Supplementary Information for: Common variants in signaling transcription factor binding sites drive phenotypic variability in red blood cell traits

Contents

Supplementary Notes

Note 1

We calculated the number of occurrences of enhancer-associated SNPs overlapping predicted TF motif hits in our list, and as a control, we randomly shuffled SNPs within enhancer regions. We observed that no random permutation of SNPs overlapped as many TF motif hits as the actual RBC trait-associated SNPs ($p<0.0001$), suggesting that non-coding SNPs, in general, tend to overlap TF motif hits.

Note 2

For each single nucleotide substitution, a 15-bp sequence with the SNP at the center 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 and alternate alleles. The E-scores obtained for each base position were subsequently compared to determine whether the alleles of the SNP may lead to altered binding efficiency of the respective transcription factor. 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. We applied this approach to available PBM datasets for all STFs across all 3263 SNPs. We then calculated the ratio of observed/expected perturbation events by comparing the events across the set of trait-associated SNPs to the events seen across a background of all common (>10% allele frequency) SNPs in dbSNP (Build 151, GRCh38p7) (presented in details in Methods).

Note 3

To examine perturbed TF binding in the context of fine-mapped SNPs, we identified the LD-based SNPs from our list that were also present in the list of fine-mapped SNPs in Ulirsch et al[1]. We found 1064 SNPs at posterior probability (PP) > 0.01, and 1911 SNPs at PP > 0.001 (Supplementary Table 8). We applied our PBM dataset-based approach in analyzing perturbed TF binding in this subset of SNPs. We found that SNPs resulting in perturbed STF binding were significantly enriched in the intersection of our LD-based and fine-mapped lists of SNPs with PP > 0.01, versus SNPs that were present only in our LD-based list (Fisher's exact test, odds ratio 1.211, p-value = 0.011, 95% CI: 1.043-1.406). This enrichment was not seen for the intersection with fine-mapped SNPs of PP > 0.001 (Fisher's exact test, odds ratio

0.996, p-value = 0.972, 95% CI: 0.864-1.147), while the enrichment for perturbed GATA binding was not significant at PP > 0.01 (Fisher's exact test, odds ratio 1.148, p-value = 0.212, 95% CI: 0.925, 1.421) nor PP > 0.001 (Fisher's exact test, odds ratio 1.158, p-value = 0.164, 95% CI: 0.940-1.430). The PBM-driven analysis coupled with the PWM (position weight matrix)-based motif analysis strongly suggest that functional alterations caused by SNPs are mediated by impaired binding of STFs to DNA.

Note 4

It is interesting to note that BMP-SMAD1 and WNT-TCF7L2, two STFs used for functional assays in this study, have been shown to be GWAS hits for RBC traits, RBC distribution width (RDW) and erythrocyte count (RBC), respectively (https://www.ebi.ac.uk/gwas/genes/SMAD1;https://www.ebi.ac.uk/gwas/genes/TCF7L2). This suggests that altered responsiveness to BMP and WNT signaling could affect RBC traits, further supporting our findings. Single examples of SNPs or larger enhancer deletions that could alter individual signaling events, mainly related to immune response or stressed-hematopoiesis, have been shown[2-4]. In a recent study[2], using tiled CRISPR activation, the authors identified one enhancer associated with the *IL2RA* gene that harbors the autoimmunity risk variant rs61839660. Using genetically engineered mouse models of this variant, the authors were able to show delayed gene activation in response to specific extracellular signals; however, further studies to identify the TF motif that is potentially altered and elucidation of the effects of the SNP on transcription factor binding remain outstanding. In another study, the authors looked for genome-wide “amalgamated E-box and GATA motifs”[3]. They identified an anemia-induced stress-regulated enhancer that appears to control the expression of *Samd14* implicated in Stem Cell Factor (SCF)/c-Kit signaling. By exploring our dataset, we observed that, in human CD34+ progenitors, the *Samd14* enhancer fits our definition of a SMAD1 TSC, and it is co-occupied by multiple STFs under stimulation, along with the MTFs GATA2 and PU.1. This supports our model that TSCs are critical nodes driving stress-responsive gene expression programs. In our current study, we show that, genome-wide, it is primarily the binding of developmental signaling-induced TFs within TSCs that is altered by enhancer variants, suggesting that many RBC trait phenotypes could result from altered stimulus response.

