# Supplemental Experimental Procedures

## Reproducible research

We have provided the raw data and R scripts necessary to reproduce the primary MPRA analyses (http://www.bloodgenes.org/RBC_MPRA/).

## Identification of relevant tissues and cellular models

We used SNPsea with standard settings to determine tissue- and cell type-specific enrichment for gene expression overlapping with LD blocks containing the 75 GWAS hits (Slowikowski et al., 2014). Hematopoietic expression profiles were downloaded from http://www.broadinstitute.org/dmap/data/ (Novershtern et al., 2011). Narrow DHS peaks for 53 cell types were downloaded from http://www.roadmapepigenomics.org/data/ (Roadmap Epigenomics et al., 2015). Adult erythroblast FAIRE-seq data was obtained the gene expression omnibus GSE36985 and narrow peaks were called using MACS2 (Xu et al., 2012; Zhang et al., 2008). In order to cluster the cell-types by open chromatin, BedTools was used to compute the Jaccard correlation statistic (total shared nucleotides between two peak sets / total unique nucleotides in two sets) on the top 50,000 peaks (Quinlan and Hall, 2010). Clustering was performed with ggdendro (http://cran.r-project.org/web/packages/ggdendro/).

## Microarrays

We infected K562 cells with an HMD-GATA1 lentiviral vector that overexpresses GATA1 as described previously (Ludwig et al., 2014). RNA was isolated 48 hours after infection and cDNA was synthesized as described below. Microarrays (GeneChip Human Gene 2.0 ST Arrays, Affymetrix) were performed on K562 for control (HMD) and GATA1 overexpression (HMD+GATA1). Raw files were processed and normalized using the RMA algorithm from the oligo package in R 3.2 (Carvalho and Irizarry, 2010). Differential expression analyses were conducted using limma (Ritchie et al., 2015). Gene set enrichment analysis (GSEA) was performed comparing K562 cells with GATA1 overexpression to K562 control cells with gene set permutation (Subramanian et al., 2005). The erythroid differentiation signature gene set was derived by identifying the top 200 genes that were expressed significantly higher in intermediate erythroblasts (CD71+/ CD235a+) compared to colony forming unit erythroid cells (CD71+/ CD235a-) (Merryweather-Clarke et al., 2011). The GATA1 target gene set was determined as previously described (Ludwig et al., 2014). Raw data have been deposited to the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) at GSE70531.

## Design and synthesis of a massively parallel reporter assay

Seventy-five GWAS hits associated with RBC traits were obtained from Table 1 in van Der Harst et al. (van der Harst et al., 2012). SNPs in high LD with sentinel GWAS hits were identified from the CEU population of the 1000 Genomes project Phase 1 Version 3 (phase1_release_v3.20101123) using Plink with the following options: "--show-tags --tag-r2 0.8 --tag-kb 500 --list-all" (Genomes Project et al., 2012; Purcell et al., 2007). All SNPs in high LD with rs12150672 were excluded since this resulted in an excessive number of constructs. 145 nucleotide constructs were designed by placing, for each of the 2756 variants, major and minor alleles into the construct such that 1/3, 1/2, and 2/3 of the total length was 5' of the variant (Table S1).

An oligonucleotide library containing the 145 nucleotide genomic regions with fourteen eleven-nucleotide barcodes each, separated by KpnI-XbaI sites and flanked by constant primer sites was obtained from Agilent (LeProust et al., 2010). The oligonucleotide library was then PCR amplified and cloned into the pMPRA1 plasmid backbone (Addgene Plasmid # 49349) with a minP-luc2 (Addgene Plasmid # 49353) insert, as previously described (Melnikov et al., 2014). The resulting plasmid library was introduced into K562 or K562+GATA1 cells using a Nucleofector II Device with Cell Line Kit V (Lonza). 48 hours later, total RNA was harvested and the barcodes were isolated by RT-PCR, as previously described (Melnikov et al., 2014). The barcodes were then sequenced and counted using an Illumina HiSeq 2500 sequencer.

