Am. J. Hum. Genet. 77:27–40, 2005

An Entropy-Based Statistic for Genomewide Association Studies

Jinying Zhao,[1] Eric Boerwinkle,[1] and Momiao Xiong[1,2]

[1]Human Genetic Center, University of Texas, Health Science Center at Houston, Houston; and [2]Laboratory of Theoretical Systems Biology, School of Life Science, Fudan University, Shanghai, China

Efficient genotyping methods and the availability of a large collection of single-nucleotide polymorphisms provide valuable tools for genetic studies of human disease. The standard χ^2 statistic for case-control studies, which uses a linear function of allele frequencies, has limited power when the number of marker loci is large. We introduce a novel test statistic for genetic association studies that uses Shannon entropy and a nonlinear function of allele frequencies to amplify the differences in allele and haplotype frequencies to maintain statistical power with large numbers of marker loci. We investigate the relationship between the entropy-based test statistic and the standard χ^2 statistic and show that, in most cases, the power of the entropy-based statistic is greater than that of the standard χ^2 statistic. The distribution of the entropy-based statistic and the type I error rates are validated using simulation studies. Finally, we apply the new entropy-based test statistic to two real data sets, one for the COMT gene and schizophrenia and one for the MMP-2 gene and esophageal carcinoma, to evaluate the performance of the new method for genetic association studies. The results show that the entropy-based statistic obtained smaller *P* values than did the standard χ^2 statistic.

Introduction

Genomewide association studies are emerging as a promising tool for genetic analysis of complex diseases (Risch and Merikangas 1996; Stumpf and Goldstein 2003; Carlson et al. 2004; Neale and Sham 2004). Such association studies depend on linkage disequilibrium (LD). However, population histories and evolutionary forces may generate aberrant LD patterns across marker loci and important differences in allele frequencies and LD patterns across populations (Morton and Collins 1998; Pritchard and Przeworsk 2001; Stephens et al. 2001; Freedman et al. 2004). An allele may show strong LD with the functional variants in one population but exhibit very weak LD in other populations (Carlson et al. 2004) or may show strong LD with one marker but exhibit weak LD with other nearby markers, even in the same population. These phenomena make it difficult to replicate genetic associations among populations and to obtain consistent results within a genomic region in the same population.

There are two ways to alleviate aberrant LD patterns, improve the power of association studies, and increase the probability of replications. One way is to construct haplotype blocks by studying LD patterns across the genome and to optimally select a set of robust tagSNPs such that all common variants are either directly genotyped or in strong LD with the genotyped tagSNPs (Goldstein 2001; Johnson et al. 2001; Stephens et al. 2001; Gabriel et al. 2002; Zhang et al. 2002, 2003*a*; Ke and Cardon 2003; Xiong et al. 2003). However, it is unclear whether the haplotype block patterns and tagSNPs are consistent among populations or among repeated sampling from within a population. Another way is to develop novel statistical methods for association studies. Currently, the two major statistical methods for association studies are to compare haplotype frequencies between affected and unaffected individuals (Chapman and Wijsman 1998; Akey et al. 2001), which is often referred to as the standard χ^2 test, and to compare haplotype similarities between affected and unaffected individuals (de Vries et al. 1996; van der Meulen and te Meerman 1997; Bourgain et al. 2000, 2001, 2002; Tzeng et al. 2003; Zhang et al. 2003*b*; Yu et al. 2004*b*). As Tzeng et al. (2003) pointed out, none of these frequency- or sharing-based methods is uniformly the most powerful.

The standard χ^2 statistic compares allele (or haplotype) frequencies—or linear transformations of allele (or haplotype) frequencies—and partially considers the variance-covariance structure of the allele or haplotype frequencies. Therefore, the standard χ^2 test statistic is not uniformly the most powerful (Tzeng et al. 2003). Amplifying the difference in allele (or haplotype) frequencies between cases and controls is key to increasing the power of current test statistics for association studies.

Received December 1, 2004; accepted for publication April 19, 2005; electronically published May 9, 2005.

Address for correspondence and reprints: Dr. Momiao Xiong, Human Genetics Center, University of Texas–Houston, P.O. Box 20334, Houston, TX 77225. E-mail: Momiao.Xiong@uth.tmc.edu

0002-9297/2005/7701-0004$15.00

One way to amplify the difference in allele (or haplotype) frequencies between cases and controls is to use nonlinear transformations of allele or haplotype frequencies, $f(P^A)$ and $f(P)$, where P^A is the frequency of the allele (or haplotype) in cases and P is the frequency in controls, and to construct new test statistics such that the statistics based on $f(P^A) - f(P)$ are larger than the statistics based on the difference in allele or haplotype frequencies, $P^A - P$. The nonlinear transformations of allele or haplotype frequencies should have the feature that the magnitude of the amplified difference in allele or haplotype frequencies will increase as the difference in allele or haplotype frequencies increases.

A typical nonlinear function of frequencies is Shannon entropy. Shannon entropy, originally defined in information theory (Shannon 1948), is used to measure the uncertainty removed or the information gained by performing an experiment. When it is applied to characterize DNA variation, entropy measures genetic diversity and extracts the maximal amount of information for a set of SNP markers (Hampe et al. 2003). Conditional entropy measures the average information gained, given the occurrence of specific events. Entropy of a genomic region carried by affected individuals or unaffected individuals is conditional entropy. The difference between affected and unaffected individuals in entropy of the SNP markers is a measure of the association of the markers with disease.

In this article, we first define the entropy of a set of SNP markers at a genomic region of interest and the partial entropy of a haplotype. We then define the conditional entropy and the partial entropy of the markers in affected individuals and describe the relationship between conditional entropy and LD. We present a novel statistic that is based on the concept of entropy, to test the association between SNP markers and disease, including a test of the association of marker alleles, haplotypes, multiple-marker loci, and haplotype blocks. The relationship between the entropy-based test statistic and the standard χ^2 test statistic is discussed.

Entropy and the Overall Measure of Multilocus LD

Entropy, proposed by Shannon (1948), measures the uncertainty of random variables or the degree of nonstructure of a system (Nothnagel et al. 2003). The entropy of a random variable X is defined as

$$S(X) = E[-\log P(X)] = -\sum_i P(x_i)\log P(x_i) ,$$

where $P(x_i)$ denotes the probability that the random variable X assumes the value x_i.

The concept of entropy can be used to study DNA variation at a marker locus and patterns of LD. First, we consider two marker loci, M_1 and M_2, with two alleles each. Assume locus M_1 has alleles A and a with frequencies P_A and P_a, respectively. Locus M_2 has alleles B and b with frequencies P_B and P_b, respectively. The frequencies of haplotypes AB, Ab, aB, and ab are denoted by P_{AB}, P_{Ab}, P_{aB}, and P_{ab}, respectively. Let δ denote the measure of LD between two loci and be defined as

$$\delta = P_{AB} - P_A P_B .$$

The entropy of the marker is defined as

$$S_{M_1} = -P_A \log P_A - P_a \log P_a ,$$

and the entropy of the haplotypes at two marker loci is defined as

$$S_{M_1M_2} = -P_{AB}\log P_{AB} - P_{Ab}\log P_{Ab} - P_{aB}\log_{aB} P - P_{ab}\log P_{ab} .$$

For convenience of presentation, a component of the entropy of the haplotypes at two marker loci is referred to as "partial entropy" of the specific haplotype. Let the partial entropy of haplotype AB, S_{AB}, be equal to $-P_{AB}\log P_{AB}$. In appendix A, we show that

$$S_{AB} \approx -P_A P_B \log(P_A P_B) - \delta(\log P_A + \log P_B + 1) - \frac{\delta^2}{2P_A P_B} .$$

The entropy of the haplotypes at marker loci M_1 and M_2 is given by

$$S_{M_1M_2} \approx S_{M_1} + S_{M_2} - \frac{\delta^2}{2P_A P_a P_B P_b}$$

(see appendix A), where the entropies of the alleles at marker loci M_1 and M_2 are $S_{M_1} = -P_A\log P_A - P_a\log P_a$ and $S_{M_2} = -P_B\log P_B - P_b\log P_b$, respectively.

