A unified method for rare variant analysis of gene-environment interactions

Elise Lim,[1] Han Chen,[2,3] Josée Dupuis,[1] Ching-Ti Liu[1]

[1] Department of Biostatistics, Boston University, Boston, Massachusetts, 02218, USA

[2] Human Genetics Center, Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas 77030, USA

[3] Center for Precision Health, School of Public Health and School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas 77030, USA

*Correspondence: elise625@bu.edu, phone #: 516-417-0771, fax #: 617-638-6484 (EL) and ctliu@bu.edu, phone #: 617-638-7752, fax #: 617-638-6484 (CTL)

Abstract

Advanced technology in whole-genome sequencing has offered the opportunity to comprehensively investigate the genetic contribution, particularly rare variants, to complex traits. Many rare variants analysis methods have been developed to jointly model the marginal effect but methods to detect gene-environment (GE) interactions are underdeveloped. Identifying the modification effects of environmental factors on genetic risk poses a considerable challenge. To tackle this challenge, we develop a unified method to detect GE interactions of a set of rare variants using generalized linear mixed effect model. The proposed method can accommodate both binary and continuous traits in related or unrelated samples. Under this model, genetic main effects, sample relatedness and GE interactions are modeled as random effects. We adopt a kernel-based method to leverage the joint information across rare variants and implement variance component score tests to reduce the computational burden. Our simulation study shows that the proposed method maintains correct type I error rates and high power under various scenarios, such as differing the direction of main genotype and GE interaction effects and the proportion of causal variants in the model for both continuous and binary traits. We illustrate our method to test gene-based interaction with smoking on body mass index or overweight status in the Framingham Heart Study and replicate the *CHRNB4* gene association reported in previous large consortium meta-analysis of single nucleotide polymorphism (SNP)-smoking interaction. Our proposed set-based GE test is computationally efficient and is applicable to both binary and continuous phenotypes, while appropriately accounting for familial or cryptic relatedness.

Keywords: rare variant analysis, gene x environment interaction, generalized linear mixed model, variance components score test, sequence kernel association test, family data

1. Introduction

Although genome-wide association studies (GWAS) have been successful in identifying genetic variants with strong association with disease, the variants evaluated in GWAS have been mostly restricted to common variants, typically defined as those with minor allele frequency (MAF) greater than 1 or 5%. Additionally, these identified variants only explain a small portion of disease heritability, which could be due to limited sample size and power in GWAS, and thus calls for performing meta-analysis.[1] However, even in large scale GWAS meta-analyses, much of the heritability remains unexplained. For example, in GWAS meta-analysis of adult height in > 93,000 East Asians, they identified 98 loci at genome-wide significance that only explain about 9% of height heritability.[2]

One plausible explanation for missing heritability is the presence of rare variants, which are not analyzed in GWAS due to their low MAF. It is well known that rare variants are responsible for many Mendelian disorders, but they have not been fully implicated in complex diseases.[3,4] With rapid advances and decreasing cost of whole-genome sequencing, attention has been shifted to investigating the potential role of low frequency variants in complex human disease etiology. There is now an abundance of growing empirical evidence that rare variants may be responsible for complex diseases, which might account for missing heritability.[4,5,6] For example, Igartua et al.[7] identified two novel rare variants associated with low-density lipoprotein cholesterol and high-density lipoprotein cholesterol with larger effects than the previously discovered variants within the known blood lipid associated loci. In another study by He et al.,[8] multiple rare variants were found to be associated with lower systolic blood pressure. Successes from rare variant association studies highlight the importance of rare variants to complex disease susceptibility.

When single variant tests are applied to rare genetic variants, they are severely underpowered, and hence several statistical methods have been developed to increase power. These methods are usually region-based tests that study the joint effects of rare variants in a specified region and can be broadly categorized into burden tests and non-burden tests. For burden tests, the cumulative effects of variants in a region are summarized into a single variable which is then tested for association with the trait of interest. Variants can be collapsed by summing up the number of rare alleles in a region or by using dichotomous variable to indicate whether an individual has any rare alleles in that region. Other extensions of burden tests have been developed. [8,9,10] Burden tests are most powerful when variants in consideration are causal and the direction of the effect on risk of the alternate/minor allele is the same. When the alternate/minor alleles of causal variants have effects on the phenotype in different direction, and/or the proportion of causal variants is small, burden tests suffer from low power.[4,12,13] To address this issue, non-burden tests have been proposed that focus on aggregating individual test statistics. Among non-burden tests, the sequence kernel association test (SKAT), which is a popular method where variant information is summarized using a kernel function, uses variance component score test to evaluate the significance.[14] SKAT is powerful when there are variants in mixed directions or when the proportion of causal variant is small, but it is less powerful than the burden tests when the majority of variants in a region are causal with effects in the same direction.[14,15]

Another plausible explanation for missing heritability is due to gene-environment (GE) interactions. Complex diseases are multifactorial and involve both genetics and environmental factors. Therefore, studying just the main effects, either genetic or environmental, cannot provide full insights into the biological mechanisms. Studying GE interactions can provide better insights into the biological mechanisms of complex diseases, which help to identify subgroups that are at high risk and eventually lead to better diagnostics and personalized treatments.[16,17] Typically, GE tests require larger sample sizes than a main effects test to achieve appropriate power, and rare variant analysis also needs bigger sample compared to common variant analysis to maintain power. [17,18,19] Tzeng et al.[19] developed similarity-based regression method (SimReg) to test GE interaction effects of rare variants for continuous traits. SimReg allows for covariate adjustments, models both main and interaction effects, and is computationally efficient. Zhao et al.[17] extended SimReg to allow for binary traits for common or rare variants. Lin et al.[20] introduced rare variant by environment interaction method using a variance component test under the generalized linear model framework. Chen et al.[18] proposed two GE interaction tests (rareGE) and a joint test of main effect and interaction effect for rare variants using variance component score tests. In rareGE, genetic variants can be included as fixed or random effects for the interaction test and rareGE works for both binary and continuous traits. The aforementioned methods are only applicable to unrelated individuals and cannot correctly account for familial correlation in their models. Recently, Mazo Lopera et al.[21]

developed SNP-set GE interaction method for family data but it was proposed in the context of common variants.

In this article, we introduce a unified approach for GE interaction tests for rare variants (famGE) to correctly incorporate sample relatedness. Our approach can accommodate both binary and continuous traits in family data, eliminating the need to restrict samples to just unrelated individuals. We adopted a kernel-based method to leverage the joint information across rare variants. We assume main genetic variants, GE interaction term, and family correlation to be random effects and implemented a variance component score test in the generalized linear mixed model (GLMM) framework to reduce the computational burden.

