nature genetics

ARTICLES

https://doi.org/10.1038/s41588-020-00738-2

Check for updates

Common variants in signaling transcription-factor-binding sites drive phenotypic variability in red blood cell traits

Avik Choudhuri[1,2,24], Eirini Trompouki[2,3,4,24], Brian J. Abraham[5,6,24], Leandro M. Colli[7,8], Kian Hong Kock[9,10], William Mallard[1,11], Min-Lee Yang[12], Divya S. Vinjamur[13], Alireza Ghamari[14], Audrey Sporrij[1], Karen Hoi[1], Barbara Hummel[3], Sonja Boatman[2], Victoria Chan[1], Sierra Tseng[1], Satish K. Nandakumar[13], Song Yang[2], Asher Lichtig[2], Michael Superdock[2], Seraj N. Grimes[9,15], Teresa V. Bowman[2,16], Yi Zhou[2], Shinichiro Takahashi[17], Roby Joehanes[18,19], Alan B. Cantor[14], Daniel E. Bauer[13], Santhi K. Ganesh[12], John Rinn[1,20], Paul S. Albert[7], Martha L. Bulyk[9,10,11,15,21], Stephen J. Chanock[7], Richard A. Young[5,22] and Leonard I. Zon[1,23] ✉

Genome-wide association studies identify genomic variants associated with human traits and diseases. Most trait-associated variants are located within cell-type-specific enhancers, but the molecular mechanisms governing phenotypic variation are less well understood. Here, we show that many enhancer variants associated with red blood cell (RBC) traits map to enhancers that are co-bound by lineage-specific master transcription factors (MTFs) and signaling transcription factors (STFs) responsive to extracellular signals. The majority of enhancer variants reside on STF and not MTF motifs, perturbing DNA binding by various STFs (BMP/TGF-β-directed SMADs or WNT-induced TCFs) and affecting target gene expression. Analyses of engineered human blood cells and expression quantitative trait loci verify that disrupted STF binding leads to altered gene expression. Our results propose that the majority of the RBC-trait-associated variants that reside on transcription-factor-binding sequences fall in STF target sequences, suggesting that the phenotypic variation of RBC traits could stem from altered responsiveness to extracellular stimuli.

A substantial fraction of worldwide mortality is attributed to erythrocyte-related disorders[1–7]. Variation in RBC traits is linked to mortality rates not related to primary hematologic disease[1,3,6]. Genome-wide association studies (GWAS) have identified numerous variable genomic regions associated with human traits and diseases, including RBC traits[8–21]. RBC-trait-associated single-nucleotide polymorphisms (SNPs) rarely affect DNA binding of MTFs, such as GATA2 and GATA1, even though they are often in close proximity to MTF target sequences[15,22,23]. Additional mechanisms by which RBC SNPs result in the phenotypic variability of human genetic traits remain to be discovered.

Environmental factors contribute to the phenotypic manifestation of complex human genetic traits[1,3,6]. Under stress conditions, growth factors and small molecules activate signaling pathways[24–26] that converge on signal-induced effector transcription factors (STFs) to control gene expression. By coordinating with MTFs, the same STFs may be active in multiple cell types but exert tissue-specific functions[27,28]. Hence, alterations in STF target sequences may lead to aberrant responses to various signals.

Here, we observed that human erythroid-trait-associated non-coding SNPs are enriched in a small subset of enhancers co-bound by MTFs and STFs, which we named transcriptional sig-

[1]Harvard Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA. [2]Stem Cell Program and Division of Hematology/Oncology, Boston Children's Hospital, Boston, MA, USA. [3]Department of Cellular and Molecular Immunology, Max Planck Institute of Immunobiology and Epigenetics, Freiburg, Germany. [4]CIBSS Centre for Integrative Biological Signaling Studies, University of Freiburg, Freiburg, Germany. [5]Whitehead Institute for Biomedical Research, Cambridge, MA, USA. [6]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN, USA. [7]Division of Cancer Epidemiology & Genetics, National Cancer Institute, Bethesda, MD, USA. [8]Department of Medical Imaging, Hematology, and Medical Oncology, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil. [9]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [10]Program in Biological and Biomedical Sciences, Harvard University, Cambridge, MA, USA. [11]The Broad Institute of the Massachusetts Institute of Technology and Harvard, Cambridge, MA, USA. [12]Division of Cardiovascular Medicine, Department of Internal Medicine and Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA. [13]Division of Hematology and Oncology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. [14]Division of Pediatric Hematology-Oncology, Boston Children's Hospital and Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. [15]Summer Institute in Biomedical Informatics, Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [16]Albert Einstein College of Medicine, Bronx, NY, USA. [17]Tohoku Medical and Pharmaceutical University, Sendai, Japan. [18]Hebrew Senior Life, Harvard Medical School, Boston, MA, USA. [19]Framingham Heart Study, National Heart, Blood, and Lung Institute, National Institutes of Health, Bethesda, MD, USA. [20]Department of Biochemistry, University of Colorado Boulder, Boulder, CO, USA. [21]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [22]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [23]Stem Cell Program and Division of Hematology/Oncology, Children's Hospital Boston, Harvard Stem Cell Institute, Harvard Medical School and Howard Hughes Medical Institute, Boston, MA, USA. [24]These authors contributed equally: Avik Choudhuri, Eirini Trompouki, Brian J. Abraham. ✉e-mail: zon@enders.tch.harvard.edu

naling centers (TSCs). Our study suggests that such SNPs alter the DNA binding of various STFs more frequently than that of blood MTFs, leading to gene expression changes induced by extracellular signaling and consequently impacting RBC phenotypes.

Results

MTFs and STFs control cell-type-specific gene expression. To understand how signaling impacts human erythropoiesis, we sought to identify genomic regions responsive to exogenous signals using in vitro erythroid differentiation of human hematopoietic progenitor cells ($CD34^+$; Extended Data Fig. 1a)[29]. By performing H3K27ac chromatin immunoprecipitation with sequencing (ChIP–seq) to identify active enhancers[30], assay for transposase-accessible chromatin using sequencing (ATAC–seq) to determine chromatin accessibility[31] and RNA sequencing (RNA-seq) to quantify gene expression in these cells at various stages of differentiation (day 0 (d0) before differentiation induction and 6 h, 3 days, 4 days and 5 days after induction of erythroid differentiation), we observed two expression clusters before and after d3, suggesting that $CD34^+$ cells commit to an erythroid fate around d3 in this system (Extended Data Fig. 1f; Supplementary Table 1 presents genome-wide RNA expression values). Thus, we considered genes that are expressed at high levels before d3 as progenitor genes and after d3 as erythroid genes.

Next, we investigated genomic occupancy of MTFs and STFs during erythroid differentiation. We chose GATA2 and GATA1 as exemplary progenitor and erythroid MTFs, respectively. To choose an STF, we tested the effect of BMP/SMAD signaling in our system, owing to its importance in stress erythropoiesis[28,32–35]. Induction of BMP signaling by recombinant BMP4 or abrogation by dorsomorphin affected the efficiency of erythroid commitment (Fig. 1a,b), so we chose SMAD1 as an exemplary erythropoietic STF.

During differentiation, genomic occupancy of GATA2, identified by ChIP–seq, steadily decreased and GATA1 occupancy progressively increased while SMAD1 gradually re-localized to new genomic sites (Fig. 1c). SMAD1 binding at progenitor stages (d0–d3) or erythroid stages (d3–d5) overlapped markedly with MTFs of the respective stages (Fig. 1c,d and Extended Data Fig. 1g). We then identified the GATA2 + SMAD1 co-occupied or GATA2-alone genomic sites at d0, hour 6 (h6) and d3 and the GATA1 + SMAD1 or GATA1-alone genomic regions at d3, d4 and d5 and assigned them to the predicted target genes (Supplementary Table 2). Notably, GATA-alone sites lack SMAD1 binding but possibly display binding of other MTFs besides GATA (refs. [36,37]). Ingenuity Pathway Analysis showed that genes co-bound by GATA1 + SMAD1 are enriched for erythroid functions, whereas genes co-bound by GATA2 + SMAD1 are enriched for progenitor functions (Extended Data Fig. 1h), indicating that GATA + SMAD1 co-bound regions regulate stage-specific genes. Next, by comparing expression between genes co-occupied by GATA + SMAD1 and genes occupied by GATA alone, we found that genes proximal to co-occupied regions showed significantly higher expression (Fig. 1e). Overlap of stage-matched ATAC–seq and ChIP–seq data demonstrated that co-bound regions exhibit enhanced chromatin accessibility compared to regions where GATA factors bind without SMAD1 (Fig. 1f). Additionally, inhibition of BMP signaling by dorsomorphin significantly decreased expression of erythroid genes such as *GLOBIN*, *ALAS* and *SLC4A1* that are co-bound by SMAD1 + GATA1 at d5 but not of genes proximal to regions where GATA1 binds alone (Fig. 1g).

SMAD1 + GATA regions are enriched for cell-type-specific MTFs. To investigate the features that distinguish co-bound from MTF-alone regions, we performed comparative motif analysis. This analysis showed over-representation of progenitor MTF sequence motifs (for example, PU.1 and FLI1 motifs[38,39]) in the GATA2 + SMAD1 regions at h6 relative to GATA2-alone regions, and erythroid factor motifs such as EKLF (also known as KLF1) and NFE motifs[40,41] in GATA1 + SMAD1 co-bound regions at the d5 erythrocyte stage relative to GATA1-alone regions (Extended Data Fig. 2a,b). Indeed, binding of PU.1 overlapped with GATA2 + SMAD1 co-bound regions at d0 while GATA1 + SMAD1 co-bound regions overlapped with KLF1 at d5. We observed at least 2.5-fold enrichment of PU.1 and KLF1 at co-occupied regions compared to the GATA-alone regions at d0 and d5, respectively (Fig. 2a,b and Supplementary Table 3a). Additionally, genomic regions where stage-specific MTFs co-localize with SMAD1 are proximal to stage-specific genes, are located in open chromatin regions and are enriched for H3K27ac (Figs. 1f and 2b and Extended Data Fig. 2c).

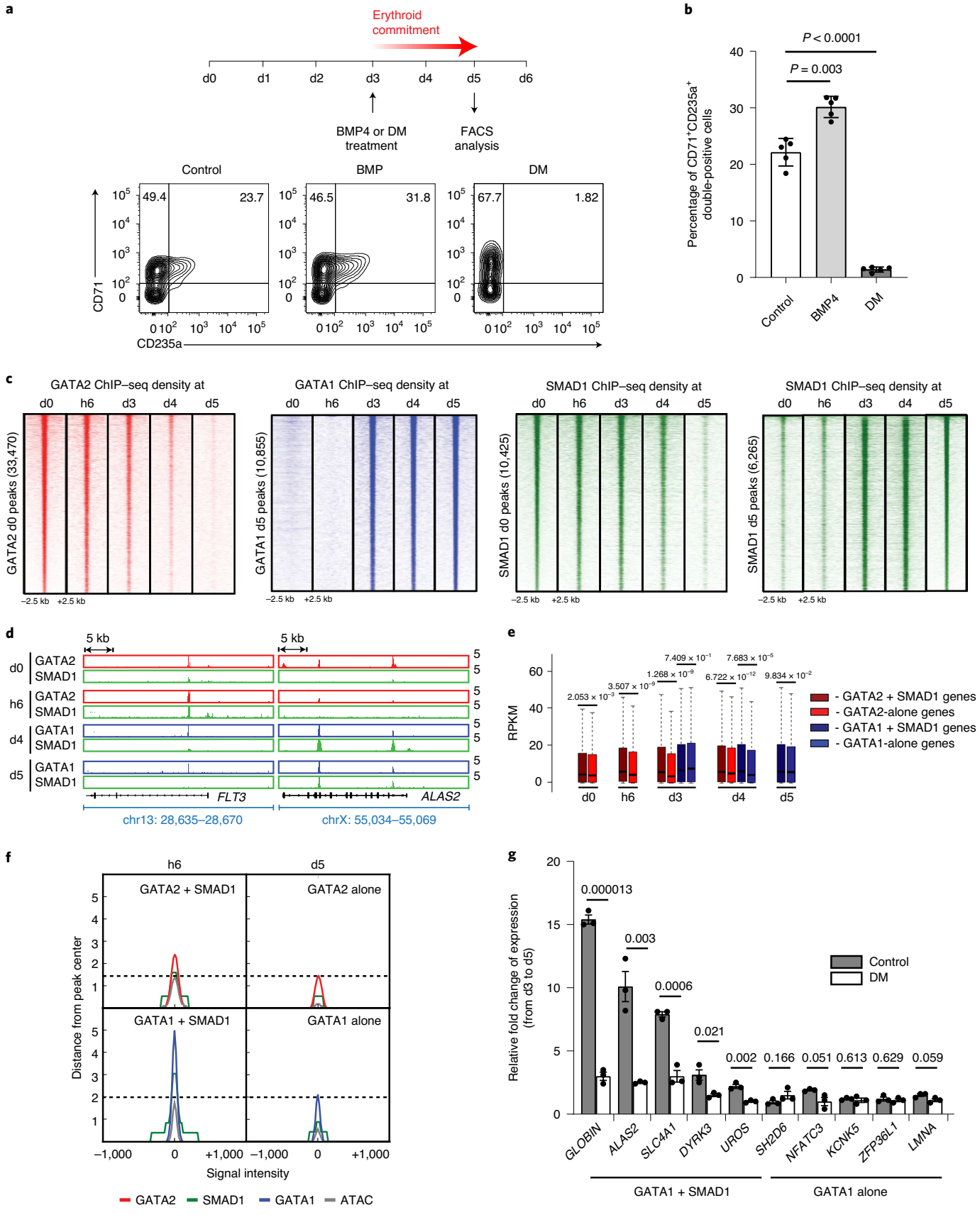
To examine the importance of binding of stage-specific MTFs within the SMAD1 + GATA co-bound regions, we investigated the change of SMAD1 binding on overexpression of PU.1 in K562 cells after BMP stimulation. PU.1-overexpressing cells showed increased binding of PU.1 in several genomic regions with a concomitant increase of SMAD1 binding within many of these regions, indicating that PU.1 can direct genomic localization of SMAD1 (Fig. 2c,d). We also confirmed that loss of PU.1 in K562 cells decreased PU.1 and SMAD1 occupancy within PU.1/SMAD1/GATA2 co-bound genomic regions while GATA2 binding did not diminish to the same extent (Fig. 2e). However, loss of PU.1 and SMAD1 binding could happen in the same or different cells. Overall, MTFs such as PU.1, enriched at GATA + SMAD1 sites, can recruit SMAD1 after stimulation to co-bound genomic regions, which likely behave as BMP-responsive enhancers.

TSCs. Next, we sought to determine whether SMAD1 + GATA co-bound regions could serve as docking sites for other STFs. We performed ChIP–seq for SMAD2 on TGF-β stimulation[42] and for TCF7L2 on WNT stimulation[28] at d0. Indeed, we observed co-localization of such STFs at GATA2-bound, ATAC–seq and H3K27ac signal-enriched enhancers, also co-occupied by SMAD1 on BMP stimulation (Fig. 3a,b). A total of 4,549 genomic regions, representing 25% of the total SMAD1-bound peaks, were co-occupied by SMAD1/2 and TCF7L2 (Extended Data Fig. 3a and

Fig. 1 | BMP/SMAD1 signaling impacts human erythroid differentiation. a, Representative FACS plots for CD71 and CD235a on BMP4- or dorsomorphin (DM)-treated $CD34^+$ cells. Numbers represent the percentages of cells in the respective quadrants. **b**, Bar plots comparing the percentage of $CD34^+CD235a^+$ double-positive cells from **a**. The mean ± s.e.m. is shown ($n = 5$; 5 biologically independent experiments). A two-sided Student's *t*-test was used. **c**, Regional heatmaps depicting the signal of the ChIP–seq reads for GATA2, GATA1 and SMAD1 at d0, h6, d3, d4 and d5 of differentiation. Signal intensities around ±2.5 kb of the peak center are shown. **d**, Representative gene tracks for a progenitor-specific gene (*FLT3*) and an erythroid-specific gene (*ALAS2*) showing binding of each TF at d0, h6, d4 and d5. **e**, Reads per kilobase of transcript per million mapped reads (RPKM) expression distribution of genes bound either by GATA + SMAD1 or by GATA alone at respective stages. The boxplots represent the median RPKM as the thickest line, the first and third quartiles as the box, and 1.5 times the interquartile range as whiskers. Two-sided Wilcoxon rank-sum tests were used. **f**, Metagene plots comparing the median signal intensities for ChIP–seq and ATAC–seq at regions co-bound by GATA2/1 + SMAD1 versus GATA2/1 alone. Signal intensities around ±1 kb of the peak center are shown. **g**, The change of expression of genes bound by GATA1 + SMAD1 (*HBB*, *ALAS2*, *SLC4A1*, *DYRK3* and *UROS*) or by GATA1 alone (*SH2D6*, *NFATC3*, *KCNK5*, *ZFP36L1* and *LMNA*) after continuous dorsomorphin treatment for two days starting from d3. The mean ± s.e.m. is shown ($n = 3$; 3 biologically independent experiments). A two-sided Student's *t*-test was used.

Supplementary Table 3b). We reasoned that enhancers where combinations of STFs would converge with hematopoietic MTFs after induction by environmental stimuli are likely signal responsive, and named them transcriptional signaling centers (TSCs; Fig. 3c).

While other STFs besides SMAD1 could define classes of TSCs, given the importance of BMP/SMAD1 signaling during stress hematopoiesis[32–35], we focused on SMAD1-bound TSCs. Genomic Regions Enrichment of Annotations Tool (GREAT) analysis[43] of



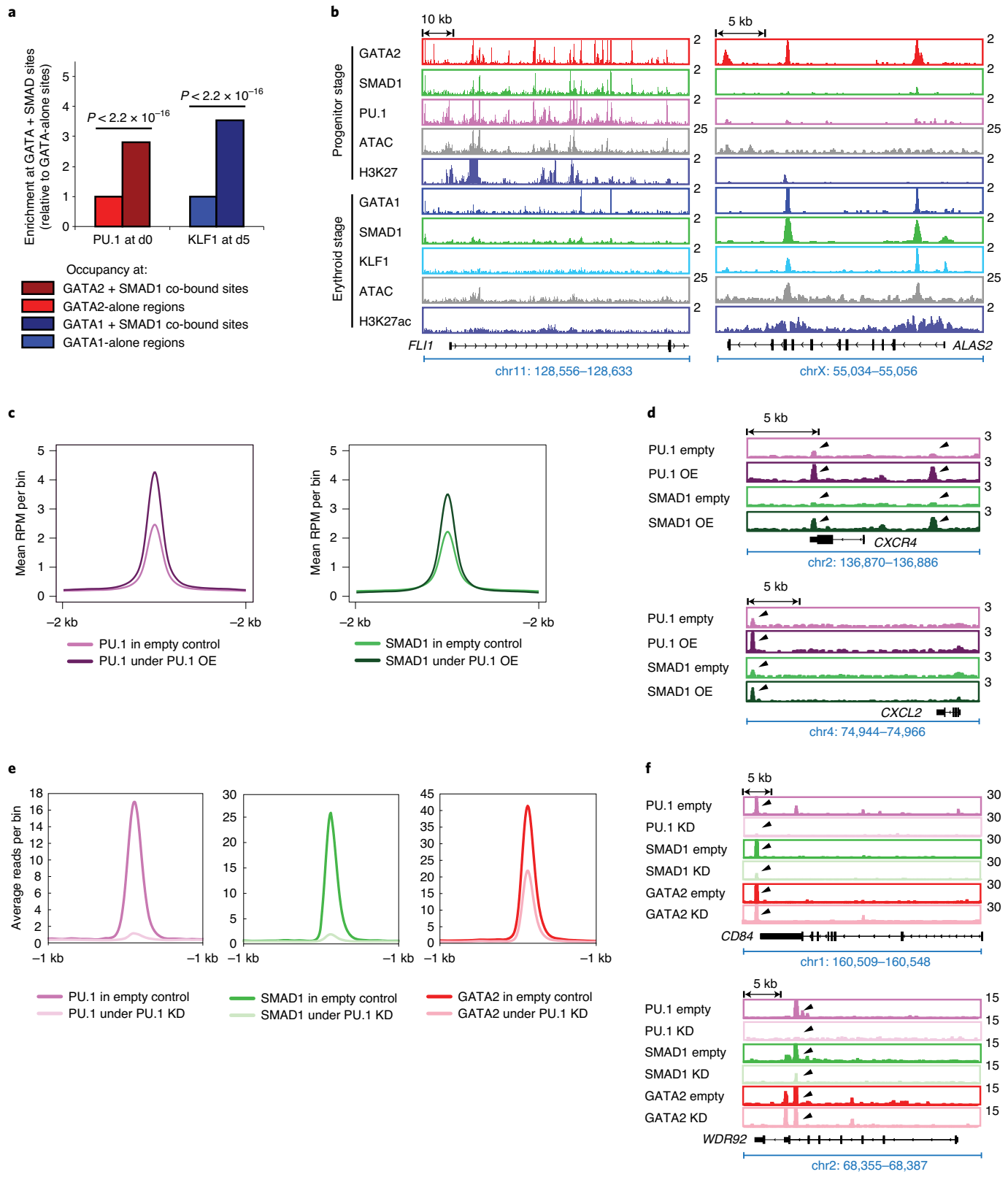
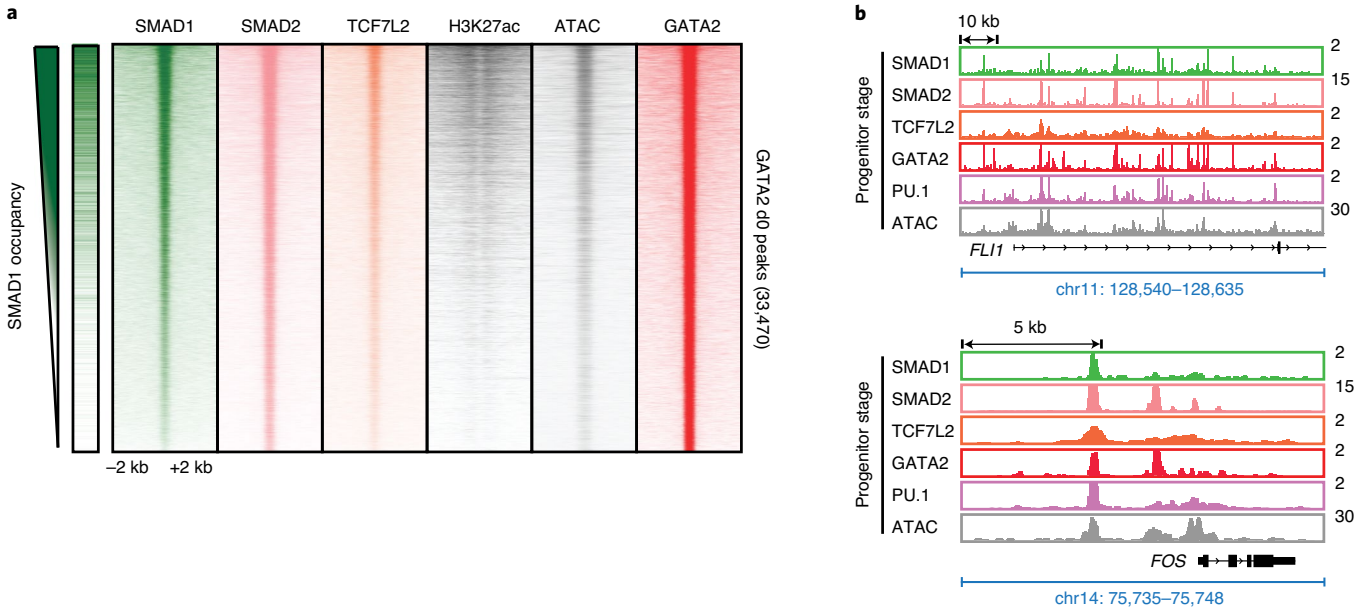
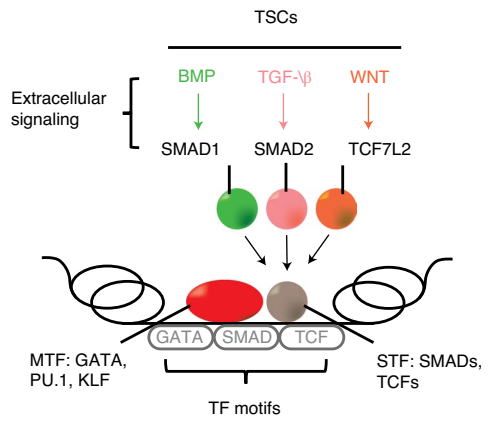


