Please cite this article in press as: Chami et al., Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits, The American Journal of Human Genetics (2016), http://dx.doi.org/10.1016/j.ajhg.2016.05.007

ARTICLE

Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits

Nathalie Chami,[1,2,91] Ming-Huei Chen,[3,91] Andrew J. Slater,[4,5,91] John D. Eicher,[3] Evangelos Evangelou,[6,7] Salman M. Tajuddin,[8] Latisha Love-Gregory,[9] Tim Kacprowski,[10,11] Ursula M. Schick,[12] Akihiro Nomura,[13,14,15,16,17] Ayush Giri,[18] Samuel Lessard,[1,2] Jennifer A. Brody,[19] Claudia Schurmann,[12,20] Nathan Pankratz,[21] Lisa R. Yanek,[22] Ani Manichaikul,[23] Raha Pazoki,[24] Evelin Mihailov,[25] W. David Hill,[26,27] Laura M. Raffield,[28] Amber Burt,[29] Traci M. Bartz,[30] Diane M. Becker,[22] Lewis C. Becker,[31] Eric Boerwinkle,[32,33] Jette Bork-Jensen,[34] Erwin P. Bottinger,[12] Michelle L. O'Donoghue,[35] David R. Crosslin,[36] Simon de Denus,[2,37] Marie-Pierre Dubé,[1,2] Paul Elliott,[6] Gunnar Engström,[38,39] Michele K. Evans,[8] James S. Floyd,[19] Myriam Fornage,[40] He Gao,[6] Andreas Greinacher,[41] Vilmundur Gudnason,[42,43] Torben Hansen,[34] Tamara B. Harris,[44] Caroline Hayward,[45] Jussi Hernesniemi,[46,47,48] Heather M. Highland,[32,49]

(Author list continued on next page)

Red blood cell (RBC) traits are important heritable clinical biomarkers and modifiers of disease severity. To identify coding genetic variants associated with these traits, we conducted meta-analyses of seven RBC phenotypes in 130,273 multi-ethnic individuals from studies genotyped on an exome array. After conditional analyses and replication in 27,480 independent individuals, we identified 16 new RBC variants. We found low-frequency missense variants in *MAP1A* (rs55707100, minor allele frequency [MAF] = 3.3%, $p = 2 \times 10^{-10}$ for hemoglobin [HGB]) and *HNF4A* (rs1800961, MAF = 2.4%, $p < 3 \times 10^{-8}$ for hematocrit [HCT] and HGB). In African Americans, we identified a nonsense variant in *CD36* associated with higher RBC distribution width (rs3211938, MAF = 8.7%, $p = 7 \times 10^{-11}$) and showed that it is associated with lower *CD36* expression and strong allelic imbalance in ex vivo differentiated human erythroblasts. We also identified a rare missense variant in *ALAS2* (rs201062903, MAF = 0.2%) associated with lower mean corpuscular volume and mean corpuscular hemoglobin ($p < 8 \times 10^{-9}$). Mendelian mutations in *ALAS2* are a cause of sideroblastic anemia and erythropoietic protoporphyria. Gene-based testing highlighted three rare missense variants in *PKLR*, a gene mutated in Mendelian non-spherocytic hemolytic anemia, associated with HGB and HCT (SKAT $p < 8 \times 10^{-7}$). These rare, low-frequency, and common RBC variants showed pleiotropy, being also associated with platelet, white blood cell, and lipid traits. Our association results and functional annotation suggest the involvement of new genes in human erythropoiesis. We also confirm that rare and low-frequency variants play a role in the architecture of complex human traits, although their phenotypic effect is generally smaller than originally anticipated.

Introduction

One in four cells in the human body is a mature enucleated red blood cell (RBC), also called an erythrocyte. RBC mean lifespan in adults is 100–120 days, requiring constant renewal. To that end, we produce on average 2.4 million RBCs per second in the bone marrow. This massive yet well-orchestrated cell proliferation process is necessary to

[1]Department of Medicine, Université de Montréal, Montréal, QC H3T 1J4, Canada; [2]Montreal Heart Institute, Montréal, QC H1T 1C8, Canada; [3]Population Sciences Branch, National Heart, Lung, and Blood Institute, The Framingham Heart Study, Framingham, MA 01702, USA; [4]Genetics Target Sciences, GlaxoSmithKline, Research Triangle Park, NC 27709, USA; [5]OmicSoft Corporation, Cary, NC 27513, USA; [6]Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, London W2 1PG, UK; [7]Department of Hygiene and Epidemiology, University of Ioannina Medical School, Ioannina 45110, Greece; [8]Laboratory of Epidemiology and Population Sciences, National Institute on Aging, NIH, Baltimore, MD 21224, USA; [9]Department of Medicine, Center of Human Nutrition, Washington University School of Medicine, St Louis, MO 63110, USA; [10]Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics, University Medicine, Greifswald and Ernst-Mortiz-Arndt University Greifswald, Greifswald 17475, Germany; [11]DZHK (German Centre for Cardiovascular Research), partner site Greifswald, Greifswald QA, Germany; [12]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10069, USA; [13]Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA; [14]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA 02142, USA; [15]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA 02114, USA; [16]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA; [17]Division of Cardiovascular Medicine, Kanazawa University, Graduate School of Medical Science, Kanazawa, Ishikawa 9200942, Japan; [18]Division of Epidemiology, Department of Medicine, Institute for Medicine and Public Health, Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN 37235, USA; [19]Department of Medicine, University of Washington, Seattle, WA 98101, USA; [20]The Genetics of Obesity and Related Metabolic Traits Program, Icahn School of Medicine at Mount Sinai, New York, NY 10069, USA; [21]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN 55454, USA; [22]Department of Medicine/Division of General Internal Medicine, Johns Hopkins University, School of Medicine, Baltimore, MD 21205, USA; [23]Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA; [24]Department of Epidemiology, Erasmus, MC Rotterdam 3000, the Netherlands; [25]Estonian Genome Center, University of Tartu, Tartu 51010, Estonia; [26]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh EH8 9JZ, UK; [27]Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, UK; [28]Department of Genetics, University of North Carolina, Chapel Hill, NC 27514, USA; [29]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA 98195, USA; [30]Department of Biostatistics, University of Washington,

(Affiliations continued on next page)

© 2016 American Society of Human Genetics.

Please cite this article in press as: Chami et al., Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits, The American Journal of Human Genetics (2016), http://dx.doi.org/10.1016/j.ajhg.2016.05.007

Joel N. Hirschhorn,[14,50] Albert Hofman,[24,51] Marguerite R. Irvin,[52] Mika Kähönen,[53,54] Ethan Lange,[55] Lenore J. Launer,[44] Terho Lehtimäki,[46,47] Jin Li,[56] David C.M. Liewald,[26,27] Allan Linneberg,[57,58,59] Yongmei Liu,[60] Yingchang Lu,[12,20] Leo-Pekka Lyytikäinen,[46,47] Reedik Mägi,[25] Rasika A. Mathias,[61] Olle Melander,[38,39] Andres Metspalu,[25] Nina Mononen,[46,47] Mike A. Nalls,[62] Deborah A. Nickerson,[63] Kjell Nikus,[48,64] Chris J. O'Donnell,[3,65] Marju Orho-Melander,[38,39] Oluf Pedersen,[34] Astrid Petersmann,[66] Linda Polfus,[32] Bruce M. Psaty,[67,68] Olli T. Raitakari,[69,70] Emma Raitoharju,[46,47] Melissa Richard,[40] Kenneth M. Rice,[30] Fernando Rivadeneira,[24,71,72] Jerome I. Rotter,[73,74] Frank Schmidt,[10] Albert Vernon Smith,[42,43] John M. Starr,[26,75] Kent D. Taylor,[73,74] Alexander Teumer,[76] Betina H. Thuesen,[57] Eric S. Torstenson,[18] Russell P. Tracy,[77] Ioanna Tzoulaki,[6,7] Neil A. Zakai,[78] Caterina Vacchi-Suzzi,[79] Cornelia M. van Duijn,[24] Frank J.A. van Rooij,[24] Mary Cushman,[78] Ian J. Deary,[26,27] Digna R. Velez Edwards,[80] Anne-Claire Vergnaud,[6] Lars Wallentin,[81] Dawn M. Waterworth,[82] Harvey D. White,[83] James G. Wilson,[84] Alan B. Zonderman,[8] Sekar Kathiresan,[13,14,15,16] Niels Grarup,[34] Tõnu Esko,[14,25] Ruth J.F. Loos,[12,20,85] Leslie A. Lange,[28] Nauder Faraday,[86] Nada A. Abumrad,[9] Todd L. Edwards,[18] Santhi K. Ganesh,[87,91] Paul L. Auer,[88,91] Andrew D. Johnson,[3,91] Alexander P. Reiner,[89,90,91,*] and Guillaume Lettre[1,2,91,*]

accommodate RBCs' main function: to transport oxygen from the lungs to the peripheral organs, and carbon dioxide from the organs to the lungs. Hemoglobin (HGB), the metalloprotein that constitutes by far the most abundant biomolecule found in mature RBCs, is responsible for oxygen transport. In addition to their critical role in the circulatory system, RBCs also have secondary, often less-appreciated, functions. Within blood vessels, they respond

Seattle, WA 98195, USA; [31]Department of Medicine/Divisions of Cardiology and General Internal Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [32]Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [33]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA; [34]The Novo Nordisk Foundation, Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2100, Denmark; [35]TIMI Study Group, Cardiovascular Division, Brigham and Women's Hospital, Boston, MA 02115, USA; [36]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA 98195, USA; [37]Faculty of Pharmacy, Université de Montréal, Montréal, QC H3T 1J4, Canada; [38]Department of Clinical Sciences, Malmö, Lund University, Malmö 221 00, Sweden; [39]Skåne University Hospital, Malmö 222 41, Sweden; [40]Institute of Molecular Medicine, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [41]Institute for Immunology and Transfusion Medicine, University Medicine Greifswald, Greifswald 17475, Germany; [42]Icelandic Heart Association, 201 Kopavogur, Iceland; [43]Faculty of Medicine, University of Iceland, 101 Reykjavik, Iceland; [44]Laboratory of Epidemiology, Demography, and Biometry, National Institute on Aging, Intramural Research Program, NIH, Bethesda, MD 20892, USA; [45]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK; [46]Department of Clinical Chemistry, Fimlab Laboratories, Tampere 33520, Finland; [47]Department of Clinical Chemistry, University of Tampere School of Medicine, Tampere 33014, Finland; [48]University of Tampere, School of Medicine, Tampere 33014, Finland; [49]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA; [50]Department of Endocrinology, Boston Children's Hospital, Boston, MA 02115, USA; [51]Department of Epidemiology, Harvard TH Chan School of Public Health, Boston, MA 02115, USA; [52]Department of Epidemiology, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35233, USA; [53]Department of Clinical Physiology, Tampere University Hospital, Tampere 33521, Finland; [54]Department of Clinical Physiology, University of Tampere School of Medicine, Tampere 33014, Finland; [55]Departments of Genetics and Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [56]Department of Medicine, Division of Cardiovascular Medicine, Stanford University, School of Medicine, Palo Alto, CA 94305, USA; [57]Research Centre for Prevention and Health, The Capital Region of Denmark, Copenhagen 2600, Denmark; [58]Department of Clinical Experimental Research, Rigshospitalet, Glostrup 2100, Denmark; [59]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2200, Denmark; [60]Center for Human Genetics, Division of Public Health Sciences, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA; [61]Department of Medicine, Divisions of Allergy and Clinical Immunology and General Internal Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [62]Laboratory of Neurogenetics, National Institute on Aging, NIH, Bethesda, MD 20892, USA; [63]Department of Genome Sciences, University of Washington, Seattle, WA 98105, USA; [64]Department of Cardiology, Heart Center, Tampere University Hospital, Tampere 33521, Finland; [65]Cardiology Section and Center for Population Genomics, Boston Veteran's Administration (VA) Healthcare, Boston, MA 02118, USA; [66]Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald 17475, Germany; [67]Cardiovascular Health Research Unit, Departments of Medicine Epidemiology and Health Services, University of Washington, Seattle, WA 98101, USA; [68]Group Health Research Institute, Group Health Cooperative, Seattle, WA 98101, USA; [69]Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku 20521, Finland; [70]Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku 20520, Finland; [71]Department of Internal Medicine, Erasmus MC, Rotterdam 3000, the Netherlands; [72]Netherlands Consortium for Healthy Ageing (NCHA), Rotterdam 3015, the Netherlands; [73]Institute for Translational Genomics and Population Sciences, Los Angeles Biomedical Research Institute, Torrance, CA 90502, USA; [74]Department of Pediatrics, Harbor-UCLA Medical Center, Torrance, CA 90502, USA; [75]Alzheimer Scotland Research Centre, Edinburgh EH8 9JZ, UK; [76]Institute for Community Medicine, University Medicine Greifswald, Greifswald 17475, Germany; [77]Departments of Pathology and Laboratory Medicine and Biochemistry, University of Vermont College of Medicine, Colchester, VT 05446, USA; [78]Departments of Medicine and Pathology, University of Vermont College of Medicine, Burlington, VT 05405, USA; [79]Department of Family Population and Preventive Medicine, Stony Brook University, Stony Brook, NY 11794, USA; [80]Vanderbilt Epidemiology Center, Department of Obstetrics & Gynecology, Institute for Medicine and Public Health, Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN 37203, USA; [81]Department of Medical Sciences, Cardiology and Uppsala Clinical Research Center, Uppsala University, Uppsala 751 85, Sweden; [82]Genetics Target Sciences, GlaxoSmithKline, King of Prussia, PA 19406, USA; [83]Green Lane Cardiovascular Service, Auckland City Hospital and University of Auckland, Auckland 1142, New Zealand; [84]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS 39216, USA; [85]The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10069, USA; [86]Department of Anesthesiology & Critical Care Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [87]Departments of Internal Medicine and Human Genetics, University of Michigan, Ann Arbor, MI 48108, USA; [88]Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI 53205, USA; [89]Department of Epidemiology, University of Washington, Seattle, WA 98195, USA; [90]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
[91]These authors contributed equally to this work
*Correspondence: apreiner@u.washington.edu (A.P.R.), guillaume.lettre@umontreal.ca (G.L.)
http://dx.doi.org/10.1016/j.ajhg.2016.05.007.

to shear stress and produce the vasodilator nitric oxide to regulate vascular tonus.[1] RBCs participate in antimicrobial strategies to fight hemolytic pathogens[2] and in the inflammatory response, acting as a reservoir for multiple chemokines.[3] Furthermore, the direct involvement of RBCs in adhering to the vascular endothelium or supporting thrombin generation may help to promote blood coagulation or thrombosis.[4,5]

Given the paramount importance of RBCs in physiology, it is not surprising that monitoring their features is common practice in medicine to assess the overall health of patients. An excessive number of circulating RBCs (erythrocytosis [MIM: 133100]) can suggest a primary bone marrow disease, a myeloproliferative neoplasm such as polycythemia vera (MIM: 263300), or chronic hypoxemia due to congenital heart defects. Low HGB concentration and hematocrit (HCT) levels (anemia) can indicate inherited HGB or RBC structural gene mutations, malnutrition, or kidney diseases. By considering the volume (mean corpuscular volume [MCV]), hemoglobin content (mean corpuscular hemoglobin [MCH] and mean corpuscular hemoglobin concentration [MCHC]) or the distribution width (RDW) of RBCs, a physician can distinguish between the different causes of anemia (e.g., microcytic/hypochromic due to iron deficiency[6]). In addition, epidemiological studies have correlated high RDW values with a worse prognosis in heart failure patients.[7] RDW is also an independent predictor of overall mortality in healthy individuals, as well as a predictor of mortality in patients with various conditions such as cardiovascular diseases, obesity, malignancies, and chronic kidney disease.[8–12]

RBC count and indices vary among individuals, and 40%–90% of this phenotypic variation is heritable.[13–16] Identifying the genes and biological pathways that contribute to this inter-individual variation in RBC traits could highlight modifiers of severity and/or therapeutic options for several hematological diseases. Already, large-scale genome-wide association studies (GWASs) have found dozens of SNPs associated with one or more of these RBC traits.[17,18] However, owing to their design, GWASs are largely insensitive to rare (minor allele frequency [MAF] $< 1\%$) and low-frequency ($1\% \leq$ MAF $< 5\%$) genetic variants. Using an exome array, we previously performed an association study for HGB and HCT in 31,340 European-ancestry individuals and identified rare coding or splice site variants in the erythropoietin and β-globin genes.[19] Within the framework of the Blood-Cell Consortium (BCX),[20,21] we now report a larger genotyping-based exome survey of seven RBC traits conducted in up to 130,273 individuals, including 23,896 participants of non-European ancestry. With this experiment, our initial goals were to expand the list of rare and low-frequency coding or splice site variants associated with RBC traits and to explore whether the exome array can complement the GWAS approach to fine map RBC causal genes.