The paper is organized as follows. In section 2, we introduce our notation, the GE interaction model, and the test statistic. In section 3, we conduct extensive simulations under various settings to assess type 1 error rates and power of our approach, comparing it to rareGE and the burden test. We apply our method in testing gene by smoking interaction on body mass index (BMI), using family data from the Framingham Heart Study (FHS) in section 4. We discuss our findings of our approach in section 5.

2. Methods

Notations and model settings for interaction test for family data

We first introduce our notation and assumptions before we derive our test statistic. Assuming a sample of size n, let Y_i be the phenotype (binary or continuous) for individual i with $E(Y_i) = \mu_i$ and $var(Y_i) = \phi v(\mu_i)$, where ϕ is the dispersion parameter (1 for binary and Poisson data), and $v(\cdot)$ is the variance function. Let $\boldsymbol{X}_i = \left(1, E_i, X_{i2}, \dots, X_{i(p-1)}\right)^{\mathrm{T}}$ be a p x 1 vector of covariates including the intercept and the environmental variable E_i, and $\boldsymbol{G}_i = \left(G_{i1}, G_{i2}, \dots, G_{iq}\right)^T$ be a q x 1 vector of genetic main effects, and d_i be the random effect of familial correlation. We consider the following GLMM for testing GE interactions:

$$g(E(Y_i)) = g(\mu_i) = \boldsymbol{X}_i^T\boldsymbol{\alpha} + \boldsymbol{G}_i^T\boldsymbol{W}_1\boldsymbol{\theta} + E_i\boldsymbol{G}_i^T\boldsymbol{W}_2\boldsymbol{\gamma} + d_i, \qquad (1)$$

where $g(\cdot)$ is the link function, $\boldsymbol{\alpha}$ is a p x 1 vector associated with the fixed covariate effects, $\boldsymbol{\theta}$ is a q x 1 vector of random effects for the genetic variants, $\boldsymbol{\gamma}$ is a q x 1 vector of random effects for GE interaction, and $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are q x q diagonal matrices with weights for genetic main effects and GE interaction effects, respectively. The matrices $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ can be any positive semi-definite matrices. We can rewrite this model in matrix notation as:

$$g(\boldsymbol{E}(\boldsymbol{Y})) = g(\boldsymbol{\mu}) = \boldsymbol{X\alpha} + \boldsymbol{GW}_1\boldsymbol{\theta} + \boldsymbol{EGW}_2\boldsymbol{\gamma} + \boldsymbol{d}, \qquad (2)$$

where $\boldsymbol{\mu} = [\mu_1 \dots \mu_n]^{\mathrm{T}}$ is the mean vector, $\boldsymbol{X} = [\boldsymbol{X}_1 \dots \boldsymbol{X}_n]^{\mathrm{T}}$ is an n x p covariate matrix, $\boldsymbol{G} = [\boldsymbol{G}_1 \dots \boldsymbol{G}_n]^T$ is an n x q genotype matrix, $\boldsymbol{EG} = [\boldsymbol{EG}_1 \dots \boldsymbol{EG}_n]^T$ is an n x q GE interaction matrix, and $\boldsymbol{d} = [d_1 \dots d_n]^T$ is an n x 1 vector for the random effects of familial correlation. For the three random effects (genetic variants, GE interactions, and relatedness in families respectively), we assume that

$$\boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{0}, \sigma_M^2\boldsymbol{I}_q)$$

$$\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}, \sigma_I^2 \boldsymbol{I_q})$$

$$\mathbf{d} \sim \mathrm{N}(\mathbf{0}, \sigma_G^2 \boldsymbol{\phi})$$

where $\boldsymbol{\phi}$ is twice the n x n kinship matrix from family relationships obtained from a pedigree or an empirical kinship calculated from genotype data to account for cryptic relatedness. The random effects $\boldsymbol{\theta}, \boldsymbol{\gamma},$ and $\boldsymbol{d}$ are assumed to be independent from each other. If $\boldsymbol{\gamma}$ is treated as a fixed effect, we would perform a q degrees of freedom score test, but this test can lead to power loss when q is large.[12] By assuming that $\boldsymbol{\gamma}$ follows a normal distribution with mean 0 and variance $\sigma_I^2 \boldsymbol{I_q}$, the null hypothesis for the interaction test: H_0: $\boldsymbol{\gamma} = 0$ is equivalent to testing H_0: $\sigma_I^2 = 0$ with a variance component score test.[12,14] Score tests only require fitting the model under the null hypothesis, so they are more computationally efficient.

Estimation and hypothesis testing

To fit the null model for binary traits, we use the penalized quasi-likelihood method. Because the integrated quasi-likelihood function for binary traits contains a high dimensional integral that is intractable, we use the Laplacian method to approximate the integral. To estimate the parameters, we define the linear working vector under the null hypothesis $\boldsymbol{Y_0} = \boldsymbol{X\alpha} + \boldsymbol{GW_1\theta} + \boldsymbol{d} + \boldsymbol{\Delta}(\boldsymbol{y} - \boldsymbol{\mu})$, where $\boldsymbol{\Delta} = diag\{g'(\mu_i)\}$, and let $\boldsymbol{D} = diag\left\{\frac{1}{\phi v(\mu_i)[g'(\mu_i)]^2}\right\}$. We iteratively fit the working vector to estimate the parameters until convergence to obtain our restricted maximum likelihood (or maximum likelihood) estimates.[22]

Let $\boldsymbol{\hat{\alpha}}$, $\hat{\sigma}_M^2$, and $\hat{\sigma}_G^2$ be the restricted maximum likelihood (or maximum likelihood) estimates under the null hypothesis. We can then calculate $\boldsymbol{\hat{\mu}}$, $\boldsymbol{\hat{\Delta}}$ and $\boldsymbol{\hat{D}}$ and $\boldsymbol{\hat{Y}_0}$ using the aforementioned restricted maximum likelihood (or maximum likelihood) estimates. The restricted maximum quasi-likelihood function is defined as:

$$ql_R = -\frac{1}{2}log|\boldsymbol{\Sigma}| - \frac{1}{2}log\left|\boldsymbol{X^T\Sigma^{-1}X}\right| - \frac{1}{2}(\boldsymbol{Y_0} - \boldsymbol{X}\alpha)^{\boldsymbol{T}}\boldsymbol{\Sigma^{-1}}(\boldsymbol{Y_0} - \boldsymbol{X}\alpha) \qquad (3)$$

where $\boldsymbol{\Sigma} = \boldsymbol{D^{-1}} + \sigma_M^2(\boldsymbol{GW_1W_1G^T}) + \sigma_G^2\boldsymbol{\phi} + \sigma_I^2\left(\boldsymbol{EGW_2W_2G^TE}\right)$.

