Title: Bounds and normalization of the composite linkage disequilibrium coefficient.

(Genetic Epidemiology 2004 27:252-257)

Author: Dmitri V. Zaykin*

* GlaxoSmithKline Inc., Research Triangle Park, NC

Contact details: Dmitri Zaykin
National Institute of Environmental Health Sciences
National Institutes of Health
Research Triangle Park, NC 27709
email: zaykind@niehs.nih.gov

Running title: Composite LD bounds

Abstract

The composite linkage disequilibrium (LD) measure is often calculated for two-locus genotypic data, especially when coupling and repulsion double heterozygotes cannot be distinguished. This measure has been reported to have good statistical properties and was suggested for routine testing of LD regardless of Hardy-Weinberg equilibrium at either of two loci [Weir, 1979, Schaid, 2004]. However, the bounds for this measure have not been yet reported. These bounds are derived here as functions of one-locus genotype or allele frequencies. They provide standardized measures of composite linkage disequilibrium defined as the proportion of its maximum attainable value given observed allele or genotype frequencies.

Introduction

Calculation of the composite disequilibrium coefficient, Δ_{AB}, for measuring association between alleles *A* and *B* at two loci is a routine step in analysis of multilocus genotypic data [Weir 1979, Weir and Cockerham, 1979, Weir, 1996]. Definition of the composite coefficient is $\Delta_{AB} = P_{AB} + P_{A/B} - 2p_A p_B$, where P_{AB} is the frequency of gamete AB, $P_{A/B}$ is the joint frequency of alleles *A* and *B* at two different gametes, and p_A, p_B are the frequencies of alleles *A* and *B* at two loci [Weir 1996]. This coefficient is directly estimated from two-locus genotypic data and under random mating corresponds to the usual measure of linkage disequilibrium, $D_{AB} = P_{AB} - p_A p_B$. This is because the non-gametic frequency reaches the equilibrium $P_{A/B} = p_A p_B$ after one generation of random mating. Weir [1996] suggests the definition in which the focus is on two alleles at a time, *A*, *B*, and other alleles at these two loci are combined and collectively referred to as a, b with frequencies $(1-p_A), (1-p_B)$. Weir [1996] defines the correlation coefficient, r_{AB}^2, which is one possible normalization of Δ_{AB},

$$r_{AB} = \frac{\hat{\Delta}_{AB}}{\sqrt{(\tilde{p}_A(1-\tilde{p}_A) + \hat{D}_A)(\tilde{p}_B(1-\tilde{p}_B) + \hat{D}_B)}}$$

where $\hat{D}_A, \hat{D}_B$ are the estimates of Hardy-Weinberg disequilibrium coefficients and $\tilde{p}_A, \tilde{p}_B$ are sample allele frequencies at loci *A*,*B*. It is most useful in that it directly relates to the asymptotic chi-square test statistic with one degree of freedom for testing the hypothesis $\Delta_{AB} = 0$. The relation is $X_{AB}^2 = nr_{AB}^2$, where *n* is the number of individuals in the sample. Constraints for the frequencies of genotypes and alleles at each locus

impose bounds for the minimum and maximum values of r_{AB}^2. As the result, the range of possible values of r_{AB}^2 is smaller than (-1 to 1). In certain situations, it might be is useful to report values of Δ_{AB} as the proportion of its maximum possible value, $\Delta_{\max}$. For example, Clark et al. [2003] evaluated the distribution of this standardized measure across 4,833 SNPs in the human genome, however they obtained the bounds ($\Delta_{\max}$) numerically. The purpose of this report is to derive the bounds on Δ_{AB}. The standardized measure is defined as $\Delta_{AB} / \Delta_{\max}$ and ranges between –1 and 1. The standardization makes the new measure independent of single locus frequencies, in the sense of the range that the coefficient can take. The allele frequencies are very much part of the usual normalized LD defined in the same way, $D'_{AB} = D_{AB} / D_{\max}$, [Hedrick, 1987, Lewontin, 1988] and it would be mistaken to interpret either D'_{AB} or $\Delta_{AB} / \Delta_{\max}$ as being free of dependencies on the allele or genotype frequencies.

