ARTICLE

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

Yao Hu,[1,56] Adrienne M. Stilp,[2,56] Caitlin P. McHugh,[2,56] Shuquan Rao,[3,56] Deepti Jain,[2] Xiuwen Zheng,[2] John Lane,[4] Sébastian Méric de Bellefon,[5] Laura M. Raffield,[6] Ming-Huei Chen,[7,8] Lisa R. Yanek,[9] Marsha Wheeler,[10] Yao Yao,[3] Chunyan Ren,[3] Jai Broome,[2] Jee-Young Moon,[11] Paul S. de Vries,[12] Brian D. Hobbs,[13] Quan Sun,[14] Praveen Surendran,[15,16,17,18] Jennifer A. Brody,[19] Thomas W. Blackwell,[20] Hélène Choquet,[21] Kathleen Ryan,[22] Ravindranath Duggirala,[23] Nancy Heard-Costa,[6,8,24] Zhe Wang,[25] Nathalie Chami,[25] Michael H. Preuss,[25] Nancy Min,[26] Lynette Ekunwe,[26] Leslie A. Lange,[27] Mary Cushman,[28] Nauder Faraday,[29] Joanne E. Curran,[23] Laura Almasy,[30] Kousik Kundu,[31,32] Albert V. Smith,[20] Stacey Gabriel,[33] Jerome I. Rotter,[34] Myriam Fornage,[35] Donald M. Lloyd-Jones,[36] Ramachandran S. Vasan,[8,37,38] Nicholas L. Smith,[39,40,41] Kari E. North,[42] Eric Boerwinkle,[12] Lewis C. Becker,[43] Joshua P. Lewis,[22] Goncalo R. Abecasis,[20] Lifang Hou,[36] Jeffrey R. O'Connell,[22] Alanna C. Morrison,[12] Terri H. Beaty,[44] Robert Kaplan,[11] Adolfo Correa,[26] John Blangero,[23] Eric Jorgenson,[21] Bruce M. Psaty,[39,40,45] Charles Kooperberg,[1]

(Author list continued on next page)

Summary

Whole-genome sequencing (WGS), a powerful tool for detecting novel coding and non-coding disease-causing variants, has largely been applied to clinical diagnosis of inherited disorders. Here we leveraged WGS data in up to 62,653 ethnically diverse participants from the NHLBI Trans-Omics for Precision Medicine (TOPMed) program and assessed statistical association of variants with seven red blood cell (RBC) quantitative traits. We discovered 14 single variant-RBC trait associations at 12 genomic loci, which have not been reported previously. Several of the RBC trait-variant associations (*RPN1*, *ELL2*, *MIDN*, *HBB*, *HBA1*, *PIEZO1*, and *G6PD*) were replicated in independent GWAS datasets imputed to the TOPMed reference panel. Most of these discovered variants are rare/low frequency, and several are observed disproportionately among non-European Ancestry (African, Hispanic/Latino, or East Asian) populations. We identified a 3 bp indel p.Lys2169del (g.88717175_88717177TCT[4]) (common only in the Ashkenazi Jewish population) of *PIEZO1*, a gene responsible for the Mendelian red cell disorder hereditary xerocytosis (MIM: 194380), associated with higher mean corpuscular hemoglobin concentration (MCHC). In stepwise conditional analysis and in gene-based rare variant aggregated association analysis, we identified several of the variants in *HBB, HBA1, TMPRSS6,* and *G6PD* that represent the carrier state for known coding, promoter, or splice site loss-of-function variants that cause inherited RBC disorders. Finally, we applied base and nuclease editing to demonstrate that the sentinel variant rs112097551 (nearest gene *RPN1*) acts through a *cis*-regulatory element that exerts long-range control of the gene *RUVBL1* which is essential for hematopoiesis. Together, these results demonstrate the utility of WGS in ethnically diverse population-based samples and gene editing for expanding knowledge of the genetic architecture of quantitative hematologic traits and suggest a continuum between complex trait and Mendelian red cell disorders.

Introduction

Red blood cells (RBCs) or erythrocytes contain hemoglobin, an iron-rich tetramer composed of two alpha-globin and two beta-globin chains. RBCs play an essential role in oxygen transport and also serve important secondary functions in nitric oxide production, regulation of vascular tone, and immune response to pathogens.[1] RBC indices, including hemoglobin (HGB), hematocrit (HCT), mean corpuscular hemoglobin (MCH), mean corpuscular

[1]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98105, USA; [2]Department of Biostatistics, University of Washington, Seattle, WA 98105, USA; [3]Division of Hematology/Oncology, Boston Children's Hospital, Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Stem Cell Institute, Broad Institute, Department of Pediatrics, Harvard Medical School, Boston, MA 02215, USA; [4]Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, Minneapolis, MN 55455, USA; [5]Montreal Heart Institute, Montréal, QC H1T 1C8, Canada; [6]Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA; [7]Population Sciences Branch, Division of Intramural Research, National Heart, Lung and Blood Institute, Bethesda, MD 20892, USA; [8]National Heart Lung and Blood Institute's and Boston University's Framingham Heart Study, Framingham, MA 01701, USA; [9]Division of General Internal Medicine, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [10]Department of Genome Sciences, University of Washington, Seattle, WA 98105, USA; [11]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY 10461, USA; [12]Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [13]Channing Division of Network Medicine and Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital and Harvard Medical School,

(Affiliations continued on next page)

Russell T. Walton,[46] Benjamin P. Kleinstiver,[46,47] Hua Tang,[48] Ruth J.F. Loos,[25] Nicole Soranzo,[16,31,32,49] Adam S. Butterworth,[15,16,17,49,50] Debbie Nickerson,[10] Stephen S. Rich,[51] Braxton D. Mitchell,[22] Andrew D. Johnson,[7,8] Paul L. Auer,[52] Yun Li,[53] Rasika A. Mathias,[54] Guillaume Lettre,[5,55] Nathan Pankratz,[4] Cathy C. Laurie,[2] Cecelia A. Laurie,[2] Daniel E. Bauer,[3] Matthew P. Conomos,[2] and Alexander P. Reiner,[39,*] and theNHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium

hemoglobin concentration (MCHC), mean corpuscular volume (MCV), RBC count, and red blood cell width (RDW), are primary indicators of RBC development, size, and hemoglobin content.[2] These routinely measured clinical laboratory assays may be altered in Mendelian genetic conditions (e.g., hemoglobinopathies such as sickle cell disease [MIM: 603903] or thalassemia [MIM: 613985, 604131], hereditary spherocytosis [MIM: 182900], or G6PD deficiency [MIM: 300908])[3] as well as by non-genetic or nutritional factors (e.g., vitamin B and iron deficiency).

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

Whole-genome sequencing (WGS) data have been generated through the NHLBI Trans-Omics for Precision Medicine (TOPMed) program in very large and ethnically diverse population samples with existing hematologic laboratory measures. These TOPMed WGS data provide novel

Boston, MA 02115, USA; [14]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [15]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK; [16]British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge CB1 8RN, UK; [17]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge CB1 8RN, UK; [18]Rutherford Fund Fellow, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK; [19]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA 98105, USA; [20]TOPMed Informatics Research Center, University of Michigan, Department of Biostatistics, Ann Arbor, MI 48109, USA; [21]Division of Research, Kaiser Permanente Northern California, Oakland, CA 94601, USA; [22]Department of Medicine, Division of Endocrinology, Diabetes & Nutrition, University of Maryland School of Medicine, Baltimore, MD 21201, USA; [23]Department of Human Genetics and South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX 78539, USA; [24]Department of Neurology, Boston University School of Medicine, Boston, MA 02118, USA; [25]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [26]Department of Medicine, University of Mississippi Medical Center, Jackson, MS 39216, USA; [27]Division of Biomedical Informatics and Personalized Medicine, School of Medicine University of Colorado, Anschutz Medical Campus, Aurora, CO 80045, USA; [28]Department of Medicine, Larner College of Medicine at the University of Vermont, Burlington, VT 05405, USA; [29]Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [30]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia and Department of Genetics University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA; [31]Department of Human Genetics, Wellcome Sanger Institute, Hinxton CB10 1SA, UK; [32]Department of Haematology, University of Cambridge, Cambridge CB2 0PT, UK; [33]Broad Institute, Boston, MA 02142, USA; [34]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA 90502, USA; [35]University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [36]Northwestern University, Chicago, IL 60208, USA; [37]Departments of Cardiology and Preventive Medicine, Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA; [38]Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA; [39]Department of Epidemiology, University of Washington, Seattle, WA 98105, USA; [40]Kaiser Permanente Washington Health Research Institute, Kaiser Permanente Washington, Seattle, WA 98105, USA; [41]Seattle Epidemiologic Research and Information Center, Department of Veterans Affairs Office of Research and Development, Seattle, WA 98105, USA; [42]Department of Epidemiology, Gillings School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [43]Division of Cardiology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [44]School of Public Health, John Hopkins University, Baltimore, MD 21205, USA; [45]Department of Medicine, University of Washington, Seattle, WA 98105, USA; [46]Center for Genomic Medicine and Department of Pathology, Massachusetts General Hospital, Boston, MA 02114, USA; [47]Department of Pathology, Harvard Medical School, Boston, MA 02115, USA; [48]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA; [49]National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Genomics, University of Cambridge, Cambridge CB1 8RN, UK; [50]National Institute for Health Research Cambridge Biomedical Research Centre, University of Cambridge and Cambridge University Hospitals, Cambridge CB1 8RN, UK; [51]Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia School of Medicine, Charlottesville, VA 22903, USA; [52]Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI 53205, USA; [53]Departments of Biostatistics, Genetics, Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [54]Division of Allergy and Clinical Immunology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MA 21205, USA; [55]Faculté de Médecine, Université de Montréal, Montréal, QC H1T 1C8, Canada
[56]These authors contributed equally
*Correspondence: apreiner@uw.edu
https://doi.org/10.1016/j.ajhg.2021.04.003.

opportunities to assess rare and common single-nucleotide and indel variants across the genome, including variants more common in African, East Asian, or Native American ancestry individuals that are not captured by existing GWAS arrays or imputation reference panels. We thereby aimed to identify previously undescribed genetic variants and genes associated with the seven RBC indices and to dissect association signals at previously reported regions through conditional analysis and fine-mapping.

Subjects and methods

TOPMed study population

The analyses reported here included 62,653 participants from 13 TOPMed studies: Genetics of Cardiometabolic Health in the Amish (Amish, n = 1,102), Atherosclerosis Risk in Communities Study VTE cohort (ARIC, n = 8,118), Mount Sinai BioMe Biobank (BioMe, n = 10,993), Coronary Artery Risk Development in Young Adults (CARDIA, n = 3,042), Cardiovascular Health Study (CHS, n = 3,490), Genetic Epidemiology of COPD Study (COPDGene, n = 5,794), Framingham Heart Study (FHS, n = 3,141), Genetic Studies of Atherosclerosis Risk (GeneSTAR, n = 1,713), Hispanic Community Health Study - Study of Latinos (HCHS_SOL, n = 7,655), Jackson Heart Study (JHS, n = 3,033), Multi-Ethnic Study of Atherosclerosis (MESA, n = 2,499), Whole Genome Sequencing to Identify Causal Genetic Variants Influencing CVD Risk - San Antonio Family Studies (SAFS, n = 1,153), and Women's Health Initiative (WHI, n = 10,920). The composition of the 62,653 participants by race/ethnicity is 54% white, 23% Black, 22% Hispanic/Latino, and 1% Asian (see Table S1 and supplemental methods for details). Further descriptions of the design of the participating TOPMed cohorts and the sampling of individuals within each cohort for TOPMed WGS are provided in the section "Participating studies" in the supplemental methods. We analyzed each of seven red blood cell traits separately, accounting for any unique sampling features within each study. The total counts of participants, mean age, and the count of male participants from each study stratified by trait are shown in Table 1. All studies were approved by the appropriate institutional review boards (IRBs), and informed consent was obtained from all participants.

RBC trait measurements and exclusion criteria in TOPMed

The seven RBC traits considered for analyses were measured from freshly collected whole blood samples at local clinical laboratories using automated hematology analyzers calibrated to manufacturer recommendations according to clinical laboratory standards. Each trait was defined as follows. HCT is the percentage of volume of blood that is composed of red blood cells. HGB is the mass per volume (grams per deciliter) of hemoglobin in the blood. MCH is the average mass in picograms of hemoglobin per red blood cell. MCHC is the average mass concentration (grams per deciliter) of hemoglobin per red blood cell. MCV is the average volume of red blood cells, measured in femtoliters. RBC count is the count of red blood cells in the blood, by number concentration in millions per microliter. RDW is the measurement of the ratio of variation in width to the mean width of the red blood cell volume distribution curve taken at ±1 CV. In studies where multiple blood cell measurements per participant were available, we selected a single measurement for each trait and each participant as described further in supplemental methods. Each trait was analyzed to identify extreme values that may have been measurement or recording errors and such observations were removed from the analysis (see supplemental methods). Table 1 displays the mean and standard deviation among participants analyzed after exclusions by study. The pairwise correlation among the seven RBC traits is shown in Table S2.

WGS data and quality control in TOPMed

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

Single-variant association analysis

Single-variant association tests were performed for each of the seven RBC traits separately using linear mixed models (LMMs). In each case, a model assuming no association between the outcome and any genetic variant was first fit; we refer to this as the "null model." In the null model, covariates modeled as fixed effects were sex; age at trait measurement; a variable indicating TOPMed study and phase of genotyping (study_phase); indicators of whether the participant is known to have had a stroke, chronic obstructive pulmonary disease (COPD), or a venous thromboembolism (VTE) event; and the first 11 PC-AiR[24] principal components (PCs) of genetic ancestry. A 4th degree sparse empirical kinship matrix (KM) computed with PC-Relate[25] was included to account for genetic relatedness among participants. Additional details on the computation of the ancestry PCs and the sparse KM are provided in the supplemental methods. Finally, we allowed for heterogeneous residual variances by study and ancestry group (e.g., ARIC_White), as this has been shown previously to control inflation.[26] The details on how we estimated the ancestry group for this adjustment are in the supplemental methods. The numbers of individuals per ancestry group per study and the respective mean and standard deviation for each trait are shown in Table S3.

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

1. Fit a LMM, with the fixed effect covariates, sparse KM, and heterogeneous residual variance model as described above. Perform a rank-based inverse-normal transformation of the marginal residuals, and subsequently rescale by their

Table 1. Characteristics of the TOPMed samples by study

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

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

variance prior to transformation. This rescaling allows for clearer interpretation of estimated genotype effect sizes from the subsequent association tests.

2. Fit a second LMM using the rank-normalized and re-scaled residuals as the outcome, with the same fixed effect covariates, sparse KM, and heterogeneous residual variance model as in stage 1.

The output of the stage 2 null model was then used to perform genome-wide score tests of genetic association for all individual variants with minor allele count (MAC) $\geq$ 5 that passed the TOPMed variant quality filters and had less than 10% of samples freeze-wide with sequencing read depth $<$ 10 at that particular variant. We tested up to 102,674,666 SNVs and 7,722,116 indels (Table S4). Genome-wide significance was determined at the $p < 5E{-}9$ level.[28] For each locus, we defined the top variant as the most significant variant within a 2 Mb window. All association analyses were performed using the GENESIS software.[29]

Conditional analysis

Because of the very large number of variants and genomic loci that have recently been associated with quantitative RBC traits, following the single-variant association analyses, we systematically performed a series of conditional association analyses for each trait to determine which genome-wide significant associations were independent of previously reported RBC variants. We gathered the variants known to be associated with each phenotype from previous publications (Table S5) and matched these to TOPMed variants using position and alleles. Then, genome-wide conditional association analyses were performed by including known variants as fixed effects covariates in the null model using the same fully adjusted two-stage LMM association testing procedure described above. We performed three types of conditional analysis, namely the trait-specific, the trait-agnostic, and the iterative, stepwise conditional analysis to identify a set of conditionally independent variants that have not been previously reported (supplemental methods).

Single-variant association analysis of chromosome 16

The alpha-globin gene region on chromosome 16p13.3 contains a large, 3.7 kb structural variant (esv3637548, chr16: 173,529–177,641) common among African ancestry individuals known to be highly significantly associated with all RBC traits.[15,18] This large copy number variant is not well-tagged by SNVs in the region. Therefore, we performed genotype calling for the alpha-globin 3.7 kb CNV in 52,772 available TOPMed whole genomes using MosDepth.[30] Since the chromosome 16 alpha-globin CNV calls were available for only a subset of the samples in the primary analyses, to assess the effect of conditioning on the alpha-globin CNV, the same set of analyses described above were run for chromosome 16 restricted to the sample set with alpha-globin CNV calls. The most probable alpha-globin copy number was included as a categorical variable to allow for potential non-linear effects on the phenotype.

Proportion of variance explained

For each trait, we estimated the proportion of variance explained (PVE) by the set of LD-pruned known associated variants, by the final set of conditionally independent variants we identified following the iterative stepwise conditional analysis, and by both sets together. These cumulative PVE values were estimated jointly from the stage 2 null model using approximations from multi-parameter score tests, thus accounting for covariance between the variant effect size estimates. The PVE estimates were calculated using the full sample set and did not include the alpha-globin CNV as a known variant but did include the set of conditionally independent SNVs and indels identified on chromosome 16 after conditioning on the alpha-globin CNV. More details are provided in the supplemental methods.

Replication studies for single-variant association findings

We sought replication of the lead variants at genome-wide significant loci identified in the trait-specific conditional analysis in

independent studies including the INTERVAL study, the Kaiser-Permanente Genetic Epidemiology Research on Aging (GERA) cohort, samples from the Women's Health Initiative - SNP Health Association Resource (WHI-SHARe)[31] not included in TOPMed, European ancestry samples from phase 1 of the UK BioBank (UKBB),[9] and African and East Asian ancestry samples from phase 2 of UKBB.[21] WGS data were used in INTERVAL while genotyping on various arrays and imputation to TOPMed WGS data or 1000 Genomes Phase 3 reference panels were performed in Kaiser, WHI-SHARe, and UKBB. Residuals were obtained by regressing the harmonized RBC traits on age, sex, the first 10 PCs in each study stratified by ancestry, followed by association analyses testing each genetic variant with the inverse-normalized residual values. Summary statistics from each study were combined through fixed-effect inverse-weighting meta-analysis using METAL.[32]

Aggregate variant association analysis of rare variants within each gene

Association tests aggregating rare variants by gene were performed for each RBC trait in order to assess the cumulative effect of rare variants within each gene and associated regulatory regions. We applied five strategies for grouping and filtering variants. Three of them aggregated coding variants and two of them aggregated coding and non-coding regulatory variants. For each aggregation strategy we filtered variants using one or more deleterious prediction scores creating relatively relaxed or stringent sets of variants (see details in supplemental methods). The five strategies are referred to as C1-S, C1-R, C2-R, C2-R+NC-S, and C2-R+NC-R by abbreviating coding to "C," non-coding to "NC," stringent to "S," and relaxed to "R." For all aggregate units, only variants with MAF < 0.01 that passed the quality filters and had less than 10% of samples with sequencing read depth < 10 were considered. The aggregate association tests were performed using the Efficient Variant-Set Mixed Model Association Test (SMMAT).[33] The SMMAT test used the same fully adjusted two-stage null model as was fit for the single variant association tests, therefore adjusting for the same covariates, kinship, and residual variance structure as the single variant association analyses. For each aggregation unit, SMMAT efficiently combines a burden test p value with an asymptotically independent adjusted "SKAT-type" test p value using Fisher's method. This testing approach is more powerful than either a burden or SKAT[34] test alone and is computationally more efficient than the SKAT-O test.[35] Wu weights[34] based on the variant MAF were used to upweight rarer variants in the aggregation units. Significance was determined using a Bonferroni threshold, adjusting for the number of gene-based aggregation units tested genome-wide with cumulative MAC ≥ 5. Two types of conditional analysis were run ("trait-specific" and "trait-agnostic), conditioning previously reported RBC trait-associated variants as well as those discovered in the TOPMed single variant tests (Table S5). In addition, any previously reported RBC trait-associated variants and the set of conditionally independent variants identified in our single variant analyses were excluded from the gene-based aggregation units.

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

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

We analyzed each study-ethnic group separately, adjusting for sex, age, and smoking status. We then normalized the residuals with each group using inverse normal transformation. We performed association testing per ethnic group with EPACTS. We adjusted all analyses using the first ten PCs and a kinship matrix (EMMAX) calculated using 150,000 common variants in LD. For pLoF, we tested an additive genetic model. For pKO, we coded individuals as "0" if they were not a pKO and as "1" if they were a pKO. We meta-analyzed association results using METAL.[32] We excluded variants located in the alpha-globin region in self-reported African-ancestry individuals. The genome-wide significant threshold for each ancestral group was defined as $p < 0.05/\text{number}$ of variants. Sensitivity analyses testing hemoglobin levels with LoF variants on chromosome 11 showed that adjustment for smoking status has minimal impact on the association results (Pearson's correlation of p values > 0.99).

Lentivirus packaging

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

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

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

Generation of AncBE4max-SpRY-expressing stable HUDEP-2 cell lines

The lentiviral plasmid for AncBE4max-SpRY[41] was generated by subcloning the coding sequence of nSpRY(D10A) into the AgeI and XcmI restriction sites of pRDA_257 (pLenti-BPNLS-AncBE4-gsXTEN gs-nSpCas9-gs-UGI-gs-BPNLS-P2A-Puro), generously provided by John Doench (Broad Institute). Lentivirus was produced as described

above. HUDEP-2 cells were transduced with lentivirus, and 1 μg/mL puromycin was added into culture medium 2 days after lentiviral transduction. After 2-week positive selection, AncBE4max-SpRY editing efficiency was tested using multiple sgRNAs with variable PAM sequence.

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

The sequence of single-guide RNA targeting rs112097551 (chr3:128,603,774, GenBank: NC_000003.12, g.128603774G>A) is summarized in Table S6. Oligos (from GENEWIZ company) were annealed and ligated into LentiGuide-Puro (Addgene plasmid 52963). Following lentiviral production and transduction into cell lines with stable SpCas9 expression, 1 μg/mL puromycin were added to select for sgRNA integrants in HUDEP-2 cells expressing AncBE4max-SpRY. C-to-T editing efficiency was determined in bulk cells 10 days after lentiviral delivery into AncBE4-max-SpRY-expressing HUDEP-2 cells (Figure S1). Briefly, genomic DNA was extracted using the QIAGEN Blood and Tissue kit. Genomic region surrounding the sgRNA targeting site was amplified using HotStarTaq DNA polymerase (QIAGEN, Cat# 203203) for other PCR reactions strictly following the manufactory instructions with variable annealing temperature. PCR products were subject to Sanger sequencing and then EditR analysis to estimate the editing efficiency based on sequencing chromatograms.[42] Single HUDEP-2 cells were plated to obtain highly edited clones. Primers for PCR were summarized in Table S7.

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

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

Determination of target gene expression

Total RNA was extracted from cell cultures 4 days after electroporation using the RNeasy Plus Mini Kit (QIAGEN) and reverse transcribed using the iScript cDNA synthesis kit (Biorad) according to the manufacturer's instructions. Expression of target genes was quantified using real-time RT-qPCR with *GAPDH* (MIM: 138400) as an internal control. All gene expression data represent the mean of at least three biological replicates. Primers for PCR are summarized in Table S7.

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

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

Results

Single-variant association analysis

In the single-variant association analyses, the genomic inflation factors ranged from 1.015 to 1.038, indicating adequate control of population stratification and relatedness (Table S8). A total of 69 loci reached genome-wide significance for any of the seven RBC traits ($p < 5E{-}9$, Figure S2 and Table S9). Of the 69 loci, 9 (*HBB*, *HBA1*, *RPN1*, *ELL2*, *EIF5-MARK3*, *MIDN*, *PIEZO1*, *TMPRSS6*, and *G6PD* [MIM: 141900, 141800, 180470, 601874, 601710, 606700, 611184, 609862, 305900, respectively]) remained significant in the conditional analysis after accounting for RBC trait-specific known loci. In addition, three more loci reached genome-wide significance following RBC trait-specific conditional analysis (*19q12*, *10q26*, and *SHANK2* [MIM: 603290], $p < 5E{-}9$, Figure S3). Therefore, a total of 12 loci showed genome-wide significance for association with at least one of the seven RBC traits in the trait-specific conditional analysis, indicating signals independent of previously reported variants ($p < 5E{-}9$) (Figure S4, Table 2).

At the 12 significant loci identified in the trait-specific conditional analyses which have not been reported previously, the number of genome-wide significant variants ranged from 1 to 162 (Figure S4 and Table S10). Six loci harbored more than one genome-wide significant variants (*HBB*, *HBA1*, *ELL2, MIDN, TMPRSS6,* and *G6PD*). The lead variants for each trait at each of these 12 loci (including, across the 7 traits, 14 distinct variants [12 SNVs and 2 small indels]) are shown in Table 2. Notably, only two lead variants (*MIDN*-rs73494666, chr19: 1,253,643, GenBank: NC_000019.10, g.1253643C>T and *TMPRSS6*-rs228914, chr22: 37,108,472, GenBank: NC_000022.11, g.37108472C>A) had MAF > 5% in TOPMed. Most of these 14 lead variants were located within non-coding regions of the genome and most were low frequency (n = 3 between MAF 0.1% and MAF 2%) or rare (n = 9 with MAF < 0.1%). The latter category included three loci (*SHANK2*-rs535577177 [chr11: 70,462,791, GenBank: NC_000011.10, g.70462791G>A], *10q26*-rs98641 5672 [chr10: 131,440,166, GenBank: NC_000010.11, g.131440166C>T], and *19q12*-rs136850044 [chr19: 28,868,

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

Trait	Variant	Chr:Pos (GRCh38)	Gene	CA/NCA	CAF(%)	N	Beta	SE	p	$p_{conditional1}$[a]	$p_{conditional2}$[b]
HCT	rs11549407	11: 5,226,774	*HBB*	A/G	0.026	62,487	−4.94	0.67	1.68E−13	3.43E−13	1.55E−12
HGB	rs11549407	11: 5,226,774	*HBB*	A/G	0.026	62,461	−2.14	0.23	2.86E−21	4.76E−21	1.75E−20
	rs1368500441	19: 28,868,893	*19q12*	A/G	0.005	62,461	2.65	0.46	1.02E−8	2.49E−9	6.64E−8
MCH	rs112097551	3:128,603,774	*RPN1*	A/G	0.398	62,461	0.78	0.12	4.01E−10	4.27E−11	4.08E−10
	rs116635225	5: 95,989,447	*ELL2*	A/G	1.307	46,241	−0.43	0.07	3.37E−9	1.18E−11	2.58E−11
	rs986415672	10: 131,440,166	*10q26*	T/C	0.006	46,241	−4.26	0.82	2.16E−7	3.06E−9	2.49E−9
	rs34598529	11: 5,227,100	*HBB*	C/T	0.083	46,241	−4.31	0.29	1.06E−49	1.37E−52	1.03E−53
	rs535577177	11: 70,462,791	*SHANK2*	A/G	0.008	46,241	−4.72	0.82	1.04E−8	8.28E−10	3.38E−9
	rs370308370	14: 103,044,696	*EIF5/MARK3*	A/G	0.011	46,241	−4.35	0.74	3.15E−9	1.42E−9	5.49E−9
	rs868351380	16: 55,649	*HBA1/2*	C/G	0.022	37,917	−3.19	0.51	4.85E−10	8.87E−11	1.49E−11
	rs73494666	19: 1,253,643	*MIDN*	T/C	16.5	46,241	−0.16	0.03	1.11E−9	4.27E−11	9.00E−9
	rs228914	22: 37,108,472	*TMPRSS6*	A/C	89.0	46,241	−0.09	0.02	3.76E−5	6.53E−10	2.76E−8
MCHC	rs11549407	11: 5,226,774	*HBB*	A/G	0.028	52,648	−1.79	0.18	4.79E−23	1.21E−23	1.87E−23
	rs763477215	16: 88,717,174	*PIEZO1*	A/ATCT	0.070	52,648	0.66	0.11	1.57E−9	2.66E−9	1.74E−9
MCV	rs112097551	3:128,603,774	*RPN1*	A/G	0.405	48,830	1.98	0.31	1.09E−10	7.65E−12	6.28E−10
	rs11549407	11: 5,226,774	*HBB*	A/G	0.028	48,830	−16.5	1.08	3.52E−53	1.00E−54	1.31E−55
	rs868351380	16: 55,649	*HBA1/2*	C/G	0.022	39,107	−7.99	1.31	1.19E−9	2.17E−10	3.20E−11
	rs73494666	19: 1,253,643	*MIDN*	T/C	16.7	48,830	−0.42	0.07	3.90E−10	2.72E−10	1.77E−11
	rs228914	22: 37,108,472	*TMPRSS6*	A/C	89.1	48,830	−0.20	0.06	3.80E−4	9.53E−10	2.52E−6
RBC	rs34598529	11: 5,227,100	*HBB*	C/T	0.084	44,470	0.55	0.06	3.59E−22	1.48E−25	1.91E−23
	rs372755452	16: 199,621	*HBA1/2*	A/AG	0.010	36,430	1.27	0.18	1.55E−12	6.08E−10	3.95E−9
RDW	rs34598529	11: 5,227,100	*HBB*	C/T	0.092	29,385	1.96	0.22	4.44E−19	1.35E−20	2.16E−20
	rs76723693	X: 154,533,025	*G6PD*	G/A	0.297	29,385	−0.91	0.10	2.38E−19	2.97E-20	2.99E−15

Conditional analysis at the *HBA1/2* locus was performed in a subset of TOPMed samples with available alpha-globin CNV data. Abbreviations are as follows: Chr, chromosome; Pos, position; CA, coded allele; NCA, non-coded allele; CAF, coded allele frequency; HCT, hematocrit; HGB, hemoglobin; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; RBC, red blood cell count; RDW, red blood cell width.
[a]In the first conditional analysis, trait-specific reported variants were adjusted in the model.
[b]In the second conditional analysis, all reported variants regardless of associated traits were adjusted in the model.

893, GenBank: NC_000019.10, g.28868893G>A]) in which the lead variant was extremely rare with MAF < 0.01%. Several of the lead variants showed large allele frequency differences between race/ethnicity groups as assessed from the genome aggregation database or gnomAD (Table S11). The *RPN1*-rs112097551 (chr3: 128,603,774, GenBank: NC_000003.12, g.128603774G>A), *HBB*-rs34598529 (chr11: 5,227,100, GenBank: NC_000011.10, g.5227100T>C), *G6PD*-rs76723693 (chrX: 154,533,025, GenBank: NC_000023.11, g.154533025A>G, NP_001346945.1, p.Leu323Pro), *MIDN*-rs73494666 (chr19: 1,253,643, GenBank: NC_000019.10, g.1253643C>T), and *ELL2*-rs116635225 (chr5: 95,989,447, GenBank: NC_000005.10, g.95989447G>A) variants are found disproportionately among individuals of African ancestry. The *EIF5/MARK3*-rs370308370 (chr14: 103,044,696, GenBank: NC_000014.9, g.103044696G>A) and chromosome 16p13.3 alpha-globin locus (rs372755452, chr16: 199,621, GenBank: NC_000016.10, g.199622del) variants are found only among East Asians. The alpha-globin locus variant rs868351380 (chr16: 55649, GenBank: NC_000016.10, g.55649G>C) and *PIEZO1* variant rs763477215 (chr16: 88,717,174, GenBank: NC_000016.10, g.88717175_88717177TCT[4], GenBank: NP_001136336.2, p.Lys2169del) are more common among Hispanics/Latinos and Europeans, respectively.

Replication of single-variant discoveries

We sought replication for each of the 14 discovered variants in INTERVAL, the Kaiser Permanente GERA Study, the WHI-SHARe study, and UKBB phase 1 European and phase 2 African and East Asian samples (Table S12). Several of the rare variants (*SHANK2*-rs535577177, *10q26*-rs986415672, *19q12*-rs1368500441, *EIF5/MARK3*-rs370308370, and *HBB*-rs11549407 [chr11: 5,226,774, GenBank: NC_000011.10, g.5226774G>A, GenBank: NP_000509.1, p.Gln40Ter]) were not available for testing in any of the replication studies due to low frequency, population specificity, and/or poor imputation quality. For eight of the nine lead variants with available genotype data for testing, we successfully replicated each of the trait-specific

associations for *HBB*-rs34598529, *HBA1*-rs868351380 (chr16: 55,649, GenBank: NC_000016.10, g.55649G>C), *HBA1*-rs372755452 (chr16: 199,622, GenBank: NC_000016.10, g.199622del), *RPN1*, *ELL2*, *PIEZO1*, *G6PD*, and *MIDN* (meta-analysis $p < 5.6E{-}3$, 0.05/9 loci, with consistent directions of effect). The replication p value for the lead variant at *TMPRSS6* did not reach the predetermined significance threshold, but the association was directionally consistent. We further note that several of our identified TOPMed single variant-RBC trait associations (*RPN1*, *HBB*-rs11549407 and -rs34598529, and *MIDN*) reached genome-wide significance in recently published very large European ancestry or multi-ethnic imputed GWASs.[19,21,46]

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

Several of the variants we discovered in the single-variant association analysis (particularly those replicated in independent samples) in Table 2 are located within genomic regions known to harbor common variants associated with RBC quantitative traits and/or variants responsible for Mendelian blood cell disorders, such as hemoglobinopathies (*HBB*, *HBA1/HBA2* [MIM: 141850]) and various hemolytic or non-hemolytic anemias (*G6PD*, *PIEZO1*, *TMPRSS6*, and *GATA2-RPN1* [MIM: 137295]). At the *HBB* locus, the lead variant associated with lower HCT, HGB, MCHC, and MCV is a LoF variant (rs11549407 encoding p.Gln40Ter, MAF = 0.026%) while the lead variant associated with lower MCH and higher RBC, and higher RDW is a variant located within the *HBB* promoter region (rs34598529, MAF = 0.083%). At the *HBA1/HBA2* locus, the lead variant for MCH and MCV, rs868351380 (MAF = 0.022%), is located ~125 kb upstream of *HBA1/HBA2* in an intron of *SNRNP25*, and the lead variant for RBC, rs372755452 (MAF = 0.010%), is located ~30 kb downstream of *HBA1/HBA2* in an intron of *LUC7L* (MIM: 607782). The *GATA2-RPN1* locus, which contains variants previously reported for association with MCH and RDW in a European-only analysis (rs2977562 [chr3:128,387,424, GenBank: NC_000003.12, g.128387424A>G] and rs147412900 [chr3:128,575,268, GenBank: NC_000003.12, g.128575268G>A]),[13] was associated with MCH and MCV in TOPMed (lead variant rs112097551, p = 4.27E−11). The MAF of the lead variant at the *GATA2-RPN1* locus in all TOPMed samples is 0.4% but is 5.9 times more common among African than non-African samples according to gnomAD. At the *G6PD* locus, the lead variant associated with lower RDW was a missense variant rs76723693, which encodes p.Leu323Pro. At the *PIEZO1* locus, the most significant variant was an in-frame 3 bp deletion rs763477215 (p.Lys2169del) associated with higher MCHC. While the index SNP rs228914 at *TMPRSS6* has not been previously associated with RBC parameters, rs228914 is a *cis*-eQTL for *TMPRSS6* and an LD surrogate rs228916 (chr22: 37,109,512, GenBank: NC_000022.11, g.37109512C>T) has been previously associated with serum iron levels.[47] The remaining genetic loci (*SHANK2*, *ELL2*, *19q12*, *10q26*, *EIF5/MARK3*, and *MIDN*) have less clear functional relationships to RBC phenotypes. Moreover, the lead variants at *EIF5/MARK3* and *MIDN* for MCH and the lead variant at *TMPRSS6* for MCH and MCV were partially attenuated in the trait-agnostic conditional analysis.

Iterative conditional analysis identifies extensive allelic heterogeneity at HBB locus

We next performed stepwise conditional analysis to dissect association signals within each of the six loci harboring more than one genome-wide significant variants in the RBC trait-specific conditional analysis. One of the six regions (*HBB*) was found to have multiple, genome-wide significant variants independent of previously reported loci. The largest number of independent signals were observed for association with MCH (11 signals, Table S13). All independent variants at the *HBB* locus had MAF < 1%. No secondary independent signals were discovered in other regions (*HBA1/2, ELL2*, *MIDN*, *TMPRSS6*, and *G6PD*). For each RBC trait, we estimated the PVE by the set of LD-pruned known variants, by the conditionally independent variants identified in stepwise conditional analysis, and by both sets together (Table S14). In total, the PVE ranged from 3.4% (HCT) to 21.3% (MCH). The identified set of genetic variants that have not been described previously explained up to 3% of phenotypic variance (for MCH and MCV).

Rare variant aggregated association analysis

We next examined rare variants with MAF < 1% in TOPMed, aggregated based on protein-coding and non-coding gene units from GENCODE. To enrich for likely causal variants in the aggregation units, we used five different variant grouping and filtering strategies based on coding sequence and regulatory (gene promoter/enhancer) functional annotations (see supplemental methods). After accounting for all previously reported RBC trait-specific single variants, a total of five loci were significantly associated with one or more RBC traits using various aggregation strategies (Tables 3 and S15). These include genes encoding *HBA1/HBA2*, *TMPRSS6*, *G6PD*, and *CD36* (MIM: 173510), as well as several genes and non-coding RNAs within the beta-globin locus on chromosome 11p15 (*HBB*, *HBG1* [MIM: 142200], *CTD-264317.6* [MIM: 604927], *OR52H1*, *RF60021*, and *OR52R1*). Some of the gene units in the chromosome 11p15 beta-globin region (*HBG1*, *OR52R1*, and *RF00621*) became non-significant after further adjustment for all known RBC variants in the trait-agnostic conditional analysis (Table 3). After additionally accounting for all 11 independent single-variant signals identified in TOPMed at the *HBB* locus in stepwise conditional analysis (Table S13), as well as all trait-specific known variants, five coding genes remained significant (*HBA1/HBA2*, *HBB*, *TMPRSS6*, *G6PD*, and *CD36*, Table S16) and two additional genes (*TFRC* [MIM: 190010] and *SLC12A7* [MIM: 604879]) reached

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

Trait	Chr (GRCh38)	Start (GRCh38)	End (GRCh38)	Gene	No. of variants	MAC	p	$p_{conditional1}$[a]	$p_{conditional2}$[b]
HCT	11	5225464	5229395	*HBB*	15	76	1.27E−23	1.35E−23	5.91E−18
	11	5224309	5225461	*AC104389.6*	94	1,395	1.85E−13	6.23E−15	3.32E−11
HGB	11	5225464	5229395	*HBB*	15	76	2.06E−35	8.99E−30	7.44E−29
	11	5224309	5225461	*AC104389.6*	94	1,394	1.29E−18	2.43E−17	1.05E−23
MCH	11	5224309	5225461	*AC104389.6*	83	1,078	6.76E−100	2.87E−104	5.51E−95
	11	5225464	5229395	*HBB*	34	126	9.53E−76	2.76E−78	3.11E−75
	11	5224448	5224639	*RF00621*	588	12,096	1.93E−20	4.02E−20	1.28E−12
	11	5544489	5548533	*OR52H1*	8	441	6.15E−16	6.13E−17	9.82E−18
	11	5248079	5249859	*HBG1*	526	7,852	9.95E−09	8.61E−9	8.36E−4
	16	176680	177522	*HBA1*	16	30	4.97E−6	5.95E−9	1.98E−9
	22	37065436	37109713	*TMPRSS6*	243	3,317	6.77E−07	9.92E−12	1.16E−9
	X	154531391	154547572	*G6PD*	59	599	2.32E−06	6.59E−7	2.50E−7
MCHC	11	5224309	5225461	*AC104389.6*	88	1,225	2.37E−64	5.01E−40	8.73E−39
	11	5225464	5229395	*HBB*	36	136	4.07E−34	1.04E−33	2.65E−31
	11	5544489	5548533	*OR52H1*	8	502	3.88E−07	2.12E−6	7.50E−7
MCV	11	5224309	5225461	*AC104389.6*	86	1,148	2.29E−153	1.40E−148	4.75E−108
	11	5225464	5229395	*HBB*	35	130	4.10E−82	6.02E−86	1.11E−81
	11	5224448	5224639	*RF00621*	597	12,848	3.11E−37	1.56E−30	2.74E−16
	11	5544489	5548533	*OR52H1*	8	468	1.07E−19	3.29E−19	4.50E−22
	11	5248079	5249859	*HBG1*	546	8,321	4.46E−15	5.71E−8	1.79E−2
	16	176680	177522	*HBA1*	16	30	5.11E−4	2.03E−6	9.24E−7
	22	37065436	37109713	*TMPRSS6*	252	3,567	8.61E−06	9.11E−10	9.90E−8
	X	154531390	154547572	*G6PD*	82	732	2.19E−12	2.70E−13	7.06E−14
RBC	11	5224309	5225461	*AC104389.6*	81	1,036	9.51E−57	5.47E−60	2.55E−44
	11	5225464	5229395	*HBB*	34	113	2.24E−24	5.35E−28	6.06E−25
	11	5224448	5224639	*RF00621*	576	11,551	6.13E−15	7.39E−15	7.31E−7
	11	4803433	4804380	*OR52R1*	72	1,551	4.48E−09	1.87E−9	9.37E−2
	11	5248079	5249859	*HBG1*	517	7,502	2.74E−07	4.09E−8	3.49E−1
	X	154531390	154547572	*G6PD*	58	574	1.29E−06	2.99E−9	3.49E−8
RDW	7	80369575	80679277	*CD36*	178	1,537	3.28E−4	6.45E−7	2.46E−6
	11	5224309	5225461	*AC104389.6*	73	702	1.55E−29	1.19E−30	2.84E−24
	11	5225464	5229395	*HBB*	13	54	2.06E−24	9.07E−27	1.14E−24
	11	5544489	5548533	*OR52H1*	7	300	1.20E−08	4.55E−9	7.08E−9
	11	5224448	5224639	*RF00621*	480	8,119	1.80E−08	1.21E−8	2.01E−4
	22	37065436	37109713	*TMPRSS6*	72	614	2.89E−07	1.38E−7	4.86E−8
	X	154531390	154547572	*G6PD*	47	449	2.13E−24	6.71E−27	8.33E−21

Conditional analysis at the *HBA1/2* locus was performed in a subset of TOPMed samples with available alpha-globin CNV data. Abbreviations are as follows: Chr, chromosome; MAC, minor allele counts; HCT, hematocrit; HGB, hemoglobin; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; RBC, red blood cell count; RDW, red blood cell width.

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

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

significance threshold (Table S16). *AC104389.6*, a non-coding gene 2 bp downstream of *HBB*, was also found significant in the aggregation approach where we included upstream regulatory variants, but the variants including in this unit are predominately the same ones tested in the *HBB* gene unit and hence we have not reported this gene unit as a distinct signal.

Notably, each of the seven genes (*HBA1/HBA2, HBB, TMPRSS6, G6PD, CD36, TFRC*, and *SLC12A7*) identified in rare variant aggregate analyses are known to harbor common non-coding or coding variants previously associated with RBC traits or disorders. We further explored the overall patterns of association, individual rare variants driving the associations, and their annotations (Figure S5 and Table S17). Several observations are noteworthy. (1) In general, for each gene, there are multiple rare missense and small indel (frameshift or stop-gain) variants contributing to the aggregate association signals, rather than a single strongly associated variant. (2) The patterns of phenotypic association are generally uni-directional and consistent with the biologic contribution of these genes to inherited RBC disorders: *HBA1* and *HBB* variants are associated with lower MCV/MCH, with *HBB* variants additionally associated with lower HCT/HGB and higher RBC/RDW, consistent with ineffective erythropoiesis and shortened red cell survival in alpha and beta thalassemia; *TMPRSS6* variants associated with lower MCH/MCV (Figures S5C16-19 and S5E13-14) and higher RDW (Figure S5G14), consistent with iron-refractory iron deficiency anemia. On the other hand, for *G6PD* rare variants, a bi-directional pattern of phenotypic association was observed for MCH, MCV, RBC, and RDW. (3) Several of the variants contributing to the *HBA1*, *HBB*, *TMPRSS6*, and *G6PD* signals are known to be pathogenic for inherited RBC disorders. Other variants that appear to contribute to the gene-based phenotypic effect are classified in ClinVar as variants of uncertain significance (VUSs) or have conflicting evidence to support their pathogenicity. (4) Three of the genes (*CD36*, *TFRC*, and *SLC12A7*) are located within regions of the genome containing common variants previously associated with RBC traits but have less clear relation to RBC biology. The presence of rare coding or LoF variants within these genes provides additional fine-mapping evidence that these three genes are causally responsible for RBC phenotypic variation.

pLoF and pKO variants associated with RBC traits

Predicted loss-of-function (pLoF) and predicted gene knockout (pKO) variants were examined in European, African, Hispanic, and Asian ancestry populations in TOPMed. The European ancestry population subset had the largest sample size and the largest number of both pLoF and pKO variants (Table S18). Two pLoF variants reached genome-wide significance, namely *CD36*-rs3211938 (chr7:80,671,133, GenBank: NM_000072.3, c.975T>G, GenBank: NP_000063.2, p.Tyr325Ter) for RDW in African participants and *HBB*-rs11549407 for multiple RBC traits in Hispanic and European participants (Table S19), which have been reported in previously published studies. No pKO variant reached genome-wide significance in any of the ancestral groups (Table S20). All pLoF and pKO variants with $p < 1E{-}4$ are presented in Tables S19 and S20.

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

In silico functional annotation of the *RPN1*-rs112097551 variant revealed a CADD-PHRED score of 20.4 and that the variant lies in a putative enhancer element bound by erythroid transcription factors GATA1 and TAL1. We therefore undertook additional experiments to investigate the causal gene underlying the association signal. First, we used cytosine base editing to modify the rs112097551 reference G to alternative A allele in HUDEP-2 erythroid precursor cells. Since there was no appropriately positioned NGG PAM motif, we utilized the recently described near-PAMless SpCas9 variant cytosine base editor AncBE4-max-SpRY,[41] achieving 33% G-to-A conversion efficiency (Figure 1A). Analysis of erythroblast promoter capture Hi-C datasets showed that the SNP interacts with *RUVBL1* (MIM: 603449) which is 500 kb upstream but not with intervening genes which include *RPN1* and the hematopoietic transcription factor *GATA2* (Figure 1B). In five G/A heterozygous HUDEP-2 clones compared to G/G clones, we observed significantly reduced expression of *RUVBL1* without significant change in expression of four more proximal genes *EEFSEC* (MIM: 607695), *GATA2*, *RPN1*, and *RAB7A* (MIM: 602298) (Figure 1C). Next, we performed SpCas9 nuclease editing to produce indels adjacent to rs112097551 in $CD34^+$ hematopoietic stem/progenitor cell (HSPC) derived primary erythroid precursors (Figures 1D and 1E). Cells bearing these short insertions and deletions centered 3 bp from the rs112097551 position demonstrated significantly reduced *RUVBL1* expression compared to control cells, while *RPN1* and *RAB7A* expression was unchanged (Figure 1F). Together, these base and nuclease editing results suggest that rs112097551-G contributes to a regulatory element that exerts long-range control of *RUVBL1* expression. Prior work has shown the mouse homolog of *RUVBL1* is required for murine hematopoiesis.[48] To test the role of *RUVBL1* in human hematopoiesis, we performed gene editing studies in $CD34^+$ HSPCs in which we targeted indels to coding sequences at *RUVBL1*. We observed 96.1% indels at *RUVBL1* compared to 84.2% indels in control cells targeted at a neutral locus. We infused edited HSPCs to immunodeficient NBSGW mice and analyzed bone marrow after 16 weeks for engrafting human hematopoietic chimerism and gene editing. Compared to $CD34^+$ HSPCs edited at a neutral locus which showed 91.6% mean human chimerism, human $CD34^+$ HSPCs edited at *RUVBL1* demonstrated only 7.7% mean chimerism (Figures 1G–1I). Engrafting human cells were marked by frequent gene edits (60.1%) when targeted at the neutral locus but only 4.8% gene edits after *RUVBL1*

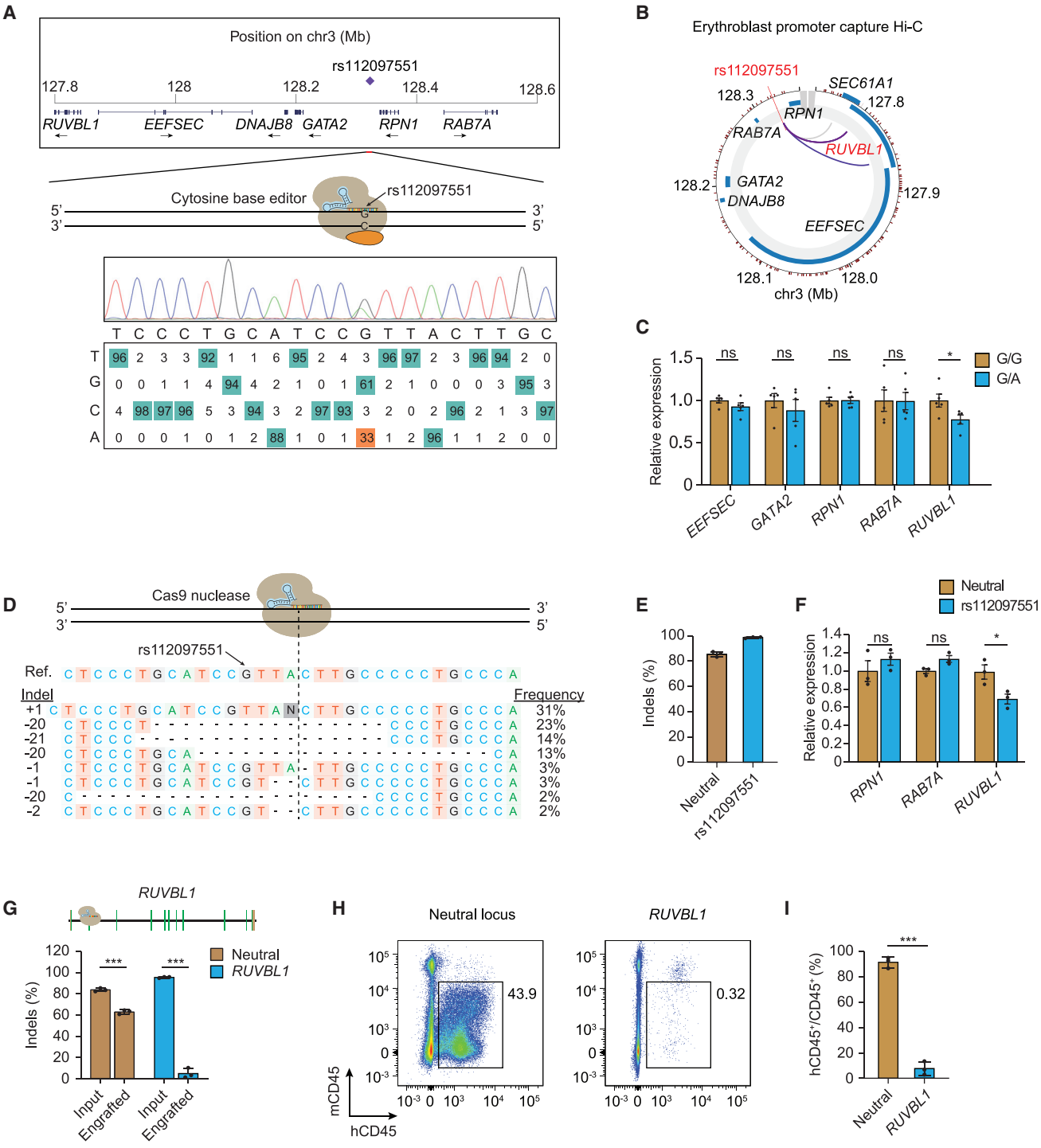


Figure 1. Gene editing implicates *RUVBL1* in rs112097551 association

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

(B) Promoter capture Hi-C from ChiCP analysis[49] of erythroblasts.[50]

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

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

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

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

(legend continued on next page)

editing, indicating that *RUVBL1* edited cells inefficiently engrafted. Together these results suggest rs112097551-G contributes to long-range enhancement of *RUVBL1* expression, which in turn supports human hematopoiesis.

Discussion

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

Our study highlights the benefits of increasing participant ethnic diversity and coverage of the genome in genetic association studies of complex polygenic traits. Among the 24 unique independent variants we identified in the single variant association analyses, 21 showed MAF $< 1\%$ in all TOPMed samples and 18 were monomorphic in at least one of the four major contributing ancestral populations in our analysis (European, African, East Asian, and Hispanic). These low-frequency or ancestry-specific variants were most likely missed by previous GWAS analysis using imputed genotype data or focusing on one ancestral population (Table S13).

GATA2-RPN1

Here we report and replicate a distinct low-frequency variant (MAF $= 0.4\%$ overall but considerably higher frequency among African [0.94%] than European [0.07%] ancestry individuals) associated with higher MCH and MCV in TOPMed (rs112097551). The region between *GATA2* and *RPN1* on chromosome 3q21 contains several common variants previously associated with various WBC-related traits in European, Asian, and Hispanic ancestry individuals and two variants previously associated with MCH and RDW in Europeans (rs2977562 and rs147412900).[13] *GATA2* is a hematopoietic transcription factor and heterozygous coding or enhancer mutations of *GATA2* are responsible for autosomal-dominant hereditary mononuclear cytopenia (MIM: 614172), immunodeficiency and myelodysplastic syndromes (MIM: 614286), as well as lymphatic dysfunction[51,52] (MIM: 137295). There was no evidence of association of the TOPMed MCH/MCV-associated rs112097551 variant with WBC-related traits in TOPMed (data not shown), though the variant was associated with higher monocyte count and percentage in Astle et al.,[9] but was not conditionally independent of other variants in the region. The MCV/MCH-associated rs112097551 variant lies in a putative enhancer element bound by erythroid transcription factors GATA-1 and TAL-1 and demonstrates physical interaction in erythroblasts with *RUVBL1* 500 kb away. Our results from gene editing of *RUVBL1* in primary human HPSCs and xenotransplantation suggest that *RUVBL1* plays a role in human hematopoiesis, consistent with data from mouse models suggesting that *RUVBL1* (which encodes the protein product pontin) to be essential for murine hematopoietic stem cell survival.[48] This finding also highlights the complexity and importance of experimentally validating the causal gene(s) underlying GWAS signals for complex traits, which are often assigned according to physical proximity (*RPN1*) or assumed on the basis of biologic function (*GATA2*).

ELL2

The chromosome *5q15* non-coding variant rs116635225 associated with lower MCH also has a low frequency in TOPMed (1.3%) and is considerably more common among African ancestry individuals (3.9%). The rs116635225 variant is located ~27 kb upstream of *ELL2*, a gene responsible for immunoglobulin mRNA production and transcriptional regulation in plasma cells. Coding and regulatory variants of *ELL2* have been associated with risk of multiple myeloma in European and African ancestry individuals as well as reduced levels of immunoglobulin A and G in healthy subjects.[53–55] Another set of genetic variants located ~200 kb away in the promoter region of *GLRX* or glutaredoxin-1 (rs10067881 [chr5:95,826,771, GenBank: NC_000005.10, g.95826771G>A], rs17462893 [chr5:95,827,733, GenBank: NC_000005.10, g.95827733A>G], rs57675369 [chr5:95,826,714, GenBank: NC_000005.10, g.95826714_95826715insG]) have been associated with

(G) Indel frequency following 3xNLS-SpCas9:sgRNA targeting *RUVBL1* coding sequence or neutral control locus in input cell 4 days after RNP electroporation or engrafted bone marrow samples 16 weeks after infusion to NBSGW mice.
(H) Representative flow cytometry of human and mouse $CD45^+$ cells from NBSGW bone marrow 16 weeks after cell infusion (representative of 3 mice).
(I) Mean human hematopoietic chimerism determined by $hCD45^+$/total $CD45^+$ cells from NBSGW bone marrow 16 weeks after cell infusion ($n = 3$ mice per group).
Student's t test (two-tailed test). ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$; ns, not significant. All error bars indicate mean and standard deviation.

higher reticulocyte count in UKBB Europeans.[9] Glutaredoxin-1 is a cytoplasmic enzyme that catalyzes the reversible reduction of glutathione-protein mixed disulfides and contributes to the antioxidant defense system. Congenital deficiencies of other members of the glutaredoxin enzyme family (*GLRX5* [MIM: 609588]) have been reported in patients with sideroblastic anemia (MIM: 300751).[56–58] Notably, our *ELL2* rs116635225 MCH-associated variant remained genome-wide significant after conditioning on the myeloma or reticulocyte-related variants. Therefore, the precise genetic regulatory mechanisms of the red cell trait associations in this region remain to be determined.

MIDN

The chromosome *19p13* African variant rs73494666 associated with lower MCV/MCH is located in an open chromatin region of an intron of *MIDN*, which encodes the midbrain nucleolar protein midnolin. The gene-rich region on chromosome *19p13* also includes *SBNO2* (MIM: 615729), *STK11* (MIM: 602216), *CBARP*, *ATP5F1D* (MIM: 603150), *CIRBP* (MIM: 602649), *EFNA2* (MIM: 602756), and *GPX4* (MIM: 138322). However, none of these genes have clear relationships to hematopoiesis or red structure/function. Other variants in the region have been associated with MCH and RBC count (rs757293, chr19:1,277,428, GenBank: NC_000019.10, g.1277428T>C)[13] or reticulocytes (rs35971149, chr19:1,164,199, GenBank: NC_000019.10, g.1164199del).[9] The *MIDN*- rs73494666 variant overlaps ENCODE *cis*-regulatory elements for CD34 stem cells and other blood cell progenitors.

