Sum Statistics for the Joint Detection of Multiple Disease Loci in Case-Control Association Studies With SNP Markers

Anja Wille,1,2* Josephine Hoh,1 and Jurg Ott1

1Laboratory of Statistical Genetics, Rockefeller University, New York, New York
2IMBIE, Bonn University, Bonn, Germany

In complex traits, multiple disease loci presumably interact to produce the disease. For this reason, even with high-resolution single nucleotide polymorphism (SNP) marker maps, it has been difficult to map susceptibility loci by conventional locus-by-locus methods. Fine mapping strategies are needed that allow for the simultaneous detection of interacting disease loci while handling large numbers of densely spaced markers. For this purpose, sum statistics were recently proposed as a first-stage analysis method for case-control association studies with SNPs. Via sums of single-marker statistics, information over multiple disease-associated markers is combined and, with a global significance value \( \alpha \), a small set of “interesting” markers is selected for further analysis. Here, the statistical properties of such approaches are examined by computer simulation. It is shown that sum statistics can often be successfully applied when marker-by-marker approaches fail to detect association. Compared with Bonferroni or False Discovery Rate (FDR) procedures, sum statistics have greater power, and more disease loci can be detected. However, in studies with tightly linked markers, simple sum statistics can be suboptimal, since the intermarker correlation is ignored. A method is presented that takes the correlation structure among marker loci into account when marker statistics are combined. Genet Epidemiol 25:350–359, 2003. © 2003 Wiley-Liss, Inc.

Key words: complex diseases; large-scale LD mapping; intermarker correlation

INTRODUCTION

With the availability of dense single-nucleotide polymorphism (SNP) maps, genome-wide association studies have been proposed for mapping complex trait loci [Risch and Merikangas, 1996]. Such association analyses can be carried out in a population-based framework (case-control studies) or using family data (TDT and extensions). The latter test design has the advantage that it is robust towards spurious association due to population stratification [Spielman and Ewens, 1996]. Still, the case-control design has enjoyed continuing popularity, since ascertainment of data can be carried out more easily and is less expensive. Recently, strategies to account for population admixture in case-control studies were also proposed [Devlin and Roeder, 1999; Bacanu et al., 2000].

When several disease genes contribute to a trait, one may detect them jointly by modeling the complex interaction pattern among loci. Data mining or pattern recognition techniques such as classification trees [Zhang and Bonney, 2000], neural networks [Bhat et al., 1999], and partitioning methods [Nelson et al., 2001] have been applied to predict interaction effects among different disease-associated marker loci. In general, these methods require the specification of parameters such as pruning criteria for the classification trees or the number of hidden layers in neural networks. With too restrictive a choice for these parameters, solutions may be overlooked, whereas with the opposite, observed data may be overfitted [Nelson et al., 2001]. Specification of parameters depends on the judgment of the investigator. Often it is not clear how to interpret results in terms of statistical significance.

Simulation studies for the detection of small numbers of interacting disease loci suggest an increase in power when association effects are measured synergistically rather than separately [Longmate, 2001]. This also applies to simple methods for the joint evaluation of marginal
effects [Longmate, 2001; Wille and Leal, 2001]. Combining genetic main effects will thus facilitate the detection of susceptibility genes while avoiding the need to characterize detailed interaction patterns among markers. Such a simultaneous search for marginal effects (called the “set association” approach) was recently proposed for case-control association studies with large numbers of SNPs [Hoh et al., 2001]. In this approach, a set of markers is selected and tested against the null hypothesis that none of the selected markers is associated with the disease. Briefly, its main features are as follows: 1) Each SNP marker is associated with a single-locus statistic $t_i$, for example, chi-square in a $2 \times 3$ table (two rows for case/control individuals, three columns for the SNP genotypes) for the $i$-th marker. 2) Markers are ordered by size of their statistic, $t_{(1)} > t_{(2)} > \ldots$ 3) Sums $s_j$ are formed over the $j$ largest single-locus statistics, for $j=1,\ldots,c$, where $c$ is a suitable limit (one of the parameters of the method) such as 20. Also, the empirical significance level $P_j$ associated with each $s_j$ is evaluated in computer-based permutation tests carried out under the null hypothesis of no association (labels “affected” and “unaffected” are permuted). At this point, the multiple testing of hundreds or thousands of SNPs has been reduced to $c$ sums. 4) The smallest of the $c$ significance levels, $P_{j_{\min}} = \min\{P_j, j = 1,\ldots,c\}$, is taken as the single statistic measuring association (the multiple testing problem has now been eliminated). Its significance level, $z$, is again determined in permutation tests. The $P_{j_{\min}}$ occurs at an SNP ($j_{\min}$) with statistic $t_{(j_{\min})}$. A significance level of $z<0.05$ leads to the rejection of the null hypothesis that none of the SNPs is associated with the disease. The SNPs with single-marker statistics larger than or equal to $t_{(j_{\min})}$ are then selected for further analysis. Although our analysis method does not specify which of the markers in the selected set are truly disease-associated, the further analysis for disease genes (second analysis stage) is reduced to a small set of markers, removing much of the multiple testing burden. Our approach can thus be applied as a first-stage analysis strategy when single-locus methods fail to provide significant results. A further discussion of two-stage designs, in particular with respect to cost effectiveness, can be found in Satagopan et al. [2002].