Statistical Methods

Using the notation of Weir [1996], the composite LD coefficient is estimated from di-locus counts and sample allele frequencies as

$$\hat{\Delta}_{AB} = \frac{1}{n}(2n_{AABB} + n_{AABb} + n_{AaBB} + \frac{1}{2}n_{AaBb}) - 2\tilde{p}_A\tilde{p}_B \qquad (1)$$

where *n* is the number of individuals in the sample, and $n_{AABB},...,n_{AaBb}$ denote di-locus sample genotype counts. One way to derive bounds for the sample value of Δ_{AB} is to use

the relation $\hat{\Delta}_{AB} = C(\mathbf{x},\mathbf{y})/(2n^2)$ where $C(\mathbf{x},\mathbf{y}) = n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$, and $\mathbf{x},\mathbf{y}$ are vectors of genotype values for two loci, re-coded as

$$x_i = \begin{cases} -1 & \text{if genotype is AA} \\ 0 & \text{if genotype is Aa} \\ 1 & \text{if genotype is aa} \end{cases} \qquad y_i = \begin{cases} -1 & \text{if genotype is BB} \\ 0 & \text{if genotype is Bb} \\ 1 & \text{if genotype is bb} \end{cases}$$

The correspondence between $\hat{\Delta}_{AB}$ and $C(\mathbf{x},\mathbf{y})$ can be shown by writing the composite LD coefficient in an alternative form:

$$\begin{aligned} \hat{\Delta}_{AB} &= \frac{1}{4}\left(\hat{\Delta}_{AB} + \hat{\Delta}_{ab} - \hat{\Delta}_{Ab} - \hat{\Delta}_{aB}\right) \\ &= \frac{1}{2n}\left(n_{AABB} + n_{aabb} - n_{AAbb} - n_{aaBB}\right) - \frac{1}{2}\left(\tilde{p}_A\tilde{p}_B + \tilde{p}_a\tilde{p}_b - \tilde{p}_A\tilde{p}_b - \tilde{p}_a\tilde{p}_B\right) \\ &= \frac{1}{2n}\left(n_{AABB} + n_{aabb} - n_{AAbb} - n_{aaBB}\right) - \frac{1}{2}\left(\tilde{p}_a - \tilde{p}_A\right)\left(\tilde{p}_b - \tilde{p}_B\right) \\ &= \frac{1}{2n}\left(n_{AABB} + n_{aabb} - n_{AAbb} - n_{aaBB}\right) - \frac{1}{2n^2}\left(n_{aa} - n_{AA}\right)\left(n_{bb} - n_{BB}\right) \end{aligned} \qquad (1)$$

Sums in $C(\mathbf{x},\mathbf{y})$ can be written in terms of the genotype counts as

$$\begin{aligned} \sum x_i y_i &= \sum_{i=1}^{n_{AABB}}(-1)(-1) + \sum_{i=1}^{n_{AAbb}}(-1)(1) + \sum_{i=1}^{n_{aaBB}}(1)(-1) + \sum_{i=1}^{n_{aabb}}(1)(1) \\ &= n_{AABB} - n_{AAbb} - n_{aaBB} + n_{aabb} \end{aligned}$$

$$\sum x_i \sum y_i = (n_{aa} - n_{AA})(n_{bb} - n_{BB})$$

Therefore, $C(\mathbf{x},\mathbf{y}) = n(n_{AABB} - n_{AAbb} - n_{aaBB} + n_{aabb}) - (n_{aa} - n_{AA})(n_{bb} - n_{BB})$. Dividing this by $2n^2$, the relation to the composite LD as defined in (1) is $\hat{\Delta}_{AB} = C(\mathbf{x},\mathbf{y})/(2n^2)$.

Sample allele frequencies are related to x_i, y_i as follows:

$$1-\tilde{p}_A=\frac{2n_{aa}+n_{Aa}}{2n}=\frac{\sum x_i+n}{2n}\Rightarrow\frac{\sum x_i}{n}=1-2\tilde{p}_A$$
$$1-\tilde{p}_B=\frac{2n_{bb}+n_{Bb}}{2n}=\frac{\sum y_i+n}{2n}\Rightarrow\frac{\sum y_i}{n}=1-2\tilde{p}_B$$

