# Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations

*A full list of authors and affiliations appears at the end of the paper.*

**Correspondence to:**

Paul L. Auer (pauer@uwm.edu)

Guillaume Lettre (guillaume.lettre@umontreal.ca)

*Summary: (148 words)*

*Main text: (2498 words)*

*4 figures, 1 table*

**SUMMARY** (148 words)

Most loci identified by GWAS have been found in populations of European ancestry (EA). In trans-ethnic meta-analyses for 15 hematological traits in 746,667 participants, including 184,535 non-EA individuals, we identified 5,552 trait-variant associations at $P<5x10^{-9}$, including 71 novel loci not found in EA populations. We also identified novel ancestry-specific variants not found in EA, including an *IL7* missense variant in South Asians associated with lymphocyte count *in vivo* and IL7 secretion levels *in vitro*. Fine-mapping prioritized variants annotated as functional, and generated 95% credible sets that were 30% smaller when using the trans-ethnic as opposed to the EA-only results. We explored the clinical significance and predictive value of trans-ethnic variants in multiple populations, and compared genetic architecture and the impact of natural selection on these blood phenotypes between populations. Altogether, our results for hematological traits highlight the value of a more global representation of populations in genetic studies.

## INTRODUCTION

Blood-cell counts and indices are quantitative clinical laboratory measures that reflect hematopoietic progenitor cell production, hemoglobin synthesis, maturation and release from the bone marrow, and clearance of mature or senescent blood cells from the circulation. Quantitative red blood cell (RBC), white blood cell (WBC) and platelet (PLT) traits exhibit strong heritability ($h^2$~30-80%)(Evans et al., 1999; Hinckley et al., 2013) and have been the subject of various genome-wide association studies (GWAS), including a large study that identified >1000 genomic loci in ~150,000 individuals of European-ancestry (EA)(Astle et al., 2016).

Importantly, the distribution of hematologic traits and prevalence of inherited hematologic conditions differs by ethnicity. For example, the prevalence of anemia and microcytosis is higher among African-ancestry (AFR) individuals compared to EA individuals in part due to the presence of globin gene mutations (e.g. sickle cell, α/β-thalassemia) more common among African, Mediterranean and Asian populations (Beutler and West, 2005; Raffield et al., 2018; Rana et al., 1993). AFR individuals tend to have lower WBC and neutrophil counts partly because of the Duffy/*DARC* null variant (Rappoport et al., 2019). Among Hispanics/Latinos (HA), a common Native American functional intronic variant of *ACTN1* is associated with lower PLT count (Schick et al., 2016).

Despite these observations, non-EA populations have been severely under-represented in most blood-cell genetic studies to date (Popejoy and Fullerton, 2016; Popejoy et al., 2018; Wojcik et al., 2019). Multiethnic GWAS have been recognized as more powerful for gene mapping due to ancestry-specific differences in allele frequency, linkage disequilibrium (LD), and effect size of

causal variants (Li and Keating, 2014). Since blood cells play a key role in pathogen invasion, defense and inflammatory responses, hematologic-associated genetic loci are particularly predisposed to be differentiated across ancestral populations as a result of population history and local evolutionary selective pressures (Ding et al., 2013; Lo et al., 2011; Raj et al., 2013). Given the essential role of blood cells in tissue oxygen delivery, inflammatory responses, atherosclerosis, and thrombosis (Byrnes and Wolberg, 2017; Chu et al., 2010; Colin et al., 2014; Tajuddin et al., 2016), factors that contribute to such inter-population differences in blood-cell traits may also play appreciable roles in the pathogenesis of chronic diseases and health disparities between populations.

## RESULTS

### Trans-ethnic and ancestry-specific blood-cell traits genetic associations

We analyzed genotype-phenotype associations at up to 45 million autosomal variants in 746,667 participants, including 184,424 individuals of non-EA descent, for 15 traits (**Figure 1**, **Supplementary Tables 1-4**, and **Methods**). The association results of the EA-specific meta-analyses are reported separately in a companion paper. In the trans-ethnic meta-analyses, we identified 5,552 trait-variant associations at $P<5\times10^{-9}$, which include 71 novel associations not reported in the EA-specific manuscript (**Supplementary Table 5**). Of the 5,552 trans-ethnic loci, 128 showed strong evidence of allelic effect heterogeneity across populations ($P_{ancestry.hetero}<5\times10^{-9}$) (**Supplementary Figure 1** and **Supplementary Table 5**). Ancestry-specific meta-analyses revealed 28 novel trait-variant associations (**Figure 1** and **Supplementary Tables 6-10**). However, 19 out of these 21 novel AFR-specific associations map to chromosome 1 and are associated with WBC or neutrophil counts, therefore reflecting long-range associations due to the admixture signal at the Duffy/*DARC* locus (Reich et al., 2009). We attempted to replicate all novel trans-ethnic or ancestry-specific genetic associations in the Million Veteran Program (MVP) cohort (Gaziano et al., 2016). Of the 89 variant-trait associations that we could test in MVP, 86 had a consistent direction of effect (binomial $P=6\times10^{-24}$), 72 had an association $P<0.05$ (binomial $P=8\times10^{-79}$), and 44 met the Bonferroni-adjusted significance threshold of $P<6\times10^{-4}$ (**Supplementary Table 11**).

For 3,552 loci with evidence of a single association signal based on conditional analyses in EA (**Supplementary Methods**), we generated fine-mapping results for each trans-ethnic or ancestry-specific dataset using an approximate Bayesian approach (**Methods**)(Wellcome Trust Case

Control et al., 2012). The 95% credible sets were smaller in the trans-ethnic meta-analyses than in the EA or EAS meta-analyses (**Figure 2A**), indicating improved resolution owing to both increased sample size and different LD patterns. When comparing loci discovered in both the trans and EA analyses, we found that the 95% credible sets were 30% smaller among the trans results (median (interquartile range) number of variants per 95% credible set was 4 (2-13) in trans vs. 5 (2-16) in EA, Wilcoxon's $P$=3x10$^{-4}$). For instance, a locus on chromosome 9 associated with PLT counts included seven variants in the EA 95% credible set but only one in the trans set, an increase in fine-mapping resolution likely driven by limited LD at the locus in EAS (**Figure 2B**). In the trans and EA results, respectively, we identified 433 and 403 loci with a single variant in the 95% credible sets (**Figure 2C**).

Next, we assessed our fine-mapped 95% credible sets for the presence of functional variants, which we defined as variants with coding consequences or those mapping to hematopoietic accessible chromatin. Genomic annotation of the 95% credible sets of the trans, EA and EAS hematological trait-associated loci revealed that the proportion of likely functional variants was higher among those with high PPI (**Figure 2D**). The enrichment within high-PPI categories was particularly notable for missense variants, but also observed for intronic and intergenic variants that map to open chromatin regions in precursor or mature blood cells (**Figure 2D**)(Corces et al., 2016). We used g-chromVAR to quantify the enrichment of trans, EA and EAS 95% credible set variants within regions of accessible chromatin identified by ATAC-seq in 18 hematopoietic populations (Corces et al., 2016; Ulirsch et al., 2019). We noted 22 significant trait-cell type enrichments using the trans-ethnic credible sets, all of which were lineage specific, including RBC traits in erythroid progenitors, platelet traits in megakaryocytes, and monocyte count in granulocyte-macrophage

progenitors (GMP) (**Figure 2E** and **Supplementary Table 12**). Cell-type enrichments were largely consistent between fine-mapped traits found in the trans, EA and EAS loci. However, we observed two noteworthy ancestry-specific differences: the EAS results revealed significant enrichments in basophil count for the common myeloid progenitor (CMP) population and eosinophil count for the GMP population, but neither pairing reached significance in the larger EA meta-analyses (**Supplementary Figure 2**). These differences persisted even after controlling for the number of loci tested in each ancestry. This is further supported by our finding that the genetic correlations for these two traits between EA and EAS are the lowest among all studied blood phenotypes (see below).

**Phenome-wide association studies (pheWAS)**

When we queried the 5,552 trans-ethnic genome-wide significant variants associated with blood-cell traits in three distinct biobanks (EA individuals from UK Biobank (UKBB), Japaneses from Biobank Japan (BBJ), African Americans from BioVU (BioVU)), we identified 1,140 phenotype-variant associations (**Supplementary Table 13**). These include 106 variants in BBJ, four variants in BioVU, and 222 variants in the UKBB. Of the four variants significant in BioVU, three were located at the β-globin locus and reflect the known clinical sequelae of sickle cell disease. Of the 1,140 associations, 246 were shared across at least two biobanks (**Methods**). Many of the associations shared between BBJ and the UKBB were related to dyslipidemia and cardiovascular diseases (**Supplementary Table 13**). Because of the large differences in sample sizes between the three biobanks (**Methods**), we reasoned that lack of power was the likely explanation for why associations were not shared across biobanks. However, it appears that differences in allele

frequencies across the three primary ancestries also played a role. Overall, unique associations had greater differences in allele frequencies than shared associations (**Supplementary Figure 3**).

**Trans-ethnic predictions of hematological traits**

Polygenic trait scores (PTS) developed in a single ethnically homogeneous population tend to underperform when tested in a different population (Grinde et al., 2019; Marquez-Luna et al., 2017; Martin et al., 2019). We explored whether we could combine the genome-wide significant trans-ethnic variants identified in our analyses into PTS that can predict blood-cell traits in a multi-ethnic setting. First, we used trans-ethnic effect sizes as weights to compute $PTS_{trans}$ for each trait, and tested their performance in independent EA, AFR and HA participants from the BioMe Biobank (**Methods**). As expected because our trans-ethnic meta-analyses are dominated by EA individuals, $PTS_{trans}$ were more predictive in EA, although their performance in HA was comparable for several traits (lymphocyte and monocyte counts, mean PLT volume)(**Supplementary Figure 4A** and **Supplementary Table 14**). Moreover, for neutrophil and WBC counts, the variance explained by the $PTS_{trans}$ was up to three times higher in AFR and HA than in EA samples due to the inclusion of the strong Duffy/*DARC* locus (**Supplementary Figure 4A**). Because these Duffy/*DARC* variants would not have been included in PTS derived uniquely from EA association results, this illustrates an interesting feature of using trans-ethnic variants for building polygenic predictors. Next, we asked if we could increase the variance explained by calculating PTS using the same trans-ethnic variants but weighting them using ancestry-specific as opposed to trans-ethnic effect sizes. $PTS_{trans}$ outperformed ancestry-specific $PTS_{AFR}$ and $PTS_{HA}$ in BioMe AFR and HA participants, respectively (**Supplementary Figure 4B-C** and **Supplementary Table 14**). This result likely indicates that the discovery sample size for

151 these two populations is still too small to provide robust estimates of the true population-specific
152 effect sizes.
153
154 **Rare coding blood-cell-traits-associated variants**
155 The identification of rare coding variants has successfully pinpointed candidate genes for many
156 complex traits, including blood-cell phenotypes (Auer et al., 2014; Chami et al., 2016; Eicher et
157 al., 2016; Justice et al., 2019; Marouli et al., 2017; Mousas et al., 2017; Tajuddin et al., 2016). Our
158 trans-ethnic and non-EA ancestry-specific meta-analyses yielded 16 coding variants with minor
159 allele frequency (MAF) <1% (**Table 1** and **Supplementary Table 15**). This list includes variants
160 of clinical significance (variants in *TUBB1*, *GFI1B*, *HBB*, *MPL* and *SH2B3*) and variants that
161 nominate candidate genes within GWAS loci (*ABCA7, GMPR*) (**Table 1**). Our analyses also
162 retrieved a known missense variant in *EGLN1* (rs186996510) that is associated with high-altitude
163 adaptation and hemoglobin levels in Tibetans (Lorenzo et al., 2014; Xiang et al., 2013). We noted
164 a missense variant in *IL7* (rs201412253, Val18Ile) associated with increased lymphocyte count in
165 South Asians (SAS)($P$=4.4x10$^{-10}$) (**Figure 3A** and **Supplementary Table 16**). This variant is low-
166 frequency in SAS (MAF=2.6%) but rare in other populations (MAF <0.4%). IL7 encodes
167 interleukin-7, a cytokine essential for B- and T-cell lymphopoiesis (Lin et al., 2017). IL7 is
168 synthesized as a proprotein that is cleaved prior to secretion, and the *IL7*-Val18Ile variant localizes
169 to the IL7 signal peptide comprising the first 25 amino acids. To determine if this variant alters
170 IL7 secretion, we engineered HEK293 cells with either IL7 allele (**Methods**). Although there was
171 no difference in *IL7* RNA expression levels ($t$-test $P$=0.63), we found that the IL7-18Ile allele,
172 which associates with higher lymphocyte counts in SAS individuals, significantly increased IL7
173 protein secretion in this heterologous cellular system (+83%, $P$=2.7x10$^{-5}$)(**Figure 3B**).

