Efficient mixed model approach for large-scale genome-wide association studies of ordinal categorical phenotypes

Wenjian Bi,[1,2,*] Wei Zhou,[3,4,5] Rounak Dey,[6] Bhramar Mukherjee,[1] Joshua N. Sampson,[7] and Seunggeun Lee[1,2,8,*]

Summary

In genome-wide association studies, ordinal categorical phenotypes are widely used to measure human behaviors, satisfaction, and preferences. However, because of the lack of analysis tools, methods designed for binary or quantitative traits are commonly used inappropriately to analyze categorical phenotypes. To accurately model the dependence of an ordinal categorical phenotype on covariates, we propose an efficient mixed model association test, proportional odds logistic mixed model (POLMM). POLMM is computationally efficient to analyze large datasets with hundreds of thousands of samples, can control type I error rates at a stringent significance level regardless of the phenotypic distribution, and is more powerful than alternative methods. In contrast, the standard linear mixed model approaches cannot control type I error rates for rare variants when the phenotypic distribution is unbalanced, although they performed well when testing common variants. We applied POLMM to 258 ordinal categorical phenotypes on array genotypes and imputed samples from 408,961 individuals in UK Biobank. In total, we identified 5,885 genome-wide significant variants, of which, 424 variants (7.2%) are rare variants with MAF < 0.01.

Introduction

Large-scale biobanks with hundreds of thousands of genotyped and extensively phenotyped subjects are valuable resources to identify genetic components of complex phenotypes.[1,2] In biobanks, ordinal categorical data, which are often collected from surveys, questionnaires, and testing to measure human behaviors, satisfaction, and preferences, are a common type of phenotype.[3–5] For example, a web questionnaire was used for 182,219 UK Biobank participants to collect 150 food and other health behavior-related preferences, all of which are ordinal categorical phenotypes based on a 9-point hedonic scale of liking from 1 (extremely dislike) to 9 (extremely like).[6] For ordinal categorical phenotypes, there is no underlying measurable scale, and therefore, it would be inappropriate to treat that phenotype as a quantitative trait and apply the linear regression methods.[7–9] Another approach is to use an arbitrary cutoff to dichotomize the ordinal categorical phenotype into two categories, followed by using a logistic regression method.[3] This approach suffers from information loss and, thus, is less powerful.

For binary and quantitative phenotype data analysis, mixed model approaches have been widely used to test genetic associations conditioning on the sample relatedness.[7,10] Some state-of-art optimization strategies have been applied to reduce memory usage and computational cost, which makes these mixed model approaches practical for incorporating a dense genetic relationship matrix (GRM) in genome-wide association studies (GWASs).[9,11] Another resource-efficient approach, fastGWA, is to use a sparse GRM to adjust for the sample relatedness.[12] For binary phenotype analysis, unbalanced case-control ratio can result in inflated type I error rates, and saddlepoint approximation (SPA) has been demonstrated to be more accurate for single-variant analysis,[8,9] region-based analysis,[13,14] and gene-environment interaction analysis.[15] Similarly, the sample size distribution in ordinal categorical data could also be highly unbalanced; that is, the sample size in one category could be dozens of times more than that in other categories. For example, of the UK Biobank participants, more than 90% extremely dislike cigarette smoking and only 1% extremely like it. In ordinal categorical data analysis, the effect of the unbalanced sample size distribution on genetic association tests should also be carefully examined.

In this paper, we propose a scalable and accurate mixed model approach for ordinal categorical data analysis in large-scale GWASs. Our approach, proportional odds logistic mixed model (POLMM), incorporates a random effect into the proportional odds logistic model to control for sample relatedness. POLMM uses penalized quasi-likelihood (PQL) and average information restricted maximum likelihood (AI-REML) algorithms[7] to efficiently fit the

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA; [2]Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA; [3]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA; [4]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [5]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [6]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; [7]Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Bethesda, MD 20892, USA; [8]Graduate School of Data Science, Seoul National University, Seoul 08826, Republic of Korea
*Correspondence: wenjianb@umich.edu (W.B.), lee7801@snu.ac.kr (S.L.)
https://doi.org/10.1016/j.ajhg.2021.03.019.

mixed model and then uses SPA to calibrate p values. We give two closely related versions, DensePOLMM and FastPOLMM. To control for the genetic relatedness between samples, DensePOLMM incorporates a dense GRM and FastPOLMM is a resource-efficient approach that uses a sparse GRM.

We demonstrated that POLMM approaches can efficiently analyze large datasets with hundreds of thousands of genetic related samples, can control type I error rates, and is statistically powerful through extensive simulations as well as real data analysis. Meanwhile, BOLT-LMM, fastGWA, and SAIGE approaches cannot control type I error rates and are less powerful, especially when the phenotypic distribution is unbalanced. DensePOLMM requires comparable computation time and memory usage to SAIGE, and FastPOLMM is more resource-efficient to fit a null mixed model. For example, FastPOLMM requires less than 0.1 h and 4.2 GB memory to fit a null mixed model with around 400,000 subjects. In most scenarios, DensePOLMM and FastPOLMM performed similarly in terms of testing. Only when the number of categories is large (e.g., 10) and polygenic effect size is large (e.g., liability heritability = 75.24%), DensePOLMM is slightly more powerful than FastPOLMM by no more than 4.67% and 7.51% when testing common (minor allele frequency, MAF = 0.3) and low-frequency variants (MAF = 0.01), respectively. We applied the FastPOLMM approach to analyze 258 ordinal categorical phenotypes in the UK Biobank data, which includes 408,961 samples from white British participants with European ancestry, and successfully identified 5,885 distinct genome-wide significant variants with clumping, of which, 424 variants (7.2%) are rare variants with MAF $<$ 0.01. All analysis results have been publicly available through a web-based visual server,[2] which provides intuitive visualizations at three levels of granularity: genome-wide summaries at the trait level and regional (LocusZoom)[16] and phenome-wide summaries at the variant level.

Material and methods

Overview of the POLMM method

The POLMM method contains two main steps: (1) fitting the null mixed model to estimate the variance component and model parameters corresponding to covariates and (2) testing for the association between each single genetic variant and ordinal categorical phenotypes. In step 1, we include covariates such as age, gender, and top SNP-derived principal components (PCs) to adjust for their effects on the phenotype. Then, we save the null model fitting results (including the residuals from the null model) in an R object. In step 2, we load the R object and use it for association testing. This strategy only requires one model fitting across a genome-wide analysis, which greatly reduces computation time.

Proportional odds logistic mixed model

We let n denote the sample size and let J denote the number of category levels. For subject $i \le n$, we let $y_i = 1, 2, \ldots, J$ denote its ordinal categorical phenotype. We consider the following proportional odds logistic mixed model (POLMM)

$$\text{logit}(\nu_{ij}) = \varepsilon_j - \eta_i = \varepsilon_j - X_i^T\beta - G_i\gamma - b_i, 1 \le i \le n, 1 \le j \le J, \quad \text{(Equation 1)}$$

where $\nu_{ij} = \Pr(y_i \le j | X_i, G_i, b_i)$ is the cumulative probability of the phenotype $y_i \le j$ conditional on a p-dimensional vector of covariates X_i and a hard called or imputed genotype G_i. The cutpoints ε: $\varepsilon_1 < \ldots < \varepsilon_J = \infty$ were used to categorize the data, and coefficients β and γ are fixed effect sizes of the covariates and genotype. To adjust for sample relatedness, we incorporate an n-dimensional random effect vector $b = (b_1, \cdots, b_n)^T$ following a multivariate normal distribution $N(0, \tau V)$ where τ is a variance component parameter and V is an $n \times n$ dimensional GRM. Equation 1 is a natural extension of a logistic mixed model as in SAIGE and GMMAT.[7,9,14] If $J = 2$, the phenotype is binary and Equation 1 is a logistic mixed model. Although POLMM is based on the proportional odds assumption, previous studies indicate that it could still be valid with respect to tests when the assumption is violated.[17] In "numeric simulations," we validate that POLMM could still control type I error rates when the ordinal categorical phenotypes were simulated following category logistic model and stereotype model.

For subject i, we define a $J \times 1$ vector $\tilde{y}_i = (y_{i1}, \cdots, y_{iJ})^T$ as an equivalent representation of the ordinal categorical phenotype y_i: if $y_i = j$, then $y_{ij} = 1$ and the other elements in $\tilde{y}_i$ are 0. The conditional log-likelihood function given random effects b is

$$l_i(\beta, \gamma; b, \varepsilon) = \log(\Pr(y_i)) = \sum_{j=1}^{J} y_{ij} \log(\mu_{ij}),$$

where μ_{ij} is the mean of y_{ij}; that is

$$\mu_{ij} = E(y_{ij}) = \Pr(y_{ij} = 1) = \Pr(y_i = j) = \Pr(y_i \le j) - \Pr(y_i \le j - 1).$$

Because random vector b follows a multivariate normal distribution $N(0, \tau V)$, the marginal log-likelihood function of (β, γ, τ) is

$$l(\beta, \gamma, \tau; \varepsilon) = \log \int \exp\{l(\beta, \gamma; b, \varepsilon)\} \times (2\pi)^{-\frac{n}{2}} |\tau V|^{-\frac{1}{2}} \times \exp\left\{ -\frac{1}{2} b^T (\tau V)^{-1} b \right\} db,$$

where log-likelihood function $l(\beta, \gamma; b, \varepsilon) = \sum_{i \le n} l_i(\beta, \gamma; b, \varepsilon)$. In Appendix A, we follow a similar framework as in GMMAT[7] to use PQL and AI-REML to simultaneously estimate the variance component $\hat{\tau}$ and other parameters $(\hat{\beta}, \hat{\varepsilon})$ that maximize $l(\beta, \gamma, \tau; \varepsilon)$ under the null model $\gamma = 0$. It is well known that PQL can generate a biased estimate for the variance component,[9,18,19] but as shown in literature, the bias does not inflate type I error rates in association tests.[7,9] Similarly, POLMM also has a biased estimate of the variance component. Through extensive simulation studies and real data analysis, we show that the bias does not inflate type I error rates.

