Title

Testing Allele Transmission of an SNP Set Using a Family-Based Generalized Genetic Random Field Method

Journal

Genetic Epidemiology, 40(4)

ISSN

0741-0395

Authors

Li, Ming
Li, Jingyun
He, Zihuai
et al.

Publication Date

2016-05-01

DOI

10.1002/gepi.21970

Peer reviewed

HHS Public Access

Author manuscript

Genet Epidemiol. Author manuscript; available in PMC 2017 May 01.

Published in final edited form as:

Genet Epidemiol. 2016 May ; 40(4): 341–351. doi:10.1002/gepi.21970.

Testing Allele Transmission of a SNP-Set using a Family-based Generalized Genetic Random Field Method

Ming Li[1,*], **Jingyun Li**[2], **Zihuai He**[3], **Qing Lu**[4], **John S Witte**[5], **Stewart L. Macleod**[2], **Charlotte A. Hobbs**[2], **Mario A. Cleves**[2], and **the National Birth Defect Prevention Study**

[1]Department of Epidemiology and Biostatistics, Indiana University at Bloomington, Bloomington, IN

[2]Department of Pediatrics, University of Arkansas for Medical Sciences, Little Rock, AR

[3]Department of Biostatistics, University of Michigan, Ann Arbor, MI

[4]Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI

[5]Department of Epidemiology and Biostatistics, University of California at San Francisco, San Francisco, CA

Abstract

Family-based association studies are commonly used in genetic research because they can be robust to population stratification (PS). Recent advances in high-throughput genotyping technologies have produced a massive amount of genomic data in family-based studies. However, current family-based association tests are mainly focused on evaluating individual variants one at a time. In this article, we introduce a family-based generalized genetic random field (FB-GGRF) method to test the joint association between a set of autosomal SNPs (i.e. Single Nucleotide Polymorphisms) and disease phenotypes. The proposed method is a natural extension of a recently developed GGRF method for population-based case-control studies. It models offspring genotypes conditional on parental genotypes, and thus, is robust to population stratification. Through simulations, we showed that under various disease scenarios the FB-GGRF has improved power over a commonly used family-based sequence kernel association test (FB-SKAT). Further, similar to GGRF, the proposed FB-GGRF method is asymptotically well behaved, and does not require empirical adjustment of the type I error rates. We illustrate the proposed method using a study of congenital heart defects (CHDs) with family trios from the National Birth Defect Prevention Study (NBDPS).

Keywords

family-based association test; generalized genetic random field; genetic similarity; allele distortion; population stratification; congenital heart defects

[*]Corresponding Author: li498@indiana.edu.

The authors declare no conflict of interest.

1. INTRODUCTION

Family-based studies are commonly used in genetic research. The simplest pedigree unit is a family trio (i.e. one offspring and the two biological parents), though much more complex pedigree structures (e.g. multi-generations) can be involved. Family-based association tests typically evaluate the transmission of alleles from parents to offspring. For example, the widely used transmission disequilibrium test (TDT) [Spielman, et al. 1993] considers an allele that is putatively associated with disease by comparing its frequency of being transmitted to affected offspring with that of its alternate allele. Since the un-transmitted alleles provide a control group with the same genetic ancestry, the TDT is robust to population stratification (PS). Other extensions of the TDT also model the offspring genotypes conditional on parental genotypes, accommodating various types of phenotypes [Fulker, et al. 1999; Lange, et al. 2003; Lazzeroni and Lange 1998; Martin, et al. 2000; Rabinowitz and Laird 2000; Spielman and Ewens 1998]. These statistical methods have emerged as promising tools facilitating association analyses in family-based studies. With the advance of high-throughput technologies and decreasing genotyping cost, family studies are now examining large numbers of variants for association with disease phenotypes.

Statistical methods for family data have routinely tested each individual genetic variant (e.g. SNP) in isolation, and a large number of disease susceptibility variants have been identified [Debette, et al. 2015; DeMeo, et al. 2002; Dunstan, et al. 2014; Lyon, et al. 2004; Smoller, et al. 2006]. Single variant analyses are most powerful when there is only a single causal variant within the gene and happens to be included in the study, but may have a number of limitations in more complex disease scenarios, such as failure to capture joint effects including epistasis interactions, and reduced power due to multiple testing correction or the small effect sizes of causal variants [Asimit and Zeggini 2010; Cordell 2009]. Though single-marker analysis remains a useful tool in genetic association studies, SNP-set-based association tests can be an important alternative with several advantages [Wang, et al. 2013]. First, they can increase power to detect association by integrating signals from multiple causal variants and reducing the burden of multiple testing (e.g. if the effect is due to a causative haplotype). Second, a multi-SNP approach can also account for possible interaction effects among variants by adopting appropriate kernel functions. Third, the findings obtained from gene- or region-based tests can also be investigated for functional or pathogenic importance as basic functional units of inheritance [Li, et al. 2011].

In the past few years, SNP-set-based association tests have gained popularity in genetic research. A major category of these methods is based on a kernel machine regression framework. These methods assume the regression coefficients of SNPs follow a distribution with shared variance components, and build corresponding score statistics for inference. The sequence kernel association test (SKAT) method was first proposed for population-based case-control studies, and then extended to family-based studies. The extensions, however, have adopted various strategies. One group of methods accounts for the kinship correlation among family members using either a linear mixed effect (LME) model or a generalized estimating equation (GEE) framework [Chen, et al. 2013; Wang, et al. 2013], referred to as famSKAT and gSKAT, respectively. Both famSKAT and gSKAT methods are not based on the conditional genotypes of the offspring, but directly model the association between

genotypes and phenotypes of all individuals. These methods can improve the statistical power by integrating both within- and between- family variations, but they may also be susceptible to population stratification for the same reason. As an alternative, the family-based sequence kernel association test (FB-SKAT) was developed. It models the offspring genotypes conditional on parental genotypes, and is robust to population stratification [Ionita-Laza, et al. 2013]. In addition, since association is tested between the phenotypes and conditional genotypes, founder phenotypes are not required in the study, as is commonly the case for studies with family trios.

Recently, we proposed a generalized genetic random field (GGRF) framework for testing phenotype-genotype associations in population-based case-control studies. A random field is a stochastic process that takes values in a Euclidean space, and the random variables are usually spatially correlated [Besag 1974]. In the GGRF framework, a *k*-dimensional Euclidean space can be constructed by *k* SNPs of interest. The phenotype of each individual can be mapped to a location in the space using his/her *k*-locus genotype as coordinates. In the presence of a gene-phenotype association, individuals tend to have similar phenotypes if their genotypes are "adjacent" in the Euclidean space [Li, et al. 2014]. Analogous to other similarity-based methods [Lee, et al. 2012; Tzeng, et al. 2009; Wu, et al. 2011], the GGRF method allows for multiple variants with different magnitude and directionality (e.g. risk and protective) of effect sizes, and is computationally efficient for high-dimensional genomic data analysis. Furthermore, it is asymptotically well behaved, and can be applied to small-scale data without the need of small-sample adjustment.

In this article, we extend the GGRF method to the analysis of family data. This family-based GGRF (FB-GGRF) method is similar to FB-SKAT in that it models the offspring genotypes conditional on parental genotypes, and thus, is robust to population stratification. Here, we assume that the phenotypes of offspring are the outcomes of interest. This method can be applied to studies with case-parent and control-parent trio designs, or family-trio designs with quantitative phenotypes. We also use a Hotelling's T^2 test for studies with case-parent trios only designs [Shi, et al. 2007]. The performance of FB-GGRF was compared with that of FB-SKAT via simulation studies, and further illustrated using a study of congenital heart defects (CHDs) with case-parent and control-parent trios.

2. METHODS

FB-GGRF for Family Trios with Quantitative or Binary Phenotypes

We and others have proposed a GGRF method for population-based case-control studies [Li, et al. 2014]. Following similar notations, we assume K variants in a gene or a genetic region and M covariates (e.g., age) are available for N family trios. Let y_i be the phenotype for the i-th offspring; $X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,M})'$ be the covariates for the i-th offspring, including maternal or paternal environmental exposures; $G_i^O = (g_{i,1}^O, g_{i,2}^O \ldots\ldots, g_{i,K}^O)'$, $G_i^F = (g_{i,1}^F, g_{i,2}^F \ldots\ldots, g_{i,K}^F)'$ and $G_i^M = (g_{i,1}^M, g_{i,2}^M \ldots\ldots, g_{i,K}^M)'$ be the K-variant genotype for the offspring, father and mother of the i-th family, respectively, coded as the minor allele counts. We can then evaluate the transmission pattern of these variants by

$$T_i=(t_{i,1},t_{i,2},\ldots\ldots,t_{i,K})\text{'; and } t_{i,k}=g^O_{i,k}-E(g^O_{i,k}|g^F_{i,k},g^M_{i,k})=g^O_{i,k}-(g^F_{i,k}+g^M_{i,k})/2; \quad \text{Eq. (1)}$$

and $t_{i,k}=g^O_{i,k}-E(g^O_{i,k}|g^F_{i,k},g^M_{i,k})=g^O_{i,k}-(g^F_{i,k}+g^M_{i,k})/2$; where the expected minor allele counts, $E(g^O_{i,k}|g^F_{i,k},g^M_{i,k})$, is calculated under the null hypothesis assuming Mendelian transmission. The $t_{i,k}$ can be interpreted as the transmission distortion at the k-th variant, measuring the difference between the observed genotype of an offspring and the expected genotype under Mendelian transmission. This metric was previously adopted in studies with case-parent trios only designs and extended for studies with case-spouse and case-offspring designs [Kistner, et al. 2009; Lee 2003]. We expect a disease risk allele to be over-transmitted among case families compared to control families (i.e. $t_{i,j}>0$), whereas a protective allele would be under-transmitted to case families compared to control families (i.e. $t_{i,j}<0$).

