

A Powerful Pathway-Based Genetic Association Test

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*

October 10, 2011

Confidential Draft

Correspondence author: Wei Pan

Telephone: (612) 626-2705

Fax: (612) 626-0660

Email: weip@biostat.umn.edu

Address: MMC 303, A460 Mayo,

Division of Biostatistics, School of Public Health,

University of Minnesota,

Minneapolis, Minnesota 55455-0392, USA.

1. INTRODUCTION

We only consider self-contained approaches, not competitive approaches, since 1) the former approaches are in general more powerful than the latter (Goeman and Buhlmann (2007) and 2) for the purpose of identifying any disease-associated SNPs, the null hypothesis of the former is more relevant.

2. METHODS

2.1. Data and notation

We consider the case-control study design, though the methods may be extended to other study designs, e.g. with a quantitative or survival trait. Suppose that for subject i , $i = 1, \dots, n$, $Y_i = 0$ or 1 is a binary trait, e.g. an indicator of disease, and $X_i = (X_{i1}, \dots, X_{ik})'$ is the genotype at k SNP loci. 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:

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

We'd like to test the null hypothesis $H_0: \beta = (\beta_1, \dots, \beta_k)' = 0$; that is, there is no association between any SNPs and the trait under H_0 .

In the current context, even after selecting some nearly independent SNPs (without using trait Y_i 's), we have a large k (and a large n). If we apply the standard score test (or its asymptotically equivalent Wald or likelihood ratio test), the power will be low. In fact, as shown theoretically in Fan (1996) and to be shown empirically later, as the dimension k increases, the power of the score test may diminish, tending to the Type I error rate α . Similarly, if we have many small $|\beta_j| \neq 0$, the most popular

single SNP-based UminP test in GWAS may be also low-powered. As a response, a polygenic test was recently proposed by the ISC (2009).

2.2. An data-adaptive test: aSPU

Our primary goal is to construct a test that is data-adaptive while maintaining the advantages of up-weighting/selecting informative SNPs.

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

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

The SPU tests cover the Sum and SSU tests as two special cases with a corresponding $\gamma = 1$ and $\gamma = 2$ respectively. Importantly, as $\gamma \rightarrow \infty$ (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:

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


Based on the standard n -asymptotics (i.e. as $n \rightarrow \infty$), we know that under H_0 , the score vector U has an asymptotic Normal distribution $N(0, V)$. Hence, in theory, we can derive the distribution of the T_{SPU} , which however may not be easy to calculate. A permutation method was proposed by Pan et al (2011), which however is not applicable in the presence of covariates; here we propose using a **parametric bootstrap** (Efron and Tibshirani 1992?; Lin and Tang 2011). Details:under H_0 , we **simulate** $U^{(b)}$ from its null distribution $N(0, V)$, then calculate the null statistic $T_{SPU}^{(b)} = T_{SPU}(U^{(b)})$; based on B simulations, we calculate the **p-value as** $\sum_{b=1}^B I(|T_{SPU}^{(b)}| > |T_{SPU}|)/B$.

As discussed in Pan (2008), there is no unfirmly 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 (2011) 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. $\Gamma = \{1, 2, 3, \dots, 8\}$ 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:

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

Of course, T_{aSPU} is no longer a genuine p-value; we recourse to 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 $b = 1, 2, \dots, B$.

 We then calculate the corresponding SPU test statistics $T_{SPU(\gamma)}^{(b)}$ and their p-values $p_\gamma^{(b)} = \sum_{b_1 \neq b} I(T_{SPU(\gamma)}^{(b_1)} > T_{SPU(\gamma)}^{(b)}) / (B - 1)$. Thus, we have $T_{aSPU}^{(b)} = \min_{\gamma \in \Gamma} p_\gamma^{(b)}$, and the final p-value of the aSPU test $p_{aSPU} = \sum_{b=1}^B I(T_{aSPU}^{(b)} < T_{aSPU}) / B$.

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). 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, while each gene might contain 1-20 SNPs, or 3-100 SNPs; there was only 1 causal SNP inside 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 $Z = (Z_1, \dots, Z_k)'$ from a multivariate Normal distribution

K : # of snps
 \bar{t} : # of λ

with a first-order auto-regressive (AR1) covariance structure: $\text{Corr}(Z_i, Z_j) = \rho^{|i-j|}$ 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, k , 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: $X_i = (X_{i1}, \dots, X_{ik})'$ for subject i . Fourth, for a non-null case, the first SNP inside the first $k_1 = 1$ or 5 or 10 genes was chosen to be causal with their corresponding $\beta_j = \log OR \neq 0$, while all other $\beta_j = 0$; for a 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 similar to SPU(∞) (data not shown).

