

# A genome-wide association study of blood cell morphology identifies cellular proteins implicated in disease aetiology

---

Received: 19 July 2021

---

Accepted: 7 August 2023

---

Published online: 18 August 2023

---

 Check for updates

---

Parsa Akbari <sup>1,2,3,4</sup>, Dragana Vuckovic<sup>2,4,5</sup>, Luca Stefanucci <sup>6,7,8</sup>, Tao Jiang<sup>1,4,9</sup>, Kousik Kundu <sup>2,6</sup>, Roman Kreuzhuber <sup>6</sup>, Erik L. Bao <sup>10,11,12,13</sup>, Janine H. Collins <sup>6,7,14</sup>, Kate Downes<sup>6,7</sup>, Luigi Grassi <sup>6,7,15</sup>, Jose A. Guerrero <sup>6,7</sup>, Stephen Kaptoge<sup>1,4,9</sup>, Julian C. Knight <sup>16</sup>, Stuart Meacham <sup>6,17</sup>, Jennifer Sambrook<sup>6,15,31</sup>, Denis Seyres<sup>6,7,15</sup>, Oliver Stegle<sup>17,18,19</sup>, Jeffrey M. Verboon<sup>10,11,12</sup>, Klaudia Walter <sup>2</sup>, Nicholas A. Watkins<sup>7</sup>, John Danesh<sup>1,2,4,8,9,20</sup>, David J. Roberts<sup>4,21,22,23</sup>, Emanuele Di Angelantonio<sup>1,4,8,9,24</sup>, Vijay G. Sankaran <sup>10,11,12</sup>, Mattia Frontini <sup>6,7,25</sup>, Stephen Burgess <sup>1,3,9</sup>, Taco Kuijpers<sup>26,27</sup>, James E. Peters <sup>20,28</sup>, Adam S. Butterworth <sup>1,4,8,9,20,32</sup> ✉, Willem H. Ouwehand <sup>6,7,29,32</sup> ✉, Nicole Soranzo <sup>2,4,6,8,30,32</sup> ✉ & William J. Astle <sup>3,4,7,32</sup> ✉

Blood cells contain functionally important intracellular structures, such as granules, critical to immunity and thrombosis. Quantitative variation in these structures has not been subjected previously to large-scale genetic analysis. We perform genome-wide association studies of 63 flow-cytometry derived cellular phenotypes—including cell-type specific measures of granularity, nucleic acid content and reactivity—in 41,515 participants in the INTERVAL study. We identify 2172 distinct variant-trait associations, including associations near genes coding for proteins in organelles implicated in inflammatory and thrombotic diseases. By integrating with epigenetic data we show that many intracellular structures are likely to be determined in immature precursor cells. By integrating with proteomic data we identify the transcription factor FOG2 as an early regulator of platelet formation and  $\alpha$ -granularity. Finally, we show that colocalisation of our associations with disease risk signals can suggest aetiological cell-types—variants in *IL2RA* and *ITGA4* respectively mirror the known effects of daclizumab in multiple sclerosis and vedolizumab in inflammatory bowel disease.

Blood cells play vital roles in human physiology, including in oxygen transport, in haemostasis, and in host defence. Many of the biological functions of blood cells, such as thrombotic aggregation and the killing of pathogens or the killing of virally infected cells, are mediated by proteins stored in cell granules that are released into the extracellular

space in response to a stimulus. At present, the functional responses of blood cells to stimuli cannot be measured using high-throughput instruments. Consequently, genetic association studies of cell function traits have been limited to small studies ( $n \leq 5000$ ) of platelet aggregation phenotypes, which have identified associations in

---

A full list of affiliations appears at the end of the paper. ✉ e-mail: [asb38@medschl.cam.ac.uk](mailto:asb38@medschl.cam.ac.uk); [who1000@cam.ac.uk](mailto:who1000@cam.ac.uk); [ns6@sanger.ac.uk](mailto:ns6@sanger.ac.uk); [wja24@cam.ac.uk](mailto:wja24@cam.ac.uk)

approximately thirty loci<sup>1</sup>. Meanwhile, high-power genome-wide association studies (GWAS) of blood cell traits have concentrated on phenotypes in classical complete blood counts (cCBCs)<sup>2–4</sup>. cCBCs are standard clinical reports, which include measurements of the counts (i.e. concentrations) and average volumes of various types of cell in the peripheral blood. However, cCBCs do not measure the properties of intracellular structures, which play important roles in many functional haematological processes. For example, pathologies of leukopoiesis that result in the absence of specific granules from neutrophils can cause immune dysfunction<sup>5</sup>, but are frequently accompanied by a normal neutrophil count and so cannot be detected by cCBCs.

cCBCs are produced by automated haematology analysers, which usually contain a built in flow-cytometer, which can measure the fluorescence and diffraction of laser light incident on individual blood cells. Variation in structural properties of blood cells—many of functional or clinical relevance—can be detected by flow-cytometry. For example, the intensity of light side-scattered (SSC) by a neutrophil is a measure of the cell's intracellular organelle complexity, including its granule content. Neutrophil granularity is an important immune phenotype. Activated neutrophils are known to exhibit greater SSC than resting neutrophils<sup>6,7</sup>. Hypo- and hyper-granulated neutrophils are observed in inflammation and also in myelodysplasia<sup>8–11</sup>. Historically, neutrophil granularity was assessed manually by microscopy of blood smears. Such estimates of granularity correlate well with optical measurements made by haematology analyser flow-cytometers<sup>9</sup>. Indeed, flow-cytometry measured neutrophil SSC is a predictor of neutrophil toxic granulation<sup>9</sup>. More generally, a range of flow-cytometry measured properties of platelets, neutrophils, lymphocytes, and monocytes have been implicated as statistical predictors of clinical outcomes, including thrombocytopenic purpura<sup>12</sup>, sepsis<sup>13–15</sup>, myelodysplastic syndromes<sup>16,17</sup>, Sezary disease<sup>18</sup> and the need for mechanical ventilation in COVID-19 patients<sup>10,19</sup>.

Here, we report the first large-scale GWAS of flow-cytometry measured non-classical CBC (ncCBC) traits (Fig. 1a–e). We call these phenotypes ncCBC traits because they were acquired using a haematology analyser, the Sysmex XN-1000, but are not included in standard cCBC reports (Fig. 1f). The XN-1000 analyser contains a flow-cytometer, which can detect variation in the intracellular organelle complexity of cells from the intensity of SSC diffracted light. It can also detect variation in the nucleic acid content and membrane permeability of cells from the fluorescence intensity (SFL) of light emitted by a cell staining dye and variation in the size of cells from the intensity of forward scattered (FSC) diffracted light (Fig. 1a–e). The analyser reports individual level summaries of the cell level distributions of SSC, SFL and FSC intensity measurements (as averages or distribution widths over a set of cells of a given type). These phenotypes can capture variation in biological processes with functional or clinical relevance<sup>7,9,20</sup>. By integrating the results of GWAS of ncCBC traits with the results of multi-omic and disease GWAS, we show how flow-cytometry phenotypes can be used to identify the secretory origins of proteins in the blood plasma and used to study the role of blood cells in mechanisms mediating disease risk variation.

## Results

### Hundreds of new genetic determinants of blood cell flow-cytometry traits

We studied 63 ncCBC phenotypes in INTERVAL (Supplementary Data 1 and 2), a cohort of blood donors, which has been described previously by Moore et al.<sup>21</sup>. Eleven of the traits explicitly summarise cell-level measurements of intracellular complexity/granularity (SSC), sixteen of the traits explicitly summarise cell level measurements of cell nucleic acid content/membrane lipid content (SFL) and fifteen of the traits explicitly summarise cell-level measurements of cell morphology/volume (FSC) (Supplementary Data 2). Ten of the traits are platelet phenotypes, twenty are red cell phenotypes and 33 are white cell

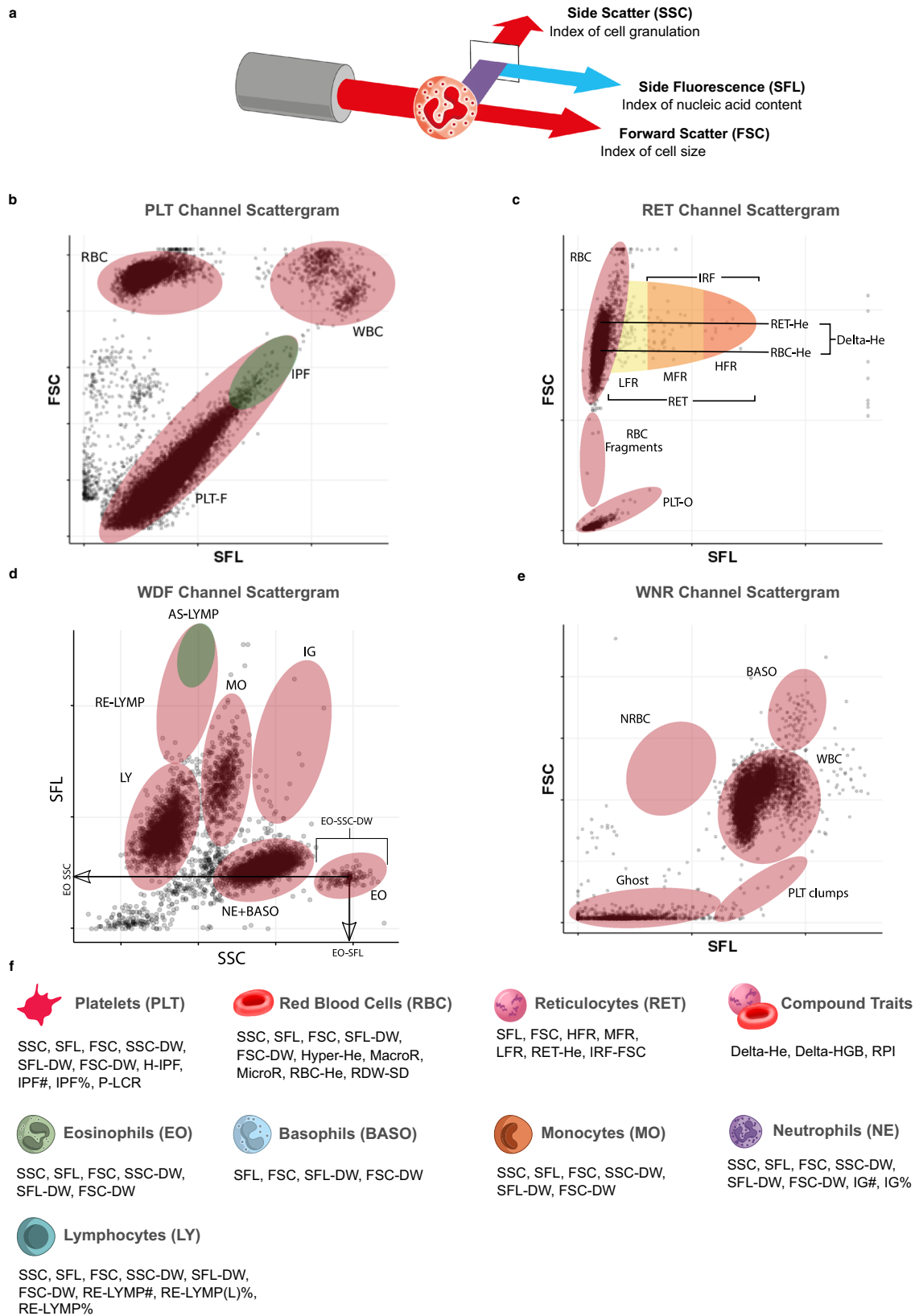
phenotypes. After performing exploratory analysis—investigating the distributions of the traits and their co-distributions with various covariates including, age, sex, menopause and BMI (Fig. 2, Supplementary Figs. 1–3, Supplementary Data 3)—we developed models to regress out extraneous variation (Methods). We also investigated covariation between the traits, observing greater correlation on average between traits specific to red cells or platelets than traits specific to white cells, principally because the white cell traits measure properties of biologically heterogeneous subtypes of cells (neutrophils, basophils, eosinophils, lymphocytes and monocytes) (Supplementary Figs. 4–7).

We performed univariable genetic association analyses (Supplementary Data 4) in subsets of a complete genetic dataset comprising imputed genotypes for 26.8 million variants in 43,059 European-ancestry participants. Stepwise regression analysis applied to each trait in turn identified 2172 distinct (unadjusted conditional  $P$ -value  $< 8.31 \times 10^{-9}$ ) variant-trait associations (Supplementary Data 4). To identify distinct association signals we pooled the conditionally significant variant-trait associations across the traits and applied a standard linkage disequilibrium (LD,  $r^2 > 0.8$ ) based greedy clumping algorithm (Methods). This clustered the 2172 variant-trait associations into 849 clumps, 231 of which represented platelet signals (i.e. contained at least one platelet cell-trait associated variant), 211 of which represented red cell signals and 432 of which represented white cell signals (Fig. 3, Supplementary Fig. 8). We compared these associations to those identified by a European-ancestry GWAS of cCBC phenotypes with a 14-fold larger sample size<sup>3</sup> and a transethnic GWAS of cCBC phenotypes with an 18-fold larger sample size<sup>4</sup>. More than half—242 (56%)—of the 432 white cell association signals were novel—i.e. the corresponding clump did not contain a variant in LD ( $r^2 > 0.8$ ) with a variant reported to be associated with a white cell phenotype by Vuckovic et al. or Chen et al.<sup>3,4</sup> (Supplementary Fig. 9). Whereas fewer than a quarter—56 (24%) and 45 (21%) respectively—of the 231 platelet and 211 red cell signals were novel, according to the same criterion. The enrichment of novel associations in white cell traits may be explained by the fact that white cells exhibit greater biological complexity than platelets or red cells, complexity which is captured by ncCBC flow cytometry traits, but not standard cCBC phenotypes. White cells are nucleated and contain complicated intracellular organelles such as granules and vacuoles which differ according to white cell subtype. Both red cells and platelets are anuclear, but only platelets contain granules, which are generally smaller than those of white cells.

### Identification of granule proteins undiscovered by GWAS of classical blood phenotypes

To identify possible molecular mediators of the cell-trait association signals, we annotated each conditionally significant variant with the genes for which VEP predicted the most severe transcriptional consequence (Methods, Supplementary Data 4). We also performed analyses to identify colocalisations of the associations with eQTL for the corresponding cell-types. Of the conditionally significant associations with a VEP gene annotation that exhibited strong evidence (Bayesian posterior probability,  $PP > 80\%$ ) for colocalisation with an eQTL, 67% had a gene annotation consistent with the transcript of the colocalising eQTL (Supplementary Data 5). An extensive literature search highlighted roles in known fundamental cellular functions including thrombus formation for platelet traits (e.g. *VWF*, *SERPINE2*), iron homeostasis for red cell traits (e.g. *HFE*, *TFRC*) and chemotaxis and adhesion for myeloid white cell traits (e.g. *P2RY2*, *SSH2*) (Fig. 4; Supplementary Data 4).

