Journal of Human Genetics (2017) 62, 777–781

www.nature.com/jhg

ORIGINAL ARTICLE

An adaptive strategy for association analysis of common or rare variants using entropy theory

Yu-Mei Li[1,2], Chao Xu[2], Yang Xiang[1], Cheng Peng[2,3] and Hong-Wen Deng[2]

Advances in DNA sequencing technology have been promoting the development of sequencing studies to identify rare variants associated with complex traits. Adaptive strategy can be effective to reduce the noise provided by non-causal variants. However, the existing adaptive strategies depend on many assumptions. In this paper, we proposed a new adaptive strategy using entropy theory for association analysis. This entropy-based strategy is based on the magnitude of association between variants and disease and does not depend on the detailed association pattern with causal variants. We considered multi-marker test and Sum test with collapsing method to construct the entropy-based adaptive strategy. Using simulation studies, we investigated the performance of our method for rare variant analyses as well as for common variant analyses with multi-marker test and compared it with several existing adaptive strategies. The results showed that our method can improve the power and achieve good performance when there is a large number of non-causal variants and effects of causal variants are in the same direction for rare variant.
Journal of Human Genetics (2017) **62,** 777–781; doi:10.1038/jhg.2017.39; published online 6 April 2017

INTRODUCTION

Genome-wide association studies have successfully identified a large number of common genetic variants involved in common diseases. However, most associations detected by current genome-wide association studies only explained a limited proportion of heritability for most complex traits.[1] Recent studies showed that rare variants (RVs) contribute to the missing heritability unexplained by the discovered common variants.[2] RVs are referred to as alternative forms of a gene that are present with a minor allele frequency (MAF) of less than 1% and have a larger effect size compared to common variants. Due to the low MAFs of RVs, traditional approaches used for analyses of common variants lack power and require large sample size to detect the variant-disease association. With the development of next-generation sequencing technologies, the availability of large quantities of sequence data provides an unprecedented opportunity for researchers to develop novel statistical methods for RV association analyses.

Due to the low MAFs and little variation information in a single RV, many methods have been explored to search for accumulative effects of a group of RVs. These include the cohort allelic sums test,[2] the combined multivariate and collapsing method (CMC),[3] The Sum test,[4] the weighted-sum method[5] and the variable threshold method.[6] The main idea of these methods is collapsing or pooling RVs across a causal region into one 'super' variant to increase allele frequency and then collectively testing their association effect as a whole. Although these methods can improve power by combining information of multiple RVs, they are developed with the assumption that all variants in the region have an effect on the phenotype and the effects are at the same direction with the same magnitude. These tests will lose power when the set of collapsed variants includes non-causal variants or the effects of causal variants have different directions. Various methods have been proposed recently to overcome these limitations. These include C-alpha score test,[7] the sequence kernel association test[8] and the adaptive sum strategy.[9] The adaptive strategy is to select the importance RVs to construct statistics under some assumptions and is considered as an effective method to overcome limitations of collapsing methods. The variable threshold method is based on the assumption that the MAFs of the causal RVs may be different from those of non-functional RVs. The series of adaptive tests proposed by Pan and Shen[9] can be considered as the extension of the variable threshold method by ordering the standardized magnitudes of a statistic U or the locations of their corresponding RVs. However, these adaptive methods are not uniformly most powerful. Major reason is that they depend on specific association effect directions and sizes, while in reality the true association pattern with causal RVs is unknown and disease-association mutations are hard to choose.[9,10] So, developing adaptive method not depending on the unknown association pattern might be particularly useful for RV association analyses.

As an important metrics in information theory, the Shannon entropy[11] is usually used to measure uncertainty of a random variable. The entropy theory has an important performance: the conditional

[1]School of Mathematics and Computational Science, Huaihua University, Hunan, China; [2]Center for Bioinformatics and Genomics, Department of Global Biostatistics and Data Science, Tulane University, New Orleans, LA, USA and [3]Department of Geriatrics, National Key Clinical Specialty, Guangzhou First People's Hospital, Guangzhou Medical University, Guangdong, China
Correspondence: Dr Y-M Li, School of Mathematics and Computational Science, Huaihua University, Yingfeng East Road 612, Huaihua, Hunan 418008, China.
E-mail: lymmail@126.com or liym74@yahoo.com
Received 4 November 2016; revised 7 February 2017; accepted 7 March 2017; published online 6 April 2017

entropy of a variable given the knowledge of another variable is less than or equal to the unconditional entropy and they are equal when the two variables are independent. We can apply the entropy theory to characterize DNA variation[12] by constructing the difference (that is, the mutual information) between the entropy of a variant and its conditional entropy given the phenotype (affected or unaffected) and then quantifying the magnitude of association between the variant and the trait.

