TECHNICAL REPORT
https://doi.org/10.1038/s41588-020-0621-6

nature genetics

Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts

Wei Zhou[1,2,3,4,13]✉, Zhangchen Zhao[1,5,13], Jonas B. Nielsen[6], Lars G. Fritsche[1,5], Jonathon LeFaive[1,5], Sarah A. Gagliano Taliun[1,5], Wenjian Bi[1,5], Maiken E. Gabrielsen[7], Mark J. Daly[2,3,4,8], Benjamin M. Neale[2,3,4], Kristian Hveem[7,9], Goncalo R. Abecasis[1,5], Cristen J. Willer[6,10,11] and Seunggeun Lee[1,5,12]✉

With very large sample sizes, biobanks provide an exciting opportunity to identify genetic components of complex traits. To analyze rare variants, region-based multiple-variant aggregate tests are commonly used to increase power for association tests. However, because of the substantial computational cost, existing region-based tests cannot analyze hundreds of thousands of samples while accounting for confounders such as population stratification and sample relatedness. Here we propose a scalable generalized mixed-model region-based association test, SAIGE-GENE, that is applicable to exome-wide and genome-wide region-based analysis for hundreds of thousands of samples and can account for unbalanced case–control ratios for binary traits. Through extensive simulation studies and analysis of the HUNT study with 69,716 Norwegian samples and the UK Biobank data with 408,910 White British samples, we show that SAIGE-GENE can efficiently analyze large-sample data ($N > 400{,}000$) with type I error rates well controlled.

In recent years, large cohort studies and biobanks, such as the Trans-Omics for Precision Medicine (TOPMed) study[1] and UK Biobank[2], have sequenced or genotyped hundreds of thousands of samples, which are invaluable resources to identify genetic components of complex traits, including rare variants (minor allele frequency (MAF) < 1%). It is well known that single-variant tests are underpowered to identify trait-associated rare variants[3]. Gene- or region-based tests, such as the burden, SKAT[4] and SKAT-O[5] tests, can be more powerful by grouping rare variants into functional units, that is, genes. To adjust for both population structure and sample relatedness, gene-based tests have been extended to mixed models[6]. For example, EMMAX[7]-based SKAT[4] approaches (EMMAX-SKAT) have been implemented and used for many rare-variant association studies, including TOPMed[1,8]. A generalized linear mixed model (GLMM) gene-based test, SMMAT, has recently been developed[6]. However, these approaches require $O(N^3)$ computation time and $O(N^2)$ memory usage, where N is the sample size, which are not scalable to large datasets.

Here we propose a new method called SAIGE-GENE for region-based association analysis that is capable of handling very large samples (>400,000 individuals) while inferring and accounting for sample relatedness. SAIGE-GENE is an extension of the previously developed single-variant association method SAIGE[9], with a modification suitable for rare variants. Like SAIGE, it uses state-of-the-art optimization strategies to reduce the computational cost for fitting null mixed models. To ensure computational efficiency while improving test accuracy for rare variants, SAIGE-GENE approximates the variance of score statistics calculated with the full genetic relationship matrix (GRM) by using the variance calculated with a sparse GRM and the ratio of these two variances estimated from a subset of genetic markers. Because the sparse GRM, which is constructed by thresholding small values in the full GRM, preserves close family structures, this approach provides more accurate variance estimation for very rare variants (minor allele count (MAC) ≤ 20) than the original approach in SAIGE[9]. By combining single-variant score statistics, SAIGE-GENE can perform burden-, SKAT- and SKAT-O-type gene-based tests. We have also developed conditional analysis to perform association tests while conditioning on a single variant or multiple variants, to identify independent rare-variant association signals. Furthermore, SAIGE-GENE can account for unbalanced case–control ratios for binary traits by adopting a robust adjustment based on saddlepoint approximation (SPA)[10–12] and efficient resampling (ER)[13]. The robust adjustment was previously developed for independent samples[14], and we have extended it for related samples in SAIGE-GENE.

[1]Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI, USA. [2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [3]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA. [4]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA. [5]Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA. [6]Division of Cardiology, Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI, USA. [7]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. [8]Institute for Molecular Medicine Finland, Helsinki Institute of Life Sciences, University of Helsinki, Helsinki, Finland. [9]HUNT Research Centre, Department of Public Health and Nursing, Norwegian University of Science and Technology (NTNU), Levanger, Norway. [10]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [11]Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI, USA. [12]Graduate School of Data Science, Seoul National University, Seoul, Republic of Korea. [13]These authors contributed equally: Wei Zhou, Zhangchen Zhao. ✉e-mail: wzhou@broadinstitute.org; leeshawn@umich.edu

We have demonstrated that SAIGE-GENE controls for type I error rates in related samples for both quantitative and binary traits through extensive simulations as well as analysis of real data, including 69,716 Norwegian samples from the Nord-Trøndelag Health Study (HUNT) study[15,16] and 408,910 White British samples from the UK Biobank[2]. By evaluating computational performance, we have shown its feasibility for large-scale genome-wide analysis. To perform exome-wide gene-based tests on 400,000 samples with on average 50 markers per gene, SAIGE-GENE requires 2,238 CPU hours and less than 36 GB of memory, whereas current methods cost more than 10 TB in memory. We have further applied SAIGE-GENE to 53 quantitative traits and 10 binary traits in the UK Biobank and identified several significantly associated genes.

Results

Overview of methods. SAIGE-GENE consists of two main steps: (1) fitting the null GLMM to estimate variance components and other model parameters and (2) testing for association between each genetic variant set, such as a gene or region, and the phenotype. Three different association tests—burden, SKAT and SKAT-O tests—have been implemented in SAIGE-GENE. The workflow is shown in Extended Data Fig. 1.

SAIGE-GENE uses optimization strategies similar to those used in the original SAIGE to fit the null GLMM in step 1. In particular, the spectral decomposition has been replaced by a preconditioning conjugate gradient (PCG) to solve linear systems without calculating and inverting the $N \times N$ GRM. To reduce memory usage, raw genotypes are stored in a binary vector and GRM elements are calculated when needed rather than being stored.

One of the most time-consuming aspects of association tests is calculating the variance of single-variant score statistics, which requires $O(N^2)$ computation. To reduce computational cost, existing approaches, such as SAIGE[9], BOLT-LMM[17] and GRAMMAR-Gamma[18], approximate the variance of single-variant score statistics with the full GRM by using the variance estimate without a GRM and the ratio of these two variances. The ratio, which is assumed to be constant, is estimated by using a subset of randomly selected genetic markers. However, for very rare variants with MAC $\leq$ 20, the assumption of a constant ratio is not satisfied (Extended Data Fig. 2, left). This is because rare variants are more susceptible to close family structures. Thus, to better approximate the variance, SAIGE-GENE incorporates close family structures through a sparse GRM in which GRM elements below a user-specified relatedness coefficient are zeroed out and close family structures are preserved. The ratio between the variance with the full GRM and the variance with the sparse GRM is much less variable (Extended Data Fig. 2, right). To construct a sparse GRM, a small subset of randomly selected genetic markers, that is, 2,000, are first used to quickly estimate which sample pairs pass the user-specified coefficient-of-relatedness cutoff, for example, $\geq$0.125 for up to third-degree relatedness. Then, the coefficients of relatedness for the related pairs are further estimated by using the full set of genetic markers, which equal values in the full GRM. Given that estimated values for variance ratios vary by MAC for extremely rare variants (Extended Data Fig. 2, left), such as singletons and doubletons, the variance ratios need to be estimated separately for different MAC categories. By default, MAC categories are set to be MAC = 1, 2, 3, 4, 5, 6 to 10, 11 to 20, and >20.

In step 2, gene-based tests are conducted with single-variant score statistics and their covariance estimates, which are approximated as the product of the covariance with the sparse GRM and the pre-estimated ratio. SAIGE-GENE can carry out burden, SKAT and SKAT-O approaches. Because SKAT-O is a combination burden and SKAT test and hence provides robust power, SAIGE-GENE performs SKAT-O tests by default.

If a gene or region is significantly associated with the phenotype of interest, it is necessary to test whether the signal is from rare variants or just a shadow of the signal for common variants in the same locus. We have developed conditional analysis using linkage disequilibrium (LD) information between conditioning markers and the tested gene[19]. Details are provided in the Methods.

SAIGE-GENE uses the same GLMM as in SMMAT, while SMMAT calculates the variances of the score statistics for all tested genes by using the full GRM directly and hence can be thought of as the 'exact' method. When the trait is quantitative, the GLMM used by SAIGE-GENE and SMMAT is equivalent to the linear mixed model (LMM) of EMMAX-SKAT. We have further shown that SAIGE-GENE provides association P values consistent with those from the two 'exact' methods, EMMAX-SKAT and SMMAT (r^2 of $-\log_{10}(P \text{ values}) > 0.99$), when using both simulation studies (Extended Data Fig. 3) and analysis of real data downsampled from the UK Biobank and HUNT (Extended Data Fig. 4), but with much smaller computational and memory costs (Fig. 1). We have also shown that SAIGE-GENE with different coefficient-of-relatedness cutoffs (0.125 and 0.2) produces nearly identical association P values for automated readings of pulse rate in the UK Biobank (Extended Data Fig. 5).

For binary phenotypes with unbalanced case–control ratios, single-variant score statistics do not follow a normal distribution, leading to inflated type I error rates for region-based tests[13]. To address this problem, we have recently developed an adjustment for independent samples[14]. The approach uses SPA[10–12] and ER[13] to calibrate the variance of single-variant score statistics. We have extended this approach to the GLMM for SAIGE-GENE, which provides greatly improved type I error control relative to the unadjusted approach of assuming normality (Extended Data Fig. 6). Details can be found in the Supplementary Note.

Computational and memory cost. To evaluate the computational performance of SAIGE-GENE, we randomly sampled subsets of the 408,144 UK Biobank participants with White British ancestry and non-missing measurements for waist-to-hip ratio[2]. We benchmarked SAIGE-GENE, EMMAX-SKAT and SMMAT for exome-wide gene-based SKAT-O tests, in which 15,342 genes were tested with the assumption that each had 50 rare variants.

Memory usage is plotted in Fig. 1a. The memory cost of SAIGE-GENE is linear to the number of markers, M_1, used for kinship estimation, but using too few markers may not be sufficient to account for subtle sample relatedness, leading to inflated type I error rates[9,20]. SAIGE-GENE uses 11.74 GB with $M_1 = 93{,}511$ and 35.59 GB with $M_1 = 340{,}447$ when the sample size N is 400,000, making it feasible for large-sample data. In contrast, with $N = 400{,}000$, the memory usage in EMMAX-SKAT and SMMAT is projected to be nearly 10 TB.

