Whole-Exome Sequencing Identifies Loci Associated with Blood Cell Traits and Reveals a Role for Alternative *GFI1B* Splice Variants in Human Hematopoiesis

Linda M. Polfus,[1,38] Rajiv K. Khajuria,[2,3,4,38] Ursula M. Schick,[5,38] Nathan Pankratz,[6] Raha Pazoki,[7] Jennifer A. Brody,[8] Ming-Huei Chen,[9] Paul L. Auer,[10] James S. Floyd,[8] Jie Huang,[11] Leslie Lange,[12] Frank J.A. van Rooij,[7] Richard A. Gibbs,[13] Ginger Metcalf,[13] Donna Muzny,[13] Narayanan Veeraraghavan,[13] Klaudia Walter,[11] Lu Chen,[11,14] Lisa Yanek,[15]

(Author list continued on next page)

Circulating blood cell counts and indices are important indicators of hematopoietic function and a number of clinical parameters, such as blood oxygen-carrying capacity, inflammation, and hemostasis. By performing whole-exome sequence association analyses of hematologic quantitative traits in 15,459 community-dwelling individuals, followed by in silico replication in up to 52,024 independent samples, we identified two previously undescribed coding variants associated with lower platelet count: a common missense variant in *CPS1* (rs1047891, $\mathrm{MAF} = 0.33$, discovery + replication $p = 6.38 \times 10^{-10}$) and a rare synonymous variant in *GFI1B* (rs150813342, $\mathrm{MAF} = 0.009$, discovery + replication $p = 1.79 \times 10^{-27}$). By performing CRISPR/Cas9 genome editing in hematopoietic cell lines and follow-up targeted knockdown experiments in primary human hematopoietic stem and progenitor cells, we demonstrate an alternative splicing mechanism by which the *GFI1B* rs150813342 variant suppresses formation of a GFI1B isoform that preferentially promotes megakaryocyte differentiation and platelet production. These results demonstrate how unbiased studies of natural variation in blood cell traits can provide insight into the regulation of human hematopoiesis.

Human genetic studies have provided important insights into hematopoiesis. Genome-wide association studies (GWASs) performed in large, population-based samples have identified associations of genomic regions and common genetic (usually non-coding) variants with inter-individual differences in blood cell traits[1–5], though the causal DNA variants and their functional mechanisms often remain elusive. Whole-exome and targeted sequencing approaches have been used to identify rare, sometimes private, loss (or gain)-of-function coding variants segregating within families with hematologic traits at the extremes of the phenotypic distribution[6–12]. As of yet, whole-exome sequencing has not been applied to large population-based cohorts well-phenotyped for hematologic traits to identify rare, functional variation with moderate-to-large phenotypic effects and to provide new biologic insight.

To this end, we performed exome sequencing in 15,459 unrelated European ancestry (EU) and African American (AA) individuals enrolled in six population-based cohort studies (see Supplemental Note). Replication of significant findings was performed in up to 52,024 additional samples via a combination of whole-exome-based or genome-based sequencing, genotyping, and imputation (Supplemental Note). Our a priori hypothesis was that systematic evaluation of coding variation detected by exome sequence analysis in samples unselected for blood cell traits would identify low-frequency variants influencing hematologic traits and could provide functional insights into hematopoiesis. We analyzed platelet count and 12 other blood cell traits (Table S1). The means of the traits were as expected in a sample of unselected healthy individuals from the population (Table S1). Association results from single-variant and from gene-based burden and sequence kernel association tests (SKATs) meeting our a priori significance thresholds in either EU, AA, or trans-ethnic discovery meta-analyses are summarized for both previously known and novel (which we define as those not reported in the available literature) loci in Table 1 and Tables S2–S5 and described further in the Supplemental Note. Lambda values showed no significant inflation (Table S6).

[1]Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [2]Division of Hematology/Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA,; [3]Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [4]Berlin-Brandenburg School for Regenerative Therapies, Charité Universitätsmedizin Berlin, Berlin 13353, Germany; [5]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [6]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN 55454, USA; [7]Department of Epidemiology, Erasmus University Medical Center, Rotterdam 3000, the Netherlands; [8]Cardiovascular Health Research Unit and Department of Medicine, University of Washington, Seattle, WA 98195, USA; [9]Department of Neurology, School of Medicine, Boston University, Boston, MA 02118, USA; [10]School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI 53205, USA; [11]Human Genetics, Wellcome Trust Sanger Institute, Hinxton CB10 1HH, UK; [12]Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA; [13]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA; [14]Department of Haematology, University of Cambridge, Cambridge CB2 0AH, UK; [15]GeneSTAR Research Program, Division of General Internal Medicine, Department of Medicine, School of Medicine, Johns Hopkins University, Baltimore, MD 21205, USA; [16]Center for Human Genetic Research,

(Affiliations continued on next page)

Lewis C. Becker,[15] Gina M. Peloso,[16] Aoi Wakabayashi,[2,3] Mart Kals,[17] Andres Metspalu,[17] Tõnu Esko,[17] Keolu Fox,[18] Robert Wallace,[19] Nora Franceshini,[20] Nena Matijevic,[21] Kenneth M. Rice,[8] Traci M. Bartz,[8] Leo-Pekka Lyytikäinen,[22] Mika Kähönen,[23] Terho Lehtimäki,[22] Olli T. Raitakari,[24] Ruifang Li-Gao,[25] Dennis O. Mook-Kanamori,[25,26] Guillaume Lettre,[27] Cornelia M. van Duijn,[28] Oscar H. Franco,[7] Stephen S. Rich,[29] Fernando Rivadeneira,[28] Albert Hofman,[28] André G. Uitterlinden,[28] James G. Wilson,[30] Bruce M. Psaty,[8,31] Nicole Soranzo,[11,14] Abbas Dehghan,[7] Eric Boerwinkle,[1] Xiaoling Zhang,[32] Andrew D. Johnson,[33] Christopher J. O'Donnell,[34] Jill M. Johnsen,[35] Alexander P. Reiner,[36,39] Santhi K. Ganesh,[37,39] and Vijay G. Sankaran[2,3,39,*]

Four gene-based associations were discovered for red blood cell (RBC) traits (*ACTN4*, *MMACHC*, *MYOM2*, and *MRPL43*). Trans-ethnic discovery meta-analyses are summarized for both previously identified loci, which we verify in this study, and previously unreported loci. A summary of these findings, and driving variants, are provided in the Supplemental Note and Table S3. None of these gene-based SKAT or burden findings could be replicated in independent samples. Nonetheless, a few of the individual rare variants driving the gene-based associations in the discovery sample showed suggestive evidence of association in the replication sample (Supplemental Note and Table S3**)**.

Among the three single-variant associations we identified (Table 1), two coding variants were associated with lower platelet count in our discovery sample: *CPS1* rs1047891, a common missense variant encoding p.Thr1412Asn (EU + AA minor-allele frequency [MAF] = 0.33, EU + AA p = 5.7×10^{-8}) and *GFI1B* rs150813342, a rare synonymous variant encoding p.Phe192 and located in alternatively spliced exon 5 (EU MAF = 0.009, EU p= 4.7×10^{-8}; EU + AA MAF = 0.008, EU + AA p = 2.64×10^{-8}). One single-nucleotide variant (SNV) result (rs9656446; EU + AA MAF = 0.03, EU + AA p = 1.48×10^{-7}) associated with basophils in trans-ethnic analyses was in the ATP/GTP binding protein-like 3 (*AGBL3*) gene. However, the allele frequencies in the discovery sample differed by ethnicity (EU MAF = 0.001 and AA MAF = 0.08), and replication in samples of EU ethnicity from the UK10K project was not significant (EU p = 0.71). In our combined replication sample, we replicated the associations of *CPS1* rs1047891 (EU + AA MAF = 0.328, EU + AA p = 1.02×10^{-4}) and *GFI1B* rs150813342 (EU + AA p = 5.71×10^{-21}) with lower platelet counts. In the combined discovery and replication samples, the p values for *CPS1* rs1047891 and *GFI1B* rs150813342 were 6.38×10^{-10} and 1.79×10^{-27}, respectively. A Manhattan plot for single-variant associations with platelet count and quantile-quantile (Q-Q) plots are shown in Figures S1–S3. Forest plots of the discovery cohorts for the two replicated findings (*GFI1B* rs150813342 and *CPS1* rs1047891) are provided in Figures S4 and S5, as well as regional plots calculating linkage disequilibrium of SNVs in the gene with respect to index SNVs (Figures S6 and S7).

AGBL3 is a metallocarboxypeptidase involved in processing tubulins of the blood cell cytoskeleton. *CPS1* encodes carbamoyl-phosphate synthase 1, a mitochondrial enzyme involved in the urea cycle. The *CPS1* variant (or its LD proxies) has been associated with various cardiometabolic traits, including high-density lipoprotein (HDL) cholesterol, homocysteine, fibrinogen, serum metabolite levels, and kidney function.[13–17] *GFI1B* is a known transcriptional repressor and a key regulator of platelet and red blood cell development. There was no evidence that either *CPS1* rs1047891 or *GFI1B* rs150813342 were significantly associated with other hematologic traits assessed in the discovery sample (Tables S7A and S7B). Moreover, neither *GFI1B* rs150813342 nor *CPS1* rs1047891 was associated with mean platelet volume, platelet aggregation, or expression of platelet surface markers, though these analyses were limited to much smaller numbers of individuals (Supplemental Note, Tables S8 and S10). However, a decrease in

Massachusetts General Hospital, Boston, MA 02114, USA; [17]Estonian Genome Center, University of Tartu, Tartu, 51010, Estonia; [18]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA; [19]College of Public Health, the University of Iowa, Iowa City, IA 52242, USA; [20]Department of Medicine, School of Medicine, University of North Carolina, Chapel Hill, NC 27599, USA; [21]Department of Surgery, University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [22]Department of Clinical Chemistry, Fimlab Laboratories and University of Tampere School of Medicine, Tampere 33520, Finland; [23]Department of Clinical Physiology, Tampere University Hospital and University of Tampere School of Medicine, Tampere 33521, Finland; [24]Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital and Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku 20520, Finland; [25]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden RC 2300, the Netherlands; [26]Epidemiology Section, Department of Biostatistics, Epidemiology, and Scientific Computing Department, King Faisal Specialist Hospital and Research Centre, Riyadh 11211 Saudi Arabia; [27]Montreal Heart Institute and Université de Montréal, Montreal, QC H1T 1C8, Canada; [28]Department of Internal Medicine, Erasmus University Medical Center, Rotterdam 3000, the Netherlands; [29]Center for Public Health Genomics, University of Virginia, Charlottesville VA 22908, USA; [30]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson MS 39216, USA; [31]Group Health Research Institute, Group Health Cooperative, Seattle, WA 98101, USA; [32]Departments of Medicine and Biostatistics, Schools of Medicine and Public Health, Boston University, Boston, MA 02118, USA; [33]Cardiovascular Epidemiology and Human Genomics Branch, Framingham Heart Study, National Heart, Lung, and Blood Institute, Framingham, MA 01702, USA; [34]Framingham Heart Study, National Heart, Lung, and Blood Institute, Framingham, MA 01702, USA; [35]Bloodworks Northwest, Seattle, WA 98102, USA; [36]Women's Health Initiative Clinical Coordinating Center, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; [37]Division of Cardiovascular Medicine, Departments of Internal Medicine and Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA

[38]These authors contributed equally to this work

[39]These authors contributed equally to this work

*Correspondence: sankaran@broadinstitute.org (V.G.S.)

http://dx.doi.org/10.1016/j.ajhg.2016.06.016.

the median fluorescence intensity of large, platelet-marker positive ($CD41^+CD61^+$) events[18] was detected by flow cytometry in *GFI1B* variant carriers even after adjustment for circulating platelet count ($p < 0.0001$), which could reflect a decrease in circulating platelet aggregates or a skewing of a platelet subpopulation with regards to platelet-surface-marker expression or size (see Supplemental Note).

We conducted bioinformatic and functional analyses to understand the impact of the *GFI1B* exon 5 synonymous variant and the *CPS1* rs1047891 variant (p.Thr1412Asn) on gene and protein function. The *CPS1* p.Thr1412Asn amino acid substitution is predicted to be benign and tolerated by SIFT and PolyPhen. Moreover, according to the GTEx Portal database, there is no evidence of an expression quantitative trait loci (eQTL) effect for rs1047891. Nonetheless, the CPS1 p.Thr1412Asn missense substitution is located within a region critical for N-acetyl-glutamate binding and has been reported to result in 20%–30% higher enzymatic activity[19] and to influence vascular function.[15]

We initially assessed the association of rs150813342 with *GFI1B* expression by using Affymetrix GeneChip Human Exon 1.0 ST Array data on whole-blood RNA available from 881 Framingham Heart Study participants.[20] There was no evidence for association of the rs150813342 genotype with expression of any *GFI1B* exon, though statistical power is likely limited by the low frequency of the rs150813342 variant allele, which was present in only 7 of the 881 individuals. According to SPANR,[21] rs150813342 had a predicted effect on splicing (difference in the percentage of transcripts with the exon spliced in [dPSI] score of −4.6). rs150813342 was predicted to disrupt a putative exon splicing enhancer (ESE) in exon 5 that contains a consensus SRSF1 binding motif.[22] To functionally evaluate the impact of this variant on GFI1B transcript splicing in a relevant cell type, we used CRISPR/Cas9 genome editing to create multiple independent isogenic K562 hematopoietic cell lines harboring the *GFI1B* synonymous single-nucleotide change (Figure 1A). These cell lines were homozygous for the variant and exhibited inclusion of less than 30% of exon 5 relative to other surrounding exons in the *GFI1B* mRNA (Figure 1B). Semi-quantitative RT-PCR showed that the presence of the synonymous variant resulted in reduced formation of the *GFI1B* isoform containing exon 5 (herein referred to as the long isoform), as well as preferential formation of the isoform lacking exon 5 (herein referred to as the short isoform) (Figures 1C and 1D). No other isoforms or intron inclusion events were detected (Figure 1C, Figure S8).

Although *GFI1B* has been implicated in both RBC and platelet production (erythropoiesis and megakaryopoiesis, respectively),[23–25] only a role for the short isoform in erythroid cells has been suggested previously.[26] We next assessed the effect of the altered splicing of *GFI1B* on lineage-specific hematopoietic differentiation. We chemically induced differentiation of the isogenic K562 cell lines with either hemin (to promote erythroid differentiation) or phorbol 12-myristate 13-acetate (PMA, to promote megakaryocytic differentiation) (Figure 2A). Although erythroid

Table 1. Single-Variant Association Findings

Trait	Discovery Ethnicity	Gene	SNP Chromosome Position, rs Number, and Function	Discovery p Value	Replication p Value	Discovery MAF	Replication MAF	Discovery Beta Coefficient (SE)	Replication Z Score[a]	Discovery N	Replication N
PLT	EU + AA	*GFI1B*	chr9: 135864513, rs150813342, synonymous	2.64×10^{-8}	5.71×10^{-21}	0.008	0.007	−0.402 (0.07)	−9.40	13,744	48,099[b]
PLT	EU + AA	*CPS1*	chr2: 211540507, rs1047891, missense	5.73×10^{-8}	1.02×10^{-4}	0.328	0.313	−0.07 (0.013)	−3.89	13,744	48,394[b]
BASO	EU + AA	*AGBL3*	chr7: 134717656, rs9656446, synonymous	1.48×10^{-7}	0.71	0.031[c]	0.002	0.271 (0.051)	−0.05 (0.13)	6,877	6,699[d]

AA, African American individuals; BASO, basophil count; EU, European ancestry individuals; MAF, minor-allele frequency; PLT, platelet count.
[a] Z score is reported from N-weighted replication meta-analyses, where more than one replication cohort was available; otherwise, beta coefficient and SE are reported.
[b] UK10K project samples and imputed EU, Cardiovascular Health Study (CHS), and Atherosclerosis Risk in Communities (ARIC) study samples.
[c] EU MAF = 0.001; AA MAF = 0.078; EU + AA MAF = 0.031.
[d] UK10K project samples and imputed EU samples.