In this paper, we will propose a new adaptive strategy using entropy theory to test the variant-disease association. Our strategy is based on the magnitude of association but is not influenced by the unknown association pattern between the variants and the trait. At the same time, we expect our method is a generally strategy which can be used for RV or common variant. So we will consider the multi-marker test which is a powerful method for association analysis of both common variant and RV and Sum test to construct test statistic. Through simulation studies, we will assess and compare the performance of our method with the existing methods.

MATERIALS AND METHODS

Preliminaries of entropy

We consider two discrete random variables X and Y. X has the state x with the probability $p(x)$ and Y has the state y with the probability $P(Y=y)$. We let $P(x|Y=y)$ be the conditional probability of X given $Y=y$. The entropy of X and the conditional entropy of X given $Y=y$ are defined with the following Equations (1) and (2), respectively.

$$H(X) = -\sum_x p(x)\cdot\log p(x) \tag{1}$$

Where $p(x)\cdot\log p(x)=0$ if $p(x)=0$.

$$H(X|Y=y) = -\sum_x p(x|Y=y)\cdot\log p(x|Y=y) \tag{2}$$

Then the conditional entropy of X given Y is

$$H(X|Y) = -\sum_y p(Y=y)\cdot H(X|Y=y) \tag{3}$$

It should be noted that $H(X)-H(X|Y)\geqslant 0$ and the equality holds only if X and Y are independent.

The concept of entropy can be used to study the relationship between variations and disease susceptibility.[13] Because multivariate test is to test all variants simultaneously, it is a powerful method for association analysis of common variants. In addition, multivariate test is considered to be more robust than collapsing method for RVs analysis in the presence of misclassification of non-functional variants. Here, first we will focus on multi-marker test and consider how to use the entropy theory to develop an adaptive strategy for association analysis. Then we extend it to collapsing method for RVs analysis.

Multiple-marker test

We first briefly review the multi-marker statistic test. Assume n individuals with n^A affected and n^C unaffected individuals ($n^A+n^C=n$) are sampled. Suppose that there are k variants, each of which has two alleles A and a. We assume that the allele A is suspected of increasing the disease risk and has the population frequency of p_i for ith variant ($i=1,\ldots,k$). To simplify our presentation, a measure with a superscript 'A' indicates a measure in affected individuals, and a measure with a superscript 'C' indicates a measure in unaffected individuals. Let X_i be the number of copies 'A' for variant 'i', $i=1,\ldots,k$. Define a k-dimensional random variable $Z=(X_1, X_2,\ldots, X_k)^T$ presenting the state of allele A at k variants. Let $\sum=(\sigma_{ij})_{k\times k}$ be the covariance matrix of Z, where σ_{ij} is the covariance of X_i and X_j. Let $Z_i^A=(X_{1i}^A, X_{2i}^A, \cdots, X_{ki}^A)^T$ and $Z_j^C=(X_{1j}^C, X_{2j}^C, \cdots, X_{kj}^C)^T$ be the state of allele A for the ith ($i=1, 2,\ldots, n^A$) affected individual and the jth ($j=1, 2,\ldots, n^C$) unaffected individual, respectively. Let $\overline{Z}^A=\frac{1}{n^A}(\sum_{i=1}^{n^A}X_{1i}^A, \sum_{i=1}^{n^A}X_{2i}^A, \cdots, \sum_{i=1}^{n^A}X_{ki}^A)^T$ and $\overline{Z}^C=\frac{1}{n^C}(\sum_{i=1}^{n^C}X_{1i}^C, \sum_{i=1}^{n^C}X_{2i}^C, \cdots, \sum_{i=1}^{n^C}X_{ki}^C)^T$ be the mean vector of Z_i^A and Z_i^C, respectively. Let $\hat{\sum}^A$ and $\hat{\sum}^C$ be the sample covariance matrix of Z_i^A and Z_i^C, respectively. Then the multi-marker statistic test is as following:[14]

$$T_M=(\overline{X}^A-\overline{X}^C)^T\left(\frac{\hat{\sum}^A}{n^A}+\frac{\hat{\sum}^C}{n^C}\right)^{-1}(\overline{X}^A-\overline{X}^C) \tag{4}$$

The statistic T_M is asymptotically a χ^2 distribution with the degree of freedom of rank for $\frac{\sum^A}{n^A}+\frac{\sum^C}{n^C}$ under the null hypothesis of no association.

