SCEBE: An Efficient and Scalable Algorithm for Genome-wide Association Studies on Longitudinal Outcomes with Mixed-Effects Modeling

Min Yuan[1*], Xu Steven Xu[2*], Yaning Yang[3], Yinsheng Zhou[3], Yi Li[3], Jinfeng Xu[4], Jose Pinheiro[5], for the Alzheimer's Disease Neuroimaging Initiative

Corresponding authors. Yaning Yang, Department of Statistics and Finance, University of Science and Technology of China, Hefei 230026 Anhui, China. Tel: 86+13655692600; Email: ynyang@ustc.edu.cn or Steven Xu, Genmab US, Inc, Princeton, NJ 08540, USA. Email: sxu@genmab.com. *Yuan and Xu contributes equally to this work.

Abstract

Genome-wide association studies (GWAS) using longitudinal phenotypes collected over time is appealing due to the improvement of power. However, computation burden has been a challenge because of the complex algorithms for modeling the longitudinal data. Approximation methods based on Empirical Bayesian Estimates (EBEs) from mixed-effects modeling have been developed to expedite the analysis. However, our analysis demonstrated that bias in both association test and estimation for the existing EBE-based methods remains an issue. We propose an incredibly fast and unbiased method (SCEBE, simultaneous correction for EBE) that can correct the bias in the naive EBE approach, and provide unbiased p-values and estimates of effect size. Through application to ADNI data with 6,414,695 single nucleotide polymorphisms, we demonstrated that SCEBE can efficiently perform large-scale GWAS with longitudinal outcomes, providing nearly 10,000 times improvement of computational efficiency and shortening the computation time from months to minutes. The SCEBE package and the example datasets are available at https://github.com/Myuan2019/SCEBE

Key Points

- Modeling GWAS data on longitudinal outcome using mixed-effects model can improve statistical power, however, computational complexity and efficiency remain difficult and challenging.
- SCEBE provides almost identical estimation and p-values compared to the standard likelihood based approach.
- SCEBE provides nearly 10,000 times improvement of computational efficiency and shortens the computation time from months to minutes.

Keywords

Genome-wide association studies; Longitudinal outcomes; Mixed-effects model; Empirical Bayesian estimates; Shrinkage

Min Yuan is Associate Professor in biostatistics in the Center for Big data in Public Health, Anhui Medical University. Her interest is developing statistical models in genomic research.

Xu Steven Xu is Data Scientist in Genmab US, Inc. His interest is developing statistical models

For Peer Review

and quantitative analysis in clinical trials.

Yinsheng Zhou and Yi Li are PhD students in biostatistics in the Department of Statistics and Finance, University of Science and Technology of China. Their interest is developing computational algorithm in haplotype-based research.

Jinfeng Xu is Associate Professor in Department of Statistics and Actuarial Science, University of Hong Kong. His interest is Statistical modeling with longitudinal data.

Jose Pinheiro is Statistician in Janssen Research and Development LLC Raritan. His interest is Statistical modeling with longitudinal data.

Yaning Yang is Professor in biostatistics in University of Science and Technology of China. His interest is genetic statistics.

Introduction

Genome-wide association studies (GWAS) with longitudinal outcomes allow higher statistical power to detect genetic variants with relatively weak effects [1-2], better identification patient populations, and better understanding of mechanisms of disease resistance and disease progression [3] etc. Mixed-effects model is a powerful and popular tool to model repeated measurements [4]. However, computation burden become challenging for such model as millions of single nucleotide polymorphisms (SNPs) are evaluated in GWAS. Currently, the most commonly used algorithm for testing association is either the Wald test or the likelihood ratio test [3-4]. In addition, local convergence may lead to biased parameter estimation and p-values for mixed-effects models.

Empirical Bayes Estimates (EBEs), derived from the base mixed-effects model without covariates has long been used as an ad hoc approach to facilitate variable selection for low-dimension data [5-6]. Efforts were made to utilize EBE-based approach (thereafter referred as naïve EBE [NEBE]) to test association in GWAS [7-8] with longitudinal outcomes. Despite of its simplicity, it is well known that the EBEs are biased as they tend to be shrunk to the corresponding population mean [6, 9], and may not be suitable for identification of significant variables [9]. Therefore, there is an urgent need to develop an

For Peer Review

efficient and scalable algorithm to compute unbiased association test statistics for GWAS with longitudinal outcomes.

We propose a novel, high throughtput algorithm to provide an efficient and scalable computation of the association test statistics for GWAS with longitudinal outcomes. This method not only corrects the bias caused by shrinkage, and provides numerically identical estimation and p-values to those from the standard mixed-effects model, but also could be 10,000 times faster than current standard approach.

Methods

Suppose the GWAS is designed from a natural population with three genotypes at each locus. Let m denote the number of individuals and q denote the number of SNPs. The ith individual has $n_i$observations $y_i = (y_{i1}, y_{i2}, \ldots, y_{ini})'$ at time points $t_i = (t_{i1}, t_{i2}, \ldots, t_{ini})'$. A typical linear mixed-effects model in GWAS can be written in a two stage form as follows,

$$y_i = Z_i\beta_i + e_i$$

$$\beta_i = \alpha + x_i\gamma + b_i,\ i = 1, 2, \cdots, m \quad (1)$$

$$e_i \sim N(0, G_i) \text{ and } b_i \sim N(0, R)$$

where β_i is the $p \times 1$ random effect vector. The design matrix Z_i is a $n_i \times p$ matrix. Covariate x_i is the genotype coded as 0, 1 or 2 for three different genotypes. α and γ are p-dimensional intercept and slope parameters. The base model corresponds to model (1) with $\gamma = 0$. Residual $e_i's$ independently folllow a multinormal distribution with mean 0 and a $n_i \times n_i$ covariance matrix G_i which chracterizes the correlation structure of within-subject variablities. b_i is the $p \times 1$ between-subject error vector following a multinormal

For Peer Review

distribution with mean 0 and a $p \times p$ covariance matrix R. R characterizes the between-subject variablilities. The standard approach of fitting model (1) is based on the likelihood function and implemented in R packages (e.g., lme4). We call the standard approach 'LME' in this article.

We propose a simultaneous correction for empirical Bayesian estimator (SCEBE) which can simultaneously correct genetic effects on all random parameters. The SCEBE method contains three steps:

Step 1: Fit a base mixed-effects model without covariates (thereafter referred as base model). In this step, maximum likelihood estimators (MLEs) or restricted maximum likelihood estimators (REMLs) are obtained for the fixed effects, between-subject variability (random effects), and within-subject variability under the base model.

Step 2: Treat the predictors of random effects (i.e., EBEs) from Step 1 as phenotypes for genome-wide association analysis using a standard linear regression model. The resulting SNP effect estimates (and corresponding p-values) are referred as the naive empirical Bayesian estimators (NEBE). The EBEs are the weighted sum of the population and sample mean, thus suffer from the shrinkage to population mean especially when longitudinal samples are sparse or/and within-subject variability is large. The shrunk EBEs tend to produce biased NEBE estimators.

Step 3: Fortunately, the degree of bias can be theoretically quantified and be used as the correction matrix to obtain the unbiased estimators and test statistics. In this step, we correct the NEBE as well as the covariance matrix of NEBE by a derived simultaneous correction matrix to obtain the unbiased estimates and testing statistics for the SNP effects. The derived correction matrix has the expression as follows

For Peer Review

$$S_c = \frac{\sum_{i=1}^{m}(x_i - \bar{x})\left[x_i(I_p - S_i) + S_i\sum_{i=1}^{m}x_i W_i\right]}{\sum_{i=1}^{m}(x_i - \bar{x})^2}$$

where $\bar{x}$ is the sample mean; I_p is the p-dimensional identity matrix; $S_i = (Z_i' G_i^{-1} Z_i + R^{-1})^{-1} R^{-1}$ is the shrinkage matrix and $W_i = (\sum_{i=1}^{m} Z_i' \Sigma_i^{-1} Z_i)^{-1} Z_i' \Sigma_i^{-1} Z_i$ with $\Sigma_i = Z_i R Z_i' + G_i$ being the covariance matrix of y_i. We proved that the expectation of NEBE under the true model (1) is $S_c\gamma$. Therefore, S_c can be used as the correction factor to correct the bias of NEBE. Details of the derivation of correction matrixes are provided in Supplementary Materials/Section 1.

While this paper was in development, Sikorska et al. also published an alternative, efficient algorithm for genome-wide analysis of longitudinal data (GALLOP) [11]. The main idea of GALLOP is to efficiently solve the Henderson equation by taking consideration of the block diagonal feature of the coefficient matrix of the Henderson equation. In this paper, we also implemented GALLOP in our R package and applied it to the simulation and real data analysis for comparison.

Results

ADNI data analysis

The data was downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.usc.edu/ADNI). The ADNI is an ongoing longitudinal multicenter study aimed at detecting and monitoring the early stage of Alzheimer's disease (AD) by investigating the magnetic resonance imaging, positron emission tomography, genetic, biochemical biomarkers, and neuropsychological and clinical assessment. Since the initial

For Peer Review

phase ADNI-1 was carried out in 2004, the ADNI has been extended to ADNI-2, ADNI-3 and ADNI-GO. There are 784 individuals enrolled in the study and a total 6,528,104 SNPs were sequenced and screened after quality control. In this paper, we used one of the most widely used imputation methods, segmented haplotype estimation and imputation tool (SHAPEIT) [12], to impute missing genotypes. After deleting SNPs with MAF being smaller than 0.05 and SNPs with only one genotype for all individuals, 6,414,695 SNPs were analyzed. We used repeatedly measured Rey Auditory Verbal Learning Test (RAVLT) forgetting scale scores over time as the longitudinal response phenotype, and investigated the SNP effects on the progression rate of RAVLT over time.

