

## ORIGINAL ARTICLE

# Block-based association tests for rare variants using Kullback–Leibler divergence

Degang Zhu<sup>1,2</sup>, Yue-Qing Hu<sup>3</sup> and Shili Lin<sup>4</sup>

Although genome-wide association studies have successfully detected numerous associations between common variants and complex diseases, these variants typically can only explain a small part of the heritable component of a disease. With the advent of next-generation sequencing, attention has turned to rare variants. Recently, a variety of approaches for detecting associations of rare variants have been proposed, including the Kullback–Leibler divergence-based tests (KLTs) for detecting genotypic differences between cases and controls. However, few of these approaches consider linkage disequilibrium (LD) structure among rare variants and common variants. In this study, we propose two block-based association tests for testing the effects of rare variants on a disease. The main idea for this approach comes from the hypothesis that a region of interest may consist of two or more LD blocks such that single-nucleotide variants (SNVs) within each block are correlated, whereas SNVs in different blocks are independent or weakly correlated. Under this hypothesis, we propose two tests that are generalizations of the KLTs by taking the block structure into account. A simulation study under various scenarios shows that the proposed methods have well-controlled type I error rates and outperform some leading methods in the literature. Moreover, application to the Dallas Heart Study data demonstrates the feasibility and performance of the two proposed methods in a realistic setting.

*Journal of Human Genetics* (2016) 61, 965–975; doi:10.1038/jhg.2016.90; published online 14 July 2016

## INTRODUCTION

Although extensive genome-wide association studies have resulted in the detection of many common variants (CVs) that are associated with complex traits or diseases, these variants tend to have modest effect on the phenotype, whereas rare variants (RVs) are likely to have stronger effects.<sup>1–3</sup> At the same time, the new sequencing technologies are providing an avenue for re-sequencing parts of, or even the entire, genome, thus leading to the effective detection of RVs. There is growing evidence supporting the role of RVs in complex trait associations. For example, four disease-associated RVs in the *IFIH1* gene, which had been proved to be protective of type 1 diabetes, were detected.<sup>4</sup> Four variants with a minor allele frequency (MAF) in the range of 0.1–0.8% in the *NOD2* gene were also found to be associated with Crohn's disease.<sup>5</sup> Also, 13 functionally screened MTNR1B variants with MAF < 0.1% were identified to be associated with type 2 diabetes.<sup>6,7</sup> However, statistical approaches for genome-wide association studies based on testing individual single-nucleotide variants (SNVs) do not work well because the power for detecting an association with a RV is low even with a very large sample.<sup>8,9</sup> Therefore, a larger number of new statistical data analysis strategies specifically targeting RVs have been proposed. Some newly proposed tests share the common idea of pooling or collapsing multiple rare SNVs and then testing for an association with some trait by combining information across multiple sites.<sup>9–13</sup> These so-called burden tests

perform well when there are no or few neutral RVs in the region of interest and most of the causal RVs have the same association direction on the trait. For ease of discussion, we refer to variants that are associated with the disease of interest as 'causal', but such variants may not be really causal but rather, merely associated. However, the effects of RVs are not always in the same direction. If the RVs to be pooled consist of both positively and negatively associated variants, the association signal may be weakened or canceled out, which may result in low power.<sup>14</sup> Tests based on model selection have then been proposed to address this.<sup>15–18</sup> The main idea of these tests is to determine whether a RV may be associated and should be collapsed, and if so, its association direction is also ascertained. As already pointed out by researchers, in spite of their strong motivation for model selection, the performance of model selection-based tests might not be as impressive as expected, especially when there are a large number of neutral RVs in the region of interest.<sup>14</sup>

Another alternative to overcoming the problem of different association directions is to use non-burden tests, including variance component tests such as the C-alpha test,<sup>19</sup> the sequence kernel association test (SKAT)<sup>20</sup> and some other related methods such as the sum of squared score test (SSU)<sup>21</sup> and the Goeman's score test.<sup>22</sup> Non-burden tests are shown to be more robust to the inclusion of causal variants with opposite association directions. However, non-burden tests have their own disadvantages. They can be less powerful

<sup>1</sup>Department of Applied Mathematics, Nanjing Forestry University, Nanjing, China; <sup>2</sup>School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China; <sup>3</sup>Institute of Biostatistics, School of Life Sciences, Fudan University, Shanghai, China and <sup>4</sup>Department of Statistics, The Ohio State University, Columbus, OH, USA

Correspondence: Dr S Lin, Department of Statistics, The Ohio State University, 1958 Neil Avenue, 404 Cokins Hall, Columbus, OH 43210–1247, USA.  
E-mail: shili@stat.osu.edu

Received 17 December 2015; revised 3 May 2016; accepted 17 June 2016; published online 14 July 2016

than burden tests if most of the causal variants have the same association direction. As such, an optimal test, SKAT-O, which combines SKAT and a burden test, was proposed and shown to perform well in a wide range of scenarios.<sup>23</sup> Nevertheless, SKAT-O has its own disadvantage in that it may lose power when the proportion of non-associated RVs in the testing region is large.<sup>24</sup> Recently, several novel methods based on Kullback–Leibler divergence for testing the effects of RVs on a trait were proposed.<sup>25</sup> These tests, referred to as Kullback–Leibler tests (KLTs), have been shown to perform better than SSU and SKAT-O through extensive simulations.

Linkage disequilibrium (LD) is known to exist among CVs, which was relied on as the fundamental principle for detecting markers that are associated with common diseases<sup>21,26,27</sup>. However, despite the increased focus on ascertaining the role of RVs in complex diseases, there has been little study on LD patterns involving RVs. The situation for RVs is clearly different from that for CVs due to low minor allele counts; therefore, it has been suggested that RVs are likely to be independent in general.<sup>9</sup> As such, many RV association methods assume that RVs are independent, either implicitly or explicitly. However, based on a study utilizing the 1000 Genomes data, Feng and Zhu<sup>28</sup> concluded that substantial LD among RVs exist, which may be explained by population admixture. Unless such LD is being appropriately accounted for, large-scale false positives may result. However, the authors also concluded that traditional family-based transmission disequilibrium test may not be able to overcome the problem if multiple RVs are analyzed together. To address this problem, a handful of methods have been proposed. Talluri and Shete<sup>29</sup> proposed a step-wise variant-selection procedure (LDSEL) that takes LD into account. An alternative variant selection procedure (CCRS) was proposed in Yazdani *et al.*,<sup>30</sup> in which LD was incorporated implicitly by selecting principal components that are most related to the phenotype being studied. On the other hand, Turkmen and Lin<sup>31</sup> proposed two family-based block approaches (rbPDT and rFBAT) that preserve the LD structure within each block while treating variants between blocks to be independent to increase statistical power without inflating the type I error.

Inspired by the block-based approach for accounting for LD and the nice properties of KLTs, in this paper, we propose two block-based association tests for RVs, which generalize the KLTs and improve the power when the LD block in a region of interest is known. These methods share the same feature as the KLTs in that they are not sensitive to different effect sizes or directions, but they can be more powerful than the original KLTs when there are indeed blocks in the test region that separate the causal variants from the majority of the non-causal ones. In addition to comparing with the original KLT, we also compare the proposed tests with SSU and SKAT-O, with the latter arguably being the most popular method to date. Note that SSU was originally proposed for CVs and takes LD into consideration. As burden tests have been shown to be less powerful than SKAT-O and SSU, they are omitted in this study. Among the handful of LD-incorporated approaches that were specifically proposed for detecting RV associations, rbPDT and rFBAT are for family data only, whereas there are no publicly available softwares for LDSEL or CCRS. Therefore, none of these methods were used in our study. In addition to the simulation, we also applied the proposed approaches and the comparison methods to the Dallas Heart Study (DHS) data to further evaluate their performances with real data.

