Supplementary Materials

1. Algorithm details

1.1 Step 1. Fitting the logistic mixed model under the null hypothesis

1.1.1 Generalized linear mixed model and penalized quasi-likelihood

Details of fitting the null logistic mixed model and estimating the parameters for fixed effects and variance components are provided in this section. Note that although we use the same restricted log likelihood and average information matrix as in GMMAT[1], we use a different approach to estimate parameters to make our method feasible for very large datasets. In particular, we use the preconditioned conjugate gradient method[2,3] to solve linear systems instead of obtaining an inverse of the covariance matrix of the phenotypes. For the derivation of the likelihood and information matrix, please refer the GMMAT paper[1]. Logistic mixed model is a part of the larger generalized linear mixed model (GLMM) with the logistic link function for binary outcome. The model can be written as

$$logit(\mu_i) = X_i\alpha + G_i\beta + b_{i,}$$

where $\mu_i = P(y_i = 1 \mid X_i, G_i, b_i)$ is the probability for the *ith* individual being a case given the covariates X_i and genotypes G_i as well as the random effect b_i, assumed to be distributed as $N(0, \tau\,\psi)$, where ψ is an $N \times N$ genetic relationship matrix (GRM)[33] and τ is an additive genetic variance. The phenotype y_i is assumed to be conditionally independent given (X_i, G_i, b_i) and follows the binomial distribution with mean $E(y_i \mid b_i) = \pi_i$ and variance $Var(y_i \mid b) = \phi v(\pi_i)$, where $v(\pi_i) = \mu_i(1-\mu_i)$ is the variance function, and the dispersion parameter $\phi = 1$.

Under the null hypothesis that $H_0: \beta = 0$, to estimate (α, ϕ, τ), the log integrated quasi-likelihood function can be written as

$$ql(\alpha, \beta = 0, \phi, \tau) = \log \int exp\{\textstyle\sum_{i=1}^{n} ql_i(\alpha, \beta = 0 \mid b)\} \times (2\pi)^{-\frac{n}{2}} |\tau\,\psi|^{-\frac{1}{2}} \times exp\left\{-\frac{1}{2} b^T (\tau\,\psi)^{-1} b\right\} db, \quad (1)$$

where $ql_{\mathbb{I}}(\alpha, \beta = 0 \mid b) = \int_{y_i}^{\mu_i} \frac{a_i(y_i - \mu)}{\phi v(\mu)} d\mu$ is the quasi-likelihood for the *ith* individual given the random effect b. Let $\kappa(b) = \sum_{i=1}^{n} ql_i(\alpha, \beta = 0 \mid b) - \frac{1}{2} b^T (\tau\,\psi)^{-1}$.Approximation for the integral $\int exp\{\kappa(b)\} db$ can be obtained using Laplace's method with the first and second derivatives. Let $\tilde{b}$ denote the solution of $\kappa'(b) = 0$, which maximizes $\kappa(b)$, and W denote the weight matrix, which is a diagonal matrix with diagonal terms $\frac{1}{\phi v(\mu_i)[g'(\mu_i)]^2}$. Note that since logistic is a canonical link function, the diagonal element of W can be simplified as $v(\mu_i)$. Equation (1) can be written as

$$ql(\alpha, \beta = 0, \phi, \tau) = \kappa\left(\tilde{b}\right) - \frac{1}{2} log|\tau\,\psi\, W + I| \quad (2)$$

1.1.2 Estimate parameters using AI-REML

Here we describe iterative steps to estimate (α, b, ϕ, τ). To obtain the estimates of the fixed effect coefficients and the random effects given (ϕ, τ), $(\hat{\alpha}(\phi, \tau), \hat{b}(\phi, \tau))$, that jointly maximize the $ql(\alpha, \beta = 0, \phi, \tau)$, we take the derivative of equation (2) with respect to α and b and get the solution for the derivatives to be zero. Assuming the weight matrix W varies slowly as a function of the conditional mean, the last term in the expression of $ql(\alpha, \beta = 0, \phi, \tau)$ in equation (2) can be ignored. Let $\Sigma = W^{-1} + \tau\psi$, $P = \Sigma^{-1} - \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$ and $\tilde{Y}$ be a working vector with the ith element being $X_i\alpha + b_i + g'(\mu_i)(y_i - \mu_i)$, and then

$$\hat{\alpha} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\tilde{Y} \quad (3)$$

$$\hat{b} = \tau\,\psi\Sigma^{-1}(\tilde{Y} - X\hat{\alpha}) \quad (4)$$

Given $\hat{\alpha}$ and $\hat{b}$ estimated,

$$ql(\hat{\alpha}(\phi,\tau),\beta=0,\phi,\tau) = c - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\tilde{Y}^T P\tilde{Y} \quad (5)$$

The restricted maximum likelihood (REML) version:

$$ql_R(\hat{\alpha}(\phi,\tau),\beta=0,\phi,\tau) = c_R - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\log|X^T\Sigma^{-1}X| - \frac{1}{2}\tilde{Y}^T P\tilde{Y} \quad (6)$$

To obtain the estimates of the variance components, (ϕ,τ), that jointly maximize the $ql(\hat{\alpha}(\phi,\tau),\beta=0,\phi,\tau)$, we take the derivative of equation (6) with respect to ϕ and τ:

$$\frac{\partial ql_R(\hat{\alpha}(\phi,\tau),\beta=0,\phi,\tau)}{\partial\phi} = \frac{1}{2\phi}\tilde{Y}^T PW^{-1}P\tilde{Y} - \frac{1}{\phi}tr(PW^{-1}) \quad (7)$$

$$\frac{\partial ql_R(\hat{\alpha}(\phi,\tau),\beta=0,\phi,\tau)}{\partial\tau} = \frac{1}{2}\tilde{Y}^T P\psi P\tilde{Y} - tr(P\psi) \quad (8)$$

To obtain the solutions to make equations (7) and (8) equal to zero, an average information matrix AI is then defined and used in place of the expected information matrix to estimate $\hat{\phi}$ and $\hat{\tau}$ iteratively[1].
Note that for the logistic mixed model, $\phi = 1$, so we do not need to obtain (7).

1.1.3 Approaches to reduce computation and memory cost.

Preconditioned Conjugate Gradient (PCG): To obtain equations (3)-(8), we need to compute expression forms containing a product of Σ^{-1} and a vector or a matrix, such as $\Sigma^{-1}X$, which is very challenging for large cohorts. Computing the $N\times N$ empirical genetic relationship matrix (GRM) $\psi = \frac{G_c^T G_c}{M_1}$ costs $O(M_1N^2)$, where G_c is an $M_1\times N$ matrix with genotypes for M_1 genetic markers of N individuals that are normalized with the means and standard deviations of raw genotypes. Moreover, the Cholesky decomposition used by GMMAT[1] to invert Σ takes $O(N^3)$ computation and very large memory space, which are not practical for studies with large sample sizes ($N > 20{,}000$).
Similar to BOLT-LMM[4], we use two strategies to reduce the computation and memory cost. First, instead of requiring the pre-computed GRM ψ as an input, we store genotypes for computing GRM in a binary vector and calculate elements of Σ as needed, which reduces the memory usage from $4N(N+1)$ bytes, given double precision floating number is used to store ψ, to $\frac{NM_1}{4}$ bytes. For instance, with N =408,961 white British participants and M_1 = 93,511 markers, the memory usage drops from 669 Gb to 9.56 Gb with this strategy. Second, the conjugate gradient method is used to calculate the product of Σ^{-1} and a vector by iteratively solving the linear system $Ax = u$, where $A = \Sigma$ and u is a known vector, such as any column vector in X matrix. The number of iterations required for convergence of the conjugate gradient algorithm is proportional to $\sqrt{\kappa(A)}$, where $\kappa(A)$ is the condition number for A[5]. To make the convergence faster, a preconditioner matrix Q is used so that $\hat{A}$=$Q^{-1}A$ and $\kappa(\hat{A}) < \kappa(A)$. Here, Q is an $N\times N$ diagonal matrix with the diagonal elements of Σ and the calculation of Q requires $O(NM_1)$.

