



A fast and powerful linear mixed model approach for genotype-environment interaction tests in large-scale GWAS

Wujuan Zhong , Aparna Chhibber, Lan Luo, Devan V. Mehrotra and Judong Shen 

Corresponding author: Judong Shen, Biostatistics and Research Decision Sciences, Merck & Co., Inc., Rahway, NJ 07065, USA. E-mail: judong.shen@merck.com

Abstract

Genotype-by-environment interaction (GEI or G×E) plays an important role in understanding complex human traits. However, it is usually challenging to detect GEI signals efficiently and accurately while adjusting for population stratification and sample relatedness in large-scale genome-wide association studies (GWAS). Here we propose a fast and powerful linear mixed model-based approach, fastGWA-GE, to test for GEI effect and G + G×E joint effect. Our extensive simulations show that fastGWA-GE outperforms other existing GEI test methods by controlling genomic inflation better, providing larger power and running hundreds to thousands of times faster. We performed a fastGWA-GE analysis of ~7.27 million variants on 452 249 individuals of European ancestry for 13 quantitative traits and five environment variables in the UK Biobank GWAS data and identified 96 significant signals (72 variants across 57 loci) with GEI test P-values $< 1 \times 10^{-9}$, including 27 novel GEI associations, which highlights the effectiveness of fastGWA-GE in GEI signal discovery in large-scale GWAS.

Introduction

Genome-wide association studies (GWAS) have reproducibly identified tens of thousands of genetic variants associated with complex human diseases. However, complex traits or diseases are usually influenced by the interplay of genetic and environmental variables such as duration of exercise, time spent watching TV, alcohol intake, cigarette smoking, new drug treatment and many others. Genotype-by-environment interaction (GEI or G×E) analysis can identify genetic loci that interact with environmental variables and influence diseases or traits via testing GEI effects or testing joint effects of genetic main effect and GEI effects (G + G×E joint effects). Large cohorts with genomic datasets of hundreds of thousands of samples, such as UK Biobank (UKB) ($N = \sim 500,000$), provide a great opportunity for GEI analyses since detection of GEI effects usually requires much larger sample size than detection of genetic main effects [1]. However, it is usually challenging to detect GEI effects efficiently and accurately while adjusting for population stratification and sample relatedness in large-scale GWAS analysis.

Multiple existing methods are available for genome-wide GEI analysis (Table S1). There are limitations of these existing methods when applied to large-scale GWAS GEI analysis. GEM (Gene-Environment interaction analysis in Millions of samples) is a fixed effects model-based approach with robust inference that can

handle both continuous and binary phenotypes [2]. It uses model-robust standard errors to test GEI effects and handles multiple environmental variables by including multiple GEI terms in the model. However, it cannot handle sample relatedness or adjust for other variants' genetic effects (i.e. it does not use a whole genome regression model). In contrast, mixed model-based methods can be used to account for sample relatedness and population stratification. Structured linear mixed model (StructLMM) is a mixed model-based approach that identifies loci interacting with one or more environmental variables [3]. It uses a mixed model to account for multiple environmental effects and adopts a variance component score test to evaluate the interaction between a single variant and multiple environmental variables. The main drawback of StructLMM is that it does not adjust for other variants' GEI effects. GENESIS is another mixed model-based approach using the average information restricted maximum likelihood (REML) procedure for single-SNP (single nucleotide polymorphism) or SNP-set association tests [4]. It provides a GEI test and adjusts for other variants' genetic main effects, but it does not adjust for other variants' GEI effects. More importantly, it does not apply model-robust standard errors when testing GEI or joint effects; thus, its GEI or joint test P-values may be inflated or deflated. LEMMA is a Bayesian whole-genome regression model and can handle multiple environmental variables by constructing an

Wujuan Zhong is a Senior Scientist, Biostatistics and Research Decision Sciences at Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA. He holds a PhD in Biostatistics and develops statistical methods for detecting biomarkers in pharmacogenomics and disease GWAS analysis and predicting DNA regulatory interactions in single-cell analysis.

Aparna Chhibber is a Senior Principal Scientist in Translational Bioinformatics at Bristol Myers Squibb. She has a PhD in Pharmaceutical Sciences and Pharmacogenomics. Her current research focuses on biomarker discovery in clinical trials.

Lan Luo is a Senior Scientist, Biostatistics and Research Decision Sciences, Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA. She holds a PhD in Biostatistics and develops statistical methods for detecting biomarkers in pharmacogenomics and disease studies.

Devan V. Mehrotra is Vice President, Biostatistics and Research Decision Sciences at Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA. His current research focus is on statistical methods for precision medicine.

Judong Shen is a Senior Director, Biostatistics and Research Decision Sciences, Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA. His research focuses on developing statistical methods for data analysis in pharmacogenomics, disease genetics, drug target sciences and translational biomarkers.

Received: July 7, 2022. Revised: October 26, 2022. Accepted: November 12, 2022

© The Author(s) 2022. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

environmental score, a linear combination of environmental variables, which interacts with genetic variants on the genome [5]. It adjusts for other variants' genetic main effects and GEI effects and adopts model-robust standard errors when testing GEI effects. However, it does not report the P-value of testing G + GxE joint effects and it is still very computationally demanding.

To overcome the limitations in the existing methods, especially the highly computational burden of conducting GEI analysis in large-scale GWAS, we propose an LMM-based approach, fastGWA-GE, which is developed on top of fastGWA [6] (a highly resource-efficient method/tool for testing the main genotype effect only) to test the GEI effect, the main G effect, and the genetic main and GEI joint effects. Compared with existing GEI analysis methods, fastGWA-GE runs hundreds to thousands of times faster. fastGWA-GE uses principal components (PCs) derived from genotype data to account for population stratification and adopts a weighted average of sparse genetic relationship matrix (sparse GRM) and the product of sparse GRM and the environment variable to account for sample relatedness. And it conducts whole-genome regression analysis. In other words, when testing each single variant, it adjusts for other variants' genetic main effects and GEI effects on the genome. Furthermore, fastGWA-GE implements model-robust (sandwich) standard errors, similar to that implemented in the LEMMA method, to correct for the inflation or deflation in P-values from testing GEI or G + GxE joint effects. Our extensive simulations reveal that fastGWA-GE outperforms alternative methods in terms of better controlling genomic inflation and providing larger power. We further demonstrate the utility of fastGWA-GE by analyzing ~7.27 million variants on 452 249 individuals of European ancestry for 13 quantitative traits and five environment variables in the UK Biobank GWAS data and discovered 96 significant signals with GEI test P-values $< 1 \times 10^{-9}$, including 27 novel GEI associations.

