**The American Journal of Human Genetics, Volume *108***

## Supplemental information

# Whole-genome sequencing in diverse subjects identifies genetic correlates of leukocyte traits: The NHLBI TOPMed program

**Anna V. Mikhaylova, Caitlin P. McHugh, Linda M. Polfus, Laura M. Raffield, Meher Preethi Boorgula, Thomas W. Blackwell, Jennifer A. Brody, Jai Broome, Nathalie Chami, Ming-Huei Chen, Matthew P. Conomos, Corey Cox, Joanne E. Curran, Michelle Daya, Lynette Ekunwe, David C. Glahn, Nancy Heard-Costa, Heather M. Highland, Brian D. Hobbs, Yann Ilboudo, Deepti Jain, Leslie A. Lange, Tyne W. Miller-Fleming, Nancy Min, Jee-Young Moon, Michael H. Preuss, Jonathon Rosen, Kathleen Ryan, Albert V. Smith, Quan Sun, Praveen Surendran, Paul S. de Vries, Klaudia Walter, Zhe Wang, Marsha Wheeler, Lisa R. Yanek, Xue Zhong, Goncalo R. Abecasis, Laura Almasy, Kathleen C. Barnes, Terri H. Beaty, Lewis C. Becker, John Blangero, Eric Boerwinkle, Adam S. Butterworth, Sameer Chavan, Michael H. Cho, Hélène Choquet, Adolfo Correa, Nancy Cox, Dawn L. DeMeo, Nauder Faraday, Myriam Fornage, Robert E. Gerszten, Lifang Hou, Andrew D. Johnson, Eric Jorgenson, Robert Kaplan, Charles Kooperberg, Kousik Kundu, Cecelia A. Laurie, Guillaume Lettre, Joshua P. Lewis, Bingshan Li, Yun Li, Donald M. Lloyd-Jones, Ruth J.F. Loos, Ani Manichaikul, Deborah A. Meyers, Braxton D. Mitchell, Alanna C. Morrison, Debby Ngo, Deborah A. Nickerson, Suraj Nongmaithem, Kari E. North, Jeffrey R. O'Connell, Victor E. Ortega, Nathan Pankratz, James A. Perry, Bruce M. Psaty, Stephen S. Rich, Nicole Soranzo, Jerome I. Rotter, Edwin K. Silverman, Nicholas L. Smith, Hua Tang, Russell P. Tracy, Timothy A. Thornton, Ramachandran S. Vasan, Joe Zein, Rasika A. Mathias, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Alexander P. Reiner, and Paul L. Auer**

## Supplementary Figures

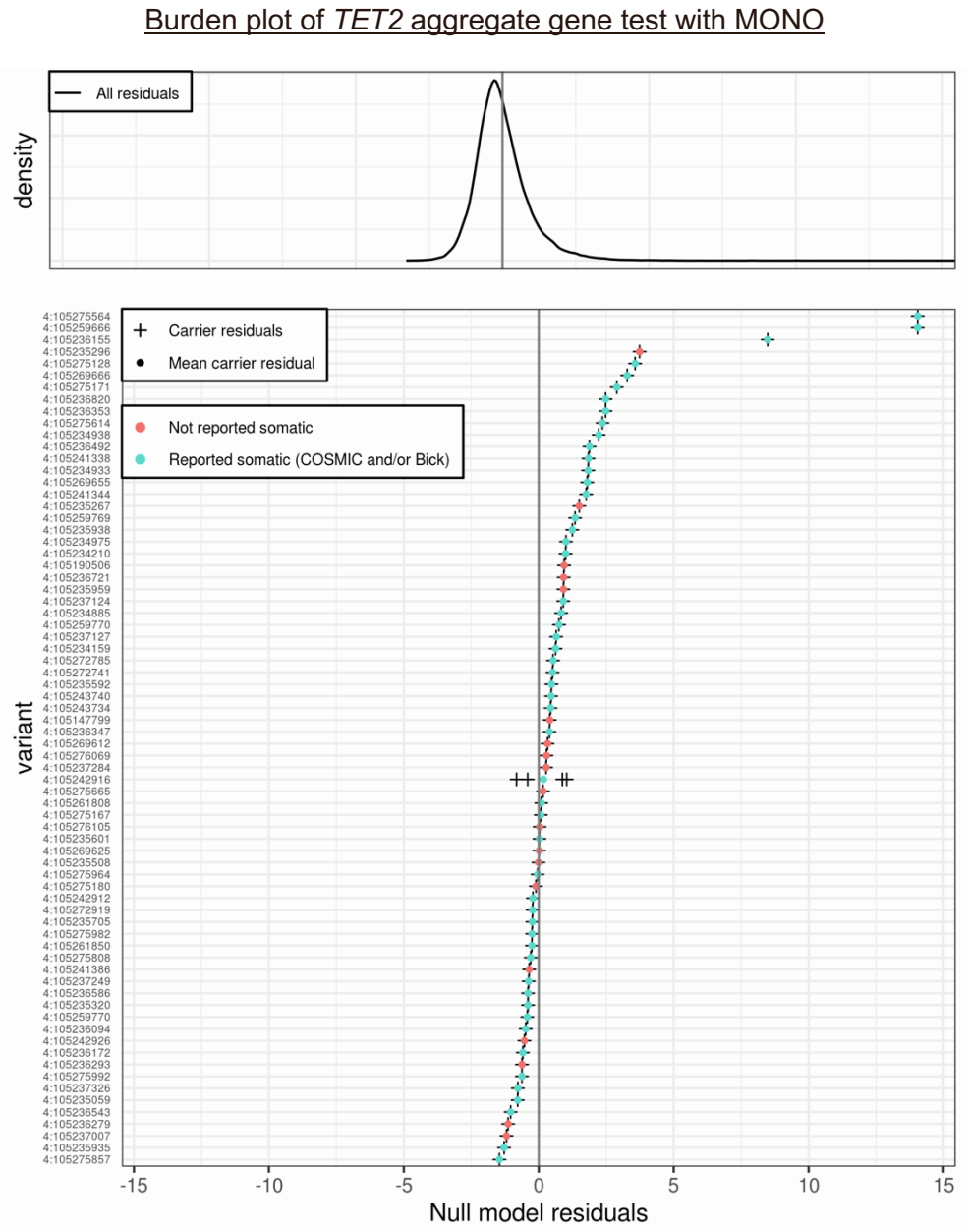


**Figure S1:** The residuals from the null model for 72 rare variants aggregated into a unit for testing the *TET2* gene region, colored by whether the variant was reported as a somatic mutation in the COSMIC database or in Bick, et al. For each carrier of the variant, a + indicates the null model residual value and a circle indicates the mean value for each variant. The density of all residuals is shown in the top panel.