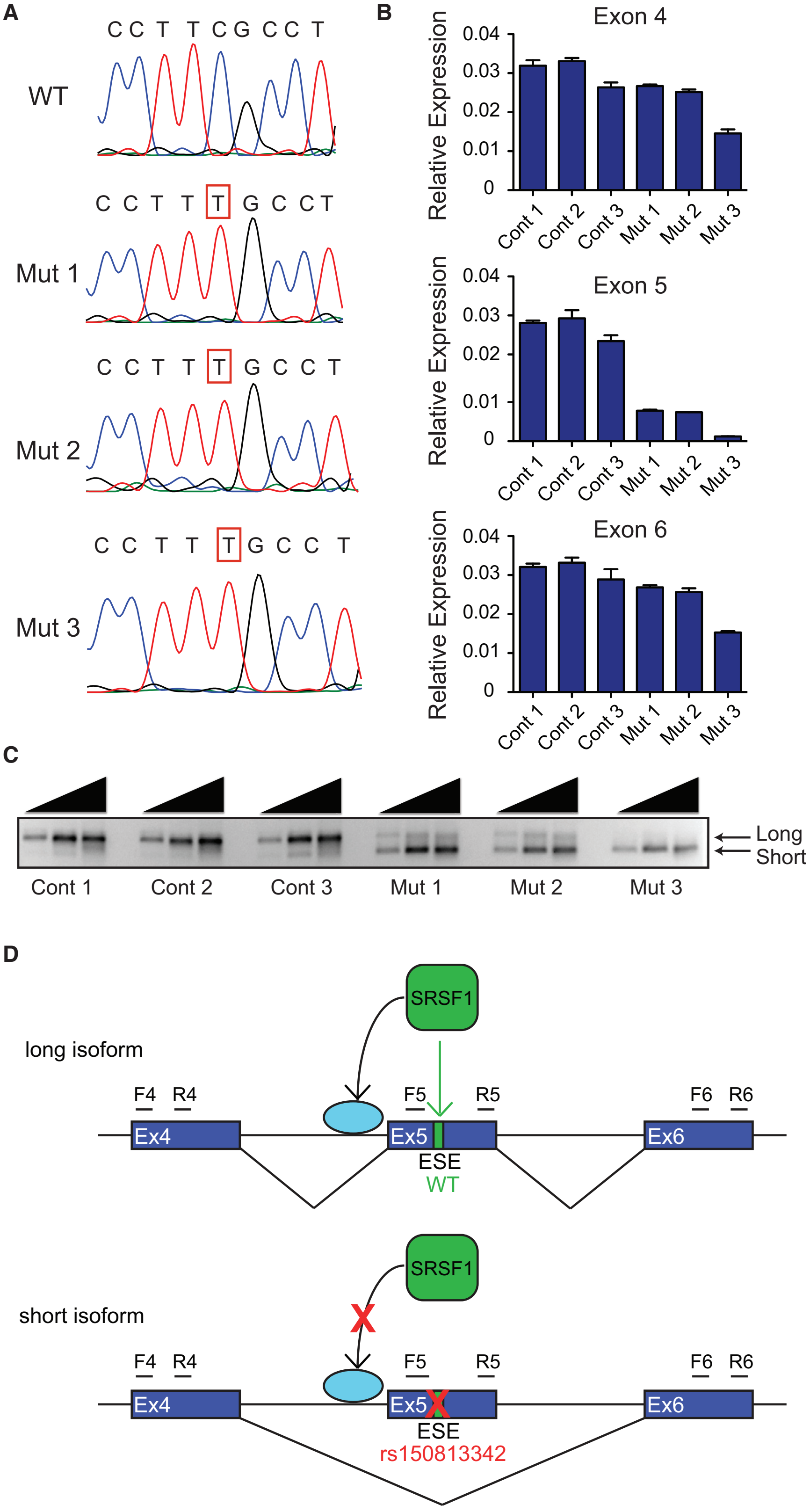


Figure 1. The Variant rs150813342 Results in Reduced Formation of the Long *GFI1B* Isoform and Preferential Formation of the Short Isoform

(A) Chromatograms of the sequence surrounding the altered nucleotide in *GFI1B* exon 5 showing the wild-type (WT) sequence and sequences of isogenic hematopoietic K562 cell mutant clones (Mut 1, Mut 2, and Mut 3) harboring the C>T single-nucleotide variant (SNV) generated via CRISPR/Cas9 mediated homologous repair.

(B) qRT-PCR of *GFI1B* exons 4, 5, and 6 measured from isogenic control (Cont) and mutant K562 cell clones showing inclusion of less than 30% of *GFI1B* exon 5 relative to the surrounding exons in *GFI1B* mRNA from mutant clones (n = 3 per group). Error bars show SD.

(C) Semi-quantitative RT-PCR with *GFI1B* exon 4 forward and exon 6 reverse primers with progressively increasing cycle numbers (26, 28, and 30 cycles) demonstrates reduced formation of the long *GFI1B* isoform and preferential formation of the short isoform, as well as no other intermediate isoforms in the clones harboring the SNV.

(D) rs150813342 is predicted to disrupt a putative exon splicing enhancer (ESE) in exon 5 that contains a consensus SRSF1 binding motif. Disruption of this binding motif results in reduced inclusion of exon 5 and preferential formation of the short isoform. The promotion of alternative splicing by SRSF1 through the spliceosome complex is indicated by an arrow to a light blue circle. Forward (F) and reverse (R) PCR primers of the respective exon are indicated.

differentiation appeared to proceed normally, as assessed morphologically (Figure 2B), and with expression of the surface marker GYPA (CD235a) (Figure 2C) and terminal erythroid marker genes (Figure 2D), megakaryocyte differentiation appeared severely impaired; the cells retained an immature blast-like morphology and failed to upregulate the surface marker of megakaryocyte differentiation, CD41a (encoded by *ITGA2B*), and mRNAs whose expression is characteristic of terminal megakaryopoiesis (Figures 2B–2D, Figure S9). The megakaryocyte genes *PPBP*, *SELP*, and *PF4* were downregulated by an average of 8.6-, 6.7-, and 41.1-fold, respectively, in the isogenic clones (p = 0.0001, 0.0013, and 0.0459, respectively) versus in the controls (Figure 2D). These results suggest that the long isoform of GFI1B is necessary for normal megakaryocyte differentiation.

To confirm a preferential role for this long GFI1B isoform in megakaryocyte differentiation, we identified two independent short hairpin RNAs (shRNAs) that specifically targeted *GFI1B* exon 5, which would thereby selectively downregulate the long but not the short isoform. We utilized

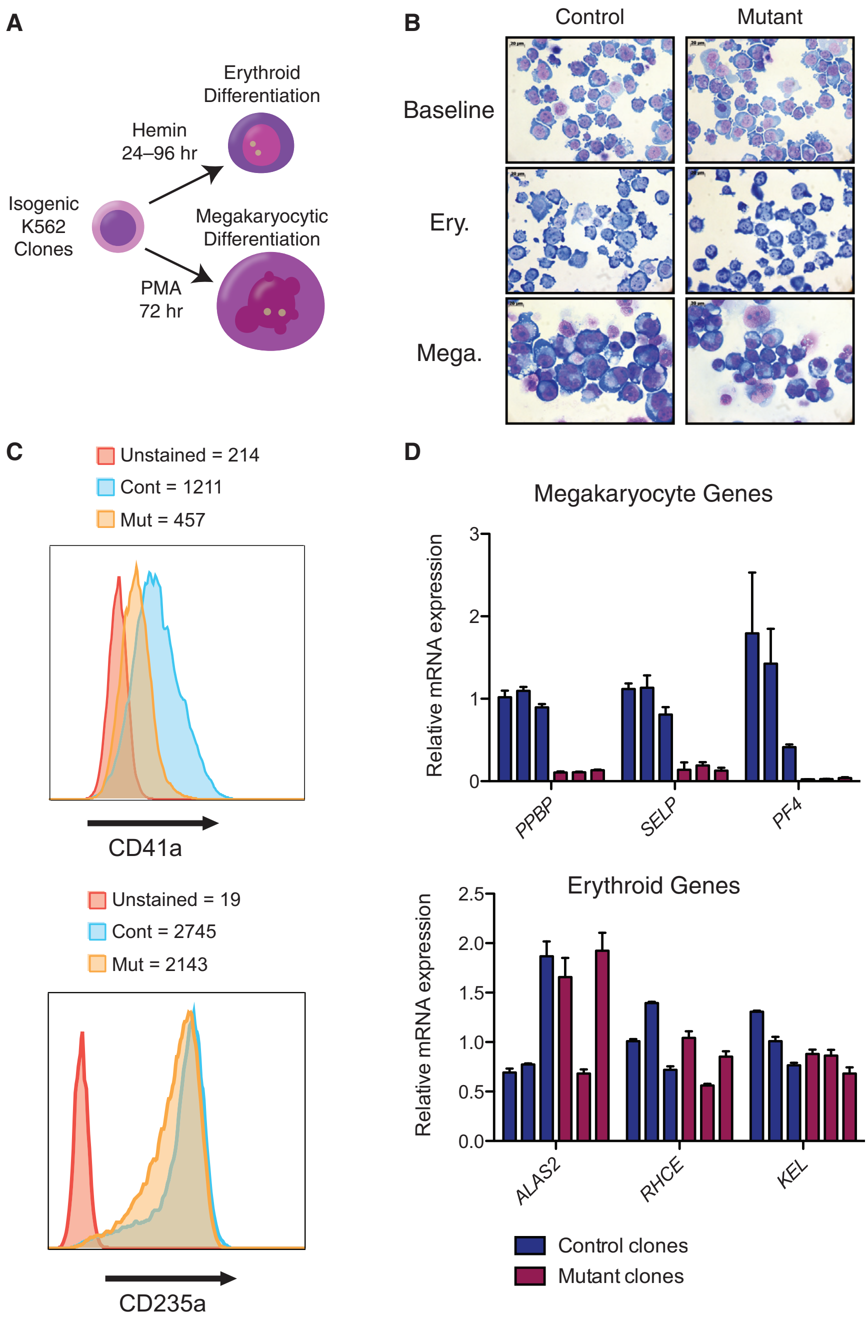


Figure 2. Impaired Megakaryopoiesis and Retained Erythropoiesis in K562 Cells Harboring the rs150813342 SNV in *GFI1B* Exon 5

(A) Scheme of phorbol 12-myristate 13-acetate (PMA)-induced megakaryocytic differentiation and hemin-induced erythroid differentiation of the hematopoietic K562 cell models.

(B) Representative May-Grünwald-Giemsa-stained cytospin images of 72 hr PMA-induced and 96 hr hemin-induced isogenic control and mutant clones showing megakaryocytic differentiation that appears severely impaired, with the cells retaining an immature blast-like morphology in the mutant clones, whereas the erythroid differentiation appears unaffected.

(C) Representative flow cytometry analysis of the megakaryocyte marker CD41a and the erythroid marker CD235a further confirmed the impaired megakaryopoiesis and the retained erythropoiesis as shown by the histogram plots with the mean fluorescence intensity (MFI) for each marker in unstained cells, control, and mutant clones, respectively.

(D) Gene expression analysis by qRT-PCR of the megakaryocyte markers *PPBP*, *SELP*, and *PF4* after 72 hr of PMA-induced differentiation and of the erythroid markers *ALAS2*, *RHCE*, and *KEL* after 24 hr of hemin-induced differentiation (n = 3 per group). Error bars show SD.

lentiviral-mediated shRNA delivery in primary human adult mobilized peripheral-blood hematopoietic stem and progenitor cells (HSPCs), which are capable of differentiation toward the erythroid and megakaryocyte lineages under appropriate culture conditions.[27] We observed a knockdown efficiency of the *GFI1B* long isoform by ~50% for both shRNAs, whereas the short isoform levels increased conversely (Figures 3A and 3B), which resulted in a 1.5- to 1.8-fold reduction in the formation of $CD41a^+$ megakaryocytic cells (relative to lineage-marker negative cells) in HSPCs undergoing differentiation (Figure 3C). In contrast, $CD235a^+$ erythroid cells appeared to be present in comparable percentages and numbers (Figure 3C). Moreover, whereas numerous morphologically mature erythroblasts could be readily visualized in both groups, fewer mature megakaryocytic cells were seen with knockdown of the long isoform than in the controls (Figure 3D, Figure S10). Overall cell growth appeared comparable between the knockdown and control cells (Figure S10). These findings are in line with our exome-sequence association findings, in which no significant effect was seen on circulating RBC levels.

GFI1B private, loss-of-function mutations (nonsense, frameshift) in the DNA-binding fifth and sixth zinc (Zn)-finger domains have recently been identified in families with an autosomal-dominant form of Gray Platelet syndrome (GPS) or related forms of thrombocytopenia, which are characterized by dysmegakaryopoiesis, thrombocytopenia, large platelets, and platelet α-granule deficiency (MIM: 187900)[28,29]. The truncating *GFI1B* mutations reported in GPS appear to have a dominant-negative effect and inhibit transcriptional activity of the *GFI1B* wild-type form. Our population study extends the allelic spectrum of naturally occurring *GFI1B* coding sequence variants associated with a lower circulating platelet count to include a more frequent, synonymous change that alters

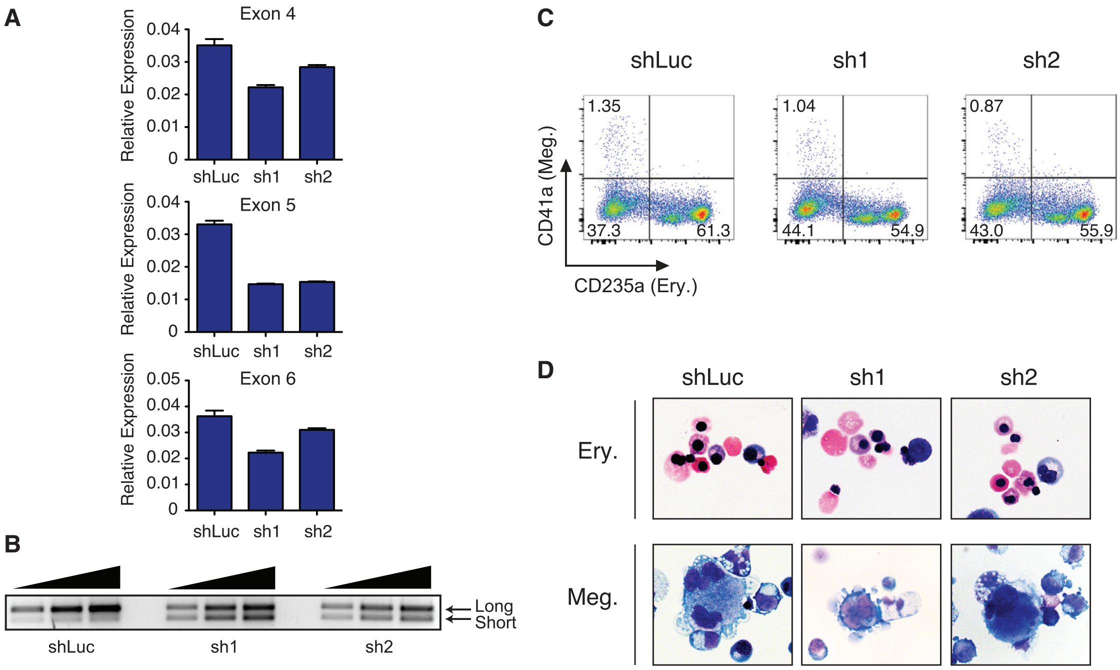


Figure 3. The Long *GFI1B* Isoform is Critical for Megakaryopoiesis in a Human Primary Cell Model

(A) qRT-PCR of *GFI1B* exons 4, 5, and 6 on day 4 after infection showing the identification of two short hairpin RNAs (shRNAs) that specifically target *GFI1B* exon 5 and thereby selectively downregulate the long isoform by ~50%, but not the short isoform (n = 3 per group). Error bars show SD.

(B) Semi-quantitative RT-PCR with *GFI1B* exon 4 forward and exon 6 reverse primers with progressively increasing cycle numbers (26, 28, and 30 cycles) demonstrates reduced formation of the long *GFI1B* isoform and increased formation of the short isoform, as well as no other intermediate isoforms in cells with targeted knockdown of GFI1B exon 5.

(C) Representative flow cytometry analysis of thrombopoietin (TPO)- and erythropoietin (EPO)-stimulated primary human hematopoietic stem and progenitor cells on day 11 of differentiation with assessment of CD41a$^+$ megakaryocytic (Meg) cells and CD235a$^+$ erythroid (Ery) cells.

(D) Representative May-Grünwald-Giemsa-stained cytospin images of megakaryocytic cells (from day 7 of differentiation) and erythroid cells (from day 13 of differentiation) showing immature megakaryocyte morphology in cells with knockdown of the long *GFI1B* isoform, in comparison with the control. In contrast, maturation of erythroblasts appears unaffected.

an exonic splicing enhancer, resulting in the skipping of exon 5, containing the first and second Zn-finger domains. Heterozygous carriers of the synonymous exon 5 variant in *GFI1B* have an average platelet count that is reduced by 25,000 to 30,000 platelets per microliter, which would be a clinically detectable effect. We also provide additional support for distinct roles of GFI1B long- and short-isoforms, which are differentially expressed at various stages of differentiation during normal hematopoiesis.[23,30] The long GFI1B isoform is expressed in HSPCs and lineage-committed myeloid, erythroid, and megakaryocytic progenitors. The abnormalities in megakaryocyte maturation with reduced formation of the *GFI1B* long isoform in the isogenic K562 cell clones containing the rs150813342 variant and in primary HSPCs with targeted suppression of the long isoform are consistent with an essential role for the GFI1B long isoform in megakaryopoiesis and platelet production. This finding is also congruent with prior work showing that the GFI1B short isoform is required for erythropoiesis[26] and provides insight into how these different splice variants function in distinct aspects of human hematopoiesis.

