Bioinformatics, 36(10), 2020, 3004–3010
doi: 10.1093/bioinformatics/btaa120
Advance Access Publication Date: 25 February 2020
Original Paper

Genome analysis

High-throughput and efficient multilocus genome-wide association study on longitudinal outcomes

Huang Xu[1,†], Xiang Li[2,†], Yaning Yang[1], Yi Li[1], Jose Pinheiro[2], Kate Sasser[3], Hisham Hamadeh[3], Xu Steven[3,*], Min Yuan [4,*] and for the Alzheimer's Disease Neuroimaging Initiative

[1]Department of Statistics and Finance, University of Science and Technology of China, Hefei 230026, China, [2]Janssen Research and Development, Raritan, NJ 08869, USA, [3]Genmab US, Inc., Princeton, NJ 08540, USA and [4]School of Public Health Administration, Anhui Medical University, Hefei 230032, China

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Alfonso Valencia

Received on October 29, 2019; revised on January 16, 2020; editorial decision on February 15, 2020; accepted on February 18, 2020

Abstract

Motivation: With the emerging of high-dimensional genomic data, genetic analysis such as genome-wide association studies (GWAS) have played an important role in identifying disease-related genetic variants and novel treatments. Complex longitudinal phenotypes are commonly collected in medical studies. However, since limited analytical approaches are available for longitudinal traits, these data are often underutilized. In this article, we develop a high-throughput machine learning approach for multilocus GWAS using longitudinal traits by coupling Empirical Bayesian Estimates from mixed-effects modeling with a novel ℓ_0-norm algorithm.

Results: Extensive simulations demonstrated that the proposed approach not only provided accurate selection of single nucleotide polymorphisms (SNPs) with comparable or higher power but also robust control of false positives. More importantly, this novel approach is highly scalable and could be approximately >1000 times faster than recently published approaches, making genome-wide multilocus analysis of longitudinal traits possible. In addition, our proposed approach can simultaneously analyze millions of SNPs if the computer memory allows, thereby potentially allowing a true multilocus analysis for high-dimensional genomic data. With application to the data from Alzheimer's Disease Neuroimaging Initiative, we confirmed that our approach can identify well-known SNPs associated with AD and were much faster than recently published approaches (≥6000 times).

Availability and implementation: The source code and the testing datasets are available at https://github.com/Myuan2019/EBE_APML0.

Contact: sxu@genmab.com or myuan@ustc.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Repeated measurements of phenotypes are usually collected to understand dynamics of diseases such as the onset of disease, treatment effect, resistance to a treatment and progression or relapse of diseases. It is well appreciated that high-dimensional genomic analysis such as genome-wide association studies (GWAS) based on repeated measures could markedly increase the statistical power, particularly for detecting genetic variants with relatively weak effects (Chiu *et al.*, 2016; Marchetti-Bowick *et al.*, 2016). In addition, multilocus GWAS analyses could further improve the power of GWAS and produce more accurate *P*-values and tests (Li *et al.*, 2011; Wu *et al.*, 2009). Therefore, it is imperative to develop a multilocus approach for GWAS using longitudinal traits.

Over the last decade, different approaches have been attempted for multilocus GWAS of longitudinal outcomes (Das *et al.*, 2013; Furlotte *et al.*, 2012; Jiang *et al.*, 2015; Li and Sillanpää, 2013; Li *et al.*, 2015; Londono *et al.*, 2013; Meirelles *et al.*, 2013; Sikorska *et al.*, 2013; Yang *et al.*, 2009). Most recently, Time-Varying Group Sparse Additive Model (TV-GroupSpAM) was proposed to provide a multilocus, functional analysis solution for high-dimensional GWAS data and longitudinal traits (Marchetti-Bowick *et al.*, 2016), which demonstrated greater statistical power than previous methods. However, although TV-GroupSpAM demonstrated

computational advantage over other published methods for functional GWAS, it is still extremely time-consuming and computationally expensive and takes >1 h for testing 1000 single nucleotide polymorphisms (SNPs). Therefore, it is not scalable and computationally infeasible for large-scale GWAS where tests of millions of SNPs are required.

Empirical Bayesian Estimates (EBEs), derived from mixed-effects models without covariates, are often used to facilitate identification of covariates for longitudinal data (Combes *et al.*, 2014; Savic and Karlsson, 2009). The EBEs-based variable selection approach is simple and quick because only simple linear regression is involved. Recently, Xu *et al.* (2017) performed extensive simulation studies and revealed that statistical tests based on EBEs not only provided almost identical power for detecting a covariate effect but also better controlled the false positive (FP) rate compared to the commonly used likelihood ratio test within the framework of non-linear mixed-effects modeling.

Regularized methods were attempted to improve the power of GWAS and produce more accurate *P*-values and tests (Li *et al.*, 2011; Wu *et al.*, 2009). These approaches usually use a ℓ_1-norm regularization penalty (Tibshirani, 1996) to identify SNPs that are predictive of phenotypical outcomes. It is well-known that the optimal penalty for the variable selection purpose is the ℓ_0-norm of the regression coefficients for all predictors. Unfortunately, due to the non-convexity and discontinuity of the ℓ_0-norm, solving such a regularized optimization is computationally challenging, known as non-deterministic polynomial-time hard (NP-hard, Natarajan, 1995). Recently, Li *et al.* (2018) developed a two-stage procedure for ℓ_0-penalty variable selection and demonstrated superior performance of the proposed method in terms of selection accuracy and computational speed as compared to ℓ_1-norm.

