PLOS | ONE

RESEARCH ARTICLE

Adaptive Combination of *P*-Values for Family-Based Association Testing with Sequence Data

Wan-Yu Lin*

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

*linwy@ntu.edu.tw

Abstract

Family-based study design will play a key role in identifying rare causal variants, because rare causal variants can be enriched in families with multiple affected subjects. Furthermore, different from population-based studies, family studies are robust to bias induced by population substructure. It is well known that rare causal variants are difficult to detect from single-locus tests. Therefore, burden tests and non-burden tests have been developed, by combining signals of multiple variants in a chromosomal region or a functional unit. This inevitably incorporates some neutral variants into the test statistics, which can dilute the power of statistical methods. To guard against the noise caused by neutral variants, we here propose an 'adaptive combination of *P*-values method' (abbreviated as '*ADA*'). This method combines per-site *P*-values of variants that are more likely to be causal. Variants with large *P*-values (which are more likely to be neutral variants) are discarded from the combined statistic. In addition to performing extensive simulation studies, we applied these tests to the Genetic Analysis Workshop 17 data sets, where real sequence data were generated according to the 1000 Genomes Project. Compared with some existing methods, *ADA* is more robust to the inclusion of neutral variants. This is a merit especially when dichotomous traits are analyzed. However, there are some limitations for *ADA*. First, it is more computationally intensive. Second, pedigree structures and founders' sequence data are required for the permutation procedure. Third, unrelated controls cannot be included. We here show that, for family-based studies, the application of *ADA* is limited to dichotomous trait analyses with full pedigree information.

Citation: Lin W-Y (2014) Adaptive Combination of *P*-Values for Family-Based Association Testing with Sequence Data. PLoS ONE 9(12): e115971. doi:10.1371/journal.pone.0115971

Editor: Yun Li, University of North Carolina, United States of America

Received: June 19, 2014

Accepted: December 1, 2014

Published: December 26, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by grants 102-2628-B-002-039-MY3 from the Ministry of Science and Technology of Taiwan, 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.

Introduction

Studies in genetic epidemiology are important to uncover the genetic architecture of complex human diseases. The development of next-generation sequencing technologies has allowed for the mapping of all genetic variants across the human genome. With this, we can search for rare causal variants (minor allele frequency (MAF) <1%), which are mostly not genotyped in genome-wide association studies (GWAS) but are related to the etiology of complex diseases. Till now, many statistical methods have been proposed for rare variant association testing. Most of them were designed for population-based studies where unrelated cases and controls were recruited and analyzed [1–22].

Despite a variety of statistical methods, there are two concerns in population-based rare variant association studies. First, population stratification may cause false-positive results. This issue was tackled since the era of genome-wide association studies (GWAS). In GWAS where most genotyped variants were common, principal component analysis (PCA) [23] and mixed models [24] were proposed as effective methods to deal with population stratification. However, studies of population stratification are still limited for next-generation sequencing data [25]. Existing methods, such as PCA and mixed models, can fail to correct for rare variant stratification [26]. Second, rare causal variants are difficult to observe in general populations, and therefore statistical methods are usually under-powered [27]. Although burden tests [2–5, 27] and non-burden tests [7–9] have been proposed to aggregate signals of multiple variants, searching for rare causal variants remains challenging.

Family-based study design will play a key role in identifying rare causal variants, because rare causal variants can be enriched in families with multiple affected subjects [28, 29]. Burden tests (such as the weighted sum approach (*WS*) [3], the cumulative minor allele test (*CMAT*) [22]), and non-burden tests (such as the sequence kernel association test (*SKAT*) [7, 8]) have been extended to family-based designs by incorporating within-family correlation structures into the statistics [30–40]. For continuous traits, Chen *et al.* [31] has proposed "*famSKAT*" that can account for members' relationships within families. This method was essentially equivalent to the method proposed by Schifano *et al.* [38] and the adjusted sequence kernel association test (abbreviated as "*ASKAT*") proposed by Oualkacha *et al.* [40], although Schifano *et al.*'s method was not originally designed for rare variant association testing. Svishcheva *et al.* [39] then developed a fast family-based *SKAT* (abbreviated as "*FFBSKAT*") that was shown to be the fastest method to perform the kernel-based association tests for continuous traits. Svishcheva *et al.* have shown a pure coincidence of the *P*-values calculated by the *famSKAT*, *ASKAT*, and the *FFBSKAT* software programs [39].

Testing for effects of rare variants individually is known to be underpowered. To strengthen association signals, both the burden tests and the non-burden tests combine information of multiple rare variants in a gene/region. This inevitably incorporates many neutral variants into the test statistics. Adaptive combination of *P*-values method (abbreviated as '*ADA*') has been shown to outperform the burden tests (e.g., *WS* and the variable threshold approach [5]) and the non-burden tests (e.g., *SKAT*) in rare variant association testing for unrelated subjects, because *ADA* is more robust to the inclusion of neutral variants [13, 21]. Taking this advantage, we here extend the *ADA* method to deal with pedigree data, and compare its power performance with that of the burden test [32], the kernel statistic [32], and the *FFBSKAT* method [39] (*famSKAT* [31] and *ASKAT* [40] are essentially equivalent to *FFBSKAT*). We also apply the method to the Genetic Analysis Workshop 17 (GAW 17) data [27, 41]. Some family-based association testing methods were designed for trio data (or trios plus unrelated controls), such as the rare-variant extensions of the transmission disequilibrium test (rvTDT) [36, 42]. Therefore, these methods are not compared here.

