Detecting Rare Variant Effects Using Extreme Phenotype Sampling in Sequencing Association Studies

Journal:	*Genetic Epidemiology*
Manuscript ID:	GenEpi-12-0096.R2
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	20-Oct-2012
Complete List of Authors:	Barnett, Ian; Harvard University, Biostatistics Lee, Seunggeun; Harvard School of Public Health, Department of Biostatistics Lin, Xihong; Harvard, Biostatistics
Key Words:	Complex trait associations, Selective sampling, Rare genetic variants, Extreme phenotype sampling

Detecting Rare Variant Effects Using Extreme Phenotype Sampling in Sequencing Association Studies

(Short title: Test Rare Variant Effects for Extreme Phenotypes)

Ian J. Barnett, Seunggeun Lee and Xihong Lin

Department of Biostatistics, Harvard School of Public Health, Boston, MA

Address for Correspondence: Xihong Lin, Ph.D.

Department of Biostatistics, Harvard School of Public Health

655 Huntington Avenue, Boston, MA 02115

Phone: (617) 432-2914

Fax: (617) 432-5619

E-mail: xlin@hsph.harvard.edu

Abstract

In the increasing number of sequencing studies aimed at identifying rare variants associated with complex traits, the power of the test can be improved by guided sampling procedures. We confirm both analytically and numerically that sampling individuals with extreme phenotypes can enrich the presence of causal rare variants and can therefore lead to an increase in power compared to random sampling. While application of traditional rare variant association tests to these extreme phenotype samples requires dichotomizing the continuous phenotypes before analysis, the dichotomization procedure can decrease the power by reducing the information in the phenotypes. To avoid this, we propose a novel statistical method based on optimal SKAT (SKAT-O) that allows us to test for rare variant effects using continuous phenotypes in the analysis of extreme phenotype samples. The increase in power of this method is demonstrated through simulation of a wide range of scenarios as well as in the triglyceride data of the Dallas Heart Study.

Key Words

Complex trait associations; Selective sampling; Rare genetic variants; Extreme phenotype sampling

Introduction

With the increase in the number of sequencing studies[Biesecker, et al. 2011], there is a newfound access to samples with low frequency (MAF 1-5%) and rare (MAF <1%) genetic variants. In the search for genetic components of complex traits, discovered common variants (MAF > 5%) from genome-wide association studies explain only a small proportion of the total heritability of these traits[Ioannidis, et al. 2009; Maher 2008; Manolio, et al. 2009]. As a result, attention has turned to low frequency and rare variants instead expecting that they could play an important role in uncovering gene-phenotype relationships[Cirulli and Goldstein 2010; Ji, et al. 2008; Nejentsev, et al. 2009; Ng, et al. 2008; Ramser, et al. 2008]. Unfortunately, rare variants are difficult to detect in even reasonably large samples. This problem can be alleviated through the development of powerful study designs. To this effort, numerous association studies have chosen to sample subjects with extreme phenotypes in the hope of increasing power to detect causal SNPs[Clement, et al. 1995; Gu, et al. 1997; Hu, et al. 2009; Khor and Goh 2010; Li and Leal 2008; Liang, et al. 2000; Price, et al. 2008; Risch and Zhang 1995]. There have also been numerous developments in methodology to detect QTLs under these extreme phenotype sampling (EPS) study designs[Chen, et al. 2005; Huang and Lin 2007; Li, et al. 2011; Slatkin 1999; Wallace, et al. 2006]. A fundamental assumption that motivates these EPS methods is that rare causal variants are more likely found in the extremes of the quantitative trait. In this paper, we support the use of this practice by showing both analytically and numerically that EPS increases the presence of rare causal variants in a variety of settings. As a result, we show that EPS is more powerful for detecting for rare variant effects than random sampling.

Various methods have been proposed to tackle the challenge of association testing for rare variants. Burden tests such as the Combined Multivariate and Collapsing method (CMC)[Li and Leal 2008], Cohort Allelic Sums Test (CAST)[Morgenthaler and Thilly 2007] and the Weighted Sum Test (WST)[Madsen and Browning 2009] combine information from all rare variants within a target region such as an exon or gene by collapsing them into a single genetic variable, which is tested for association with the phenotypes of interest. Numerous rare variants testing methods have been developed using the same strategies [Bansal, et al. 2010; Basu and Pan 2011; Lee, et al. 2012a; Morris and Zeggini 2010; Price, et al. 2010]. A limitation of all burden tests is that they could lose significant amount of power in the presence of variants with different association directions and a large fraction of non-causal variants in the region. Alternatively the Sequence Kernel Association Test (SKAT)[Wu, et al. 2011] aggregates evidence of individual variant effects across the region using a kernel function and uses a computationally efficient mixed model variance component test to test for association. SKAT can naturally adjust covariates and has robust power in the presence of variants with different association directions and a large proportion of null variants. It is also a generalization of several non burden tests such as C-alpha test[Neale, et al. 2011], the SSU test[Pan 2009], and the haplotype association test[Tzeng and Zhang 2007]. Recently the optimal SKAT (SKAT-O)[Lee, et al. 2012b] has been proposed to unify the burden test and SKAT to a single framework and to construct the optimal test within the framework.

Moreover, limited statistical methods have been developed for studying rare variant effects when extreme phenotypes are sampled. In a typical EPS study, the two extremes are treated as two different groups representing a dichotomous phenotype. For example, Hu et al.[Hu, et al. 2009] used the contrast between subjects with high HDL-C levels against those with low HDL-C levels to identify an association with the ABCA1 gene. If the same method of extreme sampling were to instead retain the continuous phenotype values, the gain in information could provide greater power to detect gene-phenotype associations. For common variants, Huang and Lin[Huang and Lin 2007] proposed testing for associations between extreme continuous phenotypes and variants using the maximum likelihood method assuming a truncated normal distribution for extreme phenotype. Recently, this approach was adapted by Li et al. [Li, et al. 2011] to accommodate testing for multiple rare variant effects with the burden CMC approach. As a burden test, this approach is powerful when most variants in a region are causal and the effects of causal variants are in the same direction. However, it loses power in the presence of variants with different association directions or a large number of non-causal variants in a region.

In this paper, we first confirm both analytically and empirically that EPS substantially increases the chance to observe rare causal variants and hence increases their observed frequencies in finite study samples. Using this result, we demonstrate that EPS provides a more powerful design strategy for testing rare variant effects compared to random sampling. We next develop a new more powerful statistical method for testing for rare variant effects in EPS. Specifically, we extend SKAT and the optimal SKAT (SKAT-O) to EPS by analyzing extreme phenotypes as continuous variables within a likelihood framework. We show that the proposed tests perform well in a wide range of situations and outperform burden tests. We further show that analysis using continuous extreme phenotypes (CEP) improves power for detecting rare variant effects compared to using dichotomized extreme phenotypes (DEP). We illustrate the finite sample performance of proposed methods by conducting extensive simulations and application to analysis of triglyceride levels from the Dallas Heart Study.[Victor, et al. 2004]

Material and Methods

Goals and notation

The goal is to find an optimum sampling strategy when resources are limited and to develop powerful association test methods to detect phenotype-genotype associations. We evaluate the effectiveness of extreme phenotype sampling (EPS) compared to random phenotype sampling.

We first confirm analytically that extreme phenotype sampling enriches causal rare variants by increasing their MAFs (**Supplementary Materials Section 1**). We consider the cases with a single causal variant and multiple causal variants and calculate the MAF in extreme phenotype sampling as a function of the population MAF, the threshold used to select extreme phenotypes, and the effect sizes of genotypes.

We next evaluate the two different methods that utilize EPS phenotypes in different ways: the method that retains continuous phenotypes and the method that dichotomizes them into cases and controls. We consider the case with a sample of n individuals who have been sequenced in a genomic region of interest containing p genetic variants. The i-th individual has covariate information over m covariates $\mathbf{X}_i=(X_{i1},\ldots, X_{im})'$, genotypes of the p variants in the region $\mathbf{G}_i=(G_{i1},\ldots,G_{ip})'$, and a continuous phenotype y_i. The genotype G_{ij} represents the number of copies of the minor allele of the j-th variant that the i-th individual has.

Model

To test for an association between the variants and continuous phenotype while controlling for covariates, consider a linear model

$$y_i = \alpha_0 + \boldsymbol{\alpha}'\boldsymbol{X}_i + \boldsymbol{\beta}'\boldsymbol{G}_i + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$. Here α_0 is an intercept term, $\boldsymbol{\alpha} = [\alpha_1, \alpha_2,\ldots, \alpha_m]'$ is a vector of regression coefficients for the m covariates, and $\boldsymbol{\beta} = [\beta_1, \beta_2,\ldots, \beta_p]'$ is a vector of regression coefficients for the p genetic variants. The null hypothesis of H_0: $\boldsymbol{\beta}=0$ corresponds to no genetic effect on the trait. Since a *p*-DF likelihood ratio test has little power to detect causal variants particularly in the presence of a large number of rare variants, the gene-phenotype relationship is instead tested for by region-based tests such as burden tests and non-burden tests, e.g., SKAT. An adaptation of the CMC[Li and Leal 2008] burden test is used that collapses genotype information by counting the number of variants in the region before applying logistic regression to the collapsed statistic. We call this test DEP-Burden.

Association tests under the extreme phenotype sampling design

Since both SKAT and burden tests are capable of handling dichotomous phenotypes in the case-control setting, they can be applied to test for associations after using EPS. Dichotomizing the high phenotypic extremes as cases and the lower phenotypic extremes as controls is a natural extension of each test's functionality. However, applying SKAT and SKAT-O to continuous phenotype data obtained from EPS requires further development, since the extreme continuous phenotypes do not follow Gaussian distribution due to the phenotypic selection. Suppose we select *n* samples with either $y_i > c_1$ or $y_i < c_2$, and denote the selected y_i as y_i^*. Then under the null hypothesis y_i^* follows truncated Gaussian distribution with a density function

$$f(y_i^*) = \frac{\phi(X_i\alpha, \sigma^2)}{\Phi(c_2, \sigma^2) + 1 - \Phi(c_1, \sigma^2)}$$

where $\phi(\mu, \sigma^2)$ and $\Phi(\mu, \sigma^2)$ are density and distribution functions of Gaussian distribution with mean μ and variance σ^2.

To increase test power and decrease the test DF, we assume β_j follows an arbitrary distribution with mean 0 and variance φw_j^2. We note that H_0: $\boldsymbol{\beta}=0$ is equivalent to H_0: $\varphi=0$. The score test statistic of $\varphi=0$ is

$$Q_s = \sum_{j=1}^{p} w_j^2 \left(\sum_{i=1}^{n} G_{ij} (y_i^* - \hat{\mu}_i) \right)^2$$

where $\hat{\mu}_j$ is an estimated mean of y_j^* under the null hypothesis. We show that Q_S asymptotically follows a mixture of chi-square distribution (**Supplementary Materials Section 2**), and p-values can be obtained by the matching the moments or inverting the characteristic function[Davies 1980].