In an application to heart disease data [Hoh et al., 2001], the set association approach appeared promising, but its statistical properties are as yet unknown. The work presented here was undertaken with the following aims: 1) To investigate the power of the set association approach for genome-wide case-control studies with uncorrelated SNP markers under oligogenic threshold models for disease. In particular, we wanted to know to what degree the power to detect disease loci degrades as a function of the number of SNPs. 2) To compare the power of the set association approach with that of conventional marker-by-marker association studies for these models. 3) To test the usefulness of the approach for densely spaced markers in linkage disequilibrium (LD) with each other. With SNP markers, most researchers currently focus on candidate gene studies with several tightly linked markers in each gene. For such studies, adding SNPs to the sum only based on their single-marker statistic may be inefficient, particularly when several highly correlated SNPs from the same gene are selected in the sum. Thus, we developed a method to “deflate” the statistic of the next SNP to be selected, based on its correlation to the SNPs already in the sum. As demonstrated below, this modified approach restores most of the power that would be lost if correlation among SNPs were ignored.

**METHODS**

**DEFLATING MARKER STATISTICS**

As statistics for single markers, either the absolute difference in allele frequencies or the genotype-based $\chi^2$ with 2 df was considered. In the first step of forming sums, $s_j$ that SNP with the largest associated statistic, $t_{(1)}$, constitutes $s_1$. As mentioned in the Introduction, the empirical significance level associated with $s_1$ is determined on the basis of $n$ randomization samples (here we chose $n=5,000$). Now consider any of the next SNPs to be added to this “sum.” Let $t_i$ be the statistic of the $i$-th SNP. Its usefulness is evaluated based on the conditional distribution $P(t_i|t_{(1)})$ with mean $E(t_i|t_{(1)})$ and standard deviation $\sigma(t_i|t_{(1)})$. A strong positive correlation between $t_i$ and $t_{(1)}$ leads to $E(t_i|t_{(1)}) > E(t_i)$. Thus, we form the normal deviate $D(t_i|t_{(1)})$ of $t_i$, i.e.,

$$D(t_i|t_{(1)}) = \frac{t_i - E(t_i|t_{(1)})}{\sigma(t_i|t_{(1)})}$$

which represents a normalized adjustment of $t_i$ with respect to its correlation with the statistic $t_{(1)}$. A “deflated” statistic $t'_i$ for the $i$-th SNP is then
obtained as
\[ t_i^j = E(t_i) + \sigma(t_i) \cdot D(t_i|t_{i(1)}) \].

The selection of the next SNP to be added to the sum is based on these deflated values rather than the original single-marker statistics, that is, the next marker added is that with largest deflated statistic \( t_i^j \). For further steps, when the sum already contains j SNPs \((1),\ldots, (j)\), we proceed in an analogous manner by computing deflated statistics \( t_i^j \) based on the distribution \( P(t_i|t_{i(1)},\ldots, t_{i(j)}) \). Details for these calculations are given in the Appendix.