PIEZO1

Mutations in the mechanosensitive ion channel *PIEZO1* on chromosome *16q24* have been reported in patients with autosomal-dominant hereditary xerocytosis (MIM: 194380), a congenital hemolytic anemia associated with increased calcium influx, red cell dehydration, and potassium efflux along with various red cell laboratory abnormalities including increased MCHC, MCH, and reticulocytosis.[59,60] Most reported hereditary xerocytosis *PIEZO1* missense mutations are associated with at least partial gain-of-function and are located within the highly conserved C-terminal region near the pore of the ion channel. In some individuals carrying *PIEZO1* missense mutations, mild red cell laboratory parameter alterations without frank hemolytic anemia have been reported.[61] The *PIEZO1* 3 bp short tandem repeat (STR) rs763477215 in-frame coding variant (p.Lys2169del) associated with higher MCHC in TOPMed is extremely rare in all populations except for the Ashkenazi Jewish population (frequency of 1.5% in gnomAD), has not been previously associated with hereditary xerocytosis, and therefore has been reported as "benign" in ClinVar. The p.Lys2169del variant is located in a highly basic -Lys-Lys-Lys-Lys- motif near the C terminus of the 36 transmembrane domain protein within a 14-residue linker region between the central ion channel pore and the peripheral propeller-like mechanosensitive domains important for modulating PIEZO1 channel function.[62,63] Interestingly, another 3 bp in-frame deletion of *PIEZO1* (E756del) reported to be highly enriched in prevalence among African populations was recently associated with dehydrated red blood cells and reduced susceptibility to malaria.[64,65] In TOPMed, however, we were unable to confirm any association between the rs59446030 (chr16:88,733,965, GenBank: NM_001142864.4, c.2247_2249GGA[7], GenBank: NP_001136336.2, p.Glu756del) putative malaria-susceptibility allele variant and phenotypic variation in MCHC (p value for trait-specific conditional analysis = 0.42).

TMPRSS6

TMPRSS6 on chromosome *22q12* encodes matriptase-2, a transmembrane serine protease that downregulates the production of hepcidin in the liver and therefore plays an essential role in iron homeostasis.[66] Rare mutations of *TMPRSS6* are associated with iron-refractory iron deficiency anemia (MIM: 206200)[67] characterized by microcytic hypochromic anemia and low transferrin saturation. Several common *TMPRSS6* variants have been associated with multiple RBC traits through prior GWASs. The common *TMPRSS6* intronic variant associated with *TMPRSS6* expression and lower MCH/MCV in TOPMed (rs228914/rs228916) was previously reported to be associated with lower iron levels,[47] and therefore likely contributes to lower MCH and MCV via iron deficiency. In rare variant aggregated association testing, we were able to identify several additional rare coding missense, stop-gain, or splice variants that appear to drive the gene-based association of *TMPRSS6* with lower MCH/MCV and higher RDW. At least one of these variants at exon 13 rs387907018 (chr22:37,073,550, GenBank: NC_000022.11, g.37073550C>T, GenBank: NP_705837.1, p.Glu522Lys, missense mutation) has been reported in a compound heterozygous iron-refractory iron deficiency anemia (IRIDA [MIM: 206200]) patient,[68] suggesting that inheritance of this or similar LoF variants in the heterozygote state may contribute to mild reductions in MCV/MCH or increased RDW.[67]

G6PD

X-linked *G6PD* mutations (glucose-6-phosphate dehydrogenase) are the most common cause worldwide of acute and chronic hemolytic anemia. The *G6PD*-rs76723693 low-frequency missense variant (p.Leu323Pro, referred to as *G6PD* Nefza[69]) is common in persons of African ancestry and is associated with lower RDW in TOPMed. In persons of African ancestry, the p.Leu323Pro variant is often co-inherited with another *G6PD* missense variant, p.Asn126Asp, encoded by rs1050829 (chrX:154,535,277, GenBank: NC_000023.11, g.154535277T>C, GenBank: NP_001346945.1, p.Asn126Asp). The 968C/376G haplotype in African ancestry individuals constitutes one of several forms of the *G6PD* variant A-.[70–73] Functional studies of the p.Leu323Pro, p.Asn126Asp, and the double

mutant suggest the p.Leu323Pro variant is the primary contributor to reduced catalytic activity.[74] In the US, another African ancestry *G6PD* A- variant is due to the haplotypic combination of rs1050829 and rs1050828 (chrX:154,536,002, GenBank: NC_000023.11, g.154536002C>T, GenBank: NP_001346945.1, p.Val68Met), which has an allele frequency of ~12%. Our finding that rs76723693 is significantly associated with lower RDW after conditioning on rs1050828 is consistent with the independence of effects of the *G6PD* Nefza and A- variants on red cell physiology and morphology. Importantly, both rs76723693 and rs1050828 *G6PD* variants were recently reported to have the effect of lowering hemoglobin A1c (HbA1c) values and therefore should be considered when screening African Americans for type 2 diabetes (MIM: 125853).[75]

In gene-based analyses, several additional *G6PD* missense variants contributed to the aggregated rare variant association signals for MCH, MCV, RBC, and RDW, including the class II Southeast Asian Mahidol variant p.Gly163Ser (rs730880992, chr12:112,453,349, GenBank: NC_000012.12, g.112453349G>A, GenBank: NP_002825.3, p.Gly163Cys)[76] and the class II Union variant p.Arg454Cys (rs398123546, chrX:154,532,390, GenBank: NC_000023.11, g.154532390G>A, GenBank: NP_001035810.1, p.Arg454Cys).[77] For a third previously reported variant associated with G6PD deficiency, the East Asian class II Gahoe variant p.His32Arg (rs137852340, chrX: 154,546,061, GenBank: NM_001360016.2, c.95A>G, GenBank: NP_001346945.1, p.His32Arg),[78] there is conflicting evidence of pathogenicity in ClinVar. Of the two female rs137852340 variant allele carriers in TOPMed, one has a normal RDW and one has an elevated RDW. These findings add to the further genotypic-phenotypic complexity and clinical spectrum of G6PD deficiency, which is influenced by its sex-linkage and zygosity, residual *G6PD* variant enzyme activity and stability, genetic background, and environmental exposures.[79]

HBB

Heterozygosity for the common African *HBB*-rs334 hemoglobin S (chr11:5,227,002, GenBank: NC_000011.10, g.5227002T>A, GenBank: NP_000509.1, p.Glu7Val) or rs33930165 hemoglobin C (chr11:5,227,003, GenBank: NC_000011.10, g.5227003C>T, GenBank: NP_000509.1, p.Glu7Lys) beta-globin structural variants have recently been associated with alterations in various red cell laboratory parameters including lower hemoglobin, MCV, MCH, and RDW, along with higher MCHC, RDW, and HbA1c.[17,18,20,80–82] In TOPMed, we were able to identify at least ten additional low-frequency or rare variants within the *HBB* locus independently associated with HGB, RBC, MCV, MCH, MCHC, and/or RDW. Notably, six of the ten variants correspond to *HBB* 5′ UTR and promoter regions previously identified in patients with beta-thalassemia: rs34598529 (chr11:5,227,100, GenBank: NC_000011.10, g.5227100T>C or −29A>G);[83] rs33944208 (chr11:5,227,159, GenBank: NC_000011.10, g.5227159G>A or −88C>T);[84–86] splice site rs33915217 (chr11:5,226,925, GenBank: NC_000011.10, g.5226925C>G or IVS1-5G>C);[84,87] rs33945777 (chr11:5,226,576, GenBank: NC_000011.10, g.5226576C>T or IVS2-1G>A);[84,87] rs35004220 (chr11:5,226,820, GenBank: NC_000011.10, g.5226820C>T or IVS-I-110 G->A),[88,89] and nonsense mutations rs11549407 (chr11:5,226,774, GenBank: NC_000011.10, g.5226774G>T, GenBank: NP_000509.1, p.Gln40Lys or p.Gln40Ter).[90,91] These findings confirm the very mild phenotype and clinically "silent" nature of the heterozygote carrier state of these beta-globin gene variants.[92] Several of these mutations occur more commonly in populations of South Asian (rs33915217), African (rs34598529, rs33944208), or Mediterranean (rs11549407) ancestry. Four additional association signals in the region—rs73404549 (*HBG2*, chr11:5,299,424, GenBank: NC_000011.10, g.5299424C>T), rs77333754 (chr11:5,001,853, GenBank: NC_000011.10, g.5001853T>C), rs1189661759 (chr11:5,183,128, GenBank: NC_000011.10, g.5183128C>A), and rs539384429 (chr11:5,106,319, GenBank: NC_000011.10, g.5106319A>G)—are all rare non-coding variants without obvious functional consequences. In addition to the *HBB* protein-coding variants identified in single-variant analyses, several of the rare variants driving the aggregate *HBB* gene-based association with lower HGB/HCT and MCH/MCV/MCHC and higher RBC/RDW are similarly previously reported missense, frameshift, or nonsense mutations previously identified in beta-thalassemia patients and categorized as pathogenic in ClinVar (Figure S5 and Table S17).

HBA1/HBA2

Several common DNA polymorphisms located in the alpha-globin gene cluster on chromosome 16p13.3 have been associated with red cell traits in large GWASs,[7,8,93] including heterozygosity for the common African ancestral 3.7 kb deletion which contributes to quantitative RBC phenotypes among African Americans and Hispanics/Latinos. In TOPMed, we identified two low-frequency variants in single-variant testing associated with MCH, MCV, and/or RBC count, independently of the 3.7 kb deletion. The rs868351380 variant is found primarily among Hispanics/Latinos while the rs372755452 variant is found primarily among East Asians. Neither of these two non-coding variants is located in any known alpha-globin regulatory region, and therefore requires further mechanistic confirmation. By contrast, in gene-based rare variant analysis, we identified several known alpha-globin variants associated in aggregate with lower MCH and MCV including the South Asian variant Hb Q India (*HBA1*, rs33984024, chr16:177,026, GenBank: NM_000558.5, c.193G>C, GenBank: NP_000549.1, p.Asp65His)[94–96] and the African variant Hb Groene Hart (*HBA1*, rs63750751, chr16:177,340, GenBank: NM_000558.5, c.358C>T, GenBank: NP_000549.1, p.Pro120Ser).[97–99] In homozygous or

compound heterozygous forms, these latter variants have been reported in probands with alpha-thalassemia, whereas heterozygotes generally have mild microcytic phenotype. Several additional variants contributing to the *HBA1* gene-based rare variant MCH/MCV signal (e.g., a 1 bp indel causing frameshift p.Asn79Ter, rs767911847, chr16:177,070, GenBank: NM_000558.5, c.237del, GenBank: NP_000549.1, p.Asn79fs) may represent previously undetected alpha-thalassemia mutations.

CD36, *TFRC*, and *SLC12A7*

The presence of rare coding or LoF variants within *CD36*, *TFRC*, and *SLC12A7* provides evidence that these genes are causally responsible for RBC phenotypic variation. A common African ancestral null variant of CD36 (rs3211938 or p.Tyr325Ter) has been previously associated with higher RDW and with lower CD36 expression in erythroblasts.[100] In TOPMed, additional *CD36* rare coding variants were associated in aggregate with higher RDW independent of rs3211938, including several nonsense and frameshift or splice acceptor mutations, which have been previously classified as VUSs. Further characterization of the genetic complexity of the CD36-null phenotype (common in African and Asian populations) may provide information relevant to the tissue-specific expression of this receptor on red cells, platelets, monocytes, and endothelial cells and its role in malaria infection and disease severity.[101] *TFRC* encodes the transferrin receptors (TfR1), which is required for iron uptake and erythropoiesis.[102] While common non-coding variants of *TFRC* have been associated with MCV and RDW, the only known *TFRC*-related Mendelian disorder is a homozygous p.Tyr20His (rs863225436, chr3:196,075,339, GenBank: NM_001128148.3, c.58T>C, GenBank: NP_001121620.1, p.Tyr20His) substitution reported to cause combined immunodeficiency affecting leukocytes and platelets but not red cells.[103] Common variants of *SLC12A7* encoding the potassium ion channel KCC4 have been associated with RDW and other RBC phenotypes. While KCC4 is expressed in erythroblasts,[104] its role in red blood cell function is not well described.[105] Further characterization of KCC4 LoF variants may illuminate the role of this ion transporter in red cell dehydration with potential implications for treatment of patients with sickle cell disease.[106]

In summary, we illustrate that expanding coverage of the genome using WGS as applied to large, population-based multi-ethnic samples can lead to discovery of variants associated with quantitative RBC traits that have not been described before. Most of the discovered variants were of low frequency and/or disproportionately observed in non-Europeans. We also report extensive allelic heterogeneity at the chromosome 11 beta-globin locus, including associations with several known beta-thalassemia carrier variants. The gene-based association of rare variants within *HBA1/HBA2*, *HBB*, *TMPRSS6*, *G6PD*, *CD36*, *TFRC*, and *SLC12A7* independent of known single variants in the same genes further suggest that rare functional variants in genes responsible for Mendelian RBC disorders contribute to the genetic architecture of RBC phenotypic variation among the population at large. Together these results demonstrate the utility of WGS in ethnically diverse population-based samples for expanding our understanding of the genetic architecture of quantitative hematologic traits and suggest a continuum between complex traits and Mendelian red cell disorders.

Data and code availability

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

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.04.003.

Consortia

The members of the NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium are Namiko Abe, Goncalo Abecasis, Francois Aguet, Christine Albert, Laura Almasy, Alvaro Alonso, Seth Ament, Peter Anderson, Pramod Anugu, Deborah Applebaum-Bowden, Kristin Ardlie, Dan Arking, Donna K Arnett, Allison Ashley-Koch, Stella Aslibekyan, Tim Assimes, Paul Auer, Dimitrios Avramopoulos, Najib Ayas, Adithya Balasubramanian, John Barnard, Kathleen Barnes, R. Graham Barr, Emily Barron-Casella, Lucas Barwick, Terri Beaty, Gerald Beck, Diane Becker, Lewis Becker, Rebecca Beer, Amber Beitelshees, Emelia Benjamin, Takis Benos, Marcos Bezerra, Larry Bielak, Joshua Bis, Thomas Blackwell, John Blangero, Eric Boerwinkle, Donald W. Bowden, Russell Bowler, Jennifer Brody, Ulrich Broeckel, Jai Broome, Deborah Brown, Karen Bunting, Esteban Burchard, Carlos Bustamante, Erin Buth, Brian Cade, Jonathan Cardwell, Vincent Carey, Julie Carrier, Cara Carty, Richard Casaburi, Juan P Casas Romero, James Casella, Peter Castaldi, Mark Chaffin, Christy Chang, Yi-Cheng Chang, Daniel Chasman, Sameer Chavan, Bo-Juen Chen, Wei-Min Chen, Yii-Der Ida Chen, Michael Cho, Seung Hoan Choi, Lee-Ming Chuang, Mina Chung, Ren-Hua Chung, Clary Clish, Suzy Comhair, Matthew Conomos, Elaine Cornell, Adolfo Correa, Carolyn Crandall, James Crapo, L. Adrienne Cupples, Joanne Curran, Jeffrey Curtis, Brian Custer, Coleen Damcott, Dawood Darbar, Sean David, Colleen Davis, Michelle Daya, Mariza de Andrade, Lisa de las Fuentes, Paul de Vries, Michael DeBaun, Ranjan Deka, Dawn DeMeo, Scott Devine, Huyen Dinh, Harsha Doddapaneni, Qing Duan, Shannon Dugan-Perez, Ravi Duggirala, Jon Peter Durda, Susan K. Dutcher, Charles Eaton, Lynette Ekunwe, Adel El Boueiz, Patrick Ellinor, Leslie Emery, Serpil Erzurum, Charles Farber, Jesse Farek, Tasha Fingerlin, Matthew Flickinger, Myriam Fornage, Nora Franceschini, Chris Frazar, Mao Fu, Stephanie M. Fullerton, Lucinda Fulton, Stacey Gabriel, Weiniu Gan, Shanshan Gao, Yan Gao, Margery Gass, Heather Geiger, Bruce Gelb, Mark Geraci,

Soren Germer, Robert Gerszten, Auyon Ghosh, Richard Gibbs, Chris Gignoux, Mark Gladwin, David Glahn, Stephanie Gogarten, Da-Wei Gong, Harald Goring, Sharon Graw, Kathryn J. Gray, Daniel Grine, Colin Gross, C. Charles Gu, Yue Guan, Xiuqing Guo, Namrata Gupta, David M. Haas, Jeff Haessler, Michael Hall, Yi Han, Patrick Hanly, Daniel Harris, Nicola L. Hawley, Jiang He, Ben Heavner, Susan Heckbert, Ryan Hernandez, David Herrington, Craig Hersh, Bertha Hidalgo, James Hixson, Brian Hobbs, John Hokanson, Elliott Hong, Karin Hoth, Chao (Agnes) Hsiung, Jianhong Hu, Yi-Jen Hung, Haley Huston, Chii Min Hwu, Marguerite Ryan Irvin, Rebecca Jackson, Deepti Jain, Cashell Jaquish, Jill Johnsen, Andrew Johnson, Craig Johnson, Rich Johnston, Kimberly Jones, Hyun Min Kang, Robert Kaplan, Sharon Kardia, Shannon Kelly, Eimear Kenny, Michael Kessler, Alyna Khan, Ziad Khan, Wonji Kim, John Kimoff, Greg Kinney, Barbara Konkle, Charles Kooperberg, Holly Kramer, Christoph Lange, Ethan Lange, Leslie Lange, Cathy Laurie, Cecelia Laurie, Meryl LeBoff, Jiwon Lee, Sandra Lee, Wen-Jane Lee, Jonathon LeFaive, David Levine, Dan Levy, Joshua Lewis, Xiaohui Li, Yun Li, Henry Lin, Honghuang Lin, Xihong Lin, Simin Liu, Yongmei Liu, Yu Liu, Ruth J.F. Loos, Steven Lubitz, Kathryn Lunetta, James Luo, Ulysses Magalang, Michael Mahaney, Barry Make, Ani Manichaikul, Alisa Manning, JoAnn Manson, Lisa Martin, Melissa Marton, Susan Mathai, Rasika Mathias, Susanne May, Patrick McArdle, Merry-Lynn McDonald, Sean McFarland, Stephen McGarvey, Daniel McGoldrick, Caitlin McHugh, Becky McNeil, Hao Mei, James Meigs, Vipin Menon, Luisa Mestroni, Ginger Metcalf, Deborah A Meyers, Emmanuel Mignot, Julie Mikulla, Nancy Min, Mollie Minear, Ryan L Minster, Braxton D. Mitchell, Matt Moll, Zeineen Momin, May E. Montasser, Courtney Montgomery, Donna Muzny, Josyf C Mychaleckyj, Girish Nadkarni, Rakhi Naik, Take Naseri, Pradeep Natarajan, Sergei Nekhai, Sarah C. Nelson, Bonnie Neltner, Caitlin Nessner, Deborah Nickerson, Osuji Nkechinyere, Kari North, Jeff O'Connell, Tim O'Connor, Heather Ochs-Balcom, Geoffrey Okwuonu, Allan Pack, David T. Paik, Nicholette Palmer, James Pankow, George Papanicolaou, Cora Parker, Gina Peloso, Juan Manuel Peralta, Marco Perez, James Perry, Ulrike Peters, Patricia Peyser, Lawrence S Phillips, Jacob Pleiness, Toni Pollin, Wendy Post, Julia Powers Becker, Meher Preethi Boorgula, Michael Preuss, Bruce Psaty, Pankaj Qasba, Dandi Qiao, Zhaohui Qin, Nicholas Rafaels, Laura Raffield, Mahitha Rajendran, Vasan S. Ramachandran, D.C. Rao, Laura Rasmussen-Torvik, Aakrosh Ratan, Susan Redline, Robert Reed, Catherine Reeves, Elizabeth Regan, Alex Reiner, Muagututi'a Sefuiva Reupena, Ken Rice, Stephen Rich, Rebecca Robillard, Nicolas Robine, Dan Roden, Carolina Roselli, Jerome Rotter, Ingo Ruczinski, Alexi Runnels, Pamela Russell, Sarah Ruuska, Kathleen Ryan, Ester Cerdeira Sabino, Danish Saleheen, Shabnam Salimi, Sejal Salvi, Steven Salzberg, Kevin Sandow, Vijay G. Sankaran, Jireh Santibanez, Karen Schwander, David Schwartz, Frank Sciurba, Christine Seidman, Jonathan Seidman, Frederic Sériès, Vivien Sheehan, Stephanie L. Sherman, Amol Shetty, Aniket Shetty, Wayne Hui-Heng Sheu, M. Benjamin Shoemaker, Brian Silver, Edwin Silverman, Robert Skomro, Albert Vernon Smith, Jennifer Smith, Josh Smith, Nicholas Smith, Tanja Smith, Sylvia Smoller, Beverly Snively, Michael Snyder, Tamar Sofer, Nona Sotoodehnia, Adrienne M. Stilp, Garrett Storm, Elizabeth Streeten, Jessica Lasky Su, Yun Ju Sung, Jody Sylvia, Adam Szpiro, Daniel Taliun, Hua Tang, Margaret Taub, Kent D. Taylor, Matthew Taylor, Simeon Taylor, Marilyn Telen, Timothy A. Thornton, Machiko Threlkeld, Lesley Tinker, David Tirschwell, Sarah Tishkoff, Hemant Tiwari, Catherine Tong, Russell Tracy, Michael Tsai, Dhananjay Vaidya, David Van Den Berg, Peter VandeHaar, Scott Vrieze, Tarik Walker, Robert Wallace, Avram Walts, Fei Fei Wang, Heming Wang, Jiongming Wang, Karol Watson, Jennifer Watt, Daniel E. Weeks, Bruce Weir, Scott T Weiss, Lu-Chen Weng, Jennifer Wessel, Cristen Willer, Kayleen Williams, L. Keoki Williams, Carla Wilson, James Wilson, Lara Winterkorn, Quenna Wong, Joseph Wu, Huichun Xu, Lisa Yanek, Ivana Yang, Ketian Yu, Seyedeh Maryam Zekavat, Yingze Zhang, Snow Xueyan Zhao, Wei Zhao, Xiaofeng Zhu, Michael Zody, and Sebastian Zoellner.

Declaration of interests

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

Received: December 19, 2020
Accepted: March 30, 2021
Published: April 21, 2021; corrected online April 28, 2021

Web resources

gnomAD, https://gnomad.broadinstitute.org
INTERVAL study, https://www.intervalstudy.org.uk/
KAISER-Permanente Genetic Epidemiology Research on Aging (GERA) cohort, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000674.v3.p3/
OMIM, https://www.omim.org/
TOPMed whole genome sequencing methods for freeze 8, https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-methods-freeze-8

References

1. Kuhn, V., Diederich, L., Keller, T.C.S., 4th, Kramer, C.M., Lückstädt, W., Panknin, C., Suvorava, T., Isakson, B.E., Kelm, M., and Cortese-Krott, M.M. (2017). Red blood cell function and dysfunction: redox regulation, nitric oxide metabolism, anemia. Antioxid. Redox Signal. *26*, 718–742.
2. Sarma, P.R. (1990). Red Cell Indices. In Clinical Methods: The History, Physical, and Laboratory Examinations, H.K. Walker, W.D. Hall, and J.W. Hurst, eds. (Boston: Butterworths).
3. Lippi, G., and Mattiuzzi, C. (2020). Updated worldwide epidemiology of inherited erythrocyte disorders. Acta Haematol. *143*, 196–203.
4. Evans, D.M., Frazer, I.H., and Martin, N.G. (1999). Genetic and environmental causes of variation in basal levels of blood cells. Twin Res. *2*, 250–257.
5. Patel, K.V. (2008). Variability and heritability of hemoglobin concentration: an opportunity to improve understanding of anemia in older adults. Haematologica *93*, 1281–1283.
6. Soranzo, N., Spector, T.D., Mangino, M., Kühnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M., et al. (2009). A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat. Genet. *41*, 1182–1190.
7. Ganesh, S.K., Zakai, N.A., van Rooij, F.J.A., Soranzo, N., Smith, A.V., Nalls, M.A., Chen, M.-H., Kottgen, A., Glazer, N.L., Dehghan, A., et al. (2009). Multiple loci influence

erythrocyte phenotypes in the CHARGE Consortium. Nat. Genet. *41*, 1191–1198.

8. van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Verweij, N., Sehmi, J., Paul, D.S., Elling, U., Allayee, H., Li, X., et al. (2012). Seventy-five genetic loci influencing the human red blood cell. Nature *492*, 369–375.
9. Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A., et al. (2016). The allelic landscape of human blood cell trait variation and links to common complex disease. Cell *167*, 1415–1429.e19.
10. Iotchkova, V., Huang, J., Morris, J.A., Jain, D., Barbieri, C., Walter, K., Min, J.L., Chen, L., Astle, W., Cocca, M., et al.; UK10K Consortium (2016). Discovery and refinement of genetic loci associated with cardiometabolic risk using dense imputation maps. Nat. Genet. *48*, 1303–1312.
11. CHARGE Consortium Hematology Working Group (2016). Meta-analysis of rare and common exome chip variants identifies S1PR4 and other loci influencing blood cell traits. Nat. Genet. *48*, 867–876.
12. Mousas, A., Ntritsos, G., Chen, M.-H., Song, C., Huffman, J.E., Tzoulaki, I., Elliott, P., Psaty, B.M., Auer, P.L., Johnson, A.D., et al.; Blood-Cell Consortium (2017). Rare coding variants pinpoint genes that control human hematological traits. PLoS Genet. *13*, e1006925.
13. Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M.K., Schoech, A., Pasaniuc, B., and Price, A.L. (2019). Leveraging polygenic functional enrichment to improve GWAS power. Am. J. Hum. Genet. *104*, 65–75.
14. van Rooij, F.J.A., Qayyum, R., Smith, A.V., Zhou, Y., Trompet, S., Tanaka, T., Keller, M.F., Chang, L.-C., Schmidt, H., Yang, M.-L., et al.; BioBank Japan Project (2017). Genome-wide Trans-ethnic Meta-analysis Identifies Seven Genetic Loci Influencing Erythrocyte Traits and a Role for RBPMS in Erythropoiesis. Am. J. Hum. Genet. *100*, 51–63.
15. Jo Hodonsky, C., Schurmann, C., Schick, U.M., Kocarnik, J., Tao, R., van Rooij, F.J., Wassel, C., Buyske, S., Fornage, M., Hindorff, L.A., et al. (2018). Generalization and fine mapping of red blood cell trait genetic associations to multi-ethnic populations: The PAGE Study. Am. J. Hematol. Published online June 15, 2018. https://doi.org/10.1002/ajh.25161.
16. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. *50*, 390–400.
17. Gurdasani, D., Carstensen, T., Fatumo, S., Chen, G., Franklin, C.S., Prado-Martinez, J., Bouman, H., Abascal, F., Haber, M., Tachmazidou, I., et al. (2019). Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. Cell *179*, 984–1002.e36.
18. Raffield, L.M., Ulirsch, J.C., Naik, R.P., Lessard, S., Handsaker, R.E., Jain, D., Kang, H.M., Pankratz, N., Auer, P.L., Bao, E.L., et al.; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Hematology & Hemostasis, Diabetes, and Structural Variation TOPMed Working Groups (2018). Common α-globin variants modify hematologic and other clinical phenotypes in sickle cell trait and disease. PLoS Genet. *14*, e1007293.
19. Vuckovic, D., Bao, E.L., Akbari, P., Lareau, C.A., Mousas, A., Jiang, T., Chen, M.-H., Raffield, L.M., Tardaguila, M., Huffman, J.E., et al.; VA Million Veteran Program (2020). The polygenic and monogenic basis of blood traits and diseases. Cell *182*, 1214–1231.e11.
20. Hodonsky, C.J., Jain, D., Schick, U.M., Morrison, J.V., Brown, L., McHugh, C.P., Schurmann, C., Chen, D.D., Liu, Y.M., Auer, P.L., et al. (2017). Genome-wide association study of red blood cell traits in Hispanics/Latinos: The Hispanic Community Health Study/Study of Latinos. PLoS Genet. *13*, e1006760.
21. Chen, M.-H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., et al.; VA Million Veteran Program (2020). Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. Cell *182*, 1198–1213.e14.
22. Beutler, E., and West, C. (2005). Hematologic differences between African-Americans and whites: the roles of iron deficiency and alpha-thalassemia on hemoglobin levels and mean corpuscular volume. Blood *106*, 740–745.
23. Regier, A.A., Farjoun, Y., Larson, D.E., Krasheninina, O., Kang, H.M., Howrigan, D.P., Chen, B.-J., Kher, M., Banks, E., Ames, D.C., et al. (2018). Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat. Commun. *9*, 4038.
24. Conomos, M.P., Miller, M.B., and Thornton, T.A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet. Epidemiol. *39*, 276–293.
25. Conomos, M.P., Reiner, A.P., Weir, B.S., and Thornton, T.A. (2016). Model-free Estimation of Recent Genetic Relatedness. Am. J. Hum. Genet. *98*, 127–148.
26. Conomos, M.P., Laurie, C.A., Stilp, A.M., Gogarten, S.M., McHugh, C.P., Nelson, S.C., Sofer, T., Fernández-Rhodes, L., Justice, A.E., Graff, M., et al. (2016). Genetic diversity and association studies in US hispanic/latino populations: applications in the hispanic community health study/study of latinos. Am. J. Hum. Genet. *98*, 165–184.
27. Sofer, T., Zheng, X., Gogarten, S.M., Laurie, C.A., Grinde, K., Shaffer, J.R., Shungin, D., O'Connell, J.R., Durazo-Arvizo, R.A., Raffield, L., et al.; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium (2019). A fully adjusted two-stage procedure for rank-normalization in genetic association studies. Genet. Epidemiol. *43*, 263–275.
28. Lin, D.-Y. (2019). A simple and accurate method to determine genomewide significance for association tests in sequencing studies. Genet. Epidemiol. *43*, 365–372.
29. Gogarten, S.M., Sofer, T., Chen, H., Yu, C., Brody, J.A., Thornton, T.A., Rice, K.M., and Conomos, M.P. (2019). Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics *35*, 5346–5348.
30. Pedersen, B.S., and Quinlan, A.R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics *34*, 867–868.
31. Reiner, A.P., Lettre, G., Nalls, M.A., Ganesh, S.K., Mathias, R., Austin, M.A., Dean, E., Arepalli, S., Britton, A., Chen, Z., et al. (2011). Genome-wide association study of white blood cell count in 16,388 African Americans: the continental origins and genetic epidemiology network (COGENT). PLoS Genet. *7*, e1002108.
32. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191.
33. Chen, H., Huffman, J.E., Brody, J.A., Wang, C., Lee, S., Li, Z., Gogarten, S.M., Sofer, T., Bielak, L.F., Bis, J.C., et al.; NHLBI Trans-

Omics for Precision Medicine (TOPMed) Consortium; and TOPMed Hematology and Hemostasis Working Group (2019). Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. Am. J. Hum. Genet. *104*, 260–274.

34. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.
35. Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Christiani, D.C., Wurfel, M.M., Lin, X.; and NHLBI GO Exome Sequencing Project—ESP Lung Project Team (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. *91*, 224–237.
36. Cummings, B.B., Karczewski, K.J., Kosmicki, J.A., Seaby, E.G., Watts, N.A., Singer-Berk, M., Mudge, J.M., Karjalainen, J., Satterstrom, F.K., O'Donnell-Luria, A.H., et al.; Genome Aggregation Database Production Team; and Genome Aggregation Database Consortium (2020). Transcript expression-aware annotation improves rare variant interpretation. Nature *581*, 452–458.
37. Lindeboom, R.G.H., Supek, F., and Lehner, B. (2016). The rules and impact of nonsense-mediated mRNA decay in human cancers. Nat. Genet. *48*, 1112–1118.
38. Lessard, S., Manning, A.K., Low-Kam, C., Auer, P.L., Giri, A., Graff, M., Schurmann, C., Yaghootkar, H., Luan, J., Esko, T., et al.; NHLBI GO Exome Sequence Project; GOT2D; T2D-GENES; and GIANT Consortium (2016). Testing the role of predicted gene knockouts in human anthropometric trait variation. Hum. Mol. Genet. *25*, 2082–2092.
39. Kurita, R., Suda, N., Sudo, K., Miharada, K., Hiroyama, T., Miyoshi, H., Tani, K., and Nakamura, Y. (2013). Establishment of immortalized human erythroid progenitor cell lines able to produce enucleated red blood cells. PLoS ONE *8*, e59890.
40. Vinjamur, D.S., and Bauer, D.E. (2018). Growing and Genetically Manipulating Human Umbilical Cord Blood-Derived Erythroid Progenitor (HUDEP) Cell Lines. Methods Mol. Biol. *1698*, 275–284.
41. Walton, R.T., Christie, K.A., Whittaker, M.N., and Kleinstiver, B.P. (2020). Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. Science *368*, 290–296.
42. Kluesner, M.G., Nedveck, D.A., Lahr, W.S., Garbe, J.R., Abrahante, J.E., Webber, B.R., and Moriarity, B.S. (2018). EditR: A Method to Quantify Base Editing from Sanger Sequencing. CRISPR J *1*, 239–250.
43. Wu, Y., Zeng, J., Roscoe, B.P., Liu, P., Yao, Q., Lazzarotto, C.R., Clement, K., Cole, M.A., Luk, K., Baricordi, C., et al. (2019). Highly efficient therapeutic gene editing of human hematopoietic stem cells. Nat. Med. *25*, 776–783.
44. Giarratana, M.-C., Rouard, H., Dumont, A., Kiger, L., Safeukui, I., Le Pennec, P.-Y., François, S., Trugnan, G., Peyrard, T., Marie, T., et al. (2011). Proof of principle for transfusion of in vitro-generated red blood cells. Blood *118*, 5071–5079.
45. Brinkman, E.K., Chen, T., Amendola, M., and van Steensel, B. (2014). Easy quantitative assessment of genome editing by sequence trace decomposition. Nucleic Acids Res. *42*, e168.
46. Kowalski, M.H., Qian, H., Hou, Z., Rosen, J.D., Tapia, A.L., Shan, Y., Jain, D., Argos, M., Arnett, D.K., Avery, C., et al.; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium; and TOPMed Hematology & Hemostasis Working Group (2019). Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. PLoS Genet. *15*, e1008500.
47. Benyamin, B., Esko, T., Ried, J.S., Radhakrishnan, A., Vermeulen, S.H., Traglia, M., Gögele, M., Anderson, D., Broer, L., Podmore, C., et al.; InterAct Consortium (2014). Novel loci affecting iron homeostasis and their effects in individuals at risk for hemochromatosis. Nat. Commun. *5*, 4926.
48. Bereshchenko, O., Mancini, E., Luciani, L., Gambardella, A., Riccardi, C., and Nerlov, C. (2012). Pontin is essential for murine hematopoietic stem cell survival. Haematologica *97*, 1291–1294.
49. Schofield, E.C., Carver, T., Achuthan, P., Freire-Pritchett, P., Spivakov, M., Todd, J.A., and Burren, O.S. (2016). CHiCP: a web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets. Bioinformatics *32*, 2511–2513.
50. Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., et al.; BLUEPRINT Consortium (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell *167*, 1369–1384.e19.
51. Crispino, J.D., and Horwitz, M.S. (2017). GATA factor mutations in hematologic disease. Blood *129*, 2103–2110.
52. Spinner, M.A., Sanchez, L.A., Hsu, A.P., Shaw, P.A., Zerbe, C.S., Calvo, K.R., Arthur, D.C., Gu, W., Gould, C.M., Brewer, C.C., et al. (2014). GATA2 deficiency: a protean disorder of hematopoiesis, lymphatics, and immunity. Blood *123*, 809–821.
53. Swaminathan, B., Thorleifsson, G., Jöud, M., Ali, M., Johnsson, E., Ajore, R., Sulem, P., Halvarsson, B.-M., Eyjolfsson, G., Haraldsdottir, V., et al. (2015). Variants in ELL2 influencing immunoglobulin levels associate with multiple myeloma. Nat. Commun. *6*, 7213.
54. Ali, M., Ajore, R., Wihlborg, A.-K., Niroula, A., Swaminathan, B., Johnsson, E., Stephens, O.W., Morgan, G., Meissner, T., Turesson, I., et al. (2018). The multiple myeloma risk allele at 5q15 lowers ELL2 expression and increases ribosomal gene expression. Nat. Commun. *9*, 1649.
55. Du, Z., Weinhold, N., Song, G.C., Rand, K.A., Van Den Berg, D.J., Hwang, A.E., Sheng, X., Hom, V., Ailawadhi, S., Nooka, A.K., et al. (2020). A meta-analysis of genome-wide association studies of multiple myeloma among men and women of African ancestry. Blood Adv. *4*, 181–190.
56. Ye, H., Jeong, S.Y., Ghosh, M.C., Kovtunovych, G., Silvestri, L., Ortillo, D., Uchida, N., Tisdale, J., Camaschella, C., and Rouault, T.A. (2010). Glutaredoxin 5 deficiency causes sideroblastic anemia by specifically impairing heme biosynthesis and depleting cytosolic iron in human erythroblasts. J. Clin. Invest. *120*, 1749–1761.
57. Peskin, A.V., Pace, P.E., Behring, J.B., Paton, L.N., Soethoudt, M., Bachschmid, M.M., and Winterbourn, C.C. (2016). Glutathionylation of the active site cysteines of peroxiredoxin 2 and recycling by glutaredoxin. J. Biol. Chem. *291*, 3053–3062.

58. Furuyama, K., and Kaneko, K. (2018). Iron metabolism in erythroid cells and patients with congenital sideroblastic anemia. Int. J. Hematol. *107*, 44–54.
59. Zarychanski, R., Schulz, V.P., Houston, B.L., Maksimova, Y., Houston, D.S., Smith, B., Rinehart, J., and Gallagher, P.G. (2012). Mutations in the mechanotransduction protein PIEZO1 are associated with hereditary xerocytosis. Blood *120*, 1908–1915.
60. Andolfo, I., Alper, S.L., De Franceschi, L., Auriemma, C., Russo, R., De Falco, L., Vallefuoco, F., Esposito, M.R., Vandorpe, D.H., Shmukler, B.E., et al. (2013). Multiple clinical forms of dehydrated hereditary stomatocytosis arise from mutations in PIEZO1. Blood *121*, 3925–3935, S1–S12.
61. Knight, T., Zaidi, A.U., Wu, S., Gadgeel, M., Buck, S., and Ravindranath, Y. (2019). Mild erythrocytosis as a presenting manifestation of *PIEZO1* associated erythrocyte volume disorders. Pediatr. Hematol. Oncol. *36*, 317–326.
62. Zhang, T., Chi, S., Jiang, F., Zhao, Q., and Xiao, B. (2017). A protein interaction mechanism for suppressing the mechanosensitive Piezo channels. Nat. Commun. *8*, 1797.
63. Zhao, Q., Zhou, H., Chi, S., Wang, Y., Wang, J., Geng, J., Wu, K., Liu, W., Zhang, T., Dong, M.-Q., et al. (2018). Structure and mechanogating mechanism of the Piezo1 channel. Nature *554*, 487–492.
64. Ma, S., Cahalan, S., LaMonte, G., Grubaugh, N.D., Zeng, W., Murthy, S.E., Paytas, E., Gamini, R., Lukacs, V., Whitwam, T., et al. (2018). Common PIEZO1 allele in african populations causes RBC dehydration and attenuates plasmodium infection. Cell *173*, 443–455.e12.
65. Nguetse, C.N., Purington, N., Ebel, E.R., Shakya, B., Tetard, M., Kremsner, P.G., Velavan, T.P., and Egan, E.S. (2020). A common polymorphism in the mechanosensitive ion channel *PIEZO1* is associated with protection from severe malaria in humans. Proc. Natl. Acad. Sci. USA *117*, 9074–9081.
66. Wang, C.-Y., Meynard, D., and Lin, H.Y. (2014). The role of TMPRSS6/matriptase-2 in iron regulation and anemia. Front. Pharmacol. *5*, 114.
67. De Falco, L., Sanchez, M., Silvestri, L., Kannengiesser, C., Muckenthaler, M.U., Iolascon, A., Gouya, L., Camaschella, C., and Beaumont, C. (2013). Iron refractory iron deficiency anemia. Haematologica *98*, 845–853.
68. Silvestri, L., Guillem, F., Pagani, A., Nai, A., Oudin, C., Silva, M., Toutain, F., Kannengiesser, C., Beaumont, C., Camaschella, C., and Grandchamp, B. (2009). Molecular mechanisms of the defective hepcidin inhibition in TMPRSS6 mutations associated with iron-refractory iron deficiency anemia. Blood *113*, 5605–5608.
69. Benmansour, I., Moradkhani, K., Moumni, I., Wajcman, H., Hafsia, R., Ghanem, A., Abbès, S., and Préhu, C. (2013). Two new class III G6PD variants [G6PD Tunis (c.920A>C: p.307Gln>Pro) and G6PD Nefza (c.968T>C: p.323 Leu>-Pro)] and overview of the spectrum of mutations in Tunisia. Blood Cells Mol. Dis. *50*, 110–114.
70. Beutler, B., and Cerami, A. (1989). The biology of cachectin/TNF–a primary mediator of the host response. Annu. Rev. Immunol. *7*, 625–655.
71. Hamel, A.R., Cabral, I.R., Sales, T.S.I., Costa, F.F., and Olalla Saad, S.T. (2002). Molecular heterogeneity of G6PD deficiency in an Amazonian population and description of four new variants. Blood Cells Mol. Dis. *28*, 399–406.
72. Monteiro, W.M., Franca, G.P., Melo, G.C., Queiroz, A.L.M., Brito, M., Peixoto, H.M., Oliveira, M.R.F., Romero, G.A.S., Bassat, Q., and Lacerda, M.V.G. (2014). Clinical complications of G6PD deficiency in Latin American and Caribbean populations: systematic review and implications for malaria elimination programmes. Malar. J. *13*, 70.
73. Reading, N.S., Ruiz-Bonilla, J.A., Christensen, R.D., Cáceres-Perkins, W., and Prchal, J.T. (2017). A patient with both methemoglobinemia and G6PD deficiency: A therapeutic conundrum. Am. J. Hematol. *92*, 474–477.
74. Ramírez-Nava, E.J., Ortega-Cuellar, D., Serrano-Posada, H., González-Valdez, A., Vanoye-Carlo, A., Hernández-Ochoa, B., Sierra-Palacios, E., Hernández-Pineda, J., Rodríguez-Bustamante, E., Arreguin-Espinosa, R., et al. (2017). Biochemical Analysis of Two Single Mutants that Give Rise to a Polymorphic G6PD A-Double Mutant. Int. J. Mol. Sci. *18*, 18.
75. Sarnowski, C., Leong, A., Raffield, L.M., Wu, P., de Vries, P.S., DiCorpo, D., Guo, X., Xu, H., Liu, Y., Zheng, X., et al.; TOPMed Diabetes Working Group; TOPMed Hematology Working Group; TOPMed Hemostasis Working Group; and National Heart, Lung, and Blood Institute TOPMed Consortium (2019). Impact of Rare and Common Genetic Variants on Diabetes Diagnosis by Hemoglobin A1c in Multi-Ancestry Cohorts: The Trans-Omics for Precision Medicine Program. Am. J. Hum. Genet. *105*, 706–718.
76. Huang, Y., Choi, M.Y., Au, S.W.N., Au, D.M.Y., Lam, V.M.S., and Engel, P.C. (2008). Purification and detailed study of two clinically different human glucose 6-phosphate dehydrogenase variants, G6PD(Plymouth) and G6PD(Mahidol): Evidence for defective protein folding as the basis of disease. Mol. Genet. Metab. *93*, 44–53.
77. Wang, X.-T., Lam, V.M.S., and Engel, P.C. (2005). Marked decrease in specific activity contributes to disease phenotype in two human glucose 6-phosphate dehydrogenase mutants, G6PD(Union) and G6PD(Andalus). Hum. Mutat. *26*, 284.
78. Chiu, D.T., Zuo, L., Chao, L., Chen, E., Louie, E., Lubin, B., Liu, T.Z., and Du, C.S. (1993). Molecular characterization of glucose-6-phosphate dehydrogenase (G6PD) deficiency in patients of Chinese descent and identification of new base substitutions in the human G6PD gene. Blood *81*, 2150–2154.
79. Luzzatto, L., Ally, M., and Notaro, R. (2020). Glucose-6-phosphate dehydrogenase deficiency. Blood *136*, 1225–1240.
80. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518.
81. Fatumo, S., Carstensen, T., Nashiru, O., Gurdasani, D., Sandhu, M., and Kaleebu, P. (2019). Complimentary Methods for Multivariate Genome-Wide Association Study Identify New Susceptibility Genes for Blood Cell Traits. Front. Genet. *10*, 334.
82. Velasco-Rodríguez, D., Alonso-Domínguez, J.-M., González-Fernández, F.-A., Muriel, A., Abalo, L., Sopeña, M., Villarrubia, J., Ropero, P., Plaza, M.P., Tenorio, M., et al. (2016). Laboratory parameters provided by Advia 2120 analyser identify structural haemoglobinopathy carriers and discriminate between Hb S trait and Hb C trait. J. Clin. Pathol. *69*, 912–920.
83. Antonarakis, S.E., Boehm, C.D., Serjeant, G.R., Theisen, C.E., Dover, G.J., and Kazazian, H.H., Jr. (1984). Origin of the beta S-globin gene in blacks: the contribution of recurrent mutation or gene conversion or both. Proc. Natl. Acad. Sci. USA *81*, 853–856.

84. Wong, C., Antonarakis, S.E., Goff, S.C., Orkin, S.H., Boehm, C.D., and Kazazian, H.H., Jr. (1986). On the origin and spread of beta-thalassemia: recurrent observation of four mutations in different ethnic groups. Proc. Natl. Acad. Sci. USA *83*, 6529–6532.
85. Orkin, S.H., Antonarakis, S.E., and Kazazian, H.H., Jr. (1984). Base substitution at position -88 in a beta-thalassemic globin gene. Further evidence for the role of distal promoter element ACACCC. J. Biol. Chem. *259*, 8679–8681.
86. Gonzalez-Redondo, J.M., Stoming, T.A., Lanclos, K.D., Gu, Y.C., Kutlar, A., Kutlar, F., Nakatsuji, T., Deng, B., Han, I.S., McKie, V.C., et al. (1988). Clinical and genetic heterogeneity in black patients with homozygous beta-thalassemia from the southeastern United States. Blood *72*, 1007–1014.
87. Treisman, R., Orkin, S.H., and Maniatis, T. (1983). Specific transcription and RNA splicing defects in five cloned beta-thalassaemia genes. Nature *302*, 591–596.
88. Westaway, D., and Williamson, R. (1981). An intron nucleotide sequence variant in a cloned beta +-thalassaemia globin gene. Nucleic Acids Res. *9*, 1777–1788.
89. Spritz, R.A., Jagadeeswaran, P., Choudary, P.V., Biro, P.A., Elder, J.T., deRiel, J.K., Manley, J.L., Gefter, M.L., Forget, B.G., and Weissman, S.M. (1981). Base substitution in an intervening sequence of a beta+-thalassemic human globin gene. Proc. Natl. Acad. Sci. USA *78*, 2455–2459.
90. Trecartin, R.F., Liebhaber, S.A., Chang, J.C., Lee, K.Y., Kan, Y.W., Furbetta, M., Angius, A., and Cao, A. (1981). beta zero thalassemia in Sardinia is caused by a nonsense mutation. J. Clin. Invest. *68*, 1012–1017.
91. Orkin, S.H., and Goff, S.C. (1981). Nonsense and frameshift mutations in beta 0-thalassemia detected in cloned beta-globin genes. J. Biol. Chem. *256*, 9782–9784.
92. Atweh, G.F., Wong, C., Reed, R., Antonarakis, S.E., Zhu, D., Ghosh, P.K., Maniatis, T., Forget, B.G., and Kazazian, H.H., Jr. (1987). A new mutation in IVS-1 of the human beta globin gene causing beta thalassemia due to abnormal splicing. Blood *70*, 147–151.
93. 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; and CHARGE Consortium (2013). Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. Hum. Mol. Genet. *22*, 2529–2538.
94. Harrison, A., Mashon, R.S., Kakkar, N., and Das, S. (2018). Clinico-Hematological Profile of Hb Q India: An Uncommon Hemoglobin Variant. Indian J. Hematol. Blood Transfus. *34*, 299–303.
95. Schmidt, R.M., Bechtel, K.C., and Moo-Penn, W.F. (1976). Hemoglobin QIndia, alpha 64 (E13) Asp replaced by His, and beta-thalassemia in a Canadian family. Am. J. Clin. Pathol. *66*, 446–448.
96. Sukumaran, P.K., Merchant, S.M., Desai, M.P., Wiltshire, B.G., and Lehmann, H. (1972). Haemoglobin Q India (alpha 64(E13) aspartic acid histidine) associated with beta-thalassemia observed in three Sindhi families. J. Med. Genet. *9*, 436–442.
97. Yu, X., Mollan, T.L., Butler, A., Gow, A.J., Olson, J.S., and Weiss, M.J. (2009). Analysis of human alpha globin gene mutations that impair binding to the alpha hemoglobin stabilizing protein. Blood *113*, 5961–5969.
98. Giordano, P.C., Zweegman, S., Akkermans, N., Arkesteijn, S.G.J., van Delft, P., Versteegh, F.G.A., Wajcman, H., and Harteveld, C.L. (2007). The first case of Hb Groene Hart [alpha119(H2)Pro–>Ser, CCT–>TCT (alpha1)] homozygosity confirms that a thalassemia phenotype is associated with this abnormal hemoglobin variant. Hemoglobin *31*, 179–182.
99. Joly, P., Lacan, P., Garcia, C., and Francina, A. (2014). Description of the phenotypes of 63 heterozygous, homozygous and compound heterozygous patients carrying the Hb Groene Hart [α119(H2)Pro®Ser; HBA1: c.358C>T] variant. Hemoglobin *38*, 64–66.
100. Chami, N., Chen, M.-H., Slater, A.J., Eicher, J.D., Evangelou, E., Tajuddin, S.M., Love-Gregory, L., Kacprowski, T., Schick, U.M., Nomura, A., et al. (2016). Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits. Am. J. Hum. Genet. *99*, 8–21.
101. Cserti-Gazdewich, C.M., Mayr, W.R., and Dzik, W.H. (2011). Plasmodium falciparum malaria and the immunogenetics of ABO, HLA, and CD36 (platelet glycoprotein IV). Vox Sang. *100*, 99–111.
102. Fillebeen, C., Charlebois, E., Wagner, J., Katsarou, A., Mui, J., Vali, H., Garcia-Santos, D., Ponka, P., Presley, J., and Pantopoulos, K. (2019). Transferrin receptor 1 controls systemic iron homeostasis by fine-tuning hepcidin expression to hepatocellular iron load. Blood *133*, 344–355.
103. Aljohani, A.H., Al-Mousa, H., Arnaout, R., Al-Dhekri, H., Mohammed, R., Alsum, Z., Nicolas-Jilwan, M., Alrogi, F., Al-Muhsen, S., Alazami, A.M., and Al-Saud, B. (2020). Clinical and immunological characterization of combined immunodeficiency due to TFRC mutation in eight patients. J. Clin. Immunol. *40*, 1103–1110.
104. Pan, D., Kalfa, T.A., Wang, D., Risinger, M., Crable, S., Ottlinger, A., Chandra, S., Mount, D.B., Hübner, C.A., Franco, R.S., and Joiner, C.H. (2011). K-Cl cotransporter gene expression during human and murine erythroid differentiation. J. Biol. Chem. *286*, 30492–30503.
105. Marcoux, A.A., Garneau, A.P., Frenette-Cotton, R., Slimani, S., Mac-Way, F., and Isenring, P. (2017). Molecular features and physiological roles of K^+-Cl^- cotransporter 4 (KCC4). Biochim. Biophys. Acta, Gen. Subj. *1861*, 3154–3166.
106. Brugnara, C. (2003). Sickle cell disease: from membrane pathophysiology to novel therapies for prevention of erythrocyte dehydration. J. Pediatr. Hematol. Oncol. *25*, 927–933.

The American Journal of Human Genetics, Volume *108*

Supplemental information

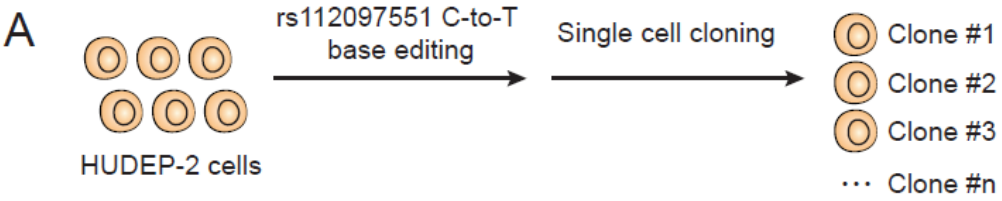
Whole-genome sequencing association analysis of quantitative red blood cell phenotypes: The NHLBI TOPMed program

Yao Hu, Adrienne M. Stilp, Caitlin P. McHugh, Shuquan Rao, Deepti Jain, Xiuwen Zheng, John Lane, Sébastian Méric de Bellefon, Laura M. Raffield, Ming-Huei Chen, Lisa R. Yanek, Marsha Wheeler, Yao Yao, Chunyan Ren, Jai Broome, Jee-Young Moon, Paul S. de Vries, Brian D. Hobbs, Quan Sun, Praveen Surendran, Jennifer A. Brody, Thomas W. Blackwell, Hélène Choquet, Kathleen Ryan, Ravindranath Duggirala, Nancy Heard-Costa, Zhe Wang, Nathalie Chami, Michael H. Preuss, Nancy Min, Lynette Ekunwe, Leslie A. Lange, Mary Cushman, Nauder Faraday, Joanne E. Curran, Laura Almasy, Kousik Kundu, Albert V. Smith, Stacey Gabriel, Jerome I. Rotter, Myriam Fornage, Donald M. Lloyd-Jones, Ramachandran S. Vasan, Nicholas L. Smith, Kari E. North, Eric Boerwinkle, Lewis C. Becker, Joshua P. Lewis, Goncalo R. Abecasis, Lifang Hou, Jeffrey R. O'Connell, Alanna C. Morrison, Terri H. Beaty, Robert Kaplan, Adolfo Correa, John Blangero, Eric Jorgenson, Bruce M. Psaty, Charles Kooperberg, Russell T. Walton, Benjamin P. Kleinstiver, Hua Tang, Ruth J.F. Loos, Nicole Soranzo, Adam S. Butterworth, Debbie Nickerson, Stephen S. Rich, Braxton D. Mitchell, Andrew D. Johnson, Paul L. Auer, Yun Li, Rasika A. Mathias, Guillaume Lettre, Nathan Pankratz, Cathy C. Laurie, Cecelia A. Laurie, Daniel E. Bauer, Matthew P. Conomos, Alexander P. Reiner, and and the NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium

Supplemental Figures and Legends

Figure S1. Rs112097551 C-to-T base editing and single cell cloning in HUDEP-2 cells. (A) Scheme of rs112097551 C-to-T base editing and FACS-based single cell separation. (B) Efficiency of rs112097551 C-to-T (G-to-A on opposing strand) base editing efficiency in all five clones. Since base editor and sgRNA are constitutively expressed, the frequency of C-to-T conversion may exceed 50% in heterozygous clones due to base editing after single cell cloning.

