Published in final edited form as:
Genet Epidemiol. 2013 May ; 37(4): 334–344. doi:10.1002/gepi.21717.

A Unified Mixed-Effects Model for Rare-Variant Association in Sequencing Studies

Jianping Sun, **Yingye Zheng**, and **Li Hsu***
Division of Public Health Sciences, Fred Hutchinson Cancer Research Center Seattle, WA, USA

Abstract

For rare-variant association analysis, due to extreme low frequencies of these variants, it is necessary to aggregate them by a prior set (e.g., genes and pathways) in order to achieve adequate power. In this paper, we consider hierarchical models to relate a set of rare variants to phenotype by modeling the effects of variants as a function of variant characteristics while allowing for variant-specific effect (heterogeneity). We derive a set of two score statistics, testing the group effect by variant characteristics and the heterogeneity effect. We make a novel modification to these score statistics so that they are independent under the null hypothesis and their asymptotic distributions can be derived. As a result, the computational burden is greatly reduced compared with permutation-based tests. Our approach provides a general testing framework for rare variants association, which includes many commonly used tests, such as the burden test [Li and Leal, 2008] and the sequence kernel association test [Wu et al., 2011], as special cases. Furthermore, in contrast to these tests, our proposed test has an added capacity to identify which components of variant characteristics and heterogeneity contribute to the association. Simulations under a wide range of scenarios show that the proposed test is valid, robust and powerful. An application to the Dallas Heart Study illustrates that apart from identifying genes with significant associations, the new method also provides additional information regarding the source of the association. Such information may be useful for generating hypothesis in future studies.

Introduction

Genome-wide association studies (GWAS) have successfully identified hundreds of variants associated with human traits [Hindorff et al., 2012]. Typically these variants are common with minor allele frequencies (MAF) > 5% and have small to moderate effects. However, despite the demonstrated successes of GWAS, these novel common variants explain only a small fraction of heritability for most complex traits. Many reasons for missing heritability have been given. These include low power for detecting common, weakly penetrant variants for current sample sizes, gene and gene interaction and copy number variation [Maher, 2008; Manolio et al., 2009]. It has also been hypothesized that rare mutations are more likely than common ones to affect structures of functions of proteins [Tennessen et al., 2012; Nelson et al., 2012] and can make significant contributions to the heritability of many traits and diseases [Maher, 2008]. Dickson et al. [2010] further argued that rare variants might not only explain some of missing heritability, but also that they may be the cause of a proportion of detected associations between complex traits and common SNPs from GWAS. Advances in high-throughput technologies have made it possible for researchers to conduct large scale sequencing studies to study the association of rare variants with complex phenotypes. For example, for the NHLBI GO Exome Sequencing Project about 7,000 subjects across diverse and richly phenotyped populations are undergoing whole-exome sequencing with particular focus on heart, lung and blood disorders [Exome variant server, December 2012 accessed].

*Corresponding author: Li Hsu, 1100 Fairview Ave. N., M2-B500, Seattle, WA 98109. lih@fhcrc.org.

Despite these increasingly large scale studies, the power for detecting individual rare-variant associations remains limited because very few individuals carry variant alleles. It is therefore necessary to aggregate rare variants by *a prior* defined sets (e.g. transcripts, genes, pathways) and assess the association of sets of variants instead of individual variants. An added benefit for aggregating rare variants is that the number of tests that need to be adjusted for multiple hypothesis correction is considerably reduced.

There is a growing literature on analyzing the association of a set of rare variants. A natural approach is to create a new variable that counts the number of risk alleles an individual carries for the set and test whether this new variable is associated with phenotypic variation. These tests, sometimes called burden tests [Morgenthaler and Thilly, 2007; Li and Leal, 2008], are motivated by population genetics evolutionary theory that most rare missense alleles are deleterious, and the effect is therefore generally considered one-sided [Kryukov et al., 2007]. It is easy to see that this variable can be incorporated into a regression model to account for potential confounders [Morris and Zeggini, 2010]. This simple count of risk alleles can also be extended to a weighted sum, where the weight may be the frequency of rare allele in the unaffected subjects [Madsen and Browning, 2009], or a function of the observed marginal effects of rare variants. For example, Han and Pan [2010] proposed using the signs of the marginal effects to account for possible protective effects of some variants. Lin and Tang [2011] formalized the idea of weighted sum into a general regression framework, where weights are estimated regression coefficients (EREC).

Another approach to testing the association of a set of rare variants is to use the kernel machine regression framework. In this framework, the effects of variants are assumed to be independently and identically distributed with a mean 0 and variance τ^2. To test whether a set of variants is associated with the phenotype, it is equivalent to test whether the variance $\tau^2 = 0$. Such a test is called the sequence kernel association test (SKAT) [Wu et al., 2011]. They also showed that SKAT is a generalization of the classical C-alpha test [Neale et al., 2011].

A simple burden test is more powerful when a large proportion of rare variants in the set is causal and the effects are mainly deleterious; whereas the SKAT test is more powerful when there is heterogeneity effect among rare variants [Neale et al., 2011]. For example there is a large number of non-causal rare variants in the set or when both protective and deleterious variants are present. Recently Lee et al. [2012] extended SKAT to allow the variant effects to have an equal correlation ρ in addition to the usual assumption of mean 0 and variance τ^2. It is shown that the variance component test (SKAT-O) under this model is a weighted linear combination of a burden test and the SKAT variance component test, with ρ being the weight. However neither SKAT nor SKAT-O accounts for the differential effects due to variant characteristics, which may lead to potential loss of power. Furthermore, the weight ρ used in SKAT-O may cover only a limited set of combinations of the two tests due to the correlation between the burden and the variance component tests. Since the underlying genetic model is unknown and it is also possible that different mechanisms coexist in genome-wide scans, it is important that a test be powerful in as many situations as possible while making use of all relevant information when it is available.

In this paper, we propose a novel association test in order to enhance the robustness and power of the existing tests in a wide variety of biological situations. When grouping a set of variants, it is common that only a subset of variants is associated with phenotype, and among the associated variants the effects may be different. However, the associated variants may have some common characteristics such as nonsense, missense, insertion or deletion. This motivates us to employ hierarchical modeling, exploiting known variant characteristics and leveraging information across loci to enhance power for identifying associated variants.

The idea of hierarchical modeling has been considered previously in genetic association [Hung et al., 2004; Chen and Witte, 2007; King et al., 2010; Capanu and Begg, 2011]. In this model, the effects of individual variants are assumed to be independently distributed with mean modeled as a function of variant characteristics and variance τ^2 to account for residual variant-specific effects or heterogeneity. Testing the association of a set of variants is equivalent to testing both mean=0 and variance $\tau^2 = 0$. Testing for mean = 0 is likely more powerful when there is a strong group effect (e.g., most variants are deleterious); however, if only a small subset of variants are associated with phenotype, testing for $\tau^2 = 0$ is more powerful. Hence, testing both mean and variance = 0 can be more powerful over a wide range of situations than testing only one of the two, because it combines information from both the mean and variance. Previous works for hierarchical model mainly focused on estimation of effect sizes and the testing for association is usually based on the likelihood ratio test (LRT). Such tests require multiple integration under the alternative model, and therefore are computationally intensive, especially when the number of variants is large. Instead of LRT, we propose to derive score statistics under the null, avoiding multiple integration. Our test reduces computational burden substantially, and therefore is more suitable for genome-wide detection of rare-variant association.