The key features of the proposed method SCEBE are time efficiency and accuracy compared to standard LME. We first compared the computation time cost for different approaches using the ADNI data (6,414,695 SNPs) (Figure 1). The computation was performed on an Ubuntu 16.04 LTS running on a server with CPU@2.9G and 8G RAM. It required approximately 145 days (single-CPU time) for LME to scan through all the SNPs, while only 2 min， 37 min and 118 min were needed for NEBE, SCEBE and GALLOP respectively (Figure 1a). Therefore, SCEBE approach was nearly 10,000 times faster than LME (Figure 1b).

The SCEBE also provide unbiased estimates and similar p-values compared to classical LME (Figure 2). In contrast, as expected, the estimates of effect size based on NEBE approach had marked biases (Figure 2b). Due to the shrinkage, the estimated effect of the SNPs on the disease progression (slope) based on NEBE was close to zero despite that the underlying genetic effects based on LME were apparent for many SNPs (Figure 2b). Furthermore, the p-values from the intermediate biased NEBE are obviously different

For Peer Review

from those of the standard LME (Figure 2a). SCEBE corrected the bias in estimation and p-values from NEBE and provided very similar p-values as the standard LME (Figure 2a). In comparison, GALLOP and SCEBE shared very similar p-values for association tests and estimation of SNP effects for the ADNI data. Nevertheless, the SCEBE was 3 – 4 times faster than GALLOP (Figure 1a and 1b).

Manhattan plot based on SCEBE for the ADNI data is presented in Figure 3. A closer look at the top 20 SNPs for both baseline disease status (intercept) and disease progression (slope) is displayed in Figure 4. Four out of the top 20 SNPs for the baseline AD status are related to genes which have been reported to be associated with AD ([13, 14]). Among them, rs429358 is within APOE, rs12721051 is within APOC1, rs4420638, rs56131196 are 500B downstream variants of APOC1.It is well known that APOE4 is involved in the pathogenesis of both late-onset familial and sporadic AD [13]. In addition, recent literature suggested that immunosuppression associated with APOC1 in the context of Aβ innate immune activation is potentially clinically relevant [14].

In addition, among the top 20 SNPs for disease progression according to RAVLT scores, rs3799160 is within PDE10A, which has been reported to be related to AD in recent literatures [15-16]. It was discovered that most PDE isoforms (including PDE10A) are expressed in the brain, and PDE inhibitors are capable to improve memory performance in different animal models of AD [15]. Additionally, expression of PDE10A was found to be upregulated after long term potentiation induction in the hippocampus of awake adult rats [16], indicating that it may have effects on memory and cognition.

Since very few GWAS association studies have been reported using RAVLT scores over time, the other SNPs identified in this study (Supplementary Table 1) may provide

For Peer Review

new insights for biology of AD and its disease progression. Further investigations are warranted in the future to better understand the biology of these SNPs.

Simulation studies

Association test

We also use extensive simulations to compare the standard LME with the NEBE, SCEBE and GALLOP approaches. Briefly, m=100, 500, 1000, or 10000 subjects were simulated for a given scenario. Two unbalanced sampling schemes, sparse (1, 2, 3, or 5 samples per subject over time) and intensive (3, 5, 7, or 9 samples per subject over time) sampling, were implemented in the simulations. Assuming that the allele frequency of risk allele p_A is randomly sampled from a uniform distribution $\mathrm{U}(0.05,0.5)$ and Hardy-Weinberg equilibrium holds in population, the probabilities of three genotypes are p_A^2, $2p_A(1-p_A)$, $(1-p_A)^2$ respectively. 100, 1000, 10000 SNPs are independently sampled from a multinomial distribution with probability $(p_A^2, 2p_A(1-p_A),(1-p_A)^2)'$. We assumed that no effects of SNPs were on baseline disease status (intercept), while the effect sizes of SNPs on disease progression (slope) were randomly sampled from a uniform distribution $\mathrm{U}(0,0.5)$. The between-subject covariance was assumed diagonal with all elements were set to 1, while the within-subject covariance was also assumed diagonal, which was set to 0.5, 1, 2, or 3 to allow different levels of shrinkage. In total, 96 scenarios were simulated and each was done for 1000 replicates.

For the association test, although the p-values calculated based on NEBE appear to be trending the same way as those based on the LME approach, the discrepancy in the p-

http://mc.manuscriptcentral.com/bib

For Peer Review

values from these two approaches was obvious as the data points scatter around the 1:1 identity line (Supplementary Figure 1). On the contrast, SCEBE provided very similar p-values for the association test on both intercept and the slope of the model compared to the LME approach regardless of the level of shrinkage.

As expected, compared to standard LME, NEBE severely underestimated the effect size due to shrinkage (Supplementary Figure 2). However, after corrections, the estimates from SCEBE are virtually identical to those based on the LME approach as the data points perfectly aligned on the 1:1 identity line. Similar to the findings based on the real ADNI data, the simulation study also demonstrated that GALLOP and SCEBE provided similar p-values and estimates for SNP effects (Supplementary Figure 1 and 2). All of the four investigated approaches can well controlled the type I error rate at the nominal level (Supplementary Figure 3).

Computation complexity

Since multiple integrations/approximations are required, the computation time for fitting a classic LME by lmer in 'lme4' package increases with the cubic of the number of individuals [17]. In addition, for a typical GWAS with LME, millions of LME model fittings are needed by adding one SNP at a time into the model.

The proposed SCEBE only requires a single run of the time-consuming LME model (ie, the base model without SNP effects) to estimate the random effects parameters (EBEs). Then the association studies are performed by treating the EBEs for a model parameter as the phenotype and SNPs as genotypes using linear regression models. This substantially reduces the per-SNP computation time as it converts the complex LME model to simple

For Peer Review

linear regression. Finally, the bias in SNP-effect estimates and test statistics caused by shrinkage of EBEs is corrected by a correction matrix. Since analytic expression for the correction matrix can be derived theoretically, the computation can be done through matrix-vector manipulation for all the SNPs together as long as the computer memory allows.

Our simulation experiments confirmed that the computation time of SCEBE was drastically improved compared to that for LME (Supplementary Figure 4a). Depending on sample size and number of SNPs, approximately 100 – 2000 folds of increase in computation efficiency was observed for SCEBE. The gain in time efficiency relative to LME improved with increasing sample size or/and increasing number of SNPs (Supplementary Figure 4b). In the GWAS analysis for ADNI data where over 6 million of SNPs were involved, the gain in time efficiency was approximately almost 10, 000 time for SCEBE (Figure 1b). Consistent with the analysis for ADNI data, the SCEBE was 3 – 4 times faster than GALLOP in the simulation studies (Supplementary Figure 4).

Confounding

Confounding due to relatedness or population stratification is one of the most challenging issues in statistical inferences for GWAS [18-21]. We conducted additional simulations to study the impact of population stratification on statistical inference based on the approaches discussed in this article. We simulated data using the Balding-Nichols model [22-24] (details are provided in Supplementary Materials/Section 2).

As expected, in the presence of population stratification, the quantiles of test statistics of the SNPs tend to deviate from the theoretical quantiles of chi-square distribution with 1

For Peer Review

degree of freedom (Supplementary Figure 5). However, SCEBE could still provide unbiased estimates and very similar p-values compared to the standard LME despite of population stratification (Supplementary Figure 6a and 6b). This suggests that population stratification has similar impact on the standard LME and SCEBE. Furthermore, it appears that genomic control [17] could correct the test statistics back to the theoretical distribution for both SCEBE and LME when all simulated SNPs had no effects, and reduce the influence of population stratification when there were SNPs with active effects (Supplementary Figure 5).

Discussion

GWAS with longitudinal outcomes based on repeated measures could markedly increase the statistical power, particularly for detecting genetic variants with relatively weak effects [1-2]. Mixed-effect modeling has been an attractive approach for GWAS with longitudinal outcomes despite of its computational challenge and cost [3, 25]. Althgouh EBE-based approaches can reduce the computational time [7-8], these approaches suffer from shrinkage-induced bias in estimation and association test (i.e., p values), particularly in presence of large measurement errors or with sparse observations per subject. We proposed a approach that can correct the bias related to NEBE but preserve the feature of high throughput for NEBE. We demonstrated that this novel approach with ADNI data and completed a GWAS with longitudinal outcomes on millions of SNPs within an hour in comparison with months using the standard LME modeling, representing nearly 10,000 times improvement of computational efficiency. In addition, our simulation shows that the improvement of time efficiency by SCEBE increases with increasing sample size

For Peer Review

(Supplementary Figure 3). This feature suggests the potential application of SCEBE to modern data with large sample size, particularly for emerging large-scale genetic data from biobanks [26].