A

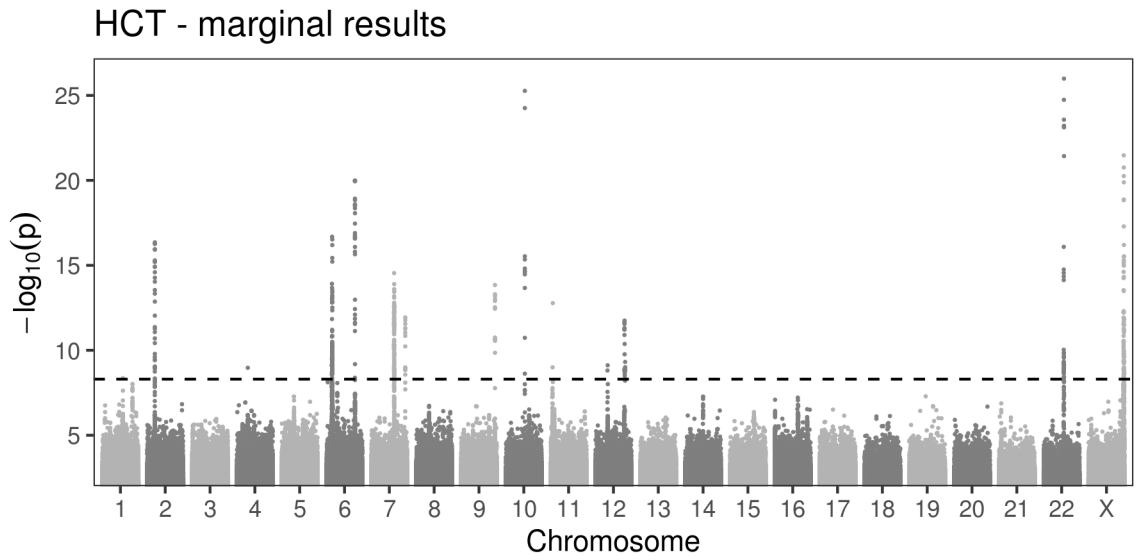


B

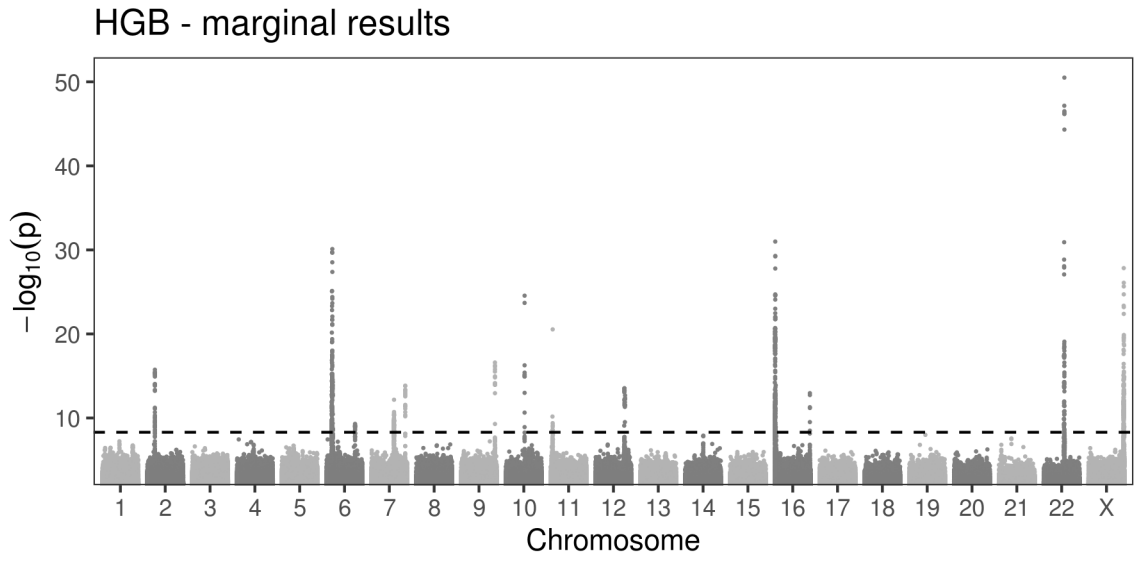
Clone ID	Allele A percentage
Clone #1	59%
Clone #2	61%
Clone #3	65%
Clone #4	54%
Clone #5	58%

Figure S2. Manhattan plots of the marginal single-variant analyses in TOPMed. (A) HCT; (B) HGB; (C) MCH; (D) MCHC; (E) MCV; (F) RBC; (G) RDW.

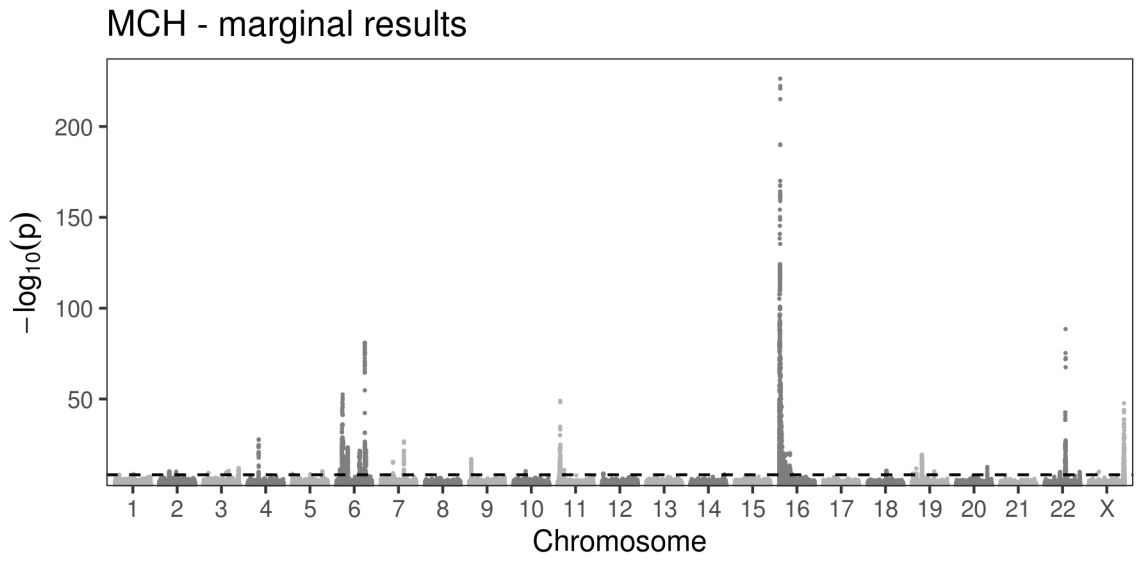
(A)



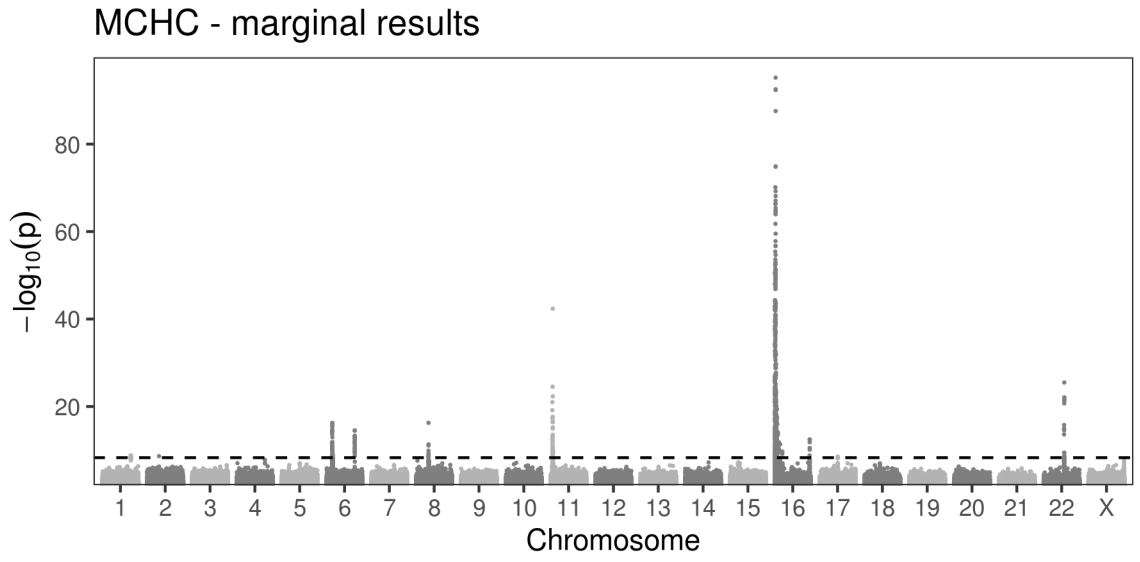
(B)



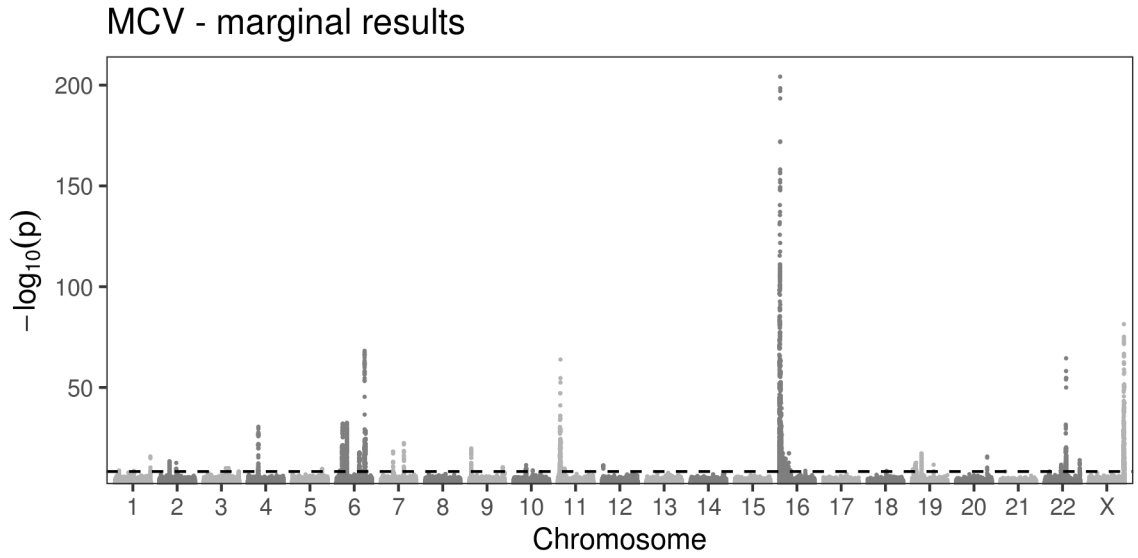
(C)


MCH - marginal results
200
150
100
50
−log10(p)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosome

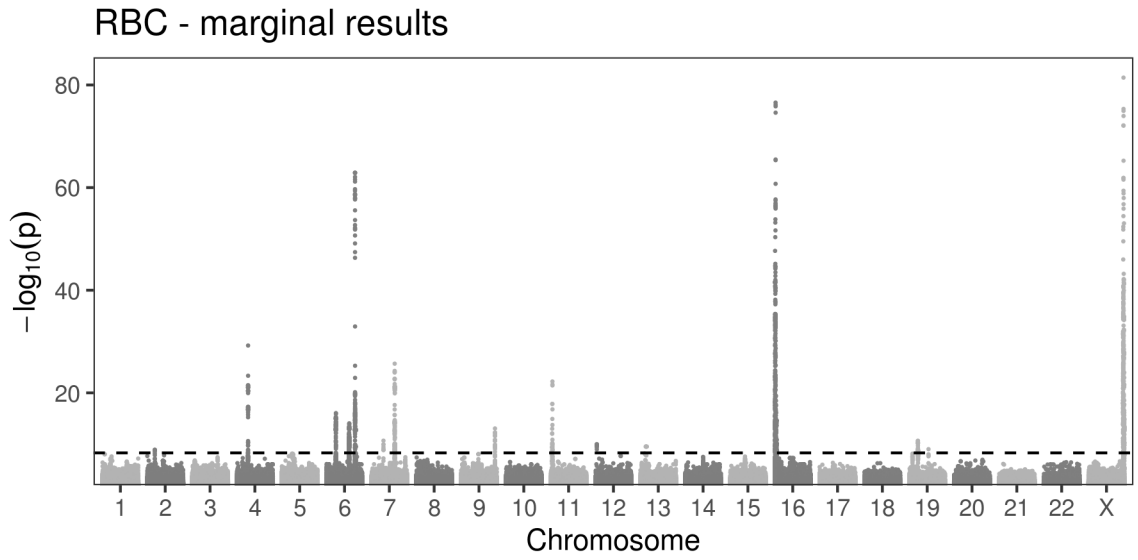
(D)


MCHC - marginal results
80
60
40
20
−log10(p)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosome

(E)


MCV - marginal results
200
150
100
50
−log10(p)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosome

(F)


RBC - marginal results
80
60
40
20
−log10(p)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosome

(G)

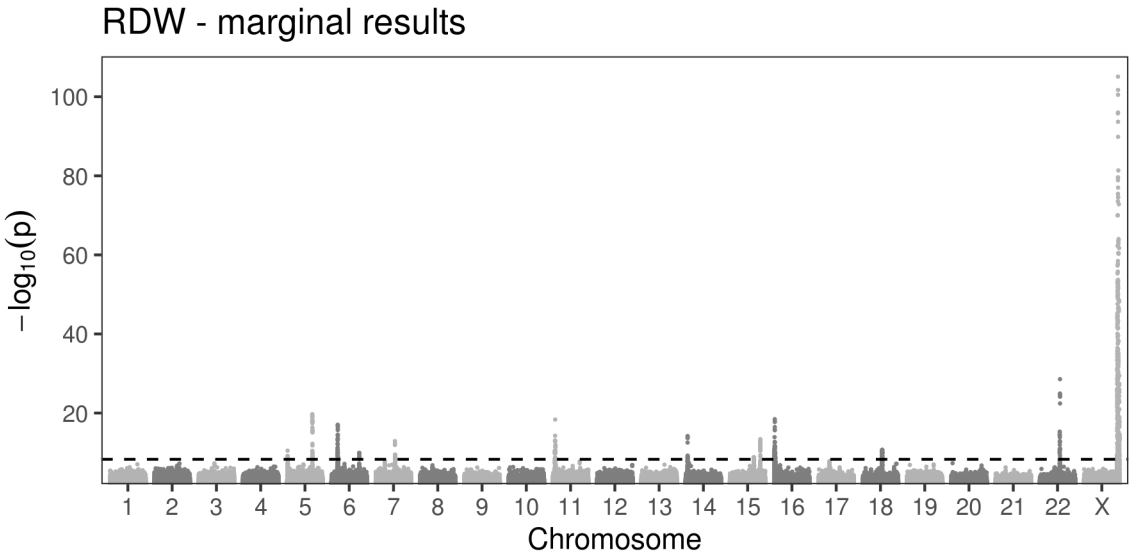
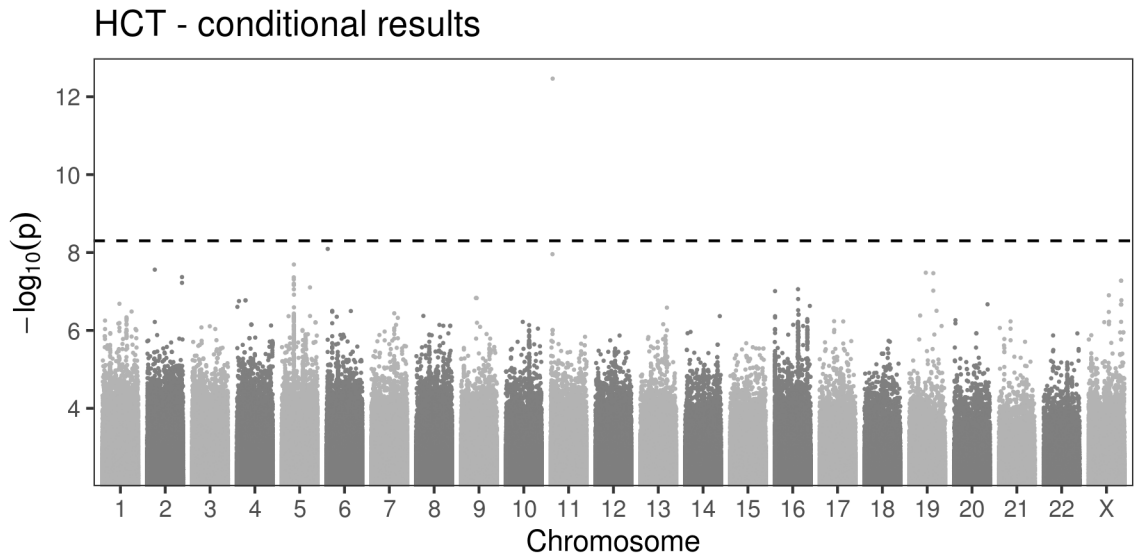
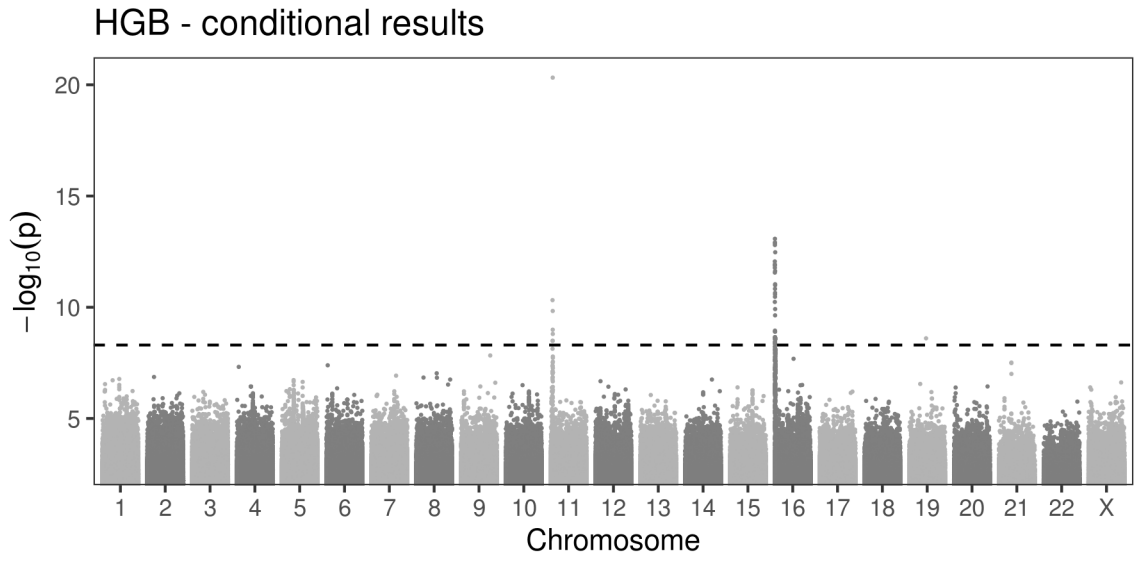

RDW - marginal results
100
80
60
40
20
−log10(p)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Chromosome

Figure S3. Manhattan plots of the trait-specific conditional single-variant analyses in TOPMed. (A) HCT; (B) HGB; (C) MCH; (D) MCHC; (E) MCV; (F) RBC; (G) RDW.

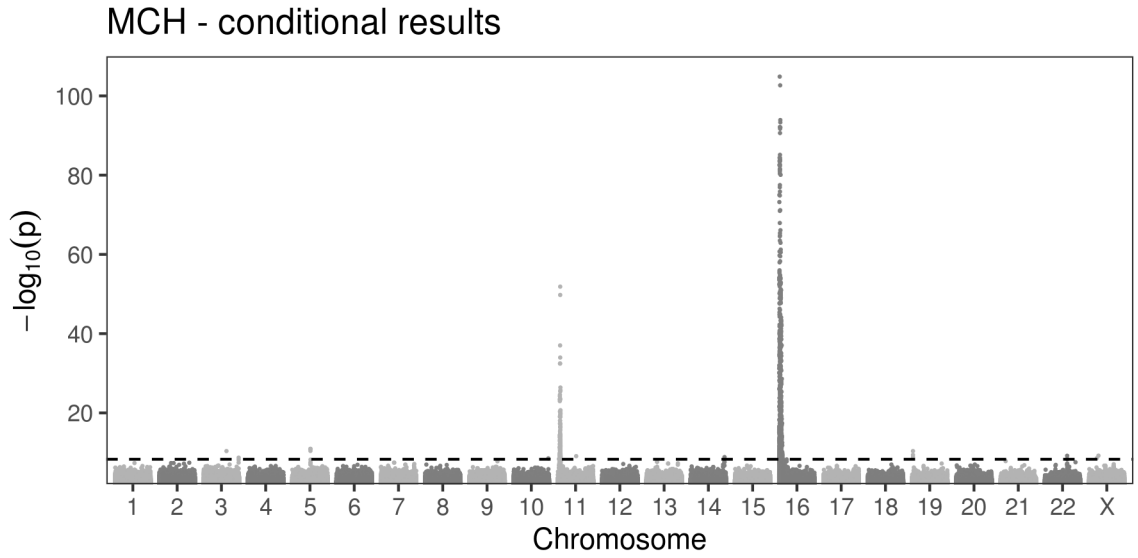
(A)



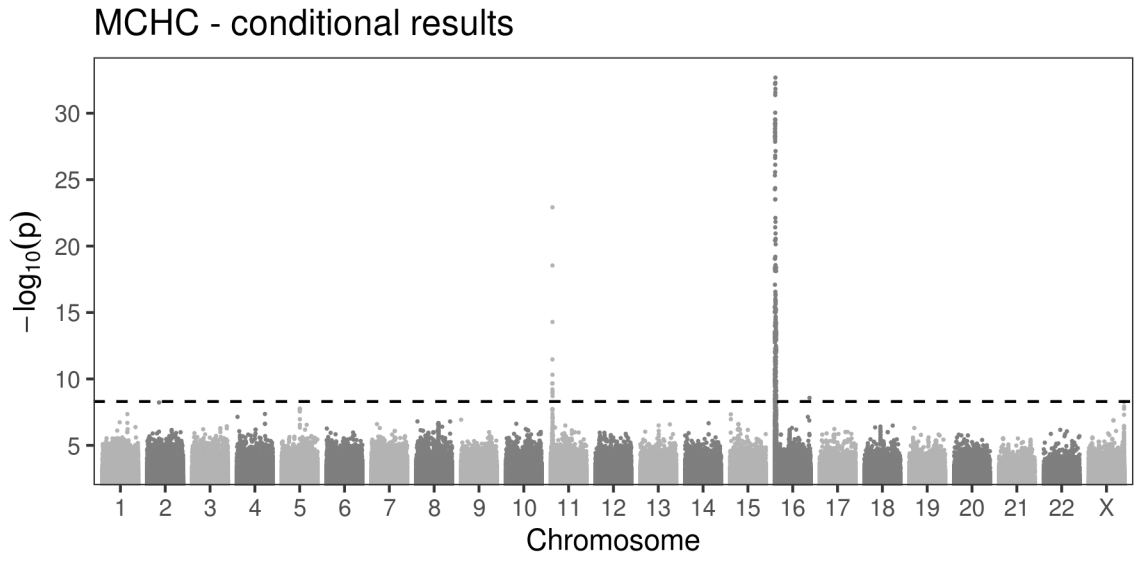
(B)



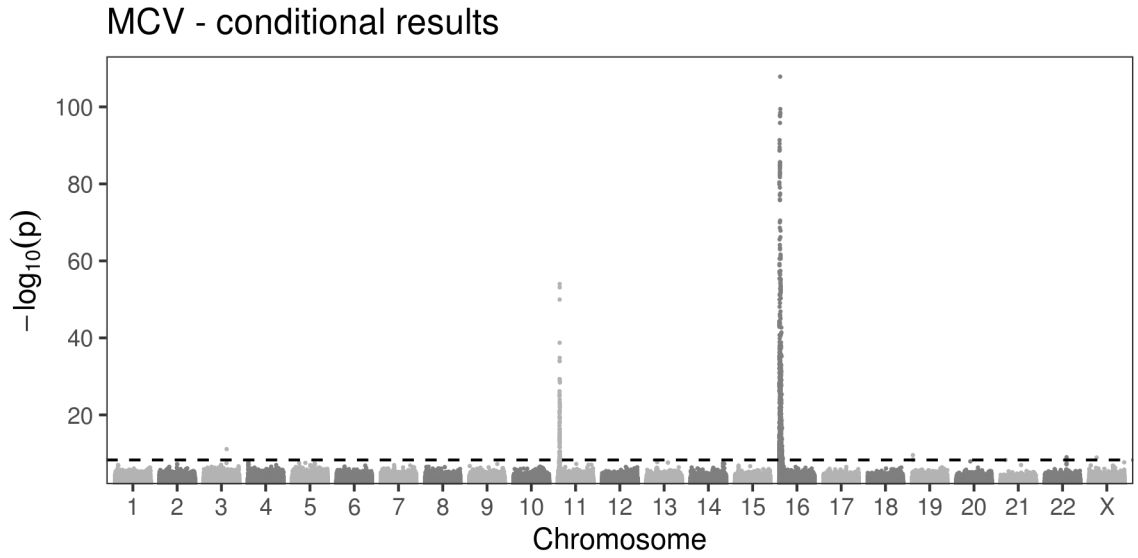
(C)


MCH - conditional results
100
80
60
40
20
−log10(p)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosome

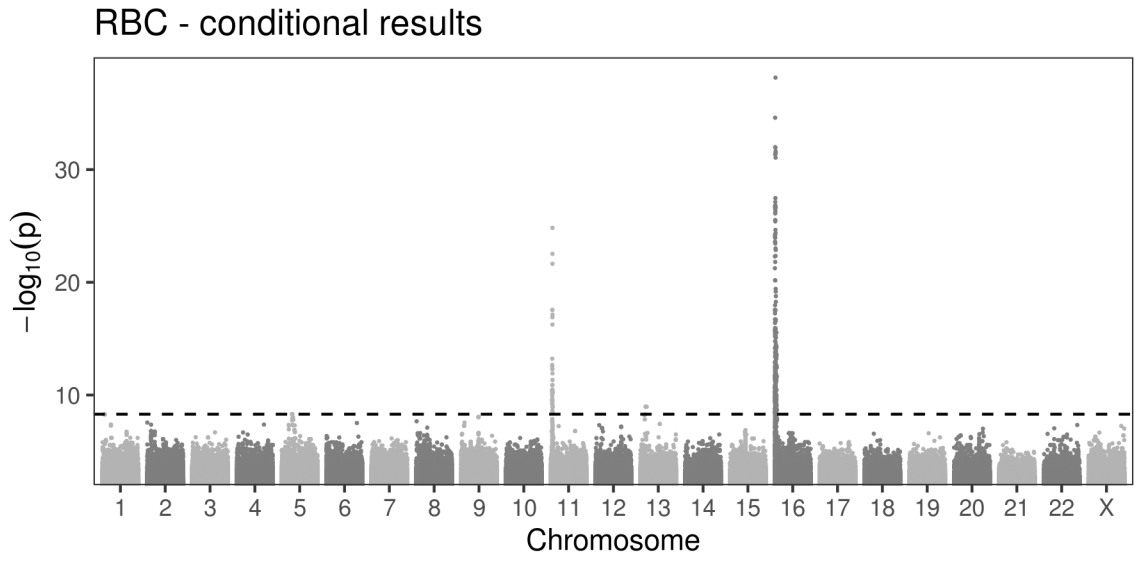
(D)


MCHC - conditional results
30
25
20
15
10
5
−log10(p)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosome

(E)


MCV - conditional results
100
80
60
40
20
−log10(p)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosome

(F)


RBC - conditional results
30
20
10
−log10(p)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosome

(G)

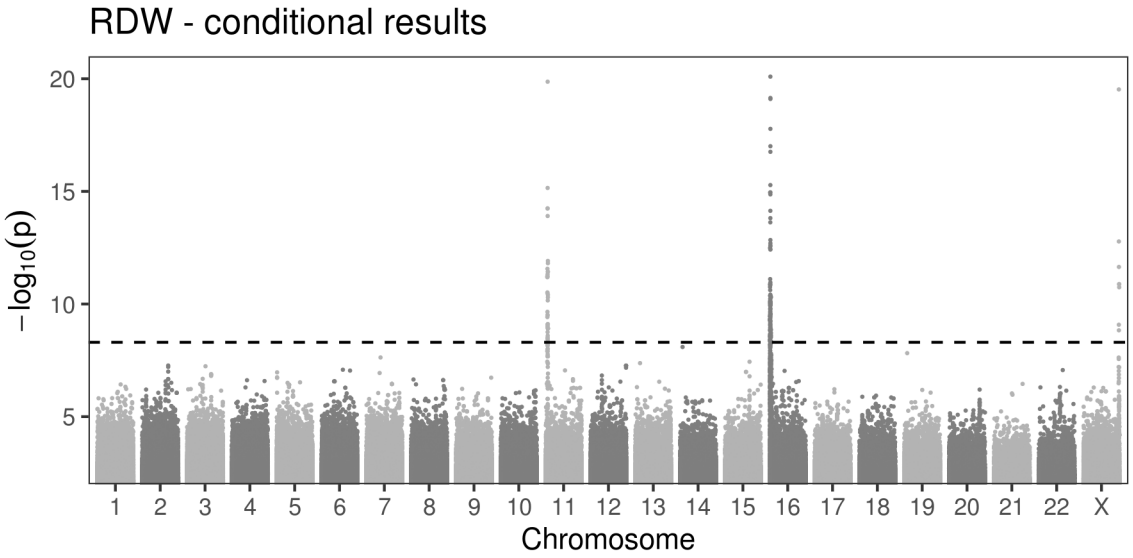
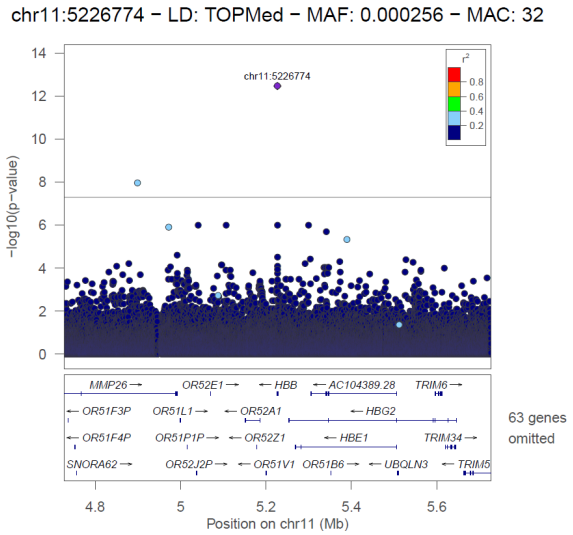
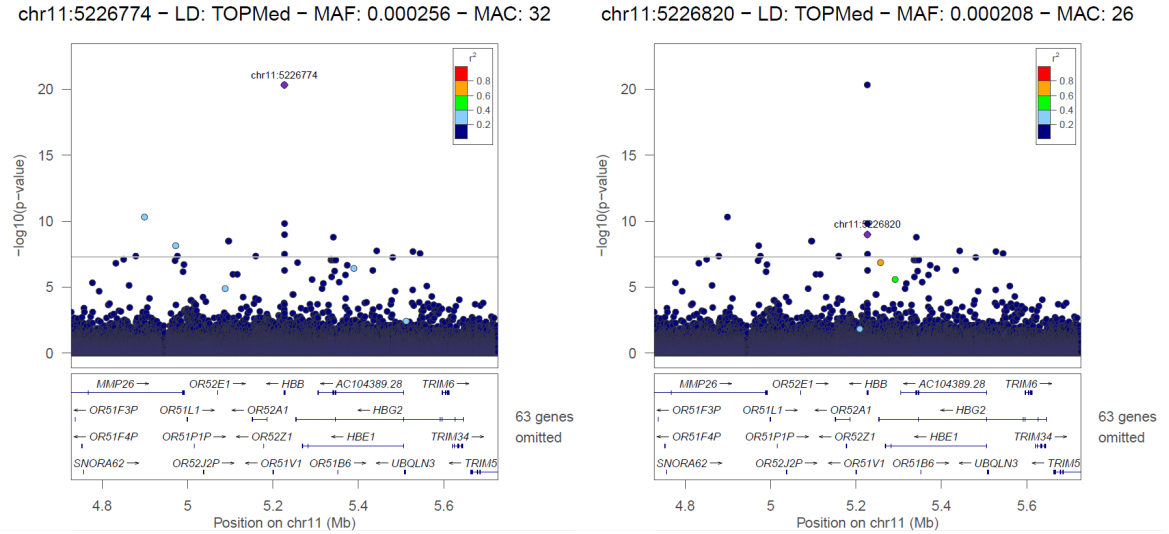

RDW - conditional results
20
15
10
5
−log10(p)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Chromosome

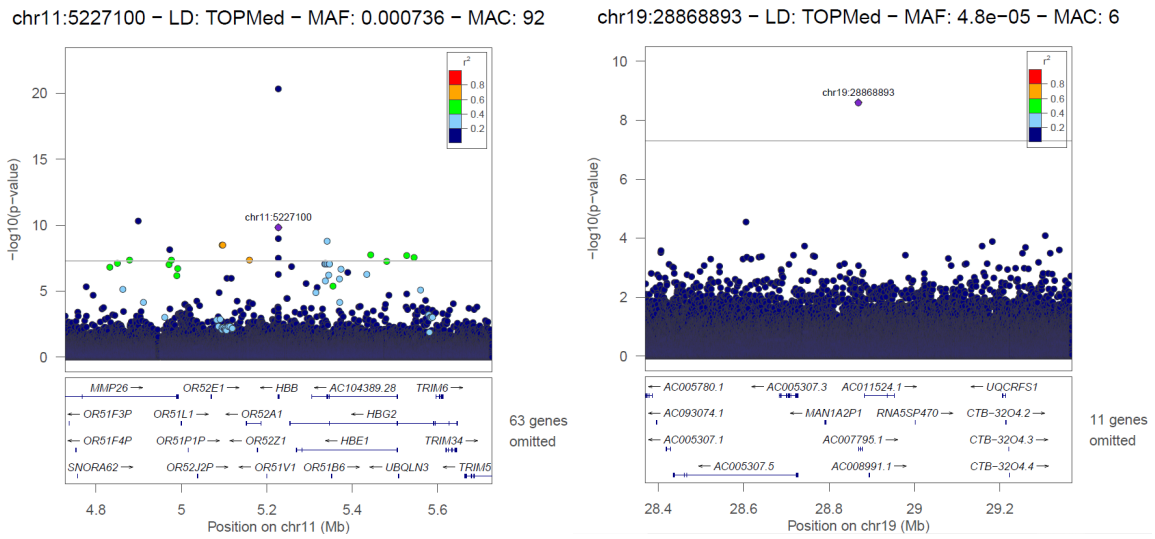
Figure S4. Locuszoom plots of the 12 novel variants and conditionally independent variants identified in TOPMed. (A) HCT; (B) HGB; (C) MCH; (D) MCHC; (E) MCV; (F) RBC; (G) RDW

(A)

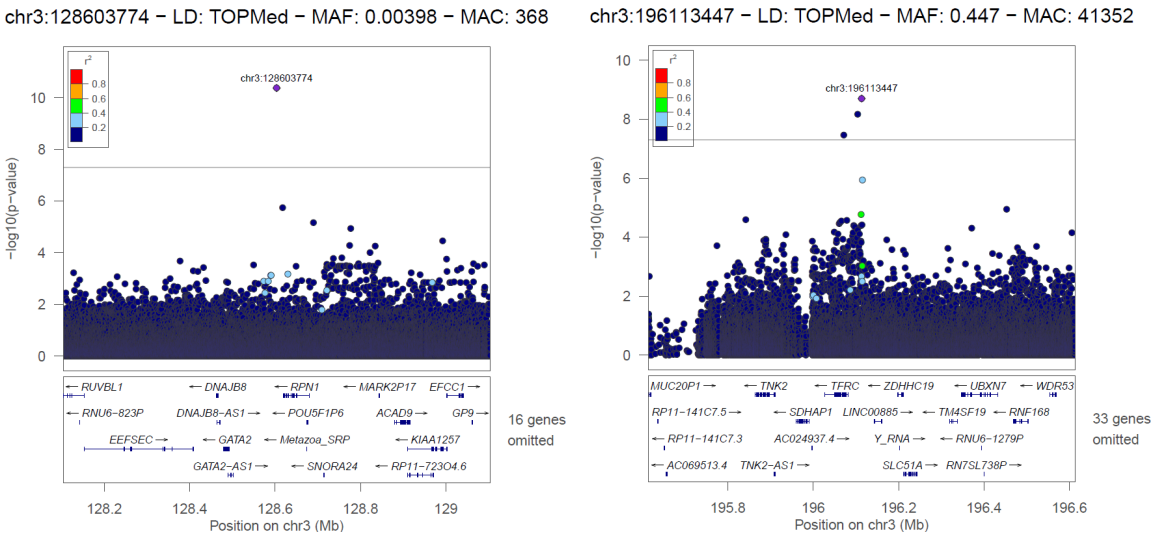


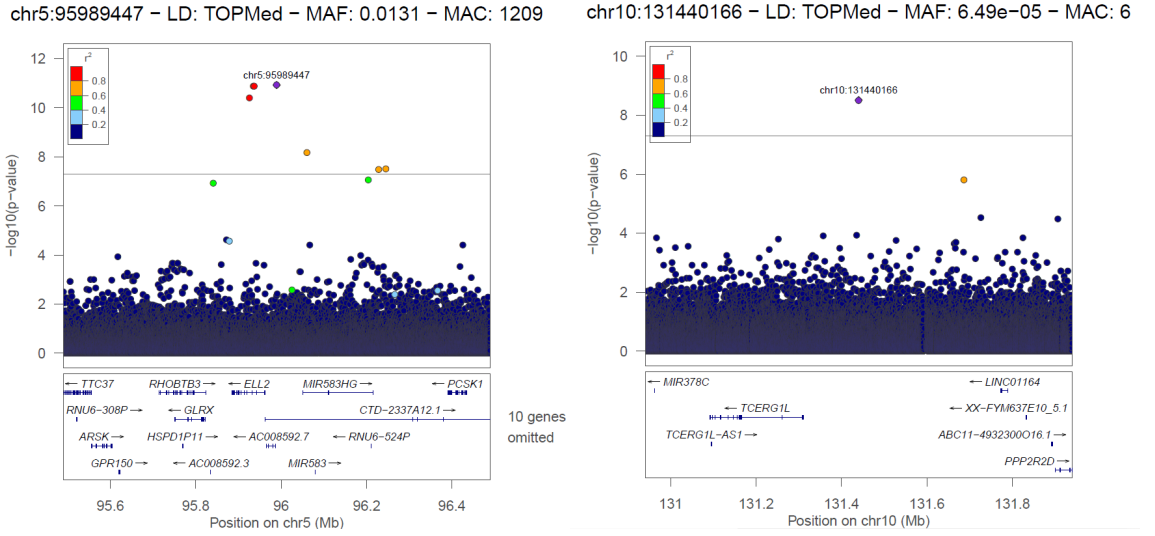
(B)



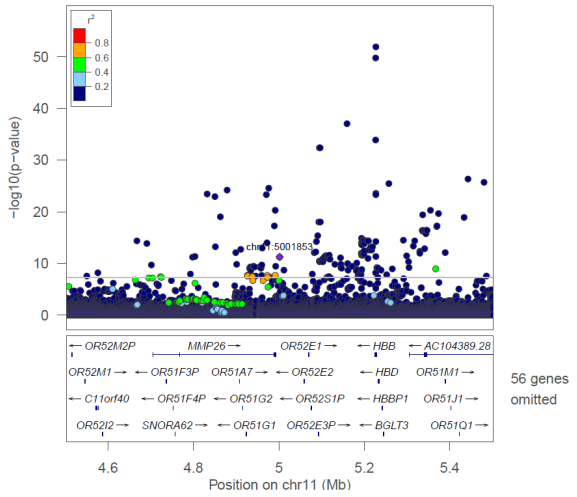

chr11:5227100 − LD: TOPMed − MAF: 0.000736 − MAC: 92
chr19:28868893 − LD: TOPMed − MAF: 4.8e−05 − MAC: 6
−log10(p−value)
r²
0.8
0.6
0.4
0.2
chr11:5227100
chr19:28868893
MMP26 → OR52E1 → ← HBB ← AC104389.28 TRIM6 →
OR51F3P OR51L1 → ← OR52A1 ← HBG2
← OR51F4P OR51P1P → ← OR52Z1 ← HBE1 TRIM34 →
SNORA62 → OR52J2P → ← OR51V1 OR51B6 → ← UBQLN3 ← TRIM5
63 genes omitted
← AC005780.1 ← AC005307.3 AC011524.1 → ← UQCRFS1
← AC093074.1 ← MAN1A2P1 RNA5SP470 → CTB−32O4.2 →
← AC005307.1 AC007795.1 → CTB−32O4.3 →
← AC005307.5 AC008991.1 → CTB−32O4.4 →
11 genes omitted
Position on chr11 (Mb)
Position on chr19 (Mb)

(C)


chr3:128603774 − LD: TOPMed − MAF: 0.00398 − MAC: 368
chr3:196113447 − LD: TOPMed − MAF: 0.447 − MAC: 41352
−log10(p−value)
r²
0.8
0.6
0.4
0.2
chr3:128603774
chr3:196113447
← RUVBL1 ← DNAJB8 ← RPN1 ← MARK2P17 EFCC1 →
← RNU6−823P DNAJB8−AS1 → ← POU5F1P6 ACAD9 → GP9 →
EEFSEC → ← GATA2 ← Metazoa_SRP ← KIAA1257
GATA2−AS1 → ← SNORA24 ← RP11−723O4.6
16 genes omitted
MUC20P1 → ← TNK2 ← TFRC ← ZDHHC19 ← UBXN7 ← WDR53
RP11−141C7.5 → ← SDHAP1 LINC00885 → ← TM4SF19 ← RNF168
← RP11−141C7.3 AC024937.4 → Y_RNA → ← RNU6−1279P
← AC069513.4 TNK2−AS1 → SLC51A → RN7SL738P →
33 genes omitted
Position on chr3 (Mb)
Position on chr3 (Mb)

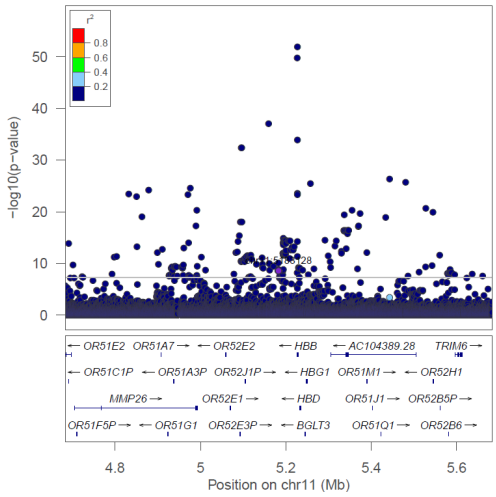

chr5:95989447 − LD: TOPMed − MAF: 0.0131 − MAC: 1209
chr10:131440166 − LD: TOPMed − MAF: 6.49e−05 − MAC: 6
−log10(p−value)
r²
0.8
0.6
0.4
0.2
chr5:95989447
chr10:131440166
← TTC37 RHOBTB3 → ← ELL2 MIR583HG → ← PCSK1
RNU6−308P → ← GLRX CTD−2337A12.1 →
ARSK → HSPD1P11 → ← AC008592.7 ← RNU6−524P
GPR150 → ← AC008592.3 MIR583 →
10 genes omitted
← MIR378C ← LINC01164
← TCERG1L ← XX−FYM637E10_5.1
TCERG1L−AS1 → ABC11−4932300O16.1 →
PPP2R2D →
Position on chr5 (Mb)
Position on chr10 (Mb)

chr11:5001853 − LD: TOPMed − MAF: 0.0044 − MAC: 407



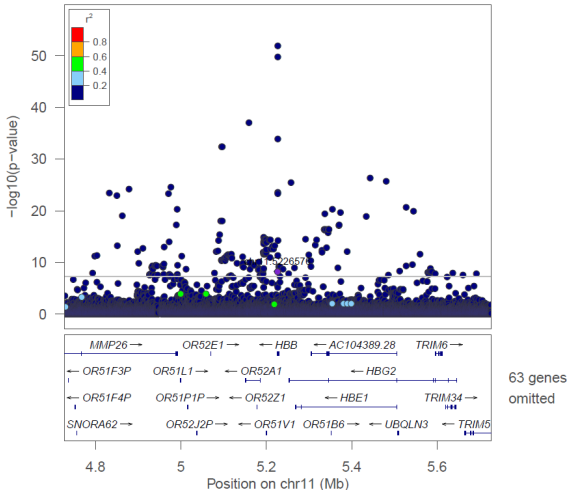
56 genes omitted

chr11:5183128 − LD: TOPMed − MAF: 5.41e−05 − MAC: 5



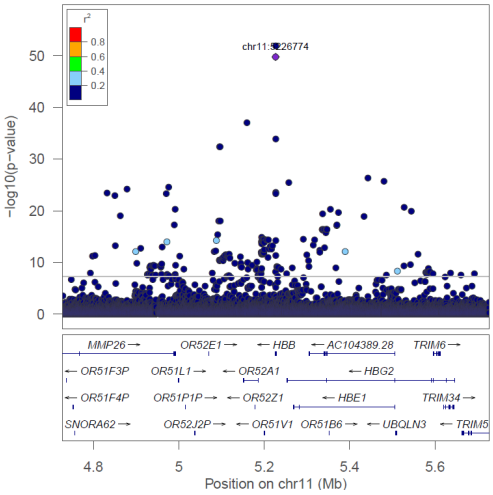
62 genes omitted

chr11:5226576 − LD: TOPMed − MAF: 7.57e−05 − MAC: 7



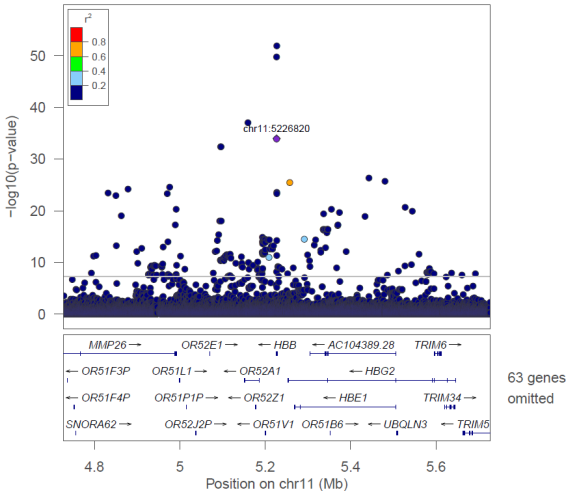
63 genes omitted

chr11:5226774 − LD: TOPMed − MAF: 0.000292 − MAC: 27



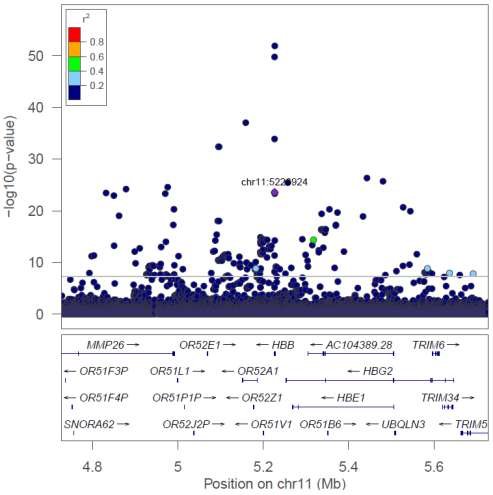
63 genes omitted

chr11:5226820 − LD: TOPMed − MAF: 0.000216 − MAC: 20



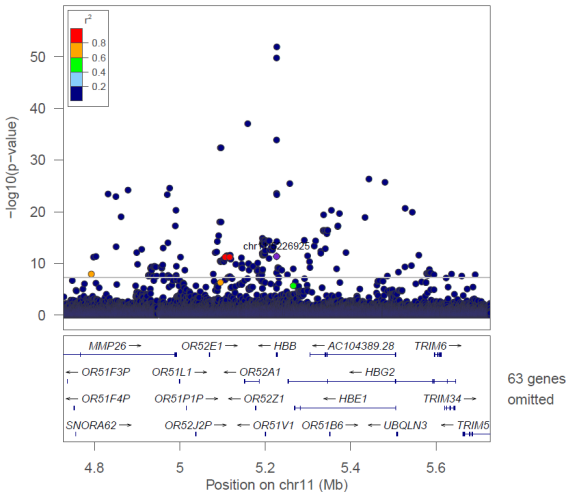
63 genes omitted

chr11:5226924 − LD: TOPMed − MAF: 0.000151 − MAC: 14

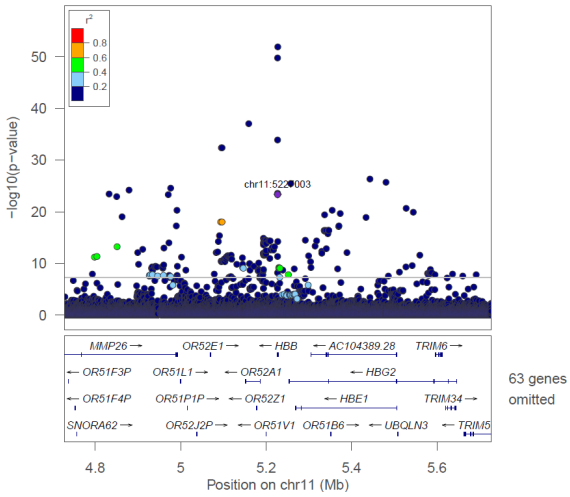


63 genes omitted

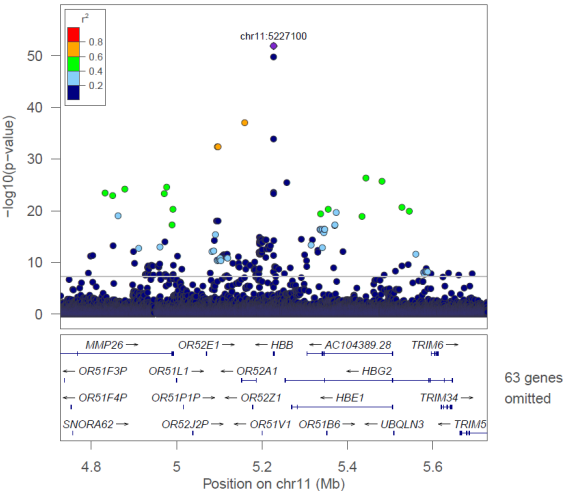
chr11:5226925 − LD: TOPMed − MAF: 0.000108 − MAC: 10



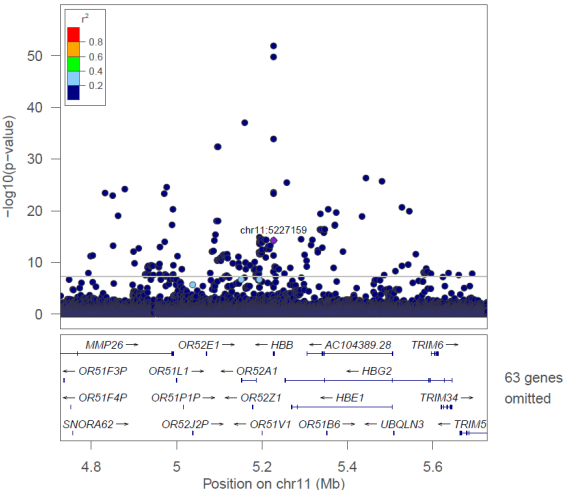
chr11:5227003 − LD: TOPMed − MAF: 0.00408 − MAC: 377



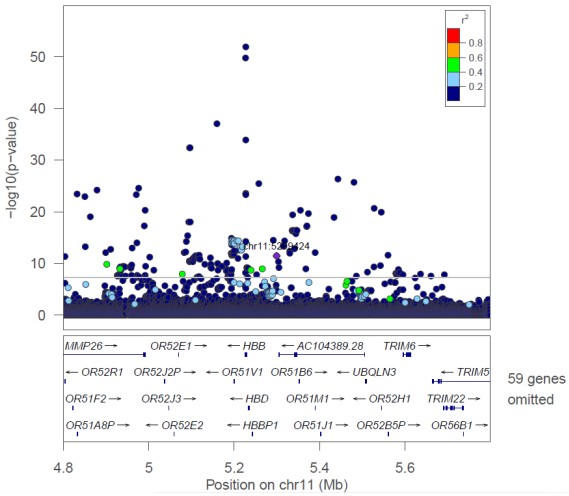
chr11:5227100 − LD: TOPMed − MAF: 0.000833 − MAC: 77



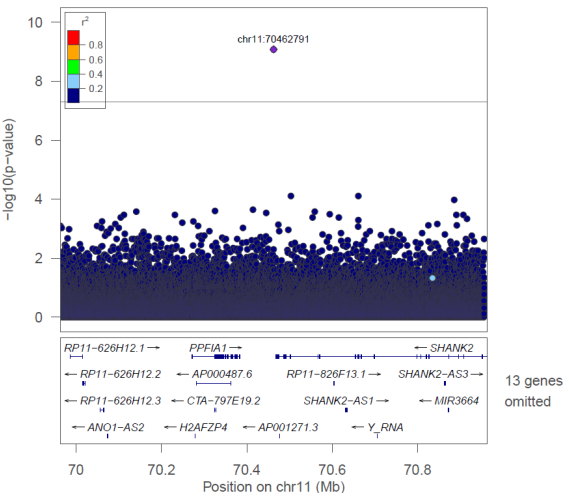
chr11:5227159 − LD: TOPMed − MAF: 0.000249 − MAC: 23



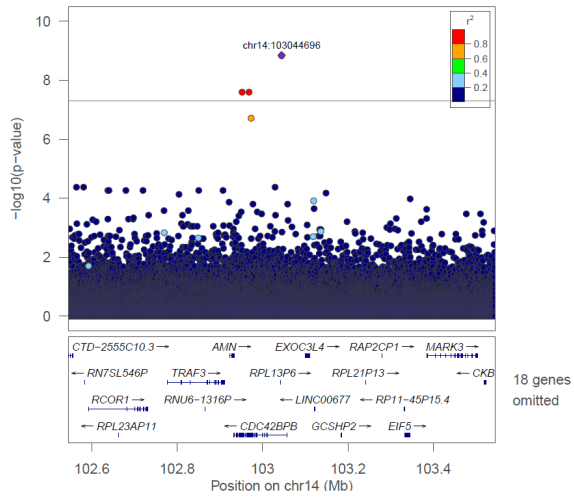
chr11:5299424 − LD: TOPMed − MAF: 0.00889 − MAC: 822



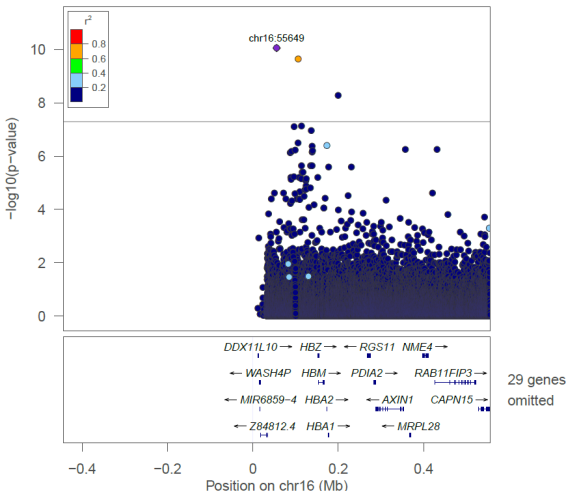
chr11:70462791 − LD: TOPMed − MAF: 7.57e−05 − MAC: 7



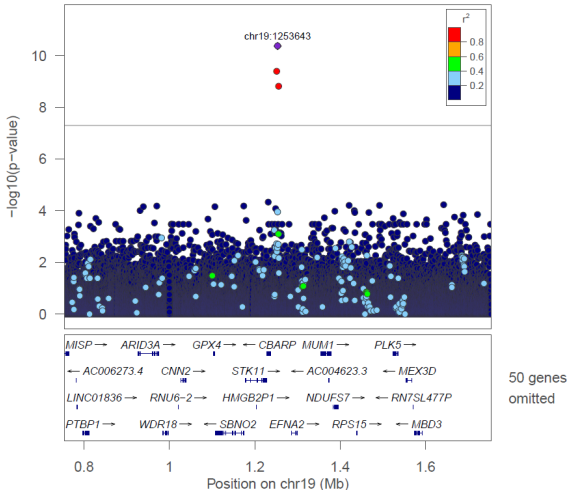
chr14:103044696 − LD: TOPMed − MAF: 0.000108 − MAC: 10



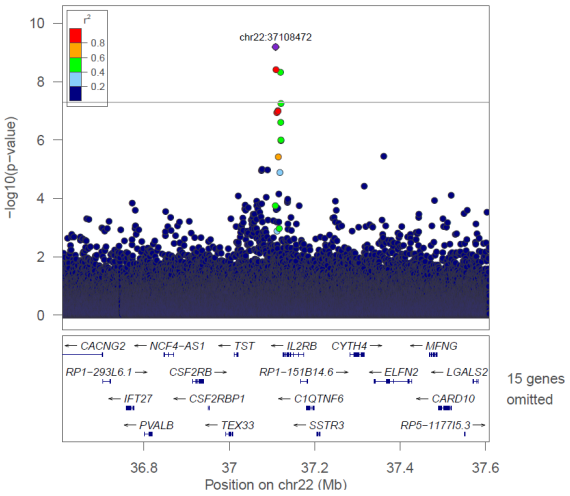
chr16:55649 − LD: TOPMed − MAF: 0.000224 − MAC: 17