Materials and methods

The standard LMM used in GWAS when testing the main genotype effect for each variant is as $\mathbf{y} = \mathbf{G}\beta_G + \mathbf{X}_c\beta_c + \mathbf{g} + \epsilon$, where \mathbf{y} is an $n \times 1$ vector of mean centered phenotypes with n being the sample size; $\mathbf{G} = (G_1, G_2, \dots, G_n)^T$ is an $n \times 1$ vector denoting genotype of the single variant being tested; β_G is this variant's genetic main effect; \mathbf{X}_c is an $n \times p$ matrix denoting covariates with p being the number of covariates; β_c is a $p \times 1$ vector of fixed effects of covariates; \mathbf{g} is an $n \times 1$ vector of total genetic effects with $\mathbf{g} \sim N(\mathbf{0}, \mathbf{K}\sigma_{\text{main}}^2)$; ϵ is an $n \times 1$ vector of residuals with $\epsilon \sim N(\mathbf{0}, \mathbf{I}_n\sigma_\epsilon^2)$ where \mathbf{I}_n is an $n \times n$ identity matrix. The most straightforward method for specifying the kinship matrix [7] \mathbf{K} is to let $\mathbf{K} = \frac{\mathbf{W}\mathbf{W}^T}{q}$, where \mathbf{W} is an $n \times q$ standardized genotype matrix (columns normalized to have a mean of zero and variance of one), and q is the number of variants. This can be equivalently written as the infinitesimal model where every variant can have a minor but non-zero effect on the outcome [8],

$$\mathbf{y} = \mathbf{G}\beta_G + \mathbf{X}_c\beta_c + \mathbf{W}\mathbf{u} + \epsilon, \quad (1)$$

where \mathbf{u} is a $q \times 1$ vector of random effect of each variant's main effect with $\mathbf{u} \sim N(\mathbf{0}, \mathbf{I}_q \frac{\sigma_{\text{main}}^2}{q})$.

We propose a fastGWA-GE model via extending the infinitesimal model (equation 1) to test each variant's genetic main effect, GEI effect, and G + GxE joint effect while simultaneously modeling genome-wide genetic main and GEI effects,

$$\mathbf{y} = \mathbf{G}\beta_G + (\mathbf{G} \circ \mathbf{E})\beta_{\text{GEI}} + \mathbf{X}_c\beta_c + \mathbf{W}\mathbf{u} + \mathbf{D}\mathbf{W}\mathbf{v} + \epsilon, \quad (2)$$

where \circ denotes Hadamard (element-wise) product; $\mathbf{E} = (E_1, E_2, \dots, E_n)^T$ is an $n \times 1$ vector of the standardized environment variable with mean zero and variance one; β_{GEI} is the GEI effect of the variant being tested; \mathbf{D} is an $n \times n$ diagonal matrix in which the j th diagonal entry is E_j ; and \mathbf{v} is a $q \times 1$ vector of each variant's GEI effect with $\mathbf{v} \sim N(\mathbf{0}, \mathbf{I}_q \frac{\sigma_{\text{GEI}}^2}{q})$. Note that \mathbf{E} is also included as one of the covariates \mathbf{X}_c to adjust for the environment variable and top 10 PCs are included as covariates to account for population stratification. For simplicity of the model, we assume variants' main effects \mathbf{u} and GEI effects \mathbf{v} are independent. The variance-covariance matrix of \mathbf{y} is $\mathbf{V} = \text{var}(\mathbf{y}) = \mathbf{W} \frac{\sigma_{\text{main}}^2}{q} \mathbf{W}^T + \mathbf{D}\mathbf{W} \frac{\sigma_{\text{GEI}}^2}{q} \mathbf{W}^T \mathbf{D}^T + \mathbf{I}_n\sigma_\epsilon^2 = \mathbf{K}\sigma_{\text{main}}^2 + \mathbf{D}\mathbf{K}\mathbf{D}^T\sigma_{\text{GEI}}^2 + \mathbf{I}_n\sigma_\epsilon^2 = \mathbf{K}_\rho\sigma_g^2 + \mathbf{I}_n\sigma_\epsilon^2$, where $\mathbf{K}_\rho = \rho\mathbf{K} + (1 - \rho)\mathbf{D}\mathbf{K}\mathbf{D}$; $\rho = \frac{\sigma_{\text{main}}^2}{\sigma_g^2}$; $\sigma_g^2 = \sigma_{\text{main}}^2 + \sigma_{\text{GEI}}^2$. If the kinship matrix \mathbf{K} is unavailable, we use a sparse genetic relationship matrix (sparse GRM) as the kinship matrix to compute the matrix \mathbf{K}_ρ (Supplementary Methods Section 1). The matrix \mathbf{K}_ρ is utilized to account for genome-wide variants' genetic effects and sample relatedness, like the kinship matrix's role in LMM. The weight parameter ρ plays a trade-off role between adjustment for genome-wide genetic main and GEI effects. The extended model (equation 2) can be equivalently written as,

$$\mathbf{y} = \mathbf{G}\beta_G + (\mathbf{G} \circ \mathbf{E})\beta_{\text{GEI}} + \mathbf{X}_c\beta_c + \mathbf{g}_\rho + \epsilon, \quad (3)$$

where \mathbf{g}_ρ is an $n \times 1$ vector of the combined genetic main and GEI effects with $\mathbf{g}_\rho \sim N(\mathbf{0}, \mathbf{K}_\rho\sigma_g^2)$. The variance components (σ_g^2 and σ_ϵ^2) and the weight parameter ρ can be estimated under the null hypothesis $\beta_G = \beta_{\text{GEI}} = 0$ by extending the grid-search-based REML algorithm implemented in fastGWA (Supplementary Methods Section 2).

To adjust genome-wide genetic effects, we first predict the total genetic effects and then construct residualized phenotype by removing the predicted genetic effects. The total genetic effects are predicted using the best linear unbiased predictor (BLUP) [9] $\hat{\mathbf{g}}_\rho = \mathbf{K}_\rho\hat{\sigma}_g^2\mathbf{V}^{-1}\mathbf{y}_{\text{adj}}$, where $\mathbf{y}_{\text{adj}} = \mathbf{y} - \mathbf{X}_c(\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{y}$ and \mathbf{V}^{-1} is computed based on the estimated variance components and the weight parameter. The residualized phenotype is $\mathbf{y}_{\text{resid}} = \mathbf{y}_{\text{adj}} - \hat{\mathbf{g}}_\rho = \hat{\sigma}_\epsilon^2\mathbf{V}^{-1}\mathbf{y}_{\text{adj}}$, which is also the BLUP of the residuals ϵ .