A new adaptive strategy for association analysis using entropy theory

We consider a homogeneous population. Under the assumption of random mating and thus Hardy–Weinberg equilibrium, X_i has the probability distribution $P_{X_{i0}}=P(X_i=0)=(1-p_i)^2$, $P_{X_{i1}}=P(X_i=1)=2p_i(1-p_i)$, $P_{X_{i2}}=P(X_i=2)=p_i^2$. From Equation (1), we calculate the entropy of X_i for variant i, $H_i=-\sum_{j=0}^{2}P_{X_{ij}}\cdot\log P_{X_{ij}}$. Define a variable Y as an individual's disease status, $Y=1$ if the individual is affected, $Y=0$ if the individual is unaffected. Then the conditional entropy of X_i given Y, denoted by H_{Ci}, is $H_{Ci}=p(Y=1)\cdot H_i^A+p(Y=0)\cdot H_i^C$, where $H_i^A=-\sum_{j=0}^{2}P_{X_{ij}}^A\cdot\log P_{X_{ij}}^A$ and $H_i^C=-\sum_{j=0}^{2}P_{X_{ij}}^C\cdot\log P_{X_{ij}}^C$ are the entropy of X_i in affected individuals and unaffected individuals, respectively. Let $\partial_i=H_i-H_{Ci}$. ∂_i is a measure of the magnitude of association: the larger the value, the stronger the association between variant i and disease and $\partial_i\geqslant 0$ with equality holding only if variant i is independent with the disease. We assume that there are L ($L\leqslant k$) variants with $\partial_i>0$. To simplify our presentation, we assume that the former L variants are those with $\partial_i>0$. We sort these L variants in descending order of ∂_i: $\partial_1\geqslant\partial_2\geqslant\ldots\geqslant\partial_L$. Let $G(L)$ be the variant set containing these L variants: $G(L)=\{i:\ \partial_1\geqslant\partial_2\geqslant\ldots\geqslant\partial_L\}$. It is noted that, obviously, variants not in $G(L)$ are those not associated with disease, and theoretically, L variants in $G(L)$ are associated with disease. However, because we calculate ∂_i with the sample data, those not associated with disease may have $\partial_i>0$. Thus, $G(L)$ contains all associated variants, and may also contain some variants not associated with the disease. Let $G(r)=\{i:\ \partial_1\geqslant\partial_2\geqslant\cdots\geqslant\partial_r\}(r=L,\ L-1,\cdots,1)$, for example, $G(L-1)=\{i:\ \partial_1\geqslant\partial_2\geqslant\cdots\geqslant\partial_{L-1}\}$, $G(L-2)=\{i:\ \partial_1\geqslant\partial_2\geqslant\cdots\geqslant\partial_{L-2}\}$, and $G(1)=\{i:\ \partial_1\}$. We obtain L variant sets $G(L),\cdots,G(1)$, containing $L,\cdots,1$ variants, respectively and the values of ∂_i in $G(r)$ are larger than those in variant sets ahead of $G(r)$. For each variant set $G(r)$, we define a statistic, denoted by $T_M^{G(r)}$, according to Equation (4). Our test statistic, here, denoted as T_M-E, is defined as following:

$$T_M-E=\min_{1\leq r\leq L}P_{T_M^{G(r)}} \tag{5}$$

where $P_{T_M^{G(r)}}$ is the P-value of $T_M^{G(r)}$. The statistical significance can be assessed by permutation.