Randomized trace estimator for $tr(PW^{-1})$ and $tr(P\psi)$: The computation of (7) and (8) requires the traces of matrices PW^{-1} and $P\psi$. For this, we use Hutchinson's randomized trace estimator[6,7]. The trace of a matrix B, such as PW^{-1} and $P\psi$, is estimated by $\frac{1}{R}\sum_{i=1}^{R} z_i^T Bz_i$, where z_i's are R independent random vectors whose entries are i.i.d Rademacher random variables ($P(z_i = \pm 1) = 0.5$). A vector z_i with size N is randomly drawn from the Rademacher distribution, followed by the calculation for $z_i^T Bz_i$. This procedure is repeated for R times and the average of the results for $z_i^T Bz_i$ is the estimate for the trace of the B matrix. The by default value for R is set to be 30.
Parallel computation for the vector multiplication: The most time-consuming step of the proposed algorithm is performing PCG, which involves computing a product of the GRM ψ and a vector x, i.e. $\psi x = G_c^T G_c x$. We use parallel computing techniques to speed up this procedure. In particular, we use Intel Threading Building Block (TBB) implemented in RcppParallel package[8] for the multi-threading

computation. Our approach utilized nearly all CPU cores allocated. For example, the CPU usages on average were 14.6 when 16 CPU cores were allocated.

1.2 Step 2. Single variant score tests with SPA

1.2.1 Score tests based on logistic mixed model

Given the estimates from step 1 for fixed effect coefficients $\hat{\alpha}$, random effects $\hat{b}$, and the variance component parameters $\hat{\phi}$ and $\hat{\tau}$ under the null hypothesis H_o: $\beta = 0$, the score test can be constructed for each genetic marker to be tested. Suppose G is the N×1 genotype vector, $\hat{\mu}$ is estimate for $P(Y = 1 \mid X, \hat{b})$, are the probabilities for study individuals being a case given the covariates X and the estimated random effect $\hat{b}$ from step 1, $\widehat{W}$ is a diagonal vector with diagonal elements $\hat{\mu}(1-\hat{\mu})$, and $\tilde{G} = G - X(X^T\widehat{W}X)^{-1}X^T\widehat{W}G$ is the covariate adjusted genotype vector with covariate effects projected out from the raw genotypes[9]. Suppose $\hat{\Sigma} = \widehat{W}^{-1} + \hat{\tau}\psi$ and $\hat{P} = \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}X\left(X^T\hat{\Sigma}^{-1}X\right)^{-1}X^T\hat{\Sigma}^{-1}$, and then $\hat{P}G = \hat{P}\tilde{G}$. The score test statistics can be written as

$$T = G^T(Y - \hat{\mu}) = G^T\hat{P}\tilde{Y} = \tilde{G}^T\hat{P}\tilde{Y} = \tilde{G}^T(Y - \hat{\mu}),$$

where $\tilde{Y}$ is the working vector previously defined. The variance of T, Var(T) $= G^T\hat{P}G = \tilde{G}^T\hat{P}\tilde{G}$.

1.2.2. Estimation of Var(*T*)

Calculating $\hat{P}\tilde{G}$ is required for the estimation of Var(T), which is computationally expensive. To avoid to calculate $\hat{P}\tilde{G}$ to all the variants, we use similar approximation approaches used in BOTL-LMM[4] and GRAMMAR-GAMMAR[10] in which we obtain the ratio between $\mathrm{Var}(T)$ and $\mathrm{Var}(T)^* = \tilde{G}^T W\tilde{G}$ using a small number of variants, and estimate variant as $r\mathrm{Var}(T)^*$, where $r = \mathrm{Var}(T)$ / $\mathrm{Var}(T)^*$. Note that $\mathrm{Var}(T)^*$ is a variance estimator without accounting the fact that the random effect *b* is estimated from data, and the calculation of $\mathrm{Var}(T)^*$ only requires O(*N*) computation.

Here we show that the ratio *r* is approximately constant across all variants. For this, we assume that $\frac{w_i}{\sum_{j=1}^{N} w_j} = o(1)$, for all i=1, …, N, where w_i is the i^{th} element of *W*. Note that this assumption can only be violated when the covariates are extremely sparse, which rarely happens in real data. First, $Var(T)$ can be written as

$$Var(T) = \tilde{G}^T P\tilde{G} = \tilde{G}^T\hat{\Sigma}^{-1}\tilde{G} - \tilde{G}^T\hat{\Sigma}^{-1}X\left(X^T\hat{\Sigma}^{-1}X\right)^{-1}X^T\hat{\Sigma}^{-1}\tilde{G} \qquad (3)$$

Since $\tilde{G}$ is adjusted by covariates, it can be shown that $\left(X^T\hat{\Sigma}^{-1}X\right)^{-1}X^T\hat{\Sigma}^{-1}\tilde{G} = O_P\left(N^{-\frac{1}{2}}\right)$, and hence $\tilde{G}^T\hat{\Sigma}^{-1}X\left(X^T\hat{\Sigma}^{-1}X\right)^{-1}X^T\hat{\Sigma}^{-1}\tilde{G} = O_p(1)$. The first term in (3) is $\tilde{G}^T\hat{\Sigma}^{-1}\tilde{G} = O_p(N)$, so (3) can be approximated by $\tilde{G}^T\hat{\Sigma}^{-1}\tilde{G}$. Let $\overline{w}$ be the mean of the diagonal element of $\widehat{W}^{-1}$ and $\xi = \widehat{W}^{-1} - \overline{w}I$. And then

$$\tilde{G}^T\hat{\Sigma}^{-1}\tilde{G} \approx \tilde{G}^T(\overline{w}I + \hat{\tau}\psi)^{-1}\tilde{G} - G^T(\overline{w}I + \hat{\tau}\psi)^{-1}\xi(\overline{w}I + \hat{\tau}\psi)^{-1}G \quad (4)$$

With the assumption $\frac{w_i}{\sum_{j=1}^{N} w_j} = o(1)$, it can be shown $\frac{G^T(\overline{w}I+\hat{\tau}\psi)^{-1}\xi(\overline{w}I+\hat{\tau}\psi)^{-1}G}{\tilde{G}^T(\overline{w}I+\hat{\tau}\psi)^{-1}\tilde{G}} = o_p(1)$, therefore (4) can be approximated by the first term, in which

$$(4) \approx \tilde{G}^T(\overline{w}I + \hat{\tau}\psi)^{-1}\tilde{G} = \tilde{G}^T\psi^{-\frac{1}{2}}\psi^{\frac{1}{2}}(\overline{w}I + \hat{\tau}\psi)^{-1}\psi^{\frac{1}{2}}\psi^{-\frac{1}{2}}\tilde{G} = a^T U\Lambda^{\frac{1}{2}}(\overline{w}I + \hat{\tau}\Lambda)^{-1}\Lambda^{\frac{1}{2}}Ua \quad (5)$$

where U and Λ are eigenvector and eigenvalue matrices of ψ, and $a = \psi^{-\frac{1}{2}}\tilde{G}$. Since correlation matrix of a is an identity matrix, asymptotically, (4) is closely approximated by the trace of $cU\Lambda^{\frac{1}{2}}(\overline{w}I + \hat{\tau}\Lambda)^{-1}\Lambda^{\frac{1}{2}}U$,

which is $c\sum_{i=1}^{n}\lambda_i/(\overline{w}+\hat{\tau}\lambda_i)$, where c=*MAF*(1-*MAF*). As the same way, $\mathrm{Var}(T)^* = \tilde{G}^T\widehat{W}\tilde{G} \approx c\sum_{i=1}^{n}\lambda_i/\overline{w}$. And hence the ratio is

$$r = \frac{\mathrm{Var}(T)}{\mathrm{Var}(T)^*} \approx \frac{\sum_{i=1}^{n}\frac{\lambda_i}{\overline{w}+\hat{\tau}\lambda_i}}{\sum_{i=1}^{n}\frac{\lambda_i}{\overline{w}}}$$

which is constant across all variants. The variance adjusted score test statistic is

$$T_{adj} = \left(\hat{r}\tilde{G}^T\widehat{W}\tilde{G}\right)^{-1/2}\tilde{G}^T(Y-\hat{\mu})$$

,where $\hat{r}$ is the estimated *r*. In simulation and real data analysis, we used 100 variants to estimate *r* . The adjusted test statistic, T_{adj}, has mean zero and variance is approximately unity.