In this article, taking advantages of simplicity of EBEs and accuracy of the novel ℓ_0-norm algorithm, we develop a two-stage machine learning approach named $\text{EBE}_{\text{APML0}}$ for multilocus GWAS using longitudinal traits. Compared to existing methods, this approach not only provides accurate selection of SNPs with comparable or higher power but also robust control of FPs. In addition, our proposed approach is able to accommodate as many SNPs as the computer memory allows, e.g. perform multilocus GWAS analysis by chromosome with ≥100 000 SNPs altogether. Importantly, this novel approach is highly scalable and could be approximately thousands times faster than recently published approaches.

2 Materials and methods

Denote by y_{ij} the longitudinal disease profile at time $t_{ij}, i = 1,\ldots,n, j = 1,\ldots,n_i$ and X_i the $p \times 1$ biomarkers (SNPs) for the ith subject. To identify prognostic biomarkers for the dynamic profile, we could model their relationship via a mixed-effects model, written as

$$\begin{aligned} y_{ij} &= \alpha_i + \beta_i t_{ij} + \epsilon_{ij}, \epsilon_{ij} \sim N\left(0, \sigma_y^2\right) \\ \alpha_i &= \alpha_0 + \eta_{i0}, \eta_{i0} \sim N\left(0, \sigma_\alpha^2\right) \\ \beta_i &= \beta_0 + \gamma' X_i + \eta_{i1}, \eta_{i1} \sim N\left(0, \sigma_\beta^2\right), \end{aligned} \tag{1}$$

where the intercept α_i and slope β_i are random effect parameters, γ is the $p \times 1$ biomarker effect on the random slope. The regression coefficient of the SNP × time interaction term characterizes the size of SNPs' influence on the evolution of the trait over time. The model could also be written as follow:

$$\begin{aligned} &y_{ij} = \alpha_0 + \beta_0 t_{ij} + \gamma X_i t_{ij} + \eta_{i0} + \eta_{i1} t_{ij} + \epsilon_{ij} \\ &\eta_{i0} \tilde{} N\left(0, \sigma_\alpha^2\right), \eta_{i1} \tilde{} N\left(0, \sigma_\beta^2\right), \epsilon_{ij} \tilde{} N\left(0, \sigma^2\right). \end{aligned}$$

It is challenging to estimate the model when the high-dimensional biomarkers are modeled interactionally with the dynamic profile and random effect is involved. To solve this problem, we propose a two-stage modeling approach to select biomarkers for longitudinal dynamic data by separating the modeling of dynamic profile and biomarker selection. A similar mixed-effects model but without biomarkers is used to model the dynamic profile alone, given by setting $\gamma = 0$ in Equation (1) which is always called 'reduced model' or 'base model'. The longitudinal profile could be summarized and simplified to the random slope. The random slope under the reduced model could be estimated and extracted by the empirical Bayesian method. EBEs are calculated as the mean or median of posterior distribution of random effects given the data and the fixed effects and are usually called the best linear unbiased predictors (BLUPs) in literature. The calculation of EBEs does not involve biomarker, thus is relatively simple and time efficient.

After the BLUPs are calculated from the reduced model, we could use a much simpler linear model to model the dynamic profile with biomarkers via this subject-specific random effect instead of directly performing biomarker selection on the longitudinal data, that is,

$$\hat{\beta}_i = \beta_0^* + \gamma^{*\prime} X_i + \eta_{i1}^* .$$

Classical biomarkers selection is usually performed by a ℓ_0 penalization on the likelihood function. Denote by $l(\gamma_0, \gamma)$ the log-likelihood function of the model above. The ℓ_0 penalized optimization problem is given by:

$$\min\left(-n^{-1} l(\gamma_0, \gamma) + \lambda |\gamma|_0\right),$$

where $||\gamma||_0 = \sum_{j=1}^{p} I(\gamma_j \neq 0)$ and γ_j is the jth component of γ. It is infeasible to solve this problem computationally, especially when the number of biomarkers is large, which is an NP-hard problem. We follow the same formulation as Li *et al.* (2018) and augment the function with a surrogate parameter θ and minimize the following constrained objective function:

$$-n^{-1} l(\gamma_0, \gamma) + \lambda ||\theta||_0, \ \text{s.t.} \sum_{j=1}^{p} \phi_j(\gamma_j - \theta_j) \leq M, \tag{2}$$

where $\phi_j(x)$ is a convex function satisfying $\phi_j(0) = 0$ and $\phi_j(|x|) \geq 0$ for any $x \neq 0$, and M is a tuning parameter. When $M = 0$, it becomes the original ℓ_0-norm problem. The objective function could be written in the equivalent Lagrange form:

$$L_\lambda(\gamma, \theta) = -n^{-1} l(\gamma_0, \gamma) + \lambda ||\theta||_0 + \rho \sum_{j=1}^{p} \phi_j(\gamma_j - \theta_j).$$

To minimize this, we could iteratively update all parameters till convergence. Parameters are initialized from all zeros and updated based on the algorithm, given by

$$\begin{aligned} \hat{\gamma}_0, \hat{\gamma} = &\ \text{argmin}\left(-n^{-1} l\left(\gamma_0, \gamma\right) + \rho \sum_{j=1}^{p} \phi_j(|\gamma_j|)\right) \\ \hat{\theta} = &\ \text{argmin}\left(\lambda |\theta|_0 + \rho \sum_{j=1}^{p} \phi_j(\hat{\gamma}_j - \theta_j)\right). \end{aligned}$$