Unfortunately, the relatively small sample size of the SAS cohort prevented us from testing whether this *IL7* variant may be associated with other relevant phenotypes, such as cancer risk or susceptibility to infections (Lin et al., 2017).

**Genetic architecture of blood-cell traits in EA and EAS populations**

We used several different approaches to quantify similarities and differences in genetic architecture of hematologic traits across populations. Focusing on the two largest studied populations, EA and EAS, we calculated heritability for all blood traits and found them to be highly concordant between ancestries (Pearson's $r$=0.75, $P$=0.0033)(**Supplementary Table 17**)(Bulik-Sullivan et al., 2015b). Likewise, within-ancestry genetic correlation coefficients ($r_g$) between pairs of hematological traits were highly concordant across ancestries (Pearson's $r$=0.97, $P<2.2\text{x}10^{-16}$)(**Supplementary Figure 5**)(Bulik-Sullivan et al., 2015a). We then used the Popcorn method to directly measure genetic correlations for blood-cell traits between EA and EAS using summary statistics for common variants (Brown et al., 2016). For all 13 traits available in both EA and EAS, genetic correlations were high (lowest for basophils ($r_g$=0.30) and highest for MCH ($r_g$=0.66)), but significantly different than 1 ($P<3\text{x}10^{-6}$)(**Supplementary Table 18**). This suggests that even when considering only common variants, the genetic architecture of blood phenotypes is substantially different between populations.

**Natural selection at blood-cell trait loci**

Natural selection can account for differences in association results between populations, as highlighted by our analyses of rare coding variants which includes several loci known to be under selection (*CD36*, *HBB*, *EGLN1*)(**Table 1**). To further explore this possibility, we assessed whether

197 variants that tag selective sweeps (tagSweeps, variants with the highest integrated haplotype score
198 (iHS)) within continental populations from the 1000 Genomes Project (1000G) are associated with
199 blood-cell phenotypes (Johnson and Voight, 2018). We found a genome-wide enrichment of
200 associations results between tagSweeps and hematological traits, particularly within EA, EAS and
201 AFR populations (**Supplementary Figure 6** and **Supplementary Table 19**). To rule out simple
202 overlaps due to the large number of sweeps and blood-cell trait loci, we compared the number of
203 genome-wide significant tagSweeps in EA, EAS and AFR with the number of significant variants
204 among 100 sets of matched variants (**Methods**). We found significant enrichment of selective
205 sweeps for WBC (EA, EAS, AFR), monocytes (EA, AFR), eosinophils (EA), neutrophils (AFR),
206 lymphocytes (EAS), and PLT (EA, EAS)(**Supplementary Table 20**).
207
208 In AFR and HA, the enrichments for WBC, neutrophils and monocytes were entirely driven by
209 selective sweeps on chromosome 1 near Duffy/*DARC* (Reich et al., 2009). Only three additional
210 loci shared evidence of associations with blood-cell traits and positive selection across
211 populations: HLA, *SH2B3* (Zhernakova et al., 2010) and *CYP3A5* (Chen et al., 2009). We found
212 eight and 100 non-overlapping selective sweeps with variants associated with hematological traits
213 in EAS and EA, respectively (**Supplementary Table 21**). Six of the eight EAS-specific tagSweeps
214 are also associated with blood-cell traits in EA participants, indicating that these regions do not
215 account for population differences in hematological trait regulation (**Supplementary Table 21**).
216 One of the remaining two variants is located at the *HBS1L-MYB* locus and, although it is not
217 associated with blood-cell traits in EA, there are many other variants near *MYB* associated with
218 blood phenotypes in EA (**Supplementary Table 6**). The remaining selective sweep highlighted
219 by this analysis is located upstream of *IL6* (**Figure 4**). The tagSweep at this locus, rs2188580, is

220 strongly associated with PLT count in EAS ($P_{EAS}$=2.8x10$^{-9}$, $P_{EA}$=0.0022), is differentiated between
221 EAS and EA as indicated by the population branch statistic (PBS)(Yi et al., 2010)(C-allele
222 frequency in EAS=44%, 4% in EA; standardized $PBS_{EAS}$=7.353), and overlaps selective sweeps
223 identified in several EAS populations from the 1000G (e.g. $iHS_{CHS}$=3.935)(**Figure 4**). The *IL6*
224 locus has previously been associated with WBC traits in EA (Astle et al., 2016), but our finding is
225 the first report of its association with PLT. *IL6* encodes interleukin-6, a cytokine that is a
226 maturation factor for megakaryocytes, the precursors of PLT (Kimura et al., 1990). Further
227 supporting the role of IL6 signaling in PLT biology, a well-characterized missense variant in the
228 IL6 receptor gene (*IL6R*-rs2228145)(van Dongen et al., 2014) is also associated with PLT count
229 in EAS ($P$=4.3x10$^{-6}$).
230

## DISCUSSION

Our meta-analyses of 15 hematological traits in up to 746,667 individuals represents one of the largest genetic study of clinically relevant complex human traits across diverse ancestral groups. We have continued to expand the repertoire of loci and genes that contribute to interindividual variation in blood-cell traits, with potential implications for hematological diseases, but also other conditions such as cancer, immune and cardiovascular diseases. Our results hold a number of implications for future human genetic studies. First, we showed that adding even a "modest" number of non-EA participants to GWAS can yield important biology, such as the identification of a lymphocyte count-associated *IL7* missense variants in 8,189 South Asians (**Figure 3**). Second, loci that underlie variation in blood-cell traits represent a broad mixture of shared associations (i.e. similar allele frequencies and effect sizes across populations) and heterogeneous associations (i.e. dissimilar allele frequencies and effect sizes across populations). This result contributes to mounting evidence that a full accounting of the genetic basis of complex human traits will require a thorough catalog of global genetic and phenotypic variation. Third, because of heterogeneity across populations in both allele frequencies and patterns of LD, fine-mapping of association signals can be substantially aided by including multiple ancestries. This will have a dramatic impact on the success of large-scale efforts aimed at functionally characterizing GWAS findings. As more studies seek to unravel the causal variants that underlie complex traits associations, we anticipate that genetic evidence from diverse ancestries will play an important role.

## SUPPLEMENTERY INFORMATION

**Supplementary Information** is linked to the online version of the paper.

## ACKNOWLEDGMENTS

We thank all participants, as well as Dr. John D. Rioux for providing the *IL7* ORF. A full list of acknowledgments appears in the **Supplementary Information**. Part of this work was conducted using the UK Biobank resource (Projects number 11707 and 13745).

## AUTHOR CONTRIBUTIONS

*Writing Group (wrote and edited manuscript)*

Ming-Huei Chen, Laura M. Raffield, Abdou Mousas, Erik L. Bao, Alexander P. Reiner, Paul L. Auer, Guillaume Lettre. All authors contributed and discussed the results, and commented on the manuscript.

*Data preparation group (checked and prepared data from contributing cohorts for meta-analyses and replication)*

Ming-Huei Chen, Laura M. Raffield, Abdou Mousas, Jennifer E. Huffman, Guillaume Lettre, Paul L. Auer.

*Meta-analyses (discovery and replication)*

Ming-Huei Chen, Laura M. Raffield, Abdou Mousas, Saori Sakaue, Jennifer E. Huffman, Tao Jiang, Parsa Akbari, Dragana Vuckovic, Erik L. Bao, Arden Moscati, Ken Sin Lo, Caleb A. Lareau,

Michael H. Guo, Tim Kacprowski, Fotis Koskeridis, Ani Manichaikul, Michael Preuss, Cassandra N. Spracklen.

*Fine-mapping and functional annotation*

Abdou Mousas, Ming-Huei Chen, Laura M. Raffield, Erik L. Bao, Caleb A. Lareau, Ken Sin Lo, Vijay G. Sankaran, Paul L. Auer, Guillaume Lettre.

*pheWAS, polygenic prediction, genetic architecture and natural selection*

Saori Sakaue, Arden Moscati, Regina Manansala, Minhui Chen, Ken Sin Lo, Huijun Qian, Yun Li, Charleston W.K. Chiang, Ruth J.F. Loos, Alexander P. Reiner, Guillaume Lettre, Paul L. Auer.

*Functional characterization of IL7*

Mélissa Beaudoin, Véronique Laplante, Guillaume Lettre, Jean-François Gauchat

## AUTHOR INFORMATION

Summary genetic association, fine-mapping and g-chromVAR results are available online: http://www.mhi-humangenetics.org/en/resources. Competing financial interests are declared in the **Supplementary Information**. Correspondence and requests should be addressed to P.L.A (pauer@uwm.edu) or G.L. (guillaume.lettre@umontreal.ca).

## References

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.

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.

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.

Beutler, E., and West, C. (2005). Hematologic differences between African-Americans and whites: the roles of iron deficiency and alpha-thalassemia on hemoglobin levels and mean corpuscular volume. Blood *106*, 740-745.

Brown, B.C., Asian Genetic Epidemiology Network Type 2 Diabetes, C., Ye, C.J., Price, A.L., and Zaitlen, N. (2016). Transethnic Genetic-Correlation Estimates from Summary Statistics. Am J Hum Genet *99*, 76-88.

Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.R., ReproGen, C., Psychiatric Genomics, C., Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control, C., Duncan, L., *et al.* (2015a). An atlas of genetic correlations across human diseases and traits. Nat Genet *47*, 1236-1241.

Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics, C., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M.

(2015b). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet *47*, 291-295.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J.*, et al.* (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203-209.

Byrnes, J.R., and Wolberg, A.S. (2017). Red blood cells in thrombosis. Blood *130*, 1795-1799.

Carroll, R.J., Bastarache, L., and Denny, J.C. (2014). R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics *30*, 2375-2376.

Chami, N., Chen, M.H., Slater, A.J., Eicher, J.D., Evangelou, E., Tajuddin, S.M., Love-Gregory, L., Kacprowski, T., Schick, U.M., Nomura, A.*, et al.* (2016). Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits. Am J Hum Genet *99*, 8-21.

Chen, X., Wang, H., Zhou, G., Zhang, X., Dong, X., Zhi, L., Jin, L., and He, F. (2009). Molecular population genetics of human CYP3A locus: signatures of positive selection and implications for evolutionary environmental medicine. Environ Health Perspect *117*, 1541-1548.

Chen, Y., Fang, F., Hu, Y., Liu, Q., Bu, D., Tan, M., Wu, L., and Zhu, P. (2016). The Polymorphisms in LNK Gene Correlated to the Clinical Type of Myeloproliferative Neoplasms. PLoS One *11*, e0154183.

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.

Colin, Y., Le Van Kim, C., and El Nemer, W. (2014). Red cell adhesion in human diseases. Current opinion in hematology *21*, 186-192.