Supplementary Figure 1 shows the ratio *r* by minor allele counts (MAC) from 1000 simulated markers. The ratio was nearly identical for markers with MAC > 20 and then variation was increased for extremely rare variants. This figure provides empirical evidence that the equal ratio assumption holds.

1.2.3 P-value calculation using SPA

The traditional score test, such as GMMAT, used the fact that the score test statistic asymptotically follows a normal distribution under the null hypothesis of no association. When the case-control ratios are unbalanced and MAC is small, this asymptotic result does not hold and type I error rates can be inflated. To obtain more accurate p-value, we use a fast-version of SPA (fastSPA)[9], which we have previously developed for logistic regression model. For this, we utilize the fact that phenotype Y_i independently follows Bernoulli distribution given π_i, and T_{adj} is a weighted sum of independent Bernoulli random variable. The approximated cumulant generating function (CGF) of T_{adj} is

$$K(t;\,\hat{\pi},c) = \sum_{i=1}^{N}\log\left(1-\hat{\pi}_i+\hat{\pi}_i e^{ct\tilde{G}_i}\right) - ct\sum_{i=1}^{N}\tilde{G}_i\,\hat{\pi}_i$$

where the constant c=Var*(T)$^{-1/2}$, which provide K'(0)=0 and K''(0) = 1, where K' and K'' are first and second derivate of K with respect to t. Note that since K uses $\hat{\pi}$, which is estimated from data, it is an approximation of the true CGF. Now we use the saddle point method to estimate the p-value. To calculate the probability that $T_{adj} < q$, where q is an observed test statistic, we use the following formula[31 35 36].

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

,where $w = sign(\hat{\zeta})\left[2\left\{\hat{\zeta}q - K(\hat{\zeta})\right\}\right]^{\frac{1}{2}}$, $v = \hat{\zeta}\left\{K''(\hat{\zeta})\right\}^{\frac{1}{2}}$ and $\hat{\zeta} = \hat{\zeta}(q)$ is the solution of the equation $K'(\hat{\zeta}) = q$. As the fastSPA[9], we exploit the sparsity of genotype vector when MAF of variants are low. In addition, since normal approximation performs well when the test statistic is close to the mean, we use normal distribution when the test statistic is within two standard deviations of the mean.

1.2.4 Effect size estimation

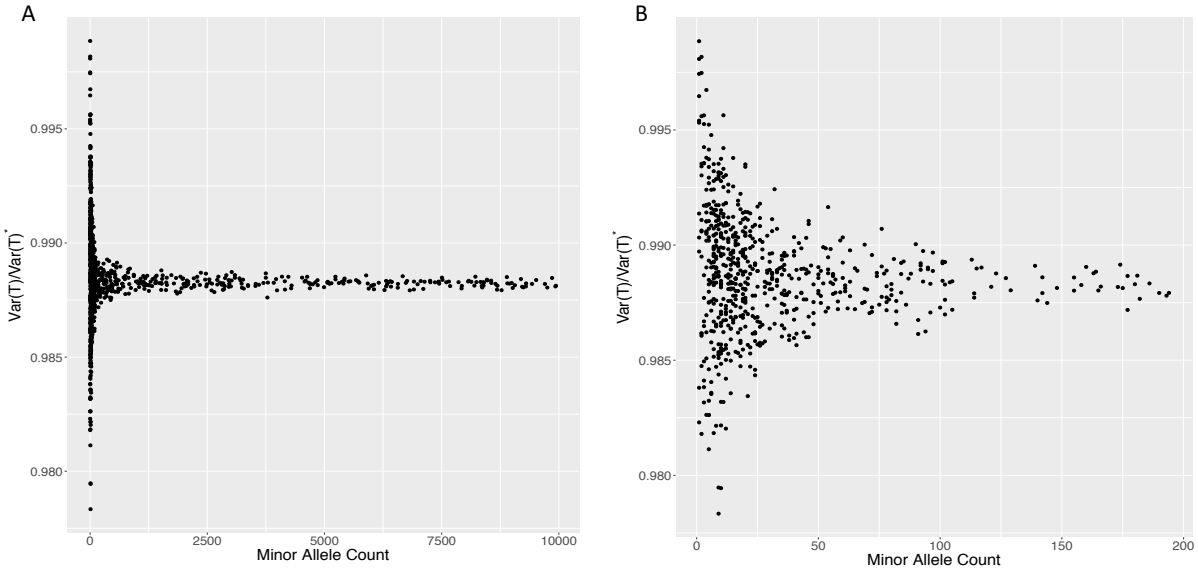
To rapidly estimate the effect size $\hat{\beta}$, which equals to the natural logarithm of the odds ratio, we use the variance component estimate under the null hypothesis. Note that a similar approach has been used in EMMAX[11] and GRAMMAR-Gamma[10]. Our $\hat{\beta}$ estimate is

$$\hat{\beta} = \left(\tilde{G}^T\hat{P}\tilde{G}\right)^{-1}\tilde{G}^T\hat{P}\tilde{Y}$$

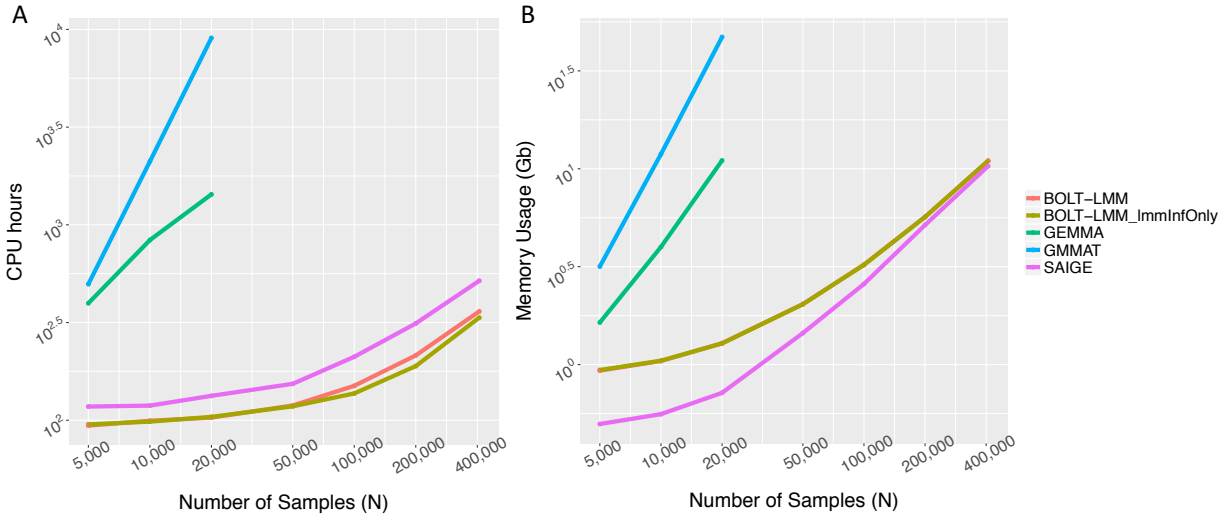
Since $T = \tilde{G}^T\hat{P}\tilde{Y}$ and $\mathrm{Var}(T) = \tilde{G}^T\hat{P}\tilde{G}$, $\hat{\beta}$ can be written as $T/\mathrm{Var}(T)$. In the section 1.2.2, we have shown that $\mathrm{Var}(T)$=$\hat{r}\mathrm{Var}(T)^* = \hat{r}\tilde{G}^T\widehat{W}\tilde{G}$. Therefore, $\hat{\beta}$ can estimated using $T, \mathrm{Var}(T)^*$, and $\hat{r}$, which have already been calculated for association p-value estimation. To estimate the standard error and confidence interval, we use p-values. The standard error of $\hat{\beta}$, $SE(\hat{\beta}) = \left|\hat{\beta}/z\right|$, where z-score corresponds to the association p-value/2.

