Rare Variant Association Testing for Sequencing Data Using the Sequence Kernel Association Test (SKAT)

Michael C. Wu,*[1] Seunggeun Lee,*[2] Tianxi Cai,[2] Yun Li,[1,3] Michael Boehnke[4] and Xihong Lin[2]

*These authors contributed equally to this work.

[1] Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[2] Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

[3] Department of Genetics, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[4] Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA

Corresponding Author:
Xihong Lin
Department of Biostatistics
Harvard School of Public Health
655 Huntington Ave, Building 2, 4th Floor
Boston, MA 02115
Tel: (617) 432-2914
Fax: (617) 432-5619
Email: xlin@hsph.harvard.edu

ABSTRACT

Sequencing studies are increasingly being conducted to identify rare variants associated with complex traits. The limited power of classical single marker association analysis for rare variants poses a central challenge in such studies. We propose the sequence kernel association test (SKAT), a supervised, flexible, computationally efficient regression method to test for association between genetic variants (common and rare) in a region and a continuous or dichotomous trait, while easily adjusting for covariates. As a score-based variance component test, SKAT can quickly calculate p-values analytically by fitting the null model containing only the covariates, and so can easily be applied to genome-wide data. Using SKAT to analyze a genome-wide sequencing study of 1000 individuals, by segmenting the whole genome into 30kb regions, requires only 7 hours on a laptop. Through analysis of simulated data across a wide range of practical scenarios and triglyceride data from the Dallas Heart Study, we show that SKAT can substantially outperform several alternative rare-variant association tests. We also provide analytic power and sample size calculations to help design candidate gene, whole exome, and whole genome sequence association studies.

INTRODUCTION

Genome-wide association studies (GWAS) have identified >1000 genetic loci associated with many human diseases and traits[1], yet common variants identified through GWAS often explain only a small proportion of trait heritability. The advent of massively parallel sequencing[2] has transformed human genetics[3-4] and has the potential to explain some of this missing heritability through identification of trait-associated rare variants[5]. Although considerable resources have been devoted to sequence mapping and genotype calling[6-9], successful application of sequencing to the study of complex traits requires novel statistical methods to test efficiently for association given data on rare variants[10] and to perform sample size and power calculations to help design sequencing-based association studies.

Rare genetic variants, here defined as alleles with frequency <1-5%, can play key roles in influencing complex disease and traits[11]. However, standard methods used to test for association with single common genetic variants are underpowered for rare variants unless sample sizes or effect sizes are very large[12-13]. A logical alternative approach is to employ burden tests that assess the cumulative effects of multiple variants in a genomic region[12-18]. Burden tests proposed to date are based on collapsing or summarizing the rare variants within a region by a single value which is then tested for association with the trait of interest. For example, the cohort allelic sum test (CAST)[14] collapses information on all rare variants within a region (e.g. exons of a gene) into a single dichotomous variable for each subject indicating whether or not the subject has any rare variants within the region, and then applies a univariate test. Instead of dichotomizing, collapsing by counting the number of rare variants within a region is also possible[18]. The Combined Multivariate and Collapsing method[12] extends CAST by collapsing rare variants within a region into subgroups based on allele frequency, collapsing subgroups as in CAST, and applying a multivariate test to the subgroups. The weighted sum test (WST)[13] specifically considers the case-control setting and collapses a set of SNPs into a single weighted average of the number of rare alleles for each individual. Numerous alternative methods are largely variations on these approaches[16-17; 19].

A limitation for all these burden tests is that they implicitly assume that all rare variants influence the phenotype in the same direction and with the same magnitude of effect (after incorporating known weights). However, one would expect most variants (common or rare) within a sequenced region to have little or no effect on phenotype, while some variants are protective and others deleterious, and the magnitude of each variant's effect is likely to vary (e.g. rarer variants might have larger effects). Hence, collapsing across all variants is likely to introduce substantial noise in the aggregated index, attenuate evidence for association, and result in power loss. Further, burden tests require either specification of thresholds for collapsing or the use of permutation to estimate the threshold[16-20]. Permutation tests are computationally expensive, especially on the whole-genome scale, and are difficult for covariate adjustment since permutation requires independence between the genotype and the covariates.

The recently proposed C-alpha test[21] is a non-burden based test and is hence robust to the direction and magnitude of effect. For case-control data, it compares the expected variance to the actual variance of the distribution of allele frequencies. These important advantages allow the C-alpha test to have improved power over burden based tests especially when the effects are in different directions. Despite these attractive features, the C-alpha test does not allow for easy covariate adjustment, such as for controlling population stratification, which is important in genetic association studies. The C-alpha test also uses permutation to obtain a p-value when linkage disequilibrium is present among the variants, which is, as noted earlier, computationally expensive for whole genome experiments. The approach has not been generalized to analysis of continuous phenotypes.

We propose in this paper the sequence kernel association test (SKAT), a flexible, computationally efficient, regression approach to test for association between variants in a region (both common and rare) and a dichotomous (e.g., case-control) or continuous phenotype, while adjusting for covariates, such as principal components to account for population stratification[22]. The kernel machine regression framework was previously considered for common variants[23-24]. In this paper, we provide several essential methodological improvements necessary for testing rare variants. SKAT uses a multiple regression model to directly regress the phenotype on genetic variants in a region and on covariates, and so allows different variants to have different directions and magnitude of effects, including no effects; SKAT also avoids selection of thresholds. We develop a kernel association test to test the regression coefficients of the variants using a variance component score test by accounting for rare variants.

SKAT is computationally efficient, especially important in genome-wide studies. This is because SKAT only requires fitting the null model in which phenotypes are regressed on the covariates alone; p-values are easily computed using simple analytic formulae. Additional features of SKAT include exploitation of local correlation structure, incorporation of flexible weights to boost power (e.g. by up-weighting rarer variants or incorporating functionality), and allowance for epistatic variant effects. As discussed in more detail below, under special cases, the SKAT, C-alpha, and individual variant tests statistics are closely related.

We demonstrate through simulation and analysis of re-sequencing data from the Dallas Heart Study that SKAT is often more powerful than existing tests across a broad range of models for both continuous and dichotomous data. We also investigate the factors that influence power for sequence association studies. Finally, we describe analytic tools to estimate statistical power and sample sizes to guide the design of new sequence association studies of rare variants using SKAT.

MATERIALS AND METHODS

Sequencing Kernel Association Test

SKAT is a supervised test for the joint effects of multiple variants in a region on a phenotype. Regions can be defined using genes (in candidate gene or whole exome studies) or moving windows across the genome (in whole genome studies). For each region, SKAT analytically calculates a p-value for association while adjusting for covariates. Adjustments for multiple comparisons are necessary for analyzing multiple regions, e.g. using the Bonferroni correction or FDR control.

Notation

Assume n subjects are sequenced in a region with p variant sites observed. Covariates might include age, gender, and top principal components of genetic variation for controlling population stratification[22]. For the i-th subject, y_i denotes the phenotype variable, $\mathbf{X}_i=(X_{i1}, X_{i2}, .., X_{im})$ the covariates, and $\mathbf{G}_i=(G_{i1}, G_{i2}, \ldots, G_{ip})$ the genotypes for the p variants within the region. Typically, we assume an additive genetic model and let $G_{ij,}=0$, 1, or 2 represent the number of copies of the minor allele. Dominant and recession models can also be considered.

Linear SKAT Model and Test

For a simple illustration of SKAT, we focus here on the testing for a relationship between the variants and the phenotype using classical multiple linear and logistic regression. We describe how the SKAT can incorporate epistatic effects later. To relate the sequence variants in a region to the phenotype, consider the linear model

$$y_i = \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_i + \boldsymbol{\beta}'\mathbf{G}_i + \varepsilon_i \qquad (1)$$

when the phenotypes are continuous traits, and the logistic model

$$\operatorname{logit} P(y_i = 1) = \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_i + \boldsymbol{\beta}'\mathbf{G}_i \qquad (2)$$

when the phenotypes are dichotomous (e.g. $y=0/1$ for case/control). Here α_0 is an intercept term, $\boldsymbol{\alpha}=[\alpha_1,\ldots,\alpha_m]'$ is the vector of regression coefficients for the m covariates, $\boldsymbol{\beta}=[\beta_1,\ldots,\beta_p]'$ is the vector of regression coefficients for the p observed gene variants in the region, and for continuous phenotypes ε_i is an error term with mean zero and variance σ^2. Under both linear and logistic models, evaluating whether the gene variants influence the phenotype, adjusting for covariates, corresponds to testing the null hypothesis H_0: $\boldsymbol{\beta}=\mathbf{0}$, i.e. $\beta_1=\beta_2=\ldots=\beta_p=0$. The standard p-DF likelihood ratio test has little power, especially for rare variants. To increase the power, SKAT tests H_0 by assuming each β_j follows an arbitrary distribution with mean zero and variance $w_j\tau$, where τ is a variance component and w_j is a pre-specified weight for variant j. One can easily see that H_0: $\boldsymbol{\beta}=\mathbf{0}$ is equivalent to testing H_0: $\tau=0$, which can be conveniently tested using a variance component score test in the corresponding mixed model; this is known to be a locally most powerful test[25]. A key advantage of the score test is that it only requires fitting the null model $y_i=\alpha_0+\boldsymbol{\alpha}_1'\mathbf{X}_i+\varepsilon_i$ for continuous traits and logit $P(y_i=1)= \alpha_0+\boldsymbol{\alpha}_1'\mathbf{X}_i$ for dichotomous traits.

Specifically, the variance component score statistic is

$$Q = (\mathbf{y} - \hat{\boldsymbol{\mu}})'\mathbf{K}(\mathbf{y} - \hat{\boldsymbol{\mu}}), \qquad (3)$$

where $\mathbf{K}=\mathbf{GWG}'$, $\hat{\boldsymbol{\mu}}$ is the predicted mean of $\boldsymbol{y}$ under H_0, i.e., $\hat{\boldsymbol{\mu}} = \hat{\alpha}_0 + \mathbf{X}\hat{\boldsymbol{\alpha}}$ for continuous traits and $\hat{\boldsymbol{\mu}} = \operatorname{logit}^{-1}(\hat{\alpha}_0 + \mathbf{X}\hat{\boldsymbol{\alpha}})$ for dichotomous traits; and $\hat{\alpha}_0$ and $\hat{\boldsymbol{\alpha}}$ are estimated under the null by regressing $\mathbf{y}$ on only the covariates $\mathbf{X}$. Here $\mathbf{G}$ is an $n\times p$ matrix with the (i,j)-th element being the genotype of variant j of subject i, and $\mathbf{W}=\operatorname{diag}(w_1,\ldots,w_p)$ contains the weights of the p variants.

In fact, $\mathbf{K}$ is an $n\times n$ matrix with (i,i')-th element equal to $K(\mathbf{G}_i,\mathbf{G}_{i'}) = \sum_{j=1}^{p} w_j G_{ij} G_{i'j}$. $K(\cdot,\cdot)$ is called the kernel function and $K(\mathbf{G}_i,\mathbf{G}_{i'})$ measures the similarity between subjects i and

i' in the region using the p markers. This particular form of $K(\cdot,\cdot)$ is called the weighted linear kernel function. We later discuss other choices of the kernel to model epistatic effects.

Good choices of weights can improve power. Each weight w_j is pre-specified, using only the genotypes, covariates and external biological information, i.e. estimated without using the outcome, and reflects the relative contribution of the j-th variant to the score statistic: if w_j is close to zero, the j-th variant makes only a small contribution to Q. Thus, down-weighting non-causal variants and up-weighting causal variants can yield improved power. Since in practice we do not know which variants are causal, we propose to set $\sqrt{w_j} = Beta(MAF_j; a_1,a_2)$, the beta distribution density function with pre-specified parameters a_1 and a_2 evaluated at the sample MAF (across cases and controls combined) for the j-th variant in the data. The Beta density is flexible and can accommodate a broad range of scenarios. For example, if rarer variants are expected to be more likely to have larger effects, then setting $0<a_1\leq1$ and $a_2\geq1$ allows for up-weighting rarer variants and down-weighting common weights. We suggest setting $a_1= 1$ and $a_2= 25$ as it up-weights rare variants while still putting decent non-zero weights for variants with MAF 1-5%. All simulations were conducted using this default choice unless stated otherwise. Note that smaller a_1 results in stronger up-weighting of rarer variants. Examples of weights across a range of a_1 and a_2 values are presented in Fig. S1. Note that $a_1=a_2=1$ corresponds to $w_j=1$, i.e., all variants are weighted equally, and $a_1=a_2=0.5$ corresponds to $\sqrt{w_j} = 1/\sqrt{MAF_j(1-MAF_j)}$, i.e., w_j being the inverse of the variance of the genotype of marker j, which puts almost zero weight for MAFs >1% and can be used if one believes only variants with MAF<1% are likely to be causal. Note that SKAT calculated using this weight is identical to the unweighted SKAT test using the standardized genotypes G's in equations (1) and (2). Other forms of the weight as a function of MAF can also be used. Since SKAT is a score test, the type I error is protected for any choice of pre-chosen weights. Note that the weights used in the weighted sum test[13] involve phenotype information and will therefore alter the null distribution of SKAT.

Under the null hypothesis, Q follows a mixture of chi-square distributions, which can be closely approximated using the computationally efficient Davies method[26]. See the Appendix for details.

A special case of SKAT arises when the outcome is dichotomous, no covariates are included, and all $w_j = 1$. Under these conditions, we show in the Appendix that the SKAT test statistic Q is equivalent to the C-alpha test statistic T. Hence, the C-alpha test can be seen as a special case of SKAT, or alternatively, SKAT as a generalization of the C-alpha test that does not require permutation but calculates the p-value analytically, allows for covariate adjustment, and accommodates either dichotomous or continuous phenotypes. Since SKAT using flat weights is also equivalent to the kernel machine regression test[23-24] and since the kernel machine regression test is in turn related to the SSU test[27], it follows transitively that SKAT under flat weights, the kernel machine regression test, the SSU test, and the C-alpha test all equivalent and special cases of SKAT. Note that the null distribution is calculated differently using these methods, with SKAT giving more accurate analytic p-values especially in the extreme tail, when sample sizes are sufficient.

Relationship Between Linear SKAT and Individual Variant Test Statistics

The test statistics Q can be computed in an efficient manner by exploiting a close connection between the SKAT score test statistic Q and the individual variant test statistics. In particular, Q is a weighted sum of the individual score statistics for testing for individual variant effects. Hence, letting $\mathbf{g}_j=[G_{1j}, G_{1j}, \ldots, G_{nj}]$ denote the $n\times 1$ vector containing the genotypes of the n subjects for variant j, then it is straightforward to see that $Q = \sum_{j=1}^{p} w_j S_j^2$ where $S_j = \mathbf{g}_j'(\mathbf{y}-\hat{\boldsymbol{\mu}}_0)$ is the individual score statistic for testing the marginal effect of the j-th marker ($H_0 : \beta_j=0$) under the individual linear/logistic regression model of y_i on $\mathbf{X}_i$ and only the j-th variant G_{ij}:

$$y_i = \alpha_0 + \mathbf{X}_i'\boldsymbol{\alpha} + \beta_j G_{ij} + \varepsilon_i$$

for continuous phenotypes and

$$\operatorname{logit} P(y_i = 1) = \alpha_0 + \mathbf{X}_i'\boldsymbol{\alpha} + \beta_j G_{ij}$$

for dichotomous phenotypes. $\hat{\boldsymbol{\mu}}_0$ is estimated as $\hat{\boldsymbol{\mu}}_0 = \hat{\alpha}_0 + \mathbf{X}_i'\hat{\boldsymbol{\alpha}}$ for continuous traits and $\hat{\boldsymbol{\mu}}_0 = \operatorname{logit}^{-1}(\hat{\alpha}_0 + \mathbf{X}_i'\hat{\boldsymbol{\alpha}})$ for dichotomous traits. As a score test, one needs fit the null model only a single time to be able to compute the S_j for all individual variants j as well as all regions to be tested. Similarly, if multiple regions are under consideration, then the same $\hat{\boldsymbol{\mu}}_0$ can be used to compute the SKAT Q statistics for each region.

