Supplementary Information for:

Rare genetic variants explain missing heritability in smoking

Seon-Kyeong Jang[1], Luke Evans[2,3], Allison Fialkowski[1], Donna K. Arnett[4], Allison E. Ashley-Koch[5], Kathleen C. Barnes[6], Diane M. Becker[7], Joshua C. Bis[8], John Blangero[9], Eugene R. Bleecker[10], Meher Preethi Boorgula[6], Donald W. Bowden[11], Jennifer A. Brody[8], Brian E. Cade[12], Brenda W. Campbell Jenkins[13], April P. Carson[14], Sameer Chavan[6], L. Adrienne Cupples[15], Brian Custer[16], Scott M. Damrauer[17,18], Sean P. David[19,20], Mariza de Andrade[21], Carla L. Dinardo[22], Tasha E. Fingerlin[23,24], Myriam Fornage[25], Barry I. Freedman[26], Melanie E. Garrett[5], Sina A. Gharib[8,27], David C Glahn[28], Jeffrey Haessler[29], Susan R. Heckbert[30,31], John E. Hokanson[32], Lifang Hou[33], Shih-Jen Hwang[34], Matthew C. Hyman[35], Renae Judy[17], Anne E. Justice[36], Robert C Kaplan[29,37], Sharon L.R. Kardia[38], Shannon Kelly[39], Wonji Kim[40], Charles Kooperberg[29], Daniel Levy[34,41], Donald M. Lloyd-Jones[33], Ruth J.F. Loos[42,43], Ani W. Manichaikul[44], Mark T. Gladwin[45], Lisa Warsinger Martin[46], Mehdi Nouraie[45], Olle Melander[47,48], Deborah A. Meyers[10], Courtney G. Montgomery[49], Kari E. North[50], Elizabeth C. Oelsner[51], Nicholette D. Palmer[11], Marinelle Payton[52], Anna L. Peljto[53], Patricia A. Peyser[38], Michael Preuss[42,43], Bruce M Psaty[54], Dandi Qiao[40], Daniel J. Rader[35,55], Nicholas Rafaels[6], Susan Redline[56], Robert M. Reed[57], Alexander P. Reiner[29], Stephen S. Rich[44], Jerome I. Rotter[58], David A. Schwartz[59,60], Aladdin H. Shadyab[61], Edwin K. Silverman[40], Nicholas L. Smith[30,31], J. Gustav Smith[62,63], Albert V. Smith[64], Jennifer A. Smith[38], Weihong Tang[65], Kent D. Taylor[58], Marilyn J. Telen[5], Ramachandran S. Vasan[66,67], Victor R. Gordeuk[68], Zhe Wang[42,43], Kerri L. Wiggins[8], Lisa R. Yanek[7], Ivana V. Yang[53], Kendra A. Young[32], Kristin L. Young[50], Yingze Zhang[45], Dajiang J. Liu[69], Matthew C. Keller[2], Scott Vrieze[1]*

Affiliations:

1 Department of Psychology, University of Minnesota, Minneapolis, MN, USA

2 Institute for Behavioral Genetics, University of Colorado Boulder, Boulder, CO, USA

3 Department of Ecology & Evolution, University of Colorado Boulder, Boulder, CO, USA

4 Dean's Office, University of Kentucky College of Public Health, Lexington, KY, USA

5 Department of Medicine, Duke University School of Medicine, Durham, NC, USA

6 Division of Biomedical Informatics & Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

7 Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

8 Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA

9 Department of Human Genetics, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX, USA

10 Department of Medicine, University of Arizona, Tucson, AZ, USA

11 Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA

12 Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA, USA

13 Jackson Heart Study Graduate Training and Education Center, Jackson State University School of Public Health, Jackson, MS, USA

14 Department of Medicine, University of Mississippi Medical Center, Jackson, MS, USA

15 Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA

16 Vitalant Research Institute, San Francisco CA, USA

17 Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

18 Department of Surgery, Corporal Michael Crescenz VA Medical Center, Philadelphia, PA, USA

19 Department of Family Medicine, Prtizker School of Medicine, University of Chicago, Chicago, IL, USA

20 NorthShore University HealthSystem, Evanston, IL, USA

21 Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA.

22 Fundação Pro Sangue, São Paulo, Brazil

23 Colorado School of Public Health, University of Colorado Denver - Anschutz Medical Campus, Aurora, CO, USA

24 Center for Genes Environment and Health, National Jewish Health, Denver, CO, USA.

25 Brown Foundation Institute of Molecular Medicine, McGovern Medical School, the University of Texas Health Science Center at Houston, Houston, TX, USA

26 Section on Nephrology, Department of Internal Medicine, Wake Forest School of Medicine, Winston-Salem, NC, USA

27 Center for Lung Biology, Division of Pulmonary, Critical Care and Sleep Medicine, University of Washington, Seattle, WA, USA

28 Department of Psychaitry, Boston Children's Hosptial and Harvard Medical School, Boston, MA, USA

29 Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

30 Department of Epidemiology, University of Washington, Seattle WA, USA

31 Kaiser Permanente Washington Health Research Institute, Kaiser Permanente Washington, Seattle WA, USA

32 Department of Epidemiology, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

33 Department of Preventive Medicine, Northwestern University, Chicago, IL, USA

34 Population Sciences Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA

35 Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

36 Department of Population Health Sciences, Geisinger Health System, Danville, PA, USA

37 Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA

38 Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA

39 Department of Pediatrics, UCSF Benioff Children's Hospital Oakland, Oakland, CA, USA

40 Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

41 Framingham Heart Study, Framingham, MA, USA

42 Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

43 The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

44 Center for Public Health Genomics, School of Medicine, University of Virginia, Charlottesville, VA, USA

45 Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

46 Division of Cardiology, School of Medicine and Health Sciences, Washington, DC, USA

47 Department of Clinical Sciences, Lund University, Malmö, Sweden

48 Department of Internal Medicine, Skåne University Hospital, Malmö, Sweden

49 Genes and Human Disease Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA

50 Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

51 Division of General Medicine, Columbia University Irving Medical Center, Columbia University, New York, NY, USA

52 Department of Epidemiology and Biostatistics, Jackson Heart Study Graduate Training and Education Center, Jackson State University School of Public Health, Jackson, MS, USA

53 Department of Medicine, University of Colorado School of Medicine, Aurora, CO, USA

54 Cardiovascular Health Research Unit, Department of Medicine, Epidemiology and Health Services, University of Washington, Seattle, WA, USA

55 Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

56 Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, US

57 University of Maryland School of Medicine, Baltimore, MD, USA

58 The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA

59 Department of Medicine, School of Medicine, University of Colorado Denver, Aurora, CO, USA

60 Department of Immunology, School of Medicine, University of Colorado Denver, Aurora, CO, USA

61 Herbert Wertheim School of Public Health and Human Longevity Science, University of California, San Diego, La Jolla, CA, USA

62 Wallenberg Laboratory/Department of Molecular and Clinical Medicine, Institute of Medicine, Gothenburg University, Sweden

63 Department of Cardiology, Sahlgrenska University Hospital, Gothenburg, Sweden

64 Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, USA

65 Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN, USA

66 Sections of Preventive medicine and Epidemiology and cardiovascular medicine, Department of medicine, Boston University School of Medicine, Boston, MA, USA

67 Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA

68 Department of Medicine, University of Illinois at Chicago, Chicago, IL, USA

69 Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA, USA

Corresponding author: vrieze@umn.edu

Supplementary Note 1. GREML-LDMS-I analysis in individuals with European ancestry

Sample selection

To select the European-ancestry sample, we applied a two-step procedure, described below, to evaluate the convergence of the two different approaches. First, an initial set of European genomes was selected by projecting 1000 Genome (1000G) PC loadings onto TOPMed genotypes (freeze 8; autosomes only) and applying K-nearest neighbors classification on the h20 projected PCs using open source software `FRAPOSA`[1]. This software uses online augmentation-decomposition-transformation (OADP) for projection to avoid potential misalignment and shrinkage bias. To calculate 1000G PC loadings, SNPs common in both 1000G and TOPMed were extracted and pruned for LD using plink 1.9 (--indep-pairwise 100kb 5 0.1) and went through other quality control filters (--geno 0.01, --mind 0.01), resulting in a total of 103,646 variants. Then, genetic PCs for the TOPMed samples were calculated using plink 1.9 (--score). Ancestry was assigned based on "votes" of the top 20 closest neighbors (uniformly weighted) in the 1000G dataset in terms of PC 1-20 scores. As in the original paper[1], those having probability of being European ancestry greater than .875 made an initial set of European ancestry (N=38,915 for EUR sample having at least one smoking phenotypes). Visual inspection of the PC 1-4 plots of this initial sample suggested residual heterogeneity within the European ancestry sample. Therefore, we further restricted samples via following way: 1), we created four increasingly restrictive samples based on Euclidean distance of PC 1-4 from the vertex (maximum of PC1, minimum of PC2, and PC3-4 values evaluated at the maximum value of PC1). Then, we calculated Euclidean distance from this vertex for each individual and removed outliers having a distance greater than 3, 2, 1.5, and 1 times Inter-Quartile Range (IQR), each comprising four increasingly restrictive sample sets. The 1 IQR sample (i.e., most conservative with respect to population structure) was used for our main analyses with the remaining samples reserved for sensitivity analysis. Sample size and PC1-4 plots of the resulting samples are presented in Supplementary Table 12 and Supplementary Fig. 2.

Phenotypes

We applied the same definition for phenotypes as used in the recent GWAS of tobacco use[2]. Age of smoking initiation (AgeSmk) is a continuous variable indicating the age at which an individual started smoking regularly. Cigarettes per day (CigDay) is an ordinal variable with 5 bins (1-5) with higher numbers indicating a higher average number of cigarettes smoked per day, either as a current smoker or former smoker. Specifically, each bin indicates the following quantity: bin1 (1~5), bin2 (6~15), bin3 (16~25), bin4 (26~35), bin5 (36+). Individuals who never smoked were set to missing. Smoking cessation (SmkCes) is a binary phenotype: current (case=1) versus former smoker (control=0). Smoking initiation (SmkInit) is also a binary variable measuring whether one reports ever being a regular smoker in their life (i.e., ever smoked 100 cigarettes in their life) versus (case=1) never smoker (control=0). Supplementary Table 8 shows cohort-wide and cohort-level descriptive statistics of four smoking phenotypes. The relationship among four smoking phenotypes was explored using linear mixed-effects regression (Supplementary Table 7). For each phenotype, they were regressed on age, sex, and the two remaining smoking phenotypes (except for SmkInit) as fixed covariates and cohort as a random intercept. The relation between SmkInit and other variables could not be estimated because every individual reporting AgeSmk, CigDay, and SmkCes, were lifetime smokers. In `R(v3.63)`, we used `lme` function in `nlme(v3.1)` package for quantitative traits and `glmer` function for binary traits in `lme4(v1.1)` package.

Local Permutation analysis

We estimated the potential amount of residual population stratification using permutation method. First, we calculated genetic distances between the i-th and the j-th individual using the following formula: $\sum_{k=1}^{11} \lambda_k \left| PC_{ki} - PC_{kj} \right|^2$, where PCs are principal components calculated with common variants and λ_k is eigenvalue corresponding to the k-th PCs. Then, we generated a list of permutations by randomly

exchanging the i-th individual with one of their 100 nearest neighbors which includes themselves. A more detailed procedure of generating permutation is described elsewhere[3]. Similar to this study, we dropped the first 50 permutations and sampled each permutation list with a step of 5 to ensure relative independence between the permutation lists. Taking a semi-empirical approach, we calculated the mean and standard deviations of 100 permutation trials and used them to construct empirical null distribution, $N(\mu, \sigma^2)$. Mean heritability departing from zero would indicate potential confounding bias by population stratification. We conducted a one-tailed z-test for 6 bins of each phenotype to test whether μ is significantly different from zero (Supplementary Table 4).

