OPEN ACCESS Freely available online

PLOS ONE

Association Testing of Clustered Rare Causal Variants in Case-Control Studies

Wan-Yu Lin*

Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan

Abstract

Biological evidence suggests that multiple causal variants in a gene may cluster physically. Variants within the same protein functional domain or gene regulatory element would locate in close proximity on the DNA sequence. However, spatial information of variants is usually not used in current rare variant association analyses. We here propose a clustering method (abbreviated as "*CLUSTER*"), which is extended from the adaptive combination of *P*-values. Our method combines the association signals of variants that are more likely to be causal. Furthermore, the statistic incorporates the spatial information of variants. With extensive simulations, we show that our method outperforms several commonly-used methods in many scenarios. To demonstrate its use in real data analyses, we also apply this *CLUSTER* test to the Dallas Heart Study data. *CLUSTER* is among the best methods when the effects of causal variants are all in the same direction. As variants located in close proximity are more likely to have similar impact on disease risk, *CLUSTER* is recommended for association testing of clustered rare causal variants in case-control studies.

Citation: Lin W-Y (2014) Association Testing of Clustered Rare Causal Variants in Case-Control Studies. PLoS ONE 9(4): e94337. doi:10.1371/journal.pone.0094337

Editor: Zhaoxia Yu, University of California, Irvine, United States of America

Received January 29, 2014; **Accepted** March 12, 2014; **Published** April 15, 2014

Copyright: © 2014 Wan-Yu Lin. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by grants 102-2628-B-002-039-MY3 and 102-2314-B-002-001-MY2 from the Ministry of Science and Technology of Taiwan, and NTU-CESRP-101R7622-8, NTU-CESRP-102R7622-8, NTU-CESRP-103R7622-8, and NTU-CDP-102R7769 from National Taiwan University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: linwy@ntu.edu.tw

Introduction

The development in next-generation sequencing technologies has allowed a comprehensive investigation of the role of rare variants (minor allele frequency (MAF) $<1\%$) on complex diseases [1]. The low frequency of rare variants decreases the statistical power of detecting individual causal variants. Many statistical methods have been proposed to test for the collective association of multiple variants in a gene or region with diseases [2–13]. However, these methods do not incorporate the information of physical positions of the variants. Biological evidence suggests that multiple causal variants in a gene may cluster physically [14]. Variants within the same protein functional domain or gene regulatory element would locate in close proximity on the DNA sequence [15–18]. Furthermore, the spatial distribution of rare variants can be used to depict population structures [19]. These all constitute the importance of spatial approaches for rare variant association analyses.

Ionita-Laza et al. [14] has proposed a likelihood ratio scan statistic, and it successfully identifies clusters of rare deleterious variants with autism spectrum disorders. This method takes into account the underlying spatial distribution of variants, and we refer it to as "*IL-K*" because it is extended from the popular Kulldorff scan statistic [20]. It allows variable window sizes and calculates a likelihood ratio statistic for each window. The sliding window with the highest likelihood ratio statistic is the most likely region to harbor a cluster of rare deleterious variants. The statistical significance is assessed by permutation *P*-values [14].

Schaid et al. [16] has extended another popular spatial clustering method, Tango's statistic [21–23], to genomic sequence data. They incorporate the distance measures between variants into a kernel matrix, and therefore this method is referred to as "Kernel distance clustering" method (abbreviated as "*KERNEL*" hereafter). The statistic is $\delta' \mathbf{A} \delta$, where $\mathbf{A}$ is the kernel matrix with spatial information, and δ is the vector of case-control differences in variant frequencies. The statistical significance is also assessed by permutation *P*-values. Schaid et al. [16] have shown that *IL-K* outperforms *KERNEL* over a range of clustering scenarios, but *KERNEL* takes approximately half the computational time of *IL-K*.

We here propose a clustering method that is extended from the adaptive combination of *P*-values [24,25]. This method truncates the variants with larger *P*-values which are more likely to be neutral variants. With extensive simulations, we have shown that our method outperforms *KERNEL* [16], the weighted-sum approach (referred to as "*WS*") [4], and the variable threshold approach (referred to as "*VT*") [6], in the majority of scenarios. It also outperforms *IL-K* [14] and the sequence kernel association test (*SKAT*) [8,9] when all the causal variants are protective. We also apply this test to the Dallas Heart Study data [26,27], to demonstrate its use in real data analyses.

Materials and Methods

Suppose that there are K variant sites in a region of interest. We name the sites with larger variant frequencies in cases than in controls "deleterious-inclined variant sites", and those with larger variant frequencies in controls than in cases "protective-inclined variant sites". For a case-control study, the association of each variant with the disease status can be tested by the Fisher's exact test [13,28] or by the logistic regression (if covariate adjustment is required). Let the per-site *P*-values of the K variants be

$p_1, p_2, \cdots p_K$, respectively. To test for the significance of the region, we combine the per-site *P*-values that are smaller than some truncation threshold. Suppose we consider $\mathcal{J}$ candidate truncation thresholds, $\theta_1, \theta_2, \cdots, \theta_J$.

