Published in final edited form as:

Genet Epidemiol. 2014 July ; 38(5): 447–456. doi:10.1002/gepi.21813.

Kernel-Machine Testing Coupled with a Rank-Truncation Method for Genetic Pathway Analysis

Qi Yan¹, Hemant K. Tiwari¹, Nengjun Yi¹, Wan-Yu Lin², Guimin Gao³, Xiang-Yang Lou¹, Xiangqin Cui¹, and Nianjun Liu^{1,*}

¹Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, 35294, USA

²Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan

³Departments of Biostatistics, School of Medicine, Virginia Commonwealth University, Richmond, VA, USA

Abstract

Traditional genome-wide association studies (GWASs) usually focus on single-marker analysis, which only accesses marginal effects. Pathway analysis, on the other hand, considers biological pathway-gene-marker hierarchical structure and therefore provides additional insights into the genetic architecture underlining complex diseases. Recently, a number of methods for pathway analysis have been proposed to assess the significance of a biological pathway from a collection of single nucleotide polymorphisms. In this study, we propose a novel approach for pathway analysis that assesses the effects of genes using the sequence-kernel-association test (SKAT) and the effects of pathways using an extended adaptive-rank-truncated product (ARTP) statistic. It has been increasingly recognized that complex diseases are caused by both common and rare variants. We propose a new weighting scheme for genetic variants across the whole allelic frequency spectrum to be analyzed together without any form of frequency cutoff for defining rare variants. The proposed approach is flexible. It is applicable to both binary and continuous traits, and incorporating covariates is easy. Furthermore, it can be readily applied to GWAS data, exome-sequencing data, and deep re-sequencing data. We evaluate the new approach on data simulated under comprehensive scenarios and show that it has the highest power in most of the scenarios while maintaining the correct Type I error rate. We also apply our proposed methodology to data from a study of the association between bipolar disorder and candidate pathways from WTCCC to show its utility.

Keywords

pathway analysis; sequence-kernel-association test; rare variants; common variants; truncation

*Correspondence to: Nianjun Liu, RPHB 420A, 1665 University Boulevard, Birmingham, AL 35294; Phone: 205-975-9190; nliu@uab.edu.

The authors also declare that they have no conflict of interest.

Introduction

Genome-wide association studies (GWASs) have been widely used for identifying common single nucleotide polymorphisms (SNPs) that are associated with complex human diseases [Hindorff, et al. 2009; Hunter, et al. 2007; Manolio, et al. 2008; Wellcome Trust Case Control 2007; Yeager, et al. 2007]. Although GWASs have successfully discovered many genetic variants for several complex diseases [Hindorff, et al. 2009], there are limitations for the single-marker analysis method commonly used in GWAS. First, common human diseases are usually caused by multiple variants that have small individual effects, and therefore dramatically important genetic variants can be overlooked by single-marker-based analysis, which only accesses marginal effects [Shi and Weinberg 2011]. Second, single-marker analysis does not handle genetic heterogeneity well. Third, the large number of genetic markers in GWAS will incur a heavy penalty for multiple testing. Consequently, single genetic variants that contribute weak but real effects on disease risks are likely to be missed after taking the multiple-comparison adjustment into account [McCarthy, et al. 2008]. In addition, single-marker association tests cannot handle genetic hierarchy (i.e., the three-level, SNP-gene-pathway hierarchical structure) appropriately. SNPs can be involved in a specific gene, and a group of SNPs could be jointly associated with diseases/traits. In addition, genes can be involved in a specific pathway that represents certain biological mechanisms of the complex diseases/traits. The traditional GWAS strategy ignores the biological pathway-gene-SNP hierarchical structure, which may result in certain genetic effects on the outcome being missed.

It has been increasingly recognized that pathway analysis offers an opportunity to complement the single-marker association test and to provide better understanding of the mechanisms underlying complex diseases [Chasman 2008; Holmans, et al. 2009; Torkamani, et al. 2008; Wang, et al. 2007]. The existing pathway analysis methods are broadly classified based on two criteria [Wang, et al. 2010]: (1) the required input data sets, and (2) the null hypothesis being tested. There are two types of input data sets: (1) a list of SNP p-values from a single-marker association test [Holmans, et al. 2009; Medina, et al. 2009; Nam, et al. 2010; Zhang, et al. 2010], and (2) raw genotype data [Chen, et al. 2010; Holden, et al. 2008; O'Dushlaine, et al. 2009; Purcell, et al. 2007; Wang, et al. 2007]. Additionally, pathway analysis can also be categorized into two types based the null hypothesis being tested: (1) a competitive hypothesis [Holden, et al. 2008; Holmans, et al. 2009; Medina, et al. 2009; Nam, et al. 2010; O'Dushlaine, et al. 2009; Wang, et al. 2007; Zhang, et al. 2010] that compares genes from the given pathway of interest with the rest of the genes from other pathways, and (2) a self-contained hypothesis [Chen, et al. 2010; Purcell, et al. 2007] that compares genes from the given pathway of interest with the null (non-associated) genomic background.

The sequence-kernel-association test (SKAT) [Wu, et al. 2010; Wu, et al. 2011] is a flexible, computationally efficient, set-based regression approach. Although SKAT was initially designed to examine the association significance at the gene level, it can be easily extended to the pathway level because a gene or a pathway is essentially a group of genetic markers. Another nice method for pathway analysis uses the product of some most significant p-

values from the single-marker association test as the summary statistics [Yu, et al. 2009]. Both methods have been widely used in genetic studies.

In this study, we propose a powerful pathway-analysis approach that combines the sequence-kernel-association test (SKAT) and an extended adaptive-rank-truncated product (ARTP) method (abbreviated as SKAT-ARTP, although this may be an abuse of the terminology), together with a new weighting scheme. This proposed SKAT-ARTP requires raw genotype data and tests self-contained null hypotheses. Specifically, SKAT is used to summarize the gene-level effect, and the extended ARTP is used to summarize pathway-level significance. This SKAT-ARTP method handles genetic variants in a natural SNP-gene-pathway hierarchical way, which allows for borrowing information across SNPs and genes in the same pathway and further clarifies the biological mechanism.

