European Journal of Human Genetics (2015) 23, 229–237

www.nature.com/ejhg

ARTICLE

Test of rare variant association based on affected sib-pairs

Qiuying Sha[1] and Shuanglin Zhang*,[1]

With the development of sequencing techniques, there is increasing interest to detect associations between rare variants and complex traits. Quite a few statistical methods to detect associations between rare variants and complex traits have been developed for unrelated individuals. Statistical methods for detecting rare variant associations under family-based designs have not received as much attention as methods for unrelated individuals. Recent studies show that rare disease variants will be enriched in family data and thus family-based designs may improve power to detect rare variant associations. In this article, we propose a novel test to test association between the optimally weighted combination of variants and trait of interests for affected sib-pairs. The optimal weights are analytically derived and can be calculated from sampled genotypes and phenotypes. Based on the optimal weights, the proposed method is robust to the directions of the effects of causal variants and is less affected by neutral variants than existing methods are. Our simulation results show that, in all the cases, the proposed method is substantially more powerful than existing methods based on unrelated individuals and existing methods based on affected sib-pairs.
European Journal of Human Genetics (2015) **23,** 229–237; doi:10.1038/ejhg.2014.43; published online 26 March 2014

INTRODUCTION

Recent studies show that the large number of disease-associated variants identified through genome-wide association studies account for only a small portion of the presumed phenotypic variation.[1] One of the potential sources of missing heritability is the contribution of rare variants.[2–7] The recent advances of sequencing technology have made directly testing rare variants possible.[8,9] Therefore, there is increasing interest to detect associations between rare variants and complex traits.

Recently, several statistical methods to detect associations between rare variants and complex traits have been developed for unrelated individuals. These methods can be roughly divided into three groups: burden tests, quadratic tests, and combined tests. Burden tests include the cohort allelic sums test,[10] the combined multivariate and collapsing method,[11] the weighted sum statistic (WSS),[12] the variable minor allele frequency (MAF) threshold method,[13] and the cumulative minor-allele test[14] among others. Burden tests implicitly assume that all the rare variants are causal and the directions of the effects are all the same. If these assumptions are true, burden tests can be powerful tests; otherwise, burden tests can perform poorly.[15–18] Quadratic tests include C-alpha test,[19] sequence kernel association test,[15] and the test for Testing the effects of the Optimally Weighted combination of variants (TOW).[17] Quadratic tests also include adaptive weighting methods[20–24] since, as pointed out by Derkach *et al*,[18] adaptive weighting methods are operationally similar to quadratic tests. Quadratic tests are robust to the directions of the effects of causal variants and are less affected by neutral variants than burden tests are. If most of the rare variants are causal and the directions of the effects of causal variants are all the same, then burden tests can outperform quadratic tests; otherwise, quadratic tests perform better. To increase the robustness of a test, Derkach *et al* and Lee *et al* proposed combined tests that combine information from burden and quadratic tests aiming to have advantages of both burden and quadratic tests.[16,18]

All of the aforementioned methods are for unrelated individuals. For any type of study design, the statistical power will be improved if rare variants can be enriched in the samples. If one parent has a copy of a rare allele, half of the offspring are expected to carry it, and hence, variants that are rare in the general population could be very common in certain families.[25] Therefore, family-based designs may have an important role in rare variant association studies. More recently, a couple of family-based rare variant association methods for quantitative traits[26,27] and for qualitative traits[28,29] have been developed.

In this article, based on affected sib-pair data, we propose a test for Testing the effects of the Optimally Weighted combination of variants (TOW-sib). TOW-sib is based on the score test for testing the optimally weighted combination of variants derived from the retrospective likelihood of affected sib-pairs, unrelated controls, and possible unrelated cases. The optimal weights are analytically derived and can be calculated from sampled genotypes and phenotypes. Based on the optimal weights, TOW-sib is robust to the directions of the effects of causal variants and is less affected by neutral variants than existing tests are. We use extensive simulation studies to compare the performance of the proposed method with that of existing methods based on unrelated individuals[12,17] and existing methods based on affected sib-pairs.[28] Our simulation results show that, in all the cases, the proposed method is substantially more powerful than existing methods based on either unrelated individuals or affected sib-pairs.

MATERIALS AND METHODS

Consider a sample of n_s affected sib-pairs, n_a unrelated cases, and n_c unrelated controls. Each individual has been genotyped at M variants in a genomic region. Denote $g_{ji}=(g_{ji1},...,g_{jiM})^T$, $g_{ai}=(g_{ai1},\ldots,g_{aiM})^T$, and $g_{ci}=(g_{ci1},\ldots,g_{ciM})^T$

[1]Department of Mathematical Sciences, Michigan Technological University, Houghton, MI, USA
*Correspondence: Professor S Zhang, Department of Mathematical Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, USA. Tel: +906 487 2146; Fax: +906 487 3133; E-mail: shuzhang@mtu.edu
Received 7 July 2013; revised 6 November 2013; accepted 30 December 2013; published online 26 March 2014

as the genotypes of the j^{th} individual in the i^{th} sib-pair, the i^{th} case, and the i^{th} control, respectively, where g_{jim}, g_{aim}, $g_{cim} \in \{0,1,2\}$ are the number of minor alleles. Let $x_{ji} = \sum_{m=1}^{M} w_m g_{jim}$ $(j = 1, 2)$, $x_{ai} = \sum_{m=1}^{M} w_m g_{aim}$, and $x_{ci} = \sum_{m=1}^{M} w_m g_{cim}$ denote the combinations of genotypic scores at the M variants of the i^{th} sib-pair, the i^{th} case, and the i^{th} control, respectively, where $w = (w_1,...,w_M)$ are weights and their values will be decided later. Denote the disease status of an individual by D with $D = 0$ indicating a normal, whereas $D = 1$ indicating a diseased individual.

The retrospective likelihood is given by

$$L = \prod_{i=1}^{n_s} \Pr(g_{1i}, g_{2i} \mid D=1, D=1) \prod_{i=1}^{n_a} \Pr(g_{ai} \mid D=1) \prod_{i=1}^{n_c} \Pr(g_{ci} \mid D=0)$$
$$= \prod_{i=1}^{n_s} \frac{\Pr(D=1 \mid g_{1i}) \Pr(D=1 \mid g_{2i}) \Pr(g_{1i}, g_{2i})}{\sum_{g_1^*, g_2^*} \Pr(D=1 \mid g_1^*) \Pr(D=1 \mid g_2^*) \Pr(g_1^*, g_2^*)} \cdot \prod_{i=1}^{n_a} \frac{\Pr(D=1 \mid g_{ai}) \Pr(g_{ai})}{\sum_{g^*} \Pr(D=1 \mid g^*) \Pr(g^*)},$$
$$\cdot \prod_{i=1}^{n_c} \frac{(1 - \Pr(D=1 \mid g_{ci})) \Pr(g_{ci})}{1 - \sum_{g^*} \Pr(D=1 \mid g^*) \Pr(g^*)}$$

where g_1^* and g_2^* represent all possible genotype pair for a sib-pair and g^* represents all possible genotypes for an individual. Choose $g_0 = (0,\ldots,0)$ as a baseline genotype. Let $r(g)$ be the relative risk of genotype g to the baseline genotype. Following Schaid,[30] we use a log-linear model to model the relative risk, ie, $r(g) = e^{x\beta}$, with x representing the combination of genotypic scores of the genotype g. Denote the risk of an individual with the baseline genotype as $\Pr(D = 1|g_0) = e^{\alpha}$. Then, the retrospective likelihood is given by

$$L = \prod_{i=1}^{n_s} \frac{e^{(x_{1i}+x_{2i})\beta} \Pr(g_{1i}, g_{2i})}{\sum_{g_1^*, g_2^*} e^{(x_1^*+x_2^*)\beta} \Pr(g_1^*, g_2^*)} \prod_{i=1}^{n_a} \frac{e^{x_{ai}\beta} \Pr(g_{ai})}{\sum_{g^*} e^{x^*\beta} \Pr(g^*)} \prod_{i=1}^{n_c} \frac{(1 - e^{\alpha + x_{ci}\beta}) \Pr(g_{ci})}{1 - \sum_{g^*} e^{\alpha + x^*\beta} \Pr(g^*)} \quad (1)$$

where x_1^* and x_2^* represent the combinations of genotypic scores of the genotypes g_1^* and g_2^*, respectively, and x^* represents the combination of genotypic scores of the genotype g^*.