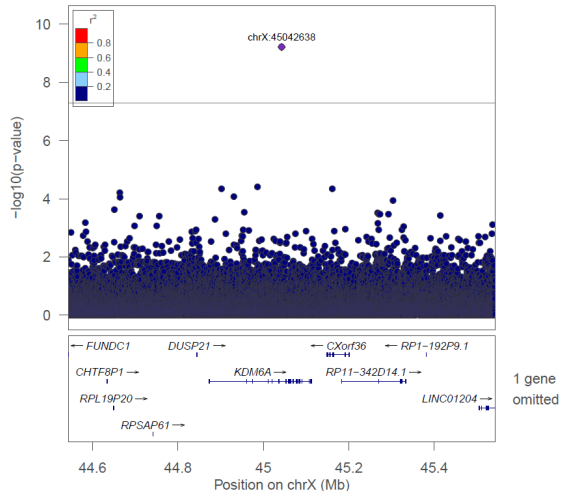
chr19:1253643 − LD: TOPMed − MAF: 0.165 − MAC: 15249



chr22:37108472 − LD: TOPMed − MAF: 0.11 − MAC: 10189

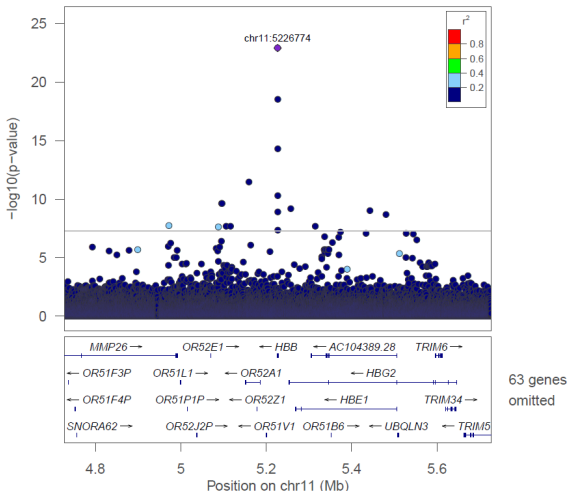


chrX:45042638 − LD: TOPMed − MAF: 0.382 − MAC: 27786

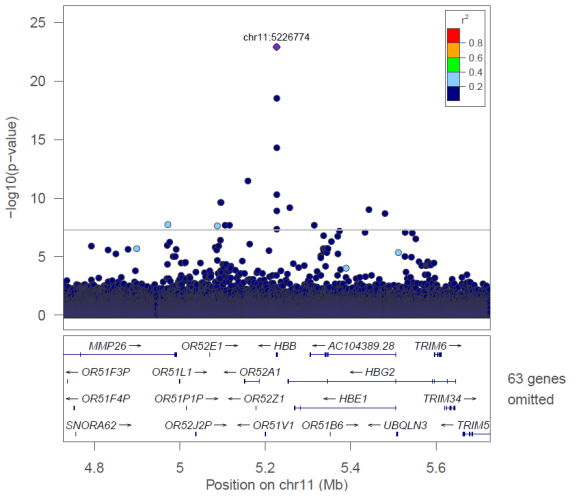


(D)

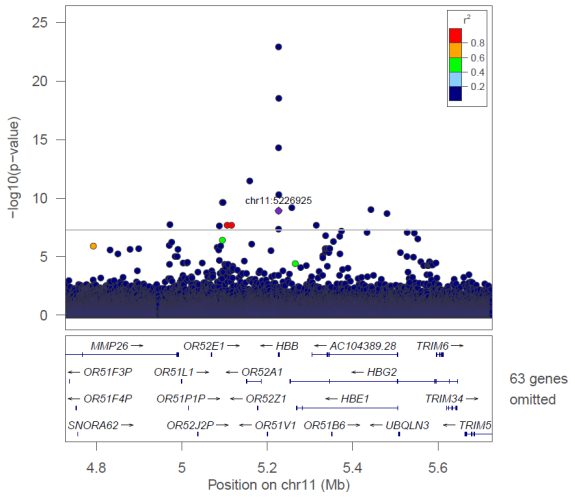
chr11:5226774 − LD: TOPMed − MAF: 0.000275 − MAC: 29



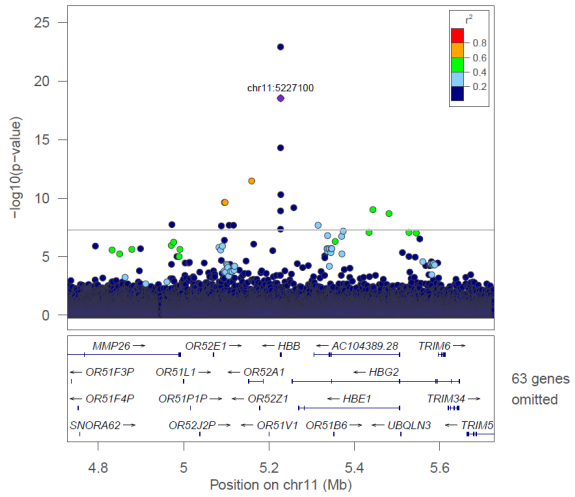
chr11:5226774 − LD: TOPMed − MAF: 0.000275 − MAC: 29



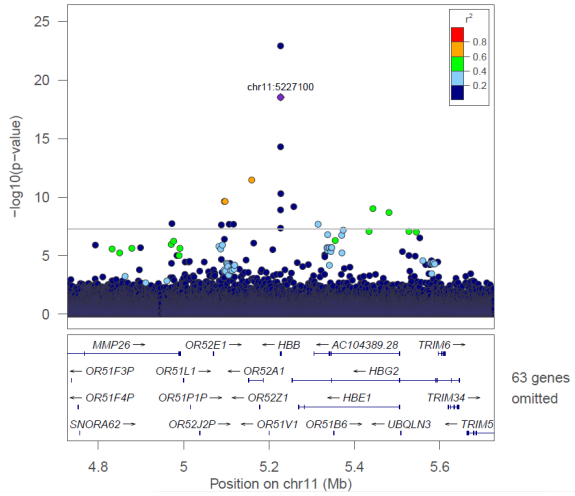
chr11:5226925 − LD: TOPMed − MAF: 9.5e−05 − MAC: 10



chr11:5227100 − LD: TOPMed − MAF: 0.000826 − MAC: 87

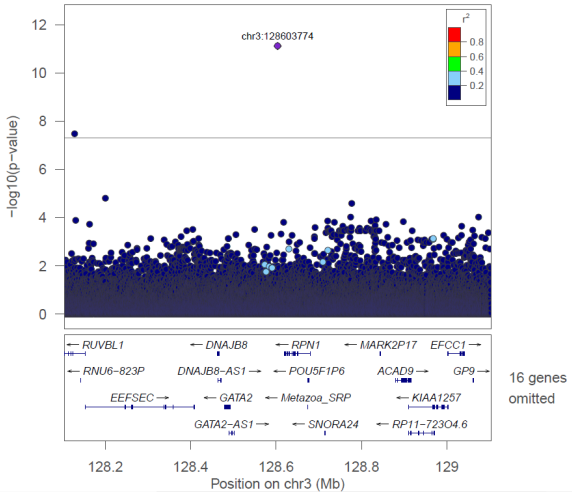


chr11:5227100 − LD: TOPMed − MAF: 0.000826 − MAC: 87

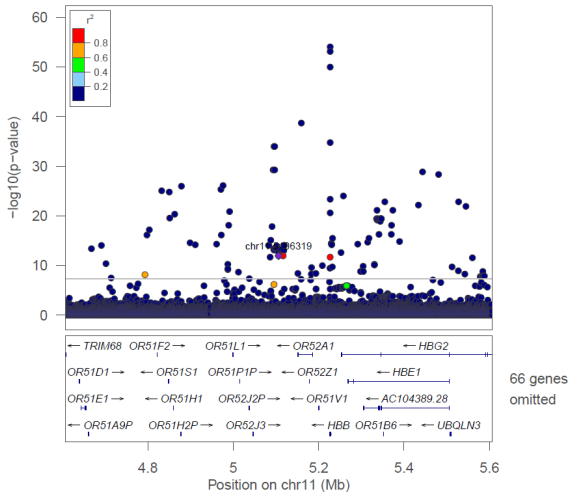


(E)

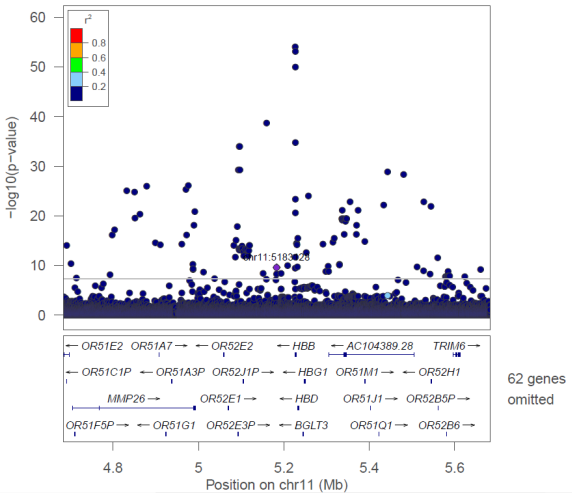
chr3:128603774 − LD: TOPMed − MAF: 0.00404 − MAC: 395



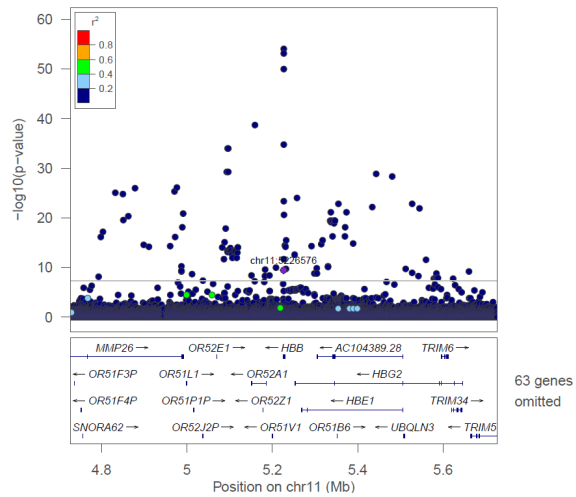
chr11:5106319 − LD: TOPMed − MAF: 0.000113 − MAC: 11



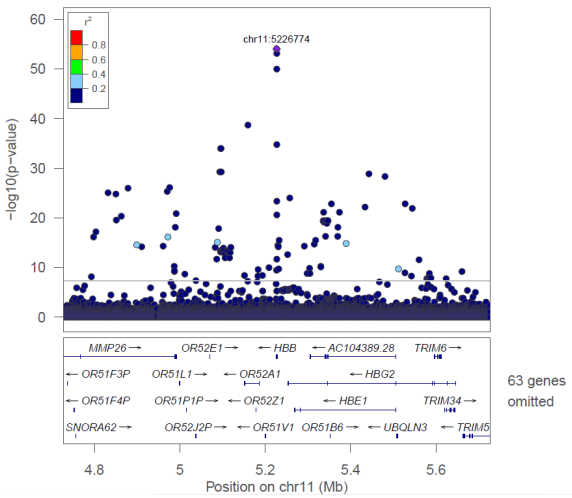
chr11:5183128 − LD: TOPMed − MAF: 5.12e−05 − MAC: 5



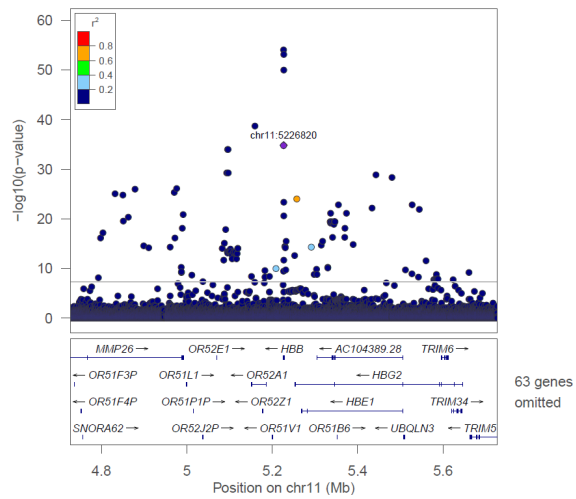
chr11:5226576 − LD: TOPMed − MAF: 7.17e−05 − MAC: 7



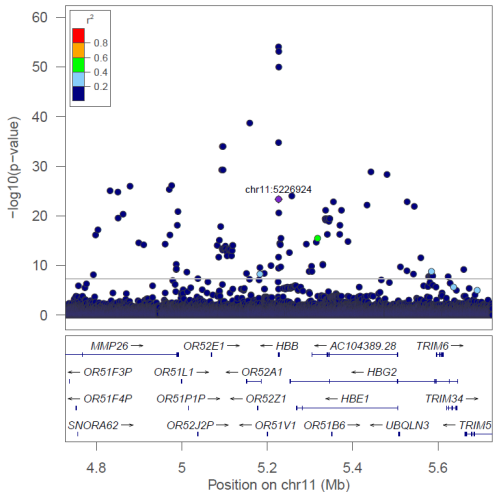
chr11:5226774 − LD: TOPMed − MAF: 0.000276 − MAC: 27



chr11:5226820 − LD: TOPMed − MAF: 0.000215 − MAC: 21

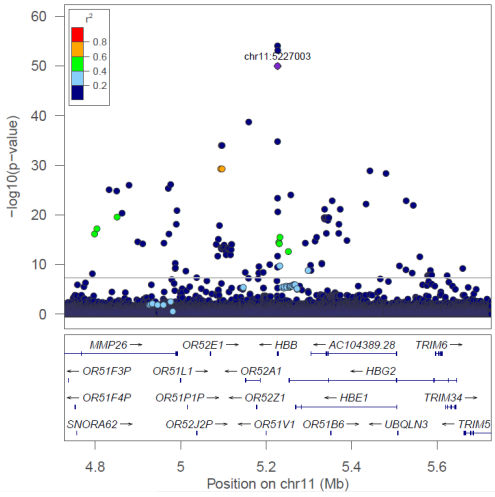


chr11:5226924 − LD: TOPMed − MAF: 0.000143 − MAC: 14



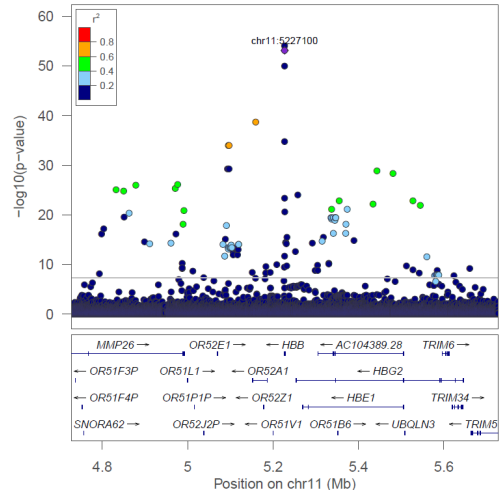
63 genes omitted

chr11:5227003 − LD: TOPMed − MAF: 0.00412 − MAC: 402



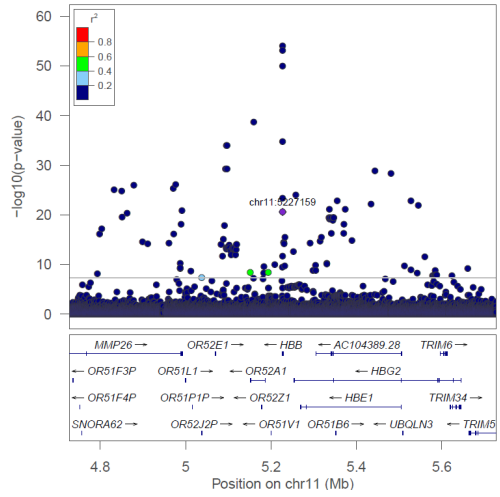
63 genes omitted

chr11:5227100 − LD: TOPMed − MAF: 0.00085 − MAC: 83



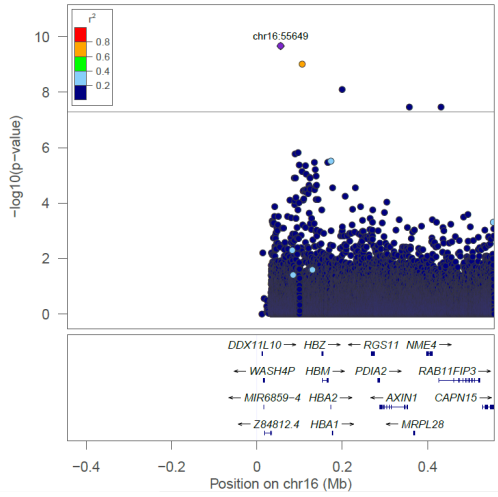
63 genes omitted

chr11:5227159 − LD: TOPMed − MAF: 0.000256 − MAC: 25



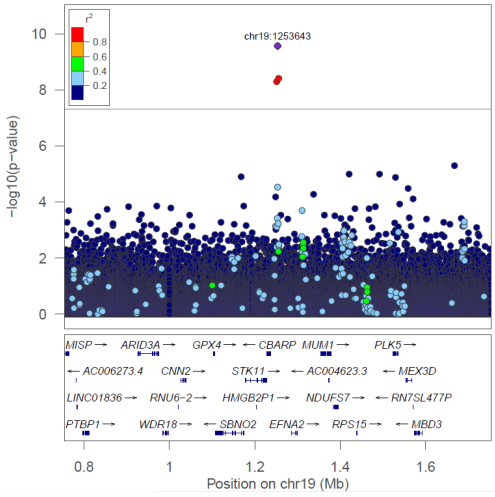
63 genes omitted

chr16:55649 − LD: TOPMed − MAF: 0.000217 − MAC: 17

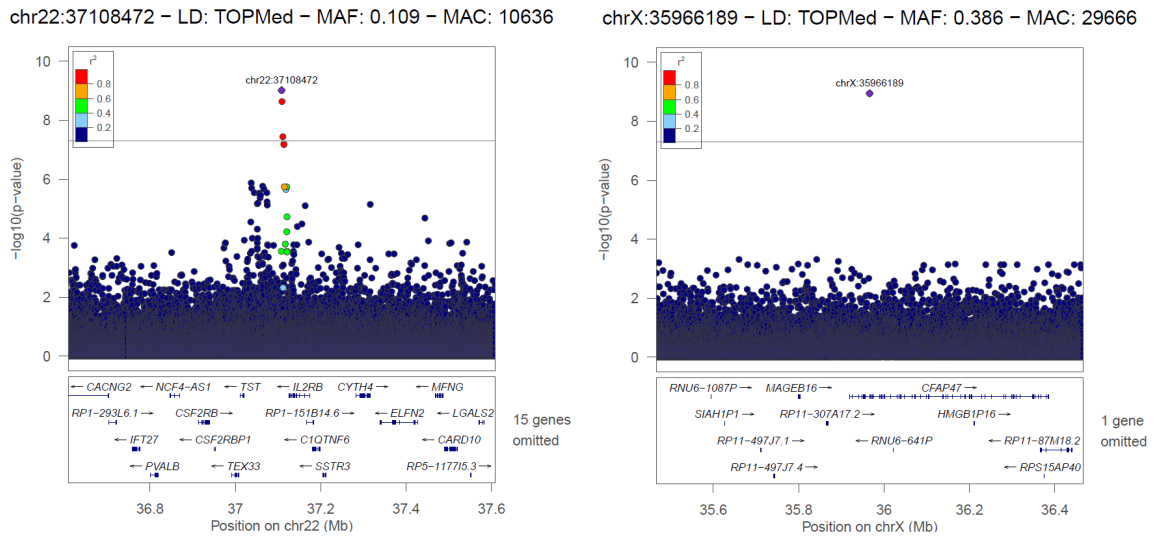


29 genes omitted

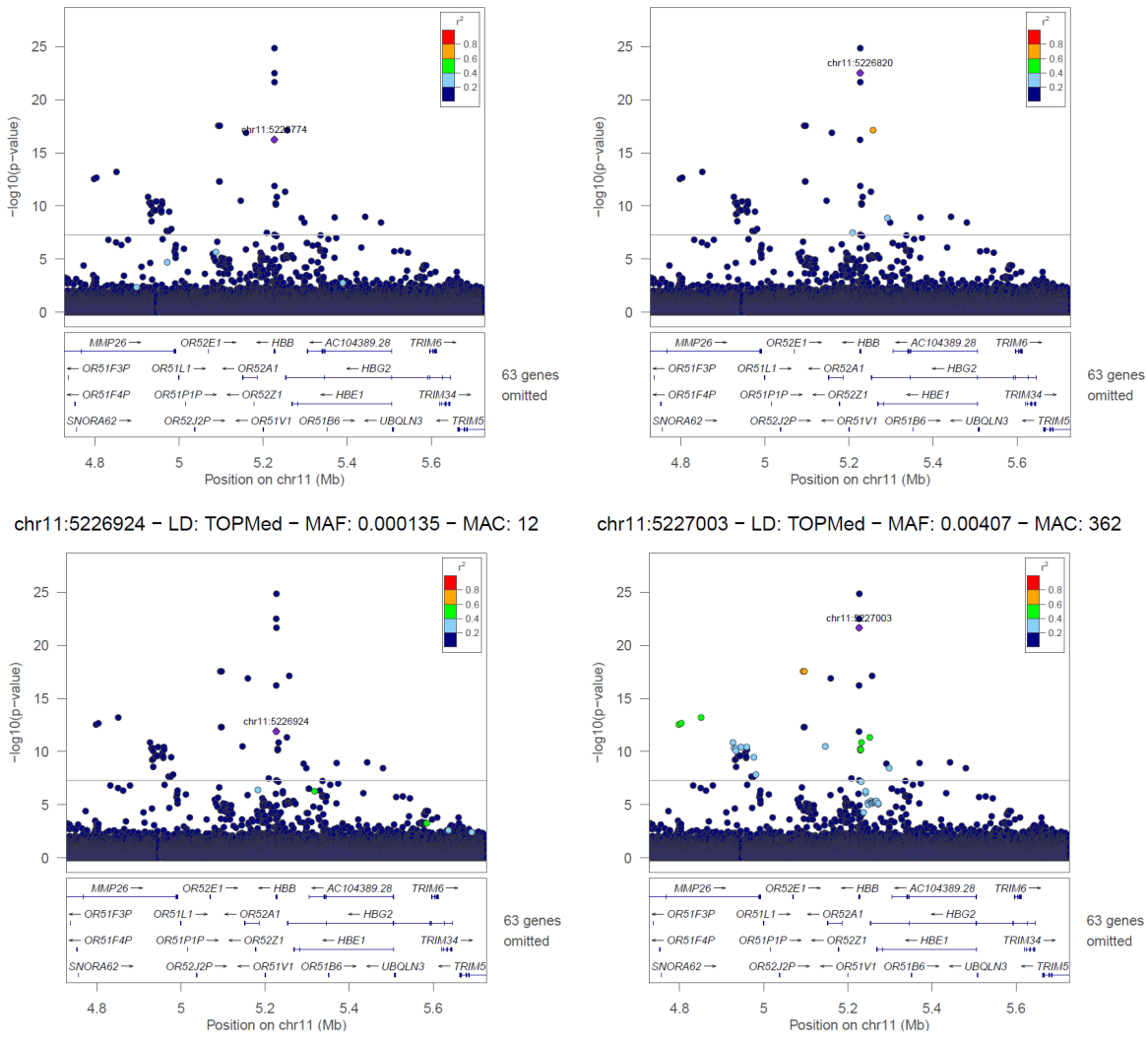
chr19:1253643 − LD: TOPMed − MAF: 0.167 − MAC: 16328

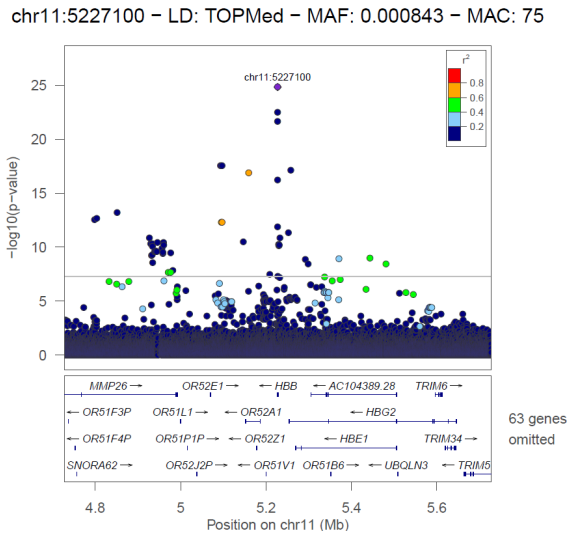


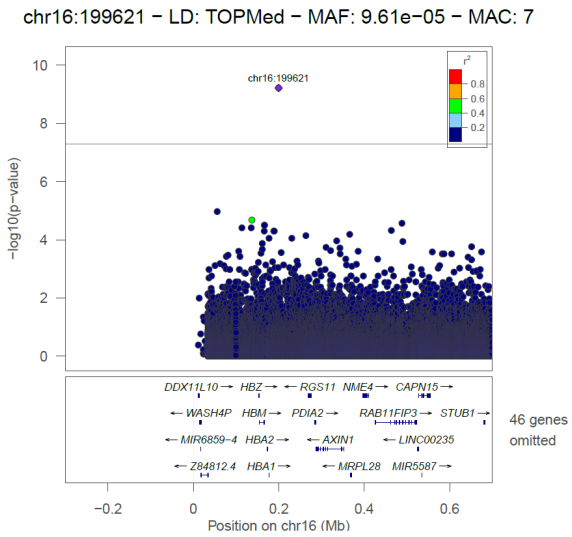
50 genes omitted


chr22:37108472 − LD: TOPMed − MAF: 0.109 − MAC: 10636
chrX:35966189 − LD: TOPMed − MAF: 0.386 − MAC: 29666
−log10(p−value)
r²
0.8
0.6
0.4
0.2
chr22:37108472
chrX:35966189
CACNG2 NCF4−AS1 TST IL2RB CYTH4 MFNG
RP1−293L6.1 CSF2RB RP1−151B14.6 ELFN2 LGALS2
IFT27 CSF2RBP1 C1QTNF6 CARD10
PVALB TEX33 SSTR3 RP5−1177I5.3
15 genes omitted
RNU6−1087P MAGEB16 CFAP47
SIAH1P1 RP11−307A17.2 HMGB1P16
RP11−497J7.1 RNU6−641P RP11−87M18.2
RP11−497J7.4 RPS15AP40
1 gene omitted
Position on chr22 (Mb)
Position on chrX (Mb)

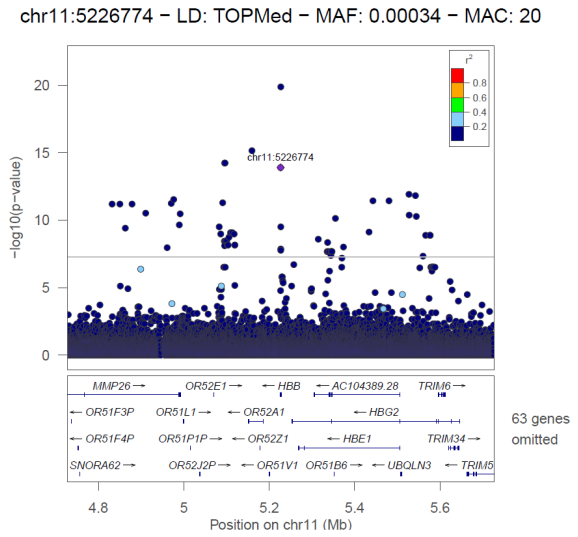
(F)

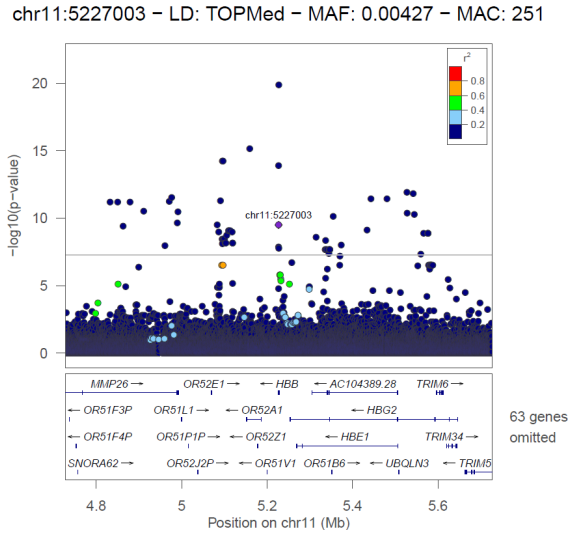

chr11:5226774 − LD: TOPMed − MAF: 0.000281 − MAC: 25
chr11:5226820 − LD: TOPMed − MAF: 0.000225 − MAC: 20
chr11:5226924 − LD: TOPMed − MAF: 0.000135 − MAC: 12
chr11:5227003 − LD: TOPMed − MAF: 0.00407 − MAC: 362
−log10(p−value)
r²
0.8
0.6
0.4
0.2
chr11:5226820
chr11:5226924
chr11:5227003
MMP26 OR52E1 HBB AC104389.28 TRIM6
OR51F3P OR51L1 OR52A1 HBG2
OR51F4P OR51P1P OR52Z1 HBE1 TRIM34
SNORA62 OR52J2P OR51V1 OR51B6 UBQLN3 TRIM5
63 genes omitted
Position on chr11 (Mb)

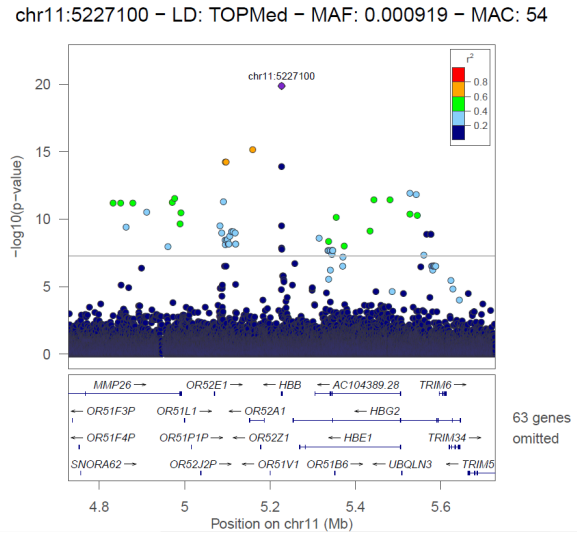

chr11:5227100 − LD: TOPMed − MAF: 0.000843 − MAC: 75
chr11:5227100
−log10(p−value)
Position on chr11 (Mb)
63 genes omitted


chr16:199621 − LD: TOPMed − MAF: 9.61e−05 − MAC: 7
chr16:199621
−log10(p−value)
Position on chr16 (Mb)
46 genes omitted

(G)


chr11:5226774 − LD: TOPMed − MAF: 0.00034 − MAC: 20
chr11:5226774
−log10(p−value)
Position on chr11 (Mb)
63 genes omitted


chr11:5227003 − LD: TOPMed − MAF: 0.00427 − MAC: 251
chr11:5227003
−log10(p−value)
Position on chr11 (Mb)
63 genes omitted


chr11:5227100 − LD: TOPMed − MAF: 0.000919 − MAC: 54
chr11:5227100
−log10(p−value)
Position on chr11 (Mb)
63 genes omitted

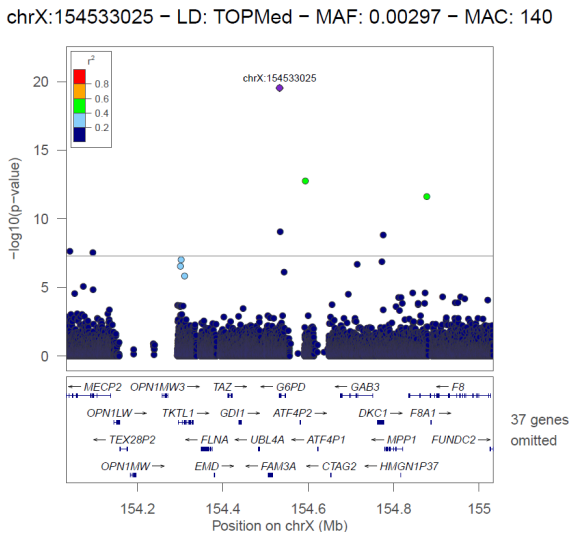
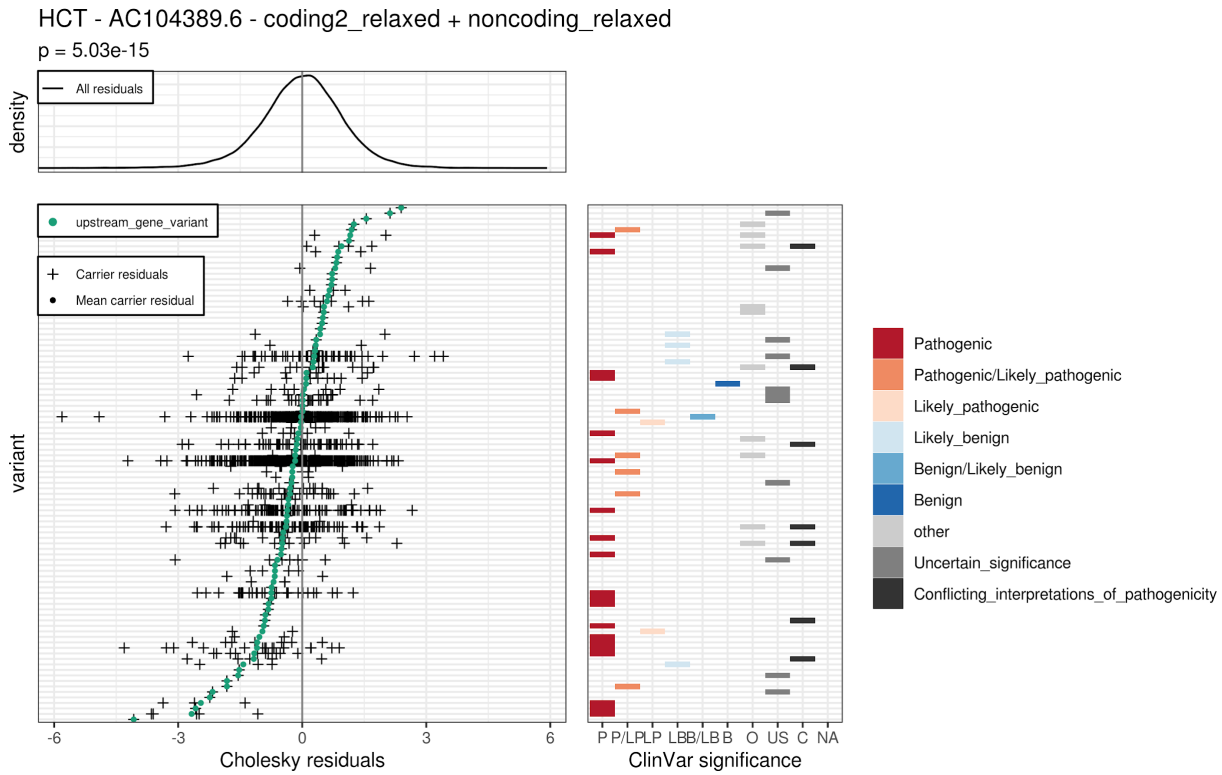

chrX:154533025 − LD: TOPMed − MAF: 0.00297 − MAC: 140
chrX:154533025
−log10(p−value)
Position on chrX (Mb)
37 genes omitted

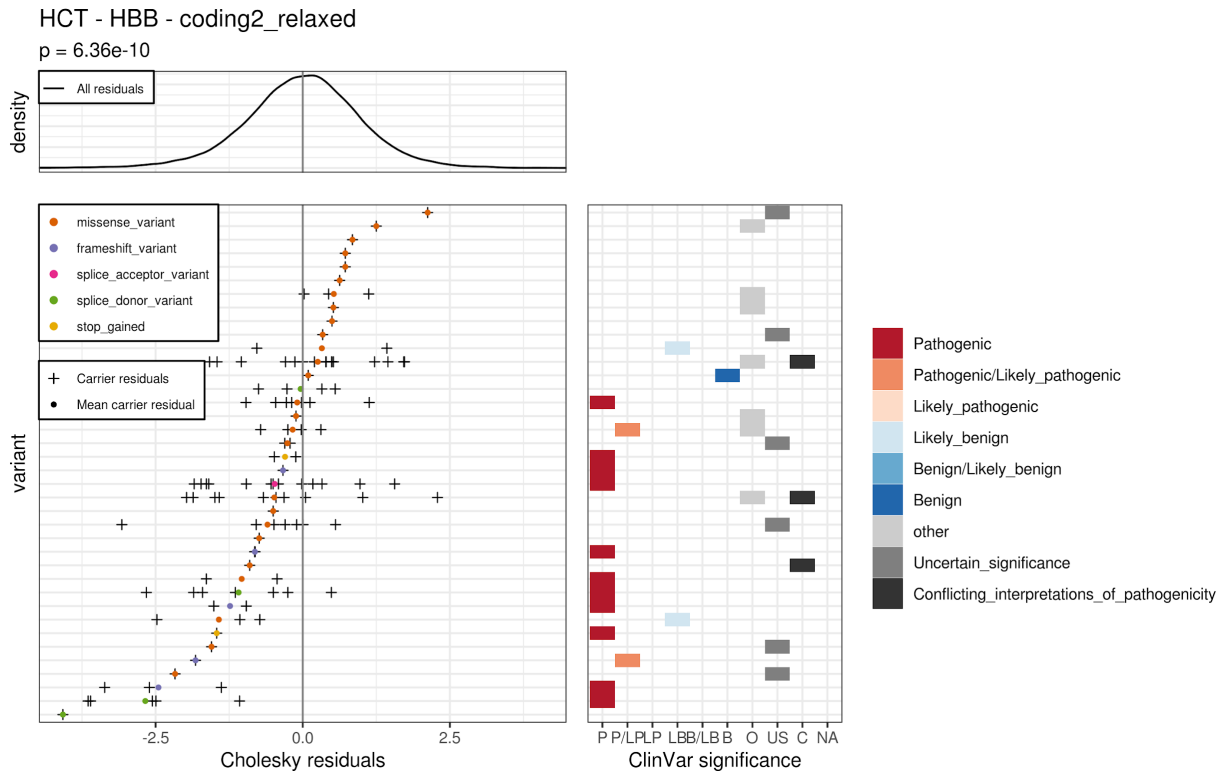
Figure S5. Rare variants identified in the aggregated analysis in TOPMed. (A) HCT; (B) HGB; (C) MCH; (D) MCHC; (E) MCV; (F) RBC; (G) RDW.

(A)

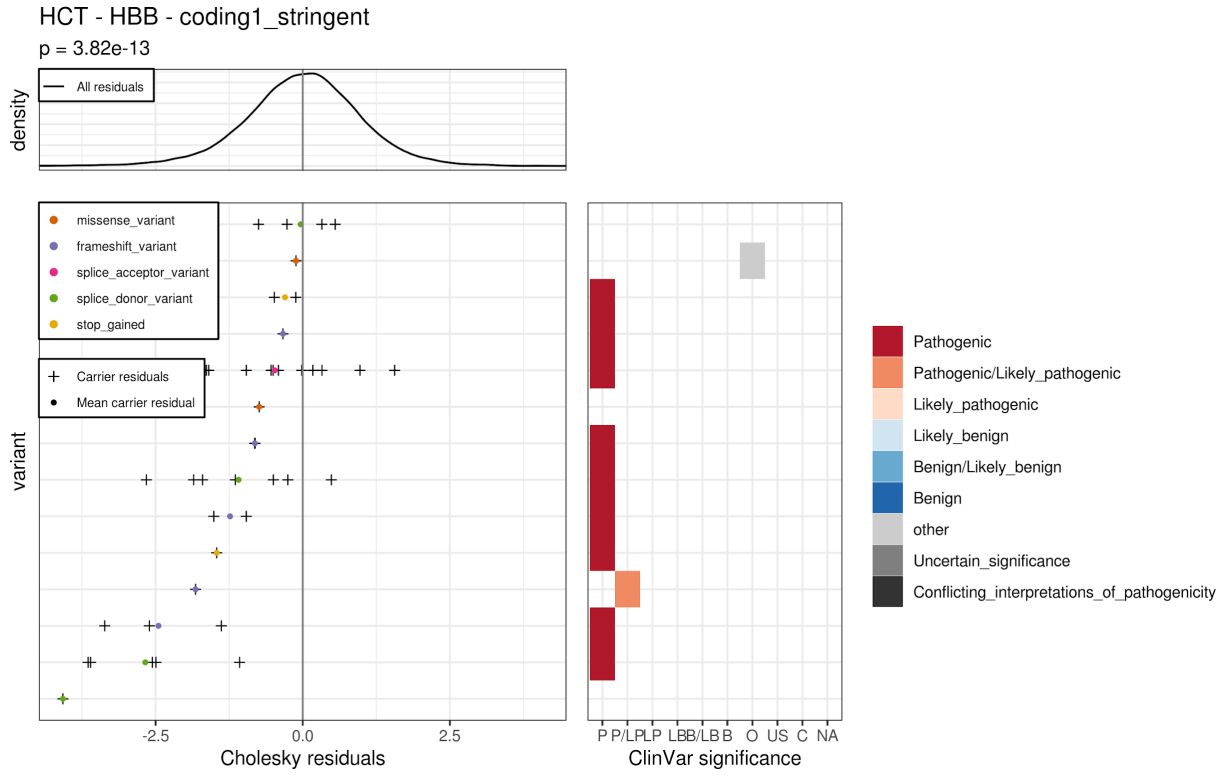
(A1)



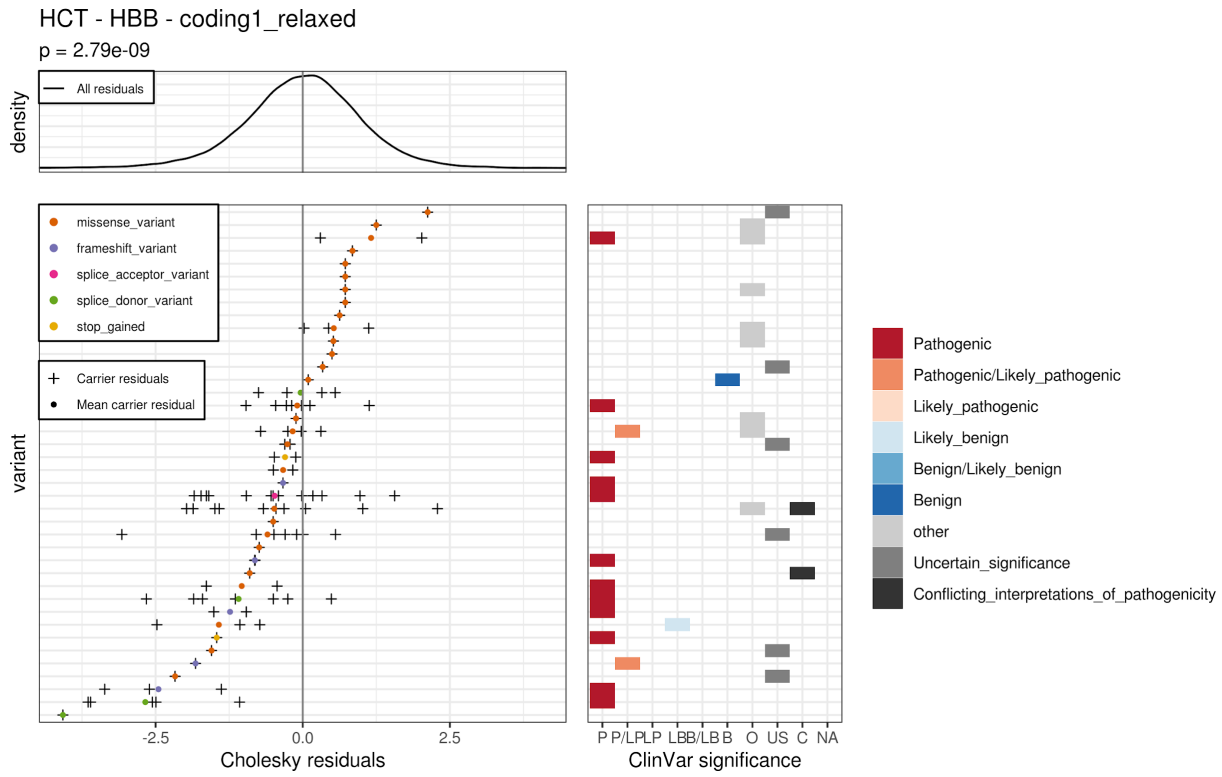
(A2)


HCT - HBB - coding2_relaxed
p = 6.36e-10
density
All residuals
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-2.5
0.0
2.5
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

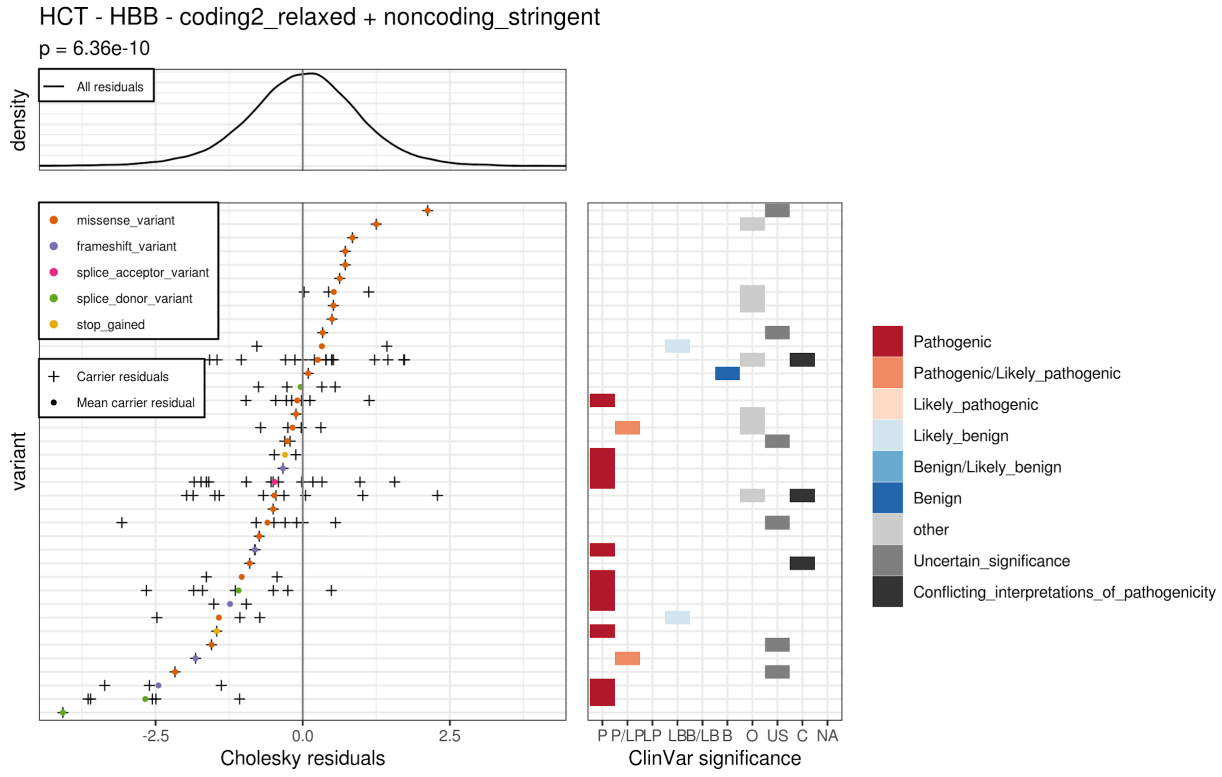
(A3)


HCT - HBB - coding1_stringent
p = 3.82e-13
density
All residuals
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-2.5
0.0
2.5
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(A4)

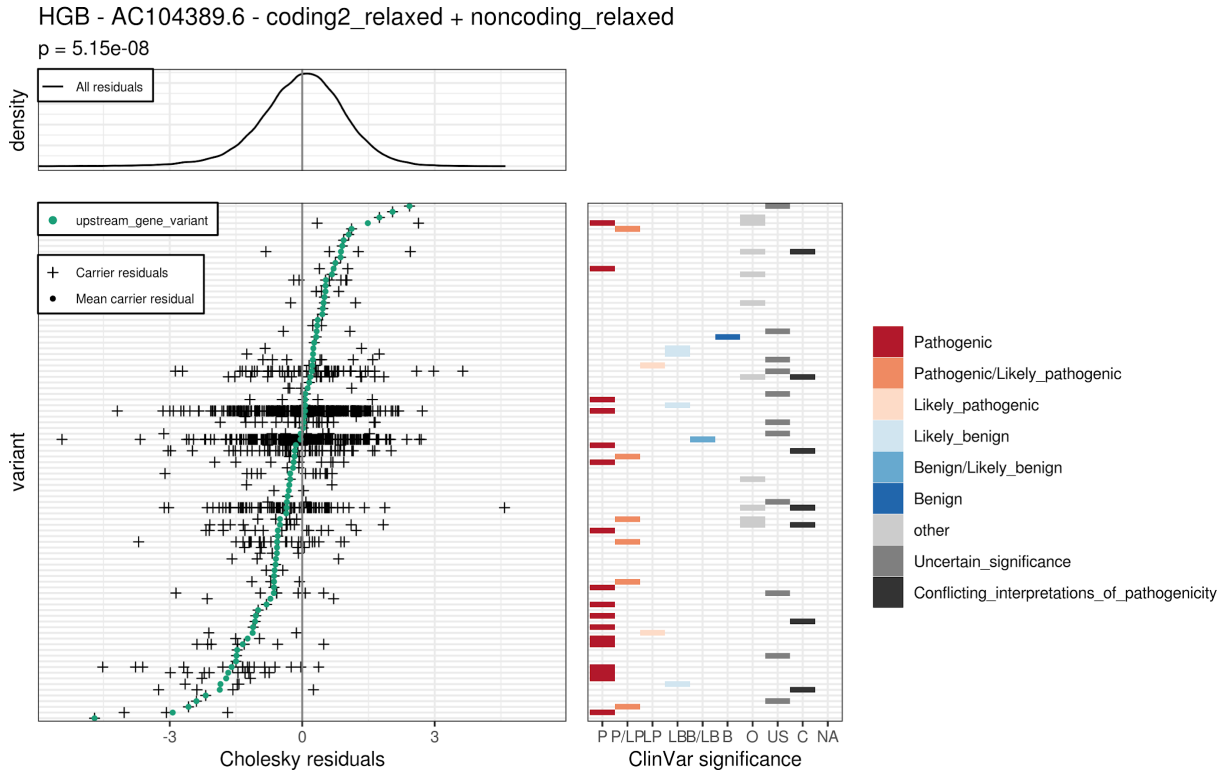

HCT - HBB - coding1_relaxed
p = 2.79e-09
density
All residuals
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-2.5
0.0
2.5
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(A5)

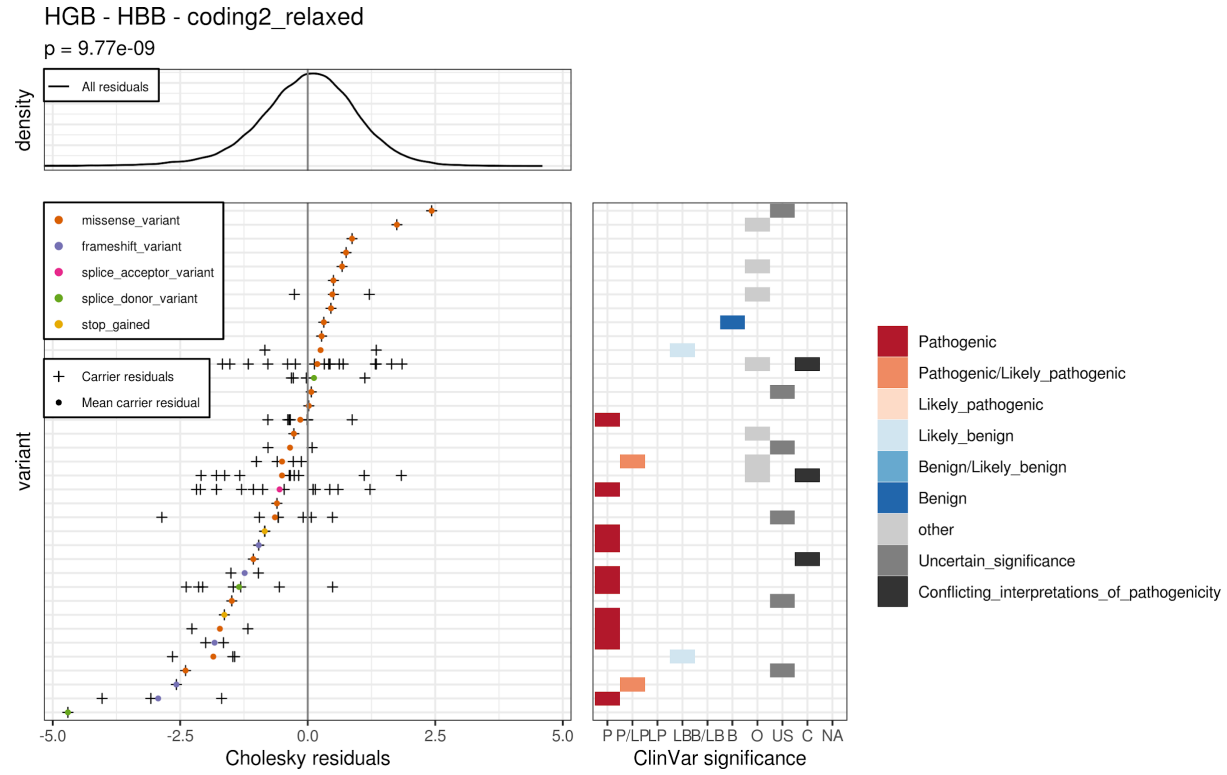

HCT - HBB - coding2_relaxed + noncoding_stringent
p = 6.36e-10
density
All residuals
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-2.5
0.0
2.5
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(B)

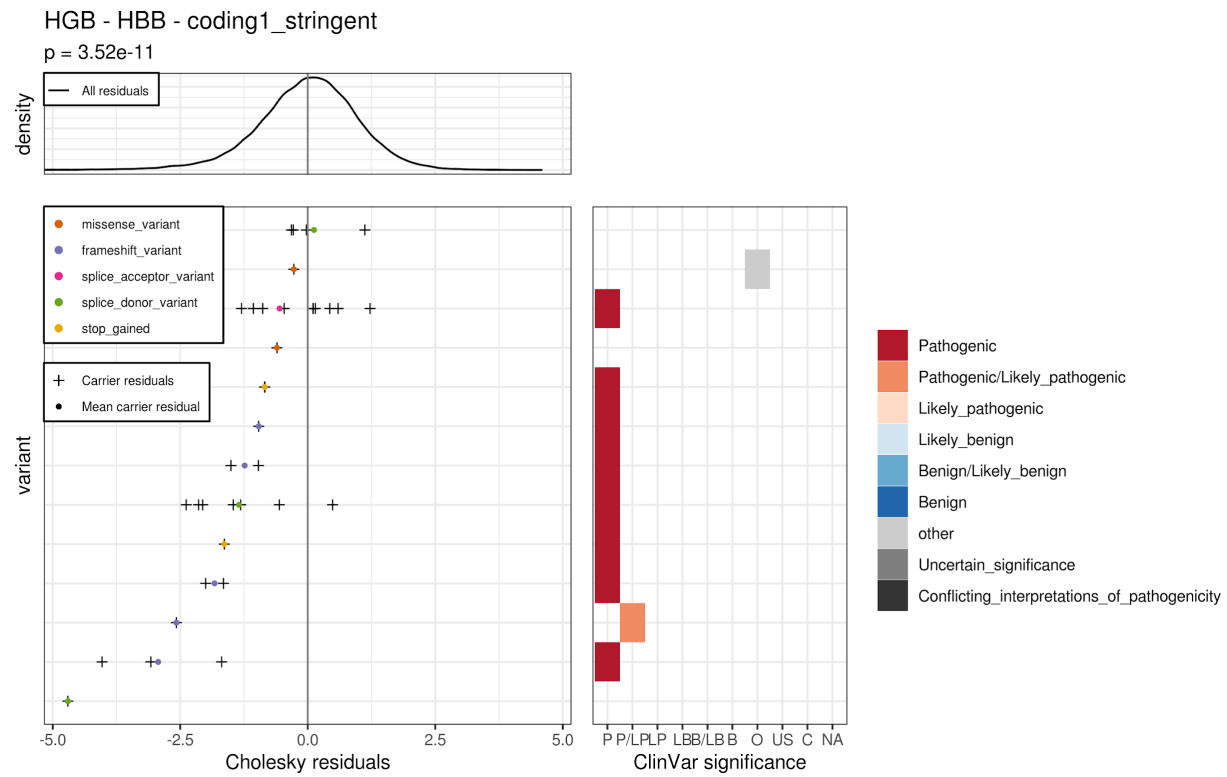
(B1)


HGB - AC104389.6 - coding2_relaxed + noncoding_relaxed
p = 5.15e-08
All residuals
density
upstream_gene_variant
Carrier residuals
Mean carrier residual
variant
-3
0
3
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