SKAT is known to perform well in detecting rare genetic variants (usually defined as minor allele frequency (MAF) less than 5%) because variants with low MAF are up-weighted and those with high MAF are down-weighted. This may lead to substantial power loss when both rare and common genetic variants influence disease susceptibility. However, increasing evidence has shown that complex diseases are often caused by both common and rare variants [Bodmer and Bonilla 2008; Ng, et al. 2009; Pritchard 2001; Pritchard and Cox 2002; Stratton and Rahman 2008; Teer and Mullikin 2010; Walsh and King 2007]. Here, we propose a new weighting scheme that allows for good power for genetic variants across the whole allelic frequency spectrum. All the methods for rare variants association tests, including burden tests [Han and Pan 2010; Li and Leal 2008; Madsen and Browning 2009; Morgenthaler and Thilly 2007; Price, et al. 2010; Zawistowski, et al. 2010] and variance-component tests [Lin, et al. 2013; Wu, et al. 2010; Wu, et al. 2011], need to have MAF thresholds, either fixed or variable, one or multiple, to define rare variants. Our weighting scheme does not need to distinguish rare or common variants, so no form of frequency cutoff is needed. Our weighting scheme can be used in any groupwise association test such as SKAT to handle both rare and common variants effectively.

Our proposed approach was evaluated with other methods via extensive simulation studies, including SKAT-ARTP, individual SKAT [Wu, et al. 2010; Wu, et al. 2011], and ARTP-ARTP [Yu, et al. 2009]. Our simulation study showed that the proposed method demonstrated the highest analytical power over other available models while preserving the Type I error rate for identifying the association between pathway and disease. Finally, our proposed method was applied to a data set from a bipolar study to illustrate its utility.

Methods

The New Method

The sequence-kernel-association test (SKAT) is shown to be powerful for analyzing sequence data with rare variants due to its efficient modeling of multiple variants jointly [Chen, et al. 2013; Johnson, et al. 2013; Kim, et al. 2013; Lee, et al. 2013; Liu, et al. 2013; Magi, et al. 2012; Wang, et al. 2013; Wu, et al. 2010; Wu, et al. 2011]. In genetic studies, only a small portion of variants is associated with phenotypes. The remaining variants are neutral and may add noise in the detection of causal variants. Excluding these neutral

variants from analysis may increase the statistical power. Truncation methods follow this idea [Lin, et al. 2014; Zaykin, et al. 2002]. Our proposed approach combines the advantages of kernel-machine testing and truncation methods. Specifically, this method uses SKAT [Wu, et al. 2010; Wu, et al. 2011] to summarize the gene-level signal and the adaptive-rank-truncated product (ARTP) algorithm [Yu, et al. 2009] to summarize the pathway-level signal (therefore it is denoted as SKAT-ARTP).

The Sequence-Kernel-Association Test (SKAT)—Let $\mathbf{y}$ be an $n \times 1$ vector coding the trait of n individuals. The link function $h(\cdot)$ is used to map the linear combination of predictors, η , to the conditional mean of $\mathbf{y}$, μ . We consider a linear model:

$$h(\mu) = \eta = X\beta + G\gamma,$$

where X is an $n \times p$ covariate matrix, β is a $p \times 1$ vector for covariates, G is an $n \times q$ genotype matrix for q genetic variants under study, and γ is a $q \times 1$ vector for the random effects of genetic variants. In order to test the significance of this set of variants, the null hypothesis is $H_0: \gamma = \mathbf{0}$, that is, $\gamma_1 = \gamma_2 = \dots \gamma_q = 0$. The random effect γ_j is assumed to be normally distributed with a mean of zero and a variance of τW_j for variant j . So the null hypothesis is equivalent to test $H_0: \tau = 0$, which can be tested with a variance-component score test, where τ is a variance component and W_j is a pre-specified weight. Specifically, the variance-component score statistic is [Wu, et al. 2011]

$$Q = (\mathbf{y} - \hat{\mu})' G W G' (\mathbf{y} - \hat{\mu}),$$

where $\hat{\mu}$ is the predicted mean of y under the null hypothesis. For dichotomous traits, $\hat{\mu} = \text{logit}^{-1}(X\hat{\beta})$. Here, $W = \text{diag}(w_1, \dots, w_q)$ contains the weights for the q variants. With matrix notation,

$$Q = [\mathbf{y}_1 - \hat{\mu}_1 \dots \mathbf{y}_n - \hat{\mu}_n]_{1 \times n} \begin{bmatrix} G_{11} & \dots & G_{1q} \\ \vdots & \ddots & \vdots \\ G_{n1} & \dots & G_{nq} \end{bmatrix}_{n \times q} \begin{bmatrix} w_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & w_q \end{bmatrix}_{q \times q} \begin{bmatrix} G_{11} & \dots & G_{n1} \\ \vdots & \ddots & \vdots \\ G_{1q} & \dots & G_{nq} \end{bmatrix}_{q \times n} \begin{bmatrix} \mathbf{y}_1 - \hat{\mu}_1 \\ \vdots \\ \mathbf{y}_n - \hat{\mu}_n \end{bmatrix}_{n \times 1}.$$

Under the null hypothesis, Q follows a mixture of Chi-square distributions, which can be closely approximated with the computationally efficient Davies method [Davies 1980].

A good choice of weights for variants can improve the power of SKAT. If the weight of the j th variant, w_j , is close to zero, it will have a small contribution to Q . Therefore, we should increase the weights of potential causal variants and decrease the weights of potentially non-causal variants. However, we do not know the causal variants in advance. Following the common disease-rare variant hypothesis, SKAT adopts the assumption that rare variants are more likely to be causal than common variants. A beta-density distribution with parameters $\alpha_1 = 1$ and $\alpha_2 = 25$ is recommended for weight function $\sqrt{w_j} = \text{Beta}(MAF_j; \alpha_1, \alpha_2)$, where MAF_j is the minor allele frequency of the j th variant. Of course, we can choose different values for α_1 and α_2 to adjust the weights on common and rare variants. Studies have shown

[Cai, et al. 2012; Yi, et al. 2011] that power could be increased if the effects of genetic variants are incorporated into the test. We propose a weight that is a function of single-marker association test p-value $\sqrt{w_j}=1/(0.1+pvalue)$ in order to take into account the information of the effects of the variants so as to assign large weights to variants with strong signals; i.e., variants with small p-values (0.1 in the denominator is to avoid the weight going to infinity when the p-values are extremely small). As mentioned in the Introduction, complex diseases are usually caused by the joint effects of both common and rare variants. Therefore, a good weight should consider both rare and common variants. The function of single-marker-based test p-values could be useful for calculating weights. However, while a single-marker test is usually powerful enough for detecting common variants, it is not powerful enough for rare variants. We need to balance the weight for the potential effects of rare variants. Thus, we propose to use the summation of beta density (e.g., $0.5 * Beta(MAF_j; 1, 25)$) and the inverse of the single-marker test p-value (e.g., $1.25/(0.1 + pvalue)$) as the square root of the weight, that is, $\sqrt{w_j}=0.5*Beta(MAF_j; 1, 25)+1.25/(0.1+pvalue)$. We calculate it this way to make the beta-density component (the range of the first part is 0~12.5) and the inverse of the p-value component (the range of the second part is 1.125~12.5) have similar ranges so that one does not dominate the other. Because phenotype information is used in the weight, in order to maintain the correct Type I error rate, permutation-based procedures are used to generate a Monte Carlo p-value.

