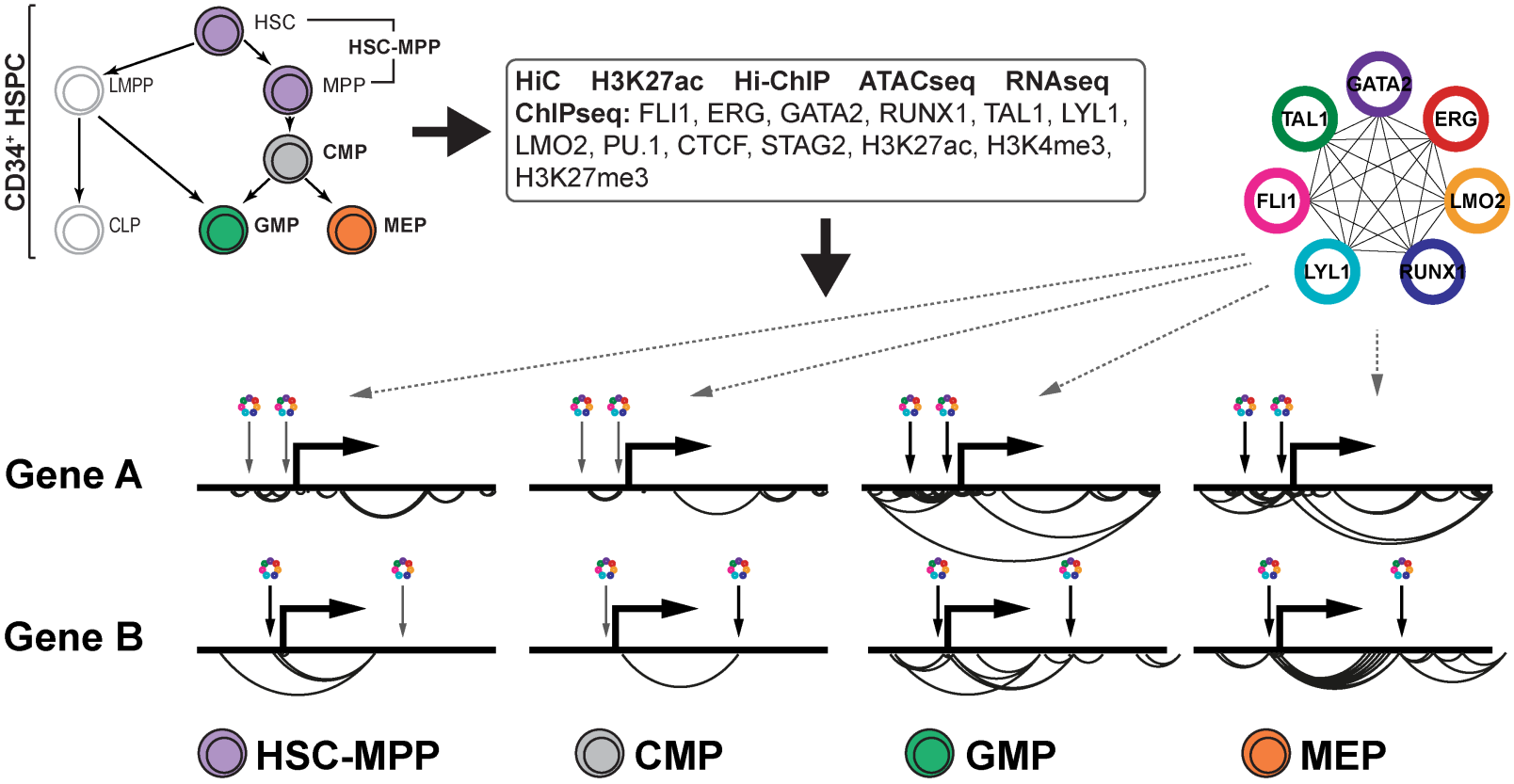

CD34+ HSPC
HSC
HSC-MPP
LMPP
MPP
CMP
CLP
GMP
MEP
HiC H3K27ac Hi-ChIP ATACseq RNAseq
ChIPseq: FLI1, ERG, GATA2, RUNX1, TAL1, LYL1, LMO2, PU.1, CTCF, STAG2, H3K27ac, H3K4me3, H3K27me3
GATA2
TAL1
ERG
FLI1
LMO2
LYL1
RUNX1
Gene A
Gene B
HSC-MPP
CMP
GMP
MEP

Genome-wide Transcription Factor binding maps reveal cell-specific changes in the regulatory architecture of human HSPC.

TF binding and 3D chromatin maps in HSPC subsets

Shruthi Subramanian[1,*], Julie A.I. Thoms[2,*,+], Yizhou Huang[3], Paola Cornejo[4], Forrest C. Koch[5], Sebastien Jacquelin[6], Sylvie Shen[7], Emma Song[7], Swapna Joshi[1], Chris Brownlee[8], Petter S. Woll[9], Diego Chacon Fajardo[3], Dominik Beck[3], David J. Curtis[10], Kenneth Yehson[11], Vicki Antonenas[11], Tracey O' Brien[12], Annette Trickett[7], Jason A. Powell[13,14], Ian D. Lewis[13], Stuart M. Pitson[13], Maher K. Gandhi[15], Steven W. Lane[16], Fatemeh Vafaee[5,17], Emily S. Wong[4,5], Berthold Göttgens[18], Hamid Alinejad Rokny[19], Jason W.H Wong[20,+] and John E. Pimanda[1,2,21,22+]

[1]*School of Clinical Medicine, University of New South Wales, Sydney, Australia*
[2]*School of Biomedical Sciences, University of New South Wales, Sydney, Australia*
[3]*Centre for Health Technologies and the School of Biomedical Engineering, University of Technology Sydney, Sydney, Australia*
[4]*Victor Chang Cardiac Research Institute, Sydney, Australia*
[5]*School of Biotechnology and Biomolecular Sciences, Faculty of Science, University of New South Wales, Sydney, Australia*
[6]*Macrophage Biology Laboratory, Mater Research, Brisbane, Australia*
[7]*Bone Marrow Transplant Lab, NSW Health Pathology, Prince of Wales Hospital, Randwick, NSW, Australia*
[8]*Mark Wainwright Analytical Centre, University of New South Wales, Sydney, Australia*
[9]*Center for Hematology and Regenerative Medicine, Department of Medicine (MedH), Karolinska Institutet, Huddinge, Sweden*
[10]*Australian Centre for Blood Diseases, Monash University, Melbourne, VIC, Australia*
[11]*Blood Transplant and Cell Therapies Laboratory, NSW Health Pathology, Westmead, NSW, Australia*
[12]*Sydney Children's Hospital, Sydney, Australia*
[13]*Centre for Cancer Biology, SA Pathology, University of South Australia, Adelaide, Australia*
[14]*Adelaide Medical School, University of Adelaide, Adelaide, Australia*
[15]*Blood Cancer Research Group, Mater Research, University of Queensland, Brisbane, QLD, Australia*
[16]*Cancer Program, QIMR Berghofer Medical Research, Brisbane, Australia*
[17]*UNSW Data Science Hub, University of New South Wales, Sydney, Australia*
[18]*Wellcome - MRC Cambridge Stem Cell Institute, Cambridge, United Kingdom*
[19]*BioMedical Machine Learning Lab, Graduate School of Biomedical Engineering, University of New South Wales, Sydney, Australia*
[20]*School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong Special Administrative Region, China*
[21]*Haematology Department, Prince of Wales Hospital, Sydney, Australia*
[22]*Lead contact*
[*]*Equal contributions*

+ Corresponding authors:
John Pimanda; jpimanda@unsw.edu.au, Level 2, Lowy Cancer Research Centre, UNSW Sydney, NSW 2052, Australia, P: +612 93828866, F:+612 93851510
Jason Wong; jwhwong@hku.hk, L1-51, Laboratory Block, 21 Sassoon Road, University of Hong Kong, Hong Kong SAR, China, P; +852 3917 9187, F: +852 2817 0857
Julie Thoms; j.thoms@unsw.edu.au, , Level 2, Lowy Cancer Research Centre, UNSW Sydney, NSW 2052, Australia, P: +612 90651412, F:+612 93851510

Word count (summary): 249, Word count (text): 4397, Figure count: 7 main, 12 supplemental, Table count: 10 supplemental, Reference count: 81 (main)/34 (supplemental)
Manuscript category: Hematopoiesis and Stem Cells
Keywords: gene regulation, stem cell differentiation, transcriptional networks, haematopoiesis, transcription factors

KEY POINTS

* Heptad TF occupancy is highly dynamic across HSPC subsets and associated with cell-type specific gene expression.

* Enhancers with cell-type-specific Heptad occupancy share a common grammar with respect to TF binding motifs.

ABSTRACT

Hematopoietic stem and progenitor cells (HSPCs) rely on a complex interplay of transcription factors (TFs) to regulate differentiation into mature blood cells. A heptad of TFs – FLI1, ERG, GATA2, RUNX1, TAL1, LYL1, LMO2 – bind regulatory elements in bulk $CD34^+$ HSPCs. However, whether specific heptad-TF combinations have distinct roles in regulating hematopoietic differentiation remained unknown. We mapped genome-wide chromatin contacts (HiC, H3K27ac HiChIP), chromatin modifications (H3K4me3, H3K27ac, H3K27me3) and 10 TF binding profiles (the Heptad, PU.1, CTCF, and STAG2) in HSPC subsets (HSC-MPP, CMP, GMP, MEP) and found that TF occupancy and enhancer-promoter interactions varied significantly across cell types and were associated with cell-type-specific gene expression. Distinct regulatory elements were enriched with specific heptad-TF combinations, including stem-cell-specific elements with ERG, and myeloid- and erythroid-specific elements with combinations of FLI1, RUNX1, GATA2, TAL1, LYL1, and LMO2. Furthermore, heptad-occupied regions in HSPCs were subsequently bound by lineage-defining TFs such as PU.1 and GATA1, suggesting that heptad factors may prime regulatory elements for use in mature cell types. We also found that enhancers with cell-type-specific heptad occupancy shared a common grammar with respect to TF binding motifs, suggesting that combinatorial binding of specific TF complexes was at least partially regulated by features encoded in specific DNA sequence motifs. Taken together, this study provides a comprehensive characterisation of the gene regulatory landscape in rare subpopulations of human HSPCs. The accompanying datasets should serve as a valuable resource for understanding adult hematopoiesis and a framework for analysing aberrant regulatory networks in leukemic cells.

INTRODUCTION

Hematopoietic stem cells (HSCs) maintain production of circulating blood cells via their capacity to either self-renew or differentiate to mature cell types [1]. The most primitive HSCs have multilineage potential but give rise to progenitor cells with increasing lineage restriction. Although single cell analyses have suggested that differentiation occurs over a continuum rather than in discrete leaps [2-4], relatively pure populations which correspond to intermediate progenitor stages can be prospectively isolated based on cell surface markers [4].

