Published in final edited form as:

Integr Biol (Camb). 2018 July 16; 10(7): 390–405. doi:10.1039/c8ib00059j.

An integrated global regulatory network of hematopoietic precursor cell self-renewal and differentiation

Yanan You^{#,1,2}, Raquel Cuevas-Diaz Duran^{#,1,2}, Lihua Jiang³, Xiaomin Dong^{1,2}, Shan Zong^{1,2,&}, Michael Snyder³, and Jia Qian Wu^{1,2,*}

¹The Vivian L. Smith Department of Neurosurgery, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

²Center for Stem Cell and Regenerative Medicine, UT Brown Foundation Institute of Molecular Medicine, Houston, TX 77030, USA

³Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

Abstract

Systematic study of the regulatory mechanisms of Hematopoietic Stem Cell and Progenitor Cell (HSPC) self-renewal is fundamentally important for understanding hematopoiesis and for manipulating HSPCs for therapeutic purposes. Previously, we have characterized gene expression and identified important transcription factors (TFs) regulating the switch between self-renewal and differentiation in a multipotent Hematopoietic Progenitor Cell (HPC) line, EML (Erythroid, Myeloid, and Lymphoid) cells. Herein, we report binding maps for additional TFs (SOX4 and STAT3) by using chromatin immunoprecipitation (ChIP)-Sequencing, to address the underlying mechanisms regulating self-renewal properties of lineage-CD34⁺ subpopulation (Lin-CD34⁺EML cells). Furthermore, we applied the Assay for Transposase Accessible Chromatin (ATAC)-Sequencing to globally identify the open chromatin regions associated with TF binding in the self-renewing Lin-CD34⁺EML cells. Mass spectrometry (MS) was also used to quantify protein relative expression levels. Finally, by integrating the protein-protein interaction database, we built an expanded transcriptional regulatory and interaction network. We found that MAPK (Mitogen-activated protein kinase) pathway and TGF- β /SMAD signaling pathway components were highly enriched among the binding targets of these TFs in Lin-CD34⁺EML cells. The present study integrates regulatory information at multiple levels to paint a more comprehensive picture of the HSPC self-renewal mechanisms.

*To whom correspondence should be addressed. Tel: +1 (713) 500-3421; Fax: +1 (713) 500-2424; jiaqian.wu@uth.tmc.edu.

[#]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Present Address: [Shan Zong], Division of Pediatrics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

CONFLICT OF INTEREST

The authors declare they have no conflict of interest.

ACCESSION NUMBERS

Raw ChIP-Seq and ATAC-Seq data as well as resulting filtered and annotated peak files have been deposited in the Gene Expression Omnibus (GEO) database under accession number GSE100689; access via <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100689> (reviewer token: kpgdgocwdzkpngv).

INTRODUCTION

The mammalian blood system is a highly heterogeneous system that contains Hematopoietic Stem Cell and Progenitor Cell (HSPCs), and more than ten differentiated cell types. HSPCs have both self-renewal capacity and the potential to differentiate into all types of hematopoietic cells¹. HSCs are among the most well-studied tissue-specific stem cells, and their analysis has improved our understanding of stem cell biology². Understanding the mechanisms regulating the switch between HSPC self-renewal and differentiation is important not only in stem cell biology, but also for manipulating HSPCs for therapeutic purposes³.

HSCs comprise only about 0.01% of all nucleated cells in the bone marrow, which makes studies such as proteomic and biochemical analyses that require large amounts of cells more difficult⁴. The mouse bone marrow-derived EML cell is a multi-potential hematopoietic precursor cell model that can differentiate into erythroid, myeloid, and lymphoid cells (EML)⁵. One subpopulation of EML cells, lineage-depleted CD34+ cells (referred as Lin-CD34+EML in this paper), can self-renew in a cell-autonomous fashion, while another subpopulation, lineage-depleted CD34- cells (referred as Lin-CD34-EML), can partially differentiate and has predominantly erythroid potential. Therefore, EML cell is an ideal model system for studying mechanisms that control the switch between self-renewal and differentiation^{3, 6}.

We have begun to study the regulatory mechanisms of self-renewal in EML cells on a genome-wide scale using Next-Generation Sequencing (NGS) technology^{3, 6-8}. In our previous study using RNA-Sequencing (RNA-Seq) gene expression analyses, chromatin immunoprecipitation in combination with high-throughput sequencing (ChIP-Seq), and gene knockdown experiments, we identified TCF7 and RUNX1 (AML1) as the key regulators of a transcriptional regulatory network that defines the Lin-CD34+EML cell state. We found that TCF7 and RUNX1 (AML1) bind to each other's promoter regions and that TCF7 and RUNX1 function coordinately to regulate self-renewal of Lin-CD34+EML cells. In addition to these two TFs, our previous RNA-Seq data showed there were other TFs with significantly different mRNA expression levels between Lin-CD34+EML and Lin-CD34-EML cells. These TFs might also play important roles in regulating Lin-CD34+EML self-renewal. However, the literature indicates that although the transcription of some genes is accompanied by concordant changes in their level of translation, the expression of mRNAs and proteins are not always correlated⁹. Also, it has been found that the joint analysis of mRNA and protein expression profiles could improve the insight when studying gene regulatory mechanisms¹⁰. Therefore, in the current study, we characterized the proteome of Lin-CD34+EML and Lin-CD34-EML cells by protein mass spectrometry (MS) and identified proteins that are differentially expressed between these two cell populations. In light of our previous RNA-Seq data, we further investigated the regulatory targets of two TFs, STAT3 (Signal Transducer and Activator of Transcription 3) and SOX4 (SRY-Box4). Previous literature indicated that STAT3 and SOX4 play important roles in regulating the proliferation and self-renewal of HSPCs¹¹⁻¹⁷.

This study focuses on characterizing the underlying mechanisms regulating self-renewal properties of Lin-CD34+EML cells. Here, we report not only new TF binding maps for Lin-CD34+EML cells, but also a global map of transcriptional regulation by including ATAC-Seq (Assay for Transposase Accessible Chromatin using Sequencing). ATAC-Seq is an ensemble method used to measure open chromatin regions in which prokaryotic Tn5 transposase inserts sequencing adapters into open chromatin to tag regulatory genomic regions¹⁸. In the present study, ChIP-Seq analysis of TFs combined with ATAC-Seq provided a genome-wide map of potential TF binding sites in Lin-CD34+EML cells.

Furthermore, although previous studies have described the regulatory network of self-renewal and differentiation in HSPCs at transcriptional or epigenetic levels^{19, 20}, the interplay among different levels of molecular regulation in the determination of stem cell fate remains unclear. In this paper, we have integrated epigenomic, transcriptomic, proteomic, and protein-protein interaction information to build a more comprehensive global regulatory network describing HPC self-renewal. This work may provide valuable clues for manipulating HSPCs and other differentiating cell systems.

MATERIALS AND METHODS

EML cell culture and separation of Lin-CD34+EML and Lin-CD34-EML populations

EML cells were cultured as previously described^{3, 6}. In brief, the culture medium contains EML basic medium (IMDM with 20% horse serum, 200 mM L-Glutamine, Penicillin/Streptomycin) supplement with 15% BHK culture supernatant. Maintain the EML cells at low cell density ($0.5-5 \times 10^5$ cells/ml) with the peak density less than 6×10^5 cells/ml. Split the cells every 2-3 days at the ratio of 1:5. For ChIP-Seq, protein mass spectrometry assays and real-time PCR, Lin-CD34+EML and Lin-CD34-EML populations were separated using magnetic-activated cell sorting (MACS)³. In brief, EML cell culture suspensions were spun down and washed twice with MACS buffer and resuspended in MACS buffer. Lineage-depleted cells were harvested using the mouse Lineage Cell Depletion Kit (Miltenyi Biotec) following the manufacturer's protocol as follows. Biotin-conjugated mouse lineage antibody cocktail was added to the cell suspension and incubated at 4 °C for 10 min and then Anti-Biotin MicroBeads were added to the suspension and incubated for an additional 15 min at 4 °C. After washing twice, cells were resuspended in MACS buffer and the magnetically labeled cells (lineage positive cells) were captured with a MACS Separator (Miltenyi Biotec). The flow-through cells were then labeled with biotin-conjugated mouse CD34 antibody and incubated with Anti-Biotin MicroBeads. CD34+ cells were captured using the MACS Separator and the CD34- cells were harvested in the flow-through buffer.

