A Comparison of Different Strategies for Computing Confidence Intervals of the Linkage Disequilibrium Measure

S.K. Kim, K. Zhang, and F. Sun

Pacific Symposium on Biocomputing 9:128-139(2004)

A COMPARISON OF DIFFERENT STRATEGIES FOR COMPUTING CONFIDENCE INTERVALS OF THE LINKAGE DISEQUILIBRIUM MEASURE

S.K. KIM, K. ZHANG, F. SUN*
Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA, 90089, USA
Email: {sungkkim, kuizhang, fsun}@usc.edu

Many linkage disequilibrium (LD) measures have been used to study LD patterns and for haplotype block partitioning. We examine the properties of one of these measures, Lewontin's $\tilde{D}'$, in order to understand the dependency of its confidence interval (CI) to allele frequency and sample size as well as its applications in defining haplotype blocks. This measure and its CIs were used to partition haplotypes into blocks by Gabriel et al. [1] as well as in many other applications. Gabriel et al. [1] utilized a bootstrap approach to calculate the CI for $\tilde{D}'$. Under this method, over 1,000 bootstrap samples may be needed to obtain an accurate estimate of the CI for each pair of single nucleotide polymorphism (SNP) markers which can be very computationally intensive, particularly when many SNP markers are involved. We develop two alternative methods for calculating the CI for $\tilde{D}'$ without bootstrap: one based on the approximate variance of $\tilde{D}'$ given by Zapata et al. [2] and the other based on a maximum likelihood estimate (MLE) of $\tilde{D}'$ together with Fisher Information theory. Both methods depend on normal approximation for the estimates of $\tilde{D}'$ for large sample sizes. We assess and compare the coverage of the CIs using the three methods through extensive simulations. We define the coverage as the fraction of times the estimated CI contains the true value of $\tilde{D}'$. In general, the average coverage of the bootstrap method is less than the pre-specified coverage. When the sample size is small $(n < 100)$, the remaining two methods slightly under estimate the coverage with MLE approach having smaller standard error compared to Zapata's method. When the sample size is large $(n > 100)$, the estimated coverage from both Zapata's and MLE methods are very close to the pre-specified coverage with the MLE method having the smallest standard error among all three methods. In most typical scenarios, we recommend the use of MLE method for all sample sizes. Only under rare specific cases, would the bootstrap method be better suited for determining the CI, i.e. small sample size, at extreme allele frequencies and $-3 < D' < 0$.

1 Introduction

Linkage analyses have been successfully used to map many simple, monogenic and high penetrant diseases that obey the rules of Mendelian inheritance [3]. However, their utilities for mapping human complex diseases are limited. Recently, the

*To whom correspondence should be addressed: Fengzhu Sun, PhD, Department of Biological Sciences, University of Southern California, 1042 W. 36th Place DRB-288, Los Angeles, CA, 90089, USA. Tel: (213) 740-2413. Fax: (213) 740-2437. Email: fsun@hto.usc.edu

analysis of linkage disequilibrium (LD) patterns has been of great interest in genome-wide association studies that attempt to identify genetic variation responsible for common human diseases [4, 5, 6, 7]. Compared to traditional linkage studies, association studies based on LD have two major advantages to achieve fine scale mapping. First, only unrelated individuals need to be genotyped, which makes it feasible to survey a large number of samples. Second, LD utilizes historical recombination events, rather than just those found within a pedigree. The interest in LD patterns have been advocated by the completion of the human genome and the establishment of large single-nucleotide polymorphisms (SNPs) collections, such as identified by the SNP Consortium [8, 9].