We then perform the hypothesis testing for the variant being tested $H_0: \beta_G = \beta_{\text{GEI}} = 0$ versus $H_1: \beta_G \neq 0$ or $\beta_{\text{GEI}} \neq 0$ by fitting a linear regression (LR) $\mathbf{y}_{\text{resid}} = \mathbf{G}\beta_G + (\mathbf{G} \circ \mathbf{E})\beta_{\text{GEI}} + \epsilon$, where ϵ is a vector of residuals that has mean of zero and variance-covariance matrix $\mathbf{\Omega}$. Ordinary least square is used to estimate β_G and β_{GEI} . More specifically, we use the sandwich approach (also known as Huber-White, or model-robust method) that has been adopted in GEI studies [10] to estimate the variance-covariance matrix of $(\hat{\beta}_G, \hat{\beta}_{\text{GEI}})^T$ since testing GEI effects is known to have inflation or deflation in P-values if the GEI effects are mis-specified in the model [10, 11]. fastGWA-GE provides three P-values using Wald test with sandwich correction: P-value for testing $\beta_G = 0$; P-value for testing $\beta_G = 0$ after adjusting for GEI effects; and P-value for testing $\beta_{\text{GEI}} = 0$ after adjusting for genetic main effects (Supplementary Methods Section 3).

Results

False positive rate and statistical power in simulations

We run extensive simulations to compare fastGWA-GE with alternative methods including LR implemented in the fastGWA-GE software, LR using sandwich standard errors, GENESIS and

Table 1. Specified proportion of variance explained by each component of the simulated phenotypes. Phenotypes were simulated via a weighted sum of six components: total genetic main effects, total GEI effects, covariates effects, population stratification effects, shared effects among relatives, and residual effects. The weights are the square root of the variance proportion of each component

Scenario	Total genetic main effects	Total GEI effects	Total effects of covariates (including E)	Population stratification	Shared effects among relatives	Residual
Strong main and weak GEI effect	40%	10%	5%	5%	10%	30%
Moderate main and moderate GEI effect	25%	25%	5%	5%	10%	30%
Weak main and strong GEI effect	10%	40%	5%	5%	10%	30%

LEMMA by quantifying the genomic inflation factor (median λ), false positive rate (FPR) and statistical power (Supplementary Methods Section 4, Figure S1). Specifically, eight methods were compared in the simulations: fastGWA-GE, fastGWA-GE-NoSandwich (fastGWA-GE without using sandwich standard errors), LR-All (simple LR method for LR analysis including all relatives), LR-Sandwich-All (LR-All using sandwich standard errors), LR-unRel (LR analysis restricted to unrelated individuals), LR-Sandwich-unRel (LR-unRel using sandwich standard errors), GENESIS and LEMMA. We did not compare with GEM in the simulations because GEM is expected to perform similarly to LR using sandwich standard errors when there is only one environment variable in the model; and we did not compare with structLMM since it was shown to be less powerful than LEMMA [5].

To generate a dataset with population stratification between two ancestries and different levels of sample relatedness, we simulated 100 000 individuals based on a subset of the UK Biobank data by following fastGWA [6] (Supplementary Methods Section 5, Figure S2). To simulate total genetic main effects and total GEI effects, we randomly selected three sets of causal variants from all the variants on the odd chromosomes: (1) variants with genetic main effects only; (2) variants with GEI effects only; (3) variants with both genetic main effects and GEI effects. Variants on the even chromosomes are treated as null variants to evaluate the median λ and FPR. The phenotypes were simulated based on six components: total genetic main effects, total GEI effects, covariates effects, population stratification effects, shared effects among relatives, and residual effects (Supplementary Methods Section 6). We simulated three scenarios representing the three different strengths of genetic main effects and GEI effects of all the causal variants on the odd chromosomes (Table 1).

When testing GEI effects, the simulation results showed that the test statistics of null variants from the GENESIS, LR-All and LR-unRel methods were severely inflated under all simulation scenarios (Figure 1A), while the fastGWA-GE-NoSandwich (in which Sandwich approach was not used) was severely deflated. This is because testing GEI effects is known to have inflation or deflation if model-based standard errors [11] are used. The test statistics of null variants from LR-Sandwich-All and LR-Sandwich-unRel were also inflated and the inflation became more severe when the GEI effects became stronger. The LR-Sandwich-All was more inflated than LR-Sandwich-unRel, which is expected because removing relatives helps control the genomic inflation. In contrast, LEMMA and fastGWA-GE control the inflation relatively well, demonstrating the robustness of fastGWA-GE under different strengths of GEI effects. Moreover, we also evaluated the FPRs at P -value < 0.05 and these

approaches' FPRs were consistent with their genomic inflation factors (Figure 1B). For the statistical power, we found that fastGWA-GE and fastGWA-GE-NoSandwich showed the highest power among all the methods (Figure 1C) and the power improvement continuously increased as the GEI effect became stronger and stronger. The LR approaches (LR-unRel, LR-Sandwich-unRel, LR-All and LR-Sandwich-All) performed similarly to each other after adjusting for their genomic inflation factors.

Our proposed method was robust for testing genetic main effects under different scenarios and had comparable testing power compared with alternative methods, although it was slightly deflated (Supplementary Results Section 1, Figure S3). As for testing G + GxE joint effects, fastGWA-GE was the only method that controls the inflation relatively well and outperformed all other methods except for fastGWA-GE-NoSandwich which was severely deflated (Supplementary Results Section 1, Figure S4). We also conducted four sensitivity analyses to test robustness of our proposed method from four perspectives (Supplementary Methods Section 7): (1) different levels of shared effects among relatives (Table S2, Figures S5–S7); (2) different numbers of causal variants (Table S3, Figures S8–S10); (3) different numbers of grid sizes for parameter ρ in the grid-search-based algorithm (Figures S11–S13); and (4) binary and non-normally distributed phenotypes (Table S4).

Proportion of phenotypic variance explained by total GEI effects in simulations

We observed that the estimated variance component for GEI effects $\hat{\sigma}_{GEI}^2$ (calculated as $(1 - \hat{\rho})\hat{\sigma}_g^2$) in simulations was generally consistent with the true proportion of phenotypic variance explained by total GEI effects in the three simulation scenarios (Figure 1D). In fastGWA-GE, one reason for its well-controlled genomic inflation while testing GEI effects was that fastGWA-GE considered genome-wide GEI effects in the model and adjusted for these effects when testing a single variant's GEI effect. Well capturing genome-wide GEI effects greatly contributed to the validity of the single variant GEI test.