To derive the score test for *H₀*: $\sigma_I^2 = 0$, we take the first derivative of the quasi-likelihood with respect to $\sigma_I^2 = 0$,

$$\frac{\partial ql_R}{\partial\sigma_I^2} = -\frac{1}{2}tr(\boldsymbol{\Sigma^{-1}EGW_2W_2G^TE}) + \frac{1}{2}(\boldsymbol{Y_0} - \boldsymbol{X\alpha})^{\boldsymbol{T}}\boldsymbol{\Sigma^{-1}EGW_2W_2G^TE\Sigma^{-1}}(\boldsymbol{Y_0} - \boldsymbol{X\alpha}) \qquad (4)$$

The first term in equation 4 is fixed and independent of the phenotype. We follow the same rationale in the derivation of the SKAT score statistic and take twice the second term to be our test statistic[14,23]

$$\boldsymbol{Q} = \left(\boldsymbol{\hat{Y}_0} - \boldsymbol{X\hat{\alpha}}\right)^{\boldsymbol{T}}\boldsymbol{\hat{\Sigma}^{-1}EGW_2W_2G^TE\hat{\Sigma}^{-1}}(\boldsymbol{\hat{Y}_0} - \boldsymbol{X\hat{\alpha}}) \qquad (5)$$

where $\boldsymbol{\hat{\Sigma}} = \boldsymbol{\hat{D}^{-1}} + \hat{\sigma}_M^2(\boldsymbol{GW_1W_1G^T}) + \hat{\sigma}_G^2\boldsymbol{\phi}$.

Under the null hypothesis, $\boldsymbol{Q} \sim \sum_{j=1}^{q}\lambda_j\chi_{1,j}^2$, where λ_j's are eigenvalues of the matrix $\boldsymbol{W_2G^TE}(\boldsymbol{\hat{\Sigma}^{-1}} - \boldsymbol{\hat{\Sigma}^{-1}X}\left(\boldsymbol{X^T\hat{\Sigma}^{-1}X}\right)^{\boldsymbol{-1}}\boldsymbol{X^T\hat{\Sigma}^{-1}})\boldsymbol{EGW_2}$.[14,24] See Appendix 1 for a detailed derivation.

3. Simulation studies

Simulation settings for type I error evaluation

We performed simulation studies to evaluate type I error rates for our proposed approach. For the null simulation study, we first considered the scenario where there are main genotype effects but no GE interaction effects. To simulate the genotypes, we used SeqSIMLA software and used reference sequence based on 1000 Genomes Project for European populations.[25] We used 7 pedigrees from FHS with family membership ranging from 120 to 640 (2030 individuals in total). In the simulated genotypes, we chose a region that spans from 1,100 base pairs to 1,140 base pairs in chromosome 1. To simulate our phenotype, we varied the proportion of low frequency (MAF<5%) causal SNPs included in our model to 20% to 40% to 60% and 80% for each of 20,000 replicates. We considered both continuous and binary phenotypes. For each of 20,000 replicates, we simulated phenotype datasets from

$$\boldsymbol{y} = 0.1 + 0.1\boldsymbol{age} + 0.5\boldsymbol{sex} + 0.3\boldsymbol{smoke} + \boldsymbol{d} + \boldsymbol{G\gamma} + \boldsymbol{\epsilon},$$

where ***age*** was generated from a normal distribution with mean of 50 and standard deviation equal to 5, ***sex*** was generated from a Bernoulli distribution with probability 0.5, ***smoke*** was generated from a Bernoulli distribution with probability 0.5, and $\boldsymbol{\epsilon}$ was generated from a standard normal distribution. Family correlation, *d*, is from a multivariate normal distribution with means 0 and covariance $\sigma_G^2\boldsymbol{\phi}$, where σ_G^2 is set to 1 and $\boldsymbol{\phi}$ is twice the kinship matrix. $\boldsymbol{\gamma}$ consist of effect sizes for the causal SNPs. For binary traits, we simulated the continuous traits and set the lower 80% as controls (0's) and the upper 20%

as cases (1's). We simulated variants where the directions of the effect on risk of the minor allele are both same and opposite and the effect sizes were determined by

$$\gamma_i = \frac{h}{2MAF_i(1-MAF_i)}$$

where MAF is the minor allele frequency of SNP *i*. The constant h is calculated as

$$h = \frac{R^2}{v^T L v}$$

where R^2, the proportion of variance explained by the causal SNPs, was fixed at 1% for causal SNPs with effect sizes in same direction and 5% for causal SNPs with effect sizes in opposite directions. The correlations between the SNPs are in matrix L, and v is a vector that indicates the direction of the SNP effects. We performed three tests for the type I error comparison: our proposed method that correctly accounts for familial correlation (famGE), rareGE where the main genetic effects are treated as random and family correlation is ignored (rareGE), and the burden test using GLMMs to account for familial correlation (BT). The summary variable used in the burden test was created using an indicator of whether or not the rare allele was present in the testing region. We simulated 20,000 replicates and used Wu weights, which are the beta density function with parameters 1 and 25 evaluated at the MAF of the variants, for famGE and rareGE tests.[14]

Simulation settings for power analysis

To assess power, we simulated data under the alternative hypothesis, where we include a gene by smoking interaction effect in addition to the main genotype effects. Similar to the type I error simulation, genotypes were simulated using the SeqSIMLA software and we varied the direction of genetic main effect and the proportion of causal SNPs with MAF less than 5%. We simulated 10,000 phenotype datasets from

$$\boldsymbol{y} = 0.1 + 0.1\boldsymbol{age} + 0.5\boldsymbol{sex} + 0.3\boldsymbol{smoke} + \boldsymbol{d} + \boldsymbol{G\gamma} + (\boldsymbol{smoke} - \mathbf{0.5})\boldsymbol{G\delta} + \boldsymbol{\epsilon},$$

where age, sex, and smoke, family correlation, and error terms were generated from the same distribution described in the type I error simulation study and the genotype effects $\boldsymbol{\gamma}$ were determined the same way as in our null simulation study. Interaction effects $\boldsymbol{\delta}$ were generated from a normal distribution with mean 2 and standard deviation of 0.3. We considered the scenarios where directions of the main effects and the interaction effects on the risk of the minor allele are same and opposite. Negative interaction effects were simulated the same way as above except we multiplied the effects by -1. To test binary outcomes, we set the lower 80% of the simulated continuous outcome to be the controls and the upper 20% to be the cases. For power comparison, we also performed a burden test, where the summary variable for each individual was created using an indicator of whether or not the rare allele was present in the testing region.