Recent studies have revealed that the human genome can be divided into long chromosomal segments with high LD punctuated by short regions with low LD [10, 11, 12]. Gabriel et al. [1] relied on the standardized gametic disequilibrium coefficient [13], a commonly used LD measure that is not significantly influenced by allele frequencies [14, 15, 16], to identify regions with high LD. Since point estimates of are unstable for low LD, especially under conditions of extreme allele frequencies or small sample size, Gabriel et al. [1] relied on the confidence intervals (CI) for instead of the estimate of LD between any two SNPs. Gabriel et al. [1] utilized a bootstrap approach [17, 18] to calculate the CI for . Under this method, over 1,000 bootstrap samples may be required to obtain an accurate estimate of the CI of for every pair of SNP loci , which can be very time consuming, particularly when many markers are involved in the analysis.

Zapata et al. derived the approximate sampling variance of between pairs of biallelic [2] and multiallelic [19] loci via large-sample theory. Through extensive simulations with various sample sizes and allele frequencies, they determined that the asymptotic sampling distribution of generally coincides with the the oretical normal distribution [19]. Therefore, the sampling variance of provides an efficient way to compute its CI under the presumption of normal approximation. Teare et al. [20] studied the properties of the sampling distributions of using simulations. However, no comparisons have been done to compare the bootstrap and Zapata's methods in estimating the CI of .

In this paper, we propose a technique based on maximum likelihood estimation (MLE) of together with Fisher Information theory for a second approach in directly estimating the sampling variance and CI [21]. Similar to Zapata's method, it is also based on the normal approximation by large-sample theory. Therefore, both Zapata's and MLE methods drastically reduces computational costs incurred by the bootstrap. We examine and compare the coverage rates for the CI estimated by bootstrap, Zapata's, and the MLE methods under various conditions of LD, allelic

frequencies and sample size. Our study provides practical guides for choosing proper methods in computing the CI of under different circumstances.

2 Methods

2.1 The LD measure

In this paper, we only consider two biallelic loci. Suppose there are two loci, A with alleles and , and B with allele s and , respectively. Let be the frequency of haplotype $A_iB_j(i=1,2;j=1,2)$, be the frequency of allele and be the frequency of allele . If haplotypes are sampled from a population, the haplotype frequencies can be estimated as follows, , where is the number of haplotypes . If diploid individuals with genotype data are sampled, is then determined by the EM algorithm [22, 23, 24]. Let , and $\hat{q}_j = \hat{q}_{1j} + \hat{q}_{2j}(j=1,2)$. For clarity and consistency in the presentation, we always assume that the observed data is , where is the number of haplotype and is the total number of haplotypes. A natural measure of gametic disequilibrium, D, which is the difference between the observed frequency of a haplotype and its expected frequency under the assumption that the alleles at two loci segregate independently, is defined as

$$D = p_{11} - p_1 q_1 . \quad (1)$$

The LD measure, , is defined by

$$D' = \frac{D}{D_{\max}}, \quad (2)$$

where

. (3)

The quantity is the maximum value that the gametic disequilibrium parameter can achieve given the marginal frequencies of the sampled observations [13]. D,

D', and $D_{\max}$ can be estimated by using $\hat{p}_{ij}$, $\hat{p}_i$ and $\hat{q}_j$ and is denoted as $\hat{D}$, $\hat{D}'$, and $\hat{D}_{\max}$, respectively.

2.2 Haplotype Data and Genotype Data Generation

Under the assumption that the population is panmictic and given , and , the expected frequency of haplotype is . The frequencies for the other haplotypes can also be computed through , and . Then haplotypes are sampled from a multinomial distribution with parameters . Pairing two haplotypes together can subsequently generate genotypes for individuals. In our simulation, we vary the haplotype sample size from 100 to 500 and the minor allele frequencies $(p_1, q_1) = (0.2,0.2)$, $(0.2,0.4)$, and (0.4,0.4), respectively. The prespecified LD measure, , ranges from –0.9 to 0.9. For each given set of parameters, we generate 1,000 replicate sets of haplotypes or genotypes.

