BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 20 2010, pages 2517–2525
doi:10.1093/bioinformatics/btq486

Genome analysis

Advance Access publication August 24, 2010

Detecting two-locus associations allowing for interactions in genome-wide association studies

Xiang Wan[1,†], Can Yang[1,†], Qiang Yang[2], Hong Xue[3], Nelson L. S. Tang[4,*] and Weichuan Yu[1,*]

[1]Department of Electronic and Computer Engineering, [2]Department of Computer Science, [3]Department of Biochemistry, The Hong Kong University of Science and Technology and [4]Laboratory for Genetics of Disease Susceptibility, Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Genome-wide association studies (GWASs) aim to identify genetic susceptibility to complex diseases by assaying and analyzing hundreds of thousands of single nucleotide polymorphisms (SNPs). Although traditional single-locus statistical tests have identified many genetic determinants of susceptibility, those findings cannot completely explain genetic contributions to complex diseases. Marchini and coauthors demonstrated the importance of testing two-locus associations allowing for interactions through a wide range of simulation studies. However, such a test is computationally demanding as we need to test hundreds of billions of SNP pairs in GWAS. Here, we provide a method to address this computational burden for dichotomous phenotypes.

Results: We have applied our method on nine datasets from GWAS, including the aged-related macular degeneration (AMD) dataset, the Parkinson's disease dataset and seven datasets from the Wellcome Trust Case Control Consortium (WTCCC). Our method has discovered many associations that were not identified before. The running time for the AMD dataset, the Parkinson's disease dataset and each of seven WTCCC datasets are 2.5, 82 and 90 h on a standard 3.0 GHz desktop with 4 G memory running Windows XP system. Our experiment results demonstrate that our method is feasible for the full-scale analyses of both single- and two-locus associations allowing for interactions in GWAS.

Availability: http://bioinformatics.ust.hk/SNPAssociation.zip

Contact: nelsontang@cuhk.edu.hk; eeyu@ust.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 24, 2010; revised on July 29, 2010; accepted on August 18, 2010

1 INTRODUCTION

In genetics, it has been well established that single nucleotide polymorphisms (SNPs) are associated with a variety of diseases. In the emerging genome-wide association studies (GWAS), the goal is to identify genetic susceptibility through assaying and analyzing SNPs at the genome-wide scale. While the analysis of susceptibility of individual SNPs has been standardized and led to many interesting findings, these findings cannot completely explain the genetic causes of complex diseases. Consequently, identifying two-locus association patterns of complex diseases has attracted more attentions (Marchini *et al.*, 2005). In general, SNP association patterns in the two-locus association study can be partitioned into four classes based on the different combinations of main effects and interaction effects, as illustrated in Figure 1.

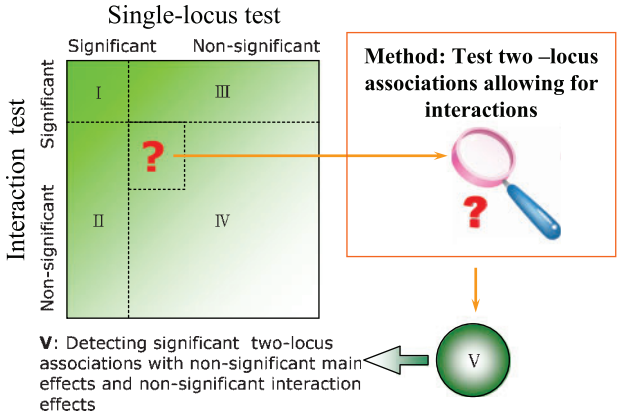

Fig. 1. The categorization of SNP association patterns in the two-locus association study. Class I: SNPs with a significant main effect and a significant interaction with a second SNP. Class II: SNPs with a significant main effect but without a significant interaction with any other SNP. Class III: SNPs without a significant main effect but with a significant interaction with some other SNP. Class IV: SNPs with neither a significant main effect nor a significant interaction with any other SNP. Class V: a subclass of Class IV, in which pairs of SNPs display significant associations via the combination of their main effects and their interaction effects. Please note that the first SNP and the second SNP in the definitions need not be in the same class.

Current works focus on finding SNPs in Classes I, II and III. As a result, many genetic determinants of susceptibility have been mapped. Two recent reviews (Balding, 2006; Cordell, 2009) presented the detailed analyses on many popular methods, including traditional single-locus-based statistical tests (Balding, 2006) and multi-locus analysis methods, such as *the ones implemented in PLINK* (Purcell *et al.*, 2007), multifactor dimensionality reduction (MDR) (Ritchie *et al.*, 2001), Tuning ReliefF (Moore and White, 2007), Random Jungle (Schwarz *et al.*, 2010) and BEAM (Zhang and Liu, 2007).

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

However, their findings cannot completely explain genetic contributions to complex diseases. Many researchers expect to detect more meaningful results by studying SNPs in Class IV. Although these SNPs neither display strong main effects nor significantly interact with other loci, some of them may influence diseases via the combination of main effects and interaction effects. This is referred to as *two-locus associations allowing for interactions* (formal definition in Section 2). Marchini *et al.* (2005) demonstrated the importance of testing associations allowing for interactions through a wide range of simulation studies. From a mathematical point of view, testing associations allowing for interactions enables us to detect a particular subset of SNPs in Class IV. We categorize this set as Class V in Figure 1, which is of our interests in this work.

