A generalized linear mixed model association tool for biobank-scale data

Longda Jiang1,2,4, Zhili Zheng1,4, Hailing Fang2,3 and Jian Yang1,2,3

Compared with 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. In the present study, leveraging efficient sparse matrix-based algorithms, we developed a GLMM-based GWA tool fastGWA-GLMM, that is severalfold to orders of magnitude faster than the state-of-the-art tools when applied to the UK Biobank (UKB) data and scalable to cohorts with millions of individuals. 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 extreme case–control ratios. We applied fastGWA-GLMM to the UKB 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.

Over the past decade, we have witnessed the tremendous growth of data from GWA studies (GWASs). For example, there are nearly half a million genotyped individuals with rich phenotypes in the UKB, which have played a pivotal role in discovering new genotype–phenotype associations in recent years1–6. Nevertheless, the scale of biobank data imposes great computational challenges on methods for GWA analysis. New methods and tools have been actively developed for biobank-scale data, including linear regression-based tools such as PLINK21 and BGENIE4, and linear mixed model (LMM)-based tools such as DISSECT5, BOLT-LMM6 and fastGWA8. LMM-based methods are usually preferred over linear regression-based methods largely because they can account for relatedness without the need to remove related individuals. Despite the fact that linear regression- and LMM-based methods are developed under normality assumptions, they are often used for binary traits in practice, even though homoscedasticity is not satisfied11–13. However, recent studies14,15 have shown that, for binary traits with low case–control ratios, test statistics from LMM-based methods are inflated under the null hypothesis, leading to an inflated false positive rate (FPR), particularly for rare variants. To avoid such inflation, a common practice is to remove rare variants (for example, minor allele frequency (MAF)<0.01) and phenotypes with highly unbalanced case–control ratios (for example, <1:99)16, resulting in unnecessary loss of data.

Compared with LMM-based approaches, GLMM-based methods are better suited for binary traits17. Unfortunately, most of the GLMM-based GWA methods are not scalable to large biobank data. SAIGE14 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 extremely computationally challenging to run SAIGE for all the thousands of binary traits in the UKB, more so in cohorts with larger sample sizes than the UKB (for example, 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×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 other LMM-based tools, mainly owing to the use of a sparse genomic relationship matrix (GRM) to capture pedigree relatedness among individuals15. However, when we applied fastGWA to the UKB data, we had to remove ~3 million variants with MAF ≤0.01 and ~1,000 binary traits with case–control ratios <1:99 to avoid the inflation in FPR mentioned above12. In the present study, we aimed to develop a GWA tool that is scalable to GWAS data with millions of individuals and applicable to both common and rare variants for all binary phenotypes, including those with highly unbalanced case–control ratios. To achieve this goal, we incorporated GLMM into the fastGWA framework and developed efficient sparse matrix-based algorithms for parameter estimation and association testing. We name the method fastGWA-GLMM (or fastGWA-B, to distinguish it from the original fastGWA method15) and have implemented it in the GCTA software package19. We 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 (for example, 0.1%). We then show by analyzing subsets of the UKB data that fastGWA-GLMM is severalfold to orders of magnitude faster than the state-of-the-art tools (for example, when n=400,000, fastGWA-GLMM is 36.8-fold faster than SAIGE using 8 CPU (central processing unit) threads). We further demonstrate the scalability of fastGWA-GLMM to GWAS data beyond the size of the UKB, based on a simulated dataset with two million individuals. We have applied fastGWA-GLMM to the UKB data for 2,989 binary traits 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:

1Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia. 2School of Life Sciences, Westlake University, Hangzhou, China. 3Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, China. 4These authors contributed equally: Longda Jiang, Zhili Zheng.

✉E-mail: jian.yang@westlake.edu.cn
Logit ($\mu$) = $x_i \beta_i + X_c \beta_c + g$

where $y$ is an $n \times 1$ vector of binary phenotypes, $\mu$ is a vector of $\mu_i = P(y_i = 1|x_i, X_c, g_i)$ with $\mu_i$ being the probability of subject $i$ being a case given the subject’s genotype $x_i$, covariates $X_c$, and random genetic effect $g_i$; $x_i$ is a vector of genotype variables of a variant of interest with its effect parameter estimation step: estimating $\beta_i$; $X_c$ is the incidence matrix of fixed-effect covariates (for example, sex, age and principal components (PCs)) with their corresponding coefficients $\beta_c$; $g$ is a vector of effects that capture genetic and common environmental effects shared among related individuals, $g \sim N(0, \pi \sigma^2_g)$ with $\pi$ being the pedigree relationship matrix and $\sigma^2_g$ being the corresponding variance component. In practice, if pedigree information is unavailable or incomplete, $\pi$ can be replaced by the sparse GRM, that is, the GRM with all the small off-diagonal elements (for example, those <0.05) set to zero.$^{16}$

The fastGW A-GLMM method comprises two steps: (1) the parameter estimation step: estimating $\sigma^2_g$ and the other parameters under the null model, that is, logit ($\mu$) = $X_c \beta_c + g$; and (2) the association test step: performing score test for each variant and, if necessary, applying saddle point approximation (SPA) to correct for potential inflation driven by case–control imbalance (Methods). In the parameter estimation step, utilizing the advantage of the sparse matrix setting, we have developed a grid search-based algorithm (named fastGW A-B-REML; see Methods) that is highly computationally efficient and robust in estimating the variance components, even for traits with an extremely unbalanced case–control ratio (for example, 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:

$$T_{score} = x_i^T (y - \hat{\mu}) \quad \text{with} \quad \text{var} (T_{score}) = x_i^T \Pi x_i$$

$$\frac{T_{score}^2}{\text{var} (T_{score})} \sim \chi^2_{df=1}$$

where $P = V^{-1} - V^{-1} X_c (X_c^T V^{-1} X_c)^{-1} X_c^T V^{-1}$ with $V = W^{-1} + \pi \sigma^2_g$ and $W$ being a diagonal matrix, that is, $W_{ii} = \mu_i (1 - \mu_i)$ (Methods). $P$ is an $n \times n$ projection matrix, that is dense except $\pi$ being sparse (Methods). Therefore, to avoid the computational bottleneck due to matrix multiplication involving $P$, a GLMM version of the GRAMMAR-GAMMA approximation method is implemented in fastGW A-GLMM (Methods).

To avoid an elevated FPR in GWAS discovery due to case–control imbalance, the SPA correction is applied to variants with $\chi^2$ test statistics $\geq 4$. To further improve computational efficiency, we have developed an approximate approach to account for covariates for variants with $\chi^2$ test statistics $< 4$ (Methods). An alternative version without using this approximation is also available, which is a fewfold less efficient depending on the number of covariates fitted (Methods). We have shown that the test statistics from the approximate approach were highly consistent with those from the exact approach (Supplementary Fig. 1). In addition, to test the aggregate effect of multiple variants (for example, all the rare variants in or around a gene), we have implemented in GCTA two gene-based methods, the fastGW A-GLMM burden test (named fastGW A-BB) and the Aggregated Cauchy Association Test, based on variant-level $P$ values (ACAT-V) (Methods).

