European Journal of Human Genetics (2017) 25, 752–757

www.nature.com/ejhg

ARTICLE

An efficient and flexible test for rare variant effects

Shonosuke Sugasawa[1], Hisashi Noma*[,2,3], Takahiro Otani[1,3], Jo Nishino[3,4] and Shigeyuki Matsui[3,4]

Since it has been claimed that rare variants with extremely small allele frequency play a crucial role in complex traits, there is great demand for the development of a powerful test for detecting these variants. However, due to the extremely low frequencies of rare variants, common statistical testing methods do not work well, which has motivated recent extensive research on developing an efficient testing procedure for rare variant effects. Many studies have suggested effective testing procedures with reasonably high power under some presumed assumptions of parametric statistical models. However, if the parametric assumptions are violated, these tests are possibly under-powered. In this paper, we develop an optimal, powerful statistical test called the aggregated conditional score test (ACST) for simultaneously testing *M* rare variant effects without restrictive parametric assumptions. The proposed test uses a test statistic aggregating the conditional score statistics of effect sizes of *M* rare variants. In simulation studies, ACST generally performed well compared with the two most commonly used tests, the optimal sequence kernel association test (SKAT-O) and Kullback–Leibler distance test. Finally, we demonstrate the performance and practical utility of ACST using the Dallas Heart Study data.

European Journal of Human Genetics (2017) **25,** 752–757; doi:10.1038/ejhg.2017.43; published online 12 April 2017

INTRODUCTION

While genome-wide association studies have identified many common variants associated with complex traits, it has also been recognized that rare variants with minor allele frequency (MAF) smaller than 1–5% play an important role in complex traits.[1] In spite of the importance of these variants, testing for their association with traits is a challenging problem. The main reason is the extremely low frequency of rare variants, which leads to low power of single-variant tests. Therefore, we collected several relevant rare variants and tested for their joint association with various traits, which is known as the region-based rare variant test and is now becoming the standard method for detecting rare variants.[2]

Many researchers have already proposed effective region-based testing methods. The earliest methods are burden-based tests, which summarize the information of each rare variant.[3–6] It is well-known that burden-based tests suffer from low power when there are large numbers of non-causal variants or both protective and deleterious variants.[7] To overcome the limitations of burden-based tests, the sequence kernel association test (SKAT)[8] using the variance component test has been proposed, along with a composite version of SKAT called the optimal sequence kernel association test (SKAT-O).[9] It has been shown that SKAT-O has higher power than both burden-based tests and SKAT in a wide range of scenarios.[9] However, SKAT-O corresponds to an efficient testing method for variance components under a logistic regression model that assumes the random-effects regression parameters are normally distributed with mean 0, so that the optimality of SKAT-O is possibly violated when unknown and irrelevant parametric assumptions are not correct.

As an alternate testing method, a new rare variant test called the Kullback–Leibler distance test (KLT)[10] using the Kullback–Leibler distance has been recently proposed. The procedure permits the straightforward comparison of two distributions over *M* rare variants divided by case and control, and uses the Kullback–Leibler distance of the two distributions as the test statistics. The authors reported that the suggested testing method performed better than SKAT-O in their simulation studies. However, there are no theoretical arguments concerning the theoretical optimality of KLT, defined for instance as having the greatest statistical power. Since the statistical power to detect disease-related rare variants is usually insufficient regardless of the specific test used, the possible non-optimality might yield serious losses of efficiency in practice. In fact, in some scenarios in our simulation studies, presented in the Results section, KLT was under-powered compared to SKAT-O.

In this paper, we propose a new and efficient testing procedure called the aggregated conditional score test (ACST). The basic idea of the test is that we jointly test the association between a single variant and disease status over *M* variants. Specifically, we calculate the conditional score statistics for effect sizes of *M* variants, and we aggregate these statistics for simultaneously testing the association between *M* variants and disease status. We propose two aggregation methods, both of which hold optimality under certain correlation structures among *M* variants. Hence, the optimality of ACST is expected to result in the efficient identification of disease-related variant sets. Moreover, ACST does not require any presumed structures of effect sizes, unlike SKAT-O, so that the theoretical optimality of ACST is assured under a wide range of structures of regression parameters. In fact, in our simulation study in the Results section, ACST generally performed well compared with SKAT-O and KLT. Moreover, through application to the Dallas Heart Study data set in the Results section, ACST was revealed to work well as a useful tool for the analysis of genome-wide sequencing data.