Next, we consider multiple loci. Suppose that there are K marker loci that generate m haplotypes. Let $P_{H_{j_1\ldots j_k}}$ be the population frequency of haplotype $H_{j_1\ldots j_k}$ with a sequence of alleles $M_{j_1}\ldots M_{j_k}$, where the allele M_{j_i} at the ith locus is either allele 1 or allele 2. Let $P_{M_{j_i}}$ be the frequency of allele M_{j_i} in the population. Define an overall measure of the haplotype LD at the K loci as

$$\delta_{j_1\ldots j_k} = P_{H_{j_1\ldots j_k}} - P_{M_{j_1}}\cdots P_{M_{j_k}} \quad (1)$$

(Xiong et al. 2003). The partial entropy of haplotype

$H_{j_1\ldots j_k}$ is defined as $S_{H_{j_1\ldots j_k}} = -P_{H_{j_1\ldots j_k}} \log P_{H_{j_1\ldots j_k}}$, and the entropy of the haplotype at K marker loci is defined as

$$S_k = -\sum_{j_1}\sum_{j_2}\cdots\sum_{j_k} P_{H_{j_1\ldots j_k}} \log P_{H_{j_1\ldots j_k}} = \sum_{j_1}\cdots\sum_{j_k} S_{H_{j_1\ldots j_k}} .$$

In appendix A, we show that the partial entropy $S_{H_{j_1\ldots j_k}}$ and the entropy of the haplotype at K marker loci, S_k, can be approximated by

$$S_{H_{j_1\ldots j_k}} \approx -P_{M_{j_1}}\cdots P_{M_{j_k}} \log\left(P_{M_{j_1}}\cdots P_{M_{j_k}}\right) + \left[1 + \log\left(P_{M_{j_1}}\cdots P_{M_{j_k}}\right)\right]\delta_{j_1\ldots j_k} - \frac{\delta^2_{j_1\ldots j_k}}{2P_{M_{j_1}}\cdots P_{M_{j_k}}}$$

and

$$S_k \approx \sum_{i=1}^{k} S_{M_i} - \frac{1}{2}\sum_{j_1}\cdots\sum_{j_k}\frac{\delta^2_{j_1\ldots j_k}}{P_{M_{j_1}}\cdots P_{M_{j_k}}} ,$$

where $S_{M_i} = -\sum_{j=1}^{2} P_{M_{j_i}} \log P_{M_{j_i}}$ is the entropy of marker M_i. This shows that the entropy of the haplotype at K marker loci is the approximation of the sum of the entropies of all K marker loci and a function of the overall measures of all haplotypes.

Entropy in Affected Individuals

If the marker allele or haplotype is in LD with the disease locus, the frequency of the marker allele or haplotype in affected individuals and unaffected individuals will be different. The entropy of the haplotypes in affected individuals and unaffected individuals will also be different, and the difference can quantify the level of LD between the marker and the disease locus.

Let H_i be one of the m haplotypes at K marker loci. Let P_{H_i} and $P^{\mathrm{A}}_{H_i}$ be the frequency of haplotype H_i in unaffected individuals and affected individuals, respectively. Let S_{H_i} and $S^{\mathrm{A}}_{H_i}$ be the partial entropy of haplotype H_i in unaffected individuals and affected individuals, respectively. Then, we have

$$S_{H_i} = -P_{H_i}\log P_{H_i} \text{ and } S^{\mathrm{A}}_{H_i} = -P^{\mathrm{A}}_{H_i}\log P^{\mathrm{A}}_{H_i} .$$

In appendix B, we show that $S^{\mathrm{A}}_{H_i}$ can be approximated by

$$S^{\mathrm{A}}_{H_i} \approx S_{H_i} - b\delta_{H_iD}(1 + \log P_{H_i}) - \frac{b^2\delta^2_{H_iD}}{2P_{H_i}} ,$$

where δ_{H_iD} is the overall measure of LD between haplotype H_i and disease allele D.

Therefore, the difference in the partial entropy of the haplotype between the affected and unaffected individuals is given by

$$\Delta S_{H_i} = S_{H_i} - S^{\mathrm{A}}_{H_i} \approx b\delta_{H_iD}(1 + \log P_{H_i}) + \frac{b^2\delta^2_{H_iD}}{2P_{H_i}} . \quad (2)$$

Clearly, the gain in information about the association of the haplotype with disease is a function of the overall measure of LD between the haplotype and the disease locus. If the haplotype or marker loci are in linkage equilibrium with the disease allele, then the difference in the partial entropy of the haplotype (ΔS_{H_i}) between affected and unaffected individuals will be zero.

The Entropy-Based Statistic for Association Tests

In this section, we present an entropy-based statistic for case-control association studies. We begin with an introduction of notation. The first partial derivatives of the partial entropy of haplotype H_i with respect to the frequency of haplotype H_j, denoted by b_{ij}, is given as

$$b_{ii} = \frac{\partial S_{H_i}}{\partial P_{H_i}} = -1 - \log P_{H_i}$$

and

$$b_{ij} = \frac{\partial S_{H_i}}{\partial P_{H_j}} = 0 \quad (i \neq j) .$$

The $m \times m$ dimensional matrix of the first partial derivatives is denoted by $B = (b_{ij})_{m\times m}$, where m is the number of haplotypes, as defined above. The number of haplotypes follows a multinomial distribution, and the variance-covariance matrix is given by $2n_{\mathrm{G}}\Sigma$, where n_{G} is the number of unaffected individuals, $\Sigma = (\sigma_{ij})_{m\times m}$, $\sigma_{ii} = P_{H_i}(1 - P_{H_i})$, and $\sigma_{ij} = -P_{H_i}P_{H_j}$ $(i \neq j)$.

The above quantities can be similarly defined for the affected individuals and will be denoted by the additional superscript "A" in the corresponding quantities—that is,

$$b^{\mathrm{A}}_{ii} = -1 - \log P^{\mathrm{A}}_{H_i} , \quad b^{\mathrm{A}}_{ij} = 0 \ (i \neq j) ,$$

and $B^{\mathrm{A}} = (b^{\mathrm{A}}_{ij})_{m\times m}$. Likewise, $\sigma^{\mathrm{A}}_{ii} = P^{\mathrm{A}}_{H_i}(1 - P^{\mathrm{A}}_{H_i})$, $\sigma^{\mathrm{A}}_{ij} = -P^{\mathrm{A}}_{H_i}P^{\mathrm{A}}_{H_j}$ $(i \neq j)$, and $\Sigma^{\mathrm{A}} = (\sigma^{\mathrm{A}}_{ij})_{m\times m}$.

Let $S = [S_{H_1}\ldots S_{H_m}]^T$, $S^{\mathrm{A}} = [S^{\mathrm{A}}_{H_1}\ldots S^{\mathrm{A}}_{H_m}]^T$, $W = B\Sigma B^T$, and $W^{\mathrm{A}} = B^{\mathrm{A}}\Sigma^{\mathrm{A}}(B^{\mathrm{A}})^T$, where $S^{\mathrm{A}}_{H_i}$ is defined as in the previous section.