Burden plot of *FLT3* aggregate gene test with MONO

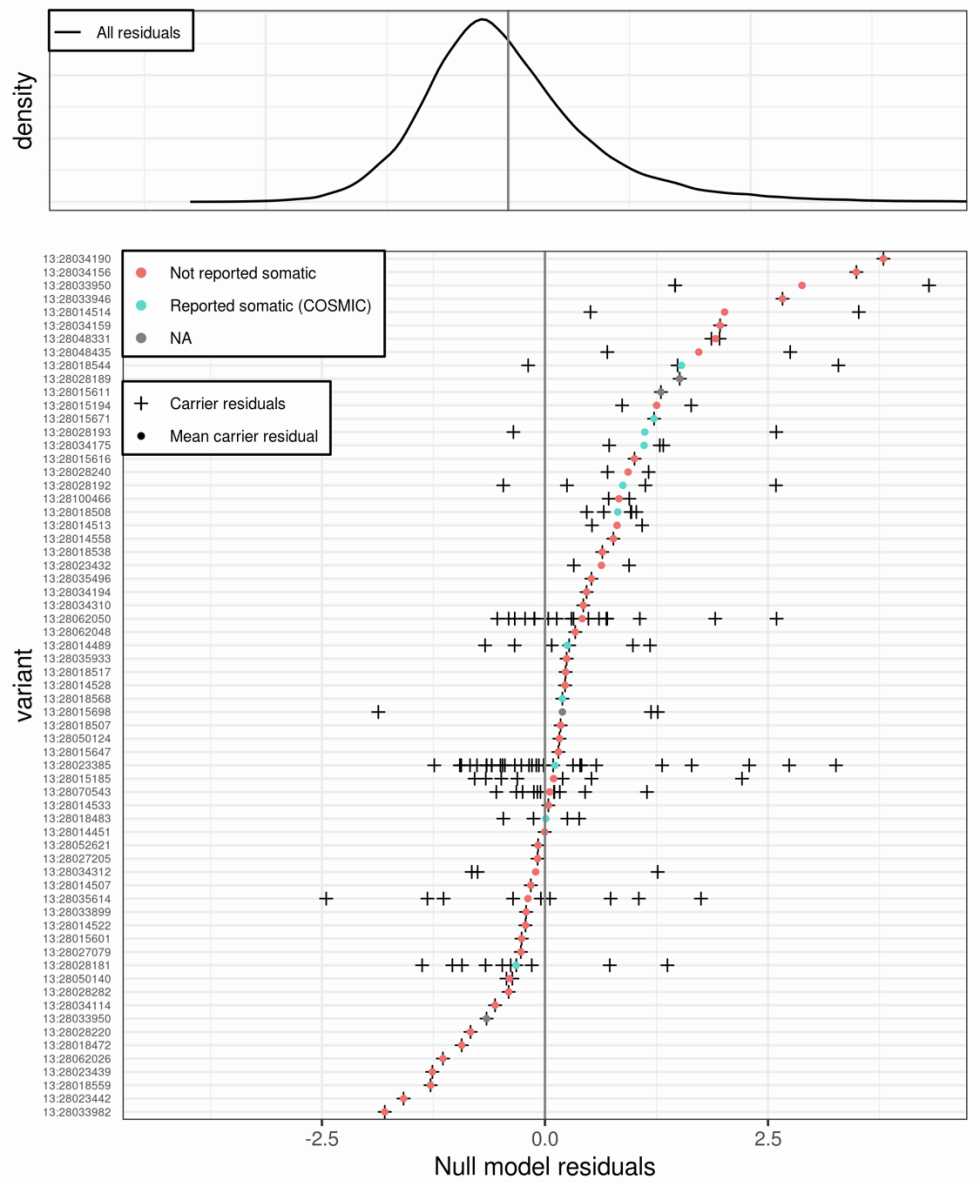


**Figure S2:** The residuals from the null model for 65 rare variants aggregated for testing the *FLT3* gene region, colored by whether the variant was reported as a somatic mutation in the COSMIC database. For each carrier of the variant, a + indicates the null model residual value and a circle indicates the mean value for each variant. The density of all residuals is shown in the top panel.

