Joint Analysis for Genome-Wide Association Studies in Family-Based Designs

Qiuying Sha¹, Zhaogong Zhang¹,², Shuanglin Zhang¹*

¹Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan, United States of America, ²School of Computer Science and Technology, Heilongjiang University, Harbin, China

Abstract

In family-based data, association information can be partitioned into the between-family information and the within-family information. Based on this observation, Steen et al. (Nature Genetics. 2005, 683–691) proposed an interesting two-stage test for genome-wide association (GWA) studies under family-based designs which performs genomic screening and replication using the same data set. In the first stage, a screening test based on the between-family information is used to select markers. In the second stage, an association test based on the within-family information is used to test association at the selected markers. However, we learn from the results of case-control studies (Skol et al. Nature Genetics. 2006, 209–213) that this two-stage approach may not be optimal. In this article, we propose a novel two-stage joint analysis for GWA studies under family-based designs. For this joint analysis, we first propose a new screening test that is based on the between-family information and is robust to population stratification. This new screening test is used in the first stage to select markers. Then, a joint test that combines the between-family information and within-family information is used in the second stage to test association at the selected markers. By extensive simulation studies, we demonstrate that the joint analysis always results in increased power to detect genetic association and is robust to population stratification.

Introduction

Currently, the family-based association tests such as the TDT and its extensions [1–6] are still the most commonly used methods to detect disease susceptibility loci in family-based GWA studies. This kind of methods uses the within-family information, but not the between-family information. The reason is that methods used between-family information may be subject to bias caused by population stratification. Recently, based on the observation that the association information in the family sample can be split into the between-family component and the within-family component, Steen et al. [7] proposed a two-stage test for family-based GWA studies. We call this method Family-based Two-Stage Approach (FTSA). In the first stage of FTSA, a test based on between-family information is used to screen markers, that is, choose R “top” markers. In the second stage of FTSA, a family-based association test based on within-family information is used to test the R selected markers for association. FTSA is robust to population stratification because the association is determined by the family-based association test in the second stage. Furthermore, since the statistic used in the first stage is statistically independent of that in the second stage, the overall significance level of the algorithm does not need to be adjusted for the first stage. In the following discussion, we call the tests used in the first stage and in the second stage screening test and association test, respectively.

In case-control studies, several authors have proposed a two-stage design which utilizes two independent samples [8,9]. The first stage that uses the first sample is to screen and select SNPs for association tests. In the second stage, the association tests are conducted on the selected SNPs by using the second sample, so that the number of association tests is diminished and the correction for multiple testing is less severe. Recently, Skol et al. [10] pointed out that joint analysis in which the test used in the second stage is the combination of the two tests based on the two samples is more powerful than replication-based analysis in which the test used in the second stage is based on the second sample only. There are some similarities between FTSA and the two-stage approach in case-control studies. One problem hindering us from doing joint analysis in FTSA as Skol et al. [10] did for the two-stage approach in case-control studies. One problem hindering us from doing joint analysis in FTSA as Skol et al. [10] did for the two-stage approach in case-control studies. Can we do joint analysis in FTSA as Skol et al. [10] did for the two-stage approach in case-control studies. Can we do joint analysis in FTSA as Skol et al. [10] did for the two-stage approach in case-control studies. Can we do joint analysis in FTSA as Skol et al. [10] did for the two-stage approach in case-control studies.
Joint Analysis for GWAS in Family-Based Designs

Consider a GWA study of \( n \) nuclear families with \( n_i \) children in the \( i^{th} \) family \((i=1,2,\ldots,n)\) and \( L \) markers have been genotyped for each sampled individual. For the \( i^{th} \) family, we use \( y_{ik} \) and \( x_{ik} \) to denote the trait values and genotypic scores at the candidate locus of the \( k^{th} \) member in the \( i^{th} \) family \((k=1,2,\ldots,L)\). We calculate the residual correlation. The PC approach is much simpler and computationally faster than the SA approach and is estimated using genotypes at genomic markers to represent the genetic background for unrelated individuals and using the genetic background to control for population stratification in population-based association studies [20–22]. We will use this idea to construct a test based on between-family information and incorporating parental phenotypes such that this test is robust to population stratification.

To construct the test, we first randomly choose \( l \) markers from the \( L \) markers in GWA panel as genomic markers. For the \( i^{th} \) family, let \( X_{ik} = (x_{ik1}, \ldots, x_{ikL})^T \) denote multiple marker genotypic scores at the \( l \) randomly chosen markers of the \( k^{th} \) member in the \( i^{th} \) family \((k=1,2,\ldots,L)\). We perform a PC analysis to summarize the genotype data at genomic markers. Because our data are family data, a naive PC analysis with all available data will result in biased directions of maximum variability for the data. Thus, the PC analysis is applied to only the parents in each family.