## MATERIALS AND METHODS

We first give a brief introduction to the KLTs.<sup>25</sup> The main idea of the KLTs is to directly compare the distributional differences of variant frequencies site

by site between cases and controls. Assume that there are  $M$  genetic variant sites within the region of interest where both RVs and CVs may be present. Two normalized variant frequencies at site  $m$  among all sites in the cases ( $n_1$  individuals) and controls ( $n_2$  individuals) are defined as  $f_m$  and  $g_m$  respectively,  $m = 1, 2, \dots, M$ . Then,  $\{f_m\} \equiv \{f_m, m = 1, 2, \dots, M\}$  and  $\{g_m\} \equiv \{g_m, m = 1, 2, \dots, M\}$  become two discrete distributions defined on the same region. We focus on one of the KLT test statistics, which is defined based on the KL divergence<sup>32</sup> as follows:

$$\text{KLT} = H(\{f_m\}, \{g_m\}) = \frac{1}{2} \left[ \sum_{m=1}^M (f_m - g_m) \log \frac{f_m}{g_m} \right]. \quad (1)$$

Obviously, the KLT statistic is 0 when the two distributions are identical and has a large (positive) observed value when one distribution is different from the other. The  $P$ -value of the KLT statistic is calculated by a permutation procedure where the case or control status of individuals is permuted. To improve the power of the KLT when there are a large number of neutral variants, they also proposed a data-adaptive screening step to distill the variants to obtain an adaptive KLTs,<sup>25</sup> which shows improvement over the KLT in some cases. Two other KLTs were also proposed. In this study, however, we only discuss the KLT as defined in (1) to be sufficiently focused, although the block versions of the other KLTs can be similarly devised.

Although KLT was shown to perform well compared with other methods, there is room for improvement in settings in which variants being investigated are in multiple blocks with disparity in signals. To address this limitation of the KLT test, we propose two block-based association tests, both of which generalize the KLT and can potentially improve the power without sacrificing type I error rate when the block structure of variants within a region being investigated is known.

### bKLTmax

The first test that we propose, bKLTmax, is a block-based KLT that attempts to find maximum information contained within a block. For ease of exposition of idea, we assume that there are two blocks, block 1 (B1) and block 2 (B2), in the region to be considered, although the method is applicable to a setting with a larger number of blocks. In each block, both causal and neutral variants may be present, and at the same time, both CVs and RVs may also be included. By LD blocks, we mean that variants within the same blocks are correlated, whereas variants between blocks are independent or only weakly correlated. In the setting where all causal variants cluster in one block, whereas the other block only contains neutral variants, the original KLT is expected to lose power due to the noise. Some numerical results shown in Figure 2 of Turkmen *et al.*<sup>25</sup> can be regarded as an illustration of this. Motivated by this shortcoming, we propose to apply the KLT test to each of the two blocks separately to increase the information contained in the block with causal variants. Specifically, we define

$$\text{bKLTmax} = \max\{\text{KLT}_1, \text{KLT}_2\}, \quad (2)$$

where  $\text{KLT}_1$  and  $\text{KLT}_2$  are the KLT statistics for B1 and B2, respectively. As the distribution of the test statistic bKLTmax under the null hypothesis of no association is not of a known form, we calculate the  $P$ -value by a permutation procedure as described in section ‘Significance Test’. The bKLTmax statistic may be defined analogously when there are more than two blocks.

### bKLT

As we can see from the definition, if causal variants are split fairly evenly among the blocks, then bKLTmax will lose power as the signals are being fragmented. Hence, we propose a second statistic, bKLT, which is an attempt to take advantage of both the robust feature of the original KLT and the greater sensitivity of bKLTmax. This is a compromise that is most appropriate when it is unknown, before hand, whether causal variants are contained in only one, or split among multiple blocks. Specifically, we define

$$\text{bKLT} = \max\{\text{KLT}_1 + \text{KLT}_2, \text{KLT}\}, \quad (3)$$

where  $\text{KLT}_1 + \text{KLT}_2$  is the sum of the numerical values of the two KLT statistics as defined in (2), and KLT is the statistic for the entire region being investigated. When there are more than two blocks, the bKLT statistic may

be defined similarly.  $P$ -value of the test statistic will also be obtained by permutation as we discuss in the following.

**Significance test**

As the distribution of either of the statistics being proposed, bKLTmax or bKLT, is not of a known form under the null hypothesis of no association, we obtain the null distribution by means of permuted samples to estimate the  $P$ -value. Specifically, we permute the case-control status of the individuals, while keeping the genotype information fixed. Without the loss of generality, let the statistic be  $T$  (which may be bKLTmax or bKLT). We first randomly assign  $n_1$  of the subjects to be cases, whereas the remaining are treated as controls. Then, we apply the test statistic to the permuted data to get the corresponding test statistic  $T^{(b)}$ . This process is repeated  $B$  times, that is,  $b = 1, 2, \dots, B$ . The  $P$ -value for the statistic  $T$  is estimated as

$$P = \frac{\sum_{b=1}^B I(T^{(b)} \geq T)}{B}, \tag{4}$$

where  $I(\cdot)$  is the usual indicator function. For the simulation study described in the following section, we generate data under a model and repeat the above procedure to obtain a  $P$ -value  $P^{(r)}$ . This process is then repeated  $R$  times, that is,  $r = 1, 2, \dots, R$ . Then, for a given significance level  $\alpha$ , we compute

$$Q = \sum_{r=1}^R I(P^{(r)} \leq \alpha) / R. \tag{5}$$

This quantity  $Q$  can be interpreted as type I error rate, if the model portrays no causal variants. On the other hand, if there are variants in the region of interest that are associated with the disease, then  $Q$  is reported as power. The codes implementing bKLT and bKLTmax are available upon request.

**RESULTS**

**Simulation study**

We first compare the two proposed methods, bKLTmax and bKLT, with three existing methods, KLT, SSU and SKAT-O, in a comprehensive simulation study. We consider both independent and correlated genetic variants.

**Data generation**

Following the paper by Turkmen *et al.*,<sup>25</sup> we generated the data under various causal mechanisms and MAFs for the causal and neutral variants, as summarized in Table 1. As can be seen from Table 1, there are a total of 32 SNVs in the region to be tested, among which 8 SNVs are causal and the others are neutral. We considered six scenarios. In scenario 1, the eight causal SNVs are all CVs, whereas in scenario 2, the eight causal SNVs are all RVs. In scenario 3, the 8 causal SNVs are all CVs, but the 24 neutral variants consist of 8 rare neutral variants and 16 common neutral variants. In scenarios 4–6, the make-up of the eight causal SNVs is two rare causal variants and six common causal

variants, but at the same time, we considered several different MAF ranges for the neutral variants.