Subjects and Methods

Study Participants

The Blood-Cell Consortium (BCX) aims to identify novel common and rare variants associated with blood-cell traits using an exome array. BCX is comprised of more than 134,021 participants from 24 discovery cohorts and five ancestries: European, African American, Hispanic, East Asian, and South Asian. Detailed description of the participating cohorts is provided in Table S1. BCX is interested in the genetics of all main hematological measures and is divided into three main working groups: RBC, white blood cell (WBC),[21] and platelet (PLT).[20] For the RBC working group, we analyzed seven traits available in up to 130,273 individuals: RBC count ($\times 10^{12}$/L), HGB (g/dL), HCT (%), MCV (fL), MCH (pg), MCHC (g/dL), and RDW (%) (Table S2). The BCX procedures were in accordance with the institutional and national ethical standards of the responsible committees and proper informed consent was obtained.

Genotyping and Quality-Control Steps

Participants from the different studies were genotyped on one of the following exome chip genotyping arrays: Illumina ExomeChip v.1.0, Illumina ExomeChip v.1.1_A, Illumina ExomeChip-12 v.1.1, Affymetrix Axiom Biobank Plus GSKBB1, or Illumina HumanOmniExpressExome Chip. Genotypes were then called either (1) with the Illumina GenomeStudio GENCALL and subsequently recalled using zCALL or (2) by the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium Exome Chip effort[22] (Table S3). The same quality-control steps were followed by each participating study. We excluded variants with low genotyping success rate (<95%, except for WHI that used a cutoff <90%) (Table S3). Samples with call rate $< 95\%$ (except for SOLID-TIMI 52 and STABILITY that used 94.5% and 93.5% cutoffs, respectively) after joint or zCALL calling and with outlying heterozygosity rate were also excluded. Other exclusions were deviation from Hardy-Weinberg equilibrium ($p < 1 \times 10^{-6}$) and gender mismatch. We performed principal-component analysis (PCA) or multidimensional scaling (MDS) and excluded sample outliers from the resulting plots through visual inspection, using populations from the 1000 Genomes Project to anchor these analyses. Keeping only autosomal and X chromosome variants for the analysis, we aligned all variants on the forward strand and created a uniform list of reference alleles using the --force alleles command in PLINK.[23] Finally, an indexed variant call format file (VCF) was created by each study and checked for allele alignment and any allele or strand flips using the checkVCF package.[24] Prior to performing meta-analyses of the association results provided by each participating study, we ran the EasyQC protocol[25] to check for population allele frequency deviations and proper trait transformation in each cohort.

Phenotype Modeling and Association Analyses

When possible, we excluded individuals with blood cancer, leukemia, lymphoma, bone marrow transplant, congenital or hereditary anemia, HIV, end-stage kidney disease, dialysis, splenectomy, or cirrhosis and those with extreme measurements of RBC traits (Table S1). We also excluded individuals on erythropoietin treatment or chemotherapy. Additionally, we excluded pregnant women and individuals with acute medical illness at the time the complete blood count (CBC) was done. For the seven RBC traits, within each study, we adjusted for age, age-squared, gender,

the first ten principal components, and, where applicable, other study-specific covariates such as study center via a linear regression model. Within each study, we then applied inverse normal transformation on the residuals and tested the phenotypes for association with the ExomeChip variants using either RVtests (v.20140416)[26] or RAREMETALWORKER.0.4.9.[27]

Discovery Meta-analyses

Score files generated by RVtests or RAREMETALWORKER from each participating study were used to carry out meta-analyses of the single variant association results using RareMETALS v.5.9.[28] All analyses were performed separately in each of European American (EA) and African American (AA) ancestries. In the multi-ancestry meta-analyses, we combined individuals of European, African American, Hispanic, East-Asian, and South-Asian ancestries (All). We included variants with allele frequency difference between the highest and lowest MAF < 0.3 for EA and AA ancestries and < 0.6 for the combined ancestry meta-analyses. For the gene-based analyses, we used score files and variance-covariance matrices from the study-specific association results and applied the sequence kernel association test (SKAT)[29] and variable threshold (VT) algorithms[30] in RareMETALS considering only missense, nonsense, and splice site variants with a MAF $< 1\%$. Gene-based analyses were also stratified by ancestry. Significance thresholds were determined using Bonferroni correction assuming ~250,000 independent variants ($p < 2 \times 10^{-7}$ for the single-variant analyses) and ~17,000 genes tested on the ExomeChip ($p < 3 \times 10^{-6}$ for the gene-based tests).

Conditional Analysis and Replication

In order to identify independent signals, we performed conditional analyses. In each round of conditional analysis, we conditioned on the most significant single variant in a 1 Mb window. These conditional analyses were performed at the meta-analysis level using RareMETALS. We repeated this step until there were no new signals identified in each region, defined as $p < 2 \times 10^{-7}$. We then checked for linkage disequilibrium (LD) within the list of variants that was retained from the conditional analyses. For variants that were in moderate-to-strong LD ($r^2 \geq 0.3$), we kept the most significant. We attempted replication of the final list of independent variants in eight additional studies that contributed a total of 27,480 individuals ($n = 21{,}473$ for EA and $n = 6{,}007$ for AA) (Table S4). The division of discovery and replication samples was dictated by timing because we collected all groups we were aware of for initial discovery and then found others who could participate only much later and hence were used for replication. These studies followed similar analytical procedures and steps as those followed by the discovery analysis (see above). A joint meta-analysis of the discovery and the replication results was carried out using a fixed-effects model and inverse-variance weighting as implemented in METAL.[31] We considered as replicated markers those with a nominal $p_{replication} < 0.05$ and an effect on phenotype in the same direction as in the discovery results.

Allelic Imbalance and Expression of *CD36*

We checked for allelic imbalance (AI) of the rs3211938 variant in *CD36* (MIM: 173510) as well as the expression of the gene in 12 samples of fetal liver erythroblasts obtained from anonymous donors. Details on the protocol including RNA extraction and sequencing can be found elsewhere.[32] We calculated the difference in the ratio of reads of the reference allele (T) and the alternate allele (G) of rs3211938. In brief, reads overlapping rs3211938 were counted with samtools (v.1.1) mpileup software using genome build hg19. We kept uniquely mapping reads using -q 50 argument (mapping quality > 50) and sites with base quality > 10. Statistical significance of the difference in the ratio of reads between the reference allele and the alternate allele was assessed with a binomial test. For each sample, we summed all reads overlapping all heterozygous SNPs and calculated the expected ratio within each SNP allele combination. Reads that fall in the top 25th coverage percentile were down-sampled so that the highest covered sites do not bias the expected ratio.[33] For rs3211938, the expected T:G ratio was 0.507.

Expression Quantitative Trait Loci Analysis

We cross-referenced our list of RBC novel variants with more than 100 separate expression quantitative trait loci (eQTL) published datasets. Datasets were collected through publications, publically available sources, and private collaborations. A general overview of a subset of >50 eQTL studies has been published,[34] with specific citations for >100 datasets included in the current query followed here. A complete list of tissues and studies used can be found in the Supplemental Data. We considered SNPs that are themselves expression SNPs (eSNP) when they meet a $p < 0.0001$ threshold or when they are in LD ($r^2 > 0.3$) with the best eSNP ($p < 0.0001$).

Results

Single-Variant Meta-analyses

We meta-analyzed ExomeChip results for seven RBC-related phenotypes (RBC count, HCT, HGB, MCH, MCHC, MCV, and RDW) available in up to 130,273 individuals from 24 studies and 5 ancestries (Tables S1–S3 and Figure S1). Across these different phenotypes, a total of 226 variants reached exome-wide significance ($p < 2 \times 10^{-7}$) in the combined ancestry analyses (Figures 1 and S2). Given that some of these RBC traits are correlated (Figure S3), these associated variants highlight 71 different loci (defined using a 1 Mb interval). Overall, we observed only modest inflation of the test statistics ($\lambda_{GC} = 1.03$–1.05), consistent with little confounding due to technical artifacts, population stratification, or cryptic relatedness.

In order to identify independent variants, we performed conditional analyses at the meta-analysis level adjusting for the effect of the most significant variant in a 1 Mb region in a stepwise manner (Subjects and Methods). After this analysis, we obtained a list of 126 independent variants associated with at least one RBC trait at $p < 2 \times 10^{-7}$ (Table S5). Selecting only variants that lie more than 1 Mb away from a known GWAS locus resulted in 23 independent variants located within 20 novel RBC loci, where novel is used to define loci not found in the existing literature (Table 1). We attempted to replicate these 126 variants in 8 independent cohorts totaling 27,480 participants (Table S5). Overall, we observed a strong replication, with 94 of the 126 variants showing consistent direction of effect between the discovery and replication analyses (binomial $p = 3 \times 10^{-8}$; Table S5). Of the 23 novel RBC variants, we replicated 16 at nominal $p < 0.05$ for at

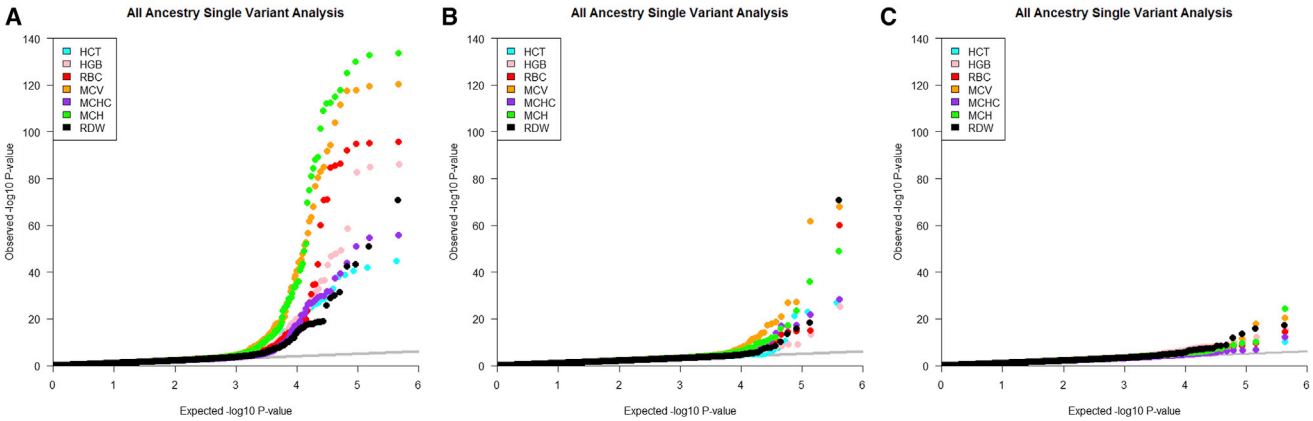


Figure 1. Quantile-Quantile Plots of Single-Variant Association Results in the All Ancestry Meta-analyses for the Seven Red Blood Cell Traits Analyzed

(A) Distribution of the single variant results for all variants tested on the exome array.

(B) Only markers with a minor allele frequency < 5% are shown here.

(C) Variants outside of known RBC GWAS regions. Variants that are within 1 Mb from a previously published RBC GWAS locus were excluded for this QQ plot.

Abbreviations are as follows: HCT, hematocrit; HGB, hemoglobin; RBC, red blood cell count; MCV, mean corpuscular volume; MCHC, mean corpuscular hemoglobin concentration; MCH, mean corpuscular hemoglobin; RDW, red blood cell distribution width.

least one RBC trait (binomial $p = 3 \times 10^{-16}$; Table 1). Out of these 16 novel and replicated RBC variants, there are five missense variants, including two variants with MAF < 5% in *MAP1A* (MIM: 600178) and *HNF4A* (MIM: 600281) and one nonsense variant in *CD36* (Table 1). Among the remaining nine novel and replicated RBC variants, there are five intronic, one synonymous, one 5′ UTR, and one intergenic marker (Table 1).

Prioritization of Candidate Genes and Genetic Variants

Our single-variant analyses in EA samples identified one rare missense variant in *ALAS2* (MIM: 301300) associated with MCV and MCH (rs201062903, p.Pro507Leu [c.1559C>T], MAF = 0.2%) (Table 1). The association with this variant did not replicate, potentially because of limited statistical power (the replication sample size for this rare marker was 5,044; see also Discussion). *ALAS2* encodes 5-aminolevulinate synthase 2, the rate-controlling enzyme of erythroid heme synthesis. Additionally, rare mutations in *ALAS2* cause X-linked sideroblastic anemia (MIM: 300751) and erythropoietic protoporphyria (MIM: 300752). Thus, despite the lack of replication, *ALAS2* remains an excellent candidate gene to modulate RBC traits. The *ALAS2* p.Pro507Leu variant, which is not reported in the ClinVar database, maps between two amino acids (Tyr506 and Thr508) that are important for catalytic activity and known to be mutated in cases of sideroblastic anemia.[35]

Two low-frequency missense variants identified in our analyses implicate *MAP1A* and *HNF4A* in RBC biology (Table 1). *MAP1A* encodes microtubule-associated protein 1A, a gene highly expressed in the nervous system and mostly studied in the context of neuronal diseases, although it is expressed in many additional tissues, including hematopoietic cells.[36] Deletion of *MAP1A* in the mouse causes defects in synaptic plasticity.[37] This observation is interesting given that inactivation of *ANK1* (MIM: 612641), another gene that encodes a cytoskeleton protein and is expressed in neurons and RBCs, is associated with neurological dysfunction in the mouse and spherocytosis and hemolytic anemia in humans (MIM: 182900). Our meta-analyses confirmed two known independent *ANK1* variants associated with MCHC: an intronic SNP (rs4737009, MAF = 19.8%, $p = 1.3 \times 10^{-8}$) and a low-frequency missense variant (rs34664882, p.Ala1462Val, MAF = 2.9%, $p = 1.7 \times 10^{-16}$) (Table S5; N.P., U.M.S., J.B.-J., and M.-H.C., unpublished data).[17]

In the accompanying BCX PLT article,[20] we report that the same *MAP1A* rs55707100 allele (p.Pro2349Leu [c.7046C>T]) associated here with decreased HGB concentration is also associated with increased PLT count. Furthermore, recent studies have identified associations between rs55707100 and HDL-cholesterol and triglyceride levels (S. Mukherjee, 2015, ASHG, conference). Adding to the complexity, the GTEx dataset indicates that rs55707100 is an expression quantitative trait locus (eQTL) for *ADAL* ($p_{eQTL} = 9 \times 10^{-11}$) but not for *MAP1A*.[38] *ADAL* is a poorly characterized adenosine deaminase-like protein that is highly expressed in human erythroblasts. However, the eQTL association between rs55707100 and *ADAL* could simply reflect "LD shadowing" from nearby markers that are much stronger eQTL variants for *ADAL*. Indeed, rs3742971 (a common variant located in *ADAL*'s 5′ UTR) is in partial LD with rs55707100 ($r^2 = 0.18$ in European populations from the 1000 Genomes Project) and strongly associated with *ADAL* expression levels ($p_{eQTL} = 6 \times 10^{-49}$).

