# Gene-centric Genomewide Association Study via Entropy

Yuehua Cui[1,*], Guolian Kang[1], Kelian Sun[2], Minping Qian[2], Roberto Romero[3], and Wenjiang Fu[2]

[1]*Department of Statistics and Probability,* [2]*Department of Epidemiology, Michigan State University, East Lansing, MI 48824*

[3]*The Perinatology Research Branch, NICHD, NIH, 48201*

**Running head:** Gene-centric genomewide association study

**Key words:** Association study, Complex disease, Entropy, Gene-centric, Genomewide

**Corresponding Author**:

Yuehua Cui, Ph.D.

A432 Wells Hall

Department of Statistics and Probability

Michigan State University

East Lansing, MI 48824

Tel: 517-432-7098

Fax: 517-432-1405

E-mail: cui@stt.msu.edu

## ABSTRACT

Genes are the functional units in most organisms. Compared to genetic variants located outside genes, genic variants are more likely to affect disease risk. The development of the human HapMap project provides an unprecedented opportunity for genetic association studies at the genomewide level for elucidating disease etiology. Currently, most association studies at the single-nucleotide-polymorphism (SNP) or the haplotype level rely on the linkage information between SNP markers and disease variants, with which association findings are difficult to replicate. Moreover, variants in genes might not be sufficiently covered by currently available methods. In this article, we present a gene-centric approach via entropy statistics for genomewide association study to identify disease genes. The new entropy-based approach considers genic variants within one gene simultaneously and is developed based on a joint genotype distribution among genetic variants for an association test. A grouping algorithm based on a penalized entropy measure is proposed to reduce the dimension of the test statistic. Type I error rates and power of the entropy test are evaluated through extensive simulation studies. The results indicate that the entropy test has stable power under different disease models with reasonable sample size. Compared to single SNP-based analysis, the gene-centric approach has greater power, especially when there is more than one disease variant in a gene. As the genomewide genic SNPs become available, our entropy-based gene-centric approach would provide a robust and computationally efficient way for gene-based genomewide association study.

## INTRODUCTION

The family-based linkage study has been the traditional means of disease gene discovery followed by a variety of fine mapping techniques. For finer resolution, larger pedigrees are required, which largely restricts its utility, especially for identifying multiple low-penetrance variants involved in common diseases (Boehnke 1994). In the past decade, population-based association mapping, as an alternative for disease gene discovery, has been rapidly developed either at the single variant or candidate gene level. Risch and Merkangas (1996) first showed that an association study has comparatively greater power than the linkage analysis in detecting disease variants with minor or modest effect size. Therefore genomewide association studies (GWAS) are feasible. With the development of recent high throughput genotyping technologies, it is now possible to conduct disease gene search with millions of SNP markers covering the whole human genome (The International HapMap Consortium 2005). This rapid escalation in disease gene search from family-based linkage scan to population-based association study has greatly facilitated the process of disease gene discovery.

The analysis of association has been historically focused on alleles and the association has been primarily referred to as allelic association. With the high density SNP marker generated by human HapMap and the availability of empirical LD information across the genome, haplotype-based association study is gaining popularity. However, both type of association studies at the SNP or haplotype level have potential pitfalls in the context of replication (Neale and Sham 2004). Due to population histories and evolutionary forces, there has been inconsistency between different studies caused by aberrant LD patterns across marker loci, different allele frequencies and LD patterns across populations (Morton and Collins 1998; Pritchard and Przeworsk 2001; Stephens et al. 2001; Freedman et al. 2004). There also has been controversy concerning statistical analysis and interpretation of association findings (Neale and Sham 2004). In a recent comprehensive review, Neale and Sham (2004) pointed out common problems associated with the single SNP and the haplotype-based analysis, and proposed the prospects of gene-based association study. To reduce reproducebility error, a shift to a gene-based analysis is necessary (Neale and Sham 2004).

With reduced genotyping costs, recent GWA studies have produced promising results (Ozaki et al. 2002; Klein et al. 2005; Maraganore et al. 2005; Skol et al. 2006; Yeager et al. 2007; Hunter et al. 2007). With limited prior knowledge of genomic regions harboring disease genes, a genomewide blind search for an association signal appears to be the most powerful approach in disease gene hunting. Most association mapping approaches based on either single SNP or haplotype can be applied in a GWA study with appropriate genomewide multiple testing adjustment. However, in a recent investigation of the HapMap ENCODE data, Jorgenson and Witte (2006) found that variants in genes are not sufficiently covered by existing GWA approach. The authors raised their concerns about the coverage of current GWA study. Some functionally important variants in genes might be missed due to the limitation of the existing SNP genotyping technology. Results in an age-related macular degeneration study also indicated that some genetic risk factors identified through candidate gene-based approach were missed by GWA approach (Conley et al. 2005; Rivera et al. 2005). Apparently, there is a lack of efficient coverage of functional variants by the current GWA approach, which could consequently result in missing association signals. With limited resources, a gene-centric GWA approach by focusing only on those SNPs located within a gene should be preferred as an alternative to current GWA approach (Jorgenson and Witte 2006).

It is well known that genes are the functional units in most organisms. Genic variants are more likely to alter gene function and affect disease risk than those that occur outside genes (Jorgenson and Witte 2006). The sequence information and function of genes are highly consistent across diverse populations, which makes the gene-based association study more robust in terms of replication compared to the single SNP analysis (Neale and Sham 2004). Moreover, multiple testing has been a major statistical issue for the single SNP-based analysis. While haplotype tagging can potentially reduce the dimension of multiple testing, there are potential risks of missing association signals due to the limitation of SNP tagging. In contrast, the multiple-testing problem can be greatly alleviated with a gene-based analysis by conveniently dividing it into two stages, dealing first with the multiple variants within a gene and then with the multiple genes in the genome (Neale and Sham 2004).

Given the feasibility for a gene-centric genomewide association study, there are pressing needs for developing appropriate statistical approaches under this framework. By treating genes as testing units, any approach developed should consider the multilocus LD information simultaneously in a gene to produce sufficient association signals to be detected by a powerful statistical test. Recently, Zhao et al. (2005) showed the power of a nonlinear transformation of haplotype frequencies through an entropy-based haplotype scoring approach for a genomewide association study. This approach can be extended for a gene-centric GWA study. However, since their approach relies on the information of haplotype frequencies, it needs an extra step to estimate haplotype phases and frequencies beyond the association testing step. Given the potentially large numbers of 'genic' SNPs involved, the haplotype-scoring approach is computationally infeasible, especially for a large scale genomewide study.

Relaxing the assumption of phase information for a haplotyping scoring approach, in this article, we propose an alternative approach by directly utilizing joint genotype information of 'genic' SNPs in a gene. The joint genotypes within one gene could be considered as 'pseudo' alleles. Following the argument of Zhao et al. (2005), we propose a nonlinear transformation of joint genotype frequencies through an entropy measure to amplify the genotype frequencies between cases and controls. The functional information for a gene is then compared between cases and controls through the entropy measure. To reduce the degrees of freedom of the entropy-based test and increase the efficiency and power of an association test, we propose a grouping algorithm to cluster rare joint genotypes to common ones with maximum similarity. An information-based penalized entropy measure is proposed to determine the clustering threshold. Extensive simulation studies are conducted to demonstrate the performance of this new entropy-based statistic. Type I error rate and power of the entropy-based nonlinear test are compared with the single SNP analysis considering different data generation schemes. The entropy test is then extended to the genomewide level and the performance of the gene-centric GWA approach is compared with the single SNP-based GWA study. Both real data and simulations confirm the powerfulness of the proposed approach.

## METHODS

**Entropy-based statistic:** In information theory, entropy measures the uncertainty of a random variable (Cover and Thomas 1991). One commonly used entropy measure is Shannon entropy (Shannon 1948). It can be used to measure the genetic diversity in DNA variation and has been applied to extract the maximal amount of information for a set of genetic loci (Hampe et al. 2003). Therefore, Shannon entropy provides a natural way to quantify relevant information gain or loss for a set of SNP markers within a gene related to certain diseases. The entropy of a discrete random variable or a system $X$ is defined in general as:

$$S(X) = -\sum_t p(x_t)\log p(x_t),$$

where $p(x_t) = \text{Prob}(X = x_t)$, $x_t$ denotes the $t$th individual element in a system. The system X represents a gene with joint genotype $x_t$ as individual element in the following model derivation.

As a functional unit, a gene contains one or more functional variants which jointly function in a coordinated manner to develop a disease phenotype. A gene can thus be defined as a genetic locus-system with jointly acting networks of functional units to maintain human homeostasis. Any systematic deviations from the homeostasis due to the perturbations of the underlying genetic variants could result in potential disease signals representing a change of the underlying locus information. This information change, when measured in genotype frequency in cases and controls, can be amplified and captured by a properly defined entropy measure.