Application to UK Biobank GWAS data

We applied fastGWA-GE to performing GWAS analysis of 7 266 032 variants with $MAF \geq 0.01$ in a subset of individuals of UKB European ancestry ($n = 452\ 249$) [12] (Supplementary Methods Section 8). We analyzed 65 trait-environment combinations from 13 quantitative traits and five environmental covariates as conducted by Wang et al. [13]. These 13 traits are HT (standing height), FVC (forced vital capacity), FEV1 (forced expiratory volume in 1 s), FFR (FEV1 and FVC ratio), BMD (heel bone mineral density T-score,

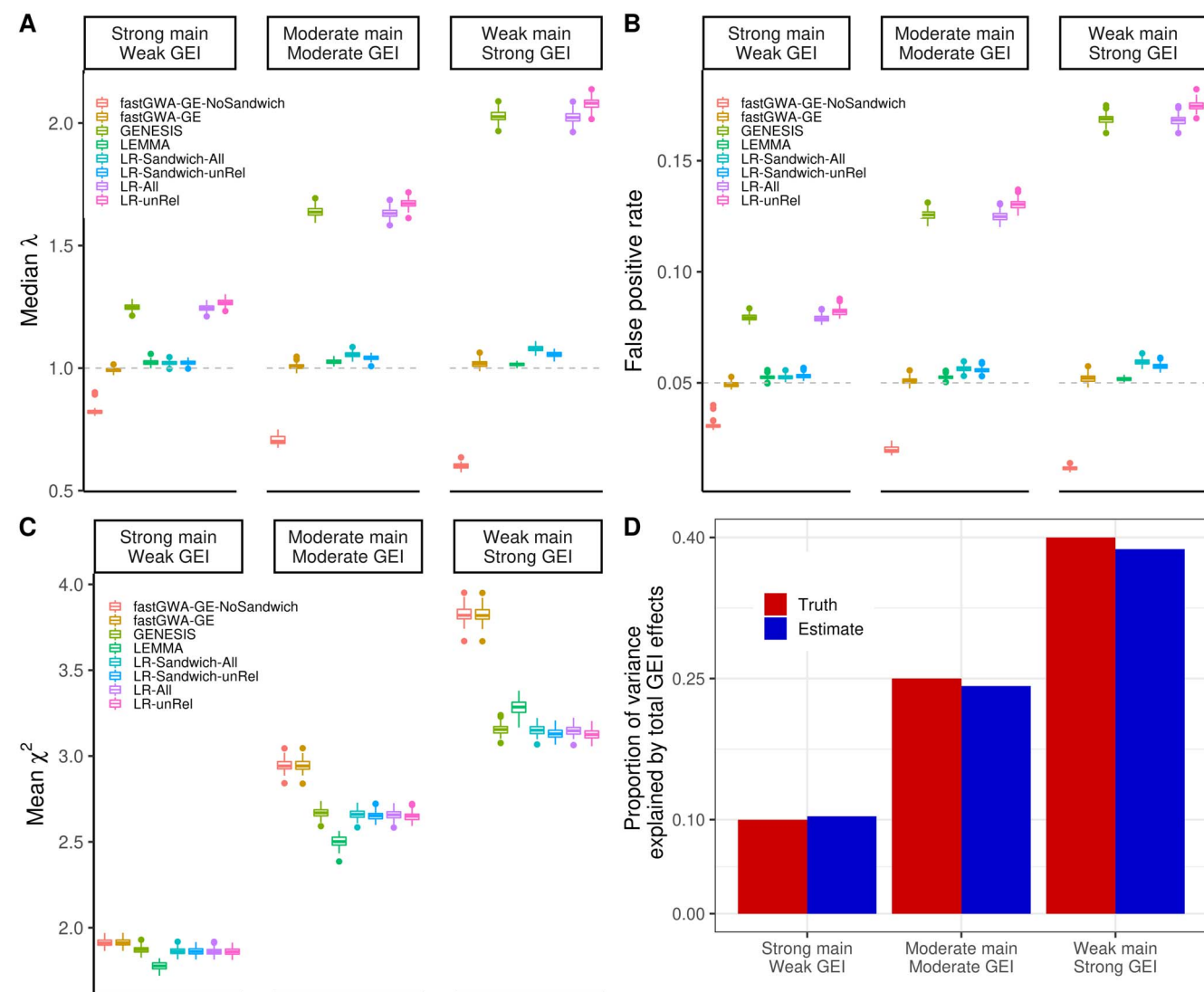


Figure 1. Genomic inflation factor (A), false positive rate (B) and statistical power (C) of testing GEI effects under different simulation scenarios. (A) The y-axis represents the genomic inflation factor (median λ) of the null variants (i.e. all variants on the even chromosomes), and the x-axis are three scenarios representing the three different strengths of genetic main effects and GEI effects of all the causal variants on the odd chromosomes as described in Table 1. Here ‘Strong main, Weak GEI’ refers to strong genetic main effects and weak GEI effects; ‘Moderate main, Moderate GEI’ refers to moderate genetic main effects and moderate GEI effects; ‘Weak main, Strong GEI’ refers to weak genetic main effects and strong GEI effects. (B) The y-axis represents the false positive rate of the null variants. (C) The y-axis represents the mean χ^2 values for all the causal variants on the odd chromosomes. Each boxplot represents the distribution of the median λ , false positive rate or mean χ^2 values across 100 simulation replicates. (D) Comparison between estimated variance component of GEI effects and true proportion of phenotypic variance explained by total GEI effects in the three scenarios. The red bar represents the true variance in simulated datasets. The blue bar represents the average of the estimated variance component for GEI effects across 100 simulation replicates.

automated), BW (birth weight), BMI (body mass index), WC (waist circumference), HC (hip circumference), WHR (waist to hip ratio), WHRadjBMI (WHR adjusted for BMI), BFP (body fat percentage) and BMR (basal metabolic rate) (see the detailed information in Table S5). The five environmental variables are sex, age, PA (physical activity), SB (sedentary behavior) and ever smoking (Table S6). We pre-adjusted these 13 traits for the top 10 PCs, age and sex, and further converted pre-adjusted phenotypes to z-scores by using the rank-based inverse-normal transformation (Supplementary Methods Section 9).

Testing genetic main effects via fastGWA-GE was generally inflated (i.e. the genomic inflation factor (median λ) > 1, Figure S14) due to the strong polygenic effects for these traits, which was consistent with the inflated fastGWA results of testing genetic effects without modeling GEI effects [6]. The genomic

inflation factors (median λ) from the fastGWA analysis results for these 13 traits were summarized in Table S7. However, fastGWA-GE controlled the inflation well for testing GEI effects in most scenarios (Figure 2A). Testing G + GxE joint effects via fastGWA-GE was less inflated than testing the main effects (Figure 2B); and the inflation was mainly driven by the larger inflation of testing genetic main effects. With an experiment-wise significance threshold of 1×10^{-9} (Supplementary Methods Section 10), we identified 96 top hits (independent at linkage disequilibrium (LD) $r^2 < 0.01$ within each trait-environment pair) in 57 loci with significant GEI effects across the 65 trait-environment combinations (Figure 2D, Table S8, Figure S15, Supplementary Methods Section 11).

