Family-based association tests for sequence data, and comparisons with population-based association tests

Iuliana Ionita-Laza[1,*], Seunggeun Lee[2], Vladimir Makarov[3],

Joseph D. Buxbaum[3,4,5], and Xihong Lin[2,*]

[1] Department of Biostatistics, Columbia University, New York, NY 10032

[2] Department of Biostatistics, Harvard University, Boston, MA 02115

[3] Seaver Autism Center for Research and Treatment, Mount Sinai School of Medicine, New York, NY 10029

[4] Department of Psychiatry, Mount Sinai School of Medicine, New York, NY 10029

[5] Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY 10029

[*] Corresponding authors: ii2135@columbia.edu, xlin@hsph.harvard.edu

Running Title: Family- and Population-based association tests

Keywords: family- and population-based association tests, sequence data

Abstract

Recent advances in high-throughput sequencing technologies make it increasingly more efficient to sequence large cohorts for many complex traits. We discuss here a class of sequence-based association tests for family-based designs that corresponds naturally to previously proposed population-based tests, including the classical burden and variance-component tests. This framework allows for a direct comparison between the power of sequence-based association tests with family- vs. population-based designs. We show that, for dichotomous traits using family-based controls results in similar power levels as the population-based design (although at an increased sequencing cost for the family-based design), while for continuous traits (in random samples, no ascertainment) the population-based design can be substantially more powerful. A possible disadvantage of population-based designs is that they can lead to increased false-positive rates in the presence of population stratification, while the family-based designs are robust to population stratification. We show also an application to a small exome-sequencing family-based study on autism spectrum disorders. The tests are implemented in publicly available software.

Introduction

Recent advances in high-throughput sequencing technologies and the availability of large study populations for many complex traits promise to lead to significant progress in understanding the genetic basis of common diseases ([1], [2]). Such progress is critically dependent on choice of efficient study design and statistical methods. In genome-wide association stud-

ies (GWAS), the population-based design has been widely used due to the intrinsic ease of collecting large datasets needed to identify disease susceptibility variants of small effects ([3]). The family-based design has therefore been less popular. However, family-based designs have important advantages, including well-known robustness to population stratification, and ability to identify technological artifacts in the data. Furthermore, family-based designs allow testing of hypotheses that are difficult to test with unrelated individuals ([4]). For example, they are indispensable in the study of de-novo variation, and can therefore be a powerful design for complex traits that have an important de-novo component, as it is believed to be the case for autism spectrum disorders ([5], [6]). They also allow testing of parent-of-origin effects ([7]).

Many tests have been proposed for population-based designs ([8]−[22]), and among them two main classes of tests can be distinguished: the burden test ([12]) and the variance-component test ([19]). Comparatively, for family-based designs there has been relatively little development. An extension of the family-based association test (FBAT [23]) to sequence data has been recently proposed, and corresponds naturally to the population-based burden test (De et al. unpublished).

We introduce here a class of family-based association tests that includes the burden and the variance-component tests as particular cases, and have natural correspondence to existing tests for population-based designs ([24]). Both the burden and the variance-component tests test the null hypothesis of no genetic variant in the region being associated with disease. However, they make different assumptions on the distribution of effect sizes, and therefore

their performance depends on the underlying disease model. In particular, the burden test tends to be more powerful when a large proportion of genetic variants in the region are associated with disease, while the variance-component test tends to be more powerful when the proportion of disease associated variants in a region is small, and/or there are both risk and protective variants in the region being tested. These tests are applicable to different family structures, including nuclear families and sibships. We therefore also perform a direct comparison of the power of sequence-based association tests for the two types of designs.

Methods

SKAT for Family-based Designs

Although the methods we present are applicable to more general family structures (including nuclear families), for the sake of simplicity we choose to show the theoretical derivations for the simplest family design, namely the trio design. We assume that n trios (one offspring and the two biological parents) have been sequenced in a region of interest, G, such as a gene. For the ith trio, we assume the offspring trait is denoted by Y_i and the offspring genotype at the jth variant in G is coded as X_{ij} $(1 \leq j \leq m)$. We assume a generalized linear model that relates the trait value Y to the genotype data:

$$h[E(Y_i)] \;=\; \alpha_0 + \alpha_1 C_{i1} + \cdots + \alpha_p C_{ip} + \beta_1 X_{i1} + \cdots + \beta_m X_{im},$$

where $h(\cdot)$ is the corresponding link function, and can be the identity function when traits are continuous, or the logistic function when traits are dichotomous; $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)'$ are regression coefficients for the covariates $\mathbf{C_i} = (C_{i1}, \ldots, C_{ip})$ that we want to adjust for. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)'$.

To test the null hypothesis of no genetic effects

$$H_0 : \quad \boldsymbol{\beta} = 0$$

we assume that each β_j follows an arbitrary distribution with mean 0 and variance $w_j^2\tau$, i.e. $E(\beta_j) = 0$ and $Var(\beta_j) = w_j^2\tau$. Then to test $H_0 : \ \tau = 0$ we can use the variance-component score statistic proposed in [25]:

$$Q \quad = \quad (\mathbf{Y} - \widehat{\boldsymbol{\mu_0}})'\tilde{\mathbf{K}}(\mathbf{Y} - \widehat{\boldsymbol{\mu_0}}),$$

where for continuous traits $\widehat{\boldsymbol{\mu_0}} = \widehat{\alpha_0} + \mathbf{C}\widehat{\boldsymbol{\alpha}}$, and for dichotomous traits $\widehat{\boldsymbol{\mu_0}} = \text{logit}^{-1}(\widehat{\alpha_0} + \mathbf{C}\widehat{\boldsymbol{\alpha}})$; $\mathbf{C}$ is the $n \times p$ covariate matrix; $\mathbf{Y}$ is the vector of phenotype values for all the offspring in the dataset. Also, for the weighted-linear kernel:

$$\tilde{\mathbf{K}} = [\mathbf{X} - E(\mathbf{X}|\mathbf{X_p})]\mathbf{W}\mathbf{W}[\mathbf{X} - E(\mathbf{X}|\mathbf{X_p})]',$$

where $\mathbf{X}$ is the (n, m) genotype matrix, $\mathbf{X_p}$ is the parental genotype data, and $\mathbf{W} = \text{diag}(w_1, \ldots, w_m)$. w_j $(j = 1 \ldots m)$ represent variant weights that can be chosen to de-

pend on the data, or can be external weights, e.g. reflecting functional predictions. As in [19] we take $w_j = \text{Beta}(\widehat{f}_j; 1, 25)$, where $\widehat{f}_j$ is the estimated variant frequency based on parental genotypes alone. Under the null hypothesis, $E(\mathbf{X}|\mathbf{X_p})$ can be calculated using the laws of Mendelian transmission. When parental genotypes are not completely known, and other family structures such as sibships are available, Rabinowitz and Laird ([26]) have developed an algorithm that specifies the distribution of offspring genotypes conditional on the sufficient statistic for the parental genotypes.