[1]Risk Analysis Research Center, The Institute of Statistical Mathematics, Tokyo, Japan; [2]Department of Data Science, The Institute of Statistical Mathematics, Tokyo, Japan; [3]Core Research for Evolutional Science and Technology (CREST), Japan Science and Technology Agency, Tokyo, Japan; [4]Department of Biostatistics, Graduate School of Medicine, Nagoya University, Aichi, Japan
*Correspondence: Dr H Noma, Department of Data Science, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan. Tel: +81 50 5533 8440; E-mail: noma@ism.ac.jp
Received 24 August 2016; revised 13 February 2017; accepted 28 February 2017; published online 12 April 2017

METHODS

Notations and model

We consider a data set with N subjects, among which n_1 are cases and n_2 are controls. For the ith subject, $i=1,\ldots,N$, we observe a phenotype y_i and a multi-site genotype $G_i=(g_{i1},\ldots,g_{iM})^{\mathrm{T}}$, where y_i takes values 0 or 1, representing control and case, respectively, and g_{ik}, $k=1,\ldots,M$, are coded as 0, 1 or 2, representing the number of minor alleles that subject i holds in the kth variant. It is noted that the MAFs of most M variants are extremely low and our goal is to test the association between G_i, a set of (rare) variants, and a disease status y_i. To this end, region-based tests have been widely studied[2] since a test for an individual variant with extremely low MAF cannot be expected to achieve sufficient power to detect its effect.

We consider the following standard logistic regression model that has been widely adopted in genetic association studies:

$$\log\left(\frac{p_{ik}}{1-p_{ik}}\right)=r_k+\beta_k g_{ik}, i=1,\ldots,N, k=1,\ldots,M$$

where p_{ik} is the probability of $y_i=1$ caused by g_{ik}, and r_k denotes a variant-specific intercept. Our interest is the effect size β_k rather than r_k, namely r_k is a nuisance parameter. To perform the region-based test for detecting rare variant effects simultaneously, we consider the testing problem:

H_0: $\beta_1=\beta_2=\cdots=\beta_M=0$ *vs* H_1: $\beta_k\neq 0$ for at least one of k $(k=1,2,\ldots,M)$.

As mentioned before, the existing methods have several restrictions and limitations such as the strong parametric assumption of $(\beta_1,\beta_2,\ldots,\beta_M)^{\mathrm{T}}$ as in SKAT-O[9] and the lack of theoretical optimality as in KLT.[10] Therefore, to solve these problems, we construct an optimal test that achieves the most powerfulness without any parametric assumptions of $\beta_1,\beta_2,\ldots,\beta_M$. To this end, we first derive the conditional score statistic of the null hypothesis $\beta_k=0$ under the logistic regression model since it is known that the conditional score test has the greatest power and that conditional inference can eliminate the effect of a variant-specific intercept r_k. Subsequently, we suggest the test statistic for simultaneously testing $\beta_1=\beta_2=\cdots=\beta_M=0$ by aggregating all the conditional score statistics.

Aggregated conditional score test (ACST)

The conditional likelihood for β_k is expressed as

$$\mathrm{CL}_k(\beta_k)=\frac{\prod_{i:y_i=1}\exp\left(r_k+\beta_k g_{ik}\right)\prod_{i=1}^{N}\left\{1+\exp\left(r_k+\beta_k g_{ik}\right)\right\}^{-1}}{\sum_{D(n_1)}\left\{\prod_{i\in D(n_1)}\exp\left(r_k+\beta_k g_{ik}\right)\prod_{i=1}^{N}\left\{1+\exp\left(r_k+\beta_k g_{ik}\right)\right\}^{-1}\right\}}$$
$$=\frac{\exp\left(\beta_k\sum_{i:y_i=1}g_{ik}\right)}{\sum_{D(n_i)}\exp\left(\beta_k\sum_{i\in D(n_1)}g_{ik}\right)}$$

where $D(n_1)$ is an arbitrary subset of $\{1,\ldots,N\}$ with n_1 elements and $\sum_{D(n_1)}$denotes the summation over all the possible subsets $D(n_1)$. Note that the conditional likelihood $\mathrm{CL}_k(\beta_k)$ is free from the variant-specific intercept r_k. For notational simplicity, we introduce 2×3 contingency tables given in Table 1, in which we summarize the quantities used in the test statistics. Based on $\mathrm{CL}_k(\beta_k)$, the conditional score statistic s_k, $k=1,\ldots,M$, for testing $\beta_k=0$ is given by

$$s_k=\frac{N\sqrt{N-1}\{2a_{1k}+b_{1k}-N^{-1}n_1(2m_{1k}+m_{2k})\}}{\sqrt{n_1n_2(4m_{1k}m_{3k}+m_{1k}m_{2k}+m_{2k}m_{3k})}}$$

where the detailed derivation is deferred to the Appendix. It is well-known that the conditional score test for β_k based on s_k is most powerful.[11]

To perform a test for $\beta_1,\ldots,\beta_M$, simultaneously, we need to aggregate these scores. The widely used method for aggregating variant-specific statistics is summing up squared statistics.[12,13] Hence, we first propose the test statistic $U_1=S^{\mathrm{T}}S$, where $S=(s_1,\ldots,s_M)^{\mathrm{T}}$, which achieves asymptotically greater power when all the variants are independent. However, it is often observed that there is a linkage disequilibrium (LD) structure among variants, so that variants are correlated; thereby the score statistics s_ks are mutually correlated. In this case, U_1 is not necessarily efficient.

