

Interrogation of human hematopoiesis at single-cell and single-variant resolution

Jacob C. Ulirsch ^{1,2,3,4,17}, Caleb A. Lareau ^{1,2,3,4,5,17}, Erik L. Bao ^{1,2,3,6,17}, Leif S. Ludwig ^{1,2,3}, Michael H. Guo ^{3,7,8,9}, Christian Benner ^{10,11}, Ansuman T. Satpathy ¹², Vinay K. Kartha ^{3,13}, Rany M. Salem ¹³, Joel N. Hirschhorn ^{3,7,8,9}, Hilary K. Finucane ¹³, Martin J. Aryee ¹³, Jason D. Buenrostro ^{13*} and Vijay G. Sankaran ^{1,2,3,16*}

Widespread linkage disequilibrium and incomplete annotation of cell-to-cell state variation represent substantial challenges to elucidating mechanisms of trait-associated genetic variation. Here we perform genetic fine-mapping for blood cell traits in the UK Biobank to identify putative causal variants. These variants are enriched in genes encoding proteins in trait-relevant biological pathways and in accessible chromatin of hematopoietic progenitors. For regulatory variants, we explore patterns of developmental enhancer activity, predict molecular mechanisms, and identify likely target genes. In several instances, we localize multiple independent variants to the same regulatory element or gene. We further observe that variants with pleiotropic effects preferentially act in common progenitor populations to direct the production of distinct lineages. Finally, we leverage fine-mapped variants in conjunction with continuous epigenomic annotations to identify trait-cell type enrichments within closely related populations and in single cells. Our study provides a comprehensive framework for single-variant and single-cell analyses of genetic associations.

Hematopoiesis is a paradigm of cellular differentiation that is highly coordinated to ensure balanced proportions of mature blood cells¹. Despite a sophisticated understanding gained primarily from model organisms, many aspects of this process remain poorly understood in humans. At the population level, there is substantial variation in commonly measured blood cell traits, such as hemoglobin levels and specific blood cell counts, which can manifest as diseases at extreme ends of the spectrum². Identifying genetic variants that drive these differences in blood cell traits in human populations may reveal regulatory mechanisms and genes critical for blood cell production and hematological diseases.

To these ends, genome-wide association studies (GWAS) have identified thousands of genomic loci linked to complex phenotypes, including blood cell traits³, but a major challenge has been the identification of causal genetic variants and relevant cell types underlying the observed associations⁴. In particular, linkage disequilibrium (LD) has confounded the precise identification of functional variants. In an effort to address these issues, several analytical approaches have been developed. The first, termed genetic fine-mapping, attempts to resolve trait-associated loci to likely causal variants by modeling LD structure and the strength of associations. In practice, a major limitation has been the computational burden imposed when allowing for multiple causal variants and methods that assume exactly one causal variant per locus are thus most commonly used^{5,6}, despite strong evidence that many loci contain multiple independent associations^{7–10}.

The second suite of approaches focus instead on identifying functional tissue enrichments. It has been well established that ~80–90% of associated loci do not tag coding variants and that ~40–80% of the narrow-sense heritability for many complex traits can be resolved to genomic regulatory regions^{11,12}. Given this observation, tissue-specific measurements of regulatory-element activity are often overlapped with significant loci (for example, in epigenomic fine-mapping) or with polygenic signal from millions of variants (for example, in partitioned heritability) to identify the variants and cell types most likely to underlie the measured trait or disease^{11,13}. These enrichment methods have identified causal tissues for diseases, including pancreatic islets for diabetes¹³ and central nervous system cells for schizophrenia¹¹, but are only beginning to be applied to highly related traits and cell types within single systems such as the hematopoietic hierarchy.

To gain insights into hematopoietic lineage commitment and differentiation, we performed GWAS and genetic fine-mapping for 16 blood cell traits on individuals from the UK Biobank (UKB)³, identifying multiple likely causal variants in hundreds of individual regions. We comprehensively annotated fine-mapped variants and identified high-confidence molecular mechanisms and putative target genes at scale. This allowed us not only to gain insights into patterns of developmental regulation but also to learn about the pleiotropic regulatory processes underlying blood cell production and maintenance. Finally, we describe and validate a new method

¹Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. ²Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. ³Broad Institute of MIT and Harvard, Cambridge, MA, USA. ⁴Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA. ⁵Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. ⁶Harvard-MIT Health Sciences and Technology, Harvard Medical School, Boston, MA, USA. ⁷Division of Endocrinology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. ⁸Department of Genetics, Harvard Medical School, Boston, MA, USA. ⁹Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA, USA. ¹⁰Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland. ¹¹Department of Public Health, University of Helsinki, Helsinki, Finland. ¹²Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. ¹³Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA. ¹⁴Schmidt Fellows Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA. ¹⁵Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. ¹⁶Harvard Stem Cell Institute, Cambridge, MA, USA. ¹⁷These authors contributed equally: Jacob C. Ulirsch, Caleb A. Lareau, Erik L. Bao.

*e-mail: jason_buenrostro@harvard.edu; sankaran@broadinstitute.org

(g-chromVAR) to discriminate between closely related cell types in an effort to identify relevant stages of hematopoiesis that are affected by these common genetic variants. Applying g-chromVAR to data from single hematopoietic cells revealed substantial heterogeneity of genetic enrichment within classically defined hematopoietic progenitor populations. Thus, we demonstrate that using a well-powered method to identify trait-relevant cell populations provides a critical step toward broadly deciphering causal mechanisms underlying phenotypic variation.

Results

Fine-mapping pinpoints hundreds of likely causal variants. We performed GWAS on ~115,000 individuals from the UKB for 16 blood cell traits representing seven distinct hematopoietic lineages (erythroid, platelet, lymphocyte, monocyte, and granulocyte (neutrophil, eosinophil, and basophil)) (Fig. 1a). Similarly to previous reports, these traits were highly heritable, with common genetic variants explaining an average of 15.4% of narrow-sense heritability (h_g^2)¹⁴ (Supplementary Fig. 1). Traits from the same lineage typically had high genetic correlations, such as red blood cell (RBC) count and hemoglobin ($r_g = 0.89$, $P = 7.1 \times 10^{-25}$), whereas traits from distinct lineages had low genetic correlations, with some exceptions such as platelet count and lymphocyte count ($r_g = 0.26$, $P = 3.8 \times 10^{-18}$) (Supplementary Fig. 1). This suggests that genetic regulation of blood production could potentially occur across various stages of hematopoiesis.

To begin to dissect the nature and stage specificity of these genetic effects, we performed genetic fine-mapping to identify high-confidence variants across 2,056 3-Mb regions containing a genome-wide-significant association. Traditional fine-mapping approaches assume only one causal variant per locus and either are agnostic to LD or use small reference panels, which are inaccurate when scaled to large sample sizes¹⁵. To overcome these limitations, we calculated LD directly from the imputed genotype probabilities (dosages) for individuals in our GWAS, rather than from a hard-called reference panel (Fig. 1b).

Across all common variants (minor allele frequency (MAF) > 0.1%, INFO¹⁶ > 0.6) in 2,056 regions, our method identified 38,654 variants with >1% posterior probability (PP) of being causal for a trait association, representing a substantial proportion of the narrow-sense heritability explained by all common variants (trait average of 24.9% of the common variant h_g^2 for PP > 0.01) (Supplementary Fig. 1 and Supplementary Table 1). 993 regions (48%) contained at least one variant with PP > 0.50 (Fig. 1c), providing strong evidence that our approach was successful in pinpointing causal variants. The posterior expected number of independent causal variants was greater than two for 35% of regions and greater than three for 13% of regions (Fig. 1d). Given their increased complexity, regions with a greater expected number of causal variants had lower top-configuration posterior probabilities (Supplementary Fig. 2 and Supplementary Table 2). The majority of variants (74%) with PP > 0.75 had MAF > 5% (Fig. 1e), consistent with the known polygenic nature of blood cell traits³. Fine-mapped variants had potentially diverse mechanisms, ranging from putative regulatory variants in accessible chromatin to coding variants, including 164 unique missense variants and 6 loss-of-function variants with PP > 0.10 (Fig. 1f, Supplementary Fig. 3, and Supplementary Table 3).

To validate our approach, we investigated the overlap of fine-mapped variants (binned by posterior probability) with several annotations previously shown to be enriched for GWAS signals (Fig. 1g)^{11,12}. To generate a null distribution, we locally shifted annotations within a 3-Mb window, similarly to the method implemented in GoShifter¹⁷. We observed minimal enrichment for intronic regions and UTRs of genes, but found strong, focal, and stepwise enrichments across bins with higher posterior probabilities for hematopoietic accessible chromatin, promoters, and coding

regions (odds ratio (OR)=4.2, 2.9, and 8.5 for PP > 0.75, respectively) (Fig. 1f)^{11,12,17}. Notably, strong enrichments persisted even after we excluded all variants with high correlation ($r^2 > 0.8$) to the sentinel variants at each locus (Supplementary Fig. 3).