**Analysis of a high-throughput variant screen**

A total of 6 replicates were performed for the MPRA screen in K562 cells and 4 replicates in K562+GATA1 cells. Left, middle, and right SWs were designed so that 1/3, 1/2, and 2/3 of the total 145 nucleotide construct was 3' of the interrogated variant. A pseudocount of 1 was added to DNA and RNA barcode counts that were subsequently normalized to counts per million (CPM) and $\log_2$ transformed. Barcodes with fewer than 8 transformed counts were removed from each replicate. Activity was calculated as the ratio of RNA and DNA counts, and the Pearson correlation coefficient was used to compare activity across replicates. Replicates were quantile normalized and combined as independent observations to increase power for each condition (K562 and K562+GATA1). The contribution of barcode bias to activity estimates was determined by randomly sorting n barcodes without replacement into two groups, where n is an integer between 1 and 7, averaging across replicates, calculating correlation coefficients, and using beta regression to estimate the bias for larger barcode numbers. ACs were defined as constructs that showed significantly higher activity (FDR < 1%, derived on all constructs and SWs) when compared with the activity distribution of all other constructs by a one-sided Mann-Whitney-U test. MFVs were identified by comparing activity between the constructs containing the major allele of the variant with constructs containing the minor allele using a two-sided Mann-Whitney-U test for K562 and K562+GATA1 MPRAs separately (FDR < 1%, derived on all constructs and SWs). GATA1-dosage dependence was determined by comparing the activity across each

construct between K562 and K562+GATA1 cellular models using a two-sided Mann-Whitney-U test.

**Additional bioinformatics analyses**

Manhattan plots were created from summary statistics and plotted using qqman (http://cran.r-project.org/web/packages/qqman/). A discriminatory k-mer based model was learned using the k-mer SVM webserver using 10-fold cross validation, a positive weight of 10 for ACs, and the regularization parameter C set to 0.5 (http://kmersvm.beerlab.org) (Lee et al., 2011). The positive training set was 555 sequences with high activity and the negative set was the remaining 11945 sequences. GATA1, TAL1, KLF1, and NFE2 ChIP-seq data in erythroid cells were processed and obtained as previously described (Ulirsch et al., 2014). Erythroblast raw ChIP-seq data of LDB1 was obtained from GSE52637, and raw data for H3K27me3 and H3K27ac were obtained from GSE52924; all data were processed similarly as previously described (Pinello et al., 2014; Stadhouders et al., 2014; Ulirsch et al., 2014). BedTools and R 3.2 were used to calculate all enrichments (Quinlan and Hall, 2010). PhastCons nucleotide conservation scores across 46 vertebrates were obtained from the Integrative Genomics Browser (IGV), and IGV was used to visualize MFVs and genome-wide sequencing data (Siepel et al., 2005; Thorvaldsdottir et al., 2013). Perturbations in DNA shape characteristics were calculated using DNAshape (Zhou et al., 2013). To investigate important measures of regulatory function for the MFVs, we compared them to cutting edge predictive algorithms including Eigen (Principal Component) (Ionita-Laza et al., 2016), DeepSea (Functional Significance Score) (Zhou and Troyanskaya, 2015), gkmer-SVM (Allelic skew, trained on K562 DNase 1 hypersensitivity) (Ghandi et al., 2014; Lee et al., 2015), and DeepSea (Allelic skew, trained on K562 DNase 1 hypersensitivity) (Zhou and Troyanskaya, 2015). Eigen and DeepSea (FunSig) are not innately directional for the allele, so a Mann-Whitney-U test was used to compare means between categories. gkmer-SVM and DeepSea (Allelic Skew) are directional, so the absolute value of predicted changes was compared. In order to identify a background set of SNPs for the distribution of tertiary DNA shape changes, we used Plink to identify all SNPs with a minor allele frequency > 5% in the CEU population of the 1000Genomes (Genomes Project et al., 2012; Purcell et al., 2007). Next, we used MACS2 to refine the set of narrow GATA1 peaks (Liu, 2014), and in these peaks (+/- 100bps from the center), we used Homer to identify SNPs that were within +/- 5bps of a GATA1 or GATA1/TAL1 motif, excluding SNPs that overlapped with the core “GATA” (Heinz et al., 2010). We obtained 382 SNPs for which we computed the total changes to DNA shape characteristics by summing up the absolute differences across each affected nucleotide. DHS skew for multiple cell-types was downloaded and analyzed as previously reported (Maurano et al., 2015). Allelic skew across erythroid TFs and open chromatin was first naively calculated by counting the number of aligned reads, and a sensitivity analysis was performed by using WASP (van de Geijn et al., 2015). Predicted motif disruptions were determined using both Transcription factor