Table 1 2×3 contingency tables between phenotype variable $y_i, i=1,\ldots,N$, representing the case ($y_i=1$) and control ($y_i=0$), and genotype variable $g_{ik}, k=1,\ldots,M$, representing the number of minor alleles

	$g_{ik}=2$	$g_{ik}=1$	$g_{ik}=0$	*Sum*
$y_i=1$	a_{1k}	b_{1k}	c_{1k}	n_1
$y_i=0$	a_{2k}	b_{2k}	c_{2k}	n_2
Sum	m_{1k}	m_{2k}	m_{3k}	N

To adapt the LD structure, we propose to allow the score vector S to have an exchangeable correlation matrix, $R_\rho=(1-\rho)I+\rho 11^{\mathrm{T}}$, as used in SKAT-O. Using the correlation structure, we consider the statistic as the function of ρ : $Q_\rho=S^{\mathrm{T}}R_\rho^{-1}S$. When $\rho=0$, Q_ρ reduces to U_1. When $\rho=1$, we define R_ρ^{-1} as the generalized inverse matrix satisfying $R_\rho R_\rho^{-1}R_\rho=R_\rho$, that is, $R_\rho^{-1}=M^{-2}11^{\mathrm{T}}$. Hence, when $\rho=1$, $Q_\rho=M^{-2}\left(\sum_{k=1}^{M}s_k\right)^2$ and M^2Q_ρ corresponds to the well-known Mantel–Haenszel test statistic in 2×3 contingency tables. For a fixed ρ, Q_ρ follows a mixture of χ^2 distributions for large N since the score vector S asymptotically follows a multivariate normal distribution with mean vector 0 and some covariance matrix C with all the diagonal elements 1. Specifically, let $\lambda_1,\ldots,\lambda_M$ be the eigenvalues of $R_\rho^{-1}\hat{C}$, then the null distribution of Q_ρ can be closely approximated by $\sum_{k=1}^{M}\lambda_k\chi_{1k}^2$, where $\chi_{11}^2,\ldots,\chi_{1M}^2$ are mutually independent χ_1^2 random variables. For estimation of the correlation matrix $C=\left(c_{kj}\right)_{1\le k,j\le M}$, we note that $2a_{1k}+b_{1k}=\sum_{i:y_i=1}g_{ik}$ and

$$c_{kj}=\mathrm{Cor}\left(\sum_{i:y_i=1}g_{ik},\sum_{i:y_i=1}g_{ij}\right)=\mathrm{Cor}\left(g_{1k},g_{1j}\right)$$

which can be estimated by the sample correlation between $(g_{1k},\ldots,g_{Nk})^{\mathrm{T}}$ and $(g_{1j},\ldots,g_{Nj})^{\mathrm{T}}$. Then we can compute the P-value of Q_ρ for each fixed ρ. However, there is little information about unknown parameter ρ in applications, thereby we propose selecting the optimal value of ρ to maximize the power similarly to SKAT-O.[9] Hence the proposed test statistic is

$$U_2=\inf_{0\le\rho\le 1}p_\rho$$

where p_ρ is the P-value of Q_ρ. In practice, the test statistic U_2 can be computed by the simple grid search, namely $U_2=\min\{p_{\rho_1},\ldots,p_{\rho_\ell}\}$ for $-1<\rho_1<\cdots<\rho_\ell<1$. Since the correlations among variants are not large, it could be enough to search the optimal ρ around 0. Hence, we set the default grids as 21 points from −0.1 to 0.1 with equal intervals.

We call the test using the conditional score statistic s_k the ACST, and, in particular, we refer to the two tests using U_1 and U_2 as the independent ACST (ACST-I) and the correlation-adjusted ACST (ACST-C), respectively. Note that ACST is asymptotically most powerful without any restriction among $\beta_1,\ldots,\beta_M$, so that ACST is flexible and theoretically efficient compared, for example, with SKAT-O and KLT. It should be noted that the score statistic s_k cannot be computed when both m_{1k} and m_{2k} are 0, namely there are no minor alleles in the kth variant. Hence, we need to omit variants with no minor alleles in advance in order to perform ACST.

In case $g_{ik}=0,1, k=1,\ldots,M$, that is, there are no subjects carrying two minor alleles, corresponding to $a_{ik}=a_{2k}=m_{1k}=0$ in Table 1, the score statistic s_k reduces to the form

$$s_k=\frac{N\sqrt{N-1}(b_{1k}-N^{-1}n_1m_{2k})}{\sqrt{n_1n_2m_{2k}m_{3k}}}$$

and the sum of these scores corresponds to the Mantel–Haenszel test statistic. Hence, the score statistic s_k can be regarded as a generalization of the score statistic used in the Mantel–Haenszel test in 2×3 contingency tables. It should be noted that ACST achieves the greatest power if the stratum-specific log odds ratios $\beta_1,\ldots,\beta_M$ are heterogeneous (for arbitrary values of $\beta_1,\ldots,\beta_M$ under H_1), while the Mantel–Haenszel test was derived as an asymptotically efficient test under the common effect assumption across the strata.[14,15] Thus, ACST, a generalization of the Mantel–Haenszel test, maintains optimality under a broad range of conditions regardless of the homogeneity of the effect measures, and this fact is quite meaningful in the context of region-based simultaneous testing to efficiently detect disease-related rare variants.

