Journal of Genetics (2022)101:12
https://doi.org/10.1007/s12041-021-01351-w

RESEARCH ARTICLE

An adaptive combination method for Cauchy variable based on optimal threshold

YUAN TANG, YAJING ZHOU* and YUNLONG BAO

Department of Statistics, School of Mathematical Sciences, Heilongjiang University, Harbin 150080, People's Republic of China
*For correspondence. E-mail: 2002099@hlju.edu.cn.

Received 31 October 2020; revised 14 October 2021; accepted 7 November 2021

Abstract. In sequence study, set-based analysis has been developed as a popular tool for analysing the association of a group rare variant with disease. However, most of the methods are sensitive to the genetic architecture. Besides, by directly combining the association signals of multiple markers within a genomic region can inevitably include a large proportion of noises. To address this issue, we extend the aggregated Cauchy association test (ACAT; Liu *et al.* 2019) and propose an adaptive Cauchy-variable combination method (ACC). Our proposed method combines Cauchy-variables, which are transformed from variant-level *P*-values in a given genomic region and adaptively truncate noises by choosing an optimal truncation threshold of the variant-level *P*-value that is determined by the data. Besides, the ACC method can use summary statistics obtained from open access database when the original genotype and phenotype data are unavailable. Extensive simulation studies and Genetic Analysis Workshop 19 real data analysis show that ACC is more powerful than the other comparative methods when only a small proportion of variants are causal, and ACC is robust to the varied genetic architecture.

Keywords. rare variants; association study; next-generation sequencing; GAW19 dataset.

Introduction

To date, genomewide association studies (GWAS) have identified a lot of common variants associated with human complex diseases (Visscher *et al.* 2012; Welter *et al.* 2014) which can explain only a small proportion of heritability (Eichler *et al.* 2010; Zuk *et al.* 2012). It has been suggested that some diseases may be caused by rare variants (Pritchard *et al.* 2002; Manolio *et al.* 2009) and now the focus of genetic studies is the identification of rare variants associated with disease. As the result of low frequencies of rare causal variants, the standard single-variant analysis is underpowered without a large-scale sample sizes. Therefore, set-based analysis by aggregating the moderate effect of each rare variant in a region becomes a popular tool for analysing the association of a group rare variant with human disease.

Over the past several years, the sequence kernel association test (SKAT) (Wu *et al.* 2011) and Burden tests (Madsen *et al.* 2009; Price *et al.* 2010) have been widely used in set-based analysis of rare variants. However, SKAT and Burden tests can loss power when only a small proportion of rare variants are causal variants (Donoho and Jin 2004; Barnett *et al.* 2017). Besides, many set-based analysis methods depend on the underlying genetic architecture, such as the effect size and the effect direction of each variant in a given region. In fact, the genetic architectures are always unknown in advance and can be different from one region to another region. One class of methods have emerged as a popular tool for analysing rare variants in sequencing studies, i.e., by combining the corresponding χ^2-statistics or the *P*-value for each rare variant in the given region. Liu *et al.* (2019) proposed an aggregated Cauchy association test, a powerful and computationally efficient *P*-value combination method. The ACAT method first transforms variant-level *P*-values into Cauchy-variables, takes the weighted sum of them as the test statistic and then evaluates the significance analytically. ACAT is powerful in the presence of a small proportion of causal variants in a genomic region. However, variants associated with the disease are sparse in a genomic region, most variants in a region have no influence on the disease. Directly combining the signals of variants inevitably brings noise. Thus, it is necessary to look for a method that can exclude noise caused by noncausal variants and can be robust to varied genetic architecture.

Published online: 19 January 2022

To guard against the noise caused by neutral variants, we extend the ACAT method in this paper and propose an adaptively combines Cauchy-variables (ACC). The ACC method transforms variant-level *P*-values into Cauchy-variables and truncate noises by choosing an optimal truncation threshold upon the variant-level *P*-values from a multiple candidate truncation thresholds (0.11, 0.12... 0.20). Variants with *P*-values larger than a threshold will be truncated, and variants highly associated with disease can be retained. Hence ACC can decrease the effect of noncausal variants and increase the power of a test. Besides, a great number of summary SNP-level statistics such as *P*-values and effect sizes (here after referred to as GWAS summary statistics) are now available for different traits and diseases in open-access databases (Pasaniuc and Price 2017), so the ACC method can use GWAS summary statistics obtained from open access database when the original genotype and phenotype data are unavailable. Moreover, the *P*-values of the significance scores under different truncation threshold can be well approximated by Cauchy distribution without estimating and accounting for the correlation among variant-level *P*-values. The extensive simulation studies show that ACC method is a powerful and robust method. The type I error rates are within 95% confidence intervals. Finally, we illustrate our proposed ACC method by analysing the Genetic Analysis Workshop 19 data (GAW19), the results show ACC method can detect some genes associated with systolic blood pressure and diastolic blood pressure.