2.3 Estimation of the Confidence Interval and the Coverage

For each simulated sample of haplotypes or genotypes, we estimate the CI of by the bootstrap, Zapata's and MLE approach. In the bootstrap method, $\hat{D}'$ is computed for each of 1,000 simulated data sets containing the same number of haplotypes or genotypes. The upper and lower confidence limits for the CI are then determined from the empirical bootstrap distribution of $\hat{D}'$ by and quantile method, respectively. For Zapata's and MLE methods, and its asymptotic sampling variance,), are computed first. Under the asymptotic normality assumption of for large sample size , the upper and the lower confidence limits are expressed as $\sqrt{}$ and $\sqrt{}$, respectively, where is the percentile of the standard normal distribution.

The entire process is repeated for 1,000 replicate data sets and the coverage is defined as the fraction of times that the CI correctly contains the pre-specified parameter, , which is used in generating the haplotype or genotype data.

2.4 Variance Estimation of D' by MLE with haplotype data

One method of approximating the variance of an unknown parameter is through the use of Fisher Information along with MLE [21]. The log-likelihood for the observed data is expressed as:

(4)

where is a function of , and , and .

The Fisher Information matrix with respect to (D', p_1, q_1) can then be calculated and is denoted as $F(D', p_1, q_1)$. The sampling variance of $\hat{D}'$ for unknown magnitude of LD and allelic frequencies is explicitly estimated by

$$V_{MLE}(\hat{D}') \approx \left[F^{-1}(D', p_1, q_1)\right]_{11}|_{\hat{D}', \hat{p}_1, \hat{q}_1} \quad (5)$$

(i.e. the first element within the inverse of the Fisher Information matrix calculated with the MLE) [21].

2.5 Variance Estimation of D' by MLE with genotype data

In order to estimate and its variance when only genotype data is available, we modify the method described in section *2.4* by computing the likelihood of the genotype data rather than the haplotype data. The log-likelihood for the observed genotypic data $(n, n_{11}, n_{12}, n_{21}, n_{22}, x)$ is expressed as:

$$\log L(n, n_{11}, n_{12}, n_{21}, n_{22}, x \mid D', p_1, q_1) = \left(\sum_{i=1}^{2} \sum_{j=1}^{2} n_{ij} \log(p_{ij}) \right) + x \log(p_{12} p_{21} + p_{11} p_{22}) \quad (6)$$

where is again a function of , and , and . Here x denotes the number of individuals who are heterozygous at both loci, and represents the total number of correctly inferred haplotypes with $n = n_{11} + n_{12} + n_{21} + n_{22} + 2x$. Similarly, the inverse of the Fisher Information matrix gives an estimate for the variance of Lewontin's LD measure [22].

2.6 Variance Estimation by Zapata et al. [2]

Zapata et al. [2] utilized the method based on the Taylor approximation to obtain the asymptotic sampling variance of . For a large sample size, variance of the gametic disequilibrium, D, is computed as

———————————————————— . (7)

Furthermore, Zapata et al. [2] approximated the variance of $\hat{D}'$ by

$$V_{Zapata}(\hat{D}') = \left[\frac{1}{n(\hat{D}_{\max})^2}\right] * \left\{(1-|\hat{D}'|)\left[nVar(\hat{D}) - |\hat{D}'|\hat{D}_{\max}(\hat{p}_1 a + \hat{p}_2 b - 2|\hat{D}|)\right] + |\hat{D}'| f(1-f)\right\} \quad (8)$$

where and if or and if . In addition, is , , , and when is , , , and , respectively.

2.7 Adjustment for the Confidence Interval

In order to obtain the CI, certain precautionary measures are taken under various conditions. Since is the normalized value of the gametic disequilibrium , its absolute value cannot be greater than 1. When computing the CI by Zapata's and MLE methods, circumstances may arise when either of the lower or upper confidence limits using the above approaches exceeds this range. This interval does not accurately depict a complete CI, thus we suggest the following tactics to ascertain the CI under different circumstances. Let be a random variable with normal distribution .