Rare variants association analysis with the entropy-based adaptive strategy

In addition to multi-marker test, collapsing methods are widely used for association analysis of RVs. In order to describe how to use the entropy-based adaptive strategy for RVs, here we focus on the statistic of Sum test proposed by Pan[4] as an example. The Sum test T_{sum} is defined as following:

$$T_{\text{Sum}}=\frac{\mathbf{1}^TU}{\sqrt{\mathbf{1}^TV\mathbf{1}}} \tag{6}$$

Here, $\mathbf{1}=(1,\cdots,1)^T$ is the k-vector of all 1's. $U=\sum_{i=1}^{n}(Y_i-\overline{Y})Z_i$ is the score vector with the covariance matrix $V=\overline{Y}(1-\overline{Y})\sum_{i=1}^{n}(Z_i-\overline{Z})(Z_i-\overline{Z})^T$, where $\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$ and $\overline{Z}=\frac{1}{n}\sum_{i=1}^{n}Z_i$. Here, $Z_i=(X_{1i}, X_{2i},\cdots,X_{ki})^T$ presents the state of allele A for ith individual. The Sum test T_{sum} belongs to the family of pooled association tests or collapsing tests. Collapsing method is to collapse all RVs across a causal region into a 'super' variant and then collectively test their association effect as a whole. This method has been widely adopted to analyze RVs. However, collapsing tests will loss power if one does not eliminate the influence of non-causal RVs and the different directions of the causal variants. In order to remove the influence of different directions of causal RVs

and a large number of non-causal RVs, Price *et al.*[6] proposed a variable threshold test (Price-VT) based on the observed MAFs,

$$\text{Price}-\text{VT}=\max_{h\in H} z(h)=\max_{h\in H}\frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\xi_j^h X_{ij}(Y_i-\overline{Y})}{[\sum_{i=1}^{n}\sum_{j=1}^{k}(\xi_j^h X_{ij})^2]^{1/2}} \quad (7)$$

where H is the set of observed MAFs across all RVs and $\xi_j^h = I(h > \text{MAF of RV}j)$. Pan and Shen[9] proposed a general class of adaptive tests aT,

$$aT = aT(U) = \min_{1\le m\le k} P_{T(U_{(m)})} \quad (8)$$

where $U_{(m)}=(U_1,\cdots,U_m)$ is the score subvector containing the first m components of U, $T(U_{(m)})$ is the statistic based on $U_{(m)}$, and $P_{T(U_{(m)})}$ is the P-value of $T(U_{(m)})$. The test aT depends on the order of the components of U. They suggested two adaptive tests aT-Loc and aT-Ord by ordering the locations of their corresponding RVs and the standardized magnitudes of a statistic U, respectively. Here, we let aSum-Ord be the adaptive test of the Sum test based on the standardized magnitudes of a statistic $U(U_j/\sqrt{v_j})$ We can use the weighting scheme to improve the performance of the statistic. The commonly used weight is $w_i = 1/\sqrt{np_i(1-p_i)}$ (here, we denote it as w_{MB}) with denominator representing the estimated standard deviation of the total number of mutations in the sample.[5] Here, $p_i = \frac{n_i^u+1}{2n^u+2}$ is the allele frequency of the ith RV in unaffected individuals, where n_i^u is the number of minor alleles of the ith variant in unaffected individuals and n^u is the number of unaffected individuals.

Following the previous symbols, we suppose that there are $L(L\leqslant k)$ variants with $\partial_i>0$. Using these L variants we construct L variant sets $G(L),\cdots$, and $G(1)$ containing $L,\cdots$ and 1 variants, respectively. Here, the values of ∂_i in $G(r)$ are larger than those in variant sets ahead of $G(r)$. Then the adaptive test for RVs analysis using entropy theory is as following:

$$aT = aT^{G(r)} = \min_{1\le r\le L} P_{T^{G(r)}} \text{ or } aT = aT^{G(r)} = \max_{1\le r\le L} T^{G(r)} \quad (9)$$

where $T^{G(r)}$ is the statistic based on the variant set $G(r)$, and $P_{T^{G(r)}}$ is the P-value of $T^{G(r)}$. We denote our method as aT-E. The variant set corresponding to aT-E can be considered as the optimal set containing variants associated with the disease. Here, aSum-E is the adaptive test of the Sum test based on the entropy-based adaptive strategy. The statistical significance for all tests can be assessed by permutation. It should be noted that the variable threshold test is based on the assumption that the MAFs of the causal RVs may be different from those of non-functional RVs.[6] The aT test also depends on the order of the components of the score vector U.[9] However, different orders of the components of the score vector U may lead to inconsistent results. Even in practice, one can not objectively determine the effects of variants. Whether the MAFs of the causal RVs are different from those of non-functional RVs is generally unknown, and even if known, the magnitude of difference are unknown. Our entropy-based adaptive strategy is based on the magnitude of association between variants and disease. It does not need any other assumption about the effects and MAFs of RVs, thus overcoming the problems associoated with earlier Sum test.