Score test and estimated variance

We let $\hat{\mu}_{ij}$ and $\hat{\nu}_{ij}$ be the fitted value μ_{ij} and ν_{ij} under the null hypothesis $\gamma = 0$, respectively. The score is

$$T = \sum_{i=1}^{n} \sum_{j=1}^{J-1} \left[G_i R_{ij} \cdot \left(y_{ij} - \hat{\mu}_{ij} \right) \right],$$

where

$$R_{ij}=\frac{1}{\widehat{\mu}_{ij}}\cdot\left(\widehat{\nu}_{i(j-1)}\cdot\left(1-\widehat{\nu}_{i(j-1)}\right)-\widehat{\nu}_{ij}\cdot\left(1-\widehat{\nu}_{ij}\right)\right)$$
$$-\frac{1}{\widehat{\mu}_{ij}}\cdot\left(\widehat{\nu}_{i(J-1)}\cdot\left(1-\widehat{\nu}_{i(J-1)}\right)-\widehat{\nu}_{ij}\cdot\left(1-\widehat{\nu}_{ij}\right)\right).$$

Because that R_{ij} and $\widehat{\mu}_{ij}$ are estimated under the null model and are the same for all variants, it takes n computations to calculate the score T for any variant. The estimated variance of the score is $\widehat{Var}(T)=\overline{G}^T\tilde{Z}^T P\tilde{Z}\overline{G}$ where n-dimensional covariate-adjusted genotype vector

$$\overline{G}=G-X\left(X^T\tilde{Z}^T R\Psi R\tilde{Z}X\right)^{-1}X^T\tilde{Z}^T R\Psi R\tilde{Z}G, G=(G_1,G_2,\ldots,G_n)^T, X=(X_1,X_2,\ldots,X_n)^T,$$

$R=diag(R_{11},\cdots,R_{1(J-1)},\cdots,R_{n1},\cdots,R_{n(J-1)})$ is an $n(J-1)\times n(J-1)$ diagonal matrix, and $\tilde{Z}=(e_1,\cdots,e_1,\cdots,e_n,\cdots,e_n)^T$ is an $n(J-1)\times n$ matrix where e_i denotes an $n\times 1$ vector with a 1 in the i-th coordinate and zeroes elsewhere. We let $n(J-1)\times n(J-1)$ block diagonal matrix Ψ denote the covariance matrix of $\tilde{y}=(y_{11},\cdots,y_{1(J-1)},\cdots,y_{n1},\cdots,y_{n(J-1)})^T$ as follows:

$$\Psi=\begin{bmatrix}\Psi_1 & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \Psi_n\end{bmatrix},\Psi_i=\begin{bmatrix}\widehat{\mu}_{i1} & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \widehat{\mu}_{i(J-1)}\end{bmatrix}-\widehat{\mu}_i\widehat{\mu}_i^T,\widehat{\mu}_i=\left(\widehat{\mu}_{i1},\cdots,\widehat{\mu}_{i(J-1)}\right)^T.$$

The $n(J-1)\times n(J-1)$ dimensional matrix $P=\Sigma^{-1}-\Sigma^{-1}\tilde{Z}X\left(X^T\tilde{Z}^T\Sigma^{-1}\tilde{Z}X\right)^{-1}X^T\tilde{Z}^T\Sigma^{-1}$ where $\Sigma=R^{-1}\Psi^{-1}R^{-1}+\tau\tilde{V}$ and $\tilde{V}=\tilde{Z}V\tilde{Z}^T$. To estimate $\widehat{Var}(T)$, we must calculate $\Sigma^{-1}\overline{G}$, which is computationally expensive for a genome-wide analysis. To reduce the computation cost, we use the same strategy as in BOLT-LMM[11] and SAIGE.[9] First, we use a small number of variants to calculate $\widehat{Var}(T)$ and $\widehat{Var}^*(T)=\overline{G}^T\tilde{Z}^T R\Psi R\tilde{Z}\overline{G}$ and estimate ratio $\widehat{r}$ by using the mean of $\widehat{Var}(T)/\widehat{Var}^*(T)$. Then, for each variant to test, we calculate $\widehat{Var}^*(T)$ and then estimate $\widehat{Var}(T)=\widehat{r}\cdot\widehat{Var}^*(T)$. The ratio has been shown approximately constant for all genetic variants with minor allele count (MAC) ≥ 20.[9,11] When estimating $\widehat{r}$, we increase the number of variants until the coefficient of variation for the ratio estimation is lower than a pre-given cutoff of 0.0025. In both simulation studies and real data analysis, the variant number is usually less than 30. Using optimized strategies, it takes $O(n)$ computations to the calculate $\widehat{Var}^*(T)$ for each variant. More details about the score test and the estimated variance can be seen in Appendix B.

Saddlepoint approximation

The regular score test assumes that T asymptotically follows a normal distribution, which uses only the first two moments. However, when the sample size distribution of different categories is highly unbalanced, the underlying distribution of T could be substantially different from a normal distribution, especially when testing low-frequency variants. To accurately calculate p values, we use SPA, which uses the entire cumulant generating function (CGF) to approximate the null distribution. Suppose that $\overline{G}_i$ is the i-th element in vector $\overline{G}$, we define

$$T_i=\sum_{j=1}^{J-1}\frac{\overline{G}_i R_{ij}\left(y_{ij}-\widehat{\mu}_{ij}\right)}{\sqrt{\overline{G}^T R\Psi R\overline{G}}}=\sum_{j=1}^{J-1}c_{ij}y_{ij}-\sum_{j=1}^{J-1}c_{ij}\widehat{\mu}_{ij}, c_{ij}=\frac{\overline{G}_i R_{ij}}{\sqrt{\overline{G}^T R\Psi R\overline{G}}},$$

then the statistic

$$T_{adj}=\frac{T}{\sqrt{\widehat{Var}(T)}}=\frac{1}{\sqrt{\widehat{r}}}\cdot\frac{T}{\sqrt{\widehat{Var}^*(T)}}=\frac{1}{\sqrt{\widehat{r}}}\cdot\sum_{i=1}^{n}T_i.$$

Because y_{ij} follows a Berounlli $(\widehat{\mu}_{ij})$ distribution, the CGF of T_i is

$$K_i(t)=\log\left[E\left(e^{tT_i}\right)\right]=\log\left(1-\sum_{j=1}^{J-1}\widehat{\mu}_{ij}+\sum_{j=1}^{J-1}e^{c_{ij}t}\widehat{\mu}_{ij}\right)-\left(\sum_{j=1}^{J-1}c_{ij}\widehat{\mu}_{ij}\right)t.$$

We use $K(t)=\sum_{i=1}^{n}K_i(t)$ to approximate CGF of T_{adj} such that the variance from CGF is 1; that is, $K''(0)=1$. The distribution of T_{adj} at the observed test statistic q can be approximated by

$$\Pr(T_{adj}<q)\approx F(q)=\Phi\left(w+\frac{1}{w}\log\left(\frac{v}{w}\right)\right),$$

where

$$w=sign(\widehat{\zeta})\sqrt{2\{\widehat{\zeta}q-K(\widehat{\zeta})\}}, v=\widehat{\zeta}\sqrt{K''(\widehat{\zeta})},$$

and $\widehat{\zeta}$ is the solution of the equation $K'(\zeta)=q$.

We apply a hybrid strategy: if $|T_{adj}|<2$, p values are calculated on the basis of normal approximation in which the variance is $\widehat{Var}(T)=\widehat{r}\cdot\widehat{Var}^*(T)$; if $|T_{adj}|\geq 2$, p values are calculated on the basis of SPA. Using this hybrid strategy, we can greatly reduce computation time while controlling type I error rates. In addition, using the fact that many elements of G are zeroes (i.e., homozygous major genotypes), we use a fast partially normal approximation method to speed up the computation. Suppose that m subjects have at least one minor allele each and the rest have homozygous major genotypes, the fast SPA takes $O(m(J-1))$ computations to calculate the CGF and its derivatives. More details about the SPA can be seen in Appendix B.

DensePOLMM and FastPOLMM

For quantitative trait analysis, Jiang et al. have demonstrated that using a sparse GRM can reduce computational time and memory usage while still being reliable to control type I error rates.[12] However, using a sparse GRM can be less powerful than using a dense GRM because a sparse GRM cannot incorporate polygenic effects. In this paper, we present two closely related versions of POLMM methods to test the null model $\gamma=0$: DensePOLMM and FastPOLMM.

DensePOLMM and FastPOLMM use dense and sparse GRMs to adjust for sample relatedness, respectively. To make DensePOLMM computationally practical for studies with large sample size n, we use strategies as in BOLT-LMM[11] and SAIGE[9] to reduce computation time and memory cost. Instead of storing an $n\times n$ dimensional dense GRM, we compactly store raw genotypes of the genetic variants into a bitwise binary vector and use them when a dense GRM is needed. When fitting the null mixed model and estimating variance $\widehat{Var}(T)$, we need to solve linear system $\Sigma\cdot x=u$, which is challenging because Cholesky decomposition takes $O(n^3)$ computation and very large memory space to invert matrix Σ. For a given vector u, we use a preconditioned conjugate

gradient (PCG) approach[9] to directly calculate $\Sigma^{-1}u$. To make the convergence faster, we use a block diagonal matrix $Q = \mathrm{diag}(Q_1, \cdots, Q_n)$ as the preconditioner matrix, where $(J-1)\times(J-1)$ matrix $Q_i = R_i^{-1}\Psi_i^{-1}R_i^{-1} + \tau V_{ii}\cdot 1_{J-1}1_{J-1}^T$, $(J-1)\times(J-1)$ dimensional matrix $R_i = \mathrm{diag}(R_{i1}, \cdots, R_{i(J-1)})$, and $(J-1)$ dimensional vector of ones $1_{J-1} = (1, 1, \cdots, 1)^T$. Given the same tolerance criterion as in SAIGE, PCG in POLMM usually takes 6–8 iterations to converge, which is ~1.5 times more than that in SAIGE. This might be because we use a block diagonal matrix in which each block corresponds to one subject as the preconditioner matrix. When updating variance component $\hat{\tau}$, we estimate $\mathrm{tr}[P\check{V}]$ by using Hutchinson's randomized trace estimator, $\sum_{i=1}^{n_R} z_i^T P\check{V} z_i$, where $z_1, \cdots, z_{n_R}$ are n_R independent random vectors whose elements are i.i.d. Rademacher random variables.[20] In addition, we use Intel Threading Building Blocks (TBB) implemented in the RcppParallel package for the multi-threading computation (see web resources). Using these strategies, DensePOLMM is of the same computation complexity as SAIGE[9] and requires memory usage $m_1 n/4$, where m_1 is the number of markers used to construct a GRM and n is the sample size. On the other hand, FastPOLMM uses a sparse GRM in which all of the small off-diagonal elements (for example, those <0.05) are set to 0. GCTA software[21] provides an efficient tool to calculate the GRM for a large-scale dataset. The sparse GRM only needs to be calculated once for one cohort study or biobank.

Leave-one-chromosome-out scheme

To avoid contamination for correlated markers, we implemented an option to apply the leave-one-chromosome-out (LOCO) scheme for DensePOLMM and FastPOLMM methods. If the LOCO scheme is used, we first use all variants to estimate the variance component $\hat{\tau}$, and then for each chromosome, we updated the estimation of $\hat{\beta}, \hat{b}$, and $\hat{\varepsilon}$ after excluding all variants in the same chromosome. This strategy is the same as SAIGE and BOLT-LMM. For FastPOLMM, we first used the tool GCTA to calculate the GRM for each chromosome and then combined them to calculate GRMs.

Liability threshold model and liability heritability

Equation 1 is equivalent to the following liability threshold model

$$z_i = \eta_i + \delta_i = X_i^T\beta + G_i\gamma + b_i + \delta_i,$$

where z_i is a latent variable and error term δ_i follows a logistic distribution with a location parameter of 0 and a scale parameter of 1. The n-dimensional random effect vector $b = (b_1, \cdots, b_n)^T$ follows a multivariate normal distribution $N(0, \tau V)$, where τ is a variance component parameter and V is an $n\times n$ dimensional GRM. The ordinal categorical phenotype $y_i = j$ if the latent variable z_i is between cutpoints ε_{j-1} and ε_j. The variances of b_i and δ_i are τ and $\pi^2/3$, respectively. Hence, similar to SAIGE,[9] we define a liability heritability $h^2_{liab} = \tau/(\tau+\pi^2/3)$. Variance components $\tau = 1$ and 10 correspond to liability heritability h^2_{liab} = 23.3% and 75.2%, respectively.

