OXFORD

SCEBE: an efficient and scalable algorithm for genome-wide association studies on longitudinal outcomes with mixed-effects modeling

Min Yuan, Xu Steven Xu, Yaning Yang, Yinsheng Zhou, Yi Li, Jinfeng Xu and Jose Pinheiro 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; E-mail: ynyang@ustc.edu.cn; Xu Steven Xu, Genmab US, Inc., Princeton 08540, NJ, USA. Tel: +1-732-277-4987; E-mail: sxu@genmab.com
†Min Yuan and Xu Steven Xu contributed 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 (simultaneous correction for EBE, SCEBE) that can correct the bias in the naive EBE approach and provide unbiased *P*-values and estimates of effect size. Through application to Alzheimer's Disease Neuroimaging Initiative 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 words: genome-wide association studies; longitudinal outcomes; mixed-effects model; empirical Bayesian estimates; shrinkage

Min Yuan is an associate professor in biostatistics in the Center for Big data in Public Health, Anhui Medical University, Anhui, China. Her interest is developing statistical models in genomic research.
Xu Steven Xu is a data scientist in Genmab US, Inc., Princeton, NJ, USA. His interest is developing statistical models and quantitative analysis in clinical trials.
Yaning Yang is a professor in biostatistics in the Department of Statistics and Finance, University of Science and Technology of China, Heifei, China. His interest is genetic statistics.
Yinsheng Zhou and **Yi Li** are PhD students in biostatistics in the Department of Statistics and Finance, University of Science and Technology of China, Hefei, China. Their interest is developing computational algorithm in haplotype-based research.
Jinfeng Xu is an associate professor in Department of Statistics and Actuarial Science, University of Hong Kong, Pok Fu Lam, Hong Kong. His interest is statistical modeling with longitudinal data.
Jose Pinheiro is a statistician in Janssen Research and Development LLC, Raritan, NJ, USA. His interest is statistical modeling with longitudinal data.

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 Bayesian estimates (EBEs) derived from the base mixed-effects model without covariates have 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, 10]. Therefore, there is an urgent need to develop an 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 the 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 $y_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})'$ at time points $t_i = (t_{i1}, t_{i2}, \ldots, t_{in_i})'$. A typical linear mixed-effects model in GWAS can be written in a two stage form as follows:

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

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 variablties. b_i is the $p \times 1$ between-subject error vector following a multinormal 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 NEBEs. 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:

