DOI: 10.1534/genetics.108.089409

Correlation-Based Inference for Linkage Disequilibrium With Multiple Alleles

Dmitri V. Zaykin,*[,1] Alexander Pudovkin[†] and Bruce S. Weir[‡]

*National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, North Carolina 27709, [†]Institute of Marine Biology, Vladivostok 690041, Russia and [‡]Department of Biostatistics, University of Washington, Seattle, Washington 98195-7232

Manuscript received March 18, 2008
Accepted for publication July 17, 2008

ABSTRACT

The correlation between alleles at a pair of genetic loci is a measure of linkage disequilibrium. The square of the sample correlation multiplied by sample size provides the usual test statistic for the hypothesis of no disequilibrium for loci with two alleles and this relation has proved useful for study design and marker selection. Nevertheless, this relation holds only in a diallelic case, and an extension to multiple alleles has not been made. Here we introduce a similar statistic, R^2, which leads to a correlation-based test for loci with multiple alleles: for a pair of loci with k and m alleles, and a sample of n individuals, the approximate distribution of $n(k-1)(m-1)/(km)R^2$ under independence between loci is $\chi^2_{(k-1)(m-1)}$. One advantage of this statistic is that it can be interpreted as the total correlation between a pair of loci. When the phase of two-locus genotypes is known, the approach is equivalent to a test for the overall correlation between rows and columns in a contingency table. In the phase-known case, R^2 is the sum of the squared sample correlations for all km 2×2 subtables formed by collapsing to one allele *vs.* the rest at each locus. We examine the approximate distribution under the null of independence for R^2 and report its close agreement with the exact distribution obtained by permutation. The test for independence using R^2 is a strong competitor to approaches such as Pearson's chi square, Fisher's exact test, and a test based on Cressie and Read's power divergence statistic. We combine this approach with our previous composite-disequilibrium measures to address the case when the genotypic phase is unknown. Calculation of the new multiallele test statistic and its P-value is very simple and utilizes the approximate distribution of R^2. We provide a computer program that evaluates approximate as well as "exact" permutational P-values.

THE phenomenon of nonrandom co-occurrence of alleles at two loci on the same haplotype is known as linkage disequilibrium (LD). It is an important population genetic concept with wide applications including theoretical studies of evolutionary dynamics (Lewontin 1974), forensic science (Evett and Weir 1998), conservation genetics and studies of effective population size (Waples 2006), evolutionary history, and human origins (Tishkoff *et al.* 1996). The extent of LD in populations has been of great interest since the development of molecular techniques allowing genotypes to be obtained at multiple loci throughout the genome. Characterization of LD in human populations has been instrumental in fine mapping of complex genetic traits in both candidate gene and whole-genome association designs. Although diallelic loci (SNPs) are utilized in most association studies, multiallelic markers (microsatellites or SNP haplotypes) will continue to be useful in genetic research, most prominently in forensic applications and studies of population size and history. Multiallelic loci provide greater precision and may yield higher power to detect and characterize LD. A simulation study by Slatkin (1994) reported an increase in power with the number of alleles to detect LD by Fisher's exact test under a finite-allele mutation model with drift and recombination. More generally, power is not a simple function of the number of alleles, as it depends on the actual disequilibria and allelic frequencies (Weir and Cockerham 1978). Formally, the LD coefficient for alleles A and B at loci **A** and **B** refers to the deviation of the joint frequency, gametic or haplotypic, from the product of allele frequencies $D_{AB} = p_{AB} - p_A p_B$. The correlation between alleles is defined as

$$\rho_{AB} = \frac{D_{AB}}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}.$$

Strictly speaking, the correlation is for the indicator variables x_A and y_B that equal 1 when the alleles are A and B and zero otherwise. This correlation coefficient has drawn much attention during recent years because the quantity $X^2_{AB} = nr^2_{AB}$, where r_{AB} is the value of ρ_{AB} in a sample of n gametes, is asymptotically distributed as $\chi^2_{(1)}$ under the hypothesis that $\rho_{AB} = 0$. This relation has obvious implications for issues of power of association studies and strategies for selecting subsets of genetic markers representative of common haplotypes for

[1]*Corresponding author:* National Institute of Environmental Health Sciences, MD A3-03, South Bldg. (101)/B356B, POB 12233, Research Triangle Park, NC 27709. E-mail: zaykind@niehs.nih.gov

genomewide analysis (Pritchard and Przeworski 2001; International Hapmap Consortium 2003; Terwilliger and Hiekkalinna 2006). However, no similar relation has been proposed for markers with more than two alleles at each locus. There is a statistical difficulty in that, beyond the two-allele case, the total squared correlation R^2 does not have a limiting chi-square distribution. Briefly, a sum of squared normal variables, $\sum Z_i^2$, has a χ^2-distribution only when the variance–covariance matrix of the Z_i's is a projection matrix. A more general result is usually stated in the matrix notation, regarding the distribution of a quadratic form, $\mathbf{Z}'\mathbf{CZ}$ (Searle 1971, Chap. 2, Theorem 2). In our case, $\mathbf{C}$ is an identity matrix. Pearson's X^2-statistic is an example of such a sum, while the sum of squared LD correlations is not. Thus, despite the vast theory on contingency tables, the distribution of R^2 has not been adopted for testing interactions. Nevertheless, different approximations by a scaled chi-square distribution are possible for a sum of dependent chi-squares (*e.g.*, Box 1954). Here we report a very simply computed chi-square approximation that appears to have good properties. This result is further applied to testing LD at a pair of multiallelic loci when only single-locus genotypes are scored unambiguously. Earlier work on characterization and testing of LD at a pair of multiallelic loci includes accounts by Hill (1975); Yamazaki (1977); Weir and Cockerham (1978); Weir (1979); Karlin and Piazza (1981); Hedrick (1987); Zaykin *et al.* (1995); Kalinowski and Hedrick (2000); Zapata (2000); Schaid (2004); and Zhao *et al.* (2005, 2007). Similar to the methods of Weir (1979) and Schaid (2004), our correlation LD approach is based on the composite disequilibrium definition. The composite disequilibrium approach has certain desirable properties. It is robust with respect to single-locus deviations from Hardy–Weinberg equilibrium (HWE). The composite disequilibrium coefficient is estimated directly from genotypic counts, and thus it is readily computed from data with the unknown gametic phase. Earlier work (Weir 1979; Schaid 2004; Zaykin 2004) demonstrated good statistical properties associated with this approach.

The correlation LD test is recommended for usage and can be readily applied for screening large numbers of pairs of multiallelic loci. It is also applicable for conducting correlation-based tests for interaction in contingency tables. Our program provides exact (permutational) *P*-values for tests based on R^2.

METHODS

Known gametic phase: When the gametic phase is unambiguous, the two-locus haplotype observations can be arranged into a $k \times m$ contingency table with the sample size N being equal to twice the number of individuals n, $N = 2n$. The cell counts in the table represent N haplotype observations: the (i, j)th cell has the number n_{ij} of haplotypes carrying allele i at the first locus and allele j at the second. We assume multinomial sampling of haplotypes. The observed haplotype frequencies are $\tilde{p}_{ij} = n_{ij}/N$. Row and column frequencies for the table of haplotype frequencies correspond to the vectors of allele frequencies at the two loci: $\{p_1, \ldots, p_k\}$ and $\{q_1, \ldots, q_m\}$. The observed correlation for the cell (i, j) is

$$r_{ij} = \frac{\tilde{p}_{ij} - \tilde{p}_i\tilde{q}_j}{\sqrt{\tilde{p}_i(1-\tilde{p}_i)\tilde{q}_j(1-\tilde{q}_i)}}. \quad (1)$$

We propose the following two correlation-based statistics, both having an approximate chi-square distribution (as shown in appendix a). The eigenvalue-based statistic is

$$T_1 = \frac{N\sum_{i=1}^{k}\sum_{j=1}^{m} r_{ij}^2}{\sigma} \overset{\text{app}}{\sim} \chi^2_{(d)}, \quad (2)$$

where

$$\sigma = \frac{\text{trace}(\mathbf{V}_R\mathbf{V}_R)}{km}$$

$$d = \frac{(km)^2}{\text{trace}(\mathbf{V}_R\mathbf{V}_R)}.$$

The statistic T_2 is much simpler, as it does not involve a computation of eigenvalues:

$$T_2 = \frac{(k-1)(m-1)N}{km}\sum_{i=1}^{k}\sum_{j=1}^{m} r_{ij}^2 \overset{\text{app}}{\sim} \chi^2_{(k-1)(m-1)}. \quad (3)$$