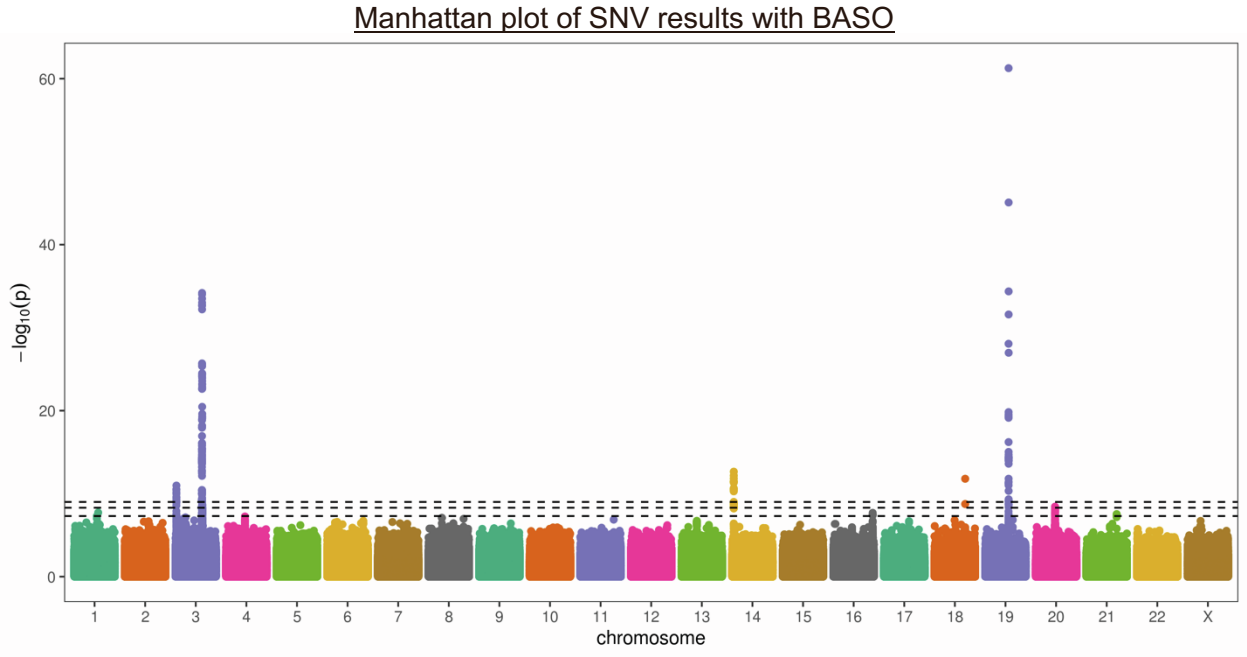


**Figure S3:** Manhattan plot of variants tested for association with binarized basophil count outcome.

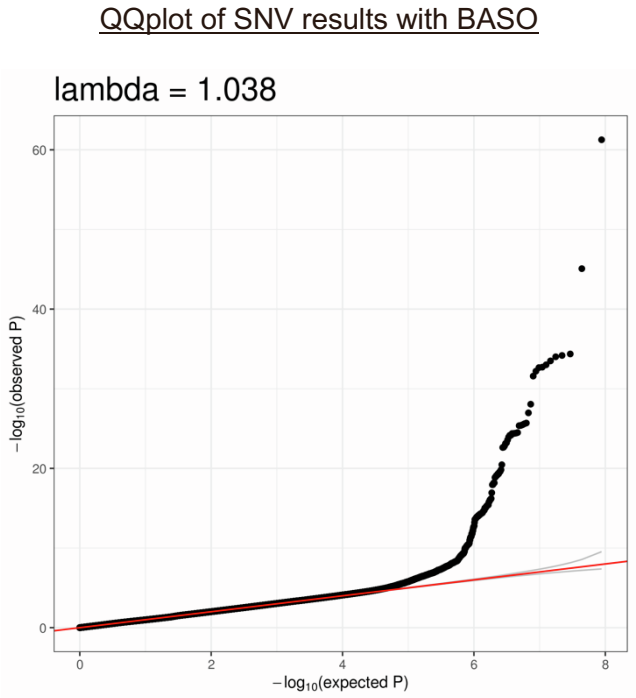


**Figure S4:** QQ plot of variants tested for association with the binarized basophil count outcome.

Manhattan plot of SNV results with EOS

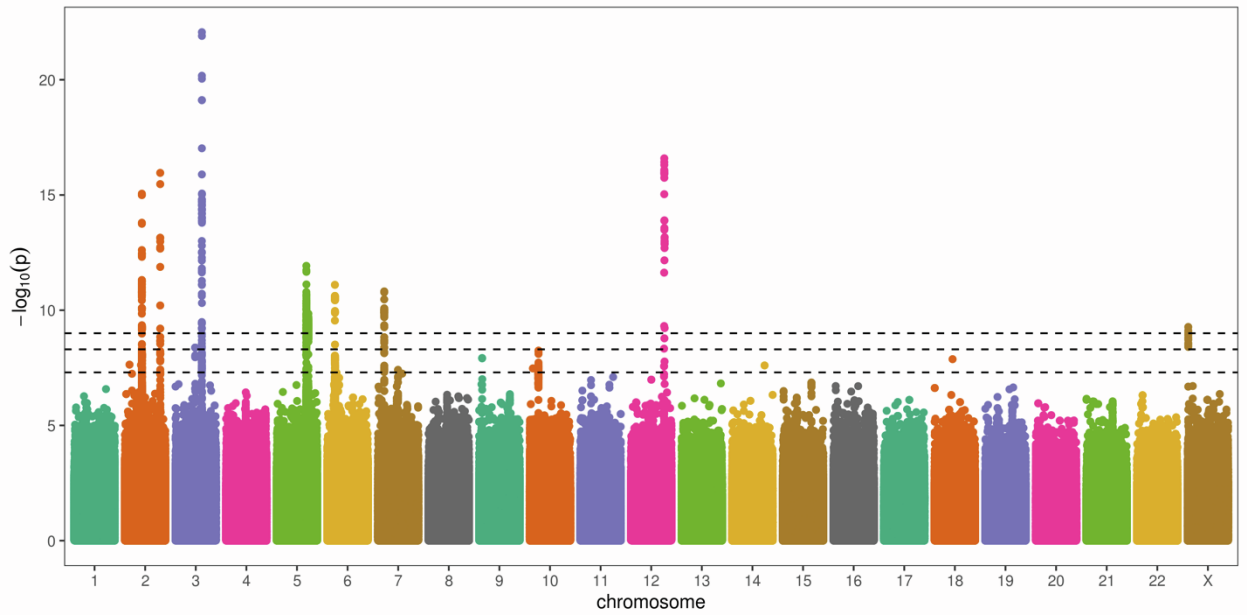


**Figure S5:** Manhattan plot of variants tested for association with the eosinophil percentage outcome.

QQplot of SNV results with EOS

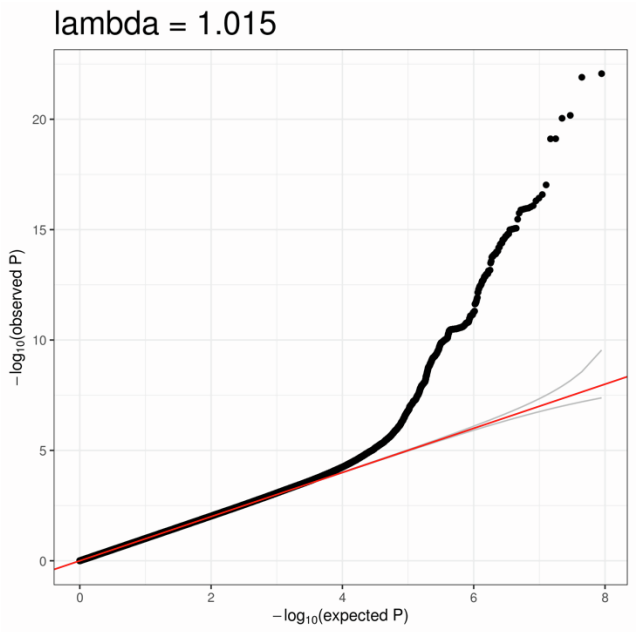


**Figure S6:** QQ plot of variants tested for association with the eosinophil percentage outcome.

Manhattan plot of SNV results with LYM

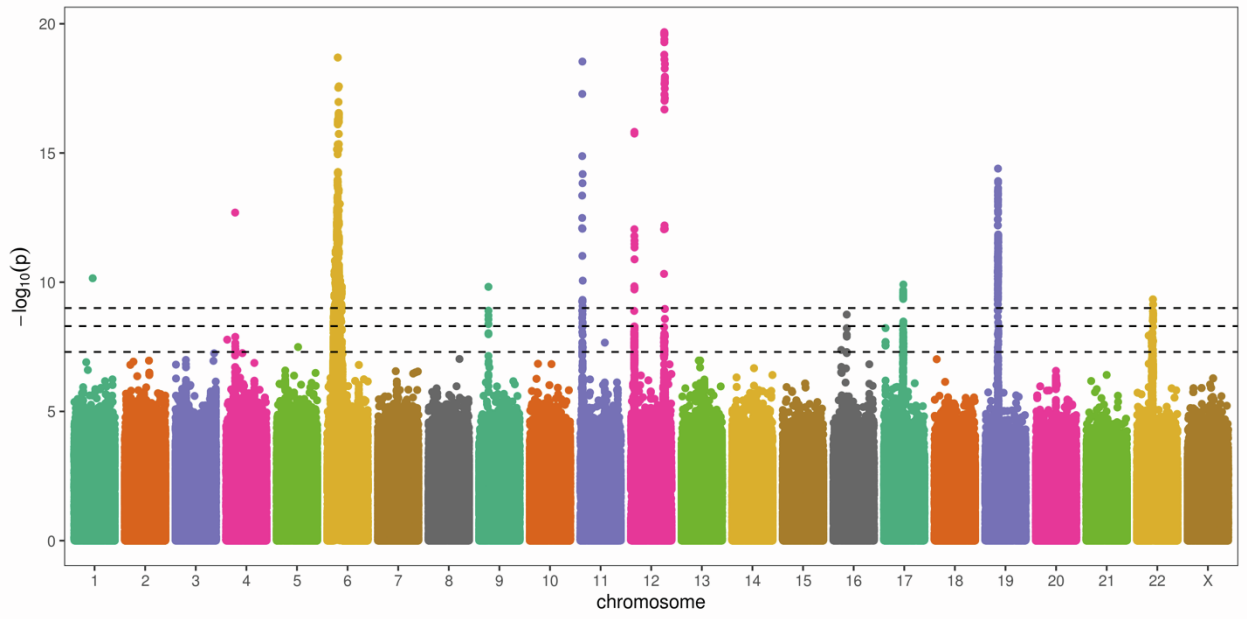


**Figure S7:** Manhattan plot of variants tested for association with the lymphocyte count outcome.

QQplot of SNV results with LYM

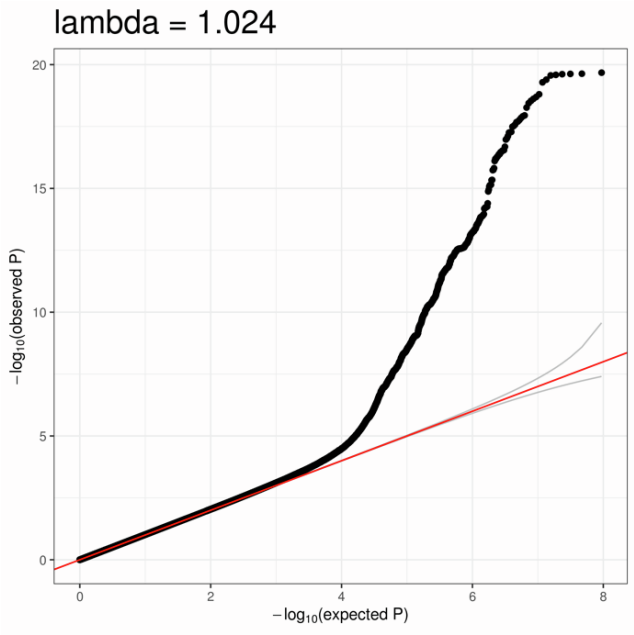


**Figure S8:** QQ plot of variants tested for association with the lymphocyte count outcome.

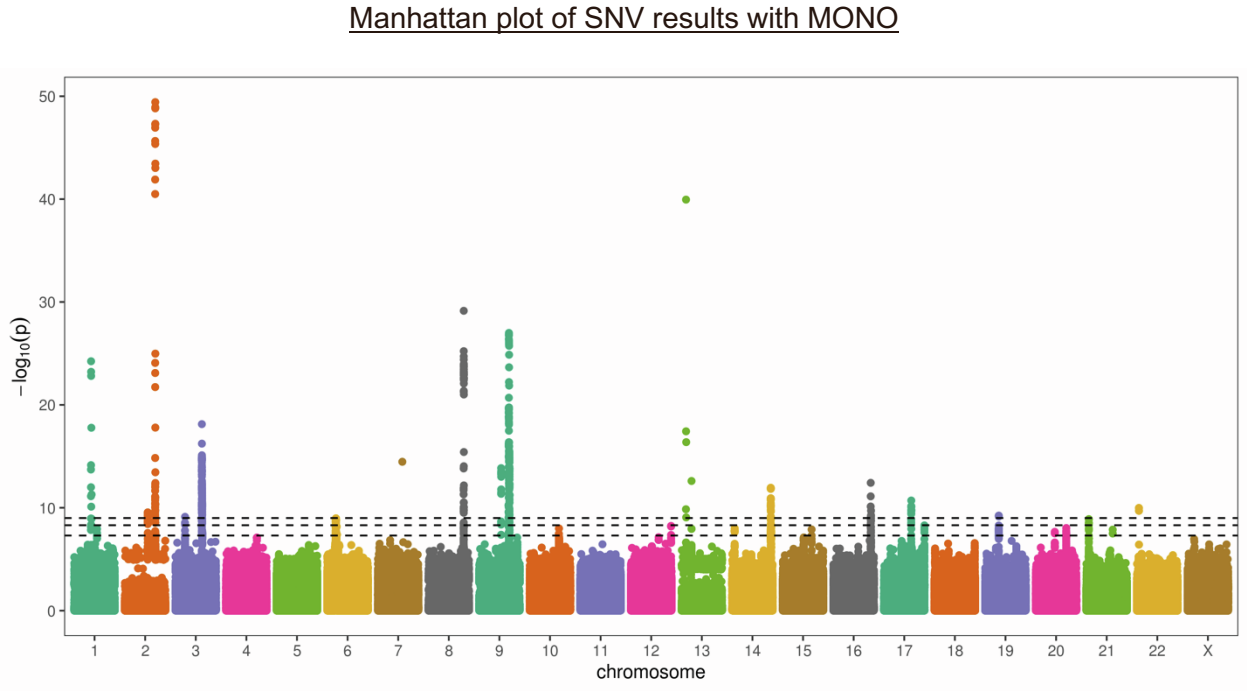