Specifically, our proposed test consists of a set of two score statistics, corresponding to grouping effects by variant characteristics and effects of individual variants. An important feature of our proposed method is that we make a novel modification to the two score statistics so that they are independent under the null hypothesis. Such modification facilitates the calculation of the joint distribution of the two statistics, which would be intractable in the conventional score test statistic setting with complex correlation structure. Furthermore, the induced independence enables us to combine the two score statistics by some common approaches such as Fisher's and Tippett's[Koziol and Perlman, 1978], and it yields an asymptotic distribution-based test rather than a permutation-based test.

Our proposed testing procedure based on hierarchical models provides a general framework for assessing the association between rare variants and phenotype traits. It includes many commonly used tests, such as the burden and the SKAT tests, as special cases. However, our test is more appealing since, in practice, it is desirable to consider a single test that encompasses a wide variety of settings rather than conducting separate tests of different types. Furthermore, our test helps identify which components, variant characteristics and/or individual variant effects contribute to the association. Such information is currently not readily available from any existing methods. Simulations under a wide range of scenarios and a real data analysis in this paper supports that our proposed test is valid, robust, powerful and informative.

Methods

Notation and Model

Consider N subjects with a phenotypic trait denoted by $Y^T = (Y_1, \ldots, Y_N)$, where Y_i can be a binary variable that indicates disease status, or a continuous variable such as body mass index or white blood count, for the ith subject, $i = 1, \ldots, N$. Let X_i denote a $m \times 1$ vector of potential confounding covariates (e.g., demographic variables, principal components to account for population stratification, environmental risk factors or other known genetic factors) for the ith subject. For simplicity in notation, we assume X_i includes the intercept. In addition, let G_i denote a $p \times 1$ vector of genotypes for p rare variants in the ith subject, where the genotypes are coded as 0, 1, or 2, representing the number of minor alleles. We assume that the (Y_i, X_i, G_i), $i = 1, \ldots, N$, are independent and identically distributed. The main interest is to test the association of G_i as a set of rare variants with the trait Y_i.

Assume the relationship between trait Y_i and covariates X_i and G_i is

$$g\{E(Y_i)\}=X_i^T\alpha+G_i^T\beta, \quad (1)$$

where $g(\cdot)$ is a link function, and α and β are the regression coefficients for X_i and G_i, respectively. The superscript ^{T} denotes the vector or matrix transpose. The two most commonly used link functions are logit function, $g\{\Pr(Y_i = 1)\} = \log\{\Pr(Y_i = 1)/\Pr(Y_i = 0)\}$, for binary traits; and the identity function, $g\{\mathrm{E}(Y_i)\} = \mathrm{E}(Y_i)$, for continuous traits. For right-skewed distributed traits, a log transformation may be taken before using the identity function.

Since these variants are rare, the power for testing each variant individually is limited. It is therefore important to leverage the information across all p variants in the set by further modeling the variant effects β_j, $j = 1, \ldots, p$. Extensive knowledge has been accumulated about variant characteristics, such as whether a variant is nonsense or missense. It is known that different characteristics affect proteins and functions differently. One way to leverage the information across variants is to assume that variants with similar characteristics have the same effect on the trait, while still allowing for potential individual variant effects. Denote the characteristics for the jth variant ($j = 1, \ldots, p$) by a $q \times 1$ vector Z_j, and δ_j the variant specific effect that can not be explained by the variant characteristics Z_j. Then the effect β_j can be specified by the model,

$$\beta_j=Z_j^T\pi+\delta_j, \quad (2)$$

where π is a vector $q \times 1$ vector of regression coefficients. Since the variants are rare, we assume δ_j, $j = 1, \ldots, p$ follow a distribution with mean 0 and variance τ^2. Plugging model (2) into model (1), we have

$$g\{E(Y_i)\}=X_i^T\alpha+(G_i^TZ)\pi+G_i^T\delta, \quad (3)$$

where $Z^T=(Z_1^T,\ldots,Z_p^T)$ is a $p \times q$ matrix of q variant characteristics for the p variants and $\delta^T = (\delta_1, \ldots, \delta_p)$ which denote the individual effects of p variants. This model is rather general and includes commonly used models for assessing the association of the set of the variants. We show below that both the burden test and the variance-component-based SKAT test are special examples of model (3).

We first show the model that corresponds to the burden test is included in model (3). Set $\delta = 0$ and Z to be a $p \times 1$ vector of $1/p$, then model (3) reduces to $g\{E(Y_i)\}=X_i^T\alpha+(r_i/p)\pi$ where $r_i=\sum_{j=1}^{p}G_{ij}$ is the count of minor alleles that the ith individual carries over the p variants. If subjects have different variants genotyped, for example due to missingness, we can set Z to be a $n_i \times 1$ vector of $1/n_i$, where n_i is the number of genotyped variants. The model becomes $g\{E(Y_i)\}=X_i^T\alpha+(r_i/n_i)\pi$, which is proposed by Morris and Ziggini [2010]. Testing $H_0 : \pi = 0$ is equivalent to testing whether there is an overall effect of the set of variants on the phenotype. The underlying assumption for this test is that all variants have the same effect on the trait and there is no heterogeneity.

To allow for individual variant effects, one may set $Z_j = w_j$, where w_j is the weight for the jth variant. One common choice for the weight is $w_j = \{f_j(1 - f_j)\}^{-1/2}$, where is the minor allele frequency (MAF) of the jth variant[Madsen and Browning, 2009]. The idea is that

rarer variants may have greater effects and also to reduce the influence of common variants in the set-based association. Another choice for weights is to use estimated regression coefficients added by a constant [Lin and Tang, 2011].

The SKAT test with linear kernel is another special case of model (3). By setting $\pi = 0$, model (3) reduces to $g\{E(Y_i)\}=X_i^T\alpha+G_i^T\delta$, where δ follows a distribution with mean 0 and variance τ^2. Testing the effect of all variants equal to 0 is equivalent to testing that the variance of δ is equal to 0, i.e., $H_0 : \tau^2 = 0$. This is essentially the model that Wu et al.[2011] used to derive the SKAT test with a linear kernel. The variance component test is particularly powerful if only a small subset of variants are associated with the trait, or if the variants have the opposite effects. If the covariance matrix of $(\delta_1, \ldots, \delta_p)$ allows for an equal correlation, then the model (3) forms the basis from which the SKAT-O score test statistic is derived [Lee et al., 2012].

Clearly, model (3) is more general than the models for any of the aforementioned tests. For example, in the Dallas Heart Study, the variant characteristics for functional variants include missense, nonsense and frameshift. For non-functional variants they include noncoding, synonymous and intronic. If the analysis is limited to functional variants only, we may set $Z = (Z_1, Z_2, Z_3)$, where Z_1 is a vector of ones, Z_2 is an indicator for missense, and Z_3 is an indicator for frameshift. The corresponding regression coefficients π_1, π_2 and π_3 describe the grouping effects of nonsense variants, and the differential effects of missense and frameshift variants compared with nonsense variants. Similarly, for analysis of all variants, we may set $Z = (Z_1, Z_2, Z_3, Z_4)$, where Z_1, Z_2 and Z_3 are same as in the functional variants analysis, and Z_4 is an indicator for nonsense variants. The corresponding regression coefficients π_1, π_2, π_3 and π_4 here describe the grouping effect of non-functional variants, and differential effects of missense, frameshift and nonsense variants compared with non-functional variants, respectively. In addition, the residual variant effects that can not be explained by Z are modeled by random effects d, which follow a distribution with mean 0 and variant τ^2. To test whether the set of variants is associated with Y, it is equivalent to test $H_0 : \pi = 0$ and $\tau^2 = 0$. It is clear that from this formulation the score statistics for π's and τ^2 can provide additional information on the sources of the association, which may help further identify subsets of variants that may be associated with phenotype.

