Xiang et al. BMC Genetics (2020) 21:130
https://doi.org/10.1186/s12863-020-00951-2

BMC Genetics

METHODOLOGY ARTICLE Open Access

Identifying rare variants for quantitative traits in extreme samples of population via Kullback-Leibler distance

Yang Xiang[1,2,3], Xinrong Xiang[4] and Yumei Li[1,2,3*]

Abstract

Background: The rapid development of sequencing technology and simultaneously the availability of large quantities of sequence data has facilitated the identification of rare variant associated with quantitative traits. However, existing statistical methods depend on certain assumptions and thus lacking uniform power. The present study focuses on mapping rare variant associated with quantitative traits.

Results: In the present study, we proposed a two-stage strategy to identify rare variant of quantitative traits using phenotype extreme selection design and Kullback-Leibler distance, where the first stage was association analysis and the second stage was fine mapping. We presented a statistic and a linkage disequilibrium measure for the first stage and the second stage, respectively. Theory analysis and simulation study showed that (1) the power of the proposed statistic for association analysis increased with the stringency of the sample selection and was affected slightly by non-causal variants and opposite effect variants, (2) the statistic here achieved higher power than three commonly used methods, and (3) the linkage disequilibrium measure for fine mapping was independent of the frequencies of non-causal variants and simply dependent on the frequencies of causal variants.

Conclusions: We conclude that the two-stage strategy here can be used effectively to mapping rare variant associated with quantitative traits.

Keywords: Quantitative trait, Rare variant, Association analysis, Fine mapping, Extreme phenotype

Background

Thanks to the rapid development of sequencing technology and the lowering of sequencing costs in the last decade, the availability of large quantities of sequence data provides an unprecedented opportunity for researchers to investigate the role of rare variants in complex traits [1–4]. But due to the low minor allele frequency (MAF < 5%) and thus resulting in weak linkage disequilibrium (LD) with nearby markers, detecting rare variant (RV) association with complex traits faces great challenges [5–8]. One challenge is that detection of rare causal variants with traditional designs usually requires a large sample, which will be the high cost [3, 6]. Thus cost-effective design should be considered to reduce sample size. Another challenge is that the statistical power with test statistics of single-marker tests is generally low in genetic association studies of rare variants with more moderate or weak genetic effects [8–10]. To date many statistical methods have been developed for rare variant association analysis, including burden tests [11–13], variance-component tests [14, 15], series of sequence kernel association tests [10, 16, 17]. Any of these methods has relative perfect performance in special

* Correspondence: lymmail@126.com; liym74@yahoo.com
[1]School of Mathematics and Computational Science, Huaihua University, Huaihua, Hunan 418008, People's Republic of China
[2]Key Laboratory of Research and Utilization of Ethnomedicinal Plant Resources of Hunan Province, Huaihua University, Huaihua 418008, China
Full list of author information is available at the end of the article

scenario, but none of them can overwhelm others in all scenarios [8, 9], especially for quantitative traits.

In fact, rare variant association analysis in the past several years mainly focused on the qualitative trait. Only a few statistical methods have been developed for the quantitative trait [13, 18–21]. One approach for rare variant association analysis of quantitative traits is the linear regression model. However, most regression-based methods rely on the normality assumption of the phenotype [8, 21]. Another commonly used approach adopts phenotype extreme selection design where one can transform the quantitative trait association study into case-control association study of qualitative traits by treating the upper extreme as cases and the lower extreme as controls in a strategy using extreme phenotype [22–25]. Extreme phenotypes of a quantitative trait are generally considered to be more informative. Moreover, a smaller sample size for extreme-phenotype sampling than that for random sampling is needed to achieve similar power [23, 24].

In this report, we use phenotype extreme selection design and Kullback-Leibler distance (KL-distance) [26] to propose a simple statistic method to identify rare variants for quantitative traits. Two stages strategies are adopted in our analysis where association analysis and fine mapping will be done in the first stage and the second stage, respectively. This method will compare the frequency distributions of rare variant in two extreme phenotypes based on KL-distance. Our method has three features: (1) it has increasing power with the stringency of the sample selection, (2) it is affected slightly by non-causal variants and the opposite effect variants.in the first stage for association analysis, and (3) it is not depend on the frequencies of non-causal variants and just dependent on the frequencies of causal variants in the second stage for fine mapping. Through simulation studies, we investigate the performance of the proposed method and compare it with three commonly used methods of the burden test [12], the sequence kernel association test (SKAT) [17], and the optimal test that combines SKAT and the burden test (SKAT-O) [10].

Results

Type I error rate and power for association analysis

Table 1 exhibits the estimated type I error rates of the statistic T_{KL} for the extreme sample with sample-selection threshold value of 20, 10, and 5% and with sample size of 1000 and 1500. It can be seen that, under various genetic parameters, type I error rates of T_{KL} are not appreciably different from the nominal alpha levels, which indicates the validity of the statistic T_{KL}.

Table 1 Estimated type I error rates of the statistic T_{KL}

Threshold values	Estimated Type I error rate			
	2 *N* = 1000		2 *N* = 1500	
	α = .05	*α* = .01	*α* = .05	*α* = .01
20%	0.048	0.012	0.048	0.011
10%	0.049	0.014	0.051	0.013
5%	0.053	0.009	0.052	0.013

