Boosting signals in gene-based association studies via efficient SNP selection

Cen Wu and Yuehua Cui

Submitted: 6th October 2012; Received (in revised form): 3rd December 2012

Abstract

Set-based association studies based on genes or pathways have shown great promise in interpreting association signals associated with complex diseases. These approaches are particularly useful when variants in a set have moderate effects and are difficult to be detected with single marker analysis, especially when variants function jointly in a complicated manner. The set-based analyses use a summary statistic such as the maximum or average of individual signal (e.g. a chi-square statistic) over all variants in a set, or consider their joint distribution to assess the significance of the set. The signal obtained with this treatment, however, could be potentially diluted when noisy variants are not taken good care of, leading to either inflated false negatives or false positives. Thus, the selection of disease informative single-nucleotide polymorphism (diSNPs) plays a crucial role in improving the power of the set-based association study. In this work, we propose an efficient diSNP selection method based on the information theory. We select diSNP variants by considering their relative information contribution to a disease status, which is different from the usual tag SNP selection. The relative merit of pre-selecting diSNPs in a set-based association analysis is demonstrated through extensive simulation studies and real data analysis.

Keywords: entropy; gene-centric association; mutual information; set-based association

INTRODUCTION

The disease gene discovery lingers on as one of the hot topics in the post-genomic era. The accessibility of high-density single-nucleotide polymorphism (SNP) data and the development of high-throughput genotyping technology enable us to broaden the hunting of genetic variants associated with complex diseases to the entire human genome. Recent development in next-generation sequencing technology even pushes the post-genomic analysis to another level via considering the contribution of rare variants. The bottleneck we are facing now is not how we can generate more data, but rather on how one can efficiently analyze these massive amount of genetic data with sound statistical approach to gain sufficient power to detect and interpret association signals. Although traditional analysis of single variants in a genome-wide scale has achieved great success in understanding the genetic basis of many complex diseases, the underlying genetic machinery of many complex diseases is still unclear, leaving large proportion of genetic variability unaccounted for by traditional single SNP analysis.

Due to the limitation of single variant analysis, much effort has been sought to find alternative strategies. Of particular interest is the recent methodology development in set-based association analysis such as the gene-centric analysis [1, 2] or pathway-based analysis [3]. These set-based analyses have been shown to provide novel interpretation of the disease signals. The set-based analysis in genome-wide association studies (GWASs) was first conceptualized by Neale and Sham [4] and further motivated by microarray gene-set analysis [5, 6]. In microarray gene-set
analyses, a gene-set score is typically obtained by averaging over individual scores in a set, as adopted by Schaid et al. [7] in their set-based association analysis. As pointed out by Schaid et al. [7], ‘The benefit of grouping genes into sets would be the greatest if many genes in a given set are associated with a trait’, which might not be the case in practice. It is less likely that all SNPs in a gene or all genes in a pathway are functional. Thus, the negative effect of averaging is that noisy signals can easily dilute the set-based signals, leading to low power. Various gene-set analyses in a GWAS content have been proposed. Pros and cons were extensively discussed (see for example [7, 8]). Therefore, it is of vital importance to choose informative SNPs first, and then to carry out the set-based disease gene association analysis.

SNP selection has been primarily done in selecting tag SNPs (tSNPs) for the purpose to reduce genotyping cost while maintaining reasonable coverage of linkage disequilibrium (LD) in a region. Daly et al. [9] and Gabriel et al. [10] originally reported the block characteristics of genomic regions which can be parsed into a series of discrete haplotype blocks with low intra-block diversity and high inter-block correlation. Various methods have been proposed to identify tSNPs since then (e.g. [11–14]). However, all these tSNP selection methods are primarily focused on covering local genomic variability. Disease phenotype information is not considered in the selection process. The selected tSNPs are useful in capturing the local LD information, but may not be useful in terms of assessing disease association. It is essential to incorporate disease status into the selection process for the purpose of association analysis.

