Published in final edited form as:

Genet Epidemiol. 2016 November ; 40(7): 579–590. doi:10.1002/gepi.21995.

Efficient Unified Rare Variant Association Test by Modeling the Population Genetic Distribution in Case-control Studies

Huilin Li^{1,*} and Jinbo Chen²

¹Department of Population Health, New York University School of Medicine, 227 East 30th Street, New York, NY 10016

²Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, 423 Guardian Dr., 203 Blockley Hall, Philadelphia, PA 18940

Abstract

Recent advancements in next-generation DNA sequencing technologies have made it plausible to study the association of rare variants with complex diseases. Due to the low frequency, rare variants need to be aggregated in association tests to achieve adequate power with reasonable sample sizes. Hierarchical modeling/kernel machine methods have gained popularity among many available methods for testing a set of rare variants collectively. Here, we propose a new score statistic based on a hierarchical model by additionally modeling the distribution of rare variants under the case-control study design. Results from extensive simulation studies show that the proposed method strikes a balance between robustness and power and outperforms several popular rare-variant association tests. We demonstrate the performance of our method using the Dallas Heart Study.

Keywords

case control study; rare variant; gene set; population genetic distribution

Introduction

Genome wide association studies (GWAS) have successfully identified hundreds of variants associated with complex human traits. Most of the identified variants are common with minor allele frequencies (MAFs) above 0.05. However, for many complex traits, there is still missing and unexplained heritability [Gibson 2012; Manolio, et al. 2009]. The 1,000 Genome Project revealed that a significant proportion of variants in any individual genome (5–10%) had never been seen before [Altshuler et al., 2012; Keinan and Clark, 2012]. It is now recognized that common and rare variants are equally important for understanding the genetic basis of human traits. Next-generation sequencing technology makes it plausible to identify these rare variants in individual samples [Altshuler et al., 2012]

Correspondence: huilin.li@nyumc.org.

All authors have no conflict of interest to declare.

By convention, SNPs with MAF between 1% and 5% are referred as low-frequency variants and SNPs with MAF less than 1% as “rare variants” [Lee et al. 2014]. Due to low frequencies, rare variants are typically found in a small fraction of sequenced samples, which poses challenge in association testing. The standard method for testing association with individual common SNPs is usually underpowered for rare variants. Some powerful test statistics have recently been proposed, which largely fall into two groups: 1) The “burden” tests [Li and Leal, 2008; Morgenthaler and Thilly, 2007; Pan, 2009], which first collapse information across rare variants (either weighted or unweighted) to create a “super” variant, then test this “super” variant in association with the interest phenotype. 2) Variance component based tests [Wu et al., 2011; Neale et al., 2011], which assume that the association parameter for each rare variant follows a distribution that has mean 0 and variance σ^2 , and the association is assessed by testing the null hypothesis $\sigma^2 = 0$. The burden tests assume a common effect for all variants, therefore have high power if the effects of different variants are in the same direction or a large proportion of the variants are associated. The variance component tests are more powerful when the directions of association across variants differ or a large proportion of the variants are null. They are not sensitive to the averaged effect of the variants in the study region. Recently, two methods have been proposed to unify the power of these two classes of tests. Lee et al. [Lee et al., 2012] proposed a Sequence Kernel Association Test-Optimum (SKAT-O) test that is a linear combination of a burden test and the Sequence Kernel Association Test (SKAT) variance component test. Sun et al [Sun et al., 2013] proposed a method that combines score statistics for the averaged group effect and variance components of individual SNP effects based on a mixed effects model (MiST). Both methods were designed to test the average and variation in individual effects of rare variants jointly.

In this paper, we propose a novel unified association test that enhances the efficiency of the current rare variant tests. It has been shown that population genetic modeling of the distribution of genetic variants could lead to improved power for assessing genetic association in retrospective case-control studies [Chen and Chatterjee, 2007]. Similar idea has been employed for improving power for assessing gene-environment interaction effects by imposing the constraint of gene-environment independence [Chatterjee and Carroll, 2005]. This motivates us to integrate the population genetic model with the generalized mixed effects model through a hierarchical modeling for both common and individual rare variant effects within one unified association test to further borrow strength across all the variants to enhance power. The proposed test is built upon the retrospective likelihood that accounts for ascertainment bias resulting from case-control sampling [Piegorsch et al., 1994].

We conducted extensive simulation studies to compare our newly proposed tests with two unified test competitors: SKAT-O and MiST tests. Together with analysis of data from the Dallas Heart Study, we showed that our test is valid and appealing because it enhances the robustness and power of the two existing unified tests across a range of realistic scenarios.

Methods

Notation and Model

Suppose that n_1 cases and n_0 controls are sequenced in a region where K variants are genotyped. Let Y_i denote case-control status for the i^{th} subject ($Y_i=1$ for cases; $Y_i=0$ for controls), $G_i=(G_{i1}, G_{i2}, \dots, G_{iK})'$ denote the numerical values of the K genotypes for the i^{th}

subject, and $\bar{G}_i = \frac{1}{K} \sum_{k=1}^K G_{ik}$ denote the averaged genotype values for the i^{th} subject. Using a logistic penetrance model, the K variants are related to the case-control status through

$$\text{logit } P(Y_i=1) = \alpha_0 + \gamma \bar{G}_i + \sum_{k=1}^K \beta_k G_{ik}. \quad (1)$$

In this model, the association of variants with the case-control status is described by two sets of parameters [Sun et al., 2013; Wang et al., 2012]: γ and $\beta=(\beta_1, \beta_2, \dots, \beta_K)'$. γ is the regression coefficient that describes the average effect of all K variants, that is, the fixed burden effect, and β_k is the deviation of the k th individual variant from the burden effect γ . β_k 's are assumed as realizations from a distribution with mean 0 and variance σ^2 . Following [Wu et al. 2011], We are interested in testing if any of these variants is associated with phenotype Y , with the null hypothesis $H_0: \gamma=0, \sigma^2=0$.

Modeling the distribution of rare variants and the likelihood function

It has been widely recognized that multilevel modeling can improve the estimation efficiency and testing power when the sample size or the number of events is small [Rao, 2003]. This is the case in testing association with a group of rare variants since the number of subjects who carry a particular mutation is small. We propose modeling the joint distribution of the K genotype variables to borrow strength across all variants to increase the power of testing the overall test under H_0 . Specifically, we assume that for subject i : (a) Variant G_{ik} follows a Bernoulli distribution with two genotypes AA ($G_{ik}=0$) and Aa ($G_{ik}=1$); (b) MAF $p_{ik}=P(G_{ik}=1)$ independently follows a Beta distribution $Beta(a_1, a_2)$, which is a conjugate prior for the Bernoulli distribution; and (c) K variants are independently distributed in controls. Under (a)~(c), the joint distribution of G_i can then be written as

$$\begin{aligned} P(G_i) &= \prod_{k=1}^K P(G_{ik}) = \prod_{k=1}^K \int P(G_{ik}|p_{ik}) f(p_{ik}) dp_k \\ &= \prod_{k=1}^K \left(\frac{a_1}{a_1+a_2} \right)^{G_{ik}} \left(\frac{a_2}{a_1+a_2} \right)^{1-G_{ik}} = p^{\sum_k G_{ik}} (1-p)^{K-\sum_k G_{ik}}. \end{aligned} \quad (2)$$

As shown in this equation, the genotype G_{ik} marginally follows a Bernoulli distribution with success probability $p = \frac{a_1}{a_1+a_2}$. Here we can see a_1 and a_2 only play their roles through p , the average proportion of the minor allele for all the variants.