Let \( \Sigma = \sum_{i=1}^{n} \sum_{k=1}^{L} (X_{ik} - \bar{X})(X_{ik} - \bar{X})^T \) denote the variance-covariance matrix of the genotype data for all of the \( 2n \) parents, where, \( \bar{X} \) is the overall mean of parental genotypic scores. Let \( e_j \) be the \( j^{th} \) eigenvector corresponding to the \( j^{th} \) largest eigenvalue of \( \Sigma \) for \( j=1, \ldots, L \). Then, the \( j^{th} \) PC for the \( k^{th} \) member of the \( i^{th} \) family is given by \( t_{ijk} = e_j^T (X_{ik} - \bar{X}) \) Here we consider only the first \( K \) PCs (in this study, we use \( K=10 \)). Because the PCs represent the genetic background information, we adjust both the trait and genotypic at candidate loci for this genetic background information by applying linear regression [23].

That is,

\[
\hat{y}_{ik} = \beta_0 + \beta_1 t_{1ik} + \cdots + \beta_K t_{Kik} + \varepsilon_{ik},
\]

and

\[
x_{ik} = a_{i0} + a_{i1} t_{1ik} + \cdots + a_{iK} t_{Kik} + \tau_{ik},
\]

where \( \varepsilon_{ik} \) and \( \tau_{ik} \) are random errors for \( i=1, \ldots, n \) and \( k=1,2,\ldots,L \). Let \( \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K, \) and \( \hat{a}_{i0}, \hat{a}_{i1}, \ldots, \hat{a}_{iK} \) be the least square estimators of \( \beta_0, \beta_1, \ldots, \beta_K \) and, \( a_{i0}, a_{i1}, \ldots, a_{iK} \), respectively. The residuals of the trait values and genotypic scores at the candidate locus for parents and children of the \( i^{th} \) family are calculated by

\[
\hat{y}_{ik} = y_{ik} - \hat{\beta}_0 - \hat{\beta}_1 t_{1ik} - \cdots - \hat{\beta}_K t_{Kik}
\]

and

\[
x_{ik}^* = x_{ik} - \hat{a}_{i0} - \hat{a}_{i1} t_{1ik} - \cdots - \hat{a}_{iK} t_{Kik},
\]

where \( i=1, \ldots, n \) and \( k=1,2,\ldots,L \). We can consider \( y_{ik}^* \) and

\[
T_{screen} = \sum_{i=1}^{n} \frac{U_i}{\sqrt{\sum_{i=1}^{n} U_i^2}}.
\]

where \( U_i = (\bar{x}_{i} - \bar{X})(y_i - \bar{Y}) + \sum_{k=1}^{K} (X_{ik} - \bar{X})(Y_{ik} - \bar{Y}); \bar{x}_{i} = \frac{1}{L} \sum_{k=1}^{L} x_{ik}, y_i = \frac{1}{n} \sum_{k=1}^{n} y_{ik}; \bar{X} \) and \( \bar{Y} \) are the overall means of genotypic scores and trait values, respectively. Feng et al. [27] have shown that this test is more powerful than one not incorporating parental phenotypes and is independent of family-based association tests based on within-family information. However, this test may be subject to bias caused by population stratification and thus cannot do joint analysis by combining \( T_{screen} \) with the test statistic of a family-based association test because the combined test may also be subject to bias caused by population stratification. We previously suggested using the PCs of genotypes at genomic markers to represent the genetic background of unrelated individuals and using the genetic background to control for population stratification in population-based association studies [20–22].

In this article, we propose a novel approach to do joint analysis within the framework of FTSA. We first perform PC analysis based on parental genotypes at a set of genomic markers and then use a PC approach to eliminate any effect of population stratification both in genotypes at the candidate marker and trait values for all family members. A screening test is then constructed based on adjusted between-family information and parental trait values. We use this screening test which is robust to population stratification to select markers at the first stage. In the second stage, we do joint analysis i.e. use a test that is a combination of the screening test and the association test to test association at the selected markers. The joint analysis is robust to population stratification because both the screening test used at the first stage and the joint test used at the second stage are robust to population stratification. We evaluate the performance of our joint analysis approach by simulation studies under a variety of population admixture models. Our simulation studies show that the proposed joint analysis approach is robust to population stratification and is consistently more powerful than FTSA proposed by Steen et al. [7].

Methods

Consider a GWA study of \( n \) nuclear families with \( n_i \) children in the \( i^{th} \) family \((i=1,2,\ldots,n)\) and \( L \) markers have been genotyped for each sampled individual. For the \( i^{th} \) family, we use \( y_{ik} \) and \( x_{ik} \) to denote the trait values and genotypic scores at the candidate marker of the \( k^{th} \) member in the \( i^{th} \) family \((k=1,2,\ldots,L)\), where genotypic score is the number of copies of minor allele.

Screening Test

We assume that the parental phenotypes are available. In this case, by incorporating parental phenotypes, Feng et al. [27] proposed a screening test statistic

\[
T_{screen} \frac{U_i}{\sqrt{\sum_{i=1}^{n} U_i^2}}.
\]
and, $x_{ik}$ as the trait value and genotypic score of the $k^{th}$ member in the, $i^{th}$ family after adjusted for population stratification.

