nature genetics

SUPPLEMENTARY INFORMATION

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

In the format provided by the authors and unedited.

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

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

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

Supplementary Note
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] 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)[3] 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 $E(y_i \mid b_i) = \mu_i$ and variance $Var(y_i \mid b) = \phi v(\mu_i)$, where $v(\mu_i) = \mu_i(1-\mu_i)$ 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 $ql\,(\alpha, \beta = 0 \mid b) = \int_{y_i}^{\mu_i} \frac{a_i(y_i - \mu)}{\phi v(\mu)} d\mu$ is the quasi-likelihood for the *ith* individual given the random effect b. Let $\kappa(b) = \sum_{i=1}^{N} ql_i(\alpha, \beta = 0 \mid b) - \frac{1}{2} b^T (\tau\,\psi)^{-1} b$. 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

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

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

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

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

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 $ql(\hat{\alpha}(\phi,\tau),\beta=0,\phi,\tau)$, let $\psi_0 = \frac{1}{\phi}W^{-1}$, we take the derivative of equation (6) with respect to ϕ and τ:

$$\frac{\partial ql_R(\hat{\alpha}(\phi,\tau),\beta=0,\phi,\tau)}{\partial\phi} = \frac{1}{2}\{\tilde{Y}^T P\psi_0 P\tilde{Y} - tr(P\psi_0)\} = \frac{1}{2\phi}\tilde{Y}^T PW^{-1}P\tilde{Y} - \frac{1}{2\phi}tr(PW^{-1}) \quad (7)$$

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

$\hat{\phi}$ and $\hat{\tau}$ are estimated by obtaining the solutions to make equations (7) and (8) equal to zero. Let $\boldsymbol{\theta}$ represents the vector of variance component parameters. In this case, $\boldsymbol{\theta}$ is a vector containing ϕ and τ. In the REML iterative process, the estimates for $\boldsymbol{\theta}$ in the (i+1)th iteration is updated by $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} + J\left(\boldsymbol{\theta}^{(i)}\right)^{-1} S(\boldsymbol{\theta}^{(i)})$, where $S(\boldsymbol{\theta}) = \frac{\partial ql_R(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$ as the equation (7) and (8) and $J(\boldsymbol{\theta}) = -\frac{\partial S(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = -\frac{\partial^2 ql_R(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}$.
The elements of the observed information matrix $J(\boldsymbol{\theta})$[4] are

$$-\frac{\partial ql_R(\hat{\alpha}(\phi,\tau),\beta=0,\phi,\tau)}{\partial\phi^2} = -\frac{1}{2}tr(P\psi_0 P\psi_0) + \tilde{Y}^T P\psi_0 P\psi_0 P\tilde{Y}$$

$$-\frac{\partial ql_R(\hat{\alpha}(\phi,\tau),\beta=0,\phi,\tau)}{\partial\phi\partial\tau} = -\frac{1}{2}tr(P\psi_0 P\psi) + \tilde{Y}^T P\psi_0 P\psi P\tilde{Y}$$

$$-\frac{\partial ql_R(\hat{\alpha}(\phi,\tau),\beta=0,\phi,\tau)}{\partial\tau^2} = -\frac{1}{2}tr(P\psi P\psi) + \tilde{Y}^T P\psi P\psi P\tilde{Y} \quad (9)$$

The elements of the expected information matrix[4] are

$$E\left(-\frac{\partial ql_R(\hat{\alpha}(\phi,\tau),\beta=0,\phi,\tau)}{\partial\phi^2}\right) = \frac{1}{2}tr(P\psi_0 P\psi_0)$$

$$E\left(-\frac{\partial ql_R(\hat{\alpha}(\phi,\tau),\beta=0,\phi,\tau)}{\partial\phi\partial\tau}\right) = \frac{1}{2}tr(P\psi_0 P\psi)$$

$$E\left(-\frac{\partial ql_R(\hat{\alpha}(\phi,\tau),\beta=0,\phi,\tau)}{\partial\tau^2}\right) = \frac{1}{2}tr(P\psi P\psi) \quad (10)$$

To avoid the trace evaluation in (9), which has high computational cost, an average information matrix AI is then defined as the average of the observed information in (9) and the expected information in (10) in place of the $J(\boldsymbol{\theta})$ matrix to estimate $\hat{\phi}$ and $\hat{\tau}$ iteratively[1,4,5].

$$AI_{\phi\phi} = \frac{1}{2}\tilde{Y}^T P\psi_0 P\psi_0 P\tilde{Y}$$

$$AI_{\phi\tau} = AI_{\tau\phi} = \frac{1}{2}\tilde{Y}^T P\psi_0 P\psi P\tilde{Y}$$

$$AI_{\tau\tau} = \frac{1}{2}\tilde{Y}^T P\psi P\psi P\tilde{Y} \quad (11)$$

Note that for the logistic mixed model, $\phi = 1$, so we do not need to obtain (7) and the first two equations in (9-11) that contain derivatives with ϕ .

1.1.3 Approaches to reduce computation and memory cost.

Preconditioned Conjugate Gradient (PCG): To obtain equations (3)-(8) and (11), 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[6], 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[7]. 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 $O(NM_1)$.

The numerical accuracy of the PCG method has been evaluated based on the Euclidean distance for the vector $\Sigma^{-1}y$ computed by PCG and by calculating Σ^{-1} for the simulated data sets as described in the Data Simulation section. With the tolerance $1\text{x}10^{-5}$ for PCG to converge, the average Euclidean distances for 100 simulated data sets with case-control ratio 1:99, 1:9 and 1:1 are $2.46\text{x}10^{-11}$, $7.70\text{x}10^{-10}$, and $1.53\text{x}10^{-9}$, respectively, suggesting the PCG method is highly accurate. The average numbers of PCG interactions to convergence are 4, 6 and 7 for case-control ratio of 1:99, 1:9, and 1:1, respectively. The average iterations for PCG to converge for the 1,283 non-sex specific binary phenotypes in the UK Biobank have been plotted in **Supplementary Figure 12.** There was no phenotype with an average number of iterations larger than 10, indicating PCG converges reasonably fast in UK Biobank data analysis.

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[8,9]. The trace of a matrix B, such as PW^{-1} and $P\psi$, is estimated by $\frac{1}{R}\sum_{i=1}^{R} z_i^T B z_i$, 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 $z_i^T B z_i$. This procedure is repeated for R times and the average of the results for $z_i^T B z_i$ is the estimate for the trace of the B matrix.

The numerical stability and convergence of the randomized trace estimator has been evaluated using data sets that were simulated as described in the Data Simulation section. During the process of fitting the null generalized logistic model iteratively, the trace of the matrix $P\psi$ was estimated using different numbers of independent random vectors (R = 10, 20, 30, 40 and 50). The estimated traces were plotted against the true traces that were computed as the sum of the elements on the main diagonal of matrix $P\psi$ in the **Supplementary Figure 13**. As the number of random vectors that were used for trace estimation increases, the estimator is more stable and more consistent to the true value. Given that the trace is estimated as the average of $z_i^T B z_i$, i=1, …., R, the coefficient of variation (CV), which is defined as the ratio of standard error to the mean (i.e. SE/Mean) and measures relative variability, is used to determine whether R independent random vectors provide stable trace estimation. When R=30, in most simulated datasets, CV < 0.0025, which indicates that trace can be accurately estimated using 30 independent random vectors for the simulated data sets. Therefore, the default number of random vectors to use (R) in SAIGE is set to be 30. But it is possible that R=30 is not enough to stably calculate the trace in some datasets. In this case, R should be increased. A function to adaptively increase R when the CV is larger than a certain threshold has been implemented in the SAIGE R package.

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[10] 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.

A low-rank GRM to correct for sample relatedness: Since the computation and memory cost of step 1 in SAIGE is linear to the number of markers (M_1) used to construct the Kinship matrix, the computation and memory cost can be reduced using a subset of markers, instead of using all available markers. In the UK Biobank data analysis, for example, 93,511 independent, high quality genotyped variants were used for the step 1 (M_1 = 93,511), which is the same set of markers used by the UK Biobank data group to estimate the kinship coefficients between samples[11]. This low-rank GRM approach was first proposed by Lippert C, *et al.*[12] and has been shown to provide similar p-values to using the more complete set of genetic markers to construct GRM[12]. Later, Yang *et al.* suggested that using a few thousand genetic markers to construct GRM would reduce the ability to correct for sample relatedness. Therefore the marker selection for step 1 should be based on careful consideration for the trade-off between computation cost and performance of adjusting for sample relatedness. In Supplementary Note Section 2, a sensitivity analysis has been reported when increasing M_1 to be 340,447. Using more markers for the step 1 produced generally similar p-values but with lambdas closer to 1.

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 $N \times 1$ 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[13]. 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[6] and GRAMMAR-Gamma[14] 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

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

Suppose $\tilde{G}_i$ is the i*th* element of $\tilde{G}$. Since $\tilde{G}$ is adjusted by covariates including the intercept, $\tilde{G}_i$ can be treated as a mean zero random variable uncorrelated with the covariates, and hence $N^{-1/2}X^T\hat{\Sigma}^{-1}\tilde{G}$ asymptotically have mean zero and variance $N^{-1}X^T\hat{\Sigma}^{-1}Var\left(\tilde{G}\right)\hat{\Sigma}^{-1}X = O(1)$. By Chebyshev's inequality $N^{-1/2}X^T\hat{\Sigma}^{-1}\tilde{G} = O_P(1)$. Since $\left(X^T\hat{\Sigma}^{-1}X\right)^{-1} = O(N^{-1})$, the second term in (3) is $\tilde{G}^T\hat{\Sigma}^{-1}X\left(X^T\hat{\Sigma}^{-1}X\right)^{-1}X^T\hat{\Sigma}^{-1}\tilde{G} = O_p(1)$. 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

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

With the assumption $\frac{w_i}{\sum_{j=1}^{N} w_j} = o(1)$, $\tilde{G}^T(\overline{w}I + \hat{\tau}\psi)^{-1}\xi(\overline{w}I + \hat{\tau}\psi)^{-1}\tilde{G} = \sum\xi_i d_i$, where ξ_i is the i*th* diagonal element of ξ and d_i is the square of the i*th* element of $(\overline{w}I + \hat{\tau}\psi)^{-1}\tilde{G}$. Since the mean of ξ_i is zero, and ξ_i and d_i are uncorrelated, $\sum\xi_i d_i = o_p(N)$. Combining a fact that $\tilde{G}^T(\overline{w}I + \hat{\tau}\psi)^{-1}\tilde{G} = O_P(N)$, $\frac{G^T(\overline{w}I+\hat{\tau}\psi)^{-1}\xi(\overline{w}I+\hat{\tau}\psi)^{-1}G}{\tilde{G}^T(\overline{w}I+\hat{\tau}\psi)^{-1}\tilde{G}} = o_p(1)$, 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*, which is estimated from 30 randomly selected genetic markers. Under the null hypothesis of no association, T_{adj} has mean zero and variance one. **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 $\geq$ 20 and then variation was increased for extremely rare variants. This figure provides empirical evidence that the equal ratio assumption holds.

In analysis of simulated and real data, 30 randomly selected genetic markers with MAC $\geq$ 20 were used to estimate $\hat{r}$. To evaluate the numerical stability of the $\hat{r}$ estimation, the coefficient of variance (CV) of $\hat{r}$ using simulated datasets was used. In most simulated datasets, the CV for $\hat{r}$ was smaller than 0.001 (**Supplementary Figure 14**) with 30 randomly selected markers, indicating that $\hat{r}$ can be accurately estimated using 30 markers. As a sensitivity analysis, $\hat{r}$ has been calculated based on 500 randomly selected markers, and the estimated $\hat{r}$ were nearly identical (**Supplementary Figure 14**). But it is also possible that using 30 markers is not enough to stably calculate $\hat{r}$ in some datasets. In this case, the number of markers for $\hat{r}$ should be increased. As the same as the random trace estimation (Section 1.1.3), a function is included in the SAIGE package, in which the number of markers for $\hat{r}$ is automatically increased if CV is larger than a given threshold (current default=0.001).

