Published in final edited form as:

Stat Methods Med Res. 2018 May ; 27(5): 1464–1475. doi:10.1177/0962280216662071.

A multi-locus genetic association test for a dichotomous trait and its secondary phenotype

Han Zhang¹, Colin O Wu², Yifan Yang³, Sonja I Berndt¹, Stephen J Chanock¹, and Kai Yu¹

¹Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, USA

²Office of Biostatistics Research, National Heart, Lung and Blood Institute, Bethesda, USA

³Department of Statistics, University of Kentucky, Lexington, USA

Abstract

Genetic association studies often collect information on secondary phenotypes related to the primary disease status. In many situations, the secondary phenotypes are only measured in subjects with the disease condition. It would be advantageous to model the primary trait and the secondary phenotype together if they share certain level of genetic heritability. We propose a family of multi-locus testing procedures to detect the composite association between a set of genetic markers and two traits (the primary trait and a secondary phenotype), in order to identify genes influencing both traits. The proposed test is derived from a random effect model with two variance components, with each presenting the genetic effect on one trait, and incorporates a model selection procedure for seeking the optimal model to represent the two sources of genetic effects. We conduct simulation studies to evaluate performance of the proposed procedure and apply the method to a genome-wide association study of prostate cancer with the Gleason score as the secondary phenotype.

Keywords

Secondary phenotype; multi-locus test; variance component; genome-wide association study; multiple testing; prostate cancer

1 Introduction

Population-based genetic association studies have been widely used for uncovering the genetic basis underlying complex diseases. Although they are typically designed to study one primary trait, information on other secondary phenotypes is often collected and is potentially valuable for the study of the primary trait. For example, besides knowing the disease status of each subject in a genetic association study of breast cancer, we might also have additional information measured on breast cancer tumor tissues, which provides more details on pathologic and molecular characteristics of the disease. Those secondary

Corresponding author: Kai Yu, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20850, USA., yuka@mail.nih.gov.

Declaration of conflicting interests

The author(s) received no financial support for the research, authorship, and/or publication of this article.

phenotypes could be helpful in identifying the disease susceptibility loci if they share certain level of genetic heritability with the primary trait.

Recently, Wu et al.¹ proposed a single-marker testing framework to assess the association between a genetic marker and two traits simultaneously in situations where the secondary phenotype is quantitative and is only measured on subjects in a particular primary trait-dependent stratum. For example, the secondary phenotype might be only available on the subjects with disease condition. The data can be collected prospectively or retrospectively with the primary trait being the disease status. Their method aims at detecting genetic markers associated with both traits and maintaining robust power even if the marker is associated with only one of traits.

Although the single-marker test has been the most commonly used approach in detecting genetic susceptibility loci, increasing evidence has suggested that multiple correlated markers within a gene could jointly influence complex diseases.² A multi-locus test that aggregates association evidence across multiple genetic markers in a considered gene or a genomic region may be more powerful than a single-marker test.^{3–6} Here, we focus on the similar setting considered by Wu et al.¹ and derive a class of multi-locus tests for the association between a set of genetic markers within a considered gene and two traits. The proposed test extends the sequence kernel association test (SKAT)⁷ to a random effect model with two variance components, with each presenting a genetic effect on one trait, and incorporates a model selection procedure for seeking the optimal model to represent the two sources of genetic effects.

Many existing multi-locus tests require complete observations on the set of genetic markers. Under the current single-nucleotide polymorphism (SNP) genotyping technology, the genotype missing rate at a given SNP is very low. However, the proportion of subjects with at least one missing genotypes within a considered genomic region could still be high if the number of SNPs in the region is relatively large. Excluding those subjects with at least one missing genotype can reduce the sample size substantially and thus diminish the power of the multi-locus test. Using statistical imputation algorithms^{8–10} to impute the missing genotype is a commonly used strategy to retain the sample, but it has several limitations. For example, it requires the knowledge of haplotype distribution on the studying population. Furthermore, it is known that the imputed genotype on a SNP with a relatively low minor allele frequency (MAF) is not very accurate.¹¹ To make the proposed test more flexible in practice, we generalize the test so that it can handle missing genotypes without resorting to imputation or removing samples.

2 Model

For a study with n samples, let $\{D_i, Z_i, X_i, G_i\}$ be the observed data on the i th sample, with D_i being the primary dichotomous trait (e.g., disease status), Z_i being the secondary quantitative trait, X_i being the set of covariates to be adjusted, and G_i being the vector of genotypes on m genetic markers in a given gene or region. We assume that the genotype is coded as 0, 1, and 2, representing the number of minor alleles at a given marker. Other coding schemes can be dealt with similarly. We will describe our method for data sampled

from a prospective cohort study, where the secondary phenotype Z_i is only available on subjects with $D_i = 1$. We then extend the application of our method to a retrospective case-control study, wherein Z_i is collected in cases. We refer to our method as MAPS, i.e., a *M*ultimulti-locus *A*ssociation test for a dichotomous *P*rimarily trait and a quantitative *S*econdary phenotype.

2.1 A random effect model for a prospective cohort study

In a prospective cohort study, we assume that the dichotomous trait D_i can be modeled by the following logistic regression model given the covariates and genotypes of multiple markers within a gene

$$\text{logitPr}(D_i = 1 | X_i, G_i) = X_i^T \alpha + G_i^T \beta, i = 1, \dots, n \quad (1)$$

with the intercept term absorbed in covariates by adding a column of 1s in X_i . We also assume that Z_i on a subject with $D_i = 1$ follows the normal distribution $\mathcal{N}(X_i^T \gamma + G_i^T \theta, \sigma^2)$. The likelihood of observing the data $\{D_i, Z_i, X_i, G_i, i = 1, \dots, n\}$ can be represented as