The SKAT-ARTP Method—The ARTP method [Yu, et al. 2009] is a general p-value combining method that can be used at any level of data (e.g., the SNP or gene level). In the original paper, Yu et al. used ARTP at both SNP level and gene level (therefore it is denoted as ARTP-ARTP). Here, we employ it to combine gene-level p-values to calculate the pathway-level p-value. Before applying ARTP, each of the gene-level p-values based on the observed data is obtained by SKAT with new weights, denoted as $p_1^{(0)}, \dots, p_L^{(0)}$, where L is the number of genes in the pathway and superscript (0) indicates p-values from observed data. Then we permute the phenotypes to generate B data sets under the null hypothesis. Based on the b^{th} permuted data, $0 < b \leq B$, we can analogously obtain the gene-level p-values using SKAT, denoted as $p_1^{(b)}, \dots, p_L^{(b)}$, where superscript (b) indicates the b^{th} permuted data. After achieving gene-level p-values for both observed and permuted data, the ARTP algorithm is used as follows:

1. Based on $p_1^{(b)}, \dots, p_L^{(b)}$ from the gene-level test approach, $0 < b \leq B$, for any given b we calculate the rank-truncated product statistic for each candidate-truncation point, denoted as $V_j^{(b)} = \prod_{i=1}^j p_{(i)}^{(b)}$, $1 \leq j \leq J$, where $p_{(i)}^{(b)}$ is the ranked p-value in the b^{th} permuted data set ($p_{(1)}^{(b)}$ is the smallest p-value). Therefore, for the b^{th} permuted data set, we obtain $V_1^{(b)}, V_2^{(b)}, \dots, V_J^{(b)}$, in which J can be any number less than or equal to L. Here, J is equal to L, which is the number of genes in the pathway in order to consider all the gene combinations.

2. Based on $V_j^{(b)}, 1 \leq j \leq J, 0 \leq b \leq B$, for any given b apply

$\hat{S}_j^{(b)} = \frac{\sum_{b^*=0}^B I(V_j^{(b^*)} \leq V_j^{(b)})}{B+1}$ to obtain the corresponding p-values for $V_j^{(b)}$. Here, $b = 0$ is for the observed data that is used as part of the null distribution data in order to avoid using a nested simulation for the calculation of the p-value of the b^{th} permuted data, following the suggestion in [Becker and Knapp 2004; Yu, et al. 2009]. The observed data serve as “permuted data” when calculating the p-value of the b^{th} permuted data.

3. For any b , let $\text{Min}P^{(b)} = \min_{1 \leq j \leq J} \hat{S}_j^{(b)}, 0 \leq b \leq B$. The adjusted p-value for the adaptive rank-truncated-product statistic $\text{Min}P^{(0)}$ is estimated as

$$\frac{\sum_{b=0}^B I(\text{Min}P^{(b)} \leq \text{Min}P^{(0)})}{B+1}, \text{ and it is the p-value for the pathway.}$$

In summary, the powerful and efficient SKAT algorithm is first used to obtain gene-level p-values for all genes within the pathway. The ARTP algorithm is then applied to evaluate the association between the pathway and the phenotype. Because the observed data are used to calculate the p-values of the permuted data, it may be appropriate to report the p-value of the pathway as $\frac{1}{B+1}$, in the extreme case when $\text{Min}P^{(0)}$ is the smallest among all $\text{Min}P^{(b)}$.

Other Pathway-Analysis Approaches

In our extensive simulation studies, we evaluated and compared several pathway-analysis approaches, including individual SKAT and ARTP-ARTP, which require raw genotype data, and we tested the self-contained null hypothesis.

The Individual SKAT Method—SKAT is directly used to test the significance of pathways. All variants in one pathway are treated as one group, and the SNP-gene-pathway hierarchical information is ignored. Because the test statistic follows a mixture of a Chi-square distribution, a permutation procedure is not necessary when weights do not depend on phenotypes.

The ARTP-ARTP Method [Yu, et al. 2009]—The SNP-level test involves performing a single-marker association test (e.g., logistic regression) for each individual variant, followed by an ARTP to combine the SNP-level evidence of association in order to obtain gene-level p-values. ARTP is then used again to combine the gene-level p-values to obtain the pathway-level p-value.

Simulation Study

To evaluate the performance of the SKAT-ARTP, individual SKAT, and ARTP-ARTP methods, we conducted a simulation study using a dichotomous trait. Under the null hypothesis, the phenotype was generated randomly, while under the alternative hypothesis, the phenotype was generated based on genetic information from the data.

Type I Error Rate—We used the data of Type I diabetes from WTCCC with 2,000 samples. SNPs in the WTCCC data set were assigned to a gene if they were located within the flank of 5 kb of the gene on either side. Overlapping genes were excluded. In addition, we linked genes to pathways in KEGG (<http://www.genome.jp/kegg/>) based on HUGO (<http://www.genenames.org/>) gene symbols, and only non-overlapping pathways were included. At the end, 6 pathways, including 99 genes and 797 SNPs (MAF distribution is shown in Fig. 1S), were selected (hsa00010: 43 genes; hsa00140: 6 genes; hsa00791: 7 genes; hsa03030: 18 genes; hsa03050: 14 genes; hsa03060: 11 genes). For all pathways, we simulated 2,000 sets of phenotypes, and hence there were 2,000 simulated data sets. The dichotomous phenotypes were generated via the following model:

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

where $\alpha_0 = -2.94$ was determined to set the prevalence, which was set to 5% in our simulation. These data sets were used to evaluate the Type I error rate.