Q has a simple expression:

$$Q = \sum_{j=1}^{m} w_j^2 \left[\sum_{i=1}^{N} (Y_i - \widehat{\mu_{i,0}})(X_{ij} - E(X_{ij}|X_{ij}^p)) \right]^2,$$

where X_{ij}^p is the parental genotype data for family i at variant j.

The main difference between this family-based test and its population-based counterpart comes from the specification of the null distribution of Q. Unlike the case for population-based tests, for the family-based test we condition on the parental genotypes $\mathbf{X_p}$ (or the sufficient statistic, when parental genotypes are not available) and on the trait values $\mathbf{Y}$ and treat the offspring genotypes $\mathbf{X}$ as random. If the assumption that $(\mathbf{X} - E(\mathbf{X}|\mathbf{X_p}))'(\mathbf{Y} - \widehat{\boldsymbol{\mu_0}})$ is multivariate normal holds, then it can be shown that the null distribution of Q can be approximated by a mixture of chi-square distributions as follows:

$$\sum_{j=1}^{m} \lambda_j \chi_{1,j}^2,$$

where $(\lambda_1, \ldots, \lambda_m)$ are the eigenvalues of matrix $\mathbf{A}^{1/2}\mathbf{L}'\mathbf{W}\mathbf{W}\mathbf{L}\mathbf{A}^{1/2}$ with

$$Cov((\mathbf{X} - E(\mathbf{X}|\mathbf{X_p}))'(\mathbf{Y} - \widehat{\boldsymbol{\mu}_\mathbf{0}})|\mathbf{X_p},\ \mathbf{Y}) = \mathbf{L}\mathbf{A}\mathbf{L}'.$$

To estimate the variance-covariance matrix $Cov((\mathbf{X} - E(\mathbf{X}|\mathbf{X_p}))'(\mathbf{Y} - \widehat{\boldsymbol{\mu}_\mathbf{0}})|\mathbf{X_p},\ \mathbf{Y})$, we can use an empirical estimator (as in [27]). In general, Davies' method ([28]) can be used to approximate the distribution of a linear combination of independent χ^2_1.

However, in our case when variants are rare (e.g. MAF≤ 0.01) and sample sizes are small to modest, the normality assumption at each variant does not necessarily hold, and the above approximation can be very conservative. Therefore, to calculate the p value for Q we use a moment matching approach. More precisely, as in Lee et al. ([24]) the p value is calculated as $1-F((Q-\mu_Q)\sqrt{2df}/\sqrt{v_Q}+df)$, where F is the distribution function for χ^2_{df}. Here, $df = 12/\gamma$ where γ is the sample kurtosis. The mean, variance and kurtosis of Q can be estimated empirically by performing Monte Carlo simulations as follows. For each family i, under the null hypothesis of no association at any of the variants in a region, we replace $\mathbf{X_i} - E(\mathbf{X_i}|\mathbf{X_p})$ with $\{\mathbf{X_i} - E(\mathbf{X_i}|\mathbf{X_p})\}$ or $-\{\mathbf{X_i} - E(\mathbf{X_i}|\mathbf{X_p})\}$ with equal probability 1/2 (under the null hypothesis and assuming an additive model, the transmitted and untransmitted haplotypes are interchangeable). Although the p value calculation involves Monte Carlo simulations, we note that only a modest number of such simulations are needed (e.g. $10,000$) to estimate the three moments of Q, regardless of the magnitude of the p value.

More General Class of Family-based Association Tests

In the previous section we have assumed that all effects β_j's are independent, and we have derived the extension of the original SKAT method ([19]) to family-based designs. To allow for possible correlation among *effects* at different variants, we introduce the following family of kernels (as in [24]):

$$\tilde{\mathbf{K}}_\rho = [\mathbf{X} - E(\mathbf{X}|\mathbf{X_p})]\mathbf{W}\mathbf{R}_\rho\mathbf{W}[\mathbf{X} - E(\mathbf{X}|\mathbf{X_p})]',$$

where $\mathbf{R}_\rho = (1-\rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}'$ specifies an exchangeable correlation matrix. As before, the test statistic is:

$$Q_\rho \;\; = \;\; (\mathbf{Y} - \widehat{\boldsymbol{\mu}_\mathbf{0}})'\tilde{\mathbf{K}}_\rho(\mathbf{Y} - \widehat{\boldsymbol{\mu}_\mathbf{0}}). \tag{1}$$

When $\rho = 0$ we get the formulation in the previous section when all effects β_j are assumed independent. When $\rho = 1$, we get

$$Q_\rho = \left[\sum_{j=1}^{m} w_j \sum_{i=1}^{N} (Y_i - \widehat{\mu_{i,0}})(X_{ij} - E(X_{ij}|X_p))\right]^2,$$

which is equivalent to the test statistic in FBAT (De et al. unpublished; a burden test).

As before, for a fixed value of ρ, the null distribution of Q_ρ can be approximated by moment matching. When $\rho = 1$, Davies' analytical method also works well.

Connection to Population-based Tests

The class of sequence-based association tests above for family-based designs has a natural correspondence to recently proposed tests for population-based designs ([24]). The score test statistic for the population-based design takes a similar form as Q_ρ in equation (1) above (for more details, see [24]). Because of this direct connection, a comparison of family-based tests and population-based tests is very natural.

Results

Simulated Data

We simulated one genomic region of length 1 Mb under a coalescent model using the software package COSI ([29]). The model used in the simulation was the calibrated model for the European population. A total of $10,000$ haplotypes were generated in this region. We then randomly sampled subregions of the size of individual genes, representative of real exonic regions.