In Appendix A, we have shown that, under the assumption that the M variants are independent (our proposed test is still valid if this assumption is not true), the score test statistic to test the null hypothesis H_0:$\beta = 0$ is given by

$$T(w_1, ..., w_M) = \frac{U^2}{V},$$

where $U = \sum_{i=1}^{n_s} (x_{1i} + x_{2i} - 4\hat{p}) + \sum_{i=1}^{n_a} (x_{ai} - 2\hat{p}) - \hat{a} \sum_{i=1}^{n_c} (x_{ci} - 2\hat{p})$, $V = (6n_s + 2n_a + 2n_c\hat{a}^2) \sum_{m=1}^{M} w_m^2 \hat{p}_m (1 - \hat{p}_m)$, $\hat{p}_m$ and $\hat{a}$ are the maximum likelihood estimates (MLEs) of p_m and $a = \frac{e^{\alpha}}{1-e^{\alpha}}$ under the null hypothesis, p_m is the MAF at the m^{th} variant, and $\hat{p} = \sum_{m=1}^{M} w_m \hat{p}_m$. Under the null hypothesis, the likelihood function becomes

$$L_0 = \prod_{i=1}^{n_s} \Pr(g_{1i}, g_{2i}) \prod_{i=1}^{n_a} \Pr(g_{ai}) \prod_{i=1}^{n_c} \Pr(g_{ci}).$$

Based on L_0, $\hat{p}_m$ has no explicit expression. Using the joint distribution of genotypes of a sib-pair given by Table 1, we can construct an expectation-maximization algorithm to calculate $\hat{p}_m$ (see Appendix B). We cannot estimate α based on L_0, because L_0 does not contain α. We propose to estimate α based on the full likelihood function

$$L_{full} = \prod_{i=1}^{n_s} \Pr(g_{1i}, g_{2i}, D=1, D=1) \prod_{i=1}^{n_a} \Pr(g_{ai}, D=1) \prod_{i=1}^{n_c} \Pr(g_{ci}, D=0).$$

Based on L_{full}, the MLE of $a = \frac{e^{\alpha}}{1-e^{\alpha}}$ under the null hypothesis is $\hat{a} = \frac{2n_s + n_a}{n_c}$. Using this estimate of a, U can be written as $U = \sum_{i=1}^{n_s} (x_{1i} + x_{2i}) + \sum_{i=1}^{n_a} x_{ai} - \hat{a} \sum_{i=1}^{n_c} x_{ci}$. Let $u_m = \sum_{i=1}^{n_s} (g_{1im} + g_{2im}) + \sum_{i=1}^{n_a} g_{aim} - \hat{a} \sum_{i=1}^{n_c} g_{cim}$, $u = (u_1, ..., u_M)^T$, $N = 6n_s + 2n_a + 2n_c\hat{a}^2$, $v = diag(N\hat{p}_1(1-\hat{p}_1), \ldots, N\hat{p}_M(1-\hat{p}_M))$, and $w = (w_1,\ldots,w_M)^T$. Then,

$$T(w_1, \ldots, w_M) = \frac{w^T u u^T w}{w^T v w}.$$

$T(w_1,\ldots,w_M)$ reaches its maxim when $w = v^{-1}u$. We define the statistic of the test for Testing the effect of an Optimally Weighted combination of variants for sib-pair data (TOW-sib) as

$$T_{TOW-sib} = \max_{w_1, \ldots, w_M} T(w_1, \ldots, w_M) = u^T v^{-1} u = \sum_{m=1}^{M} \frac{u_m^2}{N\hat{p}_m(1-\hat{p}_m)} = \sum_{m=1}^{M} T_m.$$

We use a special permutation test to evaluate P-values of TOW-sib. For each permutation, we have the following steps: (1) permute the multi-variant genotypes $g_{11}, ..., g_{1n_s}, g_{a1}, ..., g_{an_a}, g_{c1}, ..., g_{cn_c}$ and get the permuted genotypes $g_{11}^*, \ldots, g_{1n_s}^*, g_{a1}^*, \ldots, g_{an_a}^*, g_{c1}^*, \ldots, g_{cn_c}^*$. (2) In the i^{th} sib-pair, given g_{1i}^*, we generate g_{2i}^* variant by variant according to the conditional distribution $\Pr(g_2|g_1)$ from Table 1. (3) Calculate T_{TOWsib}^*, the value of $T_{TOW-sib}$ based on the permuted genotypes $g_{1i}^*, g_{2i}^* (i = 1, \ldots, n_s)$, $g_{ai}^* (i = 1, \ldots n_a)$, and $g_{ci}^* (i = 1, \ldots, n_c)$. We generate g_{2i}^* under the assumption that the M variants are independent. When the M variants are in linkage disequilibrium (LD), $T_{TOW-sib}$ and T_{TOWsib}^* may have different variances, although they have the same mean. In order to make $T_{TOW-sib}$ and T_{TOWsib}^* have the same mean and same variance, we standardize $T_{TOW-sib}$ such that $T_{TOW-sib-ST} = (T_{TOW-sib} - \mu_{TOW-sib})/\sigma_{TOW-sib}$, where $\mu_{TOW-sib}$ and σ_{TOWsib}^2 are the estimates of the mean and variance of $T_{TOW-sib}$ (see Appendix C on how to calculate $\mu_{TOW-sib}$ and σ_{TOWsib}^2). Suppose we perform B times of permutations. Let $T_{TOWsibST}^{*(b)}$ denote the value of $T_{TOW-sib-ST}$ based on data of the b^{th} permutation ($b = 0$ denotes the original data). Then, the P-value of the test is given by $Pvalue = \#\{b : T_{TOWsibST}^{*(b)} > T_{TOWsibST}^{*(0)}; b = 1, \ldots, B\}/B$.

For a simulation study with R replicates, the above procedure will be rather computationally expensive. In our simulation studies, we use the pooling permutation method proposed by Guo and Lin to evaluate P-values.[31] In the pooling permutation method, permuted samples from all the replicates are pooled together to form a joint sample from the null distribution. Suppose that we have R replicates and we perform B permutations for each replicate. Let $T_{TOW-sib-ST}^{(b,r)}$ denote the value of $T_{TOW-sib-ST}$ based on data of the b^{th} permutation of the r^{th} replicate ($b = 0$ denotes the original data). Then, the

Table 1 The joint distribution of genotypes of a sib-pair

	$Pr(g_1, g_2 \mid IBD = 0)$				$Pr(g_1, g_2 \mid IBD = 1)$		
g_1/g_2	0	1	2	g_1/g_2	0	1	2
0	q^4	$2pq^3$	p^2q^2	0	q^3	pq^2	0
1	$2pq^3$	$4p^2q^2$	$2p^3q$	1	pq^2	pq	p^2q
2	p^2q^2	$2p^3q$	p^4	2	0	p^2q	p^3

	$Pr(g_1, g_2 \mid IBD = 2)$				$Pr(g_1, g_2)$		
g1/g2	0	1	2	g1/g2	0	1	2
0	q^2	0	0	0	$q^2(1+q)^2/4$	$pq^2(q+1)/2$	$p^2q^2/4$
1	0	$2pq$	0	1	$pq^2(q+1)/2$	$pq(pq+1)$	$p^2q(p+1)/2$
2	0	0	p^2	2	$p^2q^2/4$	$p^2q(p+1)/2$	$p^2(1+p)^2/4$

Notes: g1 and g2 are the genotypes of a sib-pair at a single variant. IBD means identical by decent. p and q are the allele frequencies of the two alleles.

P-value of the test in the r^{th} replicate is given by

$$P\text{value} = \#\left\{(b, r_0) : T^{(b,r_0)}_{\text{TOWsibST}} > T^{(0,r)}_{\text{TOWsibST}}; b = 1, ..., B; r_0 = 1, ..., R\right\}/(BR).$$

As the permutation samples are pooled across all replicates to form a sample from the null, B can be set to be much smaller than the situation when only one sample is analyzed.

