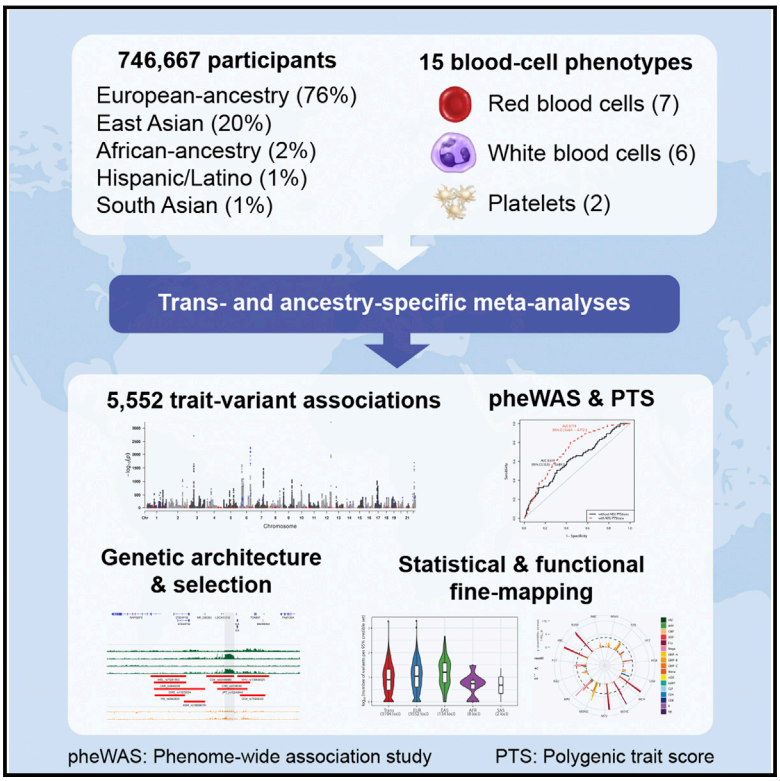
Cell

Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations

Graphical Abstract



Highlights

- Blood cell traits differ by ancestry and are subject to selective pressure
- We assessed 15 blood cell traits in 746,667 participants from 5 global populations
- We identified more than 5,500 associations, including ~100 associations not found in Europeans
- These analyses improved risk prediction and identified potential causal variants

Authors

Ming-Huei Chen, Laura M. Raffield, Abdou Mousas, ..., Alexander P. Reiner, Paul L. Auer, Guillaume Lettre

Correspondence

pauer@uwm.edu (P.L.A.), guillaume.lettre@umontreal.ca (G.L.)

In Brief

Delineation of the genetic architecture of hematological traits in a multi-ethnic dataset allows identification of rare variants with strong effects specific to non-European populations and improved fine mapping of GWAS variants using the trans-ethnic approach.

Chen et al., 2020, Cell *182*, 1198–1213
September 3, 2020
https://doi.org/10.1016/j.cell.2020.06.045

CellPress

CellPress

Cell

Article

Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations

Ming-Huei Chen,[1,2,100] Laura M. Raffield,[3,100] Abdou Mousas,[4,100] Saori Sakaue,[5,6] Jennifer E. Huffman,[7] Arden Moscati,[8] Bhavi Trivedi,[9] Tao Jiang,[10] Parsa Akbari,[10,11,12,13] Dragana Vuckovic,[12] Erik L. Bao,[14,15] Xue Zhong,[16] Regina Manansala,[17] Véronique Laplante,[18] Minhui Chen,[19] Ken Sin Lo,[4] Huijun Qian,[20] Caleb A. Lareau,[14,15] Mélissa Beaudoin,[4] Karen A. Hunt,[9] Masato Akiyama,[6,21] Traci M. Bartz,[22] Yoav Ben-Shlomo,[23] Andrew Beswick,[24] Jette Bork-Jensen,[25] Erwin P. Bottinger,[8,26] Jennifer A. Brody,[27] Frank J.A. van Rooij,[28] Kumaraswamynaidu Chitrala,[29] Kelly Cho,[7,30,31] Hélène Choquet,[32] Adolfo Correa,[33] John Danesh,[10,12,13,34,35,36] Emanuele Di Angelantonio,[10,13,34,35,36] Niki Dimou,[37,38] Jingzhong Ding,[39] Paul Elliott,[40,41,42,43,44] Tõnu Esko,[15] Michele K. Evans,[29] James S. Floyd,[27,45] Linda Broer,[46] Niels Grarup,[25] Michael H. Guo,[15,47] Andreas Greinacher,[48] Jeff Haessler,[49] Torben Hansen,[25] Joanna M.M. Howson,[10,50] Qin Qin Huang,[12] Wei Huang,[51] Eric Jorgenson,[32] Tim Kacprowski,[52,53,54]

(Author list continued on next page)

[1]The Framingham Heart Study, National Heart, Lung and Blood Institute, Framingham, MA 01702, USA
[2]Population Sciences Branch, Division of Intramural Research, National Heart, Lung and Blood Institute, Framingham, MA 01702, USA
[3]Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA
[4]Montreal Heart Institute, Montreal, QC H1T 1C8, Canada
[5]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Osaka 565-0871, Japan
[6]Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045, Japan
[7]Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, MA 02130, USA
[8]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[9]Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK
[10]BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK
[11]MRC Biostatistics Unit, University of Cambridge, Cambridge CB2 0SR, UK
[12]Human Genetics, Wellcome Sanger Institute, Hinxton CB10 1SA, UK
[13]The National Institute for Health Research Blood and Transplant Unit (NIHR BTRU) in Donor Health and Genomics, University of Cambridge, Cambridge CB2 0QQ, UK
[14]Division of Hematology/Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA
[15]Broad Institute of Harvard and MIT, Cambridge, MA 02446, USA
[16]Division of Genetic Medicine, Department of Medicine, Vanderbilt University, Nashville, TN 37232, USA
[17]Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI 53201, USA

(Affiliations continued on next page)

SUMMARY

Most loci identified by GWASs have been found in populations of European ancestry (EUR). In trans-ethnic meta-analyses for 15 hematological traits in 746,667 participants, including 184,535 non-EUR individuals, we identified 5,552 trait-variant associations at $p < 5 \times 10^{-9}$, including 71 novel associations not found in EUR populations. We also identified 28 additional novel variants in ancestry-specific, non-EUR meta-analyses, including an *IL7* missense variant in South Asians associated with lymphocyte count *in vivo* and IL-7 secretion levels *in vitro*. Fine-mapping prioritized variants annotated as functional and generated 95% credible sets that were 30% smaller when using the trans-ethnic as opposed to the EUR-only results. We explored the clinical significance and predictive value of trans-ethnic variants in multiple populations and compared genetic architecture and the effect of natural selection on these blood phenotypes between populations. Altogether, our results for hematological traits highlight the value of a more global representation of populations in genetic studies.

INTRODUCTION

Blood cell counts and indices are quantitative clinical laboratory measures that reflect hematopoietic progenitor cell production, hemoglobin synthesis, maturation, release from the bone marrow, and clearance of mature or senescent blood cells from the circulation. Quantitative red blood cell (RBC), white blood cell (WBC), and platelet (PLT) traits exhibit strong heritability (h^2, ~30%–80%) (Evans et al., 1999; Hinckley et al., 2013) and have been the subject of various genome-wide association

Mika Kähönen,[55,56] Yoichiro Kamatani,[6,57] Masahiro Kanai,[6,58] Savita Karthikeyan,[10] Fotis Koskeridis,[38] Leslie A. Lange,[59] Terho Lehtimäki,[60,61] Markus M. Lerch,[62] Allan Linneberg,[63,64] Yongmei Liu,[65] Leo-Pekka Lyytikäinen,[60,61] Ani Manichaikul,[66] Hilary C. Martin,[12] Koichi Matsuda,[67] Karen L. Mohlke,[3] Nina Mononen,[60,61] Yoshinori Murakami,[68] Girish N. Nadkarni,[8] Matthias Nauck,[54,69] Kjell Nikus,[70,71] Willem H. Ouwehand,[12,36,72,73] Nathan Pankratz,[74] Oluf Pedersen,[25] Michael Preuss,[8] Bruce M. Psaty,[27,75,76,77] Olli T. Raitakari,[78,79,80] David J. Roberts,[13,81,82,83] Stephen S. Rich,[66] Benjamin A.T. Rodriguez,[1,2] Jonathan D. Rosen,[84] Jerome I. Rotter,[85] Petra Schubert,[7]

(Author list continued on next page)

[18]Département de Pharmacologie et Physiologie, Faculty of Medicine, Université de Montréal, Montreal, QC H3T 1J4, Canada
[19]Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA
[20]Department of Statistics and Operation Research, University of North Carolina, Chapel Hill, NC 27599, USA
[21]Department of Ocular Pathology and Imaging Science, Graduate School of Medical Sciences, Kyushu University, Fukuoka 812-8581, Japan
[22]Department of Biostatistics, University of Washington, Seattle, WA 98101, USA
[23]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol BS8 2PS, UK
[24]Translational Health Sciences, Musculoskeletal Research Unit, Bristol Medical School, University of Bristol, Bristol BS10 5NB, UK
[25]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2200, Denmark
[26]Hasso-Plattner-Institut, Universität Potsdam, Potsdam 14469, Germany
[27]Department of Medicine, University of Washington, Seattle, WA 98101, USA
[28]Department of Epidemiology, Erasmus University Medical Center Rotterdam, Rotterdam 3015, the Netherlands
[29]Laboratory of Epidemiology and Population Science, National Institute on Aging/NIH, Baltimore, MD 21224, USA
[30]Department of Medicine, Division on Aging, Brigham and Women's Hospital, Boston, MA 02115, USA
[31]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA
[32]Division of Research, Kaiser Permanente Northern California, Oakland, CA 94612, USA
[33]Department of Medicine, University of Mississippi Medical Center, Jackson, MS 39216, USA
[34]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge CB10 1SA, UK
[35]National Institute for Health Research Cambridge Biomedical Research Centre, Cambridge University Hospitals, Cambridge CB2 0QQ, UK
[36]British Heart Foundation Centre of Research Excellence, Division of Cardiovascular Medicine, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK
[37]Section of Nutrition and Metabolism, International Agency for Research on Cancer, Lyon 69008, France
[38]Department of Hygiene and Epidemiology, University of Ioannina Medical School, Ioannina 45110, Greece
[39]Department of Internal Medicine, Section of Gerontology and Geriatric Medicine, Wake Forest School of Medicine, Winston-Salem, NC 27101, USA
[40]Department of Epidemiology and Biostatistics, Imperial College London, London W2 1PG, UK
[41]Imperial Biomedical Research Centre, Imperial College London and Imperial College NHS Healthcare Trust, London SW7 2AZ, UK
[42]Medical Research Council-Public Health England Centre for Environment, Imperial College London, London SW7 2AZ, UK
[43]UK Dementia Research Institute, Imperial College London, London SW7 2AZ, UK
[44]Health Data Research UK – London, London SW7 2AZ, UK
[45]Department of Epidemiology, University of Washington, Seattle, WA 98101, USA
[46]Department of Internal Medicine, Erasmus University Medical Center Rotterdam, Rotterdam 3015, the Netherlands
[47]Department of Neurology, University of Pennsylvania, Philadelphia, PA 19104, USA
[48]Institute for Immunology and Transfusion Medicine, University Medicine Greifswald, Greifswald 17475, Germany
[49]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
[50]Novo Nordisk Research Centre Oxford, Innovation Building, Old Road Campus, Oxford OX3 7FZ, UK
[51]Department of Genetics, Shanghai-MOST Key Laboratory of Health and Disease Genomics, Chinese National Human Genome Center and Shanghai Industrial Technology Institute (SITI), Shanghai 201203, China
[52]Interfaculty Institute of Genetics and Functional Genomics, University Medicine Greifswald, Greifswald 17475, Germany
[53]Chair of Experimental Bioinformatics, Research Group Computational Systems Medicine, Technical University of Munich, Freising-Weihenstephan 85354, Germany
[54]German Center for Cardiovascular Research (DZHK), Partner Site Greifswald, Greifswald 17475, Germany
[55]Department of Clinical Physiology, Tampere University Hospital, Tampere 33521, Finland
[56]Department of Clinical Physiology, Finnish Cardiovascular Research Center – Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere 33014, Finland
[57]Laboratory of Complex Trait Genomics, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo 108-8639, Japan
[58]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA
[59]Department of Medicine, University of Colorado Denver, Anschutz Medical Campus, Aurora, CO 80045, USA

(Affiliations continued on next page)

studies (GWASs), including a large study that identified more than 1,000 genomic loci in ~150,000 individuals of European ancestry (EUR) (Astle et al., 2016).

Importantly, the distribution of hematologic traits and prevalence of inherited hematologic conditions differ by ethnicity. For example, the prevalence of anemia and microcytosis is

Cassandra N. Spracklen,[3,86] Praveen Surendran,[10,87] Hua Tang,[88] Jean-Claude Tardif,[4,89] Richard C. Trembath,[90] Mohsen Ghanbari,[28,91] Uwe Völker,[52,54] Henry Völzke,[54,92] Nicholas A. Watkins,[73] Alan B. Zonderman,[29] VA Million Veteran Program, Peter W.F. Wilson,[93] Yun Li,[3,84,94] Adam S. Butterworth,[10,13,34] Jean-François Gauchat,[18] Charleston W.K. Chiang,[19,95] Bingshan Li,[96] Ruth J.F. Loos,[8] William J. Astle,[11,13,97] Evangelos Evangelou,[38,40] David A. van Heel,[9] Vijay G. Sankaran,[14,15] Yukinori Okada,[5,98] Nicole Soranzo,[12,72] Andrew D. Johnson,[1,2,101] Alexander P. Reiner,[99,101] Paul L. Auer,[17,101,*] and Guillaume Lettre[4,89,101,102,*]

[60]Department of Clinical Chemistry, Fimlab Laboratories, Tampere 33520, Finland
[61]Department of Clinical Chemistry, Finnish Cardiovascular Research Center – Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere 33014, Finland
[62]Department of Internal Medicine, University Medicine Greifswald, Greifswald 17475, Germany
[63]Center for Clinical Research and Prevention, Bispebjerg and Frederiksberg Hospital, Frederiksberg 2000, Denmark
[64]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2200, Denmark
[65]Duke Molecular Physiology Institute, Department of Medicine, Division of Cardiology, Duke University Medical Center, Durham, NC 27701, USA
[66]Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22903, USA
[67]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo 108-8639, Japan
[68]Division of Molecular Pathology, The Institute of Medical Sciences, The University of Tokyo, Tokyo 108-8639, Japan
[69]Institue of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald 17475, Germany
[70]Department of Cardiology, Heart Center, Tampere University Hospital, Tampere 33521, Finland
[71]Department of Cardiology, Finnish Cardiovascular Research Center – Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere 33014, Finland
[72]Department of Hematology, University of Cambridge, Cambridge CB2 0PT, UK
[73]National Health Service (NHS) Blood and Transplant, Cambridge Biomedical Campus, Cambridge CB2 0PT, UK
[74]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN 55455, USA
[75]Department of Epidemiology, University of Washington, Seattle, WA 98101, USA
[76]Department of Health Services, University of Washington, Seattle, WA 98101, USA
[77]Kaiser Permanente Washington Health Research Institute, Seattle, WA 98101, USA
[78]Centre for Population Health Research, University of Turku and Turku University Hospital, Turku 20521, Finland
[79]Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku 20521, Finland
[80]Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku 20521, Finland
[81]Radcliffe Department of Medicine, John Radcliffe Hospital, University of Oxford, Oxford OX3 9DU, UK
[82]Department of Haematology, Churchill Hospital, Oxford OX3 7LE, UK
[83]NHS Blood and Transplant – Oxford Centre, John Radcliffe Hospital, Oxford OX3 9DU, UK
[84]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA
[85]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation (formerly Los Angeles Biomedical Research Institute) at Harbor-UCLA Medical Center, Torrance, CA 90502, USA
[86]Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA 01002, USA
[87]Rutherford Fund Fellow, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK
[88]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA
[89]Department of Medicine, Faculty of Medicine, Université de Montréal, Montreal, QC H3T 1J4, Canada
[90]School of Basic and Medical Biosciences, Faculty of Life Sciences and Medicine, King's College London, London SE1 1UL, UK
[91]Department of Genetics, School of Medicine, Mashhad University of Medical Sciences, Mashhad 9177948564, Iran
[92]Institute for Community Medicine, University Medicine Greifswald, Greifswald 17475, Germany
[93]Atlanta VA Medical Center, Decatur, GA 30033, USA
[94]Department of Computer Science, University of North Carolina, Chapel Hill, NC 27599, USA
[95]Quantitative and Computational Biology Section, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA
[96]Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN 37232, USA
[97]NIHR Blood and Transplant Research Unit in Donor Health and Genomics, Strangeways Laboratory, Cambridge CB1 8RN, UK
[98]Laboratory of Statistical Immunology, Osaka University Graduate School of Medicine, Suita, Osaka 565-0871, Japan
[99]Department of Epidemiology, University of Washington, Seattle, WA 98109, USA
[100]These authors contributed equally
[101]Senior author
[102]Lead Contact
*Correspondence: pauer@uwm.edu (P.L.A.), guillaume.lettre@umontreal.ca (G.L.)
https://doi.org/10.1016/j.cell.2020.06.045

higher among African ancestry (AFR) individuals compared with EUR individuals, in part because of the presence of globin gene mutations (e.g., sickle cell, α/β-thalassemia) more common among African, Mediterranean, and Asian populations (Beutler and West, 2005; Raffield et al., 2018; Rana et al., 1993). AFR individuals tend to have lower WBC and neutrophil counts partly because of the Duffy/*DARC* null variant (Rappoport et al., 2019). Among Hispanic Americans (HA), a common Native

American functional intronic variant of *ACTN1* is associated with a lower PLT count (Schick et al., 2016).

Despite these observations, non-EUR populations have been severely under-represented in most blood cell genetic studies to date (Popejoy and Fullerton, 2016; Popejoy et al., 2018; Wojcik et al., 2019). Multiethnic GWASs have been recognized as more powerful for gene mapping because of ancestry-specific differences in allele frequency, linkage disequilibrium (LD), and effect size of causal variants (Li and Keating, 2014). Because blood cells play a key role in pathogen invasion, defense and inflammatory responses, hematology-associated genetic loci are particularly predisposed to be differentiated across ancestral populations as a result of population history and local evolutionary selective pressure (Ding et al., 2013; Lo et al., 2011; Raj et al., 2013). Given the essential role of blood cells in tissue oxygen delivery, inflammatory responses, atherosclerosis, and thrombosis (Byrnes and Wolberg, 2017; Chu et al., 2010; Colin et al., 2014; Tajuddin et al., 2016), factors that contribute to such inter-population differences in blood cell traits may also play appreciable roles in the pathogenesis of chronic diseases and health disparities between populations.