$$L(\beta, \theta, \alpha, \gamma, \sigma^2) = \prod_{i=1}^n \left[\frac{\exp\{X_i^T \alpha + G_i^T \beta\}}{1 + \exp\{X_i^T \alpha + G_i^T \beta\}} \phi\left(\frac{Z_i - X_i^T \gamma - G_i^T \theta}{\sigma}\right) \right]^{D_i} \left[\frac{1}{1 + \exp\{X_i^T \alpha + G_i^T \beta\}} \right]^{1-D_i} \quad (2)$$

where $\phi(\cdot)$ is the density function of the standard normal distribution. Let

$$\ell^D(\beta, \alpha) = \sum_{i=1}^n \left[D_i (X_i^T \alpha + G_i^T \beta) - \log(1 + \exp\{X_i^T \alpha + G_i^T \beta\}) \right]$$

and

$$\ell^Z(\theta, \gamma, \sigma^2) = \sum_{i=1}^n D_i \log \phi\left(\frac{Z_i - X_i^T \gamma - G_i^T \theta}{\sigma}\right)$$

therefore, the log-likelihood $\ell(\beta, \theta, \alpha, \gamma, \sigma^2) = \ell^D(\beta, \alpha) + \ell^Z(\theta, \gamma, \sigma^2)$. Denote $S_D = \frac{\partial \ell^D}{\partial \beta}$, $S_Z = \frac{\partial \ell^Z}{\partial \theta}$, $S_{DD} = \frac{\partial^2 \ell^D}{\partial \beta \partial \beta^T}$, and $S_{ZZ} = \frac{\partial^2 \ell^Z}{\partial \theta \partial \theta^T}$. It can be shown through a second-order Taylor expansion that the likelihood can be approximated around $(\beta, \theta) = (\mathbf{0}, \mathbf{0})$ as

$$\exp\{\ell(\beta, \theta, \alpha, \gamma, \sigma^2)\} \approx \exp\{\ell^D(\mathbf{0}, \alpha) + \ell^Z(\mathbf{0}, \gamma, \sigma^2)\} \left\{ 1 + S_D^T \beta + S_Z^T \theta + \frac{1}{2} (\beta^T, \theta^T) W \begin{pmatrix} \beta \\ \theta \end{pmatrix} \right\}$$

where $\mathbf{0}$ is a zero vector of length m , and

$$W = \begin{pmatrix} S_D S_D^T & S_D S_Z^T \\ S_Z S_D^T & S_Z S_Z^T \end{pmatrix} + \begin{pmatrix} S_{DD} & 0 \\ 0 & S_{ZZ} \end{pmatrix}$$

To derive a variance component test for the null hypothesis $H_0: \beta = \theta = \mathbf{0}$, we further assume that the genetic effects (β, θ) are random effects with $E\beta = E\theta = \mathbf{0}$, and

$$\text{Cov}\begin{pmatrix} \beta \\ \theta \end{pmatrix} = \tau \times \begin{pmatrix} \kappa I & \rho\sqrt{\kappa(1-\kappa)}I \\ \rho\sqrt{\kappa(1-\kappa)}I & (1-\kappa)I \end{pmatrix}$$

where I is the $m \times m$ identity matrix and the scalars $\tau \geq 0$, $\rho \in [-1, 1]$, $\kappa \in [0, 1]$. Here, the variance-covariance matrix is configured by three parameters under the following assumptions. First, genetic effects β on the primary trait are independent and identically distributed (i.i.d.) random variables with variance $\tau\kappa$. Second, genetic effects θ on the secondary phenotype are i.i.d. random variables with variance $\tau(1-\kappa)$. Third, the genetic effects from one marker on the two traits are correlated with correlation coefficient ρ . Fourth, the genetic effects from two markers on either the same or different traits are uncorrelated. One can see that testing the joint genetic effects on the two traits is equivalent to testing $H_0: \tau = 0$. Similar to Lin,¹² we can obtain the profile log-likelihood in term of variance component parameters (τ, ρ, κ) by integrating out β and θ

$$\begin{aligned} \tilde{\ell}(\tau, \rho, \kappa, \alpha, \gamma, \sigma^2) &= \log E_{\beta, \theta} \exp\{\ell(\beta, \theta, \alpha, \gamma, \sigma^2)\} \\ &\approx \ell^D(\mathbf{0}, \alpha) + \ell^Z(\mathbf{0}, \gamma, \sigma^2) + \frac{1}{2} \text{tr}\{W \text{Cov}\begin{pmatrix} \beta \\ \theta \end{pmatrix}\} \end{aligned} \quad (3)$$

Let $(\hat{\alpha}, \hat{\gamma}, \hat{\sigma}^2)$ be the maximum likelihood estimates of $\tilde{\ell}$ under the null. Define

$$Q_{\rho, \kappa} = \kappa \hat{S}_D^T \hat{S}_D + (1 - \kappa) \hat{S}_D^T \hat{S}_Z + 2\rho\sqrt{\kappa(1-\kappa)} \hat{S}_D^T \hat{S}_Z \quad (4)$$

where $\hat{S}_D$ (or $\hat{S}_Z$) is the score S_D (or S_Z) evaluated at $(\hat{\alpha}, \hat{\gamma}, \hat{\sigma}^2)$. The asymptotic null distributions of $\hat{S}_D$ and $\hat{S}_Z$ are multivariate normal distributions, with means $\mathbf{0}$ and estimated variance-covariance matrices $\hat{S}_{DD}$ and $\hat{S}_{ZZ}$, respectively. The score for τ at $\tau = 0$ is

$$\left. \frac{d\tilde{\ell}}{d\tau} \right|_{\tau=0} = \frac{1}{2} Q_{\rho, \kappa} + \frac{1}{2} \text{tr}\{\kappa \hat{S}_{DD} + (1 - \kappa) \hat{S}_{ZZ}\} \quad (5)$$

Note that the second term in equation (5) converges to some constant in probability for given (ρ, κ) , we thus can conduct a family of variance component tests based on $Q_{\rho, \kappa}$ only.

