Reexamining Dis/Similarity-Based Tests for Rare-Variant Association with Case-Control Samples

Charlotte Wang,* Jung-Ying Tzeng,^{†,*,§,***,1} Pei-Zhen Wu,[†] Martin Preisig,^{††} and Chuhsing Kate Hsiao^{†,1}

^{*}Department of Mathematics, Tamkang University, New Taipei City 251, Taiwan, [†]Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei 10055, Taiwan, [‡]Department of Statistics, and [§]Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina 27695, ^{**}Department of Statistics, National Cheng-Kung University, Tainan 701, Taiwan, and ^{††}Department of Psychiatry, University Hospital of Lausanne, Site de Cery, 1008 Lausanne, Switzerland

ORCID ID: 0000-0003-4905-8484 (C.K.H.)

ABSTRACT A properly designed distance-based measure can capture informative genetic differences among individuals with different phenotypes and can be used to detect variants responsible for the phenotypes. To detect associated variants, various tests have been designed to contrast genetic dissimilarity or similarity scores of certain subject groups in different ways, among which the most widely used strategy is to quantify the difference between the within-group genetic dissimilarity/similarity (*i.e.*, case-case and control-control similarities) and the between-group dissimilarity/similarity (*i.e.*, case-control similarities). While it has been noted that for common variants, the within-group and the between-group measures should all be included; in this work, we show that for rare variants, comparison based on the two within-group measures can more effectively quantify the genetic difference between cases and controls. The between-group measure tends to overlap with one of the two within-group measures for rare variants, although such overlap is not present for common variants. Consequently, a dissimilarity or similarity test that includes the between-group information tends to attenuate the association signals and leads to power loss. Based on these findings, we propose a dissimilarity test that compares the degree of SNP dissimilarity within cases to that within controls to better characterize the difference between two disease phenotypes. We provide the statistical properties, asymptotic distribution, and computation details for a small sample size of the proposed test. We use simulated and real sequence data to assess the performance of the proposed test, comparing it with other rare-variant methods including those similarity-based tests that use both within-group and between-group information. As similarity-based approaches serve as one of the dominating approaches in rare-variant analysis, our results provide some insight for the effective detection of rare variants.

KEYWORDS cross-sample comparison; within-sample comparison; *U*-statistics

THE advent of next-generation sequencing technology has produced an enormous amount of genetic data that scientists can use to investigate how genetic variants contribute to disease mechanisms. This has led to the development of many statistical methods designed to examine such associations, including various single-marker tests and the more powerful multiple-marker analyses (also known as gene-/set-based analyses) where the functional role of a group of

variants or a signaling pathway is of interest. One class of the marker-set methods is based on the nonparametric *U*-statistic and uses dissimilarity or similarity (dis/similarity in short) measures to test for association (*e.g.*, Pinheiro *et al.* 2005; Schaid *et al.* 2005; Wessel and Schork 2006; Wei *et al.* 2008, 2014; Tzeng *et al.* 2009; Li 2012; Wang *et al.* 2015). The formulation and derivation of these *U*-statistic-based tests are simple and intuitive; they are useful when not much information is available about the mode of inheritance and phenotype distribution.

These *U*-statistic tests contrast genetic dis/similarity scores of certain subject groups in different ways. For instance, in case-control studies, Schaid *et al.* (2005) proposed quantifying the difference between two within-group kernel measures (*i.e.*, to perform case-case vs. control-control comparisons). Their global test performed well with an asymptotic

Copyright © 2018 by the Genetics Society of America

doi: <https://doi.org/10.1534/genetics.118.300769>

Manuscript received January 30, 2018; accepted for publication March 2, 2018; published Early Online March 14, 2018.

Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.118.300769/-/DC1.

¹Corresponding authors: Department of Statistics & Bioinformatics Research Center, Campus Box 7566, North Carolina State University, Raleigh, NC 27695-7566. E-mail: jytzeng@ncsu.edu; and ckhsiao@ntu.edu.tw

distribution when functional variants were independent. Wei *et al.* (2008) further incorporated a between-group measure, specifically the case-control dissimilarity using Hamming distance (HD), and used the ratio of between-group to within-group similarity in the test. When only common variants (CVs) [i.e., the minor allele frequency (MAF) $\geq 5\%$] are considered, the inclusion of both between- and within-group measures improves the testing power. Wang *et al.* (2015) also adopted HD to first cluster SNP data and next test the phenotypic association of each cluster with the difference between the between-group and within-group dissimilarity. For a review of U -statistic-based tests and similarity-based kernel methods, see Schaid (2010) and Li (2012). These methods have demonstrated great success when dealing with causal CVs; they are robust against the ratio of neutral:causal variants and maintain consistent power when deleterious and protective variants are mixed. For variants that are not common, Wei *et al.* (2014) combined genetic similarity and phenotypic similarity into a weighted U -statistic for sequencing data (referred to as WU-SEQ) to investigate whether variants with $\text{MAF} \leq 3\%$ contribute to disease etiology. Their test includes both between- and within-group comparisons, and it outperforms the sequence kernel association test (SKAT) (Wu *et al.* 2011) when the phenotype distribution has heavy tail probability.

