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

Anna V. Mikhaylova,[1,69] Caitlin P. McHugh,[1,69] Linda M. Polfus,[2,69] Laura M. Raffield,[3] Meher Preethi Boorgula,[4] Thomas W. Blackwell,[5] Jennifer A. Brody,[6] Jai Broome,[1] Nathalie Chami,[7] Ming-Huei Chen,[8,9] Matthew P. Conomos,[1] Corey Cox,[4] Joanne E. Curran,[10] Michelle Daya,[4] Lynette Ekunwe,[11] David C. Glahn,[12] Nancy Heard-Costa,[9,13] Heather M. Highland,[14] Brian D. Hobbs,[15,16] Yann Ilboudo,[17,18] Deepti Jain,[1] Leslie A. Lange,[4] Tyne W. Miller-Fleming,[19] Nancy Min,[11] Jee-Young Moon,[20] Michael H. Preuss,[7] Jonathon Rosen,[21] Kathleen Ryan,[22] Albert V. Smith,[5] Quan Sun,[21] Praveen Surendran,[23,24,25,26] Paul S. de Vries,[27] Klaudia Walter,[28] Zhe Wang,[7] Marsha Wheeler,[29] Lisa R. Yanek,[30] Xue Zhong,[19] Goncalo R. Abecasis,[5] Laura Almasy,[31,32] Kathleen C. Barnes,[4] Terri H. Beaty,[33] Lewis C. Becker,[34] John Blangero,[10] Eric Boerwinkle,[27] Adam S. Butterworth,[23,24,25,35,36] Sameer Chavan,[4] Michael H. Cho,[15] Hélène Choquet,[37] Adolfo Correa,[11] Nancy Cox,[19] Dawn L. DeMeo,[15,16] Nauder Faraday,[38] Myriam Fornage,[39] Robert E. Gerszten,[40,41] Lifang Hou,[42] Andrew D. Johnson,[8,9] Eric Jorgenson,[43] Robert Kaplan,[20]

(Author list continued on next page)

## Summary

Many common and rare variants associated with hematologic traits have been discovered through imputation on large-scale reference panels. However, the majority of genome-wide association studies (GWASs) have been conducted in Europeans, and determining causal variants has proved challenging. We performed a GWAS of total leukocyte, neutrophil, lymphocyte, monocyte, eosinophil, and basophil counts generated from 109,563,748 variants in the autosomes and the X chromosome in the Trans-Omics for Precision Medicine (TOPMed) program, which included data from 61,802 individuals of diverse ancestry. We discovered and replicated 7 leukocyte trait associations, including (1) the association between a chromosome X, pseudo-autosomal region (PAR), noncoding variant located between cytokine receptor genes (*CSF2RA* and *CLRF2)* and lower eosinophil count; and (2) associations between single variants found predominantly among African Americans at the *S1PR3* (9q22.1) and *HBB* (11p15.4) loci and monocyte and lymphocyte counts, respectively. We further provide evidence indicating that the newly discovered eosinophil-lowering chromosome X PAR variant might be associated with reduced susceptibility to common allergic diseases such as atopic dermatitis and asthma. Additionally, we found a burden of very rare *FLT3* (13q12.2) variants associated with monocyte counts. Together, these results emphasize the utility of whole-genome sequencing in diverse samples in identifying associations missed by European-ancestry-driven GWASs.

## Introduction

Counts of circulating white blood cells (WBCs) are important clinical parameters that are used for monitoring general disease activity and tolerance to therapies for oncological and rheumatologic diseases. WBCs are derived from hematopoietic stem cells and during differentiation are committed into two distinct lineages: myeloid (neutrophils, basophils, eosinophils, and monocytes) and lymphoid (lymphocytes). By studying the genetic determinants of

[1]Department of Biostatistics, University of Washington, Seattle, WA 98105, USA; [2]Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA; [3]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [4]Division of Biomedical Informatics and Personalized Medicine, School of Medicine, University of Colorado, Anschutz Medical Campus, Aurora, CO 80045, USA; [5]TOPMed Informatics Research Center, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA; [6]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA 98105, USA; [7]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York City, NY 10029, USA; [8]Population Sciences Branch, Division of Intramural Research, National Heart, Lung and Blood Institute, Bethesda, MD 20892, USA; [9]National Heart, Lung and Blood Institute's and Boston University's Framingham Heart Study, Framingham, MA 01701, USA; [10]Department of Human Genetics and South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX 78539, USA; [11]Department of Medicine, University of Mississippi Medical Center, Jackson, MS 39216, USA; [12]Department of Psychiatry, Boston Children's Hospital and Harvard Medical School, Boston, MA 02155, USA; [13]Department of Neurology, Boston University School of Medicine, Boston, MA 02118, USA; [14]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [15]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA; [16]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA; [17]Montréal Heart Institute, Montréal, Québec H1T 1C8, Canada; [18]Faculté de Médecine, Université de Montréal, Montréal, Québec H1T 1C8, Canada; [19]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37240, USA; [20]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx,

(Affiliations continued on next page)

Charles Kooperberg,[44] Kousik Kundu,[28,45] Cecelia A. Laurie,[1] Guillaume Lettre,[17,18] Joshua P. Lewis,[22] Bingshan Li,[46] Yun Li,[47] Donald M. Lloyd-Jones,[48,49] Ruth J.F. Loos,[7] Ani Manichaikul,[50] Deborah A. Meyers,[51] Braxton D. Mitchell,[22,52] Alanna C. Morrison,[27] Debby Ngo,[41] Deborah A. Nickerson,[29] Suraj Nongmaithem,[28] Kari E. North,[14] Jeffrey R. O'Connell,[22] Victor E. Ortega,[53] Nathan Pankratz,[54] James A. Perry,[55] Bruce M. Psaty,[56,57,58] Stephen S. Rich,[50] Nicole Soranzo,[28,35,45,59] Jerome I. Rotter,[60] Edwin K. Silverman,[15,16] Nicholas L. Smith,[56,57,61] Hua Tang,[62] Russell P. Tracy,[63] Timothy A. Thornton,[1,43] Ramachandran S. Vasan,[9,64,65] Joe Zein,[66] Rasika A. Mathias,[67] NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Alexander P. Reiner,[56,*] and Paul L. Auer,[68,*]

WBC counts, we have been able to gain a more complete understanding of hematopoiesis and the complex roles of WBCs in both acute and chronic inflammation.[1,2]

Total and differential WBC counts are complex, polygenic, quantitative traits, and the genetic contribution to variance in WBC counts (heritability) is estimated at 50%–60%.[3] Numerous recent studies have characterized both common (minor-allele frequency [MAF] greater than 5%) and infrequent (MAF between 0.5% and 5%) variation contributing to WBC counts in European, African, East Asian, and Hispanic populations.[4–8] To date, most studies of the genetics of WBC counts have used a combination of study designs, including standard genome-wide genotyping arrays,[3] exome sequencing,[9] exome-chip genotyping,[4,10] and application of genome-wide imputation using reference panels.[5,6,8] An obvious gap in these study designs is a comprehensive, genome-wide interrogation of common and rare variation that could be missed by imputation-based approaches.

Whole-genome sequencing (WGS)-based analysis largely addresses these gaps, particularly in individuals of non-European origin. Importantly, WGS can assess population-specific variants,[11] including variants that are often poorly imputed with standard reference panels and genotyping arrays.[12] Here we utilized deep (∼30×) WGS data from 61,802 individuals, including African American (AA), East Asian (EAS), European American (EA), and Hispanic/Latino (HA) subjects. Data were generated as part of the National Heart, Lung and Blood Institute (NHLBI)-Trans-Omics for Precision Medicine (TOPMed) program investigating the genetics of WBC counts.

NY 10461, USA; [21]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [22]Department of Medicine, Division of Endocrinology, Diabetes, and Nutrition, University of Maryland School of Medicine, Baltimore, MD 21201, USA; [23]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK; [24]British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge CB1 8RN, UK; [25]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge CB1 8RN, UK; [26]Rutherford Fund Fellow, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK; [27]Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [28]Department of Human Genetics, Wellcome Sanger Institute, Hinxton, CB10 1SA, UK; [29]Department of Genome Sciences, University of Washington, Seattle, WA 98105, USA; [30]Division of General Internal Medicine, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [31]Department of Biomedical and Health Informatics, the Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA; [32]Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA; [33]School of Public Health, John Hopkins University, Baltimore, MD 21205, USA; [34]Division of Cardiology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [35]National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Genomics, University of Cambridge, Cambridge CB1 8RN, UK; [36]National Institute for Health Research Cambridge Biomedical Research Centre, University of Cambridge and Cambridge University Hospitals, Cambridge CB1 8RN, UK; [37]Division of Research, Kaiser Permanente Northern California, Oakland, CA 94601, USA; [38]Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [39]University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [40]Division of Cardiovascular Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA; [41]Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [42]Institute for Public Health and Medicine, Northwestern University, Chicago, IL 60661, USA; [43]Regeneron Genetics Center, Tarrytown, NY 10591, USA; [44]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; [45]Department of Haematology, University of Cambridge, Cambridge CB1 8RN, UK; [46]Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN 37232, USA; [47]Departments of Biostatistics, Genetics, and Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [48]Division of Cardiology, Department of Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL 60661, USA; [49]Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL 60661, USA; [50]Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia School of Medicine, Charlottesville, VA 22908, USA; [51]Division of Genetics, Genomics and Precision Medicine, Department of Medicine, University of Arizona, Tucson, AZ 85724, USA; [52]Geriatrics Research and Education Clinical Center, Baltimore Veterans Administration Medical Center, Baltimore, MD 21201, USA; [53]Department of Internal Medicine, Center for Precision Medicine, Wake Forest School of Medicine, Winston-Salem, NC 27101, USA; [54]Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, Minneapolis, MN 55455, USA; [55]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201, USA; [56]Department of Epidemiology, University of Washington, Seattle, WA 98105, USA; [57]Department of Health Service, University of Washington, Seattle, WA 98105, USA; [58]Department of Medicine, University of Washington, Seattle, WA 98105, USA; [59]British Heart Foundation Centre of Excellence, Division of Cardiovascular Medicine, Addenbrooke's Hospital, Cambridge CB1 8RN, UK; [60]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA 90502, USA; [61]Seattle Epidemiologic Research and Information Center, Department of Veterans Affairs Office of Research and Development, Seattle, WA 98105, USA; [62]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA; [63]Department of Pathology and Laboratory Medicine and Department of Biochemistry, University of Vermont Larner College of Medicine, Colchester, VT 05446, USA; [64]Departments of Cardiology and Preventive Medicine, Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA; [65]Department of Epidemiology, Boston University School of Public Health, Boston, MA 02118, USA; [66]Respiratory Institute, Cleveland Clinic, Cleveland, OH 44195, USA; [67]Division of Allergy and Clinical Immunology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [68]Zilber School of Public Health, University of Wisconsin, Milwaukee, Milwaukee, WI 53205, USA

[69]These authors contributed equally

*Correspondence: apreiner@uw.edu (A.P.R.), pauer@mcw.edu (P.L.A.)

https://doi.org/10.1016/j.ajhg.2021.08.007.

## Material and methods

### TOPMed samples

NHLBI's TOPMed program comprises several parent studies. The parent studies that contributed to our analyses included Atherosclerosis Risk in Communities (ARIC),[13] the Amish Complex Disease Research Program (Amish),[14] BioMe Biobank (BioMe),[15] Cardiovascular Artery Risk Development in Young Adults (CARDIA),[16] the Cardiovascular Health Study (CHS),[17] Genetic Epidemiology of COPD (COPDGene),[18] the Framingham Heart Study (FHS),[19] Genetic Study of Atherosclerosis Risk (GeneSTAR),[20] Hispanic Community Health Study/Study of Latinos (HCHS/SOL),[21] the Jackson Heart Study (JHS),[22,23] Multi-Ethnic Study of Atherosclerosis (MESA),[24] the San Antonio Family Heart Study (SAFS),[25] and the Women's Health Initiative (WHI).[26] Additional information about the design of each study and the sampling of individuals within each cohort for WGS is available in the supplemental information. Participants included in these analyses (unique $n = 61,865$) are shown in Table S1, stratified by study, ancestry group (see supplemental methods), and WBC trait. For these analyses, 1% of participants are Asian, 23% are Black, 22% are Hispanic/Latino, and 54% are white. All studies were approved by the appropriate institutional review boards (IRBs), and informed consent was obtained from all participants.

### TOPMed WGS and quality control

WGS was performed at an average depth of 38× by six sequencing centers (Broad Genomics, Northwest Genome Center, Illumina, New York Genome Center, Baylor, and McDonnell Genome Institute) with Illumina X10 technology and DNA from blood. Here we report analyses from the "Freeze 8" dataset, where reads were aligned to human-genome build GRCh38 through the use of a common pipeline across all sequencing centers. To perform variant quality control (QC) within the Freeze 8 dataset, we trained a support vector machine (SVM) classifier on known variant sites (positive labels) and Mendelian inconsistent variants (negative labels). Further variant filtering was done for variants with excess heterozygosity and Mendelian discordance. Sample QC measures included concordance between annotated and inferred genetic sex; concordance between prior array genotype data and TOPMed WGS data; and pedigree checks. Details regarding the genotype "freezes," laboratory methods, data processing, and QC are described on the TOPMed website and in a common document accompanying each study's dbGaP accession number.

### WBC phenotype measurements and exclusion criteria

White blood cells, basophils, eosinophils, neutrophils, lymphocytes, and monocytes were counted in a subset of the TOPMed freeze 8 samples (Table S1) via automated clinical hematology analyzers. Each of the phenotypes is defined as the concentration of cell type in the blood and is measured in billions/liter. Trait-specific QC excluded participants with WBC count values $> 100 \times 10^9$ cells/L ($n = 5$), neutrophil values $> 75 \times 10^9$ cells/L ($n = 1$), monocyte values $> 15 \times 10^9$ cells/L ($n = 1$), lymphocyte values $> 150 \times 10^9$ cells/L ($n = 1$), eosinophil values $> 20 \times 10^9$ cells/L ($n = 1$), and basophil values equal to $0.9 \times 10^9$ cells/L ($n = 1$). Additionally, in instances where multiple measurements were available, we kept only one measurement for each individual and each trait.

### Single-variant association tests for quantitative traits

We performed genome-wide single-variant association tests by using a two-step linear mixed model (LMM). In the first step, we fit the "null model" under the null hypothesis of no genetic association and did not include genetic variants in the model. We included sex, age, combined study by phase variable (e.g., WHI_2 refers to phase 2 of WHI study), and the first 11 PC-Air[27] principal components (PCs) of genetic ancestry as fixed effects. To account for genetic relatedness, we included a 4th-degree sparse empirical kinship matrix (KM) computed with PC-Relate.[28] In order to better control genomic inflation,[29] we allowed for heteroscedasticity in the error variances by modeling separate residual variance components, one for each study, by ancestry group (e.g., WHI_White). Details on estimating the ancestry group are in the supplemental methods.

In order to improve power and appropriately control type I error in settings with non-normal phenotype distribution, we used a fully adjusted two-stage approach for fitting the null model.[30] In stage 1, we fit an LMM with the observed phenotype values as the outcome, the fixed effects as covariates, a sparse KM, and heterogeneous residual variances. We applied a rank-based inverse-normal transformation to the residuals from the results of stage 1 and then rescaled them by the original variance. In stage 2, we fit another LMM by using the rescaled residuals obtained in stage 1 as the outcome and by using the same covariates, the same KM, and the same heterogeneous residual variance model as in stage 1. Finally, we used the output from stage 2 as the trait of interest to perform a score test of genetic association. In the association analyses, we included variants that had a minor-allele count (MAC) of at least 5, passed the TOPMed Informatics Research Center (IRC) quality filters, and had less than 10% of samples with a sequencing read depth of less than 10. A threshold level of $5 \times 10^{-8}$ was used to determine statistical significance.