Accommodating Epistatic Effects and Prior Information under the SKAT

An attractive feature of SKAT is the ability to model the epistatic effects of sequence variants on the phenotype within the flexible kernel machine regression framework[28-30]. To do so, we replace $\mathbf{G}_i'\boldsymbol{\beta}$ by a more flexible function $f(\mathbf{G}_i)$ in the linear and logistic models (1) and (2) where $f(\mathbf{G}_i)$ allows for rare variant by rare variant and common variant by rare variant interactions. Specifically, for continuous traits we use the semiparametric linear model[23; 29]

$$y_i = \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_i + f(\mathbf{G}_i) + \varepsilon_i \qquad (4)$$

and for dichotomous traits, we use the semiparametric logistic model[24; 30]

$$\text{logit}\, P(y_i = 1) = \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_i + f(\mathbf{G}_i) \qquad (5)$$

Here the variants, $\mathbf{G}_i$, are related to the phenotype through a parametric/non-parametric function $f(\cdot)$, which is assumed to lie in a functional space generated by a positive semi-definite kernel function $K(\cdot,\cdot)$. Models (1) and (2) assume linear genetic effects and are specified by $K(\mathbf{G}_i, \mathbf{G}_{i'}) = \sum_{j=1}^{p} w_j G_{ij} G_{i'j}$. By changing $K(\cdot,\cdot)$, one can allow for more complex models. Intuitively, $K(\mathbf{G}_i, \mathbf{G}_{i'})$ is a function that measures genetic similarity between the i-th and i'-th subjects using the p variants in the region, and any positive semi-definite function $K(\mathbf{G}_i, \mathbf{G}_{i'})$ may be used as a kernel function. We tailored several useful and commonly used kernels specifically for the purpose of rare variant analysis: the weighted linear kernel, the weighted quadratic kernel, and the weighted IBS kernel.

The weighted linear kernel function $K(\mathbf{G}_i, \mathbf{G}_{i'}) = \sum_{j=1}^{p} w_j G_{ij} G_{i'j}$ implies that the trait depends on the variants in a linear fashion and is equivalent to the classical linear and logistic model presented in (1) and (2). The weighted quadratic kernel $K(\mathbf{G}_i, \mathbf{G}_{i'}) = \left(1 + \sum_{j=1}^{p} w_j G_{ij} G_{i'j}\right)^2$ implicitly assumes that the model depends on the main effects and quadratic terms for the gene variants, and the first order variant by variant interactions. The weighted IBS Kernel $K(\mathbf{G}_i, \mathbf{G}_{i'})$ $= \sum_{j=1}^{p} w_j IBS(G_{ij}, G_{i'j})$ or equivalently for additively coded autosomal genotype data $K(\mathbf{G}_i, \mathbf{G}_{i'})$

= $\sum_{j=1}^{p} w_j \left(2 - \left|\mathbf{G}_{ij} - \mathbf{G}_{i'j}\right|\right)$ defines similarity between individuals as the number of alleles shared identical by state. The model implied by the weighted IBS kernel model the SNP effects nonparametrically[31]. Consequently, this allows for epistatic effects because the function $f(\cdot)$ does not assume linearity or interactions of a particular order (e.g. second order), Using the weighted IBS kernel removes the assumption of additivity since the number of alleles shared IBS is a physical quantity that does not change based on different genotype encodings..

We note that a kernel function that better captures both the similarity between individuals and the causal variant effects will increase power. In particular, if relationships are linear and no interactions are present, then the weighted linear kernel will have highest power. If interactions are present, the weighted quadratic and weighted IBS kernels can increase power. Our experience suggests using the IBS kernel when the number of interacting variants within the region is modest. As our understanding of genetic architecture improves so too will our knowledge of which kernel to use.

In each of the above kernels, w_j is an allele specific weight that controls the relative importance of the j^{th} variant and might be a function of factors such as allele frequency or anticipated functionality. Without prior information, we suggest the use of the $\sqrt{w_j}$ = *Beta(MAF*$_j$*; 1, 25)* suggested earlier. However, if prior information is available, e.g. some variants are predicted as functional or damaging via Polyphen[32] or Sift[33], weights may be selected to up-weight for likely functionality.

To test for the effects of gene variants in a region on a phenotype, one tests the null hypothesis H_0: $f(\mathbf{G}) = 0$. SKAT tests for this null hypothesis by assuming the $n\times 1$ vector $\mathbf{f}$ = $[f(\mathbf{G_1}), \ldots, f(\boldsymbol{G}_n)]'$ for the genetic effects of n subjects follows a distribution with mean zero and covariance $\tau\mathbf{K}$, where τ is a variance component that indexes the effects of the variants[29-30]. Hence, to test the null that corresponds to testing H_0: $\tau=0$ by a variance component score test. In particular, we simply replace **K** in equation (3) using the **K** discussed in this section, e.g., the weighted IBS kernel, for epistatic effect. All subsequent calculations for computing a *p*-value remain the same.

Since the SKAT evaluates significance using a score test, which operates under the null hypothesis, the SKAT is valid (in terms of protecting type I error) irrespective of the kernel and the weights used. Good choices of the kernel and the weights simply increase power.

Planning New Sequencing-based Association Studies: Power and Sample Size Estimation

Power and sample size calculations are important in designing sequencing studies of complex traits. Using a modification of the higher order moment approximation method[34], we provide an analytic method to carry out efficiently such calculations for SKAT[35]. Specifically, for a fixed sample size and α-level, given a prior hypothesis on the genetic architecture of a particular region, effect size, and proportion/number of causal variants within a region, our method provides the power to detect the region as significant using SKAT. Similarly, if the desired power is fixed, the approach can be used to find the necessary sample size.

There are key differences between the power and sample size estimation for single-variant and region (set)-based tests. For a region (set)-based test, power depends strongly on the underlying genetic architecture and its estimation requires modeling this genetic architecture and the LD between variants. Therefore, to estimate power to detect a particular region as associated with a phenotype requires specification of the significance level, sample size, which variants in the region are causal with corresponding effect size, and the LD structure of the variants in the region. Ideally, one may use prior data to assess the LD and MAF. Since prior data may be difficult to obtain, we currently recommend the use of either 1000 Genomes project data[36] or data simulated under a population genetics model [37]. Relevant preliminary data will become increasingly available as sequencing studies become more common.

Our SKAT software uses simulated data based on the coalescent population genetic model (released with the software package) as a default in performing sample size and power calculations, and instead of directly specifying the effects of any given variant, the user can input an MAF threshold for determining which variants are regarded as rare and also a proportion determining how many of the rare variants are causal. The causal variants are then randomly selected from the alleles with true MAF (based on simulated or preliminary data) less than the threshold. The magnitudes of the effects for causal variants are set to be equal to $c \times |\log_{10} \mathrm{MAF}|$ where the c is determined based on the maximum effect size the user would like to allow (described below in the power simulations section) at $\mathrm{MAF}=10^{-4}$. Since these parameters can be difficult to choose as *a priori*, power and sample size can be reasonably estimated by averaging results over a range of parameter values. Similarly, since the region architecture can vary across different regions, for genome-wide studies, one can average over multiple randomly selected regions as currently implemented in the SKAT software.

Numerical Experiments and Simulations

To validate SKAT in terms of protecting type I error, to assess its power compared to burden tests, and to assess the accuracy of our power and sample size tools, we carried out simulation studies under a range of configurations. For all simulations, sequence genotypes were determined by simulating 10,000 chromosomes for a 1Mb region based on a coalescent model that mimics the LD pattern local recombination rate, and population history for Europeans using COSI[37].

Type I Error Simulations

To investigate whether SKAT preserves the desired type I error rate at the near genome-wide threshold level, e.g., $\alpha=10^{-6}$, it is necessary to conduct simulations with hundreds of millions of simulated datasets. Although SKAT is computationally efficient, generating such a large number of datasets is challenging. To reduce the computation burden, we took the following approach. Using 10,000 randomly selected sets of 30kb subregions within a 1 Mb chromosome, we first generated 10,000 sets of genotypes $\mathbf{G}_{(n\times p)}$ from the coalescent model, with p variants on n subjects. Then for each of the 10,000 simulated genotype data sets, we simulated 10,000 sets of continuous phenotypes such that we were able to obtain 10^8 individual genotype-phenotype data sets using the model:

$$y = 0.5X_1 + 0.5X_2 + \varepsilon$$

where X_1 is a continuous covariate generated from a standard normal distribution, X_2 is a dichotomous covariate taking values 0 and 1 with probability 0.5, and ε follows a standard normal distribution. Note that the continuous trait values are not related to the genotype, so that the null model holds. The 30kb regions that the genotype values are based on contained 605 variants on average, but the number of observed variants for any given data set was considerably less and depended on the sample size n, which we set to 500, 1000, 2500, and 5000.

We repeated the type I error simulations for dichotomous phenotypes as above, except the dichotomous outcomes were generated using the model:

$$\operatorname{logit} P(y=1) = \alpha_0$$

where α_0 was determined to set the prevalence to 1%.

For both continuous and dichotomous simulations, we applied SKAT using the default weighted linear kernel to each of the 10^8 data sets and estimated the empirical type I error rate as the proportion of p-values less than $\alpha=10^{-4}$, 10^{-5}, or 10^{-6}.

We note that the estimated type I error from this approach is not completely the same as the empirical type I error when genotypes are generated randomly for each simulation, since for each of the 10000 genotype data sets, only the outcomes are resampled. However, our type I error estimator is still unbiased and results in very accurate type I error estimates. For larger α-levels (0.05 and 0.01), we directly computed the empirical type I error rate using data sets in which genotypes were randomly generated for each simulation.

Empirical Power Simulations

We simulated data sets in which 30kb subregions were randomly selected from the generated 1 Mb chromosomes and used to create causal variants and a phenotype variable together with additional simulated covariates. We generated continuous phenotypes by

$$y = 0.5X_1 + 0.5X_2 + \beta_1 G_1^c + \beta_2 G_2^c + ... + \beta_p G_p^c + \varepsilon, \quad (1)$$

where X_1, X_2, and ε are as defined for the type I error simulations , $G_1^c, G_2^c, ..., G_s^c$ are the genotypes of the s causal rare variants (a randomly selected subset of the simulated rare variants, e.g., 5% of variants that have MAF<3% in Fig. 1), and the βs are effect sizes for the causal variants. Similarly, we generated dichotomous phenotypes for case-control data under the logistic model

$$\text{logit}\, P(y=1) = \alpha_0 + 0.5X_1 + 0.5X_2 + \beta_1 G_1^c + \beta_2 G_2^c + ... + \beta_p G_p^c \;, \quad (2)$$

where $G_1^c, G_2^c, ..., G_p^c$ are again the genotypes for the causal rare variants and βs are log ORs for the causal variants. We controlled prevalence by α_0 and set to it 1% unless otherwise stated. Under both models, we set the magnitude of each β_j to $c\left|\log_{10} MAF_j\right|$ such that rarer variants had larger effects. In the simulation studies, for continuous traits, c=0.4, which gives the "maximum" effect size $|\beta_j|$=1.6 for variants with MAF=10^{-4} and small effects $|\beta_j|$=0.28 for MAF=0.2. For dichotomous traits, c=ln5/4=0.402, which gives the "maximum" OR=5.0 ($|\beta_j|$= ln5) for variants with MAF=10^{-4} and smaller OR=1.32 for MAF=0.2. The effect size curves are given in Fig. S2.

We compared SKAT, an unsupervised variation on the weighted sum test (WST)[13] which uses weighted count based collapsing, counting based collapsing[18], and CAST[14]. For each of these tests, we considered variants with observed MAF <3% as rare: CAST collapsed based on whether or not an individual exhibits any variants with allele frequency <3%, the counting method counts the number variants with MAF<3%, and the weighted count inflates the contribution of each rare variant by multiplying the genotype with the same beta density based weights as used in SKAT.

To accommodate missing genotypes commonly observed in sequence data, we consider the effect of imputing missing values by randomly setting 10% of the genotypes as missing, imputing genotypes based on observed allele frequencies and Hardy-Weinberg Equilibrium , and then applying SKAT to the imputed data. We also performed restricted SKAT (rSKAT) by applying un-weighted SKAT to rare variants with MAF<3%. Note that for dichotomous phenotypes, rSKAT is essentially equivalent to a covariate adjusted C-alpha test with the p-value calculated analytically instead of using permutation. For each of the methods, power was estimated as the proportion of p-values < α, where $\alpha = 10^{-6}$ to mimic genome-wide studies.

Power and Sample Size Formulae

To demonstrate the utility and accuracy of our power and sample size calculation method, we conducted several numerical experiments. We first illustrated the use of the methods by computing the sample size necessary to detect a 30kb region with 5% of the variants with MAF<3% being causal. We assume effect size (OR) increases with decreasing MAF, and seek 80% power at significance levels $\alpha=10^{-6}$, 10^{-3}, 10^{-2}, corresponding to approximate genome-wide sequencing significance and candidate gene sequencing studies of 50 and 5 genes, respectively. We considered both continuous and dichotomous traits.

To show that the power estimated from our sample size formula is accurate, we compared empirical power for SKAT under simulations to power estimated via our analytic method . Specifically, we simulated continuous and case-control data under the same setting as used in the power simulations and estimated power as a function of the sample size by computing the proportion of p-values < $\alpha=10^{-6}$ and compared the empirical power curve to the power estimated using our analytical method.

RESULTS

Type I Error Simulation Results

The empirical type I error rates estimated for SKAT are presented in Table 1 for $\alpha=10^{-4}$, 10^{-5}, 10^{-6}, and suggest the type I error rate is protected for continuous phenotypes, though for smaller sample sizes the SKAT can be slightly conservative. For dichotomous phenotypes, SKAT is conservative for smaller sample sizes and very small alpha-levels. Additional type I error simulation results for SKAT and the competing methods are presented in Supplementary Fig. S3 for both continuous traits and dichotomous traits, and show that at larger α levels, all of the considered tests correctly control at the $\alpha=0.05$ and 0.01 levels. These results show that SKAT is a valid method and despite being conservative at low α-levels, SKAT maintains good power relative to existing methods (see below). However, if sample sizes are small or sharp control of type I error is necessary, then standard permutation-based procedures can be used to generate a Monte Carlo p-value for significance, though this can be computationally expensive and does not work in the presence of covariates, such as controlling for population stratification and require carful modifications.

Statistical Power of the SKAT and Competing Methods

We compared the power of SKAT with three burden tests in a series of simulation studies for both continuous traits and dichotomous traits by generating sequence data in randomly selected 30kb regions using a coalescent model[37]. For our primary power simulation, within each region, 5% of variants with population minor allele frequency (MAF)<3% were randomly chosen as causal, and the effect size of causal variants was a decreasing function of MAF with 50-100% of the causal variants being positively associated with the trait (See **Materials and Methods** and Fig. S2). The simulated regions for our power analysis contained on average 605 variants (26 causal), of which, 530.9 (88%), 502.9 (83%), and 422.8 (70%) had population MAF <3%, <1%, and <0.1%, respectively. The average allele frequency spectrum across the samples is similar to that of the Dallas Heart Study data (Fig. S4). Since the majority of variants have low MAF, they may not be observed in any particular sample. The average number of observed variants (assuming no genotyping error) and the average number of observed causal variants are presented in Table 2.