337 Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M., Koenig, J.L., Snyder, M.P.,
338 Pritchard, J.K., Kundaje, A., Greenleaf, W.J., *et al.* (2016). Lineage-specific and single-cell
339 chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet *48*, 1193-
340 1203.
341 Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y.,
342 Levy, S., McGue, M., *et al.* (2016). Next-generation genotype imputation service and methods.
343 Nat Genet *48*, 1284-1287.
344 Delaneau, O., Zagury, J.F., and Marchini, J. (2013). Improved whole-chromosome phasing for
345 disease and population genetic studies. Nat Methods *10*, 5-6.
346 Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang,
347 D., Masys, D.R., Roden, D.M., and Crawford, D.C. (2010). PheWAS: demonstrating the feasibility
348 of a phenome-wide scan to discover gene-disease associations. Bioinformatics *26*, 1205-1210.
349 Ding, K., de Andrade, M., Manolio, T.A., Crawford, D.C., Rasmussen-Torvik, L.J., Ritchie, M.D.,
350 Denny, J.C., Masys, D.R., Jouni, H., Pachecho, J.A., *et al.* (2013). Genetic variants that confer
351 resistance to malaria are associated with red blood cell traits in African-Americans: an electronic
352 medical record-based genome-wide association study. G3 (Bethesda) *3*, 1061-1068.
353 Eicher, J.D., Chami, N., Kacprowski, T., Nomura, A., Chen, M.H., Yanek, L.R., Tajuddin, S.M.,
354 Schick, U.M., Slater, A.J., Pankratz, N., *et al.* (2016). Platelet-Related Variants Identified by
355 Exomechip Meta-analysis in 157,293 Individuals. Am J Hum Genet *99*, 40-55.
356 Evans, D.M., Frazer, I.H., and Martin, N.G. (1999). Genetic and environmental causes of variation
357 in basal levels of blood cells. Twin Res *2*, 250-257.

358 Fang, H., Hui, Q., Lynch, J., Honerlaw, J., Assimes, T.L., Huang, J., Vujkovic, M., Damrauer,
359 S.M., Pyarajan, S., Gaziano, J.M., *et al.* (2019). Harmonizing Genetic Ancestry and Self-identified
360 Race/Ethnicity in Genome-wide Association Studies. Am J Hum Genet *105*, 763-772.
361 Gaziano, J.M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen,
362 J., Shannon, C., Humphries, D., *et al.* (2016). Million Veteran Program: A mega-biobank to study
363 genetic influences on health and disease. J Clin Epidemiol *70*, 214-223.
364 Genomes Project, C., Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M.,
365 Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of
366 genetic variation from 1,092 human genomes. Nature *491*, 56-65.
367 Grinde, K.E., Qi, Q., Thornton, T.A., Liu, S., Shadyab, A.H., Chan, K.H.K., Reiner, A.P., and
368 Sofer, T. (2019). Generalizing polygenic risk scores from Europeans to Hispanics/Latinos. Genet
369 Epidemiol *43*, 50-62.
370 Hinckley, J.D., Abbott, D., Burns, T.L., Heiman, M., Shapiro, A.D., Wang, K., and Di Paola, J.
371 (2013). Quantitative trait locus linkage analysis in a large Amish pedigree identifies novel
372 candidate loci for erythrocyte traits. Mol Genet Genomic Med *1*, 131-141.
373 Johnson, K.E., and Voight, B.F. (2018). Patterns of shared signatures of recent positive selection
374 across human populations. Nat Ecol Evol *2*, 713-720.
375 Justice, A.E., Karaderi, T., Highland, H.M., Young, K.L., Graff, M., Lu, Y., Turcot, V., Auer, P.L.,
376 Fine, R.S., Guo, X., *et al.* (2019). Protein-coding variants implicate novel genes related to lipid
377 homeostasis contributing to body-fat distribution. Nat Genet *51*, 452-469.
378 Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N.,
379 Ikegawa, S., Hirata, M., Matsuda, K., *et al.* (2018). Genetic analysis of quantitative traits in the
380 Japanese population links cell types to complex human diseases. Nat Genet *50*, 390-400.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat Genet *42*, 348-354.

Kimura, H., Ishibashi, T., Uchida, T., Maruyama, Y., Friese, P., and Burstein, S.A. (1990). Interleukin 6 is a differentiation factor for human megakaryocytes in vitro. European journal of immunology *20*, 1927-1931.

Klarin, D., Damrauer, S.M., Cho, K., Sun, Y.V., Teslovich, T.M., Honerlaw, J., Gagnon, D.R., DuVall, S.L., Li, J., Peloso, G.M., *et al.* (2018). Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat Genet *50*, 1514-1523.

Kunishima, S., Kobayashi, R., Itoh, T.J., Hamaguchi, M., and Saito, H. (2009). Mutation of the beta1-tubulin gene associated with congenital macrothrombocytopenia affecting microtubule assembly. Blood *113*, 458-461.

Li, Y.R., and Keating, B.J. (2014). Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. Genome Med *6*, 91.

Lin, J., Zhu, Z., Xiao, H., Wakefield, M.R., Ding, V.A., Bai, Q., and Fang, Y. (2017). The role of IL-7 in Immunity and Cancer. Anticancer Res *37*, 963-967.

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.

Loh, P.R., Kichaev, G., Gazal, S., Schoech, A.P., and Price, A.L. (2018). Mixed-model association for biobank-scale datasets. Nat Genet *50*, 906-908.

403 Loh, P.R., Palamara, P.F., and Price, A.L. (2016). Fast and accurate long-range phasing in a UK
404 Biobank cohort. Nat Genet *48*, 811-816.
405 Lorenzo, F.R., Huff, C., Myllymaki, M., Olenchock, B., Swierczek, S., Tashi, T., Gordeuk, V.,
406 Wuren, T., Ri-Li, G., McClain, D.A.*, et al.* (2014). A genetic mechanism for Tibetan high-altitude
407 adaptation. Nat Genet *46*, 951-956.
408 Magi, R., Horikoshi, M., Sofer, T., Mahajan, A., Kitajima, H., Franceschini, N., McCarthy, M.I.,
409 Cogent-Kidney Consortium, T.D.G.C., and Morris, A.P. (2017). Trans-ethnic meta-regression of
410 genome-wide association studies accounting for ancestry increases power for discovery and
411 improves fine-mapping resolution. Hum Mol Genet *26*, 3639-3650.
412 Magi, R., and Morris, A.P. (2010). GWAMA: software for genome-wide association meta-
413 analysis. BMC Bioinformatics *11*, 288.
414 Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J.,
415 Steinthorsdottir, V., Scott, R.A., Grarup, N.*, et al.* (2018). Fine-mapping type 2 diabetes loci to
416 single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat
417 Genet *50*, 1505-1513.
418 Marouli, E., Graff, M., Medina-Gomez, C., Lo, K.S., Wood, A.R., Kjaer, T.R., Fine, R.S., Lu, Y.,
419 Schurmann, C., Highland, H.M.*, et al.* (2017). Rare and low-frequency coding variants alter human
420 adult height. Nature *542*, 186-190.
421 Marquez-Luna, C., Loh, P.R., South Asian Type 2 Diabetes, C., Consortium, S.T.D., and Price,
422 A.L. (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations.
423 Genet Epidemiol *41*, 811-823.
424 Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical
425 use of current polygenic risk scores may exacerbate health disparities. Nat Genet *51*, 584-591.

426 McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M.,
427 Fuchsberger, C., Danecek, P., Sharp, K.*, et al.* (2016). A reference panel of 64,976 haplotypes for
428 genotype imputation. Nat Genet *48*, 1279-1283.
429 Mousas, A., Ntritsos, G., Chen, M.H., Song, C., Huffman, J.E., Tzoulaki, I., Elliott, P., Psaty,
430 B.M., Blood-Cell, C., Auer, P.L.*, et al.* (2017). Rare coding variants pinpoint genes that control
431 human hematological traits. PLoS Genet *13*, e1006925.
432 Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T.,
433 Tamakoshi, A., Yamagata, Z., Mushiroda, T.*, et al.* (2017). Overview of the BioBank Japan
434 Project: Study design and profile. J Epidemiol *27*, S2-S8.
435 Pers, T.H., Timshel, P., and Hirschhorn, J.N. (2015). SNPsnap: a Web-based tool for identification
436 and annotation of matched SNPs. Bioinformatics *31*, 418-420.
437 Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161-164.
438 Popejoy, A.B., Ritter, D.I., Crooks, K., Currey, E., Fullerton, S.M., Hindorff, L.A., Koenig, B.,
439 Ramos, E.M., Sorokin, E.P., Wand, H.*, et al.* (2018). The clinical imperative for inclusivity: Race,
440 ethnicity, and ancestry (REA) in genomics. Hum Mutat *39*, 1713-1720.
441 Rabbolini, D.J., Morel-Kopp, M.C., Chen, Q., Gabrielli, S., Dunlop, L.C., Chew, L.P., Blair, N.,
442 Brighton, T.A., Singh, N., Ng, A.P.*, et al.* (2017). Thrombocytopenia and CD34 expression is
443 decoupled from alpha-granule deficiency with mutation of the first growth factor-independent 1B
444 zinc finger. J Thromb Haemost *15*, 2245-2258.
445 Raffield, L.M., Ulirsch, J.C., Naik, R.P., Lessard, S., Handsaker, R.E., Jain, D., Kang, H.M.,
446 Pankratz, N., Auer, P.L., Bao, E.L.*, et al.* (2018). Common alpha-globin variants modify
447 hematologic and other clinical phenotypes in sickle cell trait and disease. PLoS Genet *14*,
448 e1007293.

449 Raj, T., Kuchroo, M., Replogle, J.M., Raychaudhuri, S., Stranger, B.E., and De Jager, P.L. (2013).
450 Common risk alleles for inflammatory diseases are targets of recent positive selection. Am J Hum
451 Genet *92*, 517-529.
452 Rana, S.R., Sekhsaria, S., and Castro, O.L. (1993). Hemoglobin S and C traits: contributing causes
453 for decreased mean hematocrit in African-American children. Pediatrics *91*, 800-802.
454 Rappoport, N., Simon, A.J., Amariglio, N., and Rechavi, G. (2019). The Duffy antigen receptor
455 for chemokines, ACKR1,- 'Jeanne DARC' of benign neutropenia. Br J Haematol *184*, 497-507.
456 Reich, D., Nalls, M.A., Kao, W.H., Akylbekova, E.L., Tandon, A., Patterson, N., Mullikin, J.,
457 Hsueh, W.C., Cheng, C.Y., Coresh, J.*, et al.* (2009). Reduced neutrophil count in people of African
458 descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. PLoS
459 Genet *5*, e1000360.
460 Roden, D.M., Pulley, J.M., Basford, M.A., Bernard, G.R., Clayton, E.W., Balser, J.R., and Masys,
461 D.R. (2008). Development of a large-scale de-identified DNA biobank to enable personalized
462 medicine. Clin Pharmacol Ther *84*, 362-369.
463 Schick, U.M., Jain, D., Hodonsky, C.J., Morrison, J.V., Davis, J.P., Brown, L., Sofer, T.,
464 Conomos, M.P., Schurmann, C., McHugh, C.P.*, et al.* (2016). Genome-wide Association Study of
465 Platelet Count Identifies Ancestry-Specific Loci in Hispanic/Latino Americans. Am J Hum Genet
466 *98*, 229-242.
467 Tajuddin, S.M., Schick, U.M., Eicher, J.D., Chami, N., Giri, A., Brody, J.A., Hill, W.D.,
468 Kacprowski, T., Li, J., Lyytikainen, L.P.*, et al.* (2016). Large-Scale Exome-wide Association
469 Analysis Identifies Loci for White Blood Cell Traits and Pleiotropy with Immune-Mediated
470 Diseases. Am J Hum Genet *99*, 22-39.