Out of the 96 top hits, 36 and 50 loci were identified from genotype-sex interactions influencing WHR and WHRadjBMI,

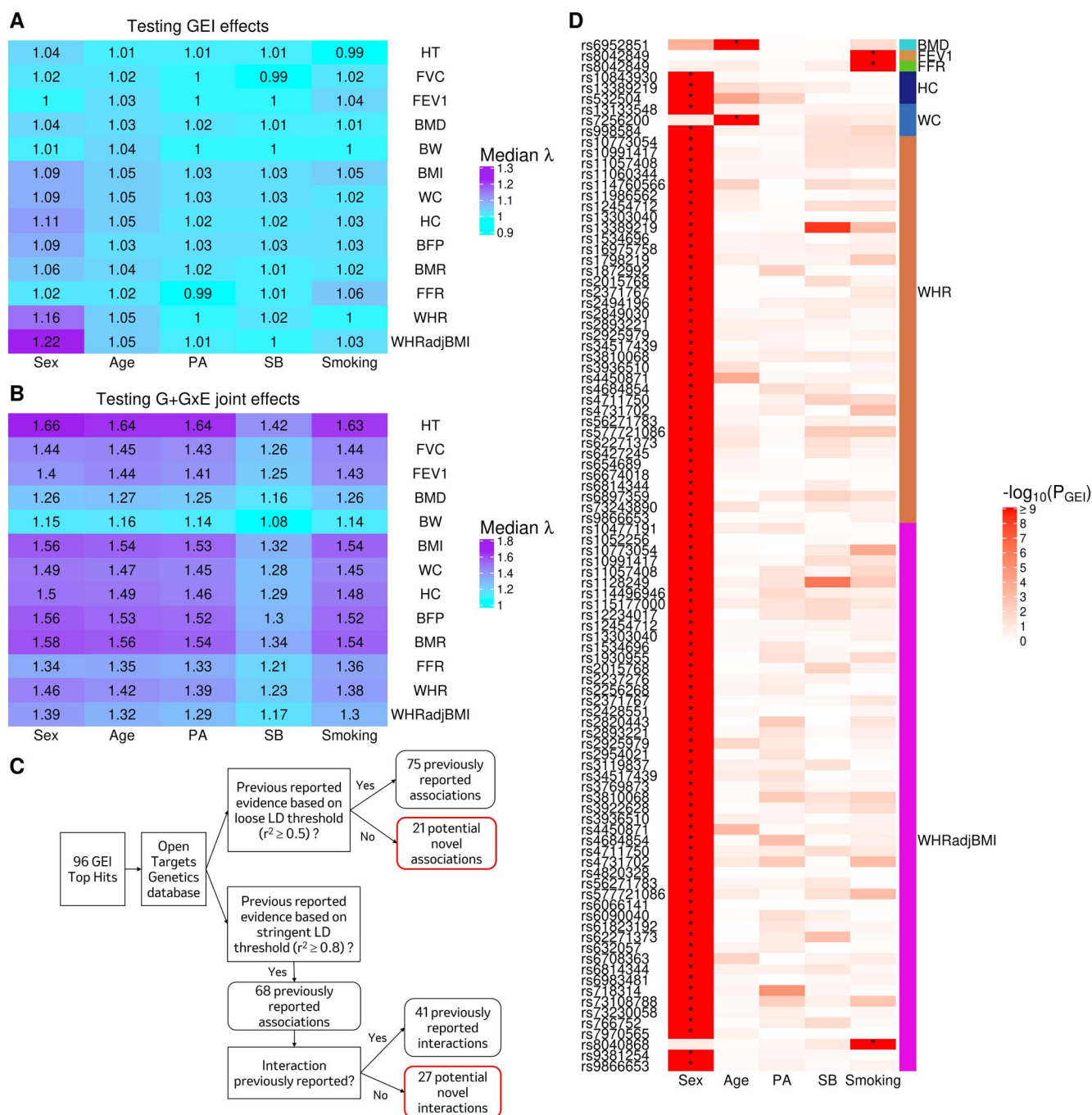


Figure 2. fastGWA-GE analysis of UK Biobank GWAS data. **(A)** Genomic inflation factor (median λ) of testing GEI effects for the 65 genome-wide GEI analyses with 13 traits (HT, FVC, FEV1, FFR, BMD, BW, BMI, WC, HC, WHR, WHRadjBMI, BFP, BMR) and five environment variables (sex, age, physical activity (PA), sedentary behavior (SB) and smoking). **(B)** Genomic inflation factor (median λ) of testing G + GxE joint effects for the 65 genome-wide GEI analyses. **(C)** Workflow of identification of the novel signals via comparison with prior GWAS studies using the Open Targets Platform. **(D)** Heatmap of GEI P -values for the 96 top hits from the 65 genome-wide GEI analyses. '*' denotes the significant GEI effects ($P < 1 \times 10^{-9}$).

respectively. As expected, given high correlation between these traits (Figure S16), 34 of these loci were identified for both traits. For the genome-wide fastGWA-GE analysis of genotype-sex interactions for WHRadjBMI, we identified 299 top hits with P -value less than significance threshold 1×10^{-9} (independent to LD $r^2 < 0.01$ within each trait-environment pair) from the G + GxE joint test. Demonstrating the power of the G + GxE joint test for discovery, 36 of these top hits did not have significant genetic marginal effects with significance threshold 1×10^{-9} in the fastGWA (main genotype effect) analysis of the

trait WHRadjBMI, and 33 did not have significant ($P < 1 \times 10^{-9}$) marginal effects from fastGWA analysis or significant GEI effects from fastGWA-GE analysis. We observed that the majority (about 62%) of the variants had lower P -values of marginal effects than that of G + GxE joint effects (Figure 3C and D). This observation was consistent with the findings in similar analyses of WHR conducted by Westerman *et al.* [2] In addition to the comparison of the joint effect test P -values and the marginal effect test P -values, we also compared the GEI test P -values and the marginal effect test P -values. Of the 50 top hits detected from the GEI test