Confounding due to relatedness or population stratification is one of the most important and challenging issues in GWAS. Our simulation studies showed that population stratification had similar impacts on all the approaches. Furthermore, our simulation showed that genomic control could correct the bias in the test statistics caused by population stratification. SCEBE reduces the LME-based GWAS for longitudinal outcomes to standard linear-regression GWAS, where EBEs are treated as phenotypes. This allows coupling SCEBE with other more sophisticated approaches, such as, EIGENSTRAT/PCA [19-20] and LD regression [21], for controlling bias due to population stratification. Future research on how to use SCEBE with these confounding-controlling approaches is warranted.

Over the last decade, different approaches have been attempted for nonlinear GWAS of longitudinal outcomes [27-29]. However, these methods are extremely time-consuming and often require hours for only 1,000 tests [1], which is not scalable for large-scale GWAS data with millions of SNPs. In the present paper, although we limited ourselves to linear mixed-effects modeling, SCEBE can be easily extended to nonlinear longitudinal data, which opens the door for efficient and scalable functional GWAS for more complex nonlinear longitudinal traits.

While this paper was in development, Sikorska et al. also present a new algorithm that expedites genome-wide analysis of longitudinal data (GALLOP) [11]. GALLOP solves the equivalent penalized least squares problem efficiently and factorizations and

For Peer Review

transformations are used to avoid inversion of large matrices. Both our simulation study and real-data analysis suggest that GALLOP and SCEBE provide similar p-values and estimation for effect size in the context of linear model for disease progression. However, SCEBE was 3 – 4 times faster than GALLOP. More importantly, when generalizing to nonlinear mixed-effects model, our preliminary simulation study indicated that the performance of GALLOP could be less consistent and exhibited suboptimal performance compared to SCEBE (Supplementary Materials/Section3 and Supplementary Figure 7). This suggests that SCEBE is robust and consistent for GWAS using both linear and nonlinear longitudinal data. Future investigation may be needed in this area.

Acknowledgement

The Authors declare that there is no conflict of interest. Prof. Yang is supported by National Science Foundation of China (NSFC), Grant No. 11671375. Dr. Min Yuan is supported by the Natural Science Foundation of Anhui Provincial Education Department, No. KJ2017A171 and Doctoral research funding of Anhui Medical University, No. XJ201710. Dr Jinfeng Xu is supported in part by the University of Hong Kong Seed Fund for Translational and Applied Research (201711160015) and The University of Hong Kong - Zhejiang Institute of Research and Innovation Seed Fund, and General Research Fund (17308018) of Hong Kong.

References

1. Marchetti-Bowick M, Yin J, Howrylak JA, et al. A time-varying group sparse additive model for genome-wide association studies of dynamic complex traits. *Bioinformatics* 2016; 32(19): 2903-2910.
2. Chiu YF, Justice AE, Melton PE. Longitudinal analytical approaches to genetic data. *BMC genetics* 2016; 17(2): S4.

For Peer Review

3. Lee E, Giovanello KS, Saykin AJ, et al. Single-nucleotide polymorphisms are associated with cognitive decline at Alzheimer's disease conversion within mild cognitive impairment patients. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 2017; 8: 86-95.
4. Xu XS, Yuan M, Yang H, et al. Further evaluation of covariate analysis using empirical Bayes estimates in population pharmacokinetics: the perception of shrinkage and likelihood ratio test. *The AAPS Journal* 2017; 19(1): 264-273.
5. Combes FP, Retout S, Frey N, et al. Powers of the likelihood ratio test and the correlation test using empirical Bayes estimates for various shrinkages in population pharmacokinetics. *CPT: pharmacometrics & systems pharmacology* 2014; 3(4): 1-9.
6. Davidian M, Giltinan DM. Nonlinear models for repeated measurement data: an overview and update. *Journal of agricultural, biological, and environmental statistics* 2003; 8(4): 387-419.
7. Londono D, Chen KM, Musolf A, et al. A novel method for analyzing genetic association with longitudinal phenotypes. *Statistical applications in genetics and molecular biology* 2013; 12(2): 241-261.
8. Meirelles OD, Ding J, Tanaka T, et al. SHAVE: shrinkage estimator measured for multiple visits increases power in GWAS of quantitative traits. *European Journal of Human Genetics* 2013; 21(6): 673.
9. Savic RM, Karlsson MO. Importance of shrinkage in empirical bayes estimates for diagnostics: problems and solutions. *The AAPS Journal* 2009; 11(3): 558-569.
10. Yuan M, Xu XS, Yang Y, et al. A quick and accurate method for the estimation of covariate effects based on empirical Bayes estimates in mixed-effects modeling: Correction of bias due to shrinkage. *Statistical Methods in Medical Research*. 2019; 28: 3568-3578.
11. Sikorska K, Lesaffre E, Groenen PJ, et al. Genome-wide Analysis of Large-scale Longitudinal Outcomes using Penalization-GALLOP algorithm. *Scientific reports* 2018; 8(1): 6815.
12. Delaneau O, Marchini J. The 1000 Genomes Project Consortium. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications* 2014; 5 3934.
13. Saunders AM. Association of apolipoprotein E allele $\epsilon 4$ with late-onset familial and sporadic Alzheimer's disease. *NEUROLOGY* 1993; 43:1467-1472.
14. Cudaback E. Apolipoprotein C-I is an APOE genotype-dependent suppressor of glial activation. *Journal of Neuroinflammation* 2012; 9:192.
15. García-Osta A. Phosphodiesterases as Therapeutic Targets for Alzheimer's Disease. *ACS Chem Neurosci* 2012; 3(11):832-844.

16. O'Connor V. Differential amplification of intron-containing transcripts reveals long term potentiation-associated up-regulation of specific Pde10A phosphodiesterase splice variants. *J Biol Chem* 2004; 279(16): 15841–15849.
17. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics* 2012; 44(7): 821.
18. Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999; 55: 997–1004.
19. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet.* 2006; 2: e190.
20. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 2006; 38: 904–909.
21. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics* 2015; 47(3): 291.
22. Balding DJ, Nichols RA. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* 1995; 96: 3-12.
23. Wright S. The genetical structure of populations. *Ann. Eugen.* 1951; 15: 323–354.
24. Cavalli-Sforza LL, Menozzi P, Piazza A. The History and Geography of Human Genes (Princeton Univ. Press, Princeton, New Jersey, 1994).
25. Xu Z, Shen X, Pan W, et al. Longitudinal analysis is more powerful than cross-sectional analysis in detecting genetic association with neuroimaging phenotypes. *PloS one* 2014; 9(8): e102312.
26. Zhou W, Nielsen JB, Fritsche LG, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature genetics* 2018; 50(9): 1335.
27. Das K, Li J, Wang Z, et al. A dynamic model for genome-wide association studies. *Human genetics* 2011; 129(6): 629-639.
28. Das K, Li J, Fu G, et al. Dynamic semiparametric Bayesian models for genetic mapping of complex trait with irregular longitudinal data. *Statistics in medicine* 2013; 32(3): 509-523.
29. Wang Z, Wang N, Wu R. fGWAS: An R package for genome-wide association analysis with longitudinal phenotypes. *Journal of genetics and genomics= Yi chuan xue bao* 2018; 45(7): 411-413.

For Peer Review

Figure Legends

Figure 1a: Running time required for LME/NEBE/GALLOP/SCEBE to complete GWAS scan of ADNI data (performed on an Ubuntu 16.04 LTS running on a server with CPU@2.9G and 8G RAM; 784 individuals and 6,414,695 SNPs).

Figure 1b : Fold change in computation time (logarithm scale) for NEBE/GALLOP/SCEBE relative to standard LME to complete GWAS scan of 23 chromosomes in ADNI data (784 individuals and 6,414,695 SNPs; fold change is calculated as time for LME over time for alternative methods; each bar represents a chromosome; performed on an Ubuntu 16.04 LTS running on a server with CPU@2.9G and 8G RAM).

Figure 2a: Scatter plots of p-values from NEBE/GALLOP/SCEBE against LME on the -log10 scale for ADNI data with 784 individuals and 6,414,695 SNPs.

Figure 2b: Scatter plots of estimates from NEBE/GALLOP/SCEBE against LME for ADNI data with 784 individuals and 6,414,695 SNPs.

Figure 3a: Manhattan Plot for testing associations on baseline disease status (intercept) by SCEBE for ADNI data with 784 individuals and 6,414,695 SNPs.

Figure 3b: Manhattan Plot for parameter estimation on disease progression (slope) by SCEBE for ADNI data with 784 individuals and 6,414,695 SNPs.

Figure 4: Lollipop Plot for top 20 SNPs selected by SCEBE for ADNI data with 784 individuals and 6,414,695 SNPs (x-axis is –log10 of p-values and y-axis is the SNP name; the number behind each bar is the chromosome ID).