For any given (ρ, κ) , denote $p_{\rho, \kappa} = \Pr(Q_{\rho, \kappa} \geq \tilde{Q}_{\rho, \kappa})$ as the p -value of $Q_{\rho, \kappa}$ evaluated at its observed value $\tilde{Q}_{\rho, \kappa}$.

Since (ρ, κ) are unknown, we propose to define the statistic for testing $H_0 : \tau = 0$ as

$$T = \min_{\rho, \kappa} p_{\rho, \kappa},$$

which measures the strongest evidence of the presence of association with (ρ, κ) turned in proper regions. The final p -value adjusted for multiple comparisons is computed from the null distribution of T .

In the following sections, we will introduce different versions of variance component tests based on T with possible choices of the tuning parameters ρ and κ . Numerical algorithms for computing the final p -value are also discussed.

2.2 The variance component test with $\rho = 0$ and $\kappa = 1/2$

One simple choice of the tuning parameters in T is to set $\rho = 0$ and $\kappa = 1/2$, which essentially assumes that the genetic effects distribute equally on either traits for each marker. The statistical significance can be evaluated by checking the distribution of

$Q_{0, 1/2} \propto \hat{S}_D^T \hat{S}_D + \hat{S}_Z^T \hat{S}_Z$, which follows a mixture of chi-square distributions under the null. Several existing algorithms are available for computing the distribution function of $Q_{0, 1/2}$, thus the p -value can be calculated accurately.^{13–15} This test is referred as MAPS_{0, 1/2}.

2.3 The variance component test with $\rho = 0$

A more flexible approach is to fix $\rho = 0$ while allowing κ to vary in $[0, 1]$. The test statistic becomes $T = \min_{\kappa \in [0, 1]} p_{0, \kappa}$. In practice, we can choose κ at the grids $\{k/20 : k = 0, \dots, 20\}$. For any given κ

$$Q_{0, \kappa} = \kappa \hat{S}_D^T \hat{S}_D + (1 - \kappa) \hat{S}_Z^T \hat{S}_Z.$$

Notice that S_D is asymptotically independent with S_Z since $\frac{\partial^2 \ell}{\partial \beta \partial \theta^T} = 0$, the final p -value of T defined can be computed explicitly by an one-dimensional numerical integration algorithm. The details are given in Appendix 1. This test is referred as MAPS₀.

2.4 The variance component test with variable ρ and κ

In real application, we usually do not have any prior knowledge on the values of ρ and κ . A robust approach is to maximize the association evidence over $(\rho, \kappa) \in [-1, 1] \times [0, 1]$, we define the statistic as $T = \min_{(\rho, \kappa) \in [-1, 1] \times [0, 1]} p_{\rho, \kappa}$. In practice, we can choose (ρ, κ) the grids $\{j/10; j = -10, \dots, 10\} * \{k/20 : k = 0, \dots, 20\}$. To assess the significance of T , we can generate the scores of $\hat{S}_D$ and $\hat{S}_Z$ under the null via the direct simulation approach.¹⁶ The final p -value of T is then estimated through the computationally efficient minP algorithm.¹⁷

We refer this optimal test as MAPS_{opt} . As a special case, the MAPS_{opt} test with κ fixed to be $1/2$, and ρ tuned in $[-1, 1]$, is referred as MAPS_{cor} . The p -value of MAPS_{cor} can be computed similarly as MAPS_{opt} .

2.5 Existing approaches

There are several alternative approaches that are applicable to the setting considered in this paper. The SKAT has been successfully applied in identifying genetic regions associated with complex diseases.⁷ In the following discussion, the SKAT tests applied to either the dichotomous or quantitative trait are referred as SKAT_D and SKAT_Z , respectively. In addition, the standard likelihood ratio test (LRT), which compares the additive model consisting of all the genetic markers with the null model, can be applied to each trait separately, leading to two tests LRT_D and LRT_Z , respectively. These two tests may loss power due to large degree-of-freedom. Finally, we generalize the single-marker test in Wu et al.¹ to a multi-locus score test. This generalized score test follows χ^2_{2m} distribution under the null.

2.6 A random effect model for a retrospective case-control study

In a case-control study, we assume the quantitative trait Z is only observed in cases. Then the likelihood of observed data $\{D_i, Z_i, X_i, G_i : i = 1, \dots, n\}$ can be written as

$$L(\beta, \theta, \alpha, \gamma, \sigma^2) = \prod_{i=1}^n \Pr(Z_i | X_i, G_i, D_i = 1)^{D_i} \Pr(X_i, G_i | D_i = 1)^{D_i} \Pr(X_i, G_i | D_i = 0)^{1-D_i} \quad (6)$$

According to Qin and Zhang,¹⁸ the joint distribution of X_i and G_i satisfies

$$\Pr(X_i, G_i | D_i = 1) \propto \exp\{X_i^T \alpha + G_i^T \beta\} \Pr(X_i, G_i | D_i = 0)$$

if the risk model is assumed as the logistic regression model in equation (1). Ignoring a constant, the profile likelihood of equation (6) is equivalent to the likelihood equation (2) in a cohort study.¹⁸ Therefore, all the tests discussed in previous subsections can be applied to case-control studies.

2.7 Missing data in multi-locus test

In the above, we have described the method assuming no missing genotypes at any considered genetic markers. In real application, we might have a substantial proportion of individuals who have at least one missing genotype in the considered region, especially when the region consists of a large number of markers. Removing those subjects can result in substantial loss of power. To make full use of observed genotypes, we propose to use following modified score statistics defined on observed genotypes.

