Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole genome sequencing studies

Han Chen,[1,2] Jennifer E. Huffman,[3] Jennifer A. Brody,[4] Chaolong Wang,[5] Seunggeun Lee,[6] Zilin Li,[7] Stephanie M. Gogarten,[8] Tamar Sofer,[9,10] Lawrence F. Bielak,[11] Joshua C. Bis,[4] John Blangero,[12] Russell P. Bowler,[13] Brian E. Cade,[9,10] Michael H. Cho,[14,15] Adolfo Correa,[16] Joanne E. Curran,[12] Paul S. de Vries,[1] David C. Glahn,[17,18] Xiuqing Guo,[19] Andrew D. Johnson,[20] Sharon Kardia,[11] Charles Kooperberg,[21] Joshua P. Lewis,[22] Xiaoming Liu,[1] Rasika A. Mathias,[23] Braxton D. Mitchell,[22,24] Jeffrey R. O'Connell,[22] Patricia A. Peyser,[11] Wendy S. Post,[25] Alex P. Reiner,[21] Stephen S. Rich,[26] Jerome I. Rotter,[19] Edwin K. Silverman,[14,15] Jennifer A. Smith,[11] Ramachandran S. Vasan,[20,27,28] James G. Wilson,[29] Lisa R. Yanek,[23] NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Hematology and Hemostasis Working Group, Susan Redline,[9,10,30] Nicholas L. Smith,[4,31-33] Eric Boerwinkle,[1,34] Ingrid B. Borecki,[8] L. Adrienne Cupples,[20,35] Cathy C. Laurie,[8] Alanna C. Morrison,[1] Kenneth M. Rice,[8] Xihong Lin[7,36]

[1] Human Genetics Center, Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA.

[2] Center for Precision Health, School of Public Health and School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA.

[3] Center for Population Genomics, Boston VA Healthcare System, Jamaica Plain, MA 02130, USA.

[4] Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA 98101, USA.

[5] Computational and Systems Biology, Genome Institute of Singapore, Singapore 138672, Singapore.

[6] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA.

[7] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA.

[8] Department of Biostatistics, University of Washington, Seattle, WA 98195, USA.

[9] Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, MA 02115, USA.

[10] Division of Sleep Medicine, Harvard Medical School, Boston, MA 02115, USA.

[11] Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA.

[12] Department of Human Genetics and South Texas Diabetes and Obesity Institute, School of Medicine, The University of Texas Rio Grande Valley, Brownsville, TX 78520, USA.

[13] Division of Pulmonary Medicine, Department of Medicine, National Jewish Health, Denver, CO 80206, USA.

[14] Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA.

[15] Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA.

[16] Jackson Heart Study, University of Mississippi Medical Center, Jackson, MS 39216, USA.

[17] Department of Psychiatry, Yale University School of Medicine, New Haven, CT 06510, USA.

[18] Olin Neuropsychiatric Research Center, Institute of Living, Hartford Hospital, Hartford, CT 06106, USA.

[19] The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA, 90502, USA.

[20] Framingham Heart Study, National Heart, Lung, and Blood Institute and Boston University, Framingham, MA 01702, USA.

[21] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA.

[22] Department of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201, USA.

[23] Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA.

[24] Geriatrics Research and Education Clinical Center, Baltimore VA Medical Center, Baltimore, MD 21201, USA.

[25] Division of Cardiology, Johns Hopkins University, Baltimore, MD 21287, USA.

[26] Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA.

[27] Sections of Preventive Medicine and Epidemiology, and of Cardiology, Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA.

[28] Department of Epidemiology, Boston University School of Public Health, Boston, MA 02118, USA.

[29] Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS 39216, USA.

[30] Division of Pulmonary, Critical Care, and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02115, USA.

[31] Kaiser Permanente Washington Health Research Institute, Seattle, WA 98101, USA.

[32] Seattle Epidemiologic Research and Information Center, Department of Veterans Affairs Office of Research and Development, Seattle, WA 98108, USA.

[33] Department of Epidemiology, University of Washington, Seattle, WA 98195, USA.

[34] Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA.

[35] Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA.

[36] Department of Statistics, Harvard University, Cambridge, MA 02138, USA

Correspondence should be addressed to X.L. (xlin@hsph.harvard.edu)

ABSTRACT

With advances in Whole Genome Sequencing (WGS) technology, more advanced statistical methods for testing genetic association with rare variants are being developed. Methods in which variants are grouped for analysis are also known as variant-set, gene-based, and aggregate unit tests. The burden test and Sequence Kernel Association Test (SKAT) are two widely used variant-set tests, which were originally developed for samples of unrelated individuals and later have been extended to family data with known pedigree structures. However, computationally-efficient and powerful variant-set tests are needed to make analyses tractable in large-scale WGS studies with complex study samples. In this paper, we propose the variant-Set Mixed Model Association Tests (SMMAT) for continuous and binary traits using the generalized linear mixed model framework. These tests can be applied to large-scale WGS studies involving samples with population structure and relatedness, such as in the National Heart, Lung, and Blood Institute's Trans-Omics for Precision Medicine (TOPMed) program. SMMAT tests share the same null model for different variant sets, and a virtue of this null model, which includes covariates only, is that it needs to be only fit once for all tests in each genome-wide analysis. Simulation studies show that all the proposed SMMAT tests correctly control type I error rates for both continuous and binary traits in the presence of population structure and relatedness. We also illustrate our tests in a real data example of analysis of plasma fibrinogen levels in the TOPMed program (n = 23,763), using the Analysis Commons, a cloud-based computing platform.

INTRODUCTION

In recent years, massive DNA sequence data have been generated. Large-scale whole genome sequencing projects, such as the National Heart, Lung, and Blood Institute's (NHLBI) Trans-Omics for Precision Medicine (TOPMed) program and the National Human Genome Research Institute's (NHGRI) Genome Sequencing Project (GSP), have produced whole genome sequences from over 120,000 samples. The designs of the studies from which participants are drawn need not be uniform or simple; for example, TOPMed includes population-based cohorts, family studies, and case-control studies, some of which are conducted in recently admixed populations, and some of which involve large pedigrees of closely-related participants.

In population-based cohorts and case-control studies, population stratification and cryptic relatedness are major sources of confounding that need to be accounted for in association tests. For common single variant analysis, linear mixed models that use an estimated genetic relationship matrix (GRM) to account for both population stratification and cryptic relatedness have been widely applied in Genome-Wide Association Studies (GWAS) to analyze structured and related samples.[1-6] For binary traits, however, we previously showed that linear mixed models may not be appropriate in the presence of population stratification due to misspecified mean-variance relationships. Therefore, we instead proposed a computationally efficient method GMMAT[7] to perform single common variant tests in GWAS by fitting generalized linear mixed models (GLMMs),[8] which simultaneously account for population structure, cryptic relatedness, and shared environmental effects, using multiple variance components and/or random effects.

Hundreds of millions of genetic variants, mostly with a low and extremely rare minor allele frequency (MAF), are being analyzed in large-scale sequencing projects such as TOPMed and GSP. Yet, single-variant tests that have been widely used in GWAS are generally underpowered for analyzing rare genetic variants from sequencing studies. To circumvent this problem, statistical tests such as the burden test,[9-12] Sequence Kernel Association Test (SKAT),[13] and their various combinations[14-16] have been proposed. These tests analyze multiple genetic variants in sets, grouped by genes, genomic regions, or other bioinformatic aggregation units. Most of these tests were originally developed to analyze samples from unrelated individuals, as well as extensions to analyze family data with known pedigree structures in the parametric mixed model and semiparametric generalized estimating equation frameworks.[17-23] However, these existing methods do not account for cryptic relatedness and have not been applied to large-scale whole genome sequencing studies with population structure, familial and/or cryptic relatedness, due to statistical and computational challenges.

One challenge is that among traditional variant set tests such as burden tests and SKAT, no single approach is uniformly most powerful. Another challenge is that existing hybrid tests that combine burden tests and SKAT, such as SKAT-O,[14] MiST[15] and aSPU,[16] are powerful but are subject to much greater computational loads than either the burden test or SKAT alone in the GLMM framework. Of note, SKAT-O is slower than SKAT because it searches on a grid for the optimal linear combination of the burden test and SKAT statistics. MiST requires adjusting for the genetic burden as a covariate in the SKAT model, and

hence needs to fit a burden model for each variant set. In large samples of possibly related individuals, extension of MiST is not as practical as in unrelated samples, since fitting a mixed effects model using the burden score for each variant set (or each test unit) is computationally intensive across the genome. Finally, aSPU uses a permutation or Monte Carlo simulation procedure to compute the p values, which can also be challenging in the context of large-scale whole genome sequencing studies with both population structure and relatedness. Therefore, there is a pressing need to develop powerful and computationally-efficient statistical methods for large-scale whole genome sequencing studies.

To address these statistical and computational challenges, we develop the variant Set Mixed Model Association Tests (SMMAT), computationally-efficient variant set tests for both continuous and binary traits, which are applicable to large-scale whole genome sequencing studies with structured and related samples. We include four tests in the SMMAT framework: the burden test (SMMAT-B), SKAT (SMMAT-S), SKAT-O (SMMAT-O), and an efficient hybrid test to combine the burden test and SKAT (SMMAT-E). All the four SMMAT tests share the same reduced model under the null hypothesis, i.e., the GLMM with only covariates, which only needs to be fit once for all genetic variant sets in an analysis. We show that all of these tests can be constructed using shared single-variant scores and their covariance matrices, thus further improving the computational efficiency in practice compared to performing these tests separately. Moreover, it has been shown that single-variant scores and their covariance matrices can also be used in the meta-analysis of variant set tests,[24, 25] thus SMMAT can be directly applied to combine multi-cohort studies ranging from unstructured independent samples, to structured and related

samples. Finally, we develop a unified analysis pipeline in our software package Generalized linear Mixed Model Association Tests (GMMAT) that implements SMMAT variant set tests in both single study (pooled analysis) and meta-analysis contexts to facilitate research on rare genetic variants from large-scale sequencing studies. We demonstrate the application of our method to the analysis of fibrinogen levels in the TOPMed study.

METHODS

Generalized Linear Mixed Models (GLMMs)

We formulate the SMMAT tests (SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E) from the same GLMM

$$g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{G}_i\boldsymbol{\beta} + b_i, \qquad \text{(Equation 1)}$$

where $g(\cdot)$ is a monotonic "link" function that connects the mean of phenotype y_i, denoted by $\mu_i = E(y_i|\boldsymbol{X}_i, \boldsymbol{G}_i, b_i)$, for subject i of n samples, to the covariate row vector $\boldsymbol{X}_i$, the genotype row vector $\boldsymbol{G}_i$ for q genetic variants in a set, and the random effects b_i that accounts for population structure and relatedness. The phenotypes y_i follow a distribution in the exponential family. For continuous traits, we usually assume y_i follow a normal distribution and use an identity link function; for binary traits, we assume y_i follow a Bernoulli distribution and use a logit link function. In Equation 1, $\boldsymbol{\alpha}$ is a $p \times 1$ vector of fixed covariate effects including an intercept, and the genotype effects $\boldsymbol{\beta}$ are assumed to be a $q \times 1$ vector whose distribution has mean $\boldsymbol{W}\mathbf{1}_q\beta_0$ and covariance $\theta\boldsymbol{W}^2$, where $\boldsymbol{W} = diag\{w_j\}$ is a pre-specified $q \times q$ matrix assigning weights to each variant, θ is a variance component parameter, and $\mathbf{1}_q$ is a column vector of length q with all elements 1. We

assume that $\boldsymbol{b} \sim N(\mathbf{0}, \sum_{k=1}^{K} \tau_k \boldsymbol{\Phi}_k)$ is an $n \times 1$ vector of random effects b_i, with variance component parameters τ_k and known $n \times n$ relatedness matrices $\boldsymbol{\Phi}_k$. We allow for multiple random effects to account for complex sampling designs such as hierarchical designs and shared environmental effects.