Note that in the first stage, the target function does not involve thetas and it can be optimized take a coordinate descent approach. For the jth component of γ we solve

$$-n^{-1} \frac{\partial l(\gamma_0, \gamma | \hat{\gamma}_{-j})}{\partial \gamma_j} + \lambda \frac{\partial \phi_j(\gamma_j)}{\partial \gamma_j} = 0, j = 1, 2, \ldots, p,$$

where $l(\gamma_0, \gamma | \hat{\gamma}_{-j})$ is the log-likelihood function with all components fixed except jth component of γ. And the second step could be solved in a closed form:

$$\hat{\theta}_j = \hat{\gamma}_j I\left(\phi_j(\hat{\gamma}_j) > \frac{\lambda}{\rho}\right),$$

which is to perform hard-thresholding on the estimate from the first step. We call this approach as one-step coordinate descent algorithm since the first stage theta is taken to be 0 and we need not to update

gamma iteratively after theta is updated in the second stage. Li *et al.* (2018) showed that the one-step updating strategy performs well and can substantially improve computational efficiency.

Many convex function includes lasso (ℓ_1-norm) and elastic net (combination of ℓ_1-norm and ℓ_2-norm) can be used in Equation (2). In Li *et al.* (2018), the number of non-zero coefficients are tuned in the second step by keeping the first few largest coefficients of $|\hat{\gamma}_j|$. In this article, we suggest to directly selecting the two tuning parameters λ and ρ based on cross-validation. We will implement this new feature in the next version of R package APML0 (https:/cran.r-project.org/web/packages/APML0/index.html).

The novelty of the proposed approach lies in its ability to perform multilocus feature selection for large-scale GWAS while effectively controlling FP rate and maintaining high power. Furthermore, the proposed approach is efficient and is scalable to ultra-high dimension of SNPs and can perform whole GWAS screening in minutes.

3 Results

3.1 Simulation study

To evaluate the performance of our method, we performed intensive simulation studies under a linear mixed-effects model with random intercept and slope. First, we generated SNPs ($p = 50$, 100, 1000 or 10 000) for a group of subjects (N= 100, 500 or 1000). The correlation r between SNPs was set to be $r = 0$ or 0.8 using function genCorData in R package 'simstudy'. Then, a subset of SNPs ($q = 5$, 10 or 20) were selected as the true variables with active effects. Next, we generated the observation y_{ij} at jth time t_{ij} for the ith subject according to the mixed-effects model with random intercept and random slope described as in (1). For each subject, 7 time points were simulated at $t_{ij} = 1$, 14, 27, 40, 53, 66 and 79. In Equation (1), $\alpha_0 = 0.8$, $\sigma_\alpha^2 = 0.01$ and 0.04 while the random errors $\sigma_y^2 = 0.01$ and 0.09, $\sigma_\beta^2 = 10^{-6}$ and 4×10^{-6}. The intercept parameter β_0 is set to be 0.002 and $\gamma_j = 0, j = 1, \ldots, p-q$; $\gamma_j = (-1)^{j-p+q+1} \times 2.5(j-p+q) \times 10^{-4}, j = p-q+1, \ldots, p$. We chose the parameters' values according to a real dataset for Alzheimer's disease (Xu *et al.*, 2013). Since the unit of the progression rate (point/day) and the progression rate for chronic disease such as Alzheimer's disease is usually slow, the parameter γ (slope for disease progression) is therefore very small.

We compared performance of our method, $\text{EBE}_{\text{APML0}}$, with the TV-GroupdSpAM in terms of number of FPs, number of true positives (TP), F1 score and running times under various scenarios. We also compared our method with a naive two-stage approach, $\text{EBE}_{\text{LASSO}}$, which applies the LASSO method in the second stage. As the number of SNPs selected is required to be specified in advance for TV-GroupSpAM, we selected exactly q SNPs assuming that number of active variables is known. For other methods, there is no need to predefine the number of SNPs to be selected. We present the results for $q = 10$ in Figure 1. Similar results were observed for $q =$ 5 and 20, and presented in Supplementary Figures S2 and S3. In total, 288 simulation scenarios were created, and we generated 20 datasets for each scenario. We also explored the sensitivity of the distribution assumptions on the random effects. Twenty datasets were generated for each of the 288 scenarios by sampling the random effect from a t distribution (df = 4). To keep the variance of the t distribution the same with the variance of normal distribution, the t distribution was multiplied by a constant. The results demonstrated that the proposed method is robust to the specification of the distribution of random effects, and the use of t distribution produced almost identical results compared to those based on the normal distribution (Supplementary Figures S1–S3). In addition, to confirm the simulation result with larger number of simulation replicates, we also generated 100 datasets under each of the following scenarios $N = 1000, q = 10, p = 1000$. The results are shown in Supplementary Table S1.