Calculation of the *P*-value

Concerning the calculation of *P*-values of ACST, we propose a permutation method since it enables us to compute adequate *P*-values regardless of sample size N and number of variants M. Without loss of generality, we demonstrate a permutation test only of ACST-I. We randomly shuffle the disease status y_i of all subjects in the sample, and we calculate the test statistic $U_1^{(b)}$ from the permuted data. We repeat this process B times to calculate the *P*-value as

$$p = B^{-1} \sum_{b=1}^{B} I\left(U_1^{(b)} \geq U_1\right)$$

where $I(\cdot)$ denotes the indicator function. It should be noted that ACST-C takes much more running time than ACST-I since ACST-C needs to compute the minimum of *P*-values in each permutation.

RESULTS

Simulation study: evaluation of type-I error rates

To evaluate the performance of ACST compared with two existing methods, SKAT-O and KLT, under realistic situations, we carried out simulation studies. To begin with, we evaluated the type-I error rate of ACST and those of KLT and SKAT-O. We considered $M = 20$ variants and they were divided into four groups with five variants in each group. The MAFs of each rare variant in the same group were set to be equal. With $(m_1,\ldots,m_4)$ as the combination of MAFs in the four groups, we considered the following three patterns:

$$(A)(0.002, 0.003, 0.004, 0.005)(B)(0.003, 0.005, 0.008, 0.010)$$
$$(C)(0.005, 0.010, 0.015, 0.020)$$

For generating genotype data, we first generated two M-dimensional binary vectors a_1, a_2 using rmvbin function in R with a correlation matrix $R = \left(\rho^{|i-j|}\right)_{i,j=1,\ldots,M}$ with $\rho = 0.05$. Then we set the genotype data $G = (g_1,\ldots,g_M)^{\mathrm{T}}$ as $G = a_1 + a_2$. For evaluating type-I error rates, we generated the disease status y of each subject from the null logistic regression model:

$$\operatorname{logit} P(y = 1) = \beta_0$$

In the above model we used $\beta_0 = -\log 4$, corresponding to a 20% background disease prevalence. From this model, we generated $5n$ samples and randomly selected $n/2$ cases and controls for $n = 2000$. Then, we applied four tests ACST-C, ACST-I, KLT and SKAT-O to the generated data set with significance level $\alpha = 0.05$,0.01 and 0.001. We used 2000 permutations for calculating *P*-values of ACST-C, ACST-I and KLT. Based on 1000 simulation runs when $\alpha = 0.05$ and 0.01, and 5000 runs when $\alpha = 0.001$, we computed the simulated type-I error rates of the four tests, which are presented in Table 2. It is observed that the type-I error rates of the four methods are around the nominal significance level, so all the four procedures can adequately control the type-I error rates.

Table 2 Simulation results: type-I error comparison among ACST-C, ACST-I, KLT and SKAT-O at significance levels $\alpha = 0.01$, 0.05 and 0.001

Pattern	α	*ACST-C*	*ACST-I*	*KLT*	*SKAT-O*
(A)	0.05	0.051	0.049	0.050	0.039
	0.01	0.011	0.009	0.012	0.012
	0.001	0.0010	0.0008	0.0010	0.0002
(B)	0.05	0.048	0.046	0.049	0.036
	0.01	0.009	0.010	0.009	0.011
	0.001	0.0008	0.0006	0.0007	0.0006
(C)	0.05	0.045	0.045	0.042	0.043
	0.01	0.015	0.013	0.013	0.012
	0.001	0.0009	0.0010	0.0009	0.0006

Abbreviations: ACST-C, aggregated conditional score test using the exchangeable correlation structure among variants; ACST-I, aggregated conditional score test assuming independence among variants; KLT, Kullback–Leibler test; SKAT-O, optimal sequence kernel association test. The type-I errors were calculated based on 1000 simulated runs when $\alpha = 0.05$ and 0.01, and 5000 runs when $\alpha = 0.001$. Two thousand permutations were used for calculating *P*-values of ACST-C, ACST-I and KLT.

Simulation study: evaluation of power

We next evaluated the statistical power of the four tests, ACST-C, ACST-I, KLT and SKAT-O, for detecting disease-related rare variants via simulation studies. We generated $M = 20$ variants in the same way as in the previous simulation, the correlation parameter ρ was set to 0, 0.02 and 0.05, and (B) pattern of MAFs was considered in this study. To evaluate the power, we used the following non-null logistic model:

$$\operatorname{logit} P(y = 1) = \beta_0 + \sum_{k=1}^{M} \beta_k g_k$$

where $\beta_0 = -\log 4$. For the setups of the non-zero effect sizes $\beta_k, k = 1, \ldots, M$, we considered the following eight scenarios:

1. $\beta_k = 0.3, k = 1, \ldots, M$
2. $(\beta_1, \ldots, \beta_M)^{\mathrm{T}} \sim 0.2N((0, \ldots, 0)^{\mathrm{T}}, 0.3I + 0.711^{\mathrm{T}})$
3. $\beta_k = 0.05\times |\log(\mathrm{MAF}_k)| \times u_k, P(u_k = 1) = P(u_k = -1)$
   $= 0.5, k = 1, \ldots, M$
4. $P(\beta_k = 0.3) = P(\beta_k = -0.3) = 0.3, k = 1, \ldots, M$
5. $\beta_{2k} = -0.5, \beta_{2k-1} = 0.5, k = 1, 2, 3, 4$
6. $\beta_1 = -0.6, \beta_2 = -0.4, \beta_3 = -0.2, \beta_4 = 0.2, \beta_1$
   $= 0.4, \beta_6 = 0.6$
7. $\beta_k = 0.5, k = 1, 2, 3, 4$
8. $\beta_1 = \beta_{15} = 1, \beta_8 = -1$

It is noted that the number of causal variants gets smaller as the scenario number gets larger. In Supplementary Table S1, we show the number of causal variants in each MAF group in the eight scenarios. In scenarios 1 and 2, all the variants are causal and deleterious, in which the parametric assumption in SKAT-O seems reasonable. In scenario 3, rarer variants have larger effect sizes while there exist both deleterious and protective variants. In scenarios 4, 5 and 6, some variants are causal, which are deleterious or protective. In scenarios 7 and 8, a small portion of variants are causal while the effect sizes are relatively large in scenario 8. In each scenario, we generated $5n$ samples and randomly selected $n/2$ cases and controls with $n = 2000$, and applied the four testing methods with a significance level $\alpha = 0.05$. In applying ACST-C, ACST-I and KLT, we used 2000 permutations to obtain *P*-values. Based on 1000 simulation runs, we computed the simulated powers in eight scenarios, which are shown in Figure 1. It is observed that SKAT-O performs quite well when all the variants are causal and deleterious like in scenarios 1 and 2. However, when protective variants are included or the number of causal variants is small, the result reveals that SKAT-O tends to be under-powered. Concerning KLT, it seems to perform well when the number of causal variants is small as in scenarios from 3 to 8. However, we can observe

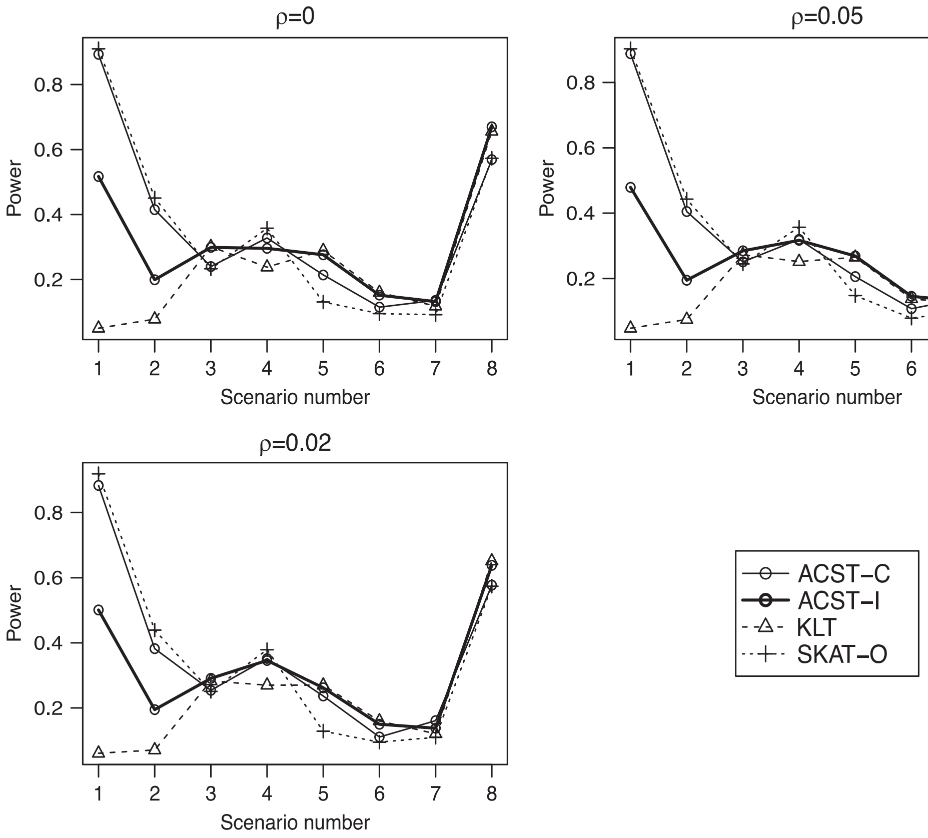

Figure 1 Simulation results: Power comparisons among ACST-C (aggregated conditional score test using the exchangeable correlation structure among variants), ACST-I (independent aggregated conditional score test assuming independence among variants), KLT (Kullback–Leibler test) and SKAT-O (optimal sequence kernel association test) at significance level $\alpha = 0.05$, number of subjects $n = 2000$ and correlation parameter $\rho = 0$, 0.02 and 0.05. The empirical powers were calculated based on 1000 simulated data sets. Two thousand permutations were used for computing *P*-values for ACST-C, ACST-I and KLT.