SMMAT-B, SMMAT-S and SMMAT-O

In Equation 1, testing the genotype effects of q variants H_0: $\boldsymbol{\beta} = \mathbf{0}$ is equivalent to testing the null hypothesis that H_0: $\beta_0 = 0$ and $\theta = 0$. The reduced GLMM under this null hypothesis specifies that

$$g(\mu_{0_i}) = \boldsymbol{X}_i \boldsymbol{\alpha} + b_i, \qquad \text{(Equation 2)}$$

where $\mu_{0_i} = E(y_i | \boldsymbol{X}_i, b_i)$. If we test H_0: $\beta_0 = 0$ under the assumption that $\theta = 0$, a burden score test SMMAT-B can be constructed as

$$T_B = \frac{(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}_0)^T \boldsymbol{G}\boldsymbol{W}\mathbf{1}_q \mathbf{1}_q^T \boldsymbol{W}\boldsymbol{G}^T (\boldsymbol{y} - \widehat{\boldsymbol{\mu}}_0)}{\hat{\phi}^2},$$

where $\boldsymbol{y} = (y_1 \quad y_2 \quad \cdots \quad y_n)^T$ is an $n \times 1$ vector of phenotypes y_i, $\widehat{\boldsymbol{\mu}}_0$ is a vector of fitted mean values under the model in Equation 2, $\boldsymbol{G} = (\boldsymbol{G}_1^T \quad \boldsymbol{G}_2^T \quad \cdots \quad \boldsymbol{G}_n^T)^T$ is an $n \times q$ genotype matrix of the variant set in the test, and $\hat{\phi}$ is an estimate of the dispersion parameter (or the residual variance) ϕ. Under H_0: $\beta_0 = 0$, the statistic T_B asymptotically follows $\xi_B \chi_1^2$, where the scalar $\xi_B = \mathbf{1}_q^T \boldsymbol{W}\boldsymbol{G}^T \widehat{\boldsymbol{P}} \boldsymbol{G}\boldsymbol{W}\mathbf{1}_q$, χ_1^2 is a chi-square distribution with 1 df, and $\widehat{\boldsymbol{P}} = \widehat{\boldsymbol{\Sigma}}^{-1} - \widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^T\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{\Sigma}}^{-1}$ is the $n \times n$ projection matrix of the null GLMM (Equation 2), $\boldsymbol{X} = (\boldsymbol{X}_1^T \quad \boldsymbol{X}_2^T \quad \cdots \quad \boldsymbol{X}_n^T)^T$ is an $n \times p$ covariate matrix, $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{V}} + \sum_{k=1}^{K} \hat{\tau}_k \boldsymbol{\Phi}_k$ with $\widehat{\boldsymbol{V}} = \hat{\phi}\boldsymbol{I}_n$ for continuous traits in linear mixed models, and $\widehat{\boldsymbol{V}} =$

$diag\left\{\frac{1}{\hat{\mu}_{0i}(1-\hat{\mu}_{0i})}\right\}$ for binary traits in logistic mixed models (where the dispersion parameter ϕ is known to be 1).

On the other hand, if we test $H_0: \theta = 0$ under the assumption $\beta_0 = 0$, a variance component score-type test SMMAT-S can be constructed as

$$T_S = \frac{(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0)^T \boldsymbol{GWWG}^T (\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0)}{\hat{\phi}^2}.$$

Under $H_0: \theta = 0$, T_S asymptotically follows $\sum_{j=1}^{q} \xi_{S_j} \chi^2_{1,j}$, where $\chi^2_{1,j}$ are independent chi-square distributions with 1 df, and ξ_{S_j} are the eigenvalues of $\boldsymbol{\Xi}_S = \boldsymbol{WG}^T \hat{\boldsymbol{P}} \boldsymbol{GW}$.

If one assumes β_0 has mean 0 and variance γ, $\boldsymbol{\beta}$ then follows a distribution 0 and covariance $\tau \boldsymbol{W}\{(1-\rho)\boldsymbol{I}_q + \rho \boldsymbol{1}_q \boldsymbol{1}_q^T\}\boldsymbol{W}$, where $\tau = \gamma + \theta$ and $\rho = \gamma/(\gamma + \theta)$, which takes values between 0 and 1. The joint null hypothesis $H_0: \beta_0 = 0$ and $\theta = 0$ is equivalent to $H_0: \tau = 0$. Given ρ, a variance component score-type test can be constructed as

$$T_\rho = \rho T_B + (1-\rho) T_S.$$

If $\rho = 1$, T_ρ becomes the SMMAT-B burden statistic T_B, which assumes $\boldsymbol{\beta}$ are the same for all q variants after weighting. If $\rho = 0$, T_ρ becomes the SMMAT-S SKAT statistic T_S. If an optimal ρ is obtained by minimizing the p-value of T_ρ, then SMMAT-O can be constructed, with its p value calculated using a one-dimensional numerical integration, following SKAT-O.[14] A key advantage of SMMAT-O is that it maximizes the power by using the optimal linear combination of the mixed model burden test SMMAT-B and the mixed model SKAT SMMAT-S. As it requires a grid search over ρ, it is computationally

considerably more expensive than SMMAT-B and SMMAT-S. We propose in the next section a computationally much more efficient method to combine SMMAT-B and SMMAT-S.

SMMAT-E

An alternative joint test to SMMAT-O for $H_0: \beta_0 = 0$ and $\theta = 0$ can be constructed using two asymptotically independent tests: a test for $H_0: \beta_0 = 0$ versus $H_1: \beta_0 \neq 0$ under the constraint $\theta = 0$, and a test for $H_0: \theta = 0$ versus $H_1: \theta > 0$ with β_0 as a nuisance parameter that is estimated under $H_0: \theta = 0$. In unrelated samples, this testing strategy is MiST,[15] which requires the burden model to be fit for each SNP set. We note that the first test is SMMAT-B T_B in the SMMAT framework, and the second test T_θ can be constructed from the null burden GLMM

$$g(\mu_{B_i}) = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{G}_i\boldsymbol{W}\boldsymbol{1}_q\beta_0 + b_i, \qquad \text{(Equation 3)}$$

where $\mu_{B_i} = E(y_i|\boldsymbol{X}_i, \boldsymbol{G}_i\boldsymbol{W}\boldsymbol{1}_q, b_i)$ is the mean of y_i in the burden GLMM. We can construct a SKAT-type statistic adjusting for the genetic burden

$$T_\theta = \frac{(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}_B)^T\boldsymbol{G}\boldsymbol{W}\boldsymbol{W}\boldsymbol{G}^T(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}_B)}{\tilde{\phi}^2},$$

where $\tilde{\boldsymbol{\mu}}_B$ is a vector of fitted values $\tilde{\mu}_{B_i}$ using the burden GLMM in Equation 3 for a given variant set. However, fitting this burden GLMM separately for each variant set is computationally expensive in large-scale whole-genome association studies.

Therefore, we propose a different computationally efficient strategy by assuming that the mean of genetic effects β_0 is not large, a reasonable assumption for most genomic regions

and most complex human diseases. Then we can construct T_θ efficiently without refitting the burden GLMMs in Equation 3 for each variant set across the genome. We show in the Appendix that T_θ can be approximated by

$$T_\theta \approx \hat{\phi}^{-2}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_0)^T\boldsymbol{GW}\left\{\boldsymbol{I}_q-\boldsymbol{1}_q\left(\boldsymbol{1}_q^T\boldsymbol{WG}^T\hat{\boldsymbol{P}}\boldsymbol{GW}\boldsymbol{1}_q\right)^{-1}\boldsymbol{1}_q^T\boldsymbol{WG}^T\hat{\boldsymbol{P}}\boldsymbol{GW}\right\}\left\{\boldsymbol{I}_q - \boldsymbol{WG}^T\hat{\boldsymbol{P}}\boldsymbol{GW}\boldsymbol{1}_q\left(\boldsymbol{1}_q^T\boldsymbol{WG}^T\hat{\boldsymbol{P}}\boldsymbol{GW}\boldsymbol{1}_q\right)^{-1}\boldsymbol{1}_q^T\right\}\boldsymbol{WG}^T(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_0).$$

Therefore, under H_0: $\theta = 0$, T_θ asymptotically approximately follows $\sum_{j=1}^{q}\xi_{\theta_j}\chi_{1,j}^2$, where $\chi_{1,j}^2$ are independent chi-square distributions with 1 df, and ξ_{θ_j} are the eigenvalues of $\boldsymbol{\Xi}_\theta = \boldsymbol{WG}^T\hat{\boldsymbol{P}}\boldsymbol{GW} - \boldsymbol{WG}^T\hat{\boldsymbol{P}}\boldsymbol{GW}\boldsymbol{1}_q\left(\boldsymbol{1}_q^T\boldsymbol{WG}^T\hat{\boldsymbol{P}}\boldsymbol{GW}\boldsymbol{1}_q\right)^{-1}\boldsymbol{1}_q^T\boldsymbol{WG}^T\hat{\boldsymbol{P}}\boldsymbol{GW}$. By the central limit theorem, both $\frac{\boldsymbol{WG}^T(\boldsymbol{y}-\tilde{\boldsymbol{\mu}}_B)}{\tilde{\phi}}$ and $\frac{\boldsymbol{1}_q^T\boldsymbol{WG}^T(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_0)}{\hat{\phi}}$ are asymptotically normal, and their covariance matrix is

$$Cov\left(\frac{\boldsymbol{WG}^T(\boldsymbol{y}-\tilde{\boldsymbol{\mu}}_B)}{\tilde{\phi}},\frac{\boldsymbol{1}_q^T\boldsymbol{WG}^T(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_0)}{\hat{\phi}}\right)$$

$$\approx \left\{\boldsymbol{I}_q-\boldsymbol{WG}^T\hat{\boldsymbol{P}}\boldsymbol{GW}\boldsymbol{1}_q\left(\boldsymbol{1}_q^T\boldsymbol{WG}^T\hat{\boldsymbol{P}}\boldsymbol{GW}\boldsymbol{1}_q\right)^{-1}\boldsymbol{1}_q^T\right\}\boldsymbol{WG}^T\hat{\boldsymbol{P}}\boldsymbol{GW}\boldsymbol{1}_q = \boldsymbol{0}.$$

Therefore, T_θ and T_B are approximately asymptotically independent. Let p_θ and p_B be the p value of the two tests respectively, then SMMAT-E p value p_E is computed using Fisher's method with a chi-square distribution with 4 df as $p_E = P(\chi_4^2 > -2\log(p_\theta p_B))$.