Figure 1 shows the results of power for 9 scenarios when sample sizes are 1000 and 1500. It is found that the power of the statistic T_{KL} with the sample size of 1500 is nearly 0.20 larger than that with the sample size of 1000, indicating that the power of the statistic T_{KL} significantly increase with the increasing of the sample size. It can be seen that, under the same sample size, the powers of the statistic T_{KL} with the low 5% samples and the up 5% samples are highest and the powers with the low 20% samples and the up 20% samples are lowest, which indicates that the powers of the statistic T_{KL} increase with the stringency of the sample selection. It is observed from scenarios {1, 2, 3} that, when rare variant effects are in the same direction, the powers of the statistic T_{KL} increase with the increasing of the number of causal variants. The same above conclusions are observed when 80% causal variants have positive effect and 20% causal variants have negative effect (scenarios {4, 5, 6}) and when there is the same number of causal variants with positive effects and negative effects (scenarios {7, 8, 9}). By comparing the powers under scenarios {1, 4, 7} with 10 causal variants, the powers under scenarios {2, 5, 8} with 20 causal variants, and the powers under scenarios {3, 6, 9} with 50 causal variants, we found that, when the number of causal variants with negative effect increases, the power of the statistic T_{KL} decreases slightly. From Fig. 1, we can observe that, among four statistics of the T_{KL}, the burden test, the SKAT, and the SKAT-O, the power of T_{KL} is higher than that of other three statistics. The burden test, the SKAT, and the SKAT-O are severely affected by the number of non-causal variants and the opposite effect variants, especially when there are the same number of opposite effect variants. Although non-causal variants and the opposite effect variants affect the power of the T_{KL}, the impact is slight. For example, when the sample size is 1500 and the number of causal variants is 50 for 10% sample-selection threshold value (B2), as the number of variants with negative effect increases from zero to 25, the powers of the burden test, the SKAT, and the SKAT-O decrease from ~ 0.80, ~ 0.79, and ~ 0.84 to ~ 0.23, ~ 0.63, and ~ 0.74, respectively, with the decline rate of 71.2, 20.2, and 12.0%. Nevertheless, when the number of variants with negative effect is 25, the T_{KL} still achieves ~ 0.83 power, with the decline rate of just 7% comparing to ~ 0.90 when the number of variants with negative effect is zero.

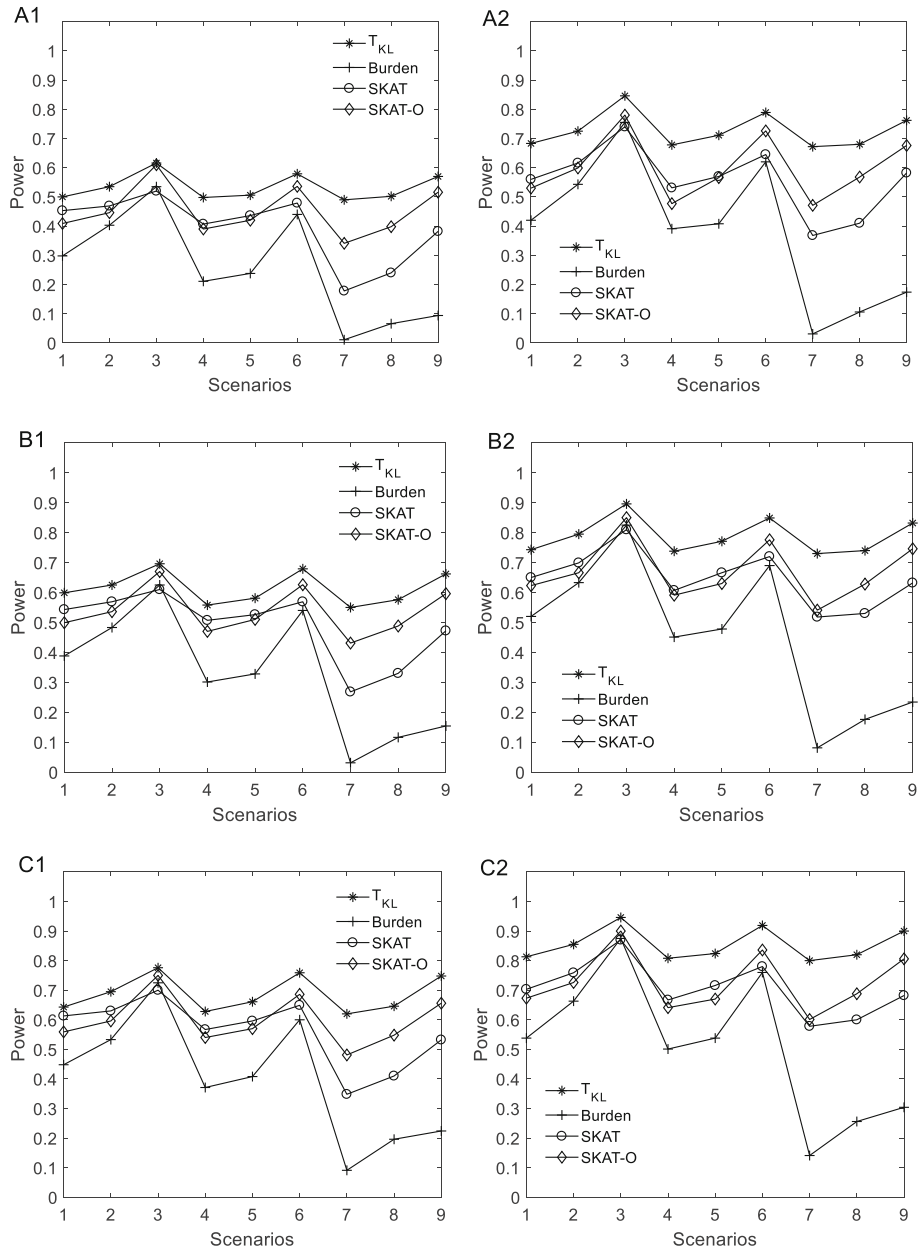

Fig. 1 Empirical power of four statistics from the extreme samples with 20% threshold value **a**, 10% threshold value **b**, and 5% threshold value **c** when the sample sizes are 1000 (**a1, b1, c1**) and 1500 (**a2, b2, c2**) at a 0.05 significance level

Power for fine mapping

In fine mapping study, the QTL can be located by the maximum value of the measure l_{KL}. So we sample 10 times from each of 100 simulation populations where each sample includes 750 individuals with the up-extreme phenotype of $Y > U$ and 750 individuals with the lower-extreme phenotype with $Y < L$. For each sample, we calculate the value of the measure l_{KL} for each variant. In order to guard against noisy distributions of the measure l_{KL}, we adopt the 5-point moving-average method to determine the maximum value. We count the number (here, we denote it B) of the maximum values that locate at variant 10 or variant 11. Then the probability that the maximum values of l_{KL} locate at variant 10 or variant 11 is B/1000. We refer this value as the power of l_{KL}, which measure the likelihood of fine mapping the QTL. Table 2 shows the results of the power for l_{KL}. It can be seen that the power of l_{KL} for fine mapping under dominant model is highest and the power of l_{KL} for fine mapping under recessive model is lowest. The power of l_{KL} increases with increasing of the heritability h^2 of the causal variant and the stringency of the sample selection. For example, power of l_{KL} under

