Biometrics 000, 000–000 DOI: 000

000 Submitted January 9, 2015

A Powerful Pathway-Based Adaptive Test for Genetic Association With Common or Rare Variants

Wei Pan[1,3], Il-Youp Kwak[1], Peng Wei[2,3]

[1] Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455

[2] Division of Biostatistics and Human Genetics Center,
University of Texas School of Public Health, Houston, TX 77030

[3] Co-corresponding authors

**email:* weip@biostat.umn.edu; peng.wei@uth.tmc.edu

Summary: In spite of the success of genome-wide association studies (GWASs), only a small proportion of heritability for each complex trait has been explained by identified genetic variants, mainly single nucleotide polymorphisms (SNPs). Likely reasons include genetic heterogeneity (i.e. multiple causal genetic variants) and small effect sizes of causal variants, for which pathway analysis has been proposed as a promising alternative to the standard single SNP-based analysis. A pathway contains a set of functionally related genes, each of which includes multiple SNPs. Here we propose a novel pathway-based test that is adaptive at both the gene and SNP levels, thus maintaining hight power across various situations with varying numbers of the genes and SNPs associated with a trait. We use both simulated and real data to compare our proposal with several existing pathway-based and SNP set-based tests, demonstrating its promising performance and its potential use in practice.

Key words: aSPU test; Genome-wide association studies (GWAS); GRASS; PLINK; SNP; SPU and SSU tests.

1. Introduction

Genome-wide association studies (GWASs) have been successful in identifying many genetic variants, mainly single nucleotide polymorphisms (SNPs), associated with complex and common disease (Hindorff et al 2010). However, only a small proportion of the estimated heritability for most human complex traits can be explained by the identified genetic variants. One possible reason is that, due to the small effects sizes and genetic heterogeneity (i.e. multiple causal variants), the standard single SNP-based analysis may not have enough power to identify many causal variants. While many human genetic diseases are caused by multiple genes, it has been increasingly recognized that, because genomic variants of these genes lead to the same or similar phenotype, these genes are likely to be functionally related, and such functional relatedness can be exploited to identify novel disease genes. One way to organize functionally related genes is through biological pathways, such as annotated in the KEGG database (Kanehisa et al 2010). Association analysis of multiple genes with related functions is here generically called *pathway analysis* (or gene set analysis), which may improve power over testing on single SNPs or single genes one by one. One convincing source of evidence is from tumor sequencing studies, e.g. The Cancer Genome Atlas (TCGA) (2011). While few cancer genes (e.g. TP53) harbor many mutations, most harbor few mutations in a tumor-dependent way. For example, a tumor may contain mutations in gene PTEN, not in gene NF1, while another tumor contains mutations in gene NF1, not in gene PTEN. Individually, each of the genes in a related pathway has only a low mutation frequency, but collectively, they have a much higher mutation frequency. Hence, for a disease (e.g. cancer) involving a few pathways, a pathway analysis by aggregating information across the genes in a relevant pathway will boost statistical power, and thus is preferred. For example, among the 316 ovarian cancer tumors studied by TCGA, 45% of them had genomic alterations (somatic mutations and DNA copy number changes) in the PI3K/RAS signaling pathway, which contains 7 genes, PTEN, PIK3CA, AKT1, AKT2, NF1, KRAS and BRAF, each with only low to moderate genomic alterations in 7%, 18%, 3%, 6%, 12%, 11% and 0.5% of the tumors respectively; hence, it should be more powerful to detect genomic alterations in the pathway level than in the individual gene level.

The importance of pathway analysis and many existing approaches have been reviewed by several authors (Wang et al 2010; Fridley and Biernacka 2011; Wang et al 2011). Many pathway-based analysis methods for GWAS data are evolved from those for gene expression data (e.g. Goeman et al 2004; Newton et al 2007); however, higher-dimensional data are involved in the former with up to hundreds to thousands of SNPs, compared to only tens to hundreds of genes in the latter. On the other hand, since it is known that not all the SNPs in any gene or any pathway are related to a disease, statistically it is most important and challenging to adaptively aggregate information over multiple unknown causal SNPs while minimizing the effects of non-causal SNPs. *Existing approaches have some limitations.* For example, the first-ever approach (Wang et al 2007) used the minimum p-value of the multiple SNPs in a gene to summarize association information for the gene, which is not efficient if there are multiple weakly associated SNPs inside the gene; a very recent approach (Schaid et al 2012) uses a variance-component test to aggregate information across multiple SNPs non-adaptively, which will lose power in the presence of many non-associated genes. Our proposal is based on a highly adaptive test called adaptive sum of powered score (aSPU) test for analysis of rare variants (RVs) (Pan et al 2014). The main idea of the aSPU test is that, since we do not know which and how many SNPs in the given set are associated with a trait, we first construct a class of tests over-weighting a sequence of increasingly smaller sets of the top ranked (i.e. most statistically significant) SNPs, then select the test with the most significant result (with a proper adjustment for multiple testing). For relatively small sets of RVs, the aSPU test may outperform other tests (Pan et al 2014). Here we extend the aSPU test to pathway analysis of GWAS data, including either common variants (CVs) or RVs. Among others, since the analysis unit of a pathway-analysis is gene while genes may contain quite different numbers of SNPs, we need to modify the aSPU test to treat each gene equally a priori. More importantly, the proposed test is adaptive with respect to both genes and SNPs, which is critical since we do not know a priori how many genes in a pathway are associated, and how many SNPs in an associated gene are associated with the given trait. We will compare our proposal with two popular pathway-based methods, one based on penalized regression called GRASS (Chen et al 2010) and the other as a representative two-step approach based on SNP-screening then combining as implemented in the popular software PLINK (called Plink simply in the sequel) (Purcell et al 2007), largely because the latter two methods have been widely used in GWAS in practice (e.g. Deelen et al 2013; Verschuren et al 2013; Wei et al 2013).

2. Methods

2.1. Data and notation

We consider the most popular case-control study design as adopted in GWAS, though the methods can be extended to other study designs, e.g. with a quantitative or survival trait. Suppose that for subject i, $i = 1,...,n$, $Y_i = 0$ or 1 is a binary trait, e.g. an indicator of disease, and $X_i = (X_{i1}, ..., X_{ik})'$ is the vector of the genotype scores for k SNPs, possibly drawn from multiple genes in a pathway. We use additive coding for each SNP; that is, X_{ij} is the number of the copies of an allele at SNP j for subject i. It is possible to include other covariates, but for simplicity we ignore them. We consider a logistic regression model:

$$\text{Logit}[Pr(Y_i = 1)] = \beta_0 + \sum_{j=1}^{k} X_{ij}\beta_j. \qquad (1)$$

We'd like to test the null hypothesis H_0: $\beta = (\beta_1, ..., \beta_k)' = 0$; that is, there is no association between any SNPs and the trait under H_0. The score vector $U = (U_1, ..., U_k)'$ for β and its covariance matrix are

$$U = \sum_i X_i(Y_i - \bar{Y}), \qquad V = Cov(U) = \bar{Y}(1-\bar{Y})\sum_i (X_i - \bar{X})(X_i - \bar{X})',$$

where $\bar{Y}$ and $\bar{X}$ are the sample means of Y_i's and X_i's respectively. The classic score test statistic is $T_{Score} = U'V^{-1}U$, which, however, in the current context with a large k relative to the sample size n, may be low powered, as its asymptotically equivalent Wald test and likelihood ratio test. As shown theoretically in Fan (1996), as the dimension k increases, the power of the score test may diminish, tending to the Type I error rate α. The most popular univariate single SNP-based test, call UminP here, is $T_{UminP} = \max_{j=1}^{k} U_j^2/V_{jj}$ with $V_{jj} = \mathrm{Var}(U_j)$, which may also be low powered if we have many small $|\beta_j| \neq 0$. Two alternatives, called the Sum and SSU tests, are

$$T_{Sum} = 1'U/\sqrt{1'V1} = \sum_{j=1}^{k} U_j/\sqrt{1'V1}, \qquad T_{SSU} = U'U = \sum_{j=1}^{k} U_j^2.$$

The Sum test is powerful when all or most $|\beta_j| \neq 0$ with the same sign, but not otherwise. As shown in Pan (2011), the SSU test can be regarded as a variance-component test (Tzeng et al 2011; Wu et al 2010), and is closely related to an empirical Bayes test for high-dimensional data (Goeman et al 2006) and a nonparametric MANOVA test (Wessel and Schork 2006). In particular, variance-component tests, including kernel machine regression (KMR), have been advocated for SNP set analysis and empirically shown to be powerful in many cases (Tzeng et al 2011; Kwee et al 2008; Wu et al 2010). Neverthelee, as shown in Pan et al (2004), since a variance-component test is not adaptive, in the presence of many non-associated SNPs as anticipated in the current context of pathway analysis, it may lose power. Accordingly, a more powerful and adaptive test was proposed (Pan et al 2014) as reviewed next.