We also considered single gene-based approaches and those based on dimension reduction. For single genebased approaches, we considered applying SPU and aSPU tests to each gene, then using the minP to combine their p-value. It is easy to see that

$$\left\{ \begin{array}{l} \text{SPU} \\ + \\ \text{aSPU} \end{array} \right. \quad k=1, \dots, g, \infty$$

the pathway-based SPU(∞) and single genebased SPU(∞) are almost the same. For

1) dimension reduction, as in GRASS, we extracted the first few principal components (PCs) accounting for at least 95% of total variation of genotypes, then we apply the SPU and aSPU tests to these PCs. It is perhaps surprising, but not difficult, to see that applying the SSU=SPU(2) test to the original genotypes or the PCs gave almost the same result, as shown below. Suppose that X is the original genotype matrix with a singular value decomposition: $X' = W\Lambda V'$. Its first L PCs are

need to clean up notation:

$$P' = W_L' X' = V \Lambda_L' = V I_{L \times k} \Lambda',$$

where $I_{L \times k}$ is a rectangular identity matrix. Now we can compare the SSU statistics when applied to X and P :

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

But for other $\gamma \neq 2$, we would expect that, in general, $SSU(\gamma)$ would give different results when applied to the original genotype X and PCs P .

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 pathway-based aSPU with GRASS, PLINK

For comparison, we also included SSU=SPU(2) and SPU(∞), which is essentially equivalent to UminP (marked as Max in the following plots).

Fig 1 or set-up B: since there was only one causal SNP, we'd reason that UminP should be most powerful, which was confirmed. In all cases the aSPU was the second most powerful. a) with about 200 independent SNPs, PLINK was third most powerful, followed by the SSU and then GRASS. b) with about 1000 independent

SNPs, $\text{SPU}(\infty)$ and aSPU showed even a more striking advantage over the other three tests, suggesting the former two's (and the latter three's) robustness (and lack of robustness) to increasing number of SNPs. In particular, the performance of SSU deteriorated with its power close to that of GRASS. c) with about 200 correlated SNPs (with the causal SNP included), the power trend was similar to that with 200 independent SNPs. d) with about 200 correlated SNPs with the causal SNP excluded, again we found that $\text{SPU}(\infty)$ and aSPU were the top two winners, while the other three tests had similar power.

Fig 2 for set-up C: with 5 causal SNPs, now it appears that the aSPU test had a slight edge over UminP since the latter uses only the single SNP with the strongest signal while ignoring the signals from other 4 causal SNPs. Again the aSPU and UminP tests were consistent winners. However, differing from set-up B, we notice that the SSU test and PLINK performed similarly in a) and c) while one was more powerful than the other in b) and d) respectively; in particular, in d) with the 5 causal SNPs excluded, GRASS could perform better than the other tests except aSPU when the causal effect size was small (and the power was low).

Fig 3 for set-up D: now with 10 causal SNPs, aSPU was the clear and sole winner; in particular, the aSPU could be much more powerful than UminP. In fact, for all three cases with only about 200 SNPs, the SSU test was more powerful than UminP, since the former combined information across multiple SNPs while the latter only used the most significant SNP. On the other hand, 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 SNPs since it did not down-weight enough the larger number of non-associated SNPs. With causal SNPs, GRASS could beat PLINK if the causal effect size was small (c) or the causal SNPs were excluded (d).

In summary, we found that the aSPU test was much more powerful than GRASS

and PLINK across all the cases.

3.2.2 Comparison of pathway-based aSPU with other approaches

Fig 4 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 adaptive choosing the tuning parameter γ to down-weight non-associated SNPs; however, since each PC is a linear combination of all the SNPs, a mixture of both associated and non-associated SNPs, hindering the ability of the aSPU test to select SNPs effectively.

Fig 5 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.

Fig 6 for set-up D: with 10 causal SNPs, the pathway-based aSPU test was by far 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 (with or without an adjustment for gene-specific mean and SD of the null statistic), but did not find it worked better than the simple PathaSPU discussed here.

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

path RV aSPU

3.3 Simulation results for RVs

We also did a simulation study to assess the performance of the proposed test for pathway analysis of RVs with sequencing data. We only considered 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=SPU(2) test that was shown by Basu and Pan (2011) to be competitive and closely related to several new 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 just appeared recent, 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 2, all the methods seem to have Type I error rates around the nominal level of 0.05.

As shown in Fig 7, the relative performance of the various tests did not strongly depend on whether there were within-gene correlations among the RVs. Clearly the aSPU test was the most powerful, followed by the SSU test, then by GRASS and EREC. 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 pooled association test); 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 (2011).

well

4. EXAMPLE

WTCCC data...

5. DISCUSSION

A summary of the results and main points...

R code will be posted on our web site at

<http://www.biostat.umn.edu/~weip/prog.html>.

ACKNOWLEDGMENTS

WP was supported by NIH grants R01GM081535, R01HL65462, R01HL105397 and R21DK089351.