Changes in cell identity across differentiation trajectories are directly related to altered transcriptional programs [5] which are controlled by lineage-specific gene regulatory networks (GRNs) [6]. At the simplest level, GRNs are comprised of genes, their associated regulatory elements (promoters and *cis*-regulatory elements (CREs) such as enhancers), and transcriptional regulators, including transcription factors (TF), which bind these elements [7,8]. Accessibility of regulatory elements is controlled by various chromatin modifications, and the DNA sequence of such elements at least partially determines which TFs can bind [9,10]. Further control is imposed by chromatin organization into topologically associated domains (TADs) [11]. Interactions between promoters and their CREs, mediated by chromatin loops and complexes of transcriptional regulators, modulate GRNs and therefore cell identity [12].

We previously showed that seven TFs (heptad: FLI1, ERG, GATA2, RUNX1, TAL1, LYL1, and LMO2), all known regulators of healthy and leukemic haematopoiesis [13-20], bind combinatorially in bulk $CD34^+$ hematopoietic stem and progenitor cells (HSPCs) [21] and leukemias [22-24]. In healthy HSPCs, heptad combinatorial binding occurs at regulatory regions associated with genes involved in stem cell maintenance and function, and also at heptad CREs such that heptad genes form a highly interconnected regulatory circuit [21,25]. However,

study of GRNs in HSPCs is hindered by heterogeneity within the $CD34^+$ population and lack of experimental evidence linking promoters to distal regulatory elements.

To address these issues and further our understanding of heptad centred GRNs in blood development, we sorted $CD34^+$ HSPCs into HSC/multipotent progenitor (MPP, collectively HSC-MPP), common myeloid progenitor (CMP), granulocyte macrophage progenitor (GMP), and megakaryocyte erythrocyte progenitor (MEP) cell types. We then used chromatin immunoprecipitation followed by sequencing (ChIP-seq) targeting 10 TFs [heptad, PU.1, CTCF, STAG2] and 3 histone modifications [H3K27ac, H3K4me3, H3K27me3], Hi-C, and H3K27ac-HiChIP in each of these cell types to chart the regulatory landscape of human HSPC differentiation. Combinatorial binding of heptad TFs was observed in all sorted populations, although specific patterns of chromatin occupancy differed between cell types. Heptad promoter looping to putative enhancers was variable across cell types, and in many cases combinatorial binding was observed at CREs in immature cells prior to formation of loops in more mature progenitors. Genome-wide occupancy of heptad TFs was also variable across cell types, with distinct sets of CREs enriched for heptad binding in MEP compared to GMP. This variation was at least partially due to sequence motifs in the CREs, with motif composition sufficient to predict cell type with high sensitivity and specificity.

METHODS

Mobilized peripheral blood was collected with patient consent in accordance with the Declaration of Helsinki and used with institutional ethics approval ref:08/190 from the South-Eastern Sydney Local Health District.

Cell sorting, preparation of nuclei, and ChIP were performed essentially as described [22,26,4] (detailed in Supplemental Methods). HiC and HiChIP libraries were generated using the Arima Genomics HiC+ kit (Arima cat#A101020).

Bioinformatic analysis used standard pipelines for ChIP [21] and HiC/HiChIP data [27]. Machine learning models were trained using XGBoost [28]. A UCSC browser session is provided at http://genome.ucsc.edu/s/PimandaLab/Heptad_Regulome. We also provide a tool for data exploration (http://unsw-data-analytics.shinyapps.io/CD34_Heptad_Regulome).

Details of standard experimental and analysis techniques are provided in the Supplemental Methods. Sequencing data have been uploaded to GEO with accession #GSE231486. Contact jpimanda@unsw.edu.au for any original data not included herewith.

RESULTS

Genome-wide heptad factor binding in HSPCs.

Primary mobilised human $CD34^+$ HSPCs were sorted (Figure 1A, Figure S1A), purity checked with colony assays (Figure S1B), and fixed for downstream assays (Figure 1B, Table S1). Cell populations were chosen to span the trajectory from early multipotent stem cells (HSC-MPP) through to progenitors committed to the myeloid (GMP) or erythroid (MEP) lineages.

High quality ChIP data were obtained for all heptad factors (Figure 1C, Figure 1D, Table S2, http://genome.ucsc.edu/s/PimandaLab/Heptad_Regulome). Total peak numbers were highly variable between heptad factors and across cell types (Figure 1D). We observed cell-type-specific trends consistent with known expression patterns and biology. For example, ERG peaks were most abundant in HSC-MPPs, consistent with its role in maintaining the stem cell state [29,30], while TAL1 peaks were most abundant in MEP, consistent with its role in erythroid development (Figure 1D). Distributions of TFs across genome features were generally conserved across cell types, but differed between heptad factors. For example, FLI1, ERG, and RUNX1 peaks were located at both promoter and non-promoter regions (Figure 1D), while GATA2, TAL1, LYL1, and LMO2 peaks were predominantly located at non-promoter regions. TAL1 peak distribution in MEP was unique, with many peaks found at intergenic regions. Motif enrichment analysis showed ETS (GGAA) and E-Box (CANNTG) motifs in TF-occupied regions from all factors. FLI1, ERG, and RUNX1 peaks were highly enriched for the ETS motif, while GATA2, TAL1, LYL1, and LMO2 peaks showed additional enrichment for the GATA motif (GATA), particularly for GATA2 and TAL1 in MEP. Consistent with observations in bulks HSPCs[21] enrichment of RUNX1 motifs was minimal. Overall, we observed conserved patterns of heptad binding, but also distinct

differences between factors and across cell types, consistent with dynamic remodeling of the heptad network across the HSPC differentiation trajectory.

Combinatorial binding of heptad TFs is cell-type-specific.

We previously described combinatorial binding of heptad factors in bulk human HSPCs [21], and now extend these observations to specific cell populations. We quantified ChIP peaks containing all possible combinations of two or more heptad factors (Figure 2Ai, ii), and evaluated the probability these occurred by random chance in each cell type (Figure 2Aiii). As in bulk HSPCs [21], combinatorial binding of all seven factors was the most significant event in any of the four cell types (z-scores for all seven factors; HSC-MPP: 8199.94, CMP:6314.92, GMP:3543.93, MEP:1877.48). Pairwise combinations had low significance scores, with exceptions such as ERG/RUNX in HSC-MPPs. Higher order complexes generally had higher significance scores, with 5/7 possible six-factor complexes showing highly significant binding in at least one cell type. Specific combinations broadly matching known TF function were also observed. For example, five- or six-factor combinations lacking GATA2, TAL1, or both, had low significance scores in MEPs (Figure 2A, stars). We also asked whether PU.1 showed combinatorial binding with heptad factors. As previously observed [25], co-binding of single heptad factors with PU.1 was minimal (Figure S2A), while addition of PU.1 to 6/7 TF combinations modestly accentuated existing patterns (Figure S2B). Overall, we find that combinatorial binding of heptad TFs is a general feature of stem and progenitor cells, with some combinations restricted to specific cell types.

Given the cell-type-specificity of some TF combinations we predicted that dynamic formation of TF complexes might play a role in priming CREs for subsequent activation. We explored heptad factor binding at promoter regions of two lineage-specific genes: the

erythroid regulator *GATA1* and the monocyte gene *MPO,* neither of which showed heptad binding in HSC-MPPs. At the *GATA1* promoter (Figure 2B; yellow region), GATA2, RUNX1, TAL1, LYL1, and LMO2 binding was observed in CMP. As cells transitioned to MEP, binding peaks became more prominent and now included FLI1 and ERG. However, there was essentially no heptad binding at the *GATA1* promoter in GMP. At the *MPO* promoter (Figure 2C; yellow region), FLI1, ERG, TAL1, LYL1, and LMO2 binding was observed in CMP. Binding peaks became more prominent as cells transitioned to GMP, with RUNX1 also bound, but no TF binding was observed at the *MPO* promoter in MEP. Taken together our data suggests that distinct patterns of heptad TF binding may prime the genomic landscape of human blood stem cells towards either an erythroid or myeloid fate.

Heptad regulatory circuits are remodelled during myeloid progenitor development.

Genes encoding heptad TFs form a densely interconnected regulatory circuit in bulk HSPCs [21], and chromatin accessibility at heptad gene CREs is sufficient to predict blood cell identity [23]. To better understand genome organization and connectivity at heptad loci during HSPC development we performed HiC and H3K27ac HiChIP experiments on HSC-MPP, CMP, GMP, MEP. While the majority of genome compartments remained stable across all cell types, some compartments underwent B to A, or A to B switching upon transition from CMP to lineage-committed progenitor (Figure S3Ai). Notably, some regions underwent compartment switching only in GMP, with corresponding changes in H3K27ac signal (Figure S3Aii, iii). TAD boundaries were highly conserved between cell types (Figure S3B), and HiC contact matrices around heptad gene loci were also highly similar (Figure S4). Global CTCF and STAG2 binding was also conserved across cell types (Figure S5). Overall, consistent with previous reports [31] we observed minimal variation in high level genome organization between HSPC subpopulations.

H3K27ac HiChIP experiments generated thousands of significant interactions with FDR ≤ 0.01 (HSC-MPP: 26210, CMP: 8170, GMP: 43448, MEP: 32773; Table S3). We focused on loops where at least one interacting region was annotated as a promoter (P). As this experiment enriched fragments with the H3K27ac active enhancer mark, we consider looped CREs as putative enhancers (E). We filtered promoter-enhancer (P – E) loops based on presence of an ATAC peak at the distal enhancer and integrated these with ChIP-seq data to create regulatory network maps at each heptad gene locus for each cell type.

Promoter-enhancer loops corresponding to the heptad spanned wide genomic regions; around 500kb for *FLI1*, *GATA2*, and *TAL1* and 1-2Mb for *ERG*, *RUNX1*, *LYL1*, and *LMO2* (Figure 3Ai, ii, Figure S6). Overall, we detected multiple putative enhancers for the heptad genes *ERG* (−610/-410/-230/+85/+88/+191/+1200), *FLI1* (+27/+32/+64), *GATA2* (−123/-92/+4), *RUNX1* (−880/+22/+100/+110/+141/+161), *TAL1* (−101/−82/−25/+0.5/+14/+45), *LYL1* (−744/−50/+165/+310), and *LMO2* (−570/−100/−67/−61/−51/−23/−22/−15/−12) (Figure 3Ai, ii, Figure S6, Table S4. Some regions have been described in humans and/or in mice [32,33,17,21,34-45] while others were novel (Table S4). Two known heptad enhancers were not directly looped to their corresponding promoter (*FLI1*-15, *GATA2*-117, Figure S6Ai, S6Bi) [21,44], although we did observe looping between *GATA2*-117 and other putative enhancers in MEPs (Figure S6Bi). Furthermore, looping at heptad genes generally increased in complexity in GMP and MEP compared to HSC-MPP, with the sparsest looping observed in CMP (Figure 3Ai, Figure S6). This additional complexity often included extensive looping to CREs not directly connected to promoters, suggesting that heptad gene expression is fine-tuned by highly interconnected cell-specific enhancer communities (e.g., TAL1 locus in MEP (Figure S6Di), LYL1 locus in GMP and MEP (Figure S6Ei)).