We compare the performance of the proposed method with three existing methods: WSS,[12] sibpair-based weighted sum statistic (SPWSS),[28] and TOW.[17] WSS and TOW are based on unrelated cases and controls, whereas SPWSS is based on affected sib-pairs, unrelated cases, and unrelated controls.

Simulation

The empirical Mini-Exome genotype data provided by the genetic analysis workshop 17 are used for simulation studies. This data set contains genotypes of 697 unrelated individuals on 3205 genes. The genotypes of the genetic analysis workshop 17 data set are extracted from the sequence alignment files provided by the 1000 Genomes Project for their pilot3 study (http://www.1000genomes.org). We choose four genes: *ELAVL4* (gene1), *MSH4* (gene2), *PDE4B* (gene3), and *ADAMTS4* (gene4) with 10, 20, 30, and 40 variants, respectively. We merge the four genes to form a super gene (Sgene) with 100 variants with 86 rare variants (MAF$<$0.01) and 14 common variants (MAF$\geq$0.01). We choose Sgene because the distributions of MAFs in the 100 variants in Sgene and in the 24 487 variants in all the 3205 genes are very similar.[17] In our simulation studies, we generate genotypes based on the genotypes of 697 individuals in Sgene. We use the program fastPHASE to infer haplotypic phase for the 697 individuals and calculate haplotype frequencies.[32] To generate the genotype of an individual, we generate two haplotypes according to the haplotype frequencies. To obtain the genotypes of a family, we first generate genotypes of parents. Then the genotypes of children are generated from parental haplotypes by random transmission. To generate a qualitative disease affection status, we use a liability threshold model based on a continuous phenotype (quantitative trait). An individual is defined to be affected if the individual's phenotype is at least one standard deviation larger than the phenotypic mean. This yields a prevalence of 16% for the simulated disease in the general population. In the following, we describe how to generate a quantitative trait.

Under the null hypothesis, we generate trait values for unrelated individuals according to the standard normal distribution. For a family with m children, let $Y_1 = (y_F, y_M)$ and $Y_2 = (y_1, y_2, \cdots, y_m)$ denote the trait values of the parents and the m children in a family, respectively. Assume that (Y_1, Y_2) follows a multivariate normal distribution with a mean vector of zero and variance-covariance matrix of

$$\sum = \begin{pmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{pmatrix}, \text{ where } \sum_{11} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \sum_{12} = \sum_{21}^{T} = \begin{pmatrix} \rho & \dots & \rho \\ \rho & \dots & \rho \end{pmatrix},$$

$$\text{and } \sum_{22} = \begin{pmatrix} 1 & \dots & \rho \\ & \ddots & \\ \rho & \dots & 1 \end{pmatrix}.$$

This variance-covariance matrix indicates that the parents in each family are independent, and the correlation coefficient between a parent and a child or between two children is constant, ρ (in this study, $\rho = 0.2$). To generate trait values of all members in each family, we first generate the trait value of a parent by using a standard normal distribution. Then, trait values of the children are generated by a normal distribution with a mean vector $\mu_c = \sum_{21} \sum_{11}^{-1} Y_1$ and a variance–covariance matrix $\sum_c = \sum_{22} - \sum_{21} \sum_{11}^{-1} \sum_{12}$.

Under the alternative hypothesis, we choose n_{cau} rare variants (MAF$<$1%) as causal variants. The value of n_{cau} is determined by p_{cau}, the percentage of causal variants in rare variants. Let pp denote the percentage of protective variants in causal variants, then the number of protective variants and the number of risk variants are $n_{\text{p}} = n_{\text{cau}} \cdot pp$ and $n_{\text{r}} = n_{\text{cau}} \cdot (1 - pp)$, respectively. For the j^{th} member in the i^{th} family, let x_{ijk_r} and x_{ijk_p} denote the genotypic scores of the k_r^{th} risk variant and the k_p^{th} protective variant, respectively. Assume that all causal variants have the same heritability. Then the disease model is given by $y_{ij} = \sum_{k_r=1}^{n_r} \beta_{k_r} x_{ijk_r} - \sum_{k_p=1}^{n_p} \beta_{k_p} x_{ijk_p} + \varepsilon_{ij}$, where β_{k_r} and β_{k_p} are coefficients and their values depend on the total heritability, and ε_{ij} is the trait value under the null hypothesis.

To generate affected sib-pairs, we generate families with two children. We keep generating families with two children until we have generated enough families with two affected children.

RESULTS

In simulation studies, P-values are estimated using a pooling permutation method in which permuted samples from all the replicates are pooled together to form a joint sample from the null distribution.[31] In each replicate, we perform 20 permutations. Type I error rates are evaluated using 10 000 replicated samples, whereas powers are evaluated using 500 replicated samples.

For type I error evaluation, we consider different haplotype structures (different genes), different sample sizes, different designs, and different significance levels. For 10 000 replicated samples, the 95% confidence intervals for type I error rates of nominal levels 0.05, 0.01, and 0.001 are (0.046, 0.054), (0.008, 0.012), and (0.0004, 0.0016), respectively. The estimated type I error rates of the proposed test are summarized in Tables 2 and 3. As shown by these tables, all the estimated type I error rates are within the 95% confidence intervals, which indicates that the proposed test is valid.

For fixed number of total cases and fixed number of total individuals, power comparisons for power as a function of the number of affected sib-pairs are given in Figure 1. As shown by Figure 1, the power of TOW-sib increases with the increase of the number of affected sib-pairs. With the increase of the number of affected sib-pairs, the power of SPWSS increases if the number of affected sib-pairs is less than 20% of total number of cases and the power of SPWSS decreases otherwise. Therefore, in the following

Table 2 Estimated type I error rates of TOW-sib for the design of affected sib-pairs and unrelated controls based on 10 000 replicated samples

	Significance level = 0.05			*Significance level = 0.01*			*Significance level = 0.001*		
	Sample size			*Sample size*			*Sample size*		
	1000	*2000*	*4000*	*1000*	*2000*	*4000*	*1000*	*2000*	*4000*
Gene 1	0.0467	0.0492	0.0472	0.0108	0.0094	0.0098	0.0014	0.0016	0.0012
Gene 2	0.0469	0.0477	0.0524	0.0086	0.0085	0.0119	0.0008	0.0008	0.0014
Gene 3	0.0469	0.0484	0.0468	0.0094	0.0113	0.0083	0.0013	0.0014	0.0007
Gene 4	0.0479	0.0478	0.0491	0.0088	0.0096	0.0092	0.0007	0.0013	0.0010
Sgene	0.0465	0.0467	0.0478	0.0091	0.0086	0.0088	0.0008	0.0008	0.0008

Note: n_{sample} is the sample size, ie, the total number of individuals in the sample. n_{sib} is the number of affected sib-pairs. n_{case} is the number of unrelated cases. n_{control} is the number of unrelated controls. $n_{\text{sib}} = n_{\text{sample}}/4$, $n_{\text{case}} = 0$, and $n_{\text{control}} = n_{\text{sample}}/2$.

Table 3 Estimated type I error rates of TOW-sib for the design of affected sib-pairs, unrelated cases, and unrelated controls based on 10 000 replicated samples

	Significance level = 0.05			*Significance level = 0.01*			*Significance level = 0.001*		
	Sample size			*Sample size*			*Sample size*		
	1000	*2000*	*4000*	*1000*	*2000*	*4000*	*1000*	*2000*	*4000*
Gene 1	0.0467	0.0507	0.0512	0.0108	0.0118	0.0091	0.0014	0.0014	0.0011
Gene 2	0.0479	0.0534	0.0529	0.0086	0.0103	0.0112	0.0008	0.0011	0.0015
Gene 3	0.0469	0.0521	0.0537	0.0094	0.0108	0.011	0.0013	0.0015	0.0014
Gene 4	0.0469	0.0539	0.0526	0.0088	0.0104	0.0112	0.0007	0.0014	0.0013
Sgene	0.0465	0.0521	0.0516	0.0091	0.0118	0.0117	0.0008	0.0014	0.0015

Note: n_{sample} is sample size, ie, the total number of individuals in the sample. n_{sib} is the number of affected sib-pairs. n_{case} is the number of unrelated cases. $n_{control}$ is the number of unrelated controls. $n_{sib} = n_{sample}/8$, $n_{case} = n_{sample}/4$, and $n_{control} = n_{sample}/2$.