Fig. 2 | Stage-specific MTFs are enriched in SMAD1 + GATA co-bound regions. a, Relative enrichment of PU.1 and KLF1 binding at GATA2/1 + SMAD1 versus GATA2/1-alone sites at d0 and d5. A two-sided Fisher's exact test was used. **b**, Representative gene tracks at *FLI1* and *ALAS2* at d0 and d5 showing occupancy of the indicated TFs relative to the ATAC–seq and H3K27ac signal. **c**, The binding intensities (mean reads per million per bin) of PU.1 and SMAD1, comparing control and PU.1-overexpressing cells. **d**, Representative gene tracks at *CXCR4* and *CXCL2* with the peak intensities of PU.1 and SMAD1 in PU.1-overexpressing versus control cells. **e**, The binding intensities (average reads per bin) of PU.1, SMAD1 and GATA2, comparing control and PU.1-knockdown cells. **f**, Representative gene tracks at *CD84* and *WDR92* showing the peak intensities of PU.1, SMAD1 and GATA2 in PU.1-knockdown cells compared to control cells.





d

Term name	Binomial FDR *Q* value
Hematopoietic system phenoype	2.62×10^{-215}
Abnormal hematopoietic system morphology/development	2.51×10^{-211}
Abnormal blood cell morphology/development	1.55×10^{-208}
Abnormal hematopoiesis	2.64×10^{-207}
Abnormal hematopoietic cell number	1.12×10^{-184}
Abnormality of blood and blood-forming tissues	6.17×10^{-34}
Abnormal bleeding	6.14×10^{-16}
Abnormality of lymphocytes	$4.4.4586 \times 10^{-16}$
Prolonged bleeding time	7.08×10^{-10}
Abnormal leukocyte count	1.85×10^{-7}

Fig. 3 | SMAD1 + GATA co-bound enhancer regions form TSCs. a, Signal heatmaps representing the ChIP–seq coverage of putative enhancers that are bound by GATA2, demonstrating co-occupancy by multiple STFs (SMAD1, green; SMAD2, magenta; and TCF7L2, orange) and an MTF (GATA2, red) in progenitor CD34$^+$ cells at d0 on stimulation with BMP4, TGF-β and WNT signaling, respectively. Regions are considered occupied if they pass a significant coverage cutoff, shown as a binary green/white for the SMAD1 heatmap on the left. The GATA2 peak numbers obtained at d0 are shown on the *y* axis (33,470). H3K27ac and ATAC–seq heatmaps are also included. **b**, Representative gene tracks showing the peak intensities of SMAD1 (BMP signaling, green), SMAD2 (TGF-β signaling, magenta) and TCF7L2 (WNT signaling, orange), with the MTF GATA2 (red) and ATAC–seq signals at the *FLI1* and *FOS* genes at d0. **c**, A schematic representation of TSCs. TSCs are genomic regions that are co-occupied by multiple STFs induced by the respective signaling pathways. TSCs could be signal specific leading to specific combinations of STFs co-occupying a given region with stage-specific MTFs. **d**, Human and mouse phenotypes associated with the peaks that are co-bound by SMAD1, SMAD2 and TCF7L2 on stimulation with BMP4, TGF-β and WNT, respectively, identified using GREAT analysis. FDR, false discovery rate.

genes associated with SMAD1 + TCF7L2 + SMAD2 co-bound regions showed enrichment for blood functions (Fig. 3d), suggesting that SMAD1, under BMP stimulation, could serve as a marker for TSCs during erythroid differentiation. Accordingly, we created a list of progenitor enhancers (merging ATAC–seq and H3K27ac ChIP–seq) and progenitor TSCs (overlapping enhancers with GATA2/SMAD1 ChIP–seq) by combining the data points d0 and h6. Similarly, erythroid enhancers and TSCs were identified by combining d4 and d5 ATAC–seq and ChIP–seq data (Supplementary Table 4). These analyses showed that TSCs represent a small fraction of ATAC/H3K27ac-positive active enhancers at each differentiation stage (7.2–21.7% of all the active enhancers; Extended Data Fig. 3b).

Perturbed STF binding at a TSC affects gene expression. To determine the functional consequences of STF occupancy within a TSC, we mutated STF- or MTF-binding sites within a representative TSC. We identified a TSC that was co-bound by GATA2, SMAD1 and PU.1 in both progenitor CD34$^+$ (d0) and K562 erythro-leukemia cells and that was located within 5 kilobases (kb) from the nearest expressed gene, *LHFPL2* (Fig. 4a). Perturbation of the GATA, PU.1 or SMAD1 motifs in K562 or the human umbilical cord blood-derived erythroid progenitor (HUDEP2) cell line[44] (Extended Data Fig. 4a,b) showed that, as for loss of MTF-binding sites (PU.1 and GATA), perturbations of binding sites of the STF SMAD1 led to downregulation of the *LHFPL2* gene under BMP

stimulation, while expression of two flanking genes (*AP3B1* and *SCAMP1*) remained relatively unaltered (Fig. 4b,c and Extended Data Fig. 4c). On differentiation of HUDEP2 cells, perturbation of the same MTF or STF motifs within the *LHFPL2* TSC led to a significantly decreased percentage of mature CD71^{low}, CD235a^{+} erythroid cells (Fig. 4d,e). Mutation of the SMAD1 motif within the *LHFPL2* TSC led to decreased occupancy of SMAD1 but not PU.1 (Fig. 4f). However, PU.1 knockdown in K562 cells led to decreased SMAD1 occupancy while GATA2 binding remained relatively unaltered (Fig. 4g). These results predict that, within a TSC, MTFs can direct the binding of an STF but not vice versa, at least in this specific TSC, while STF binding can be at least as important as an MTF in controlling gene expression.

SNPs affecting RBC traits are enriched within TSCs. Since SNPs are primarily located in non-coding genomic regions[45–50], we wondered whether TSCs harbor non-coding GWAS variants associated with RBC traits. We compiled a set of SNPs from thirteen published GWAS studies associated with seven erythrocyte traits: hemoglobin concentration (HGB), hematocrit or packed cell volume (HCT), mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), MCH concentration (MCHC), red blood cell count (RBCC) and RBC distribution width (RDW)[4,14–20,51–55]. To increase the likelihood of including functional SNPs, we used 1,270 lead SNPs for individual traits/region, together with the co-inherited SNPs in high linkage disequilibrium with them (LD $r^2 \geq 0.6$, as suggested by previous studies[56–61], designated here as lead + LD SNPs). Altogether, 29,069 lead and LD SNPs with at least 2 usable alleles across 924 loci associated with the 7 RBC traits were used (Supplementary Table 5a,b). Out of 1,270 lead SNPs and 29,069 lead + LD RBC-trait SNPs, 353 and 3,318 SNPs were located in enhancers defined by ATAC–seq and H3K27ac ChIP–seq data (Extended Data Fig. 5a and Supplementary Table 5c,d). To confirm that our criteria of selecting SNPs enriched for potentially functional variants, we used RegulomeDB (ref. [62]) and found a significant enrichment in SNPs with predicted effects on gene regulation (RegulomeDB score ≤ 4; 51.1% of ATAC + H3K27ac SNPs compared to 19.6% of all SNPs; Fig. 5a and Supplementary Table 3c).

We then investigated whether enhancer-associated SNPs are primarily located in TSCs. We assessed the number of SNPs located within TSCs out of the 353 lead enhancer SNPs or the 3,318 lead + LD enhancer SNPs (Supplementary Table 5e,f) and compared the number of SNPs in TSCs to the number of SNPs in overall enhancers or in GATA2/1-alone enhancers, normalized to the size of each region type in base pairs. We found that enhancer variants are significantly enriched within TSCs (Fig. 5b and Supplementary Table 3d). We also analyzed an independent list of fine-mapping-based SNPs associated with 16 different blood traits[23]. Indeed, fine-mapping-based SNPs (with posterior probability value PP > 0.01) are significantly enriched in TSCs compared to overall enhancers or enhancers that are occupied by GATA2/1 alone during differentiation (Fig. 5b and Supplementary Table 3d). Taken together, these results show that enhancer variants are significantly enriched within TSCs.

To test whether SNPs linked to erythroid traits and not the traits of other lineages are enriched in erythroid TSCs, we compiled SNPs linked to platelet traits as controls. We used 786 platelet-trait loci regions associated with 575 lead and 22,158 lead + LD SNPs (LD $r^2 \geq 0.6$) with at least 2 usable alleles[19] (Supplementary Table 5g–j). By comparing lead RBC-trait SNPs to lead platelet-trait SNPs, or lead + LD RBC-trait SNPs versus lead + LD platelet-trait SNPs, we observed that RBC-trait SNPs are significantly enriched within erythroid TSCs (Fig. 5c–e, Extended Data Fig. 5b and Supplementary Table 3e). In conclusion, RBC-trait SNPs, but not platelet-trait SNPs, are primarily enriched within erythroid TSCs.

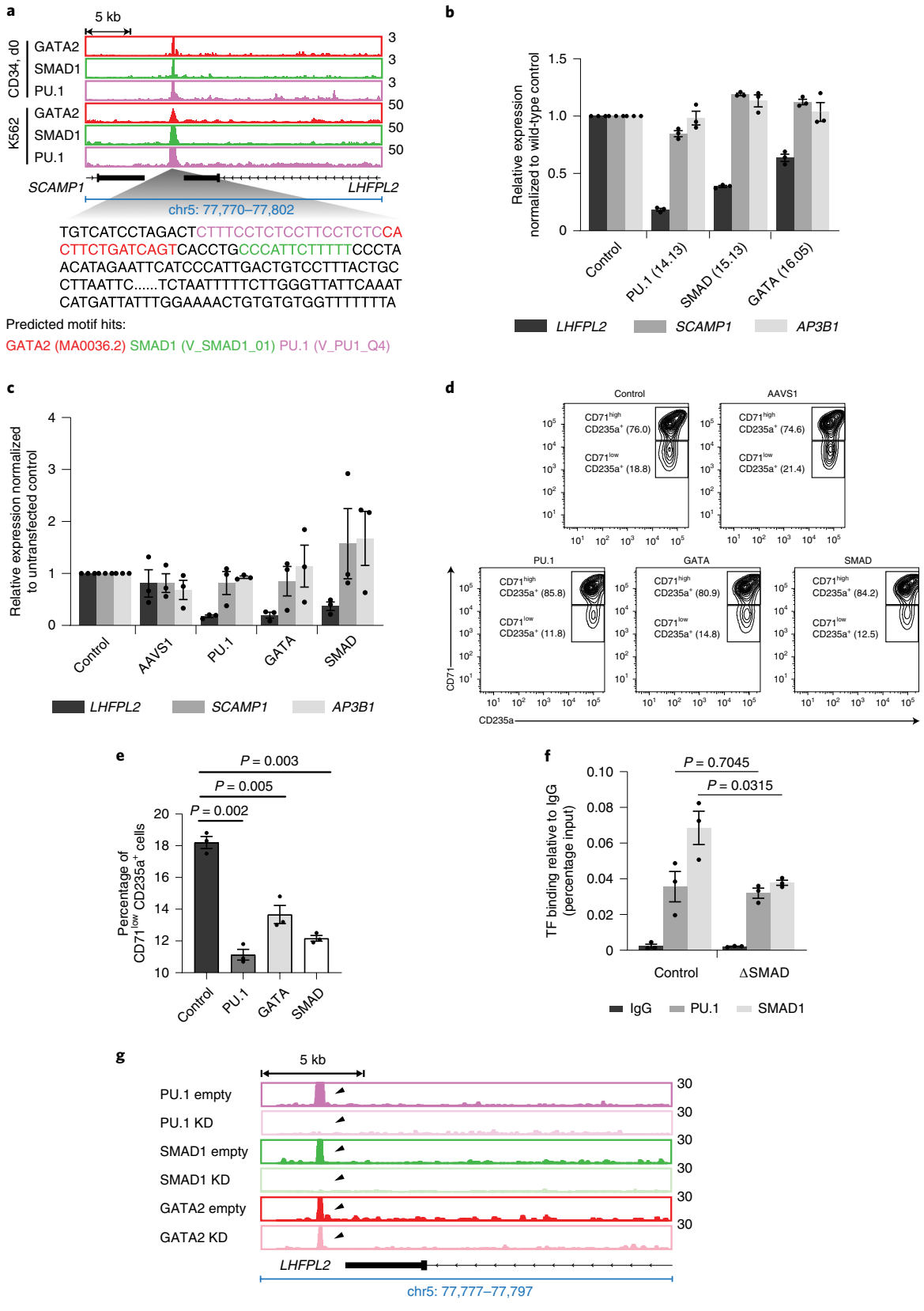
Many RBC-trait SNPs are located within STF motif hits. We then asked whether non-coding RBC-trait SNPs could modulate transcription by altering the binding of transcription factors (TFs). We predicted TF motif hits (Methods) and created lists of predicted binding sites of hematopoietic MTFs and generic STFs (Supplementary Table 6 and Supplementary Note). We calculated the number of enhancer-associated SNPs appearing in STF or MTF motif hits. We categorized the motifs as STF-alone or MTF-alone (recognized by STFs or MTFs, respectively, but not both) and STFs and MTFs (motif hits recognized by either STFs or MTFs). While 72.4% of lead SNPs within MTF or STF motif hits overlap STF-alone motif hits, only 9.8% overlap MTF-alone motif hits and 17.8% reside on ambiguous STF and MTF motif hits (Fig. 5f). Similar conclusions were true for lead + LD (Fig. 5f) and enhancer-associated fine-mapped SNPs (PP > 0.01) overlapping TF motif hits (Fig. 5f). We then investigated whether the SNPs within STF motif hits are enriched in TSCs compared to non-TSC enhancers. Using either the lead, lead + LD or the fine-mapped SNPs, we compared the number of SNPs in STF motif hits between TSCs and non-TSC enhancers, normalized to the total number of base pairs in each region type. Indeed, TSCs show a significant enrichment for SNPs associated with STF motif hits relative to non-TSC enhancers (Fig. 5g and Supplementary Table 3f). Thus, the majority of enhancer-associated RBC-trait SNPs that overlap TF-binding sequences are found in STF-binding sites, and such STF SNPs are significantly over-represented within TSCs.

Functional alteration of STF–DNA binding by RBC SNPs. We hypothesized that STF SNPs may lead to differential STF occupancy within TSCs, resulting in altered gene expression under stimulation. Thus, we analyzed protein binding microarray (PBM) datasets[63] to identify RBC-trait SNPs that affect binding of STFs within TSCs (Extended Data Fig. 6a). Using previously published PBM data for several STFs (Supplementary Table 7)[64,65], we compared the binding of in vitro–expressed SMAD between the two alleles of SNPs located in open chromatin enhancer regions as a proof of principle. Since SMAD1 PBM data were not available, we analyzed a mouse SMAD3 PBM dataset[66] (the 69.61% identity of the MH1 DNA-binding domain sequence to the human SMAD1 MH1 DNA-binding domain strongly argues that the TFs share similar sequence specificity[65]). Analysis of PBM data identified examples

Fig. 4 | STFs and MTFs at TSCs control gene expression. a, Overlap of occupancy of PU.1, GATA2 and SMAD1 at a representative TSC near the *LHFPL2* gene in progenitor CD34^{+} (d0) and K562 cells. The locations of the PU.1, GATA and SMAD1 motifs within the TSC are shown. **b**, Relative alteration of expression of *LHFPL2*, *SCAMP1* and *AP3B1* due to mutation of the respective TF motifs in specific K562 clones, as indicated. The mean ± s.e.m. is shown ($n = 3$; 3 biologically independent experiments). A two-sided Student's *t*-test was used. **c**, Relative change of expression of *LHFPL2*, *SCAMP1* and *AP3B1* in bulk-edited HUDEP2 cells transduced with single gRNAs (sgRNAs) targeting PU.1, SMAD1 and/or GATA motifs in comparison with non-transduced cells or cells transduced with a control (AAVS1). The mean ± s.e.m. is shown ($n = 3$; 3 biologically independent experiments). A two-sided Student's *t*-test was used. **d**, Representative flow cytometry plots for CD71 and CD235a for HUDEP2 cell bulk cultures from **c**. The percentage distributions of cells within CD71^{high}CD235a^{+} and CD71^{low}CD235a^{+} compartments are shown. **e**, Bar plots comparing the percentage of CD71^{low}CD235a^{+} cells from **d**. The mean ± s.e.m. is shown ($n = 3$; 3 biologically independent experiments). A two-sided Student's *t*-test was used. **f**, Alteration of binding of PU.1 and SMAD1 in K562 cells with mutation of the SMAD motif. The mean ± s.e.m. is shown ($n = 3$; 3 biologically independent experiments). A two-sided Student's *t*-test was used. **g**, Gene tracks at the *LHFPL2* TSC showing the peak intensities of PU.1, SMAD1 and GATA2 in PU.1-knockdown cells compared to control cells.

such as rs737092, where the change of the T>C allele significantly diminishes SMAD binding but causes little change in GATA binding between the two alleles of rs737092, despite its close proximity to the GATA motif[15] (Fig. 6a,b and Extended Data Fig. 6b). This result argues for the existence of RBC-trait-associated SNPs that could perturb STF–DNA binding without notably altering the binding of a hematopoietic MTF.

We then analyzed our list of enhancer-associated SNPs for their predicted effects on STF binding and gene expression. For this purpose, universal PBM octanucleotide enrichment (*E*) score datas-



ets were downloaded from the UniPROBE and CIS-BP databases (Supplementary Table 7)[64–69]. Of the 3,318 enhancer-associated lead + LD variants that included indels, we focused our analysis on the 3,263 single-nucleotide substitutions (Supplementary Note). We considered perturbed binding events for GATA-family MTFs, by using an averaged GATA binding profile from available GATA-family PBM datasets[64], for comparison against several STFs. We found that several STFs, including SMAD3, TCF4, RXRA, GLI1/2/3 and EGR1/2, showed a greater than expected frequency of perturbed binding events in this set of RBC-trait SNPs (Benjamini–Hochberg-adjusted empirical P value < 0.05), while GATA binding appeared to be perturbed less frequently than expected (Fig. 6c,d and Extended Data Fig. 6c). Inclusion of fine-mapped variants in PBM analysis further supported this conclusion (Supplementary Note and Supplementary Table 8). To further investigate the effects of STF-binding-altering SNPs on downstream gene expression, we coupled the PBM approach with expression quantitative trait locus (eQTL) analysis using microarray gene expression profiles of peripheral blood, isolated from participants in the Framingham Heart Study (FHS)[70]. In several instances, where the SNP resulted in a significant decrease in STF binding, the SNP was also identified as a *cis*-eQTL in the FHS dataset leading to a dose-dependent expression reduction of a proximal gene (Fig. 6d and Extended Data Fig. 6c). A total of 86 out of the 115 transcripts from the FHS *cis*-eQTL gene list, and 108 out of 148 transcripts from the FHS *cis*-eQTL exon list showed at least 1 *cis*-eQTL SNP that also affected STF binding. Our RNA-seq results verify that those genes show a steady increase in expression during erythroid differentiation (Fig. 6e). Notably, loss of STF binding induced by a SNP allele could also lead to increased expression of associated genes. Overall, SNP-mediated modulation of STF–DNA binding results in expression alteration of relevant trait-associated genes.