Proposed Score Test Statistics

It is obvious that from model (3) the null hypothesis of no association between G and Y is $H_0 : \pi = 0$ and $\tau^2 = 0$. We will use score statistics to test H_0 because it avoids estimating the variance components τ^2 under the alternative, as in, for example, the likelihood ratio test. Estimating τ^2 is generally very difficult due to p dimensional multiple integration. We can obtain the respective score statistics for π and τ^2 under the null. Note, however, that these two are correlated. With the score statistic for π following a normal distribution and for τ^2 following a mixture chi-square distribution as shown below, it is difficult to derive their joint distribution when they are correlated. Consequently, it is not straightforward to find adequate weights to combine these two score statistics to achieve satisfactory power under different types of alternatives.

To solve this problem, we propose to modify the score statistics so that they are independent. Specifically, we derive the score statistic for π under $H_0 : \pi = 0$, $\tau^2 = 0$ as usual, but for the variance component, we derive the score statistic for τ^2 under $\tau^2 = 0$ without constraining $\pi = 0$. By doing this, we ensure that these two score statistics are independent (see Appendix for the proof of independence). Furthermore, the independence property in addition to each of them having a well established asymptotic distribution offers many possibilities for combining two independent test statistics to maximize the power

against different types of alternatives. Below, we will show the detailed derivation for two modified score statistics and their asymptotic distributions, and discuss different combination procedures in the next section.

Let $X^T = (X_1, \ldots, X_N)$, $G^T = (G_1, \ldots, G_N)$, and let GZ be the matrix product between matrices G and Z. Following Breslow and Clayton [1993] and Lin [1997], the score statistic for π under the null $H_0 : \pi = 0,\ \tau^2 = 0$ can be derived as

$$U_\pi = (GZ)^T (Y - \tilde{\mu}), \quad (4)$$

by using penalized quasi-likelihood, where $\tilde{\mu}^T = (\tilde{\mu}_1, \ldots, \tilde{\mu}_N)$, $\tilde{\mu}_i = g^{-1}\{X_i^T \tilde{\alpha}\}$, for $i = 1, \ldots, N$, $g^{-1}(\cdot)$ is the inverse function of $g(\cdot)$, and $\tilde{\alpha}$ is the estimator for the nuisance parameter α under H_0, which can be obtained by solving the estimating equations

$$\sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \theta} \sigma_i^{-2} (Y_i - \mu_i) = 0. \quad (5)$$

Here $\theta = \alpha$. Under this situation, $\mu_i = g^{-1}(X_i^T \alpha)$, and $\sigma_i^2 = \mathrm{var}(Y_i | X_i)$ under H_0. For the logistic regression model, we have $g^{-1}(x) = \exp(x)/\{1+\exp(x)\}$, and for the linear regression model, $g^{-1}(x) = x$. For both models, the estimating equations reduce to $\sum_{i=1}^{N} X_i^T \{Y_i - g^{-1}(X_i^T \alpha)\} = 0$. By the central limit theorem and the law of large numbers, it can be shown that $\Sigma^{-1/2} U_\pi$ converges to a q-dimensional multivariate normal distribution with mean 0 and covariance matrix I, where $\Sigma = (GZ)^T \{\tilde{D} - \tilde{D}X (X^T \tilde{D} X)^{-1} X^T \tilde{D}\} (GZ)$, $\tilde{D} = \mathrm{diag}(\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_N^2)$, and $\tilde{\sigma}_i^2 = \sum_{i=1}^{N} (Y_i - \tilde{\mu}_i)^2 / N$ for a continuous trait and $\tilde{\sigma}_i^2 = \tilde{\mu}_i (1 - \tilde{\mu}_i)$ for a binary trait.

For τ^2 we obtain the score statistic under $\tau^2 = 0$ without restricting $\pi = 0$. The resulting score is $U_{\tau^2} = (Y - \mu)^T GG^T (Y - \mu) - \mathrm{tr}(GG^T)$. Following Zhang and Lin [2003], we use the first part of U_{τ^2} as our modified score statistic. That is,

$$S_{\tau^2} = (Y - \hat{\mu})^T GG^T (Y - \hat{\mu}), \quad (6)$$

where $\hat{\mu}^T = (\hat{\mu}_1, \ldots, \hat{\mu}_N)$, $\hat{\mu}_i = g^{-1}\{X_i^T \hat{\alpha} + (G_i^T Z)\hat{\pi}\}$, $i = 1, \ldots, N$, and $\hat{\theta}^T = (\hat{\alpha}^T, \hat{\pi}^T)$ can be obtained by solving equations (5) with $\mu_i = g^{-1}\{X_i^T \alpha + (G_i^T Z)\pi\}$ and $\sigma_i^2 = \mathrm{var}(Y_i | X_i, G_i, Z)$ under $\tau^2 = 0$. This proposed score statistic, S_{τ^2}, is similar to the one in Wu et al. [2011]; however, $\hat{\mu}$ is different in the sense that it also includes the grouping effects of variants. Zhang and Lin [2003] and Liu et al.[2007,2008] showed that S_{τ^2} follows a mixture of chi-squared distributions, $\sum_{t=1}^{s} \lambda_s \chi_{1,t}^2$, where $\chi_{1,t}^2$ are independent χ_1^2 random variables, and $\lambda_1 \geq \cdots \lambda_s > 0$ are the nonzero eigenvalues of $\hat{P}^{1/2} GG^T \hat{P}^{1/2}$. Here $P = D - DM (M^T DM)^{-1} M D$ and $M = (X, GZ)$. Matrix $\hat{P}$ is obtained by plugging $\hat{D}$ in P, where $\hat{D} = \mathrm{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_N^2)$, and $\hat{\sigma}_i^2 = \sum_{i=1}^{N} (Y_i - \hat{\mu}_i)^2 / N$ for a continuous trait and $\hat{\sigma}_i^2 = \hat{\mu}_i (1 - \hat{\mu}_i)$ for a binary trait. There exist several methods to approximate this mixture chi-square distribution, for example, both the Davies method and the Liu method have been shown to work well in practice [Liu et al., 2009].

The above derivations are based on assumptions in model (3), i.e. the individual variant effects, $\delta_1, \ldots, \delta_p$, follow a distribution with mean zero and the same variance τ^2. We can

further generalize this assumption to allow the variance of δ_j to be $\omega_j\tau^2$, where ω_j is a non-zero weight for the jth variant, $j = 1, \ldots, p$. Under this generalized assumption, the score test statistic for testing $\tau^2 = 0$ is $wS_{\tau^2} = (Y - \hat{\mu})^T GWG^T(Y - \hat{\mu})$, where $W = \text{diag}\{\omega_1, \ldots, \omega_p\}$. Usually, the weight for the jth variant, ω_j, is a decreasing function of its corresponding observed MAF, f_j. For example, $\sqrt{\omega_j} = beta(f_j, \alpha_1, \alpha_2)$, where $beta(\cdot)$ is a beta function. If $\alpha_1 = \alpha_2 = 0.5$, then $beta(f_j, 0.5, 0.5) = 1/\sqrt{f_j(1-f_j)}$. Thus the weighted test statistic wS_{τ^2} follows a mixture of chi-squared distributions with weights as nonzero eigenvalues of matrix $\hat{P}^{1/2}GWG^T\hat{P}^{1/2}$.