(1) If $L = \hat{D}' - Z_{\alpha/2}\sqrt{Var(\hat{D}')} < -1$ and $U = \hat{D}' + Z_{\alpha/2}\sqrt{Var(\hat{D}')} < 1$, t he lower confidence limit is defined as -1 and the upper confidence limit is defined as the smallest of 1 and U^*, where U^* is the unique value that satisfies the equation $\Pr(-1 \le X \le U^*) = 1 - \alpha$.

(2) If $L > -1$ and $U > 1$, the upper confidence limit is defined as 1 and the lower confidence limit is defined largest of -1 and L^*, where L^* is the unique value that satisfies the equation $\Pr(L^* \le X \le 1) = 1 - \alpha$.

(3) If $L < -1$ and $U > 1$, the lower and the upper confidence limits are simply defined as -1 and 1, respectively.

When $\hat{D}' = 0$, the two loci are said to be in complete linkage equilibrium and the estimation of the sampling variance is problematic. is undefined for , thus direct calculation of the estimated variance of D' is impossible for both Zapata's and the MLE methods. We suggest the following strategy to ascertain a CI for the Zapata's method and the MLE method t aking advantage of the duality between the hypothesis testing and the estimate for CIs. Let

$$\Pr\_(D', p_1, q_1) = \sum_{nn_{11} - n_{1.}n_{.1} \le 0} \binom{n}{n_{11}, n_{12}, n_{21}, n_{22}} \prod_{i=1}^{2}\prod_{j=1}^{2} (p_{ij})^{n_{ij}}, \quad (9)$$

$$\Pr\_{+}(D', p_1, q_1) = \sum_{nn_{11} - n_{1.}n_{.1} \ge 0} \binom{n}{n_{11}, n_{12}, n_{21}, n_{22}} \prod_{i=1}^{2}\prod_{j=1}^{2} (p_{ij})^{n_{ij}}, \quad (10)$$

where $n = n_{11} + n_{12} + n_{21} + n_{22}$, $n_{1.} = n_{11} + n_{12}$, and $n_{.1} = n_{11} + n_{21}$.
is calculated by given values of and . Because we do not know the true value of and , the parameters are replaced by their MLEs, and . Let $\tilde{D}_0' > 0$ be the value of D' such that $\Pr\_(D', p_1, q_1) = \alpha/2$ and $\hat{D}_0' < 0$ be the value of D' such that $\Pr\_{+}(D', p_1, q_1) = \alpha/2$. The $1-\alpha$ CI of D' is defined as $[\hat{D}_0', \tilde{D}_0']$.

$\hat{D}'$ can also take the maximum value of 1 if any one or more of the haplotypes is never observed. When this occurs, all three methods will be unable to directly determine the CI of . Bootstrap will be subject to repeated sampling distribution of $\hat{D}' = 1$ and for all three methods, the estimated sampling variance will be equal to . We employ similar tactics used for interval estimation when $\hat{D}' = 0$. Let $\hat{D}_0'$ be the value of D' that satisfies $\Pr(n, n_{11}, n_{12}, n_{21}, n_{22} \mid D', \hat{p}_1, \hat{q}_1) = \alpha$. The $1-\alpha$ CI of D' is defined as $[\hat{D}_0', 1]$. Similar methods can be used to define the $1-\alpha$ CI of D' when $\hat{D}' = -1$.

3 Results

Table 1 gives the average estimates of and its sampling variance for bootstrap, Zapata's and MLE methods using haplotype data. The results using genotype data are similar to the results based on haplotype data. However, $V_{Zapata}(\hat{D}')$ or $V_{MLE}(\hat{D}')$ were typically larger for genotype data than that for haplotype data. Our findings remain consistent with Zapata's observation [2] pertaining to the trends of the sampling variance of under different conditions of LD and allele frequencies. All three methods displayed an increase in sampling variance with a decrease in

magnitude of | | and sample size, or at extreme allele frequencies. The bootstrap method displayed the smallest variance in most cases. The MLE method typically had larger sampling variance compared to Zapata's, but the differences were minor and diminish with an increase in the sample size.