Based on the adjusted trait values and genotypic scores, we propose the following screening test, called admixture screening (Ascreen) test,

$$T_{Ascreen} = \sum_{i=1}^{n} U_i / \sqrt{\sum_{i=1}^{n} U_i^2}$$  \hspace{1cm} (2)

where $U_i = (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y}) + \sum_{k=1}^{n_i} (x_{ik} - \bar{x})(y_{ik} - \bar{y})$; $\bar{x}$ and $\bar{y}$ are the overall means of genotypic scores and trait values after adjusted for population stratification, respectively. Under the null hypothesis, $T_{Ascreen}$ follows a standard normal distribution.

Association Test

We use quantitative pedigree transmission disequilibrium (QPTD) as our family-based association test [28]. Using the notation given above, the association test statistic is given by

$$T_a = \sum_{i=1}^{n} V_i / \sqrt{\sum_{i=1}^{n} V_i^2},$$ \hspace{1cm} (3)

where $V_i = 1/\sigma_i^2 (x_{ik} - \bar{x})(y_{ik} - \bar{y})$. Under the null hypothesis of no association, $T_a$, asymptotically follows the standard normal distribution.

Joint Analysis

In the first stage, we select R markers with the smallest p-values of the admixture screening test. This selection means that there is a constant $C$ such that a marker is selected if $|T_{Ascreen}| > C$. In the joint analysis, a new test statistic

$$T_{joint} = T_{Ascreen} + T_a$$ \hspace{1cm} (4)

is used to test association in the second stage. Let $t_{joint}$ be the observed value of the statistic $T_{joint}$. Then, in the second stage, the p-value of the test $T_{joint}$ is given by

$$P_{joint} = \Pr (|T_{joint}| > |t_{joint}| | |T_{Ascreen}| > C).$$

Let $C_2 = |t_{joint}|$ and $T$ denote the event of $|T_{Ascreen}| > C$. Similarly to equation (2) in Skol et al. [10], we have

$$P_{joint} = \Pr (|T_{joint}| > C_2 | T)
= \int_{-\infty}^{-C_2} P(T_{joint} > C_2 | T_{Ascreen} = x) f(x | T) dx
+ \int_{C_2}^{\infty} P(T_{joint} < -C_2 | T_{Ascreen} = x) f(x | T) dx
+ \int_{-\infty}^{C_2} P(T_{joint} > C_2 | T_{Ascreen} = x) f(x | T) dx
+ \int_{C_2}^{\infty} P(T_{joint} < -C_2 | T_{Ascreen} = x) f(x | T) dx
= \Phi (C_2 - x) + \Phi (-C_2 - x) f(x | T) dx
= \int_{-\infty}^{-C} [1 - \Phi (C_2 - x) + \Phi (-C_2 - x)] f(x | T) dx
+ \int_{C}^{\infty} [1 - \Phi (C_2 - x) + \Phi (-C_2 - x)] f(x | T) dx,
$$

where $\Phi$ is the cumulative distribution function of the standard normal distribution; $f(x | T)$ is the probability density function of $T_{Ascreen}$ given that $|T_{Ascreen}| > C$ i.e. $f(x | T) = \frac{d(x)}{\sigma(x)}$ if $|x| > C$ and $|x| > C$ otherwise; $\phi(x)$ is the probability density function of the standard normal distribution. Thus, the p-value $P_{joint}$ can be calculated numerically. In summary, for the joint analysis, we first select R top markers using the admixture screening test $T_{Ascreen}$ and then test association for each of the R selected markers using the joint test $T_{joint}$. For one of the R selected markers, we declare that this marker is significant at level $\alpha$ if the p-value of the joint test $P_{joint}$ is less than $\alpha/R$.

Methods Compared

We compare the proposed joint analysis method with two other methods that is described below. One is FTSA proposed by Steen et al [7]. In FTSA, the screening test does not adjust for population stratification. In this study, we use $T_{screen}$ given in equation (1) as the screening test of FTSA. In the first stage of FTSA, R markers with the smallest p-values of the screening test $T_{screen}$ are selected. In the second stage, the association test $T_a$ given by equation (4) is used to test each of the R selected markers. A marker in the R selected markers is declared significant at level $\alpha$ if the p-value of the association test $T_a$ is less than $\alpha/R$. The other method we compare with is a method called Admixture Family-based Two-stage Approach (AFTSA) that is similar to FTSA but replaces the screening test $T_{screen}$ in FTSA by the admixture screening test $T_{Ascreen}$ given by equation (2).

Results

We used simulation studies to compare the performance of the joint analysis with FTSA and AFTSA. We also compared FTST with AFTSA to see if adjusting population stratification in the screening test can improve power of FTSA. The simulation setup used in this study was similar to that of Zhu et al [29]. We considered three sets of simulations: a homogeneous population, a structured population which contained two subpopulations, and an admixture population that mimicked African American population.