For ATAC-Seq assays, CD34+ and CD34- cells were separated by FACS as follows. EML cells were washed twice and resuspended with FACS buffer (0.5% FBS in PBS) and then Lineage Cell Detection Cocktail-Biotin (Miltenyi Biotec) was added to the cell suspension and incubated at 4 °C for 10 min. After washing with FACS buffer twice, Anti-Biotin APC antibody (Miltenyi Biotec) was added to the suspension and incubated for additional 15 min at 4 °C. CD34-FITC antibody (eBiosciences) was added to the cells and incubated for 1 h at 4 °C. CD34+ and CD34- cell populations were analyzed and sorted using a FACS Aria™ II Cell Sorter (BD Biosciences).

Antibodies

Antibodies SOX4 (C-20) X (sc-17326x), STAT3 (C-20) X (sc-482x), and normal Rabbit IgG (sc-2027) used in ChIP and western blot experiments were purchased from Santa Cruz Biotechnology. Anti-Biotin APC antibody (Miltenyi Biotec) and CD34-FITC antibody (eBiosciences) were used for cell separation.

Real-time PCR and Western blot

RNA extraction was performed using TRIzol™ reagent and DNase-treated RNA was reverse transcribed using the iScript cDNA Synthesis Kit according to the manufacturer's instructions (Bio-Rad). Quantitative real-time PCR was performed using SYBR® Green Master Mix (Bio-Rad) on the ABI PRISM® 7900HT Sequence Detection System. The 2-

CT method was used to calculate relative gene expression with expression of the *Gapdh* gene as reference. Primer sequences for these experiments can be found in Supplemental Table S6.

Total proteins were extracted from Lin-CD34+EML and Lin-CD34+EML cells following the methods used in our previous study³. Complete Protease Inhibitor Cocktail (Roche) was added to RIPA lysis buffer (Millipore) following the manufacturer's instructions. Western blot experiments were also performed as previously described³. After protein transfer, SOX4 (C-20) X (sc-17326x, Santa Cruz Biotechnology) and STAT3 (C-20) X (sc-482x, Santa Cruz Biotechnology) antibodies were added to the PVDF membrane at 1:1000 dilution, respectively.

Protein mass spectrometry analysis

Proteins were extracted in a buffer containing a combination of 4% SDS and urea. The in-solution tryptic digestion was performed on reduced or reduced-alkylated samples at 37 °C overnight. Digested peptides were differentially labeled using the TMTsixplex™ Isobaric Mass Tagging Kit (Thermo Fisher Scientific) according to the manufacturer's instructions. Identical quantities of labeled peptide mixtures from all six samples were pooled. The digested peptide pool was then analyzed on a Waters 2D nano liquid chromatography system (Waters NanoAquity 2DnLC) interfaced with a Linear Trap Quadrupole (LTQ) Orbitrap Velos Mass Spectrometer (Thermo Fisher Scientific). MS spectra were acquired in the Orbitrap at a resolution of 60,000 units (FWHM) using full-range zoom (300–1800 mass-to-charge ratio [m/z]). The ten most intense ions were sequentially isolated and fragmented in the HCD cell and detected in the Orbitrap at a resolution of 7500 units (FWHM). MS data for these top 10 most abundant precursor ions were acquired using a data-dependent acquisition method. MS experiments were performed in triplicates.

All MS spectra were queried using the SEQUEST program²¹ against the mouse database concatenated with reversed copies of all sequences to provide an estimate of the false-positive rate of peptide identification by SEQUEST under various parameters. FDRs for peptides and proteins were set to 1%. Only proteins represented by at least two peptides were considered reliably identified. Unique peptides were then considered for quantification.

Significance analysis of protein expression using SAM

Significance analysis of microarrays (SAM) is a statistical algorithm originally developed for finding genes with significant differential expression in a set of microarray experiments²². Previous studies have adapted this method for effective analysis of proteomic data²³. SAM calculates observed and expected scores for each protein. The observed score represents the difference in protein abundances between the Lin-CD34+EML and Lin-CD34-EML cell samples. The expected score represents the random fluctuation that occurs when there is no difference between the two samples. The differential expression of a protein is considered significant if the difference between the observed and the expected scores is beyond a certain threshold.

We used the SAMR package to perform SAM analysis (installed from <http://statweb.stanford.edu/~tibs/SAM/>) on the relative protein abundance data (log2-transformed ratios of CD34+/CD34- protein abundances) from triplicate samples as two-class unpaired data with 100 permutations and FDR of 2.2%. Imputation of missing values was performed using the K-Nearest Neighbor imputation method with K = 10 neighbors. Figure 1A shows the results from the SAM analysis in which $\lambda = 0.4$ was chosen as the significance cutoff according to the false positive rate.

Analysis of differential gene expression

To compare gene expression between Lin-CD34+EML and Lin-CD34-EML samples, we filtered genes leaving only those with RPKM >1 in at least one sample. Next, we used Limma to compare the gene expression between samples²⁴ and designated genes with fold-change >1.5 and p -value <0.001 as differentially expressed genes.

ChIP-Seq and data analysis

Chromatin immunoprecipitation (ChIP) was performed as described in our previous study²⁵. A total of 10^7 cells were used per ChIP experiment. Cells were cross-linked in 1% formaldehyde for 10 min at room temperature (RT) with rotation followed by addition of 0.125 M glycine to quench the cross-linking reaction. Cells were pelleted and washed twice with ice-cold PBS. Nuclei were isolated using nuclei lysis buffer (50 mM Tris-Cl, 10 mM EDTA, 1% SDS) and chromatin DNA was then sheared using a Bioruptor® sonicator (Diagenode). A sample of 1/100 volume of the sheared DNA was reserved for use as the input control sample. The STAT3 (C-20) X (sc-482x) and SOX4 (C-20) X (sc-17326x) and normal Rabbit IgG (sc-2027) antibodies from Santa Cruz Biotechnology were added to Dynabeads Protein A (Invitrogen) beads and incubated for 4 h at 4 °C with rotation. The Dynabeads-antibody complexes were then incubated with the sheared chromatin DNA overnight at 4 °C. After immunoprecipitation, RNase A and Proteinase K were added to the precipitated complex and the samples were incubated at 65 °C overnight with shaking to reverse the crosslinks. The ChIP-Seq libraries were prepared using the DNA SMART ChIP-Seq Kit (Clontech Laboratories) according to the manufacturer's instructions and Single-end 50-cycle sequencing reads were obtained on the Illumina HiSeq 2000 Sequencer. The input control sample was prepared in parallel with the ChIP samples.

The STAT3 and SOX4 ChIP-Seq samples and their corresponding input samples were mapped to the mm10 mouse reference genome using Bowtie2²⁶ with default parameters. Quality control plots to address clonality and autocorrelation were obtained using Homer²⁷ and the R package SPP²⁸. Peaks were called using MACS1.4²⁹ and filtered excluding ENCODE's blacklist regions (<https://sites.google.com/site/anshulkundaje/projects/blacklists>) and peaks with FDR >0.05 and fold-enrichment <10. Filtered peaks were annotated to the RefSeq mm10 annotation file using the annotatePeaks.pl function in Homer (<http://homer.ucsd.edu/homer/ngs/annotation.html>). Annotated peaks lying within 5 kb upstream of the TSS of a gene and within its gene body were retained for downstream analysis. ChIP-Seq peaks for the TFs TCF7 and RUNX1 were downloaded from³, converted to their genomic coordinates in the mm10 assembly using the UCSC LiftOver Utility³⁰, and integrated with STAT3 and SOX4 ChIP-Seq data.

ATAC-Seq and data analysis

ATAC library of Lin-CD34+EML cells was prepared following the previously published method³¹. In brief, 50,000 cells separated by FACS were spun down and washed with cold PBS and then 50 µl of cold lysis buffer was added to the cell pellet. The Nextera DNA Library Prep Kit (Illumina) was used to perform the transposase reaction. The fragmented DNA was purified using a Qiagen PCR Purification Kit (Qiagen). Purified DNA fragments were then amplified using NEBNext High-Fidelity PCR Master Mix (New England Biolabs). qPCR was performed to determine the appropriate number of additional PCR cycles, allowing amplification to be stopped prior to saturation. Purified libraries were analyzed for quality using a Bioanalyzer High-Sensitivity DNA Analysis Kit (Agilent). Libraries were quantified using the KAPA Library Quant Kit for Illumina Sequencing Platforms (KAPA Biosystems). Paired-end 100-cycle sequencing reads were obtained on the Illumina HiSeq 2000 Sequencer.

We used the ENCODE ATAC-Seq pipeline, which is described at <http://www.encodeproject.org/atac-seq/> and can be obtained from https://github.com/kundajelab/atac_dnase_pipelines, to process the ATAC-Seq samples. Paired-end reads were mapped to the mm10 reference genome using the Bowtie2 aligner²⁶ with -X2000 for the alignment of fragments up to 2 kb in size, and with -k 4 to allow up to four multiple alignments. Next, data was filtered to remove any reads that are unmapped, reads with unmapped mates, reads that are not primary alignments, low quality reads (MAPQ <30), reads that map to the mitochondrial genome, and PCR duplicates. SAMtools³² and Picard tools were used to filter reads (<http://broadinstitute.github.io/picard>). Remaining reads were offset by +4 and -5 for positive and negative strands, respectively, to accurately indicate the center of each transposon-binding event³³.