An appealing property of our proposed modified score test is that the variance component score statistic, S_{τ^2} or wS_{τ^2} for $\tau^2 = 0$, is independent of the score statistic U_π for testing $\pi = 0$, and the proof is provided in the Appendix. The benefit of independence is that we can combine U_π and S_{τ^2} (or wS_{τ^2}) together to test $H_0 : \pi = 0, \tau^2 = 0$ without dealing with complex correlation. The two most popular approaches for combining independent tests will be presented in the next section. Since our combined test is based on the mixed effects model, we term our proposed test as the Mixed effects Score Test (MiST).

Combining Independent Score Tests U_π and S_{τ^2}

The literature for combining independent tests is extensive [Koziol and Perlman, 1978]. Two commonly used procedures for combining the independent tests are Fisher's procedure and Tippett's procedure, both based on p values. Since our modified score statistics, U_π and S_{τ^2}, are independent, we can calculate the p value for each score statistic and combine them using either Fisher's or Tippett's procedures.

Specifically, let P_{τ^2} and P_π denote the p values based on S_{τ^2} and U_π, respectively. Fisher's procedure rejects H_0 if the product $P_{\tau^2}P_\pi$ is small. Under the null H_0, $-2\log(P_{\tau^2})$ and $-2\log(P_\pi)$ each follows a chi-squared distribution with 2 degrees of freedom. Fisher's procedure rejects H_0 at an overall significance level α if $-2\log(P_{\tau^2}) - 2\log(P_\pi) \geq \chi^2_{4,\alpha}$. Alternatively, one can calculate the p value for the combined test by $\Pr\{\chi^2_4 > -2\log(P_{\tau^2}) - 2\log(P_\pi)\}$. For Tippett's procedure, H_0 is rejected if the minimum of P_{τ^2} and P_π is small. Under the null H_0, Tippett's procedure rejects H_0 at an overall significance level α if $\min(P_{\tau^2}, P_\pi) \leq 1 - (1 - \alpha)^{1/2}$. The p value for the combined test using Tippett's procedure is $1 - (1 - \min(P_\pi, P_{\tau^2}))^2$

The power for these two procedures is different under different alternatives. Fisher's procedure is more powerful when both $\pi \neq 0$ and $\tau^2 \neq 0$, whereas Tippett's procedure is more powerful when either the case of $\pi \neq 0$ or the case of $\tau^2 \neq 0$ is true. In the Supplementary Materials Figures S1 and S2 show the acceptance region and power comparison of these two procedures, respectively. It can be seen that when both π and τ^2 are nonzero, the acceptance region curve for Fisher's procedure is closer to 0 than Tippett's procedure. When one of the two alternatives is true, the acceptance region curve for the Tippett's procedure is closer to 0 than for Fisher's procedure. The power curves show a similar trend. Therefore, when both alternatives are likely to be true, Fisher's procedure may be used as it is more powerful; however when only one of the alternatives is true, then Tippett's is more powerful.

Simulation Studies

List of methods for comparison

We conducted extensive simulation studies to evaluate the performance of our proposed mixed effects test (MiST) in terms of type I error and power, and compare it with currently

popular methods, including the burden test [Morris and Zeggini, 2010], SKAT [Wu et al., 2011], SKAT-O [Lee et al., 2012] and EREC [Lin and Tang, 2011]. The choice of these representative tests was in part based on the simulation results from Lin and Tang[2011], who suggested that EREC had similar power to the better one between the burden and the SKAT tests under a wide range of scenarios but more robust than either one. They also showed that EREC outperformed the HP test [Han and Pan, 2010] and the C-alpha test [Neale et al., 2011], which were not included in the comparison to save space. Like EREC, the recently proposed SKAT-O was also shown to have comparable performance of the better one between the burden and SKAT tests. For MiST, we used both Fisher's and Tippett's procedures to combine the score statistics for $\pi = 0$ and $\tau^2 = 0$. The respective combinations are denoted by MiST_F and MiST_T.

In addition, we considered a weighted version for each of the tests, and they are denoted by wBurden, wSKAT, wSKAT-O, wMiST_F and wMiST_T. For wSKAT, we adopted the default weight as suggested in the SKAT software, i.e., $\sqrt{\omega_j} = beta(f_j, 1, 25)$ where f_j is the MAF for the jth variant. To ensure a fair comparison, we used the same weight for MiST and the burden test. For MiST, we let Z in model (3) be a vector of ones or a vector of weights for wMiST, unless otherwise mentioned.

Simulation configurations

We conducted four simulation experiments to evaluate type I error and power of all methods. In the first three experiments, we mimicked the simulation set up used in Lin and Tang [2011] instead of the more natural population genetic models, because it is easy to control the number of variants and their MAFs and can give us a clear understanding of how various tests perform under different scenarios. The fourth experiment was based on the sequencing data from the Dallas Heart Study [Victor et al., 2004] to preserve the linkage disequilibrium pattern within a gene. Due to the computational intensiveness of the EREC test, we only included it in the fourth experiment. Since we used the same simulation set up in the first three experiments as in Lin and Tang [2011], readers can deduce the performance of the EREC test compared with the burden and SKAT tests from their paper.

For the first three experiments, we generated 10 variants in a region under the Hardy-Weinberg equilibrium with MAFs, $f_j = 0.005j$ for $j = 1, \ldots, 10$. In addition, we generated two covariates, one continuous covariate from $Normal(0, 1)$ and one binary covariate from $Bernoulli(0.5)$. For each individual we generated a continuous phenotype from model

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \sum_{j=1}^{10} \beta_j G_j + \varepsilon, \quad (7)$$

where $\alpha_0 = 0.1$, $\alpha_1 = 0.5$ and $\alpha_2 = 0.5$ are the regression coefficients for covariates X_1 and X_2, β_j, $j = 1, \ldots, 10$, are the regression coefficients for the genetic variants, and ε is assumed to follow $Normal(0, 1)$. Different values for the genetic effects β_j, $j = 1, \ldots, 10$, were considered to evaluate the type I error and power, and they will be provided in the following subsections on type I error and power. Unless otherwise specified, for each scenario, we considered two sample sizes: N=500 and N=1000, and we evaluated the type I error and power when the significance level was controlled at $\alpha = 10^{-2}$ and 10^{-3}. A total of 10, 000 simulated data sets were generated for each scenario.

Type I error

To evaluate the type I error of these methods, we generated the phenotype by letting $\beta_j = 0$, for $j = 1, \ldots, 10$. The results are summarized in Table **??**. Each entry is the estimated type I

error divided by the corresponding significance level α. Hence, if the value is close to 1, the estimated type I error is close to the desired level. It can be seen that all tests have nearly correct type I error. We also examined the type I error of MiST when MAFs are $0.001j$ where $j = 1, 2, \ldots, 10$. It is slightly conservative when N = 500; however when N increases to 1000, the type I error is well maintained (results not shown).

Power comparison when all variants are deleterious and when both positive and negative effects are present

We considered two rather extreme scenarios for variant effects, with one favoring the burden test and the other favoring SKAT. The purpose of this simulation setup is to evaluate the robustness of the proposed MiSTs. Specifically, in the first scenario, we set genetic effect β_j to be $H_{a1} : \beta_j = c/\{f_j(1 - f_j)\}^{1/2}$, $j = 1, \ldots, 10$, for some $c > 0$, where f_j was the MAF for the jth variant. This was to assume that all 10 variants had deleterious effects and rarer variants had stronger effects. This scenario was in favor of the (weighted) burden test. For the second one, we assumed β_j to be H_{a2}: $\beta_3 = 1.5c$, $\beta_4 = -1.5c$, $\beta_5 = c$, $\beta_6 = -c$, for some $c > 0$, and zeros for the other βs. Under this scenario, four out of ten variants were associated with the trait, and the associated variants had opposite effects on the trait. This setting was in favor of the variance components test such as SKAT. Here, in both scenarios, the constant c was selected such that the power was reasonably high at different α levels.