Unknown haplotype phase: Scoring genotypes one locus at a time creates ambiguity in determining pairs of haplotypes in individuals that are heterozygous at both loci. A maximum-likelihood solution for obtaining sample haplotype frequencies was suggested by Hill (1974, 1975) and elaborated on by Weir and Cockerham (1979). This approach was extended to multiple loci (Excoffier and Slatkin 1995) with the use of the EM algorithm incorporating the likelihood under the assumption of HWE. Weir (1979) sought to avoid making the HWE assumption and suggested estimating the composite disequilibrium defined as $\Delta_{AB} = p_{AB} + p_{A/B} - 2p_Ap_B$, where $p_{A/B}$ is the joint frequency of alleles A and B at two different gametes within individuals. The corresponding composite LD correlation is

$$\rho^c_{AB} = \frac{\Delta_{AB}}{\sqrt{[p_A(1-p_A)+D_A][p_B(1-p_B)+D_B]}}, \quad (4)$$

where D_A, D_B are the the Hardy–Weinberg disequilibrium coefficients at the two loci. Strictly speaking, this is the correlation of the number of A and B alleles carried by an individual (Weir 1979; Zaykin 2004). The composite coefficient is directly estimated from two-locus

counts by simple counting (Weir 1979). Under HWE, the intergametic disequilibrium term $D_{A/B} = p_{A/B} - p_A p_B = 0$, and the population value of $\Delta_{AB} = D_{AB}$.

The composite correlations for a pair of alleles in a multiple-allele system are

$$\rho^{c}_{ij} = \frac{\Delta_{ij}}{\sqrt{[p_i(1-p_i) + D_i][q_j(1-q_j) + D_j]}}.$$

Weir and Cockerham (1989) gave a decomposition of the two-locus genotype frequency P^{AB}_{AB} as a sum of functions of allele frequencies and two-locus disequilibria. Writing out the two-locus analog of the Hardy–Weinberg disequilibrium (HWD), $P^{AB}_{AB} - p^2_{AB}$, in these terms shows that under the two-locus HWE, only the D_{AB} and thus Δ_{AB} disequilibria are nonzero. Therefore, assuming two-locus HWE, a chi-square statistic for testing LD can be written as

$$(X^{c})^2_{AB} = n \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\tilde{\Delta}^2_{ij}}{\tilde{p}_i \tilde{q}_j}, \tag{5}$$

as was suggested by Weir (1979). Under HWE, the composite coefficient estimates the usual LD. On the basis of Fisher's formula for approximate variances, Schaid (2004) derived the covariance matrix of the sample LD coefficients (**W**). He proposed a chi-square test based on a quadratic form. The test statistic definition involves a generalized inverse, $\mathbf{W}^{-}$. This test is analogous to (19). For the vector containing all sample composite LD coefficients $\mathbf{\Delta}^T = \{\hat{\Delta}_{ij}\}$, Schaid's test statistic, $S^2 = \mathbf{\Delta}^T \mathbf{W}^{-} \mathbf{\Delta}$, has an asymptotic chi-square distribution with the degrees of freedom equal to the rank of **W**. Schaid's test explicitly incorporates deviations from HWE.

We base the unknown-phase extension of the correlation LD approach on the approximate sampling distribution of the total composite LD correlation,

$$\begin{aligned}(R^{c})^2 &= \sum_{i=1}^{k} \sum_{j=1}^{m} (r^{c})^2_{ij} \\ &= \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\hat{\Delta}^2_{ij}}{[\tilde{p}_i(1-\tilde{p}_i) + \tilde{D}_i][\tilde{q}_j(1-\tilde{q}_j) + \tilde{D}_j]},\end{aligned} \tag{6}$$

where $(r^{c})^2_{ij}$ denotes sample values of $(\rho^{c})^2_{ij}$. Comparing this statistic to (5) shows that now the deviations from HWE at both loci are explicitly incorporated into the test.

Schaid's test statistic as well as $(R^{c})^2$ assumes that trigenic and quadrigenic two-locus disequilibria can be ignored. These disequilibria compare joint frequencies of three and four alleles at two loci with the products of allele frequencies, after removing any lower-order disequilibria (Weir 1996). To obtain the Box-type approximation (for the statistic T_1), the elements of the matrix **W** are scaled as $\{W_{ij}/\sqrt{W_{ii}W_{jj}}\}$. This gives the correlation matrix $\mathbf{W}_R$. As before, the scale parameter is $\sigma = \text{trace}(\mathbf{W}_R\mathbf{W}_R)/(km)$, and the degrees of freedom are $d = (km)^2/\text{trace}(\mathbf{W}_R\mathbf{W}_R)$. Then the two statistics with their approximate distributions are

$$T_1 = \frac{n(R^{c})^2}{\sigma} \overset{\text{app}}{\sim} \chi^2_{(d)} \tag{7}$$

$$T_2 = \frac{(k-1)(m-1)}{km} n(R^{c})^2 = (k-1)(m-1)n\bar{r}^2 \overset{\text{app}}{\sim} \chi^2_{(k-1)(m-1)}, \tag{8}$$

where $\bar{r}^2 = (R^{c})^2/(km)$ is the average composite correlation.

Type-I error rates, goodness of fit to the null distribution, and power: A common way to evaluate a test performance under the null hypothesis is to report the type-I error, or the proportion of *P*-values that fall below a rejection threshold, such as $\alpha = 0.05$. An empirical estimate of the type-I error is that proportion in a large number of simulations conducted under the null hypothesis. We denote the number of simulations by *B*. For a more complete evaluation of the *P*-value distribution produced by a test, we propose to compute a statistic S_B that adds up the squares of deviations of ordered *P*-values from the respective theoretical values expected under the null distribution. A visual method of plotting ordered *P*-values against the corresponding expected values of order statistics is known as a "rankit plot" (Ipsen and Jerne 1944). Such a plot very closely corresponds to the common "Q-Q" plot (where values are plotted against quantiles instead), unless the value of *B* is small. The deviation from the null by visual inspection is judged by the deviation of actual *P*-values from the expected straight line. The essence of the statistic S_B is to capture the extent of this deviation. Since the usual type-I error reports the proportion of *P*-values below a single fixed cutoff point (a nominal level), commonly chosen to be 5%, it is possible that there would be a different degree of closeness to the nominal value at a different cutoff point. In contrast, the statistic S_B has an advantage in that it gives a summary of the correspondence of *P*-values with the null distribution for the entire (0, 1) interval.

We denote the ordered set of *P*-values obtained from *B* simulations as $\{p_{(1)}, \ldots, p_{(B)}\}$. The random variable that corresponds to the observed $p_{(i)}$ is denoted by $P_{(i)}$. The summary statistic measuring the lack of fit to the null distribution is

$$S_B = \sqrt{\frac{1}{B} \sum_{i=1}^{B} [p_{(i)} - \mathcal{E}(P_{(i)})]^2}. \tag{9}$$

Under the null hypothesis, the distribution of the order statistics $P_{(i)}$ would be Beta$(i, B - i + 1)$ if the distribution of the test statistic was continuous and exact,

rather than approximate. The computational formula for S_B is

$$S_B = \sqrt{\frac{1}{B}\sum_{i=1}^{B}\left[p_{(i)} - \frac{i}{B+1}\right]^2}. \qquad (10)$$

Larger values of S_B indicate larger deviations from the null distribution. When P-values indeed come from the null (uniform) distribution, we find the expected value of this statistic to be

$$\begin{aligned}\mathcal{E}(S_B) &\approx \sqrt{\frac{1}{B}\sum_{i=1}^{B}\mathcal{E}[P_{(i)} - \mathcal{E}(P_{(i)})]^2} \\ &= \sqrt{\frac{1}{B}\sum_{i=1}^{B}\frac{i(B-i+1)}{(B+1)^2(B+2)}} \\ &= \sqrt{\frac{1}{6(1+B)}}. \qquad (11)\end{aligned}$$

Thus, for any test statistic, the fit of P-values to the uniform distribution can be simply evaluated by computing the proposed statistic, S_B. We report and compare the values of S_B for competing methods, in addition to the usual empirical type-I error rates.

Performance of the tests under the alternative hypothesis (H_A) was characterized by statistical power. Power was estimated as the proportion of P-values that fall below the 5% rejection threshold, using data sets generated under H_A.