Therefore, by taking expectations of these expressions, $E(x_i)=1-2p_A$ and $E(y_i)=1-2p_B$. It can be further verified that

$$\mathrm{Cov}(x_i,y_i)=2\Delta_{AB}$$

$$\begin{aligned}E(x_i,y_i)&=\mathrm{Cov}(x_i,y_i)+E(x_i)E(y_i)\\&=2\Delta_{AB}+(1-2p_A)(1-2p_B)\\&=P_{AABB}-P_{AAbb}-P_{aaBB}+P_{aabb}\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(x_i)&=(P_{aa}+P_{AA})-(P_{aa}-P_{AA})^2\\&=2\left[P_{aa}P_{AA}+P_{Aa}(1-P_{Aa})/2\right]\\&=2\left[p_A(1-p_A)+D_A\right]\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(y_i)&=(P_{bb}+P_{BB})-(P_{bb}-P_{BB})^2\\&=2\left[P_{bb}P_{BB}+P_{Bb}(1-P_{Bb})/2\right]\\&=2\left[p_B(1-p_B)+D_B\right]\end{aligned}$$

where $D_A=P_{AA}-p_A^2$ and $D_B=P_{BB}-p_B^2$ are Hardy-Weinberg disequilibrium coefficients [Weir, 1996] and $P_{AA},...,P_{AABB}$ are frequencies of genotypes *AA*,…, *AABB*. Then

$$\mathrm{Corr}(x_i,y_i)=\frac{\Delta_{AB}}{\sqrt{\left(p_A(1-p_A)+D_A\right)\left(p_B(1-p_B)+D_B\right)}}$$

The indicator variables x_i, y_i are 1 minus the sums of those given by Weir [1979], $(x_{i1}+x_{i2})$ and $(y_{i1}+y_{i2})$. These variables keep track of the number of copies of *A* and *B* on the two gametes. Weir showed that the correlation of the sums is the same as the expression for $\mathrm{Corr}(x_i,y_i)$.

When $\hat{\Delta}_{AB} > 0$, the pairs $\{x_i = -1, y_i = -1\}$ and $\{x_i = 1, y_i = 1\}$ increase the resulting value of the cross-product, $\sum x_i y_i$, while the pairs $\{x_i = -1, y_i = 0\}, \{x_i = 0, y_i = -1\}$ do not add to the value, and the pairs $\{x_i = 1, y_i = -1\}, \{x_i = -1, y_i = 1\}$ decrease the value. This suggests a way to derive the bounds for $\hat{\Delta}_{AB}$ by finding the permutation of **x** relative to **y** that will maximize the absolute value of $\hat{\Delta}_{AB}$, or equivalently, $C(\mathbf{x}, \mathbf{y})$. These bounds are for fixed one-locus counts. The permutation should maximize the number of $\{x_i, y_i\}$ pairs with the same sign of x_i and y_i and distribute the remaining x_i with y_i in such a way that will not decrease the value of the cross-product. The second term in $C(\mathbf{x}, \mathbf{y})$, i.e. the product $\sum x_i \sum y_i$, is not affected by the way the vectors are permuted. When $\hat{\Delta}_{AB} < 0$, the pairs $\{x_i, y_i\}$ are arranged in a way that maximizes the number of pairs with the different sign for x_i and y_i.

Consider the calculation for the case $\hat{\Delta}_{AB} > 0$ in more detail. The maximum possible number of pairs of the same sign is

$$d = \min(n_{AA}, n_{BB}) + \min(n_{aa}, n_{bb}) \qquad (2)$$

which is the sum of the largest possible number of pairs $\{x_i = -1, y_i = -1\}$ and $\{x_i = 1, y_i = 1\}$. The remaining number of pairs is $(n - d)$, so potentially this is the amount by which the value of the cross-product given by (2) can be reduced. However, these pairs can be arranged so that as many values of $x_i = 0$ or $y_i = 0$ (heterozygotes) as

possible are matched in pairs with the second of the two values being different from zero. Then the overall reduction of the $\max(\hat{\Delta}_{AB})$ from the value in (2) is

$$\begin{aligned} s &= \max\left[(n-d)-(n_{Aa}+n_{Bb}),0\right] \\ &= n-d-\min\left(n-d, n_{Aa}+n_{Bb}\right) \end{aligned}$$

Therefore, the maximum value for the cross-product is $(d-s)$ and the bound for $\hat{\Delta}_{AB} > 0$ is

$$\begin{aligned} \max(\hat{\Delta}_{AB}) &= \frac{1}{2n}(d-s) - \frac{1}{2n^2}\left(n_{aa}-n_{AA}\right)\left(n_{bb}-n_{BB}\right) \\ &= \frac{d-s}{2n} - \frac{1}{2}\left(\tilde{p}_a-\tilde{p}_A\right)\left(\tilde{p}_b-\tilde{p}_B\right) \\ &= \frac{d-s}{2n} - \frac{1}{2}\left(1-2\tilde{p}_A\right)\left(1-2\tilde{p}_B\right) \end{aligned}$$