Method and materials

Suppose there is a sample of n unrelated individuals. The phenotype of the ith individual is denoted as Y_i (i = 1, 2… n), each individual has been sequenced in a genomic region with M rare variants. Denote $X_i = (X_{i1}, X_{i2}, \ldots, X_{iM})^T$ as the genotype score vector of ith individual, where $X_{ik} \in \{0, 1, 2\}$ is the number of minor alleles of ith individual at the kth rare variant.

The variant-level *P*-values can be obtained by meta-analysis or obtained from GWAS summary statistics when the original genotype and phenotype data are unavailable (Lee *et al*. 2013; Svishcheva *et al*. 2019), which are denoted as P_1, P_2... P_M. Here, we consider J candidate truncation threshold on per-site *P*-values, as $\theta_1, \theta_2, \ldots, \theta_J$. For the jth truncation threshold θ_j, the significance score of rare variant is:

$$T_j = \sum\nolimits_{\{i:P_i<\theta_j\}} w_i \tan\{(0.5 - P_i)\pi\},$$

w_i is the weight corresponding to P_i. The transformed variable $\tan\{(0.5 - P_i)\pi\}$ follows Cauchy distribution if P_i is from the null hypothesis of no association (Liu *et al*. 2019; Liu and Xie 2020). The tail of the null distribution of T_j can be well approximated by a Cauchy distribution with a location parameter 0 and a scale parameter $w_j = \sum_{\{i:P_i<\theta_j\}} w_i^2$ (Liu and Xie 2020). Hence, the *P*-value of T_j is approximated based on the cumulative density function of the Cauchy distribution (Liu *et al*. 2019):

$$P_j \approx \frac{1}{2} - \left\{\arctan\left(\frac{T_j}{w_j}\right)\right\}/\pi.$$

For rare-variant analysis, Wu *et al*. (2011) proposed $\sqrt{w_{i,\mathrm{SKAT}}} = Beta(\mathrm{MAF}_i; a_1, a_2)$, the beta density function with two parameters a_1 and a_2 where MAF_i is the minor allele frequency of ith variant (Wu *et al*. 2011). To make it comparable with SKAT, we choose the same weight that was used in SKAT method. Considering that the weights are used at different levels of variables in SKAT and ACC, we can standardize the weights to the same variable level $w_i = Beta(\mathrm{MAF}_i; a_1, a_2)^2$. In simulation analysis, we compare two weights. The first one is a_1=1, a_2=25, which means rarer variants have larger effect and another one is a_1=1, a_2=1, which means all variants have equal effect. The overall statistic is:

$$\mathrm{T_{ACC}} = \min_{1 \le j \le J} P_j$$

To estimate the *P*-value of $\mathrm{T_{ACC}}$, we set a total of D permutations. $\mathrm{T_{ACC}}^{(d)}$ is the value of $\mathrm{T_{ACC}}$ based on dth permuted data, d=0, 1, ..., D, where d=0 represents the original data. And the *P*-value of the statistics $\mathrm{T_{ACC}}$ is given by:

$$P_{\mathrm{ACC}} = \frac{\#\{\mathrm{T}_{\mathrm{ACC}}^{(d)} < \mathrm{T}_{\mathrm{ACC}}^{(0)}\}, d = 1, 2, \ldots, D}{D}$$

Results

Simulation design

The simulation studies work on the GAW17 dataset. This dataset contains genotype data of 697 unrelated individuals on 3205 genes. We designed the simulation studies (Sha *et al*. 2012). We specifically choose four genes: ELAVL4, MSH4, PDE4B and ADAMTS4 with 10, 20, 30, and 40 variants, respectively. We merged the four genes to form a super gene (Sgene) with 100 variants (Sha *et al*. 2012). In our simulation studies, we generate genotype of n individuals based on the genotype of 697 individuals in the Sgene by randomly combining two haplotypes of 1394 haplotypes of the 697 individuals.