Although the strategy of including both within- and between-group information in the U -statistic-based tests can increase testing power for CVs (Wei *et al.* 2008; Wang *et al.* 2015), we found that the contrast of these two factors is not optimal for rare variant (RV) ($\text{MAF} \leq 1\%$) analysis. This is because when RVs are involved the between-group dis/similarity information tends to fall in between the case-case dis/similarities and the control-control dis/similarities, or it overlaps with one of the two within-group dis/similarities (Figure 1), for RVs and for less frequent variants ($1\% \leq \text{MAF} \leq 5\%$). This observation suggests that higher power may be achieved by contrasting case-case dis/similarity with control-control dis/similarity instead of contrasting within-group comparisons with between-group comparisons. This suggestion is echoed by several recent studies which imply that contrasting two extreme phenotype groups may improve the statistical power of RV association tests (Guey *et al.* 2011; Li *et al.* 2011; Barnett *et al.* 2013; Lee *et al.* 2014).

In this work, we investigate the behavior of dissimilarity scores obtained from within-group comparisons and between-group comparisons for RVs based on HD. We show that, unlike the analysis with CVs (where the difference between the between-group dissimilarity and the within-group dissimilarity tends to be larger than the difference between case-case dissimilarity and control-control dissimilarity), the between-group dissimilarity of RVs falls between the two within-group dissimilarities. The inclusion of between-group information tends to attenuate the association signals and leads to power loss. To address this issue, we construct a U -statistic test including only the within-case and within-control comparisons; we derive its statistical properties and computation algorithms for obtaining P -values.

Finally, we show the power gain of the proposed test by comparing it with existing RV similarity tests that include both within- and between-group information, using coalescent simulations based on COSI (Schaffner *et al.* 2005) and data from the psychiatric Cohorte Lausannoise (PsyCoLaus) study (Preisig *et al.* 2009).

Materials and Methods

Let $\mathbf{X}_i$ denote the single-nucleotide variant vector of the i -th individual in the control (normal) group, where $i = 1, \dots, N_1$ and $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{iL})$ contains the genotype coding of L variants. For instance, the coding can be the number of minor alleles in the genotype. Similarly, let $\mathbf{Y}_j = (Y_{j1}, Y_{j2}, \dots, Y_{jL})$ represent the vector of genotype coding of the j -th subject in the case (disease) group, where $j = 1, \dots, N_2$. A dissimilarity measure $d(\cdot, \cdot)$ is defined for any pair of genotype vectors from two individuals. The degree of dissimilarity (heterogeneity) in the control group is measured as the average dissimilarity within this group,

$$W_{cn} = \frac{1}{C_2^{N_1}} \sum_{\substack{i, i' \in \{1, \dots, N_1\} \\ i < i'}} d(\mathbf{X}_i, \mathbf{X}_{i'}).$$

The average dissimilarity within the case groups is

$$W_{cs} = \frac{1}{C_2^{N_2}} \sum_{\substack{j, j' \in \{1, \dots, N_2\} \\ j < j'}} d(\mathbf{Y}_j, \mathbf{Y}_{j'}).$$

Specifically, a sum kernel over the L loci, $d(\mathbf{X}_i, \mathbf{X}_{i'}) = \sum_{l=1}^L d_l(X_{il}, X_{i'l})$, is recommended to account for variant-specific “dosage” effects. Here we consider the HD dissimilarity Ψ for illustration. The Ψ counts the number of discordant elements between two strings (vectors) of the same length,

$$\Psi_{ii'} = \Psi(\mathbf{X}_i, \mathbf{X}_{i'}) = \sum_{l=1}^L \psi_{l, ii'} = \sum_{l=1}^L I(X_{il} \neq X_{i'l}).$$

Note that here Ψ uses a slightly different scoring system from the allele-matching kernel (which counts the number of matching alleles among the four comparisons of two alleles per locus) besides the fact that the former scores based on dissimilarity and the latter scores based on similarity. To illustrate the difference, consider only one SNP. The allele-matching kernel assigns 4 for comparing AA vs. AA, 2 for AA vs. Aa, and 0 for AA vs. aa; in contrast, $\Psi_{ii'} = 0$ when comparing AA vs. AA and $\Psi_{ii'} = 1$ for the other two. Although HD Ψ differs from the identity-by-state (IBS) similarity measure for CVs, Ψ would be a location shift of IBS when dealing with the extremely low frequency of the minor allele homozygotes.

Distribution of dissimilarity scores for case-case, control-control, and case-control comparisons

Figure 1 displays the empirical distributions of $\Psi_{ii'} = \Psi(\mathbf{X}_i, \mathbf{X}_{i'})$ for control-control pairs, $\Psi_{jj'} = \Psi(\mathbf{Y}_j, \mathbf{Y}_{j'})$ for case-case pairs, and $\Psi_{ij} = \Psi(\mathbf{X}_i, \mathbf{Y}_j)$ for case-control pairs, based

Figure 1 Empirical distributions of dissimilarity measures under different settings. Empirical distributions of HD dissimilarity measure for pairs from the control group (green curve), from the case group (blue), and from across group (red). The three rows are for SNPs of neutral, deleterious, and protective effects, respectively. Each column represents a range of the MAFs.

on the COSI simulated sequence data (Schaffner *et al.* 2005) as detailed in the section *Simulation settings*. The data contain 520 RVs, of which 473 have MAFs $\leq 0.5\%$, 19 have MAFs in the range of (0.5, 1%), and 28 in [1, 5%]. We randomly selected 19 variants from each MAF range as causal variants and we used case-control sampling to generate data for 2000 cases and 2000 controls. More details are in Supplemental Material, [File S1](#).

