Efficient gene-environment interaction tests for large biobank-scale sequencing studies

Xinyu Wang[1], Elise Lim[2], Ching-Ti Liu[2], Yun Ju Sung[3], Dabeeru C. Rao[3], Alanna C. Morrison[4],

Eric Boerwinkle[4,5], Alisa K. Manning[6,7], Han Chen[4,8,*]

[1]*Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, United States of America*
[2]*Department of Biostatistics, Boston University, Boston, MA 02118, United States of America*
[3]*Division of Biostatistics, Washington University School of Medicine, St. Louis, MO, United States of America*
[4]*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, TX, United States of America*
[5]*Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, United States of America*
[6]*Center for Human Genetics Research, Massachusetts General Hospital, Boston, MA, United States of America*
[7]*Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, United States of America*
[8]*Center for Precision Health, School of Public Health and School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, United States of America*

*Corresponding author:

Email: Han.Chen.2@uth.tmc.edu (Han Chen)

ABSTRACT

Complex human diseases are affected by genetic and environmental risk factors and their interactions. Gene-environment interaction (GEI) tests for aggregate genetic variant sets have been developed in recent years. However, existing statistical methods become rate limiting for large biobank-scale sequencing studies with correlated samples. We propose efficient Mixed-model Association tests for GEne-Environment interactions (MAGEE), for testing GEI between an aggregate variant set and environmental exposures on quantitative and binary traits in large-scale sequencing studies with related individuals. Joint tests for the aggregate genetic main effects and GEI effects are also developed. A null generalized linear mixed model adjusting for covariates but without any genetic effects is fit only once in a whole genome GEI analysis, thereby vastly reducing the overall computational burden. Score tests for variant sets are performed as a combination of genetic burden and variance component tests by accounting for the genetic main effects using matrix projections. The computational complexity is dramatically reduced in a whole genome GEI analysis, which makes MAGEE scalable to hundreds of thousands of individuals. We applied MAGEE to the exome sequencing data of 41,144 related individuals from the UK Biobank, and the analysis of 18,970 protein coding genes finished within 10.4 CPU hours.

KEYWORDS: gene-environment interaction, generalized linear mixed model, correlated data, rare variants, joint test

1. INTRODUCTION

The variation of traits among individuals in a population results from genetic and environmental factors, as well as their interactions. Gene-environment interaction (GEI) studies can improve our understanding on the biological mechanisms of complex diseases and lead to discoveries of novel genetic associations with effects that vary with environmental exposures (Mcallister et al., 2017). The advancements in next-generation sequencing technologies over the past decade have enabled the increasing availability of large-scale data of low-frequency and rare genetic variants (with minor allele frequency (MAF) 1–5% and <1%, respectively). The single-variant tests conventionally used for testing GEI with common variants are underpowered for rare variants. For example, statistical power for testing GEI with a binary environmental exposure depends on the minor allele counts in both exposed and unexposed groups. Although computationally efficient GEI tests for biobank-scale studies have been developed recently in the context of single-variant tests on unrelated individuals (Bi et al., 2019) , critical methodological bottlenecks still exist to expand the sample size and scope of rare variant GEI analyses in large biobank-scale sequencing studies. To increase power for rare variants, various set-based methods have been developed to collapse variants in a particular gene or functional region to investigate how variants in a set affect a phenotype synergistically (Chen et al., 2019; Lee, Wu, & Lin, 2012; Pan, Kim, Zhang, Shen, & Wei, 2014; Sun, J., Zheng, & Hsu, 2013), and to demonstrate whether genetic associations with the phenotype are modified by environment factors in GEI studies (Chen, Meigs, & Dupuis, 2014; Lin et al., 2016; Su, Y., Di, & Hsu, 2017). For example, rareGE is a software tool for GEI tests on rare variants (Chen et al., 2014) that implements three variance component tests: two GEI tests that treat the genetic main effects either as fixed or random effects, and a joint test for the genetic main effects and GEI. (Lin et al., 2016) proposed the interaction sequence kernel association test

3

(iSKAT), which extends the SKAT-O test (Lee et al., 2012) for genetic main effects to a GEI test for rare variants that estimates the genetic main effects using a ridge regression model (Hastie, Tibshirani, & Friedman, 2009). Mixed effects Score Tests for interaction (MiSTi) proposed by (Su et al., 2017) is an extension to MiST (Sun et al., 2013) in the context of GEI.

Most of the aforementioned variant set-based GEI tests were developed to analyze unrelated samples and directly applying these methods to related samples will lead to invalid statistical inference and inflated type I error rates. On the other hand, extending these methods by adding a random effects term to account for relatedness in the generalized linear mixed model (GLMM) framework would result in an intensive computational complexity of $O(N^3)$ for each variant set, where N is the sample size (Lim, Chen, Dupuis, & Liu, 2020; Mazo Lopera, Coombes, & de Andrade, 2017). For large biobank-scale sequencing studies, such as the UK Biobank, related samples are often present. Therefore, there is an urgent need for computationally efficient variant set-based GEI tests that are scalable to hundreds of thousands of related samples.

Joint tests for genetic main effects and GEI effects are used to identify novel associations previously missed in genetic main effect tests, by accounting for heterogeneous genetic effects in samples with different environmental exposures (Cornelis et al., 2012). Joint tests for common variants have been developed (Chen et al., 2014; Kraft, Yen, Stram, Morrison, & Gauderman, 2007; Selinger-Leneman, Genin, Norris, & Khlat, 2003; Sun, R., Carroll, Christiani, & Lin, 2018), and a variant set-based joint test has been implemented in rareGE (Chen et al., 2014). However, the rareGE joint test relies on a Monte Carlo method to compute the p values, which changes slightly with the random number seed used in the analysis, and an analytical solution is not currently available.

We propose a computationally efficient method, Mixed-model Association test for GEne-Environment interactions (MAGEE), to test GEI effects for rare variants that can reduce the computational complexity for testing each variant set from $O(N^3)$ to at most $O(N^2)$ for related samples, where N is the sample size. For samples with well-defined family structures the use of a block diagonal correlation structure can greatly reduce the complexity to $O(nN)$, where n is the maximum number of individuals in each block. For unrelated samples, the computational complexity for testing each variant set in MAGEE is $O(N)$, with statistical power being close to the existing methods such as rareGE and MiSTi. We also propose analytical joint tests in MAGEE that do not require Monte Carlo approaches for p value calculations.

2. MATERIALS AND METHODS

2.1 Generalized linear mixed models (GLMMs) and main effect tests

We developed two GEI tests and three joint tests based on GLMMs within the MAGEE framework. The full model of MAGEE is:

$$g(\mu_i) = \boldsymbol{X_i\alpha} + \boldsymbol{G_i\beta} + \boldsymbol{K_i\gamma} + r_i \quad (1),$$

where $g(\cdot)$ is the link function of μ_i, and μ_i is the conditional mean of the phenotype for individual i given covariates $\boldsymbol{X_i}$, genotypes $\boldsymbol{G_i}$ and a random intercept r_i. $\boldsymbol{X_i}$ is a row vector of p covariates including an intercept, $\boldsymbol{G_i}$ is a row vector of q variants, and $\boldsymbol{K_i}$ is a row vector of cq pairwise GEI terms for c environmental factors (which are a subset of the p covariates in $\boldsymbol{X_i}$) and q variants. Accordingly, $\boldsymbol{\alpha}$ is a $p \times 1$ vector for the covariate effects, $\boldsymbol{\beta}$ is a $q \times 1$ vector for the genetic main effects, and $\boldsymbol{\gamma}$ is the $cq \times 1$ vector for GEI effects. The length N vector for the random intercept $\boldsymbol{r} = (r_1 \, r_2 \cdots r_N)^T \sim N(0, \sum_{l=1}^{L} \lambda_l \boldsymbol{\psi_l})$, where λ_l are the variance component parameters for L random effects, and $\boldsymbol{\psi_l}$ are $N \times N$ known relatedness matrices.

The genetic main effect model assuming no GEI ($\boldsymbol{\gamma = 0}$) is

5

$$g(\mu_i) = X_i\alpha + G_i\beta + r_i \quad (2)$$

Tests for $H_0: \beta = 0$ can be constructed in the SMMAT framework (Chen et al., 2019). If β are treated as random effects, the main effect variance component (MV) test is then SKAT (Wu et al., 2011) for related samples (Chen, Meigs, & Dupuis, 2013), from which we can acquire a p value p_{MV}. Another main effect test, the main effect hybrid test using Fisher's method (Fisher, 1928) (MF), which combines the burden test and SKAT, is the efficient hybrid test SMMAT-E (Chen et al., 2019). In this test, we get two p values, one from the burden test, p_B, and the other from the adjusted SKAT test, p_{AS}. SMMAT-E (Chen et al., 2019) combines these two p values to calculate a p value for the MF test using Fisher's method $p_{MF} = \mathrm{P}(\chi_4^2 > -2logp_B - 2logp_{AS})$.

2.2 Interaction tests

In model (2), the score vector for γ is $\tilde{S}_K = K^T(y - \tilde{\mu}_G)/\tilde{\phi}$, where y is an $N \times 1$ vector of the observed phenotype values, $\tilde{\phi}$ is the dispersion parameter estimate, $\tilde{\mu}_G$ is a vector of fitted values, and $K = (K_1^T K_2^T ... K_N^T)^T$ is the $N \times cq$ matrix for the GEI terms. Generally, testing for GEI $H_0: \gamma = 0$ requires adjusting for genetic main effects. Therefore, we need to refit model (2) for every set of genetic variants in a whole genome GEI analysis. To reduce the computational burden, we first fit a global null model without any genetic main effects:

$$g(\mu_i) = X_i\alpha + r_i \quad (3)$$

From this model, we can construct score vectors $S_G = G^T(y - \hat{\mu}_0)/\hat{\phi}$ and $S_K = K^T(y - \hat{\mu}_0)/\hat{\phi}$ for genetic main effects and GEI effects, respectively, where $\hat{\mu}_0$ is a vector of fitted values from model (3), and $\hat{\phi}$ is the dispersion parameter estimate from model (3). Assuming the main effect of genetic variants β are small, we then approximate the score vector for GEI effects by $\tilde{S}_K \approx S_K - K^T\hat{P}G(G^T\hat{P}G)^{-1}S_G$ (in appendix A), where $G = (G_1^T G_2^T ... G_N^T)^T$ is a $N \times q$ matrix of genetic variants, $X = (X_1^T X_2^T ... X_N^T)^T$ is a $N \times p$ matrix of covariates, $\hat{P} = \hat{\Sigma}^{-1} -$

$\widehat{\Sigma}^{-1} X (X^T \widehat{\Sigma}^{-1} X)^{-1} X^T \widehat{\Sigma}^{-1}$ is an $N \times N$ projection matrix, where $\widehat{\Sigma} = \widehat{V} + \sum_{l=1}^{L} \hat{\lambda}_l \psi_l$, $\widehat{V} = \hat{\phi} I_N$ for continuous traits and $diag\{\frac{1}{\hat{\mu}_{0i}(1-\hat{\mu}_{0i})}\}$ for binary traits, which we estimate from model (3).