REFERENCES

- Barrett JC, Fry B, Maller J, Daly MJ (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-265.
- Basu S, Pan W (2011) Comparison of Statistical Tests for Association with Rare Variants. To appear in *Genetic Epidemiology*.
- Chapman JM, Whittaker J (2008) Analysis of multiple SNPs in a candidate gene or region. *Genetic Epidemiology* **32**:560-566.
- Chen SX, Qin Y-L (2010) A two-sample test for high-dimensional data with applications to gene-set testing. *Ann Statist* **38**:808-835.
- Clayton D, Chapman J, Cooper J (2004) Use of unphased multilocus genotype data in indirect association studies. *Genet Epidemiol* **27**:415-428.

- Conneely KN, Boehnke M (2007). So many correlated tests, so little time! Rapid adjustment of p values for multiple correlated tests. *Am J Hum Genet* 81:1158-1168.
- Cox DR, Hinkley DV (1974) *Theoretical Statistics*, Chapman and Hall, London.
- Fan J (1996) Test of significance based on wavelet thresholding and Neyman's truncation. *JASA* 91:674-688.
- Faraway, JJ (1992) On the cost of data analysis. *J. Comp. Grap. Stat.*, 1:213-229.
- Fisher, RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Philos Trans R Soc Edinb* 52:399-433.
- 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, van Houwelingen HC (2011) Testing against a high-dimensional alternative in the generalized linear model: asymptotic type I error control. *Biometrika* 98:381-390.
- Gottesman II, Shields J (1967) A polygenic theory of schizophrenia. *Proc Natl Acad Sci USA* 58:199-205.
- Hindorff LA, Junkins HA, Hall PN, Mehta JP, and Manolio TA (2010) A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed October 31, 2010.
- 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, DY (2005) An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* 21:781-787.
- Liu D, Ghosh D, Lin X (2008) Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models, *BMC Bioinformatics* 9:292.

- Maher B (2008) Personal genomes: the case of the missing heritability. *Nature*, 456:18-21.
- 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* .
- Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 38:209-213.
- The International Schizophrenia Consortium (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460:748-752
- The Wellcome 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.
- 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.
- 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.

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

Set-up	Path-aSPU	Gene-aSPU	PC-aSPU	SSU	UminP	GRASS	PLINK
200 indep SNPs	.055	.061	.058	.046	.057	.057	.02
1000 indep SNPs	.048	.057	.041	.052	.040	.067	.03
200 corr SNPs	.054	.059	.042	.040	.062	.064	.05

Table 2: Empirical Type I error rates of the tests for RVs.