Set 1: A Homogeneous Population

In this set of simulations, we simulated samples based on the haplotype data of 120 European chromosomes (CEU) released by the HapMap project [30]. However, we used only the haplotype data on chromosome 1 at a tagging SNPs. There are 34720 SNPs in total. To generate the genotype of a parent, we generated two haplotypes that are the recombinants of the 120 HapMap chromosomes. To generate a recombinant of the 120 chromosomes, we first generated a number of crossovers across the chromosome by a Poisson process with an average of 6 crossovers per Morgan. The crossover locations were generated according to a uniform distribution. The crossover locations divided the chromosome into segments. Each segment of the recombinant was a random chosen haplotype from the HapMap chromosomes in the same segment. The offspring genotypes were generated by randomly transmitting one of the two haplotypes of the father and the mother with the crossovers occurring according to the genetic map. The LD pattern across a chromosome was generally preserved for the SNPs that are closely located.

To generate trait values under the null hypothesis, for a nuclear family with m children, let $Y_1 = (y_F, y_M)$ and
$Y_2 = (Y_1, Y_2, \ldots, Y_m)$ denote the trait values of the parents and the $m$ children. Assumed that $(Y_1, Y_2)$ followed a normal distribution with a mean vector of zero and variance-covariance matrix of

$$
\Sigma = \begin{pmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{pmatrix},
$$

where $\Sigma_{11} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$,

$$
\Sigma_{12} = \Sigma_{21}^T = \begin{pmatrix} \rho & \cdots & \rho \\ \rho & \cdots & \rho \\ \rho & \cdots & \rho \\ \end{pmatrix},
$$

and $\Sigma_{22} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}$.

This covariance structure meant that the father and mother were independent, and parents with children and children with children were correlated with the correlation coefficient of $\rho$ (in this study, we use $\rho = 0.4$). The conditional distribution of $Y_2 = (Y_1, Y_2, \ldots, Y_m)$ given parental trait values, $Y_1 = (Y_1, \ldots, Y_M)$ was a normal distribution with mean vector of $\mu_y = \Sigma_{11}^{-1} \Sigma_{12} Y_1$, and variance-covariance matrix of $\Sigma_y = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$. To generate trait values of all individuals in the family, we first generated the trait value of each parent by using a standard normal distribution. The trait values of the children can be generated by a normal distribution with a mean vector of $\mu_y$ and variance-covariance matrix of $\Sigma_y$, given the trait values of their parents.

Under the alternative hypothesis, we generated the trait values of a nuclear family with $B$ members from model $y_h = x_h \beta + \epsilon_h (b=1,2,\ldots,B)$ where $x_h$ was the numerical code of genotype $g$ at the disease locus and

$$
x_h = \begin{cases}
0, & \text{if } g = aa \\
1, & \text{if } g = aA \\
2, & \text{if } g = AA
\end{cases}
$$

and $x_h$ was

$$
x_h = \begin{cases}
0, & \text{if } g = aa \\
1, & \text{if } g = aA \\
2, & \text{if } g = AA
\end{cases}
$$

for a dominant, recessive, and additive model, respectively (a and $A$ were the two alleles at the disease locus and $A$ was the high risk allele); $\beta$ was a constant and $\epsilon_1, \ldots, \epsilon_B$ were background trait values generated under the null hypothesis using aforementioned method. The value of $\beta$ was determined by heritability $h$ and was given by

$$
\beta = \sqrt{\frac{h}{1-h(1-\rho^2)}}
$$

for a dominant, recessive, and additive model, respectively, where $p$ was the allele frequency of the high risk allele.

Set 2: A Structured Population with Two Subpopulations

In this set of simulations, we simulated samples using the haplotype data of 120 European chromosomes (CEU) and 120 African chromosomes (YRI) released by the HapMap project [30]. In these simulations, we again used only the haplotype data on chromosome 1 at the 34720 tagging SNPs. We considered that all members of a family were from one subpopulation. The genotypes can be generated in the same way as that in the simulation set 1. In this set of simulations, we sampled 70% of families from European subpopulation and 30% of families from African subpopulation. We generated the trait values of a nuclear family with $B$ members from model

$$
y_h = \begin{cases}
H2 + x_h \beta_1 + \epsilon_h (b=1,2,\ldots,B) & \text{if this family was from European subpopulation and} \\
H2 + x_h \beta_2 + \epsilon_h (b=1,2,\ldots,B) & \text{if this family was from African subpopulation}
\end{cases}
$$

where $\epsilon_1, \ldots, \epsilon_B$ were background trait values generated under the null hypothesis in simulation set 1; $x_h$ was the numerical code of genotype at the disease locus; $\mu$ was a constant that measured the difference of the average trait values between the two subpopulations; $\beta_1$ and $\beta_2$ can be determined by heritability $h$ and the relationship was given in simulation set 1. We used the same value of $h$ in the two subpopulations and thus $\beta_1$ and $\beta_2$ may be different due to the difference of allele frequencies of the high risk allele in the two subpopulations. Furthermore, we set $h = 0$ and thus $\beta_1 = \beta_2 = 0$ for generating the trait values under the null hypothesis, and $h > 0$ for generating the trait values under the alternative hypothesis.

