Original Paper

Human Heredity

Hum Hered 2011;71:234–245
DOI: 10.1159/000328842

Received: November 5, 2010
Accepted after revision: April 12, 2011
Published online: July 6, 2011

A Dimension Reduction Approach for Modeling Multi-Locus Interaction in Case-Control Studies

Saonli Basu[a] Wei Pan[a] William S. Oetting[b]

[a]Division of Biostatistics, University of Minnesota, and [b]Department of Experimental and Clinical Pharmacology, College of Pharmacy and Institute of Human Genetics, University of Minnesota, Minneapolis, Minn., USA

Key Words
Case-control study · Gene-gene interaction · Dimension reduction

Abstract
Studying one locus or one single nucleotide polymorphism (SNP) at a time may not be sufficient to understand complex diseases because they are unlikely to result from the effect of only one SNP. Each SNP alone may have little or no effect on the risk of the disease, but together they may increase the risk substantially. Analyses focusing on individual SNPs ignore the possibility of interaction among SNPs. In this paper, we propose a parsimonious model to assess the joint effect of a group of SNPs in a case-control study. The model implements a data reduction strategy within a likelihood framework and uses a test to assess the statistical significance of the effect of the group of SNPs on the binary trait. The primary advantage of the proposed approach is that the dimension reduction technique produces a test statistic with degrees of freedom significantly lower than a multiple logistic regression with only main effects of the SNPs, and our parsimonious model can incorporate the possibility of interaction among the SNPs. Moreover, the proposed approach estimates the direction of association of each SNP with the disease and provides an estimate of the average effect of the group of SNPs positively and negatively associated with the disease in the given SNP set. We illustrate the proposed model on simulated and real data, and compare its performance with a few other existing approaches. Our proposed approach appeared to outperform the other approaches for independent SNPs in our simulation studies.

Introduction

Genetic mapping of a trait involves implementation of a number of statistical strategies to identify relative position(s) of gene(s) influencing the trait in the genome. Many complex traits of medical relevance such as diabetes, asthma, and Alzheimer's disease are controlled by multiple genes. Interaction between genes, low penetrance, and environmental factors make the gene discovery difficult for these complex traits. A common study design for genetic mapping of a trait is a case-control study design where genotype data on a large number of single nucleotide polymorphisms (SNPs) are collected

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

Accessible online at:
www.karger.com/hhe

Saonli Basu
Division of Biostatistics, School of Public Health, University of Minnesota
A 460 Mayo Building MMC 303, 420 Delaware St SE
Minneapolis, MN 55455 (USA)
Tel. +1 612 624 2135, E-Mail saonli@umn.edu

on a number of cases and controls to study the association between these SNPs and the trait, with the goal of identifying SNPs that are associated with the outcome. The usual strategy to assess the effects of the SNPs on the trait is to perform a univariate logistic regression with each SNP as a predictor, and rank the SNPs based on their p values from the univariate logistic regression analyses. The top significant SNPs which satisfy the genome-wide threshold of multiple testing are reported by the studies.

In general, the single SNP association analysis, which takes only SNPs as basic units of association analysis, has a few serious limitations. Many disease-susceptibility variants typically have only mild effects [Lesnick et al., 2007]. They might be difficult to detect due to the high threshold of being genome-wide significant. The common disease often arises from the joint action of multiple genes within a pathway. If we consider only the most significant SNPs, the genetic variants that jointly have significant risk effects but individually make only a small contribution will be missed. Moreover, attempting to understand and interpret a number of significant SNPs without any unifying biological theme can be challenging and demanding and thus might lead to poor reproducibility in the validation studies.

As an alternative strategy, one could group the SNPs together into SNP sets along the genome and perform genome-wide tests for individual SNP sets instead of individual SNPs. The SNPs could be assigned to SNP sets on the basis of some meaningful biological criteria (genomic features), e.g. genes or pathways. Then, tests for the association between each genomic feature and a disease phenotype can be performed. This gene- and pathway-based association analysis could allow us to gain insight into the functional basis of the association and facilitates to unravel the mechanisms of complex diseases [Peng et al., 2009]. Moreover, by looking at a gene or pathway as a unit of analysis, one might have better chance of detecting association since this reduces the number of tests substantially compared to the genome-wide single SNP analysis. In addition, testing the simultaneous effects of multiple SNPs by considering them jointly might improve power: individual-SNP analysis considers only the marginal effect of each SNP and therefore fails to accommodate epistatic effects.

Standard methods to evaluate the association of multiple markers with disease status are based on either single-marker analyses of the selected group of markers or multi-marker multivariate analyses. For single-SNP analysis, it is common to compare the allele frequencies of each marker between cases and controls by use of a test such as Armitage's test for trend [Sasieni, 1997] or Fisher's exact test and report a test statistic representative of the extent of association of the group of SNPs with the disease. Peng et al. [2009] discussed several such choices of test statistics. We need to adjust for multiple testing by use of either the Bonferroni correction or a permutation p value for the reported test statistic. A more comprehensive modeling approach would be to model the joint effect of the group of SNPs and test for association between the selected group of SNPs and the disease of interest.

However, using multiple logistic regression with the SNPs as predictors to model the relationship between disease status and the SNPs has some obvious limitations. As each additional main effect is included in the model, the number of possible interaction terms grows exponentially. Due to the sparseness of the data in high dimensions, parameter estimates often tend to have large standard errors, making it difficult to detect interaction. Potential model instability has led many researchers to adopt variable reduction schemes (such as backward and forward selection) to decrease the number of variables included in the model. Although this approach can be more powerful than testing each marker separately [Longmate, 2001], it still suffers from weak power because of the large number of degrees of freedom. To reduce the degrees of freedom, a set of multi-marker tests that compare pairwise genetic similarity with pairwise trait similarity among individuals were proposed by Schaid et al. [2005], Wessel and Schork [2006] and Mukhopadhyay et al. [2010]. More recently, Wu et al. [2010] proposed a computationally efficient logistic kernel-machine (KM) test that scores the similarity among individuals through different choices of the kernels and proposes a score test to detect association between the SNP set and the disease. They considered several choices of kernels such as linear (KM-Linear), identity-by-descent (KM-IBS), quadratic (KM-Quad), and two-way (KM-2way) kernels which could capture the non-linear or epistatic effects of the SNPs. We have compared the performance of our approach with the KM-approach extensively through simulation studies.