Under the null hypothesis of no association (first row in Figure 1), the dissimilarity within the case group and within

the control group are nondifferentiable. When the effects of the RVs are deleterious or protective (second and third row, respectively), $\Psi_{ii'}$ and $\Psi_{jj'}$ are distant from each other, while Ψ_{ij} usually lies between the two and is closer to one of them. Also note that the distributions may contain two modes depending on the value of the MAF: When the MAF falls in the range (0.5, 5%), the second bump becomes apparent especially when the variants are deleterious. That is, when the variants have low but not rare MAF (MAF between 1 and

Figure 2 Empirical distributions of the test statistic ($W_{cs} - W_{cn}$) under different settings, including effect direction, number of subjects, number of causal variants, and MAF range of the causal variants. In the scenarios of 2000 cases and 2000 controls, the solid line is for MAF $< 0.5\%$, dashed line for MAF in $[0.5, 1\%]$, and dashed-dotted line for MAF in $[1, 5\%]$. Note that in the scenario of 28 causal RVs (i.e., the bottom-right figure), there is no dashed curve because there are not enough SNPs with MAF in $[0.5, 1\%]$. In the scenario of 5000 cases and 5000 controls (i.e., the bottom-left figure), the dotted, solid, dashed, and dashed-dotted lines are for MAF in the range of $< 0.1\%$, $[0.1, 0.5\%]$, $[0.5, 1\%]$, and $[1, 5\%]$, respectively.

5%), the bimode becomes substantial. In addition, depending on whether the effect is deleterious or protective, the $\Psi_{ii'}$ can be either larger (second row) or smaller (third row) than $\Psi_{jj'}$.

These observations motivate the consideration of a test statistic based only on the within-group information ($\Psi_{ii'}$ and $\Psi_{jj'}$) and excluding Ψ_{ij} . We refer to this test as the HD-based association test for RVs (HDAT-RV) and it is defined as $W_{cs} - W_{cn}$, where W_{cn} is equal to the average of the $\Psi_{ii'}$ and W_{cs} is the average of the $\Psi_{jj'}$.

Asymptotic distribution of HDAT-RV

First note that for any paired individuals (i, j) with L single-nucleotide variants, Ψ_{ij} takes a nonnegative integer value between 0 and L , $\Psi_{ij} = \sum_{l=1}^L I(X_{il} \neq X_{jl}) = \sum_{l=1}^L \psi_{l,ij}$, where the 0/1 indicator $\psi_{l,ij}$ follows a Bernoulli distribution with a variant-specific discordant probability $\omega_l = P(\psi_{l,ij} = 1)$.

If these L SNPs are independent, then Ψ_{ij} represents the sum of L independent Bernoulli distributions and follows a Poisson-Binomial distribution,

$$\Pr(\Psi_{ij} = k) = \sum_{A \in F_k} \prod_{l \in A} \omega_l \prod_{l' \in A^c} (1 - \omega_{l'}), \text{ for } k = 0, \dots, L.$$

Note that the index set $A = \{l : \psi_{l,ij} = 1\}$ in $F_k = \{A : \sum_{l \in A} \psi_{l,ij} = k\}$ contains the indices of SNPs corresponding to $\psi_{l,ij} = 1$ and $\sum_{l \in A} \psi_{l,ij} = k$ (more details in File S1). The mean and variance are $E(\Psi_{ij}) = \sum_{l=1}^L \omega_l$ and $\text{Var}(\Psi_{ij}) = \sum_{l=1}^L \omega_l(1 - \omega_l)$, respectively.

If these L SNPs are correlated, then the distribution of Ψ_{ij} can be derived from the joint distribution of $(\psi_{1,ij}, \psi_{2,ij}, \dots, \psi_{L,ij})$ and becomes

$$\Pr(\Psi_{ij} = k) = \sum_{A \in F_k} \left\{ \prod_{l \in A} \omega_l \prod_{l' \in A^c} (1 - \omega_{l'}) h(\psi_{1,ij}, \psi_{2,ij}, \dots, \psi_{L,ij}) \right\},$$

for $k = 0, \dots, L$;

where $h(\psi_{1,ij}, \psi_{2,ij}, \dots, \psi_{L,ij})$ contains higher order association parameters (details in File S1). The mean is $E(\Psi_{ij}) = \sum_{l=1}^L \omega_l$ and variance is

$$\text{Var}(\Psi_{ij}) = \sum_{l=1}^L \omega_l(1 - \omega_l) + \sum_{l \neq l'} \text{Cov}(\psi_{l,ij}, \psi_{l',ij}).$$

Following the asymptotic theorems for U -statistics with the HD as the kernel function, $W_{cs} - W_{cn}$ is asymptotically distributed as

$$\sqrt{N_1 + N_2} [(W_{cs} - W_{cn}) - E(W_{cs} - W_{cn})] \xrightarrow{D} N(0, \sigma_{W_{cs}}^2 + \sigma_{W_{cn}}^2).$$

The variance $\sigma_{W_{cn}}^2$ represents the variability of the average dissimilarity measure within the control group. It is four times the covariance, $\text{Cov}(\Psi_{im}, \Psi_{in})$, where index i represents the common control subject. Similarly, it is $\sigma_{W_{cs}}^2 = 4 \times \text{Cov}(\Psi_{jm}, \Psi_{jn})$ for the case group.