While various methods have been proposed to test two-locus associations allowing for interactions [see the recent review (Cordell, 2009)], most of them fail to handle hundreds of billions of SNP pairs in GWAS. The main difficulty is the heavy computational burden. For example, in order to detect associations allowing for pairwise interactions from 500 000 SNPs, we need more than 10^{11} statistical tests. Some progress has been made on this issue recently. Marchini *et al.* (2005) showed that it was computationally intensive but still possible to test two-locus associations allowing for interactions in GWAS. BEAM (Zhang and Liu, 2007) is another feasible method to detect associations allowing for interactions in the genome-wide scale. Zhang and Liu (2007) proposed a Bayesian marker partition model and used Markov chain Monte Carlo (MCMC) sampling to optimize the posterior probability of the model. BEAM has successfully demonstrated its capability of handling large datasets using the synthetic data. But it did not provide convincing evidence using the real data. When the authors applied BEAM to analyze an aged-related macular degeneration (AMD) dataset consisting of 116 204 SNPs and 146 samples (Klein *et al.*, 2005), neither main effects nor interaction effects were reported by BEAM based on its *B*-statistic test. For many larger datasets such as those from Wellcome Trust Case Control Consortium (WTCCC), BEAM is only feasible after a filtering stage based on the single-locus test (Cordell, 2009).

Here, we provide a computational method to address this issue. We show that testing two-locus associations allowing for interactions of all SNP pairs in GWAS can be done in a fast manner. We also demonstrate the statistical power of our method on simulated datasets and the applicability of our method on nine real GWAS datasets including the AMD dataset (Klein *et al.*, 2005), the Parkinson's disease dataset (Fung *et al.*, 2006) and seven datasets from WTCCC (WTCCC, 2007). Our method discovers some associations that were not identified before. Currently, we only focus on finding two-locus associations allowing for interactions in case-control studies. We do not search for higher order interactions due to the following reasons:

(1) To detect strong interactions between SNPs with weak main effects (or even in the absence of main effects), the exhaustive search is needed. Evaluating all pairs has already imposed a computational challenge. The brute-force search for higher order interactions will be computationally prohibitive. Heuristic strategies have to be applied but without guaranteed performance.

(2) Sample size is another issue for detecting higher order interactions. For a d-order interaction, the contingency table has $3^d \times 2$ cells. Due to the limited sample size, there will be many empty (or close to empty) cells, which will lead to unstable estimation (large variance). This instability makes the identified result difficult to be reproduced by other independent studies.

In the following sections, we first provide a formal definition of testing two-locus associations allowing for interactions. Then we describe our test statistics based on log-linear models. We show that our proposed test statistics can be evaluated analytically. Based on these test statistics, we develop a stepwise strategy to detect both single- and two-locus associations allowing for interactions. The type I error of the proposed strategy is validated through the null simulation study.

2 METHODOLOGY

2.1 Notation

Suppose a dataset includes $\mathcal{L}$ SNPs and n samples. X_l is used to denote the l-th SNP, $l = 1, \ldots, \mathcal{L}$ and Y is the class label ($Y = 1$ for cases and $Y = 2$ for controls). SNPs are bi-allelic genetic markers. In general, capital letters (e.g. A, B, …) denote major alleles and lowercase letters (e.g. a, b, …) denote minor alleles. For each SNP, there are three genotypes: homozygous reference genotype (AA), heterozygous genotype (Aa) and homozygous variant genotype (aa). The popular way of coding the genotype data is to use {1, 2, 3} to represent {AA, Aa, aa}, respectively.

2.2 Definition of two-locus associations allowing for interactions

There are two types of associations allowing for interactions. One is partial association allowing for interaction and the other full association allowing for interaction.

2.2.1 Two-locus partial association allowing for interaction The two-locus partial association allowing for interaction measures the association effect of a given locus X_p allowing for the interaction with another locus X_q. The likelihood ratio test is often used to test such associations. It involves three steps:

(1) Fit a full logistic regression model defined in Equation (1) to measure the full association between (X_p, X_q) and the class label Y and obtain the maximum log-likelihood value $\hat{L}_F$.

$$\log \frac{P(Y=1|X_p=i, X_q=j)}{P(Y=2|X_p=i, X_q=j)} = \beta_0 + \beta_{X_p=i} + \beta_{X_q=j} + \beta_{X_p=i, X_q=j} \quad (1)$$

Here, $\beta_{X_p=i}$ represents the coefficient of X_p at category i. The meaning of $\beta_{X_p=i}$ and $\beta_{X_p=i, X_q=j}$ are similar.

(2) Fit a logistic regression model defined in Equation (2) to measure the main effect of X_q and obtain maximum log-likelihood value $\hat{L}_M$.

$$\log \frac{P(Y=1|X_q=j)}{P(Y=2|X_q=j)} = \beta_0 + \beta_{X_q=j} \quad (2)$$

(3) Calculate the P-value using the χ^2 test on the value of $2 \cdot (\hat{L}_F - \hat{L}_M)$ with degree of freedom $df = 6$.

Table 1. The genotype counts in cases ($Y=1$) and controls ($Y=2$)

$Y=1$	$X_q=1$	$X_q=2$	$X_q=3$
$X_p=1$	n_{111}	n_{121}	n_{131}
$X_p=2$	n_{211}	n_{221}	n_{231}
$X_p=3$	n_{311}	n_{321}	n_{331}
$Y=2$	$X_q=1$	$X_q=2$	$X_q=3$
$X_p=1$	n_{112}	n_{122}	n_{132}
$X_p=2$	n_{212}	n_{222}	n_{232}
$X_p=3$	n_{312}	n_{322}	n_{332}

2.2.2 Two-locus full association allowing for interaction The two-locus full association allowing for interaction measures the association effect of two loci X_p and X_q allowing for the interaction between them. The corresponding likelihood ratio test is conducted as the following steps:

(1) Fit a full logistic regression model defined in Equation (1) and obtain the maximum log-likelihood value $\hat{L}_F$.

(2) Fit a null logistic regression model defined in Equation (3) and obtain the maximum log-likelihood value $\hat{L}_\emptyset$.

$$\log\frac{P(Y=1)}{P(Y=2)}=\beta_0 \tag{3}$$

(3) Calculate the P-value using the χ^2 test on the value of $2\cdot(\hat{L}_F-\hat{L}_\emptyset)$ with degree of freedom $df=8$.