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)[13], 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

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

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 $\hat{\mu}$, 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 $w = sign(\hat{\zeta})\left[2\{\hat{\zeta}q - K(\hat{\zeta})\}\right]^{\frac{1}{2}}$, $v = \hat{\zeta}\{K''(\hat{\zeta})\}^{\frac{1}{2}}$ and $\hat{\zeta} =\ \hat{\zeta}(q)$ is the solution of the equation $K'(\hat{\zeta}) = q$. As the fastSPA[13], 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[3] and GRAMMAR-Gamma[14]. 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 $\hat{\beta}$, $SE(\hat{\beta}) = |\hat{\beta}/z|$, where z-score corresponds to the association p-value/2.

1.2.5 Leave-one-chromosome-out

To avoid contamination from correlated markers[12], we implemented an option to apply the leave-one-chromosome-out (LOCO) scheme in SAIGE. In step 1, given the variance component parameter $\hat{\boldsymbol{\tau}}$ that was estimated using GRM constructed with genome-wide markers, the random and fixed effects were estimated for each chromosome using a GRM constructed with genetic markers excluding that chromosome. In the following step 2 for association tests, the estimates from step 1 using all other chromosomes are then used for testing genetic markers on that chromosome. We evaluated this approach by comparing p-values with and without the LOCO scheme. **Supplementary Figure 15** shows the scatter plots for the p-values of the 28 million genotyped and imputed markers for the four randomly selected phenotypes in the UK Biobank data. We found that the p-values estimated with and without LOCO are highly correlated.

2. Additional simulation and real data analysis results

2.1 Simulation studies with different τ values and heritability estimation

In SAIGE, penalized quasi-likelihood (PQL), which provides easy implementation and fast computation, is used to estimate the variance component parameter $\hat{\tau}$. Although PQL is the mostly widely used method in Generalized Linear Mixed Model and also used by the recently developed GMMAT method[1], it is known to produce biased estimate of the variance component ($\hat{\tau}$)[15–17], and therefore, the heritability estimates. This may be due to the fact that PQL approximates true-likelihood using Laplace method, and hence after the approximation, $\boldsymbol{\tau}$ in true likelihood is no longer the same as $\boldsymbol{\tau}$ in the approximated model.

Supplementary Table 8 shows $\hat{\boldsymbol{\tau}}$ estimated by PQL (as in SAIGE) for simulated data with four different τ values, 0.5, 1, 2, and 3, corresponding to h^2_{latent} =0.13, 0.23, 0.38, and 0.48, respectively, where h^2_{latent} is a liability scale heritability. The h^2_{latent} was obtained using the fact that the logistic regression can be described as a liability threshold model with standard logistic distribution, which has variance=$\pi^2/3$ = 3.23. Therefore the variance component parameter τ can be converted to the heritability on latent scale as

$$h^2_{latent} = \frac{\tau}{\pi^2/3 + \tau}$$

Using the relationship between τ and h^2_{latent}, τ can be estimated from the liability scale heritability estimates from other methods, such as phenotype correlation–genotype correlation (PCGC) regression method[18]. PCGC is a moment-based method and known to produce unbiased heritability estimation. We estimated $\hat{\boldsymbol{\tau}}$ as $\frac{\pi^2 h^2_{pcgc}}{3(1-h^2_{pcgc})}$, where h^2_{pcgc} is the latent scale heritability estimated by PCGC. **Supplementary Table 8** clearly suggests that $\hat{\boldsymbol{\tau}}$ from SAIGE is substantially biased. Therefore, $\hat{\boldsymbol{\tau}}$ estimated by SAIGE should not be used to interpret the heritability. $\hat{\boldsymbol{\tau}}$ from PCGC was more accurate than that from SAIGE; however, it was still biased especially when true τ was large. Since PCGC uses a probit model, which assumes that the liability follows a normal distribution, the bias may be caused by the difference between normal and logistic distributions.

To evaluate whether SAIGE can control type I errors in wide ranges of heritability, additional type I error simulations with four τ values (0.5, 1, 2, and 3) have been performed and the results are similar for

different τ values (**Supplementary Figure 16)**. The results with $\tau = 1$ and 2 are shown in **Supplementary Table 8**. To evaluate whether using more accurate τ estimate can have impact on type I error control, we also included approaches assuming 1) true τ is known (true- τ), and 2) estimating τ using PCGC regression (PCGC- τ). For both approaches, fixed and random effect terms were calculated from Equations (3) and (4) given τ. Note that since true τ is unknown in real data, the first approach (i.e true- τ) can be used in simulation study only. **Supplementary Figure 16** shows QQ plots when the variant MAF=0.01 and case control ratio=1:99. In all τ -values, the proposed PQL-based approach has very well calibrated QQ plots. Interestingly, both true- τ and PCGC- τ have deflated QQ plots, indicating that these approaches produce conservative results. As aforementioned, this may due to the fact that our score test statistics were derived from PQL not from original likelihood. We note that type I error simulations with different MAFs (0.3 and 0.05) and case control ratios (1:9 and 1:1) yielded nearly identical results (data not shown).

Overall these simulation studies clearly show that although PQL is biased for the heritability estimation, it works well for adjusting for sample-relatedness.

2.2 Simulation studies with Population stratification

To evaluate whether SAIGE can control type I error rates in the presence of population stratification, we have simulated two subpopulations with Fst 0.013, which corresponds the Fst between Finnish and non-Finnish Europeans[19], assuming that subpopulations have different disease prevalence. Each subpopulation has 1000 families, each with 10 family members based on the pedigree shown in **Supplementary Figure 4**. 93,511 genetic markers were simulated with the overall minor allele frequency (MAF) following the MAF spectrum of the genotyped markers that were used for constructing the GRM in the UK Biobank data. Three different disease prevalences were considered for subpopulations 1 and 2 (0.01 and 0.02; 0.1 and 0.2; 0.5 and 0.4). Four different τ values are used to simulate the phenotypes: 0.5, 1, 2, and 3. Association tests were performed on 10 million markers including the first four principle components as covariates. The overall MAF of 10 million markers follows the same MAF spectrum of the imputed genetic markers in the UK Biobank data. The plots for the PCs were presented in the **Supplementary Figure 8**, which shows that PC1 well separated two populations. QQ plots (**Supplementary Figure 9**) were well calibrated regardless of τ and prevalence. This simulation results clearly demonstrate that our approach can produce well calibrated p-values in the presence of population stratification.

2.3 UK Biobank data analysis with different M_1

As a sensitivity analysis, we used 340,447 genotyped markers for step 1, which were obtained by using the following pruning parameters on directly genotyped markers: using windows of 500,000 base pairs (bp), a step-size of 50 markers, and pairwise $r^2 < 0.2$. We compared association p-values for four randomly selected phenotypes in the UK Biobank data when GRM was constructed using the 93,511 genotyped markers and the 340,447 genotyped markers, respectively. Scatter plots comparing p-values of the 28 million tested genetic markers are presented in **Supplementary Figure 17**, suggesting highly correlated association p-values. We also note that when 340,447 markers were used for GRM , -log10 p-values were slightly lower than those of using 93,511 markers, especially for coronary artery disease (PheCode 411) (**Supplementary Figure 17**) and the genomic inflation factors (λ) at the 0.001, 0.01 p-value percentiles slightly decreased (**Supplementary Table 9**). Manhattan plots of these two approaches were largely similar (**Supplementary Figure 18**), in which colorectal cancer (PheCode 153), glaucoma (PheCode 365), and thyroid cancer (PheCode 193) have the exactly same number of GWAS hits.

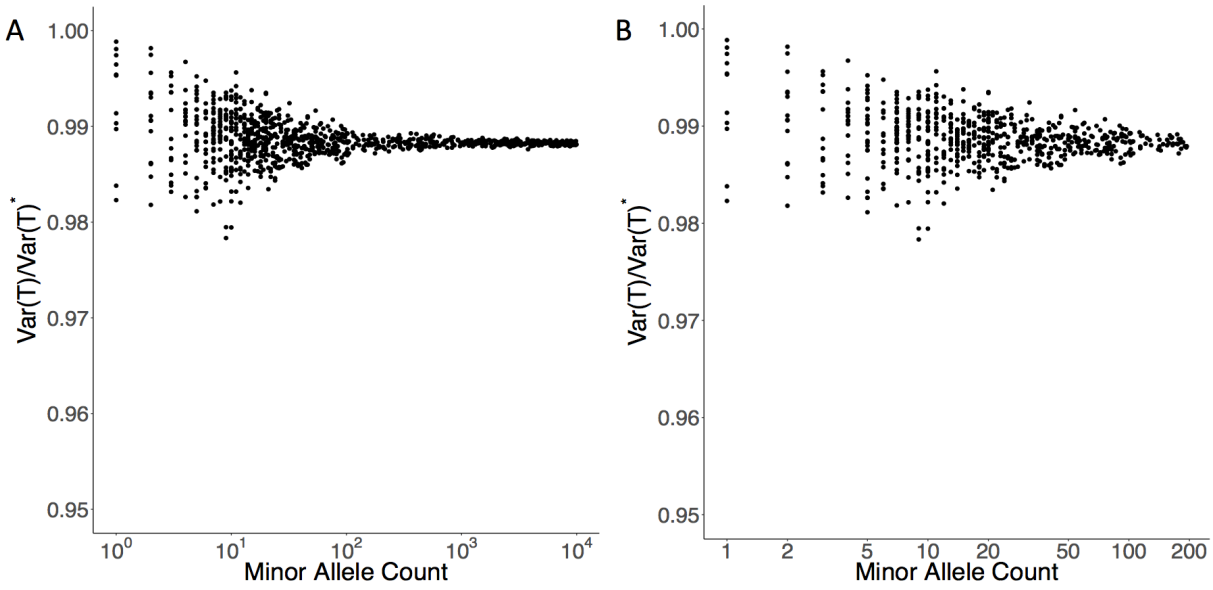
2.4 Additional rare variant associations in UK Biobank

Among SAIGE results for 1,283 non sex-specific binary phenotypes constructed based on the PheCodes in the UK Biobank data, there are total 1,609 genetic variants, including variants in the same locus, with minor allele frequency < 0.5% with SAIGE p-values < $5x10^{-8}$. Examples include the *HBB* locus (rs11549407, MAF=0.027%, p-value=$2.4x10^{-12}$) associated with hereditary hemolytic anemias (http://pheweb.sph.umich.edu:5003/pheno/282), and two different rare variants associated with breast cancer: the *ZNRF3* locus (rs6223688, MAF=0.26%, p-value=$1.8x10^{-23}$) and the *TTC28* locus (rs62237617, MAF=0.3%, p-value=$3.5x10^{-22}$) (http://pheweb.sph.umich.edu:5003/pheno/174).

