FastGWA-GLMM: a generalized linear mixed model association tool for biobank-scale data

Jian Yang (jian.yang@westlake.edu.cn)
Westlake University https://orcid.org/0000-0003-2001-2474
Longda Jiang
Institute for Molecular Bioscience, The University of Queensland
Zhili Zheng
The University of Queensland

Technical Report

Keywords: genome-wide association (GWA) methods, software, traits

Posted Date: February 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-128758/v1

Version of Record: A version of this preprint was published at Nature Genetics on November 4th, 2021. See the published version at https://doi.org/10.1038/s41588-021-00954-4.

FastGWA-GLMM: a generalized linear mixed model association tool for biobank-scale data

Longda Jiang[1,$], Zhili Zheng[1,$], Jian Yang[1,2,3,*]

[1]Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland 4072, Australia

[2]School of Life Sciences, Westlake University, Hangzhou, Zhejiang 310024, China

[3]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang 310024, China

[$]Equal contribution

*Correspondence: Jian Yang (jian.yang@westlake.edu.cn)

Abstract

Compared to linear mixed model-based genome-wide association (GWA) methods, generalized linear mixed model (GLMM)-based methods have better statistical properties when applied to binary traits but are computationally much slower. Here, leveraging efficient sparse matrix-based algorithms, we developed a GLMM-based GWA tool (called fastGWA-GLMM) that is orders of magnitude faster than the state-of-the-art tool (e.g., ~37 times faster when $n = 400{,}000$) with more scalable memory usage. We show by simulation that the fastGWA-GLMM test-statistics of both common and rare variants are well-calibrated under the null, even for traits with an extreme case-control ratio (e.g., 0.1%). We applied fastGWA-GLMM to the UK Biobank data of 456,348 individuals, 11,842,647 variants and 2,989 binary traits (full summary statistics available at http://fastgwa.info/ukbimpbin) and identified 259 rare variants associated with 75 traits, demonstrating the use of imputed genotype data in a large cohort to discover rare variants for binary complex traits.

Introduction

Over the past decade, we have witnessed the tremendous growth of data from genome-wide association studies (GWASs). For example, there are nearly half million genotyped individuals with rich phenotypes in the UK Biobank (UKB)[1], which have played a pivotal role in discovering novel genotype-phenotype associations in recent years[2-6]. Nonetheless, the scale of biobank data imposes great computational challenges on methods for genome-wide association (GWA) analysis. New methods and tools have been actively developed for biobank-scale data, including linear regression-based tools such as PLINK2 (ref.[7]) and BGENIE[1], and linear mixed model (LMM)-based tools such as DISSECT[8], BOLT-LMM[9], and fastGWA[10]. LMM-based methods are usually preferred over linear regression-based methods largely because the former can account for relatedness without the need to remove related individuals. Despite that the linear regression- and LMM-based methods are developed under normality assumption, they are often used for binary traits[11-13]. However, recent studies[9,14] show that test statistics from LMM-based methods are inflated under the null when the case-control ratio of the trait of interest is low, leading to an inflated false-positive rate (FPR), particularly for rare variants. To avoid such inflation, a common practice is to remove rare variants (e.g., minor allele frequency, MAF < 0.01) and phenotypes with a low case-control ratio (e.g., < 1:99)[9,10], resulting in unnecessary loss of data.

Compared to LMM-based approaches, generalised linear mixed model (GLMM)-based methods are better suited for GWA analysis for binary traits[14]. Unfortunately, most of the GLMM-based GWA methods are not scalable to large biobank data. SAIGE[14] is one of very few exceptions and is currently the most commonly used GLMM-based tool for biobank-scale data because of its computational efficiency and well-calibrated test-statistics of both common and rare variants for unbalanced binary traits. However, it is almost computationally prohibitive to use SAIGE to analyse all the thousands of binary traits in the UKB, more so in cohorts with larger sample sizes than the UKB (e.g., data accumulated in the direct-to-consumer genetic testing companies). The main reason why the performance of SAIGE is encumbered is because of the manipulation of full-dense $n \times n$ matrices (although not explicitly computed) with n being the sample size, which is both time- and resource-consuming.

In our previous work, we developed an LMM-based GWA tool, fastGWA, that is orders of magnitude faster than BOLT-LMM, mainly owing to the use of a sparse genomic relationship matrix (GRM) to capture pedigree relatedness among individuals[10]. However, when we applied fastGWA in the GWA analyses of all the UKB traits, we had to remove around 3 million rare variants (MAF ≤ 0.01) and ~1,000 traits with a case-control ratio < 1:99 to avoid the inflation in FPR mentioned above[10]. In this study, we aim to develop a GWA tool that is scalable to GWAS data of over a million individuals and applicable to both common and rare variants for all binary phenotypes including those with a low case-control ratio. To achieve this goal, we incorporated GLMM into the fastGWA framework and developed efficient sparse matrix-based algorithms for parameter estimation and association test. We name the method fastGWA-GLMM and demonstrate by simulation that the test-statistics from fastGWA-GLMM are not inflated for either common or rare variants even if the case-control ratio is extremely low (e.g., 0.1%). We then show by analysing subsets of the UKB data that fastGWA-GLMM is orders magnitude faster than SAIGE with more scalable memory usage (e.g., when n = 400,000, fastGWA-GLMM is ~34 times faster and uses only about a third memory compared with SAIGE). From the speed test results, we predict that fastGWA-GLMM is, in principle, applicable to GWAS data with sample sizes over a million. We have implemented fastGWA-GLMM in the GCTA software package[15]. In addition, we have used fastGWA-GLMM to perform GWA for 2,989 binary traits from the UKB and made the full summary statistics publicly accessible at the fastGWA data portal (http://fastgwa.info/ukbimpbin).

Results

Overview of the method

The fastGWA-GLMM model can be written as

$$\text{logit}(\boldsymbol{\mu}) = \boldsymbol{x}_s\beta_s + \mathbf{X}_\text{c}\boldsymbol{\beta}_c + \boldsymbol{g}$$

where $\boldsymbol{y}$ is an n×1 vector of binary phenotypes; $\boldsymbol{\mu}$ is a vector of $\mu_i = P(y_i = 1|x_{s-i}, X_{c-i}, g_i)$ with μ_i being the probability of subject *i* being a case given the subject's genotype x_{s-i}, covariates X_{c-i}, and random genetic effect g_i; $\boldsymbol{x}_s$ is a vector of genotype variables of a variant of interest with its effect β_s; $\mathbf{X}_\text{c}$ is the incidence matrix of fixed-effect covariates (e.g., sex, age and principal components) with their corresponding coefficients $\boldsymbol{\beta}_c$; $\boldsymbol{g}$ is a vector of effects that capture genetic and common environmental effects shared among related individuals, $\boldsymbol{g} \sim N(0, \boldsymbol{\pi}\sigma_g^2)$ with $\boldsymbol{\pi}$ being the pedigree relationship matrix and σ_g^2 being the corresponding variance component. In practice, if pedigree information is unavailable or incomplete, $\boldsymbol{\pi}$ can be replaced by the GRM with all the small off-diagonal elements (e.g., those<0.05) set to zero, i.e., the sparse GRM[10].

The fastGWA-GLMM method comprises two steps: 1) the estimation step: estimating σ_g^2 , $\boldsymbol{\beta}_c$, and the other parameters under the null model (i.e., $\text{logit}(\boldsymbol{\mu}) = \mathbf{X}_\text{c}\boldsymbol{\beta}_c + \boldsymbol{g}$); 2) the association test step: performing score test for each variant and, if necessary, applying saddle point approximation (SPA) to the score statistic to correct for potential inflation driven by case-control imbalance and low MAF (**Online Methods**). In the estimation step, we have developed an extraordinarily efficient method (named fastGWA-GLMM-REML; **Online Methods**) to estimate the variance components in the GLMM in a robust manner even for traits with an extreme case-control ratio (e.g., 0.1%). In the association test step, based on the estimates obtained from the step above, the score test statistic for each variant can be computed by the following equation:

$$T_{score} = \boldsymbol{x}_s^{\mathrm{T}}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) \text{ with } \mathrm{var}(T_{score}) = \boldsymbol{x}_s^{\mathrm{T}}\mathbf{P}\boldsymbol{x}_s$$

$$\frac{T_{score}^2}{\mathrm{var}(T_{score})} \sim \chi_{df=1}^2$$

where **P** is an $n \times n$ projection matrix, which is dense despite $\boldsymbol{\pi}$ being sparse. Therefore, to avoid computational bottleneck due to matrix multiplication involving $\mathbf{P}$, a GLMM version of the GRAMMAR-GAMMA approximation[14] is implemented in fastGWA-GLMM (**Online Methods**).

As for the inflation in test-statistics due to case-control imbalance, for any variant with a score test p-value larger than a threshold (e.g., $\chi_{df=1}^2 = 2$), SPA is applied to calibrate the test statistic. In addition, to further improve computational efficiency, in fastGWA-GLMM, we developed an approximate approach to account for covariates for variants with score test $\chi_{df=1}^2$ smaller than the threshold (**Online Methods**). This strategy greatly reduces the runtime, especially when the number of covariates is large. An alternative version of fastGWA-GLMM without this covariate approximation strategy is also available, which is a few times less efficient depending on the number of covariates (**Online Methods**).