Table **??** shows power comparison among different tests under alternatives H_{a1} and H_{a2}. As expected, the burden test, even though performing well under H_{a1}, performs poorly under H_{a2}; whereas SKAT performs well under H_{a2} but poorly under H_{a1}. MiST, which combines both the burden type and the variance component tests, are very robust under these two extreme alternatives. It is somewhat less powerful than the best test that the alternative favors, but generally much more powerful than the worst test when the alternative is not in favor of. Between the two combinations for MiST, Tippett's is modestly more powerful than Fisher's because the alternative is in favor of either the Burden test or SKAT. Being a linear combination of the burden test and SKAT, SKAT-O shows similar robustness to MiST. Since in this simulation rarer causal loci have greater effects, all weighted tests are more powerful than their respective unweighted ones. The relative performance of the weighted tests are similar to that of their unweighted counterparts.

Power comparison when a small subset of the variants are deleterious and the rest are neutral

To show that MiST is not only robust but also more powerful than either the burden or SKAT tests, we conducted another set of simulation experiments. In this set, we again considered two scenarios: $H_{a3} : \beta_1 = \beta_4 = \beta_7 = c$, and zeros for the other βs; and H_{a4}: $\beta_1 = c$, $\beta_4 = 0.5c$, $\beta_7 = 0.25c$, and zeros for the other βs. Under these scenarios, both neutral and causal variants coexisted in the set of variants with three out of 10 loci causal. The difference between the two scenarios were that the effect sizes were the same under H_{a3}, but inversely associated with MAFs under H_{a4}.

Table **??** summarizes the results of all tests for this experiment. Under H_{a3}, for the unweighted tests, our proposed test with Fisher's procedure MiST_F is the most powerful. Compared with the second most powerful test, SKAT, the power of MiST_F increases ranging from 13% to 63% for different sample sizes and significance levels. Similarly, among weighted tests, wMiST_F remains the most powerful, and the power gain compared to wSKAT ranges from 4% to 15%. The power of the SKAT-O (or wSKAT-O) test is slightly less than the SKAT (or wSKAT) test. The burden (or wburden) test has the least power. It is worth noting that even though all causal variants have the same effect, the weighted tests are

still more powerful than their unweighted counterparts, because the MAFs of causal variants are on average smaller than those of neutral variants.

Under the heterogeneous effect size model H_{a4}, both unweighted and weighted MiSTs with Fisher's procedure, (w)$MiST_F$, remain to be most powerful (Table **??**). For unweighted tests, $MiST_F$ gains power 37% ~ 98% compared to the second most powerful test, SKAT-O. For the weighted tests, the power gain of $wMiST_F$ compared to the second most powerful test, wSKAT, ranges from 7% to 26%.

The MiSTs (weighted or unweighted) with Tippett's procedure $MiST_T$ is less powerful than Fisher's. Under these alternatives, only a small proportion of variants are causal, which is generally in favor of the variant components test; meanwhile, all causal variants are deleterious (same direction), which is in favor of the burden test. Therefore, Fisher's procedure, which combines the information from both the heterogeneity (a mixture of causal and neutral variants) and the grouping effect of causal loci, gains power over Tippett's procedure, which tends to be more powerful when only one of the situations is true. To save space, the following simulation is only focused on Fisher's procedure.

Power comparison under a real data situation

This set of simulation for power comparison mimicked a real data situation. We used the sequencing data from the Dallas Heart Study [Vector et al., 2004]. The study included 3409 subjects who were sequenced for three candidate genes: *ANGPTL3*, *ANGPTL4* and *ANGPTL5*. We used the gene *ANGPTL5* as our template to generate the data, and will analyze the Dallas Heart Study data as a real data example in the following section.

There are a total of 93 variants in *ANGPTL5* with MAF $<3\%$, of which 27 are functional (missense, nonsense and frameshift) and 66 are non-functional (noncoding, synonymous and intronic). According to Sunyaev et. al.[2001] and Ng et. al. [2008], about 15%-20% of functional variants are causal. Hence in this simulation, we randomly selected 10%, 25%, and 50% of functional variants to be causal, respectively. Note that due to linkage disequilibrium (LD) among variants, it is likely that more variants are weakly associated with the phenotype. We generated the phenotype Y using the same model (7), with the effect sizes $\beta_j = c|\log_{10} f_j|$, where c is a constant as used in Wu et al.[2011]. We chose $c = 0.5$, 0.25, and 0.1 when 10%, 25% and 50% functional variants were causal, respectively, so that there was adequate power to discern the performance of various tests. Under this model, the effect of a causal variant with MAF 10^{-4} was twice as large as that of a causal variant with MAF 10^{-2}. A total of 1000 data sets were generated, and the power was compared for all tests at a significant level of 0.05. For this set of experiments, we included the EREC test for comparison in addition to Burden, SKAT and SKAT-O. For the weight in EREC, we used the recommended estimated variant regression coefficients with an added constant, 1, for standardized continuous trait.

We performed the analysis restricted to the functional variants only and both MiST and wMiST performed better than other tests (see Table S1 in the Supplementary Materials). Here, we only show the results of analysis using all variants (both functional and nonfunctional) to demonstrate that the MiSTs can easily accommodate possible differential effects for different types of variants. This was achieved by adding an indicator variable in the matrix Z of model (3) when applying the MiST or wMiST. The indicator variable was 1 if the variant was functional and 0 otherwise. The results for including all variants are shown in Table **??**. The proposed MiST and wMiST are considerably more powerful than all other tests. For the unweighted tests, the power gain for $MiST_F$, compared to the second most powerful test, ranges from 44% to 187%. Similarly, for the weighted tests, compared with

the second most powerful test, $wMiST_F$ increases power from 20% to 81%. The Tippett's procedure has less power than the Fisher's and the results are therefore omitted in the table.

To examine the impact of misspecifying variant characteristics on power, we randomly reassigned 10% of functional variants (3 out of 27) to non-functional status and 10% of nonfunctional variants (6 out of 66) to functional. In addition, we specifically chose a causal variant and assigned its functional status to non-functional, which results in 1 out 3, 1 out of 7 and 1 out of 15 causal variants misspecified for 10%, 25% and 50% causal cases, respectively. Under this misspecification scheme, the power for $MiST_F$ decreases from 0.600 to 0.465, losing 22.5% for 10% causal loci; from 0.765 to 0.597 (-21.9%) for 25% causal loci; and from 0.615 to 0.576 (-6.3%) for 50% causal loci. Even though misspecifying variant characteristics results in some power loss particularly when the fraction of misspecified causal variants is high, $MiST_F$ is still more powerful than all other unweighted methods.

In addition to simulations for the continuous trait, we also evaluated the performance of the MiST and wMiST in comparison with other tests for binary trait under this simulation experiment by dichotomizing the continuous trait into a binary variable. The relative performance of various methods is very similar to that of the continuous trait. The results for the binary trait are in the Supplementary Materials Table S2.

The Dallas Heart Study

To further explore the performance of MiST and illustrate how to incorporate variant characteristics in the analysis, we applied our test method to a real data set from the Dallas Heart Study [Victor et al., 2004].