Materials and Methods

Let Y_i be the trait value of the ith subject ($i=1,\cdots,n$). Suppose there are L loci in the chromosomal region of interest, and let g_{il} be the genotype score at the lth locus of the ith subject ($i=1,\cdots,n$, $l=1,\cdots,L$). Under the assumption of additive genetic model, g_{il} is the number of minor alleles, i.e., 0, 1, or 2. The statistic to test for the association between the trait and the lth marker is

$$T_l=\frac{\left[\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)'\boldsymbol{g}_{\boldsymbol{l}}\right]^2}{2MAF_l(1-MAF_l)\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)'\boldsymbol{\Omega}\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)}\sim\chi_1^2, \quad (1)$$

where $\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)$ is the vector of residuals after adjusting for covariates (e.g., age, gender), $\boldsymbol{g}_{\boldsymbol{l}}$ is the genotype score vector at the lth marker for the n subjects, MAF_l is the MAF of the lth marker calculated using founders, and $\boldsymbol{\Omega}$ is an $n\times n$ matrix of genetic correlations of these n subjects. For autosomes, the (i, j)th element of $\boldsymbol{\Omega}$ is $2\phi_{ij}$, where ϕ_{ij} is the kinship coefficient of the ith and the jth subjects. Because the kinship coefficient of subjects belonging to different pedigrees should be 0, $\boldsymbol{\Omega}$ is a block-diagonal matrix with block sizes as the sizes of pedigrees.

The test statistic in Equation (1) has an approximate χ^2 distribution with 1 degree of freedom. This statistic is essentially equivalent to the statistic proposed by Thornton and McPeek [43] (see Equation 1 in [43]). Phenotypes and genotypes are treated as fixed and random, respectively. This retrospective view allows us to correct the ascertainment bias when recruiting pedigrees through affected subjects. After performing L tests for the L markers, we have P-values $p_1,p_2,\cdots p_L$. Suppose we consider J candidate truncation thresholds, $\theta_1,\theta_2,\cdots,\theta_J$. Summarizing the L markers, the significance score under the jth truncation threshold is

$$S_j=-\sum_{l=1}^{L}I\left[p_l<\theta_j\right]\cdot\log p_l, \quad (2)$$

where $I[p_l < \theta_j]$ is an indicator variable coded as 1 if the lth marker has a P-value smaller than θ_j (the jth truncation threshold) and 0 otherwise. Throughout this study, the candidate truncation thresholds are specified as $\theta_1 = 0.10, \theta_2 = 0.11, \cdots, \theta_{11} = 0.20$, suggested by the *ADA* method for population-based studies [13, 21]. Using a wider range of P-value truncation thresholds, say, $\theta_1 = 0.05, \theta_2 = 0.06, \cdots, \theta_{21} = 0.25$, does not contribute a noticeable power gain to *ADA* (results not shown).

Because multiple P-value truncation thresholds are considered, to correct for multiple testing, statistical significance must be obtained with permutations. We first construct the distribution of the significance score S_j under the null hypothesis, where the transmission of haplotypes from parents to offspring is completely random, conditional on parental genotypes [44, 45]. For example, the family shown by Fig. 1 consists of 12 members. Conditional on the genotypes of No. 1 and No. 2, the probability that No. 3 has the observed/unobserved pattern of allelic transmission is $1/2$ under the null hypothesis. Given the founders' (Nos. 1, 2, 5, 6) genotypes, this family consists of more than $2^8 = 256$ possible patterns of allelic transmission. With a total of N families, there are at least 256^N different permutations of the genotype data.

The above permutation procedure was extended from that used for trio data [44, 45]. Unambiguous haplotype phases are not always required in the process. For example, No. 3 has a probability of $1/2$ to possess the observed pattern of allelic transmission, and then there is no need to change his genotypes. The probability of owning unobserved pattern of allelic transmission is also $1/2$. In this situation, No. 3's number of minor alleles at the lth locus (0, 1, or 2, representing three different genotype scores) is $g_{3l}^{(U)} = g_{1l} + g_{2l} - g_{3l}$, where g_{1l} and g_{2l} are the genotype scores at the lth locus of Nos. 1 and 2, respectively; g_{3l} is No. 3's original observed genotype score. Conditional on g_{3l} and g_{6l} belonging to Nos. 3 and 6, the unobserved pattern of allelic transmission for No. 7 is $g_{7l}^{(U)} = g_{3l} + g_{6l} - g_{7l}$. On the other hand, given $g_{3l}^{(U)}$ and g_{6l}, haplotype phases of Nos. 3 and 6 are required to determine the genotype scores of Nos. 7–9. An haplotype-phasing software package (such as Beagle [46]) is used to infer the most-likely haplotype pairs for Nos. 3 and 6. The genotype scores of Nos. 7–9 are then determined by randomly drawing one haplotype from No. 3 and one from No. 6.

Suppose we perform B permutations, say, B=1000. For the bth permutation, the significance score under the jth truncation threshold can be calculated with Eq. (2), denoted as $S_j^{(b)}$. The statistical significance of S_j is obtained by comparing it with $S_j^{(b)}$, $b = 1, \cdots, B$. The P-value of S_j is estimated as $\frac{\sum_{b=1}^{B} I\left(S_j^{(b)} \geq S_j\right) + 1}{B+1}$, for each truncation threshold (j=1, …, J), where $I\left[S_j^{(b)} \geq S_j\right]$ is an indicator variable coded as 1 if $S_j^{(b)} \geq S_j$ and 0 otherwise. Similarly, the P-value of $S_j^{(b')}$ for the b'th permutation is $\frac{\sum_{b \neq b'} I\left(S_j^{(b)} \geq S_j^{(b')}\right) + 1}{B}$, for j=1, …, J and $b' = 1, \cdots, B$. We