Multiple causal variants may cluster spatially in a functional region [14]. The proposed method is extended from the adaptive combination of *P*-values [24,25]. Furthermore, the spatial distribution of variants is taken into consideration. Under the *j*th truncation threshold (θ_j), the significance signal accumulated by the deleterious-inclined variant sites is $\delta_j^{+\prime}\mathrm{A}\delta_j^{+}$, where δ_j^{+} is a *K*-length vector with the *i*th element of $\sqrt{-\xi_i \cdot w_i \log p_i^{I[p_i<\theta_j]}}$. The indicator variable ξ_i is 1 if the *i*th site is deleterious-inclined and 0 otherwise, w_i is the weight given to the *i*th site (detailed in the next paragraph), and $I[p_i<\theta_j]$ is 1 if the *P*-value of the *i*th site is smaller than the *j*th truncation threshold (θ_j) and 0 otherwise. Similarly, under the *j*th truncation threshold, the significance signal accumulated by the protective-inclined variant sites is $\delta_j^{-\prime}\mathrm{A}\delta_j^{-}$, where δ_j^{-} is a *K*-length vector with the *i*th element of $\sqrt{-\varphi_i \cdot w_i \log p_i^{I[p_i<\theta_j]}}$. The indicator variable φ_i is 1 if the *i*th site is protective-inclined and 0 otherwise.

We follow Madsen and Browning [4] to determine the weights given to variant sites (w_i's). Let m_i^U be the number of mutant alleles observed for variant *i* in the unaffected subjects, and let n_i^U be the number of unaffected subjects genotyped for variant *i*. The frequency of variant *i* in the unaffected subjects is $q_i = \frac{m_i^U + 1}{2n_i^U + 2}$. The weight given to the *i*th site is $w_i = [n_i \cdot q_i(1-q_i)]^{-1}/2$, where n_i is the total number of subjects genotyped for variant *i*.

The $K \times K$ matrix A incorporates the spatial information of the variants. The (*i*, *j*)th element of A is $\left(1-d_{ij}^{'}2\right)^3$, where $d_{ij}^{'} = d_{ij}/\max_d$, d_{ij} is the physical distance between the *i*th and the *j*th variants, and $\max_d$ is a user-specified maximum distance of variants. Although the distance measure $\left(1-d_{ij}^{'}2\right)^3$ (named "tri-weight") is used throughout this work, it can be replaced by other measures (see [16]).

Under the *j*th truncation threshold, a test statistic regardless of the directions of effects (deleterious or protective) is $T_j^{(O)} = \max\left(\delta_j^{+\prime}\mathrm{A}\delta_j^{+}, \delta_j^{-\prime}\mathrm{A}\delta_j^{-}\right)$. With *B* permutations by randomly shuffling the case/control status, we obtain the permuted statistics $T_j^{(1)}, \cdots, T_j^{(B)}$. The *P*-value of the observed statistic $T_j^{(O)}$ is estimated by $\frac{\sum_{b=1}^{B} I\left(T_j^{(b)} \geq T_j^{(O)}\right)+1}{B+1}$, and the *P*-value of the b'th permuted statistic $T_j^{(b')}$ is estimated by $\frac{\sum_{b\neq b'} I\left(T_j^{(b)} \geq T_j^{(b')}\right)+1}{B}$. Across the $\mathcal{J}$ candidate truncation thresholds, the minimum *P*-value of the observed sample is $MinP^{(O)} = \min_{1\leq j\leq J} \frac{\sum_{b=1}^{B} I\left(T_j^{(b)} \geq T_j^{(O)}\right)+1}{B+1}$, and the minimum *P*-value of the b'th permuted sample is $MinP^{(b')} = \min_{1\leq j\leq J} \frac{\sum_{b\neq b'} I\left(T_j^{(b)} \geq T_j^{(b')}\right)+1}{B}$. Because we have *B* permutations, we compare $MinP^{(O)}$ with $MinP^{(1)}, \cdots, MinP^{(B)}$, and the "adjusted *P*-value" is estimated by $\frac{\sum_{b=1}^{B} I\left(MinP^{(b)} \leq MinP^{(O)}\right)+1}{B+1}$. This method is referred to as "*CLUSTER*", as it is proposed for detecting clusters of rare variants.

If we ignore the spatial information and let A be an identity matrix (all the diagonal elements are 1 and all the off-diagonal elements are 0), the statistic $T_j^{(O)}$ will be reduced to

$$T_j^{(O)} = \max\left(\delta_j^{+\prime}\mathrm{A}\delta_j^{+}, \delta_j^{-\prime}\mathrm{A}\delta_j^{-}\right) = \max\left(-\sum_{i=1}^{K} \xi_i \cdot w_i \log p_i^{I\left[p_i<\theta_j\right]}, -\sum_{i=1}^{K} \phi_i \cdot w_i \log p_i^{I\left[p_i<\theta_j\right]}\right).$$

This is equivalent to the statistic of the "adaptive combination of *P*-values for rare variant association testing" (abbreviated as "*ADA*") [24].

