nature genetics

ANALYSIS

https://doi.org/10.1038/s41588-018-0184-y

# Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies

Wei Zhou[1,2], Jonas B. Nielsen[3], Lars G. Fritsche[2,4,5], Rounak Dey[2,5], Maiken E. Gabrielsen[4], Brooke N. Wolford[1,2], Jonathon LeFaive[2,5], Peter VandeHaar[2,5], Sarah A. Gagliano[2,5], Aliya Gifford[6], Lisa A. Bastarache[6], Wei-Qi Wei[6], Joshua C. Denny[6,7], Maoxuan Lin[3], Kristian Hveem[4,8], Hyun Min Kang[2,5], Goncalo R. Abecasis[2,5], Cristen J. Willer[1,3,9,10*] and Seunggeun Lee[2,5,10*]

**In genome-wide association studies (GWAS) for thousands of phenotypes in large biobanks, most binary traits have substantially fewer cases than controls. Both of the widely used approaches, the linear mixed model and the recently proposed logistic mixed model, perform poorly; they produce large type I error rates when used to analyze unbalanced case-control phenotypes. Here we propose a scalable and accurate generalized mixed model association test that uses the saddlepoint approximation to calibrate the distribution of score test statistics. This method, SAIGE (Scalable and Accurate Implementation of GEneralized mixed model), provides accurate *P* values even when case-control ratios are extremely unbalanced. SAIGE uses state-of-art optimization strategies to reduce computational costs; hence, it is applicable to GWAS for thousands of phenotypes by large biobanks. Through the analysis of UK Biobank data of 408,961 samples from white British participants with European ancestry for > 1,400 binary phenotypes, we show that SAIGE can efficiently analyze large sample data, controlling for unbalanced case-control ratios and sample relatedness.**

Decreases in genotyping costs allow for large biobanks to genotype all participants, enabling genome-wide scale phenome-wide association studies (PheWAS) in hundreds of thousands of samples. In a typical genome-wide PheWAS, genome-wide association studies (GWAS) for tens of millions of variants are performed for thousands of phenotypes constructed from electronic health records (EHRs) and/or survey questionnaires from participants in large cohorts[1,2]. For binary traits based on disease/condition status in PheWAS, cases are typically defined as individuals with specific International Classification of Disease (ICD) diagnosis codes within the EHRs. Controls are usually all participants without the same or other related conditions[1,2]. Because of the low prevalence of many conditions/diseases, case-control ratios are often unbalanced (case:control = 1:10) or extremely unbalanced (case:control < 1:100). The scale of data and the unbalanced nature of binary traits pose substantial challenges for genome-wide PheWAS in biobanks.

Population structure and relatedness are major confounders in genetic association studies and also need to be controlled in PheWAS. Linear mixed models (LMMs) are widely used to account for these issues in GWAS for both binary and quantitative traits[3–8]. However, since LMMs are not designed to analyze binary traits, they can have inflated type I error rates, especially in the presence of unbalanced case-control ratios. Recently, Chen et al.[9] proposed to use logistic mixed models and developed a score test called the generalized mixed model association test (GMMAT). GMMAT (see URLs) assumes that score test statistics asymptotically follow a Gaussian distribution to estimate asymptotic *P* values. Although GMMAT test statistics are more robust than the LMM-based approaches, it can also suffer type I error rate inflation when case-control ratios are unbalanced because unbalanced case-control ratios invalidate asymptotic assumptions of logistic regression[10]. In addition, since GMMAT requires $O(MN^2)$ computation and $O(N^2)$ memory space, where $M$ is the number of genetic variants to be tested and $N$ is the number of individuals, it cannot handle data with hundreds of thousands of samples.

Here we propose a novel method to allow for the analysis of very large samples, for binary traits with unbalanced case-control ratios, which also infers and accounts for sample relatedness. Our method, the Scalable and Accurate Implementation of GEneralized mixed model (SAIGE) (see URLs), uses the saddlepoint approximation (SPA)[11,12] to calibrate unbalanced case-control ratios in score tests based on logistic mixed models. Since SPA uses all the cumulants, and hence all the moments, it is more accurate than using the Gaussian distribution, which uses only the first two moments. Similar to BOLT-LMM[8] (see URLs), the large sample size method for LMMs, our method uses state-of-art optimization strategies, such as the preconditioned conjugate gradient (PCG) approach[13] for

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [2]Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI, USA. [3]Department of Internal Medicine, Division of Cardiology, University of Michigan Medical School, Ann Arbor, MI, USA. [4]K. G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Norwegian University of Science and Technology, Trondheim, Norway. [5]Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA. [6]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA. [7]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [8]HUNT Research Centre, Department of Public Health and General Practice, Norwegian University of Science and Technology, Levanger, Norway. [9]Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI, USA. [10]These authors contributed equally: Cristen J. Willer and Seunggeun Lee. *e-mail: cristen@umich.edu; leeshawn@umich.edu

**Table 1 | Comparison of different methods for genome-wide association studies with mixed effect models**

| | | Method features | | | | | Algorithm complexity | | | | Benchmarks for UK Biobank data coronary artery disease (PheCode 411) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Does not require a precomputed genetic relationship matrix | Feasible for large sample sizes | Developed for binary traits | Accounts for unbalanced case-control ratio | Tests quantitative traits | Time complexity | | Memory usage (GB) | | Time CPU hours | Memory |
| | | | | | | | Step 1 | Step 2 | Step 1 | Step 2 | | |
| Logistic mixed model | SAIGE | ✓ | ✓ | ✓ | ✓ | ✓ | $O(PM_1N^{1.5})$[a] | $O(MN)$ | $M_1N/4$ | $N$ | 517 | 10.3 GB |
| | GMMAT | | | ✓ | | ✓ | $O(PN^3)$ | $O(MN^2)$ | $F\,N^2$ | $F\,N^2$ | NA | NA |
| Linear mixed model | BOLT-LMM | ✓ | ✓ | | | ✓ | $O(PM_1N^{1.5})$[a] | $O(MN)$ | $M_1N/4$ | $N$ | 360 | 10.9 GB |
| | GEMMA | | | | | ✓ | $O(N^3)$ | $O(MN^2)$ | $F\,N^2$ | $F\,N^2$ | NA | NA |

$N$, number of samples; $P$, number of iterations required to reach convergence; $M_1$, number of markers used to construct the kinship matrix; $M$, total number of markers to be tested; $F$, byte for floating number. CPU, central processing unit; NA, not applicable. [a]Number of iterations in PCG is assumed as $O(N^{0.5})$[8].

solving linear systems for large cohorts without requiring a precomputed genetic relationship matrix (GRM). The overall computation cost of this proposed method is $O(MN)$, which is substantially lower than the computation cost of GMMAT[9] and many popular LMM methods, such as GEMMA (Genome-wide Efficient Mixed Model Association)[7] (see URLs). In addition, we reduce the memory use by compactly storing raw genotypes instead of calculating and storing the GRM.