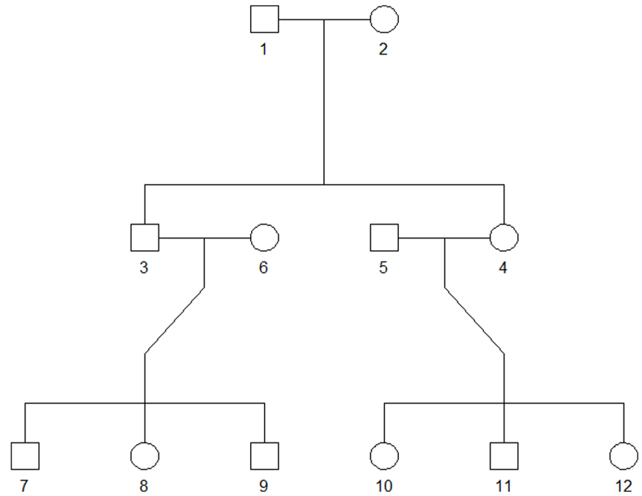

Fig. 1. The family structure simulated by SeqSIMLA.

doi:10.1371/journal.pone.0115971.g001

can then find the minimum P-values across the J candidate truncation thresholds for the observed sample and the bth permuted sample, denoted as $MinP$ and $MinP^{(b)}$, respectively. The "adjusted P-value" is estimated as $\frac{\sum_{b=1}^{B} I\left(MinP^{(b)} \leq MinP\right)+1}{B+1}$. This method is referred to as "ADA", because the optimal P-value truncation threshold is driven adaptively according to the data.

Simulation Study

We generated sequence data with the SeqSIMLA software [47], which was designed to simulate sequence data for family samples. SeqSIMLA used GENOME [48] as the default sequence generator that could efficiently simulate sequence data according to the standard coalescent model [49–52]. In this way, we aim to evaluate statistical methods with simulated data that can reflect realistic DNA sequences. In each simulation, 50, 80 or 100 three-generation families each with 12 members were generated. The family structure was shown by Fig. 1. For each subject, a chromosome region with m=50, 100, or 150 SNPs/SNVs (single-nucleotide polymorphisms/single-nucleotide variants) was simulated. Dichotomous traits and continuous traits were considered respectively.

When evaluating type-I error rates, no causal locus was specified. When evaluating power, five SNPs/SNVs (Nos. 10, 20, 30, 40, 50) were assumed to be causal. We did not restrict all causal variants to be rare/uncommon, because in reality causal variants could also be common. Let 'signal proportion' be the fraction of causal variants out of all variants. The signal proportion in our simulations was set at one of the three levels: 0.033 ($=5/150$), 0.05 ($=5/100$), 0.1 ($=5/50$).

When dichotomous traits were considered, the overall population attributable risk (PAR) for all causal loci was assumed to be 0.05, 0.10, 0.15, 0.20, 0.25, and 0.30, respectively. Therefore, the marginal PAR for each causal locus was 0.01, 0.02, 0.03, 0.04, 0.05, and 0.06, respectively. In the SeqSIMLA software [47], the genotype relative risk (GRR) of the jth causal SNP/SNV was:

$$GRR_j = 1 + \frac{PAR_j}{(1 - PAR_j) \cdot MAF_j}, \tag{3}$$

where PAR_j and MAF_j were the PAR and the population MAF of that SNP/SNV, respectively. S1 Fig. showed the distribution of GRRs of causal SNPs/SNVs when the overall PAR for all causal loci was assumed to be 0.05, 0.10, 0.15, 0.20, 0.25, and 0.30, respectively (the marginal PAR for each causal locus was 0.01, 0.02, 0.03, 0.04, 0.05, and 0.06, respectively). A variant with a smaller frequency was assumed to have a larger GRR, following the model in many previous contributions [3, 18–20, 47]. For a founder, SeqSIMLA randomly sampled two haplotypes $\{H_1, H_2\}$ from the population sequence pool created by GENOME [48]. According to the SeqSIMLA software [47], the disease status of this subject was determined by

$$P(affected|\{H_1,H_2\}) = \frac{f_0 \times \prod_{k=1}^{2} \prod_{j=1}^{5} GRR_j^{I\left(H_{k,j}=a_j\right)}}{1 - f_0 + f_0 \times \prod_{k=1}^{2} \prod_{j=1}^{5} GRR_j^{I\left(H_{k,j}=a_j\right)}}, \tag{4}$$

where f_0 was the baseline penetrance specified as 0.05, $H_{k,j}$ was the allele at the jth causal SNP/SNV on the haplotype H_k ($k=1,2$), and a_j was the minor allele at the jth causal SNP/SNV ($j=1,2,\cdots,5$) [47]. Given parental haplotypes, a child's haplotypes were formed by randomly selecting one from the father and one from the mother, and the child's disease status was again determined by Eq. (4). Based on this equation, a subject with no causal variant would have a probability of f_0 (baseline penetrance) to be diseased, while a subject with more causal variants would have a larger probability to be diseased.

When continuous traits were simulated, five SNPs/SNVs (Nos. 10, 20, 30, 40, 50) were assumed to be quantitative trait loci (QTLs). Causal variants were not restricted to be all rare or uncommon, because in reality common variants can be causal as well. Let Y_i be the trait value of the ith subject. It was determined by the model $Y_i = \mu + \sum_{l=1}^{5} G_{il} + e_i$, where μ was the overall mean of the trait, G_{il} was the genotypic value at the lth QTL of the ith subject ($i=1,\cdots,n$, $l=1,\cdots,5$) which followed a normal distribution (with a mean of μ_l, 0, or $-\mu_l$ for 2, 1, and 0 minor alleles at the lth QTL, respectively [47]), and e_i was the error term for the ith subject following a normal distribution as well. According to the default setting of the SeqSIMLA software [47], the genetic effects of QTLs were all assumed to be additive, and $Var(Y)$ and μ were both specified at 100. These two values were not critical because SeqSIMLA software [47] actually controlled the proportion of

$Var(Y)$ explained by each QTL (denoted as V_P). The value of V_P was assumed to be 0.001, 0.002, 0.003, 0.004, 0.005, and 0.006 for each QTL, respectively. The corresponding proportion of variance explained by all the five QTLs was therefore 0.005, 0.01, 0.015, 0.02, 0.025, or 0.03.