Simulation Study

To simulate real human genomic structure, we used the Cosi program [29] that was based on a coalescent process [30]. We generated 100 data sets, each containing 10,000 chromosomes of 1 Mb regions. The chromosomes were generated according to the linkage disequilibrium patterns of the HapMap CEU (Utah residents with ancestry from northern and western Europe) samples [31]. For each data set, we randomly selected a ~20 kb region. We considered two situations: (I) clustered causal variants: 20 rare causal variants were clustered within a ~6 kb region; (II) non-clustered causal variants: 20 rare causal variants were approximately equally spaced across the whole ~20 kb. The 20 causal variants were assumed to be (I) all protective; (II) 15 protective and 5 deleterious; (III) 10 protective and 10 deleterious; (IV) 5 protective and 15 deleterious; (V) all deleterious. The population attributable risk (PAR) of each causal variant was assumed to be 0%, 0.2%, 0.4%, 0.6%, 0.8%, and 1%, respectively.

Given PAR (PAR_j) and MAF (MAF_j) of the *j*th causal variant, its genotype relative risk (GRR) is:

$$GRR_j = \left(\frac{PAR_j}{(1-PAR_j)\cdot MAF_j} + 1\right)^{(-1)^{I\left(\phi_j=1\right)}},$$

[4,32–34]. The indicator function $I\left(\phi_j=1\right)$ is 1 if the *j*th causal variant is protective, and is 0 otherwise. The genotypes of a subject were formed by two chromosomes randomly drawn from the pool of 10,000 chromosomes. For a subject with chromosomes $\{H_1, H_2\}$, his/her disease status was generated by

$$P(affected|\{H_1,H_2\}) = f_0 \times \prod_{k=1}^{2}\prod_{j=1}^{d} GRR_j^{I\left(H_{k,j}=a_j\right)}$$

[32–34], where f_0 was the baseline penetrance (set at 1%), and a_j was the minor allele at the *j*th site. Chromosome pairs were randomly drawn from the chromosome pool with replacement until 500 cases and 500 controls were recruited.

Tests under Comparison

We compared *CLUSTER* with *IL-K* [14], *KERNEL* [16], *SKAT* [8,9], *WS* [4], and *VT* [6]. Single-nucleotide polymorphisms with MAF >5% in the combined sample of cases and controls were first removed from the analyses. The per-site *P*-values of individual

variants were obtained by the mid *P*-values from the Fisher's exact test [28]. The user-specified maximum distance $\max_d$ was fixed at 20 kb throughout this work. *IL-K* and *KERNEL* were implemented with the R package "vclust" [16]. The maximum window size considered by *IL-K* was set at 50% of the total region length, ~10 kb, as suggested by Ionita-Laza et al. [14]. When performing "*KERNEL*", tri-weight ($\left(1-d_{ij}'^2\right)^3$) was used as the distance measure between any two variants, because this was the default setting in the R package "vclust" [16]. To have a fair comparison, *CLUSTER* was implemented with the same tri-weight distance measure. The candidate truncation thresholds considered in *CLUSTER* were 0.10, 0.11, 0.12, …, 0.20. These are suitable *P*-value truncation thresholds for rare variant association testing [24].

Two burden tests including *WS* and *VT* were implemented with the R script by Price et al. [6] (http://genetics.bwh.harvard.edu/rare_variants/). As a representative method of non-burden tests, *SKAT* was also included into comparisons. *SKAT* was implemented with the R package "SKAT" [35]. The weight given to the *j*th variant site (with MAF of MAF_j) was set at $w_j = Beta(MAF_j,1,25)$, because this was the default weight function in the package "SKAT". Note that the *SKAT* [9] compared here is the test that optimally combines the burden tests and the original *SKAT* proposed by Wu et al. [8].

The *P*-values of *CLUSTER*, *IL-K*, *KERNEL*, *WS*, and *VT* were obtained with 10,000 permutations when evaluating type-I error rates and 1,000 permutations when evaluating power, respectively. For *SKAT*, we used the default Davies method [36] in the package "SKAT" to compute *P*-values.

Results

Type-I Error Rates

The type-I error rates were measured when PAR was set at 0%. We performed 1,000 replications for each of the 100 simulated data sets. Therefore, there were totally 100,000 ($=100\times1000$) replications. Table 1 summarizes the type-I error rates given various nominal significance levels. The type-I error rates of all the six methods match the corresponding nominal significance levels.