Power Evaluation—We used the same genotypes as generated under the null hypothesis. For all pathways, we simulated 1,000 sets of phenotypes. To see the effect of sample size, we conducted simulations for sample sizes of 1,000 and 2,000. The dichotomous phenotypes were simulated via the model:

$$\text{logit}P(y=1)=\alpha_0+0.5X_1+0.5X_2+\beta_1G_1+\beta_2G_2+\cdots+\beta_kG_k,$$

where X_1 is a continuous variable generated from a standard normal distribution; X_2 is a dichotomous variable from a Bernoulli distribution with a probability of 0.5; $G_1, G_2, \dots, G_k$ are the genotypes of causal SNPs; and $\beta_1, \beta_2, \dots, \beta_k$ are the log-odds ratios of the causal SNPs. α_0 was determined as described for the null gene sets. Furthermore, $\beta_1, \beta_2, \dots, \beta_k$ were set to be $c|\log_{10}MAF_j|$ in order to assign large weights to rare variants, and $c = 0.8$. Therefore, the effect of rare variants is larger than that of common variants, although both rare and common variants could be causal. The phenotypes for the simulation study were generated for all pathways together. We considered two scenarios. In the first, there were a total of six causal variants, and each one was randomly selected from one pathway. The MAFs of the six causal SNPs are shown in Table 2. This is the case in which only a few SNPs are associated with a phenotype in a pathway. In the second scenario, the number of randomly selected causal SNPs was proportional to the length of the pathway. Therefore, four causal SNPs are from hsa00010, one is from hsa01400, one is from hsa00791, two are from hsa03030, two are from hsa03050, and two are from hsa03060. This represents the scenario in which multiple SNPs contribute jointly to the phenotype. The MAFs of 12 total causal SNPs are shown in Table 3, and the causal variants can be either common or rare variants.

Results

Simulation Study of Type I Error Rate

We evaluated the following approaches for pathway analyses: SKAT-ARTP, individual SKAT, and ARTP-ARTP. To evaluate our proposed weights, for the SKAT-related methods we used alternative weights in addition to the default weights (e.g.,

$\sqrt{w_j} = \text{Beta}(\text{MAF}_j; 1, 25)$), that is, we used the inverse of the single marker p-value as the weight (denoted by W) and the sum of the beta density and the inverse of the single marker p-value as the weight (denoted by W2), as described in the Methods section. The Type I error rates for all six pathways at significance levels of 0.05 and 0.01 are summarized in Table 1. All of the approaches except for individual SKAT used a permutation procedure to obtain p-values, and we used 1,000 permutations for each p-value. The results show that all tests have their empirical Type I error rate close to the nominal level based on 2,000 simulation replicates for each of the six pathways. This indicates that all of these tests control the Type I error well.

Statistical Power Comparison

We first focused on the scenario with six causal SNPs from six pathways with a sample size of 2,000. The phenotype was simulated from these six causal SNPs, and they have different MAFs in order to evaluate the power for both rare and common variants. All of the power comparisons under each simulation design were based on 1,000 simulated data sets. The results for significance levels $\alpha = 0.05$ and 0.01 are presented in Table 2. The ROC curves of the average Type I error rate vs. the average power of the six causal SNPs scenario over all six pathways are shown in Fig. 1. In summary, not all of the methods are powerful enough to detect the causal variant in hsa00010 because this variant has the highest MAF, which leads to the smallest effect size in this simulation design. ARTP-ARTP suffers a major loss of power at $\alpha = 0.01$ compared with $\alpha = 0.05$, which indicates that it is less powerful when the significance level is more stringent. SKAT-ARTP and individual SKAT are both powerful enough to detect rare causal variants (e.g., causal variants in hsa00791, hsa03030, and hsa03060), and SKAT is slightly more powerful in the rare variant circumstance where SKAT is developed for. SKAT(W) only has weak power to detect the causal variant in hsa00010 and suffers a loss of power in detecting rare causal variants at $\alpha = 0.01$, although it remains powerful for detecting common causal variants. The performance of SKAT(W2) is consistently better than SKAT(W), except in testing hsa00010. Furthermore, SKAT-ARTP(W) performs better than SKAT(W), but also suffers a loss of power to detect rare causal variants at $\alpha = 0.01$. Overall, SKAT-ARTP(W2) performs consistently well across all pathways. SKAT-related methods with default weights ($\sqrt{w_j} = \text{Beta}(\text{MAF}_j; 1, 25)$) appear to be powerful for detecting rare causal variants and the use of new weights (W2) substantially helps to improve the power of detecting common causal variants. In addition, compared with the weights of W, using W2 is more powerful in detecting rare variants because the first term in its formula assigns large weights to rare variants. The ROC curves for a sample size of 1,000 are shown in Fig. 2S, which is similar to what we observed for the sample size of 2,000. In all of the settings that we considered, SKAT-ARTP(W2) had the best overall performance across various cases.

We also considered the 12 causal SNPs for the scenario with a sample size of 2,000, in which the number of causal SNPs in each pathway was made proportional to the length of the pathway so as to reduce the effect of pathway length on the final p-value. For the 12-causal-SNP scenario, multiple causal variants could be from one pathway. The phenotype was simulated from these 12 SNPs, and they have different MAFs in order to test the power for both rare and common variants. The results for significance levels $\alpha = 0.05$ and 0.01 are summarized in Table 3. The ROC curves of the average Type I error rate vs. the average power of the 12-causal-SNP scenario over all six pathways are shown in Fig. 2. ARTP-ARTP is consistently powerful for identifying causal variants across all pathways at $\alpha = 0.05$, but it suffers a loss of power for hsa00140 and hsa00791 at $\alpha = 0.01$. SKAT-ARTP is powerful except for hsa03060, which contains two common causal variants. SKAT(W) and SKAT(W2) work well for all pathways at $\alpha = 0.05$, but they suffer a substantial power loss for hsa00010 and hsa00791 at $\alpha = 0.01$, probably due to gene-level information being ignored. Not surprisingly, SKAT maintains good power for rare-variant detection but has almost no power for common-variant detection. SKAT-ARTP(W) and SKAT-ARTP(W2) both maintain good power when significance levels change and retain consistent power across all pathways, with SKAT-ARTP(W2) performing slightly better. The ROC curves for the sample size of 1,000 are shown in Fig. 3S, which is again similar to what we observed for the sample size of 2,000. In both scenarios, SKAT-ARTP(W2) was the most powerful method. We also conducted an additional set of simulations, with a sample size of 2,000, in which causal SNPs were from three pathways instead of all six pathways in order to evaluate Type I error rate and power together. The results, as shown in Table 1S and Fig. 4S, are similar to those from the simulations with Type I error rate and power being evaluated separately.