We simulate both trio and population-based data, with both dichotomous and continuous traits. We compare the two types of tests, Burden and SKAT, for both designs. Note that we are mainly interested in comparing the power of using *family-based controls* in a family-based design with the power of a population-based design, and for this purpose the trio design is a natural family design to compare against a population design. All variants (common and rare) are included in the analyses, and a weighting scheme that up-weights rare variants and down-weights common variants is used (see Methods section).

Type 1 Error

No Population Stratification To evaluate the type 1 error of the proposed tests, we have simulated datasets under the null hypothesis of no association between the offspring trait and the offspring genotypes. For dichotomous traits we simulate $n = 500$ trios, and $n = 500$ cases and an equal number of controls. For continuous traits we simulate $n = 500$ trios with a normally distributed $N(0, 1)$ offspring trait, and similarly for the population-based design we simulate $n = 500$ unrelated individuals. The results are shown in the quantile-quantile plots in Figure 1a and Supplementary Figure S1a. Both the family-based and the population-based tests result in correct type 1 error when there is no population substructure.

Population Stratification With population stratification, we assume that our sample contains individuals from two different populations. The ancestral population is simulated in COSI (as above). The two populations are simulated following the Balding-Nichols model ([30]) such that the distance between the two populations, F_{ST}, is 0.01, as would be encountered for closely related populations. More precisely, for each variant that has allele frequency p in the ancestral population, the allele frequencies in the two populations are drawn from a beta distribution with parameters $p(1 - F_{ST})/F_{ST}$ and $(1 - p)(1 - F_{ST})/F_{ST}$. For dichotomous traits, we assume the disease prevalence is 5% for population 1 and 1% in population 2. For continuous traits, $Y_1 \sim N(0, 1)$ and $Y_2 \sim N(\delta, 1)$, where $\delta = 0.5$. The results are shown in Figure 1b and Supplementary Figure S1b. While the family-based tests

maintain proper control of the type 1 error, the population-based tests show substantially inflated type 1 error rates in the presence of population substructure.

To adjust for population stratification in case-control and population-based designs, principal component analysis (PCA) has been proposed as an efficient approach in the context of common genetic variants in GWAS ([31]). We have applied such a PC analysis to our simulated data as well. PCs were calculated based on over $80,000$ variants (rare and common) that were generated across four independent chromosomes, each of size 1 Mb. The top 10 PCs were then used as covariates in our tests. We found the PCA adjustment to work well in our scenarios with a small number of discrete populations (Figure 1c and Supplementary Figure S1c), although such an adjustment may not be sufficient in more subtle scenarios, when the substructure is less discrete and the risk has a sharp spatial distribution ([32]).

Power Comparison of Family- and Population-based Designs

We compare the power of the two tests, Burden and SKAT, for family- and population-based designs on data simulated according to the following models. For a dichotomous trait, we assume the logistic model:

$$\text{logit}[P(Y_i = 1)] = \alpha_0 + \beta_1 X_{i1} + \cdots + \beta_m X_{im}.$$

For the trio design, we assume $n = 500$ trios, and $n = 500$ cases and an equal number of controls for the case-control design. The disease prevalence in the population is 0.05.

Similarly, for a continuous trait, we assume the linear model:

$$Y_i = \alpha_0 + \alpha_0 + \beta_1 X_{i1} + \cdots + \beta_m X_{im} + \epsilon_i,$$

where $\epsilon_i \sim N(0, 1)$. For the trio design, we assume $n = 500$ trios, and $n = 500$ unrelated individuals for the population-based design.

We assume that $10\% - 30\%$ of all variants are disease susceptibility variants. The β_j's are defined as

$$\beta_j = c|\log_{10} MAF_j|,$$

where $c = 0.4$ is chosen such that when $MAF = 0.0001$, $\beta = 1.6$ (i.e. $OR = 4.9$). We also simulate a scenario with only rare disease susceptibility variants and assume a constant OR of 4 for all disease susceptibility variants with MAF≤ 0.01.

Since SKAT is particularly advantageous in the presence of both risk and protective variants, we also simulate a scenario when 30% of the disease variants are protective (with $\beta_j = -c|\log_{10} MAF_j|$).

Only Risk Variants When all disease variants in a region are assumed to be risk variants, results for the two types of designs for both the Burden and SKAT tests are shown in Figure 2a. For dichotomous traits, the family-based design and the population-based design have similar power in the simulated scenarios, although at an increased sequencing cost for the family-based design. However, for continuous traits (with random ascertainment), the population-based design is much more powerful than the family-based design. For both types

of designs, the SKAT test is more powerful than the Burden test when a small proportion of the variants in a region are in fact disease susceptibility variants (e.g. 10%). The Burden test becomes slightly more powerful than the SKAT test when the percentage of causal variants in the region gets larger (e.g. 30% or larger). When only rare disease susceptibility variants are assumed with a common OR of 4, the results are qualitatively the same (Figure S2).

Mixture of Risk and Protective Variants With 30% of disease variants assumed protective, the SKAT test performs better than the Burden test for both the family- and population-based designs (Figure 2b). As before, for continuous traits the population-based design is more powerful than the family-based design. For dichotomous traits the family- and population-based designs have similar power when the Burden test is applied; however the family-based design is more powerful when the SKAT test is applied, suggesting that the family-based design with dichotomous traits has reduced sensitivity to the presence of protective variants compared to the population-based design (due to the reduced likelihood that parents of affected offspring carry protective variants).

Effect of PC Adjustment on Power We have evaluated the effect of adjusting for population stratification using PCA on the power of the population-based test. We simulated two populations as above, with an $F_{ST} = 0.01$ between the two populations, and different baseline risks as well. In particular, for dichotomous traits, the two disease prevalences are 0.05 and 0.01, while for continuous traits ϵ in the linear model above is $\sim N(0, 1)$ for population 1, and $\epsilon \sim N(0.5, 1)$ for population 2. The effect of the PC adjustment on power

was rather small in our simulations (Supplementary Figure S3).

Application to Exome-sequencing Study of 50 Trios

To illustrate these tests on real exome data, we have applied the two family-based tests to a small ongoing study of autism spectrum disorder (ASD). In total, 50 ASD children and their parents have been exome-sequenced (see Supplementary Material for more details on the data). Prior to analysis, we filtered out variants with Mendel error rate above 5%. A total of 18,303 genes were tested. Results are shown in Figure 3 for both tests, with no weighting scheme. Although the small number of trios precludes us from reporting experiment-wide significant results, it is reassuring that the observed distribution of gene P-values agrees well with the expectation.