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

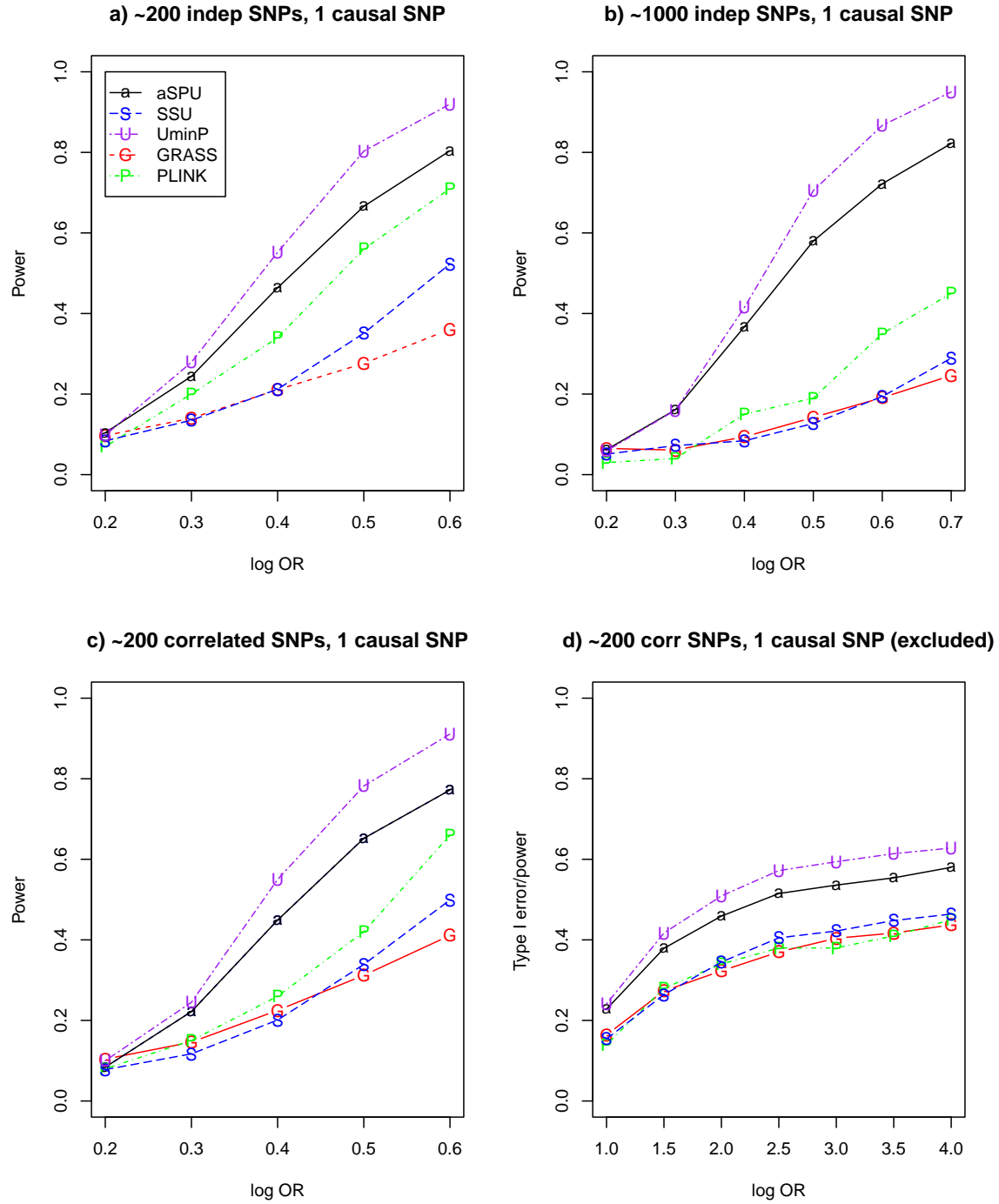


Figure 1: Empirical power for simulation set-up B with a pathway containing 20 genes, 1 of which was causal and included 1 causal SNP among 1-20 SNPs.

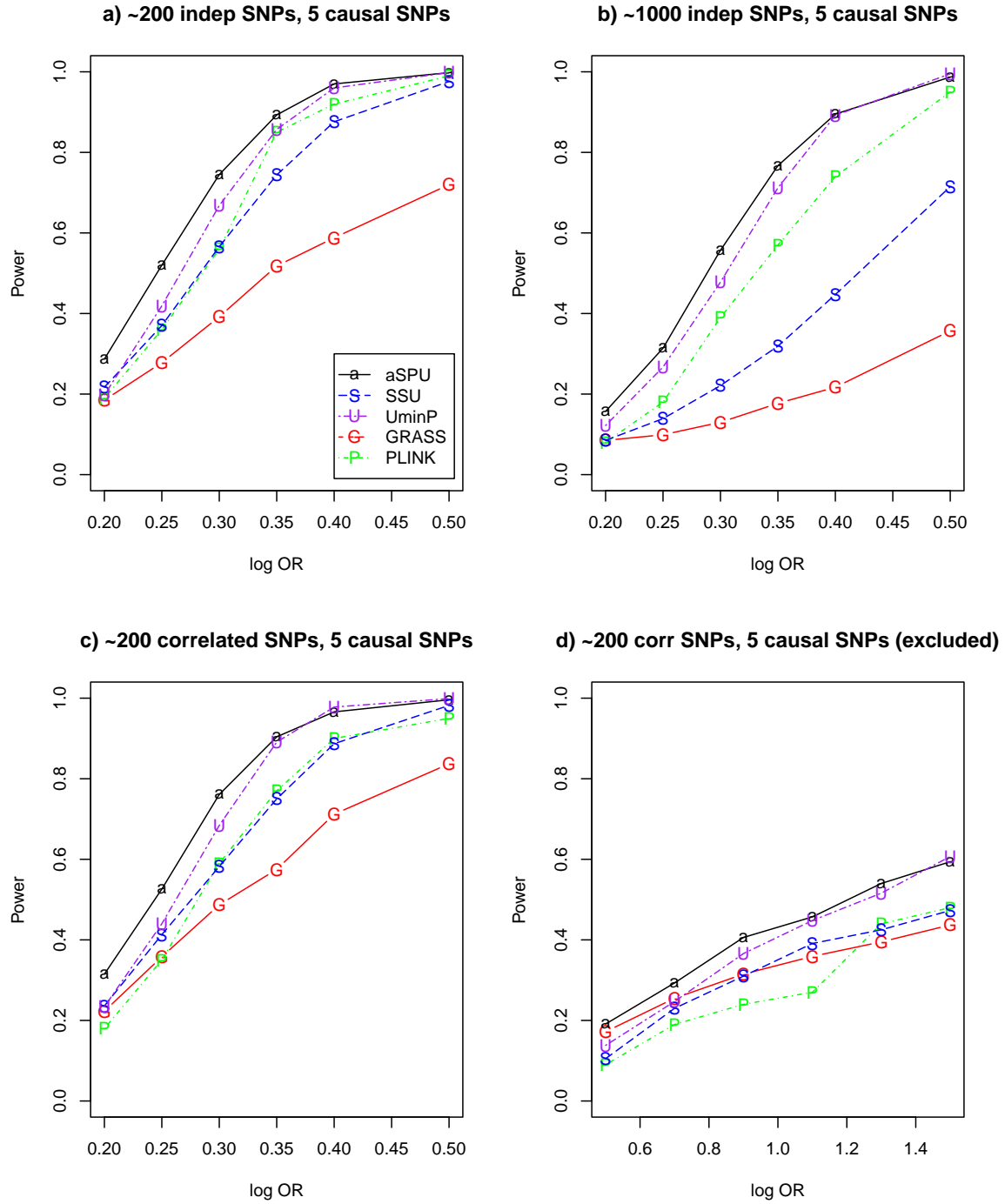


Figure 2: Empirical power for simulation set-up C with a pathway containing 20 genes, 5 of which were causal with each including 1 causal SNP among 1-20 SNPs.