For continuous traits, SKAT had much higher power than all the burden tests, with the weighted count method tending to outperform the count and CAST methods (Fig. 1). SKAT's power was robust to the proportion of causal variants that were positively associated with the trait, while the burden tests suffered substantial loss of power when causal variants had opposite effects. The simulation results examining dichotomous traits were qualitatively similar in that SKAT dominated the competing methods. However, here the power of the SKAT decreased when both protective and harmful variants were present, although less so than for the burden tests. The difference in power for SKAT for different proportions of protective variants is due to the fact that given fixed population MAFs, protective variants imply negative log ORs and lower disease risk and hence lower MAFs in cases and more difficulties in observing rare variants in cases. The larger decrease in power for the competing methods is additionally driven by sensitivity to direction of effect due to aggregation of genotypes. Across all configurations, using imputed genotypes instead of the true genotype for 10% missing genotype data led to a very small reduction in power, despite use of a very simple Hardy-Weinberg-based imputation strategy. This is true in part because most variants are rare.

Note that SKAT up-weights rare variants but does not require thresholding. To show that the superior performance of SKAT is intrinsic and is not driven by the particular choice of the weight used, we calculated rSKAT, which does not weight the rare variants but instead uses the same threshold as the burden tests. Our results presented in Fig. 1 show that rSKAT is still substantially more powerful than all three burden tests.

Power simulation results for other type I error rates (α=0.01, 0.001), lower causal variant frequencies (population MAF<1%), and other region sizes (10kb and 60kb) yielded the same conclusions (Supplementary Figs. S5-8).

In the 30kb genomic regions considered, reflecting analysis of genome wide sequencing data, it is unlikely that a large proportion of the rare variants are all causal. However, for exome scale sequencing, the number of observed rare variants can be considerably smaller and the proportion of causal rare variants may be greater. Hence, we also conducted power simulations for smaller region sizes (3kb and 5kb) and larger proportions of causal variants (10, 20, and 50%). Results for both continuous and dichotomous phenotypes are presented in Supplementary Figs. S9-12 and show that if 50% of the rare variants are causal and that all of the causal variants have effects in the same direction, then SKAT and rSKAT are less powerful compared to collapsing methods, with count based collapsing having the greatest power. This result held for both 3kb and 5kb regions and is expected since the collapsing methods implicitly assume that all of the variants are causal and have unidirectional effects. In all other settings we considered, SKAT was the most powerful method.

Power and Sample Size Estimation Results

To illustrate our power and sample size calculation method, in Fig. 2 we present the estimated sample size curves as a function a of maximum effect sizes (ORs for dichotomous traits) necessary to detect a 30kb region with 5% of the variants with MAF<3% being causal. Table 3 presents estimated sample sizes for several configurations of practical interest. Additional sample size curves when causal variants are rarer (MAF<1%) or occur more frequently (10% of variants are causal) or when prevalence is varied (5%, 0.1%) can be found in Figs. S13-15. These results show that, for a given region, one will have higher power (and lower required sample size) to detect rare causal variants if the percentage of variants that are causal is higher, the causal rare variants have higher MAFs and/or larger effect sizes (e.g. odds ratios (ORs)), and the effects are more consistently in the same direction. For case-control designs, lower prevalence yields higher power since given the same OR and population MAF, the lower prevalence results in enrichment of more harmful (ORs > 1) variants, i.e., higher MAFs, across both cases and controls, i.e. for rarer diseases, harmful rare variants are more likely to be observed. Conversely, if the prevalence is low, fewer protective variants (ORs < 1), i.e., lower MAFs, are likely to be observed in the sample.

We also compared the power/sample size formulae estimates to the empirical, simulation based power estimates for both continuous and dichotomous traits. The curves plotted in Fig. 3 show that the empirical power is accurately approximated by our analytical formula.

Application to Dallas Heart Study Data

We analyzed sequence data on 93 variants in *ANGPTL3* (MIM 604774), *ANGPTL4* (MIM 605910), and *ANGPTL5* (MIM 607666) in 3,476 individuals from the Dallas Heart Study[38] to test for association between log-transformed serum triglyceride (logTG) levels and rare variants in these genes. We adjusted for sex and ethnicity (Black, Hispanic, or White), but did not adjust for age as a large number of subjects have missing ages. In addition to testing for association using SKAT and the three burden tests considered earlier, we also applied the permutation-based varying threshold method (VT) and the Polyphen-score adjusted VT (VTP)[16], which are based on the residuals obtained from regressing the phenotype on the covariates, and assume gene-covariate independence. Since VT and VTP require permutation, they are computationally expensive when applied genome wide. For VTP, we used the Polyphen score for rare variants (MAF<0.01) and assigned a constant score=0.5 to all other variants. We also analyzed a dichotomized phenotype on the highest and lowest quartiles of each of the six sex-ethnicity groups (Table 4).

SKAT was by far the most powerful test for the dichotomous trait. For continuous traits, SKAT has much smaller p-values than two burden methods (CAST and WST) and VT, and has a slightly higher p-value than the counting-based burden test (N) and VTP. Note that SKAT was easier to apply, since it did not require prior functional information (available for only a subset of variants) or permutation, and adjusted for covariates without assuming gene-covariate independence.

Computation time

The computation time for the SKAT depends on the sample size and the number of markers. To analyze a 30kb region sequenced on 1000, 2500, or 5000 individuals, SKAT required 0.21, 0.73, and 2.3 sec for continuous traits and ~20% longer for dichotomous traits, on a 2.33 GHz Laptop with 6Gb memory. To analyze 300kb, 3Mb, or 3Gb (the entire genome) on 1000 individuals requires 2.5s, 25s, and 7h.

DISCUSSION

We propose SKAT as a supervised, flexible, and computationally efficient statistical method to test for association between a phenotype, continuous or dichotomous, and rare and common genetic variants in sequencing-based association studies. We demonstrate that SKAT's power is greater than that of several burden tests over a range of genetic models. Furthermore, we have developed analytical power and sample size calculations for SKAT to assist in designing sequencing-based association studies.

Like burden tests, SKAT performs region based testing. However, SKAT has several major advantages over the existing tests. As a supervised method, SKAT directly performs multiple regression of a phenotype on genotypes for all variants in the region, adjusting for covariates. Hence, as with conventional multiple regression models, neither directionality nor magnitudes of the associations are assumed *a priori* but are instead estimated from the data. To test efficiently for the joint effects of rare variants in the region on the phenotype, SKAT assumes a distribution for the regression coefficients of the markers, whose variances depend on flexible weights. SKAT performs a score-based variance component test, whose calculation only requires fitting the null model by regressing phenotypes on covariates alone and computing p-values analytically. The flexible regression framework also allows us to allow for epistatic effects**.**

Besides region-based analysis, SKAT can also be applied to any biologically meaningful SNP set. As SKAT is a regression based method, it can be easily extended to survival, longitudinal, and multivariate phenotypes, and hence provides a comprehensive framework for a wide variety of sequencing-based association studies.

The ability to obtain a p-value directly without the need for permutation is an attractive feature of SKAT, and allows for rapid estimation of p-values in exome and genome wide sequencing studies. Our simulations showed that for continuous phenotype, the p-values are accurate when the sample size is moderate or large; for dichotomous phenotypes, the p-values are conservative at lower α-levels (e.g. $<10^{-4}$) if the sample size is modest or small. Permutation can be used to obtain a more accurate estimate in the absence of covariates. In the presence of covariates, e.g., population stratification, standard permutations fail and require careful modifications. Despite the conservative nature of the score test, SKAT often still has higher power than competing methods at small α-levels.

SKAT can be combined with collapsing strategies to form a hybrid testing approach. If most of the variants within a range of allele frequencies are causal and have the same directionality (i.e. under settings that are optimal for burden based tests), collapsing these variants and then applying SKAT to the collapsed variants can improve power. For example, since singletons are common in sequencing studies (57 of 93 variants in the Dallas Heart Study data), a possible hybrid strategy is to first collapse all of the singletons into a single value and then apply SKAT to the collapsed value and the other 36 variants. Compared to the original SKAT, this strategy gives a slightly lower p-value 3.1×10^{-5} for the continuous trait and a slight higher p-value 1.6x10^{-4} for the dichotomous trait. Simulation studies showed that the two methods are of similar power under the settings used to generate Fig. 1.

An important feature of SKAT is that it allows for incorporation of flexible weight functions to boost analysis power, e.g. by up-weighting variants with lower MAFs and down-weighting information from variants inferred with lower confidence. Good choices of weights are likely to improve the power of the association test using SKAT, although simulations show that even equal weights can yield high power when combined with thresholding. In our simulation studies, we employed a class of flexible continuous weights as a function of MAF using the Beta function, which up-weights rare variants and does not require thresholding. Users can define other types of weight functions. To further improve analysis power, one can estimate weights by incorporating information besides MAF, e.g. using Polyphen score or integrating other annotation information, which will become increasingly available as our understanding of genome variation improves. Therefore, because of its flexibility, SKAT has the capacity to mature, and its power to increase, as the field progresses.

APPENDIX

Estimating the Null Distribution for Q.

Under the null hypothesis, Q follows a mixture of chi-square distributions[29-30]. More specifically, we define $\mathbf{P}_0 = \mathbf{V} - \mathbf{V}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\mathbf{V}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{V}$ where $\tilde{\mathbf{X}}$ is the $n\times(p+1)$ matrix equal to [**1**, **X**]. For continuous phenotypes, $\mathbf{V} = \sigma_0^2\mathbf{I}$ where σ_0 is the estimator of σ under the null model where $f(\mathbf{G}) = 0$, and I is an $n\times n$ identity matrix. For dichotomous phenotypes, $V = diag\left(\hat{\mu}_{01}(1-\hat{\mu}_{01}), \hat{\mu}_{02}(1-\hat{\mu}_{02}), \ldots, \hat{\mu}_{0n}(1-\hat{\mu}_{0n})\right)$ where $\hat{\mu}_{0i} = \operatorname{logit}^{-1}(\hat{\alpha} + \hat{\boldsymbol{\alpha}}'\mathbf{X}_i)$ is the estimated probability that the i-th subject is a case under the null model. Then under the null,

$$Q \sim \sum_{i=1}^{n} \lambda_i \chi_{1,i}^2 \qquad (6)$$

where ($\lambda_1, \lambda_2, \ldots, \lambda_n$) are the eigenvalues of $\mathbf{P}_0^{1/2}\mathbf{K}\mathbf{P}_0^{1/2}$ and $\chi_{1,i}^2$ are independent χ_1^2 random variables.

Several approximation and exact methods have been suggested to obtain the distribution of Q[39]. Among these, the Davies exact method[26], based on inverting the characteristic function of (6), appears to work well in practice and is used here.

SKAT is a Generalization of the C-alpha test

The recently proposed the C-alpha test has advantages over burden tests in that it explicitly models the possibility that minor alleles can be deleterious or protective. However, it does not currently allow for the analysis of quantitative outcomes or the inclusion of covariates and p-value calculation requires permutation. We demonstrate that for a dichotomous trait in the absence of covariates, the C-alpha test statistic is equivalent to the SKAT statistic with un-weighted linear kernel.

Suppose the *j*-th variant is observed d_j times in the cases, out of n_j times total in cases and controls, and that $p_0 = \sum_{i=1}^{n} y_i / n$. For a dichotomous trait and no covariates, the C-alpha test statistic

$$T_\alpha = \sum_{j=1}^{p} \left[\left(d_j - n_j p_0 \right)^2 - n_j p_0 \left(1 - p_0 \right) \right] \tag{7}$$

Denote $T_\alpha^1 = \sum_{j=1}^{p} \left(d_j - n_j p_0 \right)^2$. Since $\sum_{j=1}^{p} n_j p_0 \left(1 - p_0 \right)$ is the mean of T_α under the null hypothesis of no association, T_α^1 is the C-alpha test statistic without mean centering. Since $d_j = \mathbf{y}'\mathbf{G}._j$ and $n_j = \mathbf{J}'\mathbf{G}._j$, where $\mathbf{G}._j$ is the *j*-th column of the genotype matrix $\mathbf{G}$ and $\mathbf{J} = (1, 1, \ldots, 1)'$, it can be easily shown that

$$T_\alpha^1 = (\mathbf{y} - p_0 \mathbf{J})'\mathbf{G}\mathbf{G}'(\mathbf{y} - p_0 \mathbf{J}). \tag{8}$$

Note that under the un-weighted linear kernel, $\mathbf{K} = \mathbf{G}\mathbf{G}'$ and $\hat{\boldsymbol{\mu}}_0 = p_0 \mathbf{J}$ if no covariates are present. Hence, equation (8) is identical to equation (3), i.e. T_α^1 is equivalent to the SKAT test statistic with un-weighted linear kernel.

Although the SKAT statistic with unweighted linear kernel and the C-alpha test statistic are equivalent, SKAT and C-alpha test use different null distributions to assess significance: C-alpha test uses a normal approximation, while we use a mixture of chi-squares. The normal approximation gives a valid p-value when the tested rare variants are independent and sample sizes are large, and so requires an assumption of linkage equilibrium. In the presence of LD, permutation is used by the C-alpha test for significance testing. One can easily see that the test statistic takes a quadratic form of ***y***, which follows a mixture of chi-square distributions. SKAT approximates this distribution directly using the Davies method and hence gives accurate estimation of significance regardless of the LD structure when sample size is sufficient.

WEB RESOURCES

The URLs for data presented herein are as follows:

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

An implementation of SKAT and power/sample size calculations in the R language can be found at http://www.hsph.harvard.edu/~xlin/software.html.

SUPPLEMENTAL DATA

Supplemental Data includes 15 figures and can be found with this article online at http://www.cell.com/AJHG.

ACKNOWLEDGEMENTS

This work was supported by grants P30 ES010126 (M.C.W.), DMS 0854970 and R01 GM079330 (T.C.), R01 HG000376 (M.B.), and R37 CA076404 and P01 CA134294 (S.L. and X.L.). We thank Drs. Jonathan Cohen, Alkes Price, and Shamil Sunyaev for providing the Dallas Heart Study data, and Larisa Miropolsky for help with the software development.

REFERENCES

1. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106, 9362-9367.
2. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376-380.
3. Mardis, E.R. (2008). Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9, 387-402.
4. Ansorge, W.J. (2009). Next-generation DNA sequencing techniques. New Biotechnology 25, 195-203.
5. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S., Moore, J., and Nadeau, J. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews Genetics 11, 446-450.
6. Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M., et al. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456, 66-72.

7. Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research 18, 1851-1858.
8. Li, R.Q., Li, Y.R., Fang, X.D., Yang, H.M., Wang, J., and Kristiansen, K. (2009). SNP detection for massively parallel whole-genome resequencing. Genome Research 19, 1124-1132.
9. Bansal, V., Harismendy, O., Tewhey, R., Murray, S.S., Schork, N.J., Topol, E.J., and Frazer, K.A. (2010). Accurate detection and genotyping of SNPs utilizing population sequencing data. Genome Research 20, 537-545.
10. Carvajal-Carmona, L.G. (2010). Challenges in the identification and use of rare disease-associated predisposition variants. Current Opinion in Genetics & Development 20, 277-281.
11. Schork, N.J., Murray, S.S., Frazer, K.A., and Topol, E.J. (2009). Common vs. rare allele hypotheses for complex diseases. Current Opinion in Genetics & Development 19, 212-219.
12. Li, B., and Leal, S. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. The American Journal of Human Genetics 83, 311-321.
13. Madsen, B., and Browning, S. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genetics 5, e1000384.
14. Morgenthaler, S., and Thilly, W.G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). Mutation Research 615, 28-56.
15. Li, B., and Leal, S. (2009). Discovery of rare variants via sequencing: implications for the design of complex trait association studies. PLoS Genetics 5, e1000481.
16. Price, A., Kryukov, G., de Bakker, P., Purcell, S., Staples, J., Wei, L., and Sunyaev, S. (2010). Pooled association tests for rare variants in exon-resequencing studies. The American Journal of Human Genetics 86, 832-838.
17. Han, F., and Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. Human Heredity 70, 42-54.
18. Morris, A., and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genetic Epidemiology 34, 188-193.
19. Zawistowski, M., Gopalakrishnan, S., Ding, J., Li, Y., Grimm, S., and Zollner, S. (2010). Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. The American Journal of Human Genetics 87, 604-617.
20. Asimit, J., and Zeggini, E. (2010). Rare variant association analysis methods for complex traits. Annual Review of Genetics 44, 293-308.
21. Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K., and Daly, M.J. (2011). Testing for an Unusual Distribution of Rare Variants. PLoS Genetics 7, e1001322.
22. Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38, 904-909.
23. Kwee, L., Liu, D., Lin, X., Ghosh, D., and Epstein, M. (2008). A powerful and flexible multilocus association test for quantitative traits. The American Journal of Human Genetics 82, 386-397.

24. Wu, M.C., Kraft, P., Epstein, M.P., Taylor, D.M., Chanock, S.J., Hunter, D.J., and Lin, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. The American Journal of Human Genetics 86, 929-942.
25. Lin, X. (1997). Variance component testing in generalised linear models with random effects. Biometrika 84, 309-326.
26. Davies, R. (1980). The distribution of a linear combination of chi-square random variables. Applied Statistics 29, 323-333.
27. Pan, W. (2009). Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol 33, 497-507.
28. Cristianini, N., and Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods.(Cambridge Univ Pr).
29. Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics 63, 1079-1088.
30. Liu, D., Ghosh, D., and Lin, X. (2008). 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 9, 292.
31. Fleuret, F., and Sahbi, H. (2003). Scale-invariance of support vector machines based on the triangular kernel. In 3rd International Workshop on Statistical and Computational Theories of Vision.
32. Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non synonymous SNPs: server and survey. Nucleic Acids Research 30, 3894-3900.
33. Kumar, P., Henikoff, S., and Ng, P. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols 4, 1073-1081.
34. Liu, H., Tang, Y., and Zhang, H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. Computational Statistics & Data Analysis 53, 853-856.
35. Lee, S., Wu, M.C., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Power and sample size calculations for designing rare variant sequencing association studies. In Harvard University Technical Report. (http://www.hsph.harvard.edu/~xlin).
36. Durbin, R.M., Abecasis, G.R., Altshuler, D.L., Auton, A., Brooks, L.D., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073.
37. Schaffner, S., Foo, C., Gabriel, S., Reich, D., Daly, M., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. Genome Research 15, 1576-1583.
38. Romeo, S., Yin, W., Kozlitina, J., Pennacchio, L.A., Boerwinkle, E., Hobbs, H.H., and Cohen, J.C. (2009). Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. The Journal of Clinical Investigation 119, 70.
39. Duchesne, P., and Lafaye De Micheaux, P. (2010). Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. Computational Statistics & Data Analysis 54, 858-862.

FIGURE LEGENDS

Fig. 1) Power comparisons comparing SKAT with burden tests based on simulation studies

Empirical power at $\alpha=10^{-6}$ assuming 5% of the rare variants with MAF<3% within random 30kb regions were causal. Top panel considers continuous phenotypes with maximum effect size ($|\beta|$) equal to 1.6 when MAF=10^{-4}; bottom panel considers case-control studies with maximum OR=5 when MAF=10^{-4}. Regression coefficients for the *s* causal variants were assumed to be a decreasing function of MAF as $\beta_j = c\log_{10}\left|MAF_j\right|$ *(j=1,…,p* (See Fig. S2)*),* where *c* was chosen to result in these maximum effect sizes. From left to right, the plots consider settings in which the coefficients for the causal rare variants are 100% positive/0% negative, 80% positive/20% negative, and 50% positive/50% negative. Total sample sizes considered are 500, 1000, 2500, and 5000, with half being cases in case-control studies. For each setting, six methods are compared: SKAT, SKAT in which 10% of the genotypes were set to missing and then imputed (SKAT_M), restricted SKAT (rSKAT) in which un-weighted SKAT is applied to variants with MAF<3%, the weighted sum burden test (W) with the same weights as used by SKAT, counting based burden test (N), and the CAST method (C). All the burden tests used MAF<3% as the threshold. For each method, power was estimated as the proportion of p-values<α among 1000 simulated data sets.

Fig. 2) Required sample sizes to reach 80% power

Required sample sizes estimated analytically to reach 80% power to detect rare variants associated with a continuous (top panel) or dichotomous phenotype in case-control studies (half are cases) (bottom panel) at the $\alpha=10^{-6}$, 10^{-3}, and 10^{-2} levels, assuming 5% of rare variants with MAF<3% within the 30kb regions are causal. Plots correspond to 100%, 80%, and 50% of the causal variants associated with increase in the continuous phenotype or risk of the dichotomous phenotype. Regression coefficients for the *s* causal variants were assumed to be the same decreasing function of MAF as that in Fig. 1. Required total sample sizes are plotted again the “maximum” effect sizes (ORs) when MAF=10^{-4}. Estimated total sample sizes were averaged over 100 random 30kb regions.

Fig. 3) Comparisons of the power based on simulation and analytic estimation

Power as a function of total sample size estimated by simulation using 1000 replicates and by the proposed power formula for continuous and dichotomous case-control traits. Simulation configurations correspond to those used in Fig. 1 in which 80% of the regression coefficients for the causal rare variants were positive.

Figure 1

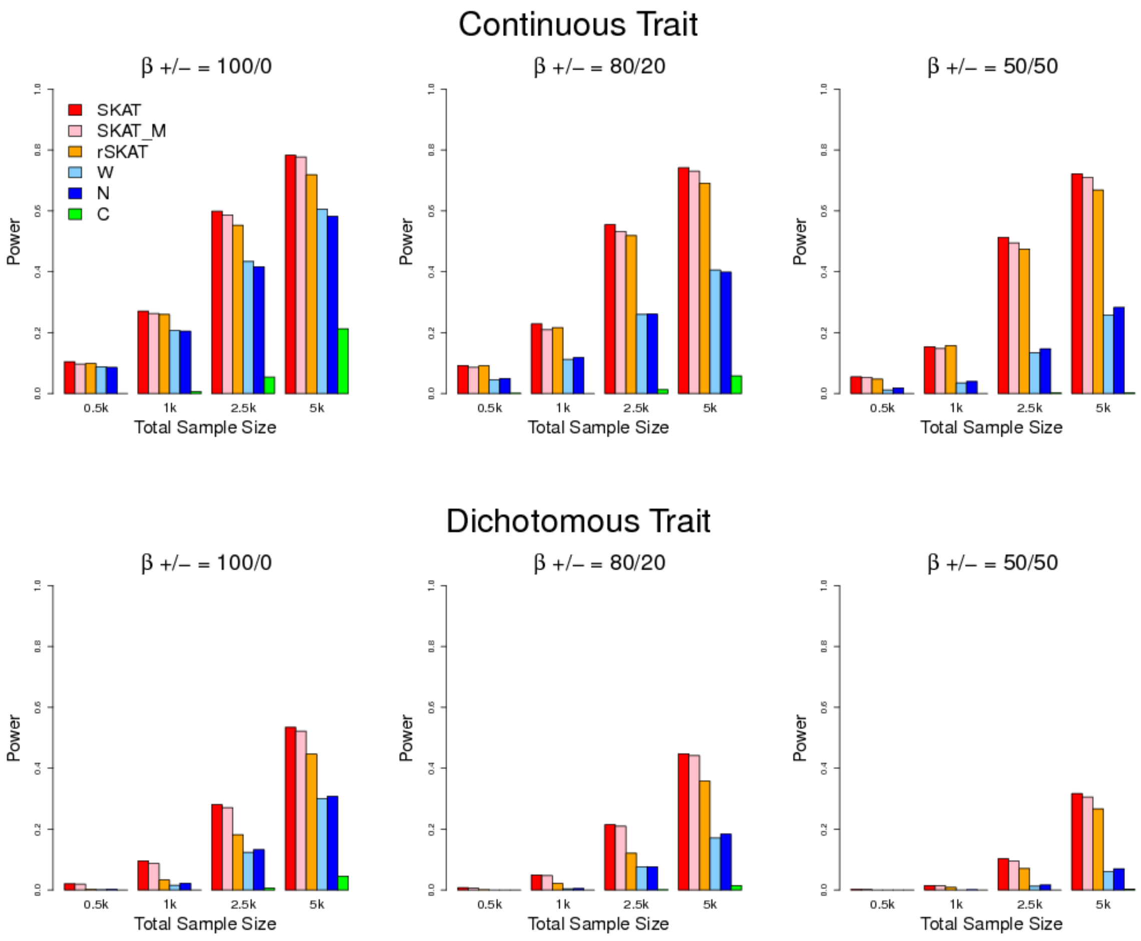

Continuous Trait
β +/– = 100/0
β +/– = 80/20
β +/– = 50/50
SKAT
SKAT_M
rSKAT
W
N
C
Power
0.0
0.2
0.4
0.6
0.8
1.0
0.5k
1k
2.5k
5k
Total Sample Size
Dichotomous Trait
β +/– = 100/0
β +/– = 80/20
β +/– = 50/50
Power
Total Sample Size

Figure 2

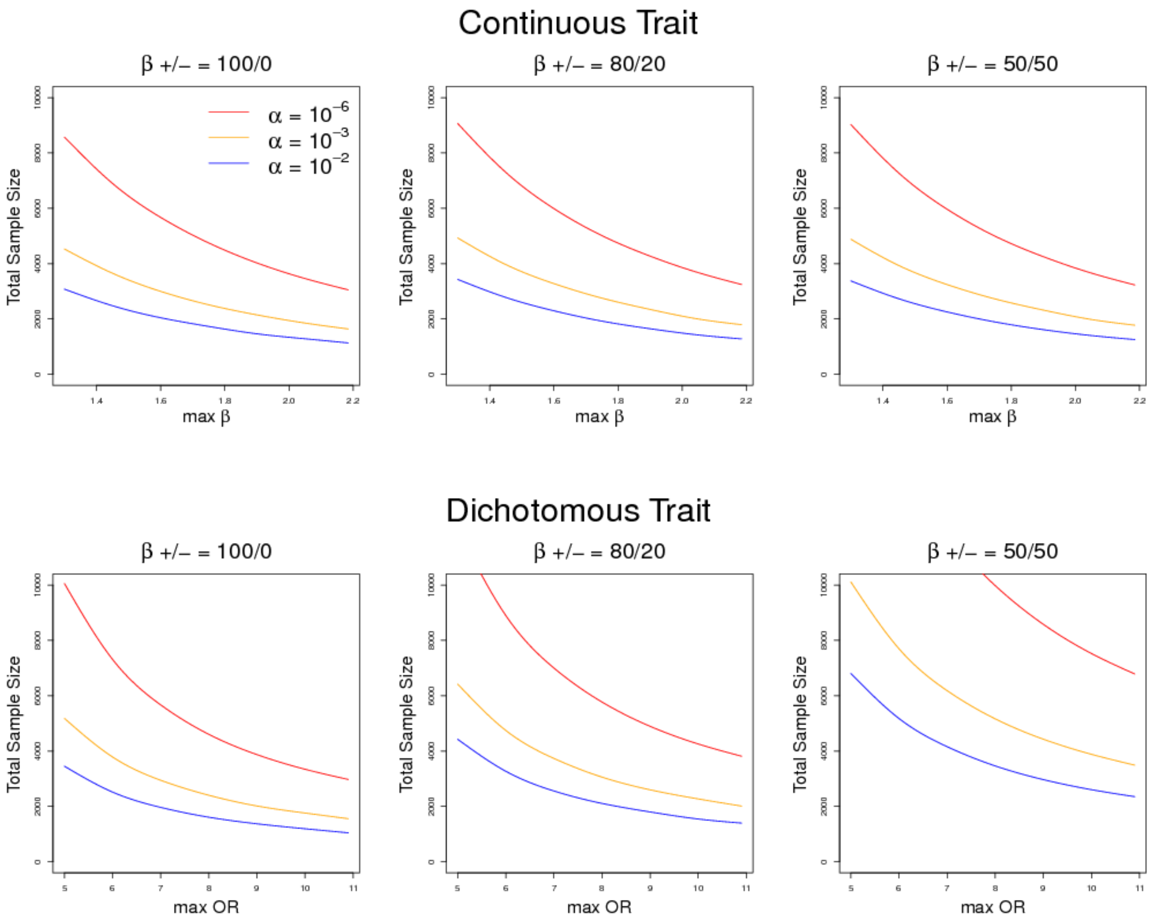

Continuous Trait
β +/− = 100/0
β +/− = 80/20
β +/− = 50/50
α = 10^−6
α = 10^−3
α = 10^−2
Total Sample Size
0
2000
4000
6000
8000
10000
1.4
1.6
1.8
2.0
2.2
max β
Dichotomous Trait
β +/− = 100/0
β +/− = 80/20
β +/− = 50/50
Total Sample Size
0
2000
4000
6000
8000
10000
5
6
7
8
9
10
11
max OR

Figure 3

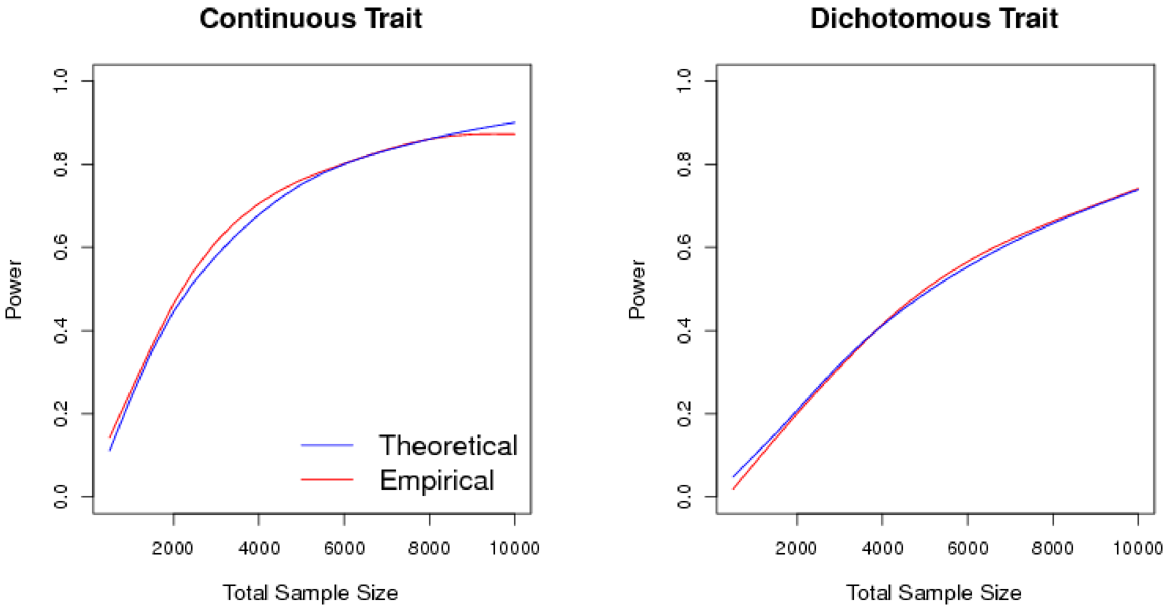

Continuous Trait
Dichotomous Trait
Power
0.0
0.2
0.4
0.6
0.8
1.0
2000
4000
6000
8000
10000
Total Sample Size
Theoretical
Empirical

Table 1 Type I error estimates of SKAT to test an association between randomly selected 30kb regions with a continuous trait at type I error rates as low as the genome-wide $\alpha=10^{-6}$ level. Each entry represents type I error rate estimates as the proportion of p-values α under the null hypothesis based on 10^8 simulated phenotypes.