Let $\hat{S}$, $\hat{S}^{\mathrm{A}}$, $\hat{W}$, and $\hat{W}^{\mathrm{A}}$ be the estimators of S, S^{A}, W,

and W^{A}, respectively. The partial entropy–based statistic for an association test, denoted by T_{PE}, is defined as

$$T_{PE} = \left(\hat{S} - \hat{S}^{A}\right)^{T}\left(\frac{\hat{W}^{A}}{2n_{A}} + \frac{\hat{W}}{2n_{G}}\right)^{-1}\left(\hat{S} - \hat{S}^{A}\right) ,$$

where n_{A} and n_{G} are the number of affected and unaffected individuals, respectively. Because the matrix

$$\left(\frac{\hat{W}^{A}}{2n_{A}} + \frac{\hat{W}}{2n_{G}}\right)$$

may not be of full rank, the generalized inverse of the matrix will be used in situations where the inverse of the matrix does not exist.

We can show that, under the null hypothesis of no association between the K marker loci and disease, when the frequencies of the haplotypes are not zero, T_{PE} is asymptotically distributed as a central $\chi^{2}_{(m-1)}$ distribution (appendix C). Because application of theorem 1.9 in Lehmann (1983, p. 344) requires that entropy be continuously differentiable with respect to the frequencies of the haplotypes, theorem 1.9 cannot be applied when the frequency of the haplotype is zero. If the frequency of the haplotype in either cases or controls is zero, then the haplotype needs to be grouped with other haplotypes. For example, rare haplotypes can be grouped with the most similar haplotype. Under the alternative hypothesis that there is an association between the K marker loci and the disease locus, T_{PE} is asymptotically distributed as a noncentral $\chi^{2}_{(m-1)}$ distribution with the following noncentrality parameter:

$$\lambda_{PE} = (S - S^{A})^{T}\left(\frac{W^{A}}{2n_{A}} + \frac{W}{2n_{G}}\right)^{-1}(S - S^{A}) .$$

By invoking the relationship between the partial entropy of the haplotypes and the measure of LD discussed in the previous section, the noncentrality parameter λ_{PE} can be further reduced to

$$\lambda_{PE} \approx b^{2}\delta_{1}^{T}\Sigma_{0}^{-1}\delta_{1} + R_{PE}$$

(see appendix C), where

$$b = \frac{P_{D}(f_{11} - f_{12}) + P_{d}(f_{12} - f_{22})}{P(A)} ,$$

$$P(A) = P_{D}^{2}f_{11} + 2P_{D}P_{d}f_{12} + P_{d}^{2}f_{22} ,$$

$$\Sigma_{0} = \frac{\Sigma}{2n_{G}} + \frac{\Sigma^{A}}{2n_{A}} ,$$

$$\delta_{1} = \left[\delta_{H_{1}D} \ldots \delta_{H_{m}D}\right]^{T} ,$$

and other parameters are as given in appendix C.

Relationship between the Entropy-Based Statistic and the Standard χ^{2} Test Statistic

Entropy-based statistics have a close relationship with the standard χ^{2} test statistic. To see this, we first derive the standard χ^{2} statistic using theorem 1.9 in Lehmann (1983, p. 344), which leads to an entropy-based statistic. Let $f_{i}(\hat{P}_{H_{1}} \ldots \hat{P}_{H_{m}}) = \hat{P}_{H_{i}}$. Then, the matrix

$$B = \left(\frac{\partial f_{i}}{\partial P_{H_{j}}}\right)_{m \times m} = I .$$

The variance-covariance matrix of the haplotype frequencies is given by $(1/n_{G})\Sigma$, where

$$\Sigma = \begin{bmatrix} P_{H_{1}}(1 - P_{H_{1}}) & -P_{H_{1}}P_{H_{2}} & \cdots & -P_{H_{1}}P_{H_{m}} \\ \vdots & \vdots & & \vdots \\ -P_{H_{m}}P_{H_{1}} & -P_{H_{m}}P_{H_{2}} & \cdots & P_{H_{m}}(1 - P_{H_{m}}) \end{bmatrix} .$$

If we ignore the terms $-P_{H_{i}}^{2}$ and $-P_{H_{i}}P_{H_{j}}(i,j = 1, \ldots, m)$ in the elements of matrix Σ, then the variance-covariance matrix Σ is reduced to

$$\Sigma \approx \begin{bmatrix} P_{H_{1}} & 0 & \cdots & 0 \\ 0 & P_{H_{2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_{H_{m}} \end{bmatrix} = \mathrm{diag}(P_{H_{i}}) .$$

Similarly, we have $\Sigma^{A} \approx \mathrm{diag}(P_{H_{i}}^{A})$ for the affected individuals. Let $P = [P_{H_{1}} \ldots P_{H_{m}}]^{T}$, and $P^{A} = [P_{H_{1}}^{A} \ldots P_{H_{m}}^{A}]^{T}$. Define the test statistic T as

$$T = (P - P^{A})^{T}\left(\frac{\Sigma}{2n_{G}} + \frac{\Sigma^{A}}{2n_{A}}\right)^{-1}(P - P^{A}) ,$$

which can be reduced to

$$T = \sum_{i=1}^{m} \frac{(P_{H_i} - P_{H_i}^{A})^2}{\frac{P_{H_i}}{2n_G} + \frac{P_{H_i}^{A}}{2n_A}} .$$

If we assume that the numbers of affected and unaffected individuals are equal (i.e., that $n_A = n_G = n$), then the χ^2 test statistic T can be further reduced to

$$T = 2n \sum_{i=1}^{m} \frac{(P_{H_i} - P_{H_i}^{A})^2}{P_{H_i} + P_{H_i}^{A}} ,$$

which is exactly the formula of the standard χ^2 test statistic (Chapman and Wijsman 1998).

Applying theorem 1.9 (Lehman 1983, p. 344) with the function form $f_i(\hat{P}_{H_1} \ldots \hat{P}_{H_m}) = -P_{H_i} \log P_{H_i}$ will yield the test statistic T_{PE}. Therefore, the difference between the standard χ^2 test statistic and the entropy-based statistic is that the χ^2 test statistic uses a linear function of haplotype frequencies, whereas the entropy-based test statistic uses a nonlinear function of haplotype frequencies. The difference between these two test statistics lies in the different mathematical forms of the haplotype frequencies when theorem 1.9 is used to construct the statistics.

In appendix C, we show that the noncentrality parameter λ_T of the standard χ^2 test statistic is given by $\lambda_T = b^2 \delta_1^T \Sigma_0^{-1} \delta_1$. This is the first term in the noncentrality parameter λ_{PE} of the test statistic based on partial entropy of the haplotypes.

Results

Distribution of the Entropy-Based Statistic

In the previous sections, we have shown that when the sample size is large enough to apply large-sample theory, the distribution of the entropy-based statistic under the null hypothesis of no association is asymptotically a central χ^2 distribution. To examine the validity of this statement, we performed a series of simulation studies. The computer program SNaP (Nothnagel 2002) was used to generate haplotypes of the sample individuals. Two data sets with a single haplotype block each were simulated. The first data set has two marker loci that generated four haplotypes with frequencies 0.2952, 0.2562, 0.1957, and 0.2529. The second data set has six marker loci that generated eight haplotypes with frequencies 0.1820, 0.1461, 0.1406, 0.1291, 0.1211, 0.1107, 0.0817, and 0.0887. For each data set, 20,000 individuals who were divided into equal groups of cases and controls were generated in the general population.