**DISEASE MODEL**

As a disease model for the power calculations, we chose an oligogenic threshold model [cf. Hartl and Clark, 1997], which has the advantage that for a prespecified prevalence \( K \), an arbitrary number of susceptibility loci with arbitrary marginal genetic contributions can be specified. The disease is taken to be expressed when the level of an underlying normally distributed quantitative phenotype \( Q \) (the liability) exceeds a certain threshold \( T \), where \( T \) determines the prevalence \( K \). The quantitative phenotype depends on the genotype at each susceptibility locus. If a disease locus \( l \) has two alleles \( a \) and \( b \), where \( a \) denotes the allele that predisposes to the disease and \( Q \sim N(0,1) \), then
\[
Q|aa \sim N(\mu_{aa}, 1 - h_l^2) \\
Q|ba \sim N(\mu_{ab}, 1 - h_l^2) \\
Q|bb \sim N(\mu_{bb}, 1 - h_l^2)
\]

with means \( \mu_{aa} \geq \mu_{ab} \geq \mu_{bb} \). The genetic contribution of locus \( l \) is described by the heritability \( h_l^2 \), the variance of \( Q \) attributable to the genotypic variation at \( l \). The overall heritability \( h^2 \) refers to the combined genotypic variation at all susceptibility loci.

In our model, we assumed all effects to be additive (no dominance effects at the genotypic and allelic levels). This simplifying assumption is often made in simulation studies, e.g., in Göring et al. [2001]. We followed this practice because, with varying prevalence and heritability parameters and a varying number of disease genes, a wide spectrum of genetic effects can be covered. For additive effects, the overall heritability \( h^2 \) is the sum of the single-locus heritabilities. In our analysis, we set \( h^2 \) to 0.2, 0.5, and 0.8, and considered 5 or 10 susceptibility loci. The genetic contribution of each locus was either set equal for all loci, or it was modeled to differ from locus to locus in a linear fashion, \( h_l^2 = l \cdot h^2 \). Once the single-locus heritabilities \( h_l^2 \) are specified, the genotypic means of locus \( l \) are calculated from
\[
\mu_{aa} = 2p_a \sqrt{h_l^2/2p_ap_b} \\
\mu_{ab} = (p_a - p_b) \sqrt{h_l^2/2p_ap_b} \\
\mu_{bb} = -2p_a \sqrt{h_l^2/2p_ap_b}
\]
where \( p_a \) and \( p_b \) are the allele frequencies at locus \( l \) [Hartl and Clark, 1997].

Values for the overall heritability \( h^2 \) and the prevalence \( K \) (set to either 0.01 or 0.05) were chosen according to parameters observed for common complex diseases such as schizophrenia or late-onset diabetes [Cavalli-Sforza and Bodmer, 1999].

**MARKER DATA**

To evaluate the power of the set association approach, 500 replicates of case-control data sets were generated with 200, 500, 1,000, and 2,000 markers. Sample sizes varied between 100–1,000 individuals, evenly split into cases and controls.

For the marker data of each individual, SNP allele frequencies were sampled from a \( \beta(0.2,0.8) \)-distribution which is “u”-shaped, i.e., values around 0.5 are less common than values near the boundaries. SNPs with minor allele frequencies smaller than 0.1 were discarded, since these SNPs are only little informative [Kruglyak, 1997], and are usually not deployed as markers in a study [Tu and Whittemore, 1999]. In addition to a given set of SNP markers, trait loci were assumed to be located in several of the genotyped regions displaying linkage disequilibrium of \( D'=0.5 \) [Lewontin, 1964] with the markers in these regions. \( D'=0.5 \) corresponds to a physical distance up to 100 kb [Reich et al., 2001; Pritchard and Przeworski, 2001]. Genotypes at the susceptibility loci were used to model the affection status, as described above, but were taken to be unknown, i.e., they were not included in the marker data. The frequencies of predisposing alleles were uniformly set to 0.2.

In order to evaluate whether in a candidate gene approach the use of statistics \( t_i \) from correlated markers is inefficient and whether our proposed adjustment is helpful, we generated haplotype data of tightly linked markers. In the sum statistics, only the amount of linkage disequilibrium, and not the genetic or physical distance among markers, is relevant. Since the relation between distance and amount of LD between markers cannot be predicted accurately [Jorde, 2000], we simulated the correlation structure of tightly linked markers based on LD values only.
We assumed 504 SNPs in 63 candidate genes, with 8 SNPs in each gene. Ten candidate genes were disease genes, i.e., each of them harbored 1 of 10 susceptibility loci.

For the SNPs in a gene, haplotype data with no, weak, medium, and strong LD were generated by setting $D'$ to 0.0, 0.2, 0.5, and 1, respectively, for all 28 SNP marker pairs. In the 10 disease genes, when no LD among markers was assumed, $D'$ was set to 0.5 between the predisposing allele and the nearest SNPs. When LD among markers was taken to exist, $D'$ was set to 0.5 between the predisposing allele and each of the 8 SNPs in the disease gene. Negative values for $D'$ need not be modeled, because we did not distinguish between minor and major alleles when simulating allele frequencies. Haplotypes were combined to form multilocus genotypes under HWE.