Similar to GGRF, a conditional auto-regressive (CAR) model is adopted:

$$E(y_i|y_{-i})=\mu_i+\gamma\sum\nolimits_{j\neq i}s_{i,j}(y_j-\mu_j), \quad \text{Eq. (2)}$$

where y_{-i} denotes phenotypes for offspring from all families other than family i, and $\mu_i=f(X_i^{'}\beta)$, where $f(\cdot)$ is the mean function as in a generalized linear model, used for adjusting covariates. Specifically, if the phenotype is quantitative, we use the identity link $f(x)=x$; if the phenotype is binary, we use the logistic link $f(x)=\frac{\exp(x)}{1+\exp(x)}$; $s_{i,j}$ is a weight representing the relative contribution of the j-th offspring in predicting the phenotype of offspring i, determined by the genetic similarity between offspring i and j. Assuming there is a gene-phenotype association, we expect that the phenotype of an offspring (i.e. y_i) can be predicted by the phenotypes of offspring from other families (i.e. y_{-i}). The contribution of offspring from the j-th family (i.e. y_j) is proportional to the genetic similarity between them (i.e. $s_{i,j}$).

Here, we define genetic similarity between offspring from two families by their transmission distortion at K-SNP loci:

$$s_{i,j}=s(T_i,T_j)=\frac{1}{K}\sum\nolimits_{k=1}^{K}(t_{i,k}-q_k)(t_{j,k}-q_k) \quad \text{Eq. (3)}$$

where q_k is the average transmission distortion at variant k in the population ($q_k=\sum\nolimits_{i=1}^{N}t_{i,k}/N$). The above similarity metric is an un-weighted version of genetic relationship which was used by GCTA for heritability estimation [Yang, et al. 2011], and is also a centered version of linear kernel which has been used in the sequence kernel association test (SKAT) and its extensions [Lee, et al. 2012; Wu, et al. 2011]. Various forms

of similarity metrics, such as the identity-by-state (IBS) metric or other p-norm distance-based similarity metric, can also be adopted [Li, et al. 2014].

In Equation (2), γ is a non-negative coefficient measuring the magnitude of joint association between the K variants and phenotypes. Under the null hypothesis of no association (i.e., $\gamma = 0$), the phenotype of offspring i will be independent of the phenotypes of offspring from other families, regardless of their genetic similarity. On the other hand, a large γ indicates a strong genetic contribution to the phenotype. Therefore, the joint association between the K variants and phenotypes can be tested through a single parameter, $H_0 : \gamma = 0$. We construct a test statistic via a generalized estimating equation (GEE) framework as:

$$\hat{\gamma}=\frac{(Y-\hat{\mu})'S(Y-\hat{\mu})}{(Y-\hat{\mu})'S^2(Y-\hat{\mu})} \quad \text{Eq. (5)}$$

with $p_{H0}(\hat{\gamma} > \hat{\gamma}_{obs}) = P((Y-\hat{\mu})'\ (S-\hat{\gamma}_{obs}S^2)(Y-\hat{\mu} > 0) = P(Q > 0)$.

In Equation (5), Y, S and $\hat{\mu}$ are the matrix form for the phenotype, genetic similarity, and non-genetic mean, respectively; $Y = (y_1, y_2, \ldots, y_N)'$; S is a $N \times N$ similarity matrix with $s_{i,j}$ as its element in row i and column j, $1 \le i \ne j \le N$, and zeros on the diagonal; and $\hat{\mu} = (X_1, \ldots\ldots\ X_M)'\ \hat{\beta}$. Given an observed value $\hat{\gamma}_{obs}$ under the null hypothesis of no association, the statistic $Q = (Y-\hat{\mu})'(S-\hat{\gamma}S^2)\ (Y-\hat{\mu})$, asymptotically follows a mixture of Chi-squares, $\sum_{k=1}^{K}\lambda_k\chi^2_{1,k}$, where $(\lambda_1, \ldots, \lambda_K)$ are the eigenvalues of the matrix $P^{1/2}(S-\hat{\gamma}S^2)\ P^{1/2}$ and $P = W - WX\,(X'\,WX)^{-1}\,X'\,W$; W is a diagonal matrix with its i-th element $w_i = 1$ for quantitative phenotypes, and $w_i = \hat{\mu}_i(1-\hat{\mu}_i)$ for binary phenotypes with a logistic link. Davies' method can then be used to obtain the significance level of the association test [Davies 1980]. The proposed FB-GGRF method has been implemented in R, and the source code can be downloaded in a public repository of GitHub at https://github.com/li498/FB_GGRF.

Hotelling's T^2 test for Case-parent Trios

In the situation when case-parent trios are available, but control families are not, the non-genetic mean, $\hat{\mu}$, cannot be estimated through a generalized linear model. However, under the null hypothesis of no association, alleles are not expected to be over- or under-transmitted in the population (i.e. Mendel's law of transmission). In such a situation, we can test the transmission distortions, T_i, as defined in Equation (1) by:

$$H_0: E(t_{i,1})=E(t_{i,2})=\ldots\ldots=E(t_{i,K})=0 \leftrightarrow H_A: \text{they are not all zeros} \quad \text{Eq. (6)}$$

A Hotelling's T^2 test was previously proposed for detecting multi-SNP effects in studies with nuclear families [Shi, et al. 2007]. The same strategy is adopted here by assuming multivariate normal distribution of T_i, so that

$$N \cdot \overline{T}' \Sigma^{-1} \overline{T} \sim \chi^2_N, \quad \text{Eq. (7)}$$

where $\overline{T} = (\overline{T}_1 + \ldots\ldots + \overline{T}_N)/N$ and Σ is the variance covariance matrix of T_i.

To account for high dimension data (e.g. $N<K$) and violations of the multivariate normality assumption, we used a modified test proposed by Dempster [Dempster 1960]:

$$\frac{N \cdot \overline{T}' \cdot \overline{T}}{\text{trace}(\Sigma)} \sim F_{r,(n-r)r}; \text{where } r = \frac{(\text{trace}(\Sigma))^2}{\text{trace}(\Sigma^2)} \quad \text{Eq. (8)}$$

A number of other modified Hotelling's T^2 test have been proposed in the literature, and may be used as well [Bai and Saranadasa 1996; Srivastava and Du 2008].

Missing Genotypes

In studies with family trios, it is common to have missing genotypes or missing members from the families (e.g. parental genotypes are not completely known). In this study, we adopt a simple strategy by imputing the missing genotypes as twice the minor allele frequency among founder populations. Various other imputation strategies have been suggested in the literature [Li, et al. 2009], which may improve the performance.

3. RESULTS

3.1 Simulation Studies

Simulation studies were conducted to evaluate the proposed method, and to compare it against the commonly used Family-based Sequence Kernel Association Test (FB-SKAT). We evaluated the performance of these two methods by examining their type I error rates, statistical power, and robustness to population stratification or missing genotypes/subjects. FB-SKAT version 2.1 was used in all simulations. Each simulation scenario was repeated 100,000 times to evaluate type I error rates, and repeated 1,000 times to evaluate statistical power.

Simulation I: Type I Error Evaluation

(a) Simulations in the Absence of Population Stratification: We first evaluated the type I error rates when population stratification was absent. We simulated a haplotype pool of 200,000 haplotypes using software HAPGEN version 2 [Su, et al. 2011]. HapMap Phase 3 CEU samples (release 2, NCBI Build 36, Utah residents with northern and western European ancestry) were used as reference panel to generate haplotypes [International HapMap, et al. 2010]. Each haplotype was simulated for the entire chromosome 22, including a total of 20,085 variants. The minor allele frequencies (MAF) ranged from an extremely low MAF (e.g. 5e-06) to a MAF close to 0.5. The distribution of MAFs is illustrated in Figure 1. For each family trio, the parental genotypes were simulated by randomly sampling two haplotypes with replacement from the haplotype pool; the offspring

genotypes were then simulated by randomly transmitting one haplotype from each parent with equal probability. In each simulation, one hundred variants located consecutively on the chromosome were randomly selected and tested as a SNP-set.

Phenotypes were simulated independently from the genotypes, in order to evaluate type I error rates. We considered phenotypes from three types of study designs: family trios with quantitative phenotypes, case-parent and control-parent trios, and case-parent trios only:

1) For family trios with quantitative phenotypes, we first simulated genotypes for 1,000 family trios. The phenotypes were then simulated from a standard normal distribution, $N(0,1)$.

