Open access • Posted Content • DOI:10.1101/2020.12.09.20246736

Whole genome sequencing association analysis of quantitative red blood cell phenotypes: the NHLBI TOPMed program — Source link

Yao Hu, Adrienne M. Stilp, Caitlin P. McHugh, Deepti Jain ...+77 more authors

Institutions: Fred Hutchinson Cancer Research Center, University of Washington, University of Minnesota, Harvard University ...+27 more institutions

Published on: 11 Dec 2020 - medRxiv (Cold Spring Harbor Laboratory Press)

Topics: Genome-wide association study, Quantitative trait locus, Population, Genetic association and Genetic architecture

Whole genome sequencing association analysis of quantitative red blood cell phenotypes: the NHLBI TOPMed program

Yao Hu* [1], Adrienne M. Stilp* [2], Caitlin P. McHugh* [2], Shuquan Rao* [4], Deepti Jain [2], Xiuwen Zheng [2], John Lane [3], Sébastian Méric de Bellefon [5], Laura M. Raffield [6], Ming-Huei Chen [7,8], Lisa R. Yanek [9], Marsha Wheeler [10], Yao Yao[4], Chunyan Ren[4], Jai Broome [2], Jee-Young Moon [11], Paul S. de Vries [12], Brian D. Hobbs [13], Quan Sun [14], Praveen Surendran [15,16,17,18], Jennifer A. Brody [19], Thomas W. Blackwell [20], Hélène Choquet [21], Kathleen Ryan [22], Ravindranath Duggirala [23], Nancy Heard-Costa [6,8,24], Zhe Wang [25], Nathalie Chami [25], Michael H. Preuss [25], Nancy Min [26], Lynette Ekunwe [26], Leslie A. Lange [27], Mary Cushman [28], Nauder Faraday [29], Joanne E. Curran [23], Laura Almasy [30], Kousik Kundu [31,32], Albert V. Smith [20], Stacey Gabriel [33], Jerome I. Rotter [34], Myriam Fornage [35], Donald M. Lloyd-Jones [36], Ramachandran S. Vasan [8,37,38], Nicholas L. Smith [39,40,41], Kari E. North [42], Eric Boerwinkle [12], Lewis C. Becker [43], Joshua P. Lewis [22], Goncalo R. Abecasis [20], Lifang Hou [36],Jeffrey R. O'Connell [22], Alanna C. Morrison [12], Terri H. Beaty [44], Robert Kaplan [11], Adolfo Correa [26], John Blangero [23], Eric Jorgenson [21], Bruce M. Psaty [39,40,45], Charles Kooperberg [1], Russell T. Walton [46], Benjamin P. Kleinstiver [46,47], Hua Tang [48], Ruth J.F. Loos [25], Nicole Soranzo [16,31,32,49], Adam S. Butterworth [15,16,17,49,50], Debbie Nickerson [10], Stephen S. Rich [51], Braxton D. Mitchell [22], Andrew D. Johnson [7,8], Paul L. Auer [52], Yun Li [53], Rasika A. Mathias [54], Guillaume Lettre [5,55], Nathan Pankratz [3], Cathy C. Laurie [2], Cecelia A. Laurie [2], Daniel E. Bauer [4], Matthew P. Conomos [2], Alexander P. Reiner [39], the NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium

1 Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98105, USA

2 Department of Biostatistics, University of Washington, Seattle, WA 98105, USA

3 Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, Minneapolis, MN 55455, USA

4 Division of Hematology/Oncology, Boston Children's Hospital, Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Stem Cell Institute, Broad Institute, Department of Pediatrics, Harvard Medical School, Boston, MA 02215, USA

5 Montreal Heart Institute, Montréal, Québec H1T 1C8, Canada

6 Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA

7 Population Sciences Branch, Division of Intramural Research, National Heart, Lung and Blood Institute, Bethesda, MD 20892, USA

8 National Heart Lung and Blood Institute's and Boston University's Framingham Heart Study, Framingham, MA 01701, USA

9 Division of General Internal Medicine, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

10 Department of Genome Sciences, University of Washington, Seattle, WA 98105, USA

11 Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY 10461, USA

12 Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas 77030, USA

13 Channing Division of Network Medicine and Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

14 Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

15 British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK

16 British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge CB1 8RN, UK

17 Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge CB1 8RN, UK

18 Rutherford Fund Fellow, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK

19 Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA 98105, USA

20 TOPMed Informatics Research Center, University of Michigan, Department of Biostatistics, Ann Arbor, MI 48109, USA

21 Division of Research, Kaiser Permanente Northern California, Oakland, CA 94601, USA

22 Department of Medicine, Division of Endocrinology, Diabetes & Nutrition, University of Maryland School of Medicine, Baltimore, MD 21201, USA

23 Department of Human Genetics and South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX 78539, USA

24 Department of Neurology, Boston University School of Medicine, Boston, MA 02118, USA

25 The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York City, NY 10029, USA

26 Department of Medicine, University of Mississippi Medical Center, Jackson, MS 39216, USA

27 Division of Biomedical Informatics and Personalized Medicine, School of Medicine University of Colorado, Anschutz Medical Campus, Aurora, CO 80045, USA

28 Department of Medicine, Larner College of Medicine at the University of Vermont, Burlington, VT 05405, USA

29 Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

30 Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia and Department of Genetics University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA

31 Department of Human Genetics, Wellcome Sanger Institute, Hinxton CB10 1SA, UK

32 Department of Haematology, University of Cambridge, Cambridge CB2 0PT, UK

33 Broad Institute, Boston, MA 02142, USA

34 The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA 90502, USA

35 University of Texas Health Science Center at Houston, Houston, TX 77030, USA

36 Northwestern University, Chicago, IL 60208, USA

37 Departments of Cardiology and Preventive Medicine, Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA

38 Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA

39 Department of Epidemiology, University of Washington, Seattle, WA 98105, USA

40 Kaiser Permanente Washington Health Research Institute, Kaiser Permanente Washington, Seattle, WA 98105, USA

41 Seattle Epidemiologic Research and Information Center, Department of Veterans Affairs Office of Research and Development, Seattle, WA 98105, USA

42 Department of Epidemiology, Gillings School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

43 Division of Cardiology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

44 School of Public Health, John Hopkins University, Baltimore, MD 21205, USA

45 Department of Medicine, University of Washington, Seattle, WA 98105, USA

46 Center for Genomic Medicine and Department of Pathology, Massachusetts General Hospital, Boston, MA 02114, USA

47 Department of Pathology, Harvard Medical School, Boston, MA 02115, USA

48 Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

49 National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Genomics, University of Cambridge, Cambridge CB1 8RN, UK

50 National Institute for Health Research Cambridge Biomedical Research Centre, University of Cambridge and Cambridge University Hospitals, Cambridge CB1 8RN, UK

51 Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia School of Medicine, Charlottesville, VA 22903, USA

52 Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI 53205, USA

53 Departments of Biostatistics, Genetics, Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

54 Division of Allergy and Clinical Immunology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MA 21205, USA

55 Faculté de Médecine, Université de Montréal, Montréal, Québec H1T 1C8, Canada

* These authors contributed equally to this work.

Abstract

Whole genome sequencing (WGS), a powerful tool for detecting novel coding and non-coding disease-causing variants, has largely been applied to clinical diagnosis of inherited disorders. Here we leveraged WGS data in up to 62,653 ethnically diverse participants from the NHLBI Trans-Omics for Precision Medicine (TOPMed) program and assessed statistical association of variants with seven red blood cell (RBC) quantitative traits. We discovered 14 single variant-RBC trait associations at 12 genomic loci. Several of the RBC trait-variant associations (*RPN1*, *ELL2*, *MIDN*, *HBB*, *HBA1*, *PIEZO1*, *G6PD*) were replicated in independent GWAS datasets imputed to the TOPMed reference panel. Most of these newly discovered variants are rare/low frequency, and several are observed disproportionately among non-European Ancestry (African, Hispanic/Latino, or East Asian) populations. We identified a 3bp indel p.Lys2169del (common only in the Ashkenazi Jewish population) of *PIEZO1*, a gene responsible for the Mendelian red cell disorder hereditary xerocytosis [OMIM 194380], associated with higher MCHC. In stepwise conditional analysis and in gene-based rare variant aggregated association analysis, we identified several of the variants in *HBB, HBA1, TMPRSS6,* and *G6PD* that represent the carrier state for known coding, promoter, or splice site loss-of-function variants that cause inherited RBC disorders. Finally, we applied base and nuclease editing to demonstrate that the sentinel variant rs112097551 (nearest gene *RPN1*) acts through a cis-regulatory element that exerts long-range control of the gene *RUVBL1* which is essential for hematopoiesis. Together, these results demonstrate the utility of WGS in ethnically-diverse population-based samples and gene editing for expanding knowledge of the genetic architecture of quantitative hematologic traits and suggest a continuum between complex trait and Mendelian red cell disorders.

Introduction

Red blood cells (RBCs) or erythrocytes contain hemoglobin, an iron-rich tetramer composed of two alpha-globin and two beta-globin chains. RBCs play an essential role in oxygen transport and also serve important secondary functions in nitric oxide production, regulation of vascular tone, and immune response to pathogens [1]. RBC indices, including hemoglobin (HGB), hematocrit (HCT), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), mean corpuscular volume (MCV), RBC count, and red blood cell width (RDW), are primary indicators of RBC development, size, and hemoglobin content[2]. These routinely measured clinical laboratory assays may be altered in Mendelian genetic conditions (e.g., hemoglobinopathies such as sickle cell disease or thalassemia, red cell cytoskeletal defects, or G6PD deficiency) [3] as well as by non-genetic or nutritional factors (e.g., vitamin B and iron deficiency).

RBC indices have estimated family-based heritability values ranging between 40% to 90% [4,5] and have been extensively studied as complex quantitative traits in genome-wide association studies (GWAS). Early GWAS identified common genetic variants with relatively large effects associated with RBC indices [6,7,8]. With improved imputation, increased sample sizes and deeper interrogation of coding regions of the genome, additional common variants associated with RBC indices with progressively smaller effect sizes and coding variants of larger effect with lower minor allele frequency (MAF) have been identified [9–19]. However, the full allelic spectrum (e.g., lower frequency non-coding variants, indels, structural variants) that explain the genetic architecture of complex traits remains incomplete [9]. In addition, non-European populations (including admixed U.S. minority populations such as African Americans and Hispanics/Latinos) have been under-represented in these studies. Since RBCs play a key role in pathogen invasion and defense, associated quantitative trait loci may be relatively isolated to a particular ancestral population due to local evolutionary selective pressures and population history. Emerging studies with greater inclusion of East Asian, African, and Hispanic ancestry populations have identified ancestry-specific variants associated with RBC quantitative traits [15–17,20,21]. These may account, at least in part, for inter-population differences in RBC indices as well as ethnic disparities in rates of hematologic and other related chronic diseases [18,22].

Whole genome sequencing (WGS) data have been generated through the NHLBI Trans-Omics for Precision Medicine (TOPMed) program in very large and ethnically-diverse population samples with existing hematologic laboratory measures. These TOPMed WGS data provide novel opportunities to assess rare and common single nucleotide and indel variants across the genome, including variants more common in African, East Asian or Native American ancestry individuals that are not captured by existing GWAS arrays or imputation reference panels. We thereby aimed to identify novel genetic variants and genes associated with the seven RBC indices, and to dissect association signals at previously reported regions through conditional analysis and fine-mapping.

Subjects and Methods

TOPMed study population

The analyses reported here included 62,653 participants from 13 TOPMed studies: Genetics of Cardiometabolic Health in the Amish (Amish, n=1,102), Atherosclerosis Risk in Communities Study VTE cohort (ARIC, n=8,118), Mount Sinai BioMe Biobank (BioMe, n=10,993), Coronary Artery Risk Development in Young Adults (CARDIA, n=3,042), Cardiovascular Health Study (CHS, n=3,490), Genetic Epidemiology of COPD Study (COPDGene, n=5,794), Framingham Heart Study (FHS, n=3,141), Genetic Studies of Atherosclerosis Risk (GeneSTAR, n=1,713), Hispanic Community Health Study - Study of Latinos (HCHS_SOL, n=7,655), Jackson Heart Study (JHS, n=3,033), Multi-Ethnic Study of Atherosclerosis (MESA, n=2,499), Whole Genome Sequencing to Identify Causal Genetic Variants Influencing CVD Risk - San Antonio Family Studies (SAFS, n=1,153) and Women's Health Initiative (WHI, n=10,920). We analyzed each of seven red blood cell traits separately; the total counts of participants, mean age, and the count of male participants, from each study stratified by trait are shown in **Table 1**. Further descriptions of the design of the participating TOPMed cohorts and the sampling of individuals within each cohort for TOPMed WGS are provided in the section “Participating studies” in the **Supplemental Methods**. All studies were approved by the appropriate institutional review boards (IRBs), and informed consent was obtained from all participants.

RBC trait measurements and exclusion criteria in TOPMed

The seven RBC traits considered for analyses were measured from freshly collected whole blood samples at local clinical laboratories using automated hematology analyzers calibrated to manufacturer recommendations according to clinical laboratory standards. Each trait was defined as follows. HCT is the percentage of volume of blood that is composed of red blood cells. HGB is the mass per volume (grams per deciliter) of hemoglobin in the blood. MCH is the average mass in picograms of hemoglobin per red blood cell. MCHC is the average mass concentration (grams per deciliter) of hemoglobin per red blood cell. MCV is the average volume of red blood cells, measured in femtoliters. RBC count is the count of red blood cells in the blood, by number concentration in millions per microliter. RDW is the measurement of the ratio of variation in width to the mean width of the red blood cell volume distribution curve taken at +/- one CV. In studies where multiple blood cell measurements per participant were available, we selected a single measurement for each trait and each participant as described further in **Supplemental Methods**. Each trait was analyzed to identify extreme values that may have been measurement or recording errors and such observations were removed from the analysis (see **Supplemental Methods**). **Table 1** displays the mean and standard deviation among participants analyzed after exclusions by study.