2.2. Review: the data-adaptive aSPU test

Pan et al (2011) proposed a class of *sum of powered score* (SPU) tests in a different context for analysis of RVs:

$$T_{\text{SPU}} = T_{SPU(\gamma)}(U) = \sum_{j=1}^{k} U_j^{\gamma} \tag{2}$$

The SPU tests cover the Sum and SSU tests as two special cases with a corresponding $\gamma = 1$ and $\gamma = 2$ respectively. Importantly, as $\gamma \to \infty$ (and as an even integer), then the SPU test would approach the UminP test if the variances of the score components are a constant (or if their varying variances are ignored, which may be advantageous in certain cases as to be shown); the reason is simple:

$$||U||_{\gamma} = \left(\sum_{j=1}^{k} |U_j|^{\gamma} \right)^{1/\gamma} \to ||U||_{\infty} = \max_{j=1}^{k} |U_j|, \qquad \text{as } \gamma \to \infty.$$

Without covariates, we propose using permutations to obtain p-values. More generally, to adjust for covariates, the parametric bootstrap (or, alternatively, permuting residuals) can be used for inference. Specifically, we will first fit a null model under H_0, then simulate a new set of traits $Y^{(b)}$'s from the fitted null model for $b = 1, ..., B$; we calculate the test statistic $T_{SPU}^{(b)}$ based on each set of simulated $Y^{(b)}$; finally we calculate the p-value as $\sum_{b=1}^{B} [I(|T_{SPU}^{(b)}| \geqslant |T_{SPU}|) + 1]/(B+1)$. We used $B = 500$ in our simulations for a nominal significance level at 5%.

There is no uniformly most powerful test in multilocus association testing; on the other hand, it has been found empirically that the Sum, SSU and UminP tests performed well under different situations, as to be confirmed. For a given dataset, to adaptively choose the value of γ for the SPU tests, Pan et al (2014) propose an adaptive SPU (aSPU) test that simply combines the results of multiple SPU tests: suppose that we have some candidate values of γ in Γ, e.g. $\Gamma = \{1, 2, 3, ..., 8\}$ as used in our later experiments, and suppose that the p-value of the $SPU(\gamma)$ test is p_{γ}, then the aSPU test simply takes the minimum p-value:

$$T_{\text{aSPU}} = \min_{\gamma \in \Gamma} p_{\gamma}.$$

Of course, T_{aSPU} is no longer a genuine p-value; we recourse to the parametric bootstrap to estimate its p-value. As before, first, we simulate B independent copies $Y^{(b)}$ from the null distribution of Y, and obtain the null score vectors $U^{(b)}$ for $b = 1, 2, ..., B$. We then calculate the corresponding SPU test statistics $T^{(b)}_{SPU(\gamma)}$ and their p-values $p^{(b)}_\gamma = \sum_{b_1 \neq b} I(T^{(b_1)}_{SPU(\gamma)} > T^{(b)}_{SPU(\gamma)})/(B-1)$. Thus, we have $T^{(b)}_{aSPU} = \min_{\gamma \in \Gamma} p^{(b)}_\gamma$, and the final p-value of the aSPU test $P_{aSPU} = \sum_{b=1}^B [I(T^{(b)}_{aSPU} \leqslant T_{aSPU}) + 1]/(B+1)$.

2.3. A data-adaptive pathway-based test: aSPUpath

Given a pathway S with $|S|$ genes, we partition the score vector according to the genes as $U = (U'_{1.}, ..., U'_{|S|.})'$ with the score subvector for gene g (with k_g SNPs) as $U_{g.} = (U_{g1}, U_{g2}, ..., U_{gk_g})'$ based on the logistic regression model (or other GLM or PHM). The gene-specific SPU statistic and the pathway-based SPU statistic are respectively

$$\text{SPU}(\gamma, w_g; g) = ||U_{g.}||_\gamma = \left(\sum_{j=1}^{k_g} (w_{gj}|U_{gj}|)^\gamma / k_g \right)^{1/\gamma}, \tag{3}$$

$$\text{PathSPU}(\gamma, \gamma_G, w, w_G; S) = \sum_{g \in S} (w_{G,g}\text{SPU}(\gamma, w_g; g))^{\gamma_G}, \tag{4}$$

where two scalars $\gamma > 0$ and $\gamma_G > 0$, gene specific weights for SNPs $w = (w'_1, .., w'_{|S|})'$ and $w_g = (w_{g1}, ..., w_{gk_g})'$, and gene-specific weights for genes $w_G = (w_{G,1}, ..., w_{G,|S|})'$ are pre-specified. w_g is used to incorporate prior information on SNPs, e.g. to up-weight SNPs associated with gene expression, while w_G can be based on gene functional annotations or gene expression data to represent prior likelihoods of their being functional (and associated with the traits); without prior knowledge or data, or for simplicity, we can simply use $w_g = 1$ and $w_G = 1$, which are to be used by default unless specified otherwise in this paper. Note that $SPU(\gamma, w_g; g)$ is standardized by the gene-specific number of SNPs, k_g, so that large genes will not dominate a pathway analysis (since the genes in a pathway are the analysis units and are thus treated equally a priori if no weighting is desired). The intuition behind using γ_G is like that for γ: a larger γ_G (or γ) is more effective if there are fewer associated genes (or SNPs) with larger effects in a pathway (or in a gene), but not otherwise. An extreme situation is to use $\gamma_G = \infty$, i.e. only using the most significant gene to form the test statistic. Since the goal of a pathway-based analysis is to take advantage of possibly multiple associated genes, we suggest to try $\gamma_G \in \{1, 2, 4, 8\}$ as shown in the below results, though this needs to be further studied.

For any given (γ, γ_G), as for $SPU(\gamma)$, we recourse to resampling to calculate its p-value $P_{\text{PathSPU}(\gamma,\gamma_G,w,w_G;S)}$. Its power depends on the choice of (γ, γ_G). A **pathway-based aSPU** test is defined as

$$\text{aSPUpath}(S) = \min_{\gamma,\gamma_G} P_{\text{PathSPU}(\gamma,\gamma_G,w,w_G;S)}, \tag{5}$$

aiming to select from multiple PathSPU tests the most powerful one. Similar to that for the aSPU test, we propose using a single layer of the permutation or parametric bootstrap to calculate the p-values.

For the possible situation where multiple causal genes in a pathway may contain quite different proportions of causal SNPs, we may use a more general pathway-based test with a gene-specific γ_g for each gene g. Denote $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_{|S|})'$, we can modify the tests as

$$\text{PathSPU2}(\boldsymbol{\gamma}, \gamma_G, w, w_G; S) = \sum_{g\in S} (w_{G,g}\text{SPU}(\gamma_g, w_g; g))^{\gamma_G}, \tag{6}$$

$$\text{aSPUpath2}(S) = \min_{\boldsymbol{\gamma},\gamma_G} P_{\text{PathSPU2}(\boldsymbol{\gamma},\gamma_G,w,w_G;S)}. \tag{7}$$

The corresponding aSPUpath2 test is computationally more demanding in searching for suitable values of more parameters in $\boldsymbol{\gamma}$ and γ_G, which will also introduce more variability to the results and thus may lead to loss of power. This needs to be further studied.

2.4. Other modifications

We also considered single gene-based approaches and those based on dimension reduction. Since they did not outperform the proposed aSPUpath, we will skip their detailed discussion except the below summary that may be interesting.