Table 1. and its average estimated sampling variance using different methods based on haplotype data.

			The sample size of haplotype											
			100				200				500			
				average variance				average variance				average variance		
D'	p	q	average D'	Bootstrap	Zapata	MLE	average D'	Bootstrap	Zapata	MLE	average D'	Bootstrap	Zapata	MLE
-0.9	0.2	0.2	-0.905	0.0211	0.0218	0.0218	-0.902	0.0115	0.0113	0.0113	-0.899	0.0048	0.0048	0.0048
	0.2	0.4	-0.900	0.0113	0.0110	0.0110	-0.901	0.0057	0.0056	0.0056	-0.900	0.0023	0.0023	0.0023
	0.4	0.4	-0.906	0.0051	0.0051	0.0051	-0.901	0.0027	0.0027	0.0027	-0.901	0.0011	0.0011	0.0011
-0.3	0.2	0.2	-0.345	0.0803	0.0905	0.0905	-0.327	0.0470	0.0540	0.0540	-0.308	0.0221	0.0246	0.0246
	0.2	0.4	-0.306	0.0490	0.0490	0.0490	-0.303	0.0254	0.0258	0.0258	-0.299	0.0105	0.0106	0.0106
	0.4	0.4	-0.298	0.0216	0.0212	0.0212	-0.302	0.0108	0.0107	0.0107	-0.298	0.0043	0.0043	0.0043
0.3	0.2	0.2	0.338	0.0265	0.0187	0.0189	0.328	0.0092	0.0087	0.0087	0.315	0.0033	0.0033	0.0033
	0.2	0.4	0.308	0.0339	0.0292	0.0292	0.303	0.0149	0.0139	0.0139	0.294	0.0056	0.0055	0.0055
	0.4	0.4	0.331	0.0135	0.0127	0.0130	0.323	0.0061	0.0060	0.0061	0.314	0.0023	0.0023	0.0024
0.9	0.2	0.2	0.940	0.0026	0.0035	0.0039	0.931	0.0015	0.0020	0.0025	0.917	0.0007	0.0010	0.0013
	0.2	0.4	0.897	0.0075	0.0073	0.0074	0.900	0.0037	0.0037	0.0037	0.901	0.0015	0.0015	0.0015
	0.4	0.4	0.933	0.0019	0.0026	0.0031	0.923	0.0011	0.0015	0.0019	0.916	0.0005	0.0006	0.0010

Table 2 shows our coverage results under haplotype sample sizes of 100, 200, and 500. Although we used , ranging from –0.9 to 0.9, we only present combinations of minor allele frequencies at 0.2 and 0.4 and values of ±0.3 and ±0.9 with 95% CIs for illustrative purposes. When LD was high ( ) and sample size was small, both Zapata's and MLE approaches tended to overestimate the coverage rates. At a haplotype sample size of 200, the average and standard error of the coverage rates for the MLE method were found to be 0.945 and 0.0011, if we consider the full spectrum of simulated conditions. Zapata's and bootstrap method averaged 0.943 and 0.929, respectively. As sample size increased, MLE-based approximations consistently were closer to the expected coverage of 95% with the least standard error than those obtained by either Zapata's or the bootstrap methods. Despite having the highest standard error for the coverage rate, the bootstrap had better coverage when we simulate data with small sample size, extreme allele frequencies and $-0.3 < D' < 0$. Zapata's and MLE methods were most analogous to each other having the closest coverage rates to the expected

coverage, relative to the bootstrap method. To further study the subtle differences between Zapata's and MLE, we calculated the mean and the standard deviation of the CI lengths based on haplotype samplings, as shown in Table 3. Although the

Table 2. The coverage rates for 95% CI with different , allele frequencies and sample size based on haplotype data. The average coverage and its standard error for each sample size condition are listed.

