Variant-to-function dissection of rare non-coding GWAS loci with high impact on blood traits

Authors: Manuel Tardaguila[1,2,*], Dominique Von Schiller[1], Michela Colombo[2], Ilaria Gori[2], Eve L. Coomber[1], Thomas Vanderstichele[1], Paola Benaglio[2], Chiara Chiereghin[2], Sebastian Gerety[1], Dragana Vuckovic[3], Arianna Landini[2], Giuditta Clerici[2], Patrick Albers[1], Helen Ray-Jones[4], Katie L. Burnham[1], Alex Tokolyi[1], Elodie Persyn[5,6,14], Mikhail Spivakov[4], Vijay G. Sankaran[7], Klaudia Walter[1], Kousik Kundu[1,13], Nicola Pirastu[2], Michael Inouye[5,6,8,9,10,14], Dirk S. Paul[5,11,14], Emma E. Davenport[1], Pelin Sahlén[12], Stephen Watt[1], Nicole Soranzo[1,2,13,14,15,*]

Abstract

Two decades of Genome Wide Association Studies (GWAS) have yielded hundreds of thousands of robust genetic associations to human complex traits and diseases. Nevertheless, the dissection of the functional consequences of variants lags behind, especially for non-coding variants (RNVs). Here we have characterised a set of rare, non-coding variants with large effects on haematological traits by integrating (i) a massively parallel reporter assay with (ii) a CRISPR/Cas9 screen and (iii) *in vivo* gene expression and transcript relative abundance analysis of whole blood and immune cells. After extensive manual curation we identify 22 RNVs with robust mechanistic hypotheses and perform an in-depth characterization of one of them, demonstrating its impact on megakaryopoiesis through regulation of the *CUX1* transcriptional cascade. With this work we advance the understanding of the translational value of GWAS findings for variants implicated in blood and immunity.

Introduction

The contribution of genetic variants to the heritability of complex human traits and diseases has been well established by GWAS, ultimately reflected by the existence of the GWAS

[1] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK.
[2] Human Technopole, Fondazione Human Technopole, Milan, Italy.
[3] Imperial College London, School of Public Health, Faculty of Medicine, London, UK
[4] MRC Laboratory of Medical Sciences, Faculty of Medicine, Imperial College, London, UK
[5] Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge UK
[6] Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
[7] Division of Hematology/Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA
[8] British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK
[9] Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK
[10] Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, 75 Commercial Rd, Melbourne 3004, Victoria, Australia
[11] Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK
[12] Royal Institute of Technology - KTH, School of Chemistry, Biotechnology and Health, Science for Life Laboratory, Stockholm, Sweden
[13] Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK.
[14] British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge UK
[15] National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Genomics, University of Cambridge, Cambridge, UK
* Corresponding authors

Catalog, comprising 400,000 curated Single Nucleotide Polymorphisms (SNPs) to trait associations across more than 5,000 human traits[1]. Nevertheless, there exist few loci where a mechanistic dissection of the mode of action of the SNPs has been completed[2]. Variant-to-function (V2F) studies are challenging because of limitations in scaling up functional validation techniques, in prioritising variants within extended blocks of variants in linkage disequilibrium (LD) and in linking unequivocally genetic variants to their candidate effector genes. Initial efforts have focused on assessing the function of rare coding variants discovered from population-scale whole exome sequencing initiatives[2–4] due to the ready availability of candidate effector genes and fine mapping sets that include typically fewer variants in LD. More recently, associations driven by rare non-coding variants (RNVs) have been discovered through use of denser imputation panels and whole genome sequencing (WGS)[5]. Compared to coding variants, that propagate their effects among all the cell types and states that express the genes they impact, RNVs (as all non-coding variants) can be exquisitely selective as they additionally require the activity of specific regulatory elements in defined cellular contexts. This selectivity promises to accelerate pharmaceutical intervention with minimal side effects, illustrating that RNVs have great potential to translate GWAS associations into actionable biology.

There are two main approaches to tackle the functional role of RNVs in GWAS for haematological parameters. 'Annotation first' strategies prioritise GWAS candidate causal variants based on high quality regulatory annotations. 'Variant first' strategies[6] rely on variant-based GWAS metrics (for instance fine-mapping posterior probabilities) to prioritise variants with a higher likelihood of causality, without prior assumption of regulatory impact. Importantly, a recent study has advised caution in the interpretation of RNVs prioritised using annotations alone, as this approach tended to prioritise variants predominantly associated through LD with nearby common variants[7].

Here we have addressed the function of a set of 94 RNVs with a massively parallel reporter assay (MPRA) for enhancer activity and a secondary CRISPR/Cas9 arrayed screening to assess the effect of the variants in the native chromatin context. We have complemented these *in vitro* approaches with analyses for differential gene expression (DE) and alternative transcript usage (ATU) modes of regulation in a large collection of samples with whole blood RNA-seq and WGS data. Importantly, we have integrated our three-pronged strategy and performed extensive manual curation of each variant to narrow down putative causal RNVs and effector genes. We identify 22 variants with compelling mechanistic evidence to regulated genes that have been robustly associated with the GWAS traits. Finally, we have carried out an in-depth functional validation of one of these variants to demonstrate that RNVs can have a sizable impact on complex traits.

Results

Variant prioritisation and MPRA in hematopoietic cell lineages

We considered 12,181 loci associated with 29 different blood indices[8]. Starting from 178,890 variants contained in 95% fine mapping credible sets, we applied sequential filters selecting: i) rare variants (minor allele frequency [MAF] ≤ 1%, leaving 5,813 variants), ii) non-coding variants, defined using the most severe consequence (MSC) in variant effect predictor (VEP, 5,248 variants) (**Figure 1A,B** and **Methods)**, iii) variants with at least one trait association fine-mapped with a posterior probability (PP) ≥ 0.9 (196 variants) and iv) variants with high effect sizes (beta < first or > third quartile of standardised trait distribution) (**Figure 1A** and **Table S2**). This left a subset of 123 RNVs matching all criteria, henceforth named 'index variants' (**Figure 1A** and **Table S2**). Index variants had higher absolute effect sizes than coding variants of comparable frequency in this study (**Figure 1C)**. Furthermore, they were comparable to reported heterozygous blood ClinVar/HGMD pathogenic variants[8] (median values 0.167 and 0.174 respectively, p value = 0.86, *Wilcoxon test*). Importantly, 95% of them remained significantly associated with at least one blood trait after conditioning on common variants, demonstrating the robustness of our prioritisation. Index variants had significantly higher orthogonal prioritisation scores; including combined annotation dependent depletion ([CADD][9]), NCBoost[10], and genomic non-coding constraint of haploinsufficient variation ([Gnocchi][11]) compared to other tiers of RNVs (**Figure 1D** and **Table S3**). They were significantly more likely than other rare variants to fall within accessible ATAC-seq peaks[12], regions predicted to carry active enhancer or promoter epigenetic marks[13] and regions with promoter capture Hi-C (PCHi-C) connections in relevant haematological cell types[14] (**Figure 1D** and **Table S3**). Finally, the genes linked to index variants via VEP MSC and PCHi-C interactions had significantly higher Capture Hi-C Omnibus Gene Score (COGS)[15,16] than genes linked to rare, poorly-fine-mapped, non-coding variants (Tier 3 variants, **Figure 1D** and **Table S3**).

Out of the 123 non-coding index variants, we could fit 94 into an MPRA library design for enhancer activity[17,18] along with 17 positive and 2 negative controls selected from previous studies[19,20]. Briefly, for each variant we synthesised and cloned five partially overlapping 148-bp tiles tagged with unique barcodes and all spanning the index variant into an MPRA reporter vector (**Methods**). The resulting MPRA library (20,340 oligos) was transfected into four cancer cell lines used as models of blood cell types: K-562 (chronic myeloid leukaemia, a model erythroid cell), CHRF-288-11 (acute megakaryoblastic leukaemia, a model for megakaryocytes, CHRF), HL-60 (acute myeloid leukaemia, a model neutrophil line) and THP-1 (acute monocytic leukaemia, a model for monocytes) (**Figure 2A**). MPRAnalyze[21] was used to estimate enhancer activity (EA, calculated as log fold change induction [Log FC]) and allele specific expression (ASE, calculated as shown previously[22], **Methods**) across each tile, index variant and cell type. Significance was set at 5% global false-discovery rate (gFDR, **Figure 2B** and **Table S4**). In control regions, we observed significant enrichment in EA and concomitant EA and ASE in the same tile for positive controls, and depletion for negative controls, indicating that our MPRA was well calibrated (**Extended Figure 1A-B**). There was moderate concordance in the directionality of the ASE for the positive controls when compared with a previous MPRA study in K-562 cells[19] (7/11) possibly due to the different

measures of ASE used in the two studies (**Extended Figure 1C**). Overall, 80 of the 94 index variants had significant evidence of EA (average of 1.5 tiles per variant, **Table S4**), demonstrating a strong enrichment of active enhancer elements in our design. Of these, 42 had parallel EA and ASE in at least one tile and cell type (MPRA positives) (**Figure 2B** and **Table S6**), illustrating allele-specific differences in enhancer function for the index variants. Among the four cell lines tested, THP-1 cells displayed the least reproducibility across replicates (average r^2 = 0.19 and 0.22 for EA and ASE respectively, **Extended Figure 1D**), and the lowest % of GFP positive cells after transfection (median values 55.5 and 6.8 for CHRF and THP1 cells, respectively, p value = 0.01, *Wilcoxon test*) as they are refractory to nucleofection[23]. Of the 42 MPRA positives, 22 were shared between two or more cell types, while 13, 6 and 1 were specific to CHRF, HL-60 and K-562 cells respectively (**Extended Figure 1E**). Variants associated with platelet traits were significantly enriched in MPRA positives in the lineage-matched CHRF cell line compared to non-platelet variants (42% vs 32%, p value < 0.05 *Log Rank test*; **Extended Figure 1F**). Further, erythroid-trait associated variants were depleted in the same cell line (26% vs 42.3%, p value < 0.05 *Log rank test*; **Extended Figure 1G**). Notably, the MPRA positive variants had significantly higher AbC scores[24] and COGS values than MPRA negative variants (p values 0.0002 and 0.0077, respectively, *Wilcoxon test*, **Table S3**) indicating that the MPRA was able to select variants that have active enhancer properties in blood cells and that are linked to causal genes for blood traits. Altogether the MPRA screening provided a strong indication of the capacity of the selected index variants to impact the activity of enhancer sequences, with a significant degree of cellular context specificity observed only for the megakaryocytic cell line CHRF.

CRISPR/Cas9 arrayed screening in the native chromatin context of the variants

The episomal setting of the MPRA lacks key features of the native chromatin context where variants exert their action, e.g. nucleosome bound DNA. It has been reported that 'context-initiator TF' (pioneering TF) sequence motifs have less cooperativity than 'context-only' (non-pioneering TF) motifs in STARR-seq experiments[25]. We predicted TF motifs overlapping the MPRA variants and found 27 and 24 variants intersecting context-initiator and context-only motifs (**Table S6** and **Methods**). Interestingly, changes in chromatin accessibility predicted with Enformer[26] correlated worse with cell-matched MPRA ASE values in variants overlapping context-initiator motifs (**Extended Figure 1H**). This evidence suggested that the MPRA was suboptimal for an important fraction of the variants we had screened. Therefore, we carried out an orthogonal screen using Genome engineering-based Interrogation of Enhancers (GenIE)[27] to assay variants in their native chromatin context. GenIE leverages CRISPR/Cas9 to assess regulatory sequences on transcription and splicing through *knock-in* of GWAS alleles and indels spanning the site of the variant[27]. The method[27,28] estimates if *knock-in* alleles and unique deletion profiles (UDPs) significantly affect transcript abundance. The caveats of this approach are that it is only applicable to intronic variants in genes with high levels of expression and it depends on sufficiently high editing rates of CRISPR/Cas9 (**Figure 1B, Extended Figure 2A-C**). We assayed 9 index variants in three of the MPRA cancer cell lines (K-562, HL-60 and THP-1), and additionally a human induced pluripotent cell line in which the technique was established (hiPSC, Kolf2[27]). Six variants passed the editing rate cutoff; rs76639840|*C2CD5*, rs35592432|*FOXP1*and rs139141690|*CUX1* had significant effects for deletions spanning the variant site while

rs76639840|*C2CD5*, rs139141690|*CUX1*, rs183393610|*BID* and rs148125759|*UGCG* had significant effects for the *knock-in* allele in at least one cell type (**Figure 2D** and **S2D-E**). Importantly, 4 of the 5 GenIE positive variants overlapped context-initiator motifs and 3 of them did not have activity in the MPRA (**Table S6**). While limited in scope, the CRISPR/Cas9 assays provided support for regulatory activity in 5/6 of the variants and supported the value of orthogonal functional validation approaches by showing activity in three of the variants overlapping TF context-initiator missed by the MPRA.

Population survey of RNA expression in healthy volunteers supports variant effects mediated by transcription

To assess regulatory impact of the index variants in a native chromatin context we used two bulk RNA-seq datasets (**Table S5**). First, the INTERVAL RNA-seq study[29], that includes gene and transcript quantification from whole blood samples of 2,971 samples with matched WGS data (15x). Overall, we identified heterozygous carriers for 88 of the 94 MPRA screened variants (median of 40 carriers per variant). Since the traditional eQTL approach[30] is under-powered for variants of low allele counts, we used better calibrated analyses based on differential gene expression (DE, including an array of experimental covariates, see **Methods**) to test for differences in expression levels between *wildtype* and heterozygous carriers of rare alleles at index variants. For each variant, we tested a median of 12 genes by combining all expressed genes within the GWAS association blocks (median size of 0.5 Mb) and those connected with index variants through PCHi-C interactions[14]. Additionally, we explored regulatory mechanisms beyond gene expression control by testing for alternative transcript usage (ATU), i.e. changes in the relative abundance within the sets of transcripts expressed for each gene[31]. To model ATU we used an additive log ratio approach that accounts for its compositional nature incorporating all the experimental covariates used in the DE analysis (**Methods**).

We detected evidence for DE or ATU at 42 of the 88 variants, involving 60 genes (**Figure 3A, and Table S5**). Out of these 42 variants, 23 (involving 29 genes) had exclusively DE effects (8 within the gene body of the regulated gene), 11 (involving 11 genes) had exclusively ATU effects (two inside the gene body of the regulated gene) and 8 showed both types of regulation (involving 14 genes with DE, 4 genes with ATU and 5 genes with simultaneous DE and ATU, **Table S5**). Eight of the 20 ATU genes harbour splicing QTLs (sQTL) signals in whole blood in the GTEx Project[32], but implicating common variants (MAF > 1%) that were conditionally independent from the ones assayed here (r^2>0.7, EUR population, window size 0.5 Mb, **Methods**). In the INTERVAL QTL repository[29] we found that three of the ATU variants were sQTLs in the correspondent genes (5 genes) and for an additional 11 ATU genes we found independent common sQTLs. This suggests complex modes of regulation at the transcript level for ATU genes (**Extended Figure 3A)**. One of the ATUs involved the synonymous cryptic splicing variant rs150813342 in the *GFI1B* gene, a known transcriptional repressor and key regulator of platelet and red blood cell development (**Extended Figure 3B**). Editing the same variant using CRISPR/Cas9 in K-562 cells induced similar transcript usage changes[33].

To explore the contribution of individual cell types, we also analysed DE and ATU in three

separate immune cell isolates (monocytes, neutrophils and naive CD4+ T-cells) from the BLUEPRINT human variation panel[13]. Overall, 25 of the 88 variants investigated in INTERVAL had at least one heterozygous carrier in the 196 BLUEPRINT donors (median number of 5 carriers per variant). We detected evidence for DE or ATU at 10 of the 25 variants, involving 11 genes (**Table S5)**. Out of these, 4 variants (involving 4 genes) had exclusively DE (one being inside the gene body of the regulated gene), 3 (involving 3 genes) had exclusively ATU (all of them being intronic in the regulated gene) and 3 showed both types of regulation (involving one gene with DE, one gene with ATU and two genes with simultaneous DE and ATU). Of the ten BLUEPRINT DE/ATU variants five were shared with the INTERVAL whole blood analysis implicating the same gene in three cases (rs147176673-*PTPN7*, rs112401631-*CCR7* and rs150649461-*GFI1*). In four of the remaining five variants, the regulation was observed exclusively in one BLUEPRINT cell type: naive CD4+ T cells for rs147453535-*PRMT5-AS1* and rs115820925-*EBP41*, monocytes for rs116577908-*PLCL2* and neutrophils for rs148125759-*C9orf84*. In the case of rs35592432 we detected ATU regulation of the *FOXP1* gene both in monocytes and naive T-CD4 cells but concomitant DE regulation of the lncRNA gene *FOXP1-IT1* was only present in the latter. Overall, the RNA-seq supported multiple MPRA positives with some differences that we explore further.

Regulatory landscapes defined by *in vivo* and *in vitro* studies

We combined evidence from the MPRA and RNAseq (DE/ATU) experiments to assess how the multiple lines of evidence converge towards a mechanistic interpretation of each association (**Figure 3B** and **Table S6**), categorising the tested variants into four groups: 19 *double positives (*variants with evidence of regulatory impact from the MPRA and RNAseq experiments), 21 MPRA+/RNA- variants, 28 MPRA-/RNA+ and 20 MPRA-/RNA- (*double negative*) variants.

The 19 *double positive* variants yielded notable biological candidates, as discussed later (see also **Table S6**). We observe here that there was high concordance in the direction of the effect allele between MPRA ASE and DE for 12/13 variants (defined as at least one MPRA ASE tile with the same directionality as the one observed *in vivo* for the regulated gene, **Methods**). At the 21 *MPRA +/ RNA-* variants (**Figure 3B**), we hypothesised that the observed lack of RNA effect might be driven by insufficient statistical power of the DE/ATU analysis at the lower end of the frequency spectrum. Supporting this, MPRA+/RNA- loci had significantly lower values of minimum allele frequency (MAF) than *MPRA-/RNA+* (p value =0.002, *Wilcoxon test*, **Table S3**), with fewer heterozygote carriers compared to double positive loci (median number of heterozygous carriers = 40 vs = 17.5), however this difference was not significant (p value = 0.065, *Wilcoxon test*). Focusing next on the 28 *MPRA-/RNA+ variants* (**Figure 3B**), we first asked whether this subset of variants might harbour predominantly transcript usage differences for which we did not expect differences in enhancer activity captured by the MPRA. This was not the case, as ATU events were equally distributed among MPRA classes (p value =0.871, *Chi square test,* **Figure 3C** and **Table S6**). We then explored if *MPRA-/RNA+* variants were enriched in context-initiator motifs, given our previous observation that the MPRA was particularly unsuited for variants overlapping this type of motifs. Indeed, we found that *MPRA-/RNA+* variants were significantly enriched in context-initiator TFs motifs (p value =0.014, *Chi square test,* **Figure**

3C and **Table S6**), and we newly observed a gene connection for two of the *MPRA-/RNA+* variants overlapping context-initiator motifs and GenIE positive (rs139141690-*CUX1* and rs148125759-*C9orf84*). Finally, we observed 20 double negative variants with no detectable effect at either MPRA or RNAseq, one of which being GenIE positive, rs183393610.

To investigate what factors underpinned the four categories we used the variant and gene-based metrics applied earlier (**Table S3**). The *MPRA+/RNA-* category had significantly higher absolute effect sizes than all the rest (**Figure 3D**) illustrating that this subgroup captures very impactful variants. Importantly, the *double positives* had significantly higher AbC scores and COGs values than the double negatives (**Figure 3D**), confirming our previous observations for MPRA positives and adding a regulated gene in whole blood and/or immune cells. Overall, this effort highlights the usefulness of orthogonal approaches for assessing functional impact of genetic variants.

Manual curation of candidate effector genes for in-depth study

We first used the GeneBass[34] and OpenTargets[35] databases, as well as a comprehensive literature search, to annotate candidate effector genes underpinning the genetic associations (**Extended Figure 4** and **Table S6**). We used all the DE/ATU genes and for RNA negative variants, we considered all genes introduced in the DE/ATU analysis. Overall, we annotated genes for 76/88 variants. Then we used INTERVAL phased WGS (not available at the time of the MPRA assay design) to identify cases where the association could be explained by a nearby coding variant in the candidate gene in tight LD with the index variant (*coding proxy*). We annotated 14 variants with at least one proxy coding variant (r^2 range 0.26-0.97); however we see evidence for regulatory activity at 7/14 of them, suggesting that either of the two nominated variants could be causal (**Extended Figure 5D)**. Furthermore, by using the haplotype phase from the same WGS data, we also identified four instances where other non-coding proxy variants (r^2 range 0.51-0.85) might drive the DE/ATU expression phenotype (*regulatory proxy*, **Extended Figure 5E**). Finally, we flagged ten variants where the regulatory effects are likely mediated in cell types or tissues different from the ones used here (*other tissue*, **Extended Figure 5F**), and seven cases where the regulated gene(s) was not a credible biological candidate for the original GWAS phenotype (*other gene*). Overall, our curation effort results in a high confidence set of 22 variants for which we can formulate robust hypotheses linking genes to the blood phenotypes impacted by the variants ('*R in candidate*' label, **Fig 3E, Table 1**). Of these, 12 implicated a DE event (**Extended Figure 5A**), 7 an ATU event (**Extended Figure 5B**) and 3 had both types of regulation (**Extended Figure 5C**). Further, 12 were MPRA positive and 2 were genie positive. Finally, five of the 22 variants had been previously described as eQTLs for the same regulated genes that we detect in our analyses by the eQTLGen[30] and/or the BLUEPRINT[13] consortia and 1 as an sQTL in the INTERVAL RNA-seq initiative[29], **Table 1**.