To examine whether the asymptotic results of the entropy-based test statistic still hold for a small sample size under the null hypothesis of no association, 200 individuals each were randomly sampled from the cases and controls. A total of 10,000 simulations were performed. In each simulation, the entropy-based test statistic T_{PE} was calculated. Figure 1*A* and 1*B* plot the histograms of the test statistic T_{PE} with the use of two-SNP haplotypes and six-SNP haplotypes, respectively. It can be seen that the distributions of T_{PE} are similar to the theoretical central χ^2 distributions, even under the scenario of a smaller sample size. Table 1 summarizes the type I error rates of the test statistic T_{PE} for sample sizes from 100 to 500 individuals with the use of two-SNP and six-SNP haplotypes. Table 1 shows that the estimated type I error rates of the entropy-based test statistic were not ap-

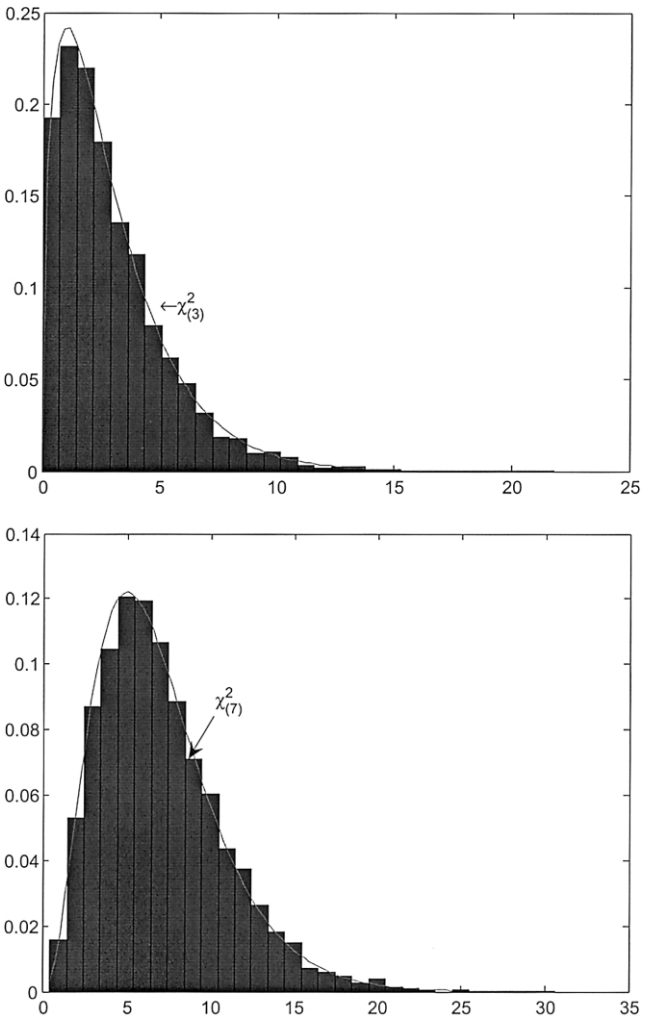

Figure 1 Distribution of the test statistic T_{PE} with the use of two-SNP haplotypes (*A*) and six-SNP haplotypes (*B*). $\chi^2_{(3)}$ and $\chi^2_{(7)}$ indicate χ^2 distribution with 3 df and 7 df, respectively.

Table 1

Estimated Type I Error Rates of the Test Statistic T_{PE} for 10,000 Simulations

Sample Size	Estimated Type I Error Rate for					
	Two-SNP Haplotypes			Six-SNP Haplotypes		
	$\alpha = .05$	$\alpha = .01$	$\alpha = .001$	$\alpha = .05$	$\alpha = .01$	$\alpha = .001$
100	.0460	.0078	.0006	.0542	.0112	.0008
200	.0480	.0088	.0012	.0488	.0106	.0010
300	.0460	.0103	.0011	.0512	.0092	.0014
400	.0478	.0100	.0008	.0542	.0102	.0008
500	.0478	.0084	.0008	.0488	.0098	.0012

preciably different from the nominal levels $\alpha = 0.05$, $\alpha = 0.01$, and $\alpha = 0.001$.

Power of the Entropy-Based Test Statistic and the Standard χ^2 Test Statistic

The power of an association test statistic depends on a number of parameters, such as the measure of LD between haplotypes and disease alleles, the sample size, and the model of disease inheritance. We compared the power of the entropy-based test statistic with that of the standard χ^2 test statistic. The markers are assumed to be biallelic (i.e., SNPs). Specifically, we considered two marker loci and a disease locus that is located in the middle of two markers. We considered three disease models: recessive, dominant, and genotype relative-risk models, in which the genotype relative risk for genotypes *Dd* and *DD* is *r* and r^2 times greater, respectively, than that of the genotype *dd* (Risch and Merikangas 1996).

Exact analytical methods were used for calculation of power. The average haplotype frequencies in the affected and unaffected individuals were calculated by equations (1) and (4) from Akey et al. (2001). The power of the entropy-based statistic and the standard χ^2 statistic (Chapman and Wijsman 1998) ($\alpha = 0.001$), with the use of four haplotypes generated by two marker loci as a function of the genetic distance between the disease locus and its flanking marker loci for recessive, dominant, and genotype relative-risk models are shown in figure 2*A,* 2*B,* and 2*C,* respectively. The data demonstrate that the power of the entropy-based statistic for all three disease models is higher than the power of the standard χ^2 statistic. In appendix C, we show that the noncentrality parameter of the entropy-based statistic is approximately equal to the summation of the noncentrality parameter of the standard χ^2 test statistic and R_{PE}. But, R_{PE} is not always positive, and the entropy-based statistic does not monotonically increase with increasing allele-frequency differences, which implies that the entropy-based statistic is not uniformly more powerful than the standard χ^2 test. In fact, when the difference in allele frequencies is very large (i.e., when $|P^A - P| > 0.7$), the χ^2 test is more powerful than the entropy-based statistic proposed here.

Application to Real-Data Examples

To evaluate its performance, we applied the entropy-based test to two real data sets. The first example is a test of association between the catechol-O-methyltransferase (COMT) gene and schizophrenia (Shifman et al. 2002). The COMT gene plays an important biochemical function in the metabolism of catecholamine neurotransmitters and is being increasingly recognized as a contributor to velocardiofacial syndrome, which is associated with a high rate of psychosis (Shifman et al. 2002). The data were from a large case-control study of schizophrenia in the Ashkenazi Jewish population. Three SNPs within the COMT gene were typed. In table 2, we present the *P* values of the entropy-based statistic for testing the association of two-SNP haplotypes (generated from two SNP markers) and three-SNP haplotypes (generated from three SNP markers) with schizophrenia. For comparison, table 2 also includes the results of Shifman et al. (2002) that were obtained by the usual χ^2 test. It is evident that *P* values of the entropy-based test are smaller than those of the χ^2 test.

The second example studied the association between functional haplotypes in the promoter of the matrix metalloproteinase-2 (MMP-2) gene and esophageal squamous cell carcinoma in the Chinese Han population (Yu et al. 2004*a*). Two SNPs (−1306C/T and −735C/T) in the MMP-2 gene were typed in 527 patients with esophageal cancer and 777 controls. Frequencies of two-SNP haplotypes and *P* values obtained by the entropy-based test and standard χ^2 test are given in table 3. Again, the entropy-based test obtained a smaller *P* value compared with that obtained by the standard χ^2 test.

Discussion

Completion of the International HapMap Project will provide a powerful tool for identifying genes that contribute to complex disease (Collins 2004). To efficiently use DNA variants for genetic studies of complex diseases, the field of human genetics needs to accomplish two tasks: choose suitable marker sets and develop robust methods of statistical analysis (Wall and Pritchard 2003). The purpose of this report is to present an analytical method of assessing the relationship between DNA variation and disease in case-control association studies.

