Genetic Epidemiology 35 : 487–498 (2011)

Contrasting Linkage Disequilibrium as a Multilocus Family-Based Association Test

Zhaoxia Yu[1]* and Shuang Wang[2]

[1]*Department of Statistics, University of California, Irvine, California*
[2]*Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, New York*

Linkage disequilibrium (LD) of genetic loci is routinely estimated and graphically illustrated in genetic association studies. It has been suggested that the information in LD is also useful for association mapping and genetic association can be detected by comparing LD patterns between cases and controls. Here, we extend this idea to analyze case-parents data by comparing LD patterns between transmitted and nontransmitted genotypes. We provide the condition when contrasting LD is valid for testing gene-gene interactions. A permutation procedure is given to assess statistical significance. One advantage of our proposed methods is that haplotype information is not required. Thus, the implementation of our methods is straightforward and the resulted tests are free from potential bias caused by assumptions made to estimate haplotypes in silico. Since our test statistics use pairwise LD measurements, they are less affected by missing data than many other multilocus methods. With simulated data, we demonstrate that examining LD patterns of case-parents data is a useful multilocus association mapping strategy and it complements existing association mapping methods. The application of our methods to a Crohn's disease data set shows that our methods can detect multilocus association that might be missed by other association methods. Our permutation procedure can also be modified to allow multiple offspring from a family to be analyzed. *Genet. Epidemiol.* 35:487–498, 2011.

Key words: composite linkage disequilibrium; multilocus analysis; nuclear family; case-parents; gene-gene interaction; association mapping

Contract grant sponsor: NIH; Contract grant number: R01 HG004960.
*Correspondence to: Zhaoxia Yu, Department of Statistics, University of California, Irvine, CA 92697. E-mail: yu.zhaoxia@ics.uci.edu
Received 14 November 2010; Revised 20 April 2011; Accepted 24 April 2011
Published online 18 July 2011 in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.20598

INTRODUCTION

Testing the association between a disease and a genetic region is critical in both candidate gene studies and genome scans. In a case-control study, the distribution of genotypes of cases is compared to that of controls. Although case-control studies have been widely used, they are not robust against potential spurious associations caused by population stratification [Li, 1969]. In contrast, data from the case-parents design or other family-based designs, when analyzed appropriately, have the advantage of avoiding such spurious associations. For example, using the transmission/disequilibrium test (TDT) [Spielman et al., 1993; Terwilliger and Ott, 1992], case-parents data are efficient in testing both linkage and association between a disease and a single nucleotide polymorphism (SNP).

Multilocus association methods have been found more powerful than single-locus association methods in many situations [Akey et al., 2001; Allen and Satten, 2007; Clayton et al., 2004; Yu and Schaid, 2007; Zaykin et al., 2002]. A complex disease might be determined by multiple genetic loci in *cis* or in *trans*. When a disease causal locus is untyped, the combined information based on its flanking markers might provide adequate information for the untyped locus. Another motivation of multilocus analyses is to test gene-gene interactions. Since the concept of interactions between genes was introduced more than a century ago [Bateson, 1909], gene-gene interactions have been defined from a wide variety of perspectives by researchers in different scientific disciplines. Statistical interaction is often defined as deviations from multiplicity or additivity, depending on the parameterization. Although interactions defined in this way might not be consistent with biological interactions, testing such defined interactions nevertheless can help uncover disease associations that might be missed by single-locus analyses.

A variety of strategies has been proposed to jointly analyze multiple loci for case-parents or more complicated designs. Broadly speaking, these strategies can be divided into two classes, haplotype-based or genotype-based, depending on whether the specific strategy uses haplotype phase or not. In a haplotype-based analysis, a nature way to handle haplotypes that consist of a set of tightly linked loci is to treat haplotypes as alleles of a multiallelic locus [Clayton, 1999; Clayton et al., 2004; Horvath et al., 2004; Kaplan et al., 1997; Knapp and Becker, 2003; Lazzeroni and Lange, 1998; Merriman et al., 1998; Sham, 1997; Wilson, 1997; Zhao et al., 2000]. One problem of haplotype-based methods is the large number of distinct haplotypes, which results in a large number of degrees of freedom and can lead to the loss of statistical power. To avoid the power loss, different strategies have been proposed, including haplotype sharing methods [Allen and Satten, 2007;

Beckmann et al., 2005; Bourgain et al., 2000; Fan et al., 2005; Lange and Boehnke, 2004; Qian and Thomas, 2001; Van der Meulen and te Meerman, 1997; Zhang et al., 2003] and cladogram-based methods [Seltman et al., 2001]. Another problem of haplotype-based methods is that phase uncertainty often complicates haplotype-based multilocus methods. Even in a tightly linked region, the chance for a case-parents trio to have phase ambiguity increases quickly with the number of SNPs analyzed. One way to avoid the unknown phase is to estimate haplotypes using observed unphased genotype data [Clayton, 1999]. However, algorithms developed to infer haplotypes usually make assumptions, such as the Hardy-Weinberg equilibrium (HWE), at the population level. As a result, those methods might not be robust against population stratification [Allen and Satten, 2007]. In addition, even if population stratification is not present, HWE may be violated in regions of genetic association. On the other hand, genotype-based multilocus methods [Chapman et al., 2003; Cordell and Clayton, 2002; Fan et al., 2005; Liang et al., 2001; Rakovski et al., 2007; Xu et al., 2006] have the advantage of not requiring phase information and in many scenarios have similar or higher power than haplotype-based methods [Chapman et al., 2003; Rakovski et al., 2007; Xu et al., 2006].

Linkage disequilibrium (LD)-based methods have been developed as a multilocus association mapping tool and can be either genotype- or haplotype-based. Yang et al. [Yang et al., 1999] proposed to test gene-gene interactions by examining LD between two independent loci in cases sampled from a random mating population. The implication of no LD between two independent SNPs in cases is interpreted in terms of haplotype penetrance parameters by Zhao et al. [2006]. Nielsen et al. [2004] proposed to contrast LD levels at two loci between cases and controls and found that their LD contrast test is powerful when the LD between functional sites and markers is weak. For multiple loci, LD patterns are often graphically displayed: the pairwise LD matrix in cases is frequently plotted and visually compared to that of a control sample or a reference sample from publically available databases such as the HapMap [The International HapMap Consortium, 2003]. Zaykin et al. [2006] suggested that the difference in pairwise LD matrices of multiple loci between cases and controls can be used as a multilocus association mapping method. In particular, to avoid the complication and potential bias resulted from estimating haplotypes in silico, they used LD measurements that can be directly estimated from unphased genotype data. Recently, several important modifications and extensions have been proposed to test association for case-control data [Pan, 2010; Wang et al., 2007, 2009; Wu et al., 2008]. However, to the best of our knowledge, no similar methods exist for family data.

Here we propose to detect association by comparing LD patterns between affected offspring and their pseudocontrols formed by nontransmitted genotypes. When there is no genetic association, the distributions of genotypes of affected offspring and nontransmitted genotypes are the same, and both are identical to the distribution of genotypes in the sampled population. Consequently, although LD in affected offspring and their pseudocontrols is shaped by factors such as evolutionary forces and data sampling scheme, the LD patterns between affected offspring and their pseudocontrols in a case-parents data set are expected to be the same when the loci under study are not associated with the disease. Since the test statistics we propose to compare LD matrices between affected offspring and their pseudocontrols are calculated from genotype data, Monte Carlo permutation procedures to evaluate significance can be implemented straightforwardly. In addition, the LD matrices used in our tests contain pairwise LD measurements. Thus, compared to other multilocus methods, power loss resulted from missing data is not a major concern.

METHODS

In a LD contrast test with the case-control design, one would first compute LD measurements for cases and controls separately, and then test whether the two LD patterns are the same. To apply this strategy to the case-parents design, we compare the affected offspring and their pseudocontrols, which are defined to be the nontransmitted genotypes in case-parents data. It is known that pseudocontrols defined in this way represent a random sample from the sampled population [Falk and Rubinstein, 1987; Spielman et al., 1993], where the sampled population can be either a single random mating population or a mixture of several random mating subpopulations with the weights depending on sampling schemes. When appropriate tests are applied, comparing affected offspring and their pseudocontrols can lead to tests that are robust against potential bias caused by population stratification. Figure 1 gives an example of a case-parents trio with an (AA,Bb) offspring, her (Aa,Bb) father, and her (Aa,BB) mother. The genotype of the pseudocontrol of the affected offspring is formed by the nontransmitted alleles, i.e., (aa,BB), as also shown in Figure 1.

This section is organized as follows. We first review LD measurements and examine the implication of no LD at two SNPs among affected subjects; we then give the test statistics to compare LD patterns between affected offspring and their pseudocontrols for multiple SNPs; finally, we present a permutation procedure to assess statistical significance for the case-parents design.