Using this approximation, in the interaction variance component (IV) test, we assume $\gamma \sim N(0, \tau W_K{}^2)$, where W_K is a $cq \times cq$ predefined diagonal weight matrix for GEI. Testing for GEI $H_0: \gamma = 0$ is then equivalent to testing the variance component parameter $H_0: \tau = 0$, with a test statistic

$$T_\gamma = \tilde{S}_K^T W_K W_K \tilde{S}_K.$$

Under the null hypothesis, T_γ asymptotically follows the distribution of $\sum_{j=1}^{cq} \xi_{\gamma,j} \chi_{1,j}^2$, where $\chi_{1,j}^2$ are independent chi-square distributions with 1 degree of freedom (df), and $\xi_{\gamma,j}$ are the eigenvalues of $W_K \Lambda W_K$, where $\Lambda = K^T \widehat{P} K - K^T \widehat{P} G (G^T \widehat{P} G)^{-1} G^T \widehat{P} K$ (in appendix A).

Alternatively, we develop the interaction hybrid test using Fisher's method (IF) to combine a burden-type test and an adjusted variance component test that are asymptotically independent, to achieve superior power than the IV test when the true mean of interaction effects γ is not close to 0. Specifically, in the IF test we assume $\gamma \sim N(W_K \mathbf{1}_{cq} \gamma_0, \ \tau W_K^2)$, where $\mathbf{1}_{cq}$ is a vector of 1's with length cq, and testing for GEI $H_0: \gamma = 0$ is equivalent to testing $H_0: \gamma_0 = \tau = 0$ simutaneously. We decompose this test into two tests (Chen et al., 2019; Sun et al., 2013). In the first test, we assume $\tau = 0$ and test $H_0: \gamma_0 = 0$ using the burden score constructed from the global null model (3): $\tilde{S}_{KB} = \mathbf{1}_{cq}^T W_K \tilde{S}_K = \mathbf{1}_{cq}^T W_K \{S_K - K^T \widehat{P} G (G^T \widehat{P} G)^{-1} S_G\}$. The test statistic

$$T_{\gamma_0} = \tilde{S}_{KB}^2$$

follows a distribution of $\xi_{\gamma_0} \chi_1^2$ under the null hypothesis $H_0: \gamma_0 = 0$, where the scalar $\xi_{\gamma_0} = \mathbf{1}_{cq}^T W_K \Lambda W_K \mathbf{1}_{cq} = \mathbf{1}_{cq}^T W_K \left\{ K^T \widehat{P} K - K^T \widehat{P} G (G^T \widehat{P} G)^{-1} G^T \widehat{P} K \right\} W_K \mathbf{1}_{cq}.$

7

In the second test, we do not make an assumption that $\gamma_0 = 0$ but we assume its true value is small and we test $H_0: \tau = 0$ using adjusted scores accounting for the burden effect $\tilde{S}_{KV} \approx \tilde{S}_K - \Lambda W_K 1_{cq}(1_{cq}^T W_K \Lambda W_K 1_{cq})^{-1}\tilde{S}_{KB}$ (in appendix B). Then the test statistic can be constructed as:

$$T_\tau = \tilde{S}_{KV}^T W_K W_K \tilde{S}_{KV}$$

and it asymptotically follows a distribution of $\sum_{j=1}^{cq} \xi_{\tau,j} \chi_{1,j}^2$ under the null hypothesis $H_0: \tau = 0$, where $\xi_{\tau,j}$ are eigenvalues for $W_K \Lambda_{KV} W_K$, and $\Lambda_{KV} = \Lambda - \Lambda W_K 1_{cq}(1_{cq}^T W_K \Lambda W_K 1_{cq})^{-1} 1_{cq}^T W_K \Lambda$.

As $\tilde{S}_{KB}$ and $\tilde{S}_{KV}$ are asymptotically independent (in appendix B), we combine p values from the two tests T_{γ_0} and T_τ using Fisher's method to compute $p_{IF} = P(\chi_4^2 > -2logp_{\gamma_0} - 2logp_\tau)$, where p_{γ_0} and p_τ are p values from the burden-type test $H_0: \gamma_0 = 0$ (under the assumption $\tau = 0$) and the adjusted variance component test $H_0: \tau = 0$, respectively.

2.3 Joint tests

As the score vector S_G for genetic main effects and the adjusted score vector $\tilde{S}_K$ for GEI effects are asymptotically normal with covariance

$$Cov(\tilde{S}_K, S_G) \approx Cov(S_K - K^T \hat{P} G(G^T \hat{P} G)^{-1} S_G, \quad S_G)$$

$$= K^T \hat{P} G - K^T \hat{P} G(G^T \hat{P} G)^{-1} G^T \hat{P} G = 0,$$

they are asymptotically independent and so are the main effect tests and interaction tests derived using S_G and $\tilde{S}_K$, respectively. Therefore, an analytical form of the joint variance component (JV) test for genetic main effects and GEI effects can be constructed by combining p values of MV (p_{MV}) and IV (p_{IV}) tests using Fisher's method as $p_{JV} = P(\chi_4^2 > -2logp_{MV} - 2logp_{IV})$, similar to the IF test above.

8

We can also construct the joint test using MF and IF since they are also derived using $\boldsymbol{S}_G$ and $\tilde{\boldsymbol{S}}_K$ respectively, and both $\tilde{\boldsymbol{S}}_{KB}$ and $\tilde{\boldsymbol{S}}_{KV}$ are linear functions of $\tilde{\boldsymbol{S}}_K$ that are asymptotically independent. The four p values − p_{γ_0} from the GEI burden-type test, p_τ from the GEI adjusted variance component test, as well as the aforementioned components in the genetic main effects MF test, p_B and p_{AS} − are asymptotically mutually independent. Therefore, the joint hybrid test using Fisher's method (JF) can be constructed by adding these four p values on the log scale, which follows a chi-square distribution with 8 df under the null hypothesis of no genetic main effects or GEI effects (Fisher, 1928) ($p_{JF} = P(\chi_8^2 > -2logp_B - 2logp_{AS} - 2logp_{\gamma_0} - 2logp_\tau)$). Alternatively, we can apply Fisher's method to combine two p values from MF and IF tests, which follows a chi-square distribution with 4 df under the null hypothesis of no genetic main effects or GEI effects ($p_{JD} = P(\chi_4^2 > -2logp_{MF} - 2logp_{IF})$). This is a joint hybrid test using double Fisher's procedures (JD), since each of MF and IF p values is already computed using Fisher's method (refer to Table S1 of the supporting information for a summary of all five new tests: GEI tests IV and IF, and joint tests JV, JF and JD).

2.4 Simulations

We conducted extensive simulations to 1) investigate MAGEE's type I error control in unrelated and related samples; 2) compare p values from MAGEE and the existing methods rareGE and MiSTi on the scale of 10^{-5} to 10^{-15} in unrelated samples; and 3) compare the power of each test within the MAGEE framework. All the simulation scenarios were performed for both quantitative and binary traits.

2.4.1 Type I error in unrelated samples

HAPGEN2 (Su, Z., Marchini, & Donnelly, 2011) was used to simulate 200,000 haplotypes on chromosome 22 based on 1000 Genomes project CEU data (Sabeti, 2015) as the reference panel.

We randomly paired them into genotypes of 100,000 unrelated individuals. We simulated 40 genotype replicates, and each genotype replicate has a total of 119,317 variants, which were assigned to 1,194 variant sets. The first 1,193 groups had 100 variants per group, while the last group contained 17 variants. The MAF for each variant ranged from 5.0×10^{-6} to 0.5, with a mean of 0.18.

We simulated 1,000 phenotype replicates for each genotype replicate. The quantitative trait for individual i was simulated from

$$y_i = 0.1 Age_i + 0.2 Sex_i + 0.1 BMI_i + \varepsilon_i,$$

where $Age_i \sim N(50, 5)$, $Sex_i \sim Bernoulli(0.5)$, $BMI_i \sim N(25, 4)$, and the random error $\varepsilon_i \sim N(0, 1)$. The binary traits for unrelated sample were simulated as a cohort study from

$$\log\left(\frac{P(y_i=1)}{1-P(y_i=1)}\right) = \alpha_0 + 0.1(Age_i - \overline{Age}) + 0.2(Sex_i - \overline{Sex}) + 0.1(BMI_i - \overline{BMI}),$$

where y_i is the observed phenotype value (either 0 or 1), Age, Sex, and body mass index (BMI) follow the same distribution as for the quantitative traits, and $\overline{Age}$, $\overline{Sex}$, and $\overline{BMI}$ are the mean values for Age, Sex, and BMI, respectively, and α_0 was set to $\log\frac{0.4}{1-0.4}$. For both quantitative and binary traits, we tested for gene-BMI interactions using beta density weight function with parameters 1 and 25 on the MAF (Wu et al., 2011) for both common and rare genetic variants, so that rare variants would have larger weights than common variants.

2.4.2 Type I error in related samples

Since MAGEE can be applied to both unrelated and related samples, we also simulated 40 genotype replicates for 25,000 families with two parents and two children with a theoretical

kinship matrix of $\begin{bmatrix} 0.5 & 0.0 & 0.25 & 0.25 \\ 0.0 & 0.5 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.5 \end{bmatrix}$, totaling 100,000 individuals. The 100,000 haplotypes for

the 50,000 founders were randomly sampled and paired into genotypes, and each child inherited

one random haplotype from each parent within the same family. The total number of variants were also 119,317 and were assigned to 1,194 variant sets.