471 Ulirsch, J.C., Lareau, C.A., Bao, E.L., Ludwig, L.S., Guo, M.H., Benner, C., Satpathy, A.T.,
472 Kartha, V.K., Salem, R.M., Hirschhorn, J.N., *et al.* (2019). Interrogation of human hematopoiesis
473 at single-cell and single-variant resolution. Nat Genet *51*, 683-693.
474 van Dongen, J., Jansen, R., Smit, D., Hottenga, J.J., Mbarek, H., Willemsen, G., Kluft, C.,
475 Collaborators, A., Penninx, B.W., Ferreira, M.A., *et al.* (2014). The contribution of the functional
476 IL6R polymorphism rs2228145, eQTLs and other genome-wide SNPs to the heritability of plasma
477 sIL-6R levels. Behav Genet *44*, 368-382.
478 Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic
479 epidemiology studies. Am J Hum Genet *81*, 208-227.
480 Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with P-
481 values. Genet Epidemiol *33*, 79-86.
482 Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population
483 Structure. Evolution *38*, 1358-1370.
484 Wellcome Trust Case Control, C., Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K.,
485 Su, Z., Howson, J.M., Auton, A., Myers, S., *et al.* (2012). Bayesian refinement of association
486 signals for 14 loci in 3 common diseases. Nat Genet *44*, 1294-1301.
487 Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: Fast and efficient meta-analysis of
488 genomewide association scans. Bioinformatics *26*, 2190-2191.
489 Williams, A.L., Patterson, N., Glessner, J., Hakonarson, H., and Reich, D. (2012). Phasing of many
490 thousands of genotyped samples. Am J Hum Genet *91*, 238-251.
491 Winkler, T.W., Day, F.R., Croteau-Chonka, D.C., Wood, A.R., Locke, A.E., Magi, R., Ferreira,
492 T., Fall, T., Graff, M., Justice, A.E., *et al.* (2014). Quality control and conduct of genome-wide
493 association meta-analyses. Nature protocols *9*, 1192-1212.

494 Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M.,
495 Patel, Y.M., Sorokin, E.P., Avery, C.L., *et al.* (2019). Genetic analyses of diverse populations
496 improves discovery for complex traits. Nature *570*, 514-518.
497 Xiang, K., Ouzhuluobu, Peng, Y., Yang, Z., Zhang, X., Cui, C., Zhang, H., Li, M., Zhang, Y.,
498 Bianba, *et al.* (2013). Identification of a Tibetan-specific mutation in the hypoxic gene EGLN1
499 and its contribution to high-altitude adaptation. Mol Biol Evol *30*, 1889-1898.
500 Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X., Pool, J.E., Xu, X., Jiang, H.,
501 Vinckenbosch, N., Korneliussen, T.S., *et al.* (2010). Sequencing of 50 human exomes reveals
502 adaptation to high altitude. Science *329*, 75-78.
503 Zhernakova, A., Elbers, C.C., Ferwerda, B., Romanos, J., Trynka, G., Dubois, P.C., de Kovel,
504 C.G., Franke, L., Oosting, M., Barisani, D., *et al.* (2010). Evolutionary and functional analysis of
505 celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. Am J Hum Genet
506 *86*, 970-977.
507 Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J.,
508 VandeHaar, P., Gagliano, S.A., Gifford, A., *et al.* (2018). Efficiently controlling for case-control
509 imbalance and sample relatedness in large-scale genetic association studies. Nat Genet *50*, 1335-
510 1341.
511

## Figure legends

**Figure 1.** Trans-ethnic and ancestry-specific meta-analyses of blood-cell traits. (**a**) Study design of the project. We used a fixed-effect meta-analysis strategy to analyze genetic associations within each of the five populations available, and a mega-regression approach that considers allele frequency heterogeneity for the trans-ethnic association tests. (**b**) Most blood-cell trait-associated loci physically overlap between populations. Despite different sample sizes between populations, we note that few loci are found in a single population, suggesting shared genetic architecture.

**Figure 2.** Fine-mapping of loci associated with hematological traits highlights likely functional variants. (**a**) We restricted fine-mapping to loci with evidence for a single association signal in European-ancestry (EA) populations. There are no such loci in Hispanic Americans. The 95% credible sets in the trans-ethnic meta-analyses are smaller than in the EA or East-Asian-ancestry (EAS) meta-analyses. (**b**) Trans-ethnic fine-mapping of a platelet locus. In EA individuals, the 95% credible set include 7 variants with posterior probability of inclusion (PPI) >0.04 and strong pairwise linkage disequilibrium (LD) with the sentinel variants rs10758481 ($r^2$>0.93 in GBR from 1000 Genomes Project, middle panel). LD is similarly strong in African-, Hispanic/South American-, and South-Asian-ancestry populations from the 1000 Genomes Project. However, LD is weaker in East Asians ($r^2$=0.68 in JPT from 1000 Genomes Project, bottom panel). In the trans-ethnic meta-analysis, rs10758481 has a PPI>0.99 (top panel). In EA and EAS, LD is color-coded based on pairwise $r^2$ with rs10758481. The dotted line indicates the genome-wide significance threshold ($P$<5x10$^{-9}$). (**c**) Number of variants in 95% credible sets in each population analyzed. In trans, EA and EAS, we identified 433, 403 and seven 95% credible sets with a single variant. (**d**) Annotation of variants in trans, EA and EAS shows a similar pattern, with a larger proportion of

likely functional variants (e.g. missense, intergenic and intronic variants within ATAC-seq peaks) among variants with higher posterior probability of inclusion (PPI). (**e**) g-chromVAR results for trans variants within 95% credible sets for 15 traits. The Bonferroni-adjusted significance level (corrected for 15 traits and 18 cell types) is indicated by the dotted line. mono, monocyte; gran, granulocyte; ery, erythroid; mega, megakaryocyte; CD4, CD4+ T cell; CD8, CD8+ T cell; B, B cell; NK, natural killer cell; mDC, myeloid dendritic cell; pDC, plasmacytoid dendritic cell; MPP, multipotent progenitor; LMPP, lymphoid-primed multipotent progenitor; CMP, common myeloid progenitor; CLP, common lymphoid progenitor; GMP, granulocyte–macrophage progenitor; MEP, megakaryocyte–erythroid progenitor.

**Figure 3.** A South-Asian-ancestry *IL7* missense variant associates with increased lymphocyte count in humans and IL7 secretion in an heterologous cellular system. (**a**) Lymphocyte count association results at the *IL7* locus in South Asians (SAS), European-ancestry participants (EA) and East Asians (EAS). In SAS, there are 7 genome-wide significant variants near *IL7*, but only rs201412253 is coding. Linkage disequilibrium (LD) $r^2$ is from 1000 Genomes Project SAS populations. In EA, the sentinel variant is located downstream of *IL7*; rs201412253 is rare (minor allele frequency=$4 \times 10^{-4}$) and not significant ($P$=0.073). In EAS, the locus is not associated with lymphocyte count. rs201412253 is monomorphic in 1000 Genomes Project EA and EAS so we could not calculate pairwise LD. (**b**) The 18Ile allele at *IL7*-rs201412253 increases IL7 secretion in a heterologous cellular system. Our ELISA assay did not detect secreted IL7 in clones generated with an empty vector. We tested eight independent clones for each *IL7* alleles. Each experiment was done in duplicate, and we performed the experiments three times (grey circles). The black

557 dots and vertical lines indicate means and standard deviations. We assess statistical significance
558 by linear regression correcting for experimental batch effects.
559
560 **Figure 4.** Selective sweep and association with platelet count at the *IL6* locus in East Asians. The
561 grey rectangle highlights a genomic region upstream of *IL6* that is strongly associated with platelet
562 (PLT) count. This association signal is driven by results from East Asians (EAS), and is absent
563 from other populations, including European- (EA) and African-ancestry (AFR) individuals
564 (green). The region overlaps several selective sweeps detected in EAS from the 1000 Genomes
565 Project (CDX, CHS, JPT, KHV). In orange, we provide standardized population branch site
566 (stdPBS) metrics in EA and EAS, indicative of allele frequency differentiation at this locus
567 between these two populations.
568

## METHODS

### Study design and participants

All participants provided written informed consent and the project was approved by each institution's ethical committee. **Supplementary Table 1** lists all participating cohorts. The SNPs we identified are available from the NCBI dbSNP database of short genetic variations (https://www.ncbi.nlm.nih.gov/projects/SNP/). No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

### Phenotypes

Complete blood count (CBC) and related blood indices were analyzed as quantitative traits. The descriptive statistics for each phenotype in each cohort analyzed are in **Supplementary Table 2**. Exclusion criteria and phenotype modeling in the UK Biobank (UKBB)(European-ancestry individuals), INTERVAL, and Biobank Japan (BBJ) have been described previously (Astle et al., 2016; Kanai et al., 2018). For all other studies, we followed the protocol developed by the Blood-Cell Consortium (Chami et al., 2016; Eicher et al., 2016; Tajuddin et al., 2016). Briefly, we excluded when possible participants with blood cancer, acute medical/surgical illness, myelodysplastic syndrome, bone marrow transplant, congenital/hereditary anemia, HIV, end-stage kidney disease, splenectomy, and cirrhosis, as well as pregnant women and those undergoing chemotherapy or erythropoietin treatment. We also excluded extreme blood-cell measures: WBC>200x$10^9$ cells/L, HGB>20 g/dL, HCT>60%, and PLT>1000x$10^9$ cells/L. For WBC subtypes, we analyzed $\log_{10}$-transformed absolute counts obtained by multiplying relative counts with total WBC count. For all phenotypes in all studies, we corrected the blood-cell phenotypes

592 for sex, age, age-squared, the 10 first genetic principal components, and other cohort-specific
593 covariates (e.g. recruitment center) using linear regression analysis. We applied rank-based inverse
594 normal transformation to the residuals form the regression analysis and used the normalized
595 residuals to test for association with genetic variants.
596
597 **Genotype quality-control and imputation**
598 The genotyping array and quality-control steps used by each cohort as well as their quality-control
599 steps are listed in **Supplementary Table 3**. Unless otherwise specified, all studies applied the
600 following criteria: samples were removed if the genotyping call rate was <95%, if they showed
601 excess heterozygosity, if we identified gender mismatches or sample duplicates, or if they appeared
602 as population outliers in principal component analyses nested with continental populations from
603 the 1000 Genomes Project (Genomes Project et al., 2012). We removed monomorphic variants, as
604 well as variants with Hardy-Weinberg $P<1x10^{-6}$ and call rate <98%.
605
606 Genotype imputation for the UKBB, INTERVAL, and BBJ have been described in details
607 elsewhere (Astle et al., 2016; Bycroft et al., 2018; Kanai et al., 2018). For all other studies, unless
608 specified in **Supplementary Table 3**, we applied the following steps for genotype imputation of
609 autosomal variants. We aligned all alleles on the forward strand of build 37/hg19 of the human
610 reference genome (http://www.well.ox.ac.uk/~wrayner/strand) and converted files into the VCF
611 format. We then applied checkVCF (http://genome.sph.umich.edu/wiki/CheckVCF.py) to confirm
612 strand and allele orientation. We carried out genotype imputation using the University of Michigan
613 (https://imputationserver.sph.umich.edu) or the Sanger Institute (https://imputation.sanger.ac.uk/)
614 imputation servers. We phased genotype data using SHAPEIT (Delaneau et al., 2013), EAGLE

(Loh et al., 2016), or HAPI-UR (Williams et al., 2012). For populations of European ancestry, we used reference haplotypes from the Haplotype Reference Consortium (HRC r1.1 2016) for imputation (McCarthy et al., 2016), whereas reference haplotypes from the 1000 Genomes Project (Phase 3, Version 5)(Genomes Project et al., 2012) were used for non-European ancestry participants.

**Study-level statistical analyses**

We tested an additive genetic model of association between genotype imputation doses and inverse normal transformed blood-cell phenotypes. We analyzed the major ancestry groups (European (EA), East Asian (EAS), African (AFR), Hispanic-Latino (HL), South Asian (SAS)) separately and used linear mixed-effect models implemented in BOLT-LMM (Loh et al., 2018), EPACTS (https://genome.sph.umich.edu/wiki/EPACTS), or EMMAX (Kang et al., 2010) to account for cryptic and known relatedness. Autosomal single nucleotide variants were analyzed in all contributing studies. For simplicity, we only analyzed insertion-deletion (indel) variants from UKBB and INTERVAL, since a similar reference panel was used for genotype imputation.