Tests under Comparison

We compared *ADA* with the broad classes of the burden test (referred to as "*Burden*") [32], the kernel test for family data (referred to as "*Kernel*") [32], and the *FFBSKAT* method [39]. *Burden* and *Kernel* were implemented with the R package "pedgene" [32]; *FFBSKAT* was performed with the package "FFBSKAT" (http://mga.bionet.nsc.ru/soft/FFBSKAT/) [39]. *FFBSKAT* was performed only when continuous traits were considered, because it could not analyze dichotomous traits. Following the default setting of *FFBSKAT* [39] and *Kernel* [32], the (*j*, *j*)th element of the diagonal weighting matrix $\boldsymbol{W}$ was set as $Beta\left(MAF_j; 1{,}25\right)$, where MAF_j was the MAF of the *j*th genetic variant. The *P*-values of *ADA* were obtained with 1,000 permutations.

Results

Type-I Error Rates

By setting the PAR or the proportion of variance explained by causal SNPs at exactly 0%, we evaluated type-I error rates by performing 10,000 replications. In each replication, 50 three-generation families each with 12 members (shown in Fig. 1) were generated. For each subject, a chromosome region containing 100 SNPs/SNVs was simulated. Table 1 shows that all the tests (three for dichotomous traits and four for continuous traits) are valid in the sense that their type-I error rates match the nominal significance levels.

Power Comparisons

When we evaluated power, 1000 replications were performed under each scenario. In total, there were 54,000 replications in power evaluation for dichotomous traits and continuous traits, respectively (three levels of m (50, 100, or 150) $\times$ three levels of family numbers (50, 80, or 100) $\times$ six levels of PAR (0.05, 0.1, ..., 0.3) or proportion of variance explained by causal SNPs (0.005, 0.01, ..., 0.03) $\times 1000$ replications for each scenario). Across all these 108,000 replications, there were totally 540,000 ($=108{,}000 \times 5$) causal SNPs/SNVs (because 5 causal SNPs/SNVs or QTLs were specified in each replication). Among these 540,000 causal variants, approximately 45% were rare (MAF$<$1%), $\sim$62% were uncommon/rare (MAF$<$5%), and $\sim$ 38% were common (MAF$\geq$5%). In our simulation setting, causal variants were not limited to be rare, because in real situations causal variants could also be common.

Table 1. Type-I error rates based on 10,000 replications.

When dichotomous traits were considered					
nominal significance level	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**
ADA[a]	0.0099	0.0201	0.0298	0.0399	0.0503
Kernel	0.0097	0.0197	0.0284	0.0375	0.0474
Burden	0.0107	0.0210	0.0311	0.0403	0.0501
When continuous traits were considered					
ADA[a]	0.0105	0.0203	0.0301	0.0403	0.0502
Kernel	0.0085	0.0180	0.0282	0.0383	0.0484
Burden	0.0103	0.0201	0.0303	0.0402	0.0498
FFBSKAT	0.0105	0.0181	0.0278	0.0359	0.0464

[a]*P*-values were estimated based on 1,000 permutations.

doi:10.1371/journal.pone.0115971.t001

Figs. 2 and 3 present the power for dichotomous traits and continuous traits, respectively. When dichotomous traits were studied (Fig. 2), *Kernel* was the most powerful method when $m=50$ (signal proportion $=0.1$), whereas *ADA* had the best performance when $m=150$ (signal proportion $=0.033$). When $m=100$ (signal proportion $=0.05$), these two methods had comparable performance. *Burden* was the uniformly least powerful test among the methods we compared, regardless of the size of m (50, 100, or 150). To conclude, when the signal proportion was larger, *Kernel* was more powerful; when the signal proportion was smaller, *ADA* took the advantage of truncating noise variants and therefore was more powerful.

When continuous traits were studied (Fig. 3), *Kernel* was again the most powerful method when $m=50$ (signal proportion $=0.1$). When the signal proportion was getting lower and lower (or, m was getting larger and larger), *Kernel* had a more substantial power loss. *Burden* was again the least powerful method, regardless of the size of m (50, 100, or 150). Compared with *Kernel*, *FFBSKAT* and *ADA* were less vulnerable to the inclusion of neutral variants. *FFBSKAT* became the most powerful method when $m=100$ (signal proportion $=0.05$) or when $m=150$ (signal proportion $=0.033$).

Application to Genetic Analysis Workshop 17 Simulated Data

We then applied the four tests to the Genetic Analysis Workshop 17 (GAW 17) simulated data [41]. The GAW 17 data set was designed to mimic a subset of data that might be generated in a full exome investigation for a complex disease. To reflect realistic human genomes, real sequence data were generated based on the 1000 Genomes Project [53]. The data set consisted of 697 subjects from eight large pedigrees, in which 202 founders had genotypes randomly selected from the 1000 Genomes Project. The MAFs ranged from 0.07% to 16.5%. These founders included 66 Tuscan, 50 Luhya, 28 Japanese, 19 Han Chinese, 18 Denver Chinese, 12 CEPH (European-descent residents of Utah), and 9 Yoruban samples. Pedigrees included four generations, and relatives were as distant as second