Table 3 Simulation results: power comparisons among ACST-C, ACST-I, KLT and SKAT-O at significance level $\alpha = 0.001$, number of subjects $n = 2000$ and $\rho = 0.05$

Scenario	*ACST-C*	*ACST-I*	*KLT*	*SKAT-O*
2	0.0860	0.0105	0.0020	0.0690
5	0.0030	0.0185	0.0170	0.0020
8	0.1040	0.1545	0.1595	0.1210

Abbreviations: ACST-C, aggregated conditional score test using the exchangeable correlation structure among variants; ACST-I, aggregated conditional score test assuming independence among variants; KLT, Kullback–Leibler test; SKAT-O, optimal sequence kernel association test. The empirical powers were calculated based on 2000 simulated data sets. Two thousand permutations were used in ACST-C, ACST-I and KLT for computing *P*-values.

that KLT is extremely under-powered in scenarios 1 and 2. On the other hand, both ACST-I and ACST-C provide reasonable powers in all scenarios. It is worth noting that ACST-C provides the almost same powers as SKAT-O, while it is under-powered compared with ACST-I in scenarios 5 and 6. In Supplementary Figure S1, we also provide a power comparison as a function of a quantity determined by the effect sizes and MAFs.

We next evaluated the statistical powers of the four tests with smaller significance level $\alpha = 0.001$ in three scenarios 2, 5 and 8 with $\rho = 0.05$. We used 2000 permutations for ACST-C, ACST-I and KLT for computing *P*-values. Based on 2000 simulation runs, the empirical powers were calculated, which are presented in Table 3. It is observed that power relations among the four tests are not different from the results with $\alpha = 0.05$.

Applications to the Dallas Heart Study

We applied ACST together with KLT and SKAT-O to the sequence data from the Dallas Heart Study[16] to test the association between serum triglyceride (TG) levels and rare variants in three genes (*ANGPTL3*, *ANGPTL4* and *ANGPTL5*). The data set was also used in both papers presenting SKAT-O[9] and KLT,[10] thereby we examined how ACST performs compared with SKAT-O and KLT. The data set has sequence information on 95 observed variants in the three genes from each of 3474 individuals, including 1830 African Americans, 1043 European Americans and 601 Hispanics. The higher levels of TG in blood are known to be related to some metabolic disease such as diabetes, coronary heart disease and fatty liver disease. It has been revealed that *ANGPTL3*, *ANGPTL4* and *ANGPTL5* are associated with lower levels of TG.[6,17–20] Most of the variants are rare: with the exception of one variant, the estimated allele frequencies of the variants are under 0.05. We considered a dichotomized trait by classifying individuals with the top *q*% of TG values as cases and the bottom *q*% as controls, and we considered three conditions, with $q = 15$, 20 or 25. For each choice of *q*, we deleted variants that had no sequence variation among the case and control samples. In Table 4, we show the sample sizes of cases and controls for each *q* and the number of variants used for testing. In applying ACST-C, ACST-I and KLT, the *P*-values were computed based on 10^6 permutations.

Table 4 Results of the analyses of the Dallas Heart Study data: The *P*-values of ACST-C, ACST-I, KLT and SKAT-O

Trait	*Gene*	*#Variants*	*ACST-C*	*ACST-I*	*KLT*	*SKAT-O*
$q=15$	ANGPTL3	17	$<1.00\times10^{-6}$	$<1.00\times10^{-6}$	$<1.00\times10^{-6}$	6.28×10^{-8}
Case: 515	ANGPTL4	10	$<1.00\times10^{-6}$	$<1.00\times10^{-6}$	$<1.00\times10^{-6}$	3.01×10^{-6}
Control: 504	ANGPTL5	14	3.06×10^{-1}	8.89×10^{-2}	5.92×10^{-2}	1.80×10^{-1}
$q=20$	ANGPTL3	20	$<1.00\times10^{-6}$	$<1.00\times10^{-6}$	2.00×10^{-5}	9.89×10^{-8}
Case: 692	ANGPTL4	12	$<1.00\times10^{-6}$	$<1.00\times10^{-6}$	$<1.00\times10^{-6}$	3.40×10^{-8}
Control: 662	ANGPTL5	16	2.62×10^{-2}	7.47×10^{-3}	4.08×10^{-3}	5.32×10^{-2}
$q=25$	ANGPTL3	22	$<1.00\times10^{-6}$	$<1.00\times10^{-6}$	$<1.00\times10^{-6}$	9.31×10^{-9}
Case: 860	ANGPTL4	15	$<1.00\times10^{-6}$	$<1.00\times10^{-6}$	$<1.00\times10^{-6}$	5.96×10^{-9}
Control: 835	ANGPTL5	18	2.98×10^{-2}	1.27×10^{-2}	6.07×10^{-3}	6.15×10^{-2}

Abbreviations: ACST-C, aggregated conditional score test using the exchangeable correlation structure among variants; ACST-I, aggregated conditional score test assuming independence among variants; KLT, Kullback–Leibler test; SKAT-O, optimal sequence kernel association test.
The permutation *P*-values for ACST-C, ACST-I and KLT were computed using 10^6 permutations. The calculated *P*-values of ACST-C, ACST-I and KLT with 0 are expressed as $<1.00\times10^{-6}$.