Without loss of generality, we consider the generalized linear model $EY = g^{-1}(X^T \alpha + G^T \beta)$ and assume that the covariates X are observed in full dataset S with sample size n . Other nuisance parameters (e.g., variance parameter σ^2 for quantitative trait Z), if any, are denoted

as ψ . The n_j individuals without missing genotypes on the j th marker are indexed as $S_j, j = 1, 2, \dots, m$, where m is the number of markers within the considered region or gene. Denote the log-likelihood as $\ell = \ell(\alpha, \beta, \psi)$ and let $\ell_\alpha = \frac{\partial \ell}{\partial \alpha}, \ell_\beta = \frac{\partial \ell}{\partial \beta}, \ell_{\alpha\alpha} = \frac{\partial^2 \ell}{\partial \alpha \partial \alpha}, \ell_{\beta\beta} = \frac{\partial^2 \ell}{\partial \beta \partial \beta},$ and

$\ell_{\beta\alpha} = \frac{\partial^2 \ell}{\partial \beta \partial \alpha}$. A superscript j on these defined term means only individuals in S_j are used. For

example, $\ell_{\beta_j}^j$ is the score of β_j defined on S_j . In contrast, the score of α can be defined on

either S (i.e., ℓ_α) or S_j (i.e., ℓ_α^j). Similarly, superscript jk means individuals in $S_j \cap S_k$ are used.

Let $(\hat{\alpha}^j, \hat{\psi}^j)$ be the maximum likelihood estimates of (α, ψ) using S_j under the null. Statistics denoted with accent $\sim$ is assessed at $(\hat{\alpha}, \hat{\psi})$ (e.g., $\hat{\ell}_{\beta_j}^j = \ell_{\beta_j}^j(\hat{\alpha}, \mathbf{0}, \hat{\psi}^j)$). We show in the Appendix

1 that, under the assumption of missing at complete randomness the modified score

$\hat{\ell}_\beta = \left\{ \hat{\ell}_{\beta_j}^j : j = 1, 2, \dots, p \right\} = \left\{ \hat{\ell}_{\beta_j}^j(\hat{\alpha}^j, \mathbf{0}, \hat{\psi}^j) : j = 1, 2, \dots, p \right\}$ asymptotically follows multivariate

normal distribution with means $\mathbf{0}$. The covariance between $\hat{\ell}_{\beta_j}^j$ and $\hat{\ell}_{\beta_k}^k$ can be consistently

estimated by $\nu_{jk} = \hat{\ell}_{\beta_j \beta_k}^{jk} + \frac{nm_{jk}}{n_j n_k} \hat{\ell}_{\beta_j \alpha}^j \hat{\ell}_{\alpha \alpha}^{-1} \hat{\ell}_{\alpha \beta_k}^k$, where n_{jk} is the sample size of $S_j \cap S_k$.

Replacing $\hat{S}_D$ and $\hat{S}_Z$ in equation (4) by the modified score $\hat{\ell}_\beta$ allows the proposed method to handle data with missing genotypes, where $(\nu_{jk})_{m \times m}$ is used as the modified variance–covariance matrix in the direct simulation for the evaluation of the p -value.

In this procedure, the score $\hat{\ell}_{\beta_j}^j$ uses all the observed genotypes at the j th SNP. The

covariance between the score $\hat{\ell}_{\beta_j}^j$ and $\hat{\ell}_{\beta_k}^k$ is estimated using information on n_{jk} subjects,

which is very close to the original total sample size if the proportion of missing genotype at each marker is low. Thus, this strategy is much more efficient than the one that requires the removal of subjects who have at least one missing genotypes on the set of considered markers.

This procedure is very general for handling missing genotypes in various multi-locus tests, as long as the test is based on score statistics derived from the generalized linear model. We therefore integrated this procedure into SKAT, LRT, and Wu's method, so that they can be applied to the real data application described below.

3 Simulation studies

We evaluated performance of the proposed variance component tests through simulation studies with genes generated under various of linkage disequilibrium (LD) structures. Similar to Wang and Elston,¹⁹ we considered a gene consisting of 20 SNPs and a study with 500 cases and 500 controls. To generate genotypes on the 20 SNPs, we first simulated continuous random variables $R = (R_1, \dots, R_{20})$ from a multivariate normal distribution with mean zero and a variance–covariance matrix $\Sigma = (\sigma_{ij})_{20 \times 20}$, where $\sigma_{ij} = r^{|i-j|}$. By properly

choosing cut-points, we then discretized R_i into a three-level genotype with levels 0, 1, and 2, so that the corresponding SNP had a MAF of 0.4. The LD within this gene was controlled by the parameter r , which was chosen as either 0 or 0.6 in our simulation. We used this algorithm to generate genotypes in controls. We assumed the risk model for the primary trait (case-control status) has the following form

$$\text{logit } \Pr(D = 1 | G) = \beta G_{10} + \beta G_{11}$$

with the 10th and 11th SNPs conferring the risk of disease. In order to simplify the simulation, we assumed the equality of the two odds ratios. Under the given risk model, we used the weighted sampling procedure¹⁸ to generate genotypes in cases from the following distribution

$$\Pr(G_1, \dots, G_{20} | D = 1) \propto \exp\{\beta G_{10} + \beta G_{11}\} \Pr(G_1, \dots, G_{20} | D = 0)$$

Within the stratum of $D = 1$, the secondary quantitative trait was simulated from

$$Z = \theta G_{10} + \theta G_{11} + \mathcal{N}(0, 1),$$

with the same risk SNPs as in the risk model of the dichotomous trait.

We also investigated the robustness of our method by additional simulations, in which the risk SNPs of the primary trait and the secondary trait were different. More specifically, $\text{logit } \Pr(D = 1 | G) = \beta G_9 + \beta G_{10}$, and $Z = \theta G_{11} + \theta G_{12} + \mathcal{N}(0, 1)$.

3.1 Type I error

We evaluated the type I errors of all tests proposed in this paper, as well as the existing approaches discussed in Section 2.5. Only the results of scenario with $r = 0.6$ are presented here. 100,000 datasets were generated under the null by setting $\beta = \theta = 0$, each with 500 cases and 500 controls. p -values of MAPS_{opt} and MAPS_{cor} were first estimated with 100,000 resampling steps. Then for those datasets with initial estimates of p -values 5×10^{-4} , more accurate estimates were obtained with 1,000,000 resampling steps. Table 1 shows that all tests can properly control the type I errors at nominal levels of 0.01, 0.001, and 0.0001.