**Figure S9:** Manhattan plot of variants tested for association with the monocyte count outcome.

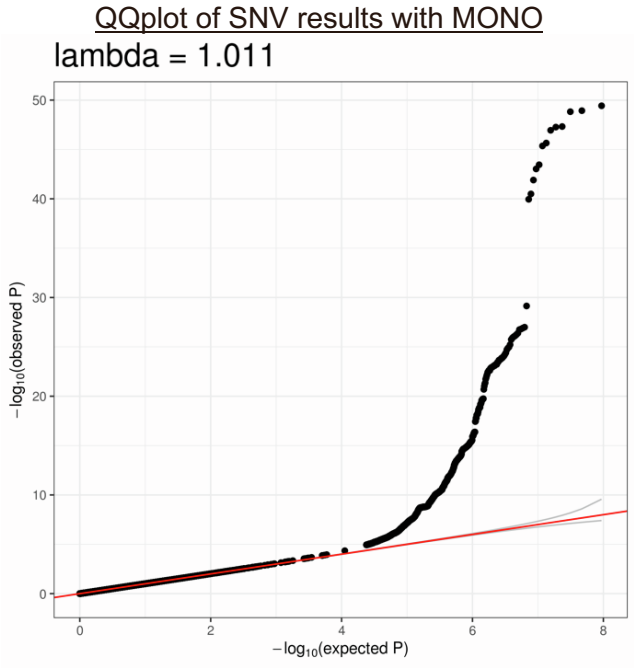


**Figure S10:** QQ plot of variants tested for association with the monocyte count outcome.

Manhattan plot of SNV results with NEU

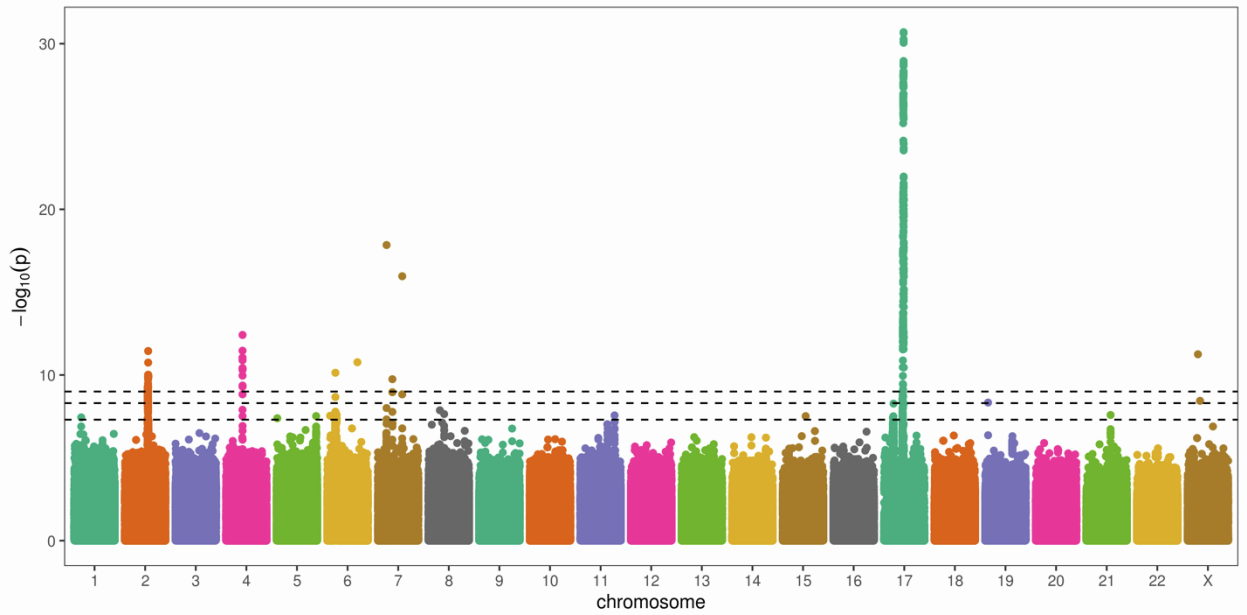


**Figure S11:** Manhattan plot of variants tested for association with the neutrophil count outcome.

QQplot of SNV results with NEU

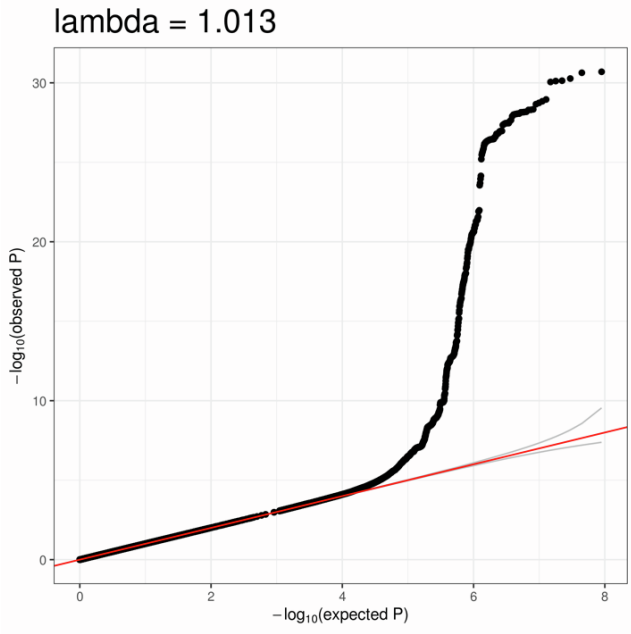


**Figure S12:** QQ plot of variants tested for association with the neutrophil count outcome.

Manhattan plot of SNV results with WBC

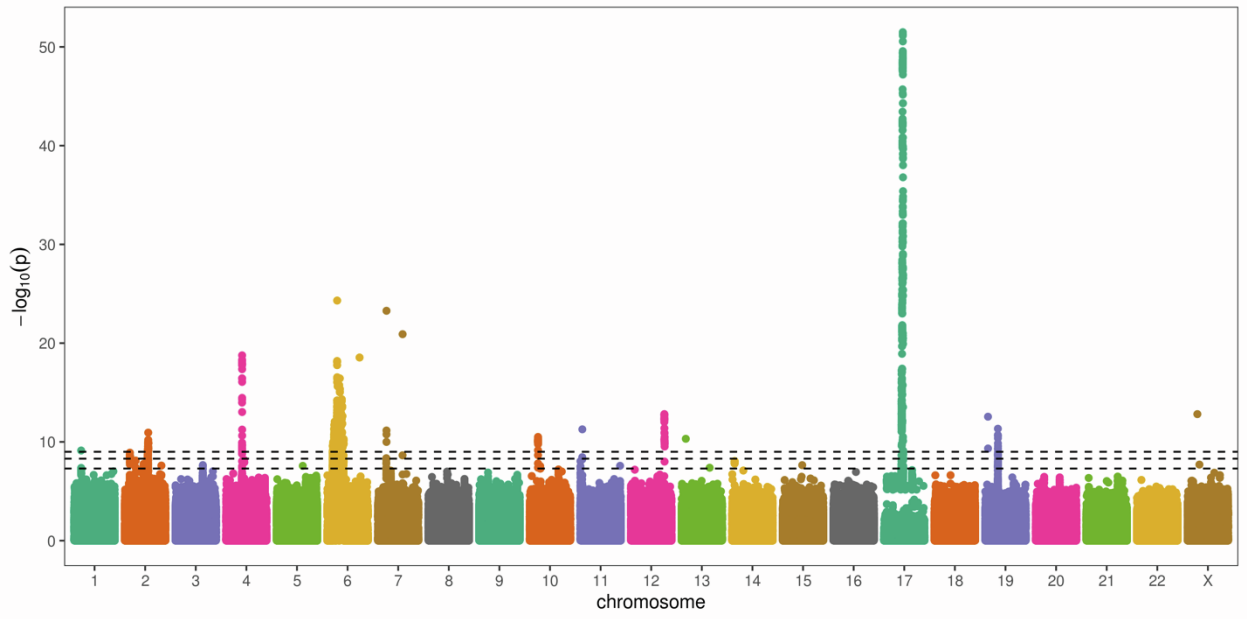


**Figure S13:** Manhattan plot of variants tested for association with the white blood cell count outcome.

QQplot of SNV results with WBC

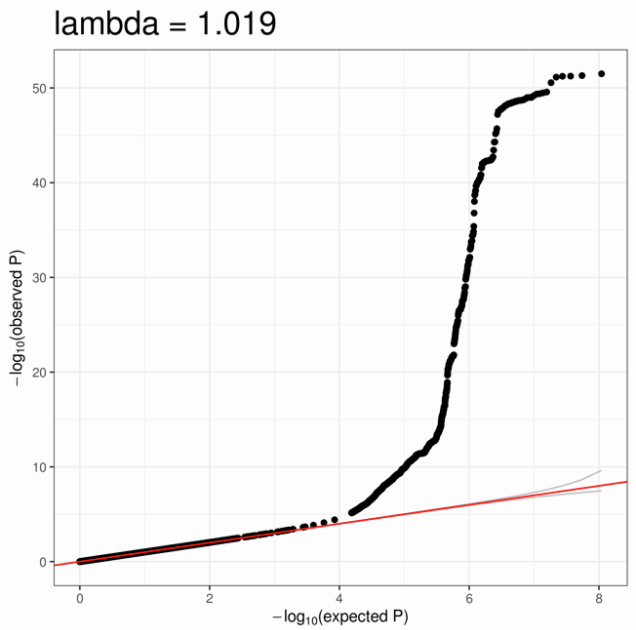


**Figure S14:** QQ plot of variants tested for association with the white blood cell count outcome.

## Supplementary Tables

See Excel file.