Generally, $\text{EBE}_{\text{APML0}}$ provides consistent and robust control for FPs genetic variants. The FPs did not change with increasing sample size or number of SNPs although it slightly increased with correlation among SNPs. $\text{EBE}_{\text{APML0}}$ consistently selected <3 FPs when SNPs are independent and it selected at most 19 SNPs (out of 1000 SNPs) when SNPs are highly correlated. Interestingly, the FPs would decrease to a relative low level in the highly correlated situations when the number of SNPs is large (e.g. $p = 10\,000$) so that the true variables are sparse enough. The FPs for $\text{EBE}_{\text{LASSO}}$ were consistently high, ranging from 4 to 140. The FP of TV-GroupSpAM varied with sample size and number of SNPs. It seems to be able to control FP when sample size is large ($N \geq 500$). However, number of FPs was high when sample size was small (e.g. $N = 100$).

The TP for $\text{EBE}_{\text{LASSO}}$ was always the highest regardless of sample size and number of SNPs. The TP for $\text{EBE}_{\text{APML0}}$ approach was comparable to $\text{EBE}_{\text{LASSO}}$ when sample size was 500 or above, but was slightly lower than when $N = 100$. The TP for TV-GroupSpAM was always the lowest when sample size $\geq$500. For all the methods, the TP increased with sample size but decreased with the number of SNPs.

Although the $\text{EBE}_{\text{LASSO}}$ has the best TP, it also has the worst FP. We therefore consider a composite measure of FP and TP, namely the F1 score. $\text{EBE}_{\text{APML0}}$ approach consistently provided the highest F1 scores in the majority of the simulation scenarios. This is particularly impressive given that TV-GroupSpAM artificially limited the number of FPs as it requires specification of the number of markers to be selected. In comparison, the other two approaches, i.e. $\text{EBE}_{\text{APML0}}$ and $\text{EBE}_{\text{LASSO}}$, did not require the limit of number of markers and could select as many markers as those in the dataset. Since the FPs for $\text{EBE}_{\text{LASSO}}$ were high, the F1 scores of this approach were generally lower than those of the other two methods, except for the scenarios where the SNPs are highly correlated and the sample size is small. For all the tested methods, the F1 score increased with sample size but decreased with the number of SNPs.

For all the three approaches, the computational time increased exponentially with sample size and more than exponentially with number of SNPs. TV-GroupSpAM took more than almost 11 days (~100 000s) to analyze 10 000 SNP for 1000 subjects when the SNPs were highly correlated while $\text{EBE}_{\text{LASSO}}$ and $\text{EBE}_{\text{APML0}}$ took only 40 s. Both $\text{EBE}_{\text{APML0}}$ and $\text{EBE}_{\text{LASSO}}$ required markedly less time compared to TV-GroupSpAM, improving the time efficiency by ~100–2500 folds, depending on the sample size and number of SNPs included in the analysis.

3.2 Genome-wide association study for Alzheimer's disease dynamics

We performed a genome-wide association study for Alzheimer's disease using TV-GroupSpAM, $\text{EBE}_{\text{LASSO}}$ and $\text{EBE}_{\text{APML0}}$. We examined association between SNPs and the change of ADAS11-cog score over time. ADAS11-cog score is an important assessment measure to assess the level of cognitive dysfunction in Alzheimer's disease. For this analysis, data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.usc.edu/ADNI). We performed quality control (QC) steps on the raw genotype data which removed missing records for ADAS11-cog score and SNPs with (i) >5% missing values, (ii) minor allele frequency below 0.05 and (iii) constant genotypes across all subjects. The missing values of the remaining SNPs were imputed by the mean value of the corresponding SNP. After the QC, we obtained 5335 observations from $n = 785$ subjects and their genotype data of $p = 145\,559$ SNPs on the 19th chromosomes. As the number of observations for every subject is required to be the same for TV-GroupSpAM, we only analyzed those subjects with >5 longitudinal measurements and keep the first 5 observations which end up with 3140 measurements for 28 subjects. Non-genetic factors, such as gender, length of education and age at baseline were also incorporated as static covariates into mixed-effects model and TV-GroupSpAM.

Following the procedure of Marchetti-Bowick *et al.* (2016), we performed a two-stage selection. First, we split the genotype data into 30 subsets, containing ~5000 SNPs each, and applied $\text{EBE}_{\text{LASSO}}$, $\text{EBE}_{\text{APML0}}$ and TV-GroupSpAM on each dataset. This