2) For case-parent and control-parent trios, we first simulated genotypes for 100,000 family trios. The phenotypes were then simulated from a Bernoulli distribution, Bernoulli (0.05). A total of 1,000 family trios, including 500 case-parent trios and 500 control-parent trios, were then randomly selected as the study sample.

3) For case-parent trios only, we first simulated genotypes and phenotypes for 100,000 family trios similar to that of binary phenotypes, and then randomly selected 500 case-parent trios as the study sample.

The type I error rates of two methods were also evaluated in the presence of missing genotypes. We first simulated all genotypes and phenotypes as previously described, and then considered three scenarios with missing genotypes: 1) ten percent of the genotypes in all subjects were randomly selected and set to missing. 2) Twenty percent of the subjects, including both parents and offspring, were randomly selected, and their entire genotypes were set to missing. 3) Thirty percent of the subjects, including both parents and offspring, were randomly selected, and their entire genotypes were set to missing.

(b) Simulations in the Presence of Population Stratification: To evaluate type I error rates in the presence of PS, we considered an admixed population with 2 ethnicity groups: CEU and ASW (African ancestry in southwest USA). A haplotype pool of 10,000 haplotypes was simulated for each ethnicity group using software HAPGEN version 2. The genotypes in each subpopulation were simulated as described for non-PS simulations. For family trios with quantitative phenotypes, the two subpopulations had a baseline phenotype difference of 0.3, while for case-parent/control parent trios and case-parent trios only, the disease prevalence was 10% and 40% for the two subpopulations, respectively. Type I error rates were also evaluated in the presence of missing genotypes or missing subjects.

Simulation II: Power Comparison for Family Trios with Quantitative Phenotypes—We first simulated genotypes for 1,000 family trios as described for non-PS simulations. The phenotype of offspring in the i-th family was then simulated as:

$$y_i=\beta_0+\sum_{j=1}^{100}\beta_j x_{ij}+\varepsilon_i; \quad \text{Eq. (9)}$$

where x_{ij} was the minor allele counts for the j-th variant of the offspring in family i; ε_i was a random error following a standard normal distribution; β_0 was the baseline level of phenotypes, which was set to 0; β_j was the effect size of the j-th variant.

We assumed the effect sizes were inversely correlated with the minor allele frequencies (MAF) of variants:

$$|\beta_j| = \begin{cases} -c \times \log(\mathrm{MAF}_j, \mathrm{base}=10) & \text{if variant } j \text{ is casual} \\ 0 & \text{otherwise} \end{cases};$$

where c was a constant adjusted to ensure the power of two methods were within a reasonable range. We examined performance with varying proportion of causal variants (i.e. 5%, 10% and 20%) and directionality of effect sizes (i.e. positive or negative effect). For unidirectional scenarios, all effect sizes were assumed to be positive, while for bi-directional scenarios, a sign factor 1 or −1 was randomly selected for each β_j with a probability of 0.5.

Simulation III: Power Comparison for Case-parent and Control-parent Trios— We first simulated genotypes for 100,000 family trios as described for non-PS simulations. The phenotypes were simulated as:

$$\text{logit } P(y_i=1) = \beta_0 + \sum_{j=1}^{100} \beta_j x_{ij}; \quad \text{Eq. (10)}$$

where β_0 was adjusted to ensure the disease prevalence was approximately 5% in the population. The effect size, β_j, was defined similarly to that in Equation (9). A total of 1,000 trios, including 500 hundred case-parent trios and 500 control-parent trios were randomly selected as the study sample. The performance of two methods was examined by varying the proportion of causal variants (i.e. 5%, 10% and 20%) and directionality of effect sizes (i.e. unidirectional and bi-directional).

Simulation IV: Power comparison for Case-parent Trios Only—Genotypes and phenotypes were simulated similar to case-parent and control-parent scenarios (Simulation III), and 500 case-parent trios were then randomly selected as the study sample. The performance of two methods was examined by varying the proportion of causal variants (i.e. 5%, 10% and 20%) and directionality of effect sizes (i.e. unidirectional and bi-directional).

3.2 Simulation Results

Simulation I: Type I Errors—The results for type I error rates are summarized in Table 1. We evaluated the type I error rates at various significance levels, including 0.05, 0.01, and 0.001. The results suggest that both FB-SKAT and FB-GGRF had well controlled type I error rates at various significance levels for all study designs (i.e. family trios with quantitative phenotypes, case-parent and control parent trios, and case-parent trios). The type I error rate of Hotelling's T^2 test was well controlled at a significance level of 0.05, but was slightly inflated at the 0.001 level. The type I error rates of all methods were robust to population stratification and missing genotypes. We also evaluated the type I error rates for

their population-based alternatives, GGRF and SKAT, by using offspring genotypes only. GGRF and SKAT had well controlled type I error rates without PS and inflated type I error rates in the presence of PS, confirming the existence of confounding effect due to population stratification in our simulated data (results not shown).

Simulation II: Power Comparison for Family Trios with Quantitative Phenotypes—Since all methods were robust to PS in terms of type I error rates, we evaluated their power without considering PS. Figure 2 summarizes the power comparison between FB-GGRF and FB-SKAT for quantitative phenotypes. The power of both methods increased as the proportion of causal variants increased. Both methods had robust power for bi-directional effect sizes (i.e. positive or negative effect). Under our simulation scenarios, FB-GGRF attained an average of 20% (SD=9.7%) increase in power over FB-SKAT. The power of FB-GGRF decreased with increasing percentages of missing genotypes. On the other hand, the power of FB-SKAT was less affected by missing genotypes or missing subjects.

Simulation III: Power Comparison for Case-parent and Control-parent Trios and Simulation IV: Power Comparison for Case-parent Trios only—Figure 3 and Figure 4 summarize the power comparison between FB-GGRF and FB-SKAT for case-parent and control-parent trios, and the power comparison between Hotelling's T^2 test and FB-SKAT for case-parent trios only, respectively. The results were highly consistent with the power comparison for quantitative phenotypes. In all simulation scenarios, FB-GGRF (Hotelling's T^2 test) had higher power than FB-SKAT. On average, FB-GGRF attained 20% (SD=9.4%) increase in power, while Hotelling's T^2 test attained 17.8% (SD=9.3%) increase in power. The power of FB-SKAT was less affected by missing genotypes than that of FB-GGRF or Hotelling's T^2 test.

3.3 Application to a Study of Congenital Heart Defects (CHDs)

To illustrate these tests on real data, we applied FB-GGRF and FB-SKAT to a study of congenital heart defects with case-parent and control-parent trios. The dataset was part of the National Birth Defects Prevention Study (NBDPS), a large-scale multi-center study covering an annual birth population of ~ 482,000, or 10% of U.S. births. CHD cases were ascertained from birth defect registries in ten participating states that had similar inclusion criteria: Arkansas, California, Georgia, Iowa, Massachusetts, New Jersey, New York, North Carolina, Texas, and Utah. A detailed description of NBDPS protocols can be found elsewhere [Gallagher, et al. 2011; Rasmussen, et al. 2002; Reefhuis, et al. 2015; Yoon, et al. 2001]. In this study, cases were singleton, live-born infants with conotruncal heart defects (CTDs), whereas controls were singleton live-born infants without any structural birth defect. To be eligible for the study, the mothers needed to have an estimated date of delivery between October 1997 and August 2008, have completed a maternal interview, and provided buccal samples for genotyping.

All eligible infants and their parents who provided DNA samples were genotyped using the Illumina GoldenGate custom genotyping platform. SNPs were selected from candidate genes in the homocysteine, folate, and transsulfuration pathways that are potentially related

to the development of CHDs. After genotyping and subsequent quality control checks, the final study population included a total of 616 case families and 1,645 control families. A large proportion of the families were incomplete with one or two missing members. Among 616 case families, there were 230 trios, 101 mother-offspring duos, 31 father-offspring duos, 90 father-mother duos, 96 mother-only, 31 father-only and 37 offspring-only families. Among 1,645 control families, there were 559 trios, 316 mother-offspring duos, 128 father-offspring duos, 143 father-mother duos, 242 mother-only, 94 father-only and 163 offspring-only families. Genotype data was available for ~921 bi-allelic SNPs in 60 candidate genes for each individual. The unavailable members from incomplete trios were handled as missing genotypes. The detailed DNA collection, genotyping, and quality control procedures can be found elsewhere [Chowdhury, et al. 2012; Hobbs, et al. 2014]. The maternal characteristics are summarized in Table 2. The case mothers were slightly older than the control mothers (P-value=0.002), but the age difference may not be of any clinical significance (28.3 years versus 27.5 years). Case and control mothers did not differ on any of the other demographics and lifestyle characteristics (P-values>0.05).