	Continuous Phenotypes			Dichotomous Phenotypes		
Total Sample Size (*n*)	$\alpha=10^{-4}$	$\alpha=10^{-5}$	$\alpha=10^{-6}$	$\alpha=10^{-4}$	$\alpha=10^{-5}$	$\alpha=10^{-6}$
500	7.4×10^{-5}	6.5×10^{-6}	5.9×10^{-7}	2.2×10^{-5}	1.0×10^{-6}	1.0×10^{-8}
1000	8.5×10^{-5}	8.2×10^{-6}	8.0×10^{-7}	5.0×10^{-5}	3.5×10^{-6}	2.3×10^{-7}
2500	9.6×10^{-5}	9.1×10^{-6}	8.4×10^{-7}	7.6×10^{-5}	6.3×10^{-6}	5.6×10^{-7}
5000	9.8×10^{-5}	9.6×10^{-6}	8.8×10^{-7}	8.9×10^{-5}	7.8×10^{-6}	7.0×10^{-7}

Table 2 Characteristics of the 30kb region data sets used in the simulation studies. The number of observed variants and the number of observed causal variants within the region are averaged over the 1000 simulated data sets.

	Sample size (n)			
	500	1000	2500	5000
Average Number of Observed Variants				
All traits	255	330	438	512
Average Number of Observed Causal Variants				
Continuous Trait	9.6	13.3	18.6	22.3
Dichotomous Trait (β+/- = 100/0)	14.4	18.7	23.5	25.2
Dichotomous Trait (β+/- = 80/20)	13.3	17.1	22.0	24.3
Dichotomous Trait (β+/- = 50/50)	11.1	14.9	19.7	22.6

Table 3 Required total sample size to achieve 80% power to detect rare variants associated with a continuous or dichotomous case-control phenotype at the genome-wide level $\alpha = 10^{-6}$. Power was estimated using the analytical formulae assuming 5% or 10% of variants with MAF<3% are causal. Regression coefficients for the *s* causal variants were assumed to be a decreasing function of MAF, $\beta_j = c\log_{10}|MAF_j|$ *(j=1,…,s)* See Fig. S2. Required total sample sizes (cases+controls) are given for different "maximum" effect sizes (or ORs) when MAF=10^{-4} and different prevalences for case-control studies. Estimated sample sizes were averaged over 100 random 30kb regions.

Total Sample Size

	Max β=1.6/ Max OR=5		Max β=1.9/ Max OR=7	
	5% causal	10% causal	5% causal	10% causal
Continuous Trait	5,990	1,800	4,260	1,290
Dichotomous Trait with Prevalence 10%	15,120	4,810	9,650	3,120
Dichotomous Trait with Prevalence 1%	12,030	3,870	7,010	2,290

Table 4 Analysis of the Dallas Heart Study sequencing data using SKAT, the weighted sum burden test (W), the counting based burden test (N), the CAST method (C), the varying threshold method (VT), and the Polyphen-score adjusted VT (VTP) method. Beta(1,25) is used as the weight in the SKAT and the weighted sum test.

	SKAT	C	N	W	VT[a]	VTP[a]
Continuous TG Level	9.5×10^{-5}	1.9×10^{-3}	7.2×10^{-5}	2.3×10^{-4}	3.5×10^{-4}	2.0×10^{-5}
Dichotomized TG Level	1.3×10^{-4}	3.2×10^{-2}	2.2×10^{-3}	3.1×10^{-3}	8.6×10^{-3}	2.1×10^{-3}

[a] P-values estimated based on 10^6 permutations.

Figure S1) Examples of flexible weight functions used in SKAT.

Examples of the weight function used in SKAT, where the weight is assumed to be a function of MAF using the beta density function *Beta*(MAF, a_1,a_2): *Beta*(1, 1), *Beta*(1, 25), *Beta*(1, 50), and *Beta*(0.5, 0.5). (A) and (B) represent the same functions, while (B) zooms into the region with MAF <0.05.

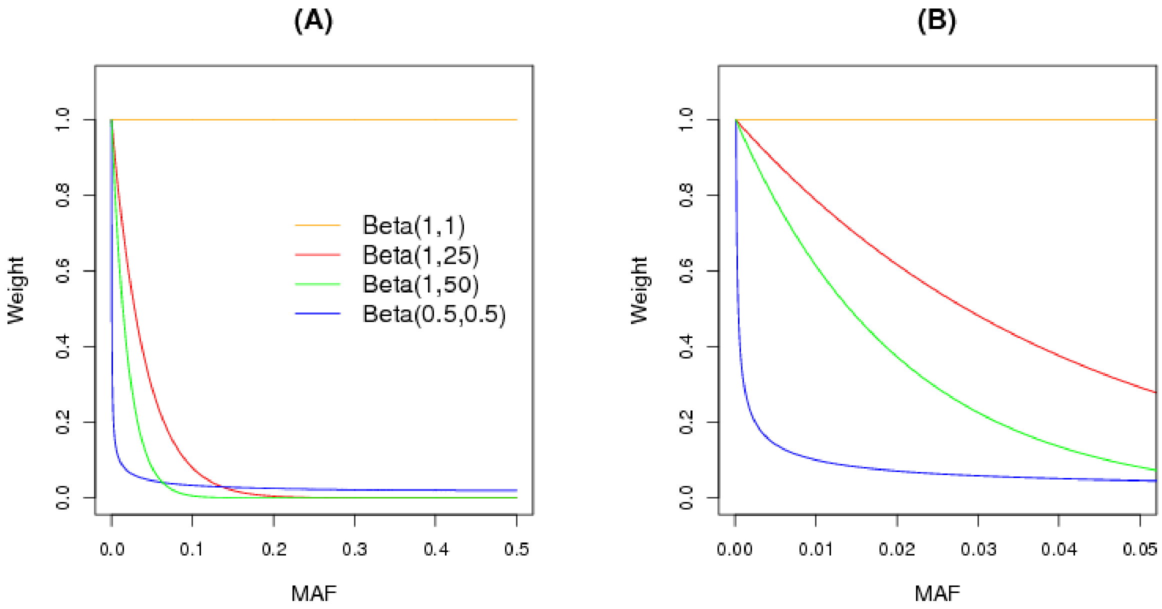


Figure S2) Effect size/Odds ratio functions used in the simulation studies.

Effect sizes for the causal variants in the continuous trait simulations (A) and odds ratios for the causal variants in case-control dichotomous trait simulations (B) as a function of population minor allele frequency. For continuous traits, the effect size is $|\beta|=c\left|\log_{10} MAF_j\right|$ with c = 0.4 (blue) and c=0.475 (red), which correspond to maximum effect size $|\beta_j|$=1.6 and 1.9, respectively, when MAF=10^{-4}. For dichotomous traits, log odds ratio is $|\beta|=c\left|\log_{10} MAF_j\right|$ with c=ln5/4 (blue) and c=ln7/4 (red), corresponding to maximum OR = 5 ($|\beta_j|$= ln5) and 7 ($|\beta_j|$= ln7), respectively, when MAF=10^{-4}.

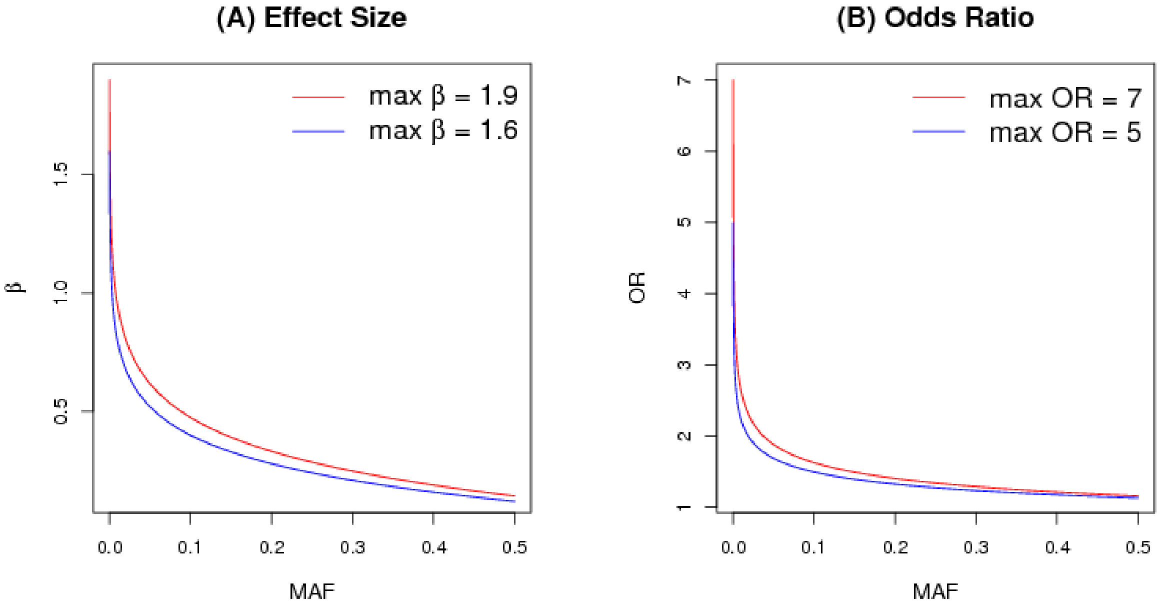


Figure S3) Type I error rates of SKAT and burden tests using simulation studies.

Empirical type I error rate to test an association between randomly selected 30kb regions with a continuous trait (top panel) or a case-control dichotomous trait (bottom panel) at the levels α=0.05 (left panel) and α=0.01 (right panel). Sample sizes considered are 500, 1000, 2500, and 5000. Each color represents a different method. For each method, type I error rate was estimated as the proportion of p-values < α under the null hypothesis using 10,000 simulated data sets. For $\alpha=10^{-4}$, $10^{-5,}$ $10^{-6,}$, the empirical type I error rates of SKAT are close to the nominal values and are given in Supplementary Table 1.

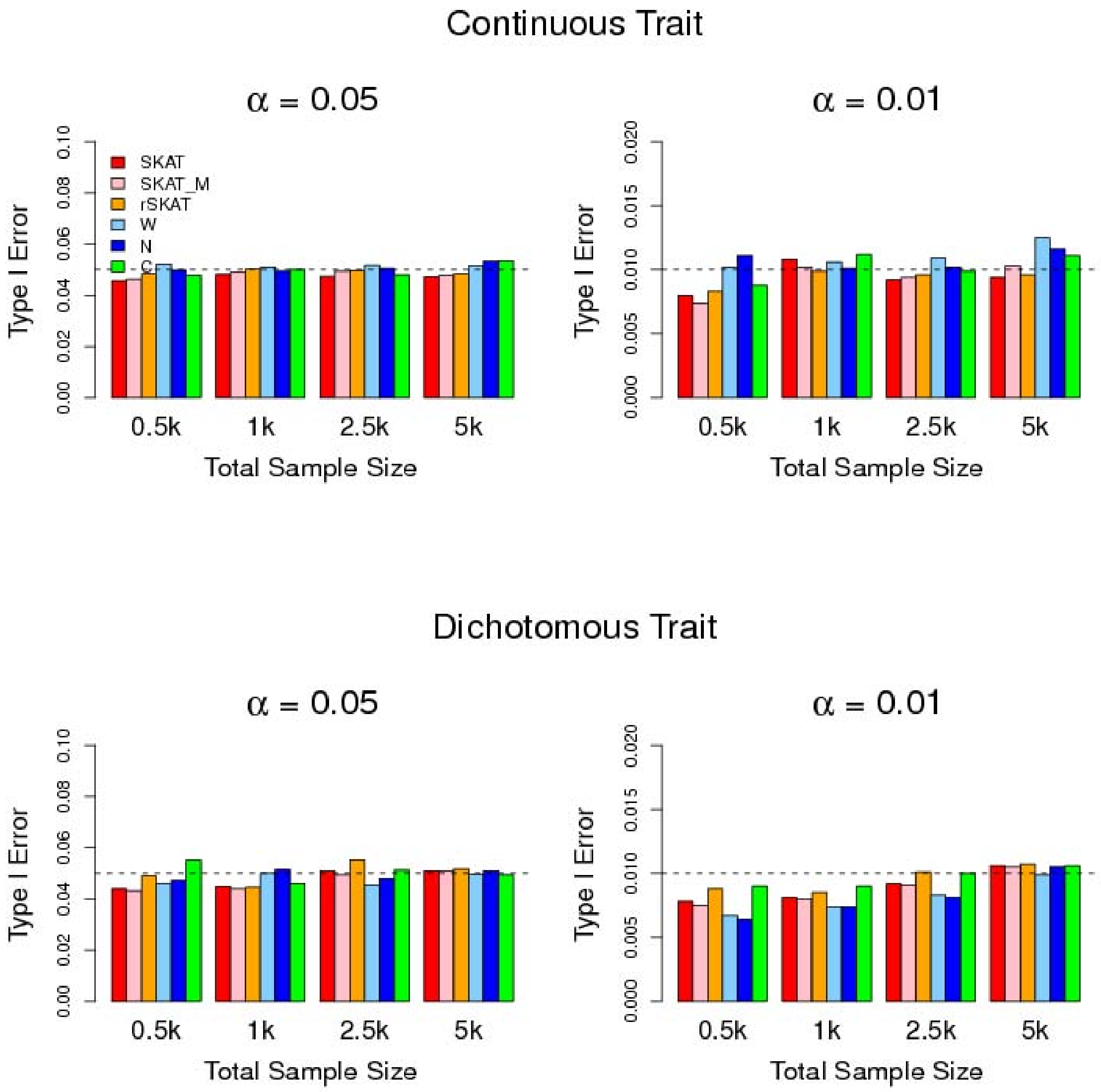


Figure S4) Distribution of allele frequencies for simulated and real data.

MAFs for simulated data are based on the population allele frequencies under the coalescent model (A) and MAFs for real data are estimated from the observed Dallas Heart study data (B).

