BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies

Xiang Wan,1,6 Can Yang,1,6 Qiang Yang,2 Hong Xue,3 Xiaodan Fan,4 Nelson L.S. Tang,5 and Weichuan Yu1,*

Gene-gene interactions have long been recognized to be fundamentally important for understanding genetic causes of complex disease traits. At present, identifying gene-gene interactions from genome-wide case-control studies is computationally and methodologically challenging. In this paper, we introduce a simple but powerful method, named “BOolean Operation-based Screening and Testing” (BOOST). For the discovery of unknown gene-gene interactions that underlie complex diseases, BOOST allows examination of all pairwise interactions in genome-wide case-control studies in a remarkably fast manner. We have carried out interaction analyses on seven data sets from the Wellcome Trust Case Control Consortium (WTCCC). Each analysis took less than 60 hr to completely evaluate all pairs of roughly 360,000 SNPs on a standard 3.0 GHz desktop with 4G memory running the Windows XP system. The interaction patterns identified from the type 1 diabetes data set display significant difference from those identified from the rheumatoid arthritis data set, although both data sets share a very similar hit region in the WTCCC report. BOOST has also identified some disease-associated interactions between genes in the major histocompatibility complex region in the type 1 diabetes data set. We believe that our method can serve as a computationally and statistically useful tool in the coming era of large-scale interaction mapping in genome-wide case-control studies.

Introduction

Genome-wide case-control studies use high-throughput genotyping technologies to assay hundreds of thousands of SNPs and relate them to clinical conditions or measurable traits. To understand underlying causes of complex disease traits, it is often necessary to consider joint genetic effects (epistasis) across the whole genome. The concept of epistasis1 was introduced around 100 years ago. It is generally defined as interactions among different genes. Recently, Phillips2 highlighted the essential role of gene-gene interactions in the structure and evolution of genetic systems. Three terminologies are used to describe gene-gene interactions:

- Functional epistasis is a functional description that addresses the molecular interactions.
- Compositional epistasis, originally defined by Bateson,1 is referred to as the blocking of one allelic effect by another allele at a different locus.
- Statistical epistasis, attributed to Fisher,3 is defined as the statistical deviation from the additive effects of two loci on the phenotype.

The existence of epistasis has been widely accepted as an important contributor to genetic variation in complex diseases such as asthma, cancer, diabetes, hypertension, and obesity.4 As a matter of fact, many researchers believe that it is critical to model complex interactions in order to elucidate the joint genetic effects that may cause complex diseases. They have demonstrated the presence of gene-gene interactions in complex diseases such as breast cancer5 and coronary heart disease.6 The problem of detecting gene-gene interactions in genome-wide case-control studies has attracted extensive research interest. The difficulty in this problem is the heavy computational burden. For example, in order to detect pairwise interactions from 500,000 SNPs genotyped in thousands of samples, we need $1.25 \times 10^{11}$ statistical tests in total. A recent review4 presented a detailed analysis on many popular methods that detect epistasis on the basis of the statistical definition, including MDR,5 PLINK,7 Tuning Relief,8 Random Jungle,9 BEAM,10 and three proposed search strategies.11

Among them, BEAM and MDR were reported to have difficulties in handling 500,000 SNPs genotyped in thousands of samples.4 Both methods need a prescreening process to reduce the number of SNPs in order to analyze the large data sets. Marchini et al.11 demonstrated that it is feasible to test association allowing for interactions in a genome-wide scale. Random Jungle can handle genome-wide data efficiently. However, both Marchini’s method and Random Jungle aim at testing associations allowing for interactions, which is easier than testing interactions (we have detailed explanations of a test of association allowing for interactions and a test of interactions in

1Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China; 2Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China; 3Department of Biochemistry, The Hong Kong University of Science and Technology, Hong Kong, China; 4Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China; 5Laboratory for Genetics of Disease Susceptibility, Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China
6These authors contributed equally to this work
*Correspondence: eeyu@ust.hk

DOI 10.1016/j.ajhg.2010.07.021. ©2010 by The American Society of Human Genetics. All rights reserved.
Definition of Interaction via Logistic Regression Models
Interactions are often defined via logistic regression models. The logistic regression model with only main effects, i.e., the main effect model, has the following form:

\[
\log \left( \frac{P(Y = 1 | X_p = i, X_q = j)}{P(Y = 2 | X_p = i, X_q = j)} \right) = \beta_0 + \beta_{i}X_p + \beta_{j}X_q
\] (Equation 1)

The logistic regression model with both main effect terms and interaction terms, i.e., the full model, has the following form:

\[
\log \left( \frac{P(Y = 1 | X_p = i, X_q = j)}{P(Y = 2 | X_p = i, X_q = j)} \right) = \beta_0 + \beta_{i}X_p + \beta_{j}X_q + \beta_{ij}X_pX_q
\] (Equation 2)

Material and Methods
Notation
Suppose we have \(L\) SNPs and \(n\) samples. We use \(X_i\) to denote the \(i\)-th SNP, \(i = 1, \ldots, L\), and \(Y\) to denote the class label (1 for case and 2 for control). SNPs are biallelic genetic markers in genome-wide case-control studies. In general, we use capital letters (e.g., \(A, B, \ldots\)) to denote major alleles and use lowercase letters (e.g., \(a, b, \ldots\)) to denote minor alleles. For each SNP, there are three genotypes: the homozygous reference genotype \((AA)\), the heterozygous genotype \((Aa)\), and the homozygous variant genotype \((aa)\). The popular way of coding the genotype data is to use \(1\), \(2\), \(3\) to represent \(AA\), \(Aa\), \(aa\), respectively.

Definition of Interaction via Logistic Regression Models
PLINK was recommended as the most computationally feasible method that is able to detect gene-gene interactions in genome-wide data. PLINK finished a pairwise interaction examination of 89,294 SNPs selected from the WTCCC Crohn disease data set in 14 days. To accelerate the analysis process in genome-wide association studies (GWAS), the parallel computation was recommended.

Here, we propose a fast method, named “BOolean Operation-based Screening and Testing” (BOOST), for the analysis of all pairwise interactions in genome-wide SNP data. In our method, we design a Boolean representation of genotype data, which promotes not only space efficiency but also CPU efficiency because it involves only Boolean values and allows for the use of fast logic (bitwise) operations to obtain contingency tables. On the basis of this data representation, we propose a two-stage (screening and testing) search method. In the screening stage, we use a noniterative method to approximate the likelihood ratio statistic in evaluating all pairs of SNPs and select those passing a specified threshold. Most nonsignificant interactions will be filtered out, and the survival of significant interactions is guaranteed. In the testing stage, we employ the classical likelihood ratio test to measure the interaction effects of selected SNP pairs. Experiments on WTCCC data sets show that our method is faster than current methods. This efficiency helps to identify interesting interaction patterns from the type 1 diabetes data set and the rheumatoid arthritis data set.