Total computation time for exome-wide gene-based tests is plotted in Fig. 1b. Computation times for step 1 and step 2 are plotted separately in Extended Data Fig. 7, with numerical data presented in Supplementary Table 1. The computation time for step 1 in SAIGE-GENE is approximately $O(M_1N^{1.5})$ and in SMMAT and EMMAX-SKAT is $O(N^3)$. In step 2, the association test for each gene costs $O(qK)$ in SAIGE-GENE, where q is the number of markers in the gene and K is the number of nonzero elements in the sparse GRM. In comparison to $O(qN^2)$ for step 2 in SMMAT and EMMAX-SKAT, SAIGE-GENE decreases the computation time dramatically. For example, in the UK Biobank ($N = 408{,}910$) with a relatedness coefficient of $\geq$0.125 (corresponding to preserving third-degree or closer relatives in the GRM), $K = 493{,}536$, which is the same order of magnitude as N, and hence $O(qK)$ is much smaller than $O(qN^2)$. As the computation time in step 2 is approximately linear to q, the number of markers in each variant set, the total computation time for exome-wide gene-based tests

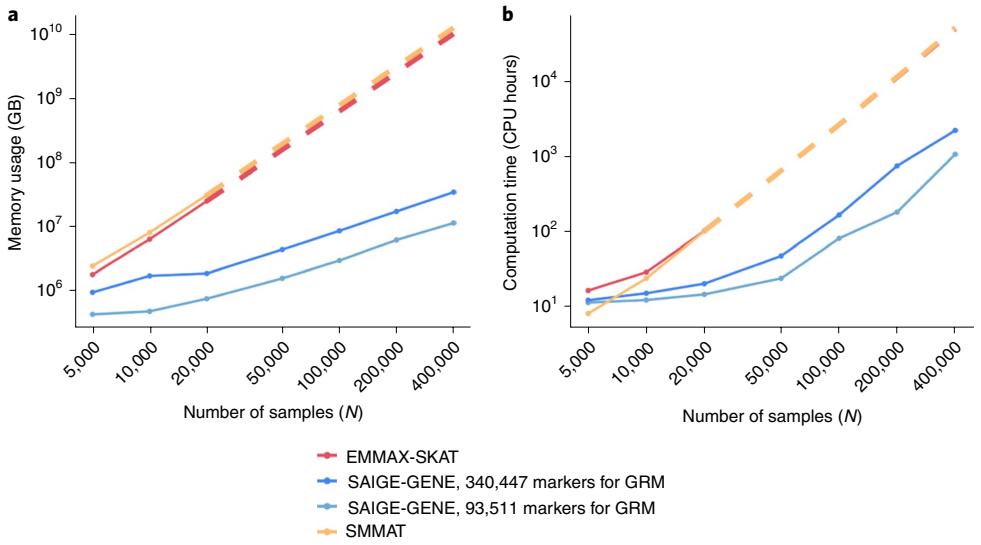


Fig. 1 | Estimated and projected computational cost by sample size (*N*) for gene-based tests of 15,342 genes, each containing 50 rare variants. Benchmarking was performed on randomly subsampled UK Biobank data with 408,144 White British participants for waist-to-hip ratio. The reported run time and memory usage are the medians of five runs with samples randomly selected from the full sample set by using different sampling seeds. The reported computation time and memory usage for EMMAX-SKAT and SMMAT were projected when $N > 20{,}000$. **a**, log–log plots of the memory usage as a function of sample size (*N*). **b**, log–log plots of the run time as a function of sample size (*N*). Numerical data are provided in Supplementary Table 1.

was projected by different q values and is plotted in Extended Data Fig. 8. In addition, we plotted the projected computation time for genome-wide region-based tests in Extended Data Fig. 9, in which 286,000 chunks of genomic DNA with 50 markers per chunk were assumed to be tested, corresponding to 14.3 million markers in Haplotype Reference Consortium (HRC)-imputed UK Biobank data with MAF ≤ 1% and imputation info score ≥ 0.8.

With $M_1 = 340{,}447$ and $N = 400{,}000$, it takes SAIGE-GENE 2,238 CPU hours for the exome-wide analysis and 3,919 CPU hours for the genome-wide analysis for waist-to-hip ratio. In comparison to EMMAX-SKAT and SMMAT, SAIGE-GENE is 25 times faster for the exome-wide analysis and 161 times faster for the genome-wide analysis. More details are presented in Supplementary Table 1. Additional steps in the robust adjustment for binary traits only slightly increases the computational cost (1,269 versus 1,232 CPU hours for exome-wide analysis with $M_1 = 93{,}511$) as compared to the unadjusted approach (Supplementary Table 2 and Extended Data Fig. 10). Details are provided in the Supplementary Note.

The computation time for constructing the sparse GRM is $O(M_1^* N^2 + M_1 K)$, where M_1^* is the number of a small set of markers used for initial determination of related sample pairs, which by default is set to be 2,000. Construction of the sparse GRM is needed once for each dataset, and the sparse GRM is then reused for all phenotypes. For example, for the UK Biobank with $N = 408{,}910$, $M_1 = 340{,}447$, $M_1^* = 2{,}000$, $K = 493{,}536$ and a relatedness coefficient of ≥0.125, corresponding to up to third-degree relatedness, it takes 312 CPU hours to create the sparse GRM.

Gene-based association analysis of quantitative traits in HUNT and UK Biobank. We applied SAIGE-GENE to analyze 13,416 genes, with at least two rare (MAF ≤ 1%) missense and/or stop-gain variants that were directly genotyped or imputed from HRC, for association with high-density lipoprotein (HDL) in 69,716 Norwegian samples from the HUNT study[9], which has substantial sample relatedness. The quantile–quantile plot for the *P* values of SKAT-O tests from SAIGE-GENE for HDL in the HUNT study is presented in Fig. 2a. As shown in Table 1, eight genes reached the exome-wide significance threshold ($P \leq 2.5 \times 10^{-6}$), all of which are located in the previously reported genome-wide association study (GWAS) loci for HDL[21,22]. After conditioning on the most significant nearby variants from single-variant association tests (500 kb upstream and downstream), all genes, except *FSD1L*, remained significant, suggesting that SAIGE-GENE identified associations of rare coding variants that were independent from nearby association signals, pointing to candidate causal genes at these loci.

We also applied SAIGE-GENE to analyze 15,342 genes for 53 quantitative traits in 408,910 UK Biobank participants with White British ancestry[2]. Heritability estimates based on the full GRM are presented in Supplementary Table 3a. Supplementary Table 4a presents all genes with *P* values reaching the exome-wide significance threshold ($P \leq 2.5 \times 10^{-6}$). The same MAF cutoff of ≤1% for missense and stop-gain variants was applied. Figure 2b shows the quantile–quantile plot for automated readings of pulse rate as an exemplary phenotype. *MYH6*, *ARHGEF40* and *DBH* remained significant after conditioning on the most significant nearby variants (Table 1). *TBX5*, *MYH6*, *TTN* and *ARHGEF40* are known genes for heart rate from previous GWAS[23–26]. To our knowledge, *KIF1C* and *DBH* have not been reported in association studies for heart rate, but $Dbh^{-/-}$ mice have decreased heart rates as compared to littermate control $Dbh^{+/-}$ mice[27]. For *DBH*, no single variant reached genome-wide significance (the most significant variant was 9:136149399 (GRCh37) with MAF = 18.7% and $P = 3.46 \times 10^{-6}$). Fifteen genes that were exome-wide significant had no genome-wide-significant single variants (Supplementary Table 5). After conditioning on the most significant nearby variants, 64 genes for 12 traits remained exome-wide significant (Supplementary Table 6a). SAIGE-GENE identified several potentially novel gene–phenotype associations, such as *DBH* for automated readings of pulse rate ($P_{\text{SKAT-O}} = 1.74 \times 10^{-6}$), and also replicated several previous findings, such as the association between *ADAMTS3* and height[28]. Details are provided in the Supplementary Note. These results demonstrate the value of gene-based tests for identifying genetic factors for complex traits.

Gene-based association analysis of binary traits in UK Biobank. We applied SAIGE-GENE to ten binary phenotypes with various case–control ratios in the UK Biobank. The heritability estimates on a liability scale are presented in Supplementary Table 3b. Nine genes

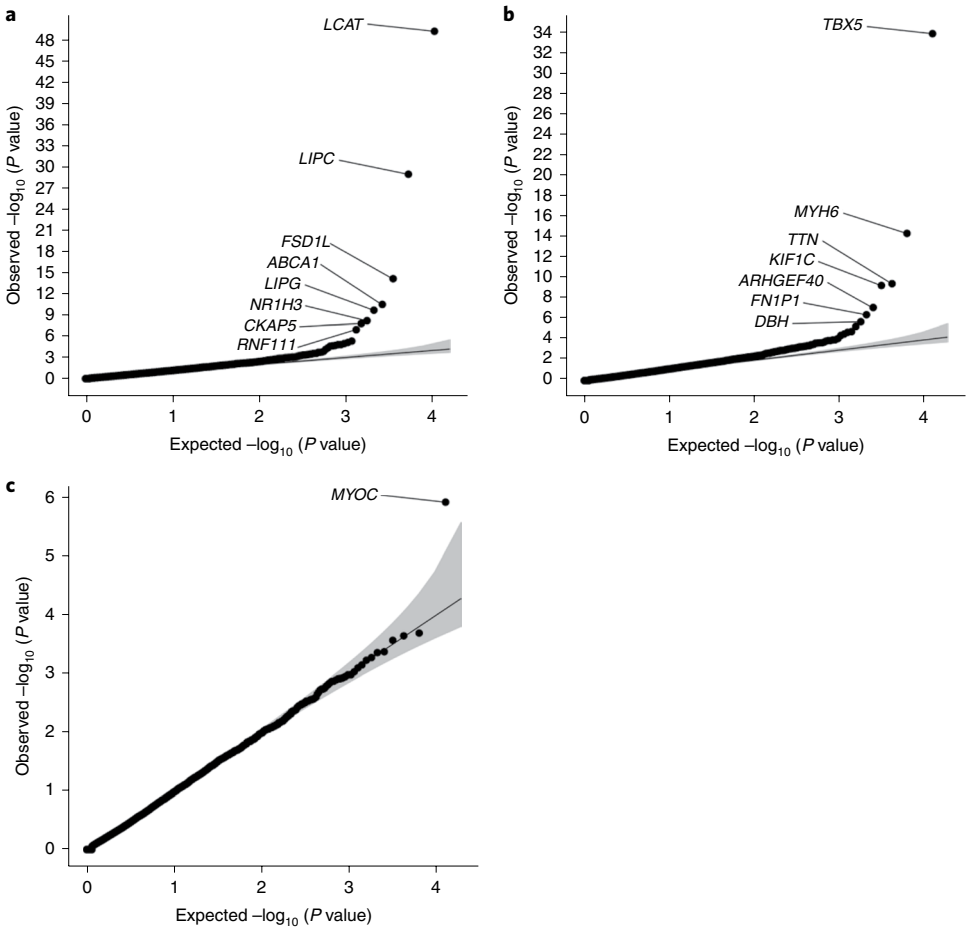


Fig. 2 | Quantile–quantile plots of exome-wide gene-based association results. a, Results for HDL in the HUNT study ($N = 69{,}214$). SKAT-O tests in SAIGE-GENE were performed for 13,416 genes with stop-gain and missense variants with MAF ≤ 1%, of which 10,600 genes having at least two variants are plotted. **b**, Results for automated readings of pulse rate in the UK Biobank ($N = 385{,}365$). SKAT-O tests in SAIGE-GENE were performed for 15,338 genes with stop-gain and missense variants with MAF ≤ 1%, of which 12,636 genes having at least two variants are plotted. **c**, Results for glaucoma in the UK Biobank (N cases = 4,462; N controls = 397,761). SKAT-O tests in SAIGE-GENE were performed for 15,338 genes with stop-gain and missense variants with MAF ≤ 1%, of which 12,638 genes having at least two variants are plotted. The gray shaded area represents the 95%-confidence band. N, sample size.

for six binary phenotypes reached the exome-wide significance threshold ($P \leq 2.5 \times 10^{-6}$; Supplementary Table 4b), all of which have been identified by both SAIGE-GENE and single-variant tests, including the gene *MYOC*, known for glaucoma[29] (Fig. 2c). Six genes for six binary phenotypes remained exome-wide significant after conditioning on the top variants (Supplementary Table 6b).

Simulation studies. We investigated the type I error rates and power of SAIGE-GENE by simulating genotypes and phenotypes for 10,000 samples in two settings. One setting had 500 families and 5,000 unrelated samples, and the other had 1,000 families. Each family had ten members, according to the pedigree shown in Supplementary Fig. 1.