In the last decades, much effort has been made in genetic association analysis from the information content perspective (reviewed in [15]). Along the line, information theory has also gained its popularity in SNP selection. The entropy-based measure for SNP selection can be traced back to Nothnagel et al. [16]. The normalized entropy difference, a new version of the traditional multilocus measure for LD, was originally developed under the framework of information theory to identify LD blocks and quantify simultaneous multilocus LD. More complex entropy measures have then been intensively studied. Hampe et al. [17] designed a distance-weighted global entropy metric to integrate the information on both marker set distribution and haplotype structure to select SNPs. Liu and Lin [18] demonstrated the utility of a weighted entropy measure of multilocus LD for the selection of tSNPs using generalized mutual information. Again, all these SNP selection methods do not take disease status into consideration. Thus, they cannot be directly applied to select disease informative SNPs (diSNPs).

Motivated by the benefit of information theory in tSNP selection, we propose to select diSNPs from an information content perspective. By selecting diSNPs, we expect to improve the power of set-based association analysis. The diSNP variants are chosen by considering their relative information contribution to a disease status, which distinguishes our strategy from the usual tSNP selection methods. We systematically investigate the proposed information algorithm via extensive simulation studies. In particular, we evaluate the efficiency and sensitivity of the selection process and show how selection will affect gene-based testing power assuming different disease models. The utility of the approach was demonstrated by applying to a real data set.

**METHOD**

**Information theory**

The Shannon Entropy denoted as $H(X)$ is the building block of information theory. It measures the uncertainty of a random variable $X$ in bits by

$$H(X) = - \sum_{i} p(x_i) \log_2 p(x_i).$$

(1)

Its unit is in nats if the base 2 logarithm is replaced by the natural logarithm in Equation (1). The entropy measure achieves the maximum if $X$ is uniformly distributed. In an association study, let $X$ be a SNP random variable with genotype frequencies $p(aa)$, $p(aA)$ and $p(AA)$, then the entropy of the SNP can be calculated as

$$H(X) = -p(aa) \log_2 p(aa) - p(aA) \log_2 p(aA) - p(AA) \log_2 p(AA).$$

(2)

Entropy can be extended to quantify the joint uncertainty of more than one random variable. For two random variables $X$ and $Y$, the joint entropy $H(X,Y)$ is defined as

$$H(X,Y) = - \sum_{i} \sum_{j} p(x_i, y_j) \log_2 p(x_i, y_j).$$

The relationship between the joint entropy of a pair of random variables and their respective entropies can be characterized by $H(X,Y) \leq H(X) + H(Y)$. The conditional entropy $H(X|Y)$ of $X$ given $Y$ can be defined as
\[ H(X|Y) = - \sum_i \sum_j p(x_i,y_j) \log_2 p(x_i|y_j) \]

\[ = H(X,Y) - H(Y). \]

It can be understood as the remaining uncertainty of \( X \) given the knowledge of \( Y \). Following the definition of entropy, one can define the mutual information \( I(X|Y) \) which measures the information contribution of \( X \) to \( Y \) (or vice versa) as

\[ I(X|Y) = - \sum_i \sum_j p(x_i,y_j) \log_2 \frac{p(x_i|y_j)}{p(x_i)p(y_j)} \]

\[ = H(X) - H(X|Y) = H(Y) - H(Y|X) \]

\[ = H(X) + H(Y) - H(X,Y). \]

\( I(X|Y) \) quantifies the decrement in entropy (or uncertainty) of \( X \) (or \( Y \)) when \( Y \) (or \( X \)) is observed. The connection between the conditional entropy and the mutual information can be established through the equation \( H(X|Y) = H(X) - I(X|Y) \). It then follows immediately \( I(X|X) = H(X) \), which manifests the idea that the mutual information of \( X \) and itself is its entropy, hence entropy can be viewed as the self-information. \( I(X|Y) \) is non-negative and reaches 0 when \( X \) and \( Y \) are statistically independent.

Mutual information was initially applied in the information theory to calculate the capacity of the channel, and then served as a useful tool to bridge information theory to many other research areas. Let \( D \) be the disease status and \( X \) represents SNP genotype. Then, the dependency between \( X \) and \( D \) can be measured by \( I(D|X) \). Unlike the usual tSNP selection strategies, we define the relative information contribution of a SNP to the disease status \( D \) by \( I(D|X) \) which measures the reduction of the uncertainty of the disease status due to the knowledge of the SNP. More generally, mutual information can quantify the dependency between a disease phenotype and a group of \( K \) SNPs as \( I(D|X_1, X_2, \ldots, X_K) \).