Dissecting mechanisms of core gene regulation in hematopoiesis. We next sought to delineate the precise mechanisms underlying the effects of fine-mapped genetic variants on hematopoietic traits. For all 140,739 variants with PP > 0.001, we combined several lines of functional and predictive evidence to better understand (i) the cell populations, (ii) the molecular mechanisms, and (iii) the target genes involved in blood cell production (Supplementary Fig. 4). First, we identified fine-mapped (PP > 0.10) nonsynonymous and loss-of-function coding variants in genes associated with RBC (77 genes), platelet (59), monocyte (20), lymphocyte (28), and granulocyte (neutrophil, basophil, and eosinophil; 46) traits (Supplementary Table 3). Within the set of genes identified from variants associated with RBC traits, we found both validated GWAS genes (*SH2B3* (ref. ¹⁸) and *TRIM58* (ref. ¹⁹) (Supplementary Fig. 5) and several genes linked to diverse Mendelian disorders involving RBCs (*HFE*, *TMPRSS6*, *PFKM*, *PKLR*, *PIEZ01*, *SPTA1*, *ANK1*, *RHD*, *GYPA*, and *KLF1*)²⁰. Genes perturbed by fine-mapped coding variants were enriched for known and novel trait-relevant biological pathways. For example, genes associated with RBC traits were involved in iron homeostasis, genes for platelet traits were involved in coagulation and wound healing, genes for lymphocyte traits were involved in T cell migration and activation, and genes for monocyte and granulocyte traits were involved in cytokine and inflammatory responses (Supplementary Fig. 6 and Supplementary Table 3). Of note, we identified several pathways corresponding to cholesterol and lipid regulation that were enriched in genes linked to RBC traits (Supplementary Fig. 6), suggesting a connection between lipid metabolism and RBCs, which are major stores of cholesterol²¹.

To investigate the exact stages of hematopoietic differentiation during which variants could regulate transcription, we overlapped fine-mapped variants (PP > 0.10) with chromatin accessibility profiles (ATAC-seq) for 18 hematopoietic progenitor, precursor, and differentiated cell populations primarily sorted from the bone marrow or blood of healthy donors (Fig. 1a, Supplementary Fig. 7, and Supplementary Table 4). Across traits representing the five major blood cell lineages, we used *k*-means clustering to categorize the developmental timing of accessible chromatin peaks containing fine-mapped variants (Fig. 2a,b and Supplementary Fig. 8). For example, across RBC traits, we identified 80 fine-mapped regulatory variants, of which 26% (21/80) were restricted to erythroid progenitors, 18% (14/80) were restricted to megakaryocyte-erythroid progenitors (MEPs) and erythroid progenitors, and 29% (23/80) could regulate transcription across the entire erythroid lineage from hematopoietic stem cells (HSCs) to erythroid progenitors, whereas 14% (11/80) could only act in other hematopoietic lineages (Fig. 2a). In some cases, we identified small clusters of variants that followed slightly different regulatory programs, such as variants that could only regulate transcription in upstream multipotent progenitors and variants associated with lymphocyte count that could regulate transcription in T cell, but not B cell, subsets (Fig. 2a,b and Supplementary Fig. 8).

Next, we investigated the molecular mechanisms underlying fine-mapped regulatory variants. To nominate a high-confidence molecular mechanism, we required that a variant (i) disrupt one of 426 motifs corresponding to known binding preferences for human transcription factors²² and (ii) show occupancy by the corresponding transcription factor in a relevant primary hematopoietic tissue or hematopoietic cell line, on the basis of 2,115 uniformly processed ChIP-seq profiles²³. In total, we identified one or more such mechanisms for 145 distinct fine-mapped noncoding variants (Fig. 2c). Specifically, we identified 13 RBC, 28 platelet, 8 monocyte, 11 lym-

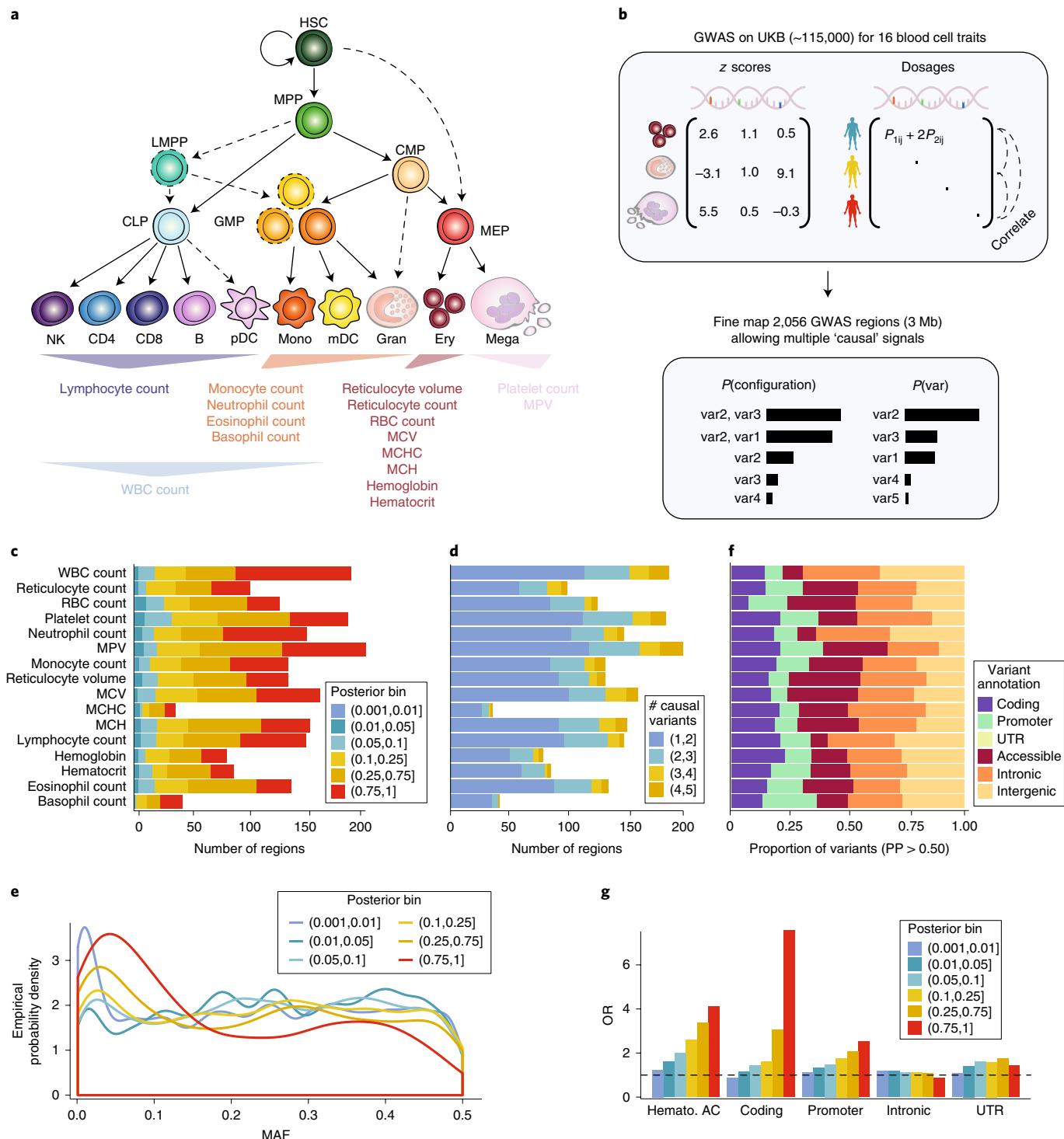


Fig. 1 | Overview of hematopoiesis, UKB GWAS, and fine-mapping. **a**, Schematic of the human hematopoietic hierarchy showing the primary cell types analyzed in this work. Colors used in this schematic are consistent throughout all figures; mono, monocyte; gran, granulocyte; ery, erythroid; mega, megakaryocyte; CD4, CD4+ T cell; CD8, CD8+ T cell; B, B cell; NK, natural killer cell; mDC, myeloid dendritic cell; pDC, plasmacytoid dendritic cell; MPP, multipotent progenitor; LMPP, lymphoid-primed multipotent progenitor; CMP, common myeloid progenitor; CLP, common lymphoid progenitor; GMP, granulocyte-macrophage progenitor; MEP, megakaryocyte-erythroid progenitor. The 16 blood traits that were genetically fine-mapped are shown below the hierarchy; WBC, white blood cell; MPV, mean platelet volume; MCV, mean corpuscular volume; MCHC, mean corpuscular hemoglobin concentration; MCH, mean corpuscular hemoglobin. **b**, Schematic of the UKB GWAS and fine-mapping approach. Briefly, blood traits for ~115,000 individuals were fine-mapped, allowing for multiple causal variants and using imputed genotype dosages as the reference for LD. **c**, Number of fine-mapped regions for each trait; the highest posterior probability of a variant being causal is indicated. **d**, Breakdown of the number of causal variants (min=1, max=5) for all regions in each trait. **e**, Empirical distribution of the MAF of variants in each posterior probability bin. **f**, Proportion of fine-mapped variants within intronic, promoter, coding, UTR, and intergenic regions. **g**, Local-shifting enrichments of fine-mapped variants across all traits for varying posterior probability bins. AC, accessible chromatin.