Figure 2 examines the empirical distribution of $W_{cs} - W_{cn}$ based on 19 randomly selected variants with MAF $\leq 0.5\%$,

19 variants with MAF between 0.5 and 1%, and 19 variants with MAF between 1 and 5%. Each plot contains 50 replications (details in [File S1](#)). It shows that the distribution is majorly affected by the range of MAFs, but it is not much affected by the number of variants or number of subjects.

P-value computation of small sample sizes

For relatively small sample sizes, as often encountered in RV tests, the above asymptotic distribution may not be applicable and obtaining *P*-values empirically based on permutations or bootstraps is often recommended (Lin and Tang 2011). To assure the validity of the test even for small sample sizes, we use permutations to calculate *P*-value. In our permutation, the pairwise dissimilarity among subjects only needs to be calculated once. Then, in each permutation, the dissimilarity between any two subjects can be retrieved as long as the indices of these subjects are identified. As explained in Wang *et al.* (2015), such computational complexity is of the order $O(L \times N^2)$, where L is the number of variants and N is the sample size.

Simulation settings

The performance of the proposed method was evaluated under various conditions of effect direction (deleterious vs. protective), control:case (CN/CS in short) ratio, total sample size, and proportion of causal RVs within a given region. A set of 10,000 haplotypes was first generated using the COSI software in Schaffner *et al.* (2005) with a coalescent model mimicking the linkage disequilibrium and population history of the European population. We focused on a region that includes 492 RVs with $MAF \leq 1\%$. Next, for each subject, a genotype sequence was constructed by randomly selecting two haplotypes with replacement from the 10,000 haplotypes. Based on the genotype sequence, the disease status Y_i of subject i was then determined via the following model:

$$\begin{aligned} \text{logit}(p_i) &= \beta_0 + \sum_{k=1}^K \beta_k S_{ik}, \\ Y_i &\sim \text{Bernoulli}(p_i), \end{aligned}$$

where S_{ik} was the number of causal alleles at the k -th locus and β_0 was set to correspond to a prevalence of 5%. To allow larger effect size (regression coefficients) for rarer causal variants, we followed Wu *et al.* (2011) and set $\beta_k = c_k |\log_{10} MAF_{S_k}|$ for causal variant k . Value c_k was chosen to make the power within a reasonable range across the scenarios considered. It was 0.4 for a deleterious or -0.4 for a protective effect. The proportion of causal loci was assumed as 5, 15, or 25%, corresponding to $p = 25, 74$, or 123 causal loci, respectively. The total sample size (N) was 600 or 1200, each with three CN/CS ratios: 1/1, 1/2, and 2/1. We considered four scenarios of causal effects, where the ratio of deleterious:protective (Del/Pro in short) variants ranged from 100/0, 80/20, 50/50, to 0/100. Every simulation setting contained 1000 replications and each *P*-value was derived based on 1000 permutations. The proposed HDAT-RV is compared with two *U*-statistic-based tests (*i.e.*, Wei's test and

Table 1 Type I error rates at the nominal level of 0.05

	CN/CS = 1/1		CN/CS = 1/2		CN/CS = 2/1	
	$N = 600$	$N = 1200$	$N = 600$	$N = 1200$	$N = 600$	$N = 1200$
HDAT-RV	0.046	0.058	0.052	0.050	0.046	0.050
SKAT	0.053	0.047	0.051	0.044	0.045	0.051
SKAT-O	0.056	0.049	0.041	0.052	0.048	0.043
Burden	0.052	0.045	0.048	0.047	0.045	0.045
Wei	0.058	0.048	0.053	0.051	0.054	0.052
WU-SEQ	0.056	0.044	0.051	0.045	0.045	0.052

WU-SEQ), SKAT, the burden test of Madsen and Browning (2009), and a hybrid method of kernel and burden tests [*i.e.*, SKAT-O of Lee *et al.* (2012)]. When quantifying genetic similarity, SKAT and SKAT-O use linear kernel, WU-SEQ uses IBS kernel, and Wei and HDAT-RV use HD. The code and data files for performing simulations are provided in [File S2](#) and [File S3](#).

Data availability

Detailed descriptions of the simulation settings are in [File S1](#). [File S2](#) contains the code for deriving HDAT-RV, conducting simulation studies, and calculating type I error rate and power. [File S3](#) contains the data needed for the simulation studies with the code in [File S2](#). CoLaus data are available at the Database of Genotypes and Phenotypes (dbGaP) with accession number: phs000145.v4.p2.

Results

Simulation studies

Table 1 lists the type I error rates at a nominal level of 0.05 for $N = 600$ and 1200 under different CN/CS ratios. The values are around the nominal level for all methods. Figure 3 shows the power performance for $N = 600$ at the 0.05 level for a spectrum of effect patterns and CN/CS ratios. The CN/CS ratios of 1/1, 1/2, and 2/1 are shown in the columns from left to right. The Del/Pro ratios of 100/0, 80/20, 50/50, and 0/100 are shown from top to bottom. Inside each of the 12 plots in Figure 3, we show the power results for the percentages of causal RVs as 5, 15, and 25%. Generally speaking, the best method varies across scenarios: When the Del/Pro ratio is not 1 (*i.e.*, 100/0, 80/20, and 0/100), HDAT-RV performs better; when the Del/Pro ratio is 50/50, WU-SEQ becomes the best method.