### Single-variant association tests for basophil count as a binary trait

Instead of testing basophils as a continuous trait, we performed genome-wide single-variant association tests of basophils as a binary trait dichotomized at $0.05 \times 10^9$ cells/L (basophil[3] $\geq 0.05$ versus basophil $< 0.05$). We fit a generalized linear mixed model (GLMM) with binomial family and logit link via the penalized quasi-likelihood[31] approach of GMMAT[32] because our outcome was no longer quantitative. The same fixed-effect covariates and sparse KM as for quantitative-trait analysis were included. Because the variance model for a GLMM is specified by the binomial family and link function, we did not use heterogeneous residual variance groups or the two-stage rank-normalization procedure. We performed genome-wide association tests based on score statistics and saddlepoint approximation (SPA) of the p values.[33,34] The SPA method has been shown to better control type I error even when the ratio of affected to control individuals is unbalanced, e.g., during the testing of low-frequency and rare variants, when the number of carriers is much lower than the sample size.

### Conditional analyses

We performed conditional single-variant association tests where, in addition to adjusting for the fixed-effect covariates and sparse KM that were used in single-variant analyses, we adjusted for variants previously known to be associated with the outcomes (Table S3). First, we matched the known variants to TOPMed variants on the basis of position and alleles and selected the variants that passed the TOPMed IRC quality filters. We then used linkage disequilibrium (LD) (with a threshold of $R^2 > 0.8$) to prune the set of matched variants for each trait separately and checked for collinearity of the pruned variants with the covariates. The final set of variants was included in the first round of conditional analyses. After the first round, we checked whether the remaining significant variants were near (within 1 Mb window) the known variants that failed the TOPMed IRC filters. These variants, in addition to the set of variants from the first round, were included in the second round of conditional analyses.

## Gene-based aggregate rare-variant tests

To improve the power to detect rare-variant associations, we implemented several strategies of aggregating variants and testing for cumulative associations of gene-based groupings with the traits. We implemented a total of five strategies of variant groupings: three strategies included coding variants only, and two strategies included coding and noncoding variants from enhancer and promoter regions, but only those with "deleterious" consequences to the corresponding gene, "deleterious" being defined by various annotation-based filters; the details are provided in the supplemental methods. We performed aggregate tests by using the efficient variant-set mixed-model association test (SMMAT),[35] which is more computationally efficient than SKAT-O (optimized SKAT [sequence kernel association test]) and more powerful than burden tests or SKAT alone. The SMMAT test used the same null model that was fit for the single-variant analyses, and the p value was constructed from a combination of the mixed-model burden p value with an asymptotically independent adjusted SKAT-like p value via Fisher's method. In our analyses, we included non-monomorphic variants that had an MAF of less than 1% and that passed the same quality filters that were used for single-variant analyses. To upweight rarer variants, we used weights that are based on MAF and given by a beta distribution with parameters 1 and 25. We determined statistical significance by using a Bonferroni correction for the number of aggregate groups tested in each aggregation strategy.

## Analyses of WBC-subtype proportions

In addition to analyzing counts of WBC subtypes, we also analyzed WBC-subtype proportions for all of the replicated, statistically significant WBC-subtype count signals. To do so, we identified samples whose WBC count and corresponding WBC-subtype count were collected at the same visit and divided the WBC-subtype count by the total measured WBC count. We excluded samples where the proportion of WBC-subtype count to WBC count was greater than 1. This proportion was treated as the phenotype and modeled similarly to the other phenotypes, i.e., with the two-step LMM described above.

## Fine-mapping analyses

After conditional analyses, we carried out statistical fine mapping by using the following approach: because our conditional analyses implicated a single independent variant at each locus, we assumed a single causal variant at each locus. We then adapted the method proposed by Maller et al.[36] to assign posterior inclusion probabilities (PIPs) to each variant and construct 95% credible sets. In brief, we considered all variants within 250 kb upstream and 250 kb downstream of the sentinel SNP and converted summary statistics into approximate Bayes factors (aBFs) as follows:

$$aBF = \sqrt{\frac{SE^2}{SE^2 + \omega}}\exp\left[\frac{\omega\beta^2}{2SE^2\left(SE^2 + \omega\right)}\right]$$

where $\beta$ and SE are the variant's effect size and standard error, respectively, and $\omega$ represents the prior variance in allelic effects. As in Maller et al.,[36] we set $\omega = 0.04$. We then calculated the PIP of each variant by dividing the variant's aBF by the sum of the aBFs for all variants at the locus. We generated the 95% credible sets by ordering all variants (at a particular locus) from largest to smallest PIP and including variants until cumulative PIPs $\geq 0.95$.

## Haplotype analyses

On the basis of the results of conditional single variant analyses, we performed haplotype analyses for the hemoglobin beta (*HBB*) region (rs334 and rs33930165) in total WBC counts and lymphocytes and the *NRIP1* region (rs28574812 and rs2823002) in monocytes and total WBC counts. We constructed 2-SNP haplotypes from phased genotype data and identified haplotypes with non-zero frequencies. We counted the number of copies of each haplotype in each subject and included the number of copies of each non-reference haplotype as covariates in the model. The haplotype with the highest frequency was considered the reference haplotype. Using the null model from the single variant analyses, we performed association tests and report haplotype-specific results.

## PheWAS analysis

We extracted phenome-wide association scanning (pheWAS) results for the seven new replicated signals from the UKBiobank (UKBB) and the BioVU biobank. The UKBB results were obtained from the UKBB ICD PheWeb hosted at the University of Michigan on the basis of 408,961 samples from white British participants. We considered the 1,261 phecodes with at least 100 affected individuals and an associated Bonferroni corrected threshold for significance of $0.05/1,261 = 3.96 \times 10^{-5}$. BioVU is the Vanderbilt University Medical Center (VUMC) biobank that houses de-identified DNA samples linked to phenotypic data derived from the electronic health record (EHR) system of VUMC. The pheWAS lookups in BioVU were restricted to African Americans (n » 5,000). For the rs334 lookups, we had access to samples from $\sim$14,000 African Americans that were either heterozygous at rs334 or had two copies of the reference allele. Phenotypes were derived from billing codes of EHRs. Association between each binary phecode and a SNP was assessed using logistic regression, while adjusting for covariates of age, sex, genotyping array batch, and 10 principal components of ancestry. We considered the 726 phecodes with at least 100 cases with an associated Bonferroni corrected threshold for significance of $0.05/726 = 6.89 \times 10^{-5}$.

## UKBiobank analysis of asthma, COPD, and atopic dermatitis with rs28532112

We constructed phenotypes for chronic obstructive pulmonary disease (COPD), asthma, and atopic dermatitis (AD), as defined in Wu et al.,[37] by using ICD10 codes to select a case group and a control group. The initial selected set of affected individuals and controls was purged of relatedness by the removal of one member of each related pair in an iterative fashion until no related subjects remained. Using the remaining group of affected individuals and remaining pool of controls, we selected a fixed number of control individuals for each affected individual, matched by sex, age, and ancestry. The fixed number used for the ratio of control individuals to affected individuals was adjusted to yield a total n in the range of 40,000 to 80,000 subjects. Association analyses were conducted with the OASIS pipeline.

### TOPMed analysis of rs28532112 with asthma and asthma severity

TOPMed generated WGS on n = 869 subjects with asthma status (410 asthmatics and 459 non-asthmatic individuals) for the Barbados Asthma Genetics Study (BAGS)[38,39] and n = 611 asthmatic subjects from the Severe Asthma Research Program (SARP)[40] study. We used GENESIS to perform association tests for rs28532112 with asthma (in BAGS), and included age, batch, and sex as covariates. We also performed tests for association with asthma severity (in SARP), as measured by pre-forced expiratory volume1 (preFEV1), for rs28532112; we controlled for age, gender, and body mass index and stratified by ancestry (EA [n = 218] and AA [n = 393]).

### ADRN analysis of rs28532112

To define genetic risk factors of AD, we performed WGS on 777 subjects from the Atopic Dermatitis Research Network (ADRN) of the National Institute of Allergy and Infectious Diseases as previously described.[41] This includes 237 unaffected individuals, 491 individuals affected with AD without eczema herpeticum, and 49 individuals affected with AD with eczema herpeticum. To perform tests for association between rs28532112 and AD, we compared 491 AD-affected individuals to 237 unaffected individuals by using generalized logistic regression (GLM) and PLINK/Seq and adjusting for the first five PCs as covariates.

### TOPMed analysis of rs28532112 with COPD and lung function

Details of the TOPMed analyses of COPD and lung function can be found in Zhao et al.[42] In brief, the analysis involved 19,996 multi-ethnic individuals, including 12,314 EAs, 6,450 AAs, and 1,232 samples classified as "other," from TOPMed. Phenotype harmonization of pulmonary-function-test measures, including pre-bronchodilator FEV1, forced vital capacity (FVC), and FEV1:FVC ratios, was conducted according to standard protocols. We incorporated covariate adjustment for age$^2$, sex, height$^2$, weight (FVC only), study, current smoking, former smoking, pack-years of smoking, first 10 PCs of ancestry, and sequencing center in an LMM framework to account for heterogenous variance across studies by using GENESIS. Case-control analyses incorporated covariate adjustment for age, sex, study, pack-years, whether an individual ever versus never smoked, first 10 PCs of ancestry, and sequencing center.

### SOMAScan proteomic profiling and *rs334 pQTL* analysis

In JHS and MESA TOPMed participants, EDTA plasma samples collected at the respective baseline exams and stored in −70°C freezers were subjected to proteomic measurements with SOMAscan, a single-stranded DNA aptamer-based proteomics platform containing 1,305 aptamers. In JHS, samples (n = 2,054 AA) were run in three separate batches. Proteins were quantified in relative fluorescent units, the concentration of which is proportional to protein concentration in the plasma sample. Proteomic measurements were standardized to a set of control samples (pooled plasma) contained within each 96-well plate, and the resulting values were log transformed and scaled to a mean of 0 and standard deviation of 1. Association between rs334 genotype and protein values were assessed via linear mixed-effects models. In JHS, proteins were standardized within each batch and then inverse normalized across batches and adjusted for age, sex, and batch. MESA samples (n = 189 AA and 301 HA) were adjusted for age, sex, ethnicity (Hispanic yes or no), plate, and site. The cohort-specific results were meta-analyzed by inverse-variance weighting. A Bonferroni-adjusted significance threshold of $3.8 \times 10^{-5}$ (0.05/1,301) was used.

### rs28532112 plasma-protein association analysis

We further performed a targeted genotype plasma-protein quantification analysis to determine the association of rs28532112 with plasma concentrations of interleukin-3 (IL-3), soluble IL3R-alpha, granulocyte-macrophage colony-stimulating factor (GM-CSF), and thymic stromal lymphopoietin (TSLP) receptor and ligand by utilizing available SOMAscan data from 2,544 multi-ethnic JHS and MESA TOPMed samples. For soluble GM-CSF receptor alpha, we utilized a separate Olink Neurology proximity extension assay (PEA) panel measured in 1,328 multi-ethnic TOPMed WHI cohort samples, adjusted for age and ancestry.

### rs334 and estimated lymphocyte subset analysis in JHS

Illumina MethylationEPIC array data (containing over 850,000 CpG methylation sites) from n = 1,756 JHS participants were generated from blood samples collected during the JHS baseline exam. Methylation levels were quantified in terms of the β value, for which the ratio of intensities between the methylated and unmethylated allele was used as the ratio of fluorescent signals. Methylation values were normalized with respect to background color intensity via the normal-exponential out-of-band (NOOB) method.[43] Cell counts (granulocytes, monocytes, natural killer [NK], CD4$^+$ T lymphocytes, naive CD8$^+$ T lymphocytes, exhausted cytotoxic CD8$^+$ T cells [defined as CD8 positive, CD28 negative, and CD45R negative], and plasmablasts) were estimated according to the method of Houseman et al.[44] and Horvath et al.[45] The association between estimated cell counts and rs334 carriers (excluding rs334 homozygotes), adjusted for age, sex, and 10 PCs of genetic ancestry, was assessed with generalized estimating equations in SAS 9.3 so that familial correlation would be accounted for.

## Results

### Single-nucleotide-variant-association results

In up to 61,802 multi-ethnic individuals (33,285 EA, 14,246 AA, 13,585 HA, and 686 EAS; Table S1), we performed genome-wide association tests for each single-nucleotide variant (SNV) or small insertion-deletion (indel) with a MAC > 5 for ~109,563,748 association tests (Table S2) with 6 different phenotypes (total WBC, neutrophil, monocyte, lymphocyte , basophil, and eosinophil counts).

Across these traits, we observed 6,993 statistically significant associations ($p < 5 \times 10^{-8}$). Inspection of QQ plots and genomic inflation factors indicated well-calibrated p value distributions (Figures S3–S14). To determine whether any of these significant loci represent new associations for WBC-count traits, we performed genome-wide analyses conditioning on all known WBC-count-associated loci

Table 1. Replicated associations with white blood cell traits in TOPMed

| Trait | Chr | Position | Ref | Alt | rsID | Annotation | Beta[a] | SE | p value | Overall AF[b] | EA AF[c] | EA β | EA p | AA AF[d] | AA β | AA p | HA AF[e] | HA β | HA p | Replication β | Replication p |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EOS | X | 1256294 | A | G | rs28532112 | intergenic<br>*CSF2RA* | −0.009 | 0.001 | $1.08 \times 10^{-11}$ | 0.2578 | 0.1579 | −0.0097 | $2.49 \times 10^{-5}$ | 0.4276 | −0.0055 | 0.0079 | 0.2571 | −0.0139 | 0.0003 | −0.0901 | $4.82 \times 10^{-6}$ |
| LYM | 11 | 5227002 | T | A | rs334 | missense<br>*HBB* | −0.237 | 0.026 | $2.76 \times 10^{-20}$ | 0.015 | 0.0003 | 0.0396 | 0.8985 | 0.0423 | −0.19 | $3.30 \times 10^{-14}$ | 0.012 | −0.2797 | $1.41 \times 10^{-6}$ | −0.1737 | $2.88 \times 10^{-13}$ |
| MON | 21 | 15012619 | A | G | rs28574812 | intronic<br>*NRIP1* | −0.011 | 0.002 | $1.54 \times 10^{-10}$ | 0.287 | 0.1592 | −0.0109 | 0.0006 | 0.5318 | −0.0102 | 0.0002 | 0.245 | −0.0101 | 0.0194 | −0.0231 | $2.19 \times 10^{-7}$ |
| MON | 9 | 88921159 | C | A | rs28450540 | intergenic<br>*S1PR3* | −0.035 | 0.004 | $3.65 \times 10^{-17}$ | 0.04 | 0.0009 | −0.0399 | 0.3284 | 0.117 | −0.033 | $3.44 \times 10^{-12}$ | 0.0245 | −0.0532 | $5.19 \times 10^{-5}$ | −0.2161 | $2.64 \times 10^{-25}$ |
| WBC | 11 | 132819661 | GAC | G | rs79353195 | intronic<br>*OPCML* | −0.135 | 0.025 | $4.00 \times 10^{-8}$ | 0.0629 | 0.0465 | −0.1285 | 0.0006 | 0.0658 | −0.1072 | 0.0283 | 0.0629 | −0.1353 | $4.00 \times 10^{-8}$ | −0.0207 | 0.0159 |
| WBC | 21 | 15001515 | A | T | rs2823002 | intronic<br>*NRIP1* | −0.079 | 0.014 | $3.75 \times 10^{-8}$ | 0.3198 | 0.1625 | −0.0823 | 0.0001 | 0.6943 | −0.0828 | 0.0021 | 0.3198 | −0.0794 | $3.75 \times 10^{-8}$ | −0.0186 | 0.0002 |
| WBC | 3 | 132710603 | C | T | rs62292471 | intronic<br>*NPHP3* | 0.1139 | 0.02 | $1.07 \times 10^{-8}$ | 0.0965 | 0.1218 | 0.1104 | $5.64 \times 10^{-6}$ | 0.025 | −0.0009 | 0.9906 | 0.0965 | 0.1139 | $1.07 \times 10^{-8}$ | 0.0223 | $7.40 \times 10^{-5}$ |