We conducted a gene-based association test for each of the sixty candidate genes using FB-GGRF and FB-SKAT. The pairwise P-values are compared in Figure 5. The P-values of two methods are generally along the diagonal, indicating that two methods are consistent with each other in terms of overall significance. Using a Bonferroni corrected threshold (i.e. 0.05/60), five and two genes were identified to have significant associations with CHD phenotype by FB-GGRF and FB-SKAT, respectively. The genes identified by either method are summarized in Table 3. Out of the seven genes identified by either FB-GGRF or FB-SKAT, six genes had P-values less than the nominal threshold of 0.05 for both methods. These results suggested that the two methods are generally consistent with each other, but may perform differently on each individual gene.

4. DISCUSSION

We have extended a previously proposed GGRF method to a FB-GGRF method for family-based association tests. The FB-GGRF method examines the transmission distortion of a set of SNPs by modeling the offspring genotypes conditional on parental genotypes, and thus is robust to population stratification. A similar strategy has previously been used to extend SKAT method to a FB-SKAT method [Ionita-Laza, et al. 2013]. In this article, we have empirically demonstrated that FB-GGRF may attain higher power than FB-SKAT under various disease scenarios. The proposed FB-GGRF method can also account for non-normally distributed phenotypes by specifying a link function through a generalized linear model framework. In the Appendix, we conducted additional simulations for phenotypes following a Poisson distribution, demonstrating that the performance of FB-GGRF can be much improved if the appropriate link function is used. Further, the score statistic of FB-SKAT asymptotically follows a mixture of Chi-square distribution when the phenotypes are normally distributed. However, the inference of FB-SKAT by using asymptotic distribution was known to be conservative for small sample size or binary phenotypes [Ionita-Laza, et al. 2013]. Therefore, a moment matching approach was implemented by FB-SKAT to obtain its testing P-values. More precisely, the variance and kurtosis of its score statistic was estimated empirically by performing Monte Carlo simulations. In our study, the number of Monte

Carlo simulations was set to 10,000, which is the default setting of FB-SKAT. On the other hand, FB-GGRF is asymptotically well behaved and can be applied efficiently. The results from the analysis of CHD dataset also appear to be consistent with these simulations.

We have focused on family-based studies with the simplest pedigree structure (i.e. family trios). In fact, both FB-SKAT and FB-GGRF can be generalized to other complex pedigree structures by breaking them down to family trios, assuming offspring genotypes are independent conditional on their parental genotypes. Both methods evaluate the transmission of alleles from parents to offspring, and gain a major advantage for being robust to population stratification. The unconditional methods (i.e. famSKAT and gSKAT) utilize both within- and between- family information for power improvement. However, they may also be susceptible to population stratification, and cannot be applied when founder phenotypes are not available. Therefore, a direct comparison between conditional methods and unconditional methods is not straightforward.

The proposed FB-GGRF has a few limitations. First, it cannot be directly applied to studies with only case-parent trios. In such a situation, we have proposed to use a modified Hotelling's T^2 test, which has been previously adopted in the literature [Shi, et al. 2007]. Second, in this study, we have imputed the missing genotypes as twice the minor allele frequency in founder populations. Imputation based on minor allele frequencies has been commonly used in existing studies [Chen, et al. 2013; Wang, et al. 2013]. Though convenient and easy, this imputation strategy tends to bias the association towards the null, resulting in power loss. The missing genotype can sometimes be partly or fully inferred based on Mendelian transmission. Incorporating other imputation methods can potentially improve the performance. The genotype imputation among related individuals is commonly based on the idea that family members share long stretches of haplotypes. Several widely used procedures have been implemented in packages such as MERLIN and MENDEL [Abecasis, et al. 2002; Abecasis and Wigginton 2005; Lange, et al. 2005].

As we discussed in a previous publication, the proposed GGRF framework has a close connection to the SKAT framework [Li, et al. 2014]. Both methods can be interpreted as similarity-based methods, but they model the correlation structure differently. When the phenotypes are normally distributed, the GGRF model can be expressed as: $Y \sim X\beta + \nu$, $\nu \sim N(0, \sigma^2(I - \gamma S)^{-1})$. On the other hand, the SKAT model can be written as a linear mixed model: $Y \sim X\beta + \upsilon$, $\upsilon \sim N(0, \sigma^2 I + \tau^2 K)$; where σ^2 is the variance of Y_i under the null hypothesis; $\tau^2 K$ represents the variance component for the genetic effects [Kwee, et al. 2008; Liu, et al. 2007; Wu, et al. 2010]. The two models are equivalent under the null hypothesis of no association (i.e., $\tau=0$ in SKAT or $\gamma=0$ in GGRF), but have different modeling of correlation structure under the alternative, which may lead to different performance.

In an application to a CHD dataset of case-parent/control-parent trios, FB-GGRF and FB-SKAT identified five and two genes, respectively. Though none of these genes overlapped, both methods achieved nominal significant *P*-values (i.e. < 0.05) for most of the identified genes. This result indicates that two methods are consistent for testing the overall association, but may perform differently for each individual gene due to the different ways

of modeling genetic effects. We also applied Hoteling T^2 test and FB-SKAT to the case families only, but none of the genes were significant after Bonferroni correction (See Supplementary Materials). The identified genes are all from the transsulfuration or homocysteine pathways. The metabolic pathway from homocysteine to glutathione is referred to as the transsulfuration pathway [Hobbs, et al. 2005]. Elevated homocysteine is associated with alterations in the transsulfuration pathway that lead to greater oxidative stress [Huang, et al. 2001]. Homocysteine is a sulfur-containing amino acid that plays a crucial role in methylation reactions. Transfer of the methyl group from betaine to homocysteine creates methionine, which donates the methyl group to methylate DNA, proteins, lipids, and other intracellular metabolites. Anomalies in homocysteine metabolism have been implicated in various disorders, such as cardiovascular diseases [Giusti, et al. 2010; McGeachie, et al. 2009; Weisberg, et al. 2003], orofacial cleft defects [Mostowska, et al. 2010a; Mostowska, et al. 2010b], and neural tube defects [Shaw, et al. 2009]. Although results from our CHD application are only suggestive, they provide promising candidates for future studies to investigate and to replicate.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We want to thank the numerous families for their generous participation in the National Birth Defects Prevention Study that made this research possible. We also thank the Centers for Birth Defects Research and Prevention in Arkansas, California, Georgia, Iowa, Massachusetts, New Jersey, New York, North Carolina, Texas, and Utah for their contribution of data and manuscript review. We also want to thank Ashley S. Block for assistance in the preparation of this manuscript, and two anonymous reviewers for their insightful comments that improved this manuscript.

This work is supported, in part, by the Translational Research Institute through the NIH National Center for Research Resources (NCRR) and the National Center for Advancing Translational Sciences (NCATS) under Award Number UL1TR000039 and KL2TR000063, the National Institute of Child Health and Human Development (NICHD) under award number 5R01HD039054, the National Center on Birth Defects and Developmental Disabilities (NCBDDD) under award number 5U01DD000491, the University of Arkansas for Medical Sciences College of Medicine Children's University Medical Group (CUMG) Fund Grant Program, the Arkansas Children's Hospital Research Institute (ACHRI), and the Arkansas Biosciences Institute (ABI). We also want to thank Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute, and Indiana METACyt Initiative. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health (NIH) or the Center for Disease Control and Prevention (CDC).

REFERENCES

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet. 2002; 30(1):97–101. [PubMed: 11731797]

Abecasis GR, Wigginton JE. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am J Hum Genet. 2005; 77(5):754–767. [PubMed: 16252236]

Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. Annu Rev Genet. 2010; 44:293–308. [PubMed: 21047260]

Bai Z, Saranadasa H. Effect of high dimension: by an example of a two sample problem. Statist. Sinica. 1996; 6:311–329.

Besag J. Spatial interaction and statistical Analysis of Lattice Systems. Journal of the Royal Statistical Society, Series B. 1974; 48:259–302.

Chen H, Meigs JB, Dupuis J. Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol. 2013; 37(2):196–204. [PubMed: 23280576]

Chowdhury S, Hobbs CA, MacLeod SL, Cleves MA, Melnyk S, James SJ, Hu P, Erickson SW. Associations between maternal genotypes and metabolites implicated in congenital heart defects. Mol Genet Metab. 2012; 107(3):596–604. [PubMed: 23059056]

Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet. 2009; 10(6):392–404. [PubMed: 19434077]

Davies R. The distribution of a linear combination of Chi-square random variables. Applied Statistics. 1980; 29:323–333.

Debette S, Kamatani Y, Metso TM, Kloss M, Chauhan G, Engelter ST, Pezzini A, Thijs V, Markus HS, Dichgans M. and others. Common variation in PHACTR1 is associated with susceptibility to cervical artery dissection. Nat Genet. 2015; 47(1):78–83. [PubMed: 25420145]

DeMeo DL, Lange C, Silverman EK, Senter JM, Drazen JM, Barth MJ, Laird N, Weiss ST. Univariate and multivariate family-based association analysis of the IL-13 ARG130GLN polymorphism in the Childhood Asthma Management Program. Genet Epidemiol. 2002; 23(4):335–348. [PubMed: 12432502]

Dempster AP. A significance test for the seperation of two highly multivariate small samples. Biometrics. 1960; 16:41–50.