In summary, whole-exome sequence association analysis performed in over 15,000 samples discovered SNVs associated with a lower platelet count in community-dwelling individuals, including a common variant in *CPS1* and a rare, synonymous variant in *GFI1B*. Follow-up genome editing and targeted knockdown experiments identified a mechanism by which alternative splicing associated with the *GFI1B* rs150813342 variant allele suppresses formation of a specific GFI1B long isoform that is required for lineage-specific megakaryocyte differentiation, while being dispensable for erythropoiesis. Functional studies coupled with an association finding demonstrated a previously unappreciated splicing-based mechanism for lineage-specific blood cell production, providing important insights into human hematopoiesis. Genes regulated by the long GFI1B isoform could provide additional understanding of downstream transcriptional events and molecular pathways required for megakaryocyte specification and platelet production. These findings hold promise for the development of therapeutics for altering platelet count without adverse effects on other blood lineages. Further characterization of the role of GFI1B isoforms could have clinical or therapeutic implications for disorders of platelet and other blood cell production or function, as well as for the prospect of improving the manufacture of ex vivo cell therapies.[31–33]

Supplemental Data

Supplemental Data include a Supplemental Note, ten figures, nine tables, and Supplemental Acknowledgments and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.06.016.

Acknowledgments

The work described in this manuscript was supported in part by NIH grants HL122684 to S.K.G. and DK103794 and HL120791 to V.G.S. N.S. is supported by the Wellcome Trust (grant codes WT098051 and WT091310), the EU 7th Framework Programme (EPIGENESYS grant code 257082 and BLUEPRINT grant code HEALTH-F5-2011-282510) and the NIH Research (NIHR) Blood and Transplant Research Unit (BTRU) in Donor Health and Genomics at the University of Cambridge in partnership with NHS Blood and Transplant (NHSBT).

Received: March 14, 2016
Accepted: June 20, 2016
Published: August 4, 2016

Web Resources:

OMIM, http://www.omim.org/

References

1. Okada, Y., and Kamatani, Y. (2012). Common genetic factors for hematological traits in humans. J. Hum. Genet. *57*, 161–169.
2. Ganesh, S.K., Zakai, N.A., van Rooij, F.J., 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.
3. 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.
4. Auer, P.L., Teumer, A., Schick, U., O'Shaughnessy, A., Lo, K.S., Chami, N., Carlson, C., de Denus, S., Dubé, M.P., Haessler, J., et al. (2014). Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. Nat. Genet. *46*, 629–634.
5. Ulirsch, J.C., Nandakumar, S.K., Wang, L., Giani, F.C., Zhang, X., Rogov, P., Melnikov, A., McDonel, P., Do, R., Mikkelsen, T.S., and Sankaran, V.G. (2016). Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. Cell *165*, 1530–1545.
6. Shiohara, M., Shigemura, T., Saito, S., Tanaka, M., Yanagisawa, R., Sakashita, K., Asada, H., Ishii, E., Koike, K., Chin, M., et al. (2009). Ela2 mutations and clinical manifestations in familial congenital neutropenia. J. Pediatr. Hematol. Oncol. *31*, 319–324.
7. Minelli, A., Maserati, E., Rossi, G., Bernardo, M.E., De Stefano, P., Cecchini, M.P., Valli, R., Albano, V., Pierani, P., Leszl, A., et al. (2004). Familial platelet disorder with propensity to acute myelogenous leukemia: genetic heterogeneity and progression to leukemia via acquisition of clonal chromosome anomalies. Genes Chromosomes Cancer *40*, 165–171.
8. Sankaran, V.G., and Gallagher, P.G. (2013). Applications of high-throughput DNA sequencing to benign hematology. Blood *122*, 3575–3582.
9. Albers, C.A., Cvejic, A., Favier, R., Bouwmans, E.E., Alessi, M.C., Bertone, P., Jordan, G., Kettleborough, R.N., Kiddle, G., Kostadima, M., et al. (2011). Exome sequencing identifies NBEAL2 as the causative gene for gray platelet syndrome. Nat. Genet. *43*, 735–737.
10. Johnsen, J.M., Nickerson, D.A., and Reiner, A.P. (2013). Massively parallel sequencing: the new frontier of hematologic genomics. Blood *122*, 3268–3275.
11. Giani, F.C., Fiorini, C., Wakabayashi, A., Ludwig, L.S., Salem, R.M., Jobaliya, C.D., Regan, S.N., Ulirsch, J.C., Liang, G., Steinberg-Shemer, O., et al. (2016). Targeted Application of Human Genetic Variation Can Improve Red Blood Cell Production from Stem Cells. Cell Stem Cell *18*, 73–78.
12. Sankaran, V.G., Ghazvinian, R., Do, R., Thiru, P., Vergilio, J.A., Beggs, A.H., Sieff, C.A., Orkin, S.H., Nathan, D.G., Lander, E.S., and Gazda, H.T. (2012). Exome sequencing identifies GATA1 mutations resulting in Diamond-Blackfan anemia. J. Clin. Invest. *122*, 2439–2443.
13. Köttgen, A., Pattaro, C., Böger, C.A., Fuchsberger, C., Olden, M., Glazer, N.L., Parsa, A., Gao, X., Yang, Q., Smith, A.V., et al. (2010). New loci associated with kidney function and chronic kidney disease. Nat. Genet. *42*, 376–384.
14. Paré, G., Chasman, D.I., Parker, A.N., Zee, R.R., Mälarstig, A., Seedorf, U., Collins, R., Watkins, H., Hamsten, A., Miletich, J.P., and Ridker, P.M. (2009). Novel associations of CPS1, MUT, NOX4, and DPEP1 with plasma homocysteine in a healthy population: a genome-wide evaluation of 13 974 participants in the Women's Genome Health Study. Circ Cardiovasc Genet *2*, 142–150.
15. Summar, M.L., Gainer, J.V., Pretorius, M., Malave, H., Harris, S., Hall, L.D., Weisberg, A., Vaughan, D.E., Christman, B.W., and Brown, N.J. (2004). Relationship between carbamoyl-phosphate synthetase genotype and systemic vascular function. Hypertension *43*, 186–191.
16. Pearson, D.L., Dawling, S., Walsh, W.F., Haines, J.L., Christman, B.W., Bazyk, A., Scott, N., and Summar, M.L. (2001). Neonatal pulmonary hypertension–urea-cycle intermediates, nitric oxide production, and carbamoyl-phosphate synthetase function. N. Engl. J. Med. *344*, 1832–1838.
17. Sabater-Lleal, M., Huang, J., Chasman, D., Naitza, S., Dehghan, A., Johnson, A.D., Teumer, A., Reiner, A.P., Folkersen, L., Basu, S., et al.; VTE Consortium; STROKE Consortium; Wellcome Trust Case Control Consortium 2 (WTCCC2); C4D Consortium; CARDIoGRAM Consortium (2013). Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated Loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease. Circulation *128*, 1310–1324.
18. Catellier, D.J., Aleksic, N., Folsom, A.R., and Boerwinkle, E. (2008). Atherosclerosis Risk in Communities (ARIC) Carotid MRI flow cytometry study of monocyte and platelet markers: intraindividual variability and reliability. Clin. Chem. *54*, 1363–1371.
19. Summar, M.L., Hall, L., Christman, B., Barr, F., Smith, H., Kallianpur, A., Brown, N., Yadav, M., Willis, A., Eeds, A., et al.

(2004). Environmentally determined genetic expression: clinical correlates with molecular variants of carbamyl phosphate synthetase I. Mol. Genet. Metab. *81* (*Suppl 1*), S12–S19.

20. Zhang, X., Joehanes, R., Chen, B.H., Huan, T., Ying, S., Munson, P.J., Johnson, A.D., Levy, D., and O'Donnell, C.J. (2015). Identification of common genetic variants controlling transcript isoform variation in human whole blood. Nat. Genet. *47*, 345–352.
21. Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.
22. Cho, S., Hoang, A., Sinha, R., Zhong, X.Y., Fu, X.D., Krainer, A.R., and Ghosh, G. (2011). Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and U1-70K snRNP protein determines early spliceosome assembly. Proc. Natl. Acad. Sci. USA *108*, 8233–8238.
23. Foudi, A., Kramer, D.J., Qin, J., Ye, D., Behlich, A.S., Mordecai, S., Preffer, F.I., Amzallag, A., Ramaswamy, S., Hochedlinger, K., et al. (2014). Distinct, strict requirements for Gfi-1b in adult bone marrow red cell and platelet generation. J. Exp. Med. *211*, 909–927.
24. Randrianarison-Huetz, V., Laurent, B., Bardet, V., Blobe, G.C., Huetz, F., and Duménil, D. (2010). Gfi-1B controls human erythroid and megakaryocytic differentiation by regulating TGF-beta signaling at the bipotent erythro-megakaryocytic progenitor stage. Blood *115*, 2784–2795.
25. Saleque, S., Cameron, S., and Orkin, S.H. (2002). The zinc-finger proto-oncogene Gfi-1b is essential for development of the erythroid and megakaryocytic lineages. Genes Dev. *16*, 301–306.
26. Laurent, B., Randrianarison-Huetz, V., Frisan, E., Andrieu-Soler, C., Soler, E., Fontenay, M., Dusanter-Fourt, I., and Duménil, D. (2012). A short Gfi-1B isoform controls erythroid differentiation by recruiting the LSD1-CoREST complex through the dimethylation of its SNAG domain. J. Cell Sci. *125*, 993–1002.
27. Ludwig, L.S., Gazda, H.T., Eng, J.C., Eichhorn, S.W., Thiru, P., Ghazvinian, R., George, T.I., Gotlib, J.R., Beggs, A.H., Sieff, C.A., et al. (2014). Altered translation of GATA1 in Diamond-Blackfan anemia. Nat. Med. *20*, 748–753.
28. Stevenson, W.S., Morel-Kopp, M.C., Chen, Q., Liang, H.P., Bromhead, C.J., Wright, S., Turakulov, R., Ng, A.P., Roberts, A.W., Bahlo, M., and Ward, C.M. (2013). GFI1B mutation causes a bleeding disorder with abnormal platelet function. J. Thromb. Haemost. *11*, 2039–2047.
29. Monteferrario, D., Bolar, N.A., Marneth, A.E., Hebeda, K.M., Bergevoet, S.M., Veenstra, H., Laros-van Gorkom, B.A., MacKenzie, M.A., Khandanpour, C., Botezatu, L., et al. (2014). A dominant-negative GFI1B mutation in the gray platelet syndrome. N. Engl. J. Med. *370*, 245–253.
30. Chen, L., Kostadima, M., Martens, J.H., Canu, G., Garcia, S.P., Turro, E., Downes, K., Macaulay, I.C., Bielczyk-Maczynska, E., Coe, S., et al.; BRIDGE Consortium (2014). Transcriptional diversity during lineage commitment of human blood progenitors. Science *345*, 1251033.
31. Vassen, L., Khandanpour, C., Ebeling, P., van der Reijden, B.A., Jansen, J.H., Mahlmann, S., Dührsen, U., and Möröy, T. (2009). Growth factor independent 1b (Gfi1b) and a new splice variant of Gfi1b are highly expressed in patients with acute and chronic leukemia. Int. J. Hematol. *89*, 422–430.
32. Koldehoff, M., Zakrzewski, J.L., Beelen, D.W., and Elmaagacli, A.H. (2013). Additive antileukemia effects by GFI1B- and BCR-ABL-specific siRNA in advanced phase chronic myeloid leukemic cells. Cancer Gene Ther. *20*, 421–427.
33. Thon, J.N., Medvetz, D.A., Karlsson, S.M., and Italiano, J.E., Jr. (2015). Road blocks in making platelets for transfusion. J. Thromb. Haemost. *13* (*Suppl 1*), S55–S62.

The American Journal of Human Genetics, Volume *99*

Supplemental Data

Whole-Exome Sequencing Identifies Loci Associated with Blood Cell Traits and Reveals a Role for Alternative *GFI1B* Splice Variants in Human Hematopoiesis

Linda M. Polfus, Rajiv K. Khajuria, Ursula M. Schick, Nathan Pankratz, Raha Pazoki, Jennifer A. Brody, Ming-Huei Chen, Paul L. Auer, James S. Floyd, Jie Huang, Leslie Lange, Frank J.A. van Rooij, Richard A. Gibbs, Ginger Metcalf, Donna Muzny, Narayanan Veeraraghavan, Klaudia Walter, Lu Chen, Lisa Yanek, Lewis C. Becker, Gina M. Peloso, Aoi Wakabayashi, Mart Kals, Andres Metspalu, Tõnu Esko, Keolu Fox, Robert Wallace, Nora Franceshini, Nena Matijevic, Kenneth M. Rice, Traci M. Bartz, Leo-Pekka Lyytikäinen, Mika Kähönen, Terho Lehtimäki, Olli T. Raitakari, Ruifang Li-Gao, Dennis O. Mook-Kanamori, Guillaume Lettre, Cornelia M. van Duijn, Oscar H. Franco, Stephen S. Rich, Fernando Rivadeneira, Albert Hofman, André G. Uitterlinden, James G. Wilson, Bruce M. Psaty, Nicole Soranzo, Abbas Dehghan, Eric Boerwinkle, Xiaoling Zhang, Andrew D. Johnson, Christopher J. O'Donnell, Jill M. Johnsen, Alexander P. Reiner, Santhi K. Ganesh, and Vijay G. Sankaran

Supplemental Note: Cohort Descriptions

Discovery Cohort Descriptions

The discovery sample included whole exome sequence data from 15,459 individuals, including 11,238 of European ancestry (EU), and 4,221 of African American ancestry (AA), participating in 6 population-based cohort studies: Atherosclerosis Risk in Communities (ARIC), Women's Health Initiative (WHI), Cardiovascular Health Study (CHS), Jackson Heart Study (JHS), Framingham Heart Study (FHS), and Rotterdam Study (RS). The whole exome sequences were generated as part of two separate projects, the CHARGE Consortium and the NHLBI Exome Sequencing Project, as described below.

Samples for complete blood count (CBC) analysis were obtained by venipuncture for collection into tubes containing ethylenediaminetetraacetic acid (EDTA). Phlebotomy, blood collection methods, hematology laboratory protocols, and blood measurement instruments have been previously described elsewhere[1-3]. Unit of measurement in all cohorts were as follows: hemoglobin (g/dl), hematocrit (%), mean corpuscular hemoglobin (pictogram), mean corpuscular hemoglobin concentration (%), mean corpuscular volume (femtoliter), erythrocyte count (1M cells/cmm), red cell distribution of width (femtoliter), white blood cell count (10^9 cells/L), neutrophil count (10^9 cells/L), eosinophil count (10^9 cells/L), basophil count (10^9 cells/L), monocyte count (10^9 cells/L), lymphocyte count (10^9 cells/L), platelet count (10^9/L), mean platelet volume (femtoliter), and mean morpuscular memoglobin moncentration = hemoglobin * 100 / hematocrit (in %). The cohort sample sizes by trait and ethnicity are shown in **Table S1** and all traits transformed as described in the Methods Section.

Atherosclerosis Risk in Communities (ARIC) Study

The ARIC study has been described in detail previously[4]. Men and women aged 45-64 years at baseline were recruited from four communities: Forsyth County, North Carolina; Jackson, Mississippi; Minneapolis, Minnesota; and Washington County, Maryland. A total of 15,792 individuals, predominantly White and African American, participated in the baseline examination in 1987-1989, with three additional triennial follow-up examinations and a fifth exam in 2011-2013.

Cardiovascular Health Study (CHS)

The CHS has been described in detail previously[5]. The CHS is a population-based cohort study of risk factors for coronary heart disease and stroke in adults ≥65 years conducted across four field centers. The original predominantly Caucasian cohort of 5,201 persons was recruited in 1989-1990 from random samples of the Medicare eligibility lists, and an additional 687 African-Americans were enrolled subsequently for a total sample of 5,888. DNA was extracted from blood samples drawn on all participants at their baseline examination in 1989-90.

Framingham Heart Study (FHS)

The FHS is a three generational prospective cohort that has been described in detail previously[6; 7]. Individuals were initially recruited in 1948 in Framingham, USA to evaluate cardiovascular disease risk factors. The second generation cohort (5,124 offspring of the original cohort) was recruited between 1971 and 1975. The third generation cohort (4,095 grandchildren of the original cohort) was collected between 2002 and 2005.

Jackson Heart Study (JHS)

The JHS is a large single-site, prospective, epidemiologic investigation of cardiovascular disease among African American adults in the three counties (Hinds, Madison, and Rankin) that comprise the Jackson, Mississippi metropolitan area.[8] The Jackson Heart Study involves a collaboration among three institutional partners, the Jackson community, and the National Institutes of Health to discover best practices for eliminating health disparities.