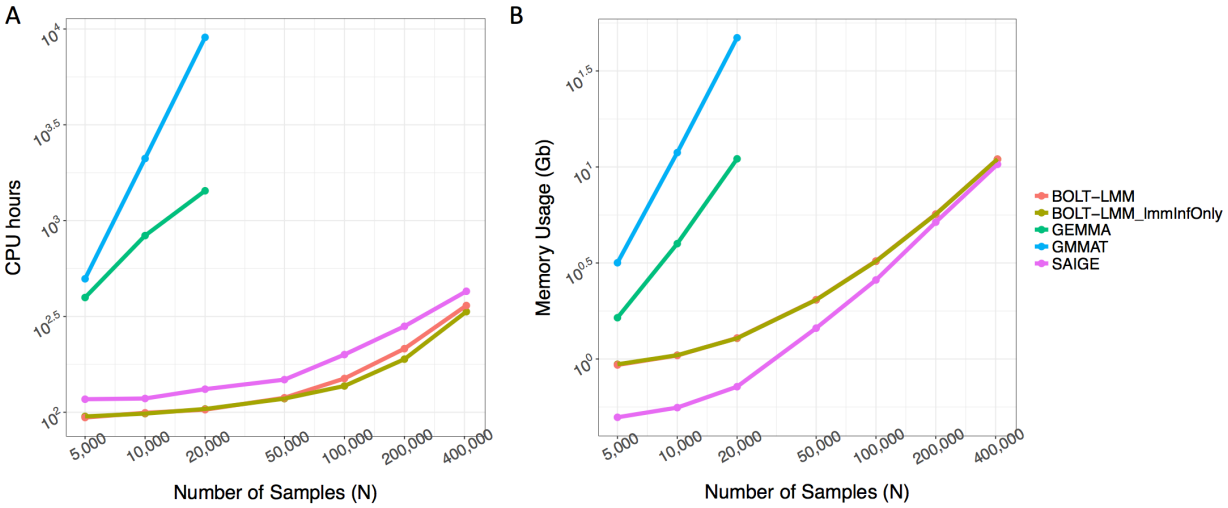
As shown in **Supplementary Table 3**, a well-known stop-gain variant rs74315329 in the gene *MYOC* for glaucoma was identified for glaucoma (PheCode 365 with 4,462 cases and 397,761 controls). This rare variant has MAF 0.14%. If rare variants were excluded from the analysis due to difficulties appropriately analyzing them, these associations would not be identified.

3. Supplementary figures

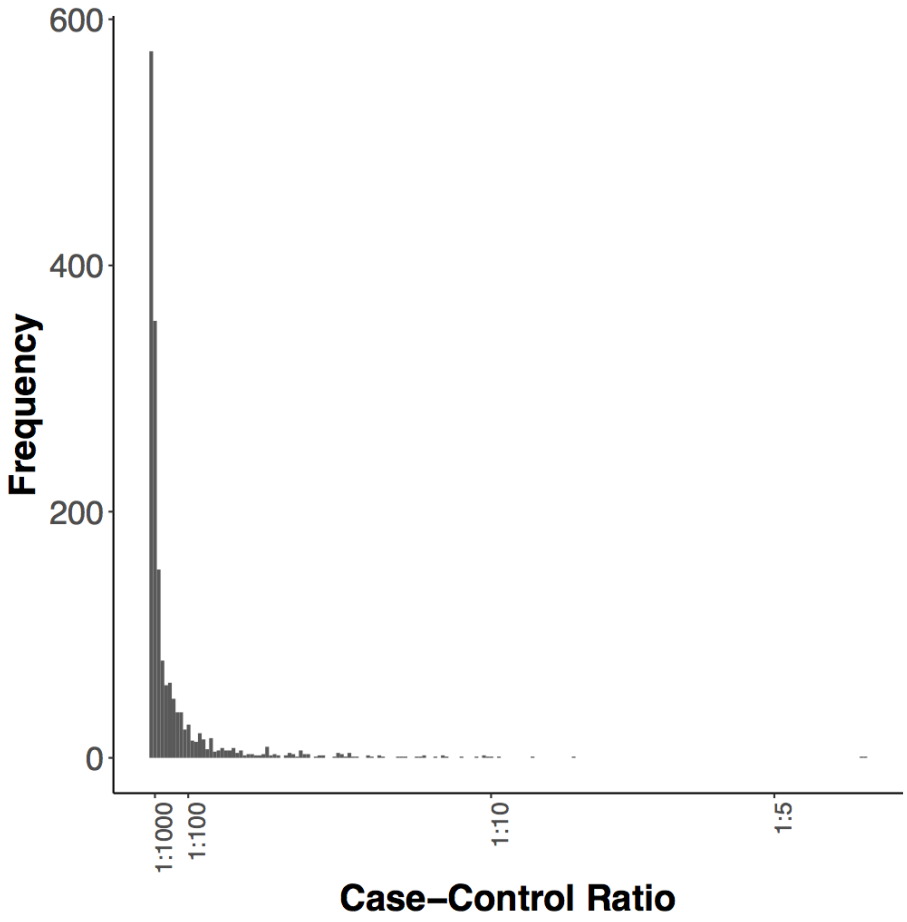
Supplementary Figure 1. Plot of the ratio of the variances of the score statistics with and without incorporating the variance components for the random effects for A. 1,000 simulated markers with MAF spectrum shown in **Supplementary Figure 11** and B. 669 out of 1,000 markers that have MAC < 200. 1,000 families were simulated based on the pedigree structure shown in **Supplementary Figure 4** with case control ratio 1:9.



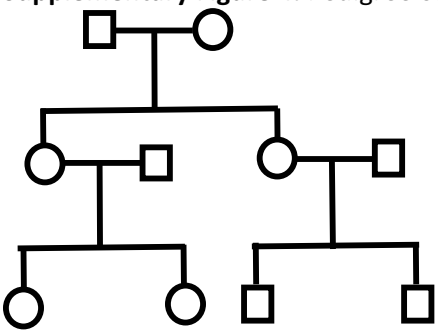
Supplementary Figure 2. Log-log plots of the estimated run time (A) and memory use (B) as a function of sample size (N). Numerical data are provided in **Supplementary Table 1.** Benchmarking was performed on randomly sub-sampled UK Biobank data with 408,458 white British participants and 200,000 markers 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



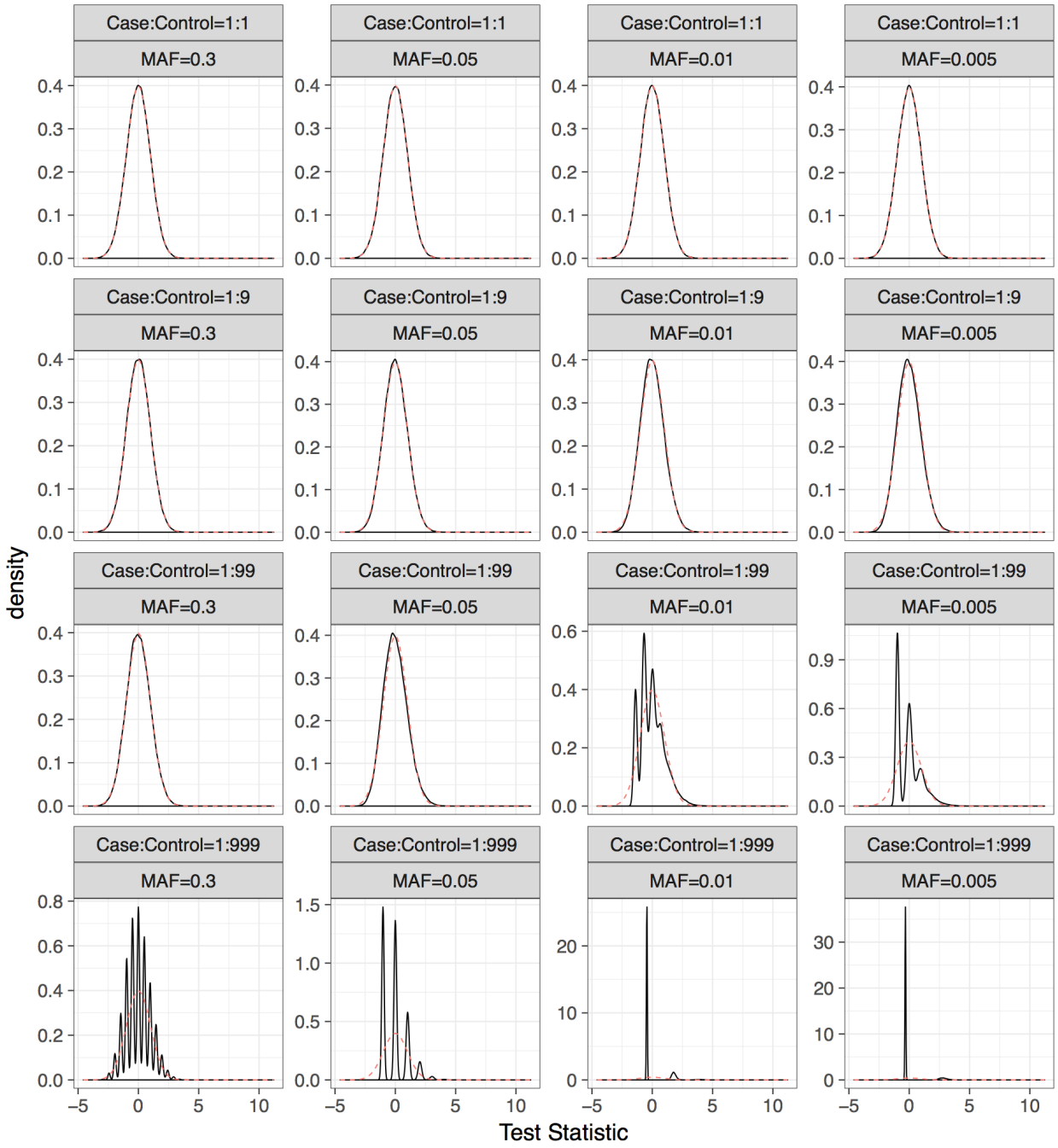
Supplementary Figure 3. Histogram of case-control ratios of 1,688 disease-specific binary phenotypes in the UK Biobank data. Phenotypes were constructed based on ICD-9 and ICD-10 codes using a previously described scheme[20].



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

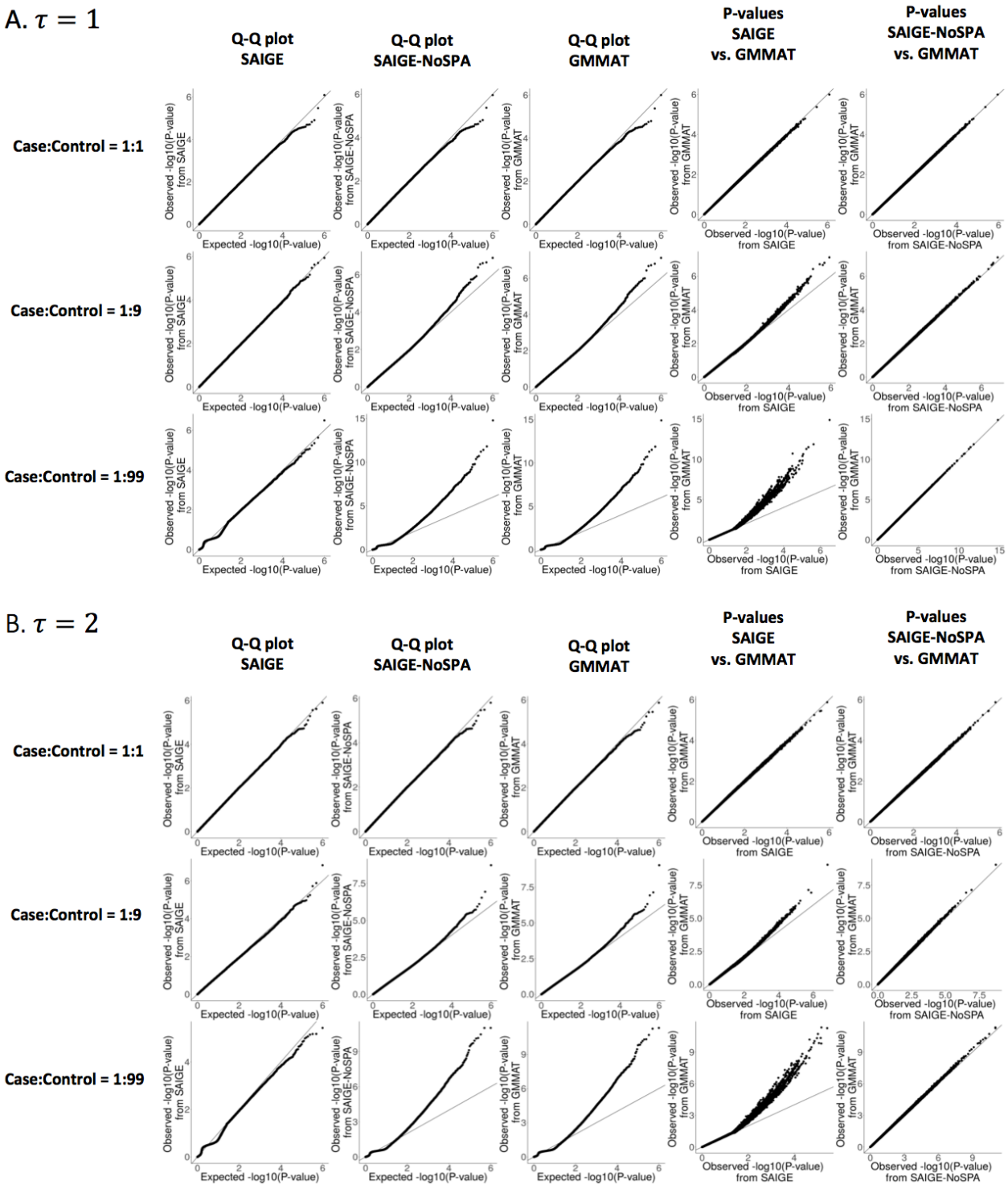


Supplementary Figure 5. Histogram of the GMMAT test statistics (solid black line) overlaid with the standard normal density curve (red dotted line) for 1,000,000 simulated genetic markers for different case-control ratios