Recently Lee et al. (2012)[Lee, et al. 2012b] proposed an optimal unified approach, which unifies SKAT and burden test to adaptively select best test structure. Suppose Q_B is the score test statistics of the weighted burden test:

$$Q_B = \left(\sum_{i=1}^{n} (y_i^* - \hat{\mu}_i) \sum_{j=1}^{p} w_j G_{ij} \right)^2$$

and then the test statistic of unified test is

$$Q_\rho = (1- \rho)Q_S + \rho\ Q_B,$$

where ρ ($0 \le \rho \le 1$) is a parameter to determine whether test is close to SKAT (ρ=0) or burden tests (ρ=1). It is based on a recent generalization of SKAT which allows the correlation among variants effects β's. Under this setting, they proposed the optimal SKAT, called SKAT-O. This test is defined by selecting the ρ that minimizes the p-value of the SKAT-O test statistic,

$$T = \min_{0 \le \rho \le 1} p_\rho$$

where p_ρ is a p-value with given ρ. The test statistic T can be obtained by simple grid search across a range of ρ: set a grid $0 = \rho_1 < \rho_2 < \ldots < \rho_b = 1$, then T=min($p_{\rho 1,\ldots,} p_{\rho b}$). In simulation studies and real data analysis, we used the equal sized grid of 11 points (from 0 to 1) to obtain T. From the fact that the Q_ρ can be decomposed to the shared random variables, asymptotic p-value of T can be obtained through computationally efficient one-dimensional numerical integration

(**Supplementary Materials Section 3**). We use this extreme phenotype optimal SKAT in our simulation studies and data analysis.

When the sample size is small, SKAT family methods (including SKAT and SKAT-O) can produce conservative results with both binary and extreme continuous phenotypes. To resolve this issue, Lee et al. (2012a, 2012b)[Lee, et al. 2012a; Lee, et al. 2012b] have proposed a method to adjust asymptotic null distribution by estimating small sample moments when the trait is dichotomous. We employ a similar approach (details in **Supplementary Materials Section 4**). For all simulation studies and real data analysis we used small sample adjustment for SKAT methods given the small to moderate sample sizes we considered. We used SKAT-O for continuous extreme phenotype SKAT (CEP-SKAT-O) and dichotomous extreme phenotype SKAT (DEP-SKAT-O), and for random sample continuous phenotypes (RS-SKAT-O). It should be noted that for larger sample sizes, the small sample adjustment is not necessary. Through simulations we found sample sizes lower than n=500 to benefit from the small sample adjustment, with sample sizes as low as n=1000 to not benefit from the adjustment.

Type 1 error simulations

We first generated haplotype data by the forward simulator, SFS_CODE[Hernandez 2008], which offers the ability to incorporate purifying selection on deleterious variants and thus provides better model to simulate variants in exomes. Data were simulated according to the European demographic model with a population bottleneck followed by exponential growth. We simulated 32,000 haplotypes each 100,000 base pairs wide as our population base. To achieve a simulated sample over a 3kb exon, a random 3kb region is selected (containing 41 variants on average) and each individual genotype is formed by combining at random two haplotypes over that region. Phenotypes for the *i*-th individual in a sample were produced from the generated genotype and covariate data according to

$$Y_i = 0.5X_{i1} + 0.5X_{i2} + \varepsilon_i$$

Where the covariate X_{i1} is 1 with probability 0.5 and 0 otherwise, and the covariate X_{i2} and the residual ε_i are both instances of a standard normal random variable.

Using the simulated genotype and phenotype data for the *N* individuals, a random sample of size *n* is selected. For random sampling of continuous traits, SKAT-O with the default w_j=Beta(1,25) weight is used to test for an association between the continuous phenotype and genotype after controlling for both covariates, producing a p-value (RS-SKAT-O) . In order to test for the association between variants and phenotype under EPS using the standard dichotomizing method, we treat the highest (*n*/2) extremes as cases and lowest (*n*/2) extremes as controls. The dichotomized phenotypes are used by both DEP-SKAT-O and DEP-Burden. This same extreme phenotype sample is used to compare with the tests that retain the continuous phenotype (CEP-SKAT-O and CEP-Burden). A p-value for the CEP-SKAT-O test and the CEP-Burden burden test are produced from these continuous phenotype values and the corresponding genotype and covariate data. The proportion of p-values below a specified α-level provides an estimate for the Type 1 error at that α significance level.

Power Simulations

Power comparisons between the various sampling methods were performed using simulated genotype data as was used in the Type 1 error simulation setting. After generating the genotypes for *N* individuals, 20% of the variants with MAF < 0.03 are selected to be causal variants. Different percentages of causal variants were also considered. Phenotypes are then generated for the *N* individuals according to:

$$y = 0.5X_1 + 0.5X_2 + \beta_1 G_1 + \beta_2 G_2 + \ldots + \beta_p G_p + \varepsilon$$

The covariate X_1 is generated as a Bernoulli random variable with p=0.5. The covariate X_2 and the added noise ε are generated independently from a standard normal distribution. Non-causal variants are assigned $\beta_j=0$, and the causal variants are generated according to:

$$|\beta_j| = -a \log_{10}(\mathrm{MAF}_j)$$

Here, $a>0$ is a parameter that specifies the strength of variant-phenotype associations, hence the strength of heritability. Large values of a lead to stronger effects of causal variants on phenotype and cause rare variants to become more enriched in the phenotypic extremes. In one simulation setting an increase in a from 0.3 to 0.4 increases the heritability of the phenotype from 0.034 to 0.042. The heritability also increases with the number of causal variants. To obtain an estimate of the heritability, the proportion of the variance in phenotype explained by the genotypes of causal variants is estimated assuming no LD between variants.

Power estimates are obtained for various (extreme phenotype) sample sizes (n=500, 1000, and 2000), percentages in each phenotypic extreme sampled (10% and 25%), percentages of causal variants with a positive effect (80%, 100%), and percentages of causal variants with MAF < 0.03 (20%, 40%, and 60%).

Results

Extreme sampling enriches rare causal variants

Our analytical calculations (See **Material and Methods** and **Supplementary Materials Section 1**) confirm that rare causal variants can be enriched in phenotypic extremes. The degree of enrichment increases when more extreme phenotypes are sampled and a higher percentage of causal variants are present in a region. To empirically validate this finding, randomly selected 3kb exonic regions were simulated using the population genetic simulation model with European demographic history (see **Material and Methods**). For each 3kb region, causal variants were randomly selected to be 100%, 70%, 40% and 0% of sufficiently rare variants (MAF < 0.03) and the *j*-th causal variant was given the effect size β_j as a function of its MAF. Note that these causal variant percentages differ from those in the power simulations so as to further accentuate the effect of causal variant percentage on the inflation of MAF due to EPS. Also for the power simulations, causal variant percentages of 10% and 20% were used instead. Phenotypes are then generated from a linear model with heritability of genetic variants being 2.6%, 1.3%, and 0%.

Because the causal variants are known in the simulation setting, the expected MAF of a causal variant using EPS can be computed analytically (see **Supplementary Materials Section 1**). The expected MAFs of causal variants using EPS matched closely with the sample MAFs of causal variants using EPS (**Figure 1**). The MAFs of simulated causal variants after EPS had an overall increased frequency over the respective population MAFs.   This trend decreases as samples are restricted to less extreme phenotypes and heritability is lower. No enrichment is found when there is no causal variant. When both the causal and non-causal variants in a region are considered simultaneously, the median MAF using EPS is much less inflated than when only causal variants are examined.

Sampling methods for comparison

Motivated by the enrichment of causal rare variants in phenotypic extremes, we expect to find that EPS methods can increase power to detect rare causal variants over random sampling methods. We extend the SKAT family methods to test for region-level rare variant effects when continuous phenotypes obtained from EPS are used in analysis. In simulation and data analysis, we only use the extreme phenotype optimal SKAT (SKAT-O), which accounts for extreme phenotype sampling and unifies the burden test and SKAT to a single framework and by constructing the optimal test within the framework. Using the simulated genotype data over 3kb regions the phenotypes were generated using the additive linear model (see **Material and Methods**). Given the same sample size, we compare the power of three tests designed for detecting rare variant effects using EPS. We first consider a burden test, DEP-Burden, that uses dichotomized extreme phenotypes along with collapsed information over genotypes by simply counting the number of rare variants with MAF<3% in the gene before applying logistic regression to the collapsed statistic. We also apply this same collapsed statistic to continuous extreme phenotypes as done in Li et al.[Li, et al. 2011] and call this test CEP-Burden. Next we consider dichotomized extreme phenotype SKAT-O (DEP-SKAT-O), which applies optimal SKAT (SKAT-O) to dichotomized extreme phenotypes while applying small sample adjustments when sample sizes are small[Lee, et al. 2012a]. Finally we consider continuous extreme phenotype SKAT (CEP-SKAT-O), which does not dichotomize and instead extends linear regression optimal SKAT over the continuous extreme phenotypes (see **Material and Methods**) by using a truncated normal distribution. We also applied the small sample adjustment to CEP-SKAT-O to obtain the correct type I error rates when sample sizes are small. To demonstrate the benefits of EPS compared to random sampling, we included in the comparison a fourth method using random sampling SKAT-O (RS-SKAT-O), which applies optimal SKAT to the continuous phenotypes of a random sample. We assume the same sample size when comparing different methods so their powers are comparable. The power of each competing method is estimated as the proportion of p-values less than $\alpha=10^{-6}$ in an effort to imitate genome-wide association studies.

The type 1 error rates for CEP-SKAT-O were accurate at $\alpha=0.01$ and $\alpha=0.05$ and slightly inflated at genome-wide significance levels $\alpha=10^{-6}$(see **Supplementary Table 1**). When all causal variants had the same direction of effect, CEP-SKAT-O and CEP-Burden had the greatest power with a substantial lead over every other method (**Figure 2**). When causal variants had effects in opposite directions all tests lost power uniformly due to less enrichment of rare variants, but CEP-SKAT-O became the most powerful by a large margin (**Figure 3**). In this case DEP-Burden had the least power. The power of the three methods employing SKAT-O (CEP-SKAT-O, DEP-SKAT-O, and RS-SKAT-O) is much more robust to changes in the proportion of causal variants that have a positive effect than the burden test's power is. This is because SKAT-O allows for each individual variant to affect phenotype in different directions and also allows for no effect. On the other hand, burden tests assume all the causal variants share the same direction of effect and that all the variants in a region are causal, and so the power of the burden tests greatly diminishes when causal variants are allowed effects in opposite directions or many causal variants are allowed no effect.

When all causal variants having the same direction of effect and as the percent of rare variants that were causal increased, the power gap between DEP-Burden and DEP-SKAT-O reduces, an observation that is also true for the relationship between CEP-Burden and CEP-SKAT-O (**Figure 2**). This is because CEP-SKAT-O includes CEP-Burden as a special case and behaves like CEP-Burden automatically when most variants are causal with effects in the same direction. To see this, we found that in simulations the estimated ρ decreased by a factor of 0.36

on average when changing from the case of having all positive causal variant effects to the case of having causal variant effects in opposite directions.