For the case $\hat{\Delta}_{AB} < 0$, the minimum is obtained the same way, but the value *d* is calculated as $d = \min(n_{AA}, n_{bb}) + \min(n_{aa}, n_{BB})$ to match $\{x_i, y_i\}$ pairs with the opposite signs. This gives the maximum absolute value:

$$\begin{aligned} \max(\hat{\Delta}_{AB}) &= -\left(\frac{-d+s}{2n} - \frac{1}{2}\left(1-2\tilde{p}_A\right)\left(1-2\tilde{p}_B\right)\right) \\ &= \frac{d-s}{2n} + \frac{1}{2}\left(1-2\tilde{p}_A\right)\left(1-2\tilde{p}_B\right) \end{aligned}$$

Putting together, the possible values of $\hat{\Delta}_{AB}$ are bounded by functions of single-locus genotype counts as

$$\left.\begin{aligned} \hat{\Delta}_{\max} &= \frac{d-s}{2n} + \frac{1}{2n^2}\left(n_{aa}-n_{AA}\right)\left(n_{bb}-n_{BB}\right) \\ \text{where } d &= \min(n_{AA}, n_{bb}) + \min(n_{aa}, n_{BB}) \end{aligned}\right\}, \quad \hat{\Delta}_{AB} < 0$$

$$\left.\begin{aligned} \hat{\Delta}_{\max} &= \frac{d-s}{2n} - \frac{1}{2n^2}\left(n_{aa}-n_{AA}\right)\left(n_{bb}-n_{BB}\right) \\ \text{where } d &= \min(n_{AA}, n_{BB}) + \min(n_{aa}, n_{bb}) \end{aligned}\right\}, \quad \hat{\Delta}_{AB} > 0 \qquad (3)$$

where $s = n - d - \min(n-d, n_{Aa} + n_{Bb})$. The second part of the bounds is a function of sample allele frequencies, $\frac{1}{2n^2}(n_{aa} - n_{AA})(n_{bb} - n_{BB}) = \frac{1}{2}(1-2\tilde{p}_A)(1-2\tilde{p}_B)$. The population value as the function of one-locus population frequencies is

$$\left.\begin{array}{l} \Delta_{\max} = \dfrac{d-s}{2} + \dfrac{1}{2}(1-2p_A)(1-2p_B) \\ \text{where } d = \min(P_{AA}, P_{bb}) + \min(P_{aa}, P_{BB}) \end{array}\right\}, \ \Delta_{AB} < 0$$

$$\left.\begin{array}{l} \Delta_{\max} = \dfrac{d-s}{2} - \dfrac{1}{2}(1-2p_A)(1-2p_B) \\ \text{where } d = \min(P_{AA}, P_{BB}) + \min(P_{aa}, P_{bb}) \end{array}\right\}, \ \Delta_{AB} > 0 \qquad (4)$$

and $s = 1 - d - \min(1-d, P_{Aa} + P_{Bb})$.

The sample standardized value with bounds as given in (3) is $\hat{\delta}'_{AB} = \hat{\Delta}_{AB} / \hat{\Delta}_{\max}$.

Comparison of Standardized Coefficients

The correspondence between the values of $\hat{D}'_{AB}$ (maximum likelihood estimator based on the HWE assumption) and the standardized composite LD in large samples should be noted. The problem with the direct comparison of $\hat{\delta}'_{AB}$ and $\hat{D}'_{AB}$ is that $\hat{D}'_{AB}$ bounds are for the fixed frequencies of alleles, whereas $\hat{\delta}'_{AB}$ bounds are for the one-locus frequencies of genotypes. In this sense, $\hat{\delta}'_{AB}$ is more comparable to r_{AB} which incorporates variances that include one-locus deviations from HWE. Numerically, $\hat{\delta}'_{AB}$ and r_{AB} coefficients are closely correlated, although $\hat{\delta}'_{AB}$ is more stretched because it can still reach extreme (-1 to 1) values even if allele frequencies at two loci are unequal. This is not necessarily a virtue

of the standardized measure ($\hat{\delta}'_{AB}$). The coefficient r_{AB} has well-defined statistical and population-genetic properties. It is useful in contexts where an allele at one locus is to be regarded as a proxy for an allele at another locus, for example during the selection of markers for association studies [e.g. Meng et al., 2003]. This requires dependency (LD) as well as the closeness of single-locus frequencies which is then reflected by large absolute values of r_{AB}.