We identified genes, not previously found by GWAS of granulocyte traits, that code for various white cell granule proteins including Arylsulfatase B, Lactoperoxidase, RNase 3 and multiple defensins. These proteins have well understood roles in immunity<sup>22</sup>, illustrating the potential for GWAS of ncCBC traits to identify genes critical to

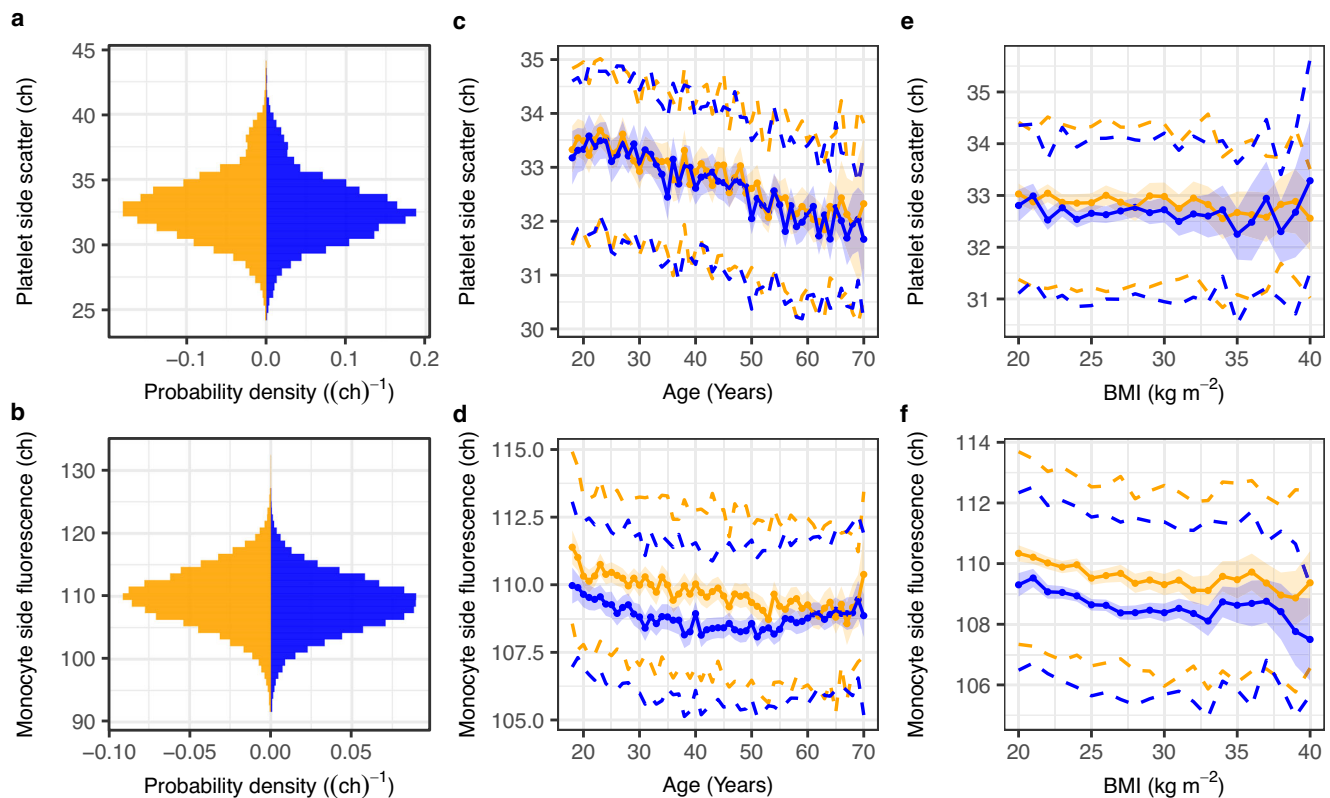


blood cell function. Specifically, we identified 35 associations with NE-SSC (neutrophil granularity), 30 associations with EO-SSC (eosinophil granularity), and twelve associations with MO-SSC (monocyte granularity) that did not correspond to any of the genetic associations with cCBC traits reported by Vuckovic et al. or Chen et al. (Supplementary Data 4)<sup>3,4</sup>. These include associations near genes with well understood

functions in blood cells, including roles in transcription and translation (*AFF1*, *RPL3P2*, *PTBP1*), in exocytosis (which is critical to granule formation and release; *APIM2*, *SMAP1*), as granule cargo (*FCN1*, *HYAL3*, *PRG2*, *RNASE3*, *ARSB*, *LPO*, *DEFA*), as lysozyme cargo (*CTNS*, *HEXB*), and in immune response (*FPRI*, *IFI30*, *KIAA0922*, *MYO1G*, *LGMN*), confirming that our approach has uncovered biologically relevant genes,

**Fig. 1 | Flow-cytometry traits measured by the Sysmex XN-1000 haematology analyser (adapted from Sysmex XN-1000 Manual<sup>104</sup>).** **a** Schematic of a granulocyte cell passing through the laser of the internal flow-cytometer of the analyser. The instrument measures the intensities of incident light scattered sidewise (SSC, cell complexity/granularity) by the cell and forward (FSC, cell volume) by the cell and the intensity of the light which is absorbed by the cell and fluoresced at a new wavelength (SFL, cell nucleic acid content). **b–e** Cytometry scattergrams from an arbitrary participant in the INTERVAL study: 2-dimensional projections of the cell level intensity data (SSC, SFL, FSC) measured in each of the four XN-1000 flow-cytometry channels active for the INTERVAL study: PLT-F (platelet flow) channel (**b**), RET (reticulocyte) channel (**c**), WDF (white cell differential) channel (**d**), WNR

(white cell and nucleated red cell) channel (**e**). Many of the traits correspond to averages or distribution widths (DWs) of cell level measurements in scattergram regions (indicated approximately by ellipses) occupied by cells of particular types. This is illustrated for three eosinophil traits (in panel **d**). Supplementary Data 2 contains a full description of the measurement procedure for each trait. **f** The 63 cytometry traits classified by the type of cells which they measure: platelets (PLT), mature red blood cells (RBC), reticulocytes (RET), neutrophils (NE), eosinophils (EO), basophils (BASO), monocytes (MO) and lymphocytes (LY). The three compound traits (Delta-HE, Delta-HGB, and RPI) depend on measurements of both mature red cells and reticulocytes. We thank Joanna Westmoreland for the artwork in (**a**) and (**f**).



**Fig. 2 | The distributions of selected ncCBC traits and their covariation with age, sex and BMI.** Summary plots for two exemplar technically adjusted traits (Methods) using data from participants who contribute to the GWAS of the respective trait. The upper row and lower row panels correspond respectively to the platelet side scatter (PLT-SSC,  $n = 29,675$ ) and monocyte side fluorescence (MO-SFL,  $n = 39,586$ ) phenotypes. **a, b** Probability density histograms stratified by sex: female (orange) and male (blue). **c, d** Covariation between the phenotype and participant age stratified by sex. Parameters of the stratified trait distributions were

estimated in bins corresponding to years of age. The linearly interpolated coloured points show estimates of the within strata-means and the underlying coloured ribbons show the corresponding 95% confidence intervals. The dashed lines show estimates of the upper and lower quartiles. **e, f** Covariation between each trait and body mass index (BMI). Estimates of sex stratified summaries were made in bins of  $1 \text{ kg m}^{-2}$ . The components of the plots are as for (**c, d**). Analogous plots for all 63 traits are presented in Supplementary Figs. 1–3.

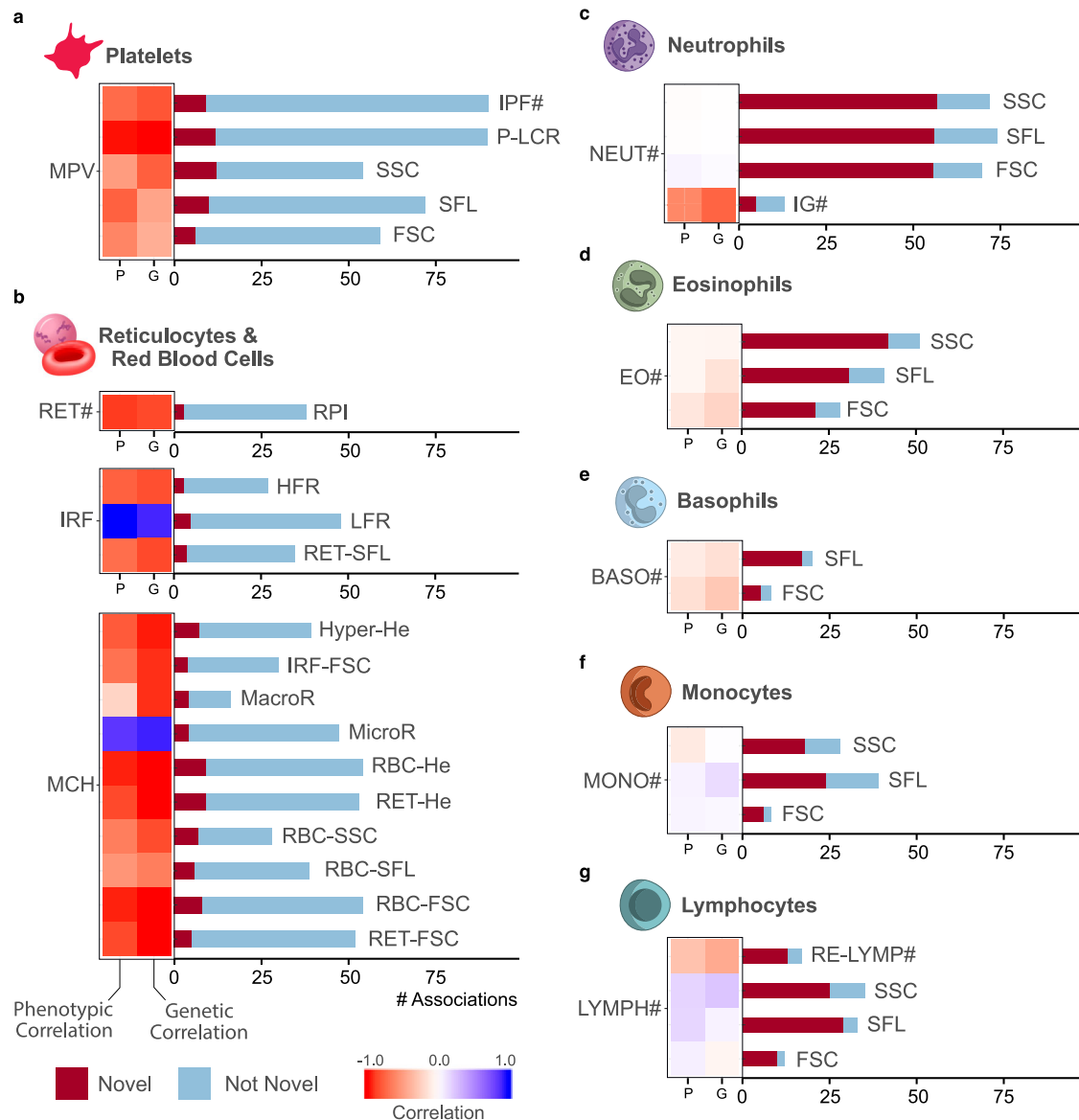
including genes in molecular pathways modulating cellular complexity and granule formation.

We used data from mass spectrometric profiling of neutrophil granules<sup>23</sup> to identify the subcellular localisation of proteins expressed by the genes identified by the neutrophil trait GWAS in a wide variety of organelles (Supplementary Data 6). Thirteen clumps contained neutrophil trait associated variants annotated by VEP with genes coding for proteins expressed in azurophilic granules and two clumps contained neutrophil trait associated variants annotated with genes coding for proteins in specific or secondary granules. Cellular granules are also prominent in eosinophils, basophils, and present in monocytes where they play critical roles in innate immune responses by releasing antimicrobial proteins. We identified signals near genes encoding several such proteins,

including *HYAL3*, *PRG2*, *PRTN3* and *RNASE3* (Supplementary Data 4)<sup>22,24–26</sup>.

### The biogenesis of cellular structures

Many of the intracellular structures generating variation in flow-cytometry phenotypes have developmental origins in the immature precursors of peripheral blood cells. Granule formation, for example, is a cell-type specific process occurring at particular stages of cellular differentiation; the granules of platelets and the granules of granulocytes begin to form in megakaryoblasts and myeloblasts, respectively<sup>27</sup>. Consequently, the absence of a colocalizing mature blood cell eQTL for a cytometry trait association may reflect the fact that the associated genetic variant exerts its variance-generating effect after lineage commitment, but before terminal differentiation. To test



**Fig. 3 | The distribution and novelty of association signals by cell-type.**

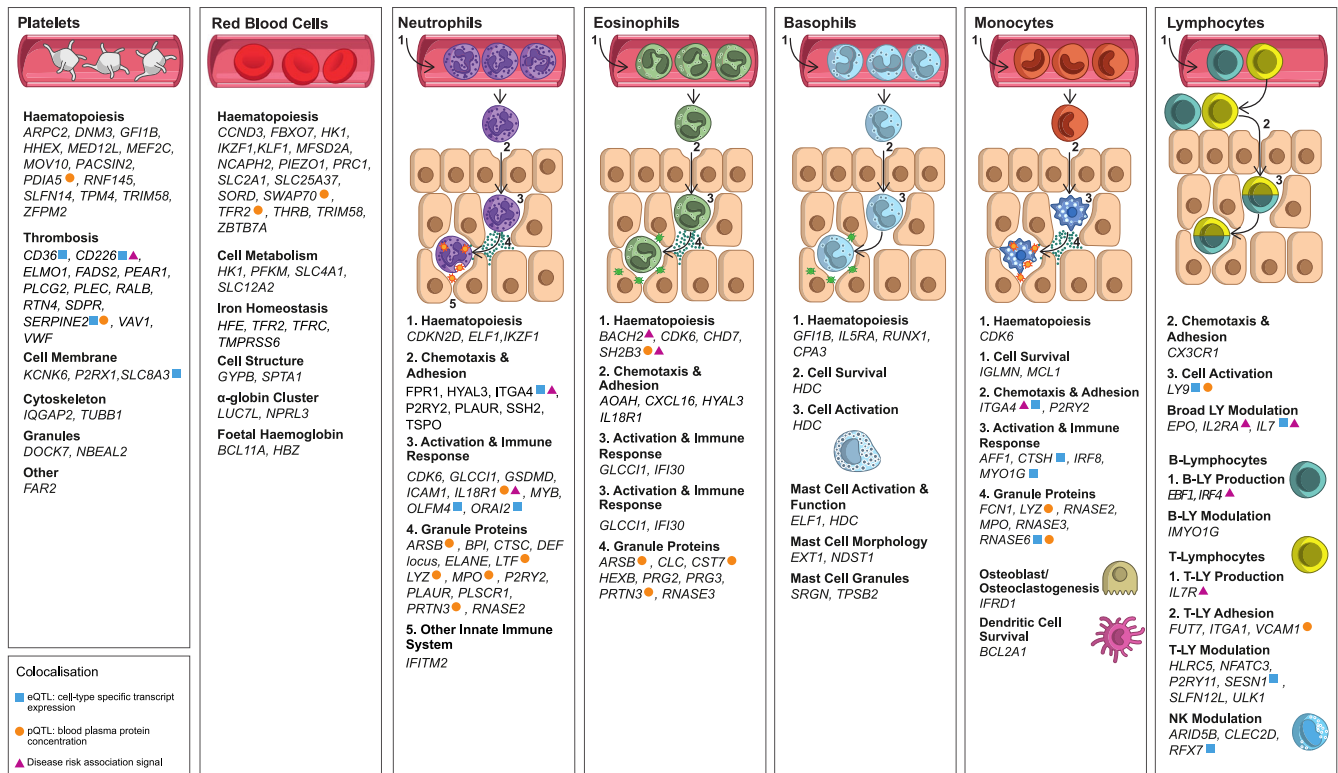
**a–g** Each panel presents statistics for selected ncCBC traits of the given cell-type. The heat map on the left of each subplot shows the estimated phenotypic (left) and genetic (right) correlation between the ccBC trait indicated to its left and the ncCBC trait corresponding to each horizontal bar. (Each ncCBC trait has been grouped with the ccBC trait studied in Vuckovic et al.<sup>3,4</sup> with which it has maximal absolute phenotypic correlation in the study sample.) The bar plot on the right of each subplot indicates the number of distinct (conditionally significant)

associations identified for each ncCBC trait and the number of distinct associations with variants that do not fall into a LD clump with a variant reported to be associated with a blood trait of the same cell-type by Vuckovic et al. or Chen et al. ('Novel')<sup>3,4</sup>. The absolute genetic correlations between the ncCBC and ccBC traits of white cells are lower than those of red cells and platelets. This is reflected in the variation between cell-types of the proportion of identified associations that are novel. We thank Joanna Westmoreland for the artwork in (a–g).

this hypothesis, we applied FINEMAP 1.3.1<sup>28</sup> to identify Bayesian posterior credible sets of variants associated with the ncCBC traits and assessed the enrichment of the variant sets in the nucleosome depleted regions (ATAC-seq) of nine types of progenitor cell (localised in the bone marrow) and nine types of the mature cell (generally localised in the peripheral blood)<sup>29</sup>.