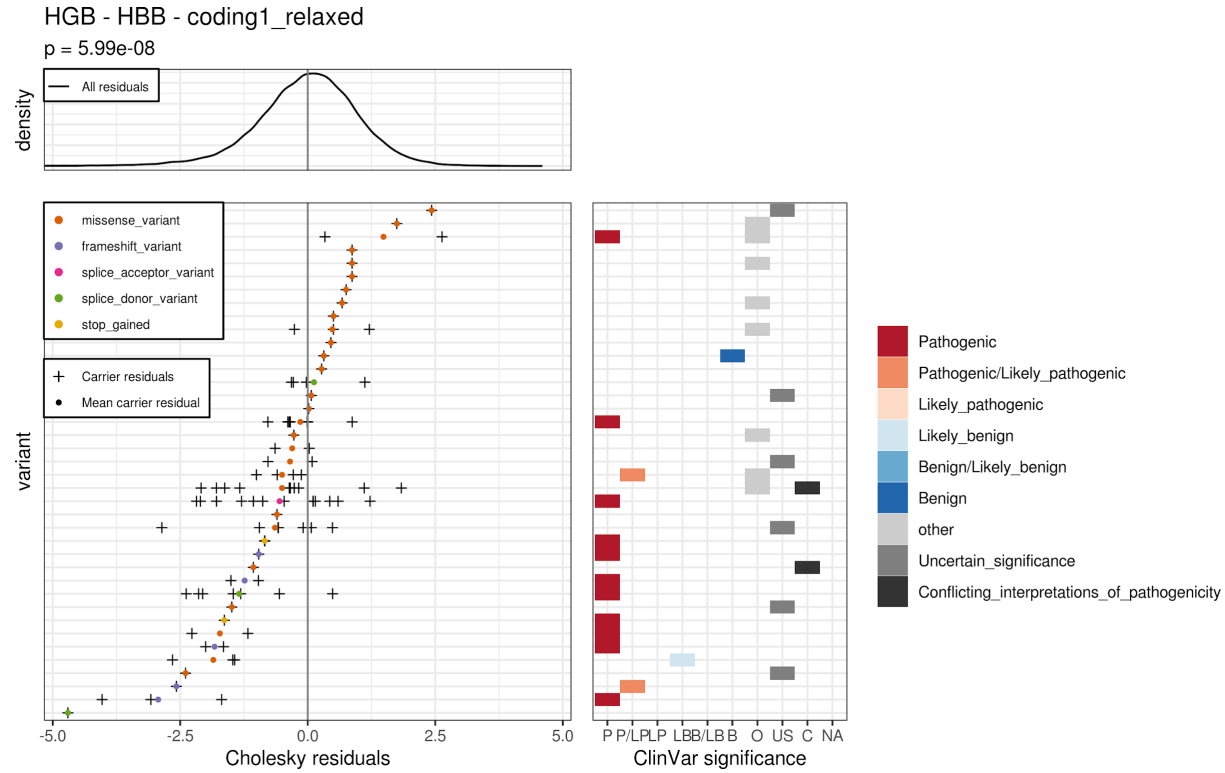
(B2)


HGB - HBB - coding2_relaxed
p = 9.77e-09
density
All residuals
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-5.0
-2.5
0.0
2.5
5.0
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

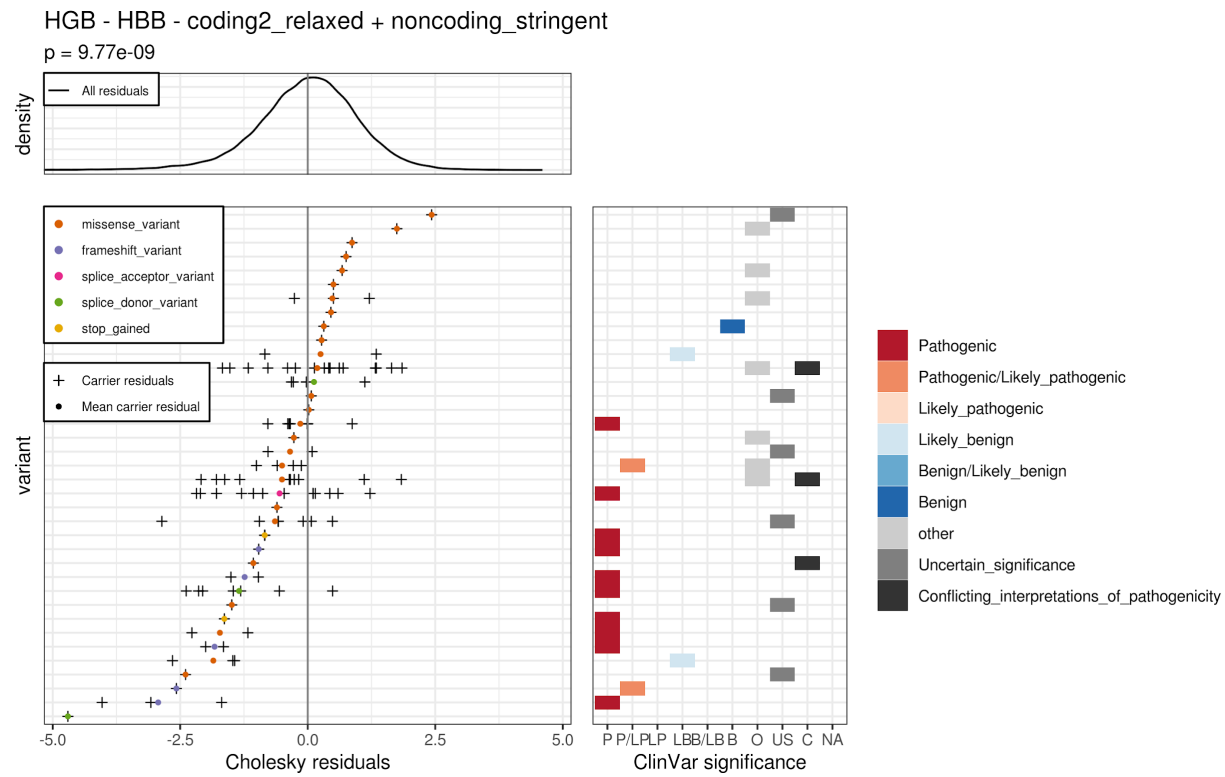
(B3)


HGB - HBB - coding1_stringent
p = 3.52e-11
density
All residuals
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-5.0
-2.5
0.0
2.5
5.0
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(B4)

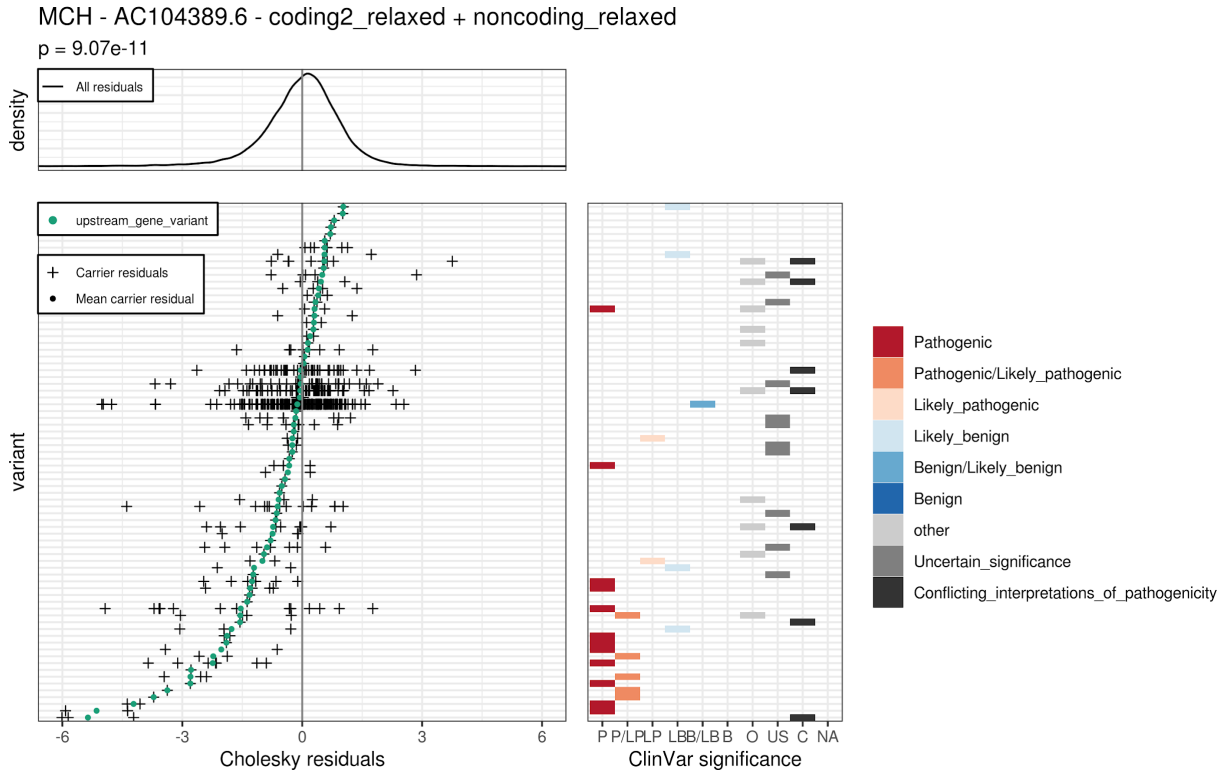

HGB - HBB - coding1_relaxed
p = 5.99e-08
density
All residuals
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-5.0
-2.5
0.0
2.5
5.0
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(B5)

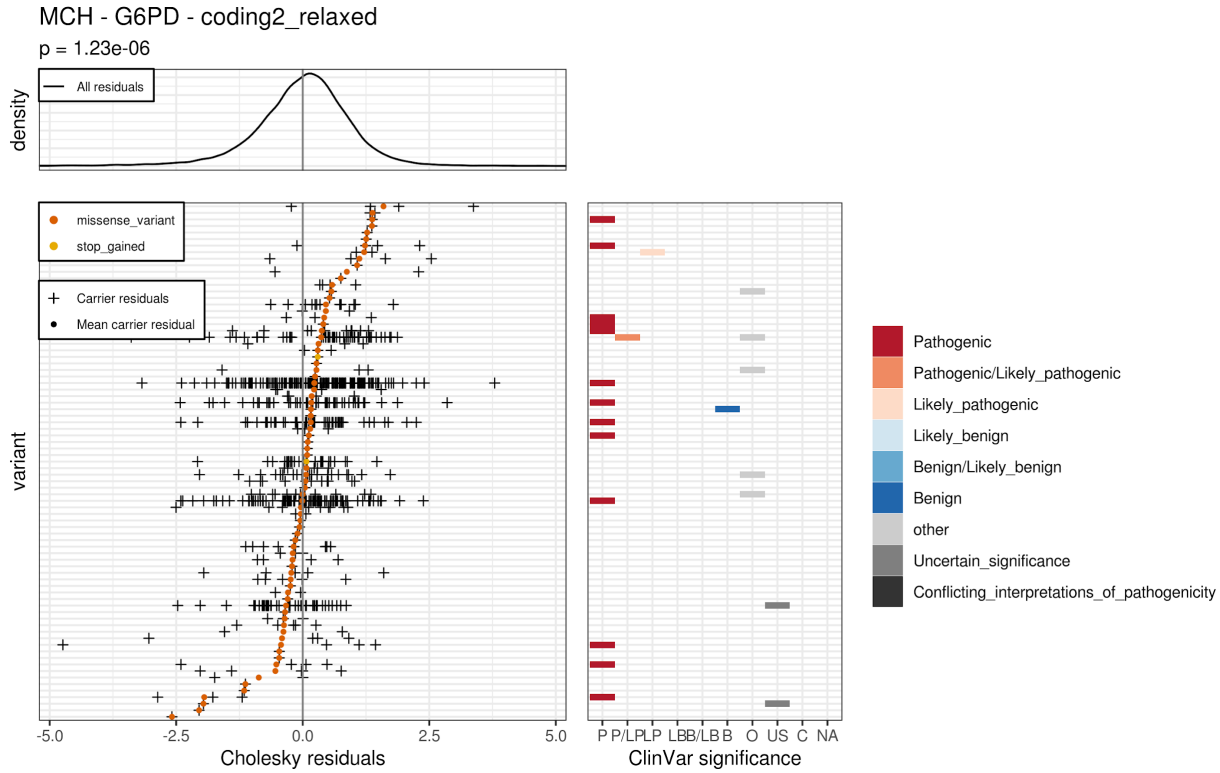

HGB - HBB - coding2_relaxed + noncoding_stringent
p = 9.77e-09
density
All residuals
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-5.0
-2.5
0.0
2.5
5.0
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(C)

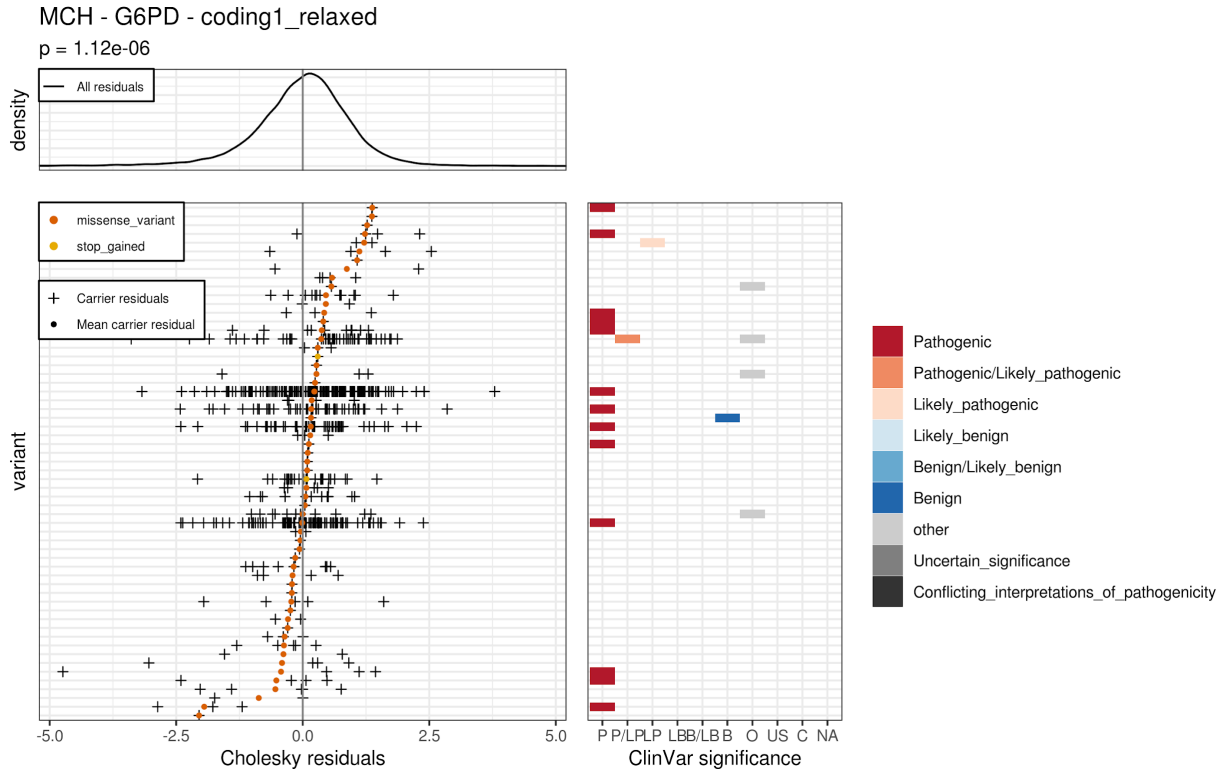
(C1)


MCH - AC104389.6 - coding2_relaxed + noncoding_relaxed
p = 9.07e-11
density
All residuals
variant
upstream_gene_variant
Carrier residuals
Mean carrier residual
-6
-3
0
3
6
Cholesky residuals
P
P/LP
LP
LB
B/LB
B
O
US
C
NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

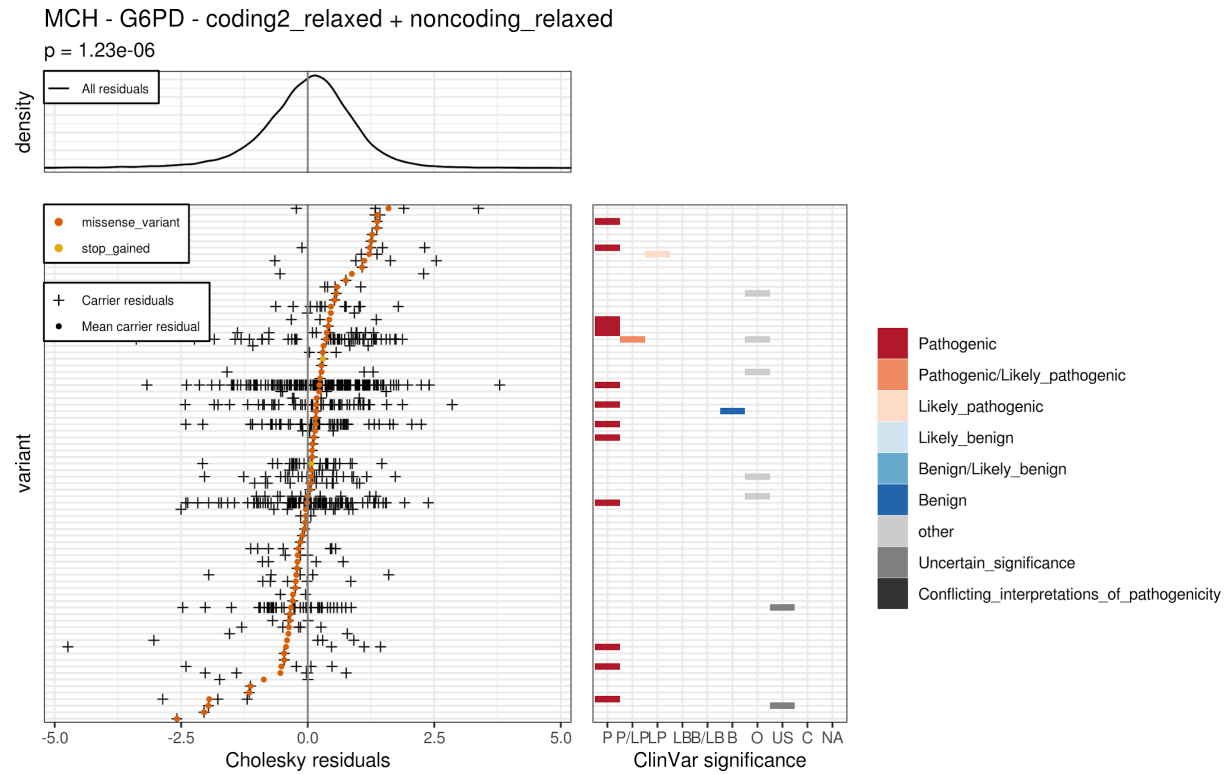
(C2)


MCH - G6PD - coding2_relaxed
p = 1.23e-06
density
All residuals
missense_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-5.0
-2.5
0.0
2.5
5.0
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

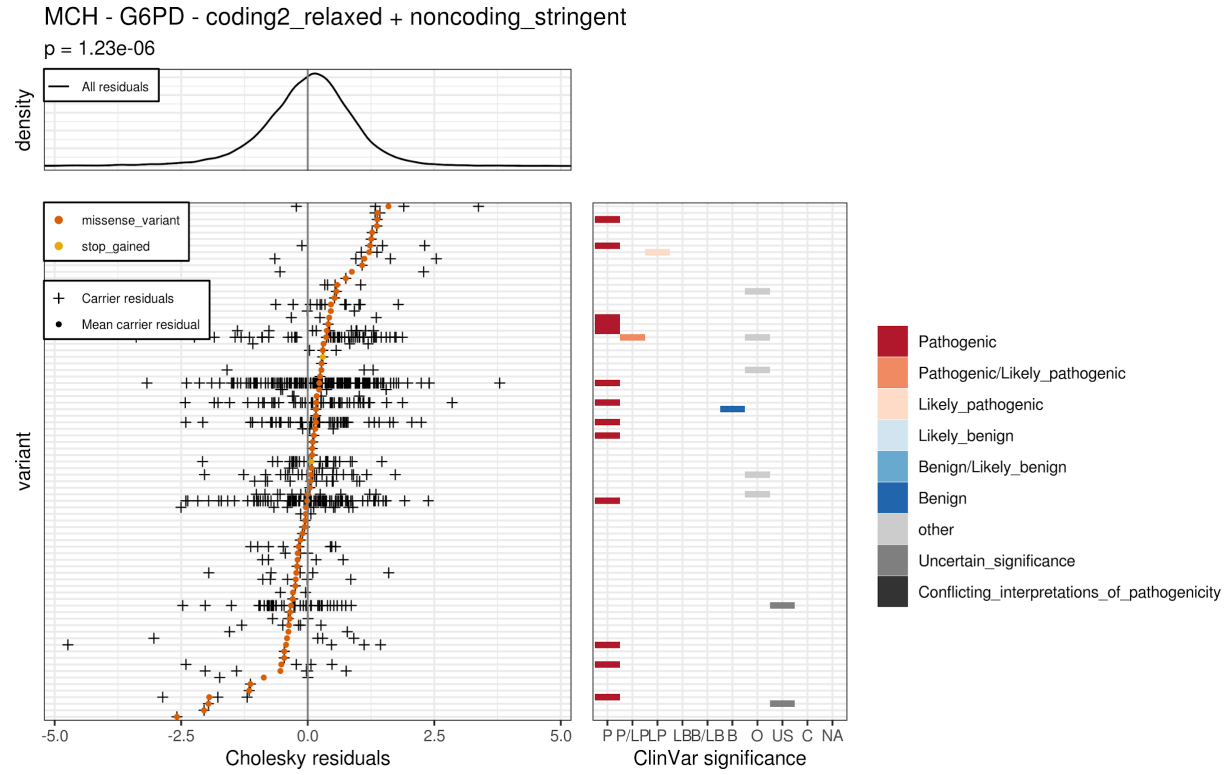
(C3)


MCH - G6PD - coding1_relaxed
p = 1.12e-06
density
All residuals
missense_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-5.0
-2.5
0.0
2.5
5.0
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

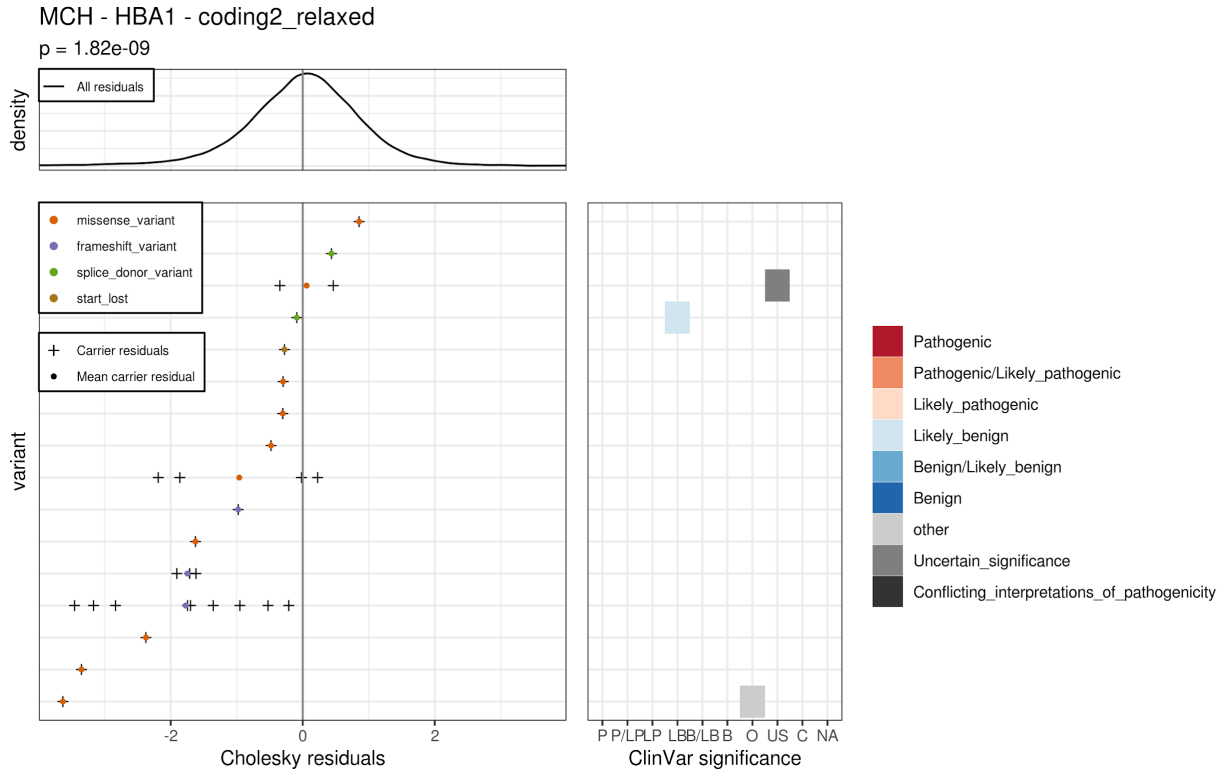
(C4)


MCH - G6PD - coding2_relaxed + noncoding_relaxed
p = 1.23e-06
density
All residuals
missense_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-5.0
-2.5
0.0
2.5
5.0
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

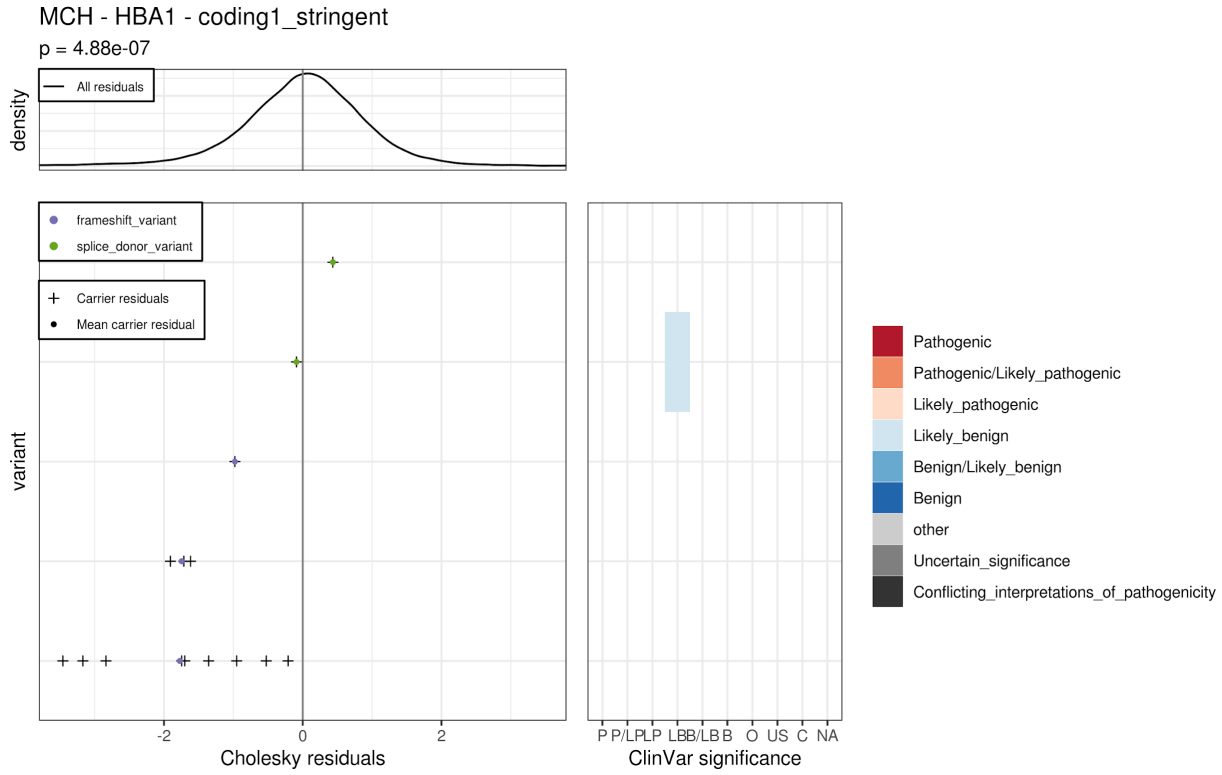
(C5)


MCH - G6PD - coding2_relaxed + noncoding_stringent
p = 1.23e-06
density
All residuals
missense_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-5.0
-2.5
0.0
2.5
5.0
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

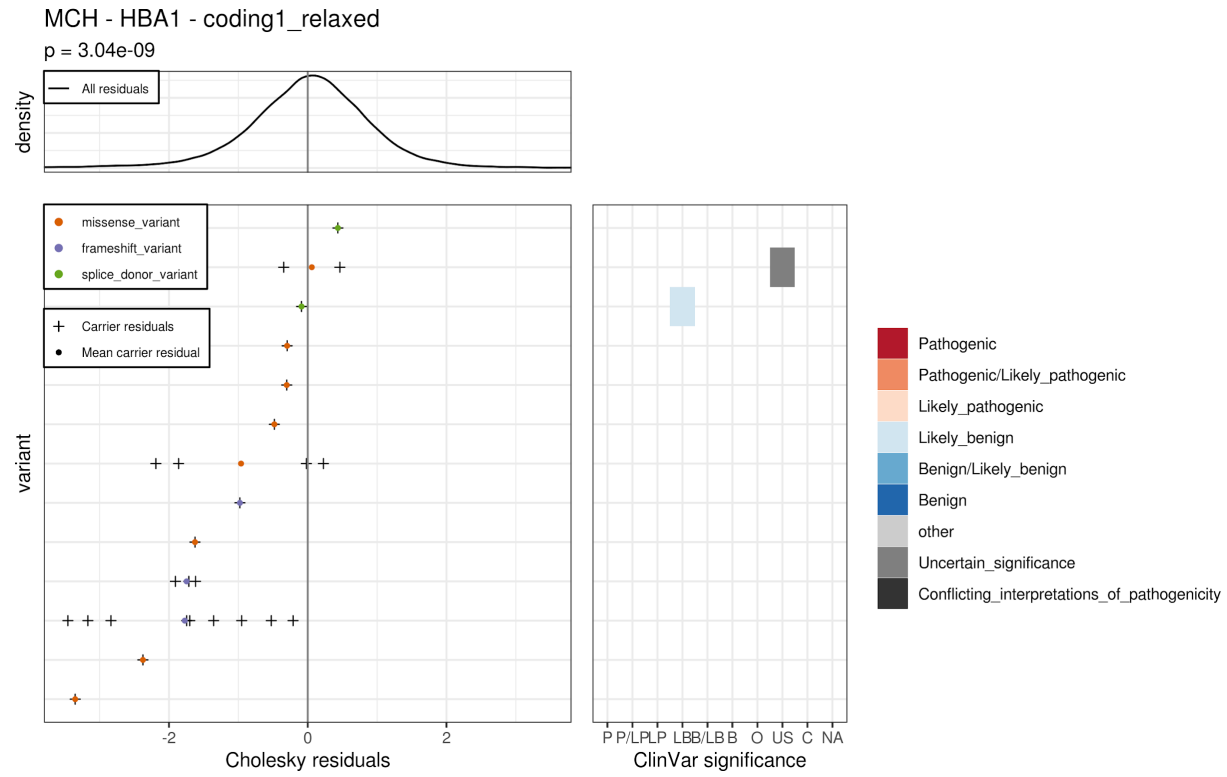
(C6)


MCH - HBA1 - coding2_relaxed
p = 1.82e-09
density
All residuals
missense_variant
frameshift_variant
splice_donor_variant
start_lost
Carrier residuals
Mean carrier residual
variant
-2
0
2
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

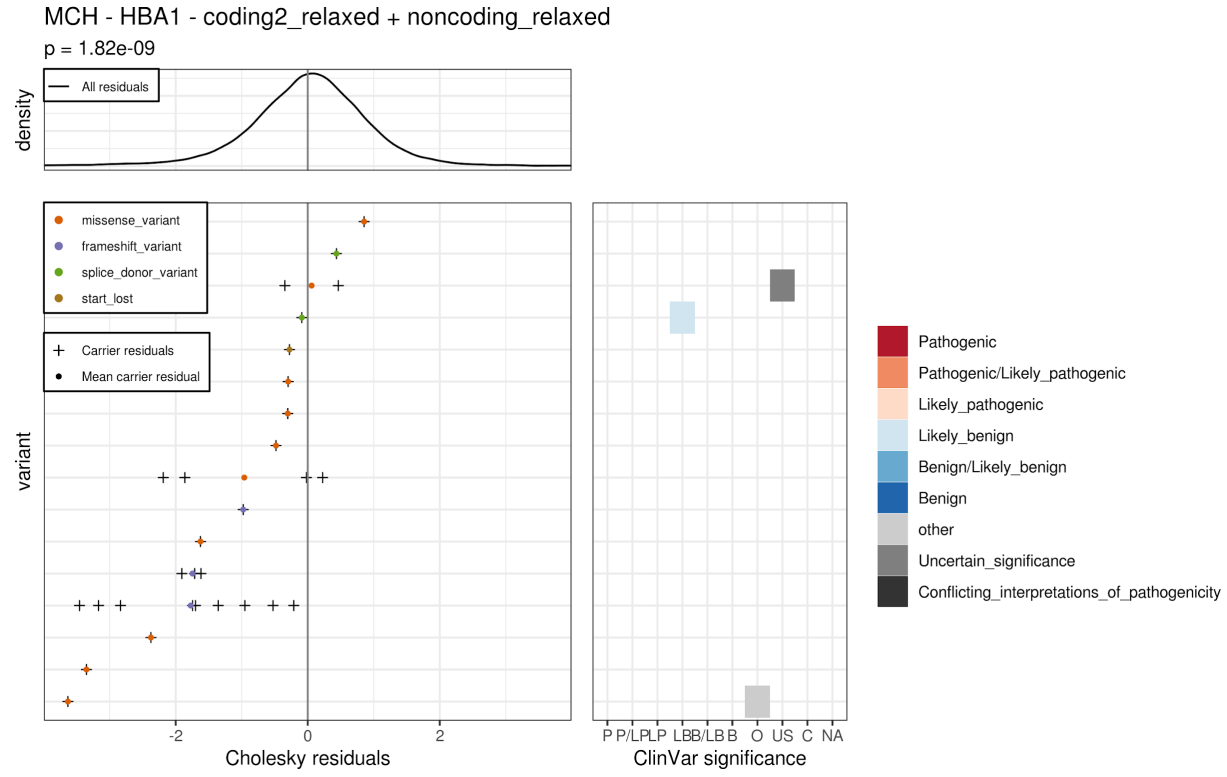
(C7)


MCH - HBA1 - coding1_stringent
p = 4.88e-07
density
All residuals
frameshift_variant
splice_donor_variant
Carrier residuals
Mean carrier residual
variant
-2
0
2
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

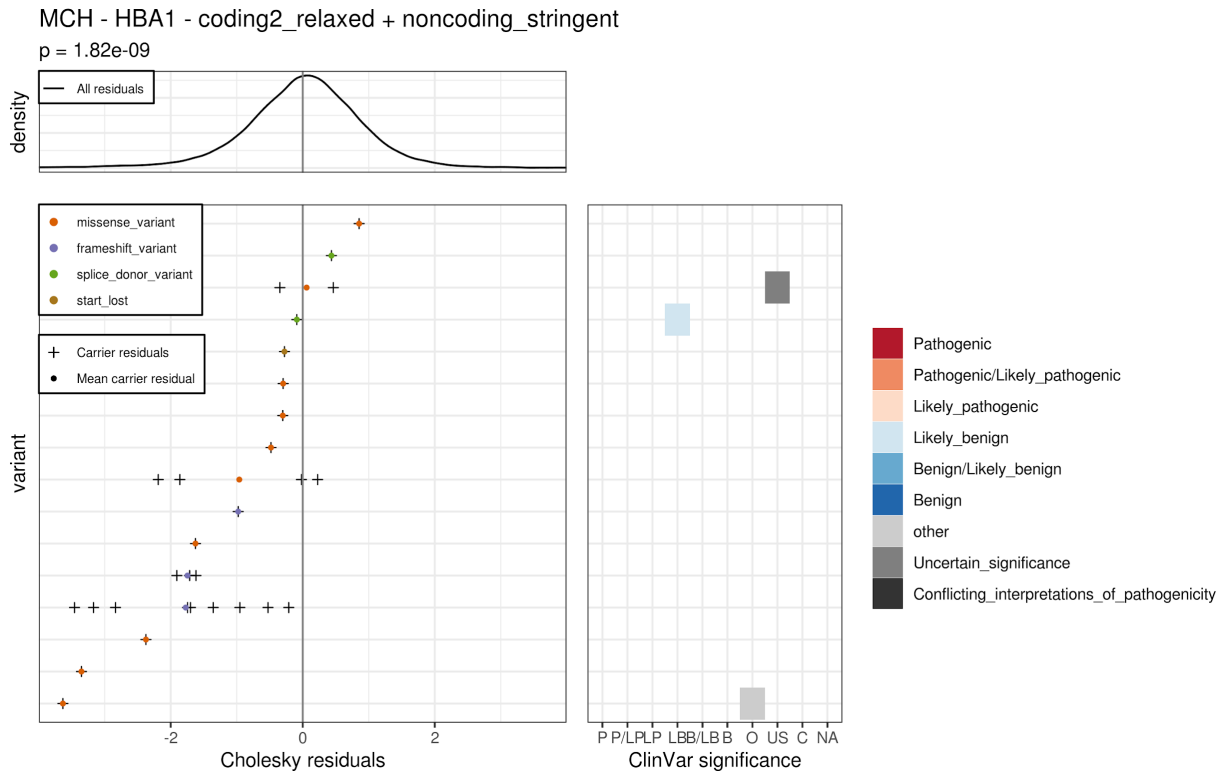
(C8)


MCH - HBA1 - coding1_relaxed
p = 3.04e-09
density
All residuals
missense_variant
frameshift_variant
splice_donor_variant
Carrier residuals
Mean carrier residual
variant
-2
0
2
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

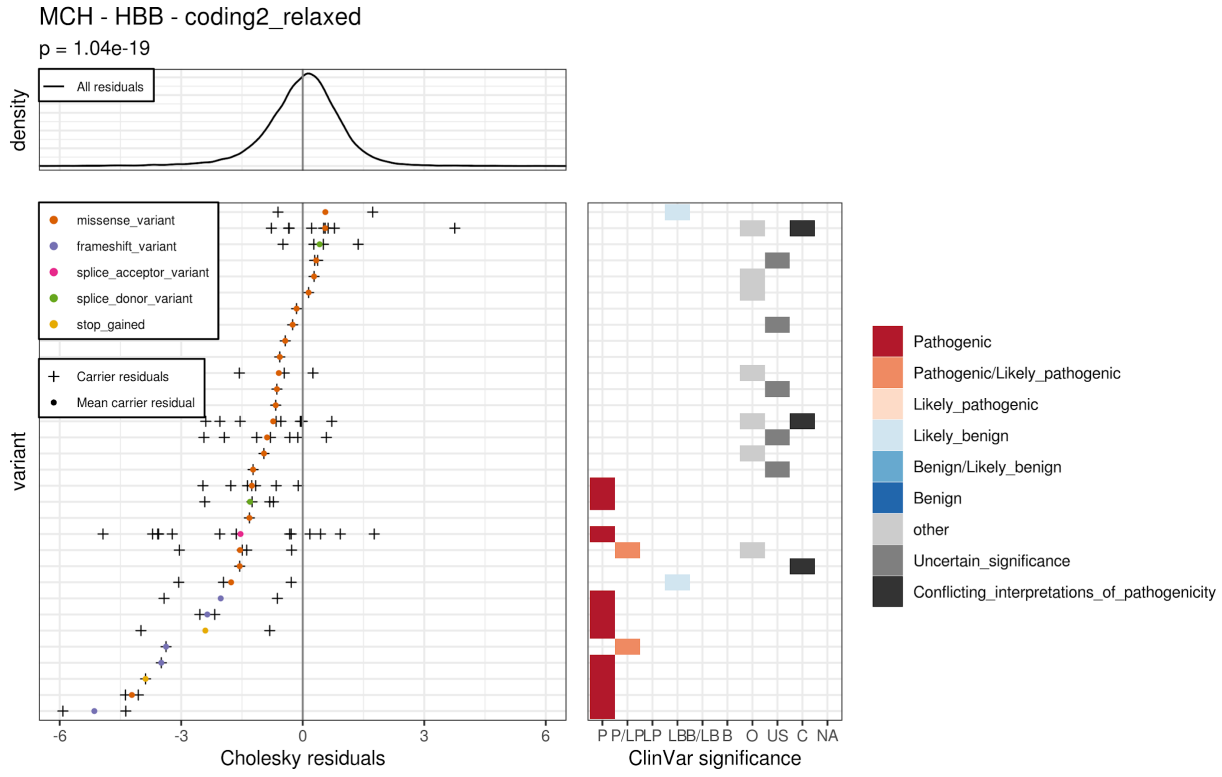
(C9)


MCH - HBA1 - coding2_relaxed + noncoding_relaxed
p = 1.82e-09
density
All residuals
missense_variant
frameshift_variant
splice_donor_variant
start_lost
Carrier residuals
Mean carrier residual
variant
-2
0
2
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

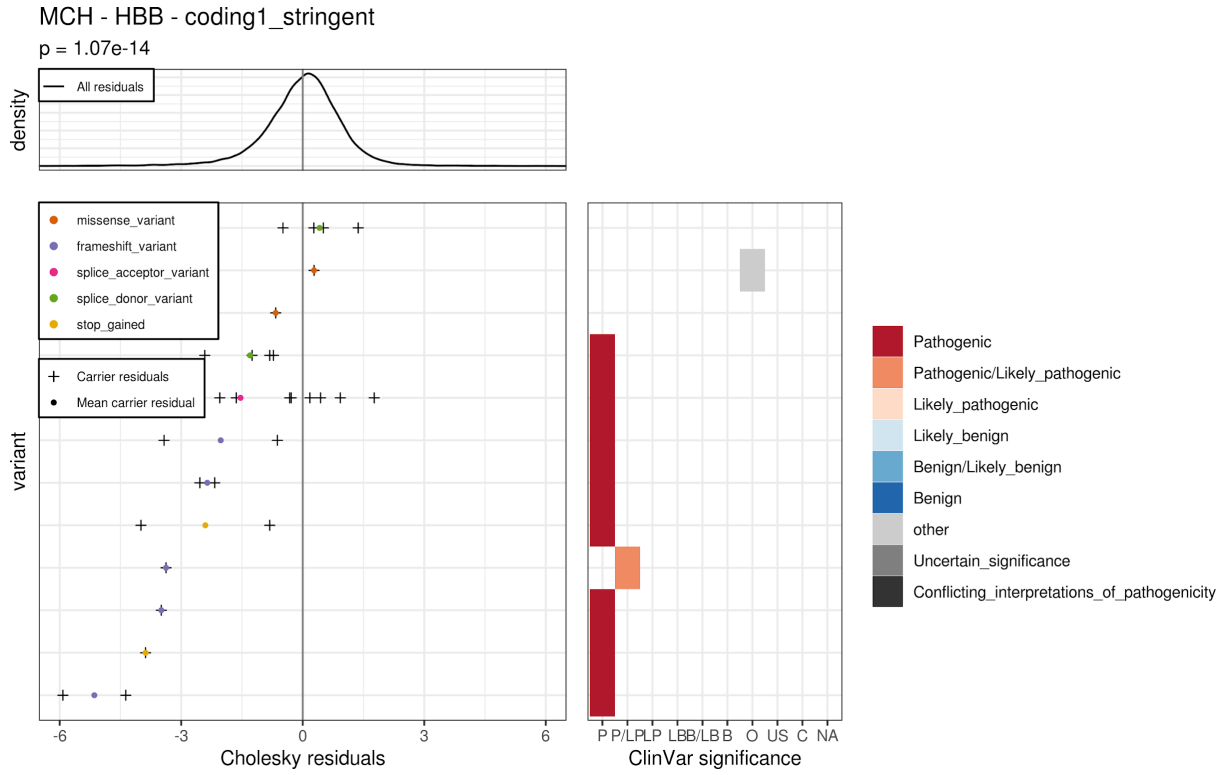
(C10)


MCH - HBA1 - coding2_relaxed + noncoding_stringent
p = 1.82e-09
density
All residuals
variant
missense_variant
frameshift_variant
splice_donor_variant
start_lost
Carrier residuals
Mean carrier residual
-2
0
2
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

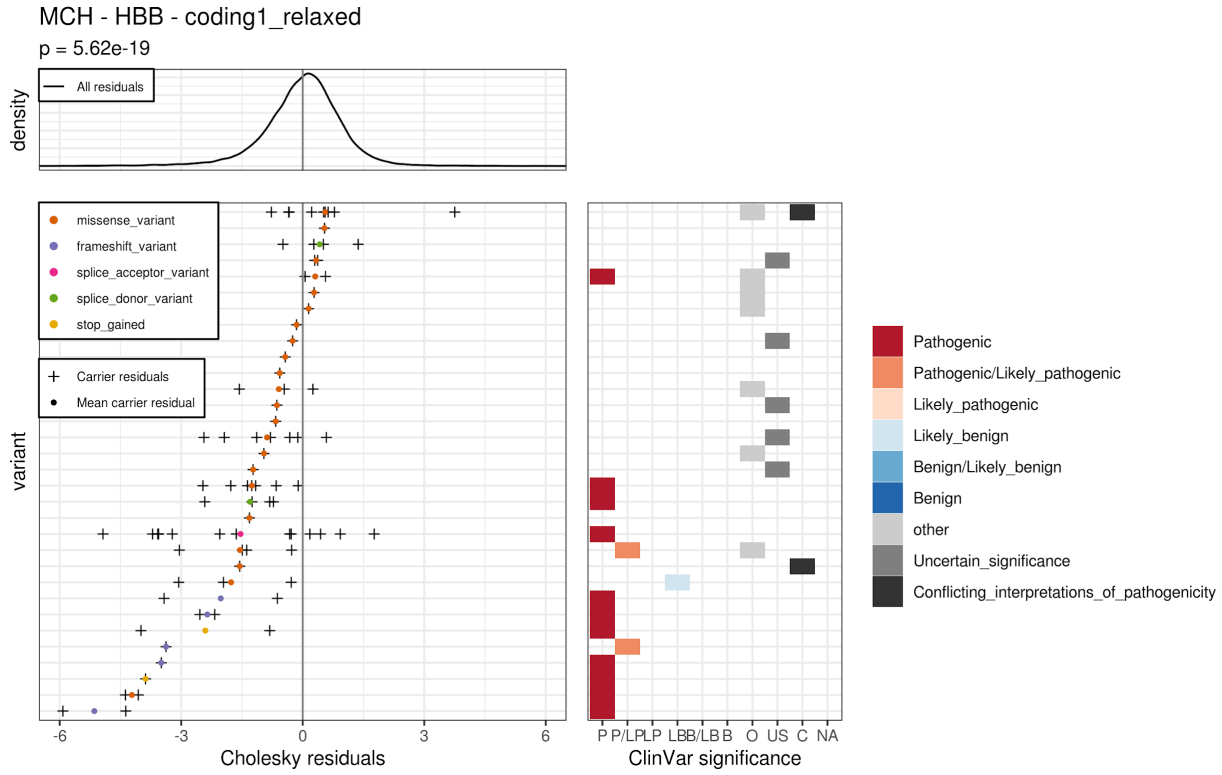
(C11)


MCH - HBB - coding2_relaxed
p = 1.04e-19
density
All residuals
variant
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-6
-3
0
3
6
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

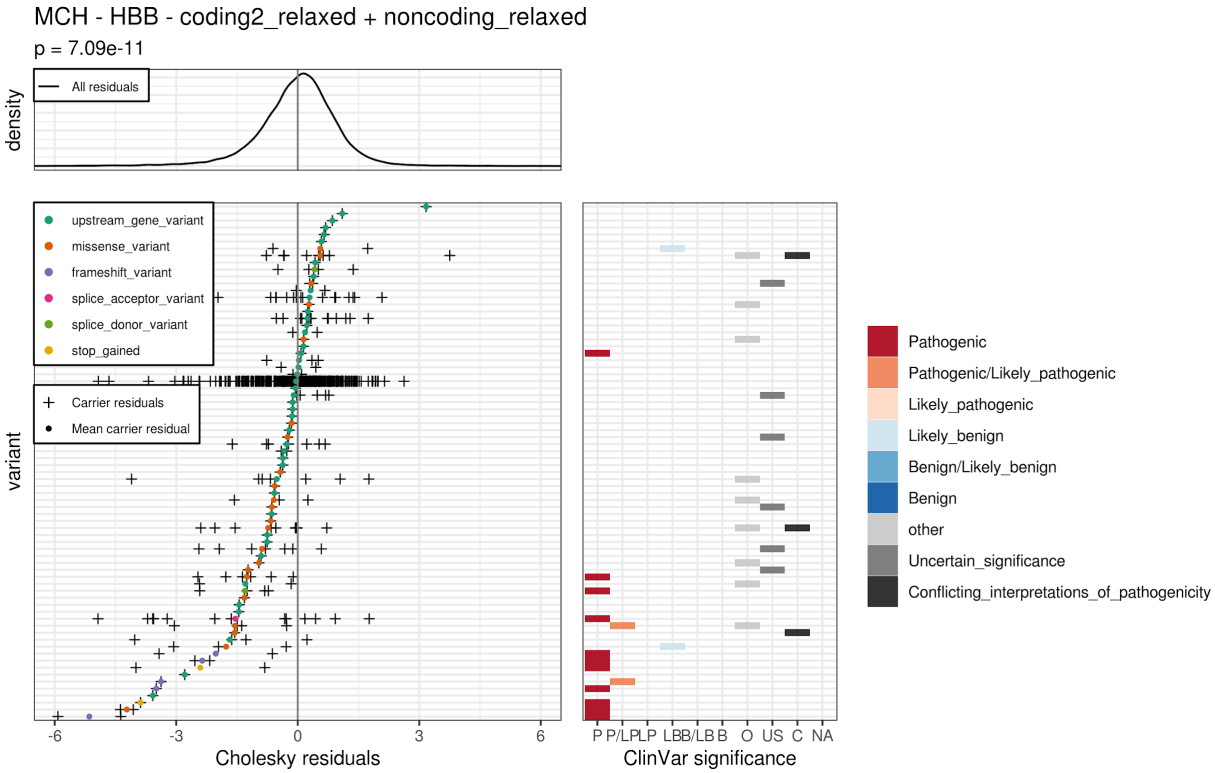
(C12)


MCH - HBB - coding1_stringent
p = 1.07e-14
density
All residuals
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-6
-3
0
3
6
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

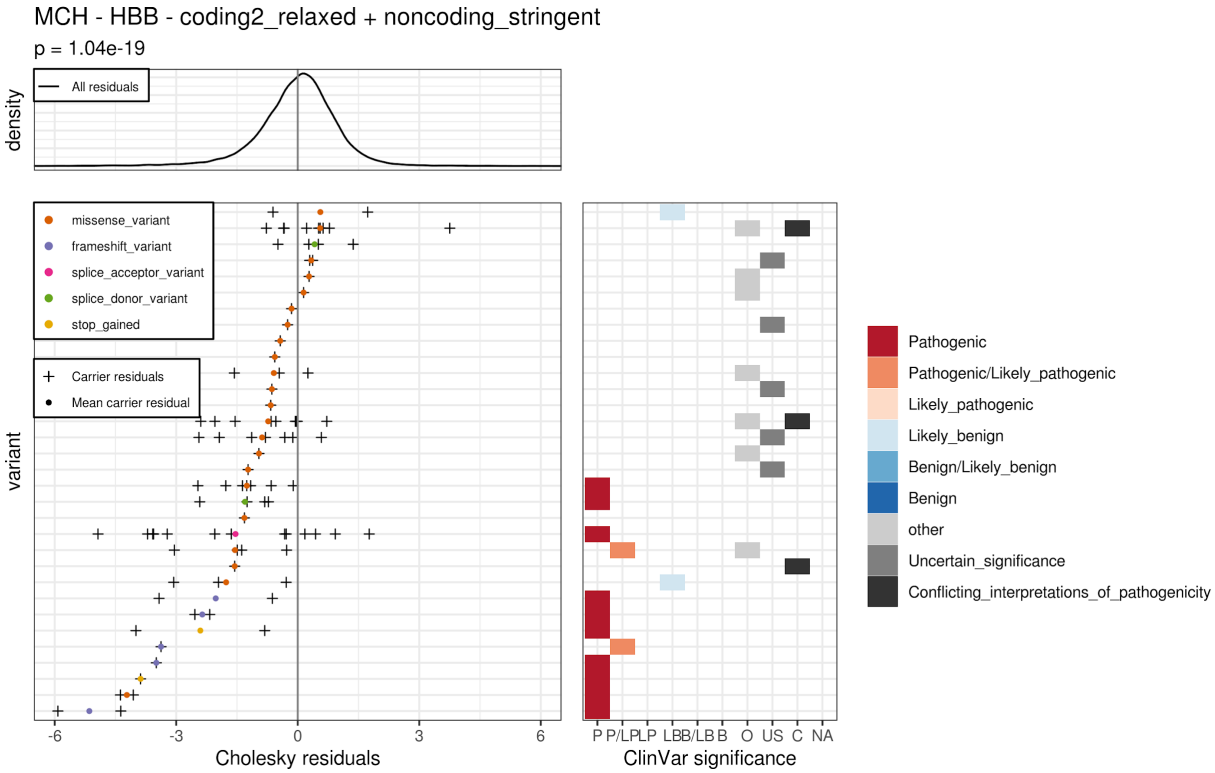
(C13)


MCH - HBB - coding1_relaxed
p = 5.62e-19
density
All residuals
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-6
-3
0
3
6
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

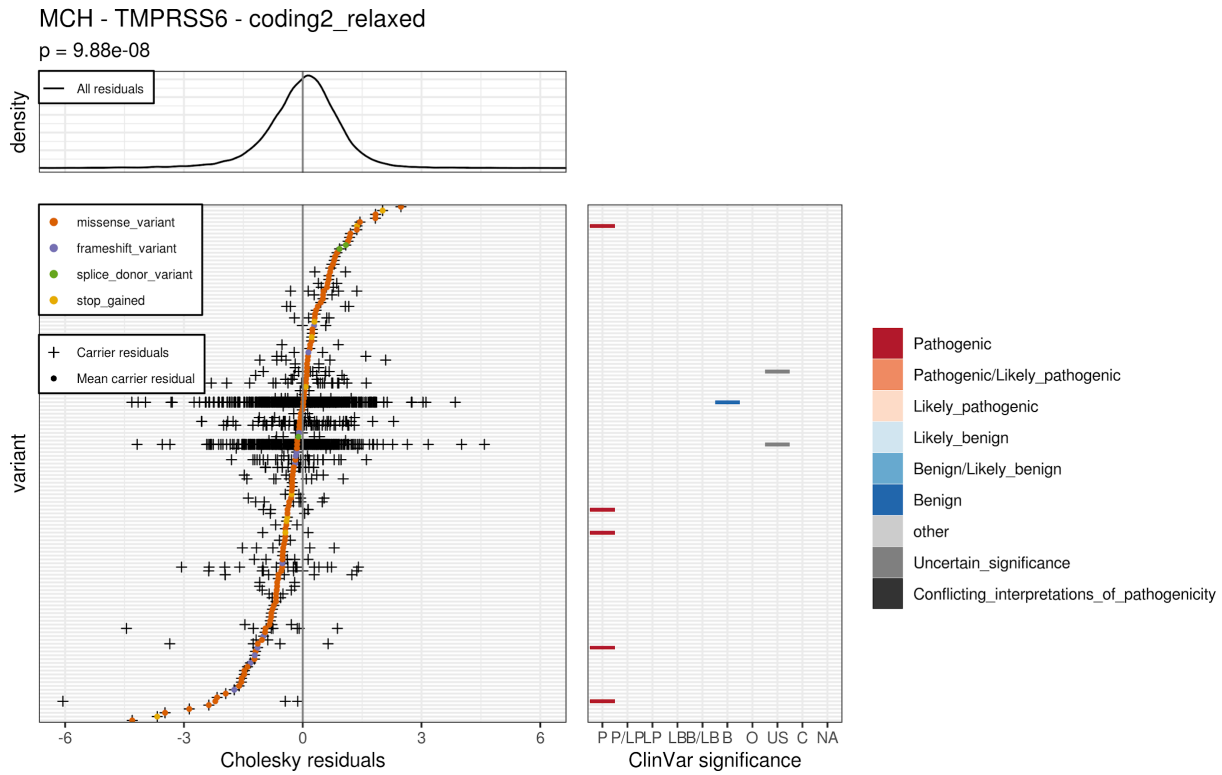
(C14)


MCH - HBB - coding2_relaxed + noncoding_relaxed
p = 7.09e-11
density
All residuals
variant
upstream_gene_variant
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-6
-3
0
3
6
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

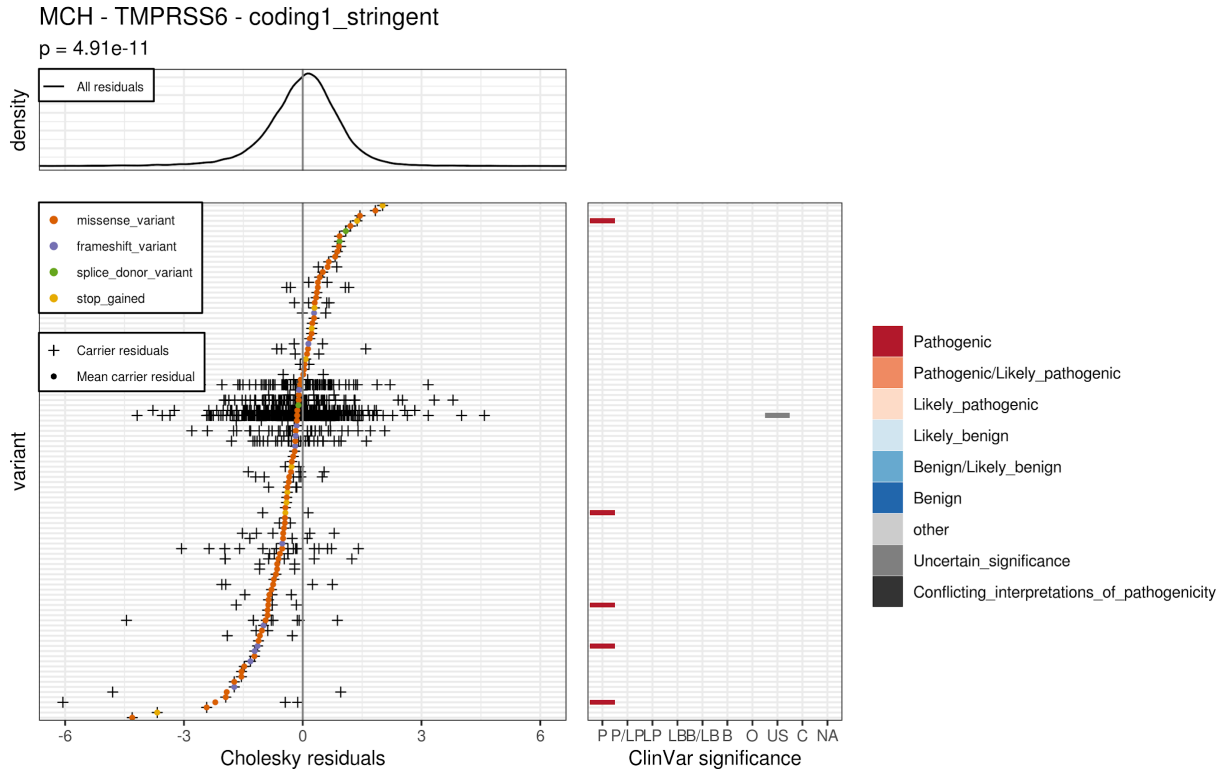
(C15)