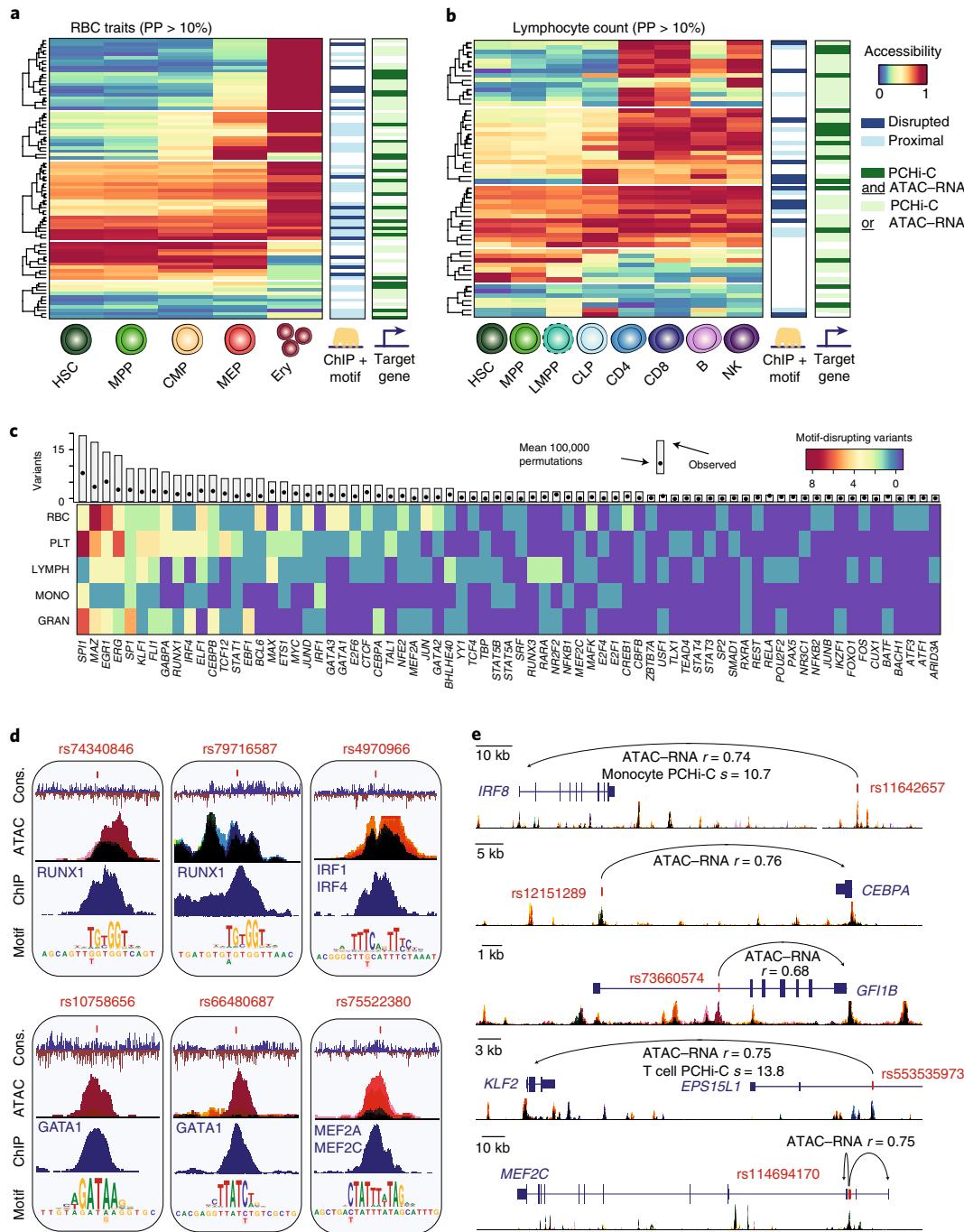


Fig. 2 | Mechanisms of core gene regulation in blood production. **a,b**, Heat maps depicting RBC-trait-associated variants (PP > 0.10) across the erythroid lineage (**a**) and lymphocyte-count-associated variants (PP > 0.10) across the lymphoid lineage (**b**), with clustering by chromatin accessibility. Each row represents a fine-mapped variant, each column denotes a cell type within the relevant lineage, and color corresponds to relative chromatin accessibility along the lineage at each variant (blue, least accessible chromatin; red, most accessible chromatin). Putative target genes (predicted by ATAC-RNA correlation and/or PCHi-C) and disrupted transcription factors (predicted by ChIP-seq occupancy and motif disruption) are indicated to the right. **c**, Transcription factor motifs disrupted in lineage-specific hematopoietic traits. Each row represents a set of traits where variants disrupt specified transcription factor motifs and are occupied by the respective transcription factor in hematopoietic cells. The unique margin sums across each lineage are shown in the bar plot for each transcription factor. The expected number of variants with evidence of ChIP-seq plus motif disruption across all posterior probabilities was estimated by using 100,000 permutations and is shown as a single point. PLT, platelet; LYMPH, lymphocyte; MONO, monocyte; GRAN, granulocyte. **d**, Examples of molecular mechanisms identified from the analysis in **c**, including putative causal variants that disrupt binding in cis of transcription factors known to be involved in regulating hematopoiesis for various blood cell traits: rs10758656 and rs66480687 are associated with RBC traits; rs75522380 and rs74340846 are associated with platelet traits; rs4970966 is associated with monocyte count; and rs79716587 is associated with lymphocyte count. In the ATAC-seq plots, black represents accessibility throughout hematopoiesis whereas other stacked colors represent accessibility for the cell types shown in Fig. 3d. Cons., conservation. **e**, Examples of putative target genes identified from the analysis in **a** and **b**: rs11642657 and rs12151289 are associated with monocyte count; rs73660574 is associated with RBC traits; rs553535973 is associated with lymphocyte count; and rs114694170 is associated with platelet traits. Colors for accessible chromatin are the same as in **d**. In PCHi-C, s indicates the CHiCAGO interaction score from ref. ³³.