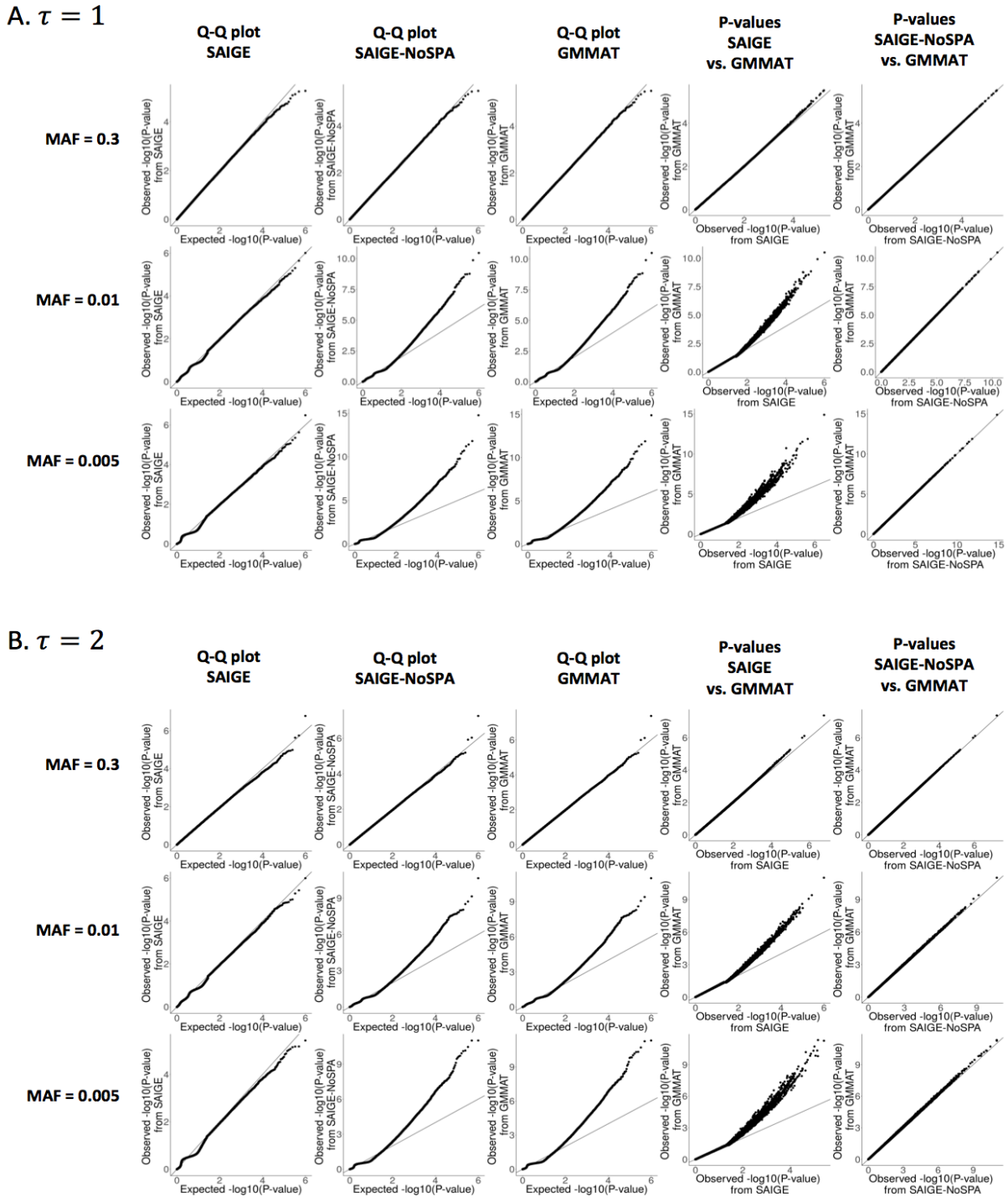


Supplementary Figure 6. Quantile-quantile plots of association p-values for 1,000,000 variants having MAF = 0.005 from the simulation study. The first column is p-values from SAIGE. The second column is for p-values from SAIGE-NoSPA. The third column is for p-values from the GMMAT[1] program. The fourth column is comparing the p-values from SAIGE and from GMMAT[1]. The fifth column is comparing the p-

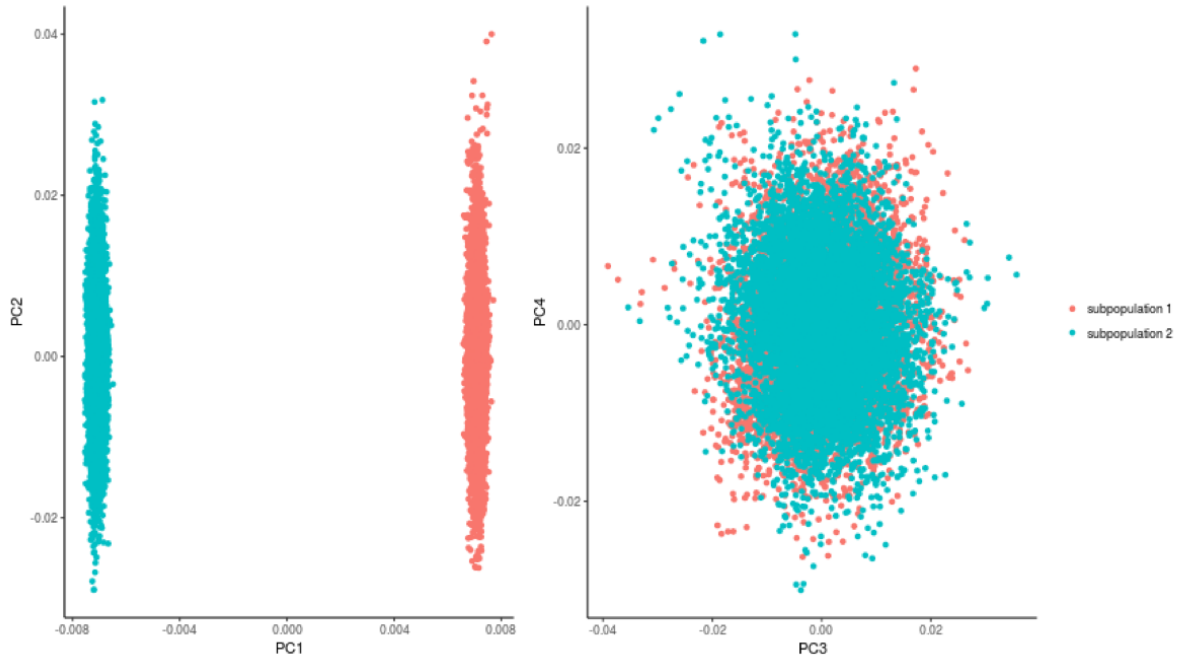
values from SAIGE-NoSPA and from GMMAT[1]. The black lines indicate x = y. τ: variance component parameter.



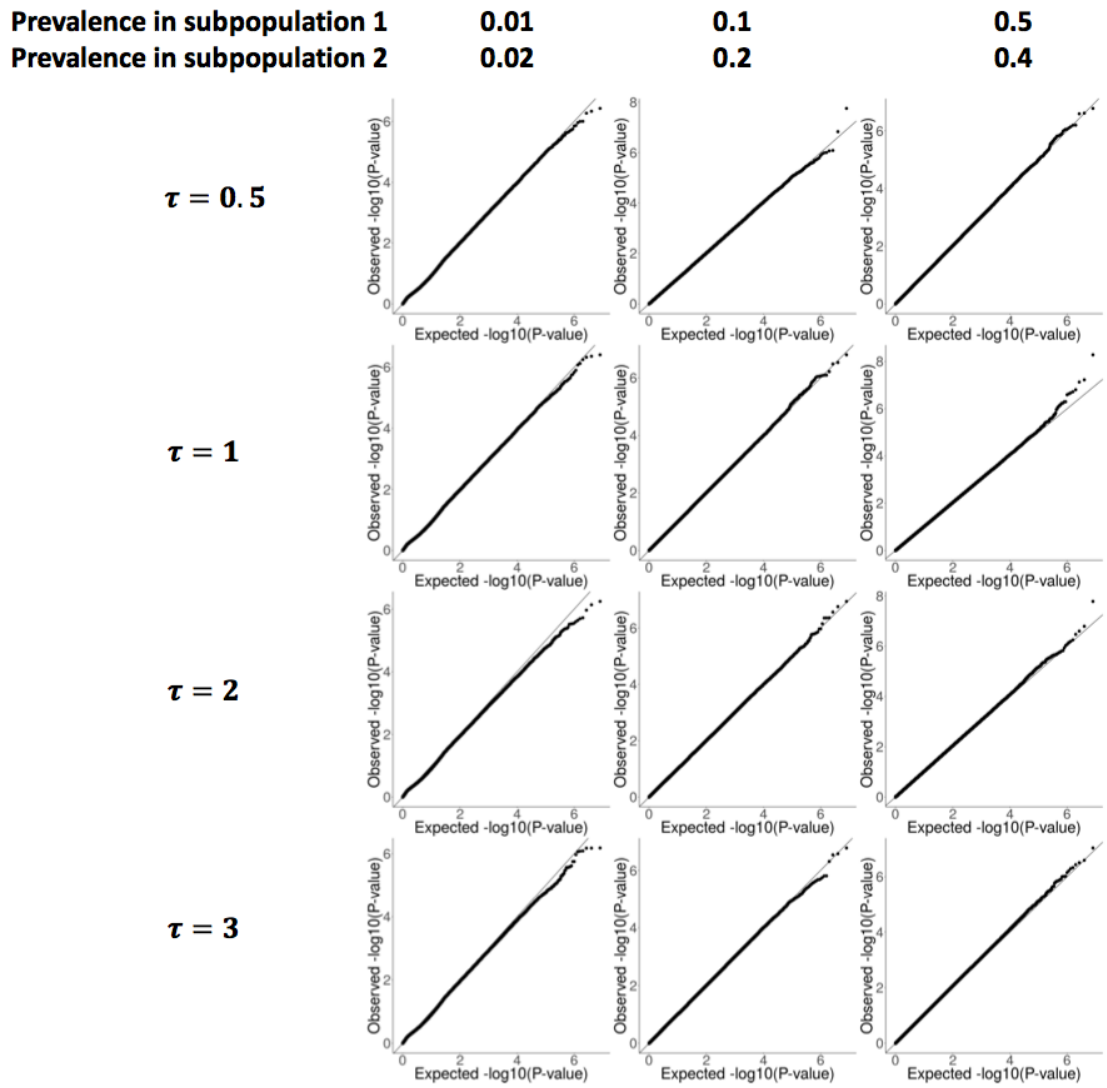
Supplementary Figure 7. Quantile-quantile plots of association p-values for 1,000,000 variants with 10,000 samples with very unbalanced case-control ratio (1:99) from the simulation study. The first column is p-values from SAIGE. The second column is for p-values from SAIGE-NoSPA. The third column is 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. τ: variance component parameter.