We simulated 1,000 phenotype replicates for each genotype replicate. The quantitative trait for individual j in family i was simulated from

$$y_{ij} = 0.1Age_{ij} + 0.2Sex_{ij} + 0.1BMI_{ij} + r_{ij} + \varepsilon_{ij},$$

where $Age_{ij} \sim N(50,5)$ for parents and $Age_{ij} \sim N(20,5)$ for children, $Sex_{ij} \sim Bernoulli(0.5)$ for both parents and children, BMI for family i $\boldsymbol{BMI_i} \sim N(\begin{bmatrix} 25 \\ 25 \\ 25 \\ 25 \end{bmatrix}, \begin{bmatrix} 4.0 & 0.0 & 1.5 & 1.5 \\ 0.0 & 4.0 & 1.5 & 1.5 \\ 1.5 & 1.5 & 4.0 & 1.5 \\ 1.5 & 1.5 & 1.5 & 4.0 \end{bmatrix})$, by assuming a

heritability of 0.75(Elks et al., 2012), the random effects for family i $\boldsymbol{r_i} \sim$

$N(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.25 & 0.0 & 0.125 & 0.125 \\ 0.0 & 0.25 & 0.125 & 0.125 \\ 0.125 & 0.125 & 0.25 & 0.125 \\ 0.125 & 0.125 & 0.125 & 0.25 \end{bmatrix})$, and the random error $\varepsilon_{ij} \sim N(0,0.75)$.

The binary traits for related sample were simulated as a cohort study from

$$\log\left(\frac{P(y_{ij}=1)}{1-P(y_{ij}=1)}\right) = \alpha_0 + 0.1(Age_{ij} - \overline{Age}) + 0.2(Sex_{ij} - \overline{Sex}) + 0.1(BMI_{ij} - \overline{BMI}) + r_{ij},$$

where Age, Sex, BMI, and the random effects for family i $\boldsymbol{r_i}$ follow the same distribution as for the quantitative traits, and $\overline{Age}$, $\overline{Sex}$, and $\overline{BMI}$ are the population mean values for them, respectively, α_0 is $\log\frac{0.4}{1-0.4}$. For both quantitative and binary traits, we tested for gene-BMI interactions using beta density weight function with parameters 1 and 25 on the MAF (Wu et al., 2011) for both common and rare genetic variants.

2.4.3 Comparison of p values

We simulated both quantitative and binary traits with unrelated individuals to compare the p value estimations for MAGEE with existing methods rareGE and MiSTi. Quantitative traits were simulated from $y_i = 0.1Age_i + 0.2Sex_i + 0.1BMI_i + \sum_{t=1}^{q}\beta_t(G_{it} - \overline{G_t}) + \sum_{t=1}^{q}\gamma_t(K_{ijt} - \overline{K_t}) + \varepsilon_i$,

11

and binary traits were simulated from $\log\left(\frac{P(y_i=1)}{1-P(y_i=1)}\right) = \alpha_0 + 0.1(Age_i - \overline{Age}) + 0.2(Sex_i - \overline{Sex}) +$

$0.1(BMI_i - \overline{BMI}) + \sum_{t=1}^{q}\beta_t(G_{it} - \overline{G_t}) + \sum_{t=1}^{q}\gamma_t(K_{it} - \overline{K_t})$, where G_{it} denoted the t-th genetic

variant for the i-th individual, and $\overline{G_t}$ was the mean of the t-th variant, $K_{it} = G_{it}(BMI_i - \overline{BMI})$ was

the GEI term for the t-th variant with BMI for individual i, and $\overline{K_t}$ was the mean of the t-th

interaction term. We conducted tests with sample sizes 2,000, 5,000, and 10,000. In scenario 1, we

simulated both genetic main effects β_t and GEI effects γ_t from $clog_{10}(MAF_t)$, with different

values of constant c. In scenario 2, we simulated only genetic main effects β_t from $clog_{10}(MAF_t)$,

and set GEI effects $\gamma_t = 0$ so that it was a null hypothesis simulation for interaction tests but an

alternative hypothesis simulation for joint tests. The specific values for constant c in each scenario

and a summary of the simulation scenarios can be found in Table S2. In all these scenarios, Age_i,

Sex_i, BMI_i and ε_i followed the same distribution as in the null simulations for unrelated samples.

We simulated one genotype replicate with 100 variants for 10,000 unrelated individuals, and 1,000

phenotype replicates in each scenario. For β_t and γ_t, we randomly assigned 80% of them to 0, 10%

to have a positive effect, and 10% to have a negative effect.

2.4.4 Power

To investigate the power of each test within the MAGEE framework, we simulated six scenarios

using related individuals with sample sizes 20,000, 50,000, and 100,000 in each scenario. In

scenarios 1-3, we randomly assigned 80% of genetic main effects β_t and GEI effects γ_t to 0, 10%

of β_t and γ_t to have a positive effect, and the remaining 10% of each of β_t and γ_t to have a

negative effect. These proportions were changed to 80% null, 16% positive, and 4% negative in

scenarios 4-6. In all of the six scenarios, Age_{ij}, Sex_{ij}, BMI_{ij}, r_{ij}, and ε_{ij} followed the same

distributions as the null simulations for related individuals. We simulated one genotype replicate

with 100 variants for 100,000 related individuals with the same family structure as the null

simulations for related individuals, and 1,000 phenotype replicates in each scenario.

Quantitative traits were simulated from $y_{ij} = 0.1 Age_{ij} + 0.2 Sex_{ij} + 0.1 BMI_{ij} + \sum_{t=1}^{q} \beta_t (G_{ijt} -$

$\overline{G_t}) + \sum_{t=1}^{q} \gamma_t (K_{ijt} - \overline{K_t}) + r_{ij} + \varepsilon_{ij}$, and binary traits were simulated from $\log \left(\frac{P(y_{ij}=1)}{1-P(y_{ij}=1)} \right) = \alpha_0 +$

$0.1(Age_{ij} - \overline{Age}) + 0.2(Sex_{ij} - \overline{Sex}) + 0.1(BMI_{ij} - \overline{BMI}) + \sum_{t=1}^{q} \beta_t (G_{ijt} - \overline{G_t}) + \sum_{t=1}^{q} \gamma_t (K_{ijt} -$

$\overline{K_t}) + r_{ij}$, where G_{ijt} was the t-th genetic variant for individual j in family i, and $K_{ijt} = G_{ijt}(BMI_{ij} -$

$\overline{BMI})$ was the GEI term for the t-th variant with BMI for individual j from family i. In scenarios 1

and 4, we simulated both genetic main effects β_t and GEI effects γ_t from $c log_{10}(MAF_t)$, with

different values of constant c. In scenario 2 and 5, we simulated only genetic main effects β_t from

$c log_{10}(MAF_t)$, and set GEI effects $\gamma_t = 0$. In scenario 3 and 6, we set genetic main effects $\beta_t =$

0, and simulated GEI effects γ_t from $c log_{10}(MAF_t)$. The detailed simulation settings and choices

for constant c in each scenario for both quantitative and binary trait are listed in Table S3.

2.5 Application to UK Biobank whole exome sequencing data

We used the first tranche of UK Biobank whole exome sequencing (WES) data released in March

2019 with 49,959 participants, which included 43,386 White British individuals with non-missing

age. We tested for gene-sex interactions on a quantitative trait BMI, as well as a dichotomized trait

obesity, which was defined as BMI ≥ 30 (24.1% of the study samples). After removing samples

with missing genetic ancestry or gender and sex mismatches, the total number of individuals was

41,144, including 18,925 men and 22,219 women. For BMI, we fit a heteroscedastic linear mixed

model (Conomos et al., 2016) that allowed the residual variance to be different in men and women,

adjusting for sex, age, age^2, the interaction terms between sex and age, age^2, and the top ten

ancestry principal components (PCs). For obesity, we adjusted for the same covariates in a GLMM.

We used the population-level genotype data generated from the Functionally Equivalent (FE)

13

pipeline (Regier et al., 2018). The variant sets were defined as protein-coding regions for the WES data and a total of 18,970 protein-coding regions with 9,760,758 variants were available for analysis. Variants with MAF less than 0.001 were excluded from the analysis. We applied a beta density weight function with parameters 1 and 25 on the MAF (Wu et al., 2011), and performed both GEI and joint tests using a single thread on a computing server.

3. RESULTS

3.1 Simulation Studies

3.1.1 Type I error

Table 1 summarizes the empirical type I error rates of MAGEE tests for quantitative and binary traits with 100,000 unrelated sample at significance levels of 0.05, 1.0×10^{-4}, and 2.5×10^{-6}. In each cell, empirical type I error rates were estimated from 47,760,000 p values under the null hypothesis (40 genotype replicates $\times$ 1,000 phenotype replicates $\times$ 1,194 variant sets) of the null model for unrelated samples. In Figure S1, the quantile-quantile (QQ) plots for both quantitative and binary traits show that all MAGEE main effect tests, GEI tests, and joint tests are well calibrated.

Similarly, Table 2 shows the empirical type I error rates of MAGEE tests for 100,000 related samples from 25,000 families at the same significance levels as Table 1. Figure 1 shows that with related individuals in the sample, existing methods that ignore the correlation structure, such as rareGE and MiSTi, give inflated type I error rates. In contrast, MAGEE tests control type I errors successfully. Due to the computational speed issue, rareGE and MiSTi do not scale up to the sample size of 100,000. Therefore, analyses in Figure 1 were performed using 10,000 related samples from 2,500 families, in which 119,400 p values (1 genotype replicate $\times$ 100 phenotype replicates $\times$ 1,194 variant sets) were estimated. Figure S2 shows the QQ plots of MAGEE tests

14

for quantitative and binary traits with 100,000 related samples from 47,760,000 p values, which are consistent with our results from 100,000 unrelated samples in Figure S1.