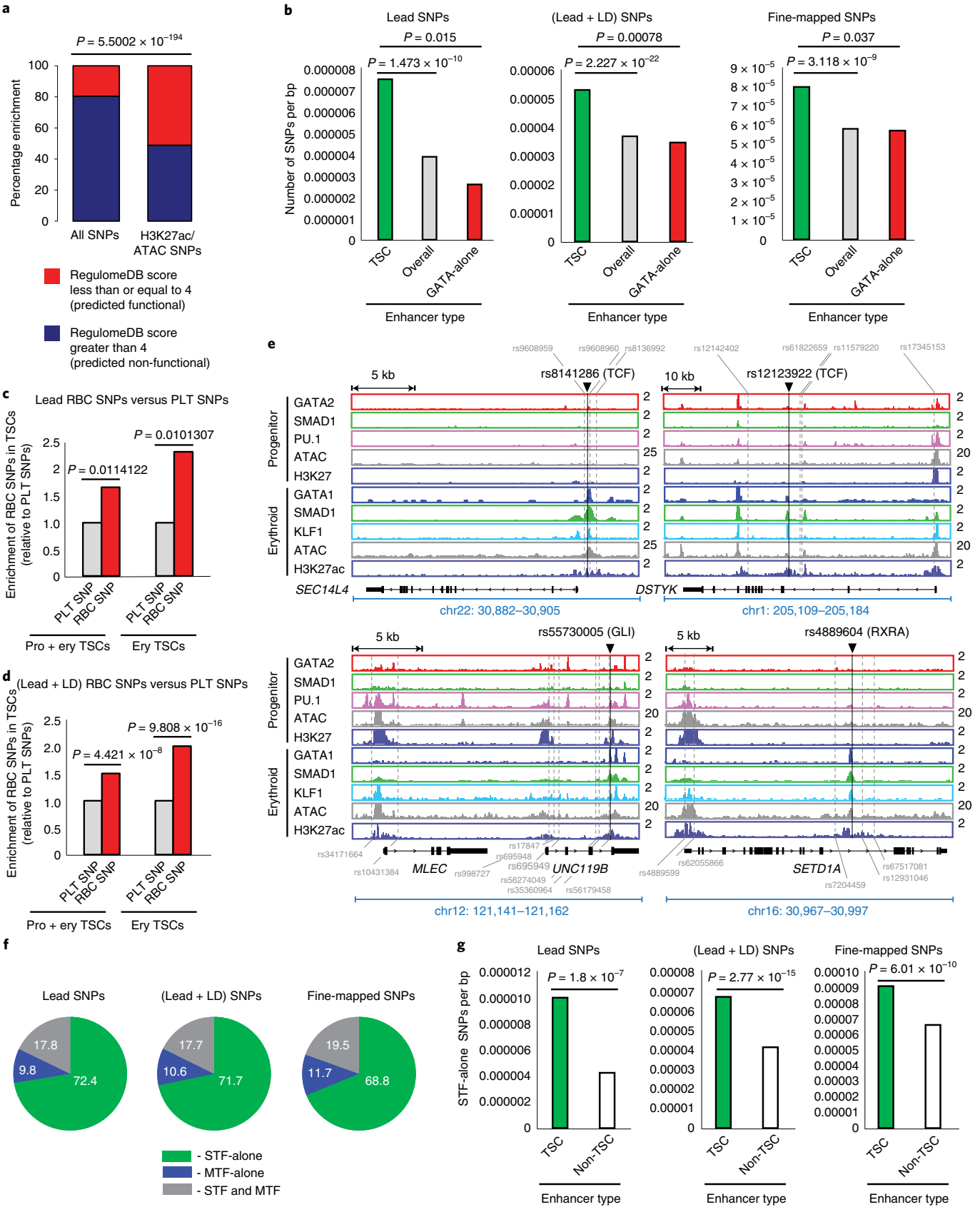
STF SNPs can perturb DNA binding and gene expression. To validate whether alternative alleles of representative SNPs govern STF occupancy and gene expression, we investigated the effects of SMAD1 binding at the MCV-associated SNP rs9467664 (T>A), residing on a SMAD target sequence within a TSC proximal to *HIST1H4A*, which shows increased expression during erythroid differentiation (Extended Data Fig. 7a,b). Electrophoretic mobility shift assays showed that oligonucleotides harboring the T but not the A allele could efficiently bind SMAD1 (Extended Data Fig. 7c,d). eQTL analysis from the FHS showed that the A allele is significantly associated with reduced levels of *HIST1H4A* messenger RNA compared to the T allele (Extended Data Fig. 7e), further supporting our hypothesis that alteration of SMAD1 binding by an RBC-trait-associated SNP may have significant effects on gene expression.

To test whether perturbed STF binding impairs signal-induced gene expression, we selected a second SNP rs737092 (T>C). This SNP resides in an erythroid-specific TSC co-bound by GATA1, SMAD1 and KLF1 within an H3K27ac-positive open chromatin region. The SNP is present within a SMAD motif flanked by two GATA motifs (Fig. 7a). PBM analysis showed that this SNP perturbed SMAD binding without altering GATA binding. rs737092 was identified in a previously published massively parallel reporter assay study as functional, regulating the expression of *RBM38* (ref. [15]), which was confirmed by eQTL analysis from the FHS study. Finally, *RBM38* is expressed at a significantly higher level in a population with the T but not the C allele and its expression is steadily increased during differentiation in our dataset (Fig. 7b,c). We obtained a CRISPR–Cas-modified K562 cell line[15], where the SMAD1 motif within the *RBM38* TSC is mutated together with the upstream but not the downstream GATA motif (Extended Data Fig. 7f). ChIP–quantitative PCR (qPCR) assays for SMAD1 binding under BMP stimulation showed significant abrogation of SMAD1 binding in these cells, but GATA1 binding remained relatively unchanged presumably owing to compensation from the other flanking GATA motif (Extended Data Fig. 7g). As a control, the binding of the WNT-responsive factor TCF7L2 (ref. [28]) to its motif in the same TSC after WNT pathway stimulation with BIO (ref. [71]) was not affected (Extended Data Fig. 7g). Concomitantly, the expression of *RBM38* was significantly reduced in the mutant cells under BMP but not under BIO treatment (Extended Data Fig. 7h). We then cloned the actual *RBM38* TSC with either the T or the C allele upstream of the firefly luciferase gene[15]. The T allele, which retains SMAD binding, showed a higher increase in luciferase expression under BMP stimulation relative to no stimulation or dorsomorphin treatment (Extended Data Fig. 7i). These results suggest that abrogation of the SMAD1 motif in the *RBM38* TSC that harbors the rs737092 SNP diminished SMAD1 binding and compromised BMP responsiveness.

Effect of STF SNPs within TSCs in primary human samples. We then investigated the effects of RBC-trait SNPs in primary human peripheral blood CD34$^+$ cells. We first validated that knockdown of SMAD1 in CD34$^+$ cells impaired activation of *RBM38* under BMP stimulation (Extended Data Fig. 7j,k). Next, we screened 18 human donors and identified individuals with homozygous alleles for preselected TSC-associated SNPs—rs737092 (T>C) and rs2154434 (C>A) (minor allele frequencies for rs737092 and rs2154434 are 47.9% and 42.9%, respectively). Similar to rs737092, rs2154434 is also located within an erythroid TSC during erythroid differentiation (Fig. 7d), and we observed a dose-dependent decrease of *ITSN1* expression in FHS when the C allele is replaced by the A allele (Fig. 7e). *ITSN1* also increases expression during erythroid differentiation (Fig. 7f). Individual donors with homozygous genotypes for alternative alleles of rs737092 and rs2154434 were confirmed by PCR and sequencing (Fig. 7g,h). We next evaluated TF binding and BMP4 responsiveness of the alleles in donor CD34$^+$ cells. rs737092 should affect SMAD but not TCF7L2 binding when the T is replaced by the C allele. Indeed, ChIP–PCR performed in

Fig. 5 | RBC-trait SNPs enriched within TSCs predominantly reside in STF motifs. a, Enrichment of predicted functional SNPs in non-exonic open enhancer regions versus all SNPs; 2 × 2 chi-squared significance tests were used. **b**, Enrichment of SNPs within TSCs versus all and GATA-alone enhancers; 2 × 2 chi-squared significance values are shown. P values for permutation tests obtained by shuffling SNP positions in TSCs <0.0001 for all SNP types; in GATA-alone enhancers: lead SNPs, $P = 0.9166$; lead + LD SNPs, $P = 1$; fine-mapped SNPs, $P = 1$. P values by permuting the TSC/non-TSC labels of enhancers <0.0001 for all SNP types. **c**,**d**, Enrichment of lead and lead + LD RBC-trait SNPs, relative to platelet-trait SNPs within progenitor + erythroid and erythroid-alone TSCs; 2 × 2 chi-squared significance tests were used. **e**, Example RBC-trait SNPs (black line) localized within stage-specific TSCs. Binding sites of STFs at these SNPs are shown. Additional SNPs with significant LD within enhancers are shown (gray dashed lines). **f**, Distribution of lead, lead + LD and fine-mapped SNPs at STF-alone MTF-alone and STF and MTFs motifs. **g**, Enrichment of SNPs overlapping STF-alone motif hits within TSC versus non-TSC enhancers; 2 × 2 chi-squared significance values are shown. P values calculated by randomly permuting SNP positions showing the enrichment of STF-alone SNPs in TSCs: lead SNPs, $P < 0.0001$; lead + LD SNPs, $P < 0.0001$; fine-mapped SNPs, $P < 0.0001$. P values calculated by randomly permuting labels of enhancers as TSC/non-TSC: lead SNPs, $P < 0.0001$; lead + LD SNPs, $P < 0.0001$; fine-mapped SNPs, $P < 0.0001$. P values calculated by randomly permuting positions of STF motif hits: lead SNPs, $P = 0.0194$; lead + LD SNPs, $P < 0.0001$; fine-mapped SNPs, $P = 0.0033$.

BMP4-treated CD34+ cells with rs737092 alleles, differentiated for five days, showed attenuated SMAD1 binding under T>C change (Fig. 7i). TCF7L2 binding under BIO stimulation or GATA1 binding did not change significantly (Fig. 7i). In contrast, rs2154434 should primarily disrupt the DNA binding of TCF7L2 but not of SMAD1. Indeed, we observed disrupted TCF7L2 but not SMAD1 or GATA1 binding following BIO stimulation when the C allele was replaced by the A allele (Fig. 7j). We also tested the allele-specific



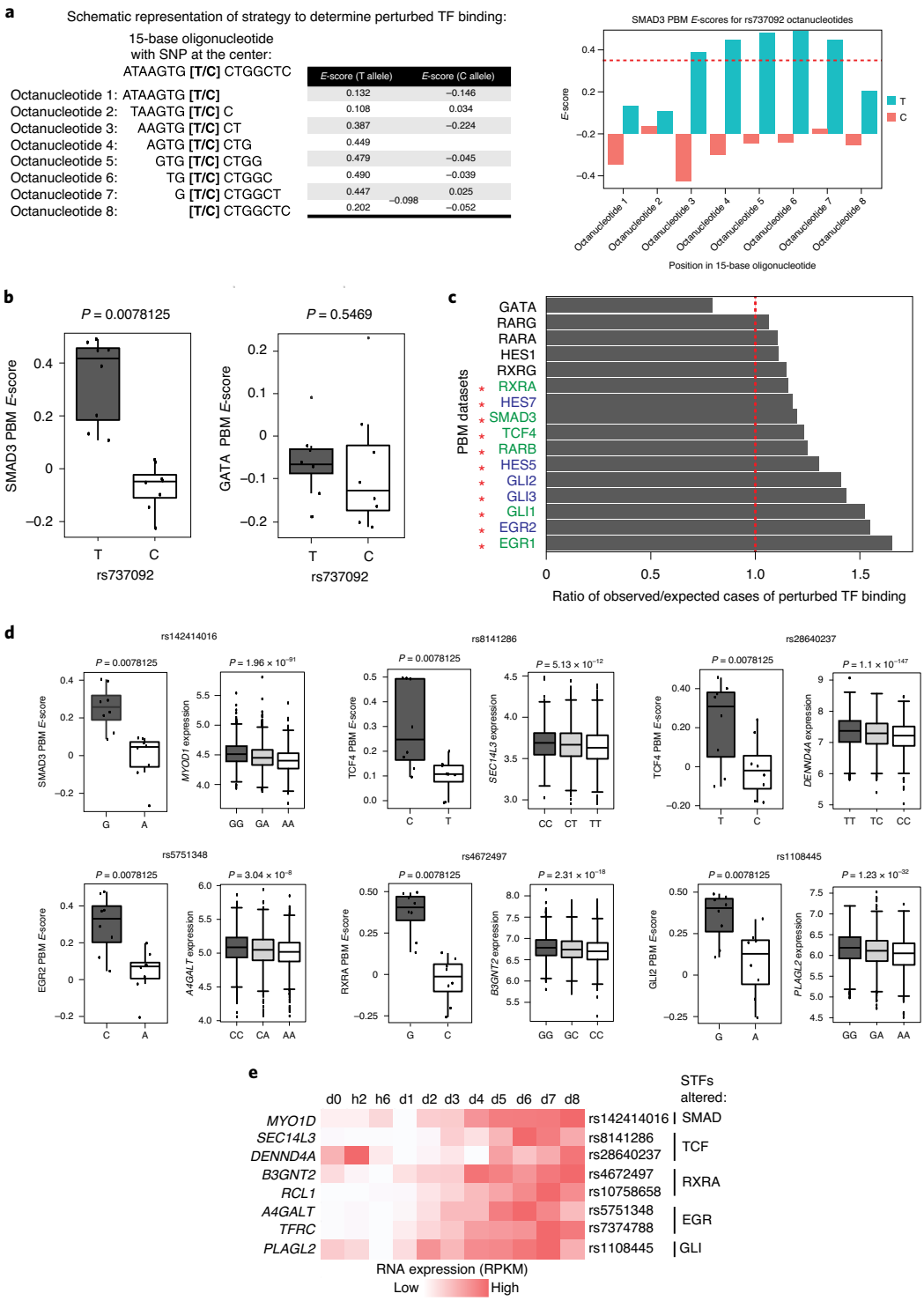
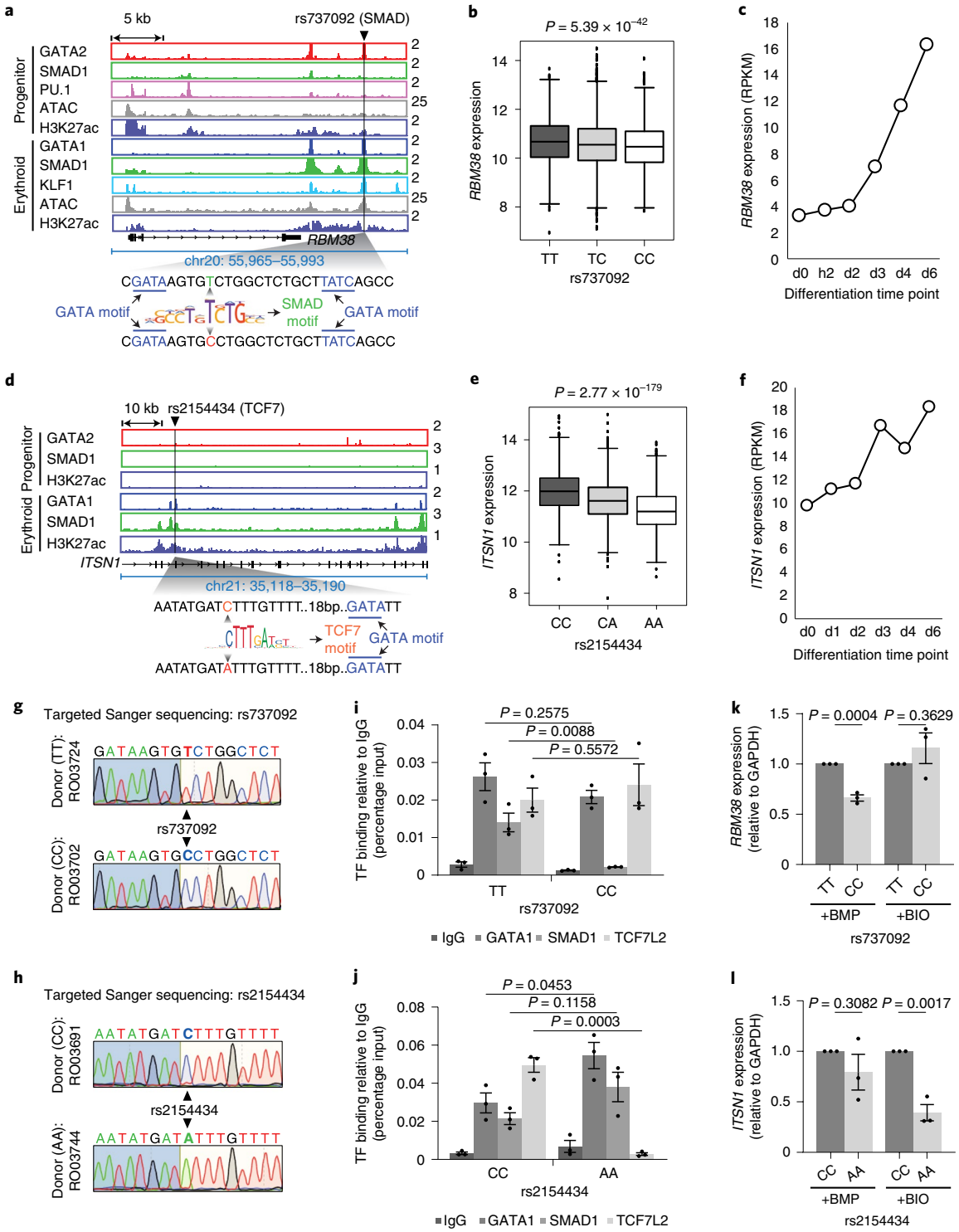


Fig. 6 | PBM identifies RBC-trait-associated SNPs that perturb STF–DNA binding. a, A schematic representation of the perturbed TF binding analysis strategy using PBM data. The red dashed line indicates a universal PBM octanucleotide *E*-score = 0.35, which is used as a threshold for specific binding by a TF. **b**, Boxplots representing SMAD3 and the average GATA PBM *E*-scores of rs737092. Two-sided Wilcoxon signed-rank tests were used. **c**, Bar plots depicting the ratio of the observed versus the expected number of perturbed TF binding events. The red dashed line indicates ratio = 1. The red asterisks indicate STFs with significantly greater than expected numbers of perturbed TF binding events (*Q* value < 0.05 after Benjamini–Hochberg adjustment of the empirical *P* values). The STFs indicated in green are expressed during CD34$^+$ differentiation (RPKM > 1). The STFs in blue have close paralogs that are expressed during erythroid differentiation. **d**, Example SNPs showing perturbed STF binding from PBM analysis and the corresponding expression distribution of the most significantly altered nearby gene in homozygous and heterozygous individuals obtained from FHS eQTL analysis. The boxplots represent the median as a line in the box, the first and third quartiles as the box, and 1.5 times the interquartile range as whiskers. Two-sided Wilcoxon signed-rank tests were used for the PBM boxplots. A two-sided test with a linear model for EffectAlleleDosage was used for eQTL analysis with Benjamini–Hochberg-adjusted *P* values. **e**, A heatmap depicting the expression of the most significantly altered nearby gene (from FHS eQTL analysis) during normal erythroid differentiation.



mRNA expression of *RBM38* and *ITSN1* for the alleles of rs737092 and rs2154434 after acutely stimulating CD34+ cells with BMP4 and BIO, respectively, at d5 of differentiation for 2 h. Change from T to C allele mediated by the rs737092 SNP led to decreased expression of *RBM38* under BMP but not BIO treatment (Fig. 7k). Similarly, *ITSN1* expression was downregulated primarily under WNT stimulation when the C allele of rs2154434 was replaced by the A allele (Fig. 7l). These results suggest that RBC-trait-associated SNPs, overlapping TF-binding sites, often abrogate DNA binding of STFs and not of MTFs to affect gene expression by respective signaling pathways in primary human samples.