We also observe that HDAT-RV tends to yield reasonable power regardless of CN/CS and Del/Pro ratio, while other methods are somewhat sensitive to these ratios. For example, SKAT has relatively lower power when CN/CS is 1/2 (more cases) and the Del/Pro ratio is high (100/0 or 80/20) or when CN/CS is 2/1 (more controls) and the Del/Pro ratio is 0/100; WU-SEQ has low power in scenarios different from those of SKAT, *e.g.*, WU-SEQ has little power when CN/CS is 2/1 and the Del/Pro ratio is high (100/0 or 80/20) or when CN/CS is 1/2 and the Del/Pro ratio is 0/100. The burden test, as expected, has less power when there is a mixture of deleterious and protective RVs.

Figure 3 Power performance under different scenarios at the 0.05 nominal level. The Del/Pro ratio of RVs includes 100%:0 (first row), 80:20% (second row), 50:50% (third row), and 0:100% (fourth row). The sample size is 600, where the CN/CS ratios are 1/1 (first column), 1/2 (second column), and 2/1 (third column). The percentages of causal RVs are 5, 15, and 25%.

We further explore the behavior of WU-SEQ. The quantile-quantile plot of P -values under the null hypothesis (Figure S1) confirms that it has the expected null distribution regardless of the CN/CS ratio. On the other hand, under the alternative hypothesis, WU-SEQ can become negative when the causal variants are rare ($MAF \leq 1\%$) and the CN/CS ratio is not 1 under certain effect patterns (e.g., when there are more controls with deleterious effects, or when there are more cases under protective effects; see Figure S2). The negative

statistics may be related to the fact that the phenotypic similarity is limited to only three values when dealing with binary traits, leading to negative WU-SEQ in this case. Such an issue is not present when traits are continuous or when variants are less frequent (MAF between 1 and 5%).

We also present the results at a nominal 0.005 level (i.e., type I error rate in Table S1 and power in Figure S3), as well as an additional analysis with $N = 1200$ at nominal levels of 0.05 (Figure S4 and Figure S5) and 0.005 (Figure S6 and

Table 2 Genes identified by the six association tests

Chromosome	Gene	No. of RVs ^a	HDAT-RV	SKAT	SKAT-O	Burden	Wei	WU-SEQ	Reference ^b
1	<i>FAAH</i>	48	0.023	0.037	0.041	0.042	0.213	0.933	^{c, d, e}
19	<i>PDE4A</i>	73	0.038	0.020	0.038	0.316	0.259	0.663	^c
3	<i>GHSR</i>	22	0.059	0.004	0.012	0.949	0.035	0.600	^{d, f}
17	<i>CCL11</i>	13	0.907	0.058	0.025	0.013	0.662	0.025	
5	<i>EGR1</i>	31	0.022	0.687	0.071	0.037	0.445	0.968	^{d, f}
7	<i>LEP</i>	56	0.819	0.294	0.087	0.046	0.658	0.002	
20	<i>CHRNA4</i>	101	0.556	0.447	0.096	0.049	0.507	0.015	
1	<i>APCS</i>	12	0.071	0.147	0.241	0.801	0.031	0.729	
1	<i>FH</i>	28	0.445	0.976	0.182	0.105	0.779	0.006	
2	<i>NRXN1</i>	159	0.017	0.933	0.880	0.670	0.838	0.692	^{c, g}
2	<i>TACR1</i>	19	0.951	0.925	0.124	0.070	0.341	0.019	
6	<i>HCRTR2</i>	29	0.044	0.871	1.000	0.791	0.823	0.383	^d
8	<i>OPRK1</i>	74	0.010	0.962	0.402	0.240	0.938	0.833	^{c, d, h}
11	<i>DRD2</i>	45	0.030	0.254	0.416	0.772	0.214	0.495	^{c, d, i}
12	<i>P2RX7</i>	73	0.848	0.037	0.068	0.103	0.099	0.393	
15	<i>CHRNA3</i>	56	0.019	0.242	0.350	0.240	0.392	0.769	^{d, j}
15	<i>CHRNA7</i>	11	0.034	0.559	0.624	0.446	0.673	0.718	
22	<i>ADORA2A</i>	78	0.007	0.438	0.629	0.415	0.094	0.878	^{c, d, i}
22	<i>MCHR1</i>	53	0.857	0.268	0.145	0.086	0.166	0.024	
Total number of genes identified			10	4	4	5	2	6	

P-values <0.05 are shown in bold font.

^a The number of RVs in each gene.

^b References listed below the table.

^c These have been considered for drug development.

^d These are reported to be associated with neurodevelopment, nicotine dependence, sleep disorders, stress, and anxiety.

^e These are reported to be associated with central nervous system function and as a treatment for anxiety.

^f It was reported that *GHSR* may mediate stress-induced, food-reward behavior in mice (Chuang *et al.* 2011), and such stress could induce anxiety as well (and is thus associated with anxiety-like behavior). Similarly, studies have pointed out that, in mice, low *EGR1* expression may be associated with social isolation and stress-induced behavior abnormalities (Matsumoto *et al.* 2012).

^g The *NRXN1* gene encodes protein in the neurexin family and is related to central nervous system function and cell adhesion. It has been considered as an antipsychotic for schizophrenia and autism. In addition, analysis based on rare copy number variation has demonstrated its association with mental diseases, especially the *NRXN1*-β isoform (Malhotra and Sebat 2012; Viñas-Jornet *et al.* 2014).