Following Goeman (2006), we assume that $\beta_1, \dots, \beta_K$ follow a common distribution with expectation zero and variance σ^2 , which is equivalent to $\beta = \sigma b$ where b does not depend on σ^2 with $E(b) = 0$ and $E(bb') = I$. Under the association model (1), population genetic model (2), and rare disease approximation, it is straightforward to show that the logarithm of the retrospective likelihood function can be written as

$$l(\sigma^2, \gamma, p; G) = \sum_{i=1}^{n_0+n_1} \log[P(G_i)] + \sum_{j=1}^{n_1} \log \left[\int_{b=(b_1, \dots, b_K)} \exp(\gamma \bar{G}_j + \sum_{k=1}^K \sigma b_k G_{jk}) f(b) db \right] - n_1 \log \left[\sum_{G_j=(G_{j1}, \dots, G_{jK})} P(G_j) \int_{b=(b_1, \dots, b_K)} \exp(\gamma \bar{G}_j + \sum_{k=1}^K \sigma b_k G_{jk}) f(b) db \right] \quad (3)$$

Score statistics under the null hypothesis $H_0: \gamma=0, \sigma^2=0$

Following Sun et al. (2013), we test $H_0: \gamma=0, \sigma^2=0$ using two score tests: the score test for the burden effect γ under the null $H_0: \gamma=0$ with the constraint $\sigma^2=0$, and the score test for σ^2 under the null $H_0: \sigma^2=0$ without any restriction on γ . First, we derive the likelihood score function for γ based on (3) under the null $H_0: \gamma=0$ given that $\sigma^2=0$, as

$$\tilde{S}_\gamma = \sum_{j=1}^{n_1} \bar{G}_j - \sum_{j=1}^{n_1} E \bar{G}_j = \sum_{j=1}^{n_1} (\bar{G}_j - p).$$

Plug the maximum likelihood estimator (MLE) of the unknown parameter p under the null,

which is given by $\hat{p} = \frac{\sum_{i=1}^{n_0+n_1} \bar{G}_i}{n_0+n_1}$, into $\tilde{S}_\gamma$ we obtain

$$S_\gamma = \frac{n_0 n_1}{n_1 + n_0} \left(\frac{1}{n_1} \sum_{j=1}^{n_1} \bar{G}_j - \frac{1}{n_0} \sum_{i=1}^{n_0} \bar{G}_i \right). \quad (4)$$

The score statistics for testing $H_0: \gamma=0$ given that $\sigma^2=0$ can then be written as

$S_\gamma / \sqrt{\widehat{\text{var}}(S_\gamma)}$, where $\widehat{\text{var}}(S_\gamma)$ is the estimated variance for S_γ . By the central limit theorem, as n_0 and n_1 go to infinity, this statistic has an asymptotic standard normal distribution $N(0,1)$ under the null hypothesis. The p-value can be calculated accordingly. We derived that

$\text{var}(S_\gamma) = \left(\frac{n_0 n_1}{n_0 + n_1} \right) \text{var}(\bar{G}_j) = \frac{n_0 n_1}{K(n_0 + n_1)} p(1-p)$ (Appendix) and obtained $\widehat{\text{var}}(S_\gamma)$ by plugging $\hat{p}$ into $\text{var}(S_\gamma)$.

We derived the score statistic for variance component σ^2 by applying Lemma 3 in Goeman et al. (2006) as

$$S_{\sigma^2} = T' T = \sum_{k=1}^K \left(\tilde{G}_k^1 - \frac{1}{K} \sum_{l=1}^K \tilde{G}_l^1 \right)^2, \quad (5)$$

where $T = \left\{ \left(\tilde{G}_k^1 - \frac{1}{K} \sum_{l=1}^K \tilde{G}_l^1 \right)^2 \right\}_{k=1}^K$ and $\tilde{G}_k^1 = \sum_{j=1}^{n_1} G_{jk}$, $k=1, \dots, K$. Following Goeman et al. 2006 and Wu et al. 2011, the asymptotic distribution of S_{σ^2} can be approximated by a

mixture of Chi-squared distribution, i.e. $\sum_{i=1}^r \lambda_i \chi_{1,i}^2$, where $\chi_{1,i}^2$ are independent χ_1^2 random variables and $\lambda_1, \lambda_2, \dots, \lambda_r$ are the non-zero eigenvalues of the estimated covariance matrix

of T , $\widehat{\text{cov}}(T)$. We derived in Appendix $\widehat{\text{cov}}(T) = 11' \left(-\frac{n_1}{K} \hat{p}(1-\hat{p}) \right) + I n_1 \hat{p}(1-\hat{p})$, where 1 is a vector of 1s with dimension K and I is the identity matrix with dimension $K \times K$, and $\hat{p}$ is the MLE estimate of nuisance parameters under the null hypothesis. Davies method [Davies, 1980] is used to approximate this mixture of chi-squared distribution and calculate the p-values. As for all the parametric methods, this asymptotic distribution is derived under the specified model and assumptions. When the model assumption is not satisfied, however, it is not feasible to obtain the valid model-based estimate of covariance matrix of T , $\widehat{\text{cov}}(T)$. Ignoring model mis-specification leads to inflated type I error rates of our asymptotic method. As an alternative, the valid p-value can be calculated by permutation, where the disease status Y_i is shuffled among subjects and the p-value is calculated as the proportion of the permuted score statistics $S_{\sigma^2}^p$ being greater than the observed score statistics S_{σ^2} .

Following Sun et al. 2013 we use the Tippett's procedure to combine these two independent score tests to produce the overall p-value for the joint test of $H_0: \gamma = 0, \sigma^2 = 0$. This combined procedure requires that S_γ and S_{σ^2} are independent, which is proved in Appendix. Specifically, let p_γ and p_{σ^2} denote the p-values from S_γ and S_{σ^2} , respectively. The overall p-value is calculated as $1 - (1 - \min(p_\gamma, p_{\sigma^2}))^2$.

Simulation Studies

We carried out extensive simulation studies to compare the performance of our case-control score method (CCS) with SKAT-O [Lee et al., 2012] and MiST [Sun et al., 2013], the two available methods for testing the group common effect and variance components jointly. We used R packages SKAT and MiST for the corresponding methods. We used the equal weight in SKAT-O method to reflect the independence of the MAF and the effect size in the simulation settings. We evaluated type I error rates and power as a function of sample size, number of neutral variants, genotype distribution, and degree of rareness of genotype data. We set the disease prevalence (p) equal to 5% under the various null scenarios. Among K variants in the model, we randomly set eight variants to be associated with the phenotype and 8 or 50 neutral variants that were not associated.