Table 2 The power of the QTL fine mapping for three LD measures by use of five-point moving average

Sample-selection threshold values	Power of the QTL fine mapping								
	Recessive model			Additive model			Dominant model		
	$h^2 = 0.01$	$h^2 = 0.05$	$h^2 = 0.10$	$h^2 = 0.01$	$h^2 = 0.05$	$h^2 = 0.10$	$h^2 = 0.01$	$h^2 = 0.05$	$h^2 = 0.10$
20%									
l_{KL}	0.39	0.50	0.58	0.44	0.53	0.62	0.52	0.59	0.70
l	0.40	0.49	0.59	0.45	0.53	0.62	0.51	0.60	0.70
p_{excess}	0.29	0.36	0.47	0.36	0.47	0.58	0.41	0.52	0.61
10%									
l_{KL}	0.49	0.59	0.66	0.52	0.64	0.70	0.62	0.69	0.80
l	0.50	0.59	0.67	0.53	0.63	0.71	0.61	0.68	0.80
p_{excess}	0.37	0.51	0.56	0.44	0.52	0.63	0.55	0.60	0.68
5%									
l_{KL}	0.56	0.64	0.71	0.61	0.67	0.77	0.67	0.75	0.83
l	0.56	0.64	0.71	0.61	0.67	0.77	0.67	0.75	0.83
p_{excess}	0.41	0.55	0.63	0.48	0.51	0.65	0.59	0.68	0.72

Note: The MAF of the causal variant is 0.01($P_a = 0.01$). The sample size is 1500 ($2N = 1500$)

dominant model with the heritability h^2 of 0.01 is 0.52, 0.62, and 0.67 at 20, 10, and 5% sample-selection threshold value, respectively; power of l_{KL} under dominant at 5% sample-selection threshold value increase from 0.67 to 0.83 with the heritability h^2 of the causal variant increasing from 0.01 to 0.10. We also investigate the effect of different sample sizes (e.g., 2n =1000, 1500, and 2000). As expected, power of l_{KL} increases with the increasing sample size (data not shown). In order to assess the performance of l_{KL}, we compare it with two LD measures l [27] and p_{excess} [28] with case-control design using extreme samples. Table 2 also lists the powers for l and p_{excess}. We found that the powers of l_{KL} and l are nearly the same and higher than those of p_{excess}.

Discussion

In this report, we present a robust approach to identify rare variant of quantitative traits. The proposed approach adopts phenotype extreme selection design and KL-distance method. We use a two-stage strategy in our analysis where the first stage is association analysis and the second stage is fine mapping of QTL if the first stage is positive result. We propose a statistic T_{KL} for association analysis and a LD measure l_{KL} for fine mapping. Simulation studies present the performance of the proposed method. We found that the power of the T_{KL} increases with the stringency of the sample selection and the increasing of the number of causal variants. The T_{KL} here has higher power for association analysis than three existing statistics. Meanwhile, the impact of non-causal variants and the opposite effect variants on the T_{KL} is slight. The LD measure l_{KL} for fine mapping in the second stage has a good feature of not dependence on the frequencies of non-causal variants and just dependence on the frequencies of causal variants. These results show that our method can be used to detect rare variant associated with quantitative traits. At the same time, we found that the proposed method can be easily extended to case-control study by treating cases and controls as samples with upper extreme phenotype and lower extreme phenotype, respectively.

In rare variant association analysis, in order to achieve high statistical power of tests, usually a large sample with high sequencing costs is needed. Thus less costly sequencing design is preferred in rare variant association study. For quantitative traits, extreme phenotypes are generally considered to be more informative because of rare causal variants enriched among them. One can use a smaller sample size for extreme-phenotype sampling to achieve similar power as that for random sampling [23, 24]. Moreover, because extreme phenotypes of quantitative traits relative to human health are of primary clinical significance and thus data set can be obtained easily for subjects with extreme phenotypic values, using extreme phenotype samples in association analysis will make our study useful and practical. Here we use KL-distance to construct the statistics T_{KL} to measure the difference between two probability distributions of rare variants in two extreme populations. Based on the principle that the larger T_{KL} value is, the more dissimilar two probability distributions of rare variants, the statistics T_{KL} can be used as a test statistic to quantify the magnitude of association between the variants and the quantitative trait in the first stage of association analysis. We found that the statistic T_{KL} here for association analysis has higher power than three existing statistics of the burden test, the SKAT, and the SKAT-O.

Moreover, whereas increasing the number of non-causal variants and the opposite effect variants result in decreasing severely the powers of the burden test, the SKAT, and the SKAT-O, non-causal variants and the opposite effect variants affect slightly on the T_{KL}. The T_{KL} has relatively stable power with small change range under various parameters set.

In the second stage of fine mapping, l_{KL} is essentially a measure of LD between the variant and the QTL. Although LD between rare variant and QTL maybe weak [24], the maximum value across all rare variants can be usually found to identify the causal variant (QTL). The measure l_{KL} here has a good performance of just dependence on the frequency of the causal variant. In practice, not dependence on the frequency of the non-causal variant can eliminate"noise" and even bias introduced by varying frequencies of non-causal variants. In our early works, we proposed the LD measure l for mapping common variant of the QTL [27]. The performance of the measure l for mapping rare variant is unknown. We found from theory analysis that the two LD measures l_{KL} and l are parallel and have the same performance, that is, both of them can quantify LD between the variant and the QTL and do not depend on frequencies of non-causal variants. The difference between them is that the measure l_{KL} here is based on KL-distance and the measure l is based on entropy theory. Another LD measure for fine mapping is p_{excess} [28]. The p_{excess} is originally developed for fine mapping common variant of qualitative trait. We compare the performance of these three LD measures for fine mapping rare variant of quantitative traits using extreme samples. We found from theory analysis and simulation study that l_{KL} is superior to p_{excess}.

