Genet. Res., Camb. (2016), vol. 98, e1. © Cambridge University Press 2016
doi:10.1017/S0016672315000269

A nonparametric method to test for associations between rare variants and multiple traits

YING ZHOU[1,2], YANGYANG CHENG[1], WENSHENG ZHU[1]* AND QIAN ZHOU[3]
[1] *Key Laboratory for Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China*
[2] *School of Mathematical Sciences, Heilongjiang University, Harbin 150080, China*
[3] *Department of Humanities, Mianyang Vocational and Technical College, Mianyang 621000, China*

(*Received 3 April 2015; revised 18 September 2015; accepted 8 December 2015*)

Summary

More and more rare genetic variants are being detected in the human genome, and it is believed that besides common variants, some rare variants also explain part of the phenotypic variance for human diseases. Due to the importance of rare variants, many statistical methods have been proposed to test for associations between rare variants and human traits. However, in existing studies, most methods only test for associations between multiple loci and one trait; therefore, the joint information of multiple traits has not been considered simultaneously and sufficiently. In this article, we present a study of testing for associations between rare variants and multiple traits, where trait value can be binary, ordinal, quantitative and/or any mixture of them. Based on the method of generalized Kendall's τ, a nonparametric method called NM-RV is proposed. A new kernel function for U-statistic, which could incorporate the information of each rare variant itself, is also presented and is expected to enhance the power of rare variant analysis. We further consider the asymptotic distribution of the proposed association test statistic. Our simulation work suggests that the proposed method is more powerful and robust than existing methods in testing for associations between rare variants and multiple traits, especially for multivariate ordinal traits.

1. Introduction

In genome-wide association studies, thousands of common variants associated with human complex diseases have been successfully identified using statistical methods over the last 10 to 20 years. However, these common variants explain only a small portion of inheritable phenotypic variance for human diseases (Maher, 2008; Manolio *et al.*, 2009; Eichler *et al.*, 2010). We expect that most missing heritability can be explained by low frequency variants. Up to now, it has been widely believed that human complex diseases are likely caused by both common and rare variants (Bodmer & Bonilla, 2008; Ng *et al.*, 2010; Robinson *et al.*, 2014). At the same time, along with the rapid development of next-generation sequencing technologies (Metzker, 2010), more and more rare genetic variants have been detected in the human genome, where rare variants (RVs) are usually defined as having minor allele frequencies (MAFs) of less than 5%.

Although statistical methods have allowed for enormous steps in testing for associations between common variants and complex traits, these methods may lead to larger bias and lower power in detecting rare variants (Li & Leal, 2008). Owing to the convenience of using large data sets of rare variants, many statistical approaches have been proposed to be used when examining rare variants that may be associated with complex traits. Currently, the idea of collapsing a group of rare variants in a gene is widely used in association tests. For instance, Morgenthaler and Thilly (2007) proposed the cohort allelic sums test (CAST); Pan (2009) developed the sum test (SUM); and Madsen and Browning (2009) presented the method of weighted sum statistic (WSS). These collapsing methods are validated to be more powerful than single-marker methods. In addition, several methods of detecting associations of rare variants have been proposed by other groups, in which the direction and magnitude of the effects of causal variants are discussed, including adaptive methods (Zhang *et al.*, 2010 *a*;

* Corresponding author: Dr Wensheng Zhu, School of Mathematics and Statistics, Northeast Normal University, 5268 Renmin Street, Changchun 130024, PR China.
E-mail: wszhu@nenu.edu.cn

https://doi.org/10.1017/S0016672315000269 Published online by Cambridge University Press

Fang *et al.*, 2012), the sequence kernel association test (SKAT; Wu *et al.*, 2011) and the sequence kernel association optimal test (SKAT-O; Lee *et al.*, 2012).

Almost all the existing methods focus on single binary or quantitative phenotype. However, ordinal responses are also very common in the investigations of complex traits, such as mental illnesses or behavioural disorders. Meanwhile, in these investigations multiple traits are often recorded as different types (e.g. binary, ordinal and quantitative). To date, researchers have already provided some methods to test for associations between multiple traits and common variants. For example, Lange *et al.* (2003) proposed the FBAT-GEE method; Zhang *et al.* (2010 *b*) proposed a nonparametric method based on generalized Kendall's τ, and Zhu *et al.* (2012) presented covariate-adjusted association tests based on generalized Kendall's τ. Investigation results show that testing for multiple traits together is more powerful than testing for one single trait at a time in association studies (Zhu & Zhang, 2009; 2013). However, it is not entirely clear as to how beneficial simultaneous testing for multiple traits is in association studies of rare variants. One drawback to consider when using multiple traits to examine rare variants is that common-variant-based approaches could not be used directly.

To circumvent this difficulty and improve power compared to single-trait tests, in this article we provide a nonparametric method to test for associations between rare variants and multiple traits, which is based on generalized Kendall's τ (Zhang *et al.*, 2010 *b*). The traits involved in our method may be binary, ordinal, quantitative and/or any mixture of them. We expect that this approach can combine the weak signals from each variant, which should provide high resolution for detecting associations. With this in mind, we define a new kernel function based on multiple rare variants to measure the genotype dis-similarity of each pair of individuals in generalized Kendall's τ, which could incorporate the information of each rare variant itself. Furthermore, our proposed test statistic has an asymptotic Chi-square distribution. Extensive simulations are performed to compare the proposed method with other existing methods and the simulation results show that our method is effective, powerful and robust for rare variant association analysis. It can control Type I error, has higher power and, therefore, can increase the chance of detecting causal variants.