These likelihood ratio tests based on logistic regressions are widely accepted. However, it is computationally demanding to evaluate hundreds of billions of SNP pairs in GWAS by directly using these tests. Therefore, faster test procedures without losing statistical powers are needed in GWAS. Noticing the equivalence between a logistic regression model and its corresponding log-linear model in categorical data analysis (Agresti, 2002; Wan *et al.*, 2010), here we propose to test two-locus associations allowing for interactions based on log-linear models. The advantage of so doing is that test statistics can be derived in the closed form.

2.3 Log-linear models

Given two loci X_p and X_q, a contingency table of (X_p, X_q) and Y will be used (Table 1) to test two-locus associations allowing for interactions. The size of the contingency table is $I\times J\times K$ with $I=3$, $J=3$ and $K=2$. In the contingency table (shown in Table 1), we use n_{ijk} to denote the observed count in the cell (i, j, k), which is a realization of a random variable N_{ijk} assumed as Poisson distributed in log-linear models. Although Poisson distribution may not be entirely suitable for the fixed sample size, it is still considered as a reasonably good approximation to a multinomial model. Such an approximation makes the mathematics tractable.

Clearly, we have $n=\sum_{i,\,j,\,k} n_{ijk}$. We use π_{ijk} to denote the probability that an observation falls in the cell (i, j, k). A natural constraint of π_{ijk} is

$$\sum_{i,j,k}\pi_{ijk}=1. \tag{4}$$

Therefore, the expectation of N_{ijk} is

$$\mu_{ijk}=n\pi_{ijk}. \tag{5}$$

The likelihood function is

$$f(\boldsymbol{\mu})=\prod_{i,j,k}\frac{e^{-\mu_{ijk}}\mu_{ijk}^{n_{ijk}}}{n_{ijk}!}. \tag{6}$$

Correspondingly, the log-likelihood function is

$$L(\boldsymbol{\mu})=\sum_{i,j,k}\left[n_{ijk}\log(\mu_{ijk})-\mu_{ijk}-\log(n_{ijk}!)\right]. \tag{7}$$

In the space of log-linear models, the saturated model matches the full logistic regression model [defined in Equation (1)]; the partial independence model is the equivalent form of the logistic regression model with only main effects [defined in Equation (2)]; and the block independence model corresponds to the null logistic regression model [defined in Equation (3)]. In the following, we first explain these three log-linear models and how they connect to the corresponding logistic regression models. Then, we describe how to test two-locus associations allowing for interactions using log-linear models.

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

2.3.1 Saturated model The saturated model M_S defines the joint distribution with all factors. The hypothesis is

$$H_0^S:\pi_{ijk}^S=\pi_{ijk}.$$

The saturated log-linear model is

$$\begin{aligned}\log\mu_{ijk}=&\lambda+\lambda_{X_p=i}+\lambda_{X_q=j}+\lambda_{Y=k}+\lambda_{X_p=i,X_q=j}\\&+\lambda_{X_p=i,Y=k}+\lambda_{X_q=j,Y=k}+\lambda_{X_p=i,X_q=j,Y=k}.\end{aligned} \tag{8}$$

Please note that once the sample size is fixed and the ratio between cases and controls is fixed, the model given in Equation (8) only has degree of freedom $df=8$ even though there are 18 coefficients in the model.

Using Equation (7), the maximum likelihood estimation (MLE) of μ_{ijk} in Equation (8) is

$$\hat{\mu}_{ijk}^S=n_{ijk}. \tag{9}$$

2.3.2 Partial independence model The partial independence model M_P factorizes the joint distribution π_{ijk} with the assumption that X_q is independent of the class label Y if X_p is given. The hypothesis is

$$H_0^P:\pi_{ijk}^P=\frac{\pi_{ij.}\pi_{i.k}}{\pi_{i..}}.$$

The partial independence log-linear model is

$$\log\mu_{ijk}=\lambda+\lambda_{X_p=i}+\lambda_{X_q=j}+\lambda_{Y=k}+\lambda_{X_p=i,X_q=j}+\lambda_{X_p=i,Y=k}. \tag{10}$$

The MLE of μ_{ijk} in Equation (10) can be obtained as

$$\hat{\mu}^P_{ijk} = \frac{n_{ij.} n_{i.k}}{n_{i..}}. \tag{11}$$

2.3.3 Block independence model The block independence model M_B defines the joint distribution with the assumption that the joint distribution of X_p and X_q is independent of the class label Y. The hypothesis is

$$H_0^B : \pi^B_{ijk} = \pi_{ij.}\pi_{..k}.$$

The block independence log-linear model is

$$\log \mu_{ijk} = \lambda + \lambda_{X_p=i} + \lambda_{X_q=j} + \lambda_{Y=k} + \lambda_{X_p=i,X_q=j}. \tag{12}$$

The MLE of μ_{ijk} in Equation (12) is

$$\hat{\mu}^B_{ijk} = \frac{n_{ij.} n_{..k}}{n}. \tag{13}$$

2.4 Connection between log-linear models and logistic models

For brevity, we use the partial independence model M_P to show the equivalence between a log-linear model and its corresponding logistic model. The logit based on the partial independence model M_P is

$$\begin{aligned}
&\log \frac{P(Y=1|X_p=i,X_q=j)}{P(Y=2|X_p=i,X_q=j)} \\
&= \log \frac{\mu_{ij1}}{\mu_{ij2}} = \log(\mu_{ij1}) - \log(\mu_{ij2}) \\
&= (\lambda + \lambda_{X_p=i} + \lambda_{X_q=j} + \lambda_{Y=1} + \lambda_{X_p=i,X_q=j} + \lambda_{X_p=i,Y=1}) \\
&\quad - (\lambda + \lambda_{X_p=i} + \lambda_{X_q=j} + \lambda_{Y=2} + \lambda_{X_p=i,X_q=j} + \lambda_{X_p=i,Y=2}) \\
&= (\lambda_{Y=1} - \lambda_{Y=2}) + (\lambda_{X_p=i,Y=1} - \lambda_{X_p=i,Y=2}).
\end{aligned} \tag{14}$$

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

$$\begin{aligned}
&\log \frac{P(Y=1|X_p=i,X_q=j)}{P(Y=2|X_p=i,X_q=j)} \\
&= \beta_0 + \beta_{X_p=i}.
\end{aligned} \tag{15}$$

Clearly, this is equivalent to the logistic regression model with only main effects defined in Equation (2).