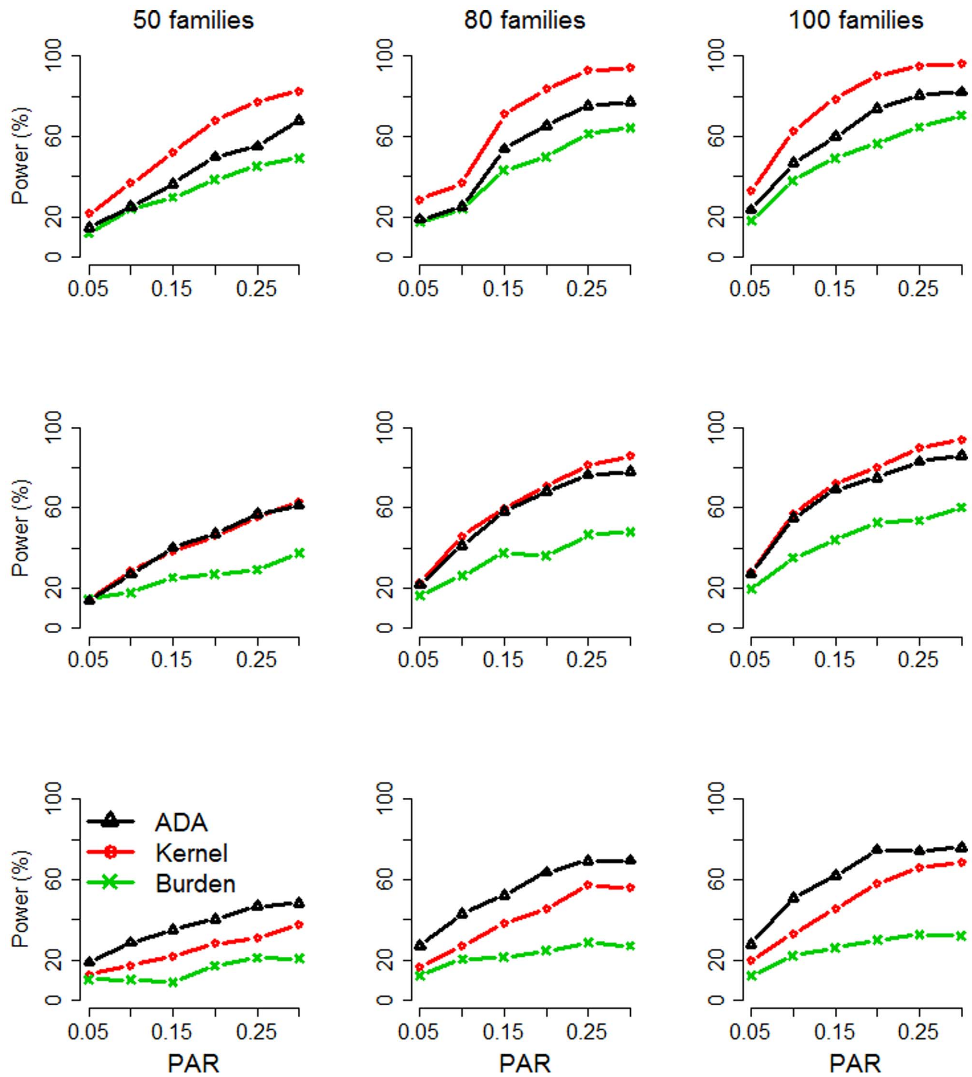

Fig. 2. Power Comparison for dichotomous traits. The figure shows the empirical power given the significance level of 0.05. Top row: 50 variants included in the tests; middle row: 100 variants; bottom row: 150 variants. The *x*-axis is the overall population attributable risk (PAR) for all causal loci, whereas the *y*-axis is the power.

doi:10.1371/journal.pone.0115971.g002

cousins. The causal SNPs/SNVs were listed in Table 2. Phenotype simulations were performed multiple times to generate 200 replications. With this simulated data set, we could evaluate the power performance of the four statistical methods, given a more general pedigree structure and a more realistic sequence composition [41]. The β column was the change in mean quantitative trait due to a copy of minor allele.

To analyze this data set, we first obtained residuals ($\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)$ in Eq. (1)) by regressing the trait values on age and smoking status. Table 2 listed the results by *FFBSKAT*, *ADA*, *Kernel*, and *Burden*. The P-values of *ADA* were estimated based on 1,000 permutations. Based on the power to detect causal genes, these methods