2. Methods

We consider data collected from n independent subjects. Let $Y_i = (Y_i^{(1)}, \ldots Y_i^{(q)})'$ denote the observed multiple traits, and $G_i = (G_{i1}, \ldots, G_{im})'$ denote the genotypic score vector at m loci for individual i $(i = 1, \ldots, n)$, where $Y_i^{(k)}$ $(k = 1, \ldots, q)$ may be binary, ordinal or quantitative; $G_{il} = 0$, 1 and 2 correspond to genotypes AA, Aa and aa $(l = 1, \ldots, m)$, and the frequency of minor allele a is less than 5%. Assume that the m rare variant loci are independent. We are concerned with the problem of testing for associations between rare variants and multiple traits. In the following, we propose our nonparametric association test method by constructing a U-statistic and further constructing a nonparametric statistic W to test for the associations. We call the proposed method NM-RV for brevity.

(i) *U-statistic*

Similar to the method of Zhang *et al.* (2010 *b*), which is based on generalized Kendall's τ, we propose a U-statistic

$$U = \binom{n}{2}^{-1} \sum_{i<j} F(Y_i, Y_j) K(G_i, G_j) \qquad (1)$$

to measure the correlation between rare variants and multiple traits. For each pair (i, j), $F(Y_i, Y_j)$ and $K(G_i, G_j)$ are the kernel functions that measure the dissimilarities of the traits and of the genotypes between individuals i and j. Following the method of Zhang *et al.* (2010 *b*), let

$$\begin{aligned} F_{ij} &= F(Y_i, Y_j) \\ &= \left(f_1\left(Y_i^{(1)} - Y_j^{(1)}\right), \ldots, f_q\left(Y_i^{(q)} - Y_j^{(q)}\right)\right)', \end{aligned}$$

where function $f_k(\cdot)$ is defined as

$$f_k\left(Y_i^{(k)} - Y_j^{(k)}\right) = Y_i^{(k)} - Y_j^{(k)}, \ k = 1, \ldots, q,$$

if the kth trait is quantitative or binary, and

$$\begin{aligned} f_k\left(Y_i^{(k)} - Y_j^{(k)}\right) &= \mathrm{sign}\left\{Y_i^{(k)} - Y_j^{(k)}\right\} \\ &= \begin{cases} 1, & Y_i^{(k)} - Y_j^{(k)} > 0, \\ -1, & Y_i^{(k)} - Y_j^{(k)} < 0 \ , k = 1, \ldots, q, \\ 0, & Y_i^{(k)} - Y_j^{(k)} = 0, \end{cases} \end{aligned}$$

if the kth trait is ordinal.

Next, we define the kernel function $K(G_i, G_j)$ such that it can be used to measure the dis-similarity of the genotypes for m rare variants between individuals $i$and j. For the purpose of simplicity, we assume that the kernel function $K(G_i, G_j)$ is a summation of $K_l(G_{il}, G_{jl})$ defined at each locus l $(l = 1, \ldots, m)$, that is,

$$K(G_i, G_j) = \sum_{l=1}^{m} K_l(G_{il}, G_{jl}),$$

where $K_l(G_{il}, G_{jl})$ is the kernel function that represents the dis-similarity of genotypes at the lth rare variant for the pair (i, j). Because the effects of all m rare variants on the traits may not be identical in practice, we should define different kernel functions for different rare variants. Then, we need to define the kernel function $K_l(G_{il}, G_{jl})$ for each locus l discriminately.

https://doi.org/10.1017/S0016672315000269 Published online by Cambridge University Press

Unfortunately, the kernel function proposed by Zhang *et al.* (2010 *b*) is not applicable to rare variants owing to the following two shortcomings. First, their kernel function assumes the dis-similarity between genotypes AA and Aa is the same as the dis-similarity between genotypes Aa and aa at a single locus, which is obviously unreasonable when allele a is a rare mutation. Second, their kernel function does not take into account MAFs for different loci, which is essential for association studies of rare variants. In this work, a new kernel function is defined as

$$K_l(G_{il}, G_{jl}) = \log\left(\frac{n_{G_{il}}}{n_{G_{jl}}}\right),$$

where $n_{G_{il}}$ represents the total number of observed genotype G_{il} for all individuals at variant l, and $n_{G_{jl}}$ has analogous explanation. This kernel function, motivated by the shrinkage of entropy-guided distance (EGS) of Jin *et al.* (2014), does take MAFs into account and could incorporate the information of the rare variant itself. We expect that this kernel function possesses very similar advantages as the EGS of Jin *et al.* (2014), although the EGS was proposed to measure the dis-similarity of haplotype pairs.

Replacing the kernel function $K(G_i, G_j)$ in formula (1), the U-statistic is given by

$$U = \binom{n}{2}^{-1} \sum_{i<j} F_{ij} \sum_{l=1}^{m} K_l(G_{il}, G_{jl}).$$

Moreover, the U-statistic can be simplified into the following form

$$U = \sum_{l=1}^{m} U_l,$$

where

$$U_l = \frac{2}{n-1} \sum_{i=1}^{n} \bar{F}_i \log(n_{G_{il}}),$$

is a U-statistic defined for the lth variant, and $\bar{F}_i = \frac{1}{n}\sum_{j=1}^{n} F_{ij}$ (see Appendix 1 for more details). The main difference between our U_l and the U-statistic given by Zhang *et al.* (2010 *b*) is that our U_l is proposed for analysing rare variants using a new kernel function. Our proposed U-statistic is a simple summation of the U-statistics for all rare variants of interest, where the number of rare variants involved in our analysis could be very large.

(ii) *Association test statistic W and its asymptotic*

According to generalized Kendall's τ (Zhang *et al.*, 2010 *b*) and based on the above proposed U-statistic, we define an association test statistic as

$$W = (U - E(U \mid Y))' Var^{-1}(U \mid Y)(U - E(U \mid Y)),$$

where

$$E(U \mid Y) = \frac{2}{n-1} \sum_{i=1}^{n} \bar{F}_i \sum_{l=1}^{m} E(\log(n_{G_{il}}) \mid Y),$$

$$Var(U \mid Y) = \left(\frac{2}{n-1}\right)^2 \sum_{i=1}^{n} \bar{F}_i \bar{F}'_i \sum_{l=1}^{m} Var(\log(n_{G_{il}}) \mid Y).$$

The detailed calculation of $E(U|Y)$ and $Var(U|Y)$ can be found in Appendix 2. Similar to the results of Zhang *et al.* (2010 *b*), it can be determined that under the null hypothesis of no association between rare variants and multiple traits, the statistic W follows an asymptotic Chi-square distribution, where the degrees of freedom are given by the rank of the variance matrix $Var(U|Y)$.