RESULTS

Known haplotype phase: The goal of this section is to compare performance of the proposed correlation-based tests. The performance was evaluated in terms of the classical type-I error and power. Additionally, the fit to the null distribution was evaluated with the usage of the coefficient S_B, as described above. The following tests were used in this study:

1. Correlation-based statistic T_1 defined by (2).
2. Correlation-based statistic T_2 defined by (3).
3. Cressie–Read's power divergence statistic,

$$C^\lambda = \frac{2}{\lambda(\lambda+1)}\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}\left\{\left(\frac{n_{ij}}{e_{ij}}\right)^\lambda - 1\right\} \qquad (12)$$

   with $\lambda = \frac{2}{3}$ (Cressie and Read 1984), where $e_{ij} = n_{i\cdot}n_{\cdot j}/n$ are the expected counts. C^λ has an asymptotic chi-square distribution with $(k-1)(m-1)$ d.f.
4. Likelihood-ratio (LR) statistic,

$$G^2 = 2\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}\ln\left(\frac{n_{ij}}{e_{ij}}\right) = \lim_{\lambda\to 0} C^\lambda. \qquad (13)$$

5. Pearson's chi-square statistic,

$$X^2 = 2\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = C^1. \qquad (14)$$

6. Permutation-based tests using statistics as defined above, which we denote as T_p, G^2_p, C^λ_p, and X^2_p. The statistics T_1 and T_2 correspond to the same permutational test, denoted by T_p.
7. Fisher's exact test F_p, with the P-value approximated by a permutation test using the statistic $\sum_{i=1}^{k}\sum_{j=1}^{m}\ln(n_{ij}!)$.

The P-value for a permutation test is defined as the proportion of times the test statistic computed from randomly sampled tables was found to be as extreme or more extreme than the statistic value for the original data. These random tables are generated with marginal counts constrained to be the same as that for the observed data set. We used $K = 19{,}999$ permutations to compute each P-value, and the number of simulations in all type-I error evaluation experiments was $B = 100{,}000$. Oden (1991) showed that the value of K in simulation experiments can be very much smaller than B. Boos and Zhang (2000) suggested that K can be as small as $8\sqrt{B}$, and if the significance level is α, the value should preferably be such that $(K+1)\alpha$ is an integer. The number of simulations to evaluate power was 10,000.

Tables 1–3 present results for the type-I error rates at the nominal 5% level and the closeness of fit to the null distribution as measured by the S_B statistic. The tables of haplotype counts in this set of simulations have fixed margins and the cell counts are generated at random to satisfy the marginal conditions. A similar approach was used in the evaluation of small sample properties of some common tests, such as Pearson's chi square (*e.g.*, Larntz 1978; Fienberg 1979). For example, the marginal frequencies in Table 2 are taken to be proportional to (2:3:5) for the rows and (2:3:4:5:6) for the columns. This matches the first setting of Table 6 in Larntz (1978). Our values for Pearson's X^2 and the LR statistic G^2 replicate the type-I error results of Larntz, who used sample sizes of 20–100. Across all simulations, our results confirm the previous observations (Larntz 1978) that the LR test (G^2) has an inflated type-I error when sample sizes are small to moderate.

Both of the proposed statistics, T_1 and T_2 show a correct type-I error for the corresponding test. Moreover, examination of S_B values indicates that small-to-moderate sample size behavior of these statistics is such that they provide the best fit to the null distribution among the asymptotic/approximate tests studied here. The simpler approximation, T_2, shows the best fit.

Tables 4–7 present both power and the behavior under H_0, given in terms of the type-I error and the S_B values. The null distribution data sets corresponding to the power results were generated by randomly shuffling

TABLE 1

Type-I error rates and values of the statistic measuring lack of fit to the null distribution, $1000 \times S_B$ for 3×5 tables: row margins, 5:3:2; column margins, 2:3:4:5:6

N	T_2	T_1	G^2	$C^{2/3}$	X^2	T_p	G^2_p	$C^{2/3}_p$	X^2_p	F_p
					Type-I error					
20	0.041	0.039	0.119	0.037	0.038	0.051	0.049	0.051	0.051	0.048
40	0.047	0.046	0.103	0.047	0.045	0.051	0.051	0.051	0.051	0.051
60	0.050	0.049	0.090	0.050	0.048	0.052	0.051	0.051	0.051	0.052
80	0.049	0.048	0.080	0.049	0.047	0.050	0.051	0.050	0.050	0.051
100	0.049	0.048	0.073	0.049	0.047	0.050	0.050	0.050	0.049	0.049
1000	0.051	0.050	0.052	0.051	0.051	0.050	0.051	0.051	0.051	0.051
					$1000 \times S_B$					
20	55	56	200	78	58	0.94	17	1.1	2.1	28
40	26	28	120	43	30	1.5	1	1.5	1.4	1.3
60	15	17	77	27	18	1.3	1.3	1.3	1.4	1.4
80	10	12	54	19	13	1.3	1.6	1.5	1.5	1.5
100	8.6	10	41	16	11	0.57	0.68	0.48	0.53	0.56
1000	1	1.6	3.8	2.1	1.7	0.53	0.76	0.78	0.78	0.73

Expected value of $1000 \times S_B$ for the uniform P-value distribution is $1000\sqrt{1/[6(1+100,000)]} = 1.29$.

the data generated under the association model, to produce new counts under the hypothesis of no association. Sample sizes for different simulations are chosen depending on the strength of the population association, to provide intermediate to high power, and highlight the difference between the tests.

The population association values for Tables 4–6 were generated as follows. The association value for the cell (i, j) can be measured in terms of LD, $D_{ij} = p_{ij} - p_i q_j$. The maximum absolute value of D_{ij} is constrained by the marginal frequencies p_i, q_j and the association values for all cells were set as proportions of the maximum attainable value, D'_{ij} (Lewontin 1964). The population frequencies, p_{ij}, and the values of D'_{ij} are given in appendix b, Tables B1–B3. Samples for each simulation experiment were obtained by multinomial sampling from these population frequencies. Both of the proposed tests (T_1, T_2) show type-I error rates close to the nominal 5% level. The simple approximation T_2 shows the best fit to the null distribution among the asymptotic/approximate tests. Moreover, the power corresponding to T_2 or its permutational equivalent T_p is somewhat higher than that for the rest of the tests. These differences in power are highly significant statistically, due to the paired nature of the data (P-values) and the large number of simulations.

As mentioned previously, in the known-phase case the test for LD is equivalent to a test for interaction in a contingency table. In principle, the tests based on the total correlation can be used in a classical setting of testing

TABLE 2

Type-I error rates and values of the statistic measuring lack of fit to the null distribution, $1000 \times S_B$ for 3×5 tables: row margins, 2:3:5; column margins, 2:3:4:5:6

N	T_2	T_1	G^2	$C^{2/3}$	X^2	T_p	G^2_p	$C^{2/3}_p$	X^2_p	F_p
					Type-I error					
20	0.041	0.039	0.119	0.036	0.038	0.051	0.048	0.051	0.051	0.044
40	0.046	0.044	0.100	0.045	0.043	0.049	0.049	0.049	0.050	0.049
60	0.049	0.047	0.088	0.049	0.046	0.050	0.051	0.050	0.050	0.050
80	0.049	0.048	0.079	0.050	0.047	0.050	0.050	0.050	0.050	0.050
100	0.049	0.047	0.072	0.049	0.047	0.049	0.049	0.050	0.049	0.049
1000	0.051	0.050	0.051	0.050	0.050	0.050	0.050	0.050	0.050	0.050
					$1000 \times S_B$					
20	55	57	200	78	59	0.82	13	0.86	1.6	23
40	26	27	120	42	29	0.85	1.5	0.85	0.87	1.9
60	17	18	79	28	20	0.67	1.2	1.1	0.91	0.93
80	11	13	54	20	13	1.1	1.3	1.2	1.2	1.2
100	9.1	11	42	16	12	0.66	0.91	0.84	0.77	0.81
1000	1.7	1	3	1.2	0.81	0.77	0.78	0.76	0.76	0.75

Expected value of $1000 \times S_B$ for the uniform P-value distribution is $1000\sqrt{1/[6(1+100,000)]} = 1.29$.