Results for type 1 error simulations

Table 1 includes the type 1 error results for famGE, rareGE, and BT at significance levels, α, of 0.05, 0.01, and 0.001 from 20,000 simulation replicates. Both famGE and BT have correct type 1 error rates at all three α levels for both continuous and binary traits. When familial correlation is not appropriately taken into account in the model, rareGE test suffers from type 1 error inflation, which is more pronounced in binary traits. FamGE has valid type 1 error rates under various scenarios, such as differing the direction of main genotype effects and increasing the proportion of causal variants in the model for both continuous and binary traits. Because rareGE does not have correct type 1 error rate when familial correlation is not taken into account, we did not include rareGE in the power comparison in the next section.

Results for power simulations

Figure 1 shows the power comparison for famGE and BT for continuous traits at α = 0.001 from 10,000 replicates. When the proportion of causal variants is low, BT has lower power than famGE (**1A)**. This is expected because burden tests are powerful when a large proportion of causal variants are included in the region. When the proportion of causal variants increases, we see that powers for both BT and famGE increase but we see a larger power increase in BT compared to famGE. However, when variants have interaction effects in different direction, we observe a power drop, with power close to 0 for BT, whereas famGE shows increasing power as proportion of causal variants increases (**1B)**. FamGE maintains fairly consistent power across different scenarios, regardless of the direction of the main effects.

Figure 2 shows power comparisons of famGE and BT for binary traits at α = 0.001 from 10,000 replicates. We come to the same conclusion as the results seen from continuous traits. Even though burden tests have higher power than famGE in the case where interaction effects are in the same direction, with the exception where there are 4 causal variants and 16 neutral variants in the model, famGE is able to maintain high power regardless of the direction or the proportion of variants (**2A**). The burden test significantly loses power in the presence of both trait-increasing and decreasing alleles, whereas performance of famGE is robust across different scenarios (**2B**).

4. Application to the Framingham Heart Study

In the real data application, we illustrate our method to test gene-based interaction with smoking on a quantitative trait, BMI, and a dichotomous trait, overweight status (BMI ≥ 25), using participants from the Framingham Heart Study (FHS). FHS is a longitudinal cardiovascular cohort study that was initiated in 1948 in Framingham, Massachusetts. The study began with 5209 participants (Original Cohort) in 1948 and in 1971, 5124 children of the original cohort and their spouses (Offspring Cohort) were enrolled. A total of 4095 additional participants, who were the grandchildren of the original cohort, were recruited in the third generation cohort (Gen 3 Cohort) in 2002.[26] To reflect ethnic diversity in Framingham, 917 additional non-white participants were recruited to form the Omni Cohort.[27]

Obesity is a world-wide problem that can lead to serious health problems, such as high blood pressure, type 2 diabetes, heart disease. Conventionally, BMI has been used to measure obesity. GWAS have identified numerous loci associated with BMI, but whether these genetic effects are modulated by environmental factors has not been extensively investigated.[28,29] Recently, Justice et al.[30] investigated the effect of smoking on genetic susceptibility to obesity in a large consortium meta-analysis of 241,258 individuals and reported two common variants reaching genome-wide significance, rs12902602 and rs336396 near the *CHRNB4* and *INPP4B* genes respectively. Here we evaluated whether there are modification effects of smoking status on genetic risk from these two loci for obesity with rare or less frequency variants.

We analyze genotype data from the Illumina V1.0 Exome Chip and select variants with MAF less than 5%. We adjust for age, sex, cohort, first two principal components, and smoking main effect in our model. Our data consists of 596 individuals from the Original Cohort (Exam 21), 2547 from the Offspring Cohort (Exam 8), 3868 from the Gen 3 Cohort (Exam 1), and 177 from the Omni Cohort (Exam 1). There are 3264 males (45.4% males), 6063 non-smokers (84.4% non-smokers), and their age ranges from 19 to 85 (median = 49). We test for gene x smoking interaction by treating BMI as continuous trait or dichotomizing BMI at 25 into a binary trait, which classified 4502 individuals as overweight and 2686 individuals in the normal range. A total of 7188 individuals are included in the gene x smoking interaction analysis.

We consider two genes, Cholinergic Receptor Nicotinic Beta 4 (*CHRNB4*) and Inositol Polyphosphate-4-Phosphatase (*INPP4B*). **Table 2** summarizes the analysis results for the two aforementioned genes. We include variants that are annotated as either stop-gain/loss, splice, or missense. Using famGE, we find *CHRNB4* to be statistically significant at $\alpha = 0.025$ for both continuous (p-value = 0.0063) and dichotomized (p-value = 0.0063) BMI, but no gene x smoking interaction were identified for gene *INPP4B.*

5. Discussion

In this article, we develop a unified method called famGE to detect GE interactions of a set of rare variants under the GLMM framework. This proposed approach can accommodate both binary and continuous traits in family data or samples with cryptic relatedness. Additionally, famGE allows weighting variants differently based on prior information such as allele frequency. Under this model, we treat the genetic variants, familial correlation, and GE interaction effects as random effects and implement a variance component score test, which only requires fitting the null model, and thus, reduces computational burden. Our simulation studies show that famGE maintains correct type 1 error and high power under various scenarios. An R package famGE is available upon request.

Ignoring familial correlation when using linear or logistic regression to analyze family data lead to inflated type I error.[23] One way to resolve this issue is to select a subset of unrelated individuals, but this may substantially reduce the sample size and lead to power loss. In the case of rare variants, especially for GE interaction test, it is important to retain as many samples as possible to have appropriate power. A preferable method is to use GLMMs that account for familial correlation as a random effect and thus, eliminates the need to restrict to unrelated individuals. In famGE, kinship coefficients can be obtained either from a pedigree or an empirical kinship matrix. Because the empirical kinship matrix can account for cryptic relatedness, it is advantageous to use the empirical kinship to estimate the level of relatedness among individuals.[14]

The proposed famGE method models the main genotypes as random effects in order to reduce the number of parameters that need to be estimated. If one wishes to model the main effect as fixed effects, the derivations of the new test statistic follows the same framework as the statistic for famGE with random genotype main effects. When the number of variants included in the model is large, however, there is a potential for multicollinearity when fitting the model. It is also shown that modeling the genetic main effects as fixed leads to slightly inflated type 1 error rate at less stringent α levels when the number of variants included in the testing region is large.[18] Therefore, modelling genetic main effects as random effects is preferable.