Log-Linear Models for Contingency Tables
To test the interaction effect between two SNPs \((X_p, X_q)\) and disease status \(Y\) by using log-linear models, a contingency table of these three variables will be used (see Table 1). The size of the contingency table is \(I \times J \times K\), where \(I = 3\), \(J = 3\) and \(K = 2\). In Table 1, \(n_{ijk}\) is used to denote the observed count in the cell \((i, j, k)\). It is considered as a realization of a random variable \(N_{ijk}\) assumed as Poisson distributed. We use \(\pi_{ijk}\) to denote the probability that an observation falls in the cell \((i, j, k)\). A natural constraint of \(\pi_{ijk}\) is

\[
\sum_{ijk} \pi_{ijk} = 1
\] (Equation 3)

We use the dot convention to indicate summation over a subscript; e.g., \(n_i = \sum_{jk} \pi_{ijk}\) is the marginal probability of \(X_p = i\), and \(n_i = \sum_{jk} n_{ijk}\) is the number of observations with \(X_p = i\). The notation extends to two dimensions as well. For example, \(\pi_{ik} = \sum_{j} \pi_{ijk}\) is the marginal probability of \(X_p = i\) and \(X_q = j\), and \(n_i = \sum_{jk} n_{ijk}\) is the corresponding count. Clearly, we have

\[
n = \sum_{ijk} n_{ijk}
\]

Log-linear models treat \(N_{ijk}\) as independent Poisson random variables with their means as follows:

\[
\mu_{ijk} = n \pi_{ijk}
\] (Equation 4)
The likelihood function is
\[ f(\mu) = \prod_{i,j,k} \frac{e^{-\mu_{ijk} n_{ijk}}}{n_{ijk}!} \]  
(Equation 5)

Correspondingly, the log-likelihood function is
\[ L(\mu) = \sum_{i,j,k} [n_{ijk} \log(\mu_{ijk}) - \mu_{ijk} - \log(n_{ijk}!)] \]  
(Equation 6)

In the space of log-linear models, the homogeneous association model is the equivalent form of the logistic regression model with only main effects (defined in Equation 1), and the saturated model matches the full logistic regression model (defined in Equation 2). Table 2 summarizes the equivalence between log-linear models and logistic models for a three-way contingency table. The details are provided in the Appendix. In the following text, we explain how these two models are used to test interactions.

### Measuring Interaction via Log-Linear Models

On the basis of the equivalence between the log-linear model and its corresponding logistic regression model, we construct our test statistic using the homogeneous association model \( M_H \) and the saturated model \( M_S \). Let \( L_H \) and \( L_S \) be the log-likelihood of \( M_H \) and \( M_S \), respectively. According to Equation 6 and the MLE of \( \mu_{ijk} \) in \( M_S \) (see Table 2 and the Appendix), the maximum log-likelihood of \( M_S \) is
\[ \hat{L}_S = \sum_{i,j,k} [n_{ijk} \log(n_{ijk}) - n_{ijk} - \log(n_{ijk}!)] \]  
(Equation 7)

The log-likelihood of \( M_H \) is maximized at its MLE \( \hat{\mu}_{ijk}^H \):
\[ \hat{\mu}_{ijk}^H = \arg \max_{\mu_{ijk}} L_H = \arg \max_{\mu_{ijk}} \sum_{i,j,k} [n_{ijk} \log\mu_{ijk} - \mu_{ijk} - \log(n_{ijk}!)] \]  
(Equation 8)

In other words,
\[ \hat{L}_H = L_H(\hat{\mu}_{ijk}) = \max_{\mu_{ijk}} L_H(\mu_{ijk}) \]  
(Equation 9)

Notice that \( \hat{\mu}_{ijk}^H \) always exists and is unique because of the concavity of \( L_H \). To measure interaction effects based on the likelihood ratio test, we have
\[ \hat{L}_S - \hat{L}_H = \sum_{i,j,k} [n_{ijk} \log(\frac{n_{ijk}}{\hat{\mu}_{ijk}^H}) - n_{ijk} + \hat{\mu}_{ijk}^H] \]  
(Equation 10)

Because Equation 4 implies that
\[ \sum_{i,j,k} \hat{\mu}_{ijk}^H = n \]  
(Equation 11)

**Table 2. Equivalence between Log-Linear Models and Logistic Models for a Three-Way Table with Binary Response Variable \( Y \)**

<table>
<thead>
<tr>
<th>Log-Linear Model</th>
<th>Logistic Model</th>
<th>MLE of ( \mu_{ijk} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block independence model (( M_0 )): ( \log\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z )</td>
<td>( \log \pi_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z )</td>
<td>Iterative estimation</td>
</tr>
<tr>
<td>Partial independence model (( M_1 )): ( \log\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_i^XY + \lambda_j^YZ + \lambda_k^XZ )</td>
<td>( \log \pi_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_i^XY + \lambda_j^YZ + \lambda_k^XZ )</td>
<td>Iterative estimation</td>
</tr>
<tr>
<td>Homogeneous association model (( M_H )): ( \log\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_i^XY + \lambda_j^YZ + \lambda_k^XZ + \lambda_i^XZ )</td>
<td>( \log \pi_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_i^XY + \lambda_j^YZ + \lambda_k^XZ + \lambda_i^XZ )</td>
<td>Iterative estimation</td>
</tr>
</tbody>
</table>

The models \( M_S \) and \( M_B \) are used in the discussion of the difference between the test of interactions and the test of associations. The details of these two models are provided in the Appendix.

**Equation 10** can be further reduced as
\[ \hat{L}_S - \hat{L}_H = \sum_{i,j,k} \frac{n_{ijk} \log(n_{ijk})}{\hat{\mu}_{ijk}^H} \]  
(Equation 12)

\[ = n \sum_{i,j,k} \frac{n_{ijk} \log(n_{ijk}/n)}{\hat{\mu}_{ijk}^H} \]  
(Equation 12)

\[ = n \sum_{i,j,k} \frac{\pi_{ijk} \log(\frac{\pi_{ijk}}{\hat{\pi}_{ijk}})}{\hat{\mu}_{ijk}^H} \]  
(Equation 12)

\[ = n \cdot D_{KL}(\hat{\pi}_{ijk} \| \hat{\pi}_{ijk}) \]  
(Equation 12)

where \( D_{KL}(\hat{\pi}_{ijk} \| \hat{\pi}_{ijk}) \) is the Kullback-Leibler divergence of \( \hat{\pi}_{ijk} \) and \( \hat{\pi}_{ijk} \).

The new measure \( D_{KL}(\hat{\pi}_{ijk} \| \hat{\pi}_{ijk}) \) provides us another interpretation of interactions. Equation 12 shows that the difference of the two log-likelihoods is proportional to the Kullback-Leibler divergence of the joint distribution \( \hat{\pi}_{ijk} \) obtained under the saturated model \( M_S \) and the distribution \( \hat{\pi}_{ijk} \) obtained under the homogeneous association model \( M_H \). The distribution \( \hat{\pi}_{ijk} \) is constructed via lower-order distributions (see the Appendix). From the perspective of log-linear models, interaction effects can be understood as the information contained in the joint distribution but not in its lower-order factorization, which is known as “synergy” in physics. If no interaction effects exist, the joint distribution can be well characterized by its lower-order factorization.