For Peer Review

Supplementary Materials

for

SCEBE: An Efficient and Scalable Algorithm for Genome-wide Association Studies on Longitudinal Outcomes with Mixed-Effects Modeling

For Peer Review

1. Estimation and test statistics for linear mixed model with single covariate

Suppose the GWAS is designed from a natural population with three genotypes at each locus. Let m denote the number of individuals and q denote the number of SNPs. The *i*th individual has $n_i$observations $y_i = (y_{i1}, y_{i2}, ..., y_{ini})'$ at time points $t_i = (t_{i1}, t_{i2}, ..., t_{ini})'$. A typical linear mixed-effects model in GWAS can be written in a two stage form as follows,

$$y_i = Z_i\beta_i + e_i$$
$$\beta_i = \alpha + x_i\gamma + b_i, i = 1, 2, \cdots, m \quad (1)$$
$$e_i \sim N(0, G_i) \text{ and } b_i \sim N(0, R)$$

where β_i is the $p \times 1$ random effect vector. The design matrix Z_i is a $n_i \times p$ matrix. Covariate x_i is the genotype coded as 0, 1 or 2 for three different genotypes. α and γ are p-dimensional intercept and slope parameters. The base model corresponds to model (1) with $\gamma = 0$. G_i is the $n_i \times n_i$ covariance matrix which chracterizes the correlation structure of within-subject variablities. R is a $p \times p$ covariance matrix which characterizes the between-subject variablilities. The standard approach of fitting model (1) is based on the likelihood function and implemented in R packages (e.g., lme4). We call the standard approach 'LME' in this article.

The best predictor of the random effects β_i, defined as the posterior mean of β_i given data y_i, equals to $\mathrm{BP}(\beta_i) = (Z_i'G_i^{-1}Z_i + R^{-1})^{-1}(Z_i'G_i^{-1}y_i + R^{-1}\alpha)$. The parametrical empirical Bayesian estimators (naive EBE) of β_i, denoted as $\hat{\beta}_i$, is then obtained by plugging the MLEs of nuisance parameters such as G_i, R and α. Let the covariance matrix of y_i be $\Sigma_i = Z_iRZ_i' + G_i$, then the MLE of α under the base model is $(\sum_{i=1}^{m} Z_i'\Sigma_i^{-1}Z_i)^{-1} Z_i'\Sigma_i^{-1}y_i$ which can be regarded as the weighted average of y_i. After the naive EBEs (NEBEs) are obtained, a simple linear regression of NEBE is carried out on

For Peer Review

the covariate x_i. The least square estimator of γ is $\hat{\gamma} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})\,\hat{\beta}_i}{\sum_{i=1}^{m}(x_i - \bar{x})^2}$ where $\bar{x}$ is the sample mean.

Under the true model (1), the expectation of y_i is $\mathrm{E}(y_i) = Z_i(\alpha + x_i\gamma)$. Therefore the expectation of $\hat{\beta}_i$, under the true model (1) is $\mathrm{E}(\hat{\beta}_i) = \alpha + [x_i(I_p - S_i) + S_i\sum_{i=1}^{m} x_i W_i]\gamma$ with $W_i = (\sum_{i=1}^{m} Z_i'\Sigma_i^{-1}Z_i)^{-1}Z_i'\Sigma_i^{-1}Z_i$, and $S_i = (Z_i'G_i^{-1}Z_i + R^{-1})^{-1}R^{-1}$. The expectation of $\hat{\gamma}$ under true model (1) can be derived based on the expectation of $\hat{\beta}_i$. Denote $S_c = \frac{\sum_{i=1}^{m}(x_i - \bar{x})\left[x_i(I_p - S_i) + S_i\sum_{i=1}^{m} x_i W_i\right]}{\sum_{i=1}^{m}(x_i - \bar{x})^2}$ where I_p is the p-dimensional identity matrix. we have $\mathrm{E}(\hat{\gamma}) = S_c\gamma$. Noticing that S_c is generally not a diagonal matrix even in the simple case that the sampling time and measuring time points are the same for all the individuals. So the elements of $\hat{\gamma}$ actually estimate the linear combination of elements of γ. Especially when at least one element of γ is not equal to 0, the EBEs-based estimator of γ is largely biased. Thus the EBEs-based estimator $\hat{\gamma}$ can only be used as an unbiased estimate and hypothesis testing after correction. We propose the simutanous correction method in this paper to correct all elements of γ at the same time. The matrix S_c defined above can be served as the simutanous correction matrix and the simutanously corrected estimator of $\hat{\gamma}$ can be expressed as $\hat{\gamma}_{sim} = S_c^{-1}\hat{\gamma}$ which is called SCEBE.

In order to derive the test statistics for hypothesis testing, we need to calculate the variance of $\hat{\gamma}_{sim}$. To show the derivation more clearly, we introduce some notations. Let

$$A_i = (Z_i'G_i^{-1}Z_i + R^{-1})^{-1}Z_i'G_i^{-1},$$

$$B_i = (Z_i'G_i^{-1}Z_i + R^{-1})^{-1}R^{-1},$$

$$C_i = (\sum_{i=1}^{m} Z_i'\Sigma_i^{-1}Z_i)^{-1}Z_i'\Sigma_i^{-1}.$$

For Peer Review

Then the covariance matrix of $\hat{\beta}_i$ can be determined by $\mathrm{var}(\hat{\beta}_i)$ and $\mathrm{cov}(\hat{\beta}_i, \hat{\beta}_j)$ which has the explicit form

$$\mathrm{var}(\hat{\beta}_i) = A_i\Sigma_i A_i' + B_i\left(\sum C_i\Sigma_i C_i'\right)B_i' + A_i\Sigma_i C_i' B_i' + B_i C_i \Sigma_i A_i'$$

$$\mathrm{cov}(\hat{\beta}_i, \hat{\beta}_j) = B_i\left(\sum C_i\Sigma_i C_i'\right)B_j' + A_i\Sigma_i C_i' B_j' + B_i C_j \Sigma_j A_j'.$$

The Variance of $\hat{\gamma}$ can be calculated as

$$\mathrm{var}(\hat{\gamma}) = \frac{\mathrm{var}(\sum(x_i - \bar{x})\hat{\beta}_i)}{(\sum(x_i - \bar{x})^2)^2} = \frac{\sum(x_i - \bar{x})^2 \mathrm{var}(\hat{\beta}_i) + \sum_{i \neq j}(x_i - \bar{x})(x_j - \bar{x})\mathrm{cov}(\hat{\beta}_i, \hat{\beta}_j)}{(\sum(x_i - \bar{x})^2)^2}.$$

The t test statistic for $H_{0i}: \gamma_i = \gamma_{i0}$ can be constrcuted as

$$t_i = \frac{\hat{\gamma}_i - \gamma_{i0}}{\sqrt{[S_c^{-1}\mathrm{var}(\hat{\gamma})(S_c^{-1})']_{i,i}}}.$$

where γ_{i0} is the true value of γ_i, $i = 1,2,\ldots, p$ and the subscript (i,i) denotes the ith diagonal of the matrix $S_c^{-1}\mathrm{var}(\hat{\gamma})(S_c^{-1})'$.

2. Simulation details for generating data with population stratification

Following Price et al. (Nature Genetics, 2006), we simulated data using Balding-Nichols model (Genetica, 1995) for two latent subpopulations. Simulation details are summarized as follows,

- Sample ancestral population allele frequency p from uniform distribution U[0.1, 0.5].
- Sample p_1 and p_2 from beta distribution $\mathrm{Beta}(\frac{p(1-F_{st})}{F_{st}}, \frac{(1-p)(1-F_{st})}{F_{st}})$. This distribution has mean p and variance $F_{st} * p(1-p)$. **The quantity F_{st} measures the genetic distance between two subpopulations** (Wright 1951 and Cavalli-Sforza et al. 1994). F_{st} was set to 0.01.
- Total sample size N=800. Sample $n_1 = 30\%N$ genotypes for the first subpopulation from the multinomial distribution $\mathrm{Mul}(n_1, ((1-p_1)^2, 2p_1(1-p_1), p_1^2)')$ and n_2

For Peer Review

$= 70\%$N genotypes for the second subpopulation from the multinomial distribution $\mathrm{Mul}\left(n_2, ((1-p_2)^2, 2p_2(1-p_2), p_2^2)'\right)$.

- Longitudinal phenotypes for *i*th subject were generated by a linear random intercept and slope model within each subpopulations and combine them to form the final dataset,

$$y_i = \alpha_i + \beta_i t_i + \varepsilon_i$$

$$\alpha_i = \alpha_0 + b_{i1} \qquad (2)$$

$$\beta_i = \gamma_0 + x_i\gamma + g_i + b_{i2},$$

where x_i is the genotype; $\varepsilon_i \sim N(0,1); b_{i1}, b_{i2} \sim N(0,1)$ independently. $g_i = 0.05$ if *i*th subject belongs to the first subgroup; $g_i = -0.05$ otherwise. $\alpha_0 = \gamma_0 = 0$.

- Make inference about γ based on Model (1) under two scenarios by ignoring: (1) all 1000 SNPs are null markers; (2) 50 out of 1000 SNPs are causal markers and the genetic effects (γ) for causal makers are set to be 0.2. We compared the quantiles of the observed test statistics of LME and SCEBE with the chi-square distribution with 1 degree of freedom to study the impact of PS on the test; we also compared the estimates of LME and SCEBE to study the impact of PS on the estimate.