In this paper, we present a new methodology for modeling the joint effects of a group of SNPs in a case-control study. The proposed approach employs a parsimonious model to capture the effect of a group of interacting SNPs on the disease. This model is motivated by the approach adopted by Basu et al. [2009], which classifies the $2f$ founder alleles in a pedigree as high-risk or low-risk alleles, and thus avoids having separate parameter for each

founder allele and thereby reduces the number of parameters from 2*f* to 2. In the proposed approach, we have used two different scoring systems to classify the SNPs into high-risk and low-risk groups, but the flexibility of this new methodology is that many other scores can be proposed in order to capture the joint effect of the SNP set on the disease. We have also proposed a test to assess the statistical significance of the effect of the group of SNPs on a binary trait. Moreover, unlike Wu et al.'s [2010] approach, our approach could provide the estimated best model that explains the relationship between the SNPs and the disease and an estimate of the average effect of the high-risk and the low-risk SNPs for the selected best model. We have compared the performance of our approach with Wu et al. [2010] through extensive simulations and have demonstrated the superiority of the proposed approach in detecting higher-order interaction among the SNP sets.

Methods

A Latent Variable Multi-Locus Model (LVMM)

Here, we propose a parsimonious latent variable model to identify the association between a group of p ($p \geq 2$) SNPs and the trait. The model employs the data reduction strategy as originally proposed in Basu et al. [2009] that tries to address the issue of estimating large number of parameters with comparatively smaller sample size. The model also allows to incorporate the interaction among the SNPs. This approach is a likelihood-based approach and we propose a formal statistical test for the significance of the effect of the group of SNPs on the risk of a disease. Below we illustrate our model for a balanced case-control study.

Consider n individuals with binary trait data Y and marker data X on a group of p SNPs. We model the minor allele of each SNP. Each individual can have 0, 1 or 2 copies of the minor allele of each SNP. Assume that the (i, j)-th entry of the matrix X represents the number of copies of the minor allele of the j-th SNP in the i-th individual ($i = 1, ..., n$ and $j = 1, ..., p$). We assume that the minor allele of each SNP can be either one of the two types, such as 'high-risk' or 'low-risk', thereby classifying all the SNPs essentially into two categories. The low-risk allele means that the minor allele is associated with the decrease in risk, whereas high-risk implies that the allele is associated with the increase in risk of the disease. We label the high-risk SNP by 1 and the low-risk SNP by 0 (table 1).

A priori, we do not know if a minor allele is a high-risk or a low-risk allele. Hence, for p SNPs, there will be 2^p possible allocations of risk statuses. If $\mathcal{A}$ denotes a risk-label allocation to the minor alleles, then there are 2^p possible values of $\mathcal{A}$, where each $\mathcal{A}$ is a vector of 1's and 0's denoting the risk statuses of the p SNPs. This allocation of risk labels to the SNPs is equivalent to the different choices of models for the SNPs (table 1). For p SNPs, there are 2^p different choices of the models or different allocation of risk labels $\mathcal{A}$ to p SNPs. Under the null hypothesis, if there is no association between the SNPs and the trait, all these allocations would be equally likely. The biggest advantage of assigning '0' and '1' statuses to the SNPs is that the approach does not require a separate parameter for each SNP, rather it classifies all the SNPs into two groups. In order to assess the effect of the SNPs on the trait, one then requires just two parameters to represent these two classes, thereby essentially reducing the degrees of freedom required to model the effect of a group of SNPs, for example SNPs within a pathway.

Table 1. Allocations of risk labels to the minor alleles of 3 SNPs and the corresponding probability under the null hypothesis of no association between the SNPs and the trait

Allocation	Configuration			Probability
$\mathcal{A}_1$	0	0	0	0.125
$\mathcal{A}_2$	0	0	1	0.125
$\mathcal{A}_3$	0	1	0	0.125
$\mathcal{A}_4$	1	0	0	0.125
$\mathcal{A}_5$	0	1	1	0.125
$\mathcal{A}_6$	1	0	1	0.125
$\mathcal{A}_7$	1	1	0	0.125
$\mathcal{A}_8$	1	1	1	0.125

For each allocation of risk statuses, one could assign a score associated with each risk class. For example, the score could be the total number of minor alleles in each class for each individual. In that case, define Z_1 = total number of alleles in the high-risk group and Z_2 = total number of alleles in the low-risk group. We call this score M-score. The following model (equation 1) is then used to assess the effect of the group of SNPs on the trait for a specific choice of risk allocation $\mathcal{A}$.

$$\log\left(\frac{\Pr_1\left[Y|X,\mathcal{A}\right]}{1-\Pr_1\left[Y|X,\mathcal{A}\right]}\right)=\beta_1(\mathcal{A})Z_1+\beta_2(\mathcal{A})Z_2, \quad (1)$$

where Y is the binary trait data on individuals, X is the design matrix corresponding to the group of p SNPs and $\Pr_1$ is the conditional probability of Y given X and $\mathcal{A}$ under the alternative hypothesis of association between p SNPs and the trait Y. The characteristic of the high-risk group is defined by the coefficient $\beta_1(\mathcal{A})$ which is restricted to be non-negative. The characteristic of the low-risk group is defined by the coefficient $\beta_2(\mathcal{A})$ which is restricted to be non-positive. In other words, by restricting $\beta_1(\mathcal{A})$ to take non-negative values, we ensure that the SNPs in that group are associated with the increase in risk of the disease. Similarly, by restricting $\beta_2(\mathcal{A})$ to take non-positive values, we ensure that the SNPs in that group are associated with the decrease in risk of the disease. A priori, we do not know if a minor allele is a high-risk or a low-risk allele, but misclassification of an SNP in a high-risk or a low-risk group would reduce the likelihood in equation 1 for a given dataset. Therefore, our goal is to find the optimal allocation $\mathcal{A}$ that maximizes the likelihood in equation 1 and construct a test for association between the group of markers and the disease. For each possible allocation of risk statuses ($\mathcal{A}$) to the minor alleles, we can determine the values of Z_1 and Z_2 for each individual given the marker data X.

Each allocation $\mathcal{A}$ and the corresponding M-score is equivalent to a multiple logistic regression model with main effects of the SNPs and a specific choice of the direction of effect for each of the p SNPs. For example, let us consider two allocations, $\mathcal{A}_1$ = (1, 0, 0, ..., 0) and $\mathcal{A}_2$ = (1, 1, 0, ..., 0). Now, if we consider the full logistic regression main-effect model

$$\log\left(\frac{\Pr_1[Y|X]}{1-\Pr_1[Y|X]}\right)=\alpha+\gamma_1X_1+\gamma_2X_2+\cdots+\gamma_pX_p, \quad (2)$$

then allocation $\mathcal{A}_2$ and M-score is equivalent to choosing $\gamma_1 = \gamma_2 = \beta_1(\mathcal{A}_2)$ and $\gamma_3 = \gamma_4 = ... = \gamma_p = \beta_2(\mathcal{A}_2)$ and allocation $\mathcal{A}_1$ equivalent to choosing $\gamma_1 = \beta_1(\mathcal{A}_1)$ and $\gamma_2 = \gamma_3 = ... = \gamma_p = \beta_2(\mathcal{A}_1)$. We select the allocation between $\mathcal{A}_1$ and $\mathcal{A}_2$, which provides a higher likelihood of the data.