**diSNP selection via entropy**

Intuitively, an efficient SNP subset selection method should ultimately lead to the discovery of disease susceptibility variants, which makes the inclusion of disease phenotype information indispensable. Given a group of SNPs \( X_1, X_2, \ldots, X_K \), the contribution of the \( i \)th SNP to the information gain about the disease status \( D \) can be measured by

\[ I(D|X_1, X_2, \ldots, X_K) - I(D|X_1, X_2, \ldots, X_{(-i)}, \ldots, X_K). \]

where the notation \( X_{(-i)} \) means that the \( i \)-th SNP is excluded when calculating the mutual information.

Intuitively, given a set of markers that have been selected, the next to be included marker \( Y_j \) should maximize the information gain about the disease status \( D \), i.e. the one that maximizes

\[ Q_j = \max \{ I(D|Y_j, X_1, X_2, \ldots, X_i) - I(D|X_1, X_2, \ldots, X_i) \}. \]

We term this as the mutual information criterion as shown in plots in later sections. If the multilocus haplotype frequency information is available, one can directly calculate the mutual information based on multilocus haplotype frequencies and obtain \( Q_j \). However, the multilocus haplotype information is generally unknown and the frequency estimation is computationally expensive when the number of markers is large. This problem can be eased by calculating the mutual information based on the joint genotype frequencies. By considering the order of SNP inclusion into the final selection set, the relative importance of a selected SNP to the disease status can also be determined. Those included into the set earlier contribute more to the disease status. For computational efficiency purpose, we choose a forward selection strategy, i.e. the selected SNPs will no longer be evaluated for further drop out. As more SNPs are included, a critical issue is to determine a stopping rule which is described later in this section.

When multilocus haplotype information is not available, Li [19] simplified the problem and proposed to consider pairwise haplotypes. Briefly, whether or not to include SNP marker \( Y \) can be determined if \( Y \) maximizes the minimum information gain of all the pairwise haplotype \((Y,X_i)\), i.e.

\[ Q_j = \max \{ \min \{ I(D|X_1, Y_j) - I(D|X_1) \} \}. \]

The above criteria could be regarded as the uniform maximization of the minimum information difference between pairwise haplotypes and the corresponding single SNPs. We further simplified this approach by using genotype frequencies between two SNPs. We term this method as maxmin. An alternative strategy would be to relax the minimum information disparity to the average of all with respect to \( Y_j \), i.e. choose the one that maximizes the following quantity:

\[ Q_j = \max \{ \text{mean} \{ I(D|X_1, Y_j) - I(D|X_1) \} \}. \]

We call this method as maxmean. Before applying the above described selection methods, the first SNP can be selected based on the maximum information
gain, i.e. the one that maximizes $\max_i I(D|X_i)$. Once the first SNP is selected, we proceed to the next one according to the above algorithm. However, all face the same issue on how to stop the selection process. Li [19] proposed a strategy by defining the quantity in Equation (5) as $\sigma_k$, $k = 1, \ldots, K$. The algorithm terminates at the first $k$ such that

$$\frac{\sigma_k}{\log(k)} < \frac{\sigma_1}{\epsilon},$$

where $\epsilon$ is a user-defined parameter which is set as $\epsilon \geq 1$ as suggested by the author. This criterion is quite arbitrary. A sensible stopping criterion should reflect the quantity or the proportion of total information gain expressed by SNPs that have already been included. Here, we propose an alternative solution to this while there might be many other possible choices. We propose that no SNPs should be added to the set at the largest $k$ such that

$$\rho = \frac{\sum_{i=1}^{k} Q_i}{\sum_{i=k+1}^{K} Q_i} \leq \epsilon,$$

where $K$ and $k$ are the total number of SNPs and the number of SNPs in the selected set, respectively. Note that $\rho$ is a normalized quantity, hence $\rho \in [0,1]$. This criterion shows the relative information contribution of SNPs in the selection set with regard to all SNPs, to the disease status. In general, one can choose $\epsilon = 0.85$ or 0.95 depending on how conservative one is. In the simulation study, we demonstrate how one can choose a reasonable cutoff based on a ratio plot. The following algorithm briefly illustrates the selection procedure.