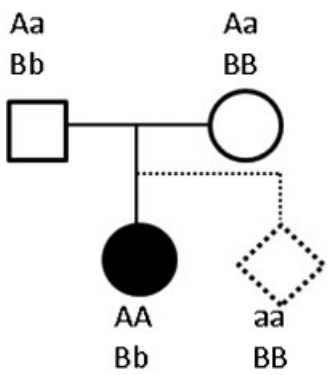

Fig. 1. The pseudocontrol of an affected offspring with (AA,Bb) genotype, (Aa,Bb) father, and (Aa,BB) mother. The solid and filled circle in the figure represents the affected offspring and the dotted diamond represents the pseudocontrol formed by the nontransmitted genotype.

CONTRASTING LD PATTERNS AT TWO SNPS IN LINKAGE EQUILIBRIUM (LE)

Consider two SNPs in LE with the first one having alleles "A" and "a" and the second one having alleles "B" and "b". To simplify our presentation, we use the following notations:

p_A: the frequency of allele "A"

p_B: the frequency of allele "B"

p_{AB}: the frequency of haplotypes carrying alleles "A" and "B"

$p_{A/B}$: the frequency that alleles "A" and "B" are on two different haplotypes

In addition, a measure with a superscript "D" indicates a measure in cases, and a measure without a superscript indicates a measure at the population level. For example, p_{AB}^D and p_{AB} are the frequencies of the haplotype with alleles "A" and "B" in cases and at the population level, respectively.

Table I summarizes genotype penetrance parameters for the nine two-locus genotypes. For example, f_{00} denotes the probability of having a disease for a subject with "aa" genotype at SNP1 and "bb" genotype at SNP2. Similar to Zhao et al. [2006], under the assumptions of random mating and LE between the two SNPs in the population, we define the penetrances of haplotypes in terms of allele frequencies and genotype penetrance parameters:

$$\begin{aligned} h_{AB} &= p_A p_B f_{22} + p_A p_b f_{21} + p_a p_B f_{12} + p_a p_b f_{11}, \\ h_{Ab} &= p_A p_B f_{21} + p_A p_b f_{20} + p_a p_B f_{11} + p_a p_b f_{10}, \\ h_{aB} &= p_A p_B f_{12} + p_A p_b f_{11} + p_a p_B f_{02} + p_a p_b f_{01}, \\ h_{ab} &= p_A p_B f_{11} + p_A p_b f_{10} + p_a p_B f_{01} + p_a p_b f_{00}, \end{aligned}$$

where h_{ij} is the penetrance of the haplotype with allele "i" at SNP1 and allele "j" at SNP2.

The LD coefficient for alleles "A" and "B" at the two SNPs is defined as

$$D = p_{AB} - p_A p_B.$$

This definition implies that haplotype frequencies are needed to calculate the LD coefficient. In practice, genetic data are usually available in the form of unphased genotype data. In this situation, one can use the composite LD coefficient [Weir, 1996], which is defined as

$$\Delta = p_{AB} + p_{A/B} - 2p_A p_B.$$

One of its standardized measures is the composite LD correlation, defined as Weir [1996]

$$r = \frac{\Delta}{\sqrt{[p_A(1-p_A)+D_A][p_B(1-p_B)+D_B]}},$$

where D_A and D_B are the Hardy-Weinberg disequilibrium coefficients at the first and second SNPs, respectively. Note that while the composite LD coefficient contains information about the correlation between the two SNPs, the Hardy-Weinberg disequilibrium coefficients are functions of parameters at individual SNPs, including heterozygote and homozygote relative risks [Wittke-Thompson et al., 2005]. In practice, it is convenient to code the genotypes of each SNP using a trinary variable. For example, at the first SNP, we can code a genotype based on the number of copies of allele "A," i.e., we code genotypes "aa," "Aa," and "AA" as 0, 1, and 2, respectively. Similarly, we can create a trinary variable for the second SNP based upon the number of copies of allele "B." It can be shown that the composite LD coefficient Δ is equal to one half of the covariance of the two trinary variables, and the composite LD correlation r is identical to the Pearson's correlation coefficient [Weir, 1979]. Specifically, if we use X_1 and X_2 to denote the trinary variables for SNP1 and SNP2, respectively, then the two LD composite measurements can be rewritten as:

$$\begin{aligned} \Delta &= \tfrac{1}{2}\mathrm{cov}(X_1, X_2) = \tfrac{1}{2}(E[X_1X_2] - E[X_1]E[X_2]), \\ r &= \mathrm{Cor}(X_1, X_2) = \frac{E[X_1X_2] - E[X_1]E[X_2]}{\sqrt{\mathrm{Var}[X_1]\mathrm{Var}[X_2]}}. \end{aligned}$$

Thus, one advantage of the composite LD measurements over other LD measurements defined on haplotype frequencies is that the haplotype phase is not required. Therefore the composite LD measurements can be rapidly computed by many standard software packages.

Under the assumption of random mating and LE between the two SNPs in the population, D, Δ, and their standardized measurements are zero. Zhao et al. [Zhao et al., 2006] used the LD coefficient among cases, denoted by D^D, as a measure of interaction. They showed that the necessary and sufficient condition for $D^D = 0$ is

$$h_{ab}h_{AB} = h_{aB}h_{Ab}.$$

Here we show in Appendix A that the necessary and sufficient condition for $\Delta^D = 0$ is the same as that for $D^D = 0$. We further show in Appendix B that the equation $h_{ab}h_{AB} = h_{aB}h_{Ab}$ indicates multiplicity of genotype penetrance parameters, i.e., the nine genotype penetrance parameters in Table I are reduced to five parameters, as shown in Table II.

Thus, the LD coefficient D, the composite LD coefficient Δ, and the composite LD correlation r can be used to test multiplicity of genotype penetrance parameters. In the literature, one definition of gene-gene interaction is the departure from multiplicative genotype risks. Therefore, when data are sampled from a random mating population where the two SNPs are in LE, both nonzero LD in cases in the case-only design [Yang et al., 1999] and differential LD between cases and controls in the case-control design [Zhao et al., 2006] provide evidence for gene-gene

TABLE I. Genotype penetrance parameters

		SNP2		
		bb	Bb	BB
SNP1	aa	f_{00}	f_{01}	f_{02}
	Aa	f_{10}	f_{11}	f_{12}
	AA	f_{20}	f_{21}	f_{22}

TABLE II. Multiplicative genotype penetrance parameters

		SNP2		
		bb	Bb	BB
SNP1	aa	f_0	$\beta_1 f_0$	$\beta_2 f_0$
	Aa	$\alpha_1 f_0$	$\alpha_1\beta_1 f_0$	$\alpha_1\beta_2 f_0$
	AA	$\alpha_2 f_0$	$\alpha_2\beta_1 f_0$	$\alpha_2\beta_2 f_0$

interaction. When case-parents trios are sampled from a random mating population and the two SNPs under study are in LE in the population, we can also test the interaction between two SNPs by comparing LD patterns between affected offspring and their pseudocontrols. Similar to Zaykin et al. [2006], to avoid the complication and potential bias caused by the estimation of the haplotype phase, we prefer to use the two composite LD measurements, i.e., Δ and r. We develop a permutation procedure (described in the subsection "Permutation procedures to assess significance") to examine whether the LD levels of affected offspring are significantly different from those of their pseudocontrols.

As will be shown in "Simulations and Results," with the conventional numerical coding (using 0, 1, and 2 to numerically code genotypes), the above LD tests as tests of gene-gene interactions are only valid when cases/affected offspring/controls are sampled from a random mating population and the two loci are in LE in the population. Violation of either of the two assumptions may lead to testing other forms of association instead. When the data are from more than one population or when the two SNPs are linked, the LD levels of affected offspring and their pseudocontrols are different even if only one of the two SNPs is associated with the disease. Thus, under the case-parents design, for two arbitrarily chosen SNPs, testing the equality of LDs between affected offspring and their pseudocontrols is not always a test of gene-gene interactions. However, when no allele or combination of alleles in a region is associated with the disease, the genotypes of affected offspring and their pseudocontrols are expected to have the same distribution. As a result, although testing equality of LD levels is not always a test of gene-gene interactions, it nevertheless provides information on whether there are some types of association between a genetic region and a disease. For case-control data, the generalization to more than two SNPs was made by Zaykin et al. [2006] and further studied by others [Pan, 2010; Wang et al., 2007, 2009; Wu et al., 2008]. In the next subsection, we present the LD contrast tests for multiple SNPs under the case-parents design.