To evaluate type I error rates, we simulate continuous trait values which are independent of genotype by using the linear model:

$$\mathrm{Y} = 0.3Z_1 + 0.4Z_2 + 0.5Z_3 + 0.4Z_4 + \varepsilon$$

and binary traits values by:

$$\mathrm{logitP(Y = 1)} = \alpha + 0.5Z_1 + 0.5Z_2 + 0.5Z_3 + 0.5Z_4$$

where $\mathrm{Z_1}$, $\mathrm{Z_2}$ and $\mathrm{Z_3}$ are continuous covariates generated independently from a standard normal distribution, $\mathrm{Z_4}$ is a

binary covariate taking values 0 and 1 with equal probability, and ε follows a standard normal distribution, α is the disease prevalence at 0.01.

To evaluate statistical power, we randomly selected a certain proportion of variants in Sgene as causal variants. We consider variants with MAF≤5%. n^r and n^p are the numbers of risk variants and protective variants, respectively. Specifically, we generate continuous traits as:

$$Y = 0.3Z_1 + 0.4Z_2 + 0.5Z_3 + 0.4Z_4 + \sum_{i=1}^{n^r} \beta_i^r G_i^r - \sum_{j=1}^{n^p} \beta_j^p G_j^p + \varepsilon$$

and binary traits by:

$$\text{logitP}(Y = 1) = \alpha + 0.5Z_1 + 0.5Z_2 + 0.5Z_3 + 0.5Z_4 + \sum_{i=1}^{n^r} \beta_i^r G_i^r - \sum_{j=1}^{n^p} \beta_j^p G_j^p$$

where G_i^r and G_i^p are genotypes of the ith risk variant and jth protective variant respectively. β^r and β^p are the effect sizes for causal variants. And Z_1, Z_2, Z_3, Z_4, ε and α are the same as the simulation of type I error rates.

Type-I error rates

To evaluate type I error rates, we set the sample sizes 2000 and 2500 under different significance levels. The *P*-values were estimated by 1000 permutations. Type I error rates were calculated as the proportion of *P*-values less than the significance level in 1000 replications. All results in different conditions are summarized in table 1. For 1000 replications, the 95% confidence intervals (CIs) for the estimated type I error rates of nominal levels 0.05 and 0.01 are (0.0362, 0.0637) and (0.0038, 0.016).

From table 1 we can see that all results are within 95% CIs, so our proposed method is valid. In table 1, ACC(1-25) indicate that the parameters used in the beta density function are a_1=1, a_2=25, ACC(1-1) indicate that the parameters used in the beta density function are a_1 =1, a_2=1. Similar expressions are used in different methods.

Power comparisons

For power comparisons, we consider various scenarios that differ in the proportions of causal variants and risk variants, sample size and weights. We also consider two different cases about the effect sizes: (i) β_i is set to be a constant b, (ii) β_i is set to be a variable based on MAFs,$\beta_i = c|\log\text{MAF}_i|$. The values of b and c depends on the proportion of causal variants and are shown in table 2. The ACC method is compared with SKAT, ACAT and Burden tests. The *P*-values are estimated by 1000 permutations at a significance level of 0.01, powers are calculated as the proportion of *P*-values less than the significance level in 1000 replications.

Table 1. Type I error rates of ACC with 1000 replications.

Significant level		0.01		0.05	
Methods with different weights					
Traits	Sample sizes	ACC(1-25)	ACC(1-1)	ACC(1-25)	ACC(1-1)
Continuous traits	2000	0.008	0.011	0.058	0.056
Continuous traits	2500	0.007	0.004	0.041	0.042
Binary traits	2000	0.015	0.011	0.049	0.050
Binary traits	2500	0.011	0.010	0.049	0.048

In figure 1, power comparisons of several methods with different percentages of causal variants are given when 50% of rare causal variants are risk variants. In all cases, the ACC is much more powerful than others. With the increasing of proportion of causal variants when $\beta_i = b$, the powers of all the methods show an overall increasing trend, but the powers of all the methods show an overall decreasing trend when $\beta_i = c|\log\text{MAF}_i|$.