We used MACS2²⁹ for peak calling with parameters “-p 0.01 -nomodel --shift 37 -extsize 73 -broad -keep-dup all”. We generated larger enrichment regions known as 'broadPeaks' and 'gappedPeaks'. To obtain punctate peaks (narrowPeaks) we used MACS2 with parameters “-p 0.01 -nomodel --shift 37 -extsize 73 -B -SMPR -keep-dup all --call-summits”. The resulting broad, gapped, and narrow peaks were further filtered based on ENCODE's blacklist. Peaks were annotated using the HOMER annotatePeaks function with

GENCODE's GRCm38.p4 basic gene annotation downloaded from <https://www.gencodegenes.org/>. The Integrative Genomics Viewer (IGV)³⁴ was used to browse and compare peak regions among samples. All genomic datasets were deposited in GEO (accession number GSE100689).

Analysis of ATAC-Seq peak regions for TF binding motifs was performed using FIMO³⁵ (FDR <0.05) with position weight matrices downloaded from <http://compbio.mit.edu/encode-motifs>. We queried motifs for 617 TF families as classified by Kheradpour et al (2014)³⁶. We further filtered the list of TF binding motifs by eliminating those whose motifs were found in less than 1% of all open chromatin regions defined by ATAC-Seq.

Gene Set enrichment analysis (GSEA)

In order to correlate the TF binding targets with gene expression data, we performed Gene Set Enrichment Analysis (GSEA)³⁷. Using a floor function, we assigned a minimum RPKM = 0.1 and we ranked the genes according to fold-change in expression such that the most upregulated genes in Lin-CD34+EML cells were at the top of the list and the most downregulated genes were at the bottom. We used this ranked list of genes and fold-change values to determine whether any of the TF binding targets were significantly enriched among genes upregulated in Lin-CD34+EML samples.

Circos plot for the visualization of ChIP-Seq binding targets and gene expression

We built a Circos plot using the RCircos R library³⁸. The outermost circular track is a cytoband that represents the chromosome ideograms in the mm10 genome. Labels on the cytoband indicate differentially expressed genes generated using Limma (p-value <0.001, fold-change >1.5, and RPKM >1 in at least one sample) 24. The second circular track represents the gene expression heatmap of log2-transformed RPKM values in Lin-CD34+EML cells. Tracks 3–6 display the fold-enrichment of ChIP-Seq binding peaks for TCF7, RUNX1, STAT3, and SOX4, respectively.

Protein-protein and protein-DNA interaction network

We downloaded the mouse known protein-protein interaction database (GeneRIF) from the NCBI repository <ftp://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/> and combined it with TCF7, RUNX1, SOX4, and STAT3 binding target information. We compiled a database of 1966 genes with 4009 interactions and constructed a network using Cytoscape³⁹. All genes (nodes) in the network and the number of interactions (edges) are listed in Supplemental Table S4. Green lines indicate known protein-protein interactions and pink lines illustrate TF-DNA interactions identified using ChIP-Seq. We integrated into the regulatory network the open and closed chromatin conformation states obtained from ATAC-Seq assay as node border color (green = open, black = closed). Additionally, we included different node shapes depicting the consistency between the protein and mRNA abundance levels (ellipse = protein and mRNA level changes are consistent, diamond = protein and mRNA changes are opposite, octagon = protein levels did not change while mRNA levels are up/downregulated, rounded rectangle = protein was not detected).

Gene Ontology analysis

We used BiNGO⁴⁰ to determine which GO terms in the Biological Process domain were statistically overrepresented (p -value $< 1 \times 10^{-10}$) among the top 100 genes with the most interactions in our protein-protein and protein-DNA interactions network. We visualized the relationships between the GO categories using Cytoscape and grouped them according to functional similarity.

Graph-based clustering

We performed graph-based clustering using the Walktrap community algorithm⁴¹ for finding subnetworks from our regulatory interaction network. Enriched gene-sets for gene clusters with more than 10 members were obtained (FDR < 0.001 , p value < 0.00001 , and number of genes in set > 10) using a hypergeometric function. Resulting clusters and enriched gene-sets are listed in Supplemental Table S5. For visualization purposes, we have selected biologically interesting local subnetworks as examples derived from clusters with higher number of enriched gene-sets and displayed them in Figure 4.

RESULTS

Differences in protein abundances between Lin-CD34+EML and Lin-CD34-EML cells identified by using mass spectrometry

To identify differentially translated proteins in Lin-CD34+EML and Lin-CD34-EML cells, we extracted proteins from these two cell populations and conducted mass spectrometry (MS) protein profiling. We adopted the SAM (Significance Analysis of Microarray) method^{22, 23} to analyze MS data for proteins from the two populations. A total of 3949 distinct peptides, including one or more peptides from the same protein, were identified from all samples. Proteins with fewer than two uniquely matched peptides in all three replicates were excluded from further analysis yielding 2646 proteins.

These 2646 proteins were then subjected to SAM analysis to test whether the log₂-transformed relative protein abundance ratios between the two subpopulations were significantly different. Figure 1A shows the resulting plot of the observed and expected SAM scores at $\lambda = 0.4$ and a false discovery rate (FDR) $< 2.2\%$. The expected score depicts random fluctuation when the difference between the two samples is not significant. The 45° line represents proteins that are not differentially expressed between samples. A protein is considered significantly differentially expressed if the difference between the observed and the expected scores is beyond a predetermined threshold. Differentially expressed proteins are highlighted in green (401 proteins with decreased expression in CD34+ cells compared to CD34- cells) or red (484 proteins with increased expression in CD34+ cells compared to CD34- cells). Figure 1B shows the distribution of q -values resulting from SAM analysis. As observed, at 2129, the q -value is approximately 30, with an FDR $< 2.2\%$. Among the 2646 proteins we identified, 2192 are derived from genes expressed at RPKM > 1 in all of the samples. RNA-Seq data were obtained from our previous study³.

Combining the RNA-Seq and the protein relative expression data, we found 742 proteins with different relative abundances between Lin-CD34+EML and Lin-CD34-EML cells and

with gene expression RPKM >1. As depicted in Figure 1C, there are 403 proteins with higher expression and 339 proteins with lower expression in Lin-CD34+EML (= 0.4 and FDR = 2.2%) relative to Lin-CD34-EML cells. In order to address whether relative protein abundance and gene expression were consistent with one another, we compared the fold-changes of the log₂-transformed relative protein abundance ratios (Lin-CD34+EML/Lin-CD34-EML) to log₂-transformed RPKM gene expression values as shown in Figure 1D. If the changes in direction of both gene expression and relative protein expression were the same, the expression of the protein was designated as 'consistent', otherwise it was designated as 'opposite'. Genes with no significant changes in mRNA or protein levels were designated as 'mRNA-No-Change' or 'Protein-No-Change'. Consistent results were obtained for 64 downregulated proteins and 66 upregulated proteins in Lin-CD34+EML relative to Lin-CD34-EML cells. We identified 31 proteins that were expressed at lower levels in Lin-CD34+EML cells and 64 proteins that were expressed at higher levels in Lin-CD34+EML cells; we characterized both of these sets of proteins as having 'opposite' expression. SERPINF1 and HMGA2 are examples of proteins with 'opposite' expression, downregulated in CD34+ cells. ITGB1, ZBTB1, and RPS6 are examples of proteins upregulated in CD34+ cells and downregulated in CD34- cells. A full list of proteins identified in each expression category is provided in Supplemental Table 1 (Table-S1).

