Human Molecular Genetics, 2021, Vol. 00, No. 00 1–15

https://doi.org/10.1093/hmg/ddab252
Advance Access Publication Date: 6 September 2021
General Article

GENERAL ARTICLE

Whole genome sequence analysis of platelet traits in the NHLBI Trans-Omics for Precision Medicine (TOPMed) initiative

Amarise Little[1], Yao Hu[2], Quan Sun[3], Deepti Jain[1], Jai Broome[1], Ming-Huei Chen[4,5], Florian Thibord[4,5], Caitlin McHugh[1], Praveen Surendran[6,7,8,9], Thomas W. Blackwell[10], Jennifer A. Brody[11], Arunoday Bhan[12], Nathalie Chami[13], Paul S. de Vries[14], Lynette Ekunwe[15], Nancy Heard-Costa[16,5], Brian D. Hobbs[17], Ani Manichaikul[18], Jee-Young Moon[19], Michael H. Preuss[13], Kathleen Ryan[20], Zhe Wang[13], Marsha Wheeler[21], Lisa R. Yanek[22], Goncalo R. Abecasis[10], Laura Almasy[23,24], Terri H. Beaty[25], Lewis C. Becker[26], John Blangero[27], Eric Boerwinkle[14], Adam S. Butterworth[6,7,8,28,29], Hélène Choquet[30], Adolfo Correa[15], Joanne E. Curran[27], Nauder Faraday[31], Myriam Fornage[32], David C. Glahn[33], Lifang Hou[34], Eric Jorgenson[30], Charles Kooperberg[2], Joshua P. Lewis[20], Donald M. Lloyd-Jones[34], Ruth J.F. Loos[13], Yuan-I Min[15], Braxton D. Mitchell[20], Alanna C. Morrison[14], Deborah A. Nickerson[21], Kari E. North[35], Jeffrey R. O'Connell[20], Nathan Pankratz[36], Bruce M. Psaty[11,37,38], Ramachandran S. Vasan[5,39,40], Stephen S. Rich[18], Jerome I. Rotter[41], Albert V. Smith[10], Nicholas L. Smith[37,38,42], Hua Tang[43], Russell P. Tracy[44], Matthew P. Conomos[1], Cecelia A. Laurie[1], Rasika A. Mathias[45], Yun Li[46], Paul L. Auer[47], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Timothy Thornton[1,†], Alexander P. Reiner[37,†], Andrew D. Johnson[4,5,†] and Laura M. Raffield[48,†,*]

[1]Department of Biostatistics, University of Washington, Seattle, WA 98105, USA, [2]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA, [3]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA, [4]Population Sciences Branch, Division of Intramural Research, National Heart, Lung and Blood Institute, Bethesda, MD 20892, USA, [5]National Heart Lung and Blood Institute's and Boston University's Framingham Heart Study, Framingham, MA 01702, USA, [6]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care,

[†]Contributed equally to this work.

Received: June 29, 2021. **Revised:** August 23, 2021. **Accepted:** August 24, 2021

University of Cambridge, Cambridge CB1 8RN, UK, [7]British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge CB1 8RN, UK, [8]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge CB1 8RN, UK, [9]Rutherford Fund Fellow, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK, [10]TOPMed Informatics Research Center, University of Michigan, Department of Biostatistics, Ann Arbor, MI 48109, USA, [11]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA 98101, USA, [12]Boston Children's Hospital, Boston, MA 02644, USA, [13]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York City, NY 10029, USA, [14]Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, Human Genetics Center, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA, [15]Department of Medicine, University of Mississippi Medical Center, Jackson, MS 39216, USA, [16]Department of Neurology, Boston University School of Medicine, Boston, MA 02118, USA, [17]Channing Division of Network Medicine and Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA, [18]Department of Public Health Sciences, Center for Public Health Genomics, University of Virginia School of Medicine, Charlottesville, VA 22908, USA, [19]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY 10461, USA, [20]Department of Medicine, Division of Endocrinology, Diabetes and Nutrition, University of Maryland School of Medicine, Baltimore, MD 21201, USA, [21]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA, [22]Division of General Internal Medicine, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA, [23]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA, [24]Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA, [25]School of Public Health, Baltimore, MD 21205, USA, [26]Division of Cardiology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA, [27]Department of Human Genetics and South Texas Diabetes and Obesity Institute, School of Medicine, University of Texas Rio Grande Valley, Brownsville, TX 78520, USA, [28]National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Genomics, University of Cambridge, Cambridge CB1 8RN, UK, [29]National Institute for Health Research Cambridge Biomedical Research Centre, University of Cambridge and Cambridge University Hospitals, Cambridge CB1 8RN, UK, [30]Division of Research, Kaiser Permanente Northern California, Oakland, CA 94612, USA, [31]Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA, [32]University of Texas Health Science Center at Houston, Houston, TX 77030, USA, [33]Department of Psychiatry, Boston Children's Hospital and Harvard Medical School, Boston, MA 02115, USA, [34]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA, [35]Department of Epidemiology, University of North Carolina, Chapel Hill, NC 27599, USA, [36]Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, Minneapolis, MN 55455, USA, [37]Department of Epidemiology, University of Washington, Seattle, WA 98195, USA, [38]Kaiser Permanente Washington Health Research Institute, Kaiser Permanente Washington, Seattle WA 98101, USA, [39]Departments of Cardiology and Preventive Medicine, Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA, [40]Department of Epidemiology, Boston University School of Public Health, Boston, MA 02118, USA, [41]Department of Pediatrics, The Institute for Translational Genomics and Population Sciences, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA 90502, USA, [42]Department of Veterans Affairs Office of Research and Development, Seattle Epidemiologic Research and Information Center, Seattle, WA 98108, USA, [43]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA, [44]Department of Pathology and Laboratory Medicine and Biochemistry, University of Vermont Larner College of Medicine, Colchester, VT 05446, USA, [45]Division of Allergy and Clinical Immunology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA, [46]Departments of Biostatistics, Genetics, Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA, [47]Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI 53201, USA and [48]Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA

*To whom correspondence should be addressed. Tel: +1 9199667255; Fax: +1 9198434682; Email: laura_raffield@unc.edu

Abstract

Platelets play a key role in thrombosis and hemostasis. Platelet count (PLT) and mean platelet volume (MPV) are highly heritable quantitative traits, with hundreds of genetic signals previously identified, mostly in European ancestry populations. We here utilize whole genome sequencing (WGS) from NHLBI's Trans-Omics for Precision Medicine initiative (TOPMed) in a large multi-ethnic sample to further explore common and rare variation contributing to PLT ($n = 61\,200$) and MPV ($n = 23\,485$). We identified and replicated secondary signals at *MPL* (rs532784633) and *PECAM1* (rs73345162), both more common in African ancestry populations. We also observed rare variation in Mendelian platelet-related disorder genes influencing variation in platelet traits in TOPMed cohorts (not enriched for blood disorders). For example, association of *GP9* with lower PLT and higher MPV was partly driven by a pathogenic Bernard-Soulier syndrome variant (rs5030764, p.Asn61Ser), and the signals at *TUBB1* and *CD36* were partly driven by loss of function variants not annotated as pathogenic in ClinVar (rs199948010 and rs571975065). However, residual signal remained for these gene-based signals after adjusting for lead variants, suggesting that additional variants in Mendelian genes with impacts in general population cohorts remain to be identified. Gene-based signals were also identified at several genome-wide association study identified loci for genes not annotated for Mendelian platelet disorders (*PTPRH*, *TET2*, *CHEK2*), with somatic variation driving the result at *TET2*. These results highlight the value of WGS in populations of diverse genetic ancestry to identify novel regulatory and coding signals, even for well-studied traits like platelet traits.

Introduction

Platelets play a critical role in thrombosis and hemostasis, and anti-platelet agents are used for secondary prevention of cardiovascular disease events, preventing thrombosis and inflammation which can lead to further acute events (1). Platelet count (PLT) and mean platelet volume (MPV) are commonly measured platelet quantitative traits used in clinical diagnosis. Very high PLT can indicate thrombocytosis, a type of myeloproliferative neoplasm (MPN) caused primarily by somatic variation in *JAK2, CALR* and *MPL*, and can lead to thrombosis/clotting (2). Very low PLT, known as thrombocytopenia, can instead lead to bleeding, for example during surgery (3). However, there is also extensive variation in PLT and MPV in healthy individuals, and this variability has been epidemiologically associated with chronic disease outcomes (such as associations of higher MPV with higher risk of myocardial infarction (4) and diabetes (5)).

Like other clinical hematologic laboratory measures, PLT and MPV are highly heritable, with hundreds of genetic signals identified in recent genome-wide association studies (GWAS), mostly in European populations (6,7). Better knowledge of the genetic determinants of platelet measures is important for informing diagnosis and understanding penetrance of inherited bleeding disorders, elucidating novel mechanisms of platelet production and clearance, and understanding of platelet biology and its connection with clinical disease endpoints (8). Along with common, noncoding variants most often identified in GWAS, analyses of exome arrays and exome sequencing have pointed to the role of lower frequency coding variants in genes like *GFI1B, CD36*, *IQGAP2, PLG* and *TUBB1,* or the African-specific *MPL* coding variant rs17292650 (p.Lys39Asn), in explaining platelet phenotypic variation (7,9,10). Mutations in many of the same genes identified by GWAS and exome analyses are known to cause Mendelian platelet disorders, often with a recessive mode of inheritance; the impact of heterozygote carrier status for such Mendelian disease variants, or of other coding variation in these genes, in general population cohorts is unclear. Studies of PLT and MPV in a significant number of individuals without known hematological disorders, or other disease case selection, have not been conducted using whole genome sequencing (WGS) data, though some analysis of WGS in a small number of individuals ($n < 4000$) in combination with densely imputed cohorts with genotype array data only has been conducted (11). The first study of WGS with platelet aggregation traits in a general population was only recently conducted (12). A WGS approach allows more complete assessment of rare and ancestry-differentiated variants than imputation-based approaches.