We show that SAIGE controls for the inflated type I error rates seen in binary traits with unbalanced case-control ratios in related samples through simulation and the UK Biobank data of 408,961 samples from white British participants[14,15]. By evaluating its computation performance, we demonstrate the feasibility of SAIGE for large-scale PheWAS.

## Results

**Overview of methods.** The SAIGE method contains two main steps: (1) fitting the null logistic mixed model to estimate the variance component and other model parameters; (2) testing for the association between each genetic variant and phenotypes by applying SPA to the score test statistics. Step 1 iteratively estimates the model parameters using the computationally efficient average information restricted maximum likelihood (AI-REML) algorithm[16], which is also used in GMMAT[9]. In step 1, several optimization strategies have been applied to make fitting the null logistic mixed model practical for large data sets, such as the UK Biobank[14,15]. First, spectral decomposition has been replaced by PCG to solve linear systems without inversing the $N\times N$ GRM[13] (as in BOLT-LMM[8]). The PCG method iteratively finds solutions of the linear system in a computation- and memory-efficient way. Thus, instead of requiring a precomputed GRM, which costs a significant amount of time to calculate when sample sizes are large, SAIGE uses the raw genotypes as input. The computation time is about $O(M_1N)$ times the number of iterations for the conjugate gradient to converge, where $M_1$ is the number of variants to be used for constructing GRM. Second, to further reduce memory usage during model fitting, the raw genotypes are stored in a binary vector and elements of the GRM are calculated when needed rather than being stored, so memory usage is $M_1N/4$ bytes (as in BOLT-LMM[8] and GenABEL[17]). For example, for the UK Biobank data with $M_1=93{,}511$ and $N=408{,}961$ (white British participants), memory usage drops from 669 GB for storing the GRM with float numbers to 9.56 GB for the raw genotypes in a binary vector.

After fitting the null logistic mixed model, the estimate of the random effects for each individual is obtained. The ratio of the variances of the score statistics with and without incorporating the variance components for the random effects is calculated using a subset of randomly selected genetic variants, similar to BOLT-LMM[8] and GRAMMAR-Gamma[18]. Svishcheva et al.[18] previously suggested that this ratio is constant for score tests based on LMMs. We have shown that the ratio is also approximately constant for all genetic variants with minor allele counts (MAC) $\geq 20$ in the scenario of the logistic mixed models through analytic derivation and simulations (see Supplementary Note and Supplementary Fig. 1).

In step 2, for each variant, the variance ratio is used to calibrate the score statistic variance that does not incorporate variance components for random effects. Since GRM is no longer needed for this step, the computation time to obtain the score statistic for each variant is $O(N)$. SAIGE next approximates the score test statistics using the SPA to obtain more accurate $P$ values than the normal distribution. A faster version of the SPA test, similar to the fastSPA method in the SPAtest R package that we recently developed[12], is used to further improve the computation time, which exploits the sparsity in low-frequency or rare variants to reduce the computation cost.

**Computation and memory cost.** The key features of SAIGE compared to other existing methods are presented in Table 1, showing that SAIGE is the only mixed model association method that accounts for the unbalanced case-control ratios while remaining computationally practical for large data sets. To further evaluate the computational performance of SAIGE, we randomly sampled subsets from the 408,458 white British UK Biobank participants who were defined as either coronary artery disease (CAD) cases (31,355) or controls (377,103) based on the PheWAS code 411[2,14,15], followed by benchmarking association tests using SAIGE and other existing methods on 200,000 genetic markers randomly selected out of the 71 million with imputation info $\geq 0.3$. The non-genetic covariates of sex, birth year and principal components 1–4 were adjusted in all tests. The $\log_{10}$ of memory usage and projected computation time for testing the full set of 71 million genetic variants were plotted against the sample size as shown in Supplementary Figure 2 and Supplementary Table 1. Although SAIGE and BOLT-LMM have the same order of computational complexity (Table 1), SAIGE was slower than BOLT-LMM across all sample sizes (for example, 517 versus 360 central processing unit (CPU) hours when $N=408{,}458$).

This is because fitting a logistic mixed model requires more iterative steps than an LMM, and applying SPA requires additional computation. SAIGE requires slightly less memory than BOLT-LMM (10–11 GB when $N = 408{,}458$); the low memory usage makes both methods feasible for a large data set. In contrast, GMMAT and GEMMA require substantially more computation time and memory usage. For example, when $N = 400{,}000$, the projected memory usages of both GMMAT and GEMMA are > 600 GB. The actual computation time and memory usage of association tests for the full UK Biobank data for CAD are given in Table 1. SAIGE required 517 CPU hours and 10.3 GB of memory to analyze 71 million variants that have an imputation info ≥ 0.3 for 408,458 samples, which means that the analysis will be done in ~26 hours with 20 CPU cores.

**Association analysis of binary traits in the UK Biobank data.** We applied SAIGE to several randomly selected binary traits defined by the PheWAS codes (PheCode) of the UK Biobank[2,14,15] and compared the association results with those obtained from the method based on LMMs, BOLT-LMM[8], and SAIGE without the SPA (SAIGE-NoSPA), which is asymptotically equivalent to GMMAT[9]. Because of computation and memory costs, the current GMMAT method cannot analyze the UK Biobank data. We restricted our analysis to markers directly genotyped or imputed by the Haplotype Reference Consortium (HRC)[19] panel because of quality control issues regarding non-HRC markers reported by the UK Biobank. Approximately 28 million markers with MAC ≥ 20 and an imputation info score ≥ 0.3 were used in the analysis. Among 408,961 white British participants in the UK Biobank, 132,179 have at least 1 up to the third-degree relative among the genotyped individuals[14,15]. We used 93,511 high-quality genotyped variants to construct the GRM. In the UK Biobank data, most binary phenotypes based on PheCodes (1,431 out of 1,688; 84.8%) have a case-control ratio lower than 1:100 (Supplementary Fig. 3) and would likely demonstrate a problematic inflation of association test statistics without SPA.

Association results of four exemplary binary traits that have various case-control ratios were plotted in the Manhattan plots shown in Fig. 1 and in the quantile–quantile (QQ) plots stratified by minor allele frequency (MAF) shown in Fig. 2. The four binary traits were: CAD (PheCode 411) with 31,355 cases and 377,103 controls (1:12); colorectal cancer (PheCode 153) with 4,562 cases and 382,756 controls (1:84); glaucoma (PheCode 365) with 4,462 cases and 397,761 controls (1:89); and thyroid cancer (PheCode 193) with 358 cases and 407,399 controls (1:1,138). In the Manhattan plots in Fig. 1, each locus that contains any variant with $P < 5 \times 10^{-8}$ has been highlighted in blue or green to indicate whether this locus has been reported by previous studies or not. Supplementary Table 2 presents the number of all significant loci and those that have not been previously reported by each method for each trait. Supplementary Table 3 lists all significant loci identified by SAIGE. A significant locus was defined to be potentially novel if it was located outside 500 kb of any previously reported ones.