From the two genes we test in the real data application, *CHRNB4* shows statistical significance in interacting with smoking on BMI. The *CHRNB4* gene has been reported to be associated with higher BMI in never smokers and lower BMI in current smokers, implying that genetics variants may influence BMI via the weight-reducing effects of smoking in opposite directions.[31] *INPP4B* gene is a novel locus identified in the meta-analysis by Justice et al.[30] but to our knowledge, this finding has not yet been replicated in other studies. We cannot exclude the possibility that there are no rare variants interacting with smoking in *INPP4B* gene.[30] However, it is also possible that reduced power due to limited sample size in our study compared to that of in the meta-analysis restrict us from finding a significant association.

With advanced development and decreasing cost of sequencing technology, rare and low frequency variants have become more easily accessible, and there is growing evidence that they are implicated in complex diseases. Therefore, more attention has been brought to analyzing and developing rare variants methods. GE interaction may explain part of the missing heritability, provide insight into etiology of disease, identify subgroups in the population that are at high risk, and help develop personalized treatments. Our proposed approach provides a powerful, robust and efficient test to identify GE interaction of rare variants for both continuous and binary traits in the sample with related or independent participants.

Acknowledgements

This work was in part supported by NIH grants: T32 GM 074905 (EL), R00 HL 130593 (HC) and U01 HL 120393 (HC), U01 DK078616 (JD and CTL), R01 HL120393 (CTL), R01 DK089256 (EL and CTL), and R01 HL118305 (EL and CTL).

Conflict of Interest

The authors declare no conflict of interest.

Appendix 1: Derivation of famGE

We used a generalized linear mixed model (GLMM) framework to derive the famGE statistic. We consider the following gene-environment (GE) model:

$$g(\boldsymbol{E}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{b})) = g(\boldsymbol{\mu}) = \boldsymbol{X\alpha} + \boldsymbol{b_1} + \boldsymbol{b_2} + \boldsymbol{b_3} = \boldsymbol{X\alpha} + \boldsymbol{Zb},$$

where $\boldsymbol{b_1} = \mathbf{GW_1\theta}$, $\boldsymbol{b_2} = \mathbf{EGW_2\gamma}$, and $\boldsymbol{b_3} = \mathbf{d}$ and **Z = [$\boldsymbol{I_n}, \boldsymbol{I_n}, \boldsymbol{I_n}$],**

$\boldsymbol{b} = (\boldsymbol{b_1^T}, \boldsymbol{b_2^T}, \boldsymbol{b_3^T})^{\boldsymbol{T}} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{B}(\boldsymbol{\sigma^2}))$**,** where

$\boldsymbol{B}(\boldsymbol{\sigma^2}) = diag\{\sigma_M^2(\boldsymbol{GW_1W_1G^T}), \sigma_I^2(\boldsymbol{EGW_2W_2G^TE}), \sigma_G^2\boldsymbol{\phi}\}$ depends on an unknown vector of variance components $\boldsymbol{\sigma^2}$.

For subject *i*, the quasi-likelihood given random effects $\boldsymbol{b}$ is defined by:

$$ql_i(\boldsymbol{\alpha}; \boldsymbol{b}) = \int_{y_i}^{\mu_i} \frac{(y_i-\mu)}{\phi v(\mu)} d\mu.$$

The integrated quasi-likelihood function used to estimate $(\boldsymbol{\alpha}, \boldsymbol{\sigma}^{\mathbf{2}})$ is written as

$$L(\alpha,\sigma^2) = e^{ql(\alpha,\sigma^2)} \propto |\boldsymbol{B}|^{-\frac{1}{2}} \int \exp\left\{\sum_{i=1}^{n} ql_i(\boldsymbol{\alpha},\boldsymbol{b}) - \frac{1}{2}\boldsymbol{b}^{\boldsymbol{T}}\boldsymbol{B}^{-\mathbf{1}}\boldsymbol{b}\right\} d\boldsymbol{b}. \qquad \text{(A1)}$$

After applying the Laplacian transformation for integral approximation, the log of A1 becomes

$$ql(\alpha,\sigma^2) = -\frac{1}{2} log|\boldsymbol{B}| - \frac{1}{2} log\left|f''\left(\tilde{\boldsymbol{b}}\right)\right| - f'\left(\tilde{\boldsymbol{b}}\right), \qquad \text{(A2)}$$

where $\tilde{\boldsymbol{b}}$ is the solution to

$$f'\left(\tilde{\boldsymbol{b}}\right) = \frac{\partial ql_i}{\partial \boldsymbol{b}} = -\frac{(y_i-\mu_i)\boldsymbol{z_i}}{\phi v(\mu_i) g\prime(\mu_i)} + \boldsymbol{B}^{-1}\boldsymbol{b} = 0. \qquad \text{(A3)}$$

For canonical link functions, the second partial derivative with respect to b is equal to

$$f''\left(\tilde{\boldsymbol{b}}\right) = \frac{\partial^2 ql_i}{\partial \boldsymbol{b} \partial \boldsymbol{b}^{\boldsymbol{T}}} = \frac{\boldsymbol{z_i z_i^T}}{\phi v(\mu_i)[g\prime(\mu_i)]^2} + \boldsymbol{B}^{-\mathbf{1}} \approx \boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{D}\boldsymbol{Z} + \boldsymbol{B}^{-\mathbf{1}}, \qquad \text{(A4)}$$

where $= diag\left\{\frac{1}{\phi v(\mu_i)[g\prime(\mu_i)]^2}\right\}$.

Combining equations (A3) – (A4), equation (A2) becomes

$$ql(\alpha,\sigma^2) = -\frac{1}{2}log|\boldsymbol{I} + \boldsymbol{Z^T DZB}| - \sum_{i=1}^{n} ql_i(\boldsymbol{\alpha},\widetilde{\boldsymbol{b}}) - \frac{1}{2}\widetilde{\boldsymbol{b}}^{\boldsymbol{T}}\boldsymbol{B}^{-1}\widetilde{\boldsymbol{b}}, \qquad \text{(A5)}$$

where $\widetilde{\boldsymbol{b}} = \widetilde{\boldsymbol{b}}(\boldsymbol{\alpha},\boldsymbol{\sigma^2})$ is the solution to

$$\frac{\partial}{\partial \boldsymbol{b}}\left\{\sum_{i=1}^{n} ql_i(\boldsymbol{\alpha},\widetilde{\boldsymbol{b}}) - \frac{1}{2}\widetilde{\boldsymbol{b}}^{\boldsymbol{T}}\boldsymbol{B}^{-1}\widetilde{\boldsymbol{b}}\right\} = 0.$$