Dallas Heart Study is a population based study aimed to investigate biological and social variables that contribute to differences of cardiovascular health situation among different ethnic groups. Information collected includes gender, ethnic group, age, and triglyceride level. In addition, three candidate genes, *ANGPTL3*, *ANGPTL4*, and *ANGPTL5*, were sequenced. In our analysis we included 3409 individuals, which involve two gender groups (male=1500 and female=1909) and four ethnic groups (black=1762, white=988, hispanic=586, other=73). We focused on rare or less frequent variants, i.e variants with observed MAF <0.05, in the sequencing data. Specifically, in gene *ANGPTL3*, *ANGPTL4*, and *ANGPTL5*, there are 85 (functional=36, nonfunctional=49), 89 (functional=31, non-functional=58), and 96 (functional=27, nonfunctional=69) rare variants, respectively. A variant is functional if it is missense, nonsense, or frameshift, and non-functional if it is noncoding, synonymous, or intronic.

We assessed the association between the log-transformed triglyceride level and each candidate gene, adjusted for gender, ethnic group, and age. We analyzed the data using functional variants only, and all variants including both functional and nonfunctional variants. Both weighted and unweighted tests were conducted. Here we only present the results from the weighted tests (Table **??**), because the results for the unweighted tests are similar. To save space, the unweighted results are provided in the Supplementary Material without further dicussion. In the following, we use $\alpha = 0.05$ as the significant level.

For MiST, there are different ways to specify the variant characteristic matrix Z. We considered two specifications: (1) Z is a vector of $beta(f_j, 1, 25)$ denoted by $wMiST_F$ and $wMiST_T$ for Fishers' and Tippett's procedures; (2) in addition to the vector in (1), for functional variants only analysis Z includes two indicators for missense and frameshift, respectively; and for all variants analysis Z includes three indicators for missense, frameshift

and nonsense. For the second specification, we denote the test by wMiST$_F$(Z) and wMiST$_T$(Z), respectively. The results are presented in Table **??**.

When including functional variants only, gene *ANGPTL5* is significant based on the weighted MiST$_F$(Z) (p=5.17e-05) and MiST$_T$(Z) (p=6.31e-5). Comparing with p values obtained from wMiST without incorporating variant characteristics, we find that wMiST$_F$(Z) and wMiST$_T$(Z) yield smaller p values, indicating a potential power gain for incorporating such information. The p value for the variance component for wMiST(Z) is 0.53 and for π is 4.95e-06, indicating that after adjusting for variant characteristics there is no evidence for the residual individual variant effect. Further examination of individual p values for Z variables shows that the missense variants have the same effect as the nonsense variants (p=0.24), and the frameshift variants have a significant different association with log triglyceride from the the nonsense variants (p=0.004) (see Table S5 in the Supplementary Materials).

When using all variants, we detected an additional significant association with gene *ANGPTL3*. The p values for wMiST$_F$(Z) and wMiST$_T$(Z) are 0.0035 and 0.0052, respectively. We further examined the individual component p values for MiST. The p value for testing $\pi = 0$ is 0.15 and for testing the heterogeneity is 0.0026 (see Table S5 in the Supplementary Material), indicating that the association is largely driven by individual variant effects. We examined the distribution of log-triglyceride level for carriers of each individual variant and found that a few variants are particularly associated with low log-triglyceride level, while majority variants are scattered around the overall mean trait value (see Figure S3 in the Supplementary Materials). For gene *ANGPTL5*, the p values for wMiST$_F$(Z) and wMiST$_T$(Z) are 0.000061 and 0.00011, respectively. The results are consistent with the functional variant analysis. Moreover, individual component p values also show a similar trend compared with the functional variant analysis (Table S5 in the Supplementary Materials).

For all other tests, only the weighted burden test shows a significant association between gene *ANGPTL5* and log triglyceride (Table **??**) (p = 0.00011 for including functional variants only; p = 0.0073 for including all variants). Taken together, our proposed tests, MiST$_F$ and MiST$_T$, are powerful and provide additional information regarding the source of the association.

Discussion

We proposed a mixed effects score test (MiST) based on hierarchical models for testing whether a set of variants is associated with phenotypes accounting for potential heterogeneous variant effects. There are several advantages to hierarchical regression modeling. It includes the usual appealing features for regression models such as adjusting for confounders and being able to accommodate different types of outcomes (e.g., continuous and binary) by using appropriate link functions. It also models the variant effects as a function of (known) variant characteristics to leverage information across loci while still allowing for individual variant effects.

The score statistic that we developed for hierarchical model provides a general framework for testing the set-based association. When the hierarchical model (3) includes only $X_i^T\alpha+(G_i^TZ)\pi$, the score statistic reduces to the burden test. When (3) includes only $X_i^T\alpha+G_i^T\delta$, the score statistic reduces to the SKAT test with linear kernel. We can further extend model (3) to replace $G_i^T\delta$ by $h(G_i)$, where $h(\cdot)$ is defined by a positive, semidefinite kernel function K($\cdot$, $\cdot$) such that $h(G_i)=\sum_{i'=1}^{N}\gamma_{i'}K(G_i,G_{i'})$ for some γ_1, …, γ_N. Assume a

subject specific random effect for γ's with mean 0 and variance τ^2, then $h(G) = (h(G_1), \ldots, h(G_N))^T$ follows a distribution with mean 0 and variance $\tau^2 \mathbb{K}$, where $\mathbb{K}$ is the kernel matrix with (i, i')th element being $K(G_i, G_{i'})$. Some commonly used kernel functions include the linear kernel, Gaussian kernel, and the identity by state kernel [Wu et al., 2011]. Our score statistics in this extended model are still the same as the ones for model (3), where the score statistic for $\pi = 0$ is derived under null ($\pi = 0$ and $\tau^2 = 0$) and the score statistic for τ^2 is derived under τ^2 without constraining $p = 0$. Simply put, the variance component test is same as SKAT except the covariates in our variance component test include not only the confounders X but also the variant characteristics G^Z. Since our proposed score statistic is a combination of the burden and kernel-based tests, it is more robust than either one. Moreover our score test can be more powerful as it makes use of variant characteristic information when available. Furthermore, since our test statistic has a well-defined and easily calculated asymptotic distribution, the computational burden is substantially reduced compared to permutation-based tests.

The proposed combined score statistic tests both $\pi = 0$ and $\tau^2 = 0$. Since it is generally unknown what type of alternatives are likely to be true particularly in a genome-wide association analysis, it is useful to have a global test that is powerful under a wide range of scenarios. However, when a combined test is statistically significant, it is not necessarily clear what may contribute to the statistical significance. Based on our score statistics, we can also test association of specific components, i.e. any parameter in π is 0 or $\tau^2 = 0$, as demonstrated in the analysis of the Dallas Heart Study. We provide p values for component association tests. A small p value may suggest that a specific component is associated with phenotype. We note that these p values are not adjusted for multiple testing and they should be interpreted with caution. Nevertheless these component tests are informative and can provide useful leads for generating hypothesis in future studies.

In rare variant analysis, it is important that common variants do not have a large impact on the group signal such that the effects of rare variants are difficult to detect [Madsen and Browning, 2009]. In the weighted MiST, the variants are weighed differently according to their allele frequencies. This weighting accentuates mutations that are rare in the subjects, so that the test is not completely dominated by common variants. By doing this, we can include variants of all frequencies. However, even though our test is motivated by rare variants association, this set-based association test can also be applied to other situations, for example, gene or pathway analysis of GWAS SNPs. Since most GWAS SNPs are common, the concern of rare variants is not relevant; in this case, unweighted MiST test may be used.

We used Fisher's and Tippett's procedures to combine the independent tests. Generally speaking, Fisher's procedure is more powerful than Tippett's procedure when both group effect and individual variant effect exist, but less powerful when only one is true. From the simulation study, we can see that Fisher's and Tippett's procedures may perform differently under different situations. A useful extension of our current work is to find an optimal weight to combine the two tests under different types of alternatives, accounting for unique features in the data. Research along this line will be presented in future work.