Affinity Prediction Tools and HaploReg v3 with standard options (Thomas-Chollier et al., 2011; Ward and Kellis, 2012).

**Sensitivity and Specificity**

Since a gold standard set of positive and negative controls functional GWAS variants is not available, the positive predictive value of the MPRA screen was estimated by two orthogonal methods. First, we calculated the PICS probability score for all variants in the CEU population of the 1000 Genomes (Farh et al., 2015). We then derived credible sets of variants by greedily summing up the highest PICS probabilities at each locus until reaching a cumulative X%, assuming only one causal variant per locus. By definition, the X% credible set is expected to contain the causal variant X% of the time (Wellcome Trust Case Control et al., 2012). We compared the prevalence of all MFVs between the 80%, 90%, and 95% credible to the prevalence of MFVs in the corresponding non-credible sets. Since these sets split the variants into reasonably large credible and non-credible sets and the enrichments in MFVs were highly similar, we subsequently calculated the PPV from this enrichment as (1 – 1 / enrichment). As an alternative estimate, the methods DeepSea and gkmer-SVM were trained on K562 DNase1 hypersensitivity and used to predict the impact of swapping alleles for each variant on regulatory function. Since we expect an agreement between the direction of effects for MPRA and these predictive methods of no more than 50% (assuming a mix of true positive and false positives), an improvement in this represents enrichment in functional variants. In order to derive the PPV from this, we calculate (observed % agreement - expected % agreement) / (expected % agreement) for the set of MFVs. As a control, we also calculate this for AC/nMFVs to show that there is little improvement in agreement of directionality for variants in active constructs that we did not call as MFVs. Sensitivity was subsequently calculated from the PPV and the prevalence of MFVs (using the 74 tested GWAS associations rather than 75 total) identified by (PPV * count(MFVs) / 74).

**Luciferase Reporter Assay**

Firefly luciferase reporter constructs (pGL4.24) were generated by cloning the variant of interest centered in 300-400 nucleotides of genomic context upstream of the minimal promoter using BglII and XhoI sites. The Firefly constructs (500ng) were co-transfected with a pRL-SV40 Renilla luciferase construct (50ng) into 100,000 K562 cells using Lipofectamine LTX (Invitrogen) according to manufacturer's protocol. After 48 hours, luciferase activity was measured by Dual-Glo Luciferase assay system (Promega) according to manufacturer's protocol. For each sample the ratio of Firefly to Renilla luminescence was measured and normalized to the empty pGL4.24 construct**.**

**Generation of isogenic clonal deletions**

The erythroleukemia K562 cell line was cultured in RPMI medium with 10% FBS, Penicillin/ Streptomycin, and Glutamine. K562 cells were transfected with 1µg total of the Cas9 nuclease (pxPR_BRD001) and sgRNA plasmids (Table S7) using the Lipofectamine® LTX Plus Reagent in a 1:3 ratio of Cas9 to sgRNA. Control clones were obtained by co-transfecting K562 cells with 1 µg total of the Cas9 nuclease and pLKO.1-GFP at a 1:3 ratio of Cas9 to plasmid. Twenty-four hours after transfection, the cells were treated with puromycin for 48 hours (2 µg/ml) and isogenic clones were obtained by limiting dilution. Potential clones were PCR screened using primer pairs flanking the guide sequence (Table S7), and positive clones were Sanger sequenced to map the deletion.