Rotterdam Study (RS)

The Rotterdam Study is an ongoing prospective population-based cohort study, focused on chronic disabling conditions of the elderly. The study comprises an outbred ethnically homogenous population of Dutch Caucasian origin. The rationale of the study has been described in detail elsewhere[9-11]. In summary, 7,983 men and women aged 55 years or older, living in Ommoord, a suburb of Rotterdam, the Netherlands, were included in the baseline exam.

Women's Health Initiative (WHI)

WHI is one of the largest (n=161,808) studies of women's health ever undertaken in the United States[12]. There are two major components of WHI: (1) a Clinical Trial (CT) that enrolled and randomized 68,132 women ages 50 – 79 into at least one of three placebo-control clinical trials (hormone therapy, dietary modification, and calcium/vitamin D); and (2) an Observational Study (OS) that enrolled 93,676 women of the same age range into a parallel prospective cohort study[13-16]. A diverse population including 26,045 (17%) women from minority groups were recruited from 1993-1998 at 40 clinical centers across the U.S. The design has been published[17; 18]. For the CT and OS participants enrolled in WHI and who had consented to genetic research, DNA was extracted by the Specimen Processing Laboratory at the Fred Hutchinson Cancer Research Center (FHCRC) using specimens that were collected at the time of enrollment in to the study.

Replication Cohort Descriptions

Replication of single variant and gene-based discovery findings was conducted in three sample sets: (1) JHS, ARIC, and CHS participants with whole exome sequencing not included in the discovery sample (N=5,536); (2) UK10K subjects with whole genome sequence (WGS) data or with WGS imputed data (N=31,551); and (3) exome chip data which was meta-analyzed from 9 cohorts (N=14,937), including non-overlapping subjects from the five discovery cohorts as well as the Cardiovascular Risk in Young Fins Study (YoungFins), Netherlands Epidemiology of Obesity (NEO) Study, and the Finnish Cardiovascular Study (FinCAVAS).

JHS exome sequence samples

Replication of single variant and gene-based discovery results was conducted using 2,437 independent African American exome sequence samples from Jackson Heart Study, not analyzed in the ESP discovery sample. Exome sequencing was performed in these additional 2,437 JHS samples either through the NHLBI Minority Health Genomics and Translational Research Bio-Repository Database (MH-GRID) (N=311) as described in Fu et al[19] or through the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Type 2 Diabetes Genetic Exploration by Next-Generation Sequencing in Multi-Ethnic Samples (T2DGENES) (N=571)[20]. Additional exome sequencing for JHS was also provided by an NHLBI funded study to Dr. Sekar Kathiresan (N=814). All samples were exome sequenced at either the Broad Institute or University of Washington and processed using the Genome Analysis ToolKit. Variant detection and genotyping were performed on both exomes and flanking 50bp of intronic sequence using the GATK v3 Haplotype Caller with recommended options. Variant recalibration was run with the GATK using the recommended resources and thresholds. Only VQRS PASS sites were used for further analysis. Samples were checked on their total number of variants, observed number of singletons and doubletons,Ti/Tv ratio, Het/Hom ratio, missingness, contamination statistics, and non-reference concordance with available exome chip data.

Additional ARIC and CHS exome sequence samples for platelet count

Subjects from ARIC (N=2404) and CHS (N=695) who were not included in the discovery sample were used for replication of the *GFI1B* and *CPS1* associations with platelet count.

UK10K whole genome and imputed samples

Replication of gene-based discovery results for hemoglobin was performed using whole genome sequence (WGS) data from 3,077 European ancestry samples from UK10K, which includes TwinsUK and ALSPAC[20]. Replication of gene-based results for other red blood cell traits (MCV, MCH, MCV) was available only in the TwinsUK sample (N=1,549). For replication of single variant results, the UK10K WGS and 1000 Genomes reference panels (Build 37) were used to impute genotypes into a total of 31,551 GWAS samples.

Blood cell exome chip consortium samples

Additional gene-based replication was performed using independent samples (N=14,937) genotyped on the exome chip array from ARIC, CHS, JHS, FHS, and Rotterdam studies as well as Cardiovascular Risk in Young Fins Study (YoungFins), Netherlands Epidemiology of Obesity (NEO), and the Finnish Cardiovascular Study (FinCAVAS) cohorts.

The Cardiovascular Risk in Young Finns Study (YFS)

The Young Finns Study is a population-based follow up-study started in 1980[21]. The main aim of the YFS is to determine the contribution made by childhood lifestyle, biological and psychological measures to the risk of cardiovascular diseases in adulthood. In 1980, over 3,500 children and adolescents all around Finland participated in the baseline study. The follow-up studies have been conducted mainly with 3-year intervals. The latest 30-year follow-up study was conducted in 2010-11 (ages 33-49 years) with 2,063 participants. The study was approved by the local ethics committees (University Hospitals of Helsinki, Turku, Tampere, Kuopio and Oulu) and was conducted following the guidelines of the Declaration of Helsinki.

The Finnish Cardiovascular Study (FinCAVAS)

The Finnish Cardiovascular Study consists of patients who undergo exercise stress tests at Tampere University Hospital and recruited between October 2001 and December 2007[22]. The study protocol was approved by the Ethical Committee of the Hospital District of Pirkanmaa, Finland, and all patients have given informed consent prior to the interview and measurements as stipulated in the Declaration of Helsinki. Follow-up data consists of information on major cardiovascular events (hospitalisation due to angina pectoris, MI or stroke), coronary procedures (angioplasties and bypass operations) and causes of deaths. The follow-up data was gathered at 2, 5 and 10 years. Information on clinical events will be collected from the Finnish National Hospital Discharge Register and data on mortality from the Causes of Death Register.

The Netherlands Epidemiology of Obesity Study (NEO)

The Netherlands Epidemiology of Obesity study is a population-based, prospective study of 6,673 individuals aged 45-65 years and oversampled in overweight or obese individuals[23]. Recruitment began in September 2008 and ended in September 2012 in the greater area of Leiden, Netherlands. NEO was designed to investigate pathways leading to obesity-related diseases and mortality with extensive phenotyping such as anthropometry, electrocardiography, spirometry, and ultrasonography to measure carotid artery intima-media thickness.

Estonian Biobank

The Estonian Biobank cohort represents approximately 5% of the Estonian adult population recruited from general practitioners throughout the country, who also donate blood samples for DNA, whole blood and plasma tests[24]. Through national electronic database and registry linkages, medical information is followed up and in some instances participants contacted for follow-up of targeted interventions. Given the database linkages and subset follow up, the Estonian Biobank is a population-based biobank with a longitudinal and prospective database. Genetic and biological samples were sequenced or genotyped at the Estonian Genome Center of the University of Tartu (EGCUT).

Description of Gene-based Association Results

A summary of the most significant association results for each trait (*P* $<10^{-4}$) are shown in **Tables S5A for gene-based burden tests** and **S5B for SKAT.**

Previously undescribed genes meeting our discovery significance threshold for either the gene-based test or a single variant within the gene are shown in **Table 1** and **Table S3.**
There were four additional gene-based rare variant SKAT (*MRPL43, ACTN4*) or burden test results (*MYOM2*, *MMACHC*) that met our pre-defined significance threshold in the discovery sample (P<2.63 x 10^{-6}), using either the broader variant annotation filter or the narrower LOF variant filter. All four of these gene-based associations were with erythrocyte-related traits (**Table S5A** and **Table S5B**). While no gene burden finding was replicated in our analyses of independent samples **(Tables S12, S13,** and **S14)**, we also attempted to replicate associations for several individual variants driving the gene-based findings, as described further below for each gene. While no gene burden finding was replicated in our analyses of independent samples, suggestive evidence was found for several individual driving variants comprising the gene burden findings, as described below for each gene.

MYOM2

The myomesin 2 (*MYOM2*) gene on 8p23.3 was significantly associated with MCHC in the race-combined meta-analysis (P=2.16 x 10^{-6}), and the lowest p-value of the missense SNP within this gene (rs150511184) was noted among non-overlapping samples of the exome chip with P=0.00895. Additionally, there were three rare missense variants in *MYOM2* that were nominally associated (P=0.02 to 0.05) with MCH in both the discovery analysis and replication analysis of exome chip data. *MYOM2* is expressed in the m bands of sarcomeres encoding for the m-protein. The relationship between mean corpuscular hemoglobin concentration and *MYOM2* is unclear.

MRPL43

The mitochondrial ribosomal protein L43 gene (*MRPL43*) was significantly associated with hemoglobin levels in EAs (P =1.16x10^{-6}). No significant replication was available for *MRPL43* (JHS exome sequence in AA's T5 burden test P=0.91).

ACTN4

Actinin, alpha 4 (*ACTN4*), a cytoskeletal protein-encoding gene, was significantly associated with MCV in our AA discovery sample (P=1.59x10^{-6}). While we were unable to replicate gene-based association in our AA replication sample (SKAT P=0.138), the P-value for association of the *ACTN4* gene burden test in the TwinsUK EU replication cohort was 0.057 (N=1,548). From non-overlapping AA exome sequenced JHS participants (N=2,437), there was borderline evidence of association for one of the individual *ACTN4* missense variants driving the gene-based association test (rs149027682, P179L; discovery P=3.10x10^{-6}; replication P=0.06) with both discovery and replication samples having a negative direction of effect.

Mutations in *ACTN4* are found in patients with the rare autosomal dominant disorder, familial focal segmental glomerulosclerosis (FSGS)[25]. Actinins belong to the spectrin gene superfamily. *ACTN4* is a cytoskeletal protein which promotes cell motility and cancer metastasis and invasion. *ACTN1* has recently been characterized as a causal gene for familial thrombocytopenia[26-28]. It is plausible that the putative association of *ACTN4* variants with MCV may be mediated through red cell cytoskeletal alterations. Two of the driving variants of the *ACTN4* association with MCV (rs149027682 (Pro179Leu) and rs113969422 (Asp890Glu)) are located within two distinct domains of the *ACTN4* gene. The more robust signal (rs149027682, MAF=0.01, P=3.10x10^{-6}) is located in the actinin-4 calponin homology domain which has a role in actin crosslinking; this variant was associated with lower MCV. The second signal (rs113969422, MAF=0.01, P=0.02) is located in the EF hand domain, which is involved in actin cytoskeleton signaling; this variant was associated with increased MCV.

MMACHC

In an analysis restricted to predicted LOF mutations, the methylmalonic aciduria and homocystinuria type C protein (*MMACHC*) gene was associated with Hb levels (P=1.26x10^{-6}) and Hct (P=1.63x10^{-6}) in the EU-only discovery meta-analysis. A different missense variant (rs192924272) achieved nominal significance in AA JHS exome sequence replication for a rare missense variant in the *MMACHC* (P=0.0018).

Mutations in *MMACHC* account for inborn errors of vitamin B12 metabolism and result in methlmalonic aciduria and homocystinuria, CblC type. In context of our LOF gene-based findings with Hb and Hct, vitamin B12 is necessary for the formation and maturation of erythrocytes. CblC is an extremely rare recessive Mendelian disorder caused by homozygous or compound heterozygous mutations; it has been identified in only ~400 individuals worldwide[29]. Treatment with a daily dose of vitamin B6 (pyridoxine), or eating a low-methionine diet may abate development of this multisystem disorder, which can involve connective tissue, mental retardation, eye anomalies, megaoblastic anemia, and vascular disease. Typically identified in infants and young children, LOF variants, including Arg132X, have been identified in individuals with late onset methylmalonic aciduria and homocystinuria, cbIC type disorder (OMIM 277400) [29], and appears to be mediated by nonsense-mediated mRNA decay (NMD). Notably, the association of *MMACHC* with lower hemoglobin in our discovery sample was identified only by restricting the analysis to predicted LOF variants. As an incidental finding, the frameshift mutation (rs398124294, 271dupA) studied in 118 cases of CblC accounted for 43% of the pathogenic alleles[29]; the same frameshift mutation occurred in twelve heterozygotes in our discovery sample. A common variant of *MMACHC* was associated with homocysteine levels (rs4660306, MAF=0.33, P=2.33x10^{-9}) [30].

Associations of GFI1B and CPS1 Variants with Platelet Surface Marker Expression, Mean Platelet Volume, and Platelet Aggregation

To further assess the relationship of the two platelet count-associated variants (*GFI1B* rs150813342 and *CPS1* rs1047891) to platelet morphologic and functional characteristics, we analyzed mean platelet volume, platelet flow cytometric, and agonist-induced aggregation measures from several available data sets.

Platelet Flow Cytometry: ARIC Carotid MRI Study

Recruitment for the ARIC Carotid MRI Study (N= 1,889) was comprised from a subset of the original ARIC cohort and described elsewhere using a stratified weighted sampling design[31; 32]. Flow cytometry was used to assess cell surface markers and size in whole blood with analysis performed on the Epics^{TM} XL^{TM} (Beckman Coulter) [31].

In ARIC, flow cytometry of peripheral blood stained with cellular markers measurements were analyzed for association with the *GFI1B* (rs150813342) and *CPS1* (rs1047891) variants. Six surface markers included CD14, CD41, CD45, CD61, CD62P (P-selectin), and CD154. Events were determined by light scatter characteristics/size and cell surface marker expression and designated as (CD41+CD61+) platelet events. Events positive for (CD61+CD45+ CD14+) were designated as platelet-monocyte events, (CD61+ CD45^{low}CD14-) events were interpreted as platelet-granulocyte events, and (CD61+ CD45^{high}CD14-) were interpreted as platelet-lymphocyte events. To account for complex survey design, the analyses utilized the inverse of the sampling fractions as weights to calculate variances and confidence intervals of estimators using SAS 9.3. Regression models were adjusted for age, sex, genetic ancestry, and platelet count. Individuals taking antiplatelet medication use including Plavix and/or aspirin, were excluded from the analysis. The Bonferroni corrected threshold for 9 flow cytometry measurements was calculated at P<0.00625.

Flow cytometry available on a subsample of 761 individuals in ARIC showed that expression of cell surface markers did not differ significantly according to *GFI1B* or *CPS1* genotypes (**Tables S8 and S10**). The rs150813342 *GFI1B* variant was associated with lower Median Fluorescence Intensity (MFI) of large platelet events, P<0.0001 (**Table S8A**); the mean difference between carriers and non-carriers was -35.6%. Analysis of MFI forward light scatter (FS) showed that the rs150813342 variant was associated with lower MFI of larger platelet events (P=0.003), but there was no association with FS at the expected size of single platelets (**Table S8B**; P=0.26). MFI was not significantly associated with the CPS1 variant, rs1047891 (**Table S10**; P=0.17).

Mean Platelet Volume (MPV)

We assessed association with MPV using a sample of 570 individuals derived from an Estonian biobank[24] as described in the Cohort Descriptions above, of whom 10 harbored the *GFI1B* rs150813342 variant. There was no statistically significant association between rs150813342 and MPV (P= 0.86). The 10 individuals heterozygous for rs150813342 had MPVs of 10.7 +/- 0.86 (9.7-12.0), suggesting that the decreased MFI for platelet-platelet events noted by flow cytometry is not an artifact due to altered platelet size in the *GFI1B* rs150813342 rare allele carriers.

Platelet Aggregometry

Framingham Heart Study participants in the Offspring Generation, examination 5 (1991-1995) were measured for platelet reactivity to three agonists [adenosine diphosphate (ADP), epinephrine and collagen] using a four-channel aggregometer (BioData Corp., Horsham, PA)[33]. Platelet aggregation of GeneSTAR participants was measured at baseline[34]. Aggregometry measures included platelet aggregation to arachidonic acid, ADP, epinephrine, and platelet function analyzer closure time. The FHS and GeneSTAR platelet aggregometry studies were both adjusted for age, sex, and ancestry, with outcome measurements undergoing Blom transformation due to non-Gaussian distributions. Sample size weighted meta-analysis of FHS and GeneSTAR associations with rs150813342 *GFI1B* variant was performed in the METAL software.

There was no detectable association of the GFI1B variant with collagen-, ADP-, or epinephrine-induced platelet aggregation measures among 1,955 GeneStar+FHS individuals, of whom 20 were carriers of the rs150813342 variant allele (**Table S9**).