**Boolean Operation-Based Screening and Testing**

**Boolean Representation of Genotype Data**

The data set containing \( L \) SNPs and \( n \) samples is usually stored in an \( L \times n \) matrix. Each cell in this matrix takes a value from \{1, 2, 3\}, the elements of which represent the homozygous reference genotype, the heterozygous genotype, and the homozygous variant genotype, respectively. In our method, we introduce a Boolean representation of genotype data (the details are provided in the Appendix). This Boolean representation enables us to collect contingency tables in a fast manner.

**Screening and Testing**

Directly using \( \hat{L}_S - \hat{L}_H \) to test interactions in GWAS still has some difficulties, because no closed-form solution exists for the homogenous association model \( M_H \). Iterative methods are needed in model fitting to compute \( \hat{L}_H \). This will be computationally intensive when we face hundreds of billions of SNP pairs.

To solve this issue, we propose to approximate the homogeneous association model \( M_H \) with the Kirkwood superposition approximation (KSA):\(^{15}\)
\[ \hat{\pi}_{ijk} = \frac{1}{\eta} \pi_i \pi_j \pi_k \]  
(Equation 13)
Comparison of the values $2(L_S - L_{KSA})$ and $2(L_S - L_H)$ based on KSA and log-linear models. KSA overestimation $2(L_S - L_H) \leq 2(L_S - L_{KSA})$ is illustrated here. For the region $[25, + \infty)$, $2(L_S - L_{KSA})$ is almost identical to $2(L_S - L_H)$.

Figure 1. KSA Performance in Simulation
(A) The LD (measured by $r^2$) pattern of simulated data from the Hapmap data. To show the block structure clearly, we show only the LD of the first 500 SNPs here. The LD block structure of all 2000 SNPs is very similar.

(B) Comparison of the values $2(L_S - L_{KSA})$ and $2(L_S - L_H)$ based on KSA and log-linear models. KSA overestimation $2(L_S - L_H) \leq 2(L_S - L_{KSA})$ is illustrated here. For the region $[25, + \infty)$, $2(L_S - L_{KSA})$ is almost identical to $2(L_S - L_H)$.

where $\eta = \sum_{ijk} \frac{p_{ijk}^K}{n_{ijk}^C} / \pi_{ijk}$ is a normalization term. The benefit of using KSA is two-fold:

First, $L_S - L_{KSA}$ is an upper bound of $L_S - L_H$; i.e.,

$$L_S - L_H \leq L_S - L_{KSA} \quad \text{(Equation 14)}$$

where $L_{KSA}$ is the log-likelihood evaluated at the MLE $\hat{p}_{ijk}^K$ of the KSA model (see the proof in the Appendix).

Noticing that the calculation of $p_{ijk}^K$ is straightforward and no iteration is involved, the approximated measure $2(L_S - L_{KSA}) = 2n_rD_{KL}(\hat{p}_{ijk}^K || \hat{p}_{ijk}^L)$ can be obtained easily on the basis of the contingency table collected by the Boolean operation. Therefore, the KSA model can be applied to evaluate hundreds of billions of SNP pairs. Because we are interested only in interactions with large interaction effect using the likelihood ratio statistic $2(L_S - L_H)$, we first filter out those SNP pairs with $2(L_S - L_{KSA}) \leq \tau$ by using a threshold $\tau$, and we can then conduct statistical tests on the remaining SNP pairs.

Second, the bound in Equation 14 is tight. When the joint distribution is $p_{ijk}^K$ (Equation 13), the equality holds; i.e., $L_S - L_{KSA} = L_S - L_H$. This bound is very close to the statistic $L_S - L_H$ of the likelihood ratio test. To illustrate the tightness of the bound, we use the simulation method proposed by Li et al.\(^\text{16}\) to generate a data set containing 2000 SNPs and 1000 samples based on HapMap data. Figure 1A shows the linkage disequilibrium (LD) pattern of the simulated data, which is very similar to the real data. Using this data, we calculate $2(L_S - L_{KSA}) = 2n_rD_{KL}(\hat{p}_{ijk}^K || \hat{p}_{ijk}^L)$ based on the KSA and $2(L_S - L_H) = 2n_rD_{KL}(\hat{p}_{ijk}^L || \hat{p}_{ijk}^L)$ based on log-linear models for all pairs of 2000 SNPs. Figure 1B shows the comparison of these two models. It can be seen that $2(L_S - L_{KSA})$ consistently overestimates $2(L_S - L_H)$. For the region $[25, + \infty)$, $2(L_S - L_{KSA})$ is almost identical to $2(L_S - L_H)$.

In summary, most nonsignificant interactions can be filtered out because of the tightness of the bound (Equation 14) and the survival of significant interactions is guaranteed. On the basis of this upper bound, we propose our method, BOOST:

Stage 1: Screening. We evaluate all pairwise interactions by using the KSA in the screening stage. For each pair, the calculation of $2(L_S - L_{KSA})$ is based on the contingency table collected by using Boolean operations. Because $2(L_S - L_H) \leq 2(L_S - L_{KSA})$, an interaction obtained by the KSA without passing a specified threshold $\tau$, i.e., $2(L_S - L_{KSA}) \leq \tau$, would not be considered in stage 2. The threshold $\tau$ corresponds to the significant threshold (with the Bonferroni correction) specified by users. Because the Bonferroni correction tends to be conservative, a smaller threshold can be used to put more SNP pairs into the testing stage. We set $\tau = 30$ in our experiments to test the computational capacity of our method. The threshold $\tau = 30$ corresponds to the unadjusted $p = 4.89 \times 10^{-6}$, which is a very weak significance level for a genome-wide study.

Stage 2: Testing. For each pair with $2(L_S - L_{KSA}) > \tau$, we test the interaction effect using the likelihood ratio statistic $2(L_S - L_H)$. We fit the log-linear models $M_0$ and $M_1$ and calculate this test statistic using Equation 12. After that, we conduct the $\chi^2$ test with four degrees of freedom ($df = 4$) to determine whether the interaction effect is significant. The $p$ value is adjusted by the Bonferroni correction, with the number of tests $L(L - 1)/2$, where $L$ is the total number of SNPs before screening.

To approximate $M_{1b}$, we may also choose some other log-linear models, such as the block independence model $M_0$ or the partial independence model $M_0$ (see Table 2). However, such approximations will lead to very loose bounds, leaving millions of SNP pairs to be examined in the testing stage. Using the KSA, we have empirically observed that 300,000–600,000 SNP pairs are examined in the testing stage when the WTCCC data are analyzed. When the partial independence model is used, the number of SNP pairs is up to $10^8$–$10^9$.

Results

Experiments on Simulation Data

The performance of our approach is evaluated through comparative studies with existing works. Our goal is to
discover epistatic interactions from genome-wide data. Among many methods recently proposed, we mainly compare BOOST with PLINK\(^2\) with respect to the power of gene-gene interaction identification. The reasons for choosing PLINK for comparison are as follows:

- A recent review\(^4\) tested many available methods and recommended PLINK as a powerful tool for testing interactions on a genome-wide scale.
- Both PLINK and BOOST use an exhaustive search strategy. The comparison of their performance is fair.