## Supplementary Methods

*Gene-based groupings for aggregate rare-variant tests*

We implemented a total of five strategies for filtering variants and aggregating them into gene-based groupings. The first three groupings included coding variants and the last two groupings included coding and noncoding variants. All coding variant groupings included high-confidence loss of function variants, protein-altering variants with Fathmm-XF score > 0.5, and synonymous variants with Fathmm-XF score > 0.5. In addition, the corresponding groupings included variants that satisfied the following criteria:

1) missense variants with MetaSVM_score>0,

2) missense variants which are predicted deleterious by ALL of these prediction approaches – SIFT4G, Polyphen2_HDIV,Polyphen2_HVAR, and LRT,

3) missense variants if they are predicted deleterious by ANY of SIFT4G, Polyphen2_HDIV, Polyphen2_HVAR, or LRT_pred.

Finally, the last two groupings included both coding and noncoding variants. These groupings were comprised of variants that satisfied the first grouping criteria and additionally:

4) variants overlapping with enhancer(s) linked to a gene using GeneHancer or overlapping with promoter(s) linked using GeneHancer or 5kb upstream region of the transcription start site, and which have Fathmm-XF score > 0.5 OR overlap with regions defined as “CTCF binding sites” or “Transcription factor binding sites” by Ensemble regulatory build annotation;
5) variants overlapping with enhancer(s) linked to a gene using GeneHancer which have Fathmm-XF score > 0.5 AND overlap with regions defined as “Promoters”, “Promoter flanking regions”, “Enhancers”, “CTCF binding sites”, “Transcription factor binding sites” or “Open chromatin regions”, specified by Ensemble regulatory build annotation; variants overlapping with promoter(s) either linked using GeneHancer or 5kb upstream region of the transcription start site, and which have Fathmm-XF score > 0.5 AND overlap with regions defined as “Promoters”, “Promoter flanking regions”, “Enhancers”, “CTCF binding sites”, “Transcription factor binding sites” or “Open chromatin regions”, specified by Ensemble regulatory build annotation.

*Genetic Ancestry and Relatedness*

Principal components (PCs) of genetic ancestry and pairwise relatedness measures were estimated for all 140,062 samples included in the TOPMed ‘freeze 8’ genotype release. Autosomal genetic variants passing the quality filter with a MAF > 0.01 and missing call rate < 0.01 were LD-pruned with an r2 threshold of 0.1 to obtain a set of 638,486 effectively independent variants for genetic ancestry and relatedness estimation. PC-AiR was used to obtain ancestry informative PCs robust to familial relatedness; the first 11 PCs showed evidence

of population structure. PC-Relate was then used to estimate pairwise kinship coefficients (KCs) for all pairs of samples, conditional on the genetic ancestry captured by PC-AiR PCs 1-11; these KC estimates reflect only recent genetic relatedness, e.g. due to pedigree structure. The PC-Relate KC estimates were used to construct a 4th degree sparse, block-diagonal, empirical kinship matrix (KM) for association testing, any pair of samples with estimated KC > 2(-11/2) ~ 0.022 were clustered in the same block; all KC estimates within a block of samples were kept, regardless of value; and all KC estimates between blocks were set to 0. By using a sparse block-diagonal KM, the association tests are more computationally efficient yet recent genetic relatedness is still accounted for. We subset the freeze-wide PCs and sparse KM to the appropriate set of participants for each analysis.

*Race imputation using HARE*

Ancestry groups were based on a combination of participants reported race/ethnicity and genetic ancestry represented by PCs from PC-AiR. To infer race/population group membership for participants with missing values, we used the HARE method, a machine learning algorithm that uses a support vector machine (SVM) to determine stratum assignment, taking as input genetically estimated PC values and reported race/ethnicity for each participant. Strata are defined by the unique reported race/ethnicity values provided, then the HARE SVM uses the input (training) data to learn the probability of stratum membership across the entire PC space. The output of HARE consists of multinomial probability vectors of stratum membership for each participant. HARE was run on a subset of samples included in the TOPMed freeze 8 genotype release; specifically, samples for participants from non-US-based studies and the Amish participants (because they were very distinct in PC space) were excluded from the HARE analysis. HARE was run using the first 9 PC-AiR PCs generated on this subset of samples to represent genetic ancestry with the following reported race/population groups: Asian, Black, Central American, Cuban, Dominican, Mexican, Puerto Rican, South American, and White. The genetic data from the 31,918 participants with either unreported or non-specific (e.g. ‘Multiple’ or ‘Other’) race and population membership was included in the HARE analysis, but they were not used to train the SVM. These participants were assigned to a population stratum based on their highest HARE output probability of membership. All other participants remained in the population stratum corresponding to their reported race/population group. Amish participants were assigned to their own stratum.

*TOPMed participating studies*

*Amish*
The Amish Complex Disease Research Program includes a set of large community-based studies focused largely on cardiometabolic health carried out in the Old Order Amish (OOA) community of Lancaster, Pennsylvania (http://medschool.umaryland.edu/endocrinology/amish/research-program.asp). The OOA population of Lancaster County, PA immigrated to the Colonies from Western Europe in the early 1700’s. There are now over 30,000 OOA individuals in the Lancaster area, nearly all of whom can trace their ancestry back 12-14 generations to approximately 700 founders. Investigators at the University of Maryland School of Medicine have been studying the genetic determinants of cardiometabolic health in this population since 1993. To date, over 7,000 Amish adults have participated in one or more of our studies.

Due to their ancestral history, the OOA are enriched for rare exonic variants that arose in the population from a single founder (or small number of founders) and propagated through genetic drift. Many of these variants have large effect sizes and identifying them can lead to new biological insights about health and disease. The parent study for this WGS project provides one (of multiple) examples. In our parent study, we identified through a genome-wide association analysis a haplotype that was highly enriched in the OOA that is associated with very high LDL-cholesterol levels. At the present time, the identity of the causative SNP – and even the implicated gene – is not known because the associated haplotype contains numerous genes, none of which are obvious lipid candidate genes. A major goal of the WGS that will be obtained through the NHLBI TOPMed Consortium will be to identify functional variants that underlie some of the large effect associations observed in this unique population.

*ARIC*
The ARIC study is a population-based cohort study consisting of 15,792 men and women that were drawn from four U.S. communities (Suburban Minneapolis, Minnesota; Washington County, Maryland; Forsyth County, North Carolina, and Jackson, Mississippi) 1. It was designed to investigate the causes of atherosclerosis and its clinical outcomes, and variation in cardiovascular risk factors, medical care, and disease by race, sex, location, and date. Participants were between age 45 and 64 years at their baseline examination in 1987-1989 when blood was drawn for DNA extraction and participants consented to genetic testing.

*BioMe*
The Charles Bronfman Institute for Personalized Medicine at Mount Sinai Medical Center (MSMC), BioMe Biobank, founded in September 2007, is an ongoing, broadly-consented electronic health record-linked clinical care biobank that enrolls participants non-selectively from the Mount Sinai Medical Center patient population. The MSMC serves diverse local communities of upper Manhattan, including Central Harlem (86% African American), East Harlem (88% Hispanic/Latino), and Upper East Side (88% Caucasian/White) with broad health disparities.

*CARDIA*
The Coronary Artery Risk Development in Young Adults (CARDIA) Study is a study examining the development and determinants of clinical and subclinical cardiovascular disease and their risk factors. It began in 1985-1986 with a group of 5,115 black and white men and women aged 18-30 years. The participants were selected so that there would be approximately the same number of people in subgroups of race, gender, education (high school or less and more than high school) and age (18-24 and 25-30) in each of 4 centers: Birmingham, AL; Chicago, IL; Minneapolis, MN; and Oakland, CA.