As a representative of single gene-based approaches, we considered applying SPU and aSPU tests to each gene, then using the minimum p-value to combine their p-values. It is easy to see that the pathway-based SPU(∞) (after ignoring the inverse weighting by the number of SNPs and the possible use of weights) and single gene-based SPU(∞) are almost the same, hence our proposed aSPUpath test is more adaptive and thus expected to be more flexible and powerful.

For dimension reduction, as in GRASS, for each gene we replaced its individual SNP genotype scores by their top few principal components (PCs) that accounted for at least 95% of total variation, then we applied the pathway-based aSPU test to these PCs. Perhaps due to the adaptivity of the original aSPUpath test and possible loss of information by PCs, we did not find improvement by the use of PCs in our simulations. However, given that PC-based tests (e.g. Wang and Abbott 2007; Chen et al 2010b) are viable competitors to variance-component tests as discussed in Schaid et al (2012), we had an interesting, perhaps surprising, observation: applying the SPU(2) (i.e. SSU) test (that is equivalent to a variance-component test) to the original genotypes or the PCs gave almost the same result; an explanation is offered below.

Suppose that X is the $n \times k$ matrix of the original genotype scores. We apply a singular value decomposition: $XX' = V\Lambda^2V'$, where we assume that the eigen-values have been put in descending order as the diagonal elements of Λ^2. The first L PCs are the columns of $P_L = V_L\Lambda_L$, where V_L is an $n \times L$ matrix containing the first L columns of V and Λ_L is an $L \times L$ diagonal matrix containing the first L eigen-values. Now we can compare the two SSU statistics when applied to X and P_L respectively:

$$\begin{aligned} SSU(X) &= U(X)'U(X) = (Y-\bar{Y})'XX'(Y-\bar{Y})' = (Y-\bar{Y})'V\Lambda\Lambda'V'(Y-\bar{Y})' \\ &\approx (Y-\bar{Y})'V_L\Lambda_L\Lambda_L'V_L'(Y-\bar{Y})' = SSU(P_L). \end{aligned}$$

But for other $\gamma \neq 2$, we would expect that, in general, SPU(γ) would give different results when applied to the original genotype scores X and its top PCs P_L respectively.

3. Simulations

3.1 Simulation set-ups

We conducted extensive simulation studies to evaluate and compare the performance of the aSPU test with several alternative methods. Our general set-ups were similar to those (set-ups A-D) in Chen et al (2010) except that we simulated SNPs, not PCs (called eigenSNPs therein) of SNPs, to mimic real data. Specifically, set-up A was the null case with no causal gene, while the other three set-ups contained 1, 5 and 10 causal genes respectively. We only considered one pathway containing 20 genes, each of which might contain 1-20 SNPs, or 3-100 SNPs; by default, there was only 1 causal SNP inside each causal gene, though we also considered the cases with 1-3 causal SNPs in each causal gene. The SNPs inside each gene might or might not be correlated while the SNPs from different genes were always independent, and the causal SNPs might or might not be included in the data.

The simulated genotypes were generated as in Wang and Elston (2008). First, we generated a latent vector $Z = (Z_1, ..., Z_k)'$ from a multivariate Normal distribution with a first-order auto-regressive (AR1) covariance structure: $Corr(Z_i, Z_j) = \rho^{|i-j|}$ between any latent components i and j; $\rho = 0$ and $\rho > 0$ randomly chosen from a uniform distribution $U(0, 0.8)$ was used to generate (neighboring) SNPs in linkage equilibrium and in linkage disequilibrium (LD) respectively. The number of SNPs inside each gene, k_g, was randomly chosen between 1 and 20, or between 3 and 100. Second, the latent vector was dichotomized to yield a haplotype with MAFs each randomly selected uniformly between 0.05 and 0.4 for CVs or between 0.001 and 0.1 for RVs. Third, we combined two independent haplotypes and obtained genotype data: $X_i = (X_{i1}, ..., X_{ik})'$ for subject i. Fourth, for a non-null case, the first SNP inside the first $k_1 = 1$ or 5 or 10 genes, corresponding to set-ups B-D, was chosen to be causal with $\beta_j = \log \mathrm{OR} \neq 0$, while all other $\beta_j = 0$; we also tried set-ups B'-D' with 1-3 randomly chosen causal SNPs. For the null case, all $\beta_j = 0$. Fifth, the disease status Y_i of subject i was generated from the logistic regression model (1). We used $\beta_0 = -\log(0.05/0.95)$ for a 5% background disease probability; that is, $Pr(Y_i = 1|X_i = 0) = 0.05$. Sixth, as in a case-control study, we sampled $n/2 = 500$ cases and $n/2 = 500$ controls in each dataset.

Throughout the simulations, we fixed the test significance level at $\alpha = 0.05$. We used the R package `SNPath` implementing GRASS and Plink. Since the program for Plink was quite slow, we only ran 100 independent replicates for Plink, but 1000 replicates for others in each set-up.

3.2 Simulation results for CVs

For comparison, we included SSU=SPU(2) and UminP tests; the former is equivalent to a global pathway-based test of Goeman et al (2004) as shown in Pan (2011), while the latter is the most popular single SNP-based test in GWAS. The UminP test often performed similarly to SPU(∞) (data not shown).

3.2.1 Type I error

As shown in Table 1, it appears that each test could control its Type I error rate satisfactorily around 0.05.

3.2.2 Comparison of the aSPUpath test with other tests

We first consider set-up B, an extreme scenario that is least favorable to pathway or SNP set analysis: since there was only one causal SNP, single SNP-based analysis as implemented in the UminP test was expected to be be most powerful, which was confirmed in Figure 1. Nevertheless, the aSPU and aSPUpath tests performed similarly and were second most powerful. In panel a) with about 200 independent SNPs, Plink was the third most powerful, followed by the SSU and then GRASS. In panel b) with about 1000 independent SNPs, the aSPU and aSPUpath tests showed even a more striking advantage over the other three pathway or SNP set-based tests, suggesting the former two's (and the latter three's) robustness (and lack of robustness) to an increasing number of SNPs. In particular, the performance of SSU deteriorated with its power close to that of GRASS. In panel c) with about 200 correlated SNPs (with the causal SNP included), the power trend was similar to that with 200 independent SNPs, though GRASS performed better than Plink and SSU with smaller ORs. In panel d) with about 200 correlated SNPs with the causal SNP excluded, again we found the UminP test, closely followed by the aSPU and aSPUpath tests, as the top winners, while the other three tests had similar power.

In set-up C with 5 causal SNPs (Figure 2), again the aSPU and aSPUpath tests performed similarly and now they had a edge over the UminP test, since the latter uses only the single SNP with the strongest signal while ignoring the signals from other 4 causal SNPs. However, differing from set-up B, we notice that the SSU test and Plink performed similarly in panels a) and c) while one was mroe powerful than the other in panels b) and d) respectively. In penal d) with the 5 causal SNPs excluded, GRASS could perform better than the other tests except the aSPU and aSPUpath tests when the causal effect size was small (and the power was low).

Now consider a case favoring pathway or SNP set analysis in set-up D with 10 causal SNPs (Figure 3), the aSPUpath test was the sole winner, having an edge over the aSPU test; in particular, the two tests could be much more powerful than the UminP test. In panels a) and c), even the SSU test was much more powerful than the UminP test, confirming the advantage of combining information across multiple SNPs. On the other hand, in panel b) with about 1000 SNPs, UminP and Plink were tied as the second most powerful, followed by SSU; the low power of SSU test was due to its non-robustness to a large number of non-associated SNPs since it did not down-weight enough the larger number of non-associated SNPs; in contrast, the two adaptive tests, aSPU and aSPUpath, did not suffer much from the presence of a large number of non-associated SNPs. With causal SNPs, GRASS could beat Plink when the causal effect size was small (in panels a and c), or when the causal SNPs were excluded (in panel d).

In all the above three situations, each causal gene contained only one causal SNP, which might be too restrictive. To cover the possible situation with more than one causal SNP inside a causal gene, we considered set-ups B'-D', in which we randomly selected 1-3 causal SNPs in each causal gene. The main results remained the same as before except the following as shown in Figure 4 for set-up D'. First, there was a larger power advantage of the aSPUpath over the aSPU test. Second, there was improved performance of GRASS: for example, for small effect sizes, GRASS was consistently more powerful than Plink, though it was still outperformed by aSPUpath.