The computation speed for running the proposed test is fast. For example, for the Dallas Heart study, we analyzed three genes, *ANGPTL3*, *ANGPTL4* and *ANGPTL5*, which included 85, 89, and 96 variants, respectively on a total of 3409 subjects. The runtimes for analyzing these genes are 5.08, 6.47, and 5.29 seconds, respectively, using a personal laptop with 2.2 GHz CPU. Based on our experience, the number of variants in a set generally does not affect computational speed unless the number is very large. The software for the proposed tests, MiST, is available upon request.

Supplemental Materials Description

The supplemental materials include two figures (Figure S1 and Figure S2) that show acceptance region and power comparison between Fisher's and Tippett's procedures, Table S1 and S2 for the power comparison for functional variants only analysis and binary trait, Table S3 for the analysis of Dallas Heart Study using unweighted test methods, Table S4 – S5 for component p values for testing $\pi = 0$ and $\tau^2 = 0$ when applying MiST or wMiST to the Dallas Heart Study with or without variant characteristics indicators, and Figure S3 to help visually examine the distribution of each individual variant when there exists significant heterogeneity.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank the investigators of the Dallas Heart Study for generously sharing their data. We also thank Drs. Chongzhi Di and Charles Kooperberg for many stimulating discussions and their critical reading of the manuscript. The work is funded in part by NIH P01AG014358, UC2HL102924, R01AG014358, R01GM085047 and R01-CA059045.

References

1. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Am Stat Assoc. 1993; 88:9–25.
2. Capanu M, Begg CB. Hierarchical modeling for estimating relative risks of rare genetic variants: properties of the pseudo-likelihood method. Biometrics. 2011; 67:371–80. [PubMed: 20707869]
3. Chen GK, Witte JS. Enriching the analysis of genome-wide association studies with hierarchical modeling. Am J Hum Genet. 2007; 81:397–404. [PubMed: 17668389]
4. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. PLoS Biol. 2010; 8:e1000294. [PubMed: 20126254]
5. Exome Variant Server. NHLBI GO Exome Sequencing Project (ESP). Seattle, WA: (URL: http://evs.gs.washington.edu/EVS/) [December, 2012 accessed]
6. Han F, Pan W. A data adaptive sum test for disease association with multiple common or rare variants. Hum Hered. 2010; 70:42–54. [PubMed: 20413981]
7. Hindorff, LA.; MacArthur, J.; Morales, J.; Junkins, HA.; Hall, PN.; Klemm, AK.; Manolio, TA. European Bioinformatics Institute. [Accessed December 7, 2012] A Catalog of Published Genome-Wide Association Studies. 2012. Available at: www.genome.gov/gwastudies
8. Hung RJ, Brennan P, Malaveille C, Porru S, Donato F, Boffetta P, Witte JS. Using hierarchical modeling in genetic association studies with multiple markers: application to a case-control study of bladder cancer. Cancer Epidemiol Biomarkers Prev. 2004; 13:1013–21. [PubMed: 15184258]
9. King CR, Rathouz PJ, Nicolae DL. An evolutionary framework for association testing in resequencing studies. PLoS Genetics. 2010; 6:e1001202. [PubMed: 21085648]
10. Koziol JA, Perlman MD. Combining independent chi-square tests. J Am Stat Assoc. 1978; 73:753–763.
11. Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet. 2007; 80:727–739. [PubMed: 17357078]
12. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012 [Epub ahead of print].
13. 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]

14. Lin DY, Tang ZZ. A general framework for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet. 2011; 89:354–367. [PubMed: 21885029]
15. Lin X. Variance component testing in generalized linear models with random effects. Biometrika. 1997; 84:309–326.
16. Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics. 2008; 9:292. [PubMed: 18577223]
17. Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics. 2007; 63:1079–1088. [PubMed: 18078480]
18. Liu H, Tang Y, Zhang H. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. Computational Statistics & Data Analysis. 2009; 53:853–856.
19. Madsen BE, Browning SR. Methods for detecting association with rare variants for common diseases: Application to analysis of sequence data. Am J Hum Genet. 2009; 83:311–321.
20. Maher B. Personal genomes: the case of the missing heritability. Nature. 2008; 456:18–21. [PubMed: 18987709]
21. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461:747–753. [PubMed: 19812666]
22. 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). Mutat Res. 2007; 615:28–56. [PubMed: 17101154]
23. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidem. 2010; 34:188–193.
24. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ, et al. Testing for an usual distribution of rare variants. PLoS Genetics. 2011; 7:161–165.
25. Nelson MR, Wegmann D, Ehm MG, et al. An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. Science. 2012; 337:100–104. [PubMed: 22604722]
26. Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, Axelrod N, Busam DA, Strausberg RL, Venter C. Genetic variation in an individual human exome. PLoS Genet. 2008; 4:e1000160. [PubMed: 18704161]
27. Sunyaev S, Ramensky V, Koch I, Lathe W, Kondrashov A, Bork P. Prediction of deleterious human alleles. Hum Molec Genet. 2001; 10:591–597. [PubMed: 11230178]
28. Tennessen JA, Bigham AW, O'Connor TD, Fu W, et al. Evolution and functional impact of rare coding variation from deep sequencing of human wxomes. Science. 2012; 337:64–69. [PubMed: 22604720]
29. Victor RG, Haley RW, Willett DL, Peshock RM, Vaeth PC, Leonard D, Basit M, Cooper RS, Iannacchione VG, Visscher WA, et al. The Dallas Heart Study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. Am J Cardiol. 2004; 93:1473–1480. [PubMed: 15194016]
30. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89:82–93. [PubMed: 21737059]
31. Zhang D, Lin X. Hypothesis testing in semiparametric additive mixed models. Biostatistics. 2003; 4:57–74. [PubMed: 12925330]

Appendix: Independence of S_{τ^2} and U_π

To prove that the score statistics $U_\pi = (GZ)^T(Y - \tilde{\mu})$ and $S_{\tau^2} = (Y - \hat{\mu})^T GG^T(Y - \hat{\mu})$, where $\tilde{\mu}$ and $\hat{\mu}$ are defined as in equation (4) and equation (6), are independent, we first note that S_{τ^2} can be written as $S_{\tau^2} = U_{\tau^2}^T U_{\tau^2}$, where $U_{\tau^2} = G^T(Y - \hat{\mu})$. In addition, if we apply the identity link for continuous traits, or logit link for binary traits, by using the Taylor

expansion we can show the following two approximation $Y-\tilde{\mu}\approx(I-DX(X^TDX)^{-1}X^T(Y-\mu)$ and $Y-\hat{\mu}\approx(I-DM(M^TDM)^{-1}M^T(Y-\mu)$, where $M=(X, GZ)$ and D is the variance-covariance matrix of Y given $\tau^2=0$. Denote $P_1=I-DX(X^TDX)^{-1}X^T$ and $P_2=I-DM(M^TDM)^{-1}M^T$, we have $U_\pi=(GZ)^TP_1(Y-\mu)$ and $U_{\tau^2}=G^TP_2(Y-\mu)$.

Define a $(q+p)\times 1$ random vector $V^T=(U_\pi^T, U_{\tau^2}^T)$. By the central limit theorem and the law of large numbers, V converges to a multivariate normal distribution with mean vector **0** and covariance matrix Σ with diagonal matrices $\sum_{11}=(GZ)^TP_1DP_1^T(GZ)$ and $\sum_{22}=G^TP_2DP_2^TG$, and the off-diagonal matrices $\sum_{12}=\sum_{21}=\sum_{21}^T=Cov(U_\pi, U_{\tau^2})$.