Numeric simulations

To evaluate the computational efficiency and memory usage of the proposed methods, we randomly sampled subjects from white British UK Biobank participants to analyze an ordinal categorical phenotype, able to confide, which consists of six levels (Figure S1). We excluded 11,163 subjects whose answer was "do not know" or "prefer not to answer" and analyzed 397,798 white British participants. We used 340,447 markers to construct the GRM and incorporated six covariates of sex, birth year, and top four SNP-derived principal components to fit the null mixed model. We compared five methods, including fastGWA, BOLT-LMM, SAIGE, DensePOLMM, and FastPOLMM. Besides the raw phenotype with six categories, we combined some levels to make a new phenotype with three categories to comprehensively evaluate POLMM methods (see Figure S1). For fastGWA and BOLT-LMM, we treated the ordinal categorical phenotype as a quantitative trait from 1 to 6. For SAIGE, we dichotomized the phenotype to a binary phenotype (see Figure S1). For fastGWA and FastPOLMM, we set the cutoff of the sparse GRM at 0.05. All analyses were conducted on CPU cores of Intel Xeon Gold 6138 at 2.00 GHz. In step 1, we used eight CPU cores and recorded the computation time. For SAIGE, fastGWA, and POLMM methods, the null mixed model fitting result can be saved and used for association testing. Hence, the genotype data to test can be divided into multiple chunks for parallel computation. In step 2, we used one CPU core and recorded the computation time. For BOLT-LMM, the model fitting and association testing cannot be separately implemented. We extracted "the time for streaming genotypes and writing output" from log files to record the computation time in step 2. Because FastPOLMM and DensePOLMM are the same when testing genetic association effect, we only recorded the computation time of DensePOLMM in step 2.

We carried out extensive simulations to investigate type I error rates and powers of POLMM approaches. We simulated genotypes of 10,000 subjects in 1,000 families on the basis of the pedigree shown in Figure S2, in which each family included 10 subjects. We performed gene-dropping simulations.[22] First, we simulated a set of "pseudo" sequences, each of which included 10,000 independent variants. Then, we used these sequences as founder haplotypes that propagated through the pedigree of 10 family members. To construct the GRM for mixed model methods, we simulated 100,000 independent variants by using the same gene-dropping scheme with MAFs ranging from 0.05 to 0.5. The estimated kinship coefficients are shown in Figure S3. For subject i, two covariates X_{i1} and X_{i2} were simulated following the standard normal distribution and a Bernoulli (0.5) distribution, respectively. Given the variance component τ, random effects $b = (b_1, b_2, \cdots, b_n)$ were simulated following a multivariate normal distribution $N(0, \tau V)$ where V is the GRM from the family structure. We followed Equation 1 to simulate ordinal categorical phenotypes by using linear predicator $\eta_i = 0.5\cdot X_{i1} + 0.5\cdot X_{i2} + \gamma\cdot G_i + b_i$, $i \le n$, in which G_i is the genotype value of one variant. We considered two common types of phenotypic distribution, bell-shaped distribution and L-shaped distribution (Figure S4), and selected cutpoints ε to correspond to the given phenotypic distribution. Under the null model $\gamma = 0$, we considered three variance components $\tau = 0.5$, 1, and 2 to evaluate type I error rates at a significance level $\alpha = 5\times10^{-8}$. For each phenotypic distribution, we simulated 100 datasets of phenotypes and covariates. We considered common, low-frequency, and rare variants with MAFs of 0.3, 0.01, and 0.005, respectively. For each MAF, we simulated 10^7 variants. Thus, for each pair of phenotypic distribution and MAF, in total 10^9 tests were performed. Under the alternative model $\gamma \neq 0$, we considered the variance component $\tau = 1$ and increased genetic effect size γ to evaluate empirical powers at a significance level $\alpha = 5\times10^{-8}$. For each γ, we simulated 200 datasets including ordinal

categorical phenotypes, covariates, and genotypes of one causal variant.

In addition to DensePOLMM and FastPOLMM, which use a hybrid of normal distribution approximation and SPA, we also evaluated DensePOLMM-NoSPA and FastPOLMM-NoSPA, both of which use normal distribution approximation to test all variants. We also evaluated some alternative methods, including SAIGE, fastGWA, and BOLT-LMM. For SAIGE, we dichotomized the categorical phenotypes (Figure S4). For fastGWA and BOLT-LMM, we treated the categorical phenotype as a quantitative trait from 1 to J, where J is the number of category levels.

To compare DensePOLMM and FastPOLMM, we added one scenario to simulate random effect vector b. First, we randomly selected 50,000 variants (i.e., 50%) from the 100,000 variants that were used to estimate the GRM. Then, for subject i, random effect $b_i = \sqrt{\tau} \cdot \sum_{h=1}^{m_2} G_{ih} \cdot \gamma_h$, where $m_2 = 50,000$, G_{ih} was the genotype of the h-th selected variant, and γ_h was simulated following a normal distribution with a mean of 0 and a standard deviation of 0.085 so that the empirical variance of the random effects is close to τ. In this scenario, the random effects were strongly related to the estimated GRM used in the null mixed models fitting. We set variance components $\tau = 1$ and 10 to simulate moderate and high heritability, respectively. Besides the bell-shaped phenotypic distribution, we also simulated phenotypes with five and ten evenly distributed categories.

We also simulated phenotypes by using real genotype data from white British participants in UK Biobank. We selected 152,951 subjects who participated the questionnaire of food (and other) preferences. Instead of simulating random effect b by using a given family structure, we randomly selected $m_2 = 50,000$ common variants with MAFs > 0.05 in chromosomes 11–22 and then simulated random effect $b_i = \sqrt{\tau} \cdot \sum_{h=1}^{m_2} G_{ih} \cdot \gamma_h$. We simulated γ_h following two distributions: (1) a normal distribution with a mean of 0 and a standard deviation of 0.085 and (2) a gamma distribution with a shape parameter of 1 and a scale parameter of 0.05. We considered three $\tau = 0.5,\ 1,\ 2$ and simulated ordinal categorical phenotypes of four L-shaped distributions by using linear predicator $\eta_i = \ 0.5 \cdot (birth\ year) + 0.5 \cdot (sex) + 0.5 \cdot (PC1 + PC2 + PC3 + PC4) + b_i,\ i \leq n$.

In section C of the supplemental methods, we simulated ordinal categorical phenotypes following some alternative models, including adjacent category logistic model and stereotype model. The simulation results showed that POLMM approaches can still control type I error rates at a stringent significance level of 5×10^{-8} even if the proportional odds ratio assumption is violated (Figure S5).

Application to UK Biobank data

We used FastPOLMM to conduct genome-wide analyses of 258 ordinal categorical phenotypes in the UK Biobank data of 408,961 white British participants. Most of the categorical phenotypes measured dietary, lifestyle and environment, and psychosocial factors (Table S2). We used 30 million Haplotype Reference Consortium[23] (HRC)-imputed variants with minor allele counts ≥ 20 and imputation R^2 greater than 0.3. More details on the quality control, genotyping, imputation, and principal components can be found elsewhere.[6] We incorporated birth year, sex (if applicable), and top four principal components as covariates and used 340,447 high-quality SNPs to calculate the sparse GRM in which all off-diagonal elements less than 0.05 were set to 0.[9,21]

For phenotypes of food (and other) preferences, the values of phenotypes were collected from 2019 to 2020; for most of the other phenotypes, we only analyzed the values on the initial assessment visit (from 2006 to 2010). In addition, some phenotypes (e.g., comparative height size at age 10) are not based on the age to answer the questions. Hence, instead of using the age to answer the questions, we incorporated birth years as covariates in all the analyses. The subjects who did not participated in the survey or without meaningful values (e.g., "do not know" or "prefer not to answer") were excluded from the analysis. For example, for the food (and other) preferences, which account for 150 of 258 phenotypes, 152,951 white British participants were analyzed. We have carefully examined the orders of different categories.

Results

Runtime and resource requirements

The computation time and memory usage of all five methods of fastGWA, BOLT-LMM, SAIGE, DensePOLMM, and FastPOLMM are presented in Figure S6 and Table S1. In step 1, to fit a null mixed model, fastGWA and FastPOLMM were much faster and required much less memory than the three methods using dense GRMs. BOLT-LMM, SAIGE, and DensePOLMM required comparable computation time and memory usage because they used the same optimized strategies to incorporate a dense GRM. SAIGE and DensePOLMM were slower than BOLT-LMM because both logistic and proportional odds models require more computation steps to adjust for covariates than linear models in step 1. DensePOLMM required more time than SAIGE when sample size was greater than 100,000. This is mainly because DensePOLMM used a block diagonal matrix as the preconditioner matrix for PCG, which took more iterations to converge than that in SAIGE given the same tolerance criterion. Interestingly, DensePOLMM was faster than SAIGE when the sample size was smaller than 40,000. This might be because we optimized C++ codes to read in genotypes for GRM construction. For POLMM methods, more computational time and slightly more memory usage were required when analyzing a phenotype with more category levels. For example, to fit a null mixed model with 397,798 subjects, if the number of levels is 3, DensePOLMM and FastPOLMM took 49.9 and 0.03 h, respectively; if the number of levels is 6, DensePOLMM and FastPOLMM took 64.2 and 0.09 h, respectively.

In step 2, we first recorded the computation time to analyze 340,447 markers and then projected them to a genome-wide analysis with 30 million markers. The genotype data were stored in BGEN format because UK Biobank uses it for the imputed data.[24] BOLT-LMM and fastGWA were faster than POLMM and SAIGE methods, which is expected because logistic regression is more complicated than linear regression. POLMM is slightly faster than SAIGE. As the number of levels increased from 3 to 6, the computation time of POLMM methods slightly increased. Suppose that we use 24 CPU cores for parallel computation: POLMM methods require around 14.2 h for a genome-wide analysis including around 30 million markers.