Pedigree-based narrow-sense heritability

For pedigree-based analysis, we only included cohorts having more than 10 pairs of close relatives (i.e., those related greater than .375) from 1 IQR European sample which has not been filtered by relatedness yet. This is to prevent the relatedness structure from becoming too low in which case heritability estimates can be deflated[4]. Then, we removed one in the pairs related to greater than .8 to prevent duplicates or monozygotic twins (which are indistinguishable) being accidentally included. Next, we created a GRM using variants with MAF greater than .05. We set pairs related less than .05 as 0 in the same GRM to prevent deflation of heritability estimates. This resulted in a GRM based on the 6th degree, or closer, relatives, including all relative pairs excluded in the SNP heritability analyses described throughout this report, as well as unrelated individuals. Such a GRM includes additive effects of common variants, but also additive effects of all variants in haplotypes shared among relatives, including common and rare SNPs. This modified GRM was entered as a random effect together with two other random effects, a cohort effect and a shared environment effect. We created the shared environment GRM by setting pairs related greater than .375 as 1 and 0 otherwise. We entered the same set of fixed effects (demographic, sequencing center, and genetic PCs) used in our main SNP-based heritability analysis. We found that fitting both genetic and shared environment GRMs did not converge or produced non-sensible

estimates due to strong dependence between the genetic GRM and the shared environment GRM. Therefore, we reported results from models with only the genetic and cohort random effects, understanding that the estimates can be upwardly biased due to shared environment effect.

Supplementary Note 2. GREML-LDMS-I analysis in individuals with African ancestry

Sample

We initially had 8,546, 8,894, 10,361, and 19,632 African admixed individuals (as predicted by OADP ancestry prediction) for AgeSmk, CigDay, SmkCes, and SmkInit, respectively. Unrelated samples were selected based on kinship coefficients calculated by PC-relate ('`pcrelate`' in GENESIS R package) which estimates kinship in admixed samples conditional on ancestry[5]. A total of 638,486 LD-pruned variants (|LD|<.32) with MAF greater than 1% were used to calculate relatedness. We applied `pcairPartition` in GENESIS package with kinship coefficient threshold < .0125 to select unrelated individuals (equivalent to $\hat{\pi} < .025$). While a majority of individuals showed two-way (African-European) admixture, some individuals showed more complex admixture patterns. We excluded these individuals with global ancestry proportion of either Central-South Asia, East Asia, Native America, and Middle East greater than 10% (N=58). Global ancestry proportions were derived from local ancestry estimates from RFMix where local ancestry of each homologous chromosome was estimated at 610,860 autosomal SNPs using Human Genome Diversity Panel (HGDP)[6] as a reference panel. Descriptive statistics of each cohort are presented in Supplementary Table 8 and the global ancestry proportion of the sample is presented in Supplementary Fig. 3. As in European ancestry analysis, we performed LD-pruning (indep-pairwise 100 5 0.2 for common, indep-pairwise 2000 10 0.02 for rare variants) and estimated 50 PCs using ~1M common (MAF 1-50%) variants and another 50 PCs from ~6M rare variants (MAF 0.01-1%). These 100 PCs were used as fixed effect covariates in GREML-LDMS-I analysis.

Local ancestry kinship

Individuals with similar global ancestry proportions can exhibit variable local ancestry compositions. Adjustment of local ancestral backgrounds was previously shown to produce less biased heritability estimates in admixed populations for other traits[7]. We modeled the effect of local ancestry as a random effect by constructing local ancestry-based kinship. Local ancestry kinship was constructed by the following procedure: 1) local ancestry genotype matrix was created with each individual assigned either 0 or 1 or 2 depending on the number of Sub-Saharan African allele at 610,860 autosomal variants from HGDP. Then, we standardized this matrix buy: $\overline{\gamma_{i,s}} = (\gamma_{i,s} - 2\hat{\theta})/\sqrt{2\hat{\theta}(1-\hat{\theta})}$ where $\gamma_{i,s}$ indicates local ancestry of the i-th individual at variant s and $\hat{\theta}$ indicates mean African ancestry proportion of current sample[8]. We created a GRM based on the standardized local ancestry matrix and included it as a random effect in GREML-LDMS-I analysis.

References

1. Zhang, D., Dey, R., and Lee, S. (2020). Fast and robust ancestry prediction using principal component analysis. Bioinformatics *36*, 3439–3446.

2. Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D.M., Chen, F., Datta, G., Davila-Velderrain, J., McGuire, D., Tian, C., et al. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet *51*, 237–244.

3. Mullaert, J., Bouaziz, M., Seeleuthner, Y., Bigio, B., Casanova, J.-L., Alcaïs, A., Abel, L., and Cobat, A. (2021). Taking population stratification into account by local permutations in rare-variant association studies on small samples. Genet Epidemiol *45*, 821–829.

4. Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., and Price, A.L. (2013). Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. PLOS Genetics *9*, e1003520.

5. Conomos, M.P., Reiner, A.P., Weir, B.S., and Thornton, T.A. (2016). Model-free Estimation of Recent Genetic Relatedness. Am J Hum Genet *98*, 127–148.

6. Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science *367*, eaay5012.

7. Zhong, Y., Perera, M.A., and Gamazon, E.R. (2019). On Using Local Ancestry to Characterize the Genetic Architecture of Human Traits: Genetic Regulation of Gene Expression in Multiethnic or Admixed Populations. Am J Hum Genet *104*, 1097–1115.

8. Zaitlen, N., Pasaniuc, B., Sankararaman, S., Bhatia, G., Zhang, J., Gusev, A., Young, T., Tandon, A., Pollack, S., Vilhjálmsson, B.J., et al. (2014). Leveraging population admixture to explain missing heritability of complex traits. Nat Genet *46*, 1356–1362.

Supplementary Note 3. Study-specific Acknowledgements and Disclaimer

Study's short name	Study name	PI	Study-specific acknowledgements details	TOPMed accession number
ARIC	Atherosclerosis Risk in Communities Study VTE cohort	Eric Boerwinkle <Eric.Boerwinkle@uth.tmc.edu>	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: Atherosclerosis Risk in Communities (ARIC)" (phs001211) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201500015C and 3U54HG003273-12S2) and the Broad Institute for MIT and Harvard (3R01HL092577-06S1). 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. The Genome Sequencing Program (GSP) was funded by the National Human Genome Research Institute (NHGRI), the National Heart, Lung, and Blood Institute (NHLBI), and the National Eye Institute (NEI). The GSP Coordinating Center (U24 HG008956) contributed to cross-program scientific initiatives and provided logistical and general study coordination. The Centers for Common Disease Genomics (CCDG) program was supported by NHGRI and NHLBI, and whole genome sequencing was performed at the Baylor College of Medicine Human Genome Sequencing Center (UM1 HG008898 and R01HL059367). 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	phs001211

			HHSN268201700005I). The authors thank the staff and participants of the ARIC study for their important contributions.	
BAGS	New Approaches for Empowering Studies of Asthma in Populations of African Descent - Barbados Asthma Genetics Study	Kathleen Barnes <Kathleen.barnes@Tempus.com>	We gratefully acknowledge the contributions of Harold Watson, Clive Landis, Desiree Walcott, Pissamai and Trevor Maul, Paul Levett, Anselm Hennis, P. Michele Lashley, Raana Naidu, Malcolm Howitt and Timothy Roach for their contributions in the field and support at the Chronic Disease Research Centre. The authors wish to thank the numerous health care providers, and community clinics and co-investigators who assisted in the phenotyping and collection of DNA samples, and the families and patients for generously donating DNA samples to the Barbados Asthma Genetics Study (BAGS). Funding for BAGS was provided by National Institutes of Health (NIH) R01HL104608, R01HL087699, HL104608 S1, R01AI132476, and R01AI114555.	phs001143
BioMe	Mount Sinai BioMe Biobank	Ruth J.F. Loos <ruth.loos@mssm.edu> Eimear Kenny <eimear.kenny@mssm.edu>	This research was supported by NIH grant X01HL134588.	phs001644
CARDIA	Coronary Artery Risk Development in Young Adults	Myriam Fornage <Myriam.Fornage@uth.tmc.edu> Lifang Hou <l-hou@northwestern.edu>	Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Whole Genome Sequencing for the NHLBI TOPMed: CARDIA Study (phs001612) was performed at the Baylor College of Medicine Human genome Sequencing Center (contract HHSN268201600033I). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The Coronary Artery Risk Development in Young Adults Study (CARDIA) is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with the	phs001612

			University of Alabama at Birmingham (HHSN268201800005I & HHSN268201800007I), Northwestern University (HHSN268201800003I), University of Minnesota (HHSN268201800006I), and Kaiser Foundation Research Institute (HHSN268201800004I).	
CFS	Cleveland Family Study - WGS Collaboration	Susan Redline <sredline1@rics.bwh.harvard.edu>	The Cleveland Family Study has been supported in part by National Institutes of Health grants [R01-HL046380, KL2-RR024990, R35-HL135818, and R01-HL113338].	phs000954
CHS	Cardiovascular Health Study	Bruce Psaty <psaty@u.washington.edu> Russell Tracy <Russell.Tracy@uvm.edu>	This CHS research was supported by NHLBI contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, 75N92021D00006; and NHLBI grants U01HL080295, R01HL087652, R01HL105756, R01HL103612, R01HL120393, and U01HL130114 with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org. 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. Dr. Psaty serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson.	phs001368
CHS_VTE	Cardiovascular Health Study	Susan Heckbert <heckbert@uw.edu>	Please contact the study PI.	phs001368
COPDGene/Eclipse	Genetic Epidemiology of COPD Study	Edwin Silverman <ed.silverman@channing.harvard.edu>	This research was supported by NIH grants U01 HL089856 and U01 HL089897. ECLIPSE was supported by GSK. Dr. Silverman has received grant support from GSK and Bayer.	phs000951
DHS	Diabetes Heart Study	Nicholette Palmer <nallred@wakehealth.edu> Donald W. Bowden <dbowden@wakehealth.edu>	This work was supported by R01 HL92301, R01 HL67348, R01 NS058700, R01 AR48797, R01 DK071891, R01 AG058921, the General Clinical Research Center of the Wake Forest University School of Medicine (M01 RR07122, F32 HL085989), the American Diabetes Association, and a pilot grant from the Claude Pepper Older Americans Independence	phs001412