Assume a sample of $n$ unrelated individuals collected from a population with $n_1$ affected and $n_2$ unaffected. In this sample, a large number of genes across the genome are selected for genotyping. Within each gene, a number of SNPs are genotyped, which are defined as the 'genic' SNPs. The goal of the study is to identify which genes are associated with the genetic disease of interest in a genomewide scale. For simplicity, we start with one gene. An extension to genomewide is given later. Consider $N_i$ 'genic' SNPs each with two alleles for the $i$th gene. Following the definition of Jorgenson and Witte (2006), 'genic' SNPs include those that are synonymous and non-synonymous coding SNPs and SNPs in 5' and 3' untranslated regions. At each locus, there are three possible genotypes coded as 0, 1, or 2 with 0 and 2 representing homozygote and 1 representing heterozygote. For the $i$th gene with $N_i$ 'genic' SNPs, there are total $3^{N_i}$ possible genotype combinations. In reality, there are fewer joint genotypes at one gene because of LD among SNPs within a gene. We define the combination of multilocus genotypes as a genetic locus-system with each joint genotype as one single element within the system.

We denote the number of observed joint genotypes for the $i$th gene as $m_i$ ($m_i < 3^{N_i}$). Let $c_{ij}$ be the $j$th joint genotype out of $m_i$ observed joint genotypes at the $i$th gene. We deal with each joint genotype as one element within the locus-system and each joint genotype captures the underlying multilocus LD information. Let $p^A_{c_{ij}}$ and $p^U_{c_{ij}}$ be the frequency of the joint genotype $c_{ij}$ in cases and controls, respectively. The entropies of the joint genotype $c_{ij}$ in cases and controls are then defined as $S^A_{c_{ij}} = -p^A_{c_{ij}} \log p^A_{c_{ij}}$ and $S^U_{c_{ij}} = -p^U_{c_{ij}} \log p^U_{c_{ij}}$, respectively. Through this nonlinear transformation of the entropy measure, the joint genotype frequencies between cases and controls are amplified and the frequency difference is easier to detect than that in the original scale (Zhao et al., 2005).

Let $S^A_i = [S^A_{c_{i1}} \quad S^A_{c_{i2}} \quad \cdots \quad S^A_{c_{im_i}}]$ and $S^U_i = [S^U_{c_{i1}} \quad S^U_{c_{i2}} \quad \cdots \quad S^U_{c_{im_i}}]$ be vectors of the entropy measure for the observed joint genotypes in cases and controls, respectively. Assuming independence among individuals, the number of joint genotypes at one gene follows a multinomial distribution. It can be shown that the variance-covariance matrix of joint genotypes in cases and controls are $n^A_i \Sigma^A_i$ and $n^U_i \Sigma^U_i$, where $n^A_i$ and $n^U_i$ are the number of cases and controls for the $i$th gene, respectively, and $\Sigma^{A/U}_i = (\sigma^{A/U}_{i|jk})_{m_i \times m_i}$ with

$$\sigma^{A/U}_{i|jk} = \begin{cases} p^{A/U}_{c_{ij}}(1 - p^{A/U}_{c_{ij}}) & \text{for } j = k \\ -p^{A/U}_{c_{ij}} p^{A/U}_{c_{ik}} & \text{for } j \neq k \end{cases}$$

where $j$ and $k$ ($j, k = 1, \cdots, m_i$) denote the $j$th column and the $k$th row of the covariance matrix $\Sigma^{A/U}_i$.

Denote the first partial derivatives of the joint genotype entropy $S_{c_{ij}}$ with respect to the joint genotype frequency $p_{c_{ij}}$ for the $i$th gene in cases and controls as $D^A_i = (d^A_{i|jk})_{m_i \times m_i}$ and $D^U_i = (d^U_{i|jk})_{m_i \times m_i}$, where

$$d^{A/U}_{i|jk} = \begin{cases} -1 - log(p^{A/U}_{c_{ij}}) & \text{for } j = k \\ 0, & \text{for } j \neq k \end{cases}$$

Then, the entropy-based statistic for an association test between the $i$th gene and the disease, is defined as

$$T^{gene}_i = (\hat{S}_i^A - \hat{S}_i^U)\hat{W}_i^{-1}(\hat{S}_i^A - \hat{S}_i^U)^T, \tag{1}$$

where $\hat{W}_i = \frac{\hat{D}_i^A \hat{\Sigma}_i^A \hat{D}_i^A}{n_i^A} + \frac{\hat{D}_i^U \hat{\Sigma}_i^U \hat{D}_i^U}{n_i^U}$, and the "$\wedge$" sign refers to estimated values.

When the frequencies of joint genotypes are not zeros, it can be shown that $T_i^{gene}$ has an asymptotical central $\chi^2_{m_i-1}$ distribution under the null hypothesis of no association between the $i$th gene and disease (Lehmann 1983; Zhao et al. 2005). When the matrix $\hat{W}_i$ is not invertible, we use the spectral decomposition to get $\hat{W}_i = P\text{diag}\{\lambda_1, \cdots, \lambda_{m_i}\}P^T$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{m_i}$ are eigenvalues of $\hat{W}_i$ and $P$ is an orthogonal matrix. In the singular case, assume the rank of $\hat{W}_i$ is $\ell$ ($\ell < m_i$), then we use the following matrix

$$P\text{diag}(\frac{1}{\lambda_1}, \cdots, \frac{1}{\lambda_\ell}, 0, \cdots, 0)P^T$$

to replace $\hat{W}_i^{-1}$ in Eq. (1). Under the alterative hypothesis that there is an association between the $i$th gene and disease, $T_i^{gene}$ follows a noncentral $\chi^2_{m_i-1}$ distribution with the noncentrality parameter given as

$$(S_i^A - S_i^U)W_i^{-1}(S_i^A - S_i^U)^T$$

When dealing with genotypes, the computation burden for estimating haplotype frequencies for a haplotype-based approach such as the one proposed by Zhao et al. (2005) can be alleviated. However, the large number of observable joint genotypes and hence the large degrees of freedom (*df*) for an association test could affect the power of test. To overcome this difficulty, we propose an information-based grouping algorithm to cluster rare genotypes into common ones.

**Genotype grouping via penalized entropy:** When the number of SNPs is large, the high dimensionality of the parameter space could reduce the power and efficiency of the proposed entropy approach. For a haplotype-based analysis, the same dimensionality problem also exists (Zhao et al. 2005). Dimension reduction is essential before conducting an association test. A number of studies have been focusing on haplotype dimension reduction either through excess haplotype sharing (Yu et al. 2004) or through haplotype grouping (Durrant et al. 2004; Tzeng 2005). To ensure the parameter space ($\mathcal{G}$) containing an appropriate set of joint genotypes, a pre-determined frequency cut-off, termed hard threshold can be set intuitively. Genotypes with frequency less than the threshold will be excluded. The so called truncated method could result in a low dimensional space $\mathcal{G}$ that consists of high-frequency components which can reduce the degrees of freedom and increase the test efficiency. The risk is lose of detailed local information leading to a consequence of power loss in detecting the difference between two distributions (Fan 1996).

To achieve optimal balance between information and dimensionality, instead of discarding rare genotypes by setting a hard threshold, a more objective way is to group rare joint genotypes with the most similar ones by choosing a soft threshold based on data. By soft threshold, we mean a threshold determined by data itself according to certain selection criterion, not based on a pre-determined value. Here we introduce an information criterion based on the Shannon entropy to select the cut-off point with the goal to find a set of common joint genotypes that are parsimonious enough yet containing maximal amount of information for the tested gene. The selected soft threshold is gene dependent and is more objective. The underlying rationale is that genotypes associated with disease should share close evolutionary character states. By assigning rare genotypes to a common cluster or group, one expects that genotypes within a cluster will preserve the original multilocus LD information and contain as much as possible the polymorphisms information and, hence, might have similar effects on disease predisposition. From the evolutionary point of view, common genotypes reflect more of the history, and hence, should be preserved in the new clustered genotype space ($\mathcal{G}$). From the statistics point of view, common genotypes represent nonsparse components which reveals the majority of information of a multinomial distribution (Zhao et al. 2005).

To incorporate both case and control information into the entropy measure, we combine cases and controls with the same joint genotype together before grouping. Consider a set of joint genotypes $\mathcal{G} = \{c_1, c_2, \cdots\}$ for gene $i$. For simplicity, we drop the gene index subscript $i$. Each category $c_j$ contains $n_{c_j}$ joint genotypes with frequency $p_{c_j}$, and $c_j$ represents the joint genotype after combining cases and controls together. The Shannon entropy of each genotype category $c_j$ is defined as $-p_{c_j}\log_2 p_{c_j}$ (Shannon 1948). The base 2 logrithm is for information measurement in binary units (Tzeng 2005). The overall entropy is the sum over all categories, $-\sum_j p_{c_j}\log_2 p_{c_j}$. As the number of joint genotype categories increases, this information measure monotonically increases. Maximum information is achieved when all joint genotype categories are included. However, this does not serve our purpose of dimension reduction. To balance between information and dimension, we introduce a penalty function following Tzeng (2005) to penalize the information increase caused by adding more genotype categories in the parameter space $\mathcal{G}$. The penalized entropy measure (PEM) is given as

$$I(\mathcal{G}_k) = -\Big(\sum_{j=1}^{k} p_{c_j}\log_2 p_{c_j}\Big) - 2\log_2 k/n_{c_k} \tag{2}$$

where $\log_2 k$ is the total number of bits needed to describe the outcome (MacKay 2003). The AIC-like penalty term $2\log_2 k$ is divided by the sample size of $c_k$ to reflect the marginal cost for including the $c_k$ category (Tzeng 2005). The penalty term $\frac{2\log_2 k}{n_{c_k}}$ penalizes the monotone increase of the entropy measure by adding new terms. Thus the one with maximum PEM measure $(I(\mathcal{G}_k))$ corresponds to the optimal parameter space. To do so, we first sort all joint genotypes according to their frequencies. Then for each truncated genotype set $\mathcal{G}_k$, $k = 1, 2, \cdots, \mathcal{G}^*$, we calculate the PEM measure $(I(\mathcal{G}_k))$. The optimal dimension $\mathcal{G}^*$ that maximizes the PEM $(I(\mathcal{G}_k))$ contains $k$ joint genotypes.