2. Supplementary figures

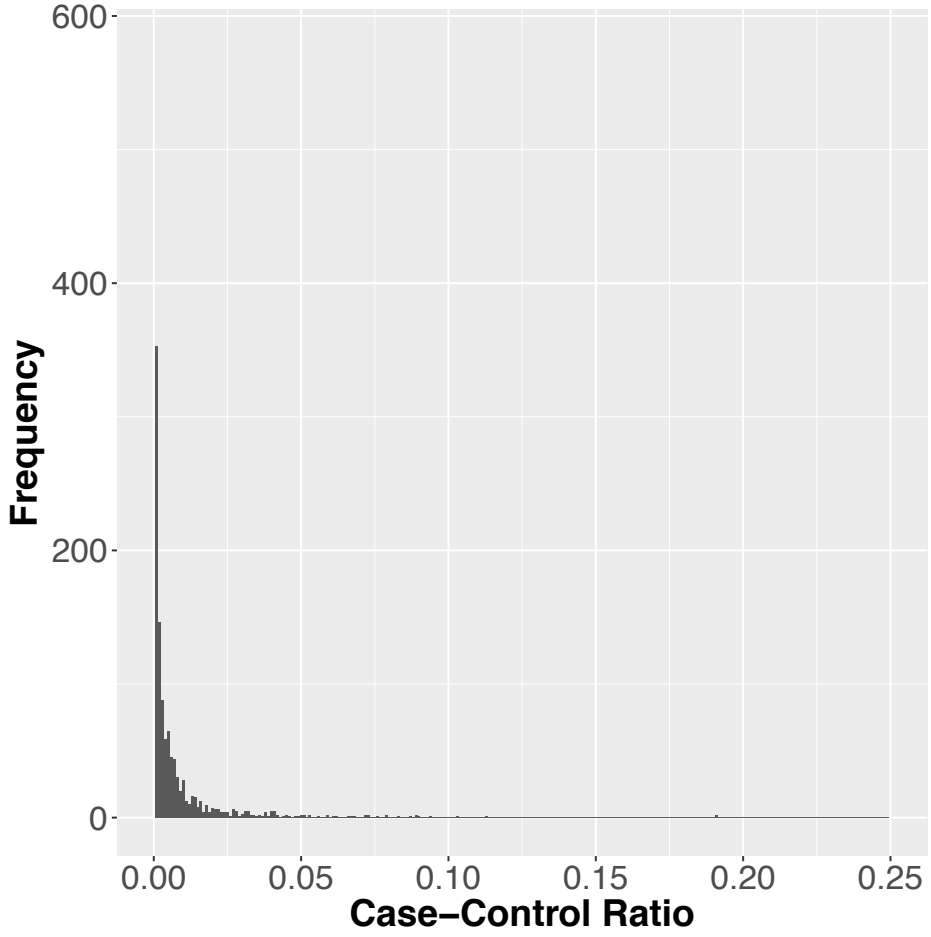
Supplementary Figure 1. Plot of the ratio of the variances of the score statistics with and without incorporating the variance components for the random effects for A. 1,000 simulated markers with MAF spectrum shown in **Supplementary Figure 8** and B. 669 out of 1,000 markers that have MAC < 200. 1,000 families were simulated based on the pedigree structure shown in **Supplementary Figure 4** with case control ratio 1:9.



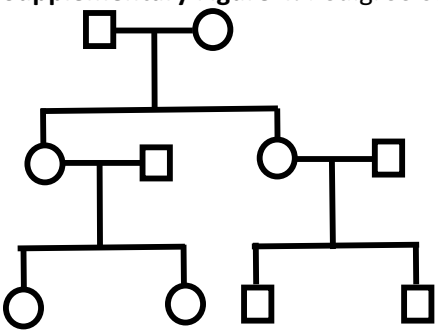
Supplementary Figure 2. Log-log plots of the estimated run time (A) and memory use (B) as a function of sample size (N). Numerical data are provided in **Supplementary Table 1.** Benchmarking was performed on randomly sub-sampled UK Biobank data with 408,458 white British participants and 200,000 markers for the cardiovascular diseases (PheCode = 411). The plotted run time is the projected computation time for testing 71 million markers with info ≥ 0.3. The reported run times are medians of five runs with samples randomly selected from the full sample set using different sampling seeds. Software versions: BOLT-LMM, v2.3; GEMMA, v0.96. BOLT-LMM: compute association statistics under the non-infinitesimal model; BOLT-LMM_lmmInfOnly: compute mixed model association statistics under the infinitesimal model



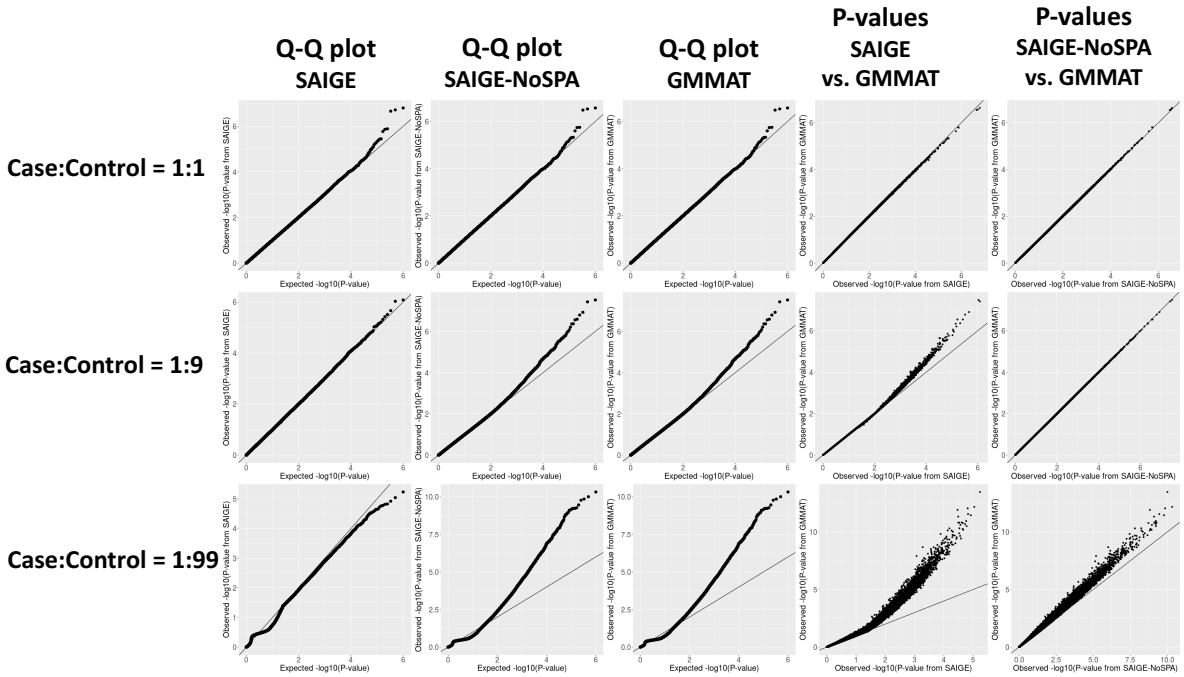
Supplementary Figure 3. Histogram of case-control ratios of 1,663 disease-specific binary phenotypes in the UK Biobank data. Phenotypes were constructed based on ICD-9 and ICD-10 codes using a previously described scheme[12].