3. Small-scale simulation with nonlinear mixed-effects model

A small-scale simulation with a nonlinear model for pharmacokinetics (PK) was performed. The PK model is defined as follows,

$$y_{ij} = \frac{De^{lka - lv_i}}{e^{lka} - e^{lcl_i - lv_i}}\left(e^{-e^{(lcl_i - lv_i)t_{ij}}} - e^{-e^{lka.t_{ij}}}\right) + \varepsilon_{ij}$$

$$lv_i = \mu_{lv} + \beta_{lv}(WT_i - 70) + \eta_{lvi}$$

$$lcl_i = \mu_{lcl} + \beta_{lcl}(WT_i - 70) + \eta_{lcli}, j = 1,2,\ldots,n_i,\ i = 1,2,\ldots,m.$$

where

- y_{ij}: the observed drug concentration for the i th individual at time t_{ij} after a single dose administration;
- D: single dose;

- lka: the logarithms of the rate of oral absorption (Ka);
- lv_i: the logarithms of volume of distribution in the central compartment (V);
- lcl_i: the logarithms of clearance (CL);
- WT_i: the body weight (covariate) sampled from normal distribution N(70kg,0.09);
- In simulation: sample size $m = 200$; measurement time is randomly drawn from (3, 5, 7, 9); dose$D = 1$;$\mu_{kl} = \mu_{lv} = \log(0.5)$;$\beta_{lv} = 0$; β_{lcl} takes values in interval [-0.5, 0.5]; ε_{ij},η_{lvi}, $\eta_{lcli} \sim N(0,0.09)$; 1000 replicates for each scenario.

Supplementary Tables

Table1: Top 20 significant SNPs and their corresponding genes for baseline disease status and disease progression.

Baseline disease status (intercept)			
SNP name	CHR ID	Corresponding Gene	Relationship
rs429358	19	APOE	within
rs2290454	17	MYO15B	within
rs61982594	14	BDKRB2	nearby
rs11629183	14	BDKRB2	nearby
rs61982595	14	BDKRB2	nearby
rs112109390	22	TBC1D22A	within
rs5767390	22	TBC1D22A	within
rs1318028	22	TBC1D22A	within
rs4823893	22	TBC1D22A	within
rs4823891	22	TBC1D22A	within
rs4823892	22	TBC1D22A	within
rs4239942	22	TBC1D22A	within
rs5767395	22	TBC1D22A	within
rs56023698	10	LOC105378335	nearby
rs4420638	19	APOC1	500B downstream variant
rs56131196	19	APOC1	500B downstream variant
chr19_32037917	19	LINC01837	within
rs12721051	19	APOC1	within
rs79963487	22	NONE	NONE
rs55658667	17	RGS9	within
Disease progression (slope)			

For Peer Review

SNP	CHR	Corresponding Gene	Relationship
rs181201525	23	SPANXN4	nearby
rs9503225	6	LINC01600	nearby
rs1001729	6	LINC01600	nearby
rs9501862	6	LINC01600	nearby
rs7770991	6	LINC01600	nearby
rs9503220	6	LINC01600	nearby
rs7755937	6	LINC01600	nearby
rs9501860	6	LINC01600	nearby
rs2054638	6	LINC01600	nearby
rs78647522	7	BPGM	nearby
rs3799160	6	PDE10A	within
rs13022686	2	LINC01121	within
rs150313784	23	SPANXN4	nearby
rs2368834	12	IQSEC3	within
rs10902747	1	ZNF683	nearby
rs4811516	20	DOK5	nearby
rs10919857	1	CCNQP1	nearby
rs11247938	1	ZNF683	2KB Upstream Variant
rs1012644	20	DOK5	nearby
rs10753872	1	CCNQP1	nearby

Supplementary Figure Legends

Supplementary Figure 1a: P-value comparison on –log10 scale for association tests on intercept among LME, NEBE and two EBE-based approaches with linear mixed-effects model.

Supplementary Figure 1b: P-value comparison on –log10 scale for association tests on slope among LME, NEBE and two EBE-based approaches with linear mixed-effects model.

Supplementary Figure 2a: Comparison of estimation for intercept among LME, NEBE and two EBE-based approaches with a linear mixed-effects model. Each symbol represents the estimation for a simulated dataset.

Supplementary Figure 2b: Comparison of estimation for slope among LME, NEBE and

For Peer Review

two EBE-based approaches with a linear mixed-effects model. Each symbol represents the estimation for a simulated dataset.

Supplementary Figure 3: Type I errors for association tests from LME, NEBE and two EBE-based approaches (sample size *m*=200, 500, 800 and 1000; between-subject error=1; within-subject error=1; 1000 replicates).

Supplementary Figure 4a: Running time for NEBE, GALLOP and SCEBE compared to LME by 96 simulation scenarios with 1000 replicates.

Supplementary Figure 4b Fold change in computation time (logarithm scale) for NEBE, GALLOP and SCEBE compared to LME for 96 simulation scenarios with 1000 replicates. Fold change is calculated as time for LME over time for alternative methods; Each bar represents a simulation scenario.

Supplementary Figure 5: Plots of theoretical quantiles of chi-square distribution with 1 degree of freedom against observed quantiles of LME and SCEBE before and after GC correction for disease progression. Scenario 1: all 1000 SNPs are null markers; Scenario 2: 50 out of 1000 SNPs are causal markers with effect sizes 0.2.

Supplementary Figure 6a: Comparison of p-values on –log10 scale based on LME and SCEBE in presence of population stratification. Scenario 1: all 1000 SNPs are null markers; Scenario 2: 50 out of 1000 SNPs are causal markers with effect sizes 0.2.

Supplementary Figure 6b: Estimation comparison between LME and SCEBE in presence of population stratification. Scenario 1: all 1000 SNPs are null markers; Scenario 2: 50 out of 1000 SNPs are causal markers with effect sizes 0.2.

Supplementary Figure 7: Estimation and p-value comparisons of GALLOP and SCEBE on clearance relative to NLME with sample size *m*=200; dose$D = 1$;$\mu_{kl} = \mu_{lv} = \log(0.5)$; β_{lcl} takes values in interval [-0.5, 0.5]; ε_{ij},η_{lvi}, $\eta_{lcli} \sim N(0,0.09)$;1000 replicates.

For Peer Review

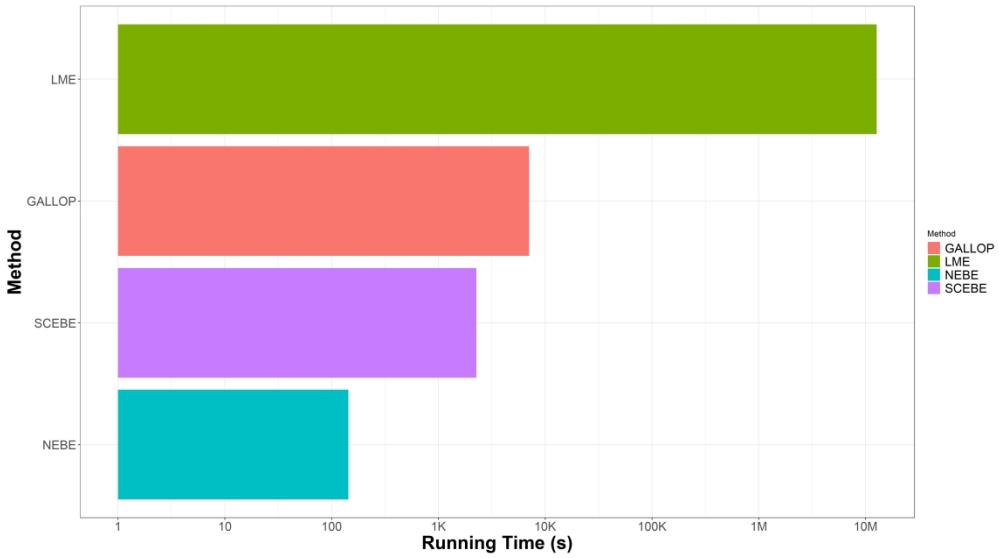


Figure 1a: Running time required for LME/NEBE/GALLOP/SCEBE to complete GWAS scan of ADNI data (performed on an Ubuntu 16.04 LTS running on a server with CPU@2.9G and 8G RAM; 784 individuals and 6,414,695 SNPs).

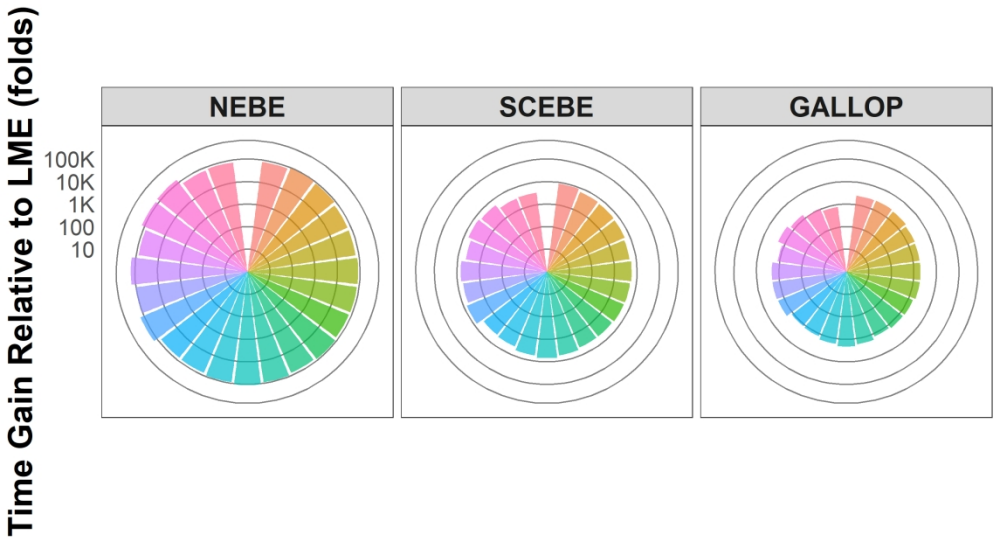


Figure 1b : Fold change in computation time (logarithm scale) for NEBE/GALLOP/SCEBE relative to standard LME to complete GWAS scan of 23 chromosomes in ADNI data (784 individuals and 6,414,695 SNPs; fold change is calculated as time for LME over time for alternative methods; each bar represents a chromosome; performed on an Ubuntu 16.04 LTS running on a server with CPU@2.9G and 8G RAM).