Here we present results from the first WGS-based analysis of PLT and MPV, two platelet quantitative traits which tend to be inversely correlated with one another. These results from a multi-ethnic population help identify new secondary signals at known loci by direct conditional analysis and novel rare variant gene-based signals. Our results demonstrate the potential of WGS, but also point to the need for larger and more diverse sample sizes to accelerate genetic discovery for platelet-related traits.

Results

Analyses of PLT were conducted in 61 200 individuals of diverse ancestral backgrounds from thirteen studies (including $n = 14\,392$ African Americans, $n = 13\,985$ Hispanic/Latino individuals, $n = 32\,129$ European ancestry, $n = 681$ East Asian ancestry, and $n = 13$ from other ancestry groups). MPV analyses were conducted in 23 485 individuals ($n = 7440$ African American, $n = 5466$ Hispanic/Latino, $n = 10\,120$ European ancestry, $n = 447$ East Asian, $n = 12$ other ancestry). Cohort demographics are described in Supplementary Material, Table S1. Average age was 56 years, with 64% women, for PLT, and average age was 57 years, with 61% women, for MPV. For the 23 331 individuals in our sample sets that have both PLT and MPV data available, the Pearson correlation coefficient between PLT and MPV is −0.3294. Distributions for PLT and MPV in each race/ethnicity group are displayed in Supplementary Material, Fig. S1. Compared to European ancestry individuals, and adjusting for age and sex, there is evidence of higher MPV and PLT in African Americans, lower PLT in East Asians, and higher MPV in Hispanics/Latinos and East Asians (Supplementary Material, Table S2). Similar racial/ethnic differences in platelet traits have been previously reported (13)), though the prior literature is inconsistent and sample sizes are generally small (14,15). Moreover, we note that these categories are derived primarily from self-reported and socially constructed race/ethnicity identifiers (with genetic ancestry clustering only used when self-report was not available). These race/ethnicity groupings include individuals with a variety of genetic ancestry backgrounds; this emphasizes the need for further understanding of the actual genetic variants where differential frequencies across genetic ancestry groups might lead to differences in reference ranges or population means for platelet traits.

Single variant results

We first conducted single variant analyses for variants across the autosomes and X chromosome with a minor allele count of at least 10 (Supplementary Material, Fig. S2). We observed little evidence of inflation in this or aggregate analyses (Supplementary Material, Table S3). We identified 44 loci (defined using ±500 kb boundaries from each sentinel variant) for PLT (Supplementary Material, Table S5) and 28 for MPV (Supplementary Material, Table S4) at a genome-wide significance threshold of $P < 1 \times 10^{-9}$. Many loci ($N = 16$) were significantly associated with both PLT and MPV, as expected given the strong inverse correlation between these two phenotypes. Where the lead sentinel variant was shared ($N = 8$ loci), the PLT and MPV effects were in the opposite direction. All 56 loci significantly associated with PLT or MPV in the Trans-Omics for Precision Medicine initiative (TOPMed) were located within 1 Mb (500 kb on either side of sentinel) of at least one variant from the GWAS, Exome Chip, and exome-sequencing literature previously reported to be associated with the respective quantitative trait (6,7,9–11,16–33). To evaluate which populations might be driving association signals in the pooled ancestry analysis, and to more finely examine association signals in European, African, Hispanic/Latino and East Asian ancestry populations with different linkage disequilibrium patterns, we also performed ancestry-stratified single variant analyses on autosomes. Ancestry stratified results are displayed for sentinel lead variants from the pooled ancestry analysis in Supplementary Material, Tables S4 and S5.

Conditional analyses

We next evaluated whether each of the 44 genome-wide significant signals for PLT and 28 for MPV included novel variants, conditionally distinct from variants reported in previous publications for any quantitative platelet trait (PLT, MPV, plateletcrit and platelet distribution width). The latter two measures were previously assessed in (6) but are not available in TOPMed (Supplementary Material, Table S6). For signals that remained significant after conditioning on variants in Supplementary Material, Table S6, we also performed analyses adjusting for variants identified in recent large GWAS from the Blood Cell Consortium (BCX) meta-analyses (Supplementary Material, Table S10) (17,16), published while this paper was in preparation, leaving two signals at a conventional GWAS threshold ($P < 5 \times 10^{-8}$) for PLT and one for MPV (Table 1). These signals can be considered as additional (secondary) independent signals at previously reported genomic loci. Ancestry stratified results for the three lead variants post conditional analysis are also displayed in Supplementary Material, Table S7.

PLT associated single variant signals

The first conditionally distinct signal (Supplementary Material, Table S7) for PLT in the ancestry-combined analysis is an

Table 1. Single variant signals still significant after conditioning on previously known variants (full results in Supplementary Material, Table S7)

rsID	Nearest Gene	Trait	Annotation	P-value	β	P-value, Original	Effect allele frequency, by population					Meta-analysis P-value
							Pooled Analysis	Europeans	African Americans	Hispanic/Latino	East Asians	
rs532784633	*MPL*	PLT	intron	1.01E−08	−20	3.36E−09	0.34%	NA	1.15%	0.26%	NA	4.74E−06
rs78022296	*RCL1*	PLT	intron	2.29E−08	−5.8	2.44E−08	3.77%	1.78%	8.17%	3.63%	7.05%	0.20
rs73345162	*PECAM1*	MPV	intron	2.38E−10	−0.15	4.70E−10	4.91%	0.05%	12.56%	3.87%	NA	1.86E−06

Original P-value is prior to conditional analysis. The meta-analysis P-value is the result of combining variant-level test statistics from various replication cohorts using the sample size-based weighting approach in METAL (full results in Supplementary Material, Table S9). The *RCL1* locus was not considered further due to the high replication P-value ($P = 0.20$). Effect allele frequency is listed as NA if the variant has a minor allele count <10 in that ancestry subset and is thus not available for analysis. PLT, platelet count, MPV, mean platelet volume. NA, not applicable.

intronic variant of *MPL* (rs532784633, $P = 1.01 \times 10^{-8}$ post-conditioning, $\beta = -19.53$, overall effect allele frequency = 0.34%, MAC = 410, more common in African ancestry populations (1.15%, MAC = 332)) associated with lower PLT. In TOPMed, this variant is not significantly associated with MPV ($P = 0.32$). *MPL* encodes the receptor for the hematopoietic growth factor thrombopoietin, which regulates platelet production. Rare loss-of-function (LoF) mutations of *MPL* underlie the autosomal recessive disorder congenital amegakaryocytic thrombocytopenia, whereas 'gain-of-function' germline and somatic *MPL* coding mutations are associated with familial thrombocytosis and predisposition to MPNs [MIM 159530]. The latter category of germline mutations includes a *MPL* coding variant rs17292650 (p.Lys39Asn) associated with higher PLT and common only in African ancestry individuals. The sentinel variant at the *MPL* locus (rs59506047), prior to conditional analysis, is a near perfect LD proxy for rs17292650 ($r^2 = 0.98$ in all TOPMed samples with measured PLT). However, the newly discovered *MPL* intronic rs532784633 variant is associated with PLT independently of the rs17292650 missense variant ($r^2 < 0.01$ with rs532784633 in all TOPMed samples with measured PLT) as well as other recently reported common or intermediate frequency PLT-associated *MPL* coding and non-coding variants in the region (17,16). The intronic *MPL* rs532784633 variant associated with lower PLT is located within an ENCODE cross-tissue enhancer (EH38E1342407) and overlaps a binding site and canonical motif for the transcription factor E2F1, which plays a role in megakaryocyte differentiation and proliferation (34). The second conditionally distinct signal, an intronic variant in *RCL1* (rs78022296), is also more common in African ancestry versus European ancestry individuals (8.2% versus 1.8%) but has little compelling functional evidence linking it to PLT. However, it does lie ~175 kb upstream from *JAK2*, a key blood lineage factor mutated in MPNs, raising the possibility of long-range interactions between these genomic loci.

MPV associated single variant signals

For MPV, we identified an association with an intronic variant in *PECAM1* common only in African ancestry populations (rs73345162, $P = 2.38 \times 10^{-10}$ post-conditioning, $\beta = -0.15$, 12.56% in African ancestry participants, and 0.05% in Europeans). This conditionally distinct variant is also nominally associated with higher PLT ($P = 4.82 \times 10^{-4}$, $\beta = 3.75$). *PECAM1* encodes a cell adhesion molecule expressed on platelets, leukocytes and vascular endothelial cells, and is involved in leukocyte transendothelial migration and regulation of platelet activation and thrombosis (35) as well as megakaryopoiesis (36). Other common non-coding variants 3′ of *PECAM1* (such as rs1050382) have been associated through prior GWAS with both higher MPV (16) and increased risk of coronary heart disease (37), providing a potential genetic link between epidemiologic observations that suggest MPV as a predictor of CVD outcomes (38).