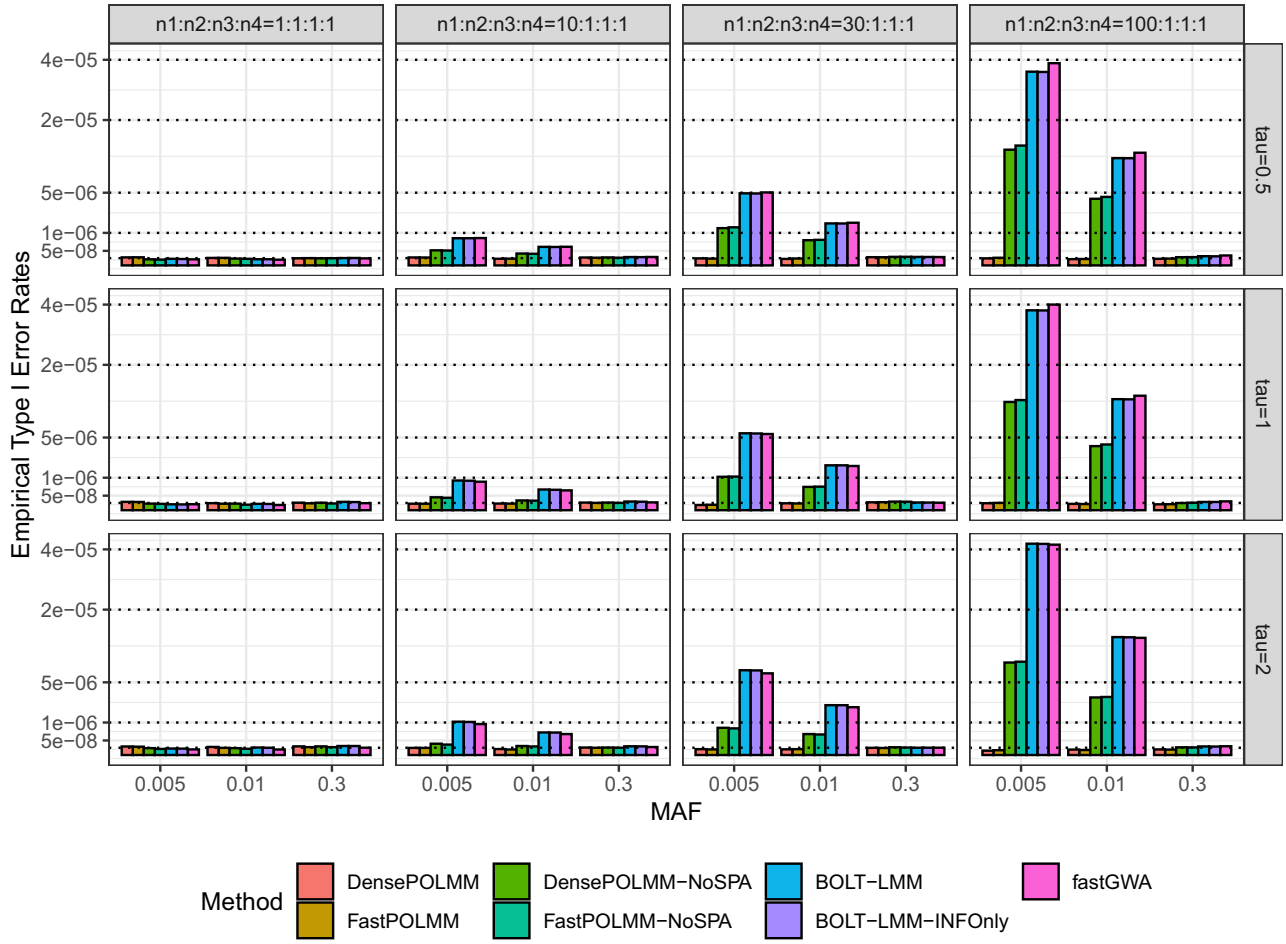

Figure 1. Empirical type I error rates of POLMM, BOLT-LMM, and fastGWA methods at a significance level 5×10^{-8}
We simulated 1,000 families with a total sample size $n = 10{,}000$ and an ordinal categorical phenotype including four levels with sample sizes n_1, n_2, n_3, and n_4. From left to right, the plots consider four scenarios: balanced ($n_1 : n_2 : n_3 : n_4 = 1 : 1 : 1 : 1$), moderately unbalanced ($n_1 : n_2 : n_3 : n_4 = 10 : 1 : 1 : 1$), unbalanced ($n_1 : n_2 : n_3 : n_4 = 30 : 1 : 1 : 1$), and extremely unbalanced ($n_1 : n_2 : n_3 : n_4 = 100 : 1 : 1 : 1$). From top to bottom, the plots consider three variance components, tau, $\tau = 0.5$, 1, and 2. We simulated common, low-frequency, and rare variants with MAFs of 0.3, 0.01, and 0.005, respectively. In total, 10^9 replications were conducted in each scenario.

False positive rate and statistical power

The simulation results showed that DensePOLMM and FastPOLMM methods can control type I error rates at a significance level of $5{\times}10^{-8}$ (Figures 1 and S7). Meanwhile, type I error rates of other methods were inflated when testing low-frequency and rare variants (MAF $\leq$ 0.01) and the phenotypic distribution was unbalanced. For example, when the variance component was $\tau = 1$ and the sample size proportion in 4 levels was 100:1:1:1, to test low-frequency variants with a MAF of 0.01, the type I error rates of POLMM methods and the other methods were less than $3.8{\times}10^{-8}$ and greater than $3.89{\times}10^{-6}$, respectively. Consistent for both bell-shaped and L-shaped phenotypic distributions, the results suggested that POLMM approaches can accurately account for ordinal categorical responses and using SPA is more accurate than using normal distribution. If we dichotomize the categorical phenotype, the POLMM is a logistic mixed model and it is expected that SAIGE can control type I error rates.[9] Hence, we did not evaluate the empirical type I error rates of SAIGE.

Next, we compared the empirical powers of POLMM methods, SAIGE, fastGWA, and BOLT-LMM at a significance level $\alpha = 5{\times}10^{-8}$ (Figures 2 and S8). Because fastGWA and BOLT-LMM cannot control type I error rates when the phenotypic distribution is unbalanced, we used empirical significance levels to evaluate powers. In all simulation scenarios, POLMM methods were the most powerful. When the phenotypic distribution is balanced, fastGWA and BOLT-LMM were similarly powerful as POLMM methods. However, when the phenotypic distribution is unbalanced, fastGWA and BOLT-LMM methods were less powerful than POLMM methods, especially when testing low-frequency variants with MAF = 0.01. Because the dichotomizing process would result in information loss, SAIGE was less powerful than POLMM methods. Figure S8 shows that different dichotomizing processes could result in significantly different powers for SAIGE.

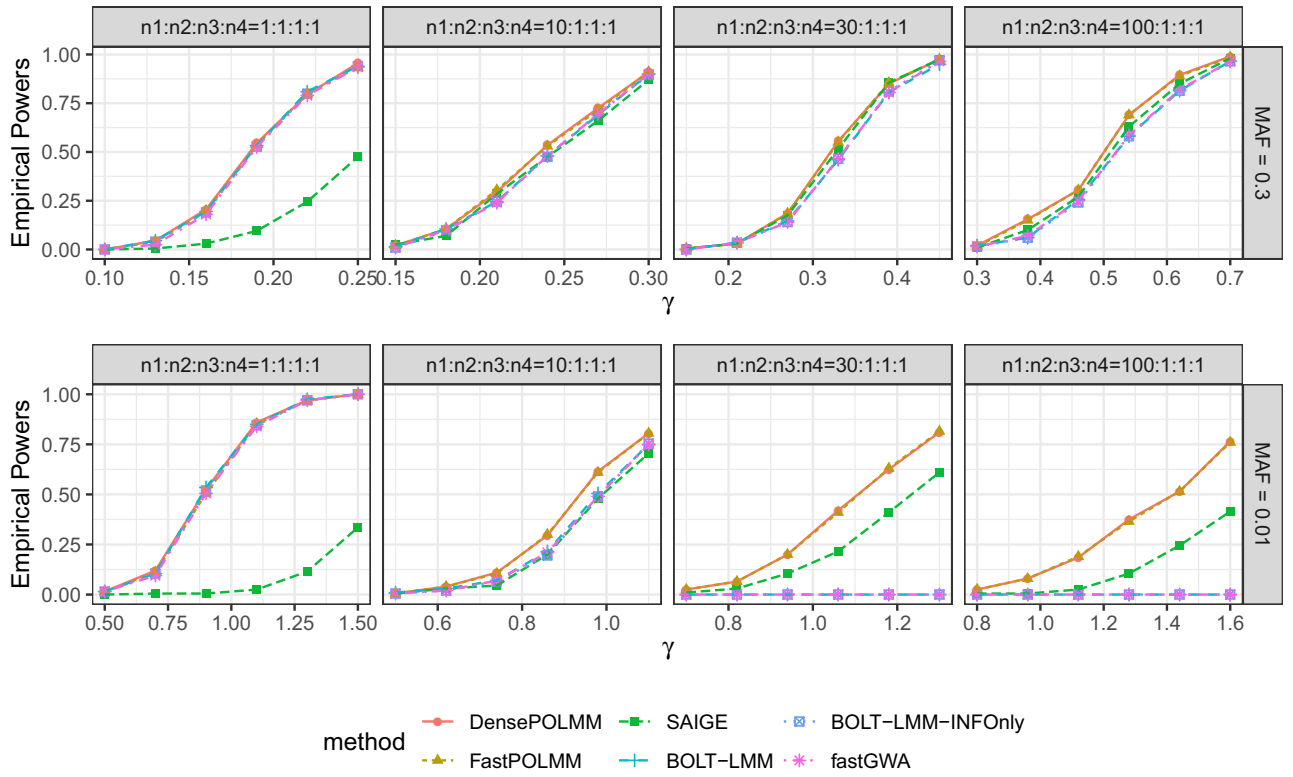

Figure 2. Empirical powers of POLMM, SAIGE, BOLT-LMM, and fastGWA methods at significance level 5×10^{-8}
We simulated 1,000 families with a total sample size $n = 10{,}000$ and an ordinal categorical phenotype including four levels with sample sizes n_1, n_2, n_3, and n_4. From left to right, the plots consider four scenarios: balanced ($n_1 : n_2 : n_3 : n_4 = 1 : 1 : 1 : 1$), moderately unbalanced ($n_1 : n_2 : n_3 : n_4 = 10 : 1 : 1 : 1$), unbalanced ($n_1 : n_2 : n_3 : n_4 = 30 : 1 : 1 : 1$), and extremely unbalanced ($n_1 : n_2 : n_3 : n_4 = 100 : 1 : 1 : 1$). From top to bottom, the plots consider two MAFs of 0.3 and 0.01 to simulate common and low-frequency variants. We let the variance component $\tau = 1$. For SAIGE, we dichotomize phenotype as 0 or 1 depending on whether the subject is in level 1 or not. For BOLT-LMM, the empirical powers were calculated on the basis of the empirical significance levels because it cannot control type I error rates for low-frequency variants.

Figures S9–S12 show the results of FastPOLMM when phenotypes were simulated with real genotypes. Because parts of genetic variants in chromosomes 11–22 are causal variants, we separately demonstrated the p value results of genetic variants in chromosomes 1–10 and chromosomes 11–22. From Figures S9 and S11, we can see POLMM methods can control type I error rates for various phenotypic distributions. On the other hand, from Figures S10 and S12, a large number of genetic variants in chromosomes 11–22 were identified. This is expected because we simulated the ordinal categorical phenotypes by using real data of variants in these chromosomes.

Comparison between DensePOLMM and FastPOLMM methods

Figures S13–S16 present the variance component estimation $\hat{\tau}$ and the empirical powers of POLMM methods. The estimation $\hat{\tau}$ of DensePOLMM and FastPOLMM, both of which deviated from true τ, were slightly different, especially when the true τ was large. The biased estimation has been widely discussed in other studies using penalized quasi-likelihood (PQL).[9] Interestingly, the estimation $\hat{\tau}$ increased and tended to the true τ as the number of levels increased from 3 to 10. This might be because more levels give more information, which results in a more accurate estimation of the variance component τ. In most scenarios, the empirical powers of DensePOLMM and FastPOLMM were similar, and the largest difference was less than 2.5%. Only when SNPs used to construct the GRM were significantly associated with the phenotype (e.g., liability heritability = 75.24%) and the number of levels is large (e.g., 10), DensePOLMM is more powerful than FastPOLMM by no more than 4.67% and 7.51% when testing SNPs with MAF = 0.3 and 0.01, respectively. This may be because only when the number of levels is large, accounting for the polygenic effects through a dense GRM can substantially improve the power. Note that in this simulation, we simulated SNPs for the dense GRM independently from the SNPs to test to prevent proximal contamination.

Compared to DensePOLMM, FastPOLMM can give a substantial improvement in terms of computation time and memory usage while only suffering a limited loss of power in restricted simulation scenarios. Hence, we recommend using FastPOLMM, especially when analyzing a large-scale dataset with sample size greater than 200,000.