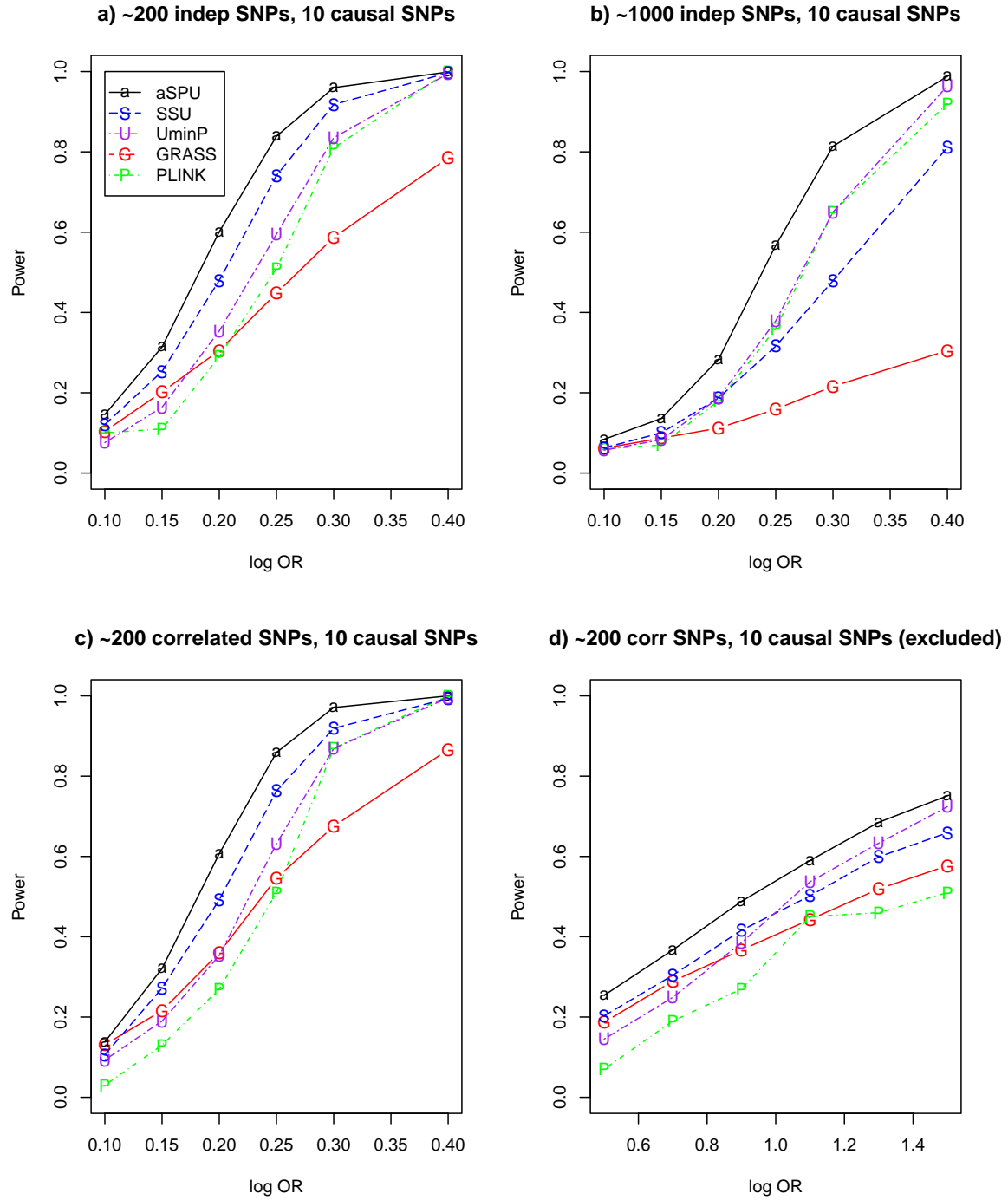


Figure 3: Empirical power for simulation set-up D with a pathway containing 20 genes, 10 of which were causal with each including 1 causal SNP among 1-20 SNPs.