Both Manhattan and QQ plots show that BOLT-LMM and SAIGE-NoSPA have greatly inflated type I error rates. The inflation problem is more severe as case-control ratios become more unbalanced and the MAF of the tested variants decreases. The genomic inflation factors (λ) at the 0.001 and 0.01 *P* value percentiles are shown for several MAF categories in Supplementary Table 4. For the colorectal cancer GWAS, which has a case-control ratio of 1:84, λ at the 0.001 *P* value percentile is 1.68 and 1.71 for variants with MAF < 0.01 by SAIGE-NoSPA and BOLT-LMM, while λ is 0.99 by SAIGE. The inflation is even more severe for the test results by SAIGE-NoSPA and BOLT-LMM for thyroid cancer, which has a case-control ratio of 1:1,138, with λ at the 0.001 *P* value percentile around 4–5 for variants with MAF < 0.01 and all variants, respectively. With the unbalanced case-control ratio accounted for in SAIGE, λ is again very close to 1.

We have generated summary statistics for all 1,403 PheCode-derived binary traits in 408,961 UK Biobank samples from white British participants using the SAIGE software and made them available in a public repository (see Supplementary Note for URL).

**Simulation studies.** Using simulation studies, we investigated the type I error control and power of two logistic mixed model approaches, SAIGE and GMMAT, and the LMM method BOLT-LMM that computes mixed model association statistics under the infinitesimal and non-infinitesimal models. We followed the steps described in the Data simulation subsection of the Online Methods section to simulate genotypes for 1,000 families, each with 10 family members ($N = 10{,}000$), based on the pedigree shown in Supplementary Figure 4.

**Type I error rates.** The type I error rates for SAIGE, SAIGE-NoSPA, GMMAT, and BOLT-LMM have been evaluated based on the association tests performed on $10^9$ simulated genetic variants. The variants were simulated using the same MAF spectrum of the UK Biobank HRC imputation data with case-control ratios 1:99, 1:9, and 1:1. Two different values of variance component parameter $\tau = 1$ and 2 were considered, which corresponded to liability-scale heritability 0.23 and 0.38, respectively. The empirical type I error rates at $\alpha = 5 \times 10^{-4}$ and $\alpha = 5 \times 10^{-8}$ are shown in Supplementary Table 5. SAIGE-NoSPA, GMMAT, and BOLT-LMM have greatly inflated type I error rates when the case-control ratios are moderately or extremely unbalanced, and slightly deflated type I error rates when the case-control ratios are balanced. This is expected since previous studies have suggested inflation of the score tests in the presence of the unbalanced case-control ratios and deflation in balanced studies[10,12]. Unlike GMMAT and BOLT-LMM, SAIGE has no inflation when case-control ratios are unbalanced. When the case-control ratios are balanced, SAIGE has correct type I error rates when $\tau = 1$ and slightly deflated type I error rates when $\tau = 2$. We also observed that GMMAT score test statistics do not follow the normal distribution when MAF is low and case-control is unbalanced (Supplementary Fig. 5).

To further investigate the type I error rates by MAF and case-control ratios, we carried out additional simulations. Supplementary Figure 6 shows QQ plots of 1,000,000 rare variants (MAF = 0.005) with various case-control ratios (1:1, 1:9, and 1:99); Supplementary Figure 7 shows QQ plots of 1,000,000 variants with different MAF (0.005, 0.01, and 0.3) when the case-control ratio was 1:99. Consistent with what has been observed in the real data study, GMMAT and SAIGE-NoSPA are more inflated for less frequent variants with more unbalanced case-control ratios. In contrast, SAIGE has successfully corrected this problem.

To evaluate whether SAIGE can control type I error rates in the presence of population stratification, we simulated two subpopulations with $F_{ST}$ (Wright's fixation index) 0.013, which corresponds to the average $F_{ST}$ between Finnish and non-Finnish Europeans[20]. We assumed that two subpopulations have comparable but different disease prevalences and three disease prevalence levels corresponding to unbalanced and balanced case-control ratios were simulated (0.01 for subpopulation 1 and 0.02 for subpopulation 2; 0.1 for subpopulation 1 and 0.2 for subpopulation 2; and 0.5 for subpopulation 1 and 0.4 for subpopulation 2). Both subpopulations have 1,000 families, each with 10 family members based on the pedigree shown in Supplementary Figure 4. Association tests were performed on 10 million simulated markers and the first four principal components were included as covariates (Supplementary Fig. 8). QQ plots (Supplementary Fig. 9) show that the test statistics were well calibrated regardless of the variance component parameter $\tau$ and prevalence. This simulation result demonstrates that SAIGE produces well-calibrated *P* values in the presence of population stratification.