Meta-analysis

SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E can all be conducted in the meta-analysis context. Assuming the single-variant scores $\boldsymbol{S} = \frac{\boldsymbol{G}^T(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_0)}{\hat{\phi}}$ and their covariance matrix $\boldsymbol{\Psi} = \boldsymbol{G}^T\hat{\boldsymbol{P}}\boldsymbol{G}$ are computed for each variant set in each study, we can reconstruct

$T_B = \boldsymbol{S}^T\boldsymbol{W}\mathbf{1}_q\mathbf{1}_q^T\boldsymbol{W}\boldsymbol{S}$ with $\xi_B = \mathbf{1}_q^T\boldsymbol{W}\boldsymbol{\Psi}\boldsymbol{W}\mathbf{1}_q$; $T_S = \boldsymbol{S}^T\boldsymbol{W}\boldsymbol{W}\boldsymbol{S}$ with $\boldsymbol{\Xi}_S = \boldsymbol{W}\boldsymbol{\Psi}\boldsymbol{W}$; $T_\rho = \rho T_B + (1-\rho)T_S$ and $T_\theta = \boldsymbol{S}^T\boldsymbol{W}\left\{\boldsymbol{I}_q - \mathbf{1}_q\left(\mathbf{1}_q^T\boldsymbol{W}\boldsymbol{\Psi}\boldsymbol{W}\mathbf{1}_q\right)^{-1}\mathbf{1}_q^T\boldsymbol{W}\boldsymbol{\Psi}\boldsymbol{W}\right\}\left\{\boldsymbol{I}_q - \boldsymbol{W}\boldsymbol{\Psi}\boldsymbol{W}\mathbf{1}_q\left(\mathbf{1}_q^T\boldsymbol{W}\boldsymbol{\Psi}\boldsymbol{W}\mathbf{1}_q\right)^{-1}\mathbf{1}_q^T\right\}\boldsymbol{W}\boldsymbol{S}$ with $\boldsymbol{\Xi}_\theta = \boldsymbol{W}\boldsymbol{\Psi}\boldsymbol{W} - \boldsymbol{W}\boldsymbol{\Psi}\boldsymbol{W}\mathbf{1}_q\left(\mathbf{1}_q^T\boldsymbol{W}\boldsymbol{\Psi}\boldsymbol{W}\mathbf{1}_q\right)^{-1}\mathbf{1}_q^T\boldsymbol{W}\boldsymbol{\Psi}\boldsymbol{W}$.

For each variant set, let $m = 1, 2, \cdots, M$ be the index of studies, $\boldsymbol{S}_m$ and $\boldsymbol{\Psi}_m$ be the single-variant scores and covariance matrix from study m, in testing the "weak" null hypothesis[26] of summary genetic effects H_0: $\boldsymbol{\beta} = \mathbf{0}$,[24, 25] we can compute meta summary statistics $\boldsymbol{S} = \sum_{m=1}^{M}\boldsymbol{S}_m$ and $\boldsymbol{\Psi} = \sum_{m=1}^{M}\boldsymbol{\Psi}_m$ and use them in SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E. When combining studies with very different sample characteristics, testing the "strong" null hypothesis[26] that genetic effects in all studies are 0 is sometimes desired. In the general case, we may choose to group studies that are similar and test if the summary genetic effects in all groups are 0, for example, in the meta-analysis of multi-ethnic samples. Let $c = 1, 2, \cdots, C$ be a partition of M studies ($C \leq M$), where C is the number of ethnicities, $\boldsymbol{S}_{c_m}$ and $\boldsymbol{\Psi}_{c_m}$ be the single-variant scores and covariance matrix from study m in partition c ($m = 1, 2, \cdots, M_c$ in partition c, and $\sum_{c=1}^{C} M_c = M$), such that genetic effects for the same variant are summarized within each partition c but heterogeneous across partitions,[24] we can also compute summary statistics $\boldsymbol{S} = \left(\sum_{m=1}^{M_1}\boldsymbol{S}_{1_m}^T \quad \sum_{m=1}^{M_2}\boldsymbol{S}_{2_m}^T \quad \cdots \quad \sum_{m=1}^{M_C}\boldsymbol{S}_{C_m}^T\right)^T$ and $\boldsymbol{\Psi} = diag\left\{\sum_{m=1}^{M_c}\boldsymbol{\Psi}_{c_m}\right\}$. Note that $\boldsymbol{S}$ is now a vector of length Cq, and $\boldsymbol{\Psi}$ is a block-diagonal matrix with C blocks of $q \times q$ matrices, one for each partition of studies (with total dimension $Cq \times Cq$), we should

replace $\boldsymbol{W}$, $\mathbf{1}_q$ and $\boldsymbol{I}_q$ by $\boldsymbol{I}_C \otimes \boldsymbol{W}$ (where $\otimes$ denotes the Kronecker product), $\mathbf{1}_{Cq}$ and $\boldsymbol{I}_{Cq}$, respectively in the above expressions for T_B, T_ρ, T_S and T_θ for meta-analysis.

Simulation studies

Type I error in single-cohort studies

We performed coalescent simulations to generate sequence data with 100 genetic variants in each set, and 10,000 independent sets for 8,000 individuals from a 20 × 20 grid of spatially continuous populations with migration rate between adjacent cells $M = 10$ (Figure 1A). Within each cell, we paired 20 individuals into 10 families and simulated 2 children for each family using gene dropping,[27] and in total we had 4,000 families and 16,000 individuals. For continuous traits, in each simulation replicate, we simulated the phenotype y_{ij} for individual j in family i under the null hypothesis of no genetic association from

$$y_{ij} = \alpha_1 Z_i + b_{ij} + \epsilon_{ij}, \qquad \text{(Equation 4)}$$

where the "population effect" $\alpha_1 = 1$, the population indicator $Z_i = 1$ if family i was from a 10 × 10 grid in the top left of the map (Population 1), and $Z_i = 0$ otherwise (Population 2). The familial random effects were simulated as

$$\boldsymbol{b}_i = \begin{pmatrix} b_{i1} \\ b_{i2} \\ b_{i3} \\ b_{i4} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.5 & 0 & 0.25 & 0.25 \\ 0 & 0.5 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.5 \end{pmatrix} \right), \qquad \text{(Equation 5)}$$

and the random error $\epsilon_{ij} \sim N(0, 1)$ for each individual j in family i. Then we randomly sampled 3,500 individuals from the 10 × 10 grid in the top left, and 6,500 individuals from the rest of the map. The family identifier was removed for all individuals in the analysis,

so that there were both population structure and cryptic relatedness in the sample. We compared SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E in analyzing 10,000 independent variant sets based on a linear mixed model using our GMMAT package, including random effects with their covariance matrix proportional to the GRM, and adjusted for the first 10 principal components (PCs) of ancestry. We repeated this 4,000 times to get p values combined from 40 million independent genetic variant sets for each test.

For binary traits, in each simulation replicate, we simulated the phenotype y_{ij} for individual j in family i under the null hypothesis of no genetic association from

$$\log\left(\frac{P(y_{ij}=1)}{1-P(y_{ij}=1)}\right)=\alpha_0+b_{ij}, \qquad \text{(Equation 6)}$$

where α_0 was chosen such that the disease prevalence was 0.01 in all populations, and the familial random effects b_{ij} were simulated in the same way as for continuous traits. Then we randomly sampled 2,500 cases and 1,000 controls from the 10 × 10 grid in the top left (Population 1), and 2,500 cases and 4,000 controls from the rest of the map (Population 2) to form a hypothetical study with balanced cases and controls in combined populations. Therefore, there was confounding by population structure resulting from unequal sampling, even though the disease prevalence was the same. We removed the family identifier, compared SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E in analyzing 10,000 independent variant sets based on a logistic mixed model using our GMMAT package, similarly as described above, and repeated this 4,000 times to get p values combined from 40 million independent genetic variant sets for each test.

Type I error in meta-analysis

We also conducted simulation studies in the meta-analysis context to evaluate the type I error rates. We considered 4 scenarios: unrelated individuals, without confounding by population structure (Scenario A studies); related individuals, with confounding by population structure (Scenario B studies); unrelated individuals, with confounding by population structure (Scenario C studies); and related individuals, without confounding by population structure (Scenario D studies).

For Scenario A studies, we simulated 16 unrelated individuals in each cell from the 10 × 10 grid in the top left of the map (Figure 1B). For continuous traits, we simulated the phenotype y_{ij} from Equation 4, with $\alpha_1 = 0$ and $b_{ij} = 0$, and randomly sampled 1,000 individuals. For binary traits, we simulated y_{ij} from Equation 6, with $b_{ij} = 0$, and randomly sampled 500 cases and 500 controls.

For Scenario B studies, we simulated 8 unrelated individuals, paired them into 4 families and simulated 2 children for each family in each cell from the 10 × 10 grid in the center of the map (Figure 1B). For continuous traits, we simulated the phenotype y_{ij} from Equation 4, with $\alpha_1 = 1$, the population indicator $Z_i = 1$ if family i was from Population 1, and $Z_i = 0$ if from Population 2, and familial random effects b_{ij} were simulated using Equation 5, and we randomly sampled 350 individuals from Population 1 and 650 individuals from Population 2. For binary traits, we simulated y_{ij} from Equation 6, with b_{ij} from Equation 5, and randomly sampled 250 cases and 100 controls from Population 1, and 250 cases and 400 controls from Population 2.

For Scenario C studies, we simulated 16 unrelated individuals in each cell from the 20 × 5 grid in the top of the map (Figure 1B). For continuous traits, we simulated the phenotype y_{ij} from Equation 4, with $\alpha_1 = 1$, the population indicator $Z_i = 1$ if family i was from Population 1, and $Z_i = 0$ if from Population 2, and $b_{ij} = 0$, and we randomly sampled 350 individuals from Population 1 and 650 individuals from Population 2. For binary traits, we simulated y_{ij} from Equation 6, with $b_{ij} = 0$, and randomly sampled 250 cases and 100 controls from Population 1, and 250 cases and 400 controls from Population 2.

For Scenario D studies, we simulated 8 unrelated individuals, paired them into 4 families and simulated 2 children for each family in each cell from the 20 × 5 grid in the bottom of the map (Figure 1B). For continuous traits, we simulated the phenotype y_{ij} from Equation 4, with $\alpha_1 = 0$, familial random effects b_{ij} simulated using Equation 5, and we randomly sampled 1,000 individuals. For binary traits, we simulated y_{ij} from Equation 6, with b_{ij} from Equation 5, and randomly sampled 500 cases and 500 controls.

In each simulation replicate, we simulated 3 studies from each scenario, totaling 12 studies with a combined sample size of 12,000 (6,000 cases and 6,000 controls for binary traits). We compared SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E using two meta-analysis strategies: all studies in the same group, and Scenario A, B, C, D studies in 4 separate groups. In the latter case, 3 studies from the same scenario were grouped in the same partition with shared genetic effects, while studies from different scenarios were

allowed to have heterogeneous genetic effects. We repeated 4,000 simulation replicates to get p values from 40 million independent genetic variant sets.

Power

We used the same genotype data as in the single-cohort type I error simulations and evaluated the empirical power of SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E (with weights equal to a beta distribution density function with parameters 1 and 25 on the MAF of each variant[13]) in 9 scenarios, with the proportion of causal variants in a test unit ranging from 10%, 20% to 50%, and the proportion of variants with negative effects out of causal variants ranging from 100%, 80% to 50%. For continuous traits, we simulated the phenotype y_{ij} for individual j in family i from

$$y_{ij} = \alpha_1 Z_i + \sum_l G_{ijl}\beta_l + b_{ij} + \epsilon_{ij},$$

where $\alpha_1 = 1$, the population indicator $Z_i = 1$ if family i was from Population 1, and $Z_i = 0$ if from Population 2, g_{ijl} was the centered genotype for causal variant l of individual j in family i, the causal effect size was $|\beta_l| = c|\log_{10} MAF_l|$ for variant l with MAF_l, where the constant c was set to 0.2, 0.1 and 0.05 when the proportion of causal variants was 10%, 20% and 50%, the familial random effects b_{ij} were simulated using Equation 5, and the random error $\epsilon_{ij} \sim N(0, 1)$. We randomly sampled 35% individuals from Population 1, and 65% individuals from Population 2.

For binary traits, we simulated the phenotype y_{ij} for individual j in family i from

$$\log\left(\frac{P(y_{ij} = 1)}{1 - P(y_{ij} = 1)}\right) = \alpha_0 + \sum_l G_{ijl}\beta_l + b_{ij},$$

where α_0 was chosen such that the disease prevalence was 0.01 in all populations, G_{ijl} was the centered genotype for causal variant l of individual j in family i, the causal effect size was $|\beta_l| = c|\log_{10} MAF_l|$ for variant l with MAF_l, where the constant c was set to 0.3, 0.2 and 0.1 when the proportion of causal variants was 10%, 20% and 50%, the familial random effects b_{ij} were simulated using Equation 5. We randomly sampled 35% individuals (with 25% cases and 10% controls out of the total sample size) from Population 1, and 65% individuals (with 25% cases and 40% controls out of the total sample size) from Population 2 to form a hypothetical study with balanced cases and controls in combined populations.