The genotype matrix is simulated as follows. First, we simulated  $M = 32$  variants with the sample size of 500 cases and 500 controls. Each variant has a MAF uniformly distributed in the intervals displayed in Table 1. Following the paper by Basu and Pan,<sup>14</sup> we generated a latent vector  $\mathbf{Z} = (Z_1, Z_2, \dots, Z_M)'$  from a multivariate normal distribution with mean  $\mathbf{0}_M = (0, 0, \dots, 0)'$  and variance  $\mathbf{1}_M = (1, 1, \dots, 1)'$ . To take the correlation between any two variants into account, we assume, within a block, that causal variants or neutral variants are correlated with themselves by a first-order auto-regressive (AR(1)) structure, but there is no correlation between these two types of causal or neutral variants. That is, there was a correlation  $\text{Corr}(Z_k, Z_j) = \rho^{|k-j|}$  between any two causal variants or any two neutral variants within the same block. The correlation coefficient  $\rho$  was set to be 0, 0.3, 0.5 and 0.7 to denote independent and various degrees of correlation. Then, each  $Z_i$  is mapped to a value between 0 and 1 through inverse transformation and then dichotomized to 1 (minor allele) or 0 (major allele) depending on the corresponding MAF of the variant. For each individual, combining two  $\mathbf{Z}$ 's lead to the vector of genotype data  $\mathbf{X}$ .

The above genotype data simulation was carried out according to three different block structure specifications, as given in Table 2. Block structure 1 specifies that all the eight causal variants are clustered in block 1. Block structure 2 depicts the situation in which six and two causal variants are contained in blocks 1 and 2, respectively. For block structure 3, both block 1 and block 2 harbor four causal variants. These three block structures were devised, in various degrees of difficulty, to test the performance of the methods. The effect sizes of the variants, collectively denoted as vector  $\boldsymbol{\beta}$ , are provided in the footnotes of the table for easy reference. Specifically, for block structure 1, the eight effect sizes in  $\boldsymbol{\beta}$  are for the eight causal variants in block 1 under all scenarios. For block structures 2 and 3, causal variants are contained in both blocks. As such, the effect sizes specified in  $\boldsymbol{\beta}$ , from left to right, correspond to the order of causal variants described in each scenario. For example, for block structure 2 under scenario 1, the first six effect sizes in  $\boldsymbol{\beta}$  are for the six common causal variants in block 1, whereas the remaining two effect sizes are for the two common causal variants in block 2.

With the specification of  $\boldsymbol{\beta}$  for each block structure and each scenario in place, we then simulated the disease status  $Y_i$  of the  $i$ th individual using the following logistic model

$$\text{logit}P(Y_i = 1) = \beta_0 + \mathbf{X}_i' \boldsymbol{\beta}, \tag{6}$$

where  $\beta_0$ , the background disease prevalence, was set to be  $\log(1/4)$ ;  $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{iM})'$  denotes the genotype vector of the  $i$ th individual ( $X_{ij} = 0, 1, 2$ ) over  $M (= 32)$  markers. In addition to

**Table 1 Summary of the six scenarios utilized in the simulation study**

Scenario	RC		RN		CC		CN	
	MR	# SNVs	MR	# SNVs	MR	# SNVs	MR	# SNVs
1	NA	0	NA	0	0.1–0.3	8	0.2–0.5	24
2	0.005–0.01	8	0.01–0.05	24	NA	0	NA	0
3	NA	0	0.005–0.01	8	0.1–0.3	8	0.2–0.5	16
4	0.005–0.01	2	0.01–0.05	8	0.1–0.3	6	0.1–0.3	16
5	0.005–0.01	2	0.005–0.01	8	0.1–0.3	6	0.2–0.5	16
6	0.005–0.01	2	0.01–0.05	8	0.1–0.3	6	0.2–0.5	16

Abbreviations: CC, common causal; CN, common neutral; MAF, minor allele frequency; NA, not applicable; MR, MAF range; RC, rare causal; RN, rare neutral; SNV, single-nucleotide variant. For each variant, the MAF value was generated uniformly from the range given.

**Table 2** Distribution of variants and their association effect sizes under three block structures for each of the six scenarios given in Table 1

Scenario	Block structure	Block 1				Block 2			
		RC	RN	CC	CN	RC	RN	CC	CN
1 <sup>a</sup>	1	0	0	8	8	0	0	0	16
	2	0	0	6	8	0	0	2	16
	3	0	0	4	8	0	0	4	16
2 <sup>b</sup>	1	8	8	0	0	0	16	0	0
	2	6	8	0	0	2	16	0	0
	3	4	8	0	0	4	16	0	0
3 <sup>a</sup>	1	0	8	8	0	0	0	0	16
	2	0	8	6	0	0	0	2	16
	3	0	8	4	0	0	0	4	16
4 <sup>c</sup>	1	2	8	6	0	0	0	0	16
	2	2	8	4	0	0	0	2	16
	3	2	8	2	0	0	0	4	16
5 <sup>c</sup>	1	2	8	6	0	0	0	0	16
	2	2	8	4	0	0	0	2	16
	3	2	8	2	0	0	0	4	16
6 <sup>c</sup>	1	2	8	6	0	0	0	0	16
	2	2	8	4	0	0	0	2	16
	3	2	8	2	0	0	0	4	16

Abbreviations: CC, common causal; CN, common neutral; MAF, minor allele frequency; MR, MAF range; RC, rare causal; RN, rare neutral.

<sup>a</sup> $\beta = (\log(3/2), \log(2/3), \log(23/20), \log(23/20), \log(23/20), \log(20/23), \log(20/23), \log(20/23))$ .

<sup>b</sup> $\beta = (\log(3), \log(1/3), \log(2), \log(2), \log(2), \log(1/2), \log(1/2), \log(1/2))$ .

<sup>c</sup> $\beta = (\log(3), \log(1/3), \log(23/20), \log(23/20), \log(23/20), \log(20/23), \log(20/23), \log(20/23))$ .

evaluating the power of the various tests using the effect size  $\beta$  provided in Table 2, we also set  $\beta = 0$  to gauge the type I error. We used 1000 replicates to evaluate type I error and power at a significance level of 0.05. For each replicate, we permuted the case-control status of the individuals 1000 times.

### Type I error rate

Our first series of the simulation study was to compare the two proposed tests, bKLTmax and bKLT with KLT, SSU and SKAT-O for all six scenarios and three block structures under the null model. Within each block, we considered four levels of correlation:  $\rho = 0, 0.3, 0.5, 0.7$ . The results are displayed in Table 3. As we can see, the type I error rates are all around the nominal level of 0.05. Thus, all the five tests appear to have properly controlled type I error rates.

### Power

Our second series of the simulation study was to compare the powers of bKLTmax, bKLT, KLT, SSU and SKAT-O. First, we considered block structure 1. In this case, all eight causal variants cluster in one block (block 1), whereas block 2 only consists of neutral variants, that is, there is no association signal in block 2. The power results under the six scenarios given in Table 2 with four different strengths of correlation are illustrated in Figure 1. As we can see from the figure, bKLTmax is more powerful than the other four tests in all the six scenarios considered. Specifically, in scenarios 2, 4, 5 and 6, when there are only rare causal variants or when there are both rare and common causal variants, bKLTmax and bKLT perform well compared with the other three tests. In scenarios 1 and 3, when only common causal variants are involved, bKLTmax and bKLT still have better performance than KLT, SSU and SKAT-O, especially with higher

correlation coefficient ( $\rho = 0.7$ ). The power results for block structure 1 are consistent with what we would expect; bKLTmax is suitable for settings in which all causal variants are within a block as it maximizes the information content and reduces the influence of noise.

For block structure 2, the results of the five tests are showed in Figure 2. In this case, six of the eight causal variants cluster in block 1 and the remaining two are located in block 2. bKLTmax and bKLT still perform well, as with block structure 1. Specifically, in scenarios 1 and 2, bKLTmax and bKLT are more powerful than KLT, SSU and SKAT-O. In scenarios 3–6, bKLTmax, bKLT and KLT all have similar power, although with a higher correlation coefficient ( $\rho = 0.5, 0.7$ ), bKLTmax and bKLT are less powerful than KLT in scenarios 5 and 6, but they are still much more powerful than SSU and SKAT-O.