To make a fair comparison with $\hat{D}'_{AB}$, equation (3) is modified so that the bounds are calculated as the maximum possible values given frequencies of alleles rather than genotypes. In this case the standardized composite LD can be compared to D'_{AB} directly. To obtain these bounds, the new values of $(d-s)/(2n)$ in (3) should be computed. This is done by replacing the single-locus counts in *d* with their maximum values given the counts of alleles and noticing that in this case $s = n - d$, so that $(d-s)/(2n) = (2d-n)/(2n)$. Then the bounds are

$$\left.\begin{aligned} &\hat{\Delta}_{\max} = \frac{2d-n}{2n} + \frac{1}{2}(1-2\tilde{p}_A)(1-2\tilde{p}_B) \\ &\text{where } d = \min(\frac{n_A}{2}, \frac{n_b}{2}) + \min(\frac{n_a}{2}, \frac{n_B}{2}) \end{aligned}\right\}, \quad \hat{\Delta}_{AB} < 0$$

$$\left.\begin{aligned} &\hat{\Delta}_{\max} = \frac{2d-n}{2n} - \frac{1}{2}(1-2\tilde{p}_A)(1-2\tilde{p}_B) \\ &\text{where } d = \min(\frac{n_A}{2}, \frac{n_B}{2}) + \min(\frac{n_a}{2}, \frac{n_b}{2}) \end{aligned}\right\}, \quad \hat{\Delta}_{AB} > 0$$

After expressing counts via sample frequencies, this further simplifies to

$$\hat{\Delta}_{\max} = \begin{cases} 2\min[\tilde{p}_A\tilde{p}_B,\ (1-\tilde{p}_A)(1-\tilde{p}_B)], & \hat{\Delta}_{AB} < 0 \\ 2\min[(1-\tilde{p}_A)\tilde{p}_B,\ \tilde{p}_A(1-\tilde{p}_B)], & \hat{\Delta}_{AB} > 0 \end{cases} \qquad (5)$$

Therefore, these bounds are twice the bounds for D'_{AB}.

Define the standardized coefficient as $\hat{\Delta}'_{AB} = \hat{\Delta}_{AB} / \hat{\Delta}_{\max}$, when the bounds $\hat{\Delta}_{\max}$ are computed for the fixed allele frequencies, using (5). When the population is in HWE, the two estimators are related as $\hat{\Delta}'_{AB} \approx \hat{D}'_{AB} / 2$. The reason for the $\hat{\Delta}'_{AB}$ value to be twice as small is that the composite disequilibrium is a sum of the usual LD, measured by D_{AB} plus the covariance between alleles at two different chromosomes, $D_{A/B}$. This second term is zero if the population is in HWE, however the maximum value of Δ_{AB} still needs to account for $D_{A/B}$. One can only claim that Δ_{AB} reached a certain proportion of its maximum value, without attributing that proportion to either the gametic or the non-gametic part.

The allele frequencies in $\hat{\Delta}'_{AB}$ and $\hat{D}'_{AB}$ are estimated in the same way, therefore the efficiency of these two coefficients in estimation of the population value of D'_{AB} is largely determined by the performance of the corresponding LD estimators. Schaid (2004) argued for superiority of the composite LD by examining properties of the test $H_0 : D_{AB} = 0$. Therefore it is expected that $\hat{\Delta}'_{AB}$ performs well compared to $\hat{D}'_{AB}$. When the population is not in HWE, $\hat{\Delta}'_{AB}$ is still a valid estimator for $(D'_{AB} + D'_{A/B})/2$ where the second term is the normalized $D_{A/B}$. Not unexpectedly, $\hat{D}'_{AB}$ is not an appropriate estimator. Figure 1 is an illustration of this. To create each data point on the graphs, ten possible di-locus population genotype frequencies were drawn from the Dirichlet

distribution with all ten parameters equal to ¼ prior to obtaining each multinomial sample of 500 individuals. The Dirichlet(¼,…,¼) sampling creates a nearly uniform (-1 to 1) distribution of D'_{AB} and $D'_{A/B}$ across populations from which each multinomial sample is obtained. Figure 2 is a similar plot with all population disequilibrium due to the gametic part, D_{AB}. Such populations are created by sampling four gamete frequencies from the Dirichlet(1,1,1,1) distribution and pairing them at random. The resulting distribution of D_{AB} is close to uniform on (-1 to 1). The estimator $\hat{D}'_{AB}$ was obtained by numerically solving the likelihood formed under the assumption of HWE. Such solution is feasible in the case of two loci and is preferred to the common alternative using an EM algorithm [Weir, Cockerham, 1979].