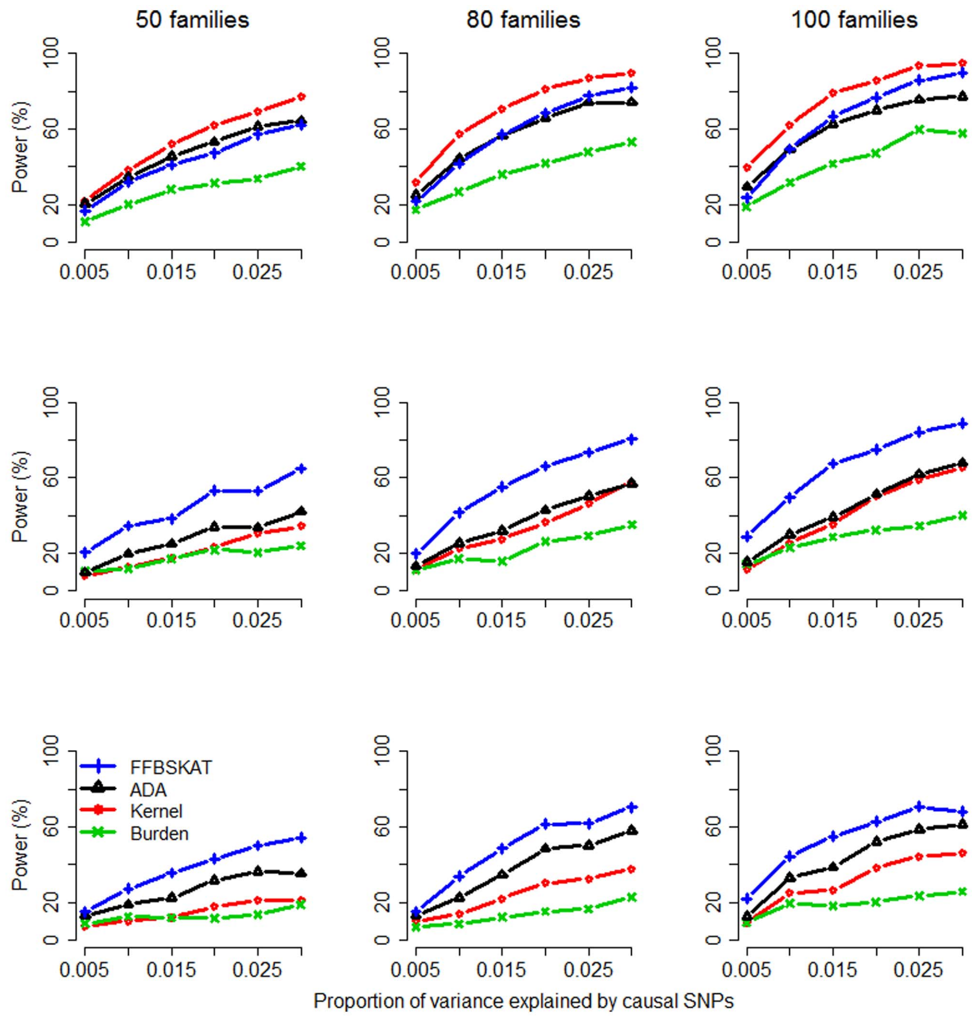

Fig. 3. Power Comparison for continuous traits. The figure shows the empirical power given the significance level of 0.05. Top row: 50 variants included in the tests; middle row: 100 variants; bottom row: 150 variants. The *x*-axis is the proportion of variance explained by causal SNPs, whereas the *y*-axis is the power.

doi:10.1371/journal.pone.0115971.g003

were roughly ranked as *FFBSKAT*> *ADA*> *Kernel* ≈ *Burden*. *FFBSKAT* was the most powerful method. It was based on the linear mixed effects model, in which the kinship relatedness was captured by the random effects terms ($\boldsymbol{b} \sim \boldsymbol{N}(\boldsymbol{0}, \sigma_b^2 \boldsymbol{\Omega})$, as described in our Introduction section). In the GAW 17 data set, the eight pedigrees were all quite large (see S2 Fig.). The linear mixed effects model ($\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{G\gamma} + \boldsymbol{b} + \boldsymbol{\varepsilon}$) partitions the total phenotypic variance into several parts. When pedigrees are larger, this model is a better choice because σ_b^2 can be estimated more accurately. (We used SeqSIMLA2 [54] to simulate 1,380 subjects based on two pedigree structures: (A) 20 pedigrees each with 69 members (Fig. 1 of [54]), and (B) 115 pedigrees each with 12 members (Fig. 1). The variance of the polygenic effects, σ_b^2, was specified at 45 in SeqSIMLA2 [54]. With 1,000 replications, the mean of $\hat{\sigma}_b^2$ estimated by the *lmekin* function [55] was 44.33 and

Table 2. Analysis of the Genetic Analysis Workshop 17 simulated data.

Causal gene	No. of SNPs/ SNVs	No. of causal SNPs/SNVs	Signal proportion	Causal SNP/ SNV	MAF	β	Power to detect the causal gene[a] (significance level =0.05)			
							FFBSKAT	*ADA*[b]	*Kernel*	*Burden*
ARNT	18	5	0.28	C1S6533	0.011478	0.589734	0.175	0.065	0.005	0
				C1S6537	0.000717	0.642689				
				C1S6540	0.001435	0.323662				
				C1S6542	0.002152	0.488219				
				C1S6561	0.000717	0.625721				
ELAVL4	10	2	0.20	C1S3181	0.000717	0.795093	0.19	0.07	0	0.005
				C1S3182	0.000717	0.328748				
FLT1	35	11	0.31	C13S320	0.001435	0.18047	1	0.76	0.335	0.38
				C13S399	0.000717	0.457361				
				C13S431	0.017217	0.732566				
				C13S479	0.000717	0.839669				
				C13S505	0.000717	0.38582				
				C13S514	0.000717	0.549816				
				C13S522	0.027977	0.623466				
				C13S523	0.066714	0.653351				
				C13S524	0.004304	0.596704				
				C13S547	0.000717	0.549214				
				C13S567	0.000717	0.090586				
FLT4	10	2	0.20	C5S5133	0.001435	0.120761	0.74	0.39	0.03	0.065
				C5S5156	0.000717	0.385374				
HIF1A	8	4	0.50	C14S1718	0.000717	0.251622	0.03	0.015	0	0
				C14S1729	0.002152	0.329088				
				C14S1734	0.012195	0.220448				
				C14S1736	0.000717	0.228202				
HIF3A	21	3	0.14	C19S4799	0.000717	0.174668	0.175	0.035	0.01	0.005
				C19S4815	0.000717	0.51468				
				C19S4831	0.000717	0.265181				
KDR	16	10	0.63	C4S1861	0.002152	0.598271	0.925	0.845	0.72	0.87
				C4S1873	0.000717	0.715613				
				C4S1874	0.000717	0.503025				
				C4S1877	0.000717	1.17194				
				C4S1878	0.164993	0.149975				
				C4S1879	0.000717	0.610938				
				C4S1884	0.020803	0.318125				
				C4S1887	0.000717	0.312058				
				C4S1889	0.000717	1.17194				
				C4S1890	0.002152	0.417977				
VEGFA	6	1	0.17	C6S2981	0.002152	1.13045	1	1	1	0.995

[a]Power to detect a causal gene $= \frac{\#\{declare\ significance\ among\ the\ 200\ replicates\}}{200}$.

[b]*P*-values were estimated based on 1,000 permutations.

doi:10.1371/journal.pone.0115971.t002

47.51 for cases (A) and (B), respectively.) Therefore, *FFBSKAT* was more powerful than other methods when analyzing the GAW 17 large pedigree data.