Replication

We pursued replication analyses in independent samples from multi-ethnic GWAS and sequencing studies (no overlap with TOPMed samples), including studies with representation of African American and Hispanic/Latino participants. Cohorts included for replication analyses were (1) participants from the INTERVAL study (WGS data), and imputed genome-wide genotype data from (2) additional, non-overlapping African Americans from the Women's Health Initiative, (3) African American and Hispanic/Latino participants from the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort and (4) African ancestry participants from the UK Biobank). For PLT, the *MPL* signal replicated in the meta-analysis of these four independent cohorts ($P = 4.74 \times 10^{-6}$) but *RCL1* did not replicate ($P = 0.20$) (Supplementary Material, Table S9). For MPV, the *PECAM1* signal also replicated ($P = 1.86 \times 10^{-6}$) (we do note that the variant/chromosomal location is not mapped in build 37, which may explain why this signal was not identified in prior genetic analyses). The replicated *MPL* and *PECAM1* signals are both largely driven by African ancestry participants and are non-significant in individuals of European ancestry in ancestry stratified marginal and conditional analyses (Supplementary Material, Table S7, Supplementary Material, Fig. S3E–H), unsurprisingly based on the allele frequency differences for *MPL* and *PECAM1* lead variants.

Gene-based aggregated rare variant results

Next, we performed gene-based aggregate association analyses for variants with a minor allele frequency <1%. In these aggregate tests, we included only variants which, based on their genomic annotation, are more likely to play a functional role. Criteria for variant inclusion are described in detail in the Methods section. We first assessed LoF, missense and synonymous variants, using three different filters with a decreasing level of stringency for inclusion of protein- or splice-altering variants. Using efficient variant-set mixed model association test (SMMAT) (39) and burden tests, and a Bonferroni corrected significance threshold for the number of genes tested with each filter, we identified 11 unique genes associated with PLT (*ITGA2B, PTPRH, TUBB1, MPL, CD36, TET2, TNFRSF13B, GP9, ITGB3, SH2B3, CHEK2*) and 4 with MPV (*TUBB1, IQGAP2, GFI1B, GP9*), with some genes identified under multiple variant filters (Table 2). Two genes were significantly associated with both traits (*GP9* and *TUBB1*). We also performed gene-based burden and SMMAT tests incorporating non-coding regulatory promoter and enhancer variants (in addition to coding variants), but no additional genes were identified that were not also identified by the coding only variant filters.

Adjusting for all single variants within 500 kb previously reported in the GWAS literature for any PLT trait (Supplementary Material, Tables S6 and S10), the gene-based rare variant association signals at two of the genes (*TNFRSF13B, IQGAP2*) were completely attenuated ($P > 0.05$). All other gene-based rare variant tests retained at least some nominally significant signal; *GFI1B*, *GP9*, *CD36*, *CHEK2*, *ITGA2B*, *ITGB3*, *SH2B3*, *TET2*, *MPL* and *TUBB1* remained genome-wide significant (based on thresholds in Supplementary Material, Table S3) for at least one of the variant filter/association models following adjustment for all previously identified coding and non-coding GWAS single variants in the region. Variant level residuals for rare variant carriers for each of the 15 associations, after adjusting for prior GWAS variants within 500 kb, are presented alongside functional annotation and ClinVar pathogenicity in Supplementary Material, Fig. S4.

Most of the associated genes are noted in the Online Mendelian Inheritance in Man (OMIM) catalog for their association with inherited or acquired quantitative platelet disorders. In each instance, the overall direction of effect for PLT and MPV is consistent with the phenotypic disease characteristics: *ITGA2B*, *ITGB3* and *CD36* with congenital thrombocytopenia (lower PLT), *GP9* and *TUBB1* with congenital macro-thrombocytopenia (lower PLT/higher MPV); *MPL* and *SH2B3* with familial and essential thrombocytosis (higher PLT). Germline *CHEK2* mutations are

Table 2. Lead results for gene-based aggregate tests, for burden and SMMAT tests, with PLT and MPV

Gene_name	OMIM Mendelian platelet disorder gene	Number of included sites	Number of alternate alleles	*P*-value	Trait	Test	Filter	Maximum *P*-value post conditioning
GFI1B	Yes	36	197	4.48E-08	MPV	SMMAT	2	3.02E-07
GP9	Yes	27	171	4.23E-08	MPV	SMMAT	3	6.62E-06
IQGAP2	No	66	298	2.52E-07	MPV	Burden	1	0.24
TUBB1	Yes	59	254	6.21E-15	MPV	SMMAT	2	5.52E-03
CD36	Yes	248	2886	5.66E-08	PLT	SMMAT	2	8.06E-08
CHEK2	No	193	2520	3.06E-07	PLT	Burden	3	3.13E-06
GP9	Yes	42	487	5.96E-17	PLT	SMMAT	3	1.03E-09
ITGA2B	Yes	232	2565	4.11E-10	PLT	SMMAT	3	5.13E-09
ITGB3	Yes	203	1693	5.40E-10	PLT	SMMAT	3	1.10E-06
MPL	Yes	73	414	2.92E-11	PLT	Burden	1	1.61E-11
PTPRH	No	91	1303	2.76E-08	PLT	Burden	1	2.39E-05
SH2B3	Yes	71	585	1.15E-07	PLT	SMMAT	2	2.17E-07
TET2	No	134	138	8.67E-07	PLT	SMMAT	1	8.52E-07
TNFRSF13B	No	89	1560	3.24E-07	PLT	Burden	1	0.17
TUBB1	Yes	107	635	1.95E-24	PLT	SMMAT	2	1.29E-06

Only the most significant test (burden or SMMAT) and variant filter combination are displayed for each gene, full results can be found in Supplementary Material, Table S8. The maximum *P*-value post-conditioning is taken from either the GWAS catalog (Supplementary Material, Table S6) or the BCX (Supplementary Material, Table S10) identified variant conditional analysis.

one of the most frequent causes of hereditary predisposition to cancer (40), which is a common acquired cause of thrombocytosis and therefore consistent with the observed burden test association with higher PLT. Along with information from the functional annotation scoring programs (such as FATHMM, PolyPhen, etc.) used to select variants, we additionally include annotation from ClinVar in Supplementary Material, Table S11 (which lists all variants included in our gene-based tests), highlighting reported Mendelian disease variants included in our TOPMed based aggregate tests (for example, rs5030764, Asn61Ser in *GP9*, which has also been reported to have a heterozygote effect for platelet traits in (17), see Supplementary Material, Table S10).

Somatic mutations of *TET2* are associated with MPNs or myelodysplasia, which often secondarily result in thrombocytosis or thrombocytopenia, respectively (41). Based on this prior literature for *TET2*, we annotated somatic variants using the COSMIC database (https://cancer.sanger.ac.uk/cosmic) for all included variants in our gene-based tests (Supplementary Material, Table S11); for *TET2*, which is a reported driver gene for clonal hematopoiesis of indeterminate potential (CHIP), we also annotated variants included in (42) as CHIP driver variants. We observed that a large percentage of variants in *TET2* were somatic (77.6% [104] of 134 variants, all but one a singleton, in the most significant gene-based test); similar results were not observed for other genes, though multiple likely somatic variants were also observed in *MPL* (14% [10] of 73 included variants, including nominally significant single variant rs141311765 [$P = 2.11E{-}06$]) and *SH2B3* (10% [7] of 71 included variants). We therefore specifically assessed whether the *TET2* gene-based signal for PLT in the bi-directional SMMAT was driven primarily by somatic variants. As seen in Supplementary Material, Fig. S5, most of the variants contributing to our aggregate signal (at extreme ends of the null model residual distribution) at *TET2* were in fact somatic. Conditional analysis of the *TET2* gene-based signal in which we conditioned on variants that were reported somatic attenuated the signal ($P = 0.227$).

The aggregate gene-based results may also aid in identification of likely causal genes at previously identified platelet GWAS loci. In particular, the *PTPRH* gene associated with lower PLT in TOPMed contains a rare LoF variant (rs147881000) recently reported (17) to be strongly associated with lower PLT as well as other quantitative blood cell traits, though some residual signal for this gene remains in TOPMed after adjusting for this variant ($P = 5.14E{-}04$). Though *PTPRH* is not highly expressed in platelets, other protein tyrosine phosphatase receptors are present in platelets/megakaryocytes and are involved in platelet signal transduction and biogenesis (43,44). Specifically, *PTPRJ* (CD148) LoF mutations have been reported in cases of familial autosomal recessive thrombocytopenia (44).

Lastly, for gene-based signals which included coding variants common enough to be tested in our single variant analyses ($MAC > 10$), we assessed whether these variants had a single variant *P*-value for association <0.001, and further adjusted for any such nominally significant single variants in our gene-based tests (Supplementary Material, Table S7). This additional adjustment step was included to evaluate whether gene-based signals were driven in large part by a small set of individual single variants. Some residual signal remained for all tested genes, but some tests were attenuated (for example *GP9*, where for filter 1 and filter 3 burden tests for PLT (the most and least stringent coding variant filters used) *P*-values were attenuated from $P < 1 \times 10^{-14}$ to $P < 1 \times 10^{-9}$ by adjustment for lead single variant rs5030764, p.Asn61Ser, $P = 3.83E{-}09$ in pooled analysis).