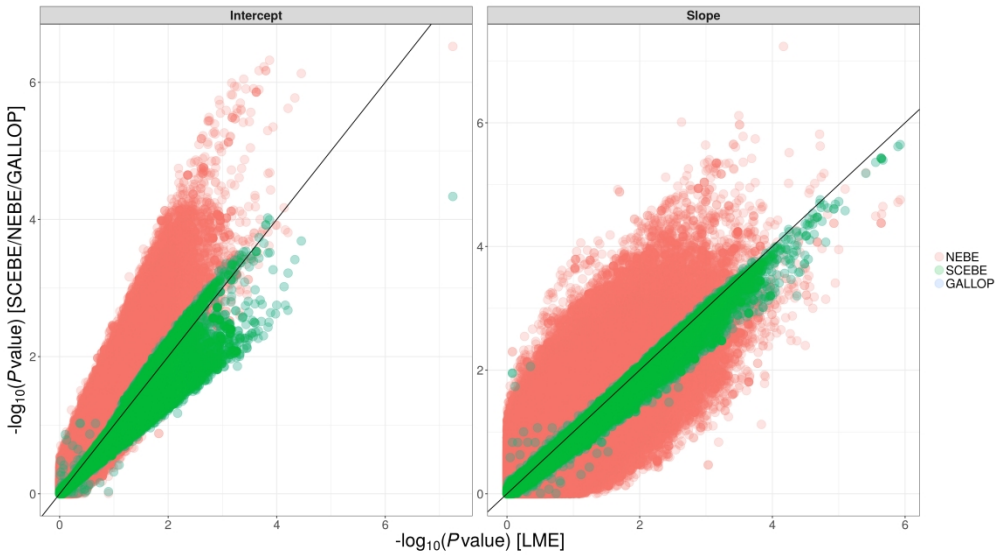


Figure 2a: Scatter plots of p-values from NEBE/GALLOP/SCEBE against LME on the -log10 scale for ADNI data with 784 individuals and 6,414,695 SNPs.

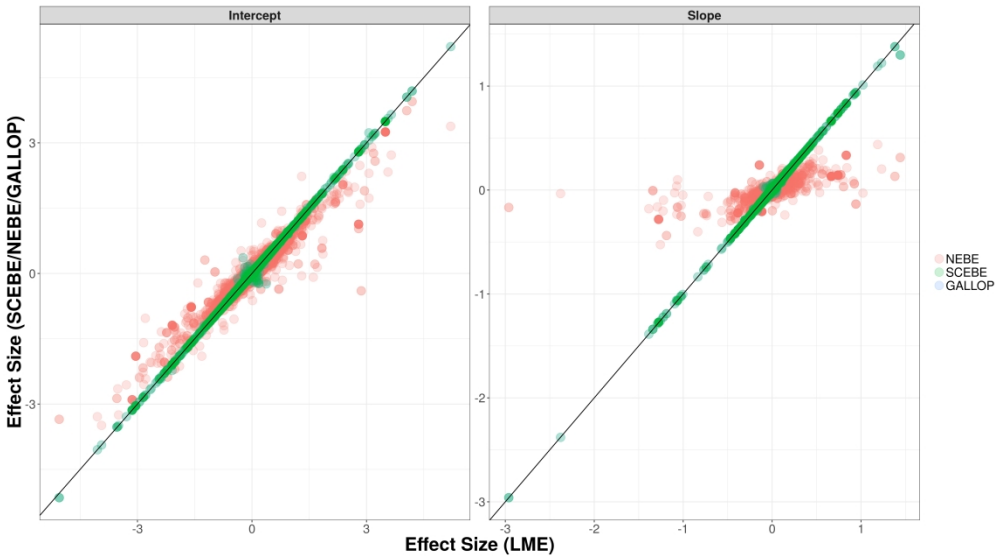


Figure 2b: Scatter plots of estimates from NEBE/GALLOP/SCEBE against LME for ADNI data with 784 individuals and 6,414,695 SNPs.

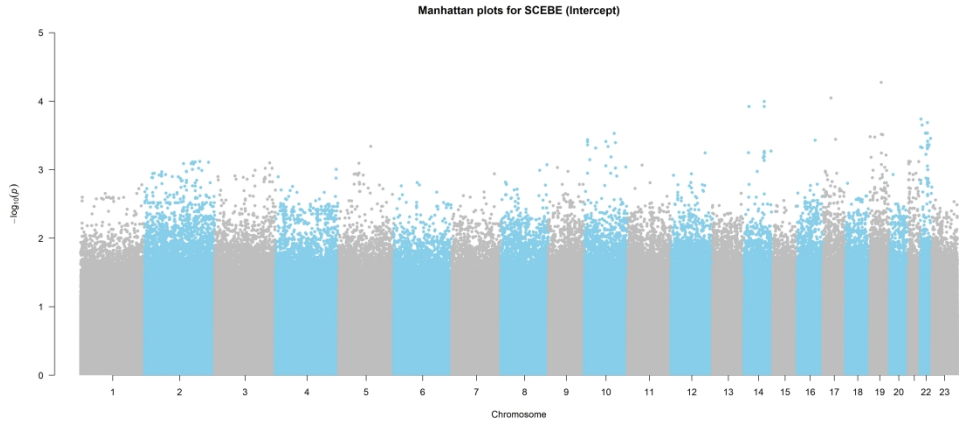


Figure 3a: Manhattan Plot for testing associations on baseline disease status (intercept) by SCEBE for ADNI data with 784 individuals and 6,414,695 SNPs.

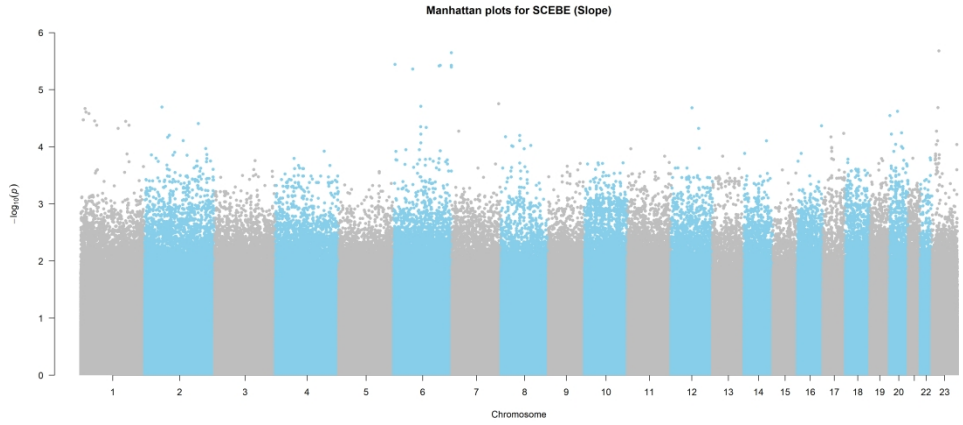


Figure 3b: Manhattan Plot for parameter estimation on disease progression (slope) by SCEBE for ADNI data with 784 individuals and 6,414,695 SNPs.