Dunstan SJ, Hue NT, Han B, Li Z, Tram TT, Sim KS, Parry CM, Chinh NT, Vinh H, Lan NP. and others. Variation at HLA-DRB1 is associated with resistance to enteric fever. Nat Genet. 2014; 46(12):1333–1336. [PubMed: 25383971]

Fulker DW, Cherny SS, Sham PC, Hewitt JK. Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet. 1999; 64(1):259–267. [PubMed: 9915965]

Gallagher ML, Sturchio C, Smith A, Koontz D, Jenkins MM, Honein MA, Rasmussen SA. Evaluation of mailed pediatric buccal cytobrushes for use in a case-control study of birth defects. Birth Defects Res A Clin Mol Teratol. 2011; 91(7):642–648. [PubMed: 21630425]

Giusti B, Saracini C, Bolli P, Magi A, Martinelli I, Peyvandi F, Rasura M, Volpe M, Lotta LA, Rubattu S. and others. Early-onset ischaemic stroke: analysis of 58 polymorphisms in 17 genes involved in methionine metabolism. Thromb Haemost. 2010; 104(2):231–242. [PubMed: 20458436]

Hobbs CA, Cleves MA, Macleod SL, Erickson SW, Tang X, Li J, Li M, Nick T, Malik S. National Birth Defects Prevention S. Conotruncal heart defects and common variants in maternal and fetal genes in folate, homocysteine, and transsulfuration pathways. Birth Defects Res A Clin Mol Teratol. 2014; 100(2):116–126. [PubMed: 24535845]

Hobbs CA, Cleves MA, Zhao W, Melnyk S, James SJ. Congenital heart defects and maternal biomarkers of oxidative stress. Am J Clin Nutr. 2005; 82(3):598–604. [PubMed: 16155273]

Huang RF, Hsu YC, Lin HL, Yang FL. Folate depletion and elevated plasma homocysteine promote oxidative stress in rat livers. J Nutr. 2001; 131(1):33–38. [PubMed: 11208935]

Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F. International HapMap C; and others. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467(7311):52–58. [PubMed: 20811451]

Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Family-based association tests for sequence data, and comparisons with population-based association tests. Eur J Hum Genet. 2013; 21(10): 1158–1162. [PubMed: 23386037]

Kistner EO, Shi M, Weinberg CR. Using cases and parents to study multiplicative gene-by-environment interaction. Am J Epidemiol. 2009; 170(3):393–400. [PubMed: 19483188]

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. Am J Hum Genet. 2008; 82(2):386–397. [PubMed: 18252219]

Lange C, Silverman EK, Xu X, Weiss ST, Laird NM. A multivariate family-based association test using generalized estimating equations: FBAT-GEE. Biostatistics. 2003; 4(2):195–206. [PubMed: 12925516]

Lange K, Sinsheimer JS, Sobel E. Association testing with Mendel. Genet Epidemiol. 2005; 29(1):36–50. [PubMed: 15834862]

Lazzeroni LC, Lange K. A conditional inference framework for extending the transmission/disequilibrium test. Hum Hered. 1998; 48(2):67–81. [PubMed: 9526165]

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA. Christiani DC, Wurfel MM, Lin X. Team NGESP-ELP. 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(2):224–237. [PubMed: 22863193]

Lee WC. Genetic association studies of adult-onset diseases using the case-spouse and case-offspring designs. Am J Epidemiol. 2003; 158(11):1023–1032. [PubMed: 14630596]

Li M, He Z, Zhang M, Zhan X, Wei C, Elston RC, Lu Q. A generalized genetic random field method for the genetic association analysis of sequencing data. Genet Epidemiol. 2014; 38(3):242–253. [PubMed: 24482034]

Li MX, Gui HS, Kwan JS, Sham PC. GATES: a rapid and powerful gene-based association test using extended Simes procedure. Am J Hum Genet. 2011; 88(3):283–293. [PubMed: 21397060]

Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009; 10:387–406. [PubMed: 19715440]

Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics. 2007; 63(4):1079–1088. [PubMed: 18078480]

Lyon H, Lange C, Lake S, Silverman EK, Randolph AG, Kwiatkowski D, Raby BA, Lazarus R, Weiland KM, Laird N. and others. IL10 gene polymorphisms are associated with asthma phenotypes in children. Genet Epidemiol. 2004; 26(2):155–165. [PubMed: 14748015]

Martin ER, Monks SA, Warren LL, Kaplan NL. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. Am J Hum Genet. 2000; 67(1):146–154. [PubMed: 10825280]

McGeachie M, Ramoni RL, Mychaleckyj JC, Furie KL, Dreyfuss JM, Liu Y, Herrington D, Guo X, Lima JA, Post W. and others. Integrative predictive model of coronary artery calcification in atherosclerosis. Circulation. 2009; 120(24):2448–2454. [PubMed: 19948975]

Mostowska A, Hozyasz KK, Biedziak B, Misiak J, Jagodzinski PP. Polymorphisms located in the region containing BHMT and BHMT2 genes as maternal protective factors for orofacial clefts. Eur J Oral Sci. 2010a; 118(4):325–332. [PubMed: 20662904]

Mostowska A, Hozyasz KK, Wojcicki P, Dziegelewska M, Jagodzinski PP. Associations of folate and choline metabolism gene polymorphisms with orofacial clefts. J Med Genet. 2010b; 47(12):809–815. [PubMed: 19737740]

Rabinowitz D, Laird N. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered. 2000; 50(4):211–223. [PubMed: 10782012]

Rasmussen SA, Lammer EJ, Shaw GM, Finnell RH, McGehee RE Jr, Gallagher M, Romitti PA, Murray JC. National Birth Defects Prevention S. Integration of DNA sample collection into a multi-site birth defects case-control study. Teratology. 2002; 66(4):177–184. [PubMed: 12353214]

Reefhuis J, Gilboa SM, Anderka M, Browne ML, Feldkamp ML, Hobbs CA, Jenkins MM, Langlois PH, Newsome KB, Olshan AF. and others. The national birth defects prevention study: A review of the methods. Birth Defects Res A Clin Mol Teratol. 2015; 103(8):656–669. [PubMed: 26033852]

Shaw GM, Lu W, Zhu H, Yang W, Briggs FB, Carmichael SL, Barcellos LF, Lammer EJ, Finnell RH. 118 SNPs of folate-related genes and risks of spina bifida and conotruncal heart defects. BMC Med Genet. 2009; 10:49. [PubMed: 19493349]

Shi M, Umbach DM, Weinberg CR. Identification of risk-related haplotypes with the use of multiple SNPs from nuclear families. Am J Hum Genet. 2007; 81(1):53–66. [PubMed: 17564963]

Smoller JW, Biederman J, Arbeitman L, Doyle AE, Fagerness J, Perlis RH, Sklar P, Faraone SV. Association between the 5HT1B receptor gene (HTR1B) and the inattentive subtype of ADHD. Biol Psychiatry. 2006; 59(5):460–467. [PubMed: 16197923]

Spielman RS, Ewens WJ. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet. 1998; 62(2):450–458. [PubMed: 9463321]

Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet. 1993; 52(3):506–516. [PubMed: 8447318]

Srivastava MS, Du M. A test for the mean vector with fewer observations than the dimension. Journal of Multivariate Analysis. 2008; 99:386–402.

Su Z, Marchini J, Donnelly P. HAPGEN2: simulation of multiple disease SNPs. Bioinformatics. 2011; 27(16):2304–2305. [PubMed: 21653516]

Tzeng JY, Zhang D, Chang SM, Thomas DC, Davidian M. Gene-trait similarity regression for multimarker-based association analysis. Biometrics. 2009; 65(3):822–832. [PubMed: 19210740]

Wang X, Lee S, Zhu X, Redline S, Lin X. GEE-based SNP set association test for continuous and discrete traits in family-based association studies. Genet Epidemiol. 2013; 37(8):778–786. [PubMed: 24166731]

Weisberg IS, Park E, Ballman KV, Berger P, Nunn M, Suh DS, Breksa AP 3rd, Garrow TA, Rozen R. Investigations of a common genetic variant in betaine-homocysteine methyltransferase (BHMT) in coronary artery disease. Atherosclerosis. 2003; 167(2):205–214. [PubMed: 12818402]

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 2010; 86(6):929–942. [PubMed: 20560208]

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89(1):82–93. [PubMed: 21737059]

Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011; 88(1):76–82. [PubMed: 21167468]

Yoon PW, Rasmussen SA, Lynberg MC, Moore CA, Anderka M, Carmichael SL, Costa P, Druschel C, Hobbs CA, Romitti PA. and others. The National Birth Defects Prevention Study. Public Health Rep. 2001; 116(Suppl 1):32–40.