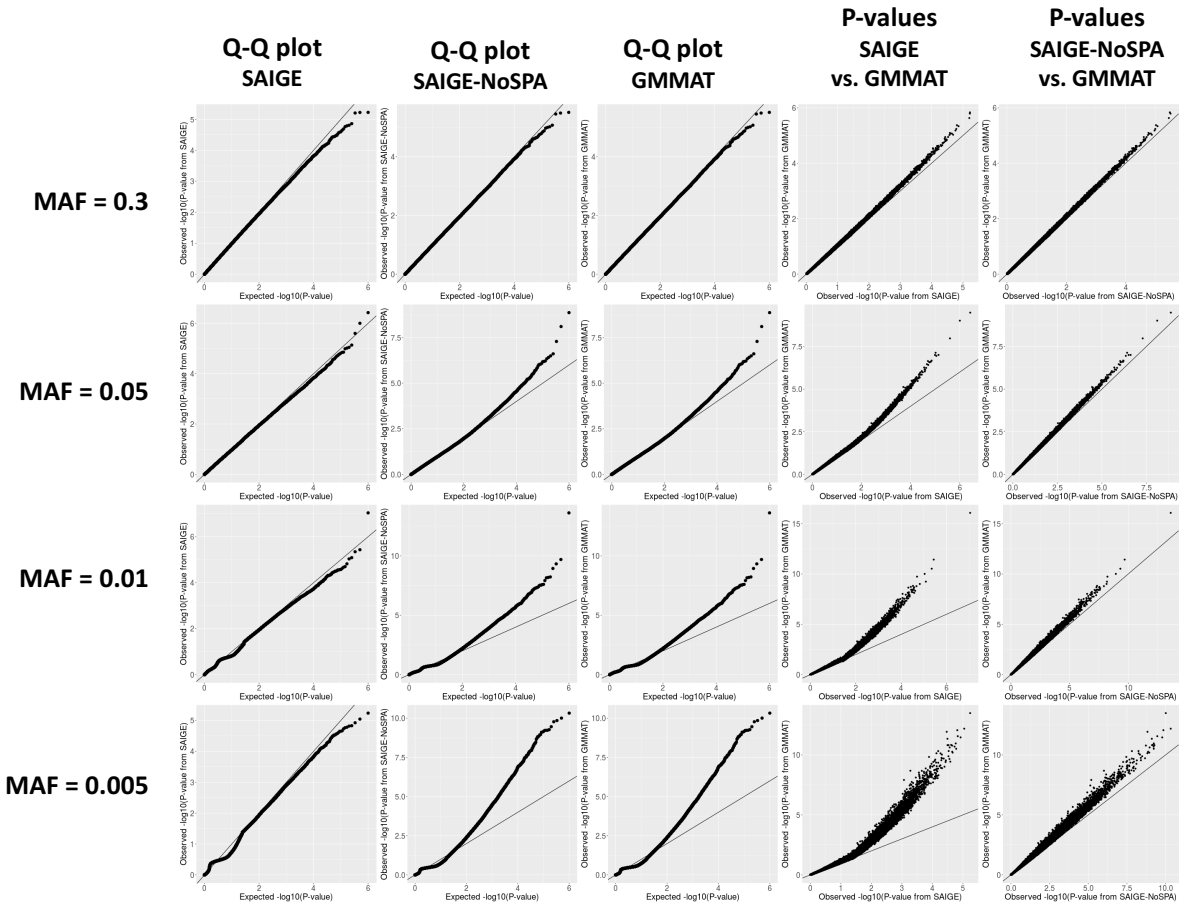
Supplementary Figure 4. Pedigree of families, each with 10 members, in the simulation study.



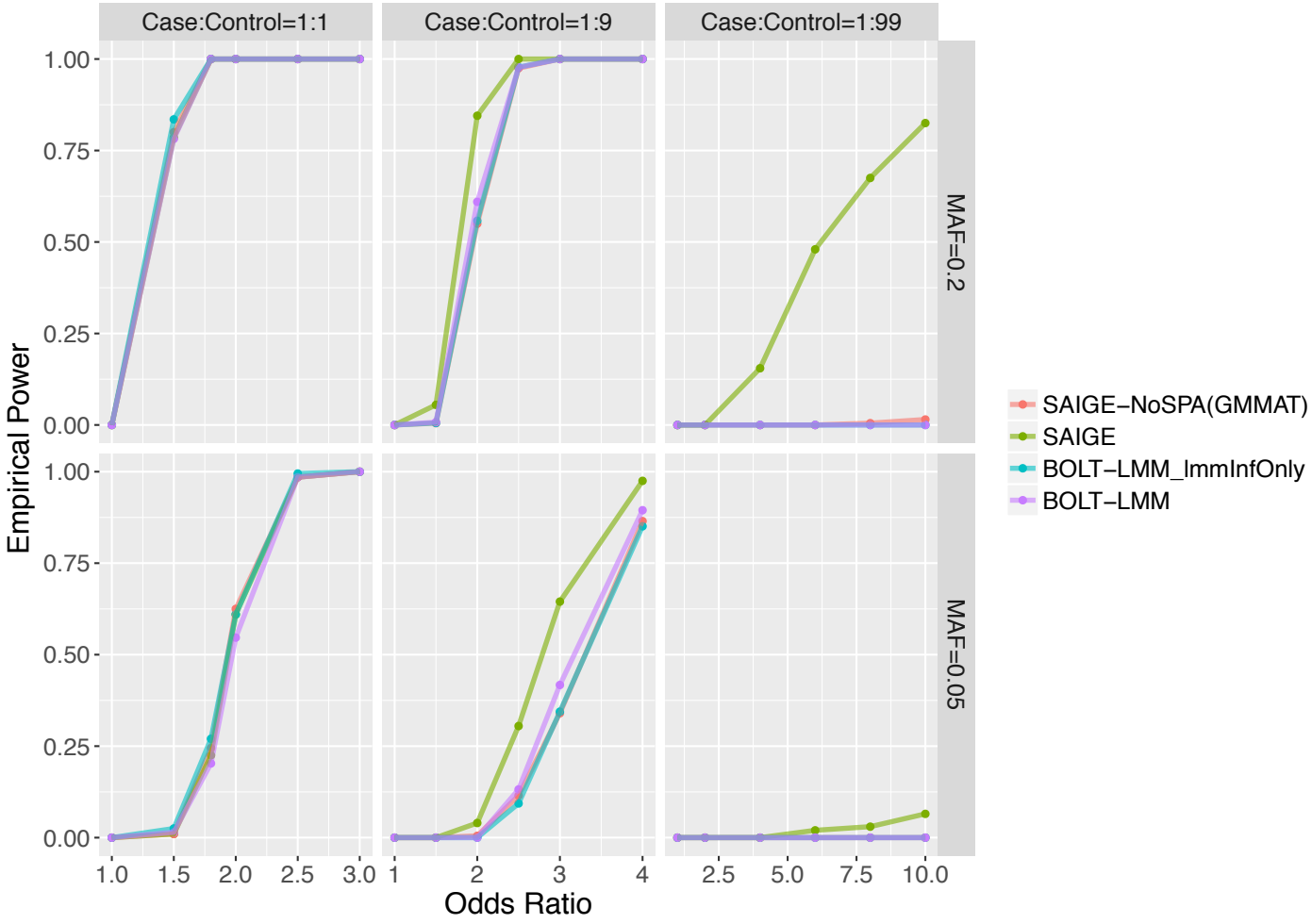
Supplementary Figure 5. Quantile-quantile plots of association p-values for 1,000,000 variants having MAF = 0.005 from the simulation study. The first column is p-values from SAIGE. The second column is for p-values from SAIGE-NoSPA. The third column is for p-values from the GMMAT[1] program. The fourth column is comparing the p-values from SAIGE and from GMMAT[1]. The fifth column is comparing the p-values from SAIGE-NoSPA and from GMMAT[1]. The black lines indicate x = y.