Studying the mechanisms of self-renewal and differentiation by identifying the *in vivo* binding targets of TFs

In our previous study using RNA-Seq, we identified genes that are differentially expressed between Lin-CD34+EML and Lin-CD34-EML cells^{3, 42}. Among these genes, *Tcf7* and *Runx1* were identified as key regulators of EML cell self-renewal. Using ChIP-Seq, We found that RUNX1 and TCF7 bind to each other's promoter regions and that they share many common target genes. We selected SOX4 and STAT3 for further study of TFs that potentially regulate self-renewal of Lin-CD34+EML cells because of reports that STAT3 plays important roles in HSC self-renewal and that overexpression of SOX4 in primary bone marrow cells induces enhanced HSC activity and repopulating activity^{11, 43}. Both SOX4 and STAT3 were differentially expressed (log₂-transformed FC = 3.9 and 2.4, respectively) in Lin-CD34+EML compared to Lin-CD34-EML cells in our RNA-Seq data. MS results were not available for SOX4. However, western blot results revealed relatively higher abundance of SOX4 protein in Lin-CD34+EML cells (Figure-S1). Similarly, we selected STAT3 because both RNA-Seq and western blot results indicated higher levels of expression of its gene and protein in CD34+ cells compared to CD34- cells. To better understand how these TFs are involved in Lin-CD34+EML self-renewal, we identified the *in vivo* binding sites for these factors using ChIP-Seq. ChIP-Seq experiments generated 4553 STAT3 binding peaks and 9660 SOX4 binding peaks in Lin-CD34+EML cells. These binding peaks were mapped to RefSeq genes in the mm10 database (<https://www.ncbi.nlm.nih.gov/refseq/>) and were then filtered according to the ENCODE blacklist (<https://sites.google.com/site/anshulkundaje/projects/blacklists>). The majority of peaks were filtered out, leaving only those peaks with FDR values <0.05 and fold-enrichment >10. Binding peaks were assigned to a particular gene if the peak was found within 5 kb upstream of the transcription start site or within the gene, including both exonic and intronic regions. We obtained 885 genes identified by binding peaks of STAT3, and 3016 genes with peaks of SOX4.

To understand the biological functions of the genes regulated by STAT3 and SOX4, we searched for enriched gene sets using a hypergeometric test. A complete collection of gene sets was downloaded from the MSigDB database (<http://software.broadinstitute.org/gsea/msigdb>). We filtered gene sets with FDR <0.25 (as recommended in the GSEA manual). The STAT3 and SOX4 targets resulted in highly enriched (p -value <0.0001 and p -value <0.001, respectively) gene sets related to hematopoiesis ('Bystrykh hematopoiesis stem cell qtl trans', 'Ivanova hematopoiesis stem cell and progenitor', 'Jaatinen hematopoietic stem cell up', 'GO hematopoietic progenitor cell differentiation', and 'GO regulation of cell cycle'). Figure 2A shows the enriched gene sets associated with STAT3 and SOX4 binding targets.

We correlated the ChIP-Seq targets of STAT3 and SOX4 with the previous TCF7 and RUNX1 ChIP-Seq peaks and identified 93 genes in common. To explore the biological processes that might be co-regulated by TCF7, RUNX1, STAT3, and SOX4, we then examined the enriched gene sets associated with the common target genes. Gene sets associated with 'cell cycle process', 'positive regulation of catalytic activity', 'positive regulation of molecular function', and 'Ivanova hematopoiesis stem cell and progenitor' were highly enriched (Figure 2A). Figure 2B shows the binding regions for TCF7, RUNX1, STAT3, and SOX4 within *Zbtb7a*, a gene that encodes a protein that is involved in regulating self-renewal of HSC by blocking Notch1-mediated T cell differentiation⁴⁴. TG-interacting factor 1 (*Tgif1*) is another gene regulated by all four of these TFs. *Tgif1*, a transcriptional repressor in TGF β signaling pathways, modulates the balance among quiescence, self-renewal, and differentiation in HSCs.

ATAC-Seq reveals nucleosome positions and open chromatin regions

We used ATAC-Seq to correlate open chromatin regions with gene expression profile in Lin-CD34+EML cells, and identify potential TF regulatory sites. ATAC-Seq provides information about nucleosome packing as well as positioning. The insert size distribution shown in Figure 2C depicts a clear periodicity of approximately 200 bp, consistent with previous results by Buenrostro et al.¹⁸, in which they inferred that several DNA fragments were protected by integer multiples of nucleosomes. The majority of the fragments in the ATAC-Seq library were smaller than 100 bp, representing internucleosome open chromatin regions. Fragments of approximately 160–240 bp were considered to be located within mononucleosome regions. Fragments of 360–440 bp span dinucleosomes, and reads of 560–640 were considered to be within trinucleosomes. We found 567,083 uniquely filtered regions classified as 'broadPeaks' because they represent large regions of enriched signal. We also obtained 398,177 regions classified as 'narrowPeaks' that represent punctate (sharp) peaks. Figure 2D shows filtered ATAC-Seq peaks that mapped within 5 kb of the transcriptional start sites (TSS) of known genes. As shown in the histogram in Figure 2D, a high proportion of the peaks in this region are within 500 bp of the TSS, indicating the accessible chromatin regions within promoters. After mapping all filtered ATAC-Seq peaks to genomic regions, we found a higher proportion of peaks located in intronic and intergenic regions (40% and 42.6%, respectively), as depicted in Figure 2E. A similar distribution of TF binding peaks within genomic regions has been reported in a different study⁴⁵. The high proportion of peaks in intronic and intergenic regions could indicate enhancers.

We searched within open chromatin regions (broadPeaks) of Lin-CD34+EML sample for the genome-wide set of motifs obtained from the ENCODE motif repository (<http://compbio.mit.edu/encode-motifs>). We obtained 28 TFs whose motifs were significantly enriched (FDR <0.05) and were found in more than 1% of the ATAC-Seq open chromatin regions. Supplemental Table S2 depicts the TFs, binding motifs, p-values, q-values, and the percentage of open chromatin regions in which these motifs were enriched. Motifs for STAT3 were most commonly found in all queried open chromatin regions. There is evidence supporting the roles of many of the resulting TF binding motifs in hematopoiesis. For example, FOXP1 signaling drives expansion of HSPCs, FOXP2 participates in erythroid differentiation, and SP1 functions at early stages of hematopoietic specification^{46–48}.

Predicting potential key regulators involved in HSC self-renewal by integration of ChIP-Seq, RNA-Seq and ATAC-Seq results

To determine whether there is a correlation between STAT3 and SOX4 binding targets and the differential expression of genes in Lin-CD34+EML compared with Lin-CD34-EML cells, we performed a Gene Set Enrichment Analysis (GSEA)³⁷. We ranked the genes according to their log2-transformed fold-change so that the top-ranked genes corresponded to genes upregulated in Lin-CD34-EML cells. GSEA showed a statistically significant enrichment of STAT3 (Enrichment Score ES <0.51, nominal *p*-value = 0, and FDR = 0) and SOX4 (Enrichment Score ES <0.45, nominal *p*-value = 0, and FDR < 0.011) targets among genes that are more highly expressed in Lin-CD34+EML cells as compared to a random distribution. The enrichment plots obtained using GSEA are shown in Figure 3A and 3B.

We also built a Circos plot to correlate ChIP-Seq binding targets and gene expression (Figure 3C). There are a number of genes that are common targets of all four TFs and are also in the open chromatin regions, such as *Eya2*, *Tgif1* and *Dapp1*. These genes play important roles in HSC/HSPC self-renewal and proliferation. Figure 3C shows *Eya2* as an example.

Limma analyses yielded 763 differentially expressed genes, with 431 and 332 genes upregulated in Lin-CD34+EML and Lin-CD34-EML, respectively. We examined the gene expression levels of TCF7, RUNX1, STAT3, and SOX4 targets and found that about two thirds of them were significantly upregulated in Lin-CD34+EML cells. Among TCF7's targets, 253 were significantly upregulated in Lin-CD34+EML and 165 were upregulated in Lin-CD34-EML cells. Similarly, 287 targets of RUNX1 were more highly expressed in Lin-CD34+EML cells and 139 in Lin-CD34-EML cells. The gene expression of 37 STAT3 targets was significantly upregulated in Lin-CD34+EML cells and that of only 17 targets was downregulated. Within SOX4 target genes, 56 were upregulated and 35 were downregulated in Lin-CD34+EML cells.

To analyze the relationship between the open chromatin regions, TF binding targets, and gene expression in self-renewing Lin-CD34+EML cells, we integrated the ATAC-Seq, ChIP-Seq, and RNA-Seq data sets. The 398,177 'narrowPeaks' were filtered to obtain only those lying within 5 kb of the TSS and within a known gene body. The remaining filtered peaks were mapped to 20,350 genes. We then searched for ATAC-Seq peaks within the gene bodies and promoters of the top 50 differentially expressed genes upregulated in CD34+

cells. We found that 41 of these 50 genes contain ATAC-Seq peaks within the gene body and 44 of these 50 genes have peaks in their promoter regions, defined as 5 kb upstream of the TSS. Among the top 25 genes most highly upregulated in Lin-CD34+EML cells relative to Lin-CD34-EML cells, many, such as *Mpo*, *Thsd1*, *Tcf7*, *Serpina3g*, *Hmga2*, *Cd34*, *Gadd45b* and *Camk2a* are highly expressed in HSCs or are involved in HSC biology^{49–52}. *Tcf7* and *Mpo* are targets of TCF7 and RUNX1 and have ATAC-Seq peaks both in their gene body and promoter region. *Cd34* is a target of TCF7, RUNX1, and SOX4 and also has ATAC-Seq peaks in both the gene body and promoter region (Table S3).