For block structure 3, half of the causal variants are in block 1 and the other half are in block 2, which represents the worse case scenario for bKLTmax. The results are displayed in Figure 3. Surprisingly, we still see that bKLTmax and bKLT are the most powerful tests for scenario 1. Under scenario 2, bKLT has the highest power among the tests. In scenarios 3–6, KLT performs the best, which is followed by bKLT, whereas bKLTmax still outperforms SSU and SKAT-O. These results are once again as expected. As the signals for the association are split between two blocks, bKLT, proposed as a compromise between bKLTmax and KLT, has better power than bKLTmax.

### Robust analysis

In the simulation study presented thus far, bKLTmax and bKLT are shown to work well in various settings based on the assumption that we know the true block structure of the region of interest. Notwithstanding the superior performance, the question arises as to how will bKLTmax and bKLT behave under the perturbation of block structure

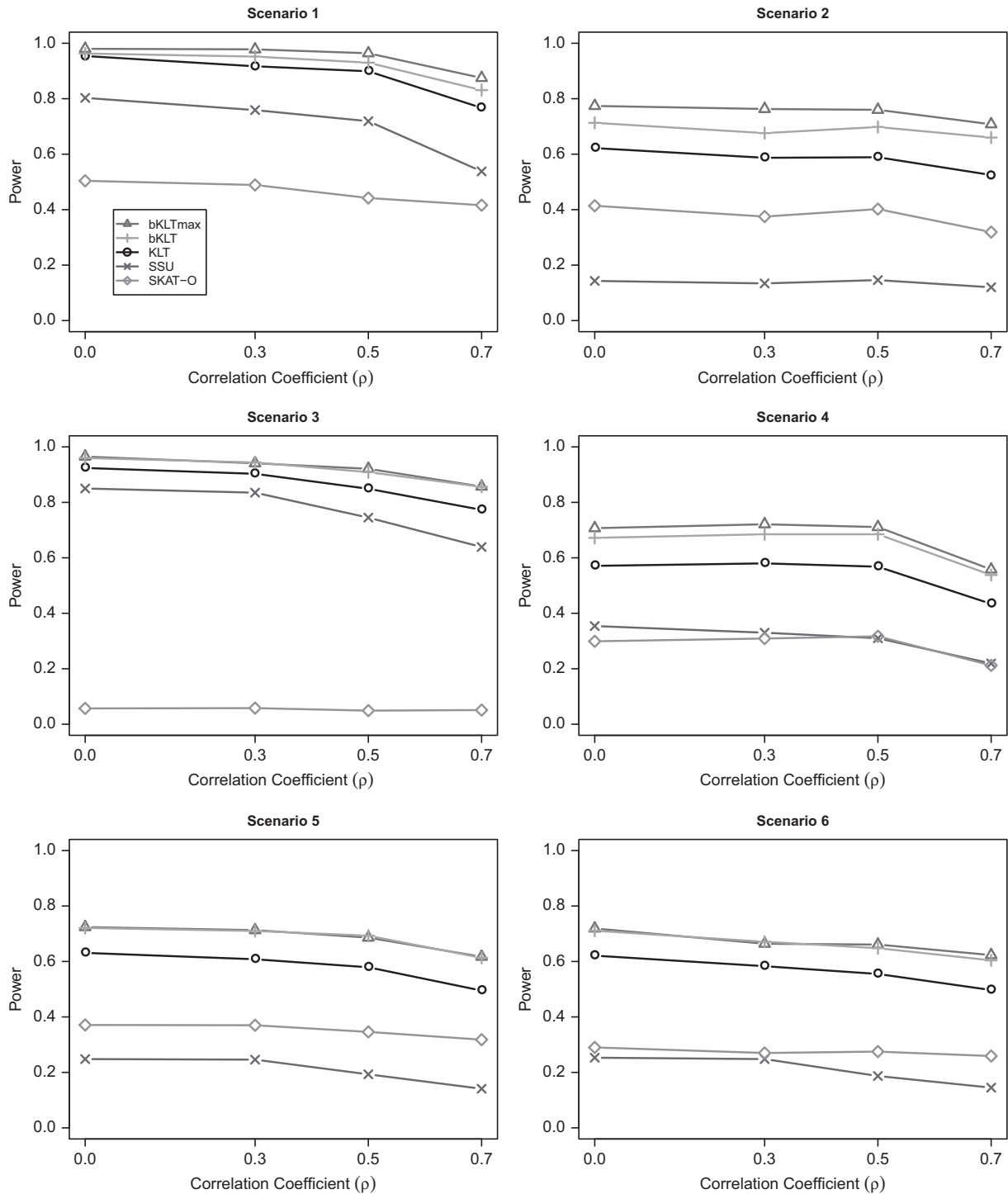
**Table 3** Empirical type I error rates (%) at the nominal significance level  $\alpha = 5\%$  based on 1000 replicates for five tests, six scenarios, three block structures and four levels of correlation

Scenario	Method	Block structure 1				Block structure 2				Block structure 3			
		$\rho$				$\rho$				$\rho$			
		0	0.3	0.5	0.7	0	0.3	0.5	0.7	0	0.3	0.5	0.7
1	bKLTmax	4.4	5.5	4.5	5.4	5.1	4.6	5.2	4.7	4.6	4.9	4.3	6
	bKLT	3.3	5.8	5.1	4.9	4.5	5.1	5.8	4.8	5	5.4	4	5.7
	KLT	3.4	5.6	4.9	4.9	5.1	5.9	5.2	4.6	5	5.4	5.1	6.1
	SSU	3.4	5.3	5.1	4.5	4.2	5.7	3.9	4.2	4.8	5.1	4.9	6.2
	SKAT-O	5.2	4.5	5.6	4.7	4.6	4.4	4.7	6	4.9	5	5.7	4
2	bKLTmax	5.3	6	5	4.3	4.6	6.1	5.3	5.4	5.8	6	5.3	4.9
	bKLT	4.9	5.9	4.9	4.7	5.4	5.6	3.9	5.1	5.2	5.7	5.8	5.5
	KLT	4.5	5.7	4	5.7	5.9	4.2	4.2	4.9	5.6	5.6	5.6	5.2
	SSU	5.3	5.1	3.7	5.6	5.6	4.1	5	4.6	3.7	6	4.8	4.8
	SKAT-O	5.6	5.8	4.3	5.9	5.9	3.7	5.5	5.1	4.6	5.7	4.6	5.4
3	bKLTmax	4.6	5.2	5.9	4.7	5.3	5	3.9	5.1	4.6	4.7	4.7	4.8
	bKLT	4.5	5	6	4.3	5	5.2	4.1	5.1	4.6	4.7	4.6	4.3
	KLT	4.7	5.8	4.7	4.8	4.6	5.4	5.7	5.7	5.2	4.2	3.7	5
	SSU	5.1	5.4	3.5	4.7	4.8	4.7	4.7	5.1	4.1	4.6	4.8	4.8
	SKAT-O	5.2	5.7	6	4.5	4.2	5.7	4.6	5.9	5.2	3.3	4.4	5.8
4	bKLTmax	4.2	4.1	4.7	4.7	3.2	4.3	3.1	5.4	5.9	6	5.3	3.9
	bKLT	4.6	4.6	4.8	4.4	3.4	4.3	3.2	5.3	5.9	5.9	5.1	4.2
	KLT	4.8	5.1	4.4	4.4	3.9	4.8	4	5	5.3	6	4.5	4.5
	SSU	5.4	4.2	4.2	5.8	5.2	3.9	4.9	5.9	5.1	4.7	4.4	5.3
	SKAT-O	4.9	5	3.8	4.7	3.7	4.6	3.7	5.6	5	4.6	4.6	4.9
5	bKLTmax	3.3	4.9	4.5	5	5.1	4.5	4.5	5.8	5.6	4.9	4.8	5.1
	bKLT	3.5	4.5	4.3	5	5	4.1	4.1	5.9	5.5	4.8	5	5.2
	KLT	3.6	4.1	4.4	5.6	5.5	4.5	4.9	5.5	6.1	4.6	5.5	5
	SSU	5.7	3.6	4.7	5.1	4.8	5.4	4.4	5.7	3.6	4.8	5.1	4.1
	SKAT-O	5.5	5.2	4.1	5.2	3.9	4.4	3.8	5	4.6	4.4	5.1	4.7
6	bKLTmax	4.4	5.9	5.4	4	5.8	4.7	4.4	5.2	4.8	5.8	5	4.4
	bKLT	4.6	5.8	5.3	3.6	5.8	4.3	4.6	5.1	4.3	6.1	4.8	4.7
	KLT	5.5	5.9	5.8	3.2	4.5	5	5.2	4.8	6.1	6	4.4	6.9
	SSU	4.2	6.4	5	3.1	4.1	4.2	5.1	5.6	4.3	5.4	4.6	5.3
	SKAT-O	4.9	4.6	4.8	4.6	4.4	5.5	5.6	5.3	4.4	6.1	4.8	6