^h *OPRK1* is associated with stress response and drug addiction (Xu *et al.* 2013).

ⁱ *DRD2* has been used in antipsychotic drugs, and both *DRD2* and *ADORA2A* have been shown to be associated with caffeine-induced anxiety (Childs *et al.* 2008).

^j *CHRNA3* has been shown to be associated with addiction behavior and neuroticism in some populations (Criado *et al.* 2014).

Figure S7). The results follow a similar pattern. In summary, the relative power of different methods suggests that the CN/CS and Del/Pro ratio together affect the performance of a method. In practice, because the Del/Pro ratio is unknown, a more robust choice would be HDAT-RV and SKAT-O, which consistently give the best or comparable power to the best-performing methods across different scenarios.

Data applications

We applied the proposed test to the Cohorte Lausannoise (CoLaus) sequence study (Firmann *et al.* 2008; Song *et al.* 2012), where the targeted sequencing genotypes on 202 drug-targeted genes were obtained for 1769 subjects. The information on anxiety disorders according to the Diagnostic and Statistical Manual of Mental Disorders, fourth edition, was available for 1689 subjects from the psychiatric evaluation (PsyCoLaus study; Preisig *et al.* 2009). Among these subjects, 580 had at least one specific adult anxiety disorder and 1109 did not meet criteria for such a disorder. This is close to a CN/CS ratio of 2/1. Using the same threshold of 1% for the definition of RVs, there were 197 genes containing at least one RV (with a total of 10,330 RVs).

We started with a PubMed search to identify any genes in the 197 genes that were presented in any abstracts containing both the disease status (anxiety) and the gene symbol. A total of 73 genes, covering 4368 RVs, were identified and defined as “candidate genes,” although most of these candidate genes were reported because of the association between CVs and disease phenotype. We conducted gene-based analyses using HDAT-RV, SKAT, SKAT-O, Burden, Wei, and WU-SEQ tests, and evaluated the performance of each method by assessing its ability to reproduce the association between anxiety disorder status and the genes. The *P*-values of HDAT-RV, Wei, and WU-SEQ were computed using 5000 permutations.

Among the 73 candidate genes found in PubMed and at the 0.05 nominal level, four genes, *fatty acid amide hydrolase* (*FAAH*), *PDE4A*, *GHSR*, and *CCL11* were identified by at least three tests (top panel in Table 2). The next three genes reached significance under two tests (second panel), while 12 genes (third panel) were identified significant with only one test. Table 2 also lists the total number of genes identified by each method (bottom row). HDAT-RV identified the most genes (*i.e.*, 10 out of the 73 PubMed genes), followed by WU-SEQ (6), Burden (5), SKAT-O (4), SKAT (4), and Wei (2).

Note that several of the genes have been considered for drug development, including *FAAH*, *neurexin-1 (NRXN1)*, *OPRK1*, *DRD2*, *PDE4A*, and *ADORA2A*. In addition, *FAAH*, *GHSR*, *EGR1*, *HCRT2*, *OPRK1*, *DRD2*, *CHRNA3*, and *ADORA2A* have been reported to be associated with neurodevelopment, nicotine dependence, sleep disorders, stress, and/or anxiety. *FAAH* is known to play an important role in the endocannabinoid system; the association between *FAAH* and anxiety behavior has been reported in Dincheva *et al.* (2015). The *FAAH* inhibitors have been suggested as a target therapy for central nervous system disorders (Ahn *et al.* 2009), for anxiety through activation of cannabinoid receptors (Panlilio *et al.* 2013), and for post-traumatic stress disorder (Bisogno and Maccarrone 2013). For *PDE4A*, studies have reported its association with anxiety-like behavior (Hansen *et al.* 2014; Keil *et al.* 2016) and it has been a target for antidepressants like rolipram (Wallace *et al.* 2005). More reported associations and references are listed in Table 2.

Discussion

In this study, we focus on comparing the dissimilarity measures of multiple RVs from two extreme phenotype groups. We show that the between-/cross-group dissimilarity measure often overlaps with one of the two within-group dissimilarity measures, and thus reduces testing power if the between-group information is considered in the test statistic. This was not an issue in earlier studies because such overlap is not present for analyses with CVs. While it has been noted for CVs that the two within-group measures and the between-group measure should all be included (Wang *et al.* 2015); for RVs, comparison of the two within-group measures can more effectively quantify the difference between the case and the control group. Using simulations, we show that the proposed HDAT-RV has reasonable and robust power performance across different CN/CS and Del/Pro ratios, while other methods that incorporate both within- and between-group information may be sensitive to these two ratios.

To address the potential concerns of insufficient sample size to apply asymptotics, we also construct a computationally efficient permutation algorithm to obtain the *P*-value of the HDAT-RV. In the procedure, the dissimilarity measure of every pair of individuals is computed only once; no calculation of the measure is needed in all subsequent permutations. Such a feature is different from WU-SEQ, where pairwise similarity scores of genotypes and phenotypes have to be recomputed in every permutation.