Set 3: An Admixture Population with Two Ancestral Populations

Again, we simulated samples based on the chromosome 1 data of 120 European chromosomes and 120 African chromosomes released by the HapMap project [30]. We first generated haplotype exchange points on the chromosome among the populations by using a Poisson process, with an average of 6 crossovers per Morgan. This is equivalent to a population that has been admixed for an average of 6 generations. In each region between two exchange points, we determined which ancestral population a haplotype came from based on a distribution of admixture proportions of Africans and Europeans, which we set to (0.8, 0.2). We then applied the same method as for simulation set 1 to generate a person’s genotypes from the selected ancestral population. The method in simulation set 1 for generating offspring genotypes was also applied.

We generated the trait values of a nuclear family with $B$ members from model $y_h = \mu_b + x_h \beta_1 + \epsilon_h (b=1,2,\ldots,B)$, where $\epsilon_1, \ldots, \epsilon_B$ and $x_h$ were the same as in simulation set 2; $\lambda_b$ was European admixture proportion of the $b$th member in the family; $\mu$ and $\beta$ were constants and $\beta$ can be determined by heritability $h$. Again, we set $h = 0$ for generating the trait values under the null hypothesis, and $h > 0$ for generating the trait values under the alternative hypothesis.

In all of the three sets of simulations, we used 1000 replicated samples to evaluate the type I error rates and power and considered nuclear family with one child i.e. trio as the family structure. To evaluate the type I error, we considered different sample sizes, different number of markers used to control for population stratification, and different values of $\mu$. However, we fixed the value of $R$, the number of markers selected at the first stage, as 10. We evaluated type I error rates of the three methods (joint analysis, FSTA, and AFSTA) as well as two screening tests ($T_{\text{screen}}$ and $T_{\text{descen}}$) for evaluating type I error rates of the joint analysis, FSTA, and AFSTA, we used 1,000 replicated samples and thus the standard deviation for the type I error rates was $0.05 \times 0.95/1000 \approx 0.0007$ and the 95% confidence interval was $(0.036, 0.064)$ for the nominal level of 0.05. For evaluating type I error rates of $T_{\text{screen}}$ and $T_{\text{descen}}$ although we still used 1,000 replicated samples, we performed 34720 tests for each sample (equivalent to 16200 independent tests calculated using the
method of Gao et al. [31]) and thus the standard deviation for type I error rates was $\sqrt{0.05 \times 0.95/1000/16200} \approx 0.000054$ and the 95% confidence interval was (4.989%, 5.011%) for the nominal level of 5%.

Table 1, 2, 3 gave type I error rates of the five tests for simulation set 1 to set 3, respectively. From the three tables, we can see that type I error rates of the joint analysis, FSTA, and AFSTA had the same pattern across simulation set 1 to set 3, i.e. the three tests were slightly conservative. This conservative was probably due to the fact that we used Bonferroni correction to adjust for multiple testing in the second stage. Table 1, 2, 3 showed that although $T_{\text{screen}}$ was a valid test in a homogeneous population (Table 1), it would lead to false positive in structured populations (Table 2, 3). Table 1, 2, 3 also showed that $T_{\text{screen}}$ was a valid test in a homogeneous population (Table 1) and it was also a valid test in structured populations if 800 or more genomic markers were used to control for population stratification (Table 2, 3). The non-inflated type I error rates of AFSTA also show that the admixture screening test $T_{\text{screen}}$ used in the first stage and $T_a$ used in the second stage are independent. If $T_{\text{screen}}$ And $T_a$ are correlated (either positive or negative correlated), a marker with a small $p$-value of $T_{\text{screen}}$ will have a high probability to have a small $p$-value of $T_a$, and thus, AFSTA will have inflated type I error rates.

For power comparison, we considered different scenarios which included different values of heritability $h$, different number of markers selected at the first stage, different values of $\mu$, and different sample sizes. To evaluate power, in each replication, we randomly chosen a marker with minor allele frequency (calculated from European subpopulation in simulation set 2) in the interval (0.1, 0.3) as the disease locus and the minor allele as the high risk allele for dominant and additive models while the major allele as the high risk allele for recessive model. Results of power comparison were summarized in Figure 1, 2, 3 for simulation set 1 to set 3, respectively. Under simulation set 1, which considered a homogeneous population, the joint analysis was consistently more powerful than FSTA and AFSTA. Also, FSTA and AFSTA had