Runtime and resource requirements

We used the UKB data consisting of 456,348 individuals of European ancestry and 11,842,647 variants (**Online Methods**) to evaluate the resource requirements of fastGWA-GLMM in GCTA v1.93.3 and benchmarked it against SAIGE v0.42.1. Note that the standard logistic regression (as implemented in PLINK2 v2.00a2.3) was not included in the runtime comparison because it is >100 times slower than fastGWA-GLMM and not applicable to some of our simulation settings. After randomly sampling subgroups of individuals (*n* ranged from 50,000 to 400,000) from the UKB, we performed a GWA in each subset of data using fastGWA-GLMM and SAIGE respectively, on a computing platform with 80 GB memory and 8 CPU cores. The trait used for comparison is "Irritability" (case-control ratio = 0.39; UKB data-field: 1940). The genotype data were stored in BGEN v1.3 format[16]. Each test was repeated 5 times for an average of runtime and memory usage. As shown in **Figure 1a**, for a GWA with *n* = 400,000, fastGWA-GLMM only required 4.9 hours, which is ~37 times more efficient than SAIGE. Besides, the runtime of the estimation step of fastGWA-GLMM is negligible compared to that of SAIGE (**Supplementary Table 1**). Moreover, the runtime of fastGWA-GLMM was generally stable for traits with different levels of case-control ratio (**Supplementary Figure 1**), so is SAIGE (**Supplementary Table 2**). As for memory requirements, the actual memory usage of fastGWA-GLMM was almost invariant to sample size (~4 GB for *n* ranged from 50,000 to 400,000), while this was not the case for SAIGE, e.g., SAIGE only required 1.88 GB memory for n = 50,000, but the memory usage soon increased to 13.0 GB when n = 400,000 (**Figure 1b** and **Supplementary Table 3**). Our observation that the runtime of fastGWA-GLMM increased almost linearly with sample size with almost invariant memory usage (**Figure 1**) suggests that fastGWA-GLMM is, in principle, scalable to sample sizes over a million given the same computing environment as used in this study.

False-positive rate (FPR) and statistical power

In order to quantify the statistical performance of fastGWA-GLMM in comparison with other methods, including SAIGE[14] and PLINK2 (logistic regression using all individuals or unrelated individuals, denoted as LR-All and LR-unRel, respectively)[7], we generated a sample of 100,000 simulated individuals with substantial population stratification and relatedness from a subset of the real UKB genotype data (**Online Methods**). Based on the simulated genotype data, we randomly sampled a number of causal variants from all variants on the odd chromosomes to simulate phenotypes, leaving the variants on the even chromosomes as the null variants to quantify the type-1 error rate. We also introduced common environmental effects (i.e., non-genetic effects shared among close relatives) and population stratification effects to the phenotype (**Online Methods**). Finally, using the simulated data, we quantified the FRP (i.e., the proportion of null variants with p-values < a threshold) and statistical power (measured by the mean χ^2 statistic at the causal variants) for different association methods.

The results showed that when the prevalence was larger than 0.05, the FPRs of the null variants at five different p-value thresholds (α=0.05, 0.005, 5×10^{-4}, 5×10^{-5}, and 5×10^{-6}) were largely consistent with the expected values for fastGWA-GLMM, SAIGE, and LR-unRel but inflated for LR-All because relatedness was not accounted for in LR-All (**Figure 2** and **Supplementary Figure 2**). When the prevalence was 0.01 or below, both LR-unRel and LR-All showed inflated FPRs, while such inflation was not observed for SAIGE and fastGWA-GLMM. The FPR of SAIGE was slightly more deflated than that of fastGWA-GLMM in all the simulation scenarios (**Figure 2** and **Supplementary Figure 2**). Particularly, in the scenario with prevalence = 0.005, the FPRs of SAIGE were more deflated than those of all the other methods because the parameter estimation process of SAIGE failed to converge in ~25% of the simulation replicates (see below for more discussion).

We partitioned all the null variants into two groups (common and rare variants) based on an MAF threshold of 0.01 and evaluated the FPR of the two groups separately in each simulation scenario. The FPRs for rare variants (MAF < 0.01) from LR-unRel and LR-All were substantially inflated in the scenarios with low prevalences, while those from fastGWA-GLMM remained consistent with the expected values for both common and rare variants regardless of the prevalence level (**Supplementary Figures 3** and **4**). SAIGE showed similar performance as fastGWA-GLMM, except that it showed more deflated FPRs than all the other methods when prevalence = 0.005 due to its convergence issue as described above.

Next, we quantified the statistical power of different methods by calculating the mean χ^2 statistic at the causal variants. We found that the power of fastGWA-GLMM was slightly higher than that of SAIGE (**Figure 3**). The mean χ^2 statistic of LR-All and LR-unRel was not informative in this case as it suffered from inflation driven by both relatedness and case-control imbalance. We then quantified the power of common and rare causal variants separately. The patterns were similar between common and rare variants, though the power to detect the rare causal variants was lower than that for the common causal variants (**Figure 3**). We also used the area under the curve (AUC) as a metric to compare the difference in power between the methods given the same level of FPR (**Online Methods**). In almost all the scenarios, SAIGE, fastGWA-GLMM, and LR-All showed similar AUCs while LR-unRel showed lower AUCs than the other methods because of its smaller sample size (**Supplementary Figure 5**). The only exception is the scenario with prevalence = 0.001, in which LR-All and LR-unRel showed higher AUCs than fastGWA-GLMM and SAIGE, possibly due to the overcorrection of GLMM and/or SPA under this extreme condition. Nevertheless, since the FPRs of LR-All and LR-unRel were heavily inflated when prevalence = 0.001 (**Figure 2**), the higher power for LR in this scenario is not practically meaningful. In addition, we showed that the test statistics of fastGWA-GLMM remained well-calibrated when cases were oversampled (**Supplementary Figures 6**-**8**). We further demonstrated that when pedigree information was fully available, fastGWA-GLMM using pedigree relationship matrix performed almost equally well as that using the sparse GRM (**Supplementary Note**; **Supplementary Figures 9** and **10**).

Application of fastGWA-GLMM to 2,989 binary traits in the UKB

We used fastGWA-GLMM to conduct GWA analyses of 11,842,647 imputed variants in all the UKB participants of European ancestry (*n*=456,348) for 2,989 binary phenotypes. These binary phenotypes were either generated from the analysis pipelines used by the Neale Lab (http://www.nealelab.is/uk-biobank) or from our in-house ICD-10-to-PheCode pipeline using map from ref.[17] (**Online methods**). To benchmark fastGWA-GLMM against SAIGE and PLINK2 LR-unRel (note: *n*=348,456 for LR-unRel), we selected eight representative phenotypes (prevalence ranging from 0.0008 to 0.45; **Supplementary Table 4**) from the 2,989 traits. Based on the summary statistics from each method for the eight traits, we noticed that overall fastGWA-GLMM identified more genome-wide significant loci than SAIGE or LR-unRel (**Supplementary**

Figure 11 and **Supplementary Table 5**). The difference was more apparent when the prevalence of the trait was moderate to high (≥ 0.1) (**Supplementary Figure 11a-c**) and became less significant as the prevalence decreased (**Supplementary Figure 11d-h**). Additionally, after clumping, the number of quasi-independent signals from fastGWA-GLMM was also higher than that from SAIGE or LR-unRel (**Supplementary Table 6**). As for case-control imbalance, the results from LR-unRel started to exhibit inflation for traits with prevalence < 0.01 (see the 3rd panels of **Supplementary Figure 11e-h**), consistent with our simulation results, and the inflation was more prominent for the rare variants (see the 3rd panels of **Supplementary Figure 12e-h**). Meanwhile, the results from fastGWA-GLMM and SAIGE remained robust for traits with low prevalence. Among all the 2,989 traits analysed, we identified 326 pairs of quasi-independent genome-wide significant associations between 259 rare variants (MAF < 0.01 and p-value $\leq 5\times10^{-9}$) and 75 traits (**Supplementary Table 7**, **Online Methods**). Of the 259 rare variants, 37 are located in either the exonic regions or the 3' or 5' UTRs (**Supplementary Table 7**), highlighting the enrichment of rare variants in the coding and UTR regions (enrichment p-value = 9.6×10^{-5}, **Supplementary Note**).

We have previously developed an online tool to query and visualize the GWAS results of over 2,000 phenotypes from the UKB[10]. Similarly, the association results of the 2,989 binary phenotypes from this study are also freely available for visualization and downloading through our fastGWA data portal at http://fastgwa.info/ukbimpbin.

Discussion

In this study, we developed an association method, fastGWA-GLMM, with extraordinary performance in computational efficiency, for GWA analyses of binary phenotypes in large cohorts such as the UKB. Tested in a dataset of 400,000 individuals and 11,842,647 variants, fastGWA-GLMM is ~37 times faster than SAIGE (the most efficient existing method for binary traits). Besides, the implementation of GLMM framework allows users to retain the maximum number of individuals in a GWA analysis in the presence of relatedness, and the incorporation of SPA correction properly calibrates the test-statistics for traits with extreme case-control ratios. The application of fastGWA-GLMM to 2,989 binary traits in the UKB further demonstrated its utility and efficiency.

The major advantage of fastGWA-GLMM over LR-unRel is that it does not need to remove related individuals from the study, as the relatedness can be well accounted for by a pedigree relatedness matrix or a sparse GRM. Take the real data application in the UKB as an example. FastGWA-GLMM was able to include all 456,348 participants into the association test, while LR-unRel could only utilize information from 348,456 unrelated participants. Since most of the large population-based cohorts rely on an assessment-centre based recruitment strategy, the proportion of relatives in the cohorts tends to be high and will keep increasing in the future[1]. In such case, it is crucial to avoid removing data of related individuals. Another advantage of fastGWA-GLMM over LR-unRel is its efficiency. FastGWA-GLMM, as many other GLMM-based methods, uses a score statistic for association test, which is computationally easy to compute (**Online Methods**). In contrast, LR-unRel as in PLINK2 is based on an iteratively reweighted least squares method and the Wald's test that solves the full model for each variant repeatedly, which is much slower than the score test especially when covariates are included.