We adopted a forward selection scheme to calculate our final test statistic in equation 4 to test the null hypothesis of no association between the SNPs and the disease. We illustrate our scheme for a 3-SNP model. Table 1 illustrates the possible risk allocations for 3 SNPs. We start by computing the likelihood in equation 1 for the allocation where every minor allele belongs to the low-risk category, that is the allocation number 1 in table 1. Next, we change the group membership of the first SNP and see if the likelihood has increased. If the Likelihood increases, we modify the risk status of that SNP, otherwise we keep it the same. Then, we move to the next SNP and repeat the same process as with the first SNP. Thus, we finally select the model $\mathcal{A}$ that maximizes the test statistic specified in equation 4, among all models considered in this forward selection scheme. The motivation behind this selection scheme is that if there is an affinity among the group of low-risk alleles, changing the group membership of an allele would cause a decrease in the likelihood. On the other hand, if an allele has more affinity towards the high-risk group, changing the membership from low-risk to high-risk would increase the likelihood in equation 1. According to our forward selection scheme, we browse through $(p + 1)$ different allocations for p SNPs. The test statistic (equation 4) is obtained by maximizing over the parameters from these $(p + 1)$ different sets of allocations and for each allocation, there is the coefficient $\beta_1(\mathcal{A})$ restricted to [0, ••) and the coefficient β_2(A) restricted to (–••, 0]. We used the glm() function in R [R Development Core Team, 2005] to implement our model. In order to ensure that the two coefficients $\beta_1(\mathcal{A})$ and $\beta_2(\mathcal{A})$ are of opposite signs, we added an intercept $\beta_1(\mathcal{A}) + \beta_2(\mathcal{A})$ to the model in equation 1. For a balanced case-control study, the intercept will be zero, which ensures the opposite signs of the coefficients in equation 1.

If there is no association between the SNPs and the trait, it is equally likely to observe these different risk allocations. In other words, all models are equally likely under the null hypothesis. Hence, under the null hypothesis, $\Pr_0[Y|X, \mathcal{A}_j] = \Pr_0[Y]$, $\forall j$ whereas $\Pr_1[Y|X, \mathcal{A}_j]$ would be different from $\Pr_0[Y]$ for at least one $\mathcal{A}_j$ $(j = 1, 2, ... (p + 1))$ under the alternative hypothesis, where $\Pr_0[Y]$ is the probability distribution of the trait Y under the null hypothesis.

In terms of the parameters, the null hypothesis of no association between the group of SNPs and the trait will be equivalent to test H_0: $\beta_1(\mathcal{A}_j) = 0$, $\beta_2(\mathcal{A}) = 0$ $\forall j$ against the alternative H_1: $\beta_1(\mathcal{A}_j) \geq 0$ or $\beta_2(\mathcal{A}_j) \leq 0$ for at least one $j = 1, 2, ..., (p + 1)$. For each allocation $\mathcal{A}$, we calculate the corresponding likelihood-ratio statistic (equation 3) given by

$$\mathrm{LR}(\mathcal{A}) = -2\log\left(\frac{\max_{\beta_1(\mathcal{A}),\beta_2(\mathcal{A})}\Pr_1[Y|X,\mathcal{A}]}{\Pr_0[Y]}\right). \quad (3)$$

Note that, under the null hypothesis, $\Pr_0$ does not depend on the risk allocation $\mathcal{A}$. We used the maximum likelihood estimates $\hat{\beta}_1(\mathcal{A})$ and $\hat{\beta}_2(\mathcal{A})$ to estimate the parameters $\beta_1(\mathcal{A})$ and $\beta_2(\mathcal{A})$ of the probability distribution $\Pr_1[Y|X, \mathcal{A}]$. As mentioned before, a priori, we do not know which minor allele can be classified as high-risk or low-risk. So we adopted the forward selection scheme mentioned above and went through the $(p + 1)$ allocations. The final test statistic T is computed by taking the maximum of the individual likelihood-ratio statistic for each allocation $\mathcal{A}$:

$$\mathrm{T} = \max_{\mathcal{A}} \mathrm{LR}(\mathcal{A}) = \max_{\beta_1(\mathcal{A}),\beta_2(\mathcal{A})\forall\mathcal{A}} \frac{\Pr_1[Y|X,\mathcal{A}]}{\Pr_0[Y]}. \quad (4)$$

We estimated the distribution of the test statistic under the null hypothesis through extensive simulations and for p SNPs, the distribution was well approximated by a χ^2 distribution with $p/2$ degrees of freedom for all different numbers of SNPs considered in our simulation study.

One important advantage of this proposed approach is the flexibility in the choice of scores. For example, one could select a score that captures the higher-order interaction among the group of SNPs while keeping the degrees of freedom of the test statistic the same. Here, we have proposed such a score, the P-score that captures the interaction among the SNPs. One important thing to note here is that one could use a more advanced scoring system such as least-squares KM scorings that have been recently described for gene-level and pathway-level analyses of both expression and SNP data that allows to estimate the joint effects of SNPs [Liu et al., 2007, 2008; Kwee et al., 2008;]. For a specific value of $\mathcal{A}$, our P-score is calculated as the total number of pairs of alleles within each risk group from the marker data X. Define Z_1 = number of pairs of alleles in the high-risk group and Z_2 = number of pairs of alleles in the low-risk group. The idea behind this choice of score is that it provides a way to incorporate higher-order interaction in the model. For example, if there is a three-way interaction associated with the increase in risk of the trait, then the allocations $\mathcal{A}_j$ which assign these three SNPs to the high-risk group would give a high value of Z_1 and the likelihood $\Pr[Y|X, \mathcal{A}_j]$ would show deviation from the likelihood under no association of the SNPs and the trait. Our scoring scheme is motivated by He et al. [2010], where we showed that pair-wise scoring can significantly improve the sparsity issue faced by the MDR approach [Ritchie et al., 2001] in high-dimensional contingency tables.

In the following section, we have performed a number of simulations to compare the performance of our approach in detecting association between a set of SNPs and a disease. We have compared the power of our approach with Wu et al.'s [2010] approach. We have also conducted a number of simulations to study the performance of our approach in detecting association between a group of SNPs which are in LD with the causal SNPs. We applied our approach on a real dataset and noticed a significant difference in the findings between our approach and the single-SNP association analysis.