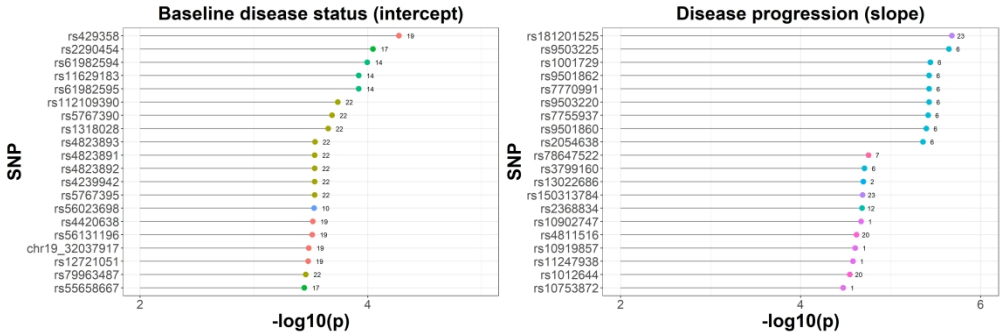
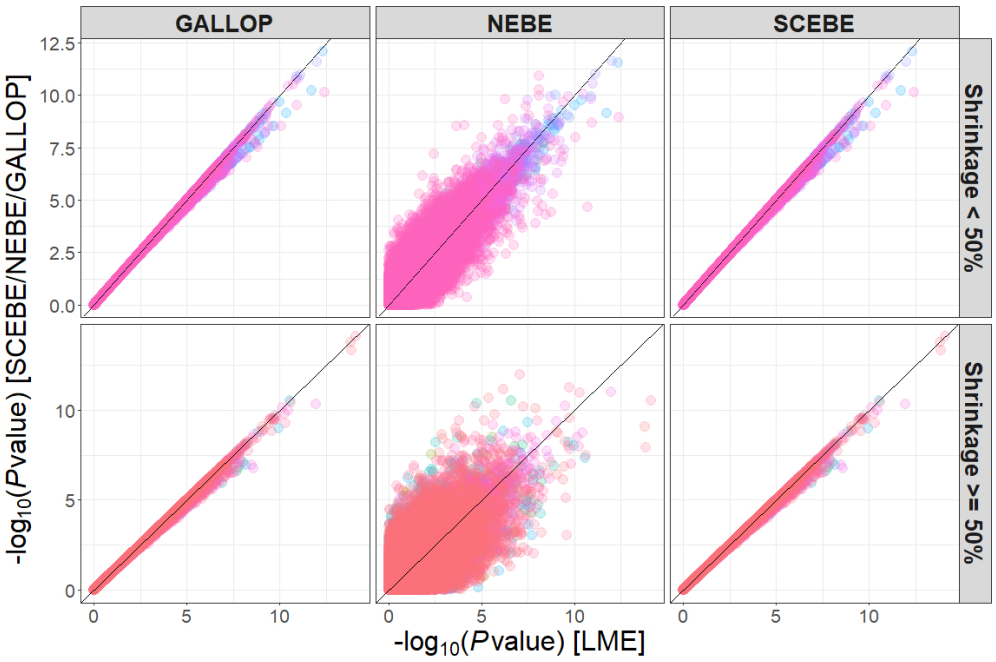
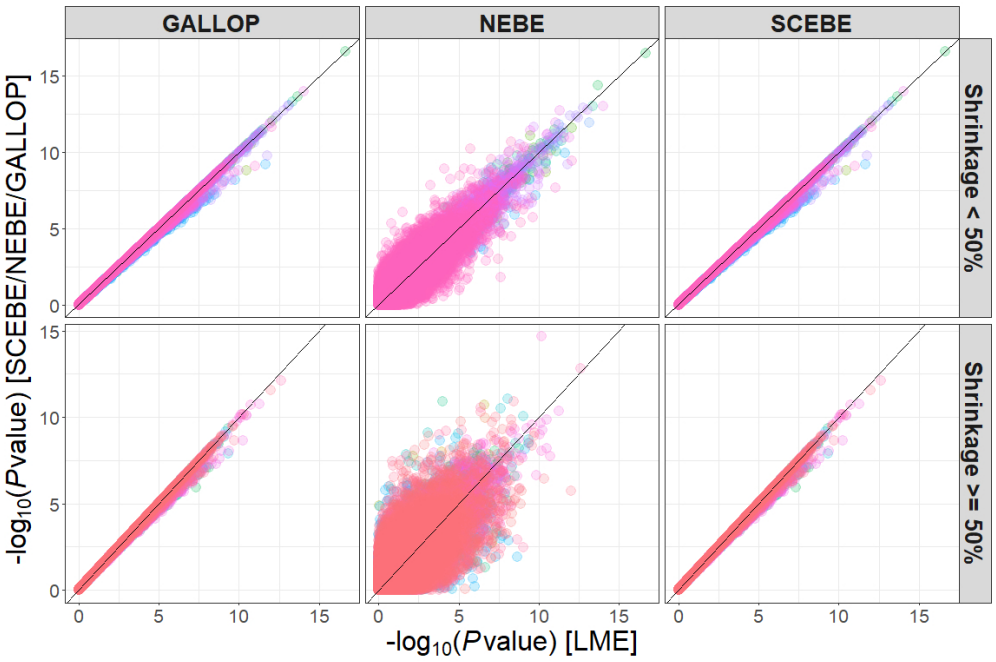


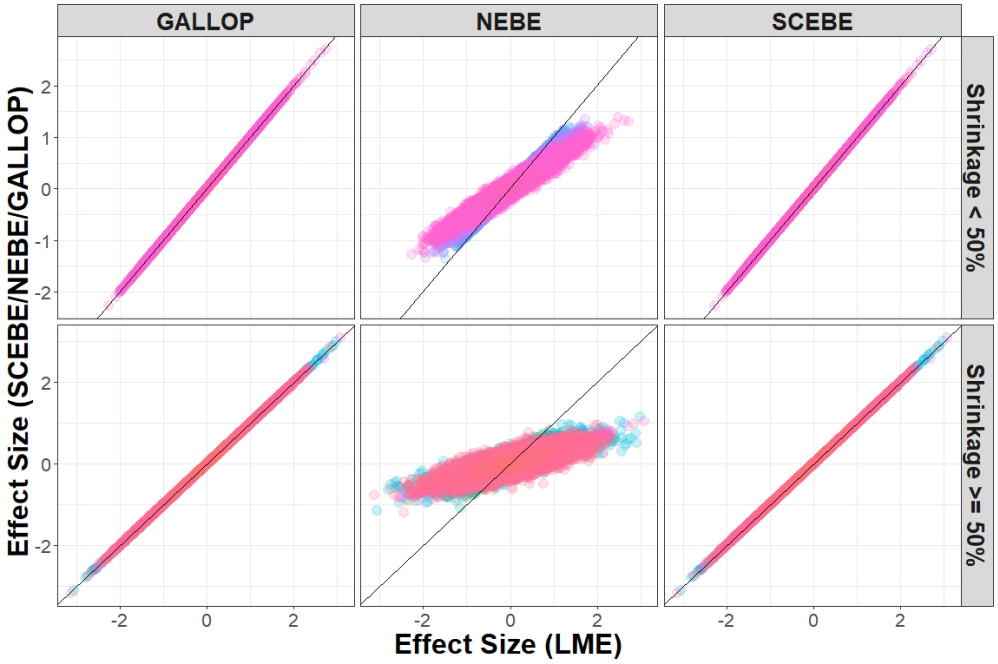
Figure 4: Lollipop Plot for top 20 SNPs selected by SCEBE for ADNI data with 784 individuals and 6,414,695 SNPs (x-axis is –log10 of p-values and y-axis is the SNP name; the number behind each bar is the chromosome ID).



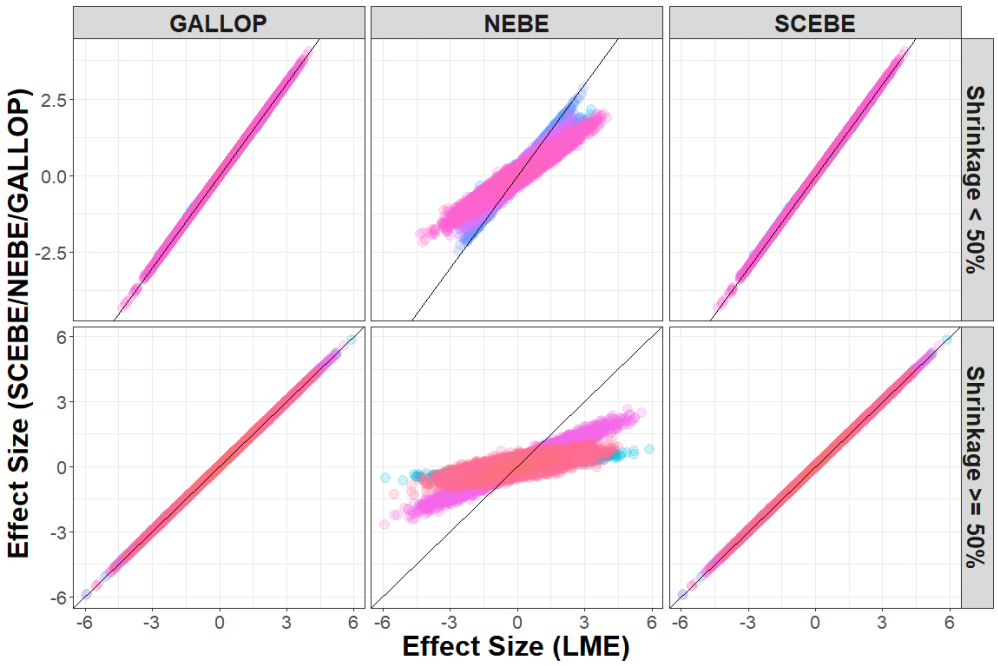
Supplementary Figure 1a: P-value comparison on –log10 scale for association tests on intercept among LME, NEBE and two EBE-based approaches with linear mixed-effects model.



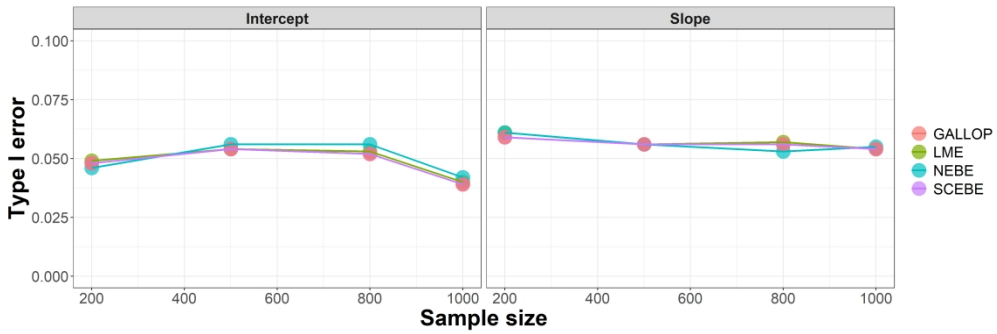
Supplementary Figure 1b: P-value comparison on –log10 scale for association tests on slope among LME, NEBE and two EBE-based approaches with linear mixed-effects model.



