Journal of Genetics, Vol. 97, No. 1, March 2018, pp. 79–85
https://doi.org/10.1007/s12041-018-0885-0

RESEARCH ARTICLE

Association analysis of multiple traits by an approach of combining *P* values

LILI CHEN[1,2]*, YONG WANG[1] and YAJING ZHOU[2]

[1] *Department of Mathematics, School of Sciences, Harbin Institute of Technology, Harbin 150001, People's Republic of China*
[2] *Department of Statistics, School of Mathematical Sciences, Heilongjiang University, Harbin 150080, People's Republic of China*
*For correspondence. E-mail: chenlili_02_06@163.com.

Received 21 October 2016; revised 17 February 2017; accepted 13 June 2017; published online 6 February 2018

Abstract. Increasing evidence shows that one variant can affect multiple traits, which is a widespread phenomenon in complex diseases. Joint analysis of multiple traits can increase statistical power of association analysis and uncover the underlying genetic mechanism. Although there are many statistical methods to analyse multiple traits, most of these methods are usually suitable for detecting common variants associated with multiple traits. However, because of low minor allele frequency of rare variant, these methods are not optimal for rare variant association analysis. In this paper, we extend an adaptive combination of *P* values method (termed ADA) for single trait to test association between multiple traits and rare variants in the given region. For a given region, we use reverse regression model to test each rare variant associated with multiple traits and obtain the *P* value of single-variant test. Further, we take the weighted combination of these *P* values as the test statistic. Extensive simulation studies show that our approach is more powerful than several other comparison methods in most cases and is robust to the inclusion of a high proportion of neutral variants and the different directions of effects of causal variants.

Keywords. association analysis; rare variant; common variant; multiple traits.

Introduction

Genomewide association studies (GWAS) have successfully detected a large number of common genetic variants in human complex diseases (Visscher *et al.* 2012; Welter *et al.* 2014). However, these common variants explain only a small proportion of disease heritability (Maher 2008; McCarthy *et al.* 2008; Bansal *et al.* 2010). Next-generation sequencing technology can identify the rare causal variants that are associated with the complex traits. Rare variants are actually responsible for part of disease heritability (Pritchard 2001; Pritchard and Cox 2002; Manolio *et al.* 2009). Because of low minor allele frequency of rare variant, it will be difficult to detect single rare variant. Hence, most of the methods used for single common variant are underpowered to detect rare variants. To improve the power of the rare-variant association test, many approaches have been proposed to test the collective effect of rare variants in a genomic region to enrich the association signal. These methods are roughly divided into burden tests and nonburden tests. Burden tests summarize the rare variants in a genomic region by a single value and then test the association between the single value and the interesting trait, such as the cohort allelic sums test (Morgenthaler and Thilly 2007), the combined multivariate and collapsing method (Li and Leal 2008), the weighted sum statistic (Madsen and Browning 2009), the variable minor allele frequency threshold method (Price *et al.* 2010), and so on. The burden tests are more powerful to the same directions of effects of variants and suffer from the loss of power with both protective and deleterious variants (Basu and Pan 2011). On the other hand, nonburden tests (also called variance-component tests), evaluate the distribution of genetic effects of a set of rare variants, such as the C-alpha (Neale *et al.* 2011), the sequence kernel association test (Wu *et al.* 2011). These methods are robust to the different directions of effects of variants.

However, almost all of the aforementioned methods analyse only the association between multiple variants and one trait. Increasing evidence shows that many human traits are highly correlated, and it is widespread for one variant to impact on multiple traits in complex diseases (Sivakumaran *et al.* 2011). Further, multiple correlated traits are usually measured in complex diseases, such as mental illness and behavioural disorders (Sattar *et al.* 2008; Sivakumaran *et al.* 2011). The joint analysis of multiple traits can increase statistical power to detect the genetic variants and to provide additional insights into the genetic architecture of the complex disease (Aschard *et al.* 2014). Currently, there are many methods for detecting genetic association in multiple traits such as regression methods (Korte *et al.* 2012; OReilly *et al.* 2012; Zhou and Stephens 2014), combining test statistics from univariate analysis (OBrien 1984; Yang *et al.* 2010; Van Der Sluis *et al.* 2013), dimension reduction methods (Ott and Rabinowitz 1999; Lange *et al.* 2004; Klei *et al.* 2008; Aschard *et al.* 2014).