In real data sets, some genotyped genes will show strong intermarker LD, and others not. To reflect this and to specifically test the effect of LD among markers in disease and nondisease genes, we distinguished four different LD settings: a) all markers are uncorrelated, b) there is LD in all 10 disease genes but no LD in nondisease genes, c) there is no LD in disease genes but there is LD in nondisease genes, and d) there is intermarker LD in disease and nondisease genes. To distinguish between the different LD settings, we introduce a notation $(D'_s, D'_ns)$ where $D'_s$ and $D'_ns$ refer to the intensity of LD in terms of $D'$ values in susceptibility and nonsusceptibility genes, respectively. For example, LD setting (c) with medium LD would be denoted as $(0.5, 0.5)$, and setting (d) with strong LD would be $(1.0, 1.0)$. Uncorrelated markers are described with $(0.0, 0.0)$.

**RESULTS**

**POWER OF THE SET ASSOCIATION APPROACH**

We define power as the proportion of replicates (generated under the disease model) in which a significant set of SNPs comprising at least one disease-associated SNP ($D' \geq 0.5$) is detected. To evaluate the efficiency of jointly detecting disease loci by forming sum statistics, power and the average number of detected disease loci in the sum were computed for 500 replicates. The latter value, together with the total number of selected markers, illustrates how the sums are composed of disease-associated and nonassociated markers. For uncorrelated SNPs, the total number of selected markers was quite stable throughout the simulations (on average, between 5–15), so that we primarily focused on the power and average number of detected disease loci for monitoring the performance of sum statistics.

For an oligogenic threshold model with prevalence $K=0.01$ and overall heritability $h^2=0.5$, results of the power analysis for samples comprising 250 cases and 250 controls are shown in Table I. Table I shows both the power for a significance level of $\alpha=0.05$, and the number of detected susceptibility loci. The probability of finding at least one susceptibility locus is high for 200 and 500 markers. If, however, the number of markers exceeds 1,000, the power drops below 50% for 10 loci. For an increasing number of markers, the decrease in power is more pronounced for 10 than for 5 disease loci. Not unexpectedly, power is better when the trait is produced by fewer disease genes, each with a stronger marginal effect.

For a heritability of 0.5 and 10 susceptibility loci, power is higher when disease genes have uneven genetic contributions (Table I). The trait loci with the higher heritability will have stronger marginal effects and will be included in the sum early in the selection process, which increases the chance of finding a significant set. On the other hand, the disease loci with smaller genetic effects may not be included in the sum. This explains why the average number of detected susceptibility loci is lower when trait loci contribute unequally to the disease, although the power is often higher (Table I).

**TABLE I. Power and average number of detected susceptibility loci for $K=0.01^a$ and $h^2=0.5^b$**

<table>
<thead>
<tr>
<th>No. of markers</th>
<th>5 trait loci$^c$</th>
<th>10 trait loci$^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Equal</td>
<td>Unequal</td>
</tr>
<tr>
<td>Power</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.88</td>
<td>0.84</td>
</tr>
<tr>
<td>500</td>
<td>0.78</td>
<td>0.77</td>
</tr>
<tr>
<td>1,000</td>
<td>0.70</td>
<td>0.76</td>
</tr>
<tr>
<td>2,000</td>
<td>0.67</td>
<td>0.69</td>
</tr>
<tr>
<td>Average number of selected susceptibility loci</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>2.99</td>
<td>2.68</td>
</tr>
<tr>
<td>500</td>
<td>2.77</td>
<td>2.40</td>
</tr>
<tr>
<td>1,000</td>
<td>2.56</td>
<td>2.14</td>
</tr>
<tr>
<td>2,000</td>
<td>2.37</td>
<td>2.00</td>
</tr>
</tbody>
</table>

$^a$Prevalence.

$^b$Heritability.

$^c$Data sets comprised 250 cases and 250 controls. “Equal” and “unequal” refer to equally and unequally contributing disease loci, as described in text.
For 500 uncorrelated markers and 10 trait loci, Figure 1 displays the performance of the set association strategy as a function of sample size. If more than 400 cases and 400 controls are included in the study, the power approaches 100%. Still, the average number of detected disease loci barely exceeds 5. This suggests that even for large sample sizes, marginal effects of some disease mutations are too weak to be detected.