Type I error rates. The type I error rates of SAIGE-GENE, EMMAX-SKAT and SMMAT were evaluated from 10^7 simulated gene–phenotype combinations, each with 20 genetic variants with MAF ≤ 1% on average. A sparse GRM with a cutoff of 0.2 for the coefficient of relatedness was used in SAIGE-GENE. Two different values of the variance component parameter corresponding to heritability $h^2 = 0.2$ and 0.4 were considered for quantitative traits (see Methods). The empirical type I error rates at the significance level (α) = 0.05, 1×10^{-4} and 2.5×10^{-6} are shown in Supplementary Table 7. Our simulation results suggest that SAIGE-GENE controls type I error rates relatively well, while the type I error rates are slightly inflated when heritability is relatively high ($h^2 = 0.4$). Similar results were observed on a larger sample with 1,000 families and 10,000 unrelated samples (Supplementary Note and Supplementary Table 8). Adjusting the test statistics by the genomic control (GC) inflation factor addressed the inflation (Supplementary Note).

Further simulations were conducted to evaluate the type I error rates of SAIGE-GENE, EMMAX-SKAT and SMMAT for phenotypes with skewed distributions, which are common in real data (Supplementary Fig. 2a). All three methods had inflated type I error rates for phenotypes with skewed distributions (Supplementary Table 9). With inverse normal transformation on phenotypes (Supplementary Fig. 2b), the inflation was reduced, but slight inflation was still observed (Supplementary Table 9). A potential reason is that inverse normal transformation disrupts sample relatedness in raw phenotypes, leading to poor fitting for the null GLMM. We then conducted a three-step phenotype transformation procedure as described in the Supplementary Note, which maintains sample relatedness in raw phenotypes, and observed well-controlled type I error rates for all three methods (Supplementary Table 10). Further simulation studies using real genotype data from the UK Biobank showed that SAIGE-GENE controlled type I error rates well in the presence of subtle population structure or non-negligible cryptic relatedness between

Table 1 | Genes significantly associated with automated readings of pulse rate ($N = 385,365$) and glaucoma (N cases = 4,462; N controls = 397,761) in the UK Biobank and HDL in the HUNT study ($N = 69,214$) at SKAT-O $P \leq 2.5 \times 10^{-6}$ from SAIGE-GENE

Gene	Number of markers	SAIGE-GENE SKAT-O test		Top hit in the locus		
		P	*P* conditional	Variant (GRCh37/hg19)	*P*	MAF
Pulse rate (UK Biobank)						
TBX5	4	9.69×10^{-35}	NA	12:114837349[C/A]	7.73×10^{-35}	0.0049
MYH6	14	3.61×10^{-15}	2.56×10^{-13}	14:23861811[A/G]	1.04×10^{-168}	0.3698
TTN	368	3.18×10^{-10}	3.41×10^{-6}	2:179721046[G/A]	8.73×10^{-100}	0.0885
KIF1C	12	4.78×10^{-10}	NA	17:4925475[C/T]	3.18×10^{-10}	0.0063
ARHGEF40	7	7.02×10^{-8}	2.57×10^{-10}	14:21542766[A/G]	3.30×10^{-52}	0.1688
FNIP1	8	3.58×10^{-7}	4.31×10^{-2}	5:131107733[C/T]	1.22×10^{-8}	0.0027
DBH	12	1.74×10^{-6}	1.74×10^{-6}	9:136149399[G/A]	3.46×10^{-6}	0.1870
HDL (HUNT)						
LCAT	3	7.34×10^{-50}	NA	16:67974303[A/T]	1.78×10^{-48}	0.0008
LIPC	4	1.25×10^{-29}	6.63×10^{-31}	15:58723939[G/A]	7.50×10^{-89}	0.1889
FSD1L	3	7.40×10^{-15}	1	9:107793713[T/C]	1.45×10^{-20}	0.0021
ABCA1	14	3.32×10^{-11}	1.28×10^{-11}	9:107620797[A/G]	3.64×10^{-48}	0.0055
LIPG	3	2.15×10^{-10}	2.41×10^{-10}	18:47156926[C/A]	5.92×10^{-40}	0.2348
NR1H3	2	6.53×10^{-9}	1.69×10^{-9}	11:47246397[G/A]	3.66×10^{-13}	0.3220
CKAP5	7	1.62×10^{-8}	1.21×10^{-9}	11:47246397[G/A]	3.66×10^{-13}	0.3220
RNF111	11	1.18×10^{-7}	1.37×10^{-9}	15:58856899[C/G]	2.82×10^{-24}	0.0047
Glaucoma (UK Biobank)						
MYOC	6	1.23×10^{-6}	NA	1:171605478[G/A]	9.13×10^{-16}	0.0014

Conditional analysis was performed when the top hit in the locus (±500 kb with respect to the start and end positions of the gene) was not included in the gene-based test. The *P* value for conditional analysis is indicated as 'NA' when the top hit was a rare missense or stop-gain variant that had been included in the gene-based test. *N*, sample size.

families (Supplementary Tables 11 and 12). Further details are provided in the Supplementary Note.

We also evaluated the type I error rates of SAIGE-GENE for binary traits with various case–control ratios. As with quantitative traits, a sparse GRM with a coefficient-of-relatedness cutoff of 0.2 was used. The variance component parameter $\tau = 1$ was assumed, corresponding to liability-scale heritability of 0.23. As expected, when case–control ratios were balanced or moderately unbalanced (for example, 1:1 and 1:9), type I error rates were well controlled, even without the robust adjustment, whereas when the ratios were extremely unbalanced (for example, 1:19 and 1:99) inflation was observed (Supplementary Table 13a and Extended Data Fig. 6). With the robust adjustment, type I error rates were relatively well controlled for the unbalanced case–control ratios (Supplementary Table 13b and Extended Data Fig. 6). However, for phenotypes with a case–control ratio of 1:99, slight inflation was still observed, although the inflation was dramatically alleviated in comparison to the unadjusted method. GC adjustment could then be used to further control type I error rates (Supplementary Table 13b). We also evaluated the empirical type I error rates of SAIGE-GENE for binary traits under case–control sampling with case–control ratios of 1:1 and 1:9, based on a disease prevalence of 1% in the population (Supplementary Note), and observed well-controlled type I error rates (Supplementary Table 14).

Power. We evaluated the empirical power of SAIGE-GENE and EMMAX-SKAT for quantitative traits. Two different settings for the proportion of causal variants were used: 10% and 40%. In each setting, among causal variants, 80% and 100% had negative effect sizes. The absolute effect sizes for causal variants were set to be $|0.3\log_{10}(\text{MAF})|$ and $|\log_{10}(\text{MAF})|$, respectively, when the proportion of causal variants was 0.4 and 0.1. Supplementary Table 15 shows that the power of both methods is nearly identical for all simulation settings for burden, SKAT and SKAT-O tests.

We also evaluated the empirical power of SAIGE-GENE for binary traits when using two different study designs: a cohort study with various values for disease prevalence (0.01–0.5) and case–control sampling with different case–control ratios (1:1–1:19), based on a disease prevalence of 1% in the population. In each setting, 40% of variants were causal variants. Among these, 80% were risk-increasing variants and 20% were risk-decreasing variants. The absolute effect sizes of causal variants were set to be $|0.55\log_{10}(\text{MAF})|$ and $|0.35\log_{10}(\text{MAF})|$ for the cohort study and case–control sampling, respectively. Supplementary Table 16 shows the empirical power of SKAT-O tests in both simulation studies. SAIGE-GENE had similar empirical power as unadjusted SAIGE-GENE in balanced case–control ratios and higher power in unbalanced scenarios. The power was low when the case–control ratio was 1:99 owing to the limited number of cases (100 cases), which can be alleviated with a larger sample size.

Discussion

In summary, we have presented a method, SAIGE-GENE, to perform gene- or region-based association tests in large cohorts or biobanks in the presence of sample relatedness. Similarly to SAIGE[9], which was previously developed for single-variant association tests, SAIGE-GENE uses a GLMM to account for sample relatedness, scalable computational approaches for large sample sizes and robust adjustment[14] to account for unbalanced case–control ratios for binary traits.

SAIGE-GENE uses several optimization strategies that are similar to those used in SAIGE to make fitting the null GLMM feasible for large sample sizes. For example, instead of storing the GRM in the memory, SAIGE-GENE stores genotypes in a binary vector and

computes the elements of the GRM as needed. PCG is used to solve linear systems instead of inverting a matrix. However, some optimization approaches are specifically applied in the gene-based tests in regard to rare variants. As estimating the variances of score statistics for rare variants is more sensitive to family structures, we use a sparse GRM to preserve close family structures rather than ignoring all sample relatedness. In addition, the variance ratios are estimated for different MAC categories, especially for extremely rare variants with MAC ≤ 20.

For binary phenotypes, SAIGE-GENE uses a robust adjustment and thereby also controls the type I error rates relatively well for both balanced and unbalanced case–control phenotypes. However, slight inflation is still observed in extremely unbalanced phenotypes (case–control ratio ≤ 1:99). To address this, we suggest using GC to further control type I error.

In numerical optimization, using good initial values can improve model convergence. In analysis of 24 quantitative traits in the UK Biobank with sample size (N) ≥ 100,000, we note that the models with the full GRM and the sparse GRM produced different variance component estimates, but they are relatively concordant (Pearson's correlation $r^2 = 0.66$; Supplementary Fig. 3). This indicates that the parameter estimates from the sparse GRM can be used as initial values to facilitate model fitting. We implemented this approach in SAIGE-GENE.

SAIGE-GENE has some limitations. First, similarly to SAIGE and other mixed-model methods, the time for algorithm convergence may vary among phenotypes and study samples given different heritability levels and sample relatedness. Second, similarly to SAIGE[9] and SMMAT[6], SAIGE-GENE uses penalized quasi-likelihood (PQL)[30] for binary traits to estimate the variance component, which is known to be biased. However, as shown in simulation studies in SAIGE[9] and SMMAT[6], PQL-based approaches work well to adjust for sample relatedness.

Overall, we have shown that SAIGE-GENE can account for sample relatedness while maintaining test power through simulation studies. By applying SAIGE-GENE to HUNT[9] and UK Biobank[2], we have demonstrated that SAIGE-GENE can identify potentially novel association signals. Currently, our method is the only available mixed-effect model approach for gene- or region-based rare-variant tests for large-sample data while accounting for unbalanced case–control ratios for binary traits. By providing a scalable solution to the current largest and future even larger datasets, our method will contribute to identifying trait-susceptibility rare variants and elucidating the genetic architecture of complex traits.

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/s41588-020-0621-6.

Received: 20 March 2019; Accepted: 31 March 2020;
Published online: 18 May 2020

References

1. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Preprint at *bioRxiv* https://doi.org/10.1101/563866 (2019).
2. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
3. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).
4. Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
5. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775 (2012).
6. Chen, H. et al. Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *Am. J. Hum. Genet.* **104**, 260–274 (2019).
7. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
8. Natarajan, P. et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat. Commun.* **9**, 3391 (2018).
9. 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).
10. 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).
11. Kuonen, D. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **4**, 7 (1999).
12. Daniels, H. E. Saddlepoint approximations in statistics. *Ann. Math. Stat.* **25**, 631–650 (1954).
13. Lee, S., Fuchsberger, C., Kim, S. & Scott, L. An efficient resampling method for calibrating single and gene-based rare variant association analysis in case–control studies. *Biostatistics* **17**, 1–15 (2016).
14. Zhao, Z. et al. UK Biobank whole-exome sequence binary phenome analysis with robust region-based rare-variant test. *Am. J. Hum. Genet.* **106**, 3–12 (2020).
15. Krokstad, S. et al. Cohort profile: the HUNT study, Norway. *Int. J. Epidemiol.* **42**, 968–977 (2013).
16. Langhammer, A., Krokstad, S., Romundstad, P., Heggland, J. & Holmen, J. The HUNT study: participation is associated with survival and depends on socioeconomic status, diseases and symptoms. *BMC Med. Res. Method.* **12**, 143 (2012).
17. Loh, P. R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
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. Liu, D. J. et al. Meta-analysis of gene-level tests for rare variant association. *Nat. Genet.* **46**, 200–204 (2014).
20. 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).
21. Willer, C. J. et al. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
22. Willer, C. J. et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet.* **40**, 161–169 (2008).
23. Holm, H. et al. Several common variants modulate heart rate, PR interval and QRS duration. *Nat. Genet.* **42**, 117–122 (2010).
24. Eijgelsheim, M. et al. Genome-wide association analysis identifies multiple loci related to resting heart rate. *Hum. Mol. Genet.* **19**, 3885–3894 (2010).
25. Eppinga, R. N. et al. Identification of genomic loci associated with resting heart rate and shared genetic predictors with all-cause mortality. *Nat. Genet.* **48**, 1557–1563 (2016).
26. Arking, D. E. et al. Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. *Nat. Genet.* **46**, 826–836 (2014).
27. Swoap, S. J., Weinshenker, D., Palmiter, R. D. & Garber, G. *Dbh*$^{-/-}$ mice are hypotensive, have altered circadian rhythms, and have abnormal responses to dieting and stress. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **286**, R108–R113 (2004).
28. Marouli, E. et al. Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).
29. Turalba, A. V. & Chen, T. C. Clinical and genetic characteristics of primary juvenile-onset open-angle glaucoma (JOAG). *Semin. Ophthalmol.* **23**, 19–25 (2008).
30. Breslow, N. E. & Clayton, D. G. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* **88**, 9–25 (1993).