TABLE 3

Type-I error rates and values of the statistic measuring lack of fit to the null distribution, $1000 \times S_B$ for 5×7 tables: row margins, 2:3:4:5:6; column margins, 1:2:3:4:5:6:7

N	T_2	T_1	G^2	$C^{2/3}$	X^2	T_p	G_p^2	$C_p^{2/3}$	X_p^2	F_p
					Type-I error					
20	0.026	0.026	0.012	0.006	0.025	0.049	0.044	0.049	0.049	0.048
40	0.038	0.037	0.078	0.022	0.036	0.050	0.050	0.050	0.050	0.050
60	0.042	0.042	0.109	0.032	0.042	0.050	0.049	0.050	0.050	0.050
80	0.045	0.044	0.117	0.039	0.044	0.050	0.049	0.049	0.050	0.049
100	0.046	0.046	0.112	0.043	0.045	0.050	0.050	0.050	0.050	0.050
1000	0.051	0.050	0.057	0.050	0.050	0.050	0.050	0.050	0.050	0.050
					$1000 \times S_B$					
20	98	99	180	84	98	0.85	32	0.81	1.3	68
40	47	48	180	47	48	0.56	3.3	0.47	0.54	5.9
60	29	30	190	39	31	0.55	0.62	0.65	0.65	0.89
80	20	21	170	34	22	1.5	1.9	1.6	1.5	1.7
100	17	18	150	30	18	1.2	1.3	1.1	1.1	1.2
1000	0.96	1.1	15	2	0.94	1.6	1.9	1.8	1.8	1.7

Expected value of $1000 \times S_B$ for the uniform P-value distribution is $1000\sqrt{1/[6(1+100,000)]} = 1.29$.

heterogeneity between several multinomial samples. Although a detailed examination of the proposed tests regarding this problem is beyond the scope of this article, a simulation study (Table 7) confirms that the proposed approach provides a competitive test. In the 5×5 tables used here, rows represent independent samples taken from five populations; and columns represent five categories (such as sample-specific allele frequencies). Population frequencies for each of the simulations were generated from the Dirichlet distribution with the common parameter, 20. A property of this sampling is such that the 1 and 99% population quantiles for the frequency of any of the five column categories are 0.1 and 0.3, with mean frequency 0.2. This range gives a measure of the between-population variability for each of the categories. Samples for each of the five populations were generated by multinomial sampling for each of the simulation runs. As before, data for the hypothesis of homogeneity (H_0) were obtained by taking the sample generated as just described and reshuffling the counts under the constraints that the marginal frequencies of a particular sample are preserved. Table 7 shows good properties of the proposed tests under the hypothesis of no association. The power values are found to be identical to those provided by Fisher's exact test. The asymptotic version of the LR test (G^2) shows a higher power; however, this value might be unreliable, because the type-I error of this test was found consistently inflated in all simulations.

Unknown haplotype phase: This section gives results of the comparison between the two "LD correlation" statistics (T_1, T_2) and a chi-square test recently described by Schaid (2004), which has similarity in that it also utilizes the composite LD definition. Schaid's test (S^2) corresponds to Pearson's chi-square in the "known-phase" case; however, there is no simple explicit expression for the test statistic in the ambiguous haplotype phase case. The calculation of S^2 involves a generalized inverse

TABLE 4

Power and the corresponding H_0 behavior for 4×3 tables at abs(D') ± 0.5 (population parameters are defined in Table B1 of appendix b)

N	T_2	T_1	G^2	$C^{2/3}$	X^2	T_p	G_p^2	$C_p^{2/3}$	X_p^2	F_p
					Power					
30	0.570	0.554	0.558	0.407	0.390	0.575	0.501	0.466	0.422	0.465
60	0.908	0.904	0.890	0.843	0.833	0.908	0.846	0.852	0.844	0.842
					Type-I error					
30	0.048	0.045	0.071	0.038	0.043	0.051	0.050	0.051	0.051	0.051
60	0.050	0.047	0.080	0.045	0.044	0.050	0.049	0.049	0.049	0.050
					$1000 \times S_B$					
30	32	38	130	51	41	2.5	3.9	2.2	2.7	6.1
60	13	18	83	29	20	0.59	0.75	0.58	0.56	0.61

Expected value of $1000 \times S_B$ for the uniform P-value distribution is $1000\sqrt{1/[6(1+100,000)]} = 1.29$.

TABLE 5

Power and the corresponding H_0 behavior for 4 × 3 tables at abs(D') ± 0.5 (population parameters are defined in Table B2 of APPENDIX B)

N	T_2	T_1	G^2	$C^{2/3}$	X^2	T_p	G_p^2	$C_p^{2/3}$	X_p^2	F_p
					Power					
30	0.526	0.510	0.488	0.342	0.318	0.529	0.435	0.397	0.343	0.401
60	0.892	0.885	0.864	0.790	0.771	0.888	0.823	0.806	0.782	0.795
					Type-I error					
30	0.049	0.045	0.068	0.038	0.046	0.050	0.050	0.051	0.050	0.049
60	0.052	0.048	0.072	0.045	0.047	0.051	0.051	0.050	0.050	0.051
					$1000 \times S_B$					
30	29	36	120	44	35	3.8	5.2	3.8	4.4	6.8
60	12	18	91	30	20	1.1	0.96	1	1.1	1.2

Expected value of $1000 \times S_B$ for the uniform P-value distribution is $1000\sqrt{1/[6(1 + 100,000)]} = 1.29$.

of the covariance matrix of the sample composite LD. We assume a common scenario when single-locus genotypes are scored at each locus, without the knowledge about arrangement of the alleles on haplotypes across the loci.

The first set of simulations was designed for a two-locus linkage equilibrium system with five and seven alleles correspondingly. Both loci have high population levels of HWD. The amount of HWD and allele frequencies for various simulation settings are given in the legend to Table 8. The homozygote HWD values for the two loci (D'_{ii}) are given as proportions of the maximum possible value. The heterozygote HW disequilibria are related to these as $\sum_{i \neq j} D_{ij} = \sum_i D_{ii}/2$. The simulation results confirm that both the correlation-based tests and Schaid's test are robust in the presence of high levels of population HWD. Similar to the known-phase results, the simple T_2 approximation shows the best fit to the null distribution (under the hypothesis of linkage equilibrium).

The second set of simulations was designed to evaluate power utilizing the population LD derived from an actual set of human short tandem repeat (STR) polymorphisms, described in ROSENBERG *et al.* (2002). We took 30 STR loci from chromosome 1, using a combined sample of 217 Middle-East and European individuals, and identified seven pairs of loci in LD by an exact test (ZAYKIN *et al.* 1995). The resulting set of loci used for these simulations had 4–6 alleles after rare alleles were grouped together. Two-locus counts of these data were further used to set the population frequencies. These fixed population frequencies were used to obtain multinomial samples of individuals for each of the simulations. Results of these simulations are shown in Table 9. The permutational ("exact") version of the correlation-based tests, T_p was included as well. The fit to the null (linkage equilibrium) distribution follows the same pattern found in the previous simulations—the simple approximation T_2 shows a better fit than other nonexact tests. The power values are found to be similar in all cases.

Correspondence between approximations and the exact test for the total correlation: Overall, we found an excellent agreement between P-values provided by either of the approximations (T_1, T_2) and the exact P-value given by the test T_p.

Figure 1, a and b, shows a very close P-value correspondence between T_2 and the its exact version, T_p. Figure 1a plots the T_2 P-values against the T_p P-values using the subset of simulations used to produce Table 1 ($N = 100$). Figure 1b is a similar plot for the unknown haplotype

TABLE 6

Power and the corresponding H_0 behavior for 5 × 5 tables at abs(D') ∈ (0.19–0.21) (population parameters are defined in Table B3 of APPENDIX B)

N	T_2	T_1	G^2	$C^{2/3}$	X^2	T_p	G_p^2	$C_p^{2/3}$	X_p^2	F_p
					Power					
150	0.752	0.749	0.747	0.719	0.726	0.754	0.672	0.723	0.733	0.709
200	0.884	0.882	0.868	0.862	0.867	0.884	0.826	0.863	0.870	0.847
					Type-I error					
150	0.050	0.048	0.084	0.048	0.047	0.050	0.050	0.050	0.050	0.050
200	0.050	0.049	0.075	0.050	0.049	0.050	0.050	0.051	0.051	0.050
					$1000 \times S_B$					
150	9.6	11	76	19	12	2.1	1.9	2.2	2.4	2
200	5	6.4	52	12	6.8	1.3	0.88	1.2	1.3	1.2

Expected value of $1000 \times S_B$ for the uniform P-value distribution is $1000\sqrt{1/[6(1 + 100,000)]} = 1.29$.