Methods utilizing extreme sampling benefit as the cutoff for extreme phenotypes increases. In particular, as the percent of the tails sampled from the distribution of phenotype decreases from 25% to 10%, all EPS tests show incremental increases in power given the same sample size due to higher enrichment of rare variants. Relative power comparisons remain unchanged after decreasing the heritability of the phenotype and after increasing the exon length to 5kb or 10kb (regions of these lengths contain 69 variants and 138 variants on average, respectively). Also, simulations were also performed where the β was selected to be a constant rather than being a decreasing function of the MAF but the relative power of the methods remained the same (**Supplementary Figure 2**). Regardless of the setting, CEP-SKAT-O was consistently robust and had the greatest overall power to detect gene-phenotype associations over the other methods.

Application to the Dallas Heart Study data

In the Dallas Heart Study [Victor, et al. 2004], 3476 individuals were sequenced over the genes *ANGPTL3* (MIM 604774)*, ANGPTL4* (MIM 605910)*,* and *ANGPTL5* (MIM 607666). A total of 93 variants are present over these genes, and the variants in all three genes were tested simultaneously for an association with log-transformed serum triglyceride levels (logTG). Analysis for each of three genes separately is also considered (see **Supplementary Materials Section 6**). Ethnicity and sex were adjusted for in the analysis. To demonstrate rare variant association test methods for extreme phenotype sampling (EPS), a total of 1,389 individuals with the highest 20% and lowest 20% of logTG levels in each age-gender stratum were selected as the ESP sample. The continuous values were used in CEP-SKAT-O while dichotomized values were used for DEP-SKAT-O and DEP-Burden. Random samples of equivalent size were selected for the RS-SKAT-O method for comparison purposes.

To compare the effects of the different cutoffs of tails, we considered to sample individuals from wider tails (30% and 40%). Since wider tails had more samples, to make p-values comparable, we randomly sub-sampled 1,389 individuals among individuals in wider tails in order to have the same sample size as compared to a 20% cut off. In these cases, median p-values calculated from multiple random samples were obtained (**Table 1**). The p-values for all EPS methods are sensitive to the extreme phenotype cutoff. CEP-SKAT-O outperforms the other methods when there is sufficient information about the continuous trait distribution. It performs similarly to DEP-SKAT-O where there is limited information in the data, e.g., when the cutoff is low or when there are a small number of rare variants in a gene (**Supplementary Materials Section 6).** When extremes were sampled from wider tails (30% and 40%) all of the tests tended to lose significance, demonstrating the strength of EPS. We also computed p-values with different cutoffs and unequal sample sizes (**Supplementary Figure 1**), and CEP-SKAT-O outperformed other competing methods overall.

Power Estimation

In the planning of new sequencing studies, it is important to be able to estimate the power to detect causal variants under various study designs. We provide such power and sample size calculations for extreme phenotype sampling designs using CEP-SKAT-O. We use analytical formulas to obtain the distribution of our statistic by allowing users to specifying desirable parameters of interest (see **Supplementary Materials Section 5**). The parameters that can be specified by the user include sample size, the percent of causal variants, the length of the genomic region, the effect size of the causal variants, the proportion of causal variants that have a positive effect, and the proportion of the tails that are sampled in EPS. We find that power is increased as sample size increases, as the proportion of causal variants increases, as the effect size increases, when causal variants have their effect in the same direction, and when we are more selective by sampling individuals with more extreme phenotypes. The power is also dependent on the genomic region as the distribution of the number of genetic variants, the MAF distribution, and the LD structure vary over the genome, so for genome-wide studies power estimations are averaged over many randomly selected regions of equivalent size.

To evaluate the accuracy of these analytic power estimations, we show a side by side comparison with empirical power simulations (**Figure 4** and **Supplementary Figure 3**). In this setting we consider 3kb regions with 20% of variants being causal with all effects in the same direction. We see that the estimated power with our analytic calculations matches the empirical power over a wide range of sample sizes.

Discussion

We confirm in this paper through analytical calculations and simulation studies that sampling phenotypic extremes of a population can enrich rare causal variants. As a consequence, we show that sampling from phenotypic extremes profit over analogous random sampling methods by showing a sizable gain in power when the same size is used. In particular, analysis using dichotomized extreme phenotype (DEP-SKAT-O) is shown to be more powerful than that using random sampling with continuous phenotypes (RS-SKAT-O) in almost all scenarios. We develop a new method, continuous extreme phenotype optimal SKAT (CEP-SKAT-O), which improves upon DEP-SKAT-O by retaining continuous phenotype information rather than dichotomizing, includes the continuous extreme phenotype burden test as a special case, and results in a significant increase in sensitivity to causal variants. We find that CEP-

SKAT-O has the overall greatest power in a wide variety of settings over DEP-SKAT-O, RS-SKAT-O, and comparable collapsing methods.

In the realm of region based association testing methodology, there already exist many methods capable of handling continuous phenotypes when a normal distribution is assumed. However in the case of extreme sampling, phenotype follows a truncated normal distribution instead, and so current methods cannot be directly applied without dichotomizing. The advantage of CEP-SKAT-O is that it adapts SKAT-O, a method that applies multiple linear regression of a phenotype on all genotypes in a region, to be able to handle phenotypes coming from a truncated normal. This adaptation allows for the usual continuous phenotype analysis without forcing the usual loss of phenotype information that occurs due to dichotomizing.

CEP-SKAT-O assumes that subjects are sampled from the extremes of a normally distributed phenotype, which in some circumstances may be inappropriate, and hence the test results can be biased when the normality of the underlying trait is violated. Dichotomizing phenotypes using DEP-SKAT-O is robust to departure from normality, although it is subject to some power loss when normality assumption is true. For candidate gene studies, permutation can be used to estimate the null distribution of the CEP-SKAT-O test statistic when the underlying trait does not follow a normal distribution. However, this is computationally difficult for GWAS where genome-wide significance levels are very stringent and a large number of genes are tested. We have found the maximum likelihood estimator of σ^2 seems to be sensitive to the distribution of the underlying trait. It is of future research interest to develop an alternative estimator of σ^2 that is more robust to deviations from normality.

In several ongoing exome sequencing studies conducted in the NHLBI Exome Sequencing Project https://esp.gs.washington.edu/drupal/, subjects were sampled using extremes of multiple phenotypes, future research is needed to develop methods for analyzing this more complex sampling setting. When subjects with extreme phenotypes are sampled for sequencing, covariate confounding effects need to be accounted for at the design phase to ensure representative samples. One strategy is to use stratified sampling, i.e. sample extreme phenotypes within key covariate stratum. For example, a phenotype distribution is likely to be gender-specific. It is desirable to sample extreme phenotypes within each gender stratum. The residual covariate confounding effects can be adjusted for at the analysis stage.

Acknowledgements

This work was supported by grants T32 GM074897 (IB) and R37 CA076404 and P01 CA134294, P54 ES016454 (SL and XL).

Web Resources

The URLs for data presented herein are as follows:

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

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

References

Bansal V, Libiger O, Torkamani A, Schork NJ. 2010. Statistical analysis strategies for association studies involving rare variants. Nature Reviews Genetics 11(11):773-785.

Basu S, Pan W. 2011. Comparison of statistical tests for disease association with rare variants. Genetic Epidemiology.

Biesecker LG, Shianna KV, Mullikin JC. 2011. Exome sequencing: the expert view. Genome Biol 12(9):128.

Chen Z, Zheng G, Ghosh K, Li Z. 2005. Linkage disequilibrium mapping of quantitative-trait Loci by selective genotyping. Am J Hum Genet 77(4):661-9.

Cirulli ET, Goldstein DB. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11(6):415-25.

Clement K, Vaisse C, Manning BS, Basdevant A, Guy-Grand B, Ruiz J, Silver KD, Shuldiner AR, Froguel P, Strosberg AD. 1995. Genetic variation in the beta 3-adrenergic receptor and an increased capacity to gain weight in patients with morbid obesity. N Engl J Med 333(6):352-4.

Davies RB. 1980. The Distribution of a Linear Combination of Chi-square Random Variables. Journal of the Royal Statistical Society 29(3):323-333.

Gu C, Todorov AA, Rao DC. 1997. Genome screening using extremely discordant and extremely concordant sib pairs. Genet Epidemiol 14(6):791-6.

Hernandez RD. 2008. A flexible forward simulator for populations subject to selection and demography. Bioinformatics 24(23):2786-7.

Hu S, Zhong Y, Hao Y, Luo M, Zhou Y, Guo H, Liao W, Wan D, Wei H, Gao Y and others. 2009. Novel rare alleles of ABCA1 are exclusively associated with extreme high-density lipoprotein-cholesterol levels among the Han Chinese. Clin Chem Lab Med 47(10):1239-45.

Huang BE, Lin DY. 2007. Efficient association mapping of quantitative trait loci with selective genotyping. Am J Hum Genet 80(3):567-76.

Ioannidis JP, Thomas G, Daly MJ. 2009. Validating, augmenting and refining genome-wide association signals. Nat Rev Genet 10(5):318-29.

Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP. 2008. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 40(5):592-9.

Khor CC, Goh DL. 2010. Strategies for identifying the genetic basis of dyslipidemia: genome-wide association studies vs. the resequencing of extremes. Curr Opin Lipidol 21(2):123-7.

Lee S, Emonds M, Bamshad M, Barnes K, Rieder M, Nickerson D, Christiani D, Wurfel M, Lin X. 2012a. Optimal unified approach for rare variant association testing with application to small sample case-control whole-exome sequencing studies. American Journal of Human Genetics, in press.

Lee S, Wu MC, Lin X. 2012b. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. doi: 10.1093/biostatistics/kxs014.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83(3):311-21.

Li D, Lewinger JP, Gauderman WJ, Murcray CE, Conti D. 2011. Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. Genet Epidemiol.

Liang KY, Huang CY, Beaty TH. 2000. A unified sampling approach for multipoint analysis of qualitative and quantitative traits in sib pairs. Am J Hum Genet 66(5):1631-41.

Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5(2):e1000384.

Maher B. 2008. Personal genomes: The case of the missing heritability. Nature 456(7218):18-21.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A and others. 2009. Finding the missing heritability of complex diseases. Nature 461(7265):747-53.

Morgenthaler S, Thilly WG. 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res 615(1-2):28-56.

Morris AP, Zeggini E. 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genetic Epidemiology 34(2):188-193.

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. 2011. Testing for an unusual distribution of rare variants. PLoS Genet 7(3):e1001322.

Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. 2009. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324(5925):387-9.

Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, Axelrod N, Busam DA, Strausberg RL, Venter JC. 2008. Genetic variation in an individual human exome. PLoS Genet 4(8):e1000160.

Pan W. 2009. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol 33(6):497-507.

Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei LJ, Sunyaev SR. 2010. Pooled association tests for rare variants in exon-resequencing studies. The American Journal of Human Genetics 86(6):832-838.

Price RA, Li WD, Zhao H. 2008. FTO gene SNPs associated with extreme obesity in cases, controls and extremely discordant sister pairs. BMC Med Genet 9:4.