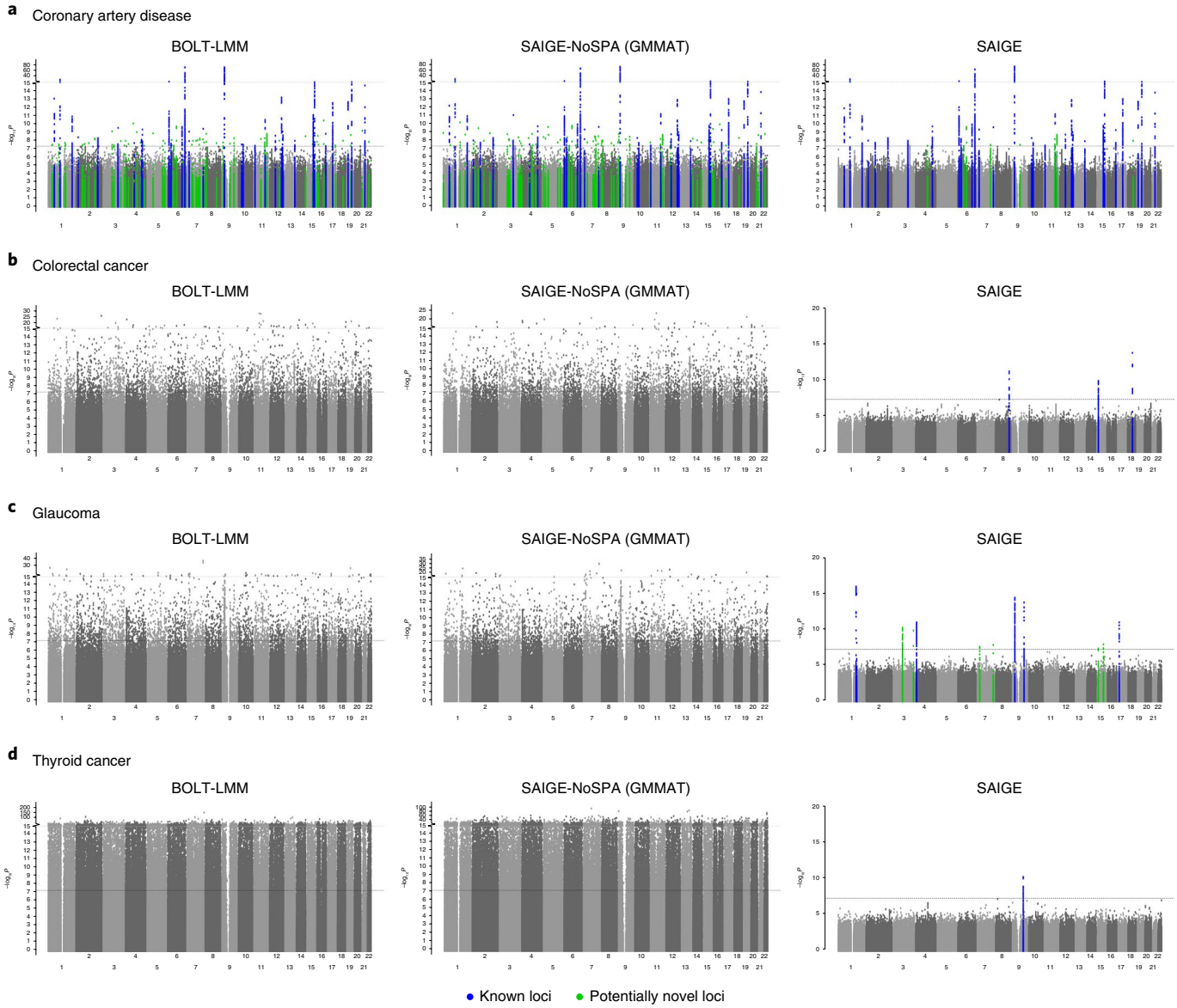


**Fig. 1 | Manhattan plots of GWAS results for four binary phenotypes with various case-control ratios in the UK Biobank. a–d**, GWAS results from SAIGE, SAIGE-NoSPA (asymptotically equivalent to GMMAT) and BOLT-LMM are shown for (**a**) coronary artery disease (PheCode 411, case:control = 1:12, $N = 408{,}458$), (**b**) colorectal cancer (PheCode 153, case:control = 1:84, $N = 387{,}318$), (**c**) glaucoma (PheCode 365, case:control = 1:89, $N = 402{,}223$), and (**d**) thyroid cancer (PheCode 193, case:control = 1:1,138, $N = 407{,}757$). *N*, sample size; blue, loci with association $P < 5 \times 10^{-8}$ that have been previously reported; green, loci that have the association $P < 5 \times 10^{-8}$ and have not been reported before. Because the results from SAIGE-noSPA and BOLT-LMM contain many false-positive signals for colorectal cancer, glaucoma, and thyroid cancer, the significant loci are not highlighted. The upper dashed line marks the breakpoint for the different scales of the *y*-axis, and the lower dashed line marks the genome-wide significance ($P = 5 \times 10^{-8}$).

**Empirical power.** Next, we evaluated empirical power. Since power simulation requires re-estimating a variance component parameter for each variant to test, to reduce computational burden, we used SAIGE-NoSPA instead of the original GMMAT software. Due to the inflated type I error rates of BOLT-LMM and GMMAT (and SAIGE-NoSPA), for a fair comparison, we estimated power at the test-specific empirical $\alpha$ levels that yield type I error rate $\alpha = 5 \times 10^{-8}$ (Supplementary Table 6). Supplementary Figure 10 shows the power curve by odds ratio for variants with MAF 0.05, 0.1, and 0.2 when $\tau = 1$. When the case-control ratio is balanced, the power of SAIGE, SAIGE-NoSPA, and BOLT-LMM was nearly identical. For studies with moderately unbalanced case-control ratios (case:control = 1:9), SAIGE has higher power than SAIGE-NoSPA and BOLT-LMM, which is due to very small empirical $\alpha$ for SAIGE-NoSPA and BOLT-LMM resulted from type I error inflation. The power gap is much larger when the case-control ratios are extremely unbalanced. Power results for $\tau = 2$ yielded the same conclusion regarding the methods comparison (data not shown).

Overall, simulation studies show that SAIGE can control type I error rates even when case-control ratios are extremely unbalanced, and it can be more powerful than GMMAT and BOLT-LMM. In contrast, GMMAT and BOLT-LMM suffer type I error inflation, and the inflation is especially severe with low MAF and unbalanced case-control ratios.

## Discussion

In this paper, we have presented a method to perform association tests for binary traits in large cohorts in the presence of sample relatedness,