The advantages of fastGWA-GLMM over LMM-based methods, including the original fastGWA method[10], can be summarized into two aspects. The first is the better interpretability of the effect sizes, as we can directly use natural logarithm to convert the $\hat{\beta}_s$ from fastGWA-GLMM into odds ratio (**Supplementary Note**). However, such transformation in LMM-based methods is indirect and requires sophisticated approximations[18]. The second aspect is the better-controlled FPR of fastGWA-GLMM by the SPA correction. Since SPA correction was only designed for GLMMs but not LMMs[19], a common strategy for LMM-based methods to mitigate such inflation is to exclude any trait with a small case-control ratio (e.g., ≤ 1:99) and any variant with a low MAF (e.g., < 0.01)[9,10]. Yet, excluding them causes significant loss of valuable information. For instance, the 3,821,959 rare variants tested in this study would have been removed from the analyses using the LMM-based methods, among which we identified hundreds of variants associated with the traits at a very stringent significance level and some of them are known (**Supplementary Table 7**). For example, we identified a rare missense variant in the *HOXB13*, rs138213197, strongly associated with prostate cancer, and this association had also been reported repeatedly in previous studies[20-22].

SAIGE is a GLMM-based method that uses a dense GRM. Apart from the GRM setting, there are another two major differences between fastGWA-GLMM and SAIGE. The first difference is that fastGWA-GLMM uses a grid search-based algorithm, fastGWA-GLMM-REML, to estimate the variance components (**Online Methods**), which is more robust and often orders of magnitude more efficient than the average information (AI) REML algorithm used in SAIGE even for traits with extreme case-control ratios. We observed that under the simulation scenario with prevalence = 0.005, the variance estimation procedure of SAIGE failed to converge for 26 out of 100 simulation replicates. The second difference is that instead of using covariate-adjusted genotype data to calculate a score test statistic for every variant, fastGWA-GLMM first uses unadjusted (but mean-centred) genotype data to calculate an approximate score test statistic, and then re-calculate the exact test statistic using the covariate-adjusted genotype data only if the p-value from the approximate score test is smaller than a threshold (by default, $\chi^2_{df=1} > 2$); note that the SPA correction is also applied when this threshold is met (**Methods**). This strategy allows fastGWA-GLMM to omit the computation of matrix multiplication between the covariate matrix and ~95% of the genotype vectors. We confirmed that the difference of test statistics between the approximate covariate-adjustment approach and the exact approach is negligible, and only variants with $\chi^2_{df=1} < 2$ might suffer from slight deflation in test-statistics which does affect the power of detecting association at a genome-wide significance level (**Supplementary Figure 13**). This strategy is particularly useful when the number of covariates is large (e.g., larger than 20). In our software tool, there is an option to allow users to switch off this approximation and force all the statistics to be calculated by the covariate-adjusted genotypes, which will cause a loss of computational efficiency by a few folds, depending on the number of covariates.

There are a few caveats when applying fastGWA-GLMM in practice. First, if pedigree data are not usable, a sparse GRM needs to be pre-computed from the SNP data. A very efficient parallelized algorithm has been implemented in GCTA to compute the sparse GRM[10]. Since the sparse GRM setting has already been adopted by fastGWA[10], once generated, the same sparse GRM of a cohort can be used for GWA analyses of all the quantitative and binary phenotypes. Therefore, the average computational cost per trait is minimal. Second, the $\hat{\sigma}^2_g$ estimated from fastGWA-GLMM-REML cannot be interpreted as genetic variance or heritability. This is mainly due to the use of the penalized quasi-likelihood and the Laplace method[14]. However, from our simulations and real data applications, it did not affect the statistical performance of the association test of fastGWA-GLMM. Third, in our previous work, we found that when analysing quantitative traits the $\hat{\sigma}^2_g$ estimated based on a sparse GRM might be a better quantity to control for relatedness than that from dense-GRM-based methods[10]. In this study, however, when analysing binary traits, we observed that fastGWA-GLMM did not have such advantage over SAIGE. Both fastGWA-GLMM and SAIGE had well-controlled FPR. Fourth, the inclusion of rare variants in the association tests increases the multiple testing burden. Hence, in this study, following the guideline from previous studies[23,24], we used p-value ≤ 5×10^{-9} instead of 5×10^{-8} as the genome-wide significance threshold.

Despite these caveats, fastGWA-GLMM is a highly efficient GLMM-based method that is applicable to GWA analyses of a large number of binary phenotypes in biobank-scale data. The extensive simulations under different parameter settings and the real-data analyses of nearly 3,000 UKB traits have together manifested its statistical robustness and computational efficiency. We believe that fastGWA-GLMM is a very useful tool for current and up-coming large-scale data, and the summary statistics released from this study will be useful for future studies to give insights into the genetic basis of many health-related outcomes.

ONLINE METHODS

Estimating the variance components

As described in the Results section, the fastGWA-GLMM model can be written as $\text{logit}(\boldsymbol{\mu}) = \boldsymbol{x}_s\beta_s + \mathbf{X}_\mathrm{c}\boldsymbol{\beta}_c + \boldsymbol{g}$. The logit function, $\text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$, is a commonly used link function in GLMM that links the expectation of the dependent binary variable $\boldsymbol{y}$ to a linear predictor that involves the independent variables. Solving this full model repeatedly for each variant is computationally unfeasible in large samples, so a common strategy is to first solve σ_g^2 as well as the other essential components under the null model, i.e., $\text{logit}(\boldsymbol{\mu}) = \mathbf{X}_\mathrm{c}\boldsymbol{\beta}_c + \boldsymbol{g}$, and then calculate the score statistic for each variant based on the estimates from the null model. This strategy has been adopted by many existing LMM and GLMM methods[14,25-33].

Following GMMAT[33] and SAIGE[14], the log quasi-likelihood of the null model is

$$ql(\boldsymbol{\beta}_c, \sigma_g^2) = \log \int \exp\left\{\sum_{i=1}^{n} ql_i(\boldsymbol{\beta}_c|\boldsymbol{g})\right\} \times (2\pi)^{-0.5n}\left|\boldsymbol{\pi}\sigma_g^2\right|^{-0.5} \times \exp\{-0.5\boldsymbol{g}^T(\boldsymbol{\pi}\sigma_g^2)^{-1}\boldsymbol{g}\}d\boldsymbol{g}$$

where $ql_i(\boldsymbol{\beta}_c|\boldsymbol{g}) = \int_{y_i}^{\mu_i}\frac{a_i(y_i-\mu)}{\mu_i(1-\mu_i)}d\mu$ is the quasi-likelihood for the i^{th} individual given the random effect $\boldsymbol{g}$, and a_i is a known constant which will be omitted during the derivation. Following the derivations in ref.[33], we have $\widehat{\boldsymbol{\beta}}_c = (\mathbf{X}_\mathrm{c}^\mathrm{T}\mathbf{V}^{-1}\mathbf{X}_\mathrm{c})^{-1}\mathbf{X}_\mathrm{c}^\mathrm{T}\mathbf{V}^{-1}\widetilde{\mathbf{Y}}$ and $\widehat{\boldsymbol{g}} = \sigma_g^2\boldsymbol{\pi}\mathbf{V}^{-1}(\widetilde{\mathbf{Y}} - \mathbf{X}_\mathrm{c}\widehat{\boldsymbol{\beta}}_c)$, where $\mathbf{V}$ is a variance-covariance matrix (i.e., $\mathbf{V} = \mathbf{W}^{-1} + \boldsymbol{\pi}\sigma_g^2$) with $\mathbf{W}$ being a diagonal matrix (i.e., $w_{ii} = \mu_i(1-\mu_i)$), and $\widetilde{\mathbf{Y}}$ is the so-called 'working vector' with $\widetilde{\mathbf{Y}} = \mathbf{X}_\mathrm{c}\boldsymbol{\beta}_c + \boldsymbol{g} + \text{logit}'(\boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})$. Given $\widehat{\boldsymbol{\beta}}_c$ and $\widehat{\boldsymbol{g}}$, the restricted maximum likelihood (REML) version of $\text{ql}(\boldsymbol{\beta}_c, \sigma_g^2)$ can be written as

$$\text{ql}(\boldsymbol{\beta}_c, \sigma_g^2) = \text{const} - 0.5\log|\mathbf{V}| - 0.5\log\left|\mathbf{X}_\mathrm{c}^\mathrm{T}\mathbf{V}^{-1}\mathbf{X}_\mathrm{c}\right| - 0.5\widetilde{\mathbf{Y}}^\mathrm{T}\mathbf{P}\widetilde{\mathbf{Y}}$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}_\mathrm{c}(\mathbf{X}_\mathrm{c}^\mathrm{T}\mathbf{V}^{-1}\mathbf{X}_\mathrm{c})^{-1}\mathbf{X}_\mathrm{c}^\mathrm{T}\mathbf{V}^{-1}$. An iterative approach is required to compute the quasi-likelihood and estimate the parameters. A commonly used algorithm is the average information REML (AI-REML)[34], which has been adopted by both GMMAT[33] and SAIGE[14]. Leveraging the sparsity of $\boldsymbol{\pi}$, we propose a grid-search-based REML approach (called fastGWA-GLMM-REML) with a special optimizer (see below) that can directly maximize $\text{ql}(\boldsymbol{\beta}_c, \sigma_g^2)$ and return a maximum likelihood estimate of σ_g^2, which is often orders of magnitude faster and more robust (especially for traits with extremely unbalanced case-control ratios) than AI REML. A brief summary of fastGWA-GLMM REML is shown as follows.