TABLE 7

Power and the corresponding H_0 behavior for the "sample heterogeneity" model (5 × 5 tables)

N	T_2	T_1	G^2	$C^{2/3}$	X^2	T_p	G_p^2	$C_p^{2/3}$	X_p^2	F_p
					Power					
100	0.655	0.655	0.670	0.658	0.655	0.657	0.656	0.658	0.657	0.657
150	0.836	0.835	0.841	0.835	0.835	0.836	0.836	0.836	0.835	0.836
					Type-I error					
100	0.050	0.050	0.054	0.050	0.050	0.050	0.050	0.050	0.050	0.050
150	0.049	0.049	0.052	0.049	0.049	0.049	0.049	0.049	0.049	0.049
					$1000 \times S_B$					
100	2.9	2.9	10	4	3	1.2	1.1	1.2	1.2	1.1
150	2.9	2.9	8.3	3.8	2.9	1.6	1.7	1.6	1.6	1.7

Expected value of $1000 \times S_B$ for the uniform P-value distribution is $1000\sqrt{1/[6(1+100,000)]} = 1.29$.

phase data (locus pairs 11 and 23 from Table 9). For comparison, Figure 1c plots T_2 P-values against those obtained by Pearson's chi-square test ($N = 100$, data from Table 1 simulations). There is no similar correspondence, which indicates that the two statistics are capturing somewhat different aspects of sample associations.

Figure 2 shows the correspondence between the two correlation-based test approximations, T_1 and T_2. Figure 2a illustrates the correspondence for the known haplotype phase case ($N = 100$; data for Table 1). Figure 2b illustrates a similar correspondence between the P-values for the unknown haplotype phase case (data from simulations to produce "setting I" in Table 8).

Due to closeness of P-values resulting from the T_1 and the T_2 tests, and much greater simplicity of the T_2-statistic computation, we recommend its usage over the test based on T_1.

TABLE 8

Type-I error and values of the statistic measuring lack of fit to the null distribution, $1000 \times S_B$, for the composite LD tests: locus A, five alleles; locus B, seven alleles; $n = 100$

Setting	T_2	T_1	S^2
		Type-I error	
I	0.050	0.046	0.045
II	0.044	0.043	0.043
III	0.050	0.046	0.045
IV	0.051	0.049	0.049
V	0.050	0.049	0.048
		$1000 \times S_B$	
I	12.2	15.2	15.8
II	14.0	16.0	16.4
III	8.2	11.6	11.2
IV	8.0	10.1	10.9
V	5.2	5.8	6.1

Setting I: locus A, p_i, 0.15, 0.22, 0.15, 0.23, 0.25 and D'_{ii}, −1, 0.13, −1, 0.20, 0.12; locus B, q_i, 0.16, 0.17, 0.11, 0.16, 0.16, 0.18, 0.05 and D'_{ii}, 0.10, 0.09, −1, 0.10, 0.10, 0.07, 1. Setting II: locus A, p_i, 0.20, 0.18, 0.20, 0.20, 0.22 and D'_{ii}, 0.25, 1, 0.25, 0.25, 0.32; locus B, q_i, 0.14, 0.14, 0.12, 0.16, 0.16, 0.14, 0.14 and D'_{ii}, 0.17, 0.17, 1, 0.23, 0.23, 0.17, 0.17. Setting III: locus A, p_i, 0.18 0.18, 0.16, 0.24, 0.25 and D'_{ii}, −0.37, −0.37, −0.20, 0.24, 0.17; locus B, q_i, 0.11, 0.15, 0.13, 0.14, 0.15, 0.16, 0.17 and D'_{ii}, 0.40, −0.31, −0.14, −0.19, 0.25, 0.22, 0.18. Setting IV: locus A, p_i, 0.23, 0.38, 0.38 and D'_{ii}, 0.57, 0.68, 0.68; locus B, q_i, 0.22, 0.25, 0.27, 0.27 and D'_{ii}, −0.35, −0.5, −0.56, −0.56. Setting V: locus A, p_i, 0.32, 0.32, 0.35 and D'_{ii}, 0.60, 0.60, 0.48; locus B, q_i, 0.26, 0.24, 0.24, 0.26 and D'_{ii}, 0.55, 0.67, 0.67, 0.55. Expected value of $1000 \times S_B$ for the uniform P-value distribution is $1000\sqrt{1/[6(1+10,000)]} = 4.08$.

TABLE 9

Human diversity panel results

Locus pair	n	T_2	T_1	S^2	T_p
			Power		
15/16	50	0.785	0.768	0.730	0.780
9/16	150	0.870	0.861	0.857	0.863
19[a]/23	150	0.882	0.874	0.874	0.880
11/23[a]	150	0.814	0.805	0.791	0.809
5/12	150	0.775	0.762	0.790	0.765
23[a]/25	150	0.957	0.955	0.952	0.956
21/26	100	0.848	0.838	0.870	0.842
			Type-I error		
15/16	50	0.050	0.044	0.043	0.044
9/16	150	0.053	0.047	0.044	0.047
19[a]/23	150	0.050	0.047	0.047	0.050
11/23[a]	150	0.051	0.047	0.047	0.050
5/12	150	0.051	0.047	0.048	0.050
23[a]/25	150	0.048	0.045	0.046	0.047
21/26	100	0.051	0.048	0.048	0.049
			$1000 \times S_B$		
15/16	50	18.9	25.3	26.6	3.5
9/16	150	3.4	7.8	7.9	2.9
19[a]/23	150	2.8	6.7	6.9	4.3
11/23[a]	150	5.1	9.0	9.1	2.7
5/12	150	5.1	9.6	9.1	4.2
23[a]/25	150	4.7	7.6	8.4	2.6
21/26	100	10.9	14.9	14.9	3.4

Expected value of $1000 \times S_B$ for the uniform P-value distribution is $1000\sqrt{1/[6(1+10,000)]} = 4.08$. Locus abbreviations: 5, ATA47D07; 9, GATA26G09; 11, GATA109; 12, GATA6A05; 15, ATA25E07; 16, ATA42G12; 19, GGAA5F09; 21, ATA4E02; 23, GATA48B01; 25, GATA4H09; 26, ATA29C07.

[a] Loci in HWD.