Most methods have primarily focussed on the common variants. In recent times there has been a gradual increase in demand to develop statistic methods for detecting rare variants associated with multiple traits. Tang and Ferreira (2012) proposed a method to simultaneously test the association between multiple continuous traits and rare variants across a gene or region in unrelated individuals, based on canonical correlation analysis (CCA). Wang *et al.* (2016) used an adaptive weighting reverse regression (AWRR) to test rare variants in a genomic region associated with multiple traits. Madsen and Browning (2009) and Wang *et al.* (2016) used the weighted sum reverse regression (WSRR) to test the association between rare variants and multiple traits. When the three methods (CCA, AWRR and WSRR) consider testing the collective effect of rare variants in the region, this strategy inevitably leads to include neutral variants, which may cause loss of power. To address this type of issue, we extend an adaptive combination of P-values method (termed ADA) (Lin *et al.* 2014) to detect association between multiple traits and rare variants in the given region, therefore the method is termed MUL-ADA. For the given region, we use reverse regression model to separately test association between each rare variant and multiple traits; in other words, we treat all the traits as predictors and the number of minor allele of rare variant as response variable. Based on score test, we obtain the P value of single-variant test. To guard against the noise caused by neutral variants, we remove the variants with P values larger than a truncation threshold. Then we take the weighted combination of the remaining P values as test statistic (Yu *et al.* 2009; Lin *et al.* 2014). Weights of our approach are based on minor allele frequencies and covariance between the first principal component of multiple traits and each variant, and we apply sign of the covariance to reasonably distinguish the directions of effects of variants. Consequently, because of removing more neutral variants and reasonably using weights to combine P values, our method is more powerful, effective and robust. Extensive simulation studies indicate that our proposed method really outperforms the other comparison methods (AWRR, CCA and WSRR) over a wide range. Because of using reverse regression model, our method can be suitable for continuous traits, binary traits and mixed types of traits, and does not need to know the complex distributions of the traits.

Materials and methods

We consider continuous traits. Suppose that n subjects are sequenced in the given region with M variants sites. Each one has K correlated traits. For the ith individual, y_{ik} denotes the kth continuous trait value, and $g_{im} \in \{0, 1, 2\}$ denotes the number of minor allele at the mth variant ($i = 1, 2, \ldots, n$; $m = 1, 2, \ldots, M$; $k = 1, 2, \ldots, K$). Due to the extreme rarity of single rare variant, in fact, g_{im} essentially is 0 or 1. Let $Y_i = (y_{i1}, \ldots, y_{iK})^{\mathrm{T}}$ denote the K traits for ith individual. The detailed steps of our method are given as follows.

***Step 1*:** Define the direction of effect of each rare variant. We compute the first principal component of the K traits, denoted as Y_{comp}. Let $g_m = (g_{1m}, \ldots, g_{nm})^{\mathrm{T}}$ for the mth variant. If $\mathrm{cov}(Y_{comp}, g_m) > 0$, the variant is called 'deleterious variant', and if $\mathrm{cov}(Y_{comp}, g_m) < 0$, the variant is called 'protective variant'.

***Step 2*:** Obtain P value of single-variant association test between each rare variant and the K traits. For the mth variant, we apply reverse regression model:

$$\mathrm{logit}(p_{im}) = a_{m0} + a_{m1}y_{i1} + \cdots + a_{mK}y_{iK}, \quad (i = 1, 2, \ldots, n) \tag{1}$$

where $p_{im} = P(g_{im} = 1)$, $m = 1, 2, \ldots, M$. Association test between the K traits and each rare variant corresponds to testing the null hypothesis $H_0 : a_{m1} = \cdots = a_{mK} = 0$. The score test statistic is given by