CONTRASTING LD PATTERNS IN GENERAL (MULTIPLE SNPS)

Assume that there are S SNPs in a genomic region. We are interested in whether this region is associated with the disease. With the two aforementioned composite LD measurements, we would like to test the equality of the composite LD correlation (or coefficient) matrices between affected offspring and their pseudocontrols, i.e., H_0:$r_Y = r_N$ or H_0:$\Delta_Y = \Delta_N$, where the subscript "Y" indicates affected offspring and "N" indicates pseudocontrols. Similar to Zaykin et al. [2006], we consider the mean of the squared differences of two matrices as the test statistic. The test statistic to compare two composite LD correlation matrices is:

$$\begin{aligned}\delta_r &= \text{trace}[(\hat{r}_Y - \hat{r}_N)^T(\hat{r}_Y - \hat{r}_N)]/(S \times (S-1)) \\ &= \sum_{i=1}^{S}\sum_{j=1}^{S}(\hat{r}_{ij,Y} - \hat{r}_{ij,N})^2/(S \times (S-1)),\end{aligned}$$

where $\hat{r}_{ij,Y}$ is the estimate of the composite LD correlation between SNPs i and j within affected offspring, and $\hat{r}_{ij,N}$ is that within their pseudocontrols. We have found that the mean of the squared differences of two composite LD coefficient matrices is also a reasonable measure, and we denote it as δ_Δ:

$$\delta_\Delta = \left[\sum_{i,j}(\hat{\Delta}_{ij,Y} - \hat{\Delta}_{ij,N})^2 + \sum_{i}(\hat{\Delta}_{ii,Y} - \hat{\Delta}_{ii,N})^2\right] \Big/ (S \times (S+1)),$$

where $\hat{\Delta}_{ij,Y}$ is the estimate of the composite LD coefficient between SNPs i and j among affected offspring, and $\hat{\Delta}_{ij,N}$ is that among their pseudocontrols.

PERMUTATION PROCEDURES TO ASSESS SIGNIFICANCE

It is known that statistical inference based on asymptotic theories regarding dispersion parameters is sensitive to the departures from the assumption of normality [Boos and Brownie, 2004; Zaykin et al., 2006]. Moreover, although it is possible to develop an asymptotic test to compare the composite LD coefficients or correlations at two SNPs between affected offspring and their pseudocontrols, it is difficult to generalize it to more than two SNPs or to allow multiple offspring from a family. Therefore, we propose to use a Monte Carlo permutation procedure to assess whether an observed statistic δ_r or δ_Δ is large enough to reject the null hypothesis. Our permutation procedure is summarized as follows:

1. Compute a test statistic based on the observed data.
2. Obtain a permutated data set by randomly switching the labels of "transmitted" and "nontransmitted" with probability 1/2 within each trio.
3. Compute the test statistic for the permutated data.
4. Repeat 2 and 3 for a large number of times. The P-value is defined as the proportion of permutated statistics that are greater than or equal to the observed statistic.

A similar permutation strategy was also described by Chapman et al. [2003] and Zhao et al. [2000]. Since the randomization is performed within a trio, our permutation tests are robust against the spurious association that is caused by population stratification.

SIMULATIONS AND RESULTS

We use simulations to study the relative efficiency of different approaches. Based on the background LD of SNPs at the population level, two scenarios are considered in our simulations. In the first scenario, we consider three genetic models at two SNPs in LE. In the second scenario, we consider six tightly linked SNPs and two genetic models. For each genetic model, we use 1,000 permutations to obtain permuted P-values and 1,000 simulations to estimate power with a P-value cutoff of 0.05. All simulations are conducted using 200 trios.

TWO SNPS IN LE

For two SNPs in LE, we considered three two-locus disease models: "dominant-dominant," "threshold," and "multiplicative" models. In the first two models, the SNPs affect disease nonmultiplicatively, i.e., there is a gene-gene interaction between the two SNPs. The third model assumes that one SNP has a dominant effect on the disease and the other SNP has no effect, which is a special

multiplicative model of the two loci. The penetrance parameters for the three models are shown in Table III. For all the three models, we sample case-parents trios from a random mating population with the following haplotype frequencies: 0.04 for haplotype "AB," 0.16 for haplotype "Ab," 0.16 for haplotype "aB," and 0.64 for haplotype "ab." Note that the LD levels between transmitted and nontransmitted genotypes are equal to each other under a multiplicative model for data sampled from a random mating population. To illustrate the impact of population stratification on our tests, we also sample case-parents trios from two random mating populations with equal probabilities for the multiplicative genetic model. The haplotype frequencies of the two populations are:

Population 1: 0.04 for AB, 0.16 for Ab, 0.16 for aB, and 0.64 for ab; Population 2: 0.49 for AB, 0.21 for Ab, 0.21 for aB, and 0.09 for ab.

We compare the two LD contrast tests proposed in the article to two other commonly used tests. The first test detects gene-gene interactions under the framework of conditional logistic regressions [Cordell and Clayton, 2002; Gauderman, 2002; Wang and Zhao, 2003]. The second test, max(TDT), computes the maximum of TDT statistics at individual SNPs and uses a permutation procedure to assess significance.

Figure 2 shows the estimated power for the simulated two-SNP data. Compared to the LD composite coefficient, the LD composite correlation measure has a similar

TABLE III. Three genetic models for two SNPs

	SNP2		
SNP1	bb	Bb	BB
Dom-Dom			
Aa	f_0	λf_0	λf_0
Aa	λf_0	λf_0	λf_0
AA	λf_0	λf_0	λf_0
Threshold			
aa	f_0	f_0	f_0
Aa	f_0	f_0	λf_0
AA	f_0	λf_0	λf_0
Multiplicative[a]			
aa	f_0	f_0	f_0
Aa	λf_0	λf_0	λf_0
AA	λf_0	λf_0	λf_0

[a]The model with dominant effect at SNP1 and multiplicative effect between two SNPs.

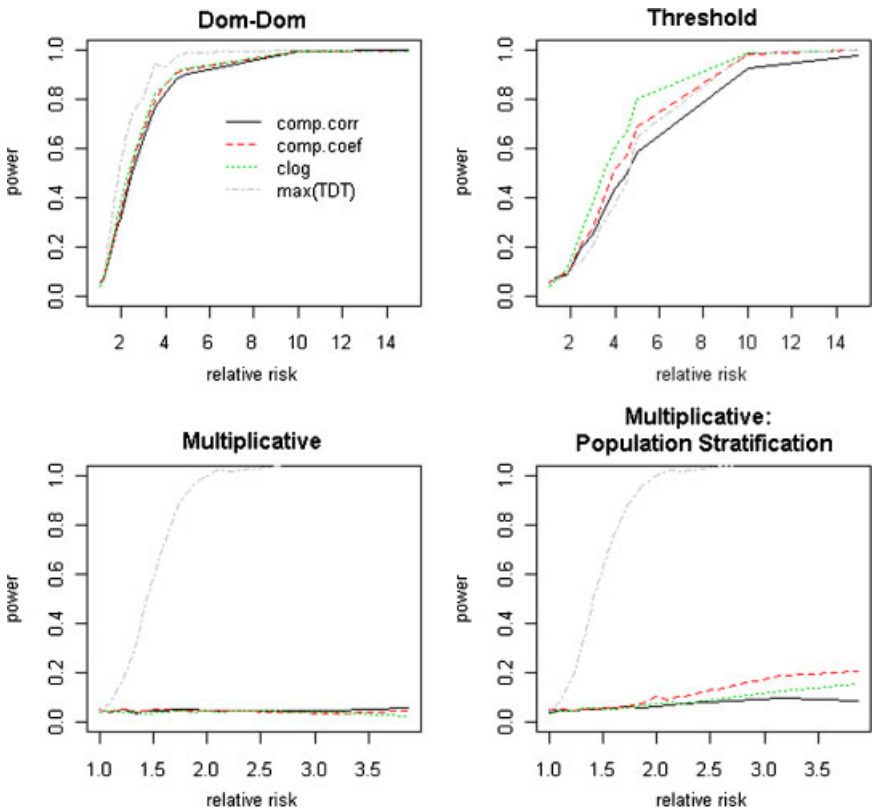

Fig. 2. Power of four tests (*comp.corr*, *comp.coef*, *clog*, and *max*(TDT)) at two independent SNPs. *comp.corr*: the test based on the composite LD correlation; *comp.coef*: the test based on the composite LD coefficient; *clog*: the test based on conditional logistic regression; *max*(TDT): the test based on the maximum TDT statistic at the two SNPs. The two upper plots and the lower left plot show results using the simulated case-parents data sampled from a random mating population. The lower right plot presents results based on data sampled from two random mating populations with distinct haplotype frequencies.

performance under the ''dominant-dominant'' model or the ''multiplicative'' model but is less powerful under the ''threshold'' model. In most situations the two LD contrast tests are less powerful than the conditional logistic regression approach. One explanation is that for two independent SNPs in the population, the conditional logistic regression compares each affected offspring to 15 possible pseudocontrols that are generated conditional on parental genotypes [Cordell and Clayton, 2002; Gauderman, 2002; Wang and Zhao, 2003], which maximizes information from data under a conditional likelihood framework. Although the conditional logistic regression tends to be more powerful, it is difficult to be generalized to study the joint effects of more than two loci.

Because the two SNPs were assumed to be independent, the null hypothesis of equal correlations is true under multiplicative genetic models if data are sampled from a random mating population. The lower left plot of Figure 2 indicates that the power of detecting gene-gene interactions, which is the Type I error rate in this situation, of the LD contrast tests and the conditional logistic regression test are all close to the nominal cutoff of 0.05. This shows that for two independent loci, both LD contrast tests and the conditional logistic regression test are valid for testing interactions when the data are sampled from a random mating population. However, the lower right plot indicates that the Type I error rates are inflated in the presence of population stratification. Note that both the LD contrast methods and the conditional logistic regression test were conducted based on the conventional numerical coding, i.e., numerical values were assigned according to the number of copies of a specific allele at each locus. This coding corresponds to assuming the multiplicative main effect model at each locus. However, in our simulation, the disease causal SNP has a dominant marginal effect, which disagrees with the way we coded genotypes. As we have shown elsewhere [Yu, 2011], when the main effects are incorrectly specified, Type I error rates might be seriously inflated in the presence of population stratification. On the other hand, when the disease is not associated with any alleles or combination of alleles (corresponding to $\lambda = 1$ in Table III), the null hypothesis of equal correlations is still true, no matter if the data are sampled from a random mating population or the two SNPs are independent. Thus, as tests of genetic association, comparing LD levels between affected offspring and their pseudocontrols are robust against population stratification.