			Center of Wake Forest University Health Sciences (P60 AG10484).	
EOCOPD	Boston Early-Onset COPD Study	Edwin Silverman <ed.silverman@channing.harvard.edu>	The Boston Early-Onset COPD Study was supported by R01 HL113264 and U01 HL089856 from the National Heart, Lung, and Blood Institute.	phs000946
FHS	Framingham Heart Study	Vasan S. Ramachandran <vasan@bu.edu> Nancy Heard-Costa <nheard@bu.edu>, Dan Levy <LevyD@nih.gov>	This investigation includes data From the Framingham Heart Study of the National Heart Lung and Blood Institute of the National Institutes of Health and Boston University School of Medicine. This project has been funded in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services, under Contract No. 75N92019D00031. The Framingham Heart Study (FHS) acknowledges the support of Contracts NO1-HC-25195, HHSN268201500001I and 75N92019D00031 from the National Heart, Lung and Blood Institute and grant supplement R01 HL092577-06S1 for this research. We also acknowledge the dedication of the FHS study participants without whom this research would not be possible. Dr. Vasan is supported in part by the Evans Medical Foundation and the Jay and Louis Coffman Endowment from the Department of Medicine, Boston University School of Medicine. 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.	phs000974
GeneSTAR	Genetic Studies of Atherosclerosis Risk	Rasika Mathias <rmathias@jhmi.edu>	GeneSTAR was supported by 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- R000052) to the Johns Hopkins General Clinical Research Center.	phs001218
GENOA	Genetic Epidemiology Network of Arteriopathy	Sharon Kardia <skardia@umich.edu> Jennifer Smith <smjenn@umich.edu>	Support for GENOA was provided by the National Heart, Lung and Blood Institute (HL054457, HL054464, HL054481, HL119443, and HL087660) of the National Institutes of Health. GENOA may also require study-specific acknowledgements depending on the phenotypes analyzed (please consult with the GENOA co-author on the paper).	phs001345

GOLDN/HyperGEN	Genetics of Lipid Lowering Drugs and Diet Network	Donna K Arnett <donna.arnett@uky.edu>	This research was supported by NIH grant 1R01 HL091357.	phs001359
HCHS_SOL	Hispanic Community Health Study - Study of Latinos	Robert Kaplan <Robert.Kaplan@einsteinmed.org> Kari North <kari_north@unc.edu>	This research was supported by NIH grants R01HG010297 and U01HG007416. All HCHS/SOL participants provided informed consent, and the study was approved by the Institutional Review Board of local field centers, coordinating center and laboratories. The Hispanic Community Health Study/Study of Latinos is a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (HHSN268201300001I / N01-HC-65233), University of Miami (HHSN268201300004I / N01-HC-65234), Albert Einstein College of Medicine (HHSN268201300002I / N01-HC-65235), University of Illinois at Chicago – HHSN268201300003I / N01-HC-65236 Northwestern Univ), and San Diego State University (HHSN268201300005I / N01-HC-65237). The following Institutes/Centers/Offices have contributed to the HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities,National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, NIH Institution-Office of Dietary Supplements.	phs001395
HVH	Heart and Vascular Health Study	Susan Heckbert <heckbert@uw.edu>	The Heart and Vascular Health Study was supported by grants HL068986, HL085251, HL095080, and HL073410 from the National Heart, Lung, and Blood Institute.	phs000993
IPF	Whole Genome Sequencing in Familial and Sporadic Idiopathic Pulmonary Fibrosis	David Schwartz <David.Schwartz@ucdenver.edu> Tasha Fingerlin <fingerlint@njhealth.org>	Please contact the study PI.	phs001607
JHS	Jackson Heart Study	April Carson <apcarson@umc.edu> Laura Raffield	The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of	phs000964

		<raffield@email.unc.edu>	Health (HHSN268201800015I) 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 staffs and participants of the JHS. 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.	
Mayo_VTE	Mayo Clinic Venous Thromboembolism Study	Mariza de Andrade <mandrade@mayo.edu>	Funded, in part, by grants from the National Institutes of Health, National Heart, Lung and Blood Institute (HL66216 and HL83141). the National Human Genome Research Institute (HG04735, HG06379), and research support provided by Mayo Foundation.	phs001402
MESA	Multi-Ethnic Study of Atherosclerosis	Jerome Rotter <jrotter@lundquist.org> Stephen Rich <ssr4n@virginia.edu>	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: Multi-Ethnic Study of Atherosclerosis (MESA)" (phs001416.v1.p1) was performed at the Broad Institute of MIT and Harvard (3U54HG003067-13S1). 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), and TOPMed MESA Multi-Omics (HHSN2682015000031/HSN26800004). The MESA projects are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for the Multi-Ethnic Study of Atherosclerosis (MESA) projects 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 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161,	phs001416

			75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420, UL1TR001881, DK063491, and R01HL105756. The authors thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions. A fill list of participating MESA investigators and institutes can be found at http://www.mesa-nhlbi.org.	
MPP	Malmo Preventative Project	Gustav Smith <gustav@broadinstitute.org>	J. Gustav Smith was supported by grants from the Swedish Heart-Lung Foundation (2016-0134, 2016-0315 and 2019-0526), the Swedish Research Council (2017-02554), the European Research Council (ERC-STG-2015-679242), the Crafoord Foundation, Skåne University Hospital, the Scania county, governmental funding of clinical research within the Swedish National Health Service, a generous donation from the Knut and Alice Wallenberg foundation to the Wallenberg Center for Molecular Medicine in Lund, and funding from the Swedish Research Council (Linnaeus grant Dnr 349-2006-237, Strategic Research Area Exodiab Dnr 2009-1039) and Swedish Foundation for Strategic Research (Dnr IRC15-0067) to the Lund University Diabetes Center.	phs001544
OMG_SCD	Outcome Modifying Genes in Sickle Cell Disease	Allison Ashley-Koch <allison.ashleykoch@duke.edu> Marilyn Telen <marilyn.telen@duke.edu>	The OMG-SCD study was administrated by Marilyn J. Telen, M.D. and Allison E. Ashley-Koch, Ph.D. from Duke University Medical Center and collection of the data set was supported by grants HL068959, HL079915, R01HL68959, HL70769, HL87681 from the National Heart, Lung, and Blood Institute (NHLBI) of the National Institute of Health (NIH).	phs001608
PMBB_AF	Early-onset Atrial Fibrillation in the Penn Medicine BioBank Cohort	Daniel Rader <rader@mail.med.upenn.edu> Scott Damrauer <Scott.Damrauer@uphs.upenn.edu>	The Penn Medicine BioBank is funded by a gift from the Smilow family, the National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA Award Number UL1TR001878, and the Perelman School of Medicine at the University of Pennsylvania. SMD is supported by IK2-CX001780. This publication does not represent the views of the Department of Veterans Affairs or the United States Government.	phs001601

			Dr. Damrauer had research support to University of Pennsylvania from RenalytixAI and personal fees from Calico Labs, both outside the current work.	
REDs-III-Brazil	Recipient Epidemiology and Donor Evaluation Study-III	Custer, Brian <BCuster@vitalant.org>; Kelly, Shannon <shannon.kelly@ucsf.edu>	REDS-III was supported by the National Heart, Lung, and Blood Institute (NHLBI) HHSN268201100007I.	phs001468
SAFS	Whole Genome Sequencing to Identify Causal Genetic Variants Influencing CVD Risk - San Antonio Family Studies	John Blangero <john.blangero@utrgv.edu> Joanne Curran <joanne.curran@utrgv.edu>	This research was supported by National Institutes of Health grant R01 HL113323.	phs001215
Sarcoidosis	Genetics of Sarcoidosis in African Americans	Courtney Montgomery <Courtney-Montgomery@omrf.org>	This research was supported by National Institutes of Health (R01HL113326, P30 GM110766-01)	phs001207
SARP	Severe Asthma Research Program	Deborah A Meyers <dameyers@email.arizona.edu>	Please contact the study PI.	phs001446
walk_PHaSST	Treatment of Pulmonary Hypertension and Sickle Cell Disease with Sildenafil Therapy	Mark Gladwin <gladwinmt@upmc.edu> Yingze Zhang <zhanyx@UPMC.EDU>	We thank the investigators of the Walk-PHasst study and the patients who participated in the study. We also thanks the walk-PHaSST clinical site team: Albert Einstein College of Medicine: Jane Little and Verlene Davis; Columbia University: Robyn Barst, Erika Rosenzweig, Margaret Lee and Daniela Brady; UCSF Benioff Children's Hospital Oakland: Claudia Morris, Ward Hagar, Lisa Lavrisha, Howard Rosenfeld, and Elliott Vichinsky; Children's Hospital of Pittsburgh of UPMC: Regina McCollum; Hammersmith Hospital, London: Sally Davies, Gaia Mahalingam, Sharon Meehan, Ofelia Lebanto, and Ines Cabrita; Howard University: Victor Gordeuk, Oswaldo Castro, Onyinye Onyekwere, Vandana Sachdev, Alvin Thomas, Gladys Onojobi, Sharmin Diaz, Margaret Fadojutimi-Akinsiku, and Randa Aladdin; Johns Hopkins University: Reda Girgis, Sophie Lanzkron and Durrant Barasa; NHLBI: Mark Gladwin, Greg Kato, James Taylor, Wynona Coles, Catherine Seamon, Mary Hall, Amy Chi, Cynthia Brenneman, Wen Li, and Erin Smith; University of Colorado: Kathryn Hassell, David Badesch, Deb McCollister and Julie McAfee; University of Illinois at Chicago: Dean Schraufnagel, Robert Molokie, George Kondos, Patricia	phs001514

			Cole-Saffold, and Lani Krauz; National Heart & Lung Institute, Imperial College London: Simon Gibbs. Thanks also to the data coordination center team from Rho, Inc.: Nancy Yovetich, Rob Woolson, Jamie Spencer, Christopher Woods, Karen Kesler, Vickie Coble, and Ronald W. Helms. We also thank Dr. Yingze Zhang for directing the Walk-PHasst repository and Dr. Mehdi Nouraie for maintaining the Walk-PHasst database and Dr. Jonathan Goldsmith as a NIH program director for this study. Special thanks to the volunteers who participated in the Walk-PHaSST study. This project was funded with federal funds from the NHLBI, NIH, Department of Health and Human Services, under contract HHSN268200617182C. This study is registered at www.clinicaltrials.gov as NCT00492531.	
WHI	Women's Health Initiative	Charles Kooperberg <clk@fredhutch.org> Alex Reiner apreiner@uw.edu Jeffrey Haessler <jhaessle@whi.org> Sean David <spdavid1@uchicago.edu>	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.	phs001237