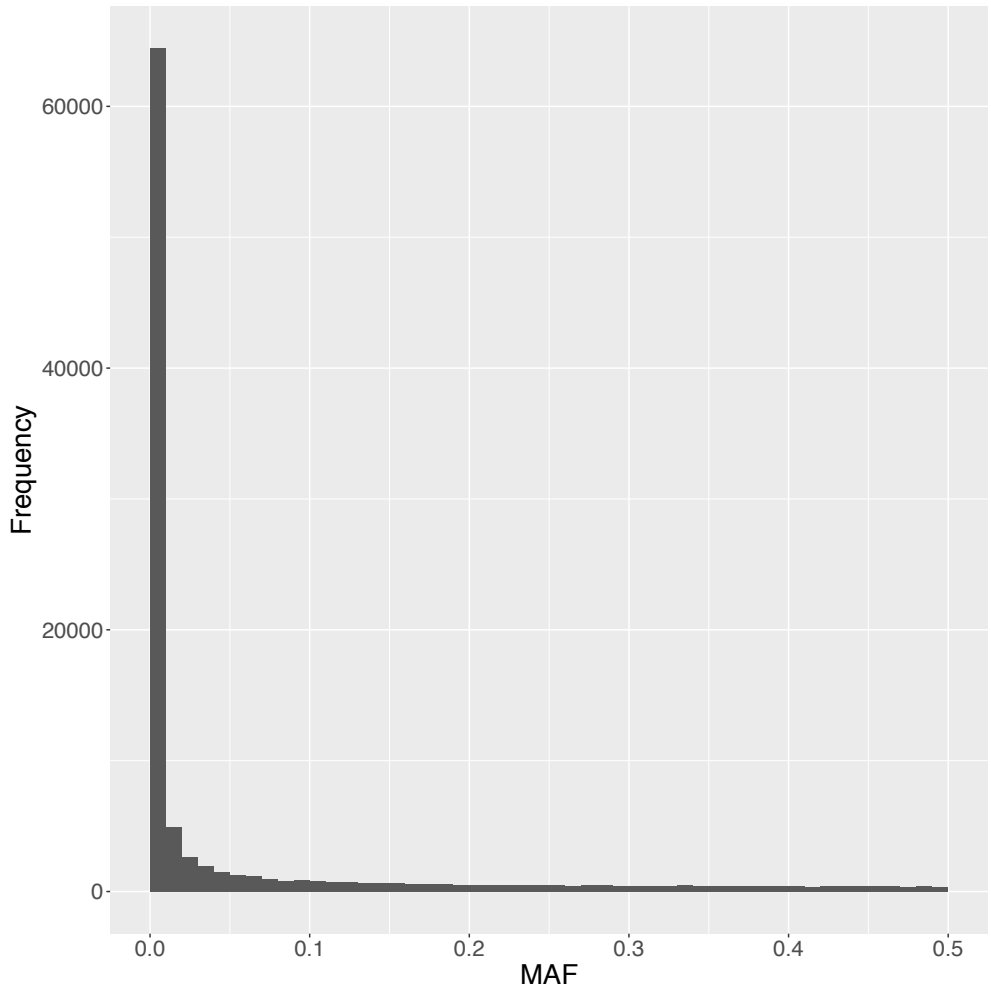
Supplementary Figure 6. Quantile-quantile plots of association p-values for 1000,000 variants with 10,000 samples with very unbalanced case-control ratio 1:99 from the simulation study. The first column is p-values from SAIGE. The second column is for p-values from SAIGE-NoSPA. The third column is for p -values from the GMMAT[1] program. The fourth column is comparing the p-values from SAIGE and from GMMAT[1]. The fifth column is comparing the p-values from SAIGE-NoSPA and from GMMAT[1]. The black lines indicate x = y.



Supplementary Figure 7. Empirical power of SAIGE, SAIGE-NoSPA (asymptotically equivalent to GMMAT), BOLT-LMM_lmmInfOnly (compute mixed model association statistics under the infinitesimal model), and BOLT-LMM (compute mixed model association statistics under the non-infinitesimal model) at the test-specific empirical α levels that yield type I error rate $\alpha = 5x10^{-8}$



Supplementary Figure 8. Distribution of the minor allele frequency spectrum for randomly selected 1,000,000 markers in the simulation study.



3. Supplementary tables

Supplementary Table 1. The estimated run time (A) and memory use (B) across different sample sizes. Benchmarking was performed on randomly sub-sampled UK Biobank data with 408,458 white British participants and 200,000 markers for the cardiovascular diseases (PheCode = 411). The plotted run time is the projected computation time for testing 71 million markers with info ≥ 0.3. The reported run times are medians of five runs with samples randomly selected from the full sample set using different sampling seeds. Software versions: BOLT-LMM, v2.3; GEMMA, v0.96. BOLT-LMM: compute non-infinitesimal association statistics; BOLT-LMM_lmmInfOnly: compute mixed model association statistics under the infinitesimal model

Sample Size(N)	Time (CPU hours)	Memory(Gb)	Tests
5,000	497.19	3.17	GMMAT
10,000	2109.83	11.88	GMMAT
20,000	9046.04	47.09	GMMAT
5,000	95.18	0.94	BOLT-LMM_lmmInfOnly
10,000	98.40	1.05	BOLT-LMM_lmmInfOnly
20,000	104.05	1.28	BOLT-LMM_lmmInfOnly
50,000	117.80	2.03	BOLT-LMM_lmmInfOnly
100,000	137.16	3.23	BOLT-LMM_lmmInfOnly
200,000	189.28	5.67	BOLT-LMM_lmmInfOnly
408,458	335.00	10.98	BOLT-LMM_lmmInfOnly
5,000	93.89	0.93	BOLT-LMM
10,000	99.39	1.04	BOLT-LMM
20,000	103.15	1.29	BOLT-LMM
50,000	119.03	2.04	BOLT-LMM
100,000	150.02	3.24	BOLT-LMM
200,000	214.71	5.69	BOLT-LMM
408,458	360.63	10.98	BOLT-LMM
5,000	397.00	1.64	GEMMA
10,000	835.59	3.99	GEMMA
20,000	1431.69	11.03	GEMMA
5,000	117.50	0.50	SAIGE
10,000	118.83	0.56	SAIGE
20,000	133.32	0.72	SAIGE
50,000	153.60	1.45	SAIGE
100,000	211.21	2.58	SAIGE
200,000	312.81	5.16	SAIGE
408,458	517.38	10.32	SAIGE

Supplementary Table 2. Number of genetic variants and loci that passed the genome-wide significant threshold ($P < 5x10^{-8}$) for the three 'real data' phenotypes identified by SAIGE, SAIGE-NoSPA(asymptotically equivalent to GMMAT), and BOLT-LMM in the UK Biobank data. Since results from SAIGE-NoSPA and BOLT-LMM contain many false positive signals for colorectal cancer and thyroid cancer, the numbers of loci are not provided.