The second low-frequency missense variant associated with HGB and HCT maps within the coding sequence of

Please cite this article in press as: Chami et al., Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits, The American Journal of Human Genetics (2016), http://dx.doi.org/10.1016/j.ajhg.2016.05.007

Table 1. Association Results of Variants in Novel Loci Associated with Red Blood Cell Traits

Marker Info						Discovery				Replication				Combined	
Trait	Position	A1/A2	SNP	Annotation	Gene	n	AF (A2)	Beta (SE)	p Value	n	AF (A2)	Beta (SE)	p Value	Beta (SE)	p Value
RDW-EA	1: 25,768,937	A/G	rs10903129*	intron	*TMEM57-RHD*	45,573	0.544	0.037 (0.007)	1.19×10^{-7}	18,475	0.560	0.023 (0.011)	0.0373	0.033 (0.006)	2.41×10^{-8}
RDW-All	1: 25,768,937	A/G	rs10903129*	intron	*TMEM57-RHD*	56,194	0.568	0.034 (0.006)	9.58×10^{-8}	24,474	0.600	0.021 (0.01)	0.0252	0.03 (0.005)	1.32×10^{-8}
HCT-All	1: 155,162,067	C/T	rs4072037*	synonymous	*MUC1*	109,875	0.554	0.025 (0.005)	5.82×10^{-8}	25,006	0.563	0.038 (0.009)	5.96×10^{-5}	0.027 (0.004)	3.47×10^{-11}
HGB-All	2: 27,741,237	T/C	rs780094	intron	*GCKR*	130,273	0.626	0.024 (0.004)	7.14×10^{-8}	3,162	0.626	−0.012 (0.026)	0.6410	0.023 (0.044)	1.68×10^{-7}
RBC-All	2: 219,509,618	C/A	rs2230115*	missense	*ZNF142*	74,488	0.509	0.033 (0.006)	9.74×10^{-9}	27,442	0.477	0.024 (0.01)	0.0167	0.031 (0.005)	7.11×10^{-10}
HCT-All	3: 56,771,251	A/C	rs3772219*	missense	*ARHGEF3*	109,875	0.338	−0.028 (0.005)	2.38×10^{-9}	25,006	0.366	−0.021 (0.01)	0.0292	−0.027 (0.004)	2.56×10^{-10}
HGB-All	3: 56,771,251	A/C	rs3772219*	missense	*ARHGEF3*	130,273	0.336	−0.026 (0.004)	3.76×10^{-9}	27,749	0.367	−0.02 (0.009)	0.0331	−0.025 (0.004)	4.33×10^{-10}
HCT-EA	4: 88,008,782	G/A	rs236985	intron	*AFF1*	87,444	0.394	0.032 (0.005)	3.89×10^{-10}	19,968	0.405	0.02 (0.011)	0.0626	0.03 (0.005)	1.14×10^{-10}
RBC-EA	4: 88,008,782	G/A	rs236985*	intron	*AFF1*	60,231	0.393	0.034 (0.006)	3.50×10^{-8}	21,435	0.405	0.023 (0.011)	0.0273	0.031 (0.005)	4.22×10^{-9}
HGB-EA	4: 88,030,261	G/T	rs442177*	intron	*AFF1*	106,377	0.595	−0.034 (0.005)	3.97×10^{-13}	21,743	0.586	−0.029 (0.01)	0.0052	−0.033 (0.004)	8.23×10^{-15}
RDW-EA	5: 127,371,588	A/G	rs10063647*	intron	*LINC01184-SLC12A2*	45,573	0.463	−0.05 (0.007)	1.72×10^{-13}	18,475	0.480	−0.033 (0.011)	0.0018	−0.045 (0.006)	2.88×10^{-15}
RDW-All	5: 127,371,588	A/G	rs10063647*	intron	*LINC01184-SLC12A2*	56,194	0.506	−0.044 (0.006)	2.11×10^{-12}	24,474	0.545	−0.03 (0.01)	0.0014	−0.04 (0.005)	2.37×10^{-14}
RDW-EA	5: 127,522,543	C/T	rs10089*	utr_5p	*LINC01184-SLC12A2*	45,573	0.21	0.051 (0.008)	8.45×10^{-10}	16,692	0.215	0.058 (0.014)	2.71×10^{-5}	0.053 (0.007)	1.15×10^{-13}
RDW-All	5: 127,522,543	C/T	rs10089*	utr_5p	*LINC01184-SLC12A2*	56,194	0.207	0.044 (0.008)	4.08×10^{-9}	22,691	0.208	0.045 (0.012)	0.0001	0.044 (0.006)	2.73×10^{-12}
HGB-All	6: 7,247,344	C/A	rs35742417*	missense	*RREB1*	130,273	0.174	0.030 (0.005)	1.17×10^{-8}	4,074	0.207	0.065 (0.028)	0.0190	0.032 (0.005)	1.50×10^{-9}
RDW-AA	7: 80,300,449	T/G	rs3211938*	nonsense	*CD36*	6,666	0.087	0.174 (0.031)	2.36×10^{-8}	5,999	0.086	0.139 (0.032)	1.83×10^{-5}	0.161 (0.025)	7.09×10^{-11}
RDW-All	7: 80,300,449	T/G	rs3211938*	nonsense	*CD36*	55,510	0.012	0.171 (0.029)	5.29×10^{-9}	22,691	0.023	0.139 (0.032)	1.61×10^{-5}	0.157 (0.022)	5.12×10^{-13}
RDW-EA	8: 126,490,972	A/T	rs2954029*	intergenic	*TRIB1*	45,573	0.46	0.036 (0.007)	1.53×10^{-7}	16,692	0.466	0.026 (0.011)	0.0210	0.034 (0.006)	1.29×10^{-8}
RDW-All	8: 126,490,972	A/T	rs2954029*	intergenic	*TRIB1*	56,194	0.439	0.032 (0.006)	1.83×10^{-7}	22,691	0.432	0.021 (0.01)	0.0298	0.029 (0.005)	2.54×10^{-8}
MCH-All	10: 105,659,826	T/C	rs2487999	missense	*OBFC1*	66,318	0.869	0.047 (0.009)	4.12×10^{-8}	26,749	0.861	0.025 (0.013)	0.0601	0.041 (0.007)	1.75×10^{-8}
MCH-AA	11: 92,722,761	G/A	rs1447352	intergenic	*MTNR1B*	8,273	0.557	0.089 (0.016)	1.85×10^{-8}	5,038	0.562	−0.022 (0.02)	0.2713	0.07 (0.014)	1.08×10^{-6}
HGB-EA	15: 43,820,717	C/T	rs55707100*	missense	*MAP1A*	106,377	0.033	−0.071 (0.013)	1.65×10^{-8}	21,743	0.0223	−0.099 (0.033)	0.0028	−0.075 (0.012)	2.31×10^{-10}
MCV-AA	16: 1,551,082	A/G	rs2667662*	intron	*TELO2*	10,849	0.725	−0.099 (0.015)	1.79×10^{-10}	5,034	0.724	−0.093 (0.022)	3.02×10^{-5}	−0.098 (0.014)	7.32×10^{-12}
MCV-AA	16: 2,812,939	C/A	rs2240140*	missense	*SRRM2*	8,525	0.118	0.134 (0.025)	7.08×10^{-8}	6,002	0.124	0.106 (0.027)	0.0001	0.128 (0.022)	5.24×10^{-9}

(*Continued on next page*)

Table 1. ***Continued***

Marker Info						Discovery				Replication				Combined	
Trait	**Position**	**A1/A2**	**SNP**	**Annotation**	**Gene**	**n**	**AF (A2)**	**Beta (SE)**	**p Value**	**n**	**AF (A2)**	**Beta (SE)**	**p Value**	**Beta (SE)**	**p Value**
HCT-EA	17: 59,017,025	T/C	rs8080784	intron	*BCAS3-TBX2*	79,344	0.158	−0.039 (0.007)	2.62×10^{-8}	19,968	0.148	0.011 (0.014)	0.4349	−0.029 (0.006)	3.39×10^{-6}
HGB-EA	17: 59,483,766	C/T	rs8068318	intron	*BCAS3-TBX2*	106,377	0.722	−0.026 (0.005)	1.53×10^{-7}	21,743	0.730	−0.021 (0.011)	0.0565	−0.025 (0.005)	2.55×10^{-8}
MCV-EA	20: 31,140,165	C/T	rs4911241*	intron	*NOL4L*	61,462	0.241	−0.04 (0.007)	1.25×10^{-8}	21,714	0.252	−0.025 (0.012)	0.0302	−0.036 (0.006)	2.01×10^{-9}
RDW-EA	20: 31,140,165	C/T	rs4911241*	intron	*NOL4L*	45,573	0.242	0.043 (0.008)	5.79×10^{-8}	18,475	0.240	0.049 (0.012)	7.44×10^{-5}	0.045 (0.007)	2.01×10^{-11}
RDW-All	20: 31,140,165	C/T	rs4911241*	intron	*NOL4L*	56,194	0.235	0.038 (0.007)	1.56×10^{-7}	24,474	0.222	0.044 (0.011)	6.10×10^{-5}	0.04 (0.006)	4.60×10^{-11}
HCT-EA	20: 43,042,364	C/T	rs1800961*	missense	*HNF4A*	79,344	0.024	0.083 (0.015)	1.44×10^{-8}	19,968	0.033	0.082 (0.028)	0.0037	0.083 (0.013)	1.91×10^{-10}
HGB-EA	20: 43,042,364	C/T	rs1800961*	missense	*HNF4A*	98,277	0.032	0.073 (0.013)	2.53×10^{-8}	21,743	0.032	0.062 (0.027)	0.0232	0.071 (0.012)	1.93×10^{-9}
HCT-All	20: 43,042,364	C/T	rs1800961*	missense	*HNF4A*	100,313	0.022	0.077 (0.014)	2.31×10^{-8}	25,006	0.027	0.091 (0.028)	0.0010	0.08 (0.012)	9.88×10^{-11}
HGB-All	22: 44,324,727	C/G	rs738409	missense	*PNPLA3*	130,273	0.223	0.028 (0.005)	2.24×10^{-8}	4,074	0.218	0.053 (0.027)	0.0504	0.029 (0.005)	4.81×10^{-9}
MCH-EA	X: 55,039,960	G/A	rs201062903	missense	*ALAS2*	52,758	0.002	−0.324 (0.053)	7.32×10^{-10}	5,855	0.001	−0.291 (0.235)	0.215	−0.323 (0.052)	5.81×10^{-10}
MCH-All	X: 55,039,960	G/A	rs201062903	missense	*ALAS2*	65,067	0.002	−0.322 (0.051)	3.36×10^{-10}	10,893	0.001	−0.276 (0.224)	0.218	−0.321 (0.051)	2.68×10^{-10}
MCV-EA	X: 55,039,960	G/A	rs201062903	missense	*ALAS2*	60,211	0.002	−0.285 (0.049)	7.11×10^{-9}	5,044	0.001	−0.178 (0.248)	0.472	−0.282 (0.049)	6.11×10^{-9}

Variants in novel loci with $p < 2 \times 10^{-7}$ and that were retained after conditional analyses are presented here. All variants are >1 Mb apart from a known GWAS signal for RBC traits. Chromosome positions are given on human genome build hg19. Allele frequency and effect size are given for the alternate (A2) allele. Replication was carried out in six cohorts for EA and two cohorts for AA and was performed in RareMetals; meta-analyses of the discovery and replication cohorts are presented under "Combined" and were carried out in METAL. Asterisks (*) indicate variants that replicated with a nominal $p < 0.05$. Abbreviations are as follows: EA, European American; AA, African American; All, combined ancestry (EA + AA + Asians + Hispanics); A1, reference allele; A2, alternate allele; N, sample size; AF, allele frequency; SE, standard error; HCT, hematocrit; HGB, hemoglobin; RBC, red blood cell count; MCV, mean corpuscular volume; MCHC, mean corpuscular hemoglobin concentration; MCH, mean corpuscular hemoglobin; RDW, red blood cell distribution width.

Please cite this article in press as: Chami et al., Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits, The American Journal of Human Genetics (2016), http://dx.doi.org/10.1016/j.ajhg.2016.05.007

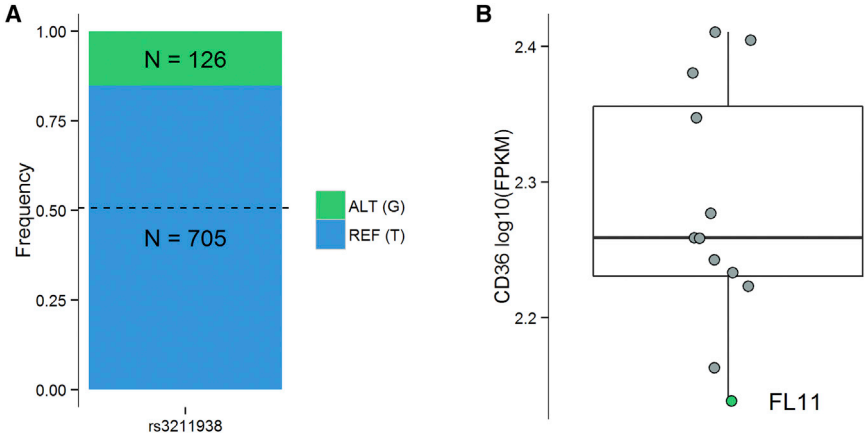


Figure 2. *CD36* Expression in Human Erythroblasts
(A) In a dataset of 12 human fetal liver erythroblasts, all samples were homozygous at rs3211938 for the reference T-allele with the exception of one heterozygous sample (FL11). FL11 demonstrated strong allelic imbalance: we observed 705 reads for the reference allele (T) and 126 reads for the alternate allele (G) (binomial $p = 4.9 \times 10^{-95}$).
(B) FL11 (in green) shows the lowest *CD36* expression level when compared to the other 11 samples. Abbreviation is as follows: FPKM, fragments per kilobase of transcript per million mapped reads.

the transcription factor HNF4A (Table 1). This marker, rs1800961 (p.Thr117Ile [c.350C>T]), has previously been associated with HDL and total cholesterol, C-reactive protein, fibrinogen, and coagulation factor VII levels.[39–42] Mutations in *HNF4A* cause maturity-onset diabetes of the young (MODY [MIM: 125851]) and a common intronic SNP in *HNF4A* (rs4812829) has been associated with type 2 diabetes (MIM: 125853) risk.[43] The missense rs1800961 associated with HGB and HCT is only in weak LD with rs4812829 ($r^2 = 0.021$ in EA populations from the 1000 Genomes Project). Querying recently released ExomeChip data from Type 2 Diabetes Genetics (Web Resources), we found that rs1800961 is also associated with T2D risk in ~82,000 participants ($p = 9.5 \times 10^{-7}$, odds ratio = 1.16). *HNF4A* is expressed in the kidney and could influence HGB and HCT through the regulation of erythropoietin production.[44] It is also abundantly expressed in the liver, where it could indirectly affect HGB and HCT levels through an effect on blood lipid levels (see Discussion). *HNF4A* is detectable at low levels in erythroblasts, and the BLUEPRINT Project has found that some *HNF4A* isoforms may be more highly expressed in this cell type (Figure S4).[45]

In AA, we identified a nonsense variant (rs3211938, p.Tyr325Ter [c.975T>G], MAF = 8.7%, $p = 7.1 \times 10^{-11}$) in *CD36* associated with RDW. This variant displays a wide variation in allele frequency between AA and EA ($MAF_{EA} = 0.01\%$). The association is slightly improved in the ancestry-combined meta-analysis ($p = 5.1 \times 10^{-13}$) because there is also evidence of association in Hispanics (MAF = 1.9%, p = 0.022) (Table 1). We examined a dataset of ex vivo differentiated human erythroblasts to determine whether this *CD36* nonsense variant shows allelic imbalance (AI).[32] All samples were homozygous at rs3211938 for the reference allele with the exception of one heterozygous sample (FL11). FL11 had the lowest level of *CD36* expression among the 12 samples tested and demonstrated strong AI where we observe 705 sequence reads for the reference allele (T) versus 126 for the alternate allele (G) ($p = 4.9 \times 10^{-95}$; Figure 2). To confirm this finding in independent samples, we queried the GTEx dataset, which has compiled RNA-sequencing and genotype information from multiple human tissues.[38] GTEx does not include data for human erythroblasts. However, it detected a strong eQTL effect of rs3211938 on *CD36* expression in whole blood ($p_{eQTL} = 1.1 \times 10^{-15}$), with carriers of the G-allele expressing less *CD36* (Figure S5). Furthermore, GTEx reported evidence for moderate AI in multiple tissues for *CD36*-rs3211938, with the G-allele being under-represented among sequence reads (Figure S5). These results strongly support our observations in human erythroblasts.

eQTL Analysis

To prioritize additional causal genes at RBC loci that contain non-coding variants, we cross-referenced our list of novel variants with more than 100 published eQTL datasets (Subjects and Methods). Overall, 12 variants were significant eQTLs in a wide variety of tissues (Table S6). The most interesting eQTL finding is the association between rs10903129, a common marker associated with RDW in our analyses and located within an intron of *TMEM57* (MIM: 610301), and the expression of *RHD* (MIM: 111680) in whole blood. *RHD* is located 112 kb downstream of *TMEM57* and encodes the D antigen of the clinically significant Rhesus (Rh) blood group. rs10903129 has also been associated with total cholesterol levels and erythrocyte sedimentation rate (ESR).[46,47] The association with ESR is particularly intriguing given that it is considered a non-specific indicator of inflammation. As described above, RDW is also abnormal in chronic diseases, such as atherosclerosis and diabetes, which have an important inflammation component.