Methods

Generalized linear mixed model. In a study with sample size N, we denote the phenotype of the ith individual by using y_i for both quantitative and binary traits. Let the $1\times(p+1)$ vector X_i represent p covariates including the intercept and the $N\times q$ matrix G_i represent the allele counts (0, 1 or 2) for q variants in the gene to test. The GLMM can be written as

$$g(\mu_i) = X_i\boldsymbol{\alpha} + G_i\beta + b_i$$

where μ_i is the mean phenotype and b_i is the random effect, which is assumed to be distributed as $N(0, \tau\psi)$, where ψ is an $N\times N$ GRM and τ is the additive genetic variance parameter. The link function g is the identity function for quantitative traits with error term $\varepsilon \sim N(0, \phi I)$ and a logistic function for binary traits. The parameter α is a $(p+1)\times 1$-coefficient vector of fixed effects and β is a $q\times 1$-coefficient vector of the genetic effect.

Estimation of the variance component and other model parameters (step 1). As in the original SAIGE[9] and GMMAT[31], to fit the null GLMM in SAIGE-GENE, the PQL method[30,32] and the computationally efficient average information restricted maximum likelihood (AI-REML) algorithm[31,33] are used to iteratively estimate $(\hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{b}})$ under the null hypothesis of $\beta=0$. At iteration k, let $(\hat{\boldsymbol{\tau}}^{(k)}, \hat{\boldsymbol{\alpha}}^{(k)}, \hat{\boldsymbol{b}}^{(k)})$ be the estimated $(\hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{b}})$, $\hat{\mu}_i^{(k)}$ be the estimated mean of y_i and $\hat{\Sigma}^{(k)} = \left\{\hat{W}^{(k)}\right\}^{-1} + \hat{\boldsymbol{\tau}}^{(k)}\psi$ be an $N\times N$ matrix of the variance of working vector $\tilde{\boldsymbol{y}}$, in which ψ is the $N\times N$ GRM. For quantitative traits, $\hat{W}^{(k)} = \hat{\phi}^{-1}I$ and $\tilde{y}_i = X_i\boldsymbol{\alpha}^{(k)} + b_i^{(k)}$. For binary traits, $\hat{W}^{(k)} = \mathrm{diag}\left[\hat{\mu}_i^{(k)}\left(1-\hat{\mu}_i^{(k)}\right)\right]$ and $\tilde{y}_i = X_i\boldsymbol{\alpha}^{(k)} + b_i^{(k)} + \left(y_i - \hat{\mu}_i^{(k)}\right)/\left\{\hat{\mu}_i^{(k)}\left(1-\hat{\mu}_i^{(k)}\right)\right\}$. To obtain the log quasi-likelihood and average information at each iteration, SAIGE and SAIGE-GENE use PCG[31,32] to obtain the product of the inverse of $\hat{\Sigma}^{(k)}$ and any other vector by iteratively solving a linear system with $\hat{\Sigma}^{(k)}$. This approach is more computationally efficient than using Cholesky decomposition to obtain $\left\{\hat{\Sigma}^{(k)}\right\}^{-1}$. The numerical accuracy of PCG was evaluated in the SAIGE paper[9].

Gene-based association tests (step 2). Test statistics for the burden, SKAT and SKAT-O tests for a gene can be constructed on the basis of score statistics from the marginal model for individual variants in the gene. Suppose there are q variants in the region or gene to test. The score statistic for variant j ($j=1, ..., q$) under H_0: $\beta_j=0$ is $T_j = g_j^T(\boldsymbol{Y} - \hat{\boldsymbol{\mu}})$, where g_j and Y are $N\times 1$ genotype and phenotype vectors, respectively, and $\hat{\boldsymbol{\mu}}$ is the estimated mean of Y under the null hypothesis.

Let u_j denote a threshold indicator or weight for variant j and $U=\mathrm{diag}(u_1, \ldots, u_q)$ be a diagonal matrix with u_j as the jth element. Similarly to the original SKAT and SKAT-O papers[4,5], to upweight rare variants, the default setting in SAIGE-GENE is $u_j = Beta\left(\mathrm{MAF}_j, 1, 25\right)$, which upweights rarer variants. Burden test statistics can be written as $Q_{\mathrm{burden}} = \left(\sum_{j=1}^{q} u_j T_j\right)^2$.

Suppose $\tilde{G} = G - X\left(X^T\hat{W}X\right)^{-1}X^T\hat{W}G$ is the covariate-adjusted genotype matrix, where $G=(g_1, \ldots, g_q)$ is the $N\times q$ genotype matrix of the q genetic variants and $\hat{P} = \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}X\left(X^T\hat{\Sigma}^{-1}X\right)^{-1}X^T\hat{\Sigma}^{-1}$ with $\hat{\Sigma} = \hat{W}^{-1} + \hat{\tau}\psi$. Under the null hypothesis of no genetic effects, Q_{burden} follows $\lambda_B\chi_1^2$, where $\lambda_B = \boldsymbol{J}^T U\tilde{G}^T\hat{P}\tilde{G}U\boldsymbol{J}$, J is a $q\times 1$ vector with all elements being unity, and χ_1^2 is a chi-squared distribution with 1 degree of freedom[3]. The SKAT test[4] statistic can be written as $Q_{SKAT} = \sum_{j=1}^{q} u_j^2 T_j^2$, which follows a mixture of chi-squared distribution $\sum_{j=1}^{q}\lambda_{Sj}\chi_1^2$, where λ_{Sj} corresponds to the eigenvalues of $U\tilde{G}^T\hat{P}\tilde{G}U$. The SKAT-O test[5] uses a linear combination of the burden and SKAT test statistics, where $Q_{SKAT-O} = (1-\rho)Q_{SKAT} + \rho Q_{burden}, 0\le\rho\le 1$. To conduct the test, the minimum P value from the grid of ρ is calculated and the P value of the test statistic based on the minimum P value is estimated through numerical integration. Following the suggestion in Lee et al.[34], we use a grid of eight values of $\rho = (0, 0.1^2, 0.2^2, 0.3^2, 0.4^2, 0.5^2, 0.5, 1)$ to find the minimum P value.

Approximation of $\tilde{\boldsymbol{G}}^T\hat{\boldsymbol{P}}\tilde{\boldsymbol{G}}$. For each gene, given $\hat{P}$, the calculation of $\tilde{G}^T\hat{P}\tilde{G}$ requires applying PCG for each variant in the gene, which can be computationally very expensive. Suppose $\tilde{\boldsymbol{g}}$ represents a covariate-adjusted single-variant genotype vector. To reduce computational cost, an approximation approach has been used in SAIGE, BOLT-LMM[17] and GRAMMAR-Gamma[18], in which the ratio between $\tilde{\boldsymbol{g}}^T\hat{P}\tilde{\boldsymbol{g}}$ and $\tilde{\boldsymbol{g}}^T\tilde{\boldsymbol{g}}$ is estimated from a small subset of randomly selected genetic markers. The ratio has been shown to be approximately constant for all variants. Given the estimated ratio $\hat{r} = \tilde{\boldsymbol{g}}^T\hat{P}\tilde{\boldsymbol{g}}/\tilde{\boldsymbol{g}}^T\tilde{\boldsymbol{g}}$, $\tilde{\boldsymbol{g}}^T\hat{P}\tilde{\boldsymbol{g}}$ for all other variants can be obtained as $\hat{r}\tilde{\boldsymbol{g}}^T\tilde{\boldsymbol{g}}$. However, the variation of the estimated $\hat{r}$ for extremely rare variants is large and including some closely related samples in the denominator helps reduce the variation of $\hat{r}$, as shown in Supplementary Fig. 2. Let ψ_S denote a sparse GRM that preserves close family structure and ψ_f denote the full GRM. We estimate the ratio $\hat{r}_s = \tilde{\boldsymbol{g}}^T\hat{P}\tilde{\boldsymbol{g}}/\tilde{\boldsymbol{g}}^T\hat{P}_s\tilde{\boldsymbol{g}}$, where $\hat{P}_s = \hat{\Sigma}_s^{-1} - \hat{\Sigma}_s^{-1}X\left(X^T\hat{\Sigma}_s^{-1}X\right)^{-1}X^T\hat{\Sigma}_s^{-1}$ and $\hat{\Sigma}_s = \hat{W}^{-1} + \hat{\tau}\psi_s$.

In ψ_s, elements below a user-specified relatedness coefficient cutoff, that is, more distantly related than third-degree relatives, are zeroed out, with only close family structures preserved. To construct ψ_s, a subset of randomly selected genetic markers, that is, 2,000, are first used to quickly estimate which related samples pass the user-specified cutoff. Then, the relatedness coefficients for these samples are further estimated by using the full set of genetic markers, which equal corresponding values in ψ_f. In the model fitting using ψ_s, $\hat{\Sigma}_s^{-1}X$ and $\hat{\Sigma}_s^{-1}\tilde{\boldsymbol{g}}$ need to be calculated. For this, we use a sparse lower-upper (LU)-based solve method[35] implemented in R. The constructed ψ_s is also used for approximating the variance of score statistics with ψ_f. For a biobank or dataset, ψ_s only needs to be constructed once and can be reused for any phenotype in the same dataset.

SAIGE-GENE estimates variance ratios for different MAC categories. By default, the MAC categories are set to be MAC = 1, 2, 3, 4, 5, 6 to 10, 11 to 20, and >20. Once the MAC categorical variance ratios are estimated, for each genetic marker in tested genes or regions, $\hat{r}_*$ can be obtained according to the MAC. Let $\hat{R}_s$ be a $q\times q$ diagonal matrix whose jth diagonal element is the ratio $\hat{r}_*$ for the jth marker in the gene (that is, $\tilde{g}_j^T\hat{P}\tilde{g}_j/\tilde{g}_j^T\hat{P}_s\tilde{g}_j$). For the tested gene with q markers, $\tilde{G}^T\hat{P}\tilde{G}$ can be approximated as $\hat{R}_s^{\frac{1}{2}}\tilde{G}^T\hat{P}_s\tilde{G}\hat{R}_s^{\frac{1}{2}}$ (see the Supplementary Note for further details).