In summary, we found that the aSPUpath and aSPU tests were much more powerful than pathway-based GRASS and Plink, and the SSU test for SNP set analysis, across all the simulation set-ups considered. In the presence of multiple causal SNPs or causal genes, as anticipated for pathway analysis, they also outperformed the single SNP-based UminP test; between the two adaptive tests, the aSPUpath test had an edge over the aSPU test in some situations, especially for the casual SNPs with small effect sizes.

3.2.2 Comparison of the aSPUpath test with its other variants

For set-up B with only one causal SNP, the single gene-based aSPU and pathway-based aSPU tests had almost identical power while being much more powerful than the PC-based aSPU test. The reason was the following. First, since there was only one single causal SNP, a single gene-based approach would not lose power as compared to a pathway-based approach aiming to combine information across multiple genes; at the same time, a pathway-based approach in general would not gain either under this situation. Second, note that the aSPU test could realize effective SNP selection by adaptively choosing the tuning parameter γ to down-weight non-associated SNPs; however, each PC is a linear combination of all the SNPs, a mixture of both associated and non-associated SNPs, hindering the ability of the PC-based aSPU test to select SNPs effectively.

For set-up C with 5 causal SNPs, the pathway-based aSPU test was more powerful than the gene-based aSPU test, while the PC-based aSPU test was still the least powerful.

For set-up D with 10 causal SNPs, the pathway-based aSPU test was by far the most powerful. For 200 SNPs, the PC-based aSPU test was more powerful than the single gene-based aSPU; however, with about 1000 SNPs, the single gene-based aSPU was more powerful than the PC-based aSPU, presumably due to the fact that each PC contained too many non-associated SNPs, diluting the association effects.

As in GRASS, we also tried to first construct gene-specific SPU test statistics before combining them across a pathway, but did not find it working better than the simple aSPUpath test discussed here.

In summary, we found that overall our proposed aSPUpath test performed better than the single gene-based aSPU and PC-based aSPU tests.

3.2.3 Other comparisons

We did a preliminary study to explore the use of informative weighting in set-up D with 10 causal genes and about 200 correlated SNPs. We applied our proposed test with $w_g = 1$, but with $w_G = 1$ or $w_G \neq 1$ to assess the effects of gene weighting (while the effects of SNP weighting could be explored similarly); we generated $w_{G,g} \sim U(0.2, 0.6)$ for non-causal genes, but $w_{G,g} \sim U(0.2 + \delta, 0.6 + \delta)$ for several values of $\delta \geqslant 0$; increasing values of δ reflected increasing informativeness of the weights while $\delta = 0$ represented completely random and non-informative weighting. As shown in Table 2, it is clear that our proposed aSPUpath test was most powerful; its weighted version was robust to mis-specified and completely random weights (with $\delta = 0$) with only small power loss, while gaining higher power with more informative weights.

3.3 Simulation results for RVs

With the increasing availability of sequencing data, it has become more important and urgent to develop and apply pathway-based analysis of RVs; currently there have been few such studies. For this purpose, we did a simulation study to assess the performance and show the promise of our proposed test for pathway analysis of RVs. To save space, we only present results for a simulation set-up similar to set-up D: a pathway contained 20 genes, 0 or 10 of which each contained one causal RV among 1-20 RVs for the null or non-null cases respectively. The MAFs for the RVs was randomly drawn between 0.1% and 1% for the control samples. We considered both independent and correlated RVs within each gene.

For comparison, we also included several existing popular or competitive tests. In particular, we included the Sum=SPU(1) as a representative pooled association (or burden) test, the SSU (i.e. SPU(2)) test that was shown by Basu and Pan (2011) to be competitive and closely related to several other association tests, C-alpha test (Neale et al 2011) and kernel machine regression or SKAT (Wu et al 2010, 2011), and three adaptive tests that appeared recently, a kernel-based adaptive clustering (KBAC) test (Liu and Leal 2010), a p-value weighted sum test (PWSU) (Zhang et al 2011) and an estimated regression coefficient (EREC) test (Lin and Tang 2011).

As shown in Table 3, all the methods seem to have Type I error rates around the nominal level of 0.05.

As shown in Figure 5, the relative performance of the various tests did not strongly depend on whether there were within-gene correlations among the RVs. Clearly the aSPUpath test was the most powerful, closely followed by the usual aSPU test, then followed by the SSU test, then GRASS, SKAT and EREC tests. Although the SSU and SKAT are closely related, since SKAT over-weights rare variants with smaller MAFs, which was not a correct assumption in our simulations, here the SSU test was more powerful than SKAT. It is worth noting that here GRASS was much more powerful than Plink, perhaps due to the latter's ineffective screening on each individual RV with only minimal information content in each individual RV.

The PWST and the single RV-based UminP test performed similarly. The KBAC had lowest power. Note that here all the causal RVs had an equal association strength (and direction), which was supposed to be ideal for the Sum test (or other burden tests); however, due to the presence of many non-associated RVs, the Sum test and several other adaptive tests did not perform well due to their non- or not-so-good selection or down-weighting of the many non-associated RVs, as discussed in Pan et al (2014).

4. Example

We applied the proposed aSPUpath tests, as well as the GRASS test, to the Wellcome Trust Case Control Consortium (WTCCC) (2007) GWAS data for Crohn's disease (CD). CD, a type of inflammatory bowel disease, is also considered as an autoimmune disease with a strong genetic component (Wang et al 2010). The WTCCC GWAS data set contains 2000 CD cases and 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 cases and 2938 controls. We further restricted the pathway analysis to SNPs with MAF of at least 1%. We retrieved a total of 214 human biological pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al 2010) To facilitate interpretation of the pathway analysis results, we excluded too small (< 10 genes) or too big (> 500 genes) pathways, leading to 197 pathways. We obtained the genomic coordinates of SNPs and genes according to human reference genome hg19, and assigned a SNP to a gene if it is located within 20000 base-pairs (20kb) upstream or downstream of the gene to include SNPs in regulatory regions. A total of 64557 SNPs were mapped to the 197 pathways including 4572 unique genes. The median number of genes in a pathway was 47 with the first and third quartiles being 27 and 76, while the median number of SNPs in a gene was 8 with the first and third quartiles being 4 and 17 respectively. We employed a stage-wise permutation strategy for both aSPUpath and GRASS tests: we first performed 5000 permutations for all pathways, and then increased to 100,000 permutations for those pathways with p-values < 0.01 in the first stage. We set the significance threshold at 0.00025 to control the family-wise error rate (FWER) at 0.05 based on the Bonferroni correction for 197 pathways.

Figure 6 shows the histograms of the p-values across the 197 KEGG pathways by the new method and GRASS; their distributions were similar, though GRASS gave a larger number of more significant p-values. Overall, the two methods gave similar and complementary results: although many common pathways were identified to be significant by both methods, each also detected some unique pathways. For example, at the significance threshold of 0.00025, aSPUpath and GRASS identified 18 and 35 significant pathways respectively, among which 11 were common. The Spearman's rank correlation coefficient between the p-values of the two methods was 0.65.

Table 4 shows 24 KEGG pathways with p-values less than .00001 by either method, i.e., none of the permutated test statistics exceeded the observed one based on 100,000 permutations. Interestingly, five 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 24 pathways. While three of them had p-values less than 0.00001 by both methods, two pathways, namely, NOD-like receptor signaling pathway (hsa04621) and Chemokine signaling pathway (hsa04062), had p-values < 0.00001 only by aSPUpath, but were not significant by GRASS (p-values > 0.00025). It is noteworthy that SNPs in the NOD2 gene in the NOD-like receptor signaling pathway were the first to be identified to be associated with CD and confer the highest risk for CD development among all CD-susceptibility SNPs discovered thus far (Jostins et al, 2012; Strober et al, 2014). The NOD-like receptor signaling pathway does not only include NOD2, but also several other CD-associated genes, including TNF, CCL2 and CCL7, making it one of the most well understood pathways underlying CD susceptibility (Billmann-Born et al, 2011). The data application here demonstrates that our proposed aSPUpath test is a competitive complementary approach to the GRASS test.