RESULTS

Trans-ethnic and Ancestry-Specific Blood Cell Trait Genetic Associations

We analyzed genotype-phenotype associations at up to 45 million autosomal variants in 746,667 participants, including 184,424 non-EUR individuals, for 15 traits (Figure 1; Figure S1; Tables S1A–S1D and S2; STAR Methods). The association results of the EUR-specific meta-analyses are reported separately in a companion paper (Vuckovic et al., 2020 [this issue of *Cell*]). In the trans-ethnic meta-analyses, we identified 5,552 trait-variant associations at $p < 5 \times 10^{-9}$, including 71 novel associations not reported in the EUR-specific manuscript (Table S3A). Of the 5,552 trans-ethnic associations, 128 showed strong evidence of allelic effect heterogeneity across populations ($P_{ancestry.hetero} < 5 \times 10^{-9}$) (Table S3A). Ancestry-specific meta-analyses revealed 28 additional novel trait-variant associations (Figure 1B; Tables S3B–S3F). However, 21 of these 28 novel associations were identified in AFR participants, and 19 of these 21 novel AFR-specific associations map to chromosome 1 and are associated with WBC or neutrophil counts, reflecting long-range associations because of the admixture signal at the Duffy/*DARC* locus that confers resistance to *Plasmodium vivax* infection (Reich et al., 2009). We attempted to replicate all novel trans-ethnic or ancestry-specific genetic associations in the Million Veteran Program (MVP) cohort (Table S1E; Gaziano et al., 2016). Of the 88 variant-trait associations we could test in the MVP, 85 had a consistent direction of effect (binomial $p = 6 \times 10^{-24}$), 83 were confirmed with a false discovery rate of less than 5%, and 44 met the Bonferroni-adjusted significance threshold of $p < 6 \times 10^{-4}$ (Table S3G).

For 3,552 loci in which conditional analyses identified a single genome-wide significant variant in EUR, we generated fine-mapping results for each trans-ethnic or ancestry-specific dataset using an approximate Bayesian approach (STAR Methods; Maller et al., 2012). The 95% credible sets were smaller in the trans-ethnic meta-analyses than in the EUR or East Asian ancestry (EAS) meta-analyses (Figure 2A). When comparing loci discovered in the trans-ethnic and EUR analyses, we found that the 95% credible sets were 30% smaller among the trans-ethnic results; the median (interquartile range) number of variants per 95% credible set was 4 (2–13) in trans-ethnic versus 5 (2–16) in EUR (Wilcoxon's $p = 3 \times 10^{-4}$). For instance, a locus on chromosome 9 associated with PLT count included seven variants in the EUR 95% credible set but only one in the trans-ethnic set, an increase in fine-mapping resolution likely driven by limited LD at the locus in EAS (Figure 2B). In the trans-ethnic and EUR results, respectively, we identified 433 and 403 loci with a single variant in the 95% credible sets (Figure 2C) and more than 300 variants with a posterior inclusion probability (PIP) of 0.99 or more (Figure 2D). To determine the reason for the improved resolution in the trans-ethnic results, we sub-sampled the data and re-ran the EUR-only (N = 141,636), EAS-only (N = 143,085), and trans-ethnic (N = 137,702) meta-analyses on similarly sized sample sets for PLT, RBC, WBC, and hemoglobin (HGB). The resulting 95% credible sets were still smaller in the trans-ethnic meta-analyses, suggesting that the improved resolution was due to LD structure rather than an increase in sample size (Figure S2).

Next we assessed our fine-mapped 95% credible sets for the presence of functional variants, which we defined as variants with coding consequences or those mapping to accessible chromatin in hematopoietic cells. Genomic annotation of the 95% credible sets of the trans-ethnic, EUR, and EAS hematological trait-associated loci revealed that the proportion of likely functional variants was higher among those with high PIP (Figure 3A). The enrichment within high-PIP categories was particularly notable for missense variants but also observed for intronic and intergenic variants that map to open chromatin regions in progenitor or mature blood cells (Figure 3A; Corces et al., 2016). We used gchromVAR to quantify the enrichment of trans-ethnic, EUR, and EAS 95% credible set variants within regions of accessible chromatin identified by the assay for transpose accessible chromatin by sequencing (ATAC-seq) in 18 hematopoietic populations (Ulirsch et al., 2019). We noted 22 significant trait-cell type enrichments using the trans-ethnic credible sets, all of which were lineage specific, including RBC traits in erythroid progenitors, PLT traits in megakaryocytes, and monocyte counts in granulocyte-macrophage progenitors (GMPs) (Figure 3B; Table S3H). Cell-type enrichments were largely consistent between fine-mapped variants found in the trans-ethnic, EUR, and EAS loci. However, we observed two noteworthy ancestry-specific differences: the EAS results revealed significant enrichments (defined as Bonferroni-corrected threshold $p < 1.9 \times 10^{-4}$) in basophil counts for common myeloid progenitors (CMPs) ($P_{EAS\text{-}BASO\text{-}CMP} = 7.6 \times 10^{-5}$) and eosinophil counts for GMPs ($P_{EAS\text{-}EOS\text{-}GMP} = 4.5 \times 10^{-6}$), but neither pairing reached significance in the larger EUR meta-analyses ($P_{EUR\text{-}BASO\text{-}CMP} = 0.08$, $P_{EUR\text{-}EOS\text{-}GMP} = 0.01$) (Figure S3). These differences persisted even after controlling for the number of loci tested in each ancestry.

Among the novel loci identified in the trans-ethnic meta-analyses, several included excellent candidate causal variants with high fine-mapping PIP and overlap with open chromatin regions found in hematopoietic cells (Figure S4; Table S3I). For instance, rs115906455, located in an intron of the RNA polymerase II

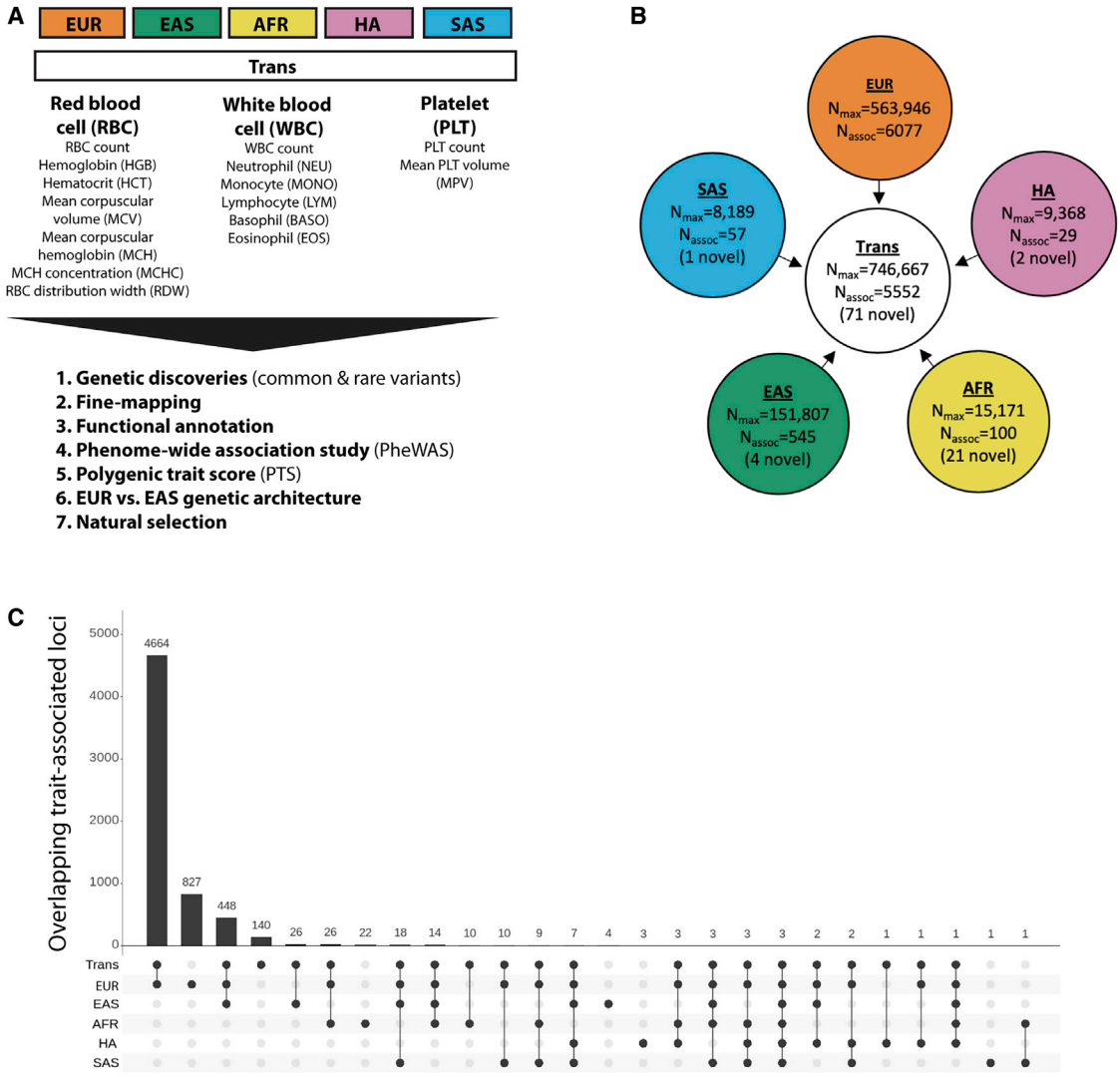


Figure 1. Trans-ethnic and Ancestry-Specific Meta-analyses of Blood Cell Traits

(A) List of blood cell phenotypes and analyses that were carried out in this project. Note that red blood cell distribution width (RDW) and mean platelet volume (MPV) were not available in EAS.

(B) Study design of the project. We used a fixed-effect meta-analysis strategy to analyze genetic associations within each of the five populations available and a mega-regression approach that considers allele frequency heterogeneity for the trans-ethnic association tests. N_{max}, maximum sample size in each meta-analysis; N_{assoc}, number of trait-variant associations. A locus is defined as novel when the 500-kb region surrounding its sentinel variant does not physically overlap with previously identified blood cell trait-associated variants (for any trait) in the corresponding population.

(C) Most blood-cell trait-associated loci physically overlap between populations. For this analysis, a locus associated with several blood cell traits was counted only once. Despite different sample sizes between populations, we note that few loci are found in a single population, suggesting shared genetic architecture. EUR, European ancestry; EAS, East Asian; AFR, African ancestry; HA, Hispanic American; SAS, South Asian.

See also Figure S1 and Tables S1A–S1D, S2, and S3A–S3F.

elongation factor *ELL2*, is strongly associated with mean corpuscular volume (MCV) ($p = 4.2 \times 10^{-12}$, PIP = 0.57) and maps to an accessible chromatin region found in RBC progenitors (CMPs, megakaryocyte-erythroid progenitors, and erythroblasts) but not megakaryocytes (Figure 3C). This variant is common in AFR populations (minor allele frequency [MAF] = 4.7%) but rare or monomorphic in non-AFR populations. A different variant at the *ELL2* locus has been associated previously with multiple myeloma and immunoglobulin G (IgG) levels (Swaminathan et al., 2015). Another example is rs941616, a common variant in an intergenic region on chromosome 14 that is associated with eosinophil count ($p = 2.4 \times 10^{-9}$, PIP = 0.2) and maps to a region of chromatin accessibility in CMPs, CD8+ lymphocytes, and natural killer cells (Figure 3D). This variant, which is in LD with another eosinophil-associated variant identified recently (Kichaev et al., 2019), is an expression quantitative trait locus (eQTL) for *PTGDR* (Võsa et al., 2018), which encodes the prostaglandin D2 receptor. Prostaglandins can activate eosinophils, which, in turn, contributes to the etiology of asthma, chronic

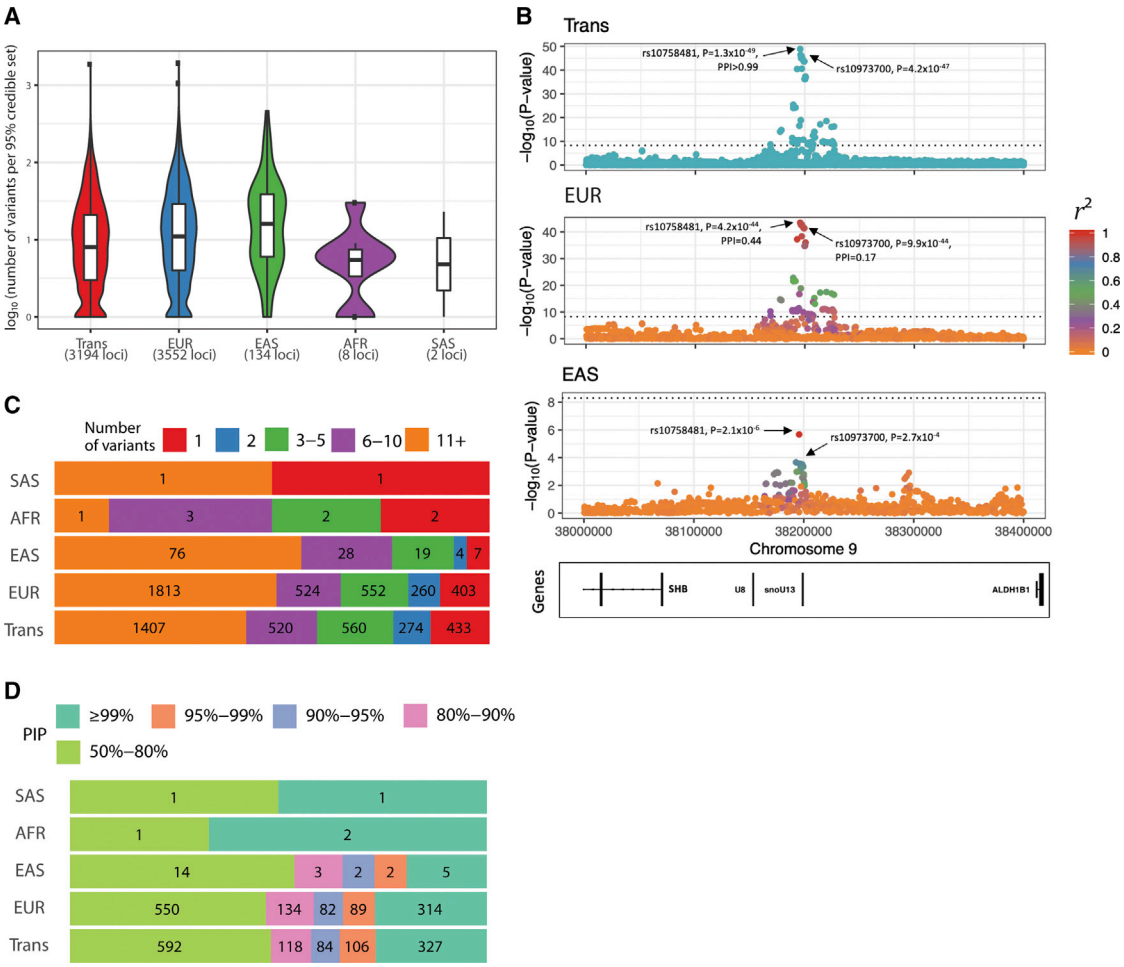


Figure 2. Fine Mapping of Genome-wide Significant Loci Associated with Hematological Traits

(A) We restricted fine mapping to loci with evidence of a single association signal in EUR populations. There are no such loci in HAs. The 95% credible sets in the trans-ethnic meta-analyses are smaller than in the EUR or EAS meta-analyses.

(B) Trans-ethnic fine mapping of a PLT locus. In EUR individuals, the 95% credible set include seven variants with a posterior inclusion probability (PIP) of more than 0.04 and a strong pairwise linkage disequilibrium (LD) with the sentinel variant rs10758481 ($r^2 > 0.93$ in British in England and Scotland [GBR] individuals from 1000G, center panel). The LD is similarly strong in AFR-, HA and South American, and South Asian ancestry populations from 1000G. However, the LD is weaker in EAS ($r^2 = 0.68$ in Japanese individuals [JPT] from 1000G, bottom panel). In the trans-ethnic meta-analysis, rs10758481 has a PIP of more than 0.99 (top panel). In EUR and EAS, the LD is color coded based on pairwise r^2 with rs10758481. The dotted line indicates the genome-wide significance threshold ($p < 5 \times 10^{-9}$).

(C) Proportion of 95% credible sets in each population with a defined number of variants. For instance, in the EUR and trans-ethnic meta-analysis results, we identified 403 and 433 95% credible sets that contain a single variant, respectively.

(D) Prioritization of causal variants using fine-mapping PIP. In each population, we provide the proportion of variants with a PIP within a specified range. For instance, in EUR and trans-ethnic populations, we found 314 and 327 variants with a PIP of 99% or more, respectively.

See also Figure S2.

obstructive pulmonary disease (COPD), and allergies (Brusselle et al., 2016). In the UK Biobank (UKBB), rs941616 is associated with allergic rhinitis ($p = 5 \times 10^{-4}$) but not asthma ($p = 0.077$) (Canela-Xandri et al., 2018).