*CHS*
The Cardiovascular Health Study (CHS) is a population-based cohort study of risk factors for coronary heart disease and stroke in adults 65 years and older conducted across four field centers 2. The original predominantly European ancestry cohort of 5,201 persons was recruited in 1989-1990 from random samples of people on Medicare eligibility lists from four US communities. Subsequently, an additional predominantly African-American cohort of 687 persons was enrolled for a total sample of 5,888. Institutional review committees at each field center approved the CHS, and participants gave informed consent. Blood samples were drawn from all participants at their baseline examination, and DNA was subsequently extracted from available samples. These analyses were limited to participants with available DNA who also consented to genetic studies. Participants were examined annually from enrollment to 1999 and continued to be under surveillance for stroke following 1999.

*COPDGene*
COPDGene (also known as the Genetic Epidemiology of COPD Study) is an NIH-funded, multicenter study. A study population of more than 10,000 smokers (1/3 African American and 2/3 non-Hispanic White) has been characterized with a study protocol including pulmonary function tests, chest CT scans, six minute walk testing, and multiple questionnaires. Five years after this initial visit, all available study participants are being brought back for a follow-up visit with a similar study protocol. This study has been used for epidemiologic and genetic studies. Previous genetic analysis in this study has been based on genome-wide SNP genotyping data. Approximately 1,900 subjects underwent whole genome sequencing in this NHLBI WGS project, including severe COPD subjects and resistant smoking controls. The COPDGene Study web site is: http://www.copdgene.org/.

*FHS*
FHS is a three-generation, single-site, community-based, ongoing cohort study that was initiated in 1948 to investigate prospectively the risk factors for CVD including stroke. It now comprises 3 generations of participants: the Original cohort followed since 1948 3; their Offspring and spouses of the Offspring, followed since 1971 4; and children from the largest Offspring families enrolled in 2002 (Gen 3) 5. The Original cohort enrolled 5,209 men and women who comprised two-thirds of the adult population then residing in Framingham, MA. Survivors continue to receive biennial examinations. The Offspring cohort comprises 5,124 persons (including 3,514 biological offspring) who have been examined approximately once every 4 years. The Gen 3 cohort contains 4,095 participants.

*GeneSTAR*
In 1982 The Johns Hopkins Sibling and Family Heart Study was created to study patterns of coronary heart disease and related risk factors in families with early-onset coronary disease, identified from10 Baltimore area hospitals. Renamed in 2003, the Genetic Study of Atherosclerosis Risk (GeneSTAR) continues to study mechanisms of coronary heart disease and stroke in families using novel models and exciting new methods. GeneSTAR is a family-based study including initially healthy brothers and sisters identified from probands with early-onset coronary disease, along with the healthy offspring of the siblings and the probands. The goal is to discover and amplify mechanisms of stroke and coronary heart disease. Our African American and European American family cohort has undergone extensive screening, genetic testing, and follow-up for new cardiovascular disease, stroke, and other clinical events for 5 to 38 years.

*HCHS/SOL*
The Hispanic Community Health Study/Study of Latinos (HCHS/SOL) is a multi-center study of Hispanic/Latino populations with the goal of determining the role of acculturation in the prevalence and development of diseases, and to identify other traits that impact Hispanic/Latino health 6 . The study is sponsored by the National Heart, Lung, and Blood Institute (NHLBI) and other institutes, centers, and offices of the National Institutes of Health (NIH). Recruitment began in 2006 with a target population of 16,000 persons of Cuban, Puerto Rican, Dominican, Mexican or Central/South American origin. Participants were recruited through four sites affiliated with San Diego State University, Northwestern University in Chicago, Albert Einstein College of Medicine in Bronx, New York, and the University of Miami. Recruitment was implemented through a two-stage area household probability design 6. The study enrolled 16,415 participants who were self-identified Hispanic/Latino and aged 18-74 years and the extensive psycho-social and clinical assessments were conducted during 2008-2011. Annual

telephone follow-up interviews are ongoing since study inception. During the 2014-2017 second visit, the participants were re-examined again of various health outcomes of interest.

*JHS*

The Jackson Heart Study (JHS, https://www.jacksonheartstudy.org/jhsinfo/) is a large, community-based, observational study whose participants were recruited from urban and rural areas of the three counties (Hinds, Madison and Rankin) that make up the Jackson, MS metropolitan statistical area (MSA). Participants were enrolled from each of 4 recruitment pools: random, 17%; volunteer, 30%; currently enrolled in the Atherosclerosis Risk in Communities (ARIC) Study, 31% and secondary family members, 22%. Recruitment was limited to non-institutionalized adult African Americans 35-84 years old, except in a nested family cohort where those 21 to 34 years of age were also eligible. The final cohort of 5,301 participants included 6.59% of all African American Jackson MSA residents aged 35-84 during the baseline exam (N-76,426, US Census 2000). Among these, approximately 3,700 gave consent that allows genetic research and deposition of data into dbGaP. Major components of three clinic examinations (Exam 1 – 2000-2004; Exam 2 – 2005-2008; Exam 3 – 2009-2013) include medical history, physical examination, blood/urine analytes and interview questions on areas such as: physical activity; stress, coping and spirituality; racism and discrimination; socioeconomic position; and access to health care. Extensive clinical phenotyping includes anthropometrics, electrocardiography, carotid ultrasound, ankle-brachial blood pressure index, echocardiography, CT chest and abdomen for coronary and aortic calcification, liver fat, and subcutaneous and visceral fat measurement, and cardiac MRI. At 12-month intervals after the baseline clinic visit (Exam 1), participants have been contacted by telephone to: update information; confirm vital statistics; document interim medical events, hospitalizations, and functional status; and obtain additional sociocultural information. Questions about medical events, symptoms of cardiovascular disease and functional status are repeated annually. Ongoing cohort surveillance includes abstraction of medical records and death certificates for relevant International Classification of Diseases (ICD) codes and adjudication of nonfatal events and deaths. CMS data are currently being incorporated into the dataset.

*MESA*

The MESA study is a study of the characteristics of subclinical cardiovascular disease (disease detected non-invasively before it has produced clinical signs and symptoms) and the risk factors that predict progression to clinically overt cardiovascular disease or progression of the subclinical disease 7. MESA researchers study a diverse, population-based sample of 6,814 asymptomatic men and women aged 45-84. Thirty-eight percent of the recruited participants are white, 28 percent African-American, 22 percent Hispanic, and 12 percent Asian, predominantly of Chinese descent. Participants were recruited from six field centers across the United States: Wake Forest University, Columbia University, Johns Hopkins University, University of Minnesota, Northwestern University and the University of California - Los Angeles.

*SAFS*

The San Antonio Family Study (SAFS) is a complex pedigree-based mixed longitudinal study designed to identify low frequency or rare variants influencing susceptibility to cardiovascular disease, using WGS information from 2,590 individuals in large Mexican American pedigrees from San Antonio, Texas. The major objectives of this study are to identify low frequency or rare variants in and around known common variant signals for CVD, as well as to find novel low frequency or rare variants influencing susceptibility to CVD.

*WHI*

The Women's Health Initiative (WHI) is a long-term, prospective, multi-center cohort study that investigates post-menopausal women's health 8. WHI was funded by the National Institutes of Health and the National Heart, Lung, and Blood Institute to study strategies to prevent heart disease, breast cancer, colon cancer, and osteoporotic fractures in women 50-79 years of age. WHI involves 161,808 women recruited between 1993 and 1998 at 40 centers across the US. The study consists of two parts: the WHI Clinical Trial which was a randomized clinical trial of hormone therapy, dietary modification, and calcium/Vitamin D supplementation, and the WHI Observational Study, which focused on many of the inequities in women's health research and provided practical information about the incidence, risk factors, and interventions related to heart disease, cancer, and osteoporotic fractures.

*Replication and pheWAS samples and Methods*

Among white blood cell trait association analyses available for replication of our genome-wide discovery variants, all data sets included autosomes only except INTERVAL only analyses which included autosomes as well as chromosome X.

*INTERVAL Whole Exome Sequencing and Whole Genome Sequencing*

For individual variant replication lookups, we used 11,822 samples from the INTERVAL study with whole genome sequencing data. For aggregate rare-variant replication lookups, we used 4,006 unrelated samples with whole exome sequencing data. The INTERVAL study was conducted in England and began recruiting in 2011 among healthy blood donors to the National Health Service Blood and Transplant team to examine whether intervals of blood donation should be tailored by age, gender, genetic profile, and other characteristics. Phenotype values were adjusted to account for the influence of environmental and technical factors. Technical variables included seasonal effects, time dependent drift of equipment, sample decay, centre of sample collection, systematic differences in equipment, and systematic changes resulting from calibration of equipment. Adjustment is also made for participant environmental variables such as participant sex, age, and lifestyle factors including smoking, alcohol consumption, and diet. Quantile inverse normalisation within groups of haematology analyser and menopausal status was carried out as post-adjustment transformation. More details can be found in Astle et al. Cell 2016.