Generate Genotype data

We generated genotype data G using three approaches. Model A) We assumed that all the variants are de novo and that their genotype were generated from a Bernoulli distribution with an independent mutation rate generated from the Beta distribution $\text{Beta}(1, a_2)$ as in model (2). Model B) To assess the robustness of our proposed test, we also generated the genotype data from a distribution that deviated from model (1). Following Basu and Pan 2011's simulation design, we first generated two independent haplotypes by dichotomizing two latent components from a multivariate normal distribution. Between any two latent components, the first-order autoregressive covariance structure was assumed. The MAF was restricted between 0.001 and 0.01. The genotype data were obtained by combining the two haplotypes. Model C) We used the software package Genome, a coalescent-based whole-genome simulator [Liang et al. 2007] to generate a population with effective size $N=60,000$ for $p=1\%$, and 1000 original variants. We retained variants which have MAF less than 0.01 (about 64%) in analyses. Starting from a randomly selected variant in the retained set, the following K variants was taken as the studied variant. The causal variants were randomly selected from the variant set. The recombination rate between consecutive fragments was taken as the default value 0.0001. For model A and C, we generated $n_0 = n_1 = 500$ or 1,000 cases and controls. For model B, since the studied methods can have meaningful power with a smaller sample size, we considered $n_0 = n_1 = 200$ or 500 cases and controls in the simulations.

Causal Variant Effect Patterns

Throughout the simulations, we set the significance level at 0.05. For type I error evaluation, we set $\gamma=0$, and $\beta_k=0$ for $k=1, \dots, K$. For the power evaluation, we considered four odds ratio (OR) parameter patterns for 8 randomly selected disease associated variants [Wang et al., 2012]. From Model (1), the marginal OR for individual variant is given as

$\text{OR} = \{\exp(\gamma/K + \beta_m)\}_{m=1}^8$. The four patterns were: 1) $\text{OR} = (2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0)$, the direction and effect sizes were the same; 2) $\text{OR} = (2.0, 2.0, 1.8, 1.8, 1.4, 1.4, 1.2, 1.2)$, the effects were in the same direction but of different magnitude; 3) $\text{OR} = (2.0, 2.0, 2.0, 2.0, 0.5, 0.5, 0.5, 0.5)$, the effect of the first four variants were in the same direction while that of the other four were in the opposite direction, and the average effect was zero; and 4) $\text{OR} = (3.0, 3.0, 2.0, 2.0, 2.0, 0.5, 0.5, 0.5)$, both the direction and strength of association varied but the average effect was positive.

Other Scenario

In order to check the robustness of the proposed method on dealing with the confounding covariates, we designed the following simulation scenario. We generated the genotype data from genotype model A with disease prevalence being 5% and the disease status using the logistic model with one extra confounding covariate, i.e.

$$\text{logit } P(Y_i=1) = \alpha_0 + \gamma \bar{G}_i + \sum_{k=1}^K \beta_k G_{ik} + \beta_X X_i,$$

where X is generated from Uniform(0,1) for each subject and β_X varies from 0 to 2. In the analysis, we pretended we did not know the confounding covariate X and ignored it in all three methods.

Results

In the simulation setting, the parameter a_2 was set as 200 or 400, where a larger value of a_2 will lead to rarer variants. Table 1 describes the Beta probability distribution of MAFs for different values of a_2 . When $a_2 = 200$, on average 14% variants had MAF between 0.01 and 0.05, 68% variants had MAF between 0.001 and 0.01, and 18% variants had MAF below 0.001. With a_2 equal to 400, 65% of the variants had MAF less than 0.01, and about 33% had MAF less than 0.001. Therefore, the considered variants were quite “rare”.

Tables 2–4 present the empirical type I error rates for all three methods under three genotype models and some other scenarios. Our CCS method based on permutation (CCS_perm) and the competitive methods all maintained the nominal type I error rates at different settings regardless of the sample size, number of neutral variants and genotype distribution. As expected, the CCS method based on the asymptotic approximation (CCS_asy) maintained the nominal type I error rate. When the model assumption for the asymptotic theory was violated, such as under genotype model B and C, the asymptotic method produced inflated type I error and the results are omitted here and in the following power section.

Power

Within each genotype generating method, we considered four patterns of causal variant effects as aforementioned. Under patterns 1 and 2, the causal variants all had positive effects, and the effect sizes were constant in pattern 1 and varied in pattern 2. These two patterns were in favor of the burden test. Under patterns 3 and 4, some causal variants had positive and some had negative effects, which were constant in pattern 3 and varied in pattern 4. Therefore, the average effect was zero in pattern 3. Patterns 3 and 4 are in favor of SKAT.

First when the genotype data was generated from $Beta(1, a_2)$ as in model (2), considering two sample sizes: 500 and 1000 cases and controls, MiST and SKAT-O's performance were comparable, and SKAT-O had slightly higher power than MiST (Figure 1, 2). Comparing to these two tests, CCS tests always had the highest power across the four patterns and two a_2 values. The power advantage of the CCS tests was particularly prominent under patterns 3 and 4. For example, under pattern 4, with a_2 equal to 200, $n_0 = n_1 = 1000$, and the number of neutral variants equal to 50 (Figure 2), the power of SKAT-O, MiST, CCS_perm and CCS_asy tests was 0.49, 0.44, 0.74, and 0.84 respectively. The power of our CCS tests was nearly twice as high as that of the other two methods. For the CCS methods, the p-value that was computed based on the asymptotic theory always produces higher power than that based on permutation. From Figure 1 and 2 we also observe that the SKAT-O_perm and SKAT_O based on asymptotic theory have almost equivalent results and the MiST_perm has always higher power than its asymptotic version.

Under genotype model B and C, we skipped CCS based on asymptotic theory because of the inflated type-I error rate due to violation of the model assumption. We only compared the performance of the SKAT-O, MiST and CCS tests using permutation in those two models.

For genotype model B, in the top row in Figure 3, when the sample size was small (200 cases and 200 controls), the CCS_perm always had higher power than SKAT-O_perm and MiST_perm. For example, in pattern 4, comparing to SKAT-O_perm, CCS_perm has 28% power improvement. When the sample size was relatively large (500 cases and 500 controls), the power of all three methods became higher in all patterns (bottom row, Figure 3). However, CCS_perm still had higher power compared with SKAT-O_perm and MiST_perm, especially when the number of neutral variants is large (50).

Under coalescent genotype model C, we considered two sample sizes: 500 and 1000 cases and controls (Figure 4). SKAT-O and MiST method had almost identical power in all scenarios, while CCS was more powerful except when the sample size was 1000 under Pattern 3 with 8 neutral variants. The advantage of CCS over the other two methods was more evident in patterns 1 and 2 than it was in patterns 3 and 4. For example, when $n_0=1000$, under pattern 1 with 50 neutral variants, the power of SKAT-O, MiST and CCS was 0.18, 0.17, and 0.22 respectively, in which CCS has gain 22% more power comparing to SKAT-O and MiST.