Local organisation of rare non-coding variants and allelic series

Having defined differential regulatory activity from MPRA and RNA assays, we next asked whether these variants exhibited matched hallmarks of chromatin organisation and lineage specificity. We first tested if the sharing of chromatin accessibility patterns matched

observed pleiotropic associations (i.e. when a variant is associated with traits implicating multiple blood cell lineages)[12]. We assigned blood phenotypes to hematopoietic lineages (**Table S1**) and then classified the 94 index variants in two groups: i) lineage restricted variants (LR), i.e. variants that are associated with blood phenotypes belonging to a single lineage (80 variants) and ii) multi-lineage variants (ML), associated with traits belonging to two or more lineages (14 variants) (**Table S6)**. In this classification we used all phenotypes associated to the variants with fine-mapping PP ≥ 0.1 to comprehensively characterise the lineages impacted by the different variants. We also defined a multi-lineage ATAC-seq statistic (multi lineage ATAC, **Methods**) ranking accessible chromatin across multiple blood lineages. Multi-lineage variants had significantly higher values of the multi-lineage ATAC-seq ranking compared to lineage restricted variants (median value 'ML' = 2.33, median value 'LR' = -0.114, p value < 0.001, *Wilcoxon test*, **Figure 4A and Table S2**). ML and LR variants also had significantly different absolute effect sizes (median value 'ML' = 4.89, median value 'LR' = 6.26, p value < 0.001, *Wilcoxon test*) and credible set sizes (median value 'ML' = -0.6, median value 'LR' = -0.7, p value < 0.001, *Wilcoxon test*, **Figure 4A and Table S2**). Overall these results indicated that lineage restricted variants overlap regions that are only accessible in a smaller subset of cell types, belonging to the same haematopoietic lineage (**Figure 4B and D**) whereas pleiotropic variants sit in accessible chromatin regions shared by different lineages (**Figure 4C, E, F and 4B**). We also investigated the extent to which RNVs were part of allelic series. Allelic series identify sets of independent genetic signals acting through a common effector gene leading to putative loss and gain of function effects that can be read as gradual phenotypic changes in the GWAS traits thus defining dose-response relationships and their mechanism of action[36]. Starting from our set of 22 curated V2F examples we found 9 examples, which are discussed in detail in **Table 1** and **Figure 5A-D**. Altogether these allelic series are important for their ability to mimic dose-response curves, illustrating the power of human genetics to reveal "experiments of nature" that are key in the target validation stages of drug discovery[36].

Cascading effects of variants regulating the expression of TFs: cis variation in *CUX1* regulates megakaryocyte size and maturation

Finally, we investigated the potential for RNVs impacting master regulatory genes such as transcription factors (TFs) to perturb the expression of multiple target effector genes[30]. We first conducted a genome-wide DE analysis comparing heterozygous and wild type carriers at each of the 22 curated examples, followed by gene set overrepresentation analysis (ORA) and enrichment analysis (GSEA). TF variants were more likely to have overrepresentation of blood and immunological gene sets (**Methods**) among their genome-wide DE genes (significant ORA in 6/7 TF variants versus 3/15 non-TF variants, p value 0.0014, Chi square test). This confirms that disruption of the expression of master regulatory genes by RNVs has a cascading effect with major changes in the transcriptome. We tested the genome wide DE genes of TF variants for overrepresentation of their target genes using the Dorothea[37] (**Methods**) but found none, which could be due to the dilution of the TF signature within the whole blood bulk RNA-seq data.

We carried out an in depth validation of one of these variant-TF pairs, rs139141690-*CUX1*, where the GWAS phenotypes associated with the variant (e.g. mean platelet volume, MPV)

are compatible with findings observed in murine knock-downs of the gene (myelodysplastic syndrome, anaemia, giant platelets and myeloproliferative neoplasms[38]), and the gene is key for hematopoietic stem cells (HSCs) maintenance. Our experimental evidence implicating this gene is summarised in **Extended Figure 6A-C**.

We used CRISPR/Cas9 to create isogenic K-562 lines: i) carrying the reference allele ('G/G' clones), ii) homozygous alternative for the rs139141690-A allele ('A/A' clones) and iii) a specific 80 bp deletion spanning the SNP designed to cover all the significantly active UDPs from the GenIE experiments ('80 bp Del' clones) (**Extended Figure 6D**). We differentiated K-562 cells to megakaryocyte-like CD41+ (*ITGA2B* gene, megakaryocyte surface marker) cells using PMA[33] (**Extended Figure 6D-E**). We monitored the changes in the cell surface abundance of CD235 (*GYPA* gene, erythrocyte surface marker) and CD41 by flow cytometry and explored to what extent the 'A/A' and 80-bp del genotypes impacted the differentiation process and detected a significant effect of the genotypes (**Extended Figure 6E** and **Methods**). The A/A clones initially accumulated more of the intermediate double positive stage ($CD235^{+}CD41^{+}$) cells but ended up having less terminally differentiated single positive CD41 cells ($CD235^{-}CD41^{+}$ cells) than the wild type clones. The 80 bp del clones reached high percentages of single positive CD41 cells more rapidly than the wild type clones. Next, we undertook a K-562 differentiation experiment capturing the RNA-seq and ATAC-seq modalities at the single cell level (**Figure 6A**). In this experiment we included two additional clone lines: one heterozygous for the SNP ('A/G') and one carrying a shorter deletion that spanned the PU.1 binding site impacted by the variant ('16 bp Del', **Extended Figure 6B**). We used a lentivirus library to barcode each clone line to perform genotype deconvolution and recovered 11,250 cells separated into 13 different clusters using both the RNA-seq and ATAC-seq modalities (**Figure 6A, B** and **Methods).**

To characterise the differentiation model at the single cell level, we selected only the 'G/G' clone lines to avoid the confounding effects of the different genotypes. The relative abundance of specific clusters changed with time; clusters 1, 3 and 6 increased in the later time points of differentiation whereas cluster 5 behaved oppositely (**Figure 6B** and **Extended Figure 6F**). Importantly, using marker genes specific of different stages in the differentiation, (*IL1B*, *IL33* and *CCR7 for* basal K-562, the megakaryocyte markers *FYB1* and *GP6*, the haemoglobin genes *HBA2*, *HBZ* and *HBQ1* and *KIF15*, *STIL* and *ANAPC7* a set of genes involved in the late stage process of polyploidization of megakaryocytes[39]) we detected heterogeneity within the broad grouping of clusters based solely on the expression of the flow cytometry markers (*ITGA2B* and *GYPA*) (**Figure 6C).** Therefore we focused on two clusters: number 3, composed of double positive cells that express high levels of megakaryocyte markers and haemoglobin genes and number 1, a single positive for *ITGA2B* expression, that co-expressed high levels of the polyploidization genes. We then analysed the impact of the four mutations relative to 'G/G' cells in gene expression (differential expression, DE) and in chromatin accessibility of the ATAC peaks (differential accessibility, DA) in pseudobulk, accounting for the effect of time and the biological variability between clone lines of the same genotype (**Methods** and **Table S8**). Given the large number of peaks (n > 300K) we restricted our DA analysis to peaks that showed significant co-accessibility with DE genes (*n =1,862* unique DE genes and *n=14,851* linked peaks of which 485 were DA).

In the heterozygous clone we observed a few DE genes and DA peaks connected to the AKT/mTOR signalling in cluster 1 (**Figure 6D)**. In contrast, the three 'A/A' clone lines showed multiple DE genes belonging to the AKT/mTOR signalling pathway (clusters 3 and 1), the "polyploidization gatekeeper" gene *EZH2*[40] and its target *XRCC2* (clusters 3 and 1) and an upregulation of *FYB1*, which is included in the human phenome ontology (HPO)[41] signature for abnormal platelet volume. Crucially, only A/A clone lines showed a significant downregulation of the *CUX1* gene in clones 3 and 1 alongside a decrease in chromatin accessibility in the peak that harbours the variant and an additional linked peak (**Figure 6D-E** and **Table S8**). Moreover, *CUX1* target genes were overrepresented in the DE genes (p value = 0.033, A/A vs G/G genotypes, *ORA analysis*) (**Figure 6D)**. In addition, the analysis of chromatin accessibility produced significant DA peaks for the genes *EZH2*, *XRCC2* and *GUSB* (**Figure 6D** and **Table S8**). Finally the two deletions showed complementary results; the long deletion had multiple DE genes involved in megakaryocyte fate (e.g. *TBXAS1*, *TBXA2R*) and platelet volume (e.g. *ITGA2B*, *TUBB1*) while the short deletion had a subset of them as DE genes with concordant directionality (**Figure 6D**). In the 80bp del clone lines we detected a significant decrease in accessibility in the *CUX1* linked peak that harbours the variant but only in cluster 1 (**Figure 6D**). For both deletions we observed significant overrepresentation of *CUX1* target genes in the absence of significant changes in the expression of *CUX1* (p value = 0.002 and p value = 8.95×10^{-5}, Del16 vs G/G and Del80 vs G/G genotypes respectively, ORA analysis) (**Figure 6D)**. Finally, to connect the cellular phenotypes observed at the single cell level with the GWAS traits associated with the variant we used as proxies the flow cytometry parameters FSC-A and SSC-A. We observed that 80-bp deletion and the 'A/A' clone lines gave rise to significantly bigger CD41 single positive cells after 72 hours of treatment when compared to G/G cells (in both cases *Wilcoxon test*, p value < 0.0001) (**Figure 6F**) but the increase was more modest for 'A/A' clone lines. Altogether these results recapitulate some of the phenotypes observed at the population level in the GWAS (MPV) albeit with putatively stronger perturbations.

Discussion

A previous comprehensive GWAS on blood indices by our group discovered 16,900 conditionally independent trait-variant associations[8]. Here we have leveraged the GWAS parameters to select a group of 94 rare non-coding variants for functional follow-up. Our work builds on -and extends- previous efforts to i) identify the correct cellular context and type of assay to explain the RNVs associations, ii) address the role of chromatin accessibility in deriving lineage specificity vs pleiotropy[12], iii) characterise GWAS allelic series and iv) assess the propagation of the impact of non-coding variants affecting TFs to their target genes[6]. Importantly, 16 out of the 22 had no previous *in vivo* evidence of regulatory effect on the candidate genes. We envision that the growth of single cell datasets to the levels required in population genetics (e.g. Onek1k[42]), will accelerate the discovery of gene and transcript targets of regulatory variants.

Our in-depth validation focused on rs139141690 and *CUX1*. Importantly, this example has been recently assayed in an independent CRISPRi-QTL study, which showed a CRISPRi-dependent *CUX1* downregulation for the region but failed to reveal a significant effect of the

SNP in K-562 cells[6]. Despite its higher throughput, this approach does not provide a genotyped readout (sgRNA detection is taken as a proxy for successful editing) and is therefore susceptible to false negatives due to incomplete editing. We propose that our approach, based on pure genotyped clonal cell lines and subsequent differentiation to a cell model relevant to the variant associated-GWAS traits, is better tailored to reveal genetic effects mediated by GWAS variants. The multiome experiment provided key insights into the mode of action of the variant revealing a causal relationship between *CUX1* expression and the chromatin accessibility in the peak that intersects rs139141690. Further, it demonstrated the general dysregulation of AKT/mTOR related genes (*CUX1* inhibition activates the *PI3K/AKT* pathway in human cancers[43]). Finally, the plausible link between the concomitant downregulation of *EZH2* in cluster 3 and the accumulation of double positive $CD235^{+}CD41^{+}$ cells as *EZH2* inhibition has been demonstrated to block polyploidization and proliferation in megakaryocyte differentiation[40]. Overall these results suggest the existence of a genetic program regulating platelet volume controlled by the hub of TF motifs in which the variant lies. The overrepresentation of *CUX1* target genes independently of *CUX1* downregulation might be related to the upregulation of *RUNX1* observed in all the deletions since the two TFs share 38.8% of Dorothea target genes.

In this work we advance the functional exploration of high impact rare non-coding GWAS variants that are often overlooked due to the inherent complexities in their analysis and the intricacies of the regulatory mechanisms. This represents an opportunity cost given the unparalleled mechanistic insights into target specificity and modulation that we have showcased here and that could prove key for pharmaceutical intervention.

.

Acknowledgements

Acknowledgements: Tristram Bellerby, Mathew Mayho, Jeremy Schwartzentruber, Sarah Cooper, Andrew Bassett from the Wellcome Sanger Institute. Cecilia Dominguez-Conde, Davide Bolognini and Edoardo Giacopuzzi from the HT Population and Medical Genomics centre. Alessio Palini, Nicolò Panini, Silvia Bombelli in the HT Flow Cytometry Applications Resource unit and Clelia Peano, Eugenia Ricciardelli, Javier Cibella, Fabio Simeoni, Niccolò Alfano, Paolo Ferrari and Luigi Lamparelli in the HT Genomic Facility.

This work has been funded by BHF Cambridge Centre for Research Excellence RE/18/1/34212, the Chan Zuckerberg foundation and MIUR/MEF through Fondazione Human Technopole. M.I. and E.P. were supported by core funding from the British Heart Foundation (RG/18/13/33946: RG/F/23/110103), NIHR Cambridge Biomedical Research Centre (NIHR203312) [*], BHF Chair Award (CH/12/2/29428), and by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and the Wellcome Trust. M.I. was also supported by the UK Economic and Social Research 878 Council (ES/T013192/1). *The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

RNA-seq in the INTERVAL study was funded as part of an alliance between the University of Cambridge and the AstraZeneca Centre for Genomics Research, and by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). Participants in the INTERVAL randomised controlled trial were recruited with the active collaboration of NHS Blood and Transplant England (www.nhsbt.nhs.uk), which has supported field work and other elements of the trial. DNA extraction and genotyping were co-funded by the National Institute for Health and Care Research (NIHR), the NIHR BioResource (http://bioresource.nihr.ac.uk) and the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014) [*]. The academic coordinating centre for INTERVAL was supported by core funding from the: NIHR Blood and Transplant Research Unit (BTRU) in Donor Health and Genomics (NIHR BTRU-2014-10024), NIHR BTRU in Donor Health and Behaviour (NIHR203337), UK Medical Research Council (MR/L003120/1), British Heart Foundation (SP/09/002; RG/13/13/30194; RG/18/13/33946), NIHR Cambridge BRC (BRC-1215-20014; NIHR203312) [*], and by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome. A complete list of the investigators and contributors to the INTERVAL trial is provided in Di Angelantonio et al[44]. The academic coordinating centre would like to thank blood donor centre staff and blood donors for participating in the INTERVAL trial.

*The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Conflicts of interest

M.I. is a trustee of the Public Health Genomics (PHG) Foundation, a member of the Scientific Advisory Board of Open Targets, and has research collaborations with AstraZeneca, Nightingale Health and Pfizer which are unrelated to this study. T.V. has received PhD studentship funding from AstraZeneca. K.K and D.P. are current employees and stockholders of AstraZeneca. P.A. is a current employee of Glaxosmithkline.

Tables and Figures

Rsid	chr:pos ref > alt[16]	Mechanism	Candidate gene	GWAS traits	GWS trait lineages	Allelic series
rs147176673[17]	chr1:202,129,205 G>A	DE	*PTPN7*	MSCV; PLT#; PCT; BASO%	ERY; MK	-
rs115820925	chr1:29,217,311 G>A	ATU	*EBP41*	MRV	ERY	MRV – *EBP41*
rs12733073[18]	chr1:198,680,015 G>A	DE	*PTPRC*	MONO%; LYMPH#; LYMPH%; WBC#	LYMPH	LYMPH# - *PTPRC*
rs150649461[2,3]	chr1:92,925,654 G>C	DE + ATU	*EVI5, GFI1*[5]	MONO#; MONO%; LYMPH%; WBC#	GM	-
rs140397066	chr2:144,162,105 A>G	DE	*ARHGAP15*	PDW; MONO#; NEUT%; LYMPH#; LYMPH%; WBC#	MK; GM; LYMPH	-
rs183347927	chr2:74,920,648 G>A	DE	*DOK1, WBP1*[5]	MPV	MK	-
rs114050631[3]	chr2:219,020,958 C>T	DE	*CXCR2*	MONO%; NEUT#; NEUT%; BASO#; LYMPH%; WBC#	GM	NEUT# - *CXCR2*
rs191451841	chr3:128,317,978 C>T	ATU	*GATA2*[5]	EO#; EO%	GM	-
rs112097551	chr3:128,322,617 G>A	ATU	*EEFSEC*	MCV; MCH; MSCV; MONO#; MONO%; EO#; EO%	ERY; GM	-
rs116577908	chr3:17,098,399 A>G	ATU	*PLCL2*	RBC#; MCV; MCH; RDW; MRV; MSCV	ERY	-
rs182089148	chr3:46,354,444 C>T	DE	*CCR2*	MONO#; MONO%	GM	MONO# - *CCR2*
rs35592432	chr3:71,355,240 G>C	DE + ATU	*FOXP1-IT1*[5], *FOXP1*[5]	NEUT%; LYMPH#; LYMPH%	LYMPH	-
rs187282666	chr5:1,093,511 G>A	DE	*SLC12A7*	MCHC; RDW	ERY	RDW – *SLC12A7*
rs186272630[3]	chr5:35,476,470 G>T	DE	*PRLR*	MONO#; MONO%	GM	-
rs139141690	chr7:101499930 G>A	DE	*CUX1*[5]	RET#; RET%; HLSR#; HLSR%;MPV;PDW;PCT	ERY; MK	-
rs149489081[19]	chr8:41589736 T>G	ATU	*ANK1*	MCV; MCHC; RDW; RET#; RET%; HLSR#; HLSR%; MRV; MSCV	ERY	RET# - *ANK1*
rs149678861	chr15:65,174,494 A>G	DE	*ZNF609*[20], *PLEKHO2*	PDW	MK	PDW - *ZNF609*, *PLEKHO2*
rs112401631	chr17:38,764,524 T>A	DE + ATU	*CCR7*	EO#;EO%; LYMPH#; LYMPH%	GM; LYMPH	LYMPH# - *CCR7*
rs17758695	chr18:60,920,854 C>T	ATU	*BCL2*[5]	RBC#; MCV; MCH; MRV; MSCV; PLT#; MPV; PCT; MONO#; MONO%; NEUT#; BASO#; BASO%; EO#; EO%; WBC#	ERY; MK; GM	-
rs17242843	chr19:11,210,157 C>T	DE	*EPOR, DOCK6*	RDW; MSCV	ERY	-
rs569022307	chr19:15,653,669 T>C	ATU	*RASAL3*	LYMPH#	LYMPH	LYMPH# - *RASAL3*
rs62237617	chr22:28761148 C>T	DE	*CHEK2*	PCT; LYMPH#	MK; LYMPH	-

[16] Genomic coordinates are in the GRCh37/hg19 assembly

[17] Described as eQTL for the candidate gene in the BLUEPRINT Consortium

[18] Described as eQTL for the candidate gene in the eQTLGen Consortium

[19] Described as sQTL for the candidate gene in the INTERVAL RNA-seq initiative

[20] Transcription factor or protein known to regulate gene expression

Table 1. Variants with curated mechanistic hypotheses. *For the abbreviations of blood phenotypes see Table S3. DE = Differential Expression, ATU = Alternative Transcript Usage.*

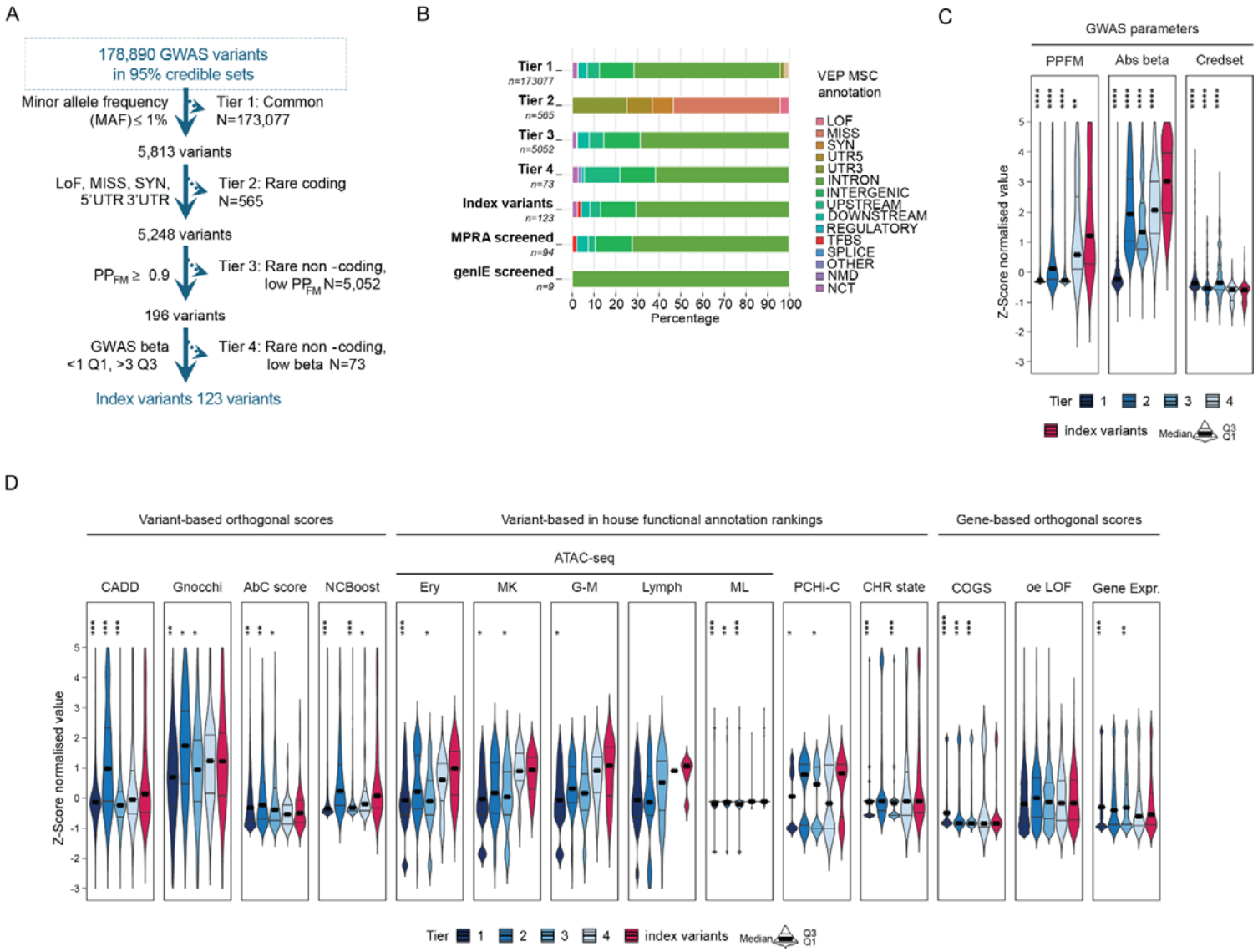

A
178,890 GWAS variants in 95% credible sets
Minor allele frequency (MAF)≤ 1%
Tier 1: Common N=173,077
5,813 variants
LoF, MISS, SYN, 5'UTR 3'UTR
Tier 2: Rare coding N=565
5,248 variants
PPFM ≥ 0.9
Tier 3: Rare non -coding, low PPFM N=5,052
196 variants
GWAS beta <1 Q1, >3 Q3
Tier 4: Rare non -coding, low beta N=73
Index variants 123 variants
B
Tier 1 n=173077
Tier 2 n=565
Tier 3 n=5052
Tier 4 n=73
Index variants n=123
MPRA screened n=94
genIE screened n=9
0 10 20 30 40 50 60 70 80 90 100
Percentage
VEP MSC annotation
LOF
MISS
SYN
UTR5
UTR3
INTRON
INTERGENIC
UPSTREAM
DOWNSTREAM
REGULATORY
TFBS
SPLICE
OTHER
NMD
NCT
C
GWAS parameters
PPFM
Abs beta
Credset
Z-Score normalised value
Tier 1 2 3 4
index variants
Median Q3 Q1
D
Variant-based orthogonal scores
Variant-based in house functional annotation rankings
Gene-based orthogonal scores
ATAC-seq
CADD
Gnocchi
AbC score
NCBoost
Ery
MK
G-M
Lymph
ML
PCHi-C
CHR state
COGS
oe LOF
Gene Expr.
Z-Score normalised value
Tier 1 2 3 4 index variants
Median Q3 Q1