Note 5: Supplementary Methods

Chromatin Immunoprecipitation (ChIP)

A summary of the bound genes determined for all ChIP-seq data is contained within the Supplementary Table 2. 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). ChIP experiments were performed as previously described [5,6]. Briefly, 20-30 million cells for each ChIP were used.Fifty microliters of cell lysates prior to addition to the beads was kept as input.

ChIP-PCR

A total of 10-20x10^6 cells were harvested and fixed with 1% formaldehyde in preparation for ChIP with 10 µg of each antibody (Smad1, Santa Cruz sc7965X; Gata1, Santa Cruz sc265X; PU.1, Santa Cruz sc352X and TCF7L2, TCF4 Santa Cruz sc8631 and the corresponding normal IgG). ChIP-DNA was then quantified against the respective loci via Light Cycler 480 II SYBR green master mix (Applied Biosystems) and the QuantStudio 12K Flex Real-Time PCR System (Applied Biosystems). The primers used for individual loci can be found in Supplementary Table 16.

RNA sequencing (RNA-seq)

RNA-seq was performed on CD34+ cells for the following time points post-hrBMP4 stimulation: D0, H2, H6 and D1-8. The cells were kept in media described above and treated with rhBMP4 for 2hrs before collection. RNA from one million cells was isolated using Trizol according to the manufacturer's instructions. The RNA was DNAse-treated using the RNase-free DNase set from Qiagen (79254) according to the instructions. The whole amount of RNA was treated with the Ribo-Zero Gold kit (Human/Mouse/Rat, Epicentre) according to the manufacturer's instructions. The cDNA was cleaned with Agencourt AMPure purification and this was used as a template to produce multiplexed libraries (see library preparation).

ChIP-Seq and RNA-seq library Preparation

Briefly, ChIP-seq libraries were prepared using the following protocol. End repair of immunoprecipitated DNA was performed using the End-It End-Repair kit (Epicentre, ER81050) and incubating the samples at 25^{0}C for 45 min. End-repaired DNA was purified using AMPure XP Beads (1.8X of the reaction volume) (Agencourt AMPure

XP – PCR purification Beads, BeckmanCoulter, A63881) and separating beads using DynaMag-96 Side Skirted Magnet (Life Technologies, 12027). A-tail was added to the end-repaired DNA using NEB Klenow Fragment Enzyme (3'-5' exo, M0212L), 1X NEB buffer 2 and 0.2 mM dATP (Invitrogen, 18252-015) and incubating the reaction mix at 37^0C for 30 min. A-tailed DNA was cleaned up using AMPure beads (1.8X of reaction volume). Subsequently, cleaned up A-tailed DNA went through Adaptor ligation reaction using Quick Ligation Kit (NEB, M2200L) following manufacturer's protocol. Adaptor-ligated DNA was first cleaned up using AMPure beads (1.8X of reaction volume), eluted in 100µl and then size-selected using AMPure beads (0.9X of the final supernatant volume, 90 µl). Adaptor-ligated DNA fragments of proper size were enriched with PCR reaction using Fusion High-Fidelity PCR Master Mix kit (NEB, M0531S) and specific index primers supplied in NEBNext Multiplex Oligo Kit for Illumina (Index Primer Set 1, NEB, E7335L). Conditions for PCR used are as follows: 98 °C, 30 sec; [98°C, 10 sec; 65 °C, 30 sec; 72 °C, 30 sec] X 15 to 18 cycles; 72°C, 5 min; hold at 4 °C. PCR enriched fragments were further size-selected by running the PCR reaction mix in 2% low-molecular weight agarose gel (Bio-Rad, 161-3107) and subsequently purifying them using QIAquick Gel Extraction Kit (28704). Libraries were eluted in 25µl elution buffer.