Supplementary Note 4. TOPMed Omics Support Table

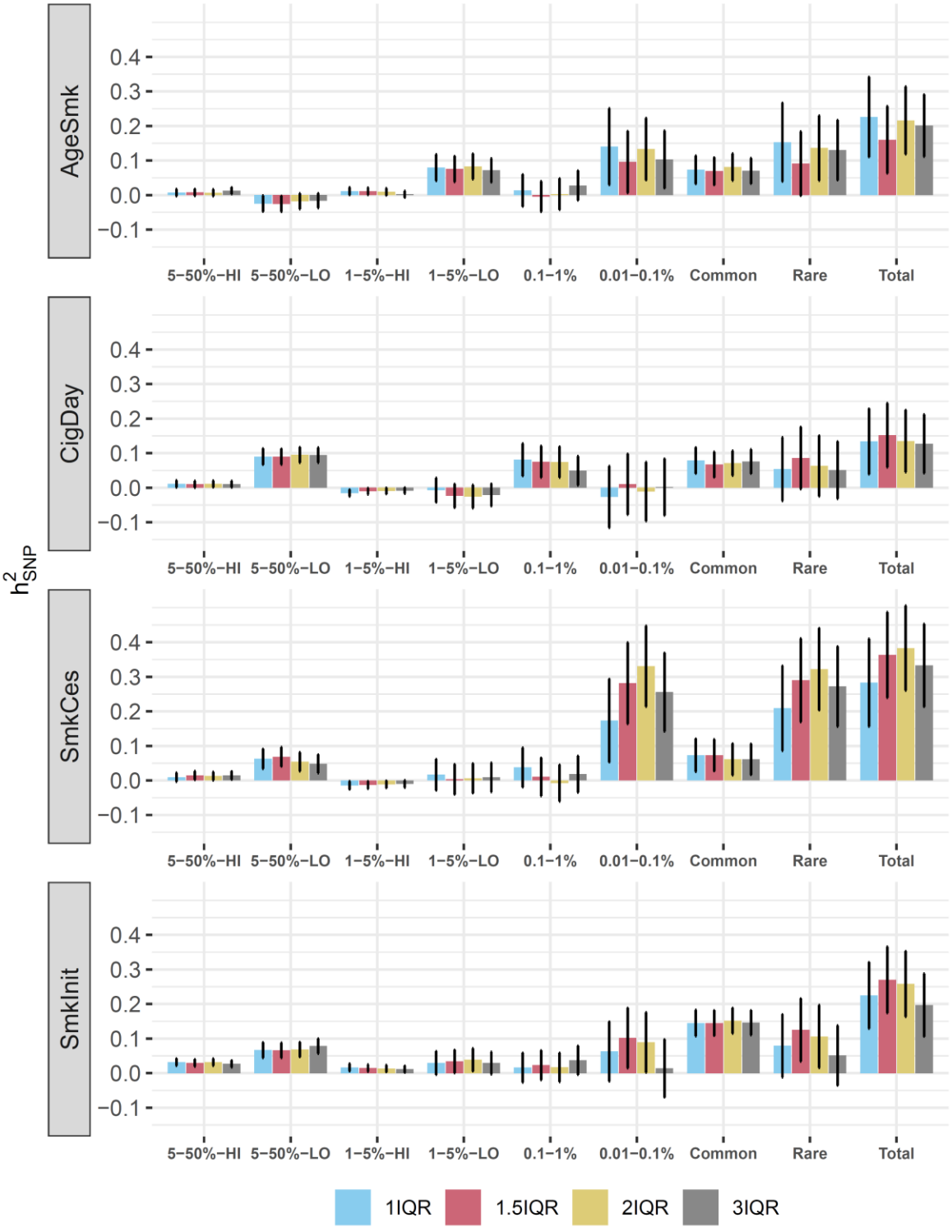
TOPMed Accession #	TOPMed Study Short Name	TOPMed Phase	TOPMed Project	Omics Center Short Name	Omics Support	Omics Type
phs001211	ARIC	2	VTE	Baylor	3U54HG003273-12S2 / HSN268201500015C	WGS
phs001143	BAGS	1	BAGS	Illumina	3U54HG003273-12S2 / HHSN268201500015C	WGS
phs001644	BioMe	3	BioMe	Baylor	HHSN268201600033I	WGS
phs001644	BioMe	3	BioMe	MGI	HHSN268201600037I	WGS
phs001612	CARDIA	3	CARDIA	Baylor	HHSN268201600033I	WGS
phs000954	CFS	1	CFS	NWGC	3R01HL098433-05S1	WGS
phs000954	CFS	3.5	CFS	NWGC	HHSN268201600032I	WGS
phs001368	CHS	3	CHS	Baylor	HHSN268201600033I	WGS
phs001368	CHS VTE	2	VTE	Baylor	3U54HG003273-12S2 / HHSN268201500015C	WGS
phs000951	COPDGene	2.5	COPD	Broad Genomics	HHSN268201500014C	WGS
phs000951	COPDGene	1	COPD	NWGC	3R01HL089856-08S1	WGS
phs000951	COPDGene	2	COPD	Broad Genomics	HHSN268201500014C	WGS
phs001412	DHS	2	AA_CAC	Broad Genomics	HHSN268201500014C	WGS
phs001472	ECLIPSE	3	ECLIPSE	MGI	HHSN268201600037I	WGS

phs000946	EOCOPD	1	COPD	NWGC	3R01HL089856-08S1	WGS
phs000974	FHS	1	FHS	Broad Genomics	3U54HG003067-12S2	WGS
phs001218	GeneSTAR	legacy	GeneSTAR	Illumina	R01HL112064	WGS
phs001218	GeneSTAR	2	GeneSTAR	Psomagen	3R01HL112064-04S1	WGS
phs001218	GeneSTAR	2	AA_CAC	Broad Genomics	HHSN268201500014C	WGS
phs001345	GENOA	2	HyperGEN_GENOA	NWGC	3R01HL055673-18S1	WGS
phs001345	GENOA_AA_CAC	2	AA_CAC	Broad Genomics	HHSN268201500014C	WGS
phs001359	GOLDN	2	GOLDN	NWGC	3R01HL104135-04S1	WGS
phs001395	HCHS_SOL	3	HCHS_SOL	Baylor	HHSN268201600033I	WGS
phs001293	HyperGEN	2	HyperGEN_GENOA	NWGC	3R01HL055673-18S1	WGS
phs000993	HVH	1	AFGen	Broad Genomics	3R01HL092577-06S1	WGS
phs001607	IPF	3	IPF	MGI	HHSN268201600037I	WGS
phs000964	JHS	1	JHS	NWGC	HHSN268201100037C	WGS
phs001416	MESA AA_CAC	2	AA_CAC	Broad Genomics	HHSN268201500014C	WGS
phs001544	MPP	2.4	AFGen	Broad Genomics	3UM1HG008895-01S2	WGS
phs001608	OMG_SCD	2	OMG_SCD	Baylor	HHSN268201500015C	WGS
phs001608	OMG_SCD	7.1	OMG_SCD	Baylor	HHSN268201600033I	WGS

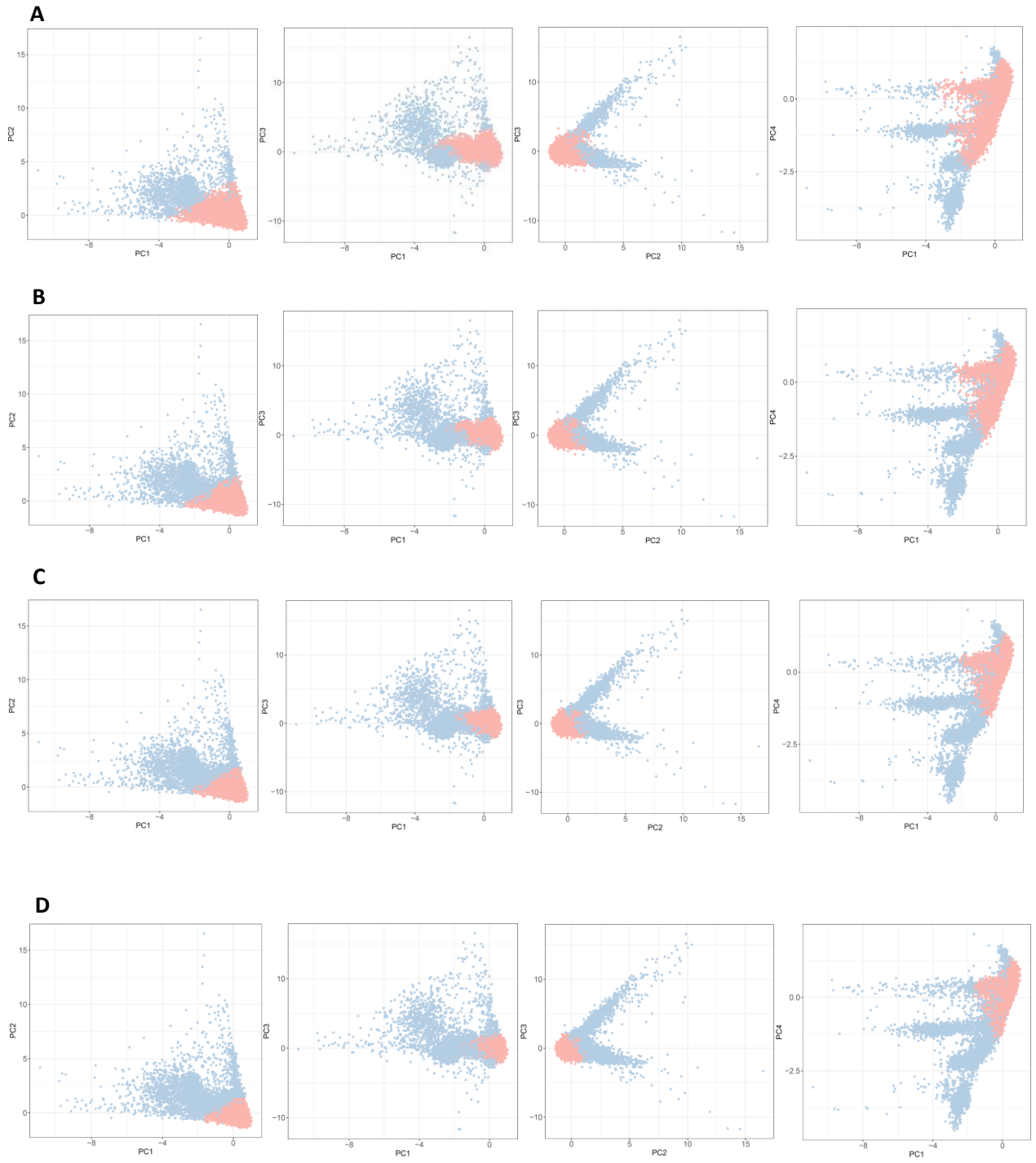
phs001601	PMBB_AF	2.4	AFGen	Broad Genomics	3UM1HG008895-01S2	WGS
phs001468	REDS-III_Brazil	2	REDS-III_Brazil	Baylor	HHSN268201500015C	WGS
phs001468	REDS-III_Brazil	5.4	REDS-III_Brazil	Baylor	HHSN268201600033I	WGS
phs001215	SAFS	1	SAFS	Illumina	3R01HL113323-03S1	WGS
phs001215	SAFS	legacy	SAFS	Illumina	R01HL113322	WGS
phs001207	Sarcoidosis	2	Sarcoidosis	Baylor	3R01HL113326-04S1	WGS
phs001207	Sarcoidosis	3.5, 5.5	Sarcoidosis	NWGC	HHSN268201600032I	WGS
phs001207	Sarcoidosis	3.5, 5.5	Sarcoidosis	Broad Genomics	HHSN268201600034I	WGS
phs001446	SARP	2	SARP	NYGC Genomics	HHSN268201500016C	WGS
phs001402	Mayo_VTE	2	VTE	Baylor	3U54HG003273-12S2 / HSN268201500015C	WGS
phs001514	walk_PHaSST	2	walk_PHaSST	Baylor	HHSN268201500015C	WGS
phs001237	WHI	2	WHI	Broad Genomics	HHSN268201500014C	WGS