Figure 1. Variant prioritisation strategy and MPRA controls and inter-replica variability. A) Workflow of the prioritisation strategy for the GWAS on blood traits. **B)** All VEP MSC per variant across the tiers of the prioritisation strategy, the variants screened in the MPRA and in the GenIE experiments. **C)** Distributions of the GWAS variant-based parameters across the prioritisation strategy tiers. **D)** Distributions of the variant-based and gene-based metrics across the different tiers of the prioritisation strategy. *Q1 = value of the first quartile, Q3 = value of the third quartile, * $p < 0.05$, ** $p < 0.001$, *** $p < 0.0001$ and **** $p < 2x10^{-16}$, Wilcoxon test. All the comparisons shown correspond to the values in the 'index variants' subset versus the rest of the tiers.*

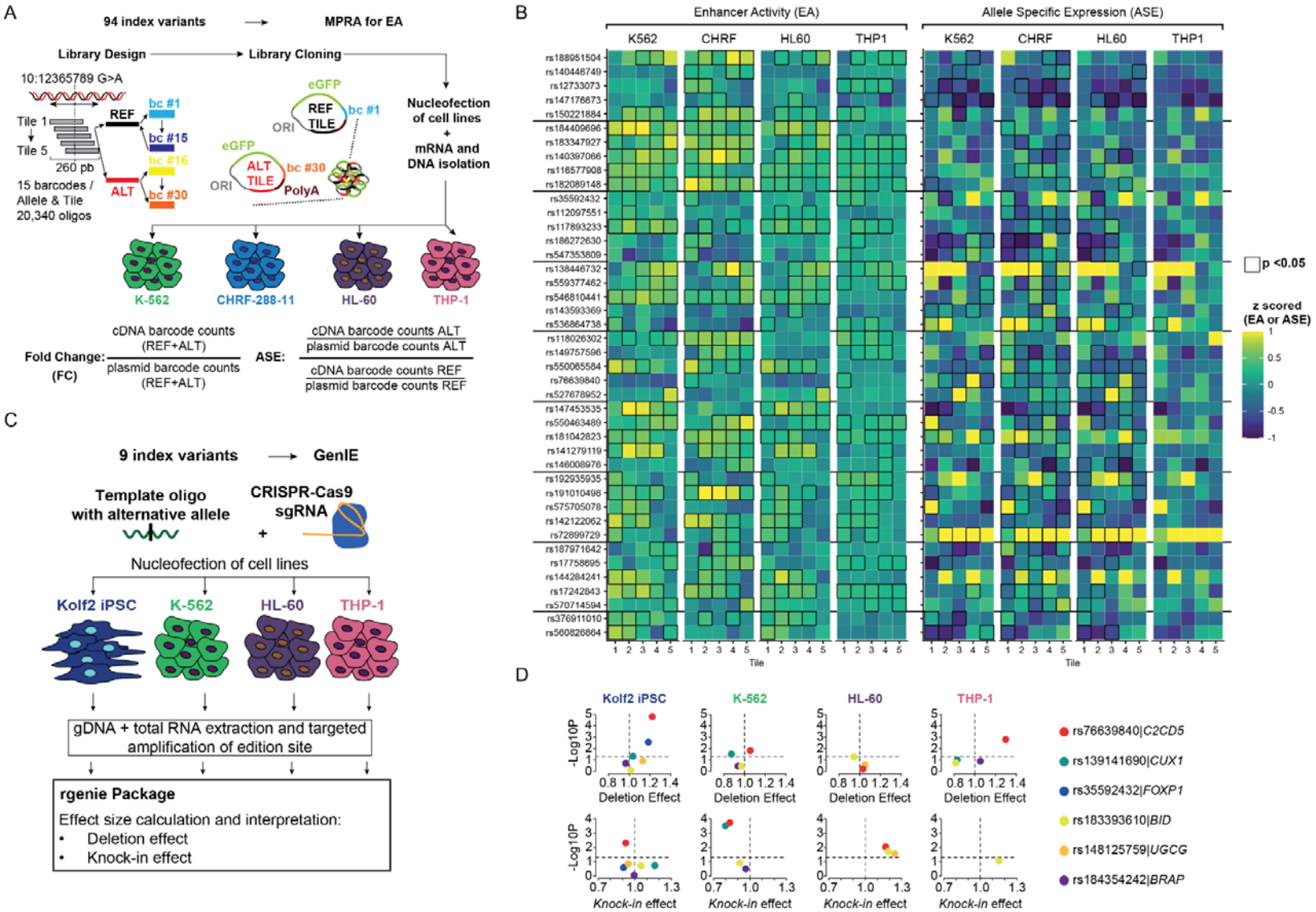

A
94 index variants
MPRA for EA
Library Design
Library Cloning
10:12365789 G>A
Tile 1
Tile 5
260 pb
15 barcodes / Allele & Tile
20,340 oligos
REF
ALT
bc #1
bc #15
bc #16
bc #30
eGFP
REF TILE
ALT TILE
ORI
PolyA
Nucleofection of cell lines
+
mRNA and DNA isolation
K-562
CHRF-288-11
HL-60
THP-1
Fold Change: (FC)
cDNA barcode counts (REF+ALT)
plasmid barcode counts (REF+ALT)
ASE:
cDNA barcode counts ALT / plasmid barcode counts ALT
cDNA barcode counts REF / plasmid barcode counts REF
B
Enhancer Activity (EA)
Allele Specific Expression (ASE)
K562
CHRF
HL60
THP1
Tile
p <0.05
z scored (EA or ASE)
C
9 index variants
GenIE
Template oligo with alternative allele
CRISPR-Cas9 sgRNA
Nucleofection of cell lines
Kolf2 iPSC
K-562
HL-60
THP-1
gDNA + total RNA extraction and targeted amplification of edition site
rgenie Package
Effect size calculation and interpretation:
• Deletion effect
• Knock-in effect
D
Kolf2 iPSC
K-562
HL-60
THP-1
-Log10P
Deletion Effect
Knock-in effect
rs76639840|C2CD5
rs139141690|CUX1
rs35592432|FOXP1
rs183393610|BID
rs148125759|UGCG
rs184354242|BRAP

Figure 2. MPRA and GenIE screenings. A) Design of the MPRA for enhancer activity. **B)** Heatmap summarising the Enhancer Activity (EA, logFC) and the Allele Specific Expression (ASE) in the MPRA screening for the 42 MPRA hits. **C)** Design of the GenIE screening. **D)** Volcano plots for deletion and *knock-in* effects per variant and cell type assayed in the GenIE screening. Significant hits are highlighted with the colour code.

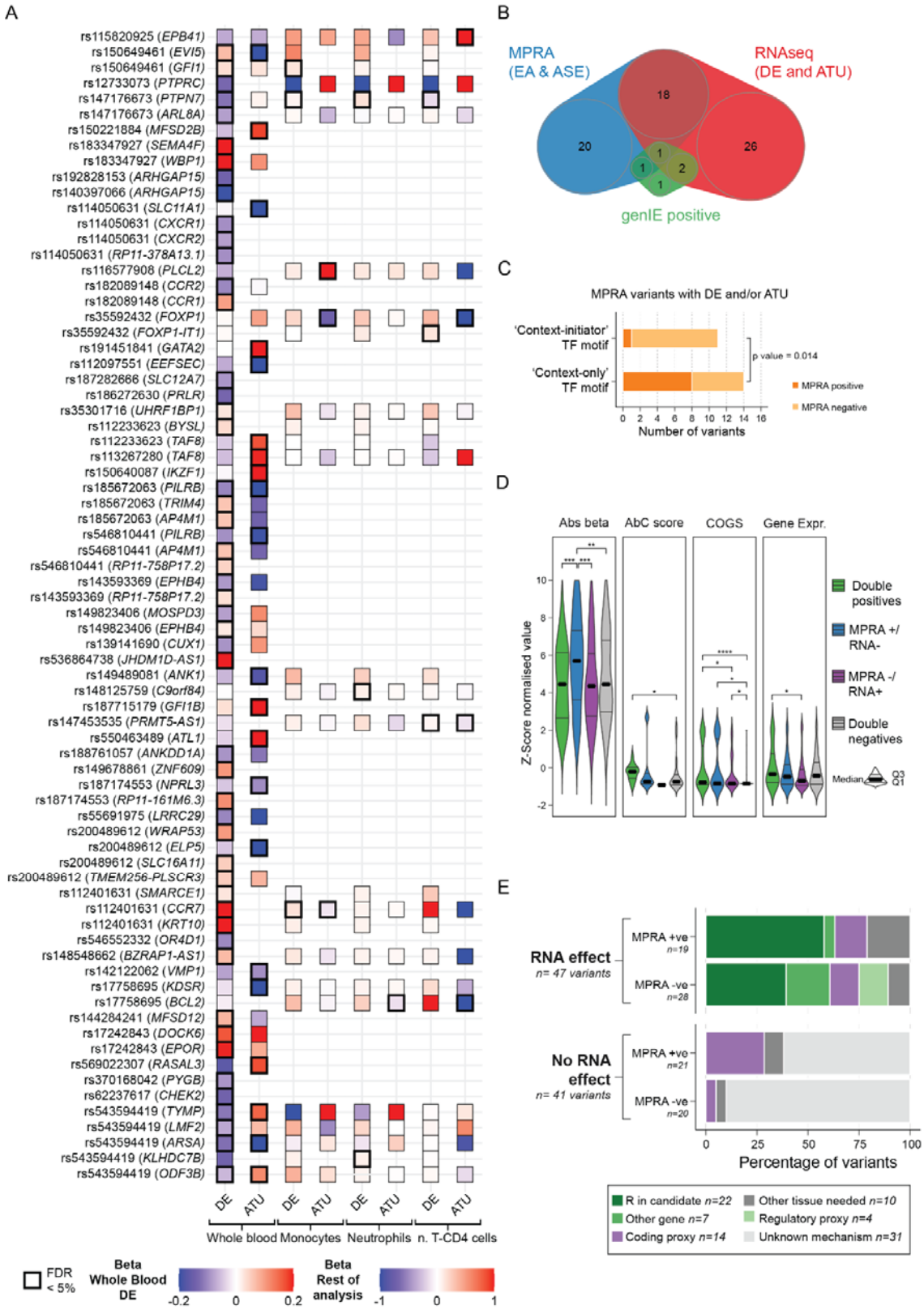


Figure 3. RNA-seq study results and curation of screened variants derived from RNA Seq, MPRA and GenIE readouts. A) Heatmap of the variant with significant DE/ATU genes.

Variants without any DE or ATU genes in either INTERVAL or BLUEPRINT were labelled 'No RNA effect (**Table S6**). **B)** Breakdown of the sharing between the results of the MPRA, the GenIE and the RNA-seq studies. **C)** Breakdown of all the DE/ATU variants between those overlapping 'context-initiator' and 'context-only' motifs and whether they were active or not in the MPRA. **D)** Distribution of the significantly different variant-based and gene-based metrics across the four groups of variants according to the MPRA and DE/ATU results. **E)** Breakdown of the mechanistic and manual curation labels by the results of the MPRA and the RNA-seq study. *DE = Differential Gene Expression, ATU = Alternative Transcript Usage, n. T-CD4 cells = naive T CD4 cells, * $p < 0.05$, ** $p < 0.001$, *** $p < 0.0001$ and **** $p < 2x10^{-16}$, Wilcoxon test.*

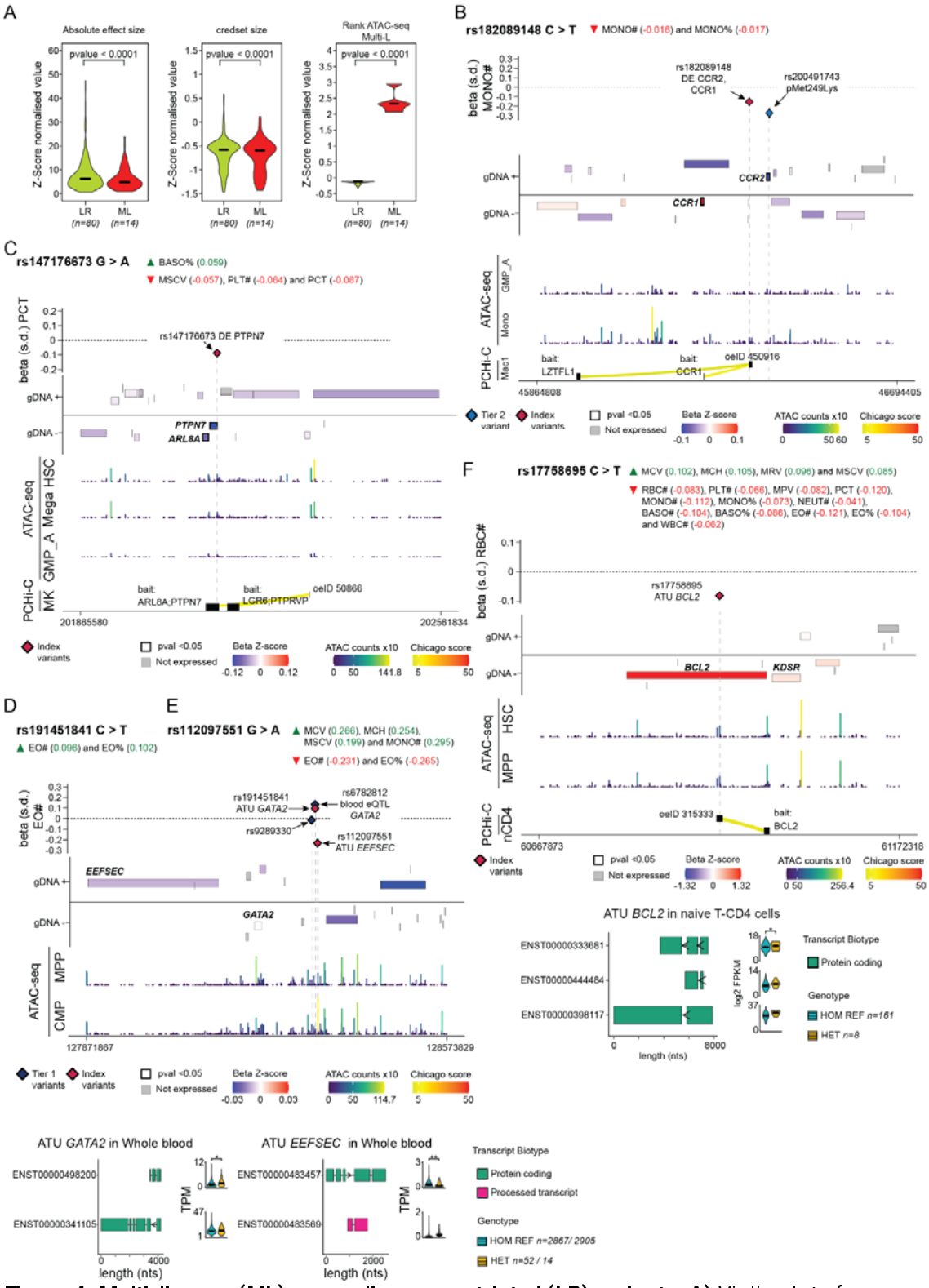


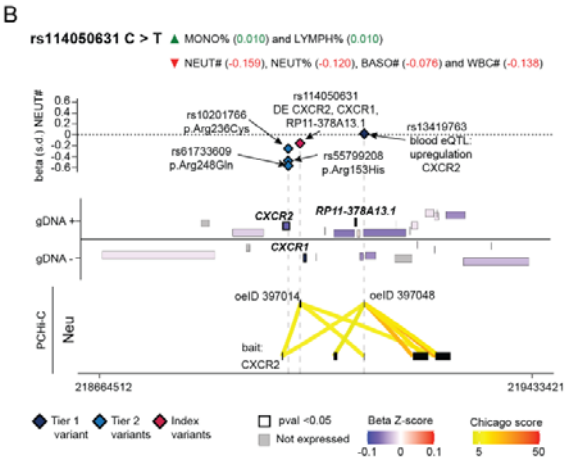
Figure 4. Multi-lineage (ML) versus lineage restricted (LR) variants. A) Violin plots for Absolute Effect Size, Credible Set Size and the Ranking for Multi-Lineage ATAC-seq separating ML and LR variants (See also **Table S3**). **B)** Locus plot for the lineage restricted variant rs182089148 with DE in the candidate gene *CCR2*. The independent variant

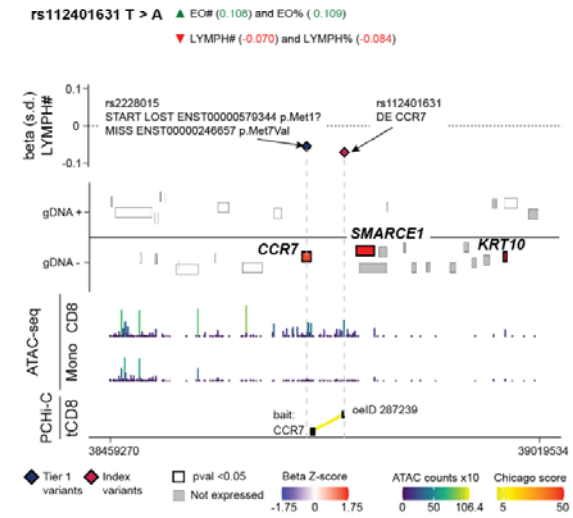
rs200491743 (Tier 2) is a coding variant affecting *CCR2*. ATAC-seq shows accessible chromatin in cell types belonging to the myeloid lineage whereas the PCHi-C track for the Mac1 cell type shows a region overlapping rs182089148 that contacts the promoters of *LZTFL1* and *CCR1* (also observed in THP-1 LPS treated cells[45]). We newly observed that rs182089148-T led to opposite effects on the expression of *CCR1* (increase) and *CCR2* (decrease) which were supported by concordant MPRA positive tiles (**Table S4**). **C)** Locus plot for the multi-lineage variant rs147176673 with DE in the candidate gene *PTPN7*. ATAC-seq shows accessible chromatin and the PCHi-C track indicates that the variant overlaps the promoters of *PTPN7* and *ARL8A* in cell types of the erythroid and megakaryocytic lineages. rs147176673-A significantly downregulated ARL8A and PTPN7 in whole blood and BLUEPRINT immune cell isolates which was supported by multiple concordant MPRA positive tiles (**Table S4**). **D)** Locus plot for the variants rs191451841 (LR) and **E)** rs112097551 (ML) that have ATU regulation in the candidate genes *GATA2* and *EEFSEC*, respectively. ATAC-seq supports the presence of accessible chromatin in progenitor cell types (MPP and CMP) only for rs112097551. rs191451841-T led to a significantly lower use of the transcript ENST00000498200 coding for a short 8 aa protein isoform and rs112097551-A led to a significant increase in the use of the processed transcript ENST00000483569. **F)** Locus plot for the ML variant rs17758695 with ATU in the candidate gene *BCL2*. rs17758695 is a very pleiotropic variant sitting in accessible chromatin in the early progenitor and stem cell stages of haematopoiesis described before[12]. ATAC-seq shows accessible chromatin in progenitor cell types and the PCHi-C track for naive CD4+ T cells indicates that the variant overlaps a region that contacts the promoters of *BCL2*. We detect the novel observation that the variant led to ATU in the *BCL2* gene in neutrophils and naive T-CD4 cells. *PCHi-C = Promoter Capture HiC. See Table S3 for all the cell type abbreviations and Table S2 for the breakdown of variant classification in tiers. * p value <= 0.05, ** p value <= 0.01, ATU additive log ratio model.*

A

Phenotype	Candidate eff. gene(s)	Coding Variants (CADD Scaled value)	Non-Coding Variants (Regulatory effect)
NEUT#	CXCR2	rs61733609 (24.9) rs55799208 (22.6) rs10201766 (15.3)	rs114050631 (DE DOWN) rs13419763 (bl. eQTL UP)
MRV	EPB41		rs10915225 (bl. eQTL DOWN) rs115820925 (ATU UP coding t.)
LYMPH#	PTPRC	rs17612648 (6.44) rs10494783 (2.2)	rs12733073 (DE DOWN) rs80099993 (T-cell eQTL DOWN)
MONO#	CCR2	rs200491743 (24.6)	rs182089148 (DE DOWN)
LYMPH#	CCR7	rs2228015 (14.8)	rs112401631 (DE UP)
RDW	SLC12A7	rs56368594 (2.98)	rs187282666 (DE DOWN) rs4975662 (bl. eQTL in SLC12A7 DOWN), rs35791340 (bl. eQTL in SLC12A7 DOWN)
PDW	ZNF609, PLEKHO2	rs143331139 (ZNF609 28.8)	rs149678861 (DE DOWN PLEKHO2, UP ZNF609) rs1719285 (bl eQTL PLEKHO2 UP)
RET#	ANK1	rs34664882 (17.9) rs142626656 (29.3) rs780729970 (15)	rs149489081 (ATU DOWN P. T.) rs113630867 (bl. eQTL UP)
LYMPH#	RASAL3	rs142945276 (25.7) rs56209154 (22.5)	rs569022307 (ATU UP RI transcripts)