Using the same inference shown above, it is straightforward to find the connection between the saturated model M_S and the full logistic regression model defined in Equation (1) and the connection between the block independence model M_B and the null logistic regression model defined in Equation (3).

2.5 Testing two-locus associations allowing for interactions using log-linear models

Based on the equivalence between the log-linear model and its corresponding logistic regression model, we construct our test statistics based on the partial independence model M_P, the saturated model M_S and the block independence model M_B. Let $\hat{L}_P$, $\hat{L}_S$ and $\hat{L}_B$ be the log-likelihood of M_P, M_S and M_B evaluated at their MLEs, respectively.

2.5.1 Testing two-locus partial associations allowing for interactions To measure two-locus partial associations allowing for interactions based on the likelihood ratio test using log-linear models, we have

$$\hat{L}_S - \hat{L}_P = \sum_{i,j,k} \left[n_{ijk} \log \frac{\hat{\mu}^S_{ijk}}{\hat{\mu}^P_{ijk}} - \hat{\mu}^S_{ijk} + \hat{\mu}^P_{ijk} \right]. \tag{16}$$

As Equation (5) implies that

$$\sum_{i,j,k} \hat{\mu}^P_{ijk} = \sum_{i,j,k} \hat{\mu}^S_{ijk} = n. \tag{17}$$

Equation (16) can be further reduced as

$$\begin{aligned}
\hat{L}_S - \hat{L}_P &= \sum_{i,j,k} \left[n_{ijk} \log \frac{\hat{\mu}^S_{ijk}}{\hat{\mu}^P_{ijk}} \right] \\
&= n \sum_{i,j,k} \left[\hat{\pi}^S_{ijk} \log \frac{\hat{\pi}^S_{ijk}}{\hat{\pi}^P_{ijk}} \right] \\
&= n \cdot D_{KL}\left(\hat{\pi}^S_{ijk} || \hat{\pi}^P_{ijk} \right),
\end{aligned} \tag{18}$$

where $D_{KL}(\hat{\pi}_{ijk} || \hat{p}_{ijk})$ is the Kullback–Leibler divergence (Kullback and Leibler, 1951) of $\hat{\pi}^S_{ijk}$ and $\hat{\pi}^P_{ijk}$. In information theory, the Kullback–Leibler divergence is considered as a type of a non-symmetric distance between two probability densities.

2.5.2 Testing two-locus full associations allowing for interactions Following the similar inference as mentioned above, we have

$$\hat{L}_S - \hat{L}_B = \sum_{i,j,k} \left[n_{ijk} \log \frac{\hat{\mu}^S_{ijk}}{\hat{\mu}^B_{ijk}} \right] = n \cdot D_{KL}\left(\hat{\pi}^S_{ijk} || \hat{\pi}^B_{ijk} \right). \tag{19}$$

As $\hat{\pi}^S_{ijk}$, $\hat{\pi}^P_{ijk}$ and $\hat{\pi}^B_{ijk}$ all have the closed-form solutions [see Equations (9, 11 and 13)], the test statistics can be quickly computed. This enables us to detect two-locus associations allowing for interactions in GWAS in a fast manner.

2.6 A stepwise strategy to detect two-locus associations allowing for interactions

We develop a stepwise method to detect two-locus associations allowing for interactions in GWAS. It involves the following steps.

- Step 1: for all of $\mathcal{L}$ SNP markers, we test the main effects using the likelihood ratio test. The details are given in the Supplementary Material. We determine the significance of main effects based on the Bonferroni correction of $\mathcal{L}$ tests.
- Step 2: for those SNPs without significant main effects, we check every pair (X_p, X_q) and evaluate its association under the following three models:
  - (a) Compute the statistic for X_p association allowing for the interaction with X_q using Equation (18), where $\hat{\mu}^S_{ijk}$ and $\hat{\mu}^P_{ijk}$ are calculated using Equations (9) and (11), respectively.
  - (b) Compute the statistic for X_q association allowing for the interaction with X_p similarly.

(c) Compute the full association statistic using Equation (19), where $\hat{\mu}^S_{ijk}$ and $\hat{\mu}^B_{ijk}$ are calculated using Equations (9) and (13), respectively.

(d) Compute Bayesian information criterions (BICs) for all three models and choose the best one to conduct the χ^2 test with degree of freedom $df = 6$ (partial association) or $df = 8$ (full association). In statistics, the BIC or Schwarz criterion (Schwarz, 1978) is a criterion for model selection and it provides a measure to favor one model over another. The Bonferroni correction is applied for $\mathcal{L}(\mathcal{L}-1)/2$ tests.

The model selection in (d) of Step 2 is critical. Marchini *et al.* (2005) did not consider this issue. According to our simulation results, ignoring the model selection would result in an unexpected high type I error rate (see our experimental result in Section 3.2).

3 RESULTS

In this section, we evaluate our proposed method using both simulation data and real data. It is also instructive to compare the empirical power of our method with some recent approaches. In Cordell (2009), BEAM was recommended as a very powerful one which could handle large-scale data and finish in a reasonable period of time. In addition, BEAM has been compared with MDR, the stepwise logistic regression and the logic regression (Zhang and Liu, 2007). BEAM outperforms all of them. Therefore, we focus on the comparison between our method and BEAM using simulated data generated from four popular epistatic models and nine real datasets, one from the AMD study (Klein *et al.*, 2005), one from the Parkinson's disease study (Fung *et al.*, 2006) and the other seven datasets from WTCCC. We also conduct the typical single-locus analysis (the details of our single-locus analysis are provided in Supplementary Section 1) and use its performance as the reference to display the improvement that the two-locus association analysis can achieve. To check the type I error of our method, we further design a null simulation study.