$$S_m = U_m^{\mathrm{T}} V_m^{-1} U_m,$$

where

$$U_m = \sum\nolimits_{i=1}^{n} Y_i(g_{im} - \bar{g}_m),$$
$$V_m = \frac{1}{n}\sum\nolimits_{i=1}^{n}(g_{im} - \bar{g}_m)^2 \sum\nolimits_{i=1}^{n}(Y_i - \overline{Y})(Y_i - \overline{Y})^{\mathrm{T}},$$
$$\bar{g}_m = \frac{1}{n}\sum\nolimits_{i=1}^{n} g_{im},$$

and

$$\overline{Y} = \frac{1}{n}\sum\nolimits_{i=1}^{n} Y_i.$$

Under the null hypothesis, S_m is approximated by χ_K^2 distribution, thus the P value of S_m can be obtained, which is denoted as p_m, $m = 1, 2, \ldots, M$.

***Step 3*:** Combine the P values of single-variant tests as statistic. To effectively guard against the noise caused by neutral variants, we first impose a truncation threshold upon the P values (p_m, $m = 1, 2, \ldots, M$), then combine the P values that are smaller than the given truncation threshold. We allow multiple truncation thresholds and assume that there are J candidate truncation thresholds $(\theta_1, \ldots, \theta_J)$ (say, 0.1, 0.11, 0.12, …, 0.2) (Lin *et al.* 2014). For the jth truncation threshold θ_j, the significance scores of the deleterious and protective variants are separate

$$SR_j^d = -\sum_{m=1}^{M} \psi_m \cdot I_{[p_m < \theta_j]} \cdot \omega_m \cdot \log p_m$$

and

$$SR_j^p = -\sum_{m=1}^{M} \varphi_m \cdot I_{[p_m < \theta_j]} \cdot \omega_m \cdot \log p_m,$$

where ψ_m is the indicator variable, which is 1 if the effect of mth variant is deleterious and 0 otherwise; φ_m is the indicator variable, which is 1 if the effect of mth variant is protective and 0 otherwise; ω_m is the weight of the mth variant; $I_{[p_m<\theta_j]}$ is 1 if the P value of the mth variant is smaller than the jth truncation threshold θ_j and 0 otherwise. We define the weight ω_m such that ω_m will be large if the mth variant is a rare one and has a strong association with the first principal component of the K traits. For these purposes, we define $\omega_m = \text{Beta}(\text{MAF}_m; 1{,}25)\,|\text{cov}(Y_{comp}, g_m)|$, $m = 1, 2, \ldots, M$, where MAF_m is the minor allele frequency of the mth variant.

***Step 4*:** Obtain the overall test statistic. For taking no account of the direction of the effects, we use the statistic, $SR_j = \max(SR_j^d, SR_j^p)$. Let P_j be the P value of the statistic SR_j, for $j = 1, 2, \ldots, J$. The overall test statistic is $T = \min_{1\le j\le J}\{P_j\}$. Because variants within a functional region are usually not independent, we use permutations to evaluate the P values of the statistic $SR_j (j = 1, 2, \ldots, J)$ and the overall test statistic T, based on the permutation process of Lin *et al.* (2014).

Simulation studies

Simulation design: The GAW17 data set provides real exome sequence data from 1000 Genomes Project, and is used as the basis for simulation studies. The data set contains genotypes of 697 unrelated individuals on 3205 genes. Based on this data set, we follow the simulation setting of Sha *et al.* (2012), and choose four genes: *ELAVL4* (gene 1), *MSH4* (gene 2), *PDE4B* (gene 3), and *ADAMTS4* (gene 4) with 10, 20, 30, and 40 variants, respectively. The four genes are emerged into a super gene (Sgene) with 100 variants. According to the genotypes of 697 individuals in the Sgene, we generate genotypes of n individuals.