### Algorithm 1 diSNP selection algorithm

if # of SNPs in gene $i \leq 2$ then
    No SNP selection.
else
    Select SNP $X$ that maximizes $I(D|X)$;
    for $j = 1$ to $(\log_{10} \text{SNPs per gene} - 1)$
        Calculate $Q_j$ for each SNP;
    end for
    Calculate the selection ratio $\rho$ for each SNP defined in (8) and include all SNPs satisfying (8).
end if

### SIMULATION STUDY

We evaluated the performance of different selection methods by extensive simulation studies. We simulated genotype data assuming known pairwise LD information. Let $p_A$ and $p_B$ be the minor allele frequencies (MAFs) of two neighboring SNPs, and the LD between them be $D$. The correlation coefficient $r$ can be calculated as $r = D/\sqrt{p_A p_B \text{MAF}_A \text{MAF}_B}$. The two SNPs form four haplotypes with frequencies given by

$$p_A \text{MAF}_A + D, \quad p_B \text{MAF}_B,$$

for genotypes $AA$, $Aa$ and $aa$, respectively. Assuming there are total $K$ SNPs in a gene, we began by simulating the first SNP with two alleles denoted as $A$ and $a$, where $A$ is the minor allele. Under the assumption of Hardy–Weinberg equilibrium (HWE), the SNP genotype can be simulated from a multinomial distribution with genotype frequencies $p_A^2$, $2p_A(1-p_A)$ and $(1-p_A)^2$ for genotypes $AA$, $Aa$ and $aa$, respectively. The genotype of the adjacent SNP can then be generated from the conditional distribution given the first SNP. For instance,

$$p(bb|aa) = \frac{p_{ab}^2}{p(aa)} = \frac{p_{ab}^2}{(1-p_A)^2}.$$ 

The detailed conditional distribution of the second marker given the first one is given in Cui et al. [1]. Once the second SNP is generated, the rest SNPs can be simulated sequentially following the same procedure. The advantage of this simulation is that one can easily control the degree of LD between two adjacent SNPs, thus leaving much flexibility in evaluating the effect of LD on testing power and SNP selection efficiency. For simplicity, we generated 10 SNPs for a gene assuming different LD block structures. SNPs are correlated within each block but are independent between blocks.

### Efficiency of diSNP selection strategies

#### The ratio plot

We simulated disease phenotype data considering several scenarios with different gene actions (Table 1). Different LD blocks were assumed in different scenarios. Specifically, we simulated two LD blocks under the null model (SNPs 1–5 and 6–10), three LD blocks for Model A (SNPs 1–3, 4–7 and 8–10) and three LD blocks for Model B (SNPs 1–4, 5–7 and 8–10). For simplicity, the same LD was assumed within each block. The LD correlation coefficient $r$ was set to 0.5 and 0.9 to investigate the performance of different selection criteria under weak and strong LD correlations. No genetic effect was considered for the null model. In Model A, three SNPs (1, 6 and 10) in each block were assumed...
The relative advantage of maxmin over the other criterion. As allele frequency and sample size increase, the transition point becomes more clear. Thus, the transition point was observed when using the maxmean criterion. A transition point is quite consistent under different LD structures. The maxmean criterion completely fails under strong LD ($r = 0.9$). Based on the ratio plot, a higher cutoff is needed in this case.

In summary, we can see that the maxmin criterion performs the best among the three. In real application, a ratio plot can help guide one to choose an appropriate cutoff. When a ratio plot as shown in Figure 1 is observed, this may indicate no gene effect or very weak gene signal. Selection may lead to no power gain and one may choose not to do any selection in this case. When there is a gene effect, one can choose a conservative cutoff such as $c = 0.95$ for all genes or choose it according to the ratio plot for each gene. Even though a constant cutoff may increase the rate of false selection for some genes, this is not an issue as we do not do association test in this selection step and we do not want to miss any potential signals. As shown in ‘Power gain by SNP selection’ section, we do observe substantial power gain by applying a fixed threshold while maintaining reasonable false-positive rate.

### Selection sensitivity analysis

The above section provides a practical guidance on how to choose an appropriate cutoff based on the ratio plot. In this section, we show the selection sensitivity by applying a fixed threshold. We fixed the selection cutoff as $c = 0.95$. The SNP models and LD blocks were the same as described above. The proportion ($t$) of each of the 10 SNPs being selected out of 1000 simulation replications was shown by a bar graph assuming different MAFs ($p = 0.1$, 0.3 and 0.5), different sample sizes ($n = 500$ and 1000) and different LD structures ($r = 0.5$ and 0.9) within each block. The results are shown in Figures 4–7. In each plot, the horizontal axis shows the actual SNP ID and the vertical axis indicates the proportion of selection out of 1000 replicates.