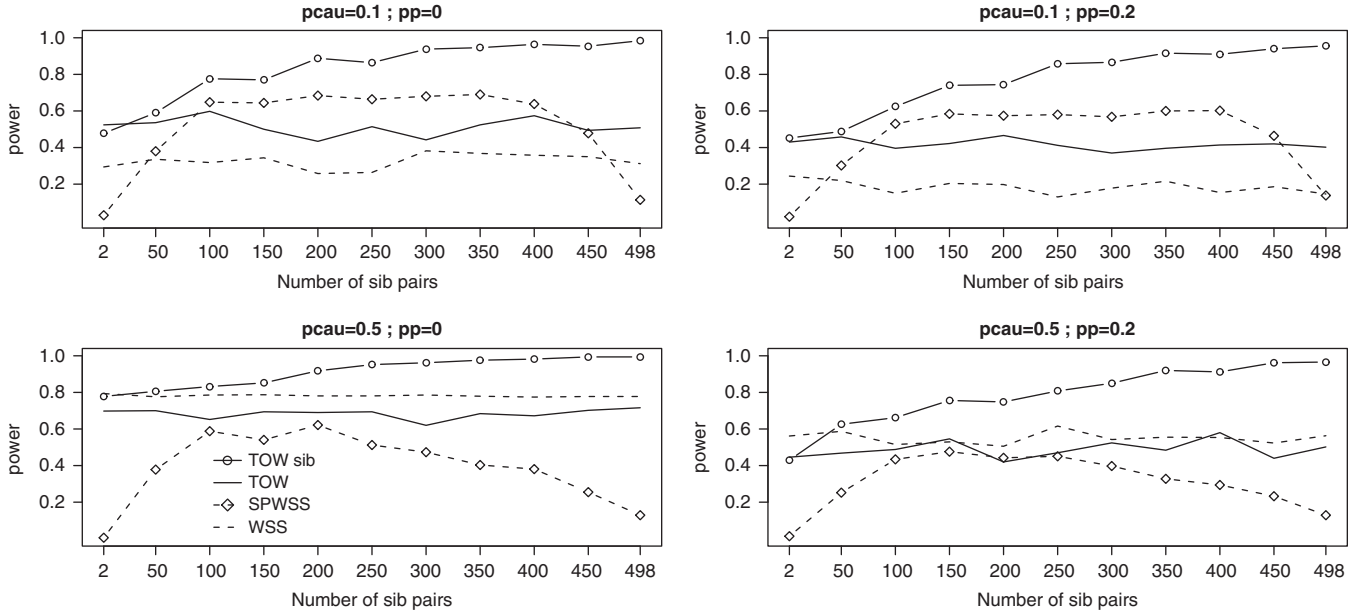

Figure 1 Power comparisons of four tests for power as a function of number of affected sib-pairs. TOW and WSS are based on 1000 unrelated cases and 1000 unrelated controls. For TOW-sib and SPWSS, the sample size is 2000, where number of unrelated controls is 1000 and number of unrelated cases plus twice of the number of affected sib-pairs is 1000. Total heritability is 0.03. pcau denotes the percentage of causal variants in rare variants; pp denotes the percentage of protective variants in causal variants. The power is evaluated at a significance level of 0.001.

discussion, the number of affected sib-pairs is equal to the half of total number of cases in the design for TOW-sib and the number of affected sib-pairs is equal to 20% of total number of cases in the design for SPWSS. The powers of TOW and WSS do not have relation with the number of affected sib-pairs. In almost all the cases, TOW-sib is the most powerful test. When the percentage of causal variants is small (10%), SPWSS is more powerful than TOW and WSS if the number of affected sib-pairs is between 10 and 45% of the total number of cases. When the percentage of causal variants is large (50%), SPWSS is the least powerful test.

As shown by power comparisons for power as a function of heritability and for power as a function of the percentage of protective variants (Figures 2 and 3), TOW-sib is the most powerful test in all the cases. When the percentage of causal variants is small (10%), SPWSS is more powerful than TOW and WSS. When the percentage of causal variants is large (50%), SPWSS and TOW have similar power and are less powerful than WSS if the percentage of protective variants is small and are more powerful than WSS if the percentage of protective variants is large.

Figure 4 shows power comparisons for power as a function of the percentage of causal variants. This figure shows that TOW-sib is the most powerful test in all the cases and the power of TOW-sib is not affected much by the percentage of causal variants. With the increase of the percentage of causal variants, the powers of WSS and TOW increase, whereas the power of SPWSS decreases. It is easy to understand that the power increases with the increase of the percentage of causal variants because larger percentage of causal variants or smaller percentage of neutral variants means smaller noise level. The reason of decrease in power of SPWSS with the increase of the percentage of causal variants probably is that it is easier to estimate weights when the percentage of causal variants is smaller. We also conduct a set of simulations to compare the powers for different values of ρ. The results (Supplementary Figure 1) show that the power comparisons have similar patterns for different values of ρ.

In summary, TOW-sib is the most powerful test in all the cases. Among other three tests: WSS, SPWSS, and TOW, none is consistently more powerful than the other two.

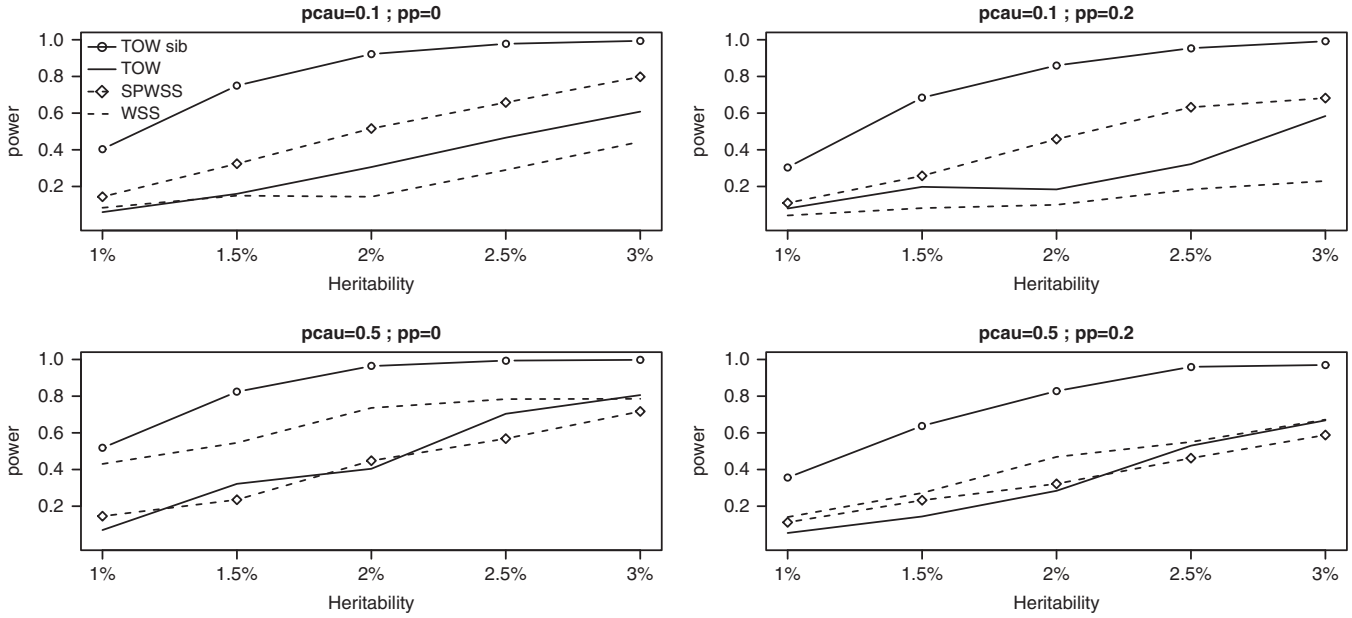

Figure 2 Powers as a function of heritability. TOW and WSS are based on 1000 unrelated cases and 1000 unrelated controls. SPWSS is based on 1000 unrelated controls, 600 unrelated cases, and 200 affected sib-pairs. TOW-sib is based on 1000 unrelated controls and 500 affected sib-pairs. pcau denotes the percentage of causal variants in rare variants; pp denotes the percentage of protective variants in causal variants. The power is evaluated at a significance level of 0.001.