An expanded transcriptional regulatory and interaction network of HSPC self-renewal

We combined the protein abundance levels, RNA-Seq expression data, ChIP-Seq binding targets, ATAC-Seq chromatin accessibility, and known mouse protein-protein interactions into a transcriptional regulatory and interaction network. The resulting network was composed of 1966 nodes with 4009 interactions. A total of 1933 nodes (98% out of 1966) correspond to genes in open chromatin regions. Table S4 shows the complete list of nodes in the protein-protein-DNA interaction network, the chromatin accessibility status, and the protein-mRNA abundance consistency. Interestingly, in addition to *Tcf7*, *Runx1*, *Stat3*, and *Sox4*, genes such as *Myc*, *Apc*, *Lmo2*, *Jun*, and *Runx2* are in open chromatin regions and are among the top 20 nodes with highest number of interactions.

We used a graph-based clustering method for finding subnetworks from our inferred regulatory network. The resulting network partition yielded 8 clusters with more than 10 genes. To determine the biological relevance of the clusters, we performed gene-set enrichment. Clusters 1, 3, and 12 had the highest number of enriched gene-sets: 81, 50, and 33 gene-sets respectively. Enriched gene-sets and clusters are listed in Table S5. Figures 4A and B show selected interesting local sub-networks derived from clusters 1 and 12. TGF- β /SMAD signaling pathway members including *Tgfb1*, *Tgfb1/2*, and *Smad 1/4/5/7* transcripts are expressed at higher levels in Lin-CD34+EML cells and according to ChIP-Seq results (Figure 4A). The expression of most of the proteins that directly interact with SMADs was also relatively higher in Lin-CD34+EML cells, including BMPR2, SMURF1, SMURF2, and RUNX2, which are involved in the BMP signaling pathway, and APC, FOXP1, and JUN, which are involved in apoptosis and regulation of cellular differentiation^{53–56}. Other interesting local sub-networks include a cluster of apoptosis pathway members, such as *Bcl3*, *Bcl9*, *Myc*, *Bcl-2l11*, *Bcl-2l1*, and also genes involved in the cell cycle, such as *Ccnd1*, *Cdk6*, and *Cdkn1a* (Figure 4B). These targets of TCF7, RUNX1, and STAT3 and their immediate neighbors were all more highly expressed in Lin-CD34+EML cells. These results show that TCF7, RUNX1, SOX4, and STAT3 might affect proliferation and self-renewal of Lin-CD34+EML cells *via* various signaling pathways. ChIP-qPCR was performed to validate the binding between TFs and a number of genes involved in important network and pathways, including *Mapk3k3*, *Tgfb2*, *Smad4*, *Smad5*, *Bcl3* and *Grb2* (Figure 4C). Significant enrichment over genomic input DNA was detected in SOX4 and STAT3 ChIP samples.

Next, we used Cytoscape³⁹ and BiNGO⁴⁰ to explore the most significantly enriched biological processes defined by the top 100 nodes with the most interactions. Figure 5 shows

the most significantly enriched Gene Ontology (GO) terms (p -value $< 1 \times 10^{-10}$) in Lin-CD34+EML and Lin-CD34-EML cells, which included 'metabolic processes', 'signaling', 'biological regulation', 'response to stimulus', and 'cell development and differentiation'.

Our integrated dataset indicated that many MAPK signaling pathway components such as *Pdgf-a/b*, *Pdgfra/β*, *Grb2*, *Ras*, *Raf*, *Mapks*, and *Elk* are binding targets for TCF-7, RUNX1, or STAT3. Further, the expression of most of these genes was relatively higher in Lin-CD34+EML cells. Many downstream target genes of the MAPK pathway that function in cell cycle regulation and cell proliferation, such as *c-Fos*, *Cdks*, and several *Cyclins*, were also upregulated. In addition to the ERK-MAPK pathway, some components of the JNK/SAPK- and p38- MAPK pathways, for example *Mapk9/11/14* and *Map3k1/5*, were also targets of TCF-7, RUNX1, STAT3, or SOX4. In addition, the ERK5-MAPK pathway gene *Map2k5* was a target of TCF-7, RUNX1, and SOX4, and its expression was also relatively high in Lin-CD34+EML cells (Figure 6).

DISCUSSION

Stem cell fate decisions involve multiple layers of molecular regulation, including epigenetic, transcriptional, and translational mechanisms, and coordinated interactions among them. Therefore, in this study, we integrated proteomic, epigenetic, and transcriptomic analyses to build a more comprehensive regulatory network of HSPC self-renewal and we identified SOX4 and STAT3 as potential important regulators in the decision between self-renewal and differentiation. Due to the relatively lower sensitivity of protein MS compared to RNA-Seq, we detected far fewer proteins using MS than we did mRNA using RNA-Seq. Combining the RNA-Seq analysis and relative protein expression data, we found that changes in the expression of mRNAs and proteins from 130 genes were 'consistent'. In contrast, 95 genes were labelled as exhibiting 'opposite' changes in expression of the mRNA and protein. The list of genes exhibiting 'opposite' changes in mRNA and protein expression included some interesting genes involved in hematopoiesis, which implies that translational or post-translational regulation is important in HSPC biology.

The EML cell system is ideal for studying bulk cultures of hematopoietic precursor cells that undergo self-renewal and differentiation without the context of a stem-cell niche⁶. We previously reported gene expression and several TF-binding profiles in the EML system using next-generation sequencing^{3, 42}. By integrating the transcriptomic and proteomic data in the current study, the potential regulatory targets of two additional TFs were investigated. STAT3, a member of the STAT family, is a signaling molecule that transduces ligand-induced signals to activate specific target genes. Studies have described the roles of STAT3 in various tissues including hematopoietic cells, such as regulation of HSPC proliferation^{11–14}. In the present study, Gene set enrichment analysis showed that gene sets related to self-renewal and proliferation were highly enriched among STAT3 targets, suggesting that STAT3 plays a role in Lin-CD34+EML cell self-renewal. SOX4 is a member of the C SOX (SOXC) TFs and has been implicated as a critical regulator in embryogenesis, tumorigenesis, and metastasis¹⁵. In human hematopoietic cell lines, SOX4 contributes to cell proliferation and self-renewal *in vitro*¹⁶. In mouse EML cells ectopically expressing

HOXB4, SOX4 was previously identified by ChIP-chip analysis as a direct downstream target of HOXB4 that might mediate HSC self-renewal through the same pathway as HOXB4¹⁷. GSEA analysis suggests that, similar to STAT3, targets of SOX4 identified by ChIP-Seq include gene sets related to HSPC functions that are highly enriched among the upregulated genes in Lin-CD34+EML cells.

ATAC-Seq provided a global view of the potential TF binding landscape in Lin-CD34+EML cells. We found a higher proportion of open chromatin regions in intronic and intergenic regions, highlighting the probable locations of enhancers. After combining ATAC-Seq and RNA-Seq data, we observed that highly upregulated genes in Lin-CD34+EML had ATAC-Seq peaks in both their promoters and their gene bodies. However, in some cases, genes not expressed in Lin-CD34+EML cells also had open chromatin regions in their promoters or gene bodies. This highlights the possibility of regulatory mechanisms involving transcription factor and cofactor complexes. Motif analysis of the open chromatin regions in Lin-CD34+EML cells revealed many interesting TFs involved in hematopoiesis. The top 28 motifs included motifs for the binding of FOX TFs (Forkhead box proteins), a family of TFs that play important roles in cell growth, proliferation, and differentiation⁵⁷ including FOXJ2, FOXC1, FOXP1, and FOXD3. FOXJ2 regulates erythroid differentiation of K562 cells, FOXC1 is a critical regulator of HSPC niche formation, and FOXP1 signaling drives expansion of normal HSPCs.^{46, 58–60} There were also three motifs for the binding of SP1 TFs. The transcription factor SP1 is ubiquitously expressed and has recently been found to function at early stages of hematopoietic specification⁴⁸. Interestingly, the most common motif found in the open chromatin regions belonged to transcription factor STAT3. These significantly enriched TF motifs in the open chromatin regions of Lin-CD34+EML cells suggest the important TFs that could be involved in self-renewal, and provides valuable clues for future studies on the function of TFs in regulation of cell fate in HSPCs.