**Centralized quality-control and meta-analyses**

We performed a centralized quality-control check on the association results of each single study using EasyQC (v9.0)(Winkler et al., 2014). By mapping variants of each study to the appropriate ethnicity reference panel (HRC for EA and 1000 Genomes Project Phase3 for non-EA participants), we were able to harmonize alleles and markers across all studies. We were also able to assess the presence of flipped alleles per study and check for excessive allele frequency discrepancies using allele frequency reference data. We also inspected quantile-quantile (QQ)

638 plots generated by EasyQC and the corresponding genomic inflation factors as well as SE-N plots
639 (inverse of the median standard error vs. the square root of the sample size) to evaluate potential
640 issues with, for example, trait transformation or unaccounted relatedness. We removed variants
641 with imputation quality metric (INFO score) ≤0.4. Except for three studies, we also removed
642 variants with minor allele count (MAC) ≤5. For UKBB EA, Women Health Initiative (WHI), and
643 GERA (EA), we instead applied a MAC ≤20 filter because empirical observations suggested that
644 unusual inflation of the test statistics (i.e. extreme effect sizes and standard errors) was due to rarer
645 variants. To simplify handling of tri-allelic and indel variants, which have the same genomic
646 coordinates but different alleles, we created a unique variant ID for each tested variant.
647 Specifically, we assigned a chromosome:position(hg19)_allele1_allele2 unique ID to each variant,
648 in which the order of the allele in the ID was based on the lexicographical order or the indel length.
649 We performed inverse variance-weighted fixed-effect meta-analyses with GWAMA
650 (v2.2.2)(Magi and Morris, 2010) and trans-ethnic meta-analyses with MR-MEGA (v0.1.5)(Magi
651 et al., 2017). For MR-MEGA, we calculated four axes of genetic variation, the default
652 recommendation, to separate global population groups.
653
654 **Statistical significance, genomic inflation and locus definition**
655 For each meta-analysis, we calculated the genomic inflation factor ($\lambda_{GC}$) for all variants, which
656 were modest when considering the large sample sizes ($\lambda_{GC}$ range: 0.9-1.2) (**Supplementary Table**
657 **4**). We used $\alpha \leq 5 \times 10^{-9}$ after GC-correction to declare statistical significance, accounting for the
658 inflation of the test statistics and the number of blood-cell traits analyzed. To count the number of
659 loci that we discovered, we first identified the most significant variants (with $P \leq 5 \times 10^{-9}$) and
660 extended the physical region around that variant 250-kb on each side. Overlapping loci were

661 merged, and we used the most significant variant within the interval as the sentinel variant. In this
662 manuscript, we defined as novel a locus if no variants were previously reported in the literature to
663 be associated with the specific blood-cell trait and if the locus is not reported in the companion
664 manuscript that focuses on EA-specific genetic discoveries.
665
666 **Million Veteran Program (MVP) blood-cell trait analyses for replication**
667 *Phenotyping.* Phenotyping methods published by the EMERGE Consortium and available on
668 PheKB (https://phekb.org/) were used for retrieving lab data and exclusion criteria for all blood
669 cell indices. This information was pulled from the VA electronic medical records for all MVP
670 participants. Lab data was subject to the Boston Lab Adjudication Protocol. This entails five steps:
671 (i) compile an initial spreadsheet of possible relevant lab tests, (ii) Subject Matter Expert (SME)
672 does an initial review of possible tests, (iii) analyst adds relevant LOINC codes for SME review,
673 (iv) second Subject Matter Expert (SME) review, (v) creation of a Lab Phenotype Table/Data Set.
674 After restricting to only outpatient labs and applying the EMERGE exclusion criteria, for each trait
675 and each person, the minimum, maximum, mean, median, SD, and number of labs was recorded.
676 Values were compared to those from UKBB (Astle et al., 2016).
677
678 *Genotyping.* DNA extracted from whole blood was genotyped using a customized Affymetrix
679 Axiom biobank array, the MVP 1.0 Genotyping Array. With 723,305 total DNA sequence variants,
680 the array is enriched for both common and rare variants of clinical importance in different ethnic
681 backgrounds (Klarin et al., 2018).
682

*Analysis.* The median lab value was the trait used for analysis. Linear regression models were run under an additive model in plink2 on 1000G (v3p5) imputed dosages. Analyses were run using models described above within each race/ethnicity stratum (AFR, ASN, EA, HA) classified based on their genotype data using HARE (Fang et al., 2019). Meta-analyses for the trans-ethnic analyses were completed in METAL (Willer et al., 2010).

**Heritabilities and genetic correlations**

We calculated heritabilities and genetic correlations between blood-cell traits within the EA and EAS populations using default parameters implemented in the LD score regression method (**Supplementary Table 17** and **Supplementary Figure 5**)(Bulik-Sullivan et al., 2015a; Bulik-Sullivan et al., 2015b). For genetic correlation of the same phenotype between ancestral populations, we used Popcorn (Brown et al., 2016). Briefly, Popcorn uses a Bayesian framework to estimate, using genome-wide summary statistics, the genetic correlation of the same phenotype but in two different populations (in our case, between EA and EAS). It reports the trans-ethnic genetic-effect correlation ($\rho_{ge}$), i.e. the correlation coefficient of per-allele SNP effect sizes, but also the trans-ethnic genetic impact correlation ($\rho_{gi}$), which includes a normalization of the effect based on allele frequency (**Supplementary Table 18**). To address whether a difference in the sample size for the EA and EAS meta-analyses could impact the Popcorn results, we repeated our analyses using the current EAS results ($N_{max}$=151,807) and EA results from preliminary analyses of the UKBB dataset ($N_{max}$=87,265)(Astle et al., 2016). These analyses confirmed that for common variants, cross-ancestry EA-EAS genetic correlations are significantly different (but non-null). Both LD score regression and Popcorn are not amenable to admixed populations, and cannot

705 handle rare variants. For these reasons, we limited these analyses to the large EA and EAS
706 populations and focused on common variants from the 1000 Genomes Project.
707
708 **Statistical fine-mapping**
709 No fine-mapping methods currently exist to handle admixed populations. Furthermore, for some
710 of the ethnic groups analyzed here, we did not have access to a sufficiently large reference panel
711 to properly account for LD, complicating conditional analyses and fine-mapping efforts. For these
712 reasons, we fine-mapped the ancestry-specific fixed-effect meta-analyses by adapting the method
713 proposed by Maller et al. (Wellcome Trust Case Control et al., 2012) in order to assign posterior
714 probability of inclusion (PPI) to each variant and construct 95% credible sets.
715
716 This method makes the strong assumption that there is a single independent causal variant at the
717 tested locus. For this reason, we limited our Bayesian fine-mapping to loci where we identified a
718 single independent association signal by conditional analysis in EA individuals from the UKBB
719 (**Supplementary Methods**). Because EA represented the largest group, we then inferred that there
720 was also a single association signal in the other populations at these loci, an inference that may not
721 always be right. Briefly, we added 250-kb on either side of genome-wide significant variants
722 ($P<5x10^{-9}$) and merged loci when they overlapped. For the loci identified in the ancestry-specific
723 meta-analyses, we converted P-values into approximate Bayes factors (aBF) using (Wakefield,
724 2009; Wellcome Trust Case Control et al., 2012):
725
726 $$aBF = \sqrt{\frac{SE^2}{SE^2 + \omega}}\, exp\left[\frac{\omega\beta^2}{2SE^2(SE^2 + \omega)}\right]$$

where β and SE are the variant's effect size and standard error, respectively, and ω denotes the prior variance in allelic effects, taken here to be 0.04 (Wakefield, 2007). For the trans-ethnic results, we directly used Bayes factors calculated by MR-MEGA (Magi et al., 2017). We calculated PPI of each variant by dividing the variant's aBF by the sum of the aBF for all the variants within the locus. We generated the 95% credible sets by ordering all variants in a given locus from the largest to the smallest PPI and by including variants until the cumulative sum of the PPI ≥95% (Mahajan et al., 2018). All variants that map to 95% credible sets are available online (see **URL**).

**Functional annotation**

To derive basic functional annotation information, we annotated all variants included in 95% credible sets from ancestry-specific and trans-ethnic meta-analyses with the Variant Effect Predictor (VEP)( https://useast.ensembl.org/info/docs/tools/vep/index.html), compiling both all consequences and the most severe consequence for Ensembl/GENCODE transcripts. We also specifically annotated rare coding variants using VEP (defined as any variant with MAF <1% in a given analysis, with a GC-corrected P-value $<5x10^{-9}$, and annotated as a missense_variant, stop_gained, stop_lost, splice_donor, or a splice_acceptor, regardless of fine-mapping results). We removed all variants with a GC-corrected P-value $<5x10^{-9}$ in EA, in the MHC region, and, in analyses including individuals with at least some African ancestry, on chromosome 1 for neutrophils and total WBC count and for RBC traits near the chromosome 11 β-globin and the chromosome 16 α-globin loci.

Bias-corrected enrichment of blood trait variants for chromatin accessibility of 18 hematopoietic populations was performed using g-chromVAR, which has been previously described in detail (Ulirsch et al., 2019). In brief, this method weights chromatin features by fine-mapped variant posterior probabilities and computes the enrichment for each cell type versus an empirical background matched for GC content and feature intensity. For chromatin feature input, we used a consensus peak set for all hematopoietic cell types with a uniform width of 500 bp centered at the summit. For variant input, we included all fine-mapped variants within 95% credible sets of the trans-ethnic GWAS. We also ran g-chromVAR for each ancestry-specific meta-analysis, keeping all other parameters the same, but using fine-mapped variants with the 95% credible sets of each ancestry-specific study. Finally, to control for the number of loci tested within each ancestry-specific study, we first ranked the loci of the largest cohort (i.e. EA) by sentinel variant p-value, and then subset only the top *n* loci, where *n* equals the number of loci in the smaller cohort (e.g. EAS) for the same trait. We then ran g-chromVAR on the subset of variants falling within these top *n* loci.

**Phenome-wide association study (pheWAS) analysis**

*UK Biobank (UKBB)*. We extracted pheWAS results for a list of 5552 variants in UKBB ICD PheWeb hosted at the University of Michigan (Accessed 21 August 2019). To account for severe imbalance in case-control ratios, we selected the output from the SAIGE analyses (http://pheweb.sph.umich.edu/SAIGE-UKB/) based on 408,961 samples from White British participants (Zhou et al., 2018). In total, 1403 phecodes were tested for association. All results were downloaded using R, and were parsed and organized into data table format using the data.table, rvest, stringr, dplyr and tidyr packages.

773

774 *Biobank Japan (BBJ).* We performed a pheWAS for the lead variants identified by the trans-ethnic
775 meta-analyses. From the list of all the significantly associated variants with blood cell-related
776 traits, we extracted those genotyped or imputed in the BBJ project ($n_{SNP}$ = 4,255). Next, we curated
777 the phenotype record of the disease status and clinical values for the same individuals analyzed in
778 the discovery phase ($n_{indiv}$ = 143,988). Then, we performed the logistic regression analyses for 22
779 binary traits (20 diseases and 2 behavioral habits) which had a sufficient number of case samples
780 ($n_{case}$ = 2,500). Regression models were adjusted for age, sex and 20 principal components as
781 covariates. Trait-specific covariates are described elsewhere (Kanai et al., 2018).

782

783 *BioVU*. BioVU is the biobank of Vanderbilt University Medical Center (VUMC) that houses de-
784 identified DNA samples linked to phenotypic data derived from electronic health records (EHRs)
785 system of VUMC. The clinical information is updated every 1-3 months for the de-identified
786 EHRs. Detailed description of program operations, ethical considerations, and continuing
787 oversight and patient engagement have been published (Roden et al., 2008). DNA samples were
788 genotyped with genome-wide arrays including the Multi-Ethnic Global **(**MEGA) array, and the
789 genotype data were imputed into the HRC reference panel(McCarthy et al., 2016) using the
790 Michigan imputation server (Das et al., 2016). Imputed data and the 1000 Genome Project data
791 were combined to carry out principal component analysis (PCA) and African-American samples
792 were extracted for analysis based on the PCA plot. PheWAS were carried out for each SNP with
793 the specified allele (Denny et al., 2010). Phenotypes were derived from billing codes of EHRs as
794 described previously (Carroll et al., 2014). Each phenotype ('phecode') has defined case, control
795 and exclusion criteria. We required two codes on different visit days to instantiate a case for each

phecode. In total, 1815 phecodes were tested for association. Association between each binary phecode and a SNP was assessed using logistic regression, while adjusting for covariates of age, sex, genotyping array type/batch and 10 principal components of ancestry.

