Genetics and population analysis

Adaptive gene- and pathway-trait association testing with GWAS summary statistics

Il-Youp Kwak and Wei Pan*

Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, USA

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on 6 August 2015; revised on 24 November 2015; accepted on 29 November 2015

Abstract

Background: Gene- and pathway-based analyses offer a useful alternative and complement to the usual single SNP-based analysis for GWAS. On the other hand, most existing gene- and pathway-based tests are not highly adaptive, and/or require the availability of individual-level genotype and phenotype data. It would be desirable to have highly adaptive tests applicable to summary statistics for single SNPs. This has become increasingly important given the popularity of large-scale meta-analyses of multiple GWASs and the practical availability of either single GWAS or meta-analyzed GWAS summary statistics for single SNPs.

Results: We extend two adaptive tests for gene- and pathway-level association with a univariate trait to the case with GWAS summary statistics without individual-level genotype and phenotype data. We use the WTCCC GWAS data to evaluate and compare the proposed methods and several existing methods. We further illustrate their applications to a meta-analyzed dataset to identify genes and pathways associated with blood pressure, demonstrating the potential usefulness of the proposed methods. The methods are implemented in R package aSPU, freely and publicly available.

Availability and implementation: https://cran.r-project.org/web/packages/aSPU/

Contact: weip@biostat.umn.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

In spite of the tremendous success of genome-wide association studies (GWASs), due to modest to small effect sizes of the majority of causal SNPs for complex and common disease, large sample sizes and more powerful statistical tests are always needed. Furthermore, even when some associated SNPs or loci are identified by the most popular single SNP-based analysis, it is often extremely difficult to offer a functional interpretation that can shed light on the underlying biology. As alternatives, gene- and pathway-based analyses have been proposed and applied for single traits, demonstrating their useful and complementary roles (Fan et al., 2015; Pan, 2009; Schaid et al., 2012; Wang et al., 2007, 2010; Wu et al., 2010). A statistical challenge is that, due to unknown true association patterns, there is no uniformly most powerful test to detect multiple SNP-single trait associations; an association test may perform well for one dataset, but not necessarily for another. For example, the presence of non-associated SNPs will largely diminish the power of a standard test if no effective SNP selection or weighting is adopted (Petersen et al., 2013). Pan et al. (2014) proposed a data-adaptive (aSPU) approach based on estimating and selecting the most powerful test among a class of so-called sum of powered score (SPU) tests, which cover several popular tests as special cases, such as the burden test, a variance component test and a univariate test. The main idea is to use various values of a parameter to construct data-driven and varying weights for the SNPs, thus adaptive to unknown signal sparsity and association directions among the multiple SNPs being tested. Furthermore, Pan et al. (2015) extended the methodology to pathway analysis (aSPUpath). In particular, two parameters are introduced such that the test is adaptive at both the SNP and gene levels. As demonstrated therein, due to their high data-adaptivity, the two adaptive tests remained powerful across a wide range of scenarios.

1178

Bioinformatics, 32(8), 2016, 1178–1184
doi: 10.1093/bioinformatics/btv719
Advance Access Publication Date: 10 December 2015
Original Paper
However, the two adaptive tests for gene- and pathway-single trait associations are only applicable to the case with individual-level genotype and phenotype data. Due to various reasons, it is often difficult for many researchers to obtain access to individual-level data. At the same time, it is easier to obtain access to some summary statistics, e.g. P-values, of individual SNPs in a GWAS. In addition, use of some summary statistics is often necessary for practical meta analyses, which have become increasingly popular and important for complex disease and traits (de Bakker et al., 2008). Here we propose extending the two highly adaptive tests to the case with only summary statistics for individual SNPs, demonstrating their applications to a meta-analyzed GWAS dataset for blood pressure (Ehret et al., 2011).