Discussion

We have proposed a class of family-based association tests that includes as particular cases the burden test and the variance-component test (SKAT). Furthermore, these family-based tests correspond directly to existing population-based tests.

We show via simulations that the SKAT test is more powerful than the Burden test when the proportion of disease susceptibility variants in a region is small, and also when there is a mixture of risk and protective variants in the region being tested. The Burden test becomes more powerful than SKAT as the proportion of disease susceptibility variants in a region increases. We have also compared the power of using family-based controls in a family (trio) design vs. the power of a pure population-based design. Comparing family-

based and population-based designs for dichotomous traits we find they have similar power, while for continuous traits the population-based design can be more powerful. Although the number of individuals that need to be sequenced is higher for the family designs, the main advantage of the proposed family-based tests is robustness to population stratification. Family-based designs also allow the possibility to test for important biological hypotheses (such as the role of de-novo variation, and parent-of-origin effects). The population-based design is not robust to population stratification and popular methods for adjustment such as principal component analysis, although effective when there is a small number of discrete sub-populations, can fail to do a proper adjustment in more subtle scenarios. In a recent study, Mathieson and McVean ([32]) have shown that PCA can fail to correct for population stratification at rare variants when the underlying population substructure is continuous, and the risk has a sharp spatial distribution.

The proposed family-based association tests can be improved in numerous ways. As with the classical family-based association tests for common variants, these tests only use the within-family information. For common variants, it has been shown that great increases in power can be achieved for continuous traits by making use of the between-family information ([33]-[35]).

The possibility that rare variants have larger effect sizes than more common variants has recently generated a lot of interest in investigating the usefulness of families enriched in affected individuals to identify such high-risk rare genetic variants. This question has been studied elsewhere ([36]). We showed there that, under a genetic heterogeneity disease

model, for complex traits with small values for the sibling risk ratio (Risch's λ_S), as it is the case for most complex traits, affected individuals that have a close affected relative can be much more advantageous than affected individuals randomly selected from the population in detecting associations with high-risk, rare variants. For the purpose of this paper, we mainly focused on family-based designs that gain robustness to population stratification through the use of family-based controls.

Acknowledgments

The research was partially supported by National Science Foundation grant DMS-1100279 and National Institutes of Health grants R01MH095797 and 1R03HG005908 (to II-L), a Seaver Foundation grant and National Institutes of Health grant MH089025 (to JDB), and National Institutes of Health grants R37 CA076404 and P01CA134294 (to SL and XL).

Software

Software implementing the family-based tests discussed in this paper is available at our website (http://www.columbia.edu/ ii2135/).

Conflict of Interest Statement

The authors declare no conflict of interest.

References

[1] Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet* 24: 133–141.

[2] Metzker ML (2010) Sequencing technologies - the next generation. *Nat Rev Genet* 11: 31–46.

[3] Amos CI (2007) Successful design and conduct of genome-wide association studies. *Hum Mol Genet* 16: R220–R225.

[4] Ott J, Kamatani Y, Lathrop M (2011) Family-based designs for genome-wide association studies. *Nat Rev Genet* 12: 465–474.

[5] Sanders SJ, Murtha MT, Gupta AR et al. (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* In press.

[6] Neale BM, Kou Y, Liu L et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* In press.

[7] Rampersaud E, Mitchell BD, Naj AC, Pollin TI (2008) Investigating Parent of Origin Effects in Studies of Type 2 Diabetes and Obesity. *Curr Diabetes Rev* 4: 329–339.

[8] 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: 311–321.

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

[10] Price AL, Kryukov GV, de Bakker PI et al. (2010) Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 86: 832–838.

[11] Liu DJ, Leal SM (2010) A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet* 6: e1001156.

[12] Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 34: 188–193.

[13] King CR, Rathouz PJ, Nicolae DL (2010) An evolutionary framework for association testing in resequencing studies. *PLoS Genet* 6: e1001202.

[14] Bhatia G, Bansal V, Harismendy O et al. (2010) A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comput Biol* 6: e1000954.

[15] Basu S, Pan W (2010) Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol* 35: 606–619.

[16] Han F, Pan W (2010) A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* 70: 42–54.

[17] Ionita-Laza I, Buxbaum J, Laird NM, Lange C (2011) A new testing strategy to identify rare variants with either risk or protective effect on disease. *Plos Genet*, 7: e1001289.

[18] Neale BM, Rivas MA, Voight BF et al. (2011) Testing for an unusual distribution of rare variants. *PLoS Genet* 7: e1001322.

[19] Wu M, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011) Rare Variant Association Testing for Sequencing Data Using the Sequence Kernel Association Test (SKAT) *Am J Hum Genet*, in press.

[20] Sul JH, Han B, He D, Eskin E (2011) An optimal weighted aggregated association test for identification of rare variants involved in common diseases. *Genetics* 188: 181–188.

[21] Lin DY, Tang ZZ (2011) A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet*, 89: 354–367.

[22] Tzeng JY, Zhang D, Pongpanich M et al. (2011) Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *Am J Hum Genet* 89: 277–288.

[23] Laird NM, Horvath S and Xu X (2000) Implementing a unified approach to family based tests of association. *Genetic Epi* 19: S36–S42.

[24] Lee S, Wu M, Lin, X (2012) Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, in press.

[25] Zhang D, Lin X (2003) Hypothesis testing in semiparametric additive mixed models. *Biostatistics* 4: 57–74.

[26] Rabinowitz D, Laird N (2000) A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered* 50: 211–223.

[27] Rakovski CS, Xu X, Lazarus R, Blacker D, Laird NM (2007) A new multimarker test for family-based association studies. *Genet Epidemiol* 31: 9–17.

[28] Davies RB (1980) Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *Applied Statistics* 29: 323–333.

[29] Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 15: 1576–1583.

[30] Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* 96: 3–12.

[31] Price AL, Patterson NJ, Plenge RM et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38: 904–909.

[32] Mathieson I, McVean G (2012) Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 44: 243–246.

[33] Van Steen K, McQueen MB, Herbert A et al. (2005) Genomic screening and replication using the same data set in family-based association testing. *Nat Genet* 37: 683–691.