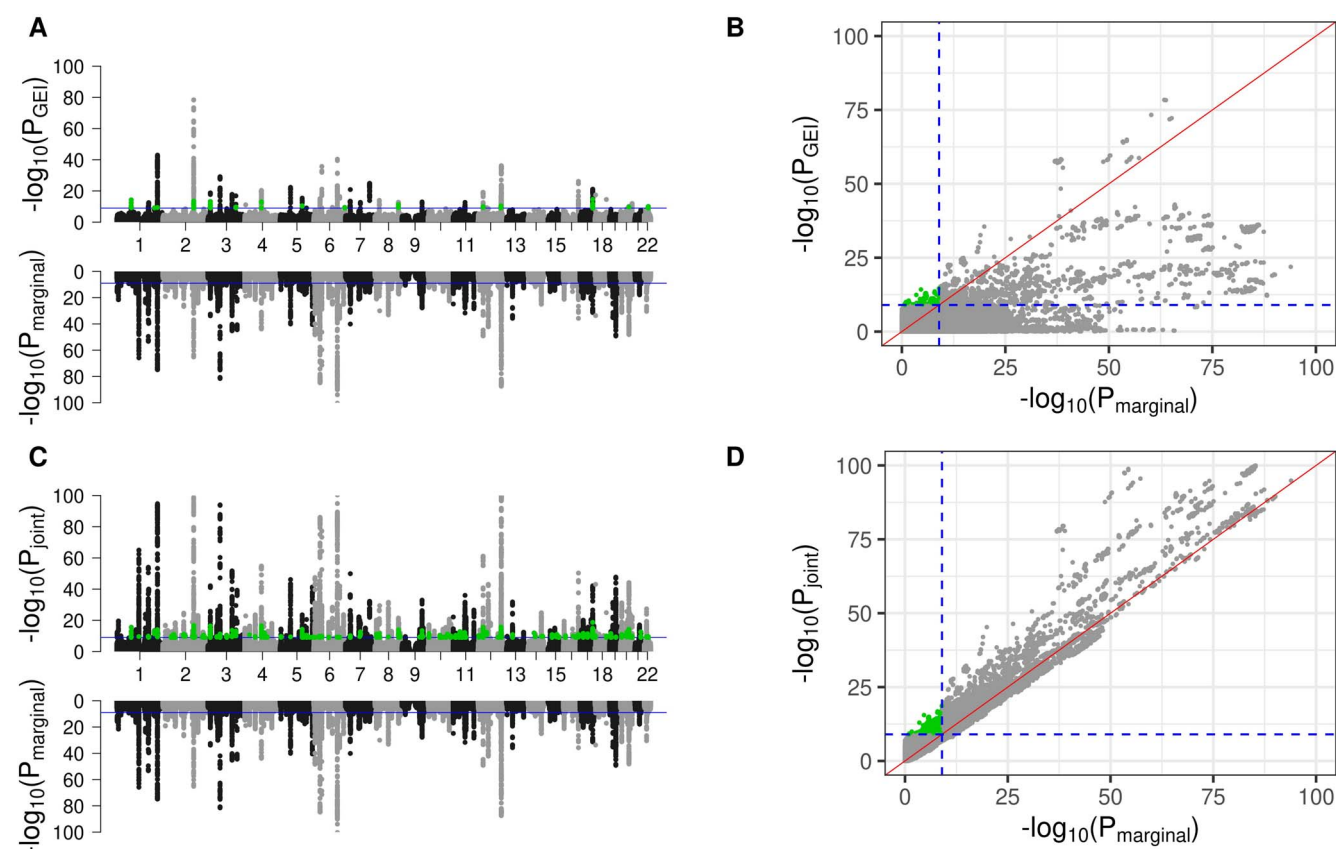


Figure 3. Comparison of genotype-sex GEI test P-values and G + GxE joint test P-values from fastGWA-GE analysis with marginal effect test P-values from fastGWA analysis of WHRadjBMI in the UK Biobank. **(A)** Manhattan plot for the P-values from the GEI test analysis (upper figure) versus marginal genetic effect test P-values (from a model without the GEI term) (lower figure). **(B)** Scatter plot of the GEI test P-values versus the marginal effect test P-values. The green dots were the variants with significant GEI effects but not significant marginal effects. **(C)** Manhattan plot of the P-values from the fastGWA-GE G + GxE joint test (upper figure) versus marginal genetic effects (from a model without the GEI term) from fastGWA analysis (lower figure). **(D)** Scatter plot of the G + GxE joint test P-values versus the marginal effects test P-values. The green dots were the variants with significant G + GxE joint effects but not significant marginal effects. The blue dashed line is the significance threshold $-\log_{10}(1 \times 10^{-9})$. The variants with P-values less than 1×10^{-100} were excluded for visualization purposes.

for genotype-sex interactions influencing the trait WHRadjBMI, six did not have significant marginal effects. Although 55% of the variants had P-values of marginal effects less than that of GEI effects, many variants showed much smaller P-values of GEI effects than that of marginal effects (i.e. the green points in the Figure 3A and B).

We compared our 96 top hits with prior GWAS studies using the Open Targets Genetics database [14] (Figure 2C, Tables S8 and S9). The associations between the tested loci and traits had previously been reported for 68 of the 96 top hits using a more stringent LD threshold ($r^2 \geq 0.8$), and for 75 of the 96 top hits at a more permissive threshold ($r^2 \geq 0.5$) (Supplementary Methods Section 12). The remaining 21 hits identified may represent novel genetic associations; of these 21, the P-value for testing the marginal genetic effects from fastGWA-GE was $<1 \times 10^{-9}$ for 11 loci. Of the 68 locus-trait pairs for which main genetic effects had been reported previously, the interaction with sex, age or smoking observed in our fastGWA-GE GEI analysis had been reported previously at 41 locus-trait pairs; for the remaining 27 pairs the interactions identified in this study may represent novel interactions.

Most (91 of 96) identified interactions were observed with sex; interestingly, we note that 59 of 91 hits with sex interactions occur at loci also associated with sex hormone binding globulin (SHBG) (and, in some cases, testosterone) levels. Most of the identified

interactions with sex were for WHR and/or WHRadjBMI, and in many of these cases, a larger genetic effect was observed among women, consistent with similar prior analyses [15]. Mendelian randomization analyses conducted by Ruth and colleagues [16] suggested that, in women (but not in men), SHBG levels may drive WHRadjBMI; thus, many of the observed associations with WHRadjBMI may in fact be driven by regulation of SHBG levels in women. In comparing loci at which we observe an interaction with sex to the published sex-QTLs [17], only one locus (with lead SNP rs2237276) was weakly linked ($r^2 = 0.44$, $D' = 0.99$ among European ancestry populations in 1000 Genomes) to a reported sex-QTL at rs4710149, for association with gene RNASET2 in breast mammary tissue. The limited overlap with known sex-QTLs is consistent with similar prior analyses [15, 18], and may reflect limited power for detection of interactions in the relatively small datasets available for expression QTL analyses. Finally, gene-set enrichment analysis (Supplementary Methods Section 12) of the 47 unique candidate causal genes for the WHR and/or WHRadjBMI-sex interactions identified enrichment of pathways related to angiogenesis and calcium signaling (Figure S17), among others. Adipose vasculature plays an important role in development and maintenance of adipose tissue [19], and calcium signaling has been linked to adipogenesis [20, 21]. The enrichment of genes involved in these processes may indicate differences between males and females in

Table 2. Computation time (hours) comparison between fastGWA-GE and software from other existing methods based on simulated data

	fastGWA-GE	LR-Sandwich-All	LEMMA	GENESIS
Strong main and weak GEI	0.08	0.28	115.46	140.46
Medium main and medium GEI	0.12	0.54	89.29	139.4
Weak main and strong GEI	0.12	0.33	60.68	143.51

their role in influencing body fat distribution. However, given the complexities inherent in mapping genetic associations to causal genes, we emphasize caution in interpretation of these gene set enrichment results, which will be highly sensitive to the input gene list used.