Power Comparisons

To evaluate power, a total of 100 replications were performed under each scenario for each of the 100 simulated data sets. Figure 1 presents the power averaged over the $100\times100=10{,}000$ replications. When all the 20 causal variants were protective, *CLUSTER* was much more powerful than other methods. Under a mixture of deleterious and protective variants, *IL-K*, *SKAT*, and *CLUSTER* were powerful methods. However, *CLUSTER* had decreased power when the causal variants were non-clustered (see the bottom row). When all the 20 causal variants were deleterious, *IL-K*, *SKAT*, and *CLUSTER* were again the more powerful methods. Note that the effect size (measured by the magnitude of odds ratio) of a deleterious variant was larger than that of a protective variant with the same PAR and MAF (as shown by Lin et al. [24]). Therefore, all the methods performed better under 20 deleterious variants (the right column of Fig. 1) than under 20 protective variants (the left column of Fig. 1).

Table 1. Type-I error rates.

nominal significance level	0.001	0.01	0.02	0.03	0.04	0.05
SKAT	0.0011	0.0102	0.0201	0.0303	0.0404	0.0503
CLUSTER	0.0008	0.0101	0.0204	0.0303	0.0401	0.0502
KERNEL	0.0011	0.0103	0.0187	0.0294	0.0404	0.0503
IL-K	0.0011	0.0100	0.0202	0.0298	0.0402	0.0501
WS	0.0008	0.0101	0.0200	0.0302	0.0404	0.0503
VT	0.0009	0.0100	0.0202	0.0304	0.0405	0.0502

doi:10.1371/journal.pone.0094337.t001

We also evaluated the power performance of these tests when the number of causal variants was 10. Figure 2 shows the results of two situations considered: (I) clustered causal variants: 10 rare causal variants were clustered within a ~3 kb region; (II) non-clustered causal variants: 10 rare causal variants were approximately equally spaced across the whole ~20 kb. The 10 causal variants were assumed to be (I) all protective; (II) 8 protective and 2 deleterious; (III) 5 protective and 5 deleterious; (IV) 2 protective and 8 deleterious; (V) all deleterious. The result was similar to that shown by Fig. 1. *CLUSTER* was among the best methods when the effects of causal variants were all in the same direction, but it had decreased power under a mixture of deleterious and protective variants (see columns 2–4 of Figs. 1 & 2). This is because the test statistic $T_j^{(O)}=\max\left(\delta_j^{+\prime}\mathrm{A}\delta_j^{+},\delta_j^{-\prime}\mathrm{A}\delta_j^{-}\right)$ facilitates the detection of variants with effects in a consistent direction. We will further discuss this in the Discussion section.

In Figs. 1 and 2, the power from the top panel (clustered situation) is generally lower than that from the bottom panel (non-clustered situation). This is because, when the causal variants are clustered in a small region (~6 kb or ~3 kb, in the simulations), the variants far from this region will have almost no correlation (or, no linkage disequilibrium) with the causal variants. Therefore, they can hardly provide any association signal when testing for the whole region (~20 kb here). When the causal variants are equally spaced across the whole region, the variants surrounding each causal variant can provide some signal because of their correlation with the causal ones. Although the correlation between rare variants is usually low [37,38], it can still boost the power to some extent. This is a general trend for all the methods. What we can compare is the performance of the methods with spatial information (*CLUSTER*, *IL-K*, and *KERNEL*) relative to that of the methods without spatial information (*SKAT*, *WS*, and *VT*), in clustered situation or in non-clustered situation.

Not surprisingly, the clustered situation favors the methods considering spatial information (*CLUSTER*, *IL-K*, and *KERNEL*). They were relatively (relative to *SKAT*, *WS*, and *VT*) more powerful when the causal variants were clustered (top panels of Figs. 1 and 2). *CLUSTER* had good performance and was more powerful than *KERNEL*. *IL-K* also had good power performance, except when all the causal variants were protective (see the left columns of Figs. 1 and 2). With a mixture of protective and deleterious variants, *IL-K* was generally more powerful than *CLUSTER*, especially when the PAR was larger (see columns 2–4 of Figs. 1 and 2).

It was worth noting that *CLUSTER* outperformed *SKAT*, even when the causal variants were non-clustered (see the bottom-left plots of Figs. 1 and 2). This may be attributed to the "noise truncation" property of *CLUSTER*. The effect size (measured by the magnitude of odds ratio) of a protective variant was smaller than that of a deleterious variant with the same PAR and MAF (as shown by Lin et al. [24]). The effects of the protective variants were rather mild, and most methods were underpowered. *CLUSTER* takes the advantage of truncating neutral variants with larger *P*-values. *CLUSTER* is an extension of *ADA*, and this