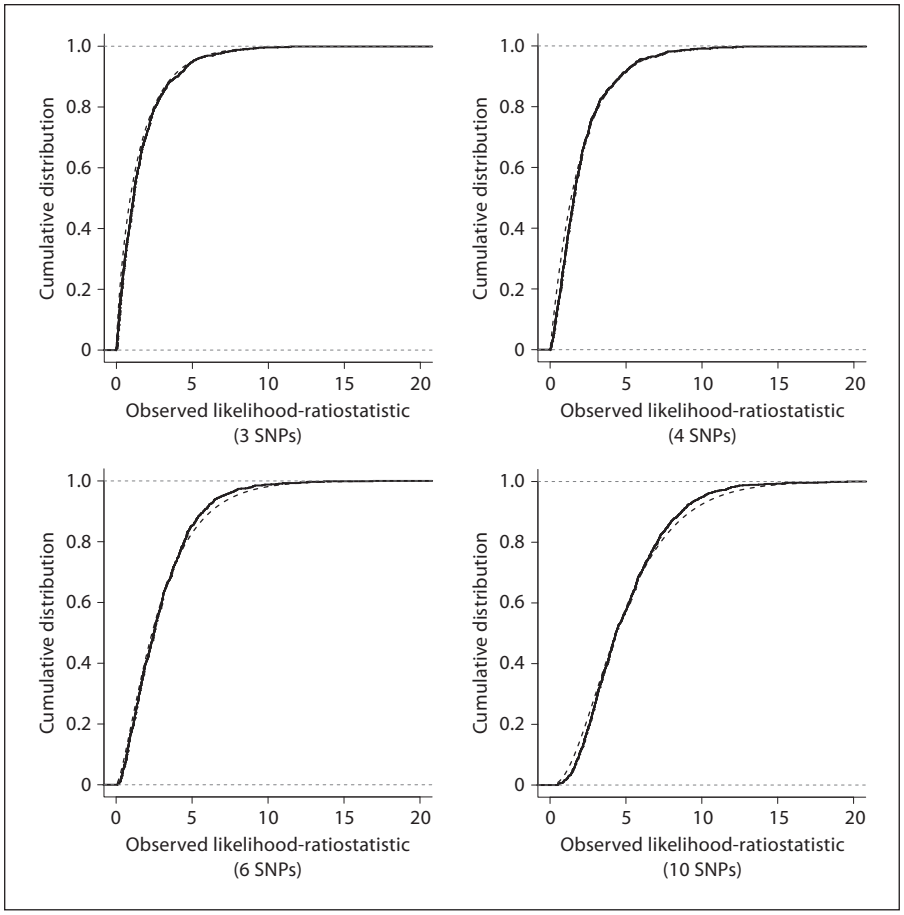

Fig. 1. Figure shows the empirical distribution of the test statistic for p SNPs (equation 4) for the proposed LVMM approach under the null hypothesis. The dashed line shows the expected CDF of the χ^2 statistic with $p/2$ degree of freedoms. The solid line shows the empirical CDF of the test statistic for 3, 4, 6 and 10 SNPs, respectively.

Results

Simulation 1

We simulated data on 200 cases and 200 controls for a different number of SNPs. We assumed that there is no LD among the SNPs. We simulated data on 3, 4, 6, and 10 SNPs under the null hypothesis that none of the SNPs are associated with the trait. We simulated 1,000 datasets for each scenario and computed the distribution of the test statistic in equation 4. We then compared the distribution of the statistic for p SNPs with the distribution of a χ^2 random variable with $p/2$ degrees of freedom. Figure 1 shows the comparison of the empirical cumulative distribution function (CDF) of the observed test statistic and the CDF of the χ^2 random variable. The estimated empirical distribution of the test statistic in equation 4 of the proposed gene-set approach matched closely with the distribution of the corresponding χ^2 statistic with $p/2$ degrees of freedom (fig. 1).

Simulation 2

We simulated 1,000 datasets from two different alternative models of phenotype-genotype association for 200 cases and 200 controls. For this simulation study, we considered multiple independent SNPs with only marginal (main) effects on the disease. We considered a 4-SNP main-effect model and a 6-SNP main-effect model for this simulation study. Each SNP had a minor allele frequency randomly drawn from a uniform distribution between 0.05–0.30. We used an additive genetic model for each SNP in this simulation study. The software PLINK [Purcell et al., 2007] was used to simulate the phenotype and the genotype data on the individuals. For the 4-SNP model, we considered SNP1 with a disease odds ratio of 1.1, SNP2 with a disease odds ratio of 0.4, SNP3 and SNP4 with a disease odds ratio of 0.6. For the 6-SNP model, 4 SNPs had a disease odds ratio of 1.2, 1 with an odds ratio of 0.3 and the 6th SNP had an odds ratio of 0.7 (fig. 2).

We compared our approach with the KM regression model proposed by Wu et al. [2010]. We also fitted a mul-

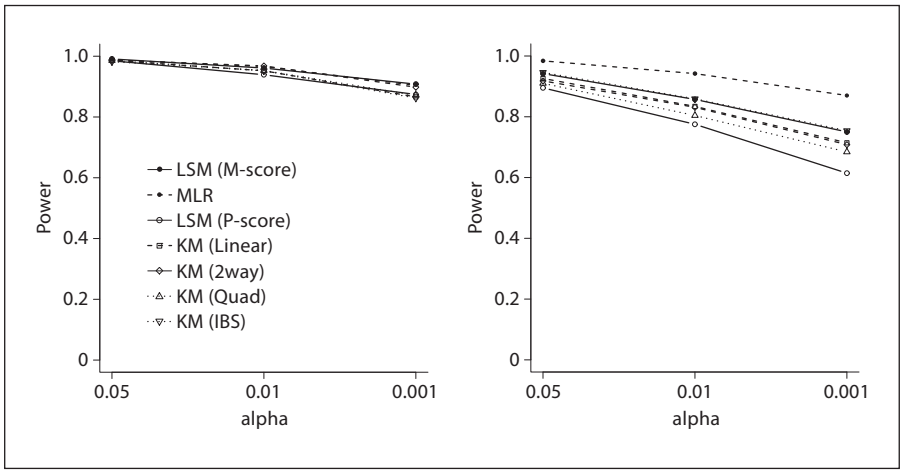

Fig. 2. The power of the KM regression [Wu et al., 2010], the MLR approach and the LVMM approach for detection of interaction under three different main-effect models (Simulation 2 in section 3.2). The power of each approach to detect the association was presented for different values of the type I error.