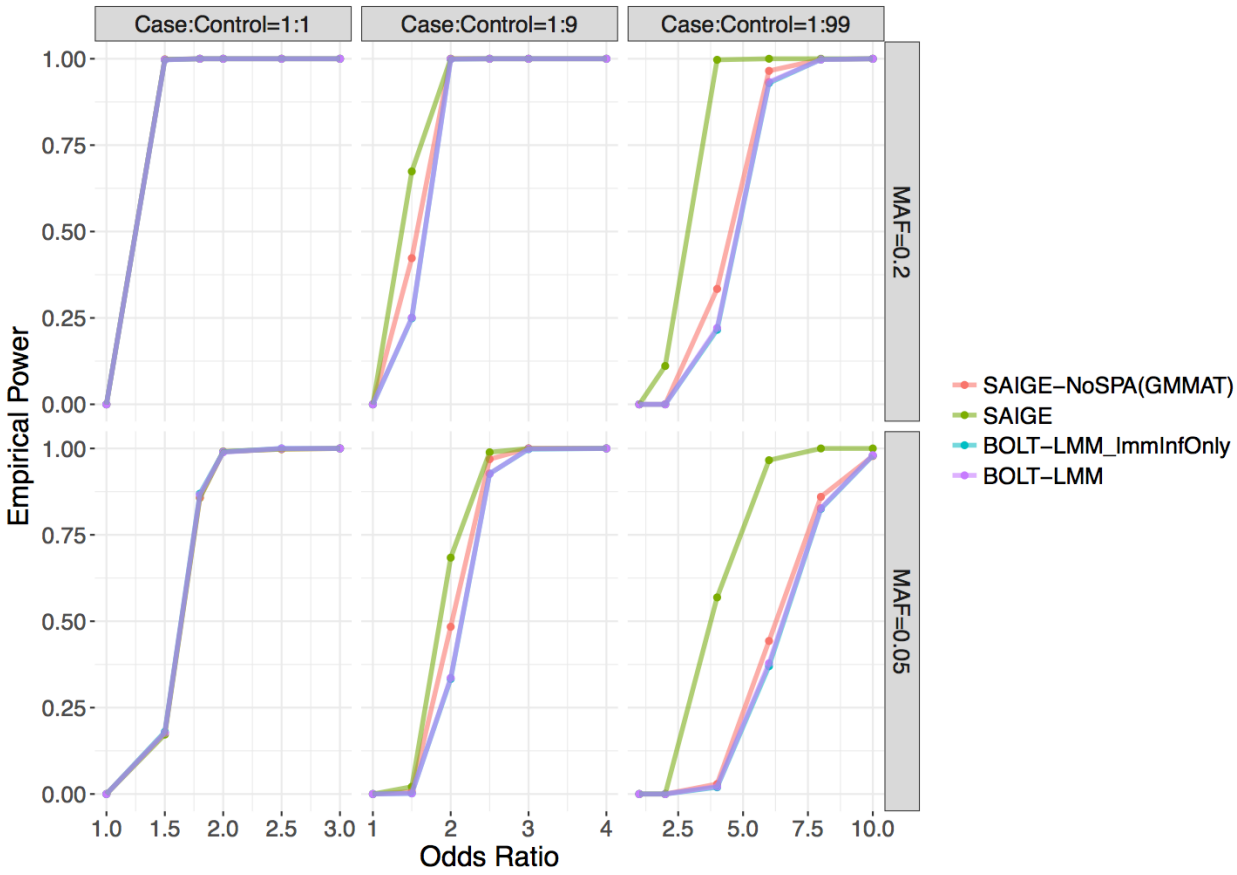
Supplementary Figure 8. Plots for the first four PCs based on the 93,511 simulated markers for samples from the two subpopulations



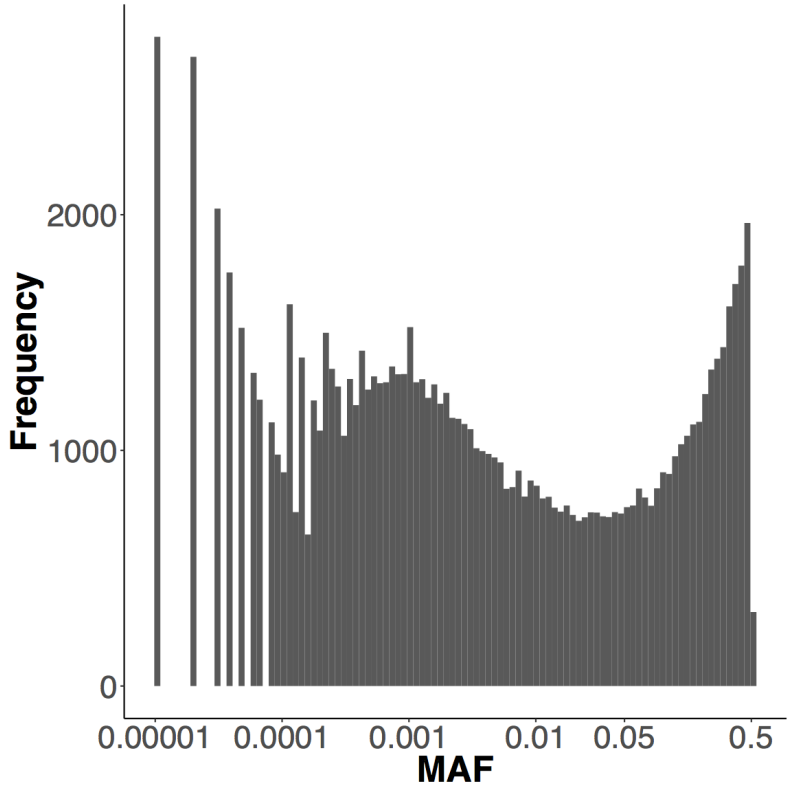
Supplementary Figure 9. Quantile-Quantile plots for the association p-values for ~10 million simulated genetic markers with MAC > 20 in presence of two subpopulations, each having 10,000 samples, with Fst = 0.013. τ: variance component parameter.



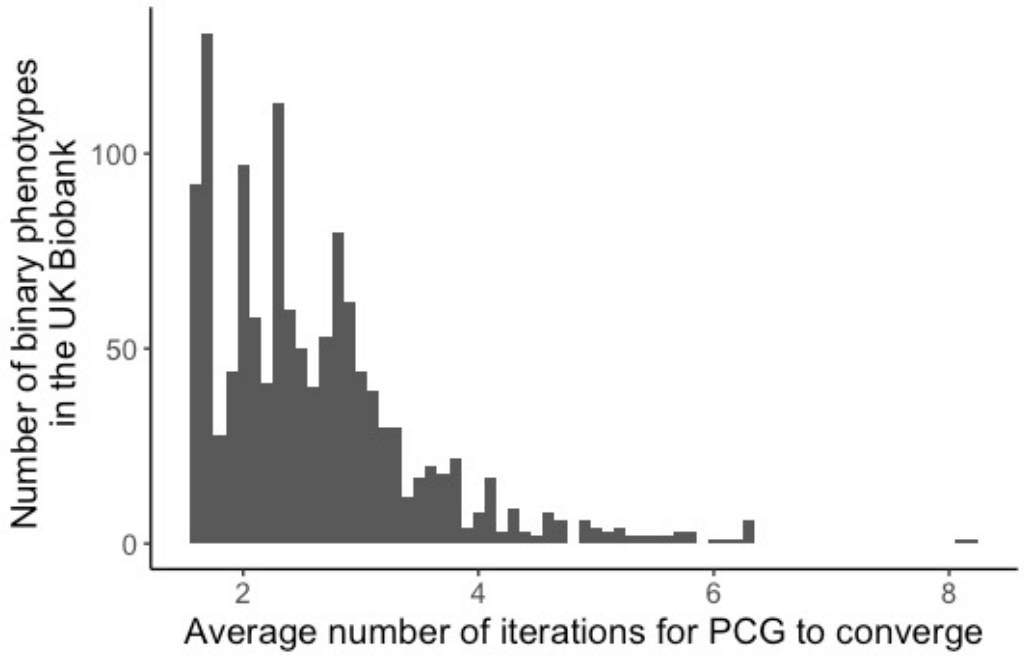
Supplementary Figure 10. Empirical power of SAIGE, SAIGE-NoSPA (asymptotically equivalent to GMMAT), BOLT-LMM_lmmInfOnly (compute mixed model association statistics under the infinitesimal model), and BOLT-LMM (compute mixed model association statistics under the non-infinitesimal model) at the test-specific empirical α levels that yield type I error rate $\alpha = 5 \times 10^{-8}$, when the variance component parameter τ=1.



Supplementary Figure 11. Distribution of the minor allele frequency spectrum for randomly selected 1,000,000 markers in the simulation study.

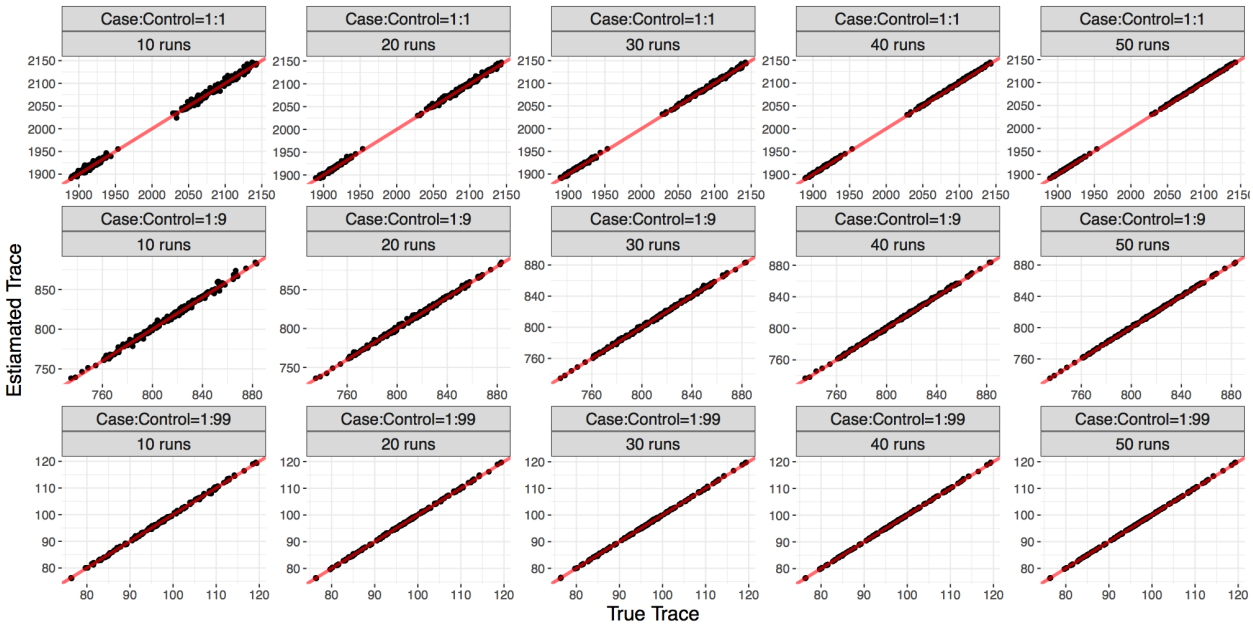


Supplementary Figure 12. Histogram of the average numbers of iterations for PCG to converge in the process of fitting the null logistic mixed model for the 1,283 non-sex specific binary phenotypes that have at least 50 cases in the UK Biobank. The PCG convergence tolerance was $1x10^{-5}$.