For the RNA-seq libraries, purified double-stranded cDNA underwent end-repair and dA-tailing reactions following manufacturer's reagents and reaction conditions. The obtained DNAs were used for Adaptor Ligation using adaptors and enzymes provided in NEBNext Multiplex Oligos for Illumina (NEB#E7335) and following kit's reaction conditions. Size selection was performed using AMPure XP Beads (starting with 0.6X of the reaction volume). DNA was eluted in 23 µl of nuclease free water. Eluted DNA was enriched with PCR reaction using Fusion High-Fidelity PCR Master Mix kit (NEB, M0531S) and specific index primers supplied in NEBNext Multiplex Oligo Kit for Illumina (Index Primer Set 1, NEB, E7335L). Conditions for PCR used are as follows: 98 °C, 30 sec; [98°C, 10 sec; 65 °C, 30 sec; 72 °C, 30 sec] X 15 cycles; 72°C, 5 min; hold at 4 °C. PCR reaction mix was purified using Agencourt AMPure XP Beads and eluted in a final volume of 20 µl. The libraries were sequenced in Illumina Hiseq 2500 platform.

ChIP-Seq data analysis

ChIP-Seq reads were aligned to the human reference genome (hg19) using bowtie[7] with parameters -k 2 -m 2 -S. WIG files for display were created using MACS[8] with parameters -w -S --space=50 --nomodel --shiftsize=200, were normalized to the millions of mapped reads, and were displayed in IGV[9,10]. High-confidence peaks of

ChIP-Seq signal were identified using MACS with parameters --keep-dup=auto -p 1e-9 and corresponding input control. The coupling of an unusually stringent p value cutoff and input control makes false positive peak identification highly unlikely, although formally possible. The bound genes that are studied in Fig. 1e and Extended Data Fig. 1h, associated with GATA2/1 and SMAD1 at each stage, are shown in Supplementary Table 2. Bound genes are defined as RefSeq genes meeting at least one of two criteria: (1) their transcription start site is most proximal to the center of the peak determined by bedtools closest, and/or (2) their promoters (transcription start site +/- 500bp) overlap a peak determined by bedtools intersect. SMAD1/GATA co-bound genes are genes that are predicted targets of a co-bound peak, defined using bedtools intersect. The analysis in Fig. 1e integrated the RNA-seq and ChIP-seq stage-matched data set from one complete round of differentiation, including ATAC-seq and H3K27ac ChIP-seq from time-point-matched CD34 cells. Each dataset from each time-point was assessed prior to analysis for its concordance with known characteristics of progenitor/erythroid cells, including signal of/near known marker genes of hematopoietic progenitors and erythroid cells. The overall quality control values, percentage occupancy of each factors at different genomic regions are mentioned in Supplementary Table 9. Region counts, collective region sizes in base pair, and gene counts associated with different genomic regions studied at respective differentiation stages are mentioned in Supplementary Table 10. The ChIP-seq peaks/enriched regions obtained from D0, H6, D3, D4 and D5 are shown in Supplementary Data Tables 11-15.

Identifying Enhancers and Transcriptional Signaling Centers (TSCs)

Enhancers were identified using H3K27ac ChIP-Seq and ATAC-Seq peak information. ATAC-seq peaks were identified using MACS 1.4 with –keep-dup=auto. Peaks were identified as described above using MACS. Coding regions were removed from H3K27ac and ATAC-Seq peaks using bedtools subtract; coding regions were defined as exons from all RefSeq transcripts. Non-exonic portions of ATAC-Seq peaks that overlapped H3K27ac peaks 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 were 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. The fraction of enhancers variously defined in Extended Data Fig. 2 was performed using H3K27ac peaks, stitched enhancers defined using H3K27ac peaks, and ATAC-seq peaks. H3K27ac peaks were defined using MACS 1.4 with -p 1e-9 and input control from the corresponding time point. ATAC-seq peaks were identified by MACS2 version 2.1.0[8] peak finding algorithm with the following parameter --nomodel --shift -100 --extsize 200. A q-value threshold of enrichment of 0.05 was used for all datasets. Promoters are defined as 1kb regions centered on RefSeq transcription start sites. Stitched enhancers were defined using ROSE with stitching distance 12500 and -t 2000. Intersected regions were determined using bedtools intersect with a single bp overlap required. Union regions were created using bedtools merge.

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 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. Heatmaps were visualized using the heatmap.2 package.

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.