tiple logistic regression with the main effects of each of the 8 SNPs and performed a Wald test for the null hypothesis that none of the SNPs are associated with the disease (MLR approach). Figure 2 shows the performance of our LVMM P-score and LVMM M-score and the different scores proposed by Wu et al. [2010] to detect association for the 4-SNP and the 6-SNP model mentioned above. We reported the empirical power of all these approaches. We calculated the 95th, 99th and 999th percentile from the null distribution of each of the test statistics. The power was computed as the number of times out of 1,000 simulations. The observed test statistic under the alternative model was higher than the empirical cut-offs. As shown in figure 2, the M-score of our LVMM approach outperformed marginally the other approaches. They all had very similar power, but the added advantage of our LVMM approach is that it also provided the allocation $\mathcal{A}$ that maximized the likelihood of the observed data. Out of 1,000 simulations, the M-score identified the allocation '1000' 986 times. In other words, even with an odds ratio of 1.1, most of the times it identified the first SNP as high-risk SNP. It also correctly identified SNP3 and SNP4 as the low-risk SNPs. For the 6-SNP model, we noticed again that the LVMM M-score approach performed marginally better than the other approaches. Moreover, the LVMM M-score identified the allocation '111100' 512 times, the allocation '101100' 212 times and the '110100' 224 times out of 1,000 simulations. As expected, the performance of the LVMM M-score approach was better the LVMM P-score approach since the underlying model was a main-effect model without any interaction among the SNPs. Moreover, the power of the LVMM M-score approach was quite close to the main-effect model with a separate parameter for each SNP. Among the KM approaches, the performance of the KM-IBS approach was better than that of the other three kernels.

Simulation 3

We simulated data under three different three-way interaction models. For each model, we simulated 1,000 datasets. Each dataset contained 400 samples (200 cases and 200 controls) and 3 SNPs. We considered the following three models of interaction for simulation (fig. 3):

1. Model 1:

logit(p) = –5 + 3 I(SNP1 = Aa, SNP2 = Bb, SNP3 = Cc) + 3 I(SNP1 = AA, SNP2 = BB, SNP3 = CC) (minor allele frequency 0.1)

2. Model 2:

logit(p) = –5 + 1.2 I(SNP1 = AA, SNP2 = BB) + 1.5 I(SNP2 = BB, SNP3 = CC) + 1.8 I(SNP1 = AA, SNP3 = CC) (minor allele frequency 0.3)

3. Model 3:

logit(p) = –5 + 4 I(SNP1 = AA, SNP2 = BB, SNP3 = CC) (minor allele frequency 0.3)

Among these 1,000 simulations, we checked how many times our LVMM P-score approach identified the correct allocation, that is all three SNPs in the high-risk group. The proportion of times our approach selected the allocation (1, 1, 1) as the best allocation (i.e. it maximized the likelihood in equation 3) was 0.954, 0.812 and 0.479 for Model 1, 2 and 3, respectively. For Model 3, 623 out of 1,000 times our approach identified at least 2 SNPs in the high-risk group out of these 3 SNPs. We also compared the performance of our approach with Wu et al.'s [2010] approach and the MLR approach in terms of detecting

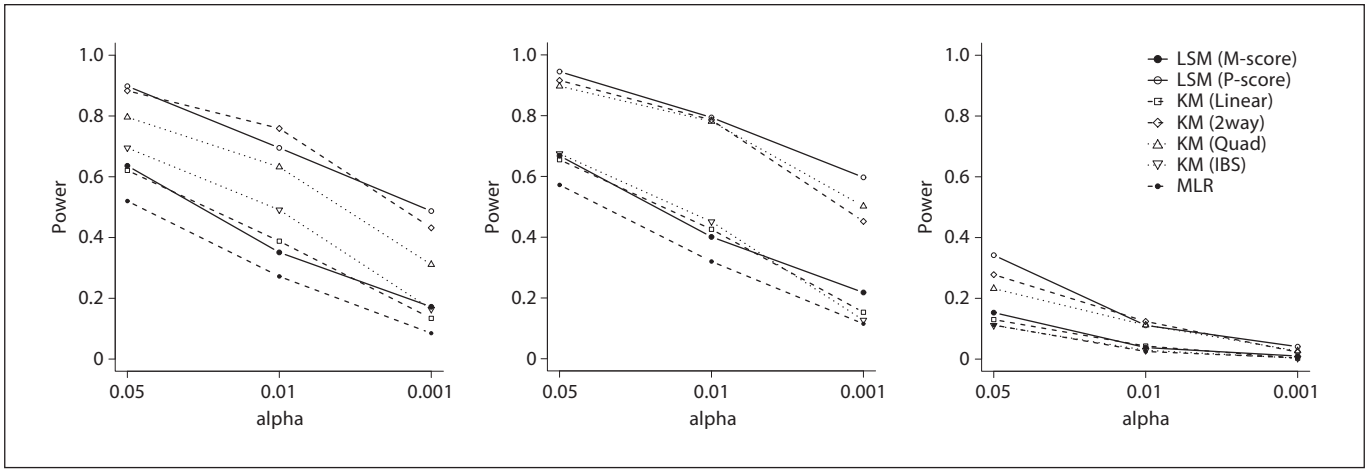

Fig. 3. The power of the KM regression [Wu et al., 2010], the MLR approach and the LVMM approach for detection of interaction under three different 3 SNP interaction models (Simulation 3 in section 3.3). The power of each approach to detect the association was presented for different values of the type I error.

Table 2. The power of KM-Linear, KM-Quad, KM-IBS, KM-2way, LVMM M-score, LVMM P-score, minP and MLR approach to detect 8 SNPs in LD at different level of type I errors

Type I error	5×10^{-3}	5×10^{-4}	5×10^{-5}
KM-Linear	0.992	0.982	0.955
KM-Quad	0.995	0.988	0.960
KM-IBS	0.991	0.982	0.953
KM-2way	0.995	0.988	0.958
LVMM M-score	0.964	0.872	0.705
LVMM P-score	0.979	0.909	0.773
minP	0.971	0.849	0.651
MLR	0.958	0.863	0.702

this set of 3 SNPs associated with the disease. Figure 3 shows the power to detect association for our LVMM approach and Wu et al.'s [2010] KM regression model for various choices of the kernel. Our approach with the P-score generally outperformed the KM regression approach for all these 3 models. The power of the LVMM approach was substantially higher for the pair-wise interaction model (Model 2).