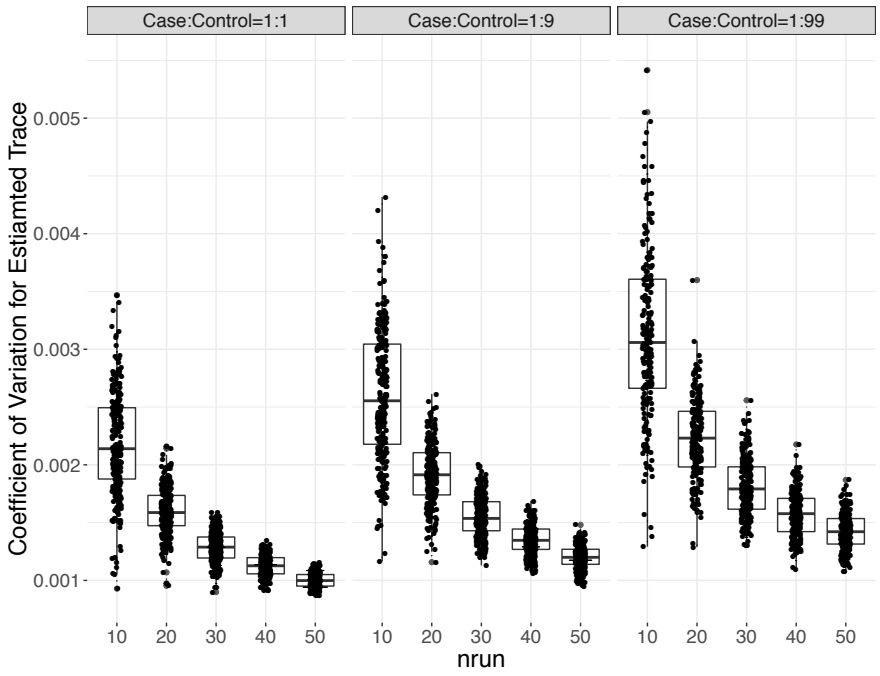


Supplementary Figure 13. Evaluation of the numerical stability for the randomized trace estimator. Simulations were performed for different numbers of random vectors: 10, 20, 30, 40, and 50 and various case-control ratios: 1:1, 1:9, and 1:99. In each plot, 200 simulations were used for every combination of the number of random vectors and the case-control ratio. A. The trace of the matrix $P\psi$ is plotted against the true traces that were computed as the sum of the elements on the main diagonal of matrix $P\psi$; B. The coefficient of variation (CV) for the trace estimator are plotted in the box plots. The line in the box represents the median, the upper hinge represents the 75% quantile, and the lower hinge represents the 25% quantile.

A

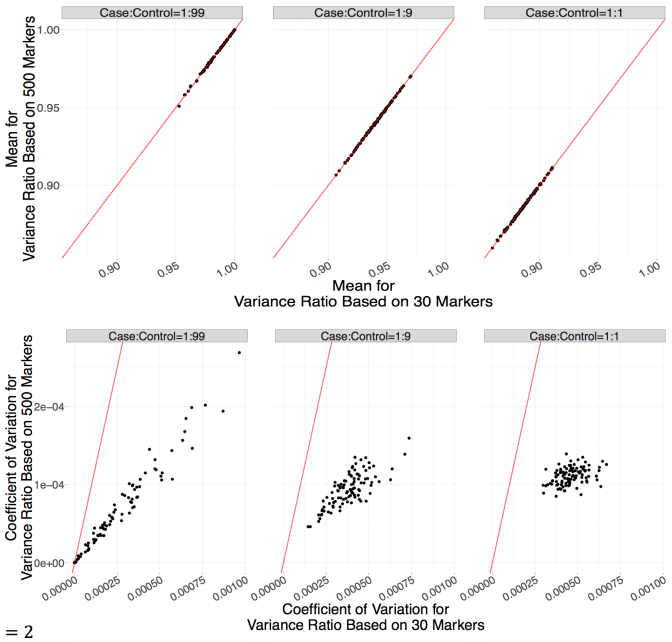


B

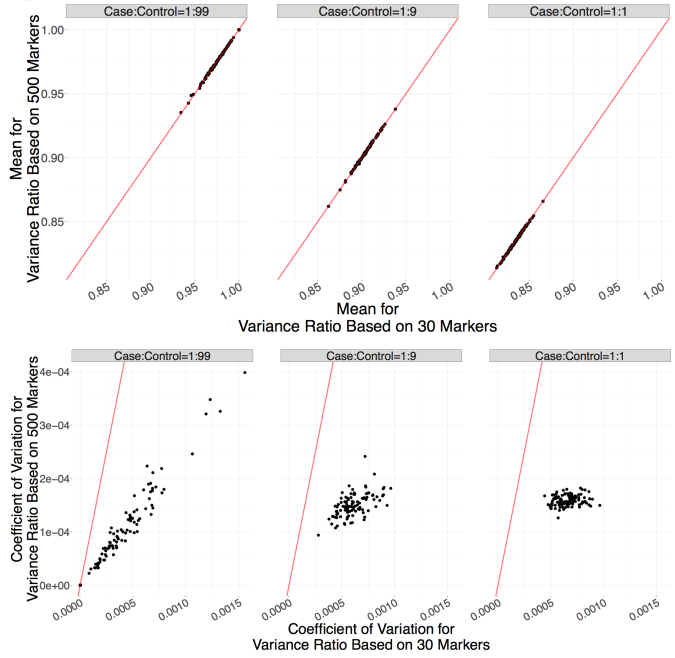


Supplementary Figure 14. The variance ratios and the coefficient of variation (CV) estimated based on 30 random genetic markers were plotted against those based on 500 markers, respectively. The thin red lines indicate x = y. τ: variance component parameter.

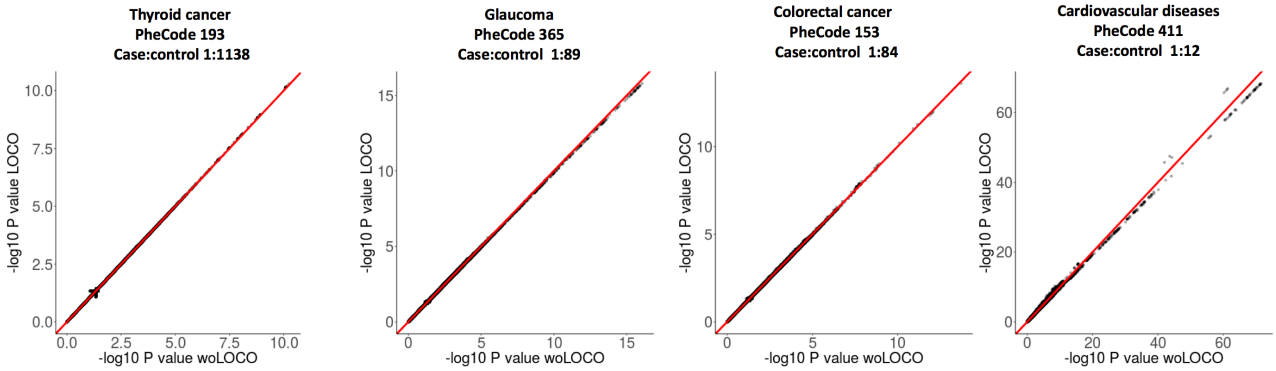
A. $\tau = 1$


Case:Control=1:99
Case:Control=1:9
Case:Control=1:1
Mean for
Variance Ratio Based on 500 Markers
Mean for
Variance Ratio Based on 30 Markers
Coefficient of Variation for
Variance Ratio Based on 500 Markers
Coefficient of Variation for
Variance Ratio Based on 30 Markers

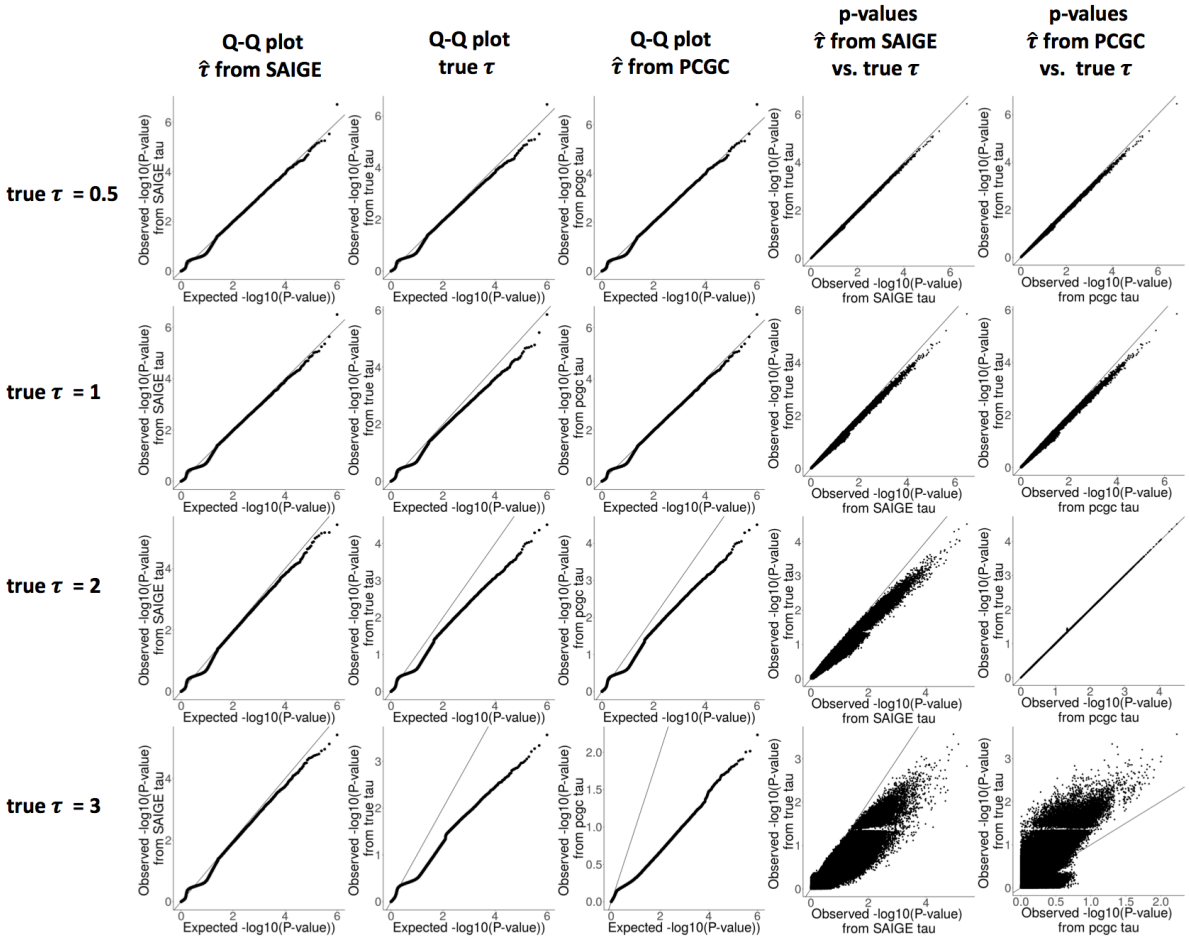
B. $\tau = 2$


Case:Control=1:99
Case:Control=1:9
Case:Control=1:1
Mean for
Variance Ratio Based on 500 Markers
Mean for
Variance Ratio Based on 30 Markers
Coefficient of Variation for
Variance Ratio Based on 500 Markers
Coefficient of Variation for
Variance Ratio Based on 30 Markers

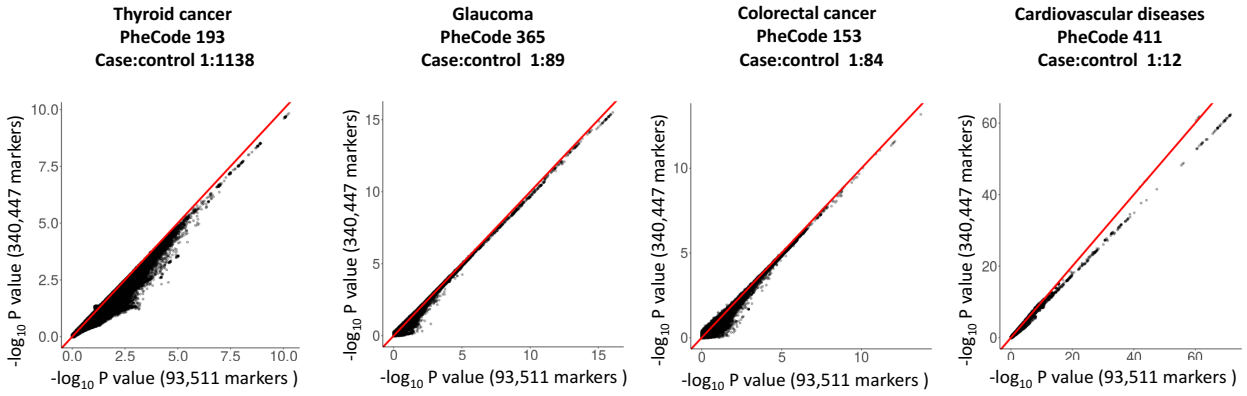
Supplementary Figure 15. Comparing association p-values from GWAS using SAIGE for the four binary phenotypes in the UK Biobank with and without the leave-one-chromosome-out (LOCO) approach. Results are shown for thyroid cancer (PheCode 193, case:control=1:1138, *N* = 407,757), glaucoma (PheCode 365, case: control = 1:89, *N* = 402,223), colorectal cancer (PheCode 153, case:control = 1:84, *N* = 387,318), and coronary artery disease (PheCode 411, case:control = 1:12, *N* = 408,458), where *N* is the sample size.