In figure 2, power comparisons of several methods with different percentages of risk variants under the same proportion of causal variants are given. As shown in figure 2, with the increasing of percentage of risk variants, the powers of all methods tend to stable under different conditions. ACC works better than SKAT, ACAT and Burden in all cases, and the power of ACC is significantly higher with continuous traits.

In figure 3, power comparisons of several methods with different sample sizes under different effect sizes of causal variants are given. As shown in figures that the power of all methods improve significantly as the sample size increases. And the different choices of weights have a weak influence on the powers of ACAT and ACC. The power of the ACC method is higher than others when they have the same weights. By comparing figure 3, we can find that the powers

Table 2. The values of parameters b and c.

Proportion of causal variants	b	c
4%	1	0.6
6%	0.9	0.5
8%	0.8	0.4
10%	0.7	0.3

The parameters b and c depend on the proportion of causal variants.

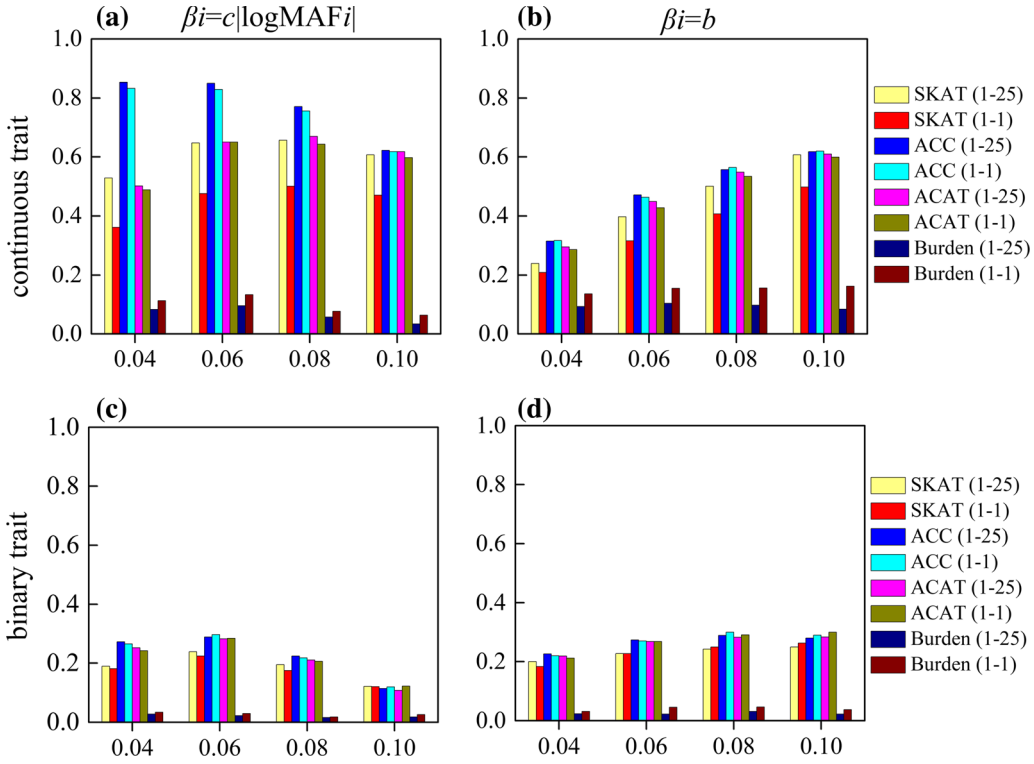

Figure 1. Power comparisons of ACC, ACAT, SKAT and Burden test in different percentages of causal variants. The *x*-axis represents the percentage of causal variants. Fifty per cent of causal variants are risk variants and the sample size is 2000. In (a), (c) panels, the effect sizes of causal variants are $\beta_i = c|\log\mathrm{MAF}_i|$ and in (b), (d) panels are $\beta_i = b$.