[a] Effect sizes are represented in transformed trait values.
[b] AF = allele frequency
[c] EA = European American
[d] AA = African American
[e] HA = Hispanic American

(see material and methods; Table S3). After conditional analysis, we observed 165 statistically significant associations, at a genome-wide threshold of $5 \times 10^{-8}$ (Table S4), implicating 18 independent loci (that were at least 500 kb apart from each other) that have not been previously reported (Table S5). To replicate the findings from the conditional analysis, we performed lookups in independent samples from five different sources, representing self-identified European and European American, Hispanic American, and African American groups in up to 199,126 individuals (see material and methods). Of the 21 independent associations (at the 18 loci) that were submitted for replication, seven signals were robustly replicated (i.e., they met the following criteria: (1) consistent direction of effect between the discovery effect estimate and the meta-analyzed replication effect estimate; (2) p value < 0.05 in at least one replication cohort; and (3) p value < 0.05 across the meta-analyzed replication cohorts) in or near *HBB*, *NRIP1*, *CSF2RA*, *S1PR3*, *NPHP3*, and *OPCML* (Table 1). Several of the replicated lead variants showed large allele-frequency differences between populations (Table S5). In particular, the lead *HBB* and *S1PR3* variants are found almost exclusively among individuals of African ancestry.

### *HBB*

At *HBB* on chromosome 11 (Figure 1A), we observed an association between lower lymphocyte counts and the missense mutation (rs334 [p.Glu6Val]) causing African sickle-cell disease ($\beta = -0.237$, $p = 2.76 \times 10^{-20}$). There was a nominal association of the rs334 sickle variant with *increased* neutrophil counts ($\beta = 0.225$, $p = 4.83 \times 10^{-6}$) and no association with total WBC counts (Table S6). The association with lymphocyte percentage (i.e., lymphocyte counts/total WBC counts) was also strong ($\beta = -0.030$, $p = 2.12 \times 10^{-25}$), as was the association with neutrophil percentage ($\beta = 0.032$, $p = 1.59 \times 10^{-21}$) (Table S7). The 95% fine-mapped credible set for lymphocyte count contained only the index variant, rs334, and PIP = 0.999 (Table S8). Exclusion of the very small number of rs334 homozygote individuals (n = 9) did not alter the association with lymphocyte count or with neutrophil percentage, suggesting that these associations are not driven by altered immune responsiveness or inflammation among individuals with sickle-cell disease. Through investigating a subset of the ~3,400 African American individuals from JHS, we further demonstrated that additional covariate adjustment for other known sickle-cell-related or inflammation-related traits (red-cell indices, kidney function, D-dimer) did not substantively alter the rs334-lymphocyte count association (Table S9), suggesting that the lymphocyte association is not mediated through these other phenotypes. In a smaller subset of 1,458 JHS TOPMed individuals with lymphocyte and immune-cell subtype proportions estimated from genome-wide methylation data, heterozygosity of the rs334 sickle cell mutation was specifically associated with lower estimated levels of $CD8^+$ T lymphocytes and NK cells (Table S10).

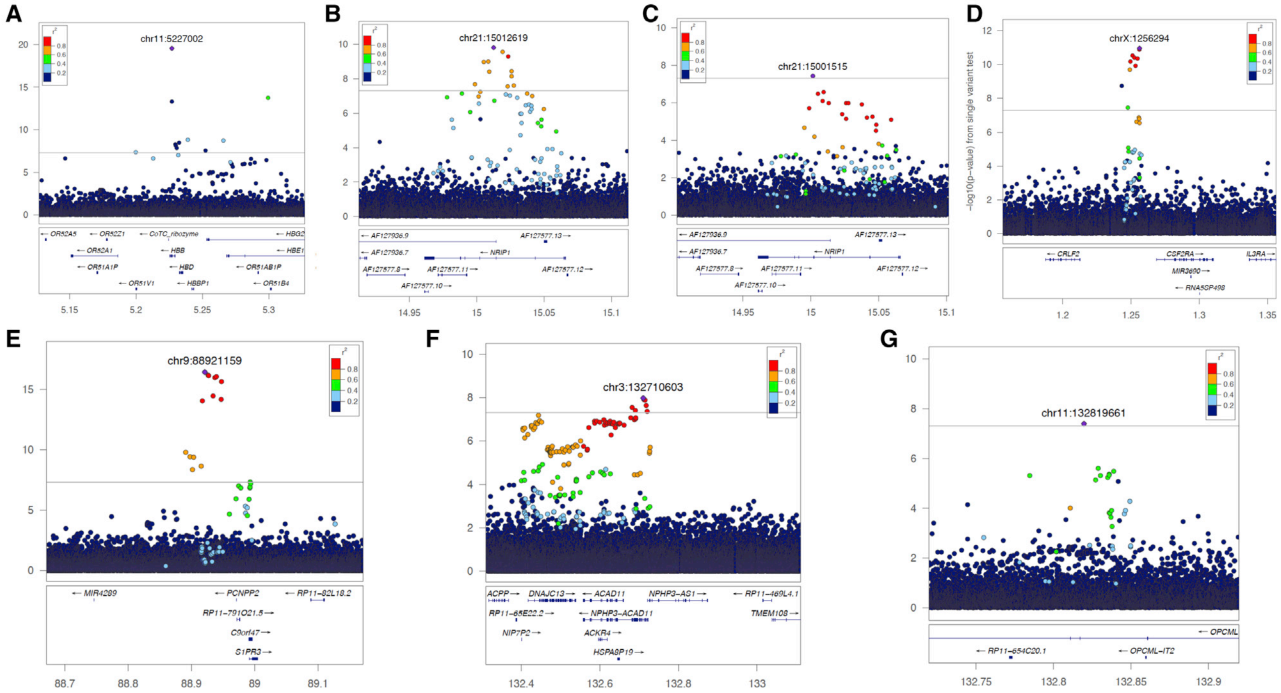

**Figure 1. Regional plots showing patterns of LD and evidence for association at the seven newly reported leukocyte-trait loci. Genomic coordinates are displayed on the horizontal axis, and −log10 p values are displayed on the vertical axis.**
(A) rs334 (chr11: 5227002) with lymphocytes; (B) rs28574812 (chr21: 15012619) with monocytes; (C) rs2823002 (chr21: 15001515) with WBCs; (D) rs28532112 (chrX: 1256294) with eosinophils; (E) rs28450540 (chr9: 88921159) with monocytes; (F) rs62292471 (chr3: 132710603) with WBCs; and (G) rs79353195 (chr11: 132819661) with WBCs.

Interestingly, at the nucleotide position adjacent to rs334, we observed an association of the African *HBB* variant rs33930165 (encoding hemoglobin C, or p.Glu6Lys) with both higher total WBC count ($\beta = 0.672, p = 8.81 \times 10^{-12}$) and higher lymphocyte count ($\beta = 0.351$, $p = 4.71 \times 10^{-14}$). The associations of the *HBB* rs33930165 missense variant with higher total WBC and lymphocyte counts are consistent with a prior analysis conducted in n = 21,513 African Americans imputed to the TOPMed freeze 5b reference panel.[12] The differing WBC-count phenotypic patterns of association between rs334 and rs33930165 in the current TOPMed analysis are summarized in Table S6. We also conducted a two-SNP haplotype analysis of rs334 and rs33930165 with both WBCs and lymphocytes (Table S12). The haplotype analysis was consistent with the single-variant association results in Table S4 (i.e., with the finding that the T allele at rs33930165 increases WBC count and that the A allele at rs334 decreases lymphocyte counts). Of note, both rs334 and rs33930165 are in very low LD ($rsq < 0.001$) and are present on distinct haplotypes in TOPMed (Table S13), consistent with prior evolutionary and population genetic data. Both variants are maintained at relatively high frequencies among populations such as those in sub-Saharan Africa, where malaria is endemic, but the geographic and population distributions and evolutionary histories of the rs334 hemoglobin S and rs33930165 hemoglobin C variant alleles are distinct.[46,47]

To further address the mechanism of the associations at *HBB*, we first assessed the relationship between rs334 and other potential immune-response-related genes in the genomic region. Although there are several type 1 interferon-inducible viral-related genes located about 400 kb centromeric to *HBB* on chromosome 11, we found no evidence of physical interaction between *HBB* and these neighboring genes when we used available promoter capture datasets in relevant blood cell types (with the HUGIN tool) nor any evidence of influence of rs334 on gene expression (eQTL) in whole blood when we used GTEx v8. It should be noted that the latter eQTL analysis might be limited by the small number of African American samples. Integration of additional functional-genomic information via the FATHMM, FANTOM5, and Roadmap annotation databases did not reveal evidence of a *cis*-regulatory role for rs334 (Table S11). Next, we performed proteome-wide analysis of rs334 genotype with ~1,300 plasma proteins measured in African Americans and Hispanics from the MESA cohort and 2,045 African Americans from the JHS cohort (Table S12). Circulating levels of several red blood cell (RBC) (plasma hemoglobin, erythropoietin, and ephrin B2[48]) and manganese proteins related to kidney function (cystatin C and β2-microglobulin) were significantly higher or lower (testican-2)[49] among rs334 variant allele carriers (Bonferroni-corrected p value $< 3.78 \times 10^{-5}$). Moreover, several other proteins related to inflammatory response and lymphocyte activation or signal transduction (fractalkine/CX3CL1, sTNFsRII, and CD59) or response to viral infection (CD59 and testican-2)[50] were significantly higher among rs334 variant carriers.

### NRIP1

Two intronic variants within nuclear receptor interacting protein (*NRIP1*) (Figure 1B, Figure 1C) were associated with monocyte count (rs28574812, $\beta = -0.011$, $p = 1.54 \times 10^{-10}$) and total WBC count (rs2823002, $\beta = -0.079$, $p = 3.75 \times 10^{-8}$). A weaker association was observed between rs28574812 and monocyte percentage ($\beta = -7.75 \times 10^{-4}$, $p = 2.36 \times 10^{-4}$) (Table S7). The rs28574812 and rs2823002 variants are in strong LD in Europeans but are highly differentiated between African and non-African populations. In the 1000 Genomes Yoruba (YRI) West African population, rs28574812 has no other SNPs in strong LD (rsq > 0.8). Analysis of two-SNP haplotypes comprising rs28574812 and rs2823002 confirm the minor alleles of the two SNVs more commonly occur together in individuals of European ancestry. In addition, haplotype association analyses show that the two haplotypes containing the rs28574812 variant G allele are each associated with lower monocyte count, whereas the two haplotypes containing the rs2823002 variant T allele are each associated with lower WBC count (Table S15). This suggests allelic heterogeneity at this locus and that different alleles might be responsible for the monocyte and total WBC-count phenotypic associations. The haplotype association results are also consistent with our fine-mapping results, which pointed to different causal variants for each signal: the 95% credible set for the locus represented by rs28574812 contained 6 variants (Table S8), and the PIP for rs28574812 was 0.422; the 95% credible set for the locus represented by rs2823002 contained 55 variants (Table S8), and the PIP for rs2823002 was 0.611.

The mechanism of association of the *NRIP1* intronic variant with hematologic traits is not immediately apparent. The protein (RIP140) encoded by *NRIP1* interacts with hormone-dependent domains of nuclear receptors and is known to modulate transcriptional activity of the estrogen receptor (ESR1). It is a key regulator of gene expression via interaction with nuclear receptors, transcription factors, and other coregulators, each of which acts as either a coactivator or corepressor. GTEx version 8 did not report either the rs28574812 or rs2823002 variants as a statistically significant eQTL. Data from the Roadmap Epigenomics Consortium indicate elevated histone chromatin immunoprecipitation sequencing (ChIP-seq) experiments acetylated for H3K4me3 and H3K9ac across several types of cell line (Table S11). It has been reported that RIP-140 plays a role in the macrophage switching between classical M1 and alternative M2 subtypes[51] and in the epigenetic responsiveness of monocytes to vitamin D,[52] both of which are important for innate immunity and inflammatory diseases. Investigation of the *NRIP1* region via Phenoscanner and the genome-wide association study (GWAS) catalog reveals that a distinct *NRIP1* missense variant rs2229742 (common only in EUR) has been associated with hemoglobin, RBC count, systolic blood pressure, vitamin D levels, birthweight, and myopia. Additionally, *NRIP1* gene expression signatures can predict survival in chronic lymphocytic leukemia.[53]

### CLRF2-CSF2RA-IL3RA

Analyses of chromosome X found an association between rs28532112 and lower eosinophil count ($\beta = -0.009$, $p = 1.08 \times 10^{-11}$) within the intergenic region between cytokine-receptor-like factor 2 (*CLRF2)* and the region ~20 kb upstream of colony-stimulating factor 2 receptor subunit alpha (*CSF2RA)*, located in the pseudo-autosomal region (PAR1) of the X and Y chromosomes (Figure 1D). The association between rs28532112 and eosinophil percentage was similarly strong ($\beta = -0.0013$, $p = 5.56 \times 10^{-12}$) (Table S7). The 95% credible set for this locus contained 7 variants (Table S8), and the PIP for rs28532112 was 0.324. This region contains the genes encoding three related cytokine receptors, *CRLF2* (the receptor for TSLP), *CSF2RA* (the receptor for colony-stimulating factor 2 (CSF-2) or GM-CSF), and *IL3RA* (the receptor for IL-3), all of which are involved in the regulation of hematopoiesis and type 2 inflammatory responses, including eosinophil production and function.[54,55] Together with the interleukin 7 receptor-alpha, *CRLF2* and *TSLP* activate STAT3 and STAT5, which polarize dendritic cells to induce type 2 inflammatory cytokines (IL-4, IL-5, and IL-13) and directly expand and/or activate Th2 cells, group 2 innate lymphoid cells, eosinophils, and basophils.[56] *CSF2RA* encodes the alpha subunit of the heterodimeric receptor for GM-CSF, a cytokine that controls the production, differentiation, and function of granulocytes and macrophages. *IL3RA* encodes the IL-3-specific alpha subunit of the heterodimeric IL-3 receptor.