*Merging across biobanks.* We defined statistical significance within each biobank to be a Bonferroni corrected level of $0.05/pq$, where $p$ is the number of phecodes tested and $q$ is the number of variants tested. We merged results across UKBB and BioVU by matching on phecode, as these two biobanks used the same phecode system for classifying outcomes. To merge with BBJ, we cross-referenced the 22 outcomes in BBJ with the phecode library used by BioVU/UKBB. Matches were determined based on phenotype similarity between the BioVU/UKBB phenotype description and the outcomes described in Nagai et *al.* (Nagai et al., 2017).

**Polygenic trait score (PTS) analyses**

We restricted these analyses to variant-trait associations that reached genome-wide significance ($P<5\text{x}10^{-9}$) in the trans-ethnic MR-MEGA meta-analyses (**Supplementary Table 5**). For each of these variant-trait pairs, we calculated an effect size – hereafter referred to as trans weights – using the fixed-effect meta-analysis method implemented in GWAMA and all cohorts available (Magi and Morris, 2010). For the same variants, we also retrieved the ancestry-specific effect sizes (or weights). We calculated the PTS using plink2 by summing up the number of trait-increasing alleles (or imputation doses) that were weighted by their corresponding trans ($PTS_{trans}$) or ancestry-specific ($PTS_{EA}$, $PTS_{AFR}$, $PTS_{HA}$) weights. The variance explained by the PTS on corrected and normalized blood-cell traits was calculated in R using linear regression. For these analyses, we

had access to 2,651 AFR, 5,048 EA and 4,281 HA BioMe participants, as well as 2546 AFR ARIC participants, that were not used in the discovery effort.

**Analysis of natural selection**

To quantify the contribution of positive selection on blood-cell trait variation, we used the recent map of selective sweeps identified in the different populations of the 1000 Genomes Project (Johnson and Voight, 2018). We grouped the sweeps identified in the 26 1000 Genomes Project populations into five larger populations that correspond to our ancestry-specific meta-analyses: Europe-ancestry (CEU, TSI, GBR, FIN, IBS); East-Asian-ancestry (CHB, JPT, CHS, CDX, KHV); African-ancestry (YRI, LWK, GWD, MSL, ESN, ASW, ACB); South-Asian-ancestry (GIH, PJL, BEB, STU, ITU); and Hispanic/Latino-ancestry (MXL, PUR, CLM, PEL). Following the nomenclature by Johnson and Voight(Johnson and Voight, 2018), each selective sweep is summarized by the variant located within the sweep that has the highest iHS value. iHS (Integrated Haplotype Score) is a statistic to quantify evidence of recent positive selection. A high positive iHS score (iHS > 2) means that haplotypes on the ancestral allele background are longer compared to derived allele background. A high negative iHS score (iHS < -2) means that the haplotypes on the derived allele background are longer compared to the haplotypes associated with the ancestral allele. We retrieved the blood-cell trait association results for these sweep-tagging SNPs from the ancestry-specific meta-analyses (**Supplementary Table 19**). To determine if the inflation observed in the QQ plots was significant, we generated 100 sets of SNPs that match the selective sweep-tagging SNPs based on allele frequency, gene proximity, and the number of LD proxies in European-ancestry, East-Asian-ancestry and African-ancestry individuals using SNPsnap (Pers et al., 2015). For these analyses, we excluded the HLA region and variants in LD ($r^2$>0.5). We

841 computed empirical significance by tallying the number of sets with the same or more genome-
842 wide significant variants than the canonical sets of selective sweep-tagging SNPs (**Supplementary**
843 **Table 20**).
844
845 We also computed the population branch statistic (PBS)(Yi et al., 2010). PBS measures the amount
846 of allele frequency change in the population since its divergence from the other two populations.
847 For a target population, PBS is calculated as:
848
849 $$PBS = \frac{T^{target,sister} + T^{target,outgroup} - T^{sister,outgroup}}{2}$$
850
851 where $T = -\log(1 - F_{ST})$ is an estimate of the divergence time between two populations. Here,
852 $F_{ST}$ between each pair of populations was estimated using Weir and Cockerham's estimate (Weir
853 and Cockerham, 1984). We then divided all variants with calculated PBS into 50 bins of equal size
854 by derived allele count in the target population, and then standardized the raw PBS values within
855 each bin. To calculate PBS for Europe-ancestry (CEU, TSI, GBR, and IBS, without FIN), we used
856 YRI as an outgroup and East-Asian-ancestry (CHB, JPT, CHS, CDX, KHV) as a sister population;
857 for East-Asian-ancestry, we used YRI as an outgroup and Europe-ancestry as a sister population;
858 for YRI, we used East-Asian-ancestry as an outgroup and Europe-ancestry as a sister population.
859
860 **IL7 functional analyses**
861 We PCR amplified and cloned the *IL7* wildtype (rs201412253-Val18) and mutant (rs201412253-
862 18Ile) open reading frame (ORF) in the pcDNA5/FRT vector (ThermoFisher Scientific) using
863 HindIII and BamHI restriction sites (see **Supplementary Table 22** for ORF and primer

sequences). We validated the sequences of the two plasmids by Sanger Sequencing. Flip-In™-293 cells (ThermoFisher Scientific) at 80% confluency were transfected with 1:10 mixes of empty pcDNA5 or pcDNA5 derivatives coding for IL7-Val8 or IL7-18Ile and pOG44 FLP recombinase coding vector (ThermoFisher Scientific) using polyethylenimine. Transfectant clones were expanded and selected in DMEM medium supplemented with 10% Foetal Bovine Serum, 4 mM L-glutamine, 100 IU penicillin, 100 µg/ml streptomycin and 100 µg/ml hygromycin. We measured the secretion of IL7 in eight independent clones for each *IL7* allele (rs201412253-Val18 and rs201412253-18Ile) as well as in four clones generated with the empty vector by ELISA assay. We used the High Sensitivity Quantikine HS ELISA kit from R & D Systems (Cat # HS750). We seeded 100,000 cells per 12-wells plates and grew them for 6 days in DMEM glutamax plus 10% FBS before doing the ELISA. We measured each supernatant in duplicate and seeded each of the clones in triplicate. The whole experiment was done on three different weeks (three complete biological replicates). We extracted total proteins from cells with RIPA buffer and we quantified the lysates by BCA. We used this quantification to normalize the ELISA assays. We extracted total RNA from ~500,000 cells using the Qiagen RNEasy kit (cat # 74136). We checked the quality of the RNA by Bioanalyzer and quantified its concentration by Nanodrop. We reverse transcribed 1 ug of total RNA into cDNA using the ABI kit (Life Technologies Cat # 4368814). We used two pairs of primers for *IL7* and assays for three normalizing genes (*HPRT*, *GAPDH*, *TBP,* **Supplementary Table 22**). We followed the MIQE recommendations and performed the qPCR reactions with the Sybergreen Platinum (Life Technologies Cat # 11733-046) on a Biorad CFX384 thermocycler.

## DATA AVAILABILITY STATEMENT

All results are available at the following URL: http://www.mhi-humangenetics.org/en/resources

## URLs

checkVCF: http://genome.sph.umich.edu/wiki/CheckVCF.py

EPACTS: https://genome.sph.umich.edu/wiki/EPACTS

Imputation servers: https://imputationserver.sph.umich.edu or https://imputation.sanger.ac.uk/

NCBI dbSNP: https://www.ncbi.nlm.nih.gov/projects/SNP/

PheKB: https://phekb.org/

Strand alignment resources: http://www.well.ox.ac.uk/~wrayner/strand

UK Biobank SAIGE results: http://pheweb.sph.umich.edu/SAIGE-UKB/

Variant Effect Predictor: https://useast.ensembl.org/info/docs/tools/vep/index.html

**Table 1**. Non-synonymous variants with a minor allele frequency (MAF) ≤1% identified in non-European-ancestry (EA) populations or in the trans-ethnic meta-analyses. The population in which each variant was discovered is listed in the first column. Complete association results for each variant are available in **Supplementary Table 15**. Genomic coordinates (chr:position) are on build hg19. For the trans-ethnic results, mMAF corresponds to the mean MAF across all studies. EA, European-ancestry; EAS, East Asians; SAS, South Asians; AFR, African-ancestry; HA, Hispanics; PLT, platelet; NEU, neutrophil; MCH, mean corpuscular hemoglobin; MCV, mean corpuscular volume; MCHC, mean corpuscular hemoglobin concentration; MPV, mean platelet volume; EOS, eosinophil; MON, monocyte; RDW, red blood cell distribution width; LYM, lymphocyte; RBC, red blood cell count; HGB, hemoglobin; WBC, white blood cell.

| Population | Position | Phenotype | rsID | mMAF (%) | P-value | Gene | Annotation | Note |
|---|---|---|---|---|---|---|---|---|
| EAS | 1:43805737 | PLT | rs117656396 | 0.9 | $5.20x10^{-12}$ | *MPL* | missense | Thrombopoietin receptor; MAF <0.1% in EA and SAS. |
| EAS | 3:133476778 | MCH | rs143019827 | 0.4 | $4.52x10^{-9}$ | *TF* | missense | Transferrin (iron binding protein); monomorphic in non-EAS populations. |
| EAS | 3:184046450 | PLT | rs112809828 | 0.08 | $2.03x10^{-35}$ | *EIF4G1* | missense | MAF <0.01% in EA. Located ~45kb from thrombopoietin (*THPO*). |
| EAS | 6:16295278 | MCH/MCV | rs78806162 | 0.4 | $4.79x10^{-14}$/ $4.82x10^{-12}$ | *GMPR* | missense | MAF <0.1% in SAS; monomorphic in EA. |
| EAS | 6:41621210 | MCH/MCV | rs201503063 | 1 | $4.28x10^{-9}$/ $1.03x10^{-9}$ | *MDFI* | missense | Monomorphic in non-EAS populations. |
| EAS | 7:100014072 | MCH/MCV | rs6957339 | 0.07 | $3.65x10^{-21}$/$1.02x10^{-10}$ | *ZCWPW1* | missense | MAF in EA <0.03%; more common in AFR and HA populations (MAF=0.2-0.6%). Located ~204kb from *TFR2*. |
| EAS | 19:1049396 | LYM/WBC | rs201347186 | 0.6 | $8.78x10^{-13}$/$6.53x10^{-22}$ | *ABCA7* | missense | Previous GWAS identified common missense SNPs in *ABCA7* associated with NEU (Astle et al., 2016); monomorphic in non-EAS populations. |
| EAS | 20:57599434 | PLT | rs121918555 | 0.06 | $1.27x10^{-13}$ | *TUBB1* | missense | Mutation previously characterized in Japanese patients with congenital macrothrombocytopenia (Kunishima et al., 2009); monomorphic in non-EAS populations. |
| SAS | 9:135863848 | MPV/PLT | rs527297896 | 0.5 | $2.24x10^{-19}$/$3.57x10^{-18}$ | *GFI1B* | missense | Mutation previously identified in patients with thrombocytopenia without α-granule or bleeding defect (Rabbolini et al., 2017); monomorphic in non-SAS populations. |
| Trans | 1:43803807 | PLT | rs17292650 | 0.2 | $2.87x10^{-16}$ | *MPL* | missense | Thrombopoietin receptor; known common variant in AFR populations (MAF=4.2%)(Auer et al., 2012). |
| Trans | 1:231557623 | HCT | rs186996510 | 1 | $4.90x10^{-15}$ | *EGLN1* | missense | Low-frequency variant in EAS (MAF=4.1%) and AFR (3.2%), rare in SAS (0.6%) and EA (<0.01%). Associated with adaptation to |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | hypoxia and HGB in Tibetans (Xiang et al., 2013). |
| Trans | 7:80300449 | RDW | rs3211938 | 0.3 | $5.75x10^{-14}$ | *CD36* | nonsense | Known common variant in AFR populations (MAF=9.3%)(Chami et al., 2016). |
| Trans | 9:136083640 | RBC | rs12336956 | 0.8 | $1.01x10^{-9}$ | *OBP2B* | missense | Association signal in EA (MAF=0.3%); MAF=3% in HL; MAF=19% in AFR. Located ~47kb from *ABO*. |
| Trans | 11:5248232 | LYM | rs334 | 0.2 | $9.50x10^{-16}$ | *HBB* | missense | Sickle cell anemia mutation, well known for RBC trait associations. |
| Trans | 12:111856673 | PLT | rs78894077 | 0.9 | $7.63x10^{-38}$ | *SH2B3* | missense | Common variant in EAS (MAF=3.6%). Associated with myeloproliferative neoplasms (Chen et al., 2016). |
| Trans | 17:38062390 | WBC | rs35266519 | 0.8 | $6.69x10^{-12}$ | *GSDMB* | missense | Rare across all populations. Previously reported for NEU count (Mousas et al., 2017). |