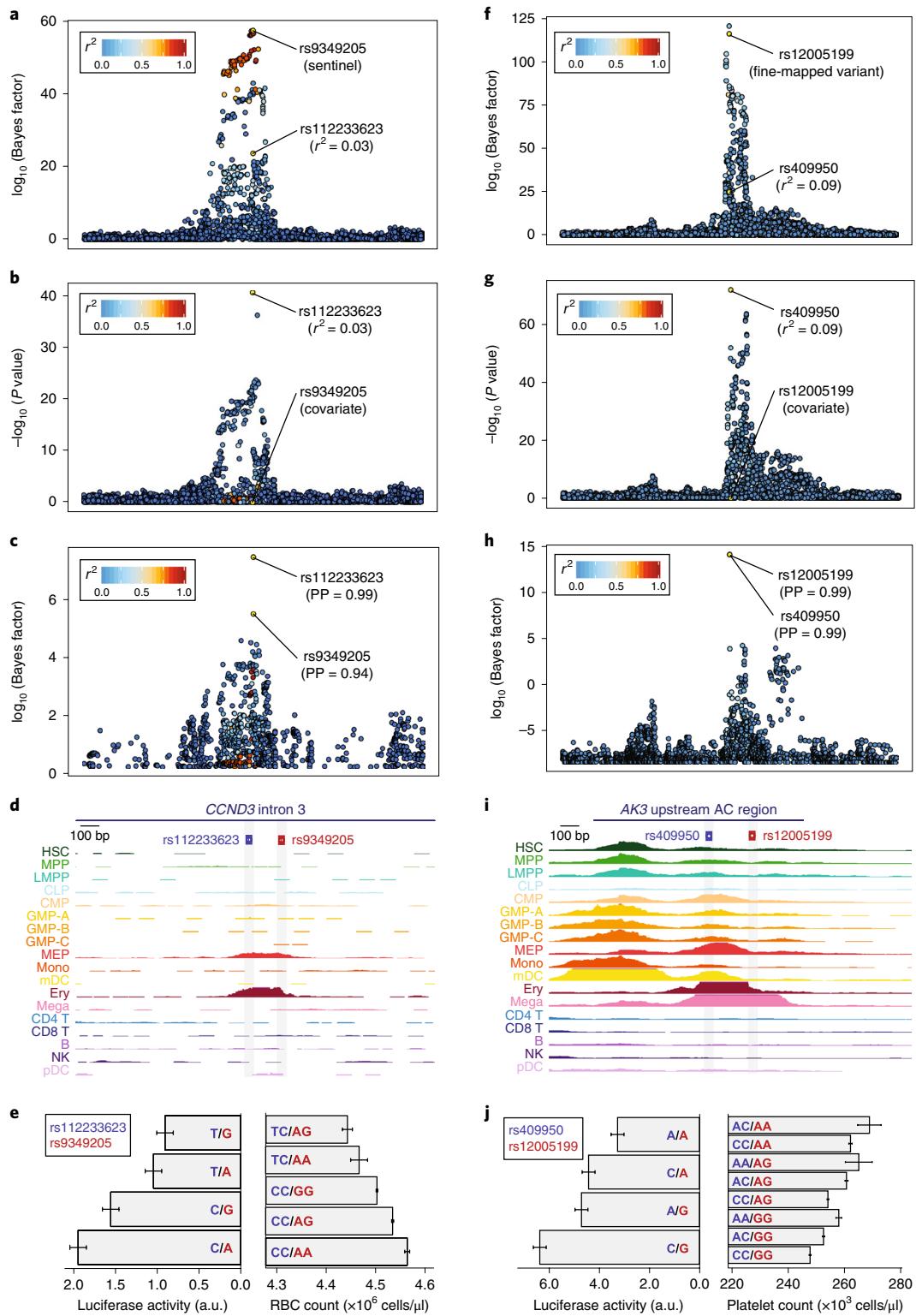


Fig. 3 | Characterization and validation of the *CCND3* and *AK3* regions with multiple causal variants. **a,b**, Regional association plots ($n=116,667$ individuals; BOLT-LMM P values) for RBC count in the *CCND3* locus from the initial GWAS (**a**) and after conditioning on the sentinel variant, rs9349205 (**b**). **c,d**, Fine-mapping identifies two putative causal variants (rs9349205, PP = 0.94; rs112233623, PP = 0.99) located 161 bp apart (**c**), both of which lie within the same erythroid-specific accessible chromatin (**d**). **e**, Luciferase reporter assays ($n=9$ biological replicates) for four haplotypes (left) corroborate independent additive effects for rs9349205 (red; two-sided Wald test $P=1.78\times 10^{-3}$) and rs112233623 (blue; two-sided Wald test $P=2.86\times 10^{-6}$) on RBC count (right). a.u., arbitrary units. **f,g**, Regional association plots ($n=116,666$ individuals, BOLT-LMM P values) for platelet count in the *AK3* locus from the initial GWAS (**f**) and after conditioning on the sentinel variant, rs12005199 (**g**). **h,i**, Fine-mapping identifies two putative causal variants (rs12005199, PP = 0.99; rs409950, PP = 0.99) 123 bp apart (**h**), both located within a strong megakaryocyte-specific accessible chromatin region (**i**). **j**, Luciferase reporter assays ($n=9$ biological replicates) for four haplotypes (left) corroborate independent additive effects for rs12005199 (red; two-sided Wald test, $P=5.19\times 10^{-4}$) and rs409950 (blue; two-sided Wald test, $P=3.57\times 10^{-5}$) on platelet count (right). In **e** and **j**, mean and standard error are shown for both phenotype and regulatory activity.

phocyte, and 18 granulocyte high-confidence molecular mechanisms for variants also in accessible chromatin in primary hematopoietic tissue (Fig. 2a,b, Supplementary Fig. 8, and Supplementary Table 5). These variants most commonly disrupted the binding sites of key transcriptional regulators of hematopoietic lineage commitment and differentiation (false-discovery rate (FDR) < 10% for 33 transcription factors). For example, we observed seven PU.1 (SPI1)^{24,25}, six ERG^{26–28}, four FLI1 (refs. ^{28,29}), three IRF4 (ref. ³⁰), and three RUNX1 (refs. ^{31,32}) binding-site-disrupting variants associated with platelet traits (Fig. 2c,d), in addition to many other compelling lineage-specific regulatory mechanisms for experimental follow-up (Supplementary Fig. 8 and Supplementary Note).

Finally, to identify high-confidence target genes for fine-mapped regulatory variants, we built hematopoietic-specific enhancer-promoter maps by using (i) measurements of physical DNA interactions in 15 primary hematopoietic cell populations from promoter capture Hi-C (PCHi-C)³³ and (ii) the correlation between chromatin accessibility and expression of genes in cis across 16 primary hematopoietic populations^{34,35}. Altogether, we identified one or more experimentally supported target genes for 415 variant-trait associations, providing testable biological hypotheses for 79% of the fine-mapped regulatory variants (Fig. 2a,b, Supplementary Figs. 5 and 8, and Supplementary Tables 6 and 7). Interestingly, a number of variants were predicted to alter the transcription of genes encoding hematopoietic transcription factors (Fig. 2d,e and Supplementary Fig. 8). For example, *IRF8* and *CEBPA*, which encode two essential transcription factors involved in monocyte differentiation^{36,37}, were targets of fine-mapped variants associated with monocyte count that fell within accessible chromatin in monocyte precursors (Fig. 2e). Similarly, we determined that *GFI1B*, *KLF2*, and *MEF2C* were targets of fine-mapped variants in progenitor-specific accessible chromatin associated with mean reticulocyte volume, lymphocyte count, and platelet count, respectively (Fig. 2e). Overall, this functional analysis will likely facilitate experimental investigation into how common genetic variants regulate hematopoietic lineage commitment and differentiation.

Regions with multiple causal variants. We next conducted a closer examination of the 785 trait-associated regions with multiple independent causal signals. Among proximal pairs of variants in which both variants had PP > 0.50, the majority were >10 kb apart (76%), although the variants in seven pairs were within fewer than 100 bp of each other (Supplementary Fig. 9 and Supplementary Table 8). Across all pairs, 42% of the variants were of the same class (for example, coding-coding variants), and pairs of variants in accessible chromatin but in different regulatory regions within 1 Mb of each other were typically lineage specific (Supplementary Fig. 9). Examples of coding-coding pairs included hemoglobin-associated rs1800730 and rs1799945 (PP > 0.66; 4 bp apart) in *HFE*, the classic gene mutated in hereditary hemochromatosis; white blood cell (WBC)-count-associated rs146125856 and rs148783236 (PP > 0.98; 24 bp apart) in *USP8*, which encodes an immune-specific ubiquitin ligase and is mutated in Cushing's disease^{38,39}; and mean platelet volume (MPV)-associated rs41303899 and rs415064 (PP > 0.76; 835 bp apart) in *TUBB1*, which encodes a β-tubulin protein important for proplatelet formation that is mutated in monogenic forms of macrothrombocytopenia⁴⁰.

Although there were several other interesting pairs of variants in accessible chromatin (Supplementary Note and Supplementary Fig. 10), we specifically investigated the association with RBC count at the *CCND3* locus, in which we previously identified a causal variant and its target gene⁴¹. At this locus, our current approach correctly identified the known causal variant (rs9349205) as the primary association, as well as ~4 additional independent signals, including a secondary imputed variant (rs112233623) associated with decreased RBC count (Fig. 3a–c). Stepwise conditional

analysis further validated these findings (Fig. 3b). Notably, these variants were missed by fine-mapping if we instead used LD estimated from either the UK10K whole-genome sequencing (WGS) reference panel or hard-called variants from the UKB population (Supplementary Fig. 11), highlighting the importance of calculating LD by using imputed genotype dosages from the GWAS population. Remarkably, rs112233623 is only 161 bp from rs9349205, and both fell within erythroid-specific accessible chromatin (Fig. 3d). Luciferase reporter assays showed that each variant affected enhancer activity independently with the minor alleles acting in opposing directions, consistent with the genetic directionality (Fig. 3e). At a separate locus associated with platelet traits, we similarly observed a large number of independent signals (approximately eight), which allowed us to identify a variant pair (rs49950 and rs12005199; PP > 0.99; 123 bp apart) within a single accessible chromatin region ~20 kb upstream of *AK3*, a gene whose zebrafish homolog is essential for platelet (thrombocyte) formation (Fig. 3f–i)⁴². Notably, we again observed that each variant significantly affected enhancer activity additively and in concordance with the population phenotypes (Fig. 3j).

Mechanisms of pleiotropic variants across distinct blood cell lineages. We next sought to examine the effects of variants associated with two or more of the seven distinct blood cell types for which phenotypes were available in the UKB. We hypothesized that these pleiotropic variants could either (i) 'tune' overall blood production by simultaneously increasing or decreasing the levels of terminal blood cells across multiple lineages or (ii) 'switch' blood cell production such that one lineage would be favored at the expense of others (Fig. 4a).

