Genet. Res., Camb. (2015), vol. 97, e20. © Cambridge University Press 2015
doi:10.1017/S0016672315000208

Detecting association of rare and common variants by adaptive combination of *P*-values

YAJING ZHOU[1,2] AND YONG WANG[1]*
[1]*Department of Mathematics, School of Science, Harbin Institute of Technology, Harbin 150001, China*
[2]*School of Mathematical Sciences, Heilongjiang University, Harbin 150080, China*

Summary

Genome-wide association studies (GWAS) can detect common variants associated with diseases. Next generation sequencing technology has made it possible to detect rare variants. Most of association tests, including burden tests and nonburden tests, mainly target rare variants by upweighting rare variant effects and downweighting common variant effects. But there is increasing evidence that complex diseases are caused by both common and rare variants. In this paper, we extend the ADA method (adaptive combination of *P*-values; Lin *et al.*, 2014) for rare variants only and propose a RC-ADA method (common and rare variants by adaptive combination of *P*-values). Our proposed method combines the per-site *P*-values with the weights based on minor allele frequencies (MAFs). The RC-ADA is robust to directions of effects of causal variants and inclusion of a high proportion of neutral variants. The performance of the RC-ADA method is compared with several other association methods. Extensive simulation studies show that the RC-ADA method is more powerful than other association methods over a wide range of models.

1. Introduction

Over the past several years, GWAS have successfully identified thousands of common genetic variants associated with complex traits and common diseases (Visscher *et al.*, 2012; Welter *et al.*, 2014). However, many common variants identified explain only a small proportion of heritability (Maher, 2008; McCarthy *et al.*, 2008; Bansal *et al.*, 2010). It has been hypothesized that some of the heritability may be explained by unidentified rare variants (Pritchard, 2001; Pritchard & Cox, 2002; Manolio *et al.*, 2009). Next generation sequencing technology is being conducted to identify rare variants associated with complex traits. Since frequencies of rare variants are very low, it will be difficult to detect single rare variants. Hence, many existing methods for single common variants would not work and are underpowered for single rare variants. Recently, to increase the power of rare variants association tests, many statistical methods have been proposed. These methods can be classified into burden tests and nonburden tests.

Burden tests collapse the multiple rare variants within a given region into a single variable, then test the association between the single variable and the trait of interest. For example, the cohort allelic sums test (Morgenthaler & Thilly, 2007), the combined multivariate and collapsing method (CMC; Li & Leal, 2008), the weighted sum statistic (Madsen & Browning, 2009), the variable minor allele frequency threshold method (Price *et al.*, 2010) and so on. The same strategy is used in many methods (Feng *et al.*, 2011; Lin & Tang, 2011; Sha *et al.*, 2011; Fang *et al.*, 2012; Sha *et al.*, 2013). These burden tests are more powerful when all rare variants within a region influence the phenotype in the same direction (Basu & Pan, 2011). Nonburden tests, such as the C-alpha (Neale *et al.*, 2011), the optimally weighted combination of variants (Sha *et al.*, 2012), and the sequence kernel association test (SKAT; Wu *et al.*, 2011), are based on the kernel machine regression method and are robust to the different directions of effects of variants.

There are several limitations for the above rare variant association methods. First, these association tests mainly target rare variants by putting large weights on rare variants and small weights on common variants. When common variants are also associated with the trait, these association methods can lead to loss of power. In fact, the relative contribution of common and rare variants is unknown for many complex traits. Recent studies show that some