Supplementary Note 5. Full TOPMed Banner Authorship list

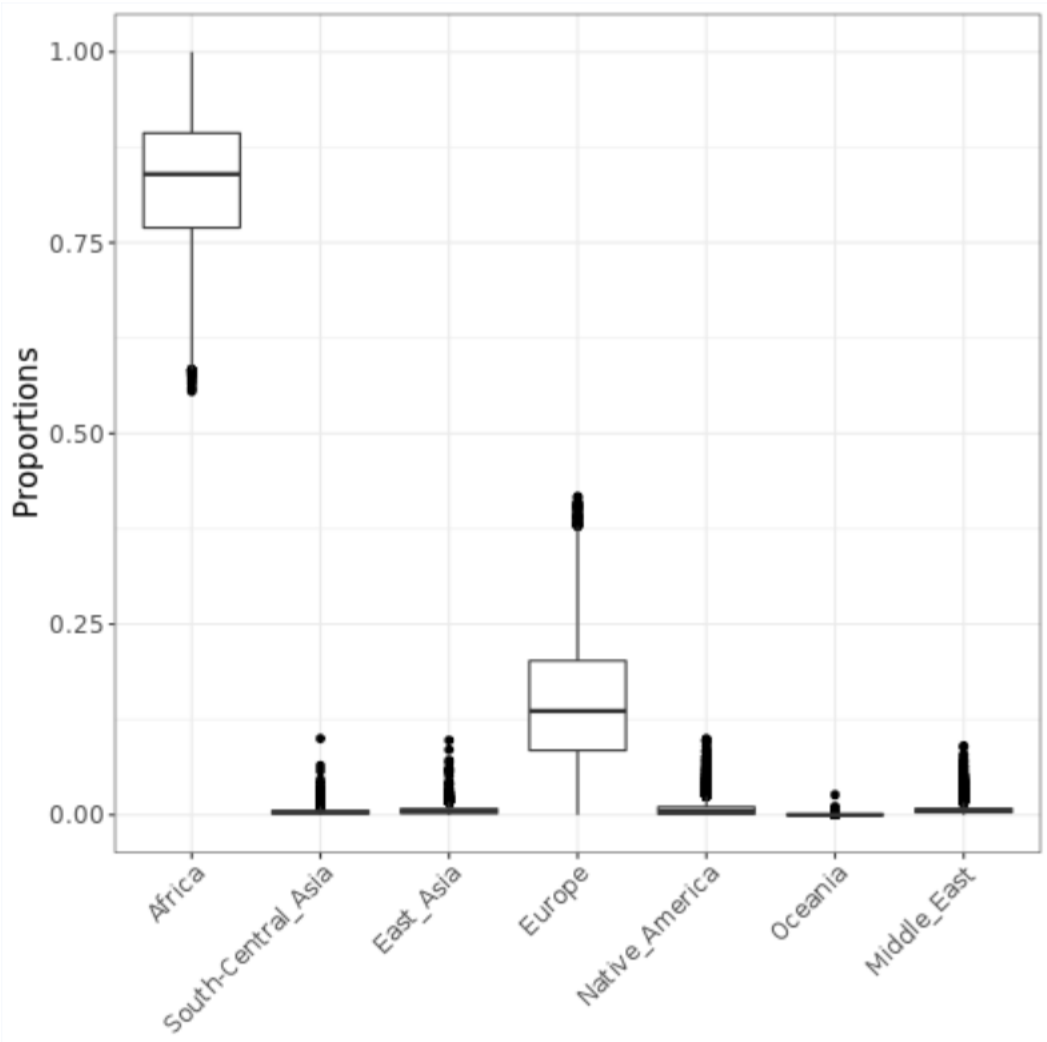
Full TOPMed Banner authorship list can be found here: https://www.nhlbiwgs.org/topmed-banner-authorship



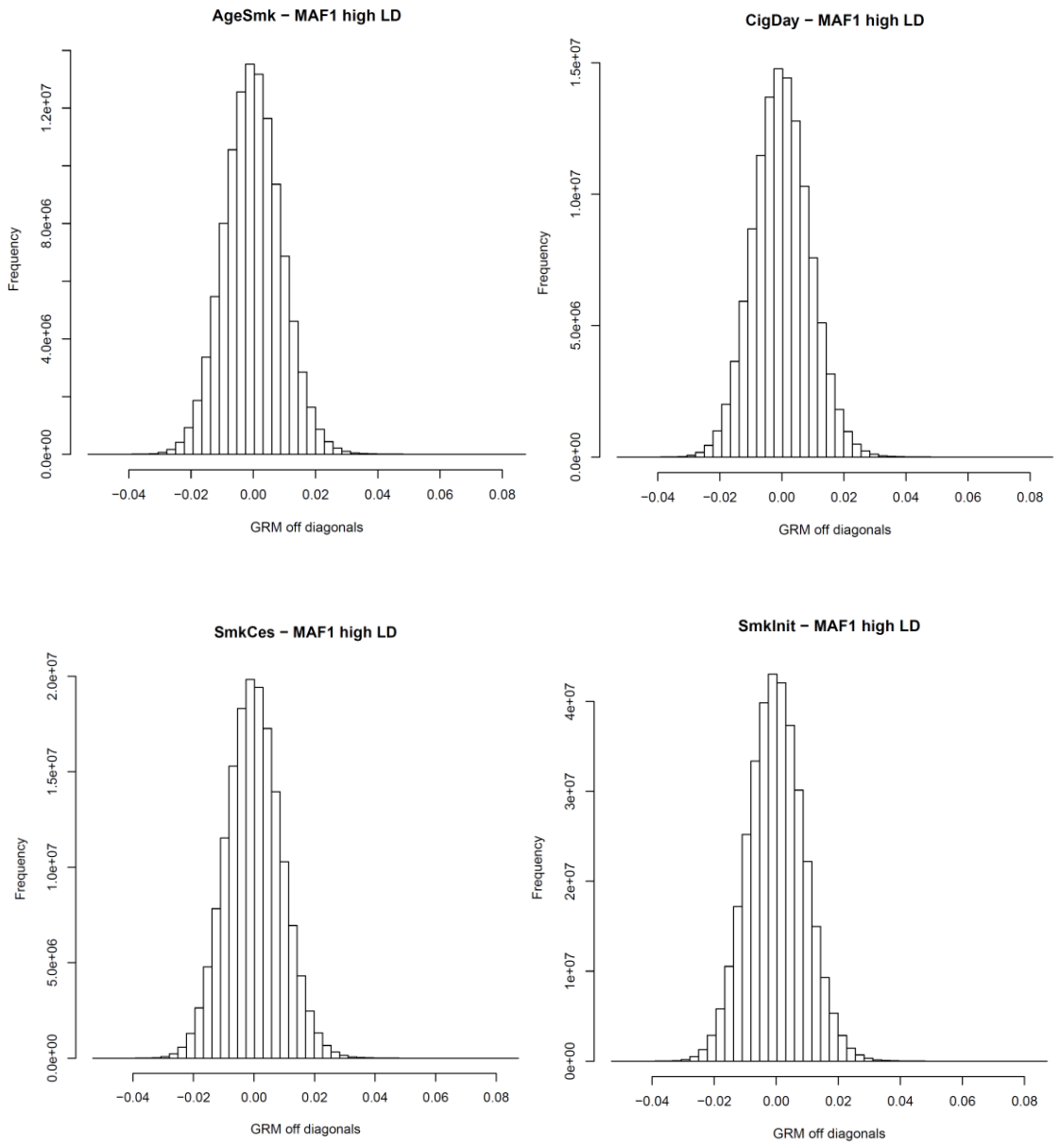
Supplementary Figure 1. SNP-based heritability estimates and standard errors from samples with increasingly less control of population structure with 1QR being the most conservative and 3IQR being the most liberal. X-axis indicates LDMS bins and their sums (Common: a sum across 5-50%-HI, 5-50%-LO, 1-5%-HI, 1-5%-LO bins; Rare: a sum across 0.1-1%, 0.01-0.1% bins; Total: a sum across all six bins). “HI” and “LO” each means variants with high or low LD scores as determined by the median LD scores of each bin.



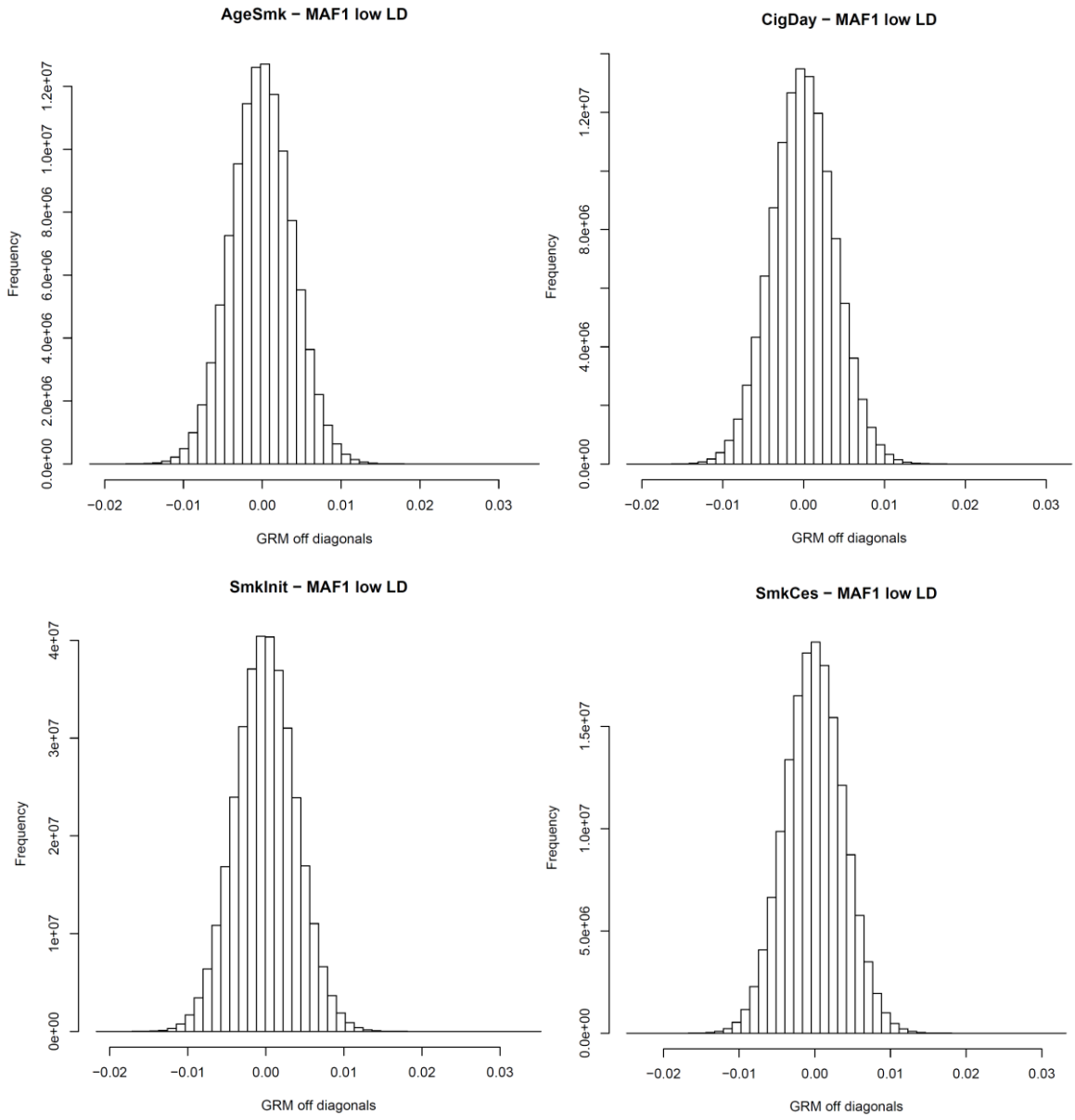
Supplementary Figure 2. Principal components 1-4 of four different IQR samples. A, B, C, D each shows samples lying inside the 3*IQR, 2*IQR, 1.5*IQR, and 1*IQR range of the sum of the Euclidian distance of the PC 1 to 4. Red and blue dots each indicate samples included and excluded, respectively.