For both continuous and binary traits, we varied the total sample size from 2,000, 5,000 to 10,000, repeated 1,000 simulation replicates for each scenario under the alternative hypothesis, and compared the empirical power at the significance level of 2.5×10^{-6}.

TOPMed example involving fibrinogen levels

Samples with both plasma fibrinogen measures and whole genome sequence data (Freeze 5b) from the following 11 TOPMed studies were included in the analysis: the Old Order Amish Study (Amish), Cleveland Family Study (CFS), Genetic Epidemiology of COPD Study (COPDGene), Framingham Heart Study (FHS), Jackson Heart Study (JHS), San Antonio Family Study (SAFS), the Atherosclerosis Risk in Communities (ARIC) Study, Genetic Studies of Atherosclerosis Risk (GeneSTAR), Genetic Epidemiology Network of Arteriopathy (GENOA), the Multi-Ethnic Study of Atherosclerosis (MESA), and Women's Health Initiative (WHI). The TOPMed studies were approved by institutional review boards at participating institutions, and informed consent was obtained from all

study participants. Amish, CFS, FHS, JHS, and SAFS are family-based studies with differing degrees of relatedness. The total sample size was 23,763. Within each study and each ethnicity, measured fibrinogen levels were adjusted for age, sex, study-specific covariates, and the residuals were rank normalized and rescaled by multiplying by the original standard deviation, so that the transformed phenotype data have the same variances as on the original scale. The transformed phenotype data were pooled together in the analysis, using a heteroscedastic linear mixed model[28] allowing for different residual variances in each study/ethnicity, adjusting for study, ethnicity, sequence center, top 10 ancestry PCs[29] as fixed-effects covariates, and including a GRM calculated by Mixed Model Analysis for Pedigrees and Populations (MMAP) to model the random effects for relatedness. Rare and low frequency genetic variants on chromosome 4 with MAF less than 5% were tested for association with fibrinogen levels in a sliding window analysis[30] of 4 kb non-overlapping windows, using SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E with weights equal to a beta distribution density function with parameters 1 and 25 on the MAF of each variant[13]. The analysis was performed using the GMMAT App (version 0.9.2) with 32 parallel threads on a single computing node with 240 GB total memory in the Analysis Commons.[31] To benchmark the computational speed in running SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E, we also ran re-analyses to perform each test separately, using summary statistics from the sliding window analysis and a single thread on a computing node with 15 GB total memory in the Analysis Commons.

RESULTS

Simulation studies

Table 1 shows the empirical type I error rates of SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E at significance levels of 0.05, 0.0001, and 2.5×10^{-6}, in the variant set analyses of continuous and binary traits in single-cohort simulation studies. All 4 tests have well-controlled type I error rates at these significance levels, suggesting that GLMMs can be effective in adjusting for population structure and cryptic relatedness in complex study samples. This is also consistent with the quantile-quantile (QQ) plots in Figure 2, which show neither inflation nor deflation in the tail.

Table 2 and Figure 3 show simulation results of SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E assuming all studies in the same group (hom) or in 4 separate groups (het) in meta-analyses for combining 4 types of studies: with and without confounding by population structure, with and without cryptic relatedness. We note that SMMAT-B statistic T_B has the same form in these two meta-analysis strategies,[24] therefore, we included 7 tests in the simulation studies. In het SMMAT-S, SMMAT-O and SMMAT-E, studies from the same scenario were grouped together to assume shared genetic effects. Under the null hypothesis of no genetic associations, hom SMMAT-O shows very mild inflation in our simulation settings, but all other 6 tests in the SMMAT framework control type I error rates well at significance levels of 0.05, 0.0001, and 2.5×10^{-6} and have well-calibrated tail probabilities, for both continuous and binary traits.

Figures 4 and 5 present the empirical power for causal variant sets at the significance level of 2.5×10^{-6} for continuous and binary traits, respectively. The power increases with the sample size. As the proportion of causal variants with effects in the same direction drops

from 100%, 80% to 50% in each row, the power drops for all tests, but most substantially for the burden test SMMAT-B. When the sample size is large (i.e., 10,000 samples), SMMAT-E has the highest power, for both continuous and binary traits in all 9 simulation scenarios.

TOPMed example involving fibrinogen levels

We compared the results from SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E in an analysis of fibrinogen levels, using chromosome 4 (including the genomic region that encodes the fibrinogen protein, *FGB*) whole genome sequence data from 11 TOPMed studies. Previous studies have reported two rare variants within *FGB* on chromosome 4, rs6054 (hg38 position 154,568,456) and rs201909029 (hg 38 position 154,567,636) associated with lower fibrinogen levels, with similar effect sizes in all ancestry groups.[32] In the sliding window analysis, we grouped low frequency and rare genetic variants with MAF less than 5% into 46,859 non-overlapping 4 kb windows containing at least one variant. The number of variants in each window passing the MAF filter ranged from 1 to 1,290, with a median of 351 (25% quartile 326 and 75% quartile 380). The QQ plot (Figure 6A) shows that all 4 tests have well-calibrated tail probabilities. Table 3 summarizes heteroscedastic linear mixed model-based SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E p values in *FGB* and flanking regions. SMMAT-S, SMMAT-O and SMMAT-E give the most significant results in the 4 kb window 154,554 – 154,558 kb, with p values 1.6×10^{-17}, 8.9×10^{-17}, and 6.2×10^{-19}, respectively, while SMMAT-B p value is much larger (6.9×10^{-5}). In the 4 kb window that covers both known association rare variants rs6054 and rs201909029 (window 154,566 – 154,570 kb), SMMAT-E gives the smallest p

value 3.1×10^{-17}), followed by SMMAT-S (p value 9.7×10^{-17}), SMMAT-O (p value 3.3×10^{-16}) and SMMAT-B (p value 1.6×10^{-8}).

Computation time

Table 4 shows the CPU time for running the sliding window analysis for 23,763 individuals with TOPMed whole genome sequence data and fibrinogen levels, using summary statistics from 46,859 non-overlapping 4 kb windows on chromosome 4. The GMMAT App (version 0.9.2) in the Analysis Commons cloud computing platform has implemented SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E, with the option of running one or more tests in an analysis. SMMAT-B results are automatically included when running SMMAT-O or SMMAT-E, and SMMAT-S p values will also be output when running SMMAT-O. Of the four tests in Table 4, SMMAT-B takes shortest time as the p value calculation does not involve any eigen-decomposition of covariance matrices. SMMAT-S takes only about 10 minutes longer than SMMAT-B for the eigen-decomposition of 46,859 covariance matrices. SMMAT-E takes about 12 minutes longer than SMMAT-S and gives both SMMAT-B and SMMAT-E p values. SMMAT-O takes 175 minutes longer than SMMAT-S, as more eigen-decompositions are performed in SMMAT-O when it searches for the optimal combination of SMMAT-B and SMMAT-S on a grid of ρ values.

DISCUSSION

We have developed and implemented SMMAT, a family of computationally-efficient variant set mixed model association tests for continuous and binary traits in large-scale whole genome sequencing studies. This framework includes extensions of three widely

used variant set tests for unrelated individuals to complex study samples with population structure and cryptic relatedness: the burden test (SMMAT-B), SKAT (SMMAT-S) and SKAT-O (SMMAT-O), as well as a new efficient hybrid test that combines the mixed model burden and SKAT tests (SMMAT-E). Specifically, SMMAT-E is constructed by combining the burden test and an adjusted mixed model SKAT statistic that is approximately asymptotically independent from the mixed model burden test statistic, in a similar spirit to MiST in non-mixed model setting,[15] but that differs from MiST in that it does not require fitting separate mixed effect burden models for each variant set with the set genetic burden as a fixed-effects covariate. Instead, we use matrix projections to approximate the adjusted SKAT statistic from a global null model without any fixed effects for the variant set-specific genetic burden. Of note, this global null model only needs to be fit once in a whole genome analysis, which greatly reduces the computational cost. We show in simulation studies and the TOPMed fibrinogen example that SMMAT-E is more powerful than the other three tests in large samples, at the computational cost almost on the same scale of SMMAT-B and SMMAT-S. Therefore, SMMAT-E is recommended in the analysis of large-scale whole genome sequencing studies.

In the SMMAT framework, different weighting strategies can be used. One can use a function of the MAF,[11, 13] or external measures based on functional annotation such as CADD,[33] Eigen,[34] FATHMM-XF,[35] or tissue-specific annotations, such as GENOSKYLINE,[36] as the weight for each variant in a set. In the analysis of fibrinogen levels in TOPMed, we used MAF-based weights. Recently, unified variant set tests allowing for multiple functional annotations have been developed,[37] and the SMMAT

framework can possibly be extended to accommodate multiple weights. Nevertheless, the optimal weighting strategy in rare variant analysis remains an open question and an active field of research.

As SMMAT-E combines the burden test p value p_B with an asymptotically independent adjusted SKAT p value p_θ using Fisher's method in our SMMAT implementation in the GMMAT App, we note that other forms of combinations may also be applied.[38] For example, previous studies have shown that Tippett's procedure based on the minimum of p_θ and p_B might be more powerful than Fisher's method in MiST when only one of the p values is small.[15] Alternatively, instead of combining the p values, weighted linear combinations of chi-square statistics have been proposed,[39-41] and they can also be applied to combine the burden test statistic T_B and the asymptotically independent SKAT statistic T_θ in the SMMAT framework.

SMMAT also has some limitations. SMMAT p values are computed based on asymptotic distributions, which may be not be accurate in small samples, especially for binary traits and heavily skewed continuous traits. For continuous traits, small-sample inference procedures have been proposed for SKAT,[42, 43] and the same methodology can be applied to SMMAT. For ultra-rare genetic variants with very low minor allele counts, the single-variant scores used to construct SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E may not be close to a normal distribution, even if the total sample size is large. If there are only ultra-rare variants (e.g. singletons, doubletons) in a test region and the number of variants is small, SMMAT-B might be the best analysis strategy as its asymptotic property depends

on the cumulative minor allele counts. Moreover, the asymptotic issue of single-variant scores also exists for binary traits with highly unbalanced case-control ratios, and a saddlepoint approximation approach has been proposed to match the cumulant generating function of the single-variant scores,[44] and it has recently been extended to GLMMs.[45]

Fitting GLMMs with a GRM has $O(n^3)$ complexity in general, where n is the sample size. We have overcome this computational challenge by fitting only one GLMM in a whole genome analysis, and using matrix multiplications with $O(n^2)$ complexity for each variant set in SMMAT. In large-scale whole genome sequencing studies, solutions to other computational challenges are being proposed. For example, when the number of variants q in SKAT is very large, eigendecomposition of the covariance matrix, which has $O(\min(n,q)^3)$ complexity, could be computationally expensive. Recently, the fastSKAT approach has been proposed to efficiently approximate the null distribution of SKAT when q is very large,[46] and the same strategy can be applied to speed up SMMAT p value calculation for very large q. On the other hand, as the sample size in ongoing large-scale sequencing projects such as TOPMed eventually expands to hundreds of thousands, using a full $n \times n$ GRM would not be computationally practical in pooled analyses, as it may take several weeks to fit even only one GLMM with $O(n^3)$ complexity, and $O(n^2)$ memory footprint. Meta-analyses may be a more appealing analysis strategy in that situation by combining summary statistics from study-specific or ancestry-specific analyses. Essentially equivalently, in pooled analyses, using a sparse and/or block-diagonal GRM with each block corresponding to an individual study in meta-analyses, will help reduce the computational cost in fitting GLMMs, providing one uses specialized routines

for manipulation of sparse matrices.[47] Although whole genome sequencing studies have not yet been conducted in large biobanks with sample sizes on the scale of millions of individuals, it is expected that calculating the GRM itself would become a major computational bottleneck. Recently, GRM-free mixed effects models such as BOLT-LMM[6, 48] and SAIGE[45] have been developed for single variant tests, and we note that extension of these methods to the SMMAT framework will further reduce the computational cost in biobank-scale whole genome sequencing studies in the future.

