Title: gsSKAT: Rapid Gene-Set Analysis and Multiple Testing Correction for Rare-variant Association Studies using Weighted Linear Kernels

Authors: Nicholas B. Larson[1§], Shannon McDonnell[1], Lisa Cannon Albright[2], Craig Teerlink[2], Janet Stanford[3], Elaine A. Ostrander[4], William B. Isaacs[5], Jianfeng Xu[6], Kathleen A. Cooney[2,7], Ethan Lange[8], Johanna Schleutker[9], John D. Carpten[10], Isaac Powell[11], Joan E. Bailey-Wilson[12], Olivier Cussenot[13], Geraldine Cancel-Tassin[13], Graham G. Giles[14], Robert J. MacInnis[14], Christiane Maier[15], Alice S. Whittemore[16], Chih-Lin Hsieh[17], Fredrik Wiklund[18], William J. Catolona[19], William Foulkes[20], Diptasri Mandal[21], Rosalind Eeles[22], Zsofia Kote-Jarai[22], Michael J. Ackerman[23], Timothy M. Olson[23], Christopher J. Klein[24], Stephen N. Thibodeau[25], Daniel J. Schaid[1]

1) Division of Biomedical Statistics and Informatics, Dept. of Health Sciences Research, Mayo Clinic, Rochester, MN

2) Dept. Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT

3) Fred Hutchinson Cancer Research Center, Seattle, WA

4) National Human Genome Research Institute, Bethesda, MD

5) Brady Urological Institute, Johns Hopkins University, School of Medicine, Baltimore, MD

6) NorthShore University HealthSystem Research Institute, Chicago, IL

7) Depts. of Internal Medicine and Urology, University of Michigan Medical School, Ann Arbor, MI

8) Dept. of Genetics, University of North Carolina, Chapel Hill, NC

9) Dept. of Medical Biochemistry and Genetics, Institute of Biomedicine, University of Turku, Finland

10) Dept. of Translational Genomics, University of Southern California, Los Angeles, CA

11) Wayne State University, Detroit, MI

12) Statistical Genetics Section, National Human Genome Research Institute, Bethesda, MD

13) CeRePP, Hopital Tenon, Paris, France

14) Cancer Epidemiology Centre, Cancer Council Victoria, and Centre for Epidemiology and Biostatistics, School of Population and Global Health, University of Melbourne, Melbourne, Australia

15) Dept. of Urology, University of Ulm, Ulm, Germany

16) Dept. Health Research and Policy, Stanford University, Stanford, CA

17) Dept. of Urology , University of Southern California, Los Angeles, CA

18) Dept. of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, SwedenNeuroLa

19) Dept. of Urology, Northwestern University Feinberg School of Medicine, Chicago, IL

20) Depts. of Oncology and Human Genetics, Montreal General Hospital, Montreal QC, Canada

Author Manuscript

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/gepi.22036.

This article is protected by copyright. All rights reserved.

21) Dept. of Genetics, LSU Health Sciences Center, New Orleans, LA

22) The Institute of Cancer Research and Royal Marsden NHS Foundation Trust, London, UK

23) Dept. of Pediatric and Adolescent Medicine, Mayo Clinic, Rochester, MN

24) Dept. of Neurology, Mayo Clinic, Rochester, MN

25) Dept. of Laboratory Medicine/Pathology, Mayo Clinic, Rochester, MN

[§]Corresponding author

Nicholas B. Larson, PhD

Department of Health Sciences Research

Mayo Clinic

200 First Street SW

Rochester, MN 55905

Email: Larson.nicholas@mayo.edu

Phone: (507) – 293 – 1700

Fax: (507) – 284 – 1516

Author Manuscript

This article is protected by copyright. All rights reserved.

Abstract: Next-generation sequencing technologies have afforded unprecedented characterization of low-frequency and rare genetic variation. Due to low power for single-variant testing, aggregative methods are commonly used to combine observed rare variation within a single gene. Causal variation may also aggregate across multiple genes within relevant biomolecular pathways. Kernel-machine regression and adaptive testing methods for aggregative rare-variant association testing have been demonstrated to be powerful approaches for pathway-level analysis, although these methods tend to be computationally intensive at high-variant dimensionality and require access to complete data. An additional analytical issue in scans of large pathway definition sets is multiple testing correction. Gene-set definitions may exhibit substantial genic overlap, and the impact of the resultant correlation in test statistics on Type I error rate control for large agnostic gene-set scans has not been fully explored. Herein, we first outline a statistical strategy for aggregative rare-variant analysis using component gene-level linear kernel score test summary statistics as well as derive simple estimators of the effective number of tests for family-wise error rate control. We then conduct extensive simulation studies to characterize the behavior of our approach relative to direct application of kernel and adaptive methods under a variety of conditions. We also apply our method to two case-control studies respectively evaluating rare variation in hereditary prostate cancer and schizophrenia. Finally, we provide open-source R code for public use to facilitate easy application of our methods to existing rare-variant analysis results.

Keywords: gene-set, next-generation sequencing, pathway, rare variation

Author Manuscript

This article is protected by copyright. All rights reserved.

INTRODUCTION

Next-generation sequencing (NGS) is rapidly emerging as a primary technology for evaluating the role of genetic variation in human phenotypes. Unlike the standardized genotyping arrays used in genome-wide association studies, which interrogate tagging single nucleotide polymorphisms (SNPs) to indirectly capture the majority of common variation (i.e., minor allele frequency (MAF) > 5%) via linkage disequilibrium, NGS studies provide nucleotide-resolution information of the targeted DNA through massively parallel short-read sequencing, capable of directly identifying causal variation. This is a critical advantage with the recent interest in the role of rare genetic variation in complex traits. Many arguments exist that not only support this model of genetic association from a population genetics perspective[1], due to a rapidly increasing population size and weak purifying selection, but suggest that rare variation may constitute a large portion of the so-called missing heritability observed in many complex traits[2, 3]. There is now a growing list of examples where high-impact rare variants have been identified in complex diseases[4, 5].

Although sequencing costs continue to decline as technology improves, NGS studies still remain quite expensive and sample sizes are often too small to be sufficiently powered for gene-based aggregative rare-variant testing[6]. However, aggregative gene-set analyses informed by molecular pathways and/or perturbation studies may be sufficiently powered by combining genetic variation across functionally related genes. Moreover, they may capture genetic heterogeneity of etiological mechanisms at a larger systematic level, as these complex traits are likely multi-factorial and could be driven by the disruption of one of any number of key genes in a particular molecular pathway. Analyses involving aggregation of rare variants across large gene-sets have recently been applied in studies of non-obstructive azoospermia[7], adolescent idiopathic scoliosis[8], and schizophrenia[9].

Various extensions of rare-variant testing to the pathway level using self-contained testing have previously been explored[6]. Yan et al. recently proposed adaptive rank-truncation product (ARTP) methods in conjunction with kernel-machine regression testing[10], which demonstrated statistical power improvements over direct application of the popular sequence kernel association test (SKAT)[11]. However, these approaches are permutation-based and consequently computationally demanding, with moderately sized (80-100 genes) gene-sets requiring thousands of minutes of compute time. Pan et al.[12] extended their adaptive score power test (aSPU) to genetic pathways (aSPUpath), which is also based upon score tests and applicable to common and rare variation, with permutation or parametric bootstrap applied to obtain gene-set p-values. While these adaptive methods have demonstrated advantages in statistical power over competing approaches, their implementation requires access to the complete datasets to conduct gene-set testing, limiting their application.