WGS data and quality control in TOPMed

WGS was performed as part of the NHLBI TOPMed program. The WGS was performed at an average depth of 38X by six sequencing centers (Broad Genomics, Northwest Genome Institute, Illumina, New York Genome Center, Baylor, and McDonnell Genome Institute) using Illumina X10 technology and DNA from blood. Here we report analyses from 'Freeze 8,' for which reads were aligned to human-genome build GRCh38 using a common pipeline across all centers. To perform variant quality control (QC), a support vector machine (SVM) classifier was trained on known variant sites (positive labels) and Mendelian inconsistent variants (negative labels). Further variant filtering was done for variants with excess heterozygosity and Mendelian discordance. Sample QC measures included: concordance between annotated and inferred genetic sex, concordance between prior array genotype data and TOPMed WGS data, and pedigree checks. Details regarding the genotype 'freezes,' laboratory methods, data processing, and quality control are described on the TOPMed website and in a common document accompanying each study's dbGaP accession (https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-methods-freeze-8) [23].

Single variant association analysis

Single variant association tests were performed for each of the seven RBC traits separately using linear mixed models (LMMs). In each case, a model assuming no association between the outcome and any genetic variant was first fit; we refer to this as the 'null model'. In the null model, covariates modelled as fixed effects were sex; age at trait measurement; a variable indicating TOPMed study and phase of genotyping (study_phase); indicators of whether the participant is known to have had a stroke, chronic obstructive pulmonary disease (COPD), or a venous thromboembolism (VTE) event; and the first 11 PC-AiR [24] principal components (PCs) of genetic ancestry. A 4th degree sparse empirical kinship matrix (KM) computed with PC-Relate [25] was included to account for genetic relatedness among participants. Additional details on the computation of the ancestry PCs and the sparse KM are provided in the **Supplemental Methods**. Finally, we allowed for heterogeneous residual variances by study and ancestry group (e.g., ARIC_White), as this has been shown previously to control inflation [26]. The details on how we estimated the ancestry group for this adjustment are in the **Supplemental Methods**. The numbers of individuals per ancestry group per study and the respective mean and standard deviation for each trait are shown in **Supplemental Table 1**.

To improve power and control of false positives when phenotypes have a non-Normal distribution, we implemented a fully-adjusted two-stage procedure for rank-Normalization when fitting the null model, for each of the 7 RBC traits in turn [27] :

1. Fit a LMM, with the fixed effect covariates, sparse KM, and heterogeneous residual variance model as described above. Perform a rank-based inverse-Normal transformation of the marginal residuals, and subsequently rescale by their variance prior to transformation. This rescaling allows for clearer interpretation of estimated genotype effect sizes from the subsequent association tests.
2. Fit a second LMM using the rank-Normalized and re-scaled residuals as the outcome, with the same fixed effect covariates, sparse KM, and heterogeneous residual variance model as in Stage 1.

The output of the Stage 2 null model was then used to perform genome-wide score tests of genetic association for all individual variants with minor allele count (MAC) ≥ 5 that passed the TOPMed variant quality filters and had less than 10% of samples freeze-wide with sequencing read depth < 10 at that particular variant. We tested up to 102,674,666 SNVs and 7,722,116 indels (**Supplemental Table 2**). Genome-wide significance was determined at the $P < 5\text{E-}9$ level [28]. For each locus, we defined the top variant as the most significant variant within a 2Mb window. All association analyses were performed using the GENESIS software [29].

Conditional analysis

Because of the very large number of variants and genomic loci that have recently been associated with quantitative RBC traits, following the single variant association analyses, we systematically performed a series of conditional association analyses for each trait to determine which genome-wide significant associations were independent of previously reported RBC variants. We gathered the variants known to be associated with each phenotype from previous publications (**Supplemental Table 3**) and matched these to TOPMed variants using position and alleles. Then, genome-wide conditional association analyses were performed by including known variants as fixed effects covariates in the null model using the same fully-adjusted two stage LMM association testing procedure described above. We performed three types of conditional analysis, namely the trait-specific, the trait-agnostic, and the iterative, step-wise conditional analysis (**Supplemental Methods**).

Single variant association analysis of chromosome 16

The alpha-globin gene region on chromosome 16p13.3 contains a large, 3.7kb structural variant common among African ancestry individuals known to be highly significantly associated with all RBC traits [15,18]. This large copy number variant is not well-tagged by SNVs in the region. Therefore, we performed genotype calling for the alpha-globin 3.7kb CNV in 52,772 available TOPMed whole genomes using MosDepth [30]. Since the chromosome 16 alpha globin CNV calls were only available for a subset of the samples in the primary analyses, to assess the effect of conditioning on the alpha globin CNV, the same set of analyses described above were run for chromosome 16 restricted to the sample set with alpha globin CNV calls. The most probable alpha globin copy number was included as a categorical variable to allow for potential non-linear effects on the phenotype.

Proportion of variance explained

For each trait, we estimated the proportion of variance explained (PVE) by the set of LD-pruned known associated variants, by the final set of conditionally-independent variants we identified following the iterative stepwise conditional analysis, and by both sets together. These cumulative PVE values were estimated jointly from the null model using approximations from multi-parameter score tests, thus accounting for covariance between the variant genotypes. The estimates were calculated using the full sample set and did not include the alpha globin CNV as a known variant but did include the set of conditionally-independent novel SNVs and indels identified on chromosome 16 after conditioning on the alpha globin CNV.

Replication studies for single variant association findings

We sought replication of the lead variants at genome-wide significant loci identified in the trait-specific conditional analysis in independent studies including the INTERVAL study (https://www.intervalstudy.org.uk/), the Kaiser-Permanente Genetic Epidemiology Research on Aging (GERA) cohort (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000674.v3.p3/), samples from the Women's Health Initiative - SNP Health Association Resource (WHI-SHARe) [31] not included in TOPMed, European ancestry samples from phase 1 of the UK BioBank (UKBB) [9] and African and East Asian ancestry samples from phase 2 of UKBB [21]. WGS data was used in INTERVAL while genotyping on various arrays and imputation to TOPMed WGS data or 1000 Genome Phase 3 reference panels were performed in Kaiser, WHI-SHARe, and UKBB. Residuals were obtained by regressing the harmonized RBC traits on age, sex, the first 10 PCs in each study stratified by ancestry, followed by association analyses testing each genetic variant with the inverse-normalized residual values. Summary statistics from each study were combined through fixed-effect inverse-weighting meta-analysis using METAL [32].

Aggregate variant association analysis of rare variants within each gene

Association tests aggregating rare variants by gene were performed for each RBC trait in order to assess the cumulative effect of rare variants within each gene and associated regulatory regions. We applied five strategies for grouping and filtering variants. Three of them aggregated coding variants and two of them aggregated coding and non-coding regulatory variants. For each aggregation strategy we filtered variants using one or more deleterious prediction scores creating relatively relaxed or stringent sets of variants (see details in **Supplemental Methods**). The five strategies are referred as C1-S,C1-R,C2-R,C2-R+NC-S and C2-R +NC-R by abbreviating , coding to"C", Non-coding to "NC", Stringent to "S" and Relaxed to "R". For all aggregate units, only variants with MAF < 0.01 that passed the quality filters and had less than 10% of samples with sequencing read depth < 10 were considered. The aggregate association tests were performed using the Efficient Variant-Set Mixed Model Association Test (SMMAT) [33]. The SMMAT test used the same fully-adjusted two-stage null model as was fit for the single variant association tests. For each aggregation unit, SMMAT efficiently combines a burden test *P* value with an asymptotically independent adjusted "SKAT-type" test *P* value using Fisher's method. This testing approach is more powerful than either a burden or SKAT [34] test alone, and is computationally more efficient than the SKAT-O test [35]. Wu weights [34] based on the variant MAF were used to upweight rarer variants in the aggregation units. Significance was determined using a Bonferroni threshold, adjusting for the number of gene-based aggregation units tested genome-wide with cumulative MAC >= 5. Two types of conditional analysis were run ("trait-specific" and "trait-agnostic), conditioning previously reported RBC trait-associated variants as well as those discovered in the TOPMed single variant tests (**Supplemental Table 3**). In addition, any previously reported RBC trait-associated variants and the set of conditionally-independent novel variants identified in our single variant analyses were excluded from the gene-based aggregation units.

Predicted loss-of-function variants and predicted gene knockouts and their association with RBC traits

Our analyses of predicted loss-of-function (pLoF) variants in TOPMed freeze 8 focused on variants annotated by ENSEMBL's Variant Effect Predictor (VEP) as nonsense, essential splice site and frameshift insertion-deletion (indel) variants. From this list, we excluded variants that map to predicted transcripts [36] and also variants located in the first and last 5% of the gene as these variants are more likely to give rise to transcripts that escape nonsense-mediated mRNA decay [37]. We used a method previously described to identify predicted gene knockouts (pKO) [38]. Briefly, we considered individuals that were homozygotes for LoF variants, but also individuals who inherited two different LoF variants in *trans* using available phased information (compound heterozygotes).

We analyzed each study-ethnic group separately, adjusting for sex, age and smoking status. We then normalized the residuals with each group using inverse normal transformation. We performed association testing per ethnic group with EPACTS. We adjusted all analyses using the first ten PCs and a kinship matrix (EMMAX) calculated using 150,000 common variants in LD. For pLoF, we tested an additive genetic model. For pKO, we coded individuals as "0" if they were not a pKO, and as "1" if they were a pKO. We meta-analyzed association results using METAL [32]. We excluded variants located in the alpha-globin region in self-reported African-ancestry individuals. The genome-wide significant threshold for each ancestral group was defined as P<0.05/number of variants.

Lentivirus packaging

HEK293T cells (ATCC, cat# CRL-3216) were cultured with DMEM with 10% fetal bovine serum and 1% Penicillin-Streptomycin solution (10,000 U/mL stock). To produce lentivirus, HEK293T cells were transfected at 70-80% confluence with 13.3 μg psPAX2, 6.7 μg VSV-G and 20 μg of the lentiviral construct plasmid of interest using 180 μg of linear polyethylenimine in 15 cm tissue culture dishes. Lentiviral supernatant was collected at both 48 h and 72 h post-transfection and concentrated by ultracentrifugation at 24,000 rpm for 4 h at 4 °C with a Beckman Coulter SW 32 Ti rotor.

HUDEP-2 cell and human CD34$^+$ hematopoietic stem and progenitor cells (HSPCs) culture

HUDEP-2 cells [39] were generously shared by Ryo Kurita (Japanese Red Cross) and Yukio Nakamura (RIKEN BioResource Research Center, University of Tsukuba, Japan) and cultured as previously described [40]. Expansion phase medium for HUDEP-2 cells consists of SFEM (Stemcell Technologies #09650) base medium supplemented with 50 ng/ml recombinant human SCF (R&D systems #255-SC), 1 μg/ml doxycycline (Sigma Aldrich #D9891), 0.4 μg/ml dexamethasone (Sigma Aldrich #D4902), 3 IU/ml EPO (Epoetin Alfa, Epogen, Amgen) and 1% Penicillin-Streptomycin solution (10,000 U/mL stock). Human CD34$^+$ HSPCs from mobilized peripheral blood of deidentified healthy donors were obtained from Fred Hutchinson Cancer Research Center, Seattle, Washington. CD34$^+$ cells were maintained in SFEM supplemented with 1x StemSpan CD34$^+$ expansion supplement (Cat# 02691, STEMCELL Technology).

Generation of AncBE4max-SpRY-expressing stable HUDEP-2 cell lines [41]

The lentiviral plasmid for AncBE4max-SpRY was generated by subcloning the coding sequence of nSpRY(D10A) into the AgeI and XcmI restriction sites of pRDA_257 (pLenti-BPNLS-AncBE4-gsXTENgs-nSpCas9-gs-UGI-gs-BPNLS-P2A-Puro), generously provided by John Doench (Broad Institute). Lentivirus was produced as described above. HUDEP-2 cells were transduced with lentivirus, and 1 μg/ml puromycin was added into culture medium 2 days after lentiviral transduction. After 2-week positive selection, AncBE4max-SpRY editing efficiency was tested using multiple sgRNAs with variable PAM sequence.

C-to-T base editing at the rs112097551 locus in HUDEP-2 cells

The sequence of single guide RNA targeting rs112097551 is summarized in **Supplemental Table 4**. Oligos (from GENEWIZ company) were annealed and ligated into LentiGuide-Puro (Addgene plasmid 52963). Following lentiviral production and transduction into cell lines with stable SpCas9 expression, 1 μg/ml puromycin were added to select for sgRNA integrants in HUDEP-2 cells expressing AncBE4max-SpRY. C-to-T editing efficiency was determined in bulk cells 10 days after lentiviral delivery into AncBE4max-SpRY-expressing HUDEP-2 cells (**Supplemental Fig. 1**). Briefly, genomic DNA was extracted using the Qiagen Blood and Tissue kit. Genomic region surrounding the sgRNA targeting site was amplified using HotStarTaq DNA polymerase (QIAGEN, Cat# 203203) for other PCR reactions strictly following the manufactory instructions with variable annealing temperature. PCR products were subject to Sanger sequencing and then EditR analysis to estimate the editing efficiency based on sequencing chromatograms [42]. Single HUDEP-2 cells were plated to obtain highly edited clones. Primers for PCR were summarized in **Supplemental Table 5**.

CRISPR/Cas genome editing in $CD34^+$ HSPCs

CD34+ cells were thawed and maintained in SFEM supplemented with 1x StemSpan $CD34^+$ expansion supplement (Cat# 02691, STEMCELL Technology) for 24 hours before electroporation. 100,000 cells per condition were electroporated using the Lonza 4D nucleofector with 100 pmol 3xNLS-SpCas9 [43] protein and 300 pmol modified sgRNA targeting the locus of interest. In addition to mock treated cells, "safe-targeting" RNPs were used as experimental controls as indicated in each figure legend. After electroporation, cells were differentiated to erythroblasts as described previously [44]. 4 days after electroporation, genomic DNA was isolated from an aliquot of cells, the sgRNA targeted locus was amplified by PCR. PCR products were subject to Sanger sequencing and then TIDE analysis to quantify indel mutations [45]. Meanwhile, total RNA was extracted from bulk cells and expression of genes of interest was determined by real time RT-qPCR as described below.