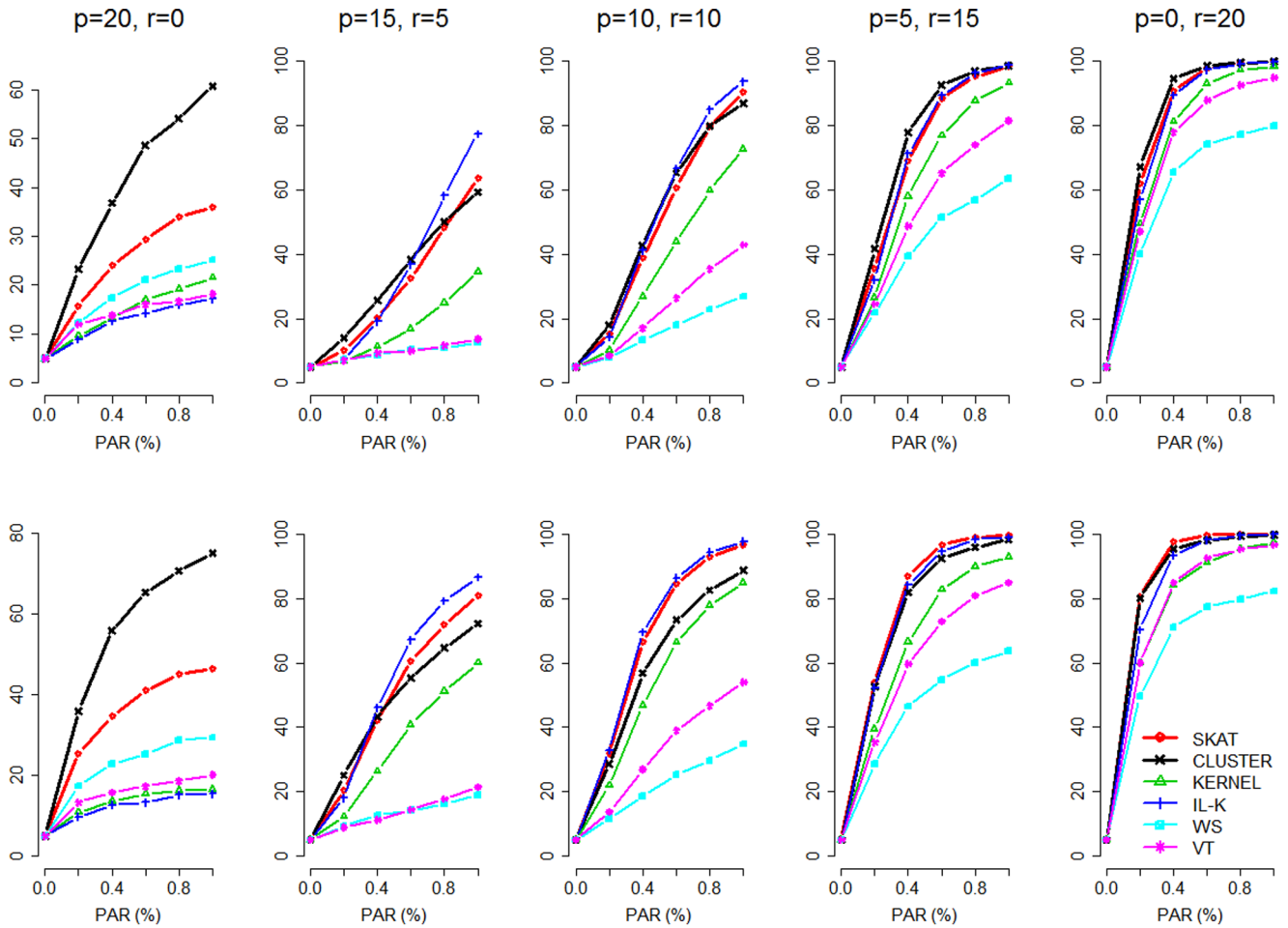

Figure 1. Simulation-Based Power Comparisons (20 rare causal variants). The figure shows the empirical power at $\alpha = 0.05$. Top panel: clustered causal variants; bottom panel: non-clustered causal variants. The letters "p" and "r" denote the numbers of protective variants and deleterious (or, risky) variants, respectively.
doi:10.1371/journal.pone.0094337.g001

outcome is consistent with that observed in the *ADA* paper (see the left column of Fig. 2 of [24]).

Application to Dallas Heart Study Data

These six tests were then applied to the Dallas Heart Study [26,27,39]. This study was to uncover the role of *ANGIOPOIETIN–LIKE 4* (*ANGPTL4*) in plasma triglyceride levels. The genotypes of 1,045 European Americans were analyzed. We first used a linear regression to adjust the log-transformed triglyceride levels for age, sex, and BMI. Subjects with residuals smaller than the 30th percentile and larger than the 70th percentile were treated as controls and cases, respectively. Then, the subjects with missing genotypes were excluded from the analysis. Finally, 179 cases and 213 controls were left.

The six tests were applied to this data set. The variants with MAF >5% were removed. To have an exhaustive search for the most likely region to harbor causal variants, the maximum window size considered by *IL-K* was set as the total region length (~10 kb). As a result, only *CLUSTER* and *SKAT* had *P*-values smaller than 0.05 (see Table 2).