In the last scenario, we generated data using the model A with one confounding covariate, which was ignored in the association tests. In this study, CCS_asy tends to have slightly inflated Type I error as β_X increases, however, CCS_perm always maintains nominal Type I error as SKAT-O and MiST methods do (Table 5). Figure 5 presents the power comparison between three methods. CCS_asy and CCS_perm always have higher power than their competitors. This shows the robustness of our proposed CCS method.

Application to Dallas Heart Study Data

We applied CCS_perm, SKAT-O, and MiST using data from Dallas Heart Study [Pritchard and Cox 2002]. Data is available for three candidate genes, *ANGPTL3* (MIM 604774), *ANGPTL4* (MIM 605910), and *ANGPTL5* (MIM 607666) in 3,476 individuals. We analyzed 35, 30 and 25 variants with $MAF < 0.05$ in the three genes, respectively. We evaluated the association between the dichotomized triglyceride level and each candidate gene. The triglyceride level was dichotomized based on the highest and lowest quartiles of each of the six sex-ethnicity groups (Table 5) [Wu, et al. 2011] with 867 cases and 873 controls. The p-values are reported in Table 4. All the three methods detected a significant association with gene *ANGPTL3*. The p-value based on CCS_perm is close to those based on SKAT-O and MiST, which confirmed the validity of our CCS test.

Discussion

In this paper, we proposed a unified score test for assessing association with rare variants under the case-control study design. Our test is novel in two aspects: 1) it integrates the population genetic model with the popular hierarchical model to increase power for testing rare variants as a set. Our method has high power when the MAFs of the test variants are

very low and sample size is small; 2) Our score test was derived based on the retrospective likelihood that appropriately accounts for the case-control sampling. Similar as other existing unified rare variant association tests, the proposed method also tests jointly both the average effect and variance component of individual variant effects. Our test outperformed the competing methods in all the simulation scenarios we considered when the genotype data arose from the Beta distribution. The power gain was even more appreciated when the sample size was small, test variants were very rare and number of neutral variant is large. When the genotype data arose from a distribution that deviated from Beta, such like in the more realistic coalescent model and Basu&Pan's genotype type model, the advantage of our test was also confirmed.

Unlike most currently available set-based rare variant association tests, which were designed assuming the prospective cohort study design, our method appropriately recognized the case-control design and derived statistics through the retrospective likelihood. In our model, the genotype data is modeled directly in the case and control groups and treated as the random variable. Therefore, we were able to build in the population genetic model into the association between rare variants and case/control status. This turns out to produce a more robust and efficient score test for the rare variant test in the case-control study.

We derived an analytic asymptotic distribution for the proposed test statistic, and showed in simulation studies that it performed well for p-value calculation when the genotype data arose from the assumed model. However, the asymptotic distribution leads to inflated type I error rates when the model assumption for the genotype data is invalid. But the p-value obtained based on permutation appeared to be robust to all of our model assumptions on the genotype data while retaining the power advantage. Therefore, we suggest that p-value be calculated using the permutation method for the proposed method in practice.

Both SKAT-O and MiST can reduce to burden test or SKAT test when certain regression parameters are taken to be zeros [Sun, Zheng and Hsu 2013]. The first component of our score statistic (4) for the common group effect is actually the same as the burden statistic before standardization (see appendix), while the second component (5) for the variance component is different from SKAT. The main factor to affect the relative performance of SKAT-O and CCS is the MAF. Since the core part of SKAT is the sum of the K squares of the difference of the MAF estimates between cases and controls in each variant, when the study variants are very rare, with small sample size, the variability of the MAF estimates for each individual variant is large, which results in large variability of the SKAT statistics and leads the power loss. In contrast, our score function for the variance component was the distance between the MAF estimates for each variant and the MAF estimates in the cases averaged across variants. The overall MAF estimate was derived from the assumption that all the variants follow the same population genetic model. As such, we were able to borrow strength across all the variants to result in smaller variance and increase the power in testing the overall effect. Similar technique has been well studied in the small area estimation field (small means sample size or event size is small) [Rao 2003]. As the review mentioned in the next comment, CCS test statistics is equivalent to the variance of MAF estimate. As K increases, the variance of MAF estimate is more accurate. While the test statistics for SKAT is the summation of K square terms, its variance goes up as K increases. This explains why

with the large number of the neutral variant, the edge of CCS over SKAT-O is more prominent.

Our method has also several limitations. First, our model implicitly assumed the independence between gene and environment independence, which has been used to increase power for testing gene-environment interaction effects. When this assumption is violated, bias and inflation in type-I error rates can be severe [Chen and Chatterjee, 2007; Chen et al., 2012; Mukherjee et al., 2012]. Our proposed method can potentially suffer from the same limitation. However, because of the rareness of the mutations, we felt that concerns for confounding would primarily be due to population stratification, not to non-genetic risk factors. When population substrata can be well characterized by a small number of population subgroups, our method can be adapted in a straightforward manner by using separate prior genetic distributions for each distinct covariate value. Secondly, our method cannot adjust the confounding covariates. When covariates do not confound the association, it has been recognized that logistic regression analysis without adjusting for covariates is valid and has higher power than that adjusting for covariates [Pirinen et al., 2012]. Along the same line of argument, our proposed method is valid. When covariates confound association, for example, when covariates include population stratification variables, these covariates must be adjusted for. Our method in general will not be applicable, although it can be generalized when covariates are discrete and have a small number of distinct values (we will then use separate prior genetic distributions corresponding to each distinct covariate value). In a simplified simulation study, we have showed that our proposed CCS method is robust to the minor confounding covariate adjustment. Thirdly, since we recommend the permutation based CCS method to avoid possible inflation in type I error rates when p-values are calculated based on the asymptotic theory due to invalid model assumption, the proposed method is infeasible for analyzing genome-wide association due to the high computation burden. The proposed CCS method therefore is best suitable for gene-based, pathway based or small scale exome sequencing studies. Fourthly, our method only has better performance over SKAT-O in the rare variant analysis (i.e. $MAF < 0.01$). When the MAF is large, SKAT-O outperforms CCS. Because when the MAF is large, with a reasonable sample size (more than 500 cases and controls), the MAF estimate for each individual in SKAT-O method is accurate enough. In such scenarios, SKAT does not impose any distribution on the MAF across the variants and is most efficient. While CCS models the distribution of MAF using a Beta prior and will likely deviate from the underlying true model. Thus when the portion of the variants with big MAF is large, SKAT will have better performance than CCS.

Acknowledgments

The authors would like to thank the reviewers and editor for their constructive suggestions. The authors also thank Dr. Iuliana Ionita-Laza for sharing the Dallas Heart Study with us. The work is funded in part by NIH U01 CA170948, R21-ES020811 and R01-ES016626.