We conduct the following simulation studies to compare BOOST with PLINK (tested with the “fast-epistasis” option and without the “case-only” option):

- Case 1: Disease loci with main effects.
- Case 2: Disease loci without main effects.
- Case 3: Genetic heterogeneity.
- Case 4: Null simulation for testing type I errors.

Case 1: Disease Loci with Main Effects

We consider four epistasis models whose odds tables are given in Table S7, available online. Model 1 is a multiplicative model.\(^1\) Model 2 is an epistasis model\(^1\) that has been used to describe handedness\(^18\) and the color of swine.\(^19\) Model 3 is a classical epistasis model.\(^20,21\) Model 4 is the well known XOR (exclusive OR) model.

Let \(p(D|G_i)\) denote the probability of an individual being affected given its genotype combination \(G_i\) (i.e., the penetrance of \(G_i\)), and let \(p(\overline{D}|G_i)\) denote the probability of an individual not being affected given its genotype \(G_i\). On the basis of the definition of the odds of a disease,

\[
\text{ODD}_{G_i} = \frac{p(D|G_i)}{p(\overline{D}|G_i)} = \frac{p(D|G_i)}{1 - p(D|G_i)} \quad \text{(Equation 15)}
\]

the penetrance \(p(D|G_i)\) of the genotype \(G_i\) can be calculated by using

\[
p(D|G_i) = \frac{\text{ODD}_{G_i}}{1 + \text{ODD}_{G_i}} \quad \text{(Equation 16)}
\]

The disease prevalence \(p(D)\) and genetic heritability \(h^2\) are given as

\[
p(D) = \sum_i p(D|G_i)p(G_i) \quad \text{(Equation 17)}
\]

\[
h^2 = \frac{\sum_i(p(D|G_i) - p(D))^2 p(G_i)}{p(D)(1 - p(D))} \quad \text{(Equation 18)}
\]

In our simulation, the prevalence \(p(D)\) and the heritability \(h^2\) are controlled by the parameters \(\alpha\) and \(\theta\) (see Table S6). We first specify the disease prevalence \(p(D)\) and the genetic heritability \(h^2\), and we then numerically solve the parameters (\(\alpha\) and \(\theta\)) on the basis of the above equations. For example, we set \(p(D) = 0.1\) and \(h^2 = 0.03\) in model 1. Then we obtain \(\alpha = 0.09989\) and \(\theta = 3.4481\) for minor allele frequency (MAF) = 0.1.

In the simulation, we set \(h^2 = 0.03\) for model 1 and \(h^2 = 0.02\) for models 2, 3, and 4. We generate genotype data on the basis of the Hardy-Weinberg principle. We set the MAFs of disease-associated SNPs to be 0.1, 0.2, and 0.4. We generate the MAFs of unassociated SNPs uniformly from \([0.05, 0.5]\). We simulate 100 data sets under each setting for each disease model. Each data set contains 1000 SNPs. To take sample size into consideration, we simulate both 800 samples and 1600 samples with the balanced design.

Figure 2 presents the comparison results with the significance thresholds selected as 0.1, 0.2, and 0.3 after the Bonferroni correction. For model 1 with MAF = 0.2, 0.4 and model 2 with MAF = 0.1, the statistical power of PLINK is higher. This is because these model settings are well captured by the allele interaction test. For all other settings, BOOST outperforms PLINK.

Case 2: Disease Loci without Main Effects

Disease models displaying no main effects\(^22\) have been carefully discussed, and a wide spectrum of these models\(^23\) has been provided. In this experiment, we use all of these 70 pure epistatic models without main effect to compare performance. For convenience, these models are listed in Tables S8–S14. The heritability \(h^2\) controls the phenotypic variation of these 70 models, which ranges from 0.01 to 0.4. The MAF ranges from 0.2 to 0.4. For each model, the statistical power is evaluated under different sample sizes, including \(n = 400, n = 800,\) and \(n = 1,600\) (half controls and half cases). For each setting, 100 data sets are generated. Each data set contains 1000 SNPs.

Please check Figures S4–S7 to see the comparison results for the 70 models. For some models, such as model ep1–5, BOOST and PLINK perform equally well. For most of these models, BOOST is superior to PLINK because the interaction patterns cannot be well characterized by allele interactions.

Case 3: Genetic Heterogeneity

Genetic heterogeneity refers to the phenomenon that a disease is affected by different subsets of genes. It plays a substantial role in complex human diseases.\(^24\) Here, we set up a simulation study to show the performance of BOOST and PLINK when genetic heterogeneity is present. We choose some epistatic models used in case 2 to generate the data. The heritability \(h^2\) of these models ranges from 0.01 to 0.4. Different sample sizes, including \(n = 400, n = 800,\) and \(n = 1600\), are simulated for each model. The details of simulation are provided in the Appendix.

The performance of both BOOST and PLINK is given in Figure S8. Genetic heterogeneity affects the performance of both BOOST and PLINK. In general, their performance degrades as heritability \(h^2\) decreases. The sample size plays an important role when genetic heterogeneity is present. When the sample size increases from 400 to 1600, the power of both BOOST and PLINK increases a lot.

Case 4: Null Simulation for Testing Type I Errors
To compare BOOST and PLINK in terms of type I errors, we conduct null simulation in two scenarios:

- Scenario 1: Without LD. We generate 1000 null data sets. Each data set contains 1000 SNPs and 1000 samples. All of the SNPs are generated independently, with MAFs uniformly distributed in [0.05, 0.5]. The result is shown in Figure 3A. It can be seen that the type I error of BOOST agrees with the nominal error rate and the type I error of PLINK is a little bit less than the nominal error rate.

- Scenario 2: With LD. The simulation program "genomeSIMLA" is used to simulate the SNP data on the basis of the marker information on the Affymetrix 500K chip from human chromosome 1. LD exists among SNPs. We generate 100 null data sets, each of which contains 38,836 SNPs and 1000 samples. The result is shown in Figure 3B. Because of the LD pattern, the error rates of both methods are lower than the nominal error rate, confirming that the Bonferroni correction is conservative. Surprisingly, unlike the situation in scenario 1, the error rate of BOOST is less than that of PLINK. The reason is that some cells of a contingency table may be empty when LD exists. This leads to the true degree of freedom $d_{true} \leq 4$.

Because we calculate p values by using the $\chi^2$ distribution with $df = 4$, BOOST has a lower type I error rate than PLINK. This simulation study also implies that it is possible to increase the power of BOOST by using a more accurate degree of freedom in statistical tests.

Experiments on WTCCC data
We have applied BOOST to analyze data (14,000 cases in total and 3000 shared controls) from the WTCCC on seven common human diseases: bipolar disorder (BD), coronary artery disease (CAD), Crohn disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D). The procedure of quality control is presented in the Appendix. The results under different constraints are reported in Table 3. For T1D, we discovered many gene-gene interactions in the MHC region (see detailed descriptions in the following section). For the other six diseases, however, we did not find nontrivial interactions (except one SNP pair in CD).