The standard χ^2 statistic is based on a linear function of haplotype or allele frequencies. Its drawback is a decrease in power as the number of degrees of freedom increases, which arises as a consequence of using a dense set of SNPs. There are two ways to increase the power

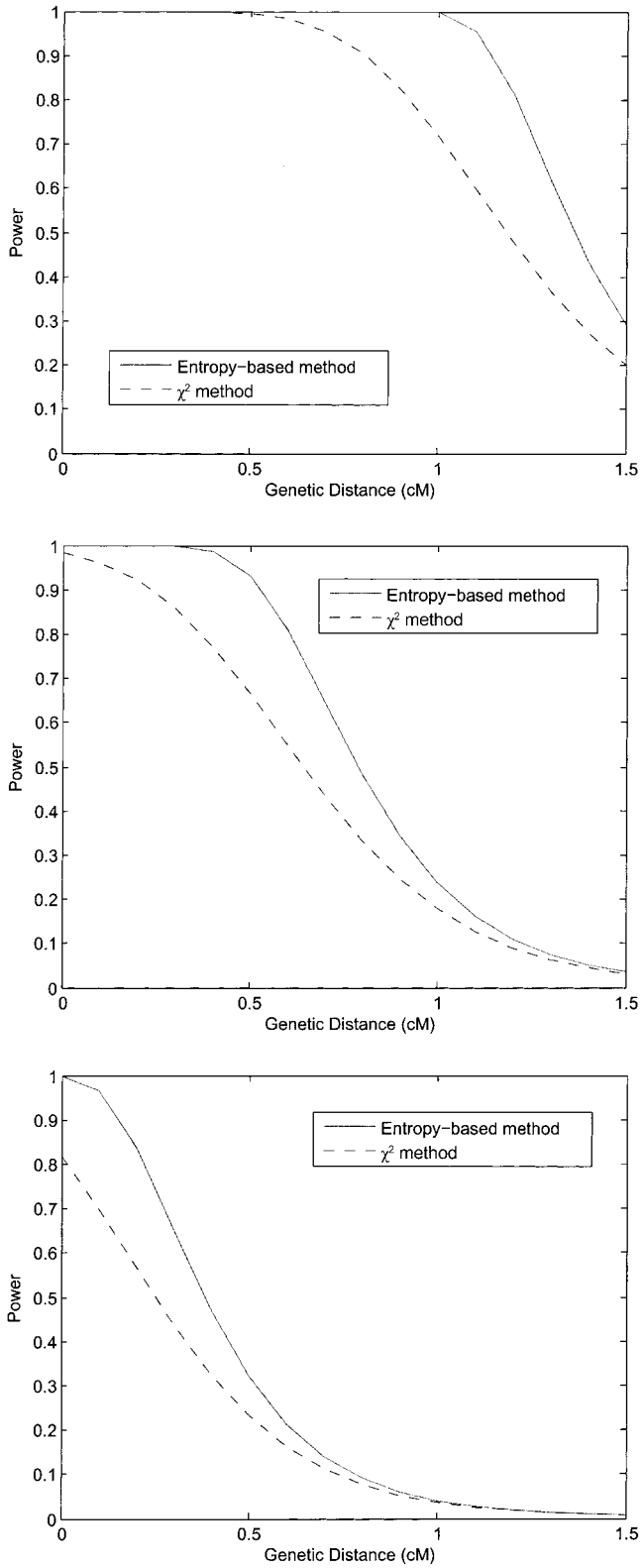

Figure 2 *A*, Power of the entropy-based test statistic and the standard χ^2 test statistic with a significance level of $\alpha = 0.001$, as a function of the genetic distance between the marker and disease loci, for a recessive disease (*A*) and a dominant disease (*B*) under the assumption that $N = 100$, $t = 100$ generations, the frequencies of the minor alleles at both of the marker loci are equal to 0.1, and $P_D = 0.1$. *C*, Power of the entropy-based test statistic and the standard χ^2 test statistic with a significance level of $\alpha = 0.001$ for a disease with genotype relative risk $r = 4$, as a function of the genetic distance between the marker and disease loci, under the assumption that $N = 200$, $t = 100$ generations, the frequencies of the minor alleles at the first and second marker loci are equal to 0.4 and 0.1, respectively, and $P_D = 0.2$.

Table 2

Tests of Association between COMT Haplotypes and Schizophrenia

Haplotype Size and Markers	*P* Value for χ^2			*P* Value for Entropy		
	Males	Females	All	Males	Females	All
Two-SNP haplotype						
rs737865, rs165599	.0093	.0014	.00014	.00012	1.5×10^{-6}	1.9×10^{-9}
rs737865, rs165688	.017	.046	.0057	.00018	.012	2.7×10^{-6}
rs165599, rs165688	.096	.0012	.0011	.0035	2.7×10^{-5}	2.9×10^{-6}
Three-SNP haplotype						
rs165688, rs737865, rs165599	.0084	.0069	.00045	8.4×10^{-9}	5.7×10^{-6}	1.5×10^{-12}

of an association test statistic. One is to find appropriate mathematical forms of the haplotype or allele frequencies that can be used to develop test statistics with higher power. Another is to reduce the degrees of freedom. Most publications in this field have focused on reducing the degrees of freedom. This report focuses on developing new entropy-based statistics to amplify the difference in allele or haplotype frequencies to increase power.

At its most fundamental level, the case-control study design provides a forum to compare allele frequencies or the transformation of allele frequencies between groups. Although the usual χ^2 test statistic is a quadratic function of the difference in allele frequencies, the difference represents a linear function of allele frequencies. There are two ways to amplify the difference in allele frequencies. One way is to make a nonlinear transformation of allele frequencies. Another way is to make a nonlinear transformation of the difference in allele frequencies between cases and controls. Although the second way can amplify the difference between allele frequencies, under the null hypothesis of zero difference, the entropy of the difference between allele frequencies is no longer differentiable at zero. Therefore, asymptotic theory of the nonlinear transformation of normal random variables cannot be applied to such statistics. The entropy-based statistic compares the difference in values of a nonlinear function of allele frequencies between cases and controls, in the hope that the difference will be larger than that of a linear function of allele frequencies between cases and controls.

Table 3

Tests of Association between MMP-2 Haplotypes and Esophageal Cancer

Haplotype		Frequency in Cases	Frequency in Controls
−1306	−735		
T	T	.0057	.0335
T	C	.1120	.1313
C	T	.2116	.2181
C	C	.6708	.6171

Note.—Overall *P* value is 7.03×10^{-6} for the χ^2 statistic and 3.24×10^{-8} for the entropy-based statistic.

Entropy measures the information contained in a stochastic process and can be used to measure haplotype diversity. In this report, we show that differences in the entropy of haplotypes between affected and unaffected individuals quantify the overall level of LD between the marker, haplotypes, and the disease locus. We note that the entropy of the observed haplotypes is a nonlinear function of haplotype frequencies. The entropy of haplotypes at *K* marker loci quantifies the information of all haplotypes generated at the *K* marker loci. The calculated entropy of the haplotypes depends on the choice of mathematical functions of the haplotype frequencies. We introduced the concept of partial entropy of a haplotype. If we compare the difference between affected and unaffected individuals in the partial entropy of a haplotype, we can test the association of that particular haplotype with disease. We can also test the association of a set of haplotypes or multiple marker loci by comparing differences in the partial entropy of the set between affected and unaffected individuals.

To use an entropy-based statistic to test the association of haplotypes with disease, we first need to study the distribution of the test statistic under the null hypothesis of no association. By simulation, we show that the distribution of the entropy-based test statistic is close to a χ^2 distribution, even for small sample sizes. To validate the test statistic and to estimate the false-positive rate, we calculated the type I error rates of the entropy-based statistic by simulation. The results showed that the type I error rates were close to the nominal significance levels, which implies that the test for association is valid for a single homogeneous population.