The significant association of *ANGPTL4* with triglyceride was previously reported. Results in over 30,000 subjects from non-diabetic and population-based studies have confirmed that variants in *ANGPTL4* reduce triglyceride and exert protective effects against hyperlipidemia [26,40,41]. With the significance level of 0.05, only *CLUSTER* and *SKAT* confirmed this association. The other two spatial approaches, *IL-K* and *KERNEL*, were shown (by simulations) to have low power when all the causal variants were protective. No wonder they failed to detect the association here. This result is consistent with the finding from our simulation study.

Discussion

Multiple rare variants may cluster in a functional region [14–18]. Variants within the same protein functional domain may locate in close proximity and have similar impact on disease risk [15,17]. Consistent with the finding from Schaid et al. [16], *KERNEL* usually has lower power than *IL-K*. However, when all the causal variants are protective, *IL-K* has very low power (see the left columns of Figs. 1 & 2). This is because *IL-K* can only identify deleterious variants [14]. When all the causal variants are protective, *CLUSTER* and *SKAT* are more powerful than other methods. No wonder only these two methods could detect the protective effect of the variants in *ANGPTL4* against hyperlipidemia [26,40,41], in the Dallas Heart Study data analysis.

As mentioned in the Methods section, a test statistic regardless of the directions of effects (deleterious or protective) is