It is noted that, in practice, we do not know how many causal variants there are in the region established through association analysis at first stage. Although we considered a region having only a single causal variant, our method works for the general case with a region consisting of multiple causal variants. In fact, when there is a region linked to a quantitative trait has multiple causal variants, we can detect all causal variants using following steps: (1) l_{KL} is used to mapping a causal variant with the maximum value of l_{KL}; (2) T_{KL} is used to do association analysis for all variants except the causal variant detected in (1). If the association analysis result is positive, then return to (1). All causal variants will be found when the association analysis result is negative. It should be noted that we use the permutation procedure to assess the statistical significance of the statistic T_{KL} for association analysis. Permutation procedure may need more computing time to conduct simulation. But with the development of high-performance computing, computing time may not be a problem in our study. In addition, it can be seen that our method involves only rare variants. A phenotype may affected by common variants or both common variants and rare variants. So our further work will involve extensive field for common variants or both common variants and rare variants.

Conclusions

The statistic T_{KL} is affected slightly by non-causal variants and the opposite effect variants. The power of the T_{KL} for association analysis of rare variants increases with the stringency of the sample selection for quantitative traits. Extreme phenotypes allow T_{KL} to achieve higher power than three commonly used methods. The LD measure l_{KL} for fine mapping is independent of the frequencies of non-causal variants and just dependent on the frequencies of causal variants.

Methods

In this study, all datasets were publically available and no research requiring ethics approval was conducted.

We consider an interesting gene region with k bialleic variants and assume that each variant has a minor allele m with the MAF P_m and a normal allele M with the allele frequency P_M ($P_m + P_M = 1$). The variants are indexed by i ($i = 1, ..., k$). The index i may or may not correspond to the variant orders. Let X_i be minor allele count at ith variant carried by a subject. Assume that there is a quantitative trait Y: $Y = \beta_0 + G + \varepsilon$, where $G = \sum_{i=1}^{k} \beta_i X_i$, β_0 is the mean baseline value, and ε is residual due to random environmental effects. Without loss of generality, we assume $\beta_0 = 0$ and $\varepsilon \sim N(0, \sigma^2)$. To simplify our presentation, we use a measure with a superscript "U" to indicate a measure in the upper extreme population that has phenotypic values of the quantitative trait $Y > U$ (U is an upper-threshold value, chosen from the continuous distribution of the study quantitative trait). We also use a measure with a superscript "L" to indicate a measure in the low extreme population that has phenotypic values of the quantitative trait $Y < L$ (L is an low-threshold value, chosen from the continuous distribution of the study quantitative trait). Assume N^U and N^L subjects are sequenced with k variants in the upper extreme population and in the low extreme population, respectively, which are indexed by j ($j = 1,..., N^U/ N^L$). Denote X_{ij}^U and X_{ij}^L as the number of copies "m" for jth subject at ith

variant in the upper extreme population and in the low extreme population, respectively. Then the frequencies of P_m and P_M at ith variant in the upper extreme population and in the low extreme population, denoted as p^{U}_{mi}, p^{U}_{Mi}, and p^{L}_{mi}, p^{L}_{Mi}, respectively, are estimated as follows:

$p^{\mathrm{U}}_{mi} = \frac{\sum_{j=1}^{N^{\mathrm{U}}} X^{\mathrm{U}}_{ij}}{2N^{\mathrm{U}}}$, $p^{\mathrm{U}}_{Mi} = 1 - p^{\mathrm{U}}_{mi}$, $p^{\mathrm{L}}_{mi} = \frac{\sum_{j=1}^{N^{\mathrm{L}}} X^{\mathrm{L}}_{ij}}{2N^{\mathrm{L}}}$, and $p^{\mathrm{L}}_{Mi} = 1 - p^{\mathrm{L}}_{mi}$.

A statistic test for association analysis in the first stage

In the first stage, we propose a statistic test for association analysis. We define a k-dimensional random vector $\tilde{p}_m = (\tilde{p}_{m1}, \cdots, \tilde{p}_{mk})^T$ as the proportion of the minor allele m among all k variants, where $\tilde{p}_{mi} = \frac{\Sigma_j X_{ij}}{\sum_{i=1}^{k} \Sigma_j X_{ij}}$ and X_{ij} is the number of copies "m" for jth subject at ith variant. In the upper extreme population and in the low extreme population, the k-dimensional random vector of the proportion of the minor allele m are denoted as $\tilde{p}^{\mathrm{U}}_m = (\tilde{p}^{\mathrm{U}}_{m1}, \cdots, \tilde{p}^{\mathrm{U}}_{mk})^T$ and $\tilde{p}^{\mathrm{L}}_m = (\tilde{p}^{\mathrm{L}}_{m1}, \cdots, \tilde{p}^{\mathrm{L}}_{mk})^T$, respectively, where $\tilde{p}^{\mathrm{U}}_{mi} = \frac{\sum_{j=1}^{N^{\mathrm{U}}} X^{\mathrm{U}}_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{N^{\mathrm{U}}} X^{\mathrm{U}}_{ij}}$ and $\tilde{p}^{\mathrm{L}}_{mi} = \frac{\sum_{j=1}^{N^{\mathrm{L}}} X^{\mathrm{L}}_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{N^{\mathrm{L}}} X^{\mathrm{L}}_{ij}}$ ($i = 1, 2, \cdots, k$). We compare the two probability distributions $\tilde{p}^{\mathrm{U}}_m$ and $\tilde{p}^{\mathrm{L}}_m$ using the KL-distance which is defined as in Kullback & Leibler [26], here, we denote it the statistic $\mathrm{T_{KL}}$:

$$\begin{aligned}\mathrm{T}_{KL} &= H\left(\tilde{p}^{\mathrm{U}}_m, \tilde{p}^{\mathrm{L}}_m\right) \\ &= \frac{1}{2}\left(\sum_{i=1}^{k} \tilde{p}^{\mathrm{U}}_{mi} \cdot \log\frac{\tilde{p}^{\mathrm{U}}_{mi}}{\tilde{p}^{\mathrm{L}}_{mi}} + \sum_{i=1}^{k} \tilde{p}^{\mathrm{L}}_{mi} \cdot \log\frac{\tilde{p}^{\mathrm{L}}_{mi}}{\tilde{p}^{\mathrm{U}}_{mi}}\right) \end{aligned} \quad (1)$$