An important property of a test statistic for genetic association studies is power. We show, by analytical methods, that the power of the entropy-based statistic is higher than the power of the standard χ^2 statistic in most cases. The power gap between these two test statistics increases as the number of haplotypes increases (data not shown). However, the entropy-based statis-

tic is not more powerful in all situations. When the difference in allele frequencies is very large, that is, $|P^A - P| > 0.7$, the χ^2 statistic is more powerful than the entropy-based statistic. However, such differences in allele frequencies between cases and controls are practically unheard of in real-world studies of common human diseases. The entropy-based statistic was applied to real data to test the association between COMT haplotypes and schizophrenia and the association between the MMP-2 gene and esophageal squamous cell carcinoma. The results of real-data analyses demonstrated that, for case-control studies, P values of the entropy-based statistic are smaller than those of the standard χ^2 test.

Asymptotical theory for nonlinear transformation of normal variables requires that entropy be continuously differentiable with respect to the frequencies of the haplotype. When the frequency of the haplotype in either cases or controls is zero, entropy is no longer continuously differentiable. In this case, the haplotype needs to be grouped with other haplotypes. However, such grouping may result in loss of statistical power, depending on the haplotype effects of the grouped haplotypes.

The entropy of a haplotype is a nonlinear function of haplotype frequencies. Other mathematical functions of haplotype frequencies should be investigated for designing novel test statistics in the future. This will open new ways of developing more powerful statistics for tests of association. The completion of the International HapMap Project and advances in genotyping technologies bode well for large-scale whole-genome association studies. Development of robust methods for relating considerable genomic information to risk of disease is urgently needed.

Acknowledgments

We thank Dr. Sagiv Shifman and Dr. Ariel Darvasi, for providing the detailed data for schizophrenic haplotype analyses. We also thank two anonymous reviewers for helpful comments on the manuscript, which led to much improvement of the article. M.X. is supported by National Institutes of Health–National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIH-NIAMS) grant IP50AR44888, grant HL74735, and NIH grant ES09912. J.Z. is supported by NIH grant ES09912. E.B. is supported by a grant from the National Heart, Lung and Blood Institute.

Appendix A

First, we consider the partial entropy of a haplotype at two marker loci. By definition, the partial entropy of haplotype AB is

$$S_{AB} = -P_{AB} \log P_{AB} \ . \tag{A1}$$

But,

$$P_{AB} = P_A P_B + \delta \ . \tag{A2}$$

Substituting P_{AB} from equation (A2) into equation (A1) yields

$$S_{AB} = -(P_A P_B + \delta) \log (P_A P_B + \delta) \ ,$$

which can be simplified to

$$\begin{aligned} S_{AB} &= -(P_A P_B + \delta) \log (P_A P_B) - (P_A P_B + \delta) \log \left(1 + \frac{\delta}{P_A P_B}\right) \\ &\approx -P_A P_B (\log P_A + \log P_B) - \delta (\log P_A + \log P_B) - (P_A P_B + \delta) \left(\frac{\delta}{P_A P_B} - \frac{\delta^2}{2 P_A^2 P_B^2}\right) \\ &\quad \left[\text{by } \log (1 + x) \approx x - \frac{x^2}{2}\right] \\ &\approx -P_A P_B (\log P_A + \log P_B) - \delta (\log P_A + \log P_B + 1) - \frac{\delta^2}{2 P_A P_B} \ . \end{aligned}$$

The above formula is derived by Taylor expansion of the logarithm function. The assumption for Taylor expansion of the logarithm function is that the argument x should be small. When the frequencies of alleles are small, δ/P_AP_B may be large. This will violate the assumption of the Taylor expansion and makes approximation inaccurate.

Similarly, we have

$$S_{Ab} \approx -P_AP_b(\log P_A + \log P_b) + \delta(\log P_A + \log P_b + 1) - \frac{\delta^2}{2P_AP_b} ,$$

$$S_{aB} \approx -P_aP_B(\log P_a + \log P_B) + \delta(\log P_a + \log P_B + 1) - \frac{\delta^2}{2P_aP_B} , \text{ and}$$

$$S_{ab} \approx -P_aP_b(\log P_a + \log P_b) - \delta(\log P_a + \log P_b + 1) - \frac{\delta^2}{2P_aP_b} .$$

The entropy of the haplotypes at marker loci M_1 and M_2 is equal to

$$\begin{aligned} S_{M_1M_2} &\approx -P_A \log P_A - P_a \log P_a - P_B \log P_B - P_b \log P_b - \frac{\delta^2}{2}\left(\frac{1}{P_AP_B} + \frac{1}{P_AP_b} + \frac{1}{P_aP_B} + \frac{1}{P_aP_b}\right) \\ &= S_{M_1} + S_{M_2} - \frac{\delta^2}{2P_AP_aP_BP_b} , \end{aligned}$$

where S_{M_1} and S_{M_2} are the entropies of marker loci M_1 and M_2, respectively.

Next, we consider the partial entropy of a haplotype at multiple marker loci. By definition, partial entropy of haplotype $H_{j_1\ldots j_k}$ is defined as

$$S_{H_{j_1\ldots j_k}} = -P_{H_{j_1\ldots j_k}} \log P_{H_{j_1\ldots j_k}} .$$

But, from equation (1), we have

$$\begin{aligned} S_{H_{j_1\ldots j_k}} &= -(P_{M_{j_1}} \cdots P_{M_{j_k}} + \delta_{j_1\ldots j_k}) \log (P_{M_{j_1}} \cdots P_{M_{j_k}} + \delta_{j_1\ldots j_k}) \\ &= -(P_{M_{j_1}} \cdots P_{M_{j_k}} + \delta_{j_1\ldots j_k}) \log (P_{M_{j_1}} \cdots P_{M_{j_k}}) - (P_{M_{j_1}} \cdots P_{M_{j_k}} + \delta_{j_1\ldots j_k}) \log\left(1 + \frac{\delta_{j_1\ldots j_k}}{P_{M_{j_1}} \cdots P_{M_{j_k}}}\right) \\ &\approx -P_{M_{j_1}} \cdots P_{M_{j_k}} \log (P_{M_{j_1}} \cdots P_{M_{j_k}}) - \delta_{j_1\ldots j_k}\left[\log (P_{M_{j_1}} \cdots P_{M_{j_k}}) + 1\right] - \frac{\delta^2_{j_1\ldots j_k}}{2P_{M_{j_1}} \cdots P_{M_{j_k}}} . \end{aligned}$$

Summarizing $S_{H_{j_1\ldots j_k}}$ over all possible $j_1 \ldots j_k$ yields

$$S_K \approx \sum_{i=1}^{K} S_{M_i} - \frac{1}{2} \sum_{j_1} \cdots \sum_{j_k} \frac{\delta^2_{j_1\ldots j_k}}{P_{M_{j_1}} \cdots P_{M_{j_k}}}$$

(note that $\sum_{j_1} \cdots \sum_{j_k} \delta_{j_1\ldots j_k} = 0$), where $S_{M_i} = -\sum_{j=1}^{2} P_{M_{j_i}} \log P_{M_{j_i}}$ is the entropy of marker M_i.

Appendix B

By definition,

$$S^{\mathrm{A}}_{H_i} = -P^{\mathrm{A}}_{H_i} \log P^{\mathrm{A}}_{H_i} , \text{ where } P^{\mathrm{A}}_{H_i} = \frac{P(H_i,A)}{P(A)} . \tag{B1}$$

If we assume Hardy-Weinberg equilibrium, then the prevalence of disease $P(A) = P_D^2 f_{11} + 2P_DP_d f_{12} + P_d^2 f_{22}$, where

P_D and P_d denote the frequencies of alleles D and d at disease locus, respectively, and f_{11}, f_{12}, and f_{22} are the penetrances of genotype DD, Dd, and dd, respectively.