In GTEx version 8, rs28532112 was an eQTL associated with the nearby *IL3RA* ($p = 4.8 \times 10^{-9}$) in thyroid tissue and with *CSF2RA* in whole blood ($p = 1.8 \times 10^{-6}$), suggesting either *IL3RA* or *CSF2RA* as the potential causal gene(s) at this locus. Additional functional genomic annotation with FATHMM, FANTOM5, and Roadmap databases did not reveal additional regulatory features for rs28532112 (Table S11). We further performed a targeted genotype-plasma protein quantification analysis of IL-3, GM-CSF, and TSLP receptors and ligands by using available SOMAscan data from 2,544 African Americans and Hispanic/Latinos from the JHS and MESA TOPMed samples. Of the 5 proteins measured within the SOMAscan panel, there was no significant association between rs28532112 and circulating concentrations of IL-3, soluble IL3R-alpha, TSLP, TSLP receptor, or GM-CSF (all Bonferroni-corrected $p > 0.05$). GM-CSF receptor alpha is not one of the proteins included in the SOMAscan platform, but it was measured through a separate Olink PEA panel in 1,328 multi-ethnic TOPMed WHI cohort samples. In the WHI TOPMed samples, there was no evidence of association between rs28532112 genotype and plasma GM-CSF receptor alpha levels ($\beta = -0.024$, $p = 0.47$).

### *S1PR3*

On chromosome 9 (Figure 1E), we identified an African-specific (MAF = 0.129) variant, rs28450540, associated with lower monocyte count ($\beta = -0.035$, $p = 5.18\times10^{-17}$). The association between rs28450540 and monocyte percentage was similarly strong ($\beta = -0.004$, $p = 5.81 \times 10^{-12}$) (Table S7). The 95% credible set for this locus contained 6 variants (Table S8), and the PIP for rs285450540 was 0.330. The rs28450540 variant is located 100 kb upstream of *S1PR3*, where a nearby SNP (rs567880204) was recently associated with monocyte count[8] (LD between these two SNPs was low in AAs; rsq = 0.03). The broader region near *S1PR3* was reported on extensively[5] and was associated with a number of different blood-cell traits in individuals of European ancestry; such traits included monocyte counts, total WBC counts, lymphocyte counts, and platelet counts. The previously reported signals near *S1PR3* include an *S1PR3* missense variant associated with higher monocyte count. However, on the basis of our conditional analyses, the African-specific rs28450540 association with monocyte counts appears to be independent of these known signals, adding to the already complex allelic heterogeneity at this locus. Although GTEx contains very few African American samples, rs28450540 appears to be an eQTL for *S1PR3* ($p = 2.9 \times 10^{-5}$) in whole blood. Functional annotation suggests that this variant is located in a putative enhancer (Table S11) in a number of different primary human cells and cell lines, including those of hematopoietic origin.

*S1PR3* encodes one of five type G-protein-coupled receptors that mediate the biologic effects of sphingosine-1-phosphate (S1P), a chemoattractant for various blood and immune cells.[57] An S1P gradient maintained between blood and other hematopoietic tissues (e.g., bone marrow and thymus lymph nodes) is an important mechanism for blood and immune-cell trafficking. *S1pr3*-deficient mice have defects in leukocyte recruitment and P-selectin-dependent leukocyte rolling, suggesting that *S1PR3* mediates the chemotactic effect of S1P in bone-marrow-derived monocyte and macrophage recruitment during inflammation,[58–61] which might affect circulating monocyte count. Additionally, S1PR3 is highly expressed on hematopoietic stem and progenitor cells (HSPCs) and might affect egress of HSPCs from the bone marrow.[62]

### NPHP3

An intronic variant (rs62292471) on chromosome 3 (Figure 1F) in *NPHP3* was associated with increased total WBC count ($\beta = 0.114$, $p = 1.07 \times 10^{-8}$). The 95% credible set for this locus contained 33 variants (Table S8), and the PIP for rs62292471 was 0.150. GTEx version 8 reports rs62292471 as an eQTL for *NPHP3* in cultured fibroblasts and in heart, adipose, lung, esophagus, and muscle tissue and as an eQTL for *DNAJC13* in adipose, esophagus, and muscle tissue. The variant is located within an ENCODE distal enhancer region (Table S11). Another *NPHP3* intronic variant, rs572076167 (not in 1000 Genomes), was associated with red-cell mean corpuscular hemoglobin (MCH) concentration,[8] whereas rs17348614 was associated with lower MCH and mean corpuscular volume (MCV).[5] Loss-of-function mutations in *NPHP3* cause the congenital cystic kidney disorder nephronophthisis (NPH). Of the other genes located within a 1 Mb window surrounding *NPHP3*, the only one with a biologic connection to blood cells is *ACKR4*, which encodes a chemokine receptor that binds to dendritic-cell- and T-cell-activated chemokines, including CCL19, CCL21, and CCL25. ACKR4 belongs to the family of atypical chemokine receptors that includes DARC, which contains a well-characterized loss-of-function promoter variant (rs2814778) that is a major genetic determinant of WBC and neutrophil count in populations with African ancestry. By scavenging chemokines, ACKR4 regulates dendritic-cell trafficking to lymph nodes during inflammation.[63]

### OPCML

We observed a 3-base indel (rs79353195) in the intron of *OPCML* (Figure 1G) associated with decreased total WBC count ($\beta = -0.135$, $p = 4.00\times10^{-9}$) with modest evidence for replication ($p = 0.0159$). The 95% credible set for this locus contained 48 variants (Table S8), and the PIP for rs79353195 was 0.868. This variant appears to be highly stratified across populations with MAF = 0.53 in EAS, MAF = 0.26 in SAS, MAF = 0.06 in AA, and MAF = 0.05 in EA populations. *OPCML* is a large gene that encodes an opioid-binding cell-adhesion molecule-like preprotein, a member of the IgLON immunoglobulin protein family highly expressed in the brain. Defects in OPCML are a cause of susceptibility to ovarian cancer (MIM: 167000). There is no apparent connection of this locus with blood cells or inflammation.

### Results of association tests for aggregate rare variants

To improve the power to detect rare-variant associations, we implemented several strategies of aggregating variants and testing for cumulative associations of gene-based groupings with the traits (see materials and methods). In so doing, we detected statistically significant ($p < 2.76 \times 10^{-6}$) associations at four different genes (Table 2): *MARCKSL1* ($p = 2.98 \times 10^{-7}$), *TET2* ($p = 6.21 \times 10^{-9}$), *FLT3* ($p = 8.96 \times 10^{-7}$), and *CNKSR2* ($p = 2.46 \times 10^{-6}$). Replication analyses in independent samples from the UK-Biobank ($p = 2.42 \times 10^{-18}$) and INTERVAL ($p = 0.0003$) studies confirmed the associations at *FLT3*. Analyses of the corresponding WBC-subtype proportions were consistent with these results (Table S16).

*TET2* encodes a demethylation enzyme and epigenetic regulator[64] that plays an important role in HSPC renewal, lineage commitment, and monocyte differentiation.[65] The *TET2* association with monocyte counts was spread across 72 rare coding variants (Figure S1). Because *TET2* is a known driver gene for clonal hematopoiesis of indeterminate potential (CHIP), myeloid, and lymphoid malignancies, we compared our rare *TET2* variants with those that are

**Table 2. Gene-based rare variant associations with white blood cell counts in TOPMed**

| Trait | Gene | Number of variants | Discovery p value | INTERVAL p value | UKBiobank p value |
|---|---|---|---|---|---|
| NEU | *MARCKSL1* | 11 | $2.98 \times 10^{-7}$ | 0.5576 | 0.1008 |
| LYM | *CNKSR2* | 13 | $2.46 \times 10^{-6}$ | 0.8067 | 0.8079 |
| MON | *TET2* | 72 | $6.21 \times 10^{-9}$ | 0.6984 | 0.6091 |
| MON | *FLT3* | 65 | $8.96 \times 10^{-7}$ | 0.0003 | $2.42 \times 10^{-18}$ |

reported as somatic mutations[66] or that appear in the COSMIC database. Indeed, we found that the majority of variants (53 of 72 variants) included in our aggregate-rare-variant test for *TET2* were also reported as somatic in TOPMed, and these mutations largely drove the association with higher monocyte counts (Figure S1). These findings are further supported by the recent observation that somatic mutations of *TET2* are commonly found among individuals referred to a hematology service for evaluation of monocytosis.[67] By contrast, reports of germline *TET2* loss-of-function variants associated with hematologic disease are quite rare.[68–70]

*FLT3* encodes a receptor tyrosine kinase that regulates early hematopoiesis as well as the development of monocytes and dendritic cells. Somatic variants of FLT3 are found commonly in individuals with acute myeloid leukemia (AML) and generally consist of gain-of-function mutations involving either internal tandem duplications of the *FLT3* juxtamembrane domain or point mutations located within the tyrosine kinase domain, both of which lead to constitutive activation of the FLT3 receptor.[71] Because some *TET2* somatic mutations had clearly passed quality control for germline variants, we also checked whether the rare variants contributing to the *FLT3* signal were somatic. The COSMIC database identified some evidence for overlap between the TOPMed rare variants in *FLT3*, yet the presumably somatic mutations (12 of 65 variants) that contributed to the *FLT3* association did not appear to drive the result (Figure S2). This observation is consistent with the recent discovery that common and low-frequency germline genetic variants of *FLT3* are associated with monocyte count as well as risk of autoimmune disease.[7,72] Interestingly, several of the rare *FLT3* missense variants driving the association with monocyte count in TOPMed are located within the juxtamembrane or tyrosine kinase domains, which are also the most common location of somatic *FLT3* mutations found in human cancers.

### Results of phenome-wide association tests

Loci associated with WBC-count traits are often pleiotropically involved in autoimmune, allergic, infectious, and other blood-related diseases.[4,5] Therefore, we additionally assessed whether any of the other newly identified WBC-count trait-associated variants are associated with clinical disease outcomes by using a combination of existing GWAS andPheWAS databases as well as evaluation of the X chromosome PAR locus and *FLT3* aggregated rare variants in WGS-based datasets.

Most of our newly discovered autosomal loci associated with WBC-count traits, *S1PR3*, *NRIP1*, *NPHP3*, *HBB*, and *OPCML*, disproportionately impact individuals of African ancestry. Therefore, in addition to the large but Euro-centric UKBiobank (UKBB), we utilized PheWAS genotype and phenotype data from the more diverse BioVU EHR-based biobank at Vanderbilt University Medical Center. The latter includes African Americans genotyped on the Illumina MEGA array imputed to Haplotype Reference Consortium (HRC) reference genomes. For evaluation of *HBB* rs334 in BioVU, we excluded homozygous individuals because of the known relationship of sickle-cell disease (SCD) to clinical outcomes. For the *S1PR3*, *NRIP1*, and *OPCML* WBC-count trait-associated index variants, we found no evidence of significant association with disease outcomes in either UKBB or BioVU, whereas the *NPHP3* index variant showed borderline association with myocardial infarction (Table S17). In the BioVU African American PheWAS, heterozygosity for rs334 was associated with several hemolytic anemia-related diagnosis codes, but not with any immune-related or infectious diseases (Table S17).

Eosinophils are classically associated with type 2 inflammation and are one of the hallmarks of allergic diseases such as asthma and AD. A subset of subjects with COPD might also be enriched for circulating eosinophils. Because the UKBiobank PheWAS lookup tool was restricted to the autosomes, we tested specifically for the X chromosome locus upstream of *CSF2RA* and its association with these three diseases in the UKBB imputed GWAS dataset (Table S18). We found a significant association with asthma in the UKBB (odds ratio [OR] = 0.94, $p = 3.52 \times 10^{-6}$); notably the association is as expected, i.e., the allele that decreases eosinophil count in our discovery is also associated with decreased risk for asthma. No associations were observed with COPD or AD in the UKBB.

Given the availability of several TOPMed lung-disease cohorts with data on asthma, COPD, and pulmonary function, we expanded our lookup of this variant in the following studies from TOPMed: BAGS for asthma, SARP for asthma severity, and ARIC, CHS, FHS, JHS, MESA, COPDGene, and EOCOPD for lung function and COPD (Table S18). In ~20,000 multi-ethnic individuals,[42] rs28532112 showed no association with spirometric measures of pulmonary function (FEV1, FVC, or FEV1:FVC ratio). There was a nominal association with COPD ($p = 0.025$) and severe COPD (p = 0.036), but paradoxically the minor allele associated with lower eosinophil count was associated with greater risk of COPD (OR = 1.09). In the BAGS asthma study there was no association of rs28532112 with asthma (p = 0.91), and no associations were noted for asthma severity in the SARP TOPMed study (p = 0.944 in 393 EAs, p = 0.144 in 218 AAs); however, we should note that these two asthma cohorts were underpowered and most likely unable to recapitulate the associations noted for asthma in the UKBB. Given the small sample size with a phecode definition of AD in the UKBB, we also took advantage of WGS data available from the ADRN, where in-depth phenotyping could help overcome the heterogeneity possible in the UKBB. In the ADRN samples, we found rs28532112 was associated with AD (OR 0.63; p = 0.008), and once again the association was as expected, i.e., the allele that decreases eosinophil count in our discovery was also associated with decreased risk for AD.

Because a low-frequency monocyte-count-associated variant of *FLT3* (rs76428106) was recently associated with autoimmune thyroid disease (AITD),[72] we used whole-exome sequencing to examine whether aggregated rare variants of *FLT3* were associated with AITD among 200,000 UKBB individuals. Among 6,686 AITD-affected individuals and 179,346 individuals without AITD, there was no evidence of association by either burden (p = 0.81) or sequence kernel association test (p = 0.38). However, because of the relatively small number of cases in the UKBB, the lack of association might be due to low statistical power.

## Discussion

By expanding coverage of the genome through deep WGS performed in a large sample of diverse individuals, we have identified several loci that are associated with WBC-count traits but that are distinct from variants previously identified through large GWASs. In each instance, the identified single variants (*HBB*, *S1PR3*, *NPHP3*, *NRIP1*, *OPCML*, and *CSF2RA*) are highly differentiated in allele frequency across ancestral populations; *HBB* and *S1PR3* are essentially monomorphic in Europeans. The eosinophil-lowering *CSF2RA* variant located on the X chromosome may be associated with reduced susceptibility to asthma and AD. We also identified a burden of rare coding variants in *FLT3* associated with monocyte count. These results demonstrate the utility of WGS in diverse cohorts of apparently healthy individuals for our further understanding of the genetic architecture of WBC traits and their relationship to immune-related disorders. Multi-omic data available from diverse TOPMed samples and other sources contributed to defining the likely causal genes or molecular mechanisms underlying these associations.

Despite its importance for hematologic traits and Mendelian disorders (e.g., Diamond-Blackfan anemia, hemophilia, and G6PD deficiency), the X chromosome has been under-studied in complex-trait genetics through GWASs, largely because of analytical issues and challenges related to imputation and sex-related differences in gene dosage. In particular, PAR1 on the X and Y chromosomes recombines in a sex-biased manner and thus has traditionally been ignored in linkage and association studies.[73] The application of WGS allowed us to circumvent the need for genotype imputation and directly identify a common-variant association that prior GWASs had missed within PAR1. We have identified and replicated a common variant associated with lower eosinophil count within the X chromosome PAR1 region between *CSF2RA* (GM-CSF) and *CRLF2* (cytokine receptor-like factor 2). The sentinel variant rs28532112 is about three times as common in African as in European populations, which might also have contributed to our ability to discover its association with eosinophil count. The recently completed gapless, end-to-end assembly of the human X chromosome reference sequence, including PARs,[74] should additionally facilitate identification of X-linked or pseudo-autosomal variants newly associated with complex traits.