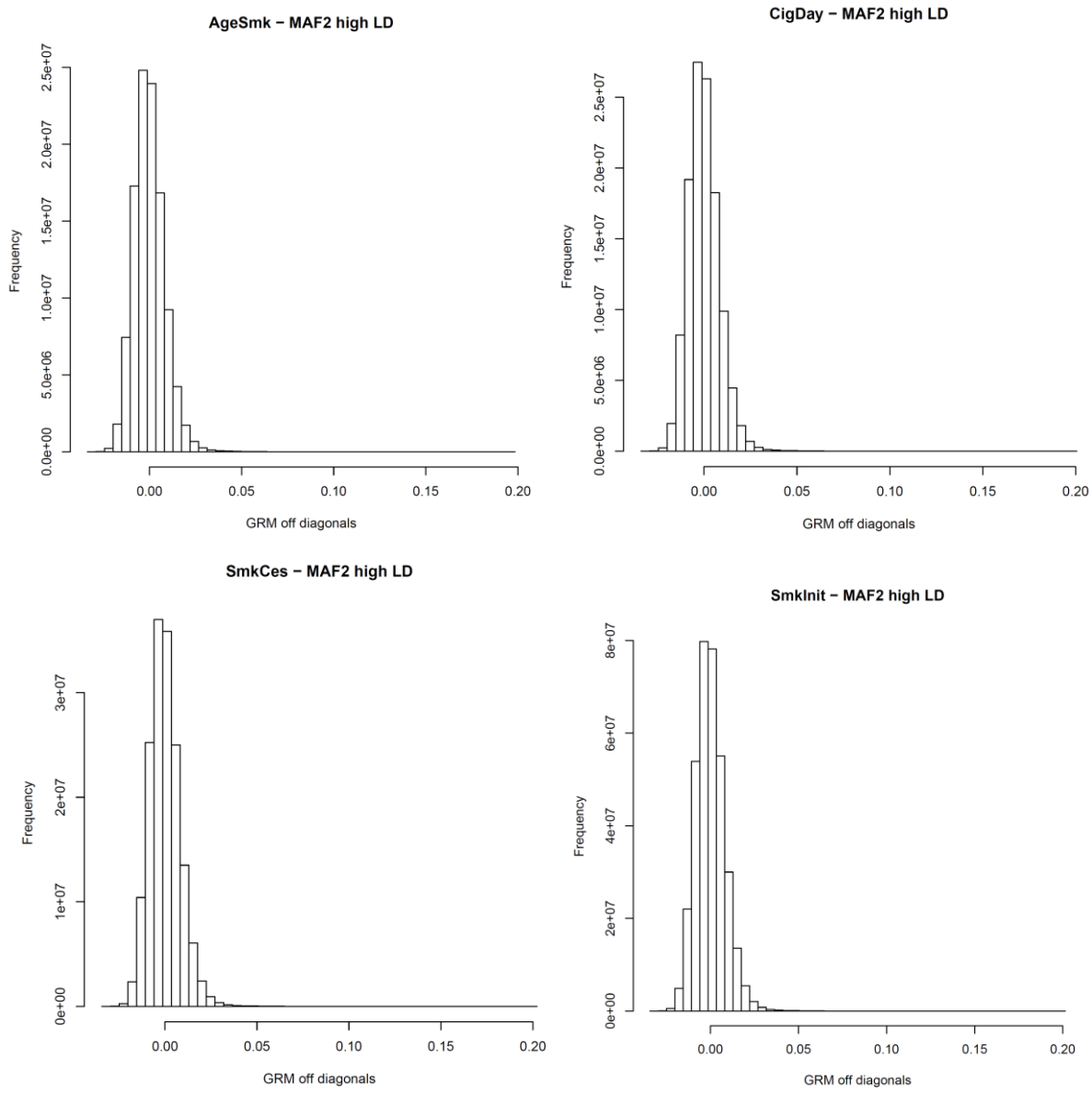
Supplementary Figure 3. Global ancestry proportions of African ancestry sample (N=11,743). Samples used in SmkInit analysis were shown here. Individuals whose ancestry proportions are above 3Q + 1.5 IQR or below 1Q - 1.5 IQR are shown in black dots.



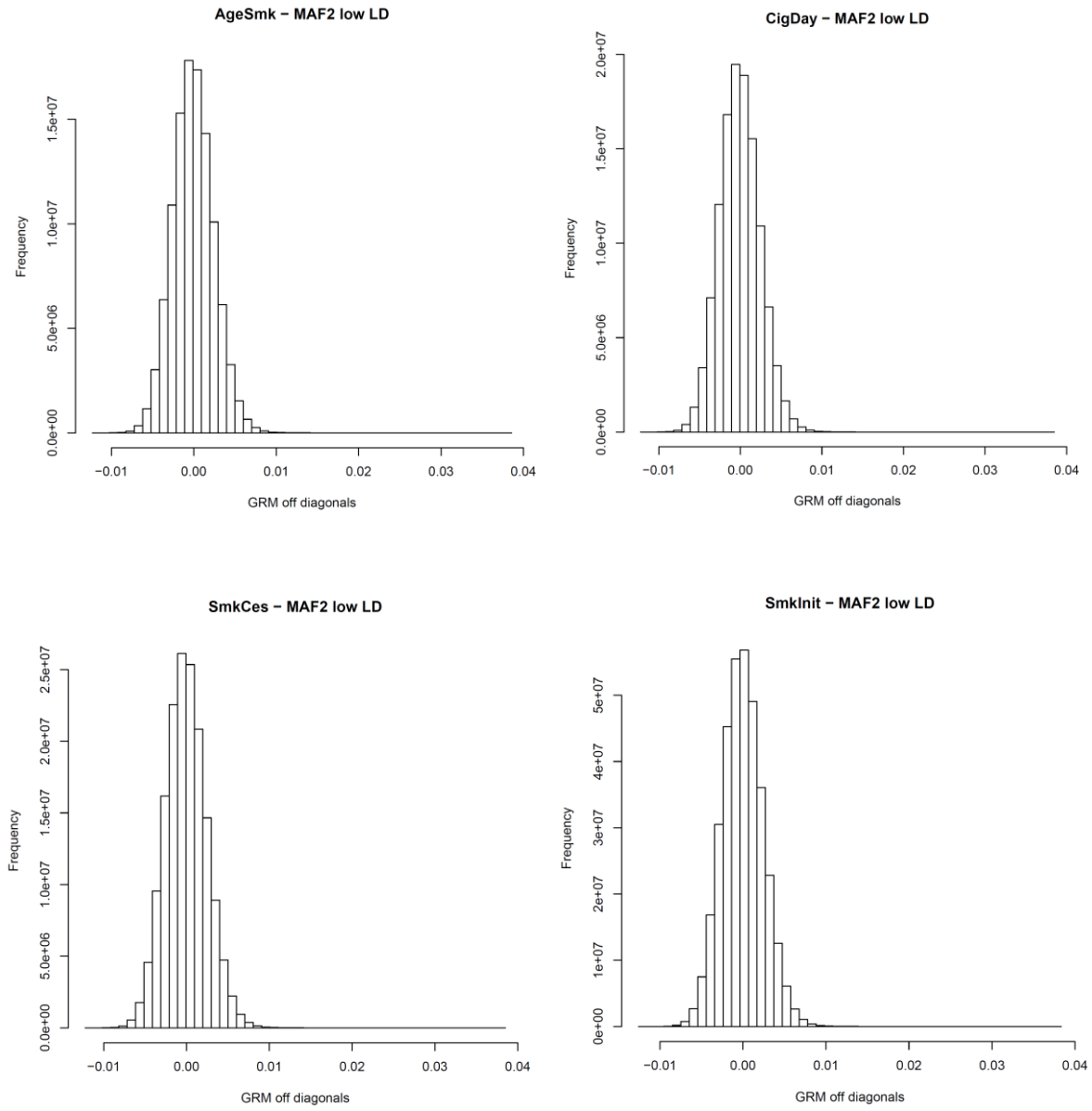
Supplementary Figure 4-1. Histogram of GRM off-diagonals from MAF 5-50% - high LD bin in European ancestry sample



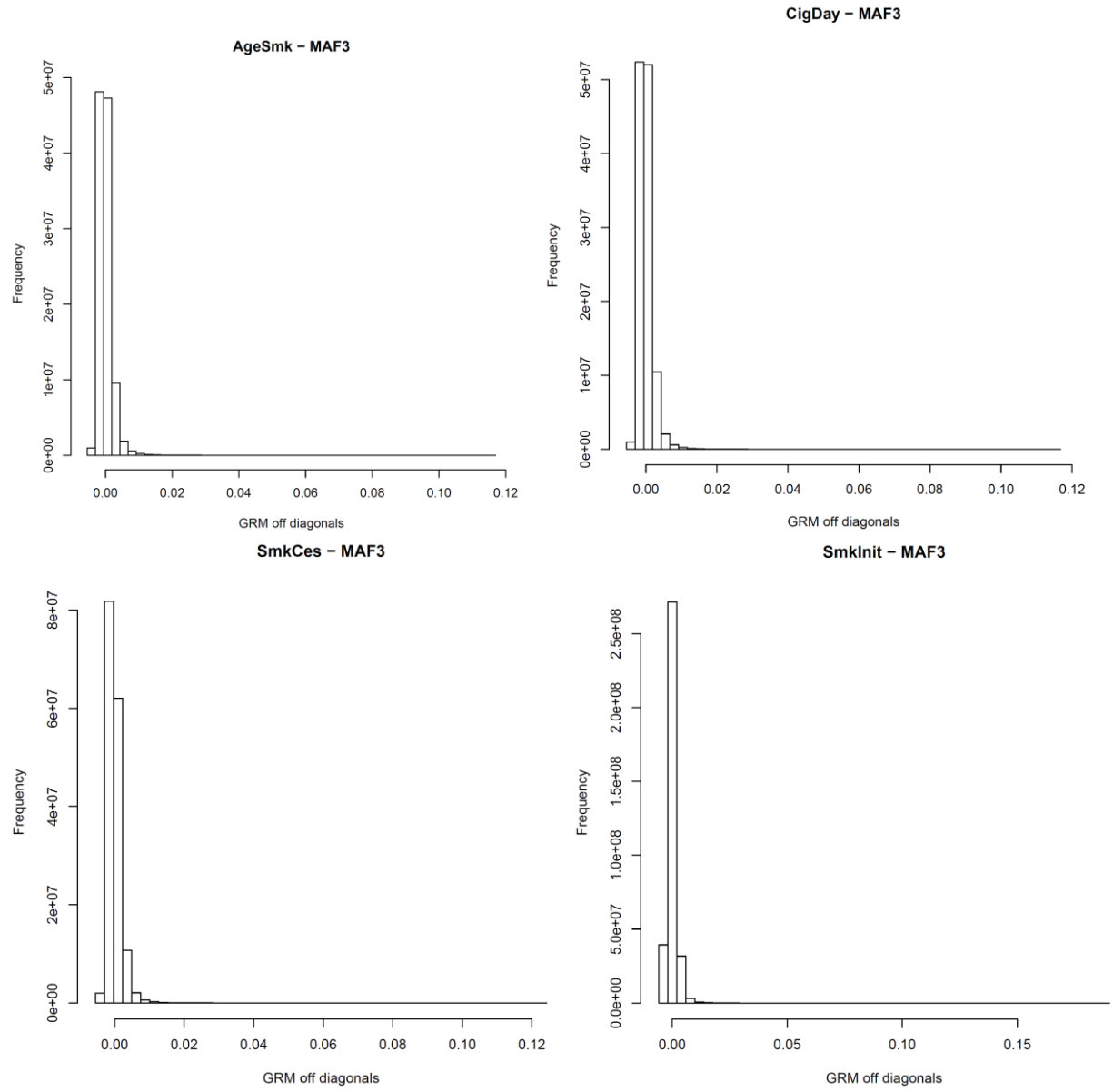
Supplementary Figure 4-2. Histogram of GRM off-diagonals from MAF 5-50% - low LD bin in European ancestry sample