Evaluation of Type I error rate control for gene-set analysis is also understudied. In contrast to gene-based rare-variant tests, which can justifiably be assumed to be independent due to anticipated low linkage disequilibrium amongst rare variation, multiple testing correction for gene-set aggregative testing is complicated by the potential for substantial genic overlap. Examination of the latest version (v5.1) of the C2 curated gene-set collection present in the Molecular Signatures Database[13, 14] (MSigDB, http://software.broadinstitute.org/gsea/msigdb) reveals 4726 total gene-sets, in addition to the 1454 Gene Ontology[15] (GO) gene-sets. A total of 13,311 gene-sets encompass the entire available list. An agnostic scan of all available gene-sets could result in thousands of correlated test statistics, for which traditional multiple testing correction methods (e.g., Bonferroni) may be overly conservative without acknowledging this correlation structure.

This article is protected by copyright. All rights reserved.

Author Manuscript

Wu et al.'s examination of rare-variant pathway analysis strategies identified the direct application of SKAT to the union of rare-variant genotypes across the gene-set (i.e., "super gene" approach) to perform well across a variety of conditions[6]. An advantage of this approach is that linear kernels can easily be combined to form composite kernels, supporting the hierarchical framework for gene-set kernel construction. Herein, we first outline a computationally efficient strategy for conducting gene-set based rare-variant association analyses and evaluate estimators of the effective number of gene-set tests based on SKAT and its burden/kernel optimized variation, SKAT-O[16]. We then conduct simulation studies under null conditions to characterize the properties of our gene-level p-value aggregation approach relative to direct application of kernel-based tests as well as demonstrate proper control of the family-wise error rate (FWER) for simultaneous analysis of large numbers of overlapping gene-sets. We also investigate comparative power relative to approaches using complete data as well as any practical impacts on study design when proper FWER control is implemented relative to conservative Bonferroni adjustment. Finally, we apply our approach to a case-control sequencing study of hereditary prostate cancer and previously published results from a case-control study of schizophrenia using exome chip array data.

METHODS

Kernel-Machine Association Testing for Rare Variants

Consider a rare-variant association study of sample size N with a given $N \times 1$ phenotype vector $\boldsymbol{Y}$, such that genotype dosages are available on N subjects for L genes. Let $\widehat{\boldsymbol{Y}}$ correspond to a vector of fitted phenotype values (if $\boldsymbol{Y}$ is continuous) or trait probabilities (if $\boldsymbol{Y}$ is dichotomous) from a relevant regression model taking into account adjusting covariates (e.g., age and sex). Hypothesis testing for gene-level rare-variant association analysis may be conducted using an aggregative variance component testing approach, like SKAT, such that a kernel test score statistic for a given gene $l \in 1, \dots, L$ is defined as $Q_l = (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})' \boldsymbol{K_l} (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})$, where $\boldsymbol{K_l}$ is a kernel representation of M_l variants within the tested region(s) of gene l. For a weighted linear kernel function, we explicitly define $\boldsymbol{K_l}$ to be represented as $\boldsymbol{K_l} = \boldsymbol{G_l W_l G_l}'$, where $\boldsymbol{G}$ is an $N \times M_l$ matrix of variant genotypes and $\boldsymbol{W_l}$ is an $M_l \times M_l$ diagonal weight matrix. The test statistic Q_l then follows a mixture chi-square distribution under the null hypothesis.

Extension to Gene-Set Analyses

Define the set $\mathcal{S} \subset \{1, \dots, L\}$ to represent a given gene-set or pathway, such that it is defined by the subset of $U \leq L$ total genes. A gene-set test statistic may be defined similarly to that conducted at the gene-level, whereby $\boldsymbol{G_S}$ is an $N \times M_S$ composite matrix formed by appending the genotype matrices of the U component genes that comprise $\mathcal{S}$, given as $(\boldsymbol{G_{S_1}} \quad \cdots \quad \boldsymbol{G_{S_U}})$. Then, we define the gene-set test statistic $Q_S = (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})' \boldsymbol{K_S} (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})$, where

$$\boldsymbol{K_S} = \boldsymbol{G_S W_S G_S'} = (\boldsymbol{G_{S_1}} \quad \cdots \quad \boldsymbol{G_{S_U}}) diag(\boldsymbol{W_{S_1}}, \dots, \boldsymbol{W_{S_U}}) \begin{pmatrix} \boldsymbol{G_{S_1}'} \\ \vdots \\ \boldsymbol{G_{S_U}'} \end{pmatrix} = \sum_{l \in \mathcal{S}} \boldsymbol{K_l}$$

Thus, $\boldsymbol{K_S}$ is a linear composite kernel matrix of the component gene-level kernel matrices, and $Q_S = (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})' (\sum_{l \in \mathcal{S}} \boldsymbol{K_l}) (\boldsymbol{Y} - \widehat{\boldsymbol{Y}}) = \sum_{l \in \mathcal{S}} Q_l$. For rare variation, it is reasonable to assume that the gene-level score statistics $\{Q_l : l \in \mathcal{S}\}$ are independent, so long as the variant sets are non-overlapping. Under these conditions, a simple solution to deriving gene-set level test p-values is to combine pre-

This article is protected by copyright. All rights reserved.

Author Manuscript

computed gene-level p-values, p_l, using p-value aggregation approach, such as Fisher's method[17] or Stouffer's combined Z-score[18]. We elect the latter, as it affords working with multivariate normality and gene-based weights, if warranted, are easily supported. Moreover, Stouffer's method tends to be more powerful than Fisher's when a large proportion of testing results are moderately significant[19], which may more likely reflect the underlying biology in pathway-directed analysis. Define $Z_l = \Phi^{-1}(1-p_l)$ where $\Phi^{-1}(\cdot)$ is the inverse cumulative density function (CDF) of the standard Gaussian distribution, such that under the null condition $Z_l \sim N(0,1)$, For a given gene-set $\mathcal{S}$ with k genes, $Z_{\mathcal{S}} = \frac{\sum_{l\in\mathcal{S}} Z_l}{\sqrt{k}}$ and $Z_{\mathcal{S}}$ also follows a standard Gaussian distribution under the gene-set null condition, from which a gene-set p-value can be easily computed. Note that in addition to SKAT, these results also generalize to SKAT-O[16], as $\boldsymbol{W_l}$ can simply be substituted by $\boldsymbol{W_l^{0.5}}\left((1-\rho_l)\boldsymbol{I} + \rho_l \boldsymbol{1}\boldsymbol{1}'\right)\boldsymbol{W_j^{0.5}}$ with the caveat that ρ_l is pre-selected gene-wise rather than optimized over the entire set of variants $\boldsymbol{G_S}$. We refer to this approach generally as gsSKAT.

Multiple Testing Correction

When testing a large number of gene-sets, R, such as via an agnostic scan of multiple curated gene-sets (e.g., GO), significance threshold adjustment such as a Bonferroni correction is very likely to be overly conservative due to substantial overlap of genes across gene-sets. Consider a second gene-set $\mathcal{S}^*$ with k^* genes such that $\mathcal{S}^* \cap \mathcal{S} \neq \emptyset$. The covariance $Cov(Z_{\mathcal{S}}, Z_{\mathcal{S}^*})$ may be derived using the bilinearity of covariance, such that

$$Cov(Z_{\mathcal{S}}, Z_{\mathcal{S}^*}) = Cov\left(\frac{1}{\sqrt{k}}\sum_{u\in\mathcal{S}} Z_u, \frac{1}{\sqrt{k^*}}\sum_{v\in\mathcal{S}^*} Z_v\right) = \sum_{u\in\mathcal{S}}\sum_{v\in\mathcal{S}^*} Cov\left(\frac{1}{\sqrt{k}}Z_u, \frac{1}{\sqrt{k^*}}Z_v\right)$$