when variants are grouped into blocks incorrectly. We seek to provide answer to this question with additional simulation. Without loss of generality, we considered three settings: block structure 1 in scenario 2, block structure 2 in scenario 4 and block structure 3 in scenario 6. All the true and analysis models in these three settings are summarized in Table 4. The true model of block structure 1 in scenario 2 is that there are 8 rare causal and 8 rare neutral variants in block 1, whereas block 2 contains 16 RVs that are all neutral. However, we may falsely ascertain the LD blocks. For example, we may group only 6 of the rare causal variants together with 8 rare neutral variants into block 1, whereas the other 2 causal variants are being grouped with the 16 neutral variants into block 2. This scenario is what we referred to as ‘analysis model’. The true and analysis models for the other two settings, block structure 2 in scenario 4 and block structure 3 in scenario 6, provided in Table 4, can be interpreted similarly. Specifically, in this study, the data are generated based on the true models, whereas the powers of tests bKLTmax and bKLT are calculated based on the analysis models, from which the robustness

of these two proposed tests can be investigated. Note that as the other three tests are not influenced by the inferred block structure, their results remain the same, but they are reproduced in Figure 4 for ease of comparison. Table 5 reports the type I error rates for bKLTmax and bKLT for the three settings. As can be seen, the empirical type I error rates are all around the nominal level of 0.05, indicating that misspecification of the block structures has minimal effect on the validity of the two proposed tests.

The power results are showed in Figure 4. For block structure 1 in scenario 2, although bKLTmax and bKLT lost some power compared with that using the true model, they were still more powerful than KLT, SSU and SKAT-O. For block structure 2 in scenario 4, as the analysis model only falsely assigned two common neutral variants originally in block 1 to block 2, there was no ‘drift’ of association information. As a result, bKLTmax and bKLT have similar performances as those based on the true model. For block structure 3 in scenario 6, the analysis model made the mistake of assigning two common causal variants in block 2 to block 1, resulting in enriching

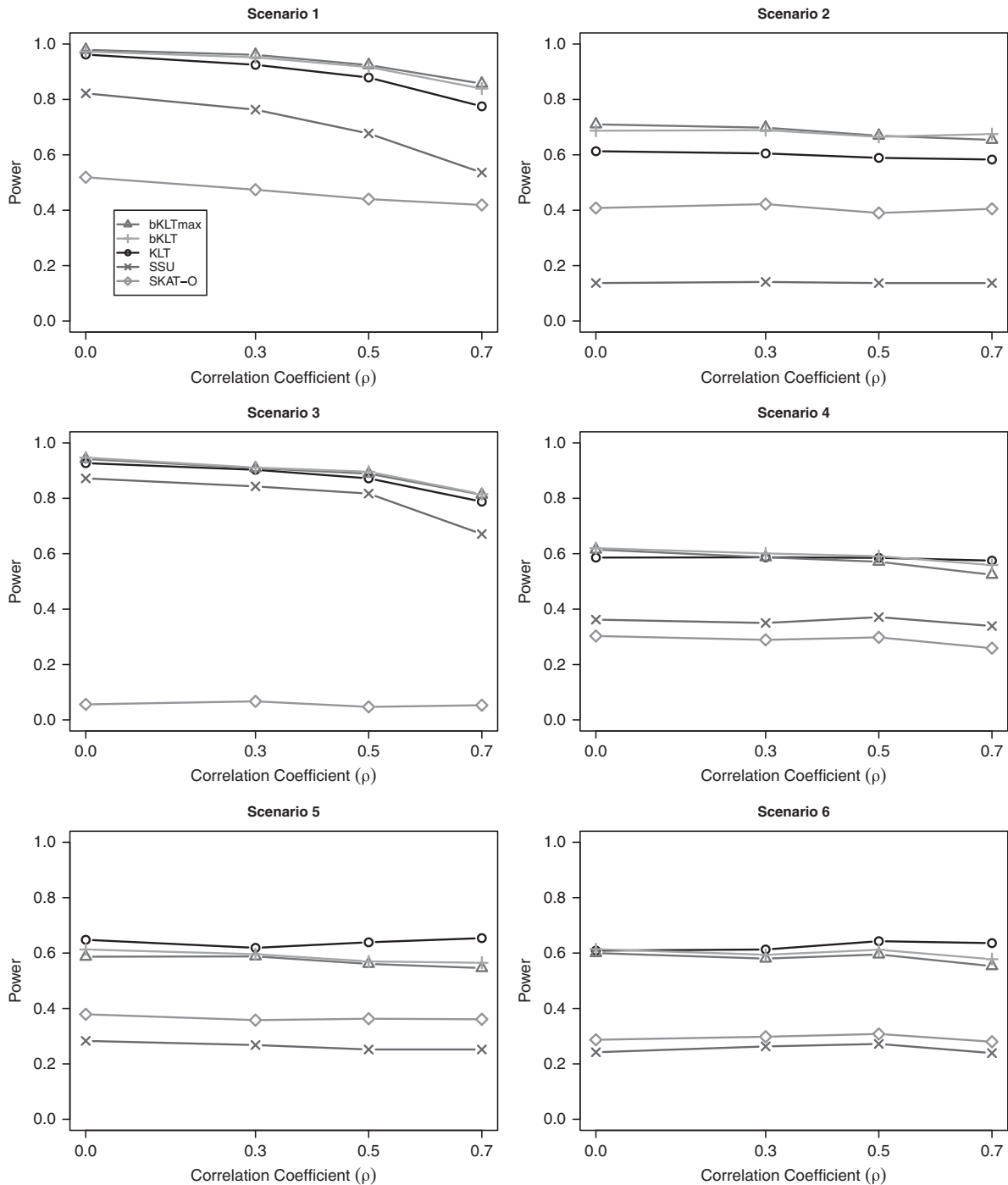


**Figure 1** Power comparisons among bKLTmax, bKLT, KLT, SSU and SKAT-O in block structure 1. Empirical power was calculated at the 5% significance level. The total sample size in each scenario is 1000 (500 cases and 500 controls). A full color version of this figure is available at the *Journal of Human Genetics* journal online.

the association signal in block 1. In this case, bKLTmax and bKLT in fact improved their power, and are seen to have the same power as KLT. Despite the slight power loss in the first two settings, bKLTmax and bKLT still outperform SSU and SKAT-O. In summary, results from this study indicate that power may be slightly influenced by the incorrect specification of blocks while type I errors are well maintained.