RESULTS

Simulation setting

In our simulation studies, we assess the type-1 error rate and compare the power of our method with several existing adaptive methods under a wide range of parameter values. The simulation parameter includes the number of variants, the MAF of each variant, the number and effect size of causal variants, and the sample size. For common variant, we consider k ($k=4$, 10, 20, 50, 100) observed variants and an unobserved causal variant in the middle. The MAFs for k common variants are uniformly determined with values ranging from 0.1 to 0.4. The MAF of unobserved common causal variant is set to be 0.2. The odds ratio (OR) = 1 for all variants under the null hypothesis of no association and OR = 1 for all non-causal variants. Under the alternative hypothesis of association, we let OR = 1.5 for the common causal variant. The sample size n ($=2N$)is chosen as 500, 1000, 1500 or 2000 with N affected individuals and N unaffected individuals. We first generate haplotypes for k+1 variants with MAFs based on a latent variable $Z=(Z_1,\cdots,Z_{k+1})$ from a multivariate normal distribution with covariance structure $\text{cov}(Z_i, Z_j)=0.8^{|i-j|}$ between any two latent components. Then we combine two haplotypes to obtain the genotype value for each individual $X_i=(X_{i1},\cdots,X_{ik+1})$. The disease status of an individual is determined by the following logistic model:[15]

$$P(\text{Affected}|X_{ij}, i=1,\cdots,k+1)=\frac{1}{1+\exp(-\gamma)},$$
$$\gamma=\ln(\frac{c}{1-c})+\sum_{i=1}^{k+1}\ln(\text{OR}_i)\cdot X_{ij} \quad (10)$$

where c is a background chance of being affected for a subject with no minor alleles, OR_i is the effect size of variant i and X_{ij} is the number of copies of minor alleles at the ith variant. In Equation (10), we let $c=0.01$. We calculate the value of statistics T_M, and T_M-E using k observed common variants.

For RVs, we consider 20 RVs with q rare causal variants and 20-q rare non-causal variants. The MAFs of all variants are randomly determined with values ranging from $0.001\sim0.01$. We obtain the genotype value for each individual in the same way as for common variant but with covariance structure $\text{cov}(Z_i, Z_j)=0.4^{|i-j|}$ between various components. In order to express possible situations for the effects of RVs, we consider three scenarios under the alternative hypothesis of association: scenario A is that variants associated with disease have the same OR value, scenario B is that variants associated with disease are all deleterious but having different effects and scenario C is that variants associated with disease can be both deleterious and protective having different effects. In scenario A, we let OR = 3 for all causal variants. In scenario B, we let $\text{OR}\in[1.2, 3]$ with increments of $\frac{1.8}{q-1}$ for causal variant 1 to variant q. In scenario C, we let $\text{OR}\in[1.2, 3]$ for half of causal variants and $\text{OR}\in[0.2, 0.8]$ for the rest causal variants. At the same time, we consider weighting scheme with weight w_{MB}. Other parameter values are similar to those for common variants. We calculate the statistics of Sum, aSum-Ord, Price-VT, aSum-E, T_M and T_M-E. For all the statistics, P-values are estimated as the proportion of the permutation-based statistics that are larger than

Table 1 The estimated type I error rates when there are 20 rare variants

Type I error rates

Sample size	*Sum*	*aSum-ord*	*Price-VT*	*aSum-E*	T_M	T_M-E
500	0.058 (0.002)	0.051 (0.009)	0.052 (0.006)	0.050 (0.007)	0.053 (0.007)	0.052 (0.005)
1000	0.051 (0.005)	0.056 (0.004)	0.054 (0.005)	0.047 (0.005)	0.047 (0.005)	0.049 (0.005)
1500	0.052 (0.005)	0.051 (0.004)	0.050 (0.005)	0.049 (0.004)	0.053 (0.005)	0.050 (0.004)
2000	0.053 (0.004)	0.051 (0.005)	0.052 (0.005)	0.052 (0.005)	0.051 (0.003)	0.051 (0.004)

Note: shown in parentheses is the standard error.