Under HWE (Figure 2) the coefficient $\hat{\Delta}'_{AB}$ is estimating half of the gametic LD term, $D'_{AB}/2$. Both figures imply that $\hat{\Delta}'_{AB}$ is performing well as an estimator of the population value, $(D'_{AB} + D'_{A/B})/2$, while $\text{abs}(\hat{D}'_{AB})$ is taking many possible values between $\text{abs}(D'_{AB} + D'_{A/B})/2$ and 1 when the population is not in equilibrium. Performance of $\hat{\Delta}'_{AB}$ further improves with the increased sample size (data not shown). Note that Figure 1 sampling of di-locus genotypes results in deviations from equilibrium at the level of two loci, which includes non-zero D_{AB}, single-locus HWE deviations, as well as the higher order disequilibria. Weir [1996] gives definitions of all corresponding coefficients. Thus, $\hat{\Delta}'_{AB}$ estimator appears to perform well even when two-locus frequencies deviate from the equilibrium conditions.

Discussion

As with the usual normalized LD coefficient, D'_{AB}, caution should be taken when values of Δ'_{AB} or δ'_{AB} are compared for two pairs of loci, or when values for the same pair of loci are compared between different populations. Similar evolutionary forces will not guarantee that normalized coefficients should attain similar values given two different sets of allele frequencies [Lewontin, 1988]. Indeed, it is the values at the bounds that are completely determined by the single-locus parameters, and values in the middle remain indeterminate. Nevertheless, the standardized coefficient can be useful in the sense of its definition – as the proportional measure of strength of association between alleles at two loci.

It is worthwhile to note that the inter-gametic coefficient $D_{A/B}$ can be non-zero if sampling is conditional on the phenotype, such as the case-control sampling. If alleles *A*, *B* are jointly predictive for the "case" category, the value $D_{A/B}$ can be non-zero among cases, even if the population value is zero. If the prevalence of the case category is *w*, the "allelic prevalence" for the allele *A* can be defined as $w_A = \sum_j P_{Aj} w_{Aj}$, where *j* indexes genotypes and P_{Aj}, w_{Aj} are the genotype frequency and its susceptibility. Suppose the case probability for the individuals carrying both alleles *A* and *B* is $w_{A,B}$. Then, among the cases, $D_{A/B} = \frac{wP_{A/B} w_{A,B} - p_A p_B w_A w_B}{w^2}$ which is non-zero if $w_{A,B} \neq w_A w_B$. Similarly, D_{AB} in cases can be away from the population value. The composite coefficient as well

as its standardized value can be examined in cases and compared to disequilibrium in controls or the population value. On the other hand, when the gametic phase in unknown, the likelihood (with HWE assumption) used for the calculation of $\hat{D}_{AB}$ and $\hat{D}'_{AB}$, e.g. by the means of EM algorithm, is generally not suitable for evaluation of disequilibrium in cases. This is because the equilibrium proportions in cases (including single-locus HWE) are likely to be distorted [Nielsen et al., 1999]. Schaid [2004] showed that in such situations the incorrect HWE assumption can lead to grossly biased results of the test $H_0 : D_{AB} = 0$, with either extremely conservative or liberal type-I error rates. On the other hand, tests based on the composite coefficient have optimal power and maintain the correct size.

References

Clark AG, Nielsen R, Signorovitch, J, Matise TC, Glanowski S, Heil J, Winn-Deen ES, Holden AL, Lai E. 2003. Linkage disequilibrium and inference of ancestral recombination in 538 single-nucleotide polymorphism clusters across the human genome. Am J Hum Genet 73:285—300.

Hedrick PH. 1987. Gametic disequilibrium measures: proceed with caution. Genetics 117:331—341.

Lewontin RC. 1988. On measures of gametic disequilibrium. Genetics 120:849—852.

Meng Z, Zaykin DV, Xu CF, Wagner M, Ehm MG. 2003 Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes Am J Hum Genet 73: 115—130.

Nielsen DM, Ehm MG, Weir BS. 1999. Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. Am J Human Genet. 63:1531—1540.

Schaid DJ 2004. Linkage disequilibrium testing when linkage phase is unknown. Genetics 166: 505—512.

Weir BS. 1996. Genetic Data Analysis II, Sinauer Associates, Inc.

Weir BS. 1979. Inferences about linkage disequilibrium. Biometrics 35: 235—254.