Ramser J, Ahearn ME, Lenski C, Yariz KO, Hellebrand H, von Rhein M, Clark RD, Schmutzler RK, Lichtner P, Hoffman EP and others. 2008. Rare missense and synonymous variants in UBE1 are associated with X-linked infantile spinal muscular atrophy. Am J Hum Genet 82(1):188-93.

Risch N, Zhang H. 1995. Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science 268(5217):1584-9.

Slatkin M. 1999. Disequilibrium mapping of a quantitative-trait locus in an expanding population. Am J Hum Genet 64(6):1764-72.

Tzeng JY, Zhang D. 2007. Haplotype-based association analysis via variance-components score test. Am J Hum Genet 81(5):927-38.

Victor RG, Haley RW, Willett DL, Peshock RM, Vaeth PC, Leonard D, Basit M, Cooper RS, Iannacchione VG, Visscher WA and others. 2004. The Dallas Heart Study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. Am J Cardiol 93(12):1473-80.

Wallace C, Chapman JM, Clayton DG. 2006. Improved power offered by a score test for linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. Am J Hum Genet 78(3):498-504.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89(1):82-93.

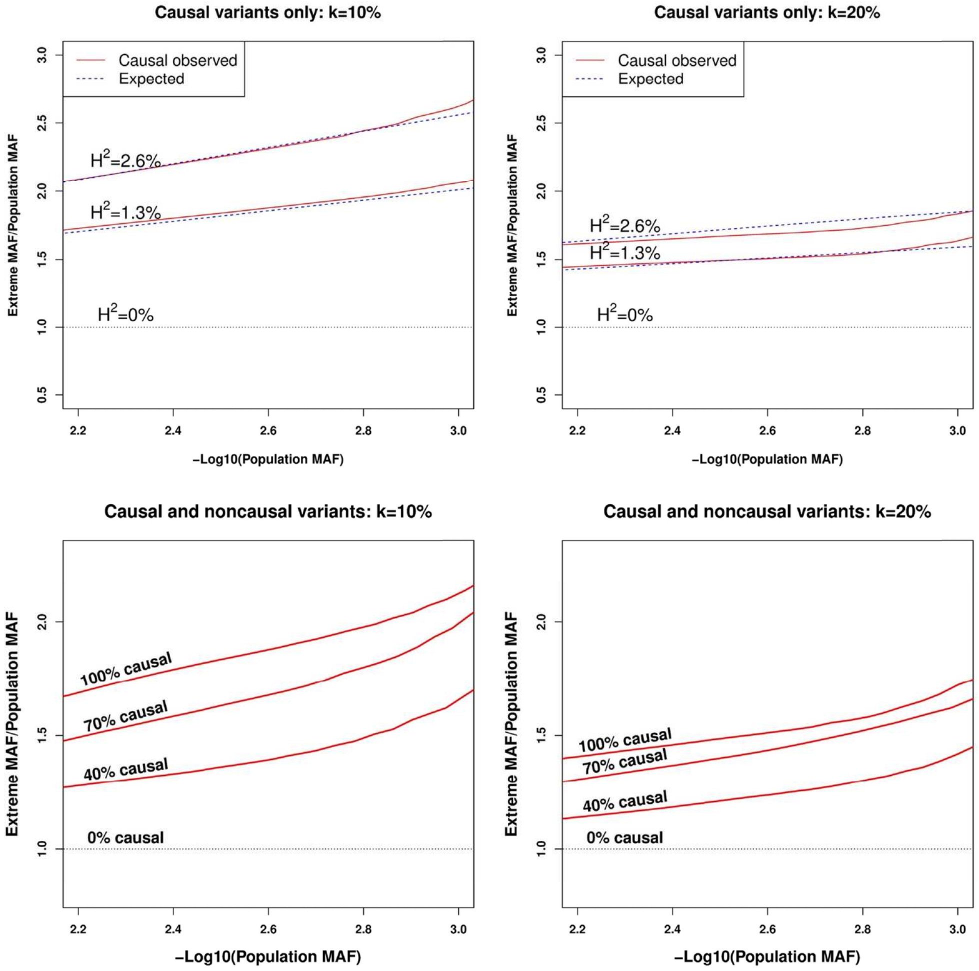


Figure 1: Enrichment of causal rare variants in phenotypic extremes

Estimated folds increase of the observed MAFs of causal variants in phenotypic extremes over population MAFs. The red lines represent the smoothed observed fold increases. The dotted lines represent the theoretical fold increase. For each causal variant, population MAF was computed using the full simulated population while extreme phenotype MAF was computed after sampling the tails. See Supplemental Materials for derivation of theoretical expected MAF for extreme phenotypes. The top two figures consider the case where all variants are causal by sampling k=10% and 20% high/low extremes. For each case, three situations were considered by heritability of causal variants: H2=2.6%, 1.3%, and 0% (no causal variant). Higher heritability gives more enrichment of rare variants. The bottom two figures consider the case where different fractions of variants in a region are causal (100%, 70%, 40% and 0%) by sampling k=10% and 20% high/low extremes. Presence of non-causal variants in a region lower the degree of enrichment of rare variants.
152x152mm (300 x 300 DPI)

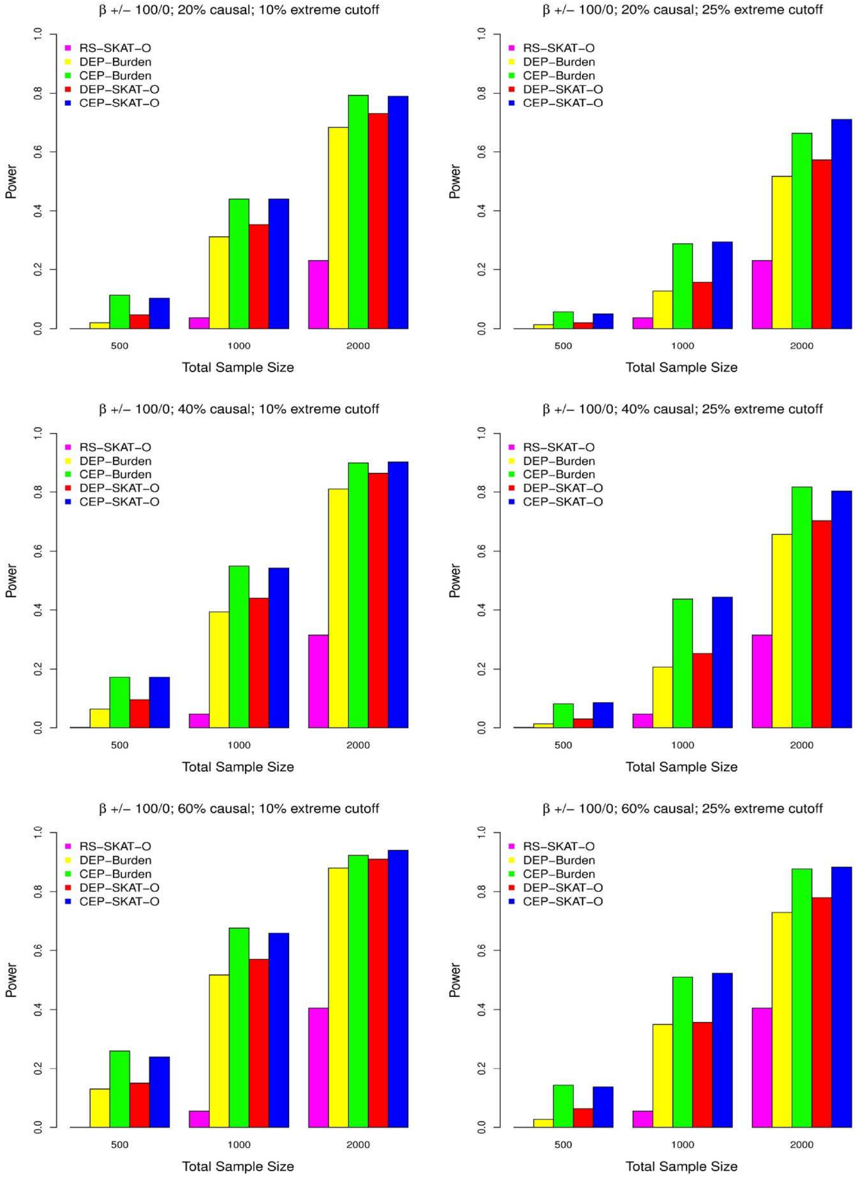


Figure 2: Power comparisons when all causal variants have the same effect direction

Simulated power comparisons between four rare variants association tests with all causal variants having a positive effect on phenotype. The five tests are random sample optimal SKAT (RS-SKAT-O), dichotomized extreme phenotype burden test (DEP-Burden), continuous extreme phenotype burden test (CEP-Burden), dichotomized extreme phenotype optimal SKAT (DEP-SKAT-O), and continuous extreme phenotype optimal SKAT (CEP-SKAT-O). The left panel considers the situation where 10% high/low extremes are sampled with the three rows corresponding to 20% (0.6% heritability), 40% (1.2% heritability) and 60% (1.8% heritability) variants in a 3kb region being causal. Three total sample sizes are considered: n=500, 1000, 2000. The right panel considers the situation where 25% high/low extremes are sampled. Exonic regions are simulated with effect sizes for each causal variant equal to β=-0.2log10MAF. Power is estimated by the proportion of tests that detect an association at the α=10-6 level.
205x277mm (300 x 300 DPI)