A decrease in heritability or increase in prevalence reduces power considerably (Table II). The dependency of power on the prevalence (threshold) is not surprising. It can be explained in analogy to the dependency of the performance of extreme-proband designs on the selection thresholds in quantitative trait analysis [Abecasis et al., 2001]. In the oligogenic threshold model, when prevalence is high (and also when heritability is low), only a few susceptibility alleles have to be present in an individual for the disease to be expressed. Consequently, we will find only a few disease-associated markers with strong marginal effects. Let \( r_j \) be the proportion of disease loci (out of the total number of disease loci) in the sum, \( s_j \), after \( j \) selection steps. For a low \( h^2 \) or a large \( K \), \( r_j \) will be small.

For different simulation settings in the study, we related power to the average value of \( r_j \). Fitting power and average of \( r_j \) for \( j=(3,5,10) \) in a regression model with logit transformation revealed that the performance of the sum statistics depends to a large extent on the magnitude of \( r_j \) (Fig. 2), regardless of the simulation parameters. The regression curve shows an approximately linear slope for power values between 0.4–0.8. For larger power values, the curve flattens, so that a further increase in \( r_j \) only marginally improves power.

### COMPARISON WITH MARKER-BY-MARKER ASSOCIATION STUDIES

We compared our approach with the false discovery rate (FDR) controlling procedure [Benjamini et al., 2001] and the Bonferroni single-marker test, with a 5% significance level for all. The FDR procedure handles multiple testing problems with varying significance levels, depending on the number of hypotheses already tested. It tends to be more powerful than the Bonferroni procedure, which applies the same significance level to each test. For our approach, as discussed above, power was calculated as the proportion of replicates in which the global \( P \)-value was at most 5% and at least one disease-associated marker was selected. For the other methods, power was similarly defined as the proportion of replicates in which at least one disease-associated marker was selected.

As the results (for uncorrelated markers) in Table II demonstrate, our method outperformed the marker-by-marker strategies for all parameter settings tested. For 200 markers, for example, for a trait with prevalence \( K=0.01 \) and heritability \( h^2=0.5 \), the power was 0.75 with our approach vs. 0.61 and 0.59 for the FDR and Bonferroni procedures, respectively. The average number of disease loci and the total number of markers in the sum were also higher. Analogous results were
obtained for data sets with correlated markers in candidate genes (Fig. 3). The better performance of the sum statistics suggests that they may be successfully applied in data sets where no association can be detected by single-marker tests. Subsequent studies can then be reduced to a small set of markers with statistically significant joint genetic effects.

### ADJUSTING FOR INTERMARKER CORRELATION BY DEFLATING MARKER STATISTICS

Before adjusting marker statistics in the “sum” for intermarker correlation, we studied the effect of LD among markers on the set association approach when no adjustment was carried out.

For this purpose, power was computed for data sets with approximately 504 markers in different intermarker correlation settings (see Methods). For the LD setting with independent markers (0.0, 0.0), we reduced the number of SNPs, since correlation leads to less variability in the marker data. The effect of intermarker association should be evaluated with respect to an “effective” number of uncorrelated markers that represents the same variability as the original marker data with LD. To characterize this number, the 95% quantile of the null distribution of the single-marker statistics of data sets with LD structure (1.0, 1.0) was matched to the 95% quantile of data sets with uncorrelated markers. We found that 63 LD segments, each with 8 SNPs in complete LD (504 markers), correspond to approximately 300 independent markers. Therefore, in the power simulations for independent markers, only 300 SNPs were used, as opposed to 504 SNPs in the simulations with intermarker LD.

#### TABLE II. Set association vs. single-marker approaches\(^a\)

<table>
<thead>
<tr>
<th>Markers</th>
<th>Sum statistic</th>
<th>FDR procedure</th>
<th>Single-marker test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Power</td>
<td>Disease loci</td>
<td>Power</td>
</tr>
<tr>
<td>K=0.01(^b)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.75</td>
<td>4.48</td>
<td>0.61</td>
</tr>
<tr>
<td>500</td>
<td>0.61</td>
<td>3.55</td>
<td>0.47</td>
</tr>
<tr>
<td>1,000</td>
<td>0.50</td>
<td>2.87</td>
<td>0.39</td>
</tr>
<tr>
<td>2,000</td>
<td>0.38</td>
<td>2.67</td>
<td>0.32</td>
</tr>
</tbody>
</table>

\(^a\)Power and average number of detected disease loci for 10 equally contributing disease loci. Data sets comprised 250 cases and 250 controls.