Discussion

Our analyses of WGS data from the TOPMed consortium highlight the role of ancestry-specific and rare/low-frequency variants in variability in platelet count and size. Our work has particularly highlighted the influence of both common and Mendelian disease rare variants in the genetic architecture of quantitative platelet traits, in general population cohorts not enriched for blood disorders (for example at *GP9* and *TUBB1* gene-based signals). This approach is complementary to efforts utilizing WGS to screen large numbers of patients with rare diseases and/or those in the tails of the quantitative phenotypic distribution for patient diagnosis and coding and non-coding causal variant discovery (45). We anticipate more information to accrue on the penetrance of high-impact rare variants or clinical importance of ancestry-specific variants on platelet disorders as larger sequencing-based analyses are conducted in diverse populations. Ultimately, we expect that sequencing studies will lead to an increased understanding of the joint importance of common, mostly noncoding variation and rare, high impact variation for platelet trait variance in the population, as demonstrated by identification of multiple Mendelian disease related genes by our rare variant gene-based aggregation tests. As platelets play roles in both bleeding and thrombotic disorders, additional rare variant associations of large effect may inform platelet biology and interpretation of future clinical genetics cases (46). The large impact of genetic variation on platelet traits (with heritability estimated at ∼50–80% (47)) in general population cohorts is often underappreciated, and improved elucidation of the genetic component of variation in platelet traits in individual patients (for example through polygenic scoring) may influence clinical inference and care. This current work in TOPMed makes several important contributions to understanding this joint impact of both common and rare variation on platelet variation, including identification of secondary, conditionally distinct signals at known loci more common in understudied African ancestry populations and identification of multiple novel aggregate rare variant signals, which are distinct from previously identified GWAS variants.

Identification of secondary signals at known loci was facilitated in our analyses by availability of individual level data in a large sample size. In most GWAS studies, summary statistics are contributed by each individual study and then meta-analyzed. Secondary signals are thus usually identified using approximate conditional analysis, with tools like GCTA. Approximate conditional analysis methods are challenging, however, with rare variants and with admixed populations, due to lack of availability of appropriate reference panels. Thus, it would be difficult to find signals such as our novel secondary signal at *MPL*, low-frequency intronic variant rs532784633. The original sentinel variant at *MPL* (rs59506047), prior to conditional analysis, is a linkage disequilibrium proxy for known African-specific coding signal rs17292650, as reflected in the very low allele frequency and non-significant association for this variant in Europeans in our TOPMed data. This secondary signal is also largely African ancestry specific (too rare to be assessed in European and East Asian ancestry subsets, frequency 1.1% African ancestry individuals, $P = 4.50E-09$, frequency 0.3% Hispanic/Latino individuals, $P = 0.18$). This type of secondary signal would be very difficult to observe using summary data, instead of direct conditional analysis using individual level data. Unfortunately, neither our novel *MPL* nor *PECAM1* variant was available in platelet specific eQTL datasets from either GeneSTAR (48) (megakaryocytes and platelets) or CEDAR (49) (platelets only), likely due to low frequency for *MPL* and lack of mapping to build 37 for *PECAM1*; however, given the known role of both *PECAM1* (50) and *MPL* (51) in platelet biology we assume these are the likely target genes.

Our gene-based tests for coding variants highlighted multiple genes known from the Mendelian platelet disorder literature. All identified genes were near a prior GWAS signal, but residual signal which remains after adjusting for these GWAS variants also provides evidence of low frequency coding variant signal independent of common, mostly noncoding variants identified by GWAS. Some of the variants partially driving these gene-based tests have been specifically reported in the clinical literature. For example, the top variant contributing to the *GP9* gene based signal is rs5030764 (chr3:129061921, p.Asn61Ser, $P = 3.83E-09$ in pooled analysis, frequency 0.09%, corresponding to 108 counts of the minor allele in our TOPMed sample, all heterozygotes). The variant is more common in European populations ($P = 7.31E-09$, 0.1% frequency, 96 counts of minor allele). Conditioning on this lead variant, which has also been reported to have a heterozygote effect in prior GWAS from UK Biobank (17), the *GP9* gene-based signal was attenuated, though still significant ($P = 9.57 \times 10^{-10}$ for top filter in Table 2, upon adjustment for rs5030764, $P = 1.03 \times 10^{-9}$ after adjusting for all variants identified in (16,17)). This suggests that multiple additional coding variants in this gene impact platelet traits in the general population. rs5030764 is listed as pathogenic/-likely pathogenic in ClinVar for Bernard-Soulier syndrome, a rare autosomal recessive platelet bleeding disorder. Bernard-Soulier syndrome is caused by a defect in or deficiency of the platelet glycoprotein membrane complex GPIb-IX-V which binds von Willebrand factor (vWF). Inability to bind vWF impairs the clotting process and causes excessive bleeding. Bernard-Soulier syndrome is clinically characterized by low PLT, large platelets, prolonged bleeding time, and abnormal platelet agglutination response to ristocetin. In a 2015 study of 30 carriers of Bernard-Soulier syndrome variants, individuals showed lower PLTs than controls, mild bleeding phenotype and higher vWF levels (52). We also note that some tests are at least partially driven by loss of function variants, including *PTPRH* (rs147881000, $P = 1.61E-04$ in single variant analyses), *TUBB1* (rs199948010, $P = 2.48E-07$, previously reported in BCX) and *CD36* (rs571975065, $P = 4.78E-05$) (Supplementary Material, Table S11). Some additional tests also include variants reported in ClinVar and nominally significant in single variant analyses, including variants of uncertain significance for the platelet disorder Glanzmann thrombasthenia rs142445733 ($P = 0.003$) and rs143967758 ($P = 2.25E-05$) in *ITGA2B*. The lead variant within *ITGB3* (integrin subunit beta 3, also known as platelet membrane glycoprotein IIIa), which was also significantly associated with PLT, was the missense variant rs5917 (chr17:47284587, R143Q), which is annotated as benign in ClinVar. Previous work has linked this polymorphism to the Pena/Penb or Human Platelet Alloantigen (HPA)-4 system (53), which is a cause of platelet alloimmune disorders such as neonatal alloimmune thrombocytopenic purpura and post transfusion purpura, highlighting the plausibility of the aggregate rare variant association for platelet related traits.

Our gene-based results showed broad concordance with previous whole exome sequencing and Exome Chip analyses of rare coding variation, while expanding the number of identified genes for platelet traits. A large Exome Chip paper including >150 000 people (7) identified five gene-based signals (*TUBB1, SH2B3, JAK2, LY75, IQGAP2*) in aggregate tests of >1 coding variant associated with either MPV or PLT; we similarly identify significant signals at three of these genes (*TUBB1, SH2B3, IQGAP2*). Coding single variant signals at *IQGAP2, TUBB1* and *SH2B3* from this Exome Chip paper were adjusted for in conditional analyses, but residual signal remained at all genes (though *IQGAP2* was attenuated to non-significance on adjustment for all previously known single variant associations from GWAS, Exome Chip and exome sequencing based analyses in general population cohorts). This highlights the additional power to detect rare variant signals using an unbiased sequencing method in TOPMed (despite smaller total sample size), as opposed to genotyping of selected variants from a small sequencing reference panel (as was done for Exome Chip). The largest existing platelet trait exome sequencing study in ∼15 000 people (25) identified two single variant signals for low frequency and rare variants (*CPS1* and *GFI1B*); *GFI1B* was also identified in our analyses for MPV, and was robust to adjustment for the lead variant identified from this prior exome sequencing study, rs150813342, as well as other mostly common noncoding variants. We also checked results in ancestry pooled PLT and MPV analyses for previously identified genes *JAK2*, *LY75* and *CPS1*, which did not meet multiple testing corrected significance thresholds in any of our gene-based analyses. All, however, showed nominal significance ($P < 0.05$) for at least one test and filter (*JAK2*, filter 2, burden, $P = 9.56\text{E-}05$, *LY75*, filter 3, burden, $P = 7.23\text{E-}05$, *CPS1*, filter 2, burden, $P = 0.013$, all for PLT). We note that the lead variant from Polfus *et al.* (25) at *CPS1* (rs1047891) was too common to be included in our gene-based analyses (MAF ∼ 32.6% in TOPMed pooled ancestry analyses, $P = 9.08\text{E-}14$ for PLT, see Supplementary Material, Table S5), and that this variant has also been identified in multiple PLT GWAS analyses (6,16).

Some genes (*PTPRH*, *TET2*, *CHEK2*) are in loci which are near GWAS signals, but which have not been previously implicated as associated with platelet traits through coding variant analysis either in general population cohorts or through Mendelian disease genetics (for example *GP9*, as discussed earlier). All are genes which have been associated with cancer risk; for example, *CHEK2* is a Mendelian cancer gene and critical for cell cycle arrest. *TET2* is a known oncogene, an epigenetic and cell differentiation regulator (54), and has a key role in regulation of telomere length. It is also a known driver gene for CHIP. Germline noncoding variants at this locus more common in African ancestry populations have also been associated with CHIP risk (42). CHIP from *TET2* driver variants has been associated with lower PLT as well in previous work from TOPMed ($P = 0.005$); this signal for *TET2* variants is stronger in our analysis, likely due to inclusion of additional somatic variants not annotated as CHIP driver variants (42). The influence of coding variants in these genes on platelet levels is less clear. In megakaryocyte RNA-sequencing data, expression of *CHEK2* ($P = 0.04$) and *TET2* ($P = 0.02$) were associated with measured platelet production (Supplementary Material, Table S12), again suggesting a plausible role in platelet biology for these genes. *PTPRH* was not associated with platelet production ($P = 0.13$).