Determination of target gene expression

Total RNA was extracted from cell cultures 4 days after electroporation using the RNeasy Plus Mini Kit (QIAGEN), and reverse transcribed using the iScript cDNA synthesis kit (Biorad) according to the manufacturer's instructions. Expression of target genes was quantified using real-time RT-qPCR with

GAPDH as an internal control. All gene expression data represent the mean of at least three biological replicates. Primers for PCR were summarized in **Supplemental Table 5**.

Immunophenotyping of human CD34$^+$ HSPCs xenograft from NBSGW mice

NOD.Cg-KitW-41J Tyr + Prkdcscid Il2rgtm1Wjl (NBSGW) mice were obtained from Jackson Laboratory (Stock 026622). CD34$^+$ HSPCs were maintained and edited as described above. After electroporation, cells were allowed to recover for 24-48 hours in SFEM medium with 1x StemSpan CD34$^+$ expansion supplement (Cat# 02691, STEMCELL Technology). Cells were then washed twice by PBS, resuspended in 200 ul DPBS per million cells, and then infused by retro-orbital injection into non-irradiated NBSGW female mice. 16 weeks post transplantation, mice were euthanized, and bone marrow was collected and analyzed as previously described [45]. Analysis of bone marrow subpopulations was performed by flow cytometry. Antibodies for flow cytometry included Human TruStainFcX (422302, BioLegend), TruStainfcX (anti-mouse CD16/32, 101320, BioLegend), anti-mouse CD45 (30-F11), anti-human CD45 (HI30), and Fixable Viability Dye eFluor 780 for live/dead staining (65-0865-14, Thermo Fisher). Percentage human engraftment was calculated as hCD45$^+$ cells/(hCD45$^+$ + mCD45$^+$ cells). Cell sorting was performed on a FACSAria II machine (BD Biosciences).

Results

Single variant association analysis

In the single variant association analyses, the genomic inflation factors ranged from 1.015 to 1.038, indicating adequate control of population stratification and relatedness (**Supplemental Table 6**). A total of 69 loci reached genome-wide significance for any of the seven RBC traits (*P*<5E-9, **Supplemental Fig. 2** and **Supplemental Table 7**). Of the 69 loci, nine (*HBB*, *HBA1*, *RPN1*, *ELL2*, *EIF5-MARK3*, *MIDN*, *PIEZO1*, *TMPRSS6*, and *G6PD*) remained significant in the conditional analysis after accounting for RBC trait-specific known loci. In addition, three more loci reached genome-wide significance following RBC trait-specific conditional analysis (*19q12*, *10q26*, and *SHANK2*, *P*<5E-9, **Supplemental Fig. 3**). Therefore, a total of 12 loci showed genome-wide significance for association with at least one of the seven RBC traits in the trait-specific conditional analysis, indicating signals independent of previously reported variants (*P*<5E-9) (**Supplemental Fig. 4, Table 2**).

At the 12 significant novel loci identified in the trait-specific conditional analyses, the number of genome-wide significant variants ranged from one to 162 (**Supplemental Fig. 4** and **Supplemental Table 8**). Six loci harbored more than one genome-wide significant variants (*HBB*, *HBA1*, *ELL2, MIDN, TMPRSS6,* and *G6PD*). The lead variants for each trait at each of these 12 loci (including 14 distinct variants across the 7 traits -- 12 SNVs and two small indels) are shown in **Table 2**. Notably, only two lead variants (*MIDN*-rs73494666 and *TMPRSS6*-rs228914) had MAF>5% in TOPMed. Most of these 14 lead variants were located within non-coding regions of the genome and most were low frequency (n=3 between MAF 0.1% and MAF 2%) or rare (n=9 with MAF<0.1%). The latter category included three loci (*SHANK2*, *10q26*, and *19q12*) in which the lead variant was extremely rare with MAF<0.01%. Several of the lead variants showed large allele frequency differences between race/ethnicity groups as assessed from the genome aggregation database or gnomAD (https://gnomad.broadinstitute.org; **Supplemental Table 9**). The *RPN1*, *HBB* -rs34598529, *G6PD*, *MIDN*, and *ELL2* variants are found disproportionately among individuals of African ancestry. The *EIF5-MARK3* and chromosome 16p13.3 alpha-globin locus rs372755452 variants are found only among East Asians. The alpha-globin locus variant rs868351380 and *PIEZO1* variant are more common among Hispanics/Latinos and Europeans, respectively.

Replication of single variant discoveries

We sought replication for each of the 14 newly discovered variants in INTERVAL, the Kaiser Permanente GERA Study, the WHI-SHARe study, and UKBB phase 1 European and phase 2 African, and East Asian samples (**Supplemental Table 10**). Several of the rare variants (*SHANK2* rs535577177, *10q26* rs986415672, *19q12* rs1368500441, *EIF5/MARK3* rs370308370, and *HBB* rs11549407), were not available for testing in any of the replication studies due to low frequency, population specificity, and/or poor imputation quality. For eight of the nine lead variants with available genotype data for testing, we successfully replicated each of the trait-specific associations for *HBB*-rs34598529, *HBA1*-rs868351380, *HBA1*-rs372755452, *RPN1*, *ELL2*, *PIEZO1*, *G6PD*, and *MIDN* (meta-analysis *P*<5.6E-3, 0.05/9 loci, with consistent directions of effect). The replication *P* value for the lead variant at *TMPRSS6* did not reach the predetermined significance threshold, but the association was directionally consistent. We further note that several of our newly identified TOPMed single variant-RBC trait associations (*RPN1*, *HBB*-rs11549407 and rs34598529, and *MIDN*) reached genome-wide significance in recently published very large European ancestry or multi-ethnic imputed GWAS [19,21,46].

Relationship of single variants discovered in TOPMed to previously known RBC genetic loci

Several of the newly discovered variants (particularly those replicated in independent samples) in **Table 2** are located within genomic regions known to harbor common variants associated with RBC quantitative traits and/or variants responsible for Mendelian blood cell disorders, such as hemoglobinopathies (*HBB*, *HBA1/HBA2*) and various hemolytic or non-hemolytic anemias (*G6PD*, *PIEZO1*, *TMPRSS6*, and *GATA2-RPN1*). At the *HBB* locus, the lead variant associated with lower HCT, HGB, MCHC, and MCV is a LoF variant (rs11549407 encoding p.Gln40Ter, MAF=0.026%) while the lead variant associated with lower MCH and higher RBC, and higher RDW is a variant located within the *HBB* promoter region (rs34598529, MAF=0.083%). At the *HBA1/HBA2* locus, the lead variant for MCH and MCV, rs868351380 (MAF=0.022%), is located ~125kb upstream of *HBA1/HBA2* in an intron of the *SNRNP25* gene, and the lead variant for RBC, rs372755452 (MAF=0.010%), is located ~30kb downstream of *HBA1/HBA2* in an intron of the *LUC7L* gene. The *GATA2-RPN1* locus, which contains variants previously reported for association with MCH and RDW in a European-only analysis (rs2977562 and rs147412900) [13], was associated with MCH and MCV in TOPMed (lead variant rs112097551, *P*=4.27E-11). The MAF of the lead variant at the *GATA2-RPN1* locus in all TOPMed samples is 0.4% but is 5.9 times more common among African than non-African samples according to gnomAD. At the *G6PD* locus, the lead variant associated with lower RDW was a missense variant rs76723693, which encodes p.Leu323Pro. At the *PIEZO1* locus, the most significant variant was an in-frame 3bp deletion rs763477215 (p.Lys2169del) associated with higher MCHC. While the index SNP rs228914 at *TMPRSS6* has not been previously associated with RBC parameters, rs228914 is a cis-eQTL for *TMPRSS6* and an LD surrogate rs228916 has been previously associated with serum iron levels [47]. The remaining genetic loci (*SHANK2*, *ELL2*, *19q12*, *10q26*, *EIF5/MARK3*, and *MIDN*) have less clear functional relationships to

RBC phenotypes. Moreover, the lead variants at *EIF5/MARK3*, and *MIDN* for MCH and the lead variant at *TMPRSS6* for MCH and MCV were partially attenuated in the trait-agnostic conditional analysis.

Iterative conditional analysis identifies extensive allelic heterogeneity at HBB locus

We next performed stepwise conditional analysis to dissect association signals within each of the six loci harboring more than one genome-wide significant variants in the RBC trait-specific conditional analysis. One of the six regions (*HBB*) was found to have multiple, genome-wide significant variants independent of previously reported loci. The largest number of independent signals were observed for association with MCH (11 signals, **Supplemental Table 11**). All independent variants at the *HBB* locus had MAF <1%. No secondary independent signals were discovered in other regions (*HBA1/2, ELL2*, *MIDN*, *TMPRSS6*, and *G6PD*). For each RBC trait, we estimated the PVE by the set of LD-pruned known variants, by the novel conditionally-independent variants identified in stepwise conditional analysis, and by both sets together (**Supplemental Table 12**). In total, the PVE ranged from 3.4% (HCT) to 21.3% (MCH). The newly identified set of genetic variants explained up to 3% of phenotypic variance (for MCH and MCV).

Rare variant aggregated association analysis

We next examined rare variants with MAF <1% in TOPMed, aggregated based on protein-coding and non-coding gene units from GENCODE. To enrich for likely causal variants in the aggregation units we used five different variant grouping and filtering strategies based on coding sequence and regulatory (gene promoter/enhancer) functional annotations (see **Supplemental Methods**). After accounting for all previously reported RBC trait-specific single variants, a total of five loci were significantly associated with one or more RBC traits using various aggregation strategies (**Table 3** and **Supplemental Table 13**). These include genes encoding *HBA1/HBA2*, *TMPRSS6*, *G6PD*, and *CD36*, as well as several genes and non-coding RNAs within the beta-globin locus on chromosome 11p15 (*HBB*, *HBG1*, *CTD-264317.6*, *OR52H1*, *RF60021*, and *OR52R1*). Some of the gene units in the chromosome 11p15 beta-globin region (*HBG1*, *OR52R1*, and *RF00621*) became non-significant after further adjustment for all known RBC variants in the trait-agnostic conditional analysis (**Table 3**). After additionally accounting for all 11 independent single variant signals identified in TOPMed at the *HBB* locus in step-wise conditional analysis (**Supplemental Table 11**), as well as all trait-specific known variants, five coding genes remained significant (*HBA1/HBA2*, *HBB*, *TMPRSS6*, *G6PD*, and *CD36*, **Supplemental Table 14**) and two additional genes (*TFRC* and *SLC12A7*) reached significance threshold (**Supplemental Table 14**). *AC104389.6* a non-coding gene 2bp downstream of *HBB* was also found significant in the aggregation approach where we included upstream regulatory variants, however the variants including in this unit are predominately the same ones tested in the *HBB* gene unit and hence we have not reported this gene unit as a distinct signal.

Notably, each of the seven genes (*HBA1/HBA2, HBB, TMPRSS6, G6PD, CD36, TFRC* and *SLC12A7*) identified in rare variant aggregate analyses are known to harbor common non-coding or coding variants previously associated with RBC traits or disorders. We further explored the overall patterns of association, individual rare variants driving the associations and their annotations (**Supplemental Fig. 5** and **Supplemental Table 15**). Several observations are noteworthy. (1) In general, for each gene, there are multiple rare missense and small indel (frameshift or stop-gain) variants contributing to the aggregate association signals, rather than a single strongly associated variant. (2) The patterns of phenotypic association are generally uni-directional, and consistent with the biologic contribution of these genes to inherited RBC disorders: *HBA1* and *HBB* variants are associated with lower MCV/MCH, with *HBB* variants additionally associated with lower HCT/HGB and higher RBC/RDW, consistent with ineffective erythropoiesis and shortened red cell survival in alpha and beta thalassemia; *TMPRSS6* variants associated with lower MCH/MCV (**Supplemental Fig. 5C16-19 and 5E13-14**) and higher RDW (**Supplemental Fig. 5G14**), consistent with iron-refractory iron deficiency anemia. On the other hand, for *G6PD* rare variants, a bi-directional pattern of phenotypic association was observed for MCH, MCV, RBC, and RDW. (3) Several of the variants contributing to the *HBA1*, *HBB*, *TMPRSS6*, and *G6PD* signals are known to be pathogenic for inherited RBC disorders. Other variants that appear to contribute to the gene-based phenotypic effect are classified in ClinVar as variants of uncertain significance (VUS) or have conflicting evidence to support their pathogenicity. (4) Three of the genes (*CD36*, *TFRC* and *SLC12A7*) are located within regions of the genome containing common variants previously associated with RBC traits but have less clear relation to RBC biology. The presence of rare coding or LoF variants within these genes provides additional fine-mapping evidence that these three genes are causally responsible for RBC phenotypic variation.

pLoF and pKO variants associated with RBC traits

Predicted loss-of-function (pLoF) and predicted gene knockout (pKO) variants were examined in European, African, Hispanic, and Asian ancestry populations in TOPMed. The European ancestry population subset had the largest sample size and the largest number of both pLoF and pKO variants (**Supplemental Table 16**). Two pLoF variants reached genome-wide significance, namely *CD36*-rs3211938 for RDW in African participants and *HBB*-rs11549407 for multiple RBC traits in Hispanic and European participants (**Supplemental Table 17**), which have been reported in previously published studies. No pKO variant reached genome-wide significance in any of the ancestral groups (**Supplemental Table 18**). All pLoF and pKO variants with P<1E-4 are presented in **Supplemental Table 17 and 18**.