			The sample size of haplotypes								
			100			200			500		
D'	p	q	Bootstrap	Zapata	MLE	Bootstrap	Zapata	MLE	Bootstrap	Zapata	MLE
-0.9	0.2	0.2	0.999	0.999	0.999	0.993	1.000	1.000	0.994	0.998	0.998
	0.2	0.4	0.993	0.999	0.999	0.988	0.992	0.992	0.904	0.974	0.974
	0.4	0.4	0.908	0.908	0.908	0.910	0.951	0.951	0.938	0.937	0.937
-0.3	0.2	0.2	0.959	0.795	0.795	0.935	0.879	0.879	0.944	0.922	0.922
	0.2	0.4	0.932	0.904	0.904	0.950	0.928	0.928	0.939	0.937	0.937
	0.4	0.4	0.954	0.946	0.946	0.951	0.947	0.947	0.960	0.959	0.959
0.3	0.2	0.2	0.932	0.932	0.933	0.949	0.952	0.952	0.942	0.958	0.959
	0.2	0.4	0.947	0.953	0.954	0.943	0.938	0.938	0.943	0.946	0.946
	0.4	0.4	0.932	0.948	0.951	0.927	0.946	0.946	0.909	0.932	0.934
0.9	0.2	0.2	0.998	1.000	1.000	0.781	0.931	0.931	0.845	0.914	0.953
	0.2	0.4	0.991	0.989	0.989	0.963	0.968	0.968	0.919	0.990	0.990
	0.4	0.4	0.875	0.965	0.965	0.836	0.987	0.987	0.810	0.890	0.924
Average			0.952	0.945	0.945	0.927	0.952	0.952	0.921	0.946	0.953
Standard Error			0.0014	0.0031	0.0031	0.0040	0.0010	0.0010	0.0032	0.0009	0.0006

Table 3. The average lengths of 95% CIs with different , allele frequencies and sample size based on haplotype data. The average length and its standard deviation for each sample size condition are listed.

			The sample size of haplotype								
			100			200			500		
			average length			average length			average length		
D'	p	q	Bootstrap	Zapata	MLE	Bootstrap	Zapata	MLE	Bootstrap	Zapata	MLE
-0.9	0.2	0.2	0.7506	1.1074	1.1074	0.4262	1.1846	1.1846	0.2503	1.2128	1.2128
	0.2	0.4	0.3985	1.2012	1.2012	0.2694	1.2723	1.2723	0.1797	0.6065	0.6065
	0.4	0.4	0.2492	1.2700	1.2700	0.1916	0.8240	0.8240	0.1285	0.1603	0.1603
-0.3	0.2	0.2	1.0208	1.4338	1.4338	0.8022	0.9795	0.9795	0.5639	0.6081	0.6080
	0.2	0.4	0.8490	0.9361	0.9360	0.6170	0.6256	0.6256	0.4007	0.4023	0.4023
	0.4	0.4	0.5716	0.5684	0.5681	0.4052	0.4036	0.4034	0.2558	0.2554	0.2553
0.3	0.2	0.2	0.6157	0.5298	0.5326	0.3735	0.3639	0.3658	0.2241	0.2246	0.2258
	0.2	0.4	0.7171	0.6729	0.6739	0.4774	0.4598	0.4604	0.2932	0.2912	0.2915
	0.4	0.4	0.4563	0.4409	0.4452	0.3063	0.3039	0.3070	0.1870	0.1881	0.1901
0.9	0.2	0.2	0.2796	1.3464	1.3551	0.1456	1.2134	1.3623	0.1048	0.1685	0.3215

	0.2	0.4	0.3180	1.2763	1.2769	0.2211	1.1111	1.1118	0.1462	0.2897	0.2903
	0.4	0.4	0.1645	1.3965	1.4776	0.1241	0.5905	0.9316	0.0858	0.1003	0.1624
Average			0.5326	1.0150	1.0231	0.3633	0.7777	0.8190	0.2350	0.3756	0.3939
Standard Deviation			0.2660	0.3687	0.3767	0.2005	0.3629	0.3772	0.1360	0.3106	0.2993

Table 4. The coverage rates for 95% CI with different , allele frequencies and sample size based on genotype data. The average coverage and its standard error for each sample size condition are listed.