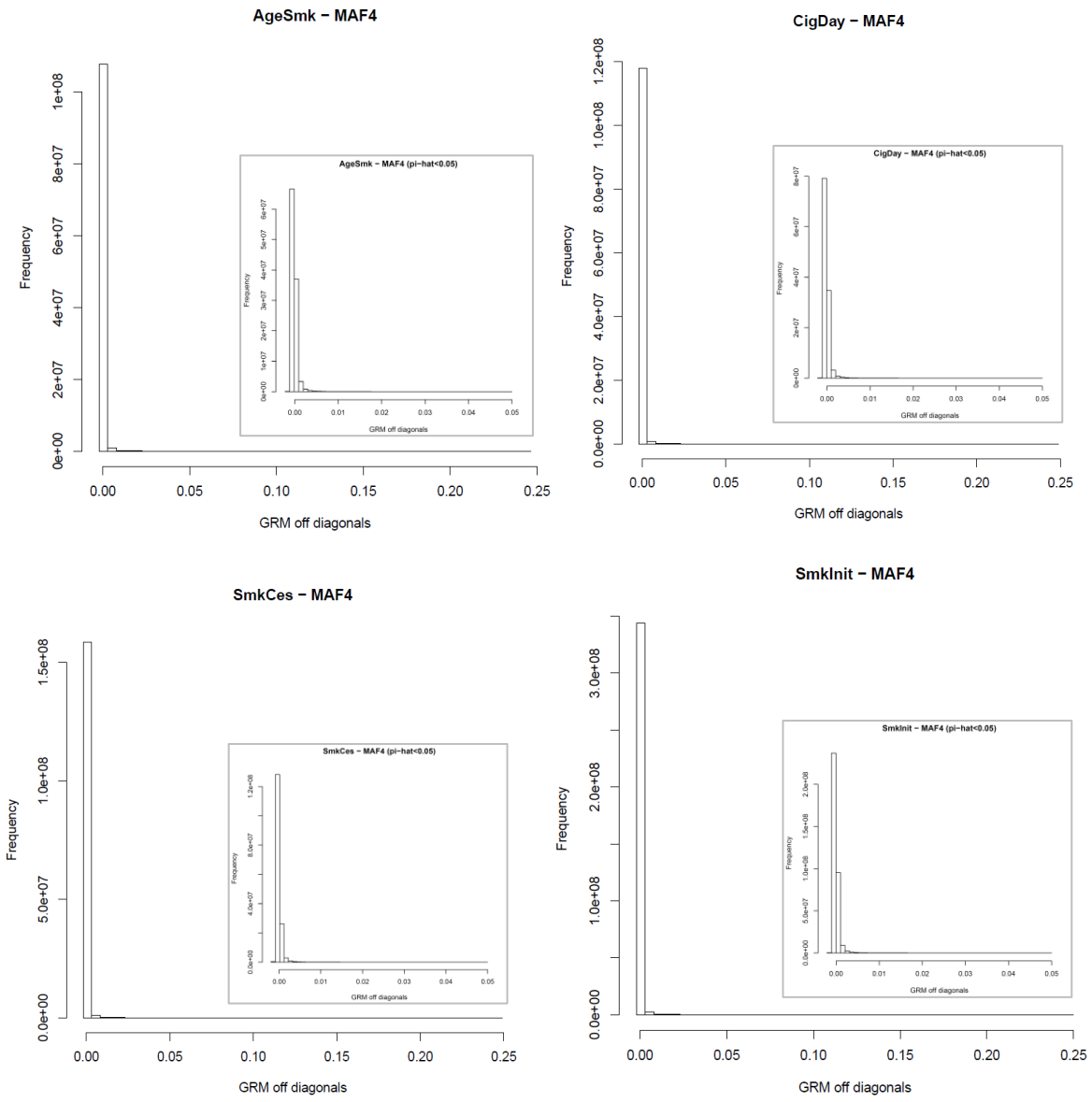
Supplementary Figure 4-3. Histogram of GRM off-diagonals from MAF 1-5% - high LD bin in European ancestry sample



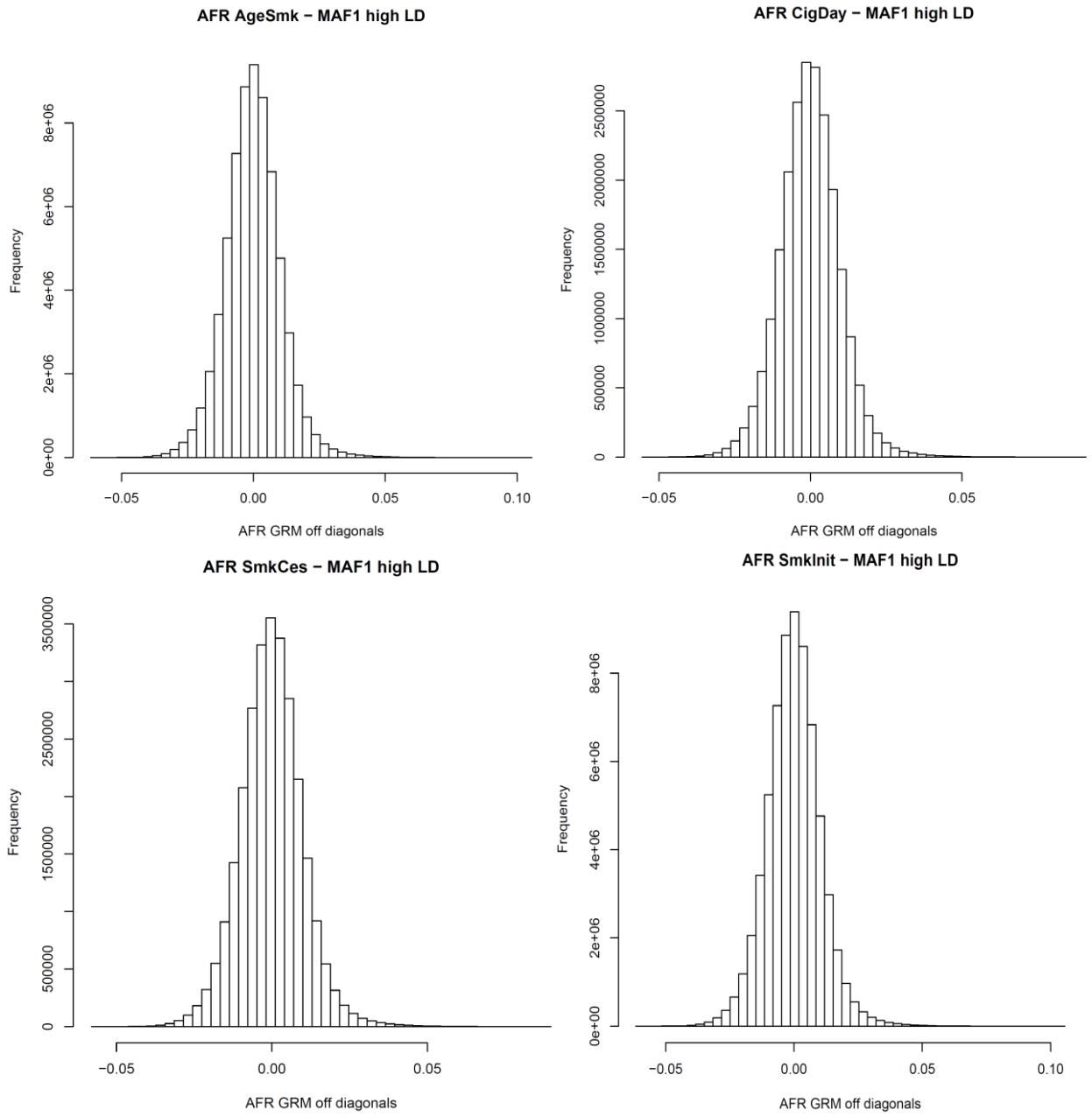
Supplementary Figure 4-4. Histogram of GRM off-diagonals from MAF 1-5% - low LD bin in European ancestry sample



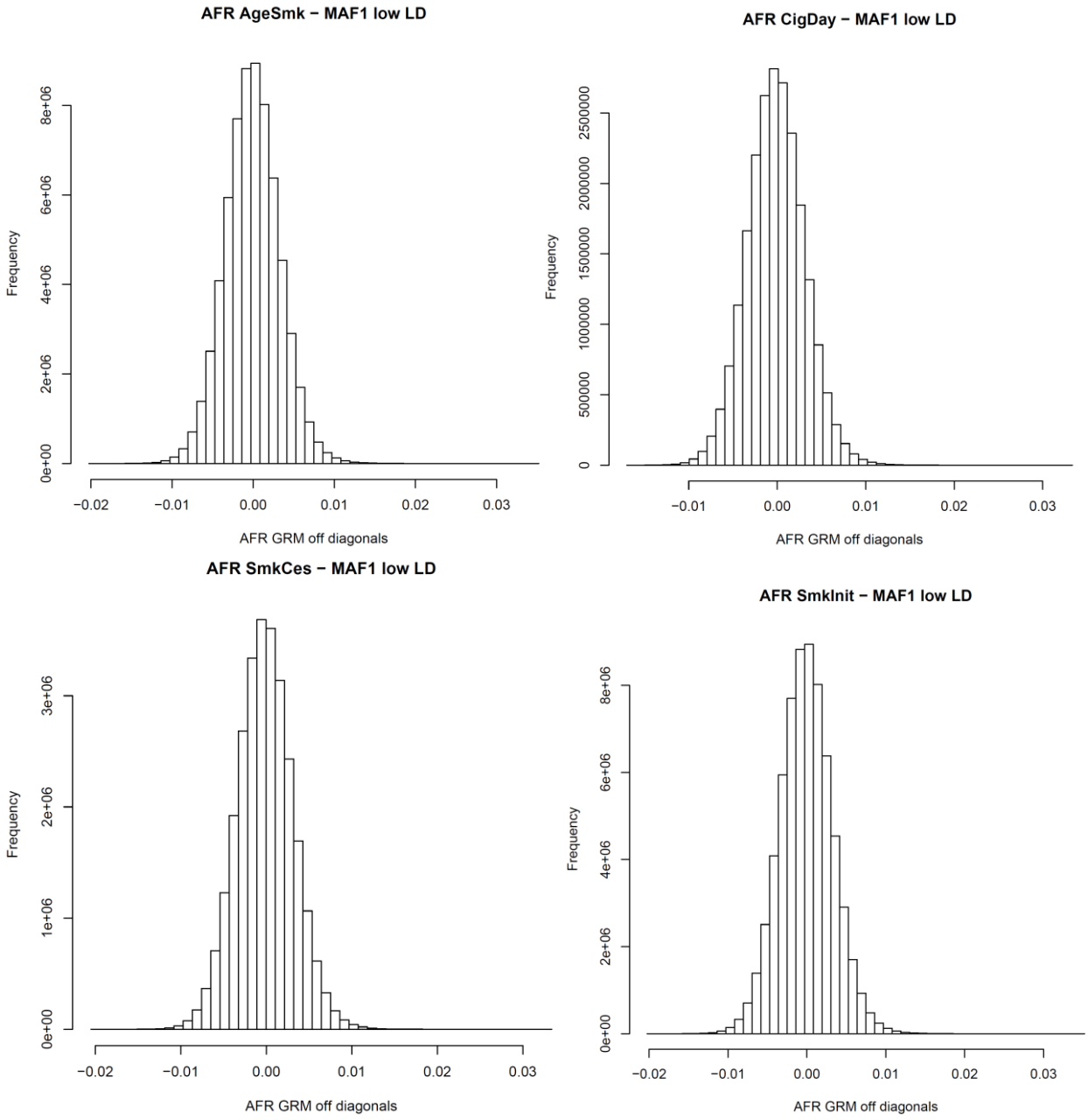
Supplementary Figure 4-5. Histogram of GRM off-diagonals from MAF 0.1-1% bin in European ancestry sample