*Corresponding author: Harbin Institute of Technology, No. 92 Xidazhi Street, Nangang District, Harbin 150001, P. R. China. E-mail: mathwy@hit.edu.cn

complex diseases are caused by both common and rare variants (Walsh & King, 2007, Bodmer & Bonilla, 2008, Stratton & Rahman, 2008, Ng *et al.*, 2009, Teer & Mullikin, 2010). So it is reasonable to assume that rare and common variants commonly influence the complex traits. Sha *et al.* (2012) analytically derived optimal weights under a certain criterion and proposed a variable weight test for testing the effect of an optimally weighted combination of variants (VW-TOW). The VW-TOW aimed to test the effects of both rare and common variants. Second, these association tests suffer from power loss with the inclusion of a large proportion of neutral variants. To guard against the noise caused by the inclusion of neutral variants, Lin *et al.* (2014) proposed an ADA method that adaptively combines per-site *P*-values with the weights based on MAFs. Before combining *P*-values, they imposed a truncation threshold upon the per-site *P*-values. However, their association method only targeted rare variants.

In this paper, we extend the ADA method and propose a RC-ADA method that detects both rare and common variants in a given region by adaptive combination of *P*-values. For the given region, each common variant or each rare variant is separately tested, to obtain per-site *P*-values. We use a suited weight scheme for rare and common variants when per-site *P*-values are combined. To guard against the noise caused by neutral variants, variants with *P*-values larger than a threshold will be truncated. We don't fix a *P*-value truncation threshold, on the contrary, we allow multiple candidate truncation thresholds (0.10, 0.11, 0.12, $\cdots$, 0.20) to choose the optimal threshold for any given data set. Our proposed method is applicable to binary traits, and is robust to the directions of effects of causal variants and the inclusion of a large proportion of neutral variants. Extensive simulation studies are used to compare the performance of the proposed method with that of other existing methods. Simulation results show that the RC-ADA is more powerful across a wide of range of scenarios.

2. Materials and methods

We consider a binary trait. Assume n individuals are sequenced in a genomic region (e.g. a candidate gene) with m variant sites. Denote Y_i as the trait value of the ith individual (1 for case and 0 for control) and denote $G_i = (G_{i1}, G_{i2}, \cdots, G_{im})^T$ as genotypic score of the ith individual, where $G_{ik} \in \{0, 1, 2\}$ is the number of minor alleles the ith individual has at the kth variant. Our analysis goal is to detect whether there is any association between the trait and the genomic region (a group of rare and common variants). We firstly test the association between the trait and each variant in the region. We divide variants into rare variants (MAF < the rare variant threshold [RVT]) and common variants (MAF > RVT). For each rare variant in the region, *P*-value is obtained by the Fisher's exact test (Fisher, 1922; Cheung *et al.*, 2012). For each common variant in the region, we consider a logistic regression model:

$$\text{logit}P(Y_i = 1) = \beta_0 + \beta_1 G_{ik}^c,$$

where superscript c represents the common variant. The score statistic of testing $\beta_1 = 0$ is:

$$T_k = \frac{\left(\sum_{i=1}^{n}(Y_i - \overline{Y})(G_{ik}^c - \overline{G_k^c})\right)^2}{\overline{Y}(1-\overline{Y})\sum_{i=1}^{n}(G_{ik}^c - \overline{G_k^c})^2},$$

where $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, $\overline{G_k^c} = \frac{1}{n}\sum_{i=1}^{n} G_{ik}^c$. T_k is approximated by χ_1^2 distribution.

Let the per-site *P*-values of the m variants be p_1, $p_2, \cdots, p_m$, respectively. We name the sites with larger variant frequencies in cases than in controls, 'deleterious-inclined variant sites', and those with larger variant frequencies in controls than in cases 'protective-inclined variant sites'. To test the significance of the region and to guard against the noise caused by neutral variants, we combine the per-site *P*-values that are smaller than a given truncation threshold. Suppose that we consider J candidate truncation thresholds $\theta_1, \theta_2, \cdots, \theta_J$. For the jth truncation threshold θ_j, the significance score of the deleterious-inclined variant sites is:

$$S_j^+ = -\sum_{i=1}^{m} \xi_i \cdot I_{[p_i < \theta_j]} \cdot w_i \cdot \log p_i,$$