Gene editing in human erythroid precursors and xenotransplantation of edited primary HSPCs identifies RUVBL1 as likely target gene of RPN1-rs112097551

In silico functional annotation of the *RPN1*-rs112097551 variant revealed a CADD-PHRED score of 20.4 and that the variant lies in a putative enhancer element bound by erythroid transcription factors GATA1 and TAL1. We therefore undertook additional experiments to investigate the causal gene underlying the association signal. First, we used cytosine base editing to modify the rs112097551 reference G to alternative A allele in HUDEP-2 erythroid precursor cells. Since there was no appropriately positioned NGG PAM motif, we utilized the recently described near-PAMless SpCas9 variant cytosine base editor AncBE4max-SpRY [41], achieving 33% G-to-A conversion efficiency (**Fig. 1A**). Analysis of erythroblast promoter capture Hi-C datasets showed that the SNP interacts with the gene *RUVBL1* which is 500 kb upstream but not with intervening genes which include *RPN1* and the hematopoietic transcription factor *GATA2* (**Fig. 1B**). In five G/A heterozygous HUDEP-2 clones compared to G/G clones, we observed significantly reduced expression of *RUVBL1* without significant change in expression of 4 more proximal genes *EEFSEC*, *GATA2*, *RPN1* and *RAB7A* (**Fig. 1C**). Next, we performed SpCas9 nuclease editing to produce indels adjacent to rs112097551 in CD34+ hematopoietic stem/progenitor cell (HSPC) derived primary erythroid precursors (**Fig. 1D and 1E**). Cells bearing these short insertions and deletions centered 3 bp from the rs112097551 position demonstrated significantly reduced *RUVBL1* expression compared to control cells, while *RPN1* and *RAB7A* expression was unchanged (**Fig. 1F**). Together, these base and nuclease editing results suggest that rs112097551-G contributes to a regulatory element that exerts long-range control of *RUVBL1* expression. Prior work has shown the mouse homolog of *RUVBL1* is required for murine hematopoiesis [48]. To test the role of *RUVBL1* in human hematopoiesis, we performed gene editing studies in CD34+ HSPCs in which we targeted indels to coding sequences at *RUVBL1*. We observed 96.1% indels at *RUVBL1* compared to 84.2% indels in control cells targeted at a neutral locus. We infused edited HSPCs to immunodeficient NBSGW mice and analyzed bone marrow after 16 weeks for engrafting human hematopoietic chimerism and gene editing. Compared to CD34+ HSPCs edited at a neutral locus which showed 91.6% mean human chimerism, human CD34+ HSPCs edited at *RUVBL1* demonstrated only 7.7% mean chimerism (**Fig. 1G-I**). Engrafting human cells were marked by frequent gene edits (60.1%) when targeted at the neutral locus but only 4.8% gene edits after *RUVBL1* editing, indicating that *RUVBL1* edited cells inefficiently engrafted. Together these results suggest rs112097551-G contributes to long-range enhancement of *RUVBL1* expression, which in turn supports human hematopoiesis.

Discussion

We report here the first WGS-based association analysis of RBC traits in an ethnically diverse sample of 62,653 participants from TOPMed. We identified 14 association signals across 12 genomic regions conditionally independent of previously reported RBC trait loci and replicated eight of these (*RPN1*, *ELL2*, *PIEZO1*, *G6PD*, *MIDN, HBB*-rs34598529, *HBA1*-rs868351380 and -rs372755452) in independent samples with available imputed genome-wide genotype data. The replicated association signals are described further below. Stepwise, iterative conditional analysis of the beta-globin gene regions on chromosomes 11 additionally identified 12 independent association signals at the *HBB* locus. Further investigation of aggregated rare variants identified seven genes (*HBA1/HBA2*, *HBB*, *TMPRSS6*, *G6PD*, *CD36*, *TFRC* and *SLC12A7*) containing significant rare variant association signals independent of previously reported and newly discovered RBC trait-associated single variants. For the *RPN1* locus, we used base and nuclease editing to demonstrate that the sentinel variant rs112097551 acts through a cis-regulatory element that exerts long-range control of the gene *RUVBL1* which is essential for hematopoiesis.

Our study highlights the benefits of increasing participant ethnic diversity and coverage of the genome in genetic association studies of complex polygenic traits. Among the 24 unique novel and independent variants we identified in the single variant association analyses, 21 showed MAF <1% in all TOPMed samples and 18 were monomorphic in at least one of the four major contributing ancestral populations in our analysis (European, African, East Asian, and Hispanic). These low-frequency or ancestry-specific variants were most likely missed by previous GWAS analysis using imputed genotype data or focusing on one ancestral population (**Supplemental Table 11**).

GATA2-RPN1

Here we report and replicate a distinct low-frequency variant [MAF=0.4% overall but considerably higher frequency among African (0.94%) than European (0.07%) ancestry individuals) associated with higher MCH and MCV in TOPMed (rs112097551). The region between *GATA2* and *RPN1* on chromosome 3q21 contains several common variants previously associated with various WBC-related traits in European, Asian, and Hispanic ancestry individuals and two variants previously associated with MCH and RDW in Europeans (rs2977562 and rs147412900) [13]. GATA2 is a hematopoietic transcription factor and heterozygous coding or enhancer mutations of *GATA2* are responsible for autosomal dominant hereditary mononuclear cytopenia, immunodeficiency and myelodysplastic syndromes, as well as lymphatic dysfunction [49,50] (MIM 137295). There was no evidence of association of the TOPMed MCH/MCV-associated rs112097551 variant with WBC-related traits in TOPMed (data not shown), though the variant was associated with higher monocyte count and percentage in Astle et al [9], but was not conditionally independent of other variants in the region. The MCV/MCH-associated rs112097551 variant lies in a putative enhancer element bound by erythroid transcription factors GATA-1 and TAL-1 and demonstrates physical interaction in erythroblasts with gene *RUVBL1* 500kb away. Our results from gene editing of *RUVBL1* in primary human HPSC and xenotransplantation suggest that *RUVBL1* plays a role in human hematopoiesis, consistent with data from mouse models suggesting that *RUVBL1* (which encodes the protein product pontin) to be essential for murine hematopoietic stem cell survival [48]. This finding also highlights the complexity and importance of experimentally validating the causal gene(s) underlying GWAS signals for complex traits, which are often assigned according to physical proximity (*RPN1*) or assumed on the basis of biologic function (*GATA2*).

ELL2

The chromosome *5q15* non-coding variant rs116635225 associated with lower MCH also has a low frequency in TOPMed (1.3%) and is considerably more common among African ancestry individuals (3.9%). The rs116635225 variant is located ~27kb upstream of *ELL2*, a gene responsible for immunoglobulin mRNA production and transcriptional regulation in plasma cells. Coding and regulatory variants of *ELL2* have been associated with risk of multiple myeloma in European and African ancestry individuals as well as reduced levels of immunoglobulin A and G in healthy subjects [51–53]. Another set of genetic variants located ~200kb away in the promoter region of *GLRX* or glutaredoxin-1 (rs10067881, rs17462893, rs57675369) have been associated with higher reticulocyte count in UKBB Europeans [9]. Glutaredoxin-1 is a cytoplasmic enzyme that catalyzes the reversible reduction of glutathione-protein mixed disulfides and contributes to the antioxidant defense system. Congenital deficiencies of other members of the glutaredoxin enzyme family (*GLRX5*) have been reported in patients with sideroblastic anemia [54–56]. Notably, our *ELL2* rs116635225 MCH-associated variant remained genome-wide significant after conditioning on the myeloma or reticulocyte-related variants. Therefore, the precise genetic regulatory mechanisms of the red cell trait associations in this region remain to be determined.

MIDN

The chromosome *19p13* African variant rs73494666 associated with lower MCV/MCH is located in an open chromatin region of an intron of *MIDN*, which encodes the midbrain nucleolar protein midnolin. The gene-rich region on chromosome *19p13* also includes *SBNO2*, *STK11*, *CBARP*, *ATP5F1D*, *CIRBP*, *EFNA2*, and *GPX4*. However, none of these genes have clear relationships to hematopoiesis or red structure/function. Other variants in the region have been associated with MCH and RBC count (rs757293) [13] or reticulocytes (rs35971149) [9]. The *MIDN*- rs73494666 variant overlaps ENCODE cis-regulatory elements for CD34 stem cells and other blood cell progenitors.

PIEZO1

Mutations in the mechanosensitive ion channel *PIEZO1* on chromosome *16q24* have been reported in patients with autosomal dominant hereditary xerocytosis (OMIM 194380), a congenital hemolytic anemia associated with increased calcium influx, red cell dehydration, and potassium efflux along with various red cell laboratory abnormalities including increased MCHC, MHC, and reticulocytosis [57,58]. Most reported hereditary xerocytosis *PIEZO1* missense mutations are associated with at least partial gain-of-function and are located within the highly conserved C-terminal region near the pore of the ion channel. In some individuals carrying *PIEZO1* missense mutations, mild red cell laboratory parameter alterations without frank hemolytic anemia have been reported [59]. The *PIEZO1* 3bp short tandem repeat (STR) rs763477215 in-frame coding variant (p.Lys2169del) associated with higher MCHC in TOPMed is extremely rare in all populations except for the Ashkenazi Jewish population (frequency of 1.5% in gnomAD), has not been previously associated with hereditary xerocytosis, and therefore has been reported as 'benign' in ClinVar. The p.Lys2169del variant is located in a highly basic -K-K-K-K- motif near the C-terminus of the 36 transmembrane domain protein within a 14-residue linker region between the central ion channel pore and the peripheral propeller-like mechanosensitive domains important for modulating PIEZO1 channel function [60,61]. Interestingly, another 3bp in-frame deletion of *PIEZO1* (p.Glu657del) reported to be highly enriched in prevalence among African populations was recently associated with dehydrated red blood cells and reduced susceptibility to malaria [62,63]. In TOPMed, however, we were unable to confirm any association between the rs59446030 (p.Glu657del) putative malaria susceptibility allele variant and phenotypic variation in MCHC (*P*-value for trait-specific conditional analysis=0.21) .

TMPRSS6

TMPRSS6 on chromosome *22q12* encodes matriptase-2, a transmembrane serine protease that down-regulates the production of hepcidin in the liver and therefore plays an essential role in iron homeostasis [64]. Rare mutations of *TMPRSS6* are associated with iron-refractory iron deficiency anemia [65] characterized by microcytic hypochromic anemia and low transferrin saturation. Several common *TMPRSS6* variants have been associated with multiple RBC traits through prior GWAS. The common *TMPRSS6* intronic variant associated with *TMPRSS6* expression and lower MCH/MCV in TOPMed (rs228914/rs228916) was previously reported to be associated with lower iron levels [47], and therefore likely contributes to lower MCH and MCV via iron deficiency. In rare variant aggregated association testing, we were able to identify several additional rare coding missense, stop-gain, or splice variants that appear to drive the gene-based association of *TMPRSS6* with lower MCH/MCV and higher RDW. At least one of these variants at exon 13 rs387907018 (p.Glu5323Lys, loss-of-function mutation) has been reported in a compound heterozygous iron-refractory iron deficiency anemia (IRIDA) patient [66], suggesting that inheritance of this or similar LoF variants in the heterozygote state may contribute to mild reductions in MCV/MCH or increased RDW [65].

G6PD

X-linked *G6PD* mutations (glucose-6-phosphate dehydrogenase) are the most common cause worldwide of acute and chronic hemolytic anemia. The *G6PD*-rs76723693 (c.968T>C) low-frequency missense variant (p.Leu323Pro, referred to as *G6PD* Nefza[67]) associated is common in persons of African ancestry and is associated with lower RDW in TOPMed. In persons of African ancestry, the p.Leu323Pro variant is often co-inherited with another *G6PD* missense variant, p.Asn126Asp, encoded by rs1050829 (c.A376>G). The 968C/376G haplotype in African ancestry individuals constitutes one of several forms of the *G6PD* variant A- [68–71]. Functional studies of the p.Leu323Pro, p.Asn126Asp, and the double mutant suggest the p.Leu323Pro variant is the primary contributor to reduced catalytic activity[72]. In the US, another African ancestry *G6PD* A- variant is due to the haplotypic combination of rs1050829 (p.Asn126Asp0) and rs1050828 (p.Val68Met), which has an allele frequency of ~12%. Our finding that rs76723693 is significantly associated with lower RDW after conditioning on rs1050828 is consistent with the independence of effects of the *G6PD* Nefza and A- variants on red cell physiology and morphology. Importantly, both rs76723693 and rs1050828 *G6PD* variants were recently reported to have the effect of lowering hemoglobin A1c (HbA1c) values and therefore should be considered when screening African Americans for type-2 diabetes [73]

In gene-based analyses, several additional *G6PD* missense variants contributed to the aggregated rare variant association signals for MCH, MCV, RBC, and RDW, including the class II Southeast Asian Mahidol variant p.Gly163Ser [74] and the class II Union variant (p.Arg454Cys) [75]. For a third previously reported variant associated with G6PD deficiency, the East Asian class II Gahoe variant (p.His32Arg) [76], there is conflicting evidence of pathogenicity in ClinVar. Of the two female rs137852340 (p.His32Arg) variant allele carriers in TOPMed, one has a normal RDW and one has an elevated RDW. These findings add to the further genotypic-phenotypic complexity and clinical spectrum of G6PD deficiency, which is influenced its sex-linkage and zygosity, residual *G6PD* variant enzyme activity and stability, genetic background, and environmental exposures [77].

HBB

Heterozygosity for the common African *HBB*-rs334 hemoglobin S (p.Glu7Val) or rs33930165 hemoglobin C (p.Glu7Gln) beta-globin structural variants have recently been associated with alterations in various red cell laboratory parameters including lower hemoglobin, MCV, MCH, and RDW, along with higher MCHC, RDW, and HbA1c [17,18,20,78–80]. In TOPMed, we were able to identify at least 10 additional low-frequency or rare variants within the *HBB* locus independently associated with HGB, RBC, MCV, MCH, MCHC, and/or RDW. Notably, six of the 10 variants correspond to *HBB* 5' UTR and promoter regions (rs34598529 or -29 A>G [81]; rs33944208 or -88C>T [82–84]; splice site (rs33915217 or IVS1-5G>C) [85] [82]; rs33945777 or IVS2-1G>A [82,85]; rs35004220 or IVS-I-110 G->A [86,87], and nonsense mutations (rs11549407 or p.Gln40Ter) [88,89] previously identified in patients with beta-thalassemia. These findings confirm the very mild phenotype and clinically "silent" nature of the heterozygote carrier state of these beta-globin gene variants [90]. Several of these mutations occur more commonly in populations of South Asian (rs33915217), African (rs34598529, rs33944208), or Mediterranean (rs11549407) ancestry. Four additional association signals in the region rs73404549 (*HBG2*), rs77333754, rs1189661759, and rs539384429 are all rare non-coding variants without obvious functional consequences. In addition to the *HBB* protein-coding variants identified in single variant analyses, several of the rare variants driving the aggregate *HBB* gene-based association with lower HGB/HCT and MCH/MCV/MCHC and higher RBC/RDW are similarly previously reported missense, frameshift or nonsense mutations previously identified in beta-thalassemia patients and categorized as pathogenic in ClinVar (**Supplemental Fig. 5** and **Supplemental Table 15**).