It is easy to find the relationship between $\mathrm{T_{KL}}$ and the frequencies of P_m and P_M as follows:

$$\mathrm{T_{KL}} = \frac{1}{2}\left(\sum_{i=1}^{k} \frac{p^{\mathrm{U}}_{mi}}{\sum_{i=1}^{k} p^{\mathrm{U}}_{mi}} \cdot \log\left(\frac{p^{\mathrm{U}}_{mi}}{p^{\mathrm{L}}_{mi}} \cdot \frac{\sum_{i=1}^{k} p^{\mathrm{L}}_{mi}}{\sum_{i=1}^{k} p^{\mathrm{U}}_{mi}}\right) + \sum_{i=1}^{k} \frac{p^{\mathrm{U}}_{mi}}{\sum_{i=1}^{k} p^{\mathrm{U}}_{mi}} \cdot \log\left(\frac{p^{\mathrm{L}}_{mi}}{p^{\mathrm{U}}_{mi}} \cdot \frac{\sum_{i=1}^{k} p^{\mathrm{U}}_{mi}}{\sum_{i=1}^{k} p^{\mathrm{L}}_{mi}}\right)\right) \quad (2)$$

$\mathrm{T_{KL}}$ is the mean between two KL-distances where one is the KL-distance between $\tilde{p}^{\mathrm{U}}_m$ and $\tilde{p}^{\mathrm{L}}_m$ and the other is the KL-distance between $\tilde{p}^{\mathrm{L}}_m$ and $\tilde{p}^{\mathrm{U}}_m$. KL-distance provides a non-symmetric measure of how big of the difference between two probability distributions are. The KL-distance is always non-negative and equal to 0 only if two distributions are identical. It can be seen that $\mathrm{T_{KL}}$ is a non-negative and symmetric measure of the two probability distributions $\tilde{p}^{\mathrm{U}}_m$ and $\tilde{p}^{\mathrm{L}}_m$. So, $\mathrm{T_{KL}}$ can be used as a statistic to quantify the magnitude of association between the variants and the quantitative trait: a much larger $\mathrm{T_{KL}}$ value will be observed under the alternative hypothesis of association compared to that under the null hypothesis of no association.

A KL-distance index for fine mapping of QTL in the second stage

Assume that a region linked to a quantitative trait has already been established through association analysis at first stage. In order to simplify our presentation, we assume that this region contains only a causal variant with a minor allele a (with frequency p_a) and a normal allele A (with frequency $p_A = 1 - p_a$), here, we call it the quantitative trait locus (QTL). We consider the quantitative trait $\mathrm{Y} = G_Q + \varepsilon$, where G_Q is the genotypic value at the QTL and $\varepsilon \sim N(0, \sigma^2)$. We hope to fine map this region by calculating the linkage disequilibrium (LD) measure between the QTL and a variant. We still use KL-distance to construct this measure through comparing the probability distributions of allele m and M at a variant in the upper extreme population and in the low extreme population. Following the previous symbols, let P_m and P_M be the frequencies of allele m and M at a variant. From Eq. (1), we have

$$\begin{aligned}&H\left(\{p^{\mathrm{U}}_m, p^{\mathrm{U}}_M\}, \{p^{\mathrm{L}}_m, p^{\mathrm{L}}_M\}\right) \\ &= \frac{1}{2}\left(p^{\mathrm{U}}_m \cdot \log\frac{p^{\mathrm{U}}_m}{p^{\mathrm{L}}_m} + p^{\mathrm{U}}_M \cdot \log\frac{p^{\mathrm{U}}_M}{p^{\mathrm{L}}_M} + p^{\mathrm{L}}_m \cdot \log\frac{p^{\mathrm{L}}_m}{p^{\mathrm{U}}_m} + p^{\mathrm{L}}_M \cdot \log\frac{p^{\mathrm{L}}_M}{p^{\mathrm{U}}_M}\right)\end{aligned} \quad (3)$$

From Appendix, $H(\{p^{\mathrm{U}}_m, p^{\mathrm{U}}_M\}, \{p^{\mathrm{L}}_m, p^{\mathrm{L}}_M\})$ can be asymptotically expressed as a function of LD (δ_{am}) between the QTL and the variant:

$$H\left(\{p^{\mathrm{U}}_m, p^{\mathrm{U}}_M\}, \{p^{\mathrm{L}}_m, p^{\mathrm{L}}_M\}\right) \approx \frac{\delta^2_{ma} \cdot (b_U - b_L)^2}{2p_m \cdot p_M} \quad (4)$$

Assume that there is an initial complete association between the variant allele m and the QTL allele a, at the 0th generation when the allele a is initially introduced into the study population. Let $\delta^{(0)}_{ma}$ be the initial complete LD between the allele a and m at the 0th generation, $\delta^{(0)}_{ma} = p_M \cdot p_a$. After n generations, the LD between the allele m and a is $\delta^{(n)}_{ma} = (1-\theta)^n \delta^{(0)}_{ma} = (1-\theta)^n \cdot p_M \cdot p_a$ [29], where, θ is the recombination between the QTL and the variant. Then we have

$$H\left(\left\{p_m^{\mathrm{U}}, p_M^{\mathrm{U}}\right\}, \left\{p_m^{\mathrm{L}}, p_M^{\mathrm{L}}\right\}\right) \approx \frac{\delta_{am}^2 \cdot (b_U - b_L)^2}{2p_M \cdot p_m} = \frac{(1-\theta)^{2n} p_a^2 \cdot p_M^2 \cdot (b_U - b_L)^2}{2p_M \cdot p_m} \quad (5)$$

Now we define a LD measure, here, we denote it as l_{KL}, as follows:

$$l_{KL} = \frac{p_m}{p_M} H\left(\left\{{p_m}^{\mathrm{U}}, {p_M}^{\mathrm{U}}\right\}, \left\{{p_m}^{\mathrm{L}}, {p_M}^{\mathrm{L}}\right\}\right) \approx \frac{1}{2}(1-\theta)^{2n} p_a^2 \cdot b^2 \quad (6)$$

where $b = b_U - b_L$. It can be seen that l_{KL} is a decreasing function of the recombination θ and reaches its maximum at $\theta = 0$. So we can use l_{KL} to find the variant closest to the QTL and thus fine map the QTL. Notice from Eq. (6) that l_{KL} is independent of the frequency of the variant, just only dependent on the frequency of the QTL.