An interaction between a locus on chromosome 15 (lead SNP rs8042849) and smoking status was observed for association with FEV1 and FFR. This locus has previously been associated with smoking behavior, particularly the number of cigarettes smoked per day [22], and lung function and disorders among smokers [23]. Consistent with prior reports [23], the association between this locus and FEV1 and FFR was specific to smokers, and it is possible that the observed relationship is driven by variability in cigarette smoking quantity among smokers. An interaction between smoking status and another linked locus in the region (lead SNP rs8040868) for association with WHRadjBMI was also observed. A similar association with BMI at this locus has been reported [24, 25]; in contrast to the results in this study, in prior reports the interaction with smoking in the association with WHR was attenuated by adjustment for BMI. In this analysis, as in prior reports, an association between the locus and trait of interest was also observed among non-smokers, suggesting an impact of the causal gene at this locus on obesity related traits or central adiposity both independent of and modified by smoking.

An interaction with age was observed for association with waist circumference with some prior evidence supporting this observation (Supplementary Results Section 2). We did not observe any interactions with PA or SB or any interactions for association with BMI, BW, BFP, BMR, FVC or HT. Only a handful of environmental variables were tested in this analysis, so the lack of detected interactions is not unexpected, especially for standing height, which would be more likely to be affected by environmental exposures in childhood than adult behaviors. The lack of any interactions for PA or SB in association with metabolic traits suggests that such interactions, if they do exist, may be weaker interactions that we were not powered to detect.

We further compared the real data analysis results of fastGWA-GE with alternative methods. However, because of the severe inflation for the GENESIS, LR (without Sandwich correction) and extremely demanding memory required for the LEMMA software, it was not feasible for us to conduct the analyses of UK Biobank GWAS data for GENESIS and LR (without Sandwich correction) and LEMMA. We ran LR-Sandwich-All for all the 13 traits and five environment variables from the UKBB GWAS data in the same way as what we did for the fastGWA-GE method and further compared the results of GEI associations between these two methods. For the traits HT, FVC, BW, BMI, BFP and BMR, neither of these two methods detected any significant (GEI P -value $< 1 \times 10^{-9}$) variants. For the traits BMD and WC, both methods detected the same variants with significant GEI effects. For the traits FEV1 (Figure S18), HC (Figure S19), FFR (Figure S20), WHR (Figure S21) and WHRadjBMI (Figure S22), fastGWA-GE detected more significant variants than LR-Sandwich-All, except for FEV1 trait's genotype-by-smoking interaction. For example, for the genotype-by-sex interactions for trait WHRadjBMI, fastGWA-GE identified

41 significant variants missed by LR-Sandwich-All, while LR-Sandwich-All detected 20 variants missed by fastGWA-GE.

Computation time

Computation time was evaluated based on the simulated data used in the main simulation. To successfully apply the methods to the simulated data, we used different computation settings for different methods. fastGWA-GE method (built on top of fastGWA) was hundreds of times more computationally efficient than alternative approaches (Table 2). More details are available in Supplementary Results Section 3.

Discussion

In this article, we propose a fast and powerful LMM-based approach, fastGWA-GE, by extending the fastGWA [6] method from testing only the genetic main effect to testing both GEI and G + GxE joint effects while taking advantage of its highly computational efficiency. Our extensive simulations show that fastGWA-GE controls inflation well and is statistically more powerful than alternative methods including LR, LEMMA and GENESIS. Another good feature of fastGWA-GE is that it does not rely on the GRAMMAR-GAMMA approximation [26] implemented in the fastGWA method, which needs to sample a set of null variants to approximately compute the summary statistics for all variants. We envision fastGWA-GE as a powerful method and tool to detect GEI effect related signals in large-scale disease genetic studies.

One limitation of our proposed method is that we adopt the infinitesimal model which assumes every variant has a small but non-zero effect on a given trait. This may not be correct when there are non-causal variants in the model, and in this case a mixture distribution may be more appropriate [5]. It is reassuring to observe well-controlled inflation and protected FPR from our simulation studies, which included a large number of noncausal variants for constructing GRM. More properly modeling the effects of genetic variants may further increase the statistical power under the alternative hypotheses. Another limitation of fastGWA-GE is that the variance components for the genetic main effects and GEI effects are estimated using a grid-search-based approach, which may not always be the most optimal solution. A more accurate method may further increase the power of detecting GEI effects.

Our proposed method can be extended in several directions. Although it can be naively applied to analyzing binary traits, it may be further improved to handle non-Gaussian traits by extending the fastGWA-GLMM method [27], which should further increase the statistical power of testing the GEI effect or G + GxE effect for the analysis of binary traits. In addition, fastGWA-GE can be also extended to handle multiple environmental variables by constructing an environmental score, a linear combination of multiple environmental variables similar as that implemented in the LEMMA method. Moreover, fastGWA-GE may be also extended to handle a more complex relationships between phenotypes such as the correlation between multiple traits and

correlation between covariates and traits, under the framework of the multivariate reaction norm model [28, 29]. Furthermore, our proposed fastGWA-GE method can be directly applicable to genotype-by-treatment interaction (GTI or GxT) test or genotype and genotype-by-treatment (G + GxT) joint test to detect predictive genetic biomarkers (i.e. with strong GTI effect) in pharmacogenomics (PGx) GWAS studies. In PGx data from randomized clinical trials, the treatment assignment information is usually binary (i.e. treatment arm versus control arm) or categorical (i.e. multiple treatment arms versus control arm). The effect sizes from drug response predictors from randomized clinical trials are usually larger than those from disease traits in disease genetic cohorts. Thus, fastGWA-GE is also useful in detecting predictive genetic biomarkers for drug response prediction and patient stratification in PGx studies.