**Runtime and resource requirements.** We used the UKB data, consisting of 456,348 individuals of European ancestry and 11,842,647 variants (Methods), to evaluate the computational performance of fastGW A-GLMM in GCTA v.1.93.3 in comparison with SAIGE v.0.42.1. Note that the standard logistic regression (as implemented in PLINK2 v.2.00a2.3) was not included in the comparison because it is >100-fold slower than fastGW A-GLMM and not applicable to some of our simulation settings. After randomly sampling subgroups of individuals (ranging from 50,000 to 400,000) from the UKB, we performed a GWA in each subset of data using fastGW A-GLMM and SAIGE, respectively, on a computing platform with 80 GB of memory and 8 CPUs. The trait used for comparison is ‘irritability’ (case–control ratio = 0.39; UKB datafield: 1940). The genotype data were stored in BGEN v.1.3 format. Each test was repeated five times for an average of runtime and peak memory usage. We observed a linear relationship of either the runtime or the memory usage of fastGW A-GLMM to sample size ($n$) (Fig. 1). For a GWA with $n = 400,000$, fastGW A-GLMM required only 4.9 h, which is 36.8-fold faster than SAIGE (Fig. 1a and Supplementary Table 1). The runtime of the parameter estimation step of fastGW A-GLMM was almost negligible compared with that of SAIGE (Supplementary Table 1). Moreover, the runtime of fastGW A-GLMM was generally stable for traits with different levels of case–control ratio.
Data Fig. 1), as was SAIGE (Supplementary Table 2). Both tools showed very efficient memory usage (Supplementary Table 3), with fastGWA-GLMM showing a better scalability (Fig. 1b). During the revision of our manuscript, we have also included in the comparison a newer version of SAIGE (v.0.44.5), which was slower than SAIGE v0.42.1 (Supplementary Tables 1 and 2). Hence, we report the main results based on SAIGE v0.42.1.

To assess the scalability of fastGWA-GLMM to GWAS data beyond the size of the UKB, we generated a pseudo-GWAS dataset with 2 million individuals and 11,842,647 variants, by duplicating the UKB genotype data multiple times (Supplementary Note). We used a threshold-liability model to simulate a binary disease trait and ran fastGWA-GLMM with the full set (n = 2 million) or a subset of the data (n ranging from 200,000 to 1.8 million) (Supplementary Note). The result showed that the relationship between the runtime (or memory usage) of fastGWA-GLMM and n was still linear when n was >400,000 (Fig. 2 and Supplementary Tables 4 and 5), suggesting the scalability of fastGWA-GLMM to data with millions of individuals. For instance, when using 16 CPUs, fastGWA-GLMM was easily scalable to data with n = 1 million (runtime = 8.3 h and memory usage = 14.5 GB), and it took less than a day (runtime = 17.0 h and memory usage = 31.6 GB) to run a GWAS with n = 2 million.

While our manuscript was under revision, a machine learning method (called REGENIE) that implements whole-genome regression of multiple traits was published14. REGENIE and fastGWA-GLMM showed highly similar performances in both simulation study (Extended Data Figs. 2–4) and real data analysis (Supplementary Figs. 2 and 3). We benchmarked the computational performance of REGENIE (v.2.0.2) using the same data (n = 400,000) and on the same platform as we used to benchmark fastGWA-GLMM and SAIGE in three scenarios: a single trait, 8 traits (the set of traits included in our real data analysis) and 50 traits (the number used in the REGENIE paper14). The runtime of REGENIE in the single trait analysis was 27.5 h, 5.6-fold slower than that of fastGWA-GLMM (4.9 h) (Supplementary Table 6). When analyzing 8 traits jointly, the runtime of REGENIE was 97.2 h, corresponding to 12.2 h per trait, which was still 2.5-fold slower than fastGWA-GLMM (Supplementary Table 6). The runtime of REGENIE per trait remained similar at 12.2 h when the number of traits increased from 8 to 50 (overall runtime of 614.0 h for 50 traits) (Supplementary Table 6). Hence, in our benchmark analysis, the runtime of fastGWA-GLMM per trait was 2.5–5.6-fold faster than that of REGENIE, depending on the number of traits analyzed jointly in REGENIE. Moreover, grouping phenotypes for a joint analysis increases the time per run, which becomes more challenging when applied to larger datasets, as demonstrated in the scenario below: When we applied REGENIE to the pseudo-cohort of 2 million individuals, on average fastGWA-GLMM was at least 4.0-fold faster than REGENIE per trait, and REGENIE could not finish the job for 50 traits because it exceeded the 4 week time limit (Supplementary Table 7). It is of note that, when using a single CPU thread, the association test step of fastGWA-GLMM was 13.0-fold faster than that of REGENIE and 7.2-fold faster than that of SAIGE in the analysis with n = 400,000 (Supplementary Table 8).

FPR and statistical power. To quantify the statistical performance of fastGWA-GLMM in comparison with other tools, including SAIGE14 and PLINK2 (logistic regression using all individuals or unrelated individuals, denoted as LR-All and LR-unRel, respectively), we generated a sample of 100,000 simulated individuals with substantial population stratification and relatedness from a subset of the real UKB genotype data (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 (that is, nongenetic effects shared beyond the size of the UKB, we generated a pseudo-GWAS dataset with 2 million individuals and 11,842,647 variants, by duplicating the UKB genotype data multiple times (Supplementary Note). We used a threshold-liability model to simulate a binary disease trait and ran fastGWA-GLMM with the full set (n = 2 million) or a subset of the data (n ranging from 200,000 to 1.8 million) (Supplementary Note). The result showed that the relationship between the runtime (or memory usage) of fastGWA-GLMM and n was still linear when n was >400,000 (Fig. 2 and Supplementary Tables 4 and 5), suggesting the scalability of fastGWA-GLMM to data with millions of individuals. For instance, when using 16 CPUs, fastGWA-GLMM was easily scalable to data with n = 1 million (runtime = 8.3 h and memory usage = 14.5 GB), and it took less than a day (runtime = 17.0 h and memory usage = 31.6 GB) to run a GWAS with n = 2 million.

While our manuscript was under revision, a machine learning method (called REGENIE) that implements whole-genome regression of multiple traits was published14. REGENIE and fastGWA-GLMM showed highly similar performances in both simulation study (Extended Data Figs. 2–4) and real data analysis (Supplementary Figs. 2 and 3). We benchmarked the computational performance of REGENIE (v.2.0.2) using the same data (n = 400,000) and on the same platform as we used to benchmark fastGWA-GLMM and SAIGE in three scenarios: a single trait, 8 traits (the set of traits included in our real data analysis) and 50 traits (the number used in the REGENIE paper14). The runtime of REGENIE in the single trait analysis was 27.5 h, 5.6-fold slower than that of fastGWA-GLMM (4.9 h) (Supplementary Table 6). When analyzing 8 traits jointly, the runtime of REGENIE was 97.2 h, corresponding to 12.2 h per trait, which was still 2.5-fold slower than fastGWA-GLMM (Supplementary Table 6). The runtime of REGENIE per trait remained similar at 12.2 h when the number of traits increased from 8 to 50 (overall runtime of 614.0 h for 50 traits) (Supplementary Table 6). Hence, in our benchmark analysis, the runtime of fastGWA-GLMM per trait was 2.5–5.6-fold faster than that of REGENIE, depending on the number of traits analyzed jointly in REGENIE. Moreover, grouping phenotypes for a joint analysis increases the time per run, which becomes more challenging when applied to larger datasets, as demonstrated in the scenario below: When we applied REGENIE to the pseudo-cohort of 2 million individuals, on average fastGWA-GLMM was at least 4.0-fold faster than REGENIE per trait, and REGENIE could not finish the job for 50 traits because it exceeded the 4 week time limit (Supplementary Table 7). It is of note that, when using a single CPU thread, the association test step of fastGWA-GLMM was 13.0-fold faster than that of REGENIE and 7.2-fold faster than that of SAIGE in the analysis with n = 400,000 (Supplementary Table 8).