MCH - HBB - coding2_relaxed + noncoding_stringent
p = 1.04e-19
density
All residuals
variant
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-6
-3
0
3
6
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

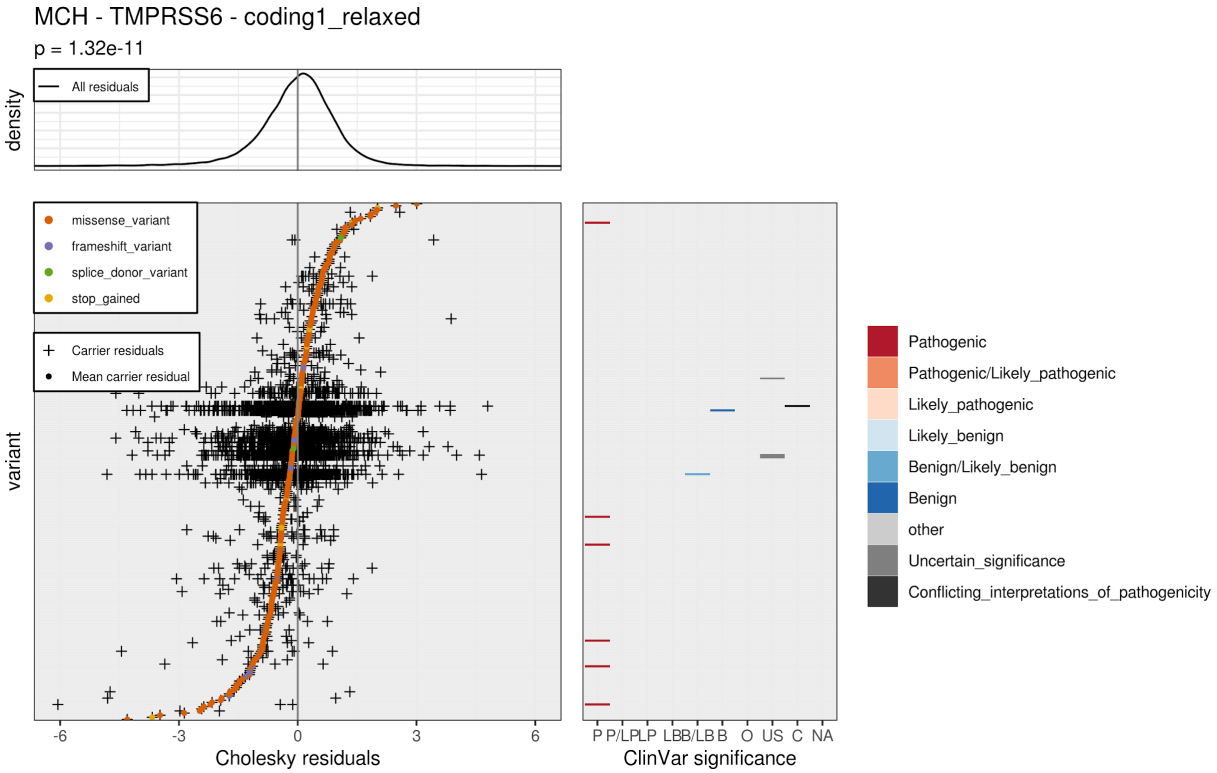
(C16)


MCH - TMPRSS6 - coding2_relaxed
p = 9.88e-08
density
All residuals
variant
missense_variant
frameshift_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-6
-3
0
3
6
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

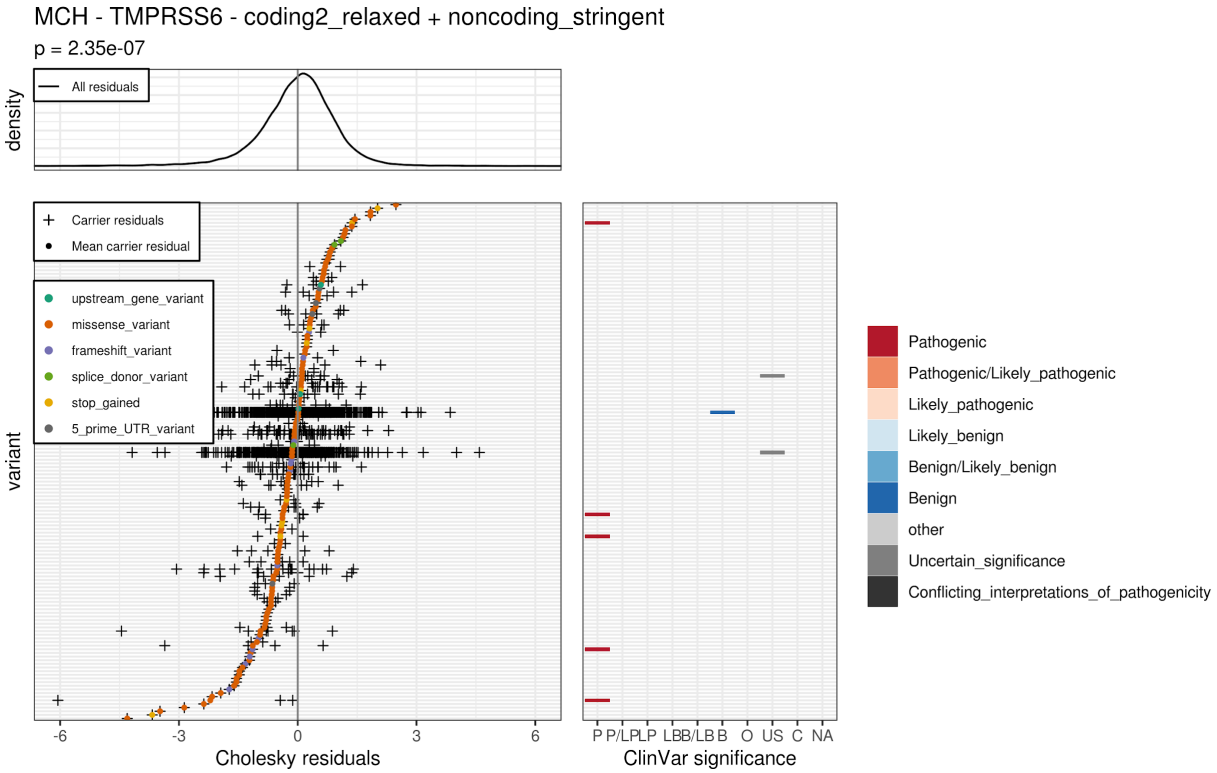
(C17)


MCH - TMPRSS6 - coding1_stringent
p = 4.91e-11
density
All residuals
variant
missense_variant
frameshift_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-6
-3
0
3
6
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(C18)

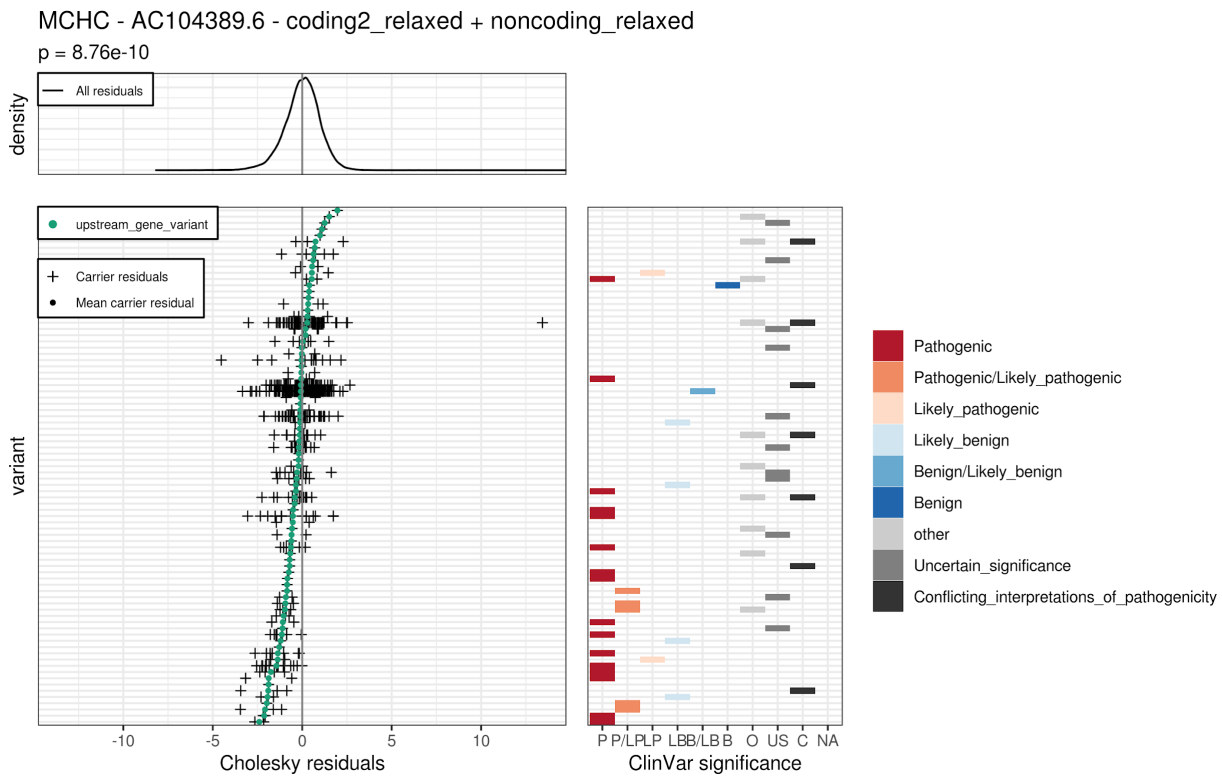

MCH - TMPRSS6 - coding1_relaxed
p = 1.32e-11
density
All residuals
variant
missense_variant
frameshift_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-6
-3
0
3
6
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(C19)

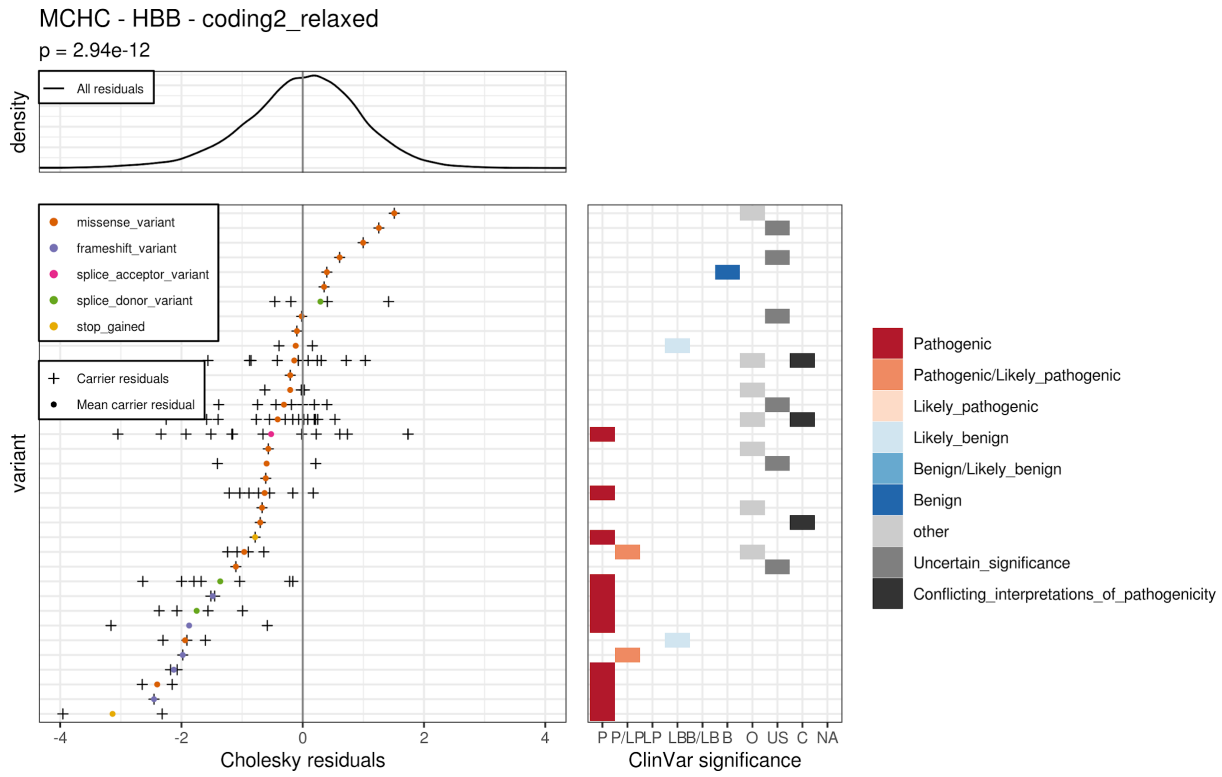

MCH - TMPRSS6 - coding2_relaxed + noncoding_stringent
p = 2.35e-07
density
All residuals
variant
Carrier residuals
Mean carrier residual
upstream_gene_variant
missense_variant
frameshift_variant
splice_donor_variant
stop_gained
5_prime_UTR_variant
-6
-3
0
3
6
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(D)

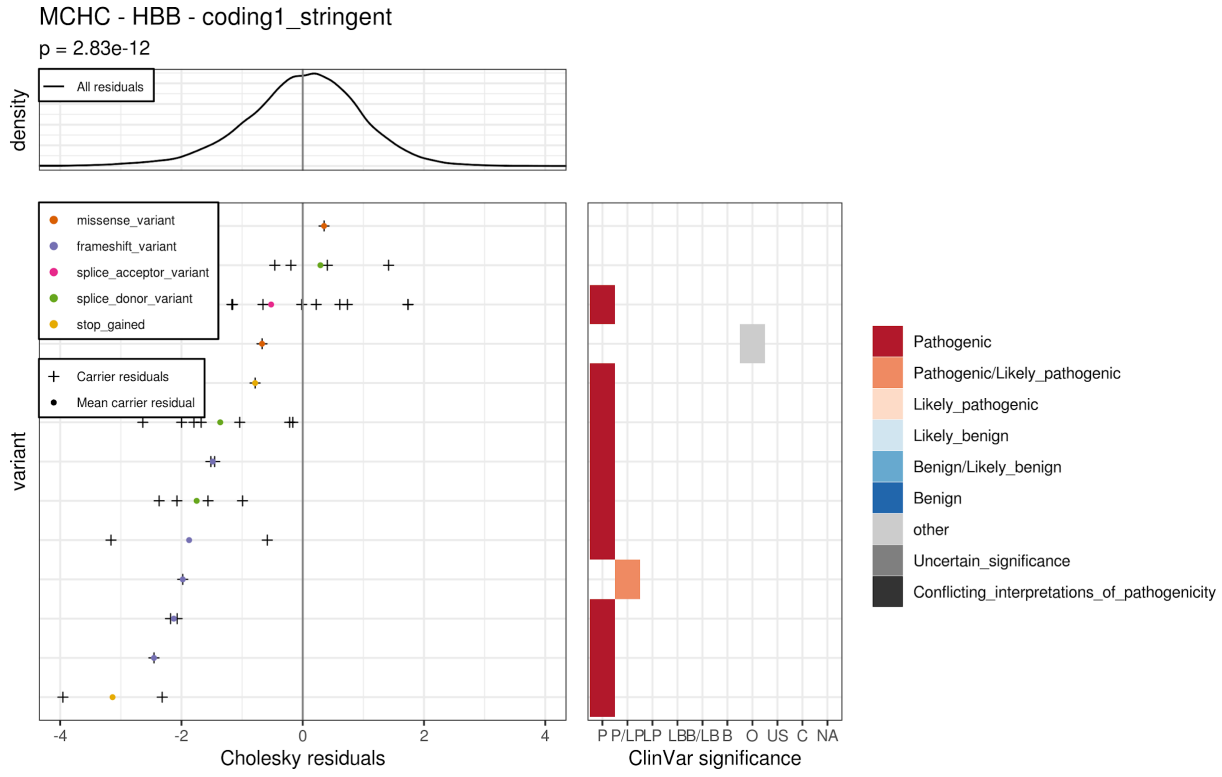
(D1)


MCHC - AC104389.6 - coding2_relaxed + noncoding_relaxed
p = 8.76e-10
density
All residuals
variant
upstream_gene_variant
Carrier residuals
Mean carrier residual
-10
-5
0
5
10
Cholesky residuals
P
P/LP
LP
LB
B/LB
B
O
US
C
NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

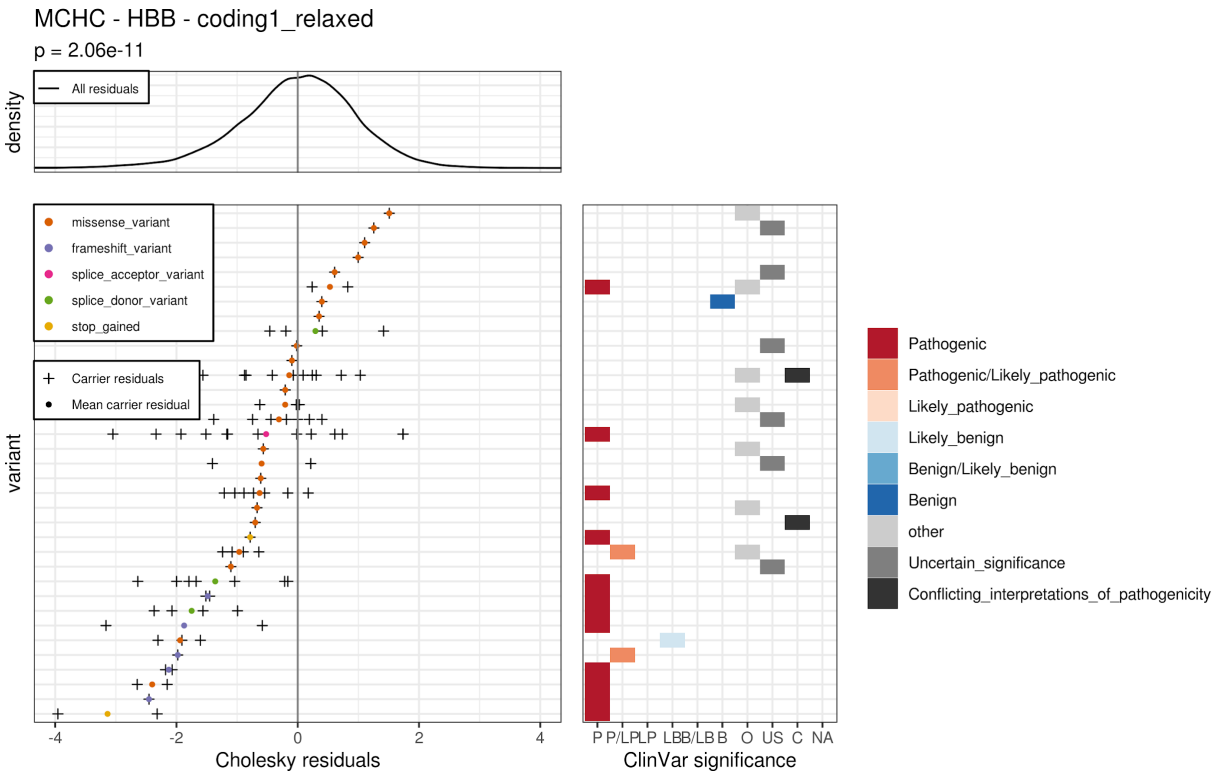
(D2)


MCHC - HBB - coding2_relaxed
p = 2.94e-12
density
All residuals
variant
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-4
-2
0
2
4
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

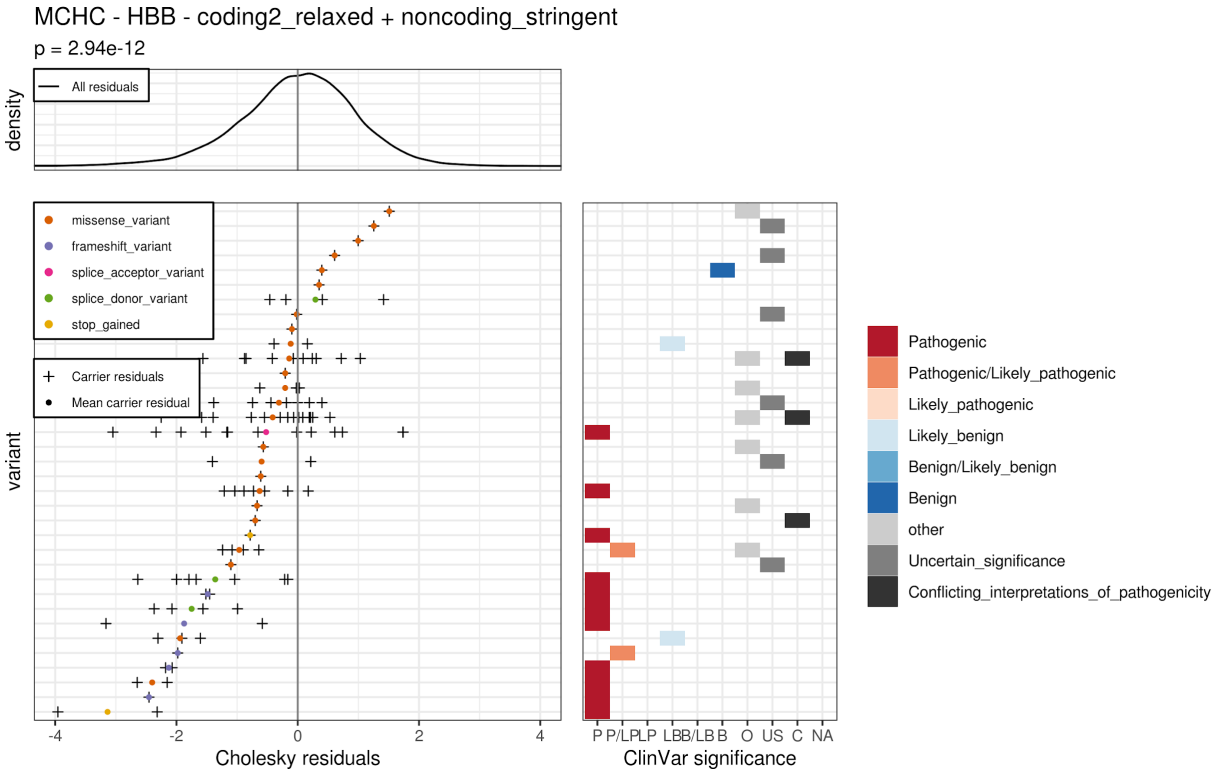
(D3)


MCHC - HBB - coding1_stringent
p = 2.83e-12
density
All residuals
variant
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-4
-2
0
2
4
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(D4)

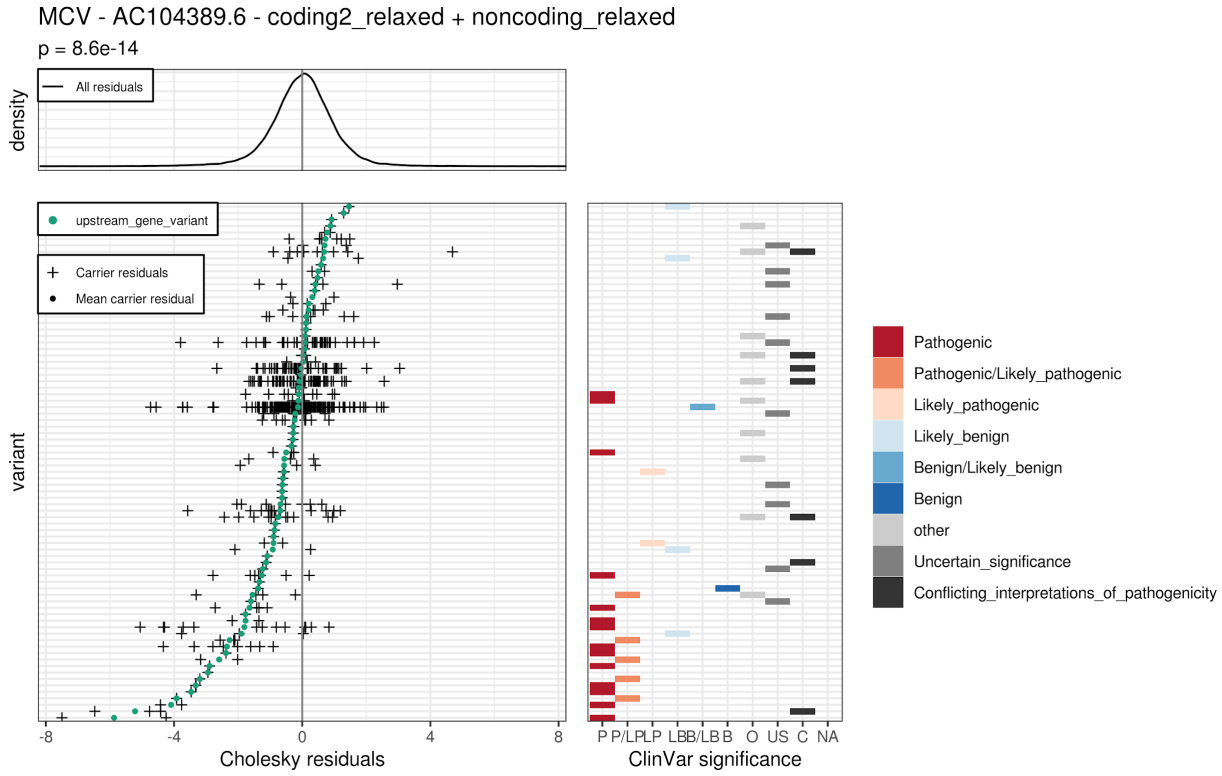

MCHC - HBB - coding1_relaxed
p = 2.06e-11
density
All residuals
variant
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-4
-2
0
2
4
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(D5)

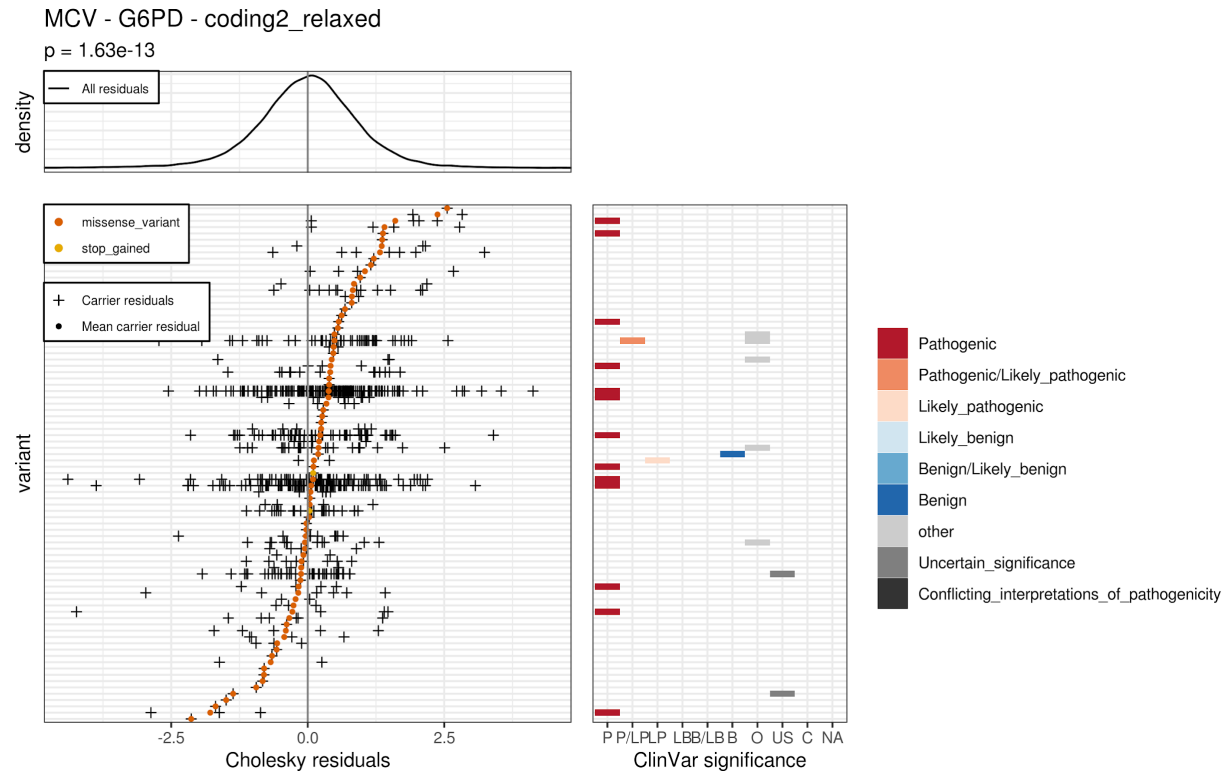

MCHC - HBB - coding2_relaxed + noncoding_stringent
p = 2.94e-12
density
All residuals
variant
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-4
-2
0
2
4
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(E)

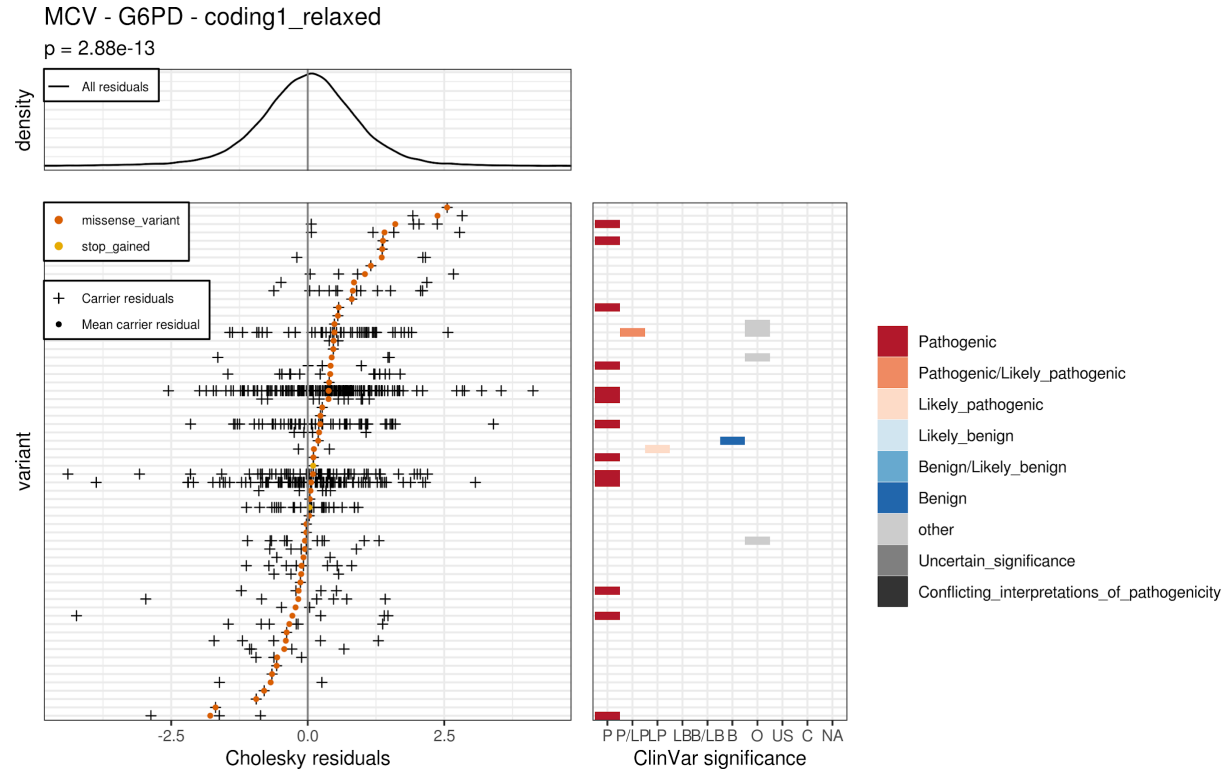
(E1)


MCV - AC104389.6 - coding2_relaxed + noncoding_relaxed
p = 8.6e-14
All residuals
density
upstream_gene_variant
Carrier residuals
Mean carrier residual
variant
-8
-4
0
4
8
Cholesky residuals
P
P/LP
LP
LB
B/LB
B
O
US
C
NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

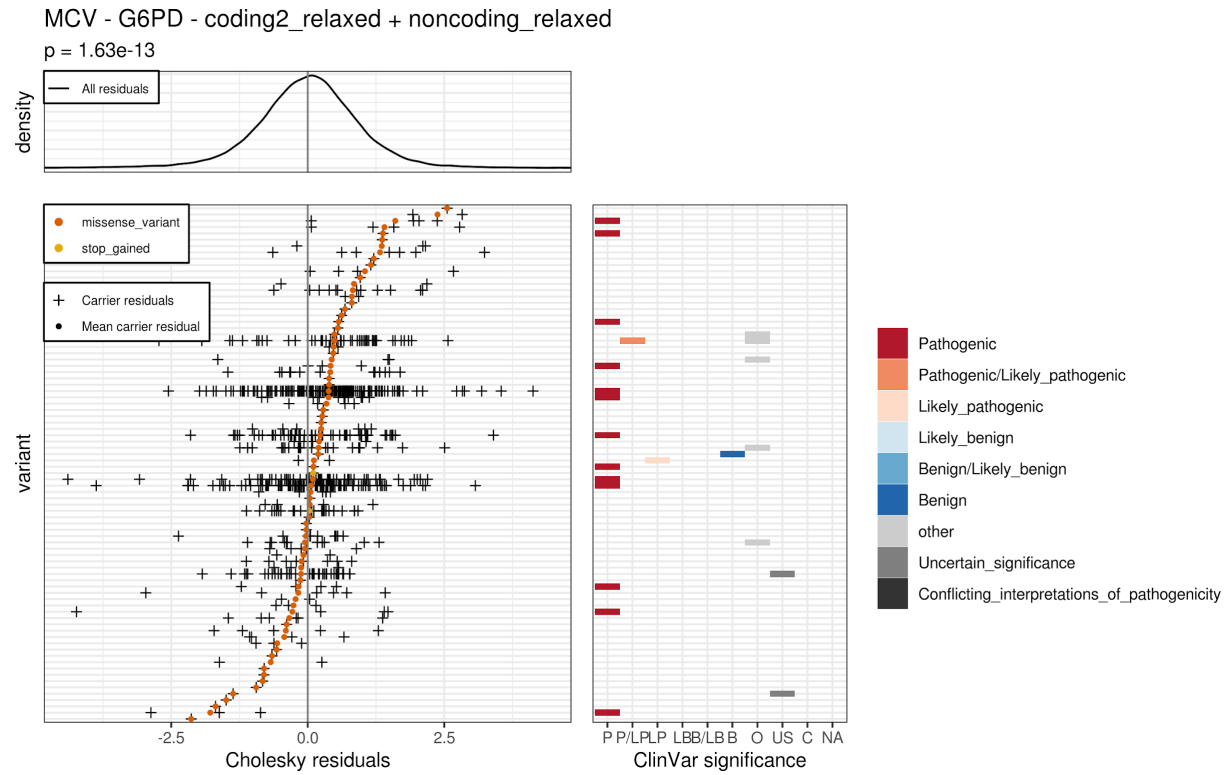
(E2)


MCV - G6PD - coding2_relaxed
p = 1.63e-13
density
All residuals
missense_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-2.5
0.0
2.5
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

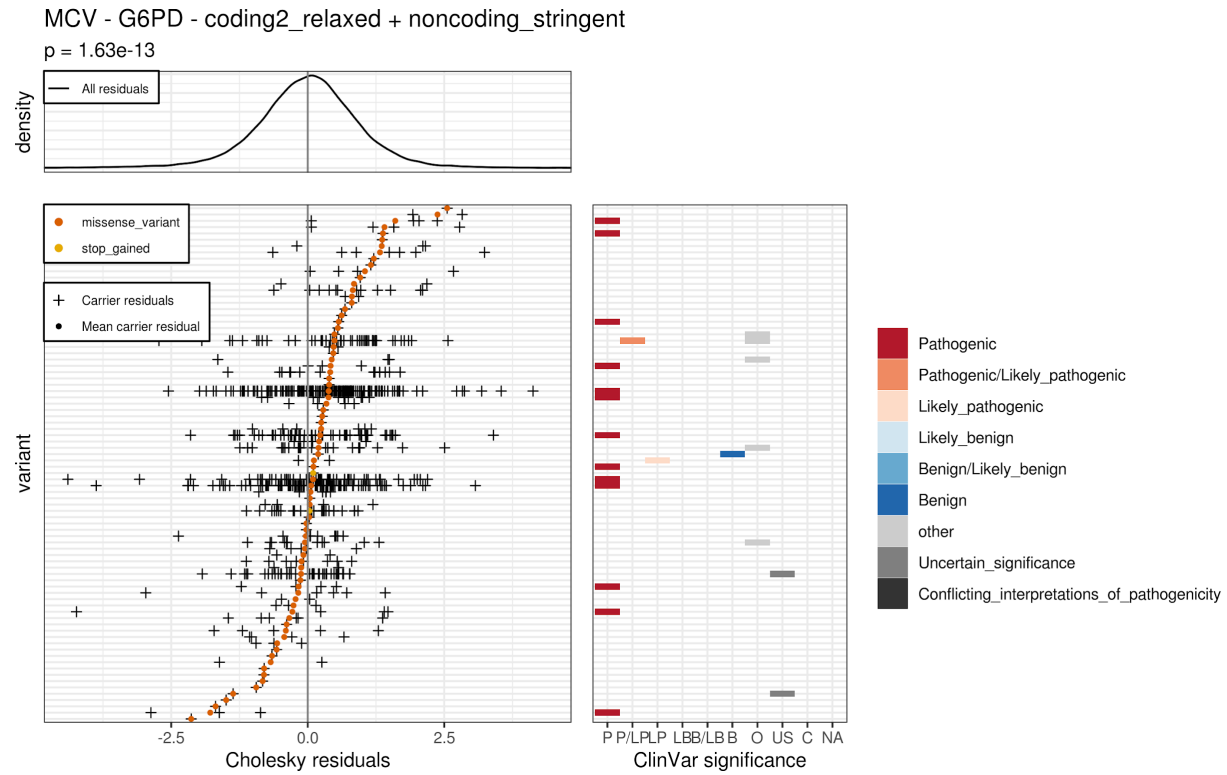
(E3)


MCV - G6PD - coding1_relaxed
p = 2.88e-13
density
All residuals
missense_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-2.5
0.0
2.5
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

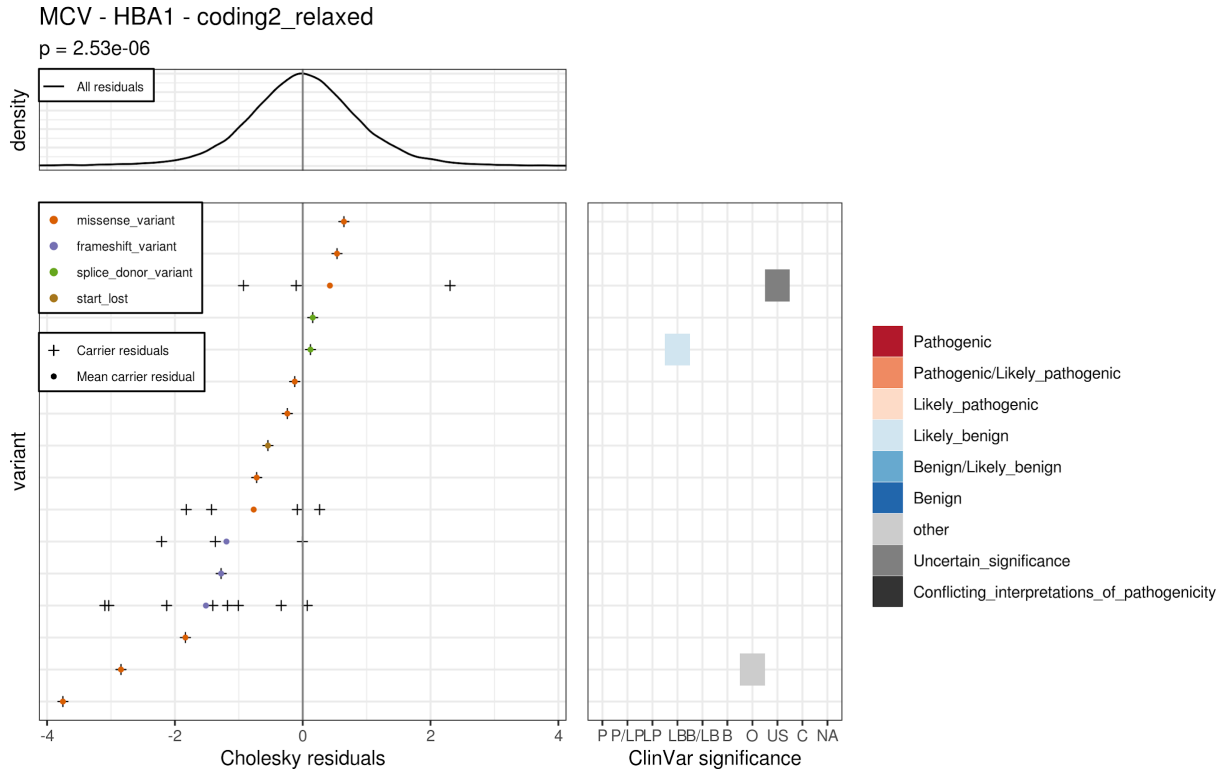
(E4)


MCV - G6PD - coding2_relaxed + noncoding_relaxed
p = 1.63e-13
density
All residuals
missense_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-2.5
0.0
2.5
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

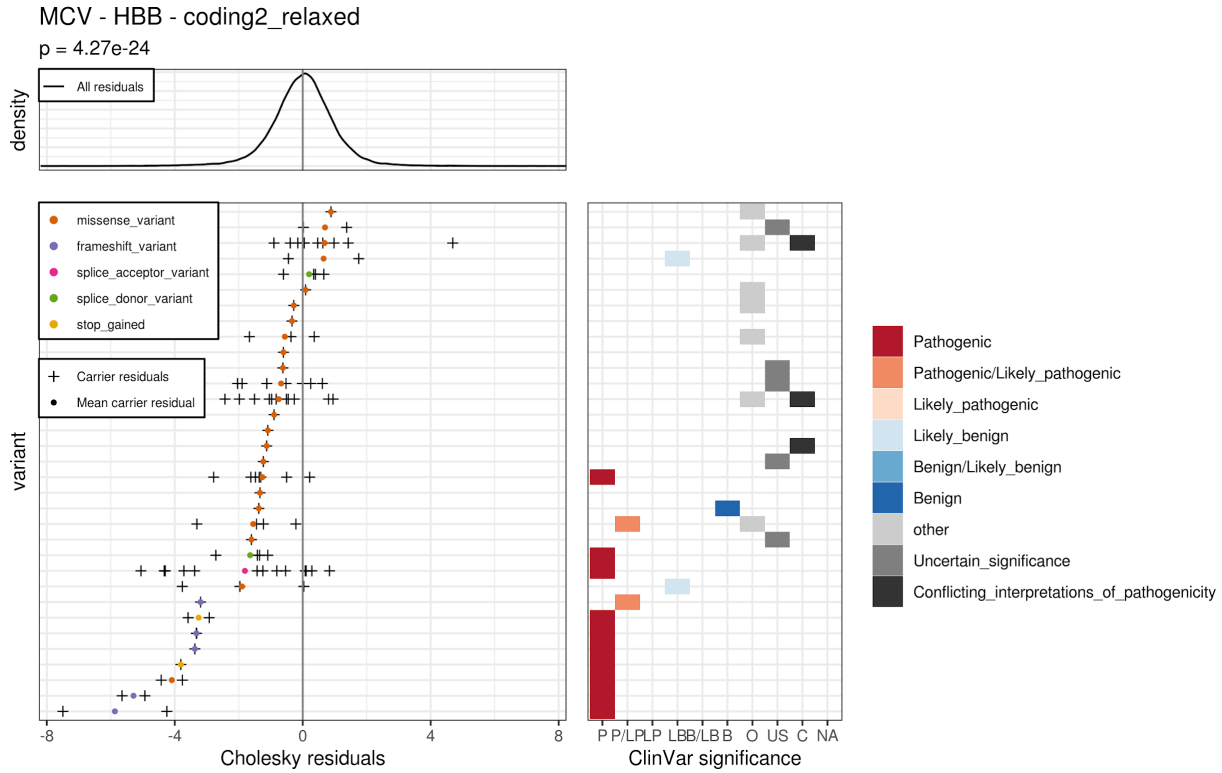
(E5)


MCV - G6PD - coding2_relaxed + noncoding_stringent
p = 1.63e-13
density
All residuals
missense_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-2.5
0.0
2.5
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

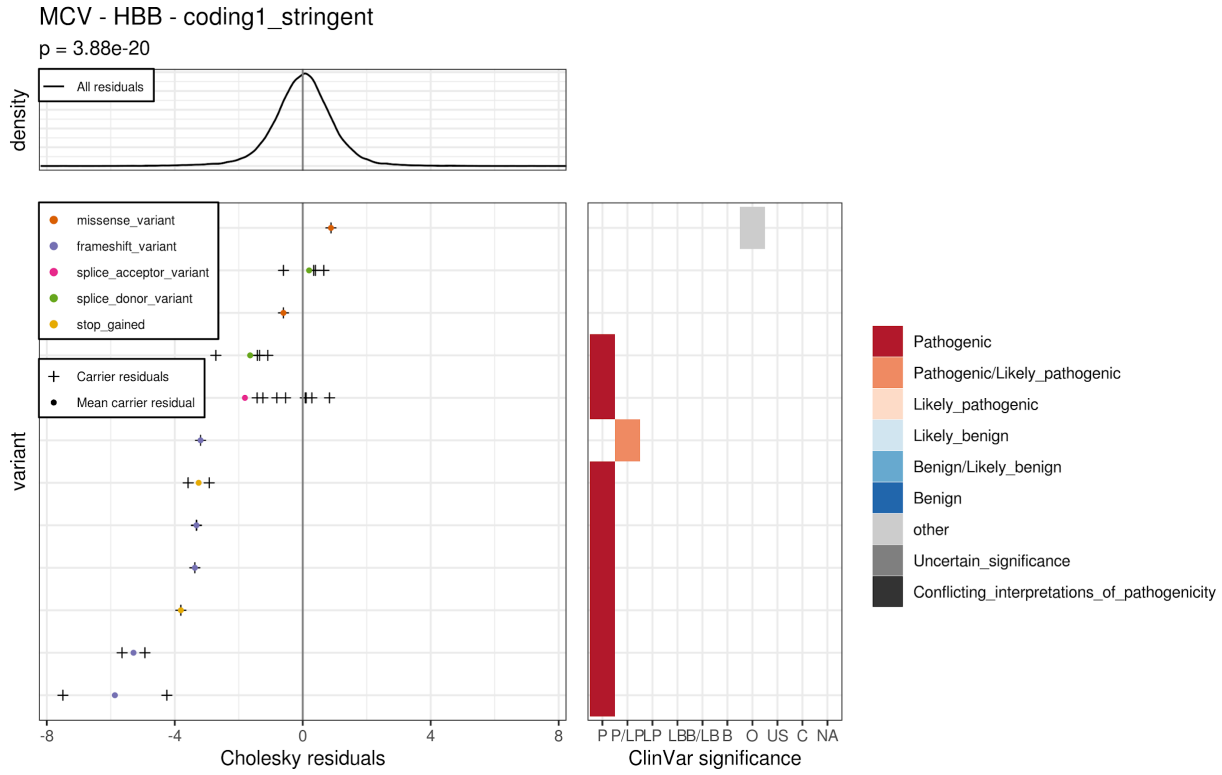
(E6)


MCV - HBA1 - coding2_relaxed
p = 2.53e-06
density
All residuals
missense_variant
frameshift_variant
splice_donor_variant
start_lost
Carrier residuals
Mean carrier residual
variant
-4
-2
0
2
4
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

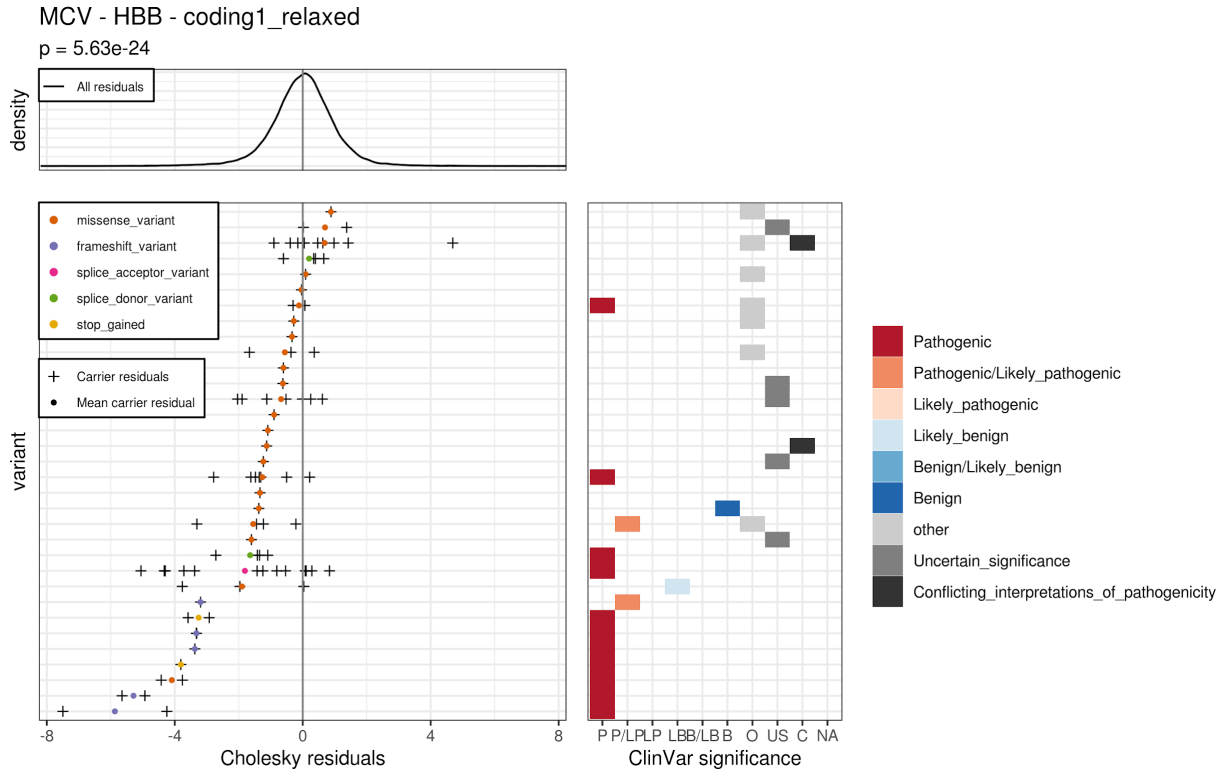
(E7)


MCV - HBB - coding2_relaxed
p = 4.27e-24
density
All residuals
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-8
-4
0
4
8
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

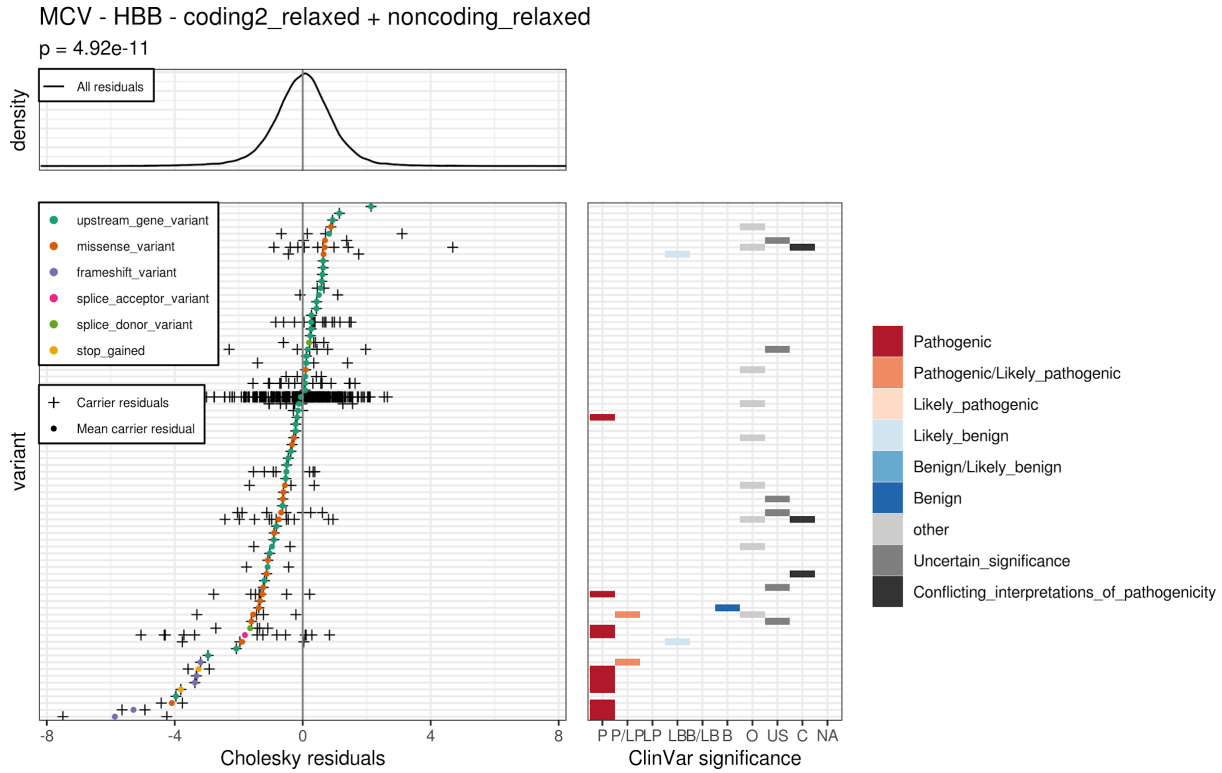
(E8)


MCV - HBB - coding1_stringent
p = 3.88e-20
density
All residuals
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-8
-4
0
4
8
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

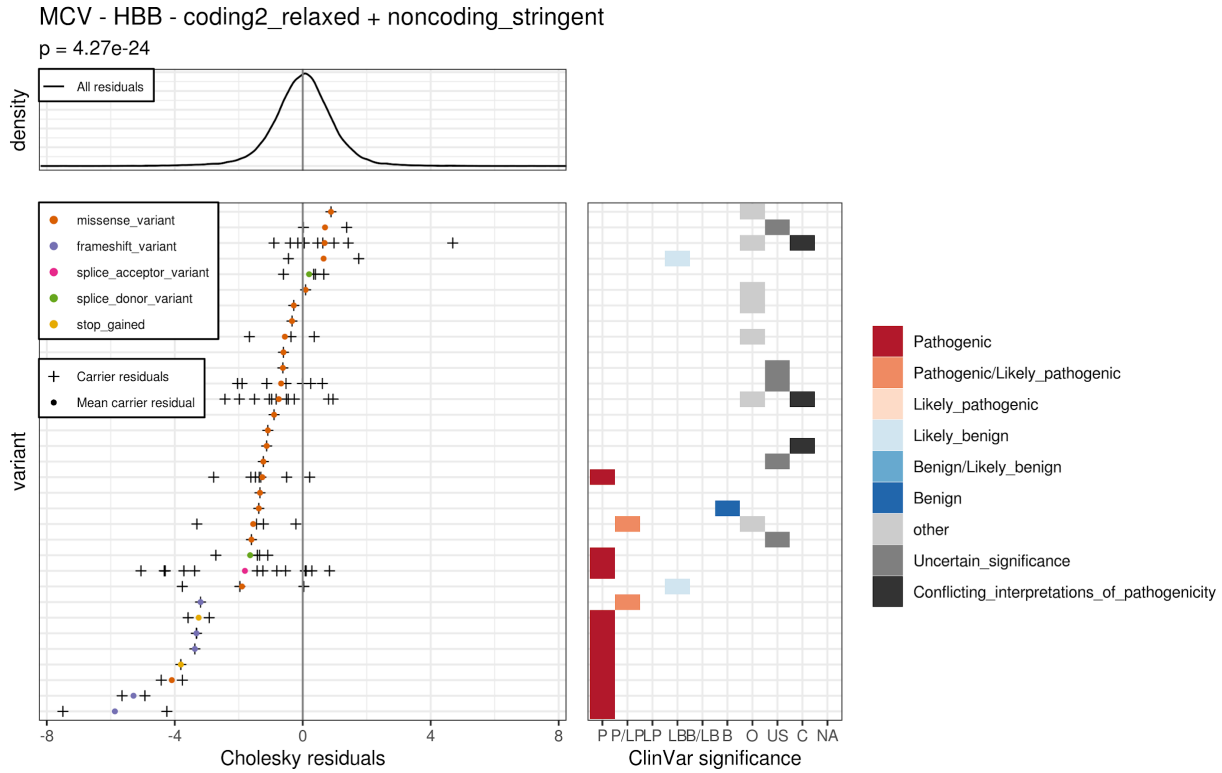
(E9)