where the indicator variable ξ_i is 1 if the ith site is deleterious-inclined and 0 otherwise, w_i is the weight of the ith site, and $I_{[p_i < \theta_j]}$ is 1 if the *P*-value of the ith site is smaller than the jth truncation threshold θ_j and 0 otherwise. Similarly, for the jth truncation threshold θ_j, the significance score of the protective-inclined variant sites is:

$$S_j^- = -\sum_{i=1}^{m} \varphi_i \cdot I_{[p_i < \theta_j]} \cdot w_i \cdot \log p_i,$$

where the indicator variable φ_i is 1 if the ith site is protective-inclined and 0 otherwise. In this paper, we specify 11 candidate truncation thresholds (0.1, 0.11, 0.12, $\cdots$, 0.2) (we will discuss the selection of candidate truncation thresholds in the Discussion section). Since the goal is to test the association regardless of the direction of the effects, we use the statistic $S_j = \max(S_j^+, S_j^-)$. Let P_j be the *P*-value of the statistic S_j, for $j = 1, 2, \cdots, J$. The overall test statistic is $T = minP_j$. Because variants within a functional region are usually not independent, we need permutations to

https://doi.org/10.1017/S0016672315000208 Published online by Cambridge University Press

obtain the P-values of the statistic $S_j (j = 1, 2, \cdots, J)$ and the overall test statistic T. The permutations process is the same as that of Lin *et al.* (2014).

Since rare variants and common variants are both likely to be associated with the trait, upweighting the contribution of rare variants and downweighting the contribution of common variants is not appropriate. So we use the same weight scheme as proposed by Ionita-Laza *et al.* (2013) for rare and common variants. For rare variants, we use the weights $w_j = Beta(\mathrm{MAF}_j; 1, 25)$. However, this weight scheme does not work for common variants because it assigns almost zero weight to common variants. For example, $w = 0{\cdot}0004$ for a MAF of 0·30, but $w = 7{\cdot}28$ for a MAF of 0·05. So for common variants, we use the weight $w_j = Beta(\mathrm{MAF}_j; 0.5, 0.5)$, which slowly decreases with increasing MAF. For example, for MAF = 0·05, $w = 1{\cdot}46$, for MAF = 0·10, $w = 1{\cdot}06$, for MAF = 0·30, $w = 0{\cdot}69$, and for MAF = 0·5, $w = 0{\cdot}64$. Our proposed method is referred to as 'RC-ADA', because the per-site P-values of rare and common variants sites are combined adaptively.

3. Simulation studies

(i) *Simulation design*

The GAW17 dataset is used for simulation studies. This dataset contains genotypes of 697 unrelated individuals on 3205 genes. We follow the simulation set-up of Sha *et al.* (2012). Specifically, we choose four genes: *ELAVL4* (gene 1), *MSH4* (gene 2), *PDE4B* (gene 3) and *ADAMTS4* (gene 4) with 10, 20, 30 and 40 variants, respectively. We merge the four genes into a super gene (Sgene) with 100 variants. In our simulation studies, we generate genotypes of n individuals based on the genotypes of 697 individuals in the Sgene. We infer haplotypic phases in the Sgene for the 697 individuals. To generate the genotypes with 100 variants of n individuals, we randomly combine two haplotypes of 1394 haplotypes of the 697 individuals. In the following, we describe how to generate trait values.

To evaluate type I error rate, we generate trait values by using the logistic model:

$$\mathrm{logit}P(Y_i = 1) = \beta_0.$$

β_0 is chosen such that the disease prevalence is 0·05. We estimate the empirical type I error rate as the proportion of P-values less than $\alpha = 0{\cdot}01$ or 0·05.