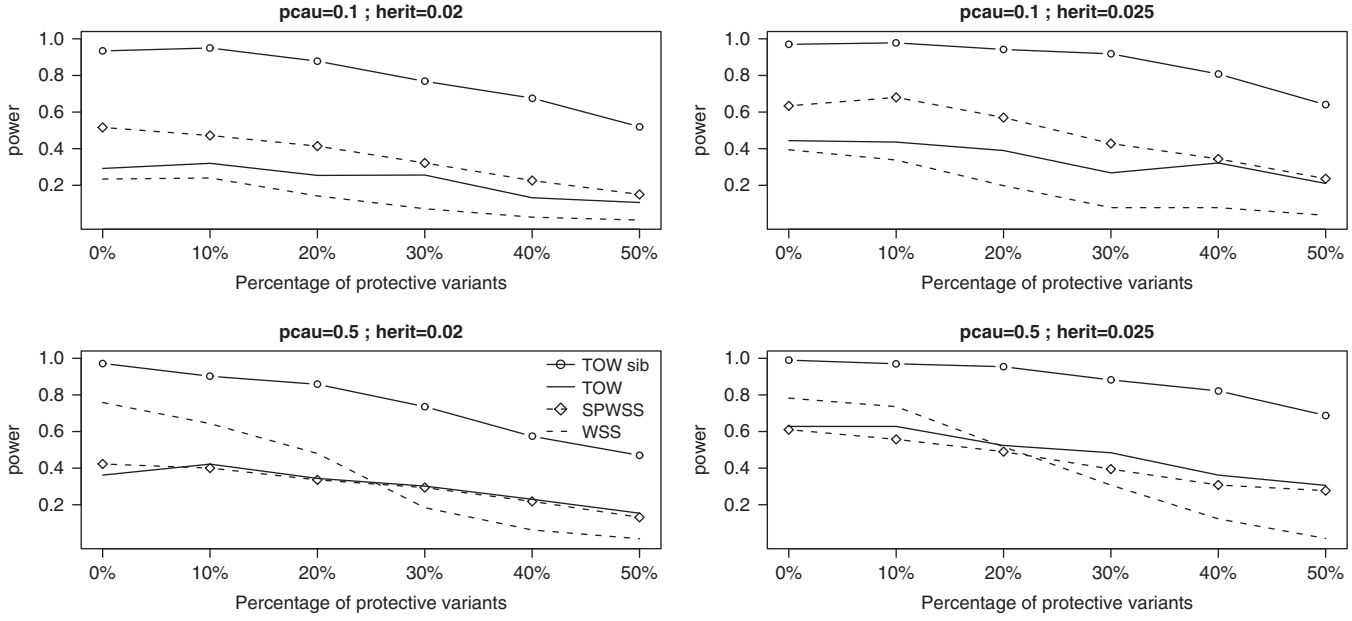

Figure 3 Powers as a function of percentage of protective variants. TOW and WSS are based on 1000 unrelated cases and 1000 unrelated controls. SPWSS is based on 1000 unrelated controls, 600 unrelated cases, and 200 affected sib-pairs. TOW-sib is based on 1000 unrelated controls and 500 affected sib-pairs. pcau denotes the percentage of causal variants in rare variants; herit denotes the total heritability. The power is evaluated at a significance level of 0.001.

DISCUSSION

There is increasing interest to detect associations between rare variants and complex traits. Recently, several statistical methods for detecting rare variant associations by jointly considering multiple variants in a genomic region have been developed for unrelated individuals. However, statistical methods for detecting rare variant associations under family-based designs have not received as much attention as methods for unrelated individuals, although family-based designs have been shown to improve power to detect rare variants.[28,29] Motivated by the facts that rare disease variants will be enriched in family data[33] and a large number of affected sib-pairs for a variety of diseases has been collected by traditional linkage studies,

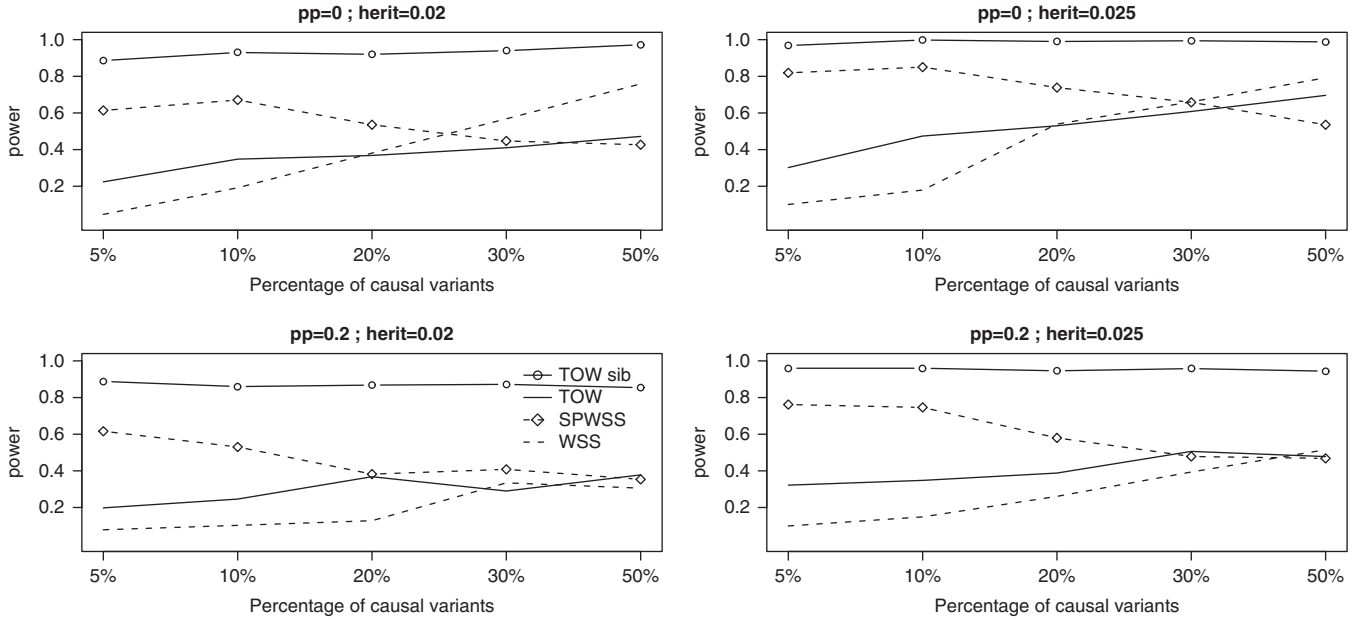

Figure 4 Powers as a function of percentage of causal variants. TOW and WSS are based on 1000 unrelated cases and 1000 unrelated controls. SPWSS is based on 1000 unrelated controls, 600 unrelated cases, and 200 affected sib-pairs. TOW-sib is based on 1000 unrelated controls and 500 affected sib-pairs. pp denotes the percentage of protective variants in causal variants; herit denotes the total heritability. The power is evaluated at a significance level of 0.001.

we develop TOW-sib to detect associations between the optimal combination of rare variants in a genomic region and complex traits based on affected sib-pairs and unrelated individuals. TOW-sib is robust to the directions of the effects of causal variants and is also relatively robust to the number of neutral variants. The proposed method does not require a MAF filtering threshold and can be applied to genomic regions that contain both rare and common variants. Our simulations demonstrated that TOW-sib using affected sib-pairs can be dramatically more powerful than the methods based on unrelated individuals and the existing methods based on affected sib-pairs.

Although TOW-sib is derived under the assumption that variants are independent, our simulation results show that TOW-sib is still a valid test when variants are in LD. Our simulations for type I error evaluation are based on the LD structures of genes 1–4 and, in each gene, there are variants in strong LD (Supplementary Tables 1–4). The correct type I error rates of TOW-sib in our simulations (Tables 2 and 3) indicate that this test is valid even if variants are in LD.