\(^b\)Prevalence.

\(^c\)Heritability.

Fig. 2. Nonlinear regression of power to detect set association vs. proportion \(r_j\) of susceptibility loci in \(j\) most elevated single-marker statistics for several disease model parameters.
In Table III, power results are listed for the different correlation settings (no correlation adjustment). In the disease model, 10 equally contributing disease loci with $h^2=0.5$ and $K=0.01$ were assumed. Whenever markers are tightly linked in susceptibility regions, the performance of our procedure improves considerably. The amount of LD among markers in nonsusceptibility regions is then not so important. For example, in the setting (1.0, 0.0), we found 100% power and 6.06 detected disease loci. Results for the setting (1.0, 1.0) were similar.

However, for the setting (0.0, 1.0), power decreases to 51%, in comparison with 71% when markers are independent. The average number of susceptibility loci per sum is reduced from 4.26 to 2.10. Thus, strong intermarker correlation in candidate regions without a disease mutation is disadvantageous for the performance of the approach.

In the next step, we adjusted marker statistics in the sum for their correlation, to evaluate whether the high power could be maintained in settings (1.0, 1.0) and (1.0, 0.0), and the low power could be improved in the setting (0.0, 1.0). For settings (1.0, 1.0) and (1.0, 0.0), when markers were tightly linked in susceptibility regions, correlation adjustment did not have a strong impact on power and the average number of detected disease loci (Fig. 4). We found that slightly more disease loci were detected when correlation adjustment was carried out. The number of markers comprised by the sum was reduced (from 19.8 to 14.6 for (1.0, 1.0)), while leaving the number of included disease loci constant. This result shows that markers originally included in $s_j$ only due to their correlation to tightly linked markers were not

---

**TABLE III. Impact of intermarker correlation pattern on sum statistics without correlation adjustment**

<table>
<thead>
<tr>
<th>Disease genes</th>
<th>Non-disease genes</th>
<th>$(D'_s , D'_0)$</th>
<th>Power</th>
<th>Detected disease loci in $s_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>No</td>
<td>(0.0, 0.0)</td>
<td>0.71</td>
<td>4.26</td>
</tr>
<tr>
<td>Weak</td>
<td>No</td>
<td>(0.2, 0.0)</td>
<td>0.99</td>
<td>7.48</td>
</tr>
<tr>
<td>Medium</td>
<td>No</td>
<td>(0.5, 0.0)</td>
<td>1.00</td>
<td>7.14</td>
</tr>
<tr>
<td>Strong</td>
<td>No</td>
<td>(1.0, 0.0)</td>
<td>1.00</td>
<td>6.06</td>
</tr>
<tr>
<td>Weak</td>
<td>Weak</td>
<td>(0.2, 0.2)</td>
<td>0.99</td>
<td>7.50</td>
</tr>
<tr>
<td>Medium</td>
<td>Medium</td>
<td>(0.5, 0.5)</td>
<td>0.99</td>
<td>7.16</td>
</tr>
<tr>
<td>Strong</td>
<td>Strong</td>
<td>(1.0, 1.0)</td>
<td>0.99</td>
<td>6.03</td>
</tr>
<tr>
<td>No</td>
<td>Weak</td>
<td>(0, 0.2)</td>
<td>0.60</td>
<td>3.50</td>
</tr>
<tr>
<td>No</td>
<td>Medium</td>
<td>(0, 0.5)</td>
<td>0.57</td>
<td>3.19</td>
</tr>
<tr>
<td>No</td>
<td>Strong</td>
<td>(0, 1.0)</td>
<td>0.51</td>
<td>2.10</td>
</tr>
</tbody>
</table>

*Power and number of detected disease loci in different LD settings $(D'_s \, D'_0)$ for 250 cases and 250 controls.*
selected after deflating their statistics. Reducing the number of selected markers simplifies further modeling procedures and is favorable.

In correlation settings (0.0, 1.0) (bottom row of Fig. 4), taking the correlation pattern among markers into account restores the performance of the approach. After the adjustment, power increased to 71%, and the number of selected disease loci to 4.39. These values are similar to those for uncorrelated markers. Thus, the impact of strong intermarker LD could be completely balanced out by the correlation correction. For medium LD among markers, the effect of correlation adjustment is less pronounced.