In numerical examples we compared our methods with gene-based GATES (Li et al., 2011) and three pathway-based approaches, GATES-Simes (Gui et al., 2011), HYST (Li et al., 2012) and MAGMA (de Leeuw et al., 2015). All methods are for a single (univariate) trait. GATES adopts an extended Simes procedure to control multiple testing while calculating the P-value quickly based on SNP summary statistics. GATES-Simes extends the main idea of GATES to extract the most significant gene-level P-value for a pathway, whereas HYST uses Fisher’s method to combine multiple genes’ P-values. MAGMA is based on multiple regression. We also considered one gene set enrichment analysis (GSEA) method, i-GSEAGWAS (Zhang et al., 2013), an extension of GSEA (Subramanian et al., 2005) to GWAS summary statistics. Note that here our goal is global association testing in the category of ‘self-contained tests’ with a null hypothesis \( H_0 \) that none of any genes in a pathway is associated with disease, in contrast to enrichment or competitive testing with a null hypothesis \( H_{0,E} \) that the proportion of associated genes in a pathway is no more than that in the rest of the genes; since rejecting \( H_{0,E} \) implies rejecting \( H_0 \), a global testing is often more powerful than an enrichment analysis (Goeman et al., 2007).

The proposed methods are available in R package aSPU, which is downloadable from CRAN.

## 2 Methods

We first assume that individual genotype and phenotype data are available, facilitating the derivation of our proposed methods later. Suppose for individual \( i = 1, \ldots, n \), we have a quantitative or binary trait \( Y_i \) and a vector of the genotype scores for \( k \) SNPs \( X_i = (X_{i1}, \ldots, X_{ik}) \), possibly drawn from either a single gene or a pathway with multiple genes. As usual we use the additive coding for each SNP: \( X_{ij} = 0, 1 \) or 2 is the count of an allele at SNP \( j \) for subject \( i \). We also assume a vector of covariates \( W_i \). We adopt a generalized linear model (GLM):

\[
g(E(Y_i)) = \beta_0 + \sum_{j=1}^{k} X_{ij} \beta_j + z' W_i,
\]

where \( g() \) is the link function (i.e. the identity function for a quantitative trait, or a logit function for a binary trait). We would like to test the null hypothesis \( H_0 : \beta = (\beta_1, \ldots, \beta_k)' = 0 \); that is, there is no association between any SNP and the trait under \( H_0 \). The score vector \( U = (U_1, \ldots, U_k)' \) for \( \beta \) is

\[
U = \sum_{i=1}^{n} (Y_i - \mu_{0,i})X_i,
\]

where \( \mu_{0,i} = E(Y_i|H_0) = g^{-1}(\beta_0 + z' W_i) \) is the estimated mean of \( Y_i \) in the null model (under \( H_0 \)). The covariance matrix of \( U \) can be estimated as

\[
\text{Cov}(U|H_0) = \sum_{i=1}^{n} (Y_i - \mu_{0,i}) (X_i - \bar{X})(X_i - \bar{X})'
\]

with its mean under \( H_0 \) as

\[
E[\text{Cov}(U|H_0)] = \sum_{i=1}^{n} \sigma_i^2 \Sigma_X,
\]

where \( \sigma_i^2 = \text{Var}(Y_i|H_0) \) and \( \Sigma_X = \text{Cov}(X) \).

### 2.1 A gene-based adaptive test with summary statistics

Now, suppose that we do not have individual-level data \( (Y_i, X_i)' \)'s but only single SNP-based summary statistics, say Z-scores \( Z = (Z_1, \ldots, Z_k)' \) with \( Z_i = \beta_j / \text{se}(\beta_j) \approx U_j / \text{se}(U_j) \) for \( j = 1, \ldots, k \); the approximation is based on the asymptotic equivalence between the Wald test and the score test. A key observation is that

\[
\text{Corr}(Z_i, Z_j) = \text{Corr}(Z_i, Z_k) \approx \text{Corr}(U_j, U_k) \approx \text{Corr}(X_{ij}, X_{ik}),
\]

where the last approximation is based on Eq. (1). Note that \( \text{Corr}(X) \) can be easily estimated from some reference panel, e.g. the CEU sample in the Hapmap data or 1000 Genome data, for a similar target population. Accordingly, we obtain an estimate of \( \text{Corr}(Z_i, Z_k) \), say \( r \).