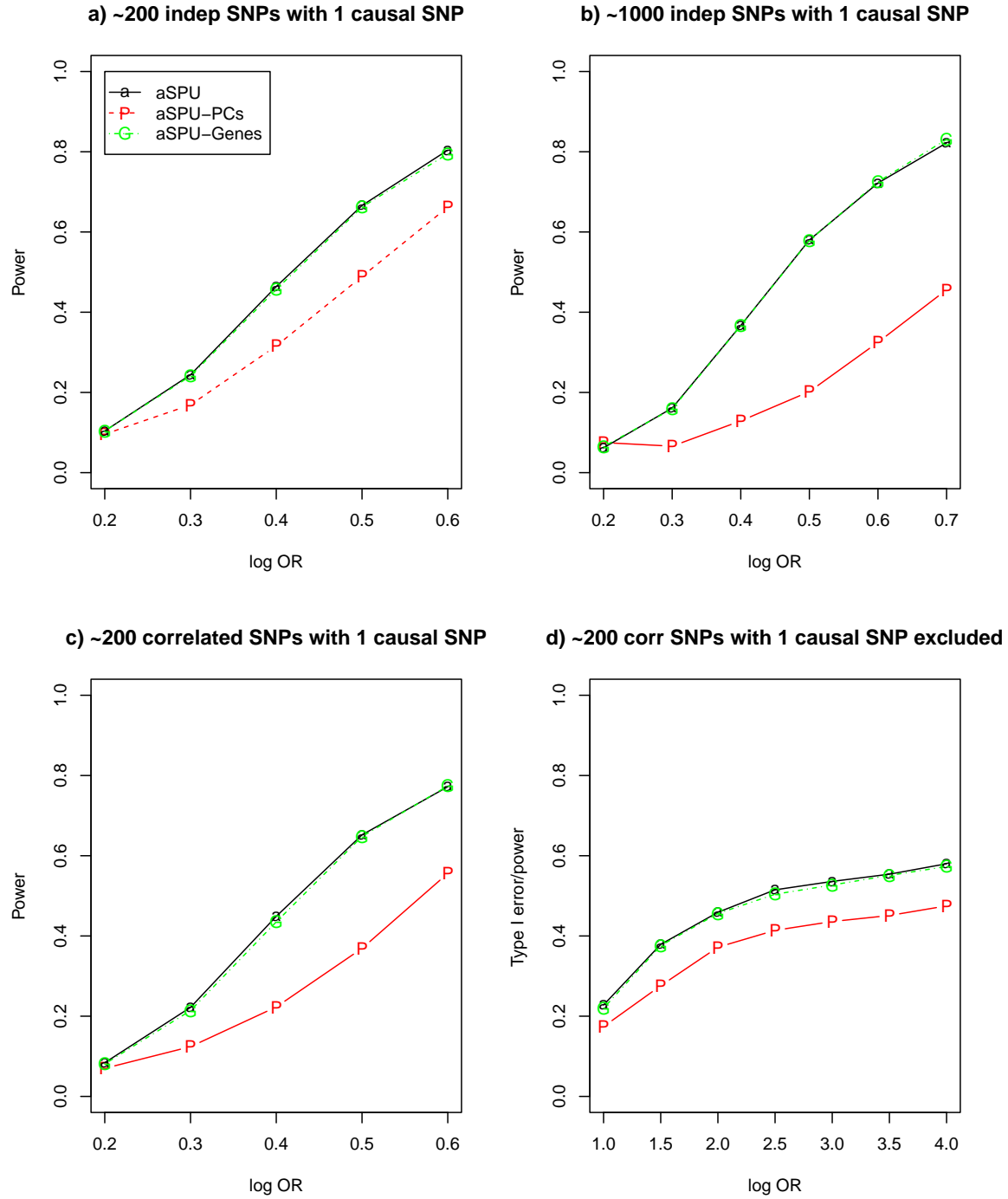


Figure 4: Empirical power for simulation set-up B with a pathway containing 20 genes, 1 of which was causal and included 1 causal SNP among 1-20 SNPs.