Application to UK Biobank data

We used FastPOLMM to conduct genome-wide analyses of 30 million SNPs in the UK Biobank data of 408,961 samples from white British participants. We analyzed 258 ordinal categorical phenotypes, most of which measured dietary, lifestyle and environment, and psychosocial factors (Table S2). All analysis results are publicly available through a visual server. The web interface provides intuitive visualizations at three levels of granularity: genome-wide summaries at the trait level and regional (LocusZoom)[16] and phenome-wide summaries at the variant level.[2]

We used PLINK[25] to conduct clumping analysis for the variants with a p value less than 5×10^{-8} (window size of 5 Mb and linkage disequilibrium threshold r^2 of 0.1). For these 258 phenotypes, we identified 5,885 clumped distinct genome-wide significant variants, of which, 424 variants (7.2%) are low-frequency variants with MAF < 0.01. We used ANNOVAR[26] to functionally annotate these genome-wide significant variants. In total, 275 clumped variants are in exon region, of which, 207 (75.3%, binomial test p value: 1.04×10^{-12}) variants are nonsynonymous variants. On the basis of the PolyPhen2 HDIV score, a score to predict functional effect via HumDiv training set,[27] 63 nonsynonymous variants (30.4%, binomial test p value: 0.506) are probably damaging (score ≥ 0.957) and 33 nonsynonymous variants (15.9%, binomial test p value: 1) are possibly damaging (score ≥ 0.453). Table S3 summarizes the functional annotation of more than 24 million SNPs in which the proportion of nonsynonymous variants, probably damaging variants, and possibly damaging variants was calculated.

We highlighted some nonsynonymous significant low-frequency variants with MAF < 0.01. For the phenotype of "morning/evening person" (UK Biobank field ID: 1180), we identified an association of a nonsynonymous SNP rs139315125 (MAF: 0.47%, p value: 5.3×10^{-21}, gene: *PER3* [MIM: 603427], PolyPhen2 HDIV score: 0.998, see Figure S17 for more details). Subjects who tend to sleep and wake up early have a higher frequency of minor allele G. *PER3* is a core component of the circadian clock and the association between this SNP and sleep-wake patterns has been reported in previous studies.[28] For the phenotype of "use of sun/UV protection" (UK Biobank field ID: 2267), we identified a nonsynonymous SNP rs121918166 (MAF: 0.9%, p value: 5.2×10^{-31}, gene: *OCA2* [MIM: 611409], PolyPhen2 HDIV score: 1, see Figure S18 for more details). Subjects who use sun/UV protection more frequently have a higher frequency of minor allele T. *OCA2* is involved in mammalian pigmentation and this SNP has been previously associated with human eye color and melanoma.[29–31] Other interesting associations include the phenotype of "comparative height size at age 10" (UK Biobank field ID: 1697) and rs78727187 (MAF: 0.6%, p value: 5.1×10^{-19}, gene: *FBN2* [MIM: 612570], PolyPhen2 HDIV score: 0.818), rs117116488 (MAF: 0.99%, p value: 1.4×10^{-18}, gene: *ACAN* [MIM: 155760], PolyPhen2 HDIV score: 0.993), and rs112892337 (MAF: 0.4%, p value: 3.0×10^{-15}, gene: *ZFAT* [MIM: 610931], PolyPhen2 HDIV score: 1) and the phenotype of "relative age of first facial hair" (UK Biobank field ID: 2375) and rs138800983 (MAF: 0.3%, p value: 8.4×10^{-10}, gene: *KRT75* [MIM: 609025], PolyPhen2 HDIV score: 0.969).

In addition, we selected four food preferences with different sample size distributions as phenotypes to compare BOLT-LMM and FastPOLMM in UK Biobank data analysis (Figure S19). The preferences were encoded from 1 (extremely dislike) to 9 (extremely like). For BOLT-LMM, we treated the phenotypes as quantitative traits and incorporated the same set of covariates and GRM as in FastPOLMM. Figures 3 and S20 present the Manhattan and QQ plots of the analysis results. When the phenotypic distribution is balanced, BOLT-LMM performed similarly to FastPOLMM. However, in other cases, BOLT-LMM could inflate type I error rates, especially when testing low-frequency and rare variants with MAF < 0.01. FastPOLMM-NoSPA was better than BOLT-LMM but still cannot control type I error rates at a genome-wide significance level, which suggests that the proportional odds logistic model and SPA both contribute to more accurate association tests. All the real data analysis results were consistent with the simulation results, which indicate that using linear models is not an ideal solution in ordinal categorical data analysis, especially when testing low-frequency variants.

Discussion

In this study, we developed a scalable and accurate genetic association analysis tool, POLMM, for ordinal categorical data analysis in a large-scale dataset with hundreds of thousands of samples. The tool can accurately account for the dependence of an ordinal categorical phenotype on covariates. Two closely related methods, DensePOLMM and FastPOLMM, were proposed to use dense and sparse GRMs to adjust for the sample relatedness, respectively. DensePOLMM uses similar optimized strategies as in SAIGE and BOLT-LMM, which makes it scalable to incorporate a dense GRM into the mixed model. However, as the sample size increases, DensePOLMM is still computationally expensive. On the other hand, FastPOLMM is more computationally efficient. Extensive simulations demonstrate that FastPOLMM is as reliable as DensePOLMM and only suffers a small amount of power loss in limited simulation scenarios. Hence, if the sample size is greater than 500,000 and hundreds of GWASs are required for a phenome-wide analysis, we recommend using FastPOLMM.

We compared our method POLMM with two commonly used strategies: (1) dichotomizing the categorical phenotype and then using SAIGE[9] and (2) treating the categorical phenotype as a quantitative trait and then using BOLT-LMM[11] and fastGWA.[12] The dichotomizing process combined multiple levels into one group, which could lose useful phenotypic information and statistical power. On the other hand, treating the categorical phenotypes as a quantitative trait violates the nature of the ordinal categorical phenotype, which could result in inflated type I error rates and power

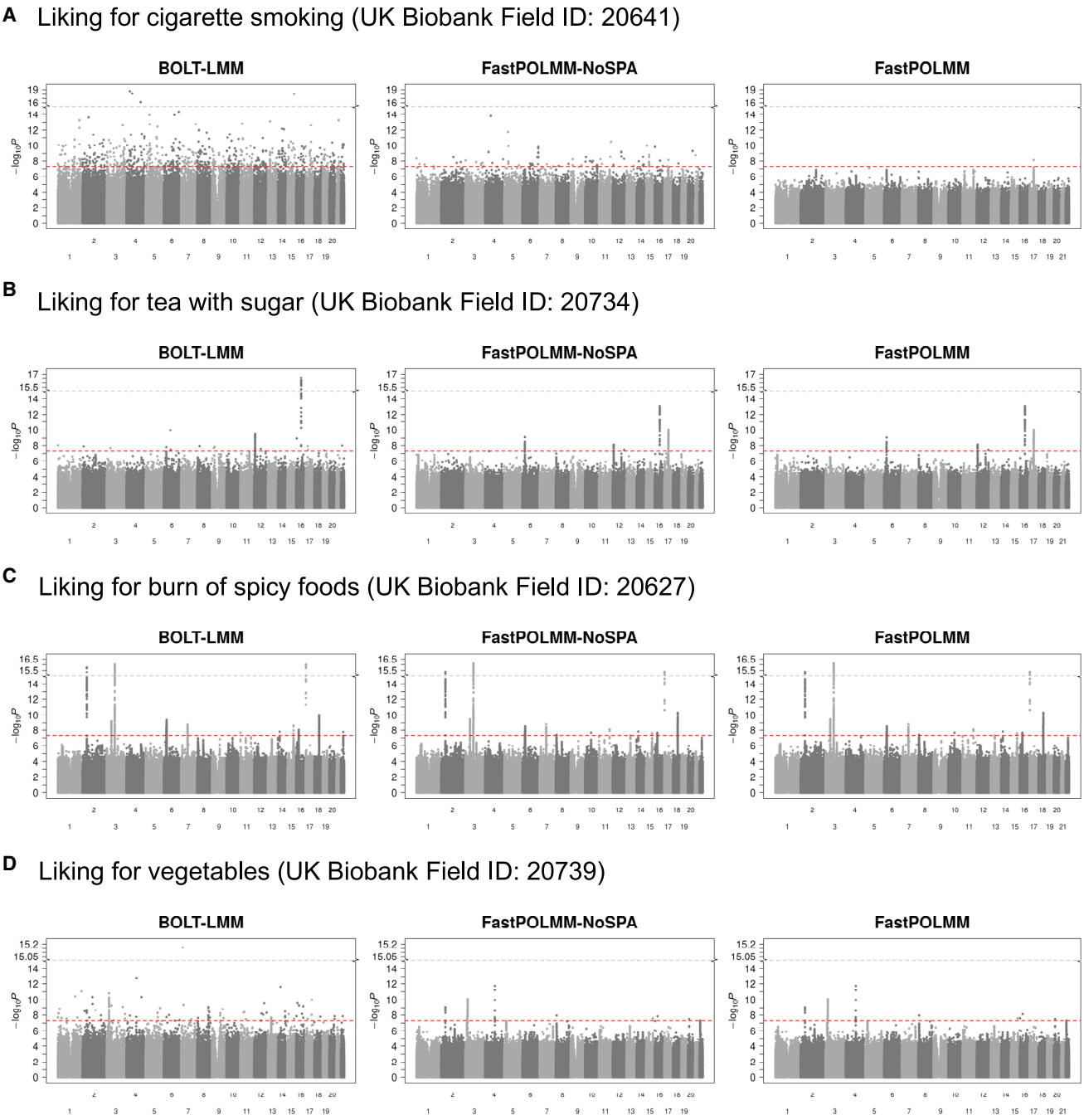

Figure 3. Manhattan plots for UK Biobank data analysis
The left panels show Manhattan plots based on BOLT-LMM, the middle panels show Manhattan plots based on FastPOLMM-NoSPA, and the right panels show Manhattan plots based on FastPOLMM. The redline represents the genome-wide significance level 5×10^{-8}.

loss. Through simulation studies and real data analysis, unless the phenotypic distribution is unbalanced, the linear mixed model approaches are still reliable when testing common variants, which suggests that fastGWA analyses limited to SNPs with MAF > 0.01 should still be valid for many of the phenotypes, whereas for low-frequency or rare variants, the linear mixed model approaches might be not valid anymore. The reliability of the linear mixed model approaches on categorical phenotypes greatly depends on the minor allele counts in the less common categories, which is relevant to both phenotypic distribution and the MAF of the marker. Considering the diversity of the phenotypic distribution, the arbitrary MAF cutoff of 0.01 still cannot ensure the results are well calibrated. In addition, we identified many phenotypes associated variants with MAF < 0.01 in the UK Biobank data analysis that were missed in the fastGWA analyses.

We applied the FastPOLMM to analyze 258 ordinal categorical phenotypes on UK Biobank, of which, 150 phenotypes are food and other preferences (UK Biobank category 1039). The preference data (v.1.1) were released in January 2020.

To the best of our knowledge, this is the first time that GWASs were applied to analyze the preference data. All analyses results have been made publicly available through a visual server. The web interface provides intuitive visualizations and is a useful resource for post-GWAS analyses. In this paper, we focus more on the development and the evaluation of the new POLMM methods. The UK Biobank data analysis has demonstrated the validity and reliability of the new methods on large-scale biobank categorical data analysis. More detailed explorations about the data analysis results are left to researchers with expertise in psychology, dietetics, etc.