 Mimicking the SPU and aSPU tests for individual genotype and phenotype data, we can now define the corresponding tests with only summary statistics \( Z \):

\[
\text{SPUs}(\gamma) = \text{SPUs}(\gamma; Z) = \sum_{i=1}^{k} Z_i^\gamma,
\]

\[
\text{aSPUs}(Z) = \min_{\gamma \in \Gamma} P_{\text{SPUs}(\gamma; Z)},
\]

where \( P_{\text{SPUs}(\gamma; Z)} \) is the \( P \)-value of the \( \text{SPUs}(\gamma; Z) \) test, and \( \Gamma = \{1, 2, \ldots, 8, \infty\} \) is often used. Note that since \( \text{SPUs}(\gamma; Z) \propto ||Z||_{\gamma} \rightarrow \max|Z| \) as an even integer \( \gamma \rightarrow \infty \), we define \( \text{SPUs}(\infty; Z) = \max|Z| \).

An optimal choice of \( \gamma \) depends on the unknown SNP-trait association patterns. For example, if most SNPs are (almost) equally associated with the trait in the same direction, then the burden test \( \text{SPUs}(1) \) will be (nearly) most powerful; if only one or few SNPs are associated with the trait, then using a larger \( \gamma \), e.g. \( \text{SPUs}(8) \) or \( \text{SPUs}(\infty) \), is expected to be more powerful; on the other hand, if the truth is between the above two extremes, then using an intermediate \( 1 < \gamma < 8 \), e.g. \( \text{SPUs}(2) \) that is similar to a variance-component test like KMR or SKAT (Liu et al., 2007; Pan, 2011; Wu et al., 2010), may have higher power. Note that \( \text{SPUs}(\infty) \) is exactly the same as the univariate minimum \( P \)-value method most often used in GWAS. Since we do not know the optimal value of \( \gamma \), we try multiple ones in \( \Gamma \) and then estimate and select the best one by aSPUs. Note that, in our experience, often \( \text{SPUs}(8) \) and \( \text{SPUs}(\infty) \) give almost identical results, hence we do not try any \( \gamma \) values between 8 and \( \infty \) in \( \Gamma \).

By the asymptotic distribution of \( U \), we know that for a large \( n \), under \( H_0 \), \( Z \) follows a multivariate normal distribution \( N(0, R) \), where \( R \) is the correlation matrix of \( X_i \) that can be estimated from some reference panel as discussed. Based on this known null distribution of \( Z \), we use Monte Carlo simulations to obtain the \( P \)-values of the \( \text{SPUs} \) and \( \text{aSPUs} \) tests. Specifically, (i) we generate
independent \( Z(b) \sim N(0, R) \) for \( b = 1, \ldots, B \); (ii) calculate the null SPU test statistics \( \text{SPUs}(\gamma; Z(b)) \); (iii) the \( P \)-value for \( \text{SPUs}(\gamma; Z) \) is \( \sum_{b=1}^{B} I(\text{SPUs}(\gamma; Z(b)) \geq |\text{SPUs}(\gamma; Z)| + 1)/(B + 1) \), and that for \( \text{SPUs}(\gamma; Z(b)) \) is \( \frac{P(b)}{\Phi(1 - p/2)}(\Phi(B - 1)) \); (iv) calculate \( a\text{SPUs}(\gamma; Z(b)) = \min_{\gamma \in B} P(b) \); (v) finally the \( P \)-value for the \( a\text{SPUs} \) test is \( \frac{P(a\text{SPUs})}{\sum_{b=1}^{B} I(a\text{SPUs}(\gamma; Z(b)) \leq a\text{SPUs}(\gamma; Z)) + 1}/(B + 1) \).