MCV - HBB - coding1_relaxed
p = 5.63e-24
density
All residuals
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-8
-4
0
4
8
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

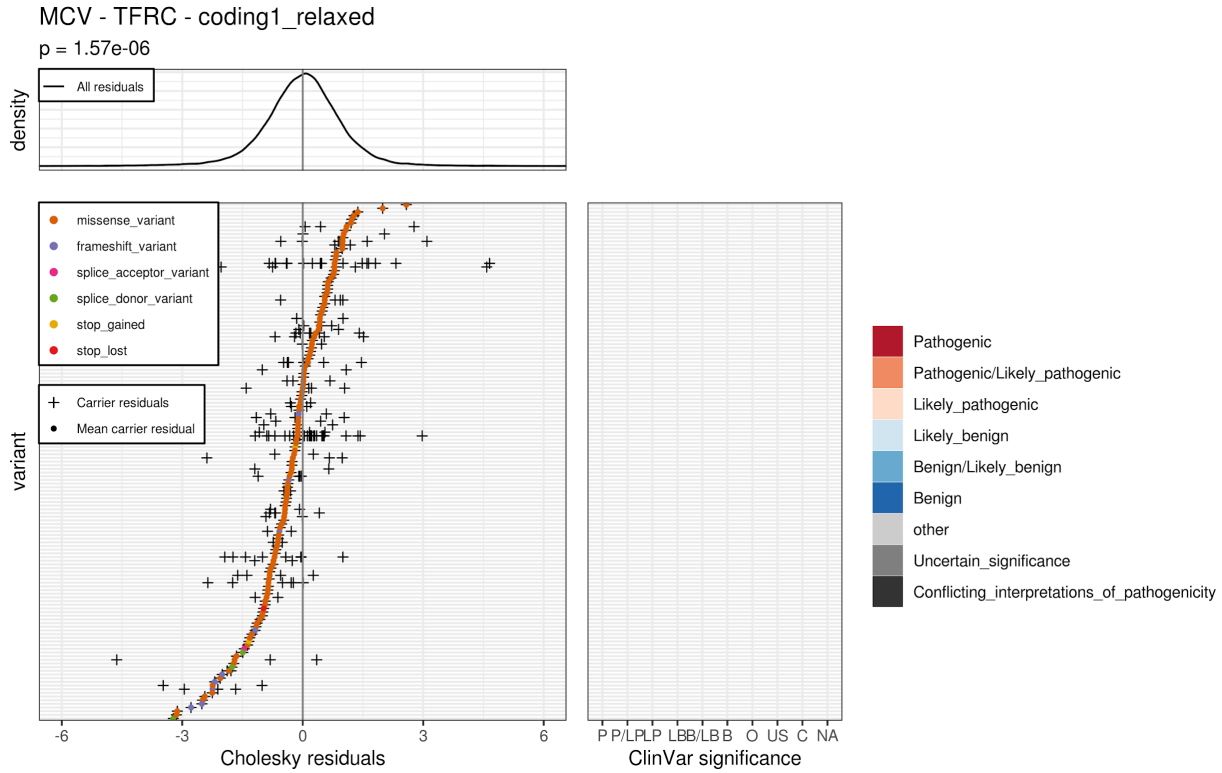
(E10)


MCV - HBB - coding2_relaxed + noncoding_relaxed
p = 4.92e-11
density
All residuals
variant
upstream_gene_variant
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-8
-4
0
4
8
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

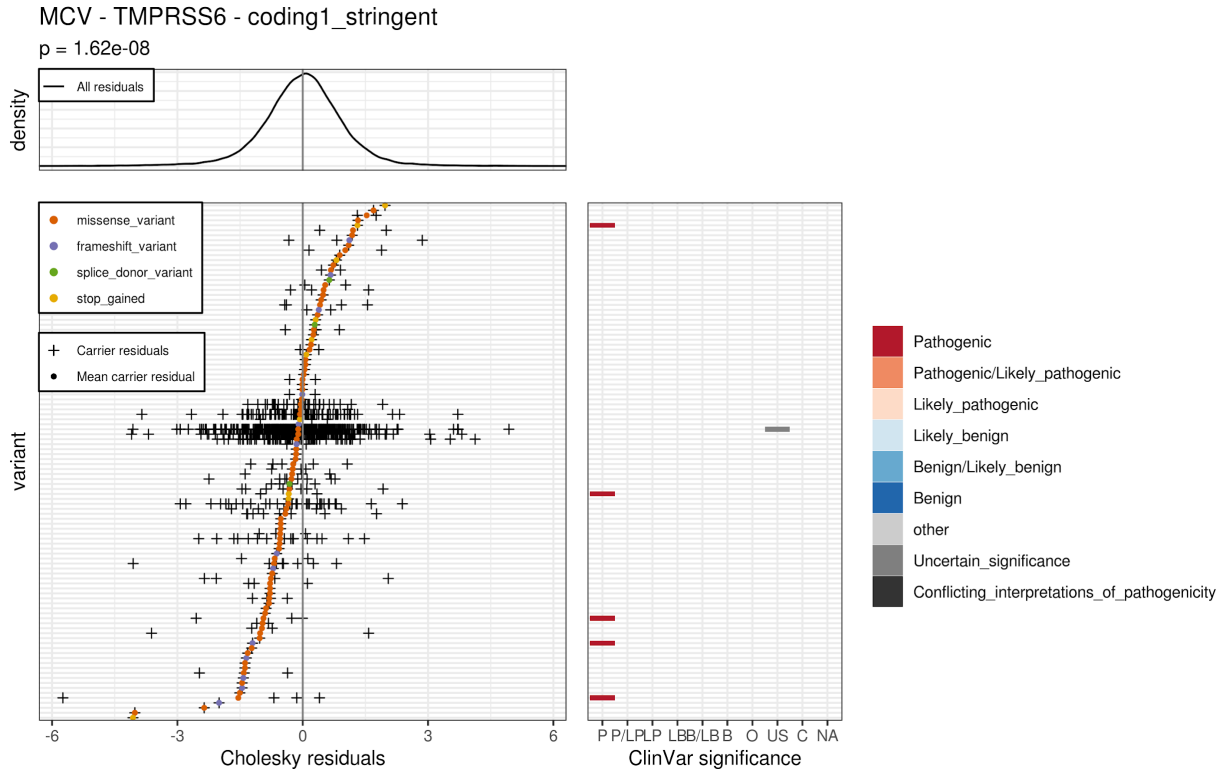
(E11)


MCV - HBB - coding2_relaxed + noncoding_stringent
p = 4.27e-24
density
All residuals
variant
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-8
-4
0
4
8
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

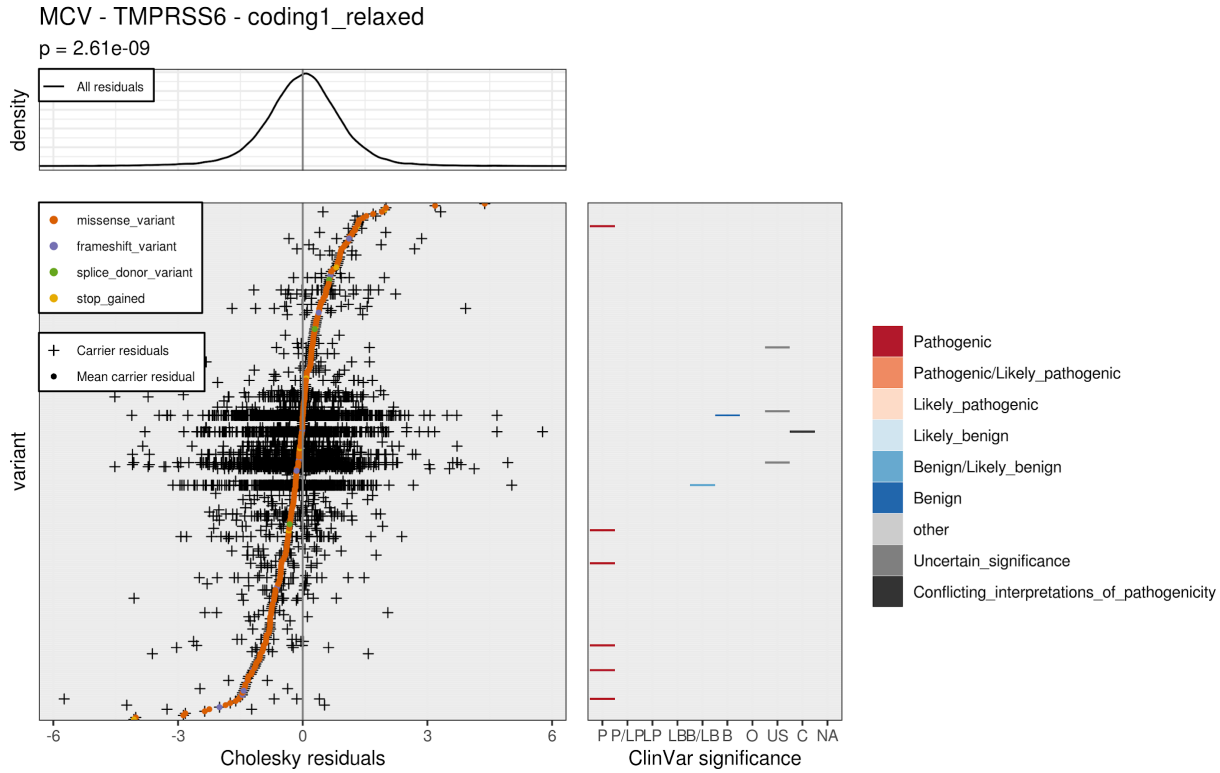
(E12)


MCV - TFRC - coding1_relaxed
p = 1.57e-06
density
All residuals
variant
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
stop_lost
Carrier residuals
Mean carrier residual
-6
-3
0
3
6
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(E13)

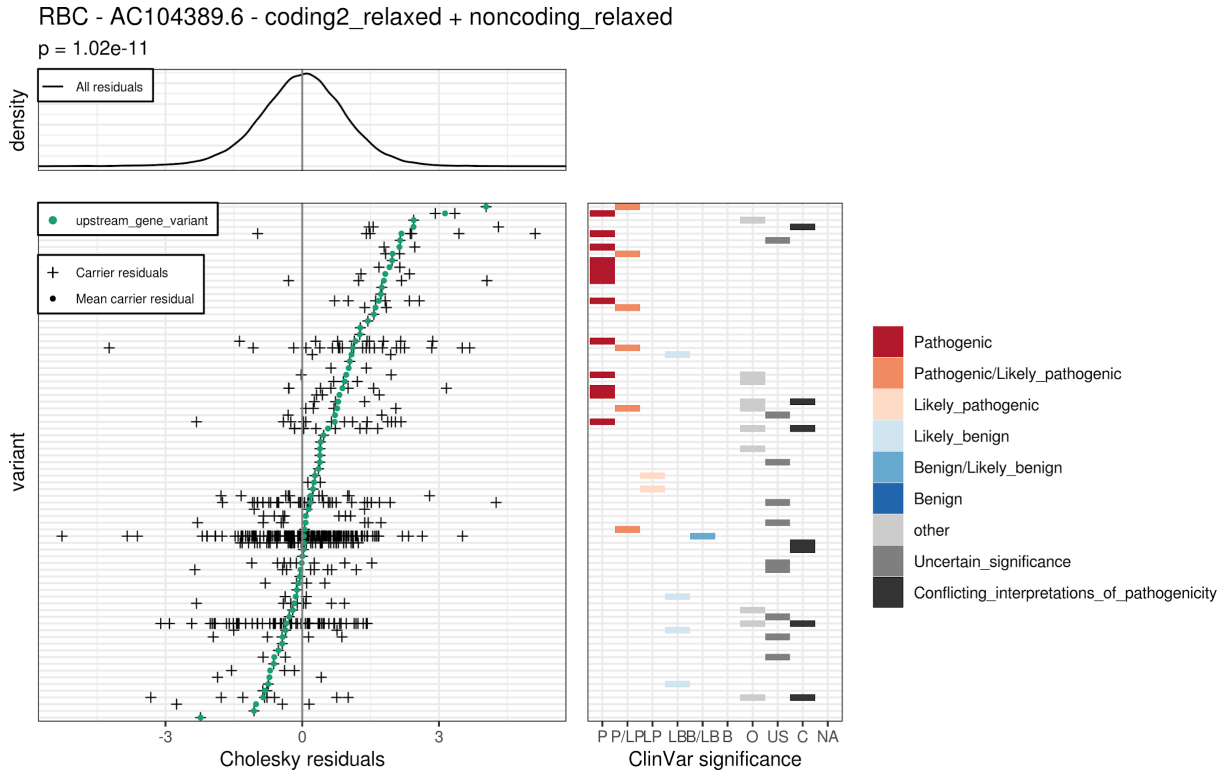

MCV - TMPRSS6 - coding1_stringent
p = 1.62e-08
density
All residuals
variant
missense_variant
frameshift_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-6
-3
0
3
6
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(E14)

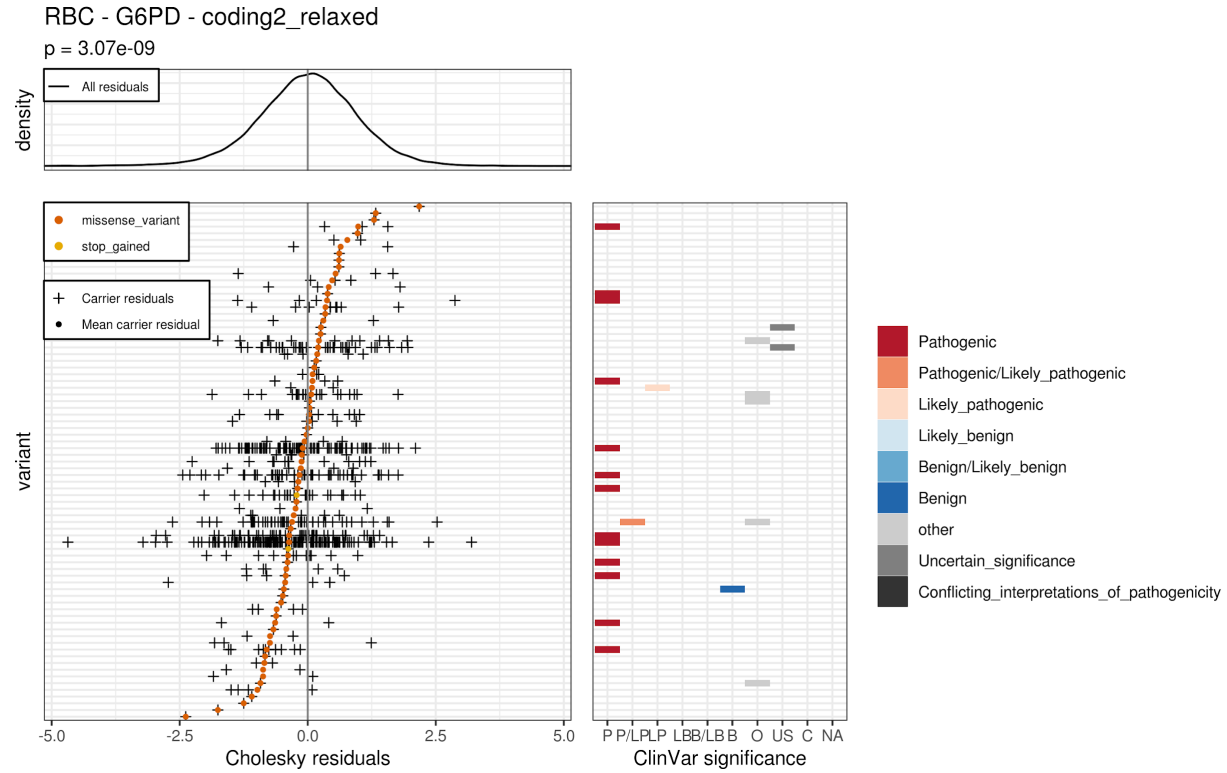

MCV - TMPRSS6 - coding1_relaxed
p = 2.61e-09
density
All residuals
variant
missense_variant
frameshift_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-6
-3
0
3
6
Cholesky residuals
P
P/LP
LP
LB
B/LB
B
O
US
C
NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(F)

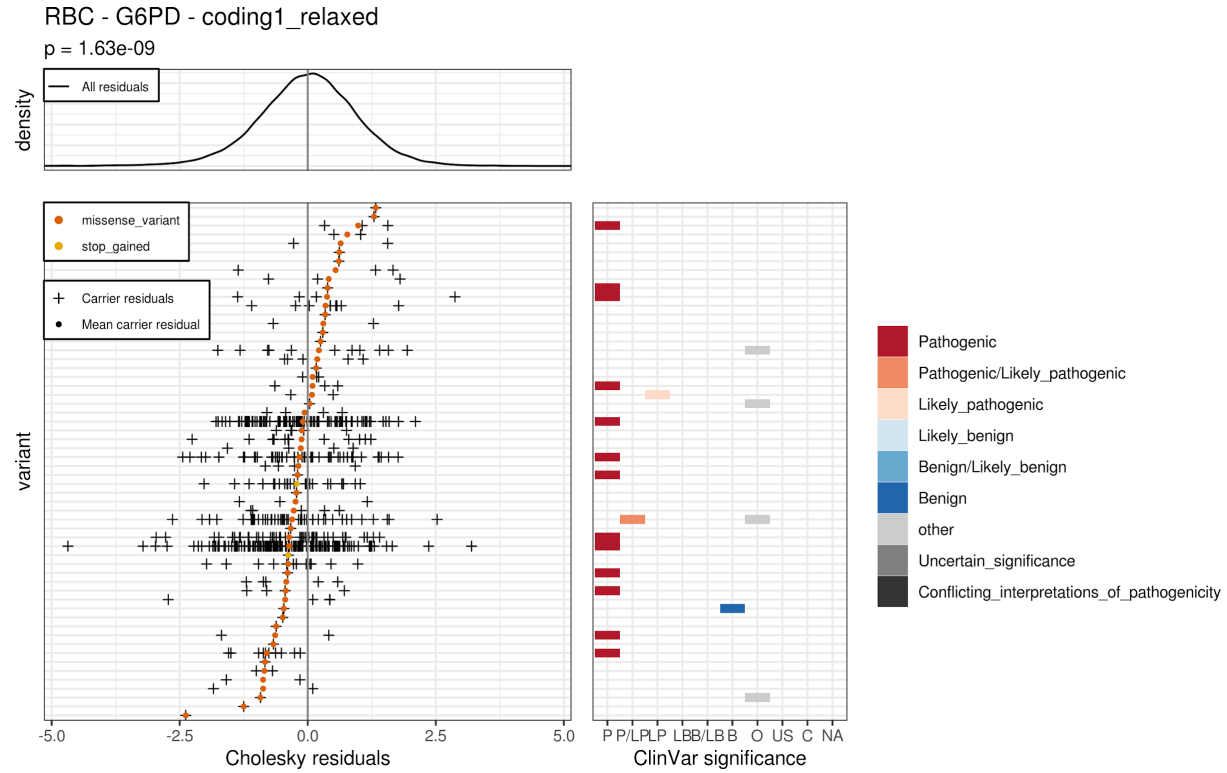
(F1)


RBC - AC104389.6 - coding2_relaxed + noncoding_relaxed
p = 1.02e-11
All residuals
density
upstream_gene_variant
Carrier residuals
Mean carrier residual
variant
-3
0
3
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

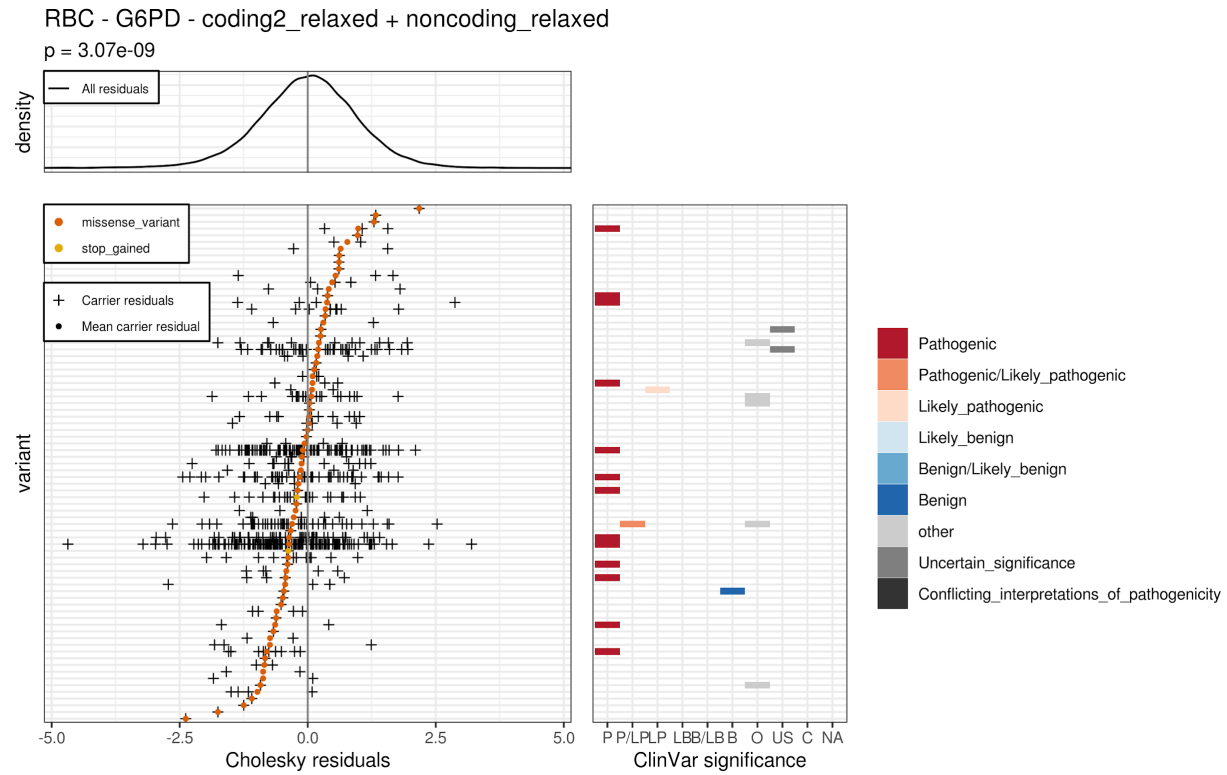
(F2)


RBC - G6PD - coding2_relaxed
p = 3.07e-09
density
All residuals
variant
missense_variant
stop_gained
Carrier residuals
Mean carrier residual
-5.0
-2.5
0.0
2.5
5.0
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

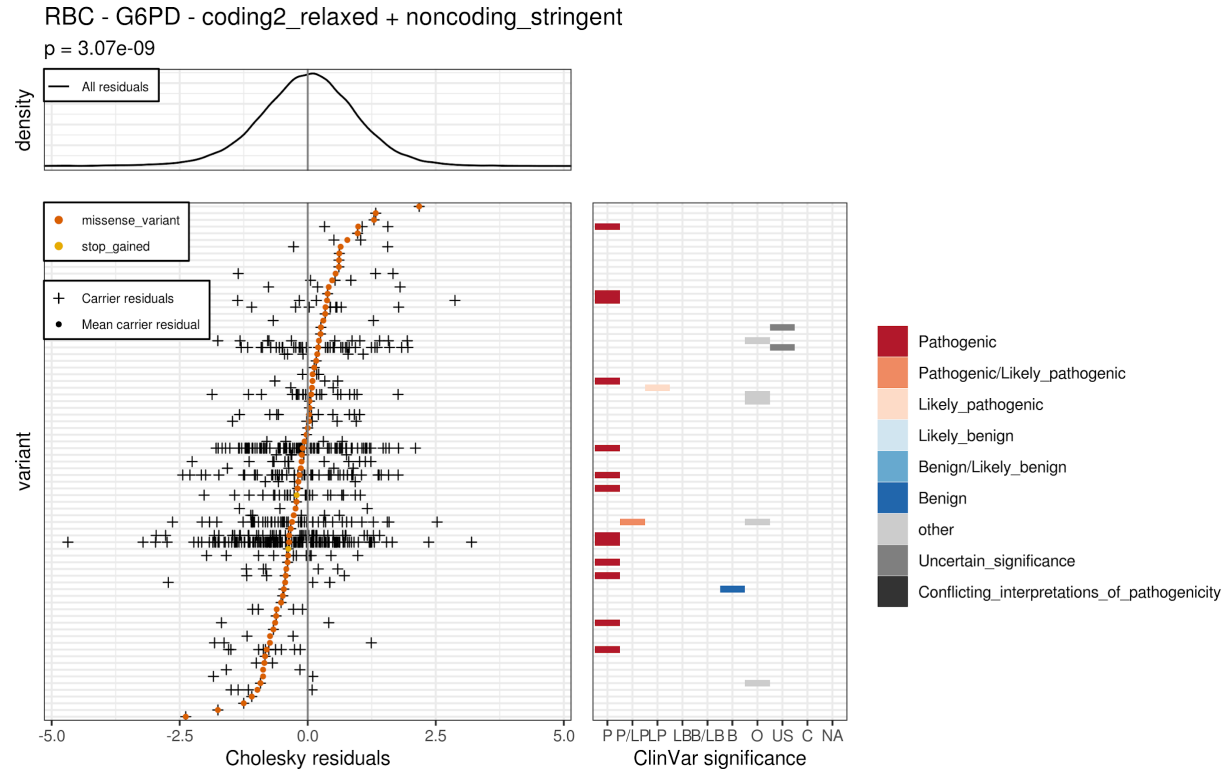
(F3)


RBC - G6PD - coding1_relaxed
p = 1.63e-09
density
All residuals
variant
missense_variant
stop_gained
Carrier residuals
Mean carrier residual
-5.0
-2.5
0.0
2.5
5.0
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

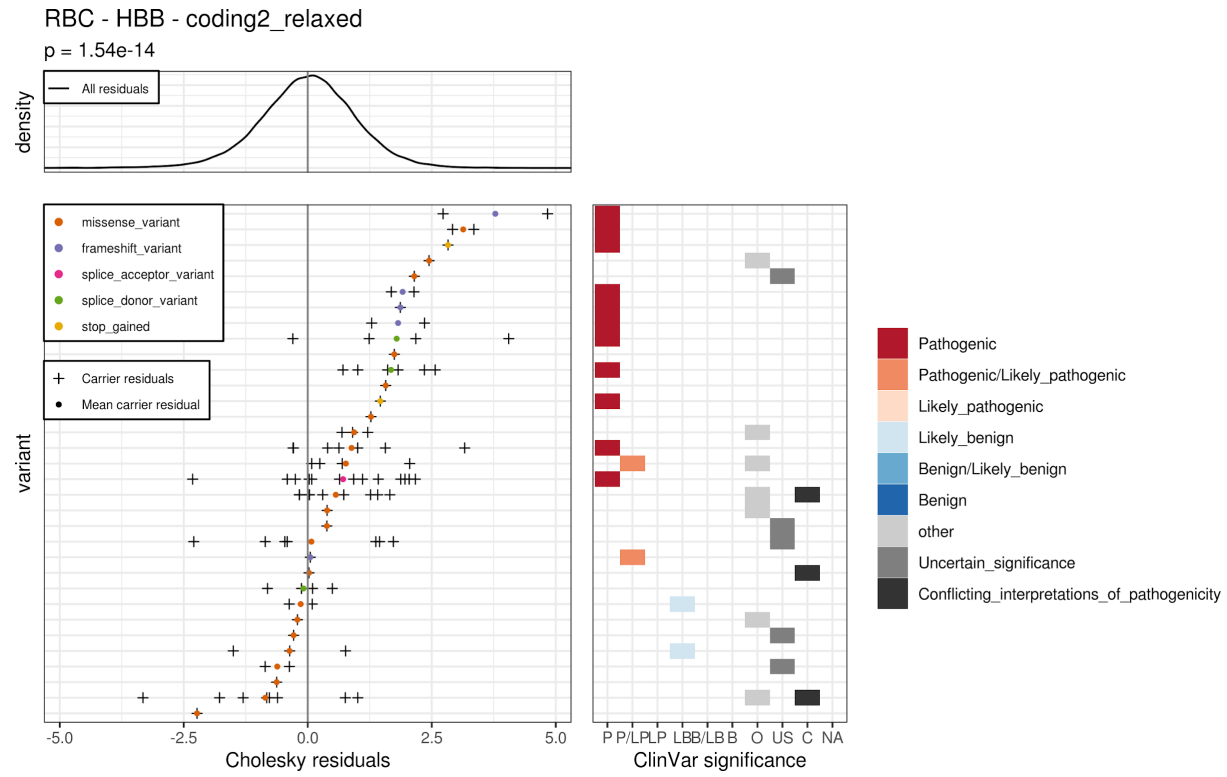
(F4)


RBC - G6PD - coding2_relaxed + noncoding_relaxed
p = 3.07e-09
density
All residuals
variant
missense_variant
stop_gained
Carrier residuals
Mean carrier residual
-5.0
-2.5
0.0
2.5
5.0
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

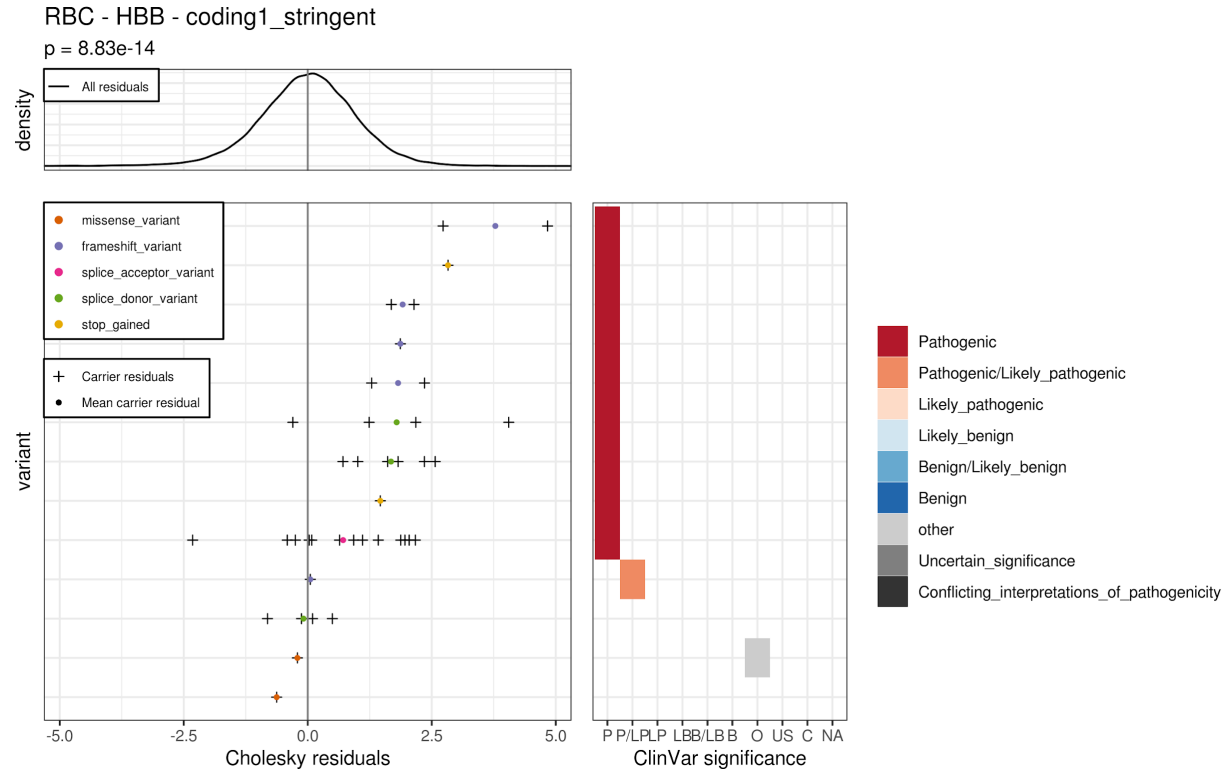
(F5)


RBC - G6PD - coding2_relaxed + noncoding_stringent
p = 3.07e-09
density
All residuals
variant
missense_variant
stop_gained
Carrier residuals
Mean carrier residual
-5.0
-2.5
0.0
2.5
5.0
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

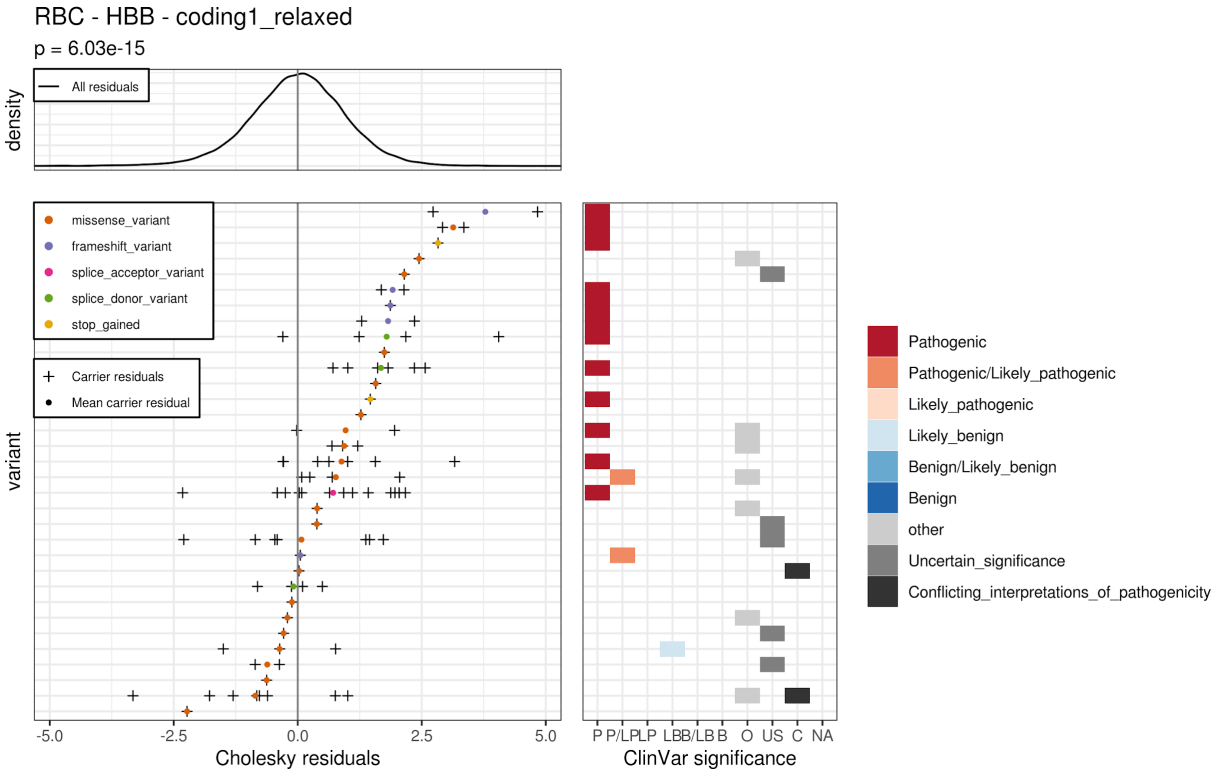
(F6)


RBC - HBB - coding2_relaxed
p = 1.54e-14
density
All residuals
variant
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-5.0
-2.5
0.0
2.5
5.0
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

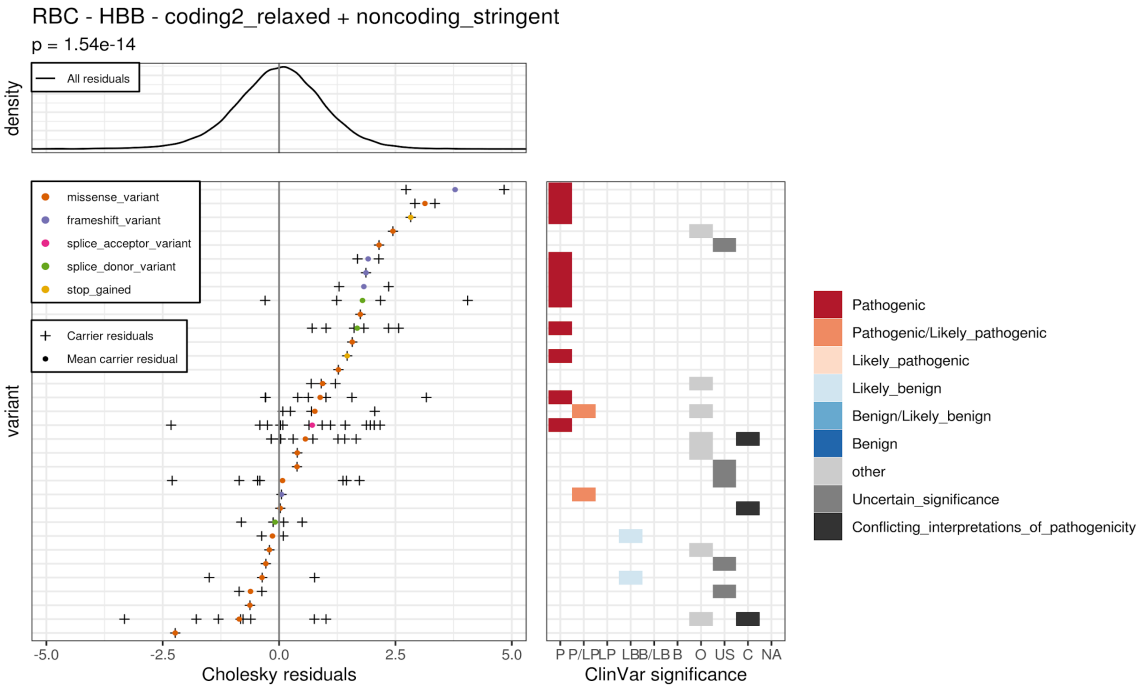
(F7)


RBC - HBB - coding1_stringent
p = 8.83e-14
density
All residuals
variant
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-5.0
-2.5
0.0
2.5
5.0
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(F8)

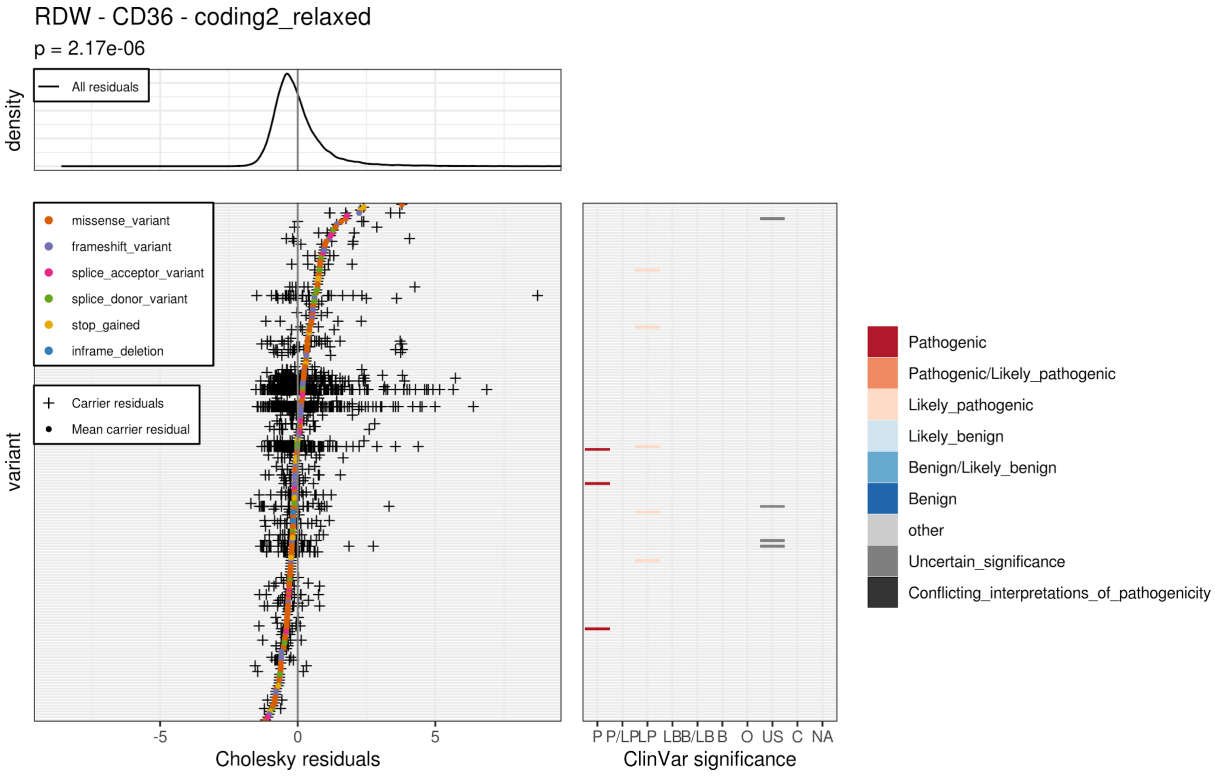

RBC - HBB - coding1_relaxed
p = 6.03e-15
density
All residuals
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-5.0
-2.5
0.0
2.5
5.0
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(F9)

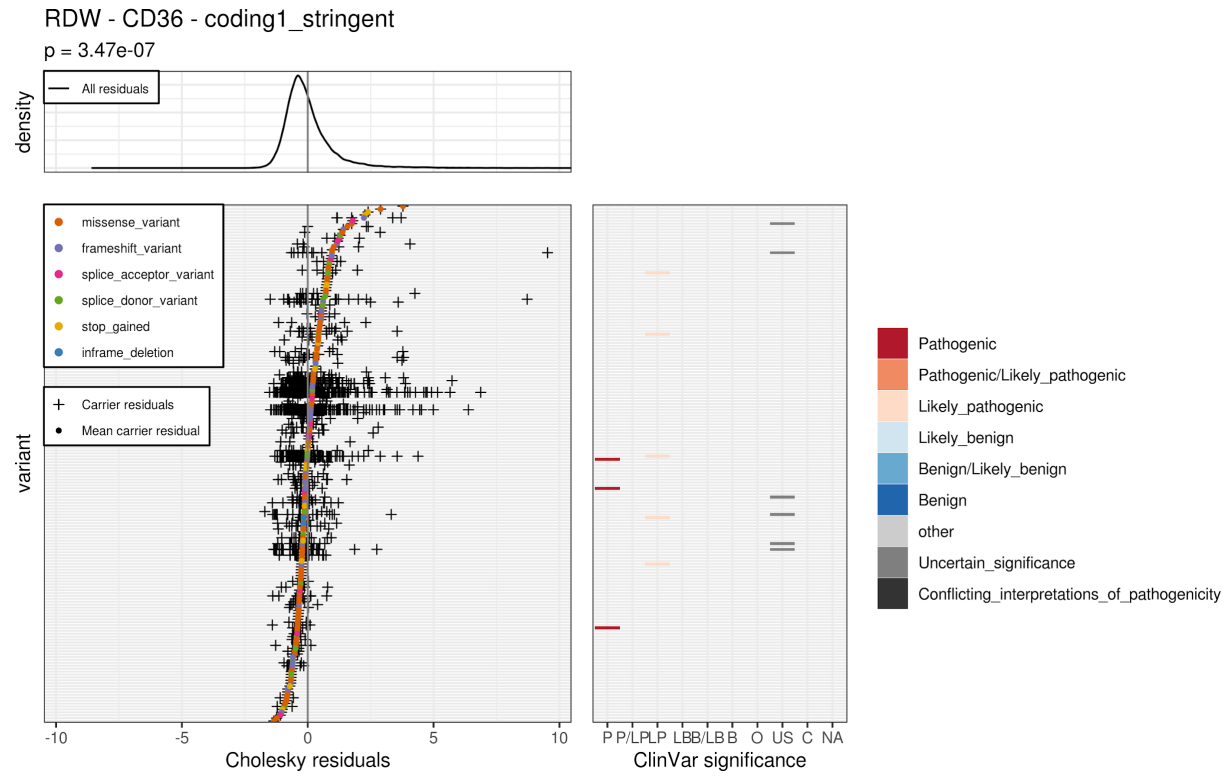

RBC - HBB - coding2_relaxed + noncoding_stringent
p = 1.54e-14
density
All residuals
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-5.0
-2.5
0.0
2.5
5.0
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(G)

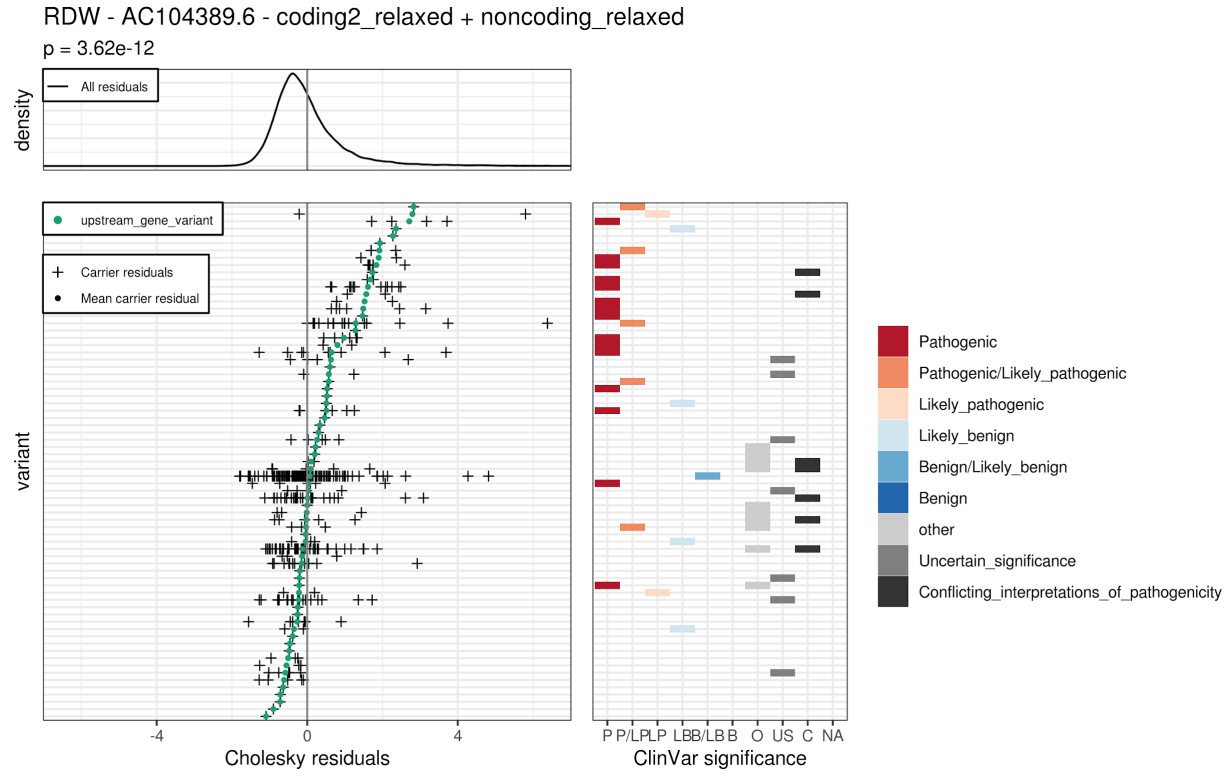
(G1)


RDW - CD36 - coding2_relaxed
p = 2.17e-06
density
All residuals
variant
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
inframe_deletion
Carrier residuals
Mean carrier residual
-5
0
5
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

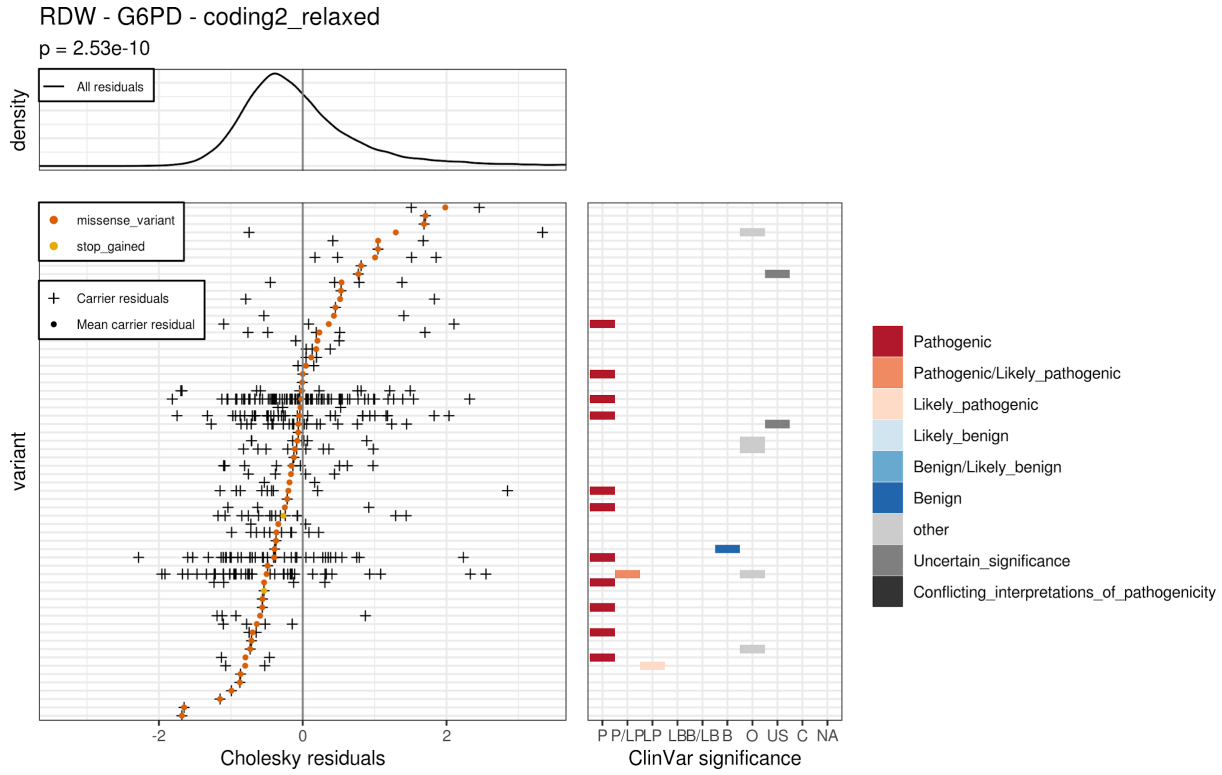
(G2)


RDW - CD36 - coding1_stringent
p = 3.47e-07
density
All residuals
variant
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
inframe_deletion
Carrier residuals
Mean carrier residual
-10
-5
0
5
10
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

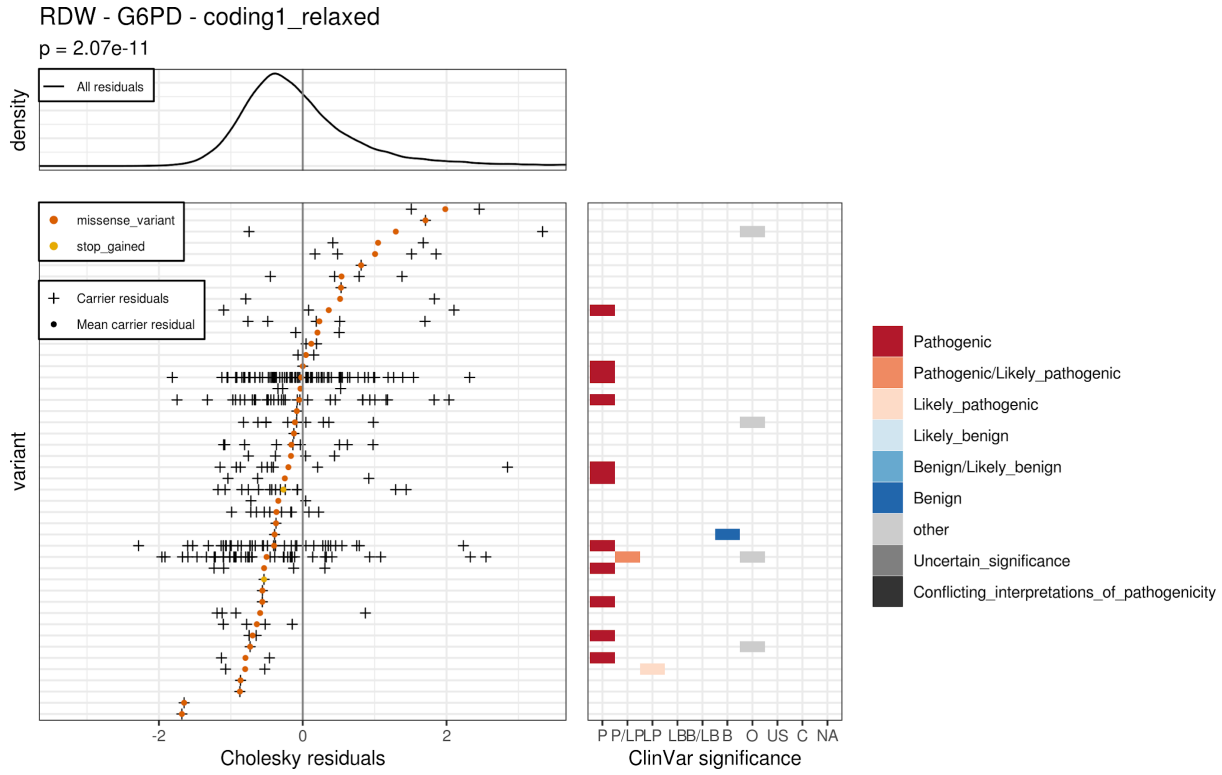
(G3)


RDW - AC104389.6 - coding2_relaxed + noncoding_relaxed
p = 3.62e-12
density
All residuals
variant
upstream_gene_variant
Carrier residuals
Mean carrier residual
-4
0
4
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

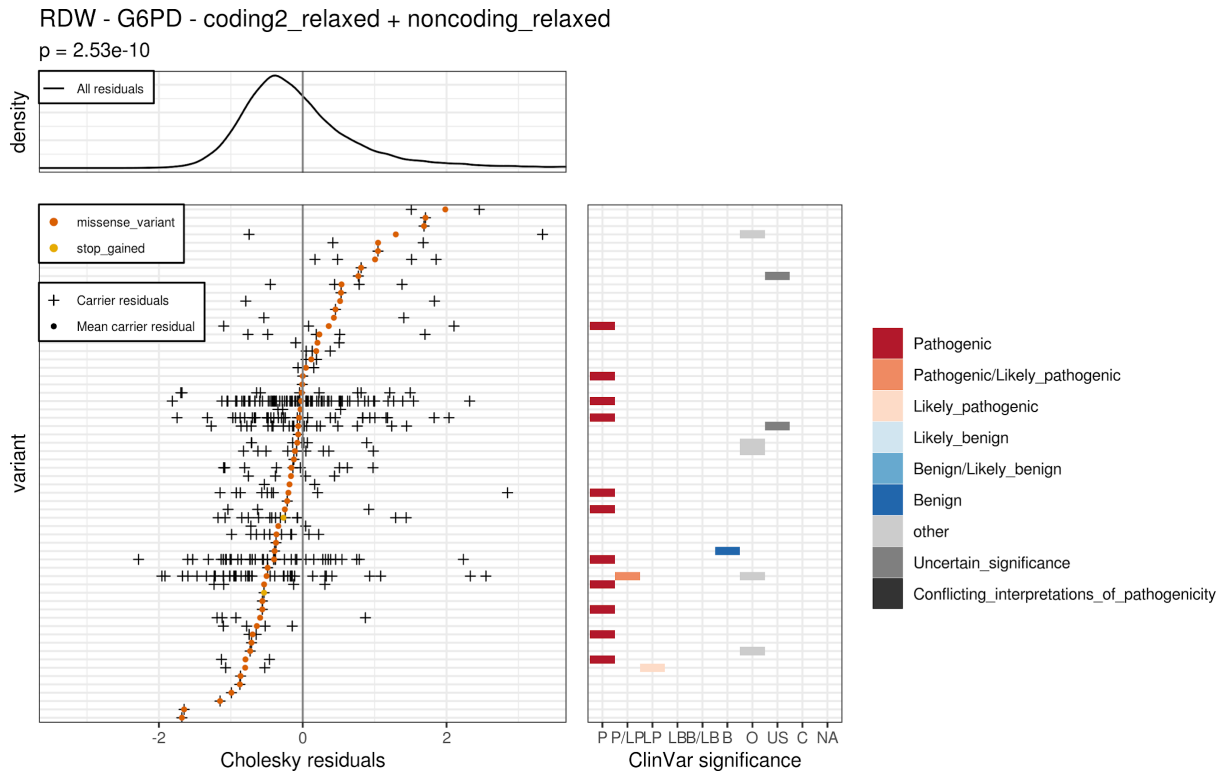
(G4)