There are several limitations in POLMM, most of which are similar to those in SAIGE and other mixed model approaches. First, DensePOLMM is still computationally expensive when fitting a null mixed model with sample size greater than 500,000. Second, POLMM assumes an infinitesimal architecture; that is, the effect sizes of genetic markers are normally distributed. If the genetic architecture is non-infinitesimal, POLMM methods may sacrifice power. Third, the variance component estimate $\hat{\tau}$ is biased and should not be used to estimate heritability. Interestingly, we observe a more accurate estimate $\hat{\tau}$ as the number of categories increases. Fourth, POLMM is based on a proportional odds model, which is not applicable to analyze unordered categorical response variables.

In the future, we plan to extend the current single-variant test to gene- or region-based multiple variants tests to better identify the rare variants. Recently, a machine learning method called REGENIE was proposed for quantitative and binary traits analysis. Instead of using a mixed effect model, REGENIE[32] uses a ridge regression model to account for polygenic effects. We plan to evaluate the strategies in REGENIE in ordinal categorical data analysis to extend POLMM. POLMM approaches are motivated to analyze large-scale biobank data collected following a cohort study design. Suppose that data are collected from a matched case-control study design, the stratified sampling for different levels could inflate the parameter estimation and genetic association testing.[33] We plan to extend the POLMM approaches to deal with the effect of the sampling. Similar to SAIGE, POLMM methods estimate odds ratios for genetic markers (supplemental methods, section A) by using the parameter estimates from the null model and might not be accurate. We plan to propose more accurate estimation by using Firth's correction on categorical data analysis.

Ordinal categorical phenotypes are widely observed in surveys, questionnaires, and testing to measure human behaviors, satisfaction, and preferences. However, because of the lack of analysis tools, methods designed for binary and quantitative traits have been used to analyze the categorical data, which is inappropriate and can result in suspicious results. Our method, POLMM, provides an accurate and scalable solution with the following features: can accurately model the ordinal categorical data by using a proportional odds logistic model, can adjust for sample relatedness by incorporating random effects, can be scalable to analyze a large-scale dataset with hundreds of thousands of subjects, and can test low-frequency variants under unbalanced phenotypic distribution by using SPA to approximate the null distribution of the test statistics. Because of all these features, POLMM is the only available unified approach for ordinal categorical data analysis in biobanks and large cohort studies.

Appendix A: Maximum likelihood estimation of POLMM

The maximum likelihood function and its derivatives

The first partial derivative of $l_i(\beta,\gamma;b,\varepsilon)$ with respect to the linear predicator η_i is

$$\begin{aligned}\frac{\partial l_i(\beta,\gamma;b,\varepsilon)}{\partial \eta_i} &= \sum_{j=1}^{J}\frac{y_{ij}}{\mu_{ij}}\cdot\frac{\partial \mu_{ij}}{\partial \eta_i} = \sum_{j=1}^{J}\frac{\left(y_{ij}-\mu_{ij}\right)}{\mu_{ij}}\cdot\frac{\partial \mu_{ij}}{\partial \eta_i}\\ &= \sum_{j=1}^{J-1}\frac{\left(y_{ij}-\mu_{ij}\right)}{\mu_{ij}}\cdot\frac{\partial \mu_{ij}}{\partial \eta_i}+\frac{\left(y_{iJ}-\mu_{iJ}\right)}{\mu_{iJ}}\cdot\frac{\partial \mu_{iJ}}{\partial \eta_i}\\ &= \sum_{j=1}^{J-1}\frac{\left(y_{ij}-\mu_{ij}\right)}{\mu_{ij}}\cdot\frac{\partial \mu_{ij}}{\partial \eta_i}-\sum_{j=1}^{J-1}\frac{\left(y_{ij}-\mu_{ij}\right)}{\mu_{iJ}}\cdot\frac{\partial \mu_{iJ}}{\partial \eta_i}\\ &= \sum_{j=1}^{J-1}\left(y_{ij}-\mu_{ij}\right)\cdot\left[\frac{1}{\mu_{ij}}\cdot\frac{\partial \mu_{ij}}{\partial \eta_i}-\frac{1}{\mu_{iJ}}\cdot\frac{\partial \mu_{iJ}}{\partial \eta_i}\right].\end{aligned}$$

The second and fourth equations hold since $\sum_{j=1}^{J} y_{ij} = \sum_{j=1}^{J}\mu_{ij} = 1$ and $\sum_{j=1}^{J}\partial\mu_{ij}/\partial\eta_i = 0$. The first derivative of log-likelihood function $l(\beta,\gamma;b,\varepsilon) = \sum_{i\le n} l_i(\beta,\gamma;b,\varepsilon)$ with respect to $\eta = (\eta_1,\cdots,\eta_n)^T$ is

$$\frac{\partial l(\beta,\gamma;b,\varepsilon)}{\partial \eta} = \tilde{Z}^T R(\tilde{y}-\tilde{\mu}),$$

and the first derivatives of $l(\beta,\gamma;b,\varepsilon)$ with respect to (β,γ,b) are

$$\begin{aligned}\frac{\partial l(\beta,\gamma;b,\varepsilon)}{\partial \beta} &= X^T\tilde{Z}^T R(\tilde{y}-\tilde{\mu}), \frac{\partial l(\beta,\gamma;b,\varepsilon)}{\partial \gamma}\\ &= G^T\tilde{Z}^T R(\tilde{y}-\tilde{\mu}), \frac{\partial l(\beta,\gamma;b,\varepsilon)}{\partial b} = \tilde{Z}^T R(\tilde{y}-\tilde{\mu}),\end{aligned}$$

where the definitions of matrix X, G, R, $\tilde{Z}$, and vectors of $\tilde{y}$ and $\tilde{\mu}$ have been given in the main text. Under certain regularity conditions,[34] the second derivative of $l(\beta,\gamma;b,\varepsilon)$ with respect to b can be approximated by

$$\begin{aligned}\frac{\partial^2 l(\beta,\gamma;b,\varepsilon)}{\partial b\partial b^T} &\approx E\left(\frac{\partial^2 l(\beta,\gamma;b,\varepsilon)}{\partial b\partial b^T}\right)\\ &\approx E\left(-\frac{\partial l(\beta,\gamma;b,\varepsilon)}{\partial b}\frac{\partial l(\beta,\gamma;b,\varepsilon)}{\partial b^T}\right)\\ &= -\tilde{Z}^T R\cdot E\left((\tilde{y}-\tilde{\mu})\cdot(\tilde{y}-\tilde{\mu})^T\right)\cdot R\tilde{Z}\\ &= -\tilde{Z}^T R\Psi R\tilde{Z}.\end{aligned}$$

Estimation of fixed covariates effects and random effects

Similar to GMMAT,[7] we use Laplace's method to approximate the n-dimensional integral, and the marginal log-

likelihood function becomes the following penalized quasi-likelihood (PQL)[35]

$$l(\beta,\gamma,\tau;\varepsilon) \approx -\frac{1}{2}\log|\tau V| - \frac{1}{2}\log\left|-f''\left(\tilde{b}\right)\right| + f\left(\tilde{b}\right), \quad \text{(Equation A1)}$$

where

$$f(b) = l(\beta,\gamma;b,\varepsilon) - \frac{1}{2}b^T(\tau V)^{-1}b,\ \tilde{b} = \text{argmax}f(b)$$

and the second derivative

$$f''(b) = \frac{\partial^2 l(\beta,\gamma;b,\varepsilon)}{\partial b\partial b^T} - (\tau V)^{-1} \approx -\tilde{Z}^T R\Psi R\tilde{Z} - (\tau V)^{-1}.$$

Following GMMAT[7] and SAIGE,[9] we assume that matrix R and Ψ change slowly with respect to η. The derivatives of Equation A1 with respect to (β,γ,b) are

$$\frac{\partial l(\beta,\gamma,\tau;\varepsilon)}{\partial\beta} = \frac{\partial l(\beta,\gamma;b,\varepsilon)}{\partial\beta} = X^T\tilde{Z}^T R(\tilde{y}-\tilde{\mu}),$$

$$\frac{\partial l(\beta,\gamma,\tau;\varepsilon)}{\partial\gamma} = \frac{\partial l(\beta,\gamma;b,\varepsilon)}{\partial\gamma} = G^T\tilde{Z}^T R(\tilde{y}-\tilde{\mu}),$$

$$\frac{\partial l(\beta,\gamma,\tau;\varepsilon)}{\partial b} = \frac{\partial l(\beta,\gamma;b,\varepsilon)}{\partial b} - (\tau V)^{-1}b = \tilde{Z}^T R(\tilde{y}-\tilde{\mu}) - (\tau V)^{-1}b.$$

Under the null hypothesis $\gamma = 0$, if ε and τ are known, we jointly choose $\widehat{\beta}(\varepsilon,\tau)$ and $\widehat{b}(\varepsilon,\tau)$ to maximize $l(\beta,\ \gamma,\ \tau;\varepsilon)$, then $\widehat{b}(\varepsilon,\tau) = \tilde{b}(\widehat{\beta}(\varepsilon,\tau),\gamma=0)$ because $\tilde{b}$ maximizes $f(b)$ for given $(\beta,\ \gamma)$.[7] Defining a working vector $\tilde{Y} = \tilde{Z}\eta + R^{-1}\Psi^{-1}(\tilde{y} - \tilde{\mu})$, the solution of

$$X^T\tilde{Z}^T R(\tilde{y}-\tilde{\mu}) = 0, \tilde{Z}^T R(\tilde{y}-\tilde{\mu}) - (\tau V)^{-1}b = 0$$

can be written as the solution to the system

$$\begin{bmatrix} X^T\tilde{Z}^T R\Psi R\tilde{Z}X & X^T\tilde{Z}^T R\Psi R\tilde{Z} \\ \tilde{Z}^T R\Psi R\tilde{Z}X & \tilde{Z}^T R\Psi R\tilde{Z} + (\tau V)^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ b \end{bmatrix} = \begin{bmatrix} X^T\tilde{Z}^T R\Psi R\tilde{Y} \\ \tilde{Z}^T R\Psi R\tilde{Y} \end{bmatrix}.$$

Let $\tilde{V} = \tilde{Z}V\tilde{Z}^T$, $\Sigma = R^{-1}\Psi^{-1}R^{-1} + \tau\tilde{V}$, and $P = \Sigma^{-1} - \Sigma^{-1}\tilde{Z}X\left(X^T\tilde{Z}^T\Sigma^{-1}\tilde{Z}X\right)^{-1}X^T\tilde{Z}^T\Sigma^{-1}$, then

$$\widehat{\beta} = \left(X^T\tilde{Z}^T\Sigma^{-1}\tilde{Z}X\right)^{-1}X^T\tilde{Z}^T\Sigma^{-1}\tilde{Y}, \widehat{b} = \tau V\cdot\tilde{Z}^T\Sigma^{-1}\left(\tilde{Y} - \tilde{Z}X\widehat{\beta}\right) \quad \text{(Equation A2)}$$

is the solution. We note that

$$\begin{aligned}\tilde{Y} - \tilde{Z}\eta &= \tilde{Y} - \tilde{Z}X\widehat{\beta} - \tilde{Z}\widehat{b} = \left\{I - \tau\tilde{V}\cdot\Sigma^{-1}\right\}\left(\tilde{Y} - \tilde{Z}X\widehat{\beta}\right) \\ &= R^{-1}\Psi^{-1}R^{-1}\Sigma^{-1}\cdot\left(\tilde{Y} - \tilde{Z}X\widehat{\beta}\right) = R^{-1}\Psi^{-1}R^{-1}P\tilde{Y}.\end{aligned}$$