3.1.2 Computation time and p value comparison

We compared the p values from MAGEE with those from rareGE and MiSTi for both quantitative and binary traits in 2,000, 5,000, and 10,000 unrelated samples. Each panel in Figure 2 (scenario 1 for quantitative trait: both genetic main effects and GEI effects) displays 1,000 p values from quantitative trait analyses with unrelated samples when both genetic main effects and GEI effects were simulated. MAGEE IV, IF, and JV tests have close p values with rareGE GEI test, MiSTi test, and rareGE JOINT test, respectively, and the accuracy increases with the sample size. Similar results are found in Figure S3 (scenario 2 for quantitative trait: genetic main effects only), in which only the genetic main effects but no GEI effects were simulated, as well as Figures S4 and S5 (scenario 1 and scenario 2 for binary traits).

Figure 3 compares the CPU time per p value on a single thread for MAGEE, rareGE, and MiSTi in quantitative trait analyses using unrelated individuals. As both rareGE and MiSTi require fitting a separate statistical model for each genetic variant set, their CPU time increases dramatically with the sample size, while MAGEE tests remain computationally efficient in large samples. Similar results are also observed in Figure S6 from binary trait analyses. In addition, MAGEE joint tests are performed simultaneously with the main effect and GEI tests. For example, when performing the JV test, MV and IV test results will be produced automatically; when performing either the JF or JD test, MF and IF test results will also be computed automatically.

3.1.3 Power

Figure 4 shows the empirical power of the seven tests within the MAGEE framework in analyzing quantitative traits (scenario 1 to 6), at the significance level of 2.5×10^{-6} with 20,000, 50,000,

15

and 100,000 related samples, respectively. The top panels show results from the scenarios 1-3, which have variants with 80% null, 10% positive, and 10% negative effects. When both genetic main effects and GEI effects are present (Figure 4A), the three joint tests are most powerful and they have very close power that increases along with the sample size. In the other two scenarios with only either genetic main effects (Figure 4B) or GEI effects (Figure 4C), the joint tests are less powerful than main effects tests or GEI tests, respectively. In general, the variance component tests (MV, IV and JV) and the hybrid tests (MF, IF, JF and JD) have close power, with the hybrid tests power being slightly more powerful in these simulation settings. The bottom panels show results from the scenarios 4-6, which have variants with 80% null, 16% positive, and 4% negative effects. The results are similar with scenarios 1-3 except that the hybrid tests have recognizable higher power compared to the variance component tests, for all main effect, GEI and joint tests. Furthermore, the JF test is slightly less powerful than the JD test, except in Figure 4A and 4D, when both genetic main effects and GEI effects were simulated. Similar results with six simulation scenarios for binary traits are shown in Figure S7. In general settings, we would recommend IF for the GEI test, and JD for the joint test, if we do not have any prior knowledge about the genetic architecture of main effects or GEI effects.

3.2 UK Biobank whole exome sequencing data

The analysis of all 18,970 protein coding regions were finished in 10.4 hours using a single thread on a computing server for a single phenotype. Figure 5 shows that the IF tests are well calibrated for both the quantitative trait BMI and the dichotomized trait obesity, while the mild inflations in the JD tests are likely attributable to the main effect tests, as expected for polygenic traits. However, from Figure 6 one can see that across the 18,970 protein-coding regions, only one significant p value in the melanocortin 4 receptor (*MC4R*) gene region on chromosome 18 was found in the JD

16

test of gene-sex interaction for BMI at the significance level of $0.05/18{,}970 = 2.64 \times 10^{-6}$ after

Bonferroni correction for multiple testing (Bland & Altman, 1995) . The IF test p value in this

region was not significant ($p = 0.26$), so the significant signal in the JD test ($p = 2.29 \times 10^{-7}$)

was mostly driven by genetic main effects (MF test $p = 4.56 \times 10^{-8}$). *MC4R* is one of the most

common monogenic cause of severe obesity in humans (Krashes, Lowell, & Garfield, 2016; Vaisse,

Clement, Guy-Grand, & Froguel, 1998). Previously, genome-wide association studies have

identified common variants in or near *MC4R* with sex-specific effects on human brain structure

and eating behavior (Horstmann et al., 2013), but it is unknown whether there are sex-specific

genetic effects on BMI in this region.

4. DISCUSSION

We have developed computationally efficient variant set-based mixed model GEI tests and joint

tests in the MAGEE framework, which can be applied to both quantitative and binary traits in large

biobank-scale sequencing studies with hundreds of thousands of possibly related individuals. We

have shown in simulation studies that existing GEI tests for variant sets developed for unrelated

samples would have inflated type I error rates when applied to related samples, while MAGEE

successfully controls type I errors in both unrelated and related samples. MAGEE requires fitting

a global null model only once for all the tests across the whole genome, and it uses two matrix

projection approaches to approximate the test statistics. MAGEE accounts for sample relatedness

using GLMMs, and it greatly reduces the computational complexity for testing each variant set

from $O(N^3)$ to no more than $O(N^2)$ with highly accurate approximations, making it the method

of choice. For samples with a block-diagonal relatedness structure (such as family studies), the

computational complexity for testing each variant set is further reduced to $O(nN)$, where n is the

maximum number of individuals in each block (e.g., the family size), which is often much smaller

17

than the total sample size N. For unrelated samples, MAGEE tests have a computational complexity of $O(N)$. The CPU time of MAGEE is much smaller than existing methods, since it does not require fitting a separate statistical model to account for genetic main effects, for each variant set in the whole genome analysis. Moreover, the MAGEE joint tests are unique because p values are computed analytically, instead of using the Monte Carlo approaches as implemented in rareGE (Chen et al., 2014). Based on our simulation studies, the hybrid test IF is more powerful than the variance component test IV for GEI effects, especially when the interaction effects do not have a mean of 0, and the hybrid test JD is usually the most powerful joint test. When both genetic main effects and GEI effect are present on approximately the same scale, the hybrid test JF is slightly more powerful than the JD test, but the power difference is negligible. In reality, with little knowledge on the genetic architecture of a quantitative or binary trait, the IF test is recommended for identifying GEI effects, and the JD test is recommended for identifying genetic associations that allow for heterogeneous effects in different environmental exposures.

We applied MAGEE to the WES data from the UK Biobank. The results showed that MAGEE p values were well calibrated in these real data applications, and we identified an association between BMI and *MC4R* gene from the joint test. However, we did not find any significant p values from the interaction test. It is possible that interaction effects may be too small to identify in 41,144 samples. As the WES project is ongoing in the UK Biobank, we hope to revisit gene-environment interaction analyses when WES data from more UK Biobank samples are released in the coming years.

Recently, StructLMM was developed to test GEI effects for a single genetic variant with high-dimensional environmental factors (Moore et al., 2019). While MAGEE can test GEI effects for multiple genetic variants with multiple environmental factors, its performance in a high-

dimensional setting (e.g., a large number of genetic variants along with hundreds of environmental factors) has not been fully investigated. Moreover, MAGEE GEI and joint tests using summary statistics in a meta-analysis setting can further boost statistical power by combining association evidence from multi-million samples from large-scale sequencing studies in upcoming years. Our MAGEE framework provides a foundation for future research in these directions.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

SOFTWARE

We have implemented MAGEE in an R package, available at https://github.com/xwang21/magee.

REFERENCES

Bi, W., Zhao, Z., Dey, R., Fritsche, L. G., Mukherjee, B., & Lee, S. (2019). A fast and accurate method for genome-wide scale phenome-wide G × E analysis and its application to UK biobank. *The American Journal of Human Genetics, 105*(6), 1182-1192. doi:10.1016/j.ajhg.2019.10.008

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: The bonferroni method. *BMJ: British Medical Journal, 310*(6973), 170. doi:10.1136/bmj.310.6973.170

Chen, H., Huffman, J. E., Brody, J. A., Wang, C., Lee, S., Li, Z., . . . Lin, X. (2019). Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *The American Journal of Human Genetics, 104*(2), 260-274. doi:10.1016/j.ajhg.2018.12.012

Chen, H., Meigs, J. B., & Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. *Genetic Epidemiology, 37*(2), 196. doi:10.1002/gepi.21703

Chen, H., Meigs, J. B., & Dupuis, J. (2014). Incorporating gene-environment interaction in testing for association with rare genetic variants. *Human Heredity, 78*(2), 81-90. doi:10.1159/000363347

Conomos, M. P., Laurie, C. A., Stilp, A. M., Gogarten, S. M., Mchugh, C. P., Nelson, S. C., . . . Laurie, C. C. (2016). Genetic diversity and association studies in US hispanic/latino populations: Applications in the hispanic community health study/study of latinos. *American Journal of Human Genetics, 98*(1), 165. doi:10.1016/j.ajhg.2015.12.001

Cornelis, M. C., Tchetgen, E. J. T., Liang, L., Qi, L., Chatterjee, N., Hu, F. B., & Kraft, P. (2012). Gene-environment interactions in genome-wide association studies: A comparative study of tests applied to empirical studies of type 2 diabetes. *American Journal of Epidemiology, 175*(3), 191. doi:10.1093/aje/kwr368

Elks, C. E., Den Hoed, M., Zhao, J. H., Sharp, S. J., Wareham, N. J., Loos, R. J. F., & Ong, K. K. (2012). Variability in the heritability of body mass index: A systematic review and meta-regression. *Frontiers in Endocrinology, 3*, 29. doi:10.3389/fendo.2012.00029

Fisher, R. A. (1928). Statistical methods for research workers. second edition, revised and enlarged. edinburgh: Oliver and boyd. *Journal of Comparative Pathology and Therapeutics, 41*, 261-262. doi:10.1016/S0368-1742(28)80029-4

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning data mining, inference, and prediction, second edition* (2nd ed.). New York, NY: Springer New York. doi:10.1007/978-0-387-84858-7

Horstmann, A., Kovacs, P., Kabisch, S., Boettcher, Y., Schloegl, H., Tonjes, A., . . . Villringer, A. (2013). Common genetic variation near MC4R has a sex-specific impact on human brain structure and eating behavior. *PLoS ONE, 8*(9), 74362. doi:10.1371/journal.pone.0074362

Kraft, P., Yen, Y., Stram, D. O., Morrison, J., & Gauderman, W. J. (2007). Exploiting gene-environment interaction to detect genetic associations. *Human Heredity, 63*(2), 111-119. doi:10.1159/000099183

Krashes, M. J., Lowell, B. B., & Garfield, A. S. (2016). Melanocortin-4 receptor-regulated energy homeostasis. *Nature Neuroscience, 19*(2), 206. doi:10.1038/nn.4202