Most directly looped elements showed heptad factor binding in at least one cell type (Figure 3Aiii, Figure S6). Some core CREs showed extensive heptad binding in all cell types, including known and functionally validated enhancers such as *ERG*+85, *GATA2*+4, *RUNX1*+22, and *TAL* +45 (Figure 3Aiii, Figure S6Biii, Ciii, Diii, Table S4), plus novel CREs such as *ERG*-410 and *LMO2*-570 that can now be linked to heptad genes (Figure 3Aiii, Figure S6Fiii). Integration of HiChIP and ChIP data at heptad gene loci showed diverse patterns of looping and TF binding across the four cell types (Figure 3Bi, ii). At the *ERG* locus, the *ERG*+85 enhancer was linked to the proximal promoter in HSC-MPP, GMP, and MEP, while other *ERG* elements showed promoter looping in GMP and/or MEP. (Figure 3Bii). Furthermore, heptad TF binding often occurred in HSC-MPPs with subsequent promoter looping of that element in more differentiated cell types (e.g./ *GATA2*−123, *TAL1*+45, and *LMO2*-570, Figure 3Bii). However, these epigenetic changes did not directly correspond to the steady state transcriptional output of heptad genes, which was relatively stable across the four cell types (Figure 3Biii).

Analysis of bulk HSPCs found that heptad genes regulate themselves and each other via a densely interconnected auto-regulatory circuit [21]. We used our expanded set of heptad CREs to construct network connectivity maps for each cell (Figure S7). These maps show that although the transcriptional output of the heptad genes was relatively stable across HSC-MPP, CMP, GMP, MEP, there were cell-type specific differences in *cis*- and *trans*-regulatory mechanisms by which stable expression was maintained. Overall, our data set allowed us to observe extensive remodeling of the regulatory connections within and between individual heptad genes during hematopoiesis and increases our understanding of the complex network regulating heptad genes during hematopoiesis.

Heptad TFs and regulation of lineage-specific gene expression.

We next asked whether heptad factor chromatin occupancy was associated with cell-type-specific transcriptional output. We identified genomic regions with cell-type-enriched binding of at least 2 heptad TFs (Differentially Enriched for Heptad [DEH]) and linked these to genes using HiChIP data (Genes associated with DEH [DEHG]) (Figure 4A, Table S5A). Approximately half of the regions were promoter-like (up to 10 kb upstream of a TSS). To characterize candidate regulatory elements (REs) and their associated genes we conducted gene set enrichment analysis (GSEA), ingenuity pathway analysis (IPA), and single-cell analysis. DEHG in HSC-MPP/GMP/MEP had greater expression in their respective cell types compared to other cell types (Figure 4B) and included previously reported lineage specific genes such as *SLAMF1* and *MPO* in GMPs [46,47] and *GATA1* and *KLF1* in MEPs[48-50]. Furthermore, genes differentially bound in GMP were enriched for pathways linked to myelopoiesis and granulopoiesis (Figure S8A) while genes linked to differentially bound regions in MEP were enriched for pathways linked to erythropoiesis (Figure S8B). Only 16 HSC-MPP-specific genes were identified precluding pathway analysis. However, these genes included stem cell regulators [51] such as *HOXB1*, *HOXB2*, *HOXB4* (Table S5A).

We also explored expression of DEHG in single cells across hematopoietic differentiation (Figure 4Ci) [3]. HSC-MPP-associated DEHG ($DEHG^{HSC}$) were enriched in cells annotated as HSC clusters (Figure 4Cii), while $DEHG^{GMP}$ were enriched in monocyte clusters (Figure 4Ciii) and $DEHG^{MEP}$ in erythroid lineage cells (Figure 4Civ). Together these data support our hypothesis that heptad occupancy at regulatory elements can be linked to cell-specific transcriptional output.

Heptad TF binding at regulators of genes crucial for myeloid and erythroid cell development

We next asked whether we could detect specific patterns of heptad occupancy at known lineage-specific genes with roles in mature monocyte and granulocyte maintenance (myeloid cell development), linked to erythroid cell development and heme metabolism (erythroid cell development), and linked to stem cell function (Table S5B). From these lists we focused on genes whose promoters were looped to a putative enhancer in any of our HiChIP datasets. We identified 40 P-E pairs from the myeloid genes, 91 from the erythroid genes, and 81 from the stem cell genes (Table S5B) and used *k*-means clustering of TF binding signals to compare heptad occupancy patterns at each P-E pair in each cell type. We observed variable TF binding across associated promoter and enhancer regions. Genes in cluster 1 (C1) showed TF enrichment at promoters, genes in cluster 2 (C2) showed TF enrichment at enhancers, and genes in clusters 3 and 4 (C3, C4) had TF enrichment at both promoters and enhancers (Figure 5A, Figure 5B, Figure S9A). Furthermore, several TF-specific observations were evident. First, ERG occupancy at stem cell, myeloid, and erythroid genes was generally highest in HSC-MPPs, and to a lesser extent GMPs, consistent with its role in maintaining stem cells (Figure S9B, Table S6) [30,23,29]. Second, FLI1 and RUNX1 were bound across both myeloid- and erythroid- gene associated regions across all differentiation stages (Figure S9C). Third, GATA2, LYL1, and LMO2 have increased occupancy at myeloid- and erythroid- specific regulatory regions during lineage commitment (Figure S9D, Table S6). Fourth, high TAL1 occupancy was observed in MEPs and their precursor CMPs, and this occurred at regions linked to erythroid genes and regions linked to myeloid genes (Figure S9E, Table S6). Binding in CMPs may indicate a role for TAL1 in priming regulatory regions and recruiting activators or repressors in downstream cell types. Finally, there was a general pattern of heptad TF occupancy at promoters and enhancers of lineage-specific genes

even in the earliest stem cells, suggesting that heptad factors bind lineage-specific regulatory regions prior to lineage commitment and subsequent differentiation.

To further explore whether heptad binding in differentiated cell types is initiated in early stem cells, we identified ChIPseq peaks bound by a single heptad factor in HSC-MPPs and compared TF binding density at these regions across cell types. Regions with peaks in HSC-MPPs had additional heptad factors bound in committed progenitors. For example, peaks bound by ERG in HSC-MPPs (Figure S10A) showed prominent binding of FLI1 and RUNX1, in GMPs, and binding of all heptad factors except for ERG in MEPs (Figure S10B, C). Furthermore, regions bound by single heptad TFs in HSC-MPPs were subsequently bound by PU.1 in dendritic cells or GATA1 in proerythroblasts (Figure S10D). Overall, these data support dynamic assembly of heptad factor complexes at sites subsequently occupied by lineage-determining TFs. Further experimentation will resolve whether heptad factors actively prime CREs for utility in more mature cell types.

Regulatory regions with cell-type-specific heptad occupancy have distinct epigenetic features.

To better understand the underlying mechanisms regulating dynamic heptad occupancy during blood formation we used our ChIP datasets to annotate and cluster ~85000 regions previously shown to be accessible in any of our four cell types [4]. The resulting UMAP projection of the accessibility landscape could be segmented into 13 clusters (Figure 6A). Clusters 1-3 had characteristics of promoters, including high H3K4me3 signal (Figure 6B). We could further classify cluster 1 as active promoters (H3K4me3 and H3K27ac), cluster 2 as bivalent promoters (H3K4me3 and H3K27me3 enriched), and cluster 3 as active promoters which were also bound by CTCF. These classifications aligned with ChromHMM annotation of the same regions (Figure 6C) [52]. Cluster 4 was enriched for CTCF alone (Figure

6B) and may contain regions involved in 3D genome organization. Clusters 5-11 could be broadly characterized as non-promoter regulatory regions which again aligned with their ChromHMM annotations (Figure 6C). Most of the variations between these regions mapped to variable TF occupancy in specific cell types (Figure S11A), which was particularly pronounced in clusters 8-11 which also showed cell-type-specific changes in chromatin accessibility (Figure 6B). For example, cluster 9 and 10 showed high accessibility in MEP compared to GMP, accompanied by high TAL1 and LYL1 occupancy (Figure S11A), suggesting that regions in these clusters may function in the erythroid lineage.

Visualizing TF signal across the accessibility landscape revealed variable occupancy patterns across cell types (Figure S11B). FLI1, RUNX1, and ERG had similar distributions across all cell types, with highest FLI1 and RUNX1 occupancy in CMP and GMP, and ERG occupancy reduced in MEP. LYL1 and LMO2 had similar distributions with enrichment in clusters 5/6/7 in GMP and clusters 9/10 in MEP (Figure S11B). TAL1 was also enriched in clusters 9/10 in MEP with lower binding in other cell types, while PU.1 was enriched in cluster 5/6/7 in CMP and GMP (Figure S11B). GATA2 had a unique occupancy pattern with enrichment in cluster 5/6 in HSC-MPP and GMP and additional occupancy in cluster 9 in CMP and MEP (Figure S11B) which may reflect distinct roles of GATA2 in early hematopoiesis and subsequent erythroid specification [53].

We then made pairwise comparisons of TF binding signals in GMP vs MEP, HSC-MPP vs GMP, and HSC-MPP vs MEP (Figure 6D, Figure S11C). Comparing GMP (Figure 6D, green) to MEP (Figure 6D, orange), there were distinct zones of TF enrichment in each cell type (Figure 6D , grey dotted line [GMP], black dotted line [MEP]) except for TAL1 which showed no region of enrichment in GMP. To confirm that regions enriched for heptad

TF binding in GMP and MEP represent lineage-specific regulators we mapped ChIP seq signal from lineage defining TFs in two mature cell types (Figure 6E, F). Regions with high heptad occupancy in GMP showed similar occupancy of PU.1 in dendritic cells (Figure 6E) [54], while regions with high heptad occupancy in MEP showed similar occupancy of GATA1 in proerythroblasts (Figure 6F)[6]. Taken together these data show that regulatory regions with heptad occupancy in progenitor populations are regions occupied by lineage-specific TFs in more mature cells.

Cell-type specificity of regulatory elements is encoded in the underlying motif composition.

TFs bind DNA via consensus binding motifs whose sequence and relative locations in each regulator determine which TF complexes can potentially bind. To better understand enhancer features underpinning lineage-specific TF occupancy we selected regions with differential accessibility and TF occupancy in HSC-MPP (Figure 7Ai, 3992 regions), GMP (Figure 7Bi, 4395 regions) and MEP (Figure 7Ci, 3469 regions) and developed machine learning models to predict cell-type associations based on DNA sequence motifs. All models showed high sensitivity and specificity to associate regions with cell types (Figure 7Aii, Bii, Cii, Table S7). To understand how specific motifs contribute to assigning regulatory elements to cell types, we calculated SHapley Additive exPlanations (SHAP) values for each cell type (Figure 7Aiii, Biii, Ciii, Table S8). Distinct combinations of motifs were found to contribute to each model, many of whihc fit the expected profile for cell-specific regulatory regions. For example, ETS motifs had positive SHAP values in the GMP model but negative values in the MEP model, consistent with known roles for ETS factors such as PU.1 in driving myeloid differentiation, while GATA motifs had positive SHAP values in HSC-MPP and MEP models but negative values in the GMP model, consistent with GATA factors such as

GATA1 driving the erythroid lineage [55,56]. Motifs not corresponding to heptad factors had high SHAP scores in all models, consistent with heptad factors binding at lineage-specific enhancers in the context of larger regulatory protein complexes. We explored expression of non-heptad TFs corresponding to motifs predicted to positively impact SHAP scores for each model (Table S8). *GATA3* (HSC-MPP model), *PU.1* and *SPI-B* (GMP model) and *GATA1* (MEP model) had lineage-specific expression patterns consistent with known expression and activity of these factors [53,56,49,57-60] (Figure S12A-C). All other motif-associated-TFs had minimal expression variation across cell types.

