Multi-SKAT: General framework to test multiple phenotype associations of rare variants

Diptavo Dutta[1,2], Laura Scott[1,2], Michael Boehnke[1,2], and Seunggeun Lee *[1,2]

[1]Department of Biostatistics

[2]Center for Statistical Genetics

University of Michigan

Ann Arbor, Michigan, USA

Abstract

In genetic association analysis, a joint test of multiple distinct phenotypes can increase power to identify sets of trait-associated variants within genes or regions of interest. Existing multi-phenotype tests for rare variants make specific assumptions about the patterns of association of underlying causal variants, and the violation of these assumptions can reduce power to detect association. Here we develop a general framework for testing pleiotropic effects of rare variants based on multivariate kernel regression (Multi-SKAT). Multi-SKAT models effect sizes of variants on the phenotypes through a kernel matrix and performs a variance component test of association. We show that many existing tests are equivalent to specific choices of kernel matrices with the Multi-SKAT framework. To increase power to detect association across tests with different kernel matrices, we developed a fast and accurate approximation of the significance of the minimum observed p-value across tests. To account for related individuals, our framework uses a random effects for the kinship matrix. Using simulated data and amino acid and exome-array data from the METSIM study, we show that Multi-SKAT can improve power over single-phenotype SKAT-O test and existing multiple phenotype tests, while maintaining type I error rate.

*Corresponding author. Email: *leeshawn@umich.edu*

Introduction

With the advent of array genotyping technologies, genome-wide association studies (GWAS) have identified numerous genetic variants associated with complex traits. Despite these discoveries, GWAS loci explain only a small proportion of heritability for most traits. This may be due, in part, to the fact that the analysis framework of such association studies is underpowered to identify associations in rare variants[1]. It has been argued that rare variants play important roles in the underlying genetics of complex traits[2;3]. To identify such rare variant associations, gene- or region-based multiple variant tests have been developed. By jointly testing rare variants in a target gene or region, these methods can increase power over a single variant test and are now a standard approach in rare variant analysis[4].

Recent GWAS results have shown that many GWAS loci are associated with multiple traits[5]. For example, 44% of autoimmune risk SNPs have been estimated to be associated with multiple autoimmune diseases[6]. Nearly 17% of genes in NHLBI GWAS categories are associated with multiple traits[7]. Detecting such pleiotropic effects is important to understand the underlying biological structure of complex traits. In addition, by leveraging cross-phenotype associations, the power for detecting trait-associated variants can be increased.

Identifying the cross-phenotype effects requires a suitable joint or multivariate analysis framework that can leverage the inter-dependency of the phenotypes. Various methods have been proposed for multiple phenotype analysis in GWAS[8;9;10;11;12]. Extending them, several groups have developed multiple phenotype tests for rare variants[13;14;15;16;17;18;19]. For example, Wang et al.[13] proposed using a multivariate functional linear model (MFLM) to this end; Broadaway et al.[14] used a non-parametric approach, called GAMuT, to test such effects and Wu et al.[15] developed a score based sequence kernel association test for multiple traits, MSKAT, which has been shown to be similar in performance to GAMuT[14].

Despite these developments, existing methods have important limitations. Since these methods were developed under specific assumptions on genetic effect structure to multiple phenotypes, specifically the relationship between the effects of different genetic variants within a region, they can lose power if the underlying assumptions are not satisfied[12]. For example, if genetic effects are heterogeneous across multiple phenotypes, methods assuming homogeneous genetic effects can lose substantial amount of power[16]. Although combining multiple models can improve power for common variants[12], no rare variant tests have aggregated different models for multiple phenotypes to yield robust results. In addition, most existing methods and softwares cannot adjust for the relatedness between individuals; thus, to apply these methods, related individuals must be removed from the analysis to maintain type I error rates. For example, in the study used for our data application, the METabolic Syndrome In Men (METSIM) study, nearly 15% of

individuals are estimated to be related up to the second degree, and removing them from the analysis would greatly reduce the power.

In this article, we develop a class of methods Multi-SKAT that extend the kernel regression framework in the mixed effect models to incorporate multiple correlated continuous phenotypes while accounting for family relatedness. Mixed effect models have been useful modeling structures for rare-variant association tests. For example, popular rare variant tests such as SKAT[20] and SKAT-O[21] are based on mixed effect models. By using kernels for the correlated phenotypes and multiple variants in a target region, Multi-SKAT allows for flexible modeling of the genetic effects on the phenotypes. Many of the existing methods for multiple phenotype rare variant tests can be viewed as special cases of Multi-SKAT with particular choices of kernels. Since Multi-SKAT is score statistics-based, it requires fitting the null model only once for genome-wide data. In addition, Multi-SKAT is computationally efficient as it can calculate asymptotic p-values. To avoid loss of power due to model misspecification, we develop computationally efficient omnibus tests, which allow aggregating tests over several kernels producing robust power of detection. Additionally, Multi-SKAT can correct for relatedness by incorporating estimated kinship information.

The article is organized as follows: in the first section, we present the multivariate mixed effect model and kernel matrices. We particularly focus on the phenotype-kernel and describe omnibus procedures that can aggregate results across different choices of kernels and kinship adjustment. We then describe the simulation experiments that clearly demonstrate that Multi-SKAT tests have increased power to detect associations compared to existing methods like GAMuT, MSKAT and others. Finally, we apply Multi-SKAT tests to detect the cross-phenotype effects of rare nonsynonymous and protein-truncating variants on a set of nine amino acids from 8,545 Finnish men from the METSIM study.

Material and Methods

Single-phenotype region-based tests

To describe the Multi-SKAT tests, we first present the existing model of the single-phenotype gene or region-based tests. Let $y_k = (y_{1k}, y_{2k}, \cdots, y_{nk})^T$ be an $n \times 1$ vector of the kth phenotype over n individuals; X an $n \times q$ matrix of the q non-genetic covariates including the intercept; $G_j = (G_{1j}, \cdots, G_{nj})^T$ is an $n \times 1$ vector of the minor allele counts (0, 1, or 2) for a binary genetic variant j; and $G = [G_1, \cdots, G_m]$ is an $n \times m$ genotype matrix for m genetic variants in a target region. The following regression model can be used to

relate m genetic variants to phenotype k,

$$y_k = X\alpha_k + G\beta_{.k} + \epsilon_k \tag{1}$$

where α_k is a $q \times 1$ vector of regression coefficients of q non-genetic covariates, $\beta_{.k} = (\beta_{1k}, \cdots, \beta_{mk})^T$ is an $m \times 1$ vector of regression coefficients of the m genetic variants, and ϵ_k is an $n \times 1$ vector of non-systematic error term with each element following $N(0, \sigma_k^2)$. To test for $H_0 : \beta_{.k} = 0$, a variance component test under the mixed effects model have been proposed to increase power over the usual F-test[20]. The variance component test assumes that the regression coefficients, $\beta_{.k}$, are random variables and follow a centered distribution with variance $\tau^2\Sigma_G$ (see below). Under these assumptions, the test for $\beta_{.k} = 0$ is equivalent to testing $\tau = 0$. The score statistic for this test is

$$Q = (y_k - \hat{\mu}_k)^T G\Sigma_G G^T (y_k - \hat{\mu}_k) \tag{2}$$

where $\hat{\mu}_k = X\hat{\alpha}_k$ is the estimated mean of y_k under the null hypothesis of no association. The test statistic Q asymptotically follows a mixture of chi-squared distributions under the null hypothesis and p-values can be computed by inverting the characteristic function[22].

The kernel matrix Σ_G plays a critical role; it models the relationship among the effect sizes of the variants on the phenotypes. Any positive semidefinite matrix can be used for Σ_G providing a unified framework for the region-based tests. A frequent choice of Σ_G is a sandwich type matrix $\Sigma_G = WR_GW$, where $W = diag(w_1, .., w_m)$ is a diagonal weighting matrix for each variant, and R_G is a correlation matrix between the effect sizes of the variants. $R_G = I_{m\times m}$ implies uncorrelated effect sizes and corresponds to SKAT, and $R_G = 1_m {1_m}^T$ corresponds to the burden test, where $I_{m\times m}$ is an $m \times m$ diagonal matrix and $1_m = (1, \cdots 1)^T$ is an $m \times 1$ vector with all element being unity. Furthermore, a linear combination of these two matrices corresponds to $R_G = \rho 1_m {1_m}^T + (1-\rho)I_{m\times m}$, which is used for SKAT-O.

Multiple-phenotype region-based tests

Extending the idea of using kernels, we build a model for multiple phenotypes. To model K correlated phenotypes in n unrelated individuals, we use the following multivariate linear model

$$Y = XA + GB + E \tag{3}$$

where $Y = (y_1, \cdots, y_K)$ denotes an $n \times K$ phenotype matrix; A is an $n \times q$ matrix of coefficients of X; $B = (\beta_{ij})$ is a $p \times K$ matrix of coefficients where β_{ij} denotes the effect of the ith variant on the jth phenotype and E is an $n \times K$ matrix of non-systematic errors. Let $vec(\cdot)$ denote the matrix vectorization function, and then $vec(E)$ follows $N(0, I_n \otimes V)$, where V is a $K \times K$ covariance matrix and $\otimes$ represents the Kronecker product.

In addition to assuming that $\beta_{.k}$ follows a centered distribution with covariance $\tau^2\Sigma_G$, we further assume that $\beta_{j.} = (\beta_{j1}, \cdots, \beta_{jK})^T$, which is the regression coefficients of variant j to K multiple phenotypes, follows a centered distribution with covariance $\tau^2\Sigma_P$, which implies that $vec(B)$ follows a centered distribution with covariance $\tau^2\Sigma_G \otimes \Sigma_P$. Similarly as before, the null hypothesis $H_0 : vec(B) = 0$ is equivalent to $\tau = 0$. The corresponding score test statistic is

$$Q = \{vec(Y) - vec(\widehat{\mu})\}^T \left\{(G\Sigma_G G^T) \otimes (\widehat{V}^{-1}\Sigma_P \widehat{V}^{-1})\right\} \{vec(Y) - vec(\widehat{\mu})\} \tag{4}$$

where $\widehat{\mu}$ and $\widehat{V}$ are the estimated mean and covariance of Y under the null hypothesis.