When quantifying genetic similarity in this work, SKAT uses linear kernel, WU-SEQ uses IBS kernel, and Wei and HDAT-RV use HD. For RVs, the linear kernel, IBS kernel, and HD become equivalent; except that HD measures distance while the other two measure similarity. So we expect that the power difference would be less likely due to the choices of distance or dissimilarity measures among SKAT, Wei, WU-SEQ, and HDAT-RV; but would largely be from the different strategies of using the between-group similarity. Certain combinations may be

vulnerable to the imbalance of sample sizes in the case and control group, as demonstrated in this research. When the MAF is <1%, the proposed HDAT-RV is less sensitive to this imbalance. Another issue of concern is the choice of genetic similarity measure. As addressed in Wu *et al.* (2011), a function that better captures the true effect of multiple loci will increase power. Unfortunately, this information is usually unavailable. Further studies focusing on different relationships and distance measures may help to sort out the effectiveness of these tests, particularly when the percentage of causal variants is not known and when deleterious and protective causal RVs are mixed with an unknown ratio.

Acknowledgments

The authors thank Peter Vollenweider and Gerard Waeber, principal investigators of the Cohorte Lausannoise study, and Meg Ehm and Matthew Nelson, collaborators at GlaxoSmithKline, for providing the CoLaus phenotype and sequence data. The CoLaus/PsyCoLaus study was supported by grants from the Swiss National Science Foundation (#3200B0-105993) and from GlaxoSmithKline (Psychiatry Center of Excellence for Drug Discovery and Genetics Division, Drug Discovery, R&D, Verona, Italy). This work was supported in part by Ministry of Science and Technology of Taiwan grants 103-2314-B-002-039-MY3, 105-2118-M-032-012-MY2, 106-2314-B-002-097-MY3, and 106-2811-B-002-006 and National Institutes of Health grant P01 CA-142538-01.

Literature Cited

- Ahn, K., D. S. Johnson, and B. F. Cravatt, 2009 Fatty acid amide hydrolase as a potential therapeutic target for the treatment of pain and CNS disorders. *Expert Opin. Drug Discov.* 4: 763–784. <https://doi.org/10.1517/17460440903018857>
- Barnett, I. J., S. Lee, and X. Lin, 2013 Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genet. Epidemiol.* 37: 142–151. <https://doi.org/10.1002/gepi.21699>
- Bisogno, T., and M. Maccarrone, 2013 Latest advances in the discovery of fatty acid amide hydrolase inhibitors. *Expert Opin. Drug Discov.* 8: 509–522. <https://doi.org/10.1517/17460441.2013.780021>
- Childs, E., C. Hohoff, J. Deckert, K. Xu, J. Badner *et al.*, 2008 Association between *ADORA2A* and *DRD2* polymorphisms and caffeine-induced anxiety. *Neuropsychopharmacology* 33: 2791–2800. <https://doi.org/10.1038/npp.2008.17>
- Chuang, J. C., M. Perello, I. Sakata, S. Osborne-Lawrence, J. M. Savitt *et al.*, 2011 Ghrelin mediates stress-induced food-reward behavior in mice. *J. Clin. Invest.* 121: 2684–2692. <https://doi.org/10.1172/JCI57660>
- Criado, J. R., I. R. Gizer, H. J. Edenberg, and C. L. Ehlers, 2014 *CHRNA5* and *CHRNA3* variants and level of neuroticism in young adult Mexican American men and women. *Twin Res. Hum. Genet.* 17: 80–88. <https://doi.org/10.1017/thg.2014.11>
- Dincheva, I., A. T. Drysdale, C. A. Hartley, D. C. Johnson, D. Jing *et al.*, 2015 *FAAH* genetic variation enhances fronto-amygdala function in mouse and human. *Nat. Commun.* 6: 6395. <https://doi.org/10.1038/ncomms7395>
- Firmann, M., V. Mayor, P. M. Vidal, M. Bochud, A. Pécoud *et al.*, 2008 The CoLaus study: a population-based study to investigate