**FIGURE 1**

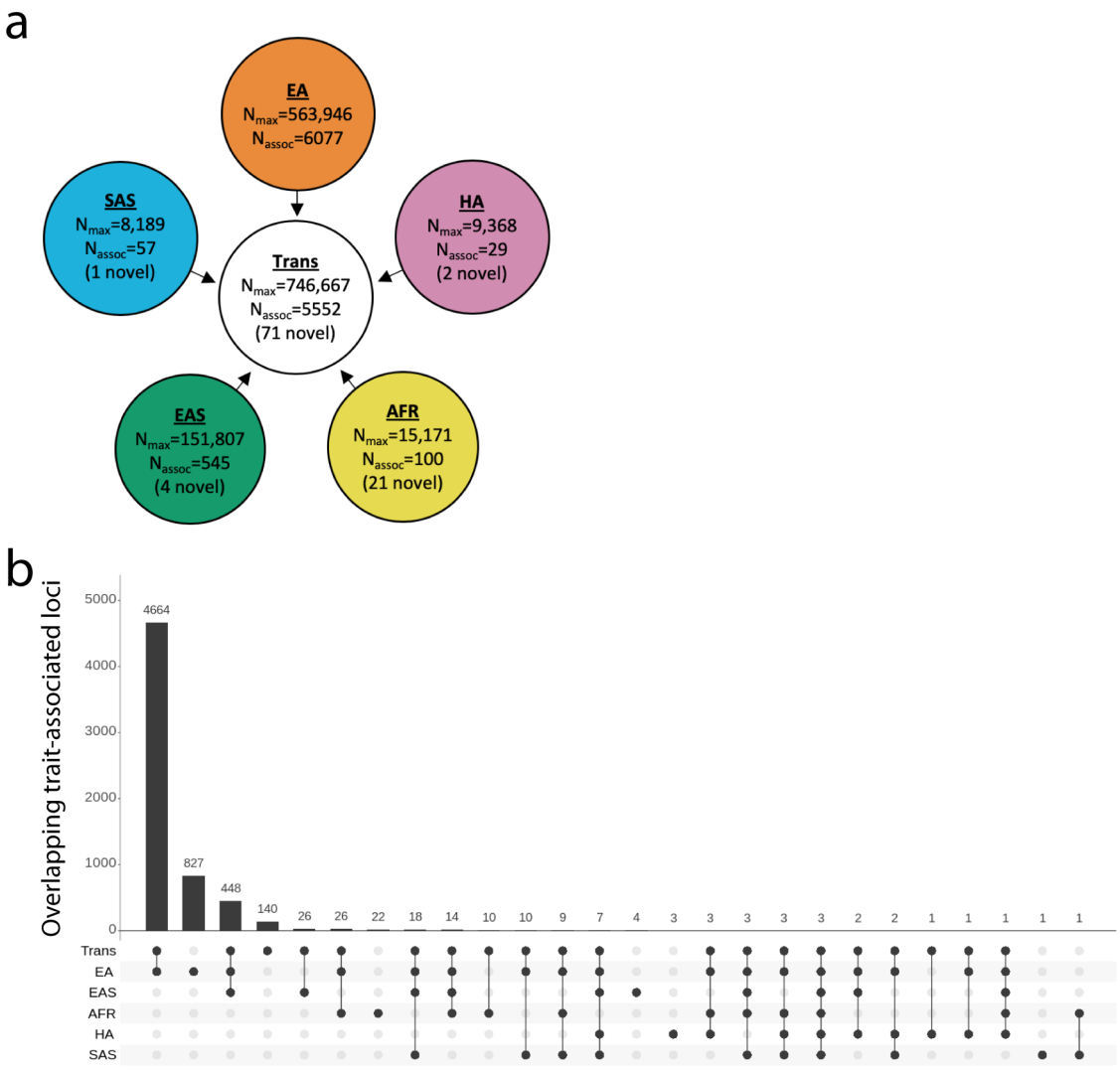


**FIGURE 2**

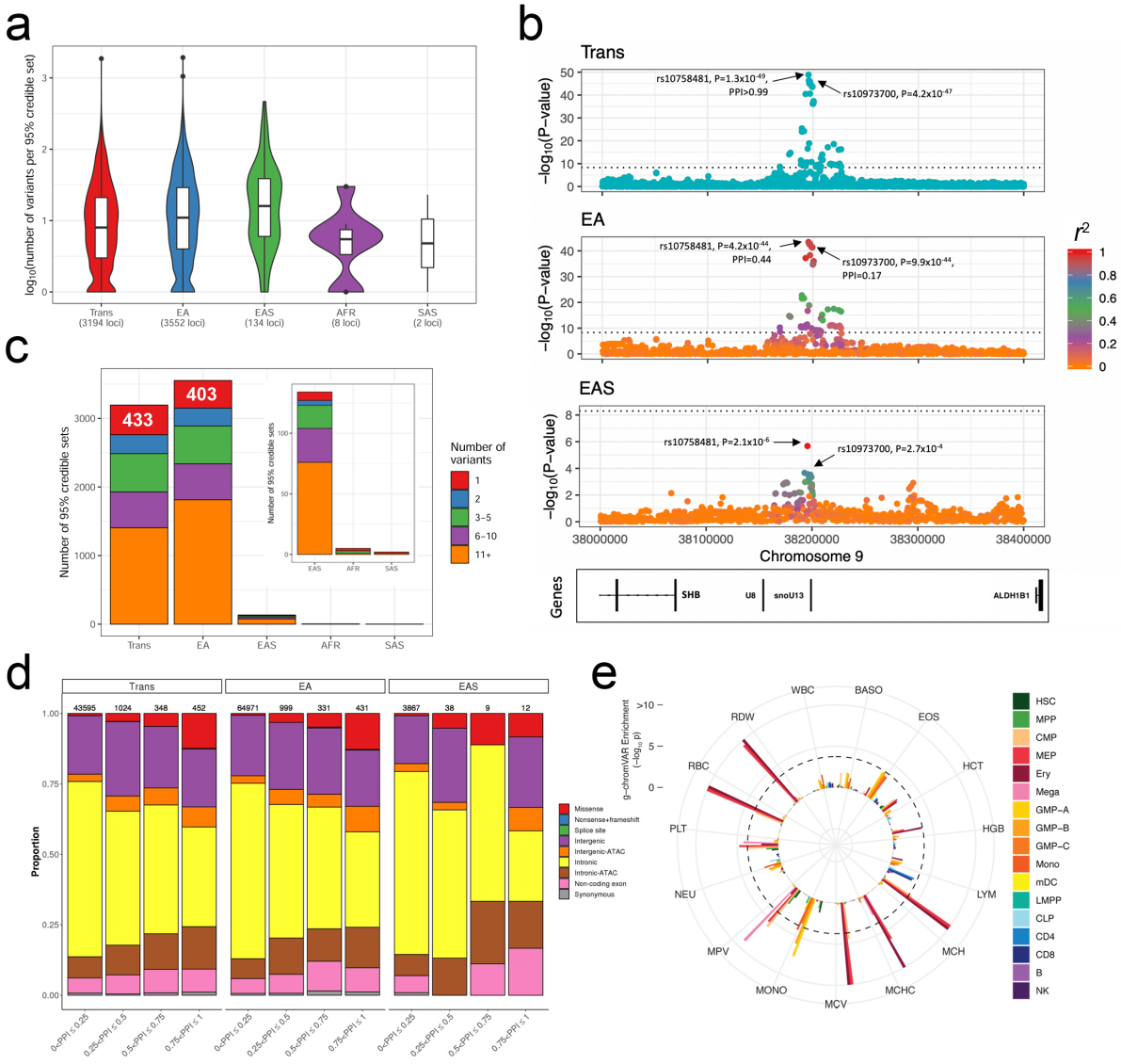


**FIGURE 3**

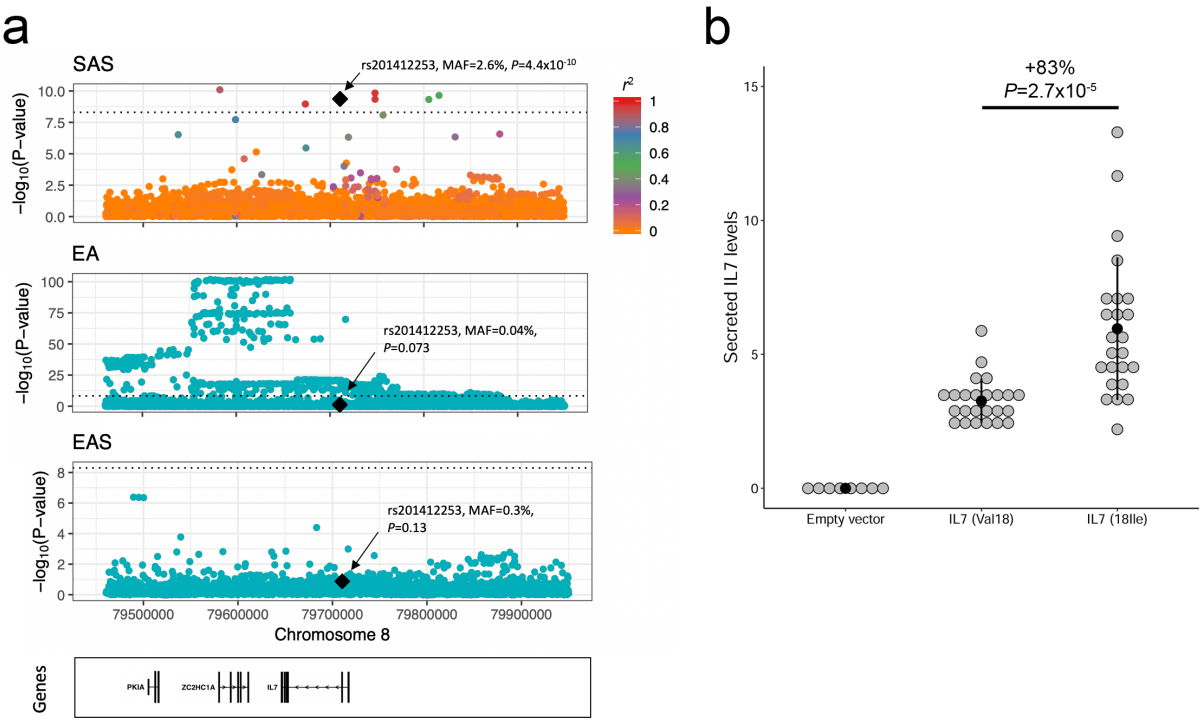
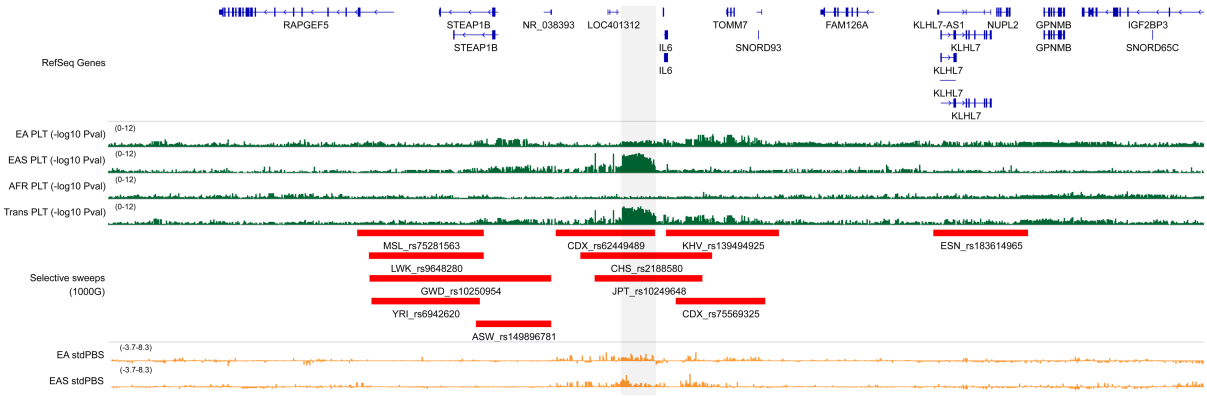