By theorem of total probability (Ross 1997),

$$\begin{aligned} P(H_i,A) &= P(H_i,DD,A) + P(H_i,Dd,A) + P(H_i,dd,A) \\ &= P_{H_iD}P_D f_{11} + (P_{H_iD}P_d + P_{H_id}P_D)f_{12} + P_{H_id}P_d f_{22} \ . \end{aligned}$$

Therefore,

$$P_{H_i}^{\mathrm{A}} = a_1 P_{H_iD} + a_2 P_{H_id} \ , \tag{B2}$$

where

$$a_1 = \frac{P_D f_{11} + P_d f_{12}}{P(A)} \quad \text{and} \quad a_2 = \frac{P_D f_{12} + P_d f_{22}}{P(A)} \ .$$

Note that

$$P_{H_iD} = P_{H_i}P_D + \delta_{H_iD} \quad \text{and} \quad P_{H_id} = P_{H_i}P_d + \delta_{H_id} \tag{B3}$$

(Xiong et al. 2003), where δ_{H_iD} and δ_{H_id} are the overall measure of LD between haplotype H_i and disease allele D and between haplotype H_i and allele d, respectively. Substituting P_{H_iD} and P_{H_id} from equation (B3) into equation (B2) leads to

$$\begin{aligned} P_{H_i}^{\mathrm{A}} &= a_1(P_{H_i}P_D + \delta_{H_iD}) + a_2(P_{H_i}P_d + \delta_{H_id}) \\ &= P_{H_i}(a_1P_D + a_2P_d) + a_1\delta_{H_iD} + a_2\delta_{H_id} \ . \end{aligned} \tag{B4}$$

Since $a_1P_D + a_2P_d = 1$ and $\delta_{H_id} = -\delta_{H_iD}$, the above equation (B4) is simplified to

$$P_{H_i}^{\mathrm{A}} = P_{H_i} + (a_1 - a_2)\delta_{H_iD} = P_{H_i} + b\delta_{H_iD} \ , \tag{B5}$$

where $b = a_1 - a_2$. Substituting $P_{H_i}^{\mathrm{A}}$ from equation (B5) into equation (B1) yields

$$\begin{aligned} S^{\mathrm{A}} = S_{H_i}^{\mathrm{A}} &= -(P_{H_i} + b\delta_{H_iD})\log(P_{H_i} + b\delta_{H_iD}) \\ &= -(P_{H_i} + b\delta_{H_iD})\log\left[P_{H_i}\left(1 + \frac{b\delta_{H_iD}}{P_{H_i}}\right)\right] \\ &= -(P_{H_i} + b\delta_{H_iD})\log P_{H_i} - (P_{H_i} + b\delta_{H_iD})\log\left(1 + \frac{b\delta_{H_iD}}{P_{H_i}}\right) \\ &\approx -P_{H_i}\log P_{H_i} - b\delta_{H_iD}\log P_{H_i} - (P_{H_i} + b\delta_{H_iD})\left(\frac{b\delta_{H_iD}}{P_{H_i}} - \frac{b^2\delta_{H_iD}^2}{2P_{H_i}^2}\right) \\ &= -P_{H_i}\log P_{H_i} - b\delta_{H_iD}(1 + \log P_{H_i}) - \frac{b^2\delta_{H_iD}^2}{2P_{H_i}} \\ &= S_{H_i} - b\delta_{H_iD}(1 + \log P_{H_i}) - \frac{b^2\delta_{H_iD}^2}{2P_{H_i}} \ . \end{aligned}$$

Appendix C

Recall that

$$S = [S_{H_1} \dots S_{H_m}]^T \text{ and } S^{\mathrm{A}} = [S^{\mathrm{A}}_{H_1} \dots S^{\mathrm{A}}_{H_m}]^T .$$

From equation (2), we have

$$S - S^{\mathrm{A}} \approx U , \tag{C1}$$

where $U = [U_1 \dots U_m]^T$ and $U_i = b(\log P_{H_i} + 1)\delta_{H_iD} + b^2\delta^2_{H_iD}/2P_{H_i}$. Next, we calculate the Jacobian matrix B^{A}. By definition,

$$\begin{aligned} b^{\mathrm{A}}_{ii} &= \frac{\partial S^{\mathrm{A}}_{H_i}}{\partial P^{\mathrm{A}}_{H_i}} \\ &= -1 - \log P^{\mathrm{A}}_{H_i} \\ &= -1 - \log (P_{H_i} + b\delta_{H_iD}) \\ &\approx -1 - \log P_{H_i} - \frac{b\delta_{H_iD}}{P_{H_i}} \quad [\text{by } \log(1+x) \approx x] \\ &= b_{ii} - \frac{b\delta_{H_iD}}{P_{H_i}} , \end{aligned}$$

where $b_{ii} = -1 - \log P_{H_i}$. Let $G_i = -b\delta_{H_iD}/P_{H_i}$ and $G = \mathrm{diag}(G_1 \dots G_m)$. Then, we have

$$B^{\mathrm{A}} \approx B + G . \tag{C2}$$

Next, we study the relationship between covariance matrix Σ^{A} in the affected individuals and covariance matrix Σ in the unaffected individuals. Note that

$$P^{\mathrm{A}}_{H_i}(1 - P^{\mathrm{A}}_{H_i}) = (P_{H_i} + b\delta_{H_iD})(1 - P_{H_i} - b\delta_{H_iD}) \approx P_{H_i}(1 - P_{H_i}) + b(1 - 2P_{H_i})\delta_{H_iD}$$

and

$$-P^{\mathrm{A}}_{H_i}P^{\mathrm{A}}_{H_j} = -(P_{H_i} + b\delta_{H_iD})(P_{H_j} + b\delta_{H_jD}) \approx -P_{H_i}P_{H_j} - b(P_{H_i}\delta_{H_jD} + P_{H_j}\delta_{H_iD}) .$$

Thus, we have $\Sigma^{\mathrm{A}} \approx \Sigma + bD$, where

$$D = \begin{bmatrix} (1 - 2P_{H_1})\delta_{H_1D} & -(P_{H_1}\delta_{H_2D} + P_{H_2}\delta_{H_1D}) & \cdots & -(P_{H_1}\delta_{H_mD} + P_{H_m}\delta_{H_1D}) \\ \vdots & \vdots & \cdots & \vdots \\ -(P_{H_m}\delta_{H_1D} + P_{H_1}\delta_{H_mD}) & -(P_{H_m}\delta_{H_2D} + P_{H_2}\delta_{H_mD}) & \cdots & (1 - 2P_{H_m})\delta_{H_mD} \end{bmatrix} .$$

Now we are ready to calculate the quantity $\frac{W}{2n_{\mathrm{G}}} + \frac{W^{\mathrm{A}}}{2n_{\mathrm{A}}}$. With the definition of W^{A} and equation (C2), we have

$$W^{\mathrm{A}} = B^{\mathrm{A}}\Sigma^{\mathrm{A}}(B^{\mathrm{A}})^T \approx (B + G)\Sigma^{\mathrm{A}}(B + G)^T = B\Sigma^{\mathrm{A}}B^T + R ,$$

where $R = B\Sigma^{A}G^{T} + G\Sigma^{A}B^{T} + G\Sigma^{A}G^{T}$. Thus,

$$\begin{aligned}\frac{W}{2n_{G}} + \frac{W^{A}}{2n_{A}} &\approx \frac{1}{2n_{G}}B\Sigma B^{T} + \frac{1}{2n_{A}}B\Sigma^{A}B^{T} + \frac{1}{2n_{A}}R \\ &= B\left(\frac{\Sigma}{2n_{G}} + \frac{\Sigma^{A}}{2n_{A}}\right)B^{T} + \frac{1}{2n_{A}}R \\ &= W_{0} + \frac{1}{2n_{A}}R\ ,\end{aligned}$$

where $\Sigma_{0} = \frac{\Sigma}{2n_{G}} + \frac{\Sigma^{A}}{2n_{A}}$ and $W_{0} = B\Sigma_{0}B^{T}$. But,

$$\left(W_{0} + \frac{1}{2n_{A}}R\right)^{-1} = W_{0}^{-1} - W_{0}^{-1}(W_{0}^{-1} + 2n_{A}R^{-1})^{-1}W_{0}^{-1}$$