Gene-Based Association Testing

Despite our large sample size, statistical power remains limited for rare variants of weak-to-moderate phenotypic effect. To try to capture these genetic factors, we performed gene-based testing by aggregating coding and splice site variants with MAF < 1% within each gene (Subjects and Methods). The SKAT analyses identified two genes: *ALAS2* associated with MCH and *PKLR* (MIM: 609712) associated with HGB and HCT (Table 2). The *ALAS2* signal was driven

Table 2. Gene-Based Association Results

Trait	Gene	n	Number of Variants Analyzed	VT p Value	SKAT p Value	Top Variant	Top-Variant MAF	Top-Variant p Value
HGB-EA	*PKLR*	106,377	15	1.92×10^{-5}	7.02×10^{-7}	rs116100695	0.003	1.17×10^{-5}
HGB-All	*PKLR*	130,273	15	0.00016	6.57×10^{-7}	rs116100695	0.003	1.94×10^{-5}
HCT-All	*PKLR*	109,875	15	3.96×10^{-5}	7.95×10^{-7}	rs116100695	0.003	2.49×10^{-5}
MCH-EA	*ALAS2*	54,009	11	4.78×10^{-6}	5.79×10^{-7}	rs201062903	0.002	7.32×10^{-10}
MCHC-All	*ALPK3*	84,841	28	1.95×10^{-6}	0.793	rs202037221	3.0×10^{-5}	0.0005

Gene-based results of the VT and SKAT algorithms for genes associated with RBC traits at $p < 3 \times 10^{-6}$. We analyzed non-synonymous coding (nonsense, missense) and splice site variants with a minor allele frequency (MAF) < 1%. Abbreviations are as follows: EA, European American; All, combined ancestry (EA + AA + Asians + Hispanics); n, sample size; HCT, hematocrit; HGB, hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCH, mean corpuscular hemoglobin.

by a single rare missense variant (rs201062903) and was described above. *PKLR* encodes the erythrocyte pyruvate kinase (PK) that catalyzes the last step of glycolysis. PK deficiency, usually caused by recessive mutations, is one of the main causes of non-spherocytic hemolytic anemia (MIM: 266200). In fact, one of the variants identified in our meta-analysis (rs116100695, p.Arg486Trp [c.1456T>G], MAF = 0.3%, $beta_{HGB} = -0.242$ g/dl, $p_{HGB} = 1.2 \times 10^{-5}$) is a frequent cause of PK deficiency in Italian and Spanish subjects.[48,49] This variant was confirmed in the replication cohorts ($p_{replication} = 0.039$; Table S7). Two additional *PKLR* rare missense variants contribute to the gene-based association statistic with HGB and HCT: rs61755431 (p.Arg569Gln [c.1706G>A], MAF = 0.2%, $beta_{HGB} = -0.179$ g/dl, $p_{HGB} = 0.006$) and rs8177988 (p.Val506Ile [c.1516G>A], MAF = 0.6%, $beta_{HGB} = +0.116$ g/dl, $p_{HGB} = 0.003$). It is noteworthy that the p.Val506Ile substitution is associated with increased HGB concentration given that this amino acid maps to a PKLR structural domain necessary for protein interaction.[50] This heterogeneity of effect among the *PKLR* missense variants might explain why SKAT's result is more significant than VT's for this gene (Table 2). A third gene, *ALPK3*, was identified only in the VT analysis for association with MCHC (Table 2). *ALPK3* encodes a kinase previously implicated in cardiomyocyte differentiation.[51] We could not test for replication because of the rarity of *ALPK3*'s coding alleles (Table S7).

RBC Variants and Pleiotropic Effects

Besides the overlap within the RBC traits themselves, we identified seven novel RBC variants associated with other blood-cell type traits or with lipid levels (Figure 3 and Table 3). To assess whether the genetic associations with RBC traits are independent of lipid levels, we performed additional analyses in a subset of BCX participants from three of our studies (FHS, MHIBB, and WHI) ranging from ~10,000 to 23,000 individuals. We repeated the association analyses for five RBC loci (*TMEM57-RHD* rs10903129, *AFF1* rs442177, *TRIB1* rs2954029, *MAP1A* rs55707100, and *HNF4A* rs1800961) additionally adjusting for the respective lipid trait and combined the results across the three studies using fixed-effect meta-analysis (Table S8). There was little or no change in the effect size or p values associated with the five RBC trait loci upon adjustment for the corresponding lipid trait, suggesting that the RBC and lipid associations are independent of one another and thus represent true "pleiotropic" genetic effects.

A correlated response to or role in inflammation might explain why some of the RBC variants are also associated with WBC, PLT, or lipid traits. Another plausible explanation for the concomitant association of several markers with RBC, WBC, and PLT phenotypes could be a more general effect of these genes on the proliferation or differentiation of hematopoietic progenitor cells. This is most likely the case for *JAK2* (MIM: 147796) and *SH2B3* (MIM: 605093), two key regulators of hematopoietic cells (Figure 3). In this category, we also observed two novel findings, *AFF1* (MIM: 159557) and *NOL4L*, which are associated with RBC and WBC phenotypes and have been previously implicated in leukemia.[53,54] Finally, we identified a novel missense variant in *ARHGEF3* (MIM: 612115) associated with HGB and HCT. In addition to its association with PLT traits, *ARHGEF3* plays a role in the regulation of iron uptake and erythroid cell maturation.[55]

Discussion

We present multi-ethnic meta-analyses of seven RBC traits using ExomeChip results of 130,273 individuals. Our statistical thresholds to declare significance at the discovery stage ($p < 2 \times 10^{-7}$ in the single-variant analyses) was adjusted for the approximate number of variants genotyped on the ExomeChip (Bonferroni correction for 250,000 variants), but we decided not to adjust it for the seven RBC phenotypes tested because of the high correlation between some of these traits (Figure S3). Instead, we relied on independent replication to distinguish true from probably false positive associations. Despite the limited size of our replication set (27,480 individuals), it was encouraging to detect a strong replication of direction of effect for known and novel RBC variants, suggesting a low false discovery rate. In total, we identified 23 novel

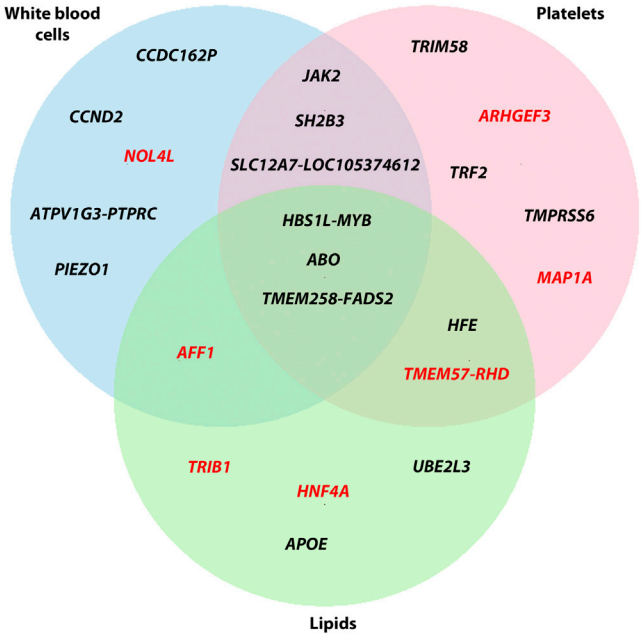


Figure 3. Venn Diagram Summarizing Pleiotropic Effects for Genetic Variants Associated with Red Blood Cell Traits
We considered variants only if their association p values with white blood cell (WBC) traits, platelet (PLT) traits, or with lipid levels was $p < 1 \times 10^{-4}$. Results for WBC and PLT are from the accompanying Blood-Cell Consortium articles.[20,21] Results for lipids have previously been published (Table 3). Genes highlighted in red are novel RBC trait findings.

variants associated with RBC traits in the single-variant analyses and a collection of three rare missense variants in *PKLR* associated with HGB and HCT in the gene-based analyses. Out of the 23 novel RBC variants, 16 were replicated at $p < 0.05$ in the independent samples (Table 1). To inform our replication criteria, we conducted a power analysis using a sample size of 20,000 and considering multiple combinations of allele frequencies and effect sizes. Based on allele frequency and effect size, one of our most difficult to replicate variants was rs1800961 (MAF = 0.022, Beta = 0.028). However, we still had approximately 56% power to detect this association in the replication stage.

We identified a nonsense variant in *CD36* associated with RDW in African Americans. CD36 is a type B scavenger receptor located on the surface of many cell types, including endothelial cells, platelets, monocytes, and erythrocytes. CD36 is a marker of erythroid progenitor differentiation[56] and might also be involved in macrophage-mediated clearance of red blood cells.[57] Furthermore, CD36 plays a role in many biological pathways such as lipid metabolism/transport and atherosclerosis, hemostasis, and inflammation.[58] The nonsense *CD36* variant identified in our RDW meta-analysis (rs3211938, Table 1) has previously been associated with platelet count, HDL cholesterol, and C-reactive protein levels in African Americans[59,60] and malaria resistance in Africans.[61,62] The *CD36* locus shows a signature of natural selection in AA populations[63] and the MAF of rs3211938 varies widely between continents: in the 1000 Genomes Project, the minor allele is absent from European populations but reaches frequency of 24%–29% in some African populations.[64] To characterize the molecular mechanism by which rs3211938 can impact RDW, we identified one heterozygous sample among a collection of ex vivo differentiated human erythroblasts.[32] In erythroblasts from this donor, we noted a strong allelic imbalance (Figure 2). Importantly, this result was confirmed in independent samples from the GTex dataset (Figure S5). At the molecular level, this *CD36* expression phenotype could be explained by nonsense-mediated mRNA decay or the regulatory effect of non-coding genetic variants in LD with rs3211938.

We observed that many new RBC variants are pleiotropic, being often associated with more than one RBC index as well as with WBC, PLT, and lipid traits (Figure 3). These shared effects could imply that the underlying causal genes at these RBC loci generally controlled blood cell proliferation or modulate inflammatory responses. An additional explanation for the link between RBC traits and lipid variants might be the cholesterol content of RBC membranes. As mentioned earlier, RBC corresponds to a large fraction (~25%) of the cells found in the human body. Genetic variation that modulates RBC count or volume could impact circulating lipid levels. In support of this hypothesis, it has been observed that a thalassemia allele is strongly associated with cholesterol levels in the Sardinian population.[65] In total, we found ten loci associated with lipid levels and RBC indices, including four novel RBC variants (*AFF1*, *TMEM57-RHD*, *TRIB1*, *HNF4A*) (Figure 3).

In summary, our multi-ethnic meta-analyses have expanded the genetic knowledge of erythrocyte biology and identified new common, low-frequency, and rare RBC variants. Many of the new RBC variants are pleiotropic, affecting other complex traits such as WBC, PLT, and blood lipid levels. Although our report demonstrates the utility of the ExomeChip for genetic discovery, it also highlights the challenge to attribute gene causality based only on association results. This is particularly evident for loci with common variants, for which coding and non-coding markers are often statistically equivalent. For instance, we found no examples of RBC coding variants that entirely explain RBC GWAS signals among the seven loci that had both a sentinel GWAS variant and ExomeChip coding markers. Although increasing sample sizes will continue to yield additional RBC loci, it has become incredibly clear that only a combination of well-powered genetic studies, transcriptomic and epigenomic surveys, and functional experiments (e.g., using genome editing) will ultimately pinpoint causal variants and genes that control RBC phenotypes.

Supplemental Data

Supplemental Data include a note on the eQTL analyses, information on supplementary funding, five figures, and eight tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.05.007.

Table 3. Overlap of Red Blood Cell Markers with Other Blood Cell Traits and/or Lipid

SNP	Position	A1/A2	AF (A2)	Annotation	Gene	Trait	Beta	p Value
rs10903129	1: 25,768,937	A/G	0.568	intron	*TMEM57-RHD*	RDW	0.037	1.19×10^{-7}
						TC[46]	0.061	5.40×10^{-10}
						PLT	−0.021	7.06×10^{-6}
rs3772219	3: 56,771,251	A/C	0.338	missense	*ARHGEF3*	HCT*	−0.028	2.38×10^{-9}
						HGB*	−0.026	3.76×10^{-9}
						PLT	0.031	5.93×10^{-10}
rs442177	4: 88,030,261	G/T	0.595	intron	*AFF1*	HGB	−0.034	3.97×10^{-13}
						TG[40]	−0.031	1.00×10^{-18}
						BASO	−0.030	1.99×10^{-5}
rs2954029	8: 126,490,972	A/T	0.439	intergenic	*TRIB1*	RDW	0.036	1.53×10^{-7}
						TG[40]	−0.076	1.00×10^{-7}
rs55707100	15: 43,820,717	C/T	0.033	missense	*MAP1A*	HGB	−0.071	1.65×10^{-8}
						PLT	0.095	7.03×10^{-14}
						TG[52]	0.090	1.40×10^{-17}
rs4911241	20: 31,140,165	C/T	0.241	intron	*NOL4L*	MCV	−0.040	1.25×10^{-8}
						RDW	0.043	5.79×10^{-8}
						BASO	−0.051	1.35×10^{-10}
						MONO	−0.033	3.57×10^{-5}
rs1800961	20: 43,042,364	C/T	0.032	missense	*HNF4A*	HCT	0.083	1.44×10^{-8}
						HGB	0.073	2.53×10^{-8}
						HDL[40]	−0.127	2.00×10^{-34}

Shown here are significant novel variants from the RBC traits association analyses that overlap with other blood-cell traits or with lipids. Results for the white blood cell and platelet traits are from the Blood Cell Consortium, and results for lipids are from the published literature. Results are presented for European-ancestry individuals, except in the presence of an asterisk (*), which stands for result from "All" ancestry. The allele frequency and direction of the effect (beta) is given for the A2 allele. Abbreviations are as follows: A1, reference allele; A2, alternate allele; AF, allele frequency; HCT, hematocrit; HGB, hemoglobin; MCV, mean corpuscular volume; RDW, red blood cell distribution width; TC, total cholesterol; PLT, platelet; TG, triglycerides; WBC, white blood cells; BASO, basophils; MONO, monocytes; HDL, HDL cholesterol.

Acknowledgments

We thank all participants, staff, and study coordinating centers. We also thank Raymond Doty and Jan Abkowitz for discussion of the *ALAS2* finding. We would like to thank Liling Warren for contributions to the genetic analysis of the SOLID-TIMI-52 and STABILITY datasets. Young Finns Study (YFS) acknowledges the expert technical assistance in the statistical analyses by Ville Aalto and Irina Lisinen. Estonian Genome Center, University of Tartu (EGCUT) thanks co-workers at the Estonian Biobank, especially Mr. V. Soo, Mr. S. Smith, and Dr. L. Milani. Airwave thanks Louisa Cavaliero who assisted in data collection and management as well as Peter McFarlane and the Glasgow CARE, Patricia Munroe at Queen Mary University of London, and Joanna Sarnecka and Ania Zawodniak at Northwick Park for their contributions to the study. This work was supported by the Fonds de Recherche du Queébec-Santeé (FRQS, scholarship to N.C.), the Canadian Institute of Health Research (Banting-CIHR, scholarship to S.L. and operating grant MOP#123382 to G.L.), and the Canada Research Chair program (to G.L.). P.L.A. was supported by NHLBI R21 HL121422-02. N.A.A. is funded by NIH DK060022. A.N. was supported by the Yoshida Scholarship Foundation. S.K. was supported by a Research Scholar award from the Massachusetts General Hospital (MGH), the Howard Goodman Fellowship from MGH, the Donovan Family Foundation, R01HL107816, and a grant from Fondation Leducq. Additional acknowledgments and funding information is provided in the Supplemental Data.