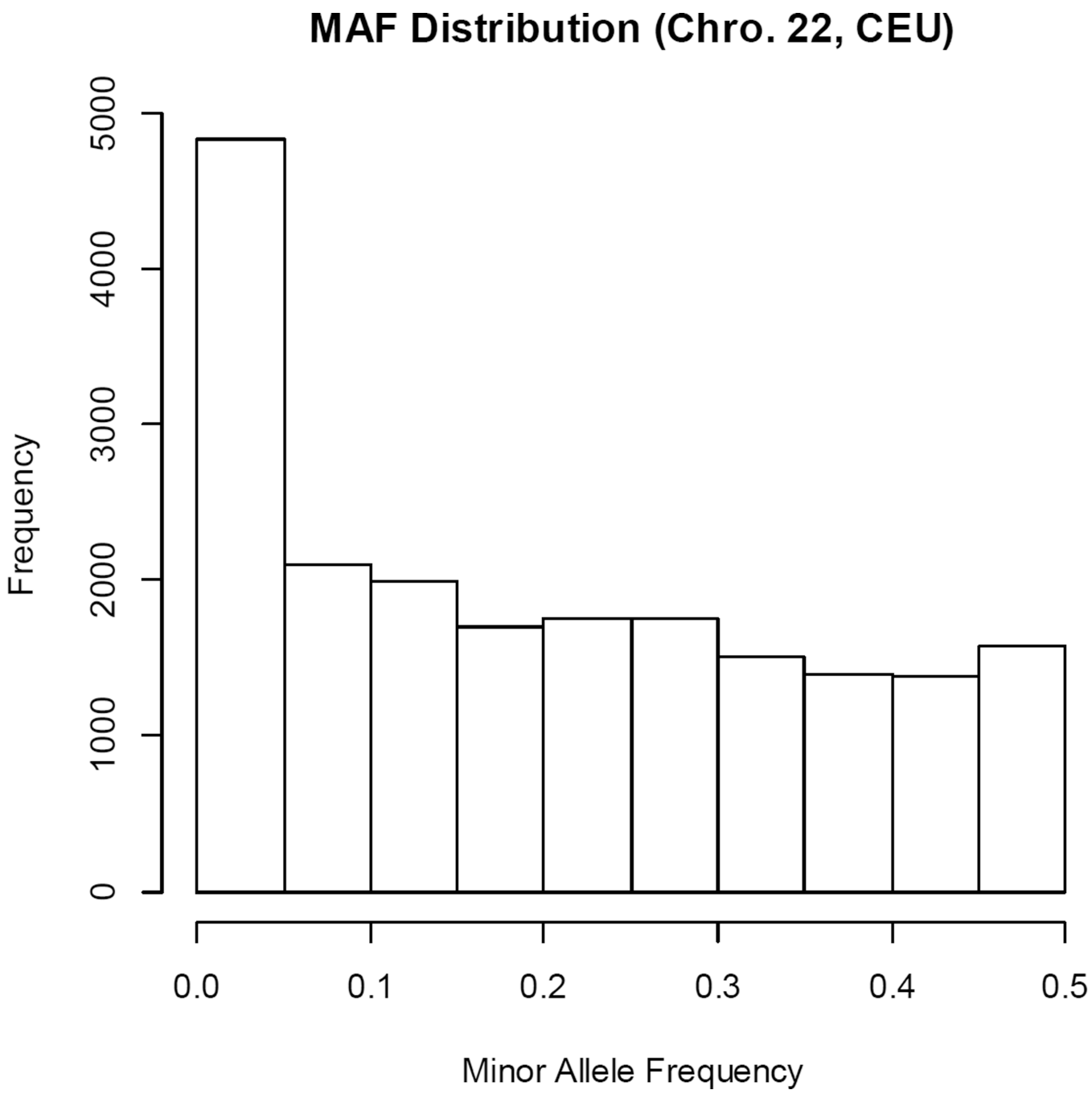

Figure 1.
Distribution of minor allele frequencies for variants on chromosome 22 of CEU populations

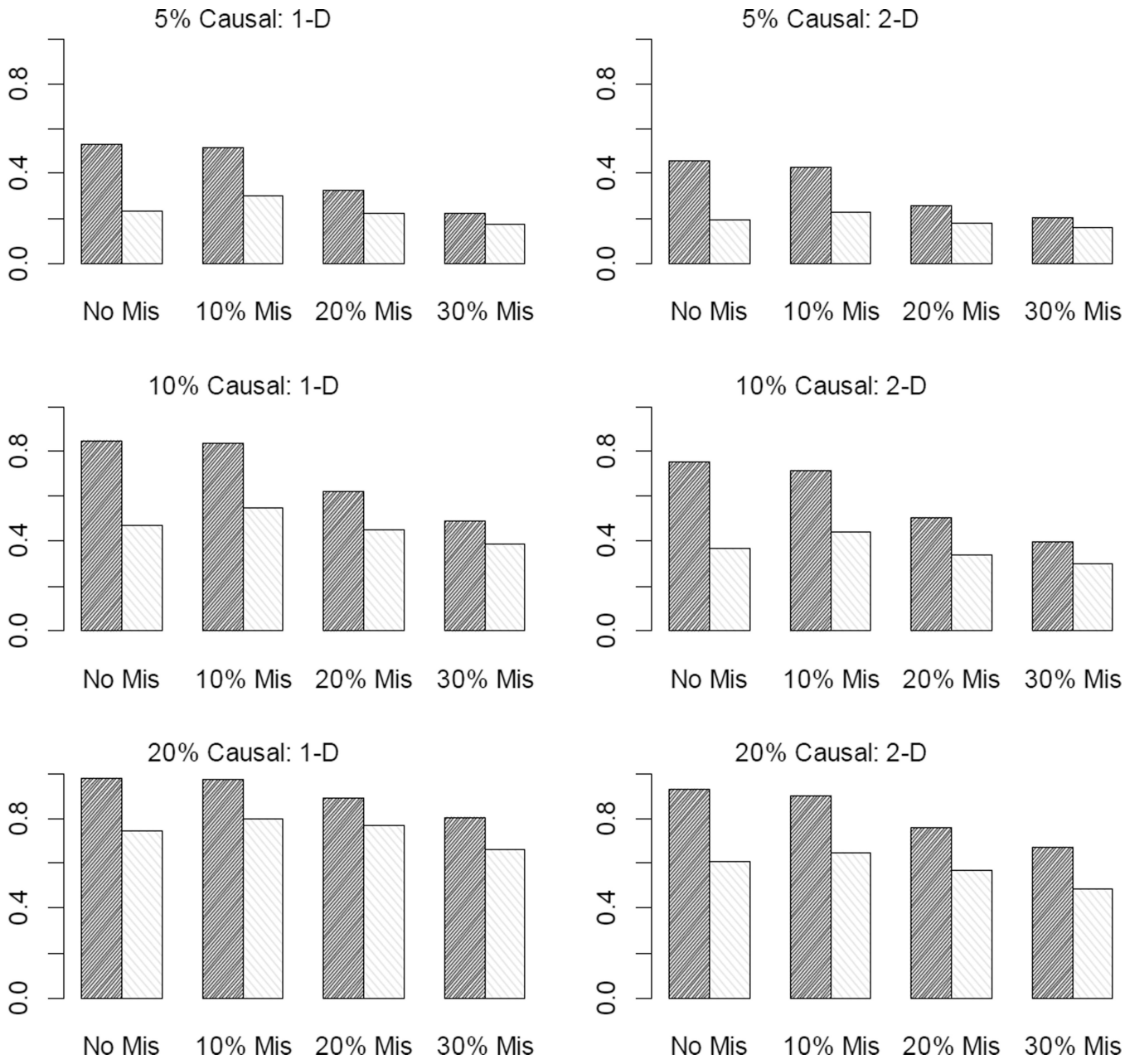

Figure 2.

Power evaluation of two methods with normally distributed phenotypes by varying the proportion of causal variants and missing genotypes. Three missing genotypes scenarios include: 1) ten percent randomly missing SNPs, 2) twenty percent randomly missing subjects, and 3) thirty percent randomly missing subjects.