We also found that max(TDT), the test based on the maximum TDT statistic at the two SNPs, performs well in some situations. Under the ''dominant-dominant'' model, the marginal effect of each SNP is closer to the dominant effect model than to other models. As it is known that when a SNP has a dominant effect and the frequency of the risk allele is small, the max(TDT) test is efficient. Thus, it is not surprising to see from the upper left plot of Figure 2 that max(TDT) is more powerful than the other tests. For the ''threshold'' model we found that the interactions can be more efficiently detected by one of the LD contrast tests and the conditional logistic regression approach. Because the simulations underlying the lower plots of Figure 2 assumed dominant effect at one SNP and no interactions between the two SNPs, the max(TDT) shows sufficient power. These comparisons imply that the LD contrast tests can complement individual SNP tests in the presence of gene-gene interactions and weak marginal effects. In the following, we will show that the LD contrast tests can be much more powerful than max(TDT) when multiple SNPs are jointly associated with diseases.

SIX TIGHTLY LINKED SNPS

For multiple tightly linked SNPs, we adopt the haplotype-driven simulation strategy used by [Zaykin et al., 2006] to simulate both haplotypes and phenotypes. We assume that there is no meiotic recombination event. We first generate haplotype frequencies from the Dirichlet(1,1,1,1,1,1) distribution and use them as the haplotype frequencies for a random mating population. We assume that there is a set of high-risk haplotypes H and a set of normal haplotypes $\bar{H}$, with the risk of the normal haplotypes defined as 1 and the risk of the high-risk haplotypes defined as $\lambda(\lambda > 1)$. The two genetic models we consider are:

Model 1: A heterogeneous model. In this model, the set of high-risk haplotypes H consists of two ''orthogonal'' haplotypes: H = {000000, 111111}, where 000000 is the haplotype consists of alleles ''0'' at all loci and, 111111 is the haplotype consists of alleles ''1'' at all loci. The relative risk for an offspring is determined by the number of copies of risk haplotypes. This type of interaction models have been examined by other researchers [Culverhouse et al., 2004; Nielsen et al., 2004; Zaykin et al., 2006].

Model 2: *A recessive haplotype model*. In this model, we assume that H comprises the haplotype with the largest frequency among the haplotype frequencies sampled from the Dirichlet(1,1,1,1,1,1) distribution. In addition, we assume only offspring with two copies of this haplotype have an increased risk for the disease.

It is interesting to investigate whether knowing the haplotype phase would gain power. Consider the paternal haplotypes ''pt'' (transmitted), ''pn'' (not transmitted), and the maternal haplotypes ''mt'' (transmitted), and ''mn'' (not transmitted). The randomization in the aforementioned permutation procedure for unphased genotype data is between the transmitted pair and the nontransmitted pair. Therefore, there are 2^N possible permutations for N trios. Using the known haplotype phase, we would choose one haplotype from each parent to create a permutated haplotype pair for an affected offspring in a permutation. Thus, the permutation procedure for data with known haplotype phase is similar to the one we just described, except that step 2 is modified as the following:

2. Obtain a permutated data set by randomly selecting one haplotype from each parent and label the chosen haplotype pair as ''transmitted'' and the remaining haplotype pair as ''nontransmitted'' within each trio.

This modification leads to 4^N possible permutations for N trios, which is much larger than the possible number of permutations based on genotype data. Since the haplotype phase is known in the simulated data, we can examine the difference of the power between treating data as unphased genotypes and phased haplotypes. In addition, we also compare the following three test statistics:

(1) δ_r: the mean of the squared differences of two composite LD correlation matrices;
(2) δ_Δ: the mean of the squared differences of two composite LD coefficient matrices;

(3) max(TDT): the maximum TDT statistic among the TDT tests for all SNPs.

For each of the three test statistics, the regional *P*-value is estimated by permutations. While the two LD contrast statistics lead to multilocus tests, the test based on max(TDT) is essentially a single-locus test with the family-wise error rate appropriately controlled by the permutations.

Table IV shows the results for the heterogeneity genetic model. Because we assumed the two "orthogonal"-like haplotypes are the risk haplotypes, the marginal effects of individual SNPs are small and it is not surprising to see the power of max(TDT) is low. On the other hand, contrasting LD patterns gives much higher power. Table V shows the results from simulations under the recessive haplotype model, in which only subjects carrying two copies of the risk haplotype have an increased risk. Again, contrasting LD is more powerful than max(TDT). Those results suggest that our LD contrast methods are an important complement to max(TDT).

It is interesting to know whether haplotype phase improves the power of the LD contrast tests. The estimated power, illustrated in Tables IV and V, shows the permutation procedure based on the haplotype data is more powerful than the one based on unphased genotype data. The difference, however, is typically quite small, with the maximum difference less than 2%. When haplotype data are not directly observed, it is known that using estimated haplotypes might lead to biased results if the estimation process relies on assumptions, such as the HWE, which may not hold in the sampled population. Therefore, we recommend not to estimate the haplotype phase if it is not directly observed.

TABLE IV. Estimated power under the heterogeneity model with six tightly linked SNPs

	Phase known			Phase unknown		
λ	δ_r	δ_Δ	max(TDT)	δ_r	δ_Δ	max(TDT)
1.2	0.065	0.059	0.058	0.059	0.058	0.060
1.5	0.169	0.170	0.064	0.168	0.167	0.068
1.8	0.464	0.434	0.089	0.455	0.433	0.090
2.0	0.753	0.720	0.084	0.756	0.720	0.087
2.5	0.999	0.994	0.139	0.999	0.996	0.139
3.0	1.000	1.000	0.260	1.000	1.000	0.256

TABLE V. Estimated power under the recessive haplotype model with six tightly linked SNPs

	Phase known			Phase unknown		
λ	δ_r	δ_Δ	max(TDT)	δ_r	δ_Δ	max(TDT)
1.2	0.066	0.076	0.067	0.065	0.068	0.068
1.5	0.099	0.106	0.067	0.102	0.102	0.068
1.8	0.196	0.216	0.095	0.182	0.198	0.094
2.0	0.345	0.358	0.145	0.322	0.335	0.147
2.5	0.742	0.792	0.335	0.730	0.773	0.340
3.0	0.978	0.987	0.623	0.974	0.981	0.624

With regard to the relative efficiency of the two LD measurements, our simulation results do not prefer one over the other. Under the heterogeneous model, the composite LD correlation is slightly more powerful than the composite LD coefficient, as shown in Table IV; however, under the recessive haplotype model, the composite LD correlation is less powerful than the composite LD coefficient, as shown in Table V.

A REAL EXAMPLE

We apply our methods to a real data set that was used to show genetic association between the 5q31 cytokine gene cluster and Crohn's disease [Rioux et al., 2001]. Crohn's disease is one of the two major types of inflammatory bowel diseases. Using a linkage analysis of nuclear families with inflammatory bowel disease patients, [Rioux et al., 2000] detected linkage signals on human chromosome 5q31. To narrow down the candidate region, they genotyped SNPs at 5q31 for 139 case-parents trios from the databases of the Mount Sinai Hospital and the Toronto Hospital [Rioux et al., 2001]. We downloaded the publically available subset, i.e., 129 trios genotyped at 103 common SNPs. These 103 SNPs cover a 500-kb region on 5q31. The LD structure of the 103 SNPs was first reported by Daly et al. [2001] and then re-analyzed by many others. Among the 129 trios, two of them have more than 40% of missing genotypes. We exclude those two and only analyze the remaining 127 trios.

We first evaluate the regional evidence of association based on two test statistics: (1) the mean of the squared differences of the composite LD correlation matrices between the affected offspring and their pseudocontrols (δ_r); (2) the maximum TDT statistic (max(TDT)). We also examine the composite LD correlation for all pairs of SNPs. Both the two regional *P*-values and the pairwise *P*-values are estimated using 100,000 permutations. For interesting subsets of SNPs that are selected based on pairwise *P*-values, we further increase the number of permutations to 1,000,000.

The composite LD correlation matrices for affected offspring (transmitted) and their pseudocontrols (nontransmitted) are shown in Figure 3. A visual inspection of the composite LD correlation plots reveals that the LD in the 127 affected offspring is higher than that in the nontransmitted. The mean values of the composite LD correlations in transmitted and nontransmitted genotypes are 0.26 and 0.17, respectively. The mean values of squared correlations are 0.20 and 0.16, respectively. The higher composite LD correlation on 5q31 in the affected offspring than that in their pseudocontrols indicates that affected subjects might be more similar to each other than subjects that are randomly chosen from the sampled population, which provides evidence that this region is associated with Crohn's disease. Based on 100,000 permutations, the regional *P*-value using the composite LD correlation is 4.30×10^{-4}. When the maximum TDT statistic was used as the test statistic, the regional *P*-value is 4.04×10^{-4},which is similar to the *P*-value based on the composite LD correlation. However, as will be shown in the following, contrasting LD provides much more information.