Phenome-wide Association Studies (pheWASs)

We queried the 5,552 trans-ethnic genome-wide significant variants associated with blood cell traits in three ancestrally distinct biobanks, including 408,961 EUR individuals from the UKBB with 1,403 disease states, 143,988 individuals of Japanese descent from BioBank Japan (BBJ) with 22 disease states, and 5,275 African Americans from the Vanderbilt University Biobank (BioVU) with 1,403 disease states (STAR Methods). We found 366 variant-disease associations in the UKBB (Table S4A). Of these 366 associations, the BBJ had matching phenotypes for 95, 26 of which were replicated. Only one of these 366 associations was replicated in the BioVU. In only three cases did we observe a variant-disease association in the UKBB that failed to replicate when the BBJ was well powered (power >80%) and the matching phenotype was available. We found 133 variant-disease associations in the BBJ (Table S4A). Of these 133 associations, the UKBB had matching phenotypes for 90, 55 of which were

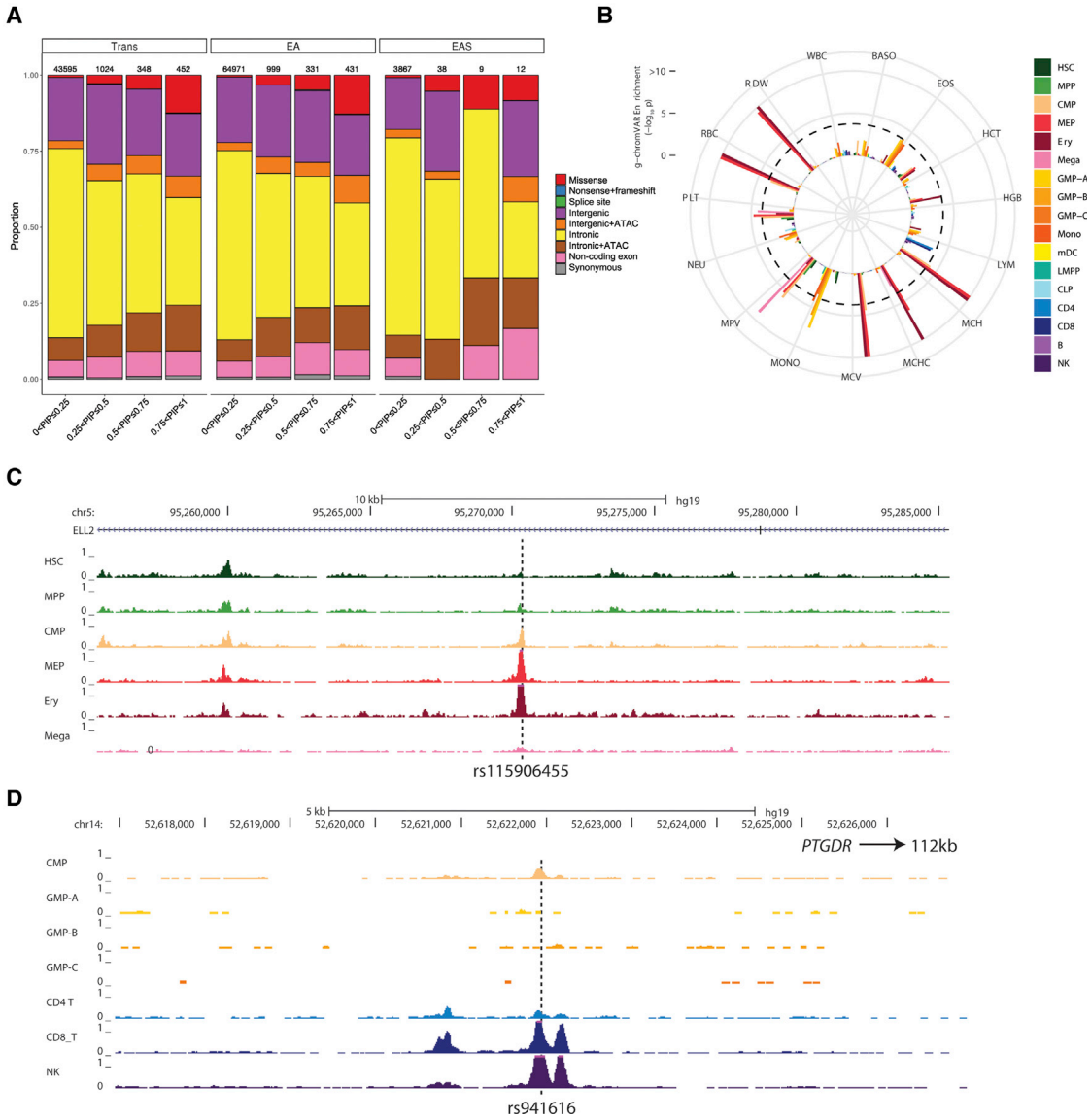


Figure 3. Functional Annotation of Possible Causal Variants Associated with Blood Cell Traits

(A) Annotation of variants in trans-ethnic, EUR, and EAS populations shows a similar pattern, with a larger proportion of likely functional variants (e.g., missense, intergenic, and intronic variants within ATAC-seq peaks) among variants with a higher PIP.

(B) g-chromVAR results for trans-ethnic variants within 95% credible sets for 15 traits. The Bonferroni-adjusted significance level (corrected for 15 traits and 18 cell types) is indicated by the dotted line. Mono, monocyte; HSC, hematopoietic stem cell; Ery, erythroid; Mega, megakaryocyte; CD4, CD4+ T lymphocyte; CD8, CD8+ T lymphocyte; B, B lymphocyte; NK, natural killer; mDC, myeloid dendritic cell; pDC, plasmacytoid dendritic cell; MPP, multipotent progenitor; LMPP, lymphoid-primed MPP; CMP, common myeloid progenitor; CLP, common lymphoid progenitor; GMP, granulocyte-macrophage progenitor; MEP, Mega-Ery progenitor.

(C) rs115906455 is a novel variant associated with mean corpuscular volume in the trans-ethnic meta-analysis ($p = 4.2 \times 10^{-12}$, PIP = 0.57). It maps to an intron of *ELL2* and overlaps with ATAC-seq peaks found in CMP, MEPs, and erythroblasts but not Megas.

(D) rs941616 is a novel variant associated with eosinophil counts in the trans-ethnic meta-analysis ($p = 2.4 \times 10^{-9}$, PIP = 0.2). It is a strong eQTL for *PTGDR* located 112 kb downstream and overlaps with ATAC-seq peaks found in CMP, CD8+ lymphocytes, and NK cells.

See also Figures S3 and Table S4 and Tables S3H and S3I.

replicated in the UKBB and one of which was replicated in the BioVU. Almost all of the non-replicated associations were well powered to replicate in the UKBB, suggesting heterogeneity across populations in genetic effects, in clinical definitions of disease states, or in disease prevalence and relevant environmental exposure. Only three of the non-replicated associations were well powered to replicate in the BioVU. Finally, in the BioVU, we observed 19 variant-disease associations (Table S4A), 18 of which were located at the β-globin locus and reflect the known clinical sequelae of sickle cell disease. Unsurprisingly, these

were not replicated in the UKBB and BBJ because the variant is monomorphic.

Many of the variant-disease associations we observed were located at well-known highly pleiotropic loci with a signal in multiple biobanks. For instance, rs1260326 in *GCKR* was associated with diabetes, dyslipidemia, alcohol consumption, gout, and urolithiasis. Multiple variants in *TERT* were associated with pre-cancerous conditions such as seborrheic keratosis, uterine leiomyoma, and myeloproliferative disease. Unsurprisingly, the major histocompatibility complex (MHC) region harbored multiple variants associated with a variety of immune-related diseases, such as celiac disease, psoriasis, asthma, rheumatoid arthritis, Graves' disease, and type 1 diabetes. Variants in and near *ABO* were associated with cardiovascular disease phenotypes as well as gastric cancer, hemorrhoids, and diverticulosis. Variants in and near *APOE* were associated with cardiovascular diseases and neurological disorders, including dementia.

We found two regions with widespread pleiotropy that were specific to a particular ancestry (in addition to the β-globin locus in AFR). Variants in and near *SH2B3* were associated with celiac disease, myocardial infarction, hypertension, and hypothyroidism in the UKBB. None of these associations were replicated in the BBJ because of these variants having very low MAFs (~0.3%) in EAS. About 2 Mb away from *SH2B3*, rs11066008 in *ACAD10* was associated with angina, myocardial infarction, arrythmia, and colorectal cancer in the BBJ. None of these associations were replicated in the UKBB because of very low MAFs in EUR and AFR (0% and 0.08%, respectively). A well-known selective sweep in this region approximately 1,200–1,700 years ago in European populations may explain why these loci display such large ancestry-specific effects (Zhernakova et al., 2010).

Trans-ethnic Predictions of Hematological Traits

Polygenic trait scores (PTSs) developed in a single ethnically homogeneous population tend to underperform when tested in a different population (Grinde et al., 2019; Márquez-Luna et al., 2017; Martin et al., 2019). We explored whether we could combine the genome-wide significant trans-ethnic variants identified in our analyses into PTSs that can predict blood cell traits in a multi-ethnic setting. First, we used trans-ethnic effect sizes as weights to compute PTS_{trans} for each trait and tested their performance in independent EUR, AFR, and HA participants from the BioMe Biobank (STAR Methods). As expected, because our trans-ethnic meta-analyses are dominated by EUR individuals, PTS_{trans} were more predictive in EUR, although their performance in HA was comparable for several traits (lymphocyte and monocyte counts and mean PLT volume) (Figure 4A; Table S4B). For neutrophil and WBC counts, the variance explained by PTS_{trans} was up to three times higher in AFR and HA than in EUR samples because of inclusion of the strong Duffy/*DARC* locus (Figure 4A; Table S4B). Because these Duffy/*DARC* variants would not have been included in PTSs derived uniquely from EUR association results, this illustrates an interesting feature of using trans-ethnic variants for building polygenic predictors. Consistent with previous reports for other human diseases, PTS_{trans} improved the precision to predict hematological disorders, defined using blood cell clinical thresholds (Figures 4B and 4C; Table S4C).

Next we asked whether we could increase the variance explained by calculating PTSs using the same trans-ethnic variants but weighting these variants using ancestry-specific as opposed to trans-ethnic effect sizes. In contrast to our expectations that a PTS calculated using ancestry-specific weights would be more accurate, we found, for most traits, that PTS_{trans} outperformed ancestry-specific PTS_{AFR} and PTS_{HA} in BioMe AFR and HA participants, respectively (Figure S5; Table S4B). This result likely indicates that the discovery sample size for these two populations is still too small to provide robust estimates of the true population-specific effect sizes and that additional ancestry-specific variants have yet to be identified.

Rare Coding Blood Cell Trait-Associated Variants

Identification of rare coding variants has successfully pinpointed candidate genes for many complex traits, including blood cell phenotypes (Auer et al., 2014; Chami et al., 2016; Eicher et al., 2016; Justice et al., 2019; Marouli et al., 2017; Mousas et al., 2017; Tajuddin et al., 2016). Our trans-ethnic and non-EUR-specific meta-analyses yielded 16 coding variants with a MAF of less than 1% (Tables S5A and S5B). This list includes variants of clinical significance (variants in *TUBB1*, *GFI1B*, *HBB*, *MPL*, and *SH2B3*) and variants that nominate candidate genes within GWAS loci (*ABCA7* and *GMPR*). Our analyses also retrieved a known missense variant in *EGLN1* (rs186996510) that is associated with high-altitude adaptation and hemoglobin levels in Tibetans (Lorenzo et al., 2014; Xiang et al., 2013).

We noted a missense variant in *IL7* (rs201412253, Val18Ile) associated with increased lymphocyte count in South Asians (SAS) ($p = 4.4 \times 10^{-10}$) (Figure 5A; Table S5C). This variant is low-frequency in SAS (MAF = 2.6%) but rare in other populations (MAF < 0.4%).

This association was replicated in 4,554 British-Pakistani and 10,638 British-Bangladeshi participants from the Genes & Health Study (combined $p = 5.7 \times 10^{-5}$) (Table S5C). *IL7* encodes interleukin-7, a cytokine essential for B and T cell lymphopoiesis (Lin et al., 2017). In large eQTL datasets such as eQTLgen and GTEx, rs201412253 is monomorphic. However, we found four heterozygote individuals among 75 Gujarati Indians that had genotypes and transcriptomics data from lymphoid cell lines (Stranger et al., 2012); in this limited dataset, rs201412253 was not associated with *IL7* expression levels (Figure 5B). Interleukin-7 (IL-7) is synthesized as a proprotein that is cleaved prior to secretion, and the IL-7-Val18Ile variant localizes to the IL-7 signal peptide comprising the first 25 amino acids. To determine whether this variant alters IL-7 secretion, we engineered HEK293 cells with either IL-7 allele (STAR Methods). Although there was no difference in *IL7* RNA expression levels (t test, $p = 0.63$), we found that the IL-7-18Ile allele, which associates with higher lymphocyte counts in SAS individuals, significantly increased IL-7 protein secretion in this heterologous cellular system (+83%, $p = 2.7 \times 10^{-5}$) (Figure 5C).

Genetic Architecture of Blood Cell Traits in EUR and EAS Populations

The genetic architecture of a trait is defined by the number, frequency, and effect size of all variants that contribute to phenotypic variation (Hansen et al., 2006). We used several

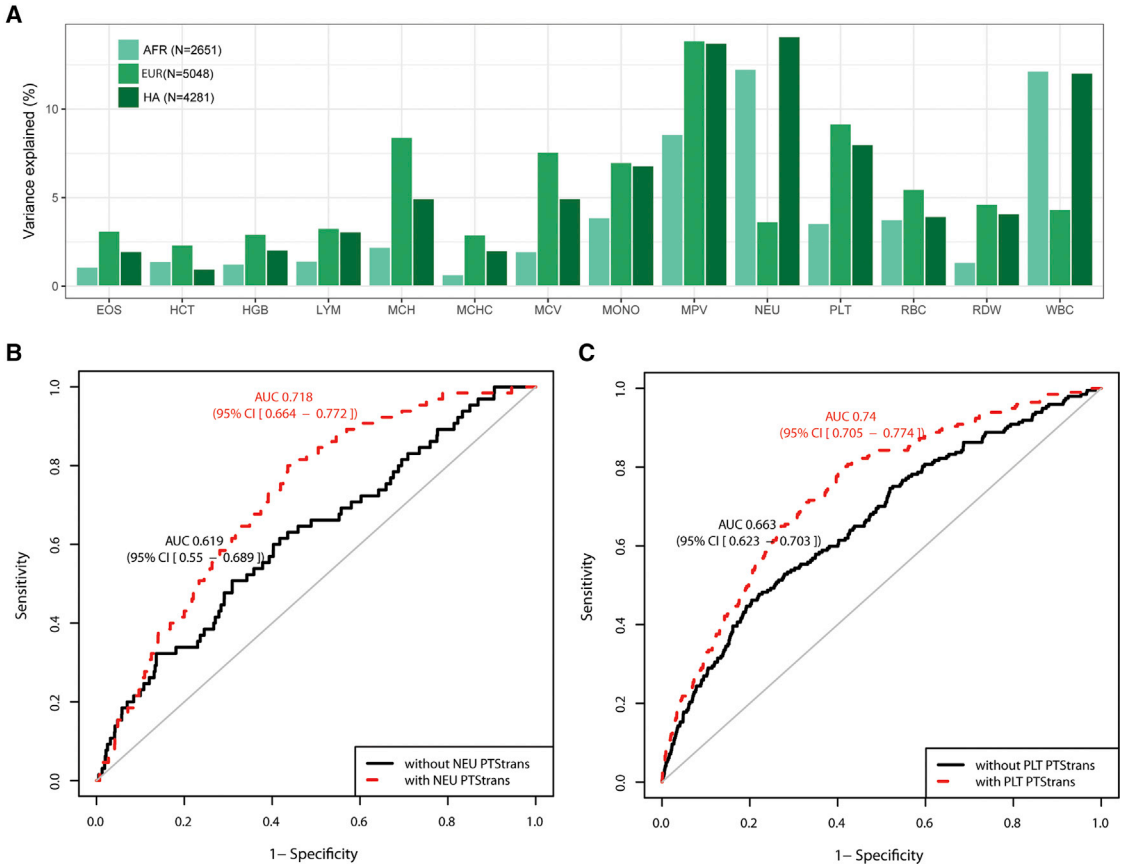


Figure 4. Phenotypic Variance and Hematological Disease Prediction Using Polygenic Trait Scores in Independent Participants from the BioMe Biobank

(A) For each blood cell trait, PTS_{trans} were calculated using genome-wide significant variants identified in the trans-ethnic meta-analyses. Trait-increasing alleles were weighted using effect sizes derived from fixed-effect trans-ethnic meta-analyses.

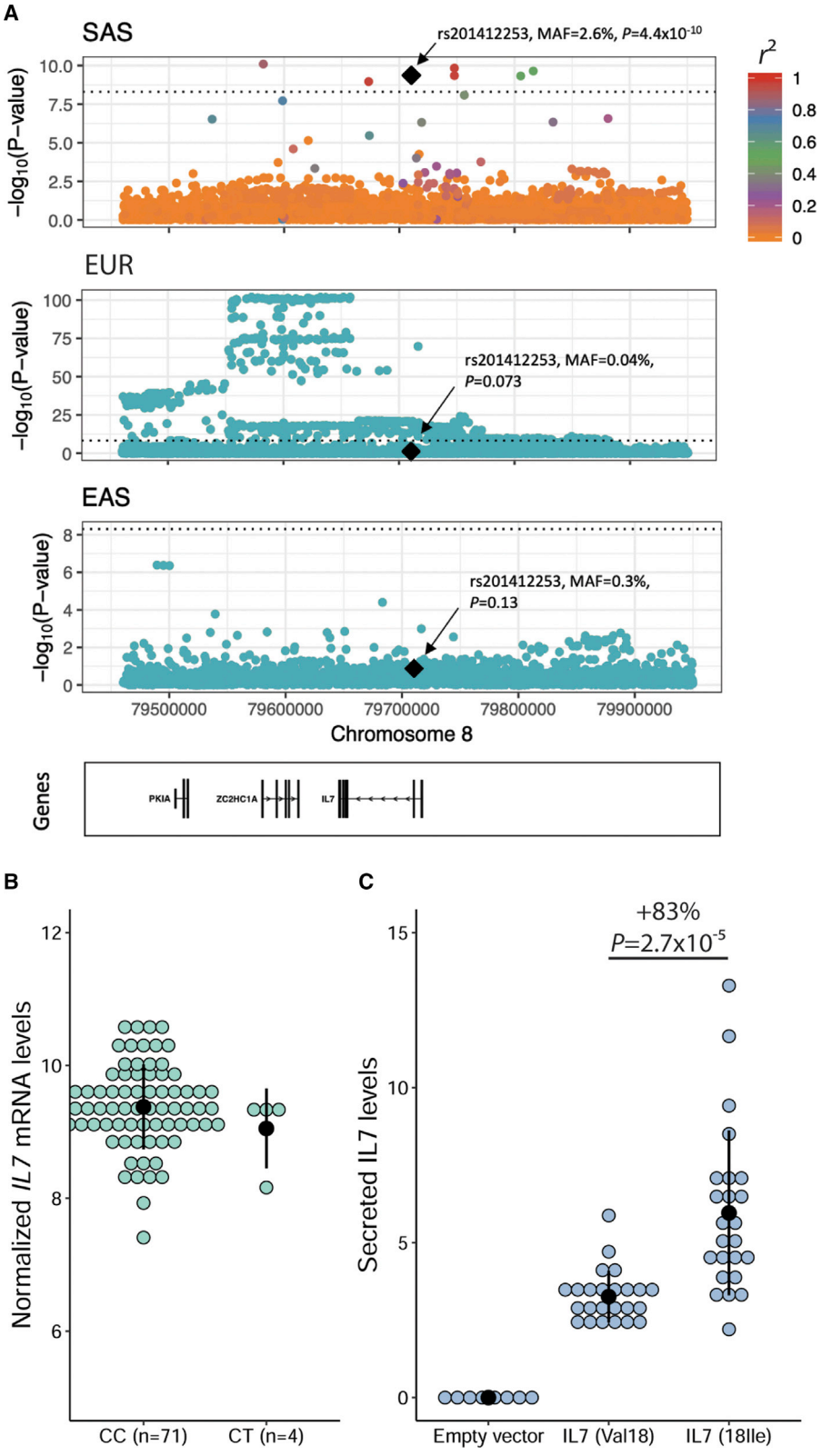
(B) Receiver operating characteristic (ROC) curve and area under the curve (AUC and 95% confidence interval) for neutropenia (defined as <1,500 neutrophils [NEU]/μL) in BioMe AFR participants without (black) or with (red) the PTS_{trans} for neutrophil count in the predictive model. Age, sex, and the first 10 principal components were used in the basic prediction model.