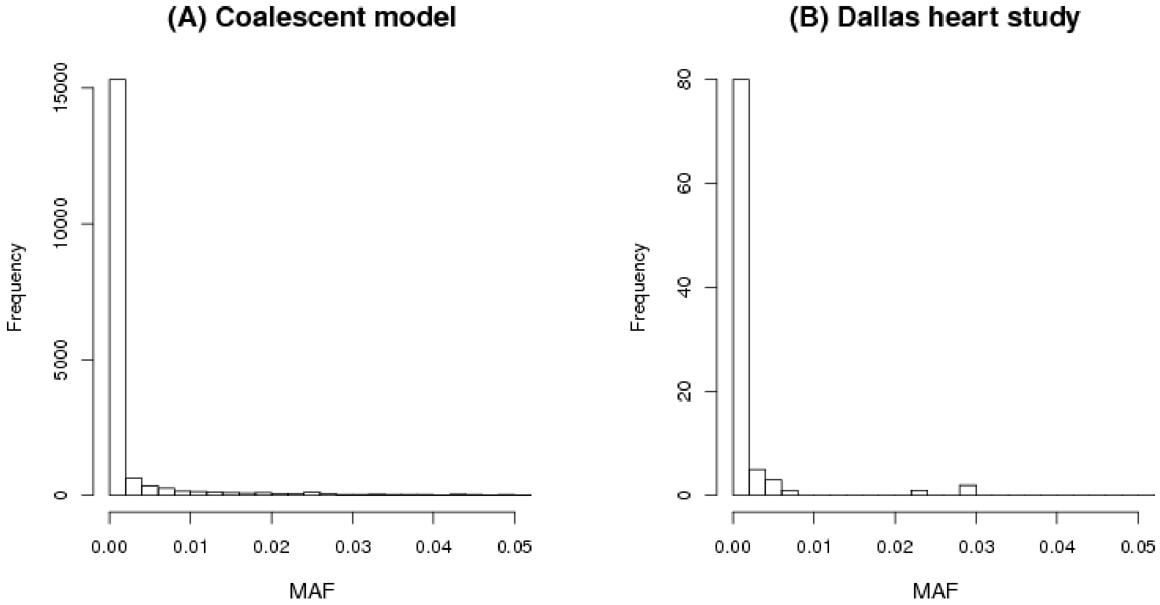


Figure S5) Power comparisons for SKAT and burden tests using simulation studies at the level α=10^{-2}.

Empirical power under the same configuration as that in Figure 1 except α= 10^{-2}. From left to right, coefficients for the causal rare variants are 100% positive/0% negative, 80% positive/20% negative, 50% positive/50% negative. Sample sizes considered are 500, 1000, 2500, and 5000. Each color represents a different method. For each method, power was estimated as the proportion of p-values < α=10^{-2} using 1000 simulated data sets.

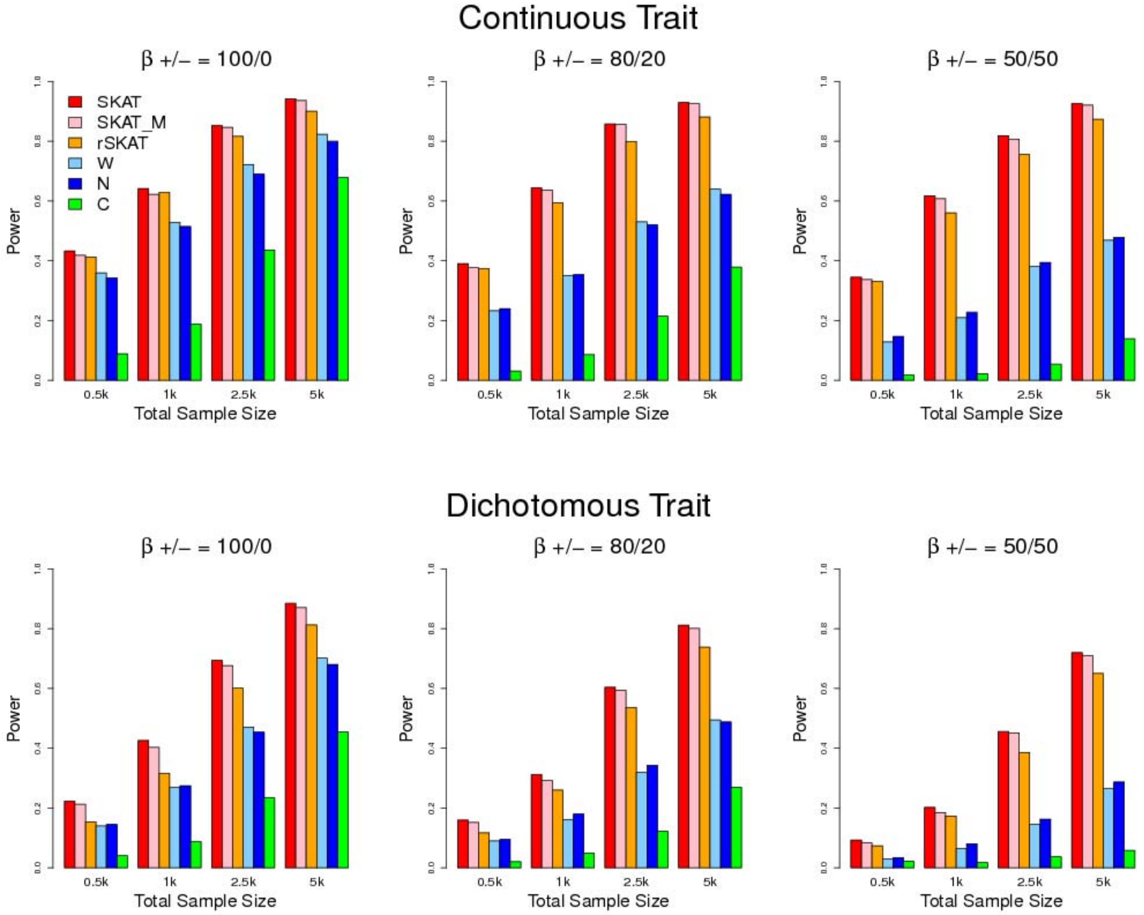


Figure S6) Power comparisons for SKAT and burden tests using simulation studies at the level α=10^{-3}.

Empirical power under the same configurations as that in Figure 1 except α=10^{-3}. From left to right, coefficients for the causal rare variants are 100% positive/0% negative, 80% positive/20% negative, 50% positive/50% negative. Sample sizes considered are 500, 1000, 2500, and 5000. Each color represents a different method. For each method, power was estimated as the proportion of p-values < α=10^{-3} using 1000 simulated data sets.

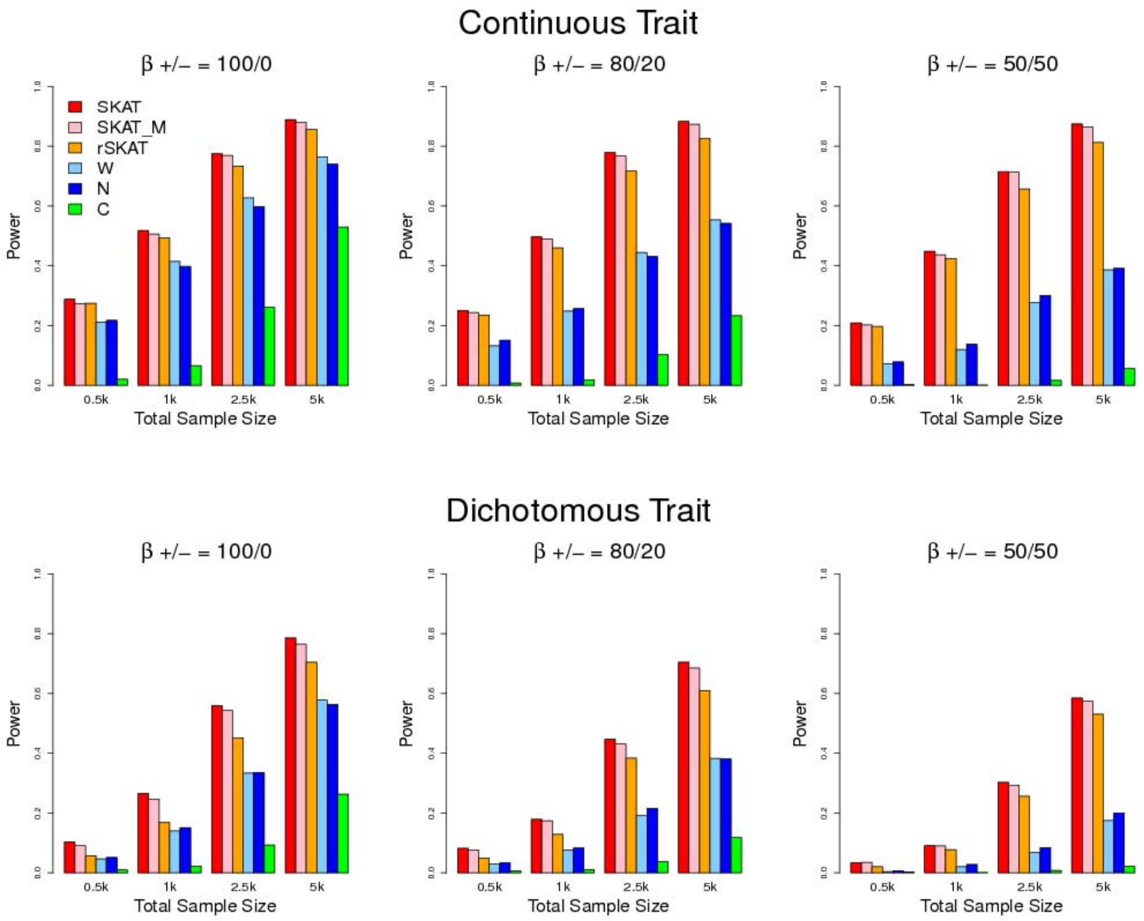


Figure S7) Power comparisons for SKAT and burden tests using simulation studies when causal variants have MAF<1%.

Empirical power at the genome-wide α=10^{-6} level under the same configuration to Figure 1 except 5% of the rare variants with MAF<1% were selected as causal variants. From left to right, coefficients for the causal rare variants are 100% positive/0% negative, 80% positive/20% negative, 50% positive/50% negative. Sample sizes considered are 500, 1000, 2500, and 5000. Each color represents a different method. For each method, power was estimated as the proportion of p-values < α=10^{-6} using 1000 simulated data sets.

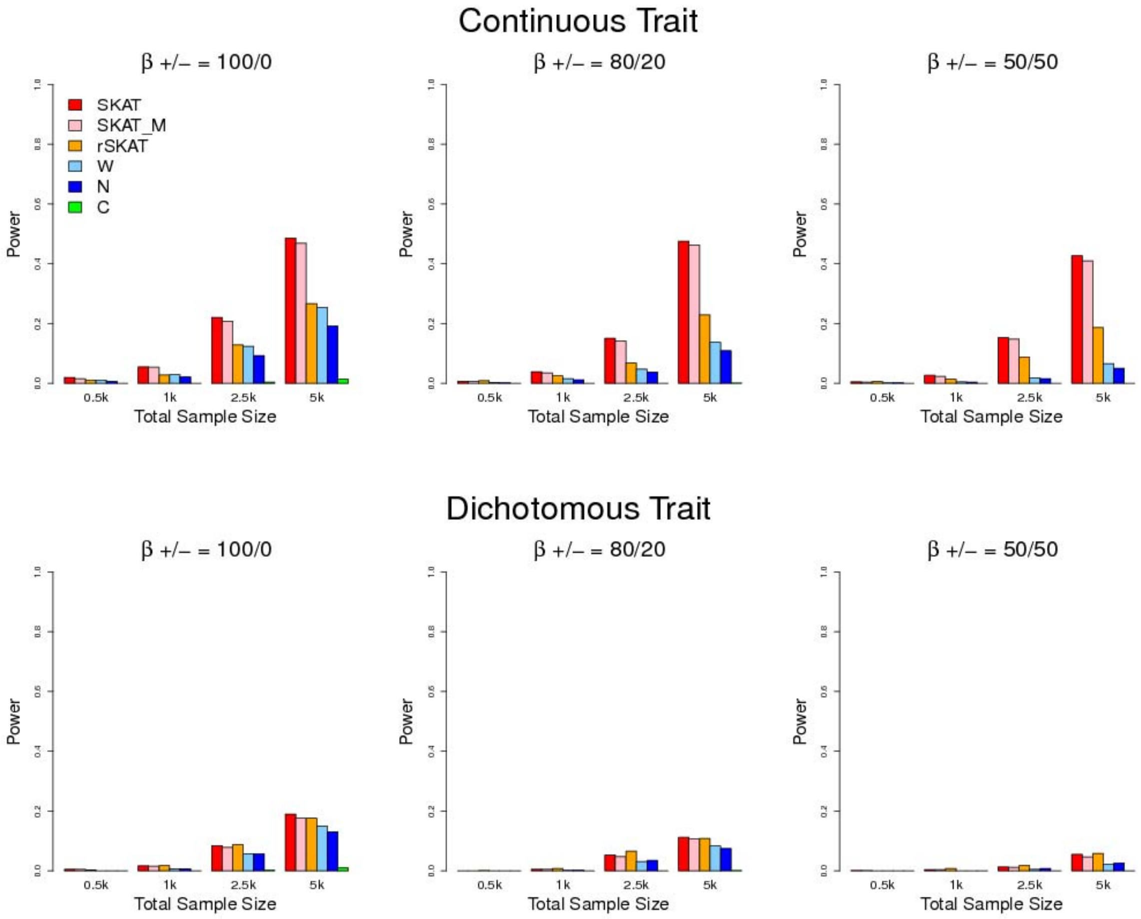


Figure S8) Power comparisons comparing SKAT with burden test using simulation studies with different region sizes.

Empirical power at the genome-wide $\alpha=10^{-6}$ level in the same setting as that in Figure 1 except the region size is 10kb (left panel) and 60kb (right panel). From left to right, the plots consider settings in which the coefficients for the causal rare variants are 100%, positive/0% negative, 80% positive/20% negative, 50% positive/50% negative. Sample sizes considered are 500, 1000, 2500, and 5000. Each color represents a different method. For each method, power was estimated as the proportion of p-values < α using 1000 simulated data sets.