To evaluate the type-I error rate and power, we generate K traits by the factor model:

$$Y = \Lambda G + \sqrt{\rho}\gamma f + \sqrt{1-\rho}\varepsilon,$$

where $Y = (y_1, y_2, \ldots, y_K)^{\mathrm{T}}$, $G = (g_1, g_2, \ldots, g_{N_c})^{\mathrm{T}}$ is the vector of the genotype scores at the causal variants, N_c is the number of rare causal variants, $\Lambda = (\beta_1, \ldots, \beta_k, \ldots, \beta_K)^{\mathrm{T}}_{K\times N_c}$, $\beta_k = (\beta_{k1}, \ldots, \beta_{kN_c})^{\mathrm{T}}$, $f = (f_1, \ldots, f_R)^{\mathrm{T}} \sim \text{MVN}(0, I)$ is a vector of R-independent standard normal latent variables, I is the identity matrix, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_K)^{\mathrm{T}} \sim \text{MVN}(0, I)$ is a vector of errors, γ is a $K \times R$ loading matrix, and ρ is constant. In simulation studies, we consider two cases: (i) there is only one factor ($R = 1$), and $\gamma = (1, \ldots, 1)^{\mathrm{T}}$; (ii) there are two factors ($R = 2$), and $\gamma = \text{diag}(D_1, D_2)$, where $D_1 = (1, \ldots, 1)^{\mathrm{T}}_{[K/2]\times 1}$, $D_2 = (1, \ldots, 1)^{\mathrm{T}}_{(K-[K/2])\times 1}$. Therefore, $Y \sim \text{MVN}(\Lambda G, \Sigma)$, where $\Sigma = \rho\gamma\gamma^{\mathrm{T}} + (1-\rho)I$.

Table 1. Type-I error rates.

	α	MUL-ADA	CCA	WSRR	AWRR
Case 1	0.01	0.006	0.004	0.012	0.008
	0.05	0.034	0.052	0.050	0.052
Case 2	0.01	0.010	0.004	0.012	0.012
	0.05	0.042	0.052	0.050	0.042

α represents the significance level.

For evaluating type-I error rates, let $\beta_{kj} = 0$, $k = 1, \ldots, K; j = 1, \ldots, N_c$. For power comparasions, we consider β_{kj}, $k = 1, \ldots, K; j = 1, \ldots, N_c$ are constants and their values depend on the total heritability. Suppose that the heritability of each rare causal variant is not always equal and rare causal variants impact all traits.

We compare our proposed method (MUL-ADA) with AWRR (Wang *et al.* 2016), CCA (Tang and Ferreira 2012), and WSRR (Madsen and Browning 2009; Wang *et al.* 2016). The AWRR and WSRR methods are implemented with their R scripts.

Evaluation on type-I error rates: For evaluating type-I error rates, sample size is set at 1000, P values are estimated by 1000 permutations and type-I error rates are evaluated by 500 replications. Table 1 summarizes the estimated type-I error rates for given different significance levels (termed α, $\alpha = 0.01, 0.05$) in two different cases, and shows that the estimated type-I error rates are not significantly different from the nominal levels.

Power comparisons: For power comparison, sample size is set at 1000, P values are estimated by 1000 permutations and powers are evaluated by 500 replications at a significance level of 0.05. In simulation, we consider different values of heritability, different percentages of protective variants, different percentages of causal variants, and different numbers of traits in two different cases.

Figure 1 shows patterns of power comparisons for different values of heritability in two cases. In figure 1, our proposed method (MUL-ADA) is more powerful than the other three methods (CCA, AWRR and WSRR). Of these three methods, AWRR is more powerful than CCA and WSRR, and WSRR is the least

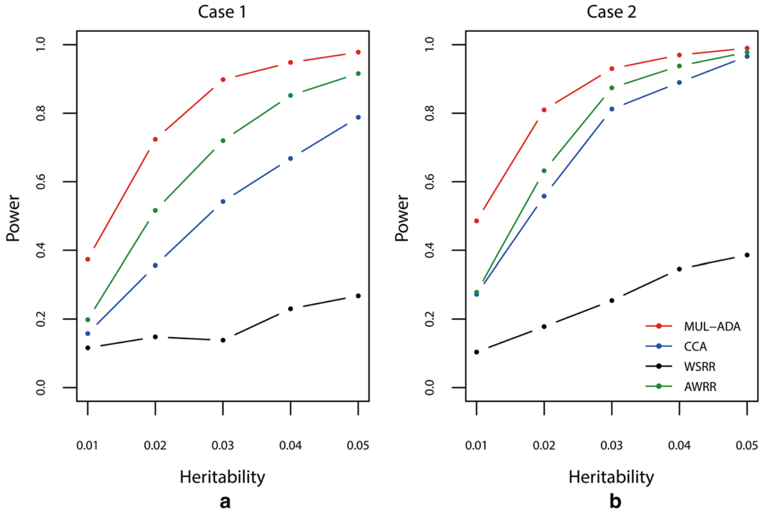

Figure 1. Power for different values of the total heritability in two cases. The sample size is 1000, and $\rho = 0.5$. 10% of rare variants are causal, and 20% of rare causal variants are protective.