FPR and statistical power. To quantify the statistical performance of fastGWA-GLMM in comparison with other tools, including SAIGE14 and PLINK2 (logistic regression using all individuals or unrelated individuals, denoted as LR-All and LR-unRel, respectively), we generated a sample of 100,000 simulated individuals with substantial population stratification and relatedness from a subset of the real UKB genotype data (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 (that is, nongenetic effects shared among close relatives) and population stratification effects to the phenotype (Methods). Finally, using the simulated data, we quantified the FPR (that is, the proportion of null variants with P values less than a threshold) and statistical power (measured by the mean χ² statistic at the causal variants) for different association methods.

When the disease prevalence was >0.05, the FPRs of the null variants at 5 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 (Fig. 3 and Extended Data Fig. 5). When the prevalence was ≤0.01, both LR-unRel and LR-All showed inflated FPRs, whereas 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 (Fig. 3 and Extended Data Fig. 5), probably owing to proximal contamination because of the use a dense GRM in SAIGE (see below for more discussion). In particular, 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.

We partitioned all the null variants into two groups (common and rare variants) based on an MAF threshold of 0.01 and evaluated the FPRs 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 prevalence levels, whereas those from fastGWA-GLMM remained consistent with the expected values for both common and rare variants regardless of the prevalence level (Supplementary Figs. 4 and 5). SAIGE showed similar performance to 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 statistical power of the methods by the mean $\chi^2$ statistic at the causal variants ($\chi^2_{\text{causal}}$). The $\chi^2_{\text{causal}}$ was similar between fastGWA-GLMM and SAIGE, although the $\chi^2_{\text{causal}}$ from fastGWA-GLMM was slightly higher than that from SAIGE for common variants (Supplementary Fig. 6). The $\chi^2_{\text{causal}}$ from LR-All or LR-unRel was noninformative in this case because it suffered from inflation due to relatedness and/or case-control imbalance. We then used the area under the curve (AUC) to compare power between the methods at the same FPR level ($\alpha = 1, 0.05, 5 \times 10^{-3}, 5 \times 10^{-4}$ or $5 \times 10^{-5}$; see Methods). Although LR-unRel tended to show lower AUCs because of its smaller sample size, SAIGE, fastGWA-GLMM and LR-All showed similar AUCs in almost all the scenarios (Supplementary Figs. 7 and 8), suggesting similar levels of power between the three methods when evaluated at the same FPR level. In addition, the test statistics of fastGWA-GLMM remained well calibrated when cases were oversampled (Supplementary Figs. 9–11). When pedigree information was fully available, fastGWA-GLMM using a pedigree relationship matrix performed almost equally well to that using the sparse GRM (Supplementary Note and Extended Data Figs. 6 and 7).

We have also used the simulation to benchmark gene-based test methods including fastGWA-BB and ACAT-V in GCTA and REGENIE-Burden (note that we did not manage to run SAIGE-Burden successfully). All the three methods showed well-calibrated FPRs under the null model (Extended Data Fig. 8). To quantify the power under the alternative model, we varied the proportion of variants being causal in a gene (5%, 20% or 50%) and the directions of variant effects (random versus consistent). When the directions of the effects of the rare alleles were simulated at random, ACAT-V was always more powerful than fastGWA-BB and REGENIE-Burden, regardless of the proportion of variants being causal (Extended Data Figs. 9a–c and 10a–c). In the scenario where the effects of all the rare alleles in a gene were in a consistent direction, ACAT-V was still more powerful than fastGWA-BB and REGENIE-Burden when the proportion of variants being causal was relatively small (that is, 5%), but had lower power than fastGWA-BB and REGENIE-Burden when the proportion of variants being causal was large (that is, 20% and 50%) (Extended Data Figs. 9d–f and 10d–f). In all the simulation scenarios, we observed an extremely strong correlation ($r > 0.98$) between the test statistics of fastGWA-BB and REGENIE-Burden (Supplementary Table 9).

Applying fastGWA-GLMM to 2,989 binary traits in the UKB. We have applied fastGWA-GLMM to real data analysis of 2,989 binary phenotypes in the UKB with 456,348 individuals of European ancestry and 11,842,647 imputed variants. The phenotypes were generated using the pipeline from the Neale Lab (http://www.nealelab.is/uk-biobank) or our in-house pipeline based on the International Classification of Disease (ICD) to Phecode maps from Wu et al. (Methods). All the summary statistics are available for visualization and downloading through the fastGWA data portal (http://fastgwa.info/ukbimpbim). To benchmark fastGWA-GLMM against SAIGE and PLINK2 LR-unRel (note: $n = 348,456$ for LR-unRel), we selected, from the 2,989 traits, 8 representative phenotypes, with prevalence ranging from 0.0008 to 0.45 (Supplementary Table 10 and Supplementary Figs. 2 and 3). After clumping, the number of quasi-independent signals from fastGWA-GLMM was larger than that from either SAIGE or LR-unRel (Supplementary Table 11). The signals missed by SAIGE were probably due to proximal contamination because most of them could be identified by SAIGE-LOCO,
that is, SAIGE with the leave-one-chromosome-out (LOCO) option (Supplementary Fig. 12). As for case-control imbalance, the results from LR-unRel started to exhibit inflation for traits with prevalence <0.01 (see the third panels of Supplementary Fig. 2e–h), consistent with our simulation results, and the inflation was more prominent for the rare variants (see the third panels of Supplementary Fig. 3e–h). Meanwhile, the results from fastGWA-GLMM and SAIGE remained robust for traits with low prevalence. Among all the 2,989 traits analyzed, we identified 326 pairs of quasi-independent associations between 259 rare variants (MAF <0.01 and P ≤ 5×10^{-9}) and 75 traits (Supplementary Table 12; see Methods). Of the 259 rare variants, 37 are located in either the exonic regions or the 3′- or 5′-UTRs (Supplementary Table 12), highlighting the enrichment of rare variants in the coding and UTR regions (enrichment P=9.6×10^{-3}; Supplementary Note).

We have applied ACAT-V and fastGWA-BB to the eight representative traits. The results were in line with our observations from simulations that ACAT-V and fastGWA-BB performed differently under different patterns of genetic architecture (Supplementary Fig. 13 and Supplementary Table 13). We identified a total of 22 genes for 3 traits, with 4 genes in common across the methods, IL4R, PBX2, NXPH4 and STAC3, all associated with asthma. IL4R was a well-known gene related to asthma16,23. PBX2 was also a previously established asthma-associated gene24. No prior evidence has been found for NXPH4 and STAC3.