References

- Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012; 491:56–65. [PubMed: 23128226]
- Basu S, Pan W. Comparison of Statistical Tests for Disease Association with Rare Variants. *Genet Epidemiol*. 2011; 35(7):606–619. DOI: 10.1002/gepi.20609 [PubMed: 21769936]

- Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*. 2005; 92(2):399–418.
- Chen J, Chatterjee N. Exploiting Hardy-Weinberg Equilibrium for efficient screening of single SNP associations from case-control studies. *Human Heredity*. 2007; 63:196–204. [PubMed: 17317968]
- Chen J, Kang G, Vanderweele T, Zhang C, Mukherjee B. Efficient designs of gene-environment interaction studies: implications of Hardy-Weinberg equilibrium and gene-environment independence. *Stat Med*. 2012 Sep 28; 31(22):2516–30. DOI: 10.1002/sim.4460 [PubMed: 22362617]
- Davies R. Algorithm as 155: The distribution of a linear combination of chi2 random variables. *JRSSC*. 1980; 29:323–333.
- Gibson G. Rare and common variants: Twenty arguments. *Nature Review Genetics*. 2012; 13:135–145.
- Goeman JJ, van de Geer S, van Houwelingen HC. Testing against a high dimensional alternative. *J Royal Stat Soc B*. 2006; 68:477–493. DOI: 10.1111/j.1467-9868.2006.00551.x
- Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*. 2012; 336:740–743. [PubMed: 22582263]
- Lee S, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet*. 2012; 91:224–237. [PubMed: 22863193]
- Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am J Hum Genet*. 2014 Jul 3; 95(1):5–23. DOI: 10.1016/j.ajhg.2014.06.009 [PubMed: 24995866]
- Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*. 2008; 83:311–321. [PubMed: 18691683]
- Liang L, Zoellner S, Abecasis GR. GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics*. 2007; 23:1565–1567. [PubMed: 17459963]
- Lin X. Variance component testing in generalized linear models with random effects. *Biometrika*. 1997; 84:309–326.
- Liu D, Lin X, Ghosh D. Semiparametric regression of multi-dimensional genetic pathway data: least squares kernel machines and linear mixed models. *Biometrics*. 2007; 63(4):1079–88. [PubMed: 18078480]
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, MacKay TFC, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. *Nature*. 2009; 461:747–753. [PubMed: 19812666]
- Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *MutatRes*. 2007; 615(1):28–56.
- Mukherjee B, Ahn J, Gruber SB, Chatterjee N. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *Am J Epidemiol*. 2012 Feb 1; 175(3):177–90. Epub 2011 Dec 22. DOI: 10.1093/aje/kwr367 [PubMed: 22199027]
- Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. Testing for an unusual distribution of rare variants. *PLoS Genet*. 2011; 7(2011):e1001322. [PubMed: 21408211]
- Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet Epidemiol*. 2009; 33(6):497–507. DOI: 10.1002/gepi.20402 [PubMed: 19170135]
- Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med*. 1994; 13(2):153–162. [PubMed: 8122051]
- Pirinen M, Donnelly P, Spencer CC. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature Genetics*. 2012; 44(8):848851.
- Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant...or not? *Human Molecular Genetics*. 2002; 11:2417–23. [PubMed: 12351577]

- Price A, Kryukov G, de Bakker P, Purcell S, Staples J, Wei L, Sunyaev S. Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics*. 2010; 86:832–838. [PubMed: 20471002]
- Rao, JNK. Small area estimation. Wiley; 2003.
- Sun J, Zheng Y, Li Hsu. A unified Mixed-effects model for rare-variant association in sequencing studies. *Genetic Epidemiology*. 2013; 37(4):334–344. [PubMed: 23483651]
- Tennessen JA, Bigham AW, O'Connor TD, Fu WQ, Kenny EE, Gravel S, McGee S, Do R, Liu XM, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM, GOB, GOS. Project NES. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. 2012; 337:64–69. [PubMed: 22604720]
- Wang YJ, Chen YS, Yong Q. Joint rare variant association test of the average and individual effects for sequencing studies. *Plos One*. 2012:e32485. [PubMed: 22468164]
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, et al. Rare variant association testing for sequencing data using the sequence kernel association test (SKAT). *Am J Hum Genet*. 2011; 891:82–93.

Appendix

A1 Derivation of the joint distribution of G_i

$$\begin{aligned}
 P(G_i) &= \prod_{k=1}^K P(G_{ik}) \\
 &= \prod_{k=1}^K \int P(G_{ik}|p_k) f(p_k) dp_k \\
 &= \prod_{k=1}^K \int p_k^{G_{ik}} (1-p_k)^{1-G_{ik}} a_2 (1-p_k)^{a_2-1} dp_k \\
 &= \prod_{k=1}^K a_2 \int p_k^{G_{ik}} (1-p_k)^{a_2-G_{ik}} dp_k = \prod_{k=1}^K a_2 B(G_{ik}+1, 1+a_2-G_{ik}) \\
 &= \prod_{k=1}^K a_2 B(G_{ik}+1, 1+a_2-G_{ik}) \\
 &= \prod_{k=1}^K a_2 \frac{\Gamma(G_{ik}+1) \Gamma(1+a_2-G_{ik})}{\Gamma(2+a_2)} \\
 &= \prod_{k=1}^K \left(\frac{a_2}{a_2+1} \right)^{1-G_{ik}} \left(\frac{1}{a_2+1} \right)^{G_{ik}} \\
 &= \left(\frac{a_2}{a_2+1} \right)^{K-\sum_k G_{ik}} \left(\frac{1}{a_2+1} \right)^{\sum_k G_{ik}}
 \end{aligned}$$

A2 Derivation of the Retrospective Likelihood Function. Under the association model (1) and population genetic model (2) and using the rare disease approximation, it is straightforward to show the retrospective likelihood function as

$$\begin{aligned}
\bar{L}(\beta, \gamma; G, p) &= \prod_{i=1}^{n_0+n_1} P(G_i|Y_i) = \prod_{i=1}^{n_0} P(G_i|Y_i=0) \prod_{j=1}^{n_1} P(G_j|Y_j=1) \\
&= \prod_{i=1}^{n_0} P(G_i) \prod_{j=1}^{n_1} \frac{P(Y_j=1|G_j)P(G_j)}{P(Y_j=1)} \\
&= \prod_{i=1}^{n_0} P(G_i) \prod_{j=1}^{n_1} \frac{P(Y_j=1|G_j, \beta) P(G_j)}{\sum_{G_j=(G_{j1}, \dots, G_{jK})} P(Y_j=1|G_j, \beta) P(G_j)} \\
&= \prod_{i=1}^{n_0} P(G_i) \prod_{j=1}^{n_1} \frac{\exp(\alpha_0 + \gamma \bar{G}_i + \sum_{k=1}^K \beta_k G_{ik}) P(G_j)}{\sum_{G_j=(G_{j1}, \dots, G_{jK})} \exp(\alpha_0 + \gamma \bar{G}_i + \sum_{k=1}^K \beta_k G_{ik}) P(G_j)} \\
&= \prod_{i=1}^{n_0} P(G_i) \prod_{j=1}^{n_1} \frac{\exp(\gamma \bar{G}_i + \sum_{k=1}^K \beta_k G_{ik}) P(G_j)}{\sum_{G_j=(G_{j1}, \dots, G_{jK})} \exp(\gamma \bar{G}_i + \sum_{k=1}^K \beta_k G_{ik}) P(G_j)}
\end{aligned}$$