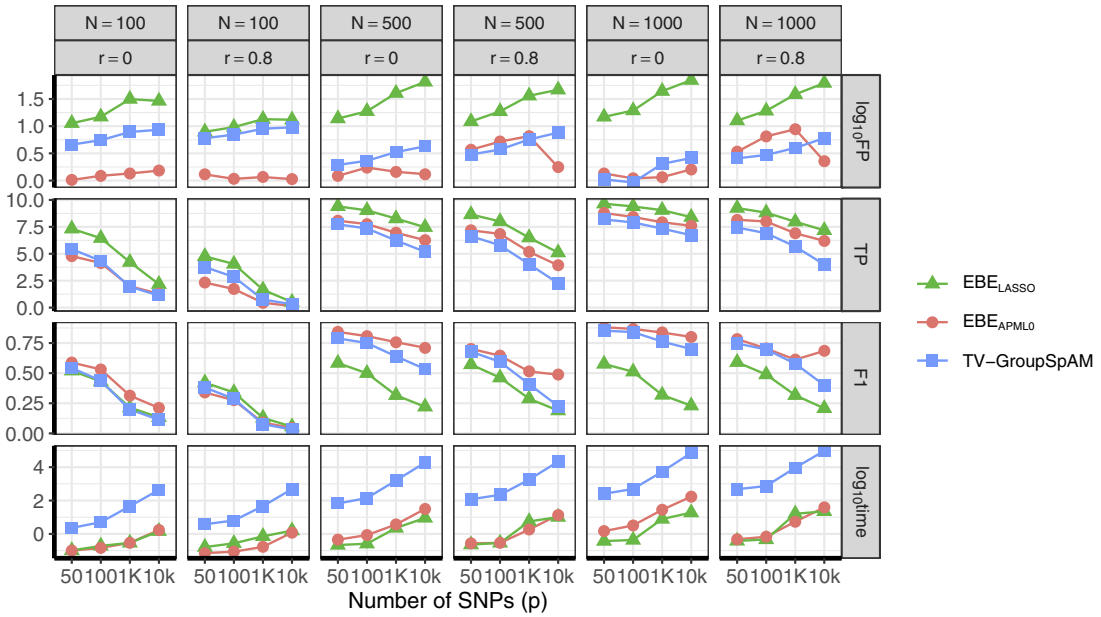

Fig. 1. Comparison of number of false positive, number of true positive, F1 score and computational time

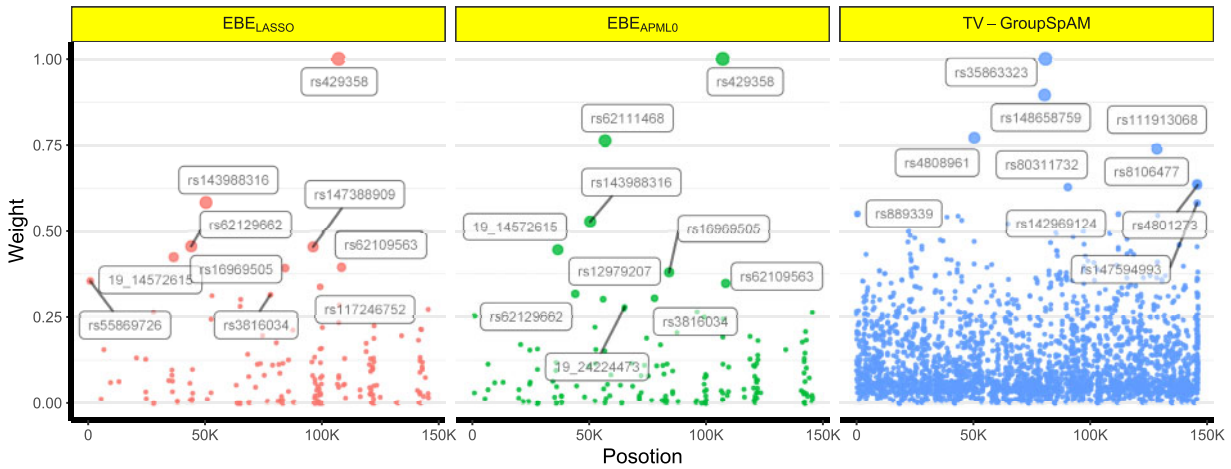

Fig. 2. Scatter plot of the normalized model weights for each SNP in the filtered set for EBE_{LASSO}, EBE_{APML0} and TV-GroupSpAM. Weight was calculated by normalizing effect size β to $\tilde{\beta} = \frac{\beta - \beta_{min}}{\beta_{max} - \beta_{min}}$

yielded a filtered set of 2943 SNPs for TV-GroupSpAM, 149 SNPs for EBE_{LASSO} and 80 SNPs for EBE_{APML0}. Figure 2 shows the model weight (an indicator of significance) of every SNP in the filtered set for each method. Next, we performed these methods again separately only with their selected SNPs in the first stage. For TV-GroupSpAM, we first selected ∼100 SNPs in the first stage and then selected ∼50 SNPs. EBE_{LASSO}, EBE_{APML0} and TV-GroupSpAM finally selected 126, 73 and 53 SNPs, respectively. Ranking these SNPs according to their fitted model weights (estimated effect sizes), the top 10 SNPs for each method are listed in Table 1. The complete sets of SNPs selected by each method are listed in Supplementary Tables S2–S4. Finally, it took TV-GroupSpAM ∼16 days to complete the analysis, whereas, EBE_{LASSO} and EBE_{APML0} only took 18.7 and 3.7 min, respectively. This suggests an impressive improvement in terms of time efficiency (∼1200- and 6200-fold) for EBE_{LASSO} and EBE_{APML0}, respectively, compared to TV-GroupSpAM.

All three methods selected SNP rs429358 located in the fourth exon of the ApoE gene, which is the most well-known genetic risk factor for Alzheimer's disease. In addition, the SNP rs7256200 located in the ApoC1 gene was also identified by both EBE_{LASSO} and EBE_{APML0}. ApoC1 has been previously reported to be associated with Alzheimer's disease by several independent studies (Zhou *et al.*, 2014). Besides, several other SNPs that are associated with diseases for the elderly were also selected. For example, rs1800468 was reported to affect osteoporosis (Langdahl *et al.*, 2008). To visualize the mean ADAS11-cog score trend for different SNP groups (the number of copies of minor allele is 0, 1 or 2), we plotted fitted locally polynomial regression (loess) curves. Figure 3 shows loess plots of three different groups: 0, 1 or 2 for rs429358 (top), rs7256200 (middle) and rs7256200 (bottom). Apparently, the time profiles of the ADAS11-cog scores were different for different genotypes. For rs429358 and rs7256200, the progression of ADAS11-cog score in subjects with one copy of minor allele appears higher faster than those with zero or two copies of minor allele. On the other hand, for SNP rs1800468, the worsening of ADAS11-cog score over time appears more severe with increasing copies of minor allele after 50 months.