**Identification of target genes**

RNA was extracted from selected clonal deletions of MFVs using the RNEasy® Plus Mini Kit (Qiagen), and cDNA was synthesized using the iScript™ cDNA Synthesis Kit (Biorad). RT-qPCR was performed with iQ™SYBR® Green Supermix (Biorad) on the CFX96TM Real-Time System (Biorad) (Table S7). Quantification was performed using the ΔΔCT method with β-actin as the reference gene.

**Primary cell culture**

Mobilized peripheral blood CD34+ cell culture was performed using a three-stage system that has been previously described (Hu et al., 2013). Cells were cultured using IMDM containing 2% human plasma, 3% human AB serum, 200 µg/ml human Holo-transferrin, 3 IU/mL heparin, and 10 mg/mL insulin (Base medium). During days 0 to 7, cells were supplemented with IL-3 (1 ng/mL), SCF (10ng/ml), and Epo (3 IU/ml). During days 7 to 12, cells were supplemented with SCF and Epo. After day 12, cells were supplemented with only Epo, and human Holo-transferrin was increased to 1mg/ml. Experiments were performed in triplicate using unique donors.

**Lentiviral vector production and transduction**

shRNA constructs targeting RBM38 were obtained from the Mission shRNA collection (Sigma-Aldrich), and constructs were in a pLKO.1-puro lentiviral vector Table S7). For the production of lentivirus, 293T cells were transfected with the appropriate viral packaging and genomic vectors (pVSV-G and pDelta8.9) using FuGene 6 reagent (Promega) according to the manufacturer's protocol. The next day, media was replaced with Base medium, and 24 hours later the lentiviral supernatant was collected and filtered using a 0.45 µm filter. On day 2 of primary cell culture, CD34+ cells were infected with lentiviral supernatants by spinfection. Between 150,000-500,000 cells were infected with viral supernatant in the presence of polybrene (8µg/ml) in a 6-well plate. Cells were spun at 2,000 r.p.m for 90 min at 22 °C and left in viral supernatant overnight. Media was replaced the morning after infection, and cells

were selected with 1 μg/ml puromycin 24 hours after infection. Puromycin selection was discontinued 48 hours later and cells were cultured as described above. For HMD-GATA1, we infected 500,000 K562 cells per well of a 6-well plate with 1.9ml of viral supernatant in the presence of polybrene (8μg/ml). Cells were spun at 2,500 rpm for 90 min at 22 °C and incubated overnight in the viral supernatant. Media was replaced 24 hours after infection, and 3 days later the cells were checked for GFP expression by flow cytometry to assess for infectivity, which was typically around 90%.

**Flow cytometry analysis**

Cells were washed in PBS and stained with human CD235a (GlyA), CD71, CD11b, CD41a, CD49d antibodies (Table S7). Propidium iodide (PI) was used as a dead cell marker. FACS analysis was conducted on a BD Bioscience LSR II and a BD LSR Fortessa. Data was analyzed using FlowJo X (TreeStar).

**May-Giemsa staining**

Approximately 100,000 – 200,000 cells were harvested, washed once at 300 x g for 5 minutes, resuspended in 130uL of FACS Buffer, and spun onto poly-L-lysine coated microscope slides with a Shandon 4 cytocentrifuge (Thermo Scientific) at 300 rpm for 4 min. Visibly dry slides were transferred into May-Grünwald solution (Sigma-Aldrich) for 5 minutes, rinsed 4 times for 30 seconds in water, and transferred to Giemsa solution (Sigma-Aldrich) for 15 min. Slides were washed as described above and mounted with coverslips. Images were taken with AxioVision software (Zeiss) at 63X oil magnification.

**Western blot Analysis**

Twenty-four hours after puromycin selection, cells were lysed in RIPA buffer and quantitated using DC Protein Assay (BioRad) according to manufacturer recommendations. To measure RBM38 knockdown, a western blot was performed on 20μg protein lysate using RBM38 antibody (C-19, sc-85873, Santa Cruz) and GAPDH (6C5; sc-32233, Santa Cruz) antibodies.