Key Points

- We propose a linear mixed model-based method fastGWA-GE for single variant genotype-by-environment interaction test and G + GxE joint test in large-scale GWAS analysis.
- fastGWA-GE uses principal components to account for population stratification, adopts a weighted average of sparse genetic relationship matrix and the product of sparse GRM and the environment variable to account for sample relatedness, implements model-robust (sandwich) standard errors and conducts whole-genome regression analysis.
- fastGWA-GE outperforms other existing GEI test methods by controlling genomic inflation better, providing larger power and running hundreds to thousands of times faster.
- Application of fastGWA-GE to UK Biobank GWAS data identifies 96 significant signals (72 variants across 57 loci) with GEI test P-values $< 1 \times 10^{-9}$, including 27 novel GEI associations.
- fastGWA-GE is demonstrated to be an effective tool for GEI signal discovery in large-scale GWAS.

Data and materials availability

Genotype data were processed using PLINK v1.90 and v2.00: <https://www.cog-genomics.org/plink/1.9/> and <https://www.cog-genomics.org/plink/2.0/>; figures were generated using ggplot2 R package v3.3.3: <https://cran.r-project.org/package=ggplot2>; Manhattan plots were generated using qqman R package v0.1.4: <https://cran.r-project.org/package=qqman>; LDlinkR R package v1.0.2: <https://cran.r-project.org/package=LDlinkR>; heatmaps were generated using ComplexHeatmap R package v2.2.0: <http://bioconductor.org/packages/ComplexHeatmap/>; Open Targets: <https://genetics.opentargets.org/>; gene set enrichment analysis results were generated using enrichR R package (v3.1): <https://CRAN.R-project.org/package=enrichR>. The individual-level data that support the findings of this study are available upon application to the UK Biobank (<https://www.ukbiobank.ac.uk/register-apply/>). Our method is implemented in the fastGWA-GE software, freely available at <https://github.com/jiayangqt/gcta.git> and the tutorial is available at <https://yanglab.westlake.edu.cn/software/gcta/#fastGWA-GE>.

Supplementary Data

Supplementary data are available online at <https://academic.oup.com/bib>.

Acknowledgments

This study has been conducted using the UK Biobank Resource under Application Number 28967. The authors thank the study participants in the UK Biobank study. The authors would also like to thank Dr Jian Yang, Dr Zhili Zheng and Dr Longda Jiang for sharing the fastGWA source code for us so that we can develop the fastGWA-GE software on top of it and Hailing Fang for merging our fastGWA-GE code as one module in GCTA software.

Funding

The authors acknowledge that they received no funding in support for this research.

References

1. Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects. *Int J Epidemiol* 1984;**13**:356–65.
2. Westerman KE, Pham DT, Hong L, et al. GEM: scalable and flexible gene-environment interaction analysis in millions of samples. *Bioinformatics* 2021;**37**:3514–20.
3. Moore R, Casale FP, Bonder MJ, et al. A linear mixed-model approach to study multivariate gene-environment interactions. *Nat Genet* 2019;**51**:180–6.
4. Gogarten SM, Sofer T, Chen H, et al. Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* 2019;**35**:5346–8.
5. Kerin M, Marchini J. Inferring gene-by-environment interactions with a Bayesian whole-genome regression model. *Am J Hum Genet* 2020;**107**:698–713.
6. Jiang L, Zheng Z, Qi T, et al. A resource-efficient tool for mixed model association analysis of large-scale data. *Nat Genet* 2019;**51**:1749–55.
7. Eu-Ahsunthornwattana J, Miller EN, Fakiola M, et al. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genet* 2014;**10**:e1004445.
8. Yang J, Lee SH, Goddard ME, et al. GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011;**88**:76–82.
9. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc* 1993;**88**:9.
10. Almli LM, Duncan R, Feng H, et al. Correcting systematic inflation in genetic association tests that consider interaction effects: application to a genome-wide association study of posttraumatic stress disorder. *JAMA Psychiat* 2014;**71**:1392–9.
11. Voorman A, Lumley T, McKnight B, et al. Behavior of QQ-plots and genomic control in studies of gene-environment interaction. *PLoS One* 2011;**6**:e19416.
12. Consortium 1000 Genomes Project, others. A map of human genome variation from population scale sequencing. *Nature* 2010;**467**:1061.
13. Wang H, Zhang F, Zeng J, et al. Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. *Sci Adv* 2019;**5**:eaaw3538.
14. Carvalho-Silva D, Pierleoni A, Pignatelli M, et al. Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res* 2019;**47**:D1056–65.

15. Shungin D, Winkler TW, Croteau-Chonka DC, et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature* 2015;**518**:187–96.
16. Ruth KS, Day FR, Tyrrell J, et al. Using human genetics to understand the disease impacts of testosterone in men and women. *Nat Med* 2020;**26**:252–8.
17. Oliva M, Muñoz-Aguirre M, Kim-Hellmuth S, et al. The impact of sex on gene expression across human tissues. *Science* 2020;**80**, 369:eaba3066.
18. Winkler TW, Justice AE, Graff M, et al. The influence of age and sex on genetic associations with adult body size and shape: a large-scale genome-wide interaction study. *PLoS Genet* 2015;**11**:e1005378.
19. Cao Y. Angiogenesis and vascular functions in modulation of obesity, adipose metabolism, and insulin sensitivity. *Cell Metab* 2013;**18**:478–89.
20. He Y-H, He Y, Liao X-L, et al. The calcium-sensing receptor promotes adipocyte differentiation and adipogenesis through PPAR γ pathway. *Mol Cell Biochem* 2012;**361**:321–8.
21. Pramme-Steinwachs I, Jastroch M, Ussar S. Extracellular calcium modulates brown adipocyte differentiation and identity. *Sci Rep* 2017;**7**:8888.
22. The Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 2010;**42**:441.
23. Wain LV, Shrine N, Miller S, et al. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir Med* 2015;**3**:769–81.
24. Taylor AE, Morris RW, Fluharty ME, 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**:e1004799.
25. Morris RW, Taylor AE, Fluharty ME, et al. Heavier smoking may lead to a relative increase in waist circumference: evidence for a causal relationship from a Mendelian randomisation meta-analysis. *CARTA Consortium BMJ Open* 2015;**5**:e008808.
26. Svishcheva GR, Axenovich TI, Belonogova NM, et al. Rapid variance components-based method for whole-genome association analysis. *Nat Genet* 2012;**44**:1166–70.
27. Jiang L, Zheng Z, Fang H, et al. A generalized linear mixed model association tool for biobank-scale data. *Nat Genet* 2021;**53**:1616–21.
28. Ni G, van der Werf J, Zhou X, et al. Genotype-covariate correlation and interaction disentangled by a whole-genome multivariate reaction norm model. *Nat Commun* 2019;**10**:2239.
29. Lee SH, van der Werf JHJ. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. *Bioinformatics* 2016;**32**:1420–2.