Supplementary Figure 16. Quantile-quantile plots of association p-values for 1,000,000 variants with MAF = 0.005 from the simulation study. The association tests were performed on a phenotype with case-control ratio 1:99 simulated for *N*=10,000 samples. The first column is p-values using $\hat{\tau}$ estimated by SAIGE. The second column is for p-values using true τ. The third column is for p-values using $\hat{\tau}$ estimated by PCGC. The fourth column is comparing the p-values using $\hat{\tau}$ estimated by SAIGE and using true τ. The fifth column is comparing the p-values using $\hat{\tau}$ estimated by PCGC and using true τ. The thin black lines indicate x = y. τ: variance component parameter.



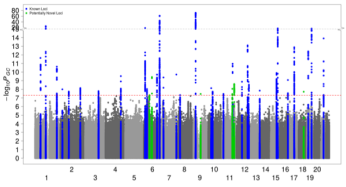
Supplementary Figure 17. Comparing association p-values for approximately 28 million genotyped or HRC imputed genetic markers for all four randomly select exemplary binary phenotypes in the UK Biobank data with a low-rank genetic relationship matrix (GRM) constructed using 93,511 genotyped markers and a GRM constructed using 340,447 directly genotyped markers. The phenotypes are thyroid cancer (PheCode 193, case: control=1:1138, *N* = 407,757), glaucoma (PheCode 365, case: control = 1:89, *N* = 402,223), colorectal cancer (PheCode 153, case: control = 1:84, *N* = 387,318), and coronary artery disease (PheCode 411, case: control = 1:12, *N* = 408,458), where *N* is the sample size.



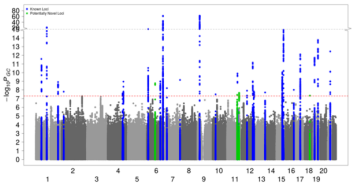
Supplementary Figure 18. Manhattan plots of association p values resulting from SAIGE with a genetic relationship matrix (GRM) constructed using 93,511 genotyped markers and a GRM constructed using 340,447 genotyped markers for A. coronary artery disease (PheCode 411, case:control = 1:12, *N* = 408,458), B. colorectal cancer (PheCode 153, case:control = 1:84, *N* = 387,318), C. glaucoma (PheCode 365, case: control = 1:89, *N* = 402,223), and D. thyroid cancer (PheCode 193, case:control=1:1138, *N* = 407,757). *N*: sample size. Blue: loci that have association p-value < $5x10^{-8}$, where the top hits are previously reported, Green: loci that have association p-value < $5x10^{-8}$ and have not been reported before.

A. Coronary Artery Disease

GRM-93,511

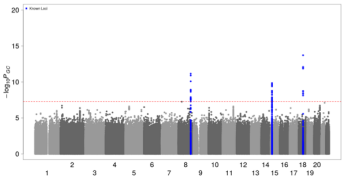


GRM-340,447

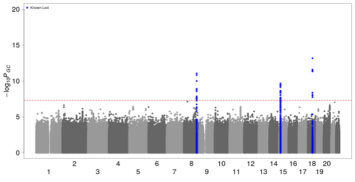


B. Colorectal Cancer

GRM-93,511

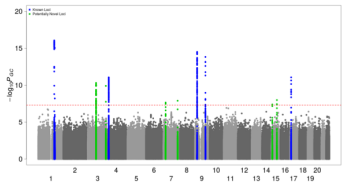


GRM-340,447

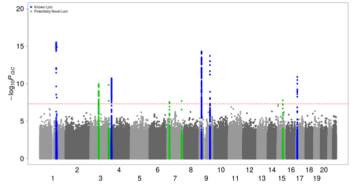


C. Glaucoma

GRM-93,511

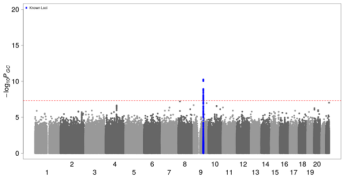


GRM-340,447

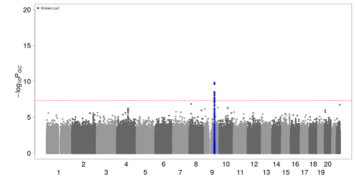


D. Thyroid Cancer

GRM-93,511



GRM-340,447



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: 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 four '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
Coronary artery disease 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
Glaucoma PheCode 365 case:control 1:89	SAIGE	362	12	6
	SAIGE-NoSPA	3,278	NA	NA
	BOLT-LMM	4,228	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 four phenotypes identified by the SAIGE in the UK Biobank data. Genomic coordinates are given in NCBI Build 37/UCSC hg19.

Phenotype	Location	Chr:Pos	rsID	Ref	Alt	Function	Gene	MAF	Sample Size	P value	Known for CAD	Previous Findings
Coronary artery disease PheCode 411 case:control =1:12	1p32.3	1:55505647	rs11591147	G	T	Exonic	*PCSK9*	0.018	408,458	2.30E-12	known	[21]
	1p32.2	1:56966350	rs17114046	A	G	Intronic	*PLPP3*	0.092	408,458	1.36E-11	known	[22]
	1p13.3	1:109817590	rs12740374	G	T	UTR3	*CELSR2*	0.222	408,458	1.68E-25	known	[22]
	1q41	1:222814442	rs2133189	C	T	Intronic	*MIA3*	0.286	408,458	2.35E-11	known	[22]
	2p24.1	2:19942473	rs16986953	G	A	Intergenic	*OSR1; LINC00954*	0.068	408,458	9.96E-09	known	[22]
	2p11.2	2:85767735	rs2028900	C	T	Intronic	*MAT2A*	0.450	408,458	1.82E-08	known	[22]
	2q33.2	2:203968973	rs72934535	T	C	Intronic	*NBEAL1*	0.108	408,458	7.14E-09	known	[22]
	3q22.3	3:136294757	rs13065626	C	G	Intronic	*STAG1*	0.137	408,458	1.63E-08	known	[22]
	4q32.1	4:156645513	rs13139571	C	A	Intronic	*GUCY1A3*	0.233	408,458	2.94E-10	known	[22]
	6p24.1	6:12903957	rs9349379	A	G	Intronic	*PHACTR1*	0.405	408,458	6.30E-19	known	[22]
	6p21.33	6:31881731	rs685031	G	A	Intronic	*C2*	0.389	408,458	9.26E-09	known	[22]
	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	[22]
	6q26	6:161005610	rs55730499	C	T	Intronic	*LPA*	0.081	408,458	4.48E-62	known	[22]
	7p21.1	7:19049388	rs2107595	G	A	Intergenic	*HDAC9; TWIST1*	0.152	408,458	4.23E-10	known	[22]
	7q36.1	7:150690176	rs3918226	C	T	Intronic	*NOS3*	0.081	408,458	1.92E-10	known	[23]
	8p21.3	8:19870271	rs35237252	C	A	Intergenic	*LPL;SLC18A1*	0.251	408,458	4.68E-08	known	[22]
	9p21.3	9:22103813	rs1333042	A	G	ncRNA_intronic	*CDKN2B-AS1*	0.496	408,458	2.29E-72	known	[22]
	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	[22]
	10q11.21	10:44687780	rs11238907	T	G	Intergenic	*LINC00841; C10orf142*	0.115	408,458	1.88E-08	known	[22]
	11p15.4	11:9766932	rs378825	A	G	Intronic	*SWAP70*	0.427	408,458	3.43E-08	known	[22]
	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	22
	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	24
	12q24.12	12:111904371	rs4766578	T	A	intronic	*ATXN2*	0.495	408,458	7.97E-14	known	22
	12q24.13	12:112486818	rs17696736	A	G	intronic	*NAA25*	0.428	408,458	7.93E-11	known	22
	12q24.31	12:121416650	rs1169288	A	C	exonic	*HNF1A*	0.313	408,458	1.37E-09	known	25
	13q34	13:110837553	rs638634	C	T	intronic	*COL4A1*	0.302	408,458	1.41E-08	known	22
	15q25.1	15:79132330	rs11072811	A	C	intergenic	*ADAMTS7; MORF4L1*	0.492	408,458	1.28E-10	known	22
	15q26.1	15:91429287	rs4932373	A	C	intronic	*FES*	0.326	408,458	1.84E-17	known	22
	16q23.3	16:83045790	rs7500448	A	G	Intronic	*CDH13*	0.254	408,458	8.32E-10	known	24
	17q21.32	17:47340297	rs2011767	C	T	Intergenic	*FLJ40194; MIR6129*	0.459	408,458	1.33E-13	known	22
	17q21.33	17:47450057	rs7209400	C	T	ncRNA_intronic	*LOC102724596*	0.453	408,458	2.25E-12	known	22
	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	22
	19q13.32	19:45412079	rs7412	C	T	Exonic	*APOE*	0.081	408,458	6.98E-17	known	22
	21q22.11	21:35593827	rs28451064	G	A	Intergenic	*LINC00310; KCNE2*	0.132	408,458	1.24E-14	known	22
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	26
	15q13.3	15:33001734	rs58658771	T	A	Intergenic	*SCG5; GREM1*	0.179	387,318	1.41E-10	known	27
	18q21.1	18:46448805	rs6507874	T	C	Intronic	*SMAD7*	0.473	387,318	1.93E-14	known	28
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	29
Glaucoma PheCode 365	1q24.1	1:165743523	rs2790049	A	G	ncRNA_exonic	*LOC100147773*	0.124003	402,223	8.71E-17	known	30
	1q24.3	1:171605478	rs74315329	G	A	exonic	*MYOC*	0.001372	402,223	9.13E-16	known	31

Case:control 1:89	3p12.1	3:85134557	rs9309969	T	G	intronic	*CADM2*	0.405599	402,223	4.94E-11	potential_novel	
	3q27.3	3:186128816	rs56233426	A	G	intergenic	*DGKG;LINC02052*	0.462593	402,223	1.25E-10	potential_novel	
	4p16.1	4:7889096	rs7663205	C	T	intronic	*AFAP1*	0.400390	402,223	8.82E-12	known	[32]
	7p15.3	7:22293117	rs113432289	A	C	intronic	*RAPGEF5*	0.000117	402,223	2.26E-08	potential_novel	
	7q35	7:146348027	rs540694424	G	C	intronic	*CNTNAP2*	0.000037	402,223	1.27E-08	potential_novel	
	9p21.3	9:22052734	rs6475604	T	C	ncRNA_intronic	*CDKN2B-AS1*	0.430593	402,223	3.12E-15	known	[30]
	9q31.1	9:107693201	rs2437812	A	C	intergenic	*ABCA1;SLC44A1*	0.423577	402,223	1.49E-14	known	[33]
	15q13.1	15:28365618	rs12913832	A	G	intronic	*HERC2*	0.214160	402,223	4.05E-08	potential_novel	
	15q24.2	15:76049154	rs187112398	C	T	intergenic	*DNM1P35;MIR4313*	0.000789	402,223	1.03E-08	potential_novel	
	17p13.1	17:10031090	rs12150284	C	T	intronic	*GAS7*	0.373370	402,223	8.70E-12	known	[34]