Robust adjustment for $\hat{\boldsymbol{R}}_s^{\frac{1}{2}}\tilde{\boldsymbol{G}}^T\hat{\boldsymbol{P}}_s\tilde{\boldsymbol{G}}\hat{\boldsymbol{R}}_s^{\frac{1}{2}}$ to account for unbalanced case–control ratios. To account for unbalanced case–control ratios of binary traits in region- or gene-based tests, we recently developed a robust adjustment for independent samples[14]. The approach first obtains well-calibrated P values for single-variant score statistics by using SPA[10–12] and ER[13]. SPA is a method to calculate P values by inverting the cumulant generating function (CGF). Because CGF completely specifies the distribution, SPA can be far more accurate than using the normal distribution. However, because SPA is still an asymptotic-based approach, it does not work well when variants are very rare (for example, MAC ≤ 10). For such variants, we use ER, which resamples the case–control status of only individuals carrying a minor allele and is extremely fast for very rare variants. To account for the fact that individuals can have different non-genetic risk of diseases (due to covariates), the resampling was done with estimated disease risk μ_i. Next, the variances of single-variant score statistics are obtained by inverting the P values, which are then used to calibrate the variances of region- or gene-based test statistics. We have extended the approach for related samples in SAIGE-GENE. For variants with MAC > 10, single-variant P values are obtained by SAIGE, which basically applies SPA to the GLMM. For variants with MAC ≤ 10, we use ER with GLMM-estimated $\hat{\mu}_i$, which includes the random effect to maintain the correlation structure among samples. After calculating P values of T_j for $j=1, \ldots, q$, the variance of T_j is calibrated by inverting the corresponding P value. Then, the calibrated variance is applied to $\hat{R}_S^{\frac{1}{2}}\tilde{G}^T\hat{P}_s\tilde{G}\hat{R}_S^{\frac{1}{2}}$ to compute a robust P value for the region- or gene-based test. Details can be found in the Supplementary Note.

Conditional analysis. In SAIGE-GENE, we implemented conditional analysis to perform gene-based tests conditioning on given markers by using the summary statistics from the unconditional gene-based tests and the LD (r^2) between testing and conditioning markers[19]. Let G be the genotypes for a gene to be tested for association, which contains q markers, and G_2 be the genotypes for the conditioning markers, which contains q_2 markers. Let β denote a $q\times 1$-coefficient vector of the genetic effect for the gene to be tested and β_2 be a $q_2\times 1$-coefficient vector of the genetic effect for the conditioning markers. The genotype matrix with non-genetic covariates projected out $\tilde{G} = G - X(X^T\hat{W}X)^{-1}X^T\hat{W}G$ and $\tilde{G}_2 = G_2 - X(X^T\hat{W}X)^{-1}X^T\hat{W}G_2$. In the unconditioned association tests, the test statistics are $\boldsymbol{T} = \tilde{G}^T(\boldsymbol{Y} - \hat{\boldsymbol{\mu}})$ and $\boldsymbol{T}_2 = \tilde{G}_2^T(\boldsymbol{Y} - \hat{\boldsymbol{\mu}})$. In conditional analysis, under the null hypothesis, $\mathrm{E}(T) = \mathrm{E}\left(\tilde{G}^T P\left(\tilde{G}_2\boldsymbol{\beta}_2\right)\right) = \tilde{G}^T\hat{P}\tilde{G}_2\boldsymbol{\beta}_2$ and $\mathrm{E}(T_2) = \mathrm{E}(\tilde{G}_2^T P(\tilde{G}_2\boldsymbol{\beta}_2)) = G_2^T\hat{P}_s\tilde{G}_2\boldsymbol{\beta}_2$. T and T_2 jointly follow the multivariate normal distribution with mean $(\mathrm{E}(T), \mathrm{E}(T_2))$ and variance $S = \begin{bmatrix}\tilde{G}^T\hat{P}\tilde{G} & \tilde{G}^T\hat{P}\tilde{G}_2\\ \tilde{G}_2^T\hat{P}\tilde{G} & \tilde{G}_2^T\hat{P}\tilde{G}_2\end{bmatrix}$.

Thus, under the null hypothesis of no association of T, that is, H_0: $\beta=0$, $T\,|\,T_2$ follows the conditional normal distribution with $\mathrm{E}(T\,|\,T_2) = \tilde{G}^T\hat{P}\tilde{G}_2\left(\tilde{G}_2^T\hat{P}\tilde{G}_2\right)^{-1}\boldsymbol{T}_2$ and $\mathrm{var}(T\,|\,T_2) = \tilde{G}^T\hat{P}\tilde{G} - \tilde{G}^T\hat{P}\tilde{G}_2\left(\tilde{G}_2^T\hat{P}\tilde{G}_2\right)^{-1}\tilde{G}_2^T\hat{P}\tilde{G}$, and P values can be calculated from the conditional distribution.

Data simulation. We carried out a series of simulations to evaluate and compare the performance of SAIGE-GENE, EMMAX-SKAT[5,7] and SMMAT[6]. We used the sequence data from 10,000 European-ancestry chromosomes over 1-Mb regions that were generated by using the calibrated coalescent model in the SKAT R package[5]. We randomly selected 10,000 regions with 3 kb from the sequence data, followed by the gene-dropping simulation[36] using these sequences as founder haplotypes that were propagated through the pedigree of ten family members shown in Supplementary Fig. 1. Only variants with MAF ≤ 1% were used for simulation studies. Quantitative phenotypes were generated from the following LMM: $y_i = X_1 + X_2 + G_i\beta + b_i + \varepsilon_i$, where G_i is the genotype value, β is the genetic effect size, b_i is the random effect simulated from $N(0, \tau\psi)$ and ε_i is the error term simulated from $N(0, (1-\tau)I)$. Two covariates, X_1 and X_2, were simulated from Bernoulli(0.5) and $N(0, 1)$, respectively. Binary phenotypes were generated from the logistic mixed model $\mathrm{logit}(\pi_{i0}) = \alpha_0 + b_i + X_1 + X_2 + G_i\beta$, where β is the genetic log odds ratio and b_i is the random effect simulated from $N(0, \tau\psi)$ with $\tau=1$. The intercept α_0 was determined by the disease prevalence (that is, case–control ratios). Given $\tau=1$, the liability-scale heritability is 0.23 (ref. [37]).

To evaluate the type I error rates at exome-wide $\alpha = 2.5 \times 10^{-6}$, we first simulated 10,000 regions and then simulated 1,000 sets of quantitative phenotypes for each simulated region with different random seeds under the null hypothesis of $\beta = 0$. Gene-based association tests were performed with SAIGE-GENE, EMMAX-SKAT and SMMAT; therefore, in total, 10^7 tests each of burden, SKAT and SKAT-O tests were carried out. Two different settings for τ were evaluated (0.2 and 0.4), and two different sample relatedness settings were used (one had 500 families and 5,000 independent samples, and the other had 1,000 families, each with 10 family members). We also simulated 1,000 sets of binary phenotypes for case–control ratios of 1:99, 1:19, 1:9, 1:4 and 1:1 for 500 families and 5,000 independent samples. Burden, SKAT and SKAT-O tests were performed on the 10,000 genomic regions with SAIGE-GENE, corresponding to a total of 10^7 tests for each method for each case–control ratio.

For power simulation, phenotypes were generated under the alternative hypothesis $\beta \neq 0$. Two different settings for the proportion of causal variants were used: 10% and 40%, corresponding to $|\beta| = |\log_{10}(\text{MAF})|$ and $|\beta| = |0.3\log_{10}(\text{MAF})|$, respectively. In each setting, 80% and 100% of the variants had negative effect sizes. We simulated 1,000 datasets in each simulation and evaluated power at test-specific empirical α, which yields nominal $\alpha = 2.5 \times 10^{-6}$. The empirical α was estimated from the type I error simulations. Similarly, 1,000 sets of binary traits were generated for 10,000 samples (500 families and 5,000 independent samples) under the alternative hypothesis $\beta \neq 0$ using two different settings: a cohort study with various disease prevalence values (0.01, 0.05, 0.1 and 0.5) and case–control sampling with three different case–control ratios (1:19, 1:9 and 1:1), based on a disease prevalence of 1% in the population (Supplementary Note). Forty percent of variants were simulated as causal variants, among which 80% were risk-increasing variants and 20% were risk-decreasing variants. The absolute effect sizes of causal variants were set to be $|0.55\log_{10}(\text{MAF})|$ and $|0.35\log_{10}(\text{MAF})|$ for the cohort study and case–control sampling, respectively.

HUNT and UK Biobank data analysis. We applied SAIGE-GENE to HDL levels in 69,500 Norwegian samples from the population-based HUNT study[15,16]. About 70,000 HUNT participants were genotyped on Illumina HumanCoreExome v1.0 and v1.1 and imputed by Minimac3 (ref. [38]) with a merged reference panel of HRC[39] and whole-genome sequencing data for 2,201 HUNT samples. Variants with imputation $r^2 < 0.8$ were excluded from further analysis. Participation in the HUNT study is based on informed consent, and the study has been approved by the Data Inspectorate and the Regional Ethics Committee for Medical Research in Norway. A total of 13,416 genes with at least two rare (MAF ≤ 1%) missense and/or stop-gain variants with imputation $r^2 \geq 0.8$ were tested. Variants were annotated with SeattleSeq annotations (http://snp.gs.washington.edu/SeattleSeqAnnotation138/). We used 249,749 pruned genotyped markers to estimate relatedness coefficients in the full GRM for step 1 and used a relatedness coefficient cutoff of ≥0.125 for the sparse GRM.

We also analyzed 53 quantitative traits and 10 binary traits with SAIGE-GENE in the UK Biobank for 408,910 participants with White British ancestry[2]. UK Biobank protocols were approved by the National Research Ethics Service Committee, and participants signed written informed consent. Markers that were imputed by the HRC[39] panel with imputation info score ≥ 0.8 were used in the analysis. A total of 15,342 genes with at least two rare (MAF ≤ 1%) missense and/or stop-gain variants that were directly genotyped or successfully imputed from HRC (imputation score ≥ 0.8) were tested. We used 340,447 pruned markers, which were pruned from the directly genotyped markers by using the following parameters, to construct the GRM: window size of 500 bp, step size of 50 bp and pairwise $r^2 < 0.2$. We used a relatedness coefficient cutoff of ≥0.125 for the sparse GRM.

Genome build. All genomic coordinates are given according to NCBI Build 37/UCSC hg19.

Statistical analysis. We performed gene-based burden, SKAT and SKAT-O tests with SAIGE-GENE on 15,342 genes for 53 quantitative traits and 10 binary traits in 408,910 UK Biobank participants with White British ancestry who passed the quality control in the UK Biobank[2]. In the linear mixed model for quantitative traits, the first four genetic principal components, gender and the age when the participant attended the assessment center were included as non-genetic covariates. In the logistic mixed model for binary traits, the first four genetic principal components, gender and birth year were included as non-genetic covariates. We also performed the same gene-based tests on 13,416 genes for HDL levels in 69,500 Norwegian samples from the HUNT study[15,16]. In the linear mixed model for HDL, age, sex, genotyping batch and the first four principal components were included as non-genetic covariates. The numbers of samples used for analysis are included in the legend of each figure.

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

Data availability

The summary statistics and quantile–quantile plots for 53 quantitative phenotypes and 10 binary phenotypes in the UK Biobank by SAIGE-GENE are available for public download at https://www.leelabsg.org/resources.

Code availability

SAIGE-GENE is implemented as an open-source R package available at https://github.com/weizhouUMICH/SAIGE/.

References

31. 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).
32. Lee, S. H. & van der Werf, J. H. An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. *Genet. Sel. Evol.* **38**, 25–43 (2006).
33. 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).
34. Lee, S., Teslovich, T. M., Boehnke, M. & Lin, X. General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.* **93**, 42–53 (2013).
35. Davis, T. A. *Direct Methods for Sparse Linear Systems (Fundamentals of Algorithms 2)* (Society for Industrial and Applied Mathematics, 2006).
36. 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).
37. 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).
38. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
39. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

Acknowledgements

This research has been conducted using the UK Biobank Resource under application number 45227. The Nord-Trøndelag Health Study (the HUNT study) is a collaboration between the HUNT Research Centre (Faculty of Medicine, Norwegian University of Science and Technology (NTNU)), the Nord-Trøndelag County Council, the Central Norway Health Authority and the Norwegian Institute of Public Health. The K.G. Jebsen Center for Genetic Epidemiology is financed by Stiftelsen Kristian Gerhard Jebsen, the Faculty of Medicine and Health Sciences at the Norwegian University of Science and Technology (NTNU) and the Central Norway Regional Health Authority. S.L. and W.B. were supported by National Institutes of Health grant R01HG008773. W.Z. was supported by the National Human Genome Research Institute of the National Institutes of Health under award number T32HG010464.