Lee, S., Wu, M. C., & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics (Oxford, England), 13*(4), 762. doi:10.1093/biostatistics/kxs014

Lim, E., Chen, H., Dupuis, J., & Liu, C. (2020). A unified method for rare variant analysis of gene-environment interactions. *Statistics in Medicine, ,* 39(6):801-813. doi:10.1002/sim.8446

Lin, X., Lee, S., Wu, M. C., Wang, C., Chen, H., Li, Z., & Lin, X. (2016). Test for rare variants by environment interactions in sequencing association studies. *Biometrics, 72*(1), 156-164. doi:10.1111/biom.12368

Mazo Lopera, M., Coombes, B. J., & de Andrade, M. (2017). An efficient test for gene-environment interaction in generalized linear mixed models with family data. *International Journal of Environmental Research and Public Health, 14*(10), 1134. doi:10.3390/ijerph14101134

Mcallister, K., Mechanic, L. E., Amos, C., Aschard, H., Blair, I. A., Chatterjee, N., . . . Witte, J. S. (2017). Current challenges and new opportunities for gene-environment interaction studies of complex diseases. *American Journal of Epidemiology, 186*(7), 753. doi:10.1093/aje/kwx227

Moore, R., Casale, F. P., Jan Bonder, M., Horta, D., Franke, L., Barroso, I., & Stegle, O. (2019). A linear mixed-model approach to study multivariate gene-environment interactions. *Nature Genetics, 51*(1), 180. doi:10.1038/s41588-018-0271-0

Pan, W., Kim, J., Zhang, Y., Shen, X., & Wei, P. (2014). A powerful and adaptive association test for rare variants. *Genetics, 197*(4), 1081. doi:10.1534/genetics.114.165035

Regier, A. A., Farjoun, Y., Larson, D. E., Krasheninina, O., Kang, H. M., Howrigan, D. P., . . . Hall, I. M. (2018). Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nature Communications, 9*(1), 4038. doi:10.1038/s41467-018-06159-4

Sabeti, P. C. (2015). A global reference for human genetic variation. *Nature, 526*(7571) doi:10.1038/nature15393

Selinger-Leneman, H., Genin, E., Norris, J. M., & Khlat, M. (2003). Does accounting for gene-environment (G×E) interaction increase the power to detect the effect of a gene in a multifactorial disease? *Genetic Epidemiology, 24*(3), 200-207. doi:10.1002/gepi.10221

Su, Y., Di, C., & Hsu, L. (2017). A unified powerful set-based test for sequencing data analysis of GxE interactions. *Biostatistics (Oxford, England), 18*(1), 119. doi:10.1093/biostatistics/kxw034

Su, Z., Marchini, J., & Donnelly, P. (2011). HAPGEN2: Simulation of multiple disease SNPs. *Bioinformatics (Oxford, England), 27*(16), 2304. doi:10.1093/bioinformatics/btr341

Sun, J., Zheng, Y., & Hsu, L. (2013). A unified Mixed-Effects model for Rare-Variant association in sequencing studies. *Genetic Epidemiology, 37*(4), 334-344. doi:10.1002/gepi.21717

Sun, R., Carroll, R. J., Christiani, D. C., & Lin, X. (2018). Testing for gene–environment interaction under exposure misspecification. *Biometrics, 74*(2), 653-662. doi:10.1111/biom.12813

Vaisse, C., Clement, K., Guy-Grand, B., & Froguel, P. (1998). A frameshift mutation in human MC4R is associated with a dominant form of obesity. *Nature Genetics, 20*(2), 113. doi:10.1038/2404

Wu, M., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics, 89*(1), 82-93. doi:10.1016/j.ajhg.2011.05.029

SUPPORTING INFORMATION

Supporting information includes seven supplemental figures and three supplemental tables.

TABLES

Table 1. Empirical type I error rates of MAGEE tests for 100,000 unrelated individuals at significance levels of 0.05, 1.0×10^{-4}, and 2.5×10^{-6}.

	Quantitative trait Significance Level			**Binary trait** Significance Level		
Test	0.05	1.0×10^{-4}	2.5×10^{-6}	0.05	1.0×10^{-4}	2.5×10^{-6}
MV	0.050	1.00×10^{-4}	2.74×10^{-6}	0.050	1.02×10^{-4}	2.35×10^{-6}
MF	0.050	1.01×10^{-4}	2.49×10^{-6}	0.050	1.00×10^{-4}	2.84×10^{-6}
IV	0.050	1.04×10^{-4}	2.47×10^{-6}	0.050	1.01×10^{-4}	2.31×10^{-6}
IF	0.050	1.01×10^{-4}	2.70×10^{-6}	0.050	9.93×10^{-5}	2.14×10^{-6}
JV	0.050	1.02×10^{-4}	2.51×10^{-6}	0.050	1.01×10^{-4}	2.44×10^{-6}
JF	0.050	1.00×10^{-4}	2.55×10^{-6}	0.050	1.02×10^{-4}	2.59×10^{-6}
JD	0.050	1.00×10^{-4}	2.53×10^{-6}	0.050	1.01×10^{-4}	2.40×10^{-6}

Table 2. Empirical type I error rates of MAGEE tests for 100,000 related individuals at significance levels of 0.05, 1.0×10^{-4}, and 2.5×10^{-6}.

	Quantitative trait Significance Level			**Binary trait** Significance Level		
Test	0.05	1.0×10^{-4}	2.5×10^{-6}	0.05	1.0×10^{-4}	2.5×10^{-6}
MV	0.050	9.97×10^{-5}	2.32×10^{-6}	0.050	9.92×10^{-5}	2.28×10^{-6}
MF	0.050	1.01×10^{-4}	2.72×10^{-6}	0.050	9.93×10^{-5}	1.86×10^{-6}
IV	0.050	9.82×10^{-5}	1.99×10^{-6}	0.048	8.90×10^{-5}	2.07×10^{-6}
IF	0.050	9.87×10^{-5}	1.97×10^{-6}	0.048	8.81×10^{-5}	2.14×10^{-6}
JV	0.050	9.99×10^{-5}	2.81×10^{-6}	0.049	9.55×10^{-5}	2.14×10^{-6}
JF	0.050	9.98×10^{-5}	2.74×10^{-6}	0.048	9.30×10^{-5}	2.11×10^{-6}
JD	0.050	1.01×10^{-4}	2.83×10^{-6}	0.048	9.14×10^{-5}	2.16×10^{-6}

FIGURE LEGENDS

Figure 1. Quantile-Quantile plots of MAGEE, rareGE and MiSTi tests on quantitative traits in 10,000 related samples. (A) MAGEE IV, rareGE and MiSTi GEI tests. (B) MAGEE JV and rareGE joint tests.
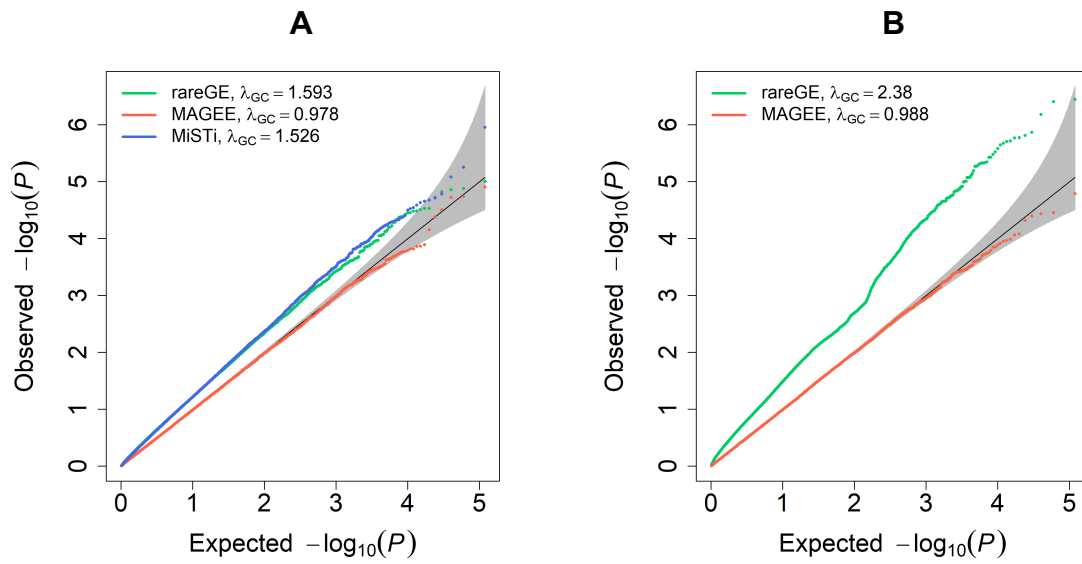
Figure 2. Comparison of p values from MAGEE versus rareGE and MiSTi tests on quantitative traits when both genetic and GEI effects were present (scenario 1) in 2,000, 5,000, and 10,000 unrelated samples. (A) MAGEE IV versus rareGE GEI tests. (B) MAGEE IF versus MiSTi GEI tests. (C) MAGEE JV versus rareGE joint tests.