We observed patterns of enrichment in the open chromatin regions of the progenitor cell-types ancestral to neutrophils, eosinophils, monocytes, and lymphocytes (Supplementary Fig. 10a–e) which suggest the possible developmental stages (Supplementary Fig. 10f) at which the cell characteristics corresponding to particular traits (SSC, SFL, FSC) are likely to develop. For instance, neutrophils contain three classes of cytotoxic granules—azurophilic, specific and gelatinase—which are formed sequentially at distinct stages of differentiation<sup>27</sup>. The

relative enrichments of genetic variants associated with NE-SSC (neutrophil granularity) in the nucleosome-depleted regions of the hematopoietic stem cell (HSC) and the four types of myeloid progenitor cell (CMP, GMP-A, GMP-B, GMP-C) are consistent with a progressive increase in the accessibility of enhancers regulating granule formation during myeloid differentiation and point to an origin of these granules in lineage-committed myeloid progenitors. In monocytes, we observed an enrichment of MO-SFL (monocyte nucleic acid content) associated variants in the nucleosome-depleted regions of granulocyte-macrophage progenitor cells (GMP)—the differentiation phase of proliferation and cell division—and an enrichment of MO-SSC (granularity) in the nucleosome depleted regions of peripheral blood monocytes, suggesting that granularity may be regulated in the ultimate stages of monocyte differentiation before egress from the bone marrow.



**Fig. 4 | Summary of the biological functions of the genes assigned to associated variants identified from a survey of the literature.** Each panel contains a list of genes assigned by VEP or by eQTL/pQTL colocalisation to genetic associations with traits corresponding to the given cell-type, for which a literature search identified evidence of known function. Each list is stratified into functional categories relevant to the cell-type. Supplementary Data 4 contains a complete list of

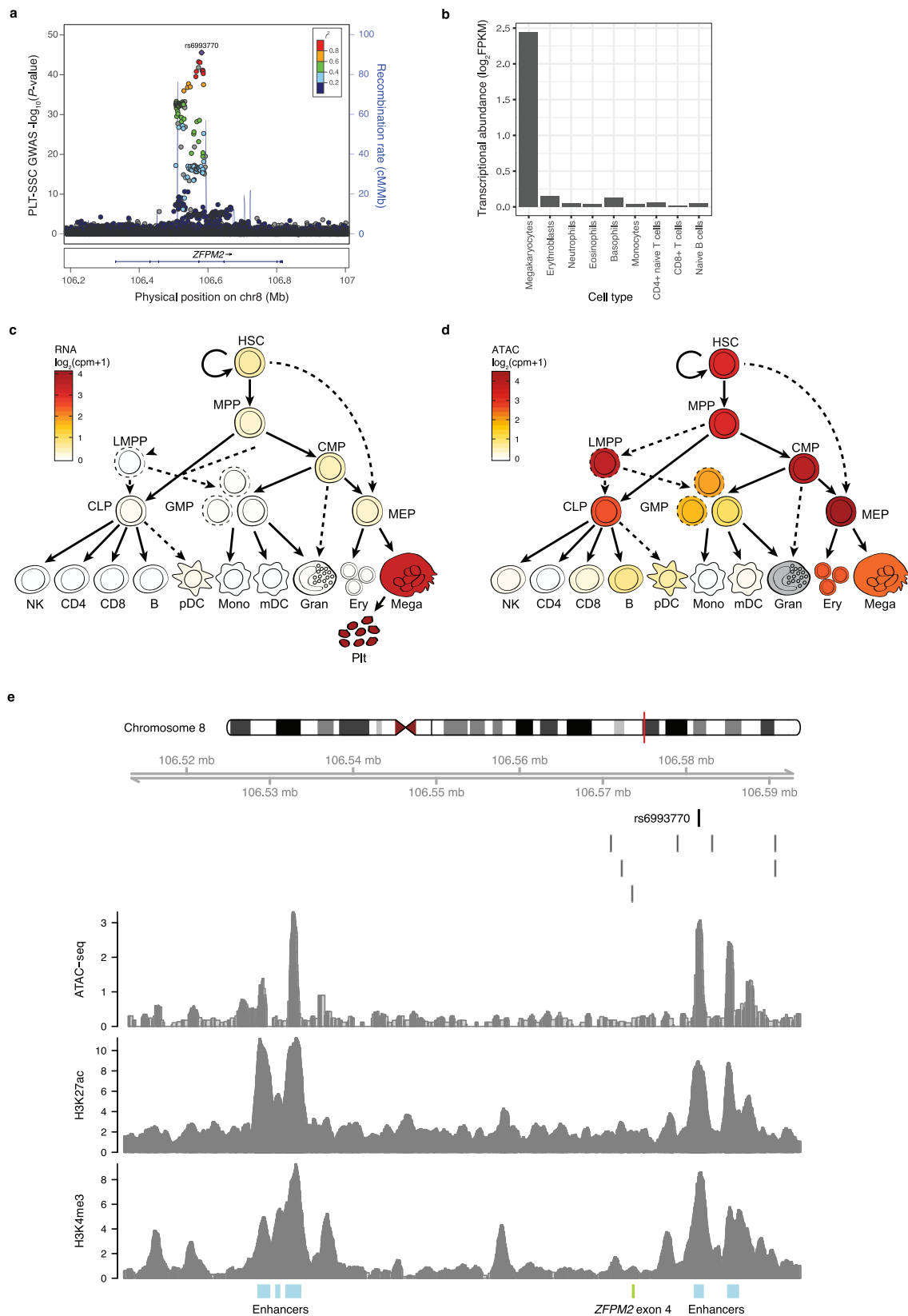
the associated variants, their VEP annotated genes, and relevant references to literature. The coloured symbolic annotations indicate genes assigned to variants which colocalise with eQTL (blue square), pQTL (orange circle), or disease GWAS associations (purple triangle). The gene(s) assigned by eQTL or pQTL colocalisation occasionally differ from the gene(s) assigned by VEP. We thank Joanna Westmoreland for the artwork.

### The cellular origins of plasma proteins

We hypothesised that some genetic variants associated with cytometric traits, in particular genetic variants associated with side scatter traits, which capture the abundance of secretory granules in cells, also influence the concentration of secretory proteins in the blood plasma. To explore this, we used the results of Sun et al., a GWAS of plasma concentrations of 1478 proteins, which identified 1927 associations (protein quantitative trait loci, pQTL)<sup>30</sup>. For 943 of these proteins, there is strong evidence that transcripts that are expressed ( $\log_2\text{FPKM} > 1.0$ , i.e.  $\log_2$  of fragments per kilobase of transcript per million mapped fragments is greater than 1.0) in at least one of the blood cell-types surveyed by the ncCBC phenotypes (megakaryocytes (MKs) and erythroblasts—the respective progenitors of platelets and red cells—and neutrophils, eosinophils, basophils, monocytes, CD4<sup>+</sup> T cells, CD8<sup>+</sup> T cells, naive B cells). We performed colocalisation analysis between the pQTL and our cytometry trait-associated conditionally significant variants and identified 61 and 1021 colocalisations (PP > 80%) with cis and trans pQTL, respectively (Fig. 4, Supplementary Data 7). There were 181 proteins with a pQTL that colocalised with ncCBC trait association signal for just one cell-type suggesting that blood cells differentially contribute to the plasma proteome (e.g. *VEGFA*, which encodes vascular endothelial growth factor A, is associated with eight different platelet traits, while *RNASE6*, which encodes RNase K6, is associated solely with monocyte side fluorescence) (Supplementary Data 4). Notably, many associations with variants assigned by VEP (gene of worse consequence, Methods) to genes encoding granule proteins colocalised with pQTL for the corresponding proteins in blood plasma (Fig. 4). Examples include *ARSB* (PP = 99%), *LY9* (PP = 98%), *MPO* (PP = 100%), *PRTN3* (PP = 100%) and *RNASE6* (PP = 99%) (Supplementary Data 7).

### Evidence that *FOG2* is a regulator of platelet $\alpha$ -granularity

Platelet activation is important in thrombus formation, wound healing, inflammation and the chemotaxis and activation of myeloid white cells. Critical to these biological processes are coagulation proteins, growth factors, proteases, chemokines and other signalling peptides that diffuse into the blood plasma when  $\alpha$ -granules are released by activated platelets. Our GWAS of PLT-SSC (platelet granularity) identified an association in *ZFPM2* (encoding the transcription factor Friend of GATA-2 or *FOG2*) (Fig. 5a), colocalizing with trans-pQTL for 24 plasma proteins, of which thirteen are platelet  $\alpha$ -granule localised proteins (Supplementary Data 6). RNA-seq data from nine differentiated nucleated blood cell-types<sup>31</sup> showed that transcripts of *ZFPM2* are substantially expressed solely in MKs (Fig. 5b). Complementary data from a study of the entire hematopoietic system showed that *ZFPM2* is specifically expressed in the MK lineage (Fig. 5c)<sup>29</sup>. A stepwise multiple regression analysis of the platelet granularity phenotype (PLT-SSC) on variants in the *ZFPM2* locus suggested a single conditionally significant association, and fine-mapping of the locus identified a single intronic SNP (rs6993770) as the most probable causal variant (PP = 95%). rs6993770 is located in a region of open chromatin (ATAC-seq) in MKs, which contains histone modifications indicative of a transcriptional enhancer (H3K4 tri-methylation, H3K27 acetylation; Fig. 5d, e). The variant is located 25 b upstream of a GATA motif on the negative strand, and 34 b upstream from the palindromic E-box binding motif CAGCTG. The juxtaposition of these motifs is characteristic of a hematopoietic co-binding site for GATA-1 and TAL1, two of the three key MK lineage determining transcription factors<sup>32–34</sup>. None of the seven variants in high LD ( $r^2 \geq 0.9$ ) with rs6993770 were located in regions for which epigenetic data supported causality in MKs (Fig. 5e).

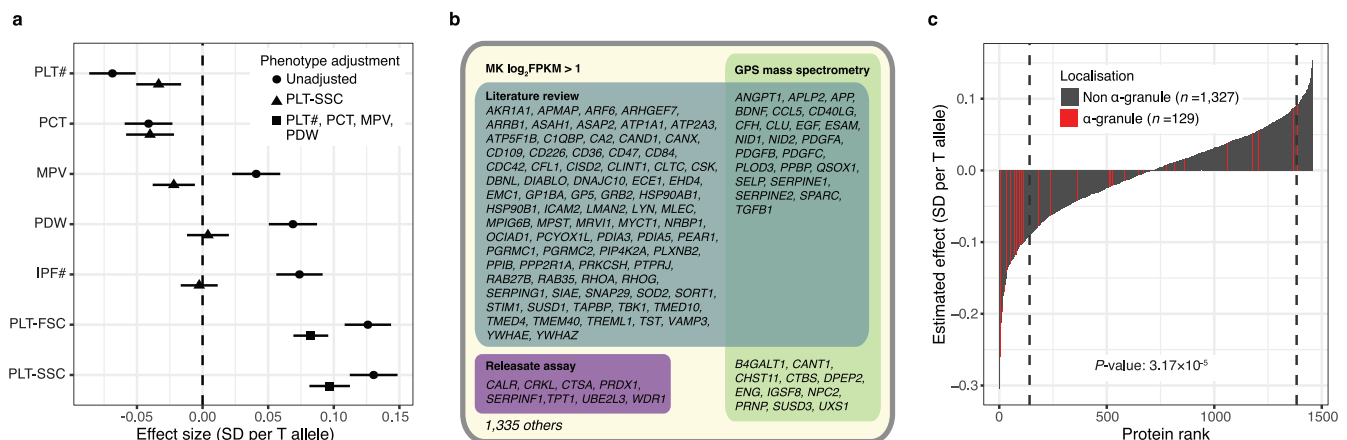


There are well-known associations between rs6993770 and the four platelet cCBC traits: platelet count (PLT), volume (MPV), crit (PCT) and volume distribution width (PDW)<sup>2,35</sup>. We identified new associations with the ncCBC traits IPF#, PLT-SSC, and PLT-FSC (Supplementary Data 4, Fig. 6a). The estimated effect of rs6993770 on mean

PLT-SSC was not significantly attenuated by a multivariable linear adjustment for the four cCBC traits, suggesting that the PLT-SSC association is independent of the association with classical platelet phenotypes. Consequently, we hypothesised that *ZFPM2* plays a role in  $\alpha$ -granule biology. To investigate this, we considered the effect of

**Fig. 5 | The association of rs6993770 with PLT-SSC is mediated by *ZFPM2* expression.** **a** A LocusZoom plot for the *ZFPM2* locus<sup>105</sup>. Each dot corresponds to a variant tested for association. The x-axis represents the physical position on chromosome 8 in GRCh37 coordinates. The (left-hand) y-axis represents the  $-\log_{10}(P\text{-value})$  from a univariable BOLT-LMM test for additive allelic association between the imputed genotypes of the variant and PLT-SSC ( $n = 29,675$ ). The colour of the dot represents the LD ( $r^2$ ) in the study sample between the corresponding variant and rs6993770. The blue line represents an estimate of the local recombination rate (right-hand y-axis). Conditional analysis identified a single association signal in the 82 kb interval of low recombination containing rs6993770. **b** The abundance of *ZFPM2* transcripts ( $\log_2\text{FPKM}$ ) in MKs, erythroblasts, neutrophils, eosinophils, basophils, monocytes, CD4+ naive T cells, CD8+ T cells, and naive B cells, in which cell-types *ZFPM2* transcription is limited to MKs<sup>31</sup>. **c** *ZFPM2* transcript expression is higher in platelets, MKs and their precursor cell-types—MEP (megakaryocyte-erythroid progenitor cells), CMP (common myeloid progenitor), MPP (multipotent progenitor), and HSC (hematopoietic stem cell)—than in other blood cell and blood cell precursor cell-types. **d** ATAC-seq applied to multiple blood cell-

types show that rs6993770 lies in an open chromatin region in the platelet precursor cell-types MK, MEP, CMP, MPP and HSC. **e** Measurements of epigenetic activity in MKs across the 82 kb recombination interval containing the association signal. The x-axis represents the physical position on chromosome 8. The dark vertical line indicates the position of rs6993770. The nearby light vertical lines indicate the locations of seven variants in high LD ( $r^2 > 0.9$ ) with rs6993770. The y-axis of each panel corresponds to the sequencing read depth of an epigenetic assay. From top to bottom the panels correspond to ATAC-seq (open chromatin), H3K27ac (a mark of active enhancers) and H3K4me3 (a mark of accessibility to transcription factors). The blue rectangles at the bottom of the figure indicate enhancer regions in MKs inferred from a set of six histone modifications (H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3 and H3K36me3) using the IDEAS chromatin segmentation algorithm<sup>106,107</sup>. The green rectangle indicates the position of exon 4 of *ZFPM2*. Panels c-d are adapted with permission from Ulirsch, J. C. et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* 51, 683–693 (2019), Springer Nature<sup>29</sup>.