Defining $\boldsymbol{\Delta} = \text{diag}\{g'(\mu_i)\}$ and assuming that the weight matrix $\boldsymbol{D}$ vary slowly as a function of the mean [22], we maximize by differentiating with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{b}$:

$$\frac{\partial ql}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)\boldsymbol{X}_i^{\boldsymbol{T}}}{\phi v(\mu_i)g'(\mu_i)} = \boldsymbol{X^T D\Delta}(\boldsymbol{y} - \boldsymbol{\mu}) \ , \qquad \text{(A6)}$$

$$\frac{\partial ql}{\partial \boldsymbol{b}} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)\boldsymbol{Z}_i^{\boldsymbol{T}}}{\phi v(\mu_i)g'(\mu_i)} - \boldsymbol{B}^{-1}\boldsymbol{b} = \boldsymbol{Z^T D\Delta}(\boldsymbol{y} - \boldsymbol{\mu}) - \boldsymbol{B}^{-1}\boldsymbol{b} \ . \qquad \text{(A7)}$$

Defining the working vector $\boldsymbol{Y_0} = \boldsymbol{X\alpha} + \boldsymbol{Zb} + \boldsymbol{\Delta}(\boldsymbol{y} - \boldsymbol{\mu})$, solutions to A6 and A7 can be expressed as an iterative solution to the system

$$\begin{bmatrix} \boldsymbol{X^T DX} & \boldsymbol{X^T DZ} \\ \boldsymbol{Z^T DX} & (\boldsymbol{Z^T DZ} + \boldsymbol{B}^{-1}) \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X^T DY_0} \\ \boldsymbol{Z^T DY_0} \end{bmatrix}.$$

The test statistic for H$_0$: $\sigma_I^2 = 0$ is

$$\boldsymbol{Q} = \left(\widehat{\boldsymbol{Y}}_{\boldsymbol{0}} - \boldsymbol{X}\widehat{\boldsymbol{\alpha}}\right)^{\boldsymbol{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{EGW_2W_2G^T E}\widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{Y}}_{\boldsymbol{0}} - \boldsymbol{X}\widehat{\boldsymbol{\alpha}})$$

where $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{D}}^{-\boldsymbol{1}} + \hat{\sigma}_M^2(\boldsymbol{GW_1W_1G^T}) + \hat{\sigma}_G^2\boldsymbol{\phi}.$

References

1. Riancho J.A. Genome-wide association studies (GWAS) in complex diseases: advantages and limitations. *Reumatol. Clin.* 2012; 8:56–57.
2. He M, Xu M, Zhang B, *et al*. Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. *Hum. Mol. Genet.* 2015; 24:1791–1800.
3. Ionita-Laza I, Makarov V, Yoon S, *et al*. Finding disease variants in mendelian disorders by using sequence data: Methods and applications. *Am. J. Hum. Genet.* 2011; 89:701–712.
4. Lee S, Abecasis G.R., Boehnke M, Lin X. Rare-variant association analysis: Study designs and statistical tests. *Am. J. Hum. Genet.* 2014; 95:5–23.
5. Kao P.Y., Leung K.H., Chan L.W., Yip S.P., Yap M.K. Pathway analysis of complex diseases for GWAS, extending to consider rare variants, multi-omics and interactions. *Biochim. Biophys.* 2017; 1861:335–353.
6. Yu C, Arcos-Burgos M, Baune B.T., *et al.* Low-frequency and rare variants may contribute to elucidate the genetics of major depressive disorder. *Transl. Psychiatry*. 2018; 8:70.
7. Igartua C, Mozaffari S.V., Nicolae D.L., Ober C. Rare non-coding variants are associated with plasma lipid traits in a founder population. *Sci. Rep.* 2017; 7:1–13.
8. He K.Y., Wang H, Cade B.E., *et al*. Rare variants in fox-1 homolog A (RBFOX1) are associated with lower blood pressure. *PLoS Genet.* 2017; 13:1–15.
9. Li B, Leal S. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* 2018; 83:311–321.
10. Madsen B.E., Browning S.R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 2009; 5:e1000384.
11. Price A.L., Kryukov G.V., de Bakker P.I., *et al.* Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. *Am. J. Hum. Genet.* 2010; 86:832–838.
12. Lin X, Lin X, Lee S, Christiani D.C. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics*. 2013; 14:667–681.
13. Lin W.Y., Lou X.Y., Gao G., Liu N. Rare variant association testing by adaptive combination of p-values. *PLoS One*. 2014; 9:e85728.
14. Wu M.C, 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; 89:82–93.
15. Lee S, Emond M.J., Bamshad M.J., *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* 2012; 91:224–237.
16. Zhang Y, Lin S, Biswas S. Detecting rare and common haplotype–environment interaction under uncertainty of gene–environment independence assumption. *Biometrics.* 2017; 73:344–355
17. Zhao G, Marceau R, Zhang D, Tzeng J.Y. Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression. *Genetics*. 2015; 199:695–710.
18. Chen H, Meigs J.B., Dupuis J. Incorporating gene-environment interaction in testing for

association with rare genetic variants. *Hum. Hered.* 2014; 78:81–90.

19. Tzeng J.Y., Zhang D, Pongpanich M, *et al.* Studying gene and gene-environment effects of uncommon and common variants on continuous traits: A marker-set approach using gene-trait similarity regression. *Am. J. Hum. Genet.* 2011; 89:277–288.
20. Lin X, Lee S, Wu M.C., *et al.* Test for Rare Variants by Environment Interactions in Sequencing Association Studies. *Biometrics.* 2016; 72:156–164.
21. Mazo Lopera M, Coombes B, de Andrade M. An Efficient Test for Gene-Environment Interaction in Generalized Linear Mixed Models with Family Data. *Int. J. Environ. Res.* Public Health. 2017; 14*:*1134.
22. Breslow N.E., Clayton D.G. Approximate Inference in Generalized Linear Mixed Models. *J. Am. Stat. Assoc.* 1993; 88:9–25.
23. Chen H, Meigs J.B., Dupuis J. Sequence Kernel Association Test for Quantitative Traits in Family Samples. *Genet. Epidemiol.* 2013; 37:196–204.
24. Zhang D, Lin X. Hypothesis testing in semiparametric additive mixed models. *Biostatistics.* 2003; 4:57–74.
25. Chung R, Shih C. SeqSIMLA□: a sequence and phenotype simulation tool for complex disease studies. 2013.
26. Mahmooda S.S., Levy D, Vasan R.S., Wang T.J. The Framingham Heart Study and the epidemiology of cardiovascular diseases: A historical perspective. *Lancet.* 2014; 383: 1933–1945.
27. Tsao C.W., Vasan R.S. Cohort Profile: The Framingham Heart Study (FHS): Overview of milestones in cardiovascular epidemiology. *Int. J. Epidemiol.* 2015; 44:1800–1813.
28. Liao C, Gao W, Cao W, *et al*. The association of cigarette smoking and alcohol drinking with body mass index: A cross-sectional, population-based study among Chinese adult male twins. *BMC Public Health.* 2016; 16:1–9.
29. Rask-Andersen M, Karlsson T, Ek W.E., Johansson Å. Gene-environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status. *PLoS Genet.* 2017; 13:1–20.
30. Justice A.E., Winkler T.W., Feitosa M.F., *et al*. Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. *Nat. Commun.* 2017; 8:1–19.
31. Taylor A.E., Morris R.W., Fluharty M.E., *et al.* Stratification by Smoking Status Reveals an Association of CHRNA5-A3-B4 Genotype with Body Mass Index in Never Smokers. *PLoS Genet.* 2014; 10:1–6.