Several of the autosomal WBC-count trait-associated loci that we report (*S1PR3*, *NPHP3*, *HBB*, and *NRIP1)* contain other nearby variants that have been previously associated with hematologic traits, suggesting broader hematopoietic lineage regulation of these genomic regions. For example, previous studies in predominantly European ancestral populations have identified several genetic variants that occur within a ~500 kb region upstream of *S1PR3* and that are associated with WBC, RBC, and platelet traits.[5,75] Although the rs334 variant encoding hemoglobin S is well known to affect RBC physiology, the mechanism of association for distinct variants within the *NPHP3* and *NRIP1* regions with RBC traits remains unclear.

Our results also extend the clinical importance of variants underlying WBC count and immune-related quantitative traits, particularly to non-European populations. The association of rs334 with a lower lymphocyte count, a higher proportion of neutrophils, and higher plasma levels of several immune- and kidney-related proteins adds to growing evidence that the carrier state of sickle-cell disease (p.Glu7Val [c.20A>T])-encoding hemoglobin S is associated with various medical phenotypes (RBC traits, higher D-dimer levels, and lower eGFR and hemoglobin A1c levels) and disease susceptibility (increased risk of chronic kidney disease and venous thromboembolic disease) in African Americans.[76] There is evidence that, in addition to having a role in leukocyte recruitment and trafficking, S1PR3 is required for myeloid cell oxidative killing of microbial pathogens,[77] is expressed on alveolar epithelial cells, and regulates epithelial integrity in lung disease.[78] In this regard, the *S1PR3* locus shows evidence of being under recent positive selection in African populations and might contribute to pulmonary edema and pathogenesis of severe malaria.[79] Although our PheWAS did not show evidence that the African-specific variant *S1PR3* rs28450540 was associated with any additional chronic-disease outcomes, the importance of rs28450540 or other variants at the *S1PR3* locus for complications from other chronic lung and infectious disease or sickle-cell disease might require larger sample sizes.

The putative association of the X chromosome eosinophil-lowering variant with reduced risk of asthma and AD is consistent with prior studies demonstrating that genetically determined eosinophil count is associated with risk of allergic diseases in UKBB.[5,8] These findings have potential implications for risk stratification or drug development for AD, asthma, or other allergic and lung diseases that disproportionately affect African Americans.[80] Eosinophils and related type 2 inflammatory responses are particularly important at the barrier surfaces of skin and the respiratory and gastrointestinal tracts. Anti-TSLP antibodies such as Tezepelumab (AMG-157/MEDI9929) reduce levels of biomarkers of type 2 inflammation; such biomarkers include, for example, blood and sputum eosinophil counts.[81] Moreover, mutations in *CSF2RA* are a cause of pulmonary surfactant metabolism dysfunction type 4 (SMDP4) (pulmonary alveolar proteinosis) (MIM: 300770), a rare lung disorder. Additional studies might be required for adequate characterization of the role of newly identified genetic variants at this important X-linked cytokine-receptor-gene-family locus for allergic, autoimmune, and pulmonary diseases, especially in the context of disease severity.

Additionally, our findings highlight one of the caveats inherent in WGS studies. The identification of a burden of rare variants in two genes (*TET2* and *FLT3*) linked to leukemogenesis and clonal myeloid expansion and associated with monocyte count among a sample of largely unscreened individuals raises the question of distinguishing between germline, somatic, or clonal hematopoietic variants in blood-based next-generation sequencing studies involving complex traits or diseases. Because clonal hematopoiesis is an age-related condition, case-control WGS germline variant studies of aging-related conditions could be particularly prone to confounding by somatic variants.[82] To avoid such potential confounding in blood-based genome-sequencing association studies, researchers might need to provide additional evidence of the germline or somatic origin by using clinical information, variant characteristics, and/or serial next-generation sequencing assays.[83]

In summary, using a WGS approach in diverse samples, we have identified and replicated 7 single-variant leukocyte-trait associations previously missed by GWASs; one of these is the association between a chromosome X PAR and both lower eosinophil count and reduced risk of allergic diseases. We extend the phenotypic profile of sickle-cell trait, which has previously been associated with RBC, kidney, and thrombosis-related biomarkers to include lymphocyte count and inflammation-related biomarkers. The identification of monocyte-specific associations, including an African-ancestry variant at the *S1PR3* locus and a burden of very rare variants in *FLT3*, might warrant additional study in the context of infectious or autoimmune diseases, respectively.

## Data and code availability

Data for each participating study can be accessed through dbGaP with the corresponding accession number (Amish, phs000956; ARIC, phs001211; BioMe, phs001644; BAGS, phs001143; CARDIA, phs001612; CHS, phs001368; COPDGene, phs000951; FHS, phs000974; GeneSTAR, phs001218; HCHS/SOL, phs001395; JHS, phs000964; MESA, phs001416; SAFS, phs001215; SARP, phs001446; and WHI, phs001237). Analysis results for the conditional single-variant analyses and the aggregate conditional analyses can be accessed through dbGaP: phs001974.

## Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.08.007.

## Acknowledgments

Molecular data for the TOPMed program was supported by the National Heart, Lung and Blood Institute (NHLBI). Study-specific omics support information can be found in the supplement. 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). The project described was also supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through grant KL2TR002490 (L.M.R.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. L.M.R. was additionally funded by the NIH grants R01HL129132 and T32 HL129982. P.L.A. was funded by the NIH grant R01HL130733. P.S.d.V. was supported by American Heart Association grant number 18CDA34110116. We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

## Declaration of interests

The authors declare no competing interests.

Received: April 16, 2021
Accepted: August 11, 2021
Published: September 27, 2021

## Web resources

COSMIC, https://cancer.sanger.ac.uk/cosmic
HUGIN, http://hugin2.genetics.unc.edu/Project/hugin
OASIS pipeline, https://omicsoasis.github.io/
TOPMed website, www.nhlbiwgs.org
TOPMed whole-genome sequencing methods for freeze 8, https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-methods-freeze-8
University of Michigan pheweb server, https://pheweb.org/UKB-SAIGE/

## References

1. Morrell, C.N., Aggrey, A.A., Chapman, L.M., and Modjeski, K.L. (2014). Emerging roles for platelets as immune and inflammatory cells. Blood *123*, 2759–2767.
2. Martinod, K., and Wagner, D.D. (2014). Thrombosis: tangled up in NETs. Blood *123*, 2768–2776.
3. Reiner, A.P., Lettre, G., Nalls, M.A., Ganesh, S.K., Mathias, R., Austin, M.A., Dean, E., Arepalli, S., Britton, A., Chen, Z., et al. (2011). Genome-wide association study of white blood cell count in 16,388 African Americans: the continental origins and genetic epidemiology network (COGENT). PLoS Genet. *7*, e1002108.
4. Tajuddin, S.M., Schick, U.M., Eicher, J.D., Chami, N., Giri, A., Brody, J.A., Hill, W.D., Kacprowski, T., Li, J., Lyytikäinen, L.P., et al. (2016). Large-scale exome-wide association identifies loci for white blood cell traits and pleiotropy with immune-mediated diseases. Am. J. Hum. Genet. *99*, 22–39.
5. Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A., et al. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell *167*, 1415–1429.e19.
6. Iotchkova, V., Huang, J., Morris, J.A., Jain, D., Barbieri, C., Walter, K., Min, J.L., Chen, L., Astle, W., Cocca, M., et al.; UK10K Consortium (2016). Discovery and refinement of genetic loci associated with cardiometabolic risk using dense imputation maps. Nat. Genet. *48*, 1303–1312.
7. Jain, D., Hodonsky, C.J., Schick, U.M., Morrison, J.V., Minnerath, S., Brown, L., Schurmann, C., Liu, Y., Auer, P.L., Laurie, C.A., et al. (2017). Genome-wide association of white blood cell counts in Hispanic/Latino Americans: the Hispanic Community Health Study/Study of Latinos. Hum. Mol. Genet. *26*, 1193–1204.
8. Chen, M.-H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., et al.; VA Million Veteran Program (2020). Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. Cell *182*, 1198–1213.e14.
9. Auer, P.L., Johnsen, J.M., Johnson, A.D., Logsdon, B.A., Lange, L.A., Nalls, M.A., Zhang, G., Franceschini, N., Fox, K., Lange, E.M., et al. (2012). Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. Am. J. Hum. Genet. *91*, 794–808.
10. Auer, P.L., Teumer, A., Schick, U., O'Shaughnessy, A., Lo, K.S., Chami, N., Carlson, C., de Denus, S., Dubé, M.P., Haessler, J., et al. (2014). Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. Nat. Genet. *46*, 629–634.
11. Raffield, L.M., Iyengar, A.K., Wang, B., Gaynor, S.M., Spracklen, C.N., Zhong, X., Kowalski, M.H., Salimi, S., Polfus, L.M., Benjamin, E.J., et al.; TOPMed Inflammation Working Group; and NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium (2020). Allelic Heterogeneity at the CRP Locus Identified by Whole-Genome Sequencing in Multi-ancestry Cohorts. Am. J. Hum. Genet. *106*, 112–120.
12. Kowalski, M.H., Qian, H., Hou, Z., Rosen, J.D., Tapia, A.L., Shan, Y., Jain, D., Argos, M., Arnett, D.K., Avery, C., et al.; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium; and TOPMed Hematology & Hemostasis Working Group (2019). Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. PLoS Genet. *15*, e1008500.
13. The ARIC investigators (1989). The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. Am. J. Epidemiol. *129*, 687–702.
14. Mitchell, B.D., McArdle, P.F., Shen, H., Rampersaud, E., Pollin, T.I., Bielak, L.F., Jaquish, C., Douglas, J.A., Roy-Gagnon, M.H., Sack, P., et al. (2008). The genetic response to short-term interventions affecting cardiovascular function: rationale and design of the Heredity and Phenotype Intervention (HAPI) Heart Study. Am. Heart J. *155*, 823–828.
15. Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W.A., Li, R., Manolio, T.A., Sanderson, S.C., Kannry, J., Zinberg, R., Basford, M.A., et al.; eMERGE Network (2013). The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet. Med. *15*, 761–771.
16. Hughes, G.H., Cutter, G., Donahue, R., Friedman, G.D., Hulley, S., Hunkeler, E., Jacobs, D.R., Jr., Liu, K., Orden, S., Pirie, P., et al. (1987). Recruitment in the coronary artery disease risk development in young adults (Cardia) study. Control. Clin. Trials *8* (4, Suppl), 68S–73S.
17. Fried, L.P., Borhani, N.O., Enright, P., Furberg, C.D., Gardin, J.M., Kronmal, R.A., Kuller, L.H., Manolio, T.A., Mittelmark, M.B., Newman, A., et al. (1991). The cardiovascular health study: design and rationale. Ann. Epidemiol. *1*, 263–276.
18. Regan, E.A., Hokanson, J.E., Murphy, J.R., Make, B., Lynch, D.A., Beaty, T.H., Curran-Everett, D., Silverman, E.K., and Crapo, J.D. (2010). Genetic epidemiology of COPD (COPDGene) study design. COPD *7*, 32–43.
19. Splansky, G.L., Corey, D., Yang, Q., Atwood, L.D., Cupples, L.A., Benjamin, E.J., D'Agostino, R.B., Sr., Fox, C.S., Larson, M.G., Murabito, J.M., et al. (2007). The third generation cohort of the national heart, lung, and blood institute's framingham heart study: design, recruitment, and initial examination. Am. J. Epidemiol. *165*, 1328–1335.
20. Becker, D.M., Segal, J., Vaidya, D., Yanek, L.R., Herrera-Galeano, J.E., Bray, P.F., Moy, T.F., Becker, L.C., and Faraday, N. (2006). Sex differences in platelet reactivity and response to low-dose aspirin therapy. JAMA *295*, 1420–1427.
21. Sorlie, P.D., Avilés-Santa, L.M., Wassertheil-Smoller, S., Kaplan, R.C., Daviglus, M.L., Giachello, A.L., Schneiderman, N., Raij, L., Talavera, G., Allison, M., et al. (2010). Design and implementation of the hispanic community health study/study of latinos. Ann. Epidemiol. *20*, 629–641.
22. Taylor, H.A., Jr., Wilson, J.G., Jones, D.W., Sarpong, D.F., Srinivasan, A., Garrison, R.J., Nelson, C., and Wyatt, S.B. (2005). Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. Ethn. Dis. *15* (4, Suppl 6), S6–S4, 17.
23. Wilson, J.G., Rotimi, C.N., Ekunwe, L., Royal, C.D., Crump, M.E., Wyatt, S.B., Steffes, M.W., Adeyemo, A., Zhou, J., Taylor, H.A., Jr., and Jaquish, C. (2005). Study design for genetic analysis in the Jackson Heart Study. Ethn. Dis. *15* (4, Suppl 6), S6–S30, 37.
24. Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R., Diez Roux, A.V., Folsom, A.R., Greenland, P., Jacob, D.R., Jr., Kronmal, R., Liu, K., et al. (2002). Multi-Ethnic Study of Atherosclerosis: objectives and design. Am. J. Epidemiol. *156*, 871–881.
25. Mitchell, B.D., Kammerer, C.M., Blangero, J., Mahaney, M.C., Rainwater, D.L., Dyke, B., Hixson, J.E., Henkel, R.D., Sharp,

R.M., Comuzzie, A.G., et al. (1996). Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans. The San Antonio Family Heart Study. Circulation *94*, 2159–2170.