We have also benchmarked the performance of fastGWA-GLMM and SAIGE in disease risk prediction by a fivefold crossvalidation using two polygenic risk score methods: the conventional pruning + thresholding (P + T) method and the more recently developed SBayesR method25. Overall, across five UKB traits, there was a difference in prediction accuracy between the polygenic risk score methods, with SBayesR performing consistently better than P + T, in line with prior work17,18, but the difference between SAIGE and fastGWA-GLMM was negligible (Supplementary Table 14).

Discussion

In the present study, we have developed a highly computationally efficient association method, fastGWA-GLMM, for GWA analyses of binary phenotypes in large cohorts such as the UKB. Tested in a dataset with 400,000 individuals and 11,842,647 variants, fastGWA-GLMM was severalfold to orders of magnitude more efficient than the existing methods, depending on sample size, the number of traits analyzed and the number of CPUs used for each analysis job. We also demonstrated the scalability of fastGWA-GLMM to GWAS data with two million individuals. The implementation of the 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, because the relatedness can be well accounted for by a pedigree relatedness matrix or a sparse GRM. Taking the real data application in the UKB as an example, fastGWA-GLMM was able to include all 456,348 participants in the association analysis, whereas LR-unRel could utilize only information from 348,456 unrelated participants. As most of the large population-based cohorts rely on an assessment center-based recruitment strategy, the proportion of relatives in the cohorts tends to be high and will keep increasing in the future1. In such cases, it is crucial to avoid removing data of related individuals. Another advantage of fastGWA-GLMM over LR-unRel is its efficiency. As with many other GLMM-based methods, fastGWA-GLMM uses a score statistic for association test, which is relatively easy to compute (Methods). In contrast, LR-unRel, such as that implemented in PLINK2, uses Wald’s test based on an iteratively reweighted least squares method, which requires repeated solution of the full model for each variant, and hence is much slower than the score test, especially for analyses with covariates.

When applied to binary traits, the advantages of fastGWA-GLMM over LMM-based methods (for example, BOLT-LMM and fastGWA10) can be summarized into two aspects. The first is the better interpretability of the effect sizes, because we can directly use natural logarithm to convert the $\beta_i$ from fastGWA-GLMM into an odds ratio (Supplementary Note). However, such transformation in LMM-based methods is indirect and requires sophisticated approximations26. The second aspect is the better-controlled FPR of fastGWA-GLMM by the SPA correction. As the SPA correction is not applicable when applying LMM-based methods to binary traits27, a common strategy is to exclude traits with low case-control ratios (for example, ≤1:99) and variants with low MAFs (for example, <0.01)28, resulting in substantial loss of valuable information. An alternative strategy is to down-sample controls, which performed reasonably well for common variants (Supplementary Fig. 14). For rare variants, however, although this strategy could reduce the inflation in the LMM test statistics, the remaining inflation was sufficiently large to generate false-positive associations (Supplementary Fig. 15). In comparison, by using fastGWA-GLMM, we obtained well-calibrated summary statistics for all 3,821,959 rare variants, among which we identified hundreds of variants associated with the traits at a very stringent significance level, including known associations (Supplementary Table 12). For example, we identified a rare missense variant in HOXB13, rs138213197, strongly associated with prostate cancer, and this association has been reported in previous studies29-32.

SAIGE is a GLMM-based method that uses a dense GRM. Apart from the GRM setting, there are three other major differences between fastGWA-GLMM and SAIGE. First, fastGWA-GLMM uses a grid search-based algorithm, fastGWA-B-REML, to estimate the variance components (Methods), which is more robust and often orders of magnitude more efficient than the average information REML algorithm used in SAIGE, even for traits with extreme case-control ratios. In our simulation with prevalence = 0.005, the variance estimation step of SAIGE failed to converge for 26 of 100 simulation replicates. Second, SAIGE may suffer from proximal contamination because of the use of a dense GRM33. This is demonstrated in our real data analysis that there were several signals missed by SAIGE for three traits with relatively high case-control ratios, that is, mood swings, irritability and asthma (Supplementary Figs. 12a–c), but most of them can be restored by SAIGE-LOCO (Supplementary Fig. 12a–c and Supplementary Table 11). Third, instead of using covariate-adjusted genotype data to calculate a score test statistic for every variant, fastGWA-GLMM first uses an approximate covariate-adjustment approach and the exact approach between the covariate matrix and ~95% of the genotype vectors. We have confirmed that the difference of test statistics between the approximate covariate-adjustment approach and the exact approach is negligible (Supplementary Fig. 1). Only variants with $\chi^2$ test statistics $>4$ (note: the SPA correction is also applied when this threshold is met; see Methods). This strategy allows fastGWA-GLMM to omit the computation of matrix multiplication between the covariate matrix and ~95% of the genotype vectors. We have confirmed that the difference of test statistics between the approximate covariate-adjustment approach and the exact approach is negligible (Supplementary Fig. 1). Only variants with $\chi^2$ test statistics $<4$ might suffer from slight deflation, which does affect the power of detecting association at a genome-wide significance level. This strategy is particularly useful when the number of covariates is large (for example, >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 genotype data, which will cause a loss of computational efficiency by a fewfold, depending on the number of covariates fitted.
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. As the sparse GRM setting has already been adopted by fastGWA, 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 \( \hat{\sigma}^2 \) estimated from fastGWA-B-REML cannot be interpreted as genetic variance or heritability. This is mainly due to the use of the penalized quasi-likelihood and Laplace's method. However, from our simulations and real data applications, it did not affect the statistical performance of the association test of fastGWA-GLMM. Third, as noted in prior work and also demonstrated in the present study (for example, Extended Data Fig. 5), the SPA correction can lead to deflation in test statistics under the null. However, we believe that the SPA correction is still necessary, because otherwise the FPR will be highly inflated for rare phenotypes (Supplementary Figs. 16 and 17). Note that the FPR of SAIGE without SPA was still deflated even when the case–control ratio was not extreme, possibly because of the proximal contamination noted above. Fourth, our previous work shows that the \( \hat{\sigma}^2 \) estimated based on a sparse GRM is a better quantity to control for relatedness than that based on a dense GRM in quantitative trait analysis, which is not, however, the case for binary traits. The FPR of either fastGWA-GLMM or SAIGE was well controlled under the null in the presence of relatedness. Fifth, the inclusion of rare variants in the association analysis increases the multiple testing burden. Hence, in the present study, following the guidelines from previous studies, we used \( P \leq 5 \times 10^{-8} \) instead of \( 5 \times 10^{-8} \) 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 many binary phenotypes in biobank-scale data. The extensive simulations under different parameter settings and the real data analyses of nearly 3,000 UKB traits demonstrate its statistical robustness and computational efficiency. We believe that fastGWA-GLMM is a very useful tool for current and upcoming large-scale data, and the summary statistics released from the present study will be useful for future efforts to gain insights into the genetic basis of many health-related outcomes.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00954-4.