Since we make the underlying assumption that Q_u and Q_v are independent for all u, v where $u \neq v$, $Cov\left(\frac{1}{\sqrt{k}}Z_u, \frac{1}{\sqrt{k^*}}Z_v\right) = 0$ for $u \neq v$. If we denote $\mathcal{S}^{INT} = \mathcal{S} \cap \mathcal{S}^*$, then

$$Cov(Z_{\mathcal{S}}, Z_{\mathcal{S}^*}) = \sum_{u\in\mathcal{S}^{INT}} Cov\left(\frac{1}{\sqrt{k}}Z_u, \frac{1}{\sqrt{k^*}}Z_u\right) = \frac{|\mathcal{S}^{INT}|}{\sqrt{k}\sqrt{k^*}}$$

where $|\mathcal{S}^{INT}|$ corresponds to the cardinality of the intersection of S and S^*, equivalent to the number of overlapping genes. More generally, we derive the correlation matrix of the pathway-level Z-scores by defining a $R \times L$ pathway design matrix, $\boldsymbol{D}$, for R unique pathways, such that matrix element $D_{ji} = \frac{1}{\sqrt{k_j}}$ if gene i is in S_j, where k_j is the total number of genes in pathway j, and 0 otherwise. Then, $\boldsymbol{Z} \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{D}'$. We consider estimates of the effective number of tests via the Galwey method[20] and the modification of the Gao approach described by Hendricks et al.[21] Let $\boldsymbol{\Lambda}$ be the vector of eigenvalues corresponding to the eigendecomposition of $\boldsymbol{\Sigma}$. A Galwey estimate for the effective number of gene-set level tests, R^{eff}_{Galwey}, is given by $R^{eff}_{Galwey} = \frac{\left(\sum\sqrt{\lambda_r}\right)^2}{\sum\lambda_r}$ for $\lambda_r > 0$, where λ_r indicates the r^{th} largest eigenvalue. For the Gao estimator, R^{eff}_{Gao} is equivalent to the smallest R^* that satisfies $\frac{\sum_{r=1}^{R^*}|\lambda_r|}{\sum_{r=1}^{R}|\lambda_r|} < C$, where C is set to the recommended value of 0.995. Estimates of R^{eff} may in turn be applied as a Bonferroni correction factor for adjusting the statistical significance threshold to maintain proper FWER control. Despite the potentially large dimensionality of $\boldsymbol{D}\boldsymbol{D}'$, efficient computation of $\boldsymbol{\Lambda}$ can easily be conducted by exploiting sparse matrix algebra.

This article is protected by copyright. All rights reserved.

Author Manuscript

Simulation Study

To simulate genotypes under realistic rare-variant distributions and frequencies, empirical MAF data were downloaded from the NHLBI-ESP Exome Variant Server (release ESP6500). We isolated the European American MAFs for all observed polymorphic sites, and all variants with MAF >0.05 were excluded as common SNPs. To simulate genotype data for genes that would likely pass traditional gene-level rare-variant testing criteria (e.g., sufficient number of minor alleles and variant positions), we first identified the subset of genes with a cumulative MAF (cMAF) > 0.01 and total number of variant positions between 10 and 100. We then randomly sampled ten gene-sets corresponding to 181 unique genes from the Reactome[22] pathway database ranging in gene-set size from 6 to 47 as well as varying degrees of overlap (Table I).

To characterize the p-value aggregative approach under a variety of study conditions, we simulated rare-variant genotypes for the 181 genes for an underlying population of 20,000 samples. Genotypes were simulated independently across samples and variant sites under assumptions of Hardy-Weinberg equilibrium. We employed a similar simulation strategy for phenotypes as defined in Wu et al.[11], such that we defined continuous outcome y_i independently as

$$y_i = X_{i1} \times 0.5 + X_{i2} \times 0.5 + \boldsymbol{G_i'\beta} + \epsilon_i$$

where $X_{i1} \sim N(0,1)$, $X_{i2} \sim Bernoulli(0.5)$, and $\epsilon_i \sim N(0,1)$ were all independently sampled for each simulated subject. For power simulations, we defined causal genetic effects $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)'$ for M total gene-set variants such that $|\beta_m| = 0.4 \times |\log_{10}(MAF_m)|$ if the variant m was designated to be causal and 0 otherwise, where MAF_m is the population MAF of the m^{th} variant. For binary traits with trait probability p_i, we modeled $logit(p_i)$ the same as continuous outcome y_i, with the exception that we additionally included an intercept β_0 such that disease prevalence was approximately 10%.

Type I Error and FWER

To characterize the behavior of gsSKAT under null conditions, we considered total sequencing study sample sizes of N = 1000 and 2000, which were randomly drawn from the larger population of 20,000 samples with replacement for each simulation iteration. SKAT and SKAT-O kernel testing was conducted for all individual genes as well as the 10 gene-sets under all simulation conditions, while Stouffer's combined p-values were also computed for the gene-sets based on the component gene-level p-values. Finally, SKAT-O has a tendency to produce p-values of 1.00, which would yield negative infinity values from $\Phi^{-1}(\cdot)$. In the event of $p = 1.00$, we set $p = 0.95$ to avoid this issue. For these simulations, 5000 iterations were conducted per analysis.

To evaluate proper control of the FWER rates for a given α using the adjusted multiple testing correction factor R^{eff}, we additionally simulated uniformly distributed p-values for 19,399 genes present in the ESP6500 dataset as the basis for gene-set analysis. We considered three separate magnitudes of analyses defined by the total number of gene-sets under consideration, denoted A_1, A_2, A_3, that were respectively comprised of all gene-sets with $5 < |\mathcal{S}| < 1000$ in Reactome (673 gene-sets), Reactome and GO (2109 total gene-sets), and MSigDB C2 and GO (6124 total gene-sets). We presumed that these analyses would be typical of most agnostic gene-set scans conducted by an investigator. For each analysis, we enumerated the number of instances a Type I error occurred at an adjusted α-level of 0.05 using each of the two R^{eff} estimates as a Bonferroni correction factor. A total of 10,000 replications were conducted for each analysis.

This article is protected by copyright. All rights reserved.

Author Manuscript

Statistical Power

There are myriad configurations of genetic architectures (e.g., causal variant proportion, direction of effects, relationship between effect magnitude and MAF) under which statistical power may be calculated for aggregate rare-variant testing methods. To investigate relative statistical power of our p-value aggregation method compared to direct applications of competing approaches using available genetic, covariate, and phenotypic data, we conducted a variety of simulations based on continuous ($N = 1000$) and dichotomous ($N = 2000$) outcomes for a limited set of realistic conditions using data for gene-sets ($S_2, S_4, S_6, S_8, S_{10}$). We defined the genetic architecture of a causal genetic pathway hierarchically, such that a gene-set is comprised of causal and null genes, and causal genes are comprised of causal and null variants. For our purposes, we randomly selected approximately 5% of the variants to be causal for a given causal gene for continuous outcomes and 10% for binary in order to achieve adequate power, and considered causal gene proportions across the gene-set to be 100%, ~50%, and ~25%. Finally, we considered effects to be either all positive or a randomly selected mixture of positive and negative. Genetic profiles were randomly selected from the population of 20,000 previously generated genotype vectors, while a larger population of 50,000 subjects was generated for binary outcomes by sampling with replacement. In addition to SKAT-O and gsSKAT using gene-level SKAT-O p-values, we also applied the aSPUpath adaptive test using the *aSPU* R package. The aSPUpath test was applied using 10,000 permutations, with all other settings set to defaults. A total of 1000 simulation iterations were performed per analysis, and all testing results were considered significant at an α level of 0.05.