Rare variant associations were obtained using the SKAT test [1] using the "Wu" weights and considering missense and loss-of-function variants (as annotated by Gencode 31, VEP, and LOFTEE) with a minor allele frequency < 1%.

*UKBiobank African ancestry*

UK Biobank recruited 500,000 people aged between 40–69 years in 2006–2010, establishing a prospective biobank study to understand the risk factors for common diseases such as cancer, heart disease, stroke, diabetes, and dementia. Participants are being followed-up through routine medical and other health-related records from the UK National Health Service. UK Biobank has genotype data on all enrolled participants, as well as extensive baseline questionnaire and physical measures and stored blood and urine samples. Hematological traits were assayed as previously described [2]. Genotyping on custom Axiom arrays and subsequent quality control has been previously described [3]. Samples were included in our analysis if ancestry self-report was "Black Caribbean", "Black African", "Black or Black British", "White and Black Caribbean", "White and Black African", or "Any Other Black Background." Variants were

selected based on call rate exceeding 95%, HWE p-value less than $10^{-8}$, and MAF exceeding 0.5%. Subsequently, variants in approximate linkage equilibrium were used to generate principal components. Samples were excluded if the principal component exceeded 0.1 and the second principal component exceeded 0.2, to exclude individuals not clustering with most African ancestry individuals. In total, 6,567 participants with blood cell traits were included in the analysis.

*UKBiobank European ancestry*

A replication source from the combined UK Biobank (N=87,265) , UK Biobank Lung Exome Variant Evaluation (UK BiLEVE) (N=45,694), and INTERVAL (N=40,521) studies tested 29.5 million genetic variants for association with 36 red cell, white cell, and platelet properties in 173,480 European-ancestry participants [2]. At UK Biocenter, the UK Biobank whole blood samples were processed using four Beckman Coulter LH700 Series instruments while the INTERVAL samples were processed using two Sysmex XN-1000 instruments. Twenty indices of myeloid and lymphoid white blood cells were tested in genetic association analyses including counts and ratios. For replication purposes, we considered a direct blood trait matching to our genome-wide associated discovery trait considered as replication. Genotyping was completed using the Applied Biosystems UK Biobank Axiom Array and the Applied Biosystems UK BiLEVE Axiom Array by Affymetrix. Imputation was carried out using IMPUTE3 [4] software to the 1000Genomes Project (phase 3) reference panel and UK10K imputation panel (confirming accurate imputation of rare variants using whole exome sequencing data from overlapping individuals).

*UKBiobank European ancestry Whole Exome Sequencing Data*

UK Biobank exome sequencing data with ~200,000 patients was downloaded. Capture details, coverage, and alignment are extensively described elsewhere [5]. Downstream analysis excluded variants with low genotyping rate (<95%) and call rate <95%. Additionally, SNPs deviating from Hardy-Weinberg equilibrium ($P<1x10^{-7}$) were not kept. We annotated variants with the ENSEMBL Variant Effect Predictor (VEP web interface, Assembly: GRCh38.p13)[6]. We queried Ensembl/GENCODE and RefSeq transcripts databases and restricted results to produce the most severe consequence per variant. Variants annotated as missense, nonsense, essential splice site, and frameshift indel were kept for further analyses. We filtered the vcf files to retain the genomic regions corresponding to the following genes: *MARCKSL1, CNKSR2, TET2,* and *FLT3*. We further limited our analysis to rare variants (MAF < 1%). Monocyte count, neutrophil count, and lymphocyte counts were normalized as described in [7] and analyzed in a linear regression model adjusting for the first ten principal components (PCs). Thyroid disease was analyzed in a logistic regression model adjusting for the first ten PCs. To define cases and controls, we used ICD10 codes. If a participant had at least one of the following ICD10 codes, it was considered a case: E059, E063, E039. If a participant had the code E032, we excluded him/her from the analysis and kept all the other participants as controls. All analyses were performed in European ancestry individuals using RVtests (v.20171009) [8]. We carried out two gene-level association tests: a burden test, which aggregates counts of rare variants (GRANVIL), and SKAT, a bidirectional approach that includes SNPs with variable effect size and direction. Gene-level associations were conducted using rareMETALS_7.1 [9].

*WHI SNP Health Association Resource (SHARe) in African Americans*

Women's Health Initiative Study participants eligible for WHI-SHARe who had consented to genetic research included 12,157 women: 8,515 (70.1%) AA and 3,642 (66.6%) HA women.

Genotyping was performed on the Affymetrix 6.0 array with 2 μg of DNA at a concentration of 100 ng/μl. Imputation was carried out with MaCH [10]. After some more stringent filtering, 829,370 genotyped SNPs were used for imputation. For the imputation in AA samples, we used 240 HapMap 2 (release 22) phased haplotypes from the CEU (Utah residents with ancestry from northern and western Europe from the CEPH collection) and YRI (Yoruba in Ibadan, Nigeria) reference panels and were left with a total of 2,203,609 SNPs. To aid imputation accuracy, we estimated parameters on a subset of 200 WHI AA subjects and then imputed all WHI AA subjects. The final SHARe AA analytic samples for white blood cell association analyses ranged from 1949 (basophil count) to 7103 (white blood cell count) individuals.

*Genetic Epidemiology Research on Aging (GERA)*

The GERA cohort includes over 100,000 adults who are members of the Kaiser Permanente Medical Care Plan, Northern California Region (KPNC) and consented to research on the genetic and environmental factors that affect health and disease, linking together clinical data from electronic health records, survey data on demographic and behavioral factors, and environmental data with genetic data. The GERA cohort was formed by including all self-reported racial and ethnic minority participants with saliva samples (19%); the remaining participants were drawn sequentially and randomly from non-Hispanic White participants (81%). Genotyping was completed as previously described [11] using 4 different custom Affymetrix Axiom arrays with ethnic-specific content to increase genomic coverage. Principal components analysis was used to characterize genetic structure in this multi-ethnic sample, as previously described [12]. Blood cell traits were extracted from medical records. In individuals with multiple measurements, the first visit with complete white blood cell differential (if any) was used for each participant. Otherwise, the first visit was used. In total, 43,475 non-Hispanic white, 4,575 Hispanic/Latino and 1,809 African American participants with blood cell traits were included in the analysis.

*ADRN*

To define genetic risk factors of atopic dermatitis (AD) whole genome sequencing (WGS) has been performed on 777 subjects from the National Institute of Allergy and Infectious Diseases/Atopic Dermatitis Research Network (ADRN) as previously described [13]. This includes 237 non-atopic controls, 491 atopic dermatitis cases without eczema herpeticum and 49 atopic dermatitis cases with eczema herpeticum. In total,491 AD cases to 237 non-atopic controls were used in the analysis.

*BAGS*

The Barbados Asthma Genetics Study is a family-based genetic study focused on asthma. Pediatric probands with asthma were initially recruited through local clinics, followed by recruitment of parents and other family members, and expansion to independent asthma cases and controls. [14,15]. Whole genome sequencing is available on N=869 subjects with asthma status (410 asthmatics and 459 non-asthmatic controls) through TOPMed.

*SARP*

The overall goal of the Severe Asthma Research Program (SARP) is to identify and characterize subjects with severe asthma to understand pathophysiologic mechanisms in severe asthma. Subjects with mild and moderate asthma were recruited for comparison but the program was enriched for subjects with severe asthma from multiple centers. Subjects were

comprehensively phenotyped for asthma related traits including lung function, atopy, questionnaires on medical and family history, exhaled nitric oxide and health care utilization including exacerbations and symptoms [16]. In total, 218 EAs and 393 AAs were included in the analysis.

INTERVAL:
Participants in the INTERVAL randomised controlled trial were recruited with the active collaboration of NHS Blood and Transplant England (www.nhsbt.nhs.uk), which has supported field work and other elements of the trial. DNA extraction and genotyping were co-funded by the National Institute for Health Research (NIHR), the NIHR BioResource (http://bioresource.nihr.ac.uk) and the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). Sequencing was funded by Wellcome Trust grant number 206194. The academic coordinating centre for INTERVAL was supported by core funding from the: NIHR Blood and Transplant Research Unit in Donor Health and Genomics (NIHR BTRU-2014-10024), UK Medical Research Council (MR/L003120/1), British Heart Foundation (SP/09/002; RG/13/13/30194; RG/18/13/33946) and NIHR Cambridge BRC (BRC-1215-20014). A complete list of the investigators and contributors to the INTERVAL trial is provided in reference [17]. The academic coordinating centre would like to thank blood donor centre staff and blood donors for participating in the INTERVAL trial.