3.1 Simulation experiment from four epistasis models

In this experiment, we select four epistasis models whose odds tables are given in the Supplementary Material. Model 1 is a multiplicative model considered in Marchini *et al.* (2005). Model 2 is discussed in Neuman and Rice (1992) to describe handedness (Levy and Nagylaki, 1992)[1] and the color of swine (Lerner, 1968). Model 3 is discussed in Li and Reich (2000) and Frankel and Schork (1996). Model 4 is the well-known XOR model. For each dataset, we generate genotype data based on the Hardy–Weinberg principle. We set the MAFs of disease-associated SNPs as 0.2 and 0.4. We generate the MAFs of unassociated SNPs uniformly from [0.05, 0.5]. The parameters of each model for each setting are calculated based on the prespecified disease prevalence $p(D)$ and the genetic heritability h^2 (please see details in the Supplementary Material). We simulate 100 datasets under each setting for each disease model. Each dataset contains 1000 SNPs. To take sample sizes into consideration, we generate 800 and 1600 samples with the balanced design.

The performance comparison of three methods is provided in Supplementary Figure 2 with the significance thresholds selected as 0.1, 0.2 and 0.3 after the Bonferroni correction. It exactly matches the analysis of variance (ANOVA) of the four disease models (Supplementary Fig. 1). The total variance of simulation models is decomposed into two parts: the variance explained by main effects and the variance explained by interactions. The performances of the single-locus analysis and BEAM are expected to be consistent as what ANOVA indicates. Specifically, when main effects are noticeable, they perform reasonably well. When interaction effects dominate, they perform poorly. For all models, our method has a higher power on average because each model has an interaction effect. For Models 1 and 3 with MAF = 0.2, BEAM outperforms our method because the generated datasets contain noticeable main effects in those settings. The MCMC used in BEAM has a high chance to sample simulated true pairs if main effects are present. The B-statistic used in BEAM often has a higher power than the statistic used in the likelihood ratio test as demonstrated in Zhang and Liu (2007). However, for other settings with weak main effects, our method outperforms BEAM. The sample size plays an important role for all methods. The power increases a lot when the sample size increases from $n = 800$ to 1600.

[1]It is interesting to note that recent papers did not accept this description any more (e.g. Vuoksimaa *et al.*, 2009).

Table 2. The number of false positives under different significant thresholds

Significant threshold	0.01	0.05	0.10	0.15	0.20
With BIC	0.010	0.048	0.085	0.129	0.186
Without BIC	0.029	0.146	0.289	0.432	0.587

3.2 Null simulation to test type I errors

To display type I errors of our method, we conduct the null simulation experiments. We generate 1000 null datasets. Each dataset consists of 1000 SNPs and 1000 samples. All SNPs are simulated independently with MAFs uniformly distributed in $[0.05, 0.5]$. We conduct two experiments on those null datasets, one using BIC in model selection and the other one without using BIC. The experiment results are provided in Table 2, from which we can see that the type I errors of our method using BIC agree with the nominal error rates while the numbers of false positives without using BIC are much higher. This results show that our model selection strategy plays an essential role in controlling the type I error at nominal error rate.

3.3 Experiment on the AMD dataset

AMD is a disease associated with aging that gradually destroys central vision. The AMD study (Klein *et al.*, 2005) genotyped 116 204 SNPs on 96 cases and 50 controls. We first apply the similar quality control process as suggested in Klein *et al.* (2005) and keep 82 144 SNPs. Klein *et al.* (2005) reported two significant loci, rs380390 and rs1329428, based on the allelic association with degree of freedom $df = 1$. In our single-locus analysis, these two loci are not significant based on the genotype association with degree of freedom $df = 2$. Their unadjust P-values are 1.75×10^{-6} and 3.64×10^{-6}. BEAM also did not detect these two SNPs because its B-statistic was used with degree of freedom $df = 2$ (Zhang and Liu, 2007). However, both our single-locus test and BEAM ranked them as the top two.

BEAM did not report any interactions in this dataset, while our method discovered a partial association between AMD and

SNP rs994542 interacting with SNP rs9298846. The *P*-value of this association is 0.070 after the Bonferroni correction. Owing to the small number of samples, the odds ratios cannot be reliably estimated. Therefore, we only provide their genotype distributions in the Supplementary Material. It is not surprising that BEAM does not report this SNP pair. The unadjusted single-locus *P*-values of rs994542 and rs9298846 are 0.009 and 0.723, respectively. BEAM is unlikely to pick up these SNPs during its MCMC iterations. How these two SNPs jointly contribute to the disease trait or whether they are tagging ungenotyped functional loci is under further investigation.

3.4 Experiment on the Parkinson's disease dataset

Parkinson's disease is a chronic neurodegenerative disease that often impairs the motor skills, speech and other functions of the patients. Fung *et al.* (2006) genotyped 396 591 SNPs in 22 autosomes of 541 samples. In their study, they did not identify any significant association using the single-locus test. BEAM was applied to analyze this dataset and did not report any association. Our method identifies a partial two-locus association involving SNP rs10519435 interacting with SNP rs849523. The *P*-value of this association is 0.061 after the Bonferroni correction. These two SNPs have high-genotyping quality with only one missing genotype.

