The American Journal of Human Genetics, Volume 104

Supplemental Data

Efficient Variant Set Mixed Model Association

Tests for Continuous and Binary Traits

in Large-Scale Whole-Genome Sequencing Studies

Han Chen, Jennifer E. Huffman, Jennifer A. Brody, Chaolong Wang, Seunggeun Lee, Zilin Li, Stephanie M. Gogarten, Tamar Sofer, Lawrence F. Bielak, Joshua C. Bis, John Blangero, Russell P. Bowler, Brian E. Cade, Michael H. Cho, Adolfo Correa, Joanne E. Curran, Paul S. de Vries, David C. Glahn, Xiuqing Guo, Andrew D. Johnson, Sharon Kardia, Charles Kooperberg, Joshua P. Lewis, Xiaoming Liu, Rasika A. Mathias, Braxton D. Mitchell, Jeffrey R. O'Connell, Patricia A. Peyser, Wendy S. Post, Alex P. Reiner, Stephen S. Rich, Jerome I. Rotter, Edwin K. Silverman, Jennifer A. Smith, Ramachandran S. Vasan, James G. Wilson, Lisa R. Yanek, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Hematology and Hemostasis Working Group, Susan Redline, Nicholas L. Smith, Eric Boerwinkle, Ingrid B. Borecki, L. Adrienne Cupples, Cathy C. Laurie, Alanna C. Morrison, Kenneth M. Rice, and Xihong Lin

Supplemental Data

Supplemental Figures

Figure S1. Empirical power of mixed model based SMMAT-B (B), SMMAT-S (S), SMMAT-O (O), SMMAT-E (E) and GLMM-MiST (M) in the presence of large genetic effects. The total sample size was 2,000, and all genetic variants were causal, with effects in the same direction. (A) Power at the significance level of 2.5×10^{-6} for continuous traits in linear mixed models. (B) Power at the significance level of 2.5×10^{-6} for binary traits in logistic mixed models. (C) P value comparison of SMMAT-E and GLMM-MiST for continuous traits in linear mixed models. (D) P value comparison of SMMAT-E and GLMM-MiST for binary traits in logistic mixed models.

Figure S2. Quantile-quantile plots of SMMAT-B, SMMAT-S, SMMAT-O and SMMAT-E in the analysis of 10,000 samples in the presence of both population-level and familial random effects, under the null hypothesis of no genetic association. (A) Continuous traits in linear mixed models with GRM random effects and no ancestry PC adjustment. (B) Binary traits in logistic mixed models with GRM random effects and no ancestry PC adjustment. (C) Continuous traits in linear mixed models with GRM random effects and 10 ancestry PCs. (D) Binary traits in logistic mixed models with GRM random effects and 10 ancestry PCs. (E) Continuous traits in linear mixed models with GRM and population random effects, but no ancestry PC adjustment. (F) Binary traits in logistic mixed models with GRM and population random effects, but no ancestry PC adjustment.

Figure S3. Empirical power of linear mixed model based SMMAT-B (B), SMMAT-S (S), SMMAT-O (O) and SMMAT-E (E) in continuous trait analysis of 5,000 samples, using three models: GRM random effects with no ancestry PC adjustment (1 RE no PC), GRM random effects with 10 ancestry PCs as fixed effects (1 RE 10 PC), GRM and population random effects with no ancestry PC adjustment (2 RE no PC). (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 S4. Empirical power of logistic mixed model based SMMAT-B (B), SMMAT-S (S), SMMAT-O (O) and SMMAT-E (E) in binary trait analysis of 5,000 samples, using three models: GRM random effects with no ancestry PC adjustment (1 RE no PC), GRM random effects with 10 ancestry PCs as fixed effects (1 RE 10 PC), GRM and population random effects with no ancestry PC adjustment (2 RE no PC). (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 S5. P value comparison of TOPMed fibrinogen level SMMAT analysis results using heteroscedastic linear mixed models with and without adjusting for 10 ancestry PCs as fixed effects, using rare variants with MAF < 5% in non-overlapping 4 kb sliding windows on chromosome 4 (n = 23,763). (A) SMMAT-B. (B) SMMAT-S. (C) SMMAT-O. (D) SMMAT-E.

Figure S6. TOPMed fibrinogen level SMMAT analysis results using a heteroscedastic linear mixed model on rare variants with MAF < 5% in non-overlapping 1 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).

Figure S7. TOPMed fibrinogen level SMMAT analysis results using a heteroscedastic linear mixed model on rare variants with MAF < 5% in non-overlapping 10 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).

Figure S8. TOPMed fibrinogen level SMMAT analysis results using a heteroscedastic linear mixed model on rare variants with MAF < 5% in non-overlapping 40 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).

Supplemental Tables

Table S1. TOPMed fibrinogen level SMMAT p values covering two known association variants rs6054 (hg38 position 154,568,456) and rs201909029 (hg 38 position 154,567,636) in gene *FGB* 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.

^a This window covers rs201909029, which has 33 minor allele counts in our TOPMed samples.

^b This window covers rs6054, which has 179 minor allele counts in our TOPMed samples.