1) Let subscript *i* denote the iteration step with *i* starting from 0;
2) $\widehat{\boldsymbol{\beta}}_{c(i)}$ is estimated from a standard logistic regression (i.e., $\text{logit}(\boldsymbol{\mu}) = \mathbf{X}_\mathrm{c}\boldsymbol{\beta}_c$), which is used as the starting value for $\boldsymbol{\beta}_c$;

3) $\hat{\sigma}^2_{g(i)}$ and $\hat{\boldsymbol{g}}_i$ are set to 0;

4) Calculate $\hat{\boldsymbol{\mu}}_i = \text{logit}^{-1}\left(\mathbf{X}_c\hat{\boldsymbol{\beta}}_{c(i)} + \hat{\boldsymbol{g}}_i\right)$;

5) Calculate $\tilde{\mathbf{Y}}_i = \mathbf{X}_c\hat{\boldsymbol{\beta}}_{c(i)} + \hat{\boldsymbol{g}}_i + \frac{\mathbf{y}-\hat{\boldsymbol{\mu}}_i}{\hat{\boldsymbol{\mu}}_i(\mathbf{1}-\hat{\boldsymbol{\mu}}_i)}$,

$\mathbf{W}_i = \text{diag}\{\hat{\boldsymbol{\mu}}_i(\mathbf{1} - \hat{\boldsymbol{\mu}}_i)\}$;

6) Perform fastGWA-GLMM-REML to estimate $\hat{\sigma}^2_{g(i+1)}$ given $\hat{\boldsymbol{\beta}}_{c(i)}$ and $\hat{\boldsymbol{g}}_i$ (see details in next section);

7) Calculate $\mathbf{V}_{i+1} = \mathbf{W}_i^{-1} + \hat{\sigma}^2_{g(i+1)}\boldsymbol{\pi}$;

8) Calculate $\hat{\boldsymbol{\beta}}_{c(i+1)} = (\mathbf{X}_c{}^{\mathrm{T}}\mathbf{V}_{i+1}^{-1}\mathbf{X}_c)^{-1}\mathbf{X}_c{}^{\mathrm{T}}\mathbf{V}_{i+1}^{-1}\tilde{\mathbf{Y}}_i$,

$\hat{\boldsymbol{g}}_{i+1} = \hat{\sigma}^2_{g(i+1)}\boldsymbol{\pi}\mathbf{V}_{i+1}^{-1}[\tilde{\mathbf{Y}}_i - \mathbf{X}_c\hat{\boldsymbol{\beta}}_{c(i+1)}]$;

9) Set $i = i + 1$ and repeat 4) to 8) until both $\left(\frac{||\hat{\boldsymbol{\beta}}_{c(i+1)}|-|\hat{\boldsymbol{\beta}}_{c(i)}||}{|\hat{\boldsymbol{\beta}}_{c(i+1)}|+|\hat{\boldsymbol{\beta}}_{c(i)}|}\right)_{max}$ and $\frac{\left|\hat{\sigma}^2_{g(i+1)}-\hat{\sigma}^2_{g(i)}\right|}{\hat{\sigma}^2_{g(i+1)}+\hat{\sigma}^2_{g(i)}} \leq$ a threshold (by default 5×10^{-5}).

We use the sparse matrix Cholesky decomposition algorithm implemented in the Eigen C++ library (http://eigen.tuxfamily.org) to compute the terms involving $|\mathbf{V}|$ or $\mathbf{V}^{-1}$ in a very efficient manner.

The grid-search-based fastGWA-GLMM-REML optimizer

As mentioned above, we have developed a grid-search-based optimizer to estimate σ^2_g, which is more robust and often orders of magnitude faster than AI-REML. In the i^{th} iteration of the estimation step of the fastGWA-GLMM REML method, the grid-search-based optimizer runs as follows.

1) Wide-range search. We set a grid of k values of $\hat{\sigma}^2_g$, i.e., $[l^{(i)},\ u^{(i)}]$, compute $\text{ql}\left(\boldsymbol{\beta}_c, \sigma^2_g\right)$ given each value of $\hat{\sigma}^2_g$, and select the flanking grids of the $\hat{\sigma}^2_g$ value that produces the maximum quasi-likelihood to form a finer-scale searching interval (denoted by $[\hat{\sigma}^2_{low,0}, \hat{\sigma}^2_{up,0}]$) for the fine-tuning step below.

2) Fine-tuning search. Similar as the process above, we divide $[\hat{\sigma}^2_{low,0}, \hat{\sigma}^2_{up,0}]$ into a grid of 16 $\hat{\sigma}^2_g$ values, compute $\text{ql}\left(\boldsymbol{\beta}_c, \sigma^2_g\right)$ given each value of $\hat{\sigma}^2_g$, and select the flanking grids of the $\hat{\sigma}^2_g$ value that produces the maximum quasi-likelihood to form a finer-scale searching interval (denoted by $[\hat{\sigma}^2_{low,1}, \hat{\sigma}^2_{up,1}]$). This fine-tuning step is repeated 4 times, and $\hat{\sigma}^2_{(max)} = (\hat{\sigma}^2_{low,5} + \hat{\sigma}^2_{up,5})/2$ is returned as an estimate of $\sigma^2_{g(i)}$ for the i^{th} iteration of fastGWA-GLMM-REML.

The $l^{(\mathrm{i})}$ is the lower bound of the grid which is set to 0 when $i \leq 3$ or $\frac{\sigma^2_{g(i-3)}+\sigma^2_{g(i-2)}+\sigma^2_{g(i-1)}}{3} \leq 0.1$, and set to $\frac{\sigma^2_{g(i-3)}+\sigma^2_{g(i-2)}+\sigma^2_{g(i-1)}}{3} \times 0.8$ when $i > 3$ and $\frac{\sigma^2_{g(i-3)}+\sigma^2_{g(i-2)}+\sigma^2_{g(i-1)}}{3} > 0.1$. Similarly, the $u^{(\mathrm{i})}$

is the upper bound of the grid which is set to $\widetilde{\mathbf{Y}}_i^2$ when $i = 1$, set to $10\sigma^2_{g(i-1)}$ when $i = 2$ and 3, and set to $\frac{\sigma^2_{g(i-3)}+\sigma^2_{g(i-2)}+\sigma^2_{g(i-1)}}{3} \times 1.2$ when $i \geq 3$. The k is the number of steps in the grid which is set to 800 when $i = 1$, set to 200 when $i = 2$ and 3, and set to 50 when $i \geq 3$. We apply the settings above to determine the boundaries and grid steps given the observation from simulations that 3 iterations are sufficient to identify a reasonable interval for σ_g^2. The reason why we do not adopt the commonly used conventional optimizers (e.g., the golden-section search) is that the domain of $\mathrm{ql}(\boldsymbol{\beta}_c, \sigma_g^2)$ does not always cover the whole range of $[0, \widetilde{\mathbf{Y}}_i^2]$. Therefore, it is difficult to choose an appropriate searching interval $[l^{(i)},\ u^{(i)}]$ for the conventional optimizers, which would lead to a local optimum. On the other hand, the main reason why we do not adopt AI-REML as implemented in SAIGE[14] is that AI-REML often fails to converge when the case-control ratio is low. For example, in our simulations, SAIGE did not converge in 26 out of 100 simulation replicates under the simulation scenario with prevalence = 0.005.

Computing the score test statistic by the GRAMMAR-GAMMA approximation

As mentioned above, the fastGWA-GLMM method comprises two steps, the estimation step and the association test step. After obtaining all the necessary estimates from the null model in the estimation step, we can test the association of each variant using the score test: $T_{score} = \widetilde{\boldsymbol{x}}_s^{\mathrm{T}}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})$ with $\mathrm{var}(T_{score}) = \widetilde{\boldsymbol{x}}_s^{\mathrm{T}}\mathbf{P}\widetilde{\boldsymbol{x}}_s$, where $\widetilde{\boldsymbol{x}}_s$ is the covariate-adjusted genotype vector with $\widetilde{\boldsymbol{x}}_s = \boldsymbol{x}_s - \mathbf{X}_c\left(\mathbf{X}_c^{\mathrm{T}}\mathbf{W}\mathbf{X}_c\right)^{-1}\mathbf{X}_c^{\mathrm{T}}\mathbf{W}\boldsymbol{x}_s$. We know from the prior work[14] that $\mathbf{X}_c^{\mathrm{T}}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) = \mathbf{0}$ and $\mathbf{P}\mathbf{X}_c\left(\mathbf{X}_c^{\mathrm{T}}\mathbf{W}\mathbf{X}_c\right)^{-1}\mathbf{X}_c^{\mathrm{T}}\mathbf{W} = \mathbf{0}$, we then have $T_{score} = \widetilde{\boldsymbol{x}}_s^{\mathrm{T}}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) = \boldsymbol{x}_s^{\mathrm{T}}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})$ and $\mathrm{var}(T_{score}) = \widetilde{\boldsymbol{x}}_s^{\mathrm{T}}\mathbf{P}\widetilde{\boldsymbol{x}}_s = \boldsymbol{x}_s^{\mathrm{T}}\mathbf{P}\boldsymbol{x}_s$. The score test p-value can be computed based on $\frac{T_{score}^2}{\mathrm{var}(T_{score})} \sim \chi^2_{df=1}$.