[34] Ionita-Laza I, McQueen MB, Laird NM, Lange C (2007) Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. *Am J Hum Genet* 81: 607–614.

[35] Won S, Wilk JB, Mathias RA et al. (2009) On the analysis of genome-wide association studies in family-based designs: a universal, robust analysis approach and an application to four genome-wide association studies. *PLoS Genet* 5: e1000741.

[36] Ionita-Laza I, Ottman R (2011) Study designs for identification of rare disease variants in complex diseases: the utility of family-based designs. *Genetics* 189: 1061–1068.

[37] DePristo MA, Banks E, Poplin R et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43: 491–498.

[38] Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760.

[39] McKenna A, Hanna M, Banks E et al. (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20: 1297–1303.

Supplementary Material

Exome-sequencing of 50 Trios

Whole exome sequencing analyses were performed in 50 autism trios. The SeqCap EZ Human Exome Library v2.0 kit was used for library enrichment (http://www.nimblegen.com/products/

seqcap/ez/v2/index.html). The captured exome libraries were then sequenced on a HiSeq2000 according to the manufacturer's instructions for paired-end 100-bp reads (www.illumina.com). The sequencing data were put through a computational pipeline for WES data processing and analysis followed the general workflow adopted by the 1000 genomes project ([37]). First, the alignment of raw sequence reads to the human reference genome sequence (NCBI GRCh37) was performed using a fast lightweight Burrows-Wheeler Alignment Tool (BWA v.0.6.1, [38]). Genome Analysis Toolkit (GATK v1.5-16-g58245bf) was then used for base-quality recalibration and local realignment to minimize base calling error and mapping error, respectively. Lastly, GATK ([39]) Unified Genotyper tool was employed to call single-nucleotide substitutions and short insertions/deletions. Only passing variants were included in the final variant set.

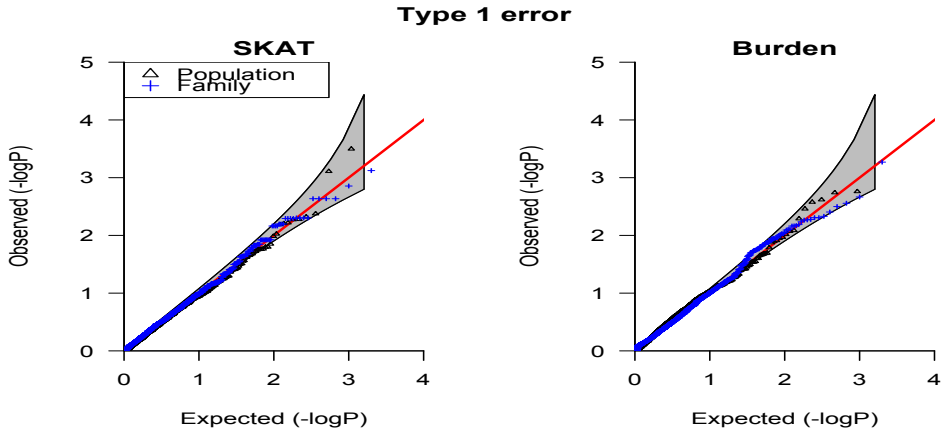
Figure Legends

Figure 1: Type 1 error, Dichotomous Trait. Results for the SKAT test and for the Burden test are shown, for both the trio design with $n = 500$ trios and the case-control design with $n = 500$ cases and $n = 500$ controls. 95% CI is also shown. a) no population stratification, b) in the presence of population stratification, c) with Eigenstrat correction for population stratification.

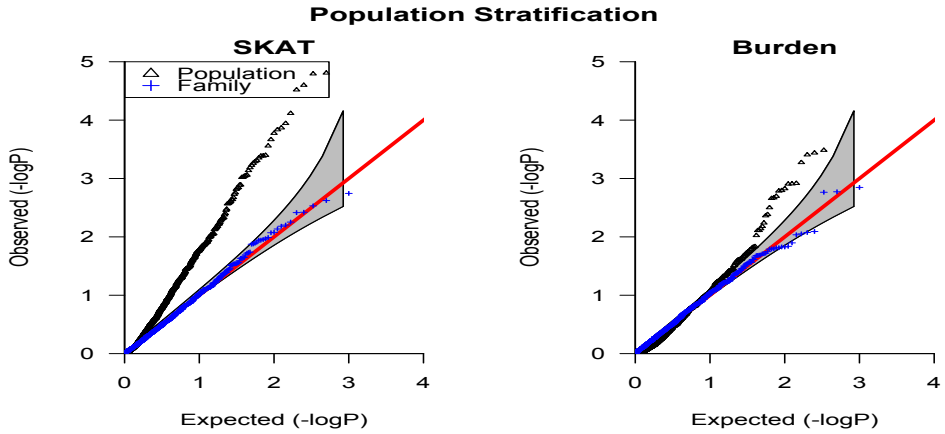
Figure 2: Power at $\alpha = 0.05$. T is the trio design ($n = 500$ trios) and P is the population-based design ($n = 500$ cases and $n = 500$ controls for the dichotomous trait, and $n = 500$ unrelated individuals for the continuous trait). T_S is the SKAT test, and T_B is the Burden test for the trio design. Similar notations for the population-based design. a) all disease susceptibility variants are risk variants, and b) 30% of the disease susceptibility variants are protective.

Figure 3: QQ plots, $n = 50$ exome-sequenced trios. Results are shown for the SKAT and Burden tests, with MAF threshold (0.05) and no MAF threshold. 95% CI is also shown.

a)


Type 1 error
SKAT
Population
Family
Burden
Observed (-logP)
Expected (-logP)

b)


Population Stratification
SKAT
Population
Family
Burden
Observed (-logP)
Expected (-logP)

c)