Discussion

Many statistical methods were proposed for rare variant association testing, but most of them were designed for population-based studies. Among the family-based rare variant association testing methods, some extend the transmission disequilibrium test [56, 57] and focus on parent-child trio data [36, 37]. Some methods are eligible for analyzing pedigree data (including but not limited to trios), and they can be categorized as the burden tests and the non-burden tests (e.g., *famSKAT* [31, 38], *FFBSKAT* [39], *Kernel* [32]). The non-burden tests were shown to be more powerful than the burden tests in most situations [32, 33].

Among the non-burden tests compared in this work, *FFBSKAT* [39] (and *famSKAT* [31, 38]) can only deal with continuous traits, but it is more powerful than the other methods (*ADA*, *Kernel*, *Burden*) when the pedigree is larger (see the GAW 17 data analysis) or when the signal proportion is smaller (see the bottom two rows of Fig. 3).

In population-based studies, *ADA* is robust to the inclusion of neutral variants, and therefore it has been shown to outperform the burden tests and the non-burden tests (e.g., *SKAT*) [13, 21]. In this work, we extend *ADA* to family-based studies and compare it with *Kernel* and *Burden*, the two commonly used methods for dichotomous traits. Simulation studies show that *ADA* is more powerful than other two competitors when the percentage of causal variants is smaller (see the bottom row of Fig. 2). On the contrary, *Kernel* is more powerful when the percentage of causal variants is larger (the top row of Fig. 2). *Burden* has the least power across the simulation scenarios we have investigated. The comparison between *Kernel* and *Burden* is consistent with that found by previous studies [32, 33].

Kernel, *Burden*, and *FFBSKAT* can provide analytical *P*-values when the sample size is large. *ADA* searches for the optimal threshold among multiple *P*-value truncation thresholds. Therefore, permutation is required to assess the statistical significance, and so *ADA* needs more computational time than other methods. To be more computationally efficient, *ADA* can be combined with a sequential Monte Carlo algorithm [58]. For simulated data sets containing 50 families and 50 SNPs/SNVs, *ADA* on average needs ~168.2 sec, *Kernel* or *Burden* takes ~37.6 sec, and *FFBSKAT* needs ~4.3 sec. This was measured on a Linux platform with an Intel Xeon E5-2690 2.9 GHz processor and 2 GB memory. Although the computation time of other competitors is much shorter than that of *ADA*, *ADA* is more robust to the inclusion of neutral variants when dichotomous traits are studied (see Fig. 2). But when continuous traits are analyzed, *FFBSKAT*/*Kernel* outperforms *ADA* when the signal proportion is smaller/larger (see Fig. 3).

Rare causal variants may play an important role in the etiology of complex diseases [59–64], but they are challenging to detect through single-locus tests

[1, 2, 65, 66]. Combining variants' signals in a chromosomal region and testing for association with a grouping statistic is a commonly used strategy. Compared with the burden test (*Burden*) and the non-burden test (*Kernel*), *ADA* is more robust to the inclusion of neutral variants when dichotomous traits are analyzed. However, there are some limitations for *ADA*. First, because this method is more computationally intensive, it is not realistic to apply it to genome-wide sequencing data. Second, pedigree structures and founders' sequence data are required for the permutation procedure implemented in *ADA*. Third, unrelated controls cannot be included in the *ADA* analyses. This work shows that, for family-based studies, the application of *ADA* is limited to dichotomous trait analyses with full pedigree information.

Supporting Information

S1 Fig. The distribution of genotype relative risk (GRR) of causal SNPs/SNVs when the overall population attributable risk (PAR) for all causal loci was assumed to be 0.05, 0.10, 0.15, 0.20, 0.25, and 0.30, respectively. Therefore, the marginal PAR for each causal SNP/SNV was 0.01, 0.02, 0.03, 0.04, 0.05, and 0.06, respectively.
doi:10.1371/journal.pone.0115971.s001 (PDF)

S2 Fig. Structures of the eight pedigrees (accordingly, from pedigree 1, 2, …, 8) in the Genetic Analysis Workshop 17 data set, plotted by the R package "kinship2".
doi:10.1371/journal.pone.0115971.s002 (PDF)

Acknowledgments

The author would like to thank the anonymous reviewers for their insightful and constructive comments, and the GAW17 workshop organizers for their permission to use their data in this research. Preparation of the Genetic Analysis Workshop 17 Simulated Exome Data Set was supported by the GAW grant, R01 GM031575, and in part by NIH R01 MH059490. The workshop used sequencing data from the 1000 Genomes Project (http://www.1000genomes.org).

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. **Bansal V, Libiger O, Torkamani A, Schork NJ** (2010) Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet 11: 773–785.

2. **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.
3. **Madsen BE, Browning SR** (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5: e1000384.
4. **Morris AP, Zeggini E** (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34: 188–193.
5. **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.
6. **Han F, Pan W** (2010) A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 70: 42–54.
7. **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.
8. **Lee S, Wu MC, Lin X** (2012) Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13: 762–775.
9. **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.
10. **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.
11. **Yi N, Zhi D** (2011) Bayesian analysis of rare variants in genetic association studies. Genet Epidemiol 35: 57–69.
12. **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.
13. **Lin WY, Lou XY, Gao G, Liu N** (2014) Rare Variant Association Testing by Adaptive Combination of P-values. PLoS One 9: e85728.
14. **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.
15. **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.
16. **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.
17. **Liu DJ, Leal SM** (2010) A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet 6: e1001156.
18. **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.
19. **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.
20. **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.
21. **Lin WY** (2014) Association testing of clustered rare causal variants in case-control studies. PLoS One 9: e94337.
22. **Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, et al.** (2010) Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. Am J Hum Genet 87: 604–617.
23. **Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al.** (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909.
24. **Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, et al.** (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203–208.