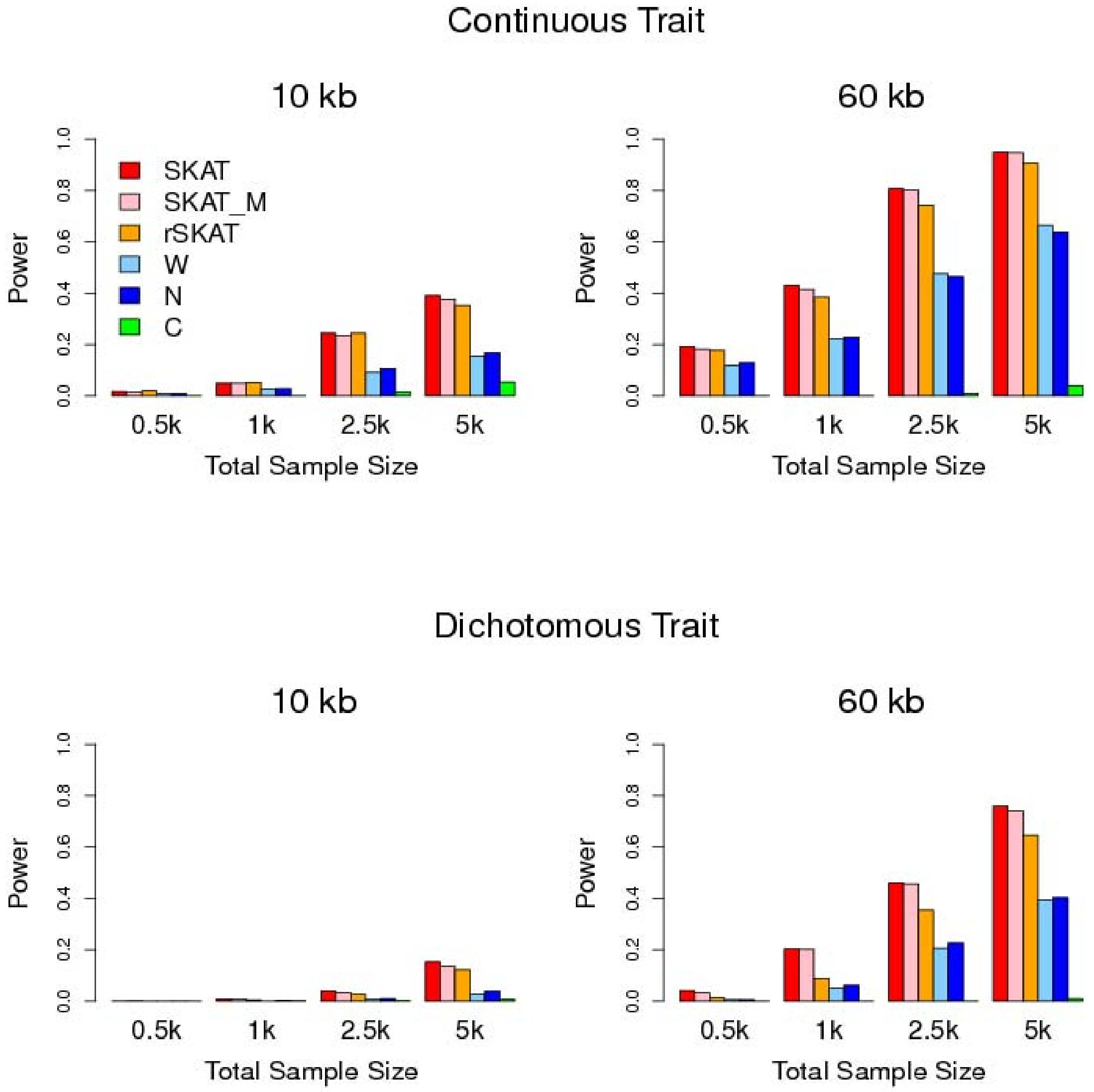


Figure S9) Power comparisons for SKAT and burden tests using simulation studies when region size=3kb and β+/-=100/0.

Empirical power at the genome-wide $\alpha=10^{-6}$ level assuming coefficients for the causal rare variants are 100% positive/0% negative. From left to right, 10%, 20%, and 50% of rare variants with MAF<3% were selected as causal variants. Top panel considers continuous phenotypes with maximum effect size equal to 2.4 when $MAF=10^{-4}$ and 10% of rare variants were causal; bottom panel considers case-control studies with maximum OR=13 when $MAF=10^{-4}$ and 10% of rare variants were causal. When more than 10% of rare variants were causal, smaller maximum effect size and maximum OR were used to adjust the increased number of causal variants. Sample sizes considered are 500, 1000, 2500, and 5000. Each color represents a different method. For each method, power was estimated as the proportion of p-values $< \alpha=10^{-6}$ using 1000 simulated data sets.

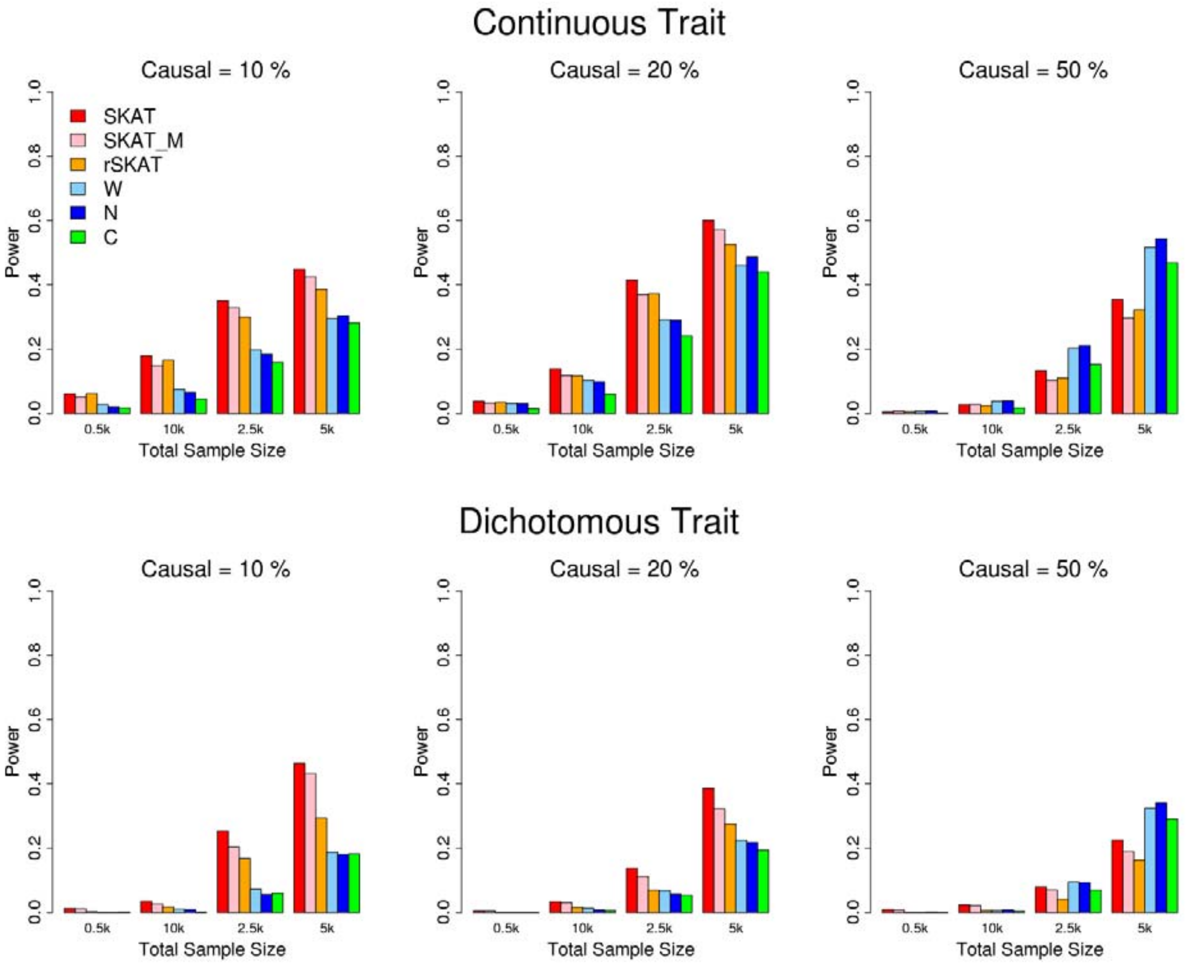


Figure S10) Power comparisons for SKAT and burden tests using simulation studies when region size=5kb and β+/-=100/0.

Empirical power at the genome-wide $\alpha=10^{-6}$ level assuming coefficients for the causal rare variants are 100% positive/0% negative. From left to right, 10%, 20%, and 50% of rare variants with MAF<3% were selected as causal variants. Top panel considers continuous phenotypes with maximum effect size equal to 2.4 when MAF=10^{-4} and 10% of rare variants were causal; bottom panel considers case-control studies with maximum OR=13 when MAF=10^{-4} and 10% of rare variants were causal. When more than 10% of rare variants were causal, smaller maximum effect size and maximum OR were used to adjust the increased number of causal variants. Sample sizes considered are 500, 1000, 2500, and 5000. Each color represents a different method. For each method, power was estimated as the proportion of p-values < $\alpha=10^{-6}$ using 1000 simulated data sets.

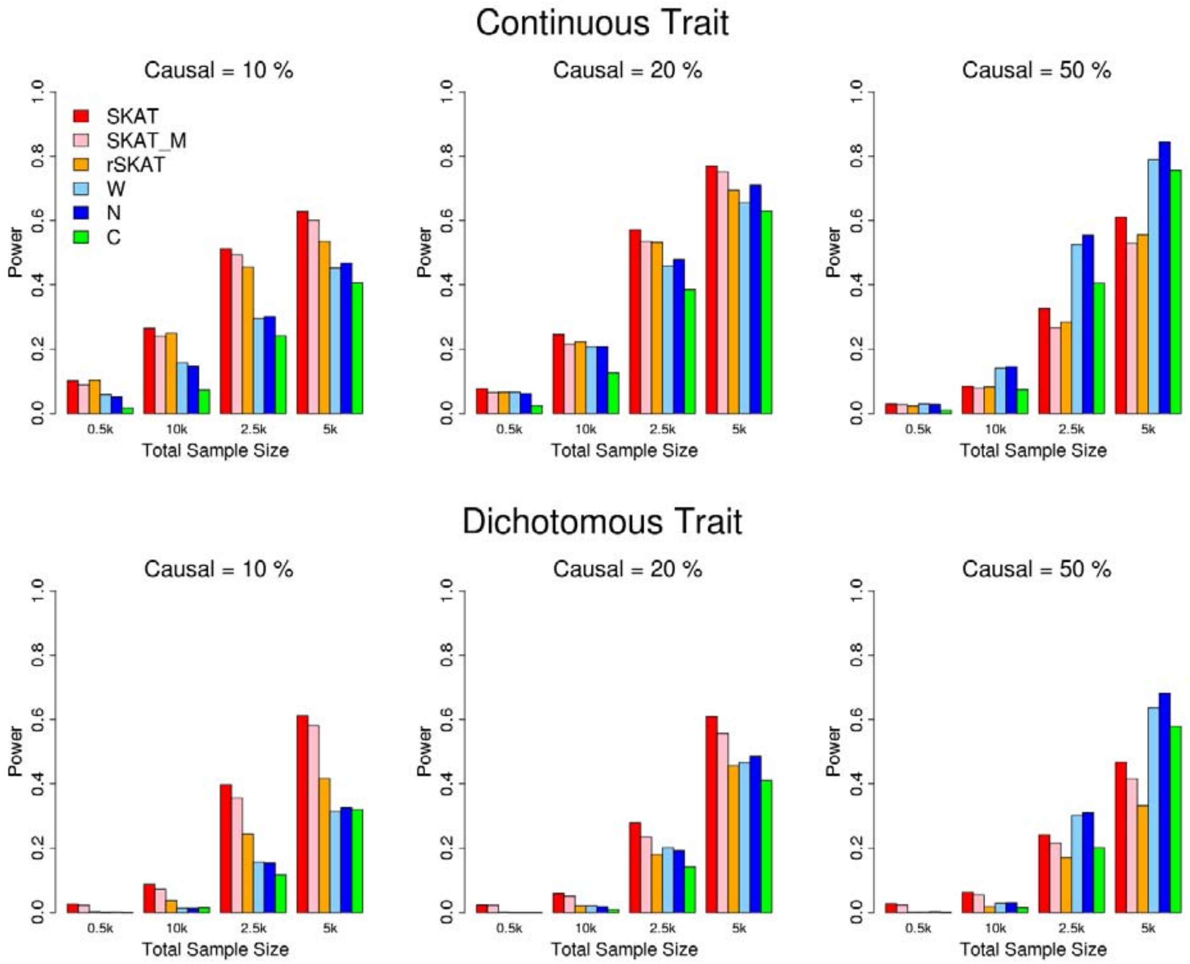


Figure S11) Power comparisons for SKAT and burden tests using simulation studies when region size=3kb and β+/-=80/20.

Empirical power at the genome-wide $\alpha=10^{-6}$ level assuming coefficients for the causal rare variants are 80% positive/20% negative. From left to right, 10%, 20%, and 50% of rare variants with MAF<3% were selected as causal variants. Top panel considers continuous phenotypes with maximum effect size equal to 2.4 when MAF=10^{-4} and 10% of rare variants were causal; bottom panel considers case-control studies with maximum OR=13 when MAF=10^{-4} and 10% of rare variants were causal. When more than 10% of rare variants were causal, smaller maximum effect size and maximum OR were used to adjust the increased number of causal variants. Sample sizes considered are 500, 1000, 2500, and 5000. Each color represents a different method. For each method, power was estimated as the proportion of p-values < $\alpha=10^{-6}$ using 1000 simulated data sets.

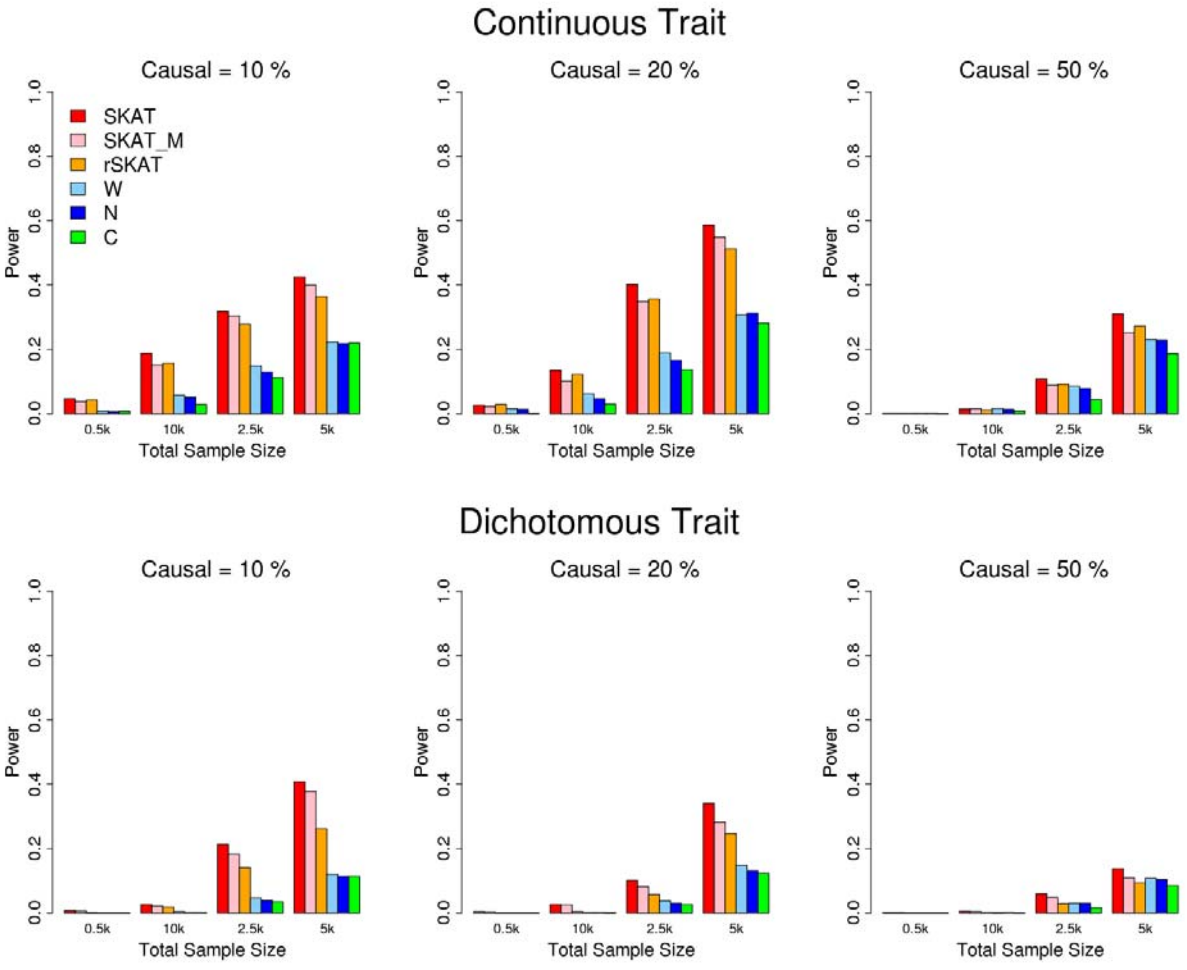


Figure S12) Power comparisons for SKAT and burden tests using simulation studies when region size=5kb and β+/-=80/20.