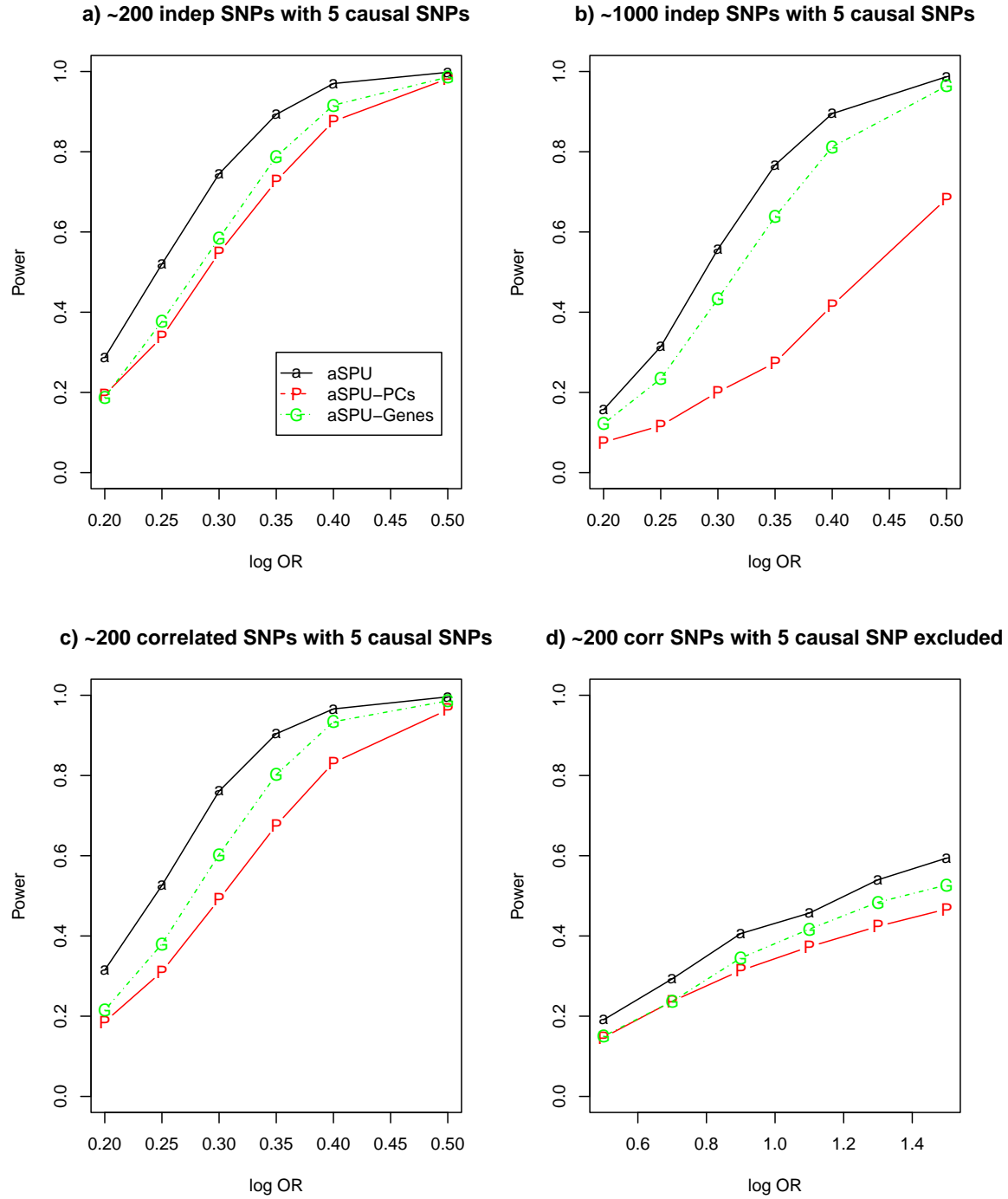


Figure 5: Empirical power for simulation set-up C with a pathway containing 20 genes, 5 of which were causal with each including 1 causal SNP among 1-20 SNPs.