Table VI shows the information for the SNP pairs with permutated *P*-values less than 0.0001. Among the 103

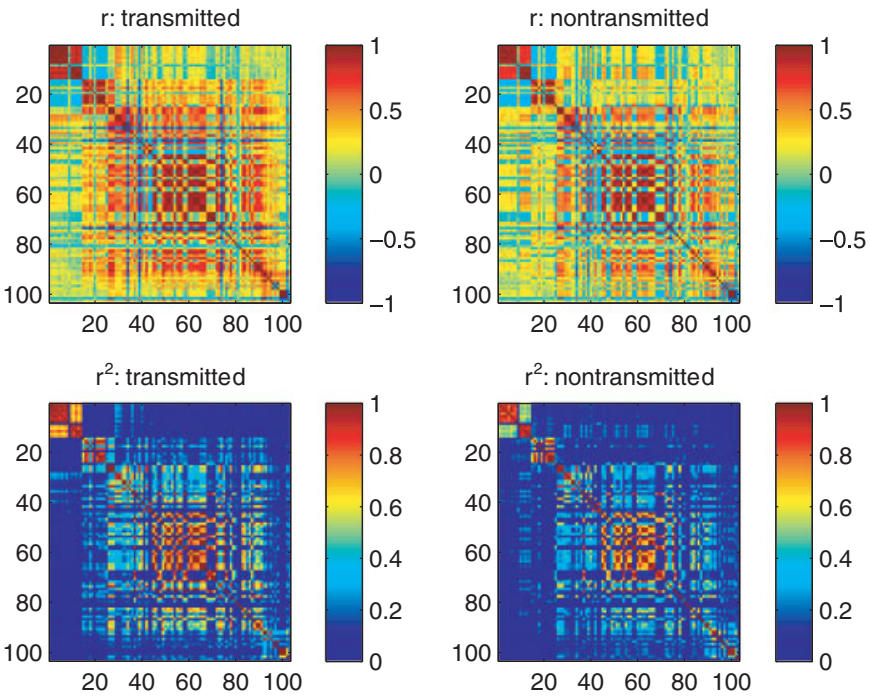

Fig. 3. Heat maps of the composite LD correlation matrices and the squared matrices computed from affected offspring (transmitted genotypes) and pseudocontrols (nontransmitted genotypes) of the Crohn data.

TABLE VI. Results of Crohn's disease data

1st snp	2nd snp	$\hat{r}_Y$	$\hat{r}_N$	p_1	p_2	Δ_r
15	91	0.66	0.19	0.023	6.70e−04	5e−05
24	91	0.62	0.10	0.049	6.70e−04	3e−05
26	47	0.37	−0.07	1.63e−05	0.239	4e−05
27	47	0.37	−0.08	5.74e−06	0.239	0e+00[a]
27	74	−0.97	−0.79	5.74e−06	3.77e−05	7e−05
28	47	0.38	−0.07	5.99e−06	0.239	0e+00[a]
28	48	0.39	−0.08	5.99e−06	0.068	7e−05
29	48	0.02	−0.40	0.094	0.068	2e−05
30	48	0.06	−0.37	0.025	0.068	2e−05
31	48	0.08	−0.33	8.83e−03	0.068	9e−05
34	74	0.97	0.79	1.93e−05	3.77e−05	5e−05
47	88	0.43	−0.14	0.239	7.59e−03	2e−05
47	91	0.42	−0.10	0.239	6.70e−04	4e−05
70	90	0.05	−0.45	0.106	0.037	1e−05
74	91	−0.92	−0.69	3.77e−05	6.70e−04	4e−05
92	98	0.59	0.02	2.38e−03	0.895	5e−05
93	98	0.56	0.08	2.31e−04	0.895	3e−05

[a]The values are truncated at zero because we used a finite number of permutations (100,000). The results for each pair are represented by the values in a row. The first two columns give the indices of paired SNPs. The third and fourth columns are the estimated composite LD correlations of the SNP pair of the affected offspring and their pseudo controls, respectively. The fifth and sixth columns are the individual TDT *P*-values of the first and second SNPs in a SNP pair, respectively. The last column gives the *P*-values based on contrasting LD pairwisely.

SNPs, the 26th, 27th, and 28th SNPs have the smallest individual *P*-values based on the TDT test. Small pairwise LD contrast *P*-values of these three SNPs and other SNPs were identified. More interestingly, there are several pairs of SNPs that have small pairwise *P*-values but in the mean time they have large individual *P*-values. For example, for the 70th and 90th SNPs, contrasting the composite LD correlations led to a *P*-value of less than 0.0001; the individual *P*-values of the two SNPs, however, are 0.106 and 0.037, respectively. As another example, the 47th SNP is not significant by itself (TDT *P*-value 0.239); however, as shown in the last column of Table VI, using the proposed LD contrast method, it has small *P*-values when paired with the 26th, 27th, 28th, 88th, or 91st SNP. We also assessed the significance of sets of multiple SNPs from the list of SNPs reported in Table VI and found that most sets show significance. As an illustration, we report the set with SNP 47 and other SNPs that show high significance (*P*-value <0.0001) with it pairwisely. Since SNPs 26, 27, and 28 are highly correlated, we considered the set of SNPs (47, 27, 88, 91). Based on 1,000,000 permutations, comparing LD matrices of the four SNPs between the affected offspring and their pseudocontrols resulted in a *P*-value of 2e-06. Because SNP 27 shows higher significance of association by itself, we then considered the triple (47, 88, 91) for contrasting LD matrices. This gave us a *P*-value of 5e-06. These results demonstrate that contrasting LD of multiple loci may reveal higher order interaction patterns that are difficult to detect by conventional methods.

The results presented here indicate that contrasting LD patterns between affected offspring and their pseudocontrols may be used as an effective association mapping tool. In particular, the LD contrast method is likely to improve the performance of individual SNP tests under many situations. However, it is worthy of noting that, because of the potential bias of multiple testing, it is

difficult to evaluate the relative efficiency of the LD contrast method and max(TDT) from this data set alone. Another issue is how to interpret interaction findings. Although statistical interactions might reflect true biological interactions, recent simulation studies show that rare causal mutations might also be responsible for the occurrence of gene-gene interaction findings [Dickson et al., 2010; Wang et al., 2010].

DISCUSSION

In this paper, we proposed a new family-based multilocus association analysis framework. The application of the proposed methods to simulated data and a real data set demonstrates that contrasting LD patterns between transmitted and nontransmitted genotypes is a complementary approach to the existing multilocus methods and the traditional one-SNP-at-a-time method. In particular, the application to Crohn's disease data shows that contrasting LD might provide information for high-order gene-gene interactions or untyped rare causal variants. Our methods are developed based on pairwise LD measurements. Therefore, the effect of missing data on our methods is less severe than on other multilocus methods. Moreover, in our methods, the haplotype phase is not required to compute the composite LD measures. This not only saves computation time but also avoids the complications resulted from estimating haplotypes.

In a region with a disease susceptibility gene, we expect that the haplotypes in cases are more similar to each other than those in controls, which is the rationale of both haplotype sharing methods and LD contrast methods. Another characteristic shared by them is that both haplotype sharing statistics and the LD coefficients can be written as quadratic forms of haplotype frequencies. Tzeng et al. [Tzeng et al., 2003] showed that three commonly used haplotype sharing measures can be written as quadratic forms of haplotype frequencies. Of particular interest is the counting measure, which is defined by the number of loci identical by state. For three SNPs, the quadratic form for the haplotype sharing statistic with the counting measure and the quadratic form for the LD coefficients are given in Appendix C. The same strategy can be applied to find quadratic forms for an arbitrary number of SNPs. The results in Appendix C demonstrate that haplotype sharing and the LD coefficients are summaries of observed data from different perspectives. While both types of statistics are quadratic forms, the matrix for haplotype sharing has zero values at the minor diagonal, while that of the LD coefficients have nonzero values at the minor diagonal. This difference makes the LD contrast methods be able to detect the association that might be missed by haplotype sharing methods. As an example, here we show a synthetic situation where haplotype sharing cannot detect the association while LD contrast methods can. Consider two SNPs and their resulted haplotypes (ab, aB, Ab, AB). Suppose that the haplotype frequencies in cases and controls (or pseudocontrols) are (0.5, 0, 0, 0.5) and (0, 0.5, 0.5, 0), respectively. It can be computed that the haplotype sharing value is 1 for both cases and controls (or pseudocontrols), but the LD coefficient in cases is 0.25 while the LD coefficient in controls (or pseudocontrols) is −0.25. In this situation, haplotype sharing methods would not be able to detect the genetic association.