Finally, we used our models to classify functional HSC regulatory elements identified through analysis of γ-retroviral integration sites (γRV-IS) in gene therapy patients [61]. γRV-IS had similar features to HSC-MPP-specific regions (Figure S7A, Figure S12D). Importantly, our models scored γRV-IS as most likely to be HSC-MPP-specific, confirming the validity of our analyses (Figure S12E).

DISCUSSION

In this study we explored genome-wide dynamics of chromatin occupancy and structure in four cell types along the HSC to myeloid/erythroid differentiation axis. Analysis of putative CREs with direct looping to heptad promoter regions revealed greater regulatory complexity than previously appreciated. Previous work identified a set of nine regions with combinatorial heptad binding in human HSPCs whose relative accessibility is sufficient to predict cell identity [23]. The current study identified more than 30 putative CREs directly looped to heptad promoters in at least one cell type (Figure 3B). While most previously described heptad regulatory elements have been tested in functional assays (Table S4), further validation is required to precisely understand individual and cooperative roles of specific CREs for each heptad gene [62,63]. Surprisingly, some previously identified enhancers were not directly looped to heptad promoters in this analysis. For example *GATA2* −117, an enhancer dysregulated in inv(3) AML [64,65], was only indirectly looped to the *GATA2* promoter, and therefore excluded from our network model (Figure S6B). However, lack of direct binding does not preclude a role for this enhancer in regulating GATA2. Indeed, previous studies have found that enhancer – enhancer interactions can stabilise and amplify TF binding, and the resulting enhancer communities can drive and coordinate cell-type specific gene expression [66,67]. Consistent with this model, we observed a trend of increasingly complex chromatin looping interactions involving heptad promoters and putative enhancers as cells committed to either erythroid or myeloid lineages (Figure 3B). Increased looping may restrict committed progenitors from accessing alternate fates or participate in shutting off stem cell programs. Intriguingly, the stem cell gene *ERG* [30,23,29] acquired loops to a +1200kb region in MEPs (Figure 3Ai), while increased long range looping and heptad binding occurred at the +45 enhancer of the erythroid gene *TAL1* [68,14] (Figure S6Di). Thus, increased looping can be associated with genes subsequently down- or up-regulated. Further

experimentation is needed to fully understand rapidly increased chromatin looping in committed progenitors.

Motif-based analysis showed that heptad-occupied CREs contain information encoding cell-type-specificity. Among motifs that contributed to cell-type specificity we found motifs for additional factors including STAT, interferon-regulatory factor (IRF), and homeodomain TFs (Figure 7, Table S8). While our focus was on heptad factors, it is clear that additional regulators co-occupy CREs and contribute to enhancer specificity [69]. To date these studies have focussed on transcriptional cofactors with broad function such as P53 and MED14 [70], and further work is required to understand how heptad factors, general cofactors, and intrinsic features of enhancers and promoters cooperate to regulate transcriptional programs throughout haematopoiesis.

Numerous bone marrow failure syndromes have an underlying germline genetic cause and may be amenable to gene therapy approaches targetting primitive HSCs to enable long term benefits. However, ectopic gene expression may lead to lineage skewing and other unwanted outcomes. Cell-type-specific enhancers are conserved across evolution and can potentially be used to restrict transgene expression [71,72,6]. Indeed, erythroid-specific *GATA1* enhancers can drive restricted expression of GATA1 to rescue erythroid differentiation in Diamond-Blackfan Anaemia [73]. However, identifying enhancers with appropriate spatiotemporal activity is not trivial. While experimental techniques for massively parallel interrogation of enhancer activity and specificity have been developed [74,63], our data provides a catalogue of putative enhancers that could fast track functional studies.

GRNs are significantly perturbed in multiple AML subtypes [75,8], often as a result of specific translocation events including RUNX1-RUNX1T1 [76,77]and inv(3) [65,64]. Recent sequencing studies found that small structural variants (SVs) can mediate enhancer hijacking and subsequent oncogenic transformation [63,78]. However, understanding the effects of SVs requires significant knowledge of cell-specific enhancers; our data set provides extensive annotation of regulatory regions that could be used to guide functional experiments exploring the effects of specific SVs.

While extensive, our dataset is limited to myeloid cell types. Sorted HSPCs, particularly CMPs, are also inherently heterogenous [79-81] which may limit our ability to resolve differences between adjacent cell types. Furthermore, while capable of reconstituting transplanted patients, mobilized cells may have subtle epigenomic differences to their unstimulated counterparts. Finally, even using low cell input techniques, our datasets represent averaged binding and looping probabilities captured across a shifting epigenetic landscape, rather than truly capturing GRNs, which operate at the level of single cells [7]. Nonetheless, our dataset represents the most extensive epigenetic characterisation of primary human blood progenitors to date and provides important insights for future studies.

ACKNOWLEDGEMENTS

The authors acknowledge members of the Pimanda group for assistance in thawing apheresis bags. The authors thank the South Australian Cancer Research Biobank (SACRB), and all patients who donated samples. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. JAIT was supported by the Anthony Rothe Memorial Trust. BG acknowledges funding from Wellcome (206328/Z/17/Z). JEP was supported by grants from the National Health and Medical Research Council of Australia (GNT1139787, GNT2011627, MRF1200271), a translational program grant from the Leukemia Lymphoma Society (LLS)-Snowdome Foundation-Leukaemia Foundation (6620-21), and the Anthony Rothe Memorial Trust.

AUTHOR CONTRIBUTIONS

Conceptualization: SS, JAIT, JWHW, JEP

Methodology: SS, JAIT, YH, SJ, SSh, ES, CB, PSW, DCF, DB, AT, SWL

Investigation: SS, JAIT, YH, SSh, ES, SJ, DJC, KY, VA, TOB, JAP, IL, SMP, MKG

Formal analysis: SS, JAIT, PC, FCK, FV, EW, BG, HAR, JWHW, JEP

Supervision: JAIT, JWHW, JEP

Writing - original draft: SS, JAIT, JEP

Writing – review and editing: JAIT, JEP

Funding acquisition: JEP

CONFLICTS of INTEREST

The authors declare no relevant competing interests.

REFERENCES

1. Doulatov S, Notta F, Laurenti E, Dick JE. Hematopoiesis: a human perspective. *Cell Stem Cell*. 2012;10(2):120-136.

2. Laurenti E, Gottgens B. From haematopoietic stem cells to complex differentiation landscapes. *Nature*. 2018;553(7689):418-426.

3. Setty M, Kiseliovas V, Levine J, Gayoso A, Mazutis L, Pe'er D. Characterization of cell fate probabilities in single-cell data with Palantir. *Nat Biotechnol*. 2019;37(4):451-460.

4. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, Snyder MP, Pritchard JK, Kundaje A, Greenleaf WJ, Majeti R, Chang HY. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet*. 2016;48(10):1193-1203.

5. Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN, McConkey ME, Habib N, Yosef N, Chang CY, Shay T, Frampton GM, Drake AC, Leskov I, Nilsson B, Preffer F, Dombkowski D, Evans JW, Liefeld T, Smutko JS, Chen J, Friedman N, Young RA, Golub TR, Regev A, Ebert BL. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*. 2011;144(2):296-309.

6. Xu J, Shao Z, Glass K, Bauer DE, Pinello L, Van Handel B, Hou S, Stamatoyannopoulos JA, Mikkola HK, Yuan GC, Orkin SH. Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Dev Cell*. 2012;23(4):796-811.

7. Davidson EH. Emerging properties of animal gene regulatory networks. *Nature*. 2010;468(7326):911-920.

8. Thoms JAI, Beck D, Pimanda JE. Transcriptional networks in acute myeloid leukemia. *Genes Chromosomes Cancer*. 2019;58(12):859-874.

9. Nasrallah R, Fast EM, Solaimani P, Knezevic K, Eliades A, Patel R, Thambyrajah R, Unnikrishnan A, Thoms J, Beck D, Vink CS, Smith A, Wong J, Shepherd M, Kent D, Roychoudhuri R, Paul F, Klippert J, Hammes A, Willnow T, Gottgens B, Dzierzak E, Zon LI, Lacaud G, Kouskoff V, Pimanda JE. Identification of novel regulators of developmental hematopoiesis using Endoglin regulatory elements as molecular probes. *Blood*. 2016;128(15):1928-1939.

10. Lara-Astiaso D, Weiner A, Lorenzo-Vivas E, Zaretsky I, Jaitin DA, David E, Keren-Shaul H, Mildner A, Winter D, Jung S, Friedman N, Amit I. Immunogenetics. Chromatin state dynamics during blood formation. *Science*. 2014;345(6199):943-949.

11. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009;326(5950):289-293.

12. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, Diao Y, Liang J, Zhao H, Lobanenkov VV, Ecker JR, Thomson JA, Ren B. Chromatin architecture reorganization during stem cell differentiation. *Nature*. 2015;518(7539):331-336.

13. Oram SH, Thoms JA, Pridans C, Janes ME, Kinston SJ, Anand S, Landry JR, Lock RB, Jayaraman PS, Huntly BJ, Pimanda JE, Gottgens B. A previously unrecognized promoter of LMO2 forms part of a transcriptional regulatory circuit mediating LMO2 expression in a subset of T-acute lymphoblastic leukaemia patients. *Oncogene*. 2010;29(43):5796-5808.

14. Curtis DJ, Salmon JM, Pimanda JE. Concise review: Blood relatives: formation and regulation of hematopoietic stem cells by the basic helix-loop-helix transcription factors stem

cell leukemia and lymphoblastic leukemia-derived sequence 1. *Stem Cells*. 2012;30(6):1053-1058.

15. Li Y, Luo H, Liu T, Zacksenhaus E, Ben-David Y. The ets transcription factor Fli-1 in development, cancer and disease. *Oncogene*. 2015;34(16):2022-2031.

16. Hahn CN, Chong CE, Carmichael CL, Wilkins EJ, Brautigan PJ, Li XC, Babic M, Lin M, Carmagnac A, Lee YK, Kok CH, Gagliardi L, Friend KL, Ekert PG, Butcher CM, Brown AL, Lewis ID, To LB, Timms AE, Storek J, Moore S, Altree M, Escher R, Bardy PG, Suthers GK, D'Andrea RJ, Horwitz MS, Scott HS. Heritable GATA2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia. *Nat Genet*. 2011;43(10):1012-1017.

17. Pimanda JE, Ottersbach K, Knezevic K, Kinston S, Chan WY, Wilson NK, Landry JR, Wood AD, Kolb-Kokocinski A, Green AR, Tannahill D, Lacaud G, Kouskoff V, Gottgens B. Gata2, Fli1, and Scl form a recursively wired gene-regulatory circuit during early hematopoietic development. *Proc Natl Acad Sci U S A*. 2007;104(45):17692-17697.