3. Simulation studies

We conduct simulation studies to evaluate the performance of our proposed method (NM-RV). In the simulations, we compare the performance of the proposed method and five other competing tests: SUM, CAST, SKAT, SKAT-O and WSS. Two traits are considered in our comparisons.

(i) *Simulation design*

To comprehensively demonstrate the validity of NM-RV, we conduct two simulations: a simulation based on designed parameters and a simulation based on a real data set. In each simulation we consider the association analyses between multiple rare variants and two kinds of bivariate traits, respectively (i.e. two ordinal traits and a mixture of binary and ordinal traits).

(a) *Simulation 1. Simulation based on designed parameters*

In this simulation, $m(= 20$ and $40)$ rare variants and two ordinal traits, as well as a mixture of binary and ordinal traits are simulated. A total of 12 causal variants out of the 20 rare variants are assigned if $m = 20$, and 20 causal variants out of the 40 rare variants are assigned if $m = 40$. Sample size $n = 500$ is considered. Firstly, genotypes of m rare variant loci are generated independently with the MAF $p_l \sim U(0 \cdot 001, 0 \cdot 01)$, $l = 1, \ldots, m$, where $U(0{\cdot}001, 0{\cdot}01)$ denotes a uniform distribution in the interval (0·001, 0·01). Under the assumption of Hardy–Weinberg equilibrium law, the frequencies of genotypes AA, Aa and aa at locus l are $(1-p_l)^2$, $2p_l(1-p_l)$ and $p_l{}^2$, respectively. Then the genotype of any individual i can be randomly generated according to the probability distribution. Based on the generated genotype (AA, Aa, aa), the corresponding genotypic score G_{il} $(= 0, 1, 2)$ can be recorded. Secondly, the trait value vector

https://doi.org/10.1017/S0016672315000269 Published online by Cambridge University Press

$T_i = \left(T_i^{(1)}, T_i^{(2)}\right)'$ of two quantitative traits of individual i is generated according to the following model,

$$T_i^{(j)} = \mu + \gamma_j' G_i + \varepsilon_{ij}, \quad j = 1, 2,$$

where $(\varepsilon_{i1}, \varepsilon_{i2})' \sim N(0, \boldsymbol{\Sigma})$, and $G_i = (G_{i1}, \cdots, G_{im})'$ is the genotypic score vector at m loci for individual i. We set $\mu = 0, \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \cdot 25 \\ 0 \cdot 25 & 1 \end{pmatrix}$, $\gamma_j = \beta_j \cdot (1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0)'$ if $m = 20$, $\gamma_j = \beta_j \cdot (1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1,0,1,0,0,1,0,1,0,1,1,1,0,0,0,0,0,0,0,0,1,1,0,0,.1)'$ if $m = 40$, where the element 1 in vector γ_j represents that the corresponding locus is a causal variant. In the simulation, the values of β_1 are taken as 0, 0·2, 0·4, 0·6 and 0·8, and we take $\beta_2 = \frac{1}{2}\beta_1$ correspondingly. It is noteworthy that the simulated data when $(\beta_1, \beta_2) = (0, 0)$ are under the null hypothesis and are used to calculate Type I errors, whereas the simulated data when $(\beta_1, \beta_2) \neq (0, 0)$ are under the alternative hypothesis and are used to caulate powers for each method. Lastly, the ordinal (or mixed) trait value vector $Y_i = \left(Y_i^{(1)}, Y_i^{(2)}\right)'$ is generated by discretizing the values of $T_i^{(1)}$ and $T_i^{(2)}$ separately. For simplicity, we set the numbers of categories of $Y_i^{(1)}$ and $Y_i^{(2)}$ to be 3 and 4 for two ordinal traits, respectively. We use the 50 and 67% sample percentiles to discretize $T_i^{(1)}$ and generate $Y_i^{(1)}$, and use the 33, 54 and 75% sample percentiles to discretize $T_i^{(2)}$ and generate $Y_i^{(2)}$. To generate the trait value vector $Y_i = \left(Y_i^{(1)}, Y_i^{(2)}\right)'$ of the mixture of binary and ordinal traits, where $Y_i^{(1)}$ $(= 0,\ 1)$ denotes the binary trait value and $Y_i^{(2)}$ $(= 1,\ 2,\ 3,\ 4)$ denotes the ordinal trait value, we use the 40% sample percentile to discretize $T_i^{(1)}$ and generate $Y_i^{(1)}$, and still use the 33, 54 and 75% sample percentiles to discretize $T_i^{(2)}$ and generate $Y_i^{(2)}$. Thus the simulated data including m rare variants and two ordinal traits, as well as the mixture of binary and ordinal traits are obtained.

(b) *Simulation 2. Simulation based on a real data set*

To better show the performance of the proposed method, we apply the proposed NM-RV and the other five methods to real genotype data from GAW17, which are extracted from the sequence alignment files provided by the 1000 Genomes Project for their pilot3 study (http://www.1000genomes.org).

We chose the real genotype data of 697 unrelated individuals from GAW17, and chose the *TG* and *COL6A3* genes as candidate genes. The *TG* gene encodes thyroglobulin, and mutation of the *TG* gene may cause hypothyroidism and autoimmune disorders (Maierhaba *et al.*, 2008); while the *COL6A3* gene encodes one component of collagen VI, and mutation of this gene will cause the occurrence of collagen disease and myopathy dystrophy (Baker *et al.*, 2005). The *TG* gene has 146 SNPs and 113 out of the 146 SNPs are rare (MAF <5%), while the *COL6A3* gene has 187 SNPs and 143 out of the 187 SNPs are rare. For each of these two genes, we randomly chose the genotype data of 20 and 40 rare variants of the 697 individuals, and we assumed that all the causal variants in the selected rare variants had the same direction of effects. To perform simulation studies based on the real genotype data set, the values of traits (two ordinal traits or a mixture of binary and ordinal traits) of each individual were simulated in the same way as in Simulation 1.