Author contributions

W.Z., Z.Z. and S.L. designed experiments. W.Z., Z.Z. and S.L. performed experiments. W.Z. implemented the software with input from W.B. and J.L. J.B.N., L.G.F. and S.A.G.T. constructed phenotypes for UK Biobank data. M.E.G. and K.H. provided data for the HUNT study. W.Z., Z.Z., C.J.W., S.L. and G.R.A. analyzed UK Biobank data. B.M.N. and M.J.D. provided helpful advice. W.Z., Z.Z. and S.L. wrote the manuscript with input from S.A.G.T. and M.E.G.

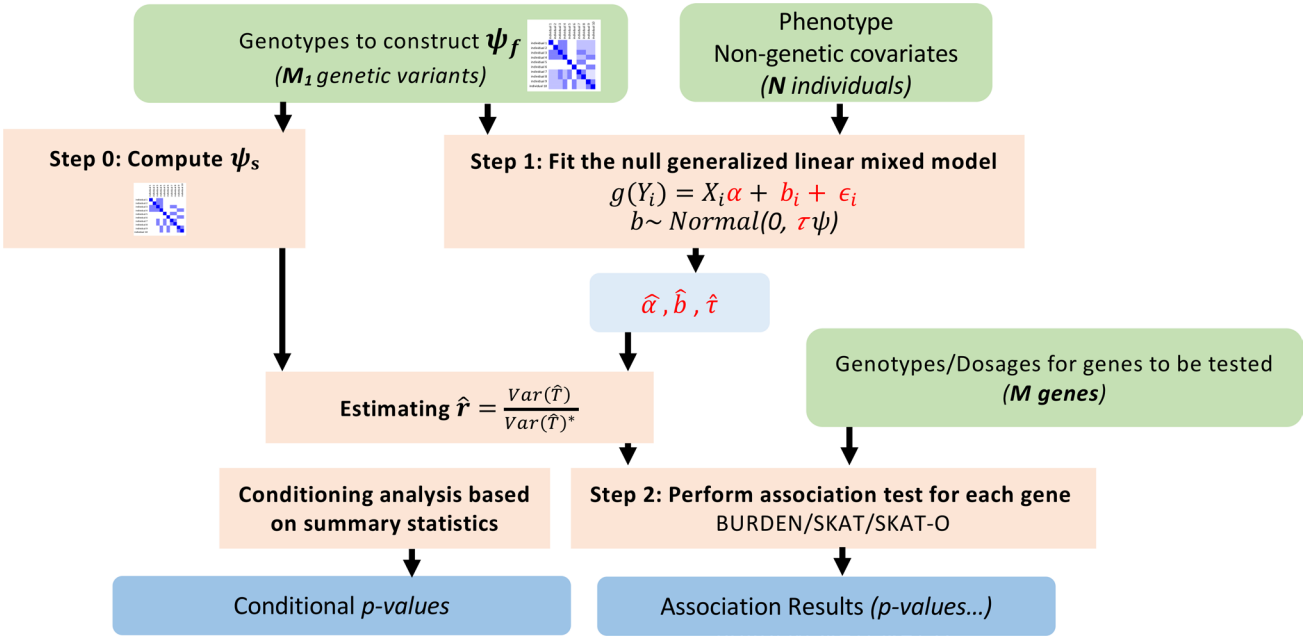
Competing interests

G.R.A. is an employee of Regeneron Pharmaceuticals. He owns stock and stock options for Regeneron Pharmaceuticals. B.M.N. is a member of the Deep Genomics Scientific Advisory Board, has received travel expenses from Illumina and also serves as a consultant for Avanir and Trigeminal solutions.

Additional information

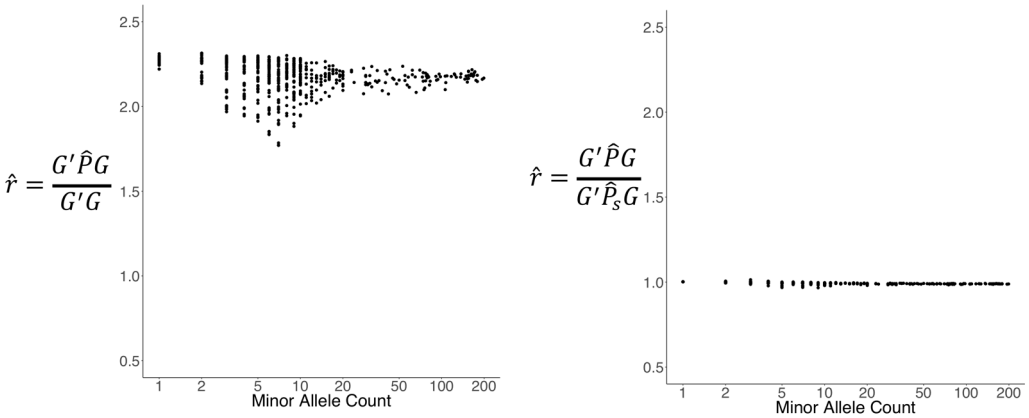
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0621-6.

Correspondence and requests for materials should be addressed to W.Z. or S.L.

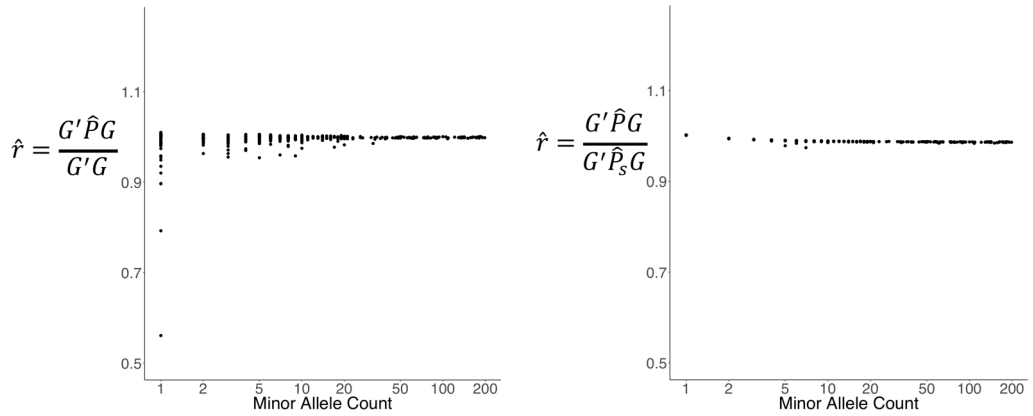


Extended Data Fig. 1 | Workflow of SAIGE-GENE. SAIGE-GENE consists of two steps: (1) Fitting the null generalized linear mixed model (GLMM) to estimate variance components and other model parameters; (2) Testing for association between each genetic variant set, such as a gene or a region, and the phenotype.

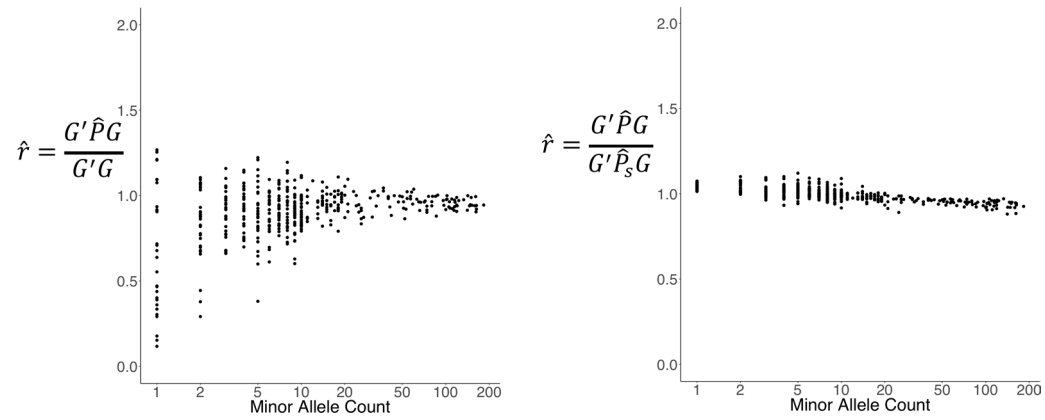
A. Simulation: 500 families and 5,000 independent individuals



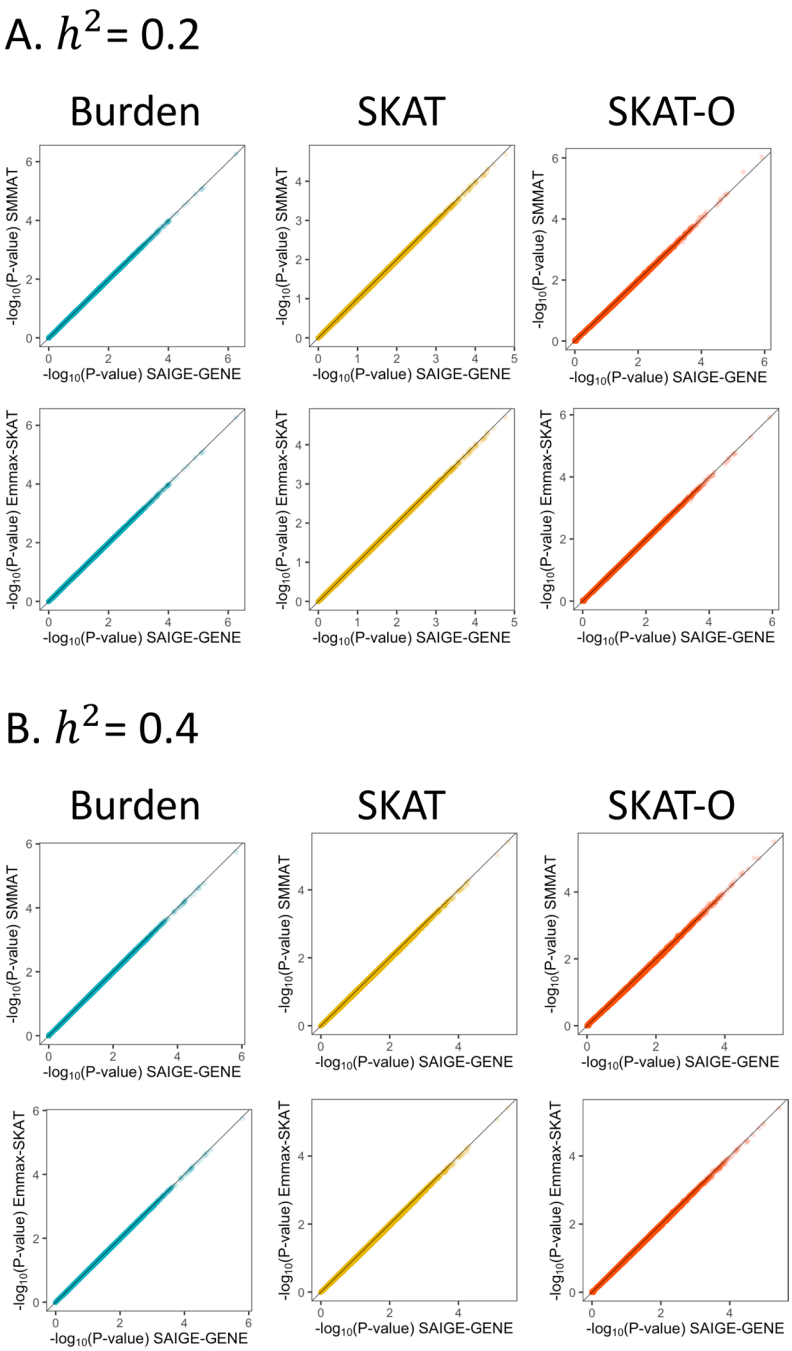
B. UK Biobank: Pulse rate automated read (mean), randomly selected 20,000 individuals