Received: February 18, 2016
Accepted: May 3, 2016
Published: June 23, 2016

Web Resources

BCX ExomeChip association results, http://www.mhi-humangenetics.org/en/resources
CheckVCF, https://github.com/zhanxw/checkVCF
ClinVar, https://www.ncbi.nlm.nih.gov/clinvar/
OMIM, http://www.omim.org/
RareMETALS, http://genome.sph.umich.edu/wiki/RareMETALS
RareMetalWorker, http://genome.sph.umich.edu/wiki/RAREMETALWORKER
RvTests, http://genome.sph.umich.edu/wiki/RvTests
Type 2 Diabetes Genetics, http://www.type2diabetesgenetics.org/

References

1. Ulker, P., Sati, L., Celik-Ozenci, C., Meiselman, H.J., and Baskurt, O.K. (2009). Mechanical stimulation of nitric oxide synthesizing mechanisms in erythrocytes. Biorheology *46*, 121–132.
2. Jiang, N., Tan, N.S., Ho, B., and Ding, J.L. (2007). Respiratory protein-generated reactive oxygen species as an antimicrobial strategy. Nat. Immunol. *8*, 1114–1122.
3. Schnabel, R.B., Baumert, J., Barbalic, M., Dupuis, J., Ellinor, P.T., Durda, P., Dehghan, A., Bis, J.C., Illig, T., Morrison, A.C., et al. (2010). Duffy antigen receptor for chemokines (Darc) polymorphism regulates circulating concentrations of monocyte chemoattractant protein-1 and other inflammatory mediators. Blood *115*, 5289–5299.
4. Colin, Y., Le Van Kim, C., and El Nemer, W. (2014). Red cell adhesion in human diseases. Curr. Opin. Hematol. *21*, 186–192.
5. Whelihan, M.F., and Mann, K.G. (2013). The role of the red cell membrane in thrombin generation. Thromb. Res. *131*, 377–382.
6. Brugnara, C. (2003). Iron deficiency and erythropoiesis: new diagnostic approaches. Clin. Chem. *49*, 1573–1578.
7. Huang, Y.L., Hu, Z.D., Liu, S.J., Sun, Y., Qin, Q., Qin, B.D., Zhang, W.W., Zhang, J.R., Zhong, R.Q., and Deng, A.M. (2014). Prognostic value of red blood cell distribution width for patients with heart failure: a systematic review and meta-analysis of cohort studies. PLoS ONE *9*, e104861.
8. Nada, A.M. (2015). Red cell distribution width in type 2 diabetic patients. Diabetes Metab. Syndr. Obes. *8*, 525–533.
9. Zalawadiya, S.K., Zmily, H., Farah, J., Daifallah, S., Ali, O., and Ghali, J.K. (2011). Red cell distribution width and mortality in predominantly African-American population with decompensated heart failure. J. Card. Fail. *17*, 292–298.
10. Zalawadiya, S.K., Veeranna, V., Panaich, S.S., and Afonso, L. (2012). Red cell distribution width and risk of peripheral artery disease: analysis of National Health and Nutrition Examination Survey 1999-2004. Vasc. Med. *17*, 155–163.
11. Patel, K.V., Semba, R.D., Ferrucci, L., Newman, A.B., Fried, L.P., Wallace, R.B., Bandinelli, S., Phillips, C.S., Yu, B., Connelly, S., et al. (2010). Red cell distribution width and mortality in older adults: a meta-analysis. J. Gerontol. A Biol. Sci. Med. Sci. *65*, 258–265.
12. Patel, H.H., Patel, H.R., and Higgins, J.M. (2015). Modulation of red blood cell population dynamics is a fundamental homeostatic response to disease. Am. J. Hematol. *90*, 422–428.
13. Whitfield, J.B., and Martin, N.G. (1985). Genetic and environmental influences on the size and number of cells in the blood. Genet. Epidemiol. *2*, 133–144.
14. Pilia, G., Chen, W.M., Scuteri, A., Orrú, M., Albai, G., Dei, M., Lai, S., Usala, G., Lai, M., Loi, P., et al. (2006). Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS Genet. *2*, e132.
15. 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.
16. Lin, J.P., O'Donnell, C.J., Jin, L., Fox, C., Yang, Q., and Cupples, L.A. (2007). Evidence for linkage of red blood cell size and count: genome-wide scans in the Framingham Heart Study. Am. J. Hematol. *82*, 605–610.
17. 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.
18. 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.; BioBank Japan Project; CHARGE Consortium (2013). Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. Hum. Mol. Genet. *22*, 2529–2538.
19. Auer, P.L., Teumer, A., Schick, U., O'Shaughnessy, A., Lo, K.S., Chami, N., Carlson, C., de Denus, S., Dubé, M.P., Haessler, J., et al. (2014). Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. Nat. Genet. *46*, 629–634.
20. Eicher, J.D., Chami, N., Kacprowski, T., Nomura, A., Chen, M.-H., Yanek, L.R., Tajuddin, S.M., Schick, U.M., Slater, A.J., Pankratz, N., et al. (2016). Platelet-related variants identified by exomechip meta-analysis in 157,293 individuals. Am. J. Hum. Genet. *99*. Published online June 23, 2016. http://dx.doi.org/10.1016/j.ajhg.2016.05.005.
21. Tajuddin, S.M., Schick, U.M., Eicher, J.D., Chami, N., Giri, A., Brody, J.A., Hill, W.D., Kacprowski, T., Li, J., Lyytikäinen, L.-P., et al. (2016). Large-scale exome-wide association analysis identifies loci for white blood cell traits and pleiotropy with immune-mediated diseases. Am. J. Hum. Genet. *99*. Published online June 23, 2016. http://dx.doi.org/10.1016/j.ajhg.2016.05.003.
22. Grove, M.L., Yu, B., Cochran, B.J., Haritunians, T., Bis, J.C., Taylor, K.D., Hansen, M., Borecki, I.B., Cupples, L.A., Fornage, M., et al. (2013). Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. PLoS ONE *8*, e68095.
23. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
24. Wells, Q.S., Becker, J.R., Su, Y.R., Mosley, J.D., Weeke, P., D'Aoust, L., Ausborn, N.L., Ramirez, A.H., Pfotenhauer, J.P., Naftilan, A.J., et al. (2013). Whole exome sequencing identifies a causal RBM20 mutation in a large pedigree with familial dilated cardiomyopathy. Circ Cardiovasc Genet *6*, 317–326.
25. Winkler, T.W., Day, F.R., Croteau-Chonka, D.C., Wood, A.R., Locke, A.E., Mägi, R., Ferreira, T., Fall, T., Graff, M., Justice, A.E., et al.; Genetic Investigation of Anthropometric Traits (GIANT) Consortium (2014). Quality control and conduct of genome-wide association meta-analyses. Nat. Protoc. *9*, 1192–1212.
26. Limongelli, G., Elliott, P., Charron, P., Mogensen, J., and McKeown, P.P. (2012). Approaching genetic testing in cardiomyopathies (ESC Council for Cardiology Practice).
27. Olson, T.M., Michels, V.V., Thibodeau, S.N., Tai, Y.S., and Keating, M.T. (1998). Actin mutations in dilated cardiomyopathy, a heritable form of heart failure. Science *280*, 750–752.
28. Liu, D.J., Peloso, G.M., Zhan, X., Holmen, O.L., Zawistowski, M., Feng, S., Nikpay, M., Auer, P.L., Goel, A., Zhang, H., et al. (2014). Meta-analysis of gene-level tests for rare variant association. Nat. Genet. *46*, 200–204.
29. 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.
30. Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. (2010). Pooled association tests

for rare variants in exon-resequencing studies. Am. J. Hum. Genet. *86*, 832–838.
31. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191.
32. Lessard, S., Beaudoin, M., Benkirane, K., and Lettre, G. (2015). Comparison of DNA methylation profiles in human fetal and adult red blood cell progenitors. Genome Med. *7*, 1.
33. Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al.; Geuvadis Consortium (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.
34. Zhang, X., Gierman, H.J., Levy, D., Plump, A., Dobrin, R., Goring, H.H., Curran, J.E., Johnson, M.P., Blangero, J., Kim, S.K., et al. (2014). Synthesis of 53 tissue and cell line expression QTL datasets reveals master eQTLs. BMC Genomics *15*, 532.
35. Astner, I., Schulze, J.O., van den Heuvel, J., Jahn, D., Schubert, W.D., and Heinz, D.W. (2005). Crystal structure of 5-aminolevulinate synthase, the first enzyme of heme biosynthesis, and its link to XLSA in humans. EMBO J. *24*, 3166–3177.
36. Halpain, S., and Dehmelt, L. (2006). The MAP1 family of microtubule-associated proteins. Genome Biol. *7*, 224.
37. Takei, Y., Kikkawa, Y.S., Atapour, N., Hensch, T.K., and Hirokawa, N. (2015). Defects in synaptic plasticity, reduced NMDA-receptor transport, and instability of postsynaptic density proteins in mice lacking microtubule-associated protein 1A. J. Neurosci. *35*, 15539–15554.
38. GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.
39. Dehghan, A., Dupuis, J., Barbalic, M., Bis, J.C., Eiriksdottir, G., Lu, C., Pellikka, N., Wallaschofski, H., Kettunen, J., Henneman, P., et al. (2011). Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels. Circulation *123*, 731–738.
40. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al.; Global Lipids Genetics Consortium (2013). Discovery and refinement of loci associated with lipid levels. Nat. Genet. *45*, 1274–1283.
41. Taylor, K.C., Lange, L.A., Zabaneh, D., Lange, E., Keating, B.J., Tang, W., Smith, N.L., Delaney, J.A., Kumari, M., Hingorani, A., et al. (2011). A gene-centric association scan for Coagulation Factor VII levels in European and African Americans: the Candidate Gene Association Resource (CARe) Consortium. Hum. Mol. Genet. *20*, 3525–3534.
42. de Vries, P.S., Chasman, D.I., Sabater-Lleal, M., Chen, M.H., Huffman, J.E., Steri, M., Tang, W., Teumer, A., Marioni, R.E., Grossmann, V., et al. (2016). A meta-analysis of 120 246 individuals identifies 18 new loci for fibrinogen concentration. Hum. Mol. Genet. *25*, 358–370.
43. Kooner, J.S., Saleheen, D., Sim, X., Sehmi, J., Zhang, W., Frossard, P., Been, L.F., Chia, K.S., Dimas, A.S., Hassanali, N., et al.; DIAGRAM; MuTHER (2011). Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. Nat. Genet. *43*, 984–989.
44. GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.
45. Pradel, L.C., Vanhille, L., and Spicuglia, S. (2015). [The European Blueprint project: towards a full epigenome characterization of the immune system]. Med. Sci. (Paris) *31*, 236–238.
46. Aulchenko, Y.S., Ripatti, S., Lindqvist, I., Boomsma, D., Heid, I.M., Pramstaller, P.P., Penninx, B.W., Janssens, A.C., Wilson, J.F., Spector, T., et al.; ENGAGE Consortium (2009). Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat. Genet. *41*, 47–55.
47. Kullo, I.J., Ding, K., Shameer, K., McCarty, C.A., Jarvik, G.P., Denny, J.C., Ritchie, M.D., Ye, Z., Crosslin, D.R., Chisholm, R.L., et al. (2011). Complement receptor 1 gene variants are associated with erythrocyte sedimentation rate. Am. J. Hum. Genet. *89*, 131–138.
48. Döbbeling, U. (1997). The effects of cyclosporin A on V(D)J recombination activity. Scand. J. Immunol. *45*, 494–498.
49. Zarza, R., Alvarez, R., Pujades, A., Nomdedeu, B., Carrera, A., Estella, J., Remacha, A., Sánchez, J.M., Morey, M., Cortes, T., et al.; Red Cell Pathology Group of the Spanish Society of Haematology (AEHH) (1998). Molecular characterization of the PK-LR gene in pyruvate kinase deficient Spanish patients. Br. J. Haematol. *103*, 377–382.
50. Valentini, G., Chiarelli, L.R., Fortin, R., Dolzan, M., Galizzi, A., Abraham, D.J., Wang, C., Bianchi, P., Zanella, A., and Mattevi, A. (2002). Structure and function of human erythrocyte pyruvate kinase. Molecular basis of nonspherocytic hemolytic anemia. J. Biol. Chem. *277*, 23807–23814.
51. Van Sligtenhorst, I., Ding, Z.M., Shi, Z.Z., Read, R.W., Hansen, G., and Vogel, P. (2012). Cardiomyopathy in α-kinase 3 (ALPK3)-deficient mice. Vet. Pathol. *49*, 131–141.
52. Peloso, G.M., Auer, P.L., Bis, J.C., Voorman, A., Morrison, A.C., Stitziel, N.O., Brody, J.A., Khetarpal, S.A., Crosby, J.R., Fornage, M., et al.; NHLBI GO Exome Sequencing Project (2014). Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. Am. J. Hum. Genet. *94*, 223–232.
53. Gu, Y., Nakamura, T., Alder, H., Prasad, R., Canaani, O., Cimino, G., Croce, C.M., and Canaani, E. (1992). The t(4;11) chromosome translocation of human acute leukemias fuses the ALL-1 gene, related to *Drosophila* trithorax, to the AF-4 gene. Cell *71*, 701–708.
54. Guastadisegni, M.C., Lonoce, A., Impera, L., Di Terlizzi, F., Fugazza, G., Aliano, S., Grasso, R., Cluzeau, T., Raynaud, S., Rocchi, M., and Storlazzi, C.T. (2010). CBFA2T2 and C20orf112: two novel fusion partners of RUNX1 in acute myeloid leukemia. Leukemia *24*, 1516–1519.
55. Serbanovic-Canic, J., Cvejic, A., Soranzo, N., Stemple, D.L., Ouwehand, W.H., and Freson, K. (2011). Silencing of RhoA nucleotide exchange factor, ARHGEF3, reveals its unexpected role in iron uptake. Blood *118*, 4967–4976.
56. Okumura, N., Tsuji, K., and Nakahata, T. (1992). Changes in cell surface antigen expressions during proliferation and differentiation of human erythroid progenitors. Blood *80*, 642–650.
57. Kiefer, C.R., and Snyder, L.M. (2000). Oxidation and erythrocyte senescence. Curr. Opin. Hematol. *7*, 113–116.
58. Nicholson, A.C., Han, J., Febbraio, M., Silversterin, R.L., and Hajjar, D.P. (2001). Role of CD36, the macrophage class B scavenger receptor, in atherosclerosis. Ann. N Y Acad. Sci. *947*, 224–228.
59. Auer, P.L., Johnsen, J.M., Johnson, A.D., Logsdon, B.A., Lange, L.A., Nalls, M.A., Zhang, G., Franceschini, N., Fox, K., Lange, E.M., et al. (2012). Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated

loci in African Americans: NHLBI GO Exome Sequencing Project. Am. J. Hum. Genet. *91*, 794–808.

60. Elbers, C.C., Guo, Y., Tragante, V., van Iperen, E.P., Lanktree, M.B., Castillo, B.A., Chen, F., Yanek, L.R., Wojczynski, M.K., Li, Y.R., et al. (2012). Gene-centric meta-analysis of lipid traits in African, East Asian and Hispanic populations. PLoS ONE *7*, e50198.
61. Ayodo, G., Price, A.L., Keinan, A., Ajwang, A., Otieno, M.F., Orago, A.S., Patterson, N., and Reich, D. (2007). Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. Am. J. Hum. Genet. *81*, 234–242.
62. Aitman, T.J., Cooper, L.D., Norsworthy, P.J., Wahid, F.N., Gray, J.K., Curtis, B.R., McKeigue, P.M., Kwiatkowski, D., Greenwood, B.M., Snow, R.W., et al. (2000). Malaria susceptibility and CD36 mutation. Nature *405*, 1015–1016.
63. Bhatia, G., Patterson, N., Pasaniuc, B., Zaitlen, N., Genovese, G., Pollack, S., Mallick, S., Myers, S., Tandon, A., Spencer, C., et al. (2011). Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. Am. J. Hum. Genet. *89*, 368–381.
64. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R.; 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
65. Sidore, C., Busonero, F., Maschio, A., Porcu, E., Naitza, S., Zoledziewska, M., Mulas, A., Pistis, G., Steri, M., Danjou, F., et al. (2015). Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nat. Genet. *47*, 1272–1281.