**RNA-seq**

RNA-seq was performed by the IDDRC Core Next-Gen Sequencing Facility of Boston Children's Hospital and Harvard Medical School in collaboration with Axeq Technologies. RNA was extracted at day 16 of culture using the RNEasy® Plus Mini Kit (Qiagen). An on-column DNase (Qiagen) digestion was performed according to the manufacturer's instructions. Approximately 1 μg of RNA from each sample was used to generate cDNA libraries for sequencing using the TruSeq RNA Sample Prep Kit v2 (Illumina, Inc., San Diego, CA). Sequencing of 101 nucleotide paired-end reads was performed on an Illumina HiSeq 2000 instrument. Adapters were removed with trimmomatic

using the following options: “PE -phred33 ILLUMINACLIP:TruSeq2-PE-2.fa:2:30:10 HEADCROP:5 LEADING:10 TRAILING:10 SLIDINGWINDOW:4:20 MINLEN:36”. Trimmed reads were aligned to the genome using Tophat2 and expression in FPKM was determined by using the Tuxedo suite as previously described (Trapnell et al., 2012; Ulirsch et al., 2014). RNA-seq from normal human erythropoiesis was obtained and processed as previously described (An et al., 2014; Li et al., 2014; Ulirsch et al., 2014). Percent spliced in (PSI) was determined for exon skipping events using SUPPA (Alamancos et al., 2015) on the gene transfer file (GTF) obtained by using Cuffmerge to create a single set of transcripts from both RNA-seq datasets created here as well as previously for normal human erythropoiesis. Differentially spliced exons were defined as exon skipping events with a >20% change in PSI between conditions (RBM38 sh1 and sh2 v. shLuc and pairwise OrthoE and PolyE v. ProE, eBasoE, and lBasoE). Gviz was used to create sashimi plot visualizations.

## Supplemental References


Alamancos, G.P., Pages, A., Trincado, J.L., Bellora, N., and Eyras, E. (2015). Leveraging transcript quantification for fast computation of alternative splicing profiles. Rna *21*, 1521-1531.

An, X., Schulz, V.P., Li, J., Wu, K., Liu, J., Xue, F., Hu, J., Mohandas, N., and Gallagher, P.G. (2014). Global transcriptome analyses of human and murine terminal erythroid differentiation. Blood *123*, 3466-3477.

Carvalho, B.S., and Irizarry, R.A. (2010). A framework for oligonucleotide microarray preprocessing. Bioinformatics *26*, 2363-2367.

Farh, K.K., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J., Shishkin, A.A*., et al.* (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature *518*, 337-343.

Genomes Project, C., Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

Ghandi, M., Mohammad-Noori, M., and Beer, M.A. (2014). Robust k-mer frequency estimation using gapped k-mers. Journal of mathematical biology *69*, 469-500.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576-589.

Hu, J., Liu, J., Xue, F., Halverson, G., Reid, M., Guo, A., Chen, L., Raza, A., Galili, N., Jaffray, J*., et al.* (2013). Isolation and functional characterization of human erythroblasts at distinct stages: implications for understanding of normal and disordered erythropoiesis in vivo. Blood *121*, 3246-3253.

Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J.D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet.

Lee, D., Gorkin, D.U., Baker, M., Strober, B.J., Asoni, A.L., McCallion, A.S., and Beer, M.A. (2015). A method to predict the impact of regulatory variants from DNA sequence. Nat Genet *47*, 955-961.

Lee, D., Karchin, R., and Beer, M.A. (2011). Discriminative prediction of mammalian enhancers from DNA sequence. Genome research *21*, 2167-2180.

LeProust, E.M., Peck, B.J., Spirin, K., McCuen, H.B., Moore, B., Namsaraev, E., and Caruthers, M.H. (2010). Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. Nucleic acids research *38*, 2522-2540.