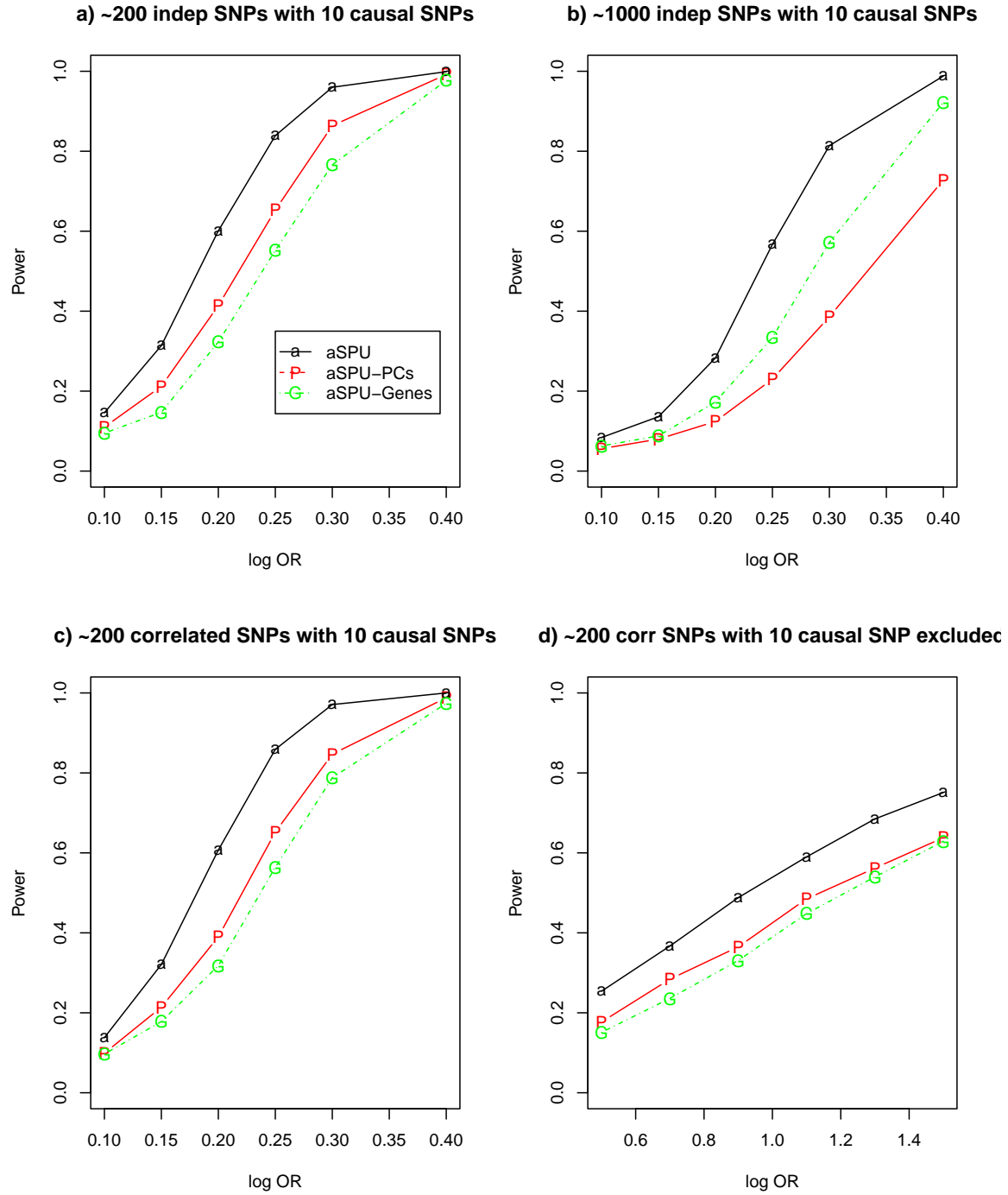


Figure 6: Empirical power for simulation set-up D with a pathway containing 20 genes, 10 of which were causal with each including 1 causal SNP among 1-20 SNPs.

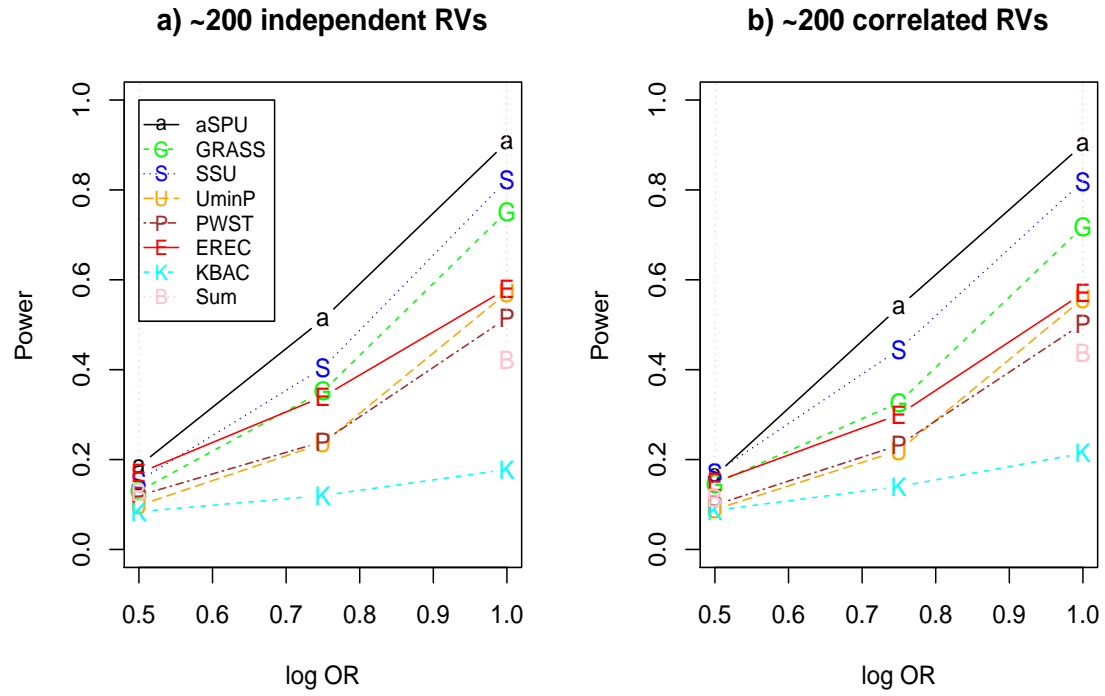


Figure 7: Empirical power for RVs in simulation set-up D2 with a pathway containing 20 genes, 10 of which were causal with each including 1 causal RV among 1-20 RVs.