**Fig. 6 | *ZFPM2* is a regulator of platelet  $\alpha$ -granularity.** **a** Forest plot showing the additive allelic effect of rs6993770-T on the means of the inverse rank normalised distributions of the platelet traits PLT# (platelet count,  $n = 29,657$ ), PCT (plateletcrit,  $n = 28,044$ ), MPV (mean platelet volume,  $n = 28,050$ ), PDW (platelet distribution width,  $n = 28,052$ ), IPF# (immature platelet fraction count,  $n = 30,587$ ), PLT-FSC (platelet volume,  $n = 29,662$ ), and PLT-SSC (platelet granularity,  $n = 29,675$ ) measured in the INTERVAL study. Circles correspond to estimates of direct effects, triangles correspond to estimates of effects adjusted for PLT-SSC and squares correspond to estimates of effects adjusted for PLT#, PCT, MPV, and PDW. The horizontal lines correspond to 95% confidence intervals. The effect of rs6993770-T on PLT-SSC and PLT-FSC does not appear to be mediated substantially through the four cCBC phenotypes. **b** A Venn diagram cross-classifying the 1456 genes coding for proteins studied by Sun et al. that are expressed in MKs (mRNA transcript

$\log_2\text{FPKM} > 1$ )<sup>30</sup>. The classifying categories indicate that the protein was implicated as an  $\alpha$ -granule protein coding gene by one of: a literature review (turquoise), detection by mass spectrometry of significant under expression in the platelets of grey platelet syndrome patients (which lack  $\alpha$ -granules) compared to those of healthy volunteers (green), identification in the platelet releasate—proteins expelled from activated platelets (purple). **c** The estimated per allele effect of rs6993770-T on the mean concentration of the 1456 plasma proteins. The y-axis measures the per allele effect size and the x-axis its rank. Bars corresponding to proteins localised to platelet  $\alpha$ -granules are coloured red. Proteins with ranks in the tails bounded by the dashed lines exhibit significant evidence for an association with rs6993770-T at a relaxed critical threshold (unadjusted  $P\text{-value} < 10^{-3}$ ).  $\alpha$ -granule proteins are significantly (embedded two-sided Fisher's exact test unadjusted  $P\text{-value}$ ) enriched in the negative compared to the positive tail.

rs6993770 on the mean plasma concentrations of 1456 of the proteins studied by Sun et al., for which there was evidence of expression in MKs (RNA-seq  $\log_2\text{FPKM} > 1$ , Fig. 6b)<sup>30</sup>. rs6993770-T had a significant effect on the mean plasma concentration of 215 of these proteins at a relaxed threshold (unadjusted  $P\text{-value} < 10^{-3}$ )<sup>30</sup>. Of the 215 associated proteins, 44 were localised to  $\alpha$ -granules and 171 were not. The direction of the effect size estimates of Sun et al. imply that rs6993770-T reduces the mean plasma concentration of 40 (91%) of the 44  $\alpha$ -granule localised proteins but only 101 (59%) of the 171 proteins not localised to  $\alpha$ -granules (Fig. 6c). This indicates that rs6993770-T significantly differentially reduces the plasma concentration of proteins expressed in MKs according to their platelet  $\alpha$ -granule localisation (Fisher's exact test unadjusted  $P\text{-value} = 3.17 \times 10^{-5}$ ).

To control for the possibility that the differential effect of rs6993770-T might be explained by differences in expression levels correlated with protein  $\alpha$ -granule localisation, we regressed the plasma protein concentration estimated effect sizes of rs6993770-T on

a dummy variable indicating  $\alpha$ -granule localisation, adjusting for the abundance of the mRNA transcripts in MKs, using data corresponding to the 1456 proteins expressed in MKs ( $\log_2\text{FPKM} > 1$ ). We estimated the additive allelic effect of rs6993770-T on plasma proteins to be 0.038 phenotypic standard deviations lower on average (two sided  $t$ -test unadjusted  $P\text{-value} = 5.6 \times 10^{-10}$ ) in proteins localised to  $\alpha$ -granules compared to other proteins, adjusting for gene expression in MKs (Supplementary Fig. 11). This analysis suggests that in addition to reducing PLT-SSC, rs6993770-T preferentially reduces the concentration of platelet  $\alpha$ -granule proteins in plasma over other platelet expressed proteins, thus implicating the *ZFPM2* in platelet granule regulation.

We studied the role of *ZFPM2* in platelet production using an in vitro model of megakaryopoiesis<sup>32</sup> applied to two independent *ZFPM2* knockouts (KO) (Methods, Supplementary Fig. 12a). In this model, the ablation of *ZFPM2* resulted in a drastic reduction in megakaryopoiesis (Supplementary Fig. 12b, c). Because of the dramatically limited



production of MKs by the KO lines, it was not possible to test the effect of *ZFPM2* KO on Sysmex platelet traits or perform further biomolecular analysis. However, our results suggest that expression of *ZFPM2* plays a critical role in platelet biology, with important consequences for platelet biogenesis and  $\alpha$ -granule regulation.

Interestingly Klarin et al., recently reported an association between the T allele of rs6993770 and decreased risk of venous thromboembolism (VTE)<sup>36</sup>. They postulated that FOG2 mediates a decrease in the plasma concentration of the principal inhibitor of plasminogen activator PAI-1, which is encoded by *SERPINE1*<sup>37</sup>. The notion that lower levels of plasma PAI-1 cause a reduced risk of VTE is biologically plausible. However, we have shown that rs6993770 is pleiotropic, modifying the process of platelet formation and platelet granule content and the concentrations of many platelet  $\alpha$ -granule derived proteins in the blood plasma (Fig. 6), highlighting the importance of broader multi-trait and multi-omic integrative analysis, for aetiological inference and the characterisation of disease risk association signals.

### Flow-cytometry phenotypes, disease aetiology and drug targets

We sought to explore more generally how genetic associations with ncCBC traits can improve or confirm our understanding of the role of blood cells in disease aetiology. We chose to focus on immune, inflammatory and cardiovascular diseases, in which blood cells are known to play a critical role. Firstly, we retrieved publicly available GWAS summary statistics for 15 diseases<sup>38</sup> and assessed the evidence for colocalisation between genetic associations with ncCBC traits and disease risks (Supplementary Data 8)<sup>39</sup>. We found strong evidence (PP > 80%) for colocalisation between 73 of the variant-ncCBC trait associations—corresponding to 29 variant clumps—and at least one disease association (Supplementary Data 8).

Eight variant clumps contained variants associated with lymphocyte traits that colocalised with genetic associations for the risk of multiple sclerosis (MS), coeliac disease, primary biliary cirrhosis, hay fever/rhinitis or coronary artery disease (Supplementary Data 8). Five of these eight clumps contained variants with lymphocyte trait associations that colocalised with associations for risk of MS, recapitulating the known importance of lymphocytes in the aetiology of MS<sup>40</sup>. The colocalizing associations were located in the gene encoding the transcription factor *BACH2*, in the genes encoding receptors for Interleukin(IL)-2 (*IL2RA*) and IL-7 (*IL7R*) and in the gene encoding IL-7 itself (*IL7*). IL-2 receptor  $\alpha$ -chain (the product of *IL2RA*) is the target of the therapeutic antibody Daclizumab, which is known to be clinically effective in the treatment of MS<sup>41,42</sup>, but has been withdrawn due to severe side effects including encephalitis<sup>43–46</sup>.

Alternative therapeutic approaches to MS are required to ameliorate the risk that subsets of patients develop severe side effects on existing drugs. The IL-7 receptor (the product of *IL7R*) is one possible target<sup>47,48</sup>, and genetic evidence suggests that a higher serum concentration of soluble IL-7R increases the risk of multiple sclerosis<sup>49</sup>. A non-synonymous coding variant in *IL7RA* (rs6897932-C) effects alternative splicing causing a two-fold increase in the skipping of exon 6 resulting in production of soluble IL-7R<sup>50,51</sup>. rs6897932-C is in high LD ( $r^2=1$ ) with rs11567705-C (AF=72.5%) which is associated with increased reactive lymphocyte count (RE-LYMP#, additive allelic effect size=0.09 SD, unadjusted  $P$ -value =  $4.4 \times 10^{-21}$ ). This association with RE-LYMP# colocalised with increased risk of MS (PP = 95.7%). Thus our analysis suggests that the effect of soluble IL-7R on MS aetiology may be mediated by modulation of lymphocyte activation as measured by RE-LYMP#, consistent with the role of IL-7R/IL-7 as an efficacious target for the treatment of MS<sup>50</sup>.

Further examples demonstrate the ability of ncCBC GWAS associations to indicate which blood cell-types may mediate disease aetiology. For instance, rs2124440-A is associated with properties of monocytes (MO-SFL-DW, additive allelic effect size = 0.048 SD,

unadjusted  $P$ -value =  $3.4 \times 10^{-11}$ ; MO-SFL additive allelic effect size =  $-0.034$  SD, unadjusted  $P$ -value =  $2.0 \times 10^{-6}$ ) and these associations colocalised with lower expression of *ITGA4* in monocytes (CD14, PP = 99.7%) and higher inflammatory bowel disease (IBD) risk (PP = 98.2%). This finding is interesting because Vedolizumab was originally used to treat IBD<sup>52,53</sup> and was assumed to reduce trafficking of  $\alpha 4\beta 7$ -positive gut-specific T-helper lymphocyte by diminishing integrin interaction with the mucosal addressin cell adhesion molecule 1 (MadCAM-1)<sup>54,55</sup>. However, a recent study has suggested that the antibody also reduces the ability of monocytes to egress into the colonic mucosa as an alternative mechanism<sup>55–57</sup> consistent with the results of our genetic analyses of monocyte flow-cytometry traits.

### Discussion

Over the last decade, GWAS have identified thousands of genetic associations with risks of common complex diseases. A major motivation for these studies has been to improve our understanding of the molecular and cellular mechanisms underlying disease aetiology, in order to develop safe and effective pharmacological treatments. Unfortunately, only a fraction of the biological mechanisms underlying the genetic associations with disease risk that have been identified are presently understood. Sometimes even the mediating cell-types and tissues are unknown. One approach to understanding a genetic association with disease risk, is to consider the context of colocalizing genetic associations with other phenotypes that have been published in the literature or online catalogues, a so-called ‘phenoscan’<sup>58,59</sup>. Genetic associations with quantitative traits measuring cell-type specific biological variation (for example, eQTL identified from studies of gene expression variation) can suggest mediating tissue types. Consequently, there is a need to catalogue genetic associations with phenotypes measuring cell-type specific biological variation at different levels of complexity, from variation in molecular abundances to variation in the properties of organelles.

We have studied a new class of flow-cytometry phenotypes, which can capture cell-type specific quantitative variation in the functional structures of blood cells, such as secretory granules. These phenotypes can be measured affordably in tens of thousands of study participants using clinical haematology analysers. GWAS analyses of these traits showed that they are heritable, have complex genetic architectures, and are affected by variation in genes implicated in a variety of molecular and cellular pathways. The broad biological sensitivity of flow-cytometry phenotypes means that their genetic associations have a coarser interpretation than those of many other intermediate quantitative phenotypes, such as molecular abundances (e.g. transcriptomics, proteomics or lipidomics). For example, cytometry measured ‘neutrophil granularity’ depends on the average number of granules in peripheral blood neutrophils, and the level of expression of genes coding for proteins in neutrophil granules (e.g. *ELANE*, *MPO*, *PRTN3*). rs138303849-C, a variant upstream of *PRTN3* (encoding proteinase-3, PR3) is associated with increased NE-SSC (granularity), NE-SFL (nucleic acid content), NE-FSC (volume), and decreased NE-SSC-DW (SSC distribution width). The allele, which is known to be associated with an increased risk of a form of vasculitis characterised by autoantibodies against PR3<sup>60</sup>, is an eQTL for *PRTN3* in whole blood<sup>61</sup>. We recently showed that the risk allele is associated with increased PR3 concentration in plasma<sup>30</sup>. In the present analysis, we observed that the risk allele and the cis-pQTL for higher plasma PR3 levels all colocalise (PP = 99.9%) with the neutrophil ncCBC associations. This suggests that the vasculitis risk allele not only increases *PRTN3* transcription and PR3 plasma protein levels, but also changes the structural and granular properties of neutrophils.

We have shown more generally how genetic associations with the new traits can be used to interpret colocalizing associations with disease risks and provide evidence to support drug targets, through examples which are corroborated by existing evidence in the

literature, including the role of lymphocyte activation in IL-7R/IL-7 mediated risk of MS and the role monocytes *ITGA4* mediated risk of inflammatory bowel disease. We have also shown that genetic associations with flow-cytometry traits can identify key genes regulating the formation and retention of intracellular structures. In particular, our multi-omic analysis of the association in *ZFPM2* with PLT-SSC (platelet granularity), has shown that the transcription factor FOG2 is a probable regulator of platelet  $\alpha$ -granularity, which influences the concentrations of a multitude of  $\alpha$ -granule proteins in the blood plasma.