RDW - G6PD - coding2_relaxed
p = 2.53e-10
density
All residuals
missense_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-2
0
2
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

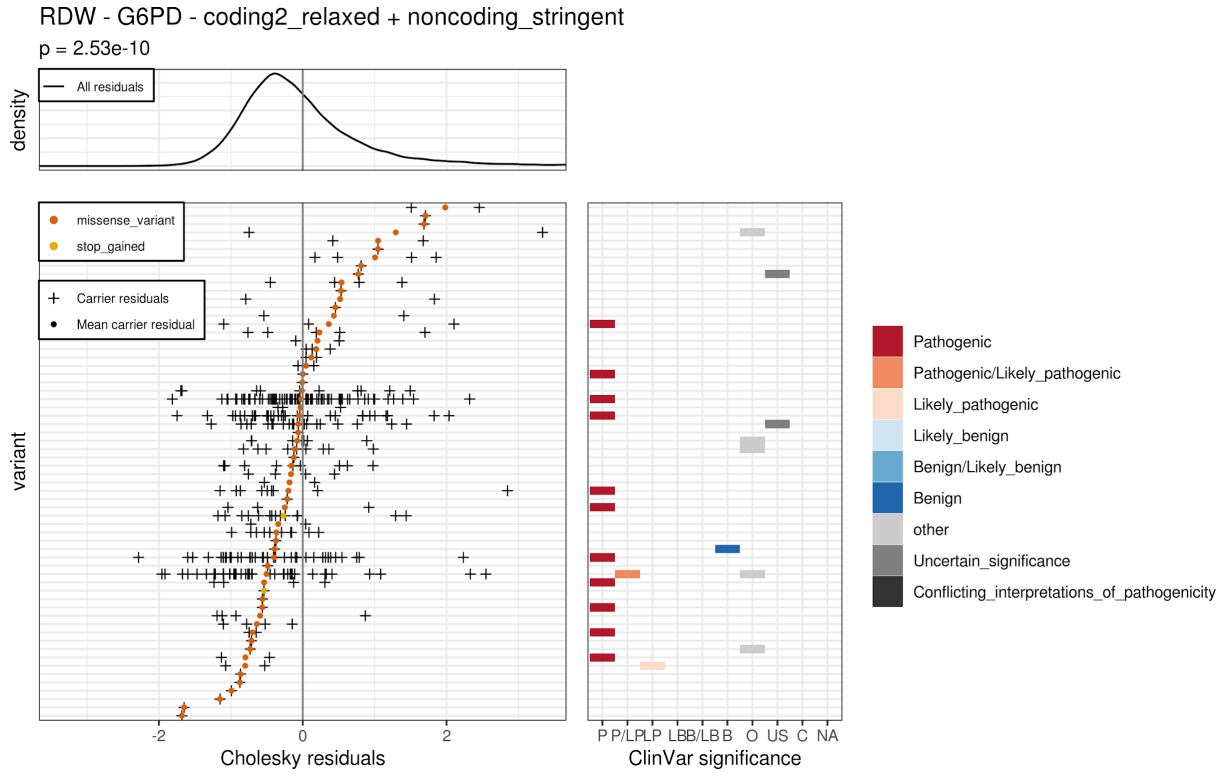
(G5)


RDW - G6PD - coding1_relaxed
p = 2.07e-11
density
All residuals
missense_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-2
0
2
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

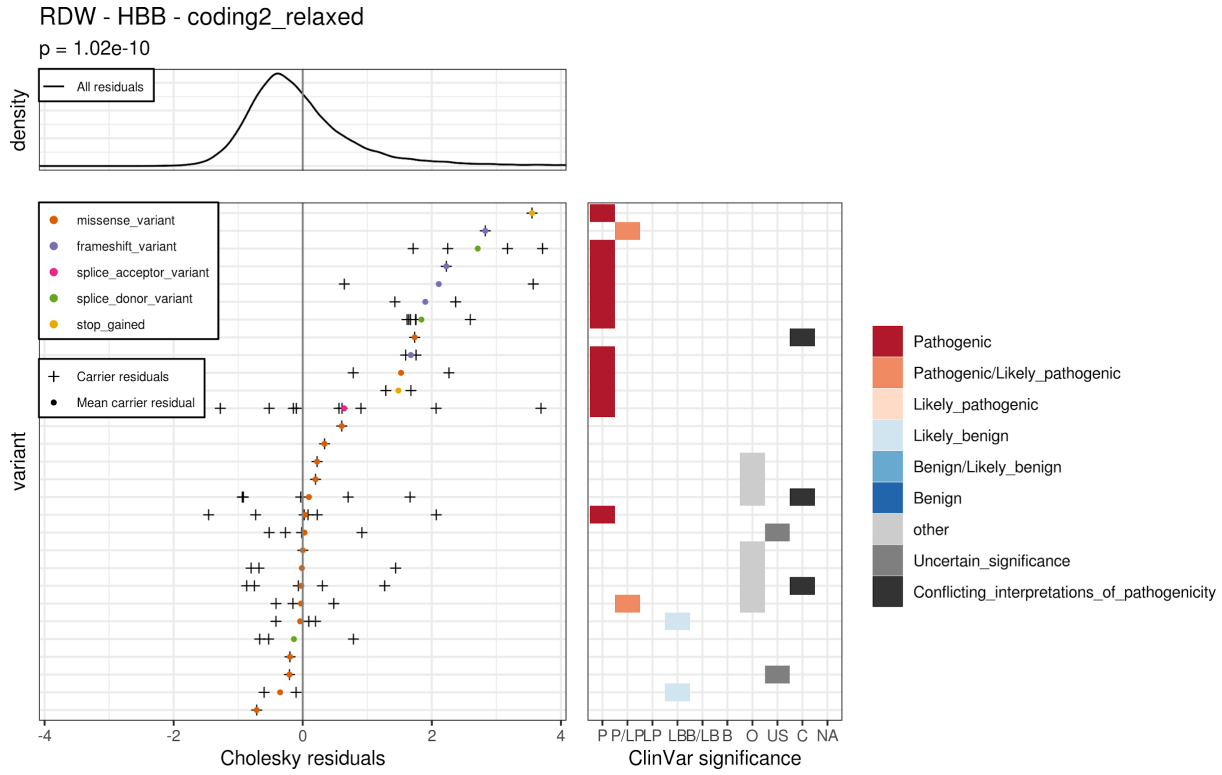
(G6)


RDW - G6PD - coding2_relaxed + noncoding_relaxed
p = 2.53e-10
density
All residuals
missense_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-2
0
2
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

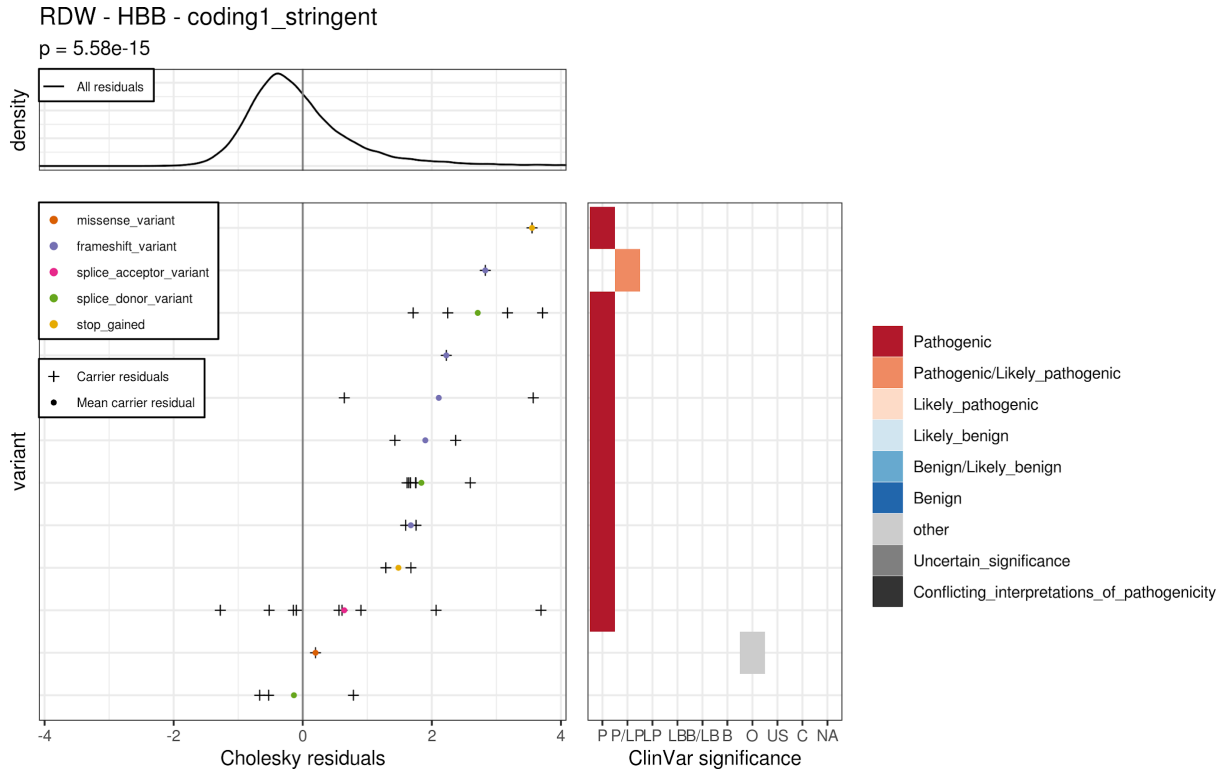
(G7)


RDW - G6PD - coding2_relaxed + noncoding_stringent
p = 2.53e-10
density
All residuals
missense_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-2
0
2
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

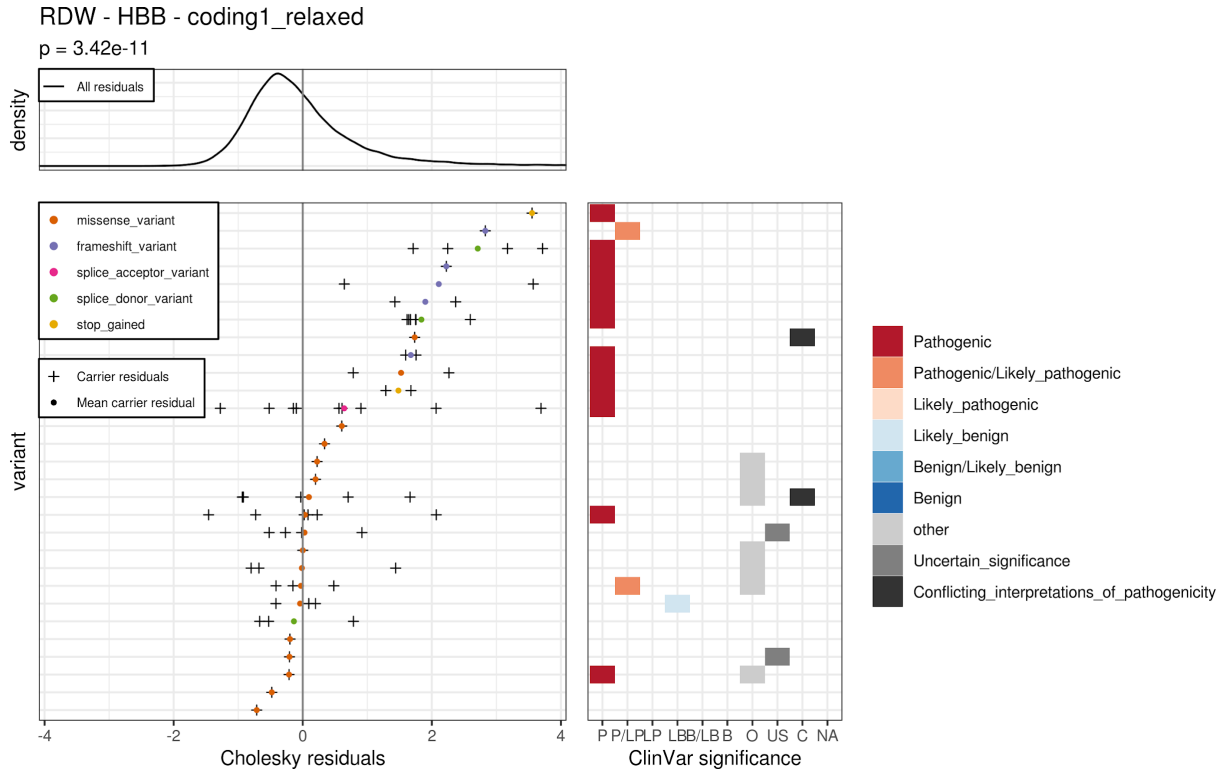
(G8)


RDW - HBB - coding2_relaxed
p = 1.02e-10
density
All residuals
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-4
-2
0
2
4
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

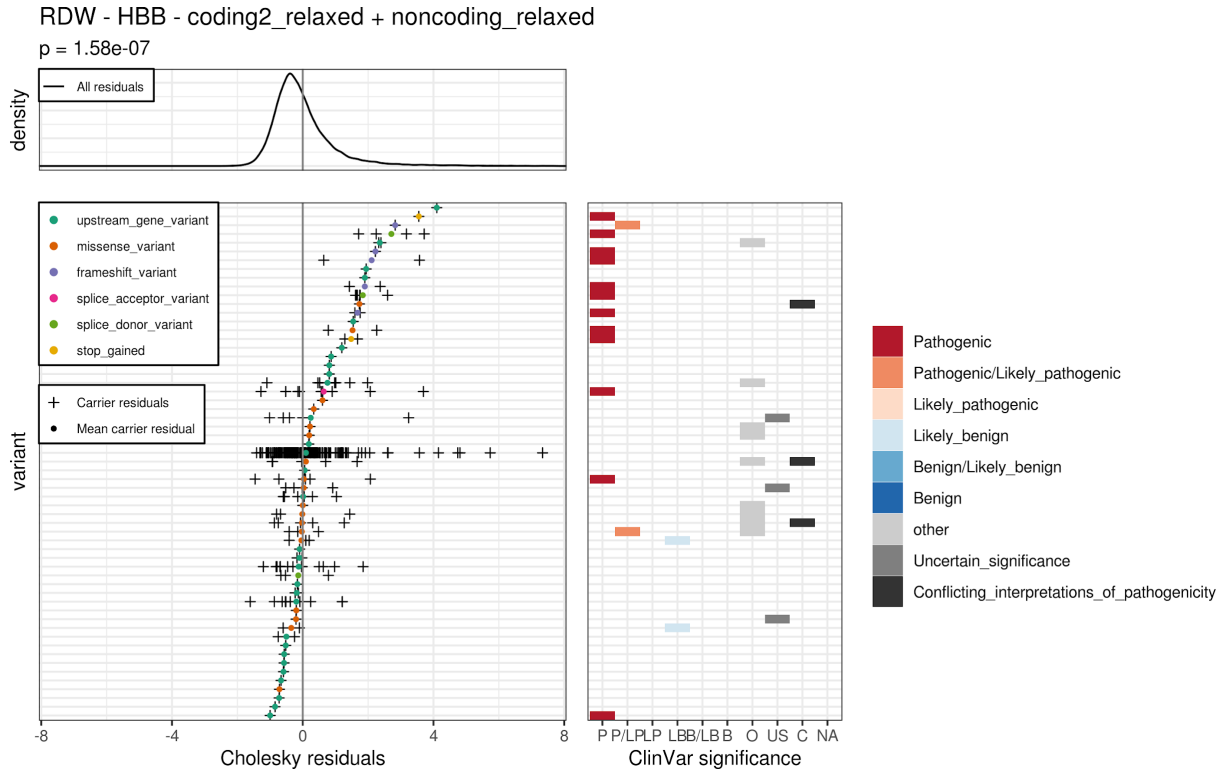
(G9)


RDW - HBB - coding1_stringent
p = 5.58e-15
density
All residuals
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
variant
-4
-2
0
2
4
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

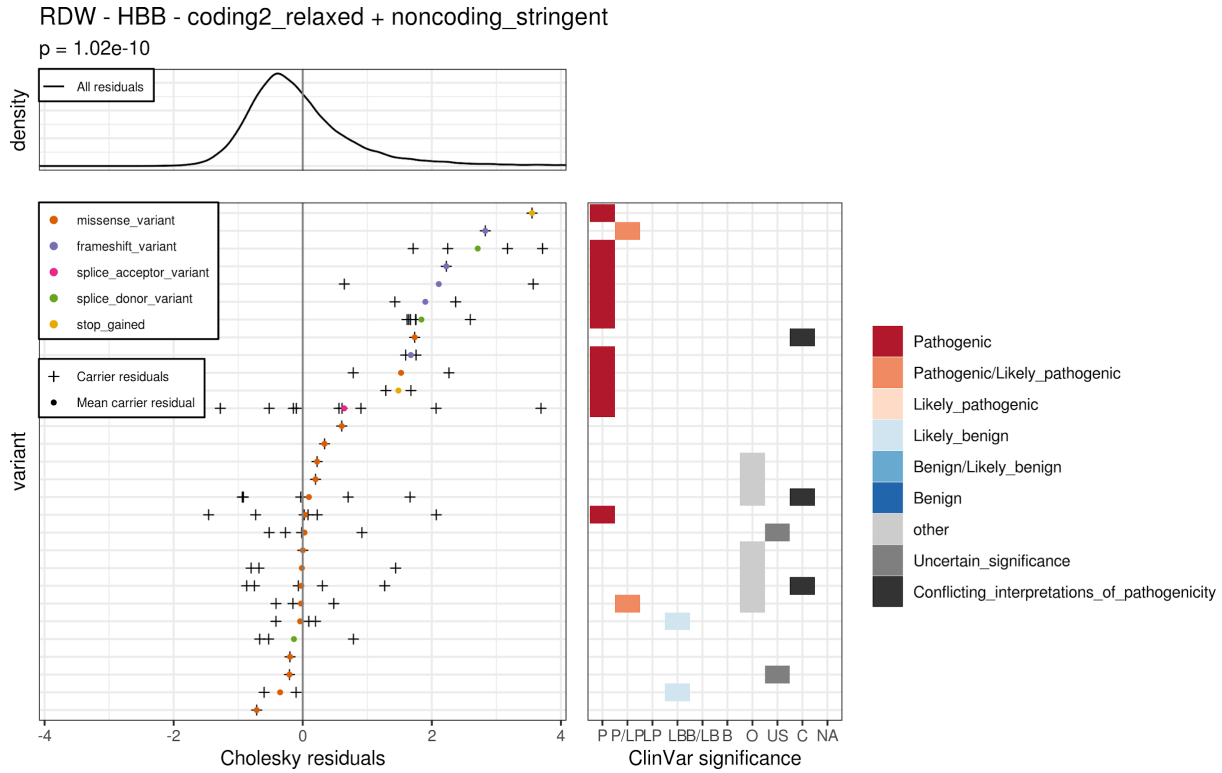
(G10)


RDW - HBB - coding1_relaxed
p = 3.42e-11
density
All residuals
variant
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-4
-2
0
2
4
Cholesky residuals
P P/LP LP LBB/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

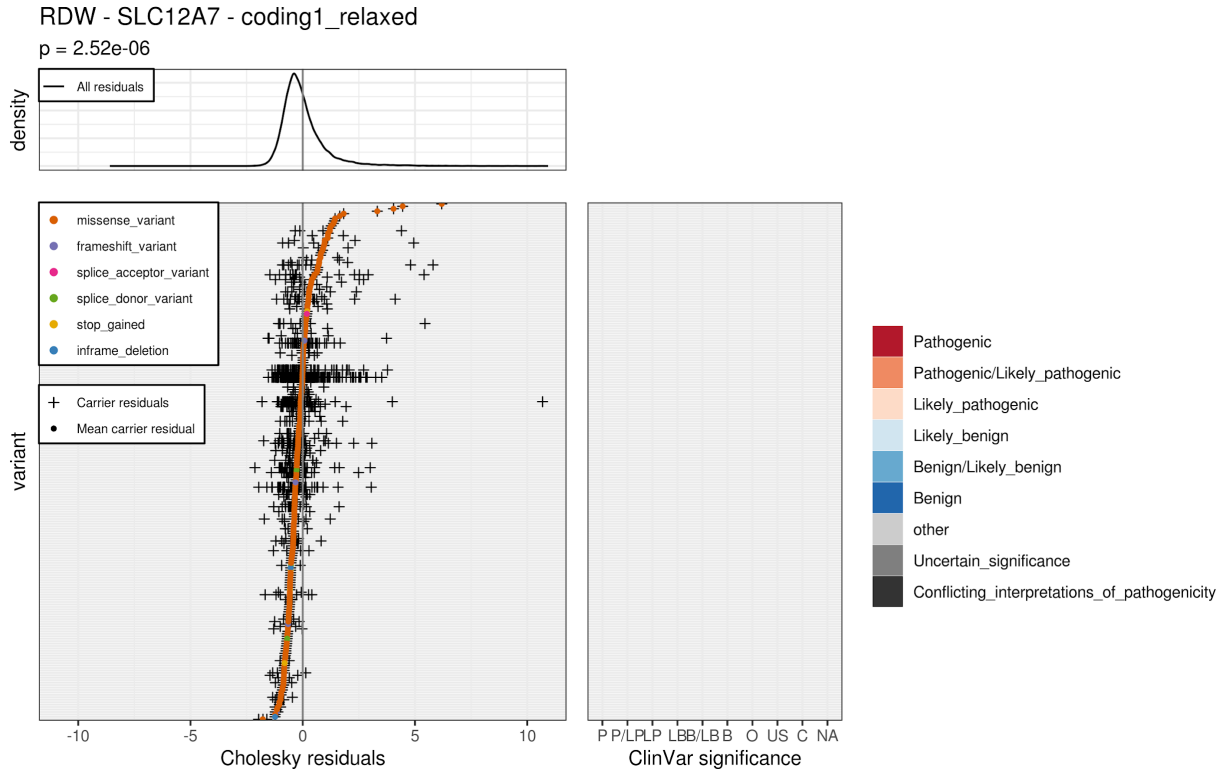
(G11)


RDW - HBB - coding2_relaxed + noncoding_relaxed
p = 1.58e-07
density
All residuals
variant
upstream_gene_variant
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-8
-4
0
4
8
Cholesky residuals
P P/LP LP LBB/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

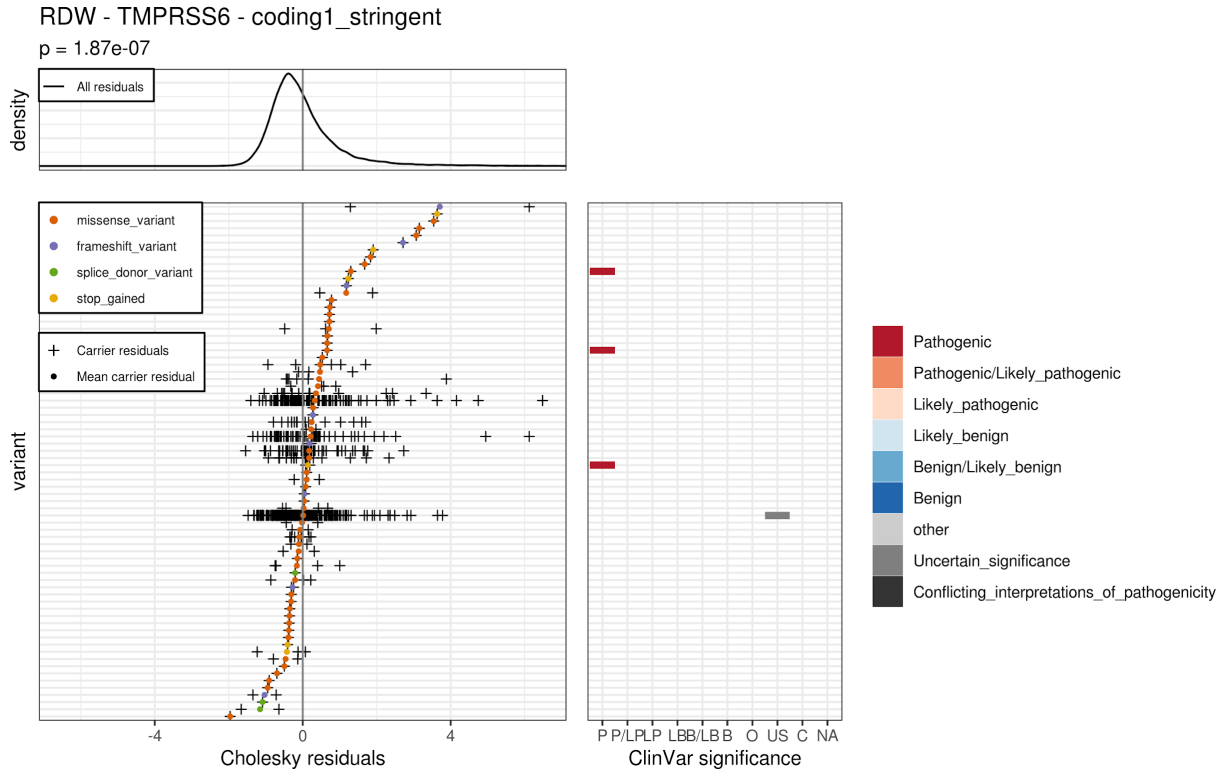
(G12)


RDW - HBB - coding2_relaxed + noncoding_stringent
p = 1.02e-10
density
All residuals
variant
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-4
-2
0
2
4
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(G13)


RDW - SLC12A7 - coding1_relaxed
p = 2.52e-06
density
All residuals
variant
missense_variant
frameshift_variant
splice_acceptor_variant
splice_donor_variant
stop_gained
inframe_deletion
Carrier residuals
Mean carrier residual
-10
-5
0
5
10
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

(G14)


RDW - TMPRSS6 - coding1_stringent
p = 1.87e-07
density
All residuals
variant
missense_variant
frameshift_variant
splice_donor_variant
stop_gained
Carrier residuals
Mean carrier residual
-4
0
4
Cholesky residuals
P P/LP LP LB B/LB B O US C NA
ClinVar significance
Pathogenic
Pathogenic/Likely_pathogenic
Likely_pathogenic
Likely_benign
Benign/Likely_benign
Benign
other
Uncertain_significance
Conflicting_interpretations_of_pathogenicity

Supplemental Tables

Table S1. Counts of participants by HARE group for each RBC phenotype

	Amish	Asian	Black	Central American	Cuban	Dominican	Mexican	Puerto Rican	South American	White
HCT	1102	654	14474	708	2037	2049	3556	4977	708	32222
HGB	1102	653	14454	708	2037	2048	3556	4974	706	32223
MCH	1102	447	11246	708	2002	2049	3435	4934	706	19612
MCHC	1102	447	13112	708	2002	2049	3434	4934	706	24154
MCV	1102	447	12285	708	2002	2049	3432	4934	706	21165
RBC	1102	384	10747	682	1984	1938	3413	4448	654	19118
RDW	0	447	6776	662	2002	1936	1898	4833	647	10184

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

Table S2. Pairwise trait correlation (upper triangle) and the number of samples used to calculate the correlations (lower triangle)

	HCT	HGB	MCH	MCHC	MCV	RBC	RDW
HCT		0.93211	0.21225	0.03469	0.23772	0.74892	-0.27781
HGB	62447		0.35214	0.31502	0.27268	0.70519	-0.38713
MCH	46099	46083		0.52655	0.87826	-0.33338	-0.44973
MCHC	52628	52612	46109		0.16304	-0.05466	-0.37418
MCV	48807	48791	46116	48816		-0.3458	-0.3531
RBC	44326	44309	44430	44334	44340		-0.0827
RDW	29244	29235	29350	29254	29261	27572	

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

Table S3. Basic characteristics of each participating study in TOPMed stratified by race/ethnicity
See Excel file.

Table S4. Number of SNVs and indels tested for each RBC trait in TOPMed

Trait	Indel	SNV	Total
RDW	5356149	70775433	76131582
RBC	6497451	86154571	92652022
MCH	6637757	88043681	94681438
MCV	6834916	90671951	97506867
MCHC	7089902	94110688	101200590
HGB	7719013	102632640	110351653
HCT	7722116	102674666	110396782

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

Table S5. Previously reported variants and indication of inclusion in the conditional analysis in TOPMed

See Excel file.

Table S6. Guide sequence used in the present study

Guide	**Sequence (5'-3')**
Guides for CRISPR/Cas9 editing	
RUVBL1	ACTACTTACCAATGGCCCTG
Neutral locus	GTAAGCTTAAAACATTAGTA
Guide for C base editing	
rs112097551_C9	GCAAGTAACGGATGCAGGGA

Table S7. Summary of PCR primers used in the present study

Gene symbol	Direction	Sequence (5'-3')
PCR primers for Sanger sequencing		
RUVBL1	Forward	ACTACTTACCAATGGCCCTG
	Reverse	GAGACAGAGAATCCCATGGG
RPN1	Forward	GTAGGTCCTCAGAGCGCGTG
	Reverse	CAGAGTCATCCAAAATAAGG
rs112097551	Forward	TCCTCTGTCCTTCCTTTCC
	Reverse	CATCTTGCCGATCTCTGAAC
Neutral locus	Forward	CCATGAGACAAGGAAGTAGTG
	Reverse	AGCAGTGGTGAGGAGAATA
Real-time qPCR primers		
EEFSEC	Forward	GAGCGGCAAGTTCAAGAT
	Reverse	GTGGGTGTCGAAGACATAAC
GATA2	Forward	TACAGCAGCGGACTCTT
	Reverse	GGTTCTGCCCATTCATCTT
RPN1	Forward	ACCAGCCACCTCCTTATT
	Reverse	GGTCCACAAACCTCATCTTC
RAB7A	Forward	CCTAGATAGCTGGAGAGATGAG
	Reverse	CTGGTCTCAAAGTAGGGAATG
RUVBL1	Forward	AAGGAGACCAAGGAAGTTTATG
	Reverse	CAGCTTCTACTCGCTCTTTC
GAPDH	Forward	ACCCAGAAGACTGTGGATGG
	Reverse	TTCAGCTCAGGGATGACCTT

Table S8. Lambda values in the single-variant association analyses in TOPMed

Trait	Lambda values		
	Unconditional analysis	Trait-specific conditional analysis	Trait-agnostic conditional analysis
HCT	1.021	1.020	1.015
HGB	1.019	1.019	1.015
MCH	1.036	1.034	1.029
MCHC	1.024	1.022	1.017
MCV	1.038	1.036	1.030
RBC	1.025	1.021	1.018
RDW	1.033	1.024	1.019

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

Table S9. Lead variants at the genome-wide significant loci of the marginal tests in TOPMed
See Excel file.

Table S10. Genome-wide significant variants at the 12 novel loci in the trait-specific conditional analysis in TOPMed

See Excel file.

Table S11. Ancestry-specific allele frequencies of the 14 novel lead variants at the 12 loci

Variant	Chr	Pos	Gene	Alternative allele	Reference allele	Alternative allele frequencies (%)			
						European	African	Hispanic Latino	East Asian
rs112097551	3	1.3E+08	*RNP1*	A	G	0.069	0.940	0.400	0
rs116635225	5	9.6E+07	*ELL2*	A	G	0.074	3.900	0.700	0
rs986415672	10	1.3E+08	*10q26*	T	C	0.011	0	0	0
rs11549407	11	5226774	*HBB*	A	G	0.016	0	0	0
rs34598529	11	5227100	*HBB*	C	T	0	0.320	0	0
rs535577177	11	7E+07	*SHANK2*	A	G	0	0	0.100	0
rs370308370	14	1E+08	*EIF5/MARK3*	A	G	0	0	0	0.910
rs868351380	16	55649	*HBA1/HBA2*	C	G	0.005	0	0.400	0
rs372755452	16	199621	*HBA1/HBA2*	A	AG	0	0	0	1.100
rs763477215	16	8.9E+07	*PIEZO1*	A	ATCT	0.355	0	0	0.050
rs73494666	19	1253643	*MIDN*	T	C	0.614	51.7	4.700	0
rs1368500441	19	2.9E+07	*19q12*	A	G	0.005	0	0	0
rs228914	22	3.7E+07	*TMPRSS6*	A	C	88.7	96.6	79.2	99.8
rs76723693	X	1.5E+08	*G6PD*	G	A	0	0.563	0.077	0

Chr, chromosome; Pos, position.

Table S12. Replication results of the novel findings and the lead independent signals
See Excel file.

Table S13. Independent signals in the step-wise conditional analysis
See Excel file.

Table S14. Phenotypic variance explained by variants identified in the single variant association analysis

Trait	All	Known	Novel
HCT	0.034	0.033	0.001
HGB	0.043	0.040	0.003
MCH	0.213	0.184	0.030
MCHC	0.047	0.041	0.006
MCV	0.179	0.153	0.028
RBC	0.126	0.117	0.010
RDW	0.118	0.109	0.009

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

Table S15. Summary of significant genes in the aggregated association analysis in TOPMed
See Excel file.

Table S16. Summary of significant genes in the aggregated association analysis adjusting for known and novel findings in TOPMed

See Excel file.

Table S17. Annotation of the rare variants identified in the aggregated analysis in TOPMed
See Excel file.

Table S18. Summary of pLoF and pKO variants in TOPMed freeze8 data [1]

Population	N [2]	No. of pLoF variants	No. of genes with at least one individual who is a pKO
African	9,870	55,750	1,617
Asian	231	4,377	395
European	25,569	114,401	1,634
Hispanic	9,757	53,105	1,557

pLoF, predicted loss-of-function; pKO, predicted gene knockout; N, sample size.

1 No minor allele frequency filter was applied.

2 Sample sizes represented the number of individuals with blood-cell traits and genotype data available.

Table S19. pLoF variants associated with RBC traits at P<1E-4 in TOPMed [1]

Population	Trait	Chr	Gene	rsID	Variant	MAF (%)	Type	Beta	SE	*P*
African	MCV	2	*WDSUB*	rs377262700	chr2:159236041_G_A	0.021	stopgain	-2.682	0.592	6.09E-06
African	RDW	7	*CD36*	rs3211938	chr7:80671133_T_G[2]	9.340	stopgain	0.244	0.043	**1.24E-08**
Hispanic	MCV	4	*SNX25*	rs1200775460	chr4:185339389_AG_A	0.022	frameshift	2.441	0.543	7.08E-06
Hispanic	MCH	11	*HBB*	rs11549407	chr11:5226774_G_A[3]	0.022	stopgain	-2.611	0.505	**2.36E-07**
Hispanic	MCV	11	*HBB*	rs11549407	chr11:5226774_G_A[3]	0.022	stopgain	-2.970	0.506	**4.58E-09**
Hispanic	RBC	11	*HBB*	rs11549407	chr11:5226774_G_A[3]	0.022	stopgain	2.272	0.506	7.18E-06
European	HCT	11	*HBB*	rs11549407	chr11:5226774_G_A[3]	0.018	stopgain	-1.545	0.333	3.39E-06
European	HGB	11	*HBB*	rs11549407	chr11:5226774_G_A[3]	0.018	stopgain	-2.052	0.332	**6.65E-10**
European	MCH	11	*HBB*	rs11549407	chr11:5226774_G_A[3]	0.017	stopgain	-2.974	0.494	**1.73E-09**
European	MCV	11	*HBB*	rs11549407	chr11:5226774_G_A[3]	0.015	stopgain	-2.981	0.494	**1.66E-09**
European	HCT	11	*CD6*	rs759187282	chr11:61017803_G_T	0.006	stopgain	-2.573	0.575	7.73E-06
European	HGB	11	*CD6*	rs759187282	chr11:61017803_G_T	0.006	stopgain	-2.575	0.574	7.39E-06
Meta-analysis	RDW	1	*SMIM1*	rs566629828	chr1:3775433_AGTCAGCCTAGGGGCTGT_A[4]	1.610	frameshift	0.303	0.068	8.22E-06
Meta-analysis	MCV	18	*SERPINB11*	rs760239610	chr18:63712688_C_CATCAGGTA	0.150	frameshift	-0.681	0.150	5.60E-06

pLoF, predicted loss-of-function; HCT, hematocrit; HGB, hemoglobin; MCH, mean corpuscular hemoglobin; MCV, mean corpuscular volume; RBC, red blood cell count; RDW, red blood cell width; Chr, chromosome; MAF, minor allele frequency.

1 The P values of pLoF variants that reached genome-wide significance were in bold (African: P<8.97E-7; Hispanic: P<9.42E-7; European: P<4.37E-7).

2 Well known CD36 null allele.

3 Well known beta-thalassemia allele.

4 This frameshift indel is responsible for the Vel blood group.

Table S20. pKO variants associated with RBC traits at P<1E-4 in TOPMed [1]

Population	Trait	Chr	Gene	rsID	Variants	MAF (%)	Type	N		Beta	SE	*P*
								Total	KO			
African	MCH	7	*ZNF3*	rs777843966	chr7:100064797_GT_G	0.008	frameshift	6042	200	0.277	0.070	8.53E-05
				rs987730433	chr7:100064875_C_CA	0.008	frameshift					
				rs71689664	chr7:100064888_GTAGT_G	18.3	frameshift					
				rs745468385	chr7:100071151_ACT_A	0.008	frameshift					
				rs774923137	chr7:100071181_TG_T	0.008	frameshift					
				rs988854061	chr7:100079535_C_T	0.008	splicing					
African	MCV	7	*ZNF3*	rs777843966	chr7:100064797_GT_G	0.007	frameshift	7198	239	0.267	0.064	3.60E-05
				rs987730433	chr7:100064875_C_CA	0.007	frameshift					
				rs71689664	chr7:100064888_GTAGT_G	18.4	frameshift					
				rs745468385	chr7:100071151_ACT_A	0.007	frameshift					
				rs774923137	chr7:100071181_TG_T	0.007	frameshift					
				rs988854061	chr7:100079535_C_T	0.007	splicing					

pKO, predicted gene knockout; MCH, mean corpuscular hemoglobin; MCV, mean corpuscular volume; Chr, chromosome; MAF, minor allele frequency.
1 No pKO variant reached genome-wide significance (African: P<3.09E-5; Hispanic: P<3.21E-5; European: P<3.06E-5).

Supplemental Methods

Participating studies

Amish

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

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

ARIC

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

BioMe

The Charles Bronfman Institute for Personalized Medicine at Mount Sinai Medical Center (MSMC), BioMe Biobank, founded in September 2007, is an ongoing, broadly-consented electronic health record-linked

clinical care biobank that enrolls participants non-selectively from the Mount Sinai Medical Center patient population. The MSMC serves diverse local communities of upper Manhattan, including Central Harlem (86% African American), East Harlem (88% Hispanic/Latino), and Upper East Side (88% Caucasian/White) with broad health disparities.

CARDIA

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

CHS

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

COPDGene

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

in this NHLBI WGS project, including severe COPD subjects and non-COPD smoking controls. The COPDGene Study web site is: http://www.copdgene.org/.

FHS

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

GeneSTAR

In 1982 The Johns Hopkins Sibling and Family Heart Study was created to study patterns of coronary heart disease and related risk factors in families with early-onset coronary disease, identified from10 Baltimore area Hospitals. GeneSTAR continues to study mechanisms of coronary heart disease and stroke in families using novel models and exciting new methods. GeneSTAR is a family-based study in initially healthy brothers and sisters, and offspring of people with early-onset coronary disease, The goal is to discover and amplify mechanisms of stroke and coronary heart disease. Our African American and European American family cohort has undergone extensive screening, genetic testing, and follow-up for new cardiovascular disease, stroke, and other clinical events for 5 to 32 years.

HCHS/SOL

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

extensive psycho-social and clinical assessments were conducted during 2008-2011. Annual telephone follow-up interviews are ongoing since study inception. During the 2014-2017 second visit, the participants were re-examined again of various health outcomes of interest.

JHS

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

MESA

The MESA study is a study of the characteristics of subclinical cardiovascular disease (disease detected non-invasively before it has produced clinical signs and symptoms) and the risk factors that predict progression to clinically overt cardiovascular disease or progression of the subclinical disease [7]. MESA researchers study a diverse, population-based sample of 6,814 asymptomatic men and women aged 45-84. Thirty-eight percent of the recruited participants are white, 28 percent African-American, 22 percent

Hispanic, and 12 percent Asian, predominantly of Chinese descent. Participants were recruited from six field centers across the United States: Wake Forest University, Columbia University, Johns Hopkins University, University of Minnesota, Northwestern University and the University of California - Los Angeles.

SAFS

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

WHI

The Women's Health Initiative (WHI) is a long-term, prospective, multi-center cohort study that investigates post-menopausal women's health [8]. WHI was funded by the National Institutes of Health and the National Heart, Lung, and Blood Institute to study strategies to prevent heart disease, breast cancer, colon cancer, and osteoporotic fractures in women 50-79 years of age. WHI involves 161,808 women recruited between 1993 and 1998 at 40 centers across the US. The study consists of two parts: the WHI Clinical Trial which was a randomized clinical trial of hormone therapy, dietary modification, and calcium/Vitamin D supplementation, and the WHI Observational Study, which focused on many of the inequities in women's health research and provided practical information about the incidence, risk factors, and interventions related to heart disease, cancer, and osteoporotic fractures. For TOPMed WGS, the study over-sampled participants with incident stroke and VTE. The remaining samples were age- and ethnicity-matched controls without stroke or VTE.

Phenotype harmonization

Because multiple studies contributed to the analysis, RBC phenotypes were harmonized across studies such that they could be analyzed together (https://www.biorxiv.org/content/10.1101/2020.06.18.146423v1). For most studies, variables were obtained from dbGaP. Data for the BioME study, the COPDGene study, and some participants in the WHI study were transferred directly to investigators. The study variables were QC'd to identify recording errors or other problematic data. After QC, variables were converted to a consistent measurement unit across studies. When possible, harmonized variables were calculated using study-derived variables. If QC uncovered data quality issues with the study-derived variable or if the study

did not provide the specific variable of interest, the harmonized variable was instead calculated from the components (e.g., MCH = 10 * hemoglobin / red cell count). Finally, QC of the harmonized variable was performed to check that no large differences between studies remained.

Both FHS and ARIC measured phenotypes at multiple times in a given participant. For these studies, a single measurement for each participant was selected for each trait. To maximize the sample size and minimize batch effects within studies, harmonized data for different RBC traits for a single subject may have been measured at different visits. Within FHS, only the Offspring cohort had measurements at multiple exams. For these participants, the measurement at the most recent exam was chosen. For the ARIC study, the visit with the most non-missing phenotype values across all participants was chosen first. For subjects without measurements at this visit, the visit with the next most non-missing values was chosen, and so forth. As a consequence, values for the same participant for different RBC phenotypes were sometimes measured at different visits."

Trait-specific QC procedures were also performed. We excluded participants with HCT values >80% and those with HCT <5% (n=14). Similarly for HGB, we excluded participants with HGB measurements >30 g/dL and <5 g/dL (n=31). For participants from WHI, for both the HCT and HGB analyses, we excluded those with an HCT/HGB ratio >7 (n=11). Participants with MCH values >75 pg were excluded from analysis (n=2). In MCHC, we excluded participants with measurements ≥60 g/dL (n=1). For the MCV analysis, outliers with values >150 fL were excluded (n=2). In the RBC and RDW analyses, no participants were excluded based upon measured trait values.

Statistical analyses

Genetic Ancestry and Relatedness

Principal components (PCs) of genetic ancestry and pairwise relatedness measures were estimated for all 140,062 samples included in the TOPMed 'freeze 8' genotype release. Autosomal genetic variants passing the quality filter with a MAF > 0.01 and missing call rate < 0.01 were LD-pruned with an r^2 threshold of 0.1 to obtain a set of 638,486 effectively independent variants for genetic ancestry and relatedness estimation. PC-AiR [9] was used to obtain ancestry informative PCs robust to familial relatedness; the first 11 PCs showed evidence of population structure. PC-Relate [10] was then used to estimate pairwise kinship coefficients (KCs) for all pairs of samples, conditional on the genetic ancestry captured by PC-AiR PCs 1-11; these KC estimates reflect only recent genetic relatedness, e.g. due to pedigree structure. The PC-Relate KC estimates were used to construct a 4th degree sparse, block-diagonal, empirical kinship matrix (KM) for association testing, using the procedure recommended in Gogarten et al [11].: any pair of samples with estimated KC > $2^{(-11/2)} \sim 0.022$ were clustered in the same block; all KC estimates within a block of samples were kept, regardless of value; and all KC estimates between blocks were set to 0. By using a sparse block-

diagonal KM, the association tests are more computationally efficient yet recent genetic relatedness is still accounted for. We subset the freeze-wide PCs and sparse KM to the appropriate set of participants for each analysis.

HARE for Imputation of Race/Population Membership using Genetic Ancestry

Ancestry groups were based on a combination of participants reported race/ethnicity and genetic ancestry represented by PCs from PC-AiR[9]. To infer race/population group membership for participants with missing values, we used the HARE method [12] . HARE is a machine learning algorithm that uses a support vector machine (SVM) to determine stratum assignment, taking as input genetically estimated PC values and reported race/ethnicity for each participant. Strata are defined by the unique reported race/ethnicity values provided, then the HARE SVM uses the input (training) data to learn the probability of stratum membership across the entire PC space. The output of HARE consists of multinomial probability vectors of stratum membership for each participant. HARE was run on a subset of samples included in the TOPMed freeze 8 genotype release; specifically, samples for participants from non-US populations (e.g. Costa Rica) and the Amish participants (because they were very distinct in PC space) were excluded from the HARE analysis. HARE was run using the first 9 PC-AiR PCs generated on this subset of samples to represent genetic ancestry with the following reported race/population groups: Asian, Black, Central American, Cuban, Dominican, Mexican, Puerto Rican, South American, and White. The genetic data from the 31,918 participants with either unreported or non-specific (e.g. 'Multiple' or 'Other') race and population membership was included in the HARE analysis, but they were not used to train the SVM. These participants were assigned to a population stratum based on their highest HARE output probability of membership. All other participants remained in the population stratum corresponding to their reported race/population group. Amish participants were assigned to their own stratum.

Fitting the Linear Mixed Model

The linear mixed model (LMM) can be written as $Y = G\beta + X\alpha + \epsilon$, where Y is the $(n\ x\ 1)$ vector of outcome values; G is an $(n\ x\ m)$ matrix of alternate allele counts for each of the n individuals at the m variants of interest ($m = 1$ for a single variant analysis) with effect sizes given by the $(m\ x\ 1)$ vector β; X is the $(n\ x\ k)$ matrix of fixed effect covariates including an intercept with effect sizes given by the $(k\ x\ 1)$ vector α; and $\epsilon \sim N(0, \Sigma)$ is the $(n\ x\ 1)$ vector of errors with covariance matrix Σ that captures both genetic covariance due to relatedness/kinship and residual variance structure. Given the true Σ, we could estimate β using generalized least squares (GLS). However, we can simplify this GLS problem to an ordinary least squares (OLS) problem by pre-multiplying both sides of the equation by the matrix C, the Cholesky-decomposition of Σ^{-1}, such that $C'C = \Sigma^{-1}$ and $C'\Sigma C = I$, where I is the $(n\ x\ n)$ identity matrix. Further,

by the Frisch-Waugh Lovell theorem[13], we can adjust for the covariates in the new OLS model, CX, by pre-multiplying CY and CG by the annihilator matrix $[I - (CX)((CX)'(CX))^{-1}(CX)']$. Ultimately, the original GLS problem can be re-written as the linear regression model $Y^* = G^*\beta + \epsilon^*$, where $Y^* = MY$, $G^* = MG$, and $M = [I - CX(X'C'CX)^{-1}X'C']C$. In practice, we use REML to estimate $\hat{\Sigma}$ under the null hypothesis that $\beta = 0$ (i.e. fit the null model) and calculate the estimate of the matrix $\widehat{M}$.

Score Tests and Approximate Variant Effect Sizes

Given Y^* and G^*, a joint score test for the set of m variants can be performed, where the score is $U = G^{*\prime}Y^*$, the variance of the score is $V = G^{*\prime}G^*$, and the test statistic is $T_G = U'V^{-1}U \sim \chi^2_m$. The score and the Wald tests are approximately asymptotically equivalent when β is small (as is typical for GWAS), so the variant effect sizes can be reasonably approximated from the score test as $\hat{\beta} \approx V^{-1}U = (G^{*\prime}G^*)^{-1}(G^{*\prime}Y^*)$, and their covariance matrix can be reasonably approximated from the score test as $\widehat{var}(\hat{\beta}) \approx V^{-1} = (G^{*\prime}G^*)^{-1}$.[14] Note that these are the score tests used for the single variant association analysis, where each variant genome-wide is tested individually (i.e. $m = 1$).

Proportion of Variance Explained Jointly by a set of variants

To estimate the proportion of phenotypic variance explained (PVE) by the m variants in G, we use the formula $PVE = 1 - RSS_1/RSS_0$, where RSS_0 and RSS_1 are the residual sums of squares computed from the null model, and the model including the m variants of interest, respectively. Under the null model, we have that $RSS_0 = Y^{*\prime}Y^*$, and from the model $Y^* = G^*\beta + \epsilon^*$, we have that $RSS_1 = (Y^* - G^*\hat{\beta})'(Y^* - G^*\hat{\beta})$. Using the approximation for $\hat{\beta}$ given above, we get that $RSS_1 \approx Y^{*\prime}Y^* - Y^{*\prime}G^*(G^{*\prime}G^*)^{-1}G^{*\prime}Y^* = Y^{*\prime}Y^* - T_G$, and the estimate $\widehat{PVE} \approx T_G/(Y^{*\prime}Y^*)$. It's worth noting that using this approach to estimate the PVE for the set of m variants jointly should provide a more accurate estimate than estimating the PVE for each variant separately and summing, as this joint approach accounts for the covariance between the variant effect sizes, as measured by V^{-1} (the separate approach is equivalent to $\widehat{PVE} = [U'diag(V)^{-1}U]/(Y^{*\prime}Y^*)$, where $diag(V)$is an $(m\,x\,m)$ matrix of just the diagonal of the V matrix). This joint PVE calculation is implemented in the GENESIS software [11] with the jointScoreTest function.

Conditional analyses

We performed three types of conditional analysis in the discovery stage. The first conditional analyses adjusted each trait for variants that were previously reported to be associated with the particular RBC trait and that passed the QC filter. These variants were pruned to a set with linkage disequilibrium (LD) r2 < 0.8 such that a variant with a more significant p-value was preferentially retained over those with higher p-

values. Known variants failing the QC filter were included if any variants within 1 MB of the known variant remained significant after adjusting for the passing variants only. We refer to this analysis as the "RBC trait-specific conditional analysis". In the second conditional analysis, we included all previously reported variants for *any* of the seven RBC traits as well as any failed variants included in any of the trait-specific conditional analyses. Variants were again pruned to LD $r^2 < 0.8$ with preferential selection based on p-value. We refer to the second conditional analysis as the "RBC trait-agnostic conditional analysis". Finally, we performed iterative conditional analysis by chromosome for each trait to identify an independent set of associated variants. For this third conditional analysis, we started with the association results from the trait-specific conditional analysis. For each chromosome, we identified the most significant variant (if any, using a $5x10^{-9}$ threshold) as the 'peak variant' and then fit a new null model adjusted for both the previous set of conditional variants from the trait-specific conditional analyses as well as this peak variant, and calculated new score test statistics. If any variant was significant at the $5x10^{-9}$ level in the new score tests (regardless of its significance level in the original trait-specific conditional results), we performed a second round of conditional analysis, re-estimating the null model and calculating the score test statistics, adjusting for the new peak variant along with the original trait-specific conditional variants and the first peak variant. We continued this procedure iteratively, adding any new 'peak variants' into the list of variants to condition on, re-fitting the null model, and calculating the updated score statistics, until no additional variants were significant at the $5x10^{-9}$ level. Finally, the variants identified across all chromosomes in this iterative conditional analysis were combined into a set of "conditionally-independent variants" for each trait.

Aggregation Strategies

For aggregate association testing, five distinct methods were used to aggregate rare variants into gene-based groups using GENCODE v29 gene model. Three strategies only included coding variants, and two strategies additionally included non-coding variants. Variants were further filtered using one or more deleterious prediction scores to enrich for likely causal variants. The detail method used for each strategy is provided below

1. **Coding filter 1 - Stringent (C1-S)**: This strategy includes high confidence predicted LoF variants inferred using LOFTEE (https://github.com/konradjk/loftee), missense variants predicted deleterious by all of SIFT4G[26633127]<=0.05, Polyphen2_HDIV>0.5[20354512], Polyphen2_HVAR>0.5[20354512], and variants predicted as "Deleterious" by LRT [19602639] and inframe indels or synonymous variants with Fathmm-XF score[28968714] > 0.5
2. **Coding filter 1 - Relaxed (C1-R)**: This strategy is same as C1-S but the missense filter was relaxed to retain variants predicted deleterious by any of SIFT4G, Polyphen2_HDIV, Polyphen2_HVAR scores

3. **Coding filter 2 - Relaxed (C2-R)**: This strategy is the same as C1-S but missense variants were filtered using MetaSVM score and a relatively relaxed set of missense variants was retained by applying MetaSVM score [25552646] > 0 filter
4. **Coding filter 2 - Relaxed & Non-coding filter-Relaxed (C2-R+NC-R)**: This strategy includes variants included in C2-R and additional regulatory variants. Regulatory variants were included if they overlapped with enhancer(s) or promoters linked to a gene using GeneHancer[28605766], or 5 Kb upstream of the Transcription start site. Within these regions only those variants were retained which had Fathmm-XF score > 0.5 **or** overlap with regions labelled as either"CTCF binding sites," "Transcription factor binding sites" as annotated by the Ensembl regulatory build annotation[25887522]
5. **Coding filter 2 - Relaxed & Non-coding filter-Stringent (C2-R +NC-S)** :This strategy includes variants included in C2-R and additional regulatory variants using a stringent filtering criteria. Even in this method regulatory variants in Genehancer[28605766] linked regulatory regions and 5 Kb upstream of the Transcription start site of gene were included. However within these regions only those variants were retained which had Fathmm-XF score > 0.5 **and** which overlapped with regions labelled as "Promoters," "Promoter flanking regions," "Enhancers," "CTCF binding sites," "Transcription factor binding sites" or "Open chromatin regions" as annotated by the Ensembl regulatory build annotation[25887522].

The annotation based variant filtering and gene based aggregation was performed using TOPMed freeze 8 WGSA Google BigQuery annotation database on the BiodataCatalyst powered by Seven Bridges platform (http://doi.org/10.5281/zenodo.3822858). The annotation database was built using variant annotations generated by Whole genome Sequence annotator version v0.8 [26395054] and formatted by WGSAParsr version 6.3.8 (https://github.com/UW-GAC/wgsaparsr). The GENCODE v29 gene model based varint consequences were obtained from Ensembl Variant effect predictor (VEP)[26683364] incorporated within WGSA. When using a deleteriousness prediction score, respective author recommended cut points were used to retain likely deleterious variants.

References

1. (1989). The atherosclerosis risk in communities (aric) study: design and objectives. Am. J. Epidemiol. *129*, 687–702.

2. Fried, L.P., Borhani, N.O., Enright, P., Furberg, C.D., Gardin, J.M., Kronmal, R.A., Kuller, L.H., Manolio, T.A., Mittelmark, M.B., and Newman, A. (1991). The Cardiovascular Health Study: design and rationale. Ann. Epidemiol. *1*, 263–276.

3. Dawber, T.R., and Kannel, W.B. (1966). The Framingham study. An epidemiological approach to coronary heart disease. Circulation *34*, 553–555.

4. Feinleib, M., Kannel, W.B., Garrison, R.J., McNamara, P.M., and Castelli, W.P. (1975). The Framingham Offspring Study. Design and preliminary data. Prev. Med. *4*, 518–525.

5. Splansky, G.L., Corey, D., Yang, Q., Atwood, L.D., Cupples, L.A., Benjamin, E.J., D'Agostino, R.B., Fox, C.S., Larson, M.G., Murabito, J.M., et al. (2007). The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. Am. J. Epidemiol. *165*, 1328–1335.

6. Lavange, L.M., Kalsbeek, W.D., Sorlie, P.D., Avilés-Santa, L.M., Kaplan, R.C., Barnhart, J., Liu, K., Giachello, A., Lee, D.J., Ryan, J., et al. (2010). Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. Ann. Epidemiol. *20*, 642–649.

7. Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R., Diez Roux, A.V., Folsom, A.R., Greenland, P., Jacob, D.R., Kronmal, R., Liu, K., et al. (2002). Multi-Ethnic Study of Atherosclerosis: objectives and design. Am. J. Epidemiol. *156*, 871–881.

8. (1998). Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. Control. Clin. Trials *19*, 61–109.

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

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

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

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

13. Frisch, R., & Waugh, F. V. (1933). Partial time regressions as compared with individual trends.

Econometrica: Journal of the Econometric Society, 387-401.
14. Zhou, B., Shi, J., and Whittemore, A.S. (2011). Optimal methods for meta-analysis of genome-wide association studies. Genet. Epidemiol. *35*, 581–591.

Acknowledgements

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

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

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

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

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

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

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

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

CHS: Cardiovascular Health Study: This research was supported by contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114

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

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

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

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

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

JHS: The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201300049C and HHSN268201300050C), Tougaloo College

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

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

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

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