We note that Σ_P plays a similar role as Σ_G but with respect to phenotypes. Σ_P represents a kernel in the phenotypes space and models the relationship among the effect sizes of a variant on each of the phenotypes. Any positive semidefinite matrix can be used as Σ_P.

The proposed approach provides a double-flexibility in modeling. Through the choice of structures for Σ_G and Σ_P, we can control the dependencies of genetic effects. Most of our hypotheses about the underlying genetic structure of a set of phenotypes can be modeled through varying structures of these two matrices.

Phenotype kernel structure Σ_P

The use of Σ_G has been extensively studied previously in literature[20;21]. Here we propose several choices for Σ_P and study their effect from a modeling perspective.

Homogeneous (Hom) Kernel

It is possible that effect sizes of a variant on different phenotypes to be homogeneous in which $\beta_{j1} = \cdots = \beta_{jK}$. Under this assumption,

$$\Sigma_{P,Hom} = 1_K 1_K^T \tag{5}$$

Under $\Sigma_{P,Hom}$, $\beta_{jk}, (k = 1, \cdots K)$ for a variant j are the same for all the phenotypes.

Heterogeneous (Het) Kernel

Effect sizes of a variant to different phenotypes can be heterogeneous in which $\beta_{j1} \neq \cdots \neq \beta_{jK}$. Under this assumption, we can construct

$$\Sigma_{P,Het} = I_{k\times k} \tag{6}$$

The $\Sigma_{P,Het}$ implies that $(\beta_{j1}, \cdots, \beta_{jK}^T)$ are uncorrelated. This also indicates that the correlation among the phenotypes is not affected by this particular region or gene.

Phenotype Covariance (PhC) Kernel

We may model Σ_P as proportional to the estimated residual covariance across the phenotypes as,

$$\Sigma_{P,PhC} = \widehat{V} \tag{7}$$

where $\widehat{V}$ is the estimated covariance matrix among the phenotypes. This model assumes that the correlation of the phenotypes is proportional to that of between the effect sizes.

Principal Component (PC) Kernel

Principal component analysis (PCA) is a popular tool for multivariate analysis. In multiple phenotype tests, PC-based approaches have been used to reduce the dimension in phenotypes[23]. Here we show that PC-based approach can be included in our framework.

Let L_i be the loading vector for the ith PC, which produces the ith PC score $P_i = YL_i$. In PCA-based analysis, PC scores are used as outcomes instead of original Y. Since the genetic information regarding the phenotypes may not be confined to the top few PCs[23], we first consider using all PCs. Let $P = (P_1, \cdots P_K)$. Since PCs are orthogonal, we assume genetic effects to multiple PCs are heterogeneous, which resulted in

$$Q = \{vec(P) - vec(\widehat{\mu}_P)\}^T \left\{ \left(G\Sigma_G G^T\right) \otimes \left(\widehat{V}_P^{-1}\widehat{V}_P^{-1}\right) \right\} \{vec(P) - vec(\widehat{\mu}_P)\} \tag{8}$$

where $\widehat{\mu}_P$ is the mean of P under the null hypothesis and $\widehat{V}_P$ is the estimated covariance matrix between the PC's. $\widehat{V}_P$ will be a diagonal matrix since PCs are orthogonal. The equation (8) can be written as

$$Q = \{vec(y) - vec(\hat{\mu})\}^T \left\{ \left(G\Sigma_G G^T\right) \otimes \left(L\widehat{V}_P^{-1}\widehat{V}_P^{-1}L^T\right) \right\} \{vec(y) - vec(\hat{\mu})\} \tag{9}$$

where $L = (L_1, \cdots, L_K)$ is a $K \times K$ PC loading matrix. The equation (9) shows that by using $\Sigma_{P,PC} = \widehat{V}L\widehat{V_P}^{-1}\widehat{V_P}^{-1}L^T\widehat{V}$, we can carry out PC-based tests.

Instead of using all the PC's, we can use selected PC's that represent the majority of cumulative variation

in phenotypes. For example, we can jointly test the PC's that have cumulative variance of 90%. If the top t PC's have been chosen for analysis using ν% cumulative variance as cutoff, we can use

$$\Sigma_{P,PC-\nu} = \widehat{V} L_{sel} \widehat{V}_P^{-1} \widehat{V}_P^{-1} L_{sel}^T \widehat{V}$$

where $L_{sel} = [L_1, \cdots, L_t, 0, \cdots, 0]$ and 0 represents a vector of 0's of appropriate length.

Relationship with other Multiple-Phenotype rare variant tests

We have proposed a uniform framework of Multi-SKAT tests that depend on Σ_G and Σ_P. There are certain specific choices of kernels that correspond to other published methods.

- Using $\Sigma_{P,PhC}$ and $\Sigma_G = WI_mW^T$ is identical to GAMuT[14](with linear kernel) and MSKAT[15].
- Using $\Sigma_{P,Hom}$ and $\Sigma_G = WI_mW^T$ is identical to hom-MAAUSS[16].
- Using $\Sigma_{P,Het}$ and $\Sigma_G = WI_mW^T$ is identical to het-MAAUSS[16] and MF-KM[19].

For the detailed proof, please see Appendix A.

Minimum p-value based omnibus tests (minP)

The model and the corresponding test of association that we propose has two parameters, Σ_G and Σ_P, which are absent in the null model of no association. Since our test is a score test, Σ_G and Σ_P cannot be estimated from the data. One possible remedy is to select Σ_G and Σ_P based on prior knowledge; however, if the selected Σ_G and Σ_P do not reflect underlying biology, the test may have substantially reduced power[12;16]. In an attempt to achieve robust power, we aggregate results across different Σ_G and Σ_P. Here we propose a general strategy to combine the p-values of a number of Multi-SKAT tests with different kernels to produce an aggregate p-value of association.

Suppose p_i is the p-value for Q_i with given i^{th} Σ_G and Σ_P, $i = 1, \cdots, b$, and $T_P = (p_1, \cdots, p_b)^T$ is an $b \times 1$ vector of p-values of b such Multi-SKAT tests. The minimum p-value test statistic after the Bonferroni adjustment is $b \times p_{min}$, where p_{min} is the minimum of the b p-values. In the presence of positive correlation among the tests, this approach is conservative and hence might lack power of detection. Rather than using Bonferroni corrected p_{min}, more accurate results can be obtained if the joint null distribution or more specifically the correlation structure of T_P can be estimated. Here we adopt a resampling based approach to estimate this correlation structure. Note that our test statistic is equivalent to

$$Q = \{vec(y) - vec(\widehat{\mu})\}^T (I_n \otimes \widehat{V}^{-1})(G \otimes I_K)(\Sigma_G \otimes \Sigma_P)(G^T \otimes I_K)(I_n \otimes \widehat{V}^{-1}) \{vec(y) - vec(\widehat{\mu})\} \qquad (10)$$

Under the null hypothesis $(I_n \otimes \widehat{V}^{-\frac{1}{2}})(vec(y) - vec(\widehat{\mu}))$ approximately follows an uncorrelated normal distribution $N(0, I_{nK})$. Using this, we propose the following resampling algorithm

- Step 1. Generate nK samples from an $N(0,1)$ distribution, say S_R.
- Step 2. Calculate b different test statistics as $T_R = S_R^T \left\{ (G\Sigma_G G^T) \otimes (\widehat{V}^{-\frac{1}{2}} \Sigma_P \widehat{V}^{-\frac{1}{2}}) \right\} S_R$ for all the choices of Σ_P and calculate p-values.
- Step 3. Repeat the previous steps independently for $R(= 1000)$ iterations, and calculate correlation between the p-values of the tests from the R resampling p-values.

With the estimated null correlation structure, we use Copula to approximate the joint distribution of T_P [24]. Copula is a statistical approach to construct joint multivariate distribution using marginal distribution of each variable and correlation structure. Since marginally each test statistic Q follows a mixture of chi-square distribution, which has a heavier tail than normal distribution, we propose to use a t-Copula to approximate the joint distribution, i.e, we assume the joint distribution of T_P to be multivariate t with the estimated correlation structure. The final p-value for association is then calculated from the distribution function of the assumed t-Copula.

When calculating the correlation across the p-values, Pearson's correlation coefficient can be unreliable since it depends on normality and homoscedasticity assumptions. To avoid such assumptions we recommend estimating the null-correlation matrix of the p-values through Kendall's tau (τ), which is a non-parametric approach based on concordance of ranks.

The minimum p-value approach can be used to combine different Σ_P given Σ_G, or combine both Σ_P and Σ_G. To differentiate the latter, we will call it minP_{com} which combines SKAT and Burden type kernels of Σ_G.

Adjusting for relatedness

We formulated model (3) and corresponding tests under the assumption of independent individuals. If individuals are related, this assumption is no longer valid, and the tests may have inflated type I error rates. Since our method is regression-based, we can relax the independence assumption by introducing a random effect term representing the correlation structure of phenotypes among related individuals.

Let Φ be the kinship matrix of the individuals and V_g is a co-heritability matrix, denoting the shared heritability between the phenotypes. Extending the model (3), we incorporate Φ and V_g as

$$Y = XA + GB + Z + E \tag{11}$$

where Z is an $n \times K$ matrix with $vec(Z)$ following $N(0, \Phi \otimes V_g)$. Z represents a matrix of random effects arising from shared genetic effects between the samples due to the relatedness. The remaining terms are the same as in (3). The corresponding score test statistic is

$$Q_{Kin} = \{vec(Y) - vec(\widehat{\mu})\}^T \widehat{V_e}^{-1} (G \otimes I_k)(\Sigma_G \otimes \Sigma_P)(G^T \otimes I_k) \widehat{V_e}^{-1} \{vec(y) - vec(\widehat{\mu})\} \tag{12}$$

where $\widehat{V_e} = \Phi \otimes \widehat{V_g} + I_n \otimes \widehat{V}$ is the estimated covariance matrix of $vec(Y)$ under the null hypothesis. Similar to the previous versions for unrelated individuals, Q_{Kin} asymptotically follows a mixture of chi-square under the assumption of no association.