For each of the two above simulations, the nominal significance levels 0·01 and 0·05 were used. The six methods (NM-RV, SUM, CAST, SKAT, SKAT-O and WSS) were used to analyse the simulated data. For Type I error evaluation and power comparisons, the simulation results were obtained from 1000 replications for all six methods. When using the other five competing methods to analyse the bivariate traits, we first tested for each trait separately, then applied Bonferroni correction to adjust for corresponding multiple testing. It should be pointed out that the significance level of each single-trait test was set to 0·01/2 and 0·05/2 based on Bonferroni adjustment for the other five competing tests, and each single-trait test was counted on its own, giving rise to 2000 tests under the null hypothesis.

(ii) *Simulation results*

(a) *Evaluation of Type I errors*

The simulated data were generated under the null hypothesis, that is, there exists no association between the rare variants and the bivariate traits when calculating Type I errors. The results of the estimated Type I errors for the two simulations are listed in Tables 1–6.

Tables 1 and 2 show the estimated Type I errors of the six methods at different nominal significance levels when analysing the two kinds of bivariate traits in Simulation 1. The traits considered in Table 1 are the mixture of binary and ordinal ones. The upper part of Table 1 lists the estimated results at the nominal significance level of 0·01, and the lower part corresponds to the nominal significance level of 0·05. From Table 1 we can see that each estimated Type I error was very close to the corresponding nominal significance level. The Type I errors can be well controlled by our proposed method (NM-RV) at different nominal significance levels for each of the two settings of rare variants. The other five methods can also control the estimated Type I errors. By comparison, the estimated Type I errors of the SKAT and the SKAT-O seem a little lower than those of the other methods, that is to say the two test methods are somewhat conservative. Table 2 lists the estimated Type I errors of the six methods when both of the two

https://doi.org/10.1017/S0016672315000269 Published online by Cambridge University Press

Table 1. *Estimated Type I errors of the six methods for a mixture of binary and ordinal traits in simulation 1.*

$n = 500$	Test	20 RVs	40 RVs
$\alpha = 0{\cdot}01$	NM-RV	0·012	0·009
	SUM	0·013	0·009
	CAST	0·008	0·011
	SKAT	0·005	0·009
	SKAT-O	0·008	0·004
	WSS	0·015	0·015
$\alpha = 0{\cdot}05$	NM-RV	0·053	0·051
	SUM	0·054	0·058
	CAST	0·051	0·051
	SKAT	0·027	0·030
	SKAT-O	0·043	0·040
	WSS	0·063	0·059

Table 2. *Estimated Type I errors of the six methods for two ordinal traits in simulation 1.*

$n = 500$	Test	20 RVs	40 RVs
$\alpha = 0{\cdot}01$	NM-RV	0·007	0·011
	SUM	0·011	0·008
	CAST	0·009	0·008
	SKAT	0·006	0·003
	SKAT-O	0·004	0·006
	WSS	0·014	0·010
$\alpha = 0{\cdot}05$	NM-RV	0·049	0·048
	SUM	0·052	0·050
	CAST	0·034	0·041
	SKAT	0·034	0·025
	SKAT-O	0·041	0·031
	WSS	0·048	0·046

Table 3. *Estimated Type I errors of the six methods in the association studies of the* TG *gene and a mixture of binary and ordinal traits at* $\alpha = 0{\cdot}05$ *in simulation 2.*

$n = 697$	Test	20 RVs	40 RVs
$\alpha = 0{\cdot}05$	NM-RV	0·050	0·049
	SUM	0·058	0·045
	CAST	0·048	0·044
	SKAT	0·033	0·025
	SKAT-O	0·044	0·035
	WSS	0·055	0·045

traits are ordinal, and it is not hard to see that the characteristics of the results are similar to those exhibited in the association analyses of the mixture of binary and ordinal traits.

Table 4. *Estimated Type I errors of the six methods in the association studies of the* TG *gene and two ordinal traits at* $\alpha = 0{\cdot}05$ *in simulation 2.*

$n = 697$	Test	20 RVs	40 RVs
$\alpha = 0{\cdot}05$	NM-RV	0·047	0·054
	SUM	0·046	0·039
	CAST	0·039	0·037
	SKAT	0·044	0·033
	SKAT-O	0·040	0·041
	WSS	0·057	0·043

Table 5. *Estimated Type I errors of the six methods in the association studies of the* COL6A3 *gene and a mixture of binary and ordinal traits at* $\alpha = 0{\cdot}05$ *in simulation 2.*

$n = 697$	Test	20 RVs	40 RVs
$\alpha = 0{\cdot}05$	NM-RV	0·048	0·055
	SUM	0·036	0·044
	CAST	0·032	0·031
	SKAT	0·030	0·032
	SKAT-O	0·029	0·036
	WSS	0·038	0·037

Table 6. *Estimated Type I errors of the six methods in the association studies of the* COL6A3 *gene and two ordinal traits at* $\alpha = 0{\cdot}05$ *in simulation 2.*

$n = 697$	Test	20 RVs	40 RVs
$\alpha = 0{\cdot}05$	NM-RV	0·052	0·052
	SUM	0·042	0·051
	CAST	0·033	0·037
	SKAT	0·038	0·030
	SKAT-O	0·040	0·036
	WSS	0·045	0·041

Tables 3 and 4 present the estimated Type I errors of the six methods for the TG gene when the significance levelα is equal to 0·05 in Simulation 2, and the estimated Type I errors of various methods for the COL6A3 gene in Simulation 2 are listed in Tables 5 and 6. It can be seen that similar conclusions can be drawn, while this time CAST seems a little conservative alongside SKAT and SKAT-O. In addition, from Tables 1–6 we can see that the number of rare variants has little impact on Type I errors for each method, as does the proportion of causal variants.