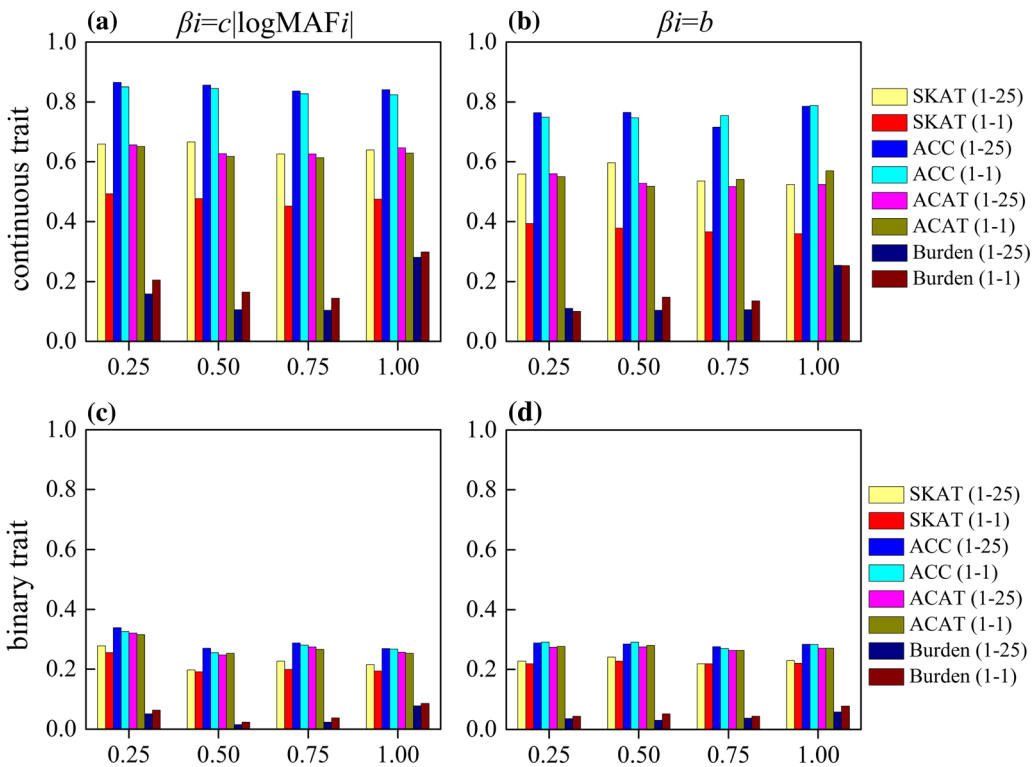

Figure 2. Power comparisons of ACC, ACAT, SKAT and Burden test in different percentages of risk variants. The *x*-axis represents the percentage of risk variants. The sample size is 2000. Six per cent of all rare variants are causal variants. In (a), (c) panels, the effect sizes of causal variants are $\beta_i = c|\log\mathrm{MAF}_i|$ and in (b), (d) panels are $\beta_i = b$.