Table 2 The estimated type I error rates and power for common variant analysis with a number of common variants where the sample size is 1000

Type I error rates						*Power*				
	# of common variants					*# of common variants*				
Test	*4*	*10*	*20*	*50*	*100*	*4*	*10*	*20*	*50*	*100*
T_M	0.05 (0.004)	0.05 (0.004)	0.052 (0.005)	0.051 (0.005)	0.053 (0.004)	0.908 (0.01)	0.807 (0.004)	0.766 (0.006)	0.725 (0.007)	0.614 (0.008)
T_M-E	0.049 (0.004)	0.053 (0.005)	0.052 (0.005)	0.053 (0.005)	0.055 (0.004)	0.931 (0.011)	0.841 (0.009)	0.806 (0.004)	0.771 (0.008)	0.635 (0.009)

Note: shown in parentheses is the standard error.

Table 3 Empirical power for RV analysis

	The number of non-causal variants in 20 RVsr									
	0		*4*		*8*		*12*		*16*	
Test	w = *1*	w = w_{MB}	w = *1*	w = w_{MB}	w = *1*	w = w_{MB}	w = *1*	w = w_{MB}	w = *1*	w = w_{MB}
Scenario A										
Sum	0.970 (0.005)	0.972 (0.006)	0.761 (0.007)	0.762 (0.008)	0.549 (0.005)	0.560 (0.007)	0.349 (0.010)	0.340 (0.009)	0.210 (0.009)	0.207 (0.010)
aSum-Ord	0.958 (0.009)	0.960 (0.007)	0.902 (0.006)	0.900 (0.006)	0.811 (0.006)	0.814 (0.005)	0.705 (0.009)	0.710 (0.006)	0.571 (0.008)	0.575 (0.007)
Price-VT	0.952 (0.012)	0.958 (0.010)	0.864 (0.011)	0.866 (0.010)	0.701 (0.006)	0.700 (0.005)	0.689 (0.008)	0.691 (0.007)	0.563 (0.009)	0.561 (0.007)
aSum-E	0.951 (0.011)	0.955 (0.010)	0.898 (0.011)	0.899 (0.011)	0.806 (0.004)	0.804 (0.004)	0.717 (0.007)	0.717 (0.006)	0.611 (0.008)	0.616 (0.009)
T_M	0.910 (0.006)	0.910 (0.006)	0.811 (0.009)	0.811 (0.009)	0.740 (0.010)	0.740 (0.010)	0.678 (0.012)	0.678 (0.012)	0.506 (0.013)	0.506 (0.013)
T_M-E	0.929 (0.011)	0.929 (0.011)	0.840 (0.010)	0.840 (0.010)	0.758 (0.011)	0.758 (0.011)	0.687 (0.011)	0.687 (0.011)	0.571 (0.012)	0.571 (0.012)
Scenario B										
Sum	0.935 (0.008)	0.936 (0.007)	0.750 (0.008)	0.768 (0.009)	0.523 (0.009)	0.529 (0.008)	0.345 (0.007)	0.343 (0.009)	0.213 (0.013)	0.212 (0.012)
aSum-Ord	0.942 (0.009)	0.947 (0.009)	0.901 (0.010)	0.919 (0.011)	0.704 (0.010)	0.702 (0.009)	0.701 (0.010)	0.707 (0.011)	0.625 (0.012)	0.630 (0.011)
Price-VT	0.918 (0.004)	0.911 (0.005)	0.850 (0.006)	0.856 (0.006)	0.669 (0.006)	0.670 (0.007)	0.678 (0.011)	0.686 (0.011)	0.579 (0.010)	0.569 (0.011)
aSum-E	0.928 (0.008)	0.931 (0.007)	0.893 (0.008)	0.895 (0.007)	0.720 (0.009)	0.722 (0.008)	0.712 (0.010)	0.716 (0.011)	0.623 (0.009)	0.628 (0.010)
T_M	0.801 (0.009)	0.801 (0.009)	0.773 (0.010)	0.773 (0.010)	0.686 (0.011)	0.686 (0.011)	0.651 (0.011)	0.651 (0.011)	0.573 (0.012)	0.573 (0.012)
T_M-E	0.818 (0.009)	0.818 (0.009)	0.800 (0.010)	0.800 (0.010)	0.714 (0.009)	0.714 (0.009)	0.702 (0.010)	0.702 (0.010)	0.593 (0.011)	0.593 (0.011)
Scenario C										
Sum	0.300 (0.006)	0.313 (0.005)	0.267 (0.008)	0.285 (0.009)	0.216 (0.006)	0.227 (0.006)	0.187 (0.009)	0.193 (0.009)	0.168 (0.008)	0.171 (0.007)
aSum-Ord	0.519 (0.006)	0.521 (0.006)	0.449 (0.012)	0.464 (0.011)	0.420 (0.008)	0.419 (0.007)	0.402 (0.009)	0.410 (0.010)	0.300 (0.008)	0.315 (0.009)
Price-VT	0.473 (0.008)	0.477 (0.007)	0.473 (0.009)	0.480 (0.009)	0.410 (0.009)	0.417 (0.010)	0.416 (0.012)	0.413 (0.011)	0.291 (0.009)	0.287 (0.008)
aSum-E	0.405 (0.003)	0.419 (0.006)	0.373 (0.008)	0.371 (0.008)	0.333 (0.008)	0.332 (0.009)	0.302 (0.009)	0.304 (0.006)	0.218 (0.007)	0.230 (0.009)
T_M	0.406 (0.003)	0.406 (0.003)	0.329 (0.008)	0.329 (0.008)	0.316 (0.006)	0.316 (0.006)	0.308 (0.008)	0.308 (0.008)	0.256 (0.010)	0.256 (0.010)
T_M-E	0.426 (0.008)	0.426 (0.008)	0.353 (0.009)	0.353 (0.009)	0.335 (0.007)	0.335 (0.007)	0.330 (0.009)	0.330 (0.009)	0.311 (0.010)	0.311 (0.010)