The American Journal of Human Genetics, Volume *99*

Supplemental Data

Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits

Nathalie Chami, Ming-Huei Chen, Andrew J. Slater, John D. Eicher, Evangelos Evangelou, Salman M. Tajuddin, Latisha Love-Gregory, Tim Kacprowski, Ursula M. Schick, Akihiro Nomura, Ayush Giri, Samuel Lessard, Jennifer A. Brody, Claudia Schurmann, Nathan Pankratz, Lisa R. Yanek, Ani Manichaikul, Raha Pazoki, Evelin Mihailov, W. David Hill, Laura M. Raffield, Amber Burt, Traci M. Bartz, Diane M. Becker, Lewis C. Becker, Eric Boerwinkle, Jette Bork-Jensen, Erwin P. Bottinger, Michelle L. O'Donoghue, David R. Crosslin, Simon de Denus, Marie-Pierre Dubé, Paul Elliott, Gunnar Engström, Michele K. Evans, James S. Floyd, Myriam Fornage, He Gao, Andreas Greinacher, Vilmundur Gudnason, Torben Hansen, Tamara B. Harris, Caroline Hayward, Jussi Hernesniemi, Heather M. Highland, Joel N. Hirschhorn, Albert Hofman, Marguerite R. Irvin, Mika Kähönen, Ethan Lange, Lenore J. Launer, Terho Lehtimäki, Jin Li, David C.M. Liewald, Allan Linneberg, Yongmei Liu, Yingchang Lu, Leo-Pekka Lyytikäinen, Reedik Mägi, Rasika A. Mathias, Olle Melander, Andres Metspalu, Nina Mononen, Mike A. Nalls, Deborah A. Nickerson, Kjell Nikus, Chris J. O'Donnell, Marju Orho-Melander, Oluf Pedersen, Astrid Petersmann, Linda Polfus, Bruce M. Psaty, Olli T. Raitakari, Emma Raitoharju, Melissa Richard, Kenneth M. Rice, Fernando Rivadeneira, Jerome I. Rotter, Frank Schmidt, Albert Vernon Smith, John M. Starr, Kent D. Taylor, Alexander Teumer, Betina H. Thuesen, Eric S. Torstenson, Russell P. Tracy, Ioanna Tzoulaki, Neil A. Zakai, Caterina Vacchi-Suzzi, Cornelia M. van Duijn, Frank J.A. van Rooij, Mary Cushman, Ian J. Deary, Digna R. Velez Edwards, Anne-Claire Vergnaud, Lars Wallentin, Dawn M. Waterworth, Harvey D. White, James G. Wilson, Alan B. Zonderman, Sekar Kathiresan, Niels Grarup, Tõnu Esko, Ruth J.F. Loos, Leslie A. Lange, Nauder Faraday, Nada A. Abumrad, Todd L. Edwards, Santhi K. Ganesh, Paul L. Auer, Andrew D. Johnson, Alexander P. Reiner, and Guillaume Lettre

Supplemental Note

Datasets used in the expression quantitative trait loci (eQTL) analyses

We queried RBC/WBC/PLT loci in over 100 separate eQTL datasets in a wide range of tissues. Datasets were collected through publications, publically available sources, or private collaboration. A general overview of a subset of >50 eQTL studies has been published (PMID: 24973796), with specific citations for >100 datasets included in the current query following here.

Blood cell related eQTL studies included fresh lymphocytes (17873875), fresh leukocytes (19966804), leukocyte samples in individuals with Celiac disease (19128478), whole blood samples (18344981, 21829388, 22692066, 23818875, 23359819, 23880221, 24013639, 23157493, 23715323, 24092820, 24314549, 24956270, 24592274, 24728292, 24740359, 25609184, 22563384, 25474530, 25816334, 25578447), lymphoblastoid cell lines (LCL) derived from asthmatic children (17873877, 23345460), HapMap LCL from 3 populations (17873874), a separate study on HapMap CEU LCL (18193047), additional LCL population samples (19644074, 22286170, 22941192, 23755361, 23995691, 25010687, 25951796), neutrophils (26151758, 26259071), CD19+ B cells (22446964), primary PHA-stimulated T cells (19644074, 23755361), CD4+ T cells (20833654), peripheral blood monocytes (19222302,20502693,22446964, 23300628, 25951796, 26019233), long non-coding RNAs in monocytes (25025429) and CD14+ monocytes before and after stimulation with LPS or interferon-gamma (24604202), CD11+ dendritic cells before and after *Mycobacterium tuberculosis* infection (22233810) and a separate study of dendritic cells before or after stimulation with LPS, influenza or interferon-beta (24604203). Micro-RNA QTLs (21691150, 26020509), DNase-I QTLs (22307276), histone acetylation QTLs (25799442), and ribosomal occupancy QTLs (25657249) were also queried for LCL. Splicing QTLs (25685889) and micro-RNA QTLs (25791433) were queried in whole blood.

Non-blood cell tissue eQTLs searched included omental and subcutaneous adipose (18344981, 21602305, 22941192, 23715323, 25578447), visceral fat (25578447) stomach (21602305), endometrial carcinomas (21226949), ER+ and ER- breast cancer tumor cells (23374354), liver (18462017,21602305,21637794, 22006096, 24665059, 25578447), osteoblasts (19654370), intestine (23474282) and normal and cancerous colon (25079323, 25766683), skeletal muscle (24306210, 25578447), breast tissue (normal and cancer)(24388359, 22522925), lung (23209423, 23715323, 24307700, 23936167, 26102239), skin (21129726, 22941192, 23715323, 25951796), primary fibroblasts (19644074, 23755361, 24555846), sputum (21949713), pancreatic islet cells (25201977), prostate (25983244), rectal mucosa (25569741), arterial wall (25578447) and heart tissue from left ventricles (23715323, 24846176) and left and right atria (24177373). Micro-RNA QTLs were also queried for gluteal and abdominal adipose (22102887) and liver (23758991). Methylation QTLs were queried in pancreatic islet cells (25375650). Further mRNA and micro-RNA QTLs were queried from ER+ invasive breast cancer samples, colon-, kidney renal clear-, lung- and prostate-adenocarcinoma samples (24907074).

Brain eQTL studies included brain cortex (19222302, 19361613, 22685416, 25609184, 25290266), cerebellar cortex (25174004), cerebellum (20485568, 22685416, 22212596, 22832957, 23622250), frontal cortex (20485568, 22832957, 25174004), gliomas (24607568), hippocampus (22832957, 25174004), inferior olivary nucleus (from medulla) (25174004), intralobular white matter (25174004), occiptal cortex (25174004), parietal lobe (22212596), pons (20485568), pre-frontal cortex (22031444, 20351726, 22832957, 23622250), putamen (at the level of anterior commussure) (25174004), substantia nigra (25174004), temporal cortex (20485568, 22685416, 22832957, 25174004), thalamus (22832957) and visual cortex (23622250).

Additional eQTL data was integrated from online sources including ScanDB, the Broad Institute GTEx Portal, and the Pritchard Lab (eqtl.uchicago.edu). Cerebellum, parietal lobe and liver eQTL data was downloaded from ScanDB and cis-eQTLs were limited to those with P<1.0E-6 and trans-eQTLs with P<5.0E-8. Results for GTEx Analysis V4 for 13 tissues were downloaded from the GTEx Portal and then additionally filtered as described below [www.gtexportal.org: thyroid, leg skin (sun exposed), tibial nerve, aortic artery, tibial artery, skeletal muscle, esophagus mucosa, esophagus muscularis, lung, heart (left ventricle), stomach, whole blood, and subcutaneous adipose (23715323)]. Splicing QTL (sQTL) results generated with sQTLseeker with false discovery rate P≤0.05 were retained. For all gene-level eQTLs, if at least 1 SNP passed the tissue-specific empirical threshold in GTEx, the best SNP for that eQTL was always retained. All gene-level eQTL SNPs with P<1.67E-11 were also retained, reflecting a global threshold correction of P=0.05/(30,000 genes X 1,000,000 tests).

Figure S1. Flow chart of the study design. Data was contributed from 24 studies for the discovery phase. We applied QC steps to remove low quality variants and samples. We also excluded individuals with extreme phenotypes. After the proper adjustments, we applied inverse normal transformation on the residuals. We then performed study-specific association analyses using RV test or RareMetalWorker followed by QC of the individual association results. We meta-analyzed the association results using RareMetals and performed single variant (SV) and gene-based analyses. Additionally, we performed conditional analyses on the SV results, and attempted replication of the significant independent markers in the replication phase which comprised 8 independent studies. HW: Hardy Weinberg; PC: principle components; SKAT: Sequence Kernel Association Test; VT: Variable threshold test; EA; European ancestry; AA: African American ancestry; All: combined ancestry (EA + AA + Asians + Hispanics).

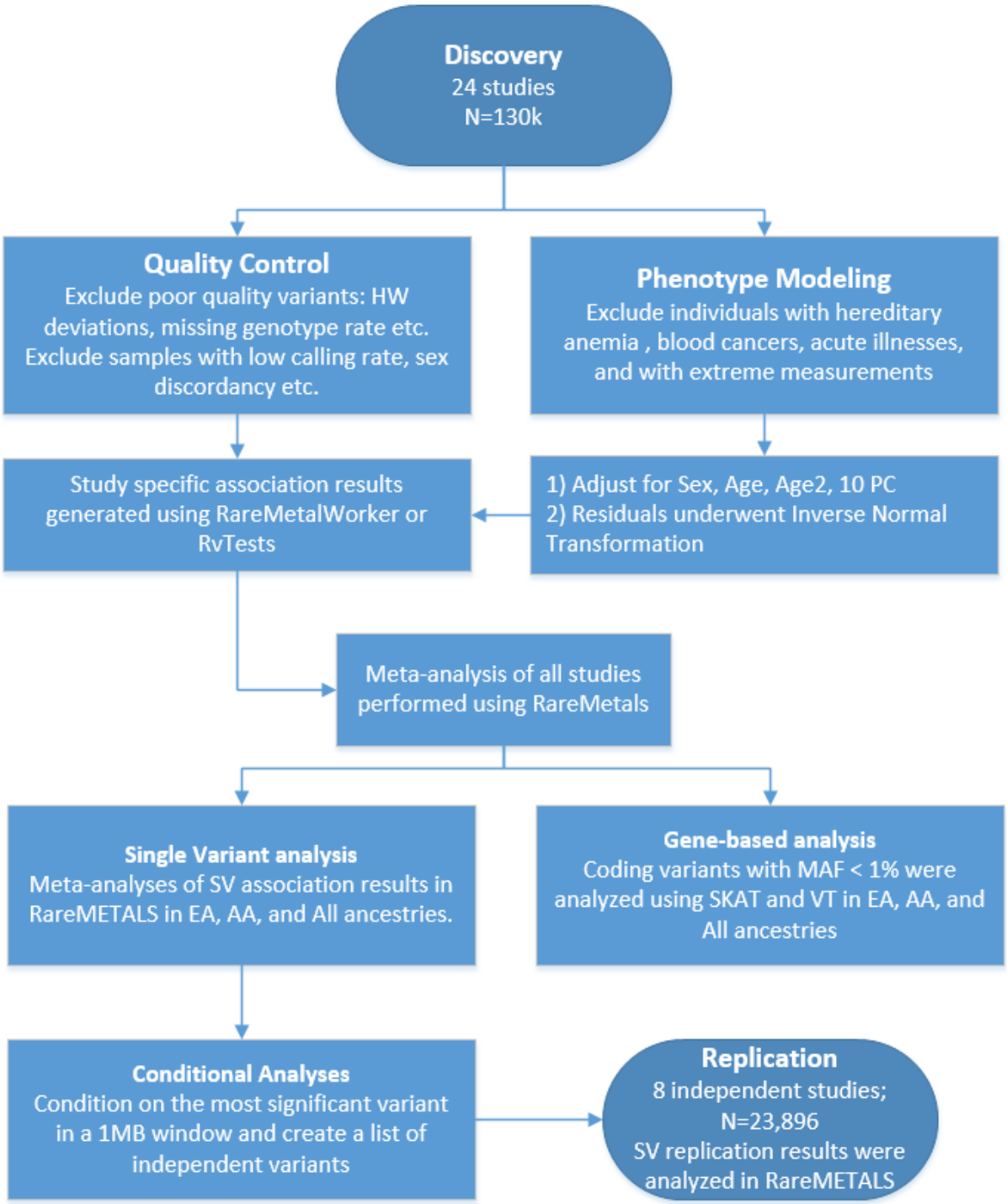


Figure S2. Manhattan plots of the all-ancestry single-variant meta-analyses results of the seven red blood cell traits of the discovery phase. Variants with $P<2x10^{-7}$ are shown in pink.

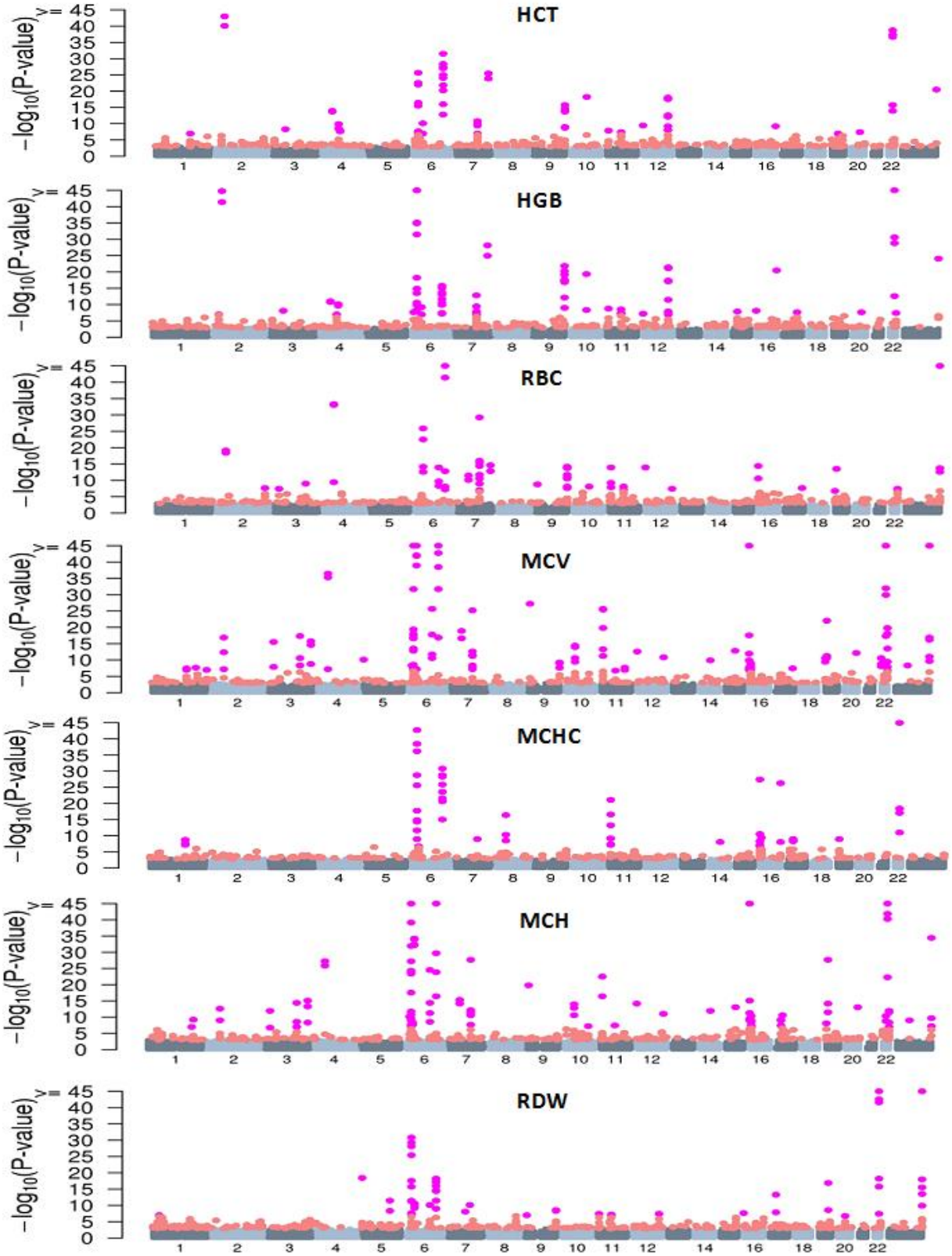


Figure S3. Scatter plots showing the correlation between the seven red blood cell traits. Shown here are raw values from the MHI Biobank cohort (N=7911). Strong correlations can be observed between HGB and HCT (r = 0.978) and between MCH and MCV (r = 0.934). RBC is highly correlated with HCT (r = 0.874) and HGB (r = 0.835).