(b) *Power comparison*

In order to better show the advantages of NM-RV in the association analysis between the rare variants and the multiple traits, we generated data under the alternative hypothesis and calculated test powers of each method under each setting. The simulation results of power comparisons for the six methods are listed in Tables 7–12.

Tables 7 and 8 present the power comparison results of the six methods for a mixture of binary and ordinal traits, and two ordinal traits in simulation 1. From

Table 7. *Power comparisons of the six methods for a mixture of binary and ordinal traits in simulation 1.*

	Number of RVs	Test	$\beta_1 = 0{\cdot}8$ $\beta_2 = 0{\cdot}4$	$\beta_1 = 0{\cdot}6$ $\beta_2 = 0{\cdot}3$	$\beta_1 = 0{\cdot}4 \beta_1 = 0{\cdot}2$	$\beta_1 = 0{\cdot}2$ $\beta_2 = 0{\cdot}1$
$\alpha = 0{\cdot}01$	20 (12 causal)	NM-RV	0·794	0·468	0·181	0·029
		SUM	0·392	0·229	0·087	0·020
		CAST	0·369	0·220	0·084	0·018
		SKAT	0·144	0·047	0·008	0·002
		SKAT-O	0·434	0·220	0·069	0·013
		WSS	0·382	0·222	0·086	0·019
	40 (20 causal)	NM-RV	0·912	0·674	0·273	0·058
		SUM	0·529	0·341	0·140	0·026
		CAST	0·458	0·274	0·115	0·022
		SKAT	0·216	0·074	0·016	0·005
		SKAT-O	0·631	0·347	0·121	0·023
		WSS	0·501	0·322	0·132	0·022
$\alpha = 0{\cdot}05$	20 (12 causal)	NM-RV	0·908	0·721	0·387	0·139
		SUM	0·562	0·403	0·215	0·072
		CAST	0·535	0·368	0·184	0·061
		SKAT	0·351	0·157	0·053	0·015
		SKAT-O	0·654	0·406	0·186	0·049
		WSS	0·558	0·391	0·198	0·070
	40 (20 causal)	NM-RV	0·968	0·864	0·542	0·164
		SUM	0·684	0·520	0·291	0·094
		CAST	0·615	0·441	0·233	0·076
		SKAT	0·446	0·202	0·067	0·023
		SKAT-O	0·792	0·558	0·265	0·071
		WSS	0·668	0·488	0·267	0·088

the two tables we can draw several conclusions. The power of NM-RV is higher than that of the other methods in any situation. With an increase in the number of rare variants or the number of true causal variants, the power of NM-RV becomes much higher. When the number of rare variants $m = 40$ is considered, although the proportion of causal variants is less than that of $m =$ 20, the power attains the greatest value (0·993) when (β_1, β_2) =(0·8, 0·4) at α = 0·05 (Table 8). Besides, a common tendency of the six methods is that the power increases with an increase in the values of β_1 and β_2. Because large β_1 and β_2 values correspond to high heritability, the simulated results confirm that heritability is an important factor that has significant impact on the test powers. By comparing the details of the simulation results in Tables 7 and 8, we found that in each case with the same parameters, the power of the proposed NM-RV is higher when the two traits are ordinal than when the traits are mixed.

Tables 9 and 10 list the simulation results of power comparisons for the TG gene when considering associations with two bivariate traits in simulation 2. We also consider two situations: 12 out of the 20 rare variants are causal variants and 20 out of the 40 rare variants are causal. Significance level is taken as 0·05. At this time, we yield the same conclusions as the aforementioned analysis, that is, NM-RV has the highest power in any simulation situation; the larger the number of rare variants or the number of the true causal variants is, the higher the power will be for each method, but the power of NM-RV is still the highest; and the proposed NM-RV is more powerful in the association analysis between rare variants and two ordinal traits.

In the power comparisons for the *COL6A3* gene in simulation 2, we consider three situations: 12 out of the 20 rare variants are causal variants, 12 out of the 40 rare variants are causal, and 20 out of the 40 rare variants are causal. From Tables 11 and 12 it is not hard to see the same conclusion; that the power of NM-RV is still the highest out of all of the six methods. Besides, as expected, the power in the situation of 20 causal variants is higher than in the situation of 12 causal variants when the number of rare variants $m = 40$ is considered; the power in the situation of 20 rare variants is higher than in the situation of 40 rare variants when the number of causal variants is 12, that is to say the proportion of causal variants is an important factor that affects power.

All in all, through comparing Type I errors and powers of the six methods, we validate that our proposed NM-RV is effective, powerful and robust; it does not matter if the genotype data $m = 40$ are randomly generated or are based on real data. When the multivariate traits are all ordinal, the advantages of the proposed method are more obvious. In fact, the main reason is that the proposed method sufficiently mines the joint information of multiple traits (even if partially or approximately true information), which helps to enhance the power of identifying associations between rare variants and multiple traits.