### Table 1. Type I error rates (in percentage) of the five tests based on simulation set 1.

| Method       | Joint | FTSA | AFTSA | $T_{\text{screen}}$ | $T_{\text{screen}}$ | $T_a$ |
|--------------|-------|------|-------|---------------------|---------------------|------|
| $S$ = 400    | 200   | 3.3  | 4     | 3.7                 | 4.99713             | 5.0991 |
| $S$ = 400    | 3.8   | 2.8  | 4.1   | 3.1                 | 4.98812             | 6.00268 |
| $S$ = 800    | 2.7   | 3.5  | 3.6   | 4.98895             | 5.01246             |
| $S$ = 600    | 3.5   | 3    | 3.4   | 4.99385             | 5.00721             |
| $S$ = 400    | 3.3   | 3.3  | 3.7   | 4.98437             | 5.00162             |
| $S$ = 800    | 3.2   | 3.7  | 3.8   | 4.98031             | 5.01397             |
| $S$ = 800    | 3.5   | 4.1  | 5.3   | 4.99007             | 5.00439             |
| $S$ = 400    | 4.4   | 4.4  | 5.0   | 4.99462             | 5.01143             |
| $S$ = 800    | 2.9   | 3.8  | 3.8   | 4.99412             | 5.01168             |

Significant level is 5%.
Note: $S$ denotes sample size in trios; $L$ denotes the number of genomic markers used to control population stratification.
doi:10.1371/journal.pone.0021957.t001

### Table 2. Type I error rates (in percentage) of the five tests based on simulation set 2.

| Method       | Joint | FTSA | AFTSA | $T_{\text{screen}}$ | $T_{\text{screen}}$ | $T_a$ |
|--------------|-------|------|-------|---------------------|---------------------|------|
| $\mu = 1$    | 3.7   | 3.3  | 4.2   | 69.9                | 5.41605             | 4.3   |
| $\mu = 2$    | 3.4   | 4.2  | 4.9   | 69.9                | 5.15228             | 3.5   |
| $\mu = 1$    | 4.1   | 2.8  | 4.6   | 69.9                | 5.00927             | 3.3   |
| $\mu = 2$    | 3.6   | 2.9  | 3.3   | 75.3                | 5.5889              | 4.0   |
| $\mu = 1$    | 3.9   | 3.8  | 3.1   | 75.3                | 5.18648             | 4.1   |
| $\mu = 2$    | 3.4   | 4.4  | 4.5   | 75.3                | 5.00261             | 4.2   |
| $\mu = 1$    | 3.5   | 3.9  | 3.6   | 78.5                | 5.77735             | 3.7   |
| $\mu = 2$    | 4.5   | 4.6  | 4.6   | 78.4                | 5.20667             | 4.8   |
| $\mu = 1$    | 3.7   | 3.2  | 3.8   | 78.4                | 5.01154             | 3.4   |
| $\mu = 2$    | 4.3   | 4.8  | 5.2   | 78.4                | 5.01154             | 3.4   |

Significant level is 5%.
Note: $S$ denotes sample size in trios; $L$ denotes the number of genomic markers used to control population stratification.
doi:10.1371/journal.pone.0021957.t002
almost the same power (Figure 1). These results indicate that the
admixture screening test, robust to population stratification, did
not lose power when compared to the traditional screening test. In
simulation set 2, we considered a structured population with two
subpopulations. In this set of simulations, the joint analysis was
again consistently more powerful than the other two methods, and
AFSTA was consistently more powerful than FSTA, which
showed that using the admixture screening test instead of
traditional screening test increased power in the presence of
population stratification (Figure 2). In simulation set 3, we
considered an admixture population with two ancestral popula-
tions which also leaded to the problem of population stratification
but not as strong as that in simulation set 2. In this set of
simulations, the pattern of power comparison was very similar to
that in simulation set 2, but the power difference between FSTA
and AFSTA was not as large as that in simulation set 2 (Figure 3).
In summary, the joint analysis was consistently the most powerful
one among the three methods we considered. Comparing the
other two methods, AFSTA had almost identical power with
FSTA in the case of no population stratification and was more
powerful than FSTA in the presence of population stratification.

Discussion

In this article, we proposed a novel method to perform joint
analysis within the framework of the family-based two-stage
analysis. In the joint analysis, we first constructed a screening test
that was based on between-family information and was robust to

Table 3. Type I error rates (in percentage) of the five tests based on simulation set 3.

| S   | L  | Joint | FTSA | AFTSA | T_{screen} | T_{Ascreen} | Joint | FTSA | AFTSA | T_{screen} | T_{Ascreen} |
|-----|----|-------|------|-------|------------|-------------|-------|------|-------|------------|-------------|
| 400 | 200| 4.1   | 2.9  | 3.3   | 5.5        | 5.31204     | 4.2   | 3.6  | 3.6   | 7.0        | 6.00905     |
| 400 | 3.8 | 4     | 4.5  | 3.3   | 5.5        | 5.17645     | 4.2   | 3.8  | 3.8   | 7.0        | 5.56551     |
| 800 | 3.4 | 3.8   | 3.8  | 5.5   | 3.3        | 5.01086     | 4.2   | 3.8  | 3.5   | 7.0        | 5.01277     |
| 600 | 200| 4.4   | 3.2  | 3.8   | 5.8        | 5.34633     | 6.5   | 4.7  | 4.7   | 8.1        | 6.37518     |
| 400 | 4.7 | 4.1   | 2.9  | 5.8   | 5.23698    | 6.5         | 5.3   | 2.9  | 8.1   | 5.75471    |
| 800 | 2.9 | 4.1   | 2.8  | 5.8   | 5.09985    | 4.1         | 7.1   | 3.5  | 8.1   | 5.01071    |
| 800 | 3.5 | 3.8   | 4.2  | 6.0   | 5.48167    | 5.9         | 5.4   | 9.1  | 6.85698   |
| 400 | 3.7 | 4.2   | 3.4  | 5.7   | 5.30941    | 5.0         | 6.6   | 4.4  | 8.7   | 6.01961    |
| 800 | 3.3 | 3.6   | 3    | 6.1   | 5.01124    | 5           | 5.1   | 4.4  | 9.1   | 5.01138    |