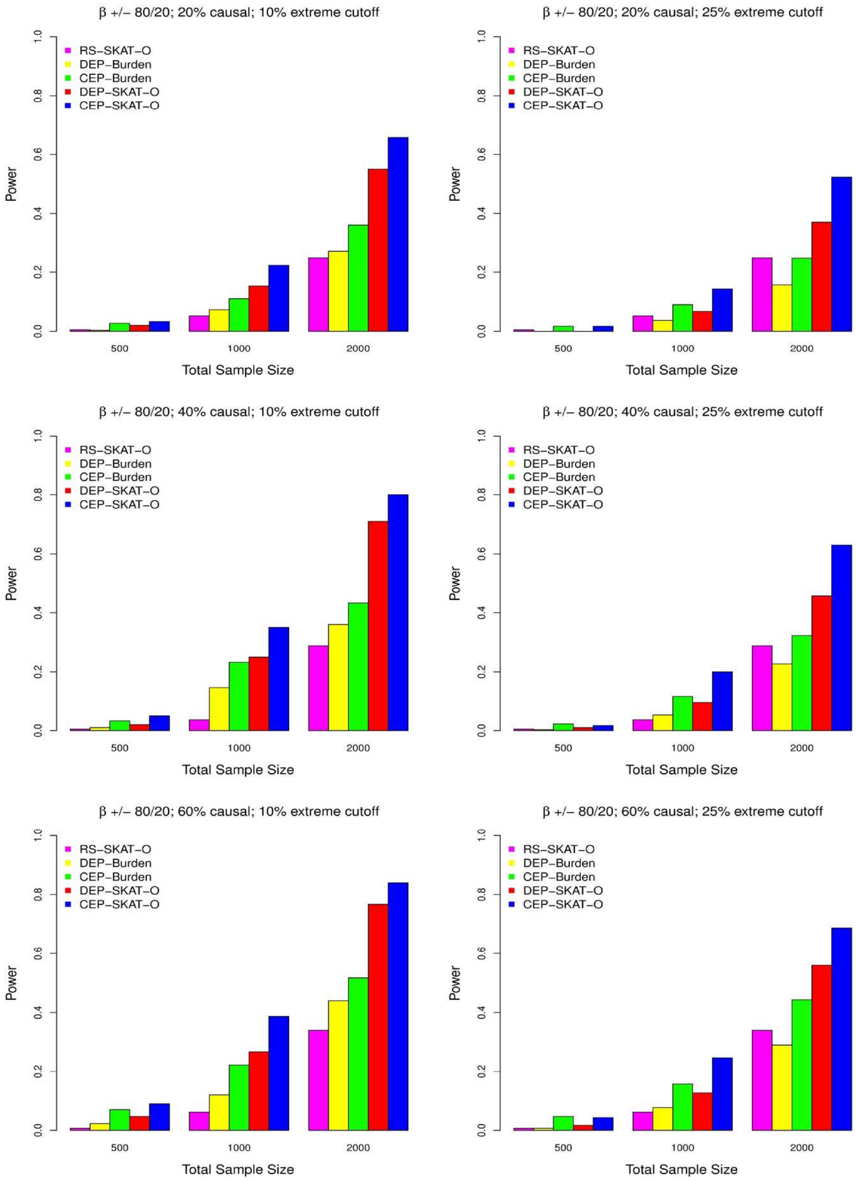


Figure 3: Power comparisons when causal variants have opposite effect directions

Simulated power comparisons between four rare variants association tests with 80% of rare causal variants selected to have a positive effect on phenotype while the remaining 20% have a negative effect. The five tests are random sample SKAT (RS-SKAT-O), dichotomized extreme phenotype burden test (DEP-Burden), continuous extreme phenotype burden test (CEP-Burden), dichotomized extreme phenotype optimal SKAT (DEP-SKAT-O), and continuous extreme phenotype optimal SKAT (CEP-SKAT-O). The left panel considers the situation where 10% high/low extremes are sampled with the three rows corresponding to 20% (0.6% heritability), 40% (1.2% heritability) and 60% (1.8% heritability) variants in a 3kb region being causal. Three total sample sizes are considered: n=500, 1000, 2000. The right panel considers the situation where 25% high/low extremes are sampled. Exonic regions are simulated with effect sizes for each causal variant equal to |β|=-0.2log10MAF with the effect being negated 20% of the time. Power is estimated by the proportion of tests that detect an association at the α=10-6 level.

205x277mm (300 x 300 DPI)

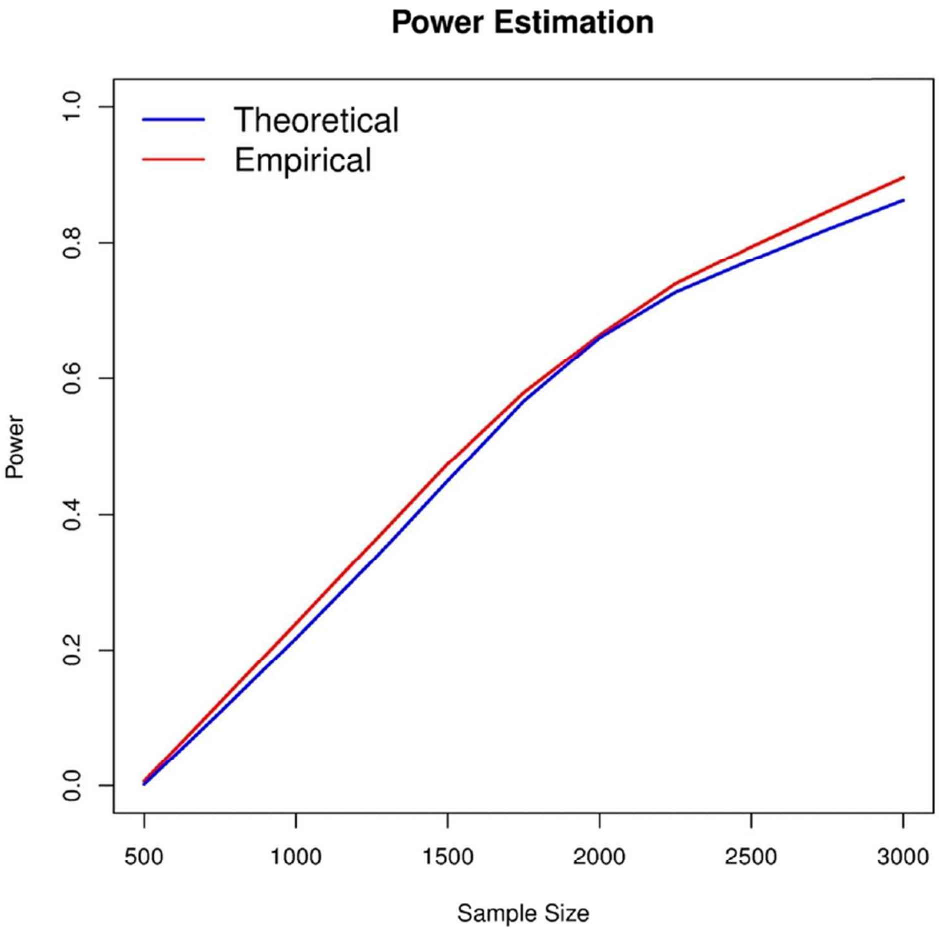


Figure 4: Comparison of theoretical and empirical powers

Estimated power of CEP-SKAT for testing 3kb regions with 20% of variants being causal with all effects in the same direction and the casual variants have effects to |β|=-0.2log10MAF. Theoretical power was calculated as described in section 5 of the Supplementary material, and empirical power was estimated by simulation using 300 replicates. No covariates were considered in either the theoretical or empirical power calculations. Furthermore empirical power was computed using CEP-SKAT without small sample adjustments.
76x76mm (300 x 300 DPI)

Table I: Analysis results of the Dallas Heart Study triglyceride data

n=1389	CEP-SKAT-O	DEP-SKAT-O	RS-SKAT-O	CEP-Burden	DEP-Burden
20%	$5.0x10^{-5}$	$3.0x10^{-5}$	$1.3x10^{-2}$	$1.2x10^{-4}$	$7.2x10^{-5}$
30%	$1.0x10^{-3}$	$1.9x10^{-3}$	$1.3x10^{-2}$	$2.2x10^{-3}$	$2.8x10^{-3}$
40%	$8.9x10^{-3}$	$1.2x10^{-2}$	$1.3x10^{-2}$	$3.4x10^{-3}$	$1.9x10^{-2}$

Analysis results of the Dallas Heart Study sequence data using various test methods and sampling schemes. A total of 3,476 subjects were sequenced. A total 1,389 individuals were selected with highest and lowest 20% logTG levels in each age-gender spectrum from the total 3,476 sequenced subjects. For sampling with higher cutoffs (30% and 40%), 1389 individuals were randomly sub-sampled among the individuals belongs to larger tails to make powers comparable. In these cases, median p-values are presented in the table from 1000 sampling iterations. Since the sample size was large, the small sample adjustment was not applied. The five tests are continuous extreme phenotype SKAT (CEP-SKAT-O), dichotomous extreme phenotype SKAT (DEP-SKAT-O), continuous extreme phenotype burden test (CEP-Burden), dichotomous extreme phenotype burden test (DEP-Burden), and random sample SKAT (RS-SKAT-O).

Detecting Rare Variant Effects Using Extreme Phenotype Sampling in Sequencing Association Studies: Supplementary Materials

Ian Barnett, Seunggeun Lee, and Xihong Lin

1 Rare causal variants are enriched in phenotypic extremes

We consider a phenotype of the ith individual to be modeled as

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

where $\epsilon_i \sim N(0, \sigma^2)$. Here α_0 is an intercept term with $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_m]'$ as the vector of regression coefficients for the covariates $\boldsymbol{X}_i$, and $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_p]'$ as the vector of regression coefficients for the p genetic variants $\boldsymbol{G}_i = (G_{i1}, \cdots, G_{ip})'$. First we will calculate the MAF when sampling extremes where there is a single causal variant and no covariate, i.e., $p = 1$ and $m = 0$. We will then extend the results to multiple causal variants and derive an analytic relationship between the MAF of a causal variant using extreme phenotype sampling to the background population MAF.

Under the single causal variant/no covariate model $y = \beta G + \epsilon$, we desire to show that an individual's probability of having at least one minor allele of the causal variant is increased when their phenotype is extreme. We assume the additive model, but the results can be easily extended to the dominant-recessive model. Without loss of generality, we consider the case where the causal variant has a positive effect on phenotype ($\beta > 0$), and show that for $c > 0$, $Pr(G > 0|y > c) > Pr(G > 0)$.

We first write

$$Pr(G > 0|y > c) = Pr(G > 0)\frac{Pr(y > c|G > 0)}{Pr(y > c)}.$$

Hence showing $Pr(G > 0|y > c) > Pr(G > 0)$ is equivalent to showing $Pr(y > c|G > 0) >$

$Pr(y > c)$. We condition on G to achieve the desired result:

$$\begin{aligned} Pr(y > c|G > 0) - Pr(y > c) &= Pr(y > c|G > 0) - Pr(y > c|G = 0)Pr(G = 0) \\ &\quad -Pr(y > c|G > 0)Pr(G > 0) \\ &= \{Pr(y > c|G > 0) - Pr(y > c|G = 0)\}Pr(G = 0) \\ &> \{Pr(y > c|G = 1) - Pr(y > c|G = 0)\}Pr(G = 0) \\ &= \{Pr(\beta + \epsilon > c) - Pr(\epsilon > c)\}Pr(G = 0) > 0 \end{aligned}$$

Hence the MAF of a causal variant among extreme phenotypes is higher than that in the population. One can easily see that if G is not a causal variant, i.e., $\beta = 0$, then $P(G > 0|y > c) = P(G > 0)$, i.e., the MAF by sampling extremes is the same as in the population.

We next calculate the expected MAF of a causal rare variant in the presence of single or multiple causal variants under extreme phenotype sampling. Specifically, we show below that the expected MAF of a causal variant in extreme phenotype samples can be written as a function of the MAFs of the p causal variants in the background population, the threshold and the regression coefficients β's. Consider the no-covariate model

$$y = \beta_1 G_1 + \cdots + \beta_p G_p + \epsilon. \tag{1}$$

We are interested in estimating $Pr(G_j = g|y > c)$ for $g = 0$, 1, and 2. For simplicity, we assume in our analytic calculations no LD between causal variants which is a plausible assumption when variants are rare. This assumption gives us:

$$Pr(G_1 = g_1, ..., G_p = g_p) = \prod_{l=1}^{p} Pr(G_l = g_l).$$

Note that each $Pr(G_l = g_l)$ can be easily estimated from the data. We model the effect size of the jth causal variant as $\beta_j = -a \cdot log_{10}(MAF_j)$ for some constant $a > 0$, i.e., we assume positive effects.