Simulation

Simulation for association analysis

To evaluate the performance of the test statistic T_{KL}, we perform a series of simulation studies. We consider k = 100 variants with MAF values of causal variants determined by a uniform distribution U (0.001, 0.01) and MAF values of non-causal variants determined by a uniform distribution U (0.001, 0.05). The genotype data are simulated similar to those in Wang and Elston [30]. We first generate haplotypes for k variants based on a latent variable $Z = (Z_1, \cdots, Z_k)$ from a multivariate normal distribution with covariance structure $\mathrm{cov}(Z_i, Z_j) = 0.4^{|i-j|}$ between any two latent components. Then we combine two haplotypes to obtain the genotype value for each individual $X_i = (X_{i1}, \cdots, X_{ik})$. A phenotype Y under the null hypothesis of no association is generated using the model $Y = \varepsilon$ with $\varepsilon \sim N(0, 1)$ ($\beta_1 = \cdots = \beta_k = 0$). Under the alternative hypothesis of association, we randomly chose s variants as causal variants while other k-*s* variants as non-causal variants having $\beta_j = 0$. Here, we let s = 10, 20, 50 in which 10, 20%, or 50% of rare variants were causal. For causal variants under the alternative hypothesis, we set $\beta_j = c \cdot \log_{10}(p_{mi})$ as used in Lee et al. [10], where c is 0.6, 0.3, and 0.2 for different values of s and different direction of the effects of causal variants. We consider 9 scenarios in the simulation study with the parameter values detailed in Table 3. We conduct 1000 simulations for each scenario. In each simulation, we select three extreme sample strategies, the low 20% and the up 20%, the low 10% and the up 10%, and, the low 5% and the up 5%, each of which consists of 2 N individuals including N individuals in an upper sample and N individuals in a lower sample. The statistical significance is assessed by a permutation procedure. We first calculate the value of the data-based statistic T_{KL} for each simulation. Then we permute the "upper sample" and "lower sample" labels with equal probability and recalculate the statistic T_{KL} for 1000 times. The estimated P value is then the proportion of permutation-based statistics that are larger than the data-based statistic in 1000 permutations for each simulation. For a given significance level α, the power/type I error rate is estimated as the proportion of rejecting the null hypothesis when p-value ≤ α in 1000 simulations. In order to compare the performance of the test statistic T_{KL} with the existing methods, we also obtain the power for the burden, SKAT, and SKAT-O tests using case-control design with the same samples as for the test statistic T_{KL}.

Table 3 The parameter values for power study

Scenario	causal variants (s)	Effect size weights (c)	Positive direction: Negative direction
1	s = 10	c = 0.6	10:0
2	s = 20	c = 0.3	20:0
3	s = 50	c = 0.2	50:0
4	s = 10	c = 0.6	8:2
5	s = 20	c = 0.3	16:4
6	s = 50	c = 0.2	40:10
7	s = 10	c = 0.6	5:5
8	s = 20	c = 0.3	10:10
9	s = 50	c = 0.2	25:25

Simulation for fine mapping

To assess the performance of the LD measure l_{KL} in fine mapping rare causal variants of quantitative traits, we conduct a simulation study using the method similar to those described in our early work [27, 28]. We consider a genetic region that has 21 variants, where only a variant locating at the middle of variant 10 and variant 11 is causal variant (that is, the QTL). The MAF of the causal variant is set to be 0.01(p_a =0.01) and the MAFs for 20 other variants are uniformly determined with values ranging from 0.001 to 0.05. Other parameters in simulation include the ratio d/v (here, v and d are the genotypic values for individuals with genotypes *aa* and *Aa*, respectively), the thresholds L and U, the heritability (h^2) of the causal variant, and the sample size (2 N) [31]. We let the ratio d/v be − 1, 0, and 1 which correspond to recessive, additive, and dominant models, respectively. Once the parameter values are chosen, a population with the effective size of 15,000 is simulated starting from the 0th generation, with an initial complete association between the minor allele a at the causal variant and m at other variants [$P(m|a) = 1$]. The population then evolved for 50 generations under random mating and genetic drift. A hundred populations are simulated for analyses.

Appendix

Firstly, we calculate $p_m^{\mathrm{U}} \cdot \log\frac{p_m^{\mathrm{U}}}{p_m^{\mathrm{L}}}$. Under the assumption of random mating and, thus, Hardy-Weinberg (HW) equilibrium holding in the population, we can get.

$p_m^{\mathrm{L}} = pr(m|Y<L) = pr(m,Y<L)/\phi_L$, where $\phi_L = pr(Y<L)$

$$\begin{aligned} pr(m,Y<L) &= pr(m,aa,Y<L) + pr(m,aA,s<T) \\ &+ pr(m,AA,Y<L) = p_{ma}\cdot p_a \cdot pr(Y<L|aa) \\ &+ (p_{ma}\cdot p_A + p_{mA}\cdot p_a)\cdot pr(Y<L|aA) \\ &+ p_{mA}\cdot p_A \cdot pr(Y<L|AA) = p_{ma}\cdot p_a\cdot \phi_{11} \\ &+ (p_{ma}\cdot p_A + p_{mA}\cdot p_a)\cdot \phi_{12} + p_{mA}\cdot p_A \cdot \phi_{22} \\ &= a_1\cdot p_{ma} + a_2 \cdot p_{mA} \end{aligned}$$

where $\phi_{11} = pr(Y<L|\,aa)$, $\phi_{12} = pr(Y<L|\,aA)$, $\phi_{22} = pr(Y<L|\,AA)$, $a_1 = (\phi_{11}\cdot p_a + \phi_{12}\cdot p_A)/\phi_L$, $a_2 = (\phi_{22}\cdot p_A + \phi_{12}\cdot p_a)/\phi_L$.