Received: 14 December 2020; Accepted: 22 September 2021; Published online: 4 November 2021

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 osteoporotic. Nat. Genet. 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. Nat. Genet. 50, 668–681 (2018).
5. Tin, A. et al. Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. Nat. Genet. 51, 1459–1474 (2019).
6. Craig, J. E. et al. Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression. Nat. Genet. 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. Nat. Commun. 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. Nat. Genet. 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. Ann. Appl. Stat. 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. Nat. Genet. 48, 1043–1048 (2016).
13. Howson, J. M. et al. Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms. Nat. Genet. 49, 1113 (2017).
14. Zhou, W. et al. Efficiently controlling for case–control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. 50, 1335–1341 (2018).
15. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for quantitative-trait linkage and association analysis. Am. J. Hum. Genet. 88, 76–82 (2011).
16. Liu, Y. et al. Acat: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. Am. J. Hum. Genet. 104, 410–421 (2019).
17. Band, G. & Marchini, J. BGEN: a binary file format for imputed genotype and haplotype data. Preprint at bioRxiv https://doi.org/10.1101/308296 (2018).
18. Mbatche, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. Nat. Genet. https://doi.org/10.1038/s41588-021-00870-7 (2021).
19. Zhou, W. et al. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. Nat. Genet. 52, 634–639 (2020).
20. Wu, P. et al. Mapping ICD-10 and ICD-10-CM codes to phenocodes: workflow development and initial validation. JMIR Med. Inform. 7, e14325 (2019).
21. Chatila, T. A. Interleukin-4 receptor signaling pathways in asthma pathogenesis. Trends Mol. Med. 10, 493–499 (2004).
22. Wenzel, S. E. et al. IL4Ra mutations are associated with asthma exacerbations and mast cell/IgE expression. Am. J. Respir. Crit. Care Med. 175, 576–576 (2007).
23. Hirotta, T. et al. Genome-wide association study identifies three new susceptibility loci for adult asthma in the Japanese population. Nat. Genet. 43, 893–896 (2011).
24. Lloyd-Jones, L. R. et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat. Commun. 10, 5088 (2019).
25. Ni, G. et al. A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. Biol. Psychiatry https://doi.org/10.1016/j.biopsych.2021.04.018 (2021).
26. 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 ratios. Genetics 208, 1397–1408 (2018).
27. 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).
28. Breiter, J. P., Avrit, T. G., McKenney, K. M., Dupont, W. D. & Smith, J. R. Confirmation of the HOXB13 G48E germline mutation in familial prostate cancer. Cancer Epidemiol. Prev. Biomark. 21, 1348–1353 (2012).
29. Ewing, C. M. et al. Germline mutations in HOXB13 and prostate-cancer risk. Hum. Genet. 146, 181–186 (2017).
30. Pirinen, M. et al. Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms. Nat. Genet. 49, 1113 (2017).
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. Nat. Genet. 46, 100–106 (2014).
32. Pultit, 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. Genet. Epidemiol. 41, 145–151 (2017).
33. 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).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. © The Author(s), under exclusive licence to Springer Nature America, Inc. 2021
Methods
Ethics approval. The present study was approved by the University of Queensland Human Research Ethics Committee B (2011001173) and the Ethics Committee of Westlake University (20200722YJ001).

Estimating the variance components. As described in Results, the fastGW A-GLMM model can be written as \( \logit(y) = \mu + X_{\text{c}} \beta + X_{\text{u}} \beta + g \). The logit function, \( \logit(y) = \frac{\mu}{\xi(1 - \mu)} \), is a commonly used link function in GLMM that links the expectation of the dependent binary variable \( y \) 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 \( \sigma^2_{\text{g}} \) as well as the other essential components under the null model, that is, \( \logit(y) = X_{\text{c}} \beta + \text{g} \) 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. Following GMMAT\(^{48}\) and SAIGE\(^{49}\), the log (quasi-likelihood) of the null model is:

\[
q_l(\beta, \sigma^2_g) = \log \left\{ \exp \left( \sum_{i=1}^{n} q_l(\beta, g) \right) \times (2\pi)^{-k/2} \left| \sigma^2_g \right|^{-1/2} \exp \left( -0.5 y_i^T \sigma^2_g^{-1} y_i \right) \right\}
\]

where \( q_l(\beta, g) = \frac{\mu_i(1 - \mu_i)}{\xi_i} \) is the quasi-likelihood (ql) for the \( i \)th individual given the random effect \( g \) and \( \mu_i \) is a known constant that will be omitted during the derivation. Following the derivations in Chen et al.\(^{30}\), we have \( \mu = (X'V)^{-1}X'y \) and \( \sigma^2_g = (X'V^{-1}X)^{-1} \). We use a variance–covariance matrix (that is, \( V = L'X_L + \sigma^2_g I \)) with \( W \) being a diagonal matrix (that is, \( W_i = \mu_i(1 - \mu_i) \)). We refer to the so-called ‘working vector’ with \( Y = X_{\text{c}} \beta + \text{g} + \logit(\mu_i) - Y_i \), given \( \beta \) and \( g \), the restricted maximum likelihood (REML) version of \( q_l(\beta, \sigma^2_g) \) can be written as:

\[
q_l(\beta, \sigma^2_g) = -0.5 \log |V| - 0.5 \log \left| X_i' V_i^{-1} X_i \right| - 0.5 \bar{y}_i^T V_i^{-1} \bar{y}_i
\]

where \( \bar{y}_i = X_i' \hat{\beta}_i + \bar{g}_i \). The runtime can be further reduced by an approximate fine-tuning step is repeated four times, and \( \sigma^2_{\text{g}} \text{est} = \frac{\sigma^2_{\text{g}} - \chi^2_{\text{g}}}{2} \) is returned as an estimate of \( \sigma^2_{\text{g}} \) for the 4th iteration of fastGW A-B-REML.

The \( \ell_{(i)} \) is the lower bound of the grid, determined by the variance components from the previous iteration. If we set \( \ell_{(i)} = \frac{\sigma_{\text{g}}^{(i-1)} + \sigma_{\text{g}}^{(i)} + 1}{2} \), then \( \ell_{(i)} \) is set to 0 when \( \ell_{(i)} \leq \sigma_{\text{g}}^{(i)} \) or \( \ell_{(i)} \geq \sigma_{\text{g}}^{(i)} \) when \( \ell_{(i)} > \sigma_{\text{g}}^{(i)} \). Similarly, the \( \ell_{(i)} \) is the upper bound of the grid, which is set to \( Y_i \) if \( \ell_{(i)} \leq 10\sigma_{\text{g}}^{(i)} \) or \( \ell_{(i)} < 10\sigma_{\text{g}}^{(i)} \). We apply the settings above to determine the boundaries and grid steps. We know from simulations that three iterations are sufficient to identify a reasonable interval for \( \sigma^2_g \). We do not adopt the commonly used conventional optimizers (for example, the golden-section search) because the domain of \( q_l(\beta, \sigma^2_g) \) does not always cover the whole range of \([0, Y_i]\). Therefore, it is difficult to choose an appropriate searching interval \([\ell_{(i)}, \ell_{(i)}]\) for the conventional optimizers, which would lead to a local optimum. On the other hand, we do not adopt AI-REML as implemented in SAIGE\(^{16}\) because AI-REML often converges when the variance 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 equal to 0.005.