Simulation 4

We performed a simulation study to compare the performance of the approaches when the causal SNP(s) is (are) not observed, but in LD with the observed set of SNPs. We followed similar set-ups as given in Wang and Elston [2007] and Pan et al. [2010] with k = 10 marker SNPs and a sample size of 400 (200 cases and 200 controls). First, we generated a latent vector of length 10 from a multivariate normal distribution with mean 0 and variance 1 and the covariance structure AR-1 with the correlation $\rho_{ij} = 0.8^{|i-j|}$. We then discretized each variate of this simulated latent vector with values 0, 1 and 2 if the absolute value of the variate was $<Z_{0.9}$, between $Z_{0.9}$ and $Z_{0.98}$ and $\geq Z_{0.98}$, respectively, where $Z_{0.9}$ and $Z_{0.98}$ are the 90th and 98th quantile of the standard normal distribution, respectively. Each of these transformed variates represented an SNP. The correlation structure introduced LD among these 10 SNPs. Our simulation also ensured the HWE and an expected allele frequency of 0.2 for each SNP. We then simulated the phenotype data from a multiple logistic regression with SNP1 and SNP8 as the causal SNPs. We considered an additive genetic model for the SNPs. The following model was used for the simulation of the phenotype data (table 2):

$$\text{logit}(p) = -4 + 2\ \text{SNP1} \times \text{SNP8}. \quad (5)$$

We then removed these 2 causal SNPs and analyzed the dataset of 200 cases and 200 controls with 8 SNPs. We considered three different approaches for testing if the SNP set is associated with the disease. We ran a single-SNP analysis with these 8 SNPs and used Bonferroni correction to adjust the p values of these SNPs. We considered the minimum of these Bonferroni adjusted p values (minP approach); in table 2, the number of times the p value was $\leq 5 \times 10^{-3}$, 5×10^{-4} and 5×10^{-5} out of 1,000

simulations is shown. We also performed our LVMM approach with M-score and P-score. In table 2, we show the asymptotic p value of the test statistic by comparing it with χ^2 distribution with four degrees of freedom. We also used the kernel logistic regression approach [Wu et al., 2010] to test for the significance of joint effect of these 8 SNPs on the disease. Table 2 shows how many times out of 1,000 simulations the p value of each test was $\leq 5 \times 10^{-3}$, 5×10^{-4} and 5×10^{-5}. We also fitted a multiple logistic regression with the main effects of each of the 8 SNPs and performed a Wald test for the null hypothesis that none of the SNPs are associated with the disease (MLR approach). In table 2, the number of times the p value of the Wald test was $\leq 5 \times 10^{-3}$, 5×10^{-4} and 5×10^{-5} out of 1,000 simulations is shown. Both the MLR approach and the LVMM M-score model the main effects of all the SNPs. The power of the MLR approach was very similar to that of the LVMM M-score approach. The KM approach generally performed better since it takes into account the LD among the SNPs and adjusts the degrees of freedom accordingly. The LVMM P-score performed well and had significantly better power than the MLR approach. All the approaches that model the joint effect of the SNPs had better power than the minP approach.

Simulation 5

In this simulation study, we simulated data from a HAPMAP CEU population. We selected the CYP3A4 gene from the ADME (absorption, distribution, metabolism and excretion) pathway for our simulation study, which is one of the genes we studied in our real data analysis. This gene, CYP3A4, encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins are monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. We considered a total of 18 tag SNPs from this CYP3A4 gene for our simulation study. We first simulated genotypes on all 18 SNPs using HAPGEN2 (http://www.well.ox.ac.uk/zhan/hapgen/hapgen2.html) and then considered an interaction model specified in Marchini et al. [2005] and used the R package simulateDiscretePhenotypes in HAPGEN2 to simulate phenotype data. We used the twoSnpInteractionModel2() function to simulate interaction between two SNPs. We then removed these two SNPs from the SNP list and considered an SNP set of 16 SNPs for our simulation. We simulated 200 cases and 200 controls for this comparison.

We compared the performance of all the methods mentioned in Simulation 4 by simulating 1,000 datasets under this 2-SNP interaction model. The power of all the methods at various levels of type I error is shown in table 3. The KM approach performed well in this simulation study as well. In general, the minP approach performed better than the MLR approach. The LVMM approach was affected by the LD in the dataset and had lower power than the KM approach, but the performance of the LVMM M-score approach was substantially better than the MLR approach and the KM-2way approach had similar performance as the LVMM M-score approach (table 3).

Table 3. The power of KM-Linear, KM-Quad, KM-IBS, KM-2way, LVMM M-score, LVMM P-score, minP and MLR approach to detect the 8 SNPs in Simulation 5 at different level of type I errors

Type I error	0.01	0.005	0.001
KM-Linear	0.695	0.607	0.419
KM-Quad	0.650	0.486	0.409
KM-IBS	0.724	0.602	0.452
KM-2way	0.635	0.482	0.412
LVMM M-score	0.620	0.510	0.392
LVMM P-score	0.614	0.474	0.359
minP	0.615	0.506	0.372
MLR	0.520	0.461	0.308

Real Data Analysis

The limitation of many existing approaches for pathway-based analysis [Wang et al., 2007; Inada et al., 2008; Peng et al., 2009] is that they only focus on the main effects of the SNPs and do not incorporate the interaction among the multiple SNPs within each pathway. Moreover, these approaches are completely nonparametric and hence cannot evaluate the significance of the joint effect of the group of SNPs within each pathway. Our parsimonious model in equation 1 provides an alternative approach to conduct a pathway-based analysis. The parameters $\beta_1(\mathcal{A})$ and $\beta_2(\mathcal{A})$ in equation 1 for the LVMM M-score approach measure average effect size of a high-risk SNP and a low-risk SNP, respectively, within each pathway. The following illustrates an implementation of pathway-based analysis using our proposed approach. We also implemented the KM approach, the minP approach and the MLR approach described in Simulation 4 to compare the findings of these approaches on a real dataset.

We studied the performance of the above mentioned approaches to detect genetic association of acute rejection (AR) in kidney transplant patients. Whole blood was ob-

Table 4. The significant subgroups of the ADME pathways reported by the different approaches

Pathway sub-group	$\hat{\beta}_1$ (SD)	$\hat{\beta}_2$ (SD)	SNP n	LVMM M-score (asym.)	LVMM M-score (permutation)	LVMM P-score (asym.)	LVMM P-score (permutation)	KM-Linear	KM-Quad	KM-IBS	KM-2way	minP	MLR
ATP-binding cassette, subfamily B (MDR/TAP)	0.176 (0.03)	−0.265 (0.05)	10	0.002	3×10^{-4}	0.003	1×10^{-4}	1×10^{-4}	4.69×10^{-5}	4.66×10^{-6}	2.57×10^{-5}	0.005	0.001
ATP-binding cassette, subfamily C (CFTR/MRP)	0.535 (0.03)	−0.248 (0.04)	16	0.124	0.022	0.302	0.144	0.085	0.530	0.002	0.531	0.485	0.246
Cytochrome P450, family 17, sub-family A	0	−0.498 (0.02)	3	0.029	0.009	0.052	0.011	0.009	0.009	0.008	0.010	0.032	0.067
Cytochrome P450, family 1, sub-family A	0	−0.510 (0.04)	4	0.091	0.037	0.152	0.008	0.018	0.032	0.024	0.032	0.059	0.077
Cytochrome P450, family 2, sub-family E	0.282 (0.03)	−0.515 (0.06)	6	0.078	0.021	0.103	0.025	0.008	0.009	0.013	0.009	0.09	0.13
UDP glycosyl-transferase 1 family	0.280 (0.01)	−0.239 (0.04)	5	0.021	0.001	0.088	0.007	0.004	0.010	0.0009	0.010	0.006	0.010