(C) As for (B) but for thrombocytopenia (defined as $<150 \times 10^9$ PLTs/L) and the PTS_{trans} for PLT count in HA participants from BioMe.

See also Figure S5 and Tables S4B and S4C.

different approaches to quantify similarities and differences in genetic architecture of hematologic traits across populations. Focusing on the two largest studied populations, EUR and EAS, we calculated heritability for all blood traits and found them to be concordant between ancestries (Pearson's $r = 0.75$, $p = 0.0033$) (Figure S6; Table S6A; Bulik-Sullivan et al., 2015b). Likewise, within-ancestry genetic correlation coefficients (r_g) between pairs of hematological traits were highly concordant across ancestries (Pearson's $r = 0.97$, $p < 2.2 \times 10^{-16}$) (Figure S6; Bulik-Sullivan et al., 2015a). We then used the Popcorn method to measure genetic correlations for blood cell traits between EUR and EAS using summary statistics for common variants (Brown et al., 2016). For all 13 traits available in EUR and EAS, genetic correlations were high (lowest for basophils [$r_g = 0.30$] and highest for MCH [$r_g = 0.66$]) but significantly different than 1 ($p < 3 \times 10^{-6}$) (Figure S6; Table S6B). This suggests that, although the effect sizes of common variants are correlated between EUR and EAS, there are significant differences between these two populations.

To further contrast the genetic architecture of blood cell traits between these two populations, we compared effect sizes for 1,423 genome-wide significant variants with a PIP of more than 0.5 in EUR or EAS (Figure 6; Table S6C). Effect sizes were correlated (Pearson's $r = 0.46$ for variants with PIP > 0.5 in EUR and $r = 0.70$ for variants with PIP > 0.5 in EAS, $p < 2.2 \times 10^{-16}$ for both) (Figure 6), which indicated largely concordant effect sizes across populations, a result consistent with the Popcorn analyses. We also noted many interesting differences. We found 70 variants with a PIP of more than 0.5 that are common (MAF > 5%) and have a similar MAF (±5%) in EUR and EAS but have at least a 2-fold difference in effect size (Table S6C). For instance, rs34651 is strongly associated with PLT in EUR ($p = 1.1 \times 10^{-30}$, PIP > 0.99, effect size = −0.0428), but the association signal is weaker in EAS ($p = 2.5 \times 10^{-7}$, effect size = −0.0336) despite the fact that the variant is more common



(legend on next page)

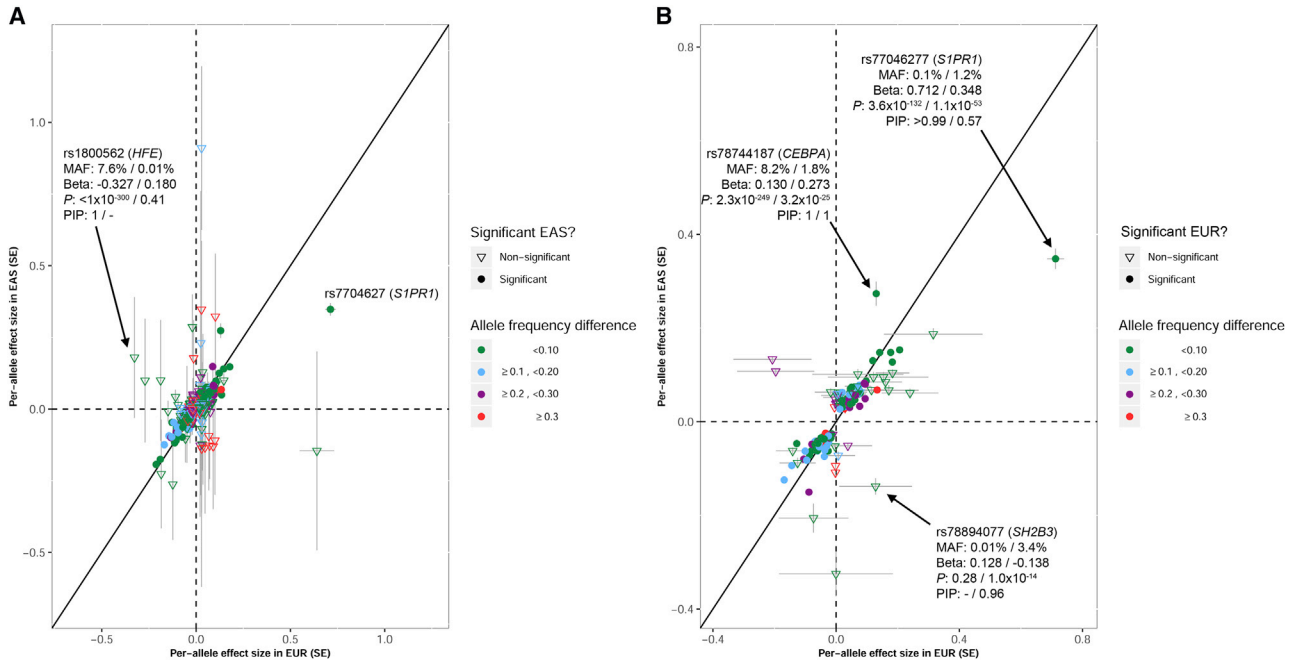


Figure 6. Comparisons of Effect Sizes for Variants with PIPs of More Than 0.5

We retained only variants with an analyzed sample size of 70,000 or more in EAS and 100,000 or more in EUR participants.

(A) We retrieved minor allele frequencies (MAFs), effect sizes (Beta), p values, and PIPs for all variants with a PIP of more than 0.5 in EUR. By definition, all of these variants are significant in EUR ($p < 5 \times 10^{-9}$). For these variants, we then retrieved the corresponding results in EAS. Effect sizes (standard error [SE]) in EUR and EAS are plotted on the x and y axis, respectively.

(B) as in (A) but for variants with a PIP of more than 0.5 in EAS.

In (A) and (B), when we provide detailed information on a specific variant, the first number always corresponds to EUR and the second to EAS (e.g., for rs77046277, $Beta_{EUR} = 0.712$ and $Beta_{EAS} = 0.348$). See also Figure S4 and Table S6C.

in EAS ($MAF_{EUR} = 8.1\%$, $MAF_{EAS} = 12.9\%$) (Table S6C). This variant maps to a region of accessible chromatin in most progenitor and mature hematopoietic cells and is a strong eQTL for *FCHO2* (Figure S4; Võsa et al., 2018). Some variants were also significant in EUR and EAS even when they had different effect sizes and MAFs. This category includes rs77046277, which is strongly associated with lymphocyte counts (LYM) in EUR and EAS despite being rare in EUR ($MAF_{EUR} = 0.1\%$, $MAF_{EAS} = 1.2\%$) (Figure 6). This variant is located near *S1PR1* and maps to regions of accessible chromatin found in T lymphocytes (Figure S4). rs78744187 is another example; it is common in EUR and less frequent in EAS ($MAF_{EUR} = 8.2\%$, $MAF_{EAS} = 1.8\%$) but strongly associated with basophil count in both populations (Figure 6). This variant is an eQTL for *CEBPA* (Võsa et al., 2018) and is located within an accessible chromatin region in CMPs, and prior studies using genome editing of this regulatory element in primary hematopoietic progenitors have validated its role in regulating *CEBPA* expression to enable basophil production (Figure S4) (Guo et al., 2017). Finally, there were also variants that were ancestry specific because they were very rare in the other population; this included the known missense variants in *SH2B3* (rs78894077) associated with LYM in EAS and in *HFE* (rs1800562) associated with RBC traits in EUR (Figure 6).

Natural Selection at Blood Cell Trait Loci

Natural selection can account for differences in association results between populations, as highlighted by our analyses of rare coding variants that includes several loci known to be under

Figure 5. An SAS *IL7* Missense Variant Associates with Increased Lymphocyte Count in Humans and IL-7 Secretion *In Vitro*

(A) Lymphocyte count association results at the *IL7* locus in SAS, EUR, and EAS. In SAS, there are seven genome-wide significant variants near *IL7*, but only rs201412253 is coding. LD r^2 is from 1000G SAS populations. In EUR, the sentinel variant is located downstream of *IL7*; rs201412253 is rare ($MAF = 4 \times 10^{-4}$) and not significant (p = 0.073). In EAS, the locus is not associated with lymphocyte count. rs201412253 is monomorphic in 1000G EUR and EAS, so we could not calculate pairwise LD.

(B) Association between genotypes at rs201412253 and normalized *IL7* expression levels in lymphoid cell lines from 75 Gujarati Indians from HapMap3. The T allele frequency is 2.7%, and the association is not significant (p = 0.62).

(C) The 18Ile allele at *IL7*-rs201412253 increases IL-7 secretion in a heterologous cellular system. Our ELISA did not detect secreted IL-7 in clones generated with an empty vector. We tested eight independent clones for each *IL7* allele. Each experiment was done in duplicate, and we performed the experiments three times. The black dots and vertical lines indicate means and standard deviations. We assessed statistical significance by linear regression, correcting for experimental batch effects.

See also Tables S5A–S5C.

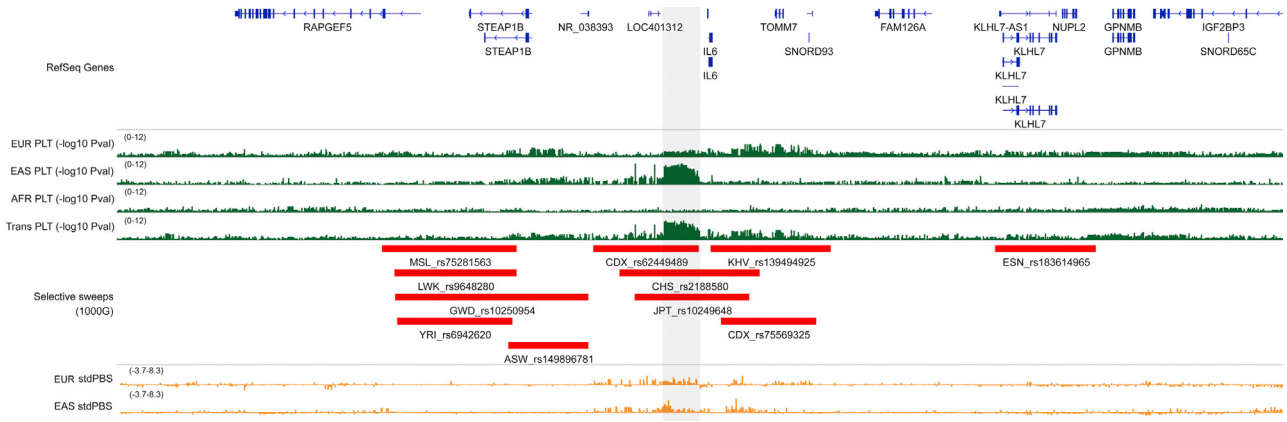


Figure 7. Selective Sweep and Association with PLT Count at the *IL6* Locus in EAS

The gray rectangle highlights a genomic region upstream of *IL6* that is strongly associated with platelet (PLT) count. This association signal is driven by results from EAS and is absent from other populations, including EUR and AFR individuals (green). The region overlaps several selective sweeps detected in EAS from 1000G (Chinese Dai in Xishuangbanna [CDX], Southern Han Chinese [CHS], and Japanese in Tokyo [JPT]). In orange, we provide standardized population branch site (stdPBS) metrics in EUR and EAS, indicative of allele frequency differentiation at this locus between these two populations. The coordinates are chr7:22-23.5Mb (hg19). See also Figure S7 and Tables S7A–S7C.

selection (*CD36*, β-globin, and *EGLN1*) (Table S5A). To further explore this possibility, we assessed whether variants that tag selective sweeps (tagSweeps, variants with the highest integrated haplotype score [iHS]) within continental populations from the 1000 Genomes Project (1000G) are associated with blood cell phenotypes (Johnson and Voight, 2018). We found genome-wide enrichment of association results between tagSweeps and hematological traits, particularly within EUR, EAS, and AFR populations (Figure S7; Table S7A). To rule out simple overlaps because of the large number of sweeps and blood cell trait loci, we compared the number of genome-wide significant tagSweeps in EUR, EAS, and AFR with the number of significant variants among 100 sets of matched variants (STAR Methods). We found significant enrichment of selective sweeps for WBCs (EUR, EAS, and AFR), monocytes (EUR and AFR), eosinophils (EUR), neutrophils (AFR), lymphocytes (EAS), and PLTs (EUR and EAS) (Table S7B).

In AFR and HA, the enrichments for WBCs, neutrophils, and monocytes were entirely driven by selective sweeps on chromosome 1 near Duffy/*DARC* (Reich et al., 2009). Only three additional loci shared evidence of associations with blood cell traits and positive selection across populations: HLA, *SH2B3* (Zhernakova et al., 2010), and *CYP3A5* (Chen et al., 2009). We found eight and 100 non-overlapping selective sweeps with variants associated with hematological traits in EAS and EUR, respectively (Table S7C). Six of the eight EAS-specific tagSweeps are also associated with blood-cell traits in EUR participants, indicating that these regions do not account for population differences in hematological trait regulation (Table S7C). One of the remaining two variants is located at the *HBS1L-MYB* locus, and although it is not associated with blood cell traits in EUR, there are many other variants near *MYB* associated with blood phenotypes in EUR (Table S3B). The remaining selective sweep highlighted by this analysis is located upstream of *IL6* (Figure 7). The tagSweep at this locus, rs2188580, is strongly associated with PLT count in EAS (P_{EAS} = 2.8 × 10^{-9}, P_{EUR} = 0.0022); is differentiated between EAS and EUR, as indicated by the population branch statistic (PBS) (Yi et al., 2010) (C-allele frequency in EAS = 44% and 4% in EUR, standardized PBS_{EAS} = 7.353); and overlaps selective sweeps identified in several EAS populations from 1000G (e.g., iHS_{CHS} = 3.935) (Figure 7). The *IL6* locus has been associated previously with WBC traits in EUR (Astle et al., 2016), but our finding is the first report of its association with PLTs. *IL6* encodes IL-6, a cytokine that is a maturation factor for megakaryocytes (Kimura et al., 1990). Further supporting the role of IL-6 signaling in PLT biology, a well-characterized missense variant in the *IL6* receptor gene (*IL6R*-rs2228145) (van Dongen et al., 2014) is also nominally associated with PLT count in EAS (p = 4.3 × 10^{-6}).

DISCUSSION

Our meta-analyses of 15 hematological traits in up to 746,667 individuals is one of the largest genetic studies of clinically relevant complex human traits across diverse ancestral groups. We continued to expand the repertoire of loci and genes that contribute to interindividual variation in blood cell traits, with potential implications for hematological diseases as well as other conditions, such as cancer and immune and cardiovascular diseases.

Differences in clinical definitions, phenotype measurements, and gene-gene and gene-environment interactions could account for some of the differences in genetic effects observed between populations. In our analyses of hematological traits in EUR and EAS, we identified extensive genetic overlaps but also significant differences in effect sizes between these two populations. Our estimates of trans-ancestry genetic correlations for blood cell traits are similar to estimates for other complex human phenotypes, such as type 2 diabetes, rheumatoid arthritis, Crohn's disease, and ulcerative colitis (Brown et al., 2016; Liu et al., 2015), although higher genetic correlations have also been reported (Lam et al., 2019; Martin et al., 2019). Despite

the shared genetic architecture, we found evidence of heterogeneity at hematological trait-associated variants with a high PIP (Figure 6). Similarly, although the genetic correlation for Crohn's disease between EUR and EAS is high (r_g = 0.76), heterogeneity was noted at causal variants in *NOD2*, *IL23R*, and *TNFSF15* because of differences in allele frequency, effect size, or both (Liu et al., 2015). This is in sharp contrast with a recent report that the genetic correlation between EUR and EAS for schizophrenia is near unity (r_g = 0.98) and that there is no evidence of locus-level heterogeneity (Lam et al., 2019). These observations, largely limited to EUR-EAS comparisons for a handful of phenotypes, already suggest that different complex human diseases and traits have different genetic architecture. These results also highlight a need for large genetic analyses in other populations and for development of methodologies amenable to admixture for genetic correlation analyses.

Our results have implications for future human genetic studies. First, we showed that adding even a "modest" number of non-EUR participants to GWASs can yield important biology, such as identification of LYM count-associated *IL7* missense variants in 8,189 SAS (Figure 5). Second, loci that underlie variation in blood cell traits represent a broad mixture of shared associations (i.e., similar allele frequencies and effect sizes across populations) and heterogeneous associations (i.e., dissimilar allele frequencies and effect sizes across populations). This result contributes to mounting evidence that a full accounting of the genetic basis of complex human traits will require a thorough catalog of global genetic and phenotypic variations. Third, because of heterogeneity across populations in allele frequencies and patterns of LD, fine-mapping of association signals can be substantially aided by including multiple ancestries. This will have a dramatic effect on the success of large-scale efforts aimed at functionally characterizing GWAS findings but also to develop polygenic predictors that transfer to multiple ancestries. As more studies seek to unravel the causal variants that underlie complex traits associations, we anticipate that genetic evidence from diverse ancestries will play an important role.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Material Availability
  - Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Study design and participants
  - Cell lines
- METHOD DETAILS
  - Phenotypes
  - Genotype quality-control and imputation
  - Study-level statistical analyses
  - Centralized quality-control and meta-analyses
  - Million Veteran Program (MVP) blood-cell trait analyses for replication
  - Heritabilities and genetic correlations
  - Functional annotation
  - Phenome-wide association study (pheWAS) analysis
  - Polygenic trait score (PTS) analyses
  - Analysis of natural selection
  - Replication of the association between *IL7*-rs201412253 and lymphocyte count in Genes & Health
  - IL7 functional analyses
- QUANTIFICATION AND STATISTICAL ANALYSES
  - Statistical significance, genomic inflation and locus definition
  - Conditional analyses in the UK Biobank European-ancestry population to identify independent variants associated with blood-cell traits
  - Statistical fine-mapping

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2020.06.045.

ACKNOWLEDGMENTS

We thank all participants as well as Dr. John D. Rioux for providing the *IL7* ORF. A full list of acknowledgments and funding support appears in Table S1F. Part of this work was conducted using the UK Biobank resource (Projects 11707 and 13745).