HBA1/HBA2

Several common DNA polymorphisms located in the α-globin gene cluster on chromosome 16p13.3 have been associated with red cell traits in large GWAS [7,8,91], including heterozygosity for the common African ancestral 3.7kb deletion which contributes to quantitative RBC phenotypes among African Americans and Hispanics/Latinos. In TOPMed, we identified two low-frequency variants in single variant testing associated with MCH, MCV, and/or RBC count, independently of the 3.7kb deletion. The rs868351380 variant is found primarily among Hispanics/Latinos while the rs372755452 variant is found primarily among East Asians. Neither of these two non-coding variants is located in any known alpha-globin regulatory region, and therefore requires further mechanistic confirmation. By contrast, in gene-based rare variant analysis, we identified several known alpha-globin variants associated in aggregate with lower MCH and MCV including the South Asian variant Hb Q India (*HBA1*, p.Asp64His) [92–94] and the African variant Hb Groene Hart (*HBA1*, p.Pro120Ser) [95–97]. In homozygous or compound heterozygous forms, these latter variants have been reported in probands with alpha-thalassemia, whereas heterozygotes generally have mild microcytic phenotype. Several additional variants contributing to the *HBA1* gene-based rare variant MCH/MCV signal (e.g., a 1 bp indel causing frameshift p.Asn79Ter) may represent previously undetected alpha-thalassemia mutations.

CD36, *TFRC* and *SLC12A7*

The presence of rare coding or LoF variants within *CD36*, *TFRC* and *SLC12A7* provides evidence that these genes are causally responsible for RBC phenotypic variation. A common African ancestral null variant of CD36 (rs3211938 or p.Tyr325Ter) has been previously associated with higher RDW and with lower CD36 expression in erythroblasts [98]. In TOPMed, additional *CD36* rare coding variants were associated in aggregate with higher RDW independent of rs3211938, including several nonsense and frameshift or splice acceptor mutations, which have been previously classified as VUS. Further characterization of the genetic complexity of the CD36□null phenotype (common in African and Asian populations) may provide information relevant to the tissue-specific expression of this receptor on red cells, platelets, monocytes, and endothelial cells and its role in malaria infection and disease severity[99]. *TFRC* encodes the transferrin receptors (TfR1), which is required for iron uptake and erythropoiesis [100]. While common non-coding variants of *TFRC* have been associated with MCV and RDW, the only known *TFRC*-related Mendelian disorder is a homozygous p.Tyr20His substitution reported to cause combined immunodeficiency affecting leukocytes and platelets but not red cells [101]. Common variants of *SLC12A7* encoding the potassium ion channel KCC4 have been associated with RDW and other RBC phenotypes. While KCC4 is expressed in erythroblasts [102], its role in red blood cell function is not well-described [103]. Further characterization of KCC4 LoF variants may illuminate the role of this ion transporter in red cell dehydration with potential implications for treatment of patients with sickle cell disease [104].

In summary, we illustrate that expanding coverage of the genome using WGS as applied to large, population-based multi-ethnic samples can lead to discovery of novel variants associated with quantitative RBC traits. Most of the newly discovered variants were of low frequency and/or disproportionately observed in non-Europeans. We also report extensive allelic heterogeneity at the chromosome 11 beta-globin locus, including associations with several known beta-thalassemia carrier variants. The gene-based association of rare variants within *HBA1/HBA2*, *HBB*, *TMPRSS6*, *G6PD*, *CD36*, *TFRC* and *SLC12A7* independent of known single variants in the same genes further suggest that rare functional variants in genes responsible for Mendelian RBC disorders contribute to the genetic architecture of RBC phenotypic variation among the population at large. Together these results demonstrate the utility of WGS in ethnically-diverse population-based samples for expanding our understanding of the genetic architecture of quantitative hematologic traits and suggest a continuum between complex traits and Mendelian red cell disorders.

Acknowledgements

Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the

National Heart, Lung and Blood Institute (NHLBI). The table below presents study specific omics support information. Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

Paul S. de Vries was supported by American Heart Association grant number 18CDA34110116. H.C. and E.J. were supported by the National Eye Institute (NEI) grant R01 EY027004, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) R01 DK116738 and by the National Cancer Institute (NCI) R01CA2416323. M.P.C and D.J were supported by NHLBI grant U01HL137162. D.E.B. was supported by NHLBI P01HL032262, DP2HL137300, R01HL130733. B.P.K. was supported by NCI R00 CA218870 and NHLBI P01HL142494.

TOPMed Accession #	TOPMed Project	Parent Study Name	TOPMed Phase	Omics Center	Omics Support
phs000956	Amish	Amish	1	Broad Genomics	3R01HL121007-01S1
phs001211	AFGen	ARIC AFGen	1	Broad Genomics	3R01HL092577-06S1
phs001211	VTE	ARIC	2	Baylor	3U54HG003273-12S2 / HHSN268201500015C
phs001644	AFGen	BioMe AFGen	2.4	MGI	3UM1HG008853-01S2
phs001644	BioMe	BioMe	3	Baylor	HHSN268201600033I
phs001644	BioMe	BioMe	3	MGI	HHSN268201600037I
phs001612	CARDIA	CARDIA	3	Baylor	HHSN268201600033I
phs001368	CHS	CHS	3	Baylor	HHSN268201600033I
phs001368	VTE	CHS VTE	2	Baylor	3U54HG003273-12S2 / HHSN268201500015C
phs000951	COPD	COPDGene	1	NWGC	3R01HL089856-08S1
phs000951	COPD	COPDGene	2	Broad Genomics	HHSN268201500014C
phs000951	COPD	COPDGene	2.5	Broad Genomics	HHSN268201500014C
phs000974	AFGen	FHS AFGen	1	Broad Genomics	3R01HL092577-06S1
phs000974	FHS	FHS	1	Broad Genomics	3U54HG003067-12S2
phs001218	AA_CAC	GeneSTAR	2	Broad	HHSN268201500014C

		AA_CAC		Genomics	
phs001218	GeneSTAR	GeneSTAR	legacy	Illumina	R01HL112064
phs001218	GeneSTAR	GeneSTAR	2	Psomagen	3R01HL112064-04S1
phs001395	HCHS_SOL	HCHS_SOL	3	Baylor	HHSN268201600033I
phs000964	JHS	JHS	1	NWGC	HHSN268201100037C
phs001416	AA_CAC	MESA AA_CAC	2	Broad Genomics	HHSN268201500014C
phs001416	MESA	MESA	2	Broad Genomics	3U54HG003067-13S1
phs001215	SAFS	SAFS	1	Illumina	3R01HL113323-03S1
phs001215	SAFS	SAFS	legacy	Illumina	R01HL113322
phs001237	WHI	WHI	2	Broad Genomics	HHSN268201500014C

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

ARIC: The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services (contract numbers HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I and HHSN268201700005I). The authors thank the staff and participants of the ARIC study for their important contributions.

BioMe: The Mount Sinai BioMe Biobank has been supported by The Andrea and Charles Bronfman Philanthropies and in part by Federal funds from the NHLBI and NHGRI (U01HG00638001; U01HG007417; X01HL134588). We thank all participants in the Mount Sinai Biobank. We also thank all our recruiters who have assisted and continue to assist in data collection and management and are grateful for the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

CARDIA: The Coronary Artery Risk Development in Young Adults Study (CARDIA) is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with the University of Alabama at Birmingham (HHSN268201800005I & HHSN268201800007I), Northwestern University (HHSN268201800003I), University of Minnesota (HHSN268201800006I), and Kaiser Foundation Research Institute (HHSN268201800004I). CARDIA was also partially supported by the Intramural Research Program of the National Institute on Aging (NIA) and an intra□agency agreement between NIA and NHLBI (AG0005).

CHS: Cardiovascular Health Study: This research was supported by contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080,

N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the National Heart, Lung, and Blood Institute (NHLBI), with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

COPDGene: The COPDGene project described was supported by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health. The COPDGene project is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Novartis, Pfizer, Siemens and Sunovion. A full listing of COPDGene investigators can be found at: http://www.copdgene.org/directory

FHS: The Framingham Heart Study (FHS) acknowledges the support of contracts NO1-HC-25195 and HHSN268201500001I from the National Heart, Lung and Blood Institute and grant supplement R01 HL092577-06S1 for this research. We also acknowledge the dedication of the FHS study participants without whom this research would not be possible.

GeneSTAR: GeneSTAR was supported by the National Institutes of Health/National Heart, Lung, and Blood Institute (U01 HL72518, HL087698, HL112064, HL11006, HL118356) and by a grant from the National Institutes of Health/National Center for Research Resources (M01-RR000052) to the Johns Hopkins General Clinical Research Center. We would like to thank our participants and staff for their valuable contributions.

HCHS/SOL: The Hispanic Community Health Study/Study of Latinos is a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (HHSN268201300001I / N01-HC-65233), University of Miami (HHSN268201300004I / N01-HC-65234), Albert Einstein College of Medicine (HHSN268201300002I / N01-HC-65235), University of Illinois at Chicago – HHSN268201300003I / N01-HC-65236 Northwestern Univ), and San Diego State University (HHSN268201300005I / N01-HC-65237). The following Institutes/Centers/Offices have contributed to the HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, NIH Institution-Office of Dietary Supplements.

JHS: The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State

University (HHSN268201300049C and HHSN268201300050C), Tougaloo College (HHSN268201300048C), and the University of Mississippi Medical Center (HHSN268201300046C and HHSN268201300047C) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute for Minority Health and Health Disparities (NIMHD). The authors also wish to thank the staffs and participants of the JHS.

MESA: MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420. MESA Family is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support is provided by grants and contracts R01HL071051, R01HL071205, R01HL071250, R01HL071251, R01HL071258, R01HL071259, by the National Center for Research Resources, Grant UL1RR033176. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

SAFS: Collection of the San Antonio Family Study data was supported in part by National Institutes of Health (NIH) grants R01 HL045522, MH078143, MH078111 and MH083824; and whole genome sequencing of SAFS subjects was supported by U01 DK085524 and R01 HL113323. We are very grateful to the participants of the San Antonio Family Study for their continued involvement in our research programs.

WHI: The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C.

Declaration of Interests

B.P.K is an inventor on patent applications filed by Mass General Brigham that describe genome engineering technologies, is an advisor to Acrigen Biosciences, and consults for Avectas Inc. and ElevateBio.

Table 1. Characteristics of the TOPMed samples by study [1]

Study	N (male)	Age	HCT	HGB	MCH	MCHC	MCV	RBC	RDW
Amish	1,102 (557)	50.6±16.9	40.6±3.5	13.8±1.2	30.9±1.3	34.1±0.8	90.7±3.4	4.5±0.4	-
ARIC	8,113 (3,577)	54.8±5.8	41.6±4.0	13.9±1.4	30.5±2.1	33.3±1.0	89.6±5.1	4.5±0.5	14.1±1.1
BioMe	10,990 (4,559)	52.1±13.5	39.5±5.2	13.1±1.7	30.3±2.8	33.7±1.0	89.0±7.2	4.4±0.6	14.2±1.8
CARDIA	3,042 (1,319)	25.0±3.6	42.1±4.4	14.2±1.5	29.8±2.1	33.8±1.0	88.1±5.4	4.8±0.5	-
CHS	3,490 (1,459)	72.6±5.4	41.8±3.9	14.0±1.3	-	33.5±1.0	-	-	-
COPDGene	5,794 (2,913)	64.8±8.8	42.0±4.1	13.9±1.5	30.3±2.3	33.2±1.1	91.4±5.8	4.6±0.5	-
FHS	3,140 (1,514)	58.4±15.0	41.6±4.0	14.1±1.3	31.1±1.8	33.9±1.0	91.9±4.9	4.5±0.5	13.1±1.0
GeneSTAR	1,713 (699)	43.7±12.9	40.9±3.9	13.5±1.4	29.6±2.1	33.0±0.8	89.5±5.3	4.6±0.4	-
HCHS/SOL	7,655 (3,186)	46.6±14.0	42.1±4.1	13.8±1.5	29.1±2.2	32.7±1.4	89.2±6.0	4.7±0.4	13.8±1.3
JHS	2,905 (1,089)	53.5±12.8	39.4±4.3	13.1±1.5	28.9±2.5	33.2±0.9	86.9±6.3	4.5±0.5	13.7±1.4
MESA	2,499 (1,211)	69.4±9.2	40.1±4.0	13.4±1.4	30.1±2.3	33.4±1.1	89.9±6.0	4.5±0.5	-
SAFS	1,152 (492)	40.6±15.9	40.3±4.5	13.1±1.5	29.0±2.3	32.6±1.4	88.9±5.4	4.5±0.5	-
WHI	10,913 (0)	66.7±6.8	40.2±2.9	13.5±1.0	29.9±2.1	32.9±1.1	90.9±5.8	4.4±0.4	14.2±1.3

HCT, hematocrit; HGB, hemoglobin; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; RBC, red blood cell count; RDW, red blood cell width.

[1] Values are shown as mean±SD.