Weir BS, Cockerham CC. 1979. Estimation of linkage disequilibrium in randomly mating populations. Heredity 42:105—111.

Acknowledgements

Two reviewers provided useful comments that improved quality of this manuscript.

Figure legends

Figure 1. Left figure: plot of the population value of $(D'_{AB} + D'_{A/B})/2$ vs. the sample value of $\hat{\Delta}'_{AB}$. Right figure: plot of the population value of $(D'_{AB} + D'_{A/B})/2$ vs. the sample value of $\hat{D}'_{AB}$.

Figure 2. Left figure: plot of the population value of D'_{AB} vs. the sample value of $\hat{\Delta}'_{AB}$. Right figure: plot of the population value of D'_{AB} vs. the sample value of $\hat{D}'_{AB}$.

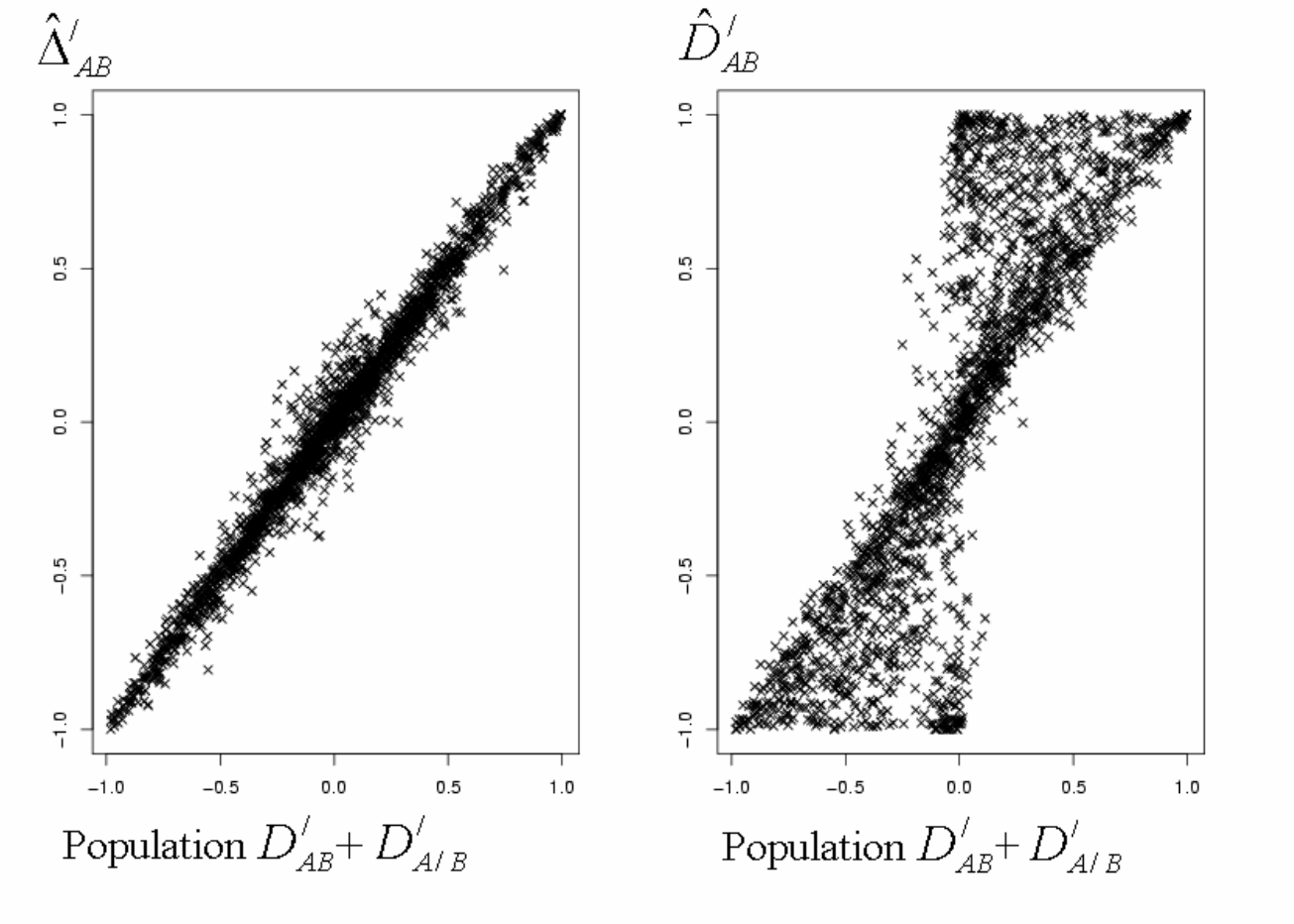
$\hat{\Delta}'_{AB}$
$\hat{D}'_{AB}$
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
Population $D'_{AB}+D'_{A/B}$
Population $D'_{AB}+D'_{A/B}$

Figure 1.