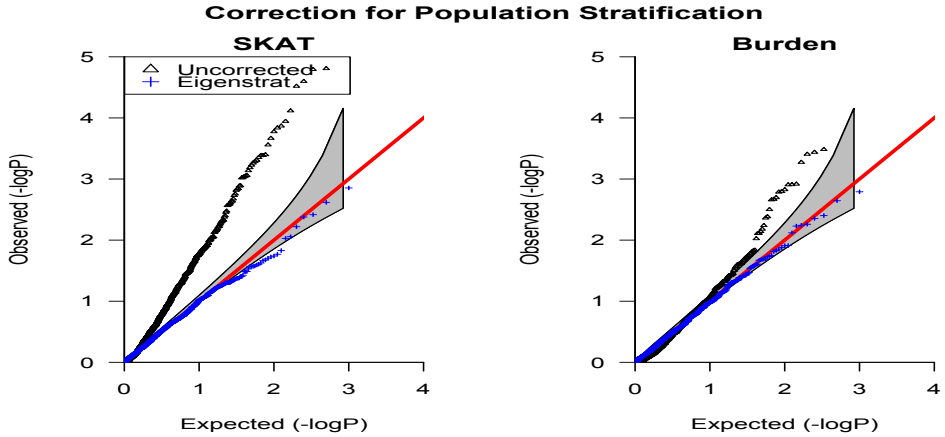
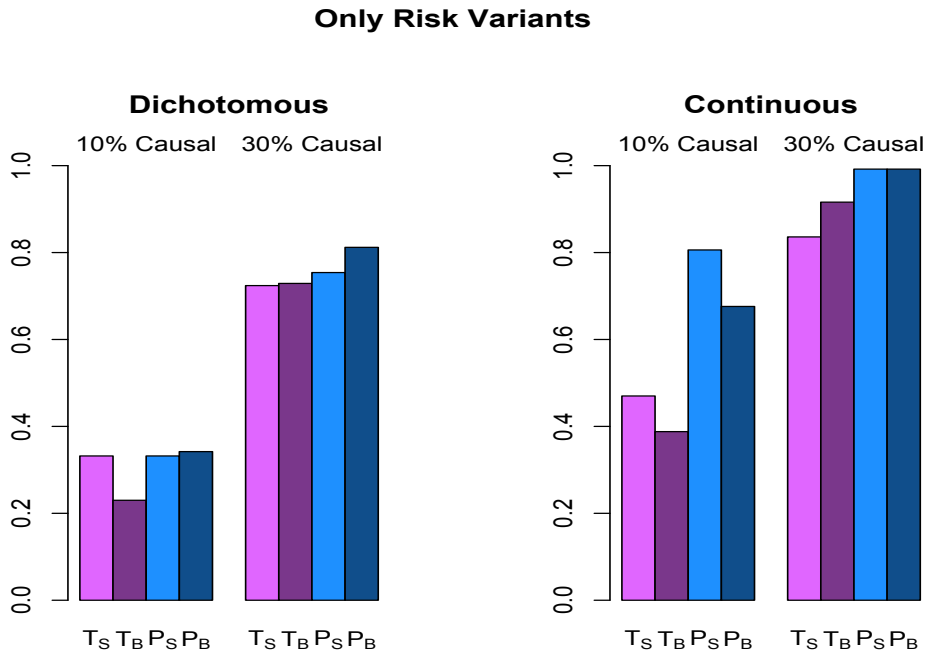

Correction for Population Stratification
SKAT
Uncorrected
Eigenstrat
Burden
Observed (-logP)
Expected (-logP)

Figure 1:

a)


Only Risk Variants
Dichotomous
10% Causal
30% Causal
Continuous
10% Causal
30% Causal
0.0
0.2
0.4
0.6
0.8
1.0
TS TB PS PB
TS TB PS PB
TS TB PS PB
TS TB PS PB

b)