Phenotype	Tests	Number of variants with p-value < $5x10^{-8}$	Number of all loci with top p-value < $5x10^{-8}$	Number of all loci with top p-value < $5x10^{-8}$ and have not been previously reported
Cardiovascular diseases PheCode 411 case:control 1:12	SAIGE	1,733	40	6
	SAIGE-NoSPA	1,820	101	68
	BOLT-LMM	1,886	89	58
Colorectal cancer PheCode 153 case:control 1:84	SAIGE	77	3	3
	SAIGE-NoSPA	2,950	NA	NA
	BOLT-LMM	3,349	NA	NA
Thyroid cancer PheCode 193 case:control=1:1138	SAIGE	125	1	1
	SAIGE-NoSPA	73,382	NA	NA
	BOLT-LMM	79,269	NA	NA

Supplementary Table 3. Loci that passed the genome-wide significant threshold ($P < 5x10^{-8}$) for the three phenotypes identified by the SAIGE in the UK Biobank data.

Phenotype	Location	Chr:Pos	rsID	Ref	Alt	Function	Gene	MAF	Sample Size	P value	Known for CAD	Previous Findings
Cardiovascular diseases PheCode 411 case:control =1:12	1p32.3	1:55505647	rs11591147	G	T	Exonic	PCSK9	0.018	408,458	2.30E-12	known	[13]
	1p32.2	1:56966350	rs17114046	A	G	Intronic	PLPP3	0.092	408,458	1.36E-11	known	[14]
	1p13.3	1:109817590	rs12740374	G	T	UTR3	CELSR2	0.222	408,458	1.68E-25	known	[14]
	1q41	1:222814442	rs2133189	C	T	Intronic	MIA3	0.286	408,458	2.35E-11	known	[14]
	2p24.1	2:19942473	rs16986953	G	A	Intergenic	OSR1; LINC00954	0.068	408,458	9.96E-09	known	[14]
	2p11.2	2:85767735	rs2028900	C	T	Intronic	MAT2A	0.450	408,458	1.82E-08	known	[14]
	2q33.2	2:203968973	rs72934535	T	C	Intronic	NBEAL1	0.108	408,458	7.14E-09	known	[14]
	3q22.3	3:136294757	rs13065626	C	G	Intronic	STAG1	0.137	408,458	1.63E-08	known	[14]
	4q32.1	4:156645513	rs13139571	C	A	Intronic	GUCY1A3	0.233	408,458	2.94E-10	known	[14]
	6p24.1	6:12903957	rs9349379	A	G	Intronic	PHACTR1	0.405	408,458	6.30E-19	known	[14]
	6p21.33	6:31881731	rs685031	G	A	Intronic	C2	0.389	408,458	9.26E-09	known	[14]
	6p11.2	6:57113816	rs430918	C	T	Intergenic	RAB23; LOC100506188	0.066	408,458	4.79E-08	potential novel	
	6q14.1	6:82459034	rs78707197	T	C	UTR3	FAM46A	0.022	408,458	3.75E-10	potential novel	
	6q23.2	6:134204247	rs12194592	A	G	ncRNA_intronic	TARID	0.307	408,458	1.95E-10	known	[14]
	6q26	6:161005610	rs55730499	C	T	Intronic	LPA	0.081	408,458	4.48E-62	known	[14]
	7p21.1	7:19049388	rs2107595	G	A	Intergenic	HDAC9; TWIST1	0.152	408,458	4.23E-10	known	[14]
	7q36.1	7:150690176	rs3918226	C	T	Intronic	NOS3	0.081	408,458	1.92E-10	known	[15]
	8p21.3	8:19870271	rs35237252	C	A	Intergenic	LPL;SLC18A1	0.251	408,458	4.68E-08	known	[14]
	9p21.3	9:22103813	rs1333042	A	G	ncRNA_intronic	CDKN2B-AS1	0.496	408,458	2.29E-72	known	[14]
	9q21.12	9:73553245	rs150282530	C	T	Intronic	TRPM3	0.001	408,458	3.45E-08	potential novel	
	10p11.23	10:30317073	rs9337951	G	A	Exonic	JCAD	0.345	408,458	7.32E-09	known	[14]
	10q11.21	10:44687780	rs11238907	T	G	Intergenic	LINC00841; C10orf142	0.115	408,458	1.88E-08	known	[14]
	11p15.4	11:9766932	rs378825	A	G	Intronic	SWAP70	0.427	408,458	3.43E-08	known	[14]

	11q22.1	11:100593538	rs633185	G	C	Intronic	ARHGAP42	0.285	408,458	8.81E-09	potential novel	
	11q22.3	11:103673294	rs2839812	T	A	Intergenic	DYNC2H1; MIR4693	0.279	408,458	1.10E-11	known	[14]
	11q23.3	11:120233626	rs7924772	A	G	intronic	ARHGEF12	0.387	408,458	2.42E-09	potential novel	
	12q13.13	12:54513915	rs11170820	C	G	ncRNA_exonic	FLJ12825	0.058	408,458	1.33E-09	known	[16]
	12q24.12	12:111904371	rs4766578	T	A	intronic	ATXN2	0.495	408,458	7.97E-14	known	[14]
	12q24.13	12:112486818	rs17696736	A	G	intronic	NAA25	0.428	408,458	7.93E-11	known	[14]
	12q24.31	12:121416650	rs1169288	A	C	exonic	HNF1A	0.313	408,458	1.37E-09	known	[17]
	13q34	13:110837553	rs638634	C	T	intronic	COL4A1	0.302	408,458	1.41E-08	known	[14]
	15q25.1	15:79132330	rs11072811	A	C	intergenic	ADAMTS7; MORF4L1	0.492	408,458	1.28E-10	known	[14]
	15q26.1	15:91429287	rs4932373	A	C	intronic	FES	0.326	408,458	1.84E-17	known	[14]
	16q23.3	16:83045790	rs7500448	A	G	Intronic	CDH13	0.254	408,458	8.32E-10	known	[16]
	17q21.32	17:47340297	rs2011767	C	T	Intergenic	FLJ40194; MIR6129	0.459	408,458	1.33E-13	known	[14]
	17q21.33	17:47450057	rs7209400	C	T	ncRNA_intronic	LOC102724596	0.453	408,458	2.25E-12	known	[14]
	18q21.2	18:52723198	rs550780826	A	G	Intergenic	CCDC68; LINC01929	0.004	408,458	1.91E-08	potential novel	
	19p13.2	19:11188164	rs56125973	T	C	Intergenic	SMARCA4; LDLR	0.118	408,458	3.99E-13	known	[14]
	19q13.32	19:45412079	rs7412	C	T	Exonic	APOE	0.081	408,458	6.98E-17	known	[14]
	21q22.11	21:35593827	rs28451064	G	A	Intergenic	LINC00310; KCNE2	0.132	408,458	1.24E-14	known	[14]
Colorectal cancer PheCode 153 case:control = 1:84	8q24.21	8:128413305	rs6983267	G	T	ncRNA_exonic	CCAT2	0.481	387,318	7.03E-12	known	[18]
	15q13.3	15:33001734	rs58658771	T	A	Intergenic	SCG5; GREM1	0.179	387,318	1.41E-10	known	[19]
	18q21.1	18:46448805	rs6507874	T	C	Intronic	SMAD7	0.473	387,318	1.93E-14	known	[20]
Thyroid cancer PheCode 193 case:control =1:1138	9q22.33	9:100546600	rs925489	C	T	ncRNA_intronic	PTCSC2	0.332	407,757	5.43E-11	known	[21]

Supplementary Table 4. Estimated inflation factors of the genomic controls at different p-value quantiles and different MAF cutoffs for SAIGE, SAIGE-NoSPA, and BOLT-LMM test applied on three different phenotypes for 28 million successfully imputed genetic markers (imputation info ≥ 0.3 and MAC ≥ 20) from the UK Biobank data