A3 The calculation of $\text{var}(\bar{G}_j)$

$$\begin{aligned}
\text{var}(\bar{G}_j) &= \frac{1}{K^2} \sum_{k=1}^K \text{var}(G_{jk}) \\
&= \frac{1}{K^2} \sum_{k=1}^K (\text{var}(E(G_{jk}|p_k)) + E(\text{var}(G_{jk}|p_k))) \\
&= \frac{1}{K^2} \sum_{k=1}^K (\text{var}(p_k) + E(p_k(1-p_k))) \\
&= \frac{1}{K^2} \sum_{k=1}^K \left(\frac{a_2}{(1+a_2)^2(2+a_2)} + \frac{1}{1+a_2} - \left(\frac{a_2}{(1+a_2)^2(2+a_2)} + \frac{1}{(1+a_2)^2} \right) \right) \\
&= \frac{1}{K^2} \sum_{k=1}^K \frac{a_2 + (1+a_2)(2+a_2) - 2(1+a_2)}{(1+a_2)^2(2+a_2)} \\
&= \frac{a_2}{K(1+a_2)^2}
\end{aligned}$$

A4 Score statistics for $H_0: \beta = (\beta_1, \dots, \beta_K)' = 0$

When sample size n is relatively small to the number of variant K , it is impossible to test $H_0: \beta = (\beta_1, \dots, \beta_K)' = 0$ directly. Following Goeman (2006), we assume $\beta_1, \dots, \beta_K$ follow a common distribution with expectation zero and variance σ^2 . Therefore, testing the hypotheses $H_0: \beta = 0$ is equivalent to testing $H_0: \tilde{H}_0: \sigma^2 = 0$ against $\tilde{H}_0: \sigma^2 > 0$. Write $\beta = \sigma b$ where b does not depend on σ^2 with $E(b) = 0$ and $E(bb') = I$, the retrospective log-likelihood can be represented in term of σ^2 as

$$\begin{aligned}
l(\sigma^2, \gamma, p; G) &= \log \{ E_{\beta|\sigma^2} \bar{L}(\beta, \gamma, p; G) \} \\
&= \log \left\{ \prod_{i=1}^{n_0} P(G_i) \prod_{j=1}^{n_1} \frac{\int_{b=(b_1, \dots, b_K)} \exp(\gamma \bar{G}_j + \sum_{k=1}^K \sigma b_k G_{jk}) f(b) db}{\sum_{G_j=(G_{j1}, \dots, G_{jK})} P(G_j) \int_{b=(b_1, \dots, b_K)} \exp(\gamma \bar{G}_j + \sum_{k=1}^K \sigma b_k G_{jk}) f(b) db} \right\} \\
&= \sum_{i=1}^{n_0+n_1} \log(P(G_i)) + \sum_{j=1}^{n_1} \log \left(\int_{b=(b_1, \dots, b_K)} \exp(\gamma \bar{G}_j + \sum_{k=1}^K \sigma b_k G_{jk}) f(b) db \right) - n_1 \log \left(\sum_{G_j=(G_{j1}, \dots, G_{jK})} P(G_j) \int_{b=(b_1, \dots, b_K)} \exp(\gamma \bar{G}_j + \sum_{k=1}^K \sigma b_k G_{jk}) f(b) db \right)
\end{aligned}$$

By implementing the Lemma 3 in Goeman et al. (2006), the score statistic is given by

$$S_{\sigma^2} = T' T$$

where

$$\begin{aligned}
T &= \frac{\partial}{\partial \beta} \log \{ \bar{L}(\beta, \gamma, p; G) \} |_{\beta=0} \\
&= \frac{\partial}{\partial \beta} \sum_{j=1}^{n_1} \left[\gamma \bar{G}_j + \sum_{k=1}^K \beta_k G_{jk} - \log \left(\sum_{G_j=(G_{j1}, \dots, G_{jK})} \exp(\gamma \bar{G}_j + \sum_{k=1}^K \beta_k G_{jk}) P(G_j) \right) \right] |_{\beta=0} \\
&= \left\{ \sum_{j=1}^{n_1} \left(G_{jk} - \frac{\sum_{G_j} \exp(\gamma \bar{G}_j + \sum_{k=1}^K \beta_k G_{jk}) G_{jk} P(G_j)}{\sum_{G_j} \exp(\gamma \bar{G}_j + \sum_{k=1}^K \beta_k G_{jk}) P(G_j)} \right) \right\}_{k=1}^K \Bigg|_{\beta=0} \\
&= \left\{ \sum_{j=1}^{n_1} \left(G_{jk} - \frac{\sum_{G_j} \exp(\gamma \bar{G}_j) G_{jk} P(G_j)}{\sum_{G_j} \exp(\gamma \bar{G}_j) P(G_j)} \right) \right\}_{k=1}^K \\
&= \left\{ \sum_{j=1}^{n_1} \left(G_{jk} - \frac{E \exp(\gamma \bar{G}_j) G_{jk}}{E \exp(\gamma \bar{G}_j)} \right) \right\}_{k=1}^K
\end{aligned}$$

When γ is not equal to zero, it needs to be replaced by $\hat{\gamma}$ which is obtained by

solving $\frac{\partial l}{\partial \gamma} \Big|_{\beta=0} = 0$ from (5) we have

$$\begin{aligned}
\frac{\partial l}{\partial \gamma} &= \sum_{j=1}^{n_1} \bar{G}_j - \sum_{j=1}^{n_1} \frac{\sum_{G_j=(G_{j1}, \dots, G_{jK})} \exp(\gamma \bar{G}_j) P(G_j) \bar{G}_j}{\sum_{G_j=(G_{j1}, \dots, G_{jK})} \exp(\gamma \bar{G}_j) P(G_j)} = 0 \\
&\Rightarrow \sum_{j=1}^{n_1} \bar{G}_j - \sum_{j=1}^{n_1} \frac{E[\exp(\gamma \bar{G}_j) \bar{G}_j]}{E[\exp(\gamma \bar{G}_j)]} = 0 \\
&\Rightarrow \sum_{j=1}^{n_1} \bar{G}_j - \frac{1}{K} \sum_{j=1}^{n_1} \sum_{k=1}^K \frac{E[\exp(\gamma \bar{G}_j) G_{jk}]}{E[\exp(\gamma \bar{G}_j)]} = 0
\end{aligned}$$