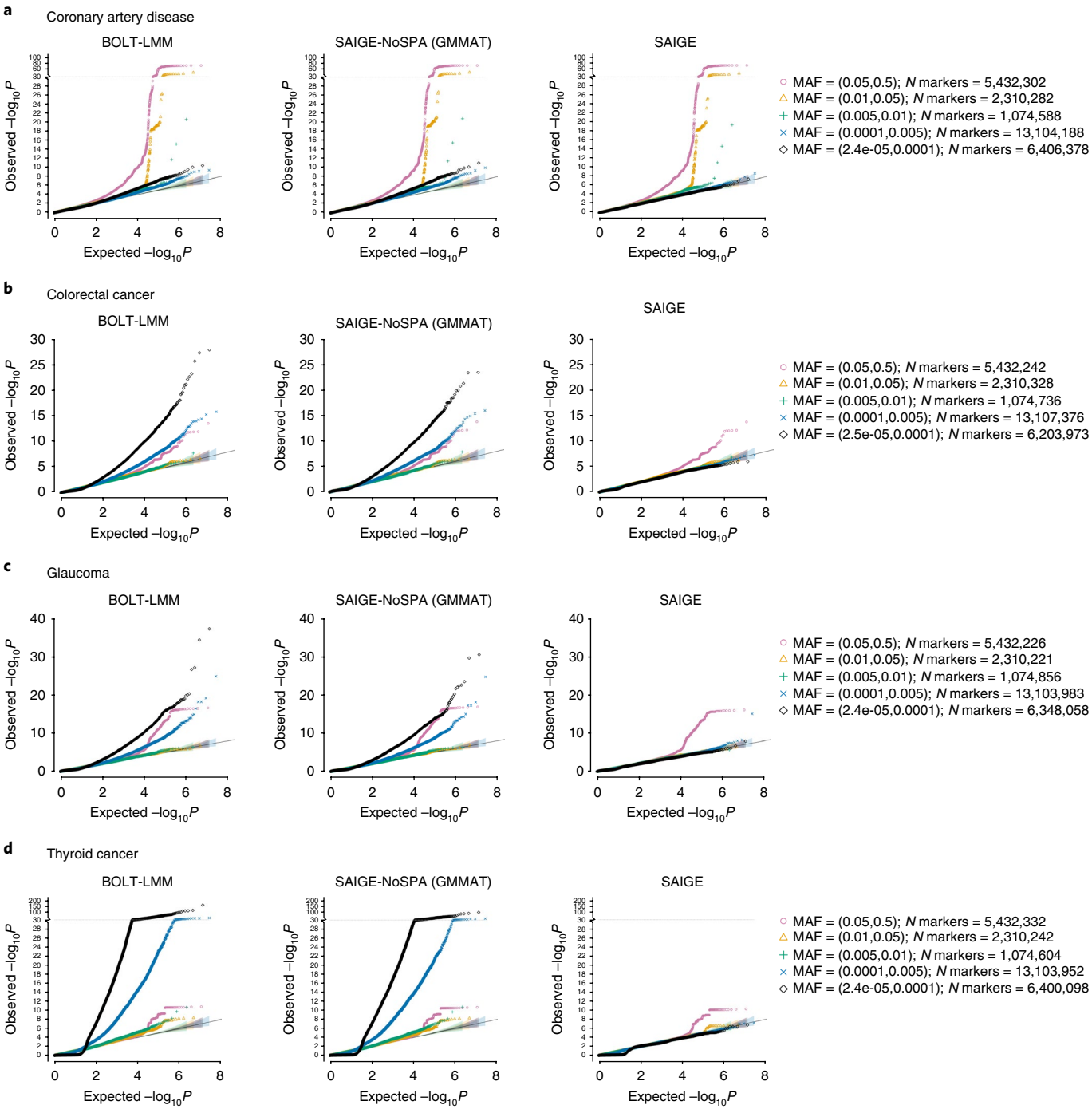


**Fig. 2 | Quantile–quantile plots of GWAS results for four binary phenotypes with various case-control ratios in the UK Biobank. a–d**, GWAS results from SAIGE, SAIGE-NoSPA (asymptotically equivalent to GMMAT), and BOLT-LMM are shown for (**a**) coronary artery disease (PheCode 411, case:control = 1:12, $N = 408{,}458$), (**b**) colorectal cancer (PheCode 153, case:control = 1:84, $N = 387{,}318$), (**c**) glaucoma (PheCode 365, case:control = 1:89, $N = 402{,}223$), and (**d**) thyroid cancer (PheCode 193, case:control = 1:1,138, $N = 407{,}757$). *N*, sample size. MAF, minor allele frequency.

which provides accurate *P* value estimates for even extremely unbalanced case-control settings (with a prevalence < 0.1%). The dramatic decrease in genotyping cost over the last decade allows increasingly larger biobanks to genotype all of their participants, followed by genome-wide PheWAS, in which GWAS analyses are performed for thousands of diseases/conditions constructed from EHRs and/or survey questionnaires to identify genetic risk factors across different phenotypes[1,2,21]. Several challenges exist for PheWAS analyses of large cohorts. Statistically, inflated type I error rates caused by unbalanced case-control ratios and sample relatedness need to be corrected. Computationally, most of existing mixed model association methods are not feasible for large sample sizes. Our method, SAIGE, uses logistic mixed modeling to account for sample relatedness and applies SPA to correct the inflation caused by the unbalanced case-control ratio in the score tests based on logistic mixed models.

SAIGE successfully corrects the inflation of type I error rates of low-frequency variants with binary traits that have unbalanced case-control ratios while also accounting for the relatedness among samples. Furthermore, our method uses several optimization strategies

that are similar to those used by BOLT-LMM to improve computational feasibility for large cohorts. For example, the PCG algorithm is used to solve linear systems instead of the spectral decomposition method so that the time complexity for fitting the null logistic model is decreased from $O(N^3)$ to approximately $O(M_1N^{1.5})$, where $M_1$ is the number of pruned markers used for estimating the GRM and $N$ is the sample size.

In the selection of genetic markers ($M_1$) for estimating the kinship matrix and the variance component, a trade-off exists between the computational cost and performance of adjusting for sample relatedness. On one hand, increasing the number of markers used for that step linearly increases computation time and memory. On the other hand, using too few markers may not be sufficient to account for all subtle sample relatedness. For example, Yang et al.[22] have shown that using a few thousand markers is not sufficient to yield correct type I error control. In the UK Biobank data analysis, we used 93,511 independent, high-quality genotyped variants, which were used by the UK Biobank data group to estimate the kinship coefficients between samples[15]. We carried out a sensitivity analysis by increasing the number of markers to 340,447 (Supplementary Note Section 2.3). Using more markers to estimate the kinship matrix for the UK Biobank data analysis produced generally similar association $P$ values but with λ closer to 1.

Using genome-wide genetic markers to adjust for sample relatedness tends to incur the proximal contamination problem, which can reduce association test power[6,8,22,23]. To avoid it, the leave-one-chromosome-out (LOCO) scheme can be used. We implemented the LOCO option in SAIGE. A sensitivity analysis (Supplementary Note Section 1.2.5) on the four exemplary binary phenotypes in the UK Biobank suggested that the proximal contamination in GWAS for diseases with relatively low prevalence, such as thyroid cancer, glaucoma, and colorectal cancer, is not as substantial as for more common diseases, for example, CAD.

Given the inflation of type I errors of LMM for rare variants (MAF < 0.5%) with unbalanced case-control phenotypes, current GWAS studies address the problem by excluding rare variants from the analysis. However, this practice can lead to false-negative results if associated rare variants are simply excluded rather than analyzed properly. For example, after using SAIGE to analyze rare variants in the UK Biobank, we identified a nonsense variant in *MYOC* (MAF = 0.14%) that was significantly associated with glaucoma. In our preliminary analysis of UK Biobank data of 1,283 non-sex-specific phenotypes, we observed 1,609 genetic variants, including variants at the same locus, with MAF < 0.5% with SAIGE $P$ values $< 5 \times 10^{-8}$ (Supplementary Note Section 2.4). The method implemented in SAIGE can control for type I error rates regardless of MAF and case-control ratios, and facilitates the identification of rare disease-associated variants.

There are several limitations in SAIGE. First, the time for algorithm convergence may vary among phenotypes and study samples given different heritability levels and sample relatedness. Second, SAIGE has been observed to be slightly conservative when case-control ratios are extremely unbalanced (Supplementary Table 5). Third, accurate estimation of the odds ratio requires fitting the model under the alternative hypothesis, which assumes non-zero effects of the tested genetic variant, and is not computationally efficient. Like several other mixed model methods[3,8,18], SAIGE estimates odds ratios for genetic markers using the parameter estimates from the null model. Fourth, SAIGE assumes that the effect sizes of genetic markers are normally distributed with a mean of zero and a standard deviation of 1, which follows an infinitesimal architecture. With this assumption, SAIGE may sacrifice power to detect genetic signals whose genetic architecture is non-infinitesimal. Finally, the variance component estimate $\tau$ from SAIGE is biased; hence, it should not be used to estimate heritability (Supplementary Note Section 2.1). This is because SAIGE uses penalized quasi-likelihood (PQL) to estimate $\tau$. However, as shown in our simulation studies and elsewhere[9], PQL-based approaches work well to adjust for sample relatedness. In the future, we plan to extend the current single variant test to gene- or region-based multiple variant tests to improve the power for identifying disease susceptibility rare variants.