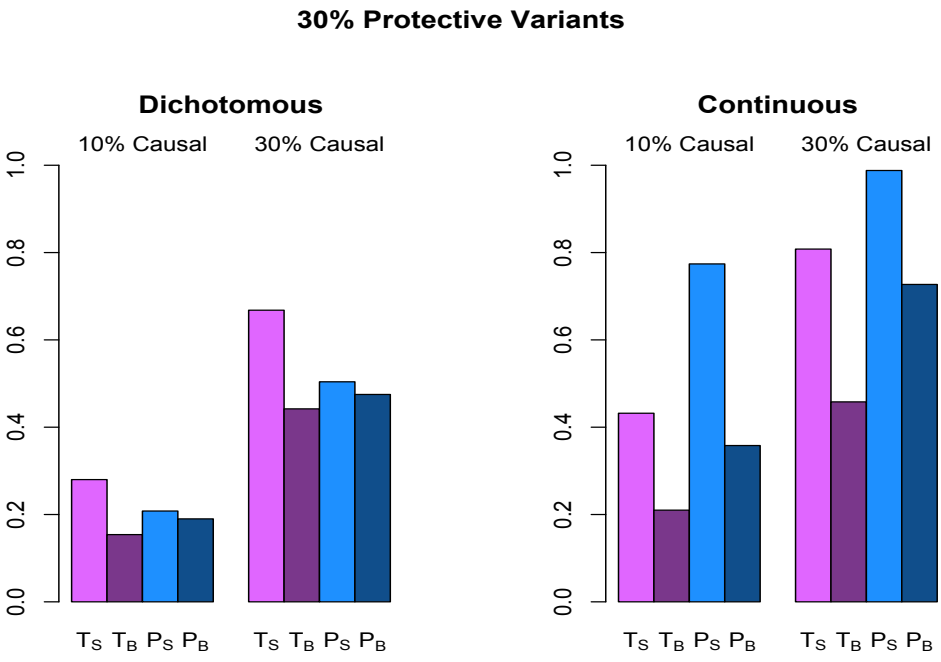
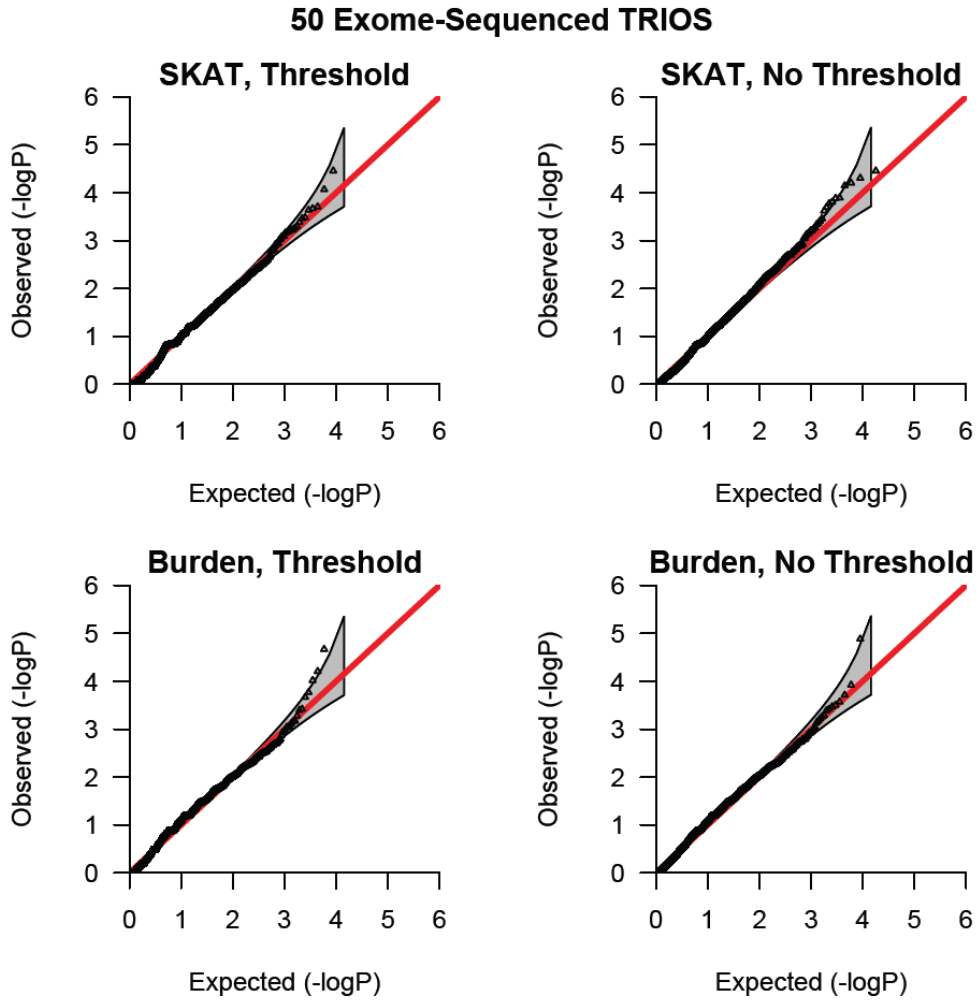
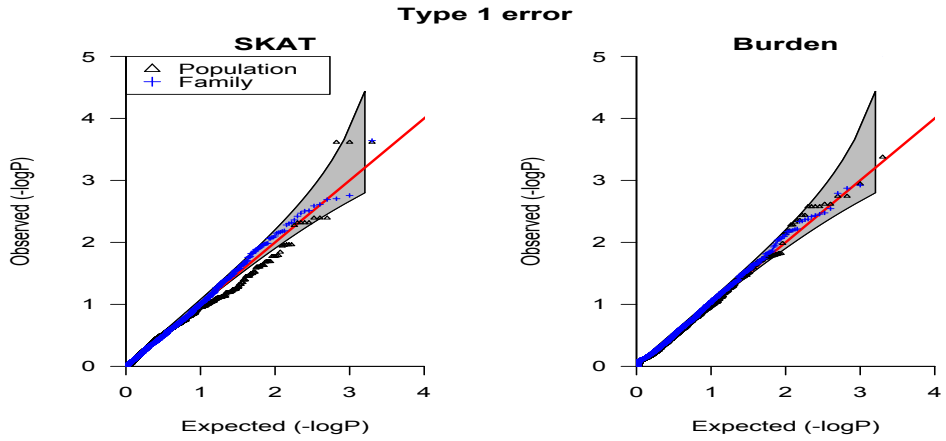

30% Protective Variants
Dichotomous
10% Causal
30% Causal
Continuous
10% Causal
30% Causal
0.0
0.2
0.4
0.6
0.8
1.0
TS TB PS PB
TS TB PS PB
TS TB PS PB
TS TB PS PB

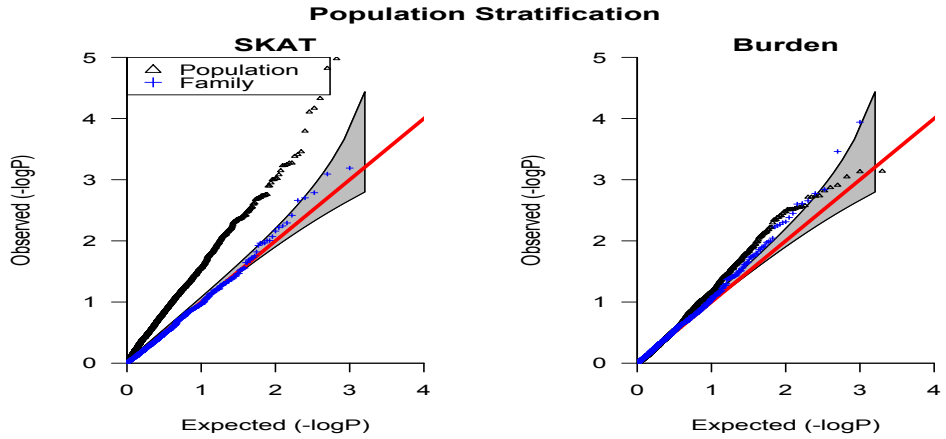
Figure 2:


50 Exome-Sequenced TRIOS
SKAT, Threshold
SKAT, No Threshold
Burden, Threshold
Burden, No Threshold
Observed (-logP)
Expected (-logP)
0
1
2
3
4
5
6

a)



b)



c)