This work was supported by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

**Acknowledgements**

| TOPMed Accession # | TOPMed Project | Parent Study Name | TOPMed Phase | Omics Center | Omics Support |
|---|---|---|---|---|---|
| phs000956 | Amish | Amish | 1 | Broad Genomics | 3R01HL121007-01S1 |
| phs001211 | AFGen | ARIC AFGen | 1 | Broad Genomics | 3R01HL092577-06S1 |
| phs001211 | VTE | ARIC | 2 | Baylor | 3U54HG003273-12S2 / HHSN268201500015C |
| phs001644 | AFGen | BioMe AFGen | 2.4 | MGI | 3UM1HG008853-01S2 |
| phs001143 | BAGS | BAGS | 1 | Illumina | 3R01HL104608-04S1 |
| phs001644 | BioMe | BioMe | 3 | Baylor | HHSN268201600033I |
| phs001644 | BioMe | BioMe | 3 | MGI | HHSN268201600037I |
| phs001612 | CARDIA | CARDIA | 3 | Baylor | HHSN268201600033I |
| phs001368 | CHS | CHS | 3 | Baylor | HHSN268201600033I |
| phs001368 | VTE | CHS VTE | 2 | Baylor | 3U54HG003273-12S2 / HHSN268201500015C |

| | | | | | |
|---|---|---|---|---|---|
| phs000951 | COPD | COPDGene | 1 | NWGC | 3R01HL089856-08S1 |
| phs000951 | COPD | COPDGene | 2 | Broad Genomics | HHSN268201500014C |
| phs000951 | COPD | COPDGene | 2.5 | Broad Genomics | HHSN268201500014C |
| phs000974 | AFGen | FHS AFGen | 1 | Broad Genomics | 3R01HL092577-06S1 |
| phs000974 | FHS | FHS | 1 | Broad Genomics | 3U54HG003067-12S2 |
| phs001218 | AA_CAC | GeneSTAR AA_CAC | 2 | Broad Genomics | HHSN268201500014C |
| phs001218 | GeneSTAR | GeneSTAR | legacy | Illumina | R01HL112064 |
| phs001218 | GeneSTAR | GeneSTAR | 2 | Psomagen | 3R01HL112064-04S1 |
| phs001395 | HCHS_SOL | HCHS_SOL | 3 | Baylor | HHSN268201600033I |
| phs000964 | JHS | JHS | 1 | NWGC | HHSN268201100037C |
| phs001416 | AA_CAC | MESA AA_CAC | 2 | Broad Genomics | HHSN268201500014C |
| phs001416 | MESA | MESA | 2 | Broad Genomics | 3U54HG003067-13S1 |
| phs001215 | SAFS | SAFS | 1 | Illumina | 3R01HL113323-03S1 |
| phs001215 | SAFS | SAFS | legacy | Illumina | R01HL113322 |
| phs001446 | SARP | 2 | SARP | NYGC Genomics | HHSN268201500016C |
| phs001237 | WHI | WHI | 2 | Broad Genomics | HHSN268201500014C |

Amish: The TOPMed component of the Amish Research Program was supported by NIH grants R01 HL121007, U01 HL072515, and R01 AG18728. Email Rhea Cosentino (rcosenti@som.umaryland.edu) for additional input.

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.

BAGS: We gratefully acknowledge the contributions of Pissamai and Trevor Maul, Paul Levett, Anselm Hennis, P. Michele Lashley, Raana Naidu, Malcolm Howitt and Timothy Roach, and 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, and HL104608 S1.

BioMe: The Mount Sinai BioMe Biobank has been supported by The Andrea and Charles Bronfman Philanthropies and in part by Federal funds from the NHLBI and NHGRI (U01HG00638001; U01HG007417; X01HL134588). We thank all participants in the Mount Sinai Biobank. We also thank all our recruiters who have assisted and continue to assist in data

collection and management and are grateful for the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

BioVU: The BioVU projects at Vanderbilt University Medical Center are supported by numerous sources: institutional funding, private agencies, and federal grants. These include the NIH funded Shared Instrumentation Grant S10OD017985 and S10RR025141; CTSA grants UL1TR002243, UL1TR000445, and UL1RR024975 from the National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent official views of the National Center for Advancing Translational Sciences or the National Institutes of Health. Genomic data are also supported by investigator-led projects that include U01HG004798, R01NS032830, RC2GM092618, P50GM115305, U01HG006378, U19HL065962, R01HD074711; and additional funding sources listed at https://victr.vumc.org/biovu-funding/.

CARDIA: 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 University of Alabama at Birmingham (HHSN268201800005I & HHSN268201800007I), Northwestern University (HHSN268201800003I), University of Minnesota (HHSN268201800006I), and Kaiser Foundation Research Institute (HHSN268201800004I). CARDIA was also partially supported by the Intramural Research Program of the National Institute on Aging (NIA) and an intra-agency agreement between NIA and NHLBI (AG0005).

CHS: Cardiovascular Health Study: This research was supported by contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the National Heart, Lung, and Blood Institute (NHLBI), with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by 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 content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

COPDGene: The COPDGene project described was supported by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health. 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

FHS: The Framingham Heart Study (FHS) acknowledges the support of contracts NO1-HC-25195 and HHSN268201500001I 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.

GeneSTAR: GeneSTAR was supported by the National Institutes of Health/National Heart, Lung, and Blood Institute (U01 HL72518, HL087698, HL112064, HL11006, HL118356) 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. We would like to thank our participants and staff for their valuable contributions.

HCHS/SOL: 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.

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.

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 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 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. MESA Family is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support is provided by grants and contracts R01HL071051, R01HL071205, R01HL071250, R01HL071251, R01HL071258, R01HL071259, by the National Center for Research Resources, Grant UL1RR033176. 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.

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.

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 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.

**References**


1 Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89(1):82-93.

2 Astle WJ, Elding H, Jiang T, et al. The allelic landscape of human blood cell trait variation and links to common complex disease. Cell. 2016;167(5):1415-1429.e19.

3 Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203-209.

4 Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3 (Bethesda). 2011;1(6):457-470.

5 Jurgens, S.J., et al., Rare Genetic Variation Underlying Human Diseases and Traits: Results from 200,000 Individuals in the UK Biobank. bioRxiv, 2020: p. 2020.11.29.402495

6 McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the ensembl api and snp effect predictor. Bioinformatics. 2010;26(16):2069-2070.

7 Mousas A, Ntritsos G, Chen M-H, et al. Rare coding variants pinpoint genes that control human hematological traits. PLoS Genet. 2017;13(8):e1006925.

8 Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics. 2016;32(9):1423-1426.

9 Liu DJ, Peloso GM, Zhan X, et al. Meta-analysis of gene-level tests for rare variant association. Nat Genet. 2014;46(2):200-204.

10 Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34(8):816-834.

11 Kvale MN, Hesselson S, Hoffmann TJ, et al. Genotyping informatics and quality control for 100,000 subjects in the genetic epidemiology research on adult health and aging (Gera) cohort. Genetics. 2015;200(4):1051-1060.

12 Banda Y, Kvale MN, Hoffmann TJ, et al. Characterizing race/ethnicity and genetic ancestry for 100,000 subjects in the genetic epidemiology research on adult health and aging (Gera) cohort. Genetics. 2015;200(4):1285-1295.

13 Bin L, Malley C, Taylor P, et al. Whole genome sequencing identifies novel genetic mutations in patients with eczema herpeticum. Allergy. Published online February 6, 2021.

14 Mathias RA, Grant AV, Rafaels N, et al. A genome-wide association study on African-ancestry populations for asthma. J Allergy Clin Immunol. 2010;125(2):336-346.e4.

15 Barnes KC, Neely JD, Duffy DL, et al. Linkage of asthma and total serum IgE concentration to markers on chromosome 12q: evidence from Afro-Caribbean and Caucasian populations. Genomics. 1996;37(1):41-50.

16 Jarjour NN, Erzurum SC, Bleecker ER, et al. Severe asthma: lessons learned from the national heart, lung, and blood institute severe asthma research program. Am J Respir Crit Care Med. 2012;185(4):356-362.

17. Di Angelantonio E, Thompson SG, Kaptoge SK, Moore C, Walker M, Armitage J, Ouwehand WH, Roberts DJ, Danesh J, INTERVAL Trial Group. Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. Lancet. 2017 Nov 25;390(10110):2360-2371.