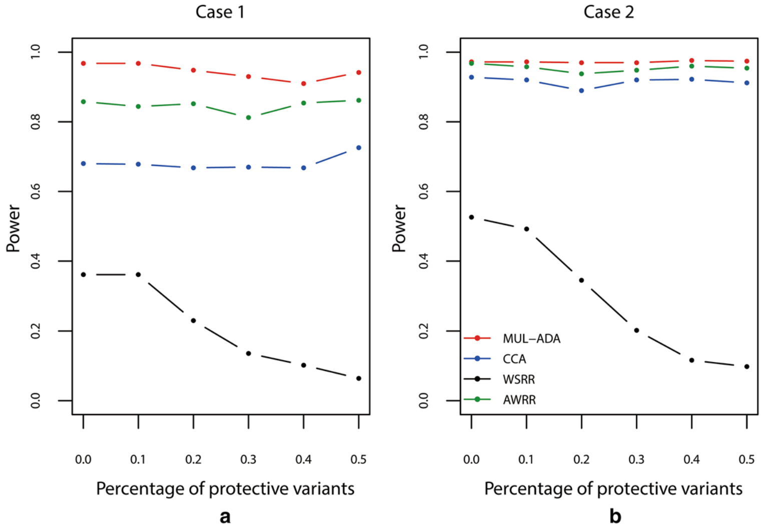

Figure 2. Power for different percentages of protective variants in two cases. The sample size is 1000, and $\rho = 0.5$. 10% of rare variants are causal, and the total heritability of all causal variants is 0.04.

powerful. In addition, our method is much more powerful than the other three methods when the heritability is very small, because our approach can exclude more neutral variants, strengthen the association signal and guard against the noise caused by neutral variants.

Figure 2 shows patterns of power comparisons for different percentages of protective variants in two cases. In figure 2, we can see that MUL-ADA is the most powerful, and WSRR is the least powerful. MUL-ADA, AWRR and CCA are robust to the percentage of protective variants. However, WSRR suffers from substantial loss of power when both deleterious and protective variants are present.

Figure 3 shows patterns of power comparisons for different percentages of causal variants in two cases. In figure 3, MUL-ADA is the most powerful. The leading cause is that our method can better guard against the noise caused by neutral variants. WSRR is still the least powerful and suffers from loss of power with a high proportion of neutral variants.

Figure 4 shows patterns of power comparisons for different numbers of traits in two cases. In figure 4, WSRR is the least powerful and MUL-ADA is the most powerful. MUL-ADA and AWRR are robust to the different numbers of traits. But CCA and WSRR suffer from loss of power with the increase of numbers of traits.

In summary, our proposed method is more powerful than the other comparison methods in most situations, and is robust to the inclusion of a high proportion of neutral variants, and different directions of effects of causal variants.

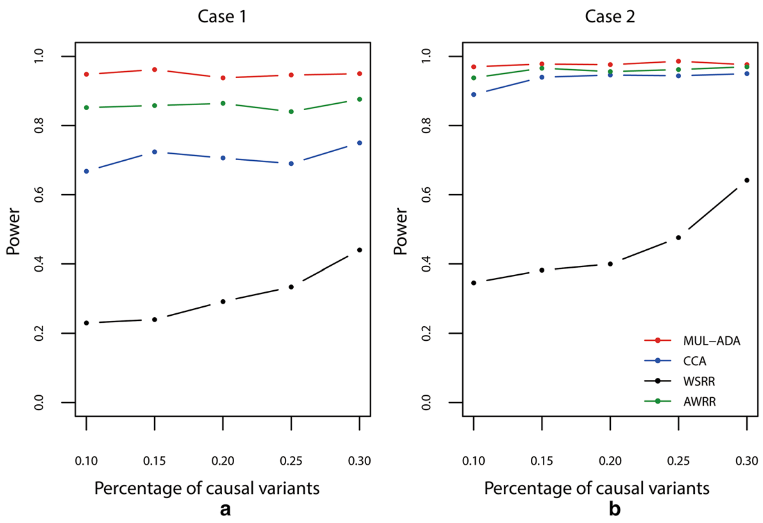

Figure 3. Power for different percentages of rare causal variants in two cases. The sample size is 1000, and $\rho = 0.5$. 20% of rare causal variants are protective, and the total heritability of all causal variants is 0.04.