Sometimes only the \( P \)-values for individual SNPs are available while we do not know the association direction for each SNP. In this case, we calculate \( Z(c) = \Phi(1 - p/2) \), where \( p \) is the \( P \)-value for SNP \( j \) and \( \Phi(t) \) is the CDF of the standard normal distribution \( N(0, 1) \). Then we will replace \( Z(c) \) by \( |Z(c)| \) in the above formula (2) (and do so similarly for the null statistics too).

### 2.2 A pathway-based adaptive test with summary statistics

The idea of substituting a Score vector with Z-scores can be similarly extended to an adaptive pathway-level test (Pan et al., 2015). Given a pathway \( S \) with \( |S| \) genes, we partition its Z-scores as \( Z = (Z_1, \ldots, Z_{|S|}) \) with subvector \( Z_g = (Z_{c1}, Z_{c2}, \ldots, Z_{ck_g}) \) for gene \( g \) with \( k_g \) SNPs. Then we define the gene- and pathway-based SPU tests as

\[
\text{SPU}(\gamma; g) = ||Z_g||_\gamma = \left( \sum_{i=1}^{k_g} Z_{gi}^\gamma \right)^{1/\gamma},
\]

\[
\text{PathSPU}(\gamma; S) = \sum_{g \in S} \text{SPU}(\gamma; g)^{1/\gamma},
\]

where two integers \( \gamma > 0 \) and \( \gamma_G > 0 \) are used to adaptively weight the SNPs and genes respectively. For example, a larger \( \gamma_G \) (or \( \gamma \)) is more effective when there are a smaller number of genes (or SNPs) associated with the trait. Since the optimal values of \( (\gamma, \gamma_G) \) are unknown, to adaptively choose \( (\gamma, \gamma_G) \), we propose

\[
a\text{SPUsPath}(\gamma, \gamma_G; S) = \min_{\gamma \in \Gamma, \gamma_G \in \Gamma_G} P(\text{PathSPU}(\gamma, \gamma_G; S)),
\]

aiming to select the most powerful one from multiple PathSPU tests. We used \( \Gamma = \{1, 2, \ldots, 8, \infty\} \) and \( \Gamma_G = \{1, 2, 4, 8\} \) as in Pan et al. (2015). A Monte Carlo simulation scheme as described for the SPUs and aSPUs tests is used to obtain the \( P \)-values. However, now the dimension of \( Z \) is possibly larger, leading to a much larger correlation matrix \( R \). For computational efficiency, we assume that any SNPs in different chromosomes are independent, leading to a block-diagonal \( R \).

Given that most genes and pathways will not be significant in most applications, we employ a stage-wise strategy to gradually increase the simulation number \( B \) to save computing time for aSPUs and aSPUsPath. In the later applications, we first performed \( B = 5000 \) simulations for all genes and pathways, and then increased \( B \) to \( 10^6 \) or \( 10^7 \) for those genes (or \( 10^5 \) for pathways) with a \( P \)-value < 0.003.

### 3 Applications

#### 3.1 WTCCC data

To demonstrate the validity and performance of our proposed method, we designed a ‘Control–Control’ experiment and a usual Case–Control experiment using Wellcome Trust Case Control Consortium (WTCCC) GWAS data for Crohn’s disease (CD) (Consortium, 2007). CD is an autoimmune disease with a strong genetic component. The WTCCC GWAS dataset contains about 2000 CD-affected cases and about 3000 controls with a total of 500 568 SNPs. Following the WTCCC’s quality control (QC) recommendations, we removed subjects and SNPs that did not pass the QC criteria, resulting in 469 612 SNPs in 1748 case subjects and 2938 control subjects. We further removed SNPs with MAF <5% since we would use a small reference panel to infer the LD structure for a set of SNPs.