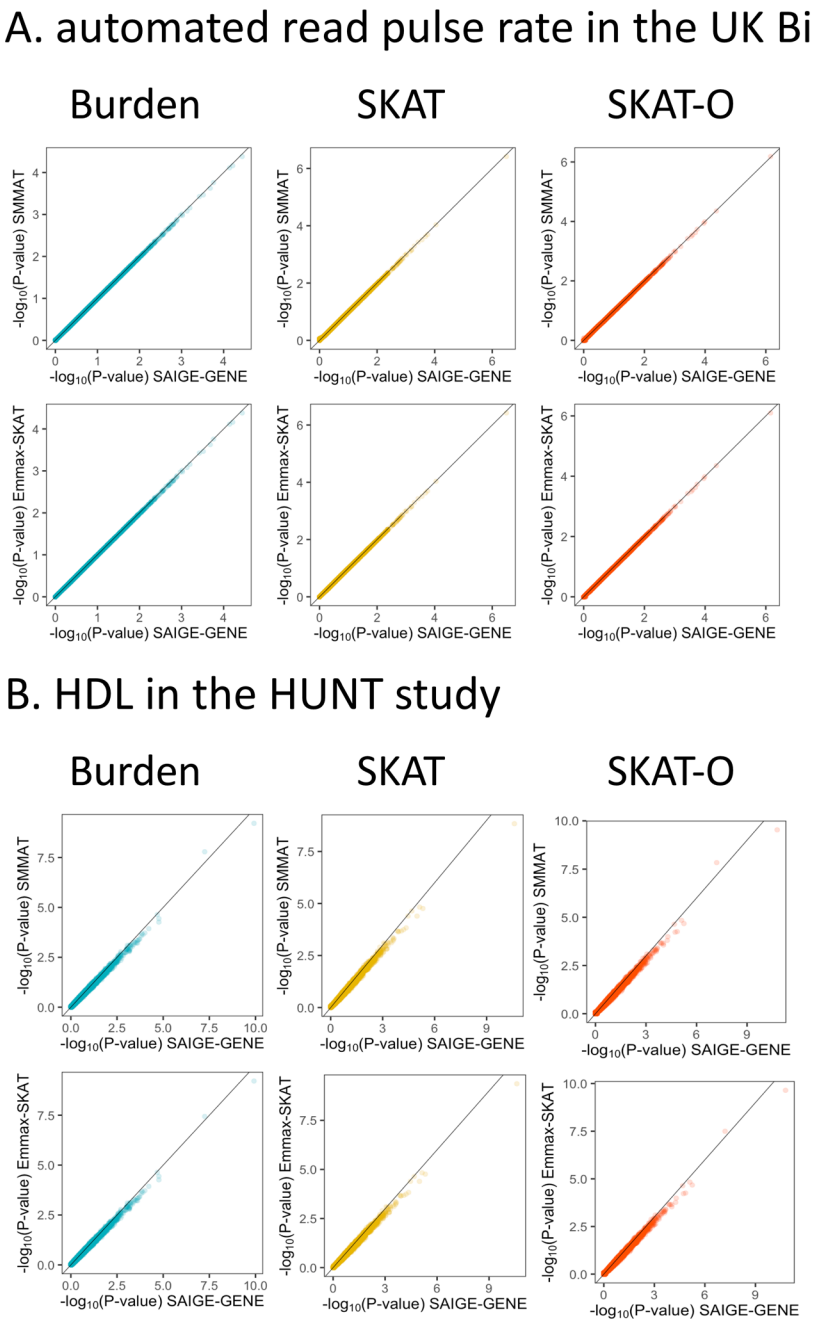
C. HUNT: HDL, randomly selected 20,000 individuals



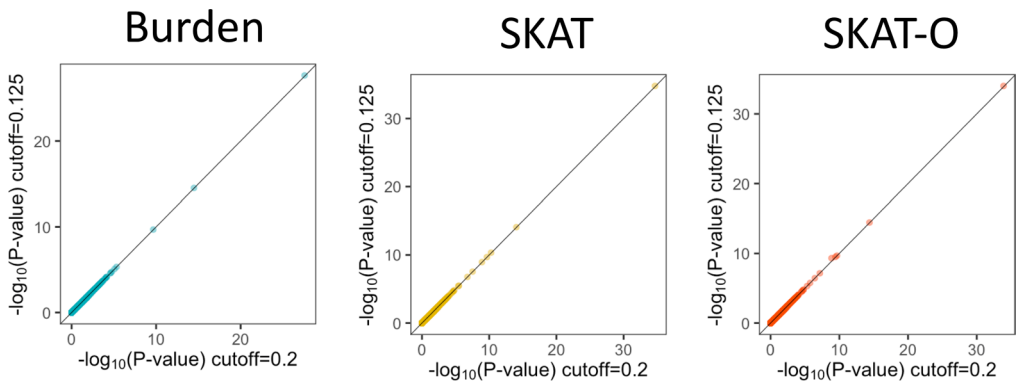
Extended Data Fig. 2 | Plots of the variance ratio of the score statistics by MAC for rare variants with and without the full GRM for sample relatedness (left) and with the full GRM and a sparse GRM for closely related samples (right). a, Genotypes were simulated for 500 families and 5,000 independent individuals based on the pedigree structure shown in Supplementary Fig. 1 and the null model was fitted for the simulated quantitative trait with $h^2 = 0.2$. The sparse GRM was constructed using a coefficient of relatedness cutoff 0.2. **b**, 20,000 samples with White British ancestry were randomly selected from the UK Biobank and the null model was fitted for the automated read pulse rate. The sparse GRM was constructed using a coefficient of relatedness cutoff 0.125. **c**, 20,000 samples were randomly selected form the HUNT study and the null model was fitted for HDL. The sparse GRM was constructed using a coefficient of relatedness cutoff 0.125.



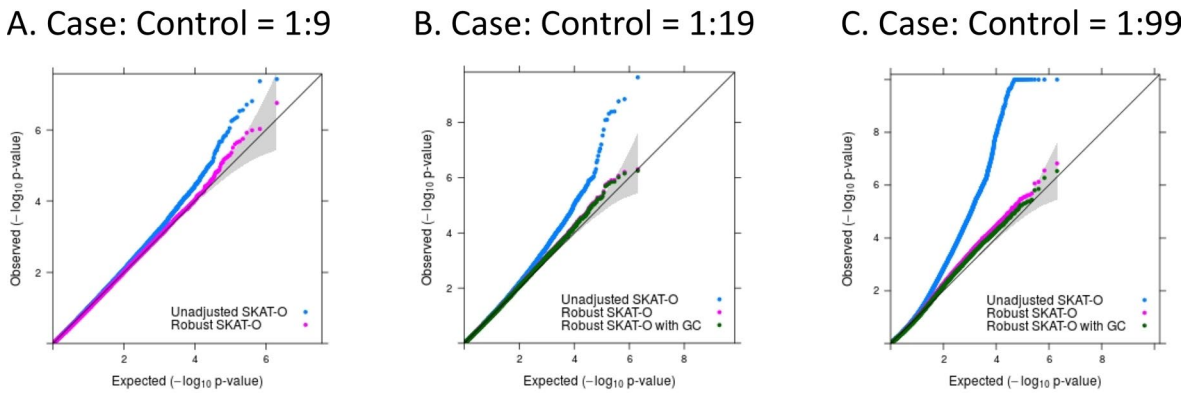
Extended Data Fig. 3 | Scatter plots of association *P*-values from SAIGE-GENE versus SMMAT and EmmaX-SKAT for the Burden, SKAT, and SKAT-O tests based on simulation data on the -log$_{10}$ scale. 1,000,000 genes were tested with 1,000 families, each having 10 members, as shown in the Supplementary Fig. 1. The Pearson's correlation coefficients $r^2 > 0.99$ for -log$_{10}$(*P*-values) between SAIGE and SMMAT and between SAIGE and EmmaX-SKAT. **a**, $h^2 = 0.2$. **b**, $h^2 = 0.4$.



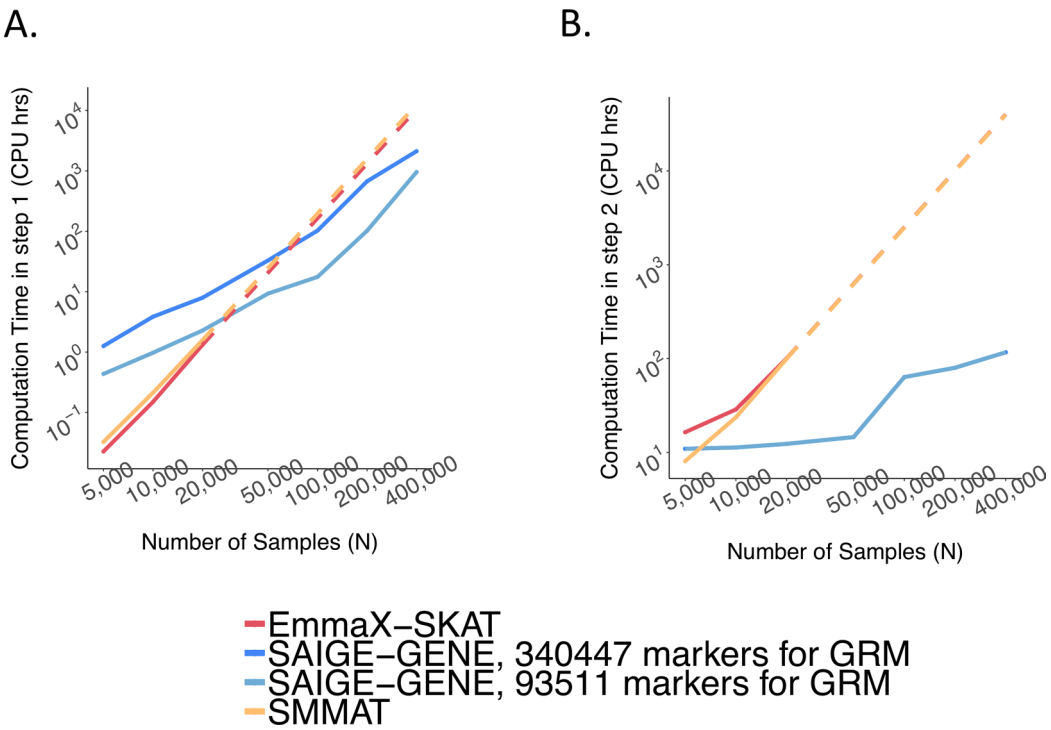
Extended Data Fig. 4 | Scatter plots of association *P*-values from SAIGE-GENE versus SMMAT and EmmaX-SKAT for the Burden, SKAT, and SKAT-O tests based on real data analysis on the -log$_{10}$ scale. a,**b**, 12,000 genes were tested for automated read pulse rate on 20,000 randomly selected white British samples in the HRC-imputed UK Biobank (**a**) and for HDL on 20,000 randomly selected samples in HUNT (**b**). Missense and stop-gain variants with MAF ≤ 1% were included. The Pearson's correlation coefficients $r^2 > 0.99$ for -log$_{10}$(*P*-values) between SAIGE and SMMAT and between SAIGE and EmmaX-SKAT.