We checked the predictive performance of linear mixed-effects models with different sets of gene features obtained by various methods by ADNI data. We put the top 50/20/10 genes selected by three different methods (EBE_{LASSO}, EBE_{APML0} and TV-GroupSpAM) into the full model as covariates, and calculated the average prediction

Table 1. Top 10 selected SNPs by three different approaches

SNP name	Gene	MAF
EBE_{LASSO}		
rs111677971	LOC390937	0.08981
rs35194062	RELB	0.05732
19_14572615	None	0.11210
rs150478685	None	0.05669
rs73488486	ZNF358 and LOC105372261	0.10064
rs143988316	None	0.07643
rs55869726	None	0.06369
rs3816034	LINC01837 and LINC01533	0.07325
rs11670478	ZNF460-AS1	0.05541
rs147388909	LOC107987267	0.07516
EBE_{APML0}		
rs55869726	None	0.06369
rs62111468	None	0.05669
rs62109563	ERCC1 and CD3EAP	0.07962
19_14572615	None	0.11210
rs111677971	LOC390937	0.08981
rs35194062	RELB	0.05732
rs16969505	SCGB1B2P	0.05032
rs73488486	ZNF358 and LOC105372261	0.10064
rs429358	ApoE	0.23694
rs143988316	None	0.07643
TV-GroupSpAM		
rs3212764	JAK3	0.26624
rs11667387	None	0.35414
rs892012	None	0.36115
rs1019945	NCLN	0.14904
rs62116566	LOC101927151	0.09873
rs1019946	NCLN	0.14841
rs56042840	ZNF160	0.09745
rs11670284	NLRP13	0.29490
rs7258847	NLRP13	0.25669
rs740568	None	0.35605

Note: SNPs that have not been identified to play a biological role is marked as 'None'.

errors using the 10-fold cross-validation method. Predictive performance for EBE_{LASSO}, EBE_{APML0} and TV-GroupSpAM are reported in Table 2. EBE_{APML0} and EBE_{LASSO} have comparable performance when top 50 or 20 SNPs are included in the linear mixed-effects model. EBE_{APML0} has smaller predictive error than EBE_{LASSO} when less SNPs are included. In all scenarios, EBE_{LASSO} and EBE_{APML0} have better predictive performance than TV-GroupSpAM.

4 Discussion

Analysis of longitudinal phenotypes in genomic studies can facilitate the understanding of complex disease status over time such as disease onset, progression and improvement after treatment, etc. In addition, genomic analysis based on longitudinal traits can markedly improve the power to detect rate variants or variants with relatively weak effects, particularly when multilocus approach is used. However, analyzing longitudinal traits collected over time is challenging as measurements from the same individual are often correlated, and therefore the approaches used for 'static traits' GWAS (e.g. linear regression models) are no longer valid. Furthermore, due to complex algorithms involved in the traditional statistical methods for longitudinal data (such as mixed-effects modeling), performing multilocus, high-dimensional whole-genome analysis using these methods is not computationally scalable and feasible. Most existing methods test each single SNP for association with the longitudinal phenotypes at a time point (Aulchenko *et al.*, 2010; Sikorska *et al.*, 2015) which can only identify the marginal effects of a locus, and use a conservative multiple hypothesis testing correction. Although this approach works well for traits that depend on strong effects from a few loci, when it comes to complex, polygenic traits that are influenced by weak effects from many different SNPs, it is less effective. In this article, we proposed a multilocus, high-throughput and efficient feature selection method by coupling EBEs from mixed-effects modeling with a regularized method based on ℓ_0-penalty variable selection.

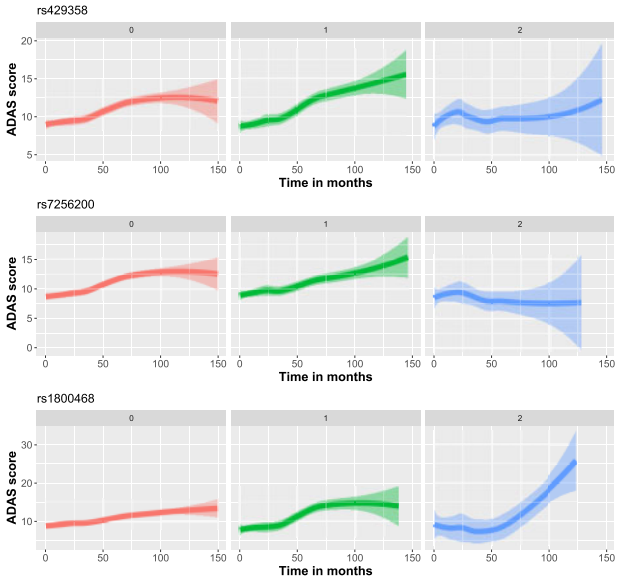

Fig. 3. The mean trend of ADAS11-cog score was plotted against time (in month) from the baseline for each group with the number of minor alleles are 0, 1, 2 (top: rs429358, middle: rs7256200, bottom: rs1800468)

Table 2. Predictive errors for linear mixed-effects models with top 50/20/10 selected SNPs with EBE_{LASSO}, EBE_{APML0} and TV-GroupSpAM, respectively