Supplemental Figures

Figure S1

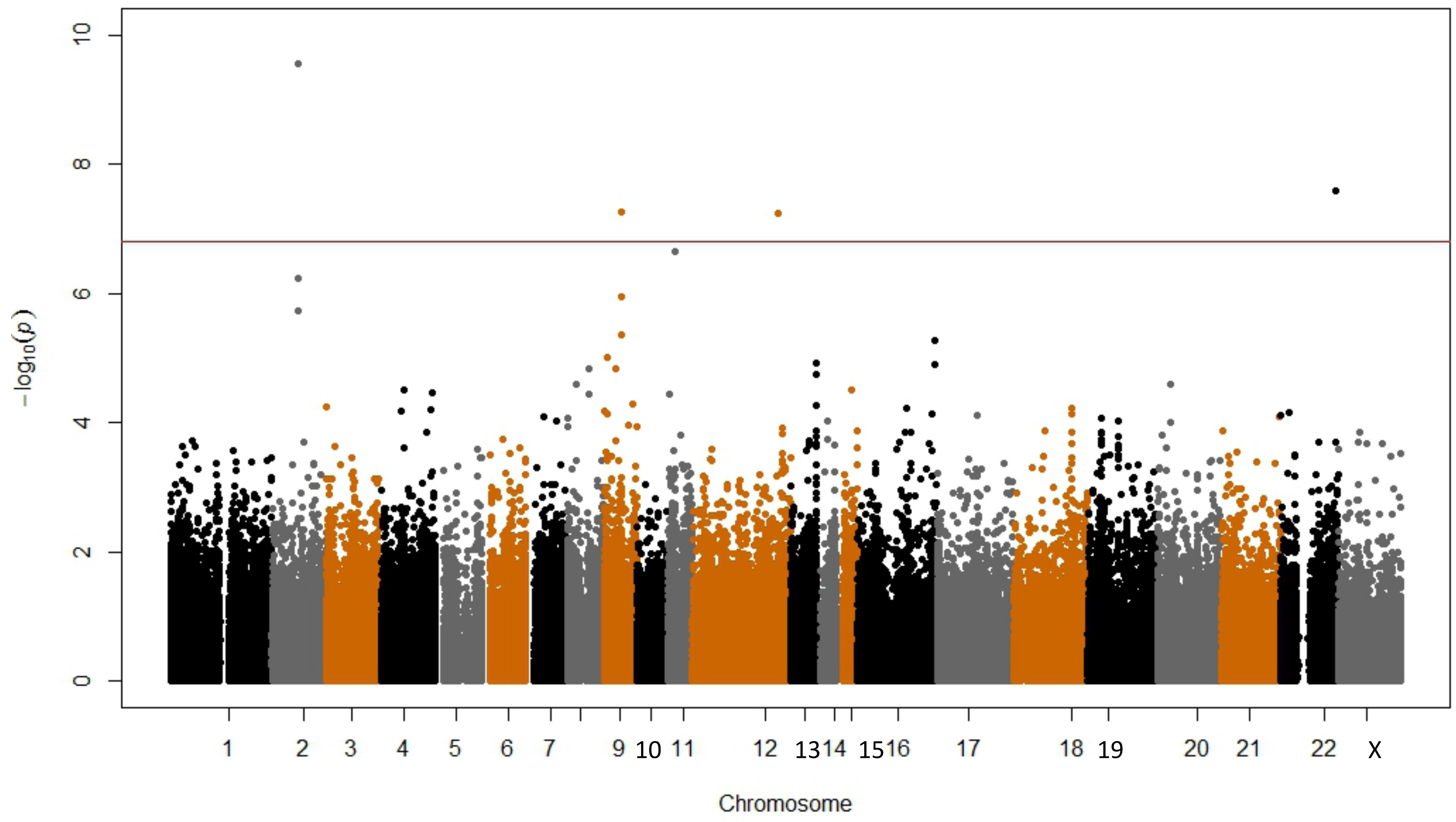


Figure S1. Manhattan Plot of Platelet Count meta-analyzed by both ethnicity and study. The red horizontal line denotes the single variant significance threshold. On the x-axis are the autosomes and X chromosome. The y-axis is the negative log p value (base 10) of the associations for platelet count and SNV. Whole Exome Sequencing and Quality Control: In the CHARGE Consortium, DNA sequencing was performed at Baylor College of Medicine's Human Genome Sequencing Center on Illumina HiSeqs (San Diego, CA) after exome capture with NimbleGen's VCRome2.1. Prior to statistical analysis, data were processed and alleles jointly called using Mercury[35]. Sequencing yields had an average of 92% of target sites, depth of coverage of 20X or greater, and a mean sequencing depth of 92X. Each SNV call was filtered based on the following criteria to produce a high-quality variant list: low SNV posterior probability (<0.95), low variant read count (<3), variant read ratio <0.25 or >0.75, strand-bias of more than 99% variant reads in a single strand direction, or total coverage less than 10-fold. All variant calls filtered by these criteria, and reference calls with less than 10-fold coverage, were set to missing. The variant call filters were the same for indels except a total coverage less than 30-fold was used for variant sites. As part of quality control procedures, variant sites were removed that had greater than 20% missingness, more than two observed alleles, mean depth at the site of greater than 500-fold, or were monomorphic. Variants not meeting Hardy-Weinberg equilibrium expectations ($P<5x10^{-6}$) in ancestry-specific groups were also excluded. Within each cohort, samples were removed that fell beyond six standard deviations in any of four measures that were calculated by cohort and ancestry group: number of singletons, heterozygote to homozygote ratio, mean depth, or transition to transversion (Ti/Tv) ratio. In the Exome Sequencing Project, the processes of library construction, exome capture, sequencing, and mapping were performed as previously described[36-38]. Sequencing was performed at the University of Washington and the Broad Institute of MIT/Harvard. Briefly, exome capture was performed using Roche Nimblegen SeqCap EZ or Agilent SureSelect Human All Exon 50 Mb. Paired end sequencing (2 x 76bp) was performed on Illumina GAII and HiSeq instruments. Single Nucleotide Variants (SNVs) were called using a maximum likelihood approach[39] implemented in the UMAKE pipeline at the University of Michigan, which allowed multiple samples to be analyzed simultaneously, both for variant calling and filtering. Binary Alignment/Map (BAM)[40] files summarizing Burrows-Wheeler Alignment (BWA)[39] alignments were mapped to the Genome Reference Consortium Human Build 37 (GRCh37), refined by duplicate removal, recalibration, and indel re-alignment

using the Genome Analysis ToolKit (GATK)[41]. We excluded all reads that were not confidently mapped (Phred-scaled mapping quality < 20) from further analysis. Mean depth was 127X in targeted regions. We then computed genotype likelihoods for exome targeted regions and 50 flanking bases, accounting for per base alignment quality using SAMtools[40]. Variable sites and their allele frequencies were identified using a maximum-likelihood model, implemented in glfMultiples[42]. The final call-set was performed on 6,823 samples. The Exome Sequencing Project used a support vector machine classifier to separate likely true positive and false- positive variant sites, applying a series of variant-level filtering metrics, including allelic balance, base quality distribution for sites supporting the reference and alternate alleles, and the distribution of supporting evidence between strands and sequencing cycle. These steps were followed by quality control on individual samples within each study using variants identified by dbSNP[43] or 1000 Genomes[44] as the positive training set and variants that failed multiple filters as the negative training set. We excluded variants with read depth greater than 500 or missingness rates ≥ 10%. Samples with discrepant self-reported race and ancestry derived from principal component analysis (PCA) performed on exome sequencing data in PLINK[45] as well as well as ancestry outliers by PCA were removed. Samples having very low concordance (<90%) with previously-obtained SNP array data were considered likely sample swaps and were also dropped from further analysis.

Figure S2

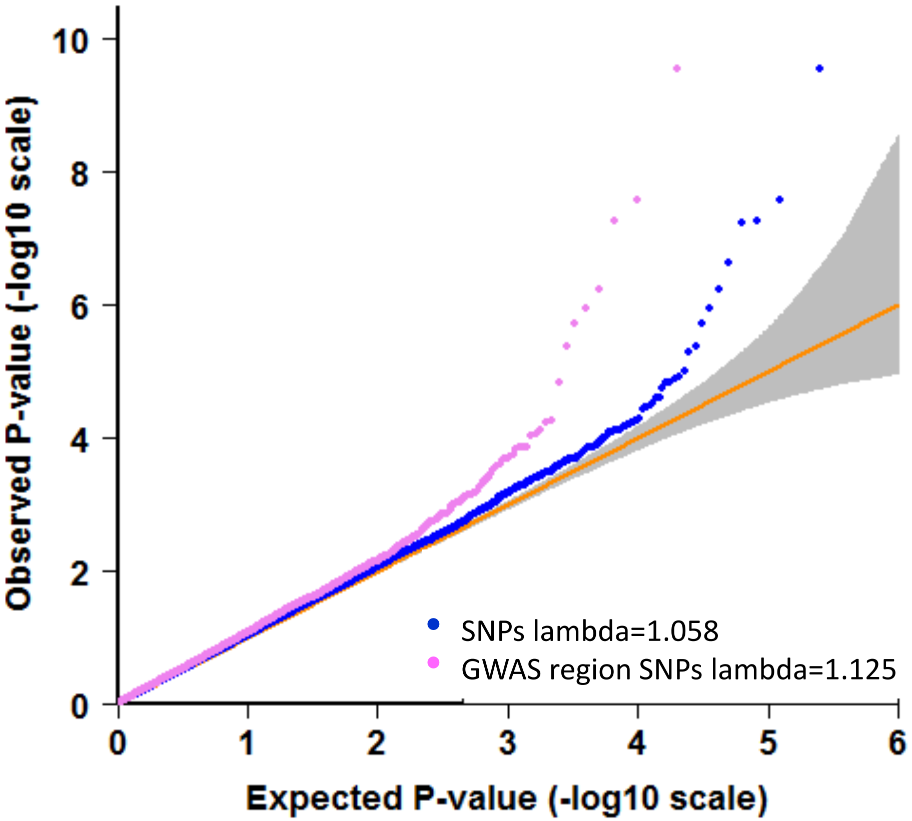
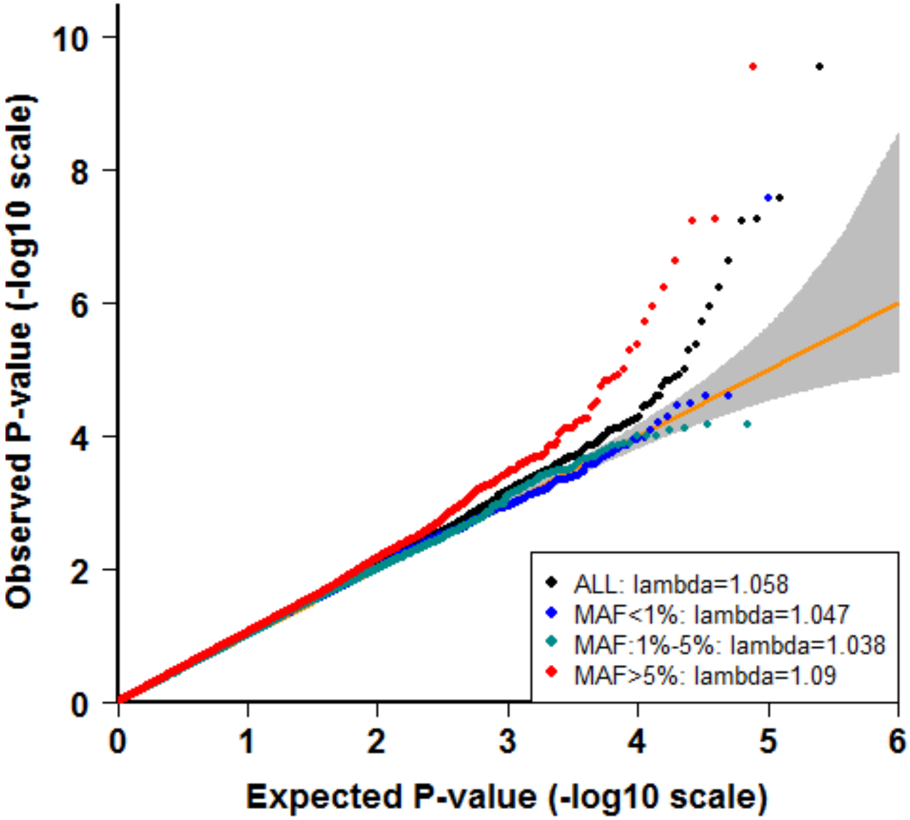

Observed P-value (-log10 scale)
10
8
6
4
2
0
0
1
2
3
4
5
6
Expected P-value (-log10 scale)
● SNPs lambda=1.058
● GWAS region SNPs lambda=1.125

Figure S2. QQ Plot of p values comparing observed (y-axis) and the expected distribution (x-axis) for all SNVs (blue dots) as well as SNVs found to be associated with blood traits in previous GWAS (pink dots). Association Analysis of Single Variants and Gene-based Tests: For each phenotype, we calculated the rank-based inverse normal transformation of the trait residuals within each cohort, adjusted for age and sex. Principal components were included as covariates in each cohort's analysis. The R seqMeta (v1.5) package was utilized to conduct individual cohort analyses, conditional analyses, and meta analyses (https://cran.r-project.org/web/packages/seqMeta/index.html). Variant functional categories were annotated using ANNOVAR[46] and included missense, synonymous, frameshift, premature stop codons, splice sites, and small indels, and missense variants predicted to disrupt protein function. Small indels were included in the gene-based test based on previous validation studies of whole exome sequencing data for the ARIC cohort[47]. Two types of association tests were conducted, using either single variants or genes as the unit of analysis. Single variant analysis was conducted for variants having a minor allele count of at least ten overall for ethnicity specific or trans-ethnic using linear regression. Gene-based tests consisted of a burden method (T5) and sequence kernel association test (SKAT)[48]. The burden method calculates an indicator variable by summing the contributing variants meeting the inclusion criteria across individuals; the indicator variable is used as an independent variable in a linear regression model. The SKAT method employs a kernel density approach to aggregate individual score test statistics of a set of variants, weighted by their allele frequencies, and allows different variants within a gene to have different directions and different magnitudes of effect. For each gene-based test, we included only variants with minor allele frequency less than 5% pooled across cohorts and ethnicities and applied two different variant annotation criteria. The broader set of variants included missense, stop gain, stop loss, frameshift, or splice site. A second annotation filter included only loss-of-function (LOF) variants (premature stop codons, splice site, and frameshift). To declare statistical significance in any of the single variant analyses (EU, AA, or EU + AA combined), a Bonferroni threshold was calculated to account for the total number of single variant tests with a minor allele count (MAC) of at least ten ($P<1.6x10^{-7}$). For gene-based tests, the Bonferroni threshold accounted for the number of genes (N=18,991) analyzed with a MAC of at least five ($P<2.63x10^{-6}$). The same gene-based significance threshold was applied whether the meta-analysis was conducted within EU, within AA, or EU + AA combined, and whether a broad variant annotation

filter or a more stringent LoF filter was applied. Three single variants and four genes meeting the discovery significance threshold were taken forward for replication. The Bonferroni-corrected significance threshold for replication analyses was 0.05 divided by seven (the total number of discovery findings), or P<0.007. The replication sample included several data sets with either whole exome sequencing (N=5,536), whole genome sequencing followed by imputation into a larger GWAS data set (N=31,551), or exome array genotyping (N=14,937). For each analysis (single variant and gene-based), cohort-level results were combined by fixed effects meta-analysis at a central analysis site using the seqMeta package. Fixed effect meta-analyses were performed separately within each ethnicity (EU or AA) and also across both ethnicities (EU + AA). Conditional analysis of DARC locus and WBC: Several SNV located on chromosome 1, including the African ancestral Duffy blood group, atypical chemokine receptor (DARC) rs2814778 promoter variant, were associated with WBC and neutrophil count in the AA-specific analysis. Conditional regression analysis of total WBC count and neutrophil count was performed in a subset of the AA discovery samples (N=4221) for all significant variants (MAC>10) in the region spanning 157,485,561 to 159,175,354. In each regression model, the functional DARC variant rs2814778 (coded in a dominant manner) was included as an additional covariate. After conditioning on rs2814778, there were no significant, independent association signals for WBC or neutrophil count in this region (**Table S11**).