In summary, SMMAT provides a flexible and practical statistical framework for large-scale whole genome sequencing studies with complex study samples, with balanced power and computational performance. With continuing advances in technology, lowering cost and development of new analytical methods, large-scale whole genome sequencing studies will facilitate human genetic research and enhance our understandings on complex diseases and traits.

Appendix: Approximations in SMMAT-E

Here we derive the approximations used in SMMAT-E to construct the SKAT-type statistic adjusting for the genetic burden

$$T_\theta = \frac{(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}_B)^T \boldsymbol{G}\boldsymbol{W}\boldsymbol{W}\boldsymbol{G}^T(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}_B)}{\tilde{\phi}^2}.$$

Let $\tilde{\phi}$, $\tilde{\boldsymbol{\alpha}}$, $\tilde{\beta}_0$, $\tilde{b}_i$, $\tilde{\boldsymbol{V}}$ and $\tilde{\boldsymbol{\Sigma}}$ be estimates for ϕ, $\boldsymbol{\alpha}$, β_0, b_i, $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$ respectively from the burden GLMM (Equation 3), we define $\tilde{\boldsymbol{Y}} = \boldsymbol{y}$ as the phenotype vector for continuous traits, and the "working vector" with components $\tilde{Y}_i = \boldsymbol{X}_i\tilde{\boldsymbol{\alpha}} + \boldsymbol{G}_i\boldsymbol{W}\boldsymbol{1}_q\tilde{\beta}_0 + \tilde{b}_i + \{\tilde{\mu}_{B_i}(1 -$

$\tilde{\mu}_{B_i}\big)\big\}^{-1}\big(y_i - \tilde{\mu}_{B_i}\big)$ at convergence of the logistic burden mixed model for binary traits (Equation 3), where $\tilde{\boldsymbol{\alpha}}$, $\tilde{\beta}_0$, $\tilde{b}_i$ are fixed-effects and random-effects estimates from the burden GLMM. We have

$$\frac{\boldsymbol{y} - \tilde{\boldsymbol{\mu}}_B}{\tilde{\phi}} = \tilde{\boldsymbol{V}}^{-1}\big(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\tilde{\boldsymbol{\alpha}} - \boldsymbol{GW}\mathbf{1}_q\tilde{\beta}_0 - \tilde{\boldsymbol{b}}\big)$$

$$= \tilde{\boldsymbol{\Sigma}}^{-1}\big(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\tilde{\boldsymbol{\alpha}} - \boldsymbol{GW}\mathbf{1}_q\tilde{\beta}_0\big)$$

$$= \tilde{\boldsymbol{\Sigma}}^{-1}\left\{\tilde{\boldsymbol{Y}} - (\boldsymbol{X} \quad \boldsymbol{GW}\mathbf{1}_q)\begin{pmatrix} \boldsymbol{X}^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{GW}\mathbf{1}_q \\ \mathbf{1}_q^T\boldsymbol{W}\boldsymbol{G}^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X} & \mathbf{1}_q^T\boldsymbol{W}\boldsymbol{G}^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{GW}\mathbf{1}_q \end{pmatrix}^{-1}\begin{pmatrix} \boldsymbol{X}^T\tilde{\boldsymbol{\Sigma}}^{-1} \\ \mathbf{1}_q^T\boldsymbol{W}\boldsymbol{G}^T\tilde{\boldsymbol{\Sigma}}^{-1} \end{pmatrix}\tilde{\boldsymbol{Y}}\right\}$$

$$= \left\{\tilde{\boldsymbol{\Sigma}}^{-1} - \tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\big(\boldsymbol{X}^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\tilde{\boldsymbol{\Sigma}}^{-1}\right\}\tilde{\boldsymbol{Y}} - \left\{\tilde{\boldsymbol{\Sigma}}^{-1} - \tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\big(\boldsymbol{X}^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\tilde{\boldsymbol{\Sigma}}^{-1}\right\}\boldsymbol{GW}\mathbf{1}_q$$

$$\left[\mathbf{1}_q^T\boldsymbol{W}\boldsymbol{G}^T\left\{\tilde{\boldsymbol{\Sigma}}^{-1} - \tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\big(\boldsymbol{X}^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\tilde{\boldsymbol{\Sigma}}^{-1}\right\}\boldsymbol{GW}\mathbf{1}_q\right]^{-1}\mathbf{1}_q^T\boldsymbol{W}\boldsymbol{G}^T\left\{\tilde{\boldsymbol{\Sigma}}^{-1} -$$

$$\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\big(\boldsymbol{X}^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\tilde{\boldsymbol{\Sigma}}^{-1}\right\}\tilde{\boldsymbol{Y}}.$$

Note that $\tilde{\phi} = 1$ for binary traits. Moreover, since the true value of β_0 is small, assuming including the genetic burden $\boldsymbol{G}_i\boldsymbol{W}\mathbf{1}_q$ in the second term in Equation 3 does not dramatically change the variance component estimates for τ_k and ϕ (and for binary traits, also the "working vector" $\tilde{\boldsymbol{Y}}$ at convergence of the model from Equation 2), we have the approximation $\tilde{\boldsymbol{\Sigma}}^{-1} - \tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\big(\boldsymbol{X}^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\tilde{\boldsymbol{\Sigma}}^{-1} \approx \hat{\boldsymbol{P}}$ and $\frac{\boldsymbol{y}-\hat{\boldsymbol{\mu}}_0}{\hat{\phi}} \approx \hat{\boldsymbol{P}}\tilde{\boldsymbol{Y}}$, then

$$\frac{\boldsymbol{W}\boldsymbol{G}^T(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}_B)}{\tilde{\phi}} \approx \boldsymbol{W}\boldsymbol{G}^T\left\{\hat{\boldsymbol{P}}\tilde{\boldsymbol{Y}} - \hat{\boldsymbol{P}}\boldsymbol{GW}\mathbf{1}_q\big(\mathbf{1}_q^T\boldsymbol{W}\boldsymbol{G}^T\hat{\boldsymbol{P}}\boldsymbol{GW}\mathbf{1}_q\big)^{-1}\mathbf{1}_q^T\boldsymbol{W}\boldsymbol{G}^T\hat{\boldsymbol{P}}\tilde{\boldsymbol{Y}}\right\}$$

$$\approx \left\{\boldsymbol{I}_q - \boldsymbol{W}\boldsymbol{G}^T\hat{\boldsymbol{P}}\boldsymbol{GW}\mathbf{1}_q\big(\mathbf{1}_q^T\boldsymbol{W}\boldsymbol{G}^T\hat{\boldsymbol{P}}\boldsymbol{GW}\mathbf{1}_q\big)^{-1}\mathbf{1}_q^T\right\}\frac{\boldsymbol{W}\boldsymbol{G}^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0)}{\hat{\phi}}.$$

Therefore,

$$T_\theta = \frac{(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}_B)^T\boldsymbol{GWW}\boldsymbol{G}^T(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}_B)}{\tilde{\phi}^2}$$

$$\approx \hat{\phi}^{-2}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_0)^T\boldsymbol{GW}\left\{\boldsymbol{I}_q-\boldsymbol{1}_q\left(\boldsymbol{1}_q^T\boldsymbol{WG}^T\hat{\boldsymbol{P}}\boldsymbol{GW}\boldsymbol{1}_q\right)^{-1}\boldsymbol{1}_q^T\boldsymbol{WG}^T\hat{\boldsymbol{P}}\boldsymbol{GW}\right\}\left\{\boldsymbol{I}_q - \boldsymbol{WG}^T\hat{\boldsymbol{P}}\boldsymbol{GW}\boldsymbol{1}_q\left(\boldsymbol{1}_q^T\boldsymbol{WG}^T\hat{\boldsymbol{P}}\boldsymbol{GW}\boldsymbol{1}_q\right)^{-1}\boldsymbol{1}_q^T\right\}\boldsymbol{WG}^T(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_0).$$

Supplemental Data

Supplemental Data include the full authorship list with affiliations of the Trans-Omics for Precision Medicine (TOPMed) Consortium.

Declaration of Interests

The authors declare no competing interests.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grants R00 HL130593 (to H.C.), U01 HL120393 (to H.C. and J.E.H.), and R35 CA197449, P01-CA134294, U01-HG009088, U19-CA203654, and R01-HL113338 (to X.L.). The authors acknowledge the Texas Advanced Computing Center (TACC, http://www.tacc.utexas.edu) at The University of Texas at Austin for providing High Performance Computing (HPC) resources that have contributed to the research results reported within this paper. Whole genome sequence analysis of fibrinogen levels in TOPMed was performed in the Analysis Commons on DNAnexus, a hosting platform that uses Amazon Web Services (AWS) to provide a cloud data management and computing environment for large genomic data projects. Phenotype harmonization and aggregation of the fibrinogen levels across TOPMed studies were supported in part by R01 HL139553. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the

National Heart, Lung, and Blood Institute, the National Institutes of Health or the U.S. Department of Health and Human Services.

TOPMed. Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Genetics of Cardiometabolic Health in the Amish" (phs000956.v1.p1) was performed at the Broad Institute of MIT and Harvard (3R01HL121007-01S1). WGS for "NHLBI TOPMed: The Cleveland Family Study" (phs000954.v1.p1) was performed at the University of Washington Northwest Genomics Center (3R01HL098433-05S1). WGS for "NHLBI TOPMed: Genetic Epidemiology of COPD (COPDGene)" (phs000951.v1.p1) was performed at the University of Washington Northwest Genomics Center (3R01HL089856-08S1). WGS for "NHLBI TOPMed: The Framingham Heart Study" (phs000974.v1.p1) was performed at the Broad Institute of MIT and Harvard (HHSN268201500014C). WGS for "NHLBI TOPMed: The Jackson Heart Study" (phs000964.v1.p1) was performed at the University of Washington Northwest Genomics Center (HHSN268201100037C). WGS for "NHLBI TOPMed: San Antonio Family Study" (phs001215.v1.p1) was performed at Illumina Genomic Service (R01 HL113323). WGS for "NHLBI TOPMed: Atherosclerosis Risk in Communities" (phs001211.v1.p1) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201500015C and 3U54HG003273-12S2) and the Broad Institute of MIT and Harvard (3R01HL092577-06S1). WGS for "NHLBI TOPMed: Genetic Studies of Atherosclerosis Risk (GeneSTAR)" (phs001218.v1.p1) were performed at Illumina Inc., Macrogen Corp., and the Broad Institute of MIT and Harvard

(3R01HL121007-01S1, HHSN268201500014C, 3R01HL092577-06S1). WGS for "NHLBI TOPMed: Genetic Epidemiology Network of Arteriopathy" (phs001345.v1.p1) was performed at the University of Washington Northwest Genomics Center for the HyperGen/GENOA project (HL055673, PI: Donna Arnett/HL119443, PI: Sharon Kardia) and the Broad Institute of MIT and Harvard for the AA-CAC project (PI: Kent Taylor). WGS for "NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA)" (phs001416.v1.p1) was performed at the Broad Institute of MIT and Harvard (3U54HG003067-13S1). WGS for "NHLBI TOPMed: Women's Health Initiative" (phs001237.v1.p1) was performed at the Broad Institute of MIT and Harvard (HHSN268201500014C). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

Old Order Amish Study (Amish). The Amish studies upon which these data are based were supported by NIH grants R01 AG18728, U01 HL072515, R01 HL088119, R01 HL121007, and P30 DK072488. See publication: PMID: 18440328.