The *P*-values of ACST-C, ACST-I, KLT, and SKAT-O are given in Table 4. The results for *ANGPTL3* and *ANGPTL4* did not differ among the four testing methods. However, it reveals that both ACST-C and ACST-I provide smaller *P*-values than SKAT-O in *ANGPTL5*, which is has been shown to be associated with serum triglycerides. On the other hand, KLT produces smaller *P*-values than both ACST-C and ACST-I while the differences are relatively small. Moreover, the *P*-value of ACST-I was smaller than that of ACST-C in this case.

Concerning the running time of these methods, it takes 336 s in ACST-I, 6 h 25 min 54 s in ACST-C, 174 s in KLT and 0.1 s in SKAT-O for computing *P*-values of ANGPTL3 with $q=15$. It is observed that ACST-I takes almost as long as KLT compared with SKAT-O since both ACST-I and KLT requires permutations for computing *P*-values. On the other hand, ACST-C takes a much more running time than ACST-I since ACST-C requires grid search for optimal correlation parameter in each permutation. The program was run on a PC with a 2.7 GHz Intel Core i5-4570R Quad Core Processor with approximately 8GB RAM.

DISCUSSION

We developed a new optimal test, ACST, to detect the association between a phenotype and a set of rare variants. We derived the conditional score statistics for testing the association between a phenotype and each single variant, and proposed two methods for aggregating these score statistics. The first method involves simply summing the squared score statistics, and the resulting test using the statistic is called ACST-I, which is most powerful if all the variants are independent. The second method is called ACST-C, and involves aggregation with a quadratic form, assuming the correlation matrix of the conditional scores is an exchangeable correlation matrix with tuning parameter ρ. We developed a grid search technique for selecting ρ to maximize the power. The *P*-values of both ACST approaches can be calculated using a permutation test.

In the simulation study of power comparison, we considered eight scenarios with various effect sizes and evaluated the power of ACST as well as KLT and SKAT-O. We found that SKAT-O tends to be under-powered when there was a large proportion of null variants or protective variants exist, while KLT was extremely under-powered when all variants are causal and deleterious. In comparison, ACST performs well in both scenarios since ACST is asymptotically most powerful under arbitrary effect sizes. On the other hand, the power of ACST is smaller than that of SKAT-O where the parametric assumption of SKAT-O seems correct while the differences are not large and ACST has still larger power than KLT. Concerning the simulation settings, the background prevalence should be smaller than that was used in our studies (20%) in terms of biological plausibility. However, the background prevalence is associated only with the intercept term and does not affect the superiority and inferiority relationship among the four tests. In the applications to DHS data set, ACST performed better than SKAT-O, but ACST produced larger *P*-values than KLT in most cases. However, since the differences were quite small and KLT might be extremely under-powered in some cases, the use of ACST is justified.

Concerning usage of ACST-I and ACST-C, we first note that ACST-C is optimal even if rare variants to be tested are correlated. Hence, ACST-C should be recommended from a theoretical point of view. However, since it cannot be assumed that the correlations among rare variants are large in this context, ACST-I is expected to perform as well as ACST-C, which can be observed from the results in simulation study.

In this paper, we considered the case without any covariates except for genotype data. However, clinical covariates are often associated with disease status, which could improve the statistical power. Since the conditional score statistics under adjustment for covariates can be computed by using the conditional logistic regression,[11] the extension of ACST seems straightforward. However, the detailed investigation is out of the scope of the paper and is left to a future study.

We used the exchangeable correlation structure in ACST-C for modeling correlations among the conditional score statistics. However, another type of correlation structure can also be implemented in quite a similar way. One possible alternative is a correlation structure defined as a function of a certain distance of variants to be tested. Finally, we note that ACST was developed under the condition where the phenotype variable y_i was binary, but the generalization of ACST to the case of multinomial variables y_i is somewhat straightforward. On the other hand, we are often faced with continuous phenotypes as well, and the extension of ACST to such cases will be a valuable future study.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

Research reported in this publication was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1TR001105. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This work was partially supported by Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (Grant numbers: 25280008 and 15K15954).