Quantifying potential improvements in power via correct selection of an adjusted α level is another statistical power consideration of direct relevance to this work. To evaluate the practical impact of proper FWER control using R^{eff} relative to a conservative Bonferroni adjustment on power, we estimated statistical power for an example gene-set, S_4, as part of a hypothetical agnostic scan of a gene-set associations for each of the three gene-set analyses (A_1, A_2, A_3) described above. Specifically, power was estimated at Bonferroni-corrected α levels of 0.05 where the correction factor was set as either R or R^{eff}. We applied the built-in power calculation functions in the *SKAT* R package to calculate power estimates under specific α levels under a causal variant percentage of 0.10, causal variant MAF threshold of 0.05, and even mixture of positive and negative effects, with all other settings set to default parameters. We calculated statistical power for continuous and dichotomous outcomes for samples sizes of (500,1000) and (2000,4000), respectively, as well as the minimum sample size necessary to achieve 80% power under both adjusted significance thresholds, based upon simulated haplotype data for variants included in S_4.

Application: Prostate Cancer Risk

We first applied our gsSKAT approach to 333 cases and 349 controls using data from a whole-exome sequencing (WES) case-control study of hereditary prostate cancer (PRCA) conducted by the International Consortium of Prostate Cancer Genetics (ICPCG), previously detailed elsewhere[23]. Briefly, WES was performed using the Agilent 50Mb SureSelect Human All Exon chip or the Agilent SureSelect V4+UTR kit. Eligible variants for association testing were then identified as corresponding to MAF < 0.05 and corresponding to functional status of nonsense, missense, or splice site variation. SKAT-O was then applied at a gene-level to all genes with ≥ 1 eligible variant position, adjusting for capture kit (dichotomous) and the first five principal components derived from the complete genotype data. Variant weighting was defined by functional impact, such that nonsense and splice-site variants received weights of 1.0 and missense variants were weighted using random

This article is protected by copyright. All rights reserved.

Author Manuscript

forest classification trees built using 15 features from dbNSFP, including seven functional prediction scores. Additional processing details are presented in the Supplemental Information. We considered all gene-sets with set size between 5 and 1000 genes within the curated pathways (CP) subset of the MSigDB C2 list, which includes 1330 gene-set definitions from Reactome, KEGG[24], PID[25], and other online gene-set databases, along with gene-sets from GO.

Application: Schizophrenia

A primary advantage of gsSKAT is that it can easily be applied to previously published gene-level rare-variant analysis results. Recently, Richards et al.[9] conducted a large case-control study of schizophrenia using the Illumina (San Diego, CA) HumanExome Array on 8103 controls and 5585 cases from the Schizophrenia Working Group of the Psychiatric Genomics Consortium (PGC) [26]. Notably, the authors had applied SKAT-O at both a gene and gene-set level, with complete results from the former reported as supplemental material. For purposes of comparison and to explore the utility of applying gsSKAT to pre-existing results of other studies, we applied gsSKAT to the reported SKAT-O p-values from Richards et al. using the same settings selected for the PRCA analysis.

RESULTS

Gene-Set P-values

Gene-set p-values derived from direct application of SKAT on $\boldsymbol{G_S}$ demonstrated high correlation with those computed using the Stouffer combined Z approach, with Spearman correlation coefficients ranging from 0.84 to 0.92. Results were comparable across gene-set size, sample size, and outcome type. In contrast, the results for SKAT-O, while still highly correlated, demonstrated a sizable reduction in rank correlation between the directly computed and Stouffer combined results (Figure 1). The degree of correlation also reduced with the gene-set size, with rank correlations for S_1 (6 genes) ranging from 0.59 to 0.62, while those for S_{10} (47 genes) ranged from 0.45 to 0.50. This is not altogether surprising, since as we previously noted, the correlation parameter ρ_j in SKAT-O is estimated independently for each component gene in the gene-set prior to p-value aggregation, thus resulting in a different underlying hypothesis. Moreover, it is expected that the correlation between direct application of SKAT-O and aggregation of gene-level SKAT-O p-values would degrade as the number of genes increases.

Type I Error Rate

Table II presents the empirical false positive rates (FPRs) for each of the 10 example Reactome gene-sets across the various conditions in our simulation study. Overall, direct application of SKAT or SKAT-O to the aggregated genotype data tended to be conservative, with binomial testing relative to the expected FPR at an α level of 0.05 resulting in significant ($P \leq \frac{0.05}{160} \approx 0.0003$) deviation for multiple simulation scenarios. This degree of conservative bias demonstrated a trend with gene-set size, with larger gene-sets trending to more conservative results. The conservative bias also appeared to be more prevalent and severe for continuous rather than binary outcomes. In contrast, the Stouffer's combined Z p-values from gsSKAT were largely within expectation under the null, with one instance of a simulation condition resulting in significantly inflated type I errors (S_{10} for a binary outcome using SKAT and N = 2000: FPR = 0.0632).

Multiple Testing Correction

This article is protected by copyright. All rights reserved.

Author Manuscript

For the three separate gene-set lists, estimates of the effective number of tests were approximately half the quantity of respective gene-sets themselves using R^{eff}_{Galwey}, while R^{eff}_{Gao} ranged between 72% and 81% of the total number of gene-sets R(Table III). With respect to FWER control using R and the two effective test number estimates, binomial testing indicated significantly ($P \leq \frac{0.05}{9} = 0.0056$) conservative FWER control using R as a Bonferroni correction factor but significantly liberal results for R^{eff}_{Galwey}. In contrast, use of R^{eff}_{Gao} resulted in FWER estimates that were approximately equal to the assumed testing size, regardless of which list of gene-sets was applied.

Power

Complete simulation results for all gene-sets under similar are presented in Table IV. Overall, we observed a trend of a reduction in power for the gsSKAT analysis of gene-level p-values relative to methods using complete data, with differentials as high as 15-20% in isolated instances. This was observed notably for continuous outcomes under simulation scenarios where gene-sets were comprised of relatively few genes and causal variant effects that were in the same direction, while differences in power for binary outcomes tended to be less pronounced. For larger gene-sets with a mixture of effect directions, the gsSKAT approach using SKAT-O p-values was competitive even with aSPUpath, especially for binary outcomes when the proportion of causal genes was lower. Given that median gene-set size for pathway collections defined by A_1, A_2, and A_3 ranged from 26-34 genes, results for gene-sets S_4, S_6, and S_8 are likely most relevant for broad characterization of relative power. Although the adaptive approach aSPUpath did not demonstrate substantially strong power advantages over the SKAT kernel-regression methods as previously reported, it is important to note that our simulation conditions supported the default variant weighting assumptions of SKAT-O by establishing a negative correlation between effect size and MAF.

In general, selection of R^{eff} (based upon the Gao approach) versus R as a Bonferroni correction factor yielded minimal advantages in statistical power over the various study conditions (Table V), with realized gains in power ranging from 0.009 to 0.017. Similarly, the minimum sample sizes to achieve 80% power using the R^{eff} correction factor resulted in 0.8-2.5% decrease in number of necessary samples across the simulations. Results were comparable across the nature of the phenotype (binary or continuous) and exhibited reduced benefit as sample size increased. Results for other example gene-sets and simulation conditions produced similar results (data not shown).

Data Applications

Gene-set analyses for both PRCA and schizophrenia respectively returned one and two significant findings (Figure 3). For PRCA, the GO gene-set Regulation of Apoptosis (GO:0042981), which interestingly would not have been declared significant by standard Bonferonni adjustment, consisted of 305 genes with observed genetic variation out of 341 total genes. A Manhattan-style plot of the component gene-level p-values is presented in Figure S1, with *FASTK* (P = 8.6E-05) corresponding to the most significant result. The two significant results from the schizophrenia analysis were also GO gene-sets, Membrane Lipid Biosynthetic Process (GO:0046467) and Lipid Biosynthetic Process (GO:0008610), the former nested within the latter. For GO:0046467, a total of 13/35 genes corresponded to a p-value < 0.1. Comparisons between the subset of reported gene-sets by Richards et al. that overlap with our results are reported in Table S2.