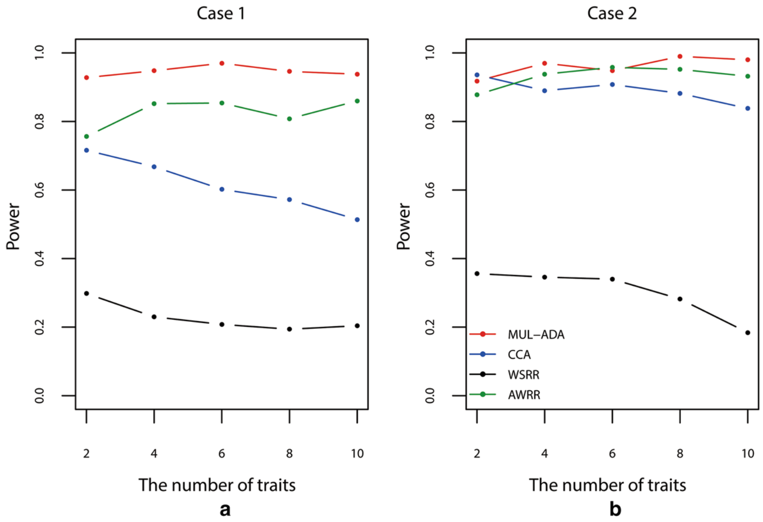

Figure 4. Power for different numbers of traits impacted by causal variants in two cases. The sample size is 1000, and $\rho = 0.5$. 10% of rare variants are causal, 20% of rare causal variants are protective, and the total heritability of all causal variants is 0.04.

Discussion

In genetic studies, joint analysis of multiple traits can increase statistical power to detect genetic variants. But the existing methods are usually suitable for common variants. Consequently, there is a surprising demand to develop statistical methods to detect rare variants associated with multiple traits. Here, we extend an adaptive combination of P-values method for single trait to test rare variants associated with multiple traits. Further, our method can be suitable for continuous traits, binary traits and mixed types of traits, and does not need to know the complex distributions of the traits.

In the presence of noise traits, our proposed method is still powerful and effective through simulation studies. For power comparison, we consider 10 traits and rare causal variants impact on four traits among the 10 traits. Simulation results given in figure 5 indicate that the powers of the three methods (MUL-ADA, CCA and AWRR) are very close in most cases, and WSRR is the least powerful one. When rare variants impact on all the traits, our method is the most powerful. However, our method suffers from loss of power in the presence of a large number of noise traits (i.e., rare variants impact on two traits among the 10 traits). The leading cause is that we use the first principal component of all the traits to define the direction of effect of each variant.