By combining proteomics, RNA-Seq, ChIP-Seq, ATAC-Seq, and protein-protein interaction datasets, we built an expanded transcriptional regulatory and interaction network for EML cells. Consistently, most of the genes in the network are in open chromatin regions. Additionally, a graph-based clustering algorithm was helpful in discovering local subnetworks with interesting enriched gene-sets such as TGF β signaling pathway, regulation of cell death, cell cycle and MAPK cascade, among others. TGF β -SMAD signaling regulates a wide spectrum of HSC biological processes, such as quiescence and self-renewal⁶¹. Our data showed that many TGF β -SMAD signaling components are the binding targets of SOX4, STAT3, TCF7, and RUNX1, and the expression of most of these targets is higher in Lin-CD34+EML than in Lin-CD34-EML cells, which is consistent with previous studies in the literature. Apoptosis plays an important role in removing aged and non-functional cells from the hematopoietic system and other tissues. Apoptosis is also involved in maintaining the appropriate balance of HSCs and mature blood cells and in maintaining the HSC pool⁶². Further, BCL-2 expression has been observed in cell populations that are generally long-lived, such as HSPCs⁶³. In the present study, genes that participate in apoptosis pathways and the cell cycle were found to be potential targets of STAT3, TCF7, and RUNX1 in EML cells.

The MAPK signaling was also an enriched gene-set in the subnetworks derived from the graph-based clustering. The MAPKs are serine/threonine kinases that transduce extracellular signals from cell-surface receptors to nuclear transcriptional events⁶⁴. The MAPK signaling pathways are involved in a wide range of cellular programs such as cell proliferation, differentiation, development, immune responses, and apoptosis⁶⁵. Three MAPK families, including ERK-, JNK/SAPK-, and p38 MAPK, have been well characterized⁶⁵. Our data showed that the expression of genes in the ERK MAPK pathway is higher in Lin-CD34+EML cells, such as *Grb2*, *Ras*, *Mapk1/3* and *Elk1*. The ability of the ERK MAPK signaling module to promote cellular proliferation and survival has been well established⁶⁵. A recent report showed that the stimulation of cord blood-derived Lin-CD34+CD133+ cells with stem cell factor (SCF) resulted in activation of the ERK-MAPK pathway and rapid induction of cell proliferation, which indicates a role for ERK signaling in the regulation of HSPC maintenance⁶⁶. ERK1/2 are required for the maintenance of hematopoietic stem cells and immature progenitors *in vivo*⁶⁷. Also, dominant negative mutants or antisense constructs of RAF-1 or ERK1 significantly inhibit cell proliferation, while stimulation of ERK1 activity results in enhanced cell proliferation^{65, 68}. The JNK/SAPK- MAPK pathway, which includes the c-Jun N-terminal kinases (JNKs) JNK1, JNK2, and JNK3, is essential for normal erythropoiesis and myelopoiesis in mammals. However, until now, the precise role of each of these different JNKs has not been well defined. Similarly, p38 MAPK signaling is also indispensable during erythropoiesis and myelopoiesis⁶⁴. Interestingly, in our study, the expression of *Mapk9* (*Jnk2*), a target of TCF7, was lower in Lin-CD34+EML cells, while the expression of *Mapk14* (*p38*), a common target of TCF7, RUNX1, and SOX4, was higher in Lin-CD34+EML. These data suggest that both JNK/SAPK- and p38 MAPK pathways may be important in HSPC development but might play different roles in this process. The p38 MAPK signaling pathway contributes to HSC exhaustion in response to reactive oxygen species (ROS)-mediated oxidative stress⁶⁹. Our data implied that p38 MAPK signaling might also play a role in normal hematopoiesis.

CONCLUSIONS

We used EML cells as a model system to study the fundamental mechanism underlying HSPC self-renewal and differentiation. For the first time, we integrated datasets from transcriptomic, proteomic, epigenomic, and protein-protein interaction levels to build a global regulatory network. We obtained binding targets for two additional transcription factors (STAT3 and SOX4) using ChIP-Seq assays. These TFs combined with previously studied transcription factors (TCF7 and RUNX1) constitute an essential part of a regulatory network with important functions controlling self-renewal of Lin-CD34+EML cells. ATAC-Seq revealed open chromatin regions on a genome-wide scale and potential TF regulatory sites within promoters of highly upregulated genes in Lin-CD34+EML subpopulations. Additionally, we also observed open chromatin regions within intronic and distal intergenic regions, highlighting the possibility of enhancer regions. Overall, this work provides valuable data resources for studying self-renewal in the early hematopoietic lineage, which we expect to ultimately be valuable for manipulating hematopoietic cells *in vivo* and to serve as a framework for investigating cell autonomous and balanced cell fate choice in mammals.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Ms. Mary Ann Cushman for editing the manuscript.

FUNDING

This work was supported by the National Institutes of Health [grant number R01 NS088353]; NHLBI [grant number K99/R00 HL093213]; The Staman Ogilvie Fund-Memorial Hermann Foundation; the UTHealth BRAIN Initiative; Clinical and Translational Science Awards [grant number TR000371]; and a grant from the University of Texas System Neuroscience and Neurotechnology Research Institute [grant number 362469]. Funding for open access charge: National Institutes of Health.

References

- Seita J, Weissman IL. Wiley Interdiscip Rev Syst Biol Med. 2010; 2:640–653. [PubMed: 20890962]
- Bryder D, Rossi DJ, Weissman IL. Am J Pathol. 2006; 169:338–346. [PubMed: 16877336]
- Wu JQ, Seay M, Schulz VP, Hariharan M, Tuck D, Lian J, Du J, Shi M, Ye Z, Gerstein M, Snyder MP, Weissman S. PLoS Genetics. 2012; 8:e1002565. [PubMed: 22412390]
- Challen GA, Boles N, Lin KK, Goodell MA. Cytometry A. 2009; 75:14–24. [PubMed: 19023891]
- Tsai S, Bartelmez S, Sitnicka E, Collins S. GENES & DEVELOPMENT. 1994; 8:11.
- Ye ZJ, Kluger Y, Weissman Z, Lian SM. Proceedings of the National Academy of Sciences. 2005; 102:18461–18466.
- Reuter JA, Spacek DV, Snyder MP. Mol Cell. 2015; 58:586–597. [PubMed: 26000844]
- Duran R, , Menon S, , Wu JQ. Transcriptomics and Gene Regulation Wu JQ, editorSpringer Publisher Dordrecht; 2016 136
- Lu R, Markowitz F, Unwin RD, Leek JT, Airolidi EM, MacArthur BD, Lachmann A, Rozov R, Ma'ayan A, Boyer LA, Troyanskaya OG, Whetton AD, Lemischka IR. Nature. 2009; 462:358–362. [PubMed: 19924215]
- Cenik C, Cenik ES, Byeon GW, Grubert F, Candille SI, Spacek D, Alsallakh B, Tilgner H, Araya CL, Tang H, Ricci E, Snyder MP. Genome Res. 2015; 25:1610–1621. [PubMed: 26297486]
- Hong SH, Yang SJ, Kim TM, Shim JS, Lee HS, Lee GY, Park BB, Nam SW, Ryoo ZY, Oh IH. Stem Cells. 2014; 32:1313–1322. [PubMed: 24446131]
- Mantel C, Messina-Graham S, Moh A, Cooper S, Hangoc G, Fu XY, Broxmeyer HE. Blood. 2012; 120:2589–2599. [PubMed: 22665934]
- Sawamiphak S, Kontarakis Z, Stainier DY. Dev Cell. 2014; 31:640–653. [PubMed: 25490269]
- Levy DE, Lee CK. J Clin Invest. 2002; 109:1143–1148. [PubMed: 11994402]
- Vervoort SJ, van Boxtel R, Coffey PJ. Oncogene. 2013; 32:3397–3409. [PubMed: 23246969]
- Sandoval S, Kraus C, Cho EC, Cho M, Bies J, Manara E, Accardi B, Landaw EM, Wolff L, Pigazzi M, Sakamoto KM. Blood. 2012; 120:155–165. [PubMed: 22627767]
- Lee HM, Zhang H, Schulz V, Tuck DP, Forget BG. Blood. 2010; 116:720–730. [PubMed: 20404135]
- Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Nature Methods. 2013; 10:1213–1218. [PubMed: 24097267]
- Wang T, Nandakumar V, Jiang XX, Jones L, Yang AG, Huang XF, Chen SY. Blood. 2013; 122:2812–2822. [PubMed: 24014243]
- Schutte J, Wang H, Antoniou S, Jarratt A, Wilson NK, Riepsaame J, Calero-Nieto FJ, Moignard V, Basilico S, Kinston SJ, Hannah RL, Chan MC, Nurnberg ST, Ouwehand WH, Bonzanni N, de Bruijn MF, Gottgens B. Elife. 2016; 5:e11469. [PubMed: 26901438]
- Eng JK, McCormack AL, Y JR. J Am Soc Mass Spectrom. 1994; 5:14.