Figure S3



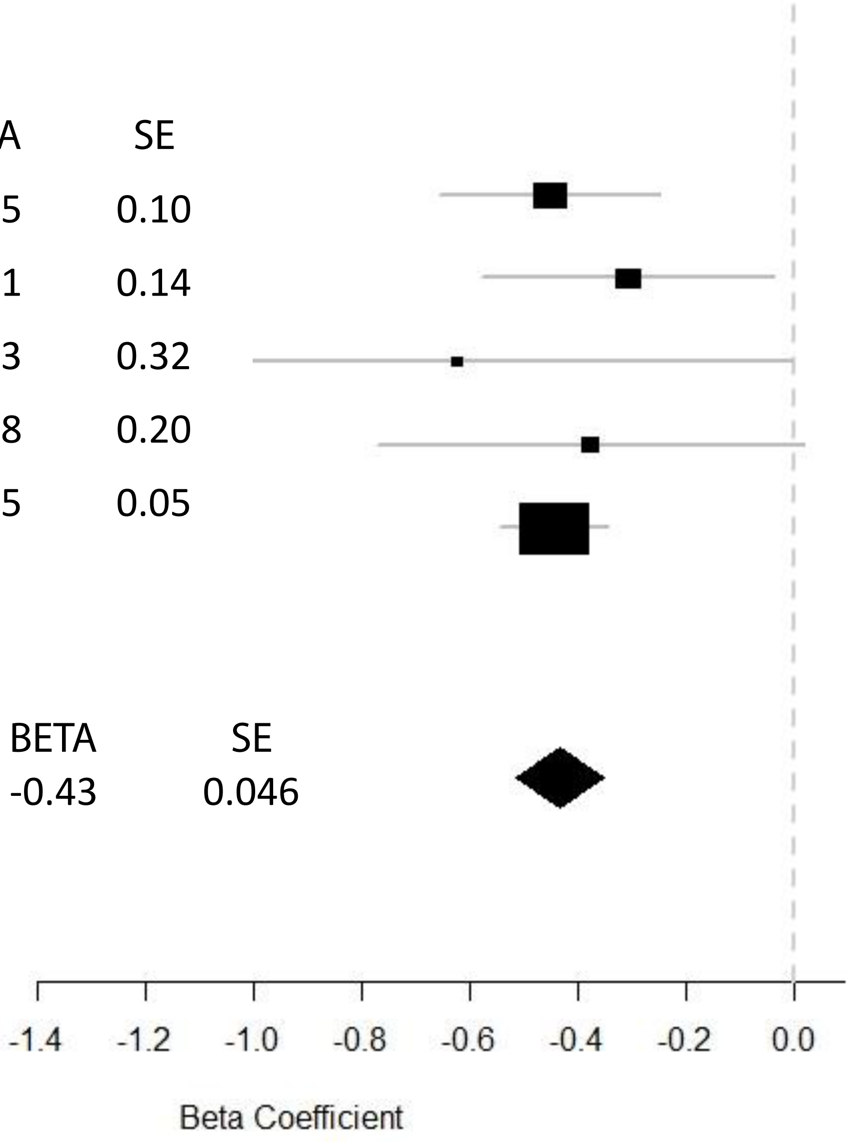
Observed P-value (-log10 scale)
0
2
4
6
8
10
Expected P-value (-log10 scale)
0
1
2
3
4
5
6
ALL: lambda=1.058
MAF<1%: lambda=1.047
MAF:1%-5%: lambda=1.038
MAF>5%: lambda=1.09

Figure S3. QQ Plot of SNV association with platelet count p values comparing observed (y-axis) and the expected distribution (x-axis) and further categorized by Minor Allele Frequency bins shown and color coded in the legend and corresponding lambda gc value.

Figure S4

Study	P value	MAF	N	BETA	SE
ARIC	1.52E-05	0.008	5673	-0.45	0.10
RS	0.026	0.017	1487	-0.31	0.14
CHS	0.049	0.007	685	-0.63	0.32
ESP	0.061	0.007	1668	-0.38	0.20
UK10K	5.07E-18	0.007	31549	-0.45	0.05

	P value	MAF	N	BETA	SE
META	1.79E-27	0.009	9513	-0.43	0.046



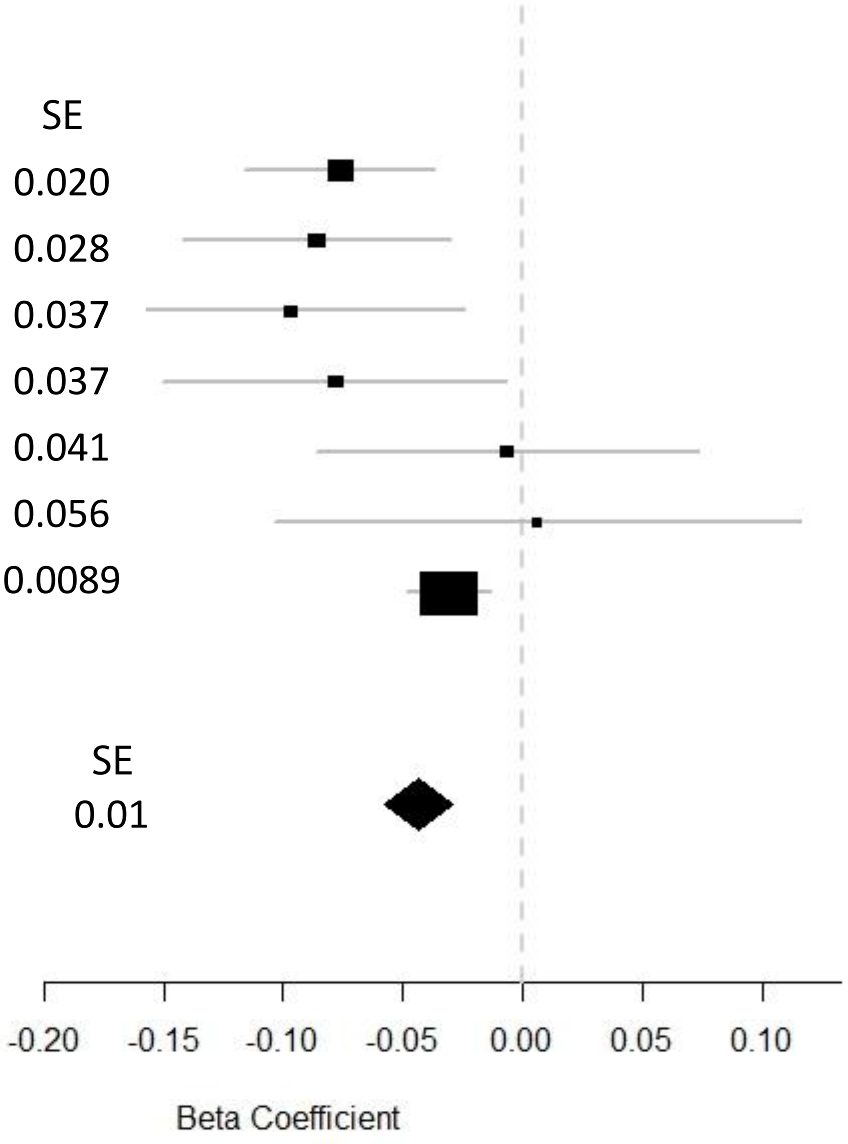
Test for heterogeneity: χ2(4)=1.36
P-value: 0.85

Figure S4. Forest Plot of platelet count association with *GFI1B* variant rs150813342 by discovery and replication sample study. MAF is minor allele frequency, N is sample size, Beta is beta coefficient, and SE (standard error). Beta Coefficients is on the x-axis with the black box denoting effect size and horizontal lines denoting standard error. The non-significant result for a chi-square test for heterogeneity is shown in the lower left corner.

Figure S5

Study	P value	MAF	N	BETA	SE
EU ARIC	1.7E-04	0.31	5673	-0.076	0.020
AA ARIC	0.0025	0.37	2683	-0.086	0.028
EA ESP	0.0088	0.32	1668	-0.097	0.037
AA ESP	0.033	0.37	1548	-0.078	0.037
EU RS	0.88	0.30	1487	-0.0064	0.041
EU CHS	0.92	0.31	685	0.0060	0.056
UK10K	1.02E-04	0.31	31551	-0.031	0.0089

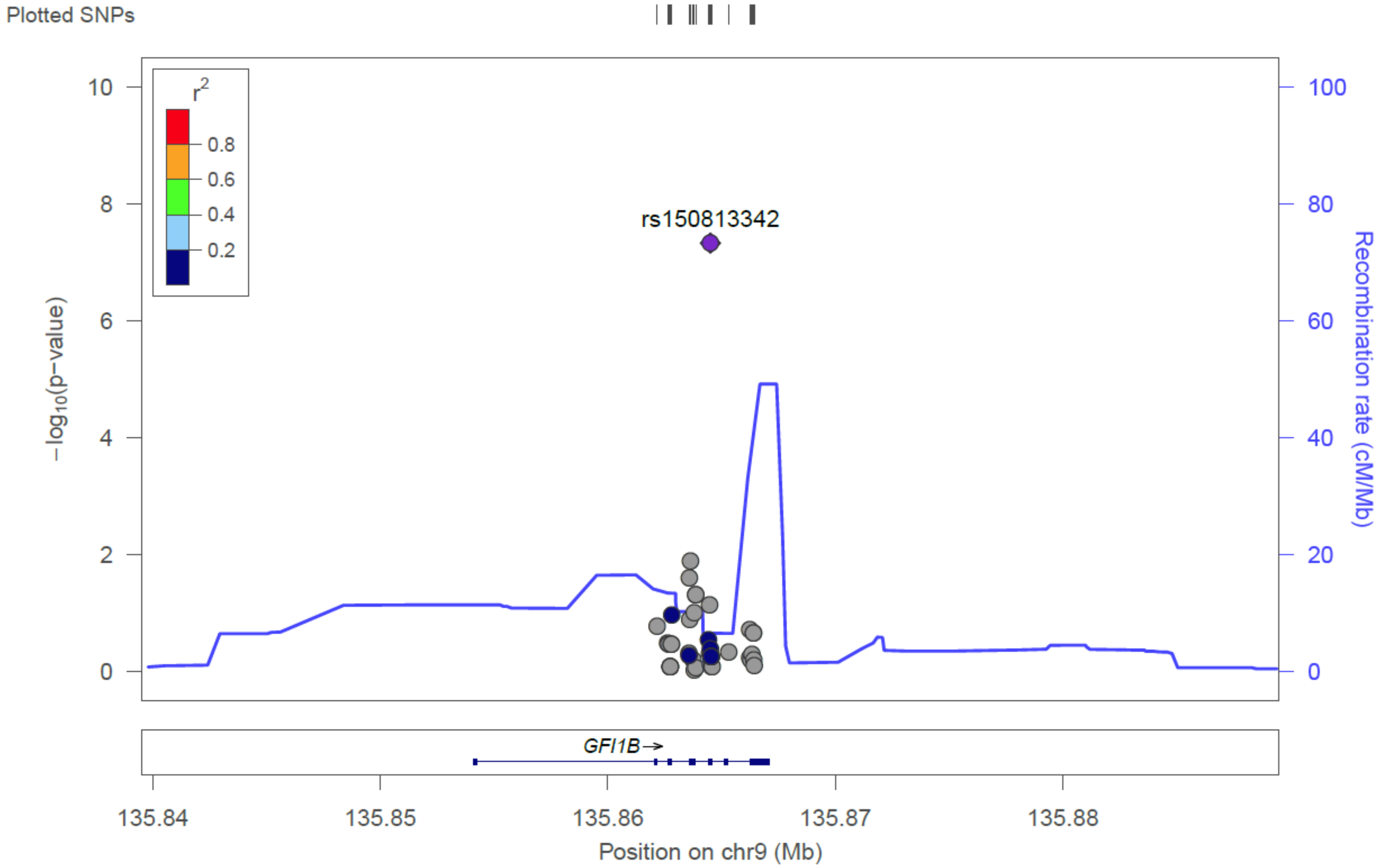
	P value	MAF	N	BETA	SE
META	6.38E-10	0.33	45295	-0.04	0.01



Test for heterogeneity: χ2(6)=11.48
P-value: 0.075

Figure S5. Forest Plot of platelet count association with *CPS1* variant rs1047891 by discovery and replication sample study. MAF is minor allele frequency, N is sample size, Beta is beta coefficient, and SE (standard error). Beta Coefficients is on the x-axis with the black box denoting effect size and horizontal lines denoting standard error. The non-significant result for a chi-square test for heterogeneity is shown in the lower left corner.

Figure S6



Plotted SNPs
r^2
0.8
0.6
0.4
0.2
rs150813342
10
8
6
4
2
0
−log10(p−value)
100
80
60
40
20
0
Recombination rate (cM/Mb)
GFI1B→
135.84
135.85
135.86
135.87
135.88
Position on chr9 (Mb)

Figure S6. Regional plot of *GFI1B* with chromosome position on the x-axis as well as a representation of *GFI1B* exon locations. On the left y-axis is the negative log p value of the association of the platelet count of European ethnicity. The right y-axis is the recombination rate as indicated by the blue line. The legend r2 shows the linkage disequilbrium with lowest p value in the gene (rs150813342) (EU hg19).

Figure S7

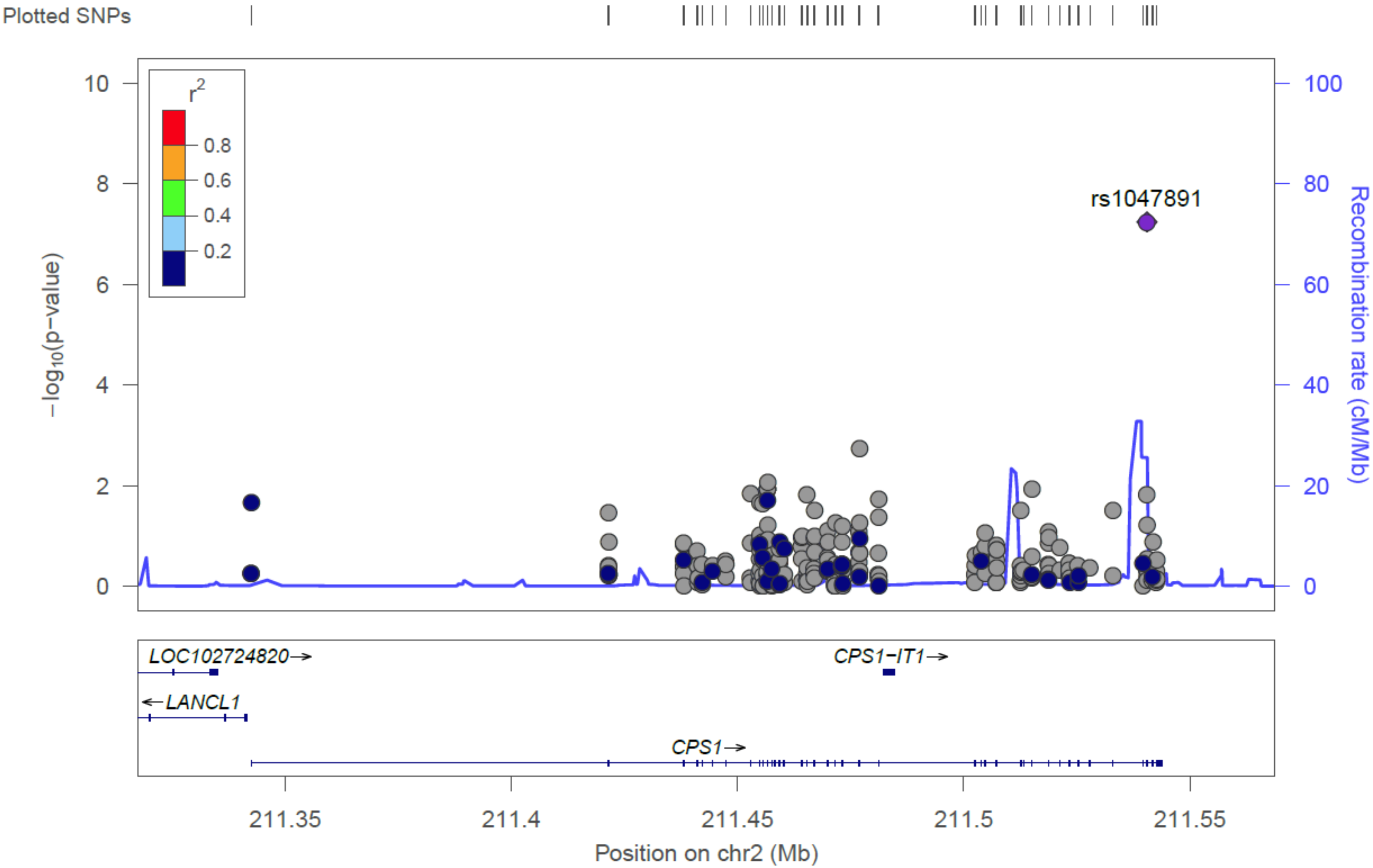

Plotted SNPs
r2
0.8
0.6
0.4
0.2
10
8
6
4
2
0
−log10(p−value)
100
80
60
40
20
0
Recombination rate (cM/Mb)
rs1047891
LOC102724820→
←LANCL1
CPS1−IT1→
CPS1→
211.35
211.4
211.45
211.5
211.55
Position on chr2 (Mb)

Figure S7. Regional plot of *CPS1* with chromosome position on the x-axis as well as a representation of *CPS1* exon locations. On the left y-axis is the negative log p value of the association of the platelet count of European ethnicity. The right y-axis is the recombination rate as indicated by the blue line. The legend r2 shows the linkage disequilbrium with lowest p value in the gene (rs1047891) (EU hg19).