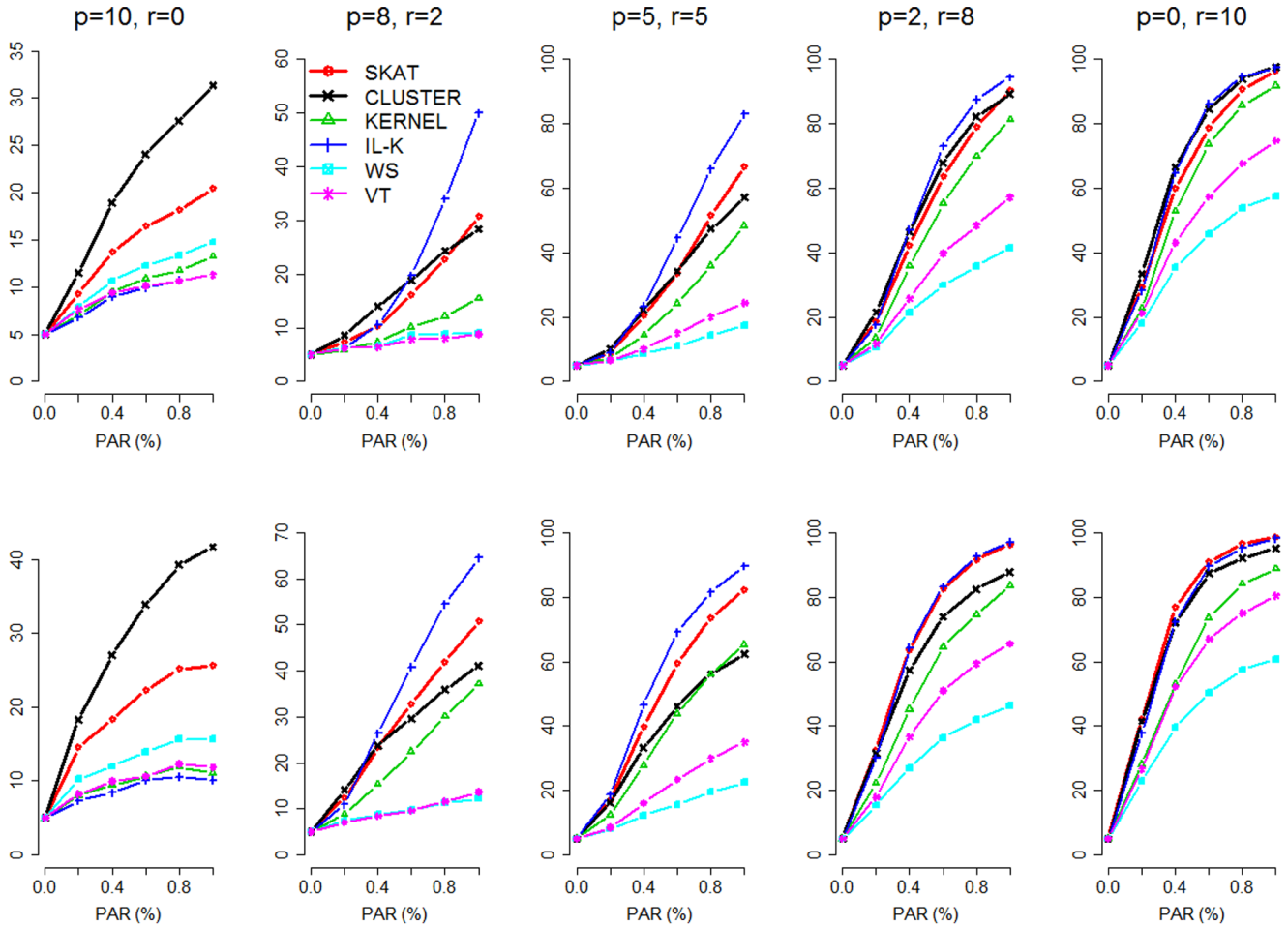

Figure 2. Simulation-Based Power Comparisons (10 rare causal variants). The figure shows the empirical power at $\alpha = 0.05$. Top panel: clustered causal variants; bottom panel: non-clustered causal variants. The letters "p" and "r" denote the numbers of protective variants and deleterious (or, risky) variants, respectively.
doi:10.1371/journal.pone.0094337.g002

$T_j^{(O)} = \max\left(\delta_j^+{}'A\delta_j^+, \delta_j^-{}'A\delta_j^-\right)$ under the jth truncation threshold. Another reasonable statistic is $T_j^{(O)} = \delta_j^+{}'A\delta_j^+ + \delta_j^-{}'A\delta_j^-$. This is more powerful than *CLUSTER* when ~50% of the causal variants are deleterious, but is less powerful when the effects of variants are all in the same direction. Because clustered variants are more likely to have effects in the same direction, we still suggest using $T_j^{(O)} = \max\left(\delta_j^+{}'A\delta_j^+, \delta_j^-{}'A\delta_j^-\right)$, instead of $T_j^{(O)} = \delta_j^+{}'A\delta_j^+ + \delta_j^-{}'A\delta_j^-$. Note that even the statistic, $\delta_j^+{}'A\delta_j^+ + \delta_j^-{}'A\delta_j^-$, is started from aggregating the information of "deleterious-inclined variants" and "protective-inclined variants", separately. Under the assumption that deleterious variants and protective variants may have their own clusters, we do not mix all the variants together in the very beginning (i.e., $\delta' A\delta$, this will incorporate the distance between "deleterious-inclined variants" and "protective-inclined variants" into the statistic).

Table 2. Application to the Dallas Heart Study data.

	SKAT	*CLUSTER*	*KERNEL*	*IL-K*	*WS*	*VT*
P-value	0.0245	0.0125[a]	0.0899[a]	0.1398[a]	0.1841[a]	0.4858[a]

[a]*P*-values were obtained by 10,000 permutations.
doi:10.1371/journal.pone.0094337.t002

All the methods evaluated here require permutations to obtain accurate *P*-values, except *SKAT* that uses the Davies method [36] to compute *P*-values. For simulated data sets each containing 500 cases and 500 controls in ~20 kb regions (including ~330 nonsynonymous variant sites), the computation time lengths were ordered as *CLUSTER* (~151.7 sec) > *SKAT* (~30.2 sec) > *IL-K* (~20.4 sec) > *KERNEL* (~6.7 sec) > *VT* or *WS* (~3.4 sec), where 1000 permutations were used for all the methods except *SKAT*. This was timed by a Linux workstation with an Intel Xeon E5-2690 2.9 GHz processor and 6 GB memory. *CLUSTER* takes a longer time to compute because it incorporates the spatial kernel matrix into the search of the optimal *P*-value truncation threshold.

Schaid *et al.* [16] showed that *IL-K* and *KERNEL* could have higher power than *SKAT*, when the variants were correlated. Without correlation, *SKAT* tended to have the highest power among the tests they compared [16]. In fact, the correlation between rare variants is usually low [37,38]. Our simulated data sets were generated from the coalescent process [30] and they reflected realistic DNA sequences. Therefore, in our simulations, the correlation between rare variants is low and *SKAT* is better than *KERNEL* (and sometimes better than *IL-K*).

KERNEL and *CLUSTER* have similar forms in test statistics ($\delta' A \delta$), and they are both implemented with the tri-weight distance measure in our simulations. However, the results showed that *CLUSTER* outperformed *KERNEL*. This is because *CLUSTER* combines the association signals (*P*-values) of variants that are more likely to be causal, i.e., truncates variants with larger *P*-values. *CLUSTER* is among the best methods when the effects of causal variants are in one direction. As variants located in close proximity are more likely to have similar impact on disease risk [15,17], *CLUSTER* is recommended for association testing of clustered rare causal variants in case-control studies.

Acknowledgments

The author would like to thank the anonymous reviewers for their insightful and constructive comments, and Drs. Jonathan C. Cohen and Helen H. Hobbs for kindly providing the Dallas Heart Study data.

Author Contributions

Conceived and designed the experiments: WYL. Performed the experiments: WYL. Analyzed the data: WYL. Contributed reagents/materials/analysis tools: WYL. Wrote the paper: WYL.