The LD contrast tests we proposed are for the case-parents design. The permutation procedures can be extended to incorporate multiple siblings from each family. It is known that when analyzing nuclear family data with multiple affected siblings in a family, failing to account for the dependence of affected siblings in a family may lead to inflated Type I error rates [Martin et al., 1997]. As suggested by Martin et al. [1997], to justify the dependence of siblings in the same family, we need to permute all siblings, instead of individual siblings, in a family simultaneously. In other words, we consider each nuclear family, instead of each offspring, as an independent randomization unit. Knapp and Becker [2003] also used this randomization strategy in their haplotype-based association mapping for nuclear family data. Another issue is how to handle unaffected siblings. One strategy is to label the genotype of an unaffected offspring as "nontransmitted" and the nontransmitted genotype as "transmitted" [Guo et al., 2007]. While this approach seems appealing in maximizing the information from the data, it is controversial regarding whether analyzing unaffected siblings improves power or not [Clayton, 1999; Lunetta et al., 2000; Schaid, 2004].

Although permutation-based approaches are more computationally intensive than asymptotic tests, several strategies can be used to expedite the permutation process and make the permutation methods applicable to genome-wide data. Boehnke and Langefeld [1998] discussed ways to determine the needed number of permutations. One particularly interesting strategy they described is the sequential permutation method to achieve a prefixed accuracy for the estimation of *P*-values [Besag and Clifford, 1991]. For genome-wide data, it is neither feasible nor meaningful to contrast LD patterns across the whole genome. Instead, as an example, we can use a sliding window method and take a two-stage approach: in stage one, we use a smaller number of permutations (such as 1000) for all sliding windows; in stage two, we apply a larger number of permutations (such as 100,000) on the sliding windows with empirical *P*-values below 0.05 (estimated from the stage one). Other strategies for reducing computational cost and terminating the permutation procedure earlier might also be applied.

ACKNOWLEDGMENTS

We thank the reviewers for their helpful comments. Z.Y. was supported in part by grant NIH/R01 HG004960.

REFERENCES

Akey JM, Zhang K, Xiong MM, Doris P, Jin L. 2001. The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. Am J Hum Genet 68:1447–1456.

Allen AS, Satten GA. 2007. Statistical models for haplotype sharing in case-parent trio data. Hum Hered 64:35–44.

Bateson W. 1909. Mendel's Principles of Heredity. Cambridge: Cambridge University Press.

Beckmann L, Thomas DC, Fischer C, Chang-Claude J. 2005. Haplotype sharing analysis using mantel statistics. Hum Hered 59:67–78.

Besag J, Clifford P. 1991. Sequential Monte-Carlo P-Values. Biometrika 78:301–304.

Boehnke M, Langefeld CD. 1998. Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am J Hum Genet 62:950–961.

Boos DD, Brownie C. 2004. Comparing variances and other measures of dispersion. Stat Sci 19:571–578.

Bourgain C, Genin E, Quesneville H, Clerget-Darpoux F. 2000. Search for multifactorial disease susceptibility genes in founder populations. Ann Hum Genet 64:255–265.

Chapman JM, Cooper JD, Todd JA, Clayton DG. 2003. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered 56:18–31.

Clayton D. 1999. A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 65:1170–1177.

Clayton D, Chapman J, Cooper J. 2004. Use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol 27:415–428.

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

Cordell HJ, Clayton DG. 2002. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. Am J Hum Genet 70:124–141.

Culverhouse R, Klein T, Shannon W. 2004. Detecting epistatic interactions contributing to quantitative traits. Genet Epidemiol 27:141–152.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. 2001. High-resolution haplotype structure in the human genome. Nat Genet 29:229–232.

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. 2010. Rare variants create synthetic genome-wide associations. PLoS Biol 8:e1000294.

Falk CT, Rubinstein P. 1987. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 51:227–233.

Fan R, Knapp M, Wjst M, Zhao C, Xiong M. 2005. High resolution T association tests of complex diseases based on family data. Ann Hum Genet 69:187–208.

Gauderman WJ. 2002. Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol 155:478–484.

Guo CY, Lunetta KL, DeStefano AL, Ordovas JM, Cupples LA. 2007. Informative-transmission disequilibrium test (i-TDT): combined linkage and association mapping that includes unaffected offspring as well as affected offspring. Genet Epidemiol 31: 115–133.

Horvath S, Xu X, Lake SL, Silverman EK, Weiss ST, Laird NM. 2004. Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. Genet Epidemiol 26:61–69.

Kaplan NL, Martin ER, Weir BS. 1997. Power studies for the transmission/disequilibrium tests with multiple alleles. Am J Hum Genet 60:691–702.

Knapp M, Becker T. 2003. Family-based association analysis with tightly linked markers. Hum Hered 56:2–9.

Lange EM, Boehnke M. 2004. The haplotype runs test: the parent-parent-affected offspring trio design. Genet Epidemiol 27:118–130.

Lazzeroni LC, Lange K. 1998. A conditional inference framework for extending the transmission/disequilibrium test. Hum Hered 48:67–81.

Li CC. 1969. Population subdivision with respect to multiple alleles. Ann Hum Genet 33:23–29.

Liang KY, Hsu FC, Beaty TH, Barnes KC. 2001. Multipoint linkage-disequilibrium-mapping approach based on the case-parent trio design. Am J Hum Genet 68:937–950.

Lunetta KL, Faraone SV, Biederman J, Laird NM. 2000. Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. Am J Hum Genet 66:605–614.

Martin ER, Kaplan NL, Weir BS. 1997. Tests for linkage and association in nuclear families. Am J Hum Genet 61:439–448.

Merriman TR, Eaves IA, Twells RC, Merriman ME, Danoy PA, Muxworthy CE, Hunter KM, Cox RD, Cucca F, McKinney PA, Shield JP, Baum JD, Tuomilehto J, Tuomilehto-Wolf E, Ionesco-Tirgoviste C, Joner G, Thorsby E, Undlien DE, Pociot F, Nerup J, Ronningen KS, Bain SC, Todd JA. 1998. Transmission of haplotypes of microsatellite markers rather than single marker alleles in the mapping of a putative type 1 diabetes susceptibility gene (IDDM6). Hum Mol Genet 7:517–524.

Nielsen DM, Ehm MG, Zaykin DV, Weir BS. 2004. Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. Genetics 168:1029–1040.

Pan W. 2010. A unified framework for detecting genetic association with multiple SNPs in a candidate gene or region: contrasting genotype scores and LD patterns between cases and controls. Hum Hered 69:1–13.

Qian D, Thomas DC. 2001. Genome scan of complex traits by haplotype sharing correlation. Genet Epidemiol 21:S582–S587.

Rakovski CS, Xu X, Lazarus R, Blacker D, Laird NM. 2007. A new multimarker test for family-based association studies. Genet Epidemiol 31:9–17.

Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z, Delmonte T, Kocher K, Miller K, Guschwan S, et al. 2001. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat Genet 29:223–228.

Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, McLeod RS, Griffiths AM, Green T, Brettin TS, Stone V, Bull SB, Bitton A, Williams CN, Greenberg GR, Cohen Z, Lander ES, Hudson TJ, Siminovitch KA. 2000. Genomewide search in Canadian families with inflammatory bowel disease reveals two novel susceptibility loci. Am J Hum Genet 66:1863–1870.

Schaid DJ. 2004. Transmission disequilibrium methods for family-based studies. Mayo Clinic Technical Report (#72).

Seltman H, Roeder K, Devlin B. 2001. Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. Am J Hum Genet 68:1250–1263.

Sham P. 1997. Transmission/disequilibrium tests for multiallelic loci. Am J Hum Genet 61:774–778.

Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516.

Terwilliger JD, Ott J. 1992. A haplotype-based "haplotype relative risk" approach to detecting allelic associations. Hum Hered 42:337–346.

Tzeng JY, Devlin B, Wasserman L, Roeder K. 2003. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am J Hum Genet 72:891–902.

Van der Meulen MA, te Meerman GJ. 1997. Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. Genet Epidemiol 14:915–920.

Wang S, Zhao H. 2003. Sample size needed to detect gene-gene interactions using association designs. Am J Epidemiol 158:899–914.

Wang T, Zhu X, Elston RC. 2007. Improving power in contrasting linkage-disequilibrium patterns between cases and controls. Am J Hum Genet 80:911–920.

Wang T, Ho G, Ye K, Strickler H, Elston RC. 2009. A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped. Genet Epidemiol 33:6–15.

Wang K, Dickson SP, Stolle CA, Krantz ID, Goldstein DB, Hakonarson H. 2010. Interpretation of Association Signals and Identification of Causal Variants from Genome-wide Association Studies. Am J Hum Genet 86:730–742.

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

Weir BS. 1996. Genetic Data Analysis II: Methods for Discrete Population Genetic Data. Sunderland, MA: Sinauer Associates.

Wilson SR. 1997. On extending the transmission/disequilibrium test (TDT). Ann Hum Genet 61:151–161.

Wittke-Thompson JK, Pluzhnikov A, Cox NJ. 2005. Rational inferences about departures from Hardy-Weinberg equilibrium. Am J Hum Genet 76:967–986.