Supplementary Table 4. Estimated inflation factors of the genomic control (λ) at different p-value quantiles and different MAF cutoffs for SAIGE, SAIGE-NoSPA, and BOLT-LMM test applied on four 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**
Coronary artery disease PheCode 411 case:control 1:12	All variants	SAIGE	1.132	1.244	1.112	1.166
		SAIGE-noSPA	1.155	1.329	1.133	1.249
		BOLT-LMM	1.129	1.306	1.108	1.225
	> 0.01	SAIGE	1.363	1.72	1.284	1.445
		SAIGE-noSPA	1.363	1.721	1.284	1.445
		BOLT-LMM	1.356	1.709	1.277	1.433
	< 0.01	SAIGE	1.046	1.041	1.045	1.04
		SAIGE-noSPA	1.069	1.162	1.069	1.16
		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
Glaucoma PheCode 365 case:control=1:89	All variants	SAIGE	1.024	1.039	1.021	1.033
		SAIGE-noSPA	1.204	1.576	1.2	1.567
		BOLT-LMM	1.222	1.634	1.216	1.621
	> 0.01	SAIGE	1.077	1.141	1.069	1.114
		SAIGE-noSPA	1.078	1.144	1.07	1.118
		BOLT-LMM	1.085	1.153	1.078	1.126
	< 0.01	SAIGE	1.004	1.003	1.003	1.003
		SAIGE-noSPA	1.266	1.702	1.265	1.702
		BOLT-LMM	1.285	1.77	1.285	1.77
Thyroid cancer PheCode 193 ase: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

Variance Component Parameter τ	Case:Control	Test	Empirical Type 1 Error Rates	
			$\alpha = 5\times10^{-4}$	$\alpha = 5\times10^{-8}$
1	1:1	SAIGE	5.11x10^{-4}	5.45x10^{-8}
		SAIGE-NoSPA	4.71x10^{-4}	4.00x10^{-8}
		GMMAT	4.66x10^{-4}	3.81x10^{-8}
		BOLT-LMM_ lmmInfOnly	4.83x10^{-4}	4.83x10^{-8}
		BOLT-LMM	4.95x10^{-4}	4.99x10^{-8}
	1:9	SAIGE	4.43x10^{-4}	4.01x10^{-8}
		SAIGE-NoSPA	6.72x10^{-4}	7.82x10^{-7}
		GMMAT	7.30x10^{-4}	1.00x10^{-6}
		BOLT-LMM_ lmmInfOnly	9.01x10^{-4}	2.73x10^{-6}
		BOLT-LMM	9.03x10^{-4}	2.71x10^{-6}
	1:99	SAIGE	3.82x10^{-4}	1.44x10^{-8}
		SAIGE-NoSPA	2.93x10^{-3}	9.76x10^{-5}
		GMMAT	3.31x10^{-3}	1.26x10^{-4}
		BOLT-LMM_ lmmInfOnly	4.02x10^{-3}	2.10x10^{-4}
		BOLT-LMM	4.02x10^{-3}	2.10x10^{-4}
2	1:1	SAIGE	5.15x10^{-4}	3.53x10^{-8}
		SAIGE-NoSPA	4.75x10^{-4}	2.72x10^{-8}
		GMMAT	4.64x10^{-4}	2.56x10^{-8}
		BOLT-LMM_ lmmInfOnly	5.03x10^{-4}	3.74x10^{-8}
		BOLT-LMM	5.21x10^{-4}	3.59x10^{-8}
	1:9	SAIGE	4.07x10^{-4}	3.20x10^{-8}
		SAIGE-NoSPA	5.96x10^{-4}	4.94x10^{-7}
		GMMAT	7.07x10^{-4}	8.01x10^{-7}
		BOLT-LMM_ lmmInfOnly	9.88x10^{-4}	3.51x10^{-6}
		BOLT-LMM	9.88x10^{-4}	3.52x10^{-6}
	1:99	SAIGE	3.53x10^{-4}	2.08x10^{-8}
		SAIGE-NoSPA	2.66x10^{-3}	7.75x10^{-5}
		GMMAT	3.13x10^{-3}	1.08x10^{-4}
		BOLT-LMM_ lmmInfOnly	4.13x10^{-3}	2.30x10^{-4}
		BOLT-LMM	4.13x10^{-3}	2.30x10^{-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

Variance Component Parameter τ	Case:Control	Test	Test-specific α levels
1	1:1	SAIGE	$4.74x10^{-8}$
		SAIGE-NoSPA	$5.70x10^{-8}$
		BOLT-LMM_ lmmInfOnly	$5.20x10^{-8}$
		BOLT-LMM	$4.80x10^{-8}$
		GMMAT	$6.79x10^{-8}$
	1:9	SAIGE	$6.08x10^{-8}$
		SAIGE-NoSPA	$6.98x10^{-10}$
		BOLT-LMM_ lmmInfOnly	$1.60x10^{-11}$
		BOLT-LMM	$1.70x10^{-11}$
		GMMAT	$5.29x10^{-10}$
	1:99	SAIGE	$1.02x10^{-7}$
		SAIGE-NoSPA	$1.54x10^{-22}$
		BOLT-LMM_ lmmInfOnly	$5.80x10^{-26}$
		BOLT-LMM	$8.40x10^{-26}$
		GMMAT	$1.50x10^{-23}$
2	1:1	SAIGE	$6.76x10^{-8}$
		SAIGE-NoSPA	$8.01x10^{-8}$
		BOLT-LMM_ lmmInfOnly	$6.40x10^{-8}$
		BOLT-LMM	$6.30x10^{-8}$
		GMMAT	$8.42x10^{-8}$
	1:9	SAIGE	$7.85x10^{-8}$
		SAIGE-NoSPA	$2.30x10^{-9}$
		BOLT-LMM_ lmmInfOnly	$1.40x10^{-11}$
		BOLT-LMM	$1.40x10^{-11}$
		GMMAT	$8.73x10^{-10}$
	1:99	SAIGE	$1.59x10^{-7}$
		SAIGE-NoSPA	$2.10x10^{-21}$
		BOLT-LMM_ lmmInfOnly	$8.10x10^{-28}$
		BOLT-LMM	$9.60x10^{-28}$
		GMMAT	$6.69x10^{-23}$

Supplementary Table 7. The variance component estimates $\hat{\tau}$ were estimated using SAIGE and PCGC for 100 simulated data sets for each combination of prevalence and the variance component parameter τ.

		$\hat{\tau}$ from PCGC		$\hat{\tau}$ from SAIGE	
Prevalence	τ	**Mean**	**SD**	**Mean**	**SD**
0.01	0.5	1.181	1.602	0.257	0.29
0.01	1	1.226	1.691	0.372	0.295
0.01	2	4.287	12.057	0.631	0.373
0.01	3	6.771	16.294	0.834	0.428
0.1	0.5	0.348	0.158	0.182	0.069
0.1	1	0.701	0.223	0.322	0.075
0.1	2	1.473	0.351	0.534	0.072
0.1	3	2.329	0.523	0.689	0.073
0.5	0.5	0.362	0.085	0.185	0.035
0.5	1	0.714	0.11	0.312	0.036
0.5	2	1.396	0.164	0.481	0.035
0.5	3	2.022	0.234	0.58	0.037

Supplementary Table 8. Estimated inflation factors of the genomic factor (λ) at different p-value quantiles and different MAF cutoffs when applying SAIGE using 93,511 genetic markers to construct GRM vs. using 340,447 genetic markers to construct GRM on four 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**
Coronary artery disease PheCode 411 case:control 1:12	All variants	GRM-93,511	1.132	1.244	1.112	1.166
		GRM-340,447	1.048	1.137	1.032	1.074
	> 0.01	GRM-93,511	1.363	1.72	1.284	1.445
		GRM-340,447	1.217	1.523	1.157	1.277
	< 0.01	GRM-93,511	1.046	1.041	1.045	1.04
		GRM-340,447	0.985	0.98	0.984	0.979
Colorectal cancer PheCode 153 case:control 1:84	All variants	GRM-93,511	1.014	1.026	1.01	1.014
		GRM-340,447	0.993	1.004	0.99	0.993
	> 0.01	GRM-93,511	1.051	1.116	1.039	1.073
		GRM-340,447	1.026	1.088	1.014	1.047
	< 0.01	GRM-93,511	0.999	0.993	0.998	0.992
		GRM-340,447	0.981	0.973	0.98	0.972
Glaucoma PheCode 365 case:control=1:89	All variants	GRM-93,511	1.024	1.039	1.021	1.033
		GRM-340,447	1.008	1.022	1.006	1.015
	> 0.01	GRM-93,511	1.077	1.141	1.069	1.114
		GRM-340,447	1.057	1.119	1.049	1.094
	< 0.01	GRM-93,511	1.004	1.003	1.003	1.003
		GRM-340,447	0.99	0.988	0.989	0.988
Thyroid cancer PheCode 193 ase:control=1:1138	All variants	GRM-93,511	1.012	0.992	1.011	0.989
		GRM-340,447	0.957	0.933	0.956	0.931
	> 0.01	GRM-93,511	1.01	1.036	1.007	1.026
		GRM-340,447	0.969	0.991	0.966	0.981
	< 0.01	GRM-93,511	1.013	0.977	1.013	0.977
		GRM-340,447	0.953	0.91	0.953	0.91

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

4. 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 (1995).

5. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88,** 76–82 (2011).

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

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

8. 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).

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

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

11. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* 166298 (2017). doi:10.1101/166298

12. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat Methods* **8,** 833–835 (2011).

13. 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).

14. 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).

15. Capanu, M., Gönen, M. & Begg, C. B. An assessment of estimation methods for generalized linear mixed models with binary outcomes. *Stat. Med.* **32,** 4550–4566 (2013).

16. Breslow, N. Whither PQL? in 1–22 (Springer, New York, NY, 2004). doi:10.1007/978-1-4419-9076-1_1

17. Breslow, N. E. & Clayton, D. G. Approximate Inference in Generalized Linear Mixed Models. *J. Am. Stat. Assoc.* **88,** 9 (1993).

18. Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: Inferring the contribution of common variants. *Proc. Natl. Acad. Sci.* **111,** E5272–E5281 (2014).

19. Nelis, M. *et al.* Genetic structure of Europeans: a view from the North-East. *PLoS One* **4,** e5472 (2009).

20. 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).

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

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

23. 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).

24. 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).

25. 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).

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

27. 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).

28. 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).

29. 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).

30. Burdon, K. P. *et al.* Genome-wide association study identifies susceptibility loci for open angle glaucoma at TMCO1 and CDKN2B-AS1. *Nat. Genet.* **43,** 574–578 (2011).

31. Stone, E. M. *et al.* Identification of a gene that causes primary open angle glaucoma. *Science* **275,** 668–70 (1997).

32. Gharahkhani, P. *et al.* Common variants near ABCA1, AFAP1 and GMDS confer risk of primary open-angle glaucoma. *Nat. Genet.* **46,** 1120–1125 (2014).

33. Chen, Y. *et al.* Common variants near ABCA1 and in PMM2 are associated with primary open-angle glaucoma. *Nat. Genet.* **46,** 1115–1119 (2014).

34. Bailey, J. N. C. *et al.* Genome-wide association analysis identifies TXNRD2, ATXN2 and FOXC1

as susceptibility loci for primary open-angle glaucoma. *Nat. Genet.* **48,** 189–194 (2016).