Predictive error	EBE_{APML0}	EBE_{LASSO}	TV-GroupSpAM
Top 50 SNPs included	54.01212	54.43673	59.75249
Top 20 SNPs included	53.97637	53.26435	57.47443
Top 10 SNPs included	55.95391	59.94889	63.50255

The EBE-based approach converts multilocus GWAS for longitudinal data to standard GWAS as only linear regression is involved, providing a simple solution for implementation of multilocus GWAS approach for longitudinal traits. As the EBEs tend to be shrunk to the corresponding population mean estimate, there is a concern that shrinkage may mask existing parameter–covariate relationships (Meirelles *et al.*, 2013; Sikorska *et al.*, 2013). However, our application to ADNI data shows that the proposed approaches have better predictive performance than the existing approach (TV-GroupSpAM) with smaller predictive error. In addition, our recent research revealed that association tests based on EBEs had comparable performance compared to the standard approach using likelihood ratio test within the framework of mixed-effects modeling (i.e. almost identical power for detecting true covariate effects, and slightly better control of the FP rate).

The 2HiGWAS (Jiang *et al.*, 2015) provides a high-dimensional varying coefficient model to chart a complete picture of the genetic architecture of complex traits that are dynamically expressed on a time-space scale. It is a two-step procedure that is dimension reduction by prescreening and predictor selection through penalized regression. In our revision, we compared 2HiGWAS with the proposed approach in terms of the FP, TP, F1 score and the running

time. The 2HiGWAS approach has a better control of FP rates compared to EBE_{APML0}. The tradeoff was that the true positive rates of 2HiGWAS are lower than those of EBE_{APML0}. Taking both FP and TP into consideration, the EBE_{APML0} had comparable and slightly better F1 scores than the 2HiGWAS approach. Finally, EBE_{APML0} is far more time efficient than 2HiGWAS (∼500- to 6000-fold). For example, for r is 0.4 and $\sigma_\gamma^2 = 0.09$, $\sigma_\beta^2 = 10^{-6}$ EBE_{APML0} only spends 28.33 s while it cost 2HiGWAS >8.5 h.

ℓ_0-norm is preferred in terms of variable selection as it can provide optimal power and control of FPs (Li *et al.*, 2018). However, due to computational challenges (non-convexity and discontinuity; Barron *et al.*, 1999; Davis *et al.*, 1997; Manyem and Ugon, 2012), currently most common approach is ℓ_1-norm regularization (or Lasso regression) proposed by Tibshirani (1996), which is a convex relaxation to ℓ_0-regularization and can provide good predictive power. In this article, we updated a recently developed two-stage procedure for ℓ_0-penalty variable selection (Li *et al.*, 2018). Our two-stage ℓ_0-norm regularization approach overcomes instability and inconsistency suffered by Lasso regression and provides much better control of FPs without sacrificing power to detect true effects.

Taking advantages of the simplicity of EBE-based approaches and the superior properties for variable selection from our ℓ_0-norm regression, the proposed novel EBE_{APML0} approach allows fast, accurate and multilocus GWAS analyses for complex longitudinal traits. Compared to the most recently proposed approach for longitudinal traits (TV-GroupSpAM), EBE_{APML0} not only provides greater power detecting true active SNPs but also markedly shortens the analysis time by >1000 times. It only takes hours for EBE_{APML0} to perform a full genome-wide multilocus GWAS (e.g. millions of SNPs) compared to months if TV-GroupSpAM is used. Furthermore, EBE_{APML0} provides consistent and robust control FPs genetic variants, whereas FP from TV-GroupSpAM varied with sample size and number of SNPs. Although TV-GroupSpAM was able to control FP rate when sample size is large ($N \geq 500$), but as the tradeoff, the power and TP detection rate was apparently lower than that from EBE_{APML0}.

In addition, due to its simplicity, a real genome-wide multilocus GWAS for longitudinal traits becomes possible with EBE_{APML0}. For example, if computing capability allows, i.e. sufficient memory, it is possible to conduct multilocus GWAS based on longitudinal traits for millions of SNPs together. However, other methods such as TV-GroupSpAM and 2HiGWAS (Jiang *et al.*, 2015) are so time-consuming that it takes hours to deal with 1000 SNPs together, let alone millions of SNPs. A true multilocus GWAS is important as it takes into account the correlation among SNPs and could effectively reduce the confounding. However, it should be mentioned that the statistical power tended to decrease with increasing number of SNPs included in the model at once. In this analysis, we also attempted to perform the EBE_{LASSO} and EBE_{APML0} by including all the 145 559 SNPs on chromosome 19 together at once, only 4 and 2 SNPs were selected, respectively (it is worth noting that both methods were still be able to select ApoE4, the most known Alzheimer's Disease-related SNP in this case). Nevertheless, with emerging of large electronic health record data and large-scale genetic data from biobanks, we may be able to attain sufficient power when including all the SNPs in the model at the same time in the future.

Though in the present study, we limited ourselves linear mixed-effects modeling, EBE_{APML0} can be easily extended to non-linear, functional dynamic data via linearization, Taylor expansion for example. Our research opens the door for efficient and scalable functional GWAS for more complex non-linear longitudinal traits. Over the last decade, different approaches have been attempted for non-linear GWAS of longitudinal outcomes (Das *et al.*, 2011, 2013; Hou *et al.*, 2008; Li *et al.*, 2015; Marchetti-Bowick *et al.*, 2016). However, these methods are extremely time-consuming and require months to scan the whole genome.