Effect allele positive effect size Effect allele negative effect size





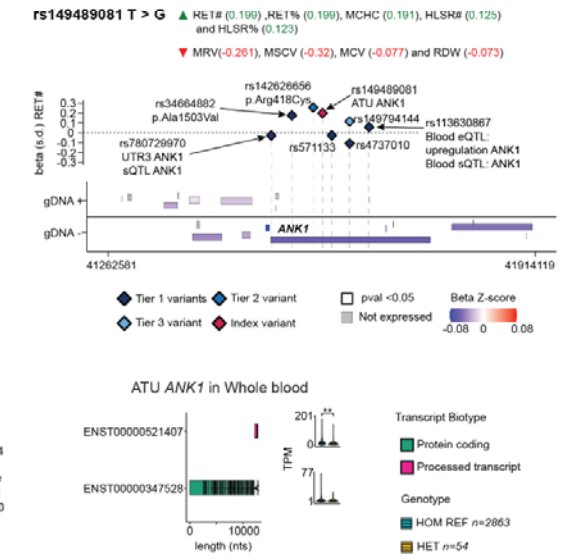


Figure 5. Allelic series (AS) derived from coding and non-coding variants acting through the same effector candidate gene. A) Summary table of the AS identified within the set of 22 'R in candidate' variants. All the conditional independent variants acting through the same candidate gene are displayed. **B)** Locus plot for the AS of variants impacting the gene *CXCR2* and the GWAS trait NEUT#. Three of the variants (rs10201766, rs55799208 and rs61733609, all Tier 2) are missenses (CADD values 15.3-24.6) affecting the *CXCR2* gene whereas a fourth independent signal (rs13419763, Tier 1) is an e-QTL in whole blood upregulating the expression of the gene. The index variant rs114050631 overlaps a PCHi-C region connected to the promoter of *CXCR2* in neutrophils and leads to the fourth biggest decrease in NEUT# in the AS. We observed a significant downregulation of three genes including *CXCR2* for the rs114050631-T allele. **C)** The AS on lymphocyte counts (LYMPH#) and *CCR7* is composed of two independent signals: rs2228015, a missense and start lost variant in *CCR7*, and the index variant rs112401631, which sits in accessible chromatin in the lymphocyte and granulocyte monocyte lineages and has a PCHi-C connection with the promoter of *CCR7* only in the lymphocyte cell lineage The index variant rs112401631 has three DE genes including a significant upregulation of *CCR7*. Both variants decrease LYMPH# in carriers of the alternative allele but only the index variant is associated with a concomitant increase in

eosinophil counts (EO#). The counter-directional effect on LYMPH# and EO# of rs112401631 and the pattern of chromatin accessibility have led others to propose this variant as a gatekeeper that switches the abundances of cell types belonging to two different lineages in hematopoiesis[12]. **D)** Locus plot for the AS of variants impacting the gene *ANK1* and the GWAS trait RET#. Three of the variants (rs780729970, rs34664882 and rs142626656, respectively Tier 1, Tier 1 and Tier 2, CADD values 15-29.3) are coding variants affecting the *ANK1* gene whereas a fourth independent signal (rs113630867, Tier 1) is an e- and s-QTL in whole blood upregulating the expression of the gene. The ATU analysis reveals that rs149489081-G entails an ATU event that results in a lower contribution of a processed transcript to the pool of transcripts of *ANK1*. We hypothesise that this change leads to a gain-of-function at the gene level resulting in an increase in RET# in keeping with the fact that the common eQTL upregulates the expression of *ANK1* and increases RET#. *PCHi-C = Promoter Capture HiC. See Table S3 for all the cell types and GWAS traits abbreviations and Table S2 for the breakdown of variant classification in tiers. ** p value <= 0.01, ATU additive log ratio model.*

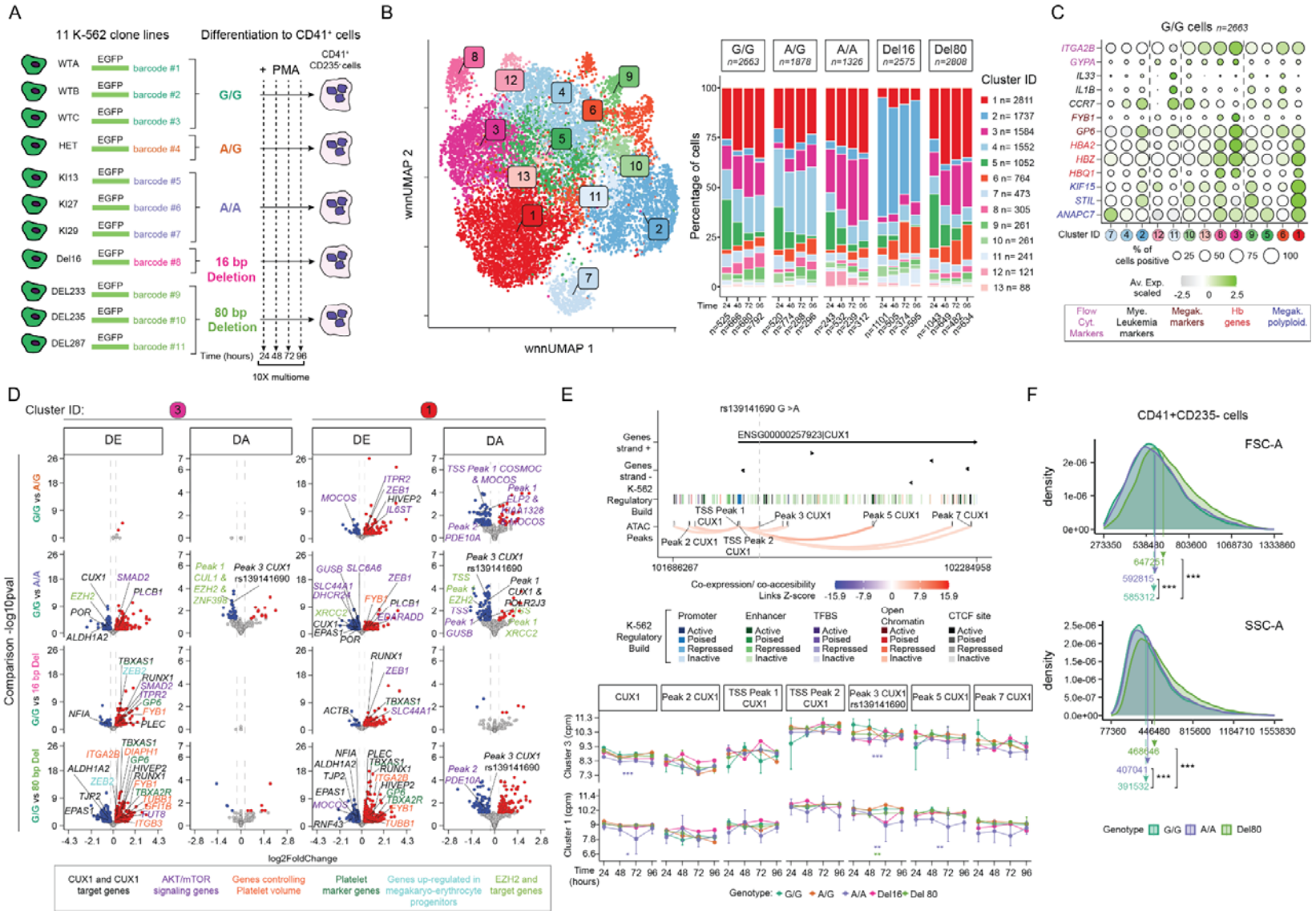

A
11 K-562 clone lines
Differentiation to CD41+ cells
WTA
WTB
WTC
HET
KI13
KI27
KI29
Del16
DEL233
DEL235
DEL287
EGFP
barcode #1
barcode #2
barcode #3
barcode #4
barcode #5
barcode #6
barcode #7
barcode #8
barcode #9
barcode #10
barcode #11
G/G
A/G
A/A
16 bp Deletion
80 bp Deletion
+ PMA
CD41+ CD235- cells
Time (hours) 24 48 72 96
10X multiome
B
wnnUMAP 1
wnnUMAP 2
G/G n=2663
A/G n=1878
A/A n=1326
Del16 n=2575
Del80 n=2808
Percentage of cells
Cluster ID
1 n= 2811
2 n= 1737
3 n= 1584
4 n= 1552
5 n= 1052
6 n= 764
7 n= 473
8 n= 305
9 n= 261
10 n= 261
11 n= 241
12 n= 121
13 n= 88
C
G/G cells n=2663
ITGA2B
GYPA
IL33
IL1B
CCR7
FYB1
GP6
HBA2
HBZ
HBQ1
KIF15
STIL
ANAPC7
Cluster ID
% of cells positive
Av. Exp. scaled
Flow Cyt. Markers
Mye. Leukemia markers
Megak. markers
Hb genes
Megak. polyploid.
D
Cluster ID:
DE
DA
Comparison -log10pval
G/G vs A/G
G/G vs A/A
G/G vs 16 bp Del
G/G vs 80 bp Del
log2FoldChange
CUX1 and CUX1 target genes
AKT/mTOR signaling genes
Genes controlling Platelet volume
Platelet marker genes
Genes up-regulated in megakaryo-erythrocyte progenitors
EZH2 and target genes
E
rs139141690 G >A
ENSG00000257923|CUX1
Genes strand +
Genes strand -
K-562 Regulatory Build
ATAC Peaks
TSS Peak 1 CUX1
Peak 3 CUX1
Peak 5 CUX1
Peak 7 CUX1
Peak 2 CUX1
TSS Peak 2 CUX1
101686267
102284958
Co-expression/ co-accesibility
Links Z-score -15.9 -7.9 0 7.9 15.9
Promoter
Enhancer
TFBS
Open Chromatin
CTCF site
Active
Poised
Repressed
Inactive
Cluster 3 (cpm)
Cluster 1 (cpm)
Time (hours)
Genotype: G/G A/G A/A Del16 Del 80
F
CD41+CD235- cells
FSC-A
SSC-A
density
Genotype G/G A/A Del80

Figure 6. Single Cell multimodal analysis confirms the effects on gene expression and chromatin accessibility elicited by rs139141690-A that underpin the cellular phenotypes affected by the variant. A) Experimental design. **B)** U-MAP combining both sc-RNAseq and sc-ATACseq modalities that display the 13 different clusters. To the right, stacked barplot illustrating the changes in their relative abundance across time points and genotypes **C)** Dotplot displaying the average scaled gene expression and the percentage of cells expressing the genes within each cluster for five sets of marker genes of different stages in the PMA differentiation of K-562 cells. The dashed lines indicate the separation of the clusters into 4 groups that we hypothesise correspond to the flow cytometry subgroups. **D)** Volcano plots corresponding to the DE and DA analysis in the clusters 3 and 1 for the comparisons of each genotype against wild type cells. Red dots and blue dots correspond to upregulated and downregulated genes in the genotype of interest compared to the 'G/G' genotype. Genes belonging to more than one of the highlighted groups are mixed coloured (e.g. *TBXAS1*). **E)**. Locus plot detailing the regulatory landscape and the DE and DA changes in clusters number 1 and 3 for the *CUX1* gene. The cpm values express either the gene expression counts for *CUX1* gene or the ATAC-seq counts for the different peaks displayed. **F)** FSC-A and SSC-A analysis across the genotypes in K-562 $CD41^{+}CD235^{-}$ cells at 72 hours. *Cyt. = cytometry, Mye. = myeloid, Megak. = megakaryocyte, Hb = Haemoglobin, polyploid = polyploidization, cpm =counts per million.See Table S8 for Peak name coordinates.*

Methods

Variant Annotation

The 178,890 variants in the 95% credible sets of the 29 blood indices from Vuckovic et al[8] were prioritised attending to MAF (1% threshold), Variant-Effect-Predictor (VEP)[46] Most Severe Consequence (MSC), Posterior Probability (PP) and effect size. The prioritised subset of index variants was below or equal 1% MAF, MSC non-coding, with at least one blood index association above or equal 0.9 PP and with an effect size for that association between the absolute minimum and first quartile (beta < Q1) or the third quartile and the absolute maximum (beta > Q3), **Figure 1A** and **Table S2**. We condensed the VEP MSC option[47] into coding and non-coding consequences. The coding group comprised the following labels: LOF (Loss Of Function, [splice_acceptor_variant, splice_donor_variant, stop_gained and frameshift_variant]), MISS (missense_variant), UTR5 (5_prime_UTR_variant), UTR3 (3_prime_UTR_variant) and SYN (synonymous_variant). The non-coding group comprised the following labels: INTRON (intron_variant), INTERGENIC (intergenic_variant), UPSTREAM (upstream_gene_variant), DOWNSTREAM (downstream_gene_variant), REGULATORY (regulatory_region_variant), TFBS (TF_binding_site_variant), SPLICE (splice_region_variant), OTHER (start_lost, stop_lost, inframe_deletion, inframe_insertion, stop_retained_variant and mature_miRNA_variant), NMD (NMD_transcript_variant) and NCT (non_coding_transcript_variant).

For the GWAS parameters the MAF across all the variants was Z-score normalised. The PP was Z-score normalised per phenotype. For the effect size first the absolute value was obtained and then Z-score normalised per phenotype. In the case of the comparison with pathogenic variants, the list was obtained from Table 1 Vuckovic et al[8] and restricted to associations with PP >= 0.9 for both pathogenic and index variants. For Constraint Z we used the intersection of all the variants with the 1Kb qc windows[48] (v31 qc). CADD[9] raw values were obtained for all the variants and Z-score normalised. For the AbC scores we downloaded the data from[49] (AllPredictions.AvgHiC.ABC0.015.minus150.ForABCPaperV3.txt.gz). Following[7] we only considered AbC scores above 0.1 from a subset of 34 blood and immune cell datasets[50] and intersect them against the variants. The resulting overlaps were Z-score normalised. SpliceAI[51] DS_DG, DS_DL, DS_AG, DS_AL, and NCBoost[10] were run for all the variants and Z-score normalised.

PCHi-C data was downloaded from Javierre et al[14] (PCHiC_peak_matrix_cutoff5.tsv). ATAC-seq data[12] intersecting our variants was provided by the laboratory of VJ Sankaran and scaled per cell type. Chromatin state predictions were obtained from Chen et al[13]. To compute the rankings for PCHi-C, chromatin states prediction, ATAC-seq and multi-lineage ATAC we excluded the following white blood cell percentage and count phenotypes: MONO%, NEUT%, EO%, BASO%, LYMPH% and WBC#. These blood indices are known to correlate with alterations of absolute counts in other white blood cells. For chromatin states the general aim was to produce one value per variant that ranked higher epigenetic states of active enhancers and promoters that occur in relevant cell types for the blood indices, see **Table S1**. The aim was the same for PCHi-C interactions but the primary magnitude was the

Chicago Score. Given that the ATAC-seq signal was obtained in the hierarchically related cells of haematopoiesis we defined lineage to trait and lineage to cell associations, **Table S1**. Once we collected the ATAC values associated to cells and lineages we aimed to produce one value per variant attending to two criteria: i) rank higher variants that have accessible chromatin in multiple cell types within the same lineage and ii) rank higher variants that have high ATAC-seq signal in at least one of the cell types belonging to the same lineage. The multi-lineage ATAC-seq ranking was derived from the ATAC-seq ranking based on two criteria: i) rank higher variants with accessible chromatin across different lineages and ii) rank higher variants with high mean ATAC-seq ranking across lineages. In all the rankings the final step to produce a continuous value from 0 to 1 was done using the desiR package[52] and the values were subsequently Z-score normalised.

Transcript quantification for blood immune cell types was obtained from the BLUEPRINT project[13] **Table S1**, and added by gene and sample to obtain gene expression. Mean gene expression was calculated for the indicated samples by cell type. COGS scores[15] were calculated running rCOGS on the summary stats for the 29 phenotypes in Vuckovic et al[8] using seven cell types of the PCHi-C dataset representative of the different blood lineages (Ery, MK, Mon, Neu, tCD4, tCD8 and tB). The lof metrics by gene (v2.1.1) were obtained from gnomAD[53]. In all cases variants were linked to genes using the VEP consequences and PCHi-C connections. For the COGS scores and the Ranking of Gene Expression we also excluded the following white blood cell percentage and count phenotypes: MONO%, NEUT%, EO%, BASO%, LYMPH% and WBC#. For the Ranking of Gene Expression the general aim was to produce one value per variant and gene that ranked higher genes that are highly expressed in relevant cell types for the blood phenotypes. We used desiR to produce a continuous value from 0 to 1 for the Ranking of Gene Expression. Then the Ranking of Gene Expression, the COGS values and the oe_lof metric from gnomAD were Z-score normalised.

We used pairwise Wilcoxon tests to assess statistically significant differences between categories, applying multiple testing correction (Benjamini–Hochberg) across all the comparisons per score or ranking, **Table S3**. The code to produce all these annotations and rankings and the list of pathogenic variants used in the comparison with index variants can be found in Github[50,54].

<u>Conditional analysis of common variants on the index variants</u>

To condition for known common variants associated with the studied traits, we run GWAS of 29 blood cell counts in ~409 UK Biobank individuals of white British ancestry, following the procedure described in[55], the scripts are available in GitHub[56]. Accordingly, we employed the same phenotype exclusions, adjustments and normalisation approach, and ran GWAS using REGENIE[57] and TOPMed imputed genotypes[58], including recruitment centre and the first ten PCs of the kinship matrix as covariates. We then obtained independently associated variants for each trait by using the GCTA (v1.94.1) cojo[59] joint model function (with parameters: collinearity = 0.9, p-value < 1e-4) and LD structure estimates from the genotypes of 30,000 unrelated individuals of white British ancestry from UK Biobank samples. Finally, for each of the 123 rare variants, we gathered all independent GWAS common variants obtained by cojo, having MAF > 1% and joint association p-value < $1x10^{-7}$, and falling within a +/- 500kb

window from a rare variant. We then performed conditional analysis by fitting a linear regression model of the blood trait and the genotype of the associated rare variant, including the genotypes of the independent GWAS hits as covariates. The model was fitted in R (using the stats::lm function) and genotypes of both common and rare variants were obtained from the WGS of UK Biobank 200k release (GraphTyper population level WGS variants, PLINK format). Out of 123 rare non-coding variants, genotypes were available for 80 of them; of these 80, 76 (95%), were significantly (p-value < 0.05) associated with at least one blood trait even after jointly conditioning for common variants.

Obtention of Allelic Series

Starting from the 22 variants from **Table 1** we selected instances in which our index variant and one or more conditionally independent signals within the AS acted through a common gene or set of genes via VEP MSC (LOF, MISS or SYN) and e- or s-QTLs in whole blood (obtained from Open Targets[35]). We detected AS matching both criteria for 9 of the 22 variants in **Table 1**.

Cell culture

K-562 (ATCC® CCL-243™), and HL60 (ATCC®CCL-240) cells were cultured as indicated by the distributor, 1x RPMI 1640 media with L-glutamine (Gibco Medium.: 52400025), supplemented with 10% FBS (Gibco, A31604-02) and 1x penicillin/streptomycin (Gibco, 15070-063). THP-1 (ATCC® TIB-202) were culture as indicated by the distributor, 1x RPMI 1640 media with L-glutamine, 2-mercaptoethanol (Sigma, M3148) was added to a final concentration of 0.05mM and supplemented with 10% FBS (Gibco, A31604-02) and 1x penicillin/streptomycin. CHRF-288-11 (a kind gift from Prof. Wilen H Ouwehand's lab) were cultured 1x RPMI 1640 media with L-glutamine, supplemented with 20% Horse Serum (Gibco 16050-122) and 1x penicillin/streptomycin. All the cell types were maintained up to a confluence of $1x10^6$ cells/ml and then reseeded at $1x10^5$ cells/ml, except for THP-1 that were reseeded at $3x10^5$ and cultured in a T75 flask in an upright position.

Induced Pluripotent Stem Cell (iPSC) line Kolf2_c1 line (Wellcome Sanger Institute's Human Induced Pluripotent Stem Cell Initiative) was cultured in TeSR-E8 complete culture media (Stem Cell Technologies #05991) (37°C, 5% CO2) on 10ng/ml Synthemax-II (Corning CLS3535) coated plates. Kolf2_c1 were thawed into TeSR-E8 + 10% CloneR (StemCell Technologies # 05888) and split at 70-80% confluence into TeSR-E8 + 10uM Y27632 Rock Inhibitor (StemCell Technologies #72302).

MPRA enhancer activity

Library design and cloning

We designed a library of 20,340 200-mer oligonucleotides that were synthesised by Twist Bioscience. The library covered 113 SNPs, each one assayed in five partially overlapping tiles, every tile having an alternative and reference allele version. Each reference or

alternative allele tile was tagged by 15 unique 11 bp barcodes. The library included 7 MPRA-ASE and 10 EA positive controls from[12] and two sequences that showed no CRISPRa activity in Fulco et al[20] as negative controls. The structure of the 200-mers included two 15 bp amplification arms at each end to amplify subpools/bins of the library based in GC content, an 11 bp barcode, the restriction enzyme sites for *BamH*I and *KpnI* and 148 bp of candidate regulatory sequence to be assayed. The amplification PCRs (Primers 1-6 **Table S7**) were done with Kapa HiFi HS Ready Mix (Kapa Biosystems 07983034001), using 20 ng input template and 50 ul final volume except for the High GC bin in which 20 ul of KAPA2G GC Buffer (ROCHE KK5516) was added to a final volume of 100 ul. Next we performed a digestion with *ExoI* (NEB) to eliminate free primers and then purified the amplified fragments using Agencourt AMPure beads (Beckman Coulter A63881).

We based the backbone vector for this assay on the hSTARR-seq_ORI vector (Addgene #99296) following the recommendations from Muerdter et al[18]. We added GFP to the vector by excising sgGFP from the pSTARR-seq_human vector (Addgene Plasmid #71509) with an o.n. digestion with *AflII* and *Age*I (New England Biolabs, NEB) at 37°C and ligating it with hSTARR-seq_ORI vector digested in the same manner (T4 DNA ligase 16℃ o.n. 3 to 1 molar ratio). Following an idea proposed in the Supplementary Figure 1 I) of Muerdter et al[18], we excised the polyA site from the hSTARR-seq_ORI GFP vector to be cloned back at a later stage of our library construction to separate the barcode and the candidate regulatory sequence. The excision was carried out by *Nae*I digestion of the vector (60' 10 Units of enzyme, NEB) and posterior blunt end ligation to obtain the hSTARR-seq_ORI GFP polyA MINUS vector. The excised SV40 poly (A) signal fragment was amplified (Primers 9-10 **Table S7**) and cloned in the pGEM T easy system (Promega A1360).