Application to the Wellcome Trust Case Control Consortium Bipolar-Disorder Data

To further validate our proposed method, we applied the methods we compared above to the Wellcome Trust Case Control Consortium (WTCCC) bipolar-disorder (BD) study shown in Table 4. There are 1,998 cases and 1,504 controls. We analyzed three pathways: cation-channel activity, gated-channel activity, and metal-ion transmembrane-transporter activity, which have been reported to be significantly associated with bipolar disorder [Chuang, et al. 2013]. Similar to the simulation section, we assigned SNPs to a gene if they were located within the flank of 5 kb of the gene on either side. Overlapping genes were deleted so that the SNPs would not be used more than once in the test. Furthermore, we linked genes to pathways in gene ontology (GO) terms downloaded from the Molecule Signature Database (MsigDB) based on HUGO gene symbols. Using 10,000 permutations, all three pathways reached the nominal-significance level with the same raw p-value of 9.99×10^{-5} calculated from SKAT(W), SKAT(W2), ARTP-ARTP, SKAT-ARTP(W), and SKAT-ARTP(W2). The raw p-value was calculated from permutation, so the lowest p-value that could be reached was 1 divided by 10,001 based on 10,000 permutations, which is equal to $1/10001 = 9.99 \times 10^{-5}$. Because this is the extreme scenario where the adaptive rank-truncated-product statistic of the observed data is smaller than that from any of the permuted data, it may be more appropriate to report the p-value as 9.99×10^{-5} . Hence, all three pathways are highly significantly associated with bipolar disorder from these five methods. On the other hand, the p-values from SKAT and SKAT-ARTP are not as low as the rest of

the methods (some of them are not even significant). This may be because the causal SNPs in the pathways are common SNPs, but SKAT and SKAT-ARTP are powerful for testing rare variants.

Computation Time

The computation time was obtained on a single computing node with 3 GHz CPU and 4 GB memory based on the analysis of the above three candidate pathways related to bipolar disorder that are shown in Fig. 3. For permutation-involved methods, we considered 1,000 and 10,000 permutations, and the number of permutations was the major factor determining the total computation time. In addition, calculating complicated weights could also be time consuming. SKAT, which does not involve permutation, is the fastest method. For permutation-based methods, SKAT-ARTP, which uses the weight as a function of the MAF of each SNP, is the fastest one. Moreover, it takes a longer time to calculate the weights (W) and (W_2), which involves calculating p-values from a single marker association test. That is why SKAT(W), SKAT(W_2), SKAT-ARTP(W), and SKAT-ARTP(W_2) take more time. In addition, SKATARTP(W , W_2) use less time than SKAT(W , W_2) in this case because SKAT-ARTP(W , W_2) consider gene-level information that can save computation time when multiple pathways are tested and have genes in common. These three pathways related to bipolar disorder share many common genes. There are two steps in SKAT-ARTP(W , W_2). The first step is to calculate gene-level statistics, which is time consuming. The second step is to calculate pathway-level p-values, which can be done quickly. In step one, genes in common are calculated only once, and in step two the statistics that have already been calculated for the genes are used in each overlapping pathway. On the other hand, if gene-level information is ignored (i.e., SKAT(W , W_2)), the statistics for the overlapping gene are calculated in every pathway that contains this gene.

Discussion

GWASs have been widely used in genetic studies in the past few years due to the availability of high-throughput genotyping technology. In GWASs, the associations between hundreds of thousands or even millions of SNPs and diseases/traits are usually evaluated, mostly one at a time. Although GWASs have successfully identified many genetic variants related to certain diseases, collectively they explain only a small proportion of the estimated heritability, leading to the so-called “missing heritability” [Eichler, et al. 2010; Manolio, et al. 2009]. The leading explanations for this missing heritability all recognize the collective effect of multiple genetic variants and emphasize the importance of analyzing them jointly [Manolio, et al. 2009; Speed, et al. 2012]. This may be because the risk of complex diseases is influenced by multiple genetic variants [Bansal, et al. 2010; Bodmer and Bonilla 2008; Pritchard and Cox 2002; Raychaudhuri 2011], and so it may take the malfunction of several variants to cause a disorder [Kimura 1985; Shi and Weinberg 2011]. In addition, genetic heterogeneity adds to the complexity, making it meaningful and necessary to analyze a group of variants jointly. Pathway-based analysis is one approach to analyzing a group of genetic variants in a hierarchical way and could complement the single-marker association test. In this study, we proposed an approach for pathway analysis that takes advantage of SKAT and ARTP, coupled with a new weighting scheme. Because of the difficulty of

deriving the exact distribution of test statistics, a permutation procedure was employed for p-value calculation.

The importance of rare variants has been increasingly recognized in identifying disease susceptibility. Recently, many statistical methods for rare-variants association detection have been developed [Chen, et al. 2011; Han and Pan 2010; Hoffmann, et al. 2010; Li and Leal 2008; Lin, et al. 2014; Lin, et al. 2013; Lin, et al. 2012; Lin, et al. 2011; Madsen and Browning 2009; Morgenthaler and Thilly 2007; Price, et al. 2010; Wu, et al. 2010; Zawistowski, et al. 2010]. Although such methods do not exclude common variants from their evaluation, they use a weighting scheme that up-weights the effects of rare variants and down-weights the effects of common variants, and they therefore mostly test for the effects of rare variants. However, recent studies have shown that complex diseases are usually caused by the joint effects of multiple genetic variants, likely both common and rare. These methods may lose power when common variants in a region under investigation are also associated with the disease. In reality, we do not know whether a disease is caused by common or rare variants, and so it is reasonable to consider both in the analysis. Currently, rare and common variants are usually tested separately using different strategies (e.g., a group test for rare variants and a single-marker test for common variants). It is desirable to have a unified method of testing the combined effect of rare and common variants. Efforts have been made toward this goal [Ionita-Laza, et al. 2013]. However, frequency cutoff is required for defining rare variants. In this report, the weight we proposed enables effective detection of both common and rare variants without the need of the frequency threshold. In genetic studies, most of the variants are not associated with the disease, and these neutral variants may add “noise” and dilute the association signal. One way to reduce such noise is to use truncation [Lin, et al. 2014; Zaykin, et al. 2002], which is equivalent to having a weighting scheme of 0 or 1. Our weighting scheme can be seen as a generalization of the truncation method. We have shown in our simulation studies that this weight can be incorporated in SKAT (SKAT(W) and SKAT (W2)) to increase the power for set-based analysis. We conducted extensive simulation studies to evaluate our proposed approach and applied it to data from a bipolar-disorder study to show the practical utility.

One common question for pathway analysis (in general, groupwise tests) is that if a pathway is found to be significantly associated with a disease, how can it be determined which genes and SNPs are related to the disease? There are two scenarios in which a pathway is associated with a disease. In the first, a small number of genes (or even SNPs) in the pathway are associated with the disease, and the association signal is mainly driven by these few genes. In the second scenario, multiple genes are associated with the disease, but each of them has only a weak effect, i.e., the association signal is driven by the joint effects of those genes. If a pathway is significantly associated with a disease, further study at the gene level should be conducted to find out whether some specific genes are associated with the disease, or if the association is due to the joint effect. Analogously, after a significant gene is identified, further SNP-level analysis should be conducted to identify whether some specific, significant SNPs are associated with the disease or if the significant effect of the gene is due to the joint effects of the SNPs. Pathway analysis helps to narrow down the location of causal variants and is one approach in the long process of deciphering the genetic mechanism of diseases.