#### Application to data from DHS

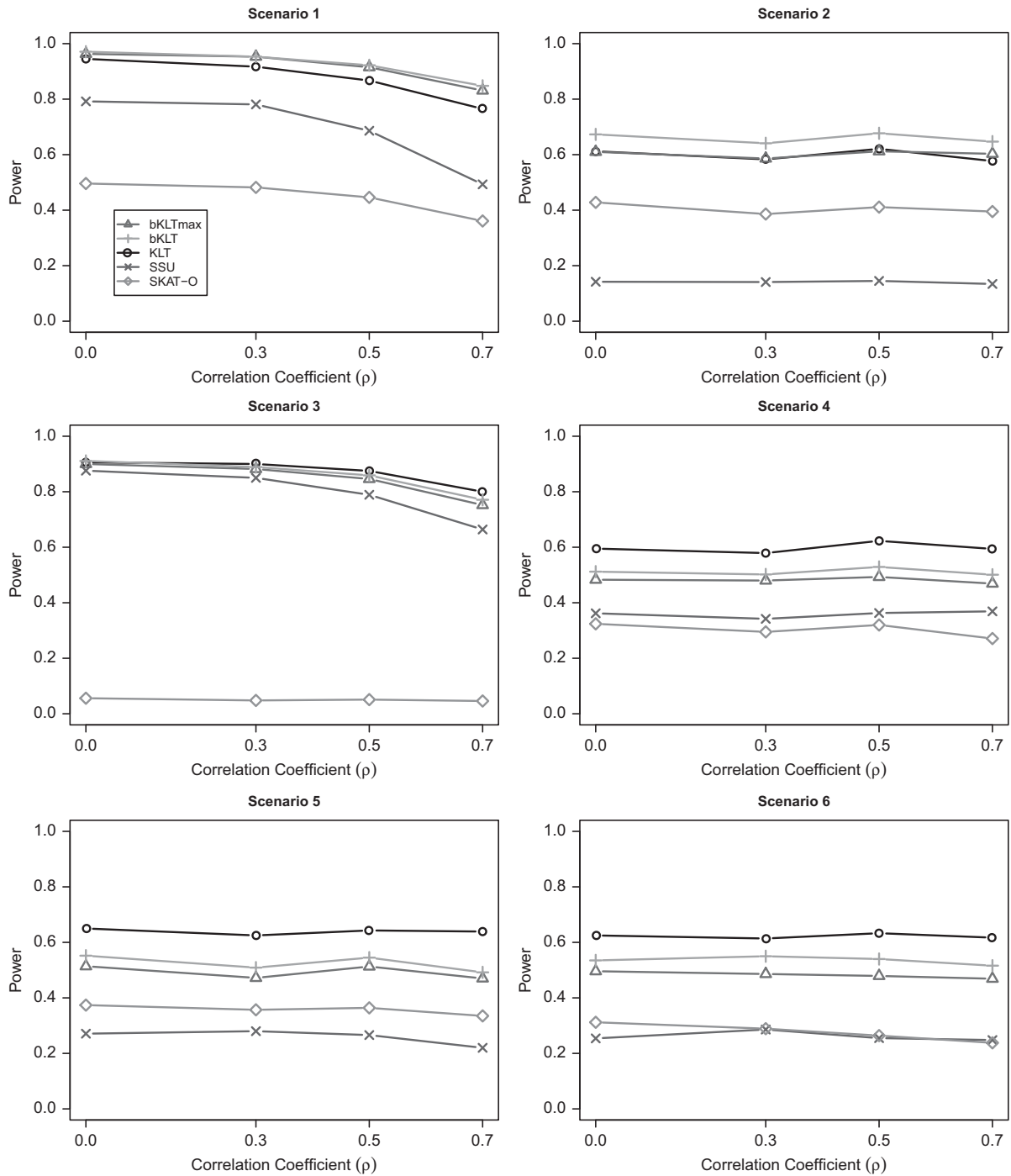
We applied the two methods proposed in this study, bKLTmax and bKLT, together with KLT, SSU and SKAT-O, to the sequence data from the DHS.<sup>33,34</sup> We focused on testing for association between serum triglyceride level and gene *ANGPTL5*. Although the data were also available for two other genes, we did not analyze them here as the association signals are extremely strong and hence are not good



**Figure 2** Power comparisons among bKLTmax, bKLT, KLT, SSU and SKAT-O in block structure 2. Empirical power was calculated at the 5% significance level. The total sample size in each scenario is 1000 (500 cases and 500 controls). A full color version of this figure is available at the *Journal of Human Genetics* journal online.

examples for comparing methods. Following the paper by Epstein *et al.*,<sup>35</sup> individuals with the top 20% of triglyceride values are treated as cases, whereas the bottom 20% are designated as controls, which results in a binary trait with 628 cases and 621 controls. After deleting variants that have no sequence variation (all homozygous for the common allele) in all cases and control samples, 15 SNVs in ANGPTL5 are left. As the bKLTmax and bKLT are block-based tests, we need to learn about the block structure of the gene if blocks do exist. Traditionally, haplotype blocks are learned through studying the

correlation structure of variants based on well-known software such as Haploview.<sup>36</sup> However, such software tools are not useful when the majority of the variants involved are rare ones, as RVs can be negatively correlated but are treated as uncorrelated based on traditional LD measures.<sup>37</sup> As such, we adopted a different strategy for finding blocks in the spirit of bKLTmax. Briefly, we first searched the region for a group of consecutive SNVs that shows the maximum association signal, that is, produces the largest KLT statistic as defined in (1). Specifically, our search resulted in the group composed of



**Figure 3** Power comparisons among bKLTmax, bKLT, KLT, SSU and SKAT-O in block structure 3. Empirical power was calculated at the 5% significance level. The total sample size in each scenario is 1000 (500 cases and 500 controls). A full color version of this figure is available at the *Journal of Human Genetics* journal online.

22604\_R269G, 22623\_L275X, 25956\_D293H, corresponding to SNVs 8, 9 and 10 among the 15 SNVs listed linearly according to their genomic locations. We then considered the correlation between SNV7 (22602\_T268M), the only variant whose MAF > 0.1, with the rest of the 14 variants. It turns out that the correlation between SNV7 and SNV1 (11727\_L98P) is positive (0.035), and much larger than that between SNV7 and any of the other variants. As such, we divided the region into three blocks: SNV1-SNV7, SNV8-SNV10 and SNV11-SNV15, for the analysis by bKLTmax and bKLT. Note that