T1D and RA
The MHC region in chromosome 6 has long been investigated as the most variable region in the human genome with respect to infection, inflammation, autoimmunity, and transplant medicine. The recent study conducted by the WTCCC has shown that both T1D and RA are
strongly associated with the MHC region via single-locus association mapping. The top-left panel of Figure 4 shows that the single-locus association map does not reveal much difference between T1D and RA. In our study, BOOST reports 4499 interactions in the T1D data set (see Table 3), in which 4489 interactions (99.8%) are in the MHC region. Clayton’s analysis on the T1D data set found that with the exception of strong interactions within the MHC region, interactions are small and have a modest effect on prediction. Our results have verified Clayton’s finding from another perspective. As a comparison, BOOST reports 350 interactions in the RA data set, in which 280 interactions (80.0%) are in the MHC region. Our genome-wide interaction map provides evidence that the MHC region is associated with these two diseases in different ways. The bottom panel of Figure 4 gives detailed interaction maps in the MHC region for T1D and RA data. We further calculate composite LD using the method by Zaykin et al. The LD map of MHC region is provided in the top-right panel of Figure 4. These interaction maps, different from the LD map, reveal a distinct pattern difference between T1D and RA. Specifically, there are three subregions in the MHC region: namely, the MHC class I region (29.8Mb–31.6Mb), the MHC class III region (31.6Mb–

Table 3. The Number of Interactions Identified from the WTCCC Data Sets of Seven Diseases under Different Constraints

<table>
<thead>
<tr>
<th></th>
<th>BD</th>
<th>CAD</th>
<th>CD</th>
<th>HT</th>
<th>RA</th>
<th>T1D</th>
<th>T2D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C^1</td>
<td>10</td>
<td>16</td>
<td>8</td>
<td>7</td>
<td>350</td>
<td>4499</td>
<td>18</td>
</tr>
<tr>
<td>C^1 &amp; C^2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>789</td>
<td>0</td>
</tr>
<tr>
<td>C^1 &amp; C^2 &amp; C^3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>91</td>
<td>0</td>
</tr>
</tbody>
</table>

Abbreviations are as follows: BD, bipolar disorder; CAD, coronary artery disease; CD, Crohn disease; HT, hypertension; RA, rheumatoid arthritis; T1D, type 1 diabetes; T2D, type 2 diabetes. C^1 is the significance threshold constraint; the significance threshold is 0.05 for the Bonferroni-corrected interaction p-value. C^2 is the distance constraint: the physical distance between two interacting SNPs is at least 1Mb. This constraint is used to avoid interactions that might be attributed to the LD effects. C^3 is the main effect constraint: the single-locus p-value should not be less than 10^-6. This constraint is used to see whether there exist strong interactions without significant main effects, because those SNPs with p ≥ 10^-6 are usually filtered out in the typical single-locus scan.
effects between these two classes may contribute to the etiology of type 1 diabetes.

Discussion

Relationship between Our Method and Other Two-Stage Methods

The analysis of GWAS data is a challenging computational problem. To speed up this process, many methods have been coupled with some prescreening algorithms to reduce the number of SNPs. Most of the currently available screening algorithms are based on single-locus tests and can be finished very quickly. However, for some SNPs with weak main effects but significant interactions, these screening algorithms will filter them out. Our screening method does not have this issue. It uses a fast approximation to evaluate all SNP pairs with the guarantee that significant interactions will not be filtered out no matter whether individual SNPs display main effects or not.

PLINK tests interactions based on alleles. Three genotype categories are collapsed into two allele categories. Correspondingly, $3 \times 3$ contingency tables are collapsed into $2 \times 2$ tables. The difference of the odds ratios from the two $2 \times 2$ tables (one for cases and the other for controls) is used to construct a $\chi^2$ test with $df = 1$.

BOOST tests interactions based on genotypes, using the $\chi^2$ test with $df = 4$.

In general, if the underlying interaction could be well characterized by an allele interaction, then the statistical power of PLINK would be higher than that of BOOST.

Figure 4. Comparison between the Single-Locus Association Mapping and the Interaction Mapping for T1D and RA

Top-left panel: Single-locus association mapping of T1D and RA. These two share a very similar hit region in chromosome 6.

Top-right panel: The LD map of the MHC region in control samples.

Bottom panel: Genome-wide interaction mapping of T1D and RA. 99.8% of T1D interactions and 80.0% of RA interactions are in the MHC region. Strong interaction effects widely exist between genes in and across the MHC class I, II, and III in T1D, whereas most significant interactions of RA involve only loci closely placed in the MHC class II region (The p values are truncated at $p = 1.0 \times 10^{-16}$).
However, the type of underlying interaction is generally unknown and may vary widely.\textsuperscript{22} BOOST is more flexible because it covers a larger model space than PLINK. BOOST can be modified to test the allelic model by collapsing $3 \times 3$ contingency tables to $2 \times 2$ contingency tables (in the same way that PLINK does). The two-stage strategy in BOOST can then be applied to these $2 \times 2$ contingency tables. The statistical power of the modified BOOST will be roughly the same as PLINK because they both are based on the same allelic model. The ignorable difference is due to the difference between the Wald test and the likelihood ratio test. In the released software of BOOST, the allelic test has also been implemented. Regarding the running time, the BOOST allelic test is similar to the BOOST genotype test.

Relationship between Our Method and INTERSNP
Recently, INTERSNP\textsuperscript{37} has implemented the interaction test in GWAS using log-linear models. Regarding the interaction test, both INTERSNP and our work are developed on the basis of the standardized definition using logistic regression models.\textsuperscript{13} INTERSNP has directly used an iterative method to fit the log-linear model $M_I$. It is still very time consuming to test interactions in GWAS. Therefore, INTERSNP suggests the use of some prior knowledge to reduce the number of SNPs, including the single-locus test, genetics criteria, and pathway information. Genetics criteria and pathway information provide biological constraints that are very useful. But using the single-locus test in the filtering, which has been discussed in the earlier section, will filter out those SNPs with weak main effects but significant interactions. Moreover, how to choose the threshold in filtering is also critical. On the contrary, we propose to use the noniterative approximation to directly examine all SNP pairs. We show the computational performance of BOOST and INTERSNP in the following section.

Computation Time
From a practical point of view, a key issue of detecting gene-gene interactions in genome-wide case-control studies is the computational efficiency. Cordell\textsuperscript{4} reported that PLINK took about 14 days to test pairwise interactions of the selected 89,294 SNPs on a single node of a computer.

---

\textbf{Figure 5. The 31350k–31390k Region of Chromosome 6} 
\textit{HLA-B} in the MHC class I is located in this region. The recombination rate and LD plot from HapMap show that a block structure exists from 31360k to 31380k. This region is mapped through the SNPs rs2524057, rs2853934, rs2524115, rs396038, rs3873385, rs2524095, and rs2524089. The SNPs rs2524095 and rs2524089 are involved in the interactions with the 32930k–32960k region shown in Figure S2.
cluster. Random Jungle can analyze the large data sets quickly. However, Random Jungle aims at detecting association allowing for interactions rather than detecting interactions (see detailed explanations in the next subsection). Besides, Random Jungle has difficulty in finding interacting SNP pairs displaying weak main effects because trees built in Random Jungle rely on the main effects of SNPs. BEAM took about 8 days to handle 47,727 SNPs using $5 \times 10^7$ Markov chain Monte Carlo iterations. Currently, BEAM has difficulties in handling 500,000 to 1,000,000 SNPs genotyped in 5000 or more samples. Cordell recommended PLINK as a powerful method of testing interactions in GWAS.