The joint distributions of these two SNPs in cases and controls and the corresponding odds ratios are provided in Figure 3, Table 3 and Table 4 in the Supplementary Material. The genotype combinations 'CC/AA', 'CT/AG' and 'TT/AG' have significantly higher odds ratios than others genotype combinations. Please note that the significantly higher disease risk of 'CT/AG' is supported by larger counts in both cells than 'CC/AA' and 'TT/AG' (see details in the Supplementary Material). In addition, the genotype 'CT/AA' has a significantly low disease risk. SNP rs10519435 resides at the intron of gene *LVRN* (laeverin). Its neighbor gene *SNCAIP* has been strongly associated with the Parkinson's disease (Engelender *et al.*, 2000). SNP rs849523 resides at the intron of gene Neuropilin-2 (*NRP2*) which is a vascular endothelial growth factor (*VEGF*) receptor. It is a well-known gene associated with many cancers. The ligands of *NRP2* are Class-3 semaphorins. In the nervous system, altered semaphorin function has been linked to Alzheimer's disease, motor neuron degeneration, schizophrenia and Parkinson's disease (Williams *et al.*, 2007). To our knowledge, this is the first identified association that can pass the statistical significance test for the study (Fung *et al.*, 2006). Although it is still unclear whether and how these genes jointly affect disease traits, this finding provides a clue to further analyze the dataset.

3.5 Experiments on seven datasets from WTCCC

The WTCCC is a collaboration of many British research groups. In the first phase, the WTCCC has examined the genetic signals (500K SNPs) of seven common human diseases: bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D) and type 2 diabetes (T2D) (14 000 cases in total and 3000 shared controls). Before we run our method on these datasets, we first apply a similar quality control procedure as suggested in (WTCCC, 2007) to preprocess the data. The numbers of remaining SNPs are roughly 360 000 (see details in the Supplementary Material). BEAM[2] cannot directly handle these datasets in current stage unless a single-locus test is used to filter out the SNPs with weak marginal effects based on their single-locus *P*-values (Cordell, 2009). In contrast, our method has discovered many associations that were not identified before.

Table 3. The number of two-locus associations allowing for interactions identified from seven diseases datasets under different constraints

	RA	T1D	T2D	CD	BD	CAD	HT
T^1	254	970	89	45	77	56	57
T^2	283	1152	104	61	86	62	59
T^3	303	1237	111	70	95	63	60
T^1 & D	3	266	2	9	0	0	0
T^2 & D	10	331	4	20	0	0	0
T^3 & D	12	364	6	26	1	0	0

T^1: the threshold of Bonferroni-corrected *P*-value is 0.10; T^2: the threshold of Bonferroni-corrected *P*-value is 0.20; T^3: the threshold of Bonferroni-corrected *P*-value is 0.30; D: the physical distance between two interacting SNPs is at least 1 Mb. This constraint is used to avoid interactions that might be attributed to the LD effect (Cordell, 2009).

The numbers of identified two-locus associations under three statistical significance thresholds with and without the distance constraint for seven diseases are reported in Table 3. The distance constraint is used to avoid associations involving two closely placed loci because the interaction effects in those associations may be attributed to the linkage disequilibrium (LD) effect (Cordell, 2009). Since people are more interested in associations induced by interactions between distantly placed loci and the significance threshold of 0.10 is widely accepted, we focus on the analysis of two-locus associations allowing for interactions under the threshold of Bonferroni-corrected *P*-value of 0.10 with the distance constraint. The details of these associations (including the statistical significance of single *P*-values and interaction *P*-values) are listed in the Supplementary Material.

3.5.1 RA All loci in the identified three two-locus associations are located at the major histocompatibility complex (MHC) region in Chromosome 6. The MHC region has been confirmed as the most variable region in the human genome with respect to infection, inflammation, autoimmunity and transplant medicine (Lechler and Warrens, 2000). The recent study (WTCCC, 2007) has replicated the strong association between gene *HLA-DRB1* and RA via single-locus association mapping. Our result further replicates another strong association between gene *HLA-B* and the disease trait, which was previously reported in (Galocha and de Castro, 2008). The identified association involves the locus rs4394275 interacting with the locus rs9276440. The *P*-value of this association is 0.059 after the Bonferroni correction. The locus rs4394275 resides at the intron region of gene *HLA-B* and the locus rs9276440 resides at the intron region of gene *HLA-DQA2*. Both genes are located at the MHC region and have shown strong associations with RA in many studies (Galocha and de Castro, 2008; Vignal *et al.*, 2009). Our results further provide the evidence that these two genes may contribute to the etiology of RA.

[2]The current version of BEAM cannot handle datasets with ∼500K SNPs and ∼5000 samples. In Wan *et al.* (2010), we tested BEAM by loading one chromosome at a time and then assembling the results together.