**DISCUSSION**

Sum statistics are designed for case-control association studies with potentially large numbers of SNPs. In this study, we evaluated the approach in various simulation settings, and compared it with conventional single-marker tests. Furthermore, we explored whether tightly linked markers (as would be used in a candidate gene study) could be combined efficiently. We found that strong correlation among tightly linked markers from nonsusceptibility genes reduced power. We presented a method that largely restores this loss of power by taking intermarker correlation into account.

For a dichotomous trait of prevalence 0.01, where 5 or 10 susceptibility loci with an overall heritability of 0.5 interact to produce the disease phenotype, our method shows a good performance for data sets with 500 individuals and hundreds of uncorrelated markers. Our method outperforms locus-by-locus approaches, since the strongest marginal genetic effects are combined in one statistic. Falsely selected markers in the sum
affect the efficiency of the approach only when they occur too often. The proportion $r_j$ of susceptibility loci in the $j$ highest single-marker statistics is a good indicator for the power of the algorithm, independent of the underlying disease model. A large proportion $r_j$ leads to high selection power, a small $r_j$ to low power.

The use of tightly linked markers in susceptibility genes has a favorable effect on our approach, which can be attributed to the way information from densely spaced markers is combined. In the marker set, several disease signals from the same susceptibility region can be accumulated. Little informativity of one marker can be balanced out by others. However, the use of tightly linked markers in nonsusceptibility genes is disadvantageous. Power is greatly reduced if nonsusceptibility regions are intensively genotyped with highly correlated SNPs, because these markers will be overly represented in the sums $s_j$.

Since, at the beginning of a study, susceptibility regions are unknown, strongly correlated marker statistics should always be deflated (adjusted) to assure good performance of the sum statistics.

The proposed adjustment for intermarker correlation eliminates the overrepresentation of non-susceptibility regions in the sums $s_j$. Marker statistics are deflated, conditional on the previously selected markers. This strategy proved to balance out the impact of interlocus LD. After adjusting marker statistics in the sum, the performance of the set association approach was mostly as good as with independent markers. It can be applied to various single-marker statistics. It worked equally well for the absolute difference in allele frequencies and the genotype-based $\chi^2$-statistic.

Since permutation sampling is used for the determination of empirical and global $P$-values, a high power in data sets with tightly linked markers does not come at the cost of an inflated type 1 error. Independent of the intermarker correlation, the type 1 error never exceeded the prespecified margin of 0.05. We tested the distribution of global $P$-values for several simulation settings (data not shown). As expected, global $P$-values followed a uniform distribution.

In our simulation, the number of selection steps was limited to $c=20$. In the data set in Hoh et al. [2001] and in our simulation study, the size of selected sets varied between 5–15 for uncorrelated markers, so that $c=20$ is an appropriate choice. If $p_{jmax}$ occurs at $c=20$, this may indicate that many disease signals are comprised in the sum, and $c$ may be chosen larger. However, even with a larger value for $c$, the global $P$-value does not necessarily decrease, since more association tests are carried out. For example, we also carried out simulations with $c=50$ and found a reduced performance of our approach compared with $c=20$. Note that tightly linked markers in disease genes increase the total number of selected markers, an effect that can be reversed if correlation adjustment is carried out.

The choice of single-marker statistics will also have an impact on our approach. When genotypic effects are composed of the sum of the allelic components, statistics based on allele frequencies will be an appropriate measurement for single-marker effects. If there is allelic interaction at a susceptibility locus, genotype-based $\chi^2$-values may provide better power. A combination of different statistics measuring different genetic phenomena can also facilitate the detection of set association. Here, we did not address the question of the optimum single-marker statistic, because we focused on the interplay of single-marker statistics in the sum $s_j$ on the genotypic level.

APPENDIX

Let $t_{(1)},...,t_{(j)}$ be the marker statistics in the sum, $s_j$, after the $j$th selection step. For all remaining markers $i$ with statistics $t_i$, we form the deflated statistics $t'_i$. The maximum deflated statistic $\max t'_i$ is next to be included in the sum constituting $s_{j+1}$.