Once the grouping threshold is determined, we can proceed to group those with frequency less than the threshold with common ones. To preserve the maximum information for those rare joint genotypes after grouping, we first calculate the similarity (Eq. (3)) between those rare joint genotypes and those with frequencies larger than the grouping threshold. Each rare joint genotype is then grouped together with the common one having the largest similarity. To illustrate the idea, we consider one gene with 10 SNPs. Denote the genotype vector at the 10 SNP loci for one rare joint genotype with frequency less than the grouping threshold as $c^0 = \begin{bmatrix} g_1^0 & g_2^0 & \cdots & g_{10}^0 \end{bmatrix}$, and denote the genotype vector for any one of the joint genotypes with frequency greater than the grouping threshold as $c^j = \begin{bmatrix} g_1^j & g_2^j & \cdots & g_{10}^j \end{bmatrix}$, where $g_s^\ell = 0, 1, 2$, $s = 1, 2, \cdots, 10$, and $\ell \in \{0, j(j \geq 1)\}$. We define a similarity measure between these two joint genotypes as

$$\psi_{(0,j)} = \sum_{s=1}^{10} I(g_s^0 = g_s^j) \tag{3}$$

where $I(g_s^0 = g_s^j) = \begin{cases} 1, & \text{if } g_s^0 = g_s^j; \\ 0, & \text{if } g_s^0 \neq g_s^j \end{cases}$ is an indicator function. For each rare joint genotype to be grouped, a similarity score will be calculated against all common joint genotypes. This rare joint genotype will be grouped with the common one with the largest similarity.

The penalized entropy measure accommodates the increase of the joint genotype dimension, and maintains the parsimony of the grouping parameter space. Fig. 1 illustrates how the information measure PEM changes as the number of joint genotype changes. Fig. 1(A) shows the bar plot of the sorted joint genotype frequencies with a case-control combined sample. The solid horizontal line indicates the grouping threshold. Fig. 1(D) displays the PEM for each truncated distribution $\mathcal{G}_k$, $k = 1, 2, \cdots ,$. We can clearly see a concave curve with the peak indicating that 6 is the optimal number of dimension, i.e., 6 common joint genotypes after grouping. As sample size increases gradually, we see the same concave pattern for the PEM and the grouped dimension gradually increases (data not shown). Fig. 1(B) and 1(C) show the bar plots of the joint genotype distribution before grouping for cases and control, respectively. The corresponding joint genotype distributions after grouping are shown in Fig. 1(E) and 1(F), resepectively. The data are simulated under the null hypothesis of no disease-gene association. We can see that the joint genotype distributions in cases and controls are very similar before and after grouping. The common distribution pattern is preserved in both cases and controls after grouping.

**Gene-centric genomewide significance:** For the single SNP based GWA study, the significance level for multiple testings is adjusted in the order of $10^5$ or higher. Traditional Bonferroni adjustment is too conservative to detect the true effects. In the past decades, a false discovery rate (FDR) controlling approach has been widely used for multiple testing adjustment (e.g., Benjamini and Hochberg 1995). As several studies have reported the problem of replicating certain genetic findings (reviewed by Neale and Sham 2004), people tend to use more stringent criteria to avoid false discoveries. This, however, could potentially result in false negatives too. The trade-off between false discovery and false negative has not been well established in a GWA study.

While the single SNP based analysis has potential multiple testing problem, a gene-based approach can greatly alleviate the multiple testing burden by dealing with multiple variants within a gene as a testing unit and then dealing with multiple genes across the genome (Neale and Sham 2004). We know that the total number of genes in whole human genome is estimated at $\sim 30,000$ and these genes are likely to represent largely functional units. By focusing on each gene as a unit, we can potentially reduce the false positives and false negatives without suffering too much from power loss due to stringent criteria. Given the large number of tested genes, the Bonferroni or FDR adjustment procedure could still be too stringent to test moderate effects. Actually, for most common diseases, even though a search for disease factors is running with thousands of genes, we know that only a small proportion of genes will trigger true effects. The large number of association tests therefore confers a considerable risk of false discoveries. In reality, if the appropriate genotyping approach is used, scientists are more interested in detecting the proportion of true effects than controlling the false discovery rate. Incorporating certain prior knowledge about the proportion of true effects, therefore, would certainly leads to a more efficient error control. From a Bayesian point of view, this prior information can be updated in a multiple testing adjustment approach. Van den Oord and Sullivan (2003) proposed an optimizing approach by eliminating and controlling false discoveries to achieve a balance of true- and false positives. Following van den Oord and Sullivan (2003), for a gene to be significant at the genomewide significant level, we have

$$\begin{aligned} \alpha_{gene} &= P_0(\text{a gene is significant}) \\ &= P_0(T^{gene} > C^{gene}) \\ &= \frac{(1-p_0)PTD}{\frac{p_0}{FDR} - p_0} \end{aligned} \tag{4}$$

where PTD is the expected proportion of genes with true effects to be detected, i.e., the average power of an association test; $p_0$ is the expected proportion of genes with no effect; and FDR is the false discovery rate. This approach can be applied for either a dependent or independent test. A prior knowledge of 0.5% true effective genes and an average 85% association test power would lead to $\alpha_{gene} \approx 2.2 \times 10^{-4}$ with an FDR of 0.05. Therefore, a genewide significance of $\sim 2.2 \times 10^{-4}$ would be considered to be genomewide significant.

The power of the gene-centric GWA approach can be calculated accordingly. If we know the genomewide significant threshold value $C^{gene}$, then the probability of selecting a disease gene under the alternative hypothesis is given by

$$\text{Power} = P_A(T^{gene} > C^{gene})$$

where $C^{gene}$ value can be found through solving Equation (4).

## SIMULATION STUDY

**Simulation design:** Monte Carlo simulations are performed to evaluate the statistical property of the proposed approach. We use two methods to simulate the genotype data. The first method is based on the MS program developed by Hudson (2002) which simulates SNP haplotypes under a coalescent model where the recombination rate varies across the SNP sequence. This method cannot control the pairwise LD pattern across multiple SNPs within a gene. The second method generates SNP genotype data based on a conditional distribution of current gentoype data given on the distribution of previous genotype data. Thus the pairwise LD pattern can be easily controlled. We call the 2nd method LD-based simulation. The details for the two methods are given as follows.

*MS program.* We use the MS program developed by Hudson (2002) to simulate haplotypes for each individual to form individual genotype data. The main parameters under the coalescent model for generating haplotypes are set as follows: 1) the effective diploid population size $n_e$ is $1 \times 10^4$; 2) the scaled recombination rate for the whole region of interest, $\rho = 4n_e\gamma/bp$, is $4 \times 10^{-3}$, where the parameter $\gamma$ is the probability of cross-over per generation between the ends of the haplotype locus being simulated; 3) the scaled mutation rate for the simulated haplotype region, $4n_e\mu/bp$, is set to be $5.6 \times 10^{-4}$ for the region of simulated haplotypes; and 4) the length of sequence within the region of simulated haplotypes, $n$ sites, is 10kb. Similar parameter settings can be found in other studies (e.g., Hudson 2002; Tzeng 2005). We set the number of SNP sequences in the simulated sample as 100 for each gene and run the MS program to generate the haplotype sample based on these parameter settings. Then we randomly select a segment of ten adjacent SNPs as a haplotype. Next two haplotypes are randomly drawn from the simulated sample containing 100 10-SNP haplotypes and are paired to form an individual genotype.

*LD-based simulation.* Under this scenario, SNP genotypes are generated according to the conditional distribution of current genotype given on the previous genotype data. Suppose the frequencies of two adjacent SNPs with risk alleles A and B are $p$ and $q$ and the linkage disequilibrium between them is $\delta$. Then the frequencies of four haplotypes are $p_{ab} = (1-p)(1-q)+\delta$, $p_{Ab} = p(1-q)-\delta$, $p_{aB} = (1-p)q-\delta$, $p_{AB} = pq+\delta$. Given the allele frequency for A at locus 1 and assume HWE, the SNP genotype data can be obtained at that locus assuming a multinomial distribution with genotype frequencies of $p^2$, $2p(1-p)$ and $(1-p)^2$ for genotype AA, Aa and aa, respectively. To simulate genotype data for SNP2 conditional on SNP1, we need to derive the conditional distribution $P(\text{SNP2}|\text{SNP1})$. To illustrate the idea, consider the simple case $P(BB|AA)$. By Bayes theorem,

$$P(BB|AA) = \frac{P(AABB)}{P(AA)} = \frac{p_{AB}^2}{p^2} = \frac{(pq+\delta)^2}{p^2}$$

Similarly, we can get the conditional distribution of other gentoypes at SNP2 given on the genotype data at SNP1. A detailed conditional distribution is tabulated in Table 1. The pairwise LD pattern between adjacent SNPs can be easily controlled with this simulation method.