Our study has some limitations. Firstly, because participants in the INTERVAL cohort are overwhelmingly of European ancestry, we have not been able to study genetic variation extraneous to the European ancestry population. Secondly, because most of the ncCBC traits are not intended for clinical use, they are subject to greater technical variability than cCBC traits. Nevertheless, we are confident that we have been able to remove much of this variation statistically. Genetic studies, which seek to estimate differences in phenotypic averages between genotype groups, do not rely on the calibration of phenotypes against an absolute standard. However, investigators who wish to perform non-genetic studies of ncCBC traits should be aware of the potential for between instrument variation. Thirdly, because we have no information about the medications taken by the participants, we were unable to exclude the possibility of confounding due to differential prescribing by genotype. However, we expect that such confounding is unlikely, given the general good health of blood donors and the fact that most common genetic variants have a modest effect on disease risk and hence prescribing risk. Fourthly, because the ncCBC phenotypes are novel to genetics, we do not have access to a replication dataset. However, CBC trait associations identified in INTERVAL using the same analysis protocol have a high rate of replicability in other datasets (e.g. UK Biobank). This gives us confidence in our results, although the ultimate validation of any genetic association must rely on cellular and functional laboratory follow up. Finally, although flow-cytometry traits are able to capture aspects of biological variation related to cell function, they lack the clean biological interpretation of phenotypes designed to measure particular cellular mechanisms, such as the response of cells to perturbation by an agonist. In future, the development of efficient cell-type specific assays may enable large-scale population studies of isolated functional processes such as activation, degranulation and cell motility likely to play a role in the aetiologies of cardiovascular and immune disorders.

## Methods

### INTERVAL study

The INTERVAL study was a randomised trial of approximately 45,000 blood donors aged eighteen years or older who were recruited at 25 NHSBT (National Health Service Blood and Transplant) static donor centres across England<sup>21</sup>. The study was approved by the Cambridge East Research Ethics Committee and we complied with the relevant ethical regulations. Informed consent was obtained from all participants during recruitment. Individuals who have suffered recent illness or infection are ineligible to donate blood. Consequently, the participants were predominantly healthy at the time of recruitment. The participants completed a baseline survey which included questions about their lifestyle and state of health, including their smoking habits, their alcohol consumption habits, whether they suffered from doctor-diagnosed anaemia, their use of medication (hormone replacement therapy, iron supplements) and their menopausal status. At baseline, participants were randomly allocated to a sex specific trial arm, according to which men were asked to donate blood every twelve, ten or eight weeks and women were asked to donate blood every sixteen, fourteen or twelve weeks. Participants gave blood samples for research purposes at baseline and at a follow up visit approximately two years later. In the intervening period they were asked to donate blood

according to the schedule of their allocated trial arm. The blood samples given for research were collected from a pouch attached to a standard blood collection unit. The samples for CBC analysis were stored in 3 ml EDTA tubes which were inverted three times before being transported in rigid boxes at ambient temperature, via three NHSBT holding sites (Manchester, Colindale, and Bristol), to UK Biocentre, Stockport, Greater Manchester. Here, extended clinical blood count reports, which include measurements of cCBC and ncCBC traits, were generated using a pair of Sysmex XN-1000 analysers. Respectively, 72% and 98% of samples were processed within 24 h and 48 h of venipuncture.

### Sysmex XN-1000 automated haematology analysis

The Sysmex XN series of haematology analysers can be configured to use various combinations of seven measurement channels: an impedance channel, a photometric channel and five flow-cytometry channels. The analysers used for the INTERVAL study were configured to use the following six channels:

- RBC/PLT (red cell and platelet counts/volumes using Coulter's impedance principle)
- HGB (photometric measurement of total blood HGB)
- WNR (white cell and nucleated red cell flow-cytometry)
- WDF (white cell differential flow-cytometry)
- RET (reticulocyte and red cell flow-cytometry)
- PLT-F (platelet flow-cytometry).

The channels used to measure each of the 63 traits we studied are indicated in Supplementary Data 2. These traits include all the continuously distributed (i.e. without a discrete component) variables reported by the analysers (when configured as described) that are not cCBC traits.

The analyser generates an aliquot of blood for each measurement channel. The flow-cytometry channel aliquots are diluted (1:200 for RET/PLT-F channels, 1:60 for other channels) and pre-treated with channel-specific reagents which differentially lyse and perforate cells by type. The aliquots are then treated with channel specific reagents containing dyes that bind to nucleic acids in organelles and the nucleus. The PLT-F flow-cytometry channel uses a reagent containing an oxazine based dye (Fluorocell PLT), while the other flow-cytometry channels use reagents containing a polymethine dye (Fluorocell WNR, Fluorocell WDF and Fluorocell RET). Each flow-cytometry channel measures three cell-level properties of each particle in the blood aliquot: side fluorescence, forward scatter and side scatter. Side fluorescence is a measure of cell nucleic acid content, but also depends on cell membrane lipid content; forward scatter is a measure of cell size; side scatter is a measure of cell granularity. The cell-level measurements are used by the analyser to classify cells by type (Fig. 1). The classification algorithms are commercially confidential. Individual-level traits are calculated by the analyser from summaries of the cell-type specific cell-level measurements.

The analysers were calibrated using an artificial blood containing stabilised human and animal cells by a Sysmex certified engineer, upon installation and subsequently every few months. The same material, prepared in three concentrations, was used by UK Biocentre staff to make daily quality control measurements. Sysmex reference analysers in Kobe, Hamburg and Chicago make routine measurements of the artificial blood to provide a reference for engineer calibration and for daily comparison with the local quality control measurements, so that any malfunction of an analyser between scheduled engineer visits can be identified.

### Selection of phenotype data

For each participant, where possible we used the ncCBC measurements derived from the blood sample donated at baseline. For a subset of traits, the data from the baseline sample were unavailable for some

participants and consequently we substituted the data derived from the analyses of the samples donated at the two year follow up visit. In particular, the PLT-F flow channel of one of the two analysers was incorrectly configured during the first 90 days of the study. Data were therefore missing for the traits measured by this channel during this period for approximately half the blood samples analysed.

We limited our analysis to samples that were analysed within 36 h of venipuncture. We excluded data points for which the analyser ‘flag’ variables indicated a sample or measurement abnormality (e.g. due to platelet clumping). We also excluded measurements if they were made on a day for which fewer than ten measurements of the corresponding trait were available from the corresponding analyser, or if they were made on a day when the median of the measurements of the trait made by the analyser deviated substantially (more than 8 median absolute deviations) from the global median of the measurements made the analyser over the course of the study.

**Removal of extraneous trait variation**

Haematology analyser data contain both technically and biologically mediated variation. Where this variation has a non-genetic origin it is a source of noise, which reduces the power of GWAS analysis. We adjusted the data from each trait in two stages using generalised additive models (GAMs) to remove phenotypic variation likely to be extraneous. We identified variables to include as predictors in the GAMs at each stage by consulting haematologists and performing exploratory visual and linear regression analyses. In the first stage, we adjusted each trait to remove technical variation and variation explained by the time at which the blood sample was analysed. The latter variation includes variation due to machine drift, variation caused by machine calibration and variation due to seasonal changes in the physiology of donors. In the second stage, we adjusted each trait for various environmental and physiological factors thought likely to generate variation in blood traits, but unlikely to be colliders of genotype and phenotype. At each stage, we adjusted the traits differently according to whether they were directly measured by the analyser or derived by the analyser from the directly measured traits (Supplementary Data 2). The directly measured traits were adjusted using regression models and the derived traits were re-calculated accordingly from the measured traits. Before each regression adjustment we log-transformed the measured traits that are positively supported (such as cell counts) and logit-transformed the measured traits that are proportions or percentages.

In the first stage, we adjusted each transformed trait for the time duration between the start of the study and the analysis of the blood sample, the time duration between venipuncture and the analysis of the blood sample, the day of the week, the number of days between the start of the year and the day of the measurement, and the instrument used for the analyse. The adjustments were made by fitting a GAM (Generalised Additive Model, R package *mgcv*) to each trait, using splines to model the dependence of the trait on the covariates jointly. The models were based on the regression Eq. (1), in which a regression coefficient corresponding to each term in square brackets has been suppressed for simplicity.

$$Ea(y_i) = s[t(i) \otimes m(i)] + c[t_{\text{year}}(i)] + tp[(t_{\text{day}}(i), t_{\text{ven}}(i)) \otimes (m(i), I(i))] + \sum_{D \in \{\text{tues, wed, \dots, sun}\}} [1_{D(i)=D}] \tag{1}$$

Here:

- $i$  is an index for the observation (the blood donor);
- $y_i$  represents the measured trait value and  $a(y_i)$  the value after its initial transformation;
- $t(i)$  denotes the number of seconds between the start of the study and the measurement of the trait value;

- $m(i)$  is a categorical variable with two levels, representing the haematology analyser used to record the measurement;
- $t_{\text{year}}(i)$  is the number of seconds between midnight on 1<sup>st</sup> of January in the year the measurement was made and the time at which the measurement was made;
- $t_{\text{day}}(i)$  is the time of day that the measurement was made, in seconds after midnight;
- $t_{\text{ven}}(i)$  is the time duration in seconds between venipuncture and the midnight immediately preceding the measurement;
- $I(i)$  is a binary variable which indicates whether  $t_{\text{ven}}(i)$  was imputed (rather than observed), imputation was by the median values of the observed  $t_{\text{ven}}(i)$  variable;
- $D(i)$  is a categorical variable with seven levels, representing the day of the week on which the measurement was made, the baseline category is monday;
- $s[\ ]$  represent a P-spline smoothing spline;
- $c[\ ]$  represents a cyclic smoothing spline;
- $tp[\ ]$  represents a thin plate spline smoothing spline;
- $\otimes$  indicates a binary operator, which expands additively to generate main effects and interaction terms.

For each trait, we took the residuals and re-centred them to have mean  $\overline{a(y)}$ , the mean of the unadjusted  $a$ -transformed traits. We called the variables obtained after applying the transformation  $a^{-1}$  to these re-centred residuals, the *technically adjusted traits*.

In the second stage, we further adjusted each trait for covariates measured at baseline and identified by haematologists as likely to generate variation in blood traits, this time using a GAM based on the regression Eq. (2), in which a regression coefficient corresponding to each term in square brackets has been suppressed for simplicity.

$$Eb(y_i) = s[age(i) \otimes meno(i)] + tp[(\log(wgt(i)), \log(hgt(i))) \otimes meno(i)] + [1_{wgt.na(i)}] + [1_{hgt.na(i)}] + \sum_{d \in K} [1_{drk(i)=d}] + \sum_{a \in A} [1_{alc(i)=a}] + s[pckysr(i)] + \sum_{m \in M} [1_{smk(i)=m}] + \sum_{f \in F} [1_{sfrq(i)=f}] + [1_{pckysr.na(i)}] + \sum_{r \in R} [1_{arm(i)=r}] \tag{2}$$

Here:

- $b(y_i)$  is the residual from the first stage regression model corresponding to observation  $i$ ;
- $age(i)$  is the age of the participant;
- $meno(i)$  is a categorical variable with five levels indicating the menopausal status of the participant, taking values in {post, pre, hysterectomy, male, NA};
- $wgt(i)$  is the weight of the participant (missing data were imputed by the population mean);
- $hgt(i)$  is the height of the participant (missing data were imputed by the population mean);
- $drk(i)$  is a categorical variable with four levels indicating the drinking status of the participant,  $K = \{\text{previous, current, NA}\}$ , the baseline category is ‘never’;
- $alc(i)$  is a categorical variable with six levels indicating the alcohol consumption rate of the participant,  $A = \{\text{rarely, 1 to 3 months, 1 to 2 weeks, 3 to 5 weeks, most days}\}$ , the baseline category is ‘never’;
- $pckysr(i)$  is the number of pack-years the participant has smoked (missing data were imputed by the population mean);
- $smk(i)$  is a categorical variable with four levels indicating the smoking status of the participant,  $M = \{\text{previous, current, NA}\}$ , the baseline category is ‘never’;
- $sfrq(i)$  is a categorical variable with six levels indicating the smoking frequency of the participant,  $F = \{\text{special occasions, rarely, occasional, most days, every day}\}$ , the baseline category is ‘never’;
- $arm(i)$  is a categorical variable with six levels indicating the INTERVAL trial arm to which the participant was assigned (see

'INTERVAL Study' above),  $R = \{M10, M12, F12, F14, F16\}$ , the baseline category is 'M8';

- $wgt.na(i)$ ,  $hgt.na(i)$ ,  $pckys.na(i)$ , evaluate to true if the participant has missing data respectively for weight, height, or pack years smoked and otherwise evaluate to false.

### Outlier removal and trait transformation

For each trait, we removed data points for which the adjustment was very large and removed data points for which the adjusted values were outliers in the adjusted distributions. Specifically, we took the residuals from the second stage GAM adjustment and removed data points for which the difference between the raw measured value (on the adjustment scale) and the adjusted value was more than 3.5 median absolute deviations from the median of the distribution of such differences. We then removed data points lying more than 4.5 median absolute deviations from the median of the distribution of the residuals. We also performed multivariate outlier removal within groups of phenotypes corresponding to each cell-type, by computing the sum of squares of the leading  $d$  principal components scores where  $d$  is the number of independent measurements required to compute the variables in each group and removing any data points for which the sum of squares fell into the upper  $10^{-7}$  tail probability of a  $\chi_d^2$  distribution. Finally, we quantile-inverse-normal transformed the trait data, stratifying by haematology analyser, sex and menopausal status. The final numbers of participants that contributed to the GWAS of each trait are given in Supplementary Data 2.

### Presentation of exploratory visual and linear regression analyses

Although the covariation between the ncCBC traits and biological variables is likely to be of interest to researchers and clinicians, it is difficult to interpret the estimates of coefficients of the GAMs used to adjust the phenotypes. Consequently, we reperformed some of the visual (Fig. 2, Supplementary Figs. 1–3) and linear regression analyses (Supplementary Data 3) used to explore the relationships between ncCBC traits and the variables age, sex, menopausal status and BMI, using the technically adjusted trait data but restricting to the set of participants contributing to the GWAS. We performed a multiple linear regression of each trait on sex, menopausal status and age to estimate:

- the mean of each trait in males, pre-menopausal females and post-menopausal females adjusted for age (estimates correspond to the mean age in the study sample, which is 43.7 years) and
- the effect of age on the mean of the trait while adjusting for sex and menopause.

We also performed a multiple linear regression of each trait on sex, menopause status, age and BMI to estimate the effect of BMI on each trait while adjusting for sex and menopause status.

### Genotyping, quality control of genetic data and imputation of unmeasured genotypes

DNA extraction was performed at LGC Genomics (UK) from buffy coat using a Kleargene method. Samples were packaged and shipped to Affymetrix (now Thermo Fisher Scientific; Mountain View, CA, USA) in 96-well barcoded plates, two wells of which were left empty for standard Affymetrix controls. A sample selection algorithm was used to ensure that the samples on each plate came from participants with a representative distribution of recruitment centre, recruitment date, regional hub and sex. A PicoGreen-based method was used to identify plates with a high proportion of low DNA concentration samples, which were replaced prior to genotyping. Genotyping was performed on the Affymetrix GeneTitan Multi-Channel Instrument according to the Affymetrix Axiom 2.0 Assay Automated Workflow. Genotypes were called in batches of approximately 50 plates (4800 samples) using the

Axiom GT1 algorithm implemented in the Affymetrix Power Tools software package.