25. **Zhang Y, Guan W, Pan W** (2013) Adjustment for population stratification via principal components in association analysis of rare variants. Genet Epidemiol 37: 99–109.

26. **Mathieson I, McVean G** (2012) Differential confounding of rare and common variants in spatially structured populations. Nat Genet 44: 243–246.

27. **Lin WY, Zhang B, Yi N, Gao G, Liu N** (2011) Evaluation of pooled association tests for rare variant identification. BMC Proc 5 Suppl 9: S118.

28. **Ionita-Laza I, Ottman R** (2011) Study designs for identification of rare disease variants in complex diseases: the utility of family-based designs. Genetics 189: 1061–1068.

29. **Ionita-Laza I, Cho MH, Laird NM** (2013) Statistical challenges in sequence-based association studies with population- and family-based designs. Statistics in Biosciences 5: 54–70.

30. **Saad M, Pierre AS, Bohossian N, Mace M, Martinez M** (2011) Comparative study of statistical methods for detecting association with rare variants in exome-resequencing data. BMC Proc 5 Suppl 9: S33.

31. **Chen H, Meigs JB, Dupuis J** (2013) Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol 37: 196–204.

32. **Schaid DJ, McDonnell SK, Sinnwell JP, Thibodeau SN** (2013) Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genet Epidemiol 37: 409–418.

33. **Saad M, Wijsman EM** (2014) Power of family-based association designs to detect rare variants in large pedigrees using imputed genotypes. Genet Epidemiol 38: 1–9.

34. **Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X** (2013) Family-based association tests for sequence data, and comparisons with population-based association tests. Eur J Hum Genet 21: 1158–1162.

35. **De G, Yip WK, Ionita-Laza I, Laird N** (2013) Rare variant analysis for family-based design. PLoS One 8: e48495.

36. **He Z, O'Roak BJ, Smith JD, Wang G, Hooker S, et al.** (2014) Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. Am J Hum Genet 94: 33–46.

37. **Cheng KF, Chen JH** (2013) Detecting rare variants in case-parents association studies. PLoS One 8: e74310.

38. **Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SL, et al.** (2012) SNP Set Association Analysis for Familial Data. Genet Epidemiol 36: 797–810.

39. **Svishcheva GR, Belonogova NM, Axenovich TI** (2014) FFBSKAT: fast family-based sequence kernel association test. PLoS One 9: e99407.

40. **Oualkacha K, Dastani Z, Li R, Cingolani PE, Spector TD, et al.** (2013) Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. Genet Epidemiol 37: 366–376.

41. **Almasy L, Dyer TD, Peralta JM, Kent JW, Jr., Charlesworth JC, et al.** (2011) Genetic Analysis Workshop 17 mini-exome simulation. BMC Proc 5 Suppl 9: S2.

42. **Jiang Y, Satten GA, Han Y, Epstein MP, Heinzen EL, et al.** (2014) Utilizing population controls in rare-variant case-parent association tests. Am J Hum Genet 94: 845–853.

43. **Thornton T, McPeek MS** (2010) ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. Am J Hum Genet 86: 172–184.

44. **Fan R, Knapp M, Wjst M, Zhao C, Xiong M** (2005) High resolution T association tests of complex diseases based on family data. Ann Hum Genet 69: 187–208.

45. **Dudbridge F** (2003) Pedigree disequilibrium tests for multilocus haplotypes. Genet Epidemiol 25: 115–121.

46. **Browning BL, Browning SR** (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84: 210–223.

47. **Chung RH, Shih CC** (2013) SeqSIMLA: a sequence and phenotype simulation tool for complex disease studies. BMC Bioinformatics 14: 199.

48. **Liang L, Zollner S, Abecasis GR** (2007) GENOME: a rapid coalescent-based whole genome simulator. Bioinformatics 23: 1565–1567.

49. **Hudson RR** (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.

50. **Donnelly P, Tavare S** (1995) Coalescents and genealogical structure under neutrality. Annu Rev Genet 29: 401–421.

51. **Hudson RR** (1983) Properties of a neutral allele model with intragenic recombination. Theor Popul Biol 23: 183–201.

52. **Hudson RR** (1990) Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology 7: 1–44.

53. **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.

54. **Chung RH, Tsai WY, Hsieh CH, Hung KY, Hsiung CA, et al.** (2014) SeqSIMLA2: Simulating Correlated Quantitative Traits Accounting for Shared Environmental Effects in User-Specified Pedigree Structure. Genet Epidemiol.

55. **Therneau T** (2012) R Package 'coxme'. Version 2.2–3.

56. **Terwilliger JD, Ott J** (1992) A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. Hum Hered 42: 337–346.

57. **Spielman RS, McGinnis RE, Ewens WJ** (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52: 506–516.

58. **Besag J, Clifford P** (1991) Sequential Monte Carlo p-values. Biometrika 78: 301–304.

59. **Azzopardi D, Dallosso AR, Eliason K, Hendrickson BC, Jones N, et al.** (2008) Multiple rare nonsynonymous variants in the adenomatous polyposis coli gene predispose to colorectal adenomas. Cancer Res 68: 358–363.

60. **Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, et al.** (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305: 869–872.

61. **Hershberger RE, Norton N, Morales A, Li D, Siegfried JD, et al.** (2010) Coding sequence rare variants identified in MYBPC3, MYH6, TPM1, TNNC1, and TNNI3 from 312 patients with familial or idiopathic dilated cardiomyopathy. Circ Cardiovasc Genet 3: 155–161.

62. **Bodmer W, Bonilla C** (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40: 695–701.

63. **Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB** (2010) Rare variants create synthetic genome-wide associations. PLoS Biol 8: e1000294.

64. **Pritchard JK** (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69: 124–137.

65. **Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI** (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82: 100–112.

66. **Altshuler D, Daly MJ, Lander ES** (2008) Genetic mapping in human disease. Science 322: 881–888.