			Genomic Control at q^{th} p-value quantile			
Phenotype	**Test**	**MAF cutoffs**	**Including previously reported loci**		**Excluding previously reported loci**	
			q=0.01	**q=0.001**	**q=0.01**	**q=0.001**
Cardiovascular diseases PheCode 411 case:control 1:12	All variants	SAIGE	1.13	1.237	1.11	1.163
		SAIGE-noSPA	1.152	1.324	1.131	1.242
		BOLT-LMM	1.129	1.306	1.108	1.225
	> 0.01	SAIGE	1.357	1.718	1.281	1.448
		SAIGE-noSPA	1.357	1.719	1.281	1.449
		BOLT-LMM	1.356	1.709	1.277	1.433
	< 0.01	SAIGE	1.044	1.039	1.043	1.038
		SAIGE-noSPA	1.067	1.159	1.066	1.158
		BOLT-LMM	1.031	1.13	1.028	1.13
Colorectal cancer PheCode 153 case:control 1:84	All variants	SAIGE	1.014	1.026	1.01	1.014
		SAIGE-noSPA	1.186	1.555	1.181	1.545
		BOLT-LMM	1.188	1.577	1.182	1.567
	> 0.01	SAIGE	1.051	1.116	1.039	1.073
		SAIGE-noSPA	1.052	1.121	1.04	1.077
		BOLT-LMM	1.057	1.126	1.044	1.085
	< 0.01	SAIGE	0.999	0.993	0.998	0.992
		SAIGE-noSPA	1.253	1.683	1.251	1.681
		BOLT-LMM	1.255	1.709	1.255	1.709
Thyroid cancer PheCode 193 case:control=1:1138	All variants	SAIGE	1.012	0.992	1.011	0.989
		SAIGE-noSPA	1.964	4.195	1.963	4.194
		BOLT-LMM	2	4.497	1.989	4.497
	> 0.01	SAIGE	1.01	1.036	1.007	1.026
		SAIGE-noSPA	1.015	1.069	1.012	1.058
		BOLT-LMM	1.02	1.074	1.017	1.064
	< 0.01	SAIGE	1.013	0.977	1.013	0.977
		SAIGE-noSPA	2.432	4.737	2.434	4.737
		BOLT-LMM	2.479	5.096	2.479	5.096

Supplementary Table 5. Empirical type 1 error rates for SAIGE, SAIGE-NoSPA, GMMAT, and BOLT-LMM estimated based on 10^9 simulated data sets. BOLT-LMM: compute non-infinitesimal association statistics; BOLT-LMM_lmmInfOnly: compute mixed model association statistics under the infinitesimal model

Case:Control	Test	Empirical Type 1 Error Rates	
		$\alpha = 5\times10^{-4}$	$\alpha = 5\times10^{-8}$
1:1	SAIGE	4.96 x 10^{-4}	4.65 x 10^{-8}
	SAIGE-NoSPA	4.54 x 10^{-4}	3.37 x 10^{-8}
	GMMAT	4.68 x 10^{-4}	3.52 x 10^{-8}
	BOLT-LMM_ lmmInfOnly	3.33 x 10^{-4}	2.18 x 10^{-8}
	BOLT-LMM	3.4 x 10^{-4}	3.04 x 10^{-8}
1:9	SAIGE	4.49 x 10^{-4}	4.22 x 10^{-8}
	SAIGE-NoSPA	7.40 x 10^{-4}	1.44 x 10^{-6}
	GMMAT	8.0 x 10^{-4}	1.72 x 10^{-6}
	BOLT-LMM_ lmmInfOnly	7.66 x 10^{-4}	1.75 x 10^{-6}
	BOLT-LMM	7.73 x 10^{-4}	1.73 x 10^{-6}
1:99	SAIGE	3.90 x 10^{-4}	1.29 x 10^{-8}
	SAIGE-NoSPA	3.25 x 10^{-3}	1.20 x 10^{-4}
	GMMAT	3.8 x 10^{-3}	1.73 x 10^{-4}
	BOLT-LMM_ lmmInfOnly	3.98 x 10^{-3}	1.95 x 10^{-4}
	BOLT-LMM	3.99 x 10^{-3}	1.94 x 10^{-4}

Supplementary Table 6. Test-specific α levels SAIGE and GMMAT where empirical type I errors were equal to $5x10^{-8}$. BOLT-LMM: compute non-infinitesimal association statistics; BOLT-LMM_lmmInfOnly: compute mixed model association statistics under the infinitesimal model

Case:Control	Test	Test-specific α levels
1:1	SAIGE	5.26×10^{-8}
	SAIGE-NoSPA	6.76×10^{-8}
	BOLT-LMM_ lmmInfOnly	1.4×10^{-7}
	BOLT-LMM	7.4×10^{-8}
	GMMAT	6.52×10^{-8}
1:9	SAIGE	5.64×10^{-8}
	SAIGE-NoSPA	8.0×10^{-11}
	BOLT-LMM_ lmmInfOnly	6.5×10^{-11}
	BOLT-LMM	1.6×10^{-10}
	GMMAT	5.0×10^{-11}
1:99	SAIGE	1.22×10^{-7}
	SAIGE-NoSPA	3.32×10^{-22}
	BOLT-LMM_ lmmInfOnly	1.8×10^{-25}
	BOLT-LMM	5.3×10^{-25}
	GMMAT	5.85×10^{-25}

References:

1. 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).

2. E.F. Kaasschieter. Preconditioned conjugate gradients for solving singular systems. *J. Comput. Appl. Math.* **24,** 265–275 (1988).

3. Hestenes Eduard, M. R. and S. *Methods of conjugate gradients for solving linear systems*. **49,** (NBS, 1952).

4. Loh, P. R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47,** 284–290 (2015).

5. Van Der Sluis, A. & Van Der Vorst, H. A. The Rate of Convergence of Conjugate Gradients. *Numer. Math* **48,** 543–560 (1986).

6. Hutchinson, M. F. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Commun. Stat. - Simul. Comput.* **19,** 433–450 (1990).

7. Avron, H. & Toledo, S. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *J. ACM* **58,** 1–34 (2011).

8. Allaire, J. *et al.* RcppParallel: Parallel Programming Tools for 'Rcpp'. (2016).

9. 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. *bioRxiv* (2017). doi:10.1101/109876

10. 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).

11. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42,** 348–354 (2010).

12. 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).

13. Kathiresan, S. A PCSK9 missense variant associated with a reduced risk of early-onset myocardial infarction. *N Engl J Med* **358,** 2299–2300 (2008).

14. CARDIoGRAMplusC4D Consortium *et al.* Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* **45,** 25–33 (2013).

15. Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* **47,** 1121–1130 (2015).

16. Verweij, N., Eppinga, R. N., Hagemeijer, Y. & van der Harst, P. Identification of 15 novel risk loci for coronary artery disease and genetic risk of recurrent events, atrial fibrillation and heart failure. *Sci. Rep.* **7,** 2761 (2017).

17. Reiner, A. P. *et al.* Common coding variants of the HNF1A gene are associated with multiple

cardiovascular risk phenotypes in community-based samples of younger and older European-American adults: the Coronary Artery Risk Development in Young Adults Study and The Cardiovascular Health Study. *Circ. Cardiovasc. Genet.* **2,** 244–54 (2009).

18. Haiman, C. A. *et al.* A common genetic risk factor for colorectal and prostate cancer. *Nat. Genet.* **39,** 954–956 (2007).

19. Jaeger, E. *et al.* Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nat. Genet.* **40,** 26–28 (2008).

20. Broderick, P. *et al.* A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat. Genet.* **39,** 1315–1317 (2007).

21. Pereira, J. S. *et al.* Identification of a novel germline FOXE1 variant in patients with familial non-medullary thyroid carcinoma (FNMTC). *Endocrine* **49,** 204–214 (2015).