Our analysis does have limitations. While these analyses were being completed, new larger GWAS analyses (17,16) have been published from the BCX consortium; these variants were included in secondary conditional analyses for significant loci only. Sample overlap between this effort and previous analyses of GWAS, Exome Chip and exome sequencing datasets make TOPMed an inappropriate replication dataset for these prior findings. Larger sample sizes and fuller integration of platelet specific regulatory information are needed to fully explore rare variation in the noncoding space; our joint analyses of coding variation with non-coding variation in enhancers and promoters only identified genes already found in coding variant only analyses and for simplicity are not presented here. Identification of somatic variants driving the observed rare variant association signal in *TET2*, with a smaller percentage of likely somatic variants also contributing to gene-based signals such as *SH2B3* and *MPL*, suggests that somatic variation, not just germline variation, is included in calls from TOPMed WGS (and likely in other sequencing studies for common complex disease). This could be a potential confounder in blood-based sequencing studies for age-related diseases (given the increases in CHIP and other somatic variation with increased age (55)), and suggests examination of somatic versus germline origin for identified variants may be warranted using clinical information, variant characteristics, or repeated sequencing (56) in future sequencing based work. We also note that we do not have consistent data on whether participants in TOPMed cohorts have conditions that cause reactive thrombocytosis (such as cancer or acute infections).

To conclude, our analyses of platelet related traits in TOPMed WGS data has added to our knowledge of ancestry-specific variants and rare variants, notably coding variants, in genetic regulation of platelets. These results highlight the utility of sequencing data in better understanding complex traits, even in relatively modest sample sizes. These results also highlight the importance of larger and more diverse cohorts for the genetic analysis of platelet-related traits, to allow identification of ancestry differentiated variants which may influence baseline levels of these highly heritable platelet traits and inform clinical practice and understanding of platelet biology.

Materials and Methods

Description of TOPMed sequencing, freeze 8

In brief, >30× WGS was completed across >70 cohorts and subcohorts and 140 062 individuals (some from CCDG sequencing initiative, but from TOPMed overlapping cohorts), and then jointly called. Details are described at https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-methods-freeze-8. Freeze 8 TOPMed data (and all positions in this paper) are on build 38. The 13 included cohorts with blood cell traits are further described in the Supplementary Material.

Single variant association analysis methods

Genome-wide single variant association tests for the PLT and MPV traits were performed using linear mixed models (LMM) implemented in GENESIS (57). A 'null model' was fit under the null hypothesis that there is no association between the trait and variant. The null model included fixed effect covariates of age at time of trait measurement, sex, study phase and the first 11 genetic principal components (PCs) estimated using PC-AiR (58). The null model incorporated a fourth degree sparse empirical kinship matrix estimated using PC-Relate (59) to account for genetic relatedness. These PCs and the kinship matrix were generated by the TOPMed Data Coordinating Center (DCC). The null model also allowed for heterogeneous residual variance for groups defined by HARE (60) strata. HARE strata with fewer than 30 subjects were merged with the largest HARE stratum for that study.

To avoid increased false positives and power reduction associated with non-normal trait distributions, the null model was fit using a fully-adjusted two-stage procedure for rank-normalization. In stage 1, a LMM was fit using the trait as the outcome, fixed effect covariates, a sparse kinship matrix and heterogeneous residual variance groups. The resulting marginal residuals were rank-based inverse-normal transformed and rescaled by their original variance. In stage 2, a second LMM was fit using the rescaled marginal residuals as the outcome with the same fixed effect covariates, sparse kinship and heterogeneous residual variance groups as in stage 1.

The model fit in stage 2 was used to perform score tests to interrogate association of each variant on chromosomes 1 through 22 that had minor allele count of at least 10 and passed the IRC quality filters. Genome-wide significance was set at the $P < 1 \times 10^{-9}$ level, based on estimates of the number of independent tests for WGS data in populations with at least some African admixture (61). The X chromosome was analyzed separately using a sex-chromosome specific kinship matrix, as in (62).

Single variant tests for ancestry stratified groups were performed similarly. Self-report race/ethnicity was used to assign individuals to ancestry stratified groups, and then inference using HARE was used when self-report was not available. Self-report outliers were not removed. HARE groups were used to stratify individuals into one of four ancestral groups: East Asian, African American, Hispanic/Latino (including Central American, Cuban, Puerto Rican, Dominican, Mexican and South American HARE strata) and European (including White and Amish HARE strata). HARE strata with five or fewer subjects were excluded.

Gene-based aggregate rare variant

Gene-based aggregate rare variant tests were performed using burden tests (63) and SMMAT (39). The same fully-adjusted two-stage null models as the single variant tests were used. Variant sets consisted of variants with minor allele frequency less than 1%. Flat weights were applied to variants. Genome-wide significance was determined using Bonferroni correction (Supplementary Material, Table S3).

Description of gene-based filters

Three gene-based filters with decreasing levels of stringency for variant inclusion were used. Coding filter 1, the strictest variant set, retained high-confidence LoF variants inferred using LOFTEE (64), missense variants with MetaSVM (65) score > 0 and protein altering or synonymous variants with FATHMM XF coding score (66) >0.5. Coding filter 2 retained high-confidence LoF variants, missense variants which were predicted deleterious by all of the included prediction approaches (SIFT4G (67), Polyphen2_HDIV (68), Polyphen2_HVAR, and LRT (69)), and protein altering or synonymous variants with FATHMM XF coding score > 0.5. Coding filter 3 retained high-confidence LoF variants, missense variants if they are predicted deleterious by any of the prediction algorithms tested (SIFT4G, Polyphen2_HDIV, Polyphen2_HVAR, or LRT_pred), and protein altering or synonymous variants with FATHMM XF coding score > 0.5.

Finally, we tested a fourth filter including both coding and noncoding variants. However, all genes identified were the same ones already identified in the first three coding filters, and these results were not considered further. This combined coding and noncoding variant filter retained high-confidence LoF variants, missense variants with MetaSVM_score > 0, protein altering or synonymous variants with Fathmm-XF score > 0.5, variants overlapping with enhancer(s) linked to a gene using GeneHancer (70), and which have Fathmm-XF score > 0.5 and overlap with regions defined as 'Promoters', 'Promoter flanking regions', 'Enhancers', 'CTCF binding sites', 'Transcription factor binding sites' or 'Open chromatin regions' by Ensembl regulatory build annotation, and variants overlapping with promoters either linked using GeneHancer or 5 Kb upstream of the Transcription start site, and which have Fathmm-XF score > 0.5 and overlap with regions defined as 'Promoters,' 'Promoter flanking regions,' 'Enhancers,' 'CTCF binding sites,' 'Transcription factor binding sites' or 'Open chromatin regions' by Ensembl regulatory build data.

Description of conditional analyses

Conditional analyses for the MPV and PLT traits were performed for genome-wide significant loci from single variant and gene-based aggregate rare variant association tests. Conditional analyses were performed separately for each locus by incorporating variants to condition on as a fixed effect in the null model (see Supplementary Material, Table S6) for previously reported variants for each locus). In some cases, variants that had been reported in previous publications did not pass all sequencing quality controls. We performed conditional analyses with and without these 'fail' variants. In general, fail variants will still capture informative genotypes for most individuals. Conditional single variant tests were performed for all variants with a MAC of at least 10, which passed the sequencing quality filters, and were within 500 kb of each sentinel variant using a score test as previously described with the adjusted null model. Conditional gene-based aggregate rare variant association tests were performed for each aggregate unit using a burden or SMMAT test with the adjusted null model.

We defined a locus as containing a conditionally distinct variant if there were any variants more significant than a Bonferroni corrected significance threshold, adjusting for the number of tested variants within 500 kb of each sentinel variant.

Additional large GWAS analyses were released while this paper was being prepared (17,16). For significant single variant and aggregate tests, we reran conditional analyses adjusting for all conditionally distinct variants identified in these publications (see Supplementary Material, Table S10).

We performed conditional analysis on the *TET2* gene-based SMMAT test under coding filter 1 to test if somatic variants drive the association between *TET2* and PLT. We included as fixed effects variants that were reported somatic and aggregated the remaining rare variants in a testing unit.

Description of replication cohorts

We sought replication of the lead variants at genome-wide significant loci identified in the trait-specific conditional analysis in independent studies including the INTERVAL study (https://www.intervalstudy.org.uk/), the Kaiser-Permanente Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort (71,72) (genotyping data on African American and Hispanic/Latino participants), non-TOPMed samples from the Women's Health Initiative—SNP Health Association Resource (WHI-SHARe) (https://www.whi.org/) (73) and African ancestry samples from phase 2 of UKBB (as defined in (74)). INTERVAL used WGS, while either TOPMed or 1000G phase 3 imputed data was used in all other cohorts. In all cohorts, we adjusted for at a minimum age, sex and cohort specific principal components/genetic relationship matrix, and assessed association of each variant with inverse normalized PLT or MPV values, adjusted for covariates. Results from each study were combined using a sample-size weighted meta-analysis approach in METAL (75). Explained briefly, for each study and for each variant, the direction of effect and *P*-value is converted into a signed *Z*-score. *Z*-scores for each variant are combined across studies as a weighted sum, with weights proportional to the square root of the sample size for each study. The overall Z-score is used to compute a meta-analysis *P*-value; this meta-analysis *P*-value tests the null hypothesis that there is no association between the variant in question and the trait of interest.