References

1. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
2. Bansal V, Libiger O, Torkamani A, Schork NJ (2010) Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet 11: 773–785.
3. Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83: 311–321.
4. Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5: e1000384.
5. Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34: 188–193.
6. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, et al. (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86: 832–838.
7. Han F, Pan W (2010) A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 70: 42–54.
8. Wu MC, Lee S, Cai T, Li Y, Boehnke M, et al. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89: 82–93.
9. Lee S, Wu MC, Lin X (2012) Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13: 762–775.
10. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, et al. (2011) Testing for an unusual distribution of rare variants. PLoS Genet 7: e1001322.
11. Yi N, Liu N, Zhi D, Li J (2011) Hierarchical generalized linear models for multiple groups of rare and common variants: jointly estimating group and individual-variant effects. PLoS Genet 7: e1002382.
12. Yi N, Zhi D (2011) Bayesian analysis of rare variants in genetic association studies. Genet Epidemiol 35: 57–69.
13. Cheung YH, Wang G, Leal SM, Wang S (2012) A fast and noise-resilient approach to detect rare-variant associations with deep sequencing data for complex disorders. Genet Epidemiol 36: 675–685.
14. Ionita-Laza I, Makarov V, Buxbaum JD (2012) 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.
15. Fier H, Won S, Prokopenko D, AlChawa T, Ludwig KU, et al. (2012) '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.
16. Schaid DJ, Sinnwell JP, McDonnell SK, Thibodeau SN (2013) Detecting genomic clustering of risk variants from sequence data: cases versus controls. Hum Genet 132: 1301–1309.
17. Krebs JE, Goldstein ES (2011) Lewin's GENES X. Jones and Bartlett Publishers, Sudbury.
18. Raab JR, Kamakaka RT (2010) Insulators and promoters: closer than we think. Nat Rev Genet 11: 439–446.
19. Mathieson I, McVean G (2012) Differential confounding of rare and common variants in spatially structured populations. Nat Genet 44: 243–246.
20. Kulldorff M (1997) A spatial scan statistic. Communications in Statistics - Theory and Methods 26: 1481–1496.
21. Tango T (1984) The detection of disease clustering in time. Biometrics 40: 15–26.
22. Tango T (2000) A test for spatial disease clustering adjusted for multiple testing. Stat Med 19: 191–204.
23. Tango T (2010) Statistical methods for disease clustering. Springer, New York.
24. Lin WY, Lou XY, Gao G, Liu N (2014) Rare Variant Association Testing by Adaptive Combination of P-values. PLoS One 9: e85728.
25. Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, et al. (2009) Pathway analysis by adaptive combination of P-values. Genet Epidemiol 33: 700–709.
26. Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, et al. (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat Genet 39: 513–516.
27. Romeo S, Yin W, Kozlitina J, Pennacchio LA, Boerwinkle E, et al. (2009) Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. J Clin Invest 119: 70–79.
28. Fisher RA (1922) On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. Journal of the Royal Statistical Society 85: 87–94.
29. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, et al. (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15: 1576–1583.
30. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.
31. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.
32. Li Y, Byrnes AE, Li M (2010) To identify associations with rare variants, just WHaIT: Weighted haplotype and imputation-based tests. Am J Hum Genet 87: 728–735.
33. Lin WY, Yi N, Lou XY, Zhi D, Zhang K, et al. (2013) Haplotype kernel association test as a powerful method to identify chromosomal regions harboring uncommon causal variants. Genet Epidemiol 37: 560–570.
34. Lin WY, Yi N, Zhi D, Zhang K, Gao G, et al. (2012) Haplotype-based methods for detecting uncommon causal variants with common SNPs. Genet Epidemiol 36: 572–582.
35. Lee S, Miropolsky L, Wu M (2013) Package 'SKAT', http://cran.r-project.org/web/packages/SKAT/index.html. Accessed Jan 2, 2013.
36. Davies RB (1980) Algorithm AS 155: the distribution of a linear combination of $\chi 2$ random variables. Journal of the Royal Statistical Society Series C (Applied Statistics) 29: 323–333.
37. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69: 124–137.
38. Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant.or not? Hum Mol Genet 11: 2417–2423.
39. Victor RG, Haley RW, Willett DL, Peshock RM, Vaeth PC, et al. (2004) The Dallas Heart Study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. Am J Cardiol 93: 1473–1480.
40. Talmud PJ, Smart M, Presswood E, Cooper JA, Nicaud V, et al. (2008) ANGPTL4 E40K and T266M: effects on plasma triglyceride and HDL levels, postprandial responses, and CHD risk. Arterioscler Thromb Vasc Biol 28: 2319–2325.
41. Smart-Halajko MC, Kelley-Hedgepeth A, Montefusco MC, Cooper JA, Kopin A, et al. (2011) ANGPTL4 variants E40K and T266M are associated with lower fasting triglyceride levels in Non-Hispanic White Americans from the Look AHEAD Clinical Trial. BMC Med Genet 12: 89.