Our proposed method has several advantages over currently available pathway-analysis approaches. First, our approach considers the SNP-gene-pathway three-level hierarchical structure. Thus, it makes the statistical method more biologically meaningful and interpretable and may help people to more clearly understand the biological mechanisms underlying diseases. In our approach, the gene-level p-values are evaluated by SKAT, which considers the joint effects of the SNPs in one gene. In contrast, some methods summarize gene-level p-values by selecting the most significant SNP to represent the significance of the gene, such as the gene-set-enrichment analysis (GSEA) algorithm [Wang, et al. 2007]. These methods assume that the associations between phenotypes and genetic variants are driven by the most significant SNP, but such methods may lose power greatly when multiple SNPs, especially those with moderate or small effects, contribute to a disease. Second, we have shown that our methodology improves the power compared with some existing pathway-based approaches while maintaining Type I error rate. The power gain comes from the new weight and ARTP, which eliminates the noise from null genes. In addition, the application of our weights helps to increase the overall power for detecting both rare and common genetic variants. Third, our proposed model is flexible, like SKAT. It has the capacity to test the associations between continuous or dichotomous phenotypes and common or rare variants, although we only evaluated the performance of dichotomous phenotypes in this paper. In our model, it is easy to incorporate environmental and clinical factors as covariates in single-marker tests (e.g., logistic regression for dichotomous traits and general linear regression for continuous traits) when calculating weights and performing SKAT to obtain gene-level p-values. Furthermore, the method can be readily applied to GWAS data, exome-sequencing data, and deep re-sequencing data. Although we only studied the performance of a linear kernel, it is straightforward to use a non-linear kernel within the flexible kernel machine-regression framework in SKAT when assuming a non-linear association between a disease and genetic variants.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by grant 1158862 from the National Science Foundation (Q.Y., H.K.T.); grants GM081488, R01HL092173, P60AR048095 (N.L.), 5R01GM069430-08 (N.Y.), 5R01DA025095 (X.L.), and GM073766 (G.G.) from the National Institutes of Health; grants NSC 102-2628-B-002-039-MY3 and NSC 102-2314-B-002-001-MY2 from the National Science Council of Taiwan; and grants NTU-CESRP-101R7622-8, NTU-CESRP-102R7622-8, NTU-CESRP-103R7622-8, and NTU-CDP-102R7769 from National Taiwan University (W-Y.L.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsors.

References

- Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet.* 2010; 11(11):773–785. [PubMed: 20940738]
- Becker T, Knapp M. A powerful strategy to account for multiple testing in the context of haplotype analysis. *Am J Hum Genet.* 2004; 75(4):561–570. [PubMed: 15290652]
- Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet.* 2008; 40(6):695–701. [PubMed: 18509313]

- Cai T, Lin X, Carroll RJ. Identifying genetic marker sets associated with phenotypes via an efficient adaptive score test. *Biostatistics*. 2012; 13(4):776–790. [PubMed: 22734045]
- Chasman DI. On the utility of gene set methods in genomewide association studies of quantitative traits. *Genet Epidemiol*. 2008; 32(7):658–668. [PubMed: 18481796]
- Chen CT, McDavid AN, Kahsai OJ, Zebari AS, Carlson CS. Efficient identification of rare variants in large populations: deep re-sequencing the CRP locus in the CARDIA study. *Nucleic Acids Res*. 2013; 41(7):e85. [PubMed: 23408856]
- Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, Peters U, Hsu L. Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *Am J Hum Genet*. 2010; 86(6):860–871. [PubMed: 20560206]
- Chen W, Gao X, Wang J, Sun C, Wan W, Zhi D, Liu N, Chen X, Gao G. Evaluation of association tests for rare variants using simulated data sets in the Genetic Analysis Workshop 17 data. *BMC Proc*. 2011; 5(Suppl 9):S86. [PubMed: 22373475]
- Chuang LC, Kao CF, Shih WL, Kuo PH. Pathway analysis using information from allele-specific gene methylation in genome-wide association studies for bipolar disorder. *PLoS One*. 2013; 8(1):e53092. [PubMed: 23326387]
- Davies R. The distribution of a linear combination of chi-square random variables. *J. R. Stat. Soc. Ser. C Appl. Stat*. 1980; 29:323–333.
- Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet*. 2010; 11(6):446–450. [PubMed: 20479774]
- Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered*. 2010; 70(1):42–54. [PubMed: 20413981]
- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*. 2009; 106(23):9362–9367. [PubMed: 19474294]
- Hoffmann TJ, Marini NJ, Witte JS. Comprehensive approach to analyzing rare genetic variants. *PLoS One*. 2010; 5(11):e13584. [PubMed: 21072163]
- Holden M, Deng S, Wojnowski L, Kulle B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics*. 2008; 24(23):2784–2785. [PubMed: 18854360]
- Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, Wellcome Trust Case-Control C, Owen MJ, O'Donovan MC, Craddock N. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am J Hum Genet*. 2009; 85(1):13–24. [PubMed: 19539887]
- Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet*. 2007; 39(7):870–874. [PubMed: 17529973]
- Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. *Am J Hum Genet*. 2013; 92(6):841–853. [PubMed: 23684009]
- Johnson AD, Hwang SJ, Voorman A, Morrison A, Peloso GM, Hsu YH, Thanassoulis G, Newton-Cheh C, Rogers IS, Hoffmann U, et al. Resequencing and clinical associations of the 9p21.3 region: a comprehensive investigation in the Framingham heart study. *Circulation*. 2013; 127(7):799–810. [PubMed: 23315372]
- Kim DS, Burt AA, Crosslin DR, Robertson PD, Ranchalis JE, Boyko EJ, Nickerson DA, Furlong CE, Jarvik GP. Novel common and rare genetic determinants of paraoxonase activity: FTO, SERPINA12, and ITGAL. *J Lipid Res*. 2013; 54(2):552–560. [PubMed: 23160181]
- Kimura, M. *The Neutral Theory of Molecular Evolution*. Cambridge University Press; 1985.
- Lee JH, Cheng R, Honig LS, Feitosa M, Kammerer CM, Kang MS, Schupf N, Lin SJ, Sanders JL, Bae H, et al. Genome wide association and linkage analyses identified three loci-4q25, 17q23.2, and 10q11.21-associated with variation in leukocyte telomere length: the Long Life Family Study. *Front Genet*. 2013; 4:310. [PubMed: 24478790]

- Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet.* 2008; 83(3):311–321. [PubMed: 18691683]
- Lin WY, Lou XY, Gao G, Liu N. Rare Variant Association Testing by Adaptive Combination of P-values. *PLoS One.* 2014; 9(1):e85728. [PubMed: 24454922]
- Lin WY, Yi N, Lou XY, Zhi D, Zhang K, Gao G, Tiwari HK, Liu N. Haplotype kernel association test as a powerful method to identify chromosomal regions harboring uncommon causal variants. *Genet Epidemiol.* 2013; 37(6):560–570. [PubMed: 23740760]
- Lin WY, Yi N, Zhi D, Zhang K, Gao G, Tiwari HK, Liu N. Haplotype-based methods for detecting uncommon causal variants with common SNPs. *Genet Epidemiol.* 2012; 36(6):572–582. [PubMed: 22706849]
- Lin WY, Zhang B, Yi N, Gao G, Liu N. Evaluation of pooled association tests for rare variant identification. *BMC Proc.* 2011; 5(Suppl 9):S118. [PubMed: 22373333]
- Liu L, Sabo A, Neale BM, Nagaswamy U, Stevens C, Lim E, Bodea CA, Muzny D, Reid JG, Banks E, et al. Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. *PLoS Genet.* 2013; 9(4):e1003443. [PubMed: 23593035]
- Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 2009; 5(2):e1000384. [PubMed: 19214210]
- Magi R, Asimit JL, Day-Williams AG, Zeggini E, Morris AP. Genome-Wide Association Analysis of Imputed Rare Variants: Application to Seven Common Complex Diseases. *Genet Epidemiol.* 2012
- Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest.* 2008; 118(5):1590–1605. [PubMed: 18451988]
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. *Nature.* 2009; 461(7265):747–753. [PubMed: 19812666]
- McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet.* 2008; 9(5):356–369. [PubMed: 18398418]
- Medina I, Montaner D, Bonifaci N, Pujana MA, Carbonell J, Tarraga J, Al-Shahrour F, Dopazo J. Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. *Nucleic Acids Res.* 2009; 37:W340–W344. (Web Server issue). [PubMed: 19502494]
- Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res.* 2007; 615(1–2):28–56. [PubMed: 17101154]
- Nam D, Kim J, Kim SY, Kim S. GSA-SNP: a general approach for gene set analysis of polymorphisms. *Nucleic Acids Res.* 2010; 38:W749–W754. (Web Server issue). [PubMed: 20501604]
- Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature.* 2009; 461(7261):272–276. [PubMed: 19684571]
- O'Dushlaine C, Kenny E, Heron EA, Segurado R, Gill M, Morris DW, Corvin A. The SNP ratio test: pathway analysis of genome-wide association datasets. *Bioinformatics.* 2009; 25(20):2762–2763. [PubMed: 19620097]
- Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet.* 2010; 86(6):832–838. [PubMed: 20471002]
- Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet.* 2001; 69(1):124–137. [PubMed: 11404818]
- Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet.* 2002; 11(20):2417–2423. [PubMed: 12351577]
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007; 81(3):559–575. [PubMed: 17701901]

- Raychaudhuri S. Mapping rare and common causal alleles for complex human diseases. *Cell*. 2011; 147(1):57–69. [PubMed: 21962507]
- Shi M, Weinberg CR. How much are we missing in SNP-by-SNP analyses of genome-wide association studies? *Epidemiology*. 2011; 22(6):845–847. [PubMed: 21878815]
- Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet*. 2012; 91(6):1011–1021. [PubMed: 23217325]
- Stratton MR, Rahman N. The emerging landscape of breast cancer susceptibility. *Nat Genet*. 2008; 40(1):17–22. [PubMed: 18163131]
- Teer JK, Mullikin JC. Exome sequencing: the sweet spot before whole genomes. *Hum Mol Genet*. 2010; 19(R2):R145–R151. [PubMed: 20705737]
- Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*. 2008; 92(5):265–272. [PubMed: 18722519]
- Walsh T, King MC. Ten genes for inherited breast cancer. *Cancer Cell*. 2007; 11(2):103–105. [PubMed: 17292821]
- Wang HZ, Qin HD, Guo W, Samuels J, Shugart YY. New insights into the genetic mechanism of IQ in autism spectrum disorders. *Front Genet*. 2013; 4:195. [PubMed: 24151499]
- Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet*. 2007; 81(6):1278–1283. [PubMed: 17966091]
- Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet*. 2010; 11(12):843–854. [PubMed: 21085203]
- Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007; 447(7145):661–678. [PubMed: 17554300]
- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet*. 2010; 86(6):929–942. [PubMed: 20560208]
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011; 89(1):82–93. [PubMed: 21737059]
- Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet*. 2007; 39(5):645–649. [PubMed: 17401363]
- Yi N, Liu N, Zhi D, Li J. Hierarchical generalized linear models for multiple groups of rare and common variants: jointly estimating group and individual-variant effects. *PLoS Genet*. 2011; 7(12):e1002382. [PubMed: 22144906]
- Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, Caporaso N, Kraft P, Chatterjee N. Pathway analysis by adaptive combination of P-values. *Genet Epidemiol*. 2009; 33(8):700–709. [PubMed: 19333968]
- Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zollner S. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am J Hum Genet*. 2010; 87(5):604–617. [PubMed: 21070896]
- Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. Truncated product method for combining P-values. *Genet Epidemiol*. 2002; 22(2):170–185. [PubMed: 11788962]
- Zhang K, Cui S, Chang S, Zhang L, Wang J. i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. *Nucleic Acids Res*. 2010; 38:W90–W95. (Web Server issue). [PubMed: 20435672]

Six Causal SNPs Scenario ROC Curve

Fig. 1. ROC curves (average Type I error rate vs. average power for six pathways) for six causal SNPs scenario for a sample size of 2,000.

Twelve Causal SNPs Scenario ROC Curve

Fig. 2. ROC curves (average Type I error rate vs. average power for six pathways) for 12 causal SNPs scenario for a sample size of 2,000.

Fig. 3.
Computation time of three candidate pathways related to bipolar disorder for 1,000 permutations and 10,000 permutations.