The current version of TOW-sib cannot adjust for covariates. It is possible to extend TOW-sib to be able to adjust for covariates. Denote z_{ji}, z_{ai}, and z_{ci} as the covariates of the j^{th} individual in the i^{th} sib-pair, the i^{th} cases, and the i^{th} controls, respectively. With covariates, the retrospective likelihood can be written as

$$L = \prod_{i=1}^{n_s} \Pr(g_{1i}, g_{2i} \mid D = 1, D{=}1, z_{1i}, z_{2i}) \prod_{i=1}^{n_a} \Pr(g_{ai} \mid D = 1, z_{ai}) \prod_{i=1}^{n_c} \Pr(g_{ci} \mid D = 0, z_{ci})$$

Let $\Pr(D \mid g, z) = e^{\alpha + x\beta + z^T\gamma}$, where x represents the combination of genotypic scores of the genotype g and z denotes covariates. Based on this likelihood, we can derive a score test statistic. However, the details of adjusting for covariates in TOW-sib need further investigation.

TOW-sib uses the optimal data-driven weights. TOW-sib belongs to quadratic tests and thus is robust to the directions of the effects of causal variants. We can use other weights. For example, in the score test statistic $T(w_1,\ldots,w_M)$, we can use the weights suggested by Madsen and Browning,[12] that is, $w_m = 1/\sqrt{p_m(1-p_m)}$, where p_m is the estimated MAF with pseudo-counts at the m^{th} variant. We call the score test $T(w_1,\ldots,w_M)$ with $w_m = 1/\sqrt{p_m(1-p_m)}$ WSS-sib. WSS-sib belongs to burden tests. When most of the rare variants are causal and the directions of the effects of causal variants are all the same, WSS-sib can outperform TOW-sib; otherwise, TOW-sib should outperform WSS-sib. To increase the robustness of the tests, we can also construct combined tests by combining information from TOW-sib and WSS-sib. One thing we want to make clear is the term 'optimal weight'. The optimal weight in this paper only means that the selected weight makes the score test statistic maximum, it does not mean that the selected weight makes the score test to have the maximum power.

In this study, we estimate $a = \frac{e^{\alpha}}{1-e^{\alpha}}$ based on the full likelihood. We can also use other estimates of a. Different estimates do not affect type I error, but do affect power. Our simulations (results not shown) show that the MLE of a based on the full likelihood is a good choice. We compare our proposed method with two methods based on the case/control design to see if the affected sib-pair design is more powerful than the case/control design. This is our main purpose. We also compare our proposed method with one of the existing methods that are applicable to the affected sib-pair design. Although several methods[28,29] developed recently are applicable to the affected sib-pair design, we only choose SPWSS[28] to compare with because SPWSS is most relevant to our proposed method.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

Research reported in this article was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number R03 HG006155. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The Genetic Analysis workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences.
Preparation of the Genetic Analysis Workshop 17 Simulated Exome Data Set was supported in part by NIH R01 MH059490 and used sequencing data from the 1000 Genomes Project (http://www.1000genomes.org).

1 McCarthy MI, Abecasis GR, Cardon LR *et al*: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008; **9**: 356–369.
2 Manolio TA, Collins FS, Cox NJ *et al*: Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.
3 Marini NJ, Gin J, Ziegle J *et al*: The prevalence of folate-remedial MTHFR enzyme variants in humans. *Proc Natl Acad Sci USA* 2008; **105**: 8055–8060.
4 Ji W, Foo JN, O'Roak BJ *et al*: Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet* 2008; **40**: 592–599.
5 Cohen JC, Pertsemlidis A, Fahmi S *et al*: Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma lowdensity lipoprotein levels. *Proc Natl Acad Sci USA* 2006; **103**: 1810–1815.
6 Nejentsev S, Walker N, Riches D, Egholm M, Todd JA: Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* 2009; **324**: 387–389.
7 Zhu X, Feng T, Li Y, Lu Q, Elston RC: Detecting rare variants for complex traits using family and unrelated data. *Genet Epidemiol* 2010; **34**: 171–187.
8 Andrés AM, Clark AG, Shimmin L, Boerwinkle E, Sing CF, Hixson JE: Understanding the accuracy of statistical haplotype inference with sequence data of known phase. *Genetic Epi* 2007; **31**: 659–671.
9 Metzker ML: Sequencing technologies – the next generation. *Nat Rev Genet* 2010; **11**: 31–46.
10 Morgenthaler S, Thilly WG: A strategy to discover genes that carry multiallelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res* 2007; **615**: 28–56.
11 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**: 311–321.
12 Madsen BE, Browning SR: A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 2009; **5**: e1000384.
13 Price AL, Kryukov GV, de Bakker PI *et al*: Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 2010; **86**: 832–838.
14 Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zollner S: Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am J Hum Genet* 2010; **87**: 604–617.
15 Wu M, Lee S, Cai T, Li Y, Boehnke M, Lin X: Rare variant association testing for sequencing data using the sequence kernel association test (SKAT). *Am J Hum Genet* 2011; **89**: 82–93.
16 Lee S, Emond MJ, Bamshad MJ *et al*: Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012; **91**: 224–237.
17 Sha Q, Wang X, Wang X, Zhang S: Detecting association of rare and common variants by testing an optimally weighted combination of variants. *Genet Epidemiol* 2012; **36**: 561–571.
18 Derkach A, Lawless J, Sun L: Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genetic Epi* 2012; **37**: 110–121.
19 Neale BM, Rivas MA, Voight BF *et al*: Testing for an unusual distribution of rare variants. *PLoS Genet* 2011; **7**: e1001322.
20 Han F, Pan W: A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* 2010; **70**: 42–54.
21 Hoffmann TJ, Marini NJ, Witte JS: Comprehensive approach to analyzing rare genetic variants. *PLoS One* 2010; **5**: e13584.
22 Lin D-Y, Tang Z-Z: A general framework for detecting disease associations with rare variants i n sequencing studies. *Am J Hum Genet* 2011; **89**: 354–367.
23 Yi N, Zhi D: Bayesian analysis of rare variants in genetic association studies. *Genet Epidemiol* 2011; **35**: 57–69.
24 Sha Q, Wang S, Zhang S: Adaptive clustering and adaptive weighting methods to detect disease associated rare variants. *Eu J Hum Genet* 2013; **21**: 332–337.
25 Shi G, Rao D: Optimum designs for next-generation sequencing to discover rare variants for common complex disease. *Genetic Epidemiol* 2011; **35**: 572–579.
26 Fang S, Sha Q, Zhang S: Two adaptive weighting methods to test for rare variant associations in family-based designs. *Genet Epidemiol* 2012; **36**: 499–507.
27 Liu D, Leal S: A unified framework for detecting rare variant quantitative trait associations in pedigree and unrelated individuals via sequence data. *Hum Hered* 2012; **73**: 105–122.
28 Feng T, Elston R, Zhu X: Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS). *Genet Epidemiol* 2011; **35**: 398–409.
29 Zhu Y, Xiong M: Family-based association studies for next-generation sequencing. *Am J Hum Genet* 2012; **90**: 1028–1045.
30 Schaid DJ: General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epidemiol* 1996; **13**: 423–449.
31 Guo W, Lin S: Generalized linear modeling with regularization for detecting common disease rare haplotype association. *Genet Epidemiol* 2009; **33**: 308–316.
32 Scheet P, Stephens M: A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 2006; **78**: 629–644.
33 Feng T, Zhu X: Genome-wide searching of rare genetic variants in WTCCC data. *Hum Genet* 2010; **128**: 269–280.