To evaluate power, we consider two cases: (1) causal variants contain both rare and common variants, (2) all causal variants are rare variants. In case 1, we randomly select one common variant and 30% of all rare variants as causal variants. In fact, our proposed method can be applied to multiple common variants. In case 2, we randomly choose 30% of all rare variants as causal variants. In the two cases, we randomly assign d% of the rare causal variants as deleterious variants, and let the remaining (100–d%) of the rare causal variants be protective variants. The value of d is set to 10, 20, 50, 80 and 100, respectively. For power comparisons, we also consider three different values of RVT (0·01, 0·03 and 0·05). We generate binary traits by:

$$\mathrm{logit}P(Y_i = 1) = \beta_0 + \sum_{i=1}^{n_d} \beta_i^d G_i^d - \sum_{j=1}^{n_p} \beta_j^p G_j^p + \beta_c G_c,$$

where n_d and n_p are the number of deleterious and protective rare variants, respectively. G_c is the genotype of the common causal variant. β_0 is chosen such that the disease prevalence is 0·05. Under the two considered cases, we set the magnitude of each β_j to $a|\log_{10}\mathrm{MAF}_j|$ such that rarer variants have larger effects, where $a = \ln 5/4 = 0{\cdot}402$. In case 2, β_c is 0.

We compare the performance of our proposed method with that of the CMC method (Li & Leal, 2008), the SKAT method (Wu *et al.*, 2011), the VW-TOW method (Sha *et al.*, 2012), and the ADA method (Lin *et al.*, 2014). The ADA method and the VW-TOW method are implemented with their respective R script.

Table 1. *The estimated type I error rates for all tests.*

RVT	α	CMC	SKAT	VW-TOW	ADA	RC-ADA
0·01	0·01	0·009	0·009	0·014	0·012	0·010
	0·05	0·043	0·047	0·047	0·049	0·046
0·03	0·01	0·010	0·013	0·012	0·013	0·011
	0·05	0·042	0·041	0·046	0·034	0·035
0·05	0·01	0·005	0·010	0·007	0·011	0·011
	0·05	0·039	0·052	0·049	0·047	0·050

Note: RVT represents the rare variant threshold; α represents the significance level.

(ii) *Evaluation on type I error rates*

For type I error evaluation, we consider different values of RVT and different significance levels. In each simulation set-up, P-values are estimated by 1000 permutations and type I error rates are evaluated by 1000 replications. Sample size is set at 1000 (500 cases and 500 controls). Table 1 summarizes the estimated type I error rates for given different values of RVT and different significance levels. From this table, we can see that the estimated type I error rates are not significantly different from the nominal levels. So all test methods are valid tests.

(iii) *Power comparisons*

To evaluate the power of the proposed approach, we consider two cases: (1) both rare variants and one

https://doi.org/10.1017/S0016672315000208 Published online by Cambridge University Press