Estimation of variance component parameters

Given random effect $\widehat{b}$, vector $\tilde{y}$ has a mean of $\tilde{\mu}$ and a covariance matrix of Ψ. Using quasi-likelihood and Pearson chi-square statistics,[35] we approximate the log-likelihood

$$\begin{aligned}l\left(\beta,\gamma;\widehat{b},\varepsilon\right) &\approx C_1 - \frac{1}{2}\cdot(\tilde{y}-\tilde{\mu})^T\Psi^{-1}(\tilde{y}-\tilde{\mu}) \\ &= C_1 - \frac{1}{2}\cdot\left(\tilde{Y} - \tilde{Z}\eta\right)^T R\Psi R\left(\tilde{Y} - \tilde{Z}\eta\right),\end{aligned}$$

where C_1 is independent from random vector $\tilde{y}$. Then, the log-likelihood function

$$l(\beta,\gamma,\tau;\varepsilon) \approx -\frac{1}{2}\log|\tau V| - \frac{1}{2}\log\left|-f''\left(\tilde{b}\right)\right| + f\left(\tilde{b}\right) \approx -\frac{1}{2}\log|\tau V| - \frac{1}{2}\log\left|\tilde{Z}^T R\Psi R\tilde{Z} + (\tau V)^{-1}\right| + l\left(\beta,\gamma;\widehat{b},\varepsilon\right)$$

$$-\frac{1}{2}\widehat{b}^T(\tau V)^{-1}\widehat{b} \approx -\frac{1}{2}\log\left|I_n + \tau V\cdot\tilde{Z}^T R\Psi R\tilde{Z}\right| + C_1 - \frac{1}{2}\cdot\left(\tilde{Y} - \tilde{Z}\eta\right)^T R\Psi R\left(\tilde{Y} - \tilde{Z}\eta\right)$$

$$-\frac{1}{2}\left(\tilde{Y} - \tilde{X}\widehat{\beta}\right)^T\Sigma^{-1}\tilde{Z}\cdot(\tau V)\cdot\tilde{Z}^T\Sigma^{-1}\left(\tilde{Y} - \tilde{X}\widehat{\beta}\right) = -\frac{1}{2}\log\left|I_n + \tau\tilde{V}\cdot R\Psi R\right| + C_1 - \frac{1}{2}\tilde{Y}^T PR^{-1}\Psi^{-1}R^{-1}P\tilde{Y}$$

$$-\frac{1}{2}\tilde{Y}^T P\cdot(\tau\tilde{V})\cdot P\tilde{Y} = -\frac{1}{2}\log\left|(R^{-1}\Psi^{-1}R^{-1} + \tau\tilde{V})\cdot R\Psi R\right| + C_1 - \frac{1}{2}\tilde{Y}^T P(R^{-1}\Psi^{-1}R^{-1} + \tau\tilde{V})P\tilde{Y} =$$

$$-\frac{1}{2}\log|\Sigma\cdot R\Psi R| + C_1 - \frac{1}{2}\tilde{Y}^T P\Sigma P\tilde{Y} = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\log|R\Psi R| + C_1 - \frac{1}{2}\tilde{Y}^T P\tilde{Y} = C - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\tilde{Y}^T P\tilde{Y}.$$

The restricted maximum likelihood (REML) version[7] is

$$l_R(\beta,\gamma,\tau;\varepsilon) \approx C_R - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\log\left|\tilde{X}^T\Sigma^{-1}\tilde{X}\right| - \frac{1}{2}\tilde{Y}^T P\tilde{Y}.$$

Because $\partial P/\partial\tau = -P\tilde{V}P$, the derivative

$$\frac{\partial l_R(\beta,\gamma,\tau;\varepsilon)}{\partial\tau} = \frac{1}{2}\tilde{Y}^T P\tilde{V}P\tilde{Y} - \frac{1}{2}\text{tr}\left[P\tilde{V}\right]$$

and the average information matrix, AI, is as below:

$$AI = \frac{1}{2}\cdot \tilde{Y}^T P\tilde{V}P\tilde{V}P\tilde{Y}.$$

Using AI-REML algorithm, we avoid the evaluation of the traces of large matrices that appear in both the expected and observed (REML) information matrices.[36]

Workflow of the model fitting algorithm

We add one intercept term with all elements of 1 to the covariate matrix and fix the first cutpoint $\varepsilon_1 = 0$. Then, after updating $\widehat{\beta}$ and $\widehat{b}$, we use the Newton-Raphson method to iteratively estimate cutpoints $\varepsilon_2, \cdots, \varepsilon_{J-1}$ until convergence.

We use the following workflow to fit the null POLMM:

(1) fit a proportional odds logistic model with $\tau = 0$ and $\gamma = 0$ to estimate $\widehat{\beta}^{(0)}, \widehat{\varepsilon}^{(0)}$, and then calculate $\tilde{Y}^{(0)}$; set initial value $\widehat{\tau}^{(0)} = 0.2$;
(2) update $\widehat{\beta}^{(1)}, \widehat{b}^{(1)}$ and $\widehat{\varepsilon}^{(1)}$ by using $\widehat{\tau}^{(0)}$ and $\tilde{Y}^{(0)}$;
(2.1) update $\widehat{\beta}, \widehat{b}$ following Equation A2;
(2.2) use the Newton-Raphson algorithm to update $\widehat{\varepsilon}$ until converges;
(2.3) repeat steps 2.1 and 2.2 until $\widehat{\beta}$ converges;
(3) update $\tilde{Y}^{(1)}$ and $\widehat{\tau}^{(1)} = \widehat{\tau}^{(0)} + \{AI^{(1)}\}^{-1}(\partial l_R(\widehat{\tau}^{(0)})/\partial\tau)$ by using $\widehat{\beta}^{(1)}, \widehat{b}^{(1)}$ and $\widehat{\varepsilon}^{(1)}$;
(4) repeat steps 2–3 until $\widehat{\tau}$ converges.

Appendix B: Score test and saddlepoint approximation

Under the null hypothesis, the score statistic

$$T = \frac{\partial l(\beta,\tau;\varepsilon)}{\partial\gamma} = G^T\tilde{Z}^T R(\tilde{y}-\tilde{\mu}) = \overline{G}^T\tilde{Z}^T R(\tilde{y}-\tilde{\mu})$$
$$= \overline{G}^T\tilde{Z}^T R\Psi R\left(\tilde{Y}-\tilde{\eta}\right) = \overline{G}^T\tilde{Z}^T P\tilde{Y}.$$

Because $\tilde{Y} = \tilde{Z}\eta + R^{-1}\Psi^{-1}(\tilde{y}-\tilde{\mu})$, its estimated variance is

$$\widehat{Var}(T) = E\left(\overline{G}^T\tilde{Z}^T P\tilde{Y}\tilde{Y}^T P\tilde{Z}\overline{G}\right) = \overline{G}^T\tilde{Z}^T P\cdot[R^{-1}\Psi^{-1}$$
$$\cdot E\left((\tilde{y}-\tilde{\mu})\cdot(\tilde{y}-\tilde{\mu})^T\right)\cdot\Psi^{-1}R^{-1} + \tilde{Z}\cdot E\left(\eta\cdot\eta^T\right)\cdot\tilde{Z}^T]\cdot P\tilde{Z}\overline{G}$$
$$= \overline{G}^T\tilde{Z}^T P\cdot\left[R^{-1}\Psi^{-1}R^{-1} + \tilde{Z}\cdot(\tau V)\cdot\tilde{Z}^T\right]\cdot P\tilde{Z}\overline{G}$$
$$= \overline{G}^T\tilde{Z}^T P\Sigma P\tilde{Z}\overline{G} = \overline{G}^T\tilde{Z}^T P\tilde{Z}\overline{G}.$$

For each variant, the variance-adjusted test statistic is

$$T_{adj} = \frac{T}{\sqrt{\widehat{Var}(T)}} = \frac{\overline{G}^T\tilde{Z}^T R(\tilde{y}-\tilde{\mu})}{\sqrt{\overline{G}^T\tilde{Z}^T P\tilde{Z}\overline{G}}} = \frac{\overline{G}^T\tilde{Z}^T R(\tilde{y}-\tilde{\mu})}{\sqrt{\widehat{\tau}\overline{G}^T\tilde{Z}^T R\Psi R\tilde{Z}\overline{G}}},$$

which has mean zero and variance one under the null hypothesis. Because the statistic

$$T_{adj} = \frac{\overline{G}^T R(\tilde{y}-\tilde{\mu})}{\sqrt{\widehat{\tau}\overline{G}^T R\Psi R\overline{G}}} = \frac{1}{\sqrt{\widehat{\tau}}}\cdot\sum_{i=1}^{n} T_i$$

and y_{ij} follows a Berounlli (μ_{ij}) distribution, the CGF of T_i is

$$K_i(t) = \log\left[E\left(e^{tT_i}\right)\right] = \log\left(1 - \sum_{j=1}^{J-1}\mu_{ij} + \sum_{j=1}^{J-1}e^{c_{ij}t}\mu_{ij}\right) - \left(\sum_{j=1}^{J-1}c_{ij}\mu_{ij}\right)t$$

and its derivatives

$$K_i'(t) = \frac{\sum_{j=1}^{J-1}e^{c_{ij}t}\mu_{ij}c_{ij}}{1-\sum_{j=1}^{J-1}\mu_{ij}+\sum_{j=1}^{J-1}e^{c_{ij}t}\mu_{ij}} - \left(\sum_{j=1}^{J-1}c_{ij}\mu_{ij}\right),$$

$$K_i''(t) = \frac{\left[\sum_{j=1}^{J-1}e^{c_{ij}t}\mu_{ij}c_{ij}^2\right]\cdot\left[1-\sum_{j=1}^{J-1}\mu_{ij}+\sum_{j=1}^{J-1}e^{c_{ij}t}\mu_{ij}\right] - \left[\sum_{j=1}^{J-1}e^{c_{ij}t}\mu_{ij}c_{ij}\right]^2}{\left[1-\sum_{j=1}^{J-1}\mu_{ij}+\sum_{j=1}^{J-1}e^{c_{ij}t}\mu_{ij}\right]^2}.$$

We use $K(t) = \sum_{i=1}^{n}K_i(t)$ to approximate the CGF of T_{adj} such that the variance from CGF is 1; that is,

$$K''(0) = \sum_{i=1}^{n}K_i''(0) = \sum_{i=1}^{n}\left\{\sum_{j=1}^{J-1}\mu_{ij}c_{ij}^2 - \left[\sum_{j=1}^{J-1}\mu_{ij}c_{ij}\right]^2\right\} = \tilde{c}^T\Psi\tilde{c}$$
$$= \frac{\overline{G}^T R\Psi R\overline{G}}{\overline{G}^T R\Psi R\overline{G}} = 1,$$

where

$$\tilde{c} = (c_{11}, c_{12}, \cdots, c_{1(J-1)}, c_{21}, c_{22}, \cdots, c_{2(J-1)}, \cdots, c_{n1}, c_{n2}, \cdots, c_{n(J-1)})^T.$$

After fitting the null model, we calculate and store the following matrix:

$$A_1 = X\left(X^T\tilde{Z}^T R\Psi R\tilde{Z}X\right)^{-1}, A_2 = X^T\tilde{Z}^T R\Psi R\tilde{Z}, A_3$$
$$= \tilde{Z}^T R(\tilde{y}-\tilde{\mu}), A_4 = \tilde{Z}^T R\Psi R\tilde{Z}.$$