26. The Women's Health Initiative Study Group (1998). Design of the Women's Health Initiative clinical trial and observational study. Control. Clin. Trials *19*, 61–109.
27. Conomos, M.P., Miller, M.B., and Thornton, T.A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet. Epidemiol. *39*, 276–293.
28. 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.
29. Conomos, M.P., Laurie, C.A., Stilp, A.M., Gogarten, S.M., McHugh, C.P., Nelson, S.C., Sofer, T., Fernández-Rhodes, L., Justice, A.E., Graff, M., et al. (2016). Genetic diversity and association studies in us hispanic/latino populations: applications in the hispanic community health study/study of latinos. Am. J. Hum. Genet. *98*, 165–184.
30. Sofer, T., Zheng, X., Gogarten, S.M., Laurie, C.A., Grinde, K., Shaffer, J.R., Shungin, D., O'Connell, J.R., Durazo-Arvizo, R.A., Raffield, L., et al.; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium (2019). A fully adjusted two-stage procedure for rank-normalization in genetic association studies. Genet. Epidemiol. *43*, 263–275.
31. Breslow, N.E., and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. *88*, 9–25.
32. Chen, H., Wang, C., Conomos, M.P., Stilp, A.M., Li, Z., Sofer, T., Szpiro, A.A., Chen, W., Brehm, J.M., Celedón, J.C., et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. Am. J. Hum. Genet. *98*, 653–666.
33. Dey, R., Schmidt, E.M., Abecasis, G.R., and Lee, S. (2017). A fast and accurate algorithm to test for binary phenotypes and its application to phewas. Am. J. Hum. Genet. *101*, 37–49.
34. Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. *50*, 1335–1341.
35. Chen, H., Huffman, J.E., Brody, J.A., Wang, C., Lee, S., Li, Z., Gogarten, S.M., Sofer, T., Bielak, L.F., Bis, J.C., et al.; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium; and TOPMed Hematology and Hemostasis Working Group (2019). Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. Am. J. Hum. Genet. *104*, 260–274.
36. Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J.M., Auton, A., Myers, S., Morris, A., et al.; Wellcome Trust Case Control Consortium (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat. Genet. *44*, 1294–1301.
37. Wu, P., Gifford, A., Meng, X., Li, X., Campbell, H., Varley, T., Zhao, J., Carroll, R., Bastarache, L., Denny, J.C., et al. (2019). Mapping icd-10 and icd-10-cm codes to phecodes: workflow development and initial evaluation. JMIR Med. Inform. *7*, e14325.
38. Mathias, R.A., Grant, A.V., Rafaels, N., Hand, T., Gao, L., Vergara, C., Tsai, Y.J., Yang, M., Campbell, M., Foster, C., et al. (2010). A genome-wide association study on African-ancestry populations for asthma. J. Allergy Clin. Immunol. *125*, 336–346.e4.
39. Barnes, K.C., Neely, J.D., Duffy, D.L., Freidhoff, L.R., Breazeale, D.R., Schou, C., Naidu, R.P., Levett, P.N., Renault, B., Kucherlapati, R., et al. (1996). Linkage of asthma and total serum IgE concentration to markers on chromosome 12q: evidence from Afro-Caribbean and Caucasian populations. Genomics *37*, 41–50.
40. Jarjour, N.N., Erzurum, S.C., Bleecker, E.R., Calhoun, W.J., Castro, M., Comhair, S.A., Chung, K.F., Curran-Everett, D., Dweik, R.A., Fain, S.B., et al.; NHLBI Severe Asthma Research Program (SARP) (2012). Severe asthma: lessons learned from the national heart, lung, and blood institute severe asthma research program. Am. J. Respir. Crit. Care Med. *185*, 356–362.
41. Bin, L., Malley, C., Taylor, P., Preethi Boorgula, M., Chavan, S., Daya, M., Mathias, M., Shankar, G., Rafaels, N., Vergara, C., et al. (2021). Whole genome sequencing identifies novel genetic mutations in patients with eczema herpeticum. Allergy *76*, 2510–2523. Published online February 6, 2021.
42. Zhao, X., Qiao, D., Yang, C., Kasela, S., Kim, W., Ma, Y., Shrine, N., Batini, C., Sofer, T., Taliun, S.A.G., et al.; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium; and TOPMed Lung Working Group (2020). Whole genome sequence analysis of pulmonary function and COPD in 19,996 multi-ethnic participants. Nat. Commun. *11*, 5182.
43. Fortin, J.-P., Triche, T.J., Jr., and Hansen, K.D. (2017). Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. Bioinformatics *33*, 558–560.
44. Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., and Kelsey, K.T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics *13*, 86.
45. Horvath, S., and Levine, A.J. (2015). Hiv-1 infection accelerates age according to the epigenetic clock. J. Infect. Dis. *212*, 1563–1573.
46. Ghansah, A., Rockett, K.A., Clark, T.G., Wilson, M.D., Koram, K.A., Oduro, A.R., Amenga-Etego, L., Anyorigiya, T., Hodgson, A., Milligan, P., et al. (2012). Haplotype analyses of haemoglobin C and haemoglobin S and the dynamics of the evolutionary response to malaria in Kassena-Nankana District of Ghana. PLoS ONE *7*, e34565.
47. Esoh, K., and Wonkam, A. (2021). Evolutionary history of sickle-cell mutation: implications for global genetic medicine. Hum. Mol. Genet. *30* (R1), R119–R128.
48. Suenobu, S., Takakura, N., Inada, T., Yamada, Y., Yuasa, H., Zhang, X.Q., Sakano, S., Oike, Y., and Suda, T. (2002). A role of EphB4 receptor and its ligand, ephrin-B2, in erythropoiesis. Biochem. Biophys. Res. Commun. *293*, 1124–1131.
49. Ngo, D., Wen, D., Gao, Y., Keyes, M.J., Drury, E.R., Katz, D.H., Benson, M.D., Sinha, S., Shen, D., Farrell, L.A., et al. (2020). Circulating testican-2 is a podocyte-derived marker of kidney health. Proc. Natl. Acad. Sci. USA *117*, 25026–25035.
50. Ahn, N., Kim, W.-J., Kim, N., Park, H.W., Lee, S.-W., and Yoo, J.-Y. (2019). The interferon-inducible proteoglycan testican-2/spock2 functions as a protective barrier against virus infection of lung epithelial cells. J. Virol. *93*, e00662-19.
51. Lin, Y.-W., Lee, B., Liu, P.-S., and Wei, L.-N. (2016). Receptor-interacting protein 140 orchestrates the dynamics of macrophage m1/m2 polarization. J. Innate Immun. *8*, 97–107.

52. Wilfinger, J., Seuter, S., Tuomainen, T.-P., Virtanen, J.K., Voutilainen, S., Nurmi, T., de Mello, V.D., Uusitupa, M., and Carlberg, C. (2014). Primary vitamin D receptor target genes as biomarkers for the vitamin D3 status in the hematopoietic system. J. Nutr. Biochem. *25*, 875–884.

53. Lapierre, M., Castet-Nicolas, A., Gitenay, D., Jalaguier, S., Teyssier, C., Bret, C., Cartron, G., Moreaux, J., and Cavaillès, V. (2015). Expression and role of RIP140/NRIP1 in chronic lymphocytic leukemia. J. Hematol. Oncol. *8*, 20.

54. Dougan, M., Dranoff, G., and Dougan, S.K. (2019). Gm-csf, il-3, and il-5 family of cytokines: regulators of inflammation. Immunity *50*, 796–811.

55. Esnault, S., and Kelly, E.A. (2016). Essential mechanisms of differential activation of eosinophils by il-3 compared to gm-csf and il-5. Crit. Rev. Immunol. *36*, 429–444.

56. Reche, P.A., Soumelis, V., Gorman, D.M., Clifford, T., Liu Mr, Travis, M., Zurawski, S.M., Johnston, J., Liu, Y.J., Spits, H., et al. (2001). Human thymic stromal lymphopoietin preferentially stimulates myeloid cells. J. Immunol. *167*, 336–343.

57. Rosen, H., and Goetzl, E.J. (2005). Sphingosine 1-phosphate and its receptors: an autocrine and paracrine network. Nat. Rev. Immunol. *5*, 560–570.

58. Nussbaum, C., Bannenberg, S., Keul, P., Gräler, M.H., Gonçalves-de-Albuquerque, C.F., Korhonen, H., von Wnuck Lipinski, K., Heusch, G., de Castro Faria Neto, H.C., Rohwedder, I., et al. (2015). Sphingosine-1-phosphate receptor 3 promotes leukocyte rolling by mobilizing endothelial P-selectin. Nat. Commun. *6*, 6416.

59. Keul, P., Lucke, S., von Wnuck Lipinski, K., Bode, C., Gräler, M., Heusch, G., and Levkau, B. (2011). Sphingosine-1-phosphate receptor 3 promotes recruitment of monocyte/macrophages in inflammation and atherosclerosis. Circ. Res. *108*, 314–323.

60. Murakami, K., Kohno, M., Kadoya, M., Nagahara, H., Fujii, W., Seno, T., Yamamoto, A., Oda, R., Fujiwara, H., Kubo, T., et al. (2014). Knock out of S1P3 receptor signaling attenuates inflammation and fibrosis in bleomycin-induced lung injury mice model. PLoS ONE *9*, e106792.

61. Yang, L., Han, Z., Tian, L., Mai, P., Zhang, Y., Wang, L., and Li, L. (2015). Sphingosine 1-phosphate receptor 2 and 3 mediate bone marrow-derived monocyte/macrophage motility in cholestatic liver injury in mice. Sci. Rep. *5*, 13423.

62. Ogle, M.E., Olingy, C.E., Awojoodu, A.O., Das, A., Ortiz, R.A., Cheung, H.Y., and Botchwey, E.A. (2017). Sphingosine-1-phosphate receptor-3 supports hematopoietic stem and progenitor cell residence within the bone marrow niche. Stem Cells *35*, 1040–1052.

63. Bryce, S.A., Wilson, R.A.M., Tiplady, E.M., Asquith, D.L., Bromley, S.K., Luster, A.D., Graham, G.J., and Nibbs, R.J. (2016). Ackr4 on stromal cells scavenges ccl19 to enable ccr7-dependent trafficking of apcs from inflamed skin to lymph nodes. J. Immunol. *196*, 3341–3353.

64. Ito, S., D'Alessio, A.C., Taranova, O.V., Hong, K., Sowers, L.C., and Zhang, Y. (2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature *466*, 1129–1133.

65. Klug, M., Schmidhofer, S., Gebhard, C., Andreesen, R., and Rehli, M. (2013). 5-Hydroxymethylcytosine is an essential intermediate of active DNA demethylation processes in primary human monocytes. Genome Biol. *14*, R46.

66. Bick, A.G., Weinstock, J.S., Nandakumar, S.K., Fulco, C.P., Bao, E.L., Zekavat, S.M., Szeto, M.D., Liao, X., Leventhal, M.J., Nasser, J., et al.; NHLBI Trans-Omics for Precision Medicine Consortium (2020). Inherited causes of clonal haematopoiesis in 97,691 whole genomes. Nature *586*, 763–768.

67. Cargo, C., Cullen, M., Taylor, J., Short, M., Glover, P., Van Hoppe, S., Smith, A., Evans, P., and Crouch, S. (2019). The use of targeted sequencing and flow cytometry to identify patients with a clinically significant monocytosis. Blood *133*, 1325–1334.

68. Stremenova Spegarova, J., Lawless, D., Mohamad, S.M.B., Engelhardt, K.R., Doody, G., Shrimpton, J., Rensing-Ehl, A., Ehl, S., Rieux-Laucat, F., Cargo, C., et al. (2020). Germline TET2 loss of function causes childhood immunodeficiency and lymphoma. Blood *136*, 1055–1066.

69. Kaasinen, E., Kuismin, O., Rajamäki, K., Ristolainen, H., Aavikko, M., Kondelin, J., Saarinen, S., Berta, D.G., Katainen, R., Hirvonen, E.A.M., et al. (2019). Impact of constitutional TET2 haploinsufficiency on molecular and clinical phenotype in humans. Nat. Commun. *10*, 1252.

70. Duployez, N., Goursaud, L., Fenwarth, L., Bories, C., Marceau-Renaut, A., Boyer, T., Fournier, E., Nibourel, O., Roche-Lestienne, C., Huet, G., et al. (2020). Familial myeloid malignancies with germline TET2 mutation. Leukemia *34*, 1450–1453.

71. Kazi, J.U., and Rönnstrand, L. (2019). Fms-like tyrosine kinase 3/flt3: from basic science to clinical implications. Physiol. Rev. *99*, 1433–1466.

72. Saevarsdottir, S., Olafsdottir, T.A., Ivarsdottir, E.V., Halldorsson, G.H., Gunnarsdottir, K., Sigurdsson, A., Johannesson, A., Sigurdsson, J.K., Juliusdottir, T., Lund, S.H., et al. (2020). FLT3 stop mutation increases FLT3 ligand level and risk of autoimmune thyroid disease. Nature *584*, 619–623.

73. Flaquer, A., Rappold, G.A., Wienker, T.F., and Fischer, C. (2008). The human pseudoautosomal regions: a review for genetic epidemiologists. Eur. J. Hum. Genet. *16*, 771–779.

74. Miga, K.H., Koren, S., Rhie, A., Vollger, M.R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G.A., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. Nature *585*, 79–84.

75. Vuckovic, D., Bao, E.L., Akbari, P., Lareau, C.A., Mousas, A., Jiang, T., Chen, M.H., Raffield, L.M., Tardaguila, M., Huffman, J.E., et al.; VA Million Veteran Program (2020). The polygenic and monogenic basis of blood traits and diseases. Cell *182*, 1214–1231.e11.

76. Naik, R.P., Smith-Whitley, K., Hassell, K.L., Umeh, N.I., de Montalembert, M., Sahota, P., Haywood, C., Jr., Jenkins, J., Lloyd-Puryear, M.A., Joiner, C.H., et al. (2018). Clinical outcomes associated with sickle cell trait: a systematic review. Ann. Intern. Med. *169*, 619–627.

77. Hou, J., Chen, Q., Wu, X., Zhao, D., Reuveni, H., Licht, T., Xu, M., Hu, H., Hoeft, A., Ben-Sasson, S.A., et al. (2017). S1pr3 signaling drives bacterial killing and is required for survival in bacterial sepsis. Am. J. Respir. Crit. Care Med. *196*, 1559–1570.

78. Gon, Y., Wood, M.R., Kiosses, W.B., Jo, E., Sanna, M.G., Chun, J., and Rosen, H. (2009). Retraction for "S1P3 receptor-induced reorganization of epithelial tight junctions compromises lung barrier integrity and is potentiated by TNF". Proc. Natl. Acad. Sci. USA *106*, 12561.

79. Punsawad, C., and Viriyavejakul, P. (2019). Expression of sphingosine kinase 1 and sphingosine 1-phosphate receptor 3 in malaria-associated acute lung injury/acute respiratory distress syndrome in a mouse model. PLoS ONE *14*, e0222098.

80. Daya, M., and Barnes, K.C. (2019). African American ancestry contribution to asthma and atopic dermatitis. Ann. Allergy Asthma Immunol. *122*, 456–462.

81. Nakajima, S., Kabata, H., Kabashima, K., and Asano, K. (2020). Anti-TSLP antibodies: Targeting a master regulator of type 2 immune responses. Allergol. Int. *69*, 197–203.

82. Holstege, H., Hulsman, M., van der Lee, S.J., and van den Akker, E.B. (2020). The role of age-related clonal hematopoiesis in genetic sequencing studies. Am. J. Hum. Genet. *107*, 575–576.

83. Kraft, I.L., and Godley, L.A. (2020). Identifying potential germline variants from sequencing hematopoietic malignancies. Blood *136*, 2498–2506.