We tested the running time of PLINK on our desktop computer. In addition, we also tested INTERSNP on the same data sets because INTERSNP also uses log-linear models to test interactions. The results are shown in Table 5. BOOST is roughly 63 times faster than PLINK and 95 times faster than INTERSNP. It can finish the analysis of all pairs of roughly 360,000 SNPs within 60 hr (around 2.5 days) on a standard desktop (3.0 GHz CPU with 4G memory running the Windows XP Professional x64 edition system). Parallel computing can be used to further improve the computation time for BOOST, PLINK, and INTERSNP. The WTCCC phase 2 study will analyze over 60,000 samples of various diseases using either the Affymetrix v6.0 chip or the Illumina 660K chip. The shared control samples will increase from 3000 to 6000. Such an increase in the number of SNPs and the sample size is more demanding on the computation efficiency. We anticipate that BOOST is still applicable for analyzing the new data sets.

### Test of interactions versus Test of Associations

To test association between a specific SNP $X_p$ and the phenotype $Y$, a typical method is to test the difference between the deviance of the null model (Equation 19) and the deviance of the alternative model (Equation 20) with $df = 2$:

$$
\log \frac{P(Y = 1)}{P(Y = 2)} = \beta_0
$$

(Equation 19)
The SNPs in the SNP 1 column reside in HLA-B, and the SNPs in the SNP 2 column are located at the block across HLA-DQA2 and HLA-DQB2. They show strong interactions without displaying significant main effects.

\[
\log \frac{P(Y = 1 | X_p = i)}{P(Y = 2 | X_p = i)} = \beta_0 + \beta_i^X
\]  
(Equation 20)

This is known as a "test of single-SNP association."

In the above test, SNP \(X_p\) is allowed to interact with other SNPs. As a matter of fact, if the disease is influenced by SNP \(X_p\) itself and its interaction effect with another SNP \(X_p\), the statistical power of detecting SNP \(X_p\) will be increased when allowing for interactions. This is known as a "test of two-locus associations allowing for interactions." Typically, this is accomplished by testing the difference between the log-likelihood of the null model (Equation 19) and that of the alternative model (Equation 21) with \(df = 8\):

\[
\log \frac{P(Y = 1 | X_p = i, X_q = j)}{P(Y = 2 | X_p = i, X_q = j)} = \beta_0 + \beta_i^X + \beta_j^Y + \beta_{ij}^{XY}
\]  
(Equation 21)

Marchini et al.\(^{11}\) highlighted the importance of testing associations allowing for interactions in a genome-wide scale and successfully demonstrated its feasibility. They reported that performing all pairwise tests of associations allowing for interactions with \(df = 8\) at 300,000 loci with 1000 cases and 1000 controls can be finished in 33 hr on a 10-node cluster. According to the equivalence between log-linear models and logistic models, it is clear that the feasibility of this exhaustive search method relies on the closed-form solution of the block independence model \(M_B\) and the closed-form solution of the saturated model \(M_s\) (see the Appendix for the details of \(M_B\) and \(M_s\)).

The differences of these tests are:

- The test of single-SNP association is to compare \(M_p\) with \(M_s\) (see Table 2 for descriptions of \(M_p\) and \(M_s\)).
- The test of associations allowing for interactions is to compare \(M_s\) with \(M_B\).
- The test of interaction is to compare \(M_s\) with \(M_{1B}\).

As we mentioned above, no closed-form solution exists for the test of interactions. In this sense, the test of interactions is more difficult than the test of associations allowing for interactions.

### On Statistical Epistasis

It is extensively debated to what extent statistical epistasis implies biological or functional epistasis.\(^3\) The statistical epistasis is exploited in the literature, perhaps because of the following reasons:

- The definition of statistical epistasis yields an appropriate measure for describing biological phenomena that one locus's effect on the phenotype depends on another locus.\(^2\) This facilitates mathematical analysis of epistasis.
- On the basis of the statistical definition, gene-gene interactions can be connected to Kullback-Leibler divergence used in the information theory (see Equation 12) and high-order mutual information in physics.\(^{15}\) This definition may bridge the gap between the biological understanding and the physical interpretation.
- Compositional epistasis, conceived by Bateson, is closer to the biological understanding of gene-gene interactions.

<table>
<thead>
<tr>
<th></th>
<th>BOOST</th>
<th>PLINK</th>
<th>INTERSNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n = 5000, L = 1000)</td>
<td>&lt; 2s</td>
<td>106s</td>
<td>160s</td>
</tr>
<tr>
<td>(n = 5000, L = 5000)</td>
<td>42s</td>
<td>2703s</td>
<td>4277s</td>
</tr>
<tr>
<td>(n = 5000, L = 10,000)</td>
<td>170s</td>
<td>10,915s</td>
<td>15,805s</td>
</tr>
</tbody>
</table>

PLINK is tested with the "–fast-epistasis" option and without the "–case-only" option. All timings are carried out on a 3.0 GHz CPU with 4G memory running the Windows XP Professional system.

---

Table 4. The Interaction SNP Pairs in the Two Regions Shown in Figure 5 and Figure 6

<table>
<thead>
<tr>
<th>SNP 1</th>
<th>Single-Locus p Value</th>
<th>SNP 2</th>
<th>Single-Locus p Value</th>
<th>Interaction p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2524057</td>
<td>4.807 × 10(^{-3})</td>
<td>rs9276448</td>
<td>8.878 × 10(^{-3})</td>
<td>5.362 × 10(^{-14})</td>
</tr>
<tr>
<td>rs2524057</td>
<td>4.807 × 10(^{-1})</td>
<td>rs5014418</td>
<td>1.116 × 10(^{-2})</td>
<td>2.738 × 10(^{-13})</td>
</tr>
<tr>
<td>rs2853934</td>
<td>8.336 × 10(^{-2})</td>
<td>rs9276448</td>
<td>8.878 × 10(^{-3})</td>
<td>2.507 × 10(^{-13})</td>
</tr>
<tr>
<td>rs2524115</td>
<td>1.215 × 10(^{-3})</td>
<td>rs9276448</td>
<td>8.878 × 10(^{-3})</td>
<td>6.456 × 10(^{-13})</td>
</tr>
<tr>
<td>rs3873385</td>
<td>3.368 × 10(^{-1})</td>
<td>rs9276448</td>
<td>8.878 × 10(^{-3})</td>
<td>3.186 × 10(^{-14})</td>
</tr>
<tr>
<td>rs3873385</td>
<td>3.368 × 10(^{-1})</td>
<td>rs5014418</td>
<td>1.116 × 10(^{-2})</td>
<td>3.841 × 10(^{-14})</td>
</tr>
<tr>
<td>rs3873385</td>
<td>3.368 × 10(^{-1})</td>
<td>rs6919798</td>
<td>6.077 × 10(^{-2})</td>
<td>4.257 × 10(^{-13})</td>
</tr>
<tr>
<td>rs396038</td>
<td>9.393 × 10(^{-2})</td>
<td>rs9276448</td>
<td>8.878 × 10(^{-3})</td>
<td>5.894 × 10(^{-12})</td>
</tr>
</tbody>
</table>

Figure 7. Potential Pathways Involving HLA-B, HLA-DQA2, and PSMB8

TIDM represents the type 1 diabetes mellitus pathway. Antigen represents the antigen processing and presentation pathway.