Dark color: Power for FB-GGRF

Light color: Power of FB-SKAT

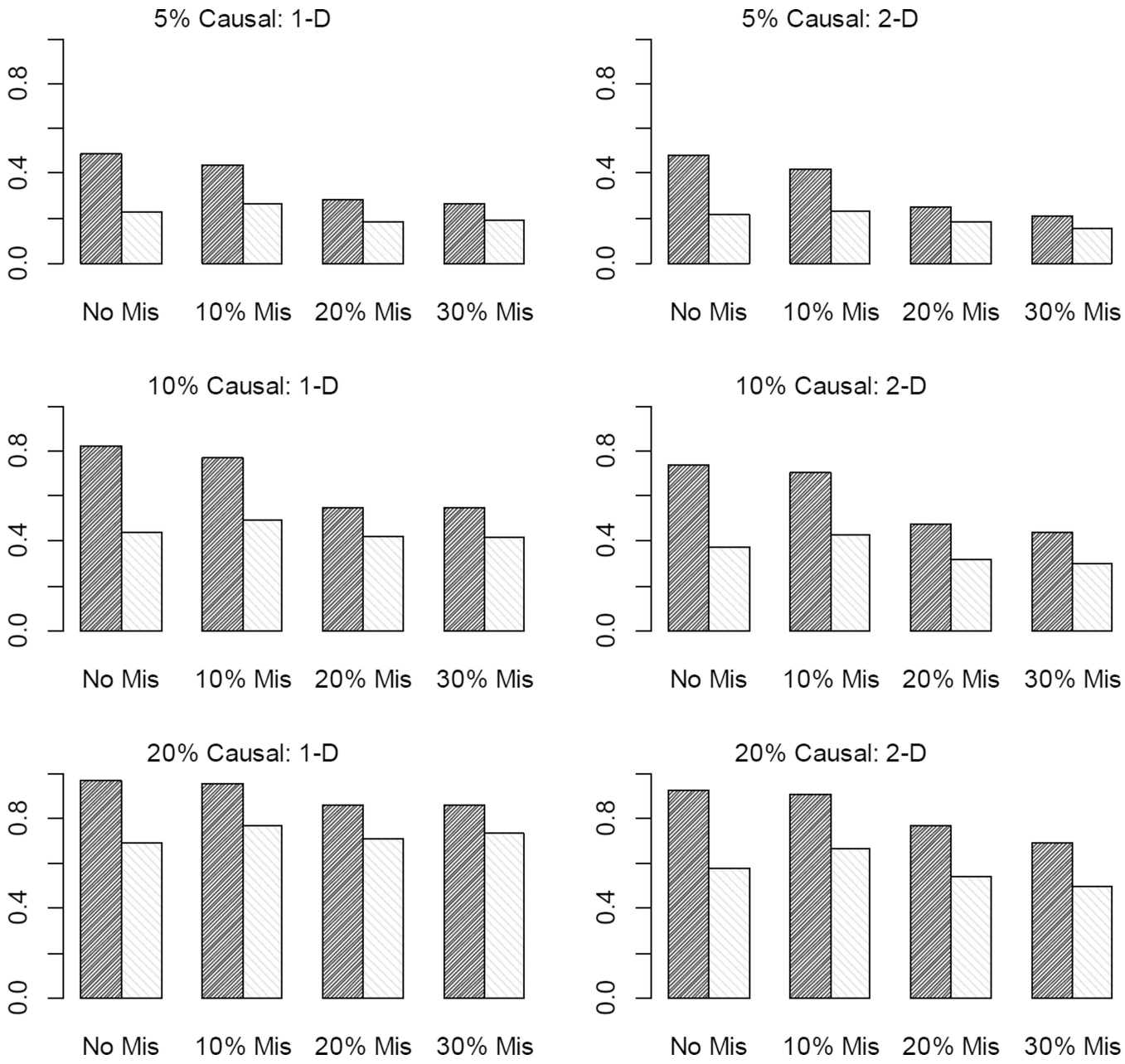

Figure 3.

Power evaluation of two methods with binary phenotypes (case-parent/control-parent trios design) by varying the proportion of causal variants and missing genotypes. Three missing genotypes scenarios include: 1) ten percent randomly missing SNPs, 2) twenty percent randomly missing subjects, and 3) thirty percent randomly missing subjects.

Dark color: Power for FB-GGRF

Light color: Power of FB-SKAT

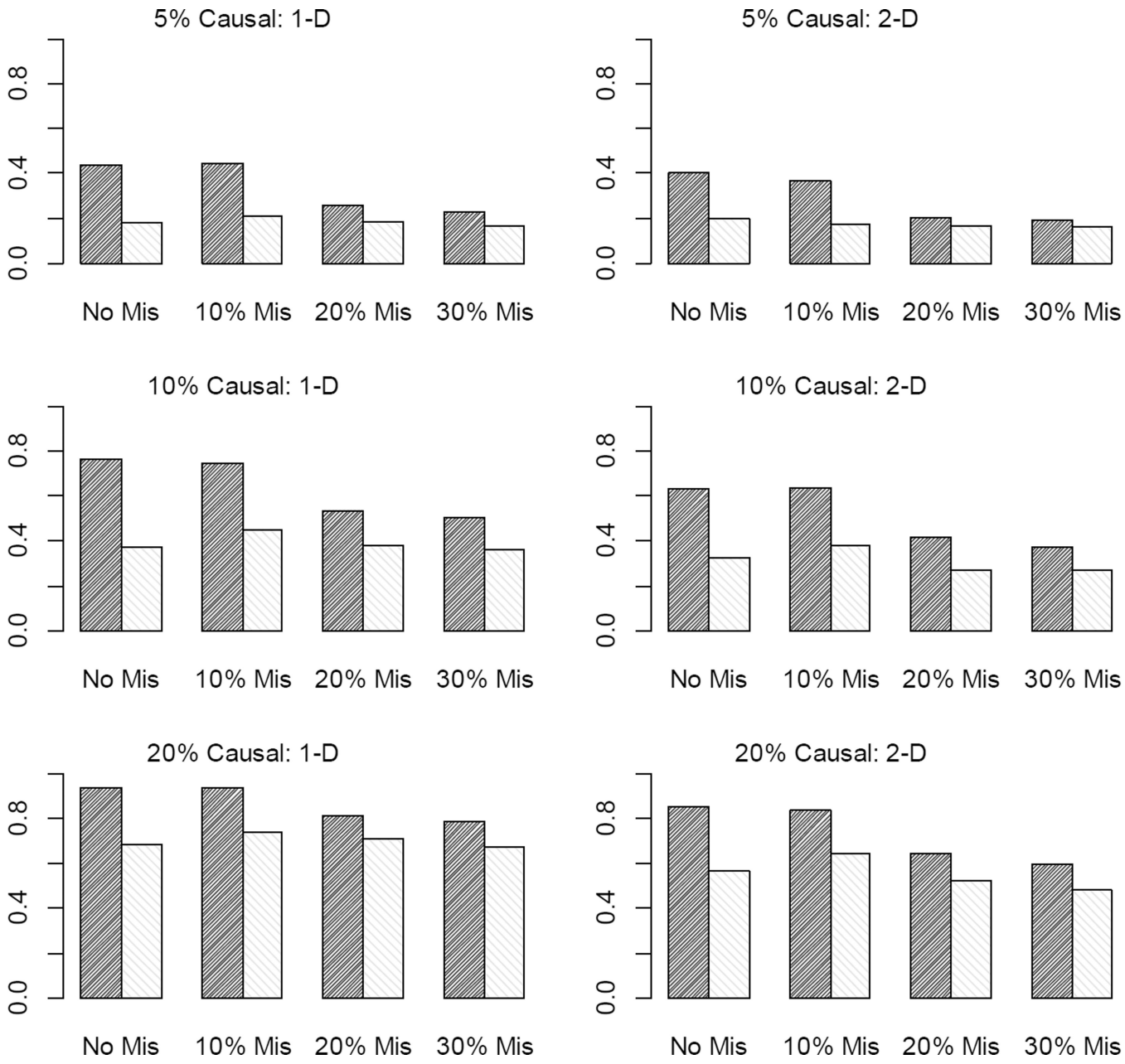

Figure 4.

Power evaluation of two methods with case-only phenotype (case-parent trios design) by varying the proportion of causal variants and missing genotypes. Three missing genotypes scenarios include: 1) ten percent randomly missing SNPs, 2) twenty percent randomly missing subjects, and 3) thirty percent randomly missing subjects.