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

Burden plot of *TET2* aggregate gene test with MONO

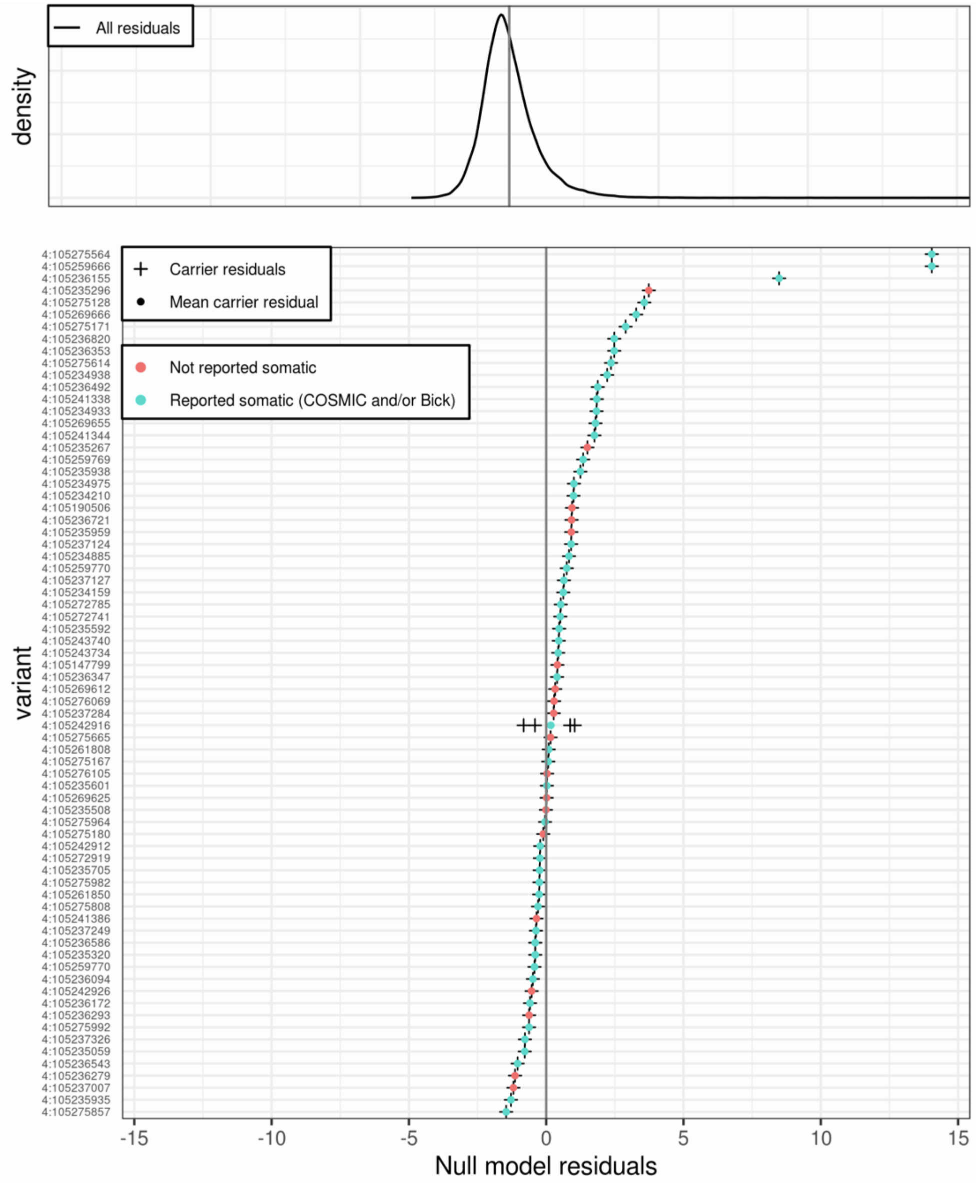

**Figure S1:** The residuals from the null model for 72 rare variants aggregated into a unit for testing the *TET2* gene region, colored by whether the variant was reported as a somatic mutation in the COSMIC database or in Bick, et al. For each carrier of the variant, a + indicates the null model residual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

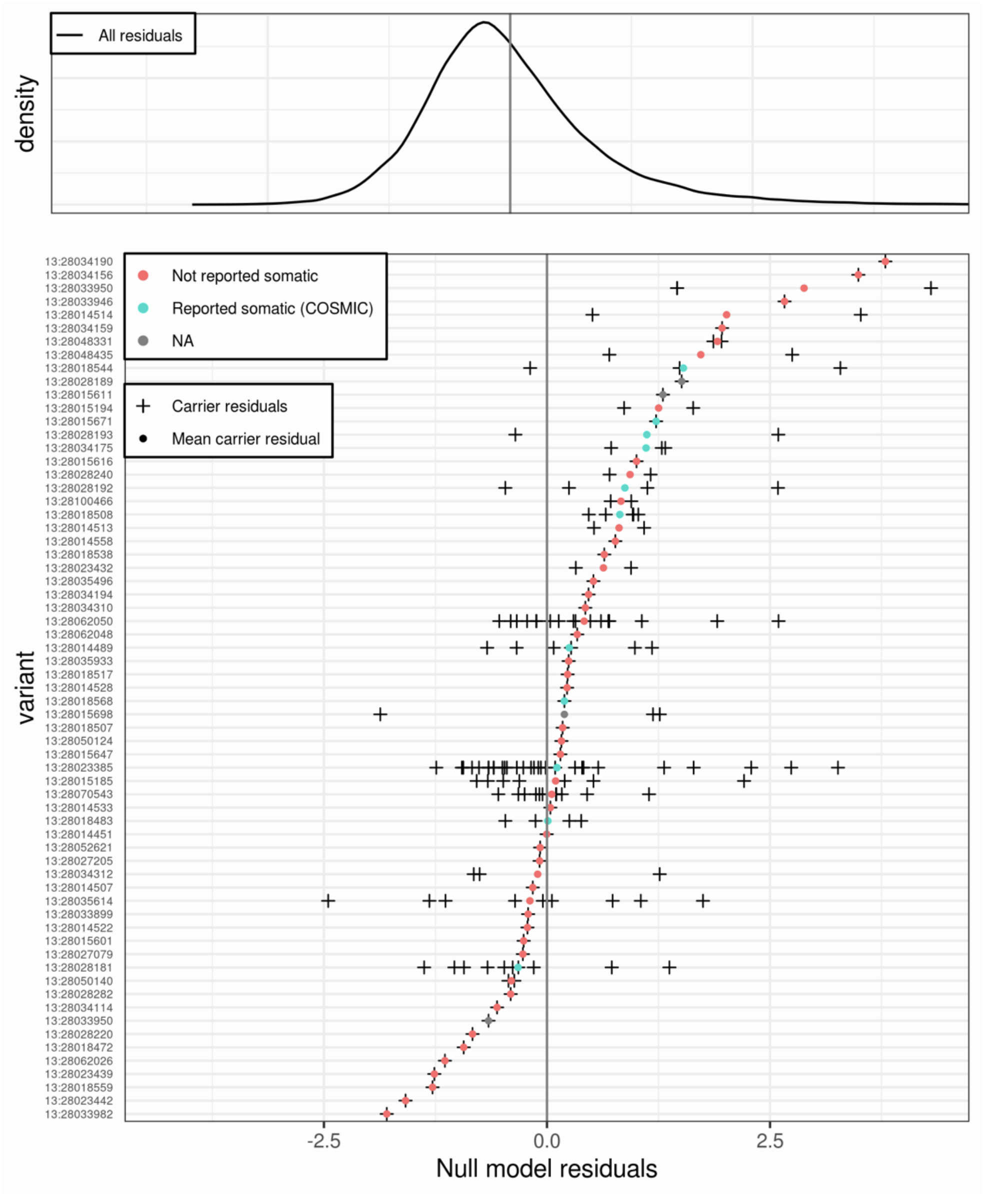

**Figure S2:** The residuals from the null model for 65 rare variants aggregated for testing the *FLT3* gene region, colored by whether the variant was reported as a somatic mutation in the COSMIC database. For each carrier of the variant, a + indicates the null model residual value and a circle indicates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