Discussion

The majority of GWAS-associated variants linked to human genetic traits and diseases are non-coding[45–50]. Using genetic fine-mapping of 16 traits associated with blood, Ulirsch et al. showed that SNPs are often located within open chromatin regions enriched for lineage-specific MTF motifs[23]. Although blood-trait-associated GWAS SNPs are often found in close proximity to MTF motifs, the majority do not disrupt their binding sites directly[15,22,23]. Here, utilizing functional and computational approaches, we show that the alteration of STF binding induced by SNPs within TSCs, which represent a subset of enhancers co-occupied by both MTFs and

Fig. 7 | STF SNPs perturb STF–DNA binding and abrogate signal responsiveness. a, Gene tracks at *RBM38* depicting the erythroid-specific TSC containing rs737092 at the SMAD motif. **b**, *RBM38* eQTL analysis for rs737092. The boxplots represent the median *RBM38* expression as a line in the box, the first and third quartiles as the box, and 1.5 times the interquartile range as whiskers. A two-sided test with a linear model for EffectAlleleDosage was used: effect estimate (β) = −0.05211; T-statistics = −13.6994, R^2 = 0.034622; $\log_{10}$[*P* value] = −41.2683, $\log_{10}$[Benjamini–Hochberg FDR] = −38.6118. **c**, Expression RPKM values for the *RBM38* gene at different stages of CD34$^+$ erythroid differentiation. **d**, Gene tracks at *ITSN1* depicting the erythroid-specific TSC containing rs2154434 at the TCF7 motif. **e**, *ITSN1* eQTL analysis for rs2154434. The boxplots represent the median *ITSN1* expression as a line in the box, the first and third quartiles as the box, and 1.5 times the interquartile range as whiskers. A two-sided test with a linear model for EffectAlleleDosage was used: effect estimate (β) = −0.0486; *t*-statistics = −29.7008, R^2 = 0.144255; $\log_{10}$[*P* value] = −178.558, $\log_{10}$[Benjamini–Hochberg FDR] = −175.322. **f**, Expression RPKM values for *ITSN1* at different stages of CD34$^+$ erythroid differentiation. **g**,**h**, Sanger sequencing chromatograms of individual donors for the SNPs rs737092 and rs2154434. Donor numbers are indicated. **i**,**j**, Binding alteration of GATA1, SMAD1 and TCF7L2 for alternative alleles of rs737092 and rs2154434. The mean ± s.e.m. is shown (n = 3; 3 biologically independent experiments). A two-sided Student's *t*-test was used. **k**,**l**, qPCR analysis comparing the expression of *RBM38* and *ITSN1*, relative to *GAPDH*, for alternative alleles of rs737092 and rs2154434, respectively, under BMP and BIO treatment. The mean ± s.e.m. is shown (n = 3; 3 biologically independent experiments). A two-sided Student's *t*-test was used.

STFs, may drive a disproportionate fraction of phenotypic variability of human RBCs. Importantly, using several systems, including primary CD34$^+$ cells isolated from human donors with specific SNP alleles, we show that SNPs altering STF binding can modulate the induction of adjacent genes by respective signaling pathways (Supplementary Note).

It is important to understand why allele-specific effects of SNPs residing in STFs are more common than in MTF motifs. We speculate that SNPs affecting MTF binding can drastically affect expression of genes essential for development, and thus be less likely to be favored by natural selection. STF SNPs, on the other hand, can cause expression variability leading to tolerable phenotypic changes in RBC traits and thereby escape evolutionary pressure. Accordingly, we evaluated the published prediction scores from NCboost (ref. [72]), which predict the pathogenicity of a variant occurring at non-coding positions of the genome based on evolutionary signals. The predicted pathogenicity of altering bases in MTF motifs appears significantly higher than that caused by alterations in STF motif hits (data not shown). Thus, an STF SNP can render enhancers and their regulated genes sub-optimally responsive to one or more signaling pathways during episodic stresses such as infections or environmental changes. The abnormal response to periodic stress signals could contribute to tissue damage and disease over time. Such altered signaling events over time could lead to 'signalopathies', ultimately resulting in phenotypic variation and susceptibility to a spectrum of human genetic diseases (Extended Data Fig. 8).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-00738-2.

Received: 15 April 2019; Accepted: 14 October 2020;
Published online: 23 November 2020

References

1. Evans, D. M., Frazer, I. H. & Martin, N. G. Genetic and environmental causes of variation in basal levels of blood cells. *Twin Res.* **2**, 250–257 (1999).
2. Guindo, A., Fairhurst, R. M., Doumbo, O. K., Wellems, T. E. & Diallo, D. A. X-linked G6PD deficiency protects hemizygous males but not heterozygous females against severe malaria. *PLoS Med.* **4**, e66 (2007).
3. Lin, J. P. et al. Evidence for linkage of red blood cell size and count: genome-wide scans in the Framingham Heart Study. *Am. J. Hematol.* **82**, 605–610 (2007).
4. Lo, K. S. et al. Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans. *Hum. Genet.* **129**, 307–317 (2011).
5. Tishkoff, S. A. et al. Haplotype diversity and linkage disequilibrium at human *G6PD*: recent origin of alleles that confer malarial resistance. *Science* **293**, 455–462 (2001).
6. Whitfield, J. B. & Martin, N. G. Genetic and environmental influences on the size and number of cells in the blood. *Genet. Epidemiol.* **2**, 133–144 (1985).
7. Koury, M. J. Abnormal erythropoiesis and the pathophysiology of chronic anemia. *Blood Rev.* **28**, 49–66 (2014).
8. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.* **93**, 779–797 (2013).
9. Guo, M. H. et al. Comprehensive population-based genome sequencing provides insight into hematopoietic regulatory mechanisms. *Proc. Natl Acad. Sci. USA* **114**, E327–E336 (2017).
10. Gusev, A. et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
11. 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).
12. Nandakumar, S. K., Ulirsch, J. C. & Sankaran, V. G. Advances in understanding erythropoiesis: evolving perspectives. *Br. J. Haematol.* **173**, 206–218 (2016).
13. Patwardhan, R. P. et al. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265–270 (2012).
14. 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).
15. Ulirsch, J. C. et al. Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165**, 1530–1545 (2016).
16. van der Harst, P. et al. Seventy-five genetic loci influencing the human red blood cell. *Nature* **492**, 369–375 (2012).
17. Ganesh, S. K. et al. Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nat. Genet.* **41**, 1191–1198 (2009).
18. van Rooij, F. J. et al. Genome-wide trans-ethnic meta-analysis identifies seven genetic loci influencing erythrocyte traits and a role for *RBPMS* in erythropoiesis. *Am. J. Hum. Genet.* **100**, 51–63 (2017).
19. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429 (2016).
20. Chami, N. et al. Exome genotyping identifies pleiotropic variants associated with red blood cell traits. *Am. J. Hum. Genet.* **99**, 8–21 (2016).
21. Pankratz, N. et al. Genomewide association study for susceptibility genes contributing to familial Parkinson disease. *Hum. Genet.* **124**, 593–605 (2009).
22. Levo, M. et al. Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res.* **25**, 1018–1029 (2015).
23. Ulirsch, J. C. et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* **51**, 683–693 (2019).
24. Dent, P. et al. Stress and radiation-induced activation of multiple intracellular signaling pathways. *Radiat. Res.* **159**, 283–300 (2003).
25. Gaki, G. S. & Papavassiliou, A. G. Oxidative stress-induced signaling pathways implicated in the pathogenesis of Parkinson's disease. *Neuromolecular Med.* **16**, 217–230 (2014).
26. Uchida, K. et al. Activation of stress signaling pathways by the end product of lipid peroxidation. 4-hydroxy-2-nonenal is a potential inducer of intracellular peroxide production. *J. Biol. Chem.* **274**, 2234–2242 (1999).
27. Mullen, A. C. et al. Master transcription factors determine cell-type-specific responses to TGF-β signaling. *Cell* **147**, 565–576 (2011).
28. Trompouki, E. et al. Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. *Cell* **147**, 577–589 (2011).
29. Sankaran, V. G. et al. Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor *BCL11A*. *Science* **322**, 1839–1842 (2008).
30. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21**, 381–395 (2011).

31. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
32. Lenox, L. E., Perry, J. M. & Paulson, R. F. BMP4 and Madh5 regulate the erythroid response to acute anemia. *Blood* **105**, 2741–2748 (2005).
33. Lenox, L. E., Shi, L., Hegde, S. & Paulson, R. F. Extramedullary erythropoiesis in the adult liver requires BMP-4/Smad5-dependent signaling. *Exp. Hematol.* **37**, 549–558 (2009).
34. McReynolds, L. J., Tucker, J., Mullins, M. C. & Evans, T. Regulation of hematopoiesis by the BMP signaling pathway in adult zebrafish. *Exp. Hematol.* **36**, 1604–1615 (2008).
35. Porayette, P. & Paulson, R. F. BMP4/Smad5 dependent stress erythropoiesis is required for the expansion of erythroid progenitors during fetal development. *Dev. Biol.* **317**, 24–35 (2008).
36. Hnisz, D. et al. Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Mol. Cell* **58**, 362–370 (2015).
37. Whyte, W. A. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
38. Fisher, R. C. & Scott, E. W. Role of PU.1 in hematopoiesis. *Stem Cells* **16**, 25–37 (1998).
39. Li, Y., Luo, H., Liu, T., Zacksenhaus, E. & Ben-David, Y. The *ets* transcription factor Fli-1 in development, cancer and disease. *Oncogene* **34**, 2022–2031 (2015).
40. Shivdasani, R. A. & Orkin, S. H. Erythropoiesis and globin gene expression in mice lacking the transcription factor NF-E2. *Proc. Natl Acad. Sci. USA* **92**, 8690–8694 (1995).
41. Siatecka, M. & Bieker, J. J. The multifunctional role of EKLF/KLF1 during erythropoiesis. *Blood* **118**, 2044–2054 (2011).
42. Nakao, A. et al. TGF-β receptor-mediated signalling through Smad2, Smad3 and Smad4. *EMBO J.* **16**, 5353–5362 (1997).
43. McLean, C. Y. et al. GREAT improves functional interpretation of *cis*-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
44. Kurita, R. et al. Establishment of immortalized human erythroid progenitor cell lines able to produce enucleated red blood cells. *PLoS ONE* **8**, e59890 (2013).
45. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102–R110 (2015).
46. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
47. Cohen, A. J. et al. Hotspots of aberrant enhancer activity punctuate the colorectal cancer epigenome. *Nat. Commun.* **8**, 14400 (2017).
48. Corradin, O. et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.* **24**, 1–13 (2014).
49. Morrow, J. J. et al. Positively selected enhancer elements endow osteosarcoma cells with metastatic competence. *Nat. Med.* **24**, 176–185 (2018).
50. Scacheri, C. A. & Scacheri, P. C. Mutations in the noncoding genome. *Curr. Opin. Pediatr.* **27**, 659–664 (2015).
51. The CHARGE Consortium Hematology Working Group. Meta-analysis of rare and common exome chip variants identifies *S1PR4* and other loci influencing blood cell traits. *Nat. Genet.* **48**, 867–876 (2016).
52. Chen, Z. et al. Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. *Hum. Mol. Genet.* **22**, 2529–2538 (2013).
53. Li, C. et al. Genome-wide association study meta-analysis of long-term average blood pressure in East Asians. *Circ. Cardiovasc. Genet.* **10**, e001527 (2017).
54. Paul, D. S. et al. Maps of open chromatin highlight cell type-restricted patterns of regulatory sequence variation at hematological trait loci. *Genome Res.* **23**, 1130–1141 (2013).
55. Paul, D. S. et al. Maps of open chromatin guide the functional follow-up of genome-wide association signals: application to hematological traits. *PLoS Genet.* **7**, e1002139 (2011).
56. Amos, C. I. et al. The oncoarray consortium: a network for understanding the genetic architecture of common cancers. *Cancer Epidemiol. Biomark. Prev.* **26**, 126–135 (2017).
57. Fachal, L. et al. Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. *Nat. Genet.* **52**, 56–73 (2020).
58. Fritsche, L. G. et al. Association of polygenic risk scores for multiple cancers in a phenome-wide study: results from the Michigan Genomics Initiative. *Am. J. Hum. Genet.* **102**, 1048–1061 (2018).
59. Jansen, I. E. et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
60. Lin, J. R. et al. Integrated post-GWAS analysis sheds new light on the disease mechanisms of schizophrenia. *Genetics* **204**, 1587–1600 (2016).
61. Vicente, C. T. et al. Long-range modulation of *PAG1* expression by 8q21 allergy risk variants. *Am. J. Hum. Genet.* **97**, 329–336 (2015).
62. Boyle, A. P. et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
63. Liu, N. et al. Direct promoter repression by BCL11A controls the fetal to adult hemoglobin switch. *Cell* **173**, 430–442 (2018).
64. Hume, M. A., Barrera, L. A., Gisselbrecht, S. S. & Bulyk, M. L. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein–DNA interactions. *Nucleic Acids Res.* **43**, D117–D122 (2015).
65. Weirauch, M. T. et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
66. Badis, G. et al. Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).
67. Barrera, L. A. et al. Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science* **351**, 1450–1454 (2016).
68. Mariani, L., Weinand, K., Vedenko, A., Barrera, L. A. & Bulyk, M. L. Identification of human lineage-specific transcriptional coregulators enabled by a glossary of binding modules and tunable genomic backgrounds. *Cell Syst.* **5**, 187–201 (2017).
69. Peterson, K. A. et al. Neural-specific Sox2 input and differential Gli-binding affinity provide context and positional information in Shh-directed neural patterning. *Genes Dev.* **26**, 2802–2816 (2012).
70. Joehanes, R. et al. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biol.* **18**, 16 (2017).
71. Tran, F. H. & Zheng, J. J. Modulating the wnt signaling pathway with small molecules. *Protein Sci.* **26**, 650–661 (2017).
72. 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**, 32 (2019).

Methods

Expansion and differentiation of CD34⁺ cells. Human CD34⁺ cells, isolated from peripheral blood of granulocyte colony-stimulating factor-mobilized healthy volunteers, were purchased from the Fred Hutchinson Cancer Research Center. The cells were maintained and differentiated as previously described[28,73]. Briefly, the cells were expanded in StemSpan medium (Stem Cell Technologies) supplemented with StemSpan CC100 cytokine mix (Stem Cell Technologies) and 2% penicillin/streptomycin for a total of 6 days. After 6 days of expansion, the cells were stimulated for 2 h with rhBMP4 (R&D) at a final concentration of 25 ng ml^{-1} and collected for use in all of the experiments corresponding to the d0 time point. For studying differentiated cells after d6 of expansion, cells were reseeded in differentiation medium (StemSpan SFEM medium with 2% penicillin/streptomycin, 20 ng ml^{-1} SCF, 1 U ml^{-1} Epo, 5 ng ml^{-1} IL-3, 2 mM dexamethasone and 1 mM β-estradiol), at a density of 0.5–1 × 10^6 cells ml^{-1}. Before collecting the cells at h2, h6 and d1–d8, they were treated with 25 ng ml^{-1} hrBMP4 for 2 h.

For testing the effects of BMP4 and dorsomorphin, cells at the beginning of the third day of differentiation were treated with either 25 ng ml^{-1} hrBMP4 or 20 µM dorsomorphin until the beginning of the fifth day of differentiation. At d5, cells were isolated for flow cytometry and qPCR analysis. Cells treated with dimethylsulfoxide were used for control experiments.

Flow cytometry analysis. Control and treated stage-matched CD34⁺ cells or CD34⁺ cells at different stages of differentiation were washed in PBS and stained with propidium iodide, 1:60 APC-conjugated CD235a (eBioscience, clone HIR2, 17-9987-42), 1:60 FITC-conjugated CD71 (eBioscience, OKT9, 11-0719-42), 1:60 PE-conjugated CD41a (eBioscience, HIP8, 12-0419-42) and 1:60 PE-conjugated CD11b (eBioscience, ICRF44, 12-0118-42). A BD Bioscience LSR II flow cytometer was used to record raw FACS data, which were analyzed subsequently using FlowJo (v10.3).

Next-generation sequencing. Methodologies for all of the massively parallel sequencing assays (ChIP–seq, RNA-seq and ATAC–seq) are described in the Supplementary Note. Overall quality control of each dataset is presented in Supplementary Table 9. Supplementary Table 10 describes counts and the genomic span of individual TF-bound regions along with counts of associated genes, as obtained from ChIP–seq and RNA-seq data. The ChIP–seq and ATAC–seq peaks/enriched regions obtained from d0, h6, d3, d4 and d5 are shown in Supplementary Tables 11–15.

qPCR analysis. RNA was extracted from CD34⁺ cells without any treatment or treated with hrBMP4 or dorsomorphin at the specified developmental stages using TRIZOL extraction (Invitrogen), followed by RNeasy column purification (QIAGEN). First-strand complementary DNA synthesis was performed using the Superscript VILO kit (Invitrogen) and equivalent amounts of starting RNA from all samples. The cDNA was analyzed with the Light Cycler 480 II SYBR green master mix (Applied Biosystems), and the QuantStudio 12K Flex (Applied Biosystems). All samples were prepared in triplicate. The PCR cycle conditions used were: 95 °C for 5 min, (95 °C for 10 s, 54 °C for 10 s, 72 °C for 15 s) × 40 cycles. The analysis of the Ct values was performed using the $2^{-\Delta\Delta Ct}$ method[74]. The PCR primer pairs used can be found in Supplementary Table 16.

Generation of CRISPR clones in K562 and checking the expression with qPCR. pSpCas9(BB)-2A-GFP (PX458; a gift from F. Zhang, Addgene plasmid no. 48138)[75] was used to generate mutations at the *LHFPL2* TSC. Guide RNAs (gRNAs) were designed using the CHOPCHOP tool[76] or the CRISPR design tool from the Zhang laboratory[77]. The sequences of the gRNAs selected are schematized in Extended Data Fig. 3. The gRNAs were cloned in pSpCas9(BB)-2A-GFP (PX458) and verified by sequencing according to the instructions by Cong et al.[77]. For the generation of mutant cell lines, 20 µg of each gRNA that was cloned into pSpCas9(BB)-2A-GFP (PX458) was electroporated into K562 cells. After 48 h, single fluorescent cells were FACS-sorted into 96-well plates. Oligonucleotide sequences corresponding to individual gRNAs (to target the PU1, GATA and SMAD1 motifs) used for cloning can be found in Supplementary Table 16.

Genome editing and differentiation of HUDEP2 cells. HUDEP2 cells were cultured as previously described[78]. Cas9-expressing HUDEP2 cells in expansion cultures were transduced with sgRNAs targeting *AAVS1* as a negative control[79], *LHFPL2*, or the PU.1, GATA or SMAD1 motif in the corresponding signaling center. The same gRNAs that were validated in K562 cells were used in this experiment. At 24 h after transduction, cells were transferred to a 'growth phase' erythroid differentiation medium containing stem cell factor and doxycycline for 3 days. Puromycin was added to this medium to select for sgRNA-transduced cells. Then cells were transferred to a 'maturation phase' erythroid differentiation medium containing doxycycline for four days. After four days in this medium, an aliquot of cells was collected and processed for RNA isolation to determine *LHFPL2* expression. The remaining cells were transferred to erythroid differentiation medium without doxycycline for two days, and the erythroid differentiation status was assessed on the final day by cell surface marker staining, using anti-CD71-PeCy7 (eBioscience, no. 25-0719-42) and anti-CD235a-APC (eBioscience, no. 17-9987-42), and flow cytometry.

Identifying human blood donors with homozygous SNP alleles. Genomic DNA from CD34⁺ cells isolated from peripheral blood of individual donors was extracted using the DNeasy Blood & Tissue kit (Qiagen, 69506) per the manufacturer's protocol. The PCR amplification of each TSC region was carried out using the Q5 High-Fidelity 2× Master Mix (M0492S) and the primers used can be found in Supplementary Table 16.

siRNA-mediated *SMAD1* knockdown. *SMAD1* knockdown was performed on nucleofecting siRNA for *SMAD1* during the expansion of CD34⁺ cells (using the Amexa 4D-Nucleofector kit from Lonza, V4XP-3024, per the manufacturer's protocol). We used confirmed *SMAD1* siRNA from Dharmacon (onTARGETplus, SMARTpool, L-012723-00-0005) and a standard non-targeting siRNA as a control (D-001810-10-05). Three different treatment doses for *SMAD1* siRNA were used—25 nM, 50 nM and 100 nM. Control siRNA was used at 100 nM concentration. After confirming *SMAD1* knockdown, we differentiated CD34⁺ cells to erythroid lineage and kept them under BMP stimulation from d3 onwards. Expression of RBM38 RNA and protein was verified at d5.

Luciferase reporter assay. Firefly luciferase reporter constructs (pGL4.24) were made by separately cloning each of the alleles of interest centered in 426 nucleotides of genomic context upstream of the minimal promoter using BglII and XhoI sites. The firefly constructs (500 ng) were co-transfected with a pRL-SV40 *Renilla* luciferase construct (50 ng) into 100,000 K562 cells using Lipofectamine LTX (Invitrogen, ref: 15338-030). After 48 h, luciferase activity was measured by the Dual-Glo Luciferase assay system (Promega, ref: E2940) according to the manufacturer's protocol. At 24 h before luciferase activity measurement, cells were treated with 25 ng ml^{-1} rhBMP4. The sequences of the constructs are in Supplementary Table 16.