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

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 Data/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 were 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 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 of 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 [12], to impute missing genotypes. After deleting SNPs with minor allele frequency being
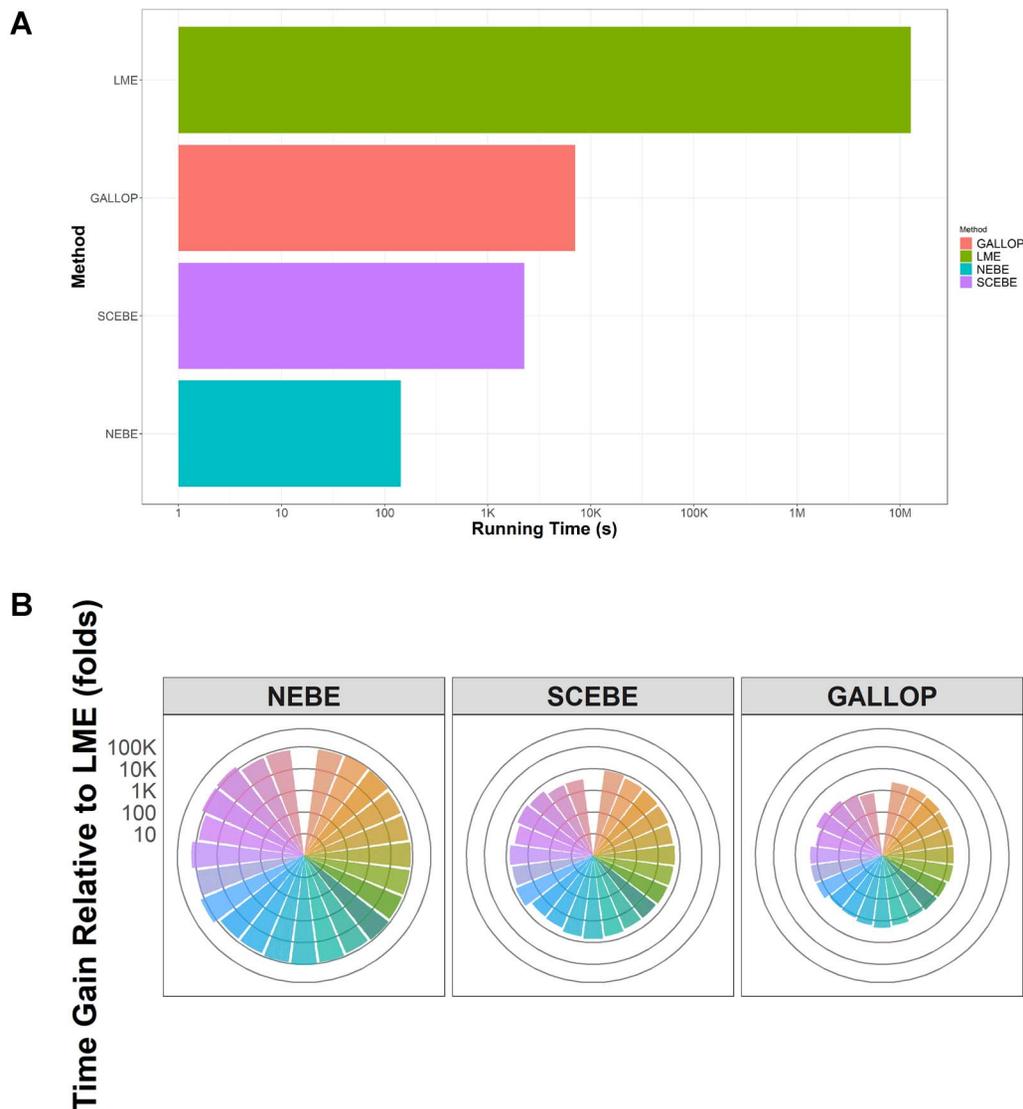
Figure 1. (A) Running time required for LME/NEBE/GALLOP/SCEBE to complete GWAS scan of ADNI data (performed on an Ubuntu 16.04 LTS running on a server with CPU@2.9 GB and 8 GB RAM; 784 individuals and 6 414 695 SNPs). (B) Fold change in computation time (logarithm scale) for NEBE/GALLOP/SCEBE relative 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@2.9 GB and 8 GB RAM).

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, 37 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 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 three to four times faster than GALLOP (Figure 1A and B).