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

APPENDIX A

Score Test Statistic

Using notations in the Method section, from Equation (1), the log retrospective likelihood is given by

$$\log L = \sum_{i=1}^{n_s}((x_{1i}+x_{2i})\beta + \log \Pr(g_{1i},g_{2i})) - n_s \log \sum_{g_1^*,g_2^*} e^{(x_1^*+x_2^*)\beta}\Pr(g_1^*,g_2^*) + \sum_{i=1}^{n_a}(x_{ai}\beta + \log\Pr(g_{ai})) - n_a \log \sum_{g^*} e^{x^*\beta}\Pr(g^*)$$
$$+ \sum_{i=1}^{n_c}(\log(1-e^{\alpha+x_{ci}\beta}) + \log\Pr(g_{ci})) - n_c \log(1-\sum_{g^*} e^{\alpha+x^*\beta}\Pr(g^*)).$$

Then,

$$\frac{\partial \log L}{\partial \beta}=\sum_{i=1}^{n_s}(x_{1i}+x_{2i})-n_s\frac{\sum\limits_{g_1^*,g_2^*}(x_1^\star+x_2^\star)e^{(x_1^\star+x_2^\star)\beta}\Pr(g_1^*,g_2^*)}{\sum\limits_{g_1^*,g_2^*}e^{(x_1^\star+x_2^\star)\beta}\Pr(g_1^*,g_2^*)}+\sum_{i=1}^{n_a}x_{ai}-n_a\frac{\sum\limits_{g^\star}x^\star e^{x^\star\beta}\Pr(g^\star)}{\sum\limits_{g^\star}e^{x^\star\beta}\Pr(g^\star)}-\sum_{i=1}^{n_c}\frac{x_{ci}e^{\alpha+x_{ci}\beta}}{1-e^{\alpha+x_{ci}\beta}}+n_c\frac{\sum\limits_{g^\star}x^\star e^{\alpha+x^\star\beta}\Pr(g^\star)}{1-\sum\limits_{g^\star}e^{\alpha+x^\star\beta}\Pr(g^\star)},$$

$$\begin{aligned}\frac{\partial^2 \log L}{\partial \beta^2}=&-n_s\frac{\sum\limits_{g_1^*,g_2^*}e^{(x_1^\star+x_2^\star)\beta}\Pr(g_1^*,g_2^*)\sum\limits_{g_1^*,g_2^*}(x_1^\star+x_2^\star)^2e^{(x_1^\star+x_2^\star)\beta}\Pr(g_1^*,g_2^*)-\left(\sum\limits_{g_1^*,g_2^*}(x_1^\star+x_2^\star)e^{(x_1^\star+x_2^\star)\beta}\Pr(g_1^*,g_2^*)\right)^2}{\left(\sum\limits_{g_1^*,g_2^*}e^{(x_1^\star+x_2^\star)\beta}\Pr(g_1^*,g_2^*)\right)^2}\\&-n_a\frac{\sum\limits_{g^\star}x^\star 2e^{x^\star\beta}\Pr(g^\star)\sum\limits_{g^\star}e^{x^\star\beta}\Pr(g^\star)-\left(\sum\limits_{g^\star}x^\star e^{x^\star\beta}\Pr(g^\star)\right)^2}{\left(\sum\limits_{g^\star}e^{x^\star\beta}\Pr(g^\star)\right)^2}-\sum_{i=1}^{n_c}\frac{x_{ci}^2e^{\alpha+x_{ci}\beta}\left(1-e^{\alpha+x_{ci}\beta}\right)+x_{ci}^2e^{2\alpha+2x_{ci}\beta}}{\left(1-e^{\alpha+x_{ci}\beta}\right)^2}\\&+n_c\frac{\sum\limits_{g^\star}x^{\star 2}e^{\alpha+x^\star\beta}\Pr(g^\star)\left(1-\sum\limits_{g^\star}e^{\alpha+x^\star\beta}\Pr(g^\star)\right)+\left(\sum\limits_{g^\star}x^\star e^{\alpha+x^\star\beta}\Pr(g^\star)\right)^2}{\left(1-\sum\limits_{g^\star}e^{\alpha+x^\star\beta}\Pr(g^\star)\right)^2},\end{aligned}$$

$$\begin{aligned}-E\frac{\partial^2 \log L}{\partial\beta\partial\alpha}\mid_{\beta=0,\alpha=\hat\alpha}&=\sum_{i=1}^{n_c}E\frac{x_{ci}e^{\hat\alpha}(1-e^{\hat\alpha})+x_{ci}e^{2\hat\alpha}}{(1-e^{\hat\alpha})^2}-n_c\frac{\sum\limits_{g^\star}x^\star e^{\hat\alpha}\Pr(g^\star)(1-\sum\limits_{g^\star}e^{\hat\alpha}\Pr(g^\star))+\sum\limits_{g^\star}x^\star e^{\hat\alpha}\Pr(g^\star)\sum\limits_{g^\star}e^{\hat\alpha}\Pr(g^\star)}{(1-\sum\limits_{g^\star}e^{\hat\alpha}\Pr(g^\star))^2}\\&=n_c\hat a(1+\hat a)E(x_{c1})-n_c\hat a(1+\hat a)E(x_{c1})=0\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \log L}{\partial\beta\partial p_m}=&-n_s\frac{\sum\limits_{g_1^*,g_2^*}(x_1^\star+x_2^\star)e^{(x_1^\star+x_2^\star)\beta}\frac{\partial}{\partial p_m}\Pr(g_1^*,g_2^*)\sum\limits_{g_1^*,g_2^*}e^{(x_1^\star+x_2^\star)\beta}\Pr(g_1^*,g_2^*)}{(\sum\limits_{g_1^*,g_2^*}e^{(x_1^\star+x_2^\star)\beta}\Pr(g_1^*,g_2^*))^2}+n_s\frac{\sum\limits_{g_1^*,g_2^*}(x_1^\star+x_2^\star)e^{(x_1^\star+x_2^\star)\beta}\Pr(g_1^*,g_2^*)\sum\limits_{g_1^*,g_2^*}e^{(x_1^\star+x_2^\star)\beta}\frac{\partial}{\partial p_m}\Pr(g_1^*,g_2^*)}{(\sum\limits_{g_1^*,g_2^*}e^{(x_1^\star+x_2^\star)\beta}\Pr(g_1^*,g_2^*))^2}\\&-n_a\frac{\sum\limits_{g^\star}x^\star e^{x^\star\beta}\frac{\partial}{\partial p_m}\Pr(g^\star)\sum\limits_{g^\star}e^{x^\star\beta}\Pr(g^\star)-\sum\limits_{g^\star}x^\star e^{x^\star\beta}\Pr(g^\star)\sum\limits_{g^\star}e^{x^\star\beta}\frac{\partial}{\partial p_m}\Pr(g^\star)}{(\sum\limits_{g^\star}e^{x^\star\beta}\Pr(g^\star))^2}\\&+n_c\frac{\sum\limits_{g^\star}x^\star e^{\alpha+x^\star\beta}\frac{\partial}{\partial p_m}\Pr(g^\star)(1-\sum\limits_{g^\star}e^{\alpha+x^\star\beta}\Pr(g^\star))+\sum\limits_{g^\star}x^\star e^{\alpha+x^\star\beta}\Pr(g^\star)(\sum\limits_{g^\star}e^{\alpha+x^\star\beta}\frac{\partial}{\partial p_m}\Pr(g^\star))}{(1-\sum\limits_{g^\star}e^{\alpha+x^\star\beta}\Pr(g^\star))^2}\end{aligned}$$

Let $P=(p_1,\ldots,p_M)^T$ and $\hat P=(\hat p_1,\ldots,\hat p_M)^T$. Note that $\sum\limits_{g_1^*,g_2^*}\Pr(g_1^*,g_2^*)=\sum\limits_{g^\star}\Pr(g^\star)=1$ and

$$\sum_{g_1^*,g_2^*}\frac{\partial}{\partial p_m}\Pr(g_1^*,g_2^*)=\sum_{g^\star}\frac{\partial}{\partial p_m}\Pr(g^\star)=0$$

for $m=1,\ldots,M$. We have

$$\frac{\partial \log L}{\partial \beta}\mid_{\beta=0,\alpha=\hat\alpha,P=\hat P}=\sum_{i=1}^{n_s}(x_{1i}+x_{2i})+\sum_{i=1}^{n_a}x_{ai}-\hat a\sum_{i=1}^{n_c}x_{ci},\ -E\frac{\partial^2 \log L}{\partial\beta^2}\mid_{\beta=0,\alpha=\hat\alpha,P=\hat P}=(6n_s+2n_a+2n_c\hat a^2)\sum_{m=1}^{M}w_m^2\hat p_m\hat q_m$$

and

$$-E\frac{\partial^2 \log L}{\partial\beta\partial p_m}\mid_{\beta=0,\alpha=\hat\alpha,P=\hat P}=0$$