In a Control–Control experiment, we randomly divided the controls into two groups, each with 1469 samples, then tested for possible association between the group indicator and a gene or a pathway. Since no association was expected, the goal was to investigate how well a method could control type I errors or false positives. In a Case–Control analysis, the two groups contained 1748 cases and 2938 controls respectively, and we’d like to see the power of a method to detect possible associations. We calculated the \( Z \)-score for each SNP based on individual-level data, then used only the \( Z \)-scores in subsequent analyses for any summary statistics-based method. We analyzed 4572 genes mapped to 186 KEGG pathways. The significance threshold after the Bonferroni correction was 0.05/4572 = 0.000108 for gene-based analysis, or 0.05/186 = 0.00027 for pathway-based analysis. The simulation number of \( B = 10^6 \) was used as a default, though we also used \( B = 10^5 \) for comparison in aSPUs and aSPUsPath.

#### 3.1.1 Choice of the reference panels

We first investigated how the choice of the reference panels in estimating LD among SNPs might influence the performance. We considered three ways to estimate a correlation matrix \( R \) among the SNPs: using (i) 90 Hapmap CEU samples, (ii) 100 randomly chosen WTCCC control samples and (iii) the whole 2938 WTCCC controls.

Figure 1 shows the QQ-plots of the aSPUs \( P \)-values based on each of the three ways to estimate the correlation matrix for the Control–Control experiment (upper row) and Case–Control experiment (lower row). Based on the Control–Control analysis, we conclude that using the Hapmap samples seemed to give a bit inflated type I errors while the other two performed well. The less desirable performance of using the Hapmap samples could be due to the small sample size, leading to unstable estimates. However, by comparison with that of using 100 WTCCC controls, it was more likely due to some inherent differences between the two panels.

Treating using the whole WTCCC controls as the gold standard, we compared in more details about the estimated \( P \)-values in Figure 2. Most points are on the identity line. However some points are above the line, implying that using the Hapmap panel over-estimated the significance levels of several genes (i.e. with smaller \( P \)-values). Nevertheless, the agreement of the aSPUs \( P \)-values based on the reference panels was high: their correlation coefficients in both experiments were 0.98. In summary, the performance using any reference panel was mostly satisfactory with an estimated inflation factor \( \hat{\lambda} \) close to 1 (e.g. \( \hat{\lambda} = 1.07 \) for the Hapmap panel for the Control–Control experiment of the WTCCC data).

For comparison, we also included the QQ-plots for GATES in Supplementary Figure S1. Since GATES uses a few most significant \( P \)-values to reach a gene-level \( P \)-value, it is more robust to estimation errors of the LD structure among SNPs, giving less inflated significance levels in the extreme. However, due to its numerical
three methods, aSPUsPath, GATES-Simes and HYST. Perhaps due

approximations in P-value calculations, it had slightly higher inflation factors (e.g. \( \lambda = 1.2 \) for the Hapmap panel for the Control–Control experiment of the WTCCC data).

3.1.2 Choice of simulation number

Next we investigated the effects of the simulation number \( B \) on the P-value of the aSPUs test. Figure 3 compares the results with \( B = 10^5 \) versus \( B = 10^6 \); we used either the 100 Hapmap samples or the whole WTCCC controls to estimate the correlation matrices in the Control–Control or Case–Control analysis. Overall there was barely any difference between using \( B = 10^5 \) and \( B = 10^6 \). As expected, the more significant a P-value, a larger \( B \) is required. In particular, \( B = 10^6 \) would be sufficient to estimate P-values \( \geq 0.00001 \) for most genes.