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 that have been reported
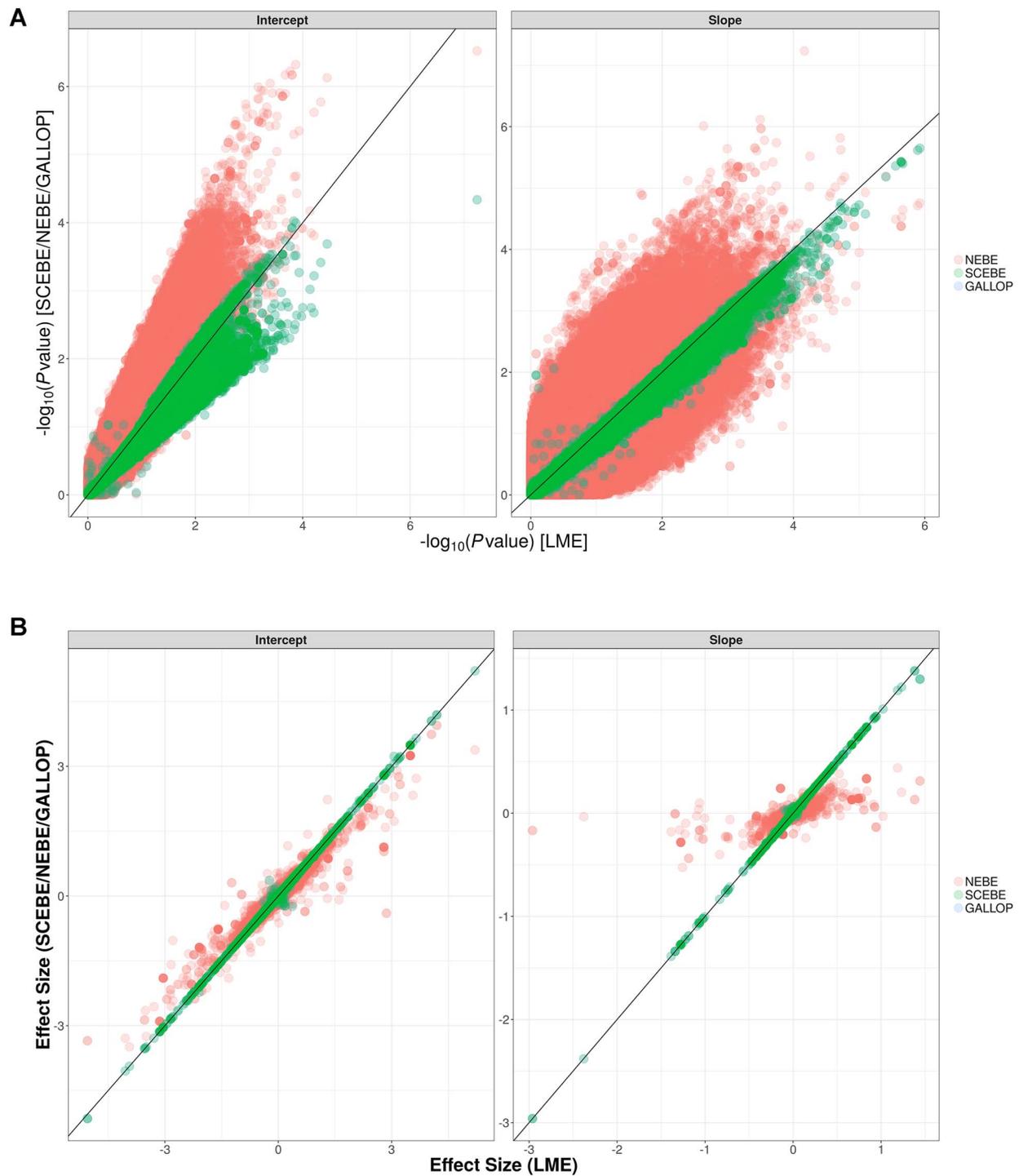
Figure 2. (A) Scatter plots of P-values from NEBE/GALLOP/SCEBE against LME on the −log10 scale for ADNI data with 784 individuals and 6 414 695 SNPs. (B) Scatter plots of estimates from NEBE/GALLOP/SCEBE against LME for ADNI data with 784 individuals and 6 414 695 SNPs.

to be associated with AD [13, 14]. Among them, rs429358 is within apolipoprotein E (APOE), rs12721051 is within APOC1 and rs4420638 and 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 phosphodiesterase (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
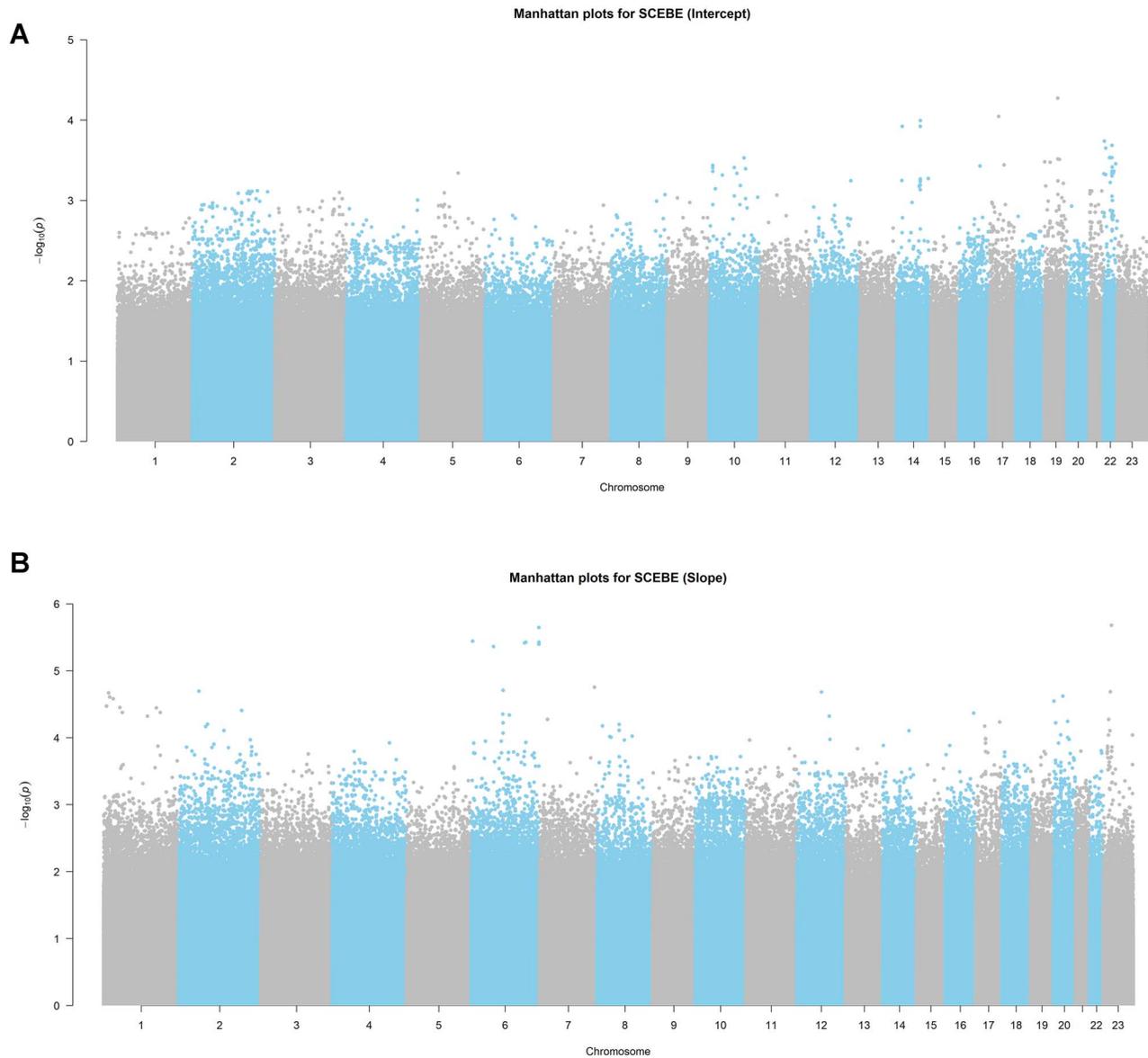
Figure 3. (A) Manhattan plot for testing associations on baseline disease status (intercept) by SCEBE for ADNI data with 784 individuals and 6 414 695 SNPs. (B) Manhattan plot for parameter estimation on disease progression (slope) by SCEBE for ADNI data with 784 individuals and 6 414 695 SNPs.