The amplified subpools of the library and the hSTARR-seq_ORI GFP polyA MINUS vector were ligated using Gibson cloning (NEB). Briefly, the vector was linearized by PCR (Primers 7-8 **Table S7**) and subsequently we carried out a digestion with *DpnI* and *BamH*I (NEB) to degrade the circular template. The ligation in Gibson mix was done with 100 ng of the linearized vector and a molar insert:vector ratio of 2:1 for each of the bins. The ligations were purified with Agencourt AMPure beads and eluted in 20 ul of Elution buffer diluted 1/10 in nuclease-free water and 10 ul of each were then used to electroporate electrocompetent E. coli bacteria (NEB C3020K) at 2000 V 25 uF 200 Ohm. We performed serial dilutions for each of the bins to ascertain the yield in colony-forming units (CFU) and aimed to keep a ratio of at least 100 CFU per oligo element. The individual colonies in each bin were lysed and the plasmids corresponding to each bin were purified using the Qiagen maxi prep kit (cat. 12162). This intermediate step in the library construction (PolyA Minus library) was sequenced to check the barcode - candidate regulatory sequence association and the design dropout rate. Briefly, we amplified 25 ngr of each of the poly (A) minus library bins (Medium GC, High GC and Low GC content) with oligos 11 and 19 (**Table S7**) for 15 cycles with annealing and extension done at 72°C in a combined step for 1'. The libraries were then quantified with KAPA Illumina SYBR Universal Lib Q. Kit. (Roche. 07960140001), adjusted at 4nM and pooled together for a final volume of 40 ul. The samples were subsequently sequenced in a MiSeq (MiSeq Reagent Kit v2 300 cycle Illumina (MS-102-2002) with the first 15 cycles of read 1 set to dark cycling and using custom primers 22 and 23 (**Table S7**). The PhiX amount was set to 10%.

Finally the polyA site was introduced back into the vector separating the barcode from the regulatory sequence using the *BamH*I and *KpnI* restriction sites that were introduced in the oligo design. We digested the PolyA Minus library with *Kpn*I (NEB, 37°C o.n.), followed by gel purification and digestion with *BamHI* in the presence of Shrimp Phosphatase (NEB M0371S) for 2h at 37°C. In parallel, we digested overnight 20 ugr of the pGEMT-polyA vector to release the SV40 polyA signal (230 bp) with *KpnI* and *BamHI* and gel purified it. We cleaned both fragments with Agencourt AMPure beads and quantified them using Qubit 1X dsDNA BR Assay Kit (cat. Q33265). For each bin we used 100 ngr of the input vector and a 2:1 insert:vector ratio in the T4 ligase 16°C overnight reactions. We purified the ligations with Agencourt AMPure beads and performed electroporations in the same conditions as the ones used for the PolyA Minus library. After evaluating the efficiency of the electroporation we seeded 0.5 million cfu per 245 mm Square BioAssayDish with Agar+Ampicillin, keeping at least 100 CFU per oligo element. A step-by-step protocol of the procedure is available in protocols.io[60].

Nucleofection and parallel mRNA and gDNA isolation

Cells were seeded at 1.5 to $2x10^5$ /ml into 2-4 T175 flasks (Corning, CLS431085-50EA) in 60 ml of fresh medium (9 to 12 million cells each flask), and cultured for 48 hours changing half of their media 24 hours after seeding. The electroporation was performed using the Neon Transfection System (ThermoFisher, Cat# MPK5000). In K-562 and HL-60 we electroporated 5 million cells with 25 ugr of a pool of the three plasmid bins per reaction (12.5 ugr of the Medium GC bin and 6.25 ugr of each of the other two bins). In the case of CHRF and THP-1 cells we electroporated 5 million cells with a four plasmid mix (10 ugr of the Medium GC bin, 5 ugr of each of the other two bins and 5 ugr of the pmaxGFP vector). The nucleofection was performed following the manufacturer instructions using buffer R. The conditions for K-562, THP-1 and CHRF cells were 1.45 v, 10 ms pulse width and 3 pulses, for HL60 1.350 v, 35 ms pulse width and 1 pulse. The efficiency of the nucleofection was evaluated at 24h and 48h post nucleofection in the pmaxGFP plate with a Countess II FL Automated Cell Counter (Cat # AMQAF1000). We routinely observed 80-90% efficiency for K-562 cells. In the case of CHRF and THP-1 due to the low transfection efficiency cells were co-transfected with pMAX-GFP and we used FACS to purify GFP positive cells to enrich cells that have incorporated plasmids in the nucleofection step.

48 h post-electroporation cells were spun down (300G 5') and resuspended in DNase I (NEB M0303L) at 37℃ for 15' using 10 U of enzyme per ug of plasmid transfected in a final volume of 2 ml of DPBS to eliminate possible carryover plasmid in the exterior of the cells. Cells were then pelleted (300 G for 5'), washed twice with DPBS and lysed in 600 ul of Buffer RLT Plus (Qiagen) with added β-mercaptoethanol and homogenised using QIAshredder columns (Qiagen 79654). DNA and total RNA were extracted using AllPrep DNA/RNA Kit (Qiagen, Cat# 80204) according to the manufacturer's instructions. In the RNA preparation, a step of on-column DNase I treatment was performed for all samples (Rnase_Free Dnase Set, Qiagen 79254). We isolated mRNA from total RNA using the Oligotex mRNA Kit (Qiagen 70042) followed by a final treatment with Turbo Dnase (Invitrogen AM1907). DNA and mRNA quantifications were done using Qubit RNA Quantification, high sensitivity assay (cat.

Q32852) . A detailed protocol can be found in[61,62].

Library preparation and sequencing

We retrotranscribed 1-1,5 ug of mRNA per replica following the protocol for SuperScript IV (ThermoFisher 18090050) using a reporter specific RT primer (Primer 24, **Table S7**) at 2uM carrying the 10-mer UMIs. Then we split retrotranscription samples for PCR amplification so as the RT template would represent 10% of the final volume of the PCR (50 ul). We performed the first round of amplification in which we introduced the sample index primer, i5-i8 (Primers 15-18 **Table S7**) for the cDNA samples of four replicas. As a reverse primer, we used P7 (Primer 21 **Table S7**). The PCR was carried out using Kapa HiFi HS Ready Mix, and 65℃ annealing temperature for a total of 3 cycles. The amplification from each replica was then pooled, purified using Agencourt AMPure XP beads and then we assessed the minimum number of cycles for a second round of PCR by q-PCR with P5 and P7 with StepOnePlus™ Real-Time PCR System (ThermoFisher cat.4376600, (Primers 20 and 21 **Table S7**). We determined between 11 and 13 cycles to be a good average range to keep the second round PCR from plateauing. Following Klein et al[63] we split each of the replicas into 8 reactions and performed the second round PCR for the cycles determined with the P5 and P7 primers and Kapa HiFi HS Ready Mix at 64℃ annealing temperature. We then pooled the reactions from each replica, purified using Agencourt AMPure XP beads and eluted in 60 ul of Elution Buffer.

The gDNA fraction from the All Prep Qiagen kit was used as a source for the episomal plasmid nucleofected in every replica. For every replica, we used 12 ugr of gDNA that was split into 24 PCR reactions following Klein et al[63]. In this first reaction we introduced the sample index and the UMIs using the primers 11-14 (**Table S7**) as forward primers and primer 24 (**Table S7**) as reverse primer. The PCR was run for 3 cycles at 65℃ annealing temperature and the reactions corresponding to each replica were pooled, purified using Agencourt AMPure XPbeads and eluted in 320 ul of Elution Buffer. As in the cDNA library preparation we assessed by qPCR the number of cycles to keep the second round PCR from plateauing. We determined 10-11 cycles. We split each of the replicas into 29 reactions and performed the second round PCR with the P5 and P7 primers and Kapa HiFi HS Ready Mix at 72°C annealing temperature. Finally, we pooled the reactions for each replica, purified 200 ul of the mix using Agencourt AMPure XP system (1.2X vol/vol of beads) and eluted in 30 ul of Elution Buffer.

All the libraries were quantified using KAPA Illumina SYBR Universal Lib Q. Kit, adjusted to 4nM and pooled afterwards in 40 ul final volume. We used a HiSeq 2500 RR for sequencing each batch with the Hiseq Rapid PE Cluster Kit V2 (Illumina BCSO1412) and the HiSeq Rapid SBS Kit v2 200 cycles (Illumina FC-402-4021). The recipe included the first 15 cycles of read 1 set to dark cycling and used custom primers 22, 25 and 26 from **Table S7**. The amount of PhiX was set to 10%. A detailed protocol can be found in[62].

Processing and alignment of reads

The first six nucleotides of the r2 reads corresponding to the *BamHI* site (see protocols io)

were removed. UMIs were added to each read ID in r1 and r2 files using UMI tools[64] and merged using flash[65] prior to aligning against the reference set of barcodes using bwa[66]. Only primary alignments were taken forward. We discarded alignments not matching perfectly the corresponding barcodes with bamtools[67] and deduplicated the UMIs per barcode using UMI tools. Finally we counted all the unique UMIs across all the tagging barcodes of a regulatory sequence.The pipeline is available in GitHub[68].

Analysis of MPRA

A series of in-house R scripts were used to put together the episomal and cDNA samples of all the replicas and fill in with 0 the missing data for dropout barcodes to obtain a regular matrix. This matrix was then filtered, only barcodes having at least 5 counts in all the replicas in the cDNA and episomal DNA samples were kept and transformed into the input format accepted by MPRAnalyze[21]. Samples were then normalised using upper Quartile normalisation. We calculated mean FoldChanges of cDNA counts over episomal DNA counts as well as a mean measure of Allele Specific Expression (ASE) described as (RNA ALT allele counts/DNA ALT allele counts) / (RNA REF allele counts/DNA REF allele counts) counts for every tile following Vockley et al[22]. MPRAnalyze[21] was used to calculate the tiles that exhibited significant EA and ASE across the biological replicates. For the enhancer analysis, the two negative controls were used to set the baseline of the analysis. For the ASE we set the reducedDesign parameter to ~1. As a quality control measure we excluded tiles with an imbalance between reference and alternative alleles in the gDNA higher or equal to 5-fold. EA tiles were required to have a LogFC value equal or greater than 0.05 and enhancer -log p value equal or above 1.3. ASE tiles were required to have an ASE value below 0.95 or above1.05 as well as an ASE -log p value equal or above 1.3 at FDR 0.05. The scripts are available in GitHub[69].

Predicting TF motif intersecting the variants

For the 94 variants screened in MPRA we selected a sequence stretch of 38 bp centred on the SNP and extracted the reference allele sequence. In the case of the 80-bp deletion we used the complete 80 bp nucleotide stretch. We then substituted the reference allele with the alternative allele in the position of the SNP to obtain the alternative allele version. We predicted TF motifs in both the reference and alternative allele versions of the 38 bp nucleotide stretch with gimme motifs[70] with the reference databases HOMER[71] and JASPAR[72], the -c option set to 0.85 and the maximum number of TF motif per nucleotide stretch set to 20. For all the SNPs only the TF motifs overlapping the SNP were taken forward in the analysis. For the 80-bp deletion we used all the motifs within the 80 bp stretch. We then classified all variants overlapping motifs for *IRF3*, *IRF7*, *IRF1*, *IRF2*, *IRF9*, *IRF4*, *IRF8*, *RUNX2*, *RUNX1*, *RUNX3*, *FOSL1*, *FOS*, *FOSB*, *FOSL2* and *PU.1* as 'context-initiator' TF motifs and variants overlapping motifs for *FOXA1*, *FOXA2*, *FOXA3*, *FOXC*, *FOXP1*, *FOXP2*, *GATA3*, *MEF2C*, *LHX1* and *NKX6-1* as 'context-only' TF motifs following Kribelbauer et al[25]. Four variants overlapped motifs for both classes and were reclassified as 'context-only'. We were able to annotate 25 and 24 variants as overlapping 'context-initiator' and 'context-only' TF motifs respectively. When restricted to variants also showing DE/ATU regulated genes in the RNA-seq analysis the number decreased to 11 and 14 respectively. The scripts are

available in GitHub[73].

Enformer run of variants intersecting 'context-initiator' and 'context-only' TF motifs

Enformer values[26] were obtained from the vcf file of the 94 variants screened in the MPRA. Of the 5,313 features output by Enformer we selected the DNASE values of resting K-562 (121_DNASE:K562, 122_DNASE:K562, 123_DNASE:K562 and 625_DNASE:K562) and of HL-60 (96_DNASE:HL-60) cell lines. For K-562 we took the mean of the four values per variant as the representative value. To plot Enformer DNASE effect sizes against MPRA ASE values following[26] we first took the ASE value of the tile centred in the variant (TILE 3) as a representative value of the variant. We then z-scored the ASE values and restricted our analysis of correlation to the 27 and 24 variants overlapping 'context-initiator' and 'context-only' TF motifs. The scripts are available in GitHub[74].

GenIE

CRISPR/Cas9 targeting and amplicon design

We followed the protocol described in Cooper et al[27]. The Wellcome Sanger Institute Genome Editing browser (WGE) (https://www.sanger.ac.uk/htgt/wge/), was used to choose a CRISPR gRNAs with NGG PAM site within 20bp of the SNP locus and with less than 1-3 mismatch off-target hits predicted. To introduce the SNP of interest a 100bp repair template oligonucleotide was designed. As a positive control for cutting we used the sgRNA against ENSG00000178927/ CYBC1/ EROS (numbers 74-78 from **Table S7**).

Primers were designed to amplify <250bp across the SNP of interest, 40-60% GC, Tm 56-65°C (NEB Tm Calculator) and with adaptor sequence tails for MiSeq Sequencing (See **Table S7**). Reverse transcriptase primers were designed downstream of the amplicon in the mRNA sequence (See **Table S7**).

Nucleofection

Guide RNAs (IDT) were annealed to tracrRNA (IDT 1072532) in duplex buffer (IDT 1072570) at 95°C for 2 min and cooled slowly to RT. Nucleofection on Kolf2_c1 was carried out as previously described[75] and recovered onto 4ng/ul Synthemax-II (Corning CLS3535) coated 6 well plates. K-562 and HL-60 were nucleofected following Lonza's protocols. THP1 were nucleofected following Lonza's primary monocyte protocol. Cells were cultured for 1-2 weeks until confluent and snap frozen as 2-3 x10^6 cell pellets.

Library preparation and sequencing

gDNA was extracted using DNA MagAttract HMW extraction kit (Qiagen #67563), following standard instructions and eluted in 100ul H20. Total RNA was extracted using Direct-zol RNA Miniprep Plus kit (Zymo R2070), TURBO DNAse treated (Thermo-fisher AM1907) and run on a 2100 RNA Nano Chip in a Bioanalyser (Agilent 5067-1511). Gene specific RT primers were annealed using 2ug RNA, 2uM RT primer and 10mM dNTPs, heated to 65°C for

5 min and placed on ice. cDNA synthesis was setup using half the annealed RNA, Superscript IV (ThermoFisher 18090050) and RNasin (Promega N261A) on ice and heated 50°C 10min, 55°C 10min, 60°C 10 min, 80°C 10min, 4°C hold.

Genomic DNA and whole RNA were amplified with PowerUP SYBR green master mix (Applied Biosystems A25742) or Q5 Hot Start polymerase (New England Biolabs M0494S), 10uM adaptor sequence primers with 4 PCR replicates for gDNA and 8 PCR replicates for cDNA. PCR conditions for PowerUP SYBR reactions were as previously described[27]. PCR conditions for Q5 Amplicons: 98°C 30s, (98°C 10s, 57°C 20s, 72°C 20s)x30, 72°C 2min. Amplicons were barcoded using WTSI PCR barcoding primers, pooled, gel extracted using Minelute kit (Qiagen 28604) and quantified by qPCR (KAPA library quantification kit KK4834). The prepared library was loaded onto a MiSeq System (Illumina) at 4nM with 20% PhiX using MiSeq Reagent Kit v2 300 cycles (Illumina BCSO0012).

Analysis of GenIE

Read pairs per amplicon were merged using flash[65] with the following parameters: read length (-r) 150, fragment sd (-s) 20, minimum overlap (-m) 10. Fragment size (-f) and maximum mismatch density (-x) depended on the targeted amplification: BID (193, 0.109 respectively), BRAP (162,0.107), C2CD5 (285, 0.15), CUX1 (223, 0.115), EPB41 (235, 0.115), ENSG00000178927/CYBC1/EROS (249, 0.12), FOXP1 (222, 0.113), NBN (188, 0.109), TNRC6A (292, 0.15) and UGCG (230, 0.114). The reads were aligned using bwa mem[76] with the following parameters: -O 24,48 -E 1 -A 4 -B 16 -T 70 -k 19 -w 200 -d 600 -L 20 -U 40. The aligned reads were filtered to discard reads with more than 10 clipped bases using samclip[77]. The GenIE results were obtained using the rgenie package[27,28] allowing for 10 bases for the required_match_right and required_match_left parameters. The scripts are available at GitHub[78].

INTERVAL WGS analysis workflow

The INTERVAL whole genome sequencing data (WGS) were generated at the Wellcome Sanger Institute. The manuscript describing WGS in full is in preparation. Briefly, WGS was performed on 12,354 samples using the Illumina HiSeq X10 platform as paired-end 151 bp reads. Raw read processing was carried out via customised pipelines at WSI. Reads were aligned with BWA MEM to the GRCh38 human reference genome with decoys (also known as HS38DH). Variants were called for each sample using GATK HaplotypeCaller version 4.0.0. Then all samples were merged, and the combined samples genotyped using GATK4.0.10.1. GATK Variant Quality Score Recalibration (VQSR) was used to identify probable false positive calls by assigning quality score log-odds (VQSLOD) separately for SNPs and INDELs using GATK VariantRecalibrator (v4.0.10.1). Sample quality control removed 491 samples in total, including 77 samples with coverage below 12x, 134 samples with > 3% non-reference discordance (NRD), 118 samples with > 3% FreeMix (VerifyBamID2) score, 221 samples failing identity checks, 30 samples swapped, 40 samples failing sex checks, 39 duplicates and 9 samples with possible contamination. Genotypes with allele read balance > 0.1 for homozygous reference variants, < 0.9 for homozygous alternative variants or not between 0.2-0.8 for heterozygous variants were removed. Genotypes were

also removed if the proportion of informative reads was < 0.9 or read depth > 100. We performed additional variant quality control and filtered out variants that failed to meet the following requirements: call rate per site > 95%, mean genotype quality (GQ) value > 20, Hardy-Weinberg equilibrium (HWE) p-value > 1×10^{-6} only for autosomes. All monomorphic variants with alternative allele count (AAC) = 0 were further removed, although we kept all monomorphic variants with reference allele count (RAC) = 0. For chrX and chrY we applied an additional step to correct allele counts and frequencies due to female and male samples accounting for diploidy/haploidy in the PAR and non-PAR regions. Finally, the WGS data set contains 116,382,870 variants (100,694,832 SNVs and 15,688,038 indels) including 6,637,420 (5.7%) multi-allelic sites across 11,863 participants.

INTERVAL RNA-Seq analysis workflow

Gene expression counts

The INTERVAL RNA-sequencing data were generated and processed as previously described[29]. We mapped the RNA-sequencing data to the GRCh38 reference and quantified read counts as described previously[79] with the difference of using GENCODE v31 across 4,731 samples passing quality control. Globin genes, rRNA genes, and pseudogenes were removed. 19,841 genes were selected with > 0.5 CPM in at least 1% of the samples. Gene expression counts were converted to FPKM, trimmed mean of M-values (TMM) normalised and $\log_2$ transformed. We used the probabilistic estimation of expression residuals (PEER) method[80], implemented in the R package peer v.1.0[81], to correct for latent batch effects and other unknown confounders. 50 PEER factors were calculated with age, sex, BMI, and 19 blood cell traits included as covariates.

Transcript quantification

We used Salmon v1.1.0[82] for transcript expression quantification. The Salmon index was built against GRCh38 cDNA. R packages tximport v1.14.2, AnnotationHub v2.18.0, BiocFileCache v1.10.2, BiocGenerics v0.32.0 were applied to obtain various count matrices from these quantifications at the transcript or gene level. We subsetted to 4,731 samples passing RNA sequencing QC and corrected sample swaps. We focused on the transcripts of the 19,841 genes that passed gene QC. From these transcripts, we selected transcripts with TPM ≥ 0.1 in at least 20% samples. Subsequently, TPM values were TMM normalised and log2 transformed.

Differential Expression linear model

We modelled FPKM normalised raw values for each gene with a linear model using as covariates the PEER factors (top 35 from 50), top 10 genotype principal components, sex, age, BMI, RIN, sequencing batch, RNA concentration, read depth, and season (based on month of blood draw). Depending on the blood phenotypes associated with the variants at PP >= 0.1 we included a set of cell-count specific covariates (see **Table S1**) to account for cell count effects on total blood composition.

The multiple testing correction was done using the Benjamini–Hochberg method at three levels for each variant: i) all transcripts of all the genes in which the variant had a direct VEP Most Severe Consequence (LOF, MISS, SYN, UTR5, UTR3, INTRON, SPLICE, UPSTREAM), ii) all transcripts of all the genes within the GWAS association block plus the genes connected to the variant via PCHi-C (Block and PCHi-C levels) and iii) Only for variants in Table 1, all the transcripts of all the genes (genome wide). The code used in this analysis is in GitHub[83].

Alternative Transcript Usage (ATU) additive logratio model

We calculated the median transcript ratio (median transcript TPM/median Expression of the gene to which the transcript belongs) for homozygous reference and heterozygous carriers separately. We discarded all transcripts with median transcript ratios below 0.1 in both genotypes to filter out transcripts whose contribution to the total expression of a given gene remains low irrespective of the genotype.

Given the compositional nature of the data we decided to transform proportions using the additive logratio model. Briefly, for all the transcripts belonging to the same gene we first estimated the transcript ratio of gene expression (expression of the transcript/sum of the expression of all the transcripts belonging to the same gene) and scaled it to a reference transcript. The most abundant transcript was chosen as the reference transcript for each gene to avoid having 0 values at the denominator. This proportion was then log transformed. To avoid having transcripts with 0 TPM value in the logratio model, for any given transcript we imputed the value of the samples with 0 TPM to 0.65 of the minimum value greater than 0 for that transcript as this value is suggested to limit the distortion of the covariance matrix[84]. The resulting log scaled ratio per transcript was used in a linear regression model with the same constitutive and cell count specific covariates (see **Table S1**) per variant as the ones used in the DE model.