Figure S8

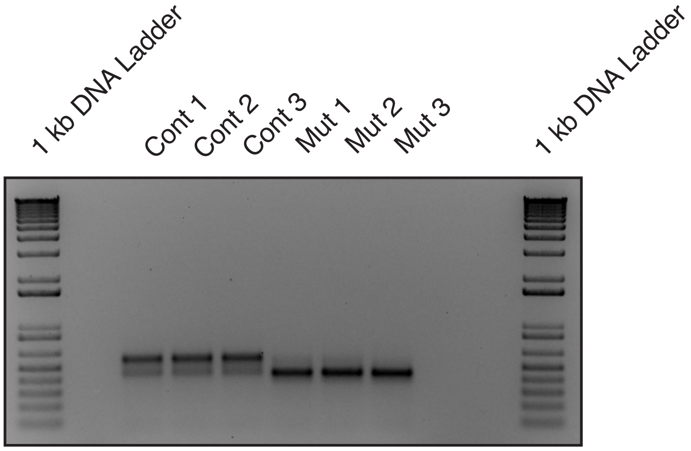

1 kb DNA Ladder
Cont 1
Cont 2
Cont 3
Mut 1
Mut 2
Mut 3
1 kb DNA Ladder

Figure S8. No detection of intermediate isoforms or intron inclusion events in *GFI1B* mRNA from K562 cells. Semi-quantitative RT-PCR showing preferential formation of the short *GFI1B* isoform with no other intermediate isoforms or intron inclusions in the isogenic mutant cell clones, while the isogenic control clones preferentially express the long *GFI1B* isoform. Clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 genome editing technology was used to introduce the C > T single nucleotide variant (SNV) into K562 hematopoietic cells via homology-directed repair (HDR)[49]. A single-guide RNA (sgRNA) (5'-CCGCAGATGTCACAGGCGAA-3') was cloned into the pSg1 vector using the BplI and NheI restriction sites. A single-stranded DNA oligonucleotide (ss-oligo) containing the SNV and flanking sequences of 90 nucleotides on each side was co-transfected with the sgRNA and PxPR001 Cas9 plasmid using the Nucleofector 2b device (Lonza) according to the manufacturer's protocol and as described previously[50]. Following puromycin selection (24-72 hours post-transfection), serial dilutions were performed and isogenic clones were screened by Sanger sequencing for the SNV 3 weeks later once individual colonies had grown to a sufficient size. Quantitative or semi-quantitative reverse-transcriptase PCR (RT-PCR) was performed with the following GFI1B primers (Forward = F, Reverse = R): Exon 4 - F: 5'-CATTGTGCTGTCCCGACC-3', R: 5'-CATAGGTTGTGGCCAAGGTG-3'; Exon 5 - F: 5'-ATGTGCGACGCTCCCATAG-3', R: 5'-CTGGGAGTGGACGTGCGT-3'; Exon 6 - F: 5'-ACCTGCTCATCCACTCAGAC-3', R: 5'-GTGTGGATGTAGGTGTGCTTC-3'. Primers F4 and R6 were used for the semi-q RT-PCR.

Figure S9

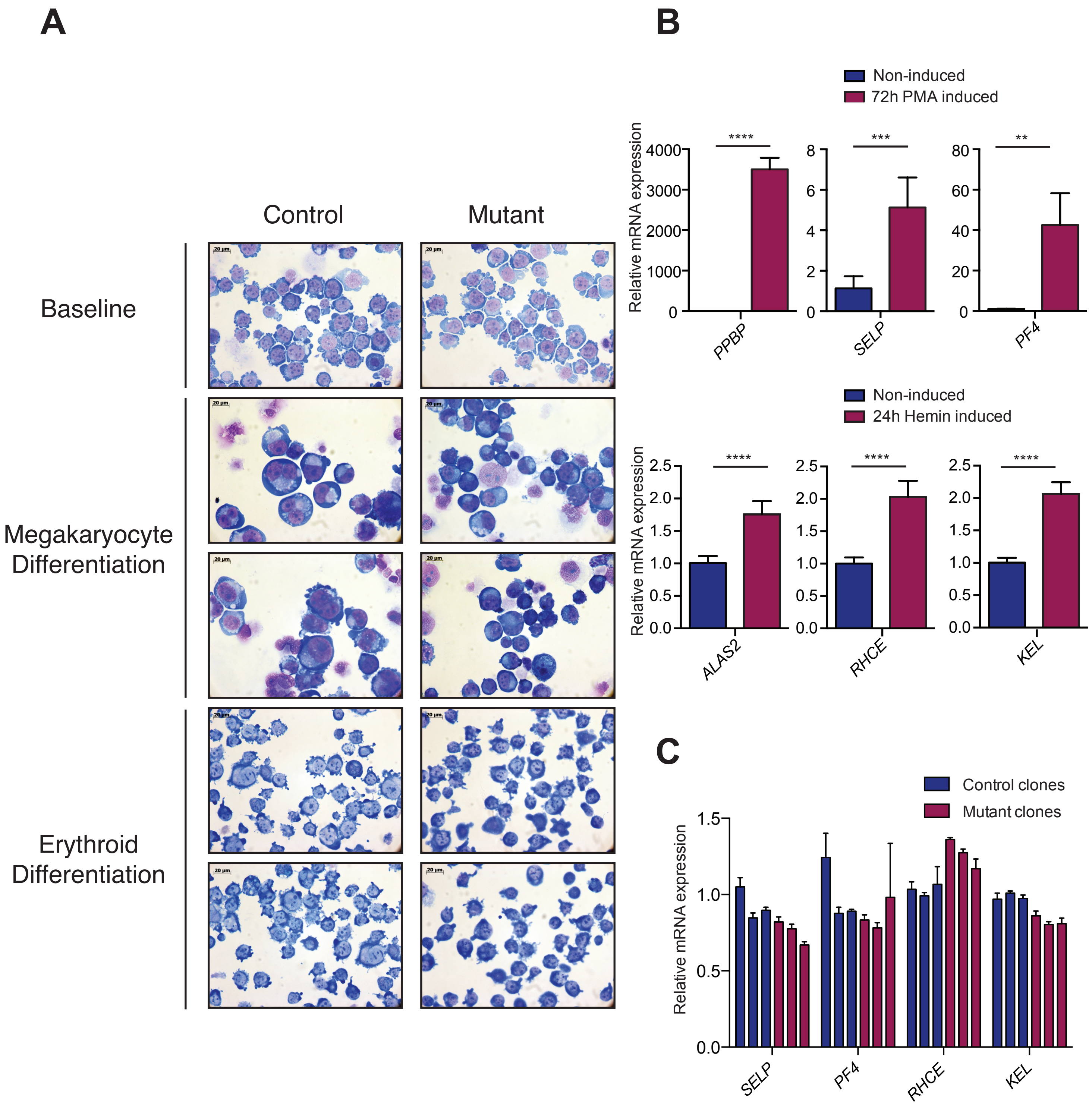

A
Control
Mutant
Baseline
Megakaryocyte Differentiation
Erythroid Differentiation
20 µm
B
Non-induced
72h PMA induced
Relative mRNA expression
0
1000
2000
3000
4000
****
PPBP
0
2
4
6
8
***
SELP
0
20
40
60
80
**
PF4
Non-induced
24h Hemin induced
Relative mRNA expression
0.0
0.5
1.0
1.5
2.0
2.5
****
ALAS2
****
RHCE
****
KEL
C
Control clones
Mutant clones
Relative mRNA expression
0.0
0.5
1.0
1.5
SELP
PF4
RHCE
KEL

Figure S9. Mutant K562 clones exhibit morphologically impaired megakaryocytic differentiation, while erythroid differentiation is unaffected. **(A)** Representative May-Grünwald-Giemsa stained cytospin images of 72 h PMA-induced megakaryocytic differentiated and 96 h hemin-induced erythroid differentiated isogenic control and mutant clones showing a retained immature blast-like morphology with smaller cell size in megakaryocytic differentiated clones harboring the C > T nucleotide change, while erythroid differentiation appears to be not affected compared to control clones. **(B)** Wild-type K562 gene expression analysis by quantitative RT-PCR of the megakaryocyte markers *PPBP*, *SELP*, and *PF4* after 72 h of PMA-induced differentiation and of the erythroid markers *ALAS2*, *RHCE*, and *KEL* after 24 h of hemin-induced differentiation compared to non-induced cells, respectively (n = 3 per group). ****P < 0.0001; ***P = 0.0005; **P = 0.001 using the unpaired two-tailed t-test. Megakaryocytic and erythroid differentiation of K562 cells was performed as described previously[51; 52]. Megakaryocytic differentiation was induced with 5 nM phorbol 12-myristate 13-acetate (PMA, Sigma-Aldrich) in DMSO. Gene expression analysis by quantitative RT-PCR for the megakaryocyte marker genes *PPBP* (forward: 5'-GAACTCCGCTGCATGTGTATAA-3'; reverse: 5'-GCAATGGGTTCCTTTCCCGAT-3'), *SELP* (forward: 5'-ATGGGTGGGAACCAAAAAGG-3', reverse: 5'-GGCTGACGGACTCTTGATGTAT-3'), and *PF4* (forward: 5'-ACAGCCGGGAATAAAACGTG-3', reverse: 5'-CTCAGTGCGATGGGAAACTC-3') was performed after 72 hours of differentiation. Flow cytometry for CD41a expression was also performed after 72 hours of differentiation. Erythroid differentiation was induced by 50 uM hemin (Sigma-Aldrich). Gene expression analysis by RT-qPCR of the erythroid markers *ALAS2* (forward: 5'-ACCTACCGTGTGTTCAAGACT-3', reverse: 5'-AGATGCCTCAGAGAAATGTTGG-3'), *RHCE* (forward: 5'-CATCTCCGTCATGCACTCCAT-3', reverse: 5'-TGAGTTCCCCAATGCTGAGGA-3'), and *KEL* (forward: 5'-CGGGTGCTGACAGCTATCC-3', reverse: 5'-ACACACAGATGTCTCACAGGG-3') was performed after 24 hours of differentiation. Flow cytometry analysis of CD235a expression was done after 96 hours of differentiation. **(C)** Gene expression analysis by quantitative RT-PCR in non-induced control clones and clones harboring the GFI1B SNV showing similar baseline levels for megakaryocytic and erythroid markers.

Figure S10

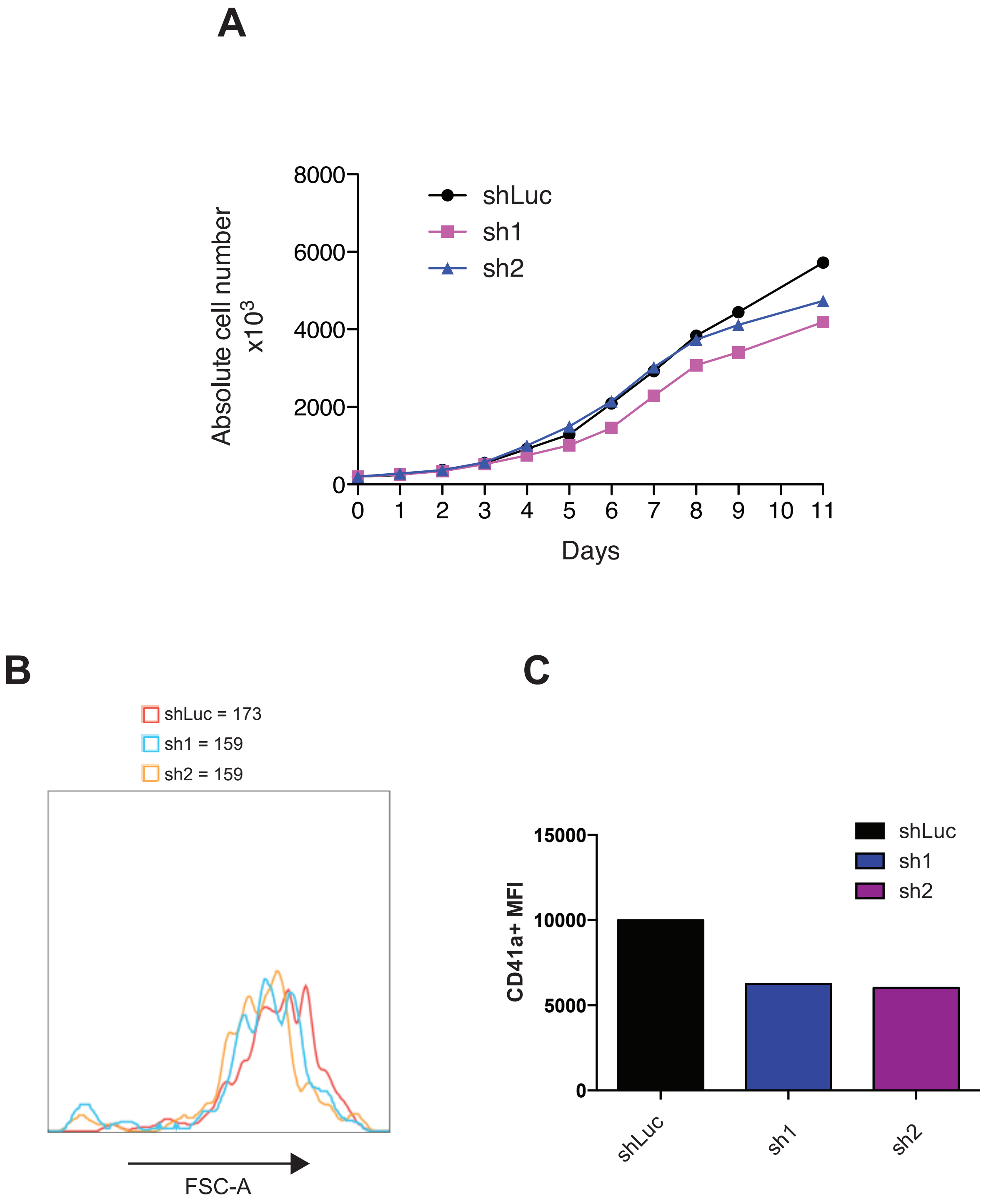

A
Absolute cell number x10^3
8000
6000
4000
2000
0
0
1
2
3
4
5
6
7
8
9
10
11
Days
shLuc
sh1
sh2
B
shLuc = 173
sh1 = 159
sh2 = 159
FSC-A
C
CD41a+ MFI
15000
10000
5000
0
shLuc
sh1
sh2
shLuc
sh1
sh2

Figure S10. Knockdown of the long *GFI1B* isoform does not affect cell growth. **(A)** Cell growth appears comparable between human primary cells with knockdown of the long *GFI1B* isoform and control cells monitored for 11 days of differentiation. Absolute cell counts during the differentiation process are shown on the y-axis (n = 3 per group). **(B)** Representative flow cytometry forward scatter histogram plots (measuring cell size) of cultured primary CD41a+ cells from shLuc control cells and cells with GFI1B knockdown on day 7 of differentiation post expansion. The mean forward scatter intensity is shown. The shRNA constructs targeting human *GFI1B* exon 5 were cloned into the pLKO.1 vector using standard cloning procedures. In brief, hairpin oligonucleotides (Invitrogen) of sh1 (5’-CCGGCATGTGCGACGCTCCCATAGTCTCGAGACTATGGGAGCGTCGCACATGTTTTTG-3’) and sh2 (5’-CCGGGCCTGTGACATCTGCGGCAAACTCGAGTTTGCCGCAGATGTCACAGGCTTTTTG-3’) were cloned into the pLKO.1 vector using the AgeI and EcoRI restriction sites. As a control, the lentiviral vector pLKO-shLuc (Sigma-Aldrich) was used. Primary human CD34+ adult HSPCs (from mobilized peripheral blood derived from G-CSF treated donors) were cultured in StemSpan SFEM II medium supplemented by 1X CC100 (containing FLT3 ligand, stem cell factor (SCF), IL-3, and IL-6) for 5 days. Lentiviral transduction on day 2 of culture and subsequent puromycin selection of infected cells was performed as described previously[53]. On day 5, the culture was changed to differentiation medium containing IMDM with 2% human AB plasma, 3% human AB serum, 1% penicillin/streptomycin, 200 ug/mL holo-transferrin, 10 ng/mL SCF, 1 ng/mL IL-3, 3 U/mL erythropoietin (EPO), and 20 ng/mL thrombopoietin (TPO). Flow cytometry analysis for CD41a and 235a was performed on days 7 and 11 of differentiation on a BD Bioscience Canto II as described[53]. **(C)** Mean fluorescence intensity of CD41a+ in primary cells treated with shLuc or GFI1B shRNAs on day 11 of differentiation corresponding to Figure 3C.