*Phenotype simulation.* Given the nature of the difficulty to know the true functional mechanism for a given gene, it is difficult to simulate the true functional variants and the true functional mechanism within a gene. Thus, we evaluate the performance of the entropy approach by considering different scenarios to mimic the situation of a complex disease for a given gene. Three situations are considered in the simulation study. The first scenario considers only one functional disease variant within a gene. The second one considers two disease variants where there are interactions between these two variants. The third one considers three disease variants and there are complex interactions among these three variants.

The simulated data are then subject to the entropy test and single SNP analysis.

Here we briefly illustrate how the disease phenotypes are simulated. Consider the first scenario in which there is only one disease variant. Let $f_i$ be the penetrance function, which is defined as the probability of being affected given possessing $i$ copies of disease alleles ($i = 0, 1, 2$). Denote $\lambda = f_1/f_0$ as the genotype relative risk (GRR). Let $p$ be the disease allele frequency. Then the penetrance function for the three genotypes at a single locus, $f_0, f_1, f_2$, can be defined for an additive and a multiplicative model (Table 2). Once $f_i$ is determined, the disease status for that particular individual is simulated according to a bernoulli distribution with the probability of success $f_i$. The same process is repeated until $n$ cases and $n$ controls are generated.

For a disease model with two interacting disease loci, we follow the settings given in Marchini et al. (2005). Two models are considered (Table 3). In model 1, the odds increase multiplicatively with genotype both within and between loci. In model 2, the odds of a disease have a baseline value ($\gamma$) and the odds increase multiplicatively both within and between genotypes once there is at least one disease allele at each disease locus.

For the three-locus interaction disease model, we denote the three-locus genotypes as $(G_A, G_B, G_C) \in 0, 1, 2$ which represents the number of risk alleles at each disease locus A, B and C. Following Table 3, we generalize the two-locus disease model to the three-locus interaction disease model as

$$\begin{aligned} \text{Model 1}: \quad Odds(G_A, G_B, G_C) = &\quad \gamma(1+\theta)^{G_A+G_B+G_C} \qquad (5) \\ \text{Model 2}: \quad Odds(G_A, G_B, G_C) = &\quad \gamma(1+\theta)^{G_A I(G_A>0)+G_B I(G_B>0)+G_C I(G_C>0)} \end{aligned}$$

Once the disease loci are determined, the case-control disease status can then be simulated according to a multinomial distribution conditional on the observed genotype data.

**Property of the new statistic:**

*Null distribution.* Under the assumption of large sample size under the null hypothesis, the entropy-based statistic $T^{gene}$ has an asymptotic $\chi^2$ distribution. To examine whether the asymptotic results of the entropy-based statistic $T^{gene}$ still hold for a small sample size under the null hypothesis of no association, we generated 200 cases and 200 controls for one gene containing 10 SNPs with 10 joint genotypes. Figure 2 plots the histogram of the test statistic $T^{gene}$. The distribution of $T^{gene}$ is very similar to a $\chi^2$ distribution with 9 df.

*Type I error.* We evaluate the type I error rate for the new entropy statistics based on the two genotype generating schemes, the MS program and the LD-based simulation. Data are simulated under the null of no association with 10 SNPs for a gene. 1000 simulation runs are conducted under different sample sizes. The pattern of LD varies among SNPs when data are simulated using the MS program. To evaluate extreme case, we use the LD-based simulation method to generate genotype data assuming high LD among SNPs ($r^2 = 0.9$). The simulated data are subject to both the entropy test and single SNP analysis to compare their performance. The single SNP analysis is based on genotype $\chi^2$ test with 2 df. Since our attention is focused on genes, for the single SNP analysis, a gene is declared to be significant if it contains at least one significant SNP. The significant level for each SNP within a gene is adjusted first with the Bonferroni correction. Results are summarized in Table 4. As we can see, the error rates are reasonably estimated for data simulated with both approaches. The disease allele frequency has no remarkable effect on type I error. Compared to the entropy test, the single SNP analysis underestimates the type I error in all scenarios. Therefore, when focused on a gene-based analysis, our approach has better error control than the single SNP analysis.

*Power analysis.* Up to now the functional mechanism of most disease related genes has not been comprehensively understood. For a gene to be associated with a disease phenotype, there must be complex functional mechanism among multiple variants within the gene. The physical location and alignment pattern of SNPs within a gene, the up-down stream signaling among SNPs, and the complex functional interactions among SNPs could result in potential functional variation and eventually lead to a disease signal. It is, therefore, difficulty to simulate the true functional mechanism of a disease gene. To compare with the single SNP analysis, we simulate a disease gene by assuming one, two or three disease loci within the gene. These scenarios represent approximations to true functional mechanism of a disease gene.

The one-locus, two-locus and three-locus disease models are given in Tables 2, 3 and Equation (5), respectively. The simulation results for these three models are summarized in Tables 5, 6, and 7, respectively. We assume 10 SNPs on average for a simulated gene and use two simulation schemes, MS program and LD-based simulation. Table 5 shows the power results under the additive and multiplicative disease model assuming one disease locus within a gene. The power for the single SNP analysis is given in the parenthesis. As can be seen, the power of association test based on statistic $T^{gene}$ increases as sample size increases. For data simulated with the MS program, we observe a slightly smaller power of the entropy test than that of the single SNP analysis. As sample size increases from 200 to 400, the difference vanishes and the two approaches are comparable. For data simulated with the LD-based approach assuming $r^2 = 0.9$, we observe consistently higher power of the entropy test than that of the single SNP analysis. Also, the power is affected by the genotype relative risk. Different allele frequency has no remarkable effect on testing power and hence is omitted.

When the number of disease loci increases within a tested gene, we observe significant differences of the two approaches under the two data simulation schemes. For the two-locus disease model, Table 6 shows that the power increases as the GE and sample size increase. As we expected, Model 1 has higher power than Model 2 due to the difference of the two models (Table 3). Noted that the entropy-based test uniformly outperforms the single SNP analysis in all the simulation scenarios. Depending on the underlying gene action mode and linkage information among SNPs, large sample size ($> 400$, say) is always preferred.

The results for the three-locus disease model are very similar to the two-locus model under different sample sizes and different GEs (Table 7). It is worthy to note that when the number of disease loci increases from 2 to 3, the power of the entropy test has a remarkable increase. For example, when data are simulated assuming Model 2 using the LD-based simulation approach, the power of the two-locus model (Table 3) is only 0.621. We observe a dramatic power increases from 0.621 to 0.960 when the number of functional disease loci increase to 3 (Eq. (5)) for fixed GE and sample size. A similar trend holds in general as the number of functional variants increase in a gene.

In summary, under different simulation scenarios, the results confirm that the entropy-bases test outperforms the single SNP analysis in general when the unit of interest is focused on a gene. The results also indicate the we may need large sample size ($> 400$, say) to obtain reasonable power to detect the disease-gene association. As the number of functional variants within a gene increases (i.e., the functional mechanism of a gene becomes more complex), the power of the entropy test to detect the association also increases.

**Power of genomewide association studies:** To check the performance of the entropy based test on a genomewide scale analysis, we simulate 1000 genes. Even though 1000 genes do not represent a genomewide level ($\sim 30,000$ genes), extension to a genomewide scale is straightforward. For a gene to be significant at the genomewide scale, a genewide significant level is caculated based on Equation (4). Assume there are 10 genes to be contributing to a disease phenotype in 1000 genes, i.e., $p_0 = 0.99$ and each gene contains 10 SNPs on average. If we further assume a minimum 80% power to detect an association ($PTD = .8$), and an FDR of 0.05, then $\alpha_{gene} \approx 0.000425$. This means that a gene significant at the $4.25 \times 10^{-4}$ level would be considered to be genomewidely significant.

We randomly simulate 800 cases and 800 controls using the MS program. Figs. 3, 4 and 5 plot the genomewide power analysis of the entropy test for one-locus, two-locus and three-locus disease models, respectively. In all three figures, the solid and the dashed lines represent the power curves for the entropy-based analysis and the single SNP analysis, respectively. The power as a function of the GRR for the one-locus disease model is plotted in Fig. 3A (additive model) and B (multiplicative model). We can see that the single SNP analysis slightly dominates the entropy test with moderate GRR. As GRR increases to 1.7, both tests converge to 100% power. The power as a function of GE is plotted in Figs. 4 and 5 for the two-locus and three-locus disease models, respectively. The entropy test has consistently higher genomewide power than the single SNP analysis under two gene action modes with moderate GE. The power difference decreases as the GE increases.

Note that the power for the single SNP test is adjusted within a gene using the Bonferroni correction, then is corrected in the genewide significant level calculated in Eq. (4) as the familywise error rate for that gene. Therefore, in the single SNP analysis, a gene is declared to be significant if the smallest p-value of any single SNP test is less than $4.25 \times 10^{-4}/10$. Compared to a SNP-wide significant level of $5 \times 10^{-6}$ if we apply the genomewide Bonferroni correction for all SNPs, this threshold is less restrictive. If we rather use the restrictive genomewide Bonferroni correction, the power for the single SNP based analysis would be much smaller compared to the gene-centric approach.