Significant level is 5%.
Note: S denotes sample size in trios; L denotes the number of genomic markers used to control population stratification.
doi:10.1371/journal.pone.0021957.t003

Figure 1. Power comparison under simulation set 1 when \( \mu = 2 \). In the first row, we compare power of the three methods for different values of heritability under the three disease models while sample size is 600 trios and the number of markers selected at the first stage is 10. In the second row, we compare power of the three methods for different numbers of markers selected at the first stage under the three disease models while sample size is 600 trios and heritability is 0.05. In the third row, we compare power of the three methods for different sample sizes under the three disease models while heritability is 0.05 and the number of markers selected at the first stage is 10. In each case, we use 800 genomic markers to control for population stratification in the admixture screening test.
doi:10.1371/journal.pone.0021957.g001
population stratification. In the first stage, we used this screening test to select markers. In the second stage, we did joint analysis i.e. used a test that was a combination of the screening test and the association test to test association at the selected markers. The joint analysis was robust to population stratification because both the screening test and the association test are robust to population stratification. Our simulation studies showed that the joint analysis was consistently more powerful than two-stage approaches in which the association test used in the second stage was only based on within-family information.

Although we have discussed the joint analysis, in which we only tested the selected markers in the second stage, it is straightforward to extend the joint analysis to the p-value weighting scheme [32,33] in which, instead of testing selected markers only, all markers are tested in the second stage and the resulting p-values are weighted using the p-values of the screening test. Using the p-value weighting scheme, the following steps can be used to perform the joint analysis. (1) Test all SNPs using the admixture screening test $T_{\text{Screen}}$ and order SNPs according to their p-values of the test. (2) Divide the SNPs into groups with the first group

---

**Figure 2. Power comparison under simulation set 2 when sample size is 600 trios.** In the first row, we compare power of the three methods for different values of heritability under the three disease models while $\mu = 2$ and the number of markers selected at the first stage is 10. In the second row, we compare power of the three methods for different numbers of markers selected at the first stage under the three disease models while $\mu = 2$ and heritability is 0.05. In the third row, we compare power of the three methods for different values of $\mu$ under the three disease models while heritability is 0.05 and the number of markers selected at the first stage is 10. In each case, we use 800 genomic markers to control for population stratification in the admixture screening test.

doi:10.1371/journal.pone.0021957.g002

---

**Figure 3. Power comparison under simulation set 3 when sample size is 600 trios.** In the first row, we compare power of the three methods for different values of heritability under the three disease models while $\mu = 2$ and the number of markers selected at the first stage is 10. In the second row, we compare power of the three methods for different numbers of markers selected at the first stage under the three disease models while $\mu = 2$ and heritability is 0.05. In the third row, we compare power of the three methods for different values of $\mu$ under the three disease models while heritability is 0.05 and the number of markers selected at the first stage is 10. In each case, we use 800 genomic markers to control for population stratification in the admixture screening test.

doi:10.1371/journal.pone.0021957.g003
containing $k^i$ SNPs and the $j$th group containing $k_j = 2^{i-1} k_i$ SNPs. (3) Let $p_{ij}$ denote the p-value of the admixture screening test at the $j$th SNP in the $i$th group and $p_{ij} = 1/p_{ij}$. Define an importance measure $I_{ij} = p_{ij}/p_j$ and a weight $w_i = I_{ij}/(2k_i)$ for the $j$th SNP in the $i$th group, where $p_i = (p_{1i} + \ldots + p_{ni})/k_i$. Test each SNP using the joint test statistic $T_j = T_{j,\text{screen}} + T_{j,\text{joint}}$.

Denote $p_{ij}$ the p-value of the joint test at the $j$th SNP in the $i$th group. Then, declare the $j$th SNP in the $i$th group to be significant at a level of $\alpha$ if $p_{ij} \leq \alpha w_i$. For the method of Ionita-Laza et al [32], $I_{ij} = 1$. Furthermore, for simplicity, we discussed our method using nuclear families. Our method can be also applied to general pedigrees. In fact, both the screening test $T_{\text{screen}}$ given by equation (1) and association test $T_{\text{joint}}$ given by equation (3) are applicable to general pedigrees [27,28].