5. Discussion

We have proposed a new adaptive test for pathway analysis of genetic SNP data as arising in GWAS (Wang et al 2010; Fridley and Biernacka 2011). Since any pathway analysis involves multiple genes, each containing multiple SNPs, it is desirable to apply a test that can maintain high power with a large number of non-associated SNPs (or genes) and many only weakly associated SNPs (or genes), an ideal case for our proposed test. On the other hand, since the genes in a pathway may contain different numbers of SNPs, to avoid undue influence from a large (or small) gene, we modify the tests to take account of varying gene lengths. Our proposed test introduces two parameters (γ and γ_G) to achieve the objective. For example, if there are only few causal genes each containing many associated SNPs (e.g. due to LD), a large value of γ and a small value of γ_G would yield a more powerful test; since the truth is unknown, we use data to adaptively estimate their optimal values. The adaptivity of the proposed test at the gene level and/or at the SNP level is missing from many existing tests for pathway or SNP set analysis, such as the SSU and SKAT tests. As supported by our numerical examples, the proposed test can gain power in many situations and serve as a tool complementary to existing methods like GRASS.

Our proposed test is general and applicable to CVs or RVs. It may be modified, e.g. through some weighting on SNPs, for analysis of both CVs and RVs, as shown for the SSU test in Basu and Pan (2011). In addition, we can also introduce some weights at the gene and/or

SNP level to incorporate biological knowledge on which genes or SNPs are more likely to be causal. We have focused on testing on a single pathway; an alternative is to take account of possible overlapping or hierarchical structures of some pathways as discussed in Schaid et al (2012). These topics warranty future investigation.

Finally, we note that our proposed approach is in the category of “self-contained tests”, not “competitive tests”, since the null hypothesis to be tested here fits the former better than the latter: we are interested in detecting any pathways with any SNPs associated with a trait, not in detecting ones that are over-enriched with associated SNPs. Furthermore, as argued by Goeman and Buhlmann (2007), the same test on the former H_0 is necessarily more powerful than on the latter H_0. Our goal also differs from that of Newton et al (2012), which goes beyond only identifying significant pathways, but also aims to uncover the common theme shared among the identified significant pathways.

R code will be posted on our web site at
`http://www.biostat.umn.edu/~weip/prog.html`.

Acknowledgment

This research was supported by NIH grant R01HL116720; WP was also supported by NIH grants R01GM113250, R01HL105397 and R01GM081535, PW by R01CA169122 and R21HL126032. “This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.”

REFERENCES

Billmann-Born S, Lipinski S, Bck J, Till A, Rosenstiel P, Schreiber S (2011) The complex interplay of NOD-like receptors and the autophagy machinery in the pathophysiology of Crohn disease. *Eur J Cell Biol*, **90**, 593-602.

Basu S, Pan W (2011) Comparison of statistical tests for association with rare variants. *Genetic Epidemiology*, **35**, 606-619.

Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, Peters U, Hsu L. (2010). Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *American Journal of Human Genetics*, **86**, 860-871.

Chen X, Wang L, Hu B, Guo M, Barnard J, Zhu X. (2010b). Pathway-based analysis for genome-wide association studies using supervised principal components. *Genet Epidemiol*, **34**, 716-724.

Deelen J, Uh H-W, Monajemi R, van Heemst D, Thijssen PE, Bohringer S, van den Akker EB, de Craen AJM, Rivadeneira F, Uitterlinden AG, Westendorp RGL, Goeman JJ, Slagboom PE, Houwing-Duistermaat JJ, Beekman M (2013). Gene set analysis of GWAS data for human longevity highlights the relevance of the insulin/IGF-1 signaling and telomere maintenance pathways. *Age (Dordr)*, **35**, 235-249.

Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, *et al* (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics*, **42**, 1118-1125.

Fridley BL, Biernacka JM (2011). Gene set analysis of SNP data: benefits, challenges, and future directions. *Eur J Hum Genet*, **19**, 837-843.

Goeman JJ, van de Geer S, de Kort F, van Houwelingen HC (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20:93-99.

Goeman JJ, van de Geer S, van Houwelingen HC (2006) Testing against a high dimensional alternative. *J R Stat Soc B* 68:477-493.

Goeman JJ, Buhlmann P (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980-987.

Jostins L, Ripke S, Weersma RK 2012 Hostmicrobe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, **491**, 119-124.

Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*, **38**, D355-360.

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP (2008) A powerful and flexible multilocus association test for quantitative traits, *Am. J. Hum. Genet.* 82:386-397.

Lin, D.Y., Tang, Z.Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet*, **89**, 354-367.

Luo L, Peng G, Zhu Y, Dong H, Amos CI, Xiong M (2010). Genome-wide gene and pathway analysis. *European Journal of Human Genetics*, **18**, 1045-1053.

Newton MA, Quintana FA, den Boon JA, Sengupta S, Ahlquist P (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Annals of Applied Statistics*, **1**, 85-106.

Newton MA, He Q, Kendziorski C (2012). A model-based analysis to infer the functional content of a gene list. *Statistical Applications in Genetics and Molecular Biology*, **11**, Article 9.

Pan W (2009) Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic Epidemiology*, **33**, 497-507.

Pan W (2011) Relationship between Genomic Distance-Based Regression and Kernel Machine Regression for Multi-marker Association Testing. *Genetic Epidemiology*, **35**, 211-216.

Pan, W., Kim, J., Zhang, Y., Shen, X., Wei, P. 2014. A powerful and adaptive association test for rare variants. *Genetics*, **197**, 1081-1095.

Purcell SM, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, and Daly MJ et al. (2007). Plink: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, **81**, 559-575.

Schaid DJ, Sinnwell JP, Jenkins GD, McDonnell SK, Ingle JN, Kubo M, Goss PE, Costantino JP, Wickerham DL, Weinshilboum RM (2012). Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies. *Genet Epidemiol*, **36**, 3-16.

Strober W, Asano N, Fuss I, Kitani A, Watanabe T (2014). Cellular and molecular mechanisms underlying NOD2 risk-associated polymorphisms in Crohns disease. *Immunological Reviews*, **260**, 249260.

The Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609-615.

The Welcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661-678.

Tzeng JY, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, Worrall BB, Hsu FC, Thomas DC, Sullivan PF (2011). Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *Am J Hum Genet*, **89**, 277-288.

Verschuren JJW, Trompet S, Sampietro ML, Heijmans BT, Koch W, Kastrati A, Houwing-Duistermaat JJ, Slagboom PE, Quax PHA, Jukema JW (2013). Pathway Analysis Using Genome-Wide Association Study Data for Coronary Restenosis – A Potential Role for the PARVB Gene. *PLoS One*, **8**, e70676.

Wang T, Elston RC (2007) Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am J Hum Genet* 80:353-360.

Wang K, Abbott D (2007). A principal components regression approach to multilocus genetic association studies. *Genetic Epidemiology*, **32**, 108-118.

Wang K, Li M, Bucan M (2007). Pathway-based approaches for analysis of genome-wide association studies. *Am J Hum Genet*, **81**, 1278-1283.

Wang K, Li M, Hakonarson H (2010). Analysing biological pathways in genome-wide association studies. *Nature Rev Genet*, **11**, 843-854.

Wang L, Jia P, Wolfinger RD, Chen X, Zhao Z (2011). Gene set analysis of genome-wide association studies: Methodological issues and perspectives *Genomics*, **98**, 1-8.

Wei P, Tang H, Li D (2012). Insights into pancreatic cancer etiology from pathway analysis of genome-wide association study data. *PLoS ONE*, **7**, e46887.

Wessel J, Schork NJ (2006) Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum Genet* 79:792-806.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X (2010) Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *Am J Hum Genet* 86:929-942.