A full description of the quality control procedures applied to the INTERVAL genotype data can be read in the Methods section of Astle et al.<sup>2</sup>. Data from samples showing evidence of DNA contamination were excluded. Data from samples that were genetic duplicates, from samples corresponding to related groups of participants and from samples from non-European ancestry participants were identified and removed. Data from samples with discordant self-reported and genetically determined sex were removed. Genetic duplicates and relatives were identified using a methods-of-moments estimator for coefficients of pairwise relatedness. Participants were considered to be of European ancestry if their scores on the first two principal components of genetic variation fell below component-specific thresholds in a dataset combining the genotyped INTERVAL participants and genotyped individuals in the major population groups of the 1000 Genomes project. These thresholds were determined by considering the bivariate distribution of principal components scores of individuals self-declaring their ethnicity as 'White'. Non-autosomal variants, variants with a low call rate, multi-allelic variants and variants showing significant (unadjusted  $P$ -value  $< 5 \times 10^{-6}$ ) deviation from the Hardy-Weinberg equilibrium were removed. The data were phased using SHAPEIT3, following which the genotypes of 87,696,910 variants were imputed from the combined 1000 Genomes Phase 3-UK10K panel using the PBWT imputation algorithm on the Sanger Imputation Server (<https://imputation.sanger.ac.uk>)<sup>62</sup>.

### Univariable genome-wide association analysis

We used BOLT-LMM version 2.3.1 to perform univariable (i.e. a single genetic variable) linear mixed model regression of each ncCBC phenotype on the imputed alternative allele dose at each of the 29.5 million genetic variants which passed quality control filters (MAF  $> 0.04\%$ , INFO score  $> 0.4$ ). We included as covariates, dummy variables indicating the donor clinic at which the blood sample was taken and the score vectors corresponding to the leading ten principal components of genetic variation in the study sample. The Genomic Control inflation factor  $\lambda$ -values ranged between 1.001 and 1.058 suggesting that the mixed model analysis and principal components covariates adequately adjusted for any confounding by relatedness or large scale population stratification<sup>63</sup>.

### Conditional analysis

For each trait, we performed stepwise multiple linear regression analyses to identify a parsimonious subset of genetic variants explaining the univariable genome-wide significant (Wald test unadjusted  $P$ -value  $< 8.31 \times 10^{-9}$ ) associations<sup>64</sup>. We partitioned the set of genome-wide significant variants for each trait into the maximum number of subsets ('blocks') such that no pair of variants in distinct subsets were separated by fewer than 10 Mb. We implemented a version of Efron's stepwise regression algorithm using the *fastLM* function in the R package *RcppEigen* and applied it to the variants in each block. We initialised the algorithm by specifying a linear regression model  $M$ , 'the current model' that included the covariates used in the univariable GWAS analysis (PC score vectors and clinic), but no genetic variants. We then performed iterations as follows:

1. For each variant not currently included in  $M$ ,
  - augment  $M$  with the variant to create  $M^*$  and compare  $M^*$  to  $M$  by a likelihood ratio test; compute the corresponding  $P$ -value.
2. Identify the variant corresponding to the lowest  $P$ -value in the set of model comparisons performed in step 1. If that  $P$ -value is lower than  $8.31 \times 10^{-9}$  and does not have a LD  $r^2$  score higher than 0.9 with any variants already in the model then update  $M$  by adding the variant to the model. If no such variant exists, terminate the algorithm.
3. For each variant currently included in  $M$ ,

- reduce  $M$  by removing the variant to create  $M^*$  and compare  $M$  to  $M^*$  by a likelihood ratio test; compute the corresponding  $P$ -value.
4. Identify the variant corresponding to the greatest  $P$ -value in the set of model comparisons performed in step 3. If that  $P$ -value is greater  $8.31 \times 10^{-9}$  then update  $M$  by removing the variant from the model and return to 3.
  5. Return to 1

Finally, for each trait, we ran the algorithm again, but beginning at step 3 and initialising  $M$  as the union of the sets of variants selected by the block-wise analyses for the trait. The terminating model of this run contains the set of ‘conditionally significant variants’ for the trait.

### Fine mapping of association signals

We performed statistical fine-mapping by applying FINEMAP 1.3.1<sup>28</sup> to each of a set of genomic windows covering the association signals. We generated the windows for each trait by centring a 500 kb precursor window on each conditionally significant variant and merging those that overlapped. The resulting windows ranged in size from 500,000 b to 599,652 b. For each application of FINEMAP, we specified the maximum number of causal variants ( $-n$ -causal-snps option) to be the number of conditionally significant variants in the window. We set the prior standard deviation on the effect size ( $-prior$ -std option) to 0.08. We computed the linkage disequilibrium (LD) correlation structure of the variants in each window from the imputed genetic dataset used for the univariable GWAS analysis. We computed a 95% Bayesian credible set of variants for each trait-window as the minimal set of variants jointly covering at least 95% of the posterior probability of association.

### Linkage disequilibrium clumping

The stepwise model selection procedure of the conditional analysis was performed independently for each ncCBC trait. Consequently, a genetic variant causally associated with multiple ncCBC traits may generate association signals that are tagged by different conditionally significant variants for different traits. To identify variants likely to represent common association signals, we sought to partition the full set of conditionally significant variants into sets of variants connected by chains of high LD. We applied the greedy clumping algorithm in PLINK 1.9.0b31, which assigns each variant to a distinct clump and then iteratively merges pairs of clumps if there exists a variant in each clump such that the pair of variants is in high LD. We used a threshold of 0.8 on the  $r^2$  measure of LD to merge clumps.

### Identification of association signals not reported by Vuckovic et al. or Chen et al.

In order to identify association signals not previously reported in large cCBC GWAS we declared a clump to represent a ‘novel’ blood cell trait signal if it did not contain a variant in strong LD ( $r^2 > 0.8$ ) with a conditionally significant variant reported by Vuckovic et al. or Chen et al.<sup>3,4</sup>. Following this, we performed an analysis at the cell-type (platelet, red cell, neutrophil, eosinophil, basophil, monocyte, lymphocyte) level. Each clump containing a variant associated with a trait of a given cell-type was declared to represent a novel association with that cell-type if it did not contain any variants in strong LD ( $r^2 > 0.8$ ) with a variant identified by the conditional analyses of Vuckovic et al. or Chen et al. to be associated with cCBC traits of the same cell-type (Table 1).

### Gene assignment

We used Variant Effect Predictor (VEP) version 84 to annotate the conditionally significant variants with the gene symbols of enveloping or proximal (within 5 kb) transcripts (Supplementary Data 4)<sup>65</sup>. All gene transcripts were considered. If a variant was enveloped by or proximal to the transcript of more than one gene, we reported only the genes for which the variant was predicted to have the maximally severe

**Table 1 | The cell-type-specific cCBC traits studied by Vuckovic et al. and Chen et al. classified by cell-type**

cCBC trait(s)	Cell-type
MPV, PCT, PDW, PLT#	Platelets
HCT, HGB, HLSR#, HLSR%, IRF, MCH, MCHC, MCV, RBC#	Red cells
NEUT#	Neutrophils
EO#	Eosinophils
BASO#	Basophils
MONO#	Monocytes
LYMPH#	Lymphocytes

transcriptional consequence according to the Ensembl ranking of transcriptional consequences. We applied the *alias2symbol* function in the R package ‘*limma*’ to map alias gene names to HGNC gene symbols. We excluded genes without an HGNC symbol.

### Genetic correlation

We estimated the pairwise genetic correlation between each pair of ncCBC traits and between each pair of ncCBC and cCBC traits in the European ancestry population by applying LD Score Regression<sup>66</sup>. We used the LD scores supplied with the LDSC software, which were pre-computed from the 1000 Genomes European data at the 1.2 million common SNPs identified in the CEU population by the HapMap 3 project. We used our own GWAS summary statistics for the ncCBC traits and those of Vuckovic et al. for the cCBC traits<sup>3</sup>.

### Colocalisation with eQTL

We applied *gwas-pw* to identify ncCBC genetic association signals that colocalised with blood cell eQTL<sup>39</sup>. We performed eQTL mapping, in windows extending 1MB from each gene boundary, in platelets ( $n = 424$ ), CD15<sup>+</sup> neutrophils ( $n = 300$ ) and CD14<sup>+</sup> monocytes ( $n = 1490$ ), CD4<sup>+</sup> T-cells, CD8<sup>+</sup> T-cells and CD19<sup>+</sup> B-cells using genetic and transcriptomic data obtained from CEDAR, Cardiogenics, the Wellcome Trust Centre for Human Genetics (WTCHG), and BLUE-PRINT. Unmeasured genotypes were imputed using the Haplotype Reference Consortium release 1.1 reference panel. These data contributed to an eQTL meta-analysis which has been published in Vösa et al.<sup>67</sup>. We performed colocalisation analysis between genetic associations from 35 cell-type matched ncCBC cytometry traits and the eQTL. Association signals from H-IPF, P-LCR, PLT-SSC, PLT-SFL, PLT-FSC, PLT-SSC-DW, PLT-SFL-DW, and PLT-FSC-DW were colocalised with eQTL of platelets. Association signals from NE-SSC, NE-SFL, NE-FSC, NE-SSC-DW, NE-SFL-DW, NE-FSC-DW, EO-SSC, EO-SFL, EO-FSC, EO-SSC-DW, EO-SFL-DW and EO-FSC-DW were colocalised with eQTL of CD15<sup>+</sup> cells. Association signals from MO-SSC, MO-SFL, MO-FSC, MO-SSC-DW, MO-SFL-DW, and MO-FSC-DW were colocalised with eQTL from CD14<sup>+</sup> cells. Finally, association signals from RE-LYMP#, RE-LYMP(L)%, RE-LYMP%, and LY-SSC, LY-SFL, LY-FSC, LY-SSC-DW, LY-SFL-DW, and LY-FSC-DW were colocalised with eQTL of CD4<sup>+</sup>, CD8<sup>+</sup>, and CD19<sup>+</sup> cells.

A colocalisation analysis was performed between the ncCBC association signal of each conditionally significant variant and each eQTL of the corresponding cell-type, providing that the variants were in strong LD ( $r^2 > 0.8$ ). Each analysis was performed using data from the recombination region of Berisa et al.<sup>68</sup> containing the conditionally significant variant. If the region contained multiple conditionally significant variants of the ncCBC trait, we regressed the ncCBC trait on the conditionally significant variants that were not in strong LD ( $r^2 > 0.8$ ) with the variant of interest and used the resulting residuals as a phenotype for univariable association tests, to obtain ncCBC association statistics for each variant in the region with the effect of secondary signals removed. We filtered the results to report only those which are

'highly likely' to represent true colocalisation: those with a posterior probability greater than 80%<sup>30</sup> (Supplementary Data 5).

### ATAC-Seq enrichment analysis

Following the procedure described by Ulirsch et al., we applied g-chromVAR to compute the enrichment of genetic variants associated with ncCBC traits in the nucleosome depleted regions of the genome of eighteen blood cell populations representing various stages of differentiation<sup>29</sup>. Briefly, regions depleted in at least one cell-type were approximated using 500 b wide windows, each centred at the summit of one of a set of ATAC-seq read peaks selected by applying a consensus finding method to the union of peaks called across the eighteen cell-types. For each cell-type, the ATAC-seq fragment counts in each region were weighted by the sum of the FINEMAP posterior probability of variants in the region, restricted to those variants for which the posterior probability of association was greater than 0.1%. The genome-wide sum of these weighted counts was compared to its expectation under a null model which assumed no enrichment and controlled for GC content and average peak intensity.

### Colocalisation with pQTL

We applied *gwas-pw* to identify ncCBC genetic association signals that colocalised with the genetic associations of 1478 blood plasma proteins identified by Sun et al.<sup>30</sup>. We followed the procedure described for the eQTL colocalisation analysis, except that the regional summary statistics for genetic association with the plasma protein concentrations were adjusted to remove local secondary signals, following the procedure previously described to remove secondary associations from the ncCBC summary statistics (Supplementary Data 7).

### Identification of platelet $\alpha$ -granule proteins

Firstly, we conducted a literature review to identify proteins localised to platelet  $\alpha$ -granules by mass spectrometry<sup>69–71</sup>, electron microscopy, immunogold labelling, or platelet sub-fractionation<sup>72–80</sup>; cytoskeletal proteins were excluded. Proteins identified in the platelet releasate were included, absent other evidence that they were  $\alpha$ -granule proteins<sup>81–83</sup>. Secondly, we performed a differential abundance analysis comparing the platelets of patients with grey platelet syndrome (GPS;  $n = 5$ ), which lack  $\alpha$ -granules, to those of healthy controls ( $n = 5$ ). Platelets were isolated, analysed by mass spectrometry, and normalised abundance values (NAVs) were calculated for each protein detected<sup>84</sup>. We performed two-sample Student's *t*-tests (without assuming equal variances) to test for a difference in the mean of the log NAVs of each protein between the two conditions. Proteins for which the test was significant (unadjusted *P*-value  $\leq 0.05$ ) and for which the log-ratios were greater than 2 standard deviations from the median were considered to be differentially abundant. Differentially abundant proteins that were less abundant in GPS platelets than controls were assumed to be localised to the platelet  $\alpha$ -granule.

### The effect of rs6993770 on the average plasma concentrations of proteins expressed in platelets

We extracted from the results of Sun et al. the statistics summarising the univariate associations of rs6993770 with 1456 proteins showing strong evidence for expression in MKs (RNA-seq  $\log_2$ FPKM  $> 1$ )<sup>30</sup>. The estimated effect size corresponding to 215 of these proteins was sufficiently large to correspond to an unadjusted *P*-value  $< 1 \times 10^{-3}$ . 44 of these were localised to  $\alpha$ -granules, of which 40 had negative estimated T-allelic effect sizes, while 171 were not localised to  $\alpha$ -granules, of which 101 had negative T-allelic effect sizes. We performed a Fisher's exact test for the corresponding 2 $\times$ 2 table.

Using a dataset in which each row corresponded to one of 1456 plasma proteins studied by Sun et al. and expressed in MKs ( $\log_2$ FPKM  $> 1$ ), we performed a linear regression of the estimated effect size of rs6993770-T on a dummy variable indicating  $\alpha$ -granule

localisation while adjusting for the abundance of the mRNA transcripts for the corresponding genes in MKs.

### Colocalisation with disease risk variants

A catalogue of summary statistics from GWAS of 15 complex diseases was collated following a literature search. The diseases considered were: allergic disease<sup>85</sup>, Alzheimer's disease<sup>86</sup>, asthma<sup>87</sup>, coeliac disease<sup>88,89</sup>, coronary artery disease<sup>90</sup>, Crohn's disease<sup>91,92</sup>, eczema<sup>93</sup>, hayfever or rhinitis<sup>85</sup>, inflammatory bowel disease<sup>91,92</sup>, multiple sclerosis<sup>94–96</sup>, primary biliary cirrhosis<sup>97,98</sup>, primary sclerosing cholangitis<sup>99</sup>, systemic lupus erythematosus<sup>100</sup>, type 1 diabetes<sup>101</sup> and ulcerative colitis<sup>91</sup>. We performed colocalisation to identify variants associated with ncCBC traits and disease risk, following the procedure described for eQTL, but without regressing out secondary signals (Supplementary Data 8).