It should be noted that the PC approach used in $T_{\text{screen}}$ to control for population stratification may be not as strongly resistant to stratification bias as the approach in Steen et al. [7] in which the significant association totally depends on the family-based association test used in the second stage. Other problems of the PC approach include (1) there is no standard as to how many PCs should be used; (2) the PC approach uses additive coding to code the population structure and also assumes additivity between the effects of the PCs and the effects of the genomic markers.

According to our experience of using the PC approach, however, if we use all markers in a GWAS as genomic markers, the first 10 PCs can capture subtle population structures such as the population structure and also assumes additivity between the effects of the PCs and the effects of the genomic markers. One remaining question is choosing the value of $R$, the number of markers selected in the first stage. Although there is no unique answer in choosing an optimal value of $R, our simulations indicate that 10 is a good choice of $R$ which is consistent with the results of Steen et al. [7]. However, we need further investigations on choosing the optimal value of $R$ in general.

Author Contributions
Conceived and designed the experiments: QS SZ. Performed the experiments: SZ ZZ. Wrote the paper: QS SZ.

References
1. Spielman RS, McGinnis RE, Ewens WJ (1993) The transmission test for linkage disequilibrium: the insulin gene and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52: 506–516.
2. Bickeboller H, Clerget-Darpoux F (1995) Statistical properties of the allelic and genotypic transmission/disequilibrium test for multi-allelic markers. Genet Epidemiol 12: 365–370.
3. Sham PC, Curtis D (1995) An extended transmission/disequilibrium test (TDT) for multi-allelic marker loci. Ann Hum Genet 59: 323–336.
4. Schaid DJ (1996) General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 13: 423–449.
5. Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the $ab$ transmission/disequilibrium test. Am J Hum Genet 62: 450–458.
6. Sun F, Flanders WD, Yang Q, Khoury MJ (1999) Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT. Am J Hum Genet 150: 97–104.
7. Steen KV, McQueen MB, Herbert A, Rady B, Lyon H, et al. (2005) Genomic studies of complex traits: evidence, controversy and strategies. Nat Rev Genet 6: 865–870.
8. Satagopan JM, Elston RC (2003) Optimal two-stage genotyping in population-based association studies. Genet Epidemiol 25: 149–157.
9. Wang L, Thomas DC, Peer I, Stram DO (2006) Optimal two-stage genotyping designs for genome-wide association scans. Genet Epidemiol 30: 356–368.
10. Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nature Genetics 38: 209–213.
11. Knowler WC, Williams RC, Pettitt DJ, Steinberg AG (1988) Gm3;5,13,14 and diabetes mellitus. Am J Hum Genet 42: 450–458.
12. Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265: 2037–2048.
13. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004.
14. Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genomic-based association studies. Theor Popul Biol 60: 155–166.
15. Reich DE, Goldstein DB (2001) Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 20: 4–16.
16. Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65: 220–228.
17. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67: 170–181.
18. Satten GA, Flanders WD, Yang Q (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 68: 466–477.
19. Li M, Reilly C, Hanson T (2008) A semi-parametric test to detect associations between quantitative traits and candidate genes in structured populations. Bioinformatics 24: 2356–2362.
20. Zhang S, Zhu X, Zhao H (2005) On a semi-parametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. Genet Epidemiol 28: 44–56.
21. Chen HS, Zhu X, Zhao H, Zhang S (2005) Qualitative semi-parametric test for genetic associations in case-controlled designs under structured populations. Ann Hum Genet 67: 250–264.
22. Zhu X, Zhang S, Zhao H, Cooper RS (2002) Association mapping, using a mixture model for complex traits. Genet Epidemiol 23: 181–196.
23. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) PGs analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909.
24. Bausch M, McEvoy B, Pearson LN, Quillen EE, Sarkissian T, et al. (2007) Measuring European population stratification with microarray genotype data. Am J Hum Genet 80: 948–956.
25. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S, et al. (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42(4): 354–358.
26. Zhang Z, Ereso E, Lai OQ, Todhunter RJ, Tiwari HK, et al. (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42: 355–360.
27. Zhang TF, Tang S, Sha Q (2007) Two-stage association tests for genome-wide association studies based on family data with arbitrary family structure. Eur J Hum Genet 15: 1169–1173.
28. Zhang SL, Zhang K, Li J, Sun FZ, Zhao H (2001) Test of linkage and association for quantitative traits in general pedigrees: the quantitative pedigree disequilibrium test. Genet Epidemiol 18 (Supp 1): 370–375.
29. Zhu X, Li S, Cooper R, Elston R (2008) A unified association analysis approach for family and unrelated samples correcting for stratification. Am J Hum Genet 82: 352–365.
30. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299–1320.
31. Gao X, Sturmer J, Martin ER (2006) A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genet Epidemiol 32: 361–369.
32. Ionita-Laza I, McQueen MB, Laird NM, Lange C (2007) Genome-wide weighted hypothesis testing in family-based association studies, with an application to a 100k scan. Am J Hum Genet 81: 607–614.
33. Qin H, Feng T, Zhang S, Sha Q (2010) A data-driven weighting scheme for family-based genome-wide association studies. Eur J Hum Genet 18: 596–603.