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

Zhang Q, Irvin MR, Arnett DK, Province MA, Borecki I (2011). A data-driven method for identifying rare variants with heterogeneous trait effects. *Genetic Epidemiology*, **35**, 679-685.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

Submitted January 9, 2015

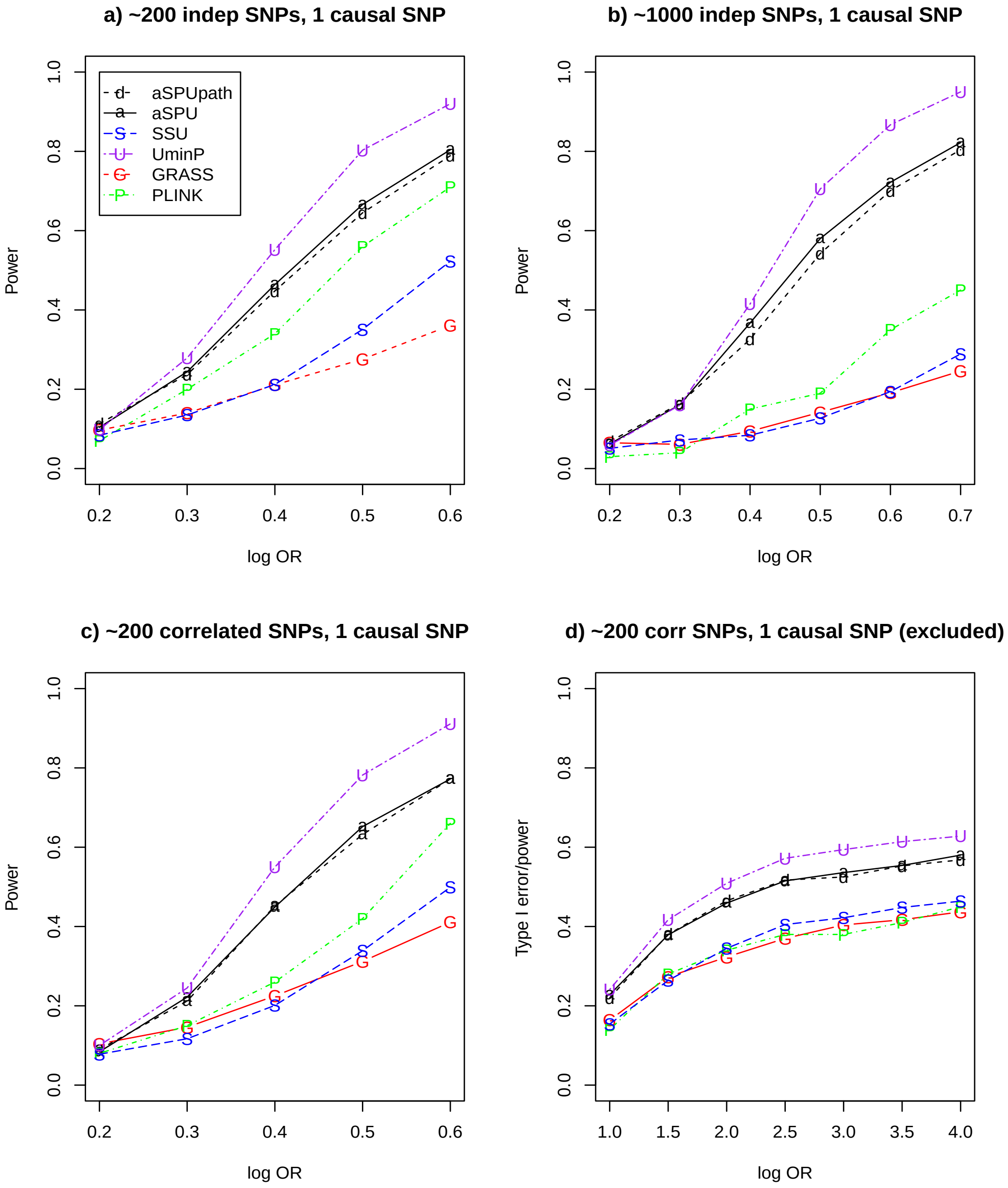

Figure 1. Empirical power for simulation set-up B with a pathway containing 20 genes, one of which was causal and included 1 causal SNP among 1-20 SNPs (panels a, c and d) or among 3-100 SNPs (panel b).

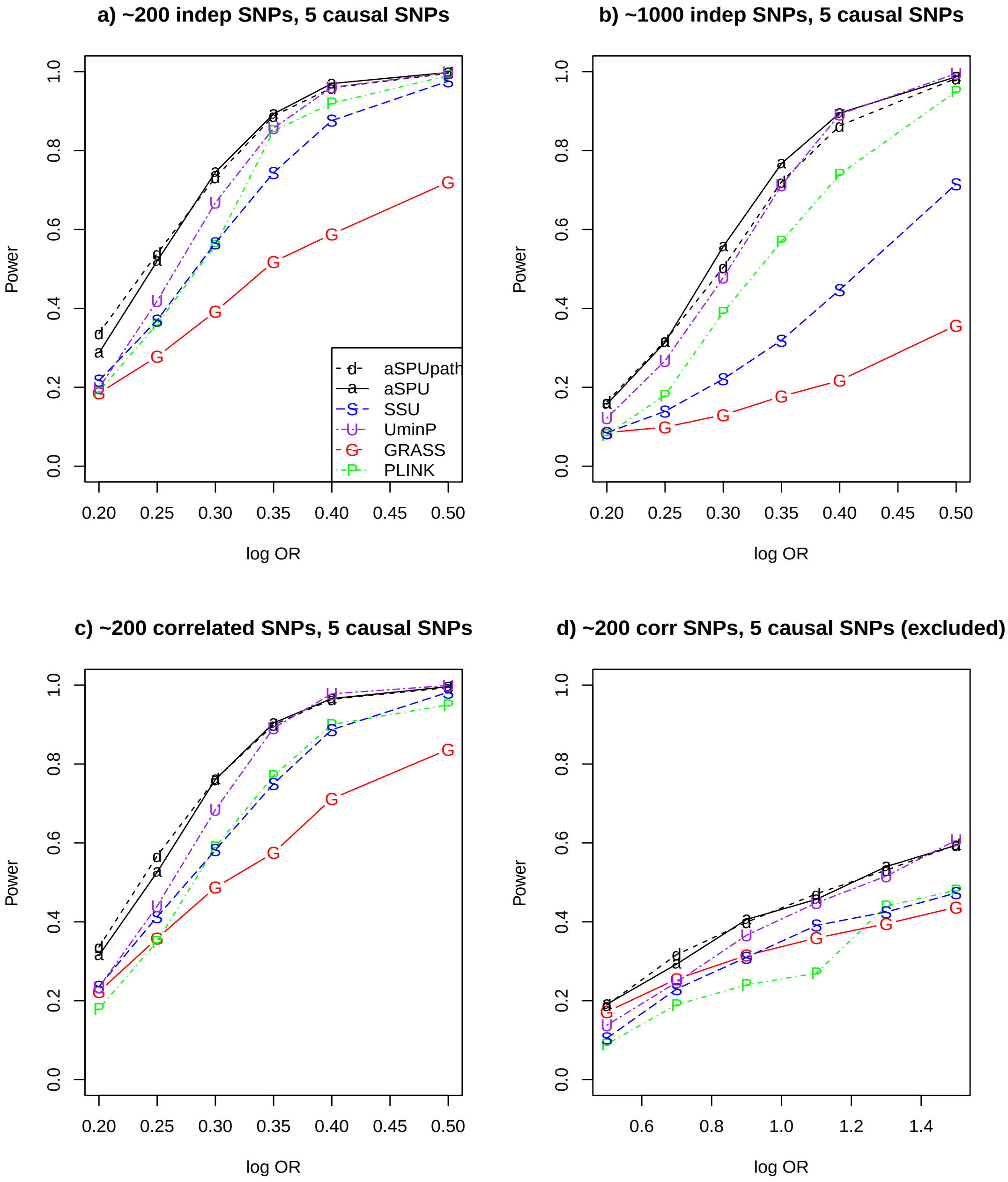

Figure 2. Empirical power for simulation set-up C with a pathway containing 20 genes, 5 of which were causal with each including 1 causal SNP among 1-20 SNPs (panels a, c and d) or among 3-100 SNPs (panel b).