Supplemental Tables

(Excel file)

Supplemental References

1. Ganesh, S.K., Zakai, N.A., van Rooij, F.J., 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. Nature genetics 41, 1191-1198.
2. 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 genetics 7, e1002108.
3. Auer, P.L., Teumer, A., Schick, U., O'Shaughnessy, A., Lo, K.S., Chami, N., Carlson, C., de Denus, S., Dube, M.P., Haessler, J., et al. (2014). Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. Nature genetics 46, 629-634.
4. (1989). The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. American journal of epidemiology 129, 687-702.
5. Fried, L.P., Borhani, N.O., Enright, P., Furberg, C.D., Gardin, J.M., Kronmal, R.A., Kuller, L.H., Manolio, T.A., Mittelmark, M.B., Newman, A., et al. (1991). The Cardiovascular Health Study: design and rationale. Ann Epidemiol 1, 263-276.
6. 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.
7. Kannel, W.B., Dawber, T.R., Kagan, A., Revotskie, N., and Stokes, J., 3rd. (1961). Factors of risk in the development of coronary heart disease--six year follow-up experience. The Framingham Study. Annals of internal medicine 55, 33-50.
8. Sempos, C.T., Bild, D.E., and Manolio, T.A. (1999). Overview of the Jackson Heart Study: a study of cardiovascular diseases in African American men and women. The American journal of the medical sciences 317, 142-146.
9. Hofman, A., Breteler, M.M., van Duijn, C.M., Janssen, H.L., Krestin, G.P., Kuipers, E.J., Stricker, B.H., Tiemeier, H., Uitterlinden, A.G., Vingerling, J.R., et al. (2009). The Rotterdam Study: 2010 objectives and design update. European journal of epidemiology 24, 553-572.
10. Westra, H.J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. Nature genetics 45, 1238-1243.
11. Hofman, A., Grobbee, D.E., de Jong, P.T., and van den Ouweland, F.A. (1991). Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. European journal of epidemiology 7, 403-422.
12. (1998). Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. Controlled clinical trials 19, 61-109.
13. Ritenbaugh, C., Patterson, R.E., Chlebowski, R.T., Caan, B., Fels-Tinker, L., Howard, B., and Ockene, J. (2003). The Women's Health Initiative Dietary Modification trial: overview and baseline characteristics of participants. Ann Epidemiol 13, S87-97.
14. Stefanick, M.L., Cochrane, B.B., Hsia, J., Barad, D.H., Liu, J.H., and Johnson, S.R. (2003). The Women's Health Initiative postmenopausal hormone trials: overview and baseline characteristics of participants. Ann Epidemiol 13, S78-86.
15. Jackson, R.D., LaCroix, A.Z., Cauley, J.A., and McGowan, J. (2003). The Women's Health Initiative calcium-vitamin D trial: overview and baseline characteristics of participants. Ann Epidemiol 13, S98-106.
16. Langer, R.D., White, E., Lewis, C.E., Kotchen, J.M., Hendrix, S.L., and Trevisan, M. (2003). The Women's Health Initiative Observational Study: baseline characteristics of participants and reliability of baseline measures. Ann Epidemiol 13, S107-121.

17. Hays, J., Hunt, J.R., Hubbell, F.A., Anderson, G.L., Limacher, M., Allen, C., and Rossouw, J.E. (2003). The Women's Health Initiative recruitment methods and results. Annals of epidemiology 13, S18-77.
18. Anderson, G.L., Manson, J., Wallace, R., Lund, B., Hall, D., Davis, S., Shumaker, S., Wang, C.Y., Stein, E., and Prentice, R.L. (2003). Implementation of the Women's Health Initiative study design. Ann Epidemiol 13, S5-17.
19. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., et al. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493, 216-220.
20. Consortium, S.T.D., Williams, A.L., Jacobs, S.B., Moreno-Macias, H., Huerta-Chagoya, A., Churchhouse, C., Marquez-Luna, C., Garcia-Ortiz, H., Gomez-Vazquez, M.J., Burtt, N.P., et al. (2014). Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. Nature 506, 97-101.
21. Raitakari, O.T., Juonala, M., Ronnemaa, T., Keltikangas-Jarvinen, L., Rasanen, L., Pietikainen, M., Hutri-Kahonen, N., Taittonen, L., Jokinen, E., Marniemi, J., et al. (2008). Cohort profile: the cardiovascular risk in Young Finns Study. International journal of epidemiology 37, 1220-1226.
22. Nieminen, T., Lehtinen, R., Viik, J., Lehtimaki, T., Niemela, K., Nikus, K., Niemi, M., Kallio, J., Koobi, T., Turjanmaa, V., et al. (2006). The Finnish Cardiovascular Study (FINCAVAS): characterising patients with high risk of cardiovascular morbidity and mortality. BMC cardiovascular disorders 6, 9.
23. de Mutsert, R., den Heijer, M., Rabelink, T.J., Smit, J.W., Romijn, J.A., Jukema, J.W., de Roos, A., Cobbaert, C.M., Kloppenburg, M., le Cessie, S., et al. (2013). The Netherlands Epidemiology of Obesity (NEO) study: study design and data collection. Eur J Epidemiol 28, 513-523.
24. Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.L., Alavere, H., Snieder, H., Perola, M., Ng, P.C., Magi, R., Milani, L., et al. (2015). Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. International journal of epidemiology 44, 1137-1147.
25. Kaplan, J.M., Kim, S.H., North, K.N., Rennke, H., Correia, L.A., Tong, H.Q., Mathis, B.J., Rodriguez-Perez, J.C., Allen, P.G., Beggs, A.H., et al. (2000). Mutations in ACTN4, encoding alpha-actinin-4, cause familial focal segmental glomerulosclerosis. Nature genetics 24, 251-256.
26. Foley, K.S., and Young, P.W. (2013). An analysis of splicing, actin-binding properties, heterodimerization and molecular interactions of the non-muscle alpha-actinins. The Biochemical journal 452, 477-488.
27. Honda, K., Yamada, T., Endo, R., Ino, Y., Gotoh, M., Tsuda, H., Yamada, Y., Chiba, H., and Hirohashi, S. (1998). Actinin-4, a novel actin-bundling protein associated with cell motility and cancer invasion. The Journal of cell biology 140, 1383-1393.
28. Bottega, R., Marconi, C., Faleschini, M., Baj, G., Cagioni, C., Pecci, A., Pippucci, T., Ramenghi, U., Pardini, S., Ngu, L., et al. (2015). ACTN1-related thrombocytopenia: identification of novel families for phenotypic characterization. Blood 125, 869-872.
29. Lerner-Ellis, J.P., Anastasio, N., Liu, J., Coelho, D., Suormala, T., Stucki, M., Loewy, A.D., Gurd, S., Grundberg, E., Morel, C.F., et al. (2009). Spectrum of mutations in MMACHC, allelic expression, and evidence for genotype-phenotype correlations. Human mutation 30, 1072-1081.
30. van Meurs, J.B., Pare, G., Schwartz, S.M., Hazra, A., Tanaka, T., Vermeulen, S.H., Cotlarciuc, I., Yuan, X., Malarstig, A., Bandinelli, S., et al. (2013). Common genetic loci influencing plasma homocysteine concentrations and their effect on risk of coronary artery disease. The American journal of clinical nutrition 98, 668-676.
31. Catellier, D.J., Aleksic, N., Folsom, A.R., and Boerwinkle, E. (2008). Atherosclerosis Risk in Communities (ARIC) Carotid MRI flow cytometry study of monocyte and platelet markers: intraindividual variability and reliability. Clin Chem 54, 1363-1371.
32. Nettleton, J.A., Matijevic, N., Follis, J.L., Folsom, A.R., and Boerwinkle, E. (2010). Associations between dietary patterns and flow cytometry-measured biomarkers of inflammation and cellular

activation in the Atherosclerosis Risk in Communities (ARIC) Carotid Artery MRI Study. Atherosclerosis 212, 260-267.

33. Johnson, A.D., Yanek, L.R., Chen, M.H., Faraday, N., Larson, M.G., Tofler, G., Lin, S.J., Kraja, A.T., Province, M.A., Yang, Q., et al. (2010). Genome-wide meta-analyses identifies seven loci associated with platelet aggregation in response to agonists. Nature genetics 42, 608-613.
34. Becker, D.M., Segal, J., Vaidya, D., Yanek, L.R., Herrera-Galeano, J.E., Bray, P.F., Moy, T.F., Becker, L.C., and Faraday, N. (2006). Sex differences in platelet reactivity and response to low-dose aspirin therapy. JAMA 295, 1420-1427.
35. Reid, J.G., Carroll, A., Veeraraghavan, N., Dahdouli, M., Sundquist, A., English, A., Bainbridge, M., White, S., Salerno, W., Buhay, C., et al. (2014). Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. BMC bioinformatics 15, 30.
36. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337, 64-69.
37. Lange, L.A., Hu, Y., Zhang, H., Xue, C., Schmidt, E.M., Tang, Z.Z., Bizon, C., Lange, E.M., Smith, J.D., Turner, E.H., et al. (2014). Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. American journal of human genetics 94, 233-245.
38. Schick, U.M., Auer, P.L., Bis, J.C., Lin, H., Wei, P., Pankratz, N., Lange, L.A., Brody, J., Stitziel, N.O., Kim, D.S., et al. (2015). Association of exome sequences with plasma C-reactive protein levels in >9000 participants. Hum Mol Genet 24, 559-571.
39. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.
40. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.
41. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297-1303.
42. Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18, 1851-1858.
43. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic acids research 29, 308-311.
44. Genomes Project, C., Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.
45. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81, 559-575.
46. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research 38, e164.
47. Li, A.H., Morrison, A.C., Kovar, C., Cupples, L.A., Brody, J.A., Polfus, L.M., Yu, B., Metcalf, G., Muzny, D., Veeraraghavan, N., et al. (2015). Analysis of loss-of-function variants and 20 risk factor phenotypes in 8,554 individuals identifies loci influencing chronic disease. Nature genetics 47, 640-642.
48. 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. American journal of human genetics 89, 82-93.
49. Gupta, R.M., and Musunuru, K. (2014). Expanding the genetic editing tool kit: ZFNs, TALENs, and CRISPR-Cas9. The Journal of clinical investigation 124, 4154-4161.

50. Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat Protoc 8, 2281-2308.

51. Huang, R., Zhao, L., Chen, H., Yin, R.H., Li, C.Y., Zhan, Y.Q., Zhang, J.H., Ge, C.H., Yu, M., and Yang, X.M. (2014). Megakaryocytic differentiation of K562 cells induced by PMA reduced the activity of respiratory chain complex IV. PloS one 9, e96246.

52. Addya, S., Keller, M.A., Delgrosso, K., Ponte, C.M., Vadigepalli, R., Gonye, G.E., and Surrey, S. (2004). Erythroid-induced commitment of K562 cells results in clusters of differentially expressed genes enriched for specific transcription regulatory elements. Physiol Genomics 19, 117-130.

53. Ludwig, L.S., Gazda, H.T., Eng, J.C., Eichhorn, S.W., Thiru, P., Ghazvinian, R., George, T.I., Gotlib, J.R., Beggs, A.H., Sieff, C.A., et al. (2014). Altered translation of GATA1 in Diamond-Blackfan anemia. Nature medicine 20, 748-753.

Acknowledgements

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

Exome Sequencing Project:
The authors wish to acknowledge the support of the National Heart, Lung, and Blood Institute (NHLBI) and the contributions of the research institutions, study investigators, field staff and study participants in creating this resource for biomedical research. Funding for GO ESP was provided by NHLBI grants RC2 HL-103010 (HeartGO), RC2 HL-102923 (LungGO) and RC2 HL-102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL-102925 (BroadGO) and RC2 HL-102926 (SeattleGO).

Cardiovascular Health Study:
This CHS research was supported by NHLBI contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086; and NHLBI grants HL080295, HL087652, HL105756, R01HL103612, and R01HL120393 with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through AG023629 from the National Institute on Aging (NIA). A full list of CHS investigators and institutions can be found at http://chs-nhlbi.org/. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR000124, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.
The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health

FHS:
The National Heart, Lung, and Blood Institute's Framingham Heart Study (FHS) is a joint project of the National Institutes of Health and Boston University School of Medicine and was supported by the National Heart, Lung, and Blood Institute's Framingham Heart Study (contract No. N01-HC-25195) and its contract with Affymetrix, Inc. for genotyping services (contract No. N02-HL-6-4278). Platelet aggregation data was collected via R01-HL-48157. NHLBI Division of Intramural Research funding supported analysis of data. Analyses reflect the efforts and

resource development from the Framingham Heart Study investigators participating in the SNP Health Association Resource (SHARe) project. A portion of this research was conducted using the Linux Cluster for Genetic Analysis (LinGAII) funded by the Robert Dawson Evans Endowment of the Department of Medicine at Boston University School of Medicine and Boston Medical Center.

JHS:
The Jackson Heart Study is supported by contracts HHSN268201300046C, HHSN268201300047C, HHSN268201300048C, HHSN268201300049C, HHSN268201300050C from the National Heart, Lung, and Blood Institute and the National Institute on Minority Health and Health Disparities. We thank the JHS participants and data collections staff.

RS I and RS II:
The GWA database of the Rotterdam Study was funded through the Netherlands Organisation of Scientific Research NWO (nr. 175.010.2005.011). The Rotterdam Study is supported by the Erasmus Medical Center and Erasmus University, Rotterdam; the Netherlands Organization for Scientific Research (NWO), the Netherlands Organization for Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam.

HeartGO:
Atherosclerosis Risk in Communities (ARIC): NHLBI (N01 HC-55015, N01 HC-55016, N01HC-55017, N01 HC-55018, N01 HC-55019, N01 HC-55020, N01 HC-55021); Cardiovascular Health Study (CHS): NHLBI (N01-HC-85239, N01-HC-85079 through N01-HC-85086, N01-HC-35129, N01 HC-15103, N01 HC-55222, N01-HC-75150, N01-HC-45133, and grant HL080295), with additional support from NINDS and from NIA (AG-023629, AG-15928, AG-20098, and AG-027058); Coronary Artery Risk Development in Young Adults (CARDIA): NHLBI (N01-HC95095 & N01-HC48047, N01-HC48048, N01-HC48049, and N01-HC48050); Framingham Heart Study (FHS): NHLBI (N01-HC-25195 and grant R01 NS17950) with additional support from NIA (AG08122 and AG033193); Jackson Heart Study (JHS): NHLBI and the National Institute on Minority Health and Health Disparities (N01 HC-95170, N01 HC-95171 and N01 HC-95172); Multi-Ethnic Study of Atherosclerosis (MESA): NHLBI (N01-HC-95159 through N01-HC-95169 and RR-024156).

Lung GO:
Cystic Fibrosis (CF): Cystic Fibrosis Foundation (GIBSON07K0, KNOWLE00A0, OBSERV04K0, RDP R026), the NHLBI (R01 HL-068890, R02 HL-095396), NIH National Center for Research Resources (UL1 RR-025014), and the National Human Genome Research Institute (NHGRI) (5R00 HG-004316). Chronic Obstructive Pulmonary Disease (COPDGene): NHLBI (U01 HL-089897, U01 HL-089856), and the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, Novartis, Pfizer, and Sunovian. The COPDGene clinical centers and investigators are available at www.copdgene.org. Acute Lung Injury (ALI): NHLBI (RC2 HL-101779). Lung Health Study (LHS): NHLBI (RC2 HL-066583), the NHGRI (HG-004738), and the NHLBI Division of Lung

Diseases (HR-46002). Pulmonary Arterial Hypertension (PAH): NIH (P50 HL-084946, K23 AR-52742), and the NHLBI (F32 HL-083714). Asthma: NHLBI (RC2 HL-101651), and the NIH (HL-077916, HL-69197, HL-76285, M01 RR-07122).

SWISS and ISGS:
Siblings with Ischemic Stroke Study (SWISS): National Institute of Neurological Disorders and Stroke (NINDS) (R01 NS039987); Ischemic Stroke Genetics Study (ISGS): NINDS (R01 NS042733)

WHISP:
Women's Health Initiative (WHI): The WHI Sequencing Project is funded by the NHLBI (HL-102924) as well as the National Institutes of Health (NIH), U.S. Department of Health and Human Services through contracts N01WH22110, 24152, 32100-2, 32105-6, 32108-9, 32111-13, 32115, 32118-32119, 32122, 42107-26, 42129-32, and 44221. The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at: http://www.whiscience.org/publications/WHI_investigators_shortlist.pdf

GeneSTAR:

GeneSTAR was supported by grants from the NIH/NHLBI (HL72518 and HL087698) and by a grant from the NIH/NCRR (M01-RR000052) to the Johns Hopkins General Clinical Research Center.

NEO:
The authors of the NEO study thank all individuals who participated in the Netherlands Epidemiology in Obesity study, all participating general practitioners for inviting eligible participants and all research nurses for collection of the data. We thank the NEO study group, Pat van Beelen, Petra Noordijk, and Ingeborg de Jonge for the coordination, lab and data management of the NEO study. The genotyping in the NEO study was supported by the Centre National de Génotypage (Paris, France), headed by Jean-Francois Deleuze. The NEO study is supported by the participating Departments, the Division and the Board of Directors of the Leiden University Medical Center, and by the Leiden University, Research Profile Area Vascular and Regenerative Medicine. Dennis Mook-Kanamori is supported by Dutch Science Organization (ZonMW-VENI Grant 916.14.023).