Figures 4 and 5 show the results for Model A under different LD structures. The causal SNPs are SNPs 1, 6 and 10. Clearly, the maxmin criterion can select the true diSNPs in all cases. Even though the other two criteria can also select the true diSNPs, the false-selection rate for those SNPs without effect is also quite high. The results for Model B are shown in Figures 6 and 7 with causal SNPs 2, 5 and 9.

### Notes

- $p = Pr(y = 1|S)$, $y$ denotes the disease status and $S_j$ denotes SNPs.
- to affect the disease status and there is no interaction among them. Model B considers both main, two-way and three-way interactions among SNPs 2, 5 and 9 from the three LD blocks.
- We evaluated different SNP selection methods defined in Equations (4–6), respectively. In each plot, the horizontal axis represents the number of SNPs included into the selection set and the vertical axis represents the ratio ($p$) averaged over 1000 simulation replicates. Figure 1 plots the selection under the null model. The plots indicate that the information contribution by all SNPs is uniform. No SNP dominates the others in terms of the information contribution to the disease status.

#### Table 1: List of data generating models

<table>
<thead>
<tr>
<th>Model</th>
<th>Gene action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>$\text{logit}(\mu) = \mu$</td>
</tr>
<tr>
<td>A</td>
<td>$\text{logit}(\mu) = \mu + 1.2S_1 + 1.2S_4 + 1.2S_{10}$</td>
</tr>
<tr>
<td>B</td>
<td>$\text{logit}(\mu) = \mu + 1.2S_1 + 1.2S_5 + 1.2S_9$ + $1.1S_2S_5 + 1.1S_3S_6 + 1.1S_2S_9 + 1.1S_3S_5S_6$</td>
</tr>
</tbody>
</table>

Notes: Where $p = Pr(y = 1|S)$, $y$ denotes the disease status and $S_j$ denotes SNPs.
Figures 1 and 2: Ratio plots for null and Model A models under different correlation structures, sample sizes and allele frequencies. Note that the numbers in the horizontal axis do not directly correspond to the SNP IDs. They represent the number of SNPs included into the selection set.
Figure 3: Ratio plot for Model B under different correlation structures, sample sizes and allele frequencies. Note that the numbers in the horizontal axis does not directly correspond to the SNP. They represent the number of SNPs included into the selection set.

Figure 4: The proportion ($\tau$) of selected SNPs for each of the 10 simulated SNPs in a gene under Model A under different sample sizes and MAFs. Pair-wise correlation within each block is assumed to be $r = 0.5$. 
Figure 5: The proportion (τ) of selected SNPs for each of the 10 simulated SNPs in a gene under Model A under different sample sizes and MAFs. Pairwise correlation within each block is assumed to be $r = 0.9$.

Figure 6: The proportion (τ) of selected SNPs for each of the 10 simulated SNPs in a gene under Model B under different sample sizes and MAFs. Pairwise correlation within each block is assumed to be $r = 0.5$. 
involving both main and interaction effects. Similar patterns are observed and the maxmin criterion performs the best among the three with high sensitivity and low false-selection rate for those non-effect SNPs. Also noted that the false-selection rate decreases as sample size increases for the maxmin in general for both models. The pattern is especially prominent when data were simulated from Model B. We also simulated data when there is no SNP effect in a gene. The results show that all the SNPs roughly have the same chance to be chosen and the plot shows a uniform pattern (data not shown). In our simulations, we also found that diSNPs with genetic effects always entered into the selection set with high priority. This information implies the robustness of the selection criterion. In summary, our simulation study shows the advantage of the maxmin approach over the other two. In the next section, we evaluate the selection power gain after SNP selection focusing on the maxmin criterion.

**Power gain by SNP selection**

The previous section demonstrates the efficiency of SNP selection by the maxmin criterion. Next, we show the power gain by selection focusing on a gene-centric analysis approach proposed by Cui et al. [1]. For simplicity, we choose a conservative cutoff \( c = 0.95 \) in all the simulation runs even though different cutoffs could be chosen based on the ratio plot for each simulation. As the real functional mechanism for a gene is generally unknown, it is generally difficult to simulate the true complex relationship between the genotype and phenotype. Here, we simulated phenotypes from a diversity of representative disease models as adopted in Cui et al. [1] to evaluate the performance of the testing power before and after SNP selection. To make the work self-contained, we briefly introduce the disease models used for our simulation study.