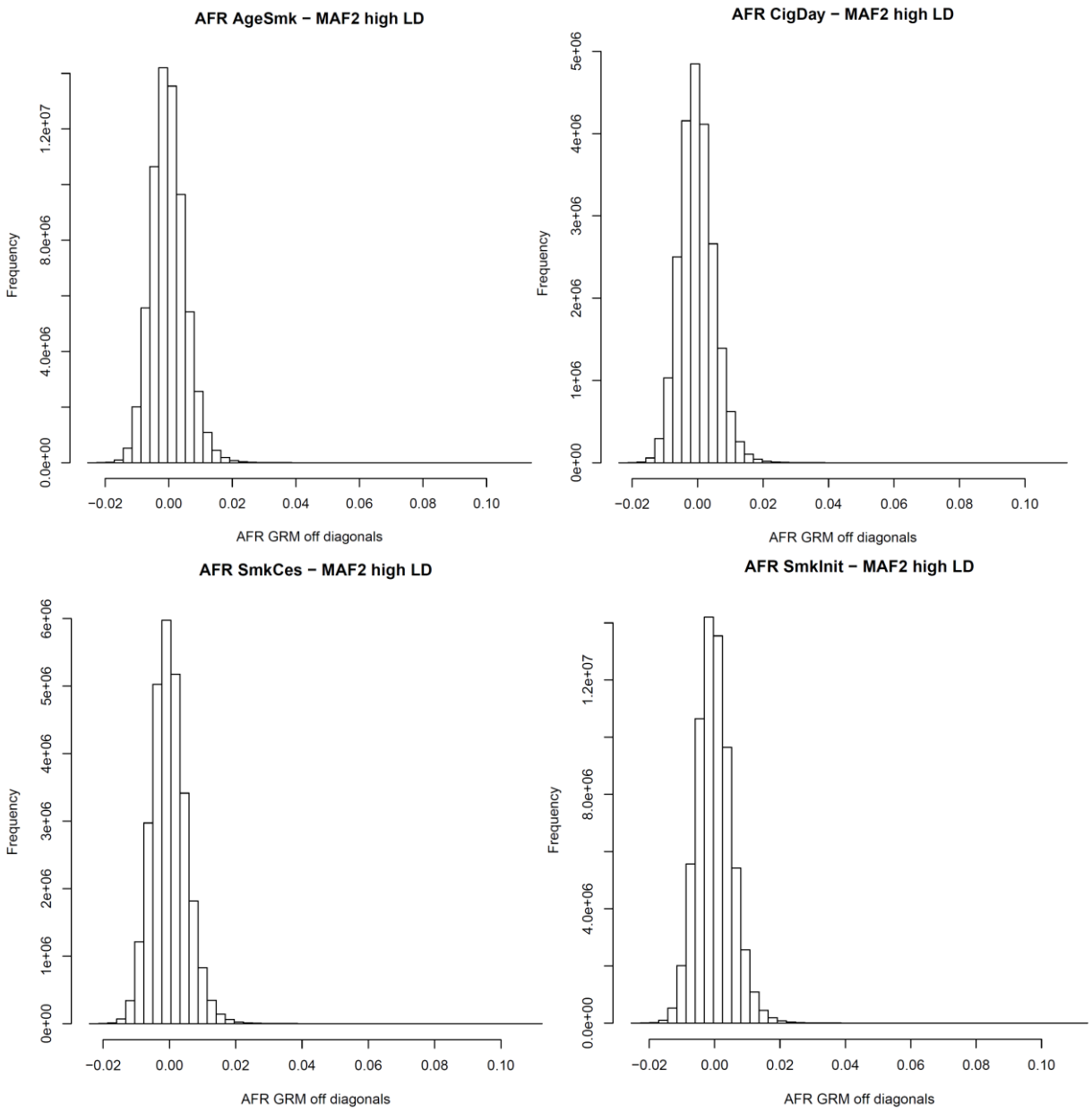
Supplementary Figure 4-6. Histogram of GRM off-diagonals from MAF 0.01-0. 1% bin in European ancestry sample. Histograms of off-diagonals < 0.05 were presented in a smaller box for finer scale distribution.



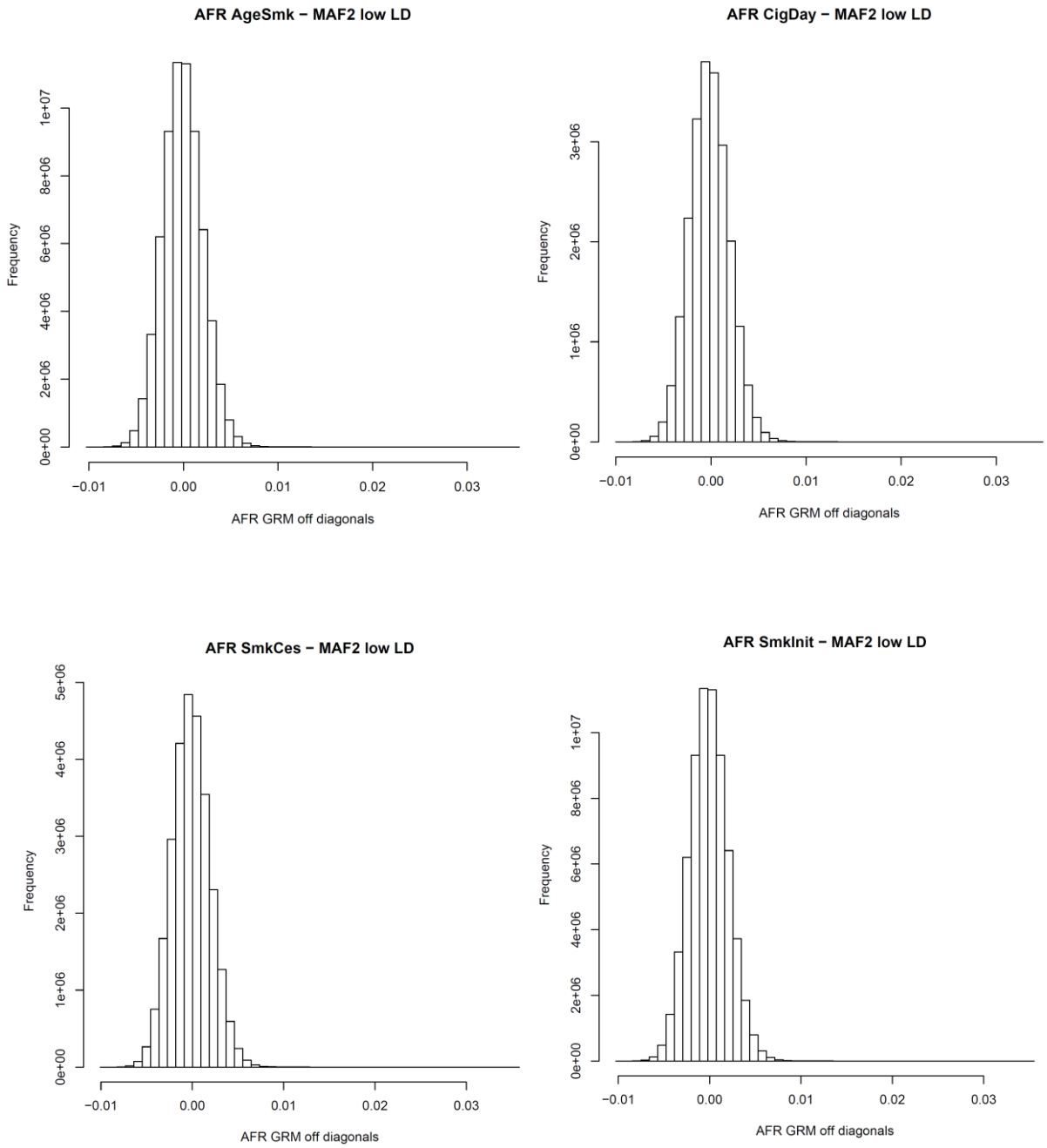
Supplementary Figure 5-1. Histogram of GRM off-diagonals from MAF 5-50% - high LD bin in African ancestry sample



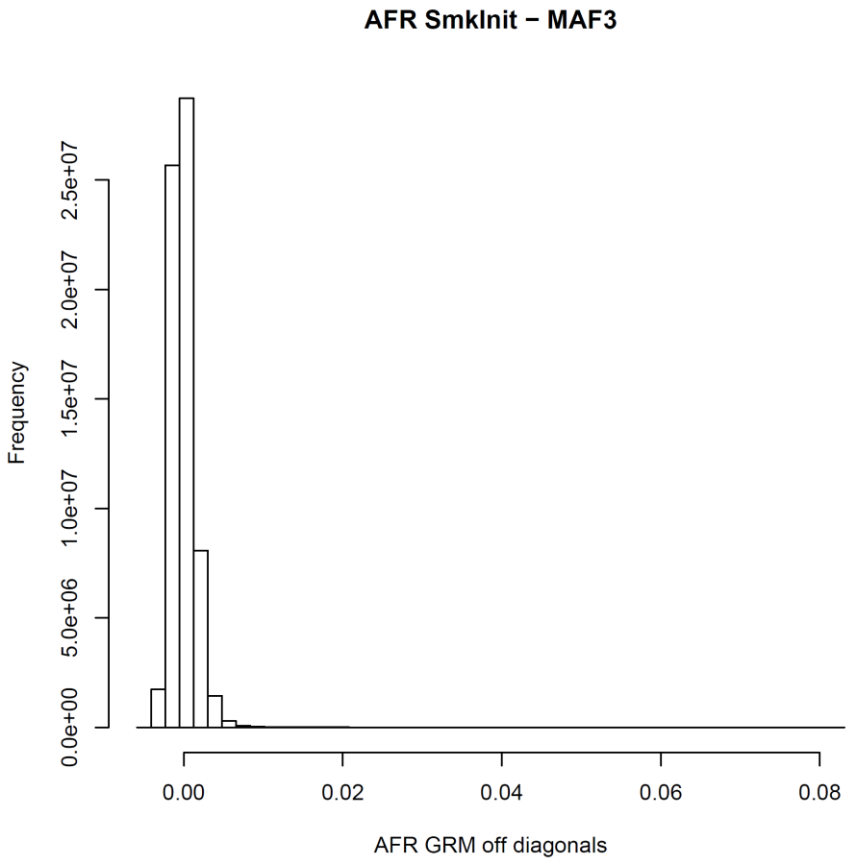
Supplementary Figure 5-2. Histogram of GRM off-diagonals from MAF 5-50% - low LD bin in African ancestry sample



Supplementary Figure 5-3. Histogram of GRM off-diagonals from MAF 1-5% - high LD bin in African ancestry sample



Supplementary Figure 5-4. Histogram of GRM off-diagonals from MAF 1-5% - low LD bin in African ancestry sample



Supplementary Figure 5-5. Histogram of GRM off-diagonals from MAF 0.1-1% in bin African ancestry sample