With the emergence of large-scale biobanks, PheWAS will be an important tool to identify genetic components of complex traits. Here we describe a scalable and accurate method, SAIGE, for the analysis of binary phenotypes in genome-wide PheWAS. Currently, SAIGE is the only available approach to adjust for both case-control imbalance and family relatedness, which are commonly observed in PheWAS data sets. In addition, the optimization approaches used in SAIGE make it scalable for the current largest (UK Biobank) and future even larger data sets. Through simulation and real data analysis, we have demonstrated that our method can efficiently analyze a data set with over 400,000 samples and adjust for type I error rates even when the case-control ratios are extremely unbalanced. Our method provides an accurate and scalable solution for large-scale biobank data analysis and ultimately contributes toward identifying the genetic mechanisms of complex diseases.

**URLs.** SAIGE (version 0.29), https://github.com/weizhouUMICH/SAIGE/. BOLT-LMM (version 2.3.2), https://data.broadinstitute.org/alkesgroup/BOLT-LMM/. GMMAT (version 0.7), https://github.com/hanchenphd/GMMAT. GEMMA (version 0.96), https://github.com/genetics-statistics/GEMMA. UKBiobank ICD PheWeb, http://pheweb.sph.umich.edu/UKBiobank. UK-Biobank analysis results (GWAS summary statistics for > 1,400 binary phenotypes in the UK Biobank by SAIGE), https://www.leelabsg.org/resources. Unified Medical Language System, https://www.nlm.nih.gov/research/umls/.

## Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0184-y.

Received: 15 November 2017; Accepted: 21 June 2018;
Published online: 13 August 2018

## References

1. Bush, W. S., Oetjens, M. T. & Crawford, D. C. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nat. Rev. Genet.* **17**, 129–145 (2016).
2. Denny, J. C. et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).
3. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
4. Zhang, Z. et al. Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
5. 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).
6. Lippert, C. et al. FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–835 (2011).
7. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
8. Loh, P. R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
9. 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).
10. Ma, C., Blackwell, T., Boehnke, M. & Scott, L. J., GoT2D investigators. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet. Epidemiol* **37**, 539–550 (2013).
11. Kuonen, D. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86**, 4 (1999).
12. 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).

13. Kaasschieter, E. F. Preconditioned conjugate gradients for solving singular systems. *J. Comput. Appl. Math.* **24**, 265–275 (1988).
14. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
15. Bycroft, C. et al. Genome-wide genetic data on ~500,000 UK Biobank participants. Preprint at *bioRxiv*, https://doi.org/10.1101/166298 (2017).
16. 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).
17. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
18. 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).
19. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
20. Nelis, M. et al. Genetic structure of Europeans: a view from the North-East. *PLoS One* **4**, e5472 (2009).
21. Shameer, K. et al. A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Hum. Genet.* **133**, 95–109 (2014).
22. 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).
23. Listgarten, J. et al. Improved linear mixed models for genome-wide association studies. *Nat. Methods* **9**, 525–526 (2012).

## Acknowledgements

This research has been conducted using the UK Biobank Resource under application number 24460. S.L. and R.D. were supported by NIH R01 HG008773. C.J.W. was supported by NIH R35 HL135824. W.Z. was supported by the University of Michigan Rackham Predoctoral Fellowship. J.B.N. was supported by the Danish Heart Foundation and the Lundbeck Foundation. J.C.D., A.G., L.A.B., and W.-Q.W. were supported by NIH R01 LM010685 and U2C OD023196.

## Author contributions

W.Z., C.J.W., and S.L. designed the experiments. W.Z. and S.L. performed the experiments. J.B.N., L.G.F., A.G., L.A.B., W.-Q.W., and J.C.D. constructed the phenotypes for the UK Biobank data. W.Z., J.L., S.A.G., B.N.W., M.L., H.M.K., C.J.W., S.L., and G.R.A. analyzed the UK Biobank data. P.V. created the PheWeb. M.E.G. and K.H. provided the data. W.Z., J.B.N., A.G., J.C.D., R.D., C.J.W., and S.L. wrote the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41588-018-0184-y.



**Correspondence and requests for materials** should be addressed to C.J.W. or S.L.

## Methods

**Generalized linear mixed model for binary traits.** In a case-control study with sample size $N$, we denote the status of the *ith* individual with $y_i = 1$ or 0, depending on whether it is a case or a control. Let the $1 \times (1+p)$ vector $X_i$ represent $p$ covariates including the intercept and $G_i$ represent the allele counts (0, 1, or 2) for the variant to test. The logistic mixed model can be written as $logit\,(\mu_i) = X_i\alpha + G_i\beta + b_i$ where $\mu_i = P(y_i = 1|X_i, G_i, b_i)$ is the probability for the *ith* individual being a case given the covariates and genotypes, as well as the random effect, which is denoted as $b_i$. The random effect $b_i$ is assumed to be distributed as $N\ (0,\ \tau\ \psi)$, where $\psi$ is an $N \times N$ GRM and $\tau$ is the additive genetic variance. The $\alpha$ is a $(1+p) \times 1$ coefficient vector of fixed effects and $\beta$ is a coefficient of genetic effect.

**Estimating the variance component and other model parameters (step 1).** To fit the null model, $logit\,(\mu_{i0}) = X_i\alpha + b_i$, the PQL method[24] and the AI-REML algorithm[16] are used to iteratively estimate $(\hat{\tau}, \hat{\alpha}, \hat{b})$. At iteration $k$, let $(\hat{\tau}^{(k)}, \hat{\alpha}^{(k)}, \hat{b}^{(k)})$ be estimated as $(\hat{\tau}, \hat{\alpha}, \hat{b})$, let $\hat{\mu}_i^{(k)}$ be the estimated mean of $y_i$, $\hat{W}^{(k)} = diag\,[\hat{\mu}_i^{(k)}(1-\hat{\mu}_i^{(k)})]$, and let $\hat{\Sigma}^{(k)} = \{\hat{W}^{(k)}\}^{-1} + \hat{\tau}^{(k)}\psi$ be an $N \times N$ matrix of the variance of working vector $\tilde{y}_i = X_i\alpha^{(k)} + b_i^{(k)} + (y_i - \hat{\mu}_i^{(k)}) / \{\hat{\mu}_i^{(k)}(1-\hat{\mu}_i^{(k)})\}$. To obtain log quasi-likelihood and average information at each iteration, the current GMMAT approach calculates the inverse of $\hat{\Sigma}^{(k)}$. Since it is computationally too expensive for large $N$, we use the PCG[13,25], which allows calculating quasi-likelihood and average information without calculating $\{\hat{\Sigma}^{(k)}\}^{-1}$ (see Supplementary Note for details). PCG is a numerical method to find solutions of linear systems. It is particularly useful when the system is very large. BOLT-LMM[8] has successfully used it to estimate the variance component in a linear mixed model.