**FIGURE 4**



# AUTHORLIST

Ming-Huei Chen[1,2,*], Laura M. Raffield[3,*], Abdou Mousas[4,*], Saori Sakaue[5,6], Jennifer E. Huffman[7], Tao Jiang[8], Parsa Akbari[9,10,11], Dragana Vuckovic[11,12], Erik L. Bao[13,14], Arden Moscati[15], Xue Zhong[16], Regina Manansala[17], Véronique Laplante[18], Minhui Chen[19], Ken Sin Lo[4], Huijun Qian[20], Caleb A. Lareau[13,14], Mélissa Beaudoin[4], Masato Akiyama[6,21], Traci M. Bartz[22], Yoav Ben-Shlomo[23], Andrew Beswick[24], Jette Bork-Jensen[25], Erwin P. Bottinger[15,26], Jennifer A. Brody[27], Frank J.A. van Rooij[28], Kumaraswamynaidu Chitrala[29], Kelly Cho[30,31,32], Hélène Choquet[33], Adolfo Correa[34], John Danesh[8,35,36,37,38,39], Emanuele Di Angelantonio[8,35,36,39], Niki Dimou[40,41], Jingzhong Ding[42], Paul Elliott[43,44,45,46,47], Tõnu Esko[14], Michele K. Evans[29], James S. Floyd[27,48], Linda Broer[49], Niels Grarup[25], Michael H. Guo[14,50], Andreas Greinacher[51], Jeff Haessler[52], Torben Hansen[25], Joanna M. M. Howson[8], Wei Huang[53], Eric Jorgenson[33], Tim Kacprowski[54,55,56], Mika Kähönen[57,58], Yoichiro Kamatani[6,59], Masahiro Kanai[6,60], Savita Karthikeyan[8], Fotis Koskeridis[41], Leslie A. Lange[61], Terho Lehtimäki[62,63], Markus M. Lerch[64], Allan Linneberg[65,66], Yongmei Liu[67], Leo-Pekka Lyytikäinen[62,63], Ani Manichaikul[68], Koichi Matsuda[69], Karen L. Mohlke[3], Nina Mononen[62,63], Yoshinori Murakami[70], Girish N. Nadkarni[15], Matthias Nauck[56,71], Kjell Nikus[72,73], Willem H. Ouwehand[11,36,39,74,75], Nathan Pankratz[76], Oluf Pedersen[25], Michael Preuss[15], Bruce M. Psaty[27,77,78,79], Olli T. Raitakari[80,81,82], David J. Roberts[83,84], Stephen S. Rich[68], Benjamin A.T. Rodriguez[1,2], Jonathan D. Rosen[85], Jerome I. Rotter[86], Petra Schubert[30], Cassandra N. Spracklen[3], Praveen Surendran[9], Hua Tang[87], Jean-Claude Tardif[4,88], Mohsen Ghanbari[28], Uwe Völker[54,56], Henry Völzke[56,89], Nicholas A. Watkins[75], Alan B. Zonderman[29], VA Million Veteran Program[†], Peter W.F. Wilson[90], Yun Li[3,85,91], Adam S. Butterworth[8,35,36], Jean-François Gauchat[18], Charleston W.K. Chiang[19,92], Bingshan Li[93], Ruth J.F. Loos[15], William J. Astle[10,12,94], Evangelos Evangelou[41,43], Vijay G. Sankaran[13,14], Yukinori Okada[5,95], Nicole Soranzo[11,74], Andrew D. Johnson[1,2,§], Alexander P. Reiner[48,§], Paul L. Auer[17,§], Guillaume Lettre[4,88,§]

[1]The Framingham Heart Study, National Heart, Lung and Blood Institute, Framingham, MA, USA, [2]Population Sciences Branch, Division of Intramural Research, National Heart, Lung and Blood Institute, Framingham, MA, USA, [3]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA, [4]Montreal Heart Institute, Montreal, Quebec, Canada, [5]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Osaka, Japan, [6]Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan, [7]Center for Population Genomics, Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, MA, USA, [8]Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK, [9]MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK, [10]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK, [11]Human Genetics, Wellcome Sanger Institute, Hinxton, UK, [12]NIHR Blood and Transplant Research Unit in Donor Health and Genomics, Strangeways Laboratory, Cambridge, UK, [13]Division of Hematology/Oncology, Boston Children's Hospital, Boston, MA, USA, [14]Broad Institute of Harvard and MIT, Cambridge, MA, USA, [15]Icahn School of Medicine at Mount Sinai, The Charles Bronfman Institute for Personalized Medicine, New York, NY, USA, [16]Department of Medicine, Division of Genetic Medicine, Vanderbilt University, Nashville, TN, USA, [17]Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA, [18]Département de Pharmacologie et Physiologie, Faculty of Medicine, Université de Montréal, Montreal, Quebec, Canada, [19]Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA,

USA, [20]Department of Statistics and Operation Research, University of North Carolina, Chapel Hill, NC, USA, [21]Department of Ocular Pathology and Imaging Science, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan, [22]Department of Biostatistics, University of Washington, Seattle, WA, USA, [23]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK, [24]Translational Health Sciences, Musculoskeletal Research Unit, Bristol Medical School, University of Bristol, Bristol, UK, [25]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark, [26]Hasso-Plattner-Institut, Universität Potsdam, Postdam, Germany, [27]Department of Medicine, University of Washington, Seattle, WA, USA, [28]Department of Epidemiology, Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands, [29]Laboratory of Epidemiology and Population Science, National Institute on Aging/NIH, Baltimore, MD, USA, [30]Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Boston Healthcare System, Boston, MA, USA, [31]Department of Medicine, Division on Aging, Brigham and Women's Hospital, Boston, MA, USA, [32]Department of Medicine, Harvard Medical School, Boston, MA, USA, [33]Division of Research, Kaiser Permanente Northern California, Oakland, CA, USA, [34]Department of Medicine, University of Mississippi Medical Center, Jackson, MS, USA, [35]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK, [36]The National Institute for Health Research Blood and Transplant Unit (NIHR BTRU) in Donor Health and Genomics at the University of Cambridge, Cambridge, UK, [37]Human Genetics, Wellcome Sanger Institute, Saffron Walden, UK, [38]National Institute for Health Research Cambridge Biomedical Research Centre, Cambridge University Hospitals, Cambridge, UK, [39]British Heart Foundation Centre of Excellence, Division of Cardiovascular Medicine, Addenbrooke's Hospital, Cambridge, UK, [40]Section of Nutrition and Metabolism, International Agency for Research on Cancer, Lyon, France, [41]Department of Hygiene and Epidemiology, University of Ioannina Medical School, Ioannina, Greece, [42]Department of Internal Medicine, Section of Gerontology and Geriatric Medicine, Wake Forest School of Medicine, Winston-Salem, NC, USA, [43]Department of Epidemiology and Biostatistics, Imperial College London, London, UK, [44]Imperial Biomedical Research Centre, Imperial College London and Imperial College NHS Healthcare Trust, London, UK, [45]Medical Research Council-Public Health England Centre for Environment, Imperial College London, London, UK, [46]UK Dementia Research Institute, Imperial College London, London, UK, [47]Health Data research UK-London, London, UK, [48]Department of Epidemiology, University of Washington, Seattle, WA, USA, [49]Department of Internal Medicine, Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands, [50]Department of Neurology, University of Pennsylvania, Philadelphia, PA, USA, [51]Institute for Immunology and Transfusion Medicine, University Medicine Greifswald, Greifswald, Germany, [52]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA, [53]Department of Genetics, Shanghai-MOST Key Laboratory of Health and Disease Genomics, Chinese National Human Genome Center and Shanghai Industrial Technology Institute (SITI), Shanghai, China, [54]Interfaculty Institute of Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany, [55]Chair of Experimental Bioinformatics, Research Group Computational Systems Medicine, Technical University of Munich, Freising-Weihenstephan, Germany, [56]German Center for Cardiovascular Research (DZHK), Partner Site Greifswald, Greifswald, Germany, [57]Department of Clinical Physiology, Tampere University Hospital, Tampere, Finland, [58]Department of Clinical Physiology, Finnish Cardiovascular Research Center - Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland, [59]Laboratory of Complex Trait Genomics, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan, [60]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston,

MA, USA, [61]Department of Medicine, University of Colorado Denver, Anschutz Medical Campus, Aurora, CO, USA, [62]Department of Clinical Chemistry, Fimlab Laboratories, Tampere, Finland, [63]Department of Clinical Chemistry, Finnish Cardiovascular Research Center - Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland, [64]Department of Internal Medicine, University Medicine Greifswald, Greifswald, Germany, [65]Center for Clinical Research and Prevention, Bispebjerg and Frederiksberg Hospital, Frederiksberg, Denmark, [66]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark, [67]Department of Medicine, Division of Cardiology, Duke Molecular Physiology Institute, Duke University Medical Center, Durham, NC, USA, [68]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA, [69]Department of Computational Biology and Medical Sciences, Graduate school of Frontier Sciences, The University of Tokyo, Tokyo, Japan, [70]Division of Molecular Pathology, the Institute of Medical Sciences, The University of Tokyo, Tokyo, Japan, [71]Institue of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Germany, [72]Department of Cardiology, Heart Center, Tampere University Hospital, Tampere, Finland, [73]Department of Cardiology, Finnish Cardiovascular Research Center - Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland, [74]Department of Hematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK, [75]National Health Service (NHS) Blood and Transplant, Cambridge Biomedical Campus, Cambridge, UK, [76]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN, USA, [77]Departments of Epidemiology, University of Washington, Seattle, WA, USA, [78]Department of Health Services, University of Washington, Seattle, WA, USA, [79]Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA, [80]Centre for Population Health Research, University of Turku and Turku University Hospital, Turku, Finland, [81]Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku, Finland, [82]Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku, Finland, [83]Radcliffe Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK, [84]Department of Haematology, Churchill Hospital, Oxford, UK, [85]Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA, [86]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation (formerly Los Angeles Biomedical Research Institute) at Harbor-UCLA Medical Center, Torrance, CA, USA, [87]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA, [88]Department of Medicine, Faculty of Medicine, Université de Montréal, Montreal, Quebec, Canada, [89]Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany, [90]Atlanta VA Medical Center, Decatur, GA, USA, [91]Department of Computer Science, University of North Carolina, Chapel Hill, NC, USA, [92]Quantitative and Computational Biology Section, Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA, [93]Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN, USA, [94]National Health Service Blood and Transplant, Cambridge, UK, [95]Laboratory of Statistical Immunology, Osaka University Graduate School of Medicine, Suita, Osaka, Japan.

[*]These authors contributed equally to this work.
[†]A list of members and their affiliations appears in the **Supplementary Information**.
[§]These authors jointly supervised this work.

Correspondence to:

**Paul L. Auer**
Joseph J. Zilber School of Public Health
KIRC 5067
PO Box 413
Milwaukee, WI, 53205-2557
414-229-2986
pauer@uwm.edu

**Guillaume Lettre**
Montreal Heart Institute
5000 Belanger St
Montreal, Quebec, Canada
H1T 1C8
514-376-3330 ext. 2657
guillaume.lettre@umontreal.ca