Since U_π and U_{τ^2} have asymptotically a joint normal distribution, to prove that they are asymptotically independent, we only need to prove that $Cov(U_\pi, U_{\tau^2})=\mathbf{0}$. This equation is true because if we denote $P_{02}=P_2D=D-DM(M^TDM)^{-1}M^TD$, then we have $M^TP_{02}=0$, which implies $X^TP_{02}=0$ and $(GZ)^TP_{02}=0$ since $M=(X, GZ)$. Hence

$$\begin{aligned}Cov(U_\pi, U_{\tau^2})&=Cov\left((GZ)^TP_1(Y-\mu), G^TP_2(Y-\mu)\right)\\&=(GZ)^TP_1Cov(Y-\mu, Y-\mu)P_2^TG\\&=(GZ)^TP_1DP_2^TG\\&=(GZ)^T\left(1-DX(X^TDX)^{-1}X^T\right)P_{02}G\\&=(GZ)^TP_{02}G\\&=0.\end{aligned}$$

Table 1

Type I error[a] for both unweighted and weighted burden, SKAT, SKAT-O, EREC, and MiST with Fisher's and Tippett's procedures. The sample size N = 500 and 1000 and the significant level $\alpha = 10^{-2}$ and 10^{-3}.

		Unweighted test					Weighted test					
N	α	Burden	SKAT	SKAT-O	$MiST_F$	$MiST_T$	wBurden	wSKAT	wSKAT-O	EREC	$wMiST_F$	$wMiST_T$
500	10^{-2}	1.062	0.999	1.073	0.997	1.001	1.051	0.952	0.961	1.048	0.983	0.964
	10^{-3}	1.050	1.120	1.050	1.100	1.020	1.040	0.920	0.880	1.020	0.920	0.980
1000	10^{-2}	1.042	0.997	1.012	1.024	0.991	1.044	0.990	0.965	1.041	1.016	0.968
	10^{-3}	0.990	1.110	1.010	1.000	1.010	1.020	0.910	0.940	0.980	0.990	0.980

[a] divided by α

Table 2

Power for both unweighted and weighted burden, SKAT, SKAT-O, EREC, and MiST with Fisher's and Tippett's procedures. The two alternatives considered are in favor of the (weighted) burden test and the variance component test (SKAT test). The sample size $N = 500$ and 1000 and the significant level $\alpha = 10^{-2}$ and 10^{-3}.

		Unweighted test					Weighted test				
N	***α***	**Burden**	**SKAT**	**SKAT-O**	**MiST$_F$**	**MiST$_T$**	**wBurden**	**wSKAT**	**wSKAT-O**	**wMiST$_F$**	**wMiST$_T$**
		$H_{a_1}: \beta_j = c/\{p_j(1-p_j)\}^{1/2}, j = 1, \ldots, 10$									
500	10^{-2}	0.515	0.170	0.444	0.392	0.423	0.534	0.174	0.455	0.408	0.441
	10^{-3}	0.247	0.048	0.199	0.159	0.187	0.264	0.047	0.206	0.169	0.201
1000	10^{-2}	0.866	0.435	0.818	0.780	0.811	0.885	0.483	0.846	0.805	0.841
	10^{-3}	0.652	0.190	0.577	0.522	0.576	0.687	0.223	0.629	0.558	0.623
		$H_{a_2}: \beta_3 = 1.5c, \beta_4 = -1.5c, \beta_5 = c, \beta_6 = -c;$									
500	10^{-2}	0.009	0.423	0.320	0.349	0.391	0.010	0.814	0.748	0.745	0.783
	10^{-3}	0.001	0.141	0.089	0.110	0.138	0.001	0.593	0.507	0.508	0.566
1000	10^{-2}	0.014	0.507	0.397	0.417	0.455	0.012	0.861	0.803	0.796	0.829
	10^{-3}	0.002	0.198	0.130	0.151	0.177	0.001	0.664	0.584	0.571	0.628

Table 3

Power for both unweighted and weighted burden, SKAT, SKAT-O, EREC, and MiST with Fisher's and Tippett's procedures. The two alternatives are such that both neutral and causal variants coexist with one that all causal loci have the same effect and the other that causal loci have heterogeneous effects. The sample size $N = 500$ and 1000 and the significant level $\alpha = 10^{-2}$ and 10^{-3}.

		Unweighted test					Weighted test				
N	α	Burden	SKAT	SKAT-O	MiST_F	MiST_T	wBurden	wSKAT	wSKAT-O	MiST_F	MiST_T
$H_{a3}: \beta_1 = \beta_4 = \beta_7 = c$											
500	10^{-2}	0.303	0.502	0.491	0.626	0.465	0.389	0.692	0.672	0.736	0.589
	10^{-3}	0.107	0.204	0.202	0.333	0.170	0.156	0.408	0.386	0.470	0.274
1000	10^{-2}	0.283	0.578	0.551	0.652	0.515	0.344	0.648	0.627	0.675	0.529
	10^{-3}	0.098	0.291	0.264	0.376	0.226	0.132	0.372	0.356	0.401	0.232
$H_{a4}: \beta_1 = c, \beta_4 = 0.5c, \beta_7 = 0.25c$											
500	10^{-2}	0.261	0.286	0.326	0.479	0.325	0.435	0.684	0.683	0.763	0.614
	10^{-3}	0.086	0.076	0.103	0.204	0.093	0.187	0.398	0.395	0.502	0.294
1000	10^{-2}	0.288	0.415	0.427	0.583	0.429	0.480	0.761	0.755	0.811	0.669
	10^{-3}	0.100	0.154	0.163	0.297	0.155	0.218	0.495	0.492	0.567	0.344

Table 4

Power for both unweighted and weighted burden, SKAT, SKAT-O, EREC, and MiST with Fisher's combinations. A real sequencing data was used in this simulation, where a continuous trait was generated with a subset of functional variants selected as causal variants and the effect sizes inversely proportional to the observed MAF, c| $\log_{10}$(MAF)|, where c = 0.5, 0.25, 0.1 for 10%, 25% and 50% causal loci. All variants are included.

	unweighted test				weighted test				
% of causal	Burden	SKAT	SKAT-O	$MiST_F$	wBurden	wSKAT	wSKAT-O	EREC	$wMiST_F$
10% (c = 0.5)	0.167	0.227	0.224	0.600	0.220	0.619	0.527	0.325	0.745
25% (c = 0.25)	0.218	0.235	0.267	0.765	0.300	0.560	0.538	0.334	0.784
50% (c = 0.1)	0.311	0.418	0.428	0.615	0.286	0.328	0.355	0.338	0.624

Table 5

p values for weighted burden test, SKAT, SKAT-O, and MiST with Fisher's and Tippett's combinations when applied to Dallas Heart Study. Associations between the set of functional or all variants and each of three genes are tested. The continuous phenotype trait is log triglyceride level.

	wBurden	wSKAT	wSKAT-O	EREC	$wMiST_F$	$wMiST_T$	$wMiST_F(Z)$	$wMiST_T(Z)$
				Functional variants only				
ANGPTL3	0.829	0.397	0.570	0.358	0.363	0.403	0.254	0.274
ANGPTL4	0.757	0.309	0.471	0.378	0.064	0.061	0.768	0.733
ANGPTL5	0.0001	0.380	0.352	0.091	0.052	0.059	0.00005	0.00006
				All variants				
ANGPTL3	0.075	0.493	0.473	0.330	0.003	0.008	0.003	0.005
ANGPTL4	0.528	0.235	0.384	0.441	0.336	0.351	0.439	0.316
ANGPTK5	0.007	0.080	0.152	0.429	0.005	0.015	0.00006	0.0001