1 Jonathan C, Robert S, Yves L *et al*: Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 2004; **305**: 869–872.
2 Lee S, Abecasis GR, Boehnke M, Lin X: Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* 2014; **95**: 5–23.
3 Li B, Leal S: Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008; **83**: 311–321.
4 Madsen BE, Browning SR: A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 2009; **5**: e1000384.
5 Morris AP, Zeggini E: An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 2010; **34**: 188–193.
6 Price AL, Kryukov GV, de Bakker PIW *et al*: Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 2010; **86**: 832–838.
7 Neale BM, Rivas MA, Voight BF *et al*: Testing for an unusual distribution of rare variants. *PLoS Genet* 2011; **7**: e1001322.
8 Wu MC, 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* 2011; **89**: 82–93.
9 Lee S, Wu MC, Lin X: Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 2012; **13**: 762–775.
10 Turkmen AS, Yan Z, Hu YQ, Lin S: Kullback-Leibler distance methods for detecting disease association with rare variants from sequencing data. *Ann Hum Genet* 2015; **79**: 199–208.
11 Day NE, Byar DP: Testing hypotheses in case-control studies—equivalence of Mantel-Haenszel statistics and logit score tests. *Biometrics* 1979; **35**: 623–630.
12 Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S: Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Comput Biol* 2016; **12**: 1–20.
13 Bakshi A, Zhu Z, Vinkhuyzen AAE *et al*: Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. *Sci Rep* 2016; **6**: 32894.
14 Nathan Mantel and William Haenszel: Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959; **22**: 719–748.
15 Rothman KJ, Greenland S, Associate TLL: *Modern Epidemiology,* 3rd edn. Lippincott Williams & Wilkins: Pennsylvania, 2012.
16 Victor RG, Haley RW, Willett DL *et al*: The Dallas Heart Study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. *Am J Cardiol* 2004; **93**: 1473–1480.
17 Zeng L, Dai J, Ying K *et al*: Identification of a novel human angiopoietin-like gene expressed mainly in heart. *J Hum Genet* 2003; **48**: 159–162.
18 Romeo S, Yin W, Kozlitina J *et al*: Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *J Clin Invest* 2009; **119**: 70–79.
19 Chen G, Yuan A, Zhou Y *et al*: Simultaneous analysis of common and rare variants in complex traits: application to SNPs (SCARVAsnp). *Bioinform Biol Insights* 2012; **6**: 177–185.
20 Mattijssen F, Kersten S: Regulation of triglyceride metabolism by angiopoietin-like proteins. *Biochim Biophys Acta* 2012; **1821**: 782–789.
21 Johnson NL, Kotz S, Balakrishnan N: *Discrete Multivariate Distributions.* Wiley: New York, 1997.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)

APPENDIX

Appendix Derivation of the Conditional Score Statistic

We here provide the derivation of the conditional score statistic s_k. To begin with, the partial derivative of the conditional log-likelihood is given by

$$\begin{aligned}\frac{\partial \log \mathrm{CL}(\beta_k)}{\partial \beta_k}\Big|_{\beta_k=0} &= \sum_{i:y_i=1} g_{ik} - |D(n_1)|^{-1}\sum_{D(n_1)}\sum_{i\in D(n_1)} g_{ik}\\ &= 2a_{1k}+b_{1k} - |D(n_1)|^{-1}\sum_{D(n_1)}\sum_{i\in D(n_1)} g_{ik}\\ &= 2a_{1k}+b_{1k} - E(2a_{1k}+b_{1k})\end{aligned}$$

where $|D(n_1)|$ denotes the number of all possible subsets $D(n_1)$. Under the null hypothesis that $\beta_k=0$, the vector (a_{1k},b_{1k},c_{1k}) follows the trivariate hypergeometric distribution[21] since the summation of (a_{1k},b_{1k},c_{1k}) is n_1 and these vector components are not greater than m_{1k}, m_{2k}, m_{3k}, respectively. It is known that the expectations and variances are given by

$$E(a_{1k}) = \tfrac{n_1 m_{1k}}{N}, E(b_{1k}) = \tfrac{n_1 m_{2k}}{N}, \mathrm{Cov}(a_{1k}, b_{1k}) = -\tfrac{m_{1k}m_{2k}n_1n_2}{N^2(N-1)},$$
$$\mathrm{Var}(a_{1k}) = \tfrac{m_{1k}n_1n_2(m_{2k}+m_{3k})}{N^2(N-1)}, \mathrm{Var}(b_{1k}) = \tfrac{m_{2k}n_1n_2(m_{1k}+m_{3k})}{N^2(N-1)}$$

Then, using the above results, the expectation $E[2a_{1k}+b_{1k}]$ is calculated as $E[2a_{1k}+b_{1k}]=N^{-1}n_1(2m_{1k}+m_{2k})$. The information of $\mathrm{CL}_k(\beta_k)$ evaluated at $\beta_k=0$ is given by

$$\begin{aligned}-\frac{\partial^2 \log \mathrm{CL}_k(\beta_k)}{\partial \beta_k^2}\Big|_{\beta_k=0} &= |D(n_1)|^{-1}\sum_{D(n_1)}\left(\sum_{i\in D(n_1)} g_{ik}\right)^2\\ &\quad - |D(n_1)|^{-2}\left(\sum_{D(n_1)}\sum_{i\in D(n_1)} g_{ik}\right)^2\\ &= E\left[(2a_{1k}+b_{1k})^2\right] - (E[2a_{1k}+b_{1k}])^2\\ &= \mathrm{Var}(2a_{1k}+b_{1k})\end{aligned}$$

Hence, we get

$$\begin{aligned}\mathrm{Var}(2a_{1k}+b_{1k}) &= 4\mathrm{Var}(a_{1k}) + 4\mathrm{Var}(b_{1k}) + 4\mathrm{Cov}(a_{1k}, b_{1k})\\ &= \frac{n_1n_2}{N^2(N-1)}(4m_{1k}m_{3k} + m_{1k}m_{2k} + m_{2k}m_{3k})\end{aligned}$$

whereby we obtain the conditional score statistic s_k.