For the computation of $t'_i$, let $V$ with elements $V_{ij}$ denote the covariance matrix of the marker statistics $t_{(1)},...,t_{(j)}$, and $t_i$

$$V = \left( \begin{array}{cccc}
\sigma^2(t_{(1)}) & \sigma(t_{(1)}, t_{(2)}) & \cdots & \sigma(t_{(1)}, t_{(j)}) & \sigma(t_{(1)}, t_i) \\
\sigma(t_{(2)}, t_{(1)}) & \sigma^2(t_{(2)}) & \cdots & \sigma(t_{(2)}, t_{(j)}) & \sigma(t_{(2)}, t_i) \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\sigma(t_{(j)}, t_{(1)}) & \sigma(t_{(j)}, t_{(2)}) & \cdots & \sigma^2(t_{(j)}) & \sigma(t_{(j)}, t_i) \\
\sigma(t_i, t_{(1)}) & \sigma(t_i, t_{(2)}) & \cdots & \sigma(t_i, t_{(j)}) & \sigma^2(t_i)
\end{array} \right),$$

and let $V^{-1}$ be its inverse with elements $V_{ij}^{-1}$. $\mu_{(1)},...\mu_{(j)}$ and $\mu_i$ refer to the mean, and $\sigma_{(1)},...\sigma_{(j)}$ and $\sigma_i$ refer to the standard deviation of the respective marker statistics $t_{(1)},...,t_{(j)}$ and $t_i$. Means, standard deviations, and the elements of $V$ are estimated via permutation sampling (affection status is permuted), which is carried out independently from ordering and selecting markers in the sum statistics.

Under the assumption of asymptotic normality of the marker statistics, the conditional
of the single-marker statistics in these cases, Equation (1) provides a transformation of normality will not be valid. However, even only the portion of included marker is uncorrelated to all the marker 

\[ D(t_i | t_1, \ldots, t_j) = \sqrt{V_{ii}^{-1}} \left( t_i - \mu_i + \sum_{k=1}^{j} \frac{V_{ik}^{-1}}{V_{ii}^{-1}} (t_k - \mu_k) \right). \]

Rescaling this normal deviate to adjust for the mean and variance of \( t_i \) leads to the deflated value \( t_i^* \) of \( t_i \)

\[ t_i^* = \mu_i + \sigma_i \sqrt{V_{ii}^{-1}} \left( t_i - \mu_i + \sum_{k=1}^{j} \frac{V_{ik}^{-1}}{V_{ii}^{-1}} (t_k - \mu_k) \right). \]

(1)

For many single-markers statistics, the assumption of normality will not be valid. However, even in these cases, Equation (1) provides a transformation of the single-marker statistics \( t_i \) in \( t_i^* \) that maintains the mean and variance of \( t_i \) and eliminates the correlation between \( t_i^* \) and the single-marker statistics \( t_1, \ldots, t_j \) already in the sum. Proofs for \( E(t_i^*) = \mu_i \) and \( \sigma(t_i^*) = \sigma_i \) are straightforward. The proof for \( \sigma(t_i^*, t_j) = 0 \) for \( l=1, \ldots, j \) is shown below:

\[
\sigma(t_i^*, t_j) = E((t_i^* - \mu_i)(t_j - \mu_j))
\]

\[
= \frac{\sigma_i}{\sqrt{V_{ii}^{-1}}} \cdot E \left( V_{ii}^{-1} (t_i - \mu_i)(t_j - \mu_j) \right) + \sum_{k=1}^{j} \frac{V_{ik}^{-1} (t_k - \mu_k)(t_j - \mu_j)}{V_{ii}^{-1}}
\]

\[
= \frac{\sigma_i}{\sqrt{V_{ii}^{-1}}} \cdot \sum_{k=1}^{j+1} \frac{V_{ik}^{-1} V_{kl}}{V_{ii}^{-1}} \delta_{ij}
\]

\[
= 0 \quad \text{for} \quad l = 1, \ldots, j.
\]

The adjusted statistic \( t_i^* \) of the newly-to-be-included marker is uncorrelated to all the marker statistics already in the sum. Independent from the distribution of the single-marker statistics, only the portion of \( t_i \) that is not yet accounted for by the markers \( (1), \ldots, (j) \) is included in the sum statistics. Therefore, we can expect the correlation adjustment to provide good results, even when the assumption of normality is not met.

The transformation from \( t_i \) to \( t_i^* \) resembles a principal component approach. However, our stepwise SNP selection procedure requires us to decompose \( V^{-1} \) into triangular matrices instead of orthogonal matrices.

REFERENCES