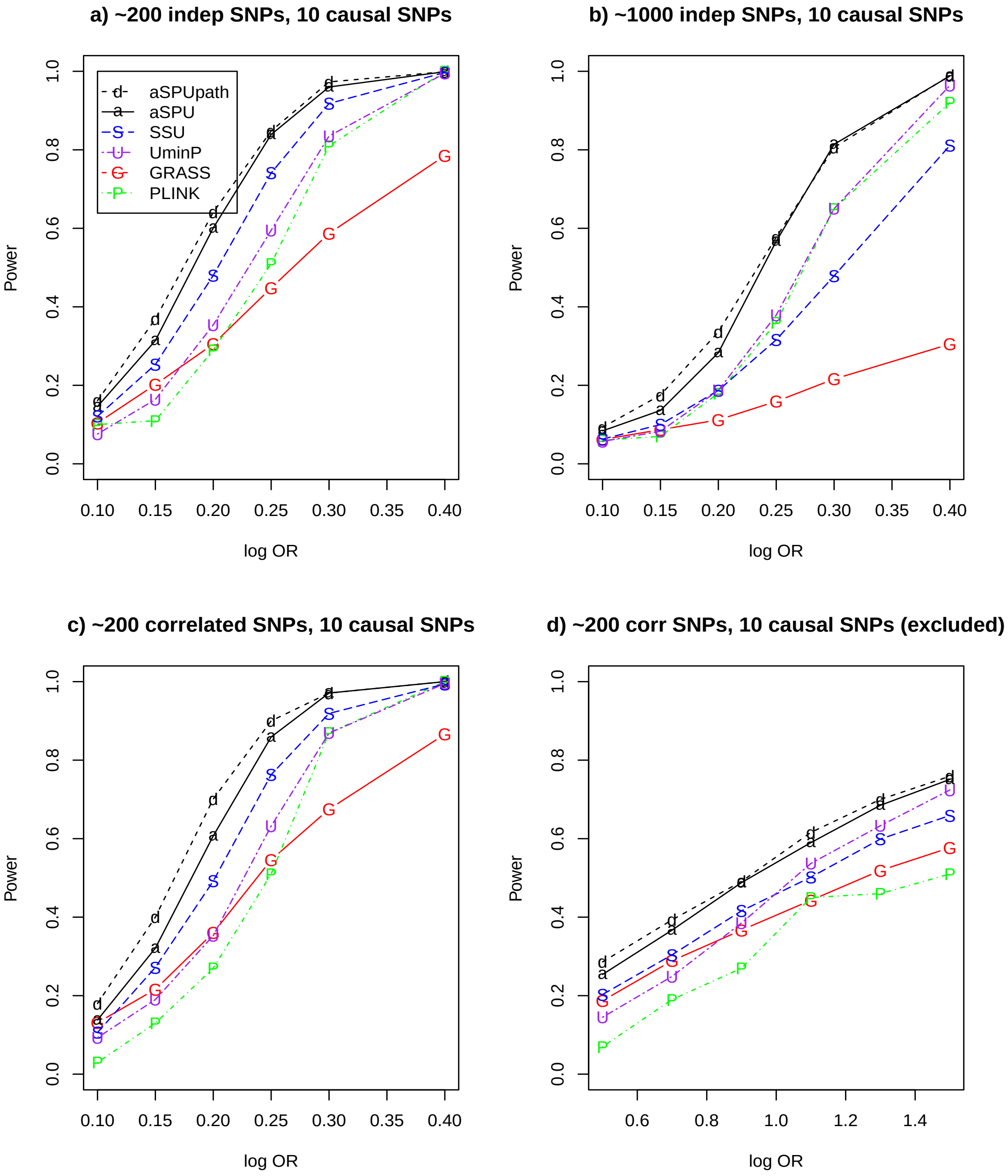

Figure 3. Empirical power for simulation set-up D with a pathway containing 20 genes, 10 of which were causal with each including 1 causal SNP among 1-20 SNPs (panels a, c and d) or among 3-100 SNPs (panel b).

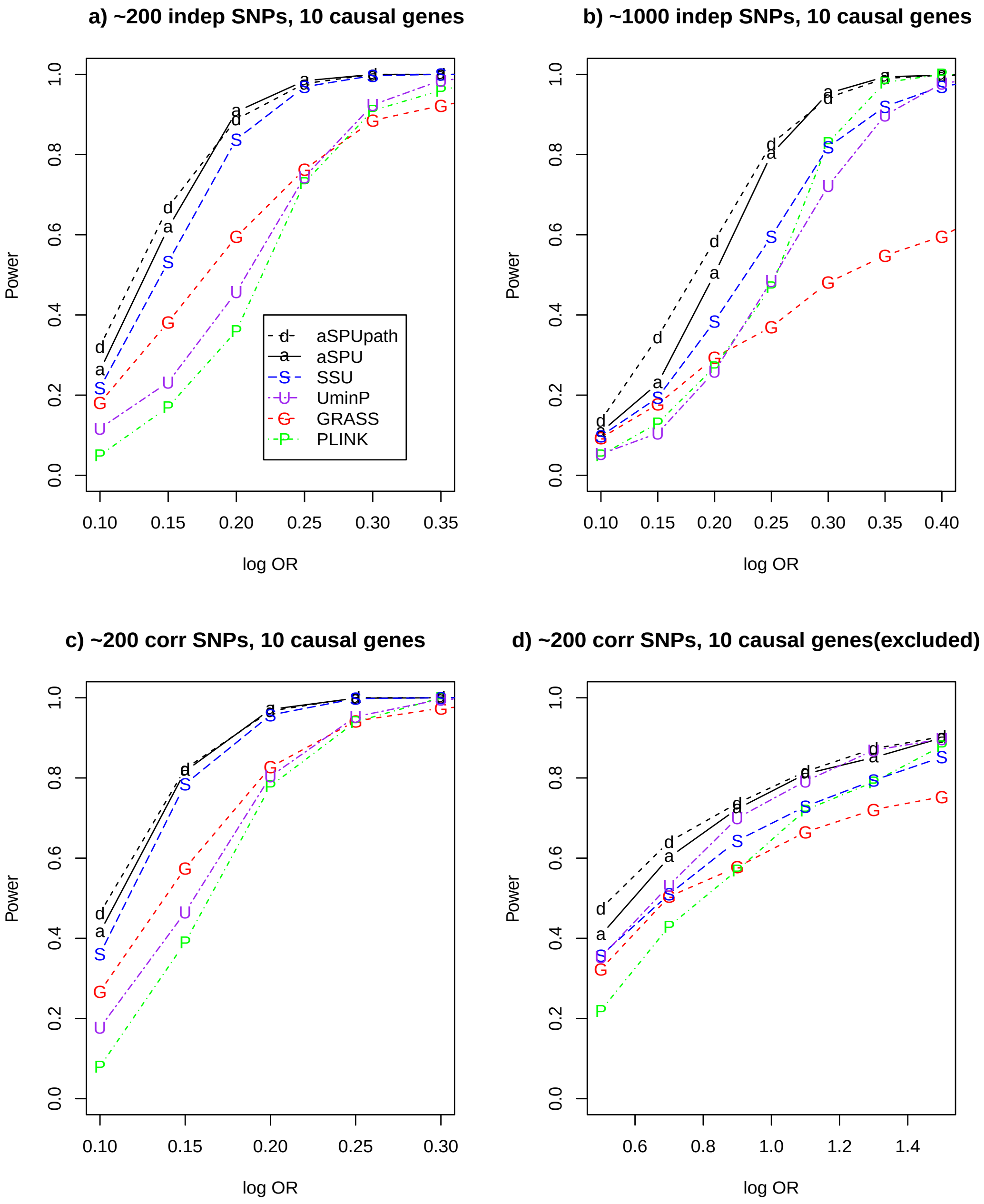

Figure 4. Empirical power for simulation set-up D' with a pathway containing 20 genes, ten of which were causal with each including 1-3 causal SNPs among 1-20 SNPs (panels a, c and d) or among 3-100 SNPs (panel b).