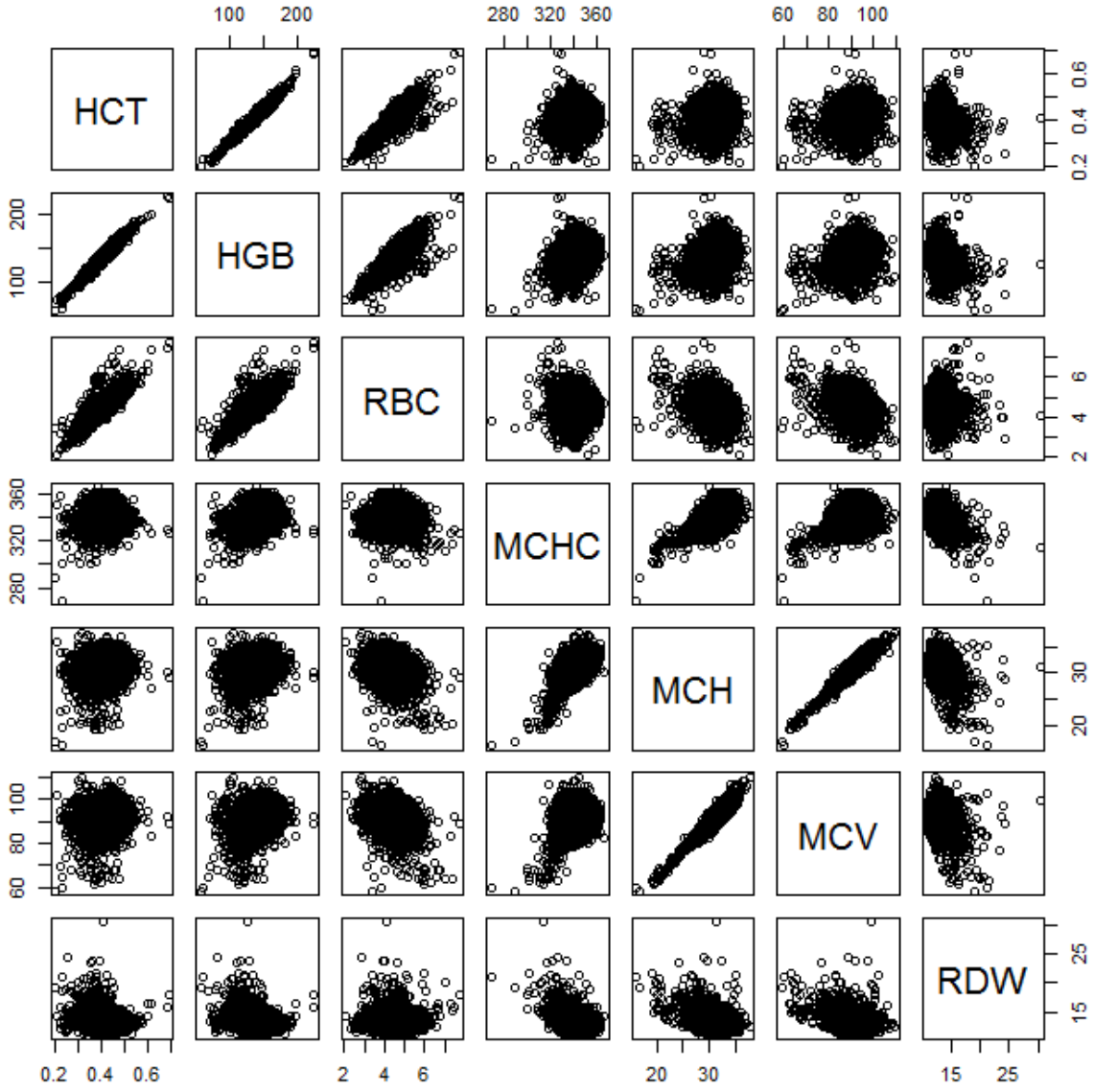


Figure S4. Expression levels of the *HNF4A* isoforms detected in hematopoietic cells by RNA-seq by the BLUEPRINT Project. *HNF4* is only detectable in erythroblasts (EB) and megakaryocytes (MK).

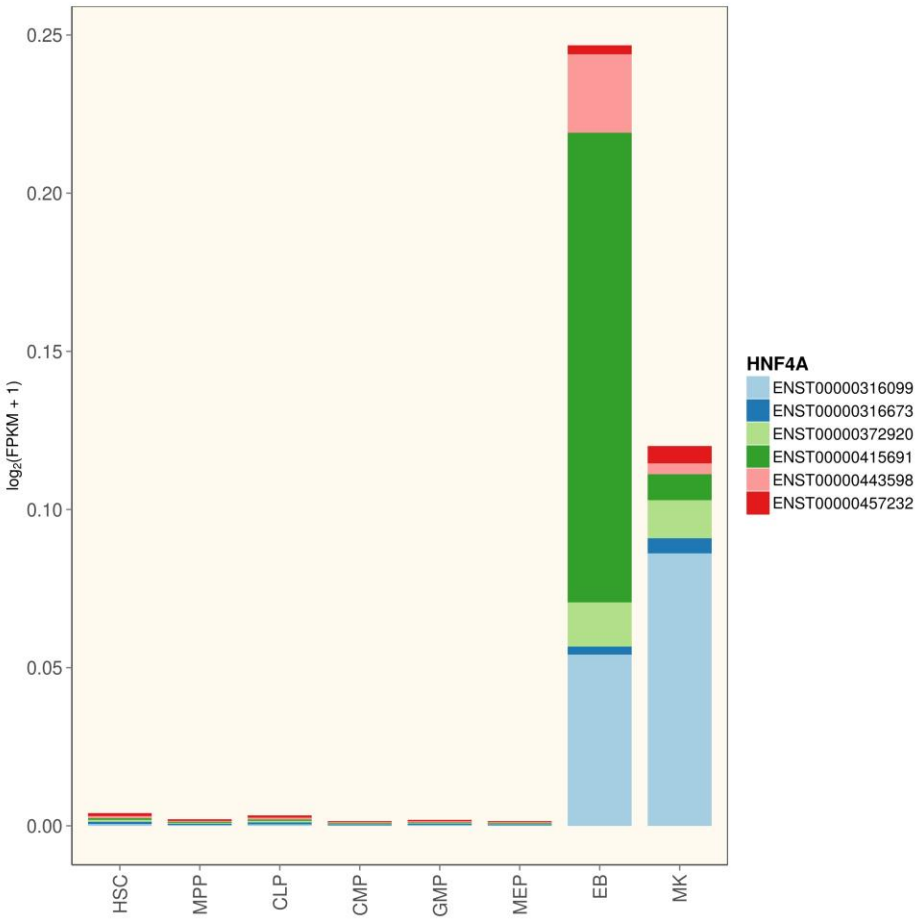
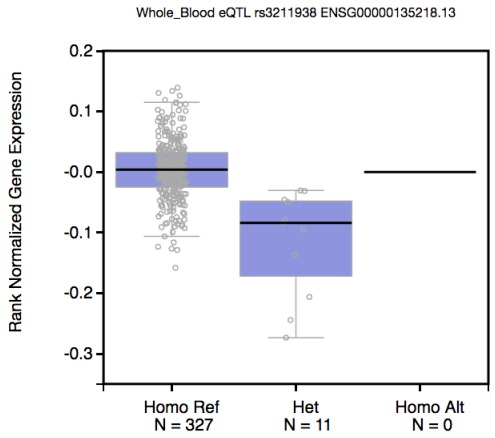


Figure S5. The nonsense variant rs3211938 in *CD36* is associated with *CD36* expression levels in the GTEx dataset. (**A**) rs3211938 is an expression quantitative trait locus (eQTL) for *CD36* in whole blood samples. The G-allele is associated with reduced expression levels (P_{eQTL}=1.1x10^{-15}) in 338 samples, including 11 heterozygous samples. (**B**) Allelic imbalance (also known as allele-specific expression or ASE) of *CD36*-rs3211938 in several available tissues. The upper panel shows the ratio of G-allele among all RNA-sequencing reads that cover the site. Consistent with our erythroblast results, several tissues show a ratio well under the expected 50:50, suggesting that the nonsense G-allele is under-represented. For instance, the ratio is 30:70 is whole blood (WHLBLD). The lower panel summarizes the statistical evidence of ASE in the same tissues based on the model developed by Rivas et *al*., Science, 2015. For instance in whole blood, the probability to observe moderate evidence of ASE is 94% whereas the probability that there is no ASE is 6%. ADPSBQ, adipose-subcutaneous; ARTAORT, artery-aorta; ARTCRN, artery-coronary; ARTTBL, artery-tibial; HRTLV, heart-left ventricle; MSCLSK, skeletal muscle; STMACH, stomach.

A



B

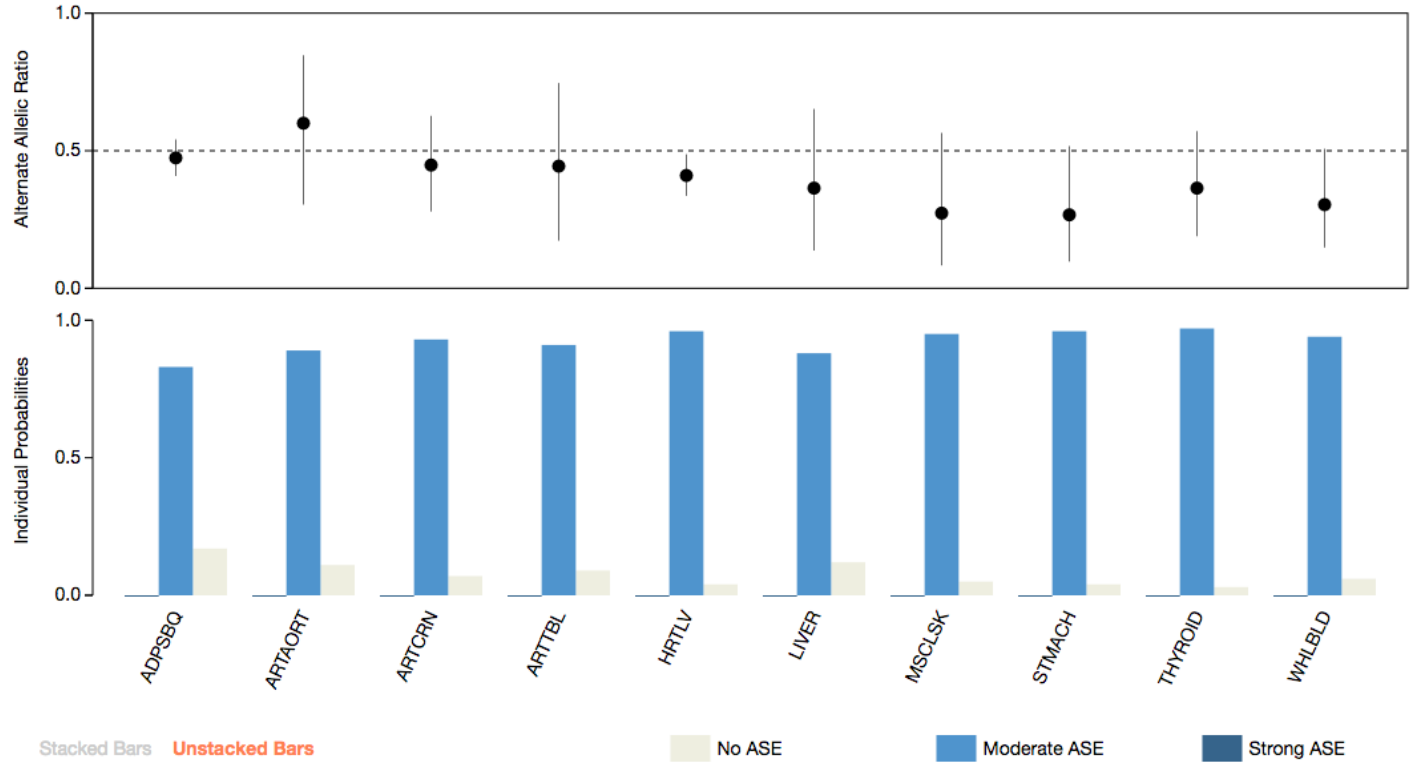


Table S4. Description of the replication cohorts.

Abbreviated Study Name	Full Study Name	Study Design	Ethnicity	Total Sample Size (N)
AIRWAVE	Airwave Health Monitoring Study	Cohort	European	14,866
GeneSTAR	Study of Atherosclerosis Risk in Families	Family-based	European	1,020
FINCAVAS	The Finnish Cardiovascular Study	Hospital-based	European	912
NWIGM	Northwest Institute of Genetic Medicine	Cohort and case-control	European	1,274
REGARDS	REasons for Geographic and Racial Differences in Stroke	Cohort	African American	5,039
YFS	The Cardiovascular Risk in Young Finns Study	Cohort	European	1,888
BioVU-Replication	Vanderbilt University DNA data bank	EMR Database at Vanderbilt University	European	1,783
BioVU	Vanderbilt University DNA data bank	EMR Database at Vanderbilt University	African American	968

Table S6. Expression quantitative trait loci (eQTL) results for variants associated with red blood cell phenotypes. eSNP, SNP associated with the gene expression phenotype; eSNP.p, eQTL association P-value; r^2, linkage disequilibrium in European populations between the eSNP (from the RBC analyses) and the best eSNP for a given gene; Best_eSNP, best reported eSNP for the gene tested; Best_eSNP.p, eQTL P-value for the best eSNP for the gene tested.

CHR	POS	eSNP	Gene	Trait	Tissue	eSNP.p	Transcript	r^2	Best_eSNP	Best_eSNP.p
1	25768937	rs10903129	*TMEM57*	RDW-EA/All	Whole blood	2.67E-128	*RHD*	0.669	rs909832	9.81E-198
1	155162067	rs4072037	*MUC1*	HCT-All	CD16+ neutrophils	2.30E-05	*THBS3*	1	rs2066981	2.30E-05
2	219509618	rs2230115	*ZNF142*	RBC-All	CD16+ neutrophils	7.26E-17	*CYP27A1*	1	rs10187066	7.26E-17
3	56771251	rs3772219	*ARHGEF3*	HCT/HGB-All	Whole blood	3.10E-21	*ARHGEF3*	0.682	rs2046823	1.16E-27
3	56771251	rs3772219	*ARHGEF3*	HCT/HGB-All	Peripheral blood mononuclear cells	4.55E-15	*ARHGEF3*	SameSNP	rs3772219	4.55E-15
4	88008782	rs236985	*AFF1*	HCT/RBC-EA	Peripheral blood mononuclear cells	4.42E-18	*AFF1*	0.932	rs442177	1.05E-18
4	88030261	rs442177	*AFF1*	HGB-EA	Peripheral blood mononuclear cells	1.05E-18	*AFF1*	same SNP	rs442177	1.05E-18
5	127371588	rs10063647	*LINC01184*	RDW-EA/All	Peripheral blood mononuclear cells	2.78E-16	*FLJ33630*	0.327	rs2250127	2.03E-40
5	127371588	rs10063647	*LINC01184*	RDW-EA/All	CD14+ monocytes	1.48E-12	*FLJ33630*	0.327	rs3749748	3.24E-38
5	127522543	rs10089	*SLC12A2*	RDW-EA/All	Whole blood	2.78E-09	*FBN2*	0.002	rs764369	9.81E-198
7	80300449	rs3211938	*CD36*	RDW-AA/All	Whole blood	6.40E-14	*CD36*	SameSNP	rs3211938	6.40E-14
10	105659826	rs2487999	*OBFC1*	MCV-All	Liver	2.05E-14	*OBFC1*	SameSNP	rs2487999	2.05E-14
17	59483766	rs8068318	*TBX2*	HGB-EA	Fibroblasts	4.09E-06	*C17ORF82*	1	rs2240736	4.09E-06
17	59483766	rs8068318	*TBX2*	HGB-EA	Monocytes (CD14+)	9.97E-07	*CCDC47*	0.527	rs9905140	2.73E-07
20	31140165	rs4911241	*NOL4L*	MCV/RDW-EA; RDW-All	Peripheral blood mononuclear cells	7.65E-11	*ASXL1*	0.293	rs6141282	1.85E-22
20	31140165	rs4911241	*NOL4L*	MCV/RDW-EA; RDW-All	Whole blood	4.37E-07	*ASXL1*	0.293	rs3746612	9.13E-18

Table S7. Single-variant replication results for the variants identified by gene-based testing. A1, reference allele; A2, alternate allele; N, sample size; AF, allele frequency; se, standard error; HCT, hematocrit; HGB, hemoglobin; RBC, red blood cell count; MCV, mean corpuscular volume; MCHC, mean corpuscular hemoglobin concentration; MCH, mean corpuscular hemoglobin; RDW, red blood cell distribution width.