Table 2. Genome-wide significant loci identified in the trait-specific conditional analysis in TOPMed [1]

Trait	Variant	Chr:Pos	Gene	CA/NCA	CAF(%)	N	Beta	SE	P	$P_{conditional1}$ [2]	$P_{conditional2}$ [3]
HCT	rs11549407	11: 5226774	*HBB*	A/G	0.026	62,487	-4.94	0.67	1.68E-13	3.43E-13	1.55E-12
HGB	rs11549407	11: 5226774	*HBB*	A/G	0.026	62,461	-2.14	0.23	2.86E-21	4.76E-21	1.75E-20
	rs1368500441	19: 28868893	*19q12*	A/G	0.005	62,461	2.65	0.46	1.02E-8	2.49E-9	6.64E-8
MCH	rs112097551	3:128603774	*RPN1*	A/G	0.398	62,461	0.78	0.12	4.01E-10	4.27E-11	4.08E-10
	rs116635225	5: 95989447	*ELL2*	A/G	1.307	46,241	-0.43	0.07	3.37E-9	1.18E-11	2.58E-11
	rs986415672	10: 131440166	*10q26*	T/C	0.006	46,241	-4.26	0.82	2.16E-7	3.06E-9	2.49E-9
	rs34598529	11: 5227100	*HBB*	C/T	0.083	46,241	-4.31	0.29	1.06E-49	1.37E-52	1.03E-53
	rs535577177	11: 70462791	*SHANK2*	A/G	0.008	46,241	-4.72	0.82	1.04E-8	8.28E-10	3.38E-9
	rs370308370	14: 103044696	*EIF5/MARK3*	A/G	0.011	46,241	-4.35	0.74	3.15E-9	1.42E-9	5.49E-9
	rs868351380	16: 55649	*HBA1/2*	C/G	0.022	37,917	-3.19	0.51	4.85E-10	8.87E-11	1.49E-11
	rs73494666	19: 1253643	*MIDN*	T/C	16.5	46,241	-0.16	0.03	1.11E-9	4.27E-11	9.00E-9
	rs228914	22: 37108472	*TMPRSS6*	A/C	89.0	46,241	-0.09	0.02	3.76E-5	6.53E-10	2.76E-8
MCHC	rs11549407	11: 5226774	*HBB*	A/G	0.028	52,648	-1.79	0.18	4.79E-23	1.21E-23	1.87E-23
	rs763477215	16: 88717174	*PIEZO1*	A/ATCT	0.070	52,648	0.66	0.11	1.57E-9	2.66E-9	1.74E-9
MCV	rs112097551	3:128603774	*RPN1*	A/G	0.405	48,830	1.98	0.31	1.09E-10	7.65E-12	6.28E-10
	rs11549407	11: 5226774	*HBB*	A/G	0.028	48,830	-16.54	1.08	3.52E-53	1.00E-54	1.31E-55
	rs868351380	16: 55649	*HBA1/2*	C/G	0.022	39,107	-7.99	1.31	1.19E-9	2.17E-10	3.20E-11
	rs73494666	19: 1253643	*MIDN*	T/C	16.7	48,830	-0.42	0.07	3.90E-10	2.72E-10	1.77E-11
	rs228914	22: 37108472	*TMPRSS6*	A/C	89.1	48,830	-0.20	0.06	3.80E-4	9.53E-10	2.52E-6
RBC	rs34598529	11: 5227100	*HBB*	C/T	0.084	44,470	0.55	0.06	3.59E-22	1.48E-25	1.91E-23
	rs372755452	16: 199621	*HBA1/2*	A/AG	0.010	36,430	1.27	0.18	1.55E-12	6.08E-10	3.95E-9
RDW	rs34598529	11: 5227100	*HBB*	C/T	0.092	29,385	1.96	0.22	4.44E-19	1.35E-20	2.16E-20
	rs76723693	X: 154533025	*G6PD*	G/A	0.297	29,385	-0.91	0.10	2.38E-19	2.97E-20	2.99E-15

Chr, chromosome; Pos, position; CA, coded allele; NCA, non-coded allele; CAF, coded allele frequency; HCT, hematocrit; HGB, hemoglobin; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; RBC, red blood cell count; RDW, red blood cell width.

[1] Conditional analysis at the *HBA1/2* locus was performed in a subset of TOPMed samples with available alpha globin CNV data.

[2] In the first conditional analysis, trait-specific reported variants were adjusted in the model.

[3] In the second conditional analysis, all reported variants regardless of associated traits were adjusted in the model.

Table 3. Genome-wide significant genes in the aggregated association analysis in TOPMed [1]

Trait	Chr	Start	End	Gene	No. of variants	MAC	P	$P_{conditional1}{}^{1}$	$P_{conditional2}{}^{2}$
HCT	11	5225464	5229395	*HBB*	15	76	1.27E-23	1.35E-23	5.91E-18
	11	5224309	5225461	*AC104389.6*	94	1395	1.85E-13	6.23E-15	3.32E-11
HGB	11	5225464	5229395	*HBB*	15	76	2.06E-35	8.99E-30	7.44E-29
	11	5224309	5225461	*AC104389.6*	94	1394	1.29E-18	2.43E-17	1.05E-23
MCH	11	5224309	5225461	*AC104389.6*	83	1078	6.76E-100	2.87E-104	5.51E-95
	11	5225464	5229395	*HBB*	34	126	9.53E-76	2.76E-78	3.11E-75
	11	5224448	5224639	*RF00621*	588	12096	1.93E-20	4.02E-20	1.28E-12
	11	5544489	5548533	*OR52H1*	8	441	6.15E-16	6.13E-17	9.82E-18
	11	5248079	5249859	*HBG1*	526	7852	9.95E-09	8.61E-9	8.36E-4
	16	176680	177522	*HBA1*	16	30	4.97E-6	5.95E-9	1.98E-9
	22	37065436	37109713	*TMPRSS6*	243	3317	6.77E-07	9.92E-12	1.16E-9
	X	154531391	154547572	*G6PD*	59	599	2.32E-06	6.59E-7	2.50E-7
MCHC	11	5224309	5225461	*AC104389.6*	88	1225	2.37E-64	5.01E-40	8.73E-39
	11	5225464	5229395	*HBB*	36	136	4.07E-34	1.04E-33	2.65E-31
	11	5544489	5548533	*OR52H1*	8	502	3.88E-07	2.12E-6	7.50E-7
MCV	11	5224309	5225461	*AC104389.6*	86	1148	2.29E-153	1.40E-148	4.75E-108
	11	5225464	5229395	*HBB*	35	130	4.10E-82	6.02E-86	1.11E-81
	11	5224448	5224639	*RF00621*	597	12848	3.11E-37	1.56E-30	2.74E-16
	11	5544489	5548533	*OR52H1*	8	468	1.07E-19	3.29E-19	4.50E-22
	11	5248079	5249859	*HBG1*	546	8321	4.46E-15	5.71E-8	1.79E-2
	16	176680	177522	*HBA1*	16	30	5.11E-4	2.03E-6	9.24E-7
	22	37065436	37109713	*TMPRSS6*	252	3567	8.61E-06	9.11E-10	9.90E-8
	X	154531390	154547572	*G6PD*	82	732	2.19E-12	2.70E-13	7.06E-14
RBC	11	5224309	5225461	*AC104389.6*	81	1036	9.51E-57	5.47E-60	2.55E-44
	11	5225464	5229395	*HBB*	34	113	2.24E-24	5.35E-28	6.06E-25
	11	5224448	5224639	*RF00621*	576	11551	6.13E-15	7.39E-15	7.31E-7
	11	4803433	4804380	*OR52R1*	72	1551	4.48E-09	1.87E-9	9.37E-2
	11	5248079	5249859	*HBG1*	517	7502	2.74E-07	4.09E-8	3.49E-1
	X	154531390	154547572	*G6PD*	58	574	1.29E-06	2.99E-9	3.49E-8
RDW	7	80369575	80679277	*CD36*	178	1537	3.28E-4	6.45E-7	2.46E-6

11	5224309	5225461	*AC104389.6*	73	702	1.55E-29	1.19E-30	2.84E-24
11	5225464	5229395	*HBB*	13	54	2.06E-24	9.07E-27	1.14E-24
11	5544489	5548533	*OR52H1*	7	300	1.20E-08	4.55E-9	7.08E-9
11	5224448	5224639	*RF00621*	480	8119	1.80E-08	1.21E-8	2.01E-4
22	37065436	37109713	*TMPRSS6*	72	614	2.89E-07	1.38E-7	4.86E-8
X	154531390	154547572	*G6PD*	47	449	2.13E-24	6.71E-27	8.33E-21

Chr, chromosome; MAC, minor allele counts; HCT, hematocrit; HGB, hemoglobin; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; RBC, red blood cell count; RDW, red blood cell width.

[1] Conditional analysis at the *HBA1/2* locus was performed in a subset of TOPMed samples with available alpha globin CNV data.

[2] In the first conditional analysis, trait-specific reported variants were adjusted in the model. All genes that reached genome-wide significance in the trait-specific conditional analysis were presented.

[3] In the second conditional analysis, all reported variants regardless of associated traits were adjusted in the model.

Figure Legends

Figure 1. Gene editing implicates *RUVBL1* in rs112097551 association.

(A) The MCV/MCH associated variant rs112097551 was targeted by cytosine base editing in HUDEP-2 cells expressing AncBE4max-SpRY and sgRNA to convert G-to-A. Sequencing chromatogram and heatmap of bulk edited HUDEP-2 cells generated by EditR analysis.

(B) Promoter capture Hi-C from ChiCP analysis [105] of erythroblasts [106].

(C) Gene expression measured by RT-qPCR in rs112097551-G/G (n=5) and -G/A (n=5) HUDEP-2 base edited clones. Expression normalized to mean of G/G clones for each gene.

(D) Representative allele table demonstrating type and frequency of indels following nuclease editing in CD34+ HSPCs following 3xNLS-SpCas9:sgRNA electroporation. Indels analyzed by TIDE analysis [45].

(E) Indel frequency measured by Sanger sequencing with TIDE analysis in CD34+ HSPCs 4 days following 3xNLS-SpCas9:sgRNA electroporation with indicated sgRNA (n = 3 biological replicates).

(F) Gene expression measured by RT-qPCR in CD34+ HSPCs 4 days following 3xNLS-SpCas9:sgRNA targeting adjacent to rs112097551 compared to neutral locus. Expression of *EEFSEC* and *GATA2* was undetectable in HSPCs.

(G) Indel frequency following 3xNLS-SpCas9:sgRNA targeting *RUVBL1* coding sequence or neutral control locus in input cell 4 days after RNP electroporation or engrafted bone marrow samples 16 weeks after infusion to NBSGW mice.

(H) Representative flow cytometry of human and mouse CD45+ cells from NBSGW bone marrow 16 weeks after cell infusion (representative of 3 mice).

(I) Mean human hematopoietic chimerism determined by hCD45+/total CD45+ cells from NBSGW bone marrow 16 weeks after cell infusion (n = 3 mice per group).

Student's t test (two-tailed test). *** P < 0.001; ** P < 0.01; * P < 0.05; ns, not significant.

References

1. Kuhn, V., Diederich, L., Keller, T.C.S., Kramer, C.M., Lückstädt, W., Panknin, C., Suvorava, T., Isakson, B.E., Kelm, M., and Cortese-Krott, M.M. (2017). Red blood cell function and dysfunction: redox regulation, nitric oxide metabolism, anemia. Antioxid. Redox Signal. *26*, 718–742.

2. Sarma, P.R. (1990). Red Cell Indices. In Clinical Methods: The History, Physical, and Laboratory Examinations, H.K. Walker, W.D. Hall, J.W. Hurst, H.K. Walker, W.D. Hall, J.W. Hurst, H.K. Walker, W.D. Hall, J.W. Hurst, H.K. Walker, et al., eds. (Boston: Butterworths), p.

3. Lippi, G., and Mattiuzzi, C. (2020). Updated worldwide epidemiology of inherited erythrocyte disorders. Acta Haematol. *143*, 196–203.

4. 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.

5. Patel, K.V. (2008). Variability and heritability of hemoglobin concentration: an opportunity to improve understanding of anemia in older adults. Haematologica *93*, 1281–1283.

6. Soranzo, N., Spector, T.D., Mangino, M., Kühnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M., et al. (2009). A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat. Genet. *41*, 1182–1190.

7. Ganesh, S.K., Zakai, N.A., van Rooij, F.J.A., Soranzo, N., Smith, A.V., Nalls, M.A., Chen, M.-H., Kottgen, A., Glazer, N.L., Dehghan, A., et al. (2009). Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nat. Genet. *41*, 1191–1198.

8. van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Verweij, N., Sehmi, J., Paul, D.S., Elling, U., Allayee, H., Li, X., et al. (2012). Seventy-five genetic loci influencing the human red blood cell. Nature *492*, 369–375.

9. 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.

10. Iotchkova, V., Huang, J., Morris, J.A., Jain, D., Barbieri, C., Walter, K., Min, J.L., Chen, L., Astle, W., Cocca, M., et al. (2016). Discovery and refinement of genetic loci associated with cardiometabolic risk using dense imputation maps. Nat. Genet. *48*, 1303–1312.

11. CHARGE Consortium Hematology Working Group (2016). Meta-analysis of rare and common exome chip variants identifies S1PR4 and other loci influencing blood cell traits. Nat. Genet. *48*, 867–876.

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

13. 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.

14. van Rooij, F.J.A., Qayyum, R., Smith, A.V., Zhou, Y., Trompet, S., Tanaka, T., Keller, M.F., Chang, L.-C., Schmidt, H., Yang, M.-L., et al. (2017). Genome-wide Trans-ethnic Meta-analysis Identifies Seven

Genetic Loci Influencing Erythrocyte Traits and a Role for RBPMS in Erythropoiesis. Am. J. Hum. Genet. *100*, 51–63.

15. Jo Hodonsky, C., Schurmann, C., Schick, U.M., Kocarnik, J., Tao, R., van Rooij, F.J., Wassel, C., Buyske, S., Fornage, M., Hindorff, L.A., et al. (2018). Generalization and fine mapping of red blood cell trait genetic associations to multi-ethnic populations: The PAGE Study. Am. J. Hematol.

16. 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.

17. Gurdasani, D., Carstensen, T., Fatumo, S., Chen, G., Franklin, C.S., Prado-Martinez, J., Bouman, H., Abascal, F., Haber, M., Tachmazidou, I., et al. (2019). Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. Cell *179*, 984-1002.e36.

18. 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. (2018). Common α-globin variants modify hematologic and other clinical phenotypes in sickle cell trait and disease. PLoS Genet. *14*, e1007293.