AUTHOR CONTRIBUTIONS

Writing and editing, M.-H.C., L.M.R., A. Mousas, E.L.B., A.D.J., A.P.R., P.L.A., and G.L.; data preparation (checking and preparing data from contributing cohorts for meta-analyses and replication), M.-H.C., L.M.R., A. Mousas, J.E.H., G.L., and P.L.A.; meta-analyses (discovery and replication), M.-H.C., L.M.R., A. Mousas, S.S., J.E.H., T.J., P.A., D.V., E.L.B., A. Moscati, K.S.L., C.A.L., M.H.G., T.K., F.K., A. Manichaikul, M.P., and C.N.S.; fine-mapping and functional annotation, A. Mousas, M.-H.C., L.M.R., E.L.B., C.A.L., K.S.L., V.G.S., A.D.J., P.L.A., and G.L.; pheWAS, polygenic prediction, genetic architecture, and natural selection, S.S., A. Moscati, R.M., M.C., K.S.L., H.Q., Y.L., C.W.K.C., R.J.F.L., A.P.R., G.L., and P.L.A.; functional characterization of IL-7, M.B., V.L., G.L., and J.-F.G.; replication of IL7- rs201412253 in Genes & Health, B.T., K.A.H., H.C.M., Q.Q.H., R.C.T., and D.A.v.H. All authors contributed and discussed the results and commented on the manuscript.

DECLARATION OF INTERESTS

Competing financial interests are declared in Table S1F.

Received: December 6, 2019
Revised: April 16, 2020
Accepted: June 29, 2020
Published: September 3, 2020

REFERENCES

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

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.

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.

Beutler, E., and West, C. (2005). Hematologic differences between African-Americans and whites: the roles of iron deficiency and alpha-thalassemia on hemoglobin levels and mean corpuscular volume. Blood *106*, 740–745.

Brown, B.C., Ye, C.J., Price, A.L., and Zaitlen, N.; Asian Genetic Epidemiology Network Type 2 Diabetes Consortium (2016). Transethnic Genetic-Correlation Estimates from Summary Statistics. Am. J. Hum. Genet. *99*, 76–88.

Brusselle, G.G., Provoost, S., and Maes, T. (2016). Prostaglandin D2 receptor antagonism: a novel therapeutic option for eosinophilic asthma? Lancet Respir. Med. *4*, 676–677.

Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.R., Duncan, L., Perry, J.R., Patterson, N., Robinson, E.B., et al.; ReproGen Consortium; Psychiatric Genomics Consortium; Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3 (2015a). An atlas of genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.

Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015b). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.

Byrnes, J.R., and Wolberg, A.S. (2017). Red blood cells in thrombosis. Blood *130*, 1795–1799.

Canela-Xandri, O., Rawlik, K., and Tenesa, A. (2018). An atlas of genetic associations in UK Biobank. Nat. Genet. *50*, 1593–1599.

Carroll, R.J., Bastarache, L., and Denny, J.C. (2014). R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics *30*, 2375–2376.

Chami, N., Chen, M.H., Slater, A.J., Eicher, J.D., Evangelou, E., Tajuddin, S.M., Love-Gregory, L., Kacprowski, T., Schick, U.M., Nomura, A., et al. (2016). Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits. Am. J. Hum. Genet. *99*, 8–21.

Chen, X., Wang, H., Zhou, G., Zhang, X., Dong, X., Zhi, L., Jin, L., and He, F. (2009). Molecular population genetics of human CYP3A locus: signatures of positive selection and implications for evolutionary environmental medicine. Environ. Health Perspect. *117*, 1541–1548.

Chu, S.G., Becker, R.C., Berger, P.B., Bhatt, D.L., Eikelboom, J.W., Konkle, B., Mohler, E.R., Reilly, M.P., and Berger, J.S. (2010). Mean platelet volume as a predictor of cardiovascular risk: a systematic review and meta-analysis. J. Thromb. Haemost. *8*, 148–156.

Colin, Y., Le Van Kim, C., and El Nemer, W. (2014). Red cell adhesion in human diseases. Curr. Opin. Hematol. *21*, 186–192.

Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M., Koenig, J.L., Snyder, M.P., Pritchard, J.K., Kundaje, A., Greenleaf, W.J., et al. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat. Genet. *48*, 1193–1203.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287.

Delaneau, O., Zagury, J.F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods *10*, 5–6.

Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D.R., Roden, D.M., and Crawford, D.C. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics *26*, 1205–1210.

Ding, K., de Andrade, M., Manolio, T.A., Crawford, D.C., Rasmussen-Torvik, L.J., Ritchie, M.D., Denny, J.C., Masys, D.R., Jouni, H., Pachecho, J.A., et al. (2013). Genetic variants that confer resistance to malaria are associated with red blood cell traits in African-Americans: an electronic medical record-based genome-wide association study. G3 (Bethesda) *3*, 1061–1068.

Eicher, J.D., Chami, N., Kacprowski, T., Nomura, A., Chen, M.H., Yanek, L.R., Tajuddin, S.M., Schick, U.M., Slater, A.J., Pankratz, N., et al.; Global Lipids Genetics Consortium; CARDIoGRAM Exome Consortium; Myocardial Infarction Genetics Consortium (2016). Platelet-Related Variants Identified by Exomechip Meta-analysis in 157,293 Individuals. Am. J. Hum. Genet. *99*, 40–55.

Evans, D.M., Frazer, I.H., and Martin, N.G. (1999). Genetic and environmental causes of variation in basal levels of blood cells. Twin Res. *2*, 250–257.

Fang, H., Hui, Q., Lynch, J., Honerlaw, J., Assimes, T.L., Huang, J., Vujkovic, M., Damrauer, S.M., Pyarajan, S., Gaziano, J.M., et al.; VA Million Veteran Program (2019). Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. Am. J. Hum. Genet. *105*, 763–772.

Finer, S., Martin, H.C., Khan, A., Hunt, K.A., MacLaughlin, B., Ahmed, Z., Ashcroft, R., Durham, C., MacArthur, D.G., McCarthy, M.I., et al. (2020). Cohort Profile: East London Genes & Health (ELGH), a community-based population genomics and health study in British Bangladeshi and British Pakistani people. Int. J. Epidemiol. *49*, 20–21i.

Gaziano, J.M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., et al. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J. Clin. Epidemiol. *70*, 214–223.

Grinde, K.E., Qi, Q., Thornton, T.A., Liu, S., Shadyab, A.H., Chan, K.H.K., Reiner, A.P., and Sofer, T. (2019). Generalizing polygenic risk scores from Europeans to Hispanics/Latinos. Genet. Epidemiol. *43*, 50–62.

Guo, M.H., Nandakumar, S.K., Ulirsch, J.C., Zekavat, S.M., Buenrostro, J.D., Natarajan, P., Salem, R.M., Chiarle, R., Mitt, M., Kals, M., et al. (2017). Comprehensive population-based genome sequencing provides insight into hematopoietic regulatory mechanisms. Proc. Natl. Acad. Sci. USA *114*, E327–E336.

Hansen, T.F., Alvarez-Castro, J.M., Carter, A.J., Hermisson, J., and Wagner, G.P. (2006). Evolution of genetic architecture under directional selection. Evolution *60*, 1523–1536.

Hinckley, J.D., Abbott, D., Burns, T.L., Heiman, M., Shapiro, A.D., Wang, K., and Di Paola, J. (2013). Quantitative trait locus linkage analysis in a large Amish pedigree identifies novel candidate loci for erythrocyte traits. Mol. Genet. Genomic Med. *1*, 131–141.

Johnson, K.E., and Voight, B.F. (2018). Patterns of shared signatures of recent positive selection across human populations. Nat. Ecol. Evol. *2*, 713–720.

Justice, A.E., Karaderi, T., Highland, H.M., Young, K.L., Graff, M., Lu, Y., Turcot, V., Auer, P.L., Fine, R.S., Guo, X., et al.; CHD Exome+ Consortium; Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium; EPIC-CVD Consortium; ExomeBP Consortium; Global Lipids Genetic Consortium; GoT2D Genes Consortium; InterAct; ReproGen Consortium; T2D-Genes Consortium; MAGIC Investigators (2019). Protein-coding variants implicate novel genes related to lipid homeostasis contributing to body-fat distribution. Nat. Genet. *51*, 452–469.

Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. *50*, 390–400.

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

Kichaev, G., Bhatia, G., Loh, P.R., Gazal, S., Burch, K., Freund, M.K., Schoech, A., Pasaniuc, B., and Price, A.L. (2019). Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am. J. Hum. Genet. *104*, 65–75.

Kimura, H., Ishibashi, T., Uchida, T., Maruyama, Y., Friese, P., and Burstein, S.A. (1990). Interleukin 6 is a differentiation factor for human megakaryocytes in vitro. Eur. J. Immunol. *20*, 1927–1931.

Klarin, D., Damrauer, S.M., Cho, K., Sun, Y.V., Teslovich, T.M., Honerlaw, J., Gagnon, D.R., DuVall, S.L., Li, J., Peloso, G.M., et al.; Global Lipids Genetics Consortium; Myocardial Infarction Genetics (MIGen) Consortium; Geisinger-Regeneron DiscovEHR Collaboration; VA Million Veteran Program (2018). Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat. Genet. *50*, 1514–1523.

Lam, M., Chen, C.Y., Li, Z., Martin, A.R., Bryois, J., Ma, X., Gaspar, H., Ikeda, M., Benyamin, B., Brown, B.C., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium; Indonesia Schizophrenia Consortium; Genetic REsearch on schizophreniA neTwork-China and the Netherlands (GREAT-CN) (2019). Comparative genetic architectures of schizophrenia in East Asian and European populations. Nat. Genet. *51*, 1670–1678.

Li, Y.R., and Keating, B.J. (2014). Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. Genome Med. *6*, 91.

Lin, J., Zhu, Z., Xiao, H., Wakefield, M.R., Ding, V.A., Bai, Q., and Fang, Y. (2017). The role of IL-7 in Immunity and Cancer. Anticancer Res. *37*, 963–967.

Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T., et al.; International Multiple Sclerosis Genetics Consortium; International IBD Genetics Consortium (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. *47*, 979–986.

Lo, K.S., Wilson, J.G., Lange, L.A., Folsom, A.R., Galarneau, G., Ganesh, S.K., Grant, S.F., Keating, B.J., McCarroll, S.A., Mohler, E.R., 3rd., et al. (2011). Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans. Hum. Genet. *129*, 307–317.

Loh, P.R., Palamara, P.F., and Price, A.L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. Nat. Genet. *48*, 811–816.

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

Lorenzo, F.R., Huff, C., Myllymäki, M., Olenchock, B., Swierczek, S., Tashi, T., Gordeuk, V., Wuren, T., Ri-Li, G., McClain, D.A., et al. (2014). A genetic mechanism for Tibetan high-altitude adaptation. Nat. Genet. *46*, 951–956.

Mägi, R., and Morris, A.P. (2010). GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics *11*, 288.

Mägi, R., Horikoshi, M., Sofer, T., Mahajan, A., Kitajima, H., Franceschini, N., McCarthy, M.I., and Morris, A.P.; COGENT-Kidney Consortium, T2D-GENES Consortium (2017). Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. Hum. Mol. Genet. *26*, 3639–3650.

Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J., Steinthorsdottir, V., Scott, R.A., Grarup, N., et al. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat. Genet. *50*, 1505–1513.

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.

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K.S., Wood, A.R., Kjaer, T.R., Fine, R.S., Lu, Y., Schurmann, C., Highland, H.M., et al.; EPIC-InterAct Consortium; CHD Exome+ Consortium; ExomeBP Consortium; T2D-Genes Consortium; GoT2D Genes Consortium; Global Lipids Genetics Consortium; ReproGen Consortium; MAGIC Investigators (2017). Rare and low-frequency coding variants alter human adult height. Nature *542*, 186–190.

Márquez-Luna, C., Loh, P.R., and Price, A.L.; South Asian Type 2 Diabetes (SAT2D) Consortium; SIGMA Type 2 Diabetes Consortium (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. Genet. Epidemiol. *41*, 811–823.

Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al.; Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. *48*, 1279–1283.

Mousas, A., Ntritsos, G., Chen, M.H., Song, C., Huffman, J.E., Tzoulaki, I., Elliott, P., Psaty, B.M., Auer, P.L., Johnson, A.D., et al.; Blood-Cell Consortium (2017). Rare coding variants pinpoint genes that control human hematological traits. PLoS Genet. *13*, e1006925.

Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., et al.; BioBank Japan Cooperative Hospital Group (2017). Overview of the BioBank Japan Project: Study design and profile. J. Epidemiol. *27* (3S), S2–S8.

Pers, T.H., Timshel, P., and Hirschhorn, J.N. (2015). SNPsnap: a Web-based tool for identification and annotation of matched SNPs. Bioinformatics *31*, 418–420.

Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161–164.

Popejoy, A.B., Ritter, D.I., Crooks, K., Currey, E., Fullerton, S.M., Hindorff, L.A., Koenig, B., Ramos, E.M., Sorokin, E.P., Wand, H., et al.; Clinical Genome Resource (ClinGen) Ancestry and Diversity Working Group (ADWG) (2018). The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. Hum. Mutat. *39*, 1713–1720.

Raffield, L.M., Ulirsch, J.C., Naik, R.P., Lessard, S., Handsaker, R.E., Jain, D., Kang, H.M., Pankratz, N., Auer, P.L., Bao, E.L., et al.; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Hematology & Hemostasis, Diabetes, and Structural Variation TOPMed Working Groups (2018). Common α-globin variants modify hematologic and other clinical phenotypes in sickle cell trait and disease. PLoS Genet. *14*, e1007293.

Raj, T., Kuchroo, M., Replogle, J.M., Raychaudhuri, S., Stranger, B.E., and De Jager, P.L. (2013). Common risk alleles for inflammatory diseases are targets of recent positive selection. Am. J. Hum. Genet. *92*, 517–529.

Rana, S.R., Sekhsaria, S., and Castro, O.L. (1993). Hemoglobin S and C traits: contributing causes for decreased mean hematocrit in African-American children. Pediatrics *91*, 800–802.

Rappoport, N., Simon, A.J., Amariglio, N., and Rechavi, G. (2019). The Duffy antigen receptor for chemokines, ACKR1,- 'Jeanne DARC' of benign neutropenia. Br. J. Haematol. *184*, 497–507.

Reich, D., Nalls, M.A., Kao, W.H., Akylbekova, E.L., Tandon, A., Patterson, N., Mullikin, J., Hsueh, W.C., Cheng, C.Y., Coresh, J., et al. (2009). Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. PLoS Genet. *5*, e1000360.

Roden, D.M., Pulley, J.M., Basford, M.A., Bernard, G.R., Clayton, E.W., Balser, J.R., and Masys, D.R. (2008). Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin. Pharmacol. Ther. *84*, 362–369.

Schick, U.M., Jain, D., Hodonsky, C.J., Morrison, J.V., Davis, J.P., Brown, L., Sofer, T., Conomos, M.P., Schurmann, C., McHugh, C.P., et al. (2016). Genome-wide Association Study of Platelet Count Identifies Ancestry-Specific Loci in Hispanic/Latino Americans. Am. J. Hum. Genet. *98*, 229–242.

Stranger, B.E., Montgomery, S.B., Dimas, A.S., Parts, L., Stegle, O., Ingle, C.E., Sekowska, M., Smith, G.D., Evans, D., Gutierrez-Arcelus, M., et al. (2012). Patterns of cis regulatory variation in diverse human populations. PLoS Genet. *8*, e1002639.

Swaminathan, B., Thorleifsson, G., Jöud, M., Ali, M., Johnsson, E., Ajore, R., Sulem, P., Halvarsson, B.M., Eyjolfsson, G., Haraldsdottir, V., et al. (2015). Variants in ELL2 influencing immunoglobulin levels associate with multiple myeloma. Nat. Commun. *6*, 7213.

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 Analysis Identifies Loci for White Blood Cell Traits

and Pleiotropy with Immune-Mediated Diseases. Am. J. Hum. Genet. *99*, 22–39.

Ulirsch, J.C., Lareau, C.A., Bao, E.L., Ludwig, L.S., Guo, M.H., Benner, C., Satpathy, A.T., Kartha, V.K., Salem, R.M., Hirschhorn, J.N., et al. (2019). Interrogation of human hematopoiesis at single-cell and single-variant resolution. Nat. Genet. *51*, 683–693.

van Dongen, J., Jansen, R., Smit, D., Hottenga, J.J., Mbarek, H., Willemsen, G., Kluft, C., Penninx, B.W., Ferreira, M.A., Boomsma, D.I., and de Geus, E.J.; AAGC Collaborators (2014). The contribution of the functional IL6R polymorphism rs2228145, eQTLs and other genome-wide SNPs to the heritability of plasma sIL-6R levels. Behav. Genet. *44*, 368–382.

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., et al. (2018). Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. bioRxiv. https://doi.org/10.1101/447367.

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. (2020). The Polygenic and Monogenic Basis of Blood Traits and Diseases. Cell *182*, this issue, 1214–1231.

Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am. J. Hum. Genet. *81*, 208–227.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with P-values. Genet. Epidemiol. *33*, 79–86.

Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure. Evolution *38*, 1358–1370.

Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191.

Williams, A.L., Patterson, N., Glessner, J., Hakonarson, H., and Reich, D. (2012). Phasing of many thousands of genotyped samples. Am. J. Hum. Genet. *91*, 238–251.

Winkler, T.W., Day, F.R., Croteau-Chonka, D.C., Wood, A.R., Locke, A.E., Mägi, R., Ferreira, T., Fall, T., Graff, M., Justice, A.E., et al.; Genetic Investigation of Anthropometric Traits (GIANT) Consortium (2014). Quality control and conduct of genome-wide association meta-analyses. Nat. Protoc. *9*, 1192–1212.

Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518.

Xiang, K., Ouzhuluobu, Peng, Y., Yang, Z., Zhang, X., Cui, C., Zhang, H., Li, M., Zhang, Y., Bianba, et al. (2013). Identification of a Tibetan-specific mutation in the hypoxic gene EGLN1 and its contribution to high-altitude adaptation. Mol. Biol. Evol. *30*, 1889–1898.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T.S., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. Science *329*, 75–78.

Zhernakova, A., Elbers, C.C., Ferwerda, B., Romanos, J., Trynka, G., Dubois, P.C., de Kovel, C.G., Franke, L., Oosting, M., Barisani, D., et al.; Finnish Celiac Disease Study Group (2010). Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. Am. J. Hum. Genet. *86*, 970–977.