Cleveland Family Study (CFS). Support for the Cleveland Family Study was provided by NHLBI grants R01 HL46380, R01 HL113338, and 1R35HL135818.

Genetic Epidemiology of COPD Study (COPDGene). This research used data generated by the COPDGene study, which was supported by NIH grants U01 HL089856 and U01 HL089897 from the National Heart, Lung, and Blood Institute. The COPDGene project is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Novartis, Pfizer, Siemens and Sunovion. A full listing of COPDGene investigators can be found at: http://www.copdgene.org/directory.

Framingham Heart Study (FHS). The Framingham Heart Study has been supported by contracts N01-HC-25195 and HHSN268201500001I and grant R01 HL092577. Fibrinogen measurement was supported by NIH R01-HL-48157. The Framingham Heart Study thanks the study participants and the multitude of investigators who over its 70 year history continue to contribute so much to further our knowledge of heart, lung, blood and sleep disorders and associated traits.

Jackson Heart Study (JHS). The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201300049C and HHSN268201300050C), Tougaloo College (HHSN268201300048C), and the University of Mississippi Medical Center (HHSN268201300046C and HHSN268201300047C) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute for Minority Health and Health Disparities (NIMHD). The authors also wish to thank the staffs and participants of the JHS.

San Antonio Family Study (SAFS). Collection of the San Antonio Family Study data was supported in part by National Institutes of Health (NIH) grants R01 HL045522, MH078143, MH078111 and MH083824; and whole genome sequencing of SAFS subjects was supported by U01 DK085524 and R01 HL113323. We are very grateful to the participants of the San Antonio Family Study for their continued involvement in our research programs.

Atherosclerosis Risk in Communities Study (ARIC). The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services (contract numbers HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I and HHSN268201700005I). The authors thank the staff and participants of the ARIC study for their important contributions.

Genetic Studies of Atherosclerosis Risk (GeneSTAR). GeneSTAR was supported by grants from the National Institutes of Health/National Heart, Lung, and Blood Institute (U01 HL72518, HL087698, HL112064) and by a grant from the National Institutes of Health/National Center for Research Resources (M01-RR000052) to the Johns Hopkins General Clinical Research Center.

Genetic Epidemiology Network of Arteriopathy (GENOA). Support for GENOA was also provided by the National Heart, Lung and Blood Institute (HL054457, HL054464,

HL054481, and HL087660) of the National Institutes of Health. We would also like to thank the GENOA participants.

Multi-Ethnic Study of Atherosclerosis (MESA). MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

Women's Health Initiative (WHI). The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C. The authors thank the WHI investigators and staff for their dedication and the study participants for making the program possible. A full listing of WHI investigators can be found at: http://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Long%20List.pdf.

WEB RESOURCES

The URLs for data presented herein are as follows:

Analysis Commons, http://analysiscommons.com/

DNAnexus, https://www.dnanexus.com/

GMMAT, https://github.com/hanchenphd/GMMAT

MMAP, https://github.com/MMAP

REFERENCES

1. Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., Eskin, E. (2008). Efficient control of population structure in model organism association mapping. Genetics 178, 1709-1723.

2. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42, 348-354.

3. Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. Nat. Methods 8, 833-835.

4. Zhou, X., Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44, 821-824.

5. Pirinen, M., Donnelly, P., Spencer, C.C.A. (2013). Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. The Annals of Applied Statistics 7, 369-390.

6. Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjalmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B. et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. 47, 284-290.

7. Chen, H., Wang, C., Conomos, M.P., Stilp, A.M., Li, Z., Sofer, T., Szpiro, A.A., Chen, W., Brehm, J.M., Celedon, J.C. et al. (2016). Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. Am. J. Hum. Genet. 98, 653-666.

8. Breslow, N.E., Clayton, D.G. (1993). Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association 88, 9-25.

9. Morgenthaler, S., Thilly, W.G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat. Res. 615, 28-56.

10. Li, B., Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. 83, 311-321.

11. Madsen, B.E., Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 5, e1000384.

12. Morris, A.P., Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet. Epidemiol. 34, 188-193.

13. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. 89, 82-93.

14. Lee, S., Wu, M.C., Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13, 762-775.

15. Sun, J., Zheng, Y., Hsu, L. (2013). A unified mixed-effects model for rare-variant association in sequencing studies. Genet. Epidemiol. 37, 334-344.

16. Pan, W., Kim, J., Zhang, Y., Shen, X., Wei, P. (2014). A powerful and adaptive association test for rare variants. Genetics 197, 1081-1095.

17. Schifano, E.D., Epstein, M.P., Bielak, L.F., Jhun, M.A., Kardia, S.L., Peyser, P.A., Lin, X. (2012). SNP set association analysis for familial data. Genet. Epidemiol. 36, 797-810.

18. Chen, H., Meigs, J.B., Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. Genet. Epidemiol. 37, 196-204.

19. Oualkacha, K., Dastani, Z., Li, R., Cingolani, P.E., Spector, T.D., Hammond, C.J., Richards, J.B., Ciampi, A., Greenwood, C.M. (2013). Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. Genet. Epidemiol. 37, 366-376.

20. Wang, X., Lee, S., Zhu, X., Redline, S., Lin, X. (2013). GEE-based SNP set association test for continuous and discrete traits in family-based association studies. Genet. Epidemiol. 37, 778-786.

21. Jiang, D., McPeek, M.S. (2014). Robust rare variant association testing for quantitative traits in samples with related individuals. Genet. Epidemiol. 38, 10-20.

22. Yan, Q., Tiwari, H.K., Yi, N., Gao, G., Zhang, K., Lin, W.Y., Lou, X.Y., Cui, X., Liu, N. (2015). A Sequence Kernel Association Test for Dichotomous Traits in Family Samples under a Generalized Linear Mixed Model. Hum. Hered. 79, 60-68.

23. Park, J.Y., Wu, C., Basu, S., McGue, M., Pan, W. (2018). Adaptive SNP-Set Association Testing in Generalized Linear Mixed Models with Application to Family Studies. Behav. Genet. 48, 55-66.

24. Lee, S., Teslovich, T.M., Boehnke, M., Lin, X. (2013). General framework for meta-analysis of rare variants in sequencing association studies. Am. J. Hum. Genet. 93, 42-53.

25. Liu, D.J., Peloso, G.M., Zhan, X., Holmen, O.L., Zawistowski, M., Feng, S., Nikpay, M., Auer, P.L., Goel, A., Zhang, H. et al. (2014). Meta-analysis of gene-level tests for rare variant association. Nat. Genet. 46, 200-204.

26. Rice, K., Higgins, J.P., Lumley, T. (2017). A re-evaluation of fixed effect(s) meta-analysis. J. R. Stat. Soc. A 181, 205-227.

27. MacCluer, J.W., VandeBerg, J.L., Read, B., Ryder, O.A. (1986). Pedigree analysis by computer simulation. Zoo Biol. 5, 147-160.

28. Conomos, M.P., Laurie, C.A., Stilp, A.M., Gogarten, S.M., McHugh, C.P., Nelson, S.C., Sofer, T., Fernandez-Rhodes, L., Justice, A.E., Graff, M. et al. (2016). Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos. Am. J. Hum. Genet. 98, 165-184.

29. Conomos, M.P., Miller, M.B., Thornton, T.A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet. Epidemiol. 39, 276-293.

30. Morrison, A.C., Huang, Z., Yu, B., Metcalf, G., Liu, X., Ballantyne, C., Coresh, J., Yu, F., Muzny, D., Feofanova, E. et al. (2017). Practical Approaches for Whole-Genome Sequence Analysis of Heart- and Blood-Related Traits. Am. J. Hum. Genet. 100, 205-215.

31. Brody, J.A., Morrison, A.C., Bis, J.C., O'Connell, J.R., Brown, M.R., Huffman, J.E., Ames, D.C., Carroll, A., Conomos, M.P., Gabriel, S. et al. (2017). Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology. Nat. Genet. 49, 1560-1563.

32. Huffman, J.E., de Vries, P.S., Morrison, A.C., Sabater-Lleal, M., Kacprowski, T., Auer, P.L., Brody, J.A., Chasman, D.I., Chen, M.H., Guo, X. et al. (2015). Rare and low-

frequency variants and their association with plasma levels of fibrinogen, FVII, FVIII, and vWF. Blood 126, e19-29.

33. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46, 310-315.

34. Ionita-Laza, I., McCallum, K., Xu, B., Buxbaum, J.D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat. Genet. 48, 214-220.

35. Rogers, M.F., Shihab, H.A., Mort, M., Cooper, D.N., Gaunt, T.R., Campbell, C. (2018). FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics 34, 511-513.

36. Lu, Q., Powles, R.L., Wang, Q., He, B.J., Zhao, H. (2016). Integrative Tissue-Specific Functional Annotations in the Human Genome Provide Novel Insights on Many Complex Traits and Improve Signal Prioritization in Genome Wide Association Studies. PLoS Genet. 12, e1005947.

37. He, Z., Xu, B., Lee, S., Ionita-Laza, I. (2017). Unified Sequence-Based Association Tests Allowing for Multiple Functional Annotations and Meta-analysis of Noncoding Variation in Metabochip Data. Am. J. Hum. Genet. 101, 340-352.

38. Koziol, J.A., Perlman, M.D. (1978). Combining Independent Chi-Squared Tests. Journal of the American Statistical Association 73, 753-763.

39. Wu, M.C., Maity, A., Lee, S., Simmons, E.M., Harmon, Q.E., Lin, X., Engel, S.M., Molldrem, J.J., Armistead, P.M. (2013). Kernel machine SNP-set testing under multiple candidate kernels. Genet. Epidemiol. 37, 267-275.

40. Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J.D., Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. Am. J. Hum. Genet. 92, 841-853.

41. Su, Y.R., Di, C., Bien, S., Huang, L., Dong, X., Abecasis, G., Berndt, S., Bezieau, S., Brenner, H., Caan, B. et al. (2018). A Mixed-Effects Model for Powerful Association Tests in Integrative Functional Genomics. Am. J. Hum. Genet. 102, 904-919.

42. Chen, J., Chen, W., Zhao, N., Wu, M.C., Schaid, D.J. (2016). Small Sample Kernel Association Tests for Human Genetic and Microbiome Association Studies. Genet. Epidemiol. 40, 5-19.

43. Zhou, J.J., Hu, T., Qiao, D., Cho, M.H., Zhou, H. (2016). Boosting Gene Mapping Power and Efficiency with Efficient Exact Variance Component Tests of Single Nucleotide Polymorphism Sets. Genetics 204, 921-931.

44. Dey, R., Schmidt, E.M., Abecasis, G.R., Lee, S. (2017). A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. Am. J. Hum. Genet. 101, 37-49.

45. Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A. et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet.

46. Lumley, T., Brody, J., Peloso, G., Morrison, A., Rice, K. (2018). FastSKAT: Sequence kernel association tests for very large sets of markers. Genet. Epidemiol.

47. Bates, D., Maechler, M., Davis, T.A., Oehlschlägel, J., Riedy, J., R Core Team. (2018). Matrix: Sparse and Dense Matrix Classes and Methods. R package Version 1.2-14.

48. Loh, P.R., Kichaev, G., Gazal, S., Schoech, A.P., Price, A.L. (2018). Mixed-model association for biobank-scale datasets. Nat. Genet. 50, 906-908.