Computing the score test statistic by the GRAMMAR-GAMMA approximation. As mentioned above, the fastGW A-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 parameter estimation step, we perform the association test for each variant using the score test: \( T_{\text{score}} = x_i' P_k x_i \) for all \( \xi \) variants, \( \xi = x_i' x_i - X_i' W X_i \), which is the covariance-adjusted genotype vector with \( x_i = x_i - X_i' W X_i x_i \). We know from prior work\(^{42}\) that \( x_i' x_i = 0 \) and \( P_k = (X_i' x_i)^{-1} \). We then have \( T_{\text{score}} = x_i' (y - \mu) + x_i' (y - \mu) + y\bar{y}_i \). The score test \( P \) value can be computed based on \( \frac{n_{\text{var}}}{n_{\text{var}}} \sim \chi^2_{n_{\text{var}}} \). \( T_{\text{score}} \) can be computed efficiently because it only involves vector multiplication and \( y - \mu \) needs to be calculated only once. However, \( T_{\text{score}} \) is difficult to obtain because \( P \) is an \( n \times n \) dense matrix, and \( x_i' P_k x_i \) 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\(^{16}\) and has been extended to cope with GLMMs in SAIGE\(^{49}\). In brief, for a random variant, its gamma ratio \( \gamma = \frac{x_i' P_k x_i}{x_i' x_i} \) is approximately constant regardless of its genotypes. The denominator is easy to compute because \( W \) is an \( n \times n \) diagonal matrix. Therefore, by randomly selecting \( m \) variants (the default \( m \) value is \( 200 \) in fastGW A-GLMM), we first estimate the mean of the gamma ratio by \( \hat{\gamma} = \frac{1}{m} \sum_{i=1}^{m} x_i' P_k x_i \) and then calculate \( T_{\text{approx}} = \frac{1}{\hat{\gamma}} x_i' P_k x_i \) for all the \( m \) variants. This strategy avoids computing \( T_{\text{score}} \) 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 covariance adjustment approach, especially when the number of covariates is large (for example, \( c > 20 \)). The full derivation of the approximate covariance 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 Fig. 1). In our software tool, users can mute the approximation method and, in that case, it is a fewfold slower than the default version, depending on the size of \( c \).

Correcting for genomic inflation by SPA. After obtaining the score test statistics, we calibrate the fastGW A-GLMM \( P \) values by SPA\(^{31-34}\) to avoid potential inflation driven by case–control imbalance. The SPA method has recently been improved to cope with GWAS data (called fastSPA\(^{31-34}\)). FastSPA was originally implemented in R\(^{31}\). To improve the computational efficiency, we implemented fastSPA by highly optimized C++ codes in fastGW A-GLMM. By default, fastGW A-GLMM applies the fastSPA correction to variants with \( \hat{\gamma} > 2 \) (Supplementary Note).

Gene-based tests. We have implemented in GCTA two gene-based test methods: the fastGW A-GLMM burden test (named fastGW A-BB) and the ACAT-V\(^{16}\). FastGW A-BB is an extension of the single-variant test in fastGW A-GLMM to assess the association of a weighted rare allele count of all the tested rare variants in or around a gene with the phenotype, conditioning on the sparse GRM (Supplementary Note). The variants weighting method follows that in SAIGE-Burden\(^{16}\). As with the single-variant association tests, the SPA correction is applied to adjust the gene-based test \( P \) values for case–control imbalance. ACAT-V is a general and flexible gene-based test method. It requires only summary statistics.
and does not rely on a specific assumption about the direction or the distribution of the effects of the rare alleles. The tail of the null distribution of the ACAT-V test statistic can be well approximated by a Cauchy distribution regardless of the linkage disequilibrium (LD) structure of the variants. Hence, LD information is not needed (Supplementary Note). We have implemented an efficient version of ACAT-V in GCTA by C/C++ code.

The UKB data. The UKB is a large cohort study consisting of approximately 500,000 participants aged between 40 and 69 years at recruitment, with extensive phenotypic records. All the UKB participants signed written informed consent, and the protocols were approved by the National Research Ethics Service Committee. In the present 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. SNP imputation was conducted by the UKB analysis team using whole-genome sequence data from the Haploview Reference Consortium and the UK10K project as the reference panels. The imputed data were filtered with standard quality control (QC) criteria in PLINK2 (ref. ), for example, MAF > 0.0001, Hardy–Weinberg equilibrium test P > 10−8, genotype 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) after QC. Note: we used 588,927 genotyped variants for the simulation study and 11,842,647 imputed variants (8,020,670 common and 3,821,977 rare) after QC. The difference of the current simulation process from that of Jiang et al.10 was the number of chromosomes simulated. Each method for common (MAF ≥0.01) and rare (MAF <0.01) variants separately.

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 ancestry backgrounds). A ‘mosaic-chromosome’ scheme modified from Loh et al.40 was used to generate the artificial individuals (Supplementary Note). Moreover, we also measured the statistical performance of different 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 labeled as European ancestry (UKB datafile 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 from the Neale Lab (https://github.com/Nealelab/UK_Biobank_GWAS) using the PHEASANT package, which were either originally dichotomous or transformed from multi-categorical traits. The other traits were generated from the ICD, 10th edn (ICD-10) records from the UKB. The original ICD-10 records provided by the UKB were unsorted text data (UKB datafile 41202), which were incompatible with the PHEASANT package. Therefore, we first extracted every unique ICD-10 code for each individual, and then grouped the ICD-10 codes into different Phenocodes based on the ICD-10 to Pheno code v.1.2 map. Any individual not labeled with a particular Phenocode was treated as a control for that Phenocode. We did not remove individuals with relevant diseases from the control group to avoid selection bias. Eventually, 835 Phenocode traits were retained for further analysis. We removed traits with n < 100 or n ≤5,000 and retained 2,989 traits in total. We fitted age, age2, sex, age×sex, age×sex and the top 20 PCs provided by the UKB as covariates in the GWA analysis (note: only age, age2 and the top 20 PCs were used for the sex-specific traits). We also applied fastGWA-GLMM and the GWAS Catalog (GCP ID: GCP000224). The individual-level genotype and phenotype data are available through formal applications to the UKB (http://www.ukbiobank.ac.uk). GWAS summary statistics for the 2,989 binary traits from our analysis of the UKB data are freely available at http://fastgwa.info/ukbimpbin and the GWAS Catalog (GCP ID: GCP000224). Source data are provided with this paper.

Code availability

fastGWA-GLMM, fastGWA-BB and ACAT-V are integrated in the GCTA software package (v.1.93.3), available at https://yanglab.westlake.edu.cn/software/gcta. The source code of GCTA v.1.93.3 is available at https://doi.org/10.5281/zenodo.5226943, and the analysis code to produce the main results presented in the paper are available at https://doi.org/10.5281/zenodo.5501110.

References

34. Yu, J. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38, 203–208 (2006).
35. Kang, H. M. et al. Efficient control of population structure in model organism association mapping. Genetics 178, 1709–1723 (2008).
36. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42, 348–354 (2010).
37. Zhang, Z. et al. Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42, 355–360 (2010).
38. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44, 821–824 (2012).
39. 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. Nat. Genet. 44, 1166–1170 (2012).
40. Loh, P. R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. 47, 284–290 (2015).
41. 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, 653–666 (2016).
42. Gilmour, A. R., Thompson, R. & Cullis, B. R. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51**, 1440–1450 (1995).

43. Breslow, N. E. & Lin, X. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82**, 81–91 (1995).

44. Kuonen, D. Miscellanea. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86**, 929–935 (1999).

45. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

46. UK10K consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).

47. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of biobank-scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).