### ZFPM2 knock-out

ZFPM2 knockout cell lines were obtained from human iPSCs (HPSI113i-qolg\_3 HipSci cell line) by CRISPR/Cas9 genome editing. The sgRNA (5' GA GTC GAC AGC AAC TTC CAG 3') was cloned into the lentiGuide-Puro (Addgene plasmid # 52963<sup>102</sup>). The sgRNA and Cas9 plasmids were co-transfected into a human iPSC line using the Lonza Nucleofector™ and the Human Stem Cell Nucleofector™ Kit 1 (Lonza, VPH-5012). Nucleofected cells were single cell sorted and knock-out variants were confirmed by Sanger sequencing (Supplementary Fig. 12).

### In vitro differentiation of MKs

MKs were differentiated from human iPSCs (HPSI113i-qolg\_3 HipSci cell line, wild-type and ZFPM2 knockout) using the protocol described in Moreau et al.<sup>32</sup>. Briefly, cells were trypsinized and seeded with 100,000 cells per well. The next day, cells were infected with lentiviruses encoding the GATA1, FLI1 and TAL1 transcription factors (Vectalis). The first two days after infection, cells were cultured with BMP4 (BioTechne) and FGF2 (Wellcome–MRC Cambridge Stem Cell Institute, Tissue Culture facility) to induce mesoderm. Eighteen days were allowed for the differentiation and maturation of MKs, during which flow-cytometry experiments were used to monitor the expression of CD41a (BD, cat 559777) and CD42b (BD, 555473) as measures of maturation. Three technical replicates were nested within each of three biological replicates: two knockouts and one wild-type.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

For ethical and legal reasons access to INTERVAL data are subject to controls. Bona fide scientists can seek access to relevant de-identified individual participant data—including genetic, haematology analyser and proteomic data—and a copy of the trial's data dictionary by applying to the INTERVAL Data Access Committee using the email address [helpdesk@intervalstudy.org.uk](mailto:helpdesk@intervalstudy.org.uk). The INTERVAL Data Access Committee (supplemented, when required, by expertise from additional external scientists) meets several times a year to review applications according to the usual academic criteria of scientific validity and feasibility. Following approval by the INTERVAL Data Access Committee, a material transfer or research collaboration agreement will be agreed and signed with the applicants. Applicants might be requested to provide reimbursement of data management or preparation costs, as the INTERVAL trial is no longer in receipt of funding. Applicants will be required to provide updates to the INTERVAL Data Access Committee on their use of the INTERVAL trial data, including provision of copies of any publications. Applicants will be required to adhere in publications with the INTERVAL trial's policy for acknowledgment of the trial's funders, stakeholders, and scientific or technical

contributors. The GRCh37 genome reference build is available for download from <https://grch37.ensembl.org/info/data/ftp/index.html>. Genomewide summary statistics may be downloaded by anonymous ftp from [ftp://ftp.sanger.ac.uk/pub/project/humgen/summary\\_statistics/sysmex\\_blood\\_cell\\_genetics](ftp://ftp.sanger.ac.uk/pub/project/humgen/summary_statistics/sysmex_blood_cell_genetics). The data from Ulirsch et al.<sup>29</sup> are available from [https://github.com/caleblareau/singlecell\\_bloodtraits/](https://github.com/caleblareau/singlecell_bloodtraits/), from the Gene Expression Omnibus (GEO) under accession GSE119453 and from the Sequence Read Archive (SRA) under accession PRJNA491478. Other MK epigenetic data were generated by the BLUEPRINT project and are available in the EGA dataset EGAD00001001871.

## Code availability

The R code used for the association analysis is available in the git repository: <https://github.com/ParsaAkbari/UKBB500K-Conditional-Analysis>.

## References

- Johnson, A. D. et al. Genome-wide meta-analyses identifies seven loci associated with platelet aggregation in response to agonists. *Nat. Genet.* **42**, 608–613 (2010).
- Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429.e19 (2016).
- Vuckovic, D. et al. The polygenic and monogenic basis of blood traits and diseases. *Cell* **182**, 1214–1231.e11 (2020).
- Chen, M.-H. et al. Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. *Cell* **182**, 1198–1213.e14 (2020).
- Amulic, B., Cazalet, C., Hayes, G. L., Metzler, K. D. & Zychlinsky, A. Neutrophil function: from mechanisms to disease. *Annu. Rev. Immunol.* **30**, 459–489 (2012).
- Kono, M. et al. Detection of activated neutrophils by reactive oxygen species production using a hematology analyzer. *J. Immunol. Methods* **463**, 122–126 (2018).
- Linssen, J. et al. Automation and validation of a rapid method to assess neutrophil and monocyte activation by routine fluorescence flow cytometry in vitro. *Cytom. B: Clin. Cytom.* **74**, 295–309 (2008).
- Zimmermann, M., Steenhuis, P., Linssen, J. & Weimann, A. Detection and quantification of hypo- and hypergranulated neutrophils on the new Sysmex XN hematology analyzer: establishment of an adapted reference interval for the neutrophil-granularity-intensity compared to XE-technology in adult patients. *Clin. Lab.* **61**, 235–241 (2015).
- Zimmermann, M., Cremer, M., Hoffmann, C., Weimann, K. & Weimann, A. Granularity Index of the SYSMEX XE-5000 hematology analyzer as a replacement for manual microscopy of toxic granulation neutrophils in patients with inflammatory diseases. *Clin. Chem. Lab. Med.* **49**, 1193–1198 (2011).
- Dennison, D. et al. Circulating activated neutrophils in COVID-19: An independent predictor for mechanical ventilation and death. *Int. J. Infect. Dis.* **106**, 155–159 (2021).
- Furundarena, J. R. et al. The utility of the Sysmex XE-2100 analyzer's NEUT-X and NEUT-Y parameters for detecting neutrophil dysplasia in myelodysplastic syndromes. *Int. J. Lab. Hematol.* **32**, 360–366 (2010).
- Miyachi, H. & Ando, Y. Diagnostic value of measurement of RNA in platelets by fluorescence flow cytometry. *Rinsho Byori* **40**, 297–302 (1992).
- Buoro, S. et al. Clinical significance of cell population data (CPD) on Sysmex XN-9000 in septic patients with our without liver impairment. *Ann. Transl. Med.* **4**, 418 (2016).
- Park, S. H. et al. Sepsis affects most routine and cell population data (CPD) obtained using the Sysmex XN-2000 blood cell analyzer: neutrophil-related CPD NE-SFL and NE-WY provide useful information for detecting sepsis. *Int. J. Lab. Hematol.* **37**, 190–198 (2015).
- Arneth, B. M. et al. Novel parameters of extended complete blood cell count under fluorescence flow cytometry in patients with sepsis. *J. Clin. Lab. Anal.* **28**, 130–135 (2014).
- Le Roux, G. et al. Routine diagnostic procedures of myelodysplastic syndromes: value of a structural blood cell parameter (NEUT-X) determined by the Sysmex XE-2100™. *Int. J. Lab. Hematol.* **32**, e237–43 (2010).
- Furundarena, J. R. et al. The utility of the Sysmex XE-2100 analyzer's NEUT-X and NEUT-Y parameters for detecting neutrophil dysplasia in myelodysplastic syndromes. *Int. J. Lab. Hematol.* **32**, 360–366 (2010).
- Brisou, G. et al. Alarms and parameters generated by hematology analyzer: new tools to predict and quantify circulating Sezary cells. *J. Clin. Lab. Anal.* **29**, 153–161 (2015).
- Harte, J. V. & Mykytiv, V. A panhaemocytometric approach to COVID-19: a retrospective study on the importance of monocyte and neutrophil population data on Sysmex XN-series analysers. *Clin. Chem. Lab. Med.: CCLM/FESCC* **59**, e169–e172 (2021).
- Henriot, I. et al. New parameters on the hematology analyzer XN-10 (Sysmex™) allow to distinguish childhood bacterial and viral infections. *Int. J. Lab. Hematol.* **39**, 14–20 (2017).
- Moore, C. et al. The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials* **15**, 363 (2014).
- Borregaard, N., Sørensen, O. E. & Theilgaard-Mönch, K. Neutrophil granules: a library of innate immunity proteins. *Trends Immunol.* **28**, 340–345 (2007).
- Rørvig, S., Østergaard, O., Heegaard, N. H. H. & Borregaard, N. Proteome profiling of human neutrophil granule subsets, secretory vesicles, and cell membrane: correlation with transcriptome profiling of neutrophil precursors. *J. Leukoc. Biol.* **94**, 711–721 (2013).
- Acharya, K. R. & Ackerman, S. J. Eosinophil granule proteins: form and function. *J. Biol. Chem.* **289**, 17406–17415 (2014).
- Becknell, B. et al. Ribonucleases 6 and 7 have antimicrobial function in the human and murine urinary tract. *Kidney Int.* **87**, 151–161 (2015).
- MacGlashan, D. W. Jr. Basophil activation testing. *J. Allergy Clin. Immunol.* **132**, 777–787 (2013).
- Grassi, L. et al. Dynamics of transcription regulation in human bone marrow myeloid differentiation to mature blood neutrophils. *Cell Rep.* **24**, 2784–2794 (2018).
- Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
- Ulirsch, J. C. et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* **51**, 683–693 (2019).
- Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
- Grassi, L. et al. Cell type-specific novel long non-coding RNA and circular RNA in the BLUEPRINT hematopoietic transcriptomes atlas. *Haematologica* **106**, 2613–2623 (2021).
- Moreau, T. et al. Large-scale production of megakaryocytes from human pluripotent stem cells by chemically defined forward programming. *Nat. Commun.* **7**, 11208 (2016).
- Han, G. C. et al. Genome-wide organization of GATA1 and TAL1 determined at high resolution. *Mol. Cell. Biol.* **36**, 157–172 (2016).
- Kassouf, M. T. et al. Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res.* **20**, 1064–1083 (2010).

35. Gieger, C. et al. New gene functions in megakaryopoiesis and platelet formation. *Nature* **480**, 201–208 (2011).
36. Klarin, D. et al. Genetic analysis of venous thromboembolism in UK Biobank identifies the *ZFPM2* locus and implicates obesity as a causal risk factor. *Circ. Cardiovasc. Genet.* **10**, e001643 (2017).
37. Klarin, D. et al. Genome-wide association analysis of venous thromboembolism identifies new risk loci and genetic overlap with arterial vascular disease. *Nat. Genet.* **51**, 1574–1579 (2019).
38. Kundu, K. et al. Genetic associations at regulatory phenotypes improve fine-mapping of causal variants for 12 immune-mediated diseases. *Nat. Genet.* **54**, 251–262 (2022).
39. Pickrell, J. K. et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).
40. Legroux, L. & Arbour, N. Multiple sclerosis and T lymphocytes: an entangled story. *J. Neuroimmune Pharmacol.* **10**, 528–546 (2015).
41. Wynn, D. et al. Daclizumab in active relapsing multiple sclerosis (CHOICE study): a phase 2, randomised, double-blind, placebo-controlled, add-on trial with interferon beta. *Lancet Neurol.* **9**, 381–390 (2010).
42. Bielekova, B. et al. Humanized anti-CD25 (daclizumab) inhibits disease activity in multiple sclerosis patients failing to respond to interferon  $\beta$ . *Proc. Natl Acad. Sci. USA* **101**, 8705–8708 (2004).
43. Giovannoni, G. et al. Safety and tolerability profile of daclizumab in patients with relapsing-remitting multiple sclerosis: An integrated analysis of clinical studies. *Mult. Scler. Relat. Disord.* **9**, 36–46 (2016).
44. Curto, E., Munteis-Olivas, E., Balcells, E. & Domínguez-Álvarez, M. M. Pulmonary eosinophilia associated to treatment with natalizumab. *Ann. Thorac. Med.* **11**, 224–226 (2016).
45. Abbas, M. et al. Hypereosinophilia in patients with multiple sclerosis treated with natalizumab. *Neurology* **77**, 1561–1564 (2011).
46. Stork, L. et al. Severe meningo-/encephalitis after daclizumab therapy for multiple sclerosis. *Mult. Scler.* **25**, 1618–1632 (2019).
47. Lee, L.-F. et al. IL-7 promotes T(H)1 development and serum IL-7 predicts clinical response to interferon- $\beta$  in multiple sclerosis. *Sci. Transl. Med.* **3**, 93ra68 (2011).
48. Bielekova, B. et al. Regulatory CD56(bright) natural killer cells mediate immunomodulatory effects of IL-2R $\alpha$ -targeted therapy (daclizumab) in multiple sclerosis. *Proc. Natl Acad. Sci. USA* **103**, 5941–5946 (2006).
49. Galarza-Muñoz, G. et al. Human epistatic interaction controls IL7R splicing and increases multiple sclerosis risk. *Cell* **169**, 72–84.e13 (2017).
50. Gregory, S. G. et al. Interleukin 7 receptor alpha chain (IL7R) shows allelic and functional association with multiple sclerosis. *Nat. Genet.* **39**, 1083–1091 (2007).
51. Majdinasab, N., Hosseini Behbahani, M., Galehdari, H. & Mohaghegh, M. Association of interleukin 7 receptor gene polymorphism rs6897932 with multiple sclerosis patients in Khuzestan. *Iran. J. Neurol.* **13**, 168–171 (2014).
52. Sandborn, W. J. et al. Vedolizumab as induction and maintenance therapy for Crohn's disease. *N. Engl. J. Med.* **369**, 711–721 (2013).
53. Feagan, B. G. et al. Vedolizumab as induction and maintenance therapy for ulcerative colitis. *N. Engl. J. Med.* **369**, 699–710 (2013).
54. Soler, D. et al. The binding specificity and selective antagonism of vedolizumab, an anti- $\alpha 4\beta 7$  integrin therapeutic antibody in development for inflammatory bowel diseases. *J. Pharmacol. Exp. Ther.* **330**, 864–875 (2009).
55. Rogler, G. Mechanism of action of vedolizumab: do we really understand it? *Gut* **68**, 4–5 (2018).
56. Zeissig, S. et al. Vedolizumab is associated with changes in innate rather than adaptive immunity in patients with inflammatory bowel disease. *Gut* **68**, 25–39 (2018).
57. Schippers, A. et al.  $\beta 7$ -Integrin exacerbates experimental DSS-induced colitis in mice by directing inflammatory monocytes into the colon. *Mucosal Immunol.* **9**, 527–538 (2016).
58. Staley, J. R. et al. PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* **32**, 3207–3209 (2016).
59. Kamat, M. A. et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics* **35**, 4851–4853 (2019).
60. Lyons, P. A. et al. Genetically distinct subsets within ANCA-associated vasculitis. *N. Engl. J. Med.* **367**, 214–223 (2012).
61. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
62. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–72 (2014).
63. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55** (1999).
64. Xu, C. et al. Estimating genome-wide significance for whole-genome sequencing studies. *Genet. Epidemiol.* **38**, 281–290 (2014).
65. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
66. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
67. Vösa, U. et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
68. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).
69. Maynard, D. M., Heijnen, H. F. G., Gahl, W. A. & Gunay-Aygun, M. The  $\alpha$ -granule proteome: novel proteins in normal and ghost granules in gray platelet syndrome. *J. Thromb. Haemost.* **8**, 1786–1796 (2010).
70. Maynard, D. M., Heijnen, H. F. G., Horne, M. K., White, J. G. & Gahl, W. A. Proteomic analysis of platelet  $\alpha$ -granules using mass spectrometry. *J. Thromb. Haemost.* **5**, 1945–1955 (2007).
71. Zufferey, A. et al. Characterization of the platelet granule proteome: evidence of the presence of MHC1 in  $\alpha$ -granules. *J. Proteom.* **101**, 130–140 (2014).
72. Huang, C.-L. et al. Disabled-2 is a novel  $\text{IIb}$ -integrin-binding protein that negatively regulates platelet-fibrinogen interactions and platelet aggregation. *J. Cell Sci.* **119**, 4420–4430 (2006).
73. Charafeddine, A. H. et al. Platelet-derived CD154: ultrastructural localization and clinical correlation in organ transplantation. *Am. J. Transpl.* **12**, 3143–3151 (2012).
74. Deuel, T. F. et al. Human platelet-derived growth factor. Purification and resolution into two active protein fractions. *J. Biol. Chem.* **256**, 8896–8899 (1981).
75. Kaplan, D. R., Chao, F. C., Stiles, C. D., Antoniades, H. N. & Scher, C. D. Platelet  $\alpha$  granules contain a growth factor for fibroblasts. *Blood* **53**, 1043–1052 (1979).
76. Chatterjee, M. et al. Distinct platelet packaging, release, and surface expression of proangiogenic and antiangiogenic factors on different platelet stimuli. *Blood* **117**, 3907–3911 (2011).
77. Chen, R., Jin, G. & McIntyre, T. M. The soluble protease ADAMDEC1 released from activated platelets hydrolyzes platelet membrane pro-epidermal growth factor (EGF) to active high-molecular-weight EGF. *J. Biol. Chem.* **292**, 10112–10122. (2017).
78. Cicha, I., Garlich, C. D., Daniel, W. G. & Goppelt-Struebe, M. Activated human platelets release connective tissue growth factor. *Thromb. Haemost.* **91**, 755–760 (2004).