22. Tusher VG, Tibshirani R, Chu G. Proceedings of the National Academy of Sciences. 2001; 98:5116–5121.
23. Roxas BAP, Li Q. BMC Bioinformatics. 2008; 9:187. [PubMed: 18402702]
24. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Nucleic Acids Res. 2015; 43:e47. [PubMed: 25605792]
25. Dong X, Chen K, Cuevas-Diaz Duran R, You Y, Sloan SA, Zhang Y, Zong S, Cao Q, Barres BA, Wu JQ. PLOS Genetics. 2015; 11:e1005669. [PubMed: 26683846]
26. Langmead B, Salzberg SL. Nat Methods. 2012; 9:357–359. [PubMed: 22388286]
27. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Mol Cell. 2010; 38:576–589. [PubMed: 20513432]
28. Kharchenko PV, Tolstorukov MY, Park PJ. Nat Biotechnol. 2008; 26:1351–1359. [PubMed: 19029915]
29. Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Genome Biol. 2008; 9:R137. [PubMed: 18798982]
30. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Raney BJ, Pohl A, Malladi VS, Li CH, Lee BT, Learned K, Kirkup V, Hsu F, Heitner S, Harte RA, Haeussler M, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Dreszer TR, Diekhans M, Cline MS, Clawson H, Barber GP, Haussler D, Kent WJ. Nucleic Acids Res. 2013; 41:D64–69. [PubMed: 23155063]
31. J. D. Buenrostro, B. Wu, H. Y. Chang and W. J. Greenleaf, 2015, DOI: 10.1002/0471142727.mb2129s109, 21.29.21–21.29.29.
32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, S. Genome Project Data Processing. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]
33. Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X, Shendure J. Genome Biol. 2010; 11
34. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Nature Biotechnology. 2011; 29:3.
35. Grant CE, Bailey TL, Noble WS. Bioinformatics. 2011; 27:1017–1018. [PubMed: 21330290]
36. Kheradpour P, Kellis M. Nucleic Acids Res. 2014; 42:2976–2987. [PubMed: 24335146]
37. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Proc Natl Acad Sci. 2005; 102:15545–15550. [PubMed: 16199517]
38. Zhang H, Meltzer P, Davis S. BMC Bioinformatics. 2013; 14:244. [PubMed: 23937229]
39. Doerks T, Copley RR, Schultz J, Ponting CP, Bork P. Genome Res. 2002; 12:47–56. [PubMed: 11779830]
40. Maere S, Heymans K, Kuiper M. Bioinformatics. 2005; 21:3448–3449. [PubMed: 15972284]
41. Pons P, Latapy M. Journal of Graph Algorithms and Applications. 2006; 10:28.
42. Zong S, Deng S, Chen K, Wu JQ. J Vis Exp. 2014; :e52104.doi: 10.3791/52104 [PubMed: 25407807]
43. Deneault E, Cellot S, Faubert A, Laverdure JP, Frechette M, Chagraoui J, Mayotte N, Sauvageau M, Ting SB, Sauvageau G. Cell. 2009; 137:369–379. [PubMed: 19379700]
44. Lee SU, Maeda M, Ishikawa Y, Li SM, Wilson A, Jubb AM, Sakurai N, Weng L, Fiorini E, Radtke F, Yan M, Macdonald HR, Chen CC, Maeda T. Blood. 2013; 121:918–929. [PubMed: 23134786]
45. Ackermann AM, Wang Z, Schug J, Naji A, Kaestner KH. Mol Metab. 2016; 5:233–244. [PubMed: 26977395]
46. Yang GH, Wang F, Yu J, Wang XS, Yuan JY, Zhang JW. J Cell Biochem. 2009; 107:548–556. [PubMed: 19350553]
47. Naudin C, Hattabi A, Michelet F, Miri-Nezhad A, Benyoucef A, Pflumio F, Guillonnet F, Fichelson S, Vigon I, Dusanter-Fourt I, Lauret E. Blood. 2017; 129:2493–2506. [PubMed: 28232582]
48. Gilmour J, Assi SA, Jaegle U, Kulu D, van de Werken H, Clarke D, Westhead DR, Philipsen S, Bonifer C. Development. 2014; 141:2391–2401. [PubMed: 24850855]

49. Ma AC, Fung TK, Lin RH, Chung MI, Yang D, Ekker SC, Leung AY. *Blood*. 2011; 118:5448–5457. [PubMed: 21937698]
50. Wilson NK, Kent DG, Buettner F, Shehata M, Macaulay IC, Calero-Nieto FJ, Sanchez Castillo M, Oedekoven CA, Diamanti E, Schulte R, Ponting CP, Voet T, Caldas C, Stingl J, Green AR, Theis FJ, Gottgens B. *Cell Stem Cell*. 2015; 16:712–724. [PubMed: 26004780]
51. Li L, Byrne SM, Rainville N, Su S, Jachimowicz E, Aucher A, Davis DM, Ashton-Rickardt PG, Wojchowski DM. *Stem Cells*. 2014; 32:2550–2556. [PubMed: 24964278]
52. Matsuoka S, Ebihara Y, Xu M, Ishii T, Sugiyama D, Yoshino H, Ueda T, Manabe A, Tanaka R, Ikeda Y, Nakahata T, Tsuji K. *Blood*. 2001; 97:419–425. [PubMed: 11154218]
53. Steigerwald K, Behbehani GK, Combs KA, Barton MC, Groden J. *Mol Cancer Res*. 2005; 3:78–89. [PubMed: 15755874]
54. Duran RC, Yan H, Zheng Y, Huang X, Grill R, Kim DH, Cao Q, Wu JQ. *Sci Rep*. 2017; 7:41008. [PubMed: 28106101]
55. van Keimpema M, Gruneberg LJ, Mokry M, van Boxtel R, Koster J, Coffey PJ, Pals ST, Spaargaren M. *Blood*. 2014; 124:3431–3440. [PubMed: 25267198]
56. Dhanasekaran DN, Reddy EP. *Oncogene*. 2008; 27:6245–6251. [PubMed: 18931691]
57. Benayoun BA, Caburet S, Veitia RA. *Trends Genet*. 2011; 27:224–232. [PubMed: 21507500]
58. Omatsu Y, Seike M, Sugiyama T, Kume T, Nagasawa T. *Nature*. 2014; 508:536–540. [PubMed: 24590069]
59. Naudin Cécile, Hattabi Aurore, Michelet Fabio, Miri-Nezhad Ayda, Benyoucef Aissa, Pflumio Françoise, Guillonnet François, Fichelson Serge, Vigon Isabelle, Dusanter-Fourt Isabelle, Lauret E. *Blood*. 2017; 129:14.
60. Holmfeldt P, Ganuza M, Marathe H, He B, Hall T, Kang G, Moen J, Pardieck J, Saulsberry AC, Cico A, Gaut L, McGoldrick D, Finkelstein D, Tan K, McKinney-Freeman S. *J Exp Med*. 2016; 213:433–449. [PubMed: 26880577]
61. Blank U, Karlsson S. *Blood*. 2015; 125:9. [PubMed: 25554745]
62. Orelia C, Dzierzak E. *Leuk Lymphoma*. 2007; 48:16–24. [PubMed: 17325844]
63. HOCKENBERY DM, ZUTTER M, HICKEY W, NAHM M, KORSMEYER SJ. *Proc Natl Acad Sci*. 1991; 88:5.
64. Geest CR, Coffey PJ. *J Leukoc Biol*. 2009; 86:237–250. [PubMed: 19498045]
65. Chang L, Karin M. *Nature*. 2001; 410:4.
66. Oostendorp RA, Gilfillan S, Parmar A, Schiemann M, Marz S, Niemeyer M, Schill S, Hammerschmid E, Jacobs VR, Peschel C, Gotze KS. *Stem Cells*. 2008; 26:2164–2172. [PubMed: 18499891]
67. Chan G, Gu S, Neel BG. *Blood*. 2013; 121:3594–3598. [PubMed: 23444405]
68. PAGES GILLES, LENORMAND PHILIPPE, L'ALLEMAIN GILLES, CHAMBARD JEAN-CLAUDE, MELOCHE SYLVAIN, POUYSSAGUR J. *Proc Natl Acad Sci*. 1993; 90:5.
69. Ito K, Hirao A, Arai F, Takubo K, Matsuoka S, Miyamoto K, Ohmura M, Naka K, Hosokawa K, Ikeda Y, Suda T. *Nat Med*. 2006; 12:446–451. [PubMed: 16565722]

Insight Box

Understanding the regulatory mechanisms controlling hematopoietic stem/progenitor cells (HSPCs) self-renewal and differentiation is important for manipulating HSPCs for therapeutic purposes. We report binding maps for two transcription factors using chromatin immunoprecipitation (ChIP)-Sequencing, to identify potential mechanisms regulating self-renewal and differentiation of a hematopoietic progenitor cell model system- the EML cells. Assay for Transposase Accessible Chromatin (ATAC)-Sequencing and Mass spectrometry were applied to globally identify open chromatin regions and quantify protein expression levels. By integrating the protein-protein interaction database and RNA-Sequencing data, we built an expanded regulatory and interaction network. MAPK and TGF- β /SMAD pathways were found to be potential regulatory pathways. We integrate regulatory information at multiple levels to better characterize the mechanisms underlying HSPCs self-renewal.