Tables

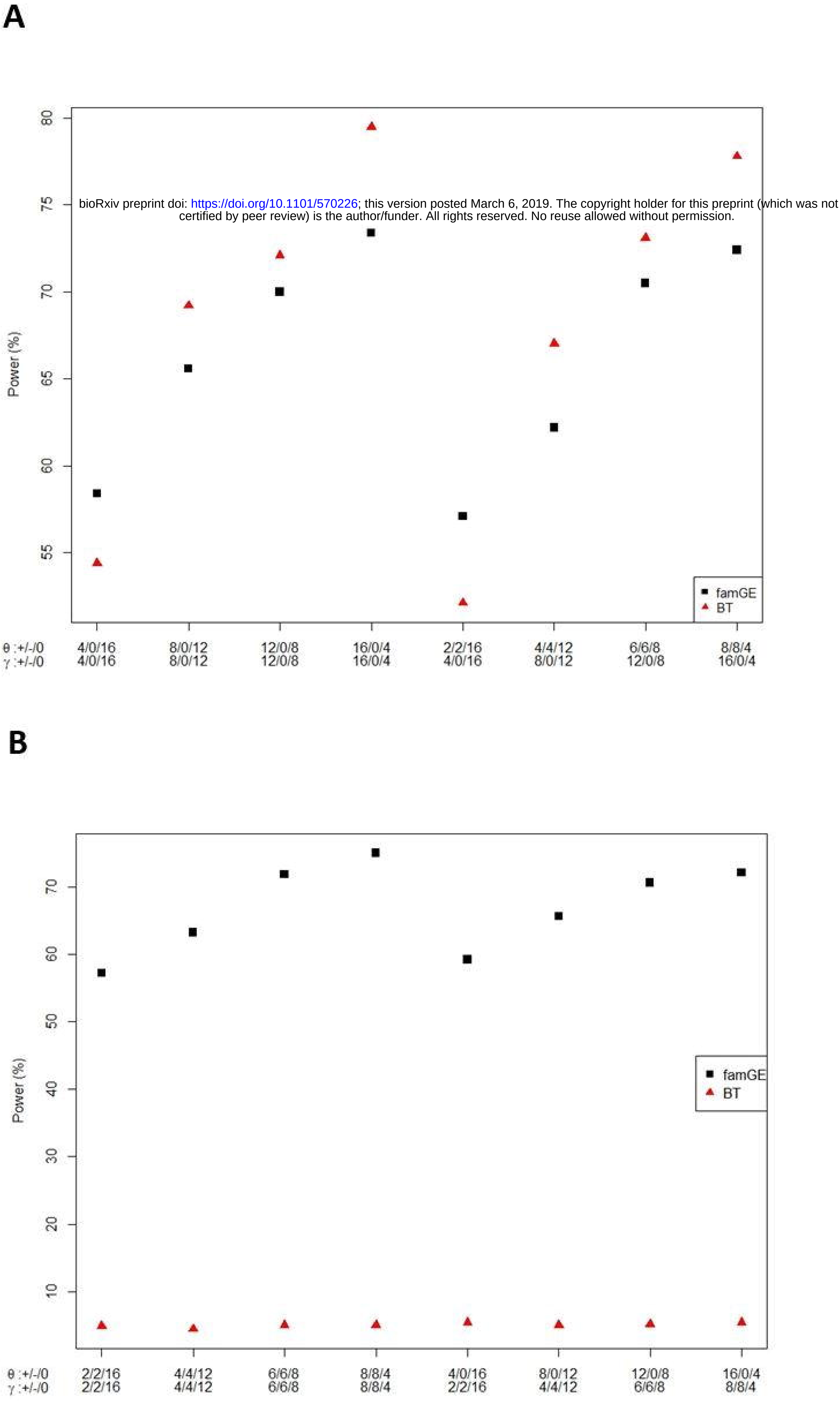
Table 1. Comparison of type 1 errors based on 20,000 replications