Write

$$Pr(G_j = g|y > c) = Pr(y > c|G_j = g)\frac{Pr(G_j = g)}{Pr(y > c)}, \tag{2}$$

which means we need just compute $P(y > c)$ and $P(y > c|G_j = g)$. This can be done by conditioning on the remaining causal variants as

$$\begin{aligned}
Pr(y > c) &= Pr(\sum_{l=1}^{p} \beta_l G_l + \epsilon > c) \\
&= \sum_{i_1=0}^{2} \cdots \sum_{i_p=0}^{2} Pr\left\{\sum_{l=1}^{p} \beta_l G_l + \epsilon > c|G_1 = g_1, ..., G_p = g_p\right\} Pr(G_1 = g_1, ..., G_p = g_p) \\
&= \sum_{i_1=0}^{2} \cdots \sum_{g_p=0}^{2} Pr(\sum_{l=1}^{p} \beta_l g_l + \epsilon > c) \prod_{l=1}^{p} P(G_l = g_l) \\
&= \sum_{g_1=0}^{2} \cdots \sum_{g_p=0}^{2} \Phi(\sum_{l=1}^{p} \beta_l g_l - c) \prod_{l=1}^{p} Pr(G_l = g_l) \qquad (3)
\end{aligned}$$

Calculations for $P(y > c|G_j = g)$ are identical except for there being no need to condition on the jth variant:

$$Pr(y > c|G_j = g) = \sum_{g_1=0}^{2} \cdots \sum_{g_{j-1}=0}^{2} \sum_{g_{j+1}=0}^{2} \cdots \sum_{g_p=0}^{2} \Phi(a\beta_j + \sum_{l \neq j} \beta_l g_l - c) \prod_{l \neq j} P(G_l = g_l) \qquad (4)$$

It follows from (2) that we can calculate the expected MAF in extreme phenotype samples as a function of the MAF of the causal variants, the threshold c, and the regression coefficients β_j's in the phenotype model (1) as

$$Pr(G_j > 1|y > c) = E(G_j|y > c)/2 = 0 * Pr(G_j = 0|y > c) + 0.5 * Pr(G_j = 1|y > c) + 1 * Pr(G_j = 2|y > c).$$

One can also use equations (2) and (4) to easily show that $Pr(G_j = g|y > c) > Pr(G_j = g)$ if β's are not equal to 0, i.e., the MAF of a causal variant is higher in extreme phenotype samples than their population counterpart.

2 Null distribution of Continuous Extreme Phenotype SKAT

Suppose the true NULL model is

$$y_i = \boldsymbol{X}_i' \boldsymbol{\alpha} + \epsilon \qquad (5)$$

where $\boldsymbol{X}_i = (x_{i0}, x_{i1}, \ldots, x_{im})$ is the covariates of i^{th} individual with $x_{i0} = 1$, and $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_m)'$ is a vector of regression coefficients of $\boldsymbol{X}_i$, and $\epsilon \sim N(0, \sigma^2)$. Note that we use a slightly different notation in which $\boldsymbol{X}_i$ includes the intercept. Suppose we select n samples with either $y_i > c_1$ or $y_i < c_2$, and denote the selected y_i as y_i^* . For notational simplicity, we use i to indicate the selected individuals. Under the null hypothesis, y_i^* follows a truncated Gaussian distribution with the density function

$$f(y_i^*) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{exp\{-(y_i^* - \boldsymbol{X}_i'\boldsymbol{\alpha})^2/2\sigma^2\}}{\Phi(t_{i2}) + 1 - \Phi(t_{i1})}, \qquad (6)$$

where $t_{i1} = (c_1 - \boldsymbol{X}_i'\boldsymbol{\alpha})/\sigma$ and $t_{i2} = (c_2 - \boldsymbol{X}_i'\boldsymbol{\alpha}))/\sigma$. The first derivative of log likelihood function is

$$u_j = \frac{\partial \ell}{\partial \alpha_j} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ij}(y_i - \boldsymbol{X}_i'\boldsymbol{\alpha} + m_i),$$

and the second derivative is

$$J_{ik} = \frac{\partial^2 \ell}{\partial \alpha_j \partial \alpha_k} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ij} x_{ik}(-1 + v_i),$$

where

$$m_i = \sigma \frac{\phi(t_{i2}) - \phi(t_{i1})}{\Phi(t_{i2}) + 1 - \Phi(t_{i1})}, \quad \text{and} \quad v_i = \frac{t_{i2}\phi(t_{i2}) - t_{i1}\phi(t_{i1})}{\Phi(t_{i2}) + 1 - \Phi(t_{i1})} + \frac{m_i^2}{\sigma^2}.$$

Define $\mathbf{S} = -\mathbf{J}$, $\mathbf{y}^* = (y_1^*, \ldots, y_n^*)'$, $\mathbf{u} = (u_0, \ldots, u_m)'$, and $\mathbf{m} = (m_1, \ldots, m_n)'$. By the Fisher Scoring (or Newton Raphson) procedure, new $\boldsymbol{\alpha}$ is

$$\boldsymbol{\alpha}^* = \boldsymbol{\alpha} + \mathbf{S}^{-1}\mathbf{u},$$

hence $\mathbf{S}(\alpha^* - \alpha) = \mathbf{u}$. Since $\mathbf{S} = \boldsymbol{X}'\mathbf{V}\boldsymbol{X}/\sigma^2$, where $\mathbf{V} = diag\{(1 - v_i)\}$ and $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots \boldsymbol{X}_n]'$,

$$\boldsymbol{X}'\mathbf{V}\boldsymbol{X}(\boldsymbol{\alpha}^* - \boldsymbol{\alpha}) = \boldsymbol{X}'(\mathbf{y}^* - \boldsymbol{X}\boldsymbol{\alpha} - \mathbf{m}).$$

Now we can treat the Fisher scoring algorithm as a weighted least square problem. Define a working vector

$$\tilde{\mathbf{y}} = \boldsymbol{X}\boldsymbol{\alpha} + \mathbf{V}^{-1}(\mathbf{y}^* - \boldsymbol{X}\boldsymbol{\alpha} - \mathbf{m}),$$

and then $\boldsymbol{\alpha}^*$ is a weighted least square estimtor of the linear model $\widetilde{\mathbf{y}} = \boldsymbol{X}\boldsymbol{\alpha} + \widetilde{\boldsymbol{\epsilon}}$ with $E(\widetilde{\boldsymbol{\epsilon}}) = 0$ and $Var(\widetilde{\boldsymbol{\epsilon}}) = \mathbf{V}^{-1}$. Since $E(y_i^*) = \boldsymbol{X}_i'\boldsymbol{\alpha} - m_i$, the SKAT test statistic with linear weighted kernel is

$$\begin{aligned} Q_S &= (\mathbf{y}^* - \widehat{\mu})'\boldsymbol{G}\mathbf{W}\boldsymbol{G}'(\mathbf{y}^* - \widehat{\mu}) = (\widetilde{\mathbf{Y}} - \boldsymbol{X}\boldsymbol{\alpha}^*)'\mathbf{V}\boldsymbol{G}\mathbf{W}\boldsymbol{G}'\mathbf{V}(\widetilde{\mathbf{Y}} - \boldsymbol{X}\boldsymbol{\alpha}^*) \\ &= \widetilde{\mathbf{Y}}\mathbf{P}_0\boldsymbol{G}\mathbf{W}\boldsymbol{G}'\mathbf{P}_0\widetilde{\mathbf{Y}}, \end{aligned}$$

where $\mathbf{P}_0 = \mathbf{V} - \mathbf{V}\boldsymbol{X}(\boldsymbol{X}'\mathbf{V}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{V}$. From $Var(\widetilde{y}_i) = (1 - v_i)\sigma^2$, the asymptotic null distribution of Q_S is

$$\sum \lambda_v \chi_v^2,$$

where λ_v is the v^{th} eigenvalue of $\hat{\sigma}^2\mathbf{P}_0^{1/2}\boldsymbol{G}\mathbf{W}\boldsymbol{G}'\mathbf{P}_0^{1/2}$.

Calculations of Q_S require fitting the null model (5) using extreme phenotypes y_i^* under truncated normal likelihood (6). The Newton-Raphson method can be used to estimate $\boldsymbol{\alpha}$ and σ^2.

3 Null distribution of the optimal unified test for continuous extreme phenotype

Suppose

$$Q_\rho = (1 - \rho)Q_S + \rho Q_B,$$

which is the test statistic of the proposed unified test, where Q_S is the SKAT statistic and Q_B is the burden test statistic. The class of test statistics Q_ρ includes both SKAT ($\rho = 0$) and the burden test ($\rho = 1$) as special cases, and p_ρ is a p-value computed based on Q_ρ. Then, the test statistic of the optimal test is

$$T = min\{p_{\rho_1}, \ldots, p_{\rho_b}\}, \quad 0 = \rho_1 < \rho_2 < \ldots < \rho_b = 1. \tag{7}$$

Define $\mathbf{Z} = \mathbf{V}^{-1/2}\boldsymbol{G}\mathbf{W}$, $\bar{\mathbf{z}} = (\bar{z}_1, \ldots, \bar{z}_n)'$, where $\bar{\mathbf{z}}_i = \sum_{j=1}^{p} z_{ij}/p$, and $\mathbf{M} = \bar{\mathbf{z}}(\bar{\mathbf{z}}'\bar{\mathbf{z}})^{-1}\bar{\mathbf{z}}'$. We further let

$$\tau(\rho) = p^2\rho\bar{\mathbf{z}}'\bar{\mathbf{z}} + \frac{1-\rho}{\bar{\mathbf{z}}'\bar{\mathbf{z}}}\sum_{j=1}^{p}(\bar{\mathbf{z}}'\mathbf{z}_{.j})^2,$$

where $\mathbf{z}_{.j}$ is the j^{th} column of $\mathbf{Z}$. Following the same argument in Lee *et al.*(2012), it can be shown that Q_ρ is asymptotically equivalent as

$$(1-\rho)(\sum_{k=1}^{m} \widetilde{\lambda}_k \eta_k + \zeta) + \tau(\rho)\eta_0, \tag{8}$$

where $\{\widetilde{\lambda}_1, \ldots \widetilde{\lambda}_m\}$ are non-zero eigenvalues of $\mathbf{Z}'(\mathbf{I}-\mathbf{M})\mathbf{Z}$, $\eta_k(k=0,\ldots,m)$ are i.i.d χ_1^2 random variables, and ζ satisfies the following conditions:

$$E(\zeta)=0, \quad Var(\zeta) = 4trace(\mathbf{Z}'\mathbf{M}\mathbf{Z}\mathbf{Z}'(\mathbf{I}-\mathbf{M})\mathbf{Z}),$$
$$Corr(\sum_{k=1}^{m} \lambda_k \eta_k, \zeta) = 0, \quad \text{and} \quad Corr(\eta_0, \zeta) = 0.$$