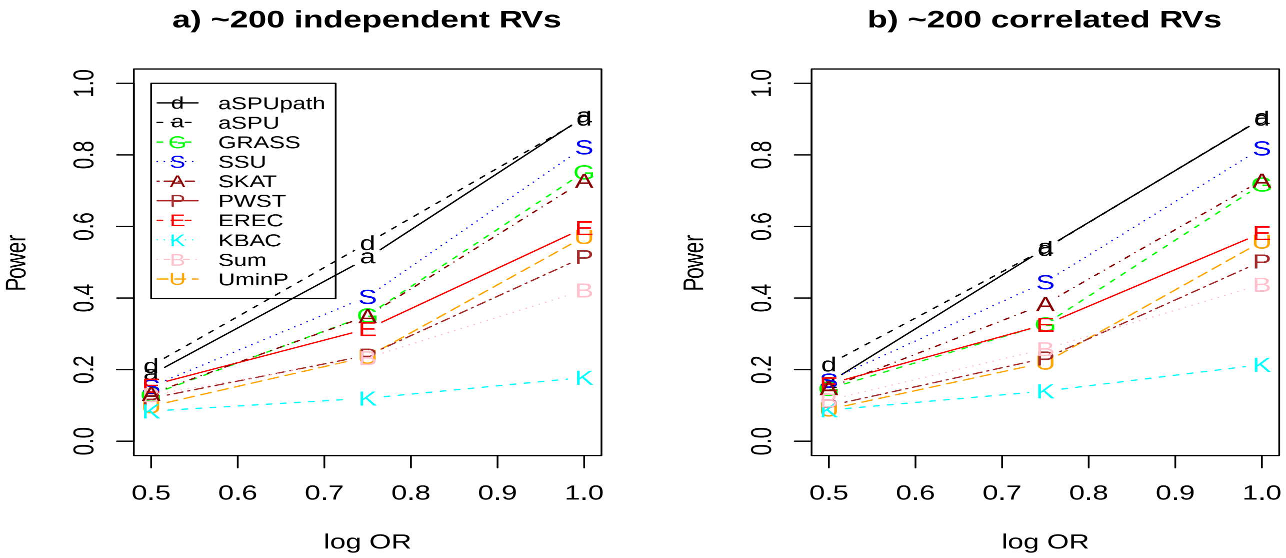

Figure 5. Empirical power for RVs in simulation set-up D2 with a pathway containing 20 genes, 10 of which were causal with each including 1 causal RV among 1-20 RVs.

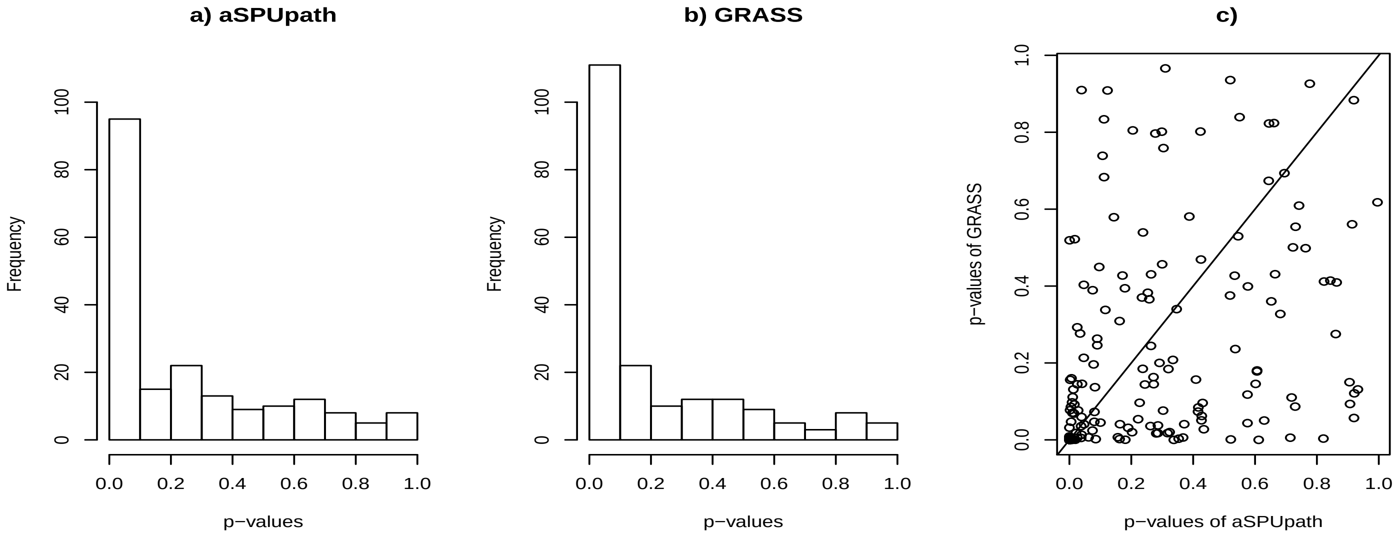

Figure 6. Distributions of the p-values from a) aSPUpath and b) GRASS and c) their comparison for the WTCCC CD data.

Table 1
Empirical Type I error rates of the tests for CVs.

Set-up	aSPUpath	GRASS	Plink	aSPU	SSU	UminP
200 indep SNPs	.055	.057	.02	.053	.046	.057
1000 indep SNPs	.048	.067	.03	.050	.052	.040
200 corr SNPs	.054	.064	.05	.048	.040	.062

Table 2

Empirical Type I error ($\log OR = 0$ *and power (*$\log OR \neq 0$*) of various tests for about 200 correlated SNPs in a 20-gene pathway (set-up D).*

	aSPUpath										
log OR	$w_G = 1$	$\delta = 0$	$\delta = .1$	$\delta = .2$	$\delta = .3$	$\delta = .4$	GRASS	Plink	aSPU	SSU	UminP
0	.054	.052	.051	.050	.048	.044	.064	.05	.059	.049	.062
0.15	**.400**	.397	.430	.468	.489	**.517**	.216	.13	.321	.272	.190
0.2	**.701**	.656	.713	.747	.769	**.791**	.360	.27	.607	.492	.353
0.25	**.900**	.873	.907	.926	.931	**.936**	.546	.51	.859	.763	.632

Table 3
Empirical Type I error rates of the tests for RVs.

Set-up	aSPUpath	GRASS	aSPU	Sum	SSU	UminP	SKAT	KBAC	PWST	EREC
200 indep SNPs	.059	.058	.060	.048	.051	.068	.050	.054	.053	.048
200 corr SNPs	.058	.065	.047	.051	.060	.045	.058	.048	.054	.052

Table 4

Results of the WTCCC CD GWAS data application: KEGG pathways with p-values < 0.00001 *by either aSPUpath or GRASS. Positive control pathways are in bold.*

				p-values	
KEGG ID	Pathway names	# of genes	# of SNPs	aSPUpath	GRASS
hsa04060	**Cytokine-cytokine receptor interaction**	247	2506	< .00001	< .00001
hsa04630	**Jak-STAT signaling pathway**	145	1410	< .00001	< .00001
hsa04660	**T cell receptor signaling pathway**	105	1373	< .00001	< .00001
hsa04310	Wnt signaling pathway	143	2087	< .00001	< .00001
hsa05310	Asthma	27	271	< .00001	< .00001
hsa05330	Allograft rejection	34	466	< .00001	< .00001
hsa05414	Dilated cardiomyopathy (DCM)	89	2605	< .00001	< .00001
hsa05416	Viral myocarditis	67	1263	< .00001	< .00001
hsa04972	Pancreatic secretion	93	2187	< .00001	.00003
hsa04621	**NOD-like receptor signaling pathway**	57	502	< .00001	.00542
hsa04062	**Chemokine signaling pathway**	174	2714	< .00001	.00061
hsa04810	Regulation of actin cytoskeleton	201	3347	< .00001	.00108
hsa05131	Shigellosis	60	784	< .00001	.00434
hsa00230	Purine metabolism	154	2810	.00759	< .00001
hsa04144	Endocytosis	180	2575	.00190	< .00001
hsa04145	Phagosome	136	1469	.00101	< .00001
hsa04270	Vascular smooth muscle contraction	113	2887	.00025	< .00001
hsa04350	TGF-beta signaling pathway	82	831	.00080	< .00001
hsa04514	Cell adhesion molecules (CAMs)	122	3312	.00120	< .00001
hsa04612	Antigen processing and presentation	63	543	.00129	< .00001
hsa04650	Natural killer cell mediated cytotoxicity	124	1464	.00199	< .00001
hsa04672	Intestinal immune network for IgA production	45	393	.00073	< .00001
hsa04940	Type I diabetes mellitus	39	714	.00031	< .00001
hsa05332	Graft-versus-host disease	33	440	.00036	< .00001