Since $G_{j1}, \dots, G_{jK}$ are interchangeable, we have

$$\Rightarrow \sum_{j=1}^{n_1} \frac{E[\exp(\gamma \bar{G}_j) G_{jk}]}{E[\exp(\gamma \bar{G}_j)]} = \sum_{j=1}^{n_1} \bar{G}_j. \quad (8)$$

Thus

$$S_{\sigma^2} = T' T = \sum_{k=1}^K (\tilde{G}_{\bullet k}^1 - \frac{1}{K} \sum_{l=1}^K \tilde{G}_{\bullet l}^1)^2, \text{ with } \tilde{G}_{\bullet k}^1 = \sum_{j=1}^{n_1} G_{jk}, \quad k=1, \dots, K.$$

A4 Independence of S_γ and T_k

$$\begin{aligned} \text{cov}(S_\gamma, T_k) &\propto \text{cov}\left(\left(\frac{1}{n_1} \sum_{j=1}^{n_1} \bar{G}_j - \frac{1}{n_0} \sum_{i=1}^{n_0} \bar{G}_i\right) \left(\tilde{G}_{\bullet k}^1 - \frac{1}{K} \sum_{l=1}^K \tilde{G}_{\bullet l}^1\right)\right) \\ &= E\left(\left(\frac{1}{n_1} \sum_{j=1}^{n_1} \bar{G}_j - \frac{1}{n_0} \sum_{i=1}^{n_0} \bar{G}_i\right) \left(\tilde{G}_{\bullet k}^1 - \frac{1}{K} \sum_{l=1}^K \tilde{G}_{\bullet l}^1\right)\right) \\ &= \frac{1}{n_1} E\left(\sum_{l=1}^K \tilde{G}_{\bullet l}^1 \left(\tilde{G}_{\bullet k}^1 - \frac{1}{K} \sum_{l=1}^K \tilde{G}_{\bullet l}^1\right)\right) = 0 \end{aligned}$$

Thus S_γ and T_k are independent.

A5 The covariance matrix of T .

- Using the case only in the second term of T

$$\begin{aligned} T &= \left\{ \tilde{G}_{\bullet k}^1 - \frac{1}{K} \sum_{l=1}^K \tilde{G}_{\bullet l}^1 \right\}_{k=1}^K = \left\{ \left(1 - \frac{1}{K}\right) \tilde{G}_{\bullet k}^1 - \frac{1}{K} \sum_{\substack{l=1 \\ l \neq k}}^K \tilde{G}_{\bullet l}^1 \right\}_{k=1}^K \\ \text{var}(T_k) &= \left(1 - \frac{1}{K}\right)^2 \text{var}(\tilde{G}_{\bullet k}^1) + \frac{K-1}{K^2} \text{var}(\tilde{G}_{\bullet l}^1) = \frac{K-1}{K} \text{var}(\tilde{G}_{\bullet k}^1). \\ \text{cov}(T_k, T_m) &= E\left[\left(1 - \frac{1}{K}\right) \tilde{G}_{\bullet k}^1 - \frac{1}{K} \sum_{\substack{l=1 \\ l \neq k}}^K \tilde{G}_{\bullet l}^1\right] \left[\left(1 - \frac{1}{K}\right) \tilde{G}_{\bullet m}^1 - \frac{1}{K} \sum_{\substack{l=1 \\ l \neq m}}^K \tilde{G}_{\bullet l}^1\right]\right] \\ &= \left(1 - \frac{1}{K}\right)^2 E \tilde{G}_{\bullet k}^1 E \tilde{G}_{\bullet m}^1 - \frac{1}{K} \left(1 - \frac{1}{K}\right) \left[E(\tilde{G}_{\bullet k}^1)^2 + (K-2) E(\tilde{G}_{\bullet k}^1 \tilde{G}_{\bullet l}^1) \right] \\ &\quad - \frac{1}{K} \left(1 - \frac{1}{K}\right) \left[E(\tilde{G}_{\bullet m}^1)^2 + (K-2) E(\tilde{G}_{\bullet m}^1 \tilde{G}_{\bullet l}^1) \right] \\ &\quad + \frac{1}{K^2} \left[(K-2) E(\tilde{G}_{\bullet l}^1)^2 + ((K-1)^2 - (K-2)) E(\tilde{G}_{\bullet k}^1 \tilde{G}_{\bullet l}^1) \right] \\ &= \left[\left(1 - \frac{1}{K}\right)^2 - \frac{2}{K} \left(1 - \frac{1}{K}\right) (K-2) + \frac{(K-1)^2 - (K-2)}{K^2}\right] E \tilde{G}_{\bullet k}^1 E \tilde{G}_{\bullet m}^1 + \left[-\frac{2}{K} \left(1 - \frac{1}{K}\right) + \frac{K-2}{K^2}\right] E(\tilde{G}_{\bullet k}^1)^2 \\ &= \frac{1}{K} [E \tilde{G}_{\bullet k}^1 E \tilde{G}_{\bullet m}^1 - E(\tilde{G}_{\bullet k}^1)^2] \\ &= -\frac{1}{K} \text{var}(\tilde{G}_{\bullet k}^1) \end{aligned}$$

A6 Other methods

Under the logistic regression model $\text{Logit } P(Y_i=1)=\beta_0+\sum_{j=1}^k G_{ij}\beta_j$. Define

$$U = \sum_{i=1}^{n_0+n_1} (Y_i - \bar{Y}) G_i \text{ which is a } K \text{ by } 1 \text{ vector with the } k\text{th element}$$

$$U_k = \sum_{j=1}^{n_1} G_{jk} - \frac{n_1}{n_1+n_0} \left(\sum_{i=1}^{n_0} G_{ik} + \sum_{j=1}^{n_1} G_{jk} \right).$$

$$\text{cov}(U) = \bar{Y}(1-\bar{Y}) \sum_{i=1}^{n_0+n_1} (G_i - \bar{G})(G_i - \bar{G})' \text{ where } \bar{Y} = \sum_{i=1}^n Y_i/n \text{ and}$$

$$\bar{G} = \sum_{i=1}^n G_i/n. \text{ For case-control study, we have } \text{cov}(U) \text{ simplified as}$$

$$\text{cov}(U) = \frac{n_0 n_1}{(n_0+n_1)^2} \sum_{i=1}^{n_0+n_1} (G_i - \bar{G})(G_i - \bar{G})'. \text{ The burden test statistics for } H_0: \beta = (\beta_1, \dots, \beta_k)' = 0 \text{ is given by}$$

$$\begin{aligned} T_B &= \frac{(\sum U_i)^2}{1' \text{cov}(U) 1} \\ &= \frac{\left(\frac{n_0}{n_0+n_1} \right)^2 \left(\sum_{k=1}^K \left(\sum_{j=1}^{n_1} G_{jk} - \frac{n_1}{n_0+1} \sum_{j=1}^{n_0} G_{jk} \right) \right)^2}{\frac{n_0 n_1}{(n_0+n_1)^2} \sum_{i=1}^{n_0+n_1} 1' (G_i - \bar{G})(G_i - \bar{G})' 1} \\ &= \frac{n_0 \left(\sum_{k=1}^K \left(\sum_{j=1}^{n_1} G_{jk} - \frac{n_1}{n_0+1} \sum_{j=1}^{n_0} G_{jk} \right) \right)^2}{n_1 \sum_{i=1}^{n_0+n_1} \left(\sum_{k=1}^K G_{ik} - \frac{1}{n_0+n_1} \sum_{i=1}^{n_0+n_1} \sum_{k=1}^K G_{ik} \right)^2} \end{aligned} \quad (\text{A6.1})$$