The American Journal of Human Genetics 87, 325–340, September 10, 2010 335
interactions than statistical epistasis.\textsuperscript{2} Compositional epistasis has recently been shown to be empirically testable via a statistical approach.\textsuperscript{38} In some cases, compositional and statistical epistasis are equivalent to each other.\textsuperscript{38} Therefore, statistical epistasis can still provide useful information for biological understanding.

Currently, PLINK, INTERSNP, and BOOST are designed to test statistical epistasis. We realize that detecting statistical epistasis in a genome-wide scale is easier than finding compositional epistasis because the test of compositional epistasis for each SNP pair requires enumerating all possible genetic interaction models.\textsuperscript{5} The detection of compositional epistasis will be investigated in our future work.

**Conclusion**

The large number of SNPs genotyped in genome-wide case-control studies poses a great computational challenge in the identification of gene-gene interactions. During the last few years, there have been fast-growing interests in developing and applying computational and statistical approaches to finding gene-gene interactions. In this paper, we present a method named ‘‘BOOST’’ to address this problem. Not only is BOOST computationally efficient, it has also shown good statistical power for a wide spectrum of epistasis models. We have successfully applied our method to analyze seven data sets from the WTCCC. Our experimental results demonstrate that interaction mapping is both computationally and statistically feasible for hundreds of thousands of SNPs genotyped in thousands of samples.

In this work, we focus mainly on the genome-wide case-control studies; i.e., the disease phenotype can be represented as a binary variable. In the current stage, our method cannot be applied to GWAS involving continuous phenotypes unless those continuous phenotypes can be discretized. There are two ways to handle covariates in our models. If the covariate is discrete or can be discretized, our method can be directly extended to handle it. If not, logistic regression can be used in the postprocessing step to adjust the covariate. In the postprocessing step, the computational burden of logistic regression is affordable because the number of selected interactions is limited.

There are some limitations of BOOST with respect to statistical power. BOOST uses a fixed degree of freedom ($df = 4$) to conduct the genotype test. When the contingency table is too sparse due to the low minor allele frequency, the degree of freedom of the statistical test should be reduced. To improve the performance of BOOST, we can first use BOOST to report interactions with a loose threshold and then use the penalized logistic regression\textsuperscript{39} with the adaptive degree of freedom to adjust these interactions. There are several other issues that we have not addressed, such as population substructures and imputation of the missed genotypes. We will investigate them in our future work.

**Appendix**

**Log-Linear models**

Here, we briefly describe four log-linear models, including the homogeneous association model $M_{HI}$, the saturated model $M_s$, the block independence model $M_b$, and the partial independence model $M_p$. These four models are used in the main text. Please see details in Agresti.\textsuperscript{14}

**Homogeneous Association Model $M_{HI}$**

The homogeneous association model $M_{HI}$ factorizes the joint distribution $\pi_{ijk}$ using the joint distributions of all pairs. The hypothesis is

$$H_0^H : \pi_{ijk} = \psi_{ij} \phi_{ik}$$  \hspace{1cm} (Equation 22)

where $\psi_{ij}$, $\phi_{ik}$ and $\omega_{jk}$ are some lower-order distributions. The name “homogeneous association” comes from the fact that the association between any two of three variables is the same at all levels of the third variable.\textsuperscript{14}

The homogeneous association model $M_{HI}$ is defined as

$$\log \mu_{ijk} = \lambda + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{ik}^{xz} + \lambda_{jk}^{yz} + \lambda_{ijk}^{xyz}$$  \hspace{1cm} (Equation 23)

Unfortunately, no closed-form expression exists for the MLE of $\mu_{ijk}$ (denoted as $\hat{\mu}_{ijk}$) in Equation 23. Iterative approaches, such as the Newton-Raphson method, are needed in order to estimate the parameters.

**Saturated Model $M_s$**

The saturated model $M_s$ defines the joint distribution with all factors. The saturated log-linear model is

$$\log \mu_{ijk} = \lambda + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{ik}^{xz} + \lambda_{jk}^{yz} + \lambda_{ijk}^{xyz}$$  \hspace{1cm} (Equation 24)

The MLE of $\mu_{ijk}$ in Equation 24 is

$$\hat{\mu}_{ijk} = n_{ijk}$$  \hspace{1cm} (Equation 25)

**Block Independence Model $M_b$**

When the joint distribution cannot be completely factorized, it may be factorized into blocks. The hypothesis is

$$H_0^B : \pi_{ijk} = \pi_{ij} \pi_k$$  \hspace{1cm} (Equation 26)

The corresponding log-linear model is

$$\log \mu_{ijk} = \lambda + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy}$$  \hspace{1cm} (Equation 27)

Under this structure, the MLE of $\mu_{ijk}$ is

$$\hat{\mu}_{ijk} = \frac{n_{ij} n_k}{n}$$  \hspace{1cm} (Equation 28)
Partial Independence Model $M_p$
The joint distribution may be factorized when some variables are given. For example, given $Y$, the hypothesis is

$$H_0^Y : \pi_{ijk} = \frac{\pi_{i_1k}\pi_{j}k}{\pi_{i_kj_k}} \quad \text{(Equation 29)}$$

The corresponding log-linear model is

$$\log \mu_{ijk} = \lambda + \lambda^X_i + \lambda^Y_j + \lambda^X^Y_{ik} + \lambda^X^Y_{ik} \quad \text{(Equation 30)}$$

Then the MLE of $\mu_{ijk}$ is

$$\hat{\mu}_{ijk} = \frac{n_{i_1k}n_{j_1k}}{n_{k}} \quad \text{(Equation 31)}$$

Connection between Log-Linear Models and Logistic Models
For convenience, we use the homogeneous association model $M_H$ as an example to describe the equivalence between a log-linear model and its corresponding logistic model. Its logit is

$$\log \frac{P(Y = 1 | X_p = i, X_q = j)}{P(Y = 2 | X_p = i, X_q = j)}$$


$$P(Y = i | X_p = i, X_q = j)$$

$$= \log \frac{\mu_{i1}}{\mu_{i2}}$$

$$= \log (\mu_{i1}) - \log (\mu_{i2})$$

$$= (\lambda + \lambda^X_i + \lambda^Y_j + \lambda^X^Y_{ik} + \lambda^X^Y_{ik})$$

$$- (\lambda + \lambda^X_i + \lambda^Y_j + \lambda^X^Y_{ik} + \lambda^X^Y_{ik})$$

$$= (\lambda^Y_i - \lambda^Y_j) + (\lambda^X^Y_{ik} - \lambda^X^Y_{ik}) + (\lambda^X^Y_{ik} - \lambda^X^Y_{ik})$$

$$= \beta_0 + \beta_i^X + \beta_j^Y$$

\text{(Equation 32)}

The first term is a constant that does not depend on $i$ or $j$. The second term depends only on the category $i$ of $X_p$. The third term depends only on the category $j$ of $X_q$. Therefore, this logit has the following form:

$$\log \frac{P(Y = 1 | X_p = i, X_q = j)}{P(Y = 2 | X_p = i, X_q = j)}$$

$$= (\lambda^Y_i - \lambda^Y_j) + (\lambda^X^Y_{ik} - \lambda^X^Y_{ik}) + (\lambda^X^Y_{ik} - \lambda^X^Y_{ik})$$

$$= \beta_0 + \beta_i^X + \beta_j^Y$$

\text{(Equation 33)}

The American Journal of Human Genetics 87, 325–340, September 10, 2010
contingency tables for all SNP pairs in a genome-wide case-control study, because hundreds of billions of SNPs pairs exist for typical genotyping chips.