tained with informed consent and DNA isolated from 271 kidney allograft recipients, 136 of whom had AR within 6 months of transplant, and 135 of whom did not have any detectable AR after at least 8 years post-transplant. All received Ab induction and CNI, with either MMF or sirolimus. DNA variants were genotyped using a Affymetrix custom genotyping chip containing 3,590 SNPs, many of which are thought to be functional variants within biologically relevant genes to AR including genes in pathways associated with immunity, cell signaling, ADME, cell growth and proliferation [Van Ness et al., 2008]. Genotyping was performed using the Affymetrix GeneChip Scanner 3000 Targeted Genotyping System (GCS 3000 TG System), which utilizes molecular inversion probes to simultaneously identify the 3,404 pre-selected SNPs. Methods for genotyping have been previously described and were performed in strict adherence to the manufacturer's protocol [Hardenbol et al., 2003]. For this comparison study, we constructed a balanced case-control study by randomly selecting 120 Caucasian patients with AR within 6 months of transplant, and 120 Caucasian patients without any detectable AR after at least 8 years post-transplant.

Of the 3,404 SNPs typed, 80 SNPs had no data and were hence excluded from the analysis. Of the remaining 3,324 SNPs, the call rate was 98.6%. Our goal here was to detect any evidence of interaction among the SNPs associated with AR in kidney allografts. Among these SNPs, we excluded those SNPs which have minor allele frequency <5%. We also excluded those SNPs which have more than 10% missing values. We then imputed the missing data for each SNP from the observed genotype distribution. For the remaining SNPs, we did Fisher's exact test and selected only the SNPs with p value <0.1 for the interaction detection purpose. Among these 340 SNPs, we only considered the SNPs in ADME pathways.

A major advancement in reducing the risk of AR in kidney allograft recipients was the introduction of immunosuppressant drugs including cyclosporine A and tacrolimus, both calcineurin inhibitors, and sirolimus, which binds to the mammalian target of the rapamycin complex 1 (mTORC1) [Kaufman et al., 2004; Meier-Kriesche et al., 2006]. How well these drugs protect the recipient from AR is due largely to how the patient absorbs and metabolizes the drug. These drugs demonstrate large inter-patient variation in pharmacokinetic (PK) parameters [Press et al., 2009]. Variation in the PK

of these drugs can result in altered serum concentrations that can reduce drug efficacy or result in toxicity. There are various proteins that alter absorbance and metabolism and are members of the ADME pathways. Few members of this pathway have been associated with immunosuppressant PK variation [Wang et al., 2009]. According to Kuypers et al. [2008], these members include hepatic cytochrome P450 3A5 (CYP3A5) and an ATP-driven efflux pump (multidrug resistance 1; MDR1) (table 4).

We implemented all the approaches on 44 different subgroups of SNPs in the ADME pathway. In table 3, only the subgroups with a reported p value ≤ 0.01 by at least one of the approaches are shown. For the LVMM approach, we reported the p values from the χ^2 approximation with appropriate degrees of freedom. We also conducted a permutation test with 10,000 permutations and the p values are shown in table 3. The LVMM approach performed substantially better than the minP and the MLR approach. We also reported the average effect sizes of a high-risk and a low-risk SNP within each SNP set as produced by the LVMM M-score approach. The KM approaches also performed really well on this real dataset. Their findings were in strong agreement with the LVMM approach. For the significant pathways, the p values of the KM approaches were generally lower the LVMM approach.

The subgroup ATP-binding cassette (sub-family B) showed a significant association with AR. The sub-family B of the ATP-binding cassette had genes such as ABCB1 and ABCB11. Both ABCB1 and ABCB11 are involved in small molecule transport and have broad substrate specificity [Sharom, 2008]. These proteins are also involved in the transport and distribution of immunosuppressants used in kidney transplantation, and reports have shown that variation within these proteins is associated with kidney transplant outcomes, especially when calcineurin inhibitors are used [Naesens et al., 2009].

Discussion

Common diseases often arise from joint action of multiple genes within a pathway. A pathway consists of a group of interacting components acting in concert to perform specific biological tasks. Although each single SNP may confer small disease risk, their joint actions will play a significant role in the development of disease. If we only consider the most significant SNPs, the genetic variants that jointly have significant risk effects, but individually make only a small contribution, will be missed. We have proposed a new methodology for assessing joint effects of a group of SNPs, incorporating the possibility of interaction among SNPs. We have implemented the data reduction strategy [Basu et al., 2009] within a likelihood framework and used a test to assess the statistical significance of a group of SNPs. The proposed approach appeared to outperform other existing approaches for independent SNPs in our simulation studies. Moreover, our proposed approach provides an estimate of average effect of the high-risk alleles in the SNP set and an estimate of the average effect of the low-risk alleles in the SNP set. Our approach also provides the best allocation or model that explains the relationship between the SNP set and the disease.

One big assumption for this LVMM approach is that it assumes that all the SNPs within each risk group have the same effect size. We investigated the performance of our proposed approach through simulation studies when this assumption is violated. In our simulation studies, for example Simulation 2, the proposed approach performed quite well as compared to the multiple logistic regression model with separate parameters for each SNP effect. Moreover, our proposed approach is computationally faster and would certainly be more useful when one deals with a large group of SNPs.

The proposed approach is designed to test for joint effects of a group of SNPs with a disease. It would get affected if there are a lot of null SNPs within the group. Hence, we need to use a pre-screening technique, for example restricting the SNP set to the SNPs below certain p value cut-offs before implementing our approach. All methods would be affected to certain degree if there are a lot of null or nearly-null SNPs in the dataset. We have investigated the effect of the null SNPs on the LVMM approach (simulations not shown here) and it appears that this approach is more sensitive to the inclusion of null SNPs as compared to the KM approach. We intend to investigate this further and propose a data-adaptive approach which would better deal with null SNPs.

The asymptotic null distribution of our proposed test statistic for p SNPs seemed to be well approximated by a χ^2 distribution with $p/2$ degrees of freedom in our simulation studies. We intend to explore this more to theoretically derive the asymptotic null distribution of the proposed test statistic. Another possible approach would be to propose a test statistic by averaging over all allocations rather than taking the maximum over the allocations. This would be similar to model averaging and the advantage of this approach is that if there are multiple possible allocations that explain the nature of association

between the SNP set and disease, the test statistic would gain power by averaging over all the allocations.

The proposed LVMM approach assigns the SNPs into high-risk and low-risk groups and provides an estimate of the average effect of each group. This strategy might be useful to model the rare variants where, due to low frequency of each SNP, it is not possible to estimate individual SNP effect.