27
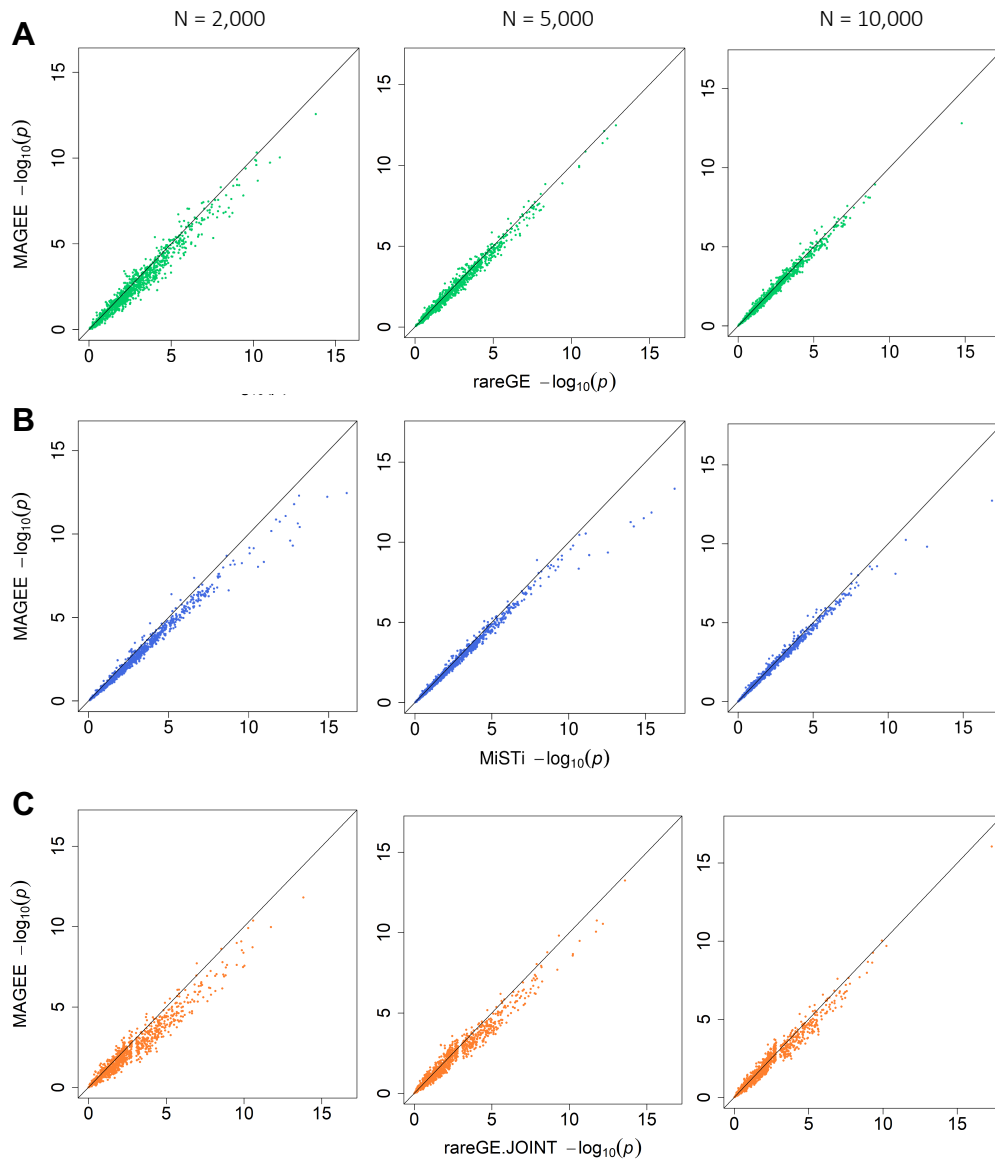
Figure 3. CPU time per *p* value of MAGEE, rareGE and MiSTi tests on quantitative traits in unrelated samples. (A) MAGEE, rareGE and MiSTi GEI tests. (B) MAGEE and rareGE joint tests.
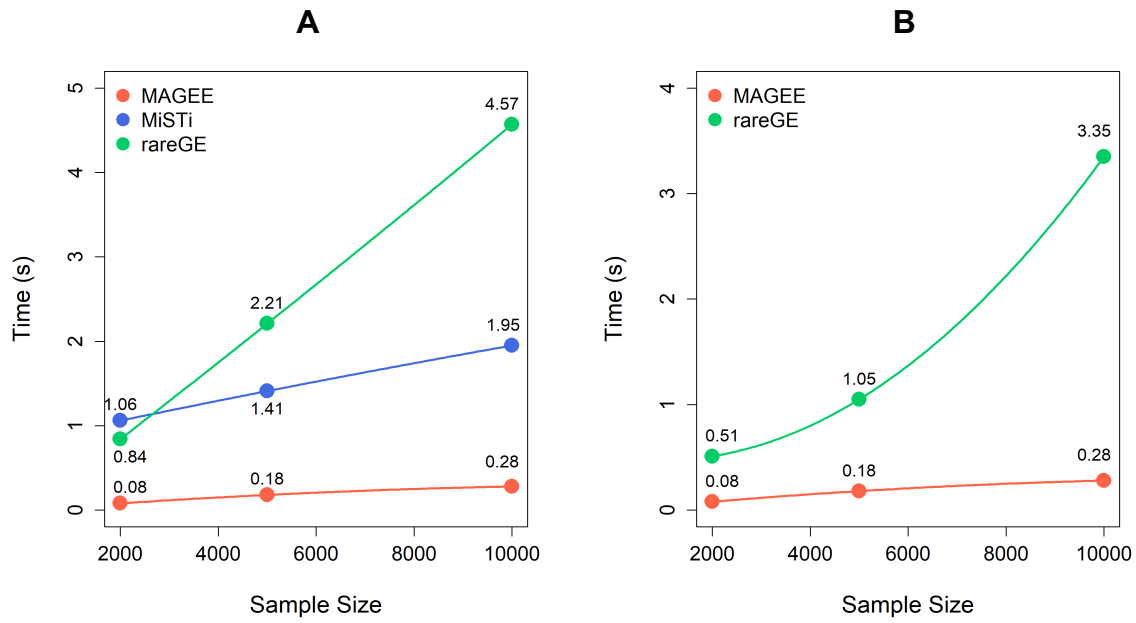
Figure 4. Empirical power of MAGEE tests on quantitative traits in 20,000, 50,000, and 100,000 related samples. (A) Scenario 1: 80% null variants, 10% causal variants with positive effects and 10% causal variants with negative effects for both genetic main effects and GEI effects. (B) Scenario 2: 80% null variants, 10% causal variants with positive effects and 10% causal variants with negative effects for genetic main effects only. (C) Scenario 3: 80% null variants, 10% causal variants with positive effects and 10% causal variants with negative effects for GEI effects only. (D) Scenario 4: 80% null variants, 16% causal variants with positive effects and 4% causal variants with negative effects for both genetic main effects and GEI effects. (E) Scenario 5: 80% null variants, 16% causal variants with positive effects and 4% causal variants with negative effects for genetic main effects only. (F) Scenario 6: 80% null variants, 16% causal variants with positive effects and 4% causal variants with negative effects for GEI effects only.
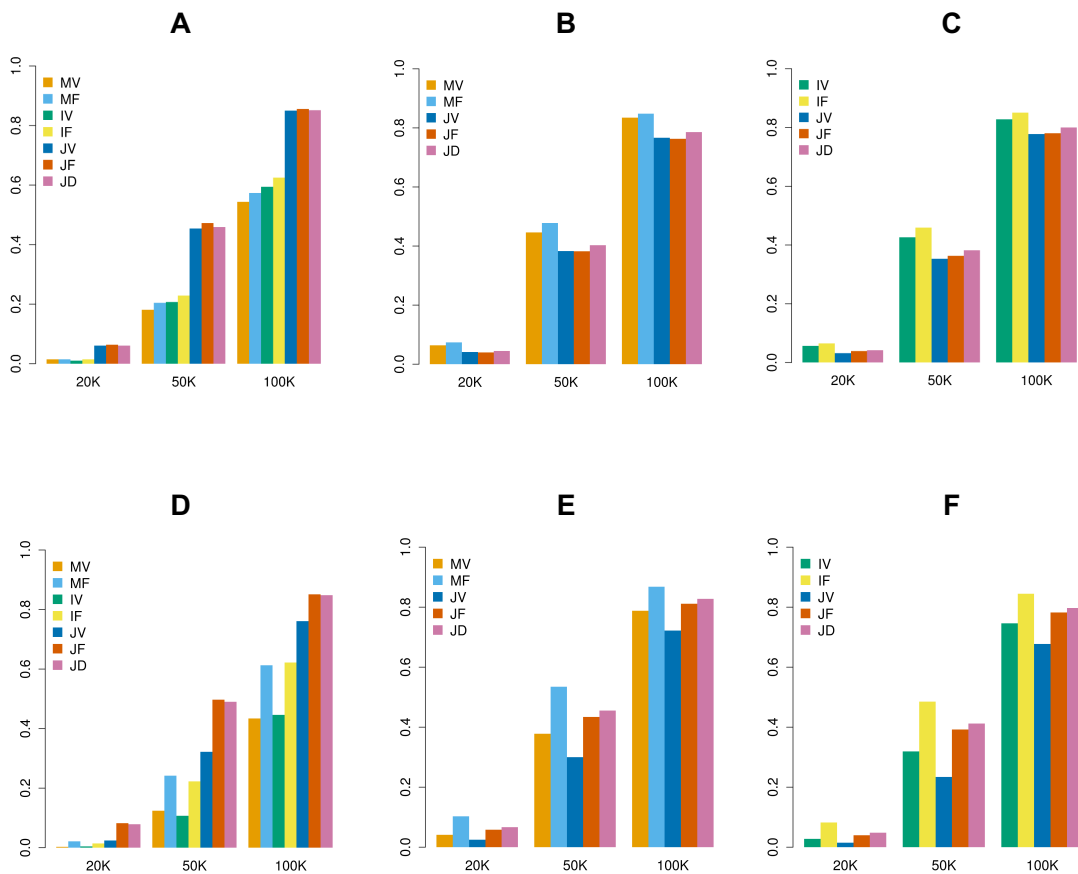
Figure 5. Quantile-Quantile plots of UK Biobank WES data analysis using MAGEE tests for gene-sex interaction effects on BMI and obesity. (A) BMI analysis. (B) Obesity analysis.
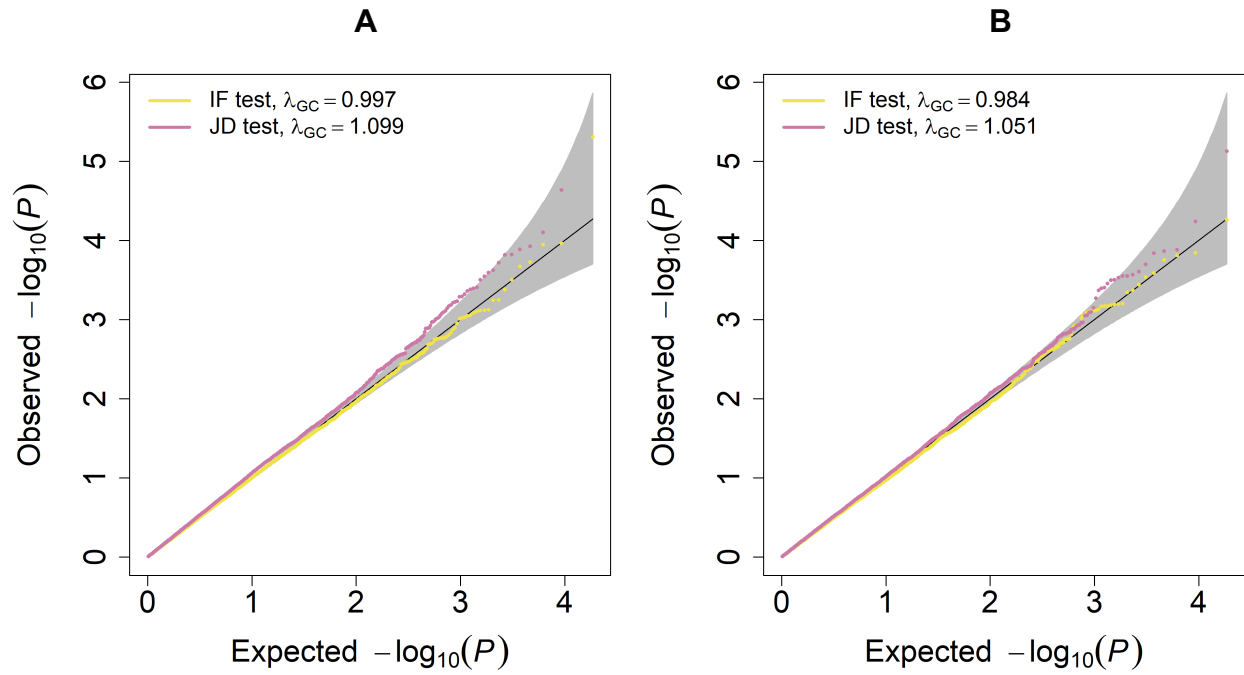
Figure 6. Manhattan plots of UK Biobank WES data analysis using MAGEE tests for gene-sex interaction effects on BMI and obesity. (A) MAGEE IF test on BMI. (B) MAGEE JD test on BMI. (C) MAGEE IF test on obesity. (D) MAGEE JD test on obesity.
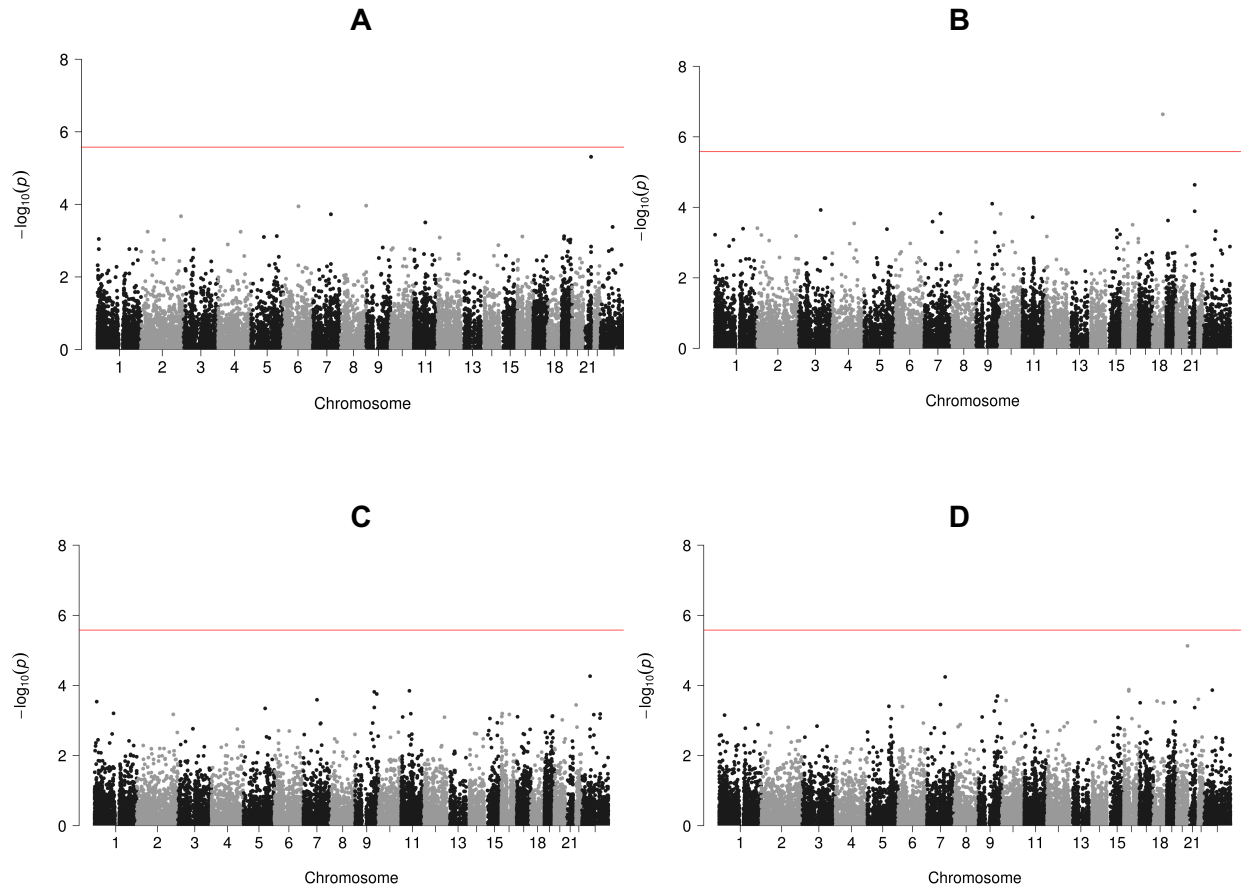
APPENDIX A: APPROXIMATIONS FOR THE SCORE VECTOR $\tilde{S}_K$