## A CASE STUDY

Until now, a complete set of genomewide genic SNPs has not been fully developed (Jorgenson and Witte, 2006). Hence no real data are available to test the method. To show the utility of the proposed approach, we rather apply our method to a large scale candidate gene study. The data set contains 190 candidate genes in a genetic association study of Pre-Eclampsia (PE). PE typically occurs after 20 weeks gestation with a syndrome of hypertension and proteinuria. The disorder is a leading cause of maternal mortality, and affects at least 5-7% of all pregnancies (Kaunitz et al., 1985). Previous study has shown that genetic factors may contribute greater than 50% of the variability in liability to PE (Esplin et al., 2001). In this study, subjects were recruited at the Sotero del Rio Hospital, in Puente Alto, Chile. Eligible mothers were enrolled either in a longitudinal cohort study designed to predict the subsequent development of PE, or the existence of PE at the time of admission to the hospital. For a detailed data description, please refer to Goddard et al. (2007). After elimination of SNPs with MAF less than 5%, 819 SNPs were subject to both the single SNP and gene-centric association study. The data set contains a total of 225 cases and 585 controls.

If we assume a prior knowledge of 19 significant genes that would be associated with PE based on previous studies (Goddard et al. 2007), the parameter $p_0$ takes a value of 19/190. If we further assume the power for a single gene test is $PTD = 85\%$, following Equation (4) with a false discovery rate of FDR= 0.05, a gene is significant at the 0.005 level would be claimed to be significant at the "genomewide" level. Similarly, we can apply the same significance criterion to the single SNP-based analysis. Based on the report in Goddard et al. (2007), we assume a prior number of 21 significant SNPs out of total 819 SNPs and a power of 0.85 for any single SNP test. Then a SNP is declared significant at the "genomewide" level if the p-value for that single test is less than 0.0012. Alteratively, if we focus on a gene-based unit, we can use the genewide significant level 0.005 as the familywise significant level for one gene and apply the Bonferroni correction by dividing it by the number of SNPs in that gene to get the significant threshold for that particular gene. For example, if a gene contains 4 SNPs, then the gene is declared to be significant if the smallest p-value for any single SNP test is less than $0.005/4 = 1.25 \times 10^{-3}$. In this case, the genewide significant level would be different for different genes.

Table 8 lists the p-values for both significant genes and SNPs. The significant genes and SNPs are shown in bold faces. For comparison, we also listed the smallest p-values of SNPs in which the corresponding gene is significant based on the gene-based test as well as the p-value for gene-based analysis if that gene contains at least one significant SNP. The test results show that the gene-based analysis detected 7 significant genes out of 190 genes. The single SNP based test detected 3 significant SNPs, and hence 3 significant genes, even though the significant level is much less stringent than the Bonferroni correction or FDR based adjustment. Among the significant genes identified, only gene F13B is consistently picked up by both the gene-based and SNP-based tests. Two polymorphisms in genes F2 and FGF4 also show marginal significant. The fact that those genes detected by the gene-based entropy test but missed by the single SNP analysis indicates that there might exist complex functional interactions among SNPs in those genes. It is not surprise that the complex systematic interactions among SNPs in those genes may not be detected by less powerful single SNP analysis by testing each SNP separately.

# DISCUSSION

The development of human HapMap provides an unprecedented opportunity for unravelling the disease etiology of complex human diseases with appropriate statistical analysis in a genomewide scale. The advancement of whole genome genotyping technology and relatively reduced genotyping cost in recent years have made it possible to detect genomewide disease variants in thousands of individuals, with an aim to find variants with modest contributions to disease risk. Traditional single SNP-based analysis approaches find regions associated with a disease based on differences in allele or genotype frequencies of the SNPs in regions between cases and controls. The methods find variants in LD with SNPs and are not comprehensive in the characterization of variation in the regions. In contrast to the SNP-based approaches, the gene-based analysis is less susceptible to erroneous findings due to genetic differences between populations (Neale and Sham 2004). Also a gene-centric approach should be more complete with regard to the coverage of genes, which is crucial to detecting causal variants (Jorgenson and Witte 2006). However, a genomewide genic SNP set has not been well established (Jorgenson and Witte 2006). Given the radical breakthrough in recent biotechnology, it will be made available to public shortly. Therefore, the proposed gene-centric GWA approach in this paper represents a timely effort to contribute to the identification of disease genes in a genomewide scale.

Increasing evidences have shown that complex interaction of genes worked in a coordinated manner is essential for normal function of any organism (Hartwell 2004; Gibson 1996; Moore 2003; Szathmary et al. 2001). The jointly acting network of functional variants, termed the genetic locus-system, is crucial to maintain the system homeostasis. Any systematic deviations from the homeostasis could result in a potential disease signal. The overall disease signal resulted from multiple interacting disease variants for a particular gene can be more efficiently amplified and captured by considering all variants in a gene simultaneously by the proposed entropy test statistic. Compared to the single SNP analysis by testing each variant separately and hence tending to ignore the complex interaction mechanism among

SNPs within a gene, the gene-centric approach is statistically more powerful and biologically more relevant. Moreover, the entropy approach considers the genic variants within a gene as a testing unit, which is in align with the conception that the gene is the functional unit of most organisms. By capturing the difference of the joint genotype distributions between cases and controls through nonlinear transformation of joint genotype frequencies, the entropy-based test shows appropriate type I error and good power in detecting disease associated genes. Simulations also confirm the powerfulness of the gene-based test, especially when the underlying testing gene is functionally complex, for instance, containing two or more interacting disease variants.

The proposed entropy-based test considers the joint genotype distribution and is simple and easy to implement. The nonlinear transformation of the joint genotype frequencies amplifies the difference between cases and controls and hence increase the testing power. Unlike the haplotype-based entropy test (Zhao et al. 2005), the calculation of joint genotype frequency is much fast and simple without estimating linkage phases as the haplotype-based approach does. The computation gain is even more significant when the number of genic SNPs is large. Intensive computation in estimating haplotype frequencies makes the haplotype-based approach practically infeasible to implement in a large genomewide scale.

Given the large number of possible joint genotypes compared to possible haplotypes, one of the disadvantages of genotype-focused analysis is the large number of *df*. One possible solution is to discard those rare joint genotypes with frequency less than a pre-specified hard threshold. This, however, will largely reduce the sample size and hence the power of the association test. Without sacrificing sample size, the problem of large *df* is overcome by an efficient grouping algorithm. To balance the trade-off between informativeness and dimensionality, we have proposed a penalized entropy measure to determine a soft threshold. Rare joint genotypes are then grouped with common ones to reconstruct a lower dimensional distribution from the original distribution. Simulations confirm the validity of the grouping approach.

The powerfulness of the proposed gene-centric approach in a genomewide scale is also confirmed by simulations (Figs. 3, 4, 5). As revealed by the genomewide simulation studies, the gene-based approach has more power than the single SNP test when there are more than one functional disease variants in a gene (Figs. 4 and 5). The proposed approach is not only restricted to a genomewide application. It also fits to a candidate gene study with appropriated defined SNP set for a candidate gene. The real example shown in the paper indicates that the gene-centric approach detects more genes than the single SNP analysis, and hence is more powerful than the single SNP based analysis. Some genes identified by the gene-based approach but missed by the single SNP analysis have been shown in the literature. For example, genes APOB, FGF4 and F2 are detected by the gene-based analysis but are missed by the single SNP test. These genes were previously reported to be associated with PE (Sarandöl et al. 2004; Anteby et al. 2004; Kosmas et al. 2004). The fact that these genes missed by the single SNP analysis but detected by the gene-based analysis may be due to higher order complex interactions among disease variants. However, when there is only one disease variant in a gene, the single SNP test slightly beats the gene-based approach by simulation (Fig. 3). As little is know about the nature of the true functional mechanism of a gene, our conservative recommendation is to report genes detected by both approaches.

By focusing on each gene as a separate module or unit, a gene-centric association study can be designed using a "direct" or "sequence-based" approach by examining candidate genes with known biological function, or can be studied using an "indirect" or "map-based" approach taking advantage of LD among variants (Jorgenson and Witte 2006). Currently, the direct approach is limited by incomplete knowledge about functional variation of a gene. The indirect approach with reliance on the information of LD among variants is preferred instead. The developed entropy approach consider the joint multilocus LD information for a given gene in an association test and fits this mission well. Also, for SNPs in evolutionarily conserved regions, once they are genotyped, can be define as a functional unit and then perform an entropy-based test to identify possible *cis* enhancers to cover all functionally important variants. Due to reduced genotyping burden with small number of genic SNPs, a gene-centric GWA approach would be a first choice to implement as an efficient initial genomewide association scan for targeting biologically plausible genomic regions with reduced multipe-testing burden.

## Acknowledgement

The authors are grateful to the associate editor Rebecca Doerge and to the two referees for their constructive comments and suggestions that led to substantial improvements of the original manuscript. This work was supported in part by an Intramural Research Program grant from NICHD/NIH, by a Michigan State University IRGP grant, and by NSF grant (DMS 0707031).

# References

Anteby E. Y., C. Greenfield, S. Natanson-Yaron, D. Goldman-Wohl, Y. Hamani, V. Khudyak, I. Ariel, S. Yagel, 2004 Vascular endothelial growth factor, epidermal growth factor and fibroblast growth factor-4 and -10 stimulate trophoblast plasminogen activator system and metalloproteinase-9. Mol. Hum. Reprod. **10**: 229-235.