We restricted our analyses to quantified blood cell counts for interpretability and identified 172 pleiotropic variants that colocalized⁴³ (PP > 0.10) to two or more traits (Fig. 4b–d, Supplementary Fig. 12, and Supplementary Table 9). Surprisingly, 91% (156/172) of these variants exhibited a tuning mechanism, modifying two or more lineages in the same direction, whereas the remaining 9% (16/172) favored one lineage at the expense of other lineages ($P = 5.08 \times 10^{-30}$, binomial test). Regardless of direction of effect, 88% of all pleiotropic variants were noncoding, and those in regions of accessible chromatin had 60% more ATAC-seq reads in progenitors than in terminal cell types (mean of 4.01 versus 2.44 counts per million; $P = 0.025$, Student's *t* test), consistent with the hypothesis that many of these variants act in common progenitor cell populations^{44,45}.

One example of a variant exhibiting a switch mechanism is rs78744187 (PP = 0.99 and 0.99), which increased RBC count while concomitantly decreasing basophil count (Fig. 4c). rs78744187 is located in an enhancer specific for common myeloid progenitors (CMPs), a heterogeneous population containing progenitors for both basophils and RBCs, approximately 36 kb downstream of *CEBPA*, which encodes a key myeloid transcription factor⁴⁶. We previously reported the association between rs78744187 and basophil count, but not RBC count, and showed that this variant was a switch for production of the closely related basophil and mast cell lineages⁴⁵. A second switch variant, rs218265 (PP = 0.99 and 0.64), located within a gene desert 1.15 Mb upstream of *KIT*, increased neutrophil count but decreased RBC count. *KIT* encodes the receptor protein for stem cell factor, a growth-stimulating cytokine involved in hematopoietic progenitor cell proliferation⁴⁷. rs218265 falls within a region of accessible chromatin that is exclusively open in multipotent and heterogeneous populations (Fig. 4d), consistent with a role for this enhancer variant in regulating *KIT* expression in the common progenitors of neutrophils and RBCs. Taken together, our results suggest that tuning the dosage of key regulatory genes in upstream progenitors may switch the production of one lineage in favor of another during the early stages of lineage commitment.

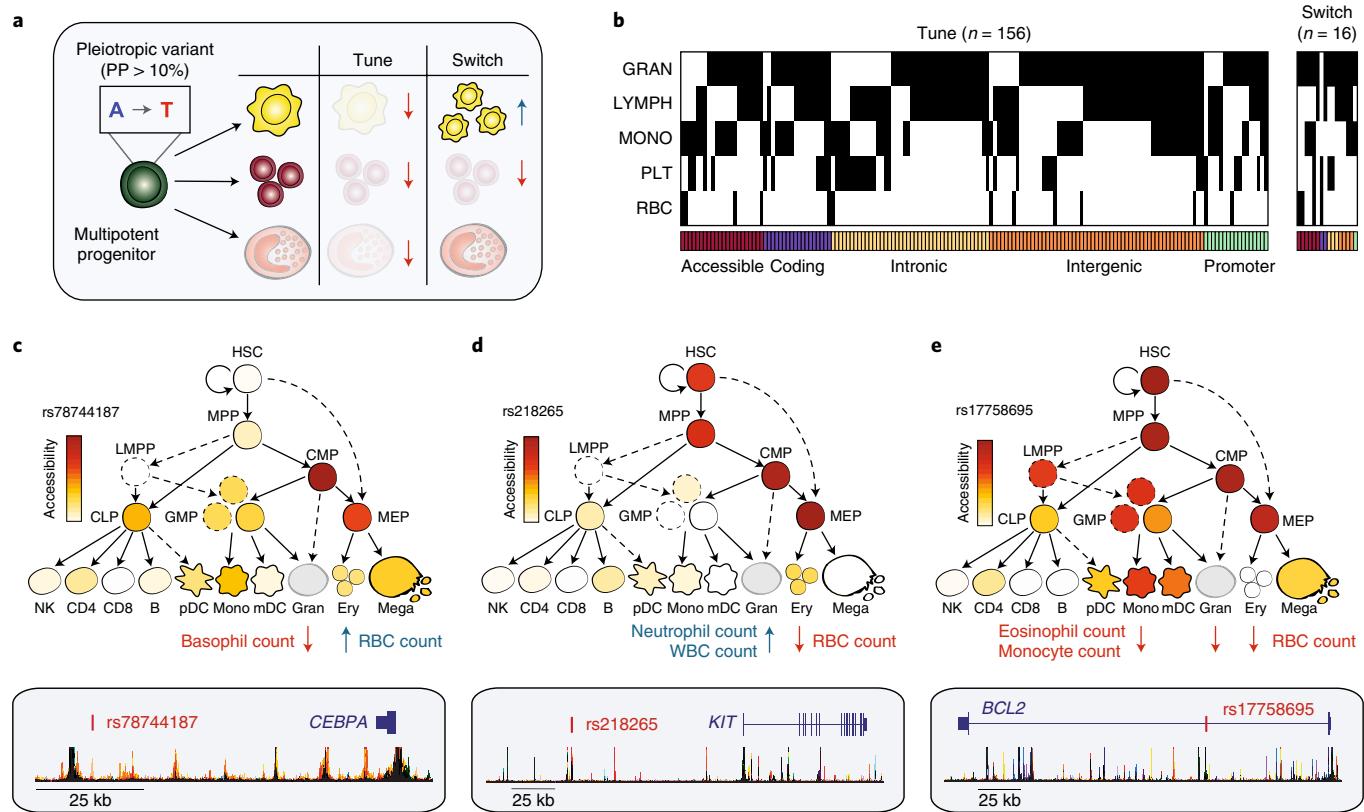


Fig. 4 | Dissecting the mechanisms of pleiotropic variants across multiple blood cell lineages. **a**, Schematic illustrating fine-mapped variants acting in multipotent or heterogeneous progenitors on distinct hematopoietic lineages, by either tuning lineages in the same direction or switching the regulation in opposite directions. **b**, Heat map depicting 172 fine-mapped variants (PP > 0.10) with pleiotropic effects on cell counts in two or more hematopoietic lineages. Effects on eosinophil, neutrophil, and basophil counts are visualized together as a single granulocyte lineage. Genomic annotation is indicated below each variant. **c**, Pleiotropic variant rs78744187, located downstream of *CEBPA*, has high chromatin accessibility in CMPs and MEPs (top) and demonstrates a switch mechanism by downregulating basophil count while upregulating RBC count (bottom). **d**, rs218265, located upstream of the *KIT* gene encoding stem cell factor receptor, has high chromatin accessibility in several early progenitors (HSCs, MPPs, CMPs, and MEPs) and demonstrates a switch mechanism by upregulating neutrophil and WBC count while downregulating RBC count. **e**, rs17758695, located within an intron of the antiapoptotic factor *BCL2*, has high chromatin accessibility in several early progenitors (HSCs, MPPs, CMPs, and MEPs) and exhibits a tuning mechanism, simultaneously downregulating eosinophil, monocyte, and RBC counts.

As an example of a pleiotropic variant exhibiting the predominant tuning mechanism, we found that rs17758695 (PP = 0.99, 0.99, and 0.99) was associated with decreases in eosinophil, monocyte, and RBC count (Fig. 4e). This variant is located within a progenitor-specific region of accessible chromatin in the intron of *BCL2*, which encodes an antiapoptotic protein known to regulate hematopoietic differentiation⁴⁸. This is consistent with the idea that regulating a general cell death protein such as *BCL2* in a common multipotent progenitor would tune the production of multiple cell types, in contrast to the switch variants proximal to key regulators of hematopoietic differentiation. An additional tuning variant is the missense variant rs12459419 (PP = 0.30, 0.28, and 0.11) in the *CD33* gene, which was associated with decreases in eosinophil, monocyte, and platelet count. *CD33* is broadly expressed in hematopoietic progenitors and encodes a surface marker of myeloid differentiation⁴⁹ (Supplementary Fig. 12). In summary, our analyses support a prominent role for pleiotropy in hematopoietic differentiation, whereby individual variants can act in upstream progenitors to simultaneously tune or switch production and maintenance of multiple lineages.

g-chromVAR, a new method to measure fine-mapped GWAS trait enrichment among closely related tissues.