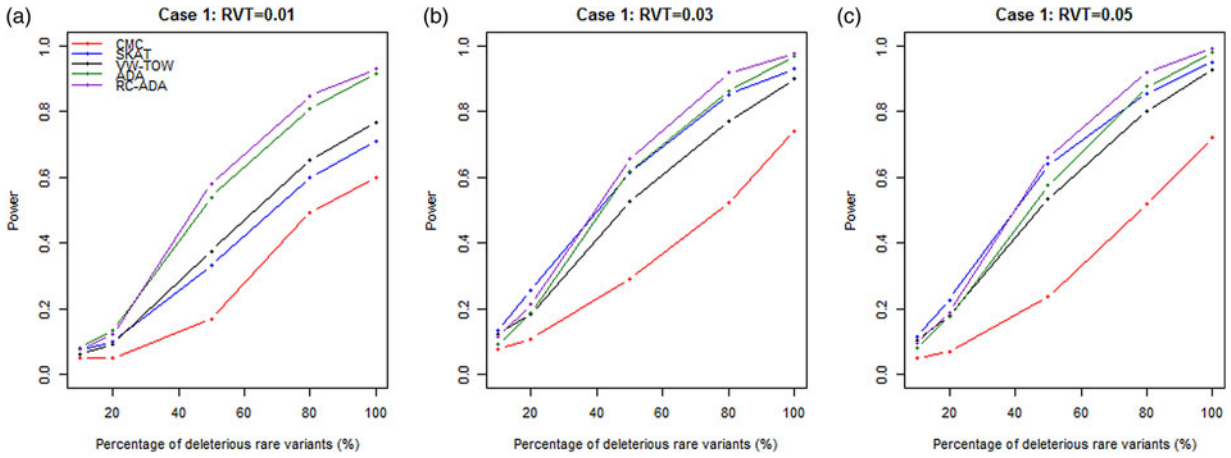

Fig. 1. Power comparisons of five tests for different percentages of deleterious rare variants based on case 1. RVT represents the rare variant threshold for (*a*) 0·01, (*b*) 0·03 and (*c*) 0·05. x-axis represents the percentage of deleterious rare variants. Sample size is 1000. Power is estimated at the 0·05 significance level.

common variant are causal variants, (2) all causal variants are rare variants. In each of two cases, we consider different values of RVT, and different percentages of deleterious rare variants. Sample size is set at 1000 (500 cases and 500 controls). In case 1, we also consider different percentages of neutral variants among all rare variants (10%, 30%, 50% and 70%) and different sample sizes (500, 1000, and 2000). In each simulation scenario, *P*-values are estimated by 1000 permutations and powers are evaluated by 500 replications at a significance level of 0·05.

In Fig. 1, we report the power of the proposed RC-ADA method and of the existing four methods (CMC, SKAT, VW-TOW and ADA) for different percentages of deleterious rare variants based on case 1. Fig.1*a* shows that the RC-ADA and the ADA are much more powerful than the other three tests when RVT is 0·01. Fig.1*b* and *c* show that the RC-ADA, the SKAT and the ADA are much more powerful than the VW-TOW and the CMC. The RC-ADA is the most powerful in many cases. Among all methods, the CMC is the least powerful one. The CMC loses power because it gives common variants the same weights as rare variants, thus common neutral variants will introduce more noise. The RC-ADA is more powerful than the ADA. This is because the ADA only considers rare variants, but the RC-ADA considers both rare and common variants and imposes proper weights. The power of all methods increases when the percentage of deleterious rare variants is increased. The reason for this, pointed out by Wu *et al.* (2011) and Sha *et al.* (2012), is that protective variants imply negative log ORs and lower disease risk and hence lower MAFs in cases and causes more difficulties in observing rare variants in cases. The power of all methods is improved when RVT is larger.

Power comparisons of five methods for different percentages of deleterious rare variants based on case 2 are given in Fig.2. By comparing Fig.2 with Fig.1, we see that patterns of power comparisons based on case 2 are very similar to that based on case 1. This is because we set smaller ORs for common causal variants.

Comparisons of power as a function of percentage of neutral variants among all rare variants based on the case 1 are given in Fig.3. As shown in Fig.3, we see that the RC-ADA and the ADA are more powerful than the other three methods. The RC-ADA is the most powerful method in all the cases. The RC-ADA and the ADA have high power because they can guard against the noise caused by neutral variants by imposing a truncation threshold upon the per-site *P*-values. The power of the RC-ADA, the ADA, the SKAT and the VW-TOW are relatively robust to the increasing of neutral variants, while the power of the CMC decreases rapidly with the increasing of neutral variants.

Power comparisons of the five methods for different sample sizes based on case 1 are given in Fig.4. This figure shows that the power of all methods increases with an increase in sample size.

In summary, the RC-ADA is the most powerful method across a wide of range of scenarios.