Table 8. *Power comparisons of the six methods for two ordinal traits in simulation 1.*

	Number of RVs	Test	$\beta_1 = 0{\cdot}8$ $\beta_2 = 0{\cdot}4$	$\beta_1 = 0{\cdot}6$ $\beta_1 = 0{\cdot}3$	$\beta_1 = 0{\cdot}4$ $\beta_2 = 0{\cdot}2$	$\beta_1 = 0{\cdot}2$ $\beta_2 = 0{\cdot}1$
$\alpha = 0{\cdot}01$	20 (12 causal)	NM-RV	0·877	0·648	0·272	0·062
		SUM	0·439	0·281	0·106	0·030
		CAST	0·403	0·237	0·097	0·023
		SKAT	0·401	0·200	0·055	0·008
		SKAT-O	0·564	0·359	0·138	0·026
		WSS	0·434	0·264	0·109	0·032
	40 (20 causal)	NM-RV	0·973	0·813	0·420	0·074
		SUM	0·564	0·394	0·171	0·033
		CAST	0·481	0·311	0·124	0·021
		SKAT	0·501	0·275	0·064	0·010
		SKAT-O	0·704	0·499	0·221	0·029
		WSS	0·537	0·369	0·152	0·027
$\alpha = 0{\cdot}05$	20 (12 causal)	NM-RV	0·957	0·806	0·509	0·179
		SUM	0·586	0·436	0·236	0·078
		CAST	0·552	0·387	0·193	0·069
		SKAT	0·554	0·345	0·146	0·039
		SKAT-O	0·718	0·516	0·275	0·082
		WSS	0·577	0·421	0·217	0·079
	40 (20 causal)	NM-RV	0·993	0·934	0·660	0·200
		SUM	0·698	0·546	0·324	0·100
		CAST	0·631	0·462	0·250	0·076
		SKAT	0·634	0·432	0·170	0·035
		SKAT-O	0·818	0·642	0·369	0·097
		WSS	0·681	0·517	0·300	0·089

Table 9. *Power comparisons of the six methods in the association studies of the* TG *gene and a mixture of binary and ordinal traits at* $\alpha = 0{\cdot}05$ *in simulation 2.*

	Number of RVs	Test	$\beta_1 = 0{\cdot}8$ $\beta_2 = 0{\cdot}4$	$\beta_1 = 0{\cdot}6$ $\beta_2 = 0{\cdot}3$	$\beta_1 = 0{\cdot}4$ $\beta_2 = 0{\cdot}2$	$\beta_1 = 0{\cdot}2$ $\beta_2 = 0{\cdot}1$
$\alpha = 0{\cdot}05$	20 (12 causal)	NM-RV	0·819	0·568	0·292	0·109
		SUM	0·520	0·352	0·171	0·060
		CAST	0·449	0·280	0·127	0·042
		SKAT	0·286	0·129	0·058	0·030
		SKAT-O	0·568	0·332	0·138	0·047
		WSS	0·479	0·299	0·139	0·048
	40 (20 causal)	NM-RV	0·888	0·688	0·356	0·116
		SUM	0·564	0·389	0·180	0·055
		CAST	0·555	0·377	0·160	0·050
		SKAT	0·315	0·151	0·050	0·022
		SKAT-O	0·642	0·396	0·162	0·046
		WSS	0·564	0·391	0·180	0·053

4. Discussion

With the innovation and development of biotechnology, it is now possible to obtain a huge amount of genetic data, among which is a massive amount of rare variant data, which impels us to find new statistical methods to be used in genetic association studies. Since comorbidity is common in mental illness and behaviour disorders, researchers are beginning to study the associations between multiple traits and genetic loci. Zhang *et al.* (2010 *b*) proposed a method to test for associations between multiple traits and marker loci based on generalized Kendall's τ. Referring to the method of Zhang *et al.* (2010 *b*), we proposed a nonparametric approach to test for associations between rare variants and multivariate phenotypes. The reason that we adopted a nonparametric statistical method was to avoid problems caused by parameter models, such as overfitting and strong collinearity. The more parameters used in the model, the more stable the model will be; however, it will bring difficulties to parameter estimation and lead to greater calculation burden. In the simulation studies, we used both simulated genotype data and real genotype data from GAW17, among which we analysed the *TG* and *COL6A3* genes. The simulation results

https://doi.org/10.1017/S0016672315000269 Published online by Cambridge University Press

Table 10. *Power comparisons of the six methods in the association studies of the* TG *gene and two ordinal traits at $\alpha = 0{\cdot}05$ in simulation 2.*

	Number of RVs	Test	$\beta_1 = 0{\cdot}8\ \beta_2 = 0{\cdot}4$	$\beta_1 = 0{\cdot}6\ \beta_2 = 0{\cdot}3$	$\beta_1 = 0{\cdot}4\ \beta_2 = 0{\cdot}2$	$\beta_1 = 0{\cdot}2\ \beta_2 = 0{\cdot}1$
$\alpha = 0{\cdot}05$	20 (12 causal)	NM-RV	0·934	0·735	0·394	0·131
		SUM	0·543	0·366	0·164	0·051
		CAST	0·479	0·298	0·134	0·036
		SKAT	0·533	0·309	0·137	0·050
		SKAT-O	0·700	0·473	0·221	0·067
		WSS	0·531	0·351	0·165	0·053
	40 (20 causal)	NM-RV	0·968	0·829	0·477	0·149
		SUM	0·593	0·415	0·195	0·058
		CAST	0·604	0·440	0·218	0·070
		SKAT	0·547	0·336	0·126	0·037
		SKAT-O	0·707	0·532	0·248	0·071
		WSS	0·599	0·431	0·214	0·069

Table 11. *Power comparisons of the six methods in the association studies of the* COL6A3 *gene and a mixture of binary and ordinal traits at* $\alpha = 0{\cdot}05$ *in simulation 2.*

	Number of RVs	Test	$\beta_1 = 0{\cdot}8\ \beta_2 = 0{\cdot}4$	$\beta_1 = 0{\cdot}6\ \beta_2 = 0{\cdot}3$	$\beta_1 = 0{\cdot}4\ \beta_2 = 0{\cdot}2$	$\beta_1 = 0{\cdot}2\ \beta_2 = 0{\cdot}1$
$\alpha = 0{\cdot}05$	20 (12 causal)	NM-RV	0·877	0·651	0·348	0·108
		SUM	0·521	0·344	0·166	0·046
		CAST	0·548	0·355	0·166	0·049
		SKAT	0·211	0·076	0·034	0·017
		SKAT-O	0·567	0·342	0·142	0·038
		WSS	0·549	0·362	0·176	0·058
	40 (12 causal)	NM-RV	0·465	0·315	0·178	0·088
		SUM	0·267	0·177	0·088	0·043
		CAST	0·203	0·124	0·060	0·031
		SKAT	0·163	0·080	0·038	0·023
		SKAT-O	0·298	0·170	0·083	0·036
		WSS	0·218	0·140	0·070	0·040
	40 (20 causal)	NM-RV	0·901	0·683	0·367	0·141
		SUM	0·557	0·385	0·202	0·068
		CAST	0·430	0·266	0·129	0·043
		SKAT	0·244	0·112	0·053	0·024
		SKAT-O	0·613	0·387	0·178	0·054
		WSS	0·456	0·303	0·151	0·054

suggested that our proposed method outperforms the existing methods when multiple traits are analysed jointly, especially for ordinal traits.