T_{score} can be computed efficiently as it only involves vector multiplication and $(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})$ only needs to be calculated once. However, $\mathrm{var}(T_{score})$ is difficult to obtain since $\mathbf{P}$ is an $n \times n$ dense matrix, and $\boldsymbol{x}_s^{\mathrm{T}}\mathbf{P}\boldsymbol{x}_s$ needs to be evaluated repeatedly for every variant. The GRAMMAR-GAMMA approximation is a method to tackle this problem in LMM-based GWA analysis for quantitative traits[30], and has been extended to cope with GLMMs in SAIGE[14]. In brief, for a random variant, its gamma ratio ($\gamma = \frac{\widetilde{\boldsymbol{x}}_s^{\mathrm{T}}\mathbf{P}\widetilde{\boldsymbol{x}}_s}{\widetilde{\boldsymbol{x}}_s^{\mathrm{T}}\mathbf{W}\widetilde{\boldsymbol{x}}_s}$) is approximately constant regardless of its genotypes. The denominator is easy to compute because $\mathbf{W}$ is an $n \times n$ diagonal matrix. Therefore, by randomly selecting m variants (the default m value is 200 in fastGWA-GLMM), we first estimate the mean of the gamma ratio by $\hat{\gamma} = \frac{1}{m}\sum\frac{\widetilde{\boldsymbol{x}}_s^{\mathrm{T}}\mathbf{P}\widetilde{\boldsymbol{x}}_s}{\widetilde{\boldsymbol{x}}_s^{\mathrm{T}}\mathbf{W}\widetilde{\boldsymbol{x}}_s}$ and then calculate $\mathrm{var}(T_{score}) \approx \hat{\gamma}\widetilde{\boldsymbol{x}}_s^{\mathrm{T}}\mathbf{W}\,\widetilde{\boldsymbol{x}}_s$ for all the variants[14]. This strategy avoids computing $\mathrm{var}(T_{score}) = \widetilde{\boldsymbol{x}}_s^{\mathrm{T}}\mathbf{P}\widetilde{\boldsymbol{x}}_s$ repeatedly for each variant and reduces the computational complexity of the association test step to nearly $O(mn)$. The runtime can be further reduced by an approximate covariate adjustment approach, especially when the number of covariates is large (e.g., $c > 20$). The full derivation of the approximate covariate adjustment approach has been described in the **Supplementary Note**. We observed from real data applications that the difference between the test statistics of the approximate and exact methods was very small (**Supplementary Figure 13**). In our software tool, users can mute the approximation method, and in that case, it is a few times slower than the default version, depending on the size of *c*.

Correcting for genomic inflation by saddle point approximation

After obtaining the score test statistics, we calibrate the fastGWA-GLMM p-values by saddle point approximation (SPA)[35,36] to avoid potential inflation driven by case-control imbalance. The SPA method has recently been improved to cope with GWAS data (called fastSPA)[14,19]. FastSPA was originally implemented in R[19]. To improve the computational efficiency, we implemented fastSPA by highly optimised C++ codes in fastGWA-GLMM. By default, fastGWA-GLMM applies the fastSPA correction to variants with $\chi^2_{df=1} \leq 2$.

The UK Biobank data

The UK Biobank (UKB) is a large cohort study consisting of approximately 500,000 participants aged between 40 and 69 at recruitment, with extensive phenotypic records[1]. In this study, 456,348 UKB participants of European ancestry were selected for simulation and real data analyses. Genetic data were genotyped by two different arrays, the Applied Biosystems™ UK Biobank Axiom™ Array and the Applied Biosystems™ UK BiLEVE Axiom™ Array[1]. SNP imputation was conducted by the UKB analysis team using whole-genome sequence data from the Haplotype Reference Consortium[37] and the UK10K project[38] as the reference panels. The imputed data were filtered with standard QC criteria in PLINK2[7], e.g., MAF ≥ 0.0001, Hardy-Weinberg Equilibrium test $P \geq 10^{-6}$, genotyping rate ≥ 0.9, and imputation info score ≥ 0.8, resulting in 11,842,647 imputed variants (8,020,670 common and 3,821,977 rare). Note: we used 588,927 genotyped variants for the simulation study and 11,842,647 imputed variants for real data analyses[1].

Simulation

To assess the statistical performance of fastGWA-GLMM, we simulated 100,000 artificial individuals with a moderate proportion of relatives (10% of all samples) and substantial population stratification (to mimic two different ancestry backgrounds). A "mosaic-chromosome" scheme modified from ref.[32] was used to generate the artificial individuals (see ref.[10] for detailed description of the simulation settings). The difference of the current simulation process with that from ref.[10] was the inclusion of 32,658 genotyped rare variants from the UKB (MAF ranging from 0.01 to 0.0001).

A set of different parameters were used to simulate a binary phenotype. We started from simulating a quantitative phenotype for the 100,000 simulated individuals based on the model below

$$\boldsymbol{y} = \boldsymbol{g}_{com} + \boldsymbol{g}_{rare} + \boldsymbol{z}b_p + \boldsymbol{e}_C + \boldsymbol{e}$$

where $\boldsymbol{g}_{com} = \sum_{i=1}^{m_1} \boldsymbol{x}_{com-i} b_{com-i}$ is the sum of the genetic effects of m_1 common causal variants (MAF ≥ 0.01) with $\boldsymbol{x}_{com-i}$ being a vector of variant genotypes and $b_{com-i} \sim N(0,1)$; $\boldsymbol{g}_{rare} = \sum_{i=1}^{m_2} \boldsymbol{x}_{rare-i} b_{rare-i}$ is the sum of the genetic effects of m_2 rare causal variants (MAF < 0.01) with $\boldsymbol{x}_{rare-i}$ being a vector of variant genotypes and $b_{rare-i} \sim N(0,1)$; $\boldsymbol{z}$ is a vector consisting of 0 (British) and 1 (Irish) to indicate ancestry with b_p being the mean difference in phenotype between the two groups; $\boldsymbol{e}_C$ is a vector of common environmental effects shared among individuals in the same families with $\boldsymbol{e}_C \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_C^2)$; and $\boldsymbol{e}$ is a vector of residuals with $e \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_e^2)$. The causal variants ($m_1 = 10{,}000$ and $m_2 = 1{,}000$) were randomly sampled from variants on the odd chromosomes, so the variants on the even chromosomes could be treated as the null variants to quantify type-1 error rate. We varied the variance of the common environmental effects in different simulation scenarios including (see **Supplementary Note** for detailed description of the parameter settings):

a) no common environmental effects (denoted by "noEnv");
b) common environmental effects explaining 10% of V_p among the 1st and 2nd degree relatives (denoted by "comEnv");

After obtaining the quantitative phenotypic value for each individual, we dichotomized the phenotype given seven sample prevalence rates (i.e., 0.3, 0.2, 0.1, 0.05, 0.01, 0.005, or 0.001) to convert it to a binary phenotype. Each simulation was repeated 100 times.

Assessing the false positive rate and statistical power

Four different methods, SAIGE, LR-All, LR-unRel, and fastGWA-GLMM, were used to conduct GWA analyses for the simulated data. The top 10 principal components (PCs) computed from a set of LD-pruned variants (MAF ≥ 0.01, window size = 1 Mb, step size = 50 variants, and LD r^2 threshold = 0.05) using flashPCA2 (ref.[39]) were included in the association analysis as fixed covariates. For fastGWA-GLMM, 538,752 common variants with MAF ≥ 0.01 were used to compute the sparse GRM, while for SAIGE, as recommended in the software documentation, a set of 78,295 LD-pruned common variants were used as "ModelSNPs" for the estimation of the additive genetic variance (**Supplementary Note**). After performing GWA analyses of the simulated data, we quantified the

FPR using the null variants on the even chromosomes and the power using the mean χ^2 of the causal variants for each method in each simulation scenario. We additionally evaluated the area under the curve (AUC) for each method, which can be interpreted as how well a method ranks true positives above true negative (**Supplementary Note**). Moreover, we also measured the statistical performance of each methods for common (MAF ≥ 0.01) and rare (MAF < 0.01) variants separately.

Real data analyses

We used fastGWA-GLMM to perform GWA analyses of 2,989 binary traits in the UKB. Participants with imputed SNP data and labelled as European ancestry (UKB data-field 1001) were included in the analyses (n=456,348 and m=11,842,647). Of all the traits, 2,154 were generated based on the QC pipeline provided by the Neale Lab (https://github.com/Nealelab/UK_Biobank_GWAS), which were either originally dichotomous or transformed from multi-categorical traits. The rest traits were generated from the ICD-10 records from the UKB. The original ICD-10 records provided by the UKB were text-based data (UKB data-field 41202), which were not easy to process. Therefore, we first extracted every unique ICD-10 code for each individual, and then grouped the ICD-10 codes into different PheCode based on the PheCode v1.2 ICD-10 map[17]. Any individual not labelled with a particular PheCode was treated as a control for that PheCode. We did not remove individuals with relevant diseases from the control group to avoid selection bias[40]. Eventually, 835 PheCode traits were retained for further analysis. We removed traits with n_{cases} < 100 or n_{total} < 5,000 and retained 2,989 traits in total. We fitted age, age^2, sex, age×sex, age^2×sex, and the top 20 PCs provided by the UKB as covariates in the GWA analysis (note: only age, age^2, and the top 20 PCs were fitted for the sex-specific traits). We also applied SAIGE to all the 456,348 individuals and PLINK2 logistic regression to 348,501 unrelated individuals for eight binary phenotypes selected from the UKB for comparison with fastGWA-GLMM. The same covariates were fitted, and details of the parameter settings of SAIGE and PLINK2 are described in the **Supplementary Note**. Clumping analyses were performed using the GWAS results from fastGWA-GLMM, SAIGE, and PLINK2, respectively (LD-clumping parameters used: p-value threshold=5×10^{-9}, window size=5Mb, and LD r^2 threshold=0.01) for each of the eight phenotypes.

Statistical testing

In all the association analyses, we used a $\chi^2_{df=1}$ statistic to test against the null hypothesis of no association (i.e., $H_0: T_{score} = 0$).

Code availability

fastGWA-GLMM is integrated in the GCTA software (http://cnsgenomics.com/software/gcta).

Acknowledgements

This research was supported by the Australian Research Council (FT180100186), the Australian National Health and Medical Research Council (1113400), and the Westlake Education Foundation. We thank J. Sidorenko for assistance in preparation of the UK Biobank data, and Alibaba Cloud - Australia and New Zealand for hosting the online tool. This study makes use of data from the UK Biobank (applications: 12505 and 66982). UK Biobank was established by the Wellcome Trust medical charity, Medical Research Council, Department of Health, Scottish Government and the Northwest Regional Development Agency. It has also had funding from the Welsh Assembly Government, British Heart Foundation and Diabetes UK.