Let $\tilde{\alpha}$, $\tilde{\beta}$, and $\tilde{r}$ be the estimates for α, β, and r from Equation 2, $\tilde{Y}$ is the working vector for the GLMM and $\tilde{Y} = y$ for quantitative trait, and $\tilde{Y}_i = X_i\tilde{\alpha} + G_i\tilde{\beta} + \tilde{r}_i + \{\tilde{\mu}_{Gi}(1 - \tilde{\mu}_{Gi})\}^{-1}(y_i - \tilde{\mu}_{Gi})$ for binary traits.

$$\frac{y - \tilde{\mu}_G}{\tilde{\phi}} = \tilde{V}^{-1}(\tilde{Y} - X\tilde{\alpha} - G\tilde{\beta} - \tilde{r}),$$

$$= \tilde{\Sigma}^{-1}(\tilde{Y} - X\tilde{\alpha} - G\tilde{\beta})$$

$$= \tilde{\Sigma}^{-1}\left\{\tilde{Y} - (X\ G)\begin{pmatrix}X^T\tilde{\Sigma}^{-1}X & X^T\tilde{\Sigma}^{-1}G \\ G^T\tilde{\Sigma}^{-1}X & G^T\tilde{\Sigma}^{-1}G\end{pmatrix}^{-1}\begin{pmatrix}X^T\tilde{\Sigma}^{-1} \\ G^T\tilde{\Sigma}^{-1}\end{pmatrix}\tilde{Y}\right\}$$

$$= \left\{\tilde{\Sigma}^{-1} - \tilde{\Sigma}^{-1}X(X^T\tilde{\Sigma}^{-1}X)^{-1}X^T\tilde{\Sigma}^{-1}\right\}\tilde{Y} - \left\{\tilde{\Sigma}^{-1} - \tilde{\Sigma}^{-1}X(X^T\tilde{\Sigma}^{-1}X)^{-1}X^T\tilde{\Sigma}^{-1}\right\}$$

$$G\left(G^T\left\{\tilde{\Sigma}^{-1} - \tilde{\Sigma}^{-1}X(X^T\tilde{\Sigma}^{-1}X)^{-1}X^T\tilde{\Sigma}^{-1}\right\}G\right)^{-1}G^T\left\{\tilde{\Sigma}^{-1} - \tilde{\Sigma}^{-1}X(X^T\tilde{\Sigma}^{-1}X)^{-1}X^T\tilde{\Sigma}^{-1}\right\}\tilde{Y}$$

Assuming the true value of β is small, including the $G_i\beta$ term in Equation (2) does not dramatically change the variance component estimates for τ and ϕ from Equation (3), we can approximate that $\tilde{\Sigma}^{-1} - \tilde{\Sigma}^{-1}X(X^T\tilde{\Sigma}^{-1}X)^{-1}X^T\tilde{\Sigma}^{-1} \approx \hat{P}$ and $\frac{y - \hat{\mu}_0}{\hat{\phi}} \approx \hat{P}\tilde{Y}$, so

$$\tilde{S}_K = \frac{K^T(y - \tilde{\mu}_G)}{\tilde{\phi}} \approx K^T\hat{P}\tilde{Y} - K^T\hat{P}G(G^T\hat{P}G)^{-1}G^T\hat{P}\tilde{Y} \approx S_K - K^T\hat{P}G(G^T\hat{P}G)^{-1}S_G$$

We can rewrite $\tilde{S}_K$ in the matrix form as $\left(-K^T\hat{P}G(G^T\hat{P}G)^{-1}\ I\right)\begin{pmatrix}S_G \\ S_K\end{pmatrix}$, where $\begin{pmatrix}S_G \\ S_K\end{pmatrix} \sim N(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}G^T\hat{P}G & G^T\hat{P}K \\ K^T\hat{P}G & K^T\hat{P}K\end{pmatrix})$, the variance of $\tilde{S}_K$ is then

$$\Lambda = \left(-K^T\hat{P}G(G^T\hat{P}G)^{-1}\ I\right)\begin{pmatrix}G^T\hat{P}G & G^T\hat{P}K \\ K^T\hat{P}G & K^T\hat{P}K\end{pmatrix}\begin{pmatrix}-(G^T\hat{P}G)^{-1}G^T\hat{P}K \\ I\end{pmatrix}$$

$$= K^T\hat{P}K - K^T\hat{P}G(G^T\hat{P}G)^{-1}G^T\hat{P}K$$

APPENDIX B: APPROXIMATIONS FOR THE SCORE VECTOR $\tilde{S}_{KV}$

To get $\tilde{S}_{KV}$, we need to fit Equation (A1)

$$g(\mu_i) = X_i\alpha + G_i\beta + K_iW_K\mathbf{1}_{cq}\gamma_0 + r_i \quad (A1)$$