4. Discussion

Many diseases are caused by both common and rare variants. However, most of the recently developed methods only detect rare variants. In this paper, we have proposed a powerful RC-ADA method for rare and common causal variant detection. We used extensive simulation studies to compare the performance of the RC-ADA with that of the existing methods. Our simulation results

https://doi.org/10.1017/S0016672315000208 Published online by Cambridge University Press

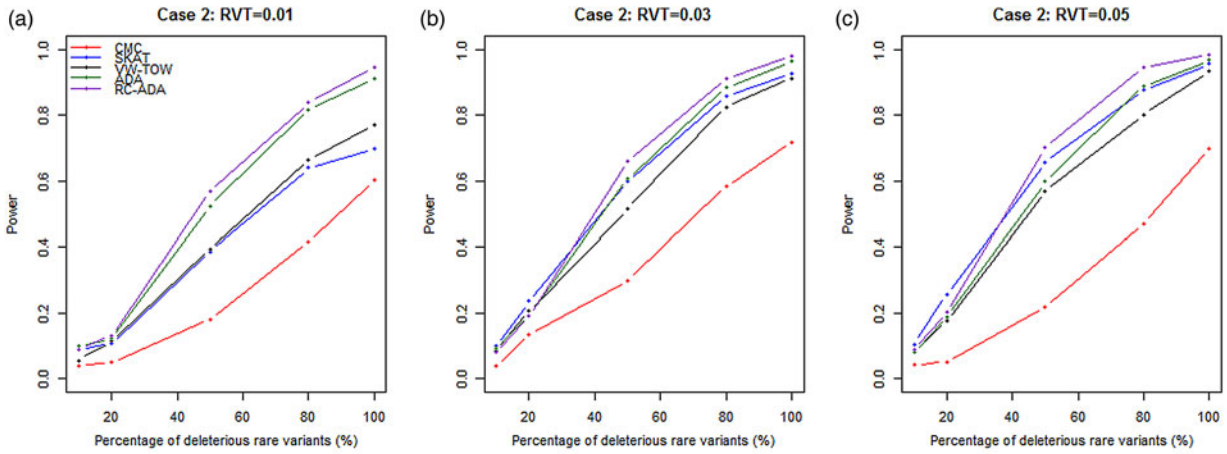

Fig. 2. Power comparisons of five tests for different percentages of deleterious rare variants based on case 2. RVT represents the rare variant threshold for (*a*) 0·01, (*b*) 0·03 and (*c*) 0·05. x-axis represents the percentage of deleterious rare variants. Sample size is 1000. Power is estimated at the 0·05 significance level.

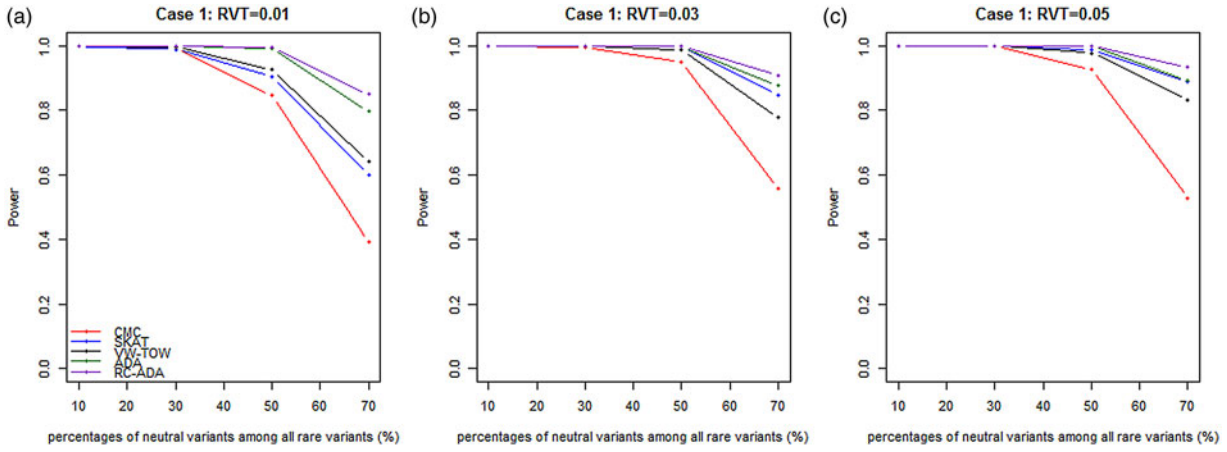

Fig. 3. Power comparisons of five tests for different percentages of neutral variants among all rare variants based on case 1. RVT represents the rare variant threshold for (*a*) 0·01, (*b*) 0·03 and (*c*) 0·05. A total of 80% of rare causal variants are deleterious variants. x-axis represents the percentage of neutral variants among all rare variants. Sample size is 1000. Power is estimated at the 0·05 significance level.