hippocampus of awake adult rats [16], indicating that it may have effects on memory and cognition.

Since very few GWAS have been reported using RAVLT scores over time, the other SNPs identified in this study (Supplementary data, Table S1) may provide 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 $10\,000$ 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 $U(0.05, 0.5)$ 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. Hundred, 1000 and 10 000 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 $U(0, 0.5)$. 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
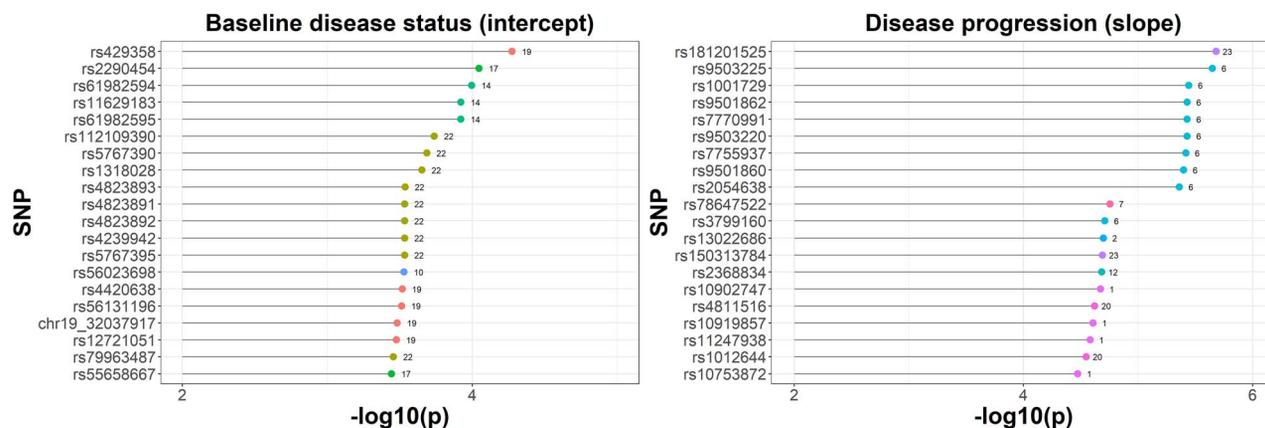
Figure 4. Lollipop plot for top 20 SNPs selected by SCEBE for ADNI data with 784 individuals and 6 414 695 SNPs (x-axis is −log10 of P-values and y-axis is the SNP name; the number behind each bar is the chromosome ID).

based on the LME approach, the discrepancy in the P-values from these two approaches was obvious as the data points scatter around the 1:1 identity line (Supplementary data, Figure S1). 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 data, Figure S2). 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 data, Figures S1 and S2). All of the four investigated approaches can well controlled the type I error rate at the nominal level (Supplementary data, Figure S3).