FIGURES

Figure 1. Map of spatially continuous populations from which genotypes were simulated based on the coalescent model. (A) Map for a single-cohort simulation study: the top left 10 × 10 grid formed Population 1, and the rest formed Population 2. (B) Map for a meta-analysis simulation study: Scenario A studies were unrelated individuals sampled from Population 1 only; Scenario B studies were related individuals sampled from specific regions in Population 1 and Population 2; Scenario C studies were unrelated individuals sampled from specific regions in Population 1 and Population 2; and Scenario D studies were related individuals sampled from specific regions in Population 2 only.

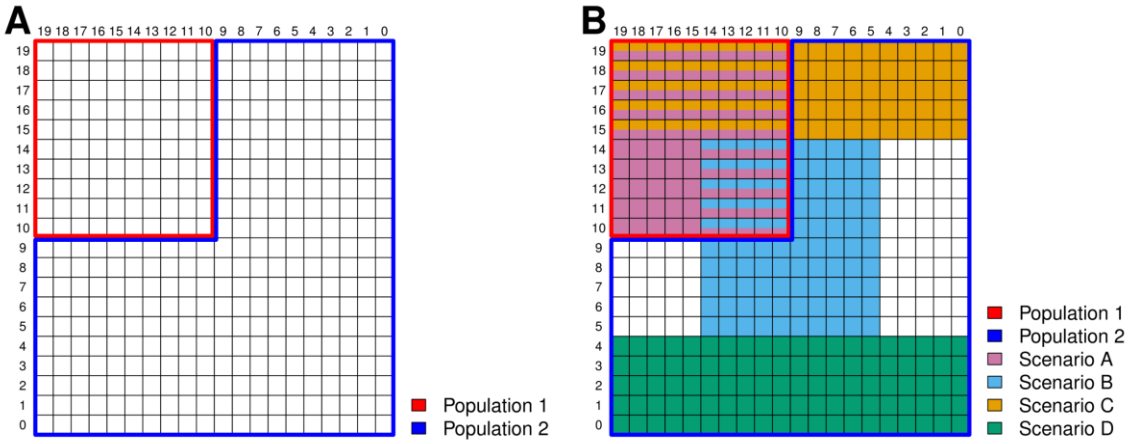


Figure 2. Quantile-quantile plots of SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E in the analysis of 10,000 samples in single-cohort studies with both population structure and cryptic relatedness, under the null hypothesis of no genetic association. (A) Continuous traits in linear mixed models. (B) Binary traits in logistic mixed models.

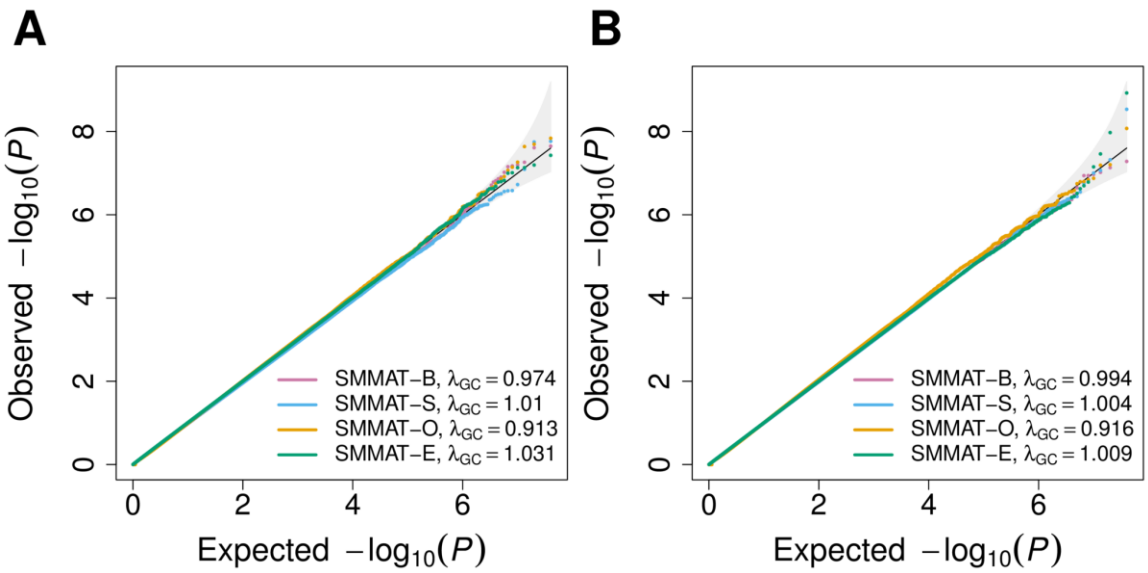


Figure 3. Quantile-quantile plots of SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E in the meta-analysis of 12 studies with a total sample size of 12,000, under the null hypothesis of no genetic association. (A) Continuous traits in linear mixed models, all studies in the same group. (B) Binary traits in logistic mixed models, all studies in the same group. (C) Continuous traits in linear mixed models, Scenario A, B, C, D studies in 4 separate groups. (D) Binary traits in logistic mixed models, Scenario A, B, C, D studies in 4 separate groups.

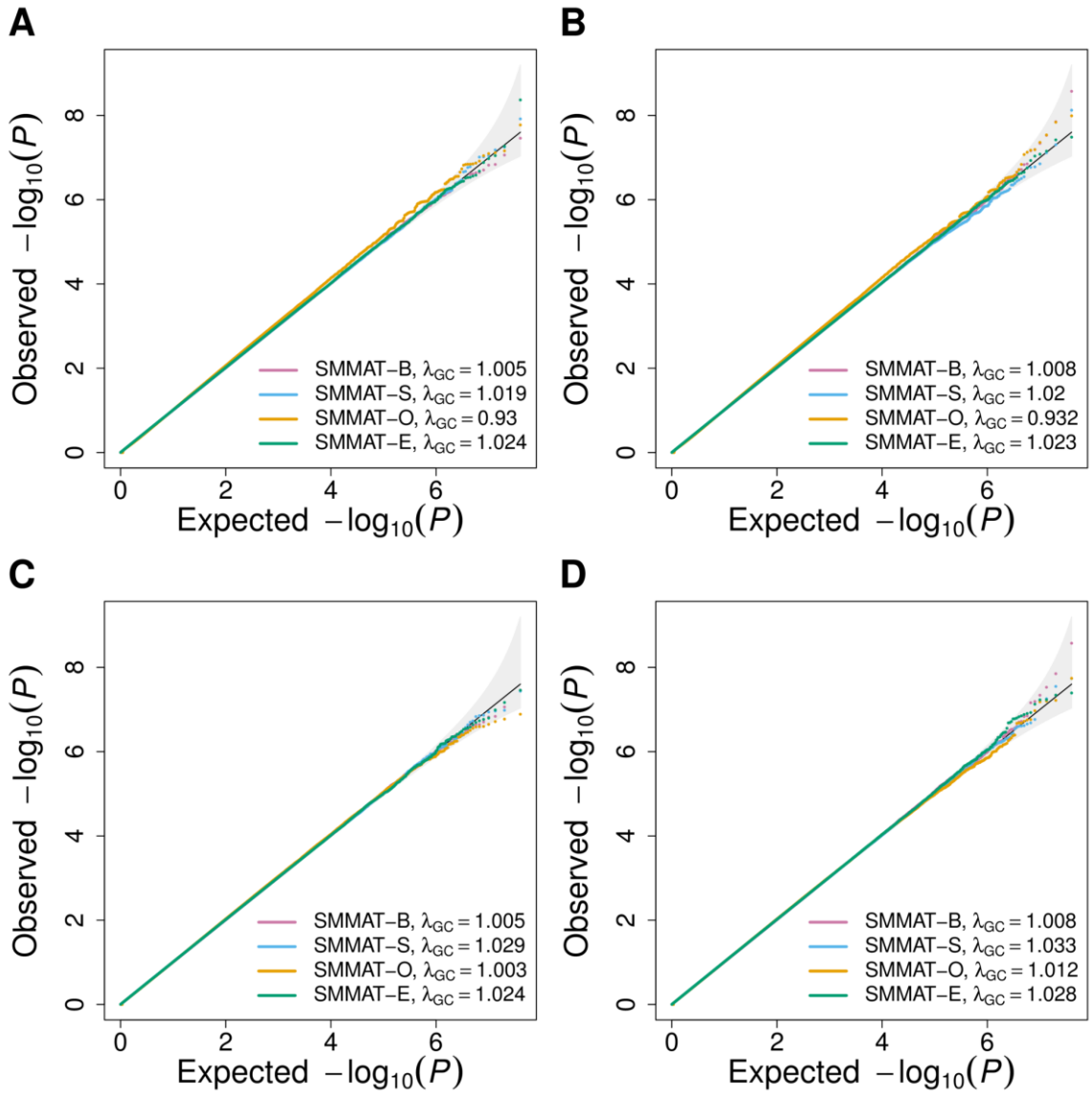


Figure 4. Empirical power of linear mixed model based SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E in continuous trait analysis of 2,000, 5,000 and 10,000 samples. (A) 10% causal variants with 100% negative effects. (B) 10% causal variants with 80% negative effects. (C) 10% causal variants with 50% negative effects. (D) 20% causal variants with 100% negative effects. (E) 20% causal variants with 80% negative effects. (F) 20% causal variants with 50% negative effects. (G) 50% causal variants with 100% negative effects. (H) 50% causal variants with 80% negative effects. (I) 50% causal variants with 50% negative effects. Effect sizes were simulated using the same parameter in each row, but different across rows.

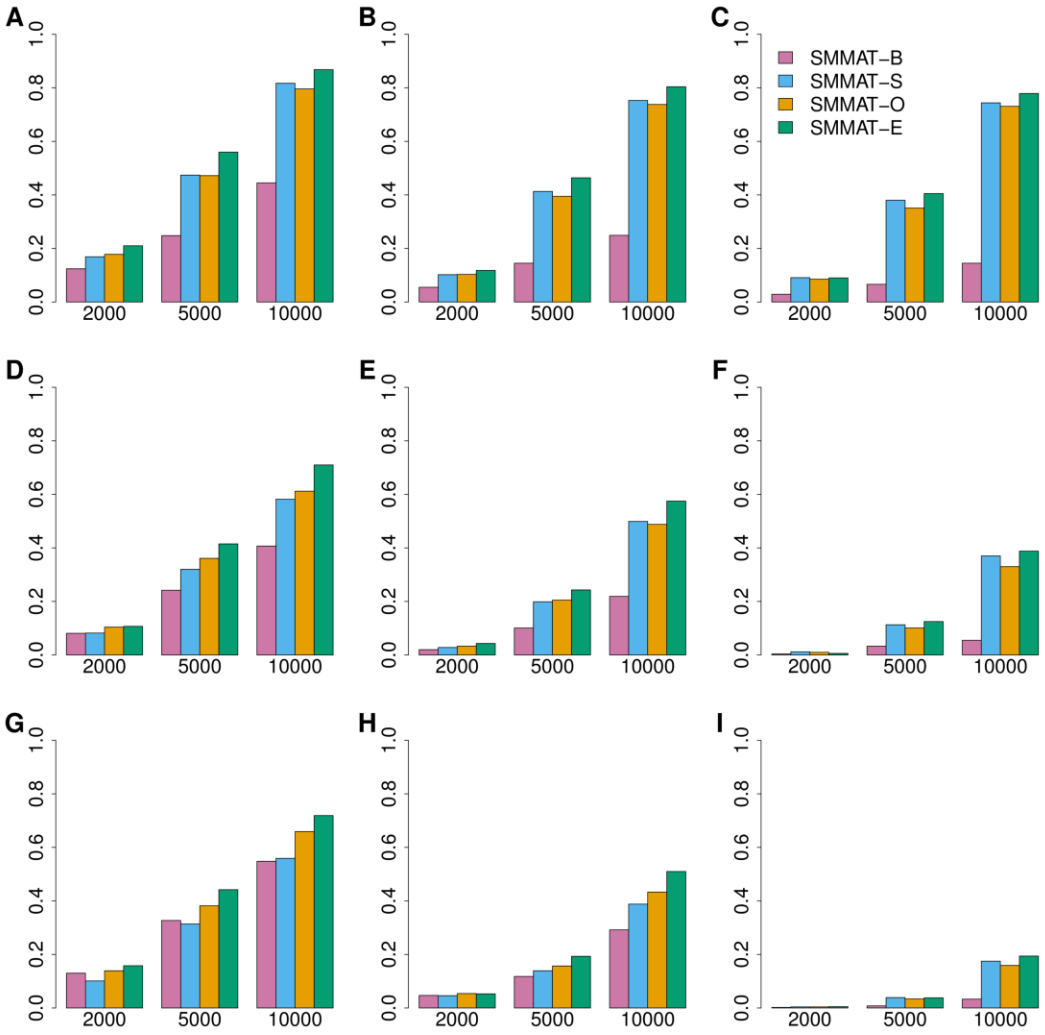


Figure 5. Empirical power of logistic mixed model based SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E in binary trait analysis of 2,000, 5,000 and 10,000 samples. (A) 10% causal variants with 100% negative effects. (B) 10% causal variants with 80% negative effects. (C) 10% causal variants with 50% negative effects. (D) 20% causal variants with 100% negative effects. (E) 20% causal variants with 80% negative effects. (F) 20% causal variants with 50% negative effects. (G) 50% causal variants with 100% negative effects. (H) 50% causal variants with 80% negative effects. (I) 50% causal variants with 50% negative effects. Effect sizes were simulated using the same parameter in each row, but different across rows.