We next shifted

our focus in the reciprocal direction—by using fine-mapping to determine the exact stages of human hematopoiesis at which the regulatory genetic variation underlying each blood cell trait is most likely acting. Although methods^{11,17} have recently been developed to calculate enrichment of genetic variation for genomic annotations, a method that takes into account both (i) the strength and specificity of the genomic annotation and (ii) the probability of variant causality, while accounting for LD structure, is needed to resolve associations within the closely related, stepwise hierarchies that define hematopoiesis. To this end, we developed a new approach called genetic-chromVAR (g-chromVAR), a generalization of the recently described chromVAR method⁵⁰, to measure the enrichment of regulatory variants in each cell state by using fine-mapped variant posterior probabilities and quantitative measurements of regulatory activity (Fig. 5a; details in Supplementary Note and Methods). We show that g-chromVAR is generally robust to variant posterior probability thresholds and numbers of background peaks (Supplementary Fig. 13), captures true enrichments in a simulated setting (Supplementary Fig. 14), is robust to the choice of fine-mapping method (Supplementary Table 10), and can identify novel enrichments in large epigenomic datasets (Supplementary Table 11; details in Supplementary Note).

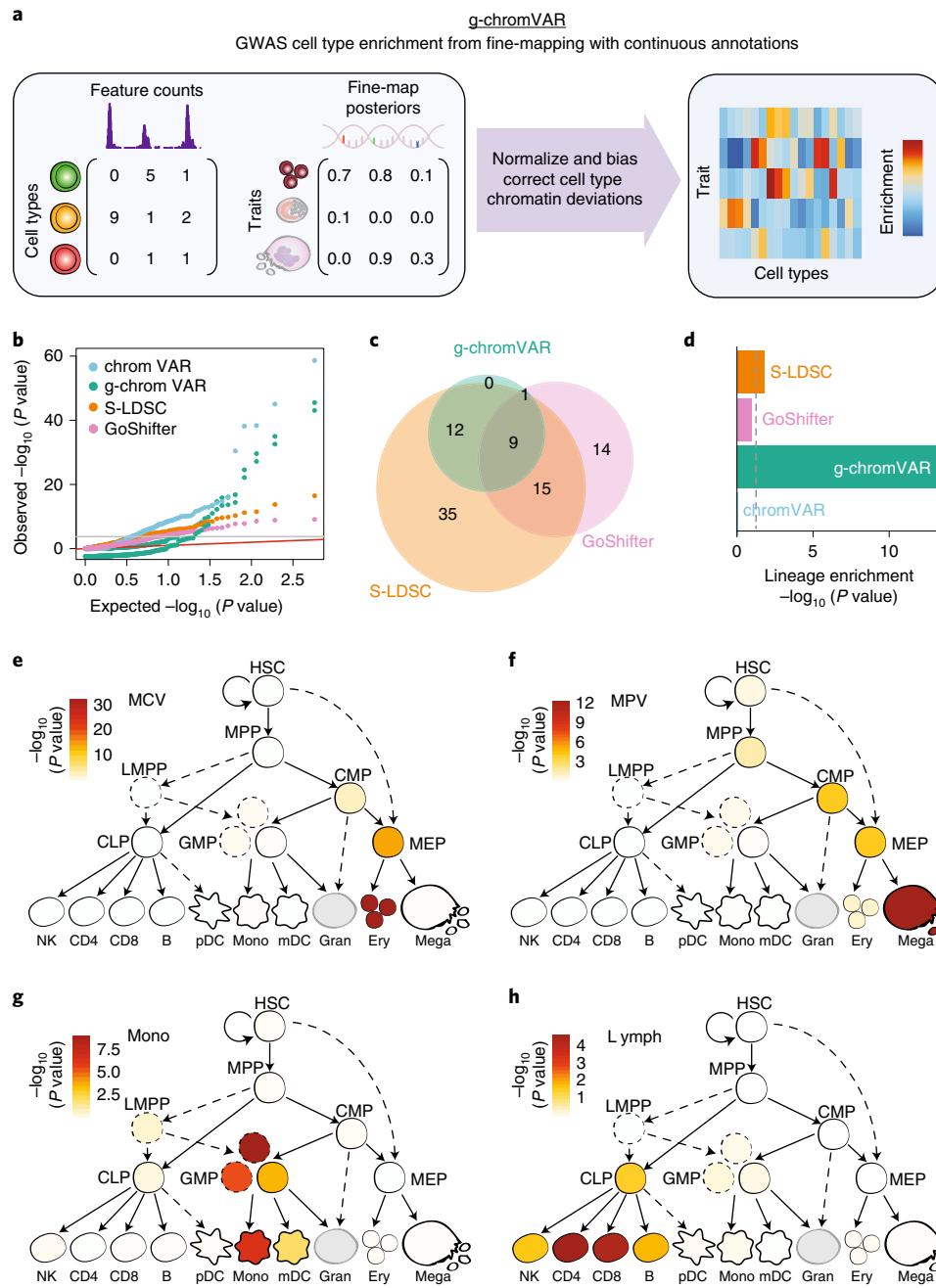


Fig. 5 | Overview of g-chromVAR and application to hematopoietic cell types. **a**, Schematic showing inputs for continuous epigenomic data for each cell type and a matrix of fine-mapped variant posterior probabilities for GWAS traits. **b-d**, Results from application of g-chromVAR and three similar methods to 16 blood cell traits for 18 hematopoietic cell types. **b**, Quantile-quantile representation of the P values from each method. **c**, Overlap between methods for Bonferroni-corrected trait enrichments. **d**, Lineage enrichment of all trait pairs ($n=288$ pairs) for each method. A two-tailed Mann-Whitney rank-sum test was used to evaluate the relative enrichment of lineage-specific trait-cell type pairs (true positives). **e-h**, Enrichments for four representative traits obtained by using g-chromVAR: mean corpuscular volume (**e**), mean platelet volume (**f**), monocyte count (**g**), and lymphocyte count (**h**).

To validate g-chromVAR in a realistic setting, we used it along with seven other methods to estimate the enrichment of each of the 16 blood cell traits within the accessible chromatin of 18 hematopoietic progenitor and terminal cell populations (Figs. 1a and 5c, Supplementary Figs. 15 and 16, and Supplementary Table 4)^{34,35}. To compare g-chromVAR's performance to that of other state-of-the-art enrichment tools, we leveraged knowledge of the hematopoietic system and devised a lineage specificity test (Supplementary Note), which is a nonparametric rank-sum test that compares the relative ranking of lineage-specific and non-lineage-specific enrichments

for each of the compared methodologies. We found that g-chromVAR was the most specific of all the tested methods while still retaining sufficient power to identify 22 trait-cell type associations (Fig. 5d and Supplementary Figs. 13a and 16).

Having validated our approach, we investigated cell type enrichments for each of the 16 traits. We found that the most lineage-restricted or terminal populations were typically most strongly enriched for a corresponding trait association (Fig. 5e-h). For example, RBC count was most strongly enriched in erythroid precursors (Fig. 5e), and lymphocyte count was most