Genome-wide occupancy comparison of PU.1 and SMAD1 at co-bound regions under PU.1 overexpression

Peaks of PU.1 ChIP-Seq signal were defined as above, and 4kb regions centered on the peak center were created for metagene analysis. Read coverage of each region was quantified using bamToGFF with parameters -r -d -m 200. The mean signal across all regions is plotted as a metagene using matplot.

Genome-wide occupancy comparison of PU.1, SMAD1 and GATA2 at co-bound regions under PU.1 knockdown

The transcription factor occupancy profile plots were generated by deeptools2 suite[11] with the computeMatrix and plotProfile command in the reference-point mode, where the TF binding profiles were plotted over the 2kb upstream and downstream regions of every TF binding sites. The y axis is the coverage score that correlated to the number of reads per bin.

RNA-seq data analysis

RNA-seq reads were mapped to the human reference genome (hg19) using TopHat v2.0.13[12] 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 v2.2.1[13] 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 (reads per kilo base per million) expression level, fold change and statistical significance for each gene. The genome-wide RPKM expression values during differentiation are mentioned in Supplementary Table 1.

Assay for Transposase Accessible Chromatin (ATAC-seq)

CD34+ cells were expanded and differentiated using the protocol mentioned above. Before collection, cells were treated with 25 ng/mL hrBMP4 for 2 hr. $5X10^4$ cells per differentiation stage were harvested by spinning at 500 x g for 5 min, 4° C. Cells

were processed exactly as was described in Buenestaro et al[14]. Using Illumina Nextera kit, 15028252, libraries were constructed according to Illumina protocol using the DNA treated with transposase, NEB PCR master mix, Sybr green, universal and library-specific Nextera index primers. Samples with appropriate nucleosomal laddering profiles were selected for next generation sequencing using Illumina Hiseq 2500 platform.

ATAC-seq data analysis

All human ATAC-Seq datasets were aligned to build version NCBI37/HG19 of the human genome using Bowtie2 (version 2.2.1)[15] 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[8] 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 [15] with default options. BedTools[16] 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.

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[17] and CIS-BP[18] databases. Please see Supplementary Table 7 for the list of PBM datasets analysed in this manuscript [18-22]. 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[23,24] 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 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[18]. We analyzed a mouse SMAD3 PBM dataset [19]; 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%. We chose this set of common dbSNP variants as our background to avoid including anything that would be particularly deleterious (with the assumption that such deleterious SNPs would likely be selected against and hence would likely appear less frequently in the population). To clarify our procedure for generating the background, we excluded indels, and SNPs giving rise to frameshift, missense, nonsense or synonymous mutations in protein-coding regions, in order to allow for comparison of the set of non-exonic single nucleotide substitution RBC trait SNPs against equally-sized samples of non-exonic single nucleotide substitution common SNPs.

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, 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[25], scanning against the Pfam profile hidden Markov model database[26] and using the default Pfam gathering threshold parameters. Pairwise global alignment of the protein sequences of these DNA binding domains was performed using EMBOSS Needle[27], with the default parameters, to allow for computation of amino acid identity between two sequences.

Supplementary Note References

1 Ulirsch, J. C. *et al.* Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat Genet* **51**, 683-693, doi:10.1038/s41588-019-0362-6 (2019).

2 Simeonov, D. R. *et al.* Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature* **549**, 111-115, doi:10.1038/nature23875 (2017).

3 Hewitt, K. J. *et al.* GATA Factor-Regulated Samd14 Enhancer Confers Red Blood Cell Regeneration and Survival in Severe Anemia. *Dev Cell* **42**, 213-225 e214, doi:10.1016/j.devcel.2017.07.009 (2017).

4 Soukup, A. A. *et al.* Single-nucleotide human disease mutation inactivates a blood-regenerative GATA2 enhancer. *J Clin Invest*, doi:10.1172/JCI122694 (2019).

5 Trompouki, E. *et al.* Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. *Cell* **147**, 577-589, doi:10.1016/j.cell.2011.09.044 (2011).

6 Lee, T. I., Johnstone, S. E. & Young, R. A. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nat Protoc* **1**, 729-748, doi:10.1038/nprot.2006.98 (2006).

7 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).

8 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).

9 Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192, doi:10.1093/bib/bbs017 (2013).

10 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).

11 Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165, doi:10.1093/nar/gkw257 (2016).