Megakaryocyte RNA-seq and platelet production

Immortalized cells were differentiated into megakaryocytic cell line (imMKCL) clones as previously described (76). Phenotypically heterogeneous single-cell subclones were cultured and expanded via doxycycline dependent expression of C-MYC, BMI-1 and BCL-XL. Clones were differentiated to generate platelets (three biological replicates of eight clones), with megakaryocytic

TOPMed accession #	TOPMed project	Parent study name	TOPMed phase	Omics center	Omics support
phs000956	Amish	Amish	1	Broad Genomics	3R01HL121007-01S1
phs001211	AFGen	ARIC AFGen	1	Broad Genomics	3R01HL092577-06S1
phs001211	VTE	ARIC	2	Baylor	3U54HG003273-12S2/HHSN268201500015C
phs001644	AFGen	BioMe AFGen	2.4	MGI	3UM1HG008853-01S2
phs001644	BioMe	BioMe	3	Baylor	HHSN268201600033I
phs001644	BioMe	BioMe	3	MGI	HHSN268201600037I
phs001612	CARDIA	CARDIA	3	Baylor	HHSN268201600033I
phs001368	CHS	CHS	3	Baylor	HHSN268201600033I
phs001368	VTE	CHS VTE	2	Baylor	3U54HG003273-12S2/HHSN268201500015C
phs000951	COPD	COPDGene	1	NWGC	3R01HL089856-08S1
phs000951	COPD	COPDGene	2	Broad Genomics	HHSN268201500014C
phs000951	COPD	COPDGene	2.5	Broad Genomics	HHSN268201500014C
phs000974	AFGen	FHS AFGen	1	Broad Genomics	3R01HL092577-06S1
phs000974	FHS	FHS	1	Broad Genomics	3U54HG003067-12S2
phs001218	AA_CAC	GeneSTAR AA_CAC	2	Broad Genomics	HHSN268201500014C
phs001218	GeneSTAR	GeneSTAR	legacy	Illumina	R01HL112064
phs001218	GeneSTAR	GeneSTAR	2	Psomagen	3R01HL112064-04S1
phs001395	HCHS_SOL	HCHS_SOL	3	Baylor	HHSN268201600033I
phs000964	JHS	JHS	1	NWGC	HHSN268201100037C
phs001416	AA_CAC	MESA AA_CAC	2	Broad Genomics	HHSN268201500014C
phs001416	MESA	MESA	2	Broad Genomics	3U54HG003067-13S1
phs001215	SAFS	SAFS	1	Illumina	3R01HL113323-03S1
phs001215	SAFS	SAFS	legacy	Illumina	R01HL113322
phs001237	WHI	WHI	2	Broad Genomics	HHSN268201500014C

RNA extraction on day 3 of differentiation and platelet production assessed by flow cytometry on day 6. Total RNA was extracted with miRNeasy Mini Kit (Qiagen), libraries prepared with NEB Ultra (PolyA) kits with 50 ng RNA input, and sequenced with 200-cycle paired-end kits on an Illumina HiSeq2500 system. Tuxedo Tools was utilized for read mapping (TopHat v. 2.1.0; Bowtie2 v. 2.2.4). Association of CHEK2, PTPRH and TET2 MK expression with platelet production was assessed by linear mixed effects (LME) models between transcript levels (FPKM) and platelet production (number of platelets/megakaryocyte) for the eight clones with replicate experiments as random effects in the models.

Cohort Acknowledgments

The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

Amish: The TOPMed component of the Amish Research Program was supported by NIH grants R01 HL121007, U01 HL072515, U01 GM074518, and R01 AG18728.

ARIC: The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services (contract numbers HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I and HHSN268201700005I). The authors thank the staff and participants of the ARIC study for their important contributions. The Genome Sequencing Program (GSP) was funded by the National Human Genome Research Institute (NHGRI), the National Heart, Lung, and Blood Institute (NHLBI), and the National Eye Institute (NEI). The GSP Coordinating Center (U24 HG008956) contributed to cross-program scientific initiatives and provided logistical and general study coordination. The Centers for Common Disease Genomics (CCDG) program was supported by NHGRI and NHLBI, and WGS was performed at the Baylor College of Medicine Human Genome Sequencing Center (UM1 HG008898 and R01HL059367).

BioMe: The Mount Sinai BioMe Biobank has been supported by The Andrea and Charles Bronfman Philanthropies and in part by Federal funds from the NHLBI and NHGRI (U01HG00638001; U01HG007417; X01HL134588). We thank all participants in the

Mount Sinai Biobank. We also thank all our recruiters who have assisted and continue to assist in data collection and management and are grateful for the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

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

CHS: Cardiovascular Health Study: This research was supported by contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, 75N92021D00006, and grants U01HL080295 and U01HL130114 from the National Heart, Lung, and Blood Institute (NHLBI), with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

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

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

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

GERA: Genotyping of the GERA cohort was funded by a grant from the National Institute on Aging, National Institute of Mental Health, and National Institute of Health Common Fund (RC2 AG036607).

HCHS/SOL: The Hispanic Community Health Study/Study of Latinos is a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (HHSN268201300001I/N01-HC-65233), University of Miami (HHSN268201300004I/N01-HC-65234), Albert Einstein College of Medicine (HHSN268201300002I/N01-HC-65235), University of Illinois at Chicago—HHSN268201300003I/N 01-HC-65236 Northwestern Univ), and San Diego State University (HHSN268201300005I/N01-HC-65237). The following Institutes/-Centers/Offices have contributed to the HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, NIH Institution-Office of Dietary Supplements.

INTERVAL: Participants in the INTERVAL randomised controlled trial were recruited with the active collaboration of NHS Blood and Transplant England (http://www.nhsbt.nhs.uk), which has supported field work and other elements of the trial. DNA extraction and genotyping were co-funded by the National Institute for Health Research (NIHR), the NIHR BioResource (http://bioresource.nihr.ac.uk) and the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014) [*]. Sequencing was funded by Wellcome Trust grant number 206194. The academic coordinating centre for INTERVAL was supported by core funding from the: NIHR Blood and Transplant Research Unit in Donor Health and Genomics (NIHR BTRU-2014-10024), UK Medical Research Council (MR/L003120/1), British Heart Foundation (SP/09/002; RG/13/13/30194; RG/18/13/33946) and NIHR Cambridge BRC (BRC-1215-20014) [*]. A complete list of the investigators and contributors to the INTERVAL trial is provided in reference (77). The academic coordinating centre would like to thank blood donor centre staff and blood donors for participating in the INTERVAL trial.

This work was supported by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome. *The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

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

MESA: MESA and the MESA SHARe projects are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278. Genotyping was performed at Affymetrix (Santa Clara, California, USA) and the Broad Institute of Harvard and MIT (Boston, Massachusetts, USA) using the Affymetrix Genome-Wide Human SNP Array 6.0. MESA Family is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support is provided by grants and contracts R01HL071051, R01HL071205, R01HL071250, R01HL071251, R01HL071258, R01HL071259, by the National Center for Research Resources, Grant UL1RR033176. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

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

WHI: The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts 75N92021D00001, 75N92021D00002, 75N92021D00003, 75N92021D00004, 75N92021D00005.

Web Resources. Full single variant and aggregate test summary statistics will be provided at the TOPMed genomic summary result dbGaP accession (phs001974).

Supplementary Material

Supplementary material is available at *HMG* online.

Acknowledgements

A complete list of TOPMed consortium authors can be found at https://www.nhlbiwgs.org/topmed-banner-authorship.

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

This research has been conducted using the UK Biobank Resource under Application Number 25953.

Conflicts of Interest statement. A.S.B. has received grants unrelated to this work from AstraZeneca, Biogen, BioMarin, Bioverativ, Novartis and Merck. P.S. is supported by a Rutherford Fund Fellowship from the Medical Research Council grant MR/S003746/2. During the course of the project P.S. also became a full-time employee of GSK. G.R.A. is an employee of Regeneron, Inc., Tarrytown NY.

Funding

T32 HL129982 and KL2TR00249 (to L.M.R.), R01HL129132 (to L.M.R. and Y.L.); American Heart Association (18CDA34110116 to P.S.d.V.); National Eye Institute (NEI) (R01 EY027004 to H.C. and E.J.), the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) (R01 DK116738 to H.C. and E.J.), the National Cancer Institute (NCI) (R01CA2416323 to H.C. and E.J.). NHLBI (U01HL137162 to M.P.C. and D.J.); HHSN268201500014C (N01-HC-25195 to M.-H.C. and A.D.J.) and NHLBI Intramural Research funds; HHSN268201800002I (to T.W.B., A.V.S., and G.R.A.); X01HL134588 (to N.C., M.H.P., Z.W., and R.J.F.L.); the Canadian Institutes of Health Research (CIHR) (to N.C.); P01 HL045522, R01 MH078143, R01 MH078111, R01 MH083824, U01 DK085524, and R01 HL113323 (to L.A., J.B., J.E.C., and D.C.G.); HHSN268201800010I, HHSN268201800011I, HHSN268201800012I, HHSN268201800013I, HHSN268201800014I, and HHSN268201800015I (to L.E., N.M., and A.C.); the National Blood Foundation/American Association of Blood Banks (FP01021164 to A.B.); NIDDK (U54DK110805); and NRSA Joint Program in Transfusion Medicine (T32 4T32HL066987-15); HHSN268201500014C and N01-HC-25197 (to N.H-C.); K08HL-136928 and U01HL089856 (to B.D.H.); R01HL137922 (to J.P.L.); R01 HL121007, U01 HL072515, and R01 AG18728 (to B.D.M.); U01 HL137181 (to J.R.O.); HHSN268201500014C and N01-HC-25199 (to V.R.); R01 HL139553 (to P.S.d.V, A.C.M, N.L.S., and L.M.R.).