Benjamini Y. and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. Ser B **57**: 289-300.

Boehnke M., 1994 Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. Am. J. Hum. Genet. **55**: 379-390.

Conley, Y. P. et al. 2005 Candidate gene analysis suggests a role for fatty acid biosynthesis and regulation of the complement system in the etiology of age-related maculopathy. Hum. Mol. Genet. **14**: 1991-2002.

Cover T. M. and J. A. Thomas, 1991 *Elements of information theory.* New York: Wiley. p12-15.

Durrant C., K. T. Zondervan, L. R. Cardon, S. Hunt, P. Deloukas, A. P. Morris, 2004 Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am. J. Hum. Genet. **75**: 35-43.

Esplin M. S., M. B. Fausett, A. Fraser, R. Kerber, G. Mineau, J. Carrillo, M. W. Varner, 2001 Paternal and maternal components of the predisposition to preeclampsia. N. Engl. J. Med. **344**: 867-872.

Fan J. 1996 Test of significance based on wavelet thresholding and Neyman's truncation. J. Am. Stat. Assoc. **91**: 674-688.

Freedman M. L., D. Reich, K. L. Penney, G. J. McDonald, A. A. Mignault, N. Patterson, S. B. Gabriel, E. J. Topol, J. W. Smoller, C. N. Pato, M. T. Pato, T. L. Petryshen, L. N. Kolonel, E. S. Lander, P. Sklar, B. Henderson, J. N. Hirschhorn, D. Altshuler, 2004 Assessing the impact of population stratification on genetic association studies. Nat. Genet. **36**: 388-393.

Gibson G., 1996 Epistasis and pleiotropy as natural properties of transcriptional regulation. Theor. Popul. Biol. **49**: 58-89.

Goddard K. A., G. Tromp, R. Romero, J. M. Olson, Q. Lu, Z. Xu, et al., 2007 Candidate-gene association study of mothers with pre-eclampsia, and their infants, analyzing 775 SNPs in 190 genes. Hum. Hered. **63**: 1-16.

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

Hartwell L., 2004 Genetics: robust interactions. Science **303**: 774-775.

Hudson R. 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18**: 337-338.

Hunter, D.J., P. Kraft, K. B. Jacobs, D. G. Cox, M. Yeager, S. E. Hankinson, et al., 2007 A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat. Genet. **39**: 870-874.

Jorgenson E. and J. S. Witte, 2006 A gene-centric approach to genome-wide association studies. Nat. Rev. Genet. **7**: 885-891.

Kaunitz A. M., J. M. Hughes, D. A. Grimes, J. C. Smith, R. W. Rochat, M. E. Kafrissen, 1985 Causes of maternal mortality in the United States. Obstet. Gynecol. **65**: 605-612.

Klein, R. J. et al. 2005 Complement factor H polymorphism in age-related macular degeneration. Science **308**: 385-389.

Kosmas I. P., A. Tatsioni, J. P. Ioannidis, 2004 Association of C677T polymorphism in the methylenetetrahydrofolate reductase gene with hypertension in pregnancy and pre-eclampsia: a meta-analysis. J. Hypertens. **22**: 1655-1662.

Lehmann, E. L., 1983 *Theory of point estimation.* John Wiley & Sons: New York, p343-344.

MacKay, D.J.C., 2003 *Information theory, inference, and learning algorithms.* Chapter 4, p73-74. London: Cambridge University Press.

Maraganore, D. M. et al., 2005 High-resolution whole-genome association study of Parkinson disease. Am. J. Hum. Genet. **77**: 685-693.

Marchini, J., P. Donnelly and L. R. Cardon, 2005 Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat. Genet. **37**: 413-417.

Moore, J.H., S. M. Williams, 2005 Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. Bioessays **27**: 637-646.

Morton, N. E. and A. Collins, 1998 Tests and estimates of allelic association in complex inheritance. Proc. Natl. Acad. Sci. USA **95**: 11389-11393.

Neale, B. M. and P. C. Sham, 2004 The future of association studies: gene-based analysis and replication. Am. J. Hum. Genet. **75**: 353-362.

Ozaki, K. et al., 2002 Functional SNPs in the lymphotoxin- gene that are associated with susceptibility to myocardial infarction. Nat. Genet. **32**: 650-654.

Pritchard, J.K., M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. **69**: 1-14.

Risch, N. and K. Merikangas, 1996 The future of genetic studies of complex human diseases. Science **273**: 1516-1517.

Rivera, A. et al., 2005 Hypothetical LOC387715 is a second major susceptibility gene for age-related macular degeneration, contributing independently of complement factor H to disease risk. Hum. Mol. Genet. **14**: 3227-3236.

Sarandöl, E, O. Safak, M. Dirican, G. Uncu, 2004 Oxidizability of apolipoprotein B-containing lipoproteins and serum paraoxonase/arylesterase activities in preeclampsia. Clin. Biochem. **37**: 990-996.

Shannon, C.E., 1948 A mathematical theory of communication. Bell Systems Tech. J. **27**: 379-423.

Skol A. D., Scott L. J., Abecasis G. R. & Boehnke M. 2006 Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat. Genet. **38**: 209-213.

Stephens, M., N. J. Smith and P. Donnelly, 2001 A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. **68**: 978-989.

Szathmary, E., F. Jordan, C. Pal, 2001 Can genes explain biological complexity? Science **292**: 1315-1316

The International HapMap Consortium, 2005 The haplotype map of the human genome. Nature **437**: 1299-1320.

Tzeng, J. Y. 2005 Evolutionary-based grouping of haplotypes in association analysis. Genet. Epidemiol. **28**: 220-231.

van den Oord, E. J. and P.E. Sullivan, 2003 False discoveries and models for gene discovery. Trends Genet. **19**: 537-542.

Yu, K., C.C. Gu, M. Province, C.J. Xiong, D.C. Rao, 2004 Genetic association mapping under founder heterogeneity via weighted haplotype similarity analysis in candidate genes. Genet. Epidemiol. **27**: 182-191.

Yeager, M., N. Orr, R.B. Hayes, K.B. Jacobs, P. Kraft, S. Wacholder et al., 2007 Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat. Genet. **39**: 645-649.

Zhao, J. Y., E. Boerwinkle and M. M. Xiong, 2005 An entropy-based statistic for genomewide association studies. Am. J. Hum. Genet. **77**: 27-40.

Table 1: Conditional genotype probability of SNP2 given on SNP1 ($P(\text{SNP2}|\text{SNP1})$)

| | bb | bB | BB |
|---|---|---|---|
| aa | $p_{ab}^2/(1-p)^2$ | $2p_{ab}p_{aB}/(1-p)^2$ | $p_{aB}^2/(1-p)^2$ |
| aA | $p_{ab}p_{Ab}/p(1-p)$ | $\frac{2p_{ab}p_{AB}+2p_{aB}p_{Ab}}{2p(1-p)}$ | $p_{AB}p_{aB}/p(1-p)$ |
| AA | $p_{Ab}^2/p^2$ | $2p_{Ab}p_{AB}/p^2$ | $p_{AB}^2/p^2$ |

Table 2: Single-locus disease model

| | $f_0$ | $f_1$ | $f_2$ |
|---|---|---|---|
| Additive | $\frac{prev}{1-2p+2p\lambda}$ | $\lambda f_0$ | $2\lambda f_0 - f_0$ |
| Multiplicative | $\frac{prev}{(1-p+\lambda p)^2}$ | $\lambda f_0$ | $\lambda^2 f_0$ |

where *prev* is the population prevalence, $\lambda$ is the genotype relative risk.

Table 3: Two-locus interaction disease model

| **Model 1** | BB | Bb | bb |
|---|---|---|---|
| AA | $\gamma(1+\theta)^4$ | $\gamma(1+\theta)^3$ | $\gamma(1+\theta)^2$ |
| Aa | $\gamma(1+\theta)^3$ | $\gamma(1+\theta)^3$ | $\gamma(1+\theta)$ |
| aa | $\gamma(1+\theta)^2$ | $\gamma(1+\theta)$ | $\gamma$ |
| | | | |
| **Model 2** | BB | Bb | bb |
| AA | $\gamma(1+\theta)^4$ | $\gamma(1+\theta)^2$ | $\gamma$ |
| Aa | $\gamma(1+\theta)^2$ | $\gamma(1+\theta)$ | $\gamma$ |
| aa | $\gamma$ | $\gamma$ | $\gamma$ |

where $\gamma$ is the baseline effect; $\theta$ is the genotypic effect.

Table 4: Type I error rates of the association test based on statistic $T^{gene}$ at the 0.05 nominal level. Data are simulated with the MS program and the LD-based approach. The type I error for the single SNP analysis is given in parenthesis.

| $n$ | MS program | LD-based | | |
|---|---|---|---|---|
| | | $P_D = 0.1$ | $P_D = 0.3$ | $P_D = 0.5$ |
| 100 | 0.081(0.025) | 0.046(0.022) | 0.061(0.025) | 0.043(0.030) |
| 200 | 0.073(0.038) | 0.066(0.015) | 0.047(0.025) | 0.046(0.025) |
| 400 | 0.052(0.027) | 0.060(0.030) | 0.046(0.027) | 0.045(0.032) |

$P_D$ refers to disease allele frequency.