18. Cai Z, de Bruijn M, Ma X, Dortland B, Luteijn T, Downing RJ, Dzierzak E. Haploinsufficiency of AML1 affects the temporal and spatial generation of hematopoietic stem cells in the mouse embryo. *Immunity*. 2000;13(4):423-431.

19. Mangan JK, Speck NA. RUNX1 mutations in clonal myeloid disorders: from conventional cytogenetics to next generation sequencing, a story 40 years in the making. *Crit Rev Oncog*. 2011;16(1-2):77-91.

20. Thoms JA, Birger Y, Foster S, Knezevic K, Kirschenbaum Y, Chandrakanthan V, Jonquieres G, Spensberger D, Wong JW, Oram SH, Kinston SJ, Groner Y, Lock R, MacKenzie KL, Gottgens B, Izraeli S, Pimanda JE. ERG promotes T-acute lymphoblastic leukemia and is transcriptionally regulated in leukemic cells by a stem cell enhancer. *Blood*. 2011;117(26):7079-7089.

21. Beck D, Thoms JA, Perera D, Schutte J, Unnikrishnan A, Knezevic K, Kinston SJ, Wilson NK, O'Brien TA, Gottgens B, Wong JW, Pimanda JE. Genome-wide analysis of transcriptional regulators in human HSPCs reveals a densely interconnected network of coding and noncoding genes. *Blood*. 2013;122(14):e12-22.

22. Diffner E, Beck D, Gudgin E, Thoms JA, Knezevic K, Pridans C, Foster S, Goode D, Lim WK, Boelen L, Metzeler KH, Micklem G, Bohlander SK, Buske C, Burnett A, Ottersbach K, Vassiliou GS, Olivier J, Wong JW, Gottgens B, Huntly BJ, Pimanda JE. Activity of a heptad of transcription factors is associated with stem cell programs and clinical outcome in acute myeloid leukemia. *Blood*. 2013;121(12):2289-2300.

23. Thoms JAI, Truong P, Subramanian S, Knezevic K, Harvey G, Huang Y, Seneviratne JA, Carter DR, Joshi S, Skhinas J, Chacon D, Shah A, de Jong I, Beck D, Gottgens B, Larsson J, Wong JWH, Zanini F, Pimanda JE. Disruption of a GATA2-TAL1-ERG regulatory circuit promotes erythroid transition in healthy and leukemic stem cells. *Blood*. 2021;138(16):1441-1455.

24. Mandoli A, Singh AA, Jansen PW, Wierenga AT, Riahi H, Franci G, Prange K, Saeed S, Vellenga E, Vermeulen M, Stunnenberg HG, Martens JH. CBFB-MYH11/RUNX1 together with a compendium of hematopoietic regulators, chromatin modifiers and basal transcription factors occupies self-renewal genes in inv(16) acute myeloid leukemia. *Leukemia*. 2014;28(4):770-778.

25. Wilson NK, Foster SD, Wang X, Knezevic K, Schutte J, Kaimakis P, Chilarska PM, Kinston S, Ouwehand WH, Dzierzak E, Pimanda JE, de Bruijn MF, Gottgens B. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell*. 2010;7(4):532-544.

26. Schmidl C, Rendeiro AF, Sheffield NC, Bock C. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat Methods*. 2015;12(10):963-965.

27. Alinejad-Rokny H, Ghavami Modegh R, Rabiee HR, Ramezani Sarbandi E, Rezaie N, Tam KT, Forrest ARR. MaxHiC: A robust background correction model to identify biologically relevant chromatin interactions in Hi-C and capture Hi-C experiments. *PLoS Comput Biol*. 2022;18(6):e1010241.

28. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA; 2016.

29. Tursky ML, Beck D, Thoms JA, Huang Y, Kumari A, Unnikrishnan A, Knezevic K, Evans K, Richards LA, Lee E, Morris J, Goldberg L, Izraeli S, Wong JW, Olivier J, Lock RB, MacKenzie KL, Pimanda JE. Overexpression of ERG in cord blood progenitors promotes expansion and recapitulates molecular signatures of high ERG leukemias. *Leukemia*. 2015;29(4):819-827.

30. Knudsen KJ, Rehn M, Hasemann MS, Rapin N, Bagger FO, Ohlsson E, Willer A, Frank AK, Sondergaard E, Jendholm J, Thoren L, Lee J, Rak J, Theilgaard-Monch K, Porse BT. ERG promotes the maintenance of hematopoietic stem cells by restricting their differentiation. *Genes Dev*. 2015;29(18):1915-1929.

31. Takayama N, Murison A, Takayanagi SI, Arlidge C, Zhou S, Garcia-Prat L, Chan-Seng-Yue M, Zandi S, Gan OI, Boutzen H, Kaufmann KB, Trotman-Grant A, Schoof E, Kron K, Diaz N, Lee JJY, Medina T, De Carvalho DD, Taylor MD, Vaquerizas JM, Xie SZ, Dick JE, Lupien M. The Transition from Quiescent to Activated States in Human Hematopoietic Stem Cells Is Governed by Dynamic 3D Genome Reorganization. *Cell Stem Cell*. 2021;28(3):488-501 e410.

32. 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. An experimentally validated network of nine

haematopoietic transcription factors reveals mechanisms of cell state stability. *Elife*. 2016;5:e11469.

33. Wilson NK, Miranda-Saavedra D, Kinston S, Bonadies N, Foster SD, Calero-Nieto F, Dawson MA, Donaldson IJ, Dumon S, Frampton J, Janky R, Sun XH, Teichmann SA, Bannister AJ, Gottgens B. The transcriptional program controlled by the stem cell leukemia gene Scl/Tal1 during early embryonic hematopoietic development. *Blood*. 2009;113(22):5456-5465.

34. Sinclair AM, Gottgens B, Barton LM, Stanley ML, Pardanaud L, Klaine M, Gering M, Bahn S, Sanchez M, Bench AJ, Fordham JL, Bockamp E, Green AR. Distinct 5' SCL enhancers direct transcription to developing brain, spinal cord, and endothelium: neural expression is mediated by GATA factor binding sites. *Dev Biol*. 1999;209(1):128-142.

35. Moignard V, Macaulay IC, Swiers G, Buettner F, Schutte J, Calero-Nieto FJ, Kinston S, Joshi A, Hannah R, Theis FJ, Jacobsen SE, de Bruijn MF, Gottgens B. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat Cell Biol*. 2013;15(4):363-372.

36. Landry JR, Bonadies N, Kinston S, Knezevic K, Wilson NK, Oram SH, Janes M, Piltz S, Hammett M, Carter J, Hamilton T, Donaldson IJ, Lacaud G, Frampton J, Follows G, Kouskoff V, Gottgens B. Expression of the leukemia oncogene Lmo2 is controlled by an array of tissue-specific elements dispersed over 100 kb and bound by Tal1/Lmo2, Ets, and Gata factors. *Blood*. 2009;113(23):5783-5792.

37. Gottgens B, Broccardo C, Sanchez MJ, Deveaux S, Murphy G, Gothert JR, Kotsopoulou E, Kinston S, Delaney L, Piltz S, Barton LM, Knezevic K, Erber WN, Begley CG, Frampton J, Green AR. The scl +18/19 stem cell enhancer is not required for hematopoiesis: identification of a 5' bifunctional hematopoietic-endothelial enhancer bound by Fli-1 and Elf-1. *Mol Cell Biol*. 2004;24(5):1870-1883.

38. Gottgens B, Ferreira R, Sanchez MJ, Ishibashi S, Li J, Spensberger D, Lefevre P, Ottersbach K, Chapman M, Kinston S, Knezevic K, Hoogenkamp M, Follows GA, Bonifer C, Amaya E, Green AR. cis-Regulatory remodeling of the SCL locus during vertebrate evolution. *Mol Cell Biol*. 2010;30(24):5741-5751.

39. Gottgens B, Nastos A, Kinston S, Piltz S, Delabesse EC, Stanley M, Sanchez MJ, Ciau-Uitz A, Patient R, Green AR. Establishing the transcriptional programme for blood: the SCL stem cell enhancer is regulated by a multiprotein complex containing Ets and GATA factors. *EMBO J*. 2002;21(12):3039-3050.

40. Sanchez M, Gottgens B, Sinclair AM, Stanley M, Begley CG, Hunter S, Green AR. An SCL 3' enhancer targets developing endothelium together with embryonic and adult haematopoietic progenitors. *Development*. 1999;126(17):3891-3904.

41. Bee T, Ashley EL, Bickley SR, Jarratt A, Li PS, Sloane-Stanley J, Gottgens B, de Bruijn MF. The mouse Runx1 +23 hematopoietic stem cell enhancer confers hematopoietic specificity to both Runx1 promoters. *Blood*. 2009;113(21):5121-5124.

42. Nottingham WT, Jarratt A, Burgess M, Speck CL, Cheng JF, Prabhakar S, Rubin EM, Li PS, Sloane-Stanley J, Kong ASJ, de Bruijn MF. Runx1-mediated hematopoietic stem-cell emergence is controlled by a Gata/Ets/SCL-regulated enhancer. *Blood*. 2007;110(13):4188-4197.

43. Chan WY, Follows GA, Lacaud G, Pimanda JE, Landry JR, Kinston S, Knezevic K, Piltz S, Donaldson IJ, Gambardella L, Sablitzky F, Green AR, Kouskoff V, Gottgens B. The paralogous hematopoietic regulators Lyl1 and Scl are coregulated by Ets and GATA factors, but Lyl1 cannot rescue the early Scl-/- phenotype. *Blood*. 2007;109(5):1908-1916.

44. Johnson KD, Kong G, Gao X, Chang YI, Hewitt KJ, Sanalkumar R, Prathibha R, Ranheim EA, Dewey CN, Zhang J, Bresnick EH. Cis-regulatory mechanisms governing stem and progenitor cell transitions. *Sci Adv*. 2015;1(8):e1500503.

45. Wozniak RJ, Boyer ME, Grass JA, Lee Y, Bresnick EH. Context-dependent GATA factor function: combinatorial requirements for transcriptional control in hematopoietic and endothelial cells. *J Biol Chem*. 2007;282(19):14665-14674.

46. Beerman I, Bhattacharya D, Zandi S, Sigvardsson M, Weissman IL, Bryder D, Rossi DJ. Functionally distinct hematopoietic stem cells modulate hematopoietic lineage potential during aging by a mechanism of clonal expansion. *Proc Natl Acad Sci U S A*. 2010;107(12):5465-5470.

47. Strobl H, Takimoto M, Majdic O, Fritsch G, Scheinecker C, Hocker P, Knapp W. Myeloperoxidase expression in CD34+ normal human hematopoietic cells. *Blood*. 1993;82(7):2069-2078.

48. Pevny L, Simon MC, Robertson E, Klein WH, Tsai SF, D'Agati V, Orkin SH, Costantini F. Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. *Nature*. 1991;349(6306):257-260.

49. Suzuki M, Kobayashi-Osaki M, Tsutsumi S, Pan X, Ohmori S, Takai J, Moriguchi T, Ohneda O, Ohneda K, Shimizu R, Kanki Y, Kodama T, Aburatani H, Yamamoto M. GATA factor switching from GATA2 to GATA1 contributes to erythroid differentiation. *Genes Cells*. 2013;18(11):921-933.