Similarly, under the case control setting, the SKAT test for $H_0: \beta = (\beta_1, \dots, \beta_k)' = 0$ is given by $T_{SKAT} = (Y - \bar{Y})' G W G' (Y - \bar{Y})$ where $W = \text{diag}(w_1, \dots, w_k)$ and $\sqrt{w_j} = \text{Beta}(\text{MAF}_j; 1, 25)$, MAF_j is the sample MAF for the j^{th} variant in the data combining both cases and controls. After the simplification, we have

$$T_{SKAT} = (Y - \bar{Y})' G W G' (Y - \bar{Y}) = U' W U = \sum_{k=1}^K w_k \left(\frac{n_0}{n_1+n_0} \sum_{j=1}^{n_1} G_{jk} - \frac{n_1}{n_1+n_0} \sum_{j=1}^{n_0} G_{jk} \right)^2. \quad (\text{A6.2})$$

Figure 1.

Power for genotype model A. (de novo rare variants). Disease prevalence=5%. $n_0=n_1=500$.

Four causal variant effect patterns: Pattern 1, $OR=(2,2,2,2,2,2,2,2)$; Pattern 2, $OR=(2,2,1.8,1.8,1.4,1.4,1.2,1.2)$; Pattern 3, $OR=(2,0.5,2,0.5,2,0.5,2,0.5)$; and Pattern 4, $OR=(3,3, 2,2,2, 0.5,0.5, 0.5)$.

Figure 2.

Power for genotype model A. (de novo rare variants). Disease prevalence=5%. $n_0=n_1=1000$.

Four causal variant effect patterns: Pattern 1, $OR=(2,2,2,2,2,2,2,2)$; Pattern 2, $OR=(2,2,1.8,1.8,1.4,1.4,1.2,1.2)$; Pattern 3, $OR=(2,0.5,2,0.5,2,0.5,2,0.5)$; and Pattern 4, $OR=(3,3, 2,2,2, 0.5,0.5, 0.5)$.

Figure 3. Power for Basu and Pan genotype model B. Disease prevalence=5%. Four causal variant effect patterns: Pattern 1, OR=(2,2,2,2,2,2,2,2); Pattern 2, OR= (2,2,1.8,1.8,1.4,1.4,1.2,1.2); Pattern 3, OR= (2,0.5,2,0.5,2,0.5,2,0.5); and Pattern 4, OR=(3,3, 2,2,2, 0.5,0.5, 0.5).

Figure 4.

Power for coalescent genotype model C. Disease prevalence=5%. Four causal variant effect patterns: Pattern 1, OR=(2,2,2,2,2,2,2,2); Pattern 2, OR= (2,2,1.8,1.8,1.4,1.4,1.2,1.2); Pattern 3, OR= (2,0.5,2,0.5,2,0.5,2,0.5); and Pattern 4, OR=(3,3, 2,2,2, 0.5,0.5, 0.5). Based on 5000 simulations

Figure 5.

Power for genotype model A with confounding covariate X. Disease prevalence=5%. $n_0=n_1=500$. 5000 simulations. Number of neutral variants is 8. Two causal variant effect patterns: Pattern 2, OR= (2,2,1.8,1.8,1.4,1.4,1.2,1.2); and Pattern 4, OR=(3,3, 2,2,2, 0.5,0.5, 0.5).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Probability distribution of $Beta(a_1=1, a_2)$

a_2	$P(MAF<0.05)$	$P(0.01<MAF<0.05)$	$P(0.001<MAF<0.01)$	$P(MAF<0.001)$
200	0.000	0.135	0.683	0.181
400	0.000	0.018	0.652	0.330

Type I Error rates at nominal significance level $\alpha=0.05$ under genotype model A where all the variants are de novo and their genotypes are independently generated from a Bernoulli distribution with an independent mutation rate generated from $Beta(1, \varphi)$ distribution as in model (2), $p=5\%$. Based on 5000 simulations.

Table 2

Genotype model A						
Methods	# of NV	n0=n1	a2=200		a2=400	
			8	50	8	50
SKAT- O_perm	500		0.043	0.044	0.052	0.054
	1000		0.056	0.048	0.041	0.045
SKAT-O	500		0.048	0.052	0.050	0.052
	1000		0.047	0.041	0.038	0.039
MiST_perm	500		0.046	0.048	0.044	0.050
	1000		0.053	0.048	0.051	0.052
MiST	500		0.048	0.038	0.036	0.037
	1000		0.044	0.047	0.041	0.045
CCS_asy	500		0.045	0.046	0.050	0.048
	1000		0.050	0.049	0.048	0.052
CCS_perm	500		0.047	0.047	0.060	0.056
	1000		0.056	0.052	0.055	0.055

Table 3

Type I Error rates at nominal significance level $\alpha=0.05$ under genotype model B) Basu and Pan model and model C) coalescent model. $p=5\%$. Based on 5000 simulations.

Methods	Genotype model B				Genotype model C			
	# of NV	n0=n1	8	50	# of NV	n0=n1	8	50
SKAT-O_perm	200		0.054	0.052	500		0.058	0.041
	500		0.055	0.045	1000		0.040	0.040
MiST_perm	200		0.046	0.045	500		0.053	0.036
	500		0.053	0.045	1000		0.045	0.041
CCS_perm	200		0.052	0.050	500		0.044	0.046
	500		0.051	0.048	1000		0.047	0.049

Table 4

Type I error rates at nominal significance level $\alpha=0.05$ under genotype model A with confounding effect β_X . Disease prevalence=5%. $n0=n1=500$. Number of neutral variants is 8. Based on 5,000 simulations.

	$\beta_X=0$	$\beta_X=0.5$	$\beta_X=1$	$\beta_X=2$
SKAT-O_perm	0.0514	0.0506	0.0582	0.0522
MiST_perm	0.0554	0.052	0.0538	0.0552
CCS_perm	0.0494	0.0454	0.0514	0.0472
CCS_asy	0.048	0.0494	0.0496	0.049

Table 5

P-values for SKAT-O, MiST and CCS_perm tests for testing associations between dichotomized triglyceride level and three genes in the Dallas Heart Study. Triglyceride level was dichotomized by the 1st and 3rd quartiles.

	nvar	SKAT-O	MiST	CCS_perm
ANGPTL3	35	0.037	0.048	0.028
ANGPTL4	30	0.019	0.024	0.083
ANGPTL5	25	0.170	0.234	0.170