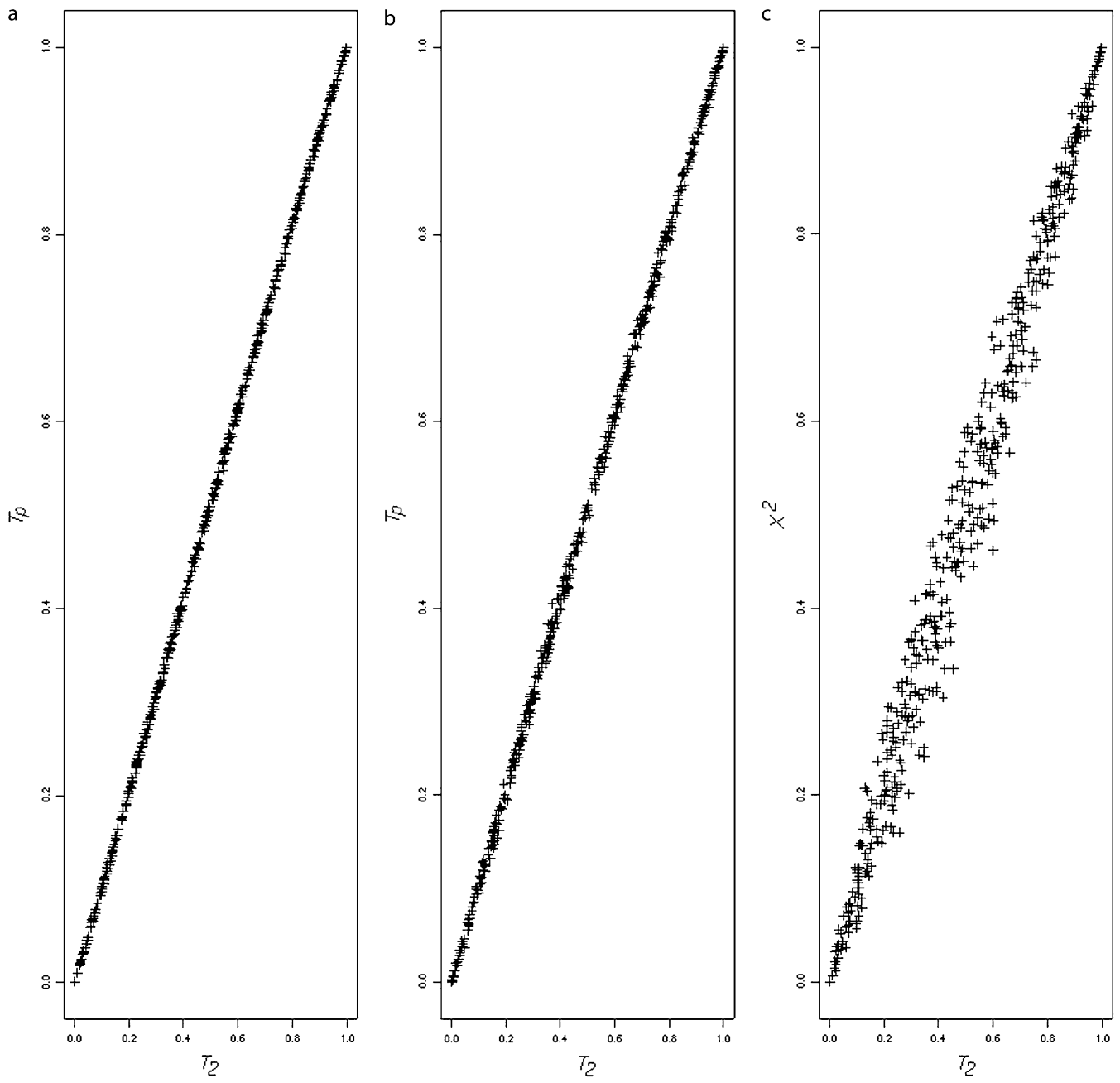

FIGURE 1.—(a) Plots of T_2 *P*-values against the T_p *P*-values for the known haplotype phase simulations. (b) Plots of T_2 *P*-values against the T_p *P*-values for the unknown haplotype phase simulations. (c) Plots of T_2 *P*-values against Pearson's χ^2 *P*-values for the known haplotype phase simulations.

DISCUSSION

We introduce correlation-based testing for linkage disequilibrium with multiple alleles. Following earlier work by WEIR (1979) and SCHAID (2004) we adopt the usage of the composite LD that provides robust inference even under conditions of high deviations from HWE. Simulations confirm that the test maintains the proper error rate even when the HWD reaches its maximum value for some of the genotypes. Our approach provides several advantages. The behavior of the proposed method under the hypothesis of no association is found to be consistently closer to the expected than that of other "nonexact" tests included in this study. Values of the statistic S_B that we introduced for evaluation of the null distribution of the studied test statistics show that in 35 of 38 experiments, the approximation T_2 was closer to its null expected value than the chi-square statistic (Tables 1–9). Power evaluations suggest that the correlation-based tests provide higher power than other tests under the alternatives where associations are present among multiple pairs of alleles (Tables 4–6). The novelty and advantages of our approach also include tractability of the corresponding

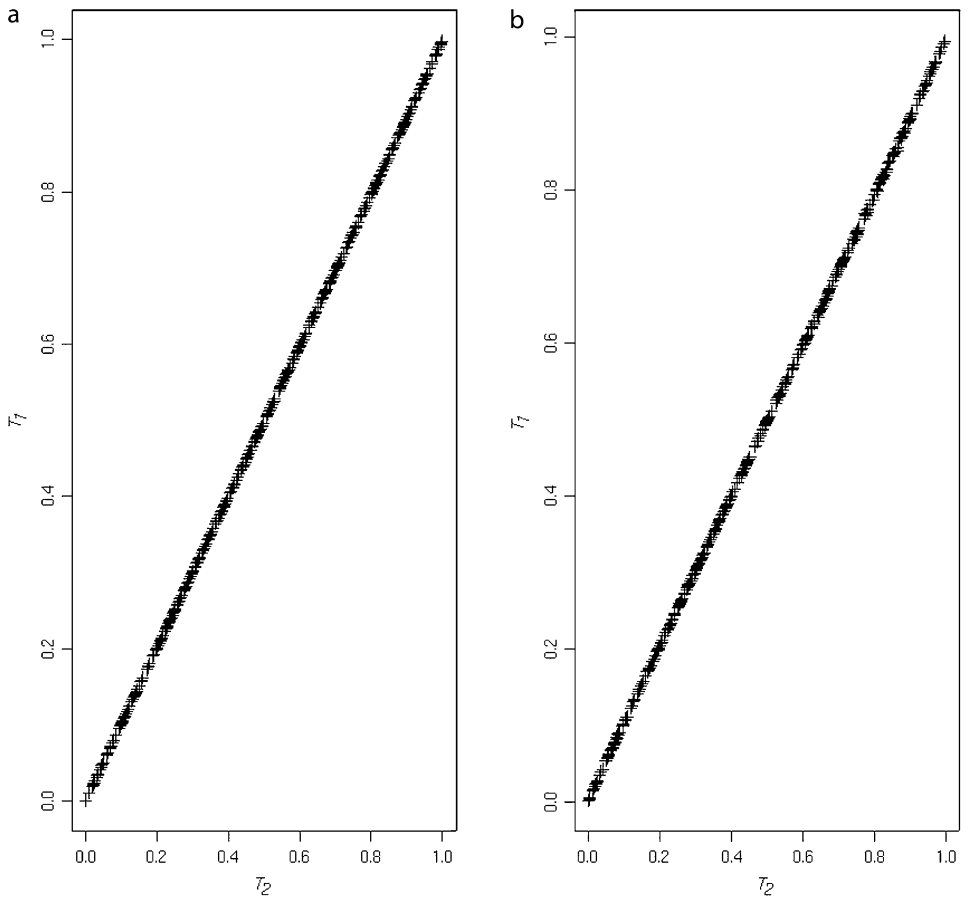

FIGURE 2.—(a) Plots of T_2 *P*-values against T_1 *P*-values for the known haplotype phase simulations. (b) Plots of T_2 *P*-values against T_1 *P*-values for the unknown haplotype phase simulations.

test statistic, simplicity, and high speed of computations. The relation of the sum of squared LD correlations to chi-square extends the well-known relation for the two-allele case and thus may have implications for the design of genetic association studies. Good power properties of the test based on a simple statistic $T_2 = (k-1)(m-1)n\bar{r}^2$ give justification for usage of the average correlation to characterize and compare multiallelic LD in various settings, including estimation of the effective population size (WAPLES 2006) and fine mapping of genetic traits, where LD coefficients could be compared between samples with and without a specific disease (NIELSEN *et al.* 2004; ZAYKIN *et al.* 2006). Further work may include investigation of confidence intervals for R^2, on the basis of the proposed chi-square approximation.

Although the method is motivated by testing the LD, the test provides high power when used to detect heterogeneity among samples in contingency tables. For example, the correlation-based test can be used to compare allele or genotype frequencies (columns) between samples from several populations, represented by rows in a contingency table. In this setting, the power is very similar to the power of common tests such as Pearson's chi-square and Fisher's exact test. Further study may be required to fully investigate properties of this test as a general purpose test for detecting interactions and heterogeneity in contingency tables.

A computer program implementing the methods described here is available at (http://www.niehs.nih.gov/research/atniehs/labs/bb/staff/zaykin/rxc.cfm) or by a request to D.V.Z. The provided implementation computes average correlations with the corresponding *P*-values on the basis of the T_2 statistic, using multilocus genotype data. For those *P*-values that fall below a user-specified threshold, a Monte Carlo *P*-value is reported as well. This approach allows rapid computations for large collections of loci. Correlation-based tests for contingency tables are implemented as well.

Shyamal Peddada and David Umbach provided useful discussion. Noah Rosenberg provided STR genotypes for deriving data sets used in the simulation study. Daniel Schaid and Jason Sinnwell provided a program implementing Schaid's S^2 test. This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Institute of Environmental Health Sciences, and by NIH GM 07591.

LITERATURE CITED

BOOS, D. D., and J. ZHANG, 2000 Monte Carlo evaluation of resampling-based hypothesis tests. J. Am. Stat. Assoc. **95:** 486–492.

BOX, G. E. P., 1954 Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effect of inequality of variance in the two-way classification. Ann. Math. Stat. **25:** 290–302.

CRESSIE, N., and T. R. C. READ, 1984 Multinomial goodness-of-fit tests. J. R. Stat. Soc. B **46:** 440–464.