50. Rossmann MP, Zon LI. 'Enhancing' red cell fate through epigenetic mechanisms. *Curr Opin Hematol*. 2021;28(3):129-137.

51. Cabezas-Wallscheid N, Klimmeck D, Hansson J, Lipka DB, Reyes A, Wang Q, Weichenhan D, Lier A, von Paleske L, Renders S, Wunsche P, Zeisberger P, Brocks D, Gu L, Herrmann C, Haas S, Essers MAG, Brors B, Eils R, Huber W, Milsom MD, Plass C, Krijgsveld J, Trumpp A. Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. *Cell Stem Cell*. 2014;15(4):507-522.

52. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*. 2012;9(3):215-216.

53. Bresnick EH, Lee HY, Fujiwara T, Johnson KD, Keles S. GATA switches as developmental drivers. *J Biol Chem*. 2010;285(41):31087-31093.

54. Seguin-Estevez Q, Dunand-Sauthier I, Lemeille S, Iseli C, Ibberson M, Ioannidis V, Schmid CD, Rousseau P, Barras E, Geinoz A, Xenarios I, Acha-Orbea H, Reith W. Extensive remodeling of DC function by rapid maturation-induced transcriptional silencing. *Nucleic Acids Res*. 2014;42(15):9641-9655.

55. Rosenbauer F, Wagner K, Kutok JL, Iwasaki H, Le Beau MM, Okuno Y, Akashi K, Fiering S, Tenen DG. Acute myeloid leukemia induced by graded reduction of a lineage-specific transcription factor, PU.1. *Nat Genet*. 2004;36(6):624-630.

56. Zhang P, Zhang X, Iwama A, Yu C, Smith KA, Mueller BU, Narravula S, Torbett BE, Orkin SH, Tenen DG. PU.1 inhibits GATA-1 function and erythroid differentiation by blocking GATA-1 DNA binding. *Blood*. 2000;96(8):2641-2648.

57. Dahl R, Ramirez-Bergeron DL, Rao S, Simon MC. Spi-B can functionally replace PU.1 in myeloid but not lymphoid development. *EMBO J*. 2002;21(9):2220-2230.

58. Ku CJ, Hosoya T, Maillard I, Engel JD. GATA-3 regulates hematopoietic stem cell maintenance and cell-cycle entry. *Blood*. 2012;119(10):2242-2251.

59. Frelin C, Herrington R, Janmohamed S, Barbara M, Tran G, Paige CJ, Benveniste P, Zuniga-Pflucker JC, Souabni A, Busslinger M, Iscove NN. GATA-3 regulates the self-renewal of long-term hematopoietic stem cells. *Nat Immunol*. 2013;14(10):1037-1044.

60. Nerlov C, Graf T. PU.1 induces myeloid lineage commitment in multipotent hematopoietic progenitors. *Genes Dev*. 1998;12(15):2403-2412.

61. Wunsche P, Eckert ESP, Holland-Letz T, Paruzynski A, Hotz-Wagenblatt A, Fronza R, Rath T, Gil-Farina I, Schmidt M, von Kalle C, Klein C, Ball CR, Herbst F, Glimm H.

Mapping Active Gene-Regulatory Regions in Human Repopulating Long-Term HSCs. *Cell Stem Cell*. 2018;23(1):132-146 e139.

62. Canver MC, Smith EC, Sher F, Pinello L, Sanjana NE, Shalem O, Chen DD, Schupp PG, Vinjamur DS, Garcia SP, Luc S, Kurita R, Nakamura Y, Fujiwara Y, Maeda T, Yuan GC, Zhang F, Orkin SH, Bauer DE. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature*. 2015;527(7577):192-197.

63. Xu J, Song F, Lyu H, Kobayashi M, Zhang B, Zhao Z, Hou Y, Wang X, Luan Y, Jia B, Stasiak L, Wong JH, Wang Q, Jin Q, Jin Q, Fu Y, Yang H, Hardison RC, Dovat S, Platanias LC, Diao Y, Yang Y, Yamada T, Viny AD, Levine RL, Claxton D, Broach JR, Zheng H, Yue F. Subtype-specific 3D genome alteration in acute myeloid leukaemia. *Nature*. 2022;611(7935):387-398.

64. Yamazaki H, Suzuki M, Otsuki A, Shimizu R, Bresnick EH, Engel JD, Yamamoto M. A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression. *Cancer Cell*. 2014;25(4):415-427.

65. Groschel S, Sanders MA, Hoogenboezem R, de Wit E, Bouwman BAM, Erpelinck C, van der Velden VHJ, Havermans M, Avellino R, van Lom K, Rombouts EJ, van Duin M, Dohner K, Beverloo HB, Bradner JE, Dohner H, Lowenberg B, Valk PJM, Bindels EMJ, de Laat W, Delwel R. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell*. 2014;157(2):369-381.

66. Madsen JGS, Madsen MS, Rauch A, Traynor S, Van Hauwaert EL, Haakonsson AK, Javierre BM, Hyldahl M, Fraser P, Mandrup S. Highly interconnected enhancer communities control lineage-determining genes in human mesenchymal stem cells. *Nat Genet*. 2020;52(11):1227-1238.

67. Zuin J, Roth G, Zhan Y, Cramard J, Redolfi J, Piskadlo E, Mach P, Kryzhanovska M, Tihanyi G, Kohler H, Eder M, Leemans C, van Steensel B, Meister P, Smallwood S,

Giorgetti L. Nonlinear control of transcription through enhancer-promoter interactions. *Nature*. 2022;604(7906):571-577.

68. Chiu SK, Orive SL, Moon MJ, Saw J, Ellis S, Kile BT, Huang Y, Chacon D, Pimanda JE, Beck D, Hamilton JR, Tremblay CS, Curtis DJ. Shared roles for Scl and Lyl1 in murine platelet production and function. *Blood*. 2019;134(10):826-835.

69. Spitz F, Furlong EE. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet*. 2012;13(9):613-626.

70. Neumayr C, Haberle V, Serebreni L, Karner K, Hendy O, Boija A, Henninger JE, Li CH, Stejskal K, Lin G, Bergauer K, Pagani M, Rath M, Mechtler K, Arnold CD, Stark A. Differential cofactor dependencies define distinct types of human enhancers. *Nature*. 2022;606(7913):406-413.

71. Cornejo-Paramo P, Roper K, Degnan SM, Degnan BM, Wong ES. Distal regulation, silencers, and a shared combinatorial syntax are hallmarks of animal embryogenesis. *Genome Res*. 2022;32(3):474-487.

72. Goode DK, Obier N, Vijayabaskar MS, Lie ALM, Lilly AJ, Hannah R, Lichtinger M, Batta K, Florkowska M, Patel R, Challinor M, Wallace K, Gilmour J, Assi SA, Cauchy P, Hoogenkamp M, Westhead DR, Lacaud G, Kouskoff V, Gottgens B, Bonifer C. Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. *Dev Cell*. 2016;36(5):572-587.

73. Voit RA, Liao X, Cohen B, Armant M, Kamal E, Huang M-M, Clarke W, Williams DA, Shimamura A, Sankaran VG. Regulated Expression of GATA1 As a Gene Therapy Cure for Diamond-Blackfan Anemia. *Blood*. 2022;140(Supplement 1):986-987.

74. Edginton-White B, Maytum A, Kellaway SG, Goode DK, Keane P, Pagnuco I, Assi SA, Ames L, Clarke M, Cockerill PN, Gottgens B, Cazier JB, Bonifer C. A genome-wide

relay of signalling-responsive enhancers drives hematopoietic specification. *Nat Commun*. 2023;14(1):267.

75. Assi SA, Imperato MR, Coleman DJL, Pickin A, Potluri S, Ptasinska A, Chin PS, Blair H, Cauchy P, James SR, Zacarias-Cabeza J, Gilding LN, Beggs A, Clokie S, Loke JC, Jenkin P, Uddin A, Delwel R, Richards SJ, Raghavan M, Griffiths MJ, Heidenreich O, Cockerill PN, Bonifer C. Subtype-specific regulatory network rewiring in acute myeloid leukemia. *Nat Genet*. 2019;51(1):151-162.

76. Loke J, Assi SA, Imperato MR, Ptasinska A, Cauchy P, Grabovska Y, Soria NM, Raghavan M, Delwel HR, Cockerill PN, Heidenreich O, Bonifer C. RUNX1-ETO and RUNX1-EVI1 Differentially Reprogram the Chromatin Landscape in t(8;21) and t(3;21) AML. *Cell Rep*. 2017;19(8):1654-1668.

77. Ptasinska A, Assi SA, Mannari D, James SR, Williamson D, Dunne J, Hoogenkamp M, Wu M, Care M, McNeill H, Cauchy P, Cullen M, Tooze RM, Tenen DG, Young BD, Cockerill PN, Westhead DR, Heidenreich O, Bonifer C. Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-wide changes in chromatin structure and transcription factor binding. *Leukemia*. 2012;26(8):1829-1841.

78. Botten GA, Zhang Y, Dudnyk K, Liu X, Sanders JT, Imanci A, Droin N, Cao H, Kaphle P, Dickerson KE, Chen M, Chen W, Solary E, Ly P, Zhou J, Xu J. Structural Variation Cooperates with Permissive Chromatin to Control Enhancer Hijacking-Mediated Oncogenic Transcription. *Blood*. 2022;140(Supplement 1):1007-1008.

79. Buenrostro JD, Corces MR, Lareau CA, Wu B, Schep AN, Aryee MJ, Majeti R, Chang HY, Greenleaf WJ. Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell*. 2018;173(6):1535-1548 e1516.

80. Karamitros D, Stoilova B, Aboukhalil Z, Hamey F, Reinisch A, Samitsch M, Quek L, Otto G, Repapi E, Doondeea J, Usukhbayar B, Calvo J, Taylor S, Goardon N, Six E, Pflumio

F, Porcher C, Majeti R, Gottgens B, Vyas P. Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells. *Nat Immunol*. 2018;19(1):85-97.

81. Drissen R, Thongjuea S, Theilgaard-Monch K, Nerlov C. Identification of two distinct pathways of human myelopoiesis. *Sci Immunol*. 2019;4(35).

FIGURE LEGENDS

Figure 1. ***Genome wide patterns of heptad transcription factor binding in fractionated primary human HSPCs.*** **A)** Human MNCs were isolated from G-CSF-stimulated donors or patients with a non-hematologic malignancy before being enriched for CD34 expression using MACS and further sub-fractionated into individual stem and progenitor cells based on surface marker expression using FACS (colored cells are those studied in this manuscript). **B)** The workflow and analysis pipeline followed for ChIPmentation and ChIP-seq experiments. **C)** UCSC browser track at the *RUNX1* locus (GRCh38 chr21:34,627,969-35,209,177) showing the RPKM-normalized signal from FLI1, ERG, GATA2, RUNX1, TAL1, LYL1, and LMO2, along with H3K4me3, H3K27ac, H3K27me3, IgG (control), and publicly available RNA-seq tracks (GSE75384) for the four cell types. Full UCSC browser tracks are available http://genome.ucsc.edu/s/PimandaLab/Heptad_Regulome **D)** Characterization of identified peaks. Number of TF peaks were identified by macs2 (p-value ≤ 1e-5) and their overall distribution along the genome (as percentages of total peaks identified) is shown. Each peak was assigned as either promoter-like (proximal [orange] or adjacent [blue], based on its distance from the TSS), intragenic [green], or intergenic [red], and enrichment (fraction of peaks containing that motif) calculated for the known ETS, GATA, RUNX, and E-Box motifs.