- the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. *BMC Cardiovasc. Disord.* 8: 6.
- Guey, L. T., J. Kravic, O. Melander, N. P. Burt, J. M. Laramie *et al.*, 2011 Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. *Genet. Epidemiol.* 35: 236–246.
- Hansen, R., M. Conti, and H. Zhang, 2014 Mice deficient in phosphodiesterase-4A display anxiogenic-like behavior. *Psychopharmacology (Berl.)* 231: 2941–2954. <https://doi.org/10.1007/s00213-014-3480-y>
- Keil, M. F., G. Briassoulis, and C. A. Stratakis, 2016 The role of protein kinase A in anxiety behaviors. *Neuroendocrinology* 103: 625–639. <https://doi.org/10.1159/000444880>
- Lee, S., M. C. Wu, and X. Lin, 2012 Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 13: 762–775. <https://doi.org/10.1093/biostatistics/kxs014>
- Lee, S., G. R. Abecasis, M. Boehnke, and X. Lin, 2014 Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* 95: 5–23. <https://doi.org/10.1016/j.ajhg.2014.06.009>
- Li, D., J. P. Lewinger, W. J. Gauderman, C. E. Murcray, and D. Conti, 2011 Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. *Genet. Epidemiol.* 35: 790–799. <https://doi.org/10.1002/gepi.20628>
- Li, H., 2012 U-statistics in genetic association studies. *Hum. Genet.* 131: 1395–1401. <https://doi.org/10.1007/s00439-012-1178-y>
- Lin, D. Y., and Z. Z. Tang, 2011 A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.* 89: 354–367. <https://doi.org/10.1016/j.ajhg.2011.07.015>
- Madsen, B. E., and S. R. Browning, 2009 A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 5: e1000384. <https://doi.org/10.1371/journal.pgen.1000384>
- Malhotra, D., and J. Sebat, 2012 CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* 148: 1223–1241. <https://doi.org/10.1016/j.cell.2012.02.039>
- Matsumoto, K., K. Ono, H. Ouchi, R. Tsushima, and Y. Murakami, 2012 Social isolation stress down-regulates cortical early growth response 1 (*Egr-1*) expression in mice. *Neurosci. Res.* 73: 257–262. <https://doi.org/10.1016/j.neures.2012.04.004>
- Panlilio, L. V., Z. Justinova, and S. R. Goldberg, 2013 Inhibition of *FAAH* and activation of *PPAR*: new approaches to the treatment of cognitive dysfunction and drug addiction. *Pharmacol. Ther.* 138: 84–102. <https://doi.org/10.1016/j.pharmthera.2013.01.003>
- Pinheiro, H. P., A. S. Pinheiro, and P. K. Sen, 2005 Comparison of genomic sequences using the Hamming distance. *J. Stat. Plan. Inference* 130: 325–339. <https://doi.org/10.1016/j.jspi.2003.03.002>
- Preisig, M., G. Waeber, P. Vollenweider, P. Bovet, S. Rothen *et al.*, 2009 The PsyCoLaus study: methodology and characteristics of the sample of a population-based survey on psychiatric disorders and their association with genetic and cardiovascular risk factors. *BMC Psychiatry* 9: 9. <https://doi.org/10.1186/1471-244X-9-9>
- Schaffner, S. F., C. Foo, S. Gabriel, D. Reich, M. J. Daly *et al.*, 2005 Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15: 1576–1583. <https://doi.org/10.1101/gr.3709305>
- Schaid, D. J., 2010 Genomic similarity and Kernel methods I: advancements by building on mathematical and statistical foundations. *Hum. Hered.* 70: 109–131.
- Schaid, D. J., S. K. McDonnell, S. J. Hebring, J. M. Cunningham, and S. N. Thibodeau, 2005 Nonparametric tests of association of multiple genes with human disease. *Am. J. Hum. Genet.* 76: 780–793. <https://doi.org/10.1086/429838>
- Song, K., M. R. Nelson, J. Aponte, E. S. Manas, and S. A. Bacanu, 2012 Sequencing of Lp-PLA2-encoding *PLA2G7* gene in 2000 Europeans reveals several rare loss-of-function mutations. *Pharmacogenomics J.* 12: 425–431. <https://doi.org/10.1038/tpj.2011.20>
- Tzeng, J. Y., D. Zhang, S. M. Chang, D. C. Thomas, and M. Davadian, 2009 Gene-trait regression for multimarker-based association analysis. *Biometrics* 65: 822–832. <https://doi.org/10.1111/j.1541-0420.2008.01176.x>
- Viñas-Jornet, M., S. Esteba-Castillo, E. Gabau, N. Ribas-Vidal, N. Baena *et al.*, 2014 A common cognitive, psychiatric, and dysmorphic phenotype in carriers of *NRXN1* deletion. *Mol. Genet. Genomic Med.* 2: 512–521. <https://doi.org/10.1002/mgg3.105>
- Wallace, D., L. Johnson, E. Huston, D. MacMaster, T. Houslay *et al.*, 2005 Identification and characterization of *PDE4A11*, a novel, widely expressed long isoform encoded by the human *PDE4A* cAMP phosphodiesterase gene. *Mol. Pharmacol.* 67: 1920–1934. <https://doi.org/10.1124/mol.104.009423>
- Wang, C., W.-H. Kao, and C. K. Hsiao, 2015 Using Hamming distance as information for SNP-sets clustering and testing in disease association studies. *PLoS One* 10: e0135918. <https://doi.org/10.1371/journal.pone.0135918>
- Wei, C., M. Li, Z. He, O. Vsevolozhskaya, D. J. Schaid *et al.*, 2014 A weighted U-statistic for genetic association analyses of sequencing data. *Genet. Epidemiol.* 38: 699–708. <https://doi.org/10.1002/gepi.21864>
- Wei, Z., M. Li, T. Rebbeck, and H. Li, 2008 U-statistics-based tests for multiple genes in genetic association studies. *Ann. Hum. Genet.* 72: 821–833. <https://doi.org/10.1111/j.1469-1809.2008.00473.x>
- Wessel, J., and N. J. Schork, 2006 Generalized genomic distance based regression methodology for multilocus association analysis. *Am. J. Hum. Genet.* 79: 792–806. <https://doi.org/10.1086/508346>
- Wu, M. C., S. Lee, T. Cai, Y. Li, K. Boehnke *et al.*, 2011 Rare-variant association testing for sequencing data with the sequence Kernel association test. *Am. J. Hum. Genet.* 89: 82–93. <https://doi.org/10.1016/j.ajhg.2011.05.029>
- Xu, K., D. Seo, C. Hodgkinson, Y. Hu, D. Goldman *et al.*, 2013 A variant on the kappa opioid receptor gene (*OPRK1*) is associated with stress response and related drug craving, limbic brain activation and cocaine relapse risk. *Transl. Psychiatry* 3: e292. <https://doi.org/10.1038/tp.2013.62>

Communicating editor: E. Hauser