Computational Efficiency

This article is protected by copyright. All rights reserved.

Author Manuscript

Conditional on the availability of pre-computed gene-level p-values, computational timing for computing the gene-set p-values and deriving R_{eff} is largely trivial. The average analysis time for the FWER simulations per replication of the A_3 scale analysis was approximately 170 seconds on a desktop computer with an Intel® Core™ i5-6500 processor and 8 GB of RAM. Computing gene-set p-values from gene-level results also precludes the need to apply SKAT to potentially high-dimensional genotype matrices under large M_S, as these analyses may incur hours of compute time per gene-set when $M_S > 1000$.

DISCUSSION

As sequencing costs diminish and the number of published rare-variant association studies continues to grow, secondary analysis of gene-level testing results may identify significant aggregation of rare-variant associations in relevant biological pathways. Herein, we have outlined a simple and effective strategy for conducting large-scale gene-set analyses using gene-level kernel testing results while simultaneously investigating multiple testing correction and correlated test statistics due to genic overlap. From our simulations, we found p-value aggregation using Stouffer's combined Z method to be highly correlated with results corresponding to direct application of kernel-machine testing across a wide variety of conditions, particularly for SKAT. These methods also demonstrated proper test size across various sizes of gene-sets. Results were expectedly less correlated for SKAT-O, for which the parameter that determines the optimal combination of burden and kernel testing is estimated per gene rather than across the entirely of the gene-set. This distinction is subtle yet clearly relevant to the end result of the hypothesis test. Accommodating gene-level control of this parameter is equivalent to allowing the distribution of variant weight parameters to vary by gene with respect to distribution of effect directionality and causal status, which may have biological justification.

Our simulation results indicated that, while proper control of Type I errors is preserved in our gsSKAT p-value aggregation testing strategy, a moderate degree of statistical power may be lost relative to direct application of leading gene-set testing methods on the raw data. Consequently, ideal analysis conditions would involve access to all genetic, covariate, and phenotype information to apply methods such as aSPUpath. However, gsSKAT did perform competitively for gene-sets that are typical of large pathway definition sets such as KEGG or Reactome, and may also prove useful as a complementary analytical approach for sizable gene-sets that present computational issues for kernel testing. Additionally, the major advantage of our approach is that it can easily be applied to existing sequencing studies for which gene-level p-values are readily available. Given the growing number of sequencing studies and wide-scale adoption of kernel testing for gene-level rare-variant testing, gsSKAT provides an efficient approach for hypothesis generation using simple summary statistics from prior studies.

Our results also suggest that the practical implications of proper estimation of the effective number of tests for large agnostic scans of available gene-sets appear largely inconsequential. Although strict Bonferroni adjustment resulted in significantly conservative control of the FWER, the degree of this deviation was relatively marginal. As demonstrated by our proposed typical analyses of up to 6124 gene-sets, the effective number of tests tends to be approximately 70-80% of the total gene-set count, based upon the Gao estimator of R^{eff}. However, targeted analyses of relatively few highly-overlapping gene-sets may induce a stronger correlation structure. Consequently, estimation of R^{eff} may be useful on a case-by-case basis to evaluate the test statistic correlation structure for a given analysis when designing a pathway-centric sequencing study.

This article is protected by copyright. All rights reserved.

Author Manuscript

We applied gsSKAT to two different case-control studies of rare variation: the ICPCG case-control WES study of hereditary PRCA as well as a previously published large case-control study of schizophrenia using exome chip data by the PGC. For the former, we identified one significant result, the GO gene-set GO:0042981, which encompasses 341 genes that modulate the occurrence and/or rate of apoptotic processes. Programmed cell death is a fundamental mechanism of tumor suppression, and genetic damage of apoptotic pathways is a key stage in tumorigenesis[27]. Although the scale of this finding in regards to number of genes renders interpretation relatively difficult, follow-up custom-capture sequencing of the gene-set members in a replication cohort could assist in resolving what elements of apoptotic pathways are relevant to hereditary PRCA.

The schizophrenia analysis results from gsSKAT are unexpected in that Richards et al.[9] also conducted exploratory gene-set analyses on these data, albeit by directly applying SKAT-O to the complete variant sets. However, neither of the two hits we identified were among the subset of moderately associated ($P<0.01$) gene-sets reported by Richards et al., despite their inclusion in the afore-mentioned analyses. Although the correlation between direct SKAT-O results and those obtained via gene-level p-value aggregation is markedly reduced in comparison to SKAT, it is nonetheless surprising to observe such a disparity. The gene-set in question is the GO set GO:0046467, a biological process gene-set corresponding to reactions or molecular pathways involved in the formation of membrane lipids. There is a large body of literature indicating membrane lipids play a role in the development of schizophrenia[28-30]. Recently, Steen et al.[31] also published research implicating SREBP-mediated lipid biosynthesis via genetic variation in transcription factor genes *SREBF1* and *SREBF2* in the development of schizophrenia. The top gene-level result within this pathway was phosphatidylinositol glycan anchor biosynthesis class G (*PIGG*: P = 1.8E-04), which belongs to a class of genes involved in glycosylphosphatidylinositol biosynthesis. Pathogenic genetic variation in *PIGG* was recently identified to cause intellectual disability with hypotonia[32]. Interestingly, multiple PIG genes were also recently identified to overlap copy number variations associated with schizophrenia[33]

A notable shortcoming of our strategy is that it necessitates the use of linear kernel functions in the kernel-based score testing to satisfy the underlying assumptions. Extensions to non-linear kernel functions, such as the Gaussian kernel, are limited by the inability to decompose the composite gene-set kernel matrix into a linear combination of gene-level component kernels. Consequently, not only must Q_S be explicitly computed from the complete data, but deriving the covariance structure of gene-set statistics requires much more rigor. However, this could still be obtained relatively efficiently using parallel computing and/or trace inequalities. Extending meta-analytic strategies for SKAT[34] to aggregate results not only across genes in gene-sets but also across different studies is another research avenue of interest. Finally, we would recommend to investigators with access to complete data to adopt previously mentioned adaptive approaches for rare-variant gene-set analyses, such as SKAT-ARTP[10] or aSPUpath[12], as the corresponding improvements in statistical power reported by the authors likely justify the computational costs. Additionally, permutation schemes that adapt to the estimated null tail probability can render these approaches feasible for larger agnostic scans, although careful consideration is necessary for sufficient permutations to appropriately control the Type I error rate under high multiple testing dimensionality[35]. We have conveniently provided an R implementation for the above-described methods with code available in the appendix. We have also currently built in support for the Gene Matrix Transposed (GMT) file format to easily import gene-set definitions and readily enable rare-variant gene-set scans of popular curated gene-sets available on the MSigDB website.

This article is protected by copyright. All rights reserved.

Author Manuscript

As our understanding of molecular pathways and their modularity continues to improve, leveraging curated gene-sets may improve rare-variant association studies of complex traits where genetic heterogeneity is plausible. Our work demonstrates that applying p-value aggregation methods to gene-level kernel testing results can be an effective strategy for large-scale evaluation of gene-sets using summary statistics from rare-variant association studies.

ACKNOWLEDGMENTS

This research was supported by the US Public Health Service, National Institutes of Health (NIH), contract Grant Number GM065450 (DJS) and National Cancer Institute, Grant number U01 CA 89600 (SNT), and the Mayo Clinic Center for Individualized Medicine. JEBW is supported by the Intramural Program of the National Human Genome Research Institute, NIH.

CONFLICTS OF INTEREST

None.