Wu X, Jin L, Xiong M. 2008. Composite measure of linkage disequilibrium for testing interaction between unlinked loci. Eur J Hum Genet 16:644–651.

Xu X, Rakovski C, Xu XP, Laird N. 2006. An efficient family-based association test using multiple markers. Genet Epidemiol 30:620–626.

Yang Q, Khoury MJ, Sun F, Flanders WD. 1999. Case-only design to measure gene-gene interaction. Epidemiology 10:167–170.

Yu Z. 2011. Testing gene-gene interactions in the case-parents design. Hum Hered. DOI: 10.1159/000327355.

Yu ZX, Schaid DJ. 2007. Sequential haplotype scan methods for association analysis. Genet Epidemiol 31:553–564.

Zaykin DV, Meng Z, Ehm MG. 2006. Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. Am J Hum Genet 78:737–746.

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. 2002. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53:79–91.

Zhang S, Sha Q, Chen HS, Dong J, Jiang R. 2003. Transmission/disequilibrium test based on haplotype sharing for tightly linked markers. Am J Hum Genet 73:566–579.

Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenauer DB, Sun F, Kidd KK. 2000. Transmission/disequilibrium tests using multiple tightly linked markers. Am J Hum Genet 67:936–946.

Zhao J, Jin L, Xiong M. 2006. Test for interaction between two unlinked loci. Am J Hum Genet 79:831–845.

APPENDIX A

Let $p_{hi/hj}$ denote the frequency of subjects with haplotype pair (h_i,h_j). According to Weir [1996, p 122],

$$\begin{aligned} p_{\mathrm{AB}} &= p_{\mathrm{AB,AB}} + 1/2(p_{\mathrm{AB,Ab}} + p_{\mathrm{AB,aB}} + p_{\mathrm{AB,ab}}), \\ p_{\mathrm{A/B}} &= p_{\mathrm{AB,AB}} + 1/2(p_{\mathrm{AB,Ab}} + p_{\mathrm{AB,aB}} + p_{\mathrm{Ab,aB}}). \end{aligned}$$

Therefore, the composite LD coefficient in cases can be rewritten to

$$\begin{aligned} \Delta^{\mathrm{D}} = {} & 2p^{\mathrm{D}}_{\mathrm{AB/AB}} + p^{\mathrm{D}}_{\mathrm{AB/Ab}} + p^{\mathrm{D}}_{\mathrm{AB/aB}} + 1/2p^{\mathrm{D}}_{\mathrm{AB/ab}} + 1/2p^{\mathrm{D}}_{\mathrm{Ab/aB}} \\ & - 2(p^{\mathrm{D}}_{\mathrm{AB}} + p^{D}_{\mathrm{Ab}})(p^{\mathrm{D}}_{\mathrm{AB}} + p^{\mathrm{D}}_{\mathrm{aB}}). \end{aligned} \tag{A1}$$

Let p_D denote the prevalence of a disease. Using the haplotype penetrance parameters defined in the main text, we have

$$p_D = p_{\mathrm{AB}}h_{\mathrm{AB}} + p_{\mathrm{Ab}}h_{\mathrm{Ab}} + p_{\mathrm{aB}}h_{\mathrm{aB}} + p_{\mathrm{ab}}h_{\mathrm{ab}}.$$

With the assumptions that the target population is a random mating population, two SNPs are in LE in the population, and no differential fertility among the phenotypes of interests, we have

$$\begin{aligned} p_D &= p_{\mathrm{A}}p_{\mathrm{B}}h_{\mathrm{AB}} + p_{\mathrm{A}}p_{\mathrm{b}}h_{\mathrm{Ab}} + p_{\mathrm{a}}p_{\mathrm{B}}h_{\mathrm{aB}} + p_{\mathrm{a}}p_{\mathrm{b}}h_{\mathrm{ab}} \\ p^{\mathrm{D}}_{\mathrm{AB/AB}} &= \frac{p^2_{\mathrm{A}}p^2_{\mathrm{B}}f_{22}}{p_D}, \quad p^{\mathrm{D}}_{\mathrm{AB/Ab}} = \frac{2p^2_{\mathrm{A}}p_{\mathrm{B}}p_{\mathrm{b}}f_{21}}{p_D}, \quad p^{\mathrm{D}}_{\mathrm{AB/aB}} = \frac{2p_{\mathrm{A}}p_{\mathrm{a}}p^2_{\mathrm{B}}f_{12}}{p_D}, \\ p^{\mathrm{D}}_{\mathrm{AB/ab}} &= p^{\mathrm{D}}_{\mathrm{Ab/aB}} = \frac{2p_{\mathrm{A}}p_{\mathrm{a}}p_{\mathrm{B}}p_{\mathrm{b}}f_{11}}{p_D}, \quad p^{\mathrm{D}}_{\mathrm{AB}} = \frac{p_{\mathrm{A}}p_{\mathrm{B}}h_{\mathrm{AB}}}{p_D}, \\ p^{\mathrm{D}}_{\mathrm{Ab}} &= \frac{p_{\mathrm{A}}p_{\mathrm{b}}h_{\mathrm{Ab}}}{p_D}, \quad p^{\mathrm{D}}_{\mathrm{aB}} = \frac{p_{\mathrm{a}}p_{\mathrm{B}}h_{\mathrm{aB}}}{p_D}, \quad p^{\mathrm{D}}_{\mathrm{ab}} = \frac{p_{\mathrm{a}}p_{\mathrm{b}}h_{\mathrm{ab}}}{p_D}. \end{aligned}$$

Plugging the above equations into (A.1) leads to

$$\begin{aligned} \Delta^{\mathrm{D}} &= \frac{2p_{\mathrm{A}}p_{\mathrm{B}}}{p_D}(p_{\mathrm{A}}p_{\mathrm{B}}f_{22} + p_{\mathrm{A}}p_{\mathrm{b}}f_{21} + p_{\mathrm{a}}p_{\mathrm{B}}f_{12} + p_{\mathrm{a}}p_{\mathrm{b}}f_{11}) \\ &\quad - \frac{2}{(p_D)^2}(p_{\mathrm{A}}p_{\mathrm{B}}h_{\mathrm{AB}} + p_{\mathrm{A}}p_{\mathrm{b}}h_{\mathrm{Ab}})(p_{\mathrm{A}}p_{\mathrm{B}}h_{\mathrm{AB}} + p_{\mathrm{a}}p_{\mathrm{B}}h_{\mathrm{aB}}) \\ &= \frac{2p_{\mathrm{A}}p_{\mathrm{B}}}{p_D}h_{\mathrm{AB}} - \frac{2}{(p_D)^2}(p_{\mathrm{A}}p_{\mathrm{B}}h_{\mathrm{AB}} + p_{\mathrm{A}}p_{\mathrm{b}}h_{\mathrm{Ab}})(p_{\mathrm{A}}p_{\mathrm{B}}h_{\mathrm{AB}} + p_{\mathrm{a}}p_{\mathrm{B}}h_{\mathrm{aB}}) \\ &= \frac{2}{(p_D)^2}[p_{\mathrm{A}}p_{\mathrm{B}}h_{\mathrm{AB}}(p_{\mathrm{A}}p_{\mathrm{B}}h_{\mathrm{AB}} + p_{\mathrm{a}}p_{\mathrm{B}}h_{\mathrm{aB}} + p_{\mathrm{A}}p_{\mathrm{b}}h_{\mathrm{Ab}} + p_{\mathrm{a}}p_{\mathrm{b}}h_{\mathrm{ab}}) \\ &\quad - (p_{\mathrm{A}}p_{\mathrm{B}}h_{\mathrm{AB}} + p_{\mathrm{A}}p_{\mathrm{b}}h_{\mathrm{Ab}})(p_{\mathrm{A}}p_{\mathrm{B}}h_{\mathrm{AB}} + p_{\mathrm{a}}p_{\mathrm{B}}h_{\mathrm{aB}})] \\ &= \frac{2p_{\mathrm{A}}p_{\mathrm{a}}p_{\mathrm{B}}p_{\mathrm{b}}}{(p_D)^2}[h_{\mathrm{ab}}h_{\mathrm{AB}} - h_{\mathrm{aB}}h_{\mathrm{Ab}}]. \end{aligned}$$

Thus $\Delta^{\mathrm{D}} = 0$ is equivalent to $h_{\mathrm{ab}}h_{\mathrm{AB}} = h_{\mathrm{aB}}h_{\mathrm{Ab}}$.