			The sample size of Genotype								
			50			100			250		
D'	p	q	Bootstrap	Zapata	MLE	Bootstrap	Zapata	MLE	Bootstrap	Zapata	MLE
-0.9	0.2	0.2	0.773	0.952	0.964	0.630	0.961	0.991	0.565	0.953	0.998
	0.2	0.4	0.445	0.981	1.000	0.617	0.978	0.999	0.909	0.972	0.997
	0.4	0.4	0.714	0.984	0.996	0.905	0.985	0.993	0.928	0.927	0.992
-0.3	0.2	0.2	0.896	0.725	0.777	0.941	0.758	0.783	0.942	0.768	0.886
	0.2	0.4	0.931	0.834	0.942	0.945	0.781	0.889	0.949	0.800	0.937
	0.4	0.4	0.938	0.817	0.910	0.941	0.833	0.939	0.940	0.851	0.946
0.3	0.2	0.2	0.931	0.886	0.939	0.940	0.883	0.939	0.946	0.895	0.946
	0.2	0.4	0.951	0.861	0.972	0.945	0.837	0.946	0.954	0.849	0.940
	0.4	0.4	0.923	0.849	0.919	0.939	0.868	0.939	0.944	0.860	0.940
0.9	0.2	0.2	0.879	1.000	1.000	0.806	0.949	1.000	0.823	0.923	0.926
	0.2	0.4	0.592	0.988	0.998	0.834	0.997	0.998	0.937	0.987	0.997
	0.4	0.4	0.833	0.975	0.975	0.783	0.989	1.000	0.841	0.911	0.911
Average			0.817	0.904	0.949	0.852	0.902	0.951	0.890	0.891	0.951
Standard Error			0.0411	0.0092	0.0036	0.0232	0.0091	0.0038	0.0149	0.0076	0.0013

Table 5. The average lengths and its standard deviation of 95% CIs under various conditions of , allele frequencies and sample size based on genotype data. Conditions were identical to Table 4.

	The sample size of Genotype								
	50			100			250		
	average length			average length			average length		
	Bootstrap	Zapata	MLE	Bootstrap	Zapata	MLE	Bootstrap	Zapata	MLE
Average	0.6575	1.1993	1.3834	0.4620	0.8745	1.0687	0.3017	0.4154	0.5746
Standard Deviation	0.3791	0.5305	0.5166	0.2881	0.4957	0.5586	0.1993	0.3708	0.5403

MLE method generally had larger CI lengths than Zapata's, it produced the least amount of variation as sample size increased. Furthermore, the bootstrap method had the shortest average and standard deviation for CIs lengths.

In many studies, genotypic data rather than haplotype data are generated. Thus, we performed the same procedures to examine the effects of genotype data to the CI

estimates and coverage rates for all three methods. Although the overall trends of performance were nearly identical, the average coverage rates from genotypic data were much less than that obtained from the haplotype-based results for the bootstrap and Zapata's methods while MLE remained unaffected. Furthermore, the standard error and CI lengths proved larger for all three routines (Tables 4 and 5). Considering all simulated conditions and a genotype sample size of 250 (consisting of 500 haplotypes), the coverage rate and standard error for the MLE was 0.940 and 0.0016, respectively. Zapata's and bootstrap methods averaged 0.864 and 0.912, respectively. The MLE routine demonstrated the closest coverage rate to 95% with the smallest standard error compared to all three techniques while Zapata's performance appears to worsen evaluated against haplotype-based findings. This drop in performance may originate from errors introduced when estimating $V_{Zapata}(\hat{D}')$ after inferring the haplotype frequencies. Furthermore, the bootstrap approach had poor coverage when the magnitude of LD was high, but improved with increase in sample size. This may be due to the problem that the EM algorithm does not always find the maximum of the likelihood function.