REFERENCES

1. Gibson G: **Rare and common variants: twenty arguments**. *Nat Rev Genet* 2011, **13**(2):135-145.
2. Bodmer W, Bonilla C: **Common and rare variants in multifactorial susceptibility to common diseases**. *Nat Genet* 2008, **40**(6):695-701.
3. Schork NJ, Murray SS, Frazer KA, Topol EJ: **Common vs. rare allele hypotheses for complex diseases**. *Curr Opin Genet Dev* 2009, **19**(3):212-219.
4. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, Boucher G, Ripke S, Ellinghaus D, Burtt N *et al*: **Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease**. *Nat Genet* 2011, **43**(11):1066-1073.
5. Holm H, Gudbjartsson DF, Sulem P, Masson G, Helgadottir HT, Zanon C, Magnusson OT, Helgason A, Saemundsdottir J, Gylfason A *et al*: **A rare variant in MYH6 is associated with high risk of sick sinus syndrome**. *Nat Genet* 2011, **43**(4):316-320.
6. Wu G, Zhi D: **Pathway-based approaches for sequencing-based genome-wide association studies**. *Genet Epidemiol* 2013, **37**(5):478-494.
7. Li Z, Huang Y, Li H, Hu J, Liu X, Jiang T, Sun G, Tang A, Sun X, Qian W *et al*: **Excess of rare variants in genes that are key epigenetic regulators of spermatogenesis in the patients with non-obstructive azoospermia**. *Sci Rep* 2015, **5**:8785.
8. Haller G, Alvarado D, McCall K, Yang P, Cruchaga C, Harms M, Goate A, Willing M, Morcuende JA, Baschal E *et al*: **A polygenic burden of rare variants across extracellular matrix genes among individuals with adolescent idiopathic scoliosis**. *Hum Mol Genet* 2016, **25**(1):202-209.
9. Richards AL, Leonenko G, Walters JT, Kavanagh DH, Rees EG, Evans A, Chambert KD, Moran JL, Goldstein J, Neale BM *et al*: **Exome arrays capture polygenic rare variant contributions to schizophrenia**. *Hum Mol Genet* 2016, **25**(5):1001-1007.
10. Yan Q, Tiwari HK, Yi N, Lin WY, Gao G, Lou XY, Cui X, Liu N: **Kernel-machine testing coupled with a rank-truncation method for genetic pathway analysis**. *Genet Epidemiol* 2014, **38**(5):447-456.
11. Wu MC, Lee S, Cai TX, Li Y, Boehnke M, Lin XH: **Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test**. *Am J Hum Genet* 2011, **89**(1):82-93.
12. Pan W, Kwak IY, Wei P: **A Powerful Pathway-Based Adaptive Test for Genetic Association with Common or Rare Variants**. *Am J Hum Genet* 2015, **97**(1):86-98.

This article is protected by copyright. All rights reserved.

Author Manuscript

13. Liberzon A: **A description of the Molecular Signatures Database (MSigDB) Web site**. *Methods Mol Biol* 2014, **1150**:153-160.
14. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP: **Molecular signatures database (MSigDB) 3.0**. *Bioinformatics* 2011, **27**(12):1739-1740.
15. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.
16. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X: **Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies**. *Am J Hum Genet* 2012, **91**(2):224-237.
17. Fisher RA: **Statistical methods for research workers**. Edinburgh, London,: Oliver and Boyd; 1925.
18. Stouffer SA: **The American soldier**. Princeton,: Princeton University Press; 1949.
19. Abelson RP: **Statistics as principled argument**. Hillsdale, N.J.: L. Erlbaum Associates; 1995.
20. Galwey NW: **A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests**. *Genet Epidemiol* 2009, **33**(7):559-568.
21. Hendricks AE, Dupuis J, Logue MW, Myers RH, Lunetta KL: **Correction for multiple testing in a gene region**. *Eur J Hum Genet* 2014, **22**(3):414-418.
22. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR *et al*: **The Reactome pathway knowledgebase**. *Nucleic Acids Res* 2014, **42**(Database issue):D472-477.
23. Larson NB, McDonnell S, Albright LC, Teerlink C, Stanford J, Ostrander EA, Isaacs WB, Xu J, Cooney KA, Lange E *et al*: **Post hoc Analysis for Detecting Individual Rare Variant Risk Associations Using Probit Regression Bayesian Variable Selection Methods in Case-Control Sequencing Studies**. *Genet Epidemiol* 2016.
24. Kanehisa M: **The KEGG database**. *Novartis Found Symp* 2002, **247**:91-101; discussion 101-103, 119-128, 244-152.
25. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH: **PID: the Pathway Interaction Database**. *Nucleic Acids Res* 2009, **37**:D674-D679.
26. Schizophrenia Working Group of the Psychiatric Genomics C: **Biological insights from 108 schizophrenia-associated genetic loci**. *Nature* 2014, **511**(7510):421-427.
27. Brown JM, Attardi LD: **The role of apoptosis in cancer development and treatment response**. *Nat Rev Cancer* 2005, **5**(3):231-237.
28. Solberg DK, Bentsen H, Refsum H, Andreassen OA: **Association between serum lipids and membrane fatty acids and clinical characteristics in patients with schizophrenia**. *Acta Psychiatr Scand* 2015, **132**(4):293-300.
29. Fenton WS, Hibbeln J, Knable M: **Essential fatty acids, lipid membrane abnormalities, and the diagnosis and treatment of schizophrenia**. *Biol Psychiatry* 2000, **47**(1):8-21.
30. Dodd GH: **The lipid membrane hypothesis of schizophrenia: implications for possible clinical breath tests**. *Prostaglandins Leukot Essent Fatty Acids* 1996, **55**(1-2):95-99.
31. Steen VM, Skrede S, Polushina T, Lopez M, Andreassen OA, Ferno J, Hellard SL: **Genetic evidence for a role of the SREBP transcription system and lipid biosynthesis in schizophrenia and antipsychotic treatment**. *Eur Neuropsychopharmacol* 2016.
32. Makrythanasis P, Kato M, Zaki MS, Saitsu H, Nakamura K, Santoni FA, Miyatake S, Nakashima M, Issa MY, Guipponi M *et al*: **Pathogenic Variants in PIGG Cause Intellectual Disability with Seizures and Hypotonia**. *Am J Hum Genet* 2016, **98**(4):615-626.
33. Morris BJ, Pratt JA: **Novel treatment strategies for schizophrenia from improved understanding of genetic risk**. *Clin Genet* 2014, **86**(5):401-411.

This article is protected by copyright. All rights reserved.

Author Manuscript

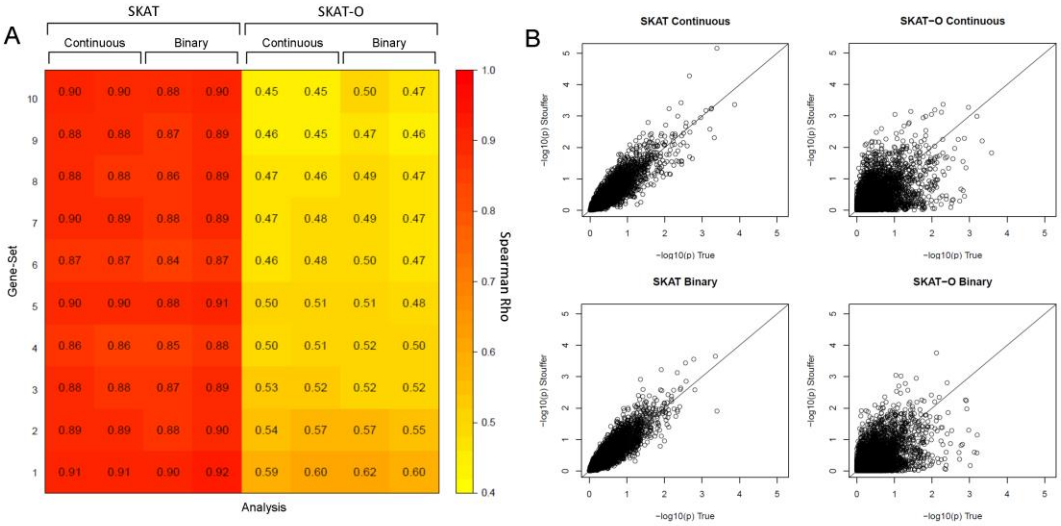
34. Lee S, Teslovich TM, Boehnke M, Lin X: **General framework for meta-analysis of rare variants in sequencing association studies**. *Am J Hum Genet* 2013, **93**(1):42-53.
35. Phipson B, Smyth GK: **Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn**. *Stat Appl Genet Mol* 2010, **9**(1).