Note that.

$p_{ma} = p_m\cdot p_a + \delta_{ma}$, $p_{mA} = p_m\cdot p_A + \delta_{mA}$, $\delta_{ma} = -\delta_{mA}$ *and* $a_1\cdot p_a + a_2\cdot p_A = 1$, here $\delta_{m\sim}$ is the measure of LD between variant allele m and the QTL allele $\sim$ and is defined as $\delta_{\sim m} = p_{\sim m} - p_{\sim}\cdot p_m$, where $p_{m\sim}$ is the frequency of haplotype $m\sim$.

Then,

$$\begin{aligned} p_m^L &= a_1\cdot(p_m\cdot p_a + \delta_{ma}) + a_2\cdot(p_m\cdot p_A + \delta_{mA}) \\ &= p_m + \delta_{ma}(a_1 - a_2) = p_m + b_L\cdot\delta_{ma} \\ b_L &= a_1 - a_2 \qquad \text{(Where)} \end{aligned}$$

Similarly, we can get $p_M^L = p_M + b_L\cdot\delta_{Ma}$, $p_m^{LI} = p_m + b_U\cdot\delta_{ma}$ and $p_M^{LI} = p_M + b_U\cdot\delta_{Ma}$, where $b_U = c_1 - c_2$, $c_1 = (\gamma_{11}\cdot p_a + \gamma_{12}\cdot p_A)/\phi_{LI}$, $c_2 = (\gamma_{22}\cdot p_A + \gamma_{12}\cdot p_a)/\phi_{LI}$, $\phi_{LI} = pr(Y>LI)$,

$\gamma_{11} = pr(Y>LI|\,aa)$, $\gamma_{12} = pr(Y>LI|\,aA)$, $\gamma_{22} = pr(Y>LI|\,AA)$, $c_1 = (\gamma_{11}\cdot p_a + \gamma_{12}\cdot p_A)/\phi_{LI}$, $c_2 = (\gamma_{22}\cdot p_A + \gamma_{12}\cdot p_a)/\phi_{LI}$.

Then

$$\begin{aligned} p_m^{\mathrm{U}}\cdot\log\frac{p_m^{\mathrm{U}}}{p_m^{\mathrm{L}}} &= p_m^{\mathrm{U}}\cdot\log p_m^{\mathrm{U}} - p_m^{\mathrm{U}}\cdot\log p_m^{\mathrm{L}} \\ &= (p_m + b_U\cdot\delta_{ma})\cdot\log\frac{p_m\cdot(1+b_U\cdot\delta_{ma}/p_m)}{p_m\cdot(1+b_L\cdot\delta_{ma}/p_m)} \\ &= (p_m + b_U\cdot\delta_{ma})\cdot[\log(1+b_U\cdot\delta_{ma}/p_m) \\ &- \log(1+b_L\cdot\delta_{ma}/p_m)]\,[by\ \log(1+x)\approx x - x^2/2] \\ &\approx (p_m + b_U\cdot\delta_{ma})\cdot\left[\left(\frac{b_U\cdot\delta_{ma}}{p_m} - \frac{{b_U}^2\cdot{\delta_{ma}}^2}{2{p_m}^2}\right)\right. \\ &\left. - \left(\frac{b_L\cdot\delta_{ma}}{p_m} - \frac{{b_L}^2\cdot{\delta_{ma}}^2}{2{p_m}^2}\right)\right] \\ &= (p_m + b_U\cdot\delta_{ma})\cdot\frac{(b_U - b_L)\cdot\delta_{ma}}{p_m}\cdot\left(1 - \frac{(b_U+b_L)\cdot\delta_{ma}}{2p_m}\right) \end{aligned}$$

Similar

$$p_M^{\mathrm{U}}\cdot\log\frac{p_M^{\mathrm{U}}}{p_M^{\mathrm{L}}} \approx (p_M + b_U\cdot\delta_{Ma})\cdot\frac{(b_U - b_L)\cdot\delta_{Ma}}{p_M}\cdot\left(1 - \frac{(b_U+b_L)\cdot\delta_{Ma}}{2p_M}\right)$$

$$p_m^{\mathrm{L}}\cdot\log\frac{p_m^{\mathrm{L}}}{p_m^{\mathrm{U}}} \approx (p_m + b_L\cdot\delta_{ma})\cdot\frac{(b_L - b_U)\cdot\delta_{ma}}{p_m}\cdot\left(1 - \frac{(b_L+b_U)\cdot\delta_{ma}}{2p_m}\right)$$

$$p_M^{\mathrm{L}}\cdot\log\frac{p_M^{\mathrm{L}}}{p_M^{\mathrm{U}}} \approx (p_M + b_L\cdot\delta_{Ma})\cdot\frac{(b_L - b_U)\cdot\delta_{Ma}}{p_M}\cdot\left(1 - \frac{(b_L+b_U)\cdot\delta_{Ma}}{2p_M}\right)$$

Then

$$\begin{aligned} H(\{p_m^{\mathrm{U}},p_M^{\mathrm{U}}\},\{p_m^{\mathrm{L}},p_M^{\mathrm{L}}\}) &= \frac{1}{2}(p_m^{\mathrm{U}}\cdot\log\frac{p_m^{\mathrm{U}}}{p_m^{\mathrm{L}}} + p_M^{\mathrm{U}}\cdot\log\frac{p_M^{\mathrm{U}}}{p_M^{\mathrm{L}}} \\ &+ p_m^{\mathrm{L}}\cdot\log\frac{p_m^{\mathrm{L}}}{p_m^{\mathrm{U}}} + p_M^{\mathrm{L}}\cdot\log\frac{p_M^{\mathrm{L}}}{p_M^{\mathrm{U}}}) \\ &\approx \frac{1}{2}\left[\frac{(b_U - b_L)^2\cdot{\delta_{ma}}^2}{p_m} + \frac{(b_U - b_L)^2\cdot{\delta_{Ma}}^2}{p_M}\right] \\ &= \frac{{\delta_{ma}}^2\cdot(b_U - b_L)^2}{2p_m\cdot p_M} \end{aligned}$$