References

1. Nagareddy, P. and Smyth, S.S. (2013) Inflammation and thrombosis in cardiovascular disease. *Curr. Opin. Hematol.*, **20**, 457–463.
2. Ashorobi, D. and Gohari, P. (2021) Essential Thrombocytosis. *StatPearls*. StatPearls Publishing Copyright © 2021, StatPearls Publishing LLC, Treasure Island (FL).
3. Jinna, S. and Khandhar, P.B. (2021) Thrombocytopenia. *StatPearls*. StatPearls Publishing Copyright © 2021, StatPearls Publishing LLC, Treasure Island (FL).
4. Chu, S.G., Becker, R.C., Berger, P.B., Bhatt, D.L., Eikelboom, J.W., Konkle, B., Mohler, E.R., Reilly, M.P. and Berger, J.S. (2010) Mean platelet volume as a predictor of cardiovascular risk: a systematic review and meta-analysis. *J. Thromb. Haemost.*, **8**, 148–156.
5. Shah, B., Sha, D., Xie, D., Mohler, E.R., 3rd and Berger, J.S. (2012) The relationship between diabetes, metabolic syndrome, and platelet activity as measured by mean platelet volume: the National Health and Nutrition Examination Survey, 1999–2004. *Diabetes Care*, **35**, 1074–1078.
6. Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A. *et al.* (2016) The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, **167**, 1415–1429 e1419.
7. Eicher, J.D., Chami, N., Kacprowski, T., Nomura, A., Chen, M.H., Yanek, L.R., Tajuddin, S.M., Schick, U.M., Slater, A.J., Pankratz, N. *et al.* (2016) Platelet-related variants identified by Exomechip meta-analysis in 157,293 individuals. *Am. J. Hum. Genet.*, **99**, 40–55.
8. Eicher, J.D., Lettre, G. and Johnson, A.D. (2018) The genetics of platelet count and volume in humans. *Platelets*, **29**, 125–130.
9. Auer, P.L., Johnsen, J.M., Johnson, A.D., Logsdon, B.A., Lange, L.A., Nalls, M.A., Zhang, G., Franceschini, N., Fox, K., Lange, E.M. *et al.* (2012) Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated

loci in African Americans: NHLBI GO exome sequencing project. *Am. J. Hum. Genet.*, **91**, 794–808.

10. Mousas, A., Ntritsos, G., Chen, M.H., Song, C., Huffman, J.E., Tzoulaki, I., Elliott, P., Psaty, B.M., Auer, P.L., Johnson, A.D. *et al.* (2017) Rare coding variants pinpoint genes that control human hematological traits. *PLoS Genet.*, **13**, e1006925.
11. Iotchkova, V., Huang, J., Morris, J.A., Jain, D., Barbieri, C., Walter, K., Min, J.L., Chen, L., Astle, W., Cocca, M. *et al.* (2016) Discovery and refinement of genetic loci associated with cardiometabolic risk using dense imputation maps. *Nat. Genet.*, **48**, 1303–1312.
12. Keramati, A.R., Chen, M.-H., Rodriguez, B.A.T., Yanek, L.R., Gaynor, B.J., Ryan, K., Brody, J.A., Kammers, K., Kanchan, K., Iyer, K. *et al.* (2021) Genome sequencing unveils a new regulatory landscape of platelet reactivity. *Nat Commun.*, **12**, 3626.
13. Segal, J.B. and Moliterno, A.R. (2006) Platelet counts differ by sex, ethnicity, and age in the United States. *Ann. Epidemiol.*, **16**, 123–130.
14. Lim, E., Miyamura, J. and Chen, J.J. (2015) Racial/ethnic-specific reference intervals for common laboratory tests: a comparison among Asians, blacks, Hispanics, and white. *Hawai'i J. Med. Public Health*, **74**, 302–310.
15. Bain, B.J. (1996) Ethnic and sex differences in the total and differential white cell count and platelet count. *J. Clin. Pathol.*, **49**, 664–666.
16. Chen, M.H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D. *et al.* (2020) Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. *Cell*, **182**, 1198–1213.e1114.
17. Vuckovic, D., Bao, E.L., Akbari, P., Lareau, C.A., Mousas, A., Jiang, T., Chen, M.H., Raffield, L.M., Tardaguila, M., Huffman, J.E. *et al.* (2020) The polygenic and monogenic basis of blood traits and diseases. *Cell*, **182**, 1214–1231.e1211.
18. Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., Serbanovic-Canic, J., Elling, U., Goodall, A.H., Labrune, Y. *et al.* (2011) New gene functions in megakaryopoiesis and platelet formation. *Nature*, **480**, 201–208.
19. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K. *et al.* (2018) Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.*, **50**, 390–400.
20. Soranzo, N., Spector, T.D., Mangino, M., Kuhnel, 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.
21. 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. *Nat. Genet.*, **46**, 629–634.
22. Kim, Y.K., Oh, J.H., Kim, Y.J., Hwang, M.Y., Moon, S., Low, S.K., Takahashi, A., Matsuda, K., Kubo, M., Lee, J. *et al.* (2015) Influence of genetic variants in EGF and other genes on Hematological traits in Korean populations by a genome-wide approach. *Biomed. Res. Int.*, **2015**, 914965.
23. Schick, U.M., Jain, D., Hodonsky, C.J., Morrison, J.V., Davis, J.P., Brown, L., Sofer, T., Conomos, M.P., Schurmann, C., McHugh, C.P. *et al.* (2016) Genome-wide association study of platelet count identifies ancestry-specific loci in Hispanic/Latino Americans. *Am. J. Hum. Genet.*, **98**, 229–242.
24. Qayyum, R., Snively, B.M., Ziv, E., Nalls, M.A., Liu, Y., Tang, W., Yanek, L.R., Lange, L., Evans, M.K., Ganesh, S. *et al.* (2012) A meta-analysis and genome-wide association study of platelet count and mean platelet volume in african americans. *PLoS Genet.*, **8**, e1002491.
25. Polfus, L.M., Khajuria, R.K., Schick, U.M., Pankratz, N., Pazoki, R., Brody, J.A., Chen, M.H., Auer, P.L., Floyd, J.S., Huang, J. *et al.* (2016) Whole-exome sequencing identifies loci associated with blood cell traits and reveals a role for alternative GFI1B splice variants in human hematopoiesis. *Am. J. Hum. Genet.*, **99**, 481–488.
26. Li, J., Glessner, J.T., Zhang, H., Hou, C., Wei, Z., Bradfield, J.P., Mentch, F.D., Guo, Y., Kim, C., Xia, Q. *et al.* (2013) GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African-American children. *Hum. Mol. Genet.*, **22**, 1457–1464.
27. Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y., Nakamura, Y. and Kamatani, N. (2010) Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat. Genet.*, **42**, 210–215.
28. Shameer, K., Denny, J.C., Ding, K., Jouni, H., Crosslin, D.R., de Andrade, M., Chute, C.G., Peissig, P., Pacheco, J.A., Li, R. *et al.* (2014) A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Hum. Genet.*, **133**, 95–109.
29. Meisinger, C., Prokisch, H., Gieger, C., Soranzo, N., Mehta, D., Rosskopf, D., Lichtner, P., Klopp, N., Stephens, J., Watkins, N.A. *et al.* (2009) A genome-wide association study identifies three loci associated with mean platelet volume. *Am. J. Hum. Genet.*, **84**, 66–71.
30. Oh, J.H., Kim, Y.K., Moon, S., Kim, Y.J. and Kim, B.J. (2014) Genome-wide association study identifies candidate loci associated with platelet count in Koreans. *Genomics Inform*, **12**, 225–230.
31. Soranzo, N., Rendon, A., Gieger, C., Jones, C.I., Watkins, N.A., Menzel, S., Döring, A., Stephens, J., Prokisch, H., Erber, W. *et al.* (2009) A novel variant on chromosome 7q22.3 associated with mean platelet volume, counts, and function. *Blood*, **113**, 3831–3837.
32. Lo, K.S., Wilson, J.G., Lange, L.A., Folsom, A.R., Galarneau, G., Ganesh, S.K., Grant, S.F., Keating, B.J., McCarroll, S.A., Mohler, E.R., 3rd *et al.* (2011) Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans. *Hum. Genet.*, **129**, 307–317.
33. Ferreira, M.A., Hottenga, J.J., Warrington, N.M., Medland, S.E., Willemsen, G., Lawrence, R.W., Gordon, S., de Geus, E.J., Henders, A.K., Smit, J.H. *et al.* (2009) Sequence variants in three loci influence monocyte counts and erythrocyte volume. *Am. J. Hum. Genet.*, **85**, 745–749.
34. Guy, C.T., Zhou, W., Kaufman, S. and Robinson, M.O. (1996) E2F-1 blocks terminal differentiation and causes proliferation in transgenic megakaryocytes. *Mol. Cell. Biol.*, **16**, 685–693.
35. Coxon, C.H., Geer, M.J. and Senis, Y.A. (2017) ITIM receptors: more than just inhibitors of platelet activation. *Blood*, **129**, 3407–3418.
36. Wu, Y., Welte, T., Michaud, M. and Madri, J.A. (2007) PECAM-1: a multifaceted regulator of megakaryocytopoiesis. *Blood*, **110**, 851–859.
37. Koyama, S., Ito, K., Terao, C., Akiyama, M., Horikoshi, M., Momozawa, Y., Matsunaga, H., Ieki, H., Ozaki, K., Onouchi, Y. *et al.* (2020) Population-specific and trans-ancestry genome-wide analyses identify distinct and shared

genetic risk loci for coronary artery disease. *Nat. Genet.*, **52**, 1169–1177.