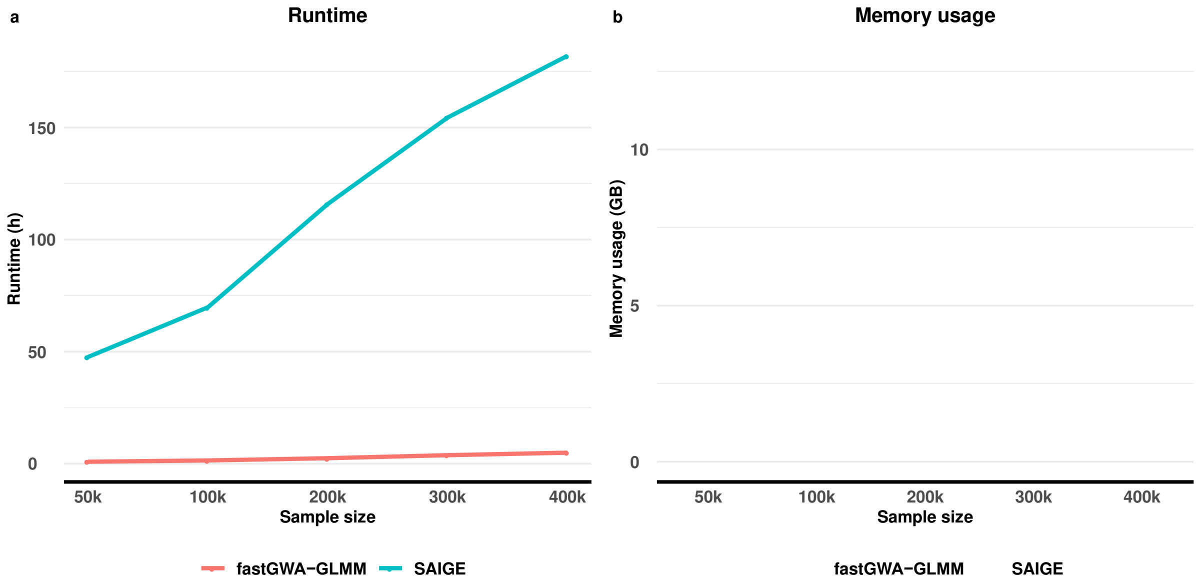

Figure 1. Comparison of runtime and memory usage between fastGWA-GLMM and SAIGE. In panel a), the x-axis represents the sample size, and the y-axis represents the runtime in hour units. For both fastGWA-GLMM and SAIGE, the runtime consists of two components: 1) the estimation of mixed model parameters ("Para. Est."), and 2) the association test ("Assoc."). In panel b), the x-axis represents the sample size, and the y-axis represents the memory usage in GB units. The data used in the tests consisted of 11,842,647 variants, of which 114,494 LD-pruned variants were used as "model SNPs" in SAIGE (**Supplementary Note**). All tests were performed in the same computing environment: 80 GB memory and 8 CPU cores (Intel Xeon Gold 6148). Each test was repeated 5 times for an average.

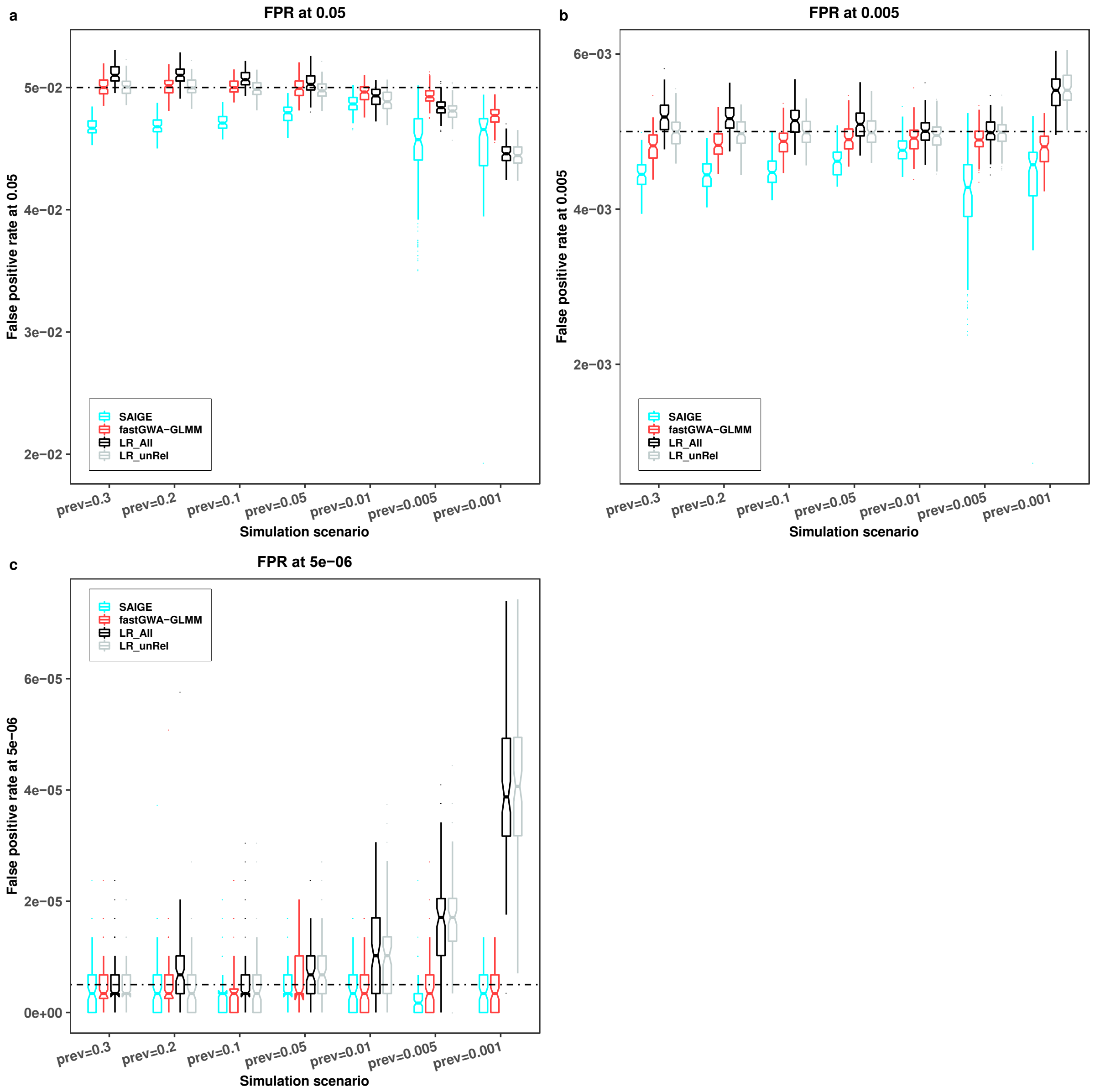

Figure 2. False positive rate (FPR) computed from the null variants. The y-axis represents the FPR computed from the null variants (i.e., all the variants on the even chromosomes), and the x-axis represents different levels of prevalence of the simulated binary phenotypes (prevalence $= n_{case}/(n_{case} + n_{control})$). FPR is evaluated at five different p-value thresholds (α=0.05, 0.005, and 5×10^{-6}), as shown from panels **a** to **c**. The dashed lines indicate the expected FPR (i.e., the alpha level). Each boxplot represents the distribution of FPR across 100 simulation replicates. The line inside each box indicates the median value, notches indicate the 95% confidence interval, central box indicates the interquartile range (IQR), whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate dots.