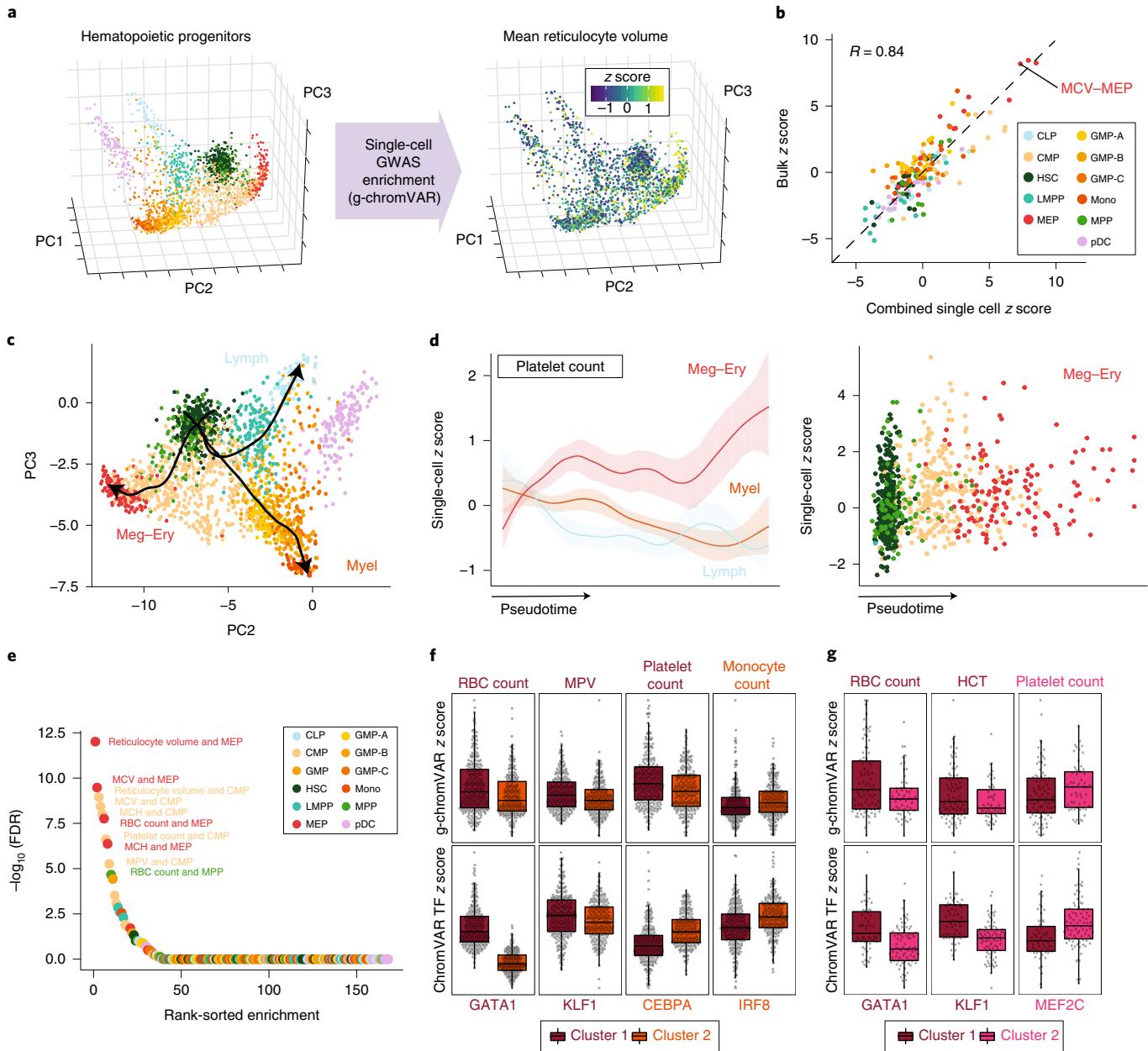


Fig. 6 | Application of g-chromVAR to single-cell chromatin accessibility data. **a**, 2,034 hematopoietic cells projected onto a 3D principal-component embedding. Single cells colored by g-chromVAR enrichment score for mean reticulocyte volume show specific regulatory enrichment in the MEP population. **b**, Validation of g-chromVAR enrichments using synthetic bulk populations obtained from sums of single cells ($n=2,034$ cells). Aggregated single-cell g-chromVAR z scores across all trait-cell type pairs (individual points) strongly correlate (Pearson's $R=0.84$) with bulk population z scores. **c**, Inferred pseudotime trajectories of three hematopoietic lineages from scATAC-seq data. **d**, Pseudotime trends (mean and 95% confidence interval) of g-chromVAR scores for platelet count across all single cells ($n=2,034$ cells) corroborate the regulatory dynamics of megakaryocyte–erythroid differentiation. **e**, Rank-order plot highlighting the trait-cell type pairs with the greatest variance over a χ^2 distribution. **f**, k -medoids partitioning of ATAC-seq counts in CMP cells ($n=502$ cells) identified two subpopulations: one that was enriched for monocyte genetic variants and one that was enriched for megakaryocyte–erythroid variants (RBC count, FDR= 1.28×10^{-4} ; mean platelet volume, FDR= 2.36×10^{-4} ; platelet count, FDR= 1.40×10^{-5} ; monocyte count, FDR= 2.21×10^{-2}). ChromVAR scores for master transcription factors (TFs) for each blood cell type support biological hypotheses for the genetic enrichments (GATA1, FDR= 1.76×10^{-82} ; KLF1, FDR= 4.33×10^{-3} ; CEBPA, FDR= 2.58×10^{-16} ; IRF8, FDR= 4.65×10^{-15}). Two-tailed t tests were used for each comparison. Box plots represent the median and interquartile range; whiskers extend 1.5 \times the interquartile range from the hinges of the box plots. **g**, Similar k -medoids partitioning of MEP cells ($n=138$ cells) identified two subpopulations with differential enrichments for megakaryocyte– and erythroid-associated genetic variants (RBC count, FDR=0.155; hematocrit, FDR= 3.98×10^{-2} ; platelet count, FDR= 7.65×10^{-2}), along with consistent differences in chromVAR transcription factor deviation scores for master transcription factors of each blood cell type (GATA1, FDR= 2.18×10^{-4} ; KLF1, FDR= 4.02×10^{-6} ; MEF2C, FDR= 2.52×10^{-3}).

strongly enriched in CD4 $^{+}$ and CD8 $^{+}$ T cells (Fig. 5h). In several instances, we observed significant enrichments for traits in earlier progenitor cells within each lineage, including enrichment for platelet traits in CMPs and enrichment for monocyte

traits in a specific subpopulation of granulocyte–macrophage progenitors (GMPs) (Supplementary Fig. 13a). We sought to investigate these enrichments in progenitor cells further at the single-cell level.

GWAS trait enrichment in single-cell chromatin accessibility data. Although the strongest g-chromVAR enrichments for blood traits were in the most lineage-restricted precursors, we reasoned that investigating progenitor populations that had robust enrichment signals, such as CMPs and MEPs, could inform principles of the genetic regulation of terminal blood cell production^{51–55}. To this end, we scored 2,034 single bone marrow-derived hematopoietic stem and progenitor cells³⁴ for GWAS enrichment by using g-chromVAR (Fig. 6a). Composite single-cell and bulk cell type enrichments were highly correlated ($R=0.84$) (Fig. 6b), and enrichments along inferred pseudotime trajectories of cellular differentiation mirrored our observations from bulk data, albeit with finer granularity (Fig. 6c,d). These results suggest that g-chromVAR is able to recover known biology from sparse single-cell ATAC-seq (scATAC-seq) profiles.

To explore potential heterogeneity within each of the 11 hematopoietic progenitor populations, we estimated the variation in regulatory genetic enrichments for each trait within the populations. We found that classically defined CMP ($n=502$ cells) and MEP ($n=138$ cells) populations exhibited significant heterogeneity in g-chromVAR enrichments for both erythroid and megakaryocyte traits (Fig. 6e). We thus hypothesized that the CMP population could be subdivided into megakaryocyte–erythrocyte-primed and monocyte-primed subtypes, whereas the MEP population could be further subdivided into erythrocyte-primed and megakaryocyte-primed subtypes. To test this hypothesis, we performed unsupervised clustering on chromatin accessibility profiles for the CMP and MEP populations (Supplementary Fig. 17) and found that the (GWAS-naive) subpopulations were indeed differentially enriched for the specific GWAS traits. In agreement with these genetic enrichments, we observed differential chromatin accessibility of motifs for lineage-specific master transcription factors between the subpopulations that corresponded to the trait enrichments, such as increased chromatin accessibility for GATA1 motifs within the clusters enriched for erythroid traits (Fig. 6f,g and Supplementary Table 12). Additional studies are needed to determine whether these differences are due to distinct lineage-biased subpopulations or whether they reflect gradations along a common axis of differentiation. Regardless, our findings demonstrate that genetic variation acts heterogeneously within classically defined progenitor populations.

Discussion

Two outstanding challenges in the post-GWAS era are (i) the precise identification of causal variants within associated loci and (ii) determination of the exact mechanisms by which these variants result in the observed phenotypes. To address the first point, we used robust genetic fine-mapping to identify hundreds of putative causal variants for 16 blood cell traits, allowing for up to five causal variants in each locus. At $PP>0.10$, we identified 240 fine-mapped coding variants as well as 647 regulatory variants in accessible chromatin in at least one of 18 primary hematopoietic populations. Several compelling anecdotes, including a number of instances in which the activity of a single regulatory element is modulated by multiple functional variants, highlight the advantages of allowing for multiple causal variants when fine-mapping.

To address the second point, we compiled and derived functional annotations to nominate regulatory mechanisms and identify putative target genes. Overall, our comprehensive approach identified a high-confidence regulatory mechanism for 145 variants and an experimentally supported target gene for 79% of variants in accessible chromatin for distinct lineages. Our investigations into fine-mapped pleiotropic variants revealed that ~90% of these variants act to tune total hematopoietic production, whereas the remaining ~10% favor production of one lineage at the expense of another (switch variants). To further improve identification of causal cell

types, we developed a new enrichment method (g-chromVAR) that can discriminate between closely related cell types and applied it to directly probe the regulatory dynamics of hematopoiesis within classically defined progenitor populations in bulk and at the single-cell level. Our ‘top loci’ method is complementary to enrichment methods that investigate polygenic signals, such as S-LDSC.

Overall, our integrated approach is designed to sequentially identify causal genetic variants, their molecular mechanisms, their target genes, and the cell types in which they act. We expect that better-powered fine-mapping studies, more numerous and higher-quality bulk and single-cell epigenomic datasets, and improved computational tools will extend the inferences discussed herein. Altogether, our study represents a paradigm for the comprehensive mapping of variants to function, which can be applied broadly to gain insights into the specific mechanisms of variants associated with a range of human traits and diseases.