The multiple testing correction of the ATU model was done using the Benjamini–Hochberg method at two levels for each variant: i) all transcripts of all the genes in which the variant had a direct VEP Most Severe Consequence (LOF, MISS, SYN, UTR5, UTR3, INTRON, SPLICE, UPSTREAM), ii) all transcripts of all the genes within the GWAS association block plus the genes connected to the variant via PCHi-C (Block and PCHi-C levels).The code used in this analysis is in GitHub[83]. To represent the transcripts we used ggtranscript[85].

GTEx and INTERVAL sQTLs

For the 20 genes with ATU in our survey of INTERVAL whole blood RNA-seq (*ATL1*, *NPRL3*, *VMP1*, *ELP5*, *KDSR*, *RASAL3*, *SLC11A1*, *MFSD2B*, *GATA2*, *EEFSEC*, *TAF8*, *IKZF1*, *PILRB*, *ANK1*, *GFI1B*, *EVI5*, *TYMP*, *ARSA*, *ODF3B* and *PILRB*) we queried the GTEx[32] web portal (release v8) and found sQTLs in whole blood for *PILRB*, *SLC11A1*, *ARSA*, *ANK1*, *MFSD2B*, *TAF8*, *GFI1B* and *RASAL3*. The same query in the INTERVAL web[86] yielded 16 genes: *MFSD2B*, *PILRB*, *TYMP*, *ANK1*, *ARSA*, *EVI5*, *NPRL3*, *VMP1*, *GFI1B*, *SLC11A1*, *IKZF1*, *EEFSEC*, *ODF3B*, *TAF8 GATA2* and *RASAL3*. We obtained all the proxy variants at R^2=0.7 in the European subpopulations (EUR)

in a window size of 0.5 Mb for all the sQTLs in these genes using LDLinkR[87]. None of the index variants leading to ATU were proxies of the GTEx sQTLs. Three of the index variants with ATU genes (rs149489081-*ANK1*, rs543594419-*TYMP, ODF3B* and *ARSA* and rs187715179-*GFI1B*) were sQTLs in the same genes in INTERVAL. The code for this analysis is available in GitHub[88,89].

GSEA and ORA

For the GSEA analysis we started from the 22 variants in **Table 1** and performed a genome wide DE analysis between carriers and non-carriers per variant in the INTERVAL whole blood dataset. Next, we used the Fold Change between WT and HET genotypes to order the genes in a decreasing manner and input the ordered gene list into the GSEA function of ClusterProfiler[90]. The minimum and maximum gene set size and the p value cutoff were set to 10, 500 and 0.05, respectively. The multiple testing was accounted for by Benjamini & Hochberg. The code used in this analysis is in GitHub[83].

For the ORA analysis we started by defining a list of gene sets for blood and immunity from the Molecular Signatures Database (MSigDB)[91]: we first selected all the gene sets that contained in their description at least one of the following blood and immunity related terms: PLATELET, ERYTHRO, MEGAKARYOCYTE, MONOCYTE, NEUTROPHIL, EOSINOPHIL, BASOPHIL, LYMPHOCYTE, T_HELPER, TH_17, TH17, TH1, TH2, BLOOD, BLOOD_COAGULATION, IMMUNE, HUMORAL_IMMUNE_RESPONSE, IMMUNOGLOBULIN and HEMATOPOIETIC. We also included two blood and immunity unrelated terms (HEPATOCYTE and NEURON). In addition, we included in the ORA analysis TF target gene sets for the TFs *GATA2, GFI1, CUX1* and *RUNX1* in the Dorothea collection[37] (A,B ,C and D confidence levels). The minimum and maximum gene set size allowed was 10 and 500, respectively, but for the Dorothea gene sets which were exempted from this filter (*CUX1* n=464 genes, *FOXP1* n=3,278, *GATA2* n=5,370, *GFI1* n=8 and *RUNX1* n=2,551). The list of selected pathways (*n=1,423 gene sets*) can be found in GitHub[92]. Next, we tested the genome wide DE analysis of the 22 variants in **Table 1** used for the GSEA for overrepresentation of DE genes applying Active Pathways[93] with the significant p value cutoff set to 0.01. As a background list of gene sets we used all the gene sets derived from the Human Phenotype ontology (c5.hpo.v2023.2.Hs.entrez.gmt, *n=5,547 gene sets*[41,91]). We obtained 64 pathways significantly overrepresented in DE genes of which 3 were from the unrelated terms. The remaining 65 pathways corresponded to 11 variants. We excluded from further analysis the gene sets from Dorothea (2 variants). A variant was labelled as 'ORA positive' if it had at least 1 ORA pathway at p value = 0.05. To test if the 9 variants were enriched in the TF annotation (**Table 1**) we used a Chi square test (6/7 variants with TF annotation vs 3/15 non TF labelled variants were 'ORA positive', p value = 0.0141). The code is in GitHub[92].

BLUEPRINT RNA-Seq analysis workflow

Differential Expression linear model

We used the normalised gene quantification - values from BLUEPRINT data [13]. We then

applied a linear model to the gene quantification values for each gene.

The multiple testing correction was done using the Benjamini–Hochberg method at two levels for each variant: i) all transcripts of all the genes in which the variant had a direct VEP Most Severe Consequence (LOF, MISS, SYN, UTR5, UTR3, INTRON, SPLICE, UPSTREAM) and ii) all transcripts of all the genes within the GWAS association block plus the genes connected to the variant via PCHi-C (Block and PCHi-C levels).

The code used in this analysis is in GitHub[94].

ATU BP

We calculated the median transcript ratio (median transcript FPKM/median Expression of the gene to which the transcript belongs) for homozygous reference and heterozygous carriers per cell type separately. For every cell type of BLUEPRINT we discarded all transcripts with median transcript ratios below 0.1 in both genotypes.

The same compositional model and multiple testing correction used for whole blood was applied in the BLUEPRINT per cell type. The code used in this analysis is in GitHub[94]

Haemvar architecture/ documentation

We created the HaemVar Database of genetic variants and blood-related traits (https://haemvar.org), containing the complete association and fine-mapping result data set of all the variants associations to 29 blood cell phenotypes assessed by Vuckovic et al[8], and including further annotation data for the prioritised subset of 178,890 variants as described above; see section 'Variant Annotation'. All information contained in the HaemVar Database is presented through an open-access website that generates comprehensive gene and variant-specific data reports with downloadable tables and figures. The software of the web-based application is written in PHP for server-side handling and optimisation of database access and data output, and uses the open-source Vega v5 (http://vega.github.io) JavaScript library for custom generation of data-driven and (partially) interactive visualisations.

Analysis of MPRA and DE ASE concordance

First, with all the variants analysed in the RNA-seq survey (both INTERVAL and BLUEPRINT studies, **Table S5**) we keep the variants that had at least 1 DE gene in at least one of the studies (34 variants). Then we classified the directionality of the beta in the DE model: beta < 0 implied that the gene was more expressed in the carriers of the homozygous reference allele (REF direction) and beta > 0 implied that the gene was more expressed in the heterozygous carriers of the alternative allele (ALT direction). For the MPRA results (**Table S4**) we selected the screened variants with at least 1 ASE or EA+ASE tile in at least one cell type (92 variants). Then we classified the selected variants and tiles according to the ASE value: ASE < 1 implied that the ASE was higher in the reference allele (REF direction) and ASE > 1 implied that the ASE was higher in the alternative allele (ALT direction). Finally, we merged both tables and classified as "CONCORDANT MPRA DE" the cases in which at least one MPRA tile had the same directionality of the DE gene(s). If a variant had more than one DE gene all the genes had to have concordant MPRA tiles for the variant to be classified as "CONCORDANT MPRA DE". In total we had 12 "CONCORDANT MPRA DE" variants and one "DISCORDANT MPRA DE". The code can be found in GitHub[95].

Work with K-562 clones of CUX1

CRISPR/Cas9 sgRNA designs and protocol

To introduce the rs139141690 (G>A) mutation into K-562 cells, a synthetic crRNA was selected using CRISPOR[96]. An 81bp ssODN carrying the mutation G>A was chemically made by IDT with 4 phosphorothioate bonds. To delete the 80 bp region encompassing the rs139141690 two crRNAs were selected one upstream and one downstream of the desired SNP.

The crRNA and trans-activating crRNA (tracrRNA) were synthesised by IDT. An electroporation enhancer was also bought by IDT and resuspended at 100 µM in water. SgRNAs are made by combining 160 µM of crRNA and tracrRNA (1:1 v/v) to get a final concentration of 80 µM and incubated at 37C for 30 min.

To assemble the Cas9/sgRNA RNPs, the sgRNAs and electroporation enhancer (0.8:1 volume ratio) were first mixed and then 40µM S.p. Cas9 Nuclease (IDT) at 1:1 v/v was added. This mixture was incubated at 37C for 15-30 min prior use.

Electroporation was performed using SF Cell Line 4D-Nucleofector™ X Kit (Lonza) according to manufacturer's instructions. The kit / program for K-562 used is SF kit / FF-120.

For the *knock-in* experiment, 50 pmol *of the RNP* was electroporated into K-562 cells together with 4µM ssODN as HDR template. 15 min after electroporation cells are incubated with a combination of 0.5µM Trichostatin A (TSA, Selleckem) and 1µM M3814 (Selleckem) to enhance the HDR as previously reported in Shy et al[97]. 24 hours post treatment drugs are removed and fresh medium is added[97]. For the deletion of 80 bp, 50 pmol of each of the two RNPs was electroporated.

After 24h cells were single-cell sorted in 96 wells plates using MoFlo Astrios (Beckman Coulter). Propidium Iodide (Sigma Aldrich) was added prior to analysis as a cell viability dye.

Isolation of clones and genotyping by amplicon sequencing

K-562 clones were let to grow from single cells for 10-14 days, then genomic DNA was extracted with the QIAamp 96 DNA QIAcube HT Kit and quantified using the QuantiFluor® dsDNA System (Promega). 2-5 ng of DNA from every colony was used as template for targeted amplicon sequencing (Fw primer and Rv primer, primers 79 and 80, respectively, **Table S7**). Samples were then indexed with Nextera XT DNA Library Preparation Kit (Illumina), pooled and sequenced on NovaSeq 6000 (500 cycles, PE).

Data were analysed by using CRISPResso2 v2.2.12[98] and clones with the desired genotypes were expanded to generate the modified cell lines.

In the same nucleofection we isolated: i) three clone lines homozygous reference for the SNP rs139141690 chr7:101499930 ('G/G'), ii) one heterozygous clone line chr7:101499930 ('A/G') and iii) three homozygous alternative clone lines chr7:101499930 ('A/A'). In addition to this we isolated: i) one homozygous clone line carrying the deletion of 16 bp (chr7:101499917 CTCACTAGAGCAAGTCC > C) and ii) three homozygous clone lines carrying the 80 bp deletion (chr7:101499894 GTTAGTGACTTCAAAAGCTGTCCTCACTAGAGCAAGTCCAACTCTTCCTCTA GTTCTGATGACTTCACGGCAGCCAACTG > G). All the coordinates are in GRCh37. The correspondent coordinates in GRCh38 are: chr7:101499930 -> chr7:101856650, chr7:101499917 -> chr7:101856637 and chr7:101499894->chr7:101856614.

Lentiviral barcoding

Plasmids carrying single unique barcodes were isolated from the Larry Barcode Library V1[99], characterised by whole plasmid sequencing (Eurofins) to identify the different barcodes and used to transfect Phoenix Ampho (ATCC, CRL-3213) cells, together with the pMD2.G and psPAX2 lentiviral packaging vectors. 48h after the transfection, the viral supernatant was collected, filtered, and used to infect the K-562 clones with the different genotypes, in presence of 6ug/ml polybrene.

GFP+ cells were sorted after 4 days and expanded to obtain 11 different cell lines (K-562-Larry), each one carrying a specific barcode, detectable both at gDNA and mRNA level.

PMA differentiation to CD41 expressing K-562 cells and Flow cytometry tracking

K-562 cells were seeded at 100.000 cells/ml in IMDM + 10% FBS and cultured in presence of 5nM phorbol 12-myristate 13-acetate (PMA, Selleckchem) or DMSO for 16, 24, 48 and 72 hours. For each time point cells were analysed by flow cytometry and collected for RNA extraction.

For flow cytometry tracking, cells were washed in PBS and incubated with anti-CD235a BUV395 (Clone GA-R2, BD Biosciences) and anti-CD41 AF700 (clone HIP8, Biolegend) antibodies for 20 min at 4°C in PBS + 1% FCS + 2mM EDTA. Cells were then washed and acquired with the CytoFLEX (Beckman Coulter) flow cytometer. DAPI (Sigma Aldrich) was added prior to analysis as a cell viability dye. Data was analysed using the FlowJo software.

ILR compositional analysis of Flow cytometry data

We collected the percentages of $CD41^{-}CD235^{-}$, $CD41^{-}CD235^{+}$, $CD41^{+}CD235^{+}$, $CD41^{+}CD235^{-}$ cells and the value of CD41 MFI and GeoMFI at the four time points of the PMA differentiations per genotype and clone line (*n=3* clone lines per genotype). The 'Basal' time point was assigned to cells mock treated for 16 hours. We used the package Compositions[100,101] to transform the cell abundance percentages into isometric log ratios (ilr)[102]. Next we model the ilr values using as covariates time (reduced model) and time and Genotype (full model) and tested the significance between adding the different terms to the models by ANOVA. The code for the analysis is in GitHub[103].

Analysis of cell size using flow cytometry FSC-A and SSC-A

We collected the values of FSC-A and SSC-A per cell, genotype and time point and restricted our analysis to $CD41^{+}CD235^{-}$ cells at 72 hours as they were the most mature cell type in our model (*n=4,892*, *8,914* and *7,553* for the wild type clone lines, *n=3,616*, *4,307* and *6,229* for the homozygous alternative clone lines and *n=9,822*, *6,476* and *9,611* for the 80 bp deletion clone lines). We used the Wilcoxon-rank test to analyse differences between the cells. The code is in GitHub[104].

Single Cell Multiome ATAC + Gene Expression

Cell culture and nuclei isolation

K-562 clones with the same genotype were pooled and seeded at 100.000 cells/ml in IMDM + 10% FBS and cultured in presence of 5nM phorbol 12-myristate 13-acetate (PMA, Selleckchem) or DMSO for 24, 48, 72 and 96 hours. For each time point, cells were analysed by flow cytometry and 50.000 cells per genotype were pooled for nuclei isolation.
Single nuclei were isolated using the Nuclei Prep Buffer (Zymo Research), counted and processed following the Chromium Single Cell Multiome ATAC + Gene Expression workflow[105].

Aligning and QC

10x genomics multiome data were processed using Cell Ranger ARC (2.0.2) using default parameters and the provided reference genome GRCh38-2020-A-2.0.0. Initial filtering steps were applied to the raw gene expression and peak (ATAC) matrices of each sample using functions from the Seurat and Signac (v5) packages[106,107]. Cells with <500 gene features were first removed and the package scDBfinder[108] was used to mark doublets in both RNA and ATAC data. Further, cells with < 1000 peak features, >10% mitochondrial reads, and multiplets marked by cellranger were removed. At this stage, doublets were additionally identified in the ATAC data using Amulet[109]. Prior to merging samples, ATAC matrices were rebuilt to reflect unique fragment counts in 5kb genomic windows instead of peaks, using a custom pipeline[110], and RNA matrices were adjusted to remove ambient contamination using CellBender[111]. The merged Seurat object was then filtered to retain cells that could be unequivocally assigned to a lentiviral Larry barcode. To deconvolute samples based on the lentiviral barcodes, reads that did not map to the reference genome were extracted from the original cellranger genome + transcriptome alignments (gex_possorted_bam.bam) using samtools view -f 4, converted to fastq, and then re-mapped to a new reference containing the 11 possible GFP+barcodes transgenes. Multimappers were then removed (samtools view -q 5) to keep only the reads uniquely mapping to each GFP barcode. For each read, molecular barcode (UMI), cell barcode and Larry barcode were recorded. Only cell/Larry barcodes combinations supported by a minimum of 3 UMI were retained. Cell barcodes were assigned to a genotype only if one cell/Larry barcodes combination was found. The RNA and ATAC reductions from the merged and filtered Seurat object were then integrated using the Weighted Nearest Neighbour (WNN) analysis following the steps and recommended parameters in guidelines[112]. For clustering analysis on the integrated WNN graph, the Leiden algorithm with resolution 2 was used. At this stage, a low-quality cluster with lower read counts and higher mitochondrial content was removed, as well as doublets, which were previously marked as such by both Amulet and scDBfinder. After these final filters, we called ATAC peaks on the remaining cells with MACS2 (2.2.9.1) using the Signac callPeaks function and generated a new feature ATAC matrix with these peak coordinates, which replaced the window matrix. WNN analysis was then repeated as above, but with 0.5 resolution for the final clusters. The final object was composed of 13 clusters, 11.250 cells, ~29.000 genes and ~357.000 peaks. Processing and analysis codes can be found in Github[113].

DE and DA analysis

The code and the detailed list of dependencies can be found in GitHub[114]. Briefly, for the DE

analysis we aggregated the counts of all the cells belonging to the same combination of clone line, time point and Seurat cluster using the function Seurat2PB (Seurat v5.01[106]). Then, per Seurat cluster, we used DESeq2[115] to calculate a linear model accounting for time (reduced model) or for time and genotype (complete model) and applied a Likelihood ratio test (LRT) to test cases in which the complete model was a significantly better fit to the data following analysis guidelines for time and condition differential analysis[116,117].

For the ORA analysis we started by defining a list of gene sets relevant for K-562 differentiation from the Molecular Signatures Database (MSigDB)[91]: we selected all the gene sets that contained in their description at least one of the following terms: PLATELET, ERYTHROCYTE, CUX1, MEGAKARYOCYTE, GATA1, GATA2, TET2, RUNX1, RUNX2, MITOSIS, ANEUPLOIDY, CYTOKINESIS, MYELOID, AML, LIPID, SPHINGOSINE, FOXM1, SPI1, PU1, PI3K, AKT, FOXP1 and GFI1. We also included two blood and immunity unrelated terms (HEPATOCYTE and NEURON). In addition, we included in the ORA analysis TF target gene sets for the TFs *GATA2, GFI1, CUX1* and *RUNX1* in the Dorothea collection[37] (A,B ,C and D confidence levels). The minimum and maximum gene set size allowed was 10 and 500, respectively, but for the Dorothea gene sets which were exempted from this filter (*CUX1* n=464 target genes, *FOXP1* n=3,278, *GATA2* n=5,370, *GFI1* n=8 and *RUNX1* n=2,551). The list of selected pathways (*n=803 gene sets*) can be found in the GitHub repository. We obtained 59 pathways significantly overrepresented in DE genes of which 2 were from the unrelated terms.

For the DA analysis we first extracted all the linked peaks to the set of DE genes and a set of marker genes (in total 1,897 genes) using the LinkPeaks function of Signac (v1.12.0)[107]. To characterise the linked peaks (n=14,211 peaks) we overlapped them with annotated gene TSS (+/- 2.5 kB of the TSS, n= 4,072 overlaps with a known TSS, ENSEMBL, release 111[118]) and with the annotated regulatory features of the basal state K-562 cells (ENSEMBL, release 111[119]). Next, we produced a reduced seurat object with the ATAC counts of the selected peaks using the function CreateChromatinAssay (Signac v1.12.0) and then obtained a new Seurat object with the assay option set to 'RNA' as the function Seurat2PB (Seurat v5.01) would not work if it was set to 'ATAC'. From this point onwards the pipeline proceeded as explained in the DE analysis. To clusterize the results of both analyses we used Pheatmap[120] and to display them in volcano plots and in detailed locus plots we used ggplot2[121].

Data accessibility

The INTERVAL study data used in this paper are available to bona fide researchers from ceu-dataaccess@medschl.cam.ac.uk. The data access policy for the data is available at http://www.donorhealth-btru.nihr.ac.uk/project/bioresource. The RNA-seq data in the INTERVAL cohort have been deposited at the European Genome-phenome Archive (EGA) under the accession number EGAD00001008015 and are available at https://IntervalRNA.org.uk. The UK Biobank genetic data used in this study were approved under application 82779 and are available to qualified researchers via the UK Biobank data access process.

Bibliography

1. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
2. Dwivedi, O. P. *et al.* Loss of ZnT8 function protects against diabetes by enhanced insulin secretion. *Nat. Genet.* **51**, 1596–1606 (2019).
3. Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat. Genet.* **37**, 161–165 (2005).
4. Spielmann, M. & Mundlos, S. Looking beyond the genes: the role of non-coding variants in human disease. *Hum. Mol. Genet.* **25**, R157–R165 (2016).
5. Turro, E. *et al.* Whole-genome sequencing of patients with rare diseases in a national health system. *Nature* **583**, 96–102 (2020).
6. Morris, J. A. *et al.* Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens. *Science* eadh7699 (2023).
7. Ribeiro, D. M. & Delaneau, O. Non-coding rare variant associations with blood traits on 166 740 UK Biobank genomes. *bioRxiv* 2023.12.01.569422 (2023) doi:10.1101/2023.12.01.569422.
8. Vuckovic, D. *et al.* The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell* **182**, 1214–1231.e11 (2020).
9. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
10. Caron, B., Luo, Y. & Rausell, A. NCBoost classifies pathogenic non-coding variants in Mendelian diseases through supervised learning on purifying selection signals in humans. *Genome Biol.* **20**, (2019).
11. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* (2023) doi:10.1038/s41586-023-06045-0.
12. Ulirsch, J. C. *et al.* Interrogation of human haematopoiesis at single-cell and single-

variant resolution. *Nat. Genet.* **51**, 683–693 (2019).

13. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398–1414.e24 (2016).
14. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369–1384.e19 (2016).
15. Thiecke, M. J., Yang, E. J., Burren, O. S., Ray-Jones, H. & Spivakov, M. Prioritisation of Candidate Genes Underpinning COVID-19 Host Genetic Traits Based on High-Resolution 3D Chromosomal Topology. *Front. Genet.* **12**, (2021).
16. Burren, O. S. *et al.* Chromosome contacts in activated T cells identify autoimmune disease candidate genes. *Genome Biol.* **18**, 165 (2017).
17. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
18. Muerdter, F. *et al.* Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods* **15**, 141–149 (2018).
19. Ulirsch, J. C. *et al.* Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **165**, 1530–1545 (2016).
20. Fulco, C. P. *et al.* Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016).
21. Ashuach, T. *et al.* MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol.* **20**, 183 (2019).
22. Vockley, C. M. *et al.* Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res.* (2015) doi:10.1101/gr.190090.115.
23. Maeß, M. B., Wittig, B. & Lorkowski, S. Highly efficient transfection of human THP-1 macrophages by nucleofection. *J. Vis. Exp.* e51960 (2014).

24. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
25. Kribelbauer, J. F. *et al.* Context transcription factors establish cooperative environments and mediate enhancer communication. *bioRxiv* 2023.05.05.539543 (2023) doi:10.1101/2023.05.05.539543.
26. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
27. Cooper, S. E., Schwartzentruber, J., Bello, E., Coomber, E. L. & Bassett, A. R. Screening for functional transcriptional and splicing regulatory variants with GenIE. *Nucleic Acids Res.* **48**, e131 (2020).
28. rgenie. https://github.com/Jeremy37/rgenie.
29. Tokolyi, A. *et al.* Genetic determinants of blood gene expression and splicing and their contribution to molecular phenotypes and health outcomes. *bioRxiv* (2023) doi:10.1101/2023.11.25.23299014.
30. Võsa, 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).
31. Love, M. I., Soneson, C. & Patro, R. Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. *F1000Res.* **7**, 952 (2018).
32. Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
33. Polfus, L. M. *et al.* Whole-Exome Sequencing Identifies Loci Associated with Blood Cell Traits and Reveals a Role for Alternative GFI1B Splice Variants in Human Hematopoiesis. *Am. J. Hum. Genet.* **99**, 785 (2016).
34. Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genomics* **2**, 100168

(2022).

35. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet.* **53**, 1527–1533 (2021).
36. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).
37. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.* **29**, 1363–1375 (2019).
38. An, N. *et al.* Gene dosage effect of CUX1 in a murine model disrupts HSC homeostasis and controls the severity and mortality of MDS. *Blood* **131**, 2682–2697 (2018).
39. Mazzi, S., Lordier, L., Debili, N., Raslova, H. & Vainchenker, W. Megakaryocyte and polyploidization. *Exp. Hematol.* **57**, 1–13 (2018).
40. Mazzi, S. *et al.* Dual role of EZH2 in megakaryocyte differentiation. *Blood* **138**, 1603–1614 (2021).
41. Gargano, M. A. *et al.* The Human Phenotype Ontology in 2024: phenotypes around the world. *Nucleic Acids Res.* **52**, D1333–D1346 (2024).
42. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
43. Wong, C. C. *et al.* Inactivating CUX1 mutations promote tumorigenesis. *Nat. Genet.* **46**, 33–38 (2014).
44. 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).
45. Pradhananga, S., Spalinskas, R., Poujade, F.-A., Eriksson, P. & Sahlén, P. Promoter anchored interaction landscape of THP-1 macrophages captures early immune

response processes. *Cell. Immunol.* **355**, 104148 (2020).

46. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, (2016).
47. VEP_most_severe. https://www.ensembl.org/info/docs/tools/vep/script/vep_options.html.
48. Chen, S. *et al.* A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. *bioRxiv* 2022.03.20.485034 (2022) doi:10.1101/2022.03.20.485034.
49. engreitzlab resources. https://www.engreitzlab.org/resources.
50. annotation_variants_extended_AbC. https://github.com/manueltar/annotation_variants_extended_AbC.
51. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–548.e24 (2019).
52. Lazic, S. E. Ranking, selecting, and prioritising genes with desirability functions. *PeerJ* **2015**, (2015).
53. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
54. Annotation_variants_pipeline. https://github.com/manueltar/Annotation_variants_pipeline.
55. 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).
56. ukbiobank_fbc. https://github.com/ariannalandini/ukbiobank_fbc/tree/master.
57. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
58. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
59. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics

identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–75, S1–3 (2012).

60. MPRA synthesis Library design and cloning Soranzo Lab. https://www.protocols.io/edit/mpra-synthesis-library-design-and-cloning-soranzo-cs3awgie.
61. MPRA synthesis Cellular work and Nucleofection Soranzo Lab. https://www.protocols.io/edit/mpra-synthesis-cellular-work-and-nucleofection-sor-cs3jwgkn.
62. MPRA synthesis DNA/RNA isolation and library preparation for sequencing Soranzo Lab. https://www.protocols.io/edit/mpra-synthesis-dna-rna-isolation-and-library-prepa-cs3mwgk6.
63. Klein, J. C. *et al.* A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods* (2020) doi:10.1038/s41592-020-0965-y.
64. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
65. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
66. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
67. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P. & Marth, G. T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691–1692 (2011).
68. MPRA_bc_synthesis_Sample_alignment_and_counts. https://github.com/manueltar/MPRA_bc_synthesis_Sample_alignment_and_counts.git.

69. MPRA_bc_synthesis_analysis. https://github.com/manueltar/MPRA_bc_synthesis_analysis/tree/main.
70. van Heeringen, S. J. & Veenstra, G. J. C. GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics* **27**, 270–271 (11 2010).
71. Heinz, S. et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576–589 (2010).
72. Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–D173 (2022).
73. TF_motif_scan. https://github.com/manueltar/TF_motif_scan.
74. Enformer_run. https://github.com/manueltar/Enformer_run/tree/main.
75. Bruntraeger, M., Byrne, M., Long, K. & Bassett, A. R. Editing the Genome of Human Induced Pluripotent Stem Cells Using CRISPR/Cas9 Ribonucleoprotein Complexes. *Methods Mol. Biol.* **1961**, 153–183 (2019).
76. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
77. samclip. https://github.com/tseemann/samclip.
78. genIE_analysis. https://github.com/manueltar/genIE_analysis/tree/main.
79. Vanderstichele, T. *et al.* Misexpression of inactive genes in whole blood is associated with nearby rare structural variants. *bioRxiv* 2023.11.17.567537 (2023) doi:10.1101/2023.11.17.567537.
80. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
81. Peer. https://github.com/PMBio/peer.

82. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
83. INTERVAL analysis code. https://github.com/manueltar/INTERVAL_ANALYSIS.
84. Martín-Fernández, J. A., Barceló-Vidal, C. & Pawlowsky-Glahn, V. Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. *Math. Geol.* **35**, 253–278 (2003).
85. Gustavsson, E. K., Zhang, D., Reynolds, R. H., Garcia-Ruiz, S. & Ryten, M. ggtranscript: an R package for the visualization and interpretation of transcript isoforms using ggplot2. *Bioinformatics* **38**, 3844–3846 (2022).
86. INTERVAL web portal. https://www.intervalrna.org.uk/.
87. Myers, T. A., Chanock, S. J. & Machiela, M. J. LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations. *Front. Genet.* **11**, 157 (2020).
88. GTEx_sQTLs. https://github.com/manueltar/GTEx_sQTLs/tree/main.
89. INTERVAL_sQTLs. https://github.com/manueltar/INTERVAL_sQTLs.
90. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141 (2021).
91. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).
92. Active_Pathways_ORA. https://github.com/manueltar/Active_Pathways_ORA_with_Dorothea.
93. Paczkowska, M. *et al.* Integrative pathway enrichment analysis of multivariate omics data. *Nat. Commun.* **11**, 735 (2020).
94. BluePrint Analysis. https://github.com/manueltar/BluePrint_ANALYSIS.
95. Concordance-DE-MPRA-analysis. https://github.com/manueltar/Concordance-DE-MPRA-analysis/tree/main.

96. Concordet, J.-P. & Haeussler, M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.* **46**, W242–W245 (2018).
97. Shy, B. R. *et al.* High-yield genome engineering in primary cells using a hybrid ssDNA repair template and small-molecule cocktails. *Nat. Biotechnol.* (2022) doi:10.1038/s41587-022-01418-8.
98. Clement, K. *et al.* CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat. Biotechnol.* **37**, 224–226 (2019).
99. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, (2020).
100. van den Boogaart, K. G. & Tolosana-Delgado, R. *Analyzing Compositional Data with R.* (Springer Science & Business Media, 2013).
101. Compositions R package. http://www.stat.boogaart.de/compositions/.
102. Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. & Barceló-Vidal, C. Isometric Logratio Transformations for Compositional Data Analysis. *Math. Geol.* **35**, 279–300 (2003).
103. ILR-compositional-analysis-of-Flow-cytometry-data. https://github.com/manueltar/ILR-compositional-analysis-of-Flow-cytometry-data.
104. FSC_A_SSC_A_analysis. https://github.com/manueltar/FSC_A_SSC_A_analysis/tree/main.
105. single-cell-multiome-atac-plus-gene-expression. https://www.10xgenomics.com/products/single-cell-multiome-atac-plus-gene-expression.
106. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2024).
107. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).

108. Germain, P.-L., Lun, A., Garcia Meixide, C., Macnair, W. & Robinson, M. D. Doublet identification in single-cell sequencing data using scDblFinder. *F1000Res.* **10**, 979 (2021).

109. Thibodeau, A. *et al.* AMULET: a novel read count-based method for effective multiplet detection from single nucleus ATAC-seq data. *Genome Biol.* **22**, 252 (2021).

110. Benaglio, P. *et al.* Mapping genetic effects on cell type-specific chromatin accessibility and annotating complex immune trait variants using single nucleus ATAC-seq in peripheral blood. *PLoS Genet.* **19**, e1010759 (2023).

111. Fleming, S. J. *et al.* Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nat. Methods* **20**, 1323–1335 (2023).

112. weighted_nearest_neighbor_analysis. https://satijalab.org/seurat/articles/weighted_nearest_neighbor_analysis.

113. Benaglio, P. *Tardaguila_etal*. (Github).

114. Multiome_downstream_analysis. *github* https://github.com/manueltar/Multiome_downstream_analysis/tree/main.

115. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

116. Single-cell RNA-seq: Pseudobulk differential expression analysis. https://hbctraining.github.io/scRNA-seq_online/lessons/pseudobulk_DESeq2_scrnaseq.html.

117. Introduction to DGE - ARCHIVED. https://hbctraining.github.io/DGE_workshop/lessons/08_DGE_LRT.html.

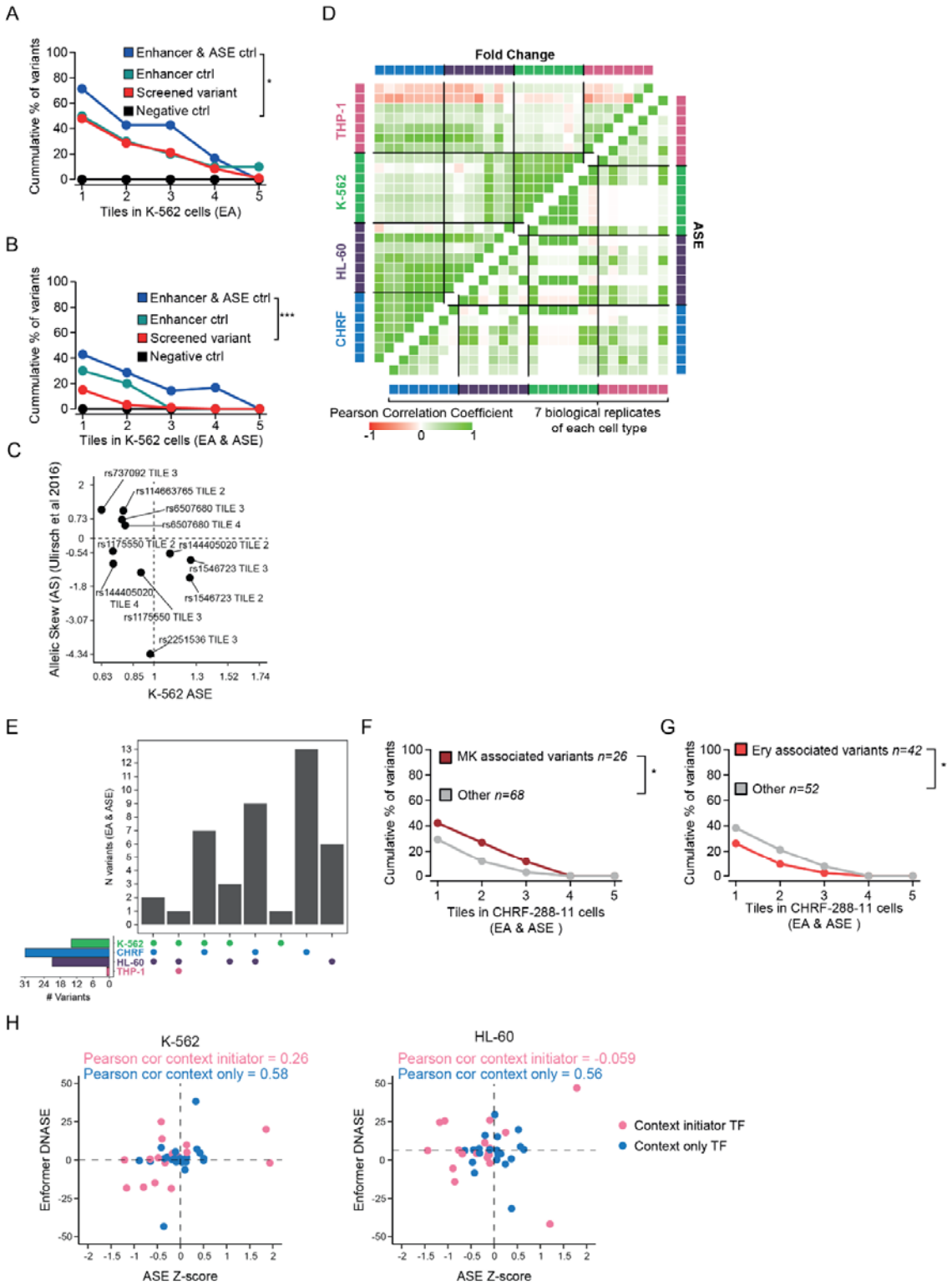
118. Harrison, P. W. *et al.* Ensembl 2024. *Nucleic Acids Res.* **52**, D891–D899 (2024).

119. RegulatoryFeatureActivity_K562. https://ftp.ensembl.org/pub/release-111/regulation/homo_sapiens/RegulatoryFeatureActivity/K562/.

120. Kolde, R. & Kolde, M. R. Package 'pheatmap'. *R package* **1**, 790 (2015).

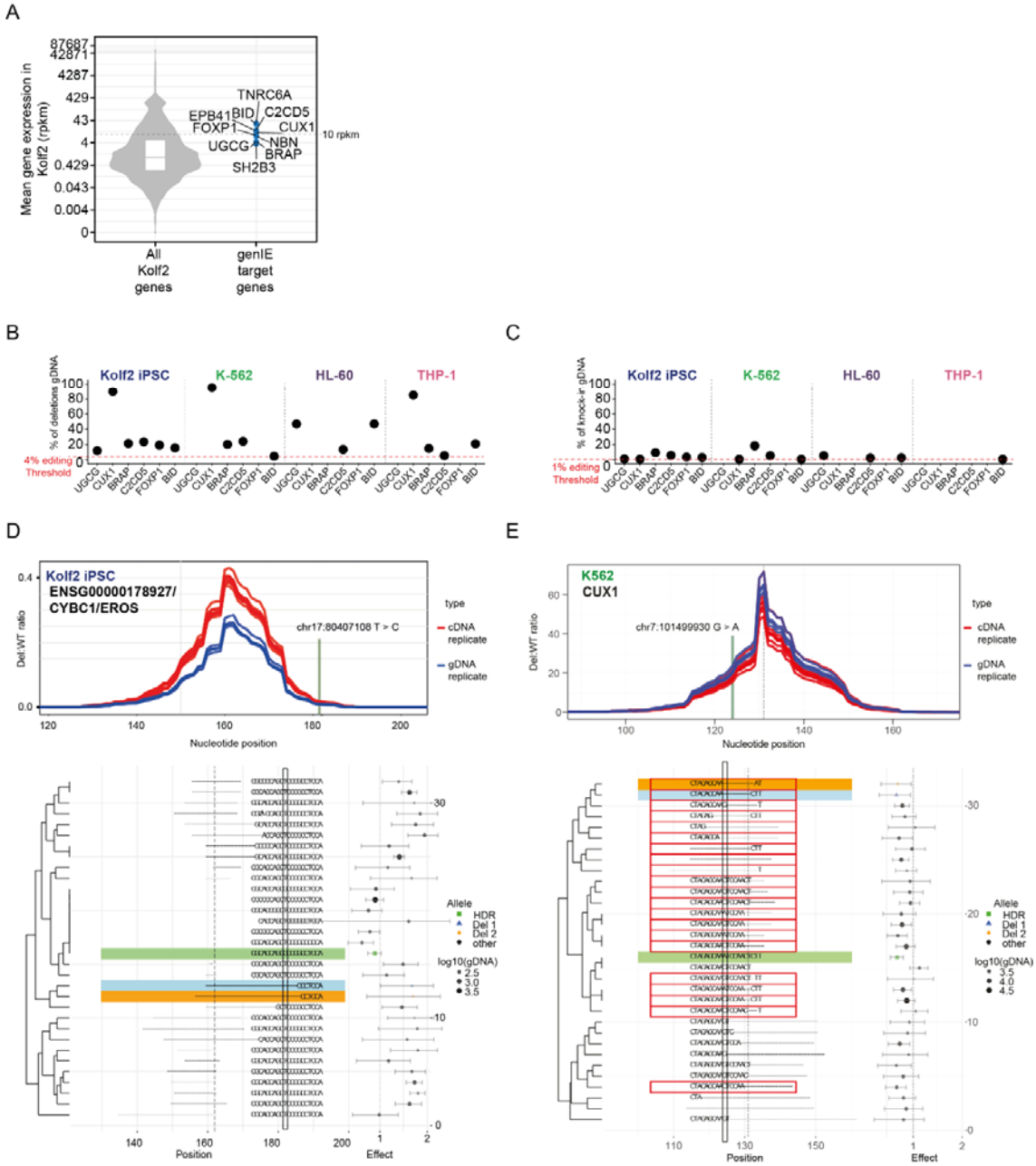
121. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer Science & Business Media, 2009).

Extended data



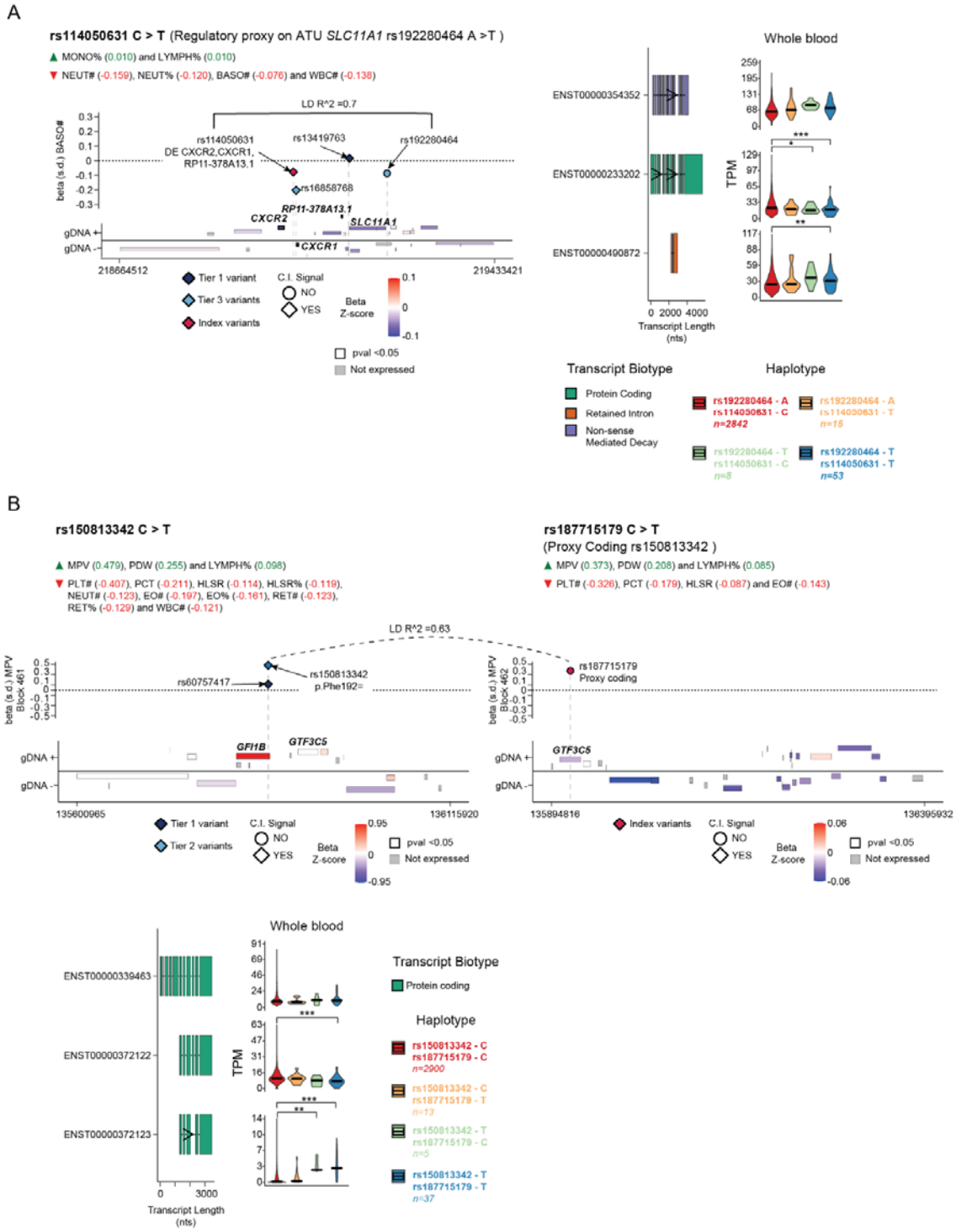
Extended Figure 1 related to Figure 2 and Figure 3. MPRA controls, selectivity and Enformer run. A) and **B)** Cumulative curves for the MPRA EA and EA +ASE in the K-562 cell

line distinguishing Enhancer control, Enhancer + ASE control and Negative control variants from Screened variants. **C)** Comparison of ASE directionality in the set of overlapping variants between our MPRA in K-562 cells and a previous MPRA in the same cell type[19]. **D)** Correlation plot for the MPRA Fold Change and ASE per tile for the seven biological replicates of each cell type. **E)** Sharing of variants with at least one tile with EA + ASE across the four cell lines in the MPRA. **F)** and **G)** The 94 variants screened in the MPRA variants were classified as affecting the Erythroid lineage, Megakaryocyte lineage, Granulocyte-Monocyte lineage, Lymphocyte lineage or impacting white blood cell counts following the criteria in Table S3. We then tested each lineage class against the variants that are not associated to the specified lineage in every cell type. We display the cumulative curves in the megakaryocytic leukaemia cell line CHRF that show significant depletion of MPRA activity for variants associated to the erythroid lineage traits and enrichment of activity for variants associated to the megakaryocyte lineage traits. **H)** The 94 variants screened in the MPRA were scanned for overlapping TF motifs and classified between 'context-initiator' and 'context-only' TFs[25] (**Methods**). 51/94 variants had TF motifs belonging to either of the two categories. We used Enformer[26] to predict the effect of the variants on the DNAse accessibility in K-562 and HL-60 cell lines and plotted these values against the Z-score values for ASE obtained in the same cell line. Enformer DNAse values correlate better with ASE values in the case of variants overlapping context-only TF motifs in both cell lines. *For the abbreviations of blood phenotypes see Table S3, Pearson cor = Pearson Correlation coefficient, * p <= 0.05, *** p <= 0.0001, Log-Rank test.*