Empirical power at the genome-wide $\alpha=10^{-6}$ level assuming coefficients for the causal rare variants are 80% positive/20% negative. From left to right, 10%, 20%, and 50% of rare variants with MAF<3% were selected as causal variants. Top panel considers continuous phenotypes with maximum effect size equal to 2.4 when MAF=10^{-4} and 10% of rare variants were causal; bottom panel considers case-control studies with maximum OR=13 when MAF=10^{-4} and 10% of rare variants were causal. When more than 10% of rare variants were causal, smaller maximum effect size and maximum OR were used to adjust the increased number of causal variants. Sample sizes considered are 500, 1000, 2500, and 5000. Each color represents a different method. For each method, power was estimated as the proportion of p-values < $\alpha=10^{-6}$ using 1000 simulated data sets.

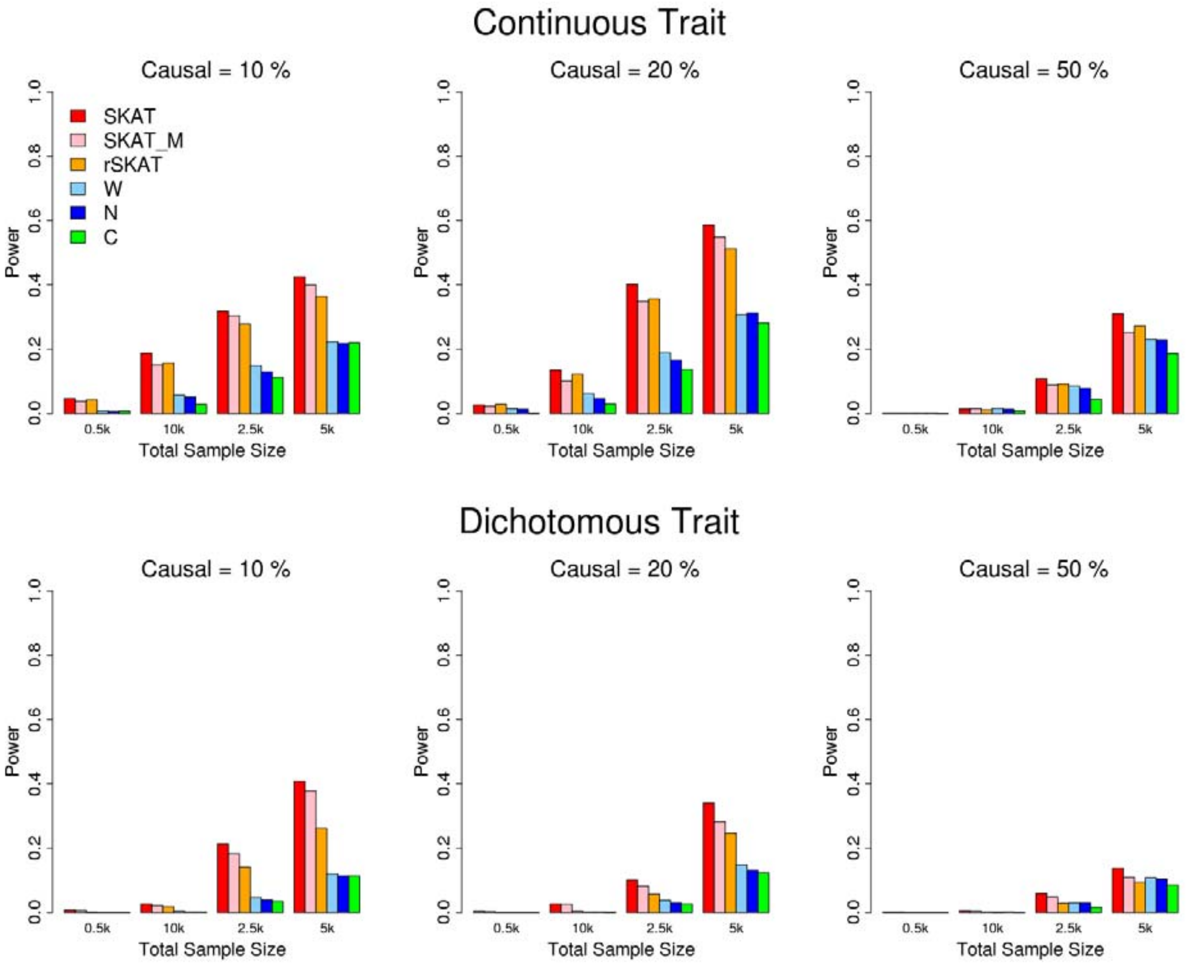


Figure S13) Required sample sizes to reach 80% power at the genome-wide level α=10^{-6} when causal variants have MAF<1%.

Required sample sizes to reach 80% power to detect rare variants associated with a continuous phenotype (top panel) or a dichotomous phenotype in case-control studies (half are cases) (bottom panel) at the α=10^{-6} (red), 10^{-3} (orange) and 10^{-2} (blue) level as estimated by the analytical formulae using beta(0.5 , 0.5) weights under the same setting as Figure 1 except that the 5% of variants with MAF<1% are generated as causal variants. Required sample sizes are plotted against the "maximum" effect sizes (ORs) when MAF=10^{-4}. Estimated total sample sizes were averaged over 100 random 30kb regions. See **Methods and Supplementary Materials.**

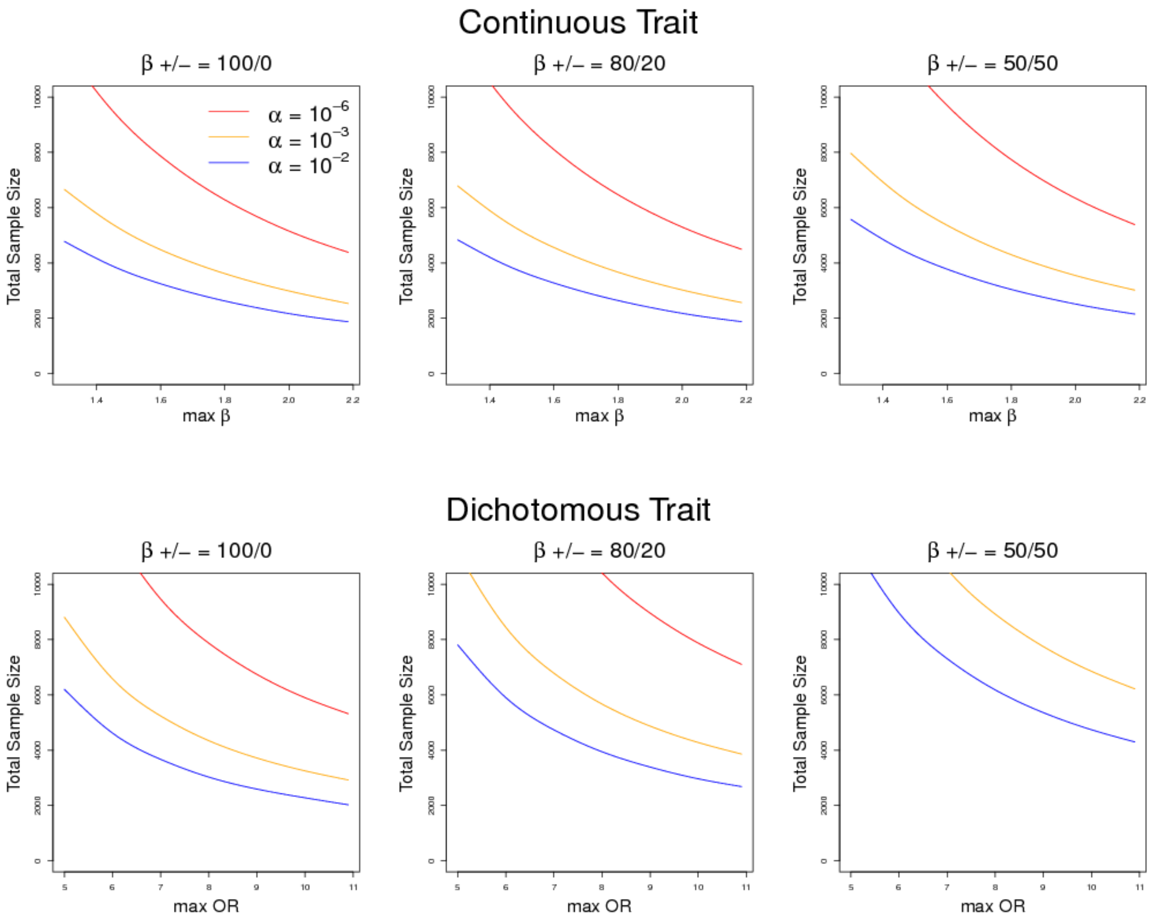


Figure S14) Required sample sizes to reach 80% power at the genome-wide level α=10^{-6} when 10% of rare variants are causal variants.

Required sample sizes estimated analytically to reach 80% power to detect rare variants associated with a continuous (top panel) or dichotomous phenotype in case-control studies (half are cases) (bottom panel) at the α=10^{-6}, 10^{-3}, and 10^{-2} level. The setting is the same as Figure 2 except that a 10% of rare variants with MAF<3% within the 30 kb regions are causal. Plots correspond to 100%, 80%, and 50% of the causal variants associated with increase in the continuous phenotype or risk of the dichotomous phenotype. Regression coefficients for the *s* causal variants were assumed to be the same decreasing function of MAF as that in Figure 1. Required total sample sizes are plotted again the "maximum" effect sizes (ORs) when MAF=10^{-4}. Estimated total sample sizes were averaged over 100 random 30kb regions.

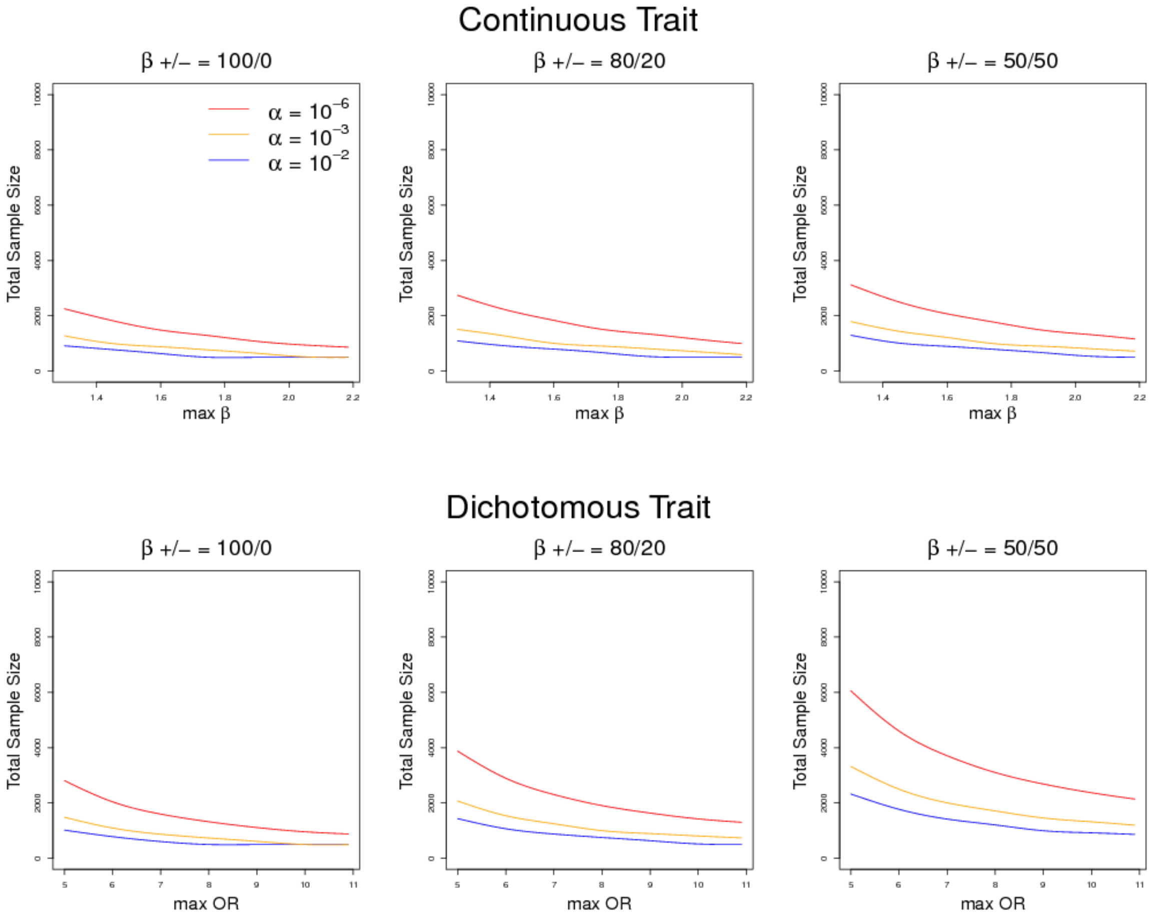
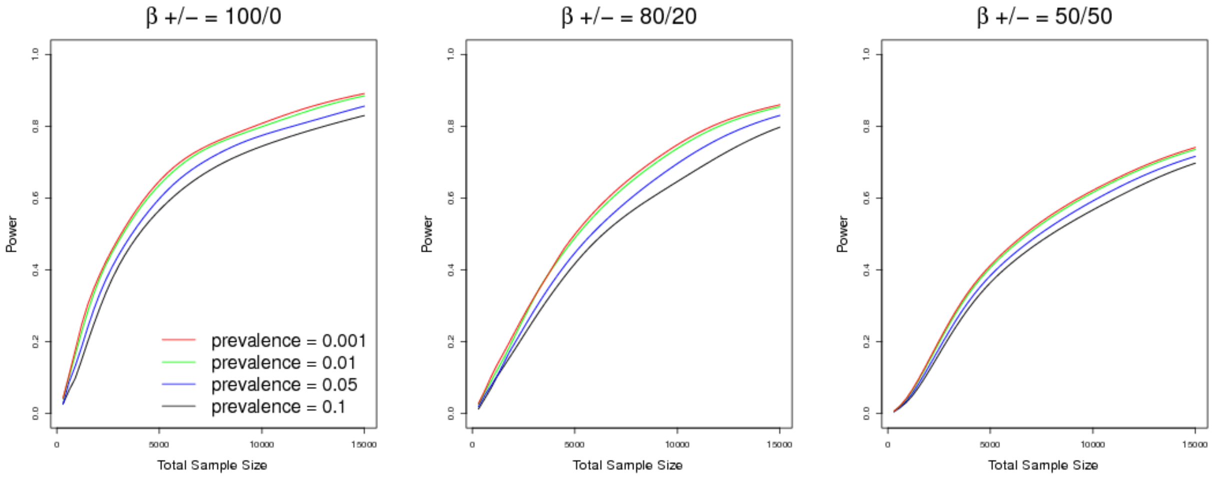


Figure S15) Estimated power to detect rare variants association with a dichotomous case-control phenotype for different prevalences.

Power to detect rare variants association with a dichotomous case-control phenotype estimated by the analytical formulae under the same setting as Figure 2 except that the disease penetrance = 0.1 (black), 0.05 (blue), 0.01 (green) and 0.001 (red). Plots correspond to the scenarios when 100%, 80%, and 50% of the causal variants are associated with increase in the risk of the dichotomous phenotype. Estimated powers were averaged over 100 random 30kb regions.