48. Millard, L. A. C., Davies, N. M., Gaunt, T. R., Davey Smith, G. & Tilling, K. Software application profile: PHESANT: a tool for performing automated phenotype scans in UK Biobank. *Int. J. Epidemiol.* **47**, 29–35 (2017).

49. World Health Organization. *International Statistical Classification of Diseases and Related Health Problems* 10th revision (ICD-10) (World Health Organization, 2016).

50. Lubin, J. H. & Gail, M. H. Biased selection of controls for case-control analyses of cohort studies. *Biometrics* **40**, 63–75 (1984).

51. Yang, J. et al. jianyangqt/gcta: GCTA (v1.93.3beta2). *Zenodo* [https://doi.org/10.5281/zenodo.5226943](https://doi.org/10.5281/zenodo.5226943) (2021).

52. Jiang, L., Zheng, Z., Fang, H. & Yang, J. A generalized linear mixed model association tool for biobank-scale data—code. *Zenodo* [https://doi.org/10.5281/zenodo.5501110](https://doi.org/10.5281/zenodo.5501110) (2021).

Acknowledgements
We thank T. Qi for helpful discussion about the gene-based test. 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. We thank the University of Queensland Research Computing Centre and the Westlake University High-Performance Computing Center for assistance in computing. J.Y. was supported by the Australian Research Council (grant no. FT180100186), the Australian National Health and Medical Research Council (grant no. 1113400) and the Westlake Education Foundation (grant no. 101566022001). The present study makes use of data from the UKB (applications: 12505 and 66982). UKB 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. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions
J.Y. conceived and supervised the study. J.Y., L.J. and Z.Z. designed the experiment and developed the methods. Z.Z. developed the software tools with input from L.J., H.F. and J.Y. L.J. and Z.Z. performed the simulations and data analyses under the assistance and guidance of J.Y. L.J. and J.Y. wrote the manuscript with the participation of Z.Z. All the authors reviewed and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Extended data Extended data are available for this paper at [https://doi.org/10.1038/s41588-021-00954-4](https://doi.org/10.1038/s41588-021-00954-4).

Supplementary information The online version contains supplementary material available at [https://doi.org/10.1038/s41588-021-00954-4](https://doi.org/10.1038/s41588-021-00954-4).

Correspondence and requests for materials should be addressed to Jian Yang.

Peer review information *Nature Genetics* thanks Bjarni Vilhjálmsson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints).
Extended Data Fig. 1 | Runtime of fastGWA-GLMM for 8 traits with different prevalence levels. The x-axis represents the sample size, and the y-axis represents the total runtime in hour units. Different traits are labelled with different colours. The data used in this test consisted of 11,842,647 variants. All tests were performed in the same computing environment: 80 GB memory and 8 CPUs (Intel Xeon Gold 6148). Each test was repeated 5 times for an average.
Extended Data Fig. 2 | See next page for caption.
Extended Data Fig. 2 | FPR for SAIGE, fastGWA-GLMM and REGENIE quantified using the null common variants in simulations. Three methods, SAIGE, fastGWA-GLMM, and REGENIE, are compared. The y-axis represents the FPR computed from the null common variants (that is, all the common 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 alpha levels (α=0.05, 0.005, 5x10^-4, 5x10^-5, and 5x10^-6), as shown in panels from a) to e), respectively. The dashed lines indicate the expected FPRs (that is, the alpha levels). 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. In all the analyses, we used a one-sided \( \chi^2 \) statistic to test against the null hypothesis of no association.
Extended Data Fig. 3 | See next page for caption.
**Extended Data Fig. 3 | FPR for SAIGE, fastGWA-GLMM and REGENIE quantified using the rare null variants in simulations.** Three methods, SAIGE, fastGWA-GLMM, and REGENIE, are compared. The y-axis represents the FPR computed from the null rare variants (that is, all the rare variants on the even chromosomes), and the x-axis represents different levels of prevalence of the simulated binary phenotypes (prevalence = $n_{\text{case}}/(n_{\text{case}} + n_{\text{control}})$).

FPR is evaluated at five different alpha levels ($\alpha=0.05, 0.005, 5 \times 10^{-4}, 5 \times 10^{-5}$, and $5 \times 10^{-6}$), as shown in panels from a) to e), respectively. The dashed lines indicate the expected FPRs (that is, the alpha levels). 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. In all the analyses, we used a one-sided $\chi^2_{df=1}$ statistic to test against the null hypothesis of no association.
Extended Data Fig. 4 | Comparison of power (as measured by the mean $\chi^2$ value of the causal variants) between SAIGE, fastGWA-GLMM and REGENIE. The y-axis represents the mean $\chi^2$ value 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 = $n_{\text{case}}/(n_{\text{case}} + n_{\text{control}})$). Apart from being evaluated for the 11,000 variants altogether in panel (a), the mean $\chi^2$ value is also evaluated for common (MAF $\geq 0.01$) and rare (MAF $< 0.01$) causal variants separately, as shown in panels b) and c), respectively. Each boxplot represents the distribution of mean $\chi^2$ 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. In all the analyses, we used a one-sided $\chi^2$ statistic to test against the null hypothesis of no association.
Extended Data Fig. 5 | See next page for caption.
Extended Data Fig. 5 | FPR for fastGWA-GLMM and other methods quantified using all the null variants in simulations. The y-axis represents the FPR computed from the null variants (that is, all the variants on the even chromosomes), and the x-axis represents different levels of prevalence of the simulated binary phenotypes (prevalence = \( \frac{n_{\text{case}}}{n_{\text{case}} + n_{\text{control}}} \)). FPR is evaluated at five different alpha (P value threshold) levels (\( \alpha = 0.05, 0.005, 5 \times 10^{-5}, 5 \times 10^{-6}, \) and \( 5 \times 10^{-7} \)), as shown in panels from a) to e), respectively. The dashed lines indicate the expected FPRs (that is, the alpha levels). 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. In all the analyses, we used a one-sided \( \chi^2_{2m-1} \) statistic to test against the null hypothesis of no association.
Extended Data Fig. 6 | See next page for caption.
Extended Data Fig. 6 | FPR for fastGWA-GLMM and fastGWA-GLMM-Ped quantified using all the null variants in simulations. FastGWA-GLMM-Ped: fastGWA-GLMM using the pedigree relatedness matrix. fastGWA-GLMM: fastGWA-GLMM using the sparse GRM. The y-axis represents the FPR computed from the null variants (that is, 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 alpha levels (α=0.05, 0.005, 5×10^{-4}, 5×10^{-5}, and 5×10^{-6}), as shown in panels from a) to e), respectively. The dashed lines indicate the expected FPRs (that is, the alpha levels). 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. In all the analyses, we used a one-sided \( \chi^2_{d.f.} \) statistic to test against the null hypothesis of no association.
Extended Data Fig. 7 | Mean χ² value of the causal variants for fastGWA-GLMM and fastGWA-GLMM-Ped under different simulation scenarios.