Table 1

Type I Error Rate for a Sample Size of 2,000

$\alpha=0.05$									
Pathway	N genes	SKAT	SKAT (W)	SKAT (W/2)	ARTP-ARTP	SKAT-ARTP (W)	SKAT-ARTP (W/2)	SKAT-ARTP (W)	SKAT-ARTP (W/2)
hsa00010	43	0.055	0.052	0.043	0.054	0.056	0.052	0.052	0.047
hsa00140	6	0.055	0.047	0.042	0.044	0.055	0.044	0.044	0.045
hsa00791	7	0.045	0.048	0.052	0.046	0.058	0.053	0.053	0.058
hsa03030	18	0.058	0.056	0.052	0.048	0.052	0.053	0.053	0.057
hsa03050	14	0.053	0.048	0.046	0.043	0.041	0.046	0.046	0.047
hsa03060	11	0.050	0.051	0.046	0.056	0.055	0.049	0.049	0.052
Average		0.052	0.050	0.047	0.048	0.053	0.050	0.050	0.051
$\alpha=0.01$									
Pathway	N genes	SKAT	SKAT (W)	SKAT (W/2)	ARTP-ARTP	SKAT-ARTP (W)	SKAT-ARTP (W/2)	SKAT-ARTP (W)	SKAT-ARTP (W/2)
hsa00010	43	0.015	0.007	0.008	0.012	0.012	0.013	0.013	0.008
hsa00140	6	0.013	0.013	0.008	0.011	0.008	0.007	0.007	0.011
hsa00791	7	0.010	0.011	0.011	0.014	0.011	0.010	0.010	0.008
hsa03030	18	0.016	0.012	0.012	0.014	0.014	0.011	0.011	0.013
hsa03050	14	0.014	0.010	0.011	0.008	0.010	0.010	0.010	0.010
hsa03060	11	0.008	0.008	0.010	0.009	0.009	0.015	0.015	0.010
Average		0.013	0.010	0.010	0.011	0.011	0.011	0.011	0.010

Table 2

Power for Six Causal SNPs Scenario for a Sample Size of 2,000

$\alpha=0.05$										
Pathway	MAF	N genes	SKAT (W)	SKAT (W2)	ARTP-ARTP	SKAT-ARTP (W)	SKAT-ARTP (W2)	SKAT-ARTP (W)	SKAT-ARTP (W2)	SKAT-ARTP (W2)
hsa00010	0.39	43	0.115	0.118	0.095	0.224	0.151	0.389	0.414	0.414
hsa00140	0.16	6	0.053	0.987	0.992	0.983	0.998	0.978	0.979	0.979
hsa00791	0.08	7	1	0.998	1	1	0.999	0.891	0.999	0.999
hsa03030	0.04	18	0.997	0.816	0.892	0.88	0.957	0.98	0.918	0.918
hsa03050	0.11	14	0.636	0.941	0.967	0.919	0.511	0.995	0.974	0.974
hsa03060	0.03	11	0.993	0.739	0.89	0.952	0.955	0.804	0.924	0.924
Average			0.632	0.766	0.806	0.826	0.762	0.84	0.868	0.868
$\alpha=0.01$										
Pathway	MAF	N genes	SKAT (W)	SKAT (W2)	ARTP-ARTP	SKAT-ARTP (W)	SKAT-ARTP (W2)	SKAT-ARTP (W)	SKAT-ARTP (W2)	SKAT-ARTP (W2)
hsa00010	0.39	43	0.026	0.036	0.027	0.055	0.04	0.189	0.201	0.201
hsa00140	0.16	6	0.006	0.865	0.889	0.652	0.991	0.793	0.804	0.804
hsa00791	0.08	7	1	0.981	0.995	0.653	0.998	0.962	0.974	0.974
hsa03030	0.04	18	0.987	0.64	0.748	0.293	0.883	0.75	0.823	0.823
hsa03050	0.11	14	0.405	0.838	0.892	0.314	0.255	0.923	0.927	0.927
hsa03060	0.03	11	0.976	0.529	0.743	0.378	0.89	0.564	0.781	0.781
Average			0.557	0.648	0.716	0.391	0.676	0.697	0.752	0.752

Table 3

Power for 12 Causal SNPs Scenario for a Sample Size of 2,000

$\alpha=0.05$										
Pathway	MAF	N genes	SKAT (W)	SKAT (W2)	ARTP- ARTP	SKAT- ARTP (W)	SKAT- ARTP (W2)			
hsa00010	0.21,0.39, 0.04,0.15	43	0.69	0.77	0.789	1	0.985	1	1	1
hsa00140	0.16	6	0.031	1	1	0.997	1	0.996	0.999	0.999
hsa00791	0.05	7	0.991	0.748	0.889	0.997	0.998	0.995	0.999	0.999
hsa03030	0.04,0.07	18	1	0.999	1	1	1	0.999	1	1
hsa03050	0.24,0.01	14	0.999	0.997	1	1	0.997	0.998	1	1
hsa03060	0.16,0.27	11	0.03	0.998	0.997	0.999	0.503	1	0.999	0.999
Average			0.624	0.919	0.946	0.999	0.914	0.998	1	1
$\alpha=0.01$										
Pathway	MAF	N genes	SKAT (W)	SKAT (W2)	ARTP- ARTP	SKAT- ARTP (W)	SKAT- ARTP (W2)			
hsa00010	0.21,0.39, 0.04,0.15	43	0.367	0.38	0.384	1	0.959	1	1	1
hsa00140	0.16	6	0.005	0.972	0.97	0.756	1	0.912	0.92	0.92
hsa00791	0.05	7	0.957	0.493	0.667	0.645	0.984	0.938	0.991	0.991
hsa03030	0.04,0.07	18	1	0.992	1	1	1	0.998	1	1
hsa03050	0.24,0.01	14	0.996	0.986	0.993	0.997	0.991	0.995	1	1
hsa03060	0.16,0.27	11	0.006	0.987	0.989	0.998	0.261	1	0.999	0.999
Average			0.555	0.802	0.834	0.899	0.866	0.974	0.985	0.985

Table 4

Results for Three Candidate Pathways Related to Bipolar Disorder

Pathway name	N genes	Raw p-value					
		SKAT	SKAT (W)	SKAT (W2)	ARTP- ARTP	SKAT- ARTP (W)	SKAT- ARTP (W2)
Cation-channel activity	91	0.0630	9.99×10^{-5}	9.99×10^{-5}	9.99×10^{-5}	9.99×10^{-5}	9.99×10^{-5}
Gated-channel activity	87	0.0348	9.99×10^{-5}	9.99×10^{-5}	9.99×10^{-5}	9.99×10^{-5}	9.99×10^{-5}
Metal ion-transmembrane-transporter activity	110	0.0062	9.99×10^{-5}	9.99×10^{-5}	9.99×10^{-5}	9.99×10^{-5}	9.99×10^{-5}