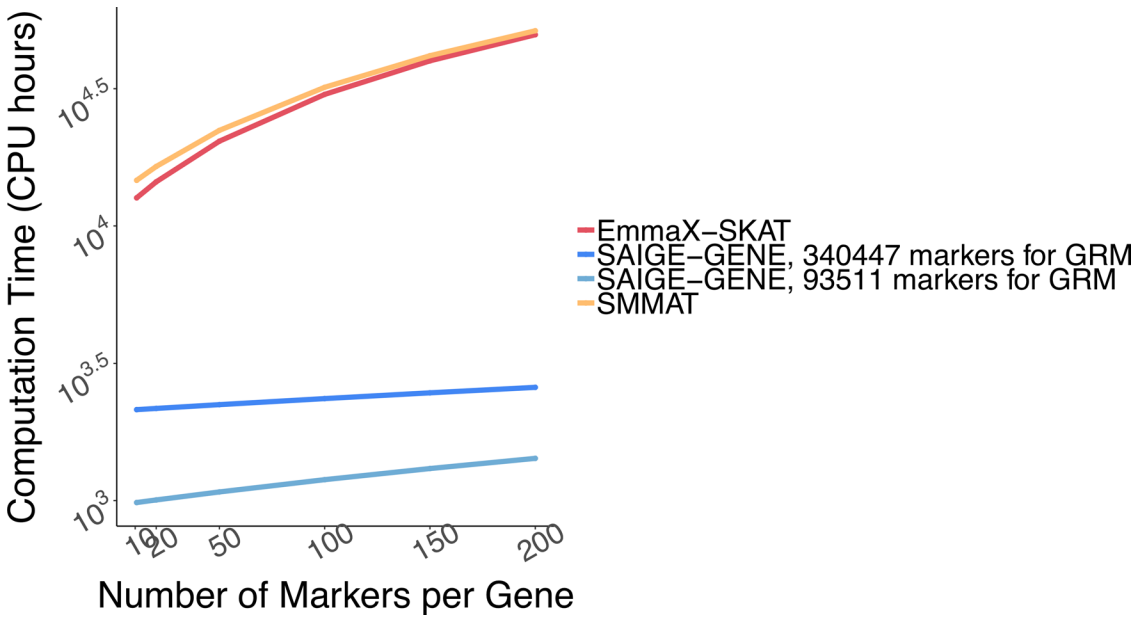
Extended Data Fig. 5 | Scatter plots of association *P*-values on the -log10 scale from SAIGE-GENE with two sample relatedness cutoffs for the sparse GRM, 0.125 and 0.2. 15,338 genes were tested for automated read pulse rate in white British samples in the HRC-imputed UK Biobank (*N* = 385,365). *N*, sample size. Missense and stop-gain variants with MAF ≤ 1% were included. **a**, Burden. **b**, SKAT. **c**, SKAT-O.



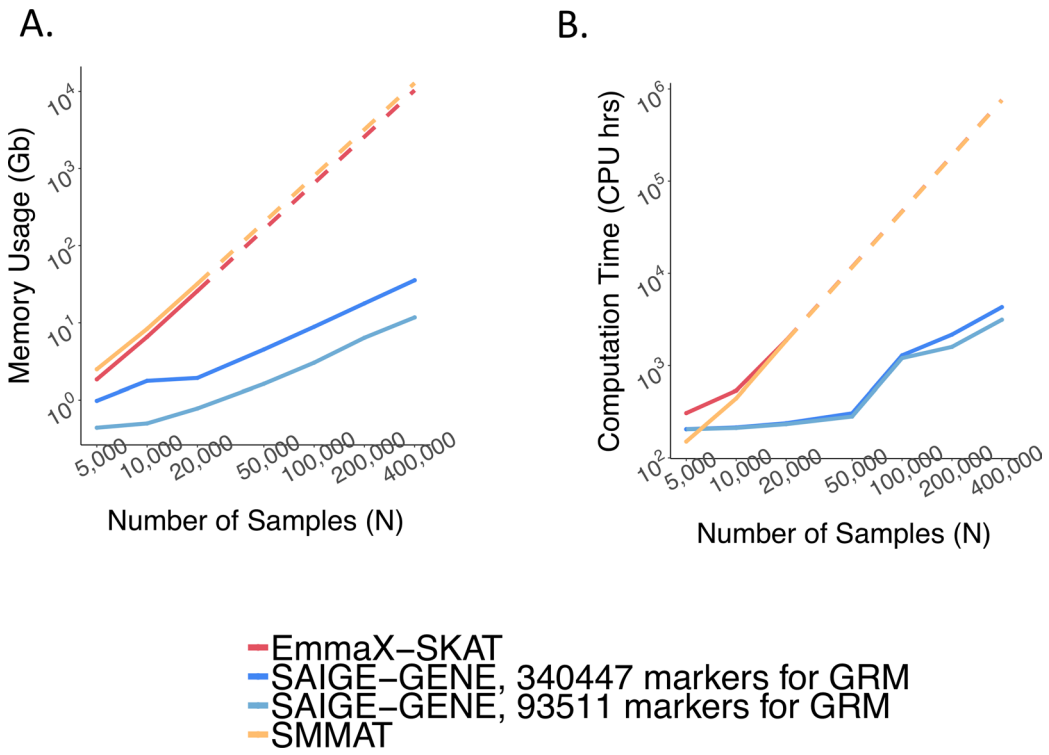
Extended Data Fig. 6 | Quantile-quantile plots of association *P*-values for 10 million variant sets from the simulation study for phenotypes with various case-control ratios (*N* = 100,000). a, Case:Control = 1:9. **b**, Case:Control = 1:19. **c**, Case:Control = 1:99. *N*, sample size.



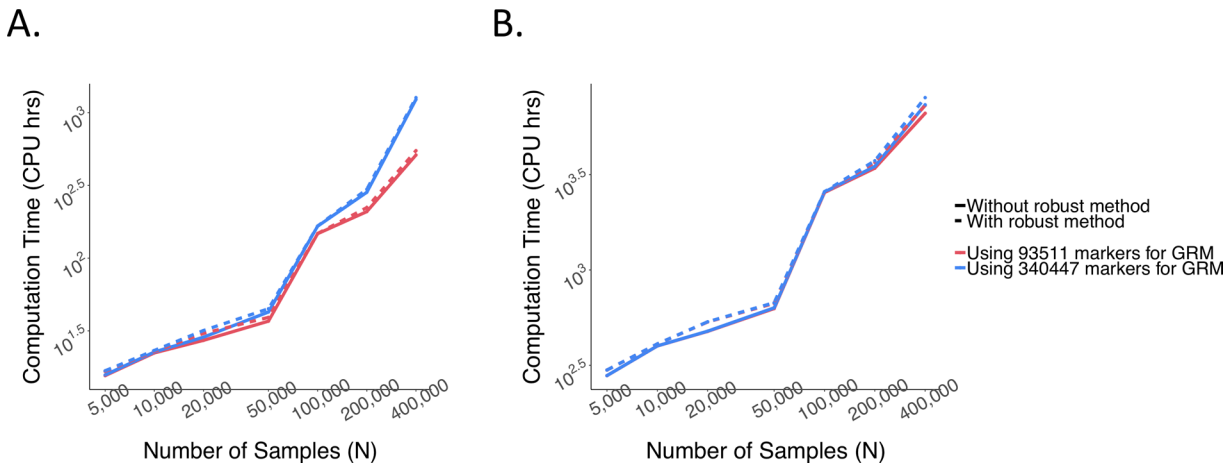
Extended Data Fig. 7 | Empirical computation time. a,**b**, Step 1 for fitting a null mixed model (**a**) and Step 2 for association tests (**b**), respectively, by sample sizes (*N*) for gene-based tests for 15,342 genes, each containing 50 rare variants. Benchmarking was performed on randomly sub-sampled UK Biobank data with 408,144 White British participants for waist-to-hip ratio. The reported run time was median of five runs with samples randomly selected from the full sample set using different sampling seeds. The reported computation time for EmmaX-SKAT and SMMAT was projected when $N > 20{,}000$. As the number of tested markers varies by sample sizes, the computation time was projected for 50 markers per gene for plotting. Numerical data are provided in Supplementary Table 1.



Extended Data Fig. 8 | Log-log plot of the estimated run time as a function of number of markers per gene. Benchmarking was performed on randomly sub-sampled 400,000 UK Biobank data with 408,144 white British participants for waist-to-hip ratio on 15,342 genes. The plotted run time was median of five runs with samples randomly selected from the full sample set using different sampling seeds. The computation time for other different number of markers per gene was projected based on the benchmarked time.



Extended Data Fig. 9 | Log-log plots of the estimated run time and memory usage as a function of sample size (*N*) for genome-wide tests for 286,000 chunks. a, Run time. **b**, Memory usage. Each chunk contains 50 variants on average, given that there are 14.3 million markers in the HRC-imputed UK Biobank with MAF ≤ 1% and imputation info score ≥ 0.8. Numerical data are provided in Supplementary Table 1. Benchmarking was performed on randomly sub-sampled UK Biobank data with 408,144 white British participants for waist-to-hip ratio. The plotted run time and memory were medians of five runs with samples randomly selected from the full sample set using different sampling seeds.



Extended Data Fig. 10 | Log-log plots of the estimated run time for as a function of sample size (*N*) for SAIGE-GENE with and without using the robust adjustment. a, Exome-wide gene-based tests for 15,871 genes. **b**, Genome-wide tests for 286,000 chunks. Each gene or chunk contains 50 variants on average. Benchmarking was performed on randomly sub-sampled UK Biobank data with 402,163 white British participants tested for glaucoma (PheCode: 365, 4,462 cases and 397,701 controls). The case-control ratio remained the same in subsampled data sets. The reported run time was median of five runs with samples randomly selected from the full sample set using different sampling seeds. As the number of tested markers varies by sample sizes, the computation time was projected for 50 markers per gene for plotting. Numerical data are provided in Supplementary Table 2.

nature research

Corresponding author(s): Seunggeun Lee

Last updated by author(s): Mar 22, 2020

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.

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 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 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	This research has been conducted using the UK Biobank Resource under application number 45227.
Data analysis	SAIGE-GENE (version 0.35.8.8), https://github.com/weizhouUMICH/SAIGE SMMAT (version 1.0.2), https://github.com/hanchenphd/GMMAT. EmmaX-SKAT (SKAT version_1.3.2.1), https://cran.r-project.org/web/packages/SKAT/index.html Minimac3, https://genome.sph.umich.edu/wiki/Minimac3

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. 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-GENE is implemented as an open-source R package available at https://github.com/weizhouUMICH/SAIGE
The SAIGE-GENE results for 53 quantitative phenotypes and 10 binary phenotypes in the UK Biobank can be download at https://www.leelabsg.org/resources.

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

Life sciences study design

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

Sample size	We analyzed publicly available UK Biobank data of samples with white British ancestry (sample size=408,910). For the study design, please refer UK Biobank(http://www.ukbiobank.ac.uk/). We also analyzed the population-based biobank, HUNT study, of Norwegian participants (sample size = 69,716)
Data exclusions	Due to QC issues in non-HRC imputed markers, we restricted our analysis to directly genotyped or HRC imputed markers with imputation score >= 0.8. Non White British samples were excluded from the analysis of the UK Biobank data. We restricted our analysis to directly genotyped or imputed markers with imputation score >= 0.8 in the HUNT study. The exclusion criteria were pre-established.
Replication	We searched published GWAS studies to check whether GWAS significant loci were known (replicated) or potentially novel.
Randomization	NA. We used publicly available UK Biobank data and the population-based HUNT study to illustrate the performance of the method.
Blinding	We used coded public data, and hence were blinded.

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

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

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	The UK Biobank study population (http://www.ukbiobank.ac.uk/) is residents of the UK aged 40-69 years at recruitment and living within a reasonable travelling distance of an assessment centre. Participants in the population-based Nord-Trøndelag Health (HUNT) study (https://www.ntnu.edu/hunt) are inhabitants of the county of Nord-Trøndelag, Norway.
Recruitment	The UK Biobank participants were selected using the NHS register and invited to volunteer for the study. Recruitment was carried out between 2007 and 2010. Full details of the recruitment process are available in reference (UK Biobank: Protocol for a large-scale prospective epidemiological resource, 2007). Every citizen of Nord-Trøndelag County in Norway being 20 years or older, have been invited to all the surveys for adults. Three phases of recruitment include HUNT1 (1984-86), HUNT2 (1995-97) and HUNT3 (2006-08). Full details of the recruitment process are available in reference (Cohort Profile: the HUNT Study, Norway., 2013)
Ethics oversight	National Research Ethics Service Committee (UK Biobank) and Regional Ethics Committee (HUNT)

Note that full information on the approval of the study protocol must also be provided in the manuscript.