In our method, we first defined a new kernel function to measure the difference of multiple rare variants between individual pairs. Then we constructed a U-statistic. By calculating the conditional expectation and variance of the U-statistic under the condition of traits, we finally proposed an association test statistic W, which has an asymptotic Chi-square distribution. Therefore we can easily calculate p-values in the simulations based on the asymptotic distribution, one advantage of which is the saving of time.

Meanwhile, our method can be easily extended to analyse family-based data, as well as to consider cases with covariates such as gender, age and environmental factors. In this article, we assume that all rare variant loci that we considered are independent and each locus has equal weight in the kernel function $K(G_i,\ G_j)$. In fact, we can also consider how to add a suitable weight ω_l to the kernel function $K(G_i,\ G_j)$ to better discriminate each locus, so that the improved association test statistic W based on the new kernel function $K\big(G_i, G_j\big) = \sum_{l=1}^{m} \omega_l\, K_l\big(G_{il}, G_{jl}\big)$ may gain higher power in the association test between rare variants and multiple traits. The weighted idea in Madsen & Browning (2009) may be a better reference for us to use when defining the weight ω_l. Of course, how to choose an optimal weight ω_l is one of the issues we will continue to consider.

Although the simulation results indicated that our method is better than other existing methods when analysing multiple traits, the proposed method still has some shortcomings. A limitation of our proposed test is that it may lose power when all rare variants influence traits in different directions. Besides, in this

Table 12. *Power comparisons of the six methods in the association studies of the COL6A3 gene and two ordinal traits at α = 0·05 in simulation 2.*

	Number of RVs	Test	$\beta_1 = 0{\cdot}8$ $\beta_2 = 0{\cdot}4$	$\beta_1 = 0{\cdot}6$ $\beta_2 = 0{\cdot}3$	$\beta_1 = 0{\cdot}4$ $\beta_2 = 0{\cdot}2$	$\beta_1 = 0{\cdot}2$ $\beta_2 = 0{\cdot}1$
$\alpha = 0{\cdot}05$	20 (12 causal)	NM-RV	0·963	0·787	0·431	0·152
		SUM	0·556	0·397	0·188	0·052
		CAST	0·599	0·431	0·208	0·066
		SKAT	0·476	0·267	0·107	0·035
		SKAT-O	0·663	0·474	0·220	0·063
		WSS	0·586	0·420	0·205	0·067
	40 (12 causal)	NM-RV	0·652	0·412	0·227	0·100
		SUM	0·341	0·221	0·119	0·059
		CAST	0·269	0·164	0·084	0·037
		SKAT	0·411	0·236	0·092	0·038
		SKAT-O	0·485	0·296	0·135	0·052
		WSS	0·275	0·173	0·087	0·040
	40 (20 causal)	NM-RV	0·977	0·832	0·496	0·167
		SUM	0·608	0·451	0·248	0·085
		CAST	0·501	0·337	0·168	0·051
		SKAT	0·533	0·323	0·136	0·044
		SKAT-O	0·708	0·528	0·269	0·078
		WSS	0·507	0·348	0·175	0·057

article we did not consider the problem of population stratification and the interactions among genes. We will carry out further investigations and develop new methods to deal with these issues in future work.

Appendix 1

The simplification of the proposed U-statistic

According to the definition of the proposed U-statistic, and further replacing the expression of the kernel function $K_l(G_{il}, G_{jl})$ in the statistic, we have

$$U = \binom{n}{2}^{-1} \sum_{i<j} F_{ij} \sum_{l=1}^{m} K_l(G_{il}, G_{jl})$$

$$= \frac{2}{n(n-1)} \sum_{l=1}^{m} \sum_{i<j} F_{ij} K_l(G_{il}, G_{jl})$$

$$= \frac{2}{n(n-1)} \sum_{l=1}^{m} \sum_{i<j} F_{ij} \log\left(\frac{n_{G_{il}}}{n_{G_{jl}}}\right)$$

$$= \frac{2}{n(n-1)} \sum_{l=1}^{m} \sum_{i<j} F_{ij} \left(\log(n_{G_{il}}) - \log(n_{G_{jl}})\right)$$

$$= \frac{2}{n(n-1)} \sum_{l=1}^{m} \sum_{i=1}^{n} \log(n_{G_{il}})$$

$$\sum_{j=1}^{n} F_{ij.} \left(\text{Since } F_{ij} = -F_{ji} \text{ and } F_{ii} = 0\right).$$

Let $\overline{F}_i = \frac{1}{n} \sum_{j=1}^{n} F_{ij}$, then

$$U = \frac{2}{(n-1)} \sum_{l=1}^{m} \sum_{i=1}^{n} \bar{F}_i \log(n_{G_{il}}).$$

Further let

$$U_l = \frac{2}{n-1} \sum_{i=1}^{n} \bar{F}_i \log(n_{G_{il}}),$$

so, the U-statistic can be expressed as

$$U = \sum_{l=1}^{m} U_l.$$

Appendix 2

The calculating process of E(U|Y) and Var(U|Y) under the null hypothesis

Because the rare variant loci are independent, it is easy to obtain

$$E(U|Y) = \frac{2}{n-1} \sum_{i=1}^{n} \bar{F}_i \sum_{l=1}^{m} E(\log(n_{G_{il}})|Y),$$

$$Var(U|Y) = \left(\frac{2}{n-1}\right)^2 \sum_{i=1}^{n} \bar{F}_i \bar{F}'_i \sum_{l=1}^{m} Var(\log(n_{G_{il}})|Y).$$