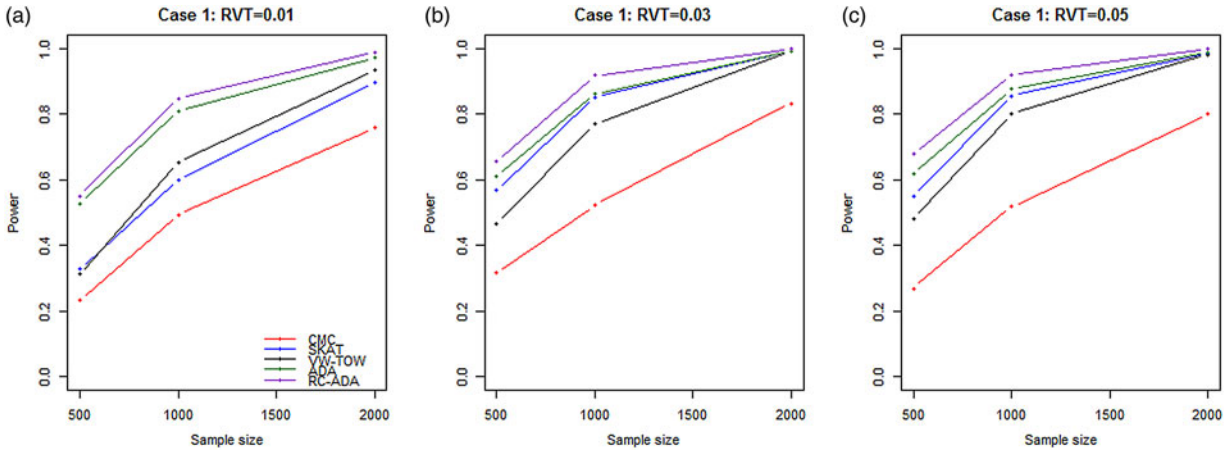

Fig. 4. Power comparisons of five tests for different sample sizes based on case 1. RVT represents the rare variant threshold for (*a*) 0·01, (*b*) 0·03 and (*c*) 0·05. A total of 80% of rare causal variants are deleterious variants. x-axis represents sample sizes. Power is estimated at the 0·05 significance level.

https://doi.org/10.1017/S0016672315000208 Published online by Cambridge University Press

Table 2. *Power(%) of RC-ADA method with three sets of candidate truncation thresholds.*

Candidate *P*-value truncation thresholds	Case 1. *d*(%)				Case 2. *d*(%)			
	20	50	80	100	20	50	80	100
(0.10, 0.11, ⋯, 0.20)	0·214	0·658	0·918	0·978	0·192	0·662	0·914	0·980
(0.05, 0.06, ⋯, 0.25)	0·218	0·704	0·902	0·972	0·218	0·680	0·918	0·980
(0.10, 0.11, ⋯, 0.20, rare; 0.05, 0.06, ⋯, 0.15, common)	0·226	0·644	0·926	0·982	0·210	0·640	0·928	0·984

Note: RVT is set at 0·03. Power is estimated at the 0·05 significance level. *d* represents the percentage of deleterious rare variants.

show that the RC-ADA is the most powerful method for testing rare and common variants in most cases.

For detecting rare variants, many methods put large weights on rare variants and small weights on common variants. Thus, these methods will lose power when the disease is caused by both rare and common variants. By choosing adaptive weights, our proposed RC-ADA shows good performance for detecting rare and common variants.