In the simple one-locus disease model, let the penetrance function \( f_j \) be the probability of being affected given possessing \( j \) copies of disease alleles \((j = 0, 1 \text{ and } 2)\). In addition, let \( \lambda = f_1/f_0 \) be the genotype relative risk (GRR) and \( p \) be the disease MAF. Then, for \( n_1 \) cases and \( n_2 \) controls, the individual disease status can be simulated from a binomial distribution with probability of success \( f_j \) calculated from the additive and multiplicative model given in Table 2, where prev denotes population prevalence.

When there are two loci function in a gene, a two-locus disease model is defined in Table 3. The model can also be found in Marchini et al. [20]. In Model 1, the odds of the disease increase with genotype multiplicatively both with and between loci.

![Figure 7: The proportion (τ) of selected SNPs for each of the 10 simulated SNPs in a gene under Model B under different sample sizes and MAFs. Pairwise correlation within each block is assumed to be \( r = 0.9 \).](image-url)
Table 2: Single-locus disease model

<table>
<thead>
<tr>
<th>Disease model</th>
<th>( f_0 )</th>
<th>( f_1 )</th>
<th>( f_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive</td>
<td>( \frac{p_{\text{prev}}}{1-p_{\text{prev}}} )</td>
<td>( f_0 \lambda )</td>
<td>( 2f_0\lambda - f_0 )</td>
</tr>
<tr>
<td>Multiplicative</td>
<td>( \frac{1}{1-p_{\text{prev}}^2} )</td>
<td>( f_0 \lambda )</td>
<td>( f_0\lambda^2 - f_0 )</td>
</tr>
</tbody>
</table>

Notes: Where \( \text{Prev} = \text{the population prevalence} \); \( p = \text{disease MAF} \) and \( \lambda = f_1/f_0 \) denotes the GRR.

Table 3: Two-locus interaction disease model

<table>
<thead>
<tr>
<th>Model</th>
<th>BB</th>
<th>Bb</th>
<th>bb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>Aa</td>
<td>( \gamma(1+0)^3 )</td>
<td>( \gamma(1+0)^2 )</td>
</tr>
<tr>
<td></td>
<td>Gb</td>
<td>( \gamma(1+0) )</td>
<td>( \gamma(1+0)^2 )</td>
</tr>
<tr>
<td></td>
<td>Gc</td>
<td>( \gamma(1+0) )</td>
<td>( \gamma(1+0)^2 )</td>
</tr>
<tr>
<td>Model 2</td>
<td>Aa</td>
<td>( \gamma(1+0)^3 )</td>
<td>( \gamma(1+0)^2 )</td>
</tr>
<tr>
<td></td>
<td>Gb</td>
<td>( \gamma(1+0) )</td>
<td>( \gamma(1+0)^2 )</td>
</tr>
<tr>
<td></td>
<td>Gc</td>
<td>( \gamma(1+0) )</td>
<td>( \gamma(1+0)^2 )</td>
</tr>
</tbody>
</table>

Notes: Where \( \gamma = \text{baseline effect} \) and \( \theta = \text{genotypic effect} \).

while in Model 2, the odds are the same within the baseline genotype and increase multiplicatively when there is at least one disease allele.

When there are three loci \((G_A, G_B, G_C) \in (0,1,2)\) interacting with each other within a gene, the previous two-locus interaction disease model can be extended to a three-locus case as defined in the following:

Model 1: Odds \((G_A, G_B, G_C) = \gamma(1+0)^{G_A+G_B+G_C} \)
Model 2: Odds \((G_A, G_B, G_C) = \gamma(1+0)^{G_A}(G_{C>0})+G_{B>0}(G_{C>0}) \)

Here, we adopted the entropy approach proposed in Cui et al. [1], to evaluate the testing power before and after SNP selection with the maxmin selection criterion. The detailed method can be found in Cui et al. [1], hence is omitted here. For simplicity, we assumed a homogeneous LD \((r^2 = 0.7)\) within each block and MAF = 0.4 in all simulations. We first evaluated the type I error rate before and after selection. We did 2000 replications to assess the error rate. The type I errors under different sample sizes before diSNP selection are 0.053 \((n = 500)\), 0.046 \((n = 1000)\) and 0.055 \((n = 2000)\). Correspondingly, the type I errors after diSNP selection are 0.054 \((n = 500)\), 0.0625 \((n = 1000)\) and 0.061 \((n = 2000)\). So the error rate is reasonably controlled after selection. Table 4 summarizes the power results for the single-locus disease model given in Table 2. A single cutoff threshold \(c = 0.95\) was applied to all 1000 simulation runs. The power after SNP selection is shown in the parenthesis. As we can see, the testing power generally improves after SNP selection.