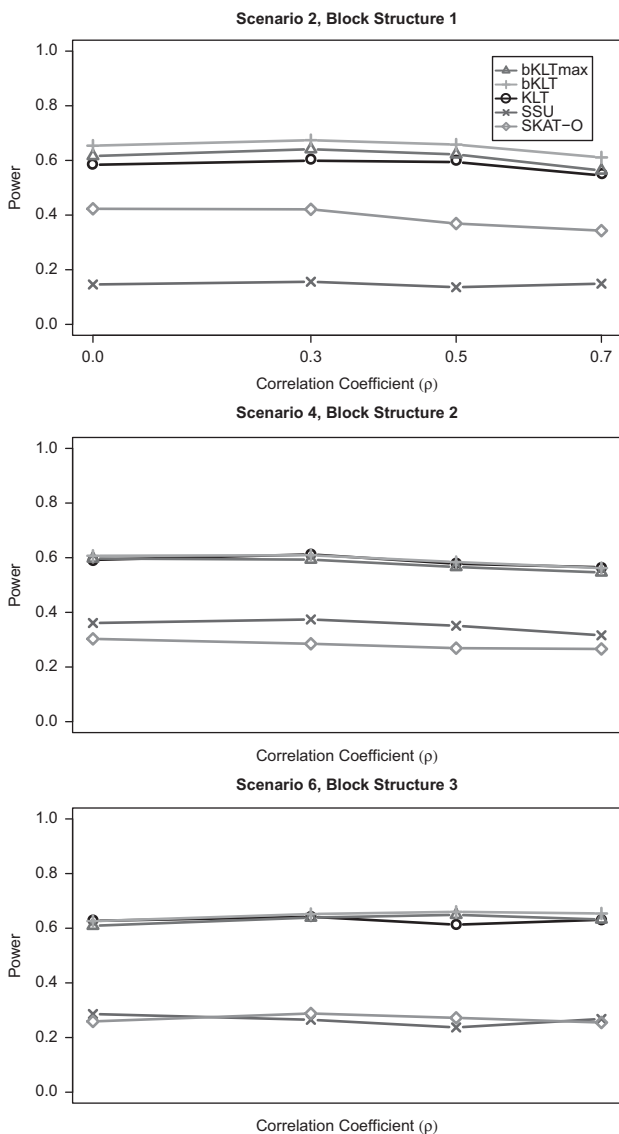
the sum in the definition of the bKLT statistic is over all three separate KTL statistics in the three blocks. For the other three methods, KLT, SSU and SKAT-O, how the blocks are formed is inconsequential as they analyze all the SNVs jointly without referring to the blocks. The results, as reported in Table 6, show that bKLTmax and bKLT achieve the smallest *P*-value (based on 10 000 permutations) compared with the other three methods. It is not surprising to see that bKLTmax and bKLT turn out to have the same *P*-value as bKLT will closely track bKLTmax if association signals are concentrated in one block, as it



**Table 4** The true and analysis models in three settings used for the robust analysis

Scenario	Block structure	Model	Block 1				Block 2			
			RC	RN	CC	CN	RC	RN	CC	CN
2	1	TRUE	8	8	0	0	0	16	0	0
		Analysis	6	8	0	0	2	16	0	0
4	2	TRUE	2	8	4	0	0	0	2	16
		Analysis	2	8	4	2	0	0	2	14
6	3	TRUE	2	8	2	0	0	0	4	16
		Analysis	2	8	4	0	0	0	2	16

Abbreviations: CC, common causal; CN, common neutral; RC, rare causal; RN, rare neutral.



**Figure 4** Power comparisons among bKLTmax, bKLT, KLT, SSU and SKAT-O for the robust analysis. Empirical power was calculated at the 5% significance level. The total sample size in each scenario is 1000 (500 cases and 500 controls). A full color version of this figure is available at the *Journal of Human Genetics* journal online.

appears to be the case in this particular data set. If it is indeed the case that most of the variants in block 1 and 3 are neutral as it is suggested from our results, then KLT will be less powerful compared with the two block versions, as KLT is more susceptible to the influence of noise. To show that our results for KLT, SSU and SKAT-O are consistent with those obtained by Turkmen *et al.*,<sup>25</sup> we also reproduced their results in Table 6. The minor discrepancy can be attributed to random variation (say, due to different random seeds).

## DISCUSSION

The importance of LD for studying associations between genetic variations and common diseases has been clearly documented in the literature for two decades since the seminal work of Risch and Merikangas.<sup>26</sup> However, LD has not been front and center in RV association studies despite the fact that their ignorance may result in false positives,<sup>28</sup> whereas accounting for them can lead to an increase in power for detecting associations. This lack of focus on RV LD is likely due to the existence of controversy views of LD<sup>37,38</sup> and the paucity of tests that are able to account for LD to increase their statistical power. To address the latter, we present two RV association tests that can make use of information available on LD block structure among the SNVs being studied. Both tests generalize the previous work of Turkmen *et al.*<sup>25</sup> by incorporating block structure information in the region of interest, thereby achieving improved power in some cases. Our first proposed test, bKLTmax, is aimed at maximizing the power of the originally proposed KLT when causal variants congregate within a segment of the region being studied. However, the advantage of bKLTmax will diminish if association signals scatter across multiple blocks; in such cases, the original KLT may gain an upper hand. To take advantage of the robust feature of KLT and the higher sensitivity of bKLTmax with more concentrated signals, we proposed our second block-based method, bKLT, which appears to perform well in a variety of settings considered. Compared with the results from KLT, SSU and SKAT-O, the performances of the two proposed tests are encouraging. Specifically, for block structures 1 and 2, when all the causal variants or most of the causal variants cluster in only one block, whereas the other block has no or a small proportion ( $\leq 25\%$ ) of causal variants, the two proposed tests have the highest power. For block structure 3 where association signals are evenly distributed, the two proposed tests are only a little less powerful than KLT and are still much more powerful than SSU and SKAT-O. As one can see from the simulation study, although bKLT is rarely the most powerful method among the three KLT-based methods, it is never too far behind the best one. As

**Table 5 Type I error rates (%) under the analysis model for the three settings in the robust analysis**

Scenario	Block structure	Method	$\rho$			
			0	0.3	0.5	0.7
2	1	bKLTmax	4.7	5.3	4.3	3.7
		bKLT	4.7	4.4	4	3.6
4	2	bKLTmax	5.1	5.2	4.7	5.6
		bKLT	4.6	5	5.3	5
6	3	bKLTmax	6	5.6	4.7	5.4
		bKLT	5.9	5.3	4.9	5.4

**Table 6 Result from the analysis of the Dallas Heart Study sequence data**

Method	P-value	
	Our analysis	Turkmen <i>et al.</i>
bKLTmax	0.0017	—
bKLT	0.0017	—
KLT	0.0689	0.0676
SSU	0.0701	0.0701
SKAT-O	0.0218	0.0237

The results from Turkmen *et al.* are also included to show consistency.

such, bKLT would be recommended in a real data analysis unless there is *a priori* information on the concentration of causal variants.

Despite the clear advantage of bKLT being robust yet having greater power, the limitation is that it relies on known or estimated LD blocks. As such, block ascertainment undoubtedly has an important role in the implementations of these methods. In the robust analysis, we investigated the performances of these two tests when we misrepresented the true underlying block structure. The results show that both bKLTmax and bKLT are not significantly affected: the type I error rate remains well controlled with small differences in power. As RVs behave quite differently than common ones, typical LD block finding programs no longer work. As such, further research is needed to find an adequate solution for uncovering LD structures when RVs are present. For example, several clustering methods focusing specifically on detecting the window in which causal variants cluster more than they do in the rest of the region have been proposed recently<sup>39–41</sup>. These methods may be promising for our purpose here, although careful consideration is needed. However, this is out of the scope of the current research and will be taken up in a future study. For the analysis of the DHS data, we adopted a search strategy coupled with traditional correlation involving a CV, and we were able to divide variants in the ANGPTL15 into three blocks. The results based on this preliminary blocking strategy are promising, as bKLTmax and bKLT were able to conclude more significant associations compared with the other tests that do not utilizing block information.