It shows that the Q_ρs are mixtures of shared random variables, and the only differences among different Q_ρs are the mixing coefficients. From this fact, a p-value of T can be efficiently computed through one dimensional numerical integration. Details can be found in Lee *et al.*(2012).[2]

4 Small Sample Adjustment

It is known that the SKAT family tests can produce conservative results when the trait is binary and the sample size is small. The same conservative pattern can be observed when we test extreme continuous phenotypes. To resolve this issue, we adopt the same strategy as that in Lee *et al.*(2012)[1] in which we adjust asymptotic null distribution of the test statistics by estimating small sample variance and kurtosis. To estimate these moments, we generate resampled test statistics using the parametric bootstrap approach. In particular, B sets of truncated normal random variables are generated from the model (6) with estimated $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\sigma}$ under the null hypothesis, and the variance and kurtosis of Q_S and Q_B are estimated using resampled phenotype sets. Then, we apply the same algorithm in Lee *et al.*(2012).[1]

5 Theoretical Power Calculation

Power calculation derivations are available for SKAT and SKAT-O, but adjustments need to be made to account for extreme phenotype sampling. Derivations for power calculations for continuous extreme phenotype SKAT (CEP-SKAT-O) mirror Lee *et al.*(2012) in their calculation for continuous phenotypes, but a key distinction in the treatment of the genotype matrix needs to be made. By sampling from the extremes, causal variants tend to occur more often in the sample than observed in the population, and the power calculation should reflect this bias appropriately. In a random sample the MAF of all sampled variants should be consistent for their respective population MAFs (note that rate of convergence is slow for rare variants). In an EPS sample, consistency is not achieved and the likelihood of sampling a genotype must be adjusted accordingly in the power calculation.

To account for this biased sampling of genotypes, a reduced genotype matrix $\boldsymbol{G}_R$ is used in the calculation of the SKAT statistic. We define $\boldsymbol{G}_R = \mathbf{B}\boldsymbol{G}$ where $\mathbf{B}$ is an n by n diagonal matrix with jth diagonal equal to $\sqrt{P(\text{Variant j exists in the EPS sample})}$. We can calculate $\mathbf{B}$ given the upper and lower cutoffs for sampling extreme phenotypes by taking the tail probabilities beyond these cutoffs of the normal distribution with a mean of the jth entry of $\boldsymbol{G}\beta$ and a variance of 1. The resulting test statistic is:

$$Q_S = (\mathbf{y}^* - \widehat{\mu}_R)'\boldsymbol{G}_R\mathbf{W}_\rho\boldsymbol{G}'_R(\mathbf{y}^* - \widehat{\mu}_R)$$

where $\widehat{\mu}_R$ is the expected value of the truncated normal $\mathbf{y}^*$ and is a function of $\boldsymbol{G}_R$. $\mathbf{W}_\rho$ is the matrix of weights adjusted by ρ through the matrix $\mathbf{R}_\rho$ as done in Lee *et al.*(2012)[2] who also recommend to approximate ρ using the percent of causal variants and percent of variants with a positive effect on phenotype. The distribution of Q_S under the alternative hypothesis can then be approximated by a non-central χ^2 distribution. The distribution of Q_S under the null hypothesis is approximated in a similar manner except under the assumption of no variant effects. Power is estimated by the area in the upper tail of the alternative distribution that lies above the critical value taken from the null distribution.

Power estimates are obtained by averaging the estimated power over many randomly selected regions of equivalent size (we selected 3kb regions) in order to account for variability of genotypes by region. In each region a new $\boldsymbol{G}_R$ is chosen with individuals selected based on their probability of being observed in the phenotypic extremes of the sample given their genotypes. We compare the theoretical powers and empirical powers in Figure 4 and Supplementary Figure 3.

6 DHS data analysis sensitivity to different cutoffs

In the Dallas Heart Study, we examine how the p-values for all the EPS methods are affected by altering the extreme phenotype cutoff, and results are presented in Supplementary Figure 1. DEP-N represents DEP-Burden test in the main manuscript. Note that two additional burden tests, DEP-W and DEP-C are included. DEP-W uses a weighted count with beta(1,25) weights while DEP-C is an adaptation of CAST for EPS. We range the cutoffs from 15% to 30% in 1% increments to capture the sensitivity of the tests to different cutoffs. From the spikes in each methods p-values over slight changes in this cutoff value, it is clear that inclusion or exclusion of certain individuals could affect the overall significance. It is important to note that because we have a fixed sample size of 3476, smaller cutoffs lead a smaller sizes through EPS. As an example, a tail cutoff of 15% leads to half the EPS sample size that we would see with a tail cutoff of 30%. Because significance is sensitive to sample size, direct comparison between p-values at different cutoffs is not appropriate in Supplemental Figure 1. This analysis is presented for each of ANGPTL3, ANGPTL4, and ANGPTL5 being tested for separately as well as a combined analysis across the three genes.

For the three-gene combined analysis, CEP-SKAT-O gives the smallest p-value than the other methods when the cutoff of selecting extreme phenotypes is less than 22% and slightly higher p-values than DEP-SKAT-O when the cutoff is greater than 22%. For ANGPL3 gene, CEP-SKAT-O overall has the smallest p-compared with the other methods. For ANGPTL4 and ANGPTL5 genes, CEP-SKAT-O has similar p-values to DEP-SKAT-O. Both outperform the burden tests.

References

[1] S. Lee, M. Emonds, M. Bamshad, K. Barnes, M. Rieder, D. Nickerson, D. Christiani, M. Wurfel, and X. Lin. Optimal unified approach for rare variant association testing with application to small sample case-control whole-exome sequencing studies. *American Journal of Human Genetics*, 2012.

[2] S. Lee, M.C. Wu, and X. Lin. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 2012.

7 Supplemental Tables and Figures

Supplemental Table 1: Type I error estimates for Continuous Extreme Phenotype (CEP-SKAT-O)

Phenotypes were simulated under the null model (5) using two covariates and added Gaussian noise, but with no genotype effects. Estimates are based on 20 million simulated p-values. Adjusted SKAT-O was used to adjust the p-value for small sample size.

α-level	0.05	0.01	1×10^{-5}	2.5×10^{-6}	1×10^{-6}
n=500	0.0471	0.0097	1.66×10^{-5}	4.22×10^{-6}	1.70×10^{-6}
n=1000	0.0480	0.0100	1.48×10^{-5}	4.70×10^{-6}	2.19×10^{-6}
n=2000	0.0497	0.0104	1.70×10^{-5}	4.75×10^{-6}	2.21×10^{-6}

Supplemental Table 2: Analysis of the Dallas Heart Study triglyceride data

Analysis results of the Dallas Heart Study (DHS) sequence data using various test methods and sampling schemes. The DHS sequenced 3,476 subjects. A total 1,389 individuals were selected with highest and lowest 20% logTG levels in each age-gender spectrum. For sampling with higher cutoffs (30% and 40%), 1389 individuals were randomly sub-sampled among the individuals belongs to larger tails. In these cases, median p-values are presented in the table from 1000 sampling iterations. Since the sample size was large, the small sample adjustment was not applied. DEP-N represents DEP-Burden test in the main manuscript. Two additional burden tests are included: DEP-W uses a weighted count while DEP-C is an adaptation of CAST for EPS. The other tests are described in Table 1.

n=1389	CEP-SKAT-O	DEP-SKAT-O	RS-SKAT-O	DEP-W	DEP-N	DEP-C
20%	5.0×10^{-5}	3.0×10^{-5}	1.3×10^{-2}	2.5×10^{-4}	7.2×10^{-5}	2.1×10^{-3}
30%	1.0×10^{-3}	1.9×10^{-3}	1.3×10^{-2}	2.0×10^{-3}	2.8×10^{-3}	1.7×10^{-2}
40%	8.9×10^{-3}	1.2×10^{-2}	1.3×10^{-2}	1.2×10^{-2}	1.9×10^{-2}	4.5×10^{-2}

Supplemental Figure 1: EPS test sensitivity to extreme cutoff in DHS triglyceride data

The p-values using six EPS association tests using different extreme cutoffs are demonstrated. Each of the three genes, ANGPTL3, ANGPTL4, and ANGPTL5 are tested separately. A test combining all three genes is also included.

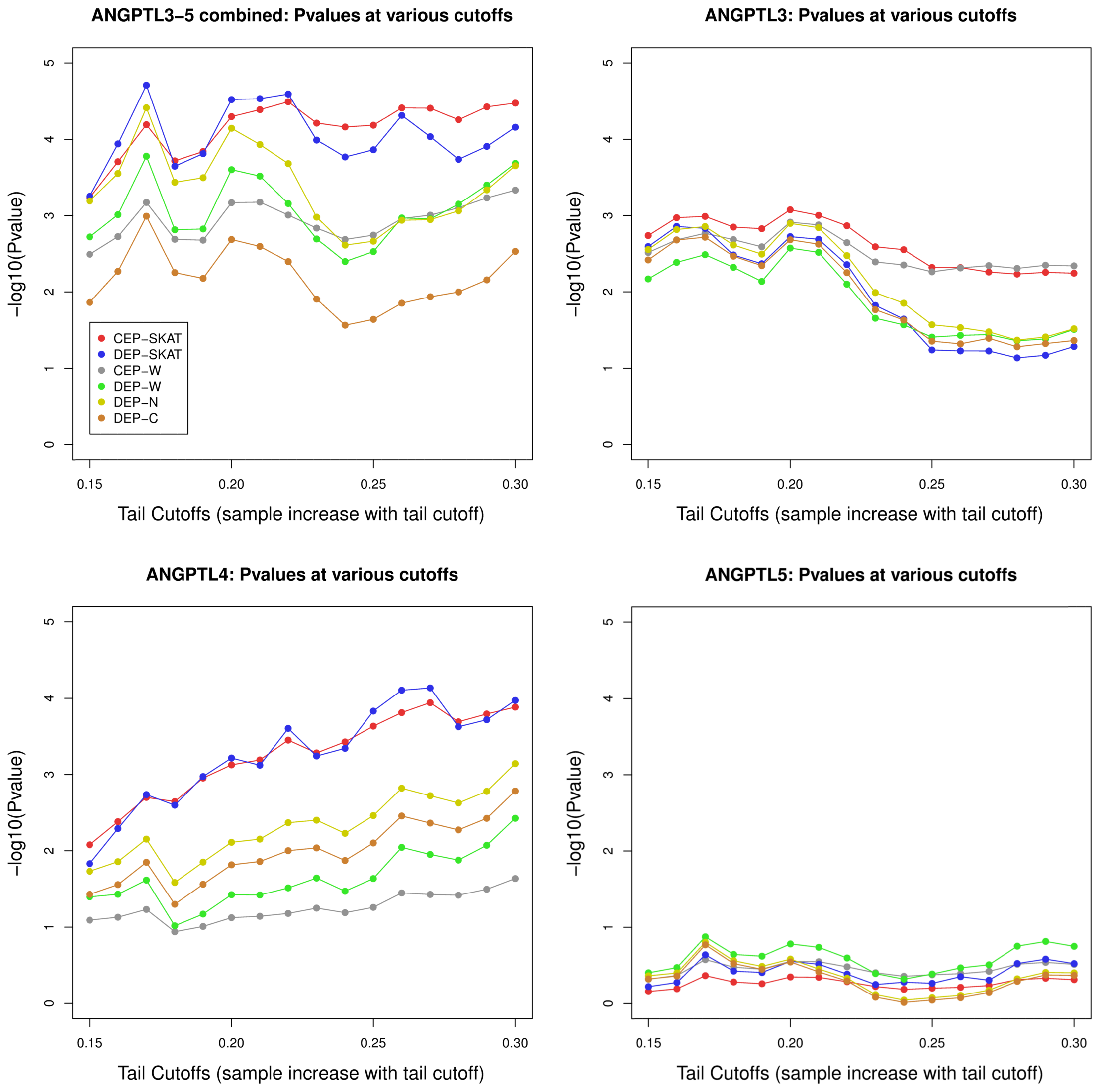
Supplemental Figure 2: Power comparisons with constant effect sizes