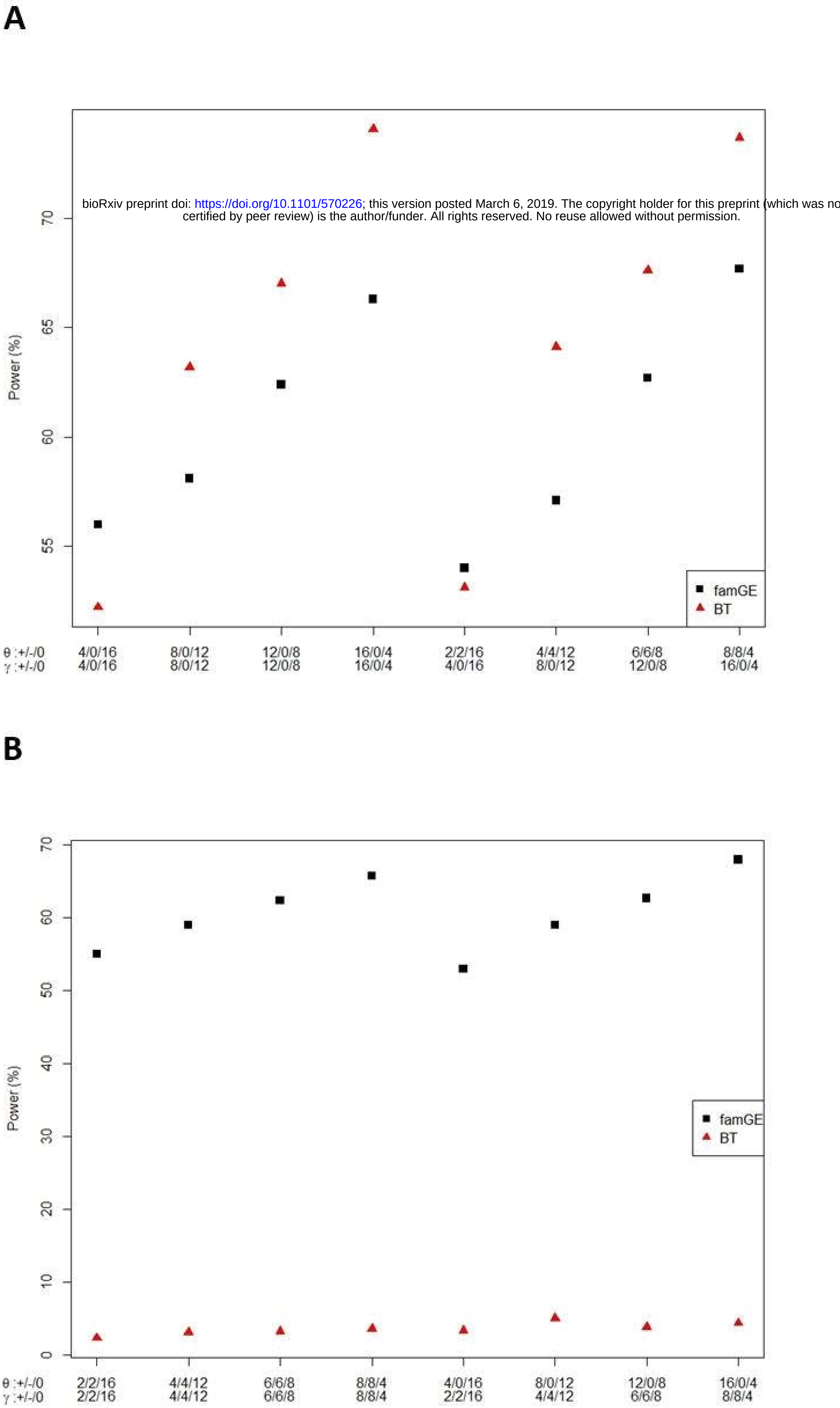
		Continuous			Binary		
Scenarios (+/-/0)	α level	famGE[a]	rareGE[b]	BT[c]	famGE[a]	rareGE[b]	BT[c]
4/0/16	0.05	0.0496	0.0530	0.0501	0.0503	0.0551	0.0507
	0.01	0.0100	0.0110	0.0102	0.0102	0.0125	0.0105
	0.001	0.0010	0.0170	0.0010	0.0012	0.0018	0.0010
8/0/12	0.05	0.0488	0.0521	0.0499	0.0484	0.0557	0.0492
	0.01	0.0100	0.0103	0.0100	0.0097	0.0134	0.0095
	0.001	0.0010	0.0011	0.0010	0.0011	0.0019	0.0009
12/0/8	0.05	0.0495	0.0520	0.0499	0.0482	0.0554	0.0498
	0.01	0.0094	0.0120	0.0096	0.0106	0.0122	0.0090
	0.001	0.0010	0.0015	0.0010	0.0008	0.0015	0.0010
16/0/4	0.05	0.0504	0.0538	0.0500	0.0484	0.0570	0.0500
	0.01	0.0091	0.0116	0.0099	0.0097	0.0141	0.0094
	0.001	0.0010	0.0016	0.0010	0.0010	0.0020	0.0010
2/2/16	0.05	0.0480	0.0525	0.0500	0.0497	0.0556	0.0506
	0.01	0.0095	0.0120	0.0101	0.0094	0.0116	0.0097
	0.001	0.0010	0.0014	0.0011	0.0011	0.0016	0.0011
4/4/12	0.05	0.0499	0.0533	0.0502	0.0484	0.0552	0.0489
	0.01	0.0093	0.0105	0.0100	0.0097	0.0116	0.0096
	0.001	0.0010	0.0019	0.0010	0.0010	0.0014	0.0009
6/6/8	0.05	0.0493	0.0541	0.0494	0.0488	0.0561	0.0488
	0.01	0.0103	0.0110	0.0104	0.0092	0.0137	0.0100
	0.001	0.0010	0.0014	0.0010	0.0011	0.0017	0.0011
8/8/4	0.05	0.0502	0.0531	0.0493	0.0489	0.0555	0.0479
	0.01	0.0090	0.0111	0.0099	0.0097	0.0127	0.0094
	0.001	0.0010	0.0017	0.0012	0.0010	0.0016	0.0010

+/-/0: number of variants with main genotype effect sizes that are positive, negative, and neutral
famGE[a]: our proposed method accounting for family correlation
rareGE[b]: interaction test ignoring family correlation
BT[c]: burden test

Table 2. Association results of gene by smoking interaction on continuous trait: BMI and binary trait: overweight status (BMI ≥ 25) for two previously identified loci using the Framingham Heart Study

Gene	Chromosome	# of variants used	famGE p-value
Continuous: BMI			
CHRNB4	15	8	0.0063
INPP4B	4	5	0.5504
Binary: BMI ≥ 25			
CHRNB4	15	8	0.0023
INPP4B	4	5	0.9166


A
Power (%)
80
75
70
65
60
55
famGE
BT
θ :+/-/0
γ :+/-/0
4/0/16
4/0/16
8/0/12
8/0/12
12/0/8
12/0/8
16/0/4
16/0/4
2/2/16
4/0/16
4/4/12
8/0/12
6/6/8
12/0/8
8/8/4
16/0/4
B
Power (%)
70
60
50
40
30
20
10
famGE
BT
θ :+/-/0
γ :+/-/0
2/2/16
2/2/16
4/4/12
4/4/12
6/6/8
6/6/8
8/8/4
8/8/4
4/0/16
2/2/16
8/0/12
4/4/12
12/0/8
6/6/8
16/0/4
8/8/4


A
Power (%)
55
60
65
70
famGE
BT
θ :+/-/0
γ :+/-/0
4/0/16 4/0/16
8/0/12 8/0/12
12/0/8 12/0/8
16/0/4 16/0/4
2/2/16 4/0/16
4/4/12 8/0/12
6/6/8 12/0/8
8/8/4 16/0/4
B
Power (%)
0
10
20
30
40
50
60
70
famGE
BT
θ :+/-/0
γ :+/-/0
2/2/16 2/2/16
4/4/12 4/4/12
6/6/8 6/6/8
8/8/4 8/8/4
4/0/16 2/2/16
8/0/12 4/4/12
12/0/8 6/6/8
16/0/4 8/8/4