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.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical Commercial Assays		
High Sensitivity Quantikine HS ELISA kit	R & D Systems	Cat#HS750
RNA extraction RNEasy kit	QIAGEN	Cat#74136
cDNA ABI kit	Life Technologies	Cat#4368814
qPCR assay Sybergreen Platinum	Life Technologies	Cat#11733-046
Deposited Data		
Summary statistics, annotation, and fine-mapping results	This paper	https://www.ebi.ac.uk/gwas/; http://www.mhi-humangenetics.org/en/resources
Experimental Models: Cell Lines		
Flip-In-293 cells	ThermoFisher Scientific	*Cat#R75007*
Oligonucleotides		
Primers for *IL7* cloning; see Table S5D	This paper	N/A
Primers for qPCR experiments; see Table S5D	This paper	N/A
Recombinant DNA		
pcDNA5/FRT	ThermoFisher Scientific	*Cat#V601020*
pOG44 FLP recombinase coding vector	ThermoFisher Scientific	Cat#V600520
Software and Algorithms		
checkVCF	N/A	https://genome.sph.umich.edu/wiki/CheckVCF.py/
eQTLgen	Võsa et al., 2018	https://www.eqtlgen.org/index.html
EPACTS	N/A	https://genome.sph.umich.edu/wiki/EPACTS
Imputation servers	McCarthy et al., 2016	http://imputationserver.sph.umich.edu/index.html
NCBI dbSNP	N/A	https://www.ncbi.nlm.nih.gov/projects/SNP/
PheKB	N/A	https://phekb.org/
Strand alignment resources	N/A	https://www.well.ox.ac.uk/~wrayner/strand/
UK Biobank SAIGE results	Zhou et al., 2018	http://pheweb.sph.umich.edu/SAIGE-UKB/
ENSEMBL Variant Effect Predictor	N/A	https://useast.ensembl.org/info/docs/tools/vep/index.html

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Guillaume Lettre (guillaume.lettre@umontreal.ca).

Material Availability

The reagents generated in this study are available without restriction.

Data and Code Availability

The genetic association results (summary statistics), functional annotations, and fine-mapping results are available at: http://www.mhi-humangenetics.org/en/resources and the GWAS catalog (https://www.ebi.ac.uk/gwas/).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Study design and participants

All participants provided written informed consent and the project was approved by each institution's ethical committee. Table S1B lists all participating cohorts. The SNPs we identified are available from the NCBI dbSNP database of short genetic variations (https://

www.ncbi.nlm.nih.gov/projects/SNP/). No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Cell lines

Flip-In^{TM}-293 cells (ThermoFisher Scientific) were grown at 80% confluency in DMEM medium supplemented with 10% Foetal Bovine Serum, 4 mM L-glutamine, 100 IU penicillin, 100 μg/ml streptomycin and 100 μg/ml hygromycin. 293 cells were likely originally derived from a female donor.

METHOD DETAILS

Phenotypes

Complete blood count (CBC) and related blood indices were analyzed as quantitative traits. The descriptive statistics for each phenotype in each cohort analyzed are in Table S1C. Exclusion criteria and phenotype modeling in UKBB (European-ancestry individuals), INTERVAL, and BBJ have been described previously (Astle et al., 2016; Kanai et al., 2018). For all other studies, we followed the protocol developed by the Blood-Cell Consortium (Chami et al., 2016; Eicher et al., 2016; Tajuddin et al., 2016). Briefly, we excluded when possible participants with blood cancer, acute medical/surgical illness, myelodysplastic syndrome, bone marrow transplant, congenital/hereditary anemia, HIV, end-stage kidney disease, splenectomy, and cirrhosis, as well as pregnant women and those undergoing chemotherapy or erythropoietin treatment. We also excluded extreme blood-cell measures: WBC > $200x10^9$ cells/L, HGB > 20 g/dL, HCT > 60%, and PLT > $1000x10^9$ cells/L. For WBC subtypes, we analyzed $\log_{10}$-transformed absolute counts obtained by multiplying relative counts with total WBC count. For all phenotypes in all studies, we corrected the blood-cell phenotypes for sex, age, age-squared, the first 10 principal components, and other cohort-specific covariates (e.g., recruitment center) using linear regression analysis. We applied rank-based inverse normal transformation to the residuals from the regression analysis and used the normalized residuals to test for association with genetic variants.

Genotype quality-control and imputation

The genotyping array and quality-control steps used by each cohort as well as their quality-control steps are listed in Table S1D. Unless otherwise specified, all studies applied the following criteria: samples were removed if the genotyping call rate was < 95%, if they showed excess heterozygosity, if we identified gender mismatches or sample duplicates, or if they appeared as population outliers in principal component analyses nested with continental populations from the 1000 Genomes Project (Abecasis et al., 2012). We removed monomorphic variants, as well as variants with Hardy-Weinberg $p < 1x10^{-6}$ and call rate < 98%.

Genotype imputation for the UKBB, INTERVAL, and BBJ have been described in details elsewhere (Astle et al., 2016; Bycroft et al., 2018; Kanai et al., 2018). For all other studies, unless specified in Table S1D, we applied the following steps for genotype imputation of autosomal variants. We aligned all alleles on the forward strand of build 37/hg19 of the human reference genome (https://www.well.ox.ac.uk/~wrayner/strand) and converted files into the VCF format. We then applied checkVCF (https://genome.sph.umich.edu/wiki/CheckVCF.py) to confirm strand and allele orientation. We carried out genotype imputation using the University of Michigan (http://imputationserver.sph.umich.edu/index.html) or the Sanger Institute (https://imputation.sanger.ac.uk/) imputation servers. We phased genotype data using SHAPEIT (Delaneau et al., 2013), EAGLE (Loh et al., 2016), or HAPI-UR (Williams et al., 2012). For populations of European ancestry, we used reference haplotypes from the Haplotype Reference Consortium (HRC r1.1 2016) for imputation (McCarthy et al., 2016) unless otherwise noted, whereas reference haplotypes from the 1000 Genomes Project (Phase 3, Version 5) (Abecasis et al., 2012) were used for non-European ancestry participants.

Study-level statistical analyses

We tested an additive genetic model of association between genotype imputation doses and inverse normal transformed blood-cell phenotypes. We analyzed the major ancestry groups (European (EUR), East Asian (EAS), African (AFR), Hispanic-American (HA), South Asian (SAS)) separately and used linear mixed-effect models implemented in BOLT-LMM (Loh et al., 2018), EPACTS (https://genome.sph.umich.edu/wiki/EPACTS), or EMMAX (Kang et al., 2010) to account for cryptic and known relatedness. Autosomal single nucleotide variants were analyzed in all contributing studies. For simplicity, we only analyzed insertion-deletion (indel) variants from UKBB and INTERVAL, since a similar reference panel was used for genotype imputation.

Centralized quality-control and meta-analyses

We performed a centralized quality-control check on the association results of each single study using EasyQC (v9.0) (Winkler et al., 2014). By mapping variants of each study to the appropriate ethnicity reference panel (HRC for EUR and 1000 Genomes Project Phase3 for non-EUR participants), we were able to harmonize alleles and markers across all studies. We were also able to assess the presence of flipped alleles per study and check for excessive allele frequency discrepancies using allele frequency reference data. We also inspected quantile-quantile (QQ) plots generated by EasyQC and the corresponding genomic inflation factors as well as SE-N plots (inverse of the median standard error versus the square root of the sample size) to evaluate potential issues with, for example, trait transformation or unaccounted relatedness. We removed variants with imputation quality metric (INFO score) $\leq$ 0.4. Except for three studies, we also removed variants with minor allele count (MAC) $\leq$ 5. For UKBB

EUR, Women Health Initiative (WHI), and GERA (EUR), we instead applied a MAC $\leq$ 20 filter because empirical observations suggested that unusual inflation of the test statistics (i.e., extreme effect sizes and standard errors) was due to rarer variants. To simplify handling of tri-allelic and indel variants, which have the same genomic coordinates but different alleles, we created a unique variant ID for each tested variant. Specifically, we assigned a chromosome:position(hg19)_allele1_allele2 unique ID to each variant, in which the order of the allele in the ID was based on the lexicographical order or the indel length. We performed inverse variance-weighted fixed-effect meta-analyses with GWAMA (v2.2.2) (Mägi and Morris, 2010) and trans-ethnic meta-analyses with MR-MEGA (v0.1.5)(Mägi et al., 2017). For MR-MEGA, we calculated four axes of genetic variation, the default recommendation, to separate global population groups.

Million Veteran Program (MVP) blood-cell trait analyses for replication

Phenotyping

Phenotyping methods published by the EMERGE Consortium and available on PheKB (https://phekb.org/) were used for retrieving lab data and exclusion criteria for all blood cell indices. This information was pulled from the VA electronic medical records for all MVP participants. Lab data was subject to the Boston Lab Adjudication Protocol. This entails five steps: (i) compile an initial spreadsheet of possible relevant lab tests, (ii) Subject Matter Expert (SME) does an initial review of possible tests, (iii) analyst adds relevant LOINC codes for SME review, (iv) second Subject Matter Expert (SME) review, (v) creation of a Lab Phenotype Table/Dataset. After restricting to only outpatient labs and applying the EMERGE exclusion criteria, for each trait and each person, the minimum, maximum, mean, median, SD, and number of labs was recorded. Values were compared to those from UKBB (Astle et al., 2016).

Genotyping

DNA extracted from whole blood was genotyped using a customized Affymetrix Axiom biobank array, the MVP 1.0 Genotyping Array. With 723,305 total DNA sequence variants, the array is enriched for both common and rare variants of clinical importance in different ethnic backgrounds (Klarin et al., 2018).

Analysis

The median lab value was the trait used for analysis. Linear regression models were run under an additive model in plink2 on 1000G (v3p5) imputed dosages. Analyses were run using models described above within each race/ethnicity stratum (AFR, ASN, EUR, HA) classified based on their genotype data using HARE (Fang et al., 2019). Meta-analyses for the trans-ethnic analyses were completed in METAL (Willer et al., 2010).

Heritabilities and genetic correlations

We calculated heritabilities and genetic correlations between blood-cell traits within the EUR and EAS populations using default parameters implemented in the LD score regression method (Figure S6; and Table S6A; Bulik-Sullivan et al., 2015a; Bulik-Sullivan et al., 2015b). For genetic correlation of the same phenotype between ancestral populations, we used Popcorn (Brown et al., 2016). Briefly, Popcorn uses a Bayesian framework to estimate, using genome-wide summary statistics, the genetic correlation of the same phenotype but in two different populations (in our case, between EUR and EAS). It reports the trans-ethnic genetic-effect correlation (ρ_{ge}), i.e., the correlation coefficient of per-allele SNP effect sizes, but also the trans-ethnic genetic impact correlation (ρ_{gi}), which includes a normalization of the effect based on allele frequency (Table S6B). To address whether a difference in the sample size for the EUR and EAS meta-analyses could impact the Popcorn results, we repeated our analyses using the current EAS results (N_{max} = 151,807) and EUR results from preliminary analyses of the UKBB dataset (N_{max} = 87,265) (Astle et al., 2016). These analyses confirmed that for common variants, cross-ancestry EUR-EAS genetic correlations are significantly different (but non null). Both LD score regression and Popcorn are not amenable to admixed populations, and cannot handle rare variants. For these reasons, we limited these analyses to the large EUR and EAS populations and focused on common variants (MAF $\geq$ 5%) from the 1000 Genomes Project.

Functional annotation

To derive basic functional annotation information, we annotated all variants included in 95% credible sets from ancestry-specific and trans-ethnic meta-analyses with the Variant Effect Predictor (VEP) (https://useast.ensembl.org/info/docs/tools/vep/index.html), compiling both all consequences and the most severe consequence for Ensembl/GENCODE transcripts. We also specifically annotated rare coding variants using VEP (defined as any variant with MAF < 1% in a given analysis, with a GC-corrected P value < 5×10^{-9}, and annotated as a missense_variant, stop_gained, stop_lost, splice_donor, or a splice_acceptor, regardless of fine-mapping results). We removed all variants with a GC-corrected P value < 5×10^{-9} in EUR, in the MHC region, and, in analyses including individuals with at least some African ancestry, on chromosome 1 for neutrophils and total WBC count and for RBC traits near the chromosome 11 β-globin and the chromosome 16 α-globin loci.

Bias-corrected enrichment of blood trait variants for chromatin accessibility of 18 hematopoietic populations was performed using g-chromVAR, which has been previously described in detail (Ulirsch et al., 2019). In brief, this method weights chromatin features by fine-mapped variant posterior probabilities and computes the enrichment for each cell type versus an empirical background matched

for GC content and feature intensity. For chromatin feature input, we used a consensus peak set for all hematopoietic cell types with a uniform width of 500 bp centered at the summit. For variant input, we included all fine-mapped variants within 95% credible sets of the trans-ethnic GWAS. We also ran g-chromVAR for each ancestry-specific meta-analysis, keeping all other parameters the same, but using fine-mapped variants with the 95% credible sets of each ancestry-specific study. Finally, to control for the number of loci tested within each ancestry-specific study, we first ranked the loci of the largest cohort (i.e., EUR) by sentinel variant p value, and then subset only the top *n* loci, where *n* equals the number of loci in the smaller cohort (e.g., EAS) for the same trait. We then ran g-chromVAR on the subset of variants falling within these top *n* loci.

Phenome-wide association study (pheWAS) analysis

UK Biobank (UKBB)

We extracted pheWAS results for a list of 5552 variants in UKBB ICD PheWeb hosted at the University of Michigan (Accessed 21 August 2019). To account for severe imbalance in case-control ratios, we selected the output from the SAIGE analyses (http://pheweb.sph.umich.edu/SAIGE-UKB/) based on 408,961 samples from White British participants (Zhou et al., 2018). In total, 1403 phecodes were tested for association. All results were downloaded using R, and were parsed and organized into data table format using the data.table, rvest, stringr, dplyr and tidyr packages.

BioBank Japan (BBJ)

We performed a pheWAS for the lead variants identified by the trans-ethnic meta-analyses. From the list of all the significantly associated variants with blood cell-related traits, we extracted those genotyped or imputed in the BBJ project (n_{SNP} = 4,255). Next, we curated the phenotype record of the disease status and clinical values for the same individuals analyzed in the discovery phase (n_{indiv} = 143,988). Then, we performed the logistic regression analyses for 22 binary traits (20 diseases and 2 behavioral habits) which had a sufficient number of case samples (n_{case} = 2,500). Regression models were adjusted for age, sex and 20 principal components as covariates. Trait-specific covariates are described elsewhere (Kanai et al., 2018).

BioVU

BioVU is the biobank of Vanderbilt University Medical Center (VUMC) that houses de-identified DNA samples linked to phenotypic data derived from electronic health records (EHRs) system of VUMC. The clinical information is updated every 1-3 months for the de-identified EHRs. Detailed description of program operations, ethical considerations, and continuing oversight and patient engagement have been published (Roden et al., 2008). DNA samples were genotyped with genome-wide arrays including the Multi-Ethnic Global **(**MEGA) array, and the genotype data were imputed into the HRC reference panel (McCarthy et al., 2016) using the Michigan imputation server (Das et al., 2016). Imputed data and the 1000 Genomes Project data were combined to carry out principal component analysis (PCA) and African-American samples were extracted for analysis based on the PCA plot. PheWAS were carried out for each SNP with the specified allele (Denny et al., 2010). Phenotypes were derived from billing codes of EHRs as described previously (Carroll et al., 2014). Each phenotype ('phecode') has defined case, control and exclusion criteria. We required two codes on different visit days to instantiate a case for each phecode. In total, 1815 phecodes were tested for association. Association between each binary phecode and a SNP was assessed using logistic regression, while adjusting for covariates of age, sex, genotyping array type/batch and 10 principal components of ancestry.

Merging across biobanks

We defined statistical significance within each biobank to be a Bonferroni corrected level of $0.05/pq$, where p is the number of phecodes tested and q is the number of variants tested. We considered an association to be replicated if the p value for the association was $< 0.05/s$ with a consistent direction of effect, where s represents the number of associations being replicated. To match phenotypes across biobanks, we merged the UKBB and BioVU by phecode, as these two biobanks used the same phecode system for classifying outcomes. To match with BBJ, we cross-referenced the 22 outcomes in BBJ with the phecode library used by BioVU/UKBB. Matches were determined based on phenotype similarity between the BioVU/UKBB phenotype description and the outcomes described in Nagai et al. (2017).

Power analysis

For all variant-disease associations that failed to replicate, we performed power analyses in the replication biobank to determine if the lack of replication was likely due to lack of signal or lack of statistical power. We assumed that the replication biobank would have the same prevalence and odds-ratio as the biobank in which the association was discovered, and we used allele frequencies from the matching population in the 1000 Genomes project. To guard against winner's curse in our power analyses, we assumed a maximum odds-ratio of 3. Power was assessed at a P value threshold of $0.05/s$, where s represents the number of associations being tested for replication.

Polygenic trait score (PTS) analyses

We restricted these analyses to variant-trait associations that reached genome-wide significance ($p < 5x10^{-9}$) in the trans-ethnic MR-MEGA meta-analyses (Table S3A). For each of these variant-trait pairs, we calculated an effect size – hereafter referred to as trans weights – using the fixed-effect meta-analysis method implemented in GWAMA and all cohorts available (Mägi and Morris, 2010). For the same variants, we also retrieved the ancestry-specific effect sizes (or weights). We calculated the PTS using plink2 by summing up the number of trait-increasing alleles (or imputation doses) that were weighted by their corresponding trans (PTS_{trans}) or ancestry-specific (PTS_{EUR}, PTS_{AFR}, PTS_{HA}) weights. The variance explained by the PTS on corrected and normalized blood-cell traits was

calculated in R using linear regression. For these analyses, we had access to 2,651 AFR, 5,048 EUR and 4,281 HA BioMe participants that were not used in the discovery effort. For the analyses of hematological diseases, we used the same independent BioMe participants and implemented logistic regression models in R. We used age, sex, and the first four principal components as covariates in all models. We used the PredictABEL package in R to calculate precision metrics. We used the following thresholds to define disease state: anemia (women < 12 g/dL HGB, men < 13 g/dL HGB), neutropenia (< 1500 NEU/μL), thrombocytosis (> 450x10^9 PLT/L), and thrombocytopenia (< 150x10^9 PLT/L).

Analysis of natural selection

To quantify the contribution of positive selection on blood-cell trait variation, we used the recent map of selective sweeps identified in the different populations of the 1000 Genomes Project (Johnson and Voight, 2018). We grouped the sweeps identified in the 26 1000 Genomes Project populations into five larger populations that correspond to our ancestry-specific meta-analyses: Europe-ancestry (CEU, TSI, GBR, FIN, IBS); East-Asian-ancestry (CHB, JPT, CHS, CDX, KHV); African-ancestry (YRI, LWK, GWD, MSL, ESN, ASW, ACB); South-Asian-ancestry (GIH, PJL, BEB, STU, ITU); and Hispanic/Latino-ancestry (MXL, PUR, CLM, PEL). Following the nomenclature by Johnson and Voight (2018), each selective sweep is summarized by the variant located within the sweep that has the highest iHS value. iHS (Integrated Haplotype Score) is a statistic to quantify evidence of recent positive selection. A high positive iHS score (iHS > 2) means that haplotypes on the ancestral allele background are longer compared to derived allele background. A high negative iHS score (iHS < −2) means that the haplotypes on the derived allele background are longer compared to the haplotypes associated with the ancestral allele. We retrieved the blood-cell trait association results for these sweep-tagging SNPs from the ancestry-specific meta-analyses (Table S7A). To determine if the inflation observed in the QQ plots was significant, we generated 100 sets of SNPs that match the selective sweep-tagging SNPs based on allele frequency, gene proximity, and the number of LD proxies in European-ancestry, East-Asian-ancestry and African-ancestry individuals using SNPsnap (Pers et al., 2015). For these analyses, we excluded the HLA region and variants in LD (r^2 > 0.5). We computed empirical significance by tallying the number of sets with the same or more genome-wide significant variants than the canonical sets of selective sweep-tagging SNPs (Table S7B).