Table 4: Power of gene-based test before and after diSNP selection under the single-locus disease model

<table>
<thead>
<tr>
<th>Disease model</th>
<th>((\text{Prev, GRR}))</th>
<th>(n = 500)</th>
<th>(n = 1000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive</td>
<td>((0.15, 1.3))</td>
<td>0.258 (0.287)</td>
<td>0.492 (0.521)</td>
</tr>
<tr>
<td>Multiplicative</td>
<td>((0.15, 1.5))</td>
<td>0.516 (0.546)</td>
<td>0.789 (0.832)</td>
</tr>
</tbody>
</table>

Notes: The fourth SNP was set as the disease locus for both models.

A CASE STUDY

We applied the selection method to a genetic association study of small for gestational age (SGA) neonates. Infants whose body weight falls below the 10th percentile for gestational age are classified as SGA [21]. A candidate gene study was conducted in which 189 genes containing total of 820 SNPs were genotyped with infant subjects recruited through the Department of Obstetrics and Gynecology at Sotero del Rio Hospital in Puente Alto, Chile. Total 753 SNPs in 183 genes were left after removing SNPs showing departure from HWE in the control sample and those with MAF < 0.05. Total 252 cases and 673 controls were analyzed. In addition, no evidence of population stratification was detected using the genomic control method [22]. For more details about the description of the data, readers are referred to Li et al. [23].

We applied the gene-based test proposed by Cui et al. [1] to all genes containing two or more SNPs, before and after diSNP selection based on the
maxmin criterion using an adaptive cutoff. Table 6 lists genes with $P$-values <0.005. Most genes show improved $P$-value except for gene COL1A2. Figure 8 shows the ratio plot for genes LPA and COL1A2. For gene LPA, it is clear that the contribution of the last three to four SNPs is quite small. The ratio plot is dominated by the first four selected diSNPs and a turning point is observed. Thus, we can select four diSNPs or conservatively choose five. Similar trends were observed for other genes except for gene COL1A2. For gene COL1A2, the ratio plot is very similar to the null ratio plot (Figure 1) in the simulation study, which indicates that there is potentially no SNP effect, hence no gene effect. Thus, the small $P$-value obtained before SNP selection for gene COL1A2 might be a false positive. It is also possible that all SNPs may function but each with a small effect. In this case, no selection is needed. However, the true function of SNPs is generally unknown in reality. A practical guidance is that no selection should be done when a ratio plot displays a pattern such as gene COL1A2 or Figure 1. The ratio plot provides a practical guidance for threshold determination.

**DISCUSSION**

Statistical analysis focusing on single SNPs in GWASs has been traditionally pursued. The recently developed gene-centric or pathway-based analysis represents a paradigm shift from single SNP analysis and has been proved to be successful in the detection of novel variants as well as the interpretation of disease signals [4, 3]. Such an analysis considers modeling the joint effect of multiple SNPs, termed set-based analysis. However, when all SNPs are included in a set, those noise SNPs with no genetic contribution could potentially dilute the set-based signal, leading to low power [7]. Thus, it is statistically important and practically meaningful to develop a procedure to exclude potential noise SNPs before applying a set-based analysis.

In this work, we proposed an efficient diSNP pre-selection method based on the information theory. Three selection criteria were proposed and evaluated by extensive simulation studies. Among the three, maxmin was shown to have the best performance. We also evaluated the selection performance of maxmin criterion in terms of power gain before and after diSNP selection via a gene-centric association method [1]. The simulation results indicate substantial power gain by eliminating potential noise SNPs. In our simulation studies, we evaluated the power using the gene-centric approach proposed by Cui et al. [1], although other gene-based methods (e.g. [2, 7]) can also be used to demonstrate the utility. When one uses an average procedure to average SNP signals in a set (e.g. [7]), the selection procedure becomes even more critical. Our method provides a quantitative framework for efficient diSNP selection in set-based association analysis.