For each variant, it takes $O(np)$ computations to calculate vector $\overline{G} = G - A_1\cdot A_2\cdot G$. Because A_4 is a diagonal matrix, it takes $O(n)$ to calculate the score statistic $T = \overline{G}^T\cdot A_3$ and

the variance $\widehat{Var}^*(T) = \overline{G}^T A_4 \overline{G}$. Thus, for normal distribution approximation, the computational complexity is still $O(np)$ and does not increase as the number of category levels J increases. For SPA, we use a partially normal approximation method to speed up the computation.[8] Suppose that the first m subjects have at least one minor allele each and the rest have homozygous major genotypes. We can express

$$T_{adj} = \frac{1}{\sqrt{\hat{r}}} \cdot \sum_{i=1}^{n} T_i = \frac{1}{\sqrt{\hat{r}}} \cdot (T_{(1)} + T_{(2)}),$$

where $T_{(1)} = \sum_{i=1}^{m} T_i$ and $T_{(2)} = \sum_{i=m+1}^{n} T_i$. Let $W = (X^T \tilde{Z}^T R \Psi R \tilde{Z} X)^{-1} X^T \tilde{Z}^T R \Psi R \tilde{Z} G$, and let W_l be the l^{th} element of W. Then, we can further express $T_{(2)}$ as

$$\begin{aligned}
T_{(2)} &= \frac{1}{\sqrt{\widehat{Var}^*(T)}} \cdot \sum_{i=m+1}^{n} \overline{G}_i \left(\sum_{j=1}^{J-1} R_{ij} \left(y_{ij} - \mu_{ij} \right) \right) \\
&= \frac{1}{\sqrt{\widehat{Var}^*(T)}} \cdot \sum_{i=m+1}^{n} (0 - X_i W) \left(\sum_{j=1}^{J-1} R_{ij} \left(y_{ij} - \mu_{ij} \right) \right) \\
&= -\frac{1}{\sqrt{\widehat{Var}^*(T)}} \cdot \sum_{i=m+1}^{n} \sum_{l=1}^{p} X_{il} W_l \left(\sum_{j=1}^{J-1} R_{ij} \left(y_{ij} - \mu_{ij} \right) \right) \\
&= -\frac{1}{\sqrt{\widehat{Var}^*(T)}} \cdot \sum_{l=1}^{p} W_l \cdot \sum_{i=m+1}^{n} X_{il} \left(\sum_{j=1}^{J-1} R_{ij} \left(y_{ij} - \mu_{ij} \right) \right) \\
&= -\frac{1}{\sqrt{\widehat{Var}^*(T)}} \cdot \sum_{l=1}^{p} W_l \cdot T_{(2l)},
\end{aligned}$$

where

$$T_{(2l)} = \sum_{i=m+1}^{n} \sum_{j=1}^{J-1} \left\{ X_{il} R_{ij} \left(y_{ij} - \mu_{ij} \right) \right\}.$$

If we assume that the non-genetic covariates are relatively balanced in the sample, then the normal approximation should be a good approximation of the null distribution of each $T_{(2l)}$. Because $T_{(2)}$ is a weighted sum of the $T_{(2l)}$ variables, we can also approximate the null distribution of $T_{(2)}$ by using a normal distribution and the CGF of $T_{(2)}$ can be approximated by

$$K_{(2)}(t) = \frac{1}{2} t^2 \cdot V_{H_0}(T_{(2)}),$$

where

$$V_{H_0}(T_{(2)}) = \sum_{i=m+1}^{n} \frac{\overline{G}_i^2 \cdot R_i^T \Psi_i R_i}{\widehat{Var}^*(T)}$$

and $R_i = (R_{i1}, R_{i2}, \cdots, R_{i(J-1)})^T$. Hence, with the partially normal approximation, the CGF of T_{adj} is $K(t) = \sum_{i=1}^{m} K_i(t) + K_{(2)}(t)$, and the SPA takes $O(m(J-1))$ computations to calculate the CGF and its derivatives.

Data and code availability

The summary statistics and PheWeb with quantile-quantile plots, Manhattan plots, and regional association plots for 258 categorical phenotypes in the UK Biobank by POLMM are available for public download (see web resources). POLMM is implemented as an open-source R package (see web resources).

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.03.019.

Acknowledgments

This research was supported by NIH grant R01-HG008773 (W.B. and S.L.) and the Brain Pool Plus Program (BP+, Brain Pool+) through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2020H1D3A2A03100666, S.L.). UK Biobank data were accessed under the accession number UKB: 45227.

Declaration of interests

The authors declare no competing interests.

Received: January 7, 2021
Accepted: March 22, 2021
Published: April 8, 2021

Web resources

ANNOVAR (April 16, 2018), https://annovar.openbioinformatics.org/en/latest/
BOLT-LMM (v.2.3.4), https://alkesgroup.broadinstitute.org/BOLT-LMM
fastGWA (GCTA, v.1.93.1beta), https://cnsgenomics.com/software/gcta/#fastGWA
POLMM (v.0.2.2), https://github.com/WenjianBI/POLMM
RcppParallel, http://rcppcore.github.io/RcppParallel/
SAIGE (v.0.36.3), https://github.com/weizhouUMICH/SAIGE
UK Biobank PheWeb and analysis results, https://polmm.leelabsg.org/

References

1. Beesley, L.J., Salvatore, M., Fritsche, L.G., Pandit, A., Rao, A., Brummett, C., Willer, C.J., Lisabeth, L.D., and Mukherjee, B. (2019). The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. Stat. Med. *39*, 773–800.
2. Gagliano Taliun, S.A., VandeHaar, P., Boughton, A.P., Welch, R.P., Taliun, D., Schmidt, E.M., Zhou, W., Nielsen, J.B., Willer, C.J., Lee, S., et al. (2020). Exploring and visualizing large-scale genetic associations by using PheWeb. Nat. Genet. *52*, 550–552.

3. Lane, J.M., Jones, S.E., Dashti, H.S., Wood, A.R., Aragam, K.G., van Hees, V.T., Strand, L.B., Winsvold, B.S., Wang, H., Bowden, J., et al.; HUNT All In Sleep (2019). Biological and clinical insights from genetics of insomnia symptoms. Nat. Genet. *51*, 387–393.
4. Agresti, A. (2003). Categorical data analysis (John Wiley & Sons).
5. Verhulst, B., Maes, H.H., and Neale, M.C. (2017). GW-SEM: A Statistical Package to Conduct Genome-Wide Structural Equation Modeling. Behav. Genet. *47*, 345–359.
6. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
7. Chen, H., Wang, C., Conomos, M.P., Stilp, A.M., Li, Z., Sofer, T., Szpiro, A.A., Chen, W., Brehm, J.M., Celedón, J.C., et al. (2016). Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. Am. J. Hum. Genet. *98*, 653–666.
8. Dey, R., Schmidt, E.M., Abecasis, G.R., and Lee, S. (2017). A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. Am. J. Hum. Genet. *101*, 37–49.
9. Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. *50*, 1335–1341.
10. Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. *44*, 821–824.
11. Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. *47*, 284–290.
12. Jiang, L., Zheng, Z., Qi, T., Kemper, K.E., Wray, N.R., Visscher, P.M., and Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data (Nature Publishing Group).
13. Zhao, Z., Bi, W., Zhou, W., VandeHaar, P., Fritsche, L.G., and Lee, S. (2020). UK Biobank Whole-Exome Sequence Binary Phenome Analysis with Robust Region-Based Rare-Variant Test. Am. J. Hum. Genet. *106*, 3–12.
14. Zhou, W., Zhao, Z., Nielsen, J.B., Fritsche, L.G., LeFaive, J., Gagliano Taliun, S.A., Bi, W., Gabrielsen, M.E., Daly, M.J., Neale, B.M., et al. (2020). Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. Nat. Genet. *52*, 634–639.
15. Bi, W., Zhao, Z., Dey, R., Fritsche, L.G., Mukherjee, B., and Lee, S. (2019). A Fast and Accurate Method for Genome-wide Scale Phenome-wide G × E Analysis and Its Application to UK Biobank. Am. J. Hum. Genet. *105*, 1182–1192.
16. Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., and Willer, C.J. (2010). LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics *26*, 2336–2337.
17. Holtbrügge, W., and Schumacher, M. (1991). A comparison of regression models for the analysis of ordered categorical data. J. R. Stat. Soc. Ser. C Appl. Stat. *40*, 249–259.
18. Gilmour, A.R., Anderson, R.D., and Rae, A.L. (1985). The Analysis of Binomial Data by a Generalized Linear Mixed Model. Biometrika *72*, 593–599.
19. Lin, X., and Breslow, N.E. (1996). Bias Correction in Generalized Linear Mixed Models With Multiple Components of Dispersion. J. Am. Stat. Assoc. *91*, 1007–1016.
20. Hutchinson, M.F. (1990). A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. Commun. Stat. Simul. Comput. *19*, 433–450.
21. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
22. Abecasis, G.R., Cherny, S.S., Cookson, W.O., and Cardon, L.R. (2002). Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat. Genet. *30*, 97–101.
23. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al.; Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. *48*, 1279–1283.
24. Band, G., and Marchini, J. (2018). BGEN: a binary file format for imputed genotype and haplotype data. bioRxiv. https://doi.org/10.1101/308296.
25. Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.
26. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.
27. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. *Chapter 7*, 20.
28. Zhang, L., Hirano, A., Hsu, P.-K., Jones, C.R., Sakai, N., Okuro, M., McMahon, T., Yamazaki, M., Xu, Y., Saigoh, N., et al. (2016). A PERIOD3 variant causes a circadian phenotype and is associated with a seasonal mood trait. Proc. Natl. Acad. Sci. USA *113*, E1536–E1544.
29. Duffy, D.L., Box, N.F., Chen, W., Palmer, J.S., Montgomery, G.W., James, M.R., Hayward, N.K., Martin, N.G., and Sturm, R.A. (2004). Interactive effects of MC1R and OCA2 on melanoma risk phenotypes. Hum. Mol. Genet. *13*, 447–461.
30. Crawford, N.G., Kelly, D.E., Hansen, M.E.B., Beltrame, M.H., Fan, S., Bowman, S.L., Jewett, E., Ranciaro, A., Thompson, S., Lo, Y., et al.; NISC Comparative Sequencing Program (2017). Loci associated with skin pigmentation identified in African populations. Science *358*, eaan8433.
31. Andersen, J.D., Pietroni, C., Johansen, P., Andersen, M.M., Pereira, V., Børsting, C., and Morling, N. (2016). Importance of nonsynonymous OCA2 variants in human eye color prediction. Mol. Genet. Genomic Med. *4*, 420–430.
32. Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J.A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., and Boutkov, B. (2020). Computationally efficient whole genome regression for quantitative and

binary traits. bioRxiv. https://doi.org/10.1101/2020.06.19.162354.
33. Mukherjee, B., Liu, I., and Sinha, S. (2007). Analysis of matched case-control data with multiple ordered disease states: possible choices and comparisons. Stat. Med. *26*, 3240–3257.
34. Casella, G., and Berger, R.L. (2002). Statistical inference (CA: Duxbury Pacific Grove).
35. Breslow, N.E., and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. *88*, 9–25.
36. Gilmour, A.R., Thompson, R., and Cullis, B.R. (1995). Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. Biometrics *51*, 1440–1450.