Table 5: Power of the association test based on statistic $T^{gene}$ assuming one disease locus in a tested gene under different sample sizes and different simulation schemes. The power for the single SNP analysis is given in parenthesis.

| Disease Model | $(prev$, GRR$)^*$ | MS program | | LD-based ($r^2 = 0.9$) | |
|---|---|---|---|---|---|
| | | $n = 200$ | $n = 400$ | $n = 200$ | $n = 400$ |
| Additive | $(0.1, 1.4)$ | 0.370(0.255) | 0.588(0.607) | 0.438(0.305) | 0.766(0.651) |
| | $(0.1, 1.6)$ | 0.602(0.555) | 0.897(0.902) | 0.681(0.577) | 0.943(0.934) |
| Multiplicative | $(0.1, 1.4)$ | 0.402(0.367) | 0.735(0.752) | 0.521(0.384) | 0.808(0.751) |
| | $(0.1, 1.6)$ | 0.751(0.687) | 0.960(0.953) | 0.812(0.743) | 0.970(0.974) |

* *prev* = population prevalence; GRR= genotype relative risk. For both simulation schemes, we set the third SNP as the disease locus. The disease allele frequency for the LD-based simulation is set as $P_D = 0.3$. The significant level is 0.05.

Table 6: Power of the association test based on statistic $T^{gene}$ assuming two interacting disease loci within one gene under different sample sizes and different simulation schemes. The power for the single SNP analysis is given in parenthesis.

| Disease Model | (BL, GE)* | MS program | | LD-based ($r^2 = 0.9$) | |
|---|---|---|---|---|---|
| | | $n = 200$ | $n = 400$ | $n = 200$ | $n = 400$ |
| Model 1 | $(1, 0.7)$ | 0.379(0.160) | 0.564(0.385) | 0.816(0.709) | 0.994(0.822) |
| | $(1, 0.9)$ | 0.708(0.272) | 0.874(0.595) | 0.909(0.817) | 1.000(0.827) |
| Model 2 | $(1, 0.7)$ | 0.269(0.152) | 0.455(0.387) | 0.468(0.349) | 0.835(0.641) |
| | $(1, 0.9)$ | 0.458(0.232) | 0.658(0.559) | 0.621(0.426) | 0.938(0.733) |

*BL=baseline effect; GE=genotypic effect. For the MS program, we set the 4th and 9th SNPs as the two interacting loci. For the LD-based simulation, we simulate two LD blocks, one containing 6 SNPs and the other containing 4 SNPs, with pairwise $r^2 = 0.9$ for SNPs within each block. Two interacting disease loci are from each block with allele frequency $P_D = 0.3$. The significant level is 0.05. Models 1 and 2 are defined in Table 3.

Table 7: Power of the association test based on statistic $T^{gene}$ assuming three interacting disease loci within one gene under different sample sizes and different simulation schemes. The power for the single SNP analysis is given in parenthesis.

| Disease Model | (BL, GE)* | MS program | | LD-based ($r^2 = 0.9$) | |
|---|---|---|---|---|---|
| | | $n = 200$ | $n = 400$ | $n = 200$ | $n = 400$ |
| Model 1 | $(1, 0.7)$ | 0.465(0.271) | 0.658(0.685) | 0.866(0.782) | 0.992(0.857) |
| | $(1, 0.9)$ | 0.802(0.481) | 0.925(0.918) | 0.956(0.867) | 1.000(0.861) |
| Model 2 | $(1, 0.7)$ | 0.471(0.299) | 0.629(0.689) | 0.858(0.774) | 0.994(0.853) |
| | $(1, 0.9)$ | 0.808(0.487) | 0.927(0.876) | 0.960(0.867) | 1.000(0.876) |

*BL=baseline effect; GE=genotypic effect. For the MS program, we set the 1st, 5th and 9th SNPs as the three interacting loci. For the LD-based simulation, we simulate three LD blocks, one containing 5 SNPs, one containing 3 SNPs and the other one containing 2 SNPs, with pairwise $r^2 = 0.9$ for SNPs within each block. Three interacting disease loci are from each block with allele frequency $P_D = 0.3$. The significant level is 0.05. Models 1 and 2 are defined in Equation (5).

Table 8: Analysis of PE data set based on the gene-based entropy test and the single SNP-based $\chi^2$ test. The number of SNPs in each gene is given in parenthesis.

| Gene<br>(SNP number) | Gene-based<br>p-value | SNP* | SNP-based<br>p-value |
|---|---|---|---|
| APOB(9) | **0.0019** | rs5456814 | *0.0165* |
| F13B(4) | **0.0025** | rs28787657 | **0.0010** |
| F2(7) | **0.0016** | rs28886771 | *0.0021* |
| FGF4(3) | **0.0033** | rs634043464 | *0.0067* |
| IGF1R(7) | *0.3600* | rs40893937 | **0.0006** |
| IGF2R(14) | **0.0004** | rs41410456 | *0.0330* |
| MMP10(8) | **0.0009** | rs634850223 | *0.0280* |
| NOS2A(10) | *0.0258* | rs9678181 | **0.0001** |
| PDGFC(2) | **0.0044** | rs634820282 | *0.0320* |

* Only SNPs with smallest p-values within the corresponding genes are listed. Significant and non-significant p-values are indicated in bold and itemized fonts, respectively.

## Legends to figures

**Figure 1**. Examples of joint genotype distributions of cases and controls within one gene under the null hypothesis of no association. **A**, The bar plots of sorted joint genotype frequencies in both cases and controls. Categories with frequencies less than the horizontal line will be grouped. **B**, Joint genotype distribution in cases. **C**, Joint genotype distribution in controls. **D**, Plot of the penalized entropy measure (PEM) against the threshold. The horizontal line in **A** labels the categories to be retained when the maximal amount of PEM is achieved. **E**, **F**, The grouped joint genotype distributions in cases and controls, respectively. Here, the numbers on the x-axis represent the categories of joint genotypes. Data are generated using the MS program with sample size 200.

**Figure 2**. Null distributions of the test statistic $T^{gene}$ from simulated 200 cases and 200 controls with 10 joint genotypes in a gene. $\chi^2_{(9)}$ indicates a $\chi^2$ distribution with 9 degree of freedom.

**Figure 3**. Power comparison of gene-based and SNP-based genomewide association studies as a function of genotype relative risk (GRR) under two single-locus disease models, additive model (**A**) and multiplicative model (**B**). The risk allele frequencies at both loci are 0.30, numbers of individuals in both cases and controls are 800 genotyped on 1000 genes with a population prevalence of 0.1.

**Figure 4**. Power comparison of gene-based and SNP-based genomewide association studies as a function of genotypic effect (GE) under two two-locus disease models, Model 1 (**A**) and Model 2 (**B**) defined in Table 3. The number of individuals in both cases and controls are 800 genotyped on 1000 genes.

**Figure 5**. Power comparison of gene-based and SNP-based genomewide association studies as a function of genotypic effect (GE) under two three-locus disease models, Model 1 (**A**) and Model 2 (**B**) defined in Equation (5). The number of individuals in both cases and controls are 800 genotyped on 1000 genes.

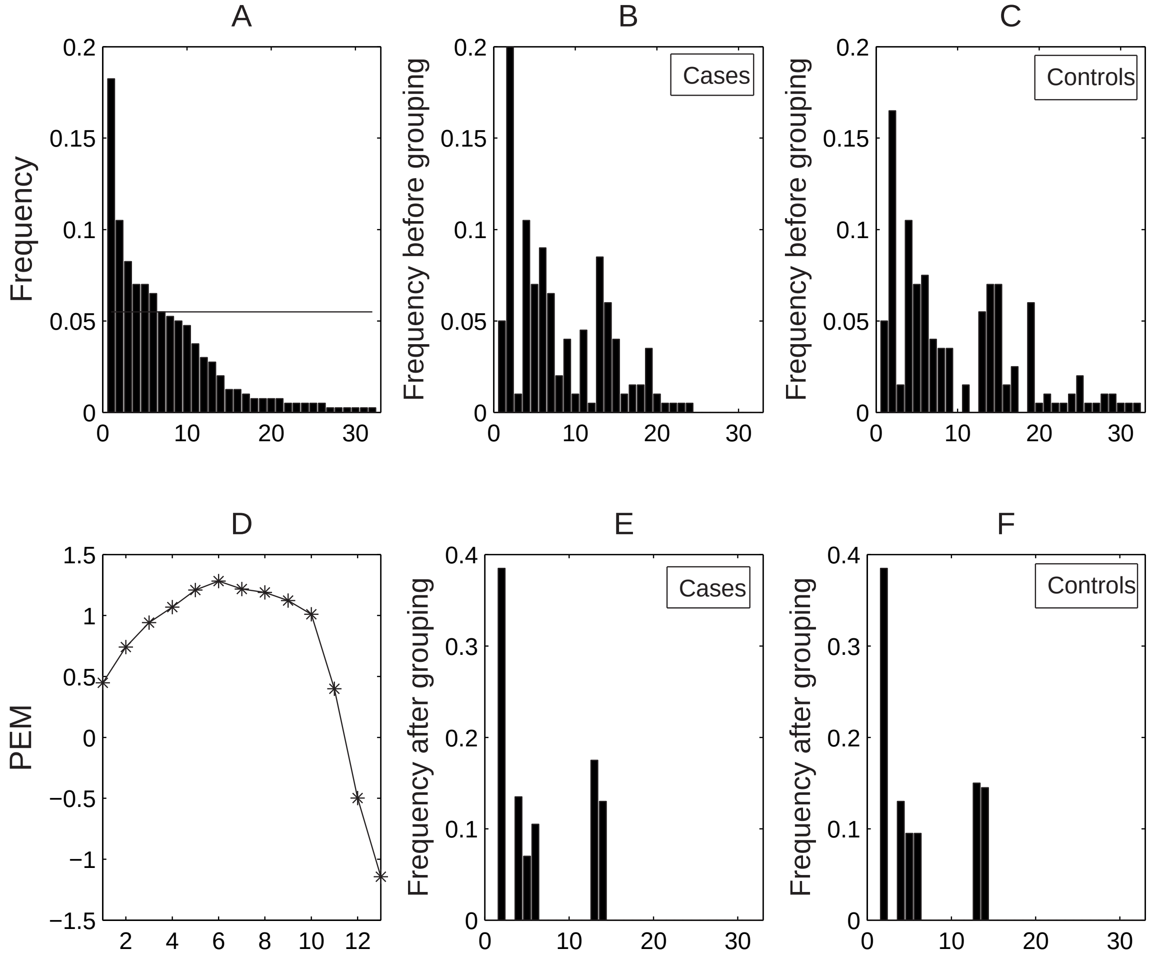
A
Frequency
0.2
0.15
0.1
0.05
0
0
10
20
30
B
Frequency before grouping
Cases
C
Frequency before grouping
Controls
D
PEM
1.5
1
0.5
0
−0.5
−1
−1.5
2
4
6
8
10
12
E
Frequency after grouping
Cases
0.4
0.3
0.2
0.1
F
Frequency after grouping
Controls