Dark color: Power of Hotelling's T^2 test

Light color: Power of FB-SKAT

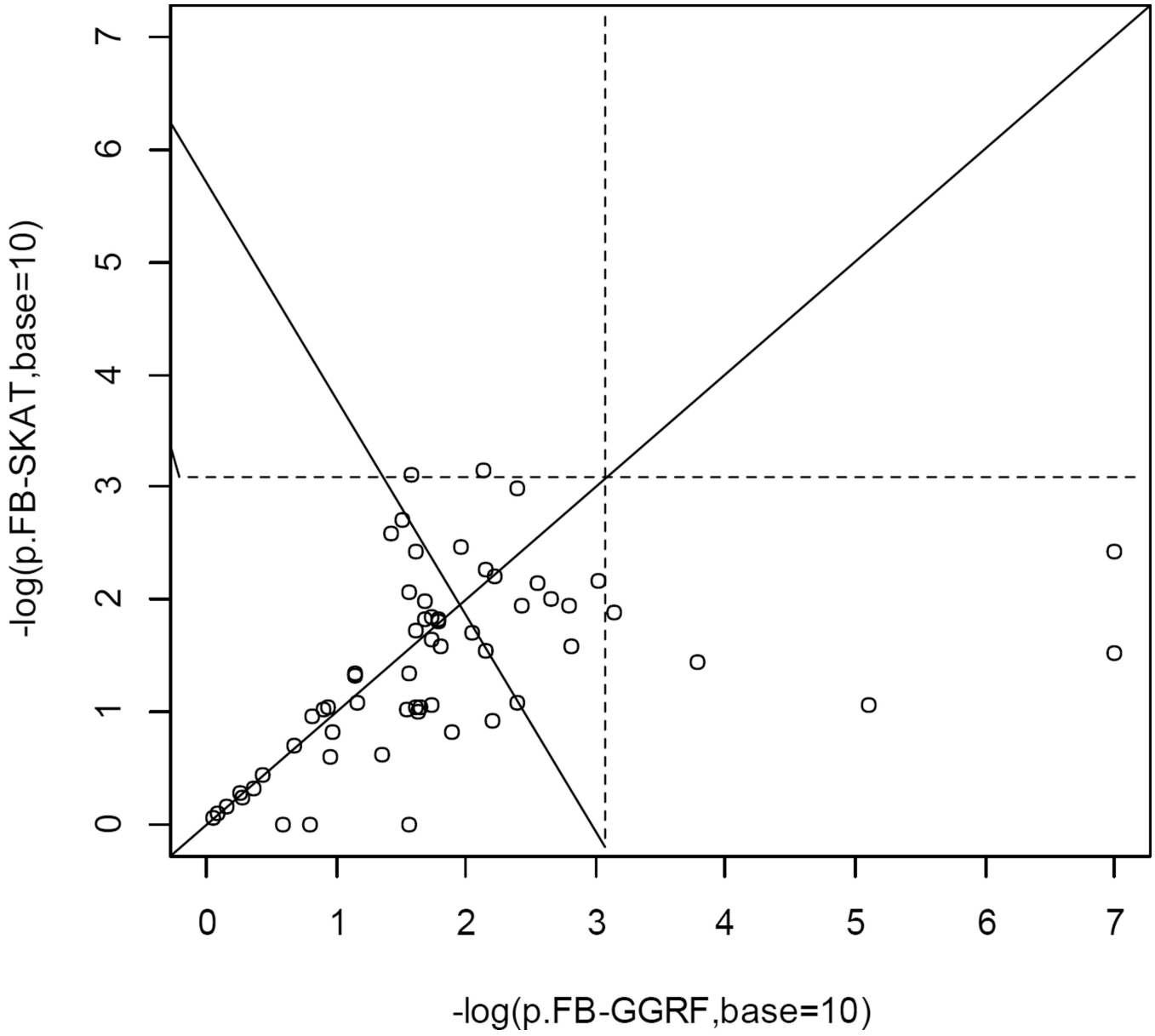

Figure 5.

Comparison of testing p-values for 60 genes in the CHD study by using FB-GGRF and FB-SKAT (ten-based negative logarithms scale)

Dashed line: Bonferroni threshold for 60 genes

Genet Epidemiol. Author manuscript; available in PMC 2017 May 01.

Table 1

Type I error evaluation for family trios with quantitative phenotypes, case-parent and control-parent trios, and case-parent trios only at various significance levels.

No PS[1]	**Quantitative Phenotype**		**Case-Parent/Control-Parent**		**Case-Parent Only**	
	FB-GGRF	**FB-SKAT**	**FB-GGRF**	**FB-SKAT**	**Hotelling T**[2]	**FB-SKAT**
No missing genotypes						
T1E=0.05	0.0505	0.0473	0.0486	0.0462	0.0501	0.0492
T1E=0.01	0.0096	0.0088	0.0097	0.0095	0.0135	0.0103
T1E=0.001	0.0009	0.0007	0.0011	0.0011	0.0025	0.0013
10% missing genotypes						
T1E=0.05	0.0493	0.0485	0.0512	0.0469	0.0517	0.0465
T1E=0.01	0.0099	0.0103	0.0106	0.0096	0.0157	0.0097
T1E=0.001	0.0009	0.0013	0.0009	0.0011	0.0027	0.0010
20% missing subjects						
T1E=0.05	0.0486	0.0481	0.0491	0.0466	0.0501	0.0482
T1E=0.01	0.0100	0.0099	0.0096	0.0099	0.0137	0.0098
T1E=0.001	0.0014	0.0012	0.0012	0.0012	0.0026	0.0011
30% missing subjects						
T1E=0.05	0.0489	0.0477	0.0514	0.0486	0.0487	0.0467
T1E=0.01	0.0102	0.0102	0.0098	0.0098	0.0137	0.0102
T1E=0.001	0.0001	0.0013	0.0009	0.0010	0.0028	0.0010
PS[2]	Quantitative Phenotype		Case-Parent/Control-Parent		Case-Parent Only	
	FB-GGRF	FB-SKAT	FB-GGRF	FB-SKAT	Hotelling T[2]	FB-SKAT
No missing genotypes						
T1E=0.05	0.0515	0.0512	0.0497	0.0493	0.0506	0.0478
T1E=0.01	0.0115	0.0100	0.0104	0.0102	0.0140	0.0096

No PS[1]	Quantitative Phenotype		Case-Parent/Control-Parent		Case-Parent Only	
	FB-GGRF	FB-SKAT	FB-GGRF	FB-SKAT	Hotelling T^2	FB-SKAT
T1E=0.001	0.0013	0.0010	0.0011	0.0014	0.0028	0.0010
10% missing genotypes						
T1E=0.05	0.0513	0.0465	0.0511	0.0476	0.0515	0.0503
T1E=0.01	0.0110	0.0096	0.0101	0.0096	0.0160	0.0147
T1E=0.001	0.0012	0.0012	0.0009	0.0011	0.0025	0.0016
20% missing subjects						
T1E=0.05	0.0511	0.0494	0.0492	0.0462	0.0507	0.0470
T1E=0.01	0.0107	0.0101	0.0095	0.0097	0.0138	0.0098
T1E=0.001	0.0010	0.0009	0.0012	0.0012	0.0024	0.0013
30% missing subjects						
T1E=0.05	0.0521	0.0433	0.0500	0.0442	0.0519	0.0500
T1E=0.01	0.0109	0.0117	0.0101	0.0121	0.0148	0.0100
T1E=0.001	0.0010	0.0012	0.0009	0.0011	0.0029	0.0010

[1] No Population Stratification, CEU samples only

[2] Population Stratification, both CEU samples and ASW samples.

Table 2

Maternal characteristics for 616 case families and 1,645 control families enrolled in National Birth Defects Prevention Study, 1999–2008

	Case (N=616)	Control (N=1,645)
Age at delivery (years)	28.3 ± 6.1	27.5 ± 6.0
Race		
African American	49 (8%)	143 (9%)
Caucasian	401 (66%)	1,136 (69%)
Hispanic	123 (20%)	285 (17%)
Others	39 (6%)	78 (5%)
Missing information	4	3
Education		
<12 years	83 (14%)	217 (13%)
High school degree or equivalent	167 (27%)	413 (25%)
1–3 years of college	173 (28%)	454 (28%)
At least 4 years of college or Bachelor degree	190 (31%)	559 (34%)
Missing information	3	2
Household income		
Less than 10 Thousand	94 (16%)	236 (15%)
10 to 30 Thousand	150 (26%)	408 (27%)
30 to 50 Thousand Dollars	118 (20%)	348 (23%)
More than 50 Thousand	217 (37%)	538 (35%)
Missing information	37	115
Folic acid supplementation		
No	299 (49%)	738 (45%)
Yes[1]	314 (51%)	907 (55%)
Missing information	3	0
BMI		
Underweight (BMI <18.5)	31 (5%)	74 (5%)
Normal weight (18.5 <=BMI <25)	298 (50%)	880 (55%)
Overweight (25 <=BMI <30)	141 (24%)	360 (23%)
Obese (>=30)	121 (20%)	281 (18%)
Missing information	25	50
Alcohol consumption		
No	460 (76%)	1,251 (76%)
Yes[2]	149 (24%)	390 (24%)
Missing information	7	4
Cigarette smoking		
No	498 (81%)	1,356 (82%)
Yes[3]	114 (19%)	288 (18%)
Missing information	4	1

[1]Folic acid supplementation yes is defined as periconceptional folate supplement use at least two months during the exposure window (i.e. one month prior to conception and two months after conception).

[2]Alcohol consumption yes is defined as drinking during the three months after pregnancy.

[3]Cigarette smoking yes is defined as smoking during the three months after pregnancy.

Table 3

Genes were found to be significantly associated with CHD risk after Bonferroni correction using FB-GGRF and FB-SKAT, respectively.

Gene	# of SNPs	Chromosome	Pathway	*P*-value (FB-GGRF)	*P*-value (FB-SKAT)
PGDS	17	4	Transsulfuration	**1.00e-07**	3.72e-03
GSR	11	8	Transsulfuration	**1.00e-07**	0.031
*BHMT*2	11	5	Homocysteine	**7.80e-06**	0.089
*TRDMT*1	32	10	Homocysteine	**1.62e-04**	0.037
*GPX*4	3	19	Transsulfuration	**7.21e-04**	0.013
*GSTA*2	7	6	Transsulfuration	0.026	**7.8e-04**
*MGST*1	28	12	Transsulfuration	7.4e-03	**7.1e-04**

PGDS hematopoietic prostaglandin d synthase; *GSR*: glutathione reductase; *BHMT*2: betaine--homocysteine S-methyltransferase 2; *TRDMT*1: tRNA aspartic acid methyltransferase 1; *GPX*4: glutathione peroxidase 4; *GSTA*2: glutathione S-transferase alpha 2; *MGSMT*1: microsomal glutathione S-transferase 1