Electrophoretic gel mobility shift assay. G1ER and G1ER-S1FB murine hematopoietic progenitor cells[80] were differentiated for 24 h with β-estradiol and treated with doxycycline to express FLAG–SMAD1. Two hours before collecting the cell extracts, cells were treated with 25 ng ml^{-1} rhBMP4 to activate the BMP pathway. Cell extracts were made using the Pierce IP lysis buffer (Thermo Scientific, 87788) according to the manufacturer's protocol. Electrophoretic gel mobility shift assays were performed using the Lightshift Chemiluminescent kit (Thermo Scientific, 20148) according to the manufacturer's instructions. Briefly, binding reactions were performed with 10 µg protein, 20 fmol biotinylated DNA probe, 1× binding buffer, 5% glycerol, 500 ng poly(dI-dC), 50 mM KCl and 1.5 mM $MgCl_2$. Reactions were incubated for 30 min at room temperature. Cold competitor reactions contained 4 pmol non-biotinylated probe. Then the reactions were run on a 10% polyacrylamide/TBE non-denaturing gel (Bio-Rad Mini-PROTEAN Precast, 456-5034). The DNA probes used for this study can be found in Supplementary Table 16.

Identification of RBC-trait-associated SNPs and related analyses. Lists of SNPs associated with blood traits were compiled from multiple studies, as mentioned in the results section. We selected 1000 Genomes European populations (CEU, TSI, FIN, GBR and IBS) for our study and filtered for SNPs associated with MCV, HGB, RBCC, MCH, HTC, MCHC and RDW as phenotypes. In total, 1,325 lead SNPs associated with any of the above RBC parameters were obtained. Using the lead GWAS SNP for each region, to increase the likelihood of including the functional SNPs from a reported hit, we also included highly associated SNPs with the lead SNP (with LD $r^2 \geq 0.6$), which we included in the 'Lead + LD' SNP list. We selected SNPs on the basis of the LD threshold of $r^2 > 0.6$ using 1000 Genomes European populations (CEU, TSI, FIN, GBR and IBS). Only SNPs with an 'rs' identifier in dbSNP version 142 were considered. SNPs can have multiple allele pairs that show differential association with traits. To account for this possibility, we broke out each allele pair for each SNP. We removed any SNP from the analysis that has different alleles reported in the publication and in the dbSNP database. Such alleles were represented as 'NA' alleles for a given SNP. Only allele pairs that had two non-NA alleles were designated as 'usable alleles' and were retained for the final analysis. Accordingly, 29,069 lead and LD SNPs with at least 2 usable alleles, across 924 loci associated with the 7 RBC traits, were used to initiate the study. Unless otherwise reported, numbers of SNPs reported refer to the positions of SNPs (that is, two allele pairs of the same SNP are reported once). We used the approach and criteria from Astle et al. (2016)[19] for selecting the platelet-trait-associated GWAS SNPs to use as negative controls. RBCs and platelets share origins from megakaryocyte and erythroblast progenitor cells, suggesting platelet-trait SNPs as the ideal negative control for our study. We used 786 quantitative trait loci regions associated with 575 lead and 22,158 lead + LD platelet-trait SNPs (LD $r^2 \geq 0.6$) with at least 2 usable alleles. The positions of these SNPs relative to the hg19 revision of the human reference genome were taken from the UCSC Genome Browser track containing dbSNP version 142. Fine-mapped SNPs for blood traits were downloaded from Ulirsch et al. (2019)[23] and were converted to BED format for downstream analyses using reported positions. Fine-mapped SNPs were filtered

for those with a PP > 0.01, which was the threshold used in the initial publication of these trait-associated SNPs, resulting in 54,255 SNP-trait associations and 39,822 SNPs with unique positions and identifiers, and that are associated with at least 1 trait. SNP–enhancer or SNP–TSC overlap was determined using bedtools intersect. SNP–motif hit overlap was determined using bedtools intersect. The lists of all the SNPs that fall within overall enhancers and within TSCs are available in Supplementary Table 5.

To predict whether either allele of a given SNP was likely to be bound by a TF of interest, we built sequences containing either allele in context. Each allele for each SNP passing the above filters was used to create short, generally 41-nt-long DNA fragments that contain hg19 reference genome sequence upstream and downstream of the SNP position (that is 20 nt of reference sequence upstream, one allele of the SNP, 20 nt of reference sequence downstream). Alleles of variants called as SNPs that were greater than 1 base pair (bp) in length generated sequences longer than 41 nt, but the vast majority of short sequences were 41 nt. Each ~41-nt sequence was scanned for the presence of predicted TF-binding sequences using FIMO 4.11.4 (ref. [81]) with a reference motif library that included multiple motif position weight matrices (PWMs). On the basis of our lists of STFs and MTFs, we identified all non-redundant PWMs from the CIS-BP database build 2.00 (ref. [65]) that had been inferred from PBM analysis and the method of systematic evolution of ligands by exponential enrichment. These PWMs learned from in vitro experiments were selected to focus on direct TF binding (versus motifs inferred from, for example, ChIP–seq, which may include information about tethering TFs). We used this set of PWMs as our motif dictionary for FIMO scans of open, non-exonic regions for identifying motif hits, and this list is available as Supplementary Table 6. Motif hits that overlapped the SNP position in the 41-nt sequence were retained and used for comparison between risk and reference alleles (that is, the SNP was required to overlap the motif hit). Thus, we also required that, for a SNP to be associated with a motif hit, the motif hit directly overlap the center of the region (that is, the SNP's position). The construction of 41-bp sequences centered on the SNP itself allowed for the SNP to appear at the extreme ends of longer motifs, such as motifs from heterodimeric TF binding. Unique SNP IDs were the unit used for counting.

To test whether our H3K27ac ChIP–seq/ATAC–seq-based approach enriches for 'functional' SNPs, we used RegulomeDB (ref. [62]). A RegulomeDB score ≤ 4 was used to predict SNPs with the minimal functional evidence. This resulted in 5,695 RBC SNPs out of the total 29,069 SNPs with 2 usable alleles.

Motif occurrence identification. Positions of predicted motif occurrences were determined across the hg19 revision of the human reference genome using FIMO (ref. [81]) with default parameters and a position weight matrix reference library built as described above. The numbers of base pairs contained within each category of motif occurrence were calculated after collapsing all occurrences of either STFs motifs or MTF motifs using bedtools merge[82]. SNPs overlapping motif occurrences were determined using bedtools intersect.

Determining significance of enrichment in SNPs. To determine the relative enrichment of SNPs in pairs of region types when accounting for the collective size of regions, we used multiple statistical analyses, including 2 × 2 chi-square tests and permutation tests.

The 2 × 2 chi-squared tests compared the numbers of SNPs falling into two categories and the number of base pairs in the collective region type after collapsing. Note that the 2 × 2 chi-squared tests assume that observations are independent, which is not always the case in this biological system, especially when multiple SNPs in LD with each other are interrogated. Hence, we performed additional simulation analysis to determine the significance of our observations.

SNP position permutation tests were performed using 10,000 iterations of SNPs from the 3 lists described above (Lead, Lead + LD, fine-mapped) shuffled randomly within specified region types using bedtools shuffle.

To determine the enrichment of SNPs in STF motif hits in enhancers using SNP position permutation, SNPs were randomly shuffled in all enhancers as defined above (bedtools shuffle -incl), and the resulting positions were used to construct 41-bp sequences that were scanned by FIMO as described above for STF motif occurrences in either allele (described in detail above). Shuffled SNPs that fell within enhancers were interrogated for whether the sequences they created are likely motif occurrences for STFs or MTFs, and occurrence-overlapping SNPs were counted. The corresponding *P* value represents the number of random permutations that meet or exceed the actual observed count.

To determine the enrichment of SNPs in TSCs versus non-TSC enhancers using SNP position permutation, SNPs were randomly shuffled within enhancers as defined above and interrogated for whether they overlap the subset of enhancers defined as TSCs. The corresponding *P* value represents the fraction of 10,000 random permutations that meet or exceed the actual observed count.

To determine the enrichment of SNPs in STF motif hits within TSCs versus STF motif hits within non-TSC enhancers using SNP position permutation, SNPs were randomly shuffled within enhancers as defined above, and interrogated for whether they fall within TSCs, and whether they are predicted to fall within motif occurrences at their original (read: not shuffled) position. The corresponding *P* value represents the fraction of 10,000 random permutations that meet or exceed the actual observed count.

Enhancer labeling permutation tests were performed by selecting a random subset of enhancers to represent TSCs to test whether SNPs are unusually concentrated in actual TSCs above background. Note that the number of observed successes differs in this approach from that of above, as the SNP position permutation analysis used both alleles of each SNP to determine whether the sequence created during shuffling was recognizable by specified TFs. A total of 7,421 of 81,636 enhancers across the system were randomly selected each of 10,000 iterations using the Unix utility shuf. The numbers of trait-associated SNPs from each of the three lists that are contained in each random TSC subset were tallied. The corresponding *P* value represents the number of random permutations that meet or exceed the actual observed count.

Motif hit permutation tests were performed by randomly shuffling the positions of unambiguous STF motif hits that fall within enhancers across all enhancer loci using bedtools shuffle -incl. Note that the number of observed successes differs in this approach from that of above, as the SNP position permutation analysis used both alleles of each SNP to determine whether the sequence created during shuffling was recognizable by specified TFs. The corresponding *P* value represents the number of random permutations that meet or exceed the actual observed count of SNPs in their real position overlapping permuted STF motif hits.

Expression analysis from FHS. Minor allele frequencies in groups of different ancestries were looked up from Hapmap CEU, YRI or CHB population data through http://snp-nexus.org/ (refs. [83–85]). eQTLs were queried using R or Perl scripting based on our selected SNP lists from the dataset downloaded from https://grasp.nhlbi.nih.gov/Updates.aspx (ref. [86]; GRASP 2.0.0.0 Expression QTLs), and the dataset downloaded from the FHS population (FHS whole-blood eQTL results) ftp://ftp.ncbi.nlm.nih.gov/eqtl/original_submissions/FHS_eQTL/ (refs. [70,87]). For FHS whole-blood eQTL results, we focus only on significant eQTLs (peer-validated results up to a log[FDR] value of −4.0, at the levels of genes and exons, respectively), and report the *cis*-eQTL with the best *P* value in each region, or all of the significant *cis*- and *trans*-eQTLs for our selected SNPs as a reference.

Statistical analysis. The detailed methodologies used for the statistical tests and the resulting significance values obtained comparing the control and test groups are described in the relevant methods sections, figures and figure legends and in Supplementary Table 3. Biological replicates and observed data-point variations are mentioned wherever applicable. All statistical analyses were carried out using the statistical computing/graphics software R and GraphPad Prism 8.

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

Data availability

The massively parallel sequencing data associated with this manuscript have been uploaded to GEO under the accession numbers GSE74483 and GSE104574 and are currently open to the public. The web links for the publicly available databases used in this study are: UniPROBE, http://thebrain.bwh.harvard.edu/uniprobe/; CIS-BP, http://cisbp.ccbr.utoronto.ca/; FHS, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v30.p11; RegulomeDB, https://regulomedb.org/regulome-search/; HEMMER, http://hmmer.org/; EMBOSS Needle, https://www.ebi.ac.uk/Tools/psa/emboss_needle/; dbSNP, https://www.ncbi.nlm.nih.gov/snp/?cmd=search. Links to all of the PBM datasets used are available in Supplementary Table 7. Source data are provided with this paper.

Code availability

The custom codes used in this study are available at https://bitbucket.org/abrahamb/workspace/projects/TSC. The code and data files for the PBM analyses are available at https://github.com/BulykLab/RBCSNPs_2020.

References

73. Sankaran, V. G., Orkin, S. H. & Walkley, C. R. *Rb* intrinsically promotes erythropoiesis by coupling cell cycle exit with mitochondrial biogenesis. *Genes Dev.* **22**, 463–475 (2008).
74. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. *Methods* **25**, 402–408 (2001).
75. Ran, F. A. et al. Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
76. Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M. & Valen, E. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Res.* **42**, W401–W407 (2014).
77. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
78. Vinjamur, D. S. & Bauer, D. E. Growing and genetically manipulating human umbilical cord blood-derived erythroid progenitor (HUDEP) cell lines. *Methods Mol. Biol.* **1698**, 275–284 (2018).

79. Canver, M. C. et al. Integrated design, execution, and analysis of arrayed and pooled CRISPR genome-editing experiments. *Nat. Protoc.* **13**, 946–986 (2018).
80. Gregory, T. et al. GATA-1 and erythropoietin cooperate to promote erythroid cell survival by regulating bcl-x_L expression. *Blood* **94**, 87–96 (1999).
81. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
82. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
83. Chelala, C., Khan, A. & Lemoine, N. R. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics* **25**, 655–661 (2009).
84. Dayem Ullah, A. Z., Lemoine, N. R. & Chelala, C. SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic Acids Res.* **40**, W65–W70 (2012).
85. Dayem Ullah, A. Z., Lemoine, N. R. & Chelala, C. A practical guide for the functional annotation of genetic variations using SNPnexus. *Brief. Bioinform.* **14**, 437–447 (2013).
86. Leslie, R., O'Donnell, C. J. & Johnson, A. D. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* **30**, i185–i194 (2014).
87. Splansky, G. L. et al. The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *Am. J. Epidemiol.* **165**, 1328–1335 (2007).

Acknowledgements

We are grateful to S. Orkin, A. Wagers and C. Santoriello for the critical reading and editing of our manuscript. We thank the HHMI high-throughput sequencing facility at the Children's Hospital Boston for generating the genome-wide raw sequencing data. We acknowledge V. G. Sankaran and S. Nandakumar for providing the luciferase plasmids containing non-coding alleles of the *RBM38* gene and the enhancer mutant for the *RBM38* gene in K562 cells. This work was supported by the following grants to L.I.Z.—R01 HL04880, P015PO1HL32262-32, 5P30 DK49216, 5R01 DK53298, 5U01 HL10001-05, R24 DK092760, 1R24OD017870-01. Additional support came from the funding by the Max Planck Society, The Fritz Thyssen Stiftung (Az 10.17.1.026MN), a Marie Curie Career Integration Grant (631432 Bloody Signals), the Deutsche Forschungsgemeinschaft DFG under Germany's Excellence Strategy (CIBSS-EXC-2189-Project-ID-390939984) and the Deutsche Forschungsgemeinschaft, Research Training Group 322977937/GRK2344 'Me*In*Bio –Bio*In*Me' to E.T., the Hope Funds for Cancer Research Grillo-Marxuach Family Fellowship and the American Lebanese Syrian Associated Charities to B.J.A. R.A.Y. is supported by NIH grants GM123511, CA213333 and CA155258. K.H.K. is supported by an A*STAR National Science Scholarship. M.L.B. is supported by the NIH grant R21 HG010200. S.K.G. is supported by R01HL139672, R01HL122684 and R01HL086694. D.E.B. was supported by the NHLBI (P01HL32262 and DP2HL137300). The FHS is funded by National Institutes of Health contract N01-HC-25195. The eQTL work for this investigation was funded by the Division of Intramural Research, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD (D. Levy, Principal Investigator). The analytical component of this project was funded by the Division of Intramural Research, National Heart, Lung, and Blood Institute, and the Center for Information Technology, National Institutes of Health, Bethesda, MD. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the US Department of Health and Human Services.

Author contributions

A.C. and E.T. designed and performed the experiments. B.J.A., L.M.C., K.H.K., W.M., S.Y., B.H., A.L., M.S. and S.N.G. performed bioinformatics and data analysis. T.V.B., A.S., D.S.V. and A.G. helped design strategies for key experiments, including CRISPR-mediated mutations. R.J. and M.-L.Y. carried out the FHS data analysis. K.H., V.C., S.B. and S. Tseng performed supervised experiments. S. Takahashi provided the PU.1-knockdown and PU.1-overexpressing K562 cell lines. S.K.N. provided the luciferase plasmids containing non-coding alleles of the *RBM38* gene and the enhancer mutant for the *RBM38* gene in K562 cells. Y.Z., A.B.C., S.K.G., J.R., D.E.B., P.S.A., S.J.C., M.L.B. and R.A.Y. provided insights on the analysis and interpretation of data. A.C., E.T., B.J.A., L.M.C., K.H.K., A.S. and L.I.Z. wrote and revised the manuscript. L.I.Z. supervised the study. All authors edited the manuscript.

Competing interests

L.I.Z. is a founder and stockholder of Fate Therapeutics, CAMP4 Therapeutics, Amagma Therapeutics, and Scholar Rock. He is a consultant for Celularity and Cellarity. R.A.Y. is a founder and shareholder of Syros Pharmaceuticals, Camp4 Therapeutics, Omega Therapeutics and Dewpoint Therapeutics. B.J.A. is a shareholder in Syros Pharmaceuticals. M.L.B. is a co-inventor on patents on PBM technology. The other authors declare no competing interests.

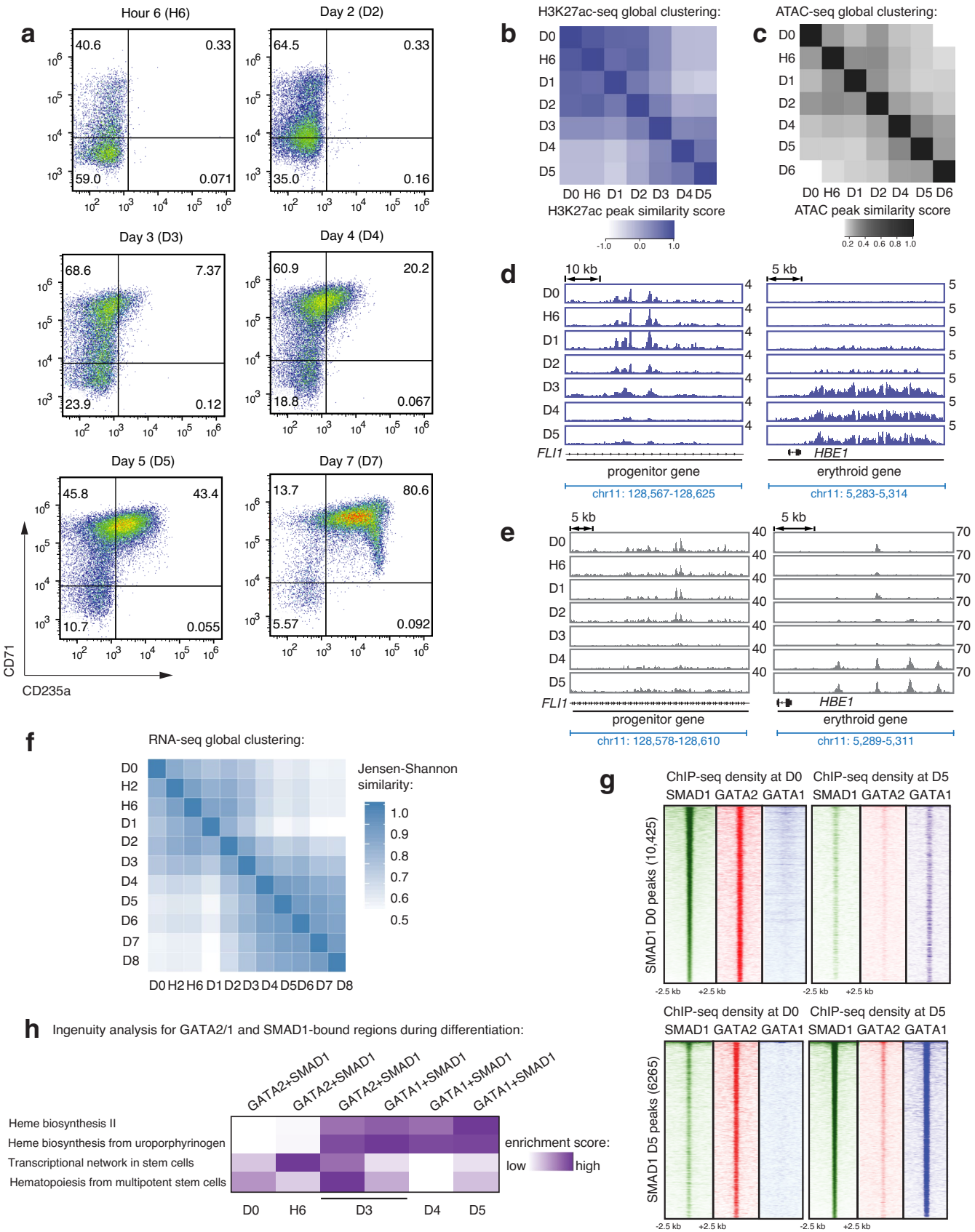
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-020-00738-2.

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-00738-2.