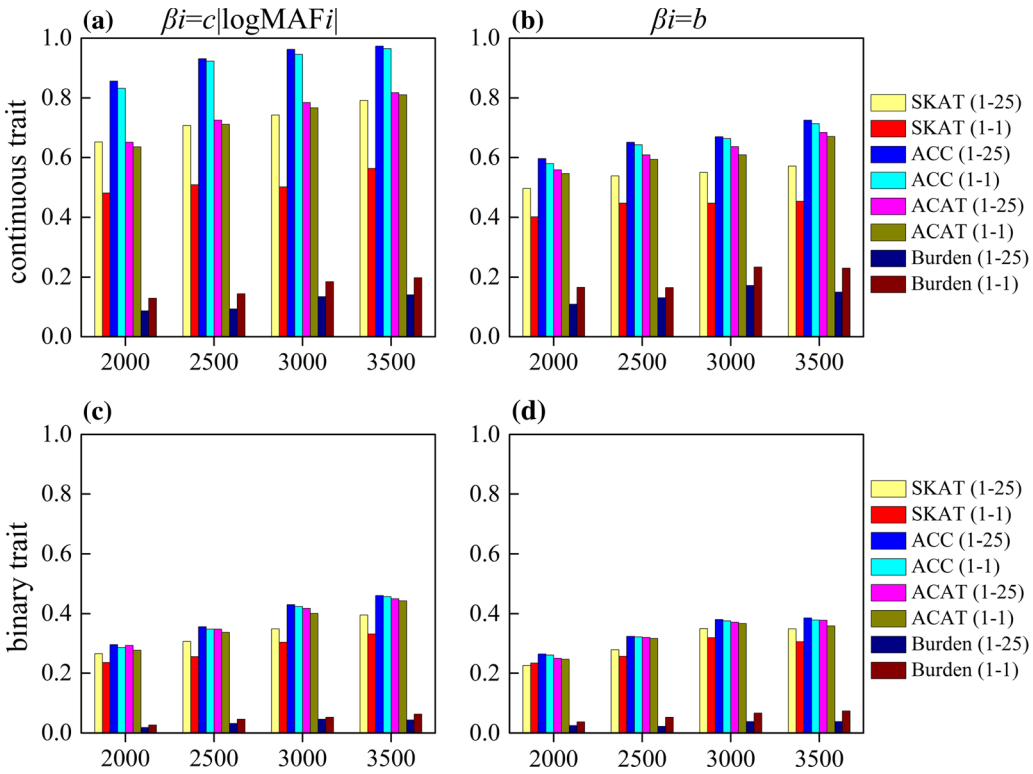

Figure 3. Power comparisons of ACC, ACAT, SKAT and Burden test in different sample sizes. The *x*-axis represents the sample sizes. In all panels, 6% of all rare variants are causal variants, 50% of rare variants are risk variants and the rest are protective variants. In (a), (c) panels, the effect sizes of causal variants are $\beta_i = c|\text{logMAF}_i|$ and in (b), (d) panels are $\beta_i = b$.

with continuous trait are higher than binary trait in all situations.

Analysis of the GAW19 dataset

To further explore the performance of SKAT, ACAT, ACC and Burden tests, we applied them on GAW19 dataset. This dataset includes 1943 Hispanic individuals with full-exome sequencing data, two phenotypes of systolic blood pressure (SBP) and diastolic blood pressure (DBP). The processing of GAW19 data follows as Chen *et al*. (2018). Moreover, we selected some genes as candidate genes to analysis which have been reported as association with the two phenotypes (SBP and DBP) according to the National Institutes of Health (NIH) GWAS catalog (Hindorff *et al*. 2014). To remove the skewed phenotype, the phenotype SBP and DBP were logarithmically transformed. Finally, the new phenotypes are log(SBP) and log(DBP). In GAW19 data, the covariates are age, sex and the status of taking medications. Because, there are many missing values in the status of taking medications, we only considered age and sex as covariates. The results of real data analysis are summarized in table 3. At a significant level of 0.05, ACC method can detect the association of the genes HIST3H2BB, MIR425, Ebf1, MY01D and SAT2 with the SBP and detect the association of the genes HIST3H2BB, TPRG1-AS and NPR3 with DBP.

Discussion

Association analysis between rare variants and disease is becoming popular. The existing analysis methods may lose power when most of the variants are noncausal in a gene region. To guard against the noises lead by the inclusion of noncausal variants, the ACC method adaptively choose the optimal truncation threshold of *P*-values to truncate those weakly associated variants. Instead of a fixed truncation threshold to truncate noises, we proposed a range of candidate truncation thresholds upon per-sit *P*-value and let the date adaptively chose an optimal truncation threshold by the data. Besides, ACC only takes the *P*-values and weights as input, and the *P*-value of the statistic T_j can be well approximated by a Cauchy distribution, which can be estimated extremely fast. Specifically, the linkage disequilibrium (LD) information in a region of the genome is not for calculating the *P*-value of T_j. By comparing ACC with other methods in different situations, we can see that the proposed ACC method is more powerful than the others for most scenarios. Through application of ACC on GAW19 dataset, the results show

Table 3. *P*-values of the tests on GAW19 dataset.

Trait	Chr	Gene	SKAT(1-25)	ACAT(1-25)	ACC(1-25)	Burden(1-25)
SBP	1	HIST3H2BB	**0.002**	0.053	**0.002**	**0.053**
	3	MIR42	**0.029**	**0.029**	**0.042**	**0.029**
	5	Ebf1	**0.015**	**0.005**	**0.031**	**0.024**
	17	MY01D	0.071	0.070	**0.001**	0.819
	17	SAT2	0.684	0.652	**0.001**	0.360
DBP	1	HIST3H2BB	**0.031**	**0.027**	**0.024**	**0.027**
	3	TPRG1-AS	0.544	0.469	**0.026**	**0.029**
	5	NPR3	0.101	0.526	**0.044**	0.228
Trait	Chr	Gene	SKAT(1-1)	ACAT(1-1)	ACC(1-1)	Burden(1-1)
SBP	1	HIST3H2BB	**0.002**	0.052	**0.002**	**0.052**
	3	MIR42	**0.029**	**0.029**	**0.032**	**0.029**
	5	Ebf1	**0.013**	**0.004**	**0.026**	**0.018**
	17	MY01D	0.077	0.069	**0.002**	0.929
	17	SAT2	0.808	0.727	**0.035**	0.596
DBP	1	HIST3H2BB	**0.032**	**0.027**	**0.015**	**0.027**
	3	TPRG1-AS	0.603	0.544	**0.019**	**0.040**
	5	NPR3	0.162	0.688	**0.044**	0.226