Figure 2. ***Combinatorial binding of heptad transcription factors is cell-type specific.*** **A)** A composite graph with three components: i) number of combinatorial binding peaks identified in the four cell types, for ii) combinations of 2, 5, 6, and 7 heptad factors and iii) heatmap showing z-scores for the combinations presented in (ii). Star indicates combinations lacking GATA2 and/or TAL1. **B-C)** UCSC browser tracks showing RPKM-normalized signal tracks of the heptad factors, H3K4me3, H3K27ac, H3K27me3, RNA-seq (public data:

GSE75384), and IgG (control) in HSC-MPP, CMP, GMP, and MEP (left to right), at **B)** the *GATA1* locus (GRCh38 chrX:48,724,037-48,839,866), a gene vital for erythroid lineage specification, and at **C)** the *MPO* locus (GRCh38 chr17:58,238,087-58,348,896), a gene specific to the monocytic lineage.

Figure 3. ***Heptad regulatory circuits are remodelled during myeloid progenitor development.*** **A)** Step-wise identification of potential regulatory regions interacting with the ERG promoter: i) Raw HiChIP contact matrix, CTCF, H3K4me3, H3K27ac, IgG, RNA-seq, and significant H3K27ac HiChIP interactions (FDR ≤ 0.01) at the *ERG* locus (GRCh38 chr21:37370238-39198738). The *ERG* promoter is indicated by the green arrow (only those HiChIP interactions where both interacting ends were found at the given locus are shown). ii) Magnified view of the *ERG* locus, with regulators identified to loop to the *ERG* proximal promoter shown as red triangles. iii) FLI1, ERG, GATA2, RUNX1, TAL1, LYL1, and LMO2 peaks at the defined regulators in each individual cell type. The peaks shown are RPKM-normalised and white boxes indicate presence of a computationally called ChIP-seq peak at the specific region. **B)** Summary plot of gene regulatory interactions across the heptad genes: i) Individual heptad gene loci with identified regulators indicated by red markers. ii) Dot plots showing regulatory regions as rows and the four cell types as columns, with size of the dot indicating number of heptad factors bound and bold color indicating the presence of an active regulatory link to the promoter (using H3K27ac HiChIP). Promoters are underlined. iii) Bar plots with individual replicates showing average log2 counts of relevant heptad gene expression in the four cell types (GSE75384).

Figure 4. ***The role of heptad transcription factors in regulating lineage-specific gene expression*****. A)** Schematic of the bioinformatic strategy used to derive regions showing

differential heptad factor binding: i) Candidate regulatory elements (REs) with binding of at least two heptad factors were chosen in the four cell types; and ii) DiffBind was used to filter for regions showing differential enrichment for heptad factors with an FDR < 0.05. To perform DiffBind analysis only HSC-MPP (HSC), GMP, and MEP populations were chosen. iii) These DEH (differentially enriched for heptad) regions were linked to genes either directly (present across a 10 kb promoter region) or indirectly (distal links using significant [FDR < 0.01] H3K27ac- HiChIP interactions), and iv) used as input for multiple characterisation assays. **B)** GSEA plots showing enrichment of derived gene sets in pairwise gene expression comparisons: i) $DEHG^{HSC}$ (genes linked to DEH regions in HSC-MPP) enriched in HSC-MPP with respect to GMP, ii) $DEHG^{GMP}$ enriched in GMP with respect to HSC-MPP, iii) $DEHG^{HSC}$ enriched in HSC-MPP with respect to MEP, and iv) $DEHG^{MEP}$ enriched in MEP with respect to HSC-MPP. NES: normalized enrichment score, ES: enrichment score, q = FDR q value from GSEA. **C)** Scoring cell-specific DEHGs along a i) single-cell expression map reveals localized enrichment of expression: ii) $DEHG^{HSC}$, iii) $DEHG^{GMP}$, and iv) $DEHG^{MEP}$.

Figure 5. ***Heptad transcription factors at promoters and distal regulators of genes crucial for myeloid and erythroid cell development.*** **A)** Genes associated with myeloid development. *Left: k*-means clustered heatmaps of TF binding intensity at promoters and distal regulatory regions. Profile plots show normalised signal for each TF in each cell type at the regions depicted in the heatmap. *Right:* z-score normalised heatmaps of RNA-seq counts (GSE75384) for the corresponding gene in each cell type. White rows are genes with no expression values in the dataset. **B)** Genes associated with erythroid development. *Left: k*-means clustered heatmaps of TF binding intensity at promoters and distal regulatory regions. Profile plots show normalised signal for each TF in each cell type at the regions depicted in

the heatmap. *Right:* z-score normalised heatmaps of RNA-seq counts (GSE75384) for the corresponding gene in each cell type.

Figure 6. ***Regulatory regions with cell-type-specific heptad occupancy have distinct epigenetic features.*** **A)** A UMAP depicting the result of clustering 85,100 accessible regions in HSPCs annotated with ChIPmentation/ChIP-seq signal strengths using the Louvain algorithm. **B)** Individual violin plots of log normalised signal derived from ATAC, 3 histone marks (H3K27ac, H3K4me3, and H3K27me3), and CTCF – accompanied by a bar plot showing the number of regions in each cluster. Inter-cluster signal variability allows annotation of individual clusters based on their regulatory potential. **C)** UMAPs overlaid with ChromHMM annotation of 85,100 individual regions show striking similarity with **B**. **D)** UMAPs coloured based on log2 fold change of binding of the heptad transcription factors in pairwise comparisons between GMP and MEP. MEP- and GMP-specific enrichment of TF binding is identified, and borders demarcated by dashed lines: black (enriched in MEP) or grey (enriched in GMP). **E)** Signal of PU.1 in dendritic cells (DC) (GSE58864) across the clustered regions. PU.1 signal enrichment in dendritic cells mirrors heptad factor enrichment patterns in GMP. **F)** Signal of GATA1 in proerythroblasts (ProE) (GSE36985) across the clustered regions. GATA1 signal enrichment in proerythroblasts mirrors heptad factor enrichment patterns at these regions in MEP.

Figure 7. ***Cell-type specificity of regulatory elements is encoded in the underlying motif composition***. **A)** (i) UMAP representation of ATAC-seq regions in $CD34^{+}$ cells (grey) with heptad TF bound HSC-MPP specific regions colored in purple. (ii) An XGBoost machine learning model was trained and tested with motif counts from a mixture of regions specified in (i) and background regions, to predict cell type with high accuracy. The ROC curve shows

the predictive performance of the constructed model to predict HSC-MPP specific regions. (iii) A beeswarm plot depicting the top 12 representative motifs in HSC-MPP specific regions - ranked based on their absolute importance in contributing to the predictive model. Each row shows the motif (and canonical TF family if known), and the corresponding SHAP (**SH**aply **A**dditive ex**P**lanations) values for the cell type in question (right) and the others (left). The feature count indicates the normalised motif counts with a range of 0-1. **B)** (i) UMAP representation of ATAC-seq regions in $CD34^+$ cells (grey) with heptad TF bound GMP specific regions colored in green. (ii) ROC curve showing the performance of the model to predict GMP specific regions. (iii) A beeswarm plot depicting the top 12 representative motifs in GMP specific regions - ranked based on their absolute importance in contributing to the predictive model. **C)** (i) UMAP representation of ATAC-seq regions in $CD34^+$ cells (grey) with heptad TF bound MEP specific regions colored in orange. (ii) ROC curve showing the performance of the model to predict MEP specific regions. (iii) A beeswarm plot depicting the top 12 representative motifs in MEP specific regions - ranked based on their absolute importance in contributing to the predictive model.