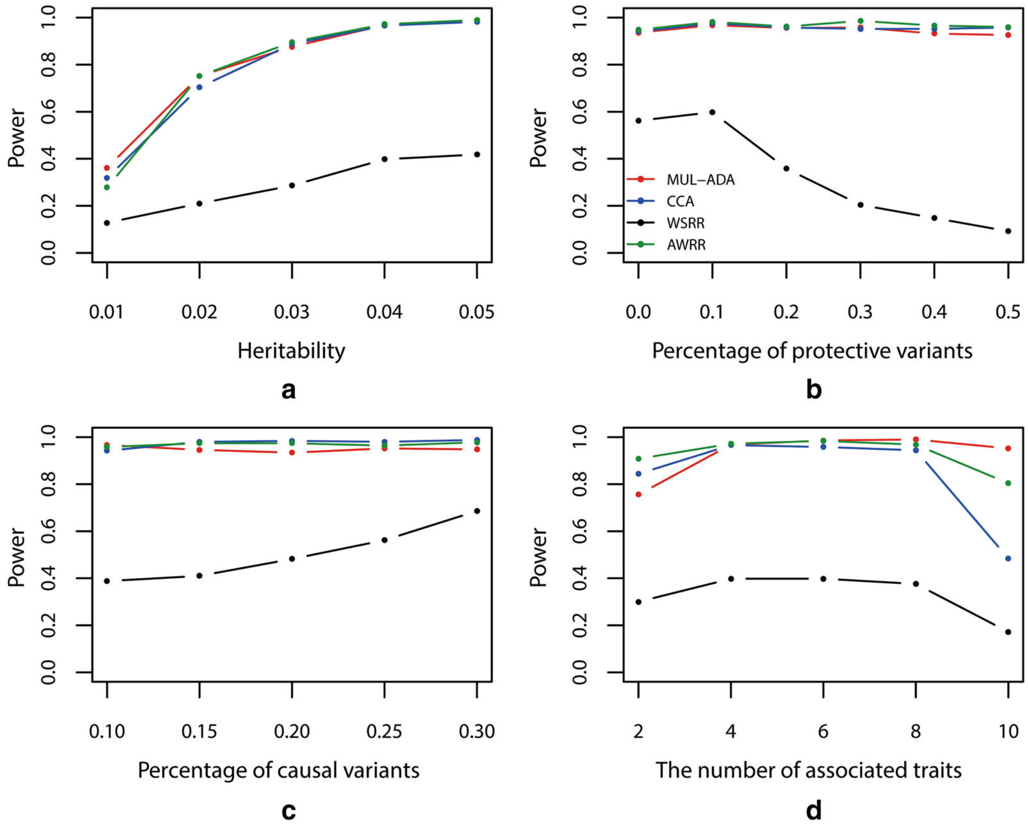

Figure 5. The total number of traits is 10, the sample size is 1000, and $\rho = 0.5$ in case 1. (a) Power for different values of the total heritability, 10% of rare variants are causal, and 20% of rare causal variants are protective. (b) Power for different percentages of protective variants, 10% of rare variants are causal, and the total heritability of all causal variants is 0.04. (c) Power for different percentages of rare causal variants, 20% of rare causal variants are protective, and the total heritability of all causal variants is 0.04. Causal variants impact on four traits among the 10 traits (a–c). (d) Power for different numbers of traits impacted by causal variants. Ten per cent of rare variants are causal, 20% of rare causal variants are protective, and the total heritability of all causal variants is 0.04.

Owing to the fact that rare and common variants can commonly cause complex disease (Walsh and King 2007; Bodmer and Bonilla 2008; Stratton and Rahman 2008; Ng *et al.* 2009; Teer and Mullikin 2010), we apply our method to detect rare and common variants associated with multiple traits. Concretely, variants are divided into rare and common variants. For rare variant, we still apply model (1) for testing single-variant association and obtain *P* value of single-variant test. Because genotype score of common variant is 0, 1 or 2, we use the proportional odds models:

$$\text{logit}\{P(g_i \leq l)\} = \alpha_l + b_1 y_{i1} + \cdots + b_K y_{iK}, \quad l = 0, 1, 2$$

where $g_i, y_{i1}, \ldots, y_{iK}$ are the same as those in model (1). We use the likelihood ratio test to test the null hypothesis H_0: $b_1 = b_2 = \cdots b_K = 0$. Under H_0, the likelihood ratio test statistic asymptotically follows χ^2_K and can be obtained by R software. Thus the *P* value of each common variant test can be obtained. Further we combine these *P* values of common and rare variant tests as statistic. Further, in future we need to investigate the performance of our proposed method for testing rare and common variants associated with multiple traits.

Acknowledgements

The Genetic Analysis workshops were 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). This work was conducted in the framework of basic research expenditure of universities in Heilongjiang Province, special fund of Heilongjiang University (no. HDJCCX-201631).

References

Aschard H., Vilhjalmsson B. J., Greliche N., Morange P. E., Tregouet D. A. and Kraft P. 2014 Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am. J. Hum. Genet.* **94**, 662–676.