Assume that there is an initial complete association between the variant allele m and the QTL allele a, at the 0th generation when the allele a is initially introduced into the study population. Let $\delta_{ma}^{(0)}$ be the initial complete LD between the allele a and m at the 0th generation, $\delta_{ma}^{(0)} = p_M\cdot p_a$. After n generations, the LD between the allele m and a is $\delta_{ma}^{(n)} = (1-\theta)^n\delta_{ma}^{(0)} = (1-\theta)^n\cdot p_M\cdot p_a$ [29], where, θ is the recombination between the QTL and the variant. Then we have

$$\begin{aligned} H(\{p_m^{\mathrm{U}},p_M^{\mathrm{U}}\},\{p_m^{\mathrm{L}},p_M^{\mathrm{L}}\}) &\approx \frac{{\delta_{ma}}^2\cdot(b_U - b_L)^2}{2p_m\cdot p_M} \\ &= \frac{(1-\theta)^{2n}p_a^2\cdot p_M^2\cdot(b_U - b_L)^2}{2p_m\cdot p_M} \end{aligned}$$

Then, we have

$$l_{KL} = \frac{p_m}{p_M}H(\{{p_m}^{\mathrm{U}},{p_M}^{\mathrm{U}}\},\{{p_m}^{\mathrm{L}},{p_M}^{\mathrm{L}}\}) \approx \frac{1}{2}(1-\theta)^{2n}p_a^2\cdot b^2$$

where $b = b_U - b_L$.

Abbreviations

LD: Linkage disequilibrium; RV: Rare variant; KL-distance: Kullback-Leibler distance; SKAT: Sequence kernel association test; SKAT-O: The optimal test that combines SKAT and the burden test; QTL: Quantitative trait locus; HW: Hardy-Weinberg

Acknowledgments

This work was supported by the Foundation of Hunan Double First-rate Discipline Construction Projects of Bioengineering.

Authors' contributions

XY developed the statistical method and wrote the manuscript. XXR developed the statistical method and revised the manuscript. LYM conceived the idea, designed the study, and revised the manuscript. All authors have read and approved the final version of the manuscript.

Funding

Not applicable.

Availability of data and materials
All data generated or analyzed during this study are included in this published article.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Mathematics and Computational Science, Huaihua University, Huaihua, Hunan 418008, People's Republic of China. [2]Key Laboratory of Research and Utilization of Ethnomedicinal Plant Resources of Hunan Province, Huaihua University, Huaihua 418008, China. [3]Key Laboratory of Hunan Higher Education for Western Hunan Medicinal Plant and Ethnobotany, Huaihua University, Huaihua 418008, China. [4]School of Mathematics and Statistics, Hunan Normal University, Changsha, Hunan 410081, People's Republic of China.

Received: 10 July 2020 Accepted: 10 November 2020
Published online: 24 November 2020

References

1. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010;11(6):415–25.
2. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90(1):7–24.
3. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hin-dorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–53.
4. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-cod-ing variants. Nature. 2013;493(7431):216–20.
5. Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. 2012;337(6090):100–4.
6. Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet. 2010;11(11):773–85.
7. Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, et al. Exome sequencing and the genetic basis of complex traits. Nat Genet. 2012;44(6):623–30.
8. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95(1):5–23.
9. Li Z, Li X, Liu Y, Shen J, Chen H, Zhou H, et al. Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. Am J Hum Genet. 2019;104(5):802–14.
10. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012;13(4):762–75.
11. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res. 2007;615(1):28–56.
12. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008;83(3):311–21.
13. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009;5(2):e1000384.
14. Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol. 2009;33(6):497–507.
15. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an unusual distribution of rare variants. PLoS Genet. 2011;7(3):e1001322.
16. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful SNP-setanalysis for case-control genome-wide association studies. Am J Hum Genet. 2010;86(6):929–42.
17. Wu MC, Lee S, Cai T, Li Y, Boehnke MC, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89(1):82–93.
18. Bacanu SA, Nelson MR, Whittaker JC. Comparison of methods and sampling designs to test for association between rare variants and quantitative traits. Genet Epidemiol. 2011;35(4):226–35.
19. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, et al. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010;86(6):832–8.
20. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol. 2010;34(2):188–93.
21. Luo L, Zhu Y, Xiong M. Quantitative trait locus (QTL) analysis for next-generation sequencing with the functional linear models. J Med Genet. 2012;49(8):513–24.
22. Guey LT, Kravic J, Melander O, Burtt NP, Laramie JM, Lyssenko V, et al. Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. Genet Epidemiol. 2011;35(4):236–46.
23. Barnett IJ, Lee S, Lin X. Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. Genet Epidemiol. 2013;37(2):142–51.
24. Li D, Lewinger JP, Gauderman WJ, Murcray CE, Conti D. Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. Genet Epidemiol. 2011;35(8):790–9.
25. Emond MJ, Louie T, Emerson J, Zhao W, Mathias RA, Knowles MR, et al. Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. Nat Genet. 2012;44(8):886–9.
26. Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat. 1951;22(1):79–86.
27. Deng HW, Chen WM, Recker RR. QTL fine mapping by measuring and test for hardy-Weinberg and linkage disequilibrium at a series of linked marker loci in extreme samples of populations. Am J Hum Genet. 2000;66(3):1027–45.
28. Li YM, Xiang Y, Sun ZQ. An entropy-based measure for QTL mapping using extreme samples of population. Hum Hered. 2008;65(3):121–8.
29. Hartl DL. A primer of population genetics, Sinauer, Sunderland, Massachusetts, 3rd; 1999.
30. Wang T, Elston RC. Improved power by use of a weighted score test for linkage disequilibrium mapping. Am J Hum Genet. 2007;80(2):353–60.
31. Lehesjoki AE, Koskiniemi M, Norio R, Tirrito S, Sistonen P, Lander E, et al. Localization of the EPM1 gene for progressive myoclonus epilepsy on chromosome 21: linkage disequilibrium allows high resolution mapping. Hum Mol Genet. 1993;2(8):1229–34.

Publisher's Note