EVETT, I. W., and B. S. WEIR, 1998 *Interpreting DNA Evidence.* Sinauer Associates, Sunderland, MA.

Excoffier, L., and M. Slatkin, 1995 Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol. Biol. Evol. **12:** 921–927.
Fienberg, S. E., 1979 The use of chi-squared statistics for categorical data problems J. R. Stat. Soc. B **41:** 54–64.
Hill, W. G., 1974 Estimation of linkage disequilibrium in randomly mating populations. Heredity **33:** 229–232.
Hill, W. G., 1975 Tests for association of gene frequencies at several loci in random mating diploid populations. Biometrics **31:** 881–888.
Hedrick, P. W., 1987 Gametic disequilibrium measures: proceed with caution. Genetics **117:** 331–341.
Ipsen, J., and N. K. Jerne, 1944 Graphical evaluation of the distribution of small experimental series. Acta Pathol. Microbiol. Scand. **21:** 343–361.
Kalinowski, S. T., and P. W. Hedrick, 2000 Estimation of linkage disequilibrium for loci with multiple alleles: basic approach and an application using data from bighorn sheep. Heredity **87:** 698–708.
Karlin, S., and A. Piazza, 1981 Statistical methods for assessing linkage disequilibrium at the HLA-A, B, C loci. Ann. Hum. Genet. **45:** 79–94.
Holt, D., A. J. Scott and P. D. Ewings, 1980 Chi-squared tests with survey data. J. R. Stat. Soc. A **143:** 303–320.
International HapMap Consortium, 2003 The International HapMap Project. Nature. **426:** 789–796.
Larntz, L., 1978 Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. J. Am. Stat. Assoc. **73:** 253–263.
Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49:** 49–67.
Lewontin, R. C., 1974 *The Genetic Basis of Evolutionary Change.* Columbia University Press, New York.
Nielsen, D. M., M. G. Ehm, D. V. Zaykin and B. S. Weir, 2004 Effect of two-and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. Genetics **168:** 1029–1040.
Oden, N. L., 1991 Allocation of effort in Monte Carlo simulation for power of permutation tests. J. Am. Stat. Assoc. **86:** 1074–1076.
Pearson, K., 1922 On the χ^2 test of goodness of fit. Biometrika **14:** 186–191.
Pritchard, J. K., and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. **69:** 1–14.
Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd *et al.*, 2002 Genetic structure of human populations. Science **298:** 2981–2985.
Schaid, D. J., 2004 Linkage disequilibrium testing when linkage phase is unknown. Genetics **166:** 505–512.
Searle, S. R., 1971 *Linear Models.* John Wiley & Sons, New York.
Slatkin, M., 1994 Linkage disequilibrium in growing and stable populations. Genetics **137:** 331–336.
Terwilliger, J. D., and T. Hiekkalinna, 2006 An utter refutation of the "Fundamental Theorem of the HapMap". Eur. J. Hum. Genet. **14:** 426–437.
Tishkoff, S. A., E. Dietzsch, W. Speed, A. J. Pakstis, J. R. Kidd *et al.*, 1996 Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science **271:** 1380–1387.
Waples, R. S., 2006 A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. Conserv. Genet. **7:** 167–184.
Weir, B. S., 1979 Inferences about linkage disequilibrium. Biometrics **35:** 235–254.
Weir, B. S., 1996 *Genetic Data Analysis II.* Sinauer Associates, Sunderland, MA.
Weir, B. S., and C. C. Cockerham, 1978 Testing hypotheses about linkage disequilibrium with multiple alleles. Genetics **88:** 633–642.
Weir, B. S., and C. C. Cockerham, 1979 Estimation of linkage disequilibrium in randomly mating populations. Heredity **42:** 105–111.
Weir, B. S., and C. C. Cockerham, 1989 Complete characterization of disequilibrium at two loci, pp. 86–110 in *Mathematical Evolutionary Theory*, edited by M. W. Feldman. Princeton University Press, Princeton, NJ.
Yamazaki, T., 1977 The effects of overdominance on linkage in a multilocus system. Genetics **86:** 227–236.
Zapata, C., 2000 The D' measure of overall gametic disequilibrium between pairs of multiallelic loci. Evolution **54:** 1809–1812.
Zaykin, D., L. Zhivotovsky and B. S. Weir, 1995 Exact tests for association between alleles at arbitrary numbers of loci. Genetica **96:** 169–178.
Zaykin, D. V., 2004 Bounds and normalization of the composite linkage disequilibrium coefficient. Genet. Epidemiol. **27:** 252–257.
Zaykin, D. V., Z. Meng and M. G. Ehm, 2006 Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. Am. J. Hum. Genet. **78:** 737–746.
Zhao, H., D. Nettleton, M. Soller and J. C. M. Dekkers, 2005 Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. Genet. Res. **86:** 77–87.
Zhao, H., D. Nettleton and J. C. M. Dekkers, 2007 Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between single nucleotide polymorphisms. Genet. Res. **89:** 1–6.

Communicating editor: A. D. Long

APPENDIX A

We denote the $km \times 1$ vectors of population and sample frequencies by $\mathbf{P}$ and $\tilde{\mathbf{P}}$; the elements of $\tilde{\mathbf{P}}$ are the observed haplotype frequencies, $\tilde{p}_{ij} = n_{ij}/N$. Under the null hypothesis, H_0: $\mathbf{P} = \mathbf{P}_0$, we have that $\sqrt{N}(\tilde{\mathbf{P}} - \mathbf{P}_0)$ converges in distribution to a multivariate normal. Row and column frequencies for the table of haplotype frequencies correspond to the vectors of allele frequencies at the two loci: $\mathbf{p}$, $\{p_1, \ldots, p_k\}^T$; $\mathbf{q}$, $\{q_1, \ldots, q_m\}^T$. For complete absence of linkage disequilibrium, the vector of frequencies is a $(km \times 1)$ Kronecker product,

$$\mathbf{P}_0 = \mathbf{p} \otimes \mathbf{q} = \{p_1 q_1, p_1 q_2, \ldots, p_i q_j, \ldots, p_k q_m\}^T$$

and the vector of expected (equilibrium) sample frequencies is based on sample values

$$\tilde{\mathbf{P}}_0 = \{\tilde{p}_1 \tilde{q}_1, \tilde{p}_1 \tilde{q}_2, \ldots, \tilde{p}_i \tilde{q}_j, \ldots, \tilde{p}_k \tilde{q}_m\}^T.$$

Under H_0, the covariance matrix of $\tilde{\mathbf{P}}$ is

$$\mathrm{Var}(\tilde{\mathbf{P}}) = \frac{1}{N}(\mathrm{diag}(\mathbf{P}_0) - \mathbf{P}_0 \mathbf{P}_0^T),$$

and the variance of $(\tilde{\mathbf{P}} - \tilde{\mathbf{P}}_0)$ is

$$\mathrm{Var}(\tilde{\mathbf{P}} - \tilde{\mathbf{P}}_0) = \frac{1}{N}(\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T) \otimes (\mathrm{diag}(\mathbf{q}) - \mathbf{q}\mathbf{q}^T)$$

(Holt *et al.* 1980). The contingency table Pearson's chi-square statistic is

$$X^2 = N \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(\tilde{p}_{ij} - \tilde{p}_i \tilde{q}_j)^2}{\tilde{p}_i \tilde{q}_j}. \qquad \text{(A1)}$$

Denote

$$\mathbf{V}_P = \mathrm{diag}(\mathbf{P}_0) - \mathbf{P}_0\mathbf{P}_0^T$$
$$\mathbf{V} = (\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T) \otimes (\mathrm{diag}(\mathbf{q}) - \mathbf{q}\mathbf{q}^T)$$
$$\mathbf{\Gamma} = \mathrm{diag}\left[\{1/\sqrt{\tilde{p}_i}\} \otimes \{1/\sqrt{\tilde{q}_i}\}\right]$$
$$\mathbf{\Psi} = \mathrm{diag}\left[\{1/\sqrt{\tilde{p}_i(1-\tilde{p}_i)}\} \otimes \{1/\sqrt{\tilde{q}_i(1-\tilde{q}_i)}\}\right]$$
$$\mathbf{Z} = \mathbf{\Gamma}(\tilde{\mathbf{P}} - \tilde{\mathbf{P}}_0)$$
$$\mathbf{R} = \mathbf{\Psi}(\tilde{\mathbf{P}} - \tilde{\mathbf{P}}_0).$$

The notation $\{\cdot\}$ above denotes vectors; *e.g.*, $\{1/\sqrt{p_i}\} \equiv \{1/\sqrt{p_1}, \ldots, 1/\sqrt{p_k}\}$. The elements of the vectors $\mathbf{Z}$ and $\mathbf{R}$ are

$$\mathbf{Z} = \left\{ \frac{\tilde{p}_{ij} - \hat{p}_i\hat{q}_j}{\sqrt{\tilde{p}_i\tilde{q}_j}} \right\} \tag{A2}$$

$$\mathbf{R} = \left\{ \frac{\tilde{p}_{ij} - \tilde{p}_i\tilde{q}_j}{\sqrt{\tilde{p}_i(1-\tilde{p}_i)\tilde{q}_j(1-\tilde{q}_i)}} \right\}. \tag{A3}$$

The elements of $\mathbf{R}$ are sample correlations for each pair of alleles, and $\mathbf{R}^T\mathbf{R} = \sum_{i,j} r_{ij}^2$ is the sum of squared correlations for the entire table.