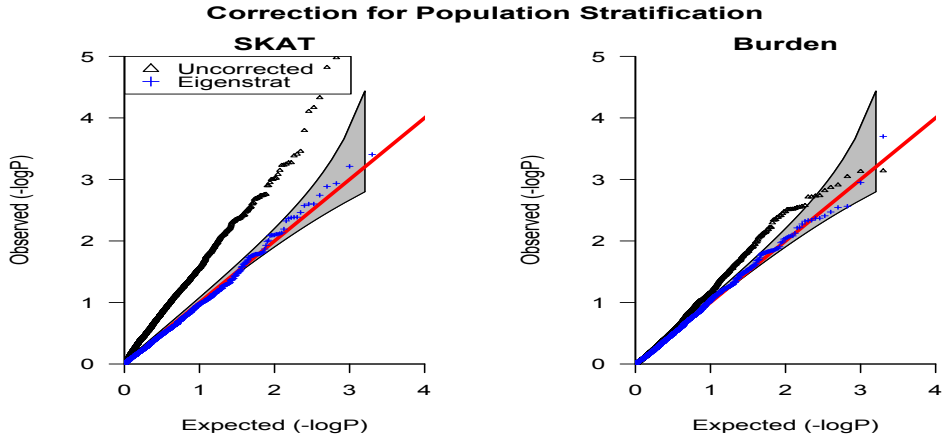
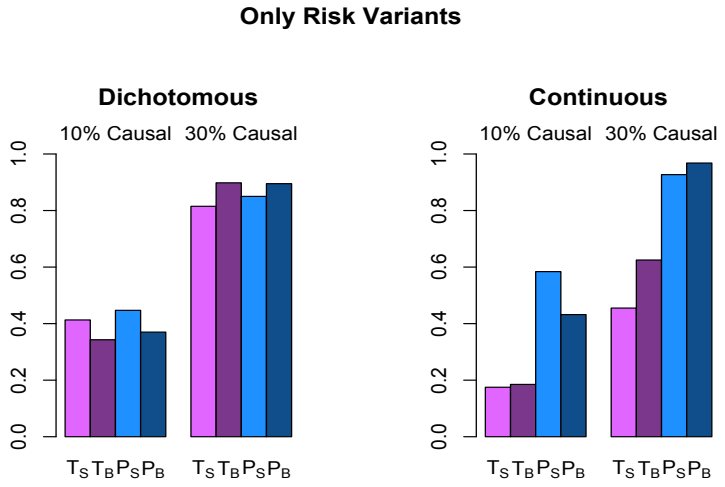


Figure S1: Type 1 error, Continuous Trait. Results for the SKAT test and for the Burden test are shown, for both the trio design with $n = 500$ trios and the population-based design with $n = 500$ unrelated individuals. 95% CI is also shown. a) no population stratification, b) in the presence of population stratification, c) with Eigenstrat correction for population stratification.

a)



b)

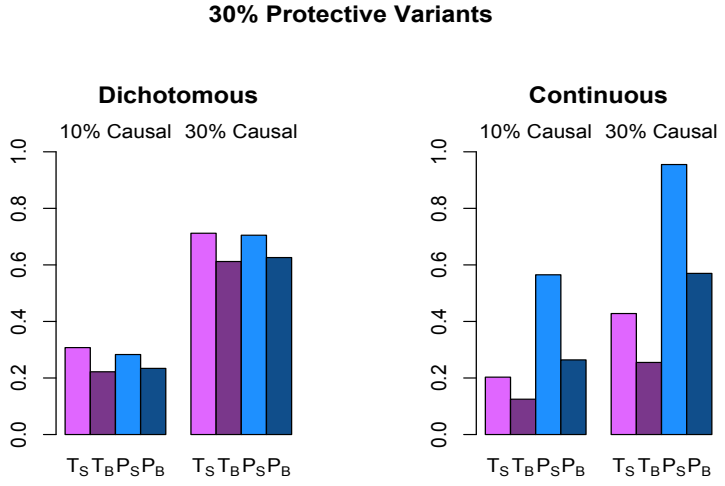


Figure S2: Constant OR= 4 for all disease risk variants with MAF$\leq$ 0.01. Power at $\alpha = 0.05$. T_S is the SKAT test, and T_B is the Burden test for the trio design. Similar notations for the population-based design. a) all disease susceptibility variants are risk variants, and b) 30% of the disease susceptibility variants are protective.

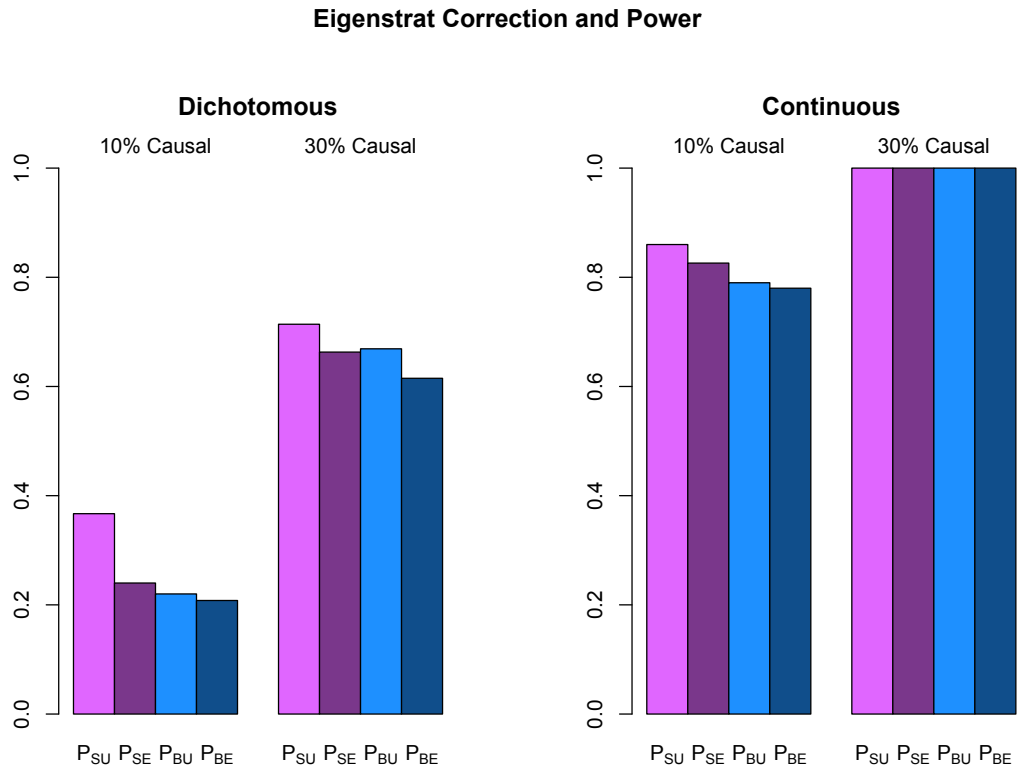


Figure S3: Power (at $\alpha = 0.05$) for the Population-based Design after PC adjustment. P_{SU} is the SKAT approach, uncorrected, and P_{SE} is the SKAT approach, with Eigenstrat correction. Similar notations hold for the Burden test.