Finally, we also limit ourselves to association tests without looking into the estimate of effect size as the primary task for GWAS is testing the association. Since EBE_{APML0} represents a high-throughput approach that can provide very high power detecting true, active SNPs and control the FPs in a consistent way, estimation of time-varying effects or prediction of trajectory of traits can be performed using standard traditional approaches after we identify the active SNPs.

Funding

This work was supported by the National Science Foundation of China (NSFC) [11671375] and Doctoral research funding of Anhui Medical University [XJ201710].

Conflict of Interest: none declared.

References

Aulchenko,Y.S. *et al.* (2010) ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics*, **11**, 134.

Barron,A. *et al.* (1999) Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields*, **113**, 301–413.

Chiu,Y.-F. *et al.* (2016). Longitudinal analytical approaches to genetic data. *BMC Genetics*, **17**, S4.

Combes,F. *et al.* (2014) Powers of the likelihood ratio test and the correlation test using empirical Bayes estimates for various shrinkages in population pharmacokinetics. *CPT Pharmacometrics Syst. Pharmacol.*, **3**, 1–9.

Das,K. *et al.* (2011) A dynamic model for genome-wide association studies. *Hum. Genet.*, **129**, 629–639.

Das,K. *et al.* (2013) Dynamic semiparametric Bayesian models for genetic mapping of complex trait with irregular longitudinal data. *Stat. Med.*, **32**, 509–523.

Davis,G. *et al.* (1997) Adaptive greedy approximations. *Constr. Approx.*, **13**, 57–98.

Furlotte,N.A. *et al.* (2012) Genome-wide association mapping with longitudinal data. *Genet. Epidemiol.*, **36**, 463–471.

Hou,W. *et al.* (2008) A nonlinear mixed-effect mixture model for functional mapping of dynamic traits. *Heredity*, **101**, 321–328.

Jiang,L. *et al.* (2015) 2HiGWAS: a unifying high-dimensional platform to infer the global genetic architecture of trait development. *Brief. Bioinform.*, **16**, 905–911.

Langdahl,B.L. *et al.* (2008) Large-scale analysis of association between polymorphisms in the transforming growth factor beta 1 gene (TGFB1) and osteoporosis: the GENOMOS study. *Bone*, **42**, 969–981.

Li,J. *et al.* (2011) The Bayesian lasso for genome-wide association studies. *Bioinformatics*, **27**, 516–523.

Li,J. *et al.* (2015) Bayesian group lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies. *Ann. Appl. Stat.*, **9**, 640–664.

Li,X. *et al.* (2018) Efficient A_0-norm feature selection based on augmented and penalized minimization. *Stat. Med.*, **37**, 473–486.

Li,Z. and Sillanpää,M.J. (2013) A Bayesian nonparametric approach for mapping dynamic quantitative traits. *Genetics*, **194**, 997–1016.

Londono,D. *et al.* (2013) A novel method for analyzing genetic association with longitudinal phenotypes. *Stat. Appl. Genet. Mol. Biol.*, **12**, 241–261.

Manyem,P. and Ugon,J. (2012) Computational complexity, NP completeness and optimization duality: a survey. *Electronic Colloquium on Computational Complexity (ECCC)*, **19**.

Marchetti-Bowick,M. *et al.* (2016) A time-varying group sparse additive model for genome-wide association studies of dynamic complex traits. *Bioinformatics*, **32**, 2903–2910.

Meirelles,O.D. *et al.* (2013) SHAVE: shrinkage estimator measured for multiple visits increases power in GWAS of quantitative traits. *Eur. J. Hum. Genet.*, **21**, 673–679.

Natarajan,B.K. (1995) Sparse approximate solutions to linear systems. *SIAM J. Comput.*, **24**, 227–234.

Savic,R.M. and Karlsson,M.O. (2009) Importance of shrinkage in empirical Bayes estimates for diagnostics: problems and solutions. *AAPS J.*, **11**, 558–569.

Sikorska,K. *et al.* (2013) GWAS on your notebook: fast semi-parallel linear and logistic regression for genome-wide association studies. *BMC Bioinformatics*, **14**, 166.

Sikorska,K. *et al.* (2015) GWAS with longitudinal phenotypes: performance of approximate procedures. *Eur. J. Hum. Genet.*, **23**, 1384–1391.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Methodol.*, **58**, 267–288.

Wu,T.T. *et al*. (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.

Xu,X.S. *et al*.; Alzheimer's Disease Neuroimaging Initiative. (2013) Mixed-effects beta regression for modeling continuous bounded outcome scores using NONMEM when data are not on the boundaries. *J. Pharmacokinet. Pharmacodyn.*, **40**, 537–544. and

Xu,X.S. *et al*. (2017) Further evaluation of covariate analysis using empirical Bayes estimates in population pharmacokinetics: the perception of shrinkage and likelihood ratio test. *AAPS J.*, **19**, 264–273.

Yang,J. *et al*. (2009) Nonparametric functional mapping of quantitative trait loci. *Biometrics*, **65**, 30–39.

Zhou,Q. *et al*. (2014) Association between APOC1 polymorphism and Alzheimer's disease: a case-control study and meta-analysis. *PLoS One*, **9**, e87017.