Pearson's X^2 in (A1) can be expressed differently, using either the vector of the chi-square contributions, $\mathbf{Z}$, or the vector of correlations, $\mathbf{R}$,

$$\frac{1}{N}X^2 = \mathbf{R}^T(\mathbf{\Psi}\tilde{\mathbf{V}}^-\mathbf{\Psi})\mathbf{R} \tag{A4}$$

$$= \mathbf{Z}^T(\mathbf{\Gamma}\tilde{\mathbf{V}}^-\mathbf{\Gamma})\mathbf{Z} \tag{A5}$$

$$= \mathbf{Z}^T\mathbf{Z}, \tag{A6}$$

where $\mathbf{V}^-$ denotes a generalized inverse of $\mathbf{V}$. The reduction to a simple sum was given by Pearson (1922).

Our primary interest is in the approximate distribution of $\mathbf{R}^T\mathbf{R}$. For any multivariate normal vector $\mathbf{Z}^*$ with covariance matrix $\mathbf{V}^*$, the distribution of $(\mathbf{Z}^*)^T\mathbf{Z}^*$ is that of $\sum_i \lambda_i\chi_i^2$, where λ_i denote the nonzero eigenvalues of $\mathbf{V}^*$ and χ_i^2 are the 1-d.f. chi-square variables (Box 1954). The asymptotic covariance matrix of $\sqrt{N}\mathbf{Z}$ is idempotent with all $(k-1)(m-1)$ nonzero eigenvalues being equal to 1. Hence, the sum of NZ_{ij}^2, that is, $(\sqrt{N}\mathbf{Z}^T)(\sqrt{N}\mathbf{Z})$, has an asymptotic χ^2-distribution. In contrast to $(\mathbf{\Gamma V \Gamma})$, the matrix $(\mathbf{\Psi V \Psi})$ with $(k-1)(m-1)$ positive eigenvalues is not idempotent. Therefore, $N\mathbf{R}^T\mathbf{R}$ does not have an asymptotic χ^2-distribution. Box (1954) suggested that the distribution of weighted chi-square variables, $\sum w_i\chi_{(v_i)}^2$, where each chi square is with the degrees of freedom v_i, can be approximated by a scaled chi-square distribution, $\sigma\chi_{(d)}^2$, where

$$\sigma = \frac{\sum v_i w_i^2}{\sum v_i w_i} \tag{A7}$$

$$d = \frac{(\sum v_i w_i)^2}{\sum v_i w_i^2}. \tag{A8}$$

The degrees of freedom d need not be integral. In the case of a sum of correlations, all $v_i = 1$, and the weights are computed from the eigenvalues of

$$\mathbf{V}_R = (\mathbf{\Psi}\tilde{\mathbf{V}}\mathbf{\Psi}). \tag{A9}$$

Since only the sums of eigenvalues or of their squares are needed, and not the eigenvalues themselves, the computations simplify substantially:

$$\sum w_i = \mathrm{trace}(\mathbf{V}_R) = km \tag{A10}$$

$$\sum w_i^2 = \mathrm{trace}(\mathbf{V}_R\mathbf{V}_R). \tag{A11}$$

This makes use of the fact that eigenvalues of a squared matrix are given by squared eigenvalues of that matrix and that the trace of a symmetric matrix is given by the sum of its eigenvalues. Therefore, for our first scaled chi-square approximation we have

$$\frac{N\mathbf{R}^T\mathbf{R}}{\sigma} \overset{\text{app}}{\sim} \chi_{(d)}^2, \tag{A12}$$

where "$\overset{\text{app}}{\sim}$" stands for "approximately distributed" and

$$\sigma = \frac{\mathrm{trace}(\mathbf{V}_R\mathbf{V}_R)}{km} \tag{A13}$$

$$d = \frac{(km)^2}{\mathrm{trace}(\mathbf{V}_R\mathbf{V}_R)}. \tag{A14}$$

In the second and much simpler approximation, we set the degrees of freedom equal to $(k-1)(m-1)$, which is the number of nonredundant disequilibrium coefficients, $D_{ij} = p_{ij} - p_iq_j$, and note the expected values

$$\mathcal{E}(N\mathbf{R}^T\mathbf{R}) = N\frac{km}{N} = km$$

$$\mathcal{E}(\sigma\chi_{(d)}^2) = \sigma d$$

$$\mathrm{Var}(\sigma\chi_{(d)}^2) = \left[E(\chi_{(d)}^2)\right]^2 \frac{2}{d}. \tag{A15}$$

By matching moments, the scale parameter is found to be $\sigma = km/[(k-1)(m-1)]$. Thus, we obtain our second approximate distribution as

$$\frac{N(k-1)(m-1)}{km}\mathbf{R}^T\mathbf{R} \overset{\text{app}}{\sim} \chi_{(k-1)(m-1)}^2. \tag{A16}$$

Note that $\mathbf{R}^T\mathbf{R}/(km)$ is just the average squared correlation. Waples (2006) noted that approximately, the distribution of such a coefficient might be a chi square and that with k alleles per locus, the

TABLE B1

The 4 × 3 table of joint frequencies with the corresponding level of association

	$p_{ij}\backslash D'$			Sum
	0.0871\0.5	0.1567\0.5	0.1134\−0.5	0.357
	0.0133\−0.5	0.0240\−0.5	0.1697\0.5	0.207
	0.0107\−0.5	0.0192\−0.5	0.1359\0.5	0.166
	0.0174\−0.5	0.0313\−0.5	0.2213\0.5	0.270
Sum	0.128	0.231	0.640	

Absolute values of D' for this set of simulations are set to be ±0.5.

TABLE B2

The 4 × 3 table of joint frequencies with the corresponding level of association

	$p_{ij}\backslash D'$			Sum
	0.0844\0.5	0.0114\−0.5	0.0106\−0.5	0.106
	0.1390\−0.5	0.1534\0.5	0.1437\0.5	0.436
	0.0803\0.5	0.0108\−0.5	0.0101\−0.5	0.101
	0.2825\0.5	0.0381\−0.5	0.0356\−0.5	0.356
Sum	0.586	0.214	0.200	

Absolute values of D' for this set of simulations are set to be ±0.5.

number of independent comparisons (and thus the degrees of freedom) for a comparison of two loci should be $(k-1)^2$. Nevertheless, he did not provide a distribution explicitly.

APPENDIX B

Tables of population joint frequencies (p_{ij}) to provide a specified amount of association (measured by D') for the power study.

TABLE B3

The 5 × 5 table of joint frequencies with the corresponding level of association, $p_{ij}\backslash D'$

	$p_{ij}\backslash D'$					Sum
	0.1183\0.20	0.0233\−0.20	0.0434\−0.21	0.0529\−0.20	0.0385\−0.20	0.277
	0.0233\−0.20	0.0086\−0.20	0.0365\0.20	0.0196\−0.20	0.0142\−0.20	0.102
	0.0228\−0.20	0.0084\−0.20	0.0357\0.20	0.0192\−0.20	0.0139\−0.20	0.100
	0.0399\−0.19	0.0147\−0.20	0.0275\−0.20	0.0335\−0.20	0.0592\0.20	0.175
	0.0791\−0.19	0.0502\0.20	0.0545\−0.20	0.1143\0.20	0.0483\−0.20	0.346
Sum	0.284	0.105	0.198	0.239	0.174	

Absolute values of D' for this set of simulations are set to be in the range 0.19–0.21.