19. 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*, 1214-1231.e11.

20. Hodonsky, C.J., Jain, D., Schick, U.M., Morrison, J.V., Brown, L., McHugh, C.P., Schurmann, C., Chen, D.D., Liu, Y.M., Auer, P.L., et al. (2017). Genome-wide association study of red blood cell traits in Hispanics/Latinos: The Hispanic Community Health Study/Study of Latinos. PLoS Genet. *13*, e1006760.

21. Chen, M.-H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., et al. (2020). Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. Cell *182*, 1198-1213.e14.

22. 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.

23. Regier, A.A., Farjoun, Y., Larson, D.E., Krasheninina, O., Kang, H.M., Howrigan, D.P., Chen, B.-J., Kher, M., Banks, E., Ames, D.C., et al. (2018). Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat. Commun. *9*, 4038.

24. Conomos, M.P., Miller, M.B., and Thornton, T.A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet. Epidemiol. *39*, 276–293.

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

26. Conomos, M.P., Laurie, C.A., Stilp, A.M., Gogarten, S.M., McHugh, C.P., Nelson, S.C., Sofer, T., Fernández-Rhodes, L., Justice, A.E., Graff, M., et al. (2016). Genetic diversity and association studies in US hispanic/latino populations: applications in the hispanic community health study/study of latinos. Am. J. Hum. Genet. *98*, 165–184.

27. Sofer, T., Zheng, X., Gogarten, S.M., Laurie, C.A., Grinde, K., Shaffer, J.R., Shungin, D., O'Connell,

J.R., Durazo-Arvizo, R.A., Raffield, L., et al. (2019). A fully adjusted two-stage procedure for rank-normalization in genetic association studies. Genet. Epidemiol. *43*, 263–275.

28. Lin, D.-Y. (2019). A simple and accurate method to determine genomewide significance for association tests in sequencing studies. Genet. Epidemiol. *43*, 365–372.

29. Gogarten, S.M., Sofer, T., Chen, H., Yu, C., Brody, J.A., Thornton, T.A., Rice, K.M., and Conomos, M.P. (2019). Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics *35*, 5346–5348.

30. Pedersen, B.S., and Quinlan, A.R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics *34*, 867–868.

31. Reiner, A.P., Lettre, G., Nalls, M.A., Ganesh, S.K., Mathias, R., Austin, M.A., Dean, E., Arepalli, S., Britton, A., Chen, Z., et al. (2011). Genome-wide association study of white blood cell count in 16,388 African Americans: the continental origins and genetic epidemiology network (COGENT). PLoS Genet. *7*, e1002108.

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

33. Chen, H., Huffman, J.E., Brody, J.A., Wang, C., Lee, S., Li, Z., Gogarten, S.M., Sofer, T., Bielak, L.F., Bis, J.C., et al. (2019). Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. Am. J. Hum. Genet. *104*, 260–274.

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

35. Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., NHLBI GO Exome Sequencing Project—ESP Lung Project Team, Christiani, D.C., Wurfel, M.M., and Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. *91*, 224–237.

36. Cummings, B.B., Karczewski, K.J., Kosmicki, J.A., Seaby, E.G., Watts, N.A., Singer-Berk, M., Mudge, J.M., Karjalainen, J., Satterstrom, F.K., O'Donnell-Luria, A.H., et al. (2020). Transcript expression-aware annotation improves rare variant interpretation. Nature *581*, 452–458.

37. Lindeboom, R.G.H., Supek, F., and Lehner, B. (2016). The rules and impact of nonsense-mediated mRNA decay in human cancers. Nat. Genet. *48*, 1112–1118.

38. Lessard, S., Manning, A.K., Low-Kam, C., Auer, P.L., Giri, A., Graff, M., Schurmann, C., Yaghootkar, H., Luan, J., Esko, T., et al. (2016). Testing the role of predicted gene knockouts in human anthropometric trait variation. Hum. Mol. Genet. *25*, 2082–2092.

39. Kurita, R., Suda, N., Sudo, K., Miharada, K., Hiroyama, T., Miyoshi, H., Tani, K., and Nakamura, Y. (2013). Establishment of immortalized human erythroid progenitor cell lines able to produce enucleated red blood cells. PLoS ONE *8*, e59890.

40. Vinjamur, D.S., and Bauer, D.E. (2018). Growing and Genetically Manipulating Human Umbilical Cord Blood-Derived Erythroid Progenitor (HUDEP) Cell Lines. Methods Mol. Biol. *1698*, 275–284.

41. Walton, R.T., Christie, K.A., Whittaker, M.N., and Kleinstiver, B.P. (2020). Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. Science *368*, 290–296.

42. Kluesner, M.G., Nedveck, D.A., Lahr, W.S., Garbe, J.R., Abrahante, J.E., Webber, B.R., and Moriarity, B.S. (2018). EditR: A Method to Quantify Base Editing from Sanger Sequencing. The CRISPR Journal *1*, 239–250.

43. Wu, Y., Zeng, J., Roscoe, B.P., Liu, P., Yao, Q., Lazzarotto, C.R., Clement, K., Cole, M.A., Luk, K., Baricordi, C., et al. (2019). Highly efficient therapeutic gene editing of human hematopoietic stem cells. Nat. Med. *25*, 776–783.

44. Giarratana, M.-C., Rouard, H., Dumont, A., Kiger, L., Safeukui, I., Le Pennec, P.-Y., François, S., Trugnan, G., Peyrard, T., Marie, T., et al. (2011). Proof of principle for transfusion of in vitro-generated red blood cells. Blood *118*, 5071–5079.

45. Brinkman, E.K., Chen, T., Amendola, M., and van Steensel, B. (2014). Easy quantitative assessment of genome editing by sequence trace decomposition. Nucleic Acids Res. *42*, e168.

46. Kowalski, M.H., Qian, H., Hou, Z., Rosen, J.D., Tapia, A.L., Shan, Y., Jain, D., Argos, M., Arnett, D.K., Avery, C., et al. (2019). Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. PLoS Genet. *15*, e1008500.

47. Benyamin, B., Esko, T., Ried, J.S., Radhakrishnan, A., Vermeulen, S.H., Traglia, M., Gögele, M., Anderson, D., Broer, L., Podmore, C., et al. (2014). Novel loci affecting iron homeostasis and their effects in individuals at risk for hemochromatosis. Nat. Commun. *5*, 4926.

48. Bereshchenko, O., Mancini, E., Luciani, L., Gambardella, A., Riccardi, C., and Nerlov, C. (2012). Pontin is essential for murine hematopoietic stem cell survival. Haematologica *97*, 1291–1294.

49. Crispino, J.D., and Horwitz, M.S. (2017). GATA factor mutations in hematologic disease. Blood *129*, 2103–2110.

50. Spinner, M.A., Sanchez, L.A., Hsu, A.P., Shaw, P.A., Zerbe, C.S., Calvo, K.R., Arthur, D.C., Gu, W., Gould, C.M., Brewer, C.C., et al. (2014). GATA2 deficiency: a protean disorder of hematopoiesis, lymphatics, and immunity. Blood *123*, 809–821.

51. 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.

52. Ali, M., Ajore, R., Wihlborg, A.-K., Niroula, A., Swaminathan, B., Johnsson, E., Stephens, O.W., Morgan, G., Meissner, T., Turesson, I., et al. (2018). The multiple myeloma risk allele at 5q15 lowers ELL2 expression and increases ribosomal gene expression. Nat. Commun. *9*, 1649.

53. Du, Z., Weinhold, N., Song, G.C., Rand, K.A., Van Den Berg, D.J., Hwang, A.E., Sheng, X., Hom, V., Ailawadhi, S., Nooka, A.K., et al. (2020). A meta-analysis of genome-wide association studies of multiple myeloma among men and women of African ancestry. Blood Adv. *4*, 181–190.

54. Ye, H., Jeong, S.Y., Ghosh, M.C., Kovtunovych, G., Silvestri, L., Ortillo, D., Uchida, N., Tisdale, J., Camaschella, C., and Rouault, T.A. (2010). Glutaredoxin 5 deficiency causes sideroblastic anemia by specifically impairing heme biosynthesis and depleting cytosolic iron in human erythroblasts. J. Clin. Invest. *120*, 1749–1761.

55. Peskin, A.V., Pace, P.E., Behring, J.B., Paton, L.N., Soethoudt, M., Bachschmid, M.M., and

Winterbourn, C.C. (2016). Glutathionylation of the active site cysteines of peroxiredoxin 2 and recycling by glutaredoxin. J. Biol. Chem. *291*, 3053–3062.

56. Furuyama, K., and Kaneko, K. (2018). Iron metabolism in erythroid cells and patients with congenital sideroblastic anemia. Int. J. Hematol. *107*, 44–54.

57. Zarychanski, R., Schulz, V.P., Houston, B.L., Maksimova, Y., Houston, D.S., Smith, B., Rinehart, J., and Gallagher, P.G. (2012). Mutations in the mechanotransduction protein PIEZO1 are associated with hereditary xerocytosis. Blood *120*, 1908–1915.

58. Andolfo, I., Alper, S.L., De Franceschi, L., Auriemma, C., Russo, R., De Falco, L., Vallefuoco, F., Esposito, M.R., Vandorpe, D.H., Shmukler, B.E., et al. (2013). Multiple clinical forms of dehydrated hereditary stomatocytosis arise from mutations in PIEZO1. Blood *121*, 3925–3935, S1.

59. Knight, T., Zaidi, A.U., Wu, S., Gadgeel, M., Buck, S., and Ravindranath, Y. (2019). Mild erythrocytosis as a presenting manifestation of PIEZO1 associated erythrocyte volume disorders. Pediatr. Hematol. Oncol. *36*, 317–326.

60. Zhang, T., Chi, S., Jiang, F., Zhao, Q., and Xiao, B. (2017). A protein interaction mechanism for suppressing the mechanosensitive Piezo channels. Nat. Commun. *8*, 1797.

61. Zhao, Q., Zhou, H., Chi, S., Wang, Y., Wang, J., Geng, J., Wu, K., Liu, W., Zhang, T., Dong, M.-Q., et al. (2018). Structure and mechanogating mechanism of the Piezo1 channel. Nature *554*, 487–492.

62. Ma, S., Cahalan, S., LaMonte, G., Grubaugh, N.D., Zeng, W., Murthy, S.E., Paytas, E., Gamini, R., Lukacs, V., Whitwam, T., et al. (2018). Common PIEZO1 allele in african populations causes RBC dehydration and attenuates plasmodium infection. Cell *173*, 443-455.e12.

63. Nguetse, C.N., Purington, N., Ebel, E.R., Shakya, B., Tetard, M., Kremsner, P.G., Velavan, T.P., and Egan, E.S. (2020). A common polymorphism in the mechanosensitive ion channel PIEZO1 is associated with protection from severe malaria in humans. Proc Natl Acad Sci USA *117*, 9074–9081.

64. Wang, C.-Y., Meynard, D., and Lin, H.Y. (2014). The role of TMPRSS6/matriptase-2 in iron regulation and anemia. Front. Pharmacol. *5*, 114.

65. De Falco, L., Sanchez, M., Silvestri, L., Kannengiesser, C., Muckenthaler, M.U., Iolascon, A., Gouya, L., Camaschella, C., and Beaumont, C. (2013). Iron refractory iron deficiency anemia. Haematologica *98*, 845–853.

66. Silvestri, L., Guillem, F., Pagani, A., Nai, A., Oudin, C., Silva, M., Toutain, F., Kannengiesser, C., Beaumont, C., Camaschella, C., et al. (2009). Molecular mechanisms of the defective hepcidin inhibition in TMPRSS6 mutations associated with iron-refractory iron deficiency anemia. Blood *113*, 5605–5608.

67. Benmansour, I., Moradkhani, K., Moumni, I., Wajcman, H., Hafsia, R., Ghanem, A., Abbès, S., and Préhu, C. (2013). Two new class III G6PD variants [G6PD Tunis (c.920A>C: p.307Gln>Pro) and G6PD Nefza (c.968T>C: p.323 Leu>Pro)] and overview of the spectrum of mutations in Tunisia. Blood Cells Mol. Dis. *50*, 110–114.

68. Beutler, B., and Cerami, A. (1989). The biology of cachectin/TNF--a primary mediator of the host response. Annu. Rev. Immunol. *7*, 625–655.

69. Hamel, A.R., Cabral, I.R., Sales, T.S.I., Costa, F.F., and Olalla Saad, S.T. (2002). Molecular heterogeneity of G6PD deficiency in an Amazonian population and description of four new variants.

Blood Cells Mol. Dis. *28*, 399–406.

70. Monteiro, W.M., Franca, G.P., Melo, G.C., Queiroz, A.L.M., Brito, M., Peixoto, H.M., Oliveira, M.R.F., Romero, G.A.S., Bassat, Q., and Lacerda, M.V.G. (2014). Clinical complications of G6PD deficiency in Latin American and Caribbean populations: systematic review and implications for malaria elimination programmes. Malar. J. *13*, 70.

71. Reading, N.S., Ruiz-Bonilla, J.A., Christensen, R.D., Cáceres-Perkins, W., and Prchal, J.T. (2017). A patient with both methemoglobinemia and G6PD deficiency: A therapeutic conundrum. Am. J. Hematol. *92*, 474–477.

72. Ramírez-Nava, E.J., Ortega-Cuellar, D., Serrano-Posada, H., González-Valdez, A., Vanoye-Carlo, A., Hernández-Ochoa, B., Sierra-Palacios, E., Hernández-Pineda, J., Rodríguez-Bustamante, E., Arreguin-Espinosa, R., et al. (2017). Biochemical Analysis of Two Single Mutants that Give Rise to a Polymorphic G6PD A-Double Mutant. Int. J. Mol. Sci. *18*,.

73. Sarnowski, C., Leong, A., Raffield, L.M., Wu, P., de Vries, P.S., DiCorpo, D., Guo, X., Xu, H., Liu, Y., Zheng, X., et al. (2019). Impact of Rare and Common Genetic Variants on Diabetes Diagnosis by Hemoglobin A1c in Multi-Ancestry Cohorts: The Trans-Omics for Precision Medicine Program. Am. J. Hum. Genet. *105*, 706–718.