For the real data analysis, we listed those genes with $P$-values <0.005 to show the utility of the

---

**Table 5:** Power of gene-based test before and after diSNP selection under the two- and three-locus interaction disease models

<table>
<thead>
<tr>
<th>Disease model</th>
<th>(BL, GE)</th>
<th>Two-locus model $n = 500$</th>
<th>$n = 1000$</th>
<th>Three-locus model $n = 500$</th>
<th>$n = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>(l, 0.6)</td>
<td>0.673 (0.785)</td>
<td>0.983 (0.992)</td>
<td>0.554 (0.609)</td>
<td>0.940 (0.990)</td>
</tr>
<tr>
<td></td>
<td>(l, 0.8)</td>
<td>0.833 (0.908)</td>
<td>0.994 (0.999)</td>
<td>0.585 (0.760)</td>
<td>0.979 (0.999)</td>
</tr>
<tr>
<td>Model 2</td>
<td>(l, 0.6)</td>
<td>0.560 (0.665)</td>
<td>0.939 (0.948)</td>
<td>0.541 (0.638)</td>
<td>0.946 (0.989)</td>
</tr>
<tr>
<td></td>
<td>(l, 0.8)</td>
<td>0.749 (0.798)</td>
<td>0.971 (0.978)</td>
<td>0.724 (0.818)</td>
<td>0.986 (0.999)</td>
</tr>
</tbody>
</table>

Notes: Two LD blocks (one contains SNPs 1–4 and the other contains SNPs 5–10) were simulated for the two-locus model, while three LD blocks (one contains SNPs 1–4, one contains SNPs 5–8 and the other contains SNPs 9–10) are simulated for the three-locus model. The interacting disease loci were assumed from each block.

**Table 6:** List of genes with number of SNPs and $P$-values before and after diSNP selection

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Before selection</th>
<th>After selection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of SNPs</td>
<td>$P$-value</td>
</tr>
<tr>
<td>PONI</td>
<td>5</td>
<td>0.0054</td>
</tr>
<tr>
<td>CETP</td>
<td>10</td>
<td>0.0035</td>
</tr>
<tr>
<td>LPL</td>
<td>8</td>
<td>0.0024</td>
</tr>
<tr>
<td>F2</td>
<td>4</td>
<td>0.0005</td>
</tr>
<tr>
<td>LPA</td>
<td>8</td>
<td>0.0042</td>
</tr>
<tr>
<td>NPPA</td>
<td>4</td>
<td>0.0037</td>
</tr>
<tr>
<td>COL1A2</td>
<td>12</td>
<td>0.0002</td>
</tr>
</tbody>
</table>
method. If we used Bonferroni correction, only gene PON1 is significant after diSNP selection. Single SNP analysis shows that the smallest $P$-value in this gene is 0.045 which cannot reach the genome-wide significance level. The gene level significance indicates that there might be complicated interactions among SNPs in this gene. Only when they are considered as a system with the elimination of one noise SNP, it can reach a significance. Gene PON1 has been shown to be associated with low birth weight and shorter gestational age [24]. Other studies also established association of gene PON1 with low birth weight [25, 26]. Without diSNP selection, this gene could be easily missed. The computational code written in Matlab to implement the proposed method is available for free download at http://www.stt.msu.edu/~cui.

**Key Points**

- We developed an efficient diSNP selection method based on the information theory content. The method considers disease status information when selecting diSNPs and is different from usual tSNP selection method.
- The diSNP selection excludes potential noise SNPs from further association analysis and is particularly useful when gene- or pathway-based association analysis is considered.
- Application to a real data set shows that some important genes could be missed without diSNP selection in a gene-centric association analysis.
- A computational tool is developed and is available upon request.

**Acknowledgements**

The authors wish to thank three anonymous referees for their constructive comments that greatly improved the article. Special thanks also go to Dr R Romero for providing the genetic data.

**FUNDING**

This work is partially supported by US National Science Foundation grants DMS-1209112 and IOS-1237969.

**References**

11. Weale ME, Depondt C, Macdonald SJ, et al. Selection and evaluation of tagging SNPs in the neuronal-sodium-channel...