URLs. UCSC Genome Browser visualization hub for all bulk ATAC-seq data, <https://s3.amazonaws.com/atachematopoeisis/hub.txt>; web app to visualize putative causal variants and corresponding annotations, <http://molpath.shinyapps.io/ShinyHeme>; functional genomic annotations, https://github.com/caleblureau/singlecell_bloodtraits/tree/master/data/annotations.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41588-019-0362-6>.

Received: 28 January 2018; Accepted: 28 January 2019;

Published online: 11 March 2019

References

- Doulatov, S., Notta, F., Laurenti, E. & Dick, J. E. Hematopoiesis: a human perspective. *Cell Stem Cell* **10**, 120–136 (2012).
- Sankaran, V. G. & Orkin, S. H. Genome-wide association studies of hematologic phenotypes: a window into human hematopoiesis. *Curr. Opin. Genet. Dev.* **23**, 339–344 (2013).
- Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429 (2016).
- Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
- Farh, K. K. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
- Wellcome Trust Case Control Consortium. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
- Lango Allen, H. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
- Flister, M. J. et al. Identifying multiple causative genes at a single GWAS locus. *Genome Res.* **23**, 1996–2002 (2013).
- Galarneau, G. et al. Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat. Genet.* **42**, 1049–1051 (2010).
- Chung, C. C. et al. Fine mapping of a region of chromosome 11q13 reveals multiple independent loci associated with risk of prostate cancer. *Hum. Mol. Genet.* **20**, 2869–2878 (2011).
- Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
- Gusev, A. et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
- Thurner, M. et al. Integration of human pancreatic islet genomic data refines regulatory mechanisms at type 2 diabetes susceptibility loci. *eLife* **7**, e31977 (2018).
- Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
- Benner, C. et al. Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *Am. J. Hum. Genet.* **101**, 539–551 (2017).

16. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
17. Trynka, G. et al. Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci. *Am. J. Hum. Genet.* **97**, 139–152 (2015).
18. Giani, F. C. et al. Targeted application of human genetic variation can improve red blood cell production from stem cells. *Cell Stem Cell* **18**, 73–78 (2016).
19. Thom, C. S. et al. Trim38 degrades dynein and regulates terminal erythropoiesis. *Dev. Cell* **30**, 688–700 (2014).
20. Wakabayashi, A. et al. Insight into GATA1 transcriptional activity through interrogation of cis elements disrupted in human erythroid disorders. *Proc. Natl Acad. Sci. USA* **113**, 4434–4439 (2016).
21. Sidore, C. et al. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat. Genet.* **47**, 1272–1281 (2015).
22. Kulakovskyi, I. V. et al. HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res.* **44**, D116–D125 (2016).
23. Oki, S. et al. ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep.* **19**, e46255 (2018).
24. Arinobu, Y. et al. Reciprocal activation of GATA-1 and PU.1 marks initial specification of hematopoietic stem cells into myeloerythroid and myelolymphoid lineages. *Cell Stem Cell* **1**, 416–427 (2007).
25. Hoppe, P. S. et al. Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. *Nature* **535**, 299–302 (2016).
26. Loughran, S. J. et al. The transcription factor Erg is essential for definitive hematopoiesis and the function of adult hematopoietic stem cells. *Nat. Immunol.* **9**, 810–819 (2008).
27. Carmichael, C. L. et al. Hematopoietic overexpression of the transcription factor Erg induces lymphoid and erythro-megakaryocytic leukemia. *Proc. Natl Acad. Sci. USA* **109**, 15437–15442 (2012).
28. Kruse, E. A. et al. Dual requirement for the ETS transcription factors Fli-1 and Erg in hematopoietic stem cells and the megakaryocyte lineage. *Proc. Natl Acad. Sci. USA* **106**, 13814–13819 (2009).
29. Vo, K. K. et al. FLI1 level during megakaryopoiesis affects thrombopoiesis and platelet biology. *Blood* **129**, 3486–3494 (2017).
30. Wang, S., He, Q., Ma, D., Xue, Y. & Liu, F. Irf4 regulates the choice between lymphoid-primed progenitor and myeloid lineage fates during embryogenesis. *Dev. Cell* **34**, 621–631 (2015).
31. Elagib, K. E. et al. RUNX1 and GATA-1 coexpression and cooperation in megakaryocytic differentiation. *Blood* **101**, 4333–4341 (2003).
32. Blyth, K. et al. Runx1 promotes B-cell survival and lymphoma development. *Blood Cells Mol. Dis.* **43**, 12–19 (2009).
33. Javierre, B. M. et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369–1384 (2016).
34. Buenrostro, J. D. et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell* **173**, 1535–1548 (2018).
35. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
36. Li, P. et al. IRF8 and IRF3 cooperatively regulate rapid interferon-β induction in human blood monocytes. *Blood* **117**, 2847–2854 (2011).
37. Hohaus, S. et al. PU.1 (Spi-1) and C/EBPα regulate expression of the granulocyte-macrophage colony-stimulating factor receptor α gene. *Mol. Cell. Biol.* **15**, 5830–5845 (1995).
38. Dufner, A. et al. The ubiquitin-specific protease USP8 is critical for the development and homeostasis of T cells. *Nat. Immunol.* **16**, 950–960 (2015).
39. Reincke, M. et al. Mutations in the deubiquitinase gene USP8 cause Cushing's disease. *Nat. Genet.* **47**, 31–38 (2015).
40. Burley, K., Westbury, S. K. & Mumford, A. D. *TUBB1* variants and human platelet traits. *Platelet* **29**, 209–211 (2018).
41. Sankaran, V. G. et al. Cyclin D3 coordinates the cell cycle during differentiation to regulate erythrocyte size and number. *Genes Dev.* **26**, 2075–2087 (2012).
42. Gieger, C. et al. New gene functions in megakaryopoiesis and platelet formation. *Nature* **480**, 201–208 (2011).
43. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
44. Giladi, A. et al. Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nat. Cell Biol.* **20**, 836–846 (2018).
45. Guo, M. H. et al. Comprehensive population-based genome sequencing provides insight into hematopoietic regulatory mechanisms. *Proc. Natl Acad. Sci. USA* **114**, E327–E336 (2017).
46. Zhang, D.-E. et al. Absence of granulocyte colony-stimulating factor signaling and neutrophil development in CCAAT enhancer binding protein α-deficient mice. *Proc. Natl Acad. Sci. USA* **94**, 569 (1997).
47. Edling, C. E. & Hallberg, B. c-Kit: a hematopoietic cell essential receptor tyrosine kinase. *Int. J. Biochem. Cell Biol.* **39**, 1995–1998 (2007).
48. Opferman, J. T. & Kothari, A. Anti-apoptotic BCL-2 family members in development. *Cell Death Differ.* **25**, 37 (2017).
49. Paul, S. P., Taylor, L. S., Stansbury, E. K. & McVicar, D. W. Myeloid specific human CD33 is an inhibitory receptor with differential ITIM function in recruiting the phosphatases SHP-1 and SHP-2. *Blood* **96**, 483 (2000).
50. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
51. Drissen, R. et al. Distinct myeloid progenitor-differentiation pathways identified through single-cell RNA sequencing. *Nat. Immunol.* **17**, 666–676 (2016).
52. Lee, J. et al. Lineage specification of human dendritic cells is marked by IRF8 expression in hematopoietic stem cells and multipotent progenitors. *Nat. Immunol.* **18**, 877–888 (2017).
53. Notta, F. et al. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351**, aab2116 (2016).
54. Paul, F. et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).
55. Khajuria, R. K. et al. Ribosome levels selectively regulate translation and lineage commitment in human hematopoiesis. *Cell* **173**, 90–103.e19 (2018).

Acknowledgements

We thank members of the Sankaran, Buenrostro, and Finucane laboratories for their helpful discussions. This work was supported by National Institutes of Health (NIH) grants R01 DK103794 and R33 HL120791 (to V.G.S.), by the New York Stem Cell Foundation (NYSCF; to V.G.S.), and by the Harvard Society and Broad Institute Fellows programs (to J.D.B.). J.C.U. is supported by an NIH training grant (5T32 GM007226-43). C.A.L. is supported by an NIH predoctoral fellowship (F31 CA232670). E.L.B. is supported by the Howard Hughes Medical Institute Medical Research Fellows Program. V.G.S. is supported as an NYSCF-Robertson Investigator. This research was conducted by using the UK Biobank resource under projects 11898 and 31063.

Author contributions

J.C.U., C.A.L., E.L.B., M.J.A., J.D.B., and V.G.S. designed the study. J.C.U., C.A.L., E.L.B., and M.H.G. analyzed data. L.S.L. performed experiments. C.B., A.T.S., V.K.K., R.M.S., and J.N.H. contributed ideas and insights. H.K.E., M.J.A., J.D.B., and V.G.S. supervised this work. J.D.B. and V.G.S. obtained funding. J.C.U., C.A.L., E.L.B., and V.G.S. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41588-019-0362-6>.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.D.B. or V.G.S.

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

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019