3.2 Empirical power

To compare the empirical powers of various tests under different LD structure configured by r , we chose β and θ so that the empirical powers of LRT_D and LRT_Z were close to specified powers (p_D, p_Z) with type I error controlled at the level of 0.01. We set (p_D, p_Z) as (0.2, 0.4), (0.3, 0.3), (0.4, 0.2), (0.6, 0.01), or (0.01, 0.6) to represent different scenarios. For example, when $(p_D, p_Z) = (0.6, 0.01)$ or (0.01, 0.6), the gene has moderate effects on both traits. When $(p_D, p_Z) = (0.6, 0.01)$ or (0.01, 0.6), the gene influenced only one trait. We used 10,000 resampling steps to evaluate p -values for MAPS_{opt} and MAPS_{cor} .

In Table 2, we summarize the empirical powers, each of which is based on 1000 simulated datasets. We can see from the table that, when the same SNPs in a causal gene influence both traits, $MAPS_{opt}$ and $MAPS_{cor}$ have the best performance among all considered tests. When the gene is associated with only one trait, $MAPS_0$ and $MAPS_{opt}$ are the most robust tests among others variance component tests, with $MAPS_0$ is slightly more powerful than $MAPS_{opt}$ due to less model selection penalty. $MAPS_{cor}$ is very sensitive to the underlying risk model. For example, its power is less than 1/3 of that of $MAPS_{opt}$ when the gene is only associated with primary trait. When two traits are influenced by different causal SNPs, the method extended from Wu et al.¹ is more powerful than other methods if all SNPs are in linkage equilibrium ($r = 0$). When SNPs are moderately correlated ($r = 0.6$), $MAPS_{opt}$ is the most robust test in discovering composite gene association.

In Figure 1, we show the optimal (ρ, κ) corresponding to $\min_{\rho, \kappa} p_{\rho, \kappa}$ for each simulated dataset, in which SNPs in a gene are moderately correlated ($r = 0.6$), and genotypes at causal SNPs shared by the two risk models are directly observed. When $MAPS_{opt}$ detects a significant association in practice (e.g., $p < 0.01$), a selected κ that is very close to 0 or 1 suggests that the gene under study is likely associated with only one trait.

We also compared the empirical powers among tests when the causal SNPs are not directly observed. The parameter r controlling the LD structure was set at 0.6. All other settings were similar to those used in previous simulation with full observations. The results are summarized in Table 3. $MAPS_{opt}$ again appears to have the most robust performance among all considered tests, especially when the gene is associated with both traits.

4. Application to a genome-wide association study of prostate cancer

We demonstrated the application of MAPS as multi-locus tests by applying them on a genome-wide association study (GWAS) of prostate cancer. We focused on 2841 controls and 4544 cases of European ancestry.²⁰ For each prostate cancer case, we used the Gleason score (2–10), which indicates how likely it is that a tumor will spread, as a quantitative trait. We hypothesized that there are genes influencing the mechanism underlying the development of prostate cancer, as well as how fast the tumor cells spread. By looking at the two traits jointly (i.e., prostate cancer status and Gleason score), we intend to increase our chance for detecting that type of genes.

Of the SNPs genotyped using the Illumina HumanOmni2.5 BeadChip, 1,531,807 passed standard quality control criteria.²⁰ We extracted SNPs within 20 kb upstream and 20 kb downstream of a gene or an annotated region. The SNPs with missing rate $> 2\%$ or MAFs $< 2\%$ were excluded from the analyses. For two SNPs with LD coefficient $r^2 > 0.95$, the one with a smaller MAF was discarded. Both traits were adjusted for center, age, and two eigenvectors.

We will provide more detailed report on the analysis of over 20,000 genes/regions elsewhere. Here, we are interested in the 69 genes with both p -values of $SKAT_Z$ and $SKAT_D$ less than 0.05, as using tests analyzing two traits jointly are most likely to be beneficial on those genes. In Table 4, we showed results of nine genes on which there were at least one

gene-level p -value less than 0.001 by all considered two-trait joint tests. Among those, KLK3 and CLDN11 are known risk genes associated with the prostate cancer in population with European ancestry.^{21,22} IRX4 has only been identified to be associated with the risk of prostate cancer in Japanese population.^{23,24} Although the other six genes have not been reported in GWAS as genes susceptible to prostate cancer risk, overexpression of PIAS3 was known to induce apoptosis in prostate cancer cells.²⁵ The forest plots in Figures 2 and 3 illustrate the marginal effects from each SNP in genes LOC643201 and PIAS3 on the prostate cancer risk and the Gleason score. It shows several SNPs in either gene are associated with both traits. This is the main reason why the joint test approach appears to be more advantageous than the single-trait test approaches. We can consider those genes in Table 4 as promising candidates underlying the development of prostate cancer, although further replications are needed.

5 Discussion

Although genetic association studies are typically designed to study one primary trait, valuable information on other secondary phenotypes is often collected. There is a growing interest to study secondary phenotypes using already measured genotypes. Several approaches have been developed to identify genetic markers associated with secondary phenotypes, taking account for the design of the original study.^{26–30} The proposed method has a different goal. It analyzes the primary trait and a secondary phenotype jointly, aiming at detecting genes influencing both traits. The family of proposed tests is derived from a random effect model with two variance components, with each presenting the genetic effect on one trait. Among its various versions, we found the one that uses observed data to adaptively model the variance–covariance matrix of genetic effects has the most robust performance. We demonstrated the application of the new method by applying it to analyze a GWAS of prostate cancer and identified several promising novel regions that appeared to influence the risk and progression of prostate cancer. An R package of the proposed test is available at <https://github.com/zhangh12/MAPS>