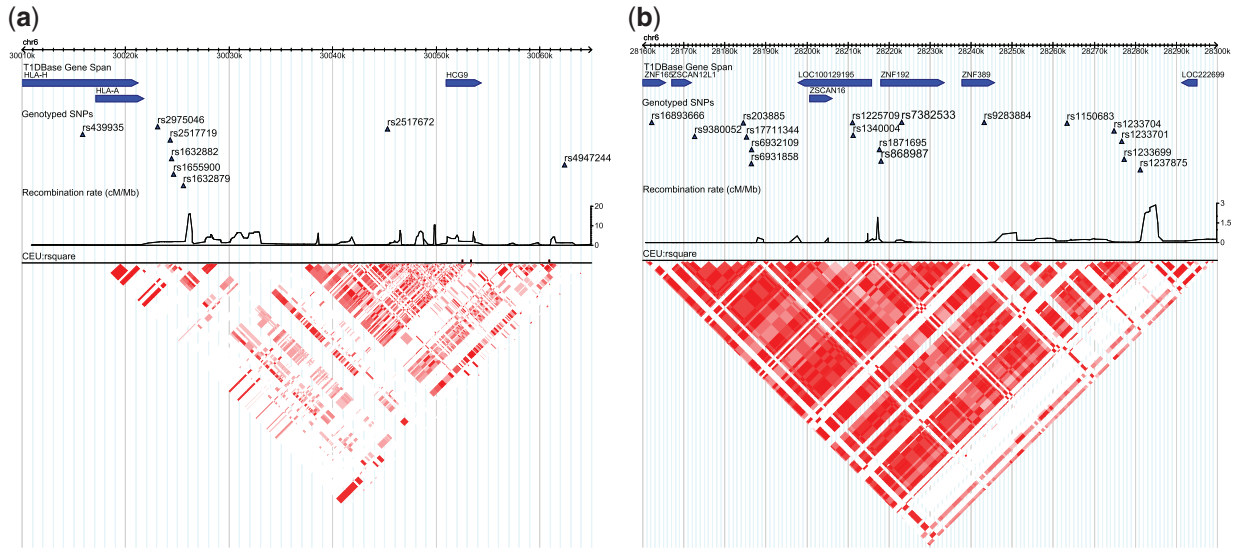

Fig. 2. The identified two interesting association regions. (**a**) The region 30 010k–30 065k of Chromosome 6. The recombination rate and LD plot from HapMap show that a block structure spans from 30 010k to 30 065k. (**b**) The region 28 160k–28 300k of Chromosome 6. The recombination rate and LD plot from HapMap show that a block structure exists from 28 160k–28 300k. The genotyped loci in our identified associations in both regions are marked in the figure. Many genes including *HLA-A* genes and *ZNF* genes in these two regions have shown strong associations with T1D [see a detailed review (Pociot and McDermott, 2002)].

3.5.2 T1D Most identified associations (except one locus rs7704018 in Chromosome 5, which is not well studied) from the T1D datasets are linked with the MHC region. Previous reports (Nejentsev *et al.*, 2007; WTCCC, 2007) have identified strong associations between many MHC genes (such as *HLA-DQB1* and *HLA-DRB1*) and T1D using the single-locus test. However, it is unclear which and how many loci within the MHC region determine T1D susceptibility. Our results provide additional information to help pinpoint disease-associated loci because SNPs involved in those associations are usually filtered out in the single-locus test. A further inspection of the 266 associations of our interest has identified two association regions which span from 30 010k bases to 30 065k bases and from 28 160k bases to 28 300k bases. Both regions are illustrated in Figure 2. Many genes including *HLA-A* genes and *ZNF* genes in these two regions have shown strong associations with T1D [see a detailed review in Pociot and McDermott (2002)]. Our finding may provide some new insights in studying the causes of T1D.

3.5.3 T2D Only two associations allowing for interactions are identified in the T2D dataset. One suspicious association involves the locus rs9586155 interacting with the locus rs7031174. The *P*-value of this association is 0.06 after the Bonferroni correction. The locus rs9586155, located at the region q33.1 in Chromosome 13, is not well studied. Nor is it in the HapMap collection. The locus rs7031174 resides at the intron of gene *MELK*. *MELK* has been found significantly over-expressed in the great majority of breast cancer cells (Lin *et al.*, 2007). It is well known that there are strong associations between T2D and breast cancer, and the insulin resistance has been connected to an increased risk of breast cancer (Michels *et al.*, 2003). Although breast cancer occurs mainly in women and thus this evidence may not indicate a direct connection between T2D and gene *MELK*, the investigation of genes interacting with gene *MELK* may help identify the causes of T2D.

3.5.4 CD Among the nine associations identified from the CD dataset, we have not found strong evidences to relate them with the disease trait. However, using the LD analysis helps us identify a well-studied gene *ANTXR2*. One identified association involves the locus rs10008294 interacting with the locus rs1450526. The *P*-value of this association is 0.042 after the Bonferroni correction. The locus rs10008294 is linked with the loci rs4312703 and rs10049995 of gene *ANTXR2*. The r^2 between rs10008294 and rs4312703 is 0.933 and the D′ is 0.966. The r^2 between rs10008294 and rs10049995 is 0.93 and the D′ is 0.964. Both rs4312703 and rs10049995 are included in the analysis. However, they are not in the identified associations. Gene *ANTXR2* encodes a receptor for anthrax toxin and mutations in this gene cause juvenile hyaline fibromatosis and infantile systemic hyalinosis.

3.6 Computation time

From a practical point of view, a key issue of two-locus association analysis in GWAS is the computational efficiency (Cordell, 2009). BEAM is a recent method which can handle large-scale data. BEAM took about 8 days to handle 47 727 SNPs genotyped on around 5000 samples using 5×10^7 MCMC iterations. A rough estimation implied that BEAM would take approximately 5 weeks to analyze 89 294 SNPs genotyped on those 5000 samples (Cordell, 2009). Currently, BEAM fails to handle 500 000 to 1 000 000 SNPs genotyped on 5000 or more samples.

Our method makes a significant progress in computational time. It evaluates two-locus associations allowing for interactions of all pairs of about 360 000 SNPs (i.e. 6.5×10^{10} pairs) within 90 h (within 4 days) on a standard desktop (3.0 GHz CPU with 4G memory running Windows XP professional x64 Edition system). A detailed running time comparison given in Table 4 shows that our method is roughly 20–70 times faster than BEAM. The ongoing WTCCC phase 2 study

Table 4. Time comparison of BEAM and our method

Setting	BEAM	Our method
$n=1000, \mathcal{L}=10000$	4507s	197s
$n=2000, \mathcal{L}=10000$	9058s	216s
$n=5000, \mathcal{L}=10000$	18 469s	259s

BEAM runs 5.0×10^7 MCMC iterations for all settings. All timings are carried out on one 3.0 GHz CPU with 4G memory running Windows XP professional x64 Edition system.

is analyzing over 60 000 samples of various diseases, which are genotyped on about one million SNPs using either the Affymetrix v6.0 chip or the Illumina 660K chip. The shared control samples will increase from 3000 to 6000. Such a growth in number of SNPs and sample size becomes even more demanding on the computation efficiency. Our new method should be useful in the analyses of new datasets.