APPENDIX B

By the definition of haplotype penetrance parameters defined in the main text, we have

$$\begin{aligned} h_{\mathrm{ab}}h_{\mathrm{AB}} = {} & (p_{\mathrm{a}}p_{\mathrm{b}}f_{00} + p_{\mathrm{a}}p_{\mathrm{B}}f_{01} + p_{\mathrm{A}}p_{\mathrm{b}}f_{10} + p_{\mathrm{A}}p_{\mathrm{B}}f_{11}) \\ & \times (p_{\mathrm{a}}p_{\mathrm{b}}f_{11} + p_{\mathrm{a}}p_{\mathrm{B}}f_{12} + p_{\mathrm{A}}p_{\mathrm{b}}f_{21} + p_{\mathrm{A}}p_{\mathrm{B}}f_{22}) \\ = {} & f_{00}f_{11}p_{\mathrm{a}}p_{\mathrm{a}}p_{\mathrm{b}}p_{\mathrm{b}} + f_{01}f_{12}p_{\mathrm{a}}p_{\mathrm{a}}p_{\mathrm{B}}p_{\mathrm{B}} + f_{10}f_{21}p_{\mathrm{A}}p_{\mathrm{A}}p_{\mathrm{b}}p_{\mathrm{b}} \\ & + f_{11}f_{22}p_{\mathrm{A}}p_{\mathrm{A}}p_{\mathrm{B}}p_{\mathrm{B}} + (f_{00}f_{12} + f_{01}f_{11})p_{\mathrm{a}}p_{\mathrm{a}}p_{\mathrm{b}}p_{\mathrm{B}} \\ & + (f_{00}f_{21} + f_{10}f_{11})p_{\mathrm{a}}p_{\mathrm{A}}p_{\mathrm{b}}p_{\mathrm{b}} \\ & + (f_{00}f_{22} + f_{11}f_{11} + f_{01}f_{21} + f_{10}f_{12})p_{\mathrm{a}}p_{\mathrm{A}}p_{\mathrm{b}}p_{\mathrm{B}} \\ & + (f_{01}f_{22} + f_{11}f_{12})p_{\mathrm{a}}p_{\mathrm{A}}p_{\mathrm{B}}p_{\mathrm{B}} + (f_{10}f_{22} + f_{11}f_{21})p_{\mathrm{A}}p_{\mathrm{A}}p_{\mathrm{b}}p_{\mathrm{B}} \\ h_{\mathrm{aB}}h_{\mathrm{Ab}} = {} & (p_{\mathrm{a}}p_{\mathrm{b}}f_{01} + p_{\mathrm{a}}p_{\mathrm{B}}f_{02} + p_{\mathrm{A}}p_{\mathrm{b}}f_{11} + p_{\mathrm{A}}p_{\mathrm{B}}f_{12}) \\ & \times (p_{\mathrm{a}}p_{\mathrm{b}}f_{10} + p_{\mathrm{a}}p_{\mathrm{B}}f_{11} + p_{\mathrm{A}}p_{\mathrm{b}}f_{20} + p_{\mathrm{A}}p_{\mathrm{B}}f_{21}) \\ = {} & f_{01}f_{10}p_{\mathrm{a}}p_{\mathrm{a}}p_{\mathrm{b}}p_{\mathrm{b}} + f_{02}f_{11}p_{\mathrm{a}}p_{\mathrm{a}}p_{\mathrm{B}}p_{\mathrm{B}} + f_{11}f_{20}p_{\mathrm{A}}p_{\mathrm{A}}p_{\mathrm{b}}p_{\mathrm{b}} \\ & + f_{12}f_{21}p_{\mathrm{A}}p_{\mathrm{A}}p_{\mathrm{B}}p_{\mathrm{B}} + (f_{01}f_{11} + f_{02}f_{10})p_{\mathrm{a}}p_{\mathrm{a}}p_{\mathrm{b}}p_{\mathrm{B}} \\ & + (f_{01}f_{20} + f_{11}f_{10})p_{\mathrm{a}}p_{\mathrm{A}}p_{\mathrm{b}}p_{\mathrm{b}} \\ & + (f_{01}f_{21} + f_{12}f_{10} + f_{02}f_{20} + f_{11}f_{11})p_{\mathrm{a}}p_{\mathrm{A}}p_{\mathrm{b}}p_{\mathrm{B}} \\ & + (f_{02}f_{21} + f_{12}f_{11})p_{\mathrm{a}}p_{\mathrm{A}}p_{\mathrm{B}}p_{\mathrm{B}} + (f_{11}f_{21} + f_{12}f_{20})p_{\mathrm{A}}p_{\mathrm{A}}p_{\mathrm{b}}p_{\mathrm{B}} \end{aligned}.$$

Thus, $h_{\mathrm{ab}}h_{\mathrm{AB}} = h_{\mathrm{aB}}h_{\mathrm{Ab}}$ holds for any value of allele frequencies if and only if

$$\begin{aligned} f_{00}f_{11} = {} & f_{01}f_{10}, f_{01}f_{12} = f_{02}f_{11}, f_{01}f_{12} = f_{02}f_{11}, f_{11}f_{22} = f_{12}f_{21}, \\ & f_{00}f_{12} + f_{01}f_{11} = f_{01}f_{11} + f_{02}f_{10}, f_{00}f_{21} + f_{10}f_{11} = f_{01}f_{20} + f_{11}f_{10}, \\ & f_{00}f_{22} + f_{11}f_{11} + f_{01}f_{21} + f_{10}f_{12} = f_{01}f_{21} + f_{12}f_{10} + f_{02}f_{20} + f_{11}f_{11}, \\ & f_{01}f_{22} + f_{11}f_{12} = f_{02}f_{21} + f_{12}f_{11}, f_{10}f_{22} + f_{11}f_{21} = f_{11}f_{21} + f_{12}f_{20}. \end{aligned}$$

Let $f_0,\alpha_1,\alpha_2,\beta_1,\beta_2$ satisfy $f_{00} = f_0$, $f_{10} = f_0\alpha_1$, $f_{20} = f_0\alpha_2$, $f_{01} = f_0\beta_1$, $f_{02} = f_0\beta_2$. It is not difficult to see the above equations imply that $f_{11} = f_0\alpha_1\beta_1$, $f_{12} = f_0\alpha_1\beta_2$, $f_{21} = f_0\alpha_2\beta_1$, $f_{22} = f_0\alpha_2\beta_2$, i.e., multiplicity.

APPENDIX C

We code the two alternative alleles of a SNP as 0 or 1. To better illustrate the relationship between the haplotype sharing statistic with the counting measure and the LD coefficients in a genomic region with three SNPs, we first introduce the following haplotype ranking system. There are eight possible unique haplotypes from three SNPs. Treating each haplotype as a binary number, we order the

eight haplotypes according to their binary values in an ascending order:

Haplotype	000	001	010	011	100	101	110	111

Let $\Pi = (\pi_{000}, \pi_{001},\ldots,\pi_{111})^{\mathrm{T}}$ denote the frequencies of the haplotypes ordered in the above table. Haplotype sharing of a sample describes the amount of sharing among haplotypes in a sample. Tzeng et al. [2003] showed that commonly used haplotype sharing statistics can be written as quadratic forms $\Pi^{\mathrm{T}}A\Pi$, where $\boldsymbol{A}$ is a symmetric matrix. When the counting measure is used to define haplotype sharing, it is not difficult to see that the symmetric matrix $\boldsymbol{A}$ in the quadratic form is

$$A = \begin{pmatrix} 3 & 2 & 2 & 1 & 2 & 1 & 1 & 0 \\ 2 & 3 & 1 & 2 & 1 & 2 & 0 & 1 \\ 2 & 1 & 3 & 2 & 1 & 0 & 2 & 1 \\ 1 & 2 & 2 & 3 & 0 & 1 & 1 & 2 \\ 2 & 1 & 1 & 0 & 3 & 2 & 2 & 1 \\ 1 & 2 & 0 & 1 & 2 & 3 & 1 & 2 \\ 1 & 0 & 2 & 1 & 2 & 1 & 3 & 2 \\ 0 & 1 & 1 & 2 & 1 & 2 & 2 & 3 \end{pmatrix}.$$

Let $\pi_{ij\cdot}$ denote the frequency of the haplotype with allele i at SNP 1 and allele j at SNP 2, i.e., $\pi_{ij\cdot} = \pi_{ij0}+\pi_{ij1}$. Similarly, we can define $\pi_{i\cdot k}$ and $\pi_{\cdot jk}$. Consider the LD coefficient between SNPs 1 and 2, we have

$$\begin{aligned} D_{12} &= \pi_{00\cdot}\pi_{11\cdot} - \pi_{01\cdot}\pi_{10\cdot} \\ &= \frac{1}{2}(\pi_{00\cdot}, \pi_{01\cdot}, \pi_{10\cdot}, \pi_{11\cdot}) \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \pi_{00\cdot} \\ \pi_{01\cdot} \\ \pi_{10\cdot} \\ \pi_{11\cdot} \end{pmatrix}. \end{aligned}$$

Note that

$$(\pi_{00\cdot}, \pi_{01\cdot}, \pi_{10\cdot}, \pi_{11\cdot}) = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \Pi,$$

which leads to

$$D_{12} = \Pi^T A_{12} \Pi,$$

with

$$A_{12} = \frac{1}{2} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Similarly, we have $D_{13}=\pi_{0\cdot0}\pi_{1\cdot1} - \pi_{0\cdot1}\pi_{1\cdot0}=\Pi^{\mathrm{T}}A_{13}\Pi$, $D_{23}=\pi_{\cdot00}\pi_{\cdot11} - \pi_{\cdot01}\pi_{\cdot10}=\Pi^{T}A_{23}\Pi$, where

$$A_{13} = \frac{1}{2} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$A_{23} = \frac{1}{2} \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$