It has been shown that multi-locus association tests can be a valuable alternative to the commonly used single-marker test. There are many multi-locus approaches for genetic association studies,^{6,7,31} most of which assume that there is no missing genotype. As a result, genotype imputation is usually needed before using the multi-locus test. However, it is not a trivial task to impute the missing genotypes, and the imputation accuracy depends on the reference genome.¹¹ The strategy for dealing with missing genotypes proposed with our test is more flexible and easy to use, as it does not need imputing missing genotypes. Also, it uses all observed genotypes, and thus is more efficient than the strategy that requires the removal of subjects with at least one missing genotypes on the set of considered markers. This strategy is especially helpful for GWAS, where genotype missing rate at a given SNP is very low. With some modifications, this strategy can be adapted to other multi-locus tests defined by score statistics.^{7,16,31}

In our method, we use three parameters to model the variance–covariance matrix for the genetic effects on a primary trait and a secondary phenotype. To extend the method to study more than two secondary phenotypes, it is important to find an appropriate model for the

variance–covariance matrix. Using too many parameters would increase the penalty for model selection and reduce efficiency of the multi-locus test. On the other hand, an over-simplified model can introduce bias into the testing procedure. Further investigations are needed to extend the proposed method to study more secondary phenotypes.

Acknowledgements

The authors would like to thank Professor Hua Liang at The George Washington University for his helpful comments. This study utilized the high-performance computational capabilities of compute cluster at the Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland.

Appendix 1

p-value of variance component test with $\rho = 0$

We derive formula of the *p*-values of MAPS₀. Let q_{κ} be the upper 100 × $p_{0,\kappa}$ % quantile of $Q_{0,\kappa}$, where $Q_{0,\kappa}$ follows a null distribution as a central mixture chi-square with non-zero weights. Under the null, $\hat{S}_D^T \hat{S}_D$ and $\hat{S}_Z^T \hat{S}_Z$ follow central mixture chi-square distributions with corresponding weights.³¹ Therefore

$$\begin{aligned} \Pr(T \leq \tilde{T}) &= 1 - \Pr(T > \tilde{T}) \\ &= 1 - \Pr(Q_{0,\kappa} < q_{\kappa}, \quad \forall_{\kappa}) \\ &= 1 - \Pr(\hat{S}_D^T \hat{S}_D < \min_{\kappa > 0} \frac{q_{\kappa} - (1 - \kappa) \hat{S}_Z^T \hat{S}_Z}{\kappa} \text{ and } \hat{S}_Z^T \hat{S}_Z < q_0) \\ &= 1 - \int_0^{q_0} \left(\min_{\kappa > 0} \frac{q_{\kappa} - (1 - \kappa)t}{\kappa} \right) f_Z(t) dt \end{aligned}$$

where $F_D(\cdot)$ is the cumulative distribution function of $\hat{S}_D^T \hat{S}_D$ and $f_Z(\cdot)$ is the probability density function of $\hat{S}_Z^T \hat{S}_Z$.

Asymptotic distribution of modified score allowing missing data

We derive modified score to deal with missing genotypes. By Taylor's Theorem and law of large numbers

$$\hat{\ell}_{\beta_j}^j \approx \ell_{\beta_j}^j - I_{\beta_j, \alpha} I_{\alpha \alpha}^{-1}, \quad j = 1, 2, \dots, p$$

where I is the Fisher information matrix. Thus

$$\begin{aligned}
\text{Cov}\left(\hat{\epsilon}_{\beta_j}^j, \hat{\epsilon}_{\beta_k}^k\right) &= \text{Cov}\left(\epsilon_{\beta_j}^j, \epsilon_{\beta_k}^k\right) + I_{\beta_j \alpha} I_{\alpha \alpha}^{-1} \text{Cov}\left(\epsilon_{\alpha}^j, \epsilon_{\alpha}^k\right) I_{\alpha \alpha}^{-1} I_{\alpha \beta_k} \\
&\quad - I_{\beta_j \alpha} I_{\alpha \alpha}^{-1} \text{Cov}\left(\epsilon_{\alpha}^j, \epsilon_{\alpha}^k\right) - \text{Cov}\left(\epsilon_{\beta_j}^j, \epsilon_{\beta_k}^k\right) I_{\alpha \alpha}^{-1} I_{\alpha \beta_k} \\
&= n_{jk} I_{\beta_j \beta_k} + I_{\beta_j \alpha} I_{\alpha \alpha}^{-1} (n_{jk} I_{\alpha \alpha}) I_{\alpha \alpha}^{-1} I_{\alpha \beta_k} \\
&\quad - I_{\beta_j \alpha} I_{\alpha \alpha}^{-1} (n_{jk} I_{\alpha \beta_k}) - (n_{jk} I_{\beta_j \alpha}) I_{\alpha \alpha}^{-1} I_{\alpha \beta_k} \\
&= n_{jk} \left(I_{\beta_j \beta_k} - I_{\beta_j \alpha} I_{\alpha \alpha}^{-1} I_{\alpha \beta_k} \right)
\end{aligned}$$

The information $I_{\beta_j \beta_k}$, $I_{\beta_j \alpha}$, $I_{\alpha \alpha}$ and $I_{\alpha \beta_k}$ can be consistently estimated by

$-n_{jk}^{-1} \ell_{\beta_j \beta_k}(\hat{\alpha}^{jk}, 0, \hat{\phi}^{jk})$, $-n_j^{-1} \ell_{\beta_j \alpha}(\hat{\alpha}^j, 0, \hat{\phi}^j)$, $-n^{-1} \ell_{\alpha \alpha}(\hat{\alpha}, 0, \hat{\phi})$ and $-n_k^{-1} \ell_{\alpha \beta_k}(\hat{\alpha}^k, 0, \hat{\phi}^k)$,

respectively. Therefore, we can estimate the covariance between scores of two genetic

effects by $-\hat{\epsilon}_{\beta_j \beta_k}^{jk} + \frac{n_{jk}}{n_j n_k} \hat{\epsilon}_{\beta_j \alpha}^j \hat{\epsilon}_{\alpha \alpha}^{-1} \hat{\epsilon}_{\alpha \beta_k}^k$. Note that we here use the most informative estimates

of (β, β) by using as much samples as possible.

References

1. Wu CO, Zheng G and Kwak M. A joint regression analysis for genetic association studies with outcome stratified samples. *Biometrics* 2013; 69: 417–426. [PubMed: 23489105]
2. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* 2012; 44: 369–375. [PubMed: 22426310]
3. Liu JZ, Mcrae AF, Nyholt DR, et al. A versatile gene-based test for genome-wide association studies. *Am J Hum Genet* 2010; 87: 139–145. [PubMed: 20598278]
4. Huang H, Chanda P, Alonso A, et al. Gene-based tests of association. *PLoS Genet* 2011; 7: e1002177. [PubMed: 21829371]
5. Li MX, Gui HS, Kwan JS, et al. GATES: a rapid and powerful gene-based association test using extended sines procedure. *Am J Hum Genet* 2011; 88: 283–293. [PubMed: 21397060]
6. Zhang H, Wheeler W, Wang Z, et al. A fast and powerful tree-based association test for detecting complex joint effects in case-control studies. *Bioinformatics* 2014; 30: 2171–2178. [PubMed: 24794927]
7. Wu MC, Kraft P, Epstein MP, et al. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* 2010; 86: 929–942. [PubMed: 20560208]
8. Browning BL and Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 2010; 84: 210–223.
9. Li Y, Willer CJ, Ding J, et al. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 2010; 34: 816–834. [PubMed: 21058334]
10. Howie B, Fuchsberger C, Stephens M, et al. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* 2012; 44: 955–959. [PubMed: 22820512]
11. Wang Z, Jacobs KB, Yeager M, et al. Improved imputation of common and uncommon SNPs with a new reference set. *Nat Genet* 2012; 44: 6–7.
12. Lin X Variance component testing in generalised linear models with random effects. *Biometrika* 1997; 84: 309–326.
13. Davies RB. The distribution of a linear combination of x2 random variables. *J R Stat Soc C* 1980; 29: 323–333.

14. Kuonen D Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* 1999; 86: 929–935.
15. Liu H, Tang Y and Zhang HH. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computat Stat Data Anal* 2009; 53: 853–856.
16. Zhang H, Shi J, Liang F, et al. A fast multilocus test with adaptive SNP selection for large-scale genetic-association studies. *Eur J Hum Genet* 2014; 22: 696–702. [PubMed: 24022295]
17. Ge Y, Dudoit S and Speed TP. Resampling-based multiple testing for microarray data analysis. *Test* 2003; 12: 1–77.
18. Qin J and Zhang B. A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* 1997; 84: 609–618.
19. Wang T and Elston RC. Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am J Hum Genet* 2007; 80: 353–360. [PubMed: 17236140]
20. Berndt SI, Wang Z, Yeager M, et al. Two susceptibility loci identified for prostate cancer aggressiveness. *Nat Commun* 2015; 6: 6889. [PubMed: 25939597]
21. Eeles RA, Kote-Jarai Z, Giles GG, et al. Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Genet* 2008; 40: 316–321. [PubMed: 18264097]
22. Kote-Jarai Z, Olama AAA, Giles GG, et al. Seven novel prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. *Nat Genet* 2011; 43: 785–791. [PubMed: 21743467]
23. Takata R, Akamatsu S, Kubo M, et al. Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population. *Nat Genet* 2010; 42: 751–754. [PubMed: 20676098]
24. Nakagawa H, Akamatsu S, Takata R, et al. Prostate cancer genomics, biology, and risk assessment through genome-wide association studies. *Cancer Sci* 2012; 103: 607–613. [PubMed: 22181854]
25. Wible BA, Wang L, Kuryshv YA, et al. Increased K⁺ efflux and apoptosis induced by the potassium channel modulatory protein KChAP/PIAS3 β in prostate cancer cells. *J Biol Chem* 2002; 277: 17852–17862. [PubMed: 11877452]
26. Lin D and Zeng D. Proper analysis of secondary phenotype data in case-control association studies. *Genet Epidemiol* 2009; 33: 256–265. [PubMed: 19051285]
27. Monsees GM, Tamimi RM and Kraft P. Genome-wide association scans for secondary traits using case-control samples. *Genet Epidemiol* 2009; 33: 717–728. [PubMed: 19365863]
28. He J, Li H, Edmondson AC, et al. A Gaussian copula approach for the analysis of secondary phenotypes in case-control genetic association studies. *Biostatistics* 2012; 13: 497–508. [PubMed: 21933777]
29. Li H and Gail MH. Efficient adaptively weighted analysis of secondary phenotypes in case-control genome-wide association studies. *Hum Hered* 2012; 73: 159–173. [PubMed: 22710642]
30. Schifano ED, Li L, Christiani DC, et al. Genome-wide association analysis for multiple continuous secondary phenotypes. *Am J Hum Genet* 2013; 92: 744–759. [PubMed: 23643383]
31. Lee S, Wu MC and Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 2012; 13: 762–775. [PubMed: 22699862]

Figure 1.

ρ and κ selected by MAPS_{opt} in simulation studies, in which two risk models share the causal SNPs that are directly observed. The SNPs in a gene are moderately correlated ($r = 0.6$). Solid points: p -values of $\text{MAPS}_{opt} < 0.01$. Circle: p -values of $\text{MAPS}_{opt} > 0.01$.

Figure 2.
The forest plot of 16 SNPs within gene LOC643201.

Figure 3.
The forest plot of 12 SNPs within gene PIAS3.

Table 1.

Type I error based on 100,000 datasets generated from the null $H_0 : \beta = \theta = \mathbf{0}$.

Level	MAPS _{opt}	MAPS _{cor}	MAPS ₀	MAPS _{0,1/2}	Wu	SKAT _D	SKAT _Z	LRT _D	LRT _Z
0.01	0.00960	0.00933	0.00902	0.01039	0.00944	0.00999	0.00967	0.01169	0.01231
0.001	0.00099	0.00094	0.00093	0.00099	0.00083	0.00093	0.00098	0.00126	0.00134
0.0001	0.00010	0.00008	0.00012	0.00005	0.00009	0.00010	0.00008	0.00010	0.00024

Table 2.