Figure 1:

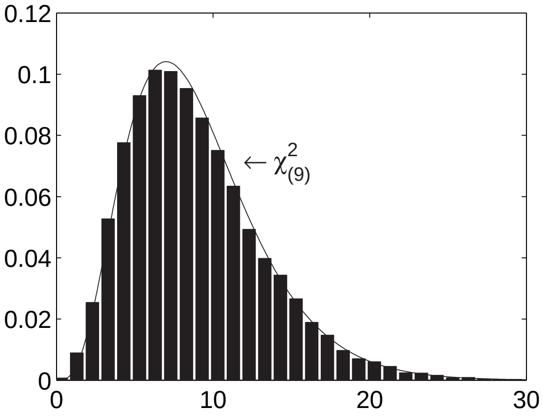

Figure 2:

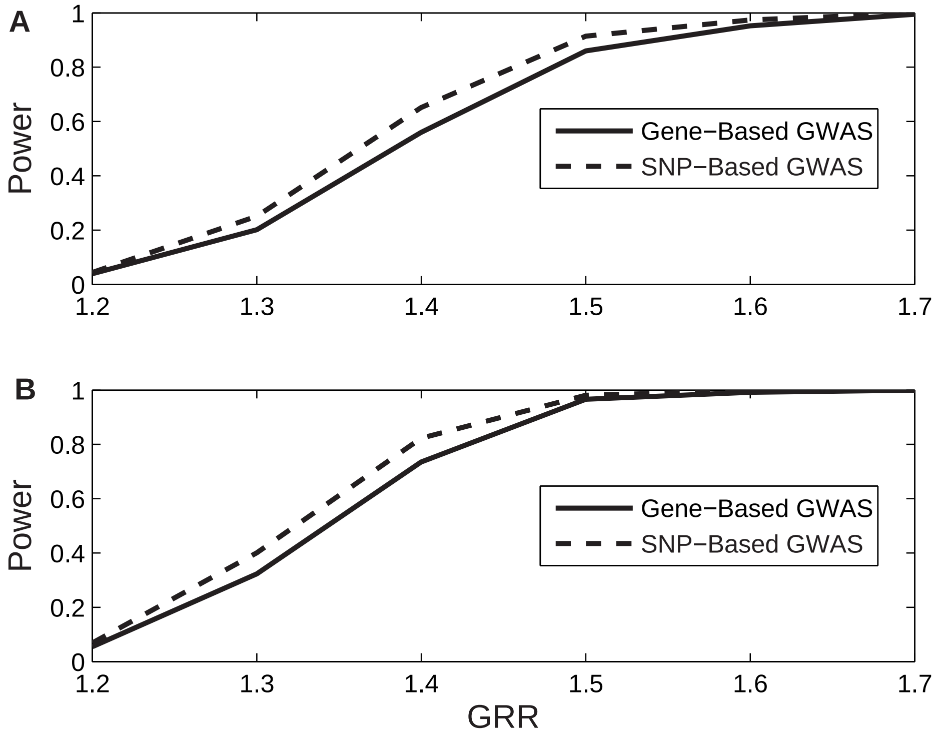
A
B
Power
Power
GRR
0
0.2
0.4
0.6
0.8
1
1.2
1.3
1.4
1.5
1.6
1.7
Gene−Based GWAS
SNP−Based GWAS

Figure 3:

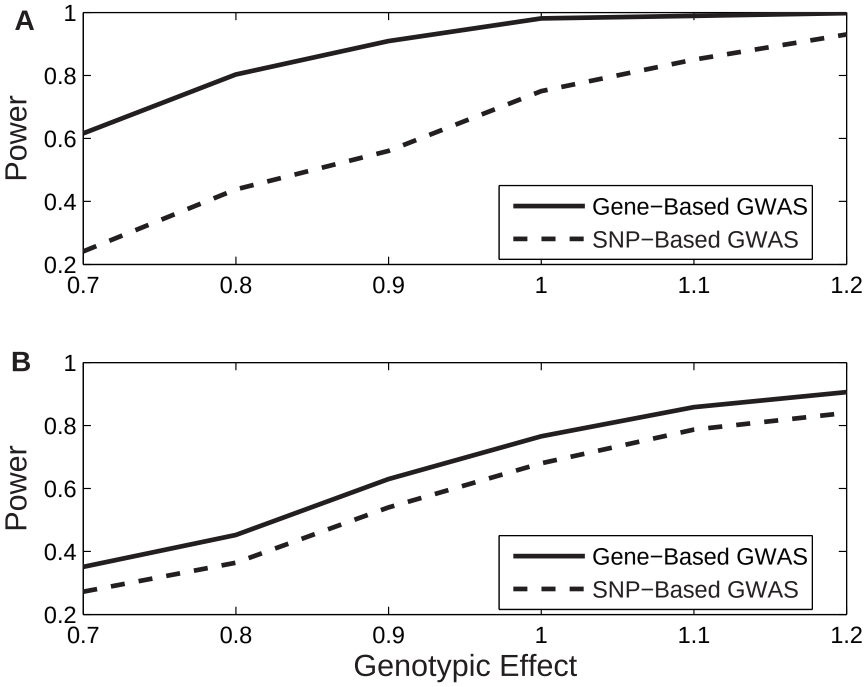
A
B
Power
Power
1
0.8
0.6
0.4
0.2
0.7
0.8
0.9
1
1.1
1.2
Gene−Based GWAS
SNP−Based GWAS
Genotypic Effect

Figure 4:

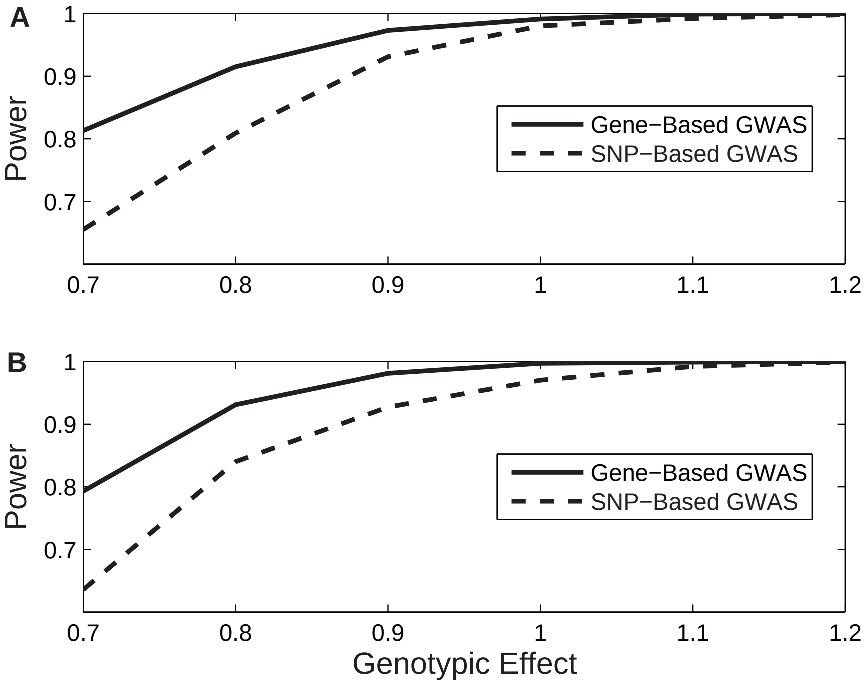
A
B
Power
Power
1
0.9
0.8
0.7
0.7
0.8
0.9
1
1.1
1.2
Genotypic Effect
Gene−Based GWAS
SNP−Based GWAS

Figure 5: