Optimal tests for rare variant effects in sequencing association studies
Supplementary Material

Seunggeun Lee
Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA
Michael C. Wu
Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA
Xihong Lin*
Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA
xlin@hsph.harvard.edu

April 5, 2012

A Sample Size/ Power calculation Formula

A.1 Continuous Traits

For simplicity, we assume no covariates are present, but we note that the results presented can be easily extended to accommodate covariates. We again suppose that there are n individuals and $\mathbf{y} = (y_1, \ldots, y_n)'$ is a vector of continuous phenotype values. To relate the variants to the phenotype, we consider the linear model

$$y_i = \alpha_0 + \mathbf{G}_i'\boldsymbol{\beta} + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$. Without loss of generality and for ease of presentation, we set each entry of $\mathbf{G}_i$ to be centered such that $E(\mathbf{G}_i) = 0$, and $\sigma = 1$ for continuous traits. The SKAT test statistic with a kernel $K(\cdot, \cdot)$ is then given by

$$Q = (\mathbf{y} - \bar{y}\mathbf{1})'\mathbf{K}(\mathbf{y} - \bar{y}\mathbf{1}), \quad \text{where} \quad \bar{y} = n^{-1}\sum_{i=1}^{n} y_i.$$

In the case of the new family of kernels with given ρ, $\mathbf{K} = \mathbf{K}_\rho = \mathbf{G}\mathbf{W}\mathbf{R}_\rho\mathbf{W}\mathbf{G}'$. Setting $\boldsymbol{\mu}_\beta = \mathbf{G}\boldsymbol{\beta}$ and $\mathbf{E} = \mathbf{y} - \bar{y}\mathbf{1} - \boldsymbol{\mu}_\beta$, then Q can be rewritten as

$$Q = (\mathbf{y} - \bar{y}\mathbf{1})'\mathbf{K}_\rho(\mathbf{y} - \bar{y}\mathbf{1}) = (\mathbf{E} + \boldsymbol{\mu}_\beta)'\mathbf{K}_\rho(\mathbf{E} + \boldsymbol{\mu}_\beta).$$

Note that by the spectral decomposition, $\mathbf{K}_\rho = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'$. Since each element of $\mathbf{E}$ is an independent Gaussian with mean 0 and asymptotic variance 1, Q asymptotically follows $\sum_{j=1}^{p}\lambda_j\chi_1^2(\delta_j)$ with $\delta_j = \boldsymbol{\mu}_\beta'\mathbf{u}_j\mathbf{u}_j'\boldsymbol{\mu}_\beta$. Here, λ_j is the j^{th} diagonal element of $\boldsymbol{\Lambda}$, and $\mathbf{u}_j$ is the j^{th} column of $\mathbf{U}$.

For computational efficiency, we approximate the mixture of chi-square distributions of Q using the non-central chi-square approximation with ν degrees of freedom and non-centrality parameter δ [1] under the null and alternative. Specifically, we compute $c_k = \sum_{j=1}^{p} \lambda_j^k$ for the null distribution, and $c_k = \sum_{j=1}^{p} \lambda_j^k + k \sum_{j=1}^{p} \lambda_j^k \delta_j$ for the alternative distribution up to $k = 4$. These values can be obtained from

$$\sum_{j=1}^{p} \lambda_j^k = trace(\mathbf{K}_\rho^k) = trace\{(\mathbf{G}'\mathbf{G}\mathbf{W}\mathbf{R}_\rho\mathbf{W})^k\} \tag{1}$$

and

$$\sum_{j=1}^{p} \lambda_j^k \delta_j = \boldsymbol{\mu}_\beta' \mathbf{K}_\rho^k \boldsymbol{\mu}_\beta = trace(\boldsymbol{\mu}_\beta' \mathbf{K}_\rho^k \boldsymbol{\mu}_\beta) = trace\{(\mathbf{G}'\mathbf{G}\mathbf{W}\mathbf{R}_\rho\mathbf{W})^{k-1}\mathbf{G}'\boldsymbol{\mu}_\beta\boldsymbol{\mu}_\beta'\mathbf{G}\mathbf{W}\mathbf{R}_\rho\mathbf{W}\}. \tag{2}$$

Suppose $\mathbf{A} = E(\mathbf{G}'\mathbf{G}\mathbf{W}\mathbf{R}_\rho\mathbf{W}/n)$ and $\mathbf{B} = E(\mathbf{G}'\boldsymbol{\mu}_\beta\boldsymbol{\mu}_\beta'\mathbf{G}\mathbf{W}\mathbf{R}_\rho\mathbf{W}/n^2)$. Since the distribution of $\mathbf{G}$ can be inferred from simulations under accepted population genetic models or existing data (e.g. 1000 genome project data), we can obtain both $\mathbf{A}$ and $\mathbf{B}$. By the continuity of trace and matrix multiplication, $trace(\mathbf{K}_\rho^k) = n^k trace(\mathbf{A}^k)$ and $trace(\boldsymbol{\mu}_\beta' \mathbf{K}_\rho^k \boldsymbol{\mu}_\beta) = n^{k+1} trace(\mathbf{A}^{k-1}\mathbf{B})$. After computing $c_1, \ldots, c_4$, we obtain following values.

$$\begin{aligned}
&\mu_Q = c_1, \quad \sigma_Q = \sqrt{2c_2}, \quad s_1 = c_3/c_2^{3/2}, s_2 = c_4/c_2^2, \\
&a = \begin{cases} 1/\left(s_1 - \sqrt{s_1^2 - s_2}\right) & \text{if } s_1^2 > s_2 \\ 1/\sqrt{s_2} & \text{if } s_1^2 \le s_2 \end{cases}, \\
&\delta = \begin{cases} s_1 a^3 - a^2 & \text{if } s_1^2 > s_2 \\ 0 & \text{if } s_1^2 \le s_2 \end{cases}, \\
&l = a^2 - 2\delta, \quad \mu_X = l + \delta, \quad \text{and } \sigma_X = \sqrt{2}\sqrt{l + 2\delta}.
\end{aligned}$$

Note that we modified the approximation of Liu *et al.*(2009) [1] when $s_1^2 \le s_2$ by matching the kurtosis, instead of the skewness, to improve the estimation of the tail probability. To estimate the power, we first compute μ_Q, μ_X, σ_Q, and σ_X under the null and find the critical value at level α is

$$q_c = (q(1-\alpha; \chi_l^2(\delta)) - \mu_X)\frac{\sigma_Q}{\sigma_X} + \mu_Q,$$

where $q(\cdot; \chi_l^2(\delta))$ is a quantile function of $\chi_l^2(\delta)$. Then, we recompute μ_Q, μ_X, σ_Q, and σ_X under the alternative and estimate the power as

$$P\left(\chi_l^2(\delta) > \frac{\sigma_X}{\sigma_Q}(q_c - \mu_Q) + \mu_X\right).$$

A.2 Dichotomous Traits in Cross-Sectional and Prospective Studies

In the absence of covariates, the logistic model we consider is

$$\text{logit}\,(\pi_i) = \alpha_0 + \mathbf{G}_i'\boldsymbol{\beta}, \tag{3}$$

where y_i is a disease status (1 = disease, 0 = non-disease). We assume that the prevalence/incidence of disease is known. Our test statistic with a kernel $\mathbf{K}_\rho$ is

$$Q = (\mathbf{y} - \hat{\pi}_0\mathbf{1})'\mathbf{K}_\rho(\mathbf{y} - \hat{\pi}_0\mathbf{1}),$$

where $\hat{\pi}_0 = n^{-1}\sum_{i=1}^n y_i$, the estimated disease probability under H_0. Denote $\boldsymbol{\mu}_\beta = (\pi_1 - \hat{\pi}_0, \ldots, \pi_n - \hat{\pi}_0)'$, where π satisfies (3), and $\mathbf{V} = diag[v_1, \ldots, v_n]$, and $v_i = \pi_i(1-\pi_i)$ is $var(y_i)$. Then Q can be written as

$$\begin{aligned} Q &= (\mathbf{y} - \hat{\pi}_0\mathbf{1})'\mathbf{K}_\rho(\mathbf{y} - \hat{\pi}_0\mathbf{1}) \\ &= (\mathbf{y} - \hat{\pi}_0\mathbf{1} - \boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta)'\mathbf{V}^{-1/2}\mathbf{V}^{1/2}\mathbf{K}_\rho\mathbf{V}^{1/2}\mathbf{V}^{-1/2}(\mathbf{y} - \hat{\pi}_0\mathbf{1} - \boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta) \\ &= (\mathbf{E} + \mathbf{V}^{-1/2}\boldsymbol{\mu}_\beta)'\widetilde{\mathbf{K}}_\rho(\mathbf{E} + \mathbf{V}^{-1/2}\boldsymbol{\mu}_\beta), \end{aligned}$$

where $\mathbf{E} = \mathbf{V}^{-1/2}(\mathbf{y} - \hat{\pi}_0\mathbf{1} - \boldsymbol{\mu}_\beta)$, and $\widetilde{\mathbf{K}}_\rho = \mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2}$. Since each element of $\mathbf{E}$ has mean 0 and variance 1, $(\mathbf{u}_j'\mathbf{E})^2$ asymptotically follows independent χ_1^2 distribution. Now we apply the same argument shown in the Section A.1 using $\widetilde{\mathbf{K}}_\rho$ instead of $\mathbf{K}_\rho$, and estimate the power.

A.3 Modifications of Power Calculations for Rare Variants

With finite sample size n, causal variants that are rare may not be observed. Our power and sample size calculations can account for this uncertainty. Suppose the population MAF for the j^{th} variant is m_j. Let $\theta_j = 1 - (1-m_j)^{2n}$ be the probability the variant j is observed (polymorphic) in sample size n. With rare variants, the model we fit is actually $y_i = \alpha_0 + \widetilde{\mathbf{G}}_i\boldsymbol{\beta} + \epsilon_i$ for continuous traits, and logit $(\pi_i) = \alpha_0 + \widetilde{\mathbf{G}}_i\boldsymbol{\beta}$ for dichotomous traits, where $\widetilde{\mathbf{G}}_i = (G_{i1}\widetilde{\Delta}_1, \ldots, G_{ip}\widetilde{\Delta}_p)' = \widetilde{\boldsymbol{\Delta}}\mathbf{G}_i$, and $\widetilde{\boldsymbol{\Delta}} = diag[\widetilde{\Delta}_1, \ldots, \widetilde{\Delta}_p]$. Here, $\widetilde{\Delta}_j$ is an indicator that variant j is observed in sample size n. Under this model, $\mathbf{K}_\rho = \widetilde{\mathbf{G}}\mathbf{W}\mathbf{R}_\rho\mathbf{W}\widetilde{\mathbf{G}} = \mathbf{G}\widetilde{\boldsymbol{\Delta}}\mathbf{W}\mathbf{R}_\rho\mathbf{W}\widetilde{\boldsymbol{\Delta}}\mathbf{G}'$. Suppose $\boldsymbol{\Pi} = diag[\theta_1, \ldots, \theta_p]$, and $\mathbf{H}$ is a $p \times p$ matrix with the (i,j)th element being $r_{ij}\theta_i\theta_j^{I(i\neq j)}$, where r_{ij} is the $(i,j)^{th}$ element of $\mathbf{R}_\rho$. Then, $trace(\mathbf{K}_\rho^k)$ and $trace(\boldsymbol{\mu}_\beta'\mathbf{K}_\rho^k\boldsymbol{\mu}_\beta)$ can be approximated as $trace(\mathbf{K}_\rho^k) \approx n^k trace((\mathbf{A}_H)^k)$ and $trace(\boldsymbol{\mu}_\beta'\mathbf{K}_\rho^k\boldsymbol{\mu}_\beta) \approx n^{k+1}trace((\mathbf{A}_H)^{k-1}\mathbf{B}_H)$, where $\mathbf{A}_H = E(\mathbf{W}\mathbf{G}'\mathbf{G}\mathbf{W})\mathbf{H}/n$ and $\mathbf{B}_H = E(\mathbf{W}\mathbf{G}'\boldsymbol{\mu}_\beta\boldsymbol{\mu}_\beta')\mathbf{H}/n^2$. Using these changes, we can compute the power using the χ^2 approximation method from the previous section.

A.4 Power and Sample Size Calculations for Retrospective Case-Control Studies

It is well known that logistic regression can be used to analyze case-control data [2]. However, it is necessary to incorporate the retrospective nature of case-control studies to properly estimate the power. Let S be a selection indicator such that $S = 1$ denotes a subject is selected in the case-control sample. Then the conditional distribution of G and y given $S = 1$, instead of the unconditional distribution of G and y, should be used to compute power. Denote by $\tilde{\pi}_i = Pr(y_i = 1|\mathbf{G}_i, S_i = 1)$ the case-control probability. If the the population disease probability follows the logistic model (3), then the case-control probability $\tilde{\pi}_i$ also follows the same logistic model except for a different intercept [2] as

$$\text{logit}\,(\tilde{\pi}_i) = \tilde{\alpha}_0 + \mathbf{G}_i'\boldsymbol{\beta}, \tag{4}$$

where

$$\tilde{\alpha}_0 = \alpha_0 + log\left\{\frac{P(S=1|y=1)}{P(S=1|y=0)}\right\} = \alpha_0 + log\left\{\frac{\hat{\pi}_0 P(y=0)}{(1-\hat{\pi}_0)P(y=1)}\right\}, \tag{5}$$

where $P(S=1|y=1)$ is the probability that a case is sampled, $P(S=1|y=0)$ is the probability that a control is sampled, and $P(y=1)$ is the population disease prevalence/incidence. Further one can show that

$$\begin{aligned} P(G|S=1) &= P(G|y=1,S=1)P(y=1|S=1) + P(G|y=0,S=1)P(y=0|S=1) \\ &= \frac{\hat{\pi}_0}{P(y=1)}P(y=1|G)P(G) + \frac{1-\hat{\pi}_0}{P(y=0)}P(y=0|G)P(G). \end{aligned} \tag{6}$$

We compute $\mathbf{A}$, $\mathbf{B}$, $\mathbf{W}$, $\mathbf{A}_2$ and $\mathbf{\Pi}$ by estimating μ_β and $\mathbf{V}$ using (5) and by using conditional distribution (6), and subsequently estimate the power.

A.5 Computing the Average Power Across Different Regions

The power to detect an association between a particular region and trait depends strongly on the LD structure of the genomic region to be investigated and the MAFs of the causal variants. If one is interested in only one known region and knows in advance which variants are causal, the power formula above can be directly applied. In practice however, one is usually interested in more than one region and one can only hypothesize as to the role of the causal variants in the disease model. For example, one may hypothesize that a certain percentage of rare variants are causal, instead of selecting the causal variants *a priori*. In this case, we propose to average the power computed across multiple regions under a genetic disease model. Specifically, the average power can be easily computed by randomly selected regions/causal variants under a particular genetic disease model and then taking the mean power across the selected regions and variants. Our experience shows that approximately 100 $\sim$ 500 sets of different regions/causal variants are sufficient to compute the average power stably.

B Derivation of the formula (6)

Denote the SKAT test statistic with true parameter $(\boldsymbol{\alpha}, \phi)$ with $\psi = 0$ as

$$Q_\rho^* = (\mathbf{y}^* - \mathbf{X}\boldsymbol{\alpha})'\mathbf{V}^{-1}\mathbf{K}_\rho\mathbf{V}^{-1}(\mathbf{y}^* - \mathbf{X}\boldsymbol{\alpha}). \tag{7}$$

Define $\mathbf{u} = \mathbf{V}^{-1/2}(\mathbf{y}^* - \mathbf{X}\boldsymbol{\alpha})$, and then each entry of $\mathbf{u}$ has mean zero and variance one. Suppose $\mathbf{Z}_0$, $\bar{\mathbf{z}}_0$ and $\mathbf{M}_0$ are $\mathbf{Z}$, $\bar{\mathbf{z}}$ and $\mathbf{M}$ with the true parameter $(\boldsymbol{\alpha}, \phi)$. Then

$$(7) = (1-\rho)\mathbf{u}'\mathbf{Z}_0\mathbf{Z}_0'\mathbf{u} + \rho\mathbf{u}'\mathbf{Z}_0\mathbf{1}\mathbf{1}'\mathbf{Z}_0'\mathbf{u} = (1-\rho)\mathbf{u}'\mathbf{Z}_0\mathbf{Z}_0'\mathbf{u} + p\rho\mathbf{u}'\bar{\mathbf{z}}_0\bar{\mathbf{z}}_0'\mathbf{u}. \tag{8}$$

From the fact

$$\mathbf{M}_0\mathbf{Z}_0\mathbf{Z}_0'\mathbf{M}_0 = \sum_{j=1}^{p}(\bar{\mathbf{z}}_0'\mathbf{z}_{0.j})^2\frac{\bar{\mathbf{z}}_0\bar{\mathbf{z}}_0'}{(\bar{\mathbf{z}}_0'\bar{\mathbf{z}}_0)^2},$$

it can be shown

$$(8) = (1-\rho)\mathbf{u}'(\mathbf{I}-\mathbf{M}_0)\mathbf{Z}_0\mathbf{Z}_0'(\mathbf{I}-\mathbf{M}_0)\mathbf{u} + 2(1-\rho)\mathbf{u}'(\mathbf{I}-\mathbf{M}_0)\mathbf{Z}_0\mathbf{Z}_0'\mathbf{M}_0\mathbf{u} + \tau(\rho)\mathbf{u}'\bar{\mathbf{z}}_0\bar{\mathbf{z}}_0'\mathbf{u}/\bar{\mathbf{z}}_0'\bar{\mathbf{z}}_0. \quad (9)$$

Denote $\zeta = 2(1-\rho)\mathbf{u}'(\mathbf{I}-\mathbf{M}_0)\mathbf{Z}_0\mathbf{Z}_0'\mathbf{M}_0\mathbf{u}$, and then $E(\zeta) = 0$. Since $\mathbf{M}_0$ is a projection matrix, $(\mathbf{I}-\mathbf{M}_0)\mathbf{u}$ and $\mathbf{M}_0\mathbf{u}$ are asymptotically independent. Therefore, the asymptotic variance of ζ is $4trace(\mathbf{Z}_0'\mathbf{M}_0\mathbf{Z}_0\mathbf{Z}_0'(\mathbf{I}-\mathbf{M}_0)\mathbf{Z}_0)$, and both $Corr(\mathbf{u}'(\mathbf{I}-\mathbf{M}_0)\mathbf{Z}_0\mathbf{Z}_0'(\mathbf{I}-\mathbf{M}_0)\mathbf{u},\zeta)$ and $Corr(\mathbf{u}'\bar{\mathbf{z}}_0\bar{\mathbf{z}}_0'\mathbf{u},\zeta)$ are zero asymptotically.

Since each elements of $\mathbf{u}$ has mean 0 and variance 1, $\mathbf{u}'(\mathbf{I}-\mathbf{M}_0)\mathbf{Z}_0\mathbf{Z}_0'(\mathbf{I}-\mathbf{M}_0)\mathbf{u}$ asymptotically follows $\sum_{k=1}^{m}\lambda_k\eta_k$, where η_ks are independent χ_1^2 random variables, and $\mathbf{u}'\bar{\mathbf{z}}_0\bar{\mathbf{z}}_0'\mathbf{u}/\bar{\mathbf{z}}_0'\bar{\mathbf{z}}_0$ asymptotically follows χ_1^2 distribution. By asymptotic independence between $(\mathbf{I}-\mathbf{M}_0)\mathbf{u}$ and $\mathbf{M}_0\mathbf{u}$, it can be shown that $\mathbf{u}'(\mathbf{I}-\mathbf{M}_0)\mathbf{Z}_0\mathbf{Z}_0'(\mathbf{I}-\mathbf{M}_0)\mathbf{u}$ and $\mathbf{u}'\bar{\mathbf{z}}_0\bar{\mathbf{z}}_0'\mathbf{u}$ are also asymptotically independent.

From the facts that $\widehat{\boldsymbol{\alpha}}$ and $\hat{\phi}$ are consistent estimators of $(\boldsymbol{\alpha},\phi)$, and Q is a continuous function of $(\boldsymbol{\alpha},\phi)$ with the finite first derivative in the neighborhood of true parameter $(\boldsymbol{\alpha},\phi)$, Q_ρ approximately follows the same asymptotic distribution of Q_ρ^*.

C Relationship between ρ and fixed β

We derived the SKAT-O test statistic by assuming the β coefficients are random with the correlation ρ among βs. It is of substantial interest to understand for fixed β coefficients, how the power of the SKAT-O test depends on the percentage of causal variants and the percentage of causal variants that have different signs. To investigate this, we derive the relationship between ρ and fixed values of β coefficients as a function of the percentages of βs being non-zero and the percentage of non-zero βs that are positive. This result also allows us to specify ρ in power calculations when β coefficients are provided by investigators.

The derivation of SKAT-O assumes β coefficients to have a distribution with mean 0 and variance $w_i\tau$, where the w_i is the weight for the i_{th} variant. Without loss of generosity, we assume $\tau = 1$. It follows that the $\beta_i^* = \beta_i/w_i$ follows a distribution with mean zero and variance τ. Since ρ is a correlation coefficient among βs, we have $\rho = E(\beta_i^*\beta_j^*)$. Hence for a given set of values of βs, we have

$$\begin{aligned}\rho &\approx \frac{1}{p(p-1)}\sum_{i\neq j}\beta_i^*\beta_j^* \\ &= \frac{1}{p(p-1)}\sum_{i\neq j}\beta_i^*\beta_j^* I(\beta_i^*>0,\beta_j^*>0) + \frac{2}{p(p-1)}\sum_{i\neq j}\beta_i^*\beta_j^* I(\beta_i^*<0,\beta_j^*>0) \\ &\quad + \frac{1}{p(p-1)}\sum_{i\neq j}\beta_i^*\beta_j^* I(\beta_i^*<0,\beta_j^*<0), \end{aligned} \quad (10)$$

where $I(\cdot)$ is an indicator function and p is the number of β coefficients. Suppose for fixed values of βs, the proportion of nonzero β_i is p_1, and among the non-zero βs, β_i is a function of MAFs as $|\beta_i| = w_i\beta_0$. The proportion of positive β_is among the nonzero β_is is p_2. Without loss of generality,

we set $\beta_0 = 1$. Thus the nonzero β_i^* take values either -1 or 1. Since

$$\frac{1}{p(p-1)} \sum_{i \neq j} I(\beta_i^* > 0, \beta_j^* > 0) \approx p_1^2 p_2^2,$$
$$\frac{1}{p(p-1)} \sum_{i \neq j} I(\beta_i^* < 0, \beta_j^* > 0) \approx p_1^2 p_2 (1-p_2), \quad \text{and}$$
$$\frac{1}{p(p-1)} \sum_{i \neq j} I(\beta_i^* < 0, \beta_j^* < 0) \approx p_1^2 (1-p_2)^2,$$

plugging these into (10), we have

$$\rho \approx p_1^2 [p_2^2 - 2p_2(1-p_2) + (1-p_2)^2] = p_1^2 (2p_2 - 1)^2. \tag{11}$$

For illustration, the following table gives the estimated ρ values with different parameter configuration of the β coefficients. It shows that when only 10% of rare variants are causal variants and all nonzero βs are positive, the estimated optimal ρ is 0.01. If 50% of rare variants are causal variants, the estimated ρ is 0.25. If there are a small percentage of variants that are in different directions, ρ is small.

	Causal=10%	Causal=20%	Causal=50%
$\beta +/- = 100/0$	0.01	0.04	0.25
$\beta +/- = 80/20$	0.0036	0.0144	0.09

We note that (11) is derived under the assumption that w_is are known and the β_is are assumed to be a function of MAFs, these might not be true in real data. However, it provides a clear insight into the behavior of ρ given fixed values of β coefficients.

D Type I error rate at genome-wide α level

Although the proposed SKAT-O is computationally efficient, it is very challenging to simulate more than 10^7 p-values, which is required to estimate the type I error rate at the genome-wide α level. For example, in whole exome sequencing studies the number of genes are around $20,000$, and thus the Bonferroni adjusted level 0.05 is 2.5×10^{-6}. To reduce the computational burden, we first generated $10,000$ sets of genotype data, each from a different randomly selected region. We then generated $1,000$ phenotype sets for each of the $10,000$ genotype sets. No additional covariates were used. Since the $1,000$ phenotype sets share common genotype set, p-values can be rapidly computed. Although the obtained type I error rates from estimated this approach is not exactly the same as the the empirical type I error rates generated under 10^7 unique phenotype/genotype sets, our estimates are still unbiased and result in very accurate type I error rate estimates. To further reduce computation burden, we restricted the sample size to be $n = 2,000$ and only considered the SKAT-O procedure.

Table 2 in the main manuscript illustrates the empirical type I error rates with three different α levels. It shows that SKAT-O can accurately control type I error with moderate α levels, but produces slightly inflated type I error rates at very small α levels.

E Effect of ρ values to the power

We investigated the effect of different ρ values to the power under varying genetic models of association. We consider the same linear and logistic models in Section 4.2 without covariates. The regression coefficient β follows the same log function in Section 4.2. Three different ρ values were used to compute the power in which one ρ was obtained from the equation (3.8) and the others were 0 and 1. The power curves were computed using the power calculation formula at the significance level $\alpha = 10^{-3}$ (Supplementary Figure 2 and 3). We also obtained the empirical powers of SKAT-O from 1,000 simulated datasets under the same genetic model. For the empirical power, we only considered three different sample sizes $n = (1000, 2000, 5000)$.

F Accuracy of the power calculation formula

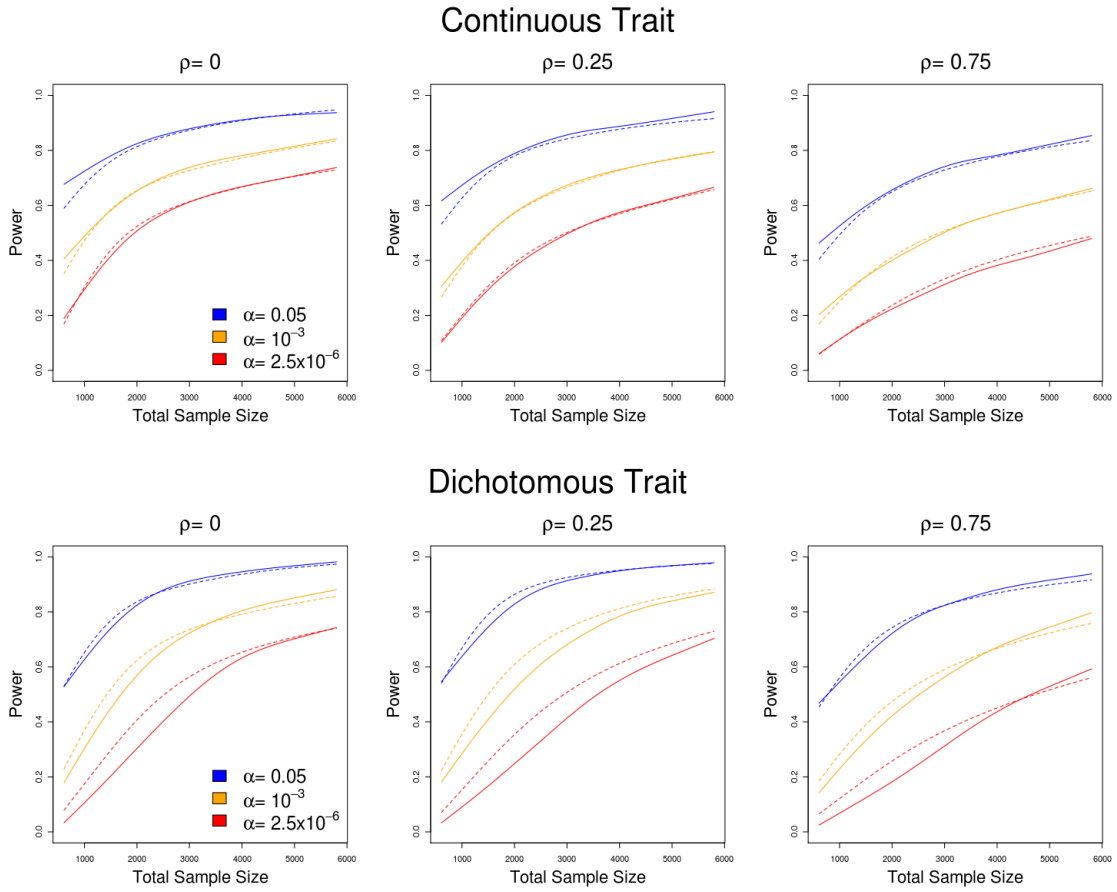
We conducted simulations to evaluate the accuracy of the power calculation formula. In particular, we estimated the statistical power to detect an arbitrary $3kb$ region as associated with a trait. We considered the setting in which 20% of the rare variants were causal variants and 20% of non-zero β coefficients were negative. 3 different level αs ($\alpha = 0.05, 10^{-3}, 2.5 \times 10^{-6}$) were considered. The power was computed by averaging the power obtained from 500 randomly selected regions and causal SNP sets. We computed the power with 3 different ρs ($\rho = 0, 0.25, 0.75$). The same log function was used for coefficient βs, and we set $c = 1/2$ for continuous trait and $c = 0.549$ for binary trait. Supplementary Figure 1 compares the estimated power from the power formula and the empirical power from 1000 simulated datasets. It clearly shows that we can accurately estimate the power using the formula.

References

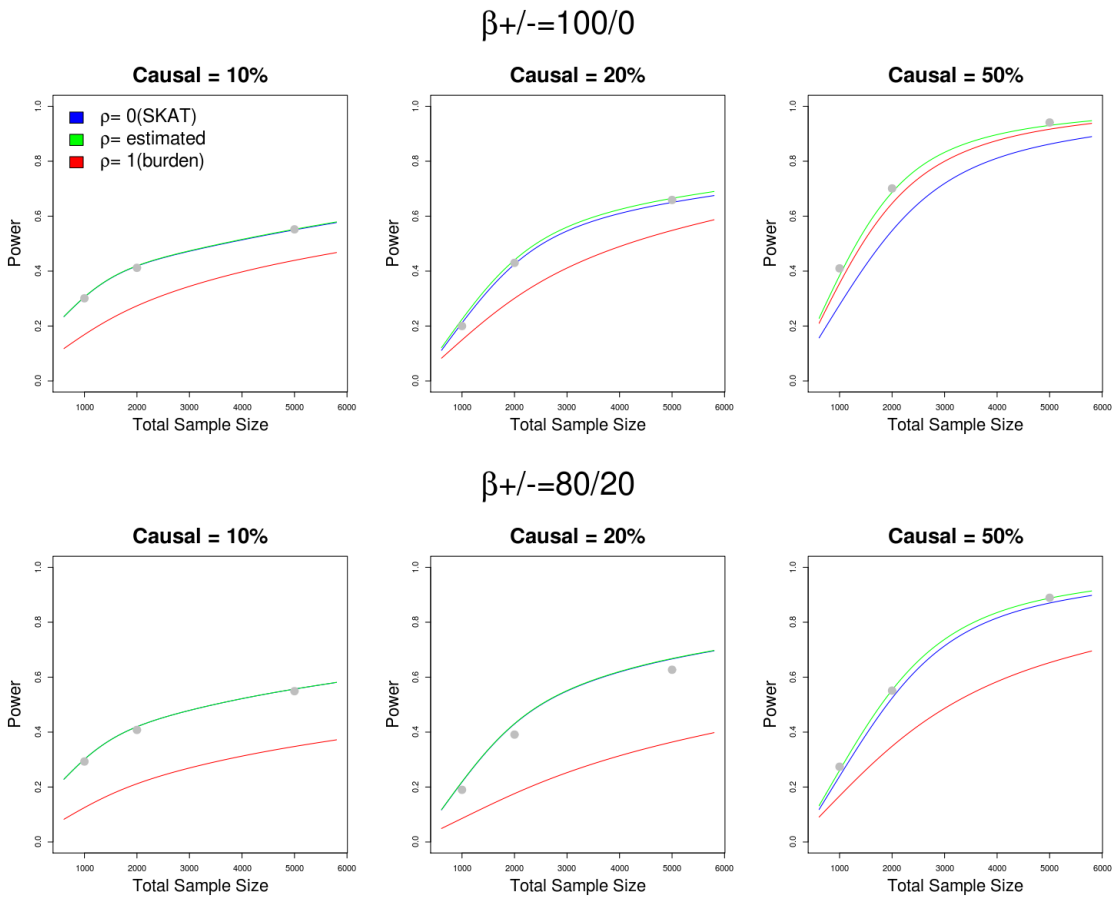
[1] H. Liu, Y. Tang, and H.H. Zhang. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*, 53(4):853–856, 2009.

[2] R. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.

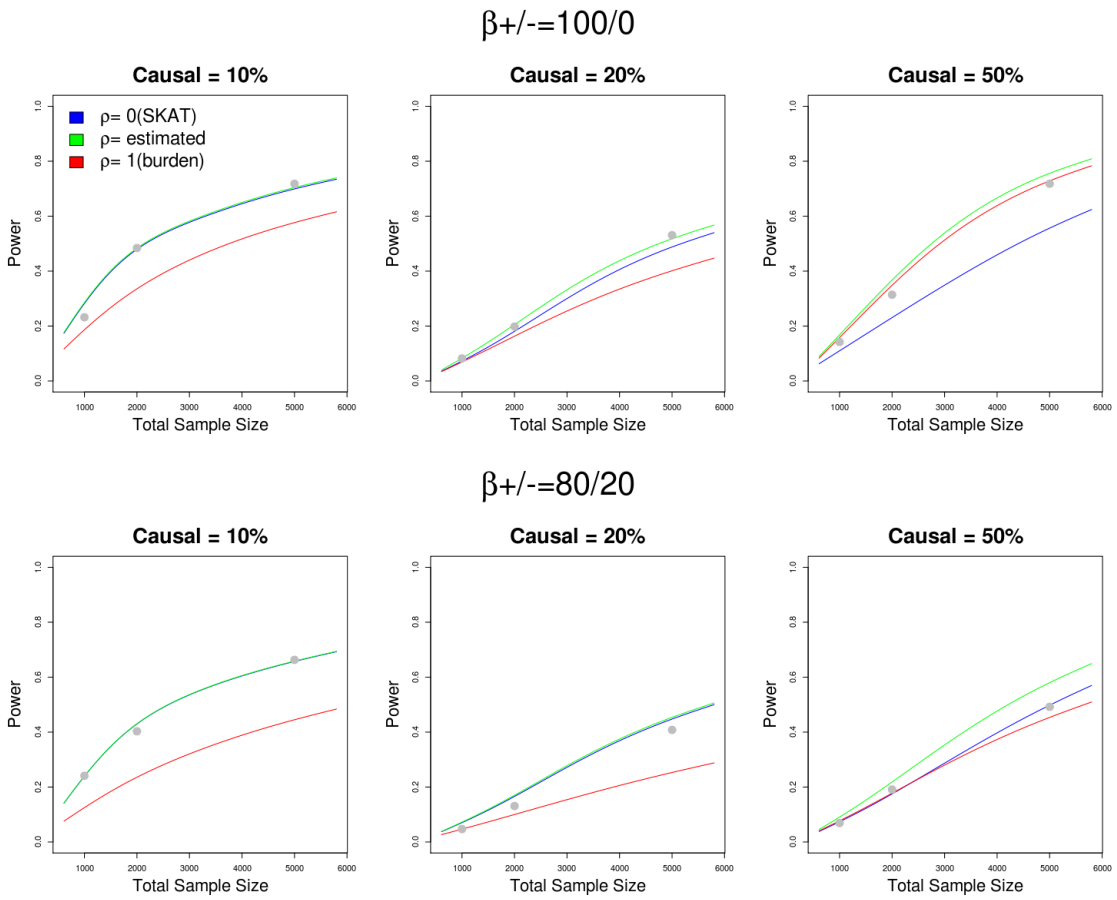
G Supplementary Figures



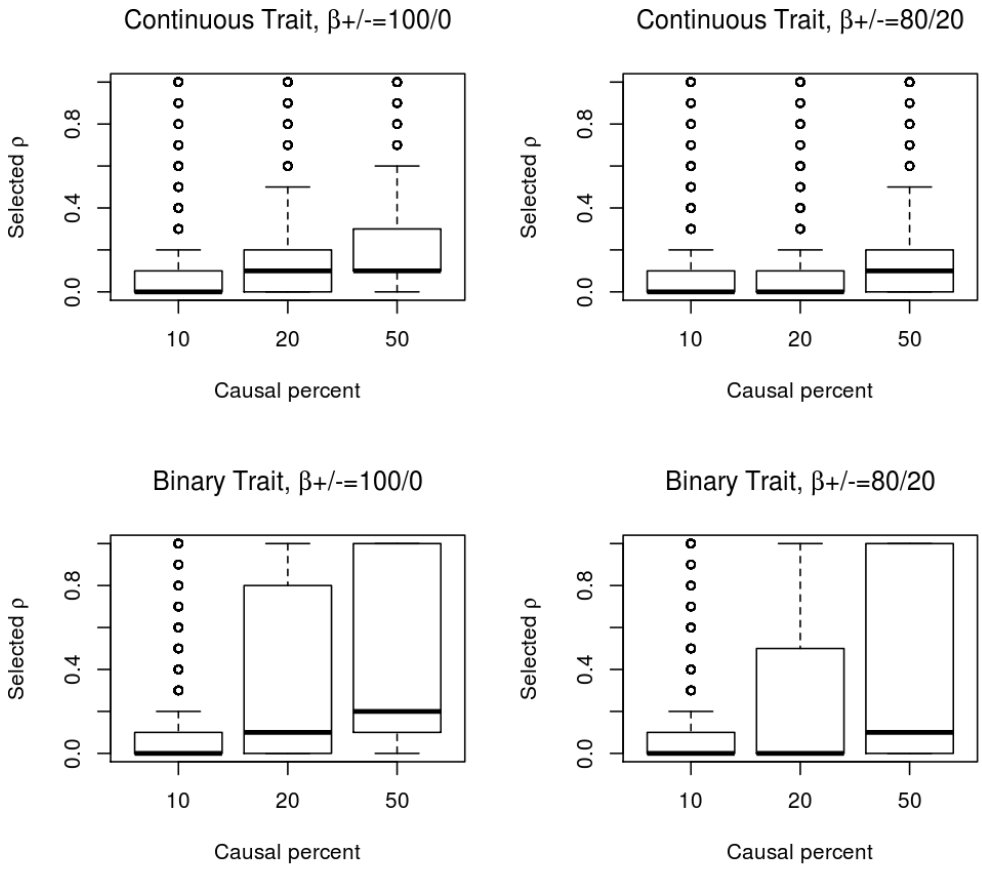
Supplementary Figure 1: Comparison of power from simulation and analytical estimation. From left to right, the plots consider the setting in which $\rho = 0$, $\rho = 0.25$, and $\rho = 0.75$. 3 different colors represents 3 different α levels. Solid line represents empirical power, and dashed line represents approximated power obtained from the power formula.



Supplementary Figure 2: Comparison of power with different ρ values at $\alpha = 10^{-3}$ using the power calculation formula when the region size was 3kb and phenotypes had continuous values. Top panel considers $\beta +/- = 100/0$ and bottom panel considers $\beta +/- = 80/20$. From left to right, the plots consider the setting in which 10% of rare variants were causal, 20% of rare variants were causal, and 50% of rare variants were causal. 3 different colors represents 3 different ρ values. "ρ=estimated" used a ρ value calculated from (C.2). Three gray dots represents empirical powers of SKAT-O obtained from 1,000 simulated datasets.



Supplementary Figure 3: Comparison of power with different ρ values at $\alpha = 10^{-3}$ using the power calculation formula when the region size was 3kb and phenotypes had binary values. Top panel considers $\beta + /- = 100/0$ and bottom panel considers $\beta + /- = 80/20$. From left to right, the plots consider the setting in which 10% of rare variants were causal, 20% of rare variants were causal, and 50% of rare variants were causal. 3 different colors represents 3 different ρ values. "ρ=estimated" used a ρ value calculated from (C.2). Three gray dots represents empirical powers of SKAT-O obtained from 1,000 simulated datasets.



Supplementary Figure 4: Box plots of estimated optimal ρ values of the power simulations in Section 4.2. Top panel considers continuous traits and bottom panel considers binary traits. From left to right, the plots consider the setting in which $\beta + /- = 100/0$ and $\beta + /- = 80/20$.