Let $\widetilde{\alpha}$, $\widetilde{\beta}$, $\tilde{r}$, and $\tilde{\gamma}_0$ be the estimates for α, β, r and γ_0 from Equation (A1), $\widetilde{Y}$ is the working vector for Equation (A1) and $\widetilde{Y} = y$ for quantitative trait, and $\widetilde{Y}_i = X_i\widetilde{\alpha} + G_i\widetilde{\beta} + K_iW_K\mathbf{1}_{cq}\tilde{\gamma}_0 + \tilde{r}_i + \{\tilde{\mu}_{Ki}(1-\tilde{\mu}_{Ki})\}^{-1}(y_i - \tilde{\mu}_{Ki})$ for binary traits, where $\tilde{\mu}_{Ki}$ is the mean for individual i after fitting Equation (A1). The mean vector after fitting Equation (A1) is $\widetilde{\mu}_K$.

$$\frac{y - \widetilde{\mu}_K}{\tilde{\phi}} = \widetilde{V}^{-1}\left(\widetilde{Y} - X\widetilde{\alpha} - G\widetilde{\beta} - KW_K\mathbf{1}_{cq}\tilde{\gamma}_0 - \tilde{r}\right)$$

$$= \widetilde{\Sigma}^{-1}\left(\widetilde{Y} - X\widetilde{\alpha} - G\widetilde{\beta} - KW_K\mathbf{1}_{cq}\tilde{\gamma}_0\right)$$

Let $Z = (G \quad KW_K\mathbf{1}_{cq})$,

$$\frac{y - \widetilde{\mu}_K}{\tilde{\phi}} = \widetilde{\Sigma}^{-1}\left\{\widetilde{Y} - (X \quad Z)\begin{pmatrix} X^T\widetilde{\Sigma}^{-1}X & X^T\widetilde{\Sigma}^{-1}Z \\ Z^T\widetilde{\Sigma}^{-1}X & Z^T\widetilde{\Sigma}^{-1}Z \end{pmatrix}^{-1}\begin{pmatrix} X^T\widetilde{\Sigma}^{-1} \\ Z^T\widetilde{\Sigma}^{-1} \end{pmatrix}\widetilde{Y}\right\}$$

$$= \left\{\widetilde{\Sigma}^{-1} - \widetilde{\Sigma}^{-1}X\left(X^T\widetilde{\Sigma}^{-1}X\right)^{-1}X^T\widetilde{\Sigma}^{-1}\right\}\widetilde{Y} - \left\{\widetilde{\Sigma}^{-1} - \widetilde{\Sigma}^{-1}X\left(X^T\widetilde{\Sigma}^{-1}X\right)^{-1}X^T\widetilde{\Sigma}^{-1}\right\}$$

$$Z\left[Z^T\left\{\widetilde{\Sigma}^{-1} - \widetilde{\Sigma}^{-1}X\left(X^T\widetilde{\Sigma}^{-1}X\right)^{-1}X^T\widetilde{\Sigma}^{-1}\right\}Z\right]^{-1}Z^T\left\{\widetilde{\Sigma}^{-1} - \widetilde{\Sigma}^{-1}X\left(X^T\widetilde{\Sigma}^{-1}X\right)^{-1}X^T\widetilde{\Sigma}^{-1}\right\}\widetilde{Y}$$

Let $\widetilde{P} = \widetilde{\Sigma}^{-1} - \widetilde{\Sigma}^{-1}X\left(X^T\widetilde{\Sigma}^{-1}X\right)^{-1}X^T\widetilde{\Sigma}^{-1}$,

$$\frac{y - \widetilde{\mu}_K}{\tilde{\phi}} = \widetilde{P}\widetilde{Y} - \widetilde{P}(G \quad K)\begin{pmatrix} G^T\widetilde{P}G & G^T\widetilde{P}KW_K\mathbf{1}_{cq} \\ \mathbf{1}_{cq}^T W_K K^T\widetilde{P}G & \mathbf{1}_{cq}^T W_K K^T\widetilde{P}KW_K\mathbf{1}_{cq} \end{pmatrix}^{-1}\begin{pmatrix} G^T\widetilde{P} \\ \mathbf{1}_{cq}^T W_K K^T\widetilde{P} \end{pmatrix}\widetilde{Y}$$

$$= \left\{\widetilde{P} - \widetilde{P}G\left(G^T\widetilde{P}G\right)^{-1}G^T\widetilde{P}\right\}\widetilde{Y} - \left\{\widetilde{P} - \widetilde{P}G\left(G^T\widetilde{P}G\right)^{-1}G^T\widetilde{P}\right\}KW_K\mathbf{1}_{cq}$$

$$\left[\mathbf{1}_{cq}^T W_K K^T\left\{\widetilde{P} - \widetilde{P}G\left(G^T\widetilde{P}G\right)^{-1}G^T\widetilde{P}\right\}KW_K\mathbf{1}_{cq}\right]^{-1}\mathbf{1}_{cq}^T W_K K^T\left\{\widetilde{P} - \widetilde{P}G\left(G^T\widetilde{P}G\right)^{-1}G^T\widetilde{P}\right\}\widetilde{Y}$$

33

Assuming the true values of $\boldsymbol{\beta}$ and γ_0 are small, including the terms $\boldsymbol{G_i\beta}$ and $\boldsymbol{K_i W_K 1_{cq}}\gamma_0$ in Equation (A1) does not dramatically change the variance component estimates for τ and ϕ from Equation (3), so we can approximate that $\widetilde{\widetilde{P}} = \widetilde{\widetilde{\Sigma}}^{-1} - \widetilde{\widetilde{\Sigma}}^{-1}X\left(X^T\widetilde{\widetilde{\Sigma}}^{-1}X\right)^{-1}X^T\widetilde{\widetilde{\Sigma}}^{-1} \approx \widehat{P}$ and $\frac{y-\widehat{\mu}_0}{\widehat{\phi}} \approx$

$\widehat{P}\widetilde{\widetilde{Y}}$, so

$$\tilde{S}_{KV} = \frac{K^T\left(y - \widehat{\widetilde{\mu}}_K\right)}{\widehat{\widetilde{\phi}}} \approx K^T\widehat{P}\widetilde{\widetilde{Y}} - K^T\widehat{P}G(G^T\widehat{P}G)^{-1}G^T\widehat{P}\widetilde{\widetilde{Y}} - K^T\widehat{P}KW_K 1_{cq}$$

$$\left[1_{cq}^T W_K\left\{K^T\widehat{P}K - K^T\widehat{P}G(G^T\widehat{P}G)^{-1}G^T\widehat{P}K\right\}W_K 1_{cq}\right]^{-1}1_{cq}^T W_K K^T\widehat{P}\widetilde{\widetilde{Y}}$$

$$+K^T\widehat{P}G(G^T\widehat{P}G)^{-1}G^T\widehat{P}KW_K 1_{cq}\left[1_{cq}^T W_K\left\{K^T\widehat{P}K - K^T\widehat{P}G(G^T\widehat{P}G)^{-1}G^T\widehat{P}K\right\}W_K 1_{cq}\right]^{-1}$$

$$1_{cq}^T W_K K^T\widehat{P}\widetilde{\widetilde{Y}} + K^T\widehat{P}KW_K 1_{cq}\left[1_{cq}^T W_K\left\{K^T\widehat{P}K - K^T\widehat{P}G(G^T\widehat{P}G)^{-1}G^T\widehat{P}K\right\}W_K 1_{cq}\right]^{-1}$$

$$1_{cq}^T W_K K^T\widehat{P}G(G^T\widehat{P}G)^{-1}G^T\widehat{P}\widetilde{\widetilde{Y}} - K^T\widehat{P}G(G^T\widehat{P}G)^{-1}G^T\widehat{P}KW_K 1_{cq}$$

$$\left[1_{cq}^T W_K\left\{K^T\widehat{P}K - K^T\widehat{P}G(G^T\widehat{P}G)^{-1}G^T\widehat{P}K\right\}W_K 1_{cq}\right]^{-1}1_{cq}^T W_K K^T\widehat{P}G(G^T\widehat{P}G)^{-1}G^T\widehat{P}\widetilde{\widetilde{Y}}$$

Notice that $\Lambda = K^T\widehat{P}K - K^T\widehat{P}G(G^T\widehat{P}G)^{-1}G^T\widehat{P}K$, $\tilde{S}_K \approx S_K - K^T\widehat{P}G(G^T\widehat{P}G)^{-1}S_G$ and $\tilde{S}_{KB} = 1_{cq}^T W_K \tilde{S}_K$,

$$\tilde{S}_{KV} \approx K^T\widehat{P}\widetilde{\widetilde{Y}} - K^T\widehat{P}G(G^T\widehat{P}G)^{-1}G^T\widehat{P}\widetilde{\widetilde{Y}} - \Lambda W_K 1_{cq}\left(1_{cq}^T W_K \Lambda W_K 1_{cq}\right)^{-1}1_{cq}^T W_K$$

$$\left\{K^T\widehat{P}\widetilde{\widetilde{Y}} - K^T\widehat{P}G(G^T\widehat{P}G)^{-1}G^T\widehat{P}\widetilde{\widetilde{Y}}\right\}$$

$$\approx \tilde{S}_K - \Lambda W_K 1_{cq}(1_{cq}^T W_K \Lambda W_K 1_{cq})^{-1}\tilde{S}_{KB}$$

$\tilde{S}_{KB}$ and $\tilde{S}_{KV}$ are asymptotically independent because they are asymptotically normal with covariance

$$Cov(\tilde{S}_{KV}, \tilde{S}_{KB}) = Cov\left(\left\{I_{cq} - \Lambda W_K 1_{cq}\left(1_{cq}^T W_K \Lambda W_K 1_{cq}\right)^{-1}1_{cq}^T W_K\right\}\tilde{S}_K, 1_{cq}^T W_K \tilde{S}_K\right)$$

$$= \left\{I_{cq} - \Lambda W_K 1_{cq}\left(1_{cq}^T W_K \Lambda W_K 1_{cq}\right)^{-1}1_{cq}^T W_K\right\}\Lambda W_K 1_{cq} = 0.$$