Manhattan plot of SNV results with BASO

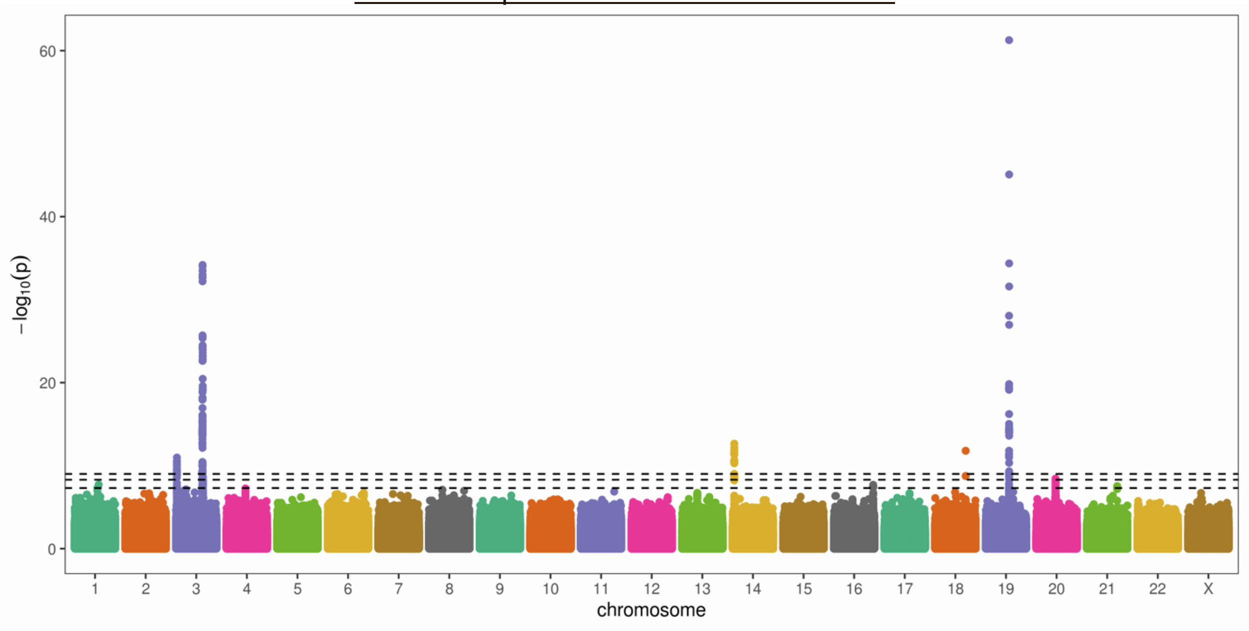

**Figure S3:** Manhattan plot of variants tested for association with binarized basophil count outcome.

QQplot of SNV results with BASO

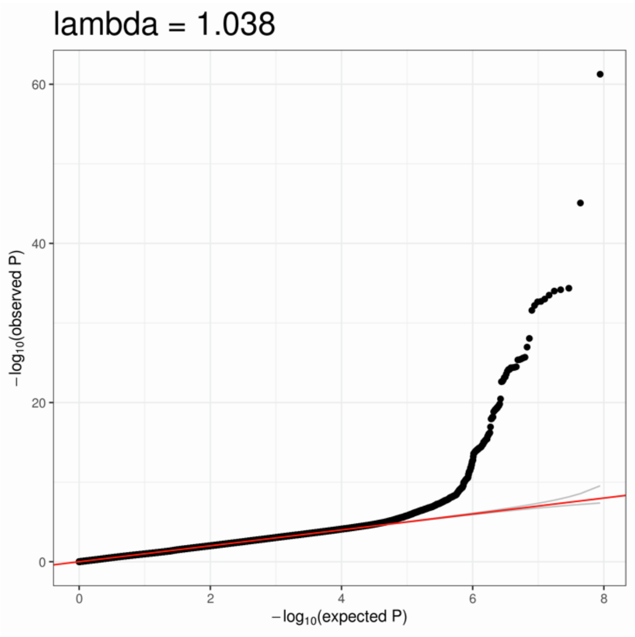

**Figure S4:** QQ plot of variants tested for association with the binarized basophil count outcome.

## Manhattan plot of SNV results with EOS

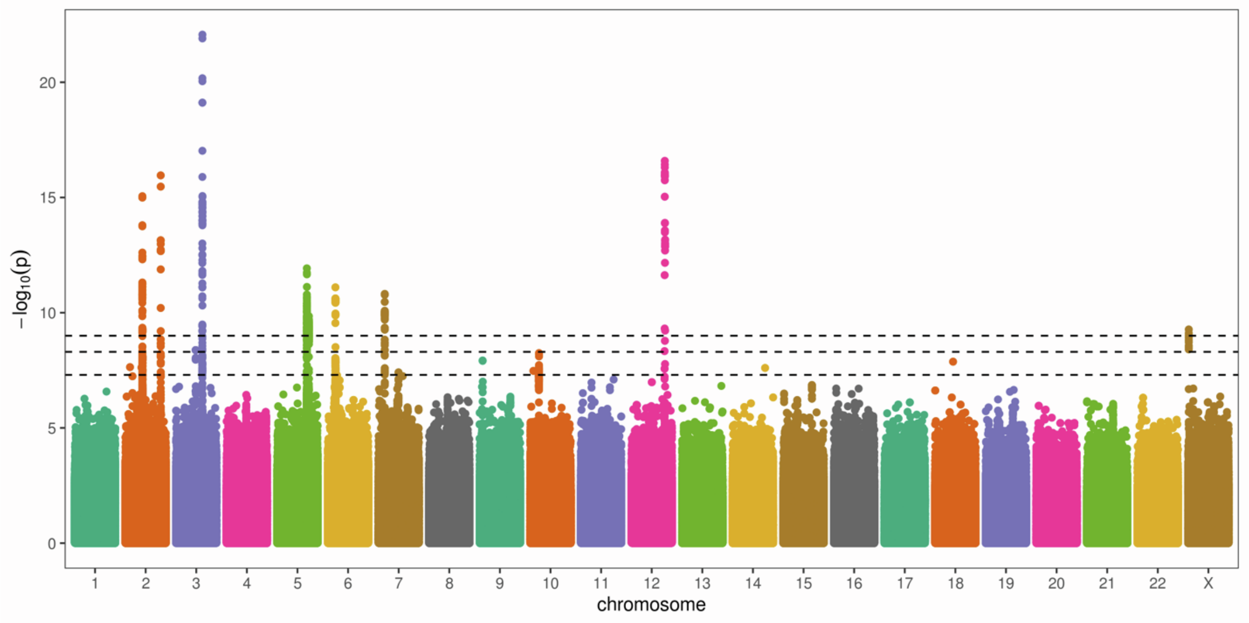

**Figure S5:** Manhattan plot of variants tested for association with the eosinophil percentage outcome.

## QQplot of SNV results with EOS

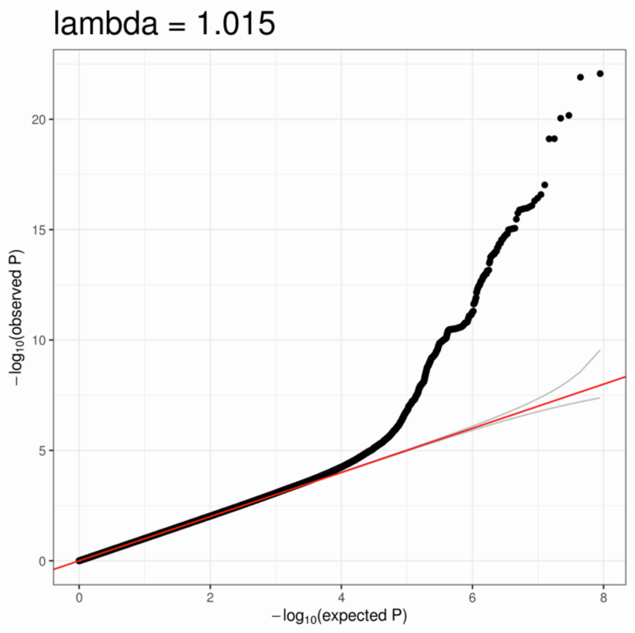

**Figure S6:** QQ plot of variants tested for association with the eosinophil percentage outcome.

Manhattan plot of SNV results with LYM

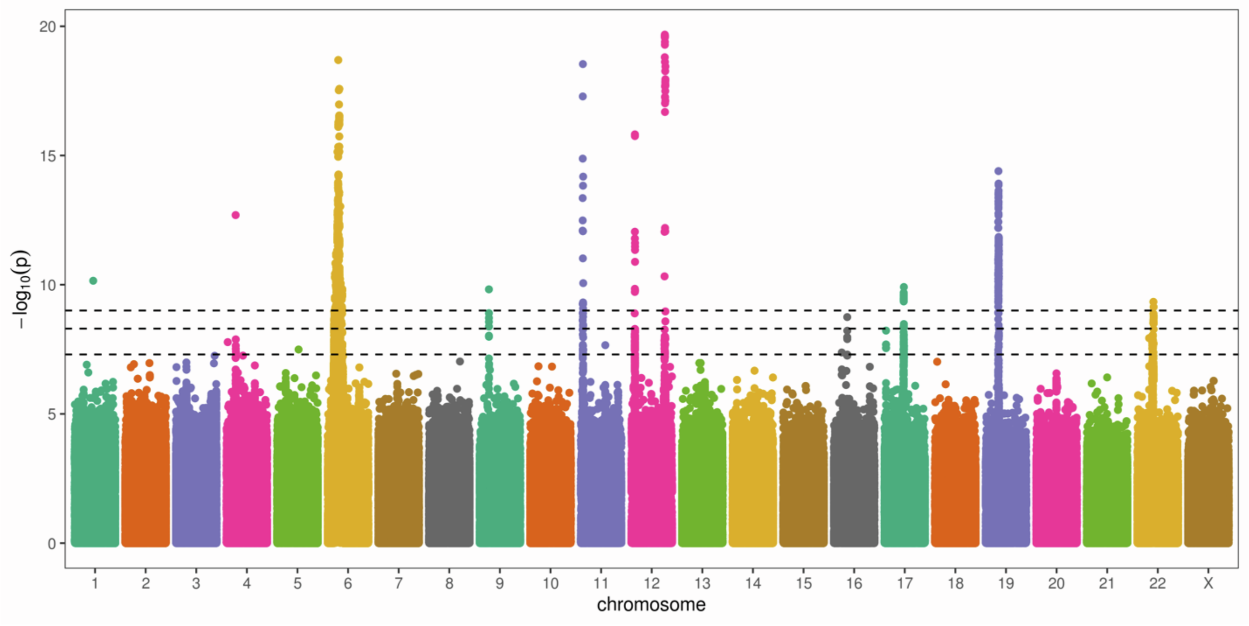

**Figure S7:** Manhattan plot of variants tested for association with the lymphocyte count outcome.

QQplot of SNV results with LYM

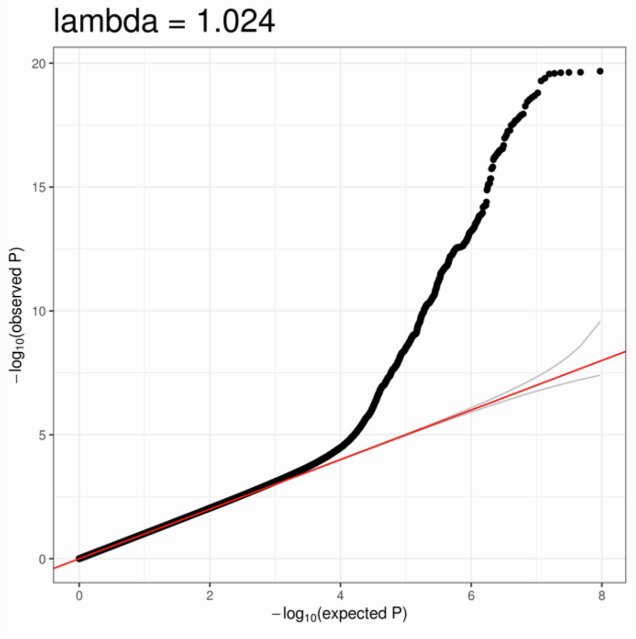

**Figure S8:** QQ plot of variants tested for association with the lymphocyte count outcome.

Manhattan plot of SNV results with MONO

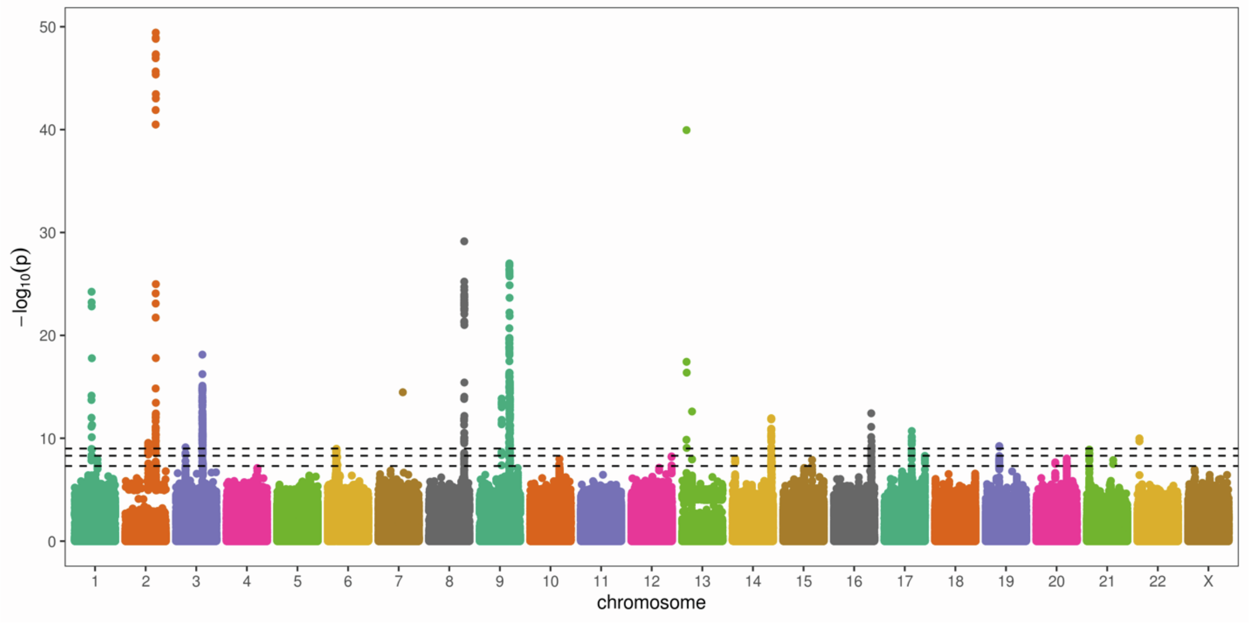

**Figure S9:** Manhattan plot of variants tested for association with the monocyte count outcome.

QQplot of SNV results with MONO

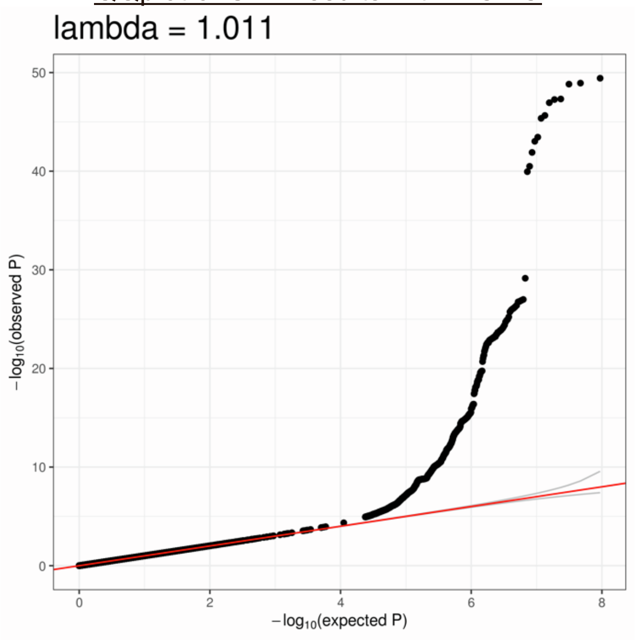

**Figure S10:** QQ plot of variants tested for association with the monocyte count outcome.

Manhattan plot of SNV results with NEU

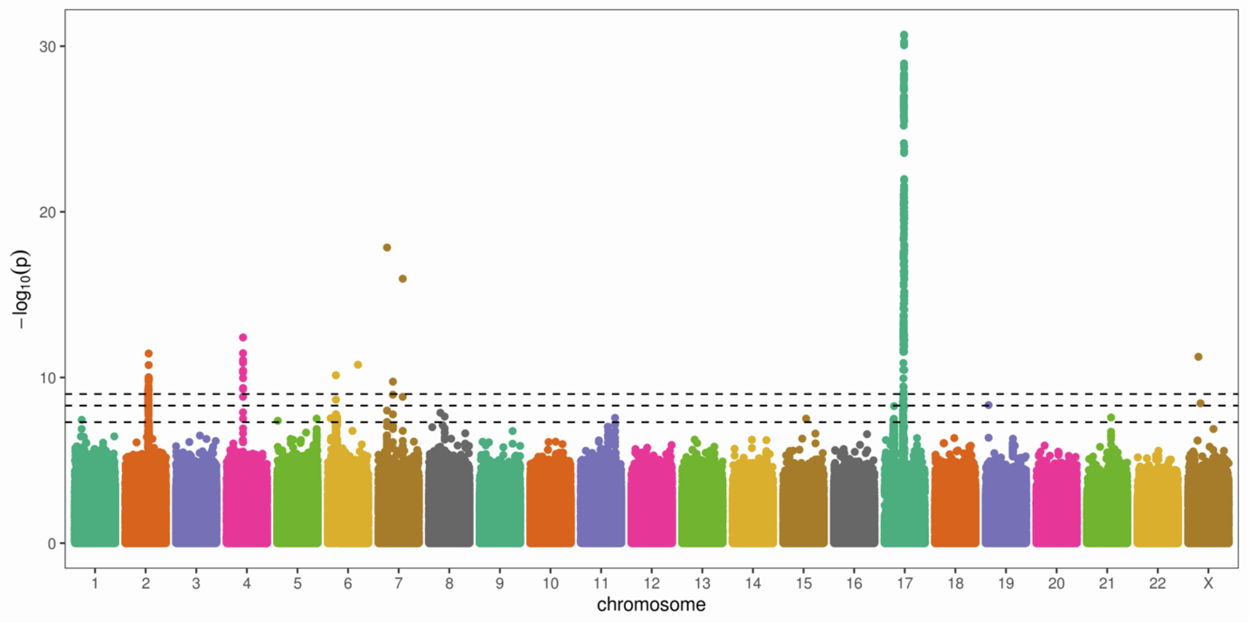

**Figure S11:** Manhattan plot of variants tested for association with the neutrophil count outcome.

QQplot of SNV results with NEU

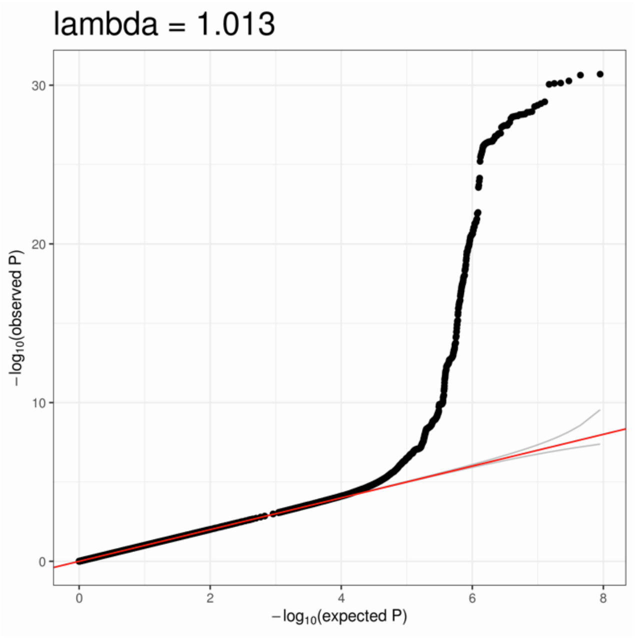

**Figure S12:** QQ plot of variants tested for association with the neutrophil count outcome.

Manhattan plot of SNV results with WBC

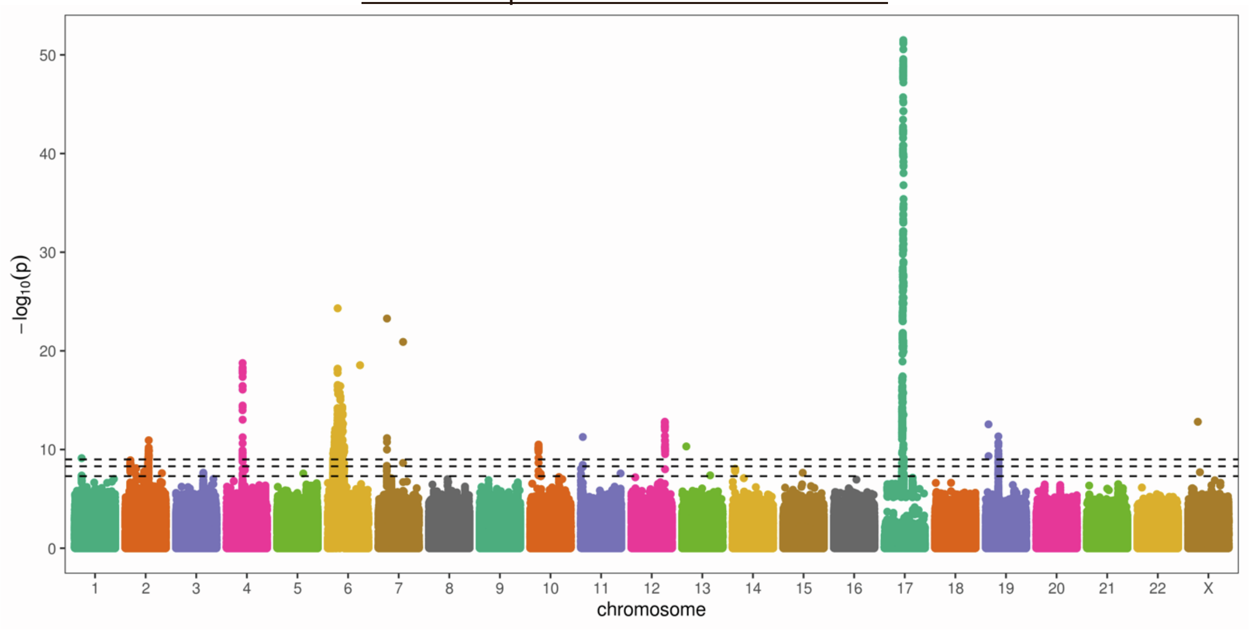

**Figure S13:** Manhattan plot of variants tested for association with the white blood cell count outcome.

QQplot of SNV results with WBC

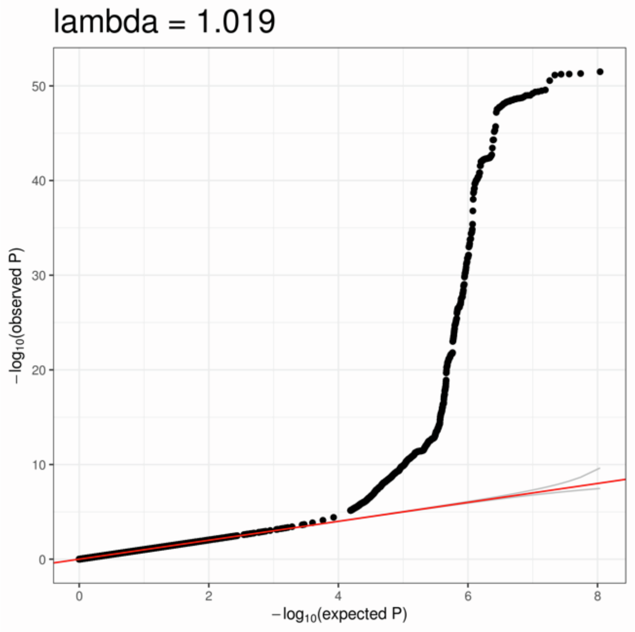

**Figure S14:** QQ plot of variants tested for association with the white blood cell count outcome.

## 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.