Bansal V., Libiger O., Torkamani A. and Schork N. J. 2010 Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.* **11**, 773–785.

Basu S. and Pan W. 2011 Comparison of statistical tests for disease association with rare variants. *Genet. Epidemiol.* **35**, 606–619.

Bodmer W. and Bonilla C. 2008 Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* **40**, 695–701.

Klei L., Luca D., Devlin B. and Roeder K. 2008 Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genet. Epidemiol.* **32**, 9–19.

Korte A., Vilhjalmsson B. J., Segura V., Platt A., Long Q. and Nordborg M. 2012 A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* **44**, 1066–1071.

Lange C., Van Steen K., Andrew T., Lyon H., Demeo D. L., Raby B. *et al.* 2004 A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. *Stat. Appl. Genet. Mol. Biol.* **3**, 1–27.

Li B. and Leal S. M. 2008 Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321.

Lin W.-Y., Lou X.-Y., Gao G. and Liu N. 2014 Rare variant association testing by adaptive combination of *P*-values. *PLoS One* **9**, e85728.

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

Maher B. 2008 Personal genomes: the case of the missing heritability. *Nature* **456**, 18–21.

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 diseases. *Nature* **461**, 747–753.

McCarthy M. I., Abecasis G. R., Cardon L. R., Goldstein D. B., Little J., Ioannidis J. P. *et al.* 2008 Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369.

Morgenthaler S. and Thilly W. G. 2007 A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res.* **615**, 28–56.

Neale B. M., Rivas M. A., Voight B. F., Altshuler D., Devlin B., Orho-Melander M. *et al.* 2011 Testing for an unusual distribution of rare variants. *PLoS Genet.* **7**, e1001322.

Ng S. B., Turner E. H. and Robertson P. D. 2009 Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276.

OBrien P. C. 1984 Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–1087.

OReilly P. F., Hoggart C. J., Pomyen Y., Calboli F. C., Elliott P., Jarvelin M. R. *et al.* 2012 MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One* **7**, e34861.

Ott J. and Rabinowitz D. 1999 A principal-components approach based on heritability for combining phenotype information. *Hum. Hered.* **49**, 106–111.

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. *Am. J. Hum. Genet.* **86**, 832–838.

Pritchard J. K. 2001 Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137.

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.

Sattar N., Mcconnachie A., Shaper A. G., Blauw G. J., Buckley B. M., De Craen A. J. *et al.* 2008 Can metabolic syndrome usefully predict cardiovascular disease and diabetes? Outcome data from two prospective studies. *Lancet* **371**, 1927–1935.

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

Sivakumaran S., Agakov F., Theodoratou E., Prendergast J. G., Zgaga L., Manolio T. *et al.* 2011 Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* **89**, 607–618.

Stratton M. R. and Rahman N. 2008 The emerging landscape of breast cancer susceptibility. *Nat. Genet.* **40**, 17–22.

Tang C. S. and Ferreira M. A. 2012 A gene-based test of association using canonical correlation analysis. *Bioinformatics* **28**, 845–850.

Teer J. K. and Mullikin J. C. 2010 Exome sequencing: the sweet spot before whole genomes. *Hum. Mol. Genet.* **19**, R145–R151.

Van Der Sluis S., Posthuma D. and Dolan C. V. 2013 TATES: Efficient multivariate genotype-phenotype analysis for genomewide association studies. *PLoS Genet.* **9**, e1003235.

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.

Walsh T. and King M. C. 2007 Ten genes for inherited breast cancer. *Cancer Cell* **11**, 103–105.

Wang Z., Wang X., Sha Q. and Zhang S. 2016 Joint analysis of multiple traits in rare variant association studies. *Ann. Hum. Genet.* **80**, 162–171.

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.

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

Yang Q., Wu H., Guo C. Y. and Fox C. S. 2010 Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genet. Epidemiol.* **34**, 444–454.

Yu K., Li Q., Bergen A. W., Pfeiffer R. M. and Rosenberg P. S. 2009 Pathway analysis by adaptive combination of *P*-values. *Genet. Epidemiol.* **33**, 700–709.

Zhou X. and Stephens M. 2014 Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* **11**, 407–409.

Corresponding editor: KUNAL RAY