79. Fang, L. et al. PDGF C is a selective alpha platelet-derived growth factor receptor agonist that is highly expressed in platelet alpha granules and vascular smooth muscle. *Arterioscler. Thromb. Vasc. Biol.* **24**, 787–792 (2004).
80. Tamura, S. et al. Release reaction of brain-derived neurotrophic factor (BDNF) through PAR1 activation and its two distinct pools in human platelets. *Thromb. Res.* **128**, e55–61 (2011).
81. Parsons, M. E. M. et al. Platelet releasate proteome profiling reveals a core set of proteins with low variance between healthy adults. *Proteomics* **18**, e1800219 (2018).
82. van Holten, T. C. et al. Quantitative proteomics analysis reveals similar release profiles following specific PAR-1 or PAR-4 stimulation of platelets. *Cardiovasc. Res.* **103**, 140–146 (2014).
83. Wijten, P. et al. High precision platelet releasate definition by quantitative reversed protein profiling—brief report. *Arterioscler. Thromb. Vasc. Biol.* **33**, 1635–1638 (2013).
84. Sims, M. C. et al. Novel manifestations of immune dysregulation and granule defects in gray platelet syndrome. *Blood* **136**, 1956–1967 (2020).
85. Ferreira, M. A. et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat. Genet.* **49**, 1752–1757 (2017).
86. Lambert, J. C. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
87. Demenais, F. et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat. Genet.* **50**, 42–53 (2018).
88. Dubois, P. C. A. et al. Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* **42**, 295–302 (2010).
89. Trynka, G. et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* **43**, 1193–1201 (2011).
90. Nikpay, M. et al. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).
91. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
92. de Lange, K. M. et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
93. Paternoster, L. et al. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nat. Genet.* **47**, 1449–1456 (2015).
94. International Multiple Sclerosis Genetics Consortium. et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).
95. International Multiple Sclerosis Genetics Consortium (IMSGC). et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* **45**, 1353–1360 (2013).
96. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365** (2019).
97. Liu, J. Z. et al. Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nat. Genet.* **44**, 1137–1141 (2012).
98. Cordell, H. J. et al. International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nat. Commun.* **6**, 8019 (2015).
99. Ji, S.-G. et al. Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. *Nat. Genet.* **49**, 269–273 (2017).
100. Bentham, J. et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* **47**, 1457–1464 (2015).
101. Onengut-Gumuscu, S. et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47**, 381–386 (2015).
102. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).
103. Di Angelantonio, E. et al. Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. *Lancet* **390**, 2360–2371 (2017).
104. Sysmex. Automated Hematology Analyzer XN series (XN-1000) Instructions for Use. *Sysmex Corporation, Kobe, Japan* (2014).
105. Pruim, R. J. et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
106. Zhang, Y., An, L., Yue, F. & Hardison, R. C. Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res.* **44**, 6721–6731 (2016).
107. Petersen, R. et al. Platelet function is modified by common sequence variation in megakaryocyte super enhancers. *Nat. Commun.* **8**, 16058 (2017).

## Acknowledgements

The authors are grateful to the participants in the INTERVAL study, whose generosity made this work possible. INTERVAL is a collaboration between the University of Cambridge, the University of Oxford and National Health Service Blood and Transplant (NHSBT; [www.nhsbt.nhs.uk](http://www.nhsbt.nhs.uk)). We thank Dr Carmel Moore and the study coordination teams at the University of Cambridge, the University of Oxford and NHSBT. In particular we recognise the contribution of Dr Jennifer Sambrook, whose work made this study possible, and who sadly died before its final publication. We thank Dr Matthew Walker of the Cardiovascular Epidemiology Unit, and the data management team of the NIHR Blood and Transplant Research Unit (BTRU) in Donor Health and Genomics. We thank Dr Jarob Saker and Dr Joachim Linszen of Sysmex Europe and Mr Stephen Garner, lately of the NHSBT Component Development Laboratory, for their invaluable advice on the Sysmex XN-1000 instrument. We thank Joanna Westmoreland of the Medical Research Council (MRC) Laboratory of Molecular Biology for preparing illustrations and artwork. We thank members of the NIHR Cambridge BioResource Scientific Advisory Board and Management Committee for their support of the study. DNA extraction and genotyping were co-funded by the National Institute for Health and Care Research (NIHR), the NIHR BioResource (<http://bioresource.nihr.ac.uk>) and the NIHR Cambridge Biomedical Research Centre (BRC) (BRC-1215-20014) [\*]. The academic coordinating centre for INTERVAL was supported by core funding from: the NIHR BTRU in Donor Health and Genomics (NIHR BTRU-2014-10024), the NIHR BTRU in Donor Health and Behaviour (NIHR203337), the NIHR Cambridge BRC (BRC-1215-20014; NIHR203312), the MRC (MR/L003120/1) and the British Heart Foundation (BHF) (SP/09/002; RG/13/13/30194; RG/18/13/33946) [\*]. A complete list of the investigators and contributors to the INTERVAL trial is given in Di Angelantonio et al.<sup>103</sup>. The academic coordinating centre would like to thank blood donor centre staff for their help and to thank blood donors for participating in the INTERVAL trial. This work was supported by Health Data Research UK (HDR UK), which is funded by the MRC, the Engineering and Physical Sciences Research Council, the Economic and Social Research Council, the Department of Health and Social Care (England), the Chief Scientist Office of the Scottish Government Health and Social Care Directorates,

the Health and Social Care Research and Development Division (Welsh Government), the Public Health Agency (Northern Ireland), the BHF and the Wellcome Trust. P.A. and D.V. were funded by the NIHR BTRU in Donor Health and Genomics (NIHR BTRU-2014-10024). T.J. is funded by a Cambridge University MRC Doctoral Training Programme (MR/S502443/1) topped up by Sackler and the NIHR Cambridge BRC [\*]. J.H.C. is supported by a MRC Clinical Research Training Fellowship (MR/P02002X/1). K.D. is supported by Health Education England as an HSST trainee. J.C.K. is funded by a Wellcome Trust Investigator Award (204969/Z/16/Z to J.C.K.), Wellcome Trust Grants (074318 to J.C.K., 090532/Z/09/Z and 203141/Z/16/Z to the Wellcome Centre for Human Genetics core facility), NIHR Oxford BRC. S.K. is funded by a BHF Chair award (CH/12/2/29428). J.D. holds a BHF Professorship and a NIHR Senior Investigator Award [\*]. V.G.S. is supported by NIH grants R01 DK103794, R01 CA265726, R01 HL146500; a gift from the Lodish Family to Boston Children's Hospital; the Mathers Foundation; and the New York Stem Cell Foundation (NYSCF). V.G.S. is a NYSCF-Robertson Investigator. M.F. is funded by the BHF (FS/18/53/33863) and is supported by the NIHR Exeter BRC. J.E.P. is supported by a UKRI Innovation Fellowship at HDR UK (MR/S004068/1). Research in the W.H.O. laboratory was funded by the International Society on Thrombosis and Haemostasis, the MRC, NHSBT and the Rosetrees Trust. W.J.A. is funded by NHSBT and the NIHR Cambridge BRC [\*]. \*The views expressed are those of the authors and not necessarily those of the NIHR, NHSBT or the Department of Health and Social Care.

## Author contributions

Conceptualisation: W.H.O.; Data curation: P.A., D.V., T.J., K.K., R.K., J.H.C., L.G., J.A.G., S.K., S.M., J.S., K.W. and W.J.A.; Formal analysis: P.A., D.V., T.J., E.L.B., L.G., D.S., J.M.V. and W.J.A.; Funding acquisition: N.A.W., J.D., D.J.R., E.D.A., A.S.B., W.H.O. and N.S.; Investigation: P.A., D.V., L.S., T.J., J.S. and W.J.A.; Methodology: P.A., D.V., T.J., S.B. and W.J.A.; Project administration: N.A.W., A.S.B., W.H.O., N.S. and W.J.A.; Resources: J.H.C., K.D., J.C.K., V.G.S., M.F., J.E.P., A.S.B. and W.H.O., N.S.; Software: P.A., D.V., T.J., K.K., R.K., S.M., J.E.P. and W.J.A.; Supervision: O.S., M.F., S.B., T.K., J.E.P., A.S.B., W.H.O., N.S. and W.J.A.; Visualisation: P.A., L.S., N.S. and W.J.A.; Writing—original draft: P.A., J.E.P., A.S.B., W.H.O., N.S. and W.J.A.; Writing—review and editing: P.A., D.V., E.L.B., J.D., D.J.R., V.G.S., M.F., S.B., T.K., J.E.P., A.S.B., W.H.O., N.S. and W.J.A.

## Competing interests

P.A. is an employee of Regeneron Pharmaceuticals and receives salary from and owns options and/or stock of the company. The remaining authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-023-40679-y>.

**Correspondence** and requests for materials should be addressed to Adam S. Butterworth, Willem H. Ouwehand, Nicole Soranzo or William J. Astle.

**Peer review information** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

<sup>1</sup>British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, Strangeways Research Laboratory, University of Cambridge, Wort's Causeway, Cambridge CB1 8RN, UK. <sup>2</sup>Department of Human Genetics, The Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1HH, UK. <sup>3</sup>Medical Research Council Biostatistics Unit, University of Cambridge, East Forvie Building, Cambridge Biomedical Campus, Forvie Site, Robinson Way, Cambridge CB2 0SR, UK. <sup>4</sup>The National Institute for Health and Care Research Blood and Transplant Unit in Donor Health and Genomics, Strangeways Research Laboratory, Strangeways Research Laboratory, University of Cambridge, Wort's Causeway, Cambridge CB1 8RN, UK. <sup>5</sup>Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. <sup>6</sup>Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Long Road, Cambridge CB2 0PT, UK. <sup>7</sup>National Health Service Blood and Transplant, Cambridge Centre, Cambridge Biomedical Campus, Long Road, Cambridge CB2 0PT, UK. <sup>8</sup>British Heart Foundation Centre of Research Excellence, University of Cambridge, Addenbrooke's Hospital, Cambridge Biomedical Campus, Cambridge CB2 0QQ, UK. <sup>9</sup>Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge CB2 0BB, UK. <sup>10</sup>Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School, 1 Blackfan Circle, Boston, MA 02115, USA. <sup>11</sup>Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, 450 Brookline Ave, Boston, MA 02115, USA. <sup>12</sup>Broad Institute of MIT and Harvard, 415 Main St, Cambridge, MA 02142, USA. <sup>13</sup>Harvard-MIT Health Sciences and Technology, Harvard Medical School, 77 Massachusetts Ave, Cambridge, MA 02139, USA. <sup>14</sup>Department of Haematology, Barts Health National Health Service Trust, London E1 1BB, UK. <sup>15</sup>National Institute for Health and Care Research Cambridge BioResource, Box 229, Addenbrooke's Hospital, Cambridge Biomedical Campus, Cambridge CB2 0QQ, UK. <sup>16</sup>Wellcome Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. <sup>17</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. <sup>18</sup>European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany. <sup>19</sup>Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. <sup>20</sup>Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK. <sup>21</sup>Nuffield Division of Clinical Laboratory Sciences, Radcliffe Department of Medicine, University of Oxford, Headley Way, Headington, Oxford OX3 9DU, UK. <sup>22</sup>National Institute for Health Research Oxford Biomedical Research Centre—Haematology Theme, John Radcliffe Hospital, Headley Way, Headington, Oxford OX3 9DU, UK. <sup>23</sup>National Health Service Blood and Transplant, Oxford Centre, John Radcliffe Hospital, Headley Way, Headington, Oxford OX3 9DU, UK. <sup>24</sup>Health Data Science Research Centre, Fondazione Human Technopole, Viale Rita Levi Montalcini 1, Milan 20157, Italy. <sup>25</sup>Department of Clinical and Biomedical Sciences, University of Exeter Medical School, Faculty of Health and Life Sciences, RILD Building, Barrack Road, Exeter EX2 5DW, UK. <sup>26</sup>Department of Pediatric

Immunology, Rheumatology and Infectious Disease, Emma Children's Hospital, Amsterdam University Medical Center, Amsterdam CB2 OPT, UK. <sup>27</sup>Department of Blood Cell Research, Sanquin Research and Landsteiner Laboratory, Sanquin, University of Amsterdam, Amsterdam, Netherlands. <sup>28</sup>Department of Immunology and Inflammation, Imperial College London, Commonwealth Building, The Hammersmith Hospital, Du Cane Road, London W12 0NN, UK. <sup>29</sup>Department of Haematology, University College London Hospitals, WC1E 6AS London, UK. <sup>30</sup>Genomics Research Centre, Fondazione Human Technopole, Viale Rita Levi Montalcini 1, Milan 20157, Italy. <sup>31</sup>Deceased: Jennifer Sambrook <sup>32</sup>These authors jointly supervised this work: Adam S. Butterworth, Willem H. Ouwehand, Nicole Soranzo, William J. Astle. ✉ e-mail: [asb38@medschl.cam.ac.uk](mailto:asb38@medschl.cam.ac.uk); [who1000@cam.ac.uk](mailto:who1000@cam.ac.uk); [ns6@sanger.ac.uk](mailto:ns6@sanger.ac.uk); [wja24@cam.ac.uk](mailto:wja24@cam.ac.uk)