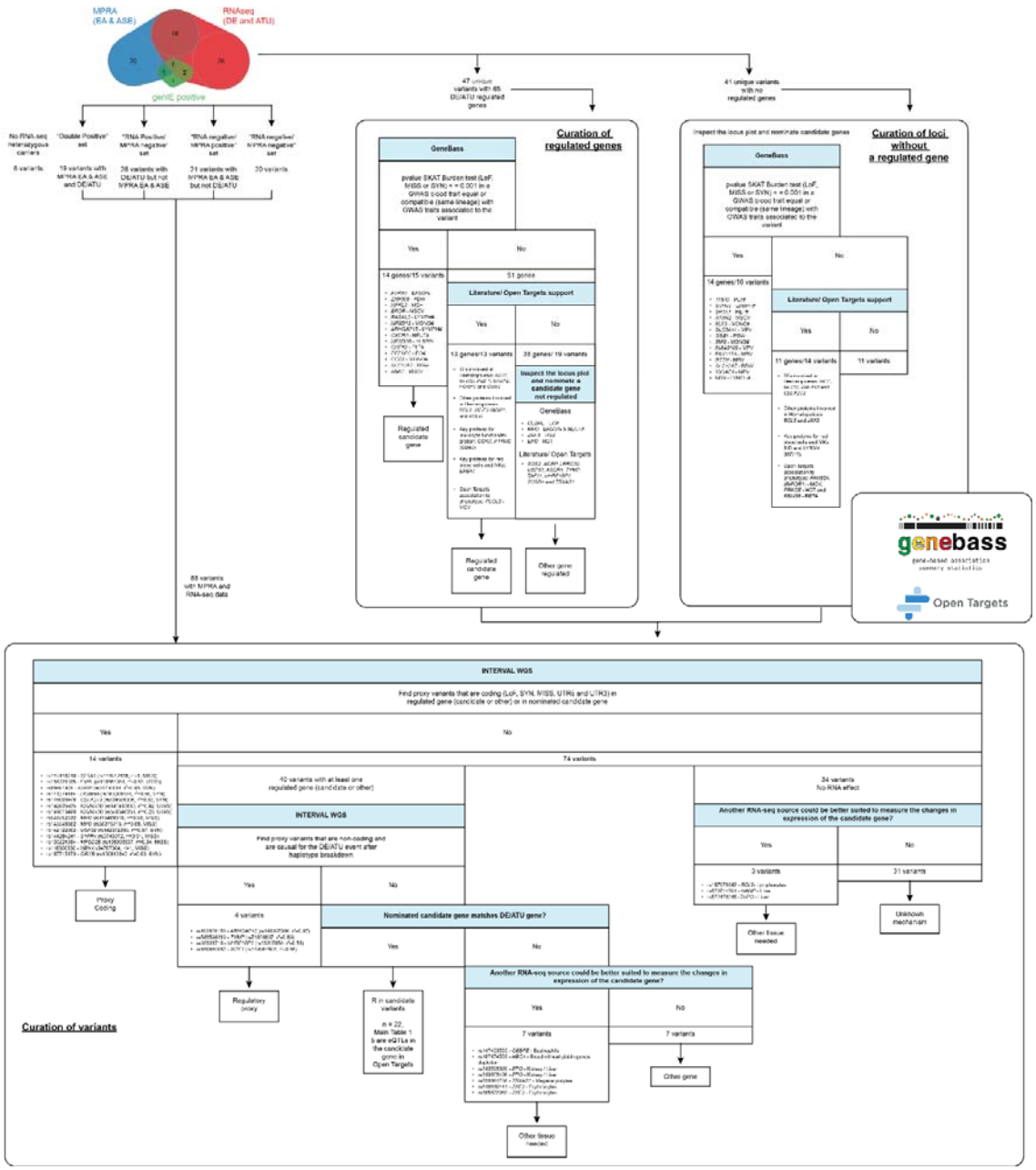
Extended Figure 2 related to Figure 2. CRISPR/Cas9 editing rates and per locus results of the GenIE screening. A) Violin plot showing the basal gene expression in the Kolf2 iPSC cell line in RPKM. The expression of the genes that host the intronic variants selected for the GenIE screening is depicted separately. The recommended cutoff value of 10 rpkm is indicated by the dashed line. **B)** and **C)** CRISPR/Cas9 deletions and knock-in mutation rates expressed as % of total reads sequenced across the different cell types assayed. The red dashed lines indicate the minimum thresholds applied for successful editing consideration. rs118026302-NBN, rs115820925-EPB41 and rs181042823-TNRC6A fell below the thresholds in all cell types and are not shown. *UDPs = Unique Deletion Profiles*, *CHRF = CHRF*. **D)** Profile of edited reads relative to *wildtype* reads vs UDPs and *wildtype* reads vs *knock-in*

reads for the locus EROS (positive control of the technique in the Kolf2 iPSC cell line) and **E)** the locus *rs139141690-CUX1*.

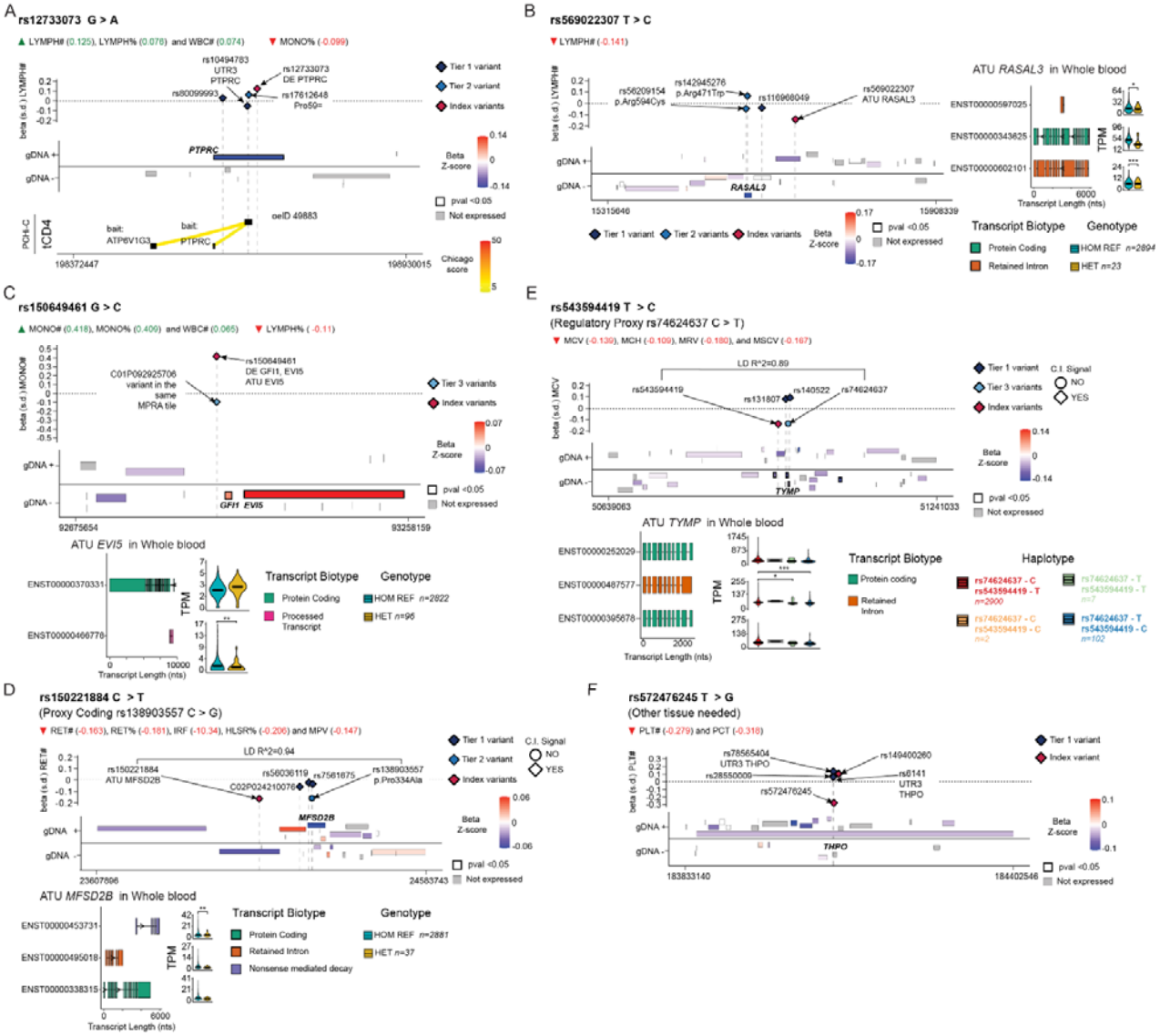


Extended Figure 3 related to Figure 3. Variants tagging coding and non-coding proxies that

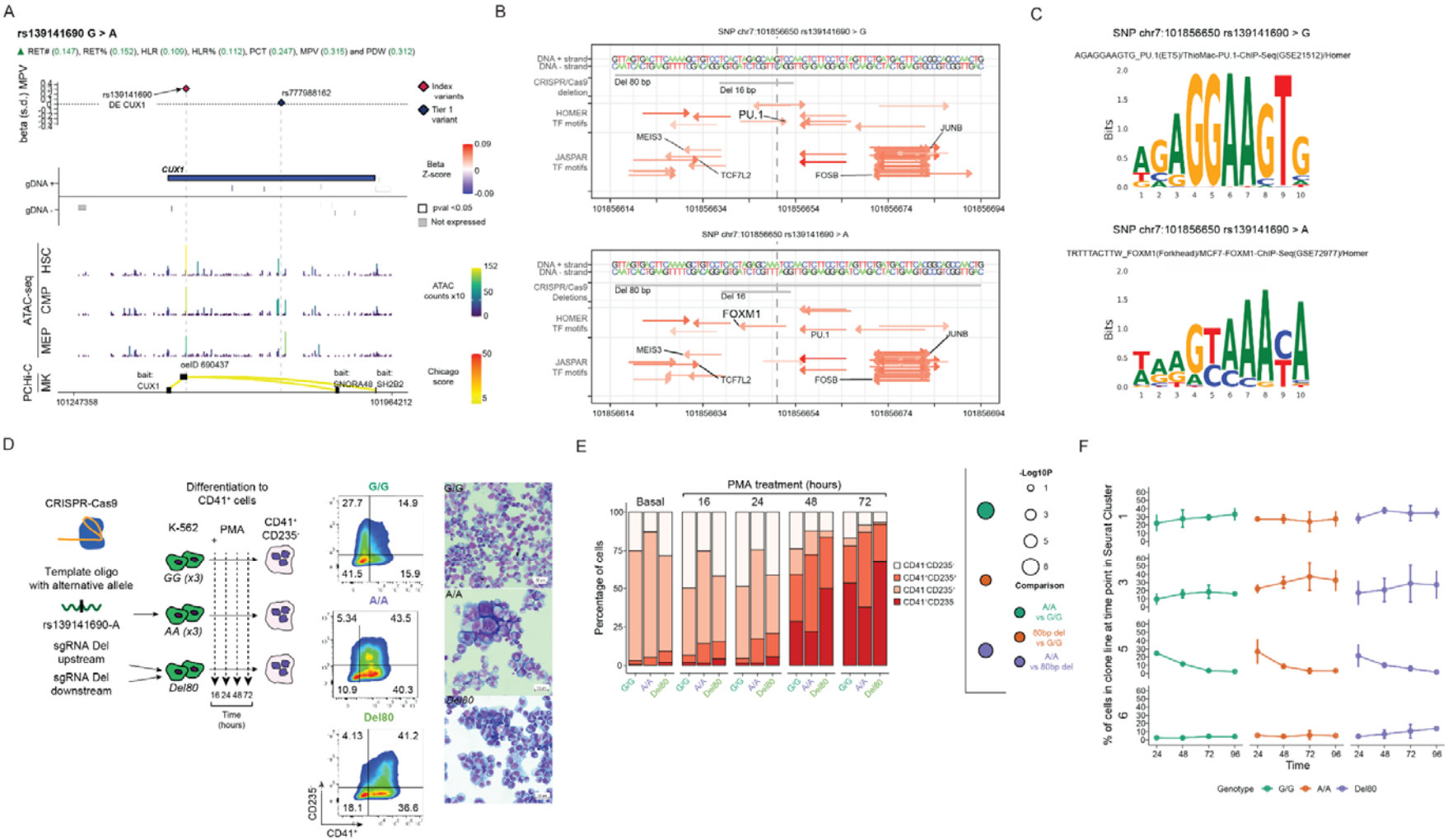
cause ATU. A) Locus plot for the variants rs114050631 and rs192280464. rs114050631 regulates the candidate gene *CXCR2* but also tags the non-coding proxy rs192280464 which in turn causes the ATU in the *SLC11A1* gene. The ATU analysis revealed that the haplotypes rs192280464-T-rs114050631-C and rs192280464-T-rs114050631-T entail an ATU event that results in a lower contribution of a protein coding transcript to the pool of transcripts of *SLC11A1*. **B)** Locus plots for the variants rs150813342 and rs187715179. rs187715179 tags the coding proxy rs150813342 which is a synonymous mutation in the *GFI1B* transcription factor. The ATU analysis revealed that the haplotypes rs150813342-T-rs187715179-C and rs150813342-T-rs187715179-T entail an ATU event that results in a lower contribution of a protein coding transcript (short protein) to the pool of transcripts of *GFI1B. For the abbreviations of blood phenotypes see Table S3 and Table S2 for the breakdown of variant classification in tiers.* p value <= 0.05, ** p value <= 0.01, *** p value <= 0.001, ATU additive log ratio model.*



Extended Figure 4 related to Figure 3. Flowchart of Manual curation.


A
rs12733073 G > A
▲ LYMPH# (0.125), LYMPH% (0.078) and WBC# (0.074)
▼ MONO% (-0.099)
rs10494783 UTR3 PTPRC
rs12733073 DE PTPRC
rs80099993
rs17612648 Pro50=
PTPRC
oeID 49883
bait: ATP6V1G3
bait: PTPRC
198372447
198930015
Tier 1 variant
Tier 2 variant
Index variants
Beta Z-score
pval <0.05
Not expressed
Chicago score
B
rs569022307 T > C
▼ LYMPH# (-0.141)
rs56209154 p.Arg594Cys
rs142945276 p.Arg471Trp
rs116968049
rs569022307 ATU RASAL3
RASAL3
15315646
15908339
ATU RASAL3 in Whole blood
ENST00000597025
ENST00000343625
ENST00000602101
Transcript Length (nts)
Transcript Biotype
Protein Coding
Retained Intron
Genotype
HOM REF n=2894
HET n=23
C
rs150649461 G > C
▲ MONO# (0.418), MONO% (0.409) and WBC# (0.065)
▼ LYMPH% (-0.11)
C01P092925706 variant in the same MPRA tile
rs150649461 DE GFI1, EVI5 ATU EVI5
GFI1 EVI5
92675654
93258159
Tier 3 variants
Index variants
ATU EVI5 in Whole blood
ENST00000370331
ENST00000466776
Transcript Length (nts)
Transcript Biotype
Protein Coding
Processed Transcript
Genotype
HOM REF n=2822
HET n=96
D
rs150221884 C > T
(Proxy Coding rs138903557 C > G)
▼ RET# (-0.163), RET% (-0.181), IRF (-10.34), HLSR% (-0.206) and MPV (-0.147)
LD R^2=0.94
rs150221884 ATU MFSD2B
rs56036119
rs7561675
rs138903557 p.Pro334Ala
C02P024210076
MFSD2B
23607806
24583743
C.I. Signal
NO
YES
ATU MFSD2B in Whole blood
ENST00000453731
ENST00000495018
ENST00000338315
Transcript Length (nts)
Transcript Biotype
Protein Coding
Retained Intron
Nonsense mediated decay
Genotype
HOM REF n=2881
HET n=37
E
rs543594419 T > C
(Regulatory Proxy rs74624637 C > T)
▼ MCV (-0.139), MCH (-0.109), MRV (-0.180), and MSCV (-0.167)
LD R^2=0.89
rs543594419
rs131807
rs140522
rs74624637
TYMP
50639063
51241033
ATU TYMP in Whole blood
ENST00000252029
ENST00000487577
ENST00000395678
Transcript Length (nts)
Transcript Biotype
Protein coding
Retained Intron
Haplotype
rs74624637 - C rs543594419 - T n=2900
rs74624637 - C rs543594419 - C n=2
rs74624637 - T rs543594419 - T n=7
rs74624637 - T rs543594419 - C n=102
F
rs572476245 T > G
(Other tissue needed)
▼ PLT# (-0.279) and PCT (-0.318)
rs78565404 UTR3 THPO
rs28550009
rs572476245
rs149400260
rs6141 UTR3 THPO
THPO
183833140
184402546

Extended Figure 5 related to Figure 3. Manual curation excludes variants tagging coding and regulatory proxies. A) Locus plot for the variants impacting the gene *PTPRC* (CD45) and the GWAS trait LYMPH#. Two of the variants (rs17612648 and rs10494783) are coding (CADD values 0.26-3.8) affecting the *PTPRC* gene. We observed a significant downregulation of *PTPRC* for the rs12733073-A allele. **B)** Locus plot for the variants impacting the gene *RASAL3* and the GWAS trait LYMPH#. Two of the variants (rs56209154 and rs142945276, Tier 2, CADD values 23.5-29.3) are coding variants affecting the *RASAL3* gene. The ATU analysis revealed that rs569022307-C entails an ATU event that results in a higher contribution of two retained intron transcripts to the pool of transcripts of *RASAL3*. **C)** Locus plot for the variant rs150649461. The conditionally independent variant C01P092925706 (chr1:92925706 ATTAGAG > A, GRCh37) overlapped some of the rs150649461 tiles in the MPRA. For these tiles we assayed all the single alternative allele and double alternative allele versions and found no tiles with concomitant EA and ASE (see **Table S4**). The *'Index'* variant rs150649461-C allele led to a significant upregulation of *GFI1* and *EVI5* and an ATU event that resulted in a lower contribution of a processed transcript to the pool of transcripts of *EVI5*. **D)** Locus plot for the variant rs150221884 which tags a coding proxy, rs138903557 (CADD value 24.7, Tier 2), in the candidate gene *MFSD2B*. The ATU analysis revealed that rs150221884-T entails an ATU event that results in a higher contribution of a nonsense mediated transcript to the pool of transcripts of *MFSD2B*. **E)** Locus plot for the variant rs543594419 which tags the regulatory proxy rs74624637 (Tier 3) for the gene *TYMP*. The ATU analysis revealed that the haplotypes rs543594419-T-rs74624637-T and rs543594419-C-rs74624637-T entail an ATU event that results in a lower contribution of a retained intron transcript to the pool of transcripts of *TYMP*. **F)** Locus plot for the variant rs572476245 for which manual curation indicates that other tissue is needed to ascertain its effect on the THPO gene (Thrombopoietin, hormone produced in the liver and the kidney and not expressed by blood cell types). Two of the variants (rs78565404 and rs6141, both Tier 1, CADD values 3.19 and 2.42) are coding variants affecting the THPO gene. See *Table S2 for the breakdown of variant classification in tiers and Table S3 for the abbreviations of blood phenotypes. C.I. = Conditionally Independent, DE = Differential Gene Expression, ATU = Differential Transcript Usage, UTR3 = variant in the 3' UTR region. LD = linkage disequilibrium. * p value <= 0.05, ** p value <= 0.01, *** p value <= 0.001, ATU additive log ratio model.*



Extended Figure 6 related to Figure 6. rs139141690-A downregulates the hematopoietic TF CUX1 leading to gene expression changes in the hematopoietic stem cell homeostasis and megakaryocyte differentiation that are recapitulated *'in vitro'*. A) Locus plot for the variant

rs139141690 with DE in the candidate gene *CUX1* in whole blood. The variant sits in an intron of *CUX1* and has PCHi-C and accessible chromatin evidence in the megakaryocytic and erythroid lineages. The variant was MPRA negative but GenIE positive (**Table S6**) and rs139141690-A led to a significant downregulation of *CUX1 in vivo* accompanied by significant enrichment of the *HSC homeostasis* and *Blood coagulation* pathways (**data not shown**). **B)** Prediction of TF motifs in the sequence stretch spanned by the 80 bp deletion for the rs139141690-G and rs139141690-A alleles. The dashed line represents the genomic coordinate for rs139141690. The point mutation rs139141690 > A elicits a change of the 'context-initiator' PU.1 TF motif to a FOXM1 motif. **C)** Position weight matrices for the rs139141690-G and rs139141690-A alleles. **D)** Design of the CRISPR/Cas9 approach to edit K-562 cells and obtain the rs139141690 'A/A' clone lines and the lines carrying an 80 bp deletion that spans the SNP and is based on the GenIE UDPs. Also shown is the time course differentiation of the K-562 cells to $CD41^+$ cells using PMA monitored by surface expression of CD235 and CD41 and change in cell morphology. **E)** Stacked barplot of the different populations in the PMA differentiation process. The percentages represent the mean of the three clones per genotype. On the left the -log p values for the genotype comparisons of cell type abundance in the ilr model (see **Methods**). **F)** Variability in the abundance of each seurat cluster per genotype expressed as the percentage of the number of cells in a specific cluster for a given clone line at a time point against all the cells of that clone line at that time point. For G/G clone lines, the clusters 1, 3 and 6 increase with time of treatment whereas cluster 5 decreases. *TF = Transcription Factor. See Table S3 for all the cell types and GWAS traits abbreviations and Table S2 for the breakdown of variant classification in tiers.*