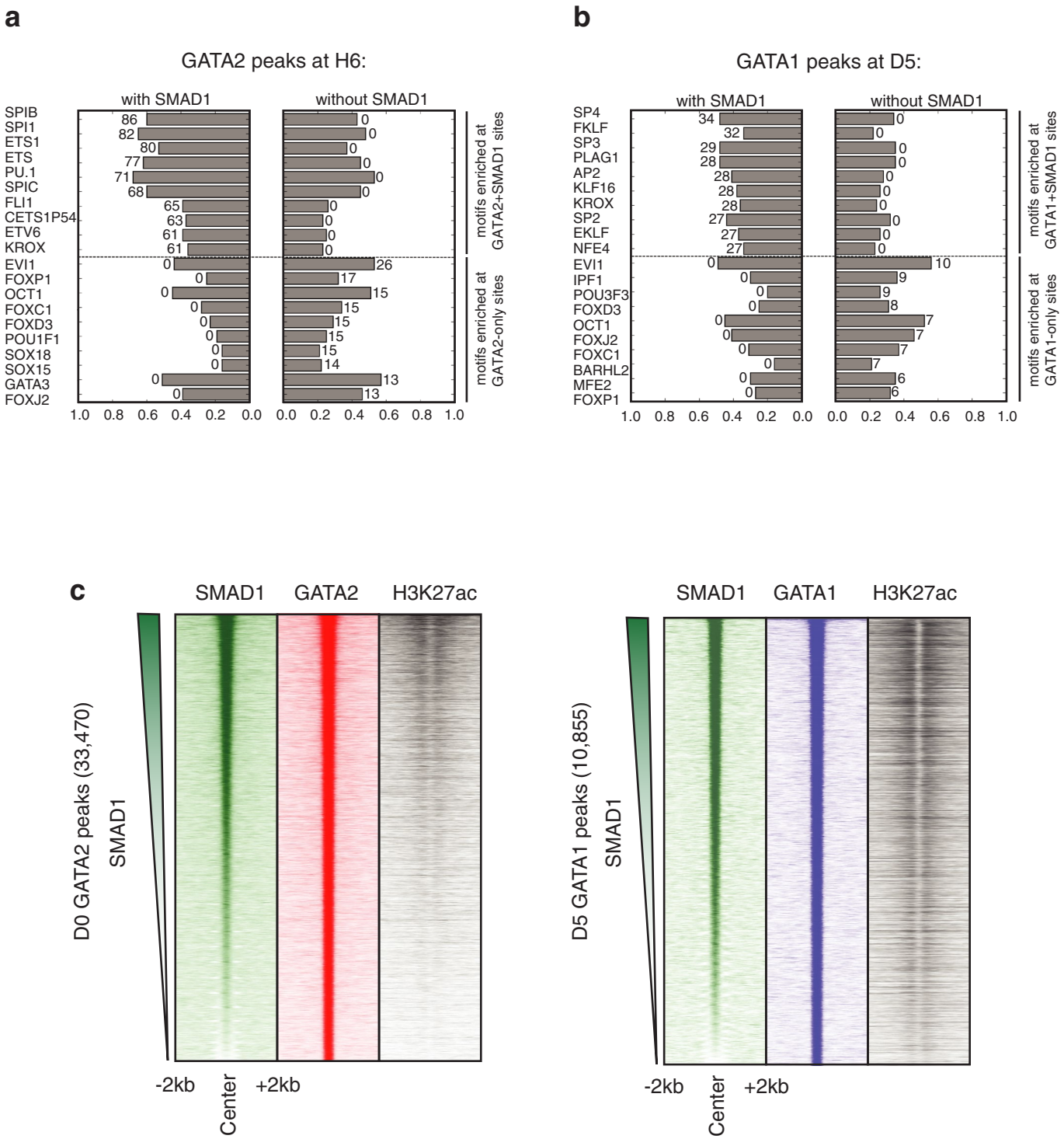
Correspondence and requests for materials should be addressed to L.I.Z.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work.



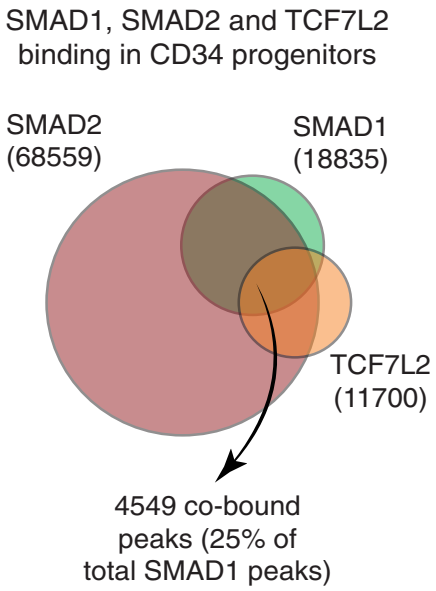
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Human CD34+ cells commit to an erythroid fate around day 3 (D3) of differentiation. a, Representative FACS plots for the erythroid markers CD71 and CD235a on CD34+ cells after induction of erythroid differentiation at hour 6 (H6), day 2 (D2), day 3 (D3), day 4 (D4), day 5 (D5) and day 7 (D7). **b-c**, Heatmaps depicting correlation of peaks from H3K27ac ChIP-seq or ATAC-seq obtained from distinct differentiation time-points. **d-e**, Gene tracks showing H3K27ac ChIP-seq signal (d) or ATAC-seq signal (e) at *FLI1* and at the β-globin locus control region at different differentiation stages. D0 = progenitor CD34+ cells before induction of differentiation; H6 = 6 hours after differentiation; and D1 through D5 = 1, 2, 3, 4 and 5 days after differentiation. **f**, Heatmap depicting correlation of gene expression profiles of all protein-coding RNAs from D0 through D8 of erythroid differentiation. D0 = progenitor CD34+ cells before induction of differentiation; H2 and H6 = 2 and 6 hours after differentiation; and D1 through D8 = 1, 2, 3, 4, 5, 6, 7 and 8 days after differentiation. **g**, Signal heatmaps comparing ChIP-seq read densities of SMAD1, GATA2, and GATA1 at SMAD1 peaks identified at D0 (upper panel) and D5 (lower panel). Signal intensities centered around +/− 2.5 kb shown. **h**, Ingenuity pathway analysis (IPA) for GATA2+SMAD1 bound genes at D0, H6, D3 and D4 and GATA1+SMAD1 bound genes at D3, D4, D5 identifying differentiation stage-specific biological properties.



Extended Data Fig. 2 | Comparative TF motif enrichment and H3K27ac signal density analysis surrounding GATA+SMAD1 versus GATA-only regions. Bar charts depicting the enrichment of transcription factor motif hits at regions co-bound by GATA+SMAD1 (left) versus by GATA only (right) at H6 **(a)** and D5 **(b)**. Length of the bar indicates the fraction of peaks containing a given motif hit, and the number associated with the bar represents the corresponding -log10(p-value) obtained from the one-tailed hyper-geometric test to assess the significance of motif enrichment. For both (a) and (b), top and bottom of the ranked lists are shown. **c**, (left panel) Region heatmaps depicting signal of ChIP-seq reads for D0 SMAD1, GATA2 and H3K27ac at 33,470 GATA2 bound peaks at D0. Peaks are ranked by the SMAD1 intensity across the row. Each plot represents signal intensities around +/− 2 kb of the peak center. (right panel) Region heatmaps depicting signal of ChIP-seq reads for D5 SMAD1, GATA1 and H3K27ac at 10,855 GATA1 bound peaks at D5. Peaks are ranked by the SMAD1 intensity across the row. Each plot represents signal intensities around +/− 2 kb of the peak center.

a

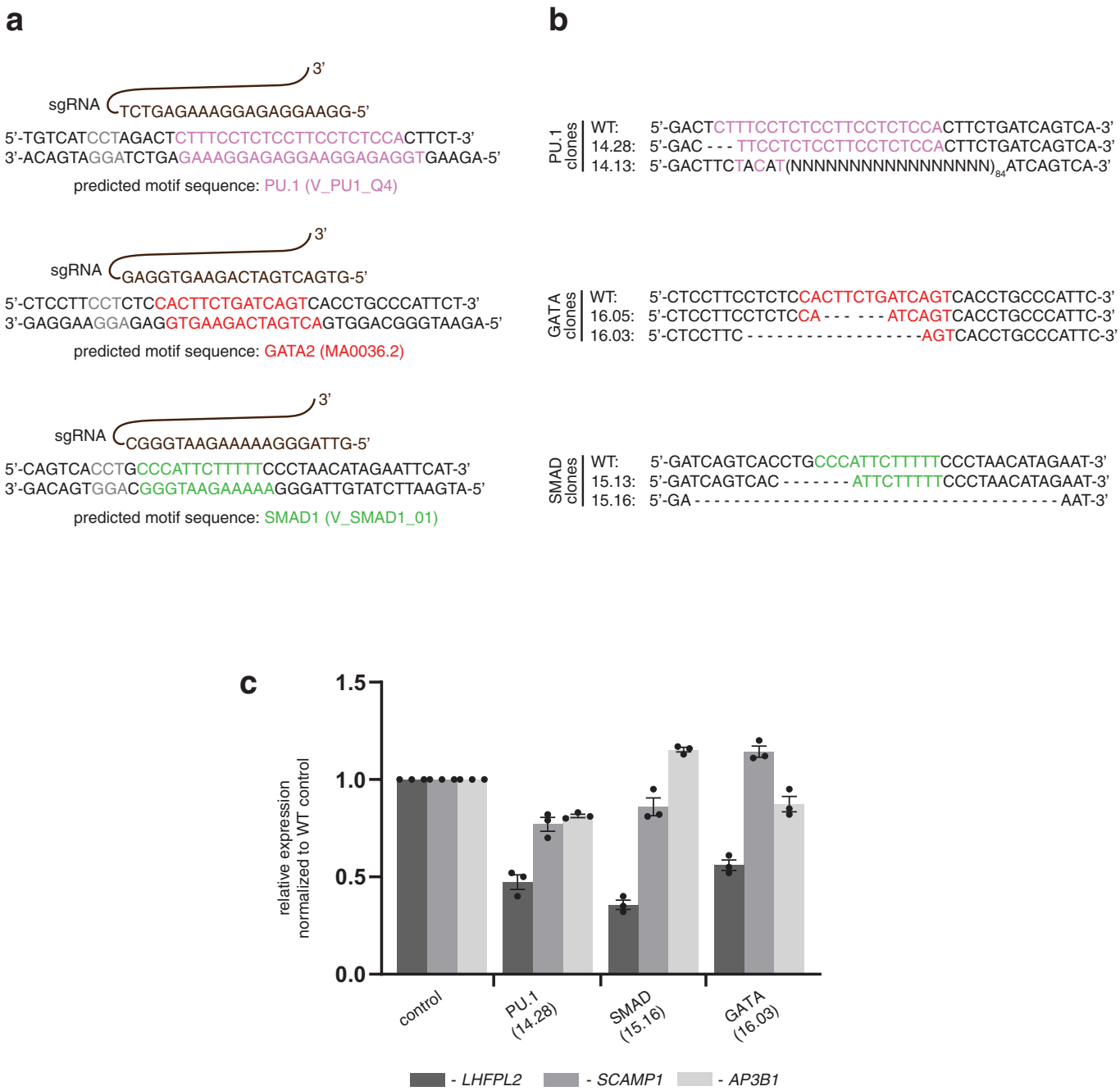


b

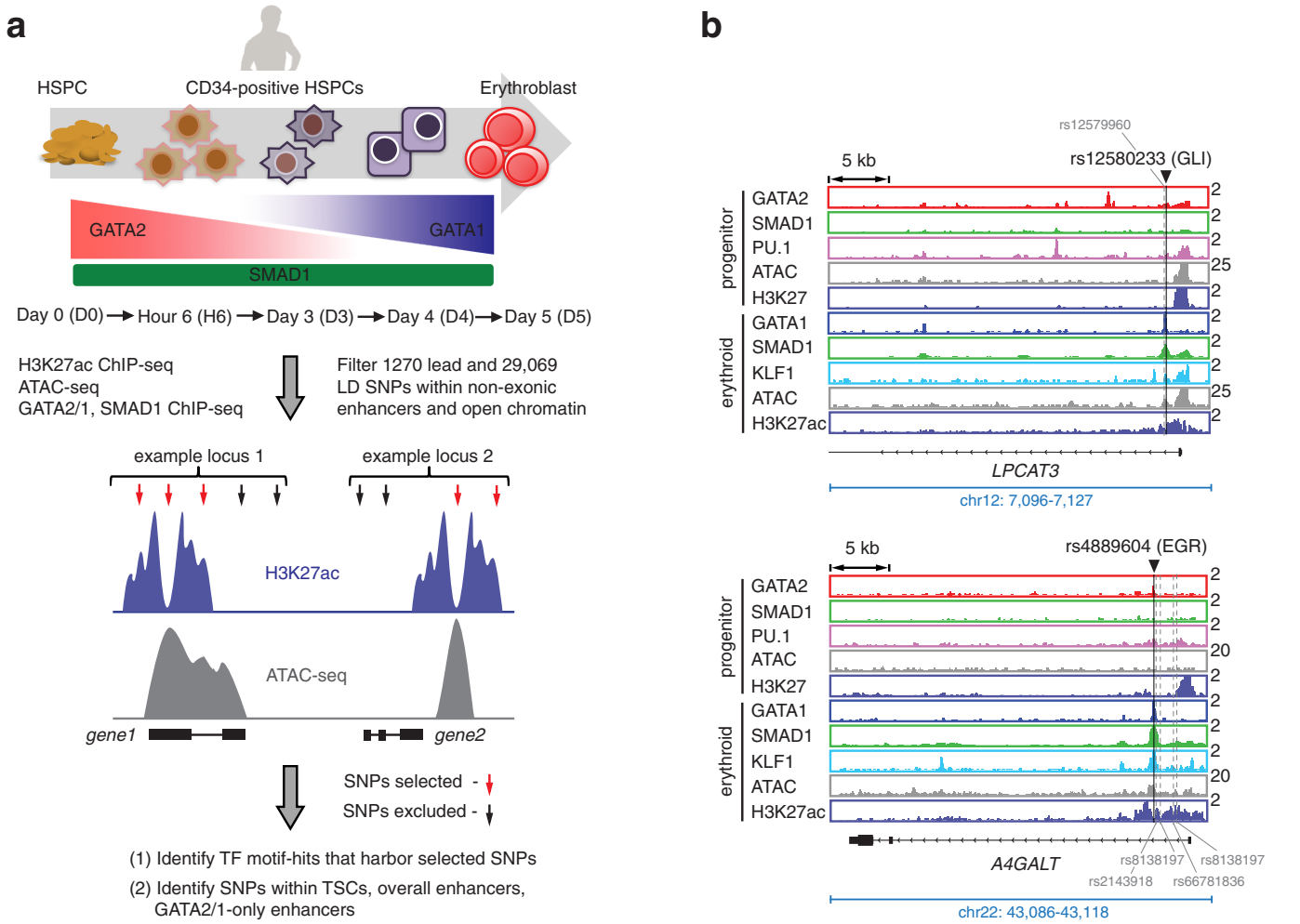
Percentage of SMAD1+GATA co-bound enhancers (TSCs) out of total active enhancers:

Enhancer Definition	Stage	# sites	TSC (GATA+SMAD1)	% TSC
H3K27ac peaks	D0	24497	1257	5.1
	D5	54756	3963	7.2
Non-promoter H3K27ac peaks	D0	12427	1091	8.8
	D5	39747	3602	9.1
ATAC/H3K27ac peak intersection	D0	17563	875	5.0
	D5	26223	2872	11.0
Non-promoter ATAC/H3K27ac peak intersection	D0	6999	813	11.6
	D5	12625	2735	21.7
ATAC/H3K27ac peak union	D0	50413	4266	8.5
	D5	60776	4063	6.7
Non-promoter ATAC/H3K27ac peak union	D0	37443	4063	10.9
	D5	48068	3697	7.7

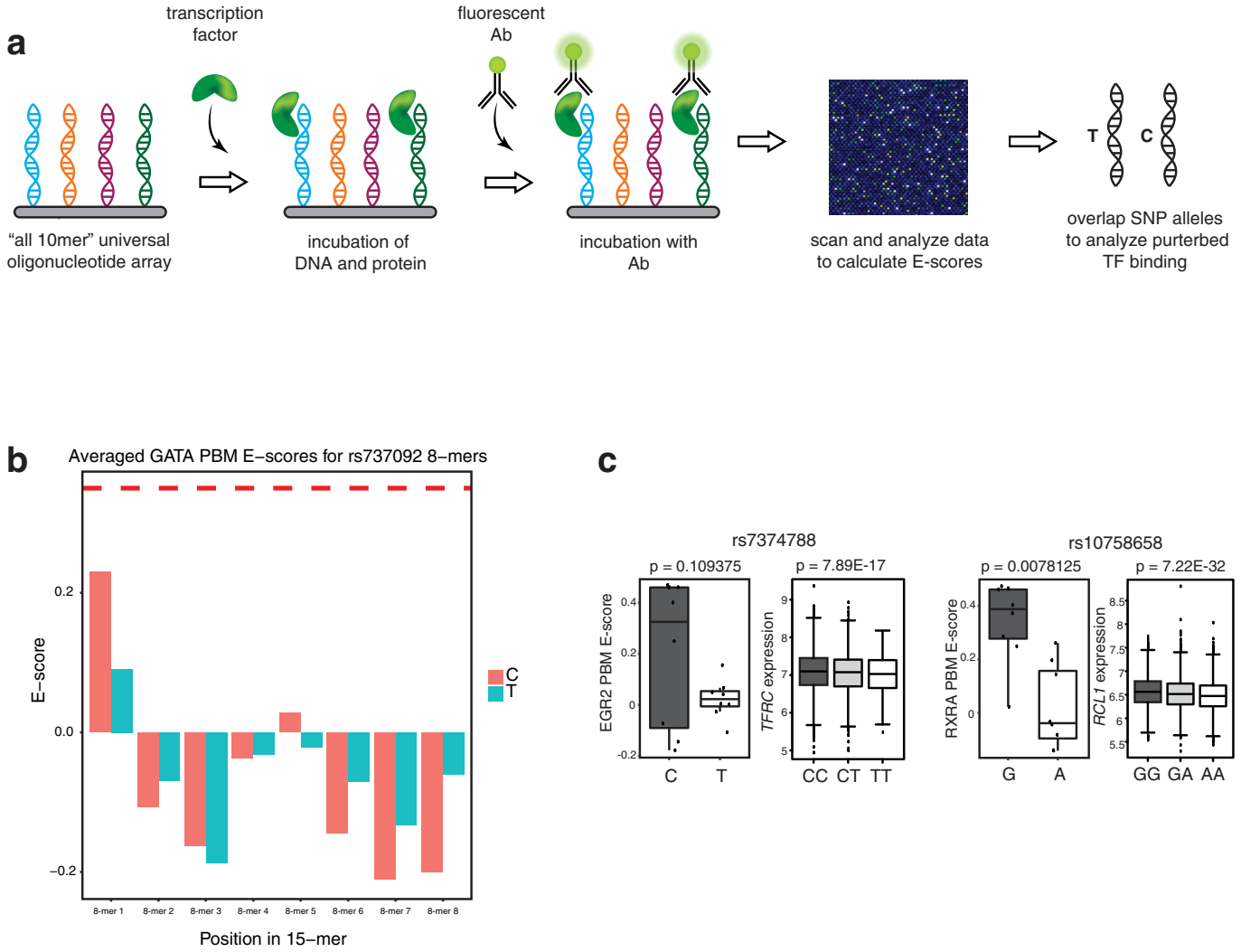
Extended Data Fig. 3 | TSCs are a small subset of overall enhancers as defined by SMAD1 and GATA co-occupancy. a, Venn diagram representing genomic regions co-occupied by different STFs - SMAD1, SMAD2 and TCF7L2 in progenitor CD34+ cells upon stimulation with BMP4, TGFβ and WNT signaling, respectively. The genomic regions bound by all three factors are 4549. The other numbers refer to the total number of peaks bound by each factor combination, as indicated. Regions are considered occupied if they pass a significant coverage cutoff. **b**, Table representing different strategies to define enhancers using H3K27ac ChIP-seq and/or ATAC-seq. The proportion of enhancers that can be classified as TSCs (GATA+SMAD1 co-bound) at progenitor (D0) or erythroid (D5) stages are as indicated.