4 Discussions

We compare three methods of estimating the CI of the commonly used normalized gametic disequilibrium measure, $\bar{}$. Aside from the bootstrap approach, we present two direct methods for determining the CI through the use of the asymptotic sampling variance of . Both later methods assume that has a normal distribution under large sample size and its sampling variance is approximated either by $V_{Zapata}(\hat{D}')$ or $V_{MLE}(\hat{D}')$. Our findings suggest that the MLE method outperforms the remaining two methods by displaying satisfactory coverage and smaller variation with respect to the length of CI and the coverage rate. However, under conditions of small sample size, extreme allele frequencies and $-0.3 < D' < 0$, it appears that the bootstrap performs best. We attribute the ill performance of Zapata's and MLE method under these conditions to high variability of $\bar{}$, as observed by Zapata *et al*. [2], and small sample size. Considering that the bootstrap method is more time consuming, we suggest using the MLE method in large-scale studies.

When the genotype data is used in estimating the CI of $\bar{}$, the trend of the performance is nearly identical with that obtained using the haplotype data. However, we notice that there are differences between the coverage rate and such differences are affected by the magnitude of the pre-specified $\bar{}$ and the allele

frequencies at the two loci. The increase in CI lengths and a reduction in average coverage rates may reflect the influence from haplotype frequency estimation.

References

1. Gabriel S.B. and Schaffner S.F., Nguyen H., et al., *Science* **296** (2002) pp. 2225-2229.
2. Zapata C., Alvarez G., Carollo C., *Am. J. Hum. Genet.* **61** (1997) pp. 771-774.
3. Hall J.M., Lee M.K., Newman B., et al., *Science* **250** (1990) pp. 1684-1689.
4. Kruglyak L., *Nat. Genet.* **22** (1999) pp. 139-144.
5. Nordborg M. and Tavaré S., *Trends Genet.* **18** (2002) pp. 83-90.
6. Risch N. and Merikangas K., *Science* **273** (1996) pp. 1516-1517.
7. Weiss K.M. and Clark A.G., *Trends Genet.* **18** (2002) pp. 19-24.
8. Sachidanandam R., Weissman D., et al., *Nature* **409** (2001) pp. 928–933.
9. Venter J.C., Adams M.D., Myers E.W., et al., *Science* **291** (2001) pp. 1304-1351.
10. Daly M.J., Rioux J.D., Schaffner S.F., et al., *Nat. Genet.* **29** (2001) pp. 229–232.
11. Jeffreys A.J., Kauppi L., Neumann R., *Nat. Genet.* **29** (2001) pp. 217-222.
12. Patil N., Berno A.J., Hinds D.A., et al., *Science* **294** (2001) pp. 1719-1723.
13. Lewontin R.C., *Genetics* **49** (1967) pp. 49-67.
14. Devlin B. and Risch N., *Genomics* **29** (1995) pp. 311-322.
15. Hedrick P.W., *Genetics* **117** (1987) pp. 331-341.
16. Lewontin R.C., *Genetics* **120** (1988) pp. 849-852.
17. Efron B., *Ann. Stat.* **7** (1979) pp. 1–26.
18. Efron B. and Tibshirani R.J., *An Introduction to the Bootstrap.* (Chapman & Hall, New York, 1993).
19. Zapata C., Alvarez G., Carollo C., *Ann. Hum. Genet.* **65** (2001) pp. 395-406.
20. Teare M.D., et al., *Ann. Hum. Genet.* **66** (2002) pp. 223-233.
21. Ferguson T.S., *A Course in Large Sample Theory.* (Chapman & Hall, London, 1996).
22. Excoffier L. and Slatkin M., *Mol. Biol. Evol.* **12** (1995) pp. 921-927.
23. Hawley M.E. and Kidd K.K., *J. Hered.* **86** (1995) pp. 409-411.
24. Long J.C., Williams R.C., Urbanek M., *Am. J. Hum. Genet.* **56** (1995) pp. 799-810.