Simulated power comparisons between four rare variants association tests with all causal variants having a positive effect on phenotype. The five tests are random sample optimal SKAT (RS-SKAT-O), dichotomized extreme phenotype burden test (DEP-Burden), continuous extreme phenotype burden test (CEP-Burden), dichotomized extreme phenotype optimal SKAT (DEP-SKAT-O), and continuous extreme phenotype optimal SKAT (CEP-SKAT-O). The left panel considers the situation where 10% high/low extremes are sampled with the three rows corresponding to 20% (0.6% heritability), 40% (1.2% heritability) and 60% (1.8% heritability) variants in a 3kb region being causal. Three total sample sizes are considered: n=500, 1000, 2000. The right panel considers the situation where 25% high/low extremes are sampled. Exonic regions are simulated with effect sizes for each causal variant equal to $\beta = 1$. Power is estimated by the proportion of tests that detect an association at the $\alpha = 10^{-6}$ level.

Supplemental Figure 3: Additional comparison of theoretical and empirical power for CEP-SKAT-O

In this setting, 60% of variants were considered causal in a 3kb region. Theoretical power for optimal continuous extreme phenotype SKAT (CEP-SKAT-O) is compared with the empirical power estimated using 300 simulations for each estimate. Four settings are considered: sampling 10% and 20% high/low extreme phenotypes; 80%/20% causal variants have positive/negative effects and 100% causal variants have positive effects.

Supplemental Figure 1

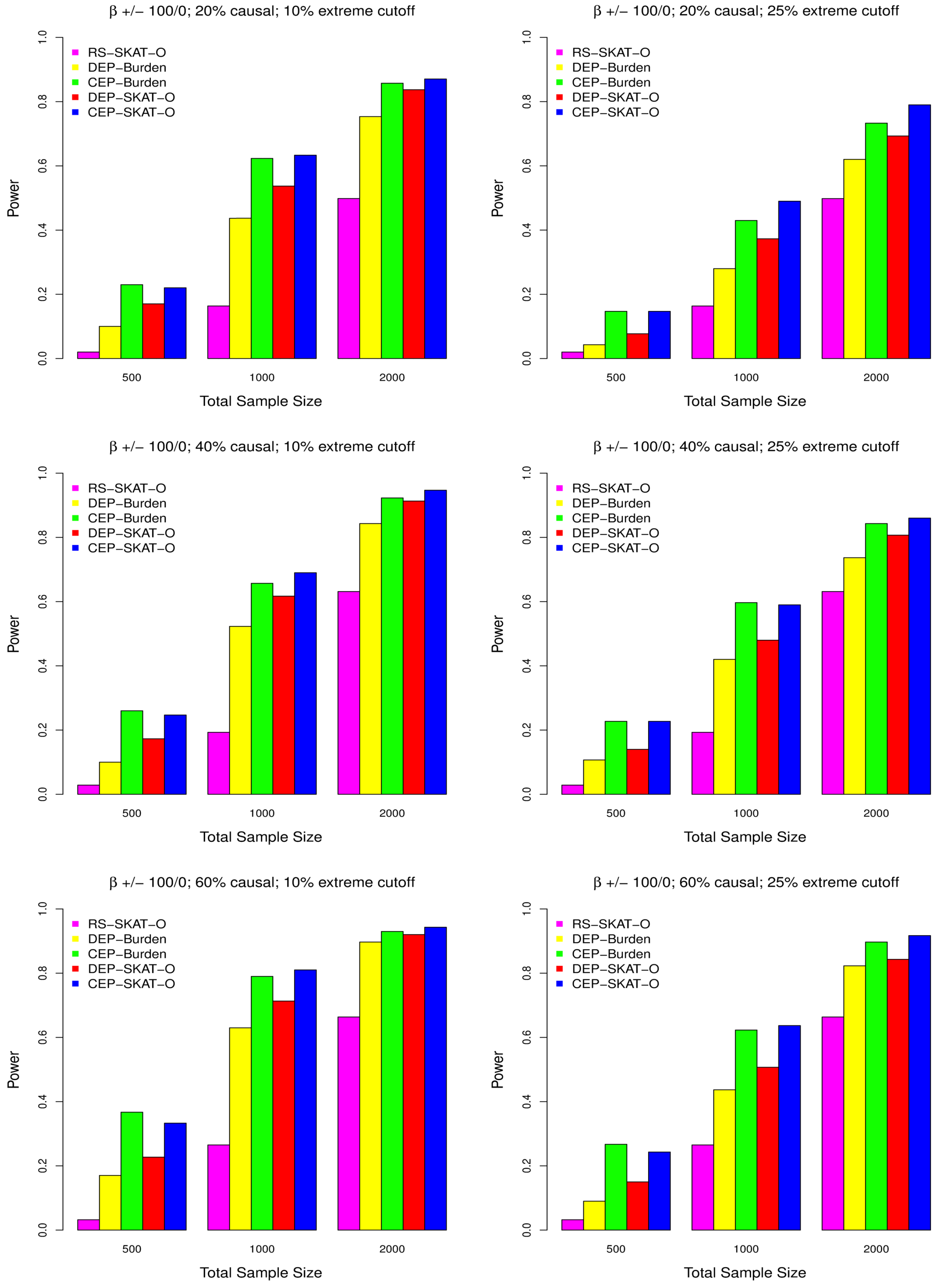

ANGPTL3–5 combined: Pvalues at various cutoffs

ANGPTL3: Pvalues at various cutoffs

ANGPTL4: Pvalues at various cutoffs

ANGPTL5: Pvalues at various cutoffs

Supplemental Figure 2



β +/– 100/0; 20% causal; 10% extreme cutoff

β +/– 100/0; 20% causal; 25% extreme cutoff

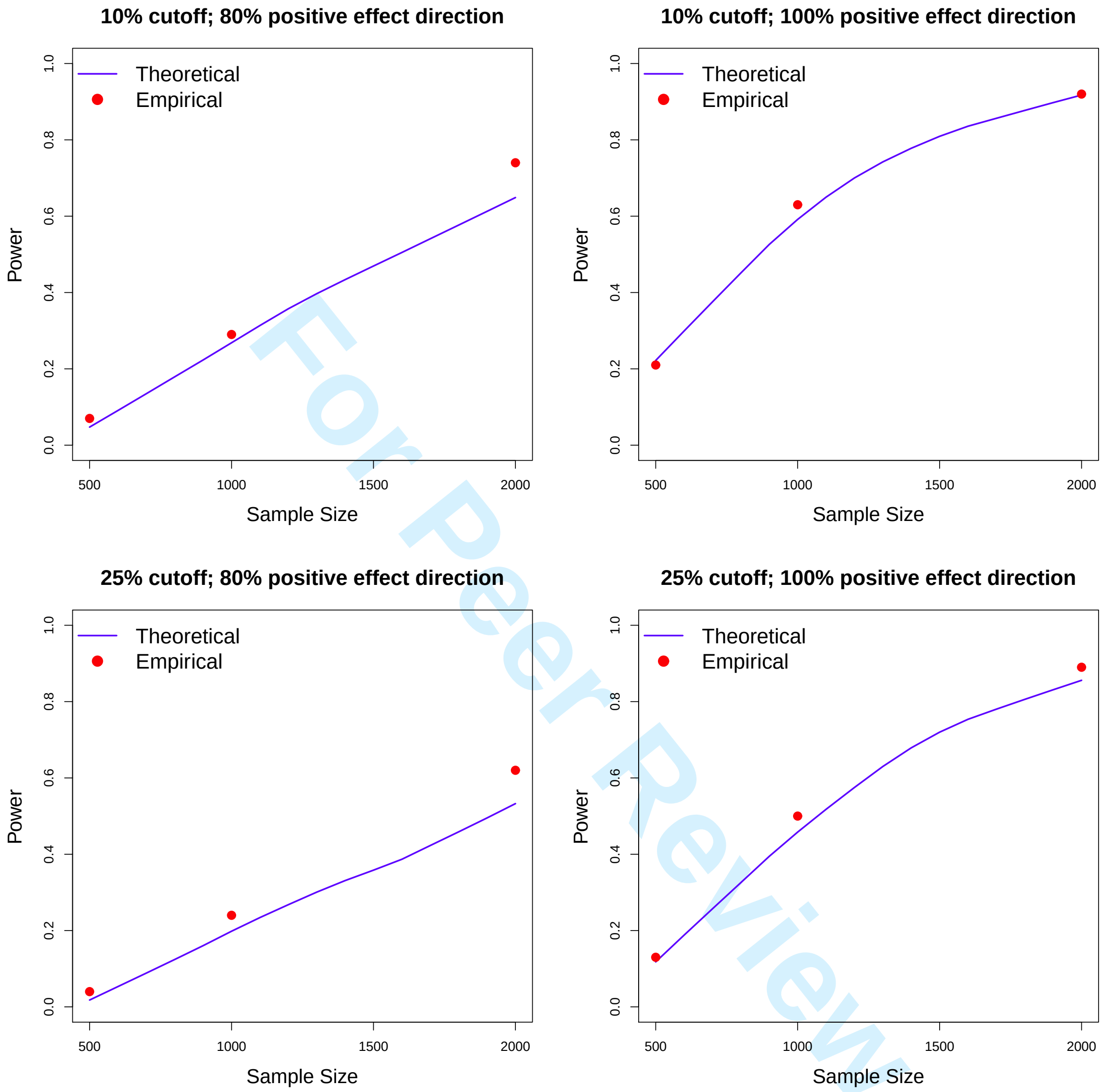
β +/– 100/0; 40% causal; 10% extreme cutoff

β +/– 100/0; 40% causal; 25% extreme cutoff

β +/– 100/0; 60% causal; 10% extreme cutoff

β +/– 100/0; 60% causal; 25% extreme cutoff

Supplemental Figure 3


10% cutoff; 80% positive effect direction
10% cutoff; 100% positive effect direction
25% cutoff; 80% positive effect direction
25% cutoff; 100% positive effect direction
Theoretical
Empirical
Power
Sample Size