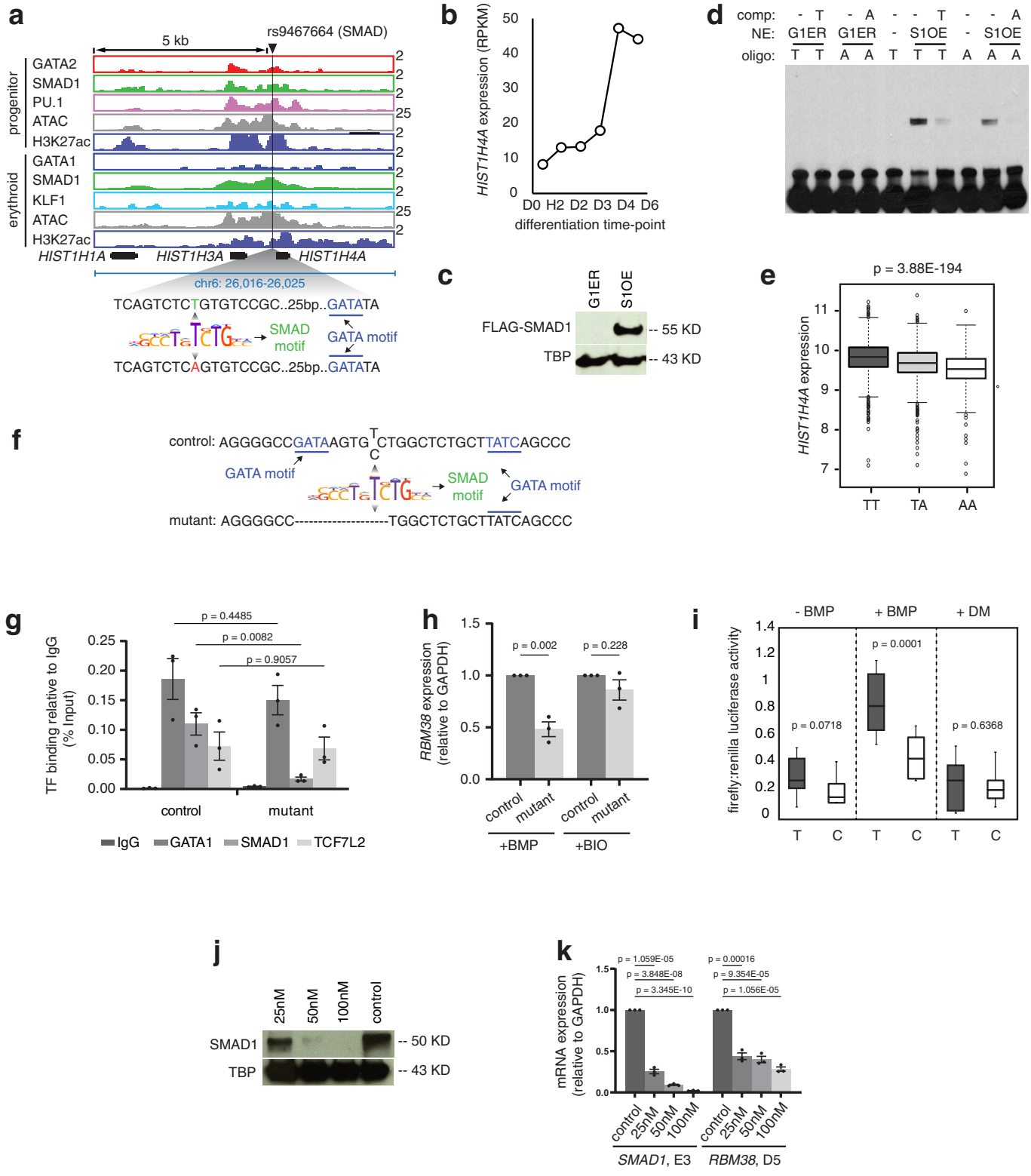
Extended Data Fig. 4 | Dissecting a TSC near the *LHFPL2* gene by CRISPR-Cas9. a, sgRNAs (shown in brown) targeting specific sequences near PU.1 (pink), GATA (red) and SMAD1 (green) motif-hits within the TSCs. **b**, Genomic sequences of the specific CRISPR-edited clones are compared against wild-type genomic sequence. Clones 14.13, 16.03 and 15.16 appear to target multiple motifs (that is PU.1 and partial GATA; partial PU.1 and partial GATA; SMAD1 and partial GATA). **c**, qPCR results depicting relative expression of *LHFPL2* (black bar) and two other flanking genes (*SCAMP1*, grey bar and *AP3B1*, white bar) are shown in different CRISPR clones compared to the WT K562 cells, as indicated. Mean ± SEM shown. ($n = 3$; 3 biologically independent experiments). Two-sided student t-tests used.



Extended Data Fig. 5 | Approach for interrogating enhancer-associated RBC-trait SNPs showing that SNPs targeting STF motif-hits are localized within TSCs. a, Schematic diagram of the strategy used to identify SNPs that may alter activity of transcriptional enhancers during human erythroid differentiation. Human CD34+ cells from mobilized peripheral blood were differentiated towards erythrocytes. Genomic experiments were performed at D0, H6, D3, D4 and D5. 1270 lead RBC-trait SNPs and additional SNPs that are in linkage disequilibrium with lead SNPs, with LD score $r^2 \geq 0.6$ (total number of SNPs = 29,069), were first overlapped with genomic regions that are defined as non-exonic enhancer (represented as violet tracks) and open chromatin peaks (represented as grey tracks) in our study. SNPs that fall within such regions (indicated with red arrows) were used to carry out motif hit analysis, and were overlapped either with TSCs, or overall enhancers or GATA-only enhancers. **b**, Gene tracks showing RBC-trait SNPs that are located within stage-specific TSCs are shown. The binding of GATA2, GATA1, SMAD1, PU.1 and KLF1 and the peaks of H3K27ac and ATAC-seq are shown in progenitor and differentiated stages. Black lines indicate the positions of representative SNPs. The potential STF motifs that these SNPs could perturb (for example GLI, EGR) are as indicated. For each representative SNP that resides in a TSC, the other associated SNPs in significant LD that fall within H3K27ac/ATAC-positive enhancers are also indicated with grey dashed lines.

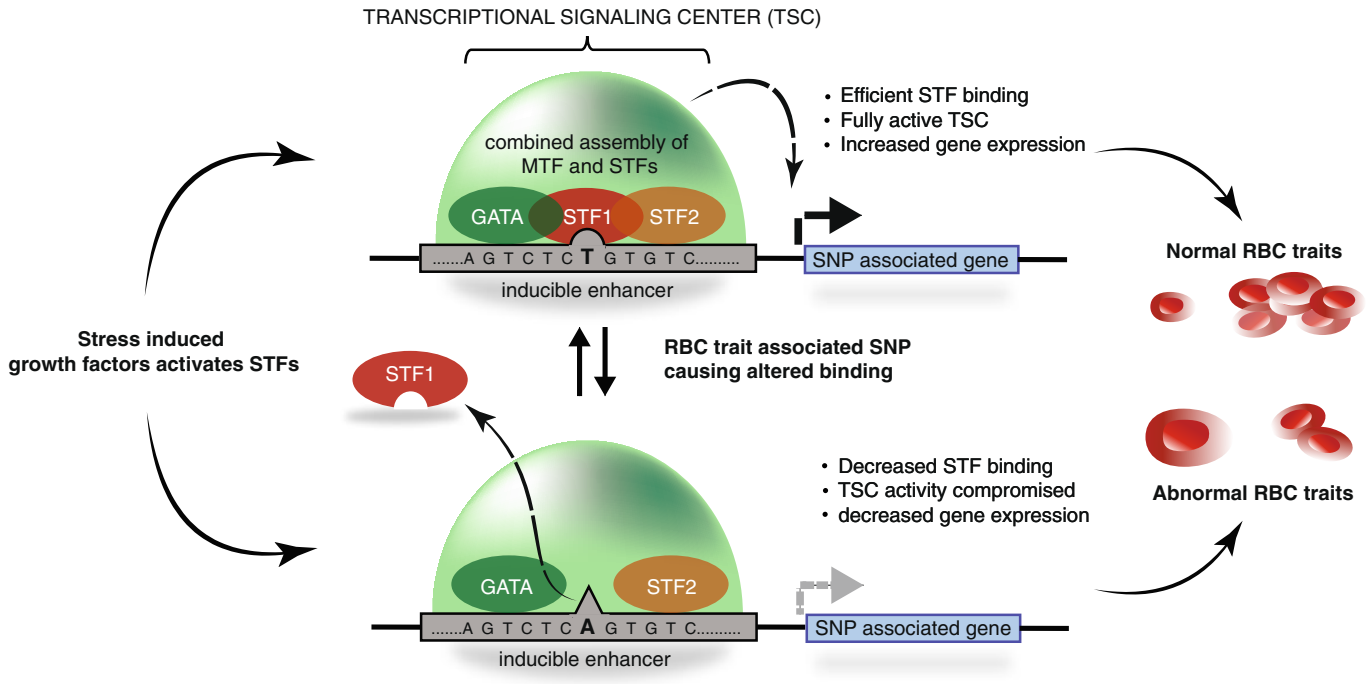


Extended Data Fig. 6 | Analysis of PBM 8-mer data identifies several RBC trait-associated SNPs that perturb STF-DNA binding. a, Schematic representation of the strategy to identify SNPs that alter STF binding utilizing protein binding microarrays. **b**, Bar charts for the GATA average PBM dataset for rs737092. The p-value (0.5469) is computed using two-sided Wilcoxon signed-rank test. **c**, Additional examples of SNPs showing perturbed TF binding of indicated STFs from PBM analysis (left) and corresponding distribution of expression values of the most significantly altered nearby gene in homozygous and heterozygous individuals obtained from FHS eQTL analysis (right). For PBM analysis, the P-values are computed using two-sided Wilcoxon signed-rank tests. Individual genotypes and the cis-eQTL gene/exon obtained from the FHS dataset are as indicated. Two-sided test with linear model for EffectAlleleDosage used for eQTL analysis with P-values adjusted using the Benjamini-Hochberg procedure. The lower and upper bounds of the box correspond to the 25th and 75th percentiles, respectively. The upper whisker extends from the upper bound of the box to the largest value no further than 1.5 * inter-quartile range (IQR) away (the IQR is defined as the distance between the 25th and 75th percentiles). The lower whisker extends from the lower bound of the box to the smallest value no further than 1.5 * IQR away.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | RBC-trait SNPs perturb STF-DNA binding. a, TSC containing rs9467664 at SMAD motif near *HIST1H4A*. **b**, *HIST1H4A* expression during CD34+ differentiation. **c**, Western blot showing the expression of FLAG-SMAD1. TBP is loading control. ($n = 3$; 3 biologically independent experiments). **d**, Representative gel-shift assay with A or T allele of rs9467664. ($n = 3$; 3 biologically independent experiments). **e**, *HIST1H4A* eQTL analysis for rs9467664: boxplots represent median *HIST1H4A64* expression as the thickest line, the first and third quartile as the box, and 1.5 times the interquartile range as whiskers. Two-sided test with linear model for EffectAlleleDosage used: effect estimate (β)=0.1562; T-statistics=31.0243, $R^2 = 0.15536$; log10(P-value)=-193.41, log10(Benjamin–Hochberg's FDR) = -190.1. **f**, Schematic representation of K562 clone with altered sequence around rs737092. **g**, Alteration of TF binding in K562 mutants from f. Mean ± SEM shown. ($n = 3$; 3 biologically independent experiments). Two-sided student t-tests used. **h**, Expression alteration of *RBM38* in K562 mutants from f, under BMP and BIO treatment. Mean ± SEM shown. ($n = 3$; 3 biologically independent experiments). Two-sided student t-tests used. **i**, Luciferase assays for alternative alleles of rs737092. Boxplots represent median as the thickest line, the first and third quartile as the box, and 1.5 times the interquartile range as whiskers. Two-sided student t-tests used. **j**, Western blot comparing SMAD1 expression in control and shRNA treated CD34+ cells with indicated doses. TBP is loading control. ($n = 3$; 3 biologically independent experiments). **k**, *RBM38* expression upon SMAD1 loss. Mean ± SEM shown. ($n = 3$; 3 biologically independent experiments). Two-sided student t-tests used.



Extended Data Fig. 8 | A model proposing how human genetic variation within TSCs induces RBC trait phenotypes. A combination of STFs and MTFs drives optimal gene expression via the TSC. The normal signal-induced expression of a red blood cell gene is perturbed due to a SNP that either eliminates an existing STF binding event or creates a new STF binding site in a critical signaling center. This can lead to a lack of response to an episodic signaling event, initiated by an exogenous stressor, and eventually manifest as phenotypic variability.

nature research

Corresponding author(s): Leonard Zon

Last updated by author(s): Sep 14, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Apart from the description given below, further details, e.g. references associated with each section can be found in "METHODS" section of the combined manuscript file.
For expression analysis from Framingham Heart Study (FHS), minor allele frequencies in different ethnic groups were looked up from Hapmap CEU, YRI, or CHB population data through http://snp-nexus.org/ 86-88. Expression QTLs (eQTLs) were queried using R or Perl scripting based on our selected SNP lists from data-set downloaded from https://grasp.nhlbi.nih.gov/Updates.aspx 89(GRASP 2.0.0.0 Expression QTLs), and data-set downloaded from Framingham Heart Study population (FHS whole blood eQTL results) ftp://ftp.ncbi.nlm.nih.gov/eqtl/original_submissions/FHS_eQTL/ 40,90. For FHS whole blood eQTL results, we only focus on significant eQTLs (peer validated results up to a logFDR value of -4.0, at the levels of genes and exons respectively), and report the cis-eQTL with best p-value in each region, or all of the significant cis and trans-eQTLs for our selected SNPs as a reference.

Data analysis

Apart from the description given below, further details, e.g. references associated with each section can be found in "METHODS" section of the combined manuscript file.
Softwares/tools used for the study along with their version numbers are as follows: R (3.4.3), MACS (2.1.0), TopHat v2.0.13 70, Cufflinks (v2.2.171), HOCOMOCO (v10), IGV (v2.8.9), corrplot (v0.84), bedtools (v2.29.2), deeptools (v2.0), Bowtie (v2.2.1 and 2.2.5), CHOPCHOP (v3.0.0), GRASP (2.0.0.0), FlowJo (V10.3), microsoft excel (v15.22), Graphprism (v8), HMMER (v3.3.10), EMBOSS (v6.6.0.0)
ChIP-Seq data analysis:
Alignment and Visualization:
ChIP-Seq reads were aligned to the human reference genome (hg19) using bowtie 66 with parameters -k 2 -m 2 -S. WIG files for display were created using MACS 67 with parameters -w -S --space=50 --nomodel --shiftsize=200, were normalized to the millions of mapped reads, and were displayed in IGV 68,69. The overall quality control values, percentage occupancy of each factors at different genomic regions are mentioned in Supplementary Data Table 9. The ChIP-seq peaks/enriched regions obtained from D0, H6, D3, D4 and D5 are shown in Supplementary Data Tables 9-13.
Peak and Bound Gene Identification:
High-confidence peaks of ChIP-Seq signal were identified using MACS with parameters --keep-dup=auto -p 1e-9 and corresponding input control. Bound genes are RefSeq genes that contact a MACS-defined peak between -10000bp from the TSS and +5000bp from the TES. The bound genes associated with GATA2/1 and SMAD1 at each stage are shown in Supplementary Data Table 2.
Identifying Enhancers and Transcriptional Signaling Centers (TSCs):
Enhancers were identified using H3K27ac ChIP-Seq and ATAC-Seq peak information. Peaks were identified as described above using

MACS. Coding regions were removed from H3K27ac and ATAC-Seq peaks using bedtools subtract; coding regions were exons from all RefSeq transcripts. Non-exonic portions of ATAC-Seq-enriched regions that overlapped H3K27ac-enriched regions by at least 1 bp were retained. H3K27ac- or ATAC-Seq/H3K27ac-enriched regions outside exons were collapsed using bedtools merge. These steps were performed for each timepoint's ATAC-Seq/H3K27ac ChIP-Seq pair. D0, H6, D4, and D5 regions were collapsed and used for "enhancers" across the time-course.
Transcription signaling centers (TSCs) were defined as those that was co-bound by SMAD1 and the corresponding GATA factor. For each time-point, regions enriched in both SMAD1 and the corresponding GATA were identified using bedtools intersect on the peaks. Enhancers, as defined above, are considered TSCs if they overlap a SMAD1/GATA-bound region by at least 1 bp. TSCs identified at each of D0, H6, D4, and D5 were collapsed and used as a canonical list of TSCs across the time-course. Progenitor signaling centers represent the union of D0 and H6; erythroid signaling centers represent the union of D4 and D5 time-points.
Lists of all the enhancers and TSCs identified using above methods are listed in Supplementary Data Table 4.
Peak Similarity Heatmaps:
The called H3K27ac ChIP-seq and ATAC-seq peaks of all samples were combined to generate peak files for heatmap analysis. Depending on the overlap of the union of peaks and peak from individual sample, a binary matrix of 0 and 1 were assigned to each peak of each sample. The similarity score was derived and correlation matrix was calculated by the cor method, and the heatmap drawn by corrplot package in R.
ChIP-Seq Read Density Heatmaps/Scatterplots:
ChIP-Seq read density heatmaps were constructed using bamToGFF (https://github.com/BradnerLab/pipeline) on 4kb regions centered on the peak center with parameters -m 200 -r -d and filtered bam files with at most one read per position.
Binary peak/not-peak "heatmaps" were determined by first taking the collapsed union of peaks defined at all five timepoints and determining whether each of these collapsed regions contacted a peak in any of the timepoints.
Co-occupancy of multiple STFs upon stimulation with respective signaling pathways:
ChIP-Seq read density heatmaps were constructed using bamToGFF (https://github.com/BradnerLab/pipeline) on 4kb regions centered on the peak center. Single-TF heatmaps were built with parameters -m 200 -r -d and filtered bam files with at most one read per position; rows were ordered by the row sums of the indicated factor. Multiple-TF heatmaps were built with parameters -m 100 -r -d, and rows were ordered by the row sum in SMAD1 signal. Binary peak/not-peak "heatmaps" were determined by asking if the original peak overlapped a SMAD1-enriched region using bedtools intersect.

RNA-seq data analysis:
RNA-seq reads were mapped to the human reference genome (hg19) using TopHat v2.0.13 70 the flags: "--no-coverage-search --GTF gencode.v19.annotation.gtf" where gencode.v19.annotation.gtf is the Gencode v19 reference transcriptome available at gencodegenes.org. Cufflinks v2.2.171 was used to quantify gene expression and assess the statistical significance of differential gene expression. Briefly, Cuffquant was used to quantify mapped reads against Gencode v19 transcripts of at least 200bp with biotypes: protein_coding, lincRNA, antisense, processed_transript, sense_intronic, sense_overlapping. Cuffdiff was run on the resulting Cuffquant .cxb files, giving a table of RPKM expression level, fold change and statistical significance for each gene.
ATAC-seq data analysis:
All human ChIP-Seq datasets were aligned to build version NCBI37/HG19 of the human genome using Bowtie2 (version 2.2.1) (Langmead et al., 2012) with the following parameters: --end-to-end, -N0, -L20. Coverage files for display were created using MACS with parameters -w -S –space=50 –nomodel –shiftsize=200. We used the MACS2 version 2.1.0 (Zhang et al., 2008) peak-finding algorithm to identify regions of ATAC-Seq peaks, with the following parameter --nomodel --shift -100 --extsize 200. A q-value threshold of enrichment of 0.05 was used for all datasets. For correlation of ATAC-seq data with ChIP-seq binding, reads were mapped to the human genome (hg19) using Bowtie v2.2.5 72 with default options. BedTools 73 was used to count the number of ATAC-seq reads under Gata/Smad peaks (+/-2.5kb from peak center; 50bp bins). Read counts were normalized by library size to get CPM. The ATAC-seq peaks/enriched regions obtained from D0, H6, D3, D4 and D5 are shown in Supplementary Data Tables 11-15.

Identification of RBC trait-associated SNPs and related analyses:
SNPs associated with RBC traits were compiled from the following GWAS studies7-13,33-38. We selected SNPs filtering for MCV, HGB, RBC#, MCH, HTC, MCHC, and RDW as phenotypes. In total, 1,325 lead SNPs associated with any of the above RBC parameters were obtained. Using the lead GWAS SNP for each region, in order to increase the likelihood of including the functional SNPs from a reported hit, we also included highly associated SNPs with the lead SNP (with linkage disequilibrium LD R2 0.6). Only SNPs with "rs" identifier numbers were considered. SNPs can have multiple allele pairs that show differential association with traits. To account for this possibility, we broke out each allele pair for each SNP; only allele pairs that had two non-NA alleles were retained. Accordingly, 29,069 lead and LD SNPs with at least two usable alleles, across 924 loci associated with the seven RBC traits, were used to initiate the study. Unless otherwise reported, numbers of SNPs reported refer to the positions of SNPs, i.e. two allele pairs of the same SNP are reported once. We used the same approach and criteria for selecting the platelet trait-associated GWAS SNPs from Astle et al., 2016 to use them as negative controls. RBCs and platelets share origin from megakaryocyte and erythroblast progenitor cells, suggesting platelet trait SNPs as the ideal negative control for our study. We used total 786 risk loci regions associated with 575 lead and 22,158 (lead+LD) platelet trait SNPs (LD R2 0.6) with at least two usable alleles. The lists of all the SNPs that fall within overall enhancers and within TSCs are mentioned in Supplementary Table 5.
Positions of these SNPs relative to the hg19 revision of the human reference genome were taken from the UCSC genome browser track containing dbSNP version 142. SNP-enhancer or SNP-TSC overlap was determined using bedtools intersect. SNP-motif hit overlap was determined using bedtools intersect. Risk and reference allele sequences for each SNP passing the above filters were used to create 41nt-long DNA fragments that contain hg19 reference genome sequence upstream and downstream of the SNP position. Each 41nt sequence was scanned for presence of predicted transcription factor-binding sequences using FIMO 4.11.4 74 with a reference motif library that included HOCOMOCOv10_HUMAN 75, JASPAR_CORE_2016_76 vertebrates and those from 77. Motif hits that overlapped the SNP position in the 41nt sequence were retained and used for comparison between risk and reference alleles, i.e. the SNP was required to overlap the motif hit. Thus, we also required that, for a SNP to be associated with a motif hit, the motif hit directly overlap the center of the region, i.e. the SNP's position. The construction of 41bp sequences centered on the SNP itself, allowed for the SNP to appear at the extreme ends of longer motifs, such as motifs from heterodimeric TF binding. Unique SNP IDs were the unit used for counting.
To test whether our H3K27ac ChIP-seq/ATAC-seq based approach enriches for "functional" SNPs, we use RegulomeDB39. A RegulomeDB score 4 was used to predict SNPs with the minimal functional evidences. This resulted in 5,695 RBC SNPs out of total 29,069 SNPs with two usable alleles.

Motif occurrence identification:

Positions of motif occurrences were determined across the hg19 revision of the human reference genome using FIMO 74 with default parameters and a position weight matrix reference library comprised of HOCOMOCOv10, jolma2013, and JASPAR_CORE_2016_vertebrates. Motifs hits were subsetted by the type of transcription factor predicted to recognize them. Motif PWMs used for downstream analyses are within Supplementary Data Table 6. The numbers of base pairs contained within each category of motif occurrence were calculated after collapsing all occurrences of either STFs motifs or MTF motifs using bedtools merge 73.