When applying the methods for detection of multilocus effects, the presence of missing observations reduces the number of observations available in the analysis. The most appropriate approach at present is to use only subjects with complete observations. However, as the number of genotypes increases, the number of subjects with complete observations decreases rapidly. Currently, there are several approaches to handle missing data and they can be used to implement any method to detect gene-gene interaction in presence of missing data [Namkung et al., 2009]. One solution to handle this situation is to impute missing observations, which we did for our real data analysis.

The proposed null distribution of the test statistic T (χ^2 with $p/2$ degrees of freedom for p SNPs) can be conservative if the p SNPs within an SNP set are correlated, i.e. if there is LD among them. In Simulation 4, we see the impact of LD among SNPs on the power of the LVMM approach. To address this issue, one could perform a permutation test or some re-simulation-based approach to estimate the null distribution of the test statistic, as we did in our real data analysis. It is computationally very feasible to perform a permutation test within our LVMM framework, but in the worst situation, the null distribution approximation (χ^2 with $p/2$ degrees of freedom for p SNPs) can be used as a screening procedure to speed up the process of SNP-set selection in pathway-based genome-wide association studies.

Acknowledgement

This research was supported by NIH grant R21DK089351.

References

Basu S, Stephens M, Pankow JS, Thompson EA: A likelihood-based trait-model-free approach for linkage detection of binary trait. Biometrics 2009;66:205–213.

Hardenbol P, Baner J, Jain M, Nilsson M, Namsaraev EA, Karlin-Neumann GA, Fakhrai-Rad H, Ronaghi M, Willis TD, Landegren U, Davis R: Multiplexed genotyping with sequence-tagged molecular inversion probes. Nat Biotechnol 2003;21:673–678.

He H, Oetting WS, Brott MJ, Basu S: Pairwise multifactor dimensionality reduction method to detect gene-gene interactions in a case-control study. Hum Hered 2010;69:60–70.

Inada T, Koga M, Ishiguro H, Horiuchi Y, Syu A, Yoshio T, Takahashi N, Ozaki N, Arinami T: Pathway-based association analysis of genome-wide screening data suggest that genes associated with the gamma-aminobutyric acid receptor signaling pathway are involved in neuroleptic-induced, treatment-resistant tardive dyskinesia. Pharmacogenet Genomics 2008;18:317–323.

Kaufman DB, Shapiro R, Lucey MR, Cherikh WS, Bustami RT, Dyke DB: Immunosuppression: practice and trends. Am J Transplant 2004;4:38–53.

Kuypers DR, de Jonge H, Naesens M, Vanrenterghem Y: Effects of cyp3a5 and mdr1 single nucleotide polymorphisms on drug interactions between tacrolimus and fluconazole in renal allograft recipients. Pharmacogenet Genomics 2008;18:861–868.

Kwee L, Liu D, Lin X, Ghosh D, Epstein MP: A powerful and flexible multilocus association test for quantitative traits. Am J Hum Genet 2008;82:386–397.

Lesnick TG, Papapetropoulos S, Mash DC, Ffrench-Mullen J, Shehadeh L, de Andrade M, Henley JR, Rocca WA, Ahlskog JE, Maraganore DM: A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. PLoS Genet 3. E98.

Liu D, Ghosh D, Lin X: Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics 2008;9:292.

Liu D, Lin X, Ghosh D: Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics 2007;63:1079–1088.

Longmate J: Complexity and power in case-control association studies. Am J Hum Genet 2001;68:1229–1237.

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

Meier-Kriesche HU, Li S, Gruessner RW, Fung JJ, Bustami RT, Barr ML, Leichtman AB: Immunosuppression: evolution in practice and trends, 1994–2004. Am J Transplant 2006;6: 1111–1131.

Mukhopadhyay I, Feingold E, Weeks D, Thalamuthu A: Association tests using kernel-based measures of multi-locus genotype similarity between individuals. Genet Epidemiol 2010;34:213–221.

Naesens M, Kuypers D, Sarwal M: Calcineurin inhibitor nephrotoxicity. Clin J Am Soc Nephrol 2009;4:481–508.

Namkung J, Elston RC, Yang JM, Park T: Identification of gene-gene interactions in the presence of missing data using the multifactor dimensionality reduction method. Genet Epidemiol 2009;33:646–656.

Pan W, Han F, Shen X: Test selection with application to detecting disease association with multiple SNPs. Hum Hered 2010;69:120–130.

Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, Zhao J, Zhou X, Reveille JD, Jin L, Amos CI: Xiong M: Gene and pathway-based second-wave analysis of genome-wide association studies. Eur J Hum Genet 2009;18:111–117.

Press RR, Ploeger BA, den Hartigh J, van der Straaten T, van Pelt J, Danhof M, de Fijter JW, Guchelaar HJ: Explaining variability in tacrolimus pharmacokinetics to optimize early exposure in adult kidney transplant recipients. Ther Drug Monit 2009;31:187–197.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: Plink: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–575.

R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2005.
Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: Multifactor-dimensionality reduction reveals highorder interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 2001;69:138–147.
Sasieni P: From genotype to genes: doubling the sample size. Biometrics 1997;53:1253–1261.
Schaid D, McDonnell S, Hebbring S, Cunningham J, Thibodeau S: Nonparametric tests of association of multiple genes with human disease. Am J Hum Genet 2005;76:780–793.
Sharom F: Abc multidrug transporters: structure, function and role in chemoresistance. Pharmacogenomics 2008;9:105–127.
Van Ness B, Ramos C, Haznadar M, Hoering A, Haessler J, Crowley J, Jacobus S, Oken M, Rajkumar V, Greipp P, Barlogie B, Durie B, Katz M, Atluri G, Fang G, Gupta R, Steinbach M, Kumar V, Mushlin R, Johnson D, Morgan G: Genomic variation in myeloma: design, content, and initial application of the Bank On A Cure SNP Panel to detect associations with progression-free survival. BMC Med 2008;6:26.
Wang K, Li M, Bucan M: Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet 2007;81:1278–1283.
Wang T, Elston R: Improved power by use of a weighted score test for linkage disequilibrium mapping. Am J Hum Genet 2007;80:353–360.
Wang Y, Wang C, Li J, Wang X, Zhu G, Chen X, Bi H, Huang M: Effect of genetic polymorphisms of cyp3a5 and mdr1 on cyclosporine concentration during the early stage after renal transplantation in chinese patients cotreated with diltiazem. Eur J Clin Pharmacol 2009;65:239–247.
Wessel J, Schork N: Generalized genomic distance-based regression methodology for multilocus association analysis. Am J Hum Genet 2006;79:792–806.
Wu M, Kraft P, Epstein M, Taylor D, Chanock S, Hunter D, Lin X: Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet 2010;86:929–942.