38. Pafili, K., Penlioglou, T., Mikhailidis, D.P. and Papanas, N. (2019) Mean platelet volume and coronary artery disease. *Curr. Opin. Cardiol.*, **34**, 390–398.
39. Chen, H., Huffman, J.E., Brody, J.A., Wang, C., Lee, S., Li, Z., Gogarten, S.M., Sofer, T., Bielak, L.F., Bis, J.C. *et al.* (2019) Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *Am. J. Hum. Genet.*, **104**, 260–274.
40. Stolarova, L., Kleiblova, P., Janatova, M., Soukupova, J., Zemankova, P., Macurek, L. and Kleibl, Z. (2020) CHEK2 germline variants in cancer predisposition: stalemate rather than checkmate. *Cells.*, **12**, 2675.
41. Ferrone, C.K., Blydt-Hansen, M. and Rauh, M.J. (2020) Age-associated TET2 mutations: common drivers of myeloid dysfunction, cancer and cardiovascular disease. *Int. J. Mol. Sci.*, **21**, 626.
42. Bick, A.G., Weinstock, J.S., Nandakumar, S.K., Fulco, C.P., Bao, E.L., Zekavat, S.M., Szeto, M.D., Liao, X., Leventhal, M.J., Nasser, J. *et al.* (2020) Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature*, **586**, 763–768.
43. Senis, Y.A., Tomlinson, M.G., Ellison, S., Mazharian, A., Lim, J., Zhao, Y., Kornerup, K.N., Auger, J.M., Thomas, S.G., Dhanjal, T. *et al.* (2009) The tyrosine phosphatase CD148 is an essential positive regulator of platelet activation and thrombosis. *Blood*, **113**, 4942–4954.
44. Marconi, C., Di Buduo, C.A., LeVine, K., Barozzi, S., Faleschini, M., Bozzi, V., Palombo, F., McKinstry, S., Lassandro, G., Giordano, P. *et al.* (2019) Loss-of-function mutations in PTPRJ cause a new form of inherited thrombocytopenia. *Blood*, **133**, 1346–1357.
45. Turro, E., Astle, W.J., Megy, K., Gräf, S., Greene, D., Shamardina, O., Allen, H.L., Sanchis-Juan, A., Frontini, M., Thys, C. *et al.* (2020) Whole-genome sequencing of patients with rare diseases in a national health system. *Nature*, **583**, 96–102.
46. Megy, K., Downes, K., Simeoni, I., Bury, L., Morales, J., Mapeta, R., Bellissimo, D.B., Bray, P.F., Goodeve, A.C., Gresele, P. *et al.* (2019) Curated disease-causing genes for bleeding, thrombotic, and platelet disorders: communication from the SSC of the ISTH. *J. Thromb. Haemost.*, **17**, 1253–1260.
47. Johnson, A.D. (2011) The genetics of common variation affecting platelet development, function and pharmaceutical targeting. *J Thromb Haemost: JTH*, **9**, 246–257.
48. Kammers, K., Taub, M.A., Rodriguez, B., Yanek, L.R., Ruczinski, I., Martin, J., Kanchan, K., Battle, A., Cheng, L., Wang, Z.Z. *et al.* (2021) Transcriptional profile of platelets and iPSC-derived megakaryocytes from whole-genome and RNA sequencing. *Blood*, **137**, 959–968.
49. Momozawa, Y., Dmitrieva, J., Théâtre, E., Deffontaine, V., Rahmouni, S., Charloteaux, B., Crins, F., Docampo, E., Elansary, M., Gori, A.S. *et al.* (2018) IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat. Commun.*, **9**, 2427.
50. Jones, K.L., Hughan, S.C., Dopheide, S.M., Farndale, R.W., Jackson, S.P. and Jackson, D.E. (2001) Platelet endothelial cell adhesion molecule-1 is a negative regulator of platelet-collagen interactions. *Blood*, **98**, 1456–1463.
51. Vainchenker, W., Plo, I., Marty, C., Varghese, L.N. and Constantinescu, S.N. (2019) The role of the thrombopoietin receptor MPL in myeloproliferative neoplasms: recent findings and potential therapeutic applications. *Expert. Rev. Hematol.*, **12**, 437–448.
52. Bragadottir, G., Birgisdottir, E.R., Gudmundsdottir, B.R., Hilmarsdottir, B., Vidarsson, B., Magnusson, M.K., Larsen, O.H., Sorensen, B., Ingerslev, J. and Onundarson, P.T. (2015) Clinical phenotype in heterozygote and biallelic Bernard-Soulier syndrome—a case control study. *Am. J. Hematol.*, **90**, 149–155.
53. Wang, R., Furihata, K., McFarland, J.G., Friedman, K., Aster, R.H. and Newman, P.J. (1992) An amino acid polymorphism within the RGD binding domain of platelet membrane glycoprotein IIIa is responsible for the formation of the Pena/Penb alloantigen system. *J. Clin. Invest.*, **90**, 2038–2043.
54. Ito, S., D'Alessio, A.C., Taranova, O.V., Hong, K., Sowers, L.C. and Zhang, Y. (2010) Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, **466**, 1129–1133.
55. Loh, P.-R., Genovese, G., Handsaker, R.E., Finucane, H.K., Reshef, Y.A., Palamara, P.F., Birmann, B.M., Talkowski, M.E., Bakhoum, S.F., McCarroll, S.A. *et al.* (2018) Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature*, **559**, 350–355.
56. Kraft, I.L. and Godley, L.A. (2020) Identifying potential germline variants from sequencing hematopoietic malignancies. *Blood*, **136**, 2498–2506.
57. Gogarten, S.M., Sofer, T., Chen, H., Yu, C., Brody, J.A., Thornton, T.A., Rice, K.M. and Conomos, M.P. (2019) Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics*, **35**, 5346–5348.
58. Conomos, M.P., Miller, M.B. and Thornton, T.A. (2015) Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet. Epidemiol.*, **39**, 276–293.
59. Conomos, M.P., Reiner, A.P., Weir, B.S. and Thornton, T.A. (2016) Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.*, **98**, 127–148.
60. Fang, H., Hui, Q., Lynch, J., Honerlaw, J., Assimes, T.L., Huang, J., Vujkovic, M., Damrauer, S.M., Pyarajan, S., Gaziano, J.M. *et al.* (2019) Harmonizing genetic ancestry and self-identified race/ethnicity in genome-wide association studies. *Am. J. Hum. Genet.*, **105**, 763–772.
61. Pulit, S.L., de With, S.A. and de Bakker, P.I. (2017) Resetting the bar: statistical significance in whole-genome sequencing-based association studies of global populations. *Genet. Epidemiol.*, **41**, 145–151.
62. Hodonsky, C.J., Jain, D., Schick, U.M., Morrison, J.V., Brown, L., McHugh, C.P., Schurmann, C., Chen, D.D., Liu, Y.M., Auer, P.L. *et al.* (2017) Genome-wide association study of red blood cell traits in Hispanics/Latinos: the Hispanic community health study/study of Latinos. *PLoS Genet.*, **13**, e1006760.
63. Lee, S., Abecasis, G.R., Boehnke, M. and Lin, X. (2014) Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.*, **95**, 5–23.
64. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P. *et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.
65. Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K. and Liu, X. (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.*, **24**, 2125–2137.
66. Rogers, M.F., Shihab, H.A., Mort, M., Cooper, D.N., Gaunt, T.R. and Campbell, C. (2018) FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*, **34**, 511–513.

67. Vaser, R., Adusumalli, S., Leng, S.N., Sikic, M. and Ng, P.C. (2016) SIFT missense predictions for genomes. *Nat. Protoc.*, **11**, 1–9.
68. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
69. Chun, S. and Fay, J.C. (2009) Identification of deleterious mutations within three human genomes. *Genome Res.*, **19**, 1553–1561.
70. Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M. *et al.* (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)*, **2017**, bax028.
71. Kvale, M.N., Hesselson, S., Hoffmann, T.J., Cao, Y., Chan, D., Connell, S., Croen, L.A., Dispensa, B.P., Eshragh, J., Finn, A. *et al.* (2015) Genotyping informatics and quality control for 100,000 subjects in the genetic epidemiology research on adult health and aging (GERA) cohort. *Genetics*, **200**, 1051–1060.
72. Banda, Y., Kvale, M.N., Hoffmann, T.J., Hesselson, S.E., Ranatunga, D., Tang, H., Sabatti, C., Croen, L.A., Dispensa, B.P., Henderson, M. *et al.* (2015) Characterizing race/ethnicity and genetic ancestry for 100,000 subjects in the genetic epidemiology research on adult health and aging (GERA) cohort. *Genetics*, **200**, 1285–1295.
73. Reiner, A.P., Lettre, G., Nalls, M.A., Ganesh, S.K., Mathias, R., Austin, M.A., Dean, E., Arepalli, S., Britton, A., Chen, Z. *et al.* (2011) Genome-wide association study of white blood cell count in 16,388 African Americans: the continental origins and genetic epidemiology network (COGENT). *PLoS Genet.*, **7**, e1002108.
74. Sun, Q., Graff, M., Rowland, B., Wen, J., Huang, L., Lee, M.P., Avery, C.L., Franceschini, N., North, K.E., Li, Y. *et al.* (2021) Analyses of biomarker traits in diverse UK biobank participants identify associations missed by European-centric analysis strategies. *J Hum Genet.*
75. Willer, C.J., Li, Y. and Abecasis, G.R. (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, **26**, 2190–2191.
76. Lee, D.H., Yao, C., Bhan, A., Schlaeger, T., Keefe, J., Rodriguez, B.A.T., Hwang, S.J., Chen, M.H., Levy, D. and Johnson, A.D. (2020) Integrative genomic analysis reveals four protein biomarkers for platelet traits. *Circ. Res.*, **127**, 1182–1194.
77. Di Angelantonio, E., Thompson, S.G., Kaptoge, S.K., Moore, C., Walker, M., Armitage, J., Ouwehand, W.H., Roberts, D.J., Danesh, J., INTERVAL Trial Group. (2017) Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. *Lancet*, **390**, 2360–2371.