(Anderson 1984). Then, the noncentrality parameter λ_{PE} can be expressed as

$$\lambda_{PE} \approx (S - S^{A})^{T}[W_{0}^{-1} - W_{0}^{-1}(W_{0}^{-1} + 2n_{A}R^{-1})^{-1}W_{0}^{-1}](S - S^{A})\ . \tag{C3}$$

Substituting $S - S^{A} \approx U$ from equation (C1) into the above equation (C3) yields

$$\begin{aligned}\lambda_{PE} &\approx U^{T}[W_{0}^{-1} - W_{0}^{-1}(W_{0}^{-1} + 2n_{A}R^{-1})^{-1}W_{0}^{-1}]U \\ &= U^{T}W_{0}^{-1}U - U^{T}W_{0}^{-1}(W_{0}^{-1} + 2n_{A}R^{-1})^{-1}W_{0}^{-1}U\ .\end{aligned} \tag{C4}$$

But,

$$\begin{aligned}U^{T}W_{0}^{-1}U &= U^{T}(B^{-1})^{T}\Sigma_{0}^{-1}B^{-1}U \\ &= (B^{-1}U)^{T}\Sigma_{0}^{-1}B^{-1}U \\ &= (b\delta_{1} + b^{2}\delta_{2})^{T}\Sigma_{0}^{-1}(b\delta_{1} + b^{2}\delta_{2})\ ,\end{aligned} \tag{C5}$$

where $\delta_{1} = [\delta_{H_{1}D} \ldots \delta_{H_{m}D}]^{T}$ and

$$\delta_{2} = \left[\frac{\delta_{H_{1}D}^{2}}{2(1 + \log P_{H_{1}})P_{H_{1}}} \cdots \frac{\delta_{H_{m}D}^{2}}{2(1 + \log P_{H_{m}})P_{H_{m}}}\right]^{T}\ .$$

Substituting equation (C5) into equation (C4) yields $\lambda_{PE} \approx b^{2}\delta_{1}^{T}\Sigma_{0}^{-1}\delta_{1} + R_{PE}$, where

$$R_{PE} = 2b^{3}\delta_{1}^{T}\Sigma_{0}^{-1}\delta_{2} + b^{4}\delta_{2}^{T}\Sigma_{0}^{-1}\delta_{2} - (b\delta_{1} + b^{2}\delta_{2})^{T}\Sigma_{0}^{-1}B^{-1}(W_{0}^{-1} + 2n_{A}R^{-1})^{-1}B^{-1}\Sigma_{0}^{-1}(b\delta_{1} + b^{2}\delta_{2})\ .$$

Then, we derive the noncentrality parameter λ_{T} of the standard χ^{2} test statistic T. From equation (C5), we know

$$\lambda_{T} = (P - P^{A})^{T}\left(\frac{\Sigma}{2n_{G}} + \frac{\Sigma^{A}}{2n_{A}}\right)^{-1}(P - P^{A}) = (P - P^{A})^{T}\Sigma_{0}^{-1}(P - P^{A})\ .$$

But, from equation (B5), we have $P^{A} - P = b\delta_{1}$. Therefore, $\lambda_{T} = b^{2}\delta_{1}^{T}\Sigma_{0}^{-1}\delta_{1}$.

References

Akey J, Jin L, Xiong M (2001) Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet 9:291–300

Anderson TW (1984) An introduction to multivariate statistical analysis. John Wiley & Sons, New York

Bourgain C, Genin E, Margaritte-Jeannin P, Clerget-Darpoux F (2001) Maximum identity length contrast: a powerful method for susceptibility gene detection in isolated populations. Genet Epidemiol Suppl 21:S560–S564

Bourgain C, Genin E, Ober C, Clerget-Darpoux F (2002) Missing data in haplotype analysis: a study on the MILC method. Ann Hum Genet 66:99–108

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

Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. Nature 429:446–452

Chapman NH, Wijsman EM (1998) Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. Am J Hum Genet 63:1872–1885

Collins FS (2004) The case for a US prospective cohort study of genes and environment. Nature 429:475–477

de Vries HG, van der Meulen MA, Rozen R, Halley DJ, Scheffer H, ten Kate LP, Buys CH, te Meerman GJ (1996) Haplotype identity between individuals who share a CFTR mutation allele "identical by descent": demonstration of the usefulness of the haplotype-sharing concept for gene mapping in real populations. Hum Genet 98:304–309

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D (2004) Assessing the impact of population stratification on genetic association studies. Nat Genet 36:388–393

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229

Goldstein DB (2001) Islands of linkage disequilibrium. Nat Genet 29:109–111

Hampe J, Schreiber S, Krawczak M (2003) Entropy-based SNP selection for genetic association studies. Hum Genet 114: 36–43

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA (2001) Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237

Ke X, Cardon LR (2003) Efficient selective screening of haplotype tag SNPs. Bioinformatics 19:287–288

Lehmann EL (1983) Theory of point estimation. John Wiley & Sons, New York

Morton NE, Collins A (1998) Tests and estimates of allelic association in complex inheritance. PNAS 95:11389–11393

Neale BM, Sham PC (2004) The future of association studies: gene-based analysis and replication. Am J Hum Genet 75: 353–362

Nothnagel M (2002) Simulation of LD block-structured SNP haplotype data and its use for the analysis of case-control data by supervised learning methods. Am J Hum Genet Suppl 71:A2363

Nothnagel M, Furst R, Rohde K (2002) Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. Hum Hered 54:186–198

Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69:1–14

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Ross SM (1997) Introduction to probability models. Academic Press, New York

Shannon CE (1948) A mathematical theory of communication. Bell Systems Tech J 27:379–423

Shifman S, Bronstein M, Sternfeld M, Pisante-Shalom A, Lev-Lehman E, Weizman A, Reznik I, Spivak B, Grisaru N, Karp L, Schiffer R, Kotler M, Strous RD, Swartz-Vanetik M, Knobler HY, Shinar E, Beckmann JS, Yakir B, Risch N, Zak NB, Darvasi A (2002) A highly significant association between a COMT haplotype and schizophrenia. Am J Hum Genet 71:1296–1302

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, et al (2001) Haplotype variation and linkage disequilibrium in 313 human genes. Science 293:489–493

Stumpf MP, Goldstein DB (2003) Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. Curr Biol 13:1–8

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

Wall JD, Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet 4:587–597

Xiong M, Zhao J, Boerwinkle E (2003) Haplotype block linkage disequilibrium mapping. Front Biosci 8:a85–a93

Yu C, Zhou Y, Miao X, Xiong P, Tan W, Lin D (2004*a*) Functional haplotypes in the promoter of matrix metalloproteinase-2 predict risk of the occurrence and metastasis of esophageal cancer. Cancer Res 64:7622–7628

Yu K, Gu CC, Province M, Xiong CJ, Rao DC (2004*b*) Genetic association mapping under founder heterogeneity via weighted haplotype similarity analysis in candidate genes. Genet Epidemiol 27:182–191

Zhang K, Calabrese P, Nordborg M, Sun F (2002) Haplotype block structure and its applications to association studies: power and study designs. Am J Hum Genet 71:1386–1394

Zhang K, Sun F, Waterman MS, Chen T (2003*a*) Haplotype block partition with limited resources and applications to human chromosome 21 haplotype data. Am J Hum Genet 73:63–73

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