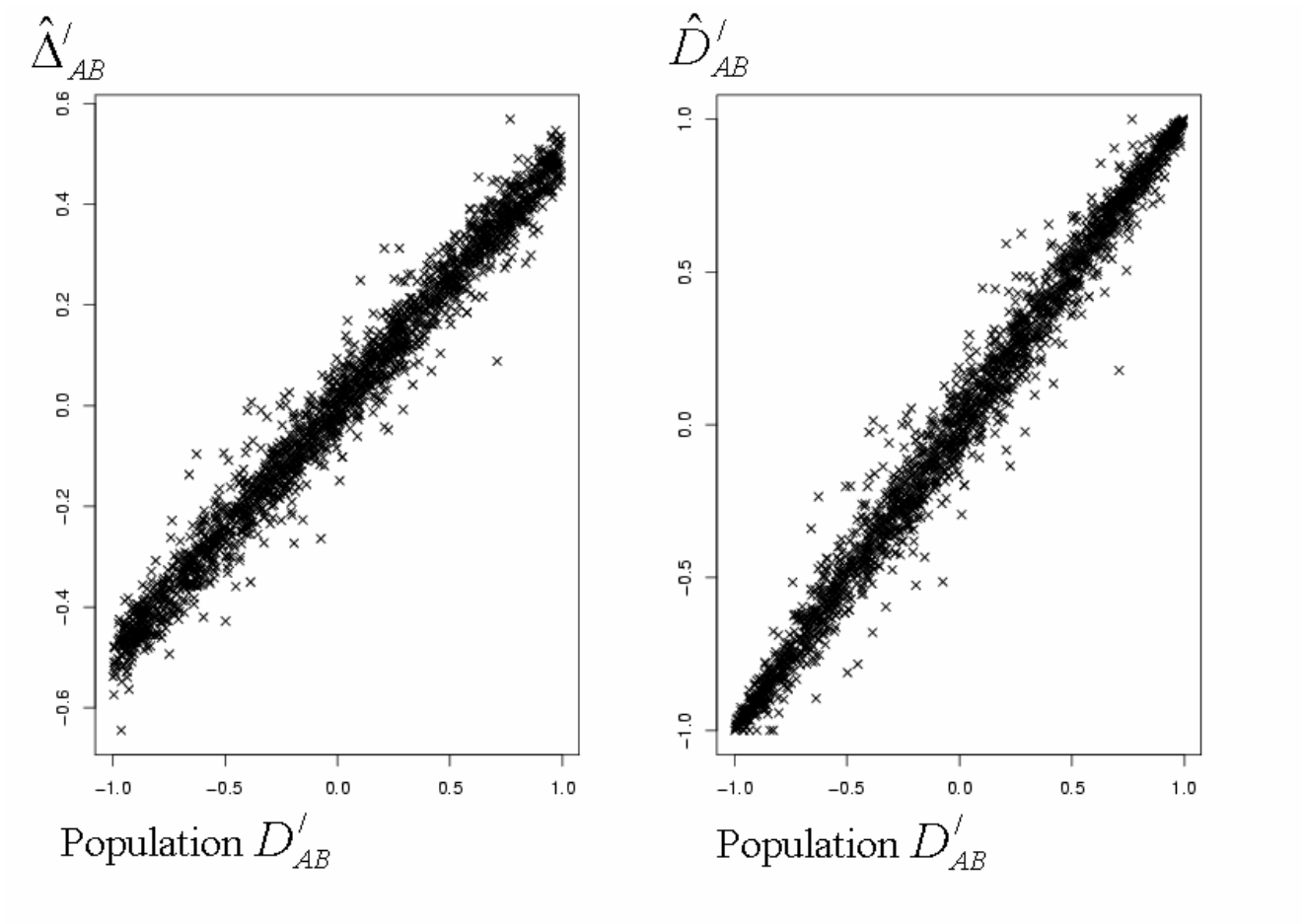
$\hat{\Delta}'_{AB}$
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−1.0
−0.5
0.0
0.5
1.0
Population D'_{AB}
$\hat{D}'_{AB}$
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
Population D'_{AB}

Figure 2.