Author Manuscript

This article is protected by copyright. All rights reserved.

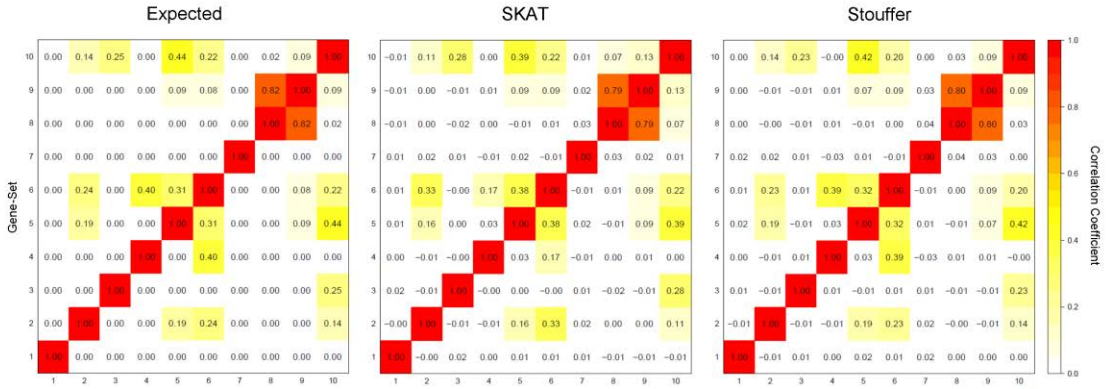
FIGURE LEGENDS

Figure 1: (A) Tileplot of Spearman correlation coefficients between p-values derived from the true kernel test statistic (i.e., SKAT or SKAT-O applied to $\boldsymbol{G_S}$) with the p-values from the Stouffer combined Z approach based on gene-level p-values. Tiles are colored by degree of correlation as indicated by the key, with correlation estimates included in the tiles. Adjacent columns corresponding to a particular analysis configuration (indicated by brackets on top of the figure) are for $N = 1000$ (left) and $N = 2000$ (right). **(B)** Example scatterplots for –log10 transformed p-values corresponding to the true kernel gene-set analysis test (x-axis) and the Stouffer combined Z p-value (y-axis) for gene-set S_5 under $N = 2000$.



This article is protected by copyright. All rights reserved.

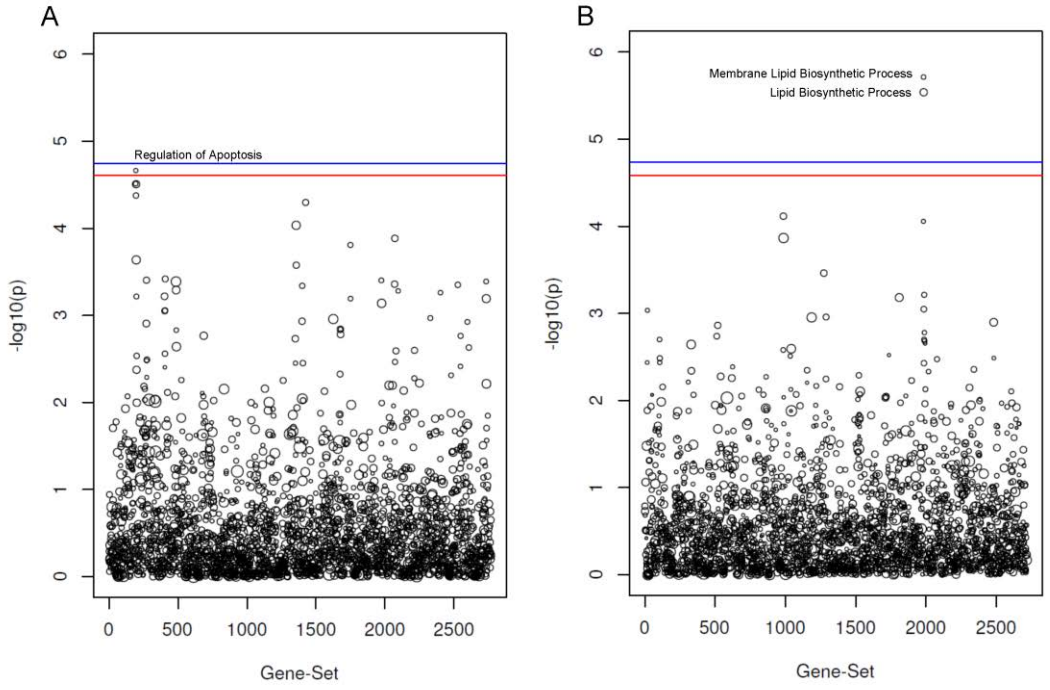
Figure 2: Visualization of expected and example observed correlation structures from the gene-set rare-variant association simulation study. Correlation matrices corresponding to (from left to right) the design matrix $\boldsymbol{D}$ for the 10 Reactome gene-sets, Spearman rank correlations from the SKAT analysis p-values of continuous outcomes based on $N = 2000$, and the Spearman rank correlations for the Stouffer combined Z p-values under the same simulations conditions.



Author

This article is protected by copyright. All rights reserved.

Figure 3: Gene-set Manhattan-style plot based upon gene-level SKAT-O analysis results for **(A)** hereditary PRCA and **(B)** schizophrenia case-control studies. Gene-sets for each figure are initially sorted by significance and then re-ordered on the basis of expected correlation structure due to overlapping genes as well as p-value. The blue and red lines indicate Bonferroni-adjusted significance thresholds based upon the total number and effective number of gene-set tests, respectively, while the size of each point is a function of the total number of genes per gene-set. Gene-sets declared significantly associated by the latter are highlighted by gene-set name.



				Gene Overlap									
Gene-Set	L_S	M_S	cMAF	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
S_1	6	227	0.443	-	0	0	0	0	0	0	0	0	0
S_2	10	473	0.895	-	-	0	0	3	4	0	0	0	3
S_3	12	634	0.790	-	-	-	0	0	0	0	0	0	6
S_4	19	868	1.386	-	-	-	-	0	9	0	0	0	0
S_5	25	1255	1.888	-	-	-	-	-	8	0	0	3	15
S_6	27	1174	1.997	-	-	-	-	-	-	0	0	3	8
S_7	32	1879	5.477	-	-	-	-	-	-	-	0	0	0
S_8	37	2130	3.027	-	-	-	-	-	-	-	-	34	1
S_9	47	2611	4.097	-	-	-	-	-	-	-	-	-	4
S_{10}	47	2621	4.591	-	-	-	-	-	-	-	-	-	-

Table I: Gene-Set properties and gene overlap for 10 randomly selected REACTOME gene-sets.

This article is protected by copyright. All rights reserved.