Figure 1.

Comparison of relative protein and mRNA abundances. Relative protein abundances were measured as $\log_2$ (Lin-CD34+EML/Lin-CD34-EML). a) SAM plot of observed and expected scores for each protein at $\Delta = 0.4$ and $FDR < 2.2\%$. Green dots represent proteins with significantly lower expression in Lin-CD34+EML cells compared with Lin-CD34-EML cells, while red dots represent proteins with significantly higher expression in Lin-CD34+EML cells. b) Distribution of q -values calculated by SAM for each protein and the consensus relative abundances of each protein. The numbers of proteins with at least a particular q -value and protein relative expression are labelled in the plot. c) Scatter plot depicting the relationship between SAM scores and relative protein expression on a $\log_2$ scale. Blue and red dots represent downregulated and upregulated proteins respectively when comparing Lin-CD34+EML cells with Lin-CD34-EML cells. SAM scores were obtained at

= 0.4 and FDR <2.2%. Only proteins with RPKM >1 in at least one of the samples were considered for further analysis. d) Scatter plot comparing the log2-transformed fold-changes in both relative protein abundances and gene expression. Labels indicate the colors and categories assigned to each protein according to whether changes in relative protein abundances and gene expression were consistent or opposite. Cyan: 'downregulated opposite', turquoise: 'downregulated, mRNA no change', dark blue: 'downregulated consistent', light gray: 'protein-no-change, upregulated mRNA', dark gray: 'protein-no-change, mRNA no-change', green: 'protein-no-change, downregulated mRNA', red: 'upregulated consistent', orange: 'upregulated mRNA-no-change', and magenta: 'upregulated opposite'.

Figure 2.

CD34⁺ cells have common binding target genes of TCF7, STAT3, and SOX4 TFs within open chromatin regions. a) Highly enriched target gene sets for i) STAT3, ii) SOX4, and iii) 93 common TF binding sites for TCF7, RUNX1, STAT3, and SOX4. Yellow lines indicate the $-\log_{10}$ (q-value) in the upper scale. Black bar plots (lower scale) refer to the number of genes found belonging to each gene set. b) TCF7, RUNX1, STAT3, and SOX4 bind common gene targets. TCF7 and RUNX1 peaks were called using PeakSeq, whereas targets of STAT3 and SOX4 were identified using MACS1.4. Genome browser views displaying ATAC-Seq

open chromatin regions (broadPeaks), punctate peaks (narrowPeaks), and common transcription factor binding sites within a window 5 kb upstream of the TSS of a gene and within its gene body (introns and exons). Upper: *Zbtb7a* gene locus; Lower: *Tgif1* gene locus. c) ATAC-Seq fragment size distribution provides genome-wide information on chromatin compaction. Fragments smaller than 100 bp represent sequence reads in open chromatin. The peak at 200 bp represents reads that span one nucleosome. Larger peaks indicate regions of more compact chromatin, such as dinucleosomes (400 bps) and trinucleosomes (600 bp). d) Histogram of distances from the ATAC-Seq narrowPeaks to the nearest transcriptional start site (TSS). e) Pie chart depicting the distributions of the genome annotations of all peak regions identified using ATAC-Seq.

Figure 3. Integration of ChIP-Seq, RNA-Seq, and ATAC-Seq datasets. a, b) Correlation of transcription factor binding targets with RNA-Seq differential gene expression data using GSEA. Genes were ranked according to their fold-change in expression. The most upregulated genes in Lin-CD34+EML cells were placed at the top of the list, while the most upregulated genes in Lin-CD34-EML cells were placed at the bottom. GSEA was performed to determine if transcription factor binding targets were overrepresented among the genes upregulated in Lin-CD34+EML cells. The lower scale shows the most upregulated (red) and

downregulated (blue) transcription factor binding targets in Lin-CD34+EML cells compared with Lin-CD34-EML cells. Black bars represent the position of targets in the ranked gene list. Enrichment scores (ES) indicate the degree of enrichment of the transcription factor binding targets. a) Sox4 binding targets have a high enrichment score (ES = 0.45) among genes upregulated in Lin-CD34+EML cells (nominal p -value = 0, FDR <0.011). b) Statistically significant enrichment (ES = 0.51) of Stat3 binding peaks among upregulated genes in Lin-CD34+EML cells (nominal p -value = 0, FDR = 0). c) Circos plot representing transcription factor binding sites and gene expression. The outermost layer is a cytoband representing the chromosome ideograms of the mm10 genome. The second track represents the gene expression heatmap of log2-transformed RPKM values of expressed genes in Lin-CD34+EML cells. Tracks 3–6 display the ChIP-Seq binding peaks in Lin-CD34+EML cells for TCF7, RUNX1, STAT3, and SOX4 respectively. Bar heights indicate the fold-enrichment of each peak. Labels indicate differentially expressed genes as analyzed using Limma (p -value <0.001, fold-change >1.5, and RPKM >1 in at least one sample). Right upper panel zooms into *Eya2* loci as an example of an upregulated gene which is also a binding target of all four transcription factors.

Figure 4. Examples of biologically interesting local subnetworks derived from graph-based clustering. The Walktrap algorithm was used to cluster genes from the integrated regulatory network (protein-protein and protein-DNA interactions) and ChIP-qPCR validation. Node fill colors represent gene expression fold-changes (red = genes with higher expression in Lin-CD34+EML cells compared to Lin-CD34-EML cells, blue = genes with higher expression in Lin-CD34-EML cells compared to Lin-CD34+EML cells). Node border colors depict chromatin conformation (green = open chromatin, black = closed chromatin). Node shapes highlight the relation between protein relative abundance and mRNA expression (ellipse = protein and mRNA level changes are consistent, diamond = protein and mRNA level

changes are opposite, octagon = protein levels don't change, mRNA levels are up or down-regulated, rounded rectangle = protein was not detected). For visualization purposes, cutoffs of protein consensus ratio of 0.01 and a gene expression fold-change of 1.5 were applied. Edge colors indicate the nature of the interaction (green = known protein-protein interaction, pink = ChIP-Seq-derived interactions). Arrows are indicative of direction of interaction. a) A Local subnetwork includes TGF- β /SMAD signaling pathway members. Most TGF β s, SMADs, and their interacting proteins were more highly expressed in Lin-CD34+EML cells compared to Lin-CD34-EML cells. b) Bcl-2 pathway members enriched in another local subnetwork. The expression of genes involved in apoptosis regulation such as *Myc*, *Bcl-2l1l*, and *Bcl-2l1*, and that of genes involved in cell cycle processes, such as *Ccnd1*, *Cdk6*, and *Cdkn1a* was relatively higher in Lin-CD34+EML cells. c) ChIP-qPCR was performed to validate the binding of TFs and a number of genes in biologically important sub-networks and pathways. Enrichment over genomic input DNA was calculated. Experiments were performed in triplicate and error bars indicate standard error. *t*-test analysis * $p < 0.05$ and ** $p < 0.01$.

Figure 5.

Gene Ontology (GO) network of biological process categories analyzed using BiNGO in Cytoscape. Data for enrichment was obtained from an integrated protein-DNA and protein-protein interaction dataset. The integrated dataset consisted of TCF7, RUNX1, SOX4, and STAT3 ChIP-Seq targets combined with known protein-protein interactions from NCBI. Genes were filtered for those with RPKM >1 and fold-change >1.5 (up or downregulated in Lin-CD34+EML cells). We then selected the top 100 nodes (genes) with the most interactions. Enriched GO terms in Lin-CD34+EML cells include ‘metabolic processes’, ‘signaling’, ‘biological regulation’, ‘response to stimulus’, and ‘cell development and differentiation’. *P*-value of the enriched terms is indicated in the color scale shown.

Figure 6.

Enriched target genes of the TFs TCF7, RUNX1, STAT3, and SOX4 in the MAPK signaling pathway. Data were analyzed using the KEGG Pathway feature of the Database for Annotation, Visualization, and Integrated Discovery (DAVID) (accessed at <https://david.ncifcrf.gov/>). Red asterisks indicate genes that are targets of at least one of four TFs including TCF7, RUNX1, STAT3, and SOX4.