We also computed the population branch statistic (PBS) using whole-genome sequencing information from the 1000 Genomes Project (Yi et al., 2010). PBS measures the amount of allele frequency change in the population since its divergence from the other two populations. For a target population, PBS is calculated as:

$$PBS = \frac{T^{target,\ sister} + T^{target,\ outgroup} - T^{sister,\ outgroup}}{2}$$

where $T = -log(1 - F_{ST})$ is an estimate of the divergence time between two populations. Here, F_{ST} between each pair of populations was estimated using Weir and Cockerham's estimate (Weir and Cockerham, 1984). We then divided all variants with calculated PBS into 50 bins of equal size by derived allele count in the target population, and then standardized the raw PBS values within each bin. To calculate PBS for Europe-ancestry (CEU, TSI, GBR, and IBS, without FIN), we used YRI as an outgroup and East-Asian-ancestry (CHB, JPT, CHS, CDX, KHV) as a sister population; for East-Asian-ancestry, we used YRI as an outgroup and Europe-ancestry as a sister population; for YRI, we used East-Asian-ancestry as an outgroup and Europe-ancestry as a sister population.

Replication of the association between *IL7*-rs201412253 and lymphocyte count in Genes & Health

Genes & Health is a population cohort study of British-Bangladeshi and British-Pakistani adult volunteers recruited from London and Bradford UK (http://www.genesandhealth.org; Finer et al., 2020). Participant saliva DNA samples (Oragene, DNA Genotek) were genotyped on the Illumina GSAMD-24v3-0-EA genotyping chip. Several rounds of data filtering and quality control were undertaken in Genome Studio using cluster separation scores (< 0.57), Gentrain score (≤0.7) and with increasingly stringent per-variant call rate threshold across remaining samples, and per-sample call rate threshold across remaining variants. Final dataset had call rate of > 0.992 per female-, and > 0.995 per male-sample across all 637,829 variants (which included Y chromosome). PLINK gender calls were compared to self-stated questionnaire gender information and where discordant, samples were removed from analyses. For individuals that had taken part on multiple occasions the sample with highest call rate was retained, while all samples were removed for an individual if duplicate samples were not concordant. Where exome data was available, sample genotypes were compared across platforms and highly discordant samples removed for further work.

Genome-wide imputation using the genotype chip data was carried out on the Michigan Imputation Server using reference panel Genome Asia Pilot (GAsP). This panel performed better than other available reference panels in the south Asian samples. Variants with minimac4 imputation Rsq < 0.3 were removed, as were variants with MAF < 0.1%.

Genotyped volunteer samples with Barts Health NHS Trust hospital clinical pathology laboratory full blood count data - to obtain lymphocyte count data – were selected. This included tests ordered on hospital patients, and also from primary care GP surgeries using the hospital laboratory. We split the data into Pakistani and Bangladeshi populations based on those samples with complete DNA genotype principal component and questionnaire ethnicity agreement (N = 5,912 Pakistani Individuals, N = 13,611 Bangladeshi Individuals). Intersex individuals, and related individuals (one from each pair of samples with piHat > 0.1875) were removed to leave 4,554 Pakistani and 10,638 Bangladeshi samples.

Absolute lymphocyte counts (x10^9 cells/L) were extracted from Barts Health NHS Trust pathology data warehouse. The median count, and age at test for that measurement, were taken when multiple measurements were available on an individual. Log10 transformation of cell counts was undertaken in RStudio(v1.1.453), before correcting for median age at test, median age at test squared and gender using linear regression analysis on each population separately. Residuals from the regression analyses were extracted and rank-based inverse normalization was performed. These normalized residuals were used as the phenotype in association analysis which was undertaken in PLINK2.0 (–glm) using bgen files from Imputation and only default settings. Pakistani and Bangladeshi populations were analyzed separately.

IL7 functional analyses

We PCR amplified and cloned the *IL7* wild-type (rs201412253-Val18) and mutant (rs201412253-18Ile) open reading frame (ORF) in the pcDNA5/FRT vector (ThermoFisher Scientific) using HindIII and BamHI restriction sites (see Table S5D for ORF and primer sequences). We validated the sequences of the two plasmids by Sanger Sequencing. Flip-In^{TM}-293 cells (ThermoFisher Scientific) at 80% confluency were transfected with 1:10 mixes of empty pcDNA5 or pcDNA5 derivatives coding for IL7-Val8 or IL7-18Ile and pOG44 FLP recombinase coding vector (ThermoFisher Scientific) using polyethylenimine. Transfectant clones were expanded and selected in DMEM medium supplemented with 10% Foetal Bovine Serum, 4 mM L-glutamine, 100 IU penicillin, 100 μg/ml streptomycin and 100 μg/ml hygromycin. We measured the secretion of IL7 in eight independent clones for each *IL7* allele (rs201412253-Val18 and rs201412253-18Ile) as well as in four clones generated with the empty vector by ELISA assay. We used the High Sensitivity Quantikine HS ELISA kit from R & D Systems (Cat # HS750). We seeded 100,000 cells per 12-wells plates and grew them for 6 days in DMEM glutamax plus 10% FBS before doing the ELISA. We measured each supernatant in duplicate and seeded each of the clones in triplicate. The whole experiment was done on three different weeks (three complete biological replicates). We extracted total proteins from cells with RIPA buffer and we quantified the lysates by BCA. We used this quantification to normalize the ELISA assays. We extracted total RNA from ~500,000 cells using the QIAGEN RNEasy kit (cat # 74136). We checked the quality of the RNA by Bioanalyzer and quantified its concentration by Nanodrop. We reverse transcribed 1 ug of total RNA into cDNA using the ABI kit (Life Technologies Cat # 4368814). We used two pairs of primers for *IL7* and assays for three normalizing genes (*HPRT*, *GAPDH*, *TBP*; Table S5D). We followed the MIQE recommendations and performed the qPCR reactions with the Sybergreen Platinum (Life Technologies Cat # 11733-046) on a Biorad CFX384 thermocycler.

QUANTIFICATION AND STATISTICAL ANALYSES

Statistical significance, genomic inflation and locus definition

For each meta-analysis, we calculated the genomic inflation factor (λ_{GC}) for all variants, which were modest when considering the large sample sizes (λ_{GC} range: 0.9-1.2) (Table S2). We used $\alpha \leq 5x10^{-9}$ after GC-correction to declare statistical significance, accounting for the inflation of the test statistics and the number of blood-cell traits analyzed. To count the number of loci that we discovered, we first identified the most significant variants (with $p \leq 5x10^{-9}$) and extended the physical region around that variant 250-kb on each side. Overlapping loci were merged, and we used the most significant variant within the interval as the sentinel variant. In this manuscript, we defined as novel a locus if no variants were previously reported in the literature to be associated with the specific blood-cell trait and if the locus is not reported in the companion manuscript that focuses on EUR-specific genetic discoveries.

Conditional analyses in the UK Biobank European-ancestry population to identify independent variants associated with blood-cell traits

This method is described in details in the companion manuscript (Vuckovic et al., 2020). Briefly, we applied the following four steps: (1) Initialisation step: From the list of all variants in the block, add the variant with the lowest P value that is also below the significance threshold ($8.31x10^{-9}$). (2) Dropping: Study the P values for all variants in the model, if any of these are above the significance threshold we iteratively prune and rebuild model starting with the variant with the highest P value. Once a variant is pruned it is returned to the list of variants not currently in the parsimonious model and may rejoin at a later iteration. (3) Addition: Test each variant not currently in the block sequentially in the model, add the variant with the lowest P value which is below the threshold. Any tested variants which have a P value of higher than 0.01 are not tested again in future iterations. Variants are not permitted to be tested in the model if they have a LD $r^2 > 0.9$ with any variant currently in the model. (4) Completion: If the algorithm could neither add a variant into the model nor remove a variant from the model then we abort the iteration with the model at this stage representing the parsimonious model for this block. Following identification of conditionally significant variants in each block, all conditionally significant variants within each chromosome are put into a single linear model and tested with the same multiple stepwise linear regression algorithm as that defined above. The resultant set is the 'conditionally significant' list of variants for the blood cell index. Full results from these conditional analyses are described in the companion European focused manuscript. We will note that this conditional analysis model for selecting loci for fine-mapping would not allow for the detection of non-European ancestry specific secondary signals, with these direct conditional analyses only feasible at most loci in a very large single cohort like the UK Biobank.

Statistical fine-mapping

No fine-mapping methods currently exist to handle admixed populations. Furthermore, for some of the ethnic groups analyzed here, we did not have access to a sufficiently large reference panel to properly account for LD, complicating conditional analyses and fine-mapping efforts. For these reasons, we fine-mapped the ancestry-specific fixed-effect meta-analyses by adapting the method proposed by Maller et al. (2012) in order to assign posterior probability of inclusion (PIP) to each variant and construct 95% credible sets.

This method makes the strong assumption that there is a single independent causal variant at the tested locus. For this reason, we limited our Bayesian fine-mapping to loci where we identified a single independent association signal by conditional analysis in EUR individuals from the UKBB (Vuckovic et al., 2020). Because EUR represented the largest group, we then inferred that there was also a single association signal in the other populations at these loci, an inference that may not always be right. Briefly, we added 250-kb on either side of genome-wide significant variants ($p < 5x10^{-9}$) and merged loci when they overlapped. For the loci identified in the ancestry-specific meta-analyses, we converted summary statistics into approximate Bayes factors (aBF) using (Wakefield, 2009; Maller et al., 2012):

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

where β and SE are the variant's effect size and standard error, respectively, and ω denotes the prior variance in allelic effects, taken here to be 0.04 (Wakefield, 2007). For the trans-ethnic results, we directly used Bayes factors calculated by MR-MEGA (Mägi et al., 2017). We calculated PIP of each variant by dividing the variant's aBF by the sum of the aBF for all the variants within the locus. We generated the 95% credible sets by ordering all variants in a given locus from the largest to the smallest PIP and by including variants until the cumulative sum of the PIP $\geq$ 95% (Mahajan et al., 2018). All variants that map to 95% credible sets are available online (http://www.mhi-humangenetics.org/en/resources).

Supplemental Figures

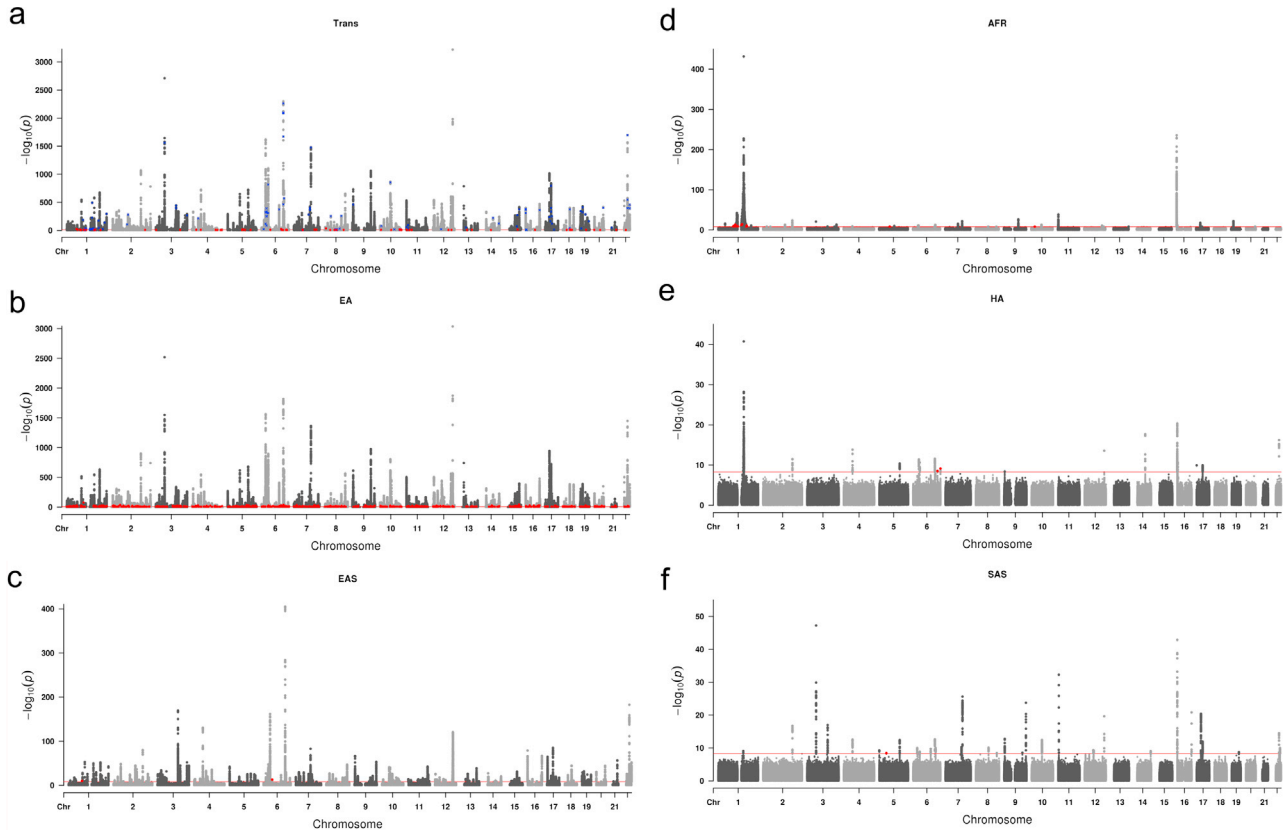


Figure S1. Manhattan Plots of Association Results in Each Population and in the Trans-ethnic Meta-analyses, Related to Figure 1 and STAR Methods

At each variant, we report the smallest P value across all 15 traits analyzed. In blue and red, we highlight novel loci and loci with heterogeneity P value < $5x10^{-9}$ respectively. EUR, European-ancestry; EAS, East-Asian-ancestry; HA, Hispanic-American-ancestry; AFR, African-ancestry; SAS, South-Asian ancestry.

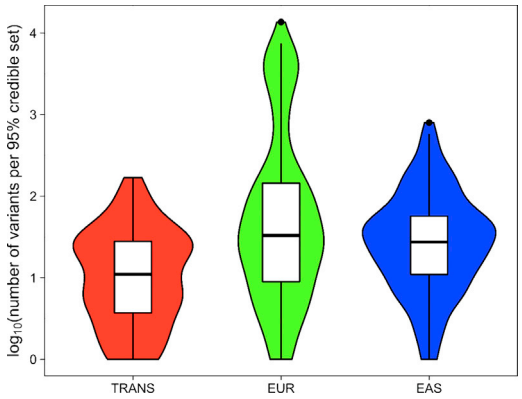


Figure S2. Distribution of the Number of Variants in the 95% Credible Sets for Three Approximately Equal-Sized Meta-analyses for trans-ethnic (N = 137,702), EUR-only (N = 141,636), and EAS-only (N = 143,085) for RBC, PLT, and WBC Counts and Hemoglobin Levels, Related to Figure 2 and STAR Methods

We restricted to loci with evidence of a single association signal based on conditional analyses in EUR populations. The 95% credible sets in the trans-ethnic meta-analyses (median (interquartile range) number of variants per 95% credible set = 11 (4-38)) are smaller than in the EUR (33 (9-144), Wilcoxon's $p = 1.4x10^{-11}$) and the EAS meta-analyses (27 (11-57), Wilcoxon's $p = 8.1x10^{-8}$).

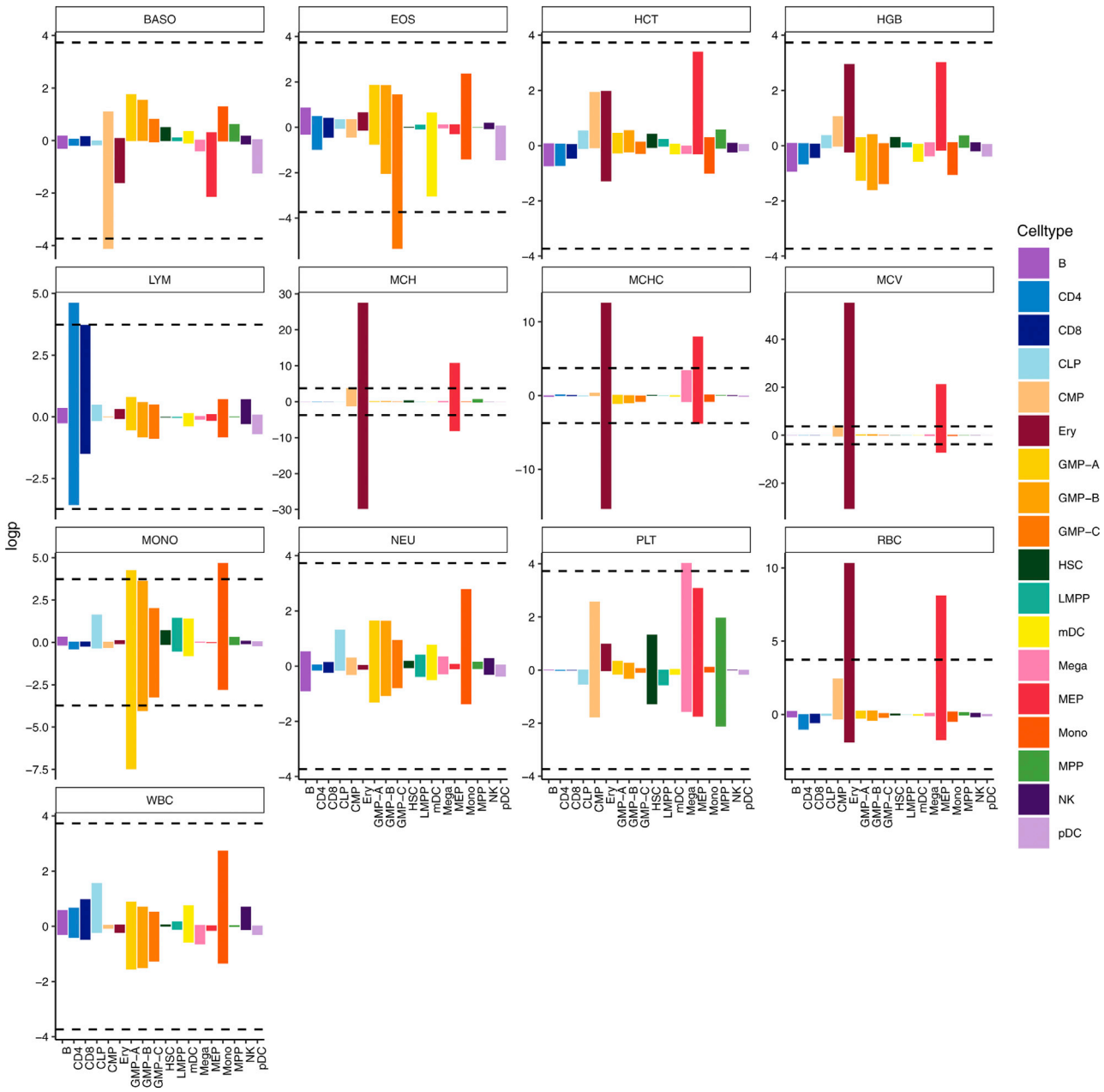


Figure S3. Miami Plots Contrasting Cell Type Enrichments of 95% Credible Set Variants for 13 Hematological Traits Studied in the EUR and EAS GWASs, Computed Using g-chromVAR, Related to Figure 3 and STAR Methods

The Bonferroni-adjusted significance level (one sided z-test) is indicated by the dotted line. mono, monocyte; gran, granulocyte; ery, erythroid; mega, megakaryocyte; CD4, $CD4^+$ T cell; CD8, $CD8^+$ T cell; B, B cell; NK, natural killer cell; mDC, myeloid dendritic cell; pDC, plasmacytoid dendritic cell; MPP, multipotent progenitor; LMPP, lymphoid-primed multipotent progenitor; CMP, common myeloid progenitor; CLP, common lymphoid progenitor; GMP, granulocyte–macrophage progenitor; MEP, megakaryocyte–erythroid progenitor.