So we only need to calculate $E(\log(n_{G_{il}})|Y)$ and $Var(\log(n_{G_{il}})|Y)$. Under the null hypothesis and the assumption of Hardy–Weinberg equilibrium law for each locus, we have

$$E(\log(n_{G_{il}})|Y) = P(G_{il} = 0) \log n_0 + P(G_{il} = 1) \log n_1 + P(G_{il} = 2) \log n_2$$

$$= (1 - p_l)^2 \log n_0 + 2p_l(1 - p_l) \log n_1 + p_l{}^2 \log n_2$$

$$= \log n + 2(1 - p_l)^2 \log(1 - p_l),$$

$$+ 2p_l(1 - p_l) \log[2p_l(1 - p_l)] + 2p_l{}^2 \log p_l$$

https://doi.org/10.1017/S0016672315000269 Published online by Cambridge University Press

where $n_{G_{il}}$ represents the total number of observed genotype G_{il} (= 0, 1, 2) for all individuals at variant l. Similarly,

$$\begin{aligned}
&Var\left(\log\left(n_{G_{il}}\right)|Y\right)\\
&= E\left(\log^2\left(n_{G_{il}}\right)|Y\right) - \left[E\left(\log\left(n_{G_{il}}\right)|Y\right)\right]^2\\
&= \left[\log(1-p_l)\right]^2\left[4(1-p_l)^2\left(1-(1-p_l)^2\right)\right]\\
&+ \left[\log 2p_l(1-p_l)\right]^2\left[2p_l(1-p_l)\left[1-2p_l(1-p_l)\right]\right]\\
&+ (\log p_l)^2\left[4p_l{}^2(1-p_l{}^2)\right]\\
&- 8p_l(1-p_l)^3\log(1-p_l)\cdot\log 2p_l(1-p_l)\\
&- 8p_l{}^2(1-p_l)^2\log p_l\cdot\log(1-p_l)\\
&- 8p_l{}^3(1-p_l)\log p_l\cdot\log 2p_l(1-p_l)
\end{aligned},$$

where p_l is the MAF of the lth rare variant.

The authors are indebted to the editor and the referees for their constructive comments and suggestions, which made the improvement of the original manuscript of this article possible. This research was supported by the National Natural Science Foundation of China (nos. 11371083, 11001044 and 11201129), the Natural Science Foundation of Jilin Province (no. 201215007), the Fundamental Research Funds for the Central Universities (no. 11CXPY007), China Scholarship Council (201406625026), Program for Changjiang Scholars and Innovative Research Team in University, the Natural Science Foundation of Heilongjiang Province (A201207), the Scientific Research Foundation of Department of Education of Heilongjiang Province (no. 1253G044) and the Science and Technology Innovation Team in Higher Education Institutions of Heilongjiang Province (no. 2014TD005).

Declaration of interest

None.

References

Baker, N. L., Mörgelin, M., Peat, R., Goemans, N., North, K. N., Bateman, J. F. & Lamandé, S. R. (2005). Dominant collagen VI mutations are a common cause of Ullrich congenital muscular dystrophy. *Human Molecular Genetics* **14**, 279–293.

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

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H. & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics* **11**, 446–450.

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

Jin, L. N., Zhu, W. S., Yu, Y. Q., Kou, C., Meng, X., Tao, Y. & Guo, J. (2014). Nonparametric tests of associations with disease based on U-Statistic. *Annals of Human Genetics* **78**, 141–153.

Lange, C., Silverman, E. K., Xu, X., Weiss, S. T. & Laird, N. M. (2003). A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics* **4**, 195–206.

Lee, S., Wu, M. & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **4**, 762–775.

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

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.

Maierhaba, M., Zhang, J. A., Yu, Z. Y., Wang, Y., Xiao, W. X., Quan, Y. & Dong, B. N. (2008). Association of the thyroglobulin gene polymorphism with autoimmune thyroid disease in Chinese population. *Endocrine* **33**, 294–299.

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., McCarroll, S. A. & Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature* **461**, 747–753.

Metzker, M. L. (2010). Sequencing technologies – the next generation. *Nature Reviews Genetics* **11**, 31–46.

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.

Ng, S. B., Buckingham, K. J. & Lee, C. (2010). Exome sequencing identifies the cause of a Mendelian disorder. *Nature Genetics* **42**, 30–35.

Pan, W. (2009). Asymptotic tests of association with multiple SNP in linkage disequilibrium. *Genetic Epidemiology* **5**, e1000384.

Robinson, M. R., Wray, N. R. & Visscher, P. M. (2014). Explaining additional genetic variation in complex traits. *Trends in Genetics* **30**, 124–132.

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

Zhang, L., Pei, Y. F., Li, J., Papasian, C. J. & Deng, H. W. (2010*a*). Efficient utilization of rare variants for detection of disease-related genomic regions. *PLoS ONE* **5**, e14288.

Zhang, H. P., Liu, C. T. & Wang, X. Q. (2010*b*). An association test for multiple traits based on the generalized Kendall's tau. *Journal of the American Statistical Association* **105**, 473–481.

Zhu, W. S. & Zhang, H. P. (2009). Why do we test multiple traits in genetic association studies? *Journal of the Korean Statistical Society* **38**, 1–10.

Zhu, W. S. & Zhang, H. P. (2013). A nonparametric regression method for multiple longitudinal phenotypes using multivariate adaptive splines. *Frontiers of Mathematics in China* **3**, 731–743.

Zhu, W. S., Jiang, Y., & Zhang, H. P. (2012). Nonparametric covariate-adjusted association tests based on the generalized Kendall's tau. *Journal of the American Statistical Association* **107**, 1–11.

https://doi.org/10.1017/S0016672315000269 Published online by Cambridge University Press