3.1.3 Pathway analysis

We first investigated how the choice of the reference panel might influence the performance of a pathway-level test. Again we used either the 90 Hapmap samples or all WTCCC controls to estimate the correlation matrices. Figure 4 compares the results for each of the three methods, aSPUsPath, GATES-Simes and HYST. Perhaps due to more SNPs in a gene set (pathway), aSPUsPath appeared to be more robust than aSPUs to estimation errors using a small reference panel. Again GATES-Simes was very robust due to its use of only a few top SNPs and genes to construct its test statistic. GATES-Simes identified the same 5 pathways with any choice of the reference panels based on the threshold of \( P < 0.000025 \). HYST seemed to be more influenced by incorrectly estimated correlation matrices (due to its using Fisher’s method to combine the P-values of all the genes in a pathway). HYST identified the same 3 pathways using either the Hapmap panel or 100 WTCCC controls. However, it identified 12 pathways using all WTCCC controls as the reference panel. This indicates that the accuracy of estimating the correlation matrix is critical to HYST. In comparison, aSPUsPath identified 17 significant pathways using either the 100 or all WTCCC controls as the reference panel, among which 15 pathways were common; the Pearson correlation between the two sets of the P-values was 0.998. Using the Hapmap panel, aSPUsPath identified 16 significant pathways, among which 11 pathways were common with those identified by using the whole WTCCC control as the reference panel.

Fig. 1. QQ plots of aSPUs P-values from a Control–Control analysis (top rows) and a Case–Control analysis (bottom row) of the WTCCC CD data. The correlation matrices were estimated using the 90 Hapmap CEU samples, 100 and all 2938 WTCCC controls respectively.

Fig. 2. Comparison of \(-\log_{10}(p)\) transformed P-values of aSPUs with correlation matrices estimated from the Hapmap panel versus from all WTCCC controls. The dotted lines indicate the significance threshold. The Pearson correlation coefficients in the two panels are both 0.98.

Fig. 3. Comparison of \(-\log_{10}(p)\) transformed P-values of aSPUs with the simulation number \( B = 10^5 \) versus \( B = 10^6 \). The dotted lines indicate the Bonferroni-corrected significance threshold.

Fig. 4. Comparison of \(-\log_{10}(p)\) transformed P-values with the Hapmap samples versus all WTCCC controls for three pathway-level tests. The dotted lines indicate the Bonferroni-corrected significance threshold. The P-values < 0.00001 pointed as 0.00001 for easier comparisons among plots.
Table 1 shows 17 highly significant KEGG pathways with P-values less than 0.00001 by any of the three methods. The aSPUsPath identified nine such highly significant pathways using either all or 100 WTCCC controls as the reference panel, a much larger number than those of the other two methods, suggesting possibly higher power of aSPUsPath for the WTCCC data. Furthermore, the close performance between using the two sets of the WTCCC controls as reference panels suggested that using 100 samples from the corresponding population worked well enough in identifying significant pathways.

The five KEGG pathways that have been confirmed to be associated with susceptibility to CD by meta-analysis and replication studies (Franke et al., 2010; Jostins et al., 2012; Wang et al., 2010) are all among the 17 significant pathways.

We also performed analysis with one GSEA method, i-GSEA4GWAS, but it did not find any significant pathway.

### 3.2 ICBP data

We illustrate the application of the methods to the summary statistics of a meta analysis by the International Consortium for Blood Pressure Genome-Wide Association Studies (ICBP GWAS) (Ehret et al., 2011). The data include the P-values of 2.6 million SNPs for the diastolic blood pressure (DBP) based on the discovery sample of 29 studies with 69,395 individuals of European ancestry. We obtained the genomic coordinates of the SNPs and genes according to 29 studies with 69,395 individuals of European ancestry. We obtained the genomic coordinates of the SNPs and genes according to the human reference genome hg19, and assigned SNPs within 2 kb upstream or downstream a gene to the gene using software MAGMA (de Leeuw et al., 2013). About 1,095,843 (41.04%) SNPs were mapped into at least one of 17,543 genes. We set the gene-level significance threshold at 0.05/17,543 = 2.85 × 10^{-6} based on the Bonferroni correction. We used the Hapmap CEU reference panel to estimate the correlations among the SNPs, downloaded from the Plink website (Purcell et al., 2007).