This approach depends on the estimation of the matrices Φ, V_g and V. The kinship matrix Φ can be estimated using the genome-wide genotype data[25]. Several of the published methods like LD-Score[26], PHENIX[27] and GEMMA[10 28] can jointly estimate V_g and V. In our numerical analysis, we have used PHENIX. This is an efficient method to fit local maximum likelihood variance components in a multiple phenotype mixed model through an E-M algorithm.

Once the matrices Φ, V_g and V are estimated, we compute the asymptotic p-values for Q_{Kin} by using a mixture of chi-square distribution. The computation of Q_{Kin} requires large matrix multiplications, which can be time and memory consuming. To reduce computational burden, we employ several transformations. We perform an eigen-decomposition on the kinship matrix Φ as $\Phi = U \Lambda U^T$, where U is an orthogonal matrix of eigenvectors and Λ is a diagonal matrix of corresponding eigenvalues. We obtain the transformed phenotype matrix as $\widetilde{Y} = YU$, the transformed covariate matrix as $\widetilde{X} = XU$, the transformed random effects matrix $\widetilde{Z} = ZU$ and transformed residual error matrix $\widetilde{E} = EU$. The equation (11) can be transformed into

$$\widetilde{Y} = \widetilde{X}A + \widetilde{G}B + \widetilde{Z} + \widetilde{E}; \quad vec(\widetilde{Z}) \sim N(0, \Lambda \otimes V_g); \quad vec(\widetilde{e}) \sim N(0, I \otimes V) \tag{13}$$

For each individual l, the transformed phenotypes given the transformed covariates asymptotically follow independent (but not identical) multivariate normal distributions.

$$\widetilde{y}_l = A\widetilde{x}_l + B\widetilde{g}_l + \widetilde{z}_l + \widetilde{e}_l; \quad \widetilde{z}_l \sim N(0, d_l \otimes V_g); \quad \widetilde{e}_l \sim N(0, V); \tag{14}$$

All the properties of the tests in the model (3) are directly applicable to those for (13). Q_{Kin} can be computed from this transformed model as,

$$Q_{Kin} = \left\{vec(\widetilde{Y}) - vec(\widehat{\mu})\right\}^T \widetilde{V}_e^{-1} (\widetilde{G} \otimes I_k)(\Sigma_G \otimes \Sigma_P)(\widetilde{G}^T \otimes I_k) \widetilde{V}_e^{-1} \left\{vec(\widetilde{Y}) - vec(\widehat{\mu})\right\} \tag{15}$$

where $\widetilde{V}_e = \Lambda \otimes \widehat{V_g} + I_n \otimes \widehat{V}$. Asymptotic p-values can be obtained from the corresponding mixture of chi-squares distribution. Further, omnibus strategies developed for (3) are applicable in this case with similar modifications. For example, the resampling algorithm for minimum p-value based omnibus test can be implemented here as well by noting that $\widetilde{V}_e^{-\frac{1}{2}} \left\{ vec(\widetilde{Y}) - vec(\widehat{\mu}) \right\}$ approximately follows an uncorrelated normal distribution.

Simulations

We carried out extensive simulation studies to evaluate the type I error control and power of Multi-SKAT tests. Our simulation studies focus on evaluating the proposed tests when the number of phenotypes are 5, 6 or 9. For type I error simulations without related individuals and all power simulations, we generated 10,000 chromosomes over 1Mbp regions using a coalescent simulator with European demographic model[29]. Since the average length of the gene is 3 kbps we randomly selected a 3 kbps region for each simulated dataset to test for associations. For the type I error simulations with related individuals, to have a realistic kinship structure, we used the METSIM study genotype data.

Phenotypes were generated from the multivariate normal distribution as

$$y_i \sim MVN\{(\beta_1 G_1 + \cdots + \beta_m G_m) I, V\}, \tag{16}$$

where $y_i = (y_{i1}, \cdots, y_{iK})^T$ is the outcome vector, G_j is the genotype of the jth variant, and β_j is the corresponding effect size, and V is a covariance of the non-systematic error term. I is a $k \times 1$ indicator vector, which has 1 when the corresponding phenotype is associated with the region and 0 otherwise. For example, if there are 5 phenotypes and the last three are associated with the region, $I = (0, 0, 1, 1, 1)^T$.

To evaluate whether Multi-SKAT can control type I error under realistic scenarios, we simulated a dataset with 9 phenotypes with a correlation structure identical to that of 9 amino acid phenotypes in the METSIM data (See Supplementary Figure S1). Phenotypes were generated using equation (16) with $\beta = 0$. Total $5,000,000$ datasets with $5,000$ individuals were generated to obtain the empirical type-I error rates at $\alpha = 10^{-4}, 10^{-5}$ and 2.5×10^{-6}, which are corresponding to candidate gene studies to test for 500 and 5000 genes and exome-wide studies to test for all $20,000$ protein coding genes, respectively.

Next, we evaluated type I error controls in the presence of related individuals. To have a realistic related structure, as aforementioned, we used the METSIM study genotype data. We generated a random subsample of 5000 individuals from the METSIM study individuals and generated null values for the 9 phenotypes from $MVN(0, V_e)$, where $V_e = \Phi_{5k} \otimes \widehat{V}_{g;5k} + I \otimes \widehat{V}_{5k}$, Φ_{5k} is the estimated kinship matrix of the 5000 selected individuals, $\widehat{V}_{g;5k}$ and $\widehat{V}_{5k}$ are estimated co-heritability and residual variance matrices respectively for these

samples. For each set of 9 phenotypes, we performed the Multi-SKAT tests for a randomly selected 5000 genes in the METSIM data. For the details about the data, see next section. We carried out this procedure 1000 times and obtained $5,000,000$ p-values, and estimated type I error rates as proportions of p-values smaller than the given level α.

Power simulations were performed both in situations when there was no pleiotropy (i.e., only one of the phenotypes was associated with the causal variants) and also when there was pleiotropy. Under pleiotropy, since it is unlikely that all the phenotypes are associated with genotypes in the region, we varied the number of phenotypes associated. For each associated phenotype, 30% or 50% of the rare variants (MAF $< 1\%$) were randomly selected to be causal variants. We modeled the rarer variants to have stronger effect, as $|\beta_j| = c|log_{10}(MAF_j)|$. We used $c = 0.3$ which yields $|\beta_j| = 0.9$ for variants with MAF$= 10^{-3}$. Our choice of β yielded the average heritability of associated phenotypes between 1% to 4%. We also considered situations that all causal variants were trait-increasing variants (i.e. positive β) or 20 % of causal variants were trait-decreasing variants (i.e. negative β). Empirical power was estimated from 1000 independent datasets at exome-wide $\alpha = 2.5 \times 10^{-6}$.

In type I error and power simulations, we compared the following tests: Bonferroni adjusted minimum p-values from SKAT on each phenotype (minPhen); Multi-SKAT with $\Sigma_{P,Hom}$ (Hom); Multi-SKAT with $\Sigma_{P,Het}$ (Het); Multi-SKAT with $\Sigma_{P,PhC}$ (PhC); Multi-SKAT with $\Sigma_{P,PC-0.9}$ (PC-Sel); Minimum P-value of Hom, Het, PhC and PC-Sel using Copula (minP). For the Multi-SKAT tests, we used two different Σ_G's corresponding to SKAT (i.e. $\Sigma_G = WW$) and Burden tests (i.e. $\Sigma_G = W1_m1_m^TW$). We further evaluated minP_{com}, by combining two Σ_G's (SKAT and Burden) and four Σ_P's (Hom, Het, PhC and PC-Sel), and hence a total of 8 tests. For the variant weighting matrix $W = diag(w_1, \cdots, w_m)$, we used $w_j = Beta(MAF_j, 1, 25)$ function to upweight rarer variants, as recommended by Wu et al.[20].

Analysis of the METSIM study exomechip data

To investigate the cross-phenotype roles of low frequency and rare variants on amino acids, we analyzed data on 8545 participants of the METSIM study on whom all 9 Amino acids (Alanine, Leucine, Isoleucine, Glycine,Valine, Tyrosine, Phenylalanine, Glutamine, Histidine) were measured by proton nuclear magnetic resonance spectroscopy. Samples were genotyped on the Illumina ExomeChip and OmniExpress arrays and we included samples that passed sample QC filters[30]. The kinship between the samples was estimated via KING (version 2.0). We adjusted the amino acid levels (AA) for age, age^2 and BMI and inverse-normalized the residuals. The phenotype correlation matrix after covariate adjustment is shown in Figure 3.

We included rare (MAF $< 1\%$) nonsynonymous and protein-truncating variants with a total rare minor

allele count of at least 5 for genes that had at least 3 rare variants leaving 5207 genes for analysis.

Results

Type I Error simulations

We estimated empirical type I error rates of the Multi-SKAT tests with and without related individuals. For unrelated individuals, we simulated $5,000$ individuals and 9 phenotypes based on the correlation structure for the amino acids phenotypes in the METSIM study data. For related individuals, we simulated $5,000$ individuals using the kinship matrix for randomly chosen METSIM individuals (see the Method section). We performed association tests and type I error rates were estimated by the proportion of p-values less than the specified α levels. Type I error rates of the Multi-SKAT tests were well maintained at $\alpha = 10^{-4}, 10^{-5}$ and 2.5×10^{-6} for both unrelated and related individuals (Table 1). For example, at level $\alpha = 2.5 \times 10^{-6}$, the largest empirical type I error rates from any of the Multi-SKAT tests was 3.4×10^{-6}, which was within the 95% confidence interval (CI = $(1.6 \times 10^{-6}, 4 \times 10^{-6})$).

Power simulations

We compared the empirical power of the minPhen and Multi-SKAT tests. For each simulation setting, we generated $1,000$ datasets with $5,000$ unrelated individuals and for each test estimated empirical power as the proportion of p-values less than $\alpha = 2.5 \times 10^{-6}$, reflecting Bonferroni correction for testing $20,000$ independent genes. Since Hom and Het are identical to hom-MAAUSS and het-MAAUSS, respectively, and using PhC is identical to GAMuT (with linear kernel) and MSKAT, our power simulation studies effectively compare the existing multiple phenotype tests.

In Figure 1, we show the results for 5 phenotypes with compound symmetric correlation structure with the correlation 0.3 or 0.7, where 30% of rare variants (MAF < 0.01) were positively associated with $1, 2$ or 3 phenotypes. In the most scenarios, PhC, PC-Sel and Het have greater power among the Multi-SKAT tests with pre-specified phenotype kernels (i.e. Hom, Het, PhC and PC-Sel) while the Multi-SKAT omnibus test, minP, maintained high power as well. For example, when the correlation between the phenotypes is 0.3 (i.e. $\rho = 0.3$) and SKAT kernel is used for Σ_G, if 3 phenotypes are associated with the region, minP and PhC are more powerful than the other tests. If the correlation between the phenotypes is $\rho = 0.7$ and Burden kernel is used for Σ_G, Het, PC-Sel and minP had higher power than the rest of the tests when 2 phenotypes were associated. It is noteworthy that Hom had the lowest power in almost all the scenarios of Figure 1.

As aforementioned, PhC is equivalent to GAMuT and MSKAT. Figure 2 demonstrates scenarios involving

6 phenotypes and clustered correlation structures where PhC can be outperformed by other choices of Σ_P. For example, Figure 2 upper panel shows a situation when the phenotypes have clustered correlation structure and the correlation within the clusters is low ($\rho = 0.3$); phenotypes $1, 2, 3, 4$ and 6 were associated with the causal variants. In this situation, Hom and minP outperformed PhC when SKAT kernel was used. This may be because the phenotype correlation structure does not reflect the genetic association pattern. Similarly in Figure 2 lower panel the phenotypes consist of 2 small high correlation($\rho = 0.7$) cluster and one larger low correlation ($\rho = 0.3$) cluster as shown in right column; if phenotypes $1, 3$ and 6 are associated with the region, Het and minP had higher power than PhC.

When 20% of causal variants were trait-decreasing variants, the power of Multi-SKAT tests with Burden Σ_G was substantially reduced (Supplementary Figure S2 and S3). This is because the association signals were canceled out due to the presence of negative genetic effect coefficients. Since SKAT is robust regardless of the association direction, power with SKAT Σ_G was largely maintained. The relative performance of methods with different Σ_P given Σ_G was quantitatively similar to the results without trait-decreasing variants.

Further, we estimated power of minP_{com}, which combines both Σ_P and Σ_G. The power of minP_{com} was evaluated for the compound symmetric phenotype correlation structure presented in Figure 1 and has been compared with the two minP tests of SKAT (minP-SKAT) and Burden (minP-Burden) Σ_G kernels. Figure 3 shows empirical power with and without trait-decreasing variants. When all genetic effect coefficients were positive (Figure 3, left panel), and the correlation among phenotypes were low (i.e. $\rho = 0.3$), minP-SKAT was slightly more powerful than minP-Burden. When the correlation was high (i.e. $\rho = 0.7$), minP-Burden was more powerful. When 20% of genetic effect coefficients were negative (Figure 3, right panel), as expected, the power of minP-Burden was substantially decreased. Across all the situations, the power of minP_{com} was either similar or slightly higher than the most powerful minP with fixed Σ_G. When 50% of variants were causal variants and all genetic effect coefficients were positive (Supplementary Figure S4, left panel), minP-Burden was more powerful than minP-SKAT, and minP_{com} had similar or slightly higher power than minP-Burden.

Overall, our simulation results show that the omnibus tests, especially minP_{com}, had robust power throughout all the simulation scenarios considered. When Σ_G and Σ_P were fixed, power depended on the model of association and the correlation structure of the phenotypes. It is also noteworthy that the proposed Multi-SKAT tests generally outperformed the single phenotype test (minPhen), even when only one phenotype was associated with genetic variants.

Application to the METSIM study exomechip data

Inborn errors of amino acid metabolism cause mild to severe symptoms like mental retardation, seizure, coma, and osteoporosis. Amino acid levels are perturbed in other disease states, e.g., Glutamic and aspartic acid levels are reduced in Alzheimer disease brains[31]; Isoleucine, glycine, alanine, phenylalanine, and threonine levels are increased in cerebo-spinal fluid (CSF) of motor neuron disease patients[32]. To find rare variants associated with 9 amino acid levels, we applied the Multi-SKAT tests to the METSIM study data. We estimated the relatedness between individuals by KING[25], and coheritability and residual variance using PHENIX[27] (Supplementary Figure S1). Among the 8,545 METSIM participants with non-missing phenotypes and covariates, 1332 individuals were found to be as closely related as degree 2 or closer. A total of 5207 genes with at least three rare variants were included in our analysis. The Bonferroni corrected significance threshold was $\alpha = 0.05/5207 = 9.6 \times 10^{-6}$. Further we used a less significant cutoff of $\alpha = 10^{-4}$ for a gene to be suggestive. To investigate the validity of our results, we applied SKAT-O to test for single phenotype associations. After identifying associated genes, we carried out backward elimination procedure (Appendix B) to investigate which phenotypes are associated with the gene. This procedure iteratively removes phenotypes based on minP_{com} p-values.

QQ plots for the p-values obtained by minPhen and Multi-SKAT omnibus tests are displayed in Figure 4. Due to the presence of several strong associations, for the ease of viewing, any p-value $< 10^{-12}$ was collapsed to 10^{-12}. Table 2 shows significant and suggestive genes by either minPhen or minP_{com}, and Supplementary Table S3 shows SKAT-O p-values between each of these genes and each phenotype. Among the eight significant or suggestive genes displayed in Table 2, minP_{com} provides more significant p-values than minPhen for six genes: Glycine decarboxylase (*GLDC* [MIM: 238300]), Histidine ammonia-lyase (*HAL* [MIM: 609457]), Phenylalanine hydroxylase (*PAH* [MIM: 612349]), Dihydroorotate dehydrogenase (*DHODH* [MIM: 126064]), Mediator of RNA polymerase II transcription subunit 1 (*MED1* [MIM: 604311]), Serine/Threonine Kinase 33 (*STK33* [MIM: 607670]). Interestingly, *PAH* and *MED1* are significant by minP_{com}, but not significant by minPhen. *PAH* encodes an Phenylalanine hydroxylase, which catalyzes the hydroxylation of the aromatic side-chain of phenylalanine to generate tyrosine. Backward elimination shows that Phenylalanine and Tyrosine are two last phenotypes to be eliminated (Supplementary Table S2). *MED1* is involved in the regulated transcription of nearly all RNA polymerase II-dependent genes. This gene does not show any single phenotype association, but the cross-phenotype analysis produced the evidence of association.

Among other genes, *GLDC* has the smallest p-value. Variants in *GLDC* are known to cause glycine encephalopathy (MIM: 605899)[33]. Univariate SKAT-O test with each of these phenotype reveals that this gene has a strong association with Glycine (p-value $= 2.5 \times 10^{-64}$, Supplementary Table S3). Among the

variants genotyped in this gene, rs138640017 (MAF = 0.009) appears to drive the association (single variant p-value = 1.0×10^{-64}). Variants in *HAL* cause histidinemia (MIM: 235800) in human and mouse. This gene manifests significant univariate association with Histidine (SKAT-O p-value = 3.2×10^{-8}, Supplementary Table S3) which in turn is influenced by the association of rs141635447 (MAF = 0.005) with Histidine (single variant p-value = 3.7×10^{-13}). Similarly, variants in *DHODH*, which have been previously found to be associated with postaxial acrofacial dysostosis (MIM: 263750), have significant cross-phenotype association though mostly driven by the association with Alanine (SKAT-O p-value $=1.4 \times 10^{-07}$, Supplementary Table S3). *ALDH1L1* catalyzes conversion of 10-formyltetrahydrofolate to tetrahydrofolate. Published results show that common variant rs1107366, 5kb upstream of *ALDH1L1*, is associated with Glycine-Seratinine ratio[34]. Down-regulation of *BCAT2* in mice causes elevated serum branched chain amino acid levels and features of maple syrup urine disease.

Table 3 shows p-values of Multi-SKAT kernel and omnibus tests with two genotype kernels (SKAT and Burden). Among phenotype kernels, PhC and Het generally produced the smallest p-values. We further applied Multi-SKAT tests without kinship adjustment on the whole METSIM study sample. As expected, this produced inflation in QQ plots (Supplementary Figure S5).

Overall, our METSIM amino acid data analysis suggests that the proposed method can be more powerful than the single phenotype tests, while maintaining type I error rates even in the presence of the sample relatedness. It also shows that the omnibus test provides robust performance by effectively aggregating results of various kernels.

Computation Time

When Σ_P and Σ_G are given, p-values of Multi-SKAT are computed by the Davies method[22], which inverts the characteristic function of the mixture of chi-squares. On average, Multi-SKAT tests given Σ_P and Σ_G required less than 1 CPU sec (Intel Xeon 2.80 GHz) when applied to a dataset with 5000 independent samples, 50 variants and 5 phenotypes (Supplementary Table S1). With the kinship adjustment for 5000 related individuals, computation time was increased to 5 CPU sec. Since minP_{com} requires only a small number of resampling to estimate the correlation among tests, it is still scalable for genome-wide analysis. In the same dataset, minP_{com} required 16 and 19 CPU sec on average without and with the kinship adjustment, respectively. Analyzing the METSIM dataset required 10 hours when parallelized into 5 processes.

Discussion

In this article, we have introduced a new class of methods, Multi-SKAT, for rare variant test for multiple phenotypes. As demonstrated, Multi-SKAT gains flexibility in terms of modeling the complex interaction between the phenotypes and the genotypes through the use of the kernels Σ_P and Σ_G. Many published methods of gene-based analysis of multiple phenotypes, including GAMuT, MSKAT and MAAUSS, can be viewed as special cases of the Multi-SKAT test, corresponding to different values of Σ_P and Σ_G. Thus, Multi-SKAT, in essence, provides a general framework for region based tests for cross-phenotype effects.

It is natural to assume that different genes follow different models of association. For some genes, the effect of the variants on the phenotypes might be independent of each other, resembling the Het kernel for Σ_P, while for others, the effects might be nearly the same, which is similar to the Hom kernel. The omnibus test, which combines various choices of kernels, can be a useful method in this regard. The minimum p-value approaches using copula (minP and minP_{com}) combines the tests from a p-value perspective and provide robust power regardless of underlying genetic models.

Multi-SKAT retains most of the desirable properties of SKAT. The asymptotic p-values of all the Multi-SKAT tests, other than minP and minP_{com}, can be analytically obtained via Davies' method. The p-value calculations for minP and minP_{com} depend on a resampling based approach but a reliable estimate can be obtained from a small number of resampling. Thus, computationally all the Multi-SKAT tests are scalable at the genome-wide level. This method also allows the inclusion of prior information through weighing of variants.

Additionally, Multi-SKAT can adjust for the relatedness among study individuals by accounting for the kinship matrix. As shown in Supplementary Figure S5, in the presence of related individuals, not adjusting for the relatedness can produce inflated type I error rates. Since Multi-SKAT is a regression based approach, it effectively incorporates the relatedness by including a random effect term due to kinship. Type I error simulation and METSIM data analysis show that our approach produced more accurate p-values than alternative methods while controlling type I error rates.

Although it provides a general framework for gene-based multiple phenotype tests, the current approach is limited to continuous phenotypes. In the future, using a generalized mixed effect model framework, we aim to extend Multi-SKAT to binary phenotypes.

In summary, we have developed a powerful multiple phenotype test for rare variants. The proposed method has robust power regardless of the underlying biology and can adjust for family relatedness. Our method can be a scalable and practical solution to test for multiple phenotypes and will contribute to detecting rare variants with pleiotropic effects. All our methods are implemented in the R package Multi-

SKAT.

Supplemental Data

Five figures and three tables that were referred to in this paper are provided in the Supplemental tables and figures.

Acknowledgments

This work was supported by grants R01 HG008773 (D.D. and S.L.) from the National Institutes of Health. We would like to thank investigators of the METSIM study for access to the genotype and amino-acid phenotype data.

Web Resources

MultiSKAT R-package: https://github.com/diptavo/MultiSKAT

Online Mendelian Inheritance in Man (OMIM): http://www.omim.org

Appendices

A. Relationship between Multi-SKAT and existing methods

A.1. MSKAT

The MSKAT test statistic is

$$Q_{MSKAT} = vec(S)^T (WW \otimes \widehat{V}^{-1}) vec(S), \tag{17}$$

where $S = G^T(Y - \widehat{\mu})$ is a matrix of score statistics[15]. From the fact

$$(I \otimes \widehat{V}^{-1}) vec(S) = (G^T \otimes I)(I \otimes \widehat{V}^{-1}) \left\{ vec(Y) - vec(\widehat{\mu}) \right\}$$

the Q_{MSKAT} can be written as

$$\left\{ vec(Y) - vec(\widehat{\mu}) \right\}^T \left\{ (GWWG^T) \otimes \widehat{V}^{-1} \right\} \left\{ vec(Y) - vec(\widehat{\mu}) \right\},$$

which is a Multi-SKAT test statistics with $\Sigma_G = WW$ and $\Sigma_P = \widehat{V}$.

A.2. GAMuT

Suppose $Y_{adj} = HY$ and $G_{adj} = HG$ are covariate adjusted phenotype and genotype matrices where $H = I - X(X^TX)^{-1}X^T$. We note that with the intercept in X, Y_{adj} and G_{adj} are mean centered. The covariate adjusted GAMuT test statistics is

$$Q_{GAMuT} = \frac{tr(P_cX_c)}{n}$$

where $P_c = Y_{adj}(Y_{adj}^TY_{adj})^{-1}Y_{adj}^T$ and $X_c = G_{adj}WWG_{adj}^T$. Using the fact that $Y_{adj}^TY_{adj}/n = \widehat{V}$ is the estimate of variance after adjusting covariates and $G_{adj}^TY_{adj} = GHY = G^TY_{adj}$, we show

$$\begin{aligned}
tr(P_cX_c)/n &= tr(Y_{adj}\widehat{V}^{-1}Y_{adj}^TG_{adj}WWG_{adj}^T) \\
&= tr(\widehat{V}^{-\frac{1}{2}}Y_{adj}^TGWWG^TY_{adj}\widehat{V}^{-\frac{1}{2}}) \\
&= vec(WG^TY_{adj}\widehat{V}^{-\frac{1}{2}})^Tvec(WG^TY_{adj}\widehat{V}^{-\frac{1}{2}}) \\
&= ((WG^T \otimes \widehat{V}^{-\frac{1}{2}})vec(Y_{adj}))^T((WG^T \otimes \widehat{V}^{-\frac{1}{2}})vec(Y_{adj})) \\
&= (vec(Y) - vec(\widehat{\mu}))^T(GWWG^T \otimes \widehat{V}^{-1})(vec(Y) - vec(\widehat{\mu}))
\end{aligned}$$

which the Multi-SKAT test statistic with $\Sigma_G = WW$ and $\Sigma_P = \widehat{V}$.

A.3. MAAUSS and MF-KM

There exists two different version of the MAAUSS tests. The homogeneous version of MAAUSS assumes that the effects of a variant on multiple phenotypes are identical and uses the following test statistics

$$Q_{MAAUSS-HOM} = (vec(Y) - vec(\widehat{\mu}))^T(I_n \otimes \widehat{V}^{-1})(G \otimes I)(WW \otimes 1_m1_m^T)(G^T \otimes I)(I_n \otimes \widehat{V}^{-1})(vec(Y) - vec(\widehat{\mu})) \tag{18}$$

which is identical to the Multi-SKAT test statistic with $\Sigma_G = WW$ and $\Sigma_P = 1_m1_m^T$. The heterogeneous version of MAAUSS assumes that the effects of a variant on multiple phenotypes are independent, and use the following test statistics

$$Q_{MAAUSS-HET} = (vec(Y) - vec(\widehat{\mu}))^T(I_n \otimes \widehat{V}^{-1})(G \otimes I)(WW \otimes I)(G^T \otimes I)(I_n \otimes \widehat{V}^{-1})(vec(Y) - vec(\widehat{\mu})) \tag{19}$$

which is identical to the Multi-SKAT test statistic with $\Sigma_G = WW$ and $\Sigma_P = I$. Note that the test statistic of MF-KM is exactly the same as $Q_{MAAUSS-HET}$.

B. Backward elimination procedure to identify associated phenotypes

After identifying the gene or region associated with multiple phenotypes, next question would be identifying truly associated phenotypes. Here we present a simple backward elimination algorithm to iteratively remove relatively less important phenotypes. Similar method has previously been applied to identify rare causal variants in an associated gene[35].

- Step 1. Start with a set of k phenotypes $Phen_{Current} = \{y_1, y_2, \cdots y_k\}$ and compute a Multi-SKAT test association p-value for the set $Phen_{Current}$ denoted by $p_{Current}$.
- Step 2. Remove each of the phenotypes one at a time from the set $Phen_{Current}$. The resulting set is $Phen_{-i} = \{y_1, y_2, \cdots y_{i-1}, y_{i+1} \cdots y_k\}$ for $i = 1, 2, \cdots, k$ and compute the corresponding p-values p_{-i} for that same Multi-SKAT test.
- Step 3. Remove the phenotype j that leads to the smallest p-value, i.e. $j = argmin\{p_{-1}, p_{-2}, \cdots, p_{-k}\}$. Update $Phen_{Current}$ to $Phen_{-j}$.
- Step 4. Continue removing phenotypes till only 1 phenotype is left.

Supplementary Table S2 shows the backward elimination results of 5 most significant and suggestive genes in the METSIM study data analysis as per the p-values reported by minP_{com}. Although this procedure does not provide a set of phenotypes truly associated, it provides the relative importance of the phenotypes in driving association signals. For example, the minP_{com} p-value for *GLDC* was 2.3×10^{-72}. When each of the phenotypes were removed one at a time and the minP_{com} p-values were calculated on the remaining 8 phenotypes, we found that eliminating Isoleucine (Ile) actually improved the signal. The minP_{com} p-value of the set of 8 amino acids leaving out Ile was 2.8×10^{-73}. This indicates that Isoleucine has very minimal contribution to the association between the amino acids and *GLDC*. Subsequently, Valine was the next phenotype to be eliminated indicating that it has the next lowest contribution after Isoleucine. Carrying out this procedure further, we find that Glycine is the last phenotype to remain indicating that it is the strongest driver of the signal. This is in agreement to the single phenotype SKAT-O results (Supplementary Table S3). Similary for genes *HAL*, *DHODH*, *PAH* and *MED1*, Histine, Alanine, Phenylalanine and Tyrosine were the most associated phenotypes, respectively. Interestingly for *PAH* and *MED1*, single phenotype p-values are not significant, which suggests that multiple phenotypes are associated with these genes.

References

[1] Arthur Korte and Ashley Farlow. The advantages and limitations of trait analysis with gwas: a review. *Plant Methods*, 9:9–29, 2013.

[2] Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, Andrew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F. C. Mackay, Steven A. McCarroll, , and Peter M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461:747–753, 2009.

[3] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65, 2012.

[4] Seunggeung Lee, Gonçalo R Abecasis, Michael Boehnke, and Xihong Lin. Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, 95(1):5–23, 2014.

[5] Nadia Solovieff, Chris Cotsapas, Phil H Lee, Shaun M Purcell, and Jordan W Smoller. Pleiotropy in complex traits: challenges and strategies. *Nature reviews. Genetics*, 14(7):483, 2013.

[6] Chris Cotsapas, Benjamin F Voight, Elizabeth Rossin, Kasper Lage, Benjamin M Neale, Chris Wallace, Gonçalo R Abecasis, Jeffrey C Barrett, Timothy Behrens, Judy Cho, et al. Pervasive sharing of genetic effects in autoimmune disease. *PLoS genetics*, 7(8):e1002254, 2011.

[7] Shanya Sivakumaran, Felix Agakov, Evropi Theodoratou, James G Prendergast, Lina Zgaga, Teri Manolio, Igor Rudan, Paul McKeigue, James F Wilson, and Harry Campbell. Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics*, 89(5):607–618, 2011.

[8] M.A. Ferreira and S.M. Purcell. A multivariate test of association. *Bioinformatics*, 25:132–133, 2009.

[9] J. Huang, A.D. Johnson, and C.J. O'Donnell. Prime: a method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies. *Bioinformatics*, 27:1201–1206, 2011.

[10] Xiang Zhou and Matthew Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, 11:407–409, 2014.

[11] J. S. Ried, A. Doring, K. Oexle, C. Meisinger, J. Winkelmann, N. Klopp, T. Meitinger, A. Peters, K. Suhre, H.E. Wichmann, and C. Gieger. Psea: Phenotype set enrichment analysis–a new method for analysis of multiple phenotypes. *Genetic Epidemiology*, 36:244–252, 2012.

[12] Debashree Ray, James S Pankow, and Saonli Basu. Usat: A unified score-based association test for multiple phenotype-genotype analysis. *Genetic epidemiology*, 40(1):20–34, 2016.

[13] Y. Wang, A. Liu, J.L. Mills, M. Boehnke, A.F. Wilson, J.E. Bailey-Wilson, M. Xiong, C.O. Wu, and R. Fan. Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. *Genetic Epidemiology*, 39:259–275, 2015.

[14] K. Alaine Broadaway, David J. Cutler, Richard Duncan, Jacob L. Moore, Erin B. Ware, Lawrence F. Bielak Min A. Jhun, Wei Zhao, Jennifer A. Smith, Patricia A. Peyser, Sharon L.R. Kardia, Debashis Ghosh, and Michael P. Epstein. A statistical approach for testing cross-phenotype effects of rare variants. *American Journal of Human Genetics*, 98(3):525–540, 2016.

[15] Baolin Wu and J.S. Pankow. Sequence kernel association test of multiple continuous phenotypes. *Genetic Epidemiology*, 40(2):91–100, 2016.

[16] Selyeong Lee, Sungho Won, Young Jin Kim, Yongkang Kim, Bong-Jo Kim, and Taesung Park. Rare variant association test with multiple phenotypes. *Genetic Epidemiology*, 2016.

[17] Jianping Sun, Karim Oualkacha, Vincenzo Forgetta, Hou-Feng Zheng, J Brent Richards, Antonio Ciampi, and Celia MT Greenwood. A method for analyzing multiple continuous phenotypes in rare variant association studies allowing for flexible correlations in variant effects. *European Journal of Human Genetics*, 24(9):1344–1351, 2016.

[18] A. Maity, P.F. Sullivan, and J.Y. Tzeng. Multivariate phenotype association analysis by marker-set kernel machine regression. *Genetic Epidemiology*, 36:686–695, 2012.

[19] Qi Yan, Daniel E Weeks, Juan C Celedón, Hemant K Tiwari, Bingshan Li, Xiaojing Wang, Wan-Yu Lin, Xiang-Yang Lou, Guimin Gao, Wei Chen, et al. Associating multivariate quantitative phenotypes with genetic variants in family samples with a novel kernel machine regression method. *Genetics*, 201(4):1329–1339, 2015.

[20] Michael C. Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *Americal Journal of Human Genetics*, 89:82–93, 2011.

[21] Seunggeun Lee, Michael C Wu, and Xihong Lin. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775, 2012.

[22] Robert B Davies. The distribution of a linear combination of x2 random variables. *Applied Statistics*, 29(3):323–333, 1980.

[23] Hugues Aschard, Bjarni J Vilhjálmsson, Nicolas Greliche, Pierre-Emmanuel Morange, David-Alexandre Trégouët, and Peter Kraft. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *The American Journal of Human Genetics*, 94(5):662–676, 2014.

[24] Stefano Demarta and Alexander J McNeil. The t copula and related copulas. *International Statistical Review/Revue Internationale de Statistique*, pages 111–129, 2005.

[25] A Manichaikul, J C Mychaleckyj, S S Rich, K Daly, M Sale, and W M Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010.

[26] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47:291–295, 2015.

[27] Andrew Dahl, Valentina Iotchkova, Amelie Baud, Asa Johansson, Ulf Gyllensten, Nicole Soranzo, Richard Mott, Andreas Kranis, and Jonathan Marchini. A multiple-phenotype imputation method for genetic studies. *Nature Genetics*, 48:466–472, 2016.

[28] Xiang Zhou, Peter Carbonetto, and Stephens Matthew. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genetics*, 9(2):e1003264, 2013.

[29] Stephen F Schaffner, Catherine Foo, Stacey Gabriel, David Reich, Mark J Daly, and David Altshuler. Calibrating a coalescent simulation of human genome sequence variation. *Genome research*, 15(11):1576–1583, 2005.

[30] Jeroen R Huyghe, Anne U Jackson, Marie P Fogarty, Martin L Buchkivoch, Alena Stancakova, Heather M Stringham, Xueling Sim, Lingyao Yang, Christian Fuchsberger, Henna Cederberg, and et al. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nature Genetics*, 45:197–201, 2013.

[31] D. Allan Butterfield and Chava B. Pocernich. The glutamatergic system and alzheimer's disease. *CNS Drugs*, 17(9):641–652, 2003.

[32] J. de Belleroche, A. Recordati, and F. Clifford Rose. Elevated levels of amino acids in the csf of motor neuron disease patients. *CNS Drugs*, 17(9):641–652, 2003.

[33] Ian Hughes. Pediatric endocrinology and inborn errors of metabolism. *Journal of Paediatrics and Child Health*, 45(11):668–669, 2009.

[34] Weijia Xie, Andrew R. Wood, Valeria Lyssenko, Michael N. Weedon, Joshua W. Knowles, Sami Alkayyali, Themistocles L. Assimes, Thomas Qaurtermous, Fahim Abbasi, Jussi Paananen, and et al. Genetic variants associated with glycine metabolism and their role in insulin sensitivity and type 2 diabetes. *Diabetes*, 62:2141–2150, 2013.

[35] Iuliana Ionita-Laza, Marinela Capanu, Silvia De Rubeis, Kenneth McCallum, and Joseph D Buxbaum. Identification of rare causal variants in sequence-based studies: methods and applications to vps13b, a gene involved in cohen syndrome and autism. *PLoS Genet*, 10(12):e1004729, 2014.

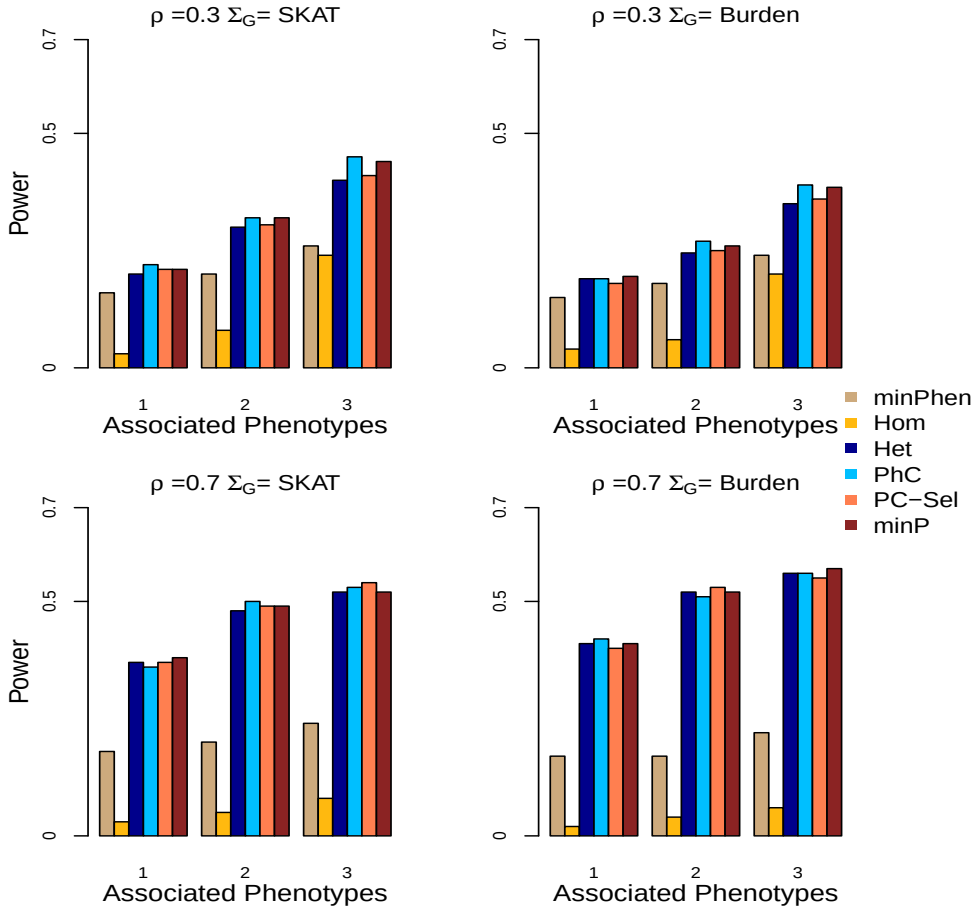

Figure 1: Power for Multi-SKAT tests when phenotypes have compound symmetric correlation structures. Empirical power for minPhen, Hom, Het, PhC, PC-Sel, minP plotted against the number of phenotypes associated with the gene of interest with a total of 5 phenotypes under consideration. Upper row shows the results for $\rho = 0.3$ and lower row for $\rho = 0.7$. Left column shows results with SKAT kernel Σ_G, and right columns shows results with Burden kernel. All the causal variants were trait-increasing variants.

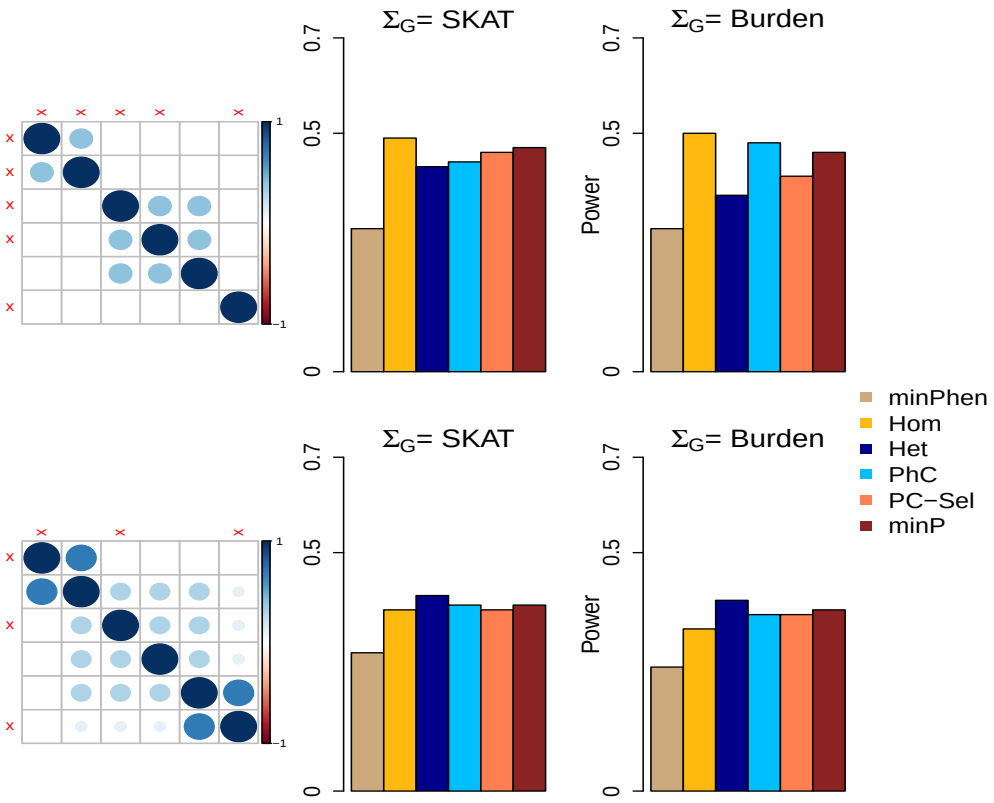

Figure 2: Power for Multi-SKAT tests when phenotypes have clustered correlation structures. Empirical powers for minPhen, Hom, Het, PhC, PC-Sel, minP are plotted under different levels of association with a total of 6 phenotypes and with clustered correlation structures. Middle column shows the empirical powers for different combinations of phenotypes associated with SKAT kernel Σ_G; the rightmost column shows the corresponding results with Burden kernel; left column shows the corresponding correlation matrices for the phenotypes. The associated phenotypes are indicated in red cross marks across the correlation matrices. All the causal variants were trait-increasing variants.

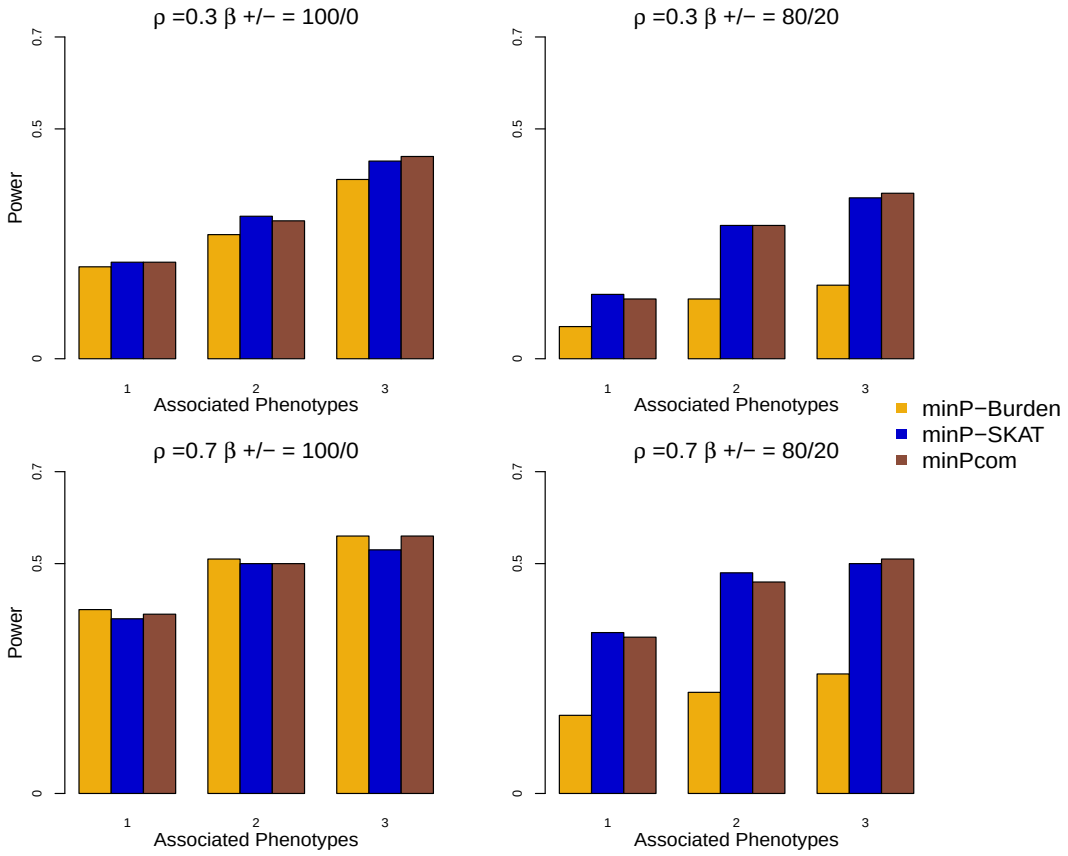

Figure 3: Power for Multi-SKAT by combining tests with Σ_P as Hom, Het, PhC, PC-Sel and Σ_G as SKAT and Burden when phenotypes have compound symmetric correlation structures. Empirical powers for minP-Burden, minP-SKAT and minP_{com} are plotted against the number of phenotypes associated with the gene of interest with a total of 5 phenotypes under consideration. Upper row shows the results for $\rho = 0.3$ and lower row for $\rho = 0.7$. Left column shows results when all the causal variants were trait-increasing variants, and right column shows results when 80%/20% of the causal variants were trait-increasing/trait-decreasing variants.

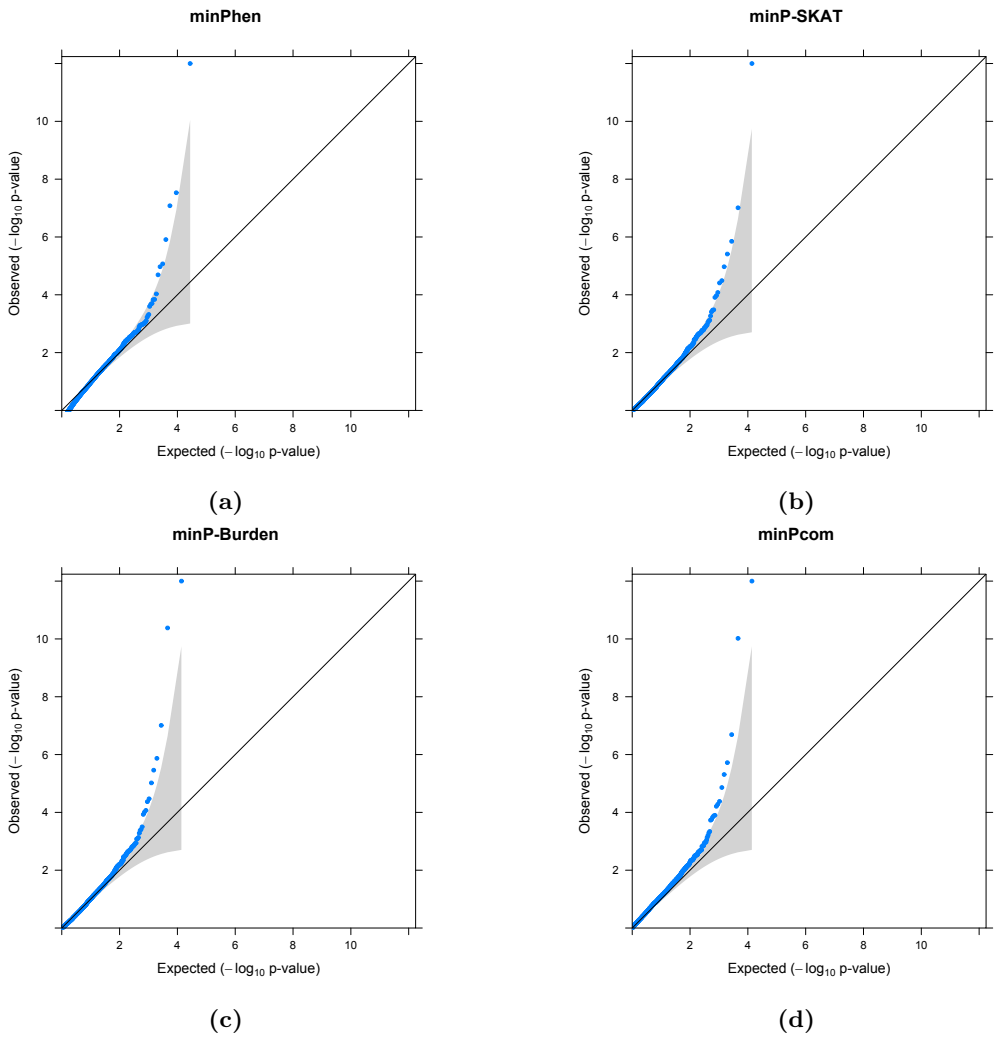

Figure 4: QQplot of the p-values of minPhen and Multi-SKAT omnibus tests for the METSIM data. For the ease of viewing, any associations with p-values $< 10^{-12}$ have been collapsed to 10^{-12}

Table 1: Empirical type I error rates of the Multi-SKAT tests. The number of phenotypes were nine and the correlation structure among the phenotypes were similar to that of the amino acid phenotypes in the METSIM study data. Type I error rates were estimated based on 5,000,000 simulated null datasets.

	Level	Σ_G	Hom	Het	PhC	PC-Sel	minP	minP_{com}
Independent samples (without kinship adjustment)	2.5×10^{-06}	*SKAT*	2.6×10^{-06}	2.6×10^{-06}	2.8×10^{-06}	2.6×10^{-06}	3.4×10^{-06}	2.8×10^{-06}
		Burden	2.8×10^{-06}	2.6×10^{-06}	2.6×10^{-06}	2.6×10^{-06}	3.0×10^{-06}	
	10^{-05}	*SKAT*	9.6×10^{-06}	9.8×10^{-06}	9.8×10^{-06}	9.8×10^{-06}	9.4×10^{-06}	1.2×10^{-05}
		Burden	9.6×10^{-06}	9.8×10^{-06}	9.6×10^{-06}	9.6×10^{-06}	9.6×10^{-06}	
	10^{-04}	*SKAT*	9.5×10^{-05}	9.6×10^{-05}	9.8×10^{-05}	9.8×10^{-05}	9.8×10^{-05}	1.1×10^{-04}
		Burden	9.6×10^{-05}	9.4×10^{-05}	9.7×10^{-05}	9.6×10^{-05}	9.7×10^{-05}	
Related samples (with kinship adjustment)	2.5×10^{-06}	*SKAT*	2.4×10^{-06}	2.4×10^{-06}	2.8×10^{-06}	2.6×10^{-06}	3.0×10^{-06}	2.6×10^{-06}
		Burden	2.6×10^{-06}	2.4×10^{-06}	2.6×10^{-06}	2.4×10^{-06}	3.2×10^{-65}	
	10^{-05}	*SKAT*	9.8×10^{-06}	9.8×10^{-06}	9.8×10^{-06}	9.8×10^{-06}	9.4×10^{-06}	1.4×10^{-05}
		Burden	9.6×10^{-06}	9.4×10^{-06}	9.4×10^{-06}	9.4×10^{-06}	9.4×10^{-65}	
	10^{-04}	*SKAT*	9.6×10^{-05}	9.7×10^{-05}	9.8×10^{-05}	9.4×10^{-05}	9.8×10^{-05}	1.2×10^{-04}
		Burden	9.7×10^{-05}	9.6×10^{-05}	9.8×10^{-05}	9.8×10^{-05}	9.7×10^{-05}	

Table 2: Significant and suggestive genes associated with 9 amino acid phenotypes. Genes with p-values $< 10^{-4}$ by minPhen or any Multi-SKAT tests (minP-SKAT, minP-Burden and minP_{com}) were reported in this table. Multi-SKAT tests were applied with the kinship adjustment, and 5207 genes with at least three rare (MAF < 0.01) nonsynonymous and protein-truncating variants were used in this analysis. The total sample size was n=8545. P-values smaller than the Bonferroni corrected significance $\alpha = 9.6 \times 10^{-06}$ were marked as bold. Smallest p-value for each gene among all the tests have been underlined.

Gene	Chromosome	Rare SNPs	MAC	minPhen	minP-SKAT	minP-Burden	minP_{com}
GLDC	9	4	183	$\mathbf{2.2 \times 10^{-63}}$	$\underline{\mathbf{1.1 \times 10^{-72}}}$	$\mathbf{9.7 \times 10^{-64}}$	$\mathbf{2.3 \times 10^{-72}}$
HAL	3	6	42	$\mathbf{2.9 \times 10^{-08}}$	1.1×10^{-05}	$\underline{\mathbf{4.2 \times 10^{-11}}}$	$\mathbf{9.5 \times 10^{-11}}$
DHODH	16	5	90	$\mathbf{1.3 \times 10^{-06}}$	$\underline{\mathbf{9.7 \times 10^{-08}}}$	$\mathbf{7.7 \times 10^{-06}}$	$\mathbf{1.1 \times 10^{-07}}$
PAH	12	6	27	6.1×10^{-04}	$\mathbf{1.4 \times 10^{-06}}$	$\underline{\mathbf{1.3 \times 10^{-06}}}$	$\mathbf{1.9 \times 10^{-06}}$
MED1	17	3	147	6.2×10^{-01}	$\underline{\mathbf{3.9 \times 10^{-06}}}$	3.4×10^{-04}	$\mathbf{4.9 \times 10^{-06}}$
STK33	11	6	180	8.9×10^{-01}	$\underline{3.2 \times 10^{-05}}$	3.3×10^{-02}	4.2×10^{-05}
ALDH1L1	3	8	103	$\underline{\mathbf{8.4 \times 10^{-08}}}$	3.8×10^{-04}	4.2×10^{-05}	5.4×10^{-05}
BCAT2	19	3	133	$\underline{1.1 \times 10^{-05}}$	9.3×10^{-03}	3.4×10^{-05}	6.2×10^{-05}

Table 3: P-values for MultiSKAT tests (Hom, Het, PhC,PC-Sel, minP) with SKAT and burden kernels for the genes reported in Table 2. Multi-SKAT tests were applied with the kinship adjustment. Only genes with at least three rare (MAF < 0.01) nonsynonymous and protein-truncating variants were used in this analysis. The total sample size was n=8545. P-values smaller than the Bonferroni corrected significance $\alpha = 9.6 \times 10^{-06}$ were marked as bold.

	Gene					Multi-SKAT				
Σ_G	Name	Chromosome	Rare SNPs	MAC	minPhen	Hom	Het	PhC	PC-Sel	minP
	GLDC	9	4	183	$\mathbf{6.8 \times 10^{-63}}$	1.3×10^{-03}	$\mathbf{1.3 \times 10^{-17}}$	$\mathbf{2.9 \times 10^{-73}}$	$\mathbf{5.3 \times 10^{-59}}$	$\mathbf{1.1 \times 10^{-72}}$
	DHODH	16	5	90	$\mathbf{5.1 \times 10^{-08}}$	1.4×10^{-01}	3.2×10^{-03}	$\mathbf{3.1 \times 10^{-08}}$	$\mathbf{7.1 \times 10^{-06}}$	$\mathbf{9.7 \times 10^{-08}}$
	PAH	12	6	27	4.3×10^{-05}	7.9×10^{-02}	1.8×10^{-05}	$\mathbf{3.9 \times 10^{-07}}$	6.3×10^{-04}	$\mathbf{1.4 \times 10^{-06}}$
SKAT	*MED1*	17	3	147	8.7×10^{-02}	3.4×10^{-01}	$\mathbf{9.3 \times 10^{-07}}$	2.8×10^{-04}	1.9×10^{-04}	$\mathbf{3.9 \times 10^{-06}}$
	HAL	12	6	42	3.8×10^{-05}	5.5×10^{-01}	3.5×10^{-03}	$\mathbf{2.9 \times 10^{-06}}$	1.5×10^{-04}	1.1×10^{-05}
	STK33	11	6	180	4.0×10^{-01}	2.3×10^{-01}	$\mathbf{8.9 \times 10^{-06}}$	1.1×10^{-03}	5.7×10^{-03}	3.2×10^{-05}
	ALDH1L1	3	8	103	4.4×10^{-02}	1.1×10^{-01}	2.2×10^{-02}	1.6×10^{-04}	4.2×10^{-04}	3.8×10^{-04}
	BCAT2	19	3	133	9.7×10^{-04}	5.0×10^{-01}	8.1×10^{-02}	2.4×10^{-03}	3.5×10^{-03}	9.3×10^{-03}
	GLDC	9	4	183	$\mathbf{1.4 \times 10^{-59}}$	7.5×10^{-04}	$\mathbf{1.5 \times 10^{-15}}$	$\mathbf{1.3 \times 10^{-64}}$	$\mathbf{2.1 \times 10^{-54}}$	$\mathbf{9.7 \times 10^{-64}}$
	HAL	12	6	42	$\mathbf{2.4 \times 10^{-08}}$	2.8×10^{-05}	4.6×10^{-01}	$\mathbf{4.2 \times 10^{-12}}$	7.9×10^{-03}	$\mathbf{4.2 \times 10^{-11}}$
	PAH	12	6	27	4.3×10^{-05}	2.0×10^{-01}	5.3×10^{-04}	$\mathbf{7.2 \times 10^{-06}}$	9.3×10^{-03}	$\mathbf{1.3 \times 10^{-06}}$
Burden	*DHODH*	16	5	90	$\mathbf{1.3 \times 10^{-07}}$	5.3×10^{-02}	1.3×10^{-02}	$\mathbf{2.9 \times 10^{-06}}$	7.6×10^{-04}	$\mathbf{7.7 \times 10^{-06}}$
	BCAT2	19	3	133	$\mathbf{8.4 \times 10^{-06}}$	4.0×10^{-01}	1.7×10^{-02}	1.1×10^{-05}	6.1×10^{-03}	3.7×10^{-05}
	ALDH1L1	3	8	103	$\mathbf{8.3 \times 10^{-08}}$	1.7×10^{-02}	5.1×10^{-03}	$\mathbf{2.1 \times 10^{-06}}$	1.7×10^{-05}	4.2×10^{-05}
	MED1	17	3	147	9.4×10^{-01}	1.8×10^{-01}	8.7×10^{-05}	2.1×10^{-03}	7.1×10^{-01}	3.4×10^{-04}
	STK33	11	6	180	4.8×10^{-01}	7.1×10^{-01}	7.0×10^{-04}	3.8×10^{-02}	6.4×10^{-01}	6.3×10^{-03}