Theoretically, both bKLT and bKLTmax can handle any number of blocks and any number of variants within each block. There are two practical considerations, though. The first is that we need to compute the KLT statistic within each block and with all the variants combined. If there are many blocks and the number of variants within each block is large, then the computational demand would be an important factor

to consider. Using the computational time documented for KLT,<sup>25</sup> for a sample with 1000 individuals and 1000 permutations for significance assessment, the computational times would be ~1.5 and 3.2 s for an analysis of 32 and 64 variants, respectively. In the work by Yazdani *et al.*,<sup>30</sup> it is believed that 50 variants are sufficient to capture LD blocks. Whereas in our analysis of the DHS data, we determined three LD blocks with <10 variants per block. Therefore, it is likely that the computational time may be just seconds for each block in a real data analysis. As the whole genome is typically broken up into regions (for example, genomic regions of 5 kb), the number of blocks is typically small, say up to 10 blocks using existing work as a guideline.<sup>31</sup> As such, the number of blocks and the number of variants within each block in a real data scenario does not seem to pose undue computational burden on the block KLT methods.

Although the proposed method can handle any number of blocks, the tests may not be powerful enough if there are indeed multiple (>2) blocks and there is a proper subset of the blocks (but greater than one block) that contains ‘causal’ variants. To address this problem, one may expand the set of KLT statistics in the definition of bKLT to include those that are over a combination of blocks. This is a non-trivial extension and its feasibility will be studied in a future investigation.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## ACKNOWLEDGEMENTS

We thank two anonymous reviewers for their constructive comments and suggestions, which, we believe, have led to an improved manuscript. This research was supported in part by the Youth Science and Technology Innovation Fund of Nanjing Forestry University (CX2015027), Co-supervised model for PhD candidates program of Shanghai University of Finance and Economics, National Natural Science Foundation of China (11571082 and 11171075), National Basic Research Program of China (2012CB316505), the Scientific Research Foundation of Fudan University and the United States National Science Foundation (DMS-1208968).

- 1 Maher, B. Personal genomes: the case of the missing heritability. *Nature* **456**, 18–21 (2008).
- 2 Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* **40**, 695–701 (2008).
- 3 Gorlov, I. P., Gorlova, O. Y., Sunyaev, S. R., Spitz, M. R. & Amos, C. I. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **82**, 100–112 (2008).
- 4 Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387–389 (2009).
- 5 Rivas, M. A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C. K. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.* **43**, 1066–1073 (2011).
- 6 Bonnefond, A., Clément, N., Fawcett, K., Yengo, L., Vaillant, E., Guillaume, J.-L. *et al.* Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nat. Genet.* **44**, 297–301 (2012).
- 7 Moutsianas, L., Agarwala, V., Fuchsberger, C., Flannick, J., Rivas, M. A., Gaulton, K. J. *et al.* The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet.* **11**, e1005165 (2015).
- 8 Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
- 9 Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
- 10 Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res-Fund. Mol. M* **615**, 28–56 (2007).
- 11 Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5**, e1000384 (2009).

- 12 Price, A. L., Kryukov, G. V., de Bakker, P. I., 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).
- 13 Lin, D.-Y. & Tang, Z.-Z. A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.* **89**, 354–367 (2011).
- 14 Basu, S. & Pan, W. Comparison of statistical tests for disease association with rare variants. *Genet. Epidemiol.* **35**, 606–619 (2011).
- 15 Han, F. & Pan, W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum. Hered.* **70**, 42–54 (2010).
- 16 Hoffmann, T. J., Marini, N. J. & Witte, J. S. Comprehensive approach to analyzing rare genetic variants. *PLoS ONE* **5**, e13584 (2010).
- 17 Bhatia, G., Bansal, V., Harismendy, O., Schork, N. J., Topol, E. J., Frazer, K. *et al*. A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comput. Biol.* **6**, e1000954 (2010).
- 18 Zhang, L., Pei, Y.-F., Li, J., Pappasian, C. J. & Deng, H.-W. Efficient utilization of rare variants for detection of disease-related genomic regions. *PLoS ONE* **5**, e14288 (2010).
- 19 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).
- 20 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).
- 21 Pan, W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet. Epidemiol.* **33**, 497–507 (2009).
- 22 Goeman, J. J., Van De Geer, S. A. & Van Houwelingen, H. C. Testing against a high dimensional alternative. *J. Roy. Stat. Soc. B* **68**, 477–493 (2006).
- 23 Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775 (2012).
- 24 Pan, W., Kim, J., Zhang, Y., Shen, X. & Wei, P. A powerful and adaptive association test for rare variants. *Genetics* **197**, 1081–1095 (2014).
- 25 Turkmen, A. S., Yan, Z., Hu, Y.-Q. & Lin, S. Kullback-Leibler distance methods for detecting disease association with rare variants from sequencing data. *Ann. Hum. Genet.* **79**, 199–208 (2015).
- 26 Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **271**, 1516–1517 (1996).
- 27 Wang, T. & Elston, R. C. Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am. J. Hum. Genet.* **80**, 353–360 (2007).
- 28 Feng, T. & Zhu, X. Whole genome sequencing data from pedigrees suggests linkage disequilibrium among rare variants created by population admixture. *BMC Proc.* **8**, S44 (2014).
- 29 Talluri, R. & Shete, S. A linkage disequilibrium-based approach to selecting disease-associated rare variants. *PLoS ONE* **8**, 1–6 (2013).
- 30 Yazdani, A., Yazdani, A. & Boerwinkle, E. Rare variants analysis using penalization methods for whole genome sequence data. *BMC Bioinformatics* **16**, 405 (2015).
- 31 Turkmen, A. & Lin, S. Blocking approach for identification of rare variants in family-based association studies. *PLoS ONE* **9**, 1–11 (2014).
- 32 Kullback, S. & Leibler, R. A. On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951).
- 33 Victor, R. G., Haley, R. W., Willett, D. L., Peshock, R. M., Vaeth, P. C., Leonard, D. *et al*. The Dallas Heart Study: a population-based probability sample for the multi-disciplinary study of ethnic differences in cardiovascular health. *Am. J. Cardiol.* **93**, 1473–1480 (2004).
- 34 Romeo, S., Yin, W., Kozlitina, J., Pennacchio, L. A., Boerwinkle, E., Hobbs, H. H. *et al*. Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *J. Clin. Invest.* **119**, 70–79 (2009).
- 35 Epstein, M. P., Duncan, R., Jiang, Y., Conneely, K. N., Allen, A. S. & Satten, G. A. A permutation procedure to correct for confounders in case-control studies, including tests of rare variation. *Am. J. Hum. Genet.* **91**, 215–223 (2012).
- 36 Barrett, J. C. Haploview: visualization and analysis of SNP genotype data. *Cold Spring Harb Protoc.* **2009**, pdb.ip71 (2009).
- 37 Kinnamon, D. D., Hershberger, R. E. & Martin, E. R. Reconsidering association testing methods using single-variant test statistics as alternatives to pooling tests for sequence data with rare variants. *PLoS ONE* **7**, e30238 (2012).
- 38 Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).
- 39 Ionita-Laza, I., Makarov, V., Buxbaum, J. D. & Consortium, A. A. S. Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked and associated with autism spectrum disorders, in three datasets. *Am. J. Hum. Genet.* **90**, 1002–1013 (2012).
- 40 Fier, H., Won, S., Prokopenko, D., AlChawa, T., Ludwig, K. U., Fimmers, R. *et al*. 'Location, Location, Location': a spatial approach for rare variant analysis and an application to a study on non-syndromic cleft lip with or without cleft palate. *Bioinformatics* **28**, 3027–3033 (2012).
- 41 Schaid, D. J., Sinnwell, J. P., McDonnell, S. K. & Thibodeau, S. N. Detecting genomic clustering of risk variants from sequence data: cases versus controls. *Hum. Genet.* **132**, 1301–1309 (2013).