74. Huang, Y., Choi, M.Y., Au, S.W.N., Au, D.M.Y., Lam, V.M.S., and Engel, P.C. (2008). Purification and detailed study of two clinically different human glucose 6-phosphate dehydrogenase variants, G6PD(Plymouth) and G6PD(Mahidol): Evidence for defective protein folding as the basis of disease. Mol. Genet. Metab. *93*, 44–53.

75. Wang, X.-T., Lam, V.M.S., and Engel, P.C. (2005). Marked decrease in specific activity contributes to disease phenotype in two human glucose 6-phosphate dehydrogenase mutants, G6PD(Union) and G6PD(Andalus). Hum. Mutat. *26*, 284.

76. Chiu, D.T., Zuo, L., Chao, L., Chen, E., Louie, E., Lubin, B., Liu, T.Z., and Du, C.S. (1993). Molecular characterization of glucose-6-phosphate dehydrogenase (G6PD) deficiency in patients of Chinese descent and identification of new base substitutions in the human G6PD gene. Blood *81*, 2150–2154.

77. Luzzatto, L., Ally, M., and Notaro, R. (2020). Glucose-6-phosphate dehydrogenase deficiency. Blood *136*, 1225–1240.

78. 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.

79. Fatumo, S., Carstensen, T., Nashiru, O., Gurdasani, D., Sandhu, M., and Kaleebu, P. (2019). Complimentary Methods for Multivariate Genome-Wide Association Study Identify New Susceptibility Genes for Blood Cell Traits. Front. Genet. *10*, 334.

80. Velasco-Rodríguez, D., Alonso-Domínguez, J.-M., González-Fernández, F.-A., Muriel, A., Abalo, L., Sopeña, M., Villarrubia, J., Ropero, P., Plaza, M.P., Tenorio, M., et al. (2016). Laboratory parameters provided by Advia 2120 analyser identify structural haemoglobinopathy carriers and discriminate between Hb S trait and Hb C trait. J. Clin. Pathol. *69*, 912–920.

81. Antonarakis, S.E., Boehm, C.D., Serjeant, G.R., Theisen, C.E., Dover, G.J., and Kazazian, H.H.

(1984). Origin of the beta S-globin gene in blacks: the contribution of recurrent mutation or gene conversion or both. Proc Natl Acad Sci USA *81*, 853–856.

82. Wong, C., Antonarakis, S.E., Goff, S.C., Orkin, S.H., Boehm, C.D., and Kazazian, H.H. (1986). On the origin and spread of beta-thalassemia: recurrent observation of four mutations in different ethnic groups. Proc Natl Acad Sci USA *83*, 6529–6532.

83. Orkin, S.H., Antonarakis, S.E., and Kazazian, H.H. (1984). Base substitution at position -88 in a beta-thalassemic globin gene. Further evidence for the role of distal promoter element ACACCC. J. Biol. Chem. *259*, 8679–8681.

84. Gonzalez-Redondo, J.M., Stoming, T.A., Lanclos, K.D., Gu, Y.C., Kutlar, A., Kutlar, F., Nakatsuji, T., Deng, B., Han, I.S., and McKie, V.C. (1988). Clinical and genetic heterogeneity in black patients with homozygous beta-thalassemia from the southeastern United States. Blood *72*, 1007–1014.

85. Treisman, R., Orkin, S.H., and Maniatis, T. (1983). Specific transcription and RNA splicing defects in five cloned beta-thalassaemia genes. Nature *302*, 591–596.

86. Westaway, D., and Williamson, R. (1981). An intron nucleotide sequence variant in a cloned beta +-thalassaemia globin gene. Nucleic Acids Res. *9*, 1777–1788.

87. Spritz, R.A., Jagadeeswaran, P., Choudary, P.V., Biro, P.A., Elder, J.T., deRiel, J.K., Manley, J.L., Gefter, M.L., Forget, B.G., and Weissman, S.M. (1981). Base substitution in an intervening sequence of a beta+-thalassemic human globin gene. Proc Natl Acad Sci USA *78*, 2455–2459.

88. Trecartin, R.F., Liebhaber, S.A., Chang, J.C., Lee, K.Y., Kan, Y.W., Furbetta, M., Angius, A., and Cao, A. (1981). beta zero thalassemia in Sardinia is caused by a nonsense mutation. J. Clin. Invest. *68*, 1012–1017.

89. Orkin, S.H., and Goff, S.C. (1981). Nonsense and frameshift mutations in beta 0-thalassemia detected in cloned beta-globin genes. J. Biol. Chem. *256*, 9782–9784.

90. Atweh, G.F., Wong, C., Reed, R., Antonarakis, S.E., Zhu, D., Ghosh, P.K., Maniatis, T., Forget, B.G., and Kazazian, H.H. (1987). A new mutation in IVS-1 of the human beta globin gene causing beta thalassemia due to abnormal splicing. Blood *70*, 147–151.

91. Chen, Z., Tang, H., Qayyum, R., Schick, U.M., Nalls, M.A., Handsaker, R., Li, J., Lu, Y., Yanek, L.R., Keating, B., et al. (2013). Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. Hum. Mol. Genet. *22*, 2529–2538.

92. Harrison, A., Mashon, R.S., Kakkar, N., and Das, S. (2018). Clinico-Hematological Profile of Hb Q India: An Uncommon Hemoglobin Variant. Indian J. Hematol. Blood Transfus. *34*, 299–303.

93. Schmidt, R.M., Bechtel, K.C., and Moo-Penn, W.F. (1976). Hemoglobin QIndia, alpha 64 (E13) Asp replaced by His, and beta-thalassemia in a Canadian family. Am. J. Clin. Pathol. *66*, 446–448.

94. Sukumaran, P.K., Merchant, S.M., Desai, M.P., Wiltshire, B.G., and Lehmann, H. (1972). Haemoglobin Q India (alpha 64(E13) aspartic acid histidine) associated with beta-thalassemia observed in three Sindhi families. J. Med. Genet. *9*, 436–442.

95. Yu, X., Mollan, T.L., Butler, A., Gow, A.J., Olson, J.S., and Weiss, M.J. (2009). Analysis of human alpha globin gene mutations that impair binding to the alpha hemoglobin stabilizing protein. Blood *113*, 5961–5969.

96. Giordano, P.C., Zweegman, S., Akkermans, N., Arkesteijn, S.G.J., van Delft, P., Versteegh, F.G.A., Wajcman, H., and Harteveld, C.L. (2007). The first case of Hb Groene Hart [alpha119(H2)Pro-->Ser, CCT-->TCT (alpha1)] homozygosity confirms that a thalassemia phenotype is associated with this abnormal hemoglobin variant. Hemoglobin *31*, 179–182.

97. Joly, P., Lacan, P., Garcia, C., and Francina, A. (2014). Description of the phenotypes of 63 heterozygous, homozygous and compound heterozygous patients carrying the Hb Groene Hart [α119(H2)Pro→Ser; HBA1: c.358C>T] variant. Hemoglobin *38*, 64–66.

98. 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.

99. Cserti-Gazdewich, C.M., Mayr, W.R., and Dzik, W.H. (2011). Plasmodium falciparum malaria and the immunogenetics of ABO, HLA, and CD36 (platelet glycoprotein IV). Vox Sang. *100*, 99–111.

100. Fillebeen, C., Charlebois, E., Wagner, J., Katsarou, A., Mui, J., Vali, H., Garcia-Santos, D., Ponka, P., Presley, J., and Pantopoulos, K. (2019). Transferrin receptor 1 controls systemic iron homeostasis by fine-tuning hepcidin expression to hepatocellular iron load. Blood *133*, 344–355.

101. Aljohani, A.H., Al-Mousa, H., Arnaout, R., Al-Dhekri, H., Mohammed, R., Alsum, Z., Nicolas-Jilwan, M., Alrogi, F., Al-Muhsen, S., Alazami, A.M., et al. (2020). Clinical and immunological characterization of combined immunodeficiency due to TFRC mutation in eight patients. J. Clin. Immunol.

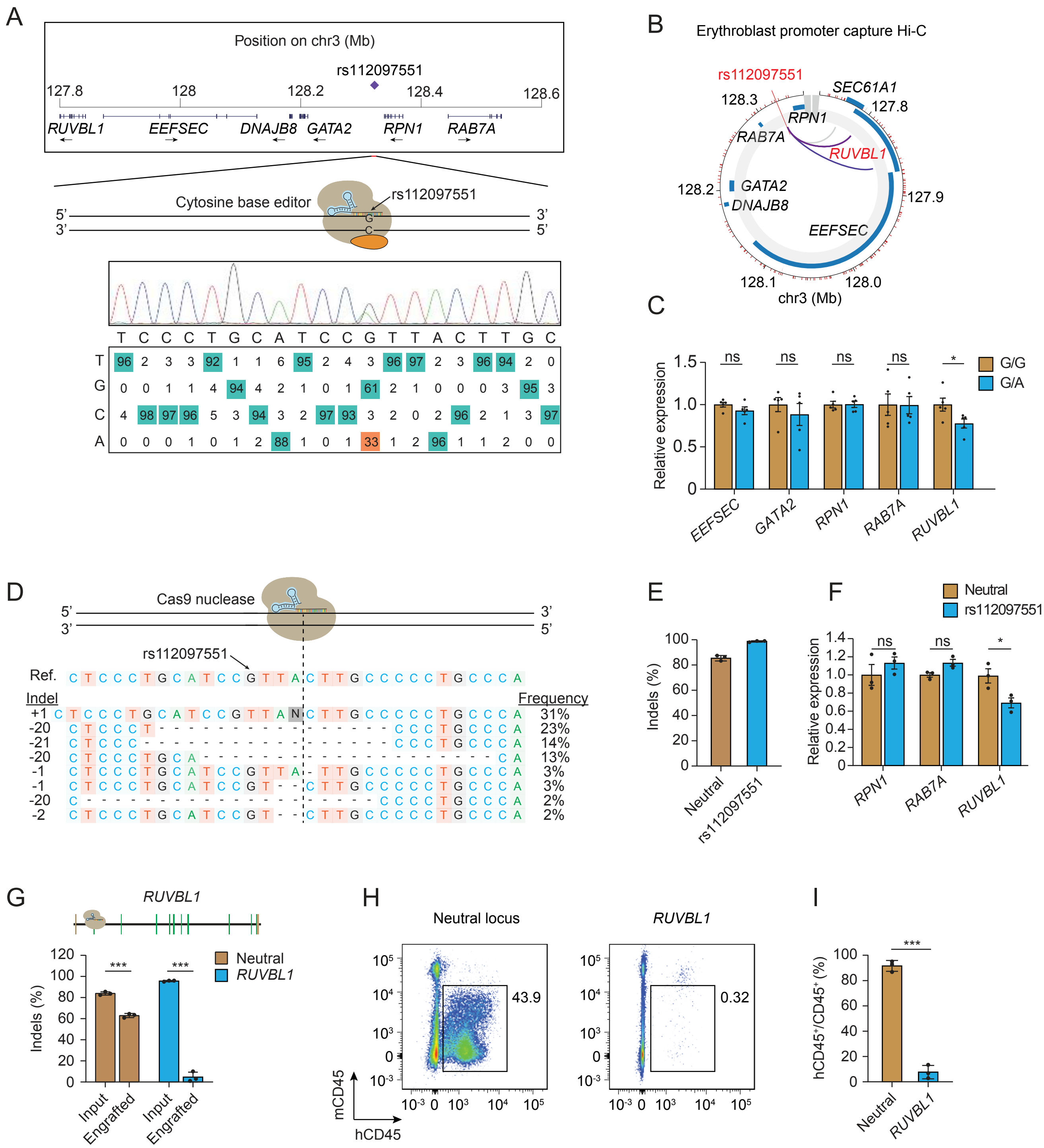
102. Pan, D., Kalfa, T.A., Wang, D., Risinger, M., Crable, S., Ottlinger, A., Chandra, S., Mount, D.B., Hübner, C.A., Franco, R.S., et al. (2011). K-Cl cotransporter gene expression during human and murine erythroid differentiation. J. Biol. Chem. *286*, 30492–30503.

103. Marcoux, A.A., Garneau, A.P., Frenette-Cotton, R., Slimani, S., Mac-Way, F., and Isenring, P. (2017). Molecular features and physiological roles of K+-Cl- cotransporter 4 (KCC4). Biochim. Biophys. Acta Gen. Subj. *1861*, 3154–3166.

104. Brugnara, C. (2003). Sickle cell disease: from membrane pathophysiology to novel therapies for prevention of erythrocyte dehydration. J. Pediatr. Hematol. Oncol. *25*, 927–933.

105. Schofield, E.C., Carver, T., Achuthan, P., Freire-Pritchett, P., Spivakov, M., Todd, J.A., and Burren, O.S. (2016). CHiCP: a web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets. Bioinformatics *32*, 2511–2513.

106. Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., et al. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell *167*, 1369-1384.e19.


A
Position on chr3 (Mb)
rs112097551
127.8
128
128.2
128.4
128.6
RUVBL1
EEFSEC
DNAJB8
GATA2
RPN1
RAB7A
Cytosine base editor
rs112097551
B
Erythroblast promoter capture Hi-C
rs112097551
SEC61A1
128.3
127.8
RPN1
RAB7A
RUVBL1
128.2
GATA2
DNAJB8
127.9
EEFSEC
128.1
128.0
chr3 (Mb)
C
Relative expression
G/G
G/A
ns
*
EEFSEC
GATA2
RPN1
RAB7A
RUVBL1
D
Cas9 nuclease
rs112097551
Ref.
Indel
Frequency
+1
-20
-21
-20
-1
-1
-20
-2
31%
23%
14%
13%
3%
3%
2%
2%
E
Indels (%)
Neutral
rs112097551
F
Neutral
rs112097551
Relative expression
RPN1
RAB7A
RUVBL1
G
RUVBL1
Indels (%)
Neutral
RUVBL1
***
Input
Engrafted
H
Neutral locus
RUVBL1
43.9
0.32
mCD45
hCD45
I
hCD45⁺/CD45⁺ (%)
***
Neutral
RUVBL1