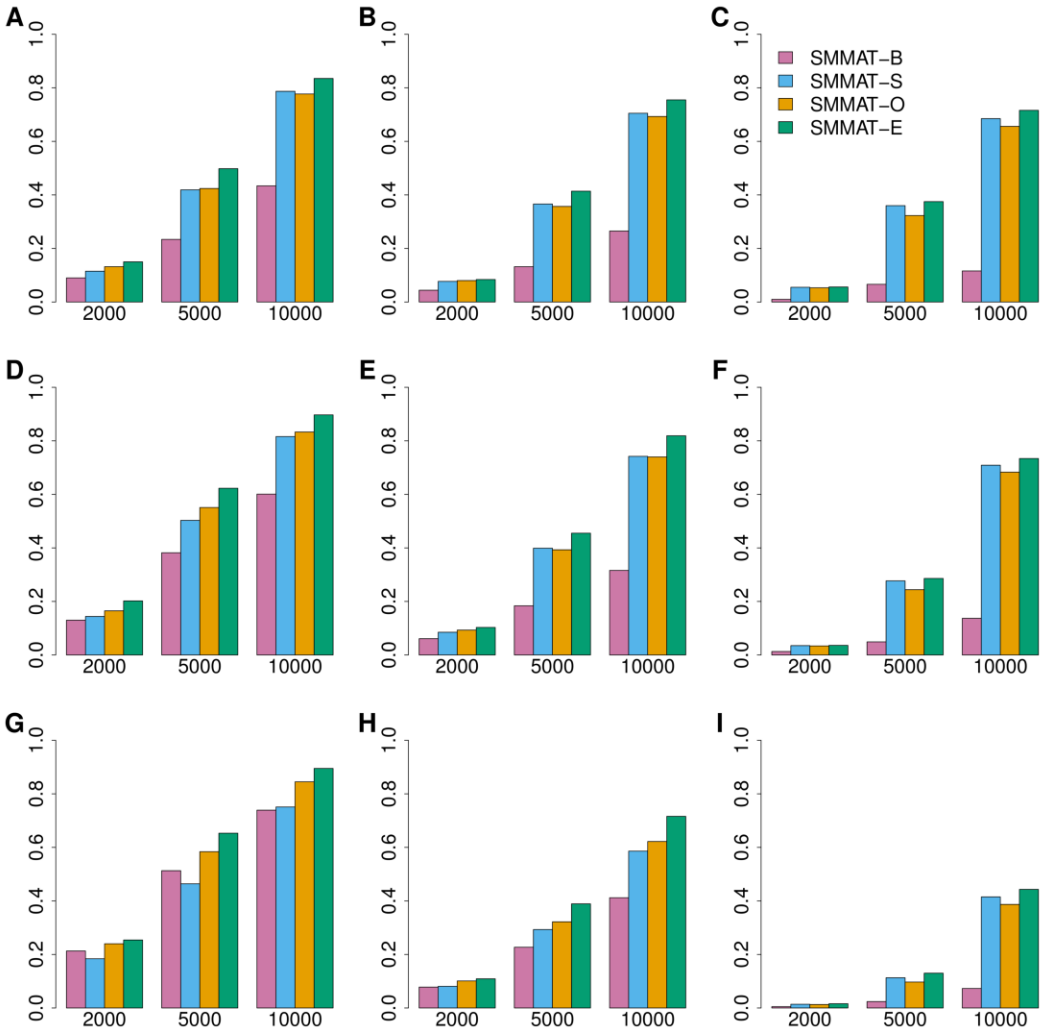
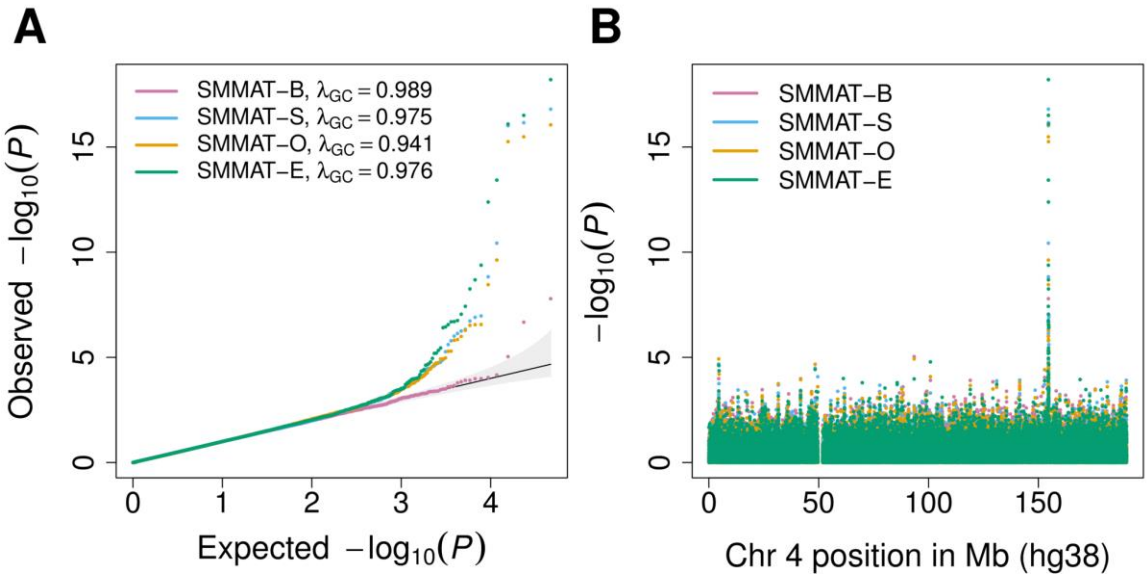


Figure 6. TOPMed fibrinogen level SMMAT analysis results using a heteroscedastic linear mixed model on rare variants with MAF < 5% in non-overlapping 4 kb sliding windows on chromosome 4 (n = 23,763). (A) Quantile-quantile plot. (B) P values on the log scale versus physical positions of the windows on chromosome 4 (build hg38).



TABLES

Table 1. Empirical type I error rates of SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E in single-cohort simulation studies at significance levels of 0.05, 0.0001, and 2.5×10^{-6}. The total sample size was 10,000, and results from 4,000 simulation replicates were combined to get 40 million genetic variant sets.

	Continuous Traits			Binary Traits		
Level	0.05	0.0001	2.5×10^{-6}	0.05	0.0001	2.5×10^{-6}
SMMAT-B	0.047	8.7×10^{-5}	2.0×10^{-6}	0.049	9.6×10^{-5}	2.0×10^{-6}
SMMAT-S	0.048	8.7×10^{-5}	2.0×10^{-6}	0.049	9.5×10^{-5}	2.3×10^{-6}
SMMAT-O	0.050	1.1×10^{-4}	3.0×10^{-6}	0.052	1.2×10^{-4}	3.0×10^{-6}
SMMAT-E	0.050	1.0×10^{-4}	3.0×10^{-6}	0.050	9.9×10^{-5}	2.0×10^{-6}

Table 2. Empirical type I error rates of SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E assuming all studies in the same group (hom) and Scenario A, B, C, D studies in 4 separate groups (het), in meta-analysis simulation studies at significance levels of 0.05, 0.0001, and 2.5×10^{-6}. The total sample size was 12,000 from 12 studies, and results from 4,000 simulation replicates were combined to get 40 million genetic variant sets.

	Continuous Traits			Binary Traits		
Level	0.05	0.0001	2.5×10^{-6}	0.05	0.0001	2.5×10^{-6}
SMMAT-B	0.051	1.0×10^{-4}	2.6×10^{-6}	0.051	1.1×10^{-4}	2.5×10^{-6}
Hom SMMAT-S	0.051	1.0×10^{-4}	2.6×10^{-6}	0.051	1.1×10^{-4}	2.1×10^{-6}
Het SMMAT-S	0.051	1.0×10^{-4}	2.8×10^{-6}	0.052	1.0×10^{-4}	2.4×10^{-6}
Hom SMMAT-O	0.053	1.3×10^{-4}	4.0×10^{-6}	0.053	1.4×10^{-4}	3.4×10^{-6}
Het SMMAT-O	0.052	1.1×10^{-4}	2.6×10^{-6}	0.052	1.1×10^{-4}	2.2×10^{-6}
Hom SMMAT-E	0.051	1.0×10^{-4}	2.5×10^{-6}	0.051	1.1×10^{-4}	2.6×10^{-6}
Het SMMAT-E	0.051	1.0×10^{-4}	2.8×10^{-6}	0.052	1.1×10^{-4}	3.0×10^{-6}

Table 3. TOPMed fibrinogen level SMMAT p values in known association gene *FGB* and flanking regions on chromosome 4, using a heteroscedastic linear mixed model on rare variants with MAF < 5% (n = 23,763). Physical positions of each window are on build hg38.

Start (kb)	End (kb)	No. of variants	SMMAT-B	SMMAT-S	SMMAT-O	SMMAT-E
154,554	154,558	348	6.9×10^{-5}	1.6×10^{-17}	8.9×10^{-17}	6.2×10^{-19}
154,558	154,562	370	0.078	3.7×10^{-11}	2.4×10^{-10}	3.7×10^{-14}
154,562	154,566	326	0.76	1.5×10^{-9}	3.5×10^{-9}	4.2×10^{-10}
154,566	154,570	309	1.6×10^{-8}	9.7×10^{-17}	3.3×10^{-16}	3.1×10^{-17}
154,570	154,574	332	0.030	1.9×10^{-7}	5.2×10^{-7}	8.9×10^{-8}
154,574	154,578	349	2.1×10^{-7}	7.3×10^{-7}	2.8×10^{-7}	4.1×10^{-13}
154,578	154,582	342	1.7×10^{-4}	2.7×10^{-5}	2.8×10^{-5}	2.1×10^{-9}

Table 4. CPU time in the TOPMed fibrinogen level SMMAT using summary statistics from a sliding window analysis using non-overlapping 4 kb windows on chromosome 4 (n = 23,763). Tests were performed using the GMMAT App (version 0.9.2) with one single thread on a computing node with 15 GB total memory in the Analysis Commons.

Test	Time (min)
SMMAT-B	81
SMMAT-S	91
SMMAT-O	266
SMMAT-E	103