Marker Info				Discovery					Replication					
Gene	**MarkerName**	**A1/A2**	**Trait**	**EAF**	**N**	**beta**	**se**	**Pvalue**	**N**	**EAF**	**beta**	**se**	**Pvalue**	**Study**
PKLR	**1:155260382**	C/T	HGB_EA	0.002	106377	-0.128	0.047	0.006	20723	0.003	-0.287	0.100	0.004	Meta (NWIGM, BIOVU-replication, FINCAVAS, YFS, AIRWAVE)
PKLR	1:155261649	C/T	HGB_EA	0.006	106377	0.083	0.028	0.003	20723	0.005	0.0144	0.156	0.926	Meta (NWIGM, BIOVU-replication, FINCAVAS, YFS, AIRWAVE)
PKLR	**1:155261709**	G/A	HGB_EA	0.003	106377	-0.174	0.040	1.23E-05	20723	0.003	-0.182	0.088	0.039	Meta (NWIGM, BIOVU-replication, FINCAVAS, YFS, AIRWAVE)
PKLR	1:155260382	C/T	HCT_EA	0.002	87444	-0.155	0.052	0.003	18945	0.003	0.069	0.314	0.839	Meta (NWIGM, FINCAVAS, YFS, AIRWAVE)
PKLR	1:155261709	G/A	HCT_EA	0.003	87444	-0.187	0.041	4.53E-06	18945	0.003	-0.130	0.194	0.503	Meta (NWIGM, FINCAVAS, YFS, AIRWAVE)
PKLR	1:155264320	G/T	HCT_EA	0.0004	87444	-0.359	0.174	0.039	NA	NA	NA	NA	NA	Meta (NWIGM, FINCAVAS, YFS)
ALAS2	X:55035659	G/A	MCH_EA	0.0005	54009	0.287	0.125	0.021	3056	0.0015	-0.067	0.333	0.840	Meta (NWIGM, BIOVU-replication)
ALAS2	X:55039960	G/A	MCH_EA	0.002	52758	-0.324	0.053	7.32E-10	5855	0.001	-0.291	0.235	0.215	Meta (NWIGM , BIOVU-replication, YFS, GeneSTAR)
ALPK3	15:85382313	C/T	MCHC_EA	7.53E-06	67917	2.211	0.995	0.026	NA	NA	NA	NA	NA	BIOVU-replication
ALPK3	15:85399650	C/T	MCHC_EA	3.18E-05	67917	1.360	0.499	0.006	1783	0.0003	-0.131	0.994	0.895	BIOVU-replication
ALPK3	15:85400482	G/T	MCHC_EA	3.79E-05	67917	1.555	0.447	0.001	1783	0.0003	1.280	0.995	0.198	BIOVU-replication

Table S8. Association results for five novel RBC trait loci, before and after adjustment for lipids. HGB, hemoglobin; RDW, red blood cell distribution width, TC, total cholesterol, TG, triglycerides, HDL, high-density lipoprotein cholesterol

SNP	RBC trait	Lipid trait	No adjustment				Adjustment for corresponding lipid trait			
			beta	se	p-value	N	beta	se	p-value	N
TMEM57-RHD rs10903129	RDW	TC	-0.0266	0.0141	0.0588	10406	-0.0276	0.0140	0.0490	10406
AFF1 rs442177	HGB	TG	0.0257	0.0096	0.0072	22872	0.0264	0.0095	0.0055	22872
TRIB1 rs2954029	RDW	TG	0.0188	0.0141	0.1818	10405	0.0004	0.0002	0.0356	10405
MAP1A rs55707100	HGB	TG	-0.0529	0.0263	0.0442	22858	-0.0678	0.0262	0.0096	22858
HNF4A rs1800961	HGB	HDL	0.0745	0.0275	0.0068	23441	0.0662	0.0275	0.0161	23441

Additional Funding Information

AIRWAVE

The Airwave Study is funded by the Home Office (grant number 780-TETRA) with additional support from the National Institute for Health Research (NIHR) Imperial College Healthcare NHS Trust (ICHNT) and Imperial College Biomedical Research Centre (BRC) (Grant number BRC-P38084). Paul Elliott is an NIHR Senior Investigator and is supported by the ICHNT and Imperial College BRC, the MRC-PHE Centre for Environment and Health and the NIHR Health Protection Research Unit on Health Impact of Environmental Hazards.

ARIC

The Atherosclerosis Risk in Communities (ARIC) Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research. The meta-analysis and meta-regression analyses were funded by grant R01 HL086694 from the National Heart, Lung, and Blood Institute. The authors thank the staff and participants of the ARIC study for their important contributions.

BioMe

The Mount Sinai IPM Biobank Program is supported by The Andrea and Charles Bronfman Philanthropies.

BIOVU

The dataset used in the analyses described were obtained from Vanderbilt University Medical Center's BioVU which is supported by institutional funding and by the Vanderbilt CTSA grant UL1 TR000445 from NCATS/NIH. Genome-wide genotyping was funded by NIH grants RC2GM092618 from NIGMS/OD and U01HG004603 from NHGRI/NIGMS. Funding for TLE and DRVE was provided by 1R21HL12142902 from NHLBI/NIH. Funding for the BioVU replication cohort was provided by 5R01HD074711 from NICHD/NIH.

CARDIA

The CARDIA Study is conducted and supported by the National Heart, Lung, and Blood Institute in collaboration with the University of Alabama at Birmingham (HHSN268201300025C & HHSN268201300026C), Northwestern University (HHSN268201300027C), University of Minnesota (HHSN268201300028C), Kaiser Foundation Research Institute (HHSN268201300029C), and Johns Hopkins University School of Medicine (HHSN268200900041C). CARDIA is also partially supported by the Intramural Research Program of the National Institute on Aging. Exome Chip genotyping was supported from grants R01-HL093029 and U01- HG004729 to MF. This manuscript has been reviewed and approved by CARDIA for scientific content.

CHS

This CHS research was supported by NHLBI contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086; and NHLBI grants HL080295, HL087652, HL103612, HL105756, HL120393, R01HL068986 with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through AG023629 from the National Institute on Aging (NIA). The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR000124, 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. A full list of CHS investigators and institutions can be found at http://chs-nhlbi.org/.

EGCUT
This study was supported by EU H2020 grants 692145, 676550, 654248, Estonian Research Council Grant IUT20-60, NIASC, EIT – Health and NIH-BMI grant 2R01DK075787-06A1.

FINCAVAS
This work was supported by the Competitive Research Funding of the Tampere University Hospital (Grant 9M048 and 9N035), the Finnish Cultural Foundation, the Finnish Foundation for Cardiovascular Research, the Emil Aaltonen Foundation, Finland, and the Tampere Tuberculosis Foundation. FINCAVAS thanks the staff of the Department of Clinical Physiology for collecting the exercise test data.

Framingham Heart Study
Genotyping, quality control and calling of the Illumina HumanExome BeadChip in the Framingham Heart Study was supported by funding from the National Heart, Lung and Blood Institute Division of Intramural Research (Daniel Levy and Christopher J. O'Donnell, Principle Investigators). Support for the centralized genotype calling was provided by Building on GWAS for NHLBI-diseases: the U.S. CHARGE consortium through the National Institutes of Health (NIH) American Recovery and Reinvestment Act of 2009 (5RC2HL102419). The NHLBI's Framingham Heart Study is a joint project of the National Institutes of Health and Boston University School of Medicine and was supported by contract N01-HC-25195. The FHS authors are pleased to acknowledge that the computational work reported on in this paper was performed on the Shared Computing Cluster, which is administered by Boston University's Research Computing Services. URL: www.bu.edu/tech/support/research/. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

GeneSTAR
GeneSTAR was supported by the National Institutes of Health/National Heart, Lung, and Blood Institute (U01 HL72518, HL087698, and HL112064) 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. Genotyping services were provided through the RS&G Service by the Northwest Genomics Center at the University of Washington, Department of Genome Sciences, under U.S. Federal Government contract number HHSN268201100037C from the National Heart, Lung, and Blood Institute.

HABC
HABC funding/acknowledgement: The Health ABC Study was supported by NIA contracts N01AG62101, N01AG62103, and N01AG62106 and, in part, by the NIA Intramural Research Program. The genome-wide association study was funded by NIA grant 1R01AG032098-01A1 to Wake Forest University Health Sciences and genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268200782096C. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Md. (http://biowulf.nih.gov).

HANDLS
The Healthy Aging in Neighborhoods of Diversity across the Life Span Study (HANDLS) research was supported by the Intramural Research Program of the NIH, National Institute on Aging and the National Center on Minority Health and Health Disparities (project # Z01-AG000513 and human subjects protocol # 2009-149). Data analyses for the HANDLS study utilized the computational resources of the NIH HPC Biowulf cluster at the National Institutes of Health, Bethesda, MD (http://hpc.nih.gov).

Health2006/2008
The Health2006 was financially supported by grants from the Velux Foundation; The Danish Medical Research Council, Danish Agency for Science, Technology and Innovation; The Aase and Ejner Danielsens Foundation; ALK-Abello A/S, Hørsholm, Denmark, and Research Centre for Prevention and Health, the Capital Region of Denmark. The Novo Nordisk Foundation Center for Basic Metabolic Research is an independent Research Center at the University of Copenhagen partially funded by an unrestricted donation from the Novo Nordisk Foundation (www.metabol.ku.dk). This work was supported by the Timber Merchant Vilhelm Bang's Foundation, the Danish Heart Foundation (Grant number 07-10-R61-A1754-B838-22392F), and the Health Insurance Foundation (Helsefonden) (Grant number 2012B233).

JHS
The JHS is supported by contracts HHSN268201300046C, HHSN268201300047C, HHSN268201300048C, HHSN268201300049C, HHSN268201300050C from the National Heart, Lung, and Blood Institute and the National Institute on Minority Health and Health Disparities.

LBC1921/1936
Phenotype collection in the Lothian Birth Cohort 1921 was supported by the UK's Biotechnology and Biological Sciences Research Council (BBSRC), The Royal Society and The Chief Scientist Office of the Scottish Government. Phenotype collection in the Lothian Birth Cohort 1936 was supported by Age UK (The Disconnected Mind project). Genotyping was supported by Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award), Age UK, and the Royal Society of Edinburgh. The work was undertaken by The University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology, part of the cross council Lifelong Health and Wellbeing Initiative (MR/K026992/1). Funding from the BBSRC and Medical Research Council

(MRC) is gratefully acknowledged. WDH is supported by a grant from Age UK (Disconnected Mind Project).

MDCC
The Malmö Diet and Cancer cohort studies were supported by grants from the Swedish Research Council, the Swedish Heart and Lung Foundation, the Påhlsson Foundation, the Novo Nordic Foundation and European Research Council starting grant StG-282255.

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 N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-001079, UL1-TR-000040, and DK063491. 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, and the National Center for Advancing Translational Sciences, Grant UL1TR000124. Funding support for the inflammation dataset was provided by grant HL077449. The MESA Epigenomics & Transcriptomics Study was funded by NIA grant 1R01HL101250-01 to Wake Forest University Health Sciences. MESA thanks its Coordinating Center, MESA investigators, and study staff for their valuable contributions. A full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org.

Montreal Heart Institute Biobank (MHIBB)
We thank all participants and staff of the André and France Desmarais MHIBB. The MHI Biobank acknowledges the technical support of the Beaulieu-Saucier MHI Pharmacogenomic Center. Genotyping of the MHIBB participants was funded by the MHI Foundation. Jean-Claude Tardif holds the Canada Research Chair in translational and personalized medicine and the Université de Montréal endowed research chair in atherosclerosis.

NWIGM
This phase of the eMERGE Network was initiated and funded by the NHGRI through the following grants: U01HG8657 (Group Health Cooperative/University of Washington); U01HG8685 (Brigham and Women's Hospital); U01HG8672 (Vanderbilt University Medical Center); U01HG8666 (Cincinnati Children's Hospital Medical Center); U01HG6379 (Mayo Clinic); U01HG8679 (Geisinger Clinic); U01HG8680 (Columbia University Health Sciences); U01HG8684 (Children's Hospital of Philadelphia); U01HG8673 (Northwestern University); U01HG8701 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG8676 (Partners Healthcare/Broad Institute); and U01HG8664 (Baylor College of Medicine). NWIGM dataset please also add "Additional support was provided by the University of Washington's Northwest Institute of Genetic Medicine from Washington State Life Sciences Discovery funds (Grant 265508).

REGARDS
This research was supported by cooperative agreement U01 NS041588 from the National Institute of Neurological Disorders and Stroke, National Institutes of Health, Department of

Health and Human Services. Additional funding was provided by an investigator-initiated grant-in-aid from Amgen Corporation (Thousand Oaks, California). Amgen did not have any role in the design and conduct of the study or in the collection and management of the data. This research project is supported by a cooperative agreement U01 NS041588 from the National Institute of Neurological Disorders and Stroke, National Institutes of Health, Department of Health and Human Service. The authors thank the other investigators, the staff, and the participants of the REGARDS study for their valuable contributions. A full list of participating REGARDS investigators and institutions can be found at http://www.regardsstudy.org. The genotyping for this project was provided by NIH/NCRR center grant 5U54RR026137-03.
We thank the investigators, staff and participants of the REGARDS study for their valuable contributions. A full list of participating REGARDS investigators and institutions can be found at http://www.regardsstudy.org.

RS
The generation and management of the Illumina exome chip v1.0 array data for the Rotterdam Study (RS-I) was executed by the Human Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands. The Exome chip array data set was funded by the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, from the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO)-sponsored Netherlands Consortium for Healthy Aging (NCHA; project nr. 050-060-810); the Netherlands Organization for Scientific Research (NWO; project number 184021007) and by the Rainbow Project (RP10; Netherlands Exome Chip Project) of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL; www.bbmri.nl). We thank Ms. Mila Jhamai, Ms. Sarah Higgins, and Mr. Marijn Verkerk for their help in creating the exome chip database, and Carolina Medina-Gomez, MSc, Lennard Karsten, MSc, and Linda Broer PhD for QC and variant calling. Variants were called using the best practice protocol developed by Grove et al. as part of the CHARGE consortium exome chip central calling effort.

The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists.

SHIP and SHIP-TREND
SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grants no. 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania, and the network 'Greifswald Approach to Individualized Medicine (GANI_MED)' funded by the Federal Ministry of Education and Research (grant 03IS2061A). ExomeChip data have been supported by the Federal Ministry of Education and Research (grant no. 03Z1CN22) and the Federal State of Mecklenburg-West Pomerania. SHIP and SHIP-TREND are part of the Research Network of Community Medicine of the University Medicine Greifswald (www-community-medicine.de). The studies were supported by the Federal Ministry of Education and Research, by the Federal State of

Mecklenburg-Pomerania and the Siemens AG. The University Medicine Greifswald is a member of the Caché Campus program of the InterSystems GmbH. The SHIP and SHIP-TREND samples were genotyped at the Helmholtz Zentrum München.

STABILITY and SOLID TIMI-52
The STABILITY and SOLID TIMI-52 studies were funded by GlaxoSmithKline. Michelle L. O'Donoghue acknowledges receiving grants from GSK, Merck, AstraZeneca, and Eisa.

WHI
The WHI program is funded by the National Heart, Lung, and Blood Institute, the US National Institutes of Health and the US Department of Health and Human Services (HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C and HHSN271201100004C). Exome chip data and analysis were supported through the Exome Sequencing Project (NHLBI RC2 HL-102924, RC2 HL-102925 and RC2 HL-102926), the Genetics and Epidemiology of Colorectal Cancer Consortium (NCI CA137088), and the Genomics and Randomized Trials Network (NHGRI U01-HG005152). The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at: http://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Short%20List.pdf.

YFS
The Young Finns Study has been financially supported by the Academy of Finland: grants 285902, 286284, 134309 (Eye), 126925, 121584, 124282, 129378 (Salve), 117787 (Gendi), and 41071 (Skidi); the Social Insurance Institution of Finland; Kuopio, Tampere and Turku University Hospital Medical Funds (grant X51001); Juho Vainio Foundation; Paavo Nurmi Foundation; Finnish Foundation of Cardiovascular Research ; Finnish Cultural Foundation; Tampere Tuberculosis Foundation ; Emil Aaltonen Foundation ; and Yrjö Jahnsson Foundation. The expert technical assistance in the statistical analyses by Irina Lisinen is gratefully acknowledged.