Similarly, we have

$$\frac{\partial \log L}{\partial \alpha}\mid_{\beta=0,\alpha=\hat\alpha,P=\hat P}=\frac{\partial \log L}{\partial p_m}\mid_{\beta=0,\alpha=\hat\alpha,P=\hat P}=-E\frac{\partial^2 \log L}{\partial\alpha\partial p_m}\mid_{\beta=0,\alpha=\hat\alpha,P=\hat P}=-E\frac{\partial^2 \log L}{\partial\alpha\partial\beta}\mid_{\beta=0,\alpha=\hat\alpha,P=\hat P}=0.$$

Let $U=\sum_{i=1}^{n_s}(x_{1i}+x_{2i})+\sum_{i=1}^{n_a}x_{ai}-\hat a\sum_{i=1}^{n_c}x_{ci}$, $V=(6n_s+2n_a+2n_c\hat a^2)\sum_{m=1}^{M}w_m^2\hat p_m\hat q_m$, $U^\star=(U,0,0)^T$ denote the score vector, and I denote the information matrix. Then, the score test statistic is given by

$$T=U^{\star T}I^{-1}U^\star=\frac{U^2}{V}.$$

APPENDIX B

Expectation-maximization Algorithm to Estimate Allele Frequency Based on Sib-pairs and Unrelated Individuals

Consider a variant with two alleles. Let B denote the minor allele and p denote the frequency of allele B. We use the following notations.

N: the number of unrelated individuals
N_f: the number of sib-pairs
n: the number of minor alleles in genotypes of the N unrelated individuals
n_{ij}: the number of sib-pairs with genotype pair (i,j) or (j,i)
n_{ij}^k: the number of sib-pairs with genotype pair (i,j) or (j,i) and the pair of genotypes has k alleles IBD

E-step:

$$n_{00}^0 = \frac{q^2}{(1+q)^2} n_{00}, n_{00}^1 = \frac{2q}{(1+q)^2} n_{00}, n_{00}^2 = \frac{1}{(1+q)^2} n_{00}$$

$$n_{01}^0 = \frac{q}{1+q} n_{01}, n_{01}^1 = \frac{1}{1+q} n_{01}, n_{01}^2 = 0$$

$$n_{02}^0 = n_{02}, n_{02}^1 = n_{02}^0 = 0$$

$$n_{11}^0 = \frac{pq}{1+pq} n_{11}, n_{11}^1 = \frac{1}{2(1+pq)} n_{11}, n_{11}^2 = \frac{1}{2(1+pq)} n_{11}$$

$$n_{12}^0 = \frac{p}{1+p} n_{12}, n_{12}^1 = \frac{1}{1+p} n_{12}, n_{12}^2 = 0$$

$$n_{22}^0 = \frac{p^2}{(1+p)^2} n_{22}, n_{22}^1 = \frac{2p}{(1+p)^2} n_{22}, n_{22}^2 = \frac{1}{(1+p)^2} n_{22}$$

M-step: $p = \frac{m_2}{m_1+m_2}$,

where

$$m_1 = 2N - n + 4n_{00}^0 + 3n_{00}^1 + 2n_{00}^2 + 3n_{01}^0 + 2n_{01}^1 + 2n_{02}^0 + 2n_{11}^0 + n_{11}^1 + n_{11}^2 + n_{12}^0 + n_{12}^1,$$

$$m_2 = n + n_{01}^0 + n_{01}^1 + 2n_{02}^0 + 2n_{11}^0 + n_{11}^1 + n_{11}^2 + 3n_{12}^0 + 2n_{12}^1 + 4n_{22}^0 + 3n_{22}^1 + 2n_{22}^2.$$

APPENDIX C

Mean and Variance of TOW-sib

It is easy to know that $\mu_{\mathrm{TOW-sib}} = E(T_{\mathrm{TOW-sib}}) = \sum_{m=1}^{M} I\{p_m > 0\}$. In the following, we will calculate the variance of $T_{\mathrm{TOW-sib}}$.

Let g_1 and g_2 denote genotypes of a sib-pair, $x = g_1 + g_2$, and p ($q = 1-p$) denote the MAF. Using the distribution given by Table 1, we have $E(g_1 - 2p)^4 = 2pq$, $\mathrm{var}(x) = 6\ pq$, and $E(x-4p)^4 = 6pq(pq+3)$.

We know that $\mathrm{var}(T_{\mathrm{TOW-sib}}) = \mathrm{var}(\sum_{m=1}^{M} T_m) = \sum_{m=1}^{M} \mathrm{var}(T_m) + \sum_{m \neq k} \mathrm{cov}(T_m, T_k)$, $\mathrm{var}(T_m) = \frac{E(u_m^4)}{v_m^2} - 1$, and $\mathrm{cov}(T_m, T_k) = \frac{E(u_m^2 u_k^2)}{v_m v_k} - 1$.

Let $n = n_s + n_a + n_c$, $x_i = g_{1im} + g_{2im} - 4p_m$ for $i = 1,\ldots,n_s$, $x_{i+n_s} = g_{aim} - 2p_m$ for $i = 1,\ldots,n_a$, $x_{i+n_s+n_a} = -\hat{a}(g_{cim} - 2p_m)$ for $i = 1,\ldots,n_c$, and y_i is similarly defined for the k^{th} variant as x_i for the m^{th} variant.

We can calculate the variance of $T_{\mathrm{TOW-sib}}$ if we note that

$$E(u_m^4) = E(x_1 + x_2 + \ldots + x_n)^4 = E(\sum_{i=1}^{n} x_i^4 + \sum_{i \neq j} 3x_i^2 x_j^2) = n_s\sigma_s^4 + (n_a + n_c\hat{a}^4)\sigma_c^4 + 3[(6n_s + 2n_a + 2n_c\hat{a}^2)^2 - 36n_s - 4n_a - 4n_c\hat{a}^4]p^2q^2$$

$$= n_s(18pq) + (n_a + n_c\hat{a}^4)2pq + 3[(6n_s + 2n_a + 2n_c\hat{a}^2)^2 - 34n_s - 4n_a - 4n_c\hat{a}^4]p^2q^2 = (9n_s + n_a + n_c\hat{a}^4)2pq$$

$$+ 3[(6n_s + 2n_a + 2n_c\hat{a}^2)^2 - 34n_s - 4n_a - 4n_c\hat{a}^4]p^2q^2 = 2pqN_1 + 3p^2q^2N_2,$$

where $N_1 = 9n_s + n_a + n_c\hat{a}^4$, $N_2 = (6n_s + 2n_a + 2n_c\hat{a}^2)^2 - 34n_s - 4n_a - 4n_c\hat{a}^4$, and $\hat{a} = \frac{2n_s + n_a}{n_c}$;

$$E(u_m^2 u_k^2) = E((\sum_{i=1}^{n} x_i)^2 (\sum_{i=1}^{n} y_i)^2) = E(\sum_{i=1}^{n} x_i^2 y_i^2 + \sum_{i \neq j} x_i^2 y_j^2 + 2\sum_{i \neq j} x_i y_i x_j y_j) = n_s E(x_1^2 y_1^2) + (n_a + n_c\hat{a}^4)E(x_n^2 y_n^2)$$

$$+ (6n_s + 2n_a + 2n_c\hat{a}^2)^2 p_m q_m p_k q_k + 2((4n_s + n_a + n_c\hat{a}^2)^2 - 16n_s - n_a - n_c\hat{a}^4)\mathrm{cov}^2(x_n, y_n) = n_s E(x_1^2 y_1^2) + (n_a + n_c\hat{a}^4)E(x_n^2 y_n^2)$$

$$+ N_3 p_m q_m p_k q_k + N_4 \mathrm{cov}^2(x_n, y_n)$$

where $N_3 = (6n_s + 2n_a + 2n_c\hat{a}^2)^2$, $N_4 = 2((4n_s + n_a + n_c\hat{a}^2)^2)$, $E({x_1}^2{y_1}^2)$ is estimated with $\frac{1}{n_s}\sum_{i=1}^{n_s} x_i^2 y_i^2$, $E({x_n}^2{y_n}^2)$ is estimated with $\frac{1}{n_a + n_c}\sum_{i=1+n_s}^{n} x_i^2 y_i^2$, and $\mathrm{cov}(x_n, y_n)$ is estimated with $\frac{1}{n_a + n_c}\sum_{i=1+n_s}^{n} x_i y_i$.