A score test statistics for $H_o$: $\beta = 0$ is $T = G^T(Y - \hat{\mu}) = \tilde{G}^T(Y - \hat{\mu})$ where $G$ and $Y$ are $N \times 1$ genotype and phenotype vectors, respectively, $\hat{\mu}$ is the estimated mean of Y under the null hypothesis, and $\tilde{G} = G - X(X^T\hat{W}X)^{-1}X^T\hat{W}G$ is the covariate-adjusted genotype vector. The variance of $T$ is $\mathrm{Var}(T) = \tilde{G}^T P \tilde{G}$, where $\hat{P} = \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}X(X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}$. For each variant, given $\hat{P}$, the calculation of $\mathrm{Var}(T)$ requires $O(N^2)$ computation. In addition, since our approach does not calculate $\hat{\Sigma}^{-1}$, and hence $\hat{P}$, obtaining $\mathrm{Var}(T)$ requires applying PCG for each variant, which can be computationally very expensive. To reduce computation costs, we use the same approximation approach used in BOLT-LMM and GRAMMAR-Gamma[18]; we estimate a variance of $T$ assuming that true random effect $b$ is given and then calculate the ratio between these two variances. Let us suppose that $\mathrm{Var}(T)^* = \tilde{G}^T\hat{W}\tilde{G}$, which is a variance estimate of $T$, assuming $\hat{b}$ is given. Let $r = \mathrm{Var}(T)/\mathrm{Var}(T)^*$ be the ratio of these two different types of variance estimates. In the Supplementary Note, we have shown that the ratio is approximately constant for all variants. Using this fact, we can estimate $r$ using a relatively small number of variants. In all the numerical studies in this paper, we used 30 variants to estimate $r$.

**Score testing with SPA (step 2).** Suppose that $\hat{r}$ is the estimated ratio (that is, $r$) in step 1. Now the variance-adjusted test statistics is $T_{adj} = \frac{\tilde{G}^T(Y-\hat{\mu})}{\sqrt{\hat{r}\tilde{G}^T\hat{W}\tilde{G}}}$, which has mean zero and variance 1 under the null hypothesis. The computation of $T_{adj}$ requires $O(N)$ computation. The traditional score tests assume that $T$ (and hence $T_{adj}$) asymptotically follows a Gaussian distribution under $H_o$:$\beta = 0$, which is using only the first two moments (mean and variance). When the case-control ratios are unbalanced and variants have low MAC, the underlying distribution of $T_{adj}$ can be substantially different from the Gaussian distribution. To obtain accurate $P$ values, we use SPA[11,12,26], which approximates the distribution using the entire cumulant generating function (CGF). A fast version of SPA (fastSPA)[12] has recently been developed and applied to PheWAS; it provides accurate $P$ values even when case-control ratios are extremely unbalanced (for example, case:control = 1:600).

To apply fastSPA to $T_{adj}$, we need to obtain the CGF of $T_{adj}$ first. To do this, we use the fact that given $\hat{b}$, $T_{adj}$ is a weighted sum of independent Bernoulli random variables. The approximated CGF is $K(t; \hat{\mu}, c) = \sum_{i=1}^{N} \log(1 - \hat{\mu}_i + \hat{\mu}_i e^{ct\tilde{G}_i}) - ct\sum_{i=1}^{N}\tilde{G}_i\hat{\mu}_i$, where the constant $c = \mathrm{Var}^*(T)^{-1/2}$. Let us assume that $K'(t)$ and $K''(t)$ are the first and second derivatives of $K$ with respect to $t$. To calculate the probability that $T_{adj} < q$, where $q$ is an observed test statistic, we use the following formula[11]: $pr(T_{adj} < q) \simeq F(q) = \Phi\left\{w + \frac{1}{w}\log\left(\frac{v}{w}\right)\right\}$, where $w = sign(\hat{\zeta})[2\{\hat{\zeta}q - K(\hat{\zeta})\}]^{\frac{1}{2}}$, $v = \hat{\zeta}\{K''(\hat{\zeta})\}^{\frac{1}{2}}$, and $\hat{\zeta} = \hat{\zeta}(q)$ is the solution of the equation $K'(\hat{\zeta}) = q$. Like fastSPA[12], we exploit the sparsity of genotype vector when MAFs of variants are low. In addition, since normal approximation works well when the test statistic is close to the mean, we use the normal distribution when the test statistic is within two standard deviations of the mean.

**Data simulation.** We carried out a series of simulations to evaluate SAIGE and compare its performance with that of GMMAT. We randomly simulated a set of 1,000,000 bp "pseudo" sequences, in which variants are independent to each other. Alleles for each variant were randomly drawn from Binomial ($n = 2$, $p$ = MAF). Then, we performed the gene-dropping simulation[27] using these sequences as founder haplotypes that were propagated through the pedigree of ten family members shown in Supplementary Figure 4. Binary phenotypes were generated from the following logistic mixed model $logit\,(\mu_{i0}) = \alpha_0 + b_i + X_1 + X_2 + G_i\beta$, where $G_i$ is the genotype value, $\beta$ is the genetic log odds ratio, and $b_i$ is the random effect simulated from $N(0, \tau\psi)$. Two covariates, $X_1$ and $X_2$, were simulated from Bernoulli (0.5) and $N$ (0,1), respectively. The intercept $\alpha_0$ was determined by the given prevalence (that is, the case-control ratios).

To evaluate the type I error rates at genome-wide $\alpha = 5 \times 10^{-8}$, 10 million markers along with 100 sets of phenotypes with different random seeds for case-control ratios 1:99, 1:9, and 1:1 were simulated with $\beta = 0$. Given that $\tau = 1$ and 2, the liability-scale heritability is 0.23 and 0.38, respectively[28] (Supplementary Note Section 2.1). Association tests were performed on the 10 million genetic markers for each of the 100 sets of phenotypes using SAIGE, GMMAT, and BOLT-LMM; therefore, in total $10^9$ tests were performed. To have a realistic MAF spectrum, MAFs were randomly sampled from the MAF spectrum in the UK Biobank data (Supplementary Fig. 11). Additional type I error simulations were carried out for five different MAFs (0.005, 0.01, 0.05, 0.1, and 0.3) to evaluate type I error rates by MAFs and in the presence of population stratification (Supplementary Note Section 2.2).

For the power simulation, phenotypes were generated under the alternative hypothesis $\beta \neq 0$. We simulated 1,000 data sets for MAF 0.05 and 0.2 respectively. Power was evaluated at test-specific empirical $\alpha$, which yields nominal $\alpha = 5 \times 10^{-8}$. The empirical $\alpha$ was estimated from the previous type I error simulations. Like type I error simulations, three different case-control ratios (1:99, 1:9, and 1:1) were considered.