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 (i.e. the base model without SNP effects) to estimate the random effects parameters. 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 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 data, Figure S4A). Depending on sample size and number of SNPs, approximately 100–2000 folds of increase in computation efficiency were observed for SCEBE. The

gain in time efficiency relative to LME improved with increasing sample size or/and increasing number of SNPs (Supplementary data, Figure S4B). 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 three to four times faster than GALLOP in the simulation studies (Supplementary data, Figure S4).

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 Data/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 degree of freedom (Supplementary data, Figure S5). However, SCEBE could still provide unbiased estimates and very similar P-values compared to the standard LME despite of population stratification (Supplementary data, Figure S6A and B). 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 data, Figure S5).

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 an

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 (Supplementary data, Figure S3). 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/principal components analysis [19, 20] and linkage disequilibrium 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 1000 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 open 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 transformations are used to avoid inversion of large matrices. Both of 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 three to four 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 Data/Section 3 and Supplementary data, Figure S7). 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.

> **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.

Supplementary Data

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

Funding

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):2903–10.
2. Chiu YF, Justice AE, Melton PE. Longitudinal analytical approaches to genetic data. *BMC Genet* 2016;**17**(2):S4.
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 Dement* 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. *AAPS J* 2017;**19**(1):264–73.
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 Syst Pharmacol* 2014;**3**(4):1–9.
6. Davidian M, Giltinan DM. Nonlinear models for repeated measurement data: an overview and update. *J Agr Biol Envir St* 2003;**8**(4):387–419.
7. Londono D, Chen KM, Musolf A, *et al*. A novel method for analyzing genetic association with longitudinal phenotypes. *Stat Appl Genet Mol Biol* 2013;**12**(2):241–61.
8. Meirelles OD, Ding J, Tanaka T, *et al*. SHAVE: shrinkage estimator measured for multiple visits increases power in GWAS of quantitative traits. *Eur J Hum Genet* 2013;**21**(6):673.
9. Savic RM, Karlsson MO. Importance of shrinkage in empirical Bayes estimates for diagnostics: problems and solutions. *AAPS J* 2009;**11**(3):558–69.
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. *Stat Methods Med Res* 2019;**28**:3568–78.
11. Sikorska K, Lesaffre E, Groenen PJ, *et al*. Genome-wide analysis of large-scale longitudinal outcomes using penalization-GALLOP algorithm. *Sci Rep* 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. *Nat Commun* 2014;**5**:3934.
13. Saunders AM. Association of apolipoprotein E allele (Fx1) with late-onset familial and sporadic Alzheimer's disease. *Neurology* 1993;**43**:1467–72.
14. Cudaback E. Apolipoprotein C-I is an APOE genotype-dependent suppressor of glial activation. *J Neuroinflammation* 2012;**9**:192.
15. García-Osta A. Phosphodiesterases as therapeutic targets for Alzheimer's disease. *ACS Chem Nerosci* 2012;**3**(11): 832–44.
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):15841–9.
17. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 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. *Nat Genet* 2006;**38**:904–9.
21. Bulik-Sullivan BK, Loh PR, Finucane HK, *et al*. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 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**:323–54.
24. Cavalli-Sforza LL, Menozzi P, Piazza A. *The History and Geography of Human Genes*. Princeton, NJ: Princeton University Press, 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. *Nat Genet* 2018;**50**(9):1335.
27. Das K, Li J, Wang Z, *et al*. A dynamic model for genome-wide association studies. *Hum Genet* 2011;**129**(6):629–39.
28. Das K, Li J, Fu G, *et al*. Dynamic semiparametric Bayesian models for genetic mapping of complex trait with irregular longitudinal data. *Stat Med* 2013;**32**(3):509–23.
29. Wang Z, Wang N, Wu R. fGWAS: an R package for genome-wide association analysis with longitudinal phenotypes. *J Genet Genomics* 2018;**45**(7):411–3.