FastGWA-GLMM-Ped: fastGWA-GLMM using the pedigree relatedness matrix. fastGWA-GLMM: fastGWA-GLMM using the sparse GRM. The y-axis represents the mean χ² value 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 χ² value is also 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 χ² 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. In all the analyses, we used a one-sided χ² statistic to test against the null hypothesis of no association.
Extended Data Fig. 8 | See next page for caption.
Extended Data Fig. 8 | False positive rate (FPR) for ACAT-V, fastGWA-BB, and REGENIE-Burden under different simulation scenarios. Three gene-based test methods are compared in this analysis, that is, ACAT-V (implemented in GCTA), fastGWA-BB, and REGENIE-Burden. The y-axis represents the FPR computed from the null genes (that is, all the 1,224 genes on chromosome 1 under the null simulation scenarios), and “Prev” on the x-axis represents different levels of simulated prevalence of the binary trait. The prevalence is defined as \( \frac{n_{\text{case}}}{n_{\text{case}} + n_{\text{control}}} \). FPR is evaluated at five different alpha levels (\( \alpha = 0.05, 0.01, 0.005, 0.001, \) and \( 5 \times 10^{-4} \)), as shown in panels from a) to e), respectively. The dashed lines indicate the expected FPRs (that is, the alpha levels). 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 dot. In all the analyses, we used a one-sided \( \chi^2 \) statistic to test against the null hypothesis of no association.
Extended Data Fig. 9 | Statistical power for ACAT-V, fastGWA-BB, and REGENIE-Burden under different simulation scenarios. Three gene-based test methods are compared in this analysis, that is, ACAT-V (implemented in GCTA), fastGWA-BB, and REGENIE-Burden. The y-axis represents the power, defined as the proportion of the 100 simulated causal genes on chromosome 1 with P values less than the significance threshold after Bonferroni correction (that is, 0.05/1,224=4.1x10^{-5}, where 1,224 is the number of genes used in the simulation), and “Prev” on the x-axis represents different levels of simulated prevalence of the binary trait. The prevalence is defined as ncase/(ncase + ncontrol). We varied the proportion of variants being causal in a gene (5%, 20%, or 50%) and the directions of variant effects (random or consistent), as labelled in the title of each panel. Each boxplot represents the distribution of power 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 dot. In all the analyses, we used a one-sided $\chi^2_{\text{df}=1}$ statistic to test against the null hypothesis of no association.
Extended Data Fig. 10 | Area under the curve (AUC) for ACAT-V, fastGWA-BB, and REGENIE-Burden under different simulation scenarios. Three gene-based test methods are compared in this analysis, that is, ACAT-V (implemented in GCTA), fastGWA-BB, and REGENIE-Burden. The y-axis represents the AUC (that is, the area under the receiver operating characteristic (ROC) curve), and “Prev” on the x-axis represents different levels of simulated prevalence of the binary trait. The prevalence is defined as \( \frac{n_{\text{case}}}{n_{\text{case}} + n_{\text{control}}} \). We varied the proportion of variants being causal in a gene (5%, 20% or 50%) and the directions of variant effects (random vs. consistent), as labelled in the title of each panel. Each boxplot represents the distribution of AUC 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 dot. In all the analyses, we used a one-sided \( \chi^2 \) statistic to test against the null hypothesis of no association.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ The statistical test(s) used AND whether they are one- or two-sided

☐ Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ A description of any covariates tested

☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted. Give P values as exact values whenever suitable.

☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection No software was used for data collection.

Data analysis We used fastGWA-GLMM [implemented in GCTA v1.93.3], SAIGE v0.42.1, PLINK v2.00a2.3, REGENIE v2.0.2, BOLT-LMM v2.3.2, GCTB v2.0.3, codes from the Neale Lab, R [package PHESANT v0.15], and flashPCA v2.0 to perform simulations and statistical analyses of the UK Biobank data. We have also mentioned the following software tools in the paper: fastGWA [implemented in GCTA v1.93.3], ACAT-V [implemented in ACA1 v0.91], fastSPa [implemented in SPAtest v2.0.0], DISSECT v1.15.2e, GMMAT v1.3.1, and BGENIE v1.2.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. Github). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The individual-level genotype and phenotype data are available through formal application to the UK Biobank [http://www.ukbiobank.ac.uk]. The summary-level statistics of the 2,989 binary traits from our real data analyses are available at our data portal [https://yanglab.westlake.edu.cn/resources/ukb_fastgwa/imp_binary/].
Field-specific reporting
Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☑ Life sciences  ☐ Behavioural & social sciences  ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](http://nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size  We used 456,348 individuals of European ancestry from the UKB for real data analysis and subsets of the data for simulations. Given an effective sample size of >350,000, we have >99.8% power to detect a genetic variant explaining only 0.02% of the phenotypic variance at p < 5e-8.

Data exclusions  We excluded UK Biobank participants of non-European ancestries because the sample sizes were insufficient to compare the methods’ computing performance or identify associated variants for most traits. Standard quality control criteria were applied to exclude genetic variants with minor allele frequencies < 0.0001, Hardy-Weinberg Equilibrium test P-value < 1e-6, imputation info score < 0.8, or missingness rate > 0.1 from the UK Biobank data to avoid the inclusion of low-quality variants in the association analyses.

Replication  We replicated the simulation for 100 times in each simulation scenario to investigate the performance of fastGWA-GLMM in comparison with existing methods. All replications were successfully performed. The real data analyses were not replicated because of the use of the whole UKB sample to maximize the statistical power.

Randomization  NA. We did not use any study design that required randomization.

Blinding  NA. We did not use any study design that required blinding.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems       | Methods                                      |
|----------------------------------------|----------------------------------------------|
| n/a  Involved in the study             | n/a  Involved in the study                   |
| ☑ Antibodies                           | ☑ ChiP-seq                                   |
| ☑ Eukaryotic cell lines                | ☐ Flow cytometry                             |
| ☑ Palaeontology and archaeology        | ☑ MRI-based neuroimaging                     |
| ☑ Animals and other organisms          |                                              |
| ☑ Human research participants          |                                              |
| ☑ Clinical data                        |                                              |
| ☑ Dual use research of concern         |                                              |

Human research participants

Policy information about [studies involving human research participants](https://www.ukbiobank.ac.uk/key-documents/

Population characteristics  All the UK Biobank participants used in this study are of European ancestry, aged between 40 and 69 at recruitment. The female to male ratio is approximately 1.2 (see https://www.ukbiobank.ac.uk/key-documents/ for more details of the population characteristics of the UK Biobank sample).

Recruitment  This study used data from participants recruited by the UKB. Please see https://www.ukbiobank.ac.uk/key-documents/ for more details of the recruitment. In summary, during 2006-2010, the UKB conducted its recruitment phase of more than 500,000 participants who gave their consent, answered questions, had physical measurements and gave samples (blood, urine and saliva) at a baseline assessment visit. It should be noted that the UKB is not representative of the general population on a variety of sociodemographic, physical, lifestyle and health-related characteristics, with evidence of a ‘healthy volunteer’ selection bias. As a result, the UKB is not a suitable resource for deriving generalizable disease prevalence and incidence rates. The selection bias in the UKB would not affect our simulation results because all the phenotypes were simulated based on randomly generated genetic and environmental effects. However, our real data analyses results should be interpreted with caution as those from many other genome-wide association analyses of the UKB data.

Ethics oversight  This study was approved by the University of Queensland Human Research Ethics Committee B (2011001173) and the Ethics Committee of Westlake University [20200722YH001].
Note that full information on the approval of the study protocol must also be provided in the manuscript.