#### 3.2.1 Gene-based analysis

Figure 5 shows the significant genes identified by aSPUs, GATES (Li et al., 2011), VEGAS (Liu et al., 2010) and MAGMA (de Leeuw et al., 2013). We implemented the first two methods but used the original software for the latter two. In total 49 significant genes were identified by at least one method, in which 14 were common across all the methods. In comparison with a single SNP-based analysis, Ehret et al. (2011) identified 29 significant independent SNPs in 28 loci or genes for SBP and/or diastolic BP (DBP) based on a much larger sample combining both the discovery and validation samples with up to 133,661 additional individuals; even so, if only SBP was considered, 4 of the 29 SNPs were no longer significant. In contrast, as shown in Figure 3, 12 genes out of the 28 loci were detected by gene-based testing with a much smaller sample size of the discovery sample. In fact, based on single SNP analysis of only the discovery sample as used here, only 8 of the 29 independent SNPs/loci in Table 1 of Ehret et al. (2011) were significant. The Manhattan plots for the various methods are shown in Supplementary Figure S5.

Since GATES is essentially a univariate method with a modified Simes procedure for multiple testing adjustment, as expected, its results were similar to that of SPUs(<i>p</i>), which is the univariate minimum P-value method: SPUs(<i>p</i>) detected 32 associated genes, among which 28 genes overlapped with those of GATES; in comparison, SPUs(1) and SPUs(2) identified 25 and 28 significant genes, among which only 17 and 20 genes were common to those of GATES. On the other hand, the test statistic of VEGAS is \( \sum_{j=1}^{k} Z_j^2 = \text{SPUs}(2) \), though VEGAS was implemented differently in

### Table 1. Pathways of 17 KEgg pathways with \( P \)-values \( < 0.00001 \) by at least one of the aSPUsPath, GATES-Simes and HYST tests using the 90 Hapmap CEU samples, 100 WTCCC controls or all WTCCC controls to estimate the SNP correlation matrices.

<table>
<thead>
<tr>
<th>KEGG ID</th>
<th>Pathway Names</th>
<th>Hapmap CEU (90)</th>
<th>WTCCC (100)</th>
<th>WTCCC (100)</th>
<th>WTCCC (100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa04060</td>
<td>Cytokine-cytokine receptor interaction*</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>hsa04630</td>
<td>Jak-STAT signaling pathway*</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>hsa05131</td>
<td>NOD-like receptor signaling pathway*</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>hsa04660</td>
<td>T cell receptor signaling pathway*</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>hsa04972</td>
<td>Pancreatic secretion</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>hsa04640</td>
<td>Hematopoietic cell lineage</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>hsa04520</td>
<td>Adherens junction</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>hsa05320</td>
<td>Autoimmune thyroid disease</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>hsa04310</td>
<td>Wnt signaling pathway</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>hsa05330</td>
<td>Allograft rejection</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>hsa04622</td>
<td>RIG-I-like receptor signaling pathway*</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>hsa04621</td>
<td>Chemokine signaling pathway*</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>hsa04620</td>
<td>Axon guidance</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>hsa04619</td>
<td>Vascular smooth muscle contraction</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>hsa05416</td>
<td>Viral myocarditis</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
</tr>
</tbody>
</table>

Note: * indicates that the pathway is also significant by single SNP analysis.
SNP/gene mapping and covariance estimation (Liu et al., 2010), we would still expect that VEGAS and SPU(2) performed similarly: it was confirmed that 21 of the 28 genes identified by SPU(2) overlapped with that of VEGAS.