Li, J., Hale, J., Bhagia, P., Xue, F., Chen, L., Jaffray, J., Yan, H., Lane, J., Gallagher, P.G., Mohandas, N.*, et al.* (2014). Isolation and transcriptome analyses of human erythroid progenitors: BFU-E and CFU-E. Blood *124*, 3636-3645.

Liu, T. (2014). Use model-based Analysis of ChIP-Seq (MACS) to analyze short reads generated by sequencing protein-DNA interactions in embryonic stem cells. Methods in molecular biology *1150*, 81-95.

Ludwig, L.S., Gazda, H.T., Eng, J.C., Eichhorn, S.W., Thiru, P., Ghazvinian, R., George, T.I., Gotlib, J.R., Beggs, A.H., Sieff, C.A.*, et al.* (2014). Altered translation of GATA1 in Diamond-Blackfan anemia. Nature medicine *20*, 748-753.

Maurano, M.T., Haugen, E., Sandstrom, R., Vierstra, J., Shafer, A., Kaul, R., and Stamatoyannopoulos, J.A. (2015). Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. Nat Genet.

Melnikov, A., Zhang, X., Rogov, P., Wang, L., and Mikkelsen, T.S. (2014). Massively parallel reporter assays in cultured mammalian cells. Journal of visualized experiments : JoVE.

Merryweather-Clarke, A.T., Atzberger, A., Soneji, S., Gray, N., Clark, K., Waugh, C., McGowan, S.J., Taylor, S., Nandi, A.K., Wood, W.G.*, et al.* (2011). Global gene expression analysis of human erythroid progenitors. Blood *117*, e96-108.

Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T.*, et al.* (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell *144*, 296-309.

Pinello, L., Xu, J., Orkin, S.H., and Yuan, G.C. (2014). Analysis of chromatin-state plasticity identifies cell-type-specific regulators of H3K27me3 patterns. Proc Natl Acad Sci U S A *111*, E344-353.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J.*, et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics *81*, 559-575.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research.

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J.*, et al.* (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317-330.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S.*, et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome research *15*, 1034-1050.

Slowikowski, K., Hu, X., and Raychaudhuri, S. (2014). SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. Bioinformatics *30*, 2496-2497.

Stadhouders, R., Aktuna, S., Thongjuea, S., Aghajanirefah, A., Pourfarzad, F., van Ijcken, W., Lenhard, B., Rooks, H., Best, S., Menzel, S.*, et al.* (2014). HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. J Clin Invest *124*, 1699-1710.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Thomas-Chollier, M., Hufton, A., Heinig, M., O'Keeffe, S., Masri, N.E., Roider, H.G., Manke, T., and Vingron, M. (2011). Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. Nature protocols *6*, 1860-1869.

Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics *14*, 178-192.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols *7*, 562-578.

Ulirsch, J.C., Lacy, J.N., An, X.L., Mohandas, N., Mikkelsen, T.S., and Sankaran, V.G. (2014). Altered Chromatin Occupancy of Master Regulators Underlies Evolutionary Divergence in the Transcriptional Landscape of Erythroid Differentiation. Plos Genetics *10*, e1004890.

van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J.K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nature methods *12*, 1061-1063.

van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Verweij, N., Sehmi, J., Paul, D.S., Elling, U., Allayee, H., Li, X.*, et al.* (2012). Seventy-five genetic loci influencing the human red blood cell. Nature *492*, 369-375.

Ward, L.D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic acids research *40*, D930-934.

Wellcome Trust Case Control, C., Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J.M., Auton, A., Myers, S.*, et al.* (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet *44*, 1294-1301.

Xu, J., Shao, Z., Glass, K., Bauer, D.E., Pinello, L., Van Handel, B., Hou, S., Stamatoyannopoulos, J.A., Mikkola, H.K., Yuan, G.C.*, et al.* (2012). Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. Developmental cell *23*, 796-811.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology *9*, R137.

Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nature methods *12*, 931-934.

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R., and Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic acids research *41*, W56-62.