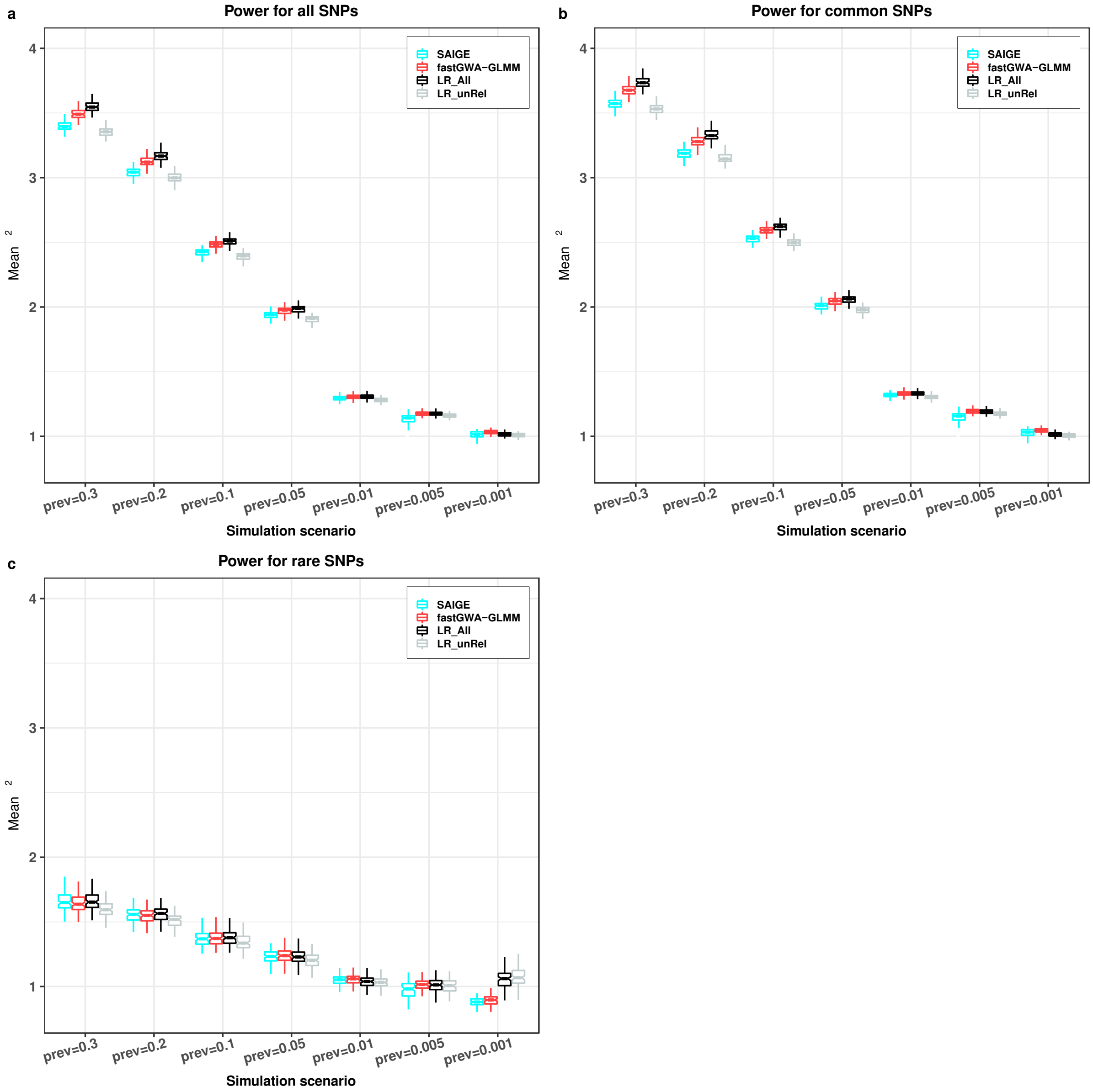

Figure 3. Comparison in power between the methods. Here, power is measured by the mean χ^2 of the causal variants. The y-axis represents the mean χ^2 of the causal variants (10,000 common and 1,000 rare causal variants on the odd chromosomes), and the x-axis represents different levels of prevalence of the simulated binary phenotypes (prevalence $= n_{case}/(n_{case} + n_{control})$). Apart from being evaluated for the 11,000 variants altogether in panel (**a**), the mean χ^2 is evaluated for common (MAF ≥ 0.01) and rare (MAF < 0.01) causal variants separately, as shown in panels (**b**) and (**c**), respectively. Each boxplot represents the distribution of mean χ^2 across 100 simulation replicates. The line inside each box indicates the median value, notches indicate the 95% confidence interval, central box indicates the interquartile range (IQR), whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate dots.

References

1. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
2. Astle, W.J. *et al.* The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415-1429. e19 (2016).
3. Kemp, J.P. *et al.* Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nature genetics* **49**, 1468 (2017).
4. Wray, N.R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature genetics* **50**, 668-681 (2018).
5. Tin, A. *et al.* Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. *Nature genetics*, 1-16 (2019).
6. Craig, J.E. *et al.* Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression. *Nature genetics* **52**, 160-166 (2020).
7. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
8. Canela-Xandri, O., Law, A., Gray, A., Woolliams, J.A. & Tenesa, A. A new tool called DISSECT for analysing large genomic data sets using a Big Data approach. *Nature communications* **6**, 10162 (2015).
9. Loh, P.R., Kichaev, G., Gazal, S., Schoech, A.P. & Price, A.L. Mixed-model association for biobank-scale datasets. *Nat Genet* **50**, 906-908 (2018).
10. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics* **51**, 1749-1755 (2019).
11. Pirinen, M., Donnelly, P. & Spencer, C.C. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *The Annals of Applied Statistics* **7**, 369-390 (2013).
12. Van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature genetics* **48**, 1043-1048 (2016).
13. Howson, J.M. *et al.* Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms. *Nature genetics* **49**, 1113 (2017).
14. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics* **50**, 1335-1341 (2018).
15. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
16. Band, G. & Marchini, J. BGEN: a binary file format for imputed genotype and haplotype data. *preprint at bioRxiv*, 308296 (2018).
17. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. *JMIR Medical Informatics* **7**, e14325 (2019).
18. Lloyd-Jones, L.R., Robinson, M.R., Yang, J. & Visscher, P.M. Transformation of Summary Statistics from Linear Mixed Model Association on All-or-None Traits to Odds Ratio. *Genetics* **208**, 1397-1408 (2018).
19. Dey, R., Schmidt, E.M., Abecasis, G.R. & Lee, S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *Am J Hum Genet* **101**, 37-49 (2017).
20. Breyer, J.P., Avritt, T.G., McReynolds, K.M., Dupont, W.D. & Smith, J.R. Confirmation of the HOXB13 G84E germline mutation in familial prostate cancer. *Cancer Epidemiology and Prevention Biomarkers* **21**, 1348-1353 (2012).
21. Ewing, C.M. *et al.* Germline mutations in HOXB13 and prostate-cancer risk. *New England Journal of Medicine* **366**, 141-149 (2012).
22. Karlsson, R. *et al.* A population-based assessment of germline HOXB13 G84E mutation and prostate cancer risk. *European urology* **65**, 169-176 (2014).

23. Pulit, S.L., de With, S.A. & de Bakker, P.I. Resetting the bar: Statistical significance in whole-genome sequencing-based association studies of global populations. *Genetic epidemiology* **41**, 145-151 (2017).
24. Wu, Y., Zheng, Z., Visscher, P.M. & Yang, J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome Biol* **18**, 86 (2017).
25. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics* **38**, 203-208 (2006).
26. Kang, H.M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709-1723 (2008).
27. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42**, 348-354 (2010).
28. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nature genetics* **42**, 355-360 (2010).
29. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics* **44**, 821-824 (2012).
30. Svishcheva, G.R., Axenovich, T.I., Belonogova, N.M., van Duijn, C.M. & Aulchenko, Y.S. Rapid variance components-based method for whole-genome association analysis. *Nature genetics* **44**, 1166-1170 (2012).
31. Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M. & Price, A.L. Advantages and pitfalls in the application of mixed-model association methods. *Nature genetics* **46**, 100-106 (2014).
32. Loh, P.R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**, 284-90 (2015).
33. Chen, H. *et al.* Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *Am J Hum Genet* **98**, 653-66 (2016).
34. Gilmour, A.R., Thompson, R. & Cullis, B.R. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 1440-1450 (1995).
35. Breslow, N.E. & Lin, X. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82**, 81-91 (1995).
36. Kuonen, D. Miscellanea. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86**, 929-935 (1999).
37. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics* **48**, 1279 (2016).
38. UK10K consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).
39. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776-2778 (2017).
40. Lubin, J.H. & Gail, M.H. Biased selection of controls for case-control analyses of cohort studies. *Biometrics*, 63-75 (1984).

Figures

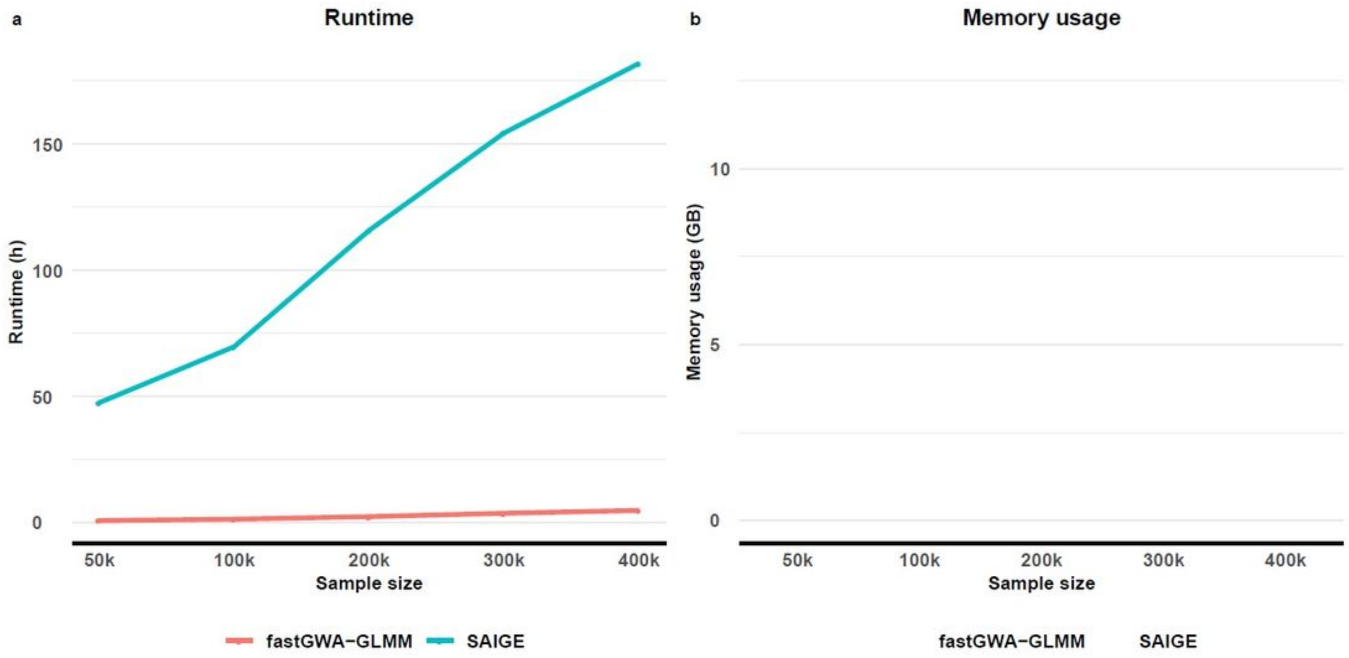

Figure 1

Comparison of runtime and memory usage between fastGWA-GLMM and SAIGE. In panel a), the x-axis represents the sample size, and the y-axis represents the runtime in hour units. For both fastGWA-GLMM and SAIGE, the runtime consists of two components: 1) the estimation of mixed model parameters (“Para. Est.”), and 2) the association test (“Assoc.”). In panel b), the x-axis represents the sample size, and the y-axis represents the memory usage in GB units. The data used in the tests consisted of 11,842,647 variants, of which 114,494 LD-pruned variants were used as “model SNPs” in SAIGE (Supplementary Note). All tests were performed in the same computing environment: 80 GB memory and 8 CPU cores (Intel Xeon Gold 6148). Each test was repeated 5 times for an average.