Note that since we evaluated the empirical type I error rates and power based on genetic variants that were simulated independently, the LD Score regression[29] calibration and the LOCO scheme were not applied in BOLT-LMM or SAIGE.

**Performance evaluation of approaches used to obtain computational scalability.** We evaluated the numerical stability and convergence for numerical and asymptotic approximations that we use to achieve computational scalability: PCG (Supplementary Fig. 12); randomized trace estimation (Supplementary Fig. 13); variance ratio estimation (Supplementary Fig. 14); LOCO (Supplementary Fig. 15); and variance component parameter estimation (Supplementary Table 7 and Supplementary Fig. 16). In addition, we compared the results of UK Biobank data analysis using a GRM constructed from a larger number of markers ($M_1$ = 340,447) (Supplementary Table 8 and Supplementary Figs. 17 and 18).

**Phenotype definition in the UK Biobank.** We used a previously published scheme to defined disease-specific binary phenotypes by combining hospital ICD-9 codes into hierarchical PheCodes, each representing a more or less specific disease group[2].

ICD-10 codes were mapped to PheCodes using a combination of available maps through the Unified Medical Language System (see URLs) and other sources, string matching, and manual review. Study participants were assigned a PheCode if they had one or more of the PheCode-specific ICD codes. Cases were all study participants with the PheCode of interest and controls were all study participants without the PheCode of interest or any related PheCodes. Gender checks were performed, so PheCodes specific for one sex could not mistakenly be assigned to the other sex.

**Genome build.** All genomic coordinates are given in NCBI Build 37/UCSC hg19.

**Statistical analysis.** With the assumption of the additive genetic model, we performed GWAS using SAIGE (version 0.13) on 28 million genetic markers for 1,403 binary phenotypes of 408,961 white British participants, who passed the quality control in the UK Biobank[14]. In the logistic mixed model by SAIGE, the first four principal components, sex, and birth year were included as the non-genetic covariates. To evaluate the performance of SAIGE and account for sample relatedness and unbalanced case-control ratio for binary phenotypes, GWAS results from SAIGE were compared to those from SAIGE-NoSPA (asymptotically equivalent to GMMAT version 0.7) and BOLT-LMM (version 2.3) for four exemplary binary phenotypes with various case-control ratios. The number of samples used for analysis are included in the legend of each figure. The genomic inflation factors ($\lambda$) were calculated as the ratio of observed and expected chi-squared statistic at the 0.001 and 0.01 $P$ value percentiles.

**Reporting summary.** Further information on study design is available in the Nature Research Reporting Summary linked to this article.

**Code availability.** SAIGE is implemented as an open source R package available at https://github.com/weizhouUMICH/SAIGE/.

**Data availability.** The GWAS results for 1,403 binary phenotypes with the PheCodes[2] based on the ICD codes in the UK Biobank using SAIGE are currently available for public download at https://www.leelabsg.org/resources. Information for

all phenotypes can be found at https://github.com/weizhouUMICH/SAIGE/. The results are also displayed in the Michigan PheWeb (http://pheweb.sph.umich.edu/UKBiobank); they consist of Manhattan plots, QQ plots, and regional association plots for each phenotype, as well as the PheWAS plots for every genetic marker.

## References


24. Breslow, N. E. & Clayton, D. G. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc* **88**, 9–25 (1993).
25. Hestenes, M. R. & Stiefel, E. Methods of conjugate gradients for solving linear systems. *J. Res. Natl Bur. Stand.* **49**, 409–436 (1952).
26. Imhof, J. P. Computing the distribution of quadratic forms in normal variables. *Biometrika* **48**, 419–426 (1961).
27. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**, 97–101 (2002).
28. de Villemereuil, P., Schielzeth, H., Nakagawa, S. & Morrissey, M. General methods for evolutionary quantitative genetic inference from generalized mixed models. *Genetics* **204**, 1281–1294 (2016).
29. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

nature research

Corresponding author(s): Seunggeun Lee
Cristen Willer

# 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 Authors & Referees and the Editorial Policy Checklist.

## Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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 |
| ☐ | ☒ | An indication of 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<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☐ | ☒ | A description of all covariates tested |
| ☐ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistics 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<br>*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 |
| ☐ | ☒ | Clearly defined error bars<br>*State explicitly what error bars represent (e.g. SD, SE, CI)* |

*Our web collection on statistics for biologists may be useful.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | SAIGE is implemented as an open-source R package available at https://github.com/weizhouUMICH/SAIGE/. |
| Data analysis | BOLT-LMM v2.3 https://data.broadinstitute.org/alkesgroup/BOLT-LMM/ and GMMAT v0.7 https://github.com/hanchenphd/GMMAT were used in both simulation and real data analysis. In addition, GEMMA 0.96 https://github.com/genetics-statistics/GEMMA was used in the performance comparison. |

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/reviewers upon request. 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

SAIGE is implemented as an open-source R package available at https://github.com/weizhouUMICH/SAIGE/.
The GWAS results for > 1400 binary phenotypes with the PheCodes constructed based on ICD codes in UK Biobank using SAIGE are currently available for public download at https://www.leelabsg.org/resources
Information for all the phenotypes can be found in the website above. We also display the results in the Michigan PheWeb http://pheweb.sph.umich.edu/UKBiobank, which consists of Manhattan plots, Q-Q plots, and regional association plots for each phenotype as well as the PheWAS plots for every genetic marker.

# Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

# Life sciences study design

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

| | |
|---|---|
| Sample size | We analyzed publicly available UKBiobank data of white British samples (N=408961). We expect this large sample size can provide enough power to detect small and moderate effects of genetic associations. Since we analyzed a large number of phenotypes with wide ranges of case-control ratios, we didn't perform sample size calculation. Power analysis with a much smaller sample size (N=10,000) was performed to compare the performance of different methods, and show that the proposed approach is more powerful than the existing approaches. |
| Data exclusions | Due to QC issues in non-HRC imputed markers, we restricted our analysis to directly genotyped or HRC imputed markers. Non White British samples were excluded from the analysis. |
| Replication | We searched GWAS catalog to check whether GWAS significant loci were known (replicated) or novel. Experimental replication was not attempted. |
| Randomization | Randomization of experimental groups were not required to this study. Participants were allocated into experimental groups according to their disease status. |
| Blinding | We used coded public data, and hence were blinded. |

# Reporting for specific materials, systems and methods

### Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Unique biological materials |
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology |
| ☒ | ☐ Animals and other organisms |
| ☐ | ☒ Human research participants |

### Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Human research participants

Policy information about studies involving human research participants

| | |
|---|---|
| Population characteristics | Participant in the UK Biobank studies (http://www.ukbiobank.ac.uk/), coded publicily available data |

| | |
|---|---|
| Recruitment | UK Biobank recruited 500,000 people aged between 40-69 years in 2006-2010 from across UK. They have undergone measures, provided blood, urine and saliva samples for future analysis, detailed information about themselves. More details can be found at http://www.ukbiobank.ac.uk/ |