12 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36, doi:10.1186/gb-2013-14-4-r36 (2013).

13 Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31**, 46-53, doi:10.1038/nbt.2450 (2013).

14 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, doi:10.1038/nmeth.2688 (2013).

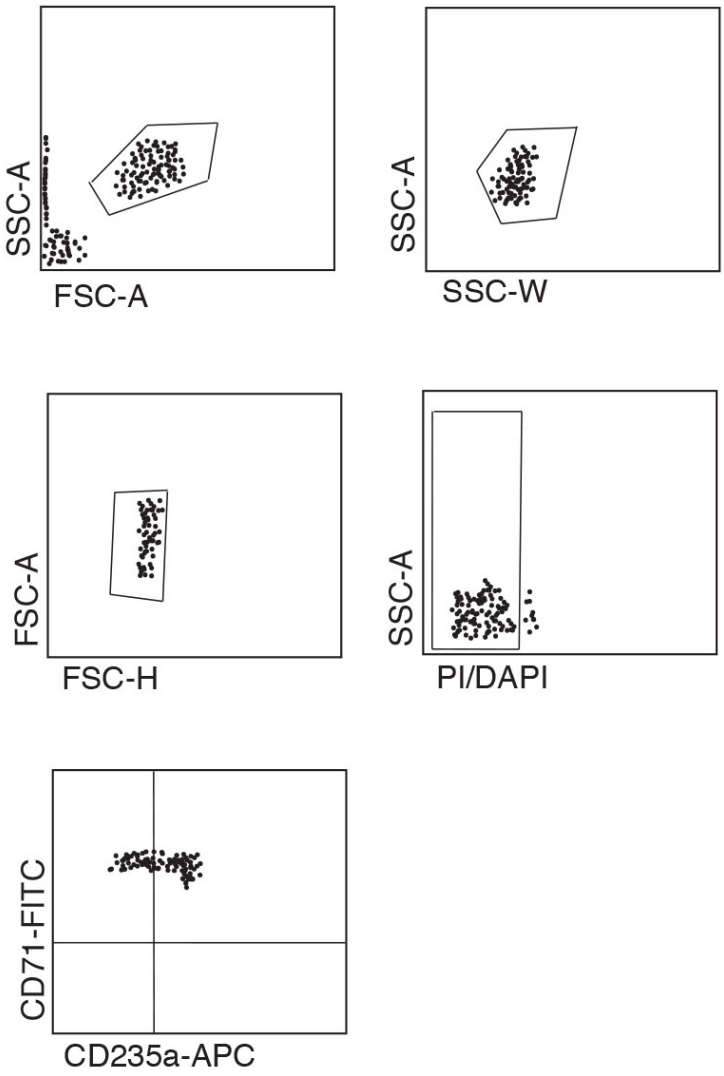
15 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

16 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

17 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**, D117-122, doi:10.1093/nar/gku1045 (2015).

18 Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443, doi:10.1016/j.cell.2014.08.009 (2014).

19 Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720-1723, doi:10.1126/science.1162327 (2009).
20 Barrera, L. A. *et al.* Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science* **351**, 1450-1454, doi:10.1126/science.aad2257 (2016).
21 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**, 654, doi:10.1016/j.cels.2017.12.011 (2017).
22 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, doi:10.1101/gad.207142.112 (2012).
23 Busser, B. W. *et al.* Molecular mechanism underlying the regulatory specificity of a Drosophila homeodomain protein that specifies myoblast identity. *Development* **139**, 1164-1174, doi:10.1242/dev.077362 (2012).
24 Grove, C. A. *et al.* A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. *Cell* **138**, 314-327, doi:10.1016/j.cell.2009.04.058 (2009).
25 Potter, S. C. *et al.* HMMER web server: 2018 update. *Nucleic Acids Res* **46**, W200-W204, doi:10.1093/nar/gky448 (2018).
26 El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res*, doi:10.1093/nar/gky995 (2018).
27 Chojnacki, S., Cowley, A., Lee, J., Foix, A. & Lopez, R. Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. *Nucleic Acids Res* **45**, W550-W553, doi:10.1093/nar/gkx273 (2017).



Supplementary Fig. 1. FACS 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.