In our proposed RC-ADA, to guard against the noise caused by the inclusion of neutral variants, we imposed a truncation threshold upon the per-site *P*-values. Instead of fixing a threshold, we search for the optimal threshold among multiple candidate truncation thresholds. In this paper, we consider 11 candidate *P*-value truncation thresholds, 0.10, 0.11, 0.12, ⋯, 0.20. In fact, we also consider two other cases. In the first case, we use 21 candidate *P*-value truncation thresholds, 0.05, 0.06, 0.07, ⋯, 0.25. In the second case, we consider respective *P*-value truncation thresholds for rare variants and common variants. We consider a more stringent threshold for common variants. For example, 0.05, 0.06, 0.07, ⋯, 0.15 for common variants, and 0.10, 0.11, 0.12, ⋯, 0.20 for rare variants. Table 2 lists the power of the RC-ADA with three sets of candidate *P*-value truncation thresholds. Table 2 shows that using the other two cases doesn't contribute a noticeable power gain to the RC-ADA.

The authors would like to thank the joint Editor and referees for comments that greatly improved the presentation of the paper. This research was supported by the National Natural Science Foundation of China (no.11201129). 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 1,000 Genomes Project (http://www.1000genomes.org).

Declaration of Interest

None.

References

Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics* **11**, 773–785.

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

Bodmer, W. & Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics* **40**, 695–701.

Cheung, Y. H., Wang, G., Leal, S. M. & Wang, S. (2012). A fast and noise-resilient approach to detect rare-variant associations with deep sequencing data for complex disorders. *Genetic Epidemiology* **36**, 675–685.

Fang, S., Sha, Q. & Zhang, S. (2012). Two adaptive weighting methods to test for rare variant associations in family-based designs. *Genetic Epidemiology* **36**, 499–507.

Feng, T., Elston, R. C. & Zhu, X. (2011). Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS). *Genetic Epidemiology* **35**, 398–409.

Fisher, R. A. (1922). On the interpretation of x2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* **85**, 87–94.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D. & Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics* **92**, 841–853.

Li, B. & Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American Journal of Human Genetics* **83**, 311–321.

Lin, D. Y. & Tang, Z. Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *American Journal of Human Genetics* **89**, 354–367.

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

Madsen, B. E. & Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics* **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., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E.,

Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A. & Visscher, P. M. (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. & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9**, 356–369.

Morgenthaler, S. & 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). *Mutation Research* **615**, 28–56.

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K. & Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS Genetics* **7**, e1001322.

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

Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L. J. & Sunyaev, S. R. (2010). Pooled association tests for rare variants in exon-resequencing studies. *American Journal of Human Genetics* **86**, 832–838.

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics* **69**, 124–137.

Pritchard, J. K. & Cox, N. J. (2002). The allelic architecture of human disease genes: common disease-common variant … or not? *Human Molecular Genetics* **11**, 2417–2423.

Sha, Q., Wang, S. & Zhang, S. (2013). Adaptive clustering and adaptive weighting methods to detect disease associated rare variants. *European Journal of Human Genetics* **21**, 332–337.

Sha, Q., Wang, X., Wang, X. & Zhang, S. (2012). Detecting association of rare and common variants by testing an optimally weighted combination of variants. *Genetic Epidemiology* **36**, 561–571.

Sha, Q., Zhang, Z. & Zhang, S. (2011). An improved score test for genetic association studies. *Genetic Epidemiology* **35**, 350–359.

Stratton, M. R. & Rahman, N. (2008). The emerging landscape of breast cancer susceptibility. *Nature Genetics* **40**, 17–22.

Teer, J. K. & Mullikin, J. C. (2010). Exome sequencing: the sweet spot before whole genomes. *Human Molecular Genetics* **19**, R145–R151.

Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics* **90**, 7–24.

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

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. & Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* **42**, D1001–D1006.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics* **89**, 82–93.