Analysis of single nucleotide polymorphisms (SNPs) using protein binding microarray (PBM) data:
Universal protein binding microarray (PBM) 8-mer enrichment (E) score datasets were downloaded from the UniPROBE 44 and CIS-BP 45 databases. Please see Supplementary Data Table 7for the list of PBM datasets analysed in this manuscript 45,46,78-80. Of the 3318 RBC trait SNPs mapped within the non-exonic enhancer regions in this manuscript (defined, as above in Identifying Enhancers and Transcriptional Signaling Centers (TSCs), using the H3K27ac ChIP-seq and ATAC-seq peak information), 3,263 SNPs involving single nucleotide substitutions were considered in the analytical workflow. For each SNP, a 15-bp window, with the SNP at the centre, was obtained, using the GRCh38 version of the human reference sequence. For each of the eight 8-mers spanning the 15-bp window, contiguous ungapped PBM 8-mer E-scores for a transcription factor of interest were obtained for both the reference allele and the SNP-containing allele. Wilcoxon signed-rank tests were performed for the eight reference 8-mers versus their corresponding eight SNP-containing 8-mers to evaluate the statistical significance of any change in E-scores per 15-bp window associated with a SNP. For heightened stringency, the RBC trait SNP examples presented in this manuscript contain at least two consecutive 8-mers within the 15-bp window 81,82 in which the reference allele 8-mers have E-scores of >0.35 and the SNP-containing allele 8-mers have E-scores <0.3, or vice-versa.

Analysis of perturbed transcription factor binding events associated with the set of single nucleotide substitution RBC trait SNPs:
For this analysis, individual PBM datasets (Supplementary Data Table 7) were considered, with the exception of GATA – the average E-score for each contiguous ungapped 8-mer from GATA3, GATA4, GATA5 and GATA6 PBM datasets was used; results were similar to this averaged GATA binding profile when individual GATA factor PBM datasets were analysed. The GATA zinc fingers in these mouse GATA3, GATA4, GATA5 and GATA6 TFs show between 80.00% to 91.43% amino acid identity when compared to the corresponding DNA binding domains in human GATA1, and 82.86% to 97.14% amino acid identity to that in human GATA2. A threshold of ~70% amino acid identity in the DNA binding domain has previously been proposed for TFs to share similar sequence specificity 45. We analyzed a mouse SMAD3 PBM dataset 46; mouse SMAD3 shows 69.61% identity in the amino acid sequence of the MH1 DNA binding domain when compared to the human SMAD1 MH1 DNA binding domain (please see below, in Sequence alignment of transcription factor DNA binding domains, for the methodologies used to calculate percent amino acid identity of DNA binding domains of TFs considered from the same family).
To consider whether the set of 3,263 single nucleotide substitution RBC trait SNPs mapped within enhancers were enriched for perturbation of binding by GATA factors versus putative signal transcription factors or by GATA factors, this set of SNPs was compared against a background set of SNPs, comprising all common SNPs from dbSNP (Build 151, GRCh38p7) that had an allele frequency >10%. For each PBM dataset of interest, the E-scores for reference allele 8-mers versus SNP-containing allele 8-mers were obtained according to the method described in Analysis of single nucleotide polymorphisms (SNPs) using protein binding microarray (PBM) data. For each pair of reference allele 8-mer and corresponding SNP-containing 8-mer, if one allele had an E-score >0.35, while the other allele had an E-score < 0.3, binding by the corresponding transcription factor was considered to be perturbed by the SNP. This procedure considered both SNPs that resulted in a gain of binding by the transcription factor of interest, and SNPs that abrogated or diminished transcription factor binding. This computation was performed for all PBM datasets of interests, to compare all 3,263 foreground SNPs against the background of ~5.4 million SNPs. Bootstrapping of the background SNPs was performed to obtain an empirical background distribution: 100,000 iterations of the background were obtained by sampling, with replacement, 3263 SNPs from the background SNPs. Each of these 100,000 iterations resulted in a distribution of values corresponding to the number of perturbed transcription factor binding events per 3263 SNPs * eight 8-mers per SNP = 26,104 8-mers; the mean value of these 100,000 iterations was taken as the expected number of perturbed binding events per transcription factor of interest. The empirical p-value for each transcription factor of interest was computed by ranking the number of perturbed transcription factor binding events for the foreground set of 3263 SNPs * eight 8-mers per SNP against the 100,000 values from the empirical background distribution. The Benjamini-Hochberg procedure was applied (Benjamini and Hochberg, 1995), using the p.adjust function in R, to correct for multiple hypothesis testing.

Sequence alignment of transcription factor DNA binding domains:
DNA binding domains in transcription factors were identified by using hmmscan on the HMMER web server 83, scanning against the Pfam profile hidden Markov model database84 and using the default Pfam gathering threshold parameters. Pairwise global alignment of the protein sequences of these DNA binding domains was performed using EMBOSS Needle85, with the default parameters, to allow for computation of amino acid identity between two sequences.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data Availability Statement:
The massively parallel sequencing data associated with this manuscript have been uploaded to GEO under the accession numbers GSE74483 and GSE104574 and are currently open to public. The web links for the publicly available databases used in this study are: UniPROBE: http://thebrain.bwh.harvard.edu/uniprobe/, CIS-BP: http://cisbp.ccbr.utoronto.ca/, FHS: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v30.p11, RegulomeDB: https://regulomedb.org/regulome-search/, HEMMER: http://hmmer.org/, EMBOSS Needle: https://www.ebi.ac.uk/Tools/psa/emboss_needle/. dbSNP: https://www.ncbi.nlm.nih.gov/snp/?cmd=search. Links to all the PBM datasets used are available in Supplementary Data Table 7.

Code Availability:
Custom codes used in this study are available at https://bitbucket.org/abrahamb/workspace/projects/TSC. The code and data files for the PBM analyses are available at https://github.com/BulykLab/RBCSNPs_2020.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size power calculations were used. Since independent human donor derived CD34 cells or clones were used, three independent replicates were used for most of the experiments.
Replication	Data were not excluded from the analysis, rather Values from each replicate were considered to test if the final results were statistically significant or not
Randomization	Samples were randomized for calculation of frequency of SNPs altering binding of specific TFs using protein binding microarray (PBM) and to calculate frequency of SNPs in TSCs and in STF motif hits. Detailed methodologies can be found in the online and supplementary methods section of the manuscript.No randomizations were performed for other studies.
Blinding	Blinding was not used as most of the resalts relied on electronically-derived values and not on visualization through human eyes.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
X	☐ Palaeontology
X	☐ Animals and other organisms
X	☐ Human research participants
X	☐ Clinical data

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☐	☒ Flow cytometry
X	☐ MRI-based neuroimaging

Antibodies

Antibodies used	The antibodies, along with the supplier name, catalog number, clone/lot number, are mentioned below: SMAD1: Santacruz, sc7965X, A4, E1314; GATA1: Santacruz, sc265X, N6, J0511; GATA2: Santacruz, sc9008X, H-116, F0315; H3K27ac: Abcam, Ab4729, GR3231988-1; PU.1: Santacruz, sc352X, D1316, T-21; KLF1: Abcam, Ab2483, GR37687-20; TCF7L2: Cell Signaling, 2569S, C48H11, 07/2017. All these antibodies were used for ChIP-seq and 10 ug antibody was used for each ChIP experiment.
Validation	Details of the antibodies used for the FACS are as follows and dilutions used are also metioned: 1:60 APC-conjugated CD235a (eBioscience, clone HIR2, 17-9987-42), 1:60 FITC-conjugated CD71 (eBioscience, OKT9, 11-0719-42), 1:60 PE-conjugated CD41a (eBioscience, HIP8, 12-0419-42) and 1:60 PE-conjugated CD11b (eBioscience, ICRF44, 12-0118-42) All antibodies in this study were used according to manufacturer's instructions and dilutions. For ChIP-seq, only ChIP-seq grade antibodies were used either by searching each manufacturer's website, or by searching previously published successful ChIP-seq results. Reliability of ChIP-seq peaks obtained was validated by known nearby erythroid genes. Wherever possible, ChIP-seq peaks were further validated by performing ChIP-seq in a relevant knockout cell lines. Furthermore, The overall quality control values, percentage occupancy of each factors at different genomic regions are mentioned in Supplementary Data Table 9.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	All the cell lines used in this study have long been used in this lab and are validated from multiple publications. Fresh batch of primary CD34 cells are bought from Fred Hutch Center in Seattle, and they always provide the authentication of each vial. K562 cell lines were bought from ATCC. G1ER cells were obtained from Alan Cantor's lab, PU.1 modified K562 cells were obtained from Sinichiro Takahashi's lab. They are the co-authors of the manuscript.
Authentication	Identity of cell lines were validated by STR analysis. For mutated parental cell lines, mutations were verified by PCR with primers that are all mentioned in Supplementary Data Table 16 and procedure is described in methods section.
Mycoplasma contamination	Mycoplasma tests were performed on a frequent basis and confirmed to be negative. Tests were done using the luminiscence based MycoAlert Mycoplasma Detection Kit, Lonza LT07-318. Media collected from all the growing cells were tested for a luminiscence ratio of less than 0.9 to be confirmed as mycoplasma negative as per manufacturer's protocol. Ratio of greater than 1.2 is considered as mycoplasma positive. If ratio falls between 0.9 and 1.2, samples were tested further by PCR-based LookOut Mycoplasma PCR Detection Kit, Sigma MP0035. All the cell lines used in this study were associated with a ratio of well below 0.9 and hence confirmed to be mycoplasma negative.

Commonly misidentified lines (See ICLAC register)	None of the cell lines used are listed in the ICLAC database.

Palaeontology

Specimen provenance	Not Applicable
Specimen deposition	Not Applicable
Dating methods	Not Applicable

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Not Applicable
Wild animals	Not Applicable
Field-collected samples	Not Applicable
Ethics oversight	Not Applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Not Applicable
Recruitment	Not Applicable
Ethics oversight	Not Applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	Not Applicable
Study protocol	Not Applicable
Data collection	Not Applicable
Outcomes	Not Applicable

Methodology

Replicates

Apart from ChIPseq, we conducted multiple extensive genome-wide assays from the time-point-matched CD34 cells for each time-point of differentiation. The genome-wide analyses of all these assays showed high concordance among each other establishing each stage of differentiation. We reasoned that each genome-wide assay served as perfect circumstantial evidence for replicating the results concluded from a different assay, and hence, we didn't use replicates for the same assay. For example, the comparison of genome-wide H3K27ac ChIP-seq and ATAC-seq signal intensities and RNA-seq RPKM values between different differentiation stages formed a progenitor-specific and an erythroid-specific cluster before and after day 3 (D3) (Extended Data Fig. b, c, d, e and f). These results independently reveal D3 as the erythroid commitment time-point in our cell culture system during the gradual transition of cells from progenitor to erythroid stage. This observation also correlated with genome-wide ChIP-seq analysis of differentiation stage-specific GATA2 and GATA1 that showed a gradual GATA2 to GATA1 genomic switch around D3 marking the erythroid fate change (Fig. 1c, d). SMAD1 ChIP-seq peaks corroborated the gradual shift of stages around D3 (Fig. 1c). The Spearman correlation co-efficient values comparing the RPM-normalized (reads per million) read densities for SMAD1 ChIP-seq data sets, at each time-point, suggested that ChIP-seq peaks before D3 (i.e. D0 and H6) and after D3 (i.e. D4 and D5) are more similar to each other than pairs across D3 (This result is not included in the current figure for space restriction, but can be provided if requested). The individual gene-tracks of known hematopoietic genes (An et al., 2014) supported the conclusions from genome-wide analyses; progenitor-specific genes FLI1 and FLT3 show gradual decrease of H3K27ac- and ATAC-seq peaks along with RNA-seq RPKM values, whereas erythroid gene ALAS2 showed a steady increase of open/active chromatin marks and RPKM values during the course of differentiation (Fig. 2b, Extended Data Fig. 1d, e). (RPKM values of FLI1: D0 = 20.86, H6 = 25.99, D2 = 14.33, D3 = 7.38, D4 = 5.47, D5 = 4.99; RPKM values for FLT3: H6 = 25.99, D1 = 26.52, D3 = 10.72, D4 = 4.37, D5 = 2.47; RPKM values for ALAS2: D0 = 4.86, H6 = 7.84, D2 = 7.0, D3 = 23.11, D4 = 54.37, D5 = 174.69). Given the strong correlation between samples, we believe that our data-sets for each assay are of high quality, accurately reflecting the identity of differentiation stages.

To validate this further, we have correlated the datasets used in this manuscript with selected time-point-matched replicates. We have compared the ChIP-seq data for GATA2 at D0, GATA1 at D5 and SMAD1 at D0 with previously generated ChIP-seq peaks in our lab as representative key samples. Replicate peak sets show appreciable similarity (around 70% of peaks in common, on average), with gene tracks revealing highly similar signal distributions (This result is not included in the current figure for space restriction, but can be provided if requested). Additionally, we have now replicated RNA-seq datasets at D0, H6, D1, D2, D3 and D4 using CD34 cells from a different donor. Clustering datasets from each donor reveals a similar kinetic profile, as both donors' samples clearly separate into erythroid cluster after D3 (This result is not included in the current figure for space restriction, but can be provided if requested). This observation is underscored looking at a subset of genes that show 2-fold or 4-fold increases in expression from D3 to D8 in the previous RNA-seq data-set. These genes replicated similar expression dynamics, showing a comparable increase in expression starting around D3 of differentiation (This result is not included in the current figure for space restriction, but can be provided if requested). Given these observations are reproducible between distinct biological replicates, we are highly confident about the validity of our results and the conclusions that we have drawn.

Sequencing depth

Average sequencing depth was 20-25 million reads per sample

Antibodies

For ChIP-seq experiments the following antibodies were used: Smad1 (Santa Cruz sc7965X), Gata1 (Santa Cruz sc265X), Gata2 (Santa Cruz sc9008X), H3K27ac (Abcam ab4729), PU1 (Santa Cruz sc352X) and KLF1 (Abcam ab2483).

Peak calling parameters

Alignment and Visualization:
ChIP-Seq reads were aligned to the human reference genome (hg19) using bowtie 66 with parameters -k 2 -m 2 -S. WIG files for display were created using MACS 67 with parameters -w -S --space=50 --nomodel --shiftsize=200, were normalized to the millions of mapped reads, and were displayed in IGV 68,69. The overall quality control values, percentage occupancy of each factors at different genomic regions are mentioned in Supplementary Data Table 9. The ChIP-seq peaks/enriched regions obtained from D0, H6, D3, D4 and D5 are shown in Supplementary Data Tables 11-15.
Peak and Bound Gene Identification:
High-confidence peaks of ChIP-Seq signal were identified using MACS with parameters --keep-dup=auto -p 1e-9 and corresponding input control. Bound genes are RefSeq genes that contact a MACS-defined peak between -10000bp from the TSS and +5000bp from the TES. The bound genes associated with GATA2/1 and SMAD1 at each stage are shown in Supplementary Data Table 2.

Data quality

For individual assays performed at each time-point, Supp Data Table 9 enlists: (a) number of peaks/enriched regions, the percentage of peaks that are within promoter and non-promoter regions (along with the definitions of promoter and non-promoter regions); (b) Supp Data Tables 11-15 shows genomic localization of peaks for all the genome-wide assays performed at day 0 (D0), hour 6 (H6), day 3 (D3), day 4 (D4) and day 5 (D5); (c) Supp Data Table S9 also depicts "Fraction Non-Redundant Reads in Peaks/Regions" (FRiP) that we have now used to detect the overall quality of our data. ENCODE use FRiP parameter as part of their QC definition. ENCODE data sets have a FRiP enrichment of 1% or more when peaks are called using MACS with default parameters. The ENCODE Consortium scrutinizes experiments in which the FRiP falls below 1% (Landt et al., 2012). All of our samples meet this 1% cutoff, except for the ChIP-seq for GATA1at D0 (FRiP = 0.9%), GATA1 at H6 (FRiP = 0.7%) and GATA2 at D5 (FRiP = 0.4%). This is likely due to the low genomic occupancy of GATA2 and GATA1 at respective stages, that can be explained by GATA-switch, a well-known phenomenon during erythropoiesis.

Software

All the softwares used for this study are listed in the methods section and also in the earlier "software and code" section of this reporting summary.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

To access GSE74483, please Go to https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74483
To access GSE104574, please Go to https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE104574

Files in database submission

Raw FASTA files, BED files and wig or bigwig files

Genome browser session
(e.g. UCSC)

IGV

ChIP-seq peak alignment and Visualization:
ChIP-Seq reads were aligned to the human reference genome (hg19) using bowtie 66 with parameters -k 2 -m 2 -S. WIG files for display were created using MACS 67 with parameters -w -S --space=50 --nomodel --shiftsize=200, were normalized to the millions of mapped reads, and were displayed in IGV 68,69. The overall quality control values, percentage occupancy of each factors at different genomic regions are mentioned in Supplementary Data Table 9. The ChIP-seq peaks/enriched regions obtained from D0, H6, D3, D4 and D5 are shown in Supplementary Data Tables 11-15.
ChIP-seq peak and Bound Gene Identification:
High-confidence peaks of ChIP-Seq signal were identified using MACS with parameters --keep-dup=auto -p 1e-9 and corresponding input control. Bound genes are RefSeq genes that contact a MACS-defined peak between -10000bp from the TSS and +5000bp from the TES. The bound genes associated with GATA2/1 and SMAD1 at each stage are shown in Supplementary Data Table 2.
Identifying Enhancers and Transcriptional Signaling Centers (TSCs):
Enhancers were identified using H3K27ac ChIP-Seq and ATAC-Seq peak information. Peaks were identified as described above using MACS. Coding regions were removed from H3K27ac and ATAC-Seq peaks using bedtools subtract; coding regions were exons from all RefSeq transcripts. Non-exonic portions of ATAC-Seq-enriched regions that overlapped H3K27ac-enriched regions by at least 1 bp were retained. H3K27ac- or ATAC-Seq/H3K27ac-enriched regions outside exons were collapsed using bedtools merge. These steps were performed for each timepoint's ATAC-Seq/H3K27ac ChIP-Seq pair. D0, H6, D4, and D5 regions were collapsed and used for "enhancers" across the time-course.
Transcription signaling centers (TSCs) were defined as those that was co-bound by SMAD1 and the corresponding GATA factor. For each time-point, regions enriched in both SMAD1 and the corresponding GATA were identified using bedtools intersect on the peaks. Enhancers, as defined above, are considered TSCs if they overlap a SMAD1/GATA-bound region by at least 1 bp. TSCs identified at each of D0, H6, D4, and D5 were collapsed and used as a canonical list of TSCs across the time-course. Progenitor signaling centers represent the union of D0 and H6; erythroid signaling centers represent the union of D4 and D5 time-points.

Lists of all the enhancers and TSCs identified using above methods are listed in Supplementary Data Table 4.
Peak Similarity Heatmaps:
The called H3K27ac ChIP-seq and ATAC-seq peaks of all samples were combined to generate peak files for heatmap analysis. Depending on the overlap of the union of peaks and peak from individual sample, a binary matrix of 0 and 1 were assigned to each peak of each sample. The similarity score was derived and correlation matrix was calculated by the cor method, and the heatmap drawn by corrplot package in R.
ChIP-Seq Read Density Heatmaps/Scatterplots:
ChIP-Seq read density heatmaps were constructed using bamToGFF (https://github.com/BradnerLab/pipeline) on 4kb regions centered on the peak center with parameters -m 200 -r -d and filtered bam files with at most one read per position. Binary peak/not-peak "heatmaps" were determined by first taking the collapsed union of peaks defined at all five timepoints and determining whether each of these collapsed regions contacted a peak in any of the timepoints.
Co-occupancy of multiple STFs upon stimulation with respective signaling pathways:
ChIP-Seq read density heatmaps were constructed using bamToGFF (https://github.com/BradnerLab/pipeline) on 4kb regions centered on the peak center. Single-TF heatmaps were built with parameters -m 200 -r -d and filtered bam files with at most one read per position; rows were ordered by the row sums of the indicated factor. Multiple-TF heatmaps were built with parameters -m 100 -r -d, and rows were ordered by the row sum in SMAD1 signal. Binary peak/not-peak "heatmaps" were determined by asking if the original peak overlapped a SMAD1-enriched region using bedtools intersect.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Control and treated stage-matched CD34 cells, or CD34 cells at different stages of differentiation were washed in PBS and stained with propidium iodide (PI), 1:60 APC-conjugated CD235a (eBioscience, clone HIR2, 17-9987-42), 1:60 FITC-conjugated

	CD71 (eBioscience, OKT9, 11-0719-42), 1:60 PE-conjugated CD41a (eBioscience , HIP8, 12-0419-42) and 1:60 PE-conjugated CD11b (eBioscience, ICRF44, 12-0118-42).
Instrument	BD Bioscience LSR II flow cytometer was used to record raw FACS data
Software	FlowJo 8.6.9 10.0.7 (TreeStar).
Cell population abundance	Cells were not sorted using flow cytometer
Gating strategy	Cells of interest were separated from dead cell debris using forward scatter versus side scatter. Single cells were separated from doublets through forward scatter height (FSC-H) versus forward scatter area (FSC-A). Subsequent live-dead differentiation was done using Propidium Iodide (PI) stain. The live cells were then stained for the differentiation markers, such as CD235a-APC and CD71-FITC for the CD34+ and the HUDEP2 cells.

[X] Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI [ ] Used [ ] Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
Normalization template	*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
Noise and artifact removal	*Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*
Volume censoring	*Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings	*Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*
Effect(s) tested	*Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: [ ] Whole brain [ ] ROI-based [ ] Both

Statistic type for inference (See Eklund et al. 2016)	*Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.