Empirical power comparison when causal SNPs are observed.

(r, β, θ)	MAPS _{opt}	MAPS _{cor}	MAPS ₀	MAPS _{0,1/2}	Wu	SKAT _D	SKAT _Z	LRT _D	LRT _Z
Two risk models share the same causal SNPs									
(0.0, 0.21, 0.18)	0.686	0.702	0.531	0.584	0.561	0.232	0.421	0.209	0.404
(0.6, 0.17, 0.15)	0.853	0.897	0.767	0.809	0.542	0.404	0.646	0.193	0.397
(0.0, 0.23, 0.17)	0.672	0.688	0.524	0.527	0.559	0.318	0.320	0.304	0.298
(0.6, 0.19, 0.14)	0.857	0.889	0.770	0.789	0.582	0.523	0.552	0.302	0.315
(0.0, 0.26, 0.15)	0.656	0.626	0.515	0.447	0.537	0.433	0.200	0.409	0.194
(0.6, 0.21, 0.12)	0.836	0.853	0.776	0.747	0.584	0.651	0.407	0.409	0.215
Gene associates with one trait									
(0.0, 0.30, 0.00)	0.478	0.145	0.512	0.151	0.392	0.619	0.010	0.593	0.008
(0.6, 0.24, 0.00)	0.726	0.350	0.746	0.376	0.418	0.820	0.007	0.607	0.011
(0.0, 0.00, 0.22)	0.458	0.475	0.496	0.539	0.391	0.004	0.601	0.004	0.600
(0.6, 0.00, 0.18)	0.726	0.760	0.752	0.794	0.391	0.001	0.821	0.007	0.599
Two risk models contain different causal SNPs									
(0.0, 0.21, 0.18)	0.492	0.522	0.533	0.576	0.580	0.246	0.409	0.227	0.395
(0.6, 0.17, 0.15)	0.774	0.804	0.758	0.803	0.555	0.384	0.651	0.187	0.404
(0.0, 0.23, 0.17)	0.479	0.464	0.517	0.515	0.570	0.317	0.307	0.301	0.302
(0.6, 0.19, 0.14)	0.779	0.778	0.758	0.762	0.569	0.524	0.547	0.293	0.307
(0.0, 0.26, 0.15)	0.480	0.398	0.515	0.453	0.561	0.445	0.190	0.409	0.191
(0.6, 0.21, 0.12)	0.766	0.738	0.760	0.727	0.564	0.638	0.381	0.411	0.217

Table 3.Empirical power comparison when causal SNPs are not directly observed ($r = 0.6$).

(β, θ)	MAPS _{opt}	MAPS _{cor}	MAPS ₀	MAPS _{0,1/2}	Wu	SKAT _D	SKAT _Z	LRT _D	LRT _Z
Two risk models share the same causal SNPs									
(0.27, 0.25)	0.780	0.803	0.674	0.724	0.558	0.295	0.566	0.205	0.401
(0.31, 0.23)	0.781	0.781	0.675	0.689	0.568	0.440	0.453	0.300	0.293
(0.34, 0.20)	0.791	0.765	0.679	0.639	0.551	0.572	0.325	0.402	0.201
Gene associates with one trait									
(0.39, 0.00)	0.604	0.243	0.650	0.270	0.372	0.755	0.009	0.599	0.014
(0.00, 0.30)	0.613	0.626	0.635	0.686	0.397	0.003	0.738	0.011	0.598
Two risk models contain different causal SNPs									
(0.34, 0.32)	0.593	0.630	0.573	0.639	0.542	0.252	0.464	0.199	0.389
(0.38, 0.30)	0.604	0.616	0.581	0.617	0.560	0.360	0.373	0.291	0.309
(0.42, 0.26)	0.620	0.588	0.594	0.578	0.566	0.473	0.273	0.398	0.213

Table 4.

The suggestive genes with at least one gene-level p -value $< 10^{-3}$, and the p -values of SKAT_D and SKAT_Z are both < 0.05 .

Gene	MAPS _{opt}	MAPS _{cor}	MAPS ₀	MAPS _{0,1/2}	Wu	SKAT _D	SKAT _Z	LRT _D	LRT _Z
SENP6	3.0E-5	2.3E-4	3.7E-5	1.3E-3	2.9E-2	3.7E-5	4.8E-2	1.8E-2	2.8E-1
LOC643201	8.1E-5	5.0E-5	1.8E-4	2.6E-4	1.5E-3	5.4E-3	1.0E-3	2.1E-3	8.7E-2
KLK3	8.6E-5	3.8E-5	8.1E-5	4.5E-5	1.6E-4	3.7E-2	1.0E-4	2.0E-2	1.0E-3
PIAS3	9.6E-5	7.7E-5	2.8E-4	4.8E-4	6.3E-3	3.8E-3	2.2E-3	2.4E-2	4.8E-2
IRX4	3.9E-4	4.9E-4	9.9E-4	2.7E-3	2.0E-1	1.9E-3	2.1E-2	5.0E-2	7.0E-1
MRPS31	4.1E-4	2.1E-4	7.1E-4	4.4E-4	4.3E-2	4.0E-2	8.4E-4	4.0E-1	1.8E-2
CLDN11	5.3E-4	1.3E-3	5.6E-4	3.1E-3	2.7E-4	7.4E-4	3.5E-2	4.8E-4	5.8E-2
ZNF526	2.0E-3	1.7E-3	8.8E-4	7.3E-4	6.9E-3	1.5E-2	2.3E-3	3.4E-2	3.3E-2
AGXT2L1	6.0E-3	4.4E-3	8.4E-3	6.9E-3	1.7E-4	1.7E-2	2.0E-2	1.1E-3	1.9E-2

The values marked in bold are known susceptibility genes of the risk of prostate cancer.