The significant level of the test is 0.05 and are in bold.

the ACC method can effectively detect genes associated with diseases.

Joint analysis of related traits has become a trend of association analysis, which improves the power of analysis by using the correlation between traits. Therefore, our next research goal is to apply this method to the analysis of multiple phenotypes.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant No. 12071114), and basic research expenditure of universities in Heilongjiang Province, special fund of Heilongjiang University (KJCX201803 and KJCX201804). The Genetic Analysis Workshops are supported by GAW Grant R01 GM031575 from the National Institute of General Medical Sciences. Preparation of the Genetic Analysis Workshop 17 Simulated Exome Dataset was supported in part by NIH R01 MH059490 and used sequencing data from the 1000 Genomes Project (http://www.1000genomes.org). The GAW19 unrelated data were provided by Type 2 Diabetes Genetic Exploration by Next-generation sequencing in Ethnic Samples (T2D-GENES) Project 1.

References

Barnett I., Mukherjee R. and Lin X. 2017 The generalized higher criticism for testing SNP-set effects in genetic association studies. *J. Am. Stat. Assoc.* **112**, 64–76.

Chen L., Wang Y. and Zhou Y. 2018 Association analysis of multiple traits by an approach of combining P values. *J. Genet.* **97**, 79–85.

Donoho D. and Jin J. 2004 Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Stat.* **32**, 962–994.

Eichler E. E., Flint J., Gibson G., Kong A., Leal S. M., Moore J. H. *et al.* 2010 Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446–450.

Hindorff L. A., MacArthur J. and Junkins H. A. 2014 A catalog of published genome-wide association studies (available at: http://www.genome.gov/gwastudies).

Lee S., Teslovich T. M., Boehnke M. and Lin X. H. 2013 General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet*. **93**, 42–53.

Liu Y. and Xie J. 2020 Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *J. Am. State Assoc.* **115**, 393–402.

Liu Y., Chen S., Li Z., Morrison A. C., Boerwinkle E. and Lin X. H. 2019 ACAT: A fast and powerful P value combination method for rare-variant analysis in sequencing studies. *Am. J. Hum. Genet.* **104**, 410–421.

Madsen B. E. and Browning S. R. 2009 A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5**, e1000384.

Manolio T. A., Collins F. S., Cox N. J., Goldstein D. B., Hindorff L. A., Hunter D. J. *et al.* 2009 Finding the missing heritability of complex disease. *Nature* **461**, 747–753.

Pritchard J. K. and Cox N. J. 2002 The allelic architecture of human disease genes: common disease-common variant … or not? *Hum. Mol. Genet.* **11**, 2417–2423.

Price A. L., Kryukov G. V., de Bakker P. I., Purcell S. M., Staples J., Wei L. J. *et al.* 2010 Pooled association tests for rare variants in exon-resequencing studies. *Bioinformatics* **86**, 832–838.

Pasaniuc B. and Price A. L. 2017 Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18**, 117–127.

Sha Q., Wang X., Wang X. and Zhang S. 2012 Detecting association of rare and common variants by testing an optimally weighted combination of variant. *Genet. Epidemiol.* **36**, 561–571.

Svishcheva G. R., Belonogova N. M., Zorkoltseva I. V., Kirichenko A. V. and Axenovich T. I. 2019 Gene-based association tests using GWAS summary statistics. *Bioinformatics.* **35**, 3701–3708.

Visscher P. M., Brown M. A., McCarthy M. I. and Yang J. 2012 Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24.

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

Welter D., MacArthur J., Morales J., Burdett T., Hall P., Junkins H. *et al.* 2014 The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006.

Zuk O., Hechter E., Sunyaev S. R. and Lander E. S. 2012 The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. USA* **109**, 1193–1198.

Corresponding editor: SHRISH TIWARI