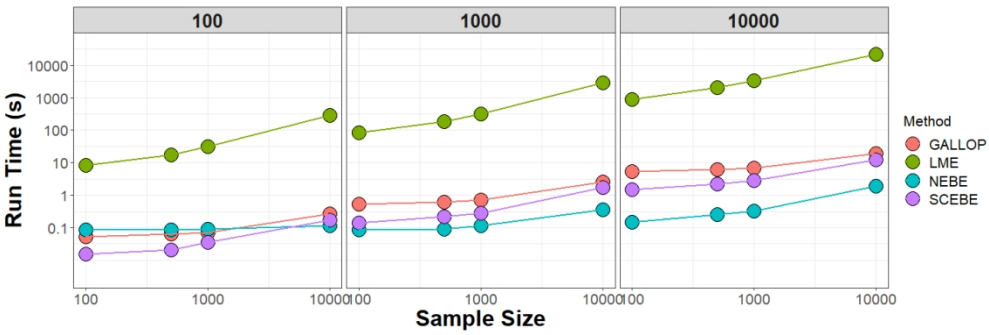
Supplementary Figure 2a: Comparison of estimation for intercept among LME, NEBE and two EBE-based approaches with a linear mixed-effects model. Each symbol represents the estimation for a simulated dataset.



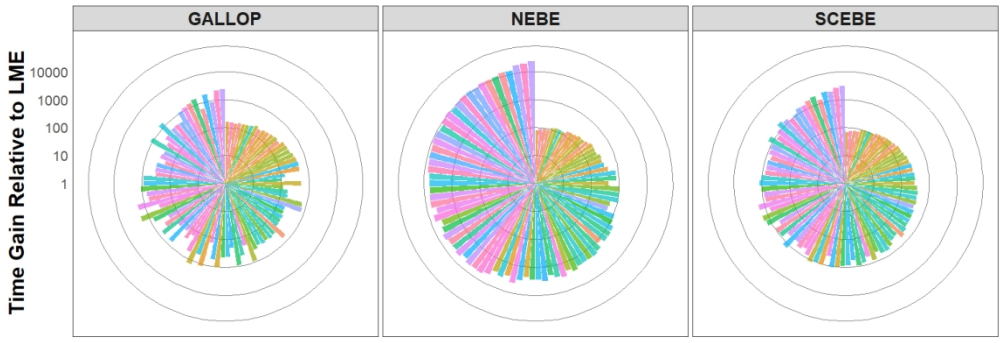
Supplementary Figure 2b: Comparison of estimation for slope among LME, NEBE and two EBE-based approaches with a linear mixed-effects model. Each symbol represents the estimation for a simulated dataset.



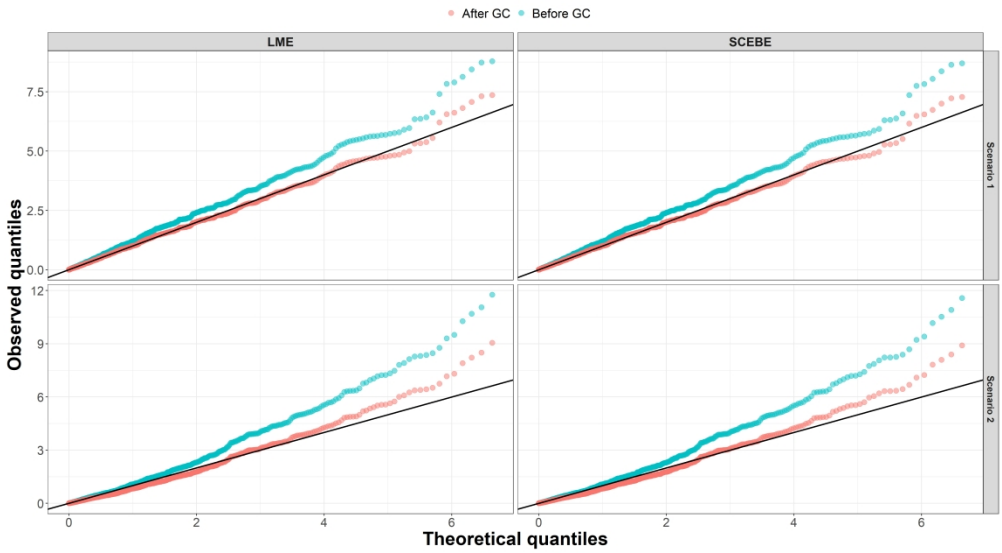
Supplementary Figure 3: Type I errors for association tests from LME, NEBE and two EBE-based approaches (sample size m=200, 500, 800 and 1000; between-subject error=1; within-subject error=1; 1000 replicates).



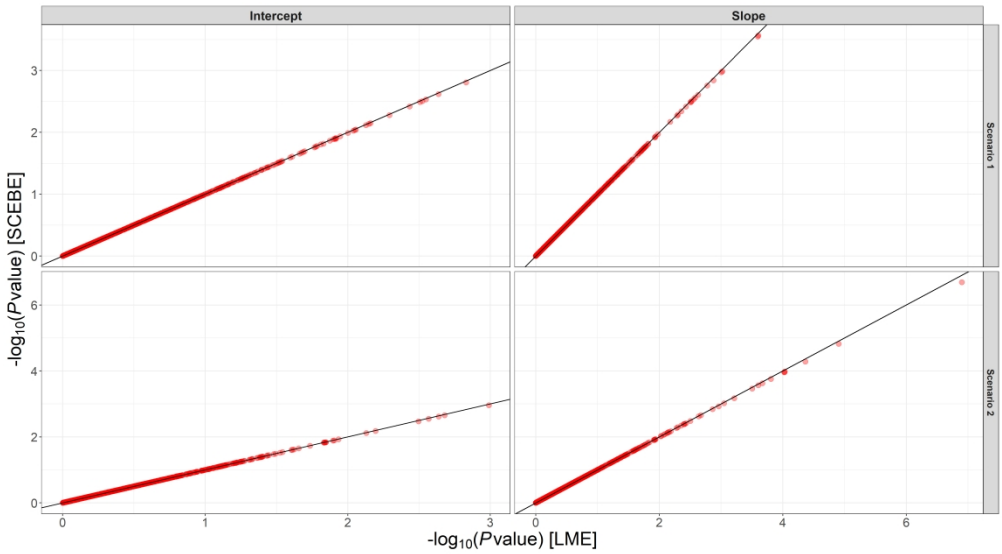
Supplementary Figure 4a: Running time for NEBE, GALLOP and SCEBE compared to LME by 96 simulation scenarios with 1000 replicates.



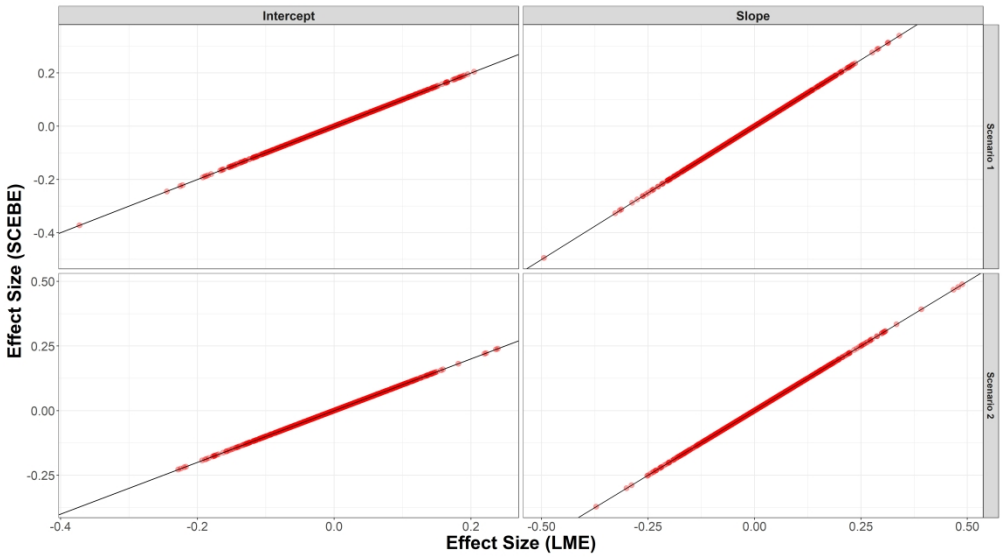
Supplementary Figure 4b Fold change in computation time (logarithm scale) for NEBE, GALLOP and SCEBE compared to LME for 96 simulation scenarios with 1000 replicates. Fold change is calculated as time for LME over time for alternative methods; Each bar represents a simulation scenario.



: Plots of theoretical quantiles of chi-square distribution with 1 degree of freedom against observed quantiles of LME and SCEBE before and after GC correction for disease progression. Scenario 1: all 1000 SNPs are null markers; Scenario 2: 50 out of 1000 SNPs are causal markers with effect sizes 0.2.



Comparison of p-values on –log10 scale based on LME and SCEBE in presence of population stratification. Scenario 1: all 1000 SNPs are null markers; Scenario 2: 50 out of 1000 SNPs are causal markers with effect sizes 0.2.



Estimation comparison between LME and SCEBE in presence of population stratification. Scenario 1: all 1000 SNPs are null markers; Scenario 2: 50 out of 1000 SNPs are causal markers with effect sizes 0.2.