Supplemental Methods: Additional Simulation Studies

Impact of Large Genetic Effects on SMMAT-E Power

The efficient hybrid test SMMAT-E is developed based on the assumption that the mean of genetic effects β_0 is not large. To investigate the impact of large genetic effects on SMMAT-E power, we performed additional simulation studies in which all variants in a test unit are causal, with the effects in the same direction. This simulation setting is in favor of SMMAT-B. We used the same genotype data as that in the single-cohort type I error simulations and evaluated the empirical power of SMMAT-B, SMMAT-S, SMMAT-O, SMMAT-E, and GLMM-MiST that combines the p value of SMMAT-B (Equation 2 in Methods) and the p value of SMMAT-S (Equation 3 in Methods) using Fisher's method. All tests were performed using weights equal to a beta distribution density function with parameters 1 and 25 on the MAF of each variant.

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

$$
y_{ij} = \alpha_1 Z_i + \beta_0 T_{ij} + b_{ij} + \epsilon_{ij},
$$

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, the genetic effect $\beta_0 = 0.15$. The burden score T_{ij} of individual j in family i was the weighted sum of causal variant genotypes (with weights equal to a beta distribution density function with parameters 1 and 25 on the MAF of each variant), normalized to have mean 0 and variance 1. The familial random effects b_{ij} were simulated using Equation 5 in Methods, and the random error $\epsilon_{ij} \sim N(0, 1)$. In this parameter setting, the burden score T_{ij} explains 1.3% of the total phenotypic variance. 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

$$
\log\left(\frac{P(y_{ij} = 1)}{1 - P(y_{ij} = 1)}\right) = \alpha_0 + \beta_0 T_{ij} + b_{ij},
$$

where α_0 was chosen such that the disease prevalence was 0.01 in all populations, the genetic effect $\beta_0 = 0.3$. The burden score T_{ij} of individual j in family *i* was calculated in the same way as for continuous traits, and the familial random effects b_{ij} were simulated using Equation 5 in Methods. 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.

We used the sample size of 2,000 for both continuous and binary traits, since the simulated genetic effect was large. We repeated 1,000 simulation replicates, and compared the empirical power at the significance level of 2.5×10^{-6} , as well as the p values from SMMAT-E and GLMM-MiST (Figure S1).

Impact of Multiple Random Effects

One advantage of the SMMAT framework is that it can flexibly use multiple random effects to model between-subject correlation from different sources in the model. We considered two sources of genetic relatedness in our simulation, one from population structure and one from family membership. The population membership matrix has 400 population blocks corresponding to the 20×20 grid in Figure 1A, with elements equal to 1 if two individuals are from the same population grid, 0 otherwise. We included a population-level random intercept with mean 0 and variance 1, which is the same for individuals from the same population grid, in the true models for both continuous and binary traits. We also included the familial random effects from Equation 5 in Methods.

We evaluated the performance of three analytical strategies in the presence of sample correlation due to both population-level and familial random effects: 1) including a single random effects term with the covariance matrix proportional to the GRM, without adjusting for ancestry PCs; 2) including a single random effects term with the covariance matrix proportional to the GRM, and adjusting for the first 10 ancestry PCs as fixed effects; and 3) including two random effects terms, one with the covariance matrix proportional to the GRM, and the other with the covariance matrix proportional to the block-diagonal population membership matrix, without adjusting for ancestry PCs. We evaluated the type I error rate and power as described in Methods.

Supplemental Acknowledgments

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.](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 (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I/HHSN26800001) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) 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 staff and participants of the JHS. Dr. Wilson is supported by U54GM115428 from the National Institute of General Medical Sciences.

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](http://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Long%20List.pdf) [%20Long%20List.pdf.](http://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Long%20List.pdf)

Full Authorship List of the Trans-Omics for Precision Medicine (TOPMed) Consortium

Name Institution(s)

Ryan, Kathleen University of Maryland Sakornsakolpat, Phuwanat Brigham & Women's Hospital Salimi, Shabnam University of Maryland Salzberg, Steven Johns Hopkins University Sandow, Kevin Los Angeles Biomedical Research Institute Sankaran, Vijay Harvard University Scheller, Christopher University of Michigan Schmidt, Ellen University of Michigan Schwander, Karen Washington University in St Louis Schwartz, David University of Colorado at Denver Sciurba, Frank University of Pittsburgh Seidman, Christine Harvard Medical School Sheehan, Vivien Baylor College of Medicine Shetty, Amol University of Maryland Shetty, Aniket University of Colorado at Denver Sheu, Wayne Hui-Heng Taichung Veterans General Hospital Taiwan Shoemaker, M. Benjamin Vanderbilt University Silver, Brian UMass Memorial Medical Center Silverman, Edwin Brigham & Women's Hospital Smith, Jennifer University of Michigan Smith, Josh University of Washington Smith, Nicholas University of Washington Smith, Tanja New York Genome Center Smoller, Sylvia Albert Einstein College of Medicine Snively, Beverly Wake Forest Baptist Health Sofer, Tamar Brigham & Women's Hospital Sotoodehnia, Nona University of Washington Stilp, Adrienne University of Washington Streeten, Elizabeth University of Maryland Sung, Yun Ju Washington University in St Louis Sylvia, Jody Brigham & Women's Hospital Szpiro, Adam University of Washington Sztalryd, Carole University of Maryland Taliun, Daniel University of Michigan Tang, Hua Stanford University Taub, Margaret Johns Hopkins University

Taylor, Kent Los Angeles Biomedical Research Institute

Zhi, Degui University of Texas Health at Houston Zhou, Xiang University of Michigan Zody, Michael New York Genome Center Zoellner, Sebastian University of Michigan

Name Institution(s)