The gene ULK4 was uniquely identified by aSPUs. As shown in Figure 6, gene ULK4 contains many nearly significant SNPs and many non-significant SNPs, requiring an adaptive test like aSPUs with effective SNP weighting (or selection) to detect it. For gene HIST1H4C that was identified by GATES but not by other methods, it contained 11 SNPs; only one of the SNPs had a P-value around 2e-7 while those for other SNPs were all around 0.1 or 0.01. The P-values of aSPUs, GATES, MAGMA and VEGAS were 4.70e-6, 1.29e-6, 0.0015 and 0.00023 respectively. Due to the sparse signal with only one significant SNP, it favored a univariate method like GATES; nevertheless, due to the inclusion of SPU(∞) (with a P-value of 2.0e-6), the P-value of the aSPUs test was almost significant. Now consider another gene BRAP that was identified by aSPUs, VEGAS and MAGMA but not by GATES. It contained 10 SNPs, 6 of which were with a P-value around 2e-6 while other SNPs were much less significant. The P-values of aSPUs, GATES, MAGMA and VEGAS were 1.3e-6, 5.06e-6, 2.065e-7 and 1e-7 respectively, confirming an advantage of aSPUs in accumulating non-sparse and weak signals with multiple moderately associated SNPs.

### 3.2.2 Pathway-based analysis
Finally we conducted a pathway-level analysis. We used the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa et al., 2010) downloaded from MSigDB (Subramanian et al., 2005). There were 186 pathways with 4874 genes and 309 347 SNPs mapped for trait DBP. We restricted the analysis to SNPs with MAF > 5%. We randomly removed some adjacent SNPs within 50kb in high LD with \( r > 0.9 \) based on the genotype data of the Hapmap CEU sample (release 2). We set the significance threshold at 0.05/186 = 0.000268 based on the Bonferroni correction.

In total 42 pathways were identified to be significant, including 24 by aSPUsPath, 38 by GATES-Simes, 34 by Hyst and 7 by MAGMA, as shown in Figure 7. Here GATES-Simes and Hyst identified more pathways than aSPUsPath and MAGMA. Nevertheless, aSPUsPath uniquely identified 16 genes, including 4 known BP-related genes in Table 1 of Ehret et al. (2011); out of a total of 16503 SNPs in this pathway, 55, 21 and 6 SNPs were significant at P < 1e-5, 1e-6 and 1e-7 respectively, demonstrating multiple moderate associations (in multiple genes) and thus lower power of a univariate test like GATES-Simes.

### 4 Discussion
We have proposed two adaptive tests, aSPUs and aSPUsPath, respectively for gene- and pathway-level association analyses of a univariate trait with the availability of only summary statistics, such as Z-statistics or P-values, for individual SNPs. With only summary statistics like Z-statistics or P-values, one has to recourse to some reference panel to estimate LD or correlations among SNPs, which
might influence the performance of subsequent association testing. Overall, the P-values calculated from using different panels were similar as shown in our analysis of the WTCCC data. However, depending on the reference panels being used, the results for a small number of the genes or pathways could be different. Hence, caution must be taken, and an independent validation on any detected associations based on summary statistics becomes more important.

We have further illustrated the application of the proposed methods to a meta-analyzed GWAS data, demonstrating their usefulness as compared to the popular single SNP-based analysis. In addition, we have compared their performance with several other existing tests. Since there is no uniformly most powerful test for multiple SNPs, it would be desirable to have an adaptive test that can robustly maintain high, not necessarily highest, power across various scenarios, which motivated our proposed two tests (Pan et al., 2014, 2015). Our proposed tests offer an alternative and complement to existing gene- and pathway-level association testing. We have developed an R package aSPU to facilitate their use.

Acknowledgements

The authors thank the reviewers for helpful comments to improve the manuscript. This research was supported by NIH grants R01GM113250, RO1HL105397 and RO1HL116720, and by the Minnesota Supercomputing Institute at University of Minnesota.

Conflict of Interest: none declared.

References