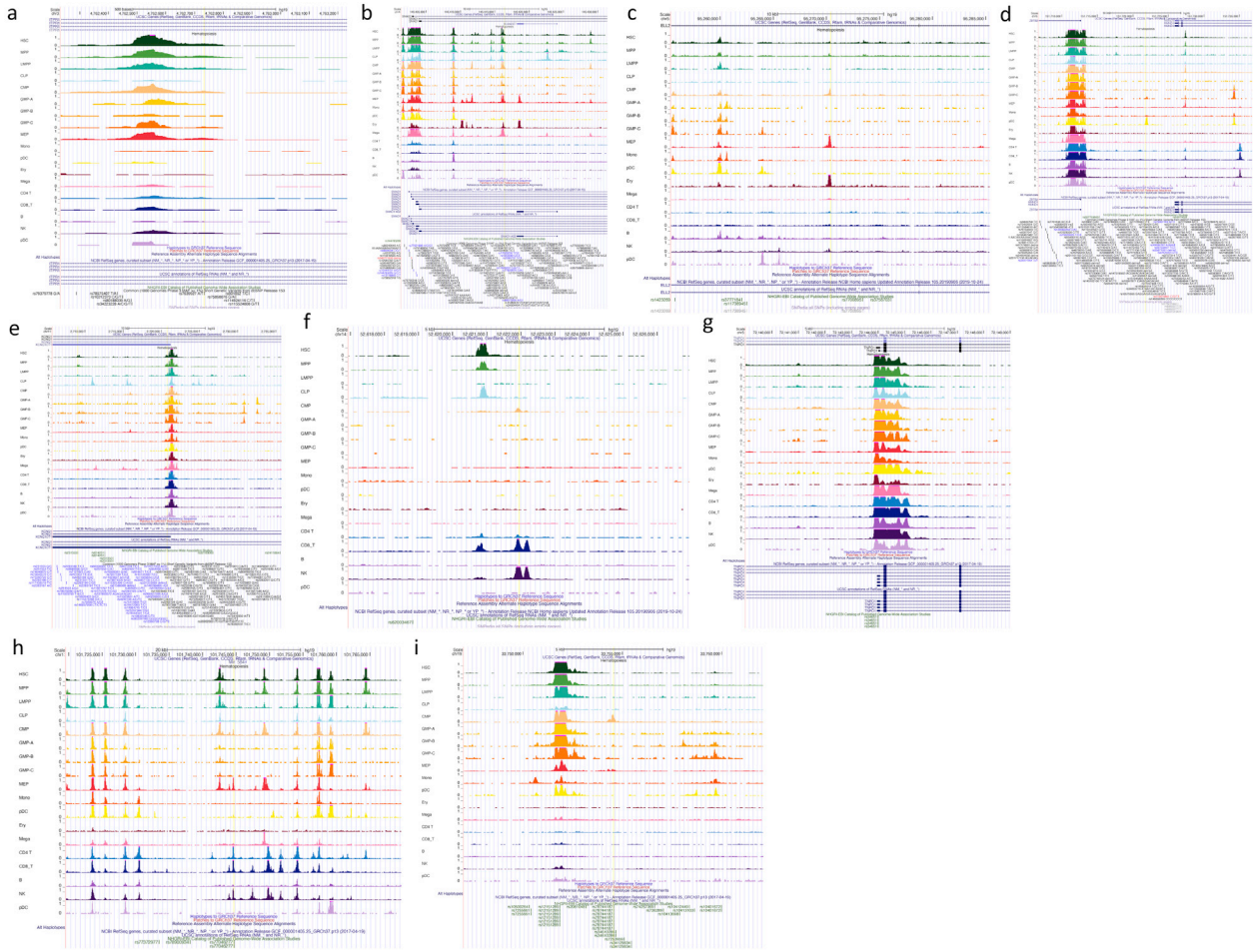


Figure S4. Functional Annotation of Top Novel Variants Identified in the Trans-ethnic Meta-analyses, Related to Figures 3 and 6 and STAR Methods

Genomic landscapes are from build hg19 on the UCSC Genome Browser. ATAC-seq peaks from different hematopoietic cell types are from Corces et al. (2016). In all six plots (A–F), the vertical yellow line indicates the location of the top variant. See text and Table S6I for details. (a) rs7639927 associated with lymphocyte counts. (b) rs6537356 associated with mean corpuscular hemoglobin (MCH). (c) rs115906455 associated with mean cell volume (MCV). (d) rs7771156 associated with white blood cell counts. (e) rs368427206 associated with mean platelet volume (MPV). (f) rs941616 associated with eosinophil counts. Variants with high posterior inclusion probability (PIP) and evidence of differentiation in effect size across populations that map to ATAC-seq open chromatin regions identified in progenitor and mature hematopoietic cells (Corces et al., 2016). In all cases, the variant of interest is located in the middle of the plot. (g) rs34651 is associated with platelet counts. (h) rs77046277 is associated with lymphocyte counts. (i) rs78744187 is associated with basophil counts.

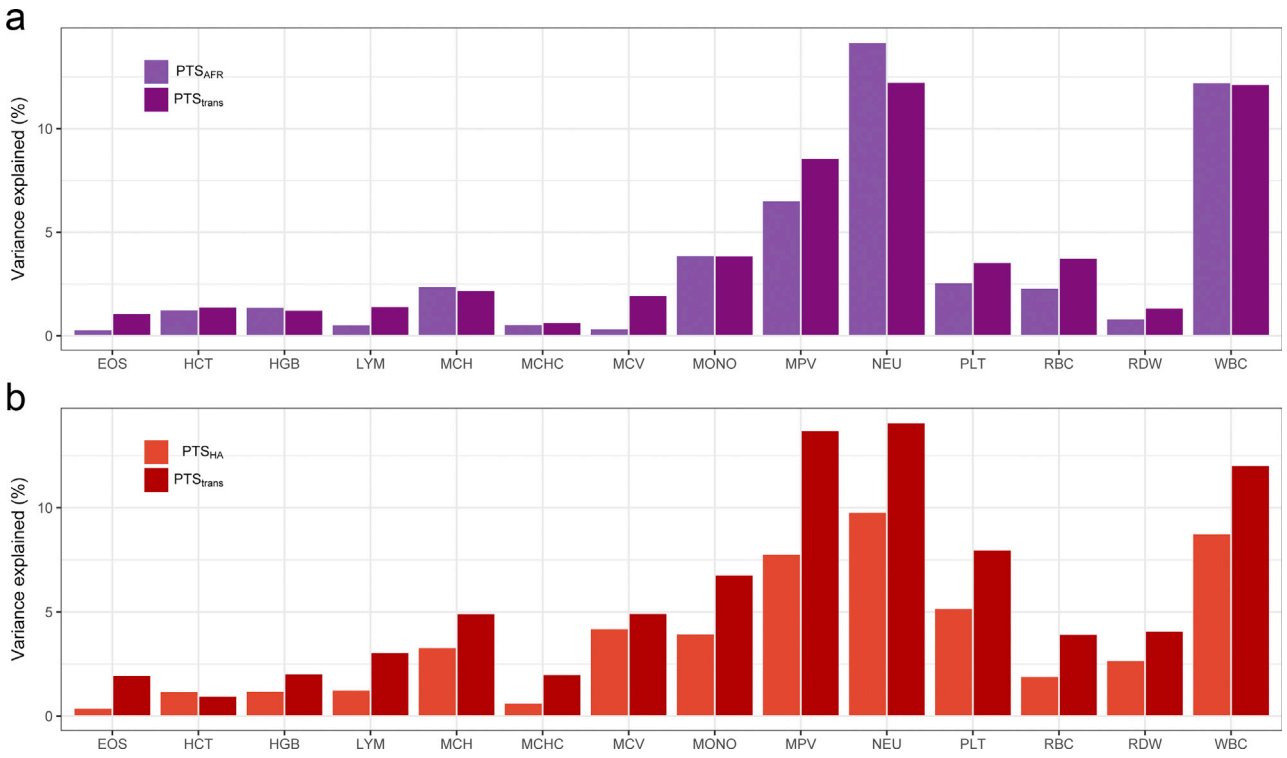


Figure S5. Phenotypic Variance Explained by PTS in Independent Participants from the BioMe Biobank, Related to Figure 4 and STAR Methods

Basophil counts were not tested as the PTS were not significant in any of the BioMe populations. (a) In AFR BioMe participants, we compared the variance explained by PTS_{trans} or PTS_{AFR}, a polygenic predictor calculated using the same trans-ethnic genome-wide significant variants, but weighted with AFR-specific effect sizes. (b) As for a, but for HA BioMe participants and using HA-specific effect sizes to weight PTS_{HA}.

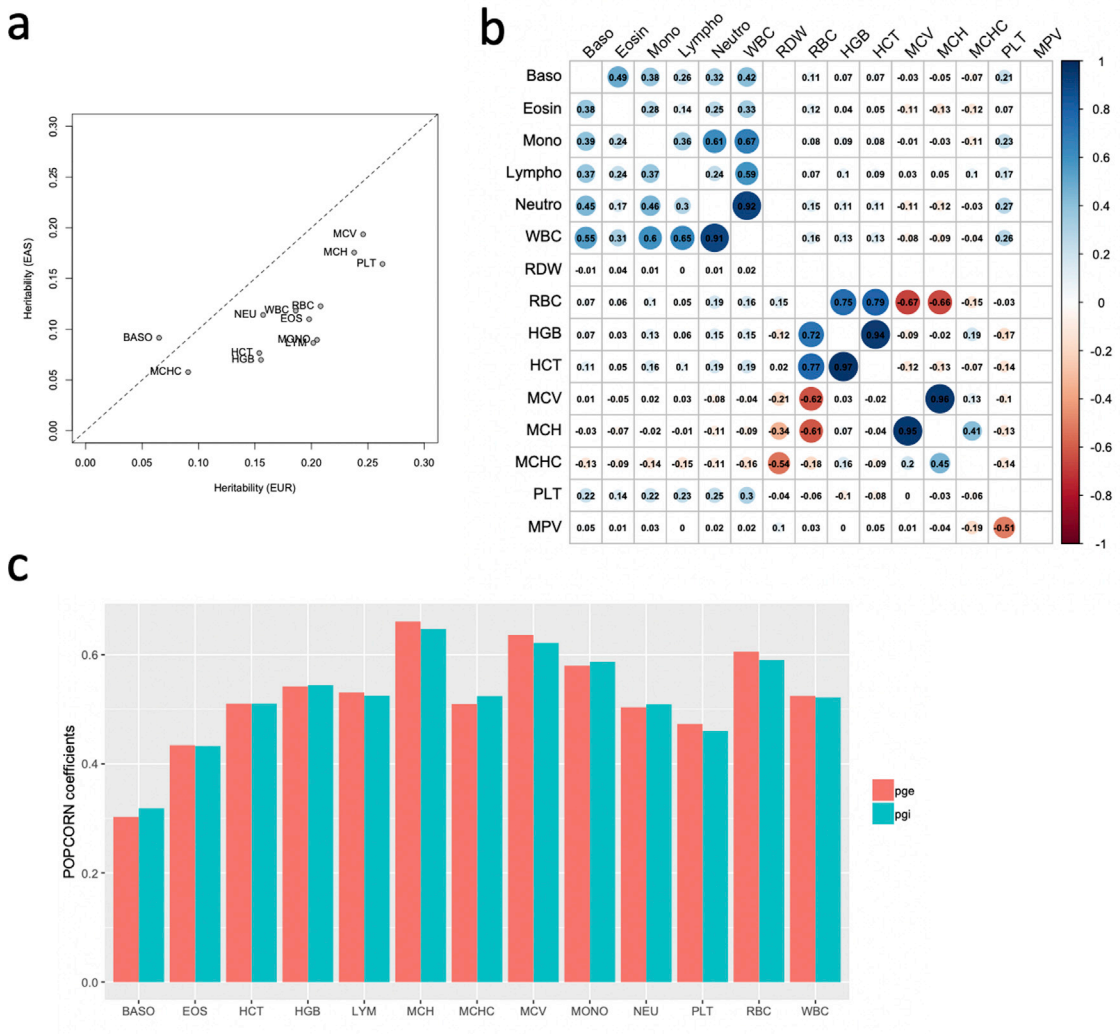


Figure S6. Heritabilities and Genetic Correlations of Blood Cell Traits in EUR and EAS Populations, Related to STAR Methods

(a) Heritabilities estimated using common variants and the linkage disequilibrium (LD) score regression method. (b) Genetic correlations estimated using LD score regression between each pairs of phenotypes within ancestry. The number in each cell correspond to the genetic correlation coefficient (r_g) between the pair of traits analyzed. Results over the diagonal (right side of the square) are for EAS whereas results under the diagonal (left side of the square) are for European-ancestry EUR individuals. We note that the results on one side of the diagonal form an almost perfect mirror image of the results on the other side, indicating that the genetic correlations between pairs of blood-cell traits are very similar between EAS and EA. RDW and MPV results were not available in EAS. (c) Genetic correlations for each blood-cell trait estimated between EUR and EAS using Popcorn (Brown et al., 2016). pge is the correlation coefficient of per-allele SNP effect sizes, whereas pgi is the genetic impact correlation, which includes a normalization of the effect based on allele frequency.

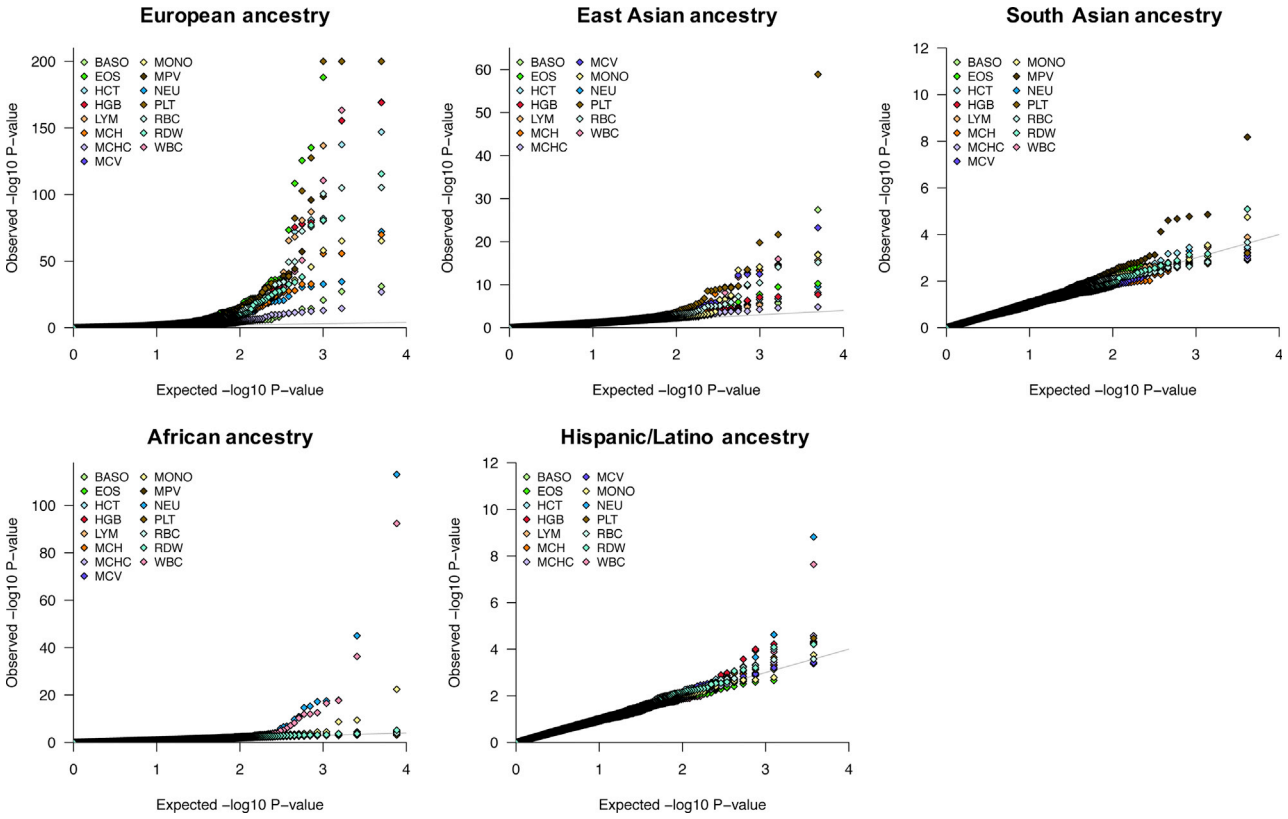


Figure S7. Quantile-Quantile Plots of SNPs that "Tag" Selective Sweeps in 1000G Populations, Related to Figure 7 and STAR Methods

For each of these variants, we retrieved the corresponding blood-cell trait association results from the ancestry-specific meta-analyses.