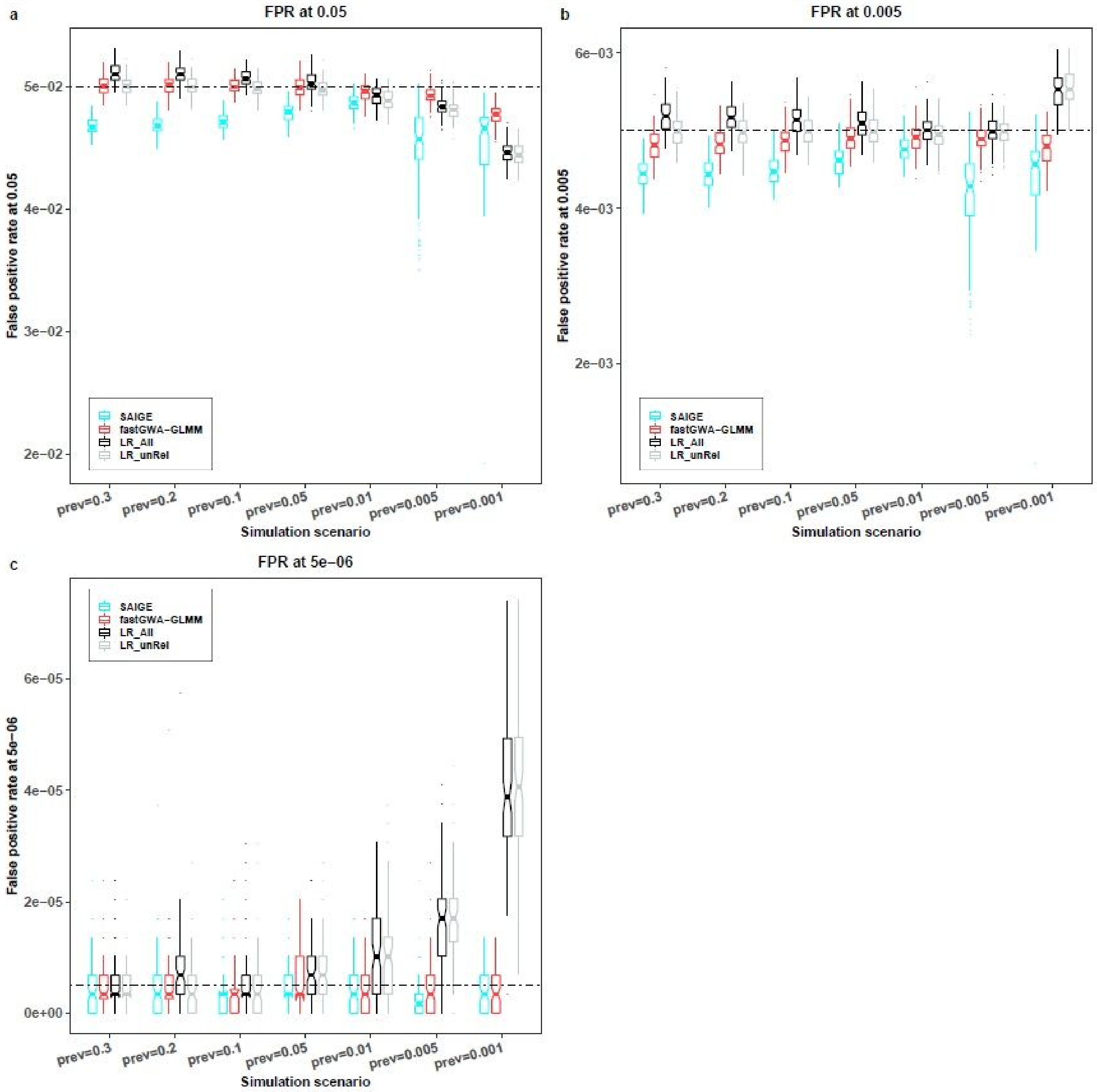

Figure 2

False positive rate (FPR) computed from the null variants. The y-axis representsthe FPR computed from the null variants (i.e., all the variants on the even chromosomes), and the x-axis represents different levels of prevalence of the simulated binary phenotypes (prevalence =⊠⊠⊠⊠⊠/(⊠⊠⊠⊠⊠+⊠⊠⊠⊠⊠⊠⊠⊠)). FPR is evaluated at five different p-value thresholds (a=0.05, 0.005, and 5 x 10-6), as shown from panels a to c. The dashed lines indicate the expected FPR (i.e., the alpha level). Each boxplot represents the distribution of FPR across 100 simulation replicates. The line inside each box indicates the median value, notches indicate the 95% confidence interval, central box indicates the interquartile range (IQR), whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate dots.

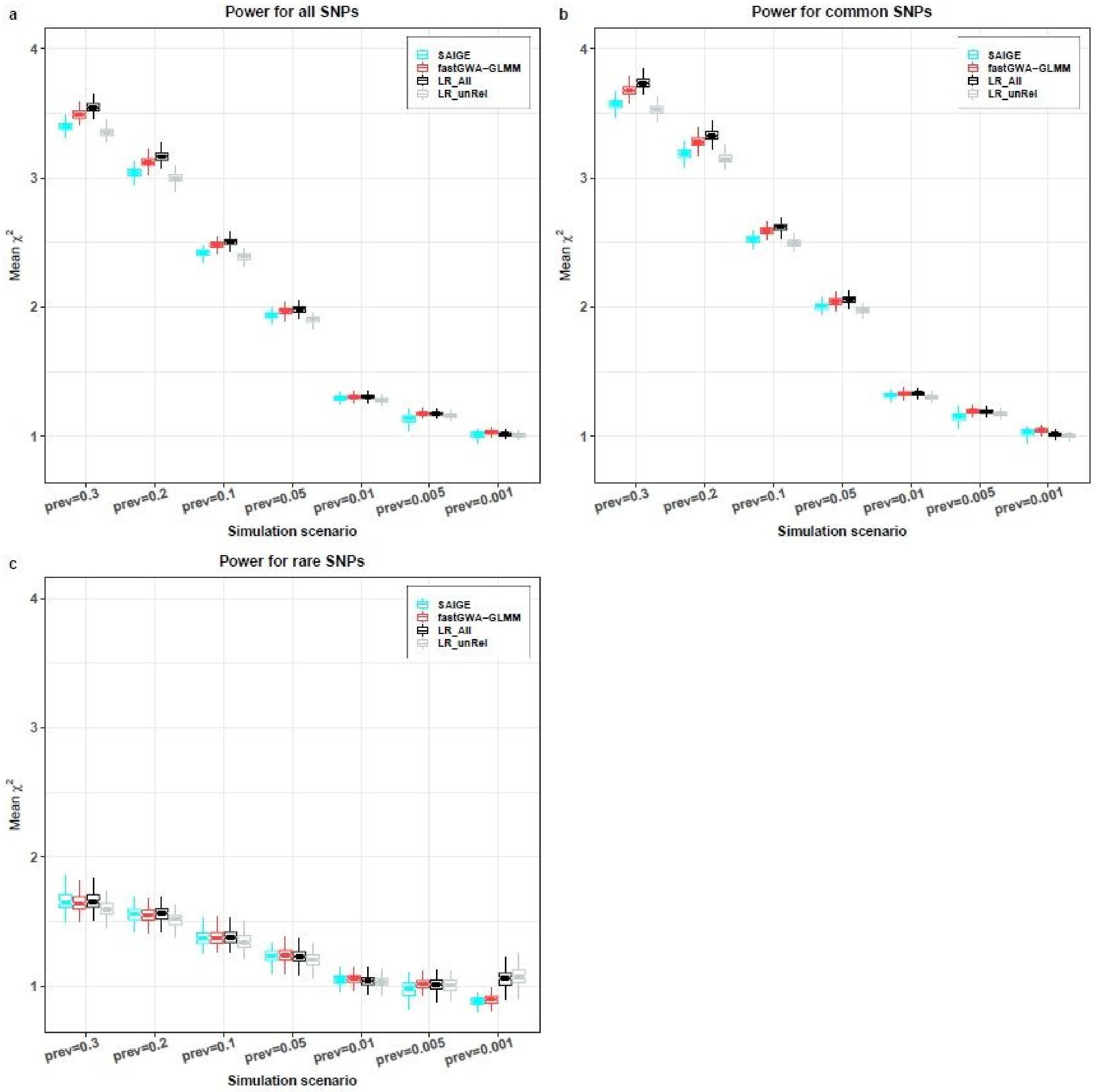

Figure 3

Comparison in power between the methods. Here, power is measured by the mean x2 of the causal variants. The y-axis represents the mean x2 of the causal variants (10,000 common and 1,000 rare causal variants on the odd chromosomes), and the x-axis represents different levels of prevalence of the simulated binary phenotypes (prevalence =□□□□□/(□□□□□+□□□□□□□□)). Apart from being evaluated for the 11,000 variants altogether in panel (a), the mean x2 is evaluated for common (MAF ≥ 0.01) and rare (MAF

< 0.01) causal variants separately, as shown in panels (b) and (c), respectively. Each boxplot represents the distribution of mean x2 across 100 simulation replicates. The line inside each box indicates the median value, notches indicate the 95% confidence interval, central box indicates the interquartile range (IQR), whiskers indicate data up to 1.5 times the IQR, and outliers are shown as separate dots.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- fastGWAGLMMSupplementry14Dec2020.pdf