Pheno.	N	Method	P-value	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
Cont.	1000	SKAT	True	0.0440	0.0442	0.0416	0.0396	**0.0356**	**0.0348**	**0.0258**	**0.0272**	**0.0228**	**0.0224**
			Stouffer	0.0518	0.0526	0.0498	0.0484	0.0536	0.0542	0.0512	0.0474	0.0530	0.0472
	2000	SKAT	True	0.0466	0.0474	0.0444	0.0412	0.0422	0.0396	**0.0352**	**0.0334**	**0.0312**	**0.0352**
			Stouffer	0.0452	0.0464	0.0486	0.0514	0.0550	0.0502	0.0494	0.0462	0.0446	0.0558
	1000	SKAT-O	True	0.0536	0.0494	0.0502	0.0448	0.0490	0.0404	**0.0362**	0.0396	0.0392	**0.0318**
			Stouffer	0.0546	0.0510	0.0462	0.0446	**0.0382**	**0.0368**	0.0396	**0.0376**	**0.0348**	**0.0346**
	2000	SKAT-O	True	0.0494	0.0530	0.0464	0.0458	0.0508	0.0446	0.0430	0.0414	**0.0372**	0.0430
			Stouffer	0.0444	0.0480	0.0436	0.0450	0.0462	0.0392	0.0434	**0.0354**	**0.0332**	**0.0382**
Binary	1000	SKAT	True	**0.0626**	0.0512	0.0548	0.0526	0.0528	0.0566	0.0484	0.0586	0.0574	0.0566
			Stouffer	0.0594	0.0518	0.0572	0.0504	0.0500	0.0564	0.0494	0.0552	0.0556	0.0520
	2000	SKAT	True	0.0472	0.0448	0.0444	0.0416	0.0392	0.0432	**0.0348**	**0.0374**	**0.0366**	**0.0360**
			Stouffer	0.0492	0.0468	0.0510	0.0536	0.0530	0.0558	0.0496	0.0520	0.0584	**0.0632**
	1000	SKAT-O	True	**0.0622**	0.0534	0.0558	0.0536	0.0442	0.0534	0.0482	0.0504	0.0482	0.0560
			Stouffer	0.0522	0.0500	0.0552	0.0556	0.0512	0.0558	0.0454	0.0532	0.0532	0.0530
	2000	SKAT-O	True	0.0494	0.0440	0.0458	0.0520	0.0428	0.0474	0.0456	0.0466	0.0412	0.0428
			Stouffer	0.0480	0.0450	0.0432	0.0448	0.0422	**0.0366**	**0.0364**	0.0402	**0.0340**	**0.0348**

Table II: Empirical false positive rates for the direct application of SKAT/SKAT-O (i.e., "True") and Stouffer combined Z gene-set p-values from the null simulation study based upon an alpha level of 0.05. Error rates significantly different from the expectation of 0.05 based upon a Bonferroni-adjusted (160 tests) two-sided binomial test are highlighted in bold.

Author Manuscript

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/gepi.22036.

This article is protected by copyright. All rights reserved.

Analysis	Number of Gene-Sets			Number of Unique Genes	FWER by Correction Factor ($\alpha = 0.05$)		
	R	R^{eff}_{Galwey}	R^{eff}_{Gao}		R	R^{eff}_{Galwey}	R^{eff}_{Gao}
A_1	673	368.7	500	5637	**0.0416**	**0.0731**	0.0550
A_2	2109	1084.2	1515	9701	**0.0398**	**0.0727**	0.0537
A_3	6124	3885.1	4966	17272	**0.0418**	**0.0674**	0.0511

Table III: Gene-set analysis summaries and empirical FWER results by Bonferroni correction factor using the total number of gene-set test R as well as two effective test number estimates, R^{eff}_{Gao} and R^{eff}_{Galwey}. FWERs significantly different from the expectation of 0.05 based upon a Bonferroni-adjusted (9 tests) two-sided binomial test are highlighted in bold.

Author Manuscript

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/gepi.22036.

This article is protected by copyright. All rights reserved.

Effect Direction	Causal Gene %		Continuous Outcome			Binary Outcome		
		Gene-set	SKAT-O	aSPUpath	gsSKAT	SKAT-O	aSPUpath	gsSKAT
One-way	100	S_2	**0.570**	0.487	0.450	**0.635**	0.492	0.480
		S_4	**0.581**	0.463	0.453	**0.664**	0.458	0.478
		S_6	**0.825**	0.666	0.608	**0.903**	0.715	0.725
		S_8	**0.991**	0.971	0.915	**0.984**	0.960	0.939
		S_{10}	**0.972**	0.898	0.889	**0.993**	0.917	0.927
	50	S_2	0.275	**0.285**	0.218	**0.305**	0.257	0.257
		S_4	**0.298**	0.267	0.209	**0.282**	0.204	0.221
		S_6	**0.486**	0.385	0.359	**0.521**	0.385	0.401
		S_8	0.712	**0.742**	0.600	**0.764**	0.707	0.658
		S_{10}	**0.674**	0.596	0.542	**0.755**	0.595	0.628
	25	S_2	0.138	**0.161**	0.118	0.091	**0.119**	0.088
		S_4	0.171	**0.190**	0.128	0.138	0.127	**0.145**
		S_6	0.194	**0.226**	0.176	**0.215**	0.213	0.180
		S_8	0.328	**0.479**	0.293	0.316	**0.364**	0.309
		S_{10}	0.280	**0.362**	0.239	**0.356**	0.315	0.312
Mixed	100	S_2	**0.529**	0.473	0.447	**0.403**	0.349	0.359
		S_4	**0.526**	0.457	0.447	**0.401**	0.315	0.332
		S_6	**0.752**	0.627	0.608	**0.638**	0.492	0.561
		S_8	**0.955**	0.957	0.915	**0.924**	0.883	0.872
		S_{10}	**0.937**	0.882	0.889	**0.899**	0.775	0.871
	50	S_2	**0.274**	0.266	0.229	**0.216**	0.212	0.175
		S_4	**0.261**	0.250	0.206	0.176	**0.191**	0.171
		S_6	**0.418**	0.380	0.339	**0.314**	0.257	0.298
		S_8	0.630	**0.733**	0.547	0.533	**0.545**	0.518
		S_{10}	0.601	**0.607**	0.559	0.469	0.418	**0.492**
	25	S_2	0.131	**0.164**	0.122	0.091	**0.105**	0.082
		S_4	0.158	**0.167**	0.148	**0.113**	0.109	0.106
		S_6	0.192	**0.210**	0.158	0.144	0.148	**0.150**
		S_8	0.299	**0.464**	0.316	0.193	**0.267**	0.227
		S_{10}	0.217	**0.342**	0.234	0.189	0.238	**0.257**

Table IV: Power simulation results for binary and continuous outcomes for analysis conditions defined by effect direction (one-way or randomly mixed), causal gene percentage (100%, 50%, or 25%), and gene-set. Empirical power estimates are derived at the α level of 0.05, with highest level per scenario highlighted in bold.

Author Manuscript

This article is protected by copyright. All rights reserved.

		Continuous			Binary		
Analysis	Corr. Factor	N = 500	N = 1000	Min N ($\beta = 0.8$)	N = 2000	N = 4000	Min N ($\beta = 0.8$)
A_1	R	0.214	0.666	1345	0.356	0.789	4118
	R_{Gao}^{eff}	0.229	0.680	1320	0.373	0.799	4014
A_2	R	0.165	0.614	1427	0.295	0.749	4458
	R_{Gao}^{eff}	0.178	0.629	1405	0.311	0.760	4367
A_3	R	0.129	0.566	1491	0.245	0.711	4781
	R_{Gao}^{eff}	0.136	0.575	1480	0.254	0.719	4721

Table V: Estimates of statistical power and minimum sample size to achieve 80% power for gene-set S_4 under a variety of sample size and total gene-set conditions. Results are presented at alpha levels defined by Bonferroni correction using total number of gene-sets R per analysis A_1,A_2, and A_3, as well as corresponding estimate of effective number of tests R_{Gao}^{eff}.

Author Manuscript

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/gepi.22036.

This article is protected by copyright. All rights reserved.