Note: scenario A, causal variants have the same effect. OR = 3; scenario B, causal variants have different effects with the same direction. OR $\in$ [1.2, 3] for causal variants; scenario C, causal variants have different effects. OR $\in$ [1.2, 3] for half of causal variants and OR $\in$ [0.2, 0.8] for the rest causal variants. $w = 1$ means no weighting and $w = w_{MB}$ means weighting. MAF of causal variants $\in$ [0.001, 0.01]. The sample size is 1000. Shown in parentheses is the standard error.

the data-based statistic by 1000 permutations. For a given significance level α (0.05), type I error rates and power are then estimated as the proportion of rejecting the null hypothesis when P-value $\leqslant \alpha$ in 1000 replications. Here, we repeat this simulation process 100 times and present the mean and the standard error for the estimated type I error rates and power.

Type I error rate and power

Table 1 exhibits the estimated type I error rates of Sum, aSum-Ord, Price-VT, aSum-E, T_M and T_M-E for RV, where the sample size n is 500, 1000, 1500 and 2000, respectively. As shown in Table 1, the type I error rates are all well-controlled. We list the results of T_M and T_M-E for common variant in Table 2 with the sample size of 1000. We found that the Type I error rates are also reasonable.

The results of power are presented in Table 2 for CV and Table 3 for RV when the sample size is 1000. From Table 2, we can see that the power of the multi-marker test T_M decreases with the increasing of the number of common variants. The entropy-based adaptive strategy can improve the power of T_M. Table 3 presents the power for six statistics, Sum, aSum-Ord, Price-VT, aSum-E, T_M and T_M-E. For each scenario, the power of these statistics decreases with the increasing of the number of non-causal variants. For collapsing method, there are four statistics, one is the Sum test and the other three are adaptive methods. We observed that, when there are rare non-causal variants, the Sum test has the lowest power, indicating that the Sum test is most seriously affected by non-causal variants. When the number of non-causal variants is <12, the statistic aSum-Ord has the highest power. We noted that, for the first two scenarios, with the number of non-causal variants increasing, the power of the aSum-E is gradually close to that of aSum-Ord and almost the same as that of aSum-Ord when the number of non-causal variants is 16, indicating that the entropy-based adaptive strategy can improve the power for the collapsing method. However, we found that, for scenario C where causal RVs have opposite association directions, the power of aSum-E is less than that of aSum-Ord.

For multi-marker test for RV, the power is higher than that of the Sum test when there are rare non-causal variants. Although the power is lower than that of collapsing method with adaptive strategy, the deference gradually decreases when the number of non-causal variants is increased. It can be found that the power improves by using the entropy-based adaptive strategy and the entropy-based adaptive strategy further decreases the difference between the multi-marker

test and the collapsing method with adaptive strategy. We also found that, although the power of multi-marker test decreases with the increasing of the number of non-causal variants, multi-marker test is least affected by non-causal variants. For example, with the number of non-causal variants increasing from 4 to 8, the power of T_M-E decreases from 0.801 to 0.714 with the decline rate of 10.86% while the decline rates of power for Sum, aSum-Ord, Price-VT and aSum-E are 30.27, 21.86, 21.29 and 19.37%, respectively.

It can also be seen from Table 3 that there exists difference for the power between three scenarios. The power in scenario A is close to that in scenario B, and powers in scenario A and scenario B are far higher than those in scenario C. This result showed that different direction of the effects of causal variants severely affect the power. Moreover, we also consider the smaller significance level. When we let the significance level be 0.001, we found that the estimated type I error rates are also close to the nominal levels and the results of power are similar to those in Table 2, Table 3 and as reflected by more data not shown here.

DISCUSSION

In this paper, we proposed a novel adaptive strategy using entropy theory for association analysis. We used the mutual information in entropy theory to measure the association between RVs and the disease. The mutual information can capture all linear and nonlinear dependencies between random variables and not just linear dependence as the correlation coefficient measures. In practice, the number of non-causal variants and the effects of causal variants are unknown. Misclassification of non-functional variants can seriously affect the power of collapsing methods for RV association analysis. Here, we proposed a strategy to diminish the influence of non-causal variants and search the optimal variants set associated with the disease in the studied genetic region to construct the statistical test.

Different from several existing adaptive methods which depend on the association pattern with causal variants, our method is based on the magnitude of association between variants and disease provided by the data. It can be used not only for common variants but also for RVs. For common variant, we considered the multi-marker test to construct the entropy-based adaptive strategy. We choose multivariate test mainly because it is a powerful method for association analysis of common variants or RVs and it is considered to be more robust than collapsing method for RVs analysis in the presence of misclassification of non-functional variants.[3] For RV, we considered the Sum test, a collapsing method to conduct RVs analysis. Using simulation study, we investigated the performance of our method and compared it with several existing adaptive methods. The results showed that our entropy-based adaptive strategy can improve the power of multi-marker test. At the same time, for RV analysis, our method can improve the power for the Sum test when there are non-causal variants and, achieve good performance similar to that of the Sum test with adaptive strategy proposed by Pan and Shen[9] when there is a large number of non-causal variants and causal variants have positive effects. These results indicate that our method is a general approach to reduce the noise incurred by non-causal variants.

Although our method is for population-based design, it can be easily extended to family-based analysis. For example, when we obtain case-parents data, we use nontransmitted genotypes as complement of affected offspring and construct a difference vector calculated by comparing the genotypes of affected offspring with their corresponding 'complements'. In this way, we can transform the family-based data and apply case–control statistical tests. In a future study, we will focus on family-based analysis.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

LYM was supported by National Natural Science Foundation of China (11301206), Foundation of Hunan Educational Committee (16A166), Hunan Provincial Natural Science Foundation of China (2017JJ2212) and China Scholarship Council. HWD was partially supported by grants from the National Institutes of Health [R01AR057049, R01AR059781, D43TW009107, P20 GM109036, R01MH107354, R01MH104680 and R01GM109068], the Edward G. Schlieder Endowment fund to Tulane University.

1 Maher, B. Personal genomes: the case of the missing heritability. *Nature* **456**, 18–21 (2008).
2 Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multiallelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res.* **615**, 28–56 (2007).
3 Li, B. S. & Leal, S. M. Methods for detecting association with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
4 Pan, W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Epidemiology* **33**, 497–507 (2009).
5 Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighter sum statistic. *PLoS Genet.* **5**, e1000384 (2009).
6 Price, A. L., Kryukov, G. V., Bakker, P. I. W., Purcell, S. M., Staples, J., Wei, L. J. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86**, 832–838 (2010).
7 Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet.* **7**, e1001322 (2011).
8 Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. & Lin, X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
9 Pan, W. & Shen, X. T. Adaptive tests for association analysis of rare variants. *Genet. Epidemiol.* **35**, 381–388 (2011).
10 Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).
11 Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
12 Hampe, J., Schreiber, S. & Krawczak, M. Entropy-based SNP selection for genetic association studies. *Hum. Genet.* **114**, 36–43 (2003).
13 Zhao, J., Boerwinkle, E. & Xiong, M. An entropy-based statistic for genomewide association studies. *Am. J. Hum. Genet.* **77**, 27–40 (2005).
14 Li, Y. M. & Xiang, Y. Genotype-based association analysis via entropy. *J. Hum. Genet.* **57**, 734–737 (2012).
15 Preston, M. D. & Dudbridge, F. Utilising family-based designs for detecting rare variant disease associations. *Ann. Hum. Genet.* **78**, 129–140 (2014).