In our method, we introduce a Boolean representation of genotype data. Instead of using one row for each SNP, the new representation uses three rows, with each row for one specific genotype. Each row consists of two-bit strings, one for control samples and the other for case samples. Each bit in the string represents one sample, and its value (0 or 1) indicates whether the sample has the corresponding genotype. For the above toy example, the corresponding Boolean representation is as follows:

\[
W_{bit} = \begin{bmatrix}
X_1 &= 1 & \text{Control} & \text{Case} \\
& 10000011 & 00001011 \\
X_2 &= 2 & 00110000 & 11000100 \\
X_3 &= 3 & 10001100 & 00110000 \\
X_{bit} &= 2 & 00001110 & 00111010 \\
X_{bit} &= 3 & 01110001 & 00000101 \\
X_{bit} &= 4 & 10000000 & 11000000 \\
X_{bit} &= 5 & 11000001 & 00111110 \\
X_{bit} &= 6 & 00001110 & 01110000 \\
X_{bit} &= 7 & 00110000 & 00000001
\end{bmatrix}
\]

Both \(W\) and \(W_{bit}\) contain the same amount of information. To demonstrate this equivalence, we underline some matched items between \(W\) and \(W_{bit}\). For example, the five 2’s in the first row of \(W\) are represented as five 1’s in the second row of \(W_{bit}\). Although the dimension of \(W_{bit}\) is three times as large as that of \(W\), its space usage in the computer is smaller because each byte can store 8 bits. For a data set with 4000 samples and 500,000 SNPs (about the same size as the WTCCC data set), the new data representation needs around 700M bytes, whereas the general data representation requires 1900M bytes. More importantly, using \(W_{bit}\) is more CPU efficient than using \(W\) in collecting the contingency table (Table 1). This is because we can directly carry out the fast logic (bitwise) operation with \(W_{bit}\). For example, to collect \(n_{121}\) in Table 1 (\(n_{121}\) represents the number of cases with \(X_q = 1\) and \(X_p = 2\)), we just need to conduct the logical AND operation on the case bit strings of row \(X_p = 1\) and \(X_q = 2\), then count the number of 1’s in the result. The 64-bit registers can perform 64-bit AND operation in one instruction, and the counting of “1” bits in a bit string (also called hamming weight) can be accomplished with an efficient algorithm (see http://en.wikipedia.org/wiki/Hamming_weight).

Genetic Heterogeneity Simulation

The simulation models are chosen on the basis of the performance of BOOST and PLINK in case 2. For each setting of \(h^2\) and \(MAF\), there are five models. We choose the one under which BOOST and PLINK perform best (i.e., have the highest statistical power). For example, both BOOST and PLINK have the best performance on model epi33 among models epi31–epi35 (with the same setting of \(h^2 = 0.05\) and \(MAF = 0.2\)). Therefore, for this setting of \(h^2\) and \(MAF\), we select model epi33. The reason for so doing is to make sure that both BOOST and PLINK have reasonably good performance when genetic heterogeneity is absent. Then we can observe how genetic heterogeneity degrades their performance. All selected models are given in Table S5. In the simulation, 100 data sets are generated under each model setting. In each data set, 1000 SNPs are simulated. Different sample sizes (\(n = 400, 800,\) and \(1600\)) are simulated. To simulate genetic heterogeneity, 50% case samples are generated at loci \(X_1\) and \(X_2\) and another 50% case samples are generated at loci \(X_3\) and \(X_4\). The distribution of case samples is based on a specific disease model given in Table S6. Each data set has two pairs of associated SNPs. Therefore, there are 200 pairs of SNPs for each parameter setting. We set the counter \(T\) to be zero initially. If one pair of these 200 pairs is detected (on the basis of the Bonferroni correction), then \(T = T + 1\). After testing 100 data sets, the power is calculated as \(T/200\).

Quality Control

We first check the quality of control samples:

- Those genotype data with a Chiamo score \(\leq 0.95\) are considered as missing data. SNPs with more than 10% missing data are removed.
- Those SNPs with a minor allele frequency \(< 0.05\) are removed.
- We also perform the Hardy-Weinberg Equilibrium (HWE) test for each SNP. Those SNPs with a \(p\) value \(\leq 0.001\) are removed.

Next, we check the quality of case samples. The strategy is similar to that for control samples except that the HWE test is not performed. The number of remaining SNPs is given in Table S1.

More Results of T1D Data Analysis

We have identified 91 interactions in which all loci are nonsignificant in the single-locus scan. These 91 interactions show two interesting interaction patterns between MHC class I and class II. We have shown one pattern in the main article. We have also identified another interaction pattern in chromosome 6 in the 31350k–31390k region (shown in Figure S1) and the 32930k–32960k region (shown in Figure S2). The six interactions between these two regions are listed in Table S2. It can be observed again that all SNPs in this table display weak main effects whereas their joint effects are statistically significant. We further report the odds ratios for those interactions in Table S3 and Table S4. For the first interaction group given in Table S3, the genotype combinations Aa/Bb, Aa/ bb, aa/Bb, and aa/ bb, where the uppercase and lowercase letters represent the major alleles and minor alleles, respectively, have significantly higher disease risks than others. The interaction effect of these genotypes can generally approximate the multiplicative model (see the left panel.
of Figure S3). For the second interaction group given in Table S4, the genotype combination aa/bb has a significantly higher disease risk than others. The interaction effect of this genotype is considered as a joint recessive effect (see the right panel of Figure S3).

Supplemental Data

Supplemental Data include eight figures and 14 tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

We thank the editor and the anonymous reviewers for their constructive suggestions and comments. This work was partially supported with grant GRF621707 from the Hong Kong Research Grant Council, grants RPC06/07.EG09, RPC07/08.EG25, and RPC10EG04 from the Hong Kong University of Science and Technology, and a grant from Sir Michael and Lady Kadoorie Funded Research Into Cancer Genetics.

Received: May 10, 2010
Revised: July 9, 2010
Accepted: July 29, 2010
Published online: September 2, 2010

Web Resources

The URL for data presented herein is as follows:


References