4 DISCUSSION AND CONCLUSION

The development of our method is triggered by the limitations of existing works on detecting two-locus associations allowing for interactions from GWAS. Our method shares the same goal with Marchini's method (Marchini *et al.*, 2005). Both methods take interaction effects into account when detecting two-locus associations. Compared with Marchini's method, the novelties of our method are as follows:

- We develop a stepwise strategy to detect significant two-locus associations allowing for interactions. Our stepwise strategy involves three models: one full model and two partial models. Marchini's method only considers the full model.
- We use BIC to choose the best model among the three models to test two-locus associations allowing for interactions. This model selection process is critical to control the type I error.
- Instead of working with logistic regression, we make use of the analytical solutions of some log-linear models, i.e. Equations (11, 9 and 13), so that the test statistics can be obtained in the closed form.

We have applied our method to analyze many datasets from GWAS and identified many undiscovered associations. Our method displays many advantages over existing methods:

- It can detect two-locus associations allowing for interactions from genome-wide data in a fast manner.
- It does not assume any particular epistasis model. This is very important for real studies because the patterns of SNP interactions are generally unknown and could be complex.
- It can be extended into distributed and parallel computing to analyze the phase 2 datasets from WTCCC.

One limitation of our method is that it can only detect associations between two SNPs and disease traits, while BEAM is able to detect higher order associations as well. There are several other issues that we have not addressed, such as population substructures, imputation of the missed genotypes and incorporation of covariates.

In summary, our experiment results demonstrate that our method is more powerful than existing works in detecting two-locus associations allowing for interactions. It is also computationally feasible to test hundreds of thousands of SNPs genotyped in thousands of samples.

ACKNOWLEDGEMENTS

The authors thank the anonymous reviewers for their comments that greatly helped us improve the manuscript.

Funding: Hong Kong Research Grant Council (grant GRF621707, in parts); The Hong Kong University of Science and Technology (grant RPC06/07.EG09, RPC07/08.EG25 and RPC10EG04, in parts); grant from Sir Michael and Lady Kadoorie Funded Research Into Cancer Genetics.

Conflict of Interest: none declared.

REFERENCES

Agresti,A. (2002) *Categorical Data Analysis*, 2nd edn. John Wiley & Sons, Hoboken, NJ, USA.

Balding,D. (2006) A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.*, **7**, 781–791.

Cordell,H. (2009) Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **10**, 392–404.

Engelender,S. *et al*. (2000) Organization of the human synphilin-1 gene, a candidate for parkinson's disease. *Mamm. Genome*, **11**, 763–766.

Frankel,W.N. and Schork,N. (1996) Who's afraid of epistasis? *Nat. Genet.*, **14**, 371–373.

Fung,H. *et al.* (2006) Genome-wide genotyping in parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol.*, **5**, 911–916.

Galocha,B. and de Castro,J. (2008) Folding of HLA-B27 subtypes is determined by the global effect of polymorphic residues and shows incomplete correspondence to ankylosing spondylitis. *Arthritis Rheum.*, **58**, 401–412.

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

Kullback,S. and Leibler,R.A. (1951) On information and sufficiency. *Ann. Math. Stat.*, **22**, 79–86.

Lechler,R. and Warrens,A. (2000) *HLA in Health and Disease*. Academic Press, San Diego, CA, USA.

Lerner,I. (1968) *Heredity, Evolution, and Society.* W.H. Freeman, San Francisco.

Levy,J. and Nagylaki,T. (1992) A model for the genetics of handedness. *Genetics*, **72**, 117–128.

Li,W. and Reich,J. (2000) A complete enumeration and classification of two-locus disease models. *Hum. Hered.*, **50**, 334–349.

Lin,M. *et al.* (2007) Involvement of maternal embryonic leucine zipper kinase (MELK) in mammary carcinogenesis through interaction with Bcl-G, a pro-apoptotic member of the Bcl-2 family. *Breast Cancer Res.*, **9**, R17.

Marchini,J. *et al.* (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.*, **37**, 413–417.

Michels,K. *et al.* (2003) Type 2 diabetes and subsequent incidence of breast cancer in the nurses' health study. *Diabetes Care*, **26**, 1752–1758.

Moore,J. and White,B. (2007) Tuning reliefF for genomewide genetic analysis. *Lect. Notes Comput. Sci.*, **4447**, 166–175.

Nejentsev,S. *et al*. (2007) Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. *Nature*, **450**, 887–892.

Neuman,R. and Rice,J. (1992) Two-locus models of disease. *Genet. Epidemiol.*, **9**, 347–365.

Pociot,F. and McDermott,M. (2002) Genetics of type 1 diabetes mellitus. *Genes Immun.*, **3**, 235–249.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Ritchie,M. *et al.* (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogenmetabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, **69**, 138–147.

Schwarz,D. *et al.* (2010) On safari to random jungle: a fast implementation of random forests for high dimensional data. *Bioinformatics*, **26**, 1752–1758.

Schwarz,G.F. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

Vignal,C. *et al.* (2009) Genetic association of the major histocompatibility complex with rheumatoid arthritis implicates two non-DRB1 loci. *Arthritis Rheum.*, **60**, 53–62.
Vuoksimaa,E. *et al.* (2009) Origins of handedness: a nationwide study of 30,161 adults. *Neuropsychologia*, **47**, 1294–1301.
Wan,X. *et al*. (2010) BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.*, **87** (in press).
Wan,X. *et al.* (2010) SNPRuler: predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics*, **26**, 30–37.
Williams,A. *et al.* (2007) Semaphorin 3A and 3F: key players in myelin repair in multiple sclerosis? *Brain*, **130**, 2554–2565.
WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
Zhang,Y. and Liu,J. (2007) Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet.*, **39**, 1167–1173.