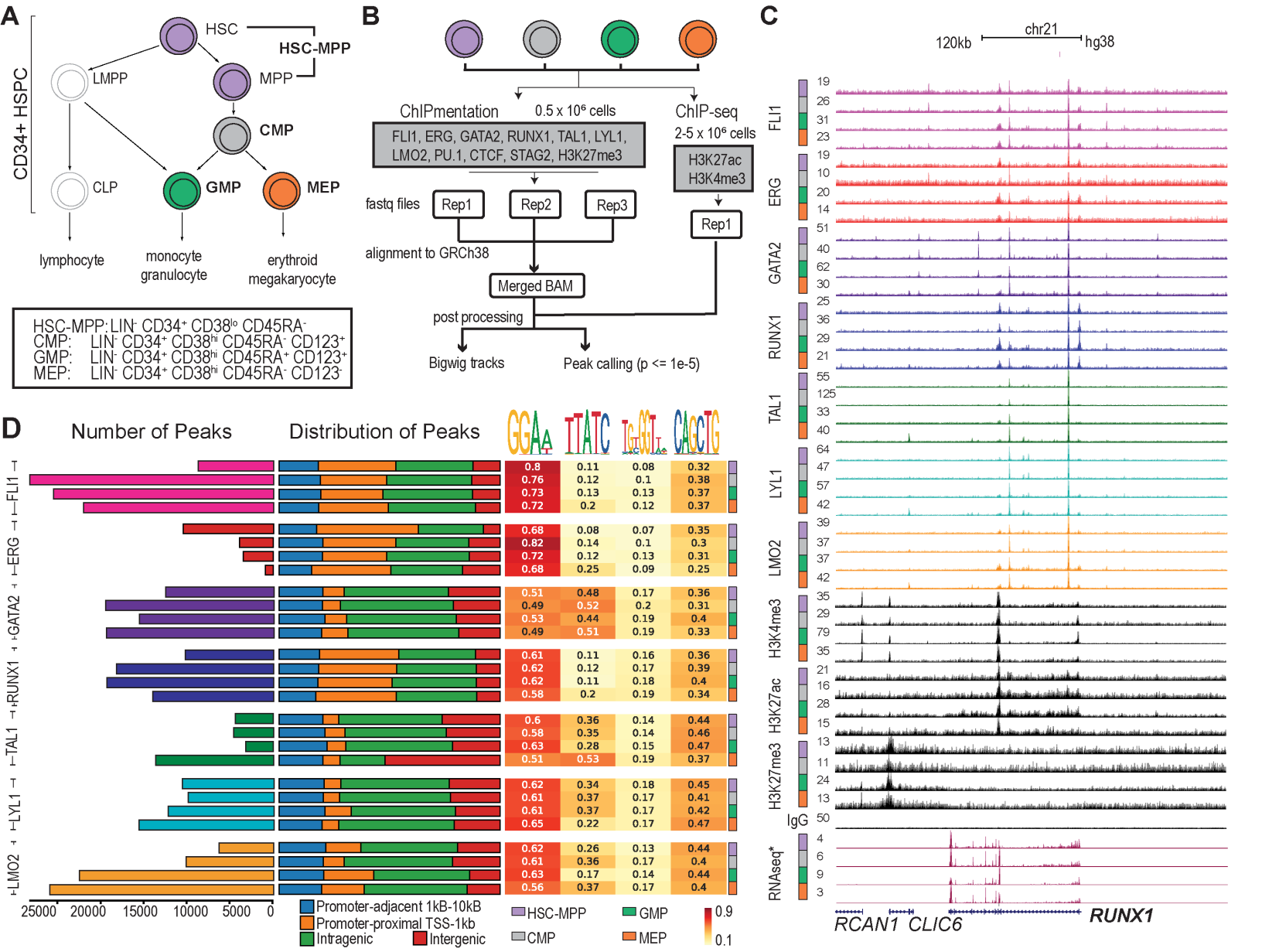

A
CD34+ HSPC
HSC
HSC-MPP
LMPP
MPP
CMP
CLP
GMP
MEP
lymphocyte
monocyte granulocyte
erythroid megakaryocyte
HSC-MPP: LIN- CD34+ CD38lo CD45RA-
CMP: LIN- CD34+ CD38hi CD45RA- CD123+
GMP: LIN- CD34+ CD38hi CD45RA+ CD123+
MEP: LIN- CD34+ CD38hi CD45RA- CD123-
B
ChIPmentation
0.5 x 10⁶ cells
FLI1, ERG, GATA2, RUNX1, TAL1, LYL1, LMO2, PU.1, CTCF, STAG2, H3K27me3
ChIP-seq
2-5 x 10⁶ cells
H3K27ac H3K4me3
fastq files
Rep1
Rep2
Rep3
Rep1
alignment to GRCh38
Merged BAM
post processing
Bigwig tracks
Peak calling (p <= 1e-5)
C
120kb
chr21
hg38
FLI1
ERG
GATA2
RUNX1
TAL1
LYL1
LMO2
H3K4me3
H3K27ac
H3K27me3
IgG
RNAseq*
RCAN1
CLIC6
RUNX1
D
Number of Peaks
Distribution of Peaks
FLI1
ERG
GATA2
RUNX1
TAL1
LYL1
LMO2
25000
20000
15000
10000
5000
0
Promoter-adjacent 1kB-10kB
Promoter-proximal TSS-1kb
Intragenic
Intergenic
HSC-MPP
CMP
GMP
MEP
0.9
0.1

Figure 1

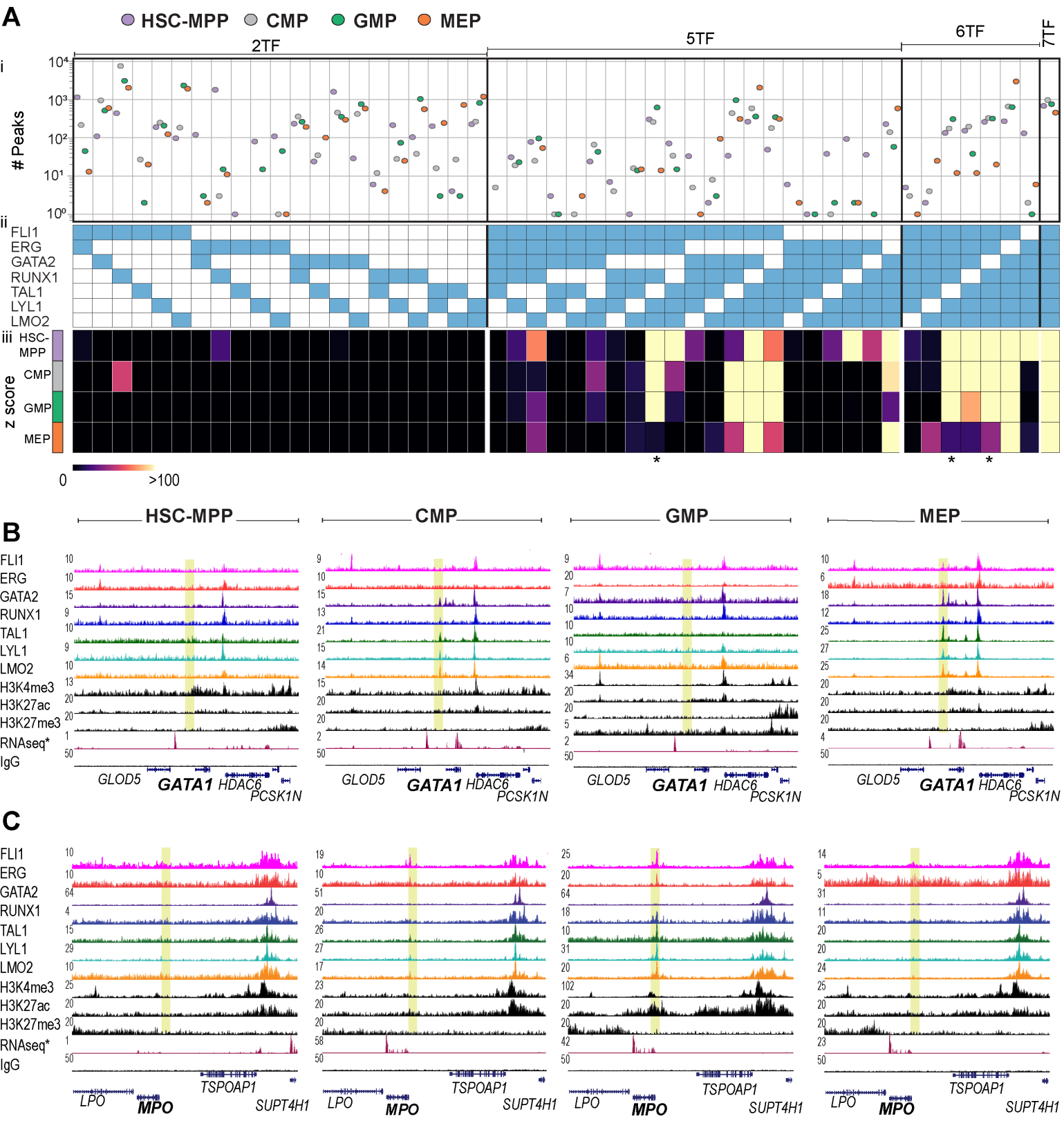

A
HSC-MPP
CMP
GMP
MEP
2TF
5TF
6TF
7TF
i
Peaks
ii
FLI1
ERG
GATA2
RUNX1
TAL1
LYL1
LMO2
iii
HSC-MPP
CMP
GMP
MEP
z score
0
>100
B
HSC-MPP
CMP
GMP
MEP
FLI1
ERG
GATA2
RUNX1
TAL1
LYL1
LMO2
H3K4me3
H3K27ac
H3K27me3
RNAseq*
IgG
GLOD5
GATA1
HDAC6
PCSK1N
C
TSPOAP1
LPO
MPO
SUPT4H1

Figure 2

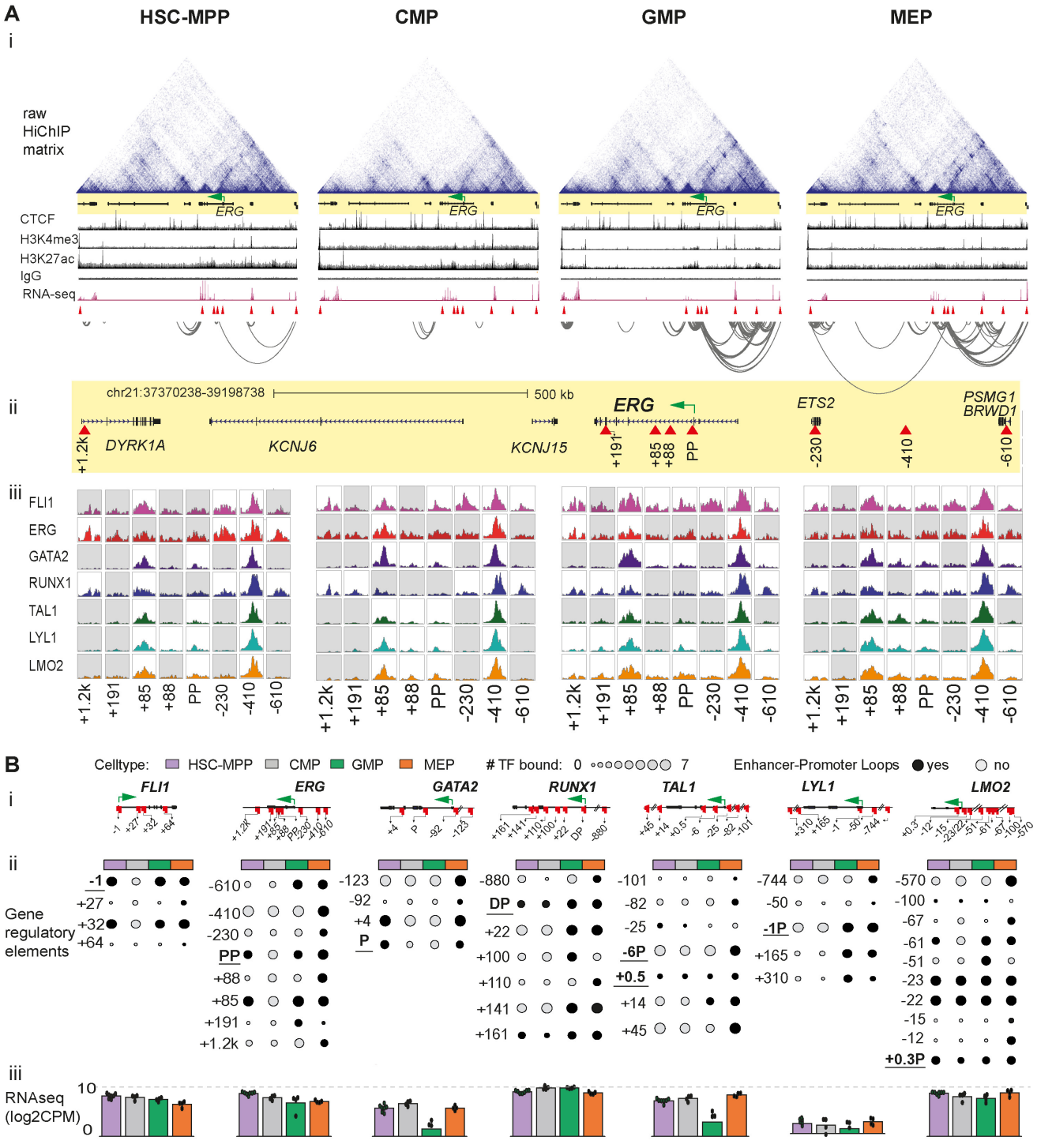

A
HSC-MPP
CMP
GMP
MEP
i
raw HiChIP matrix
ERG
CTCF
H3K4me3
H3K27ac
IgG
RNA-seq
ii
chr21:37370238-39198738
500 kb
DYRK1A
KCNJ6
KCNJ15
ERG
ETS2
PSMG1
BRWD1
+1.2k
+191
+85
+88
PP
-230
-410
-610
iii
FLI1
ERG
GATA2
RUNX1
TAL1
LYL1
LMO2
B
Celltype: HSC-MPP CMP GMP MEP
TF bound: 0 7
Enhancer-Promoter Loops yes no
FLI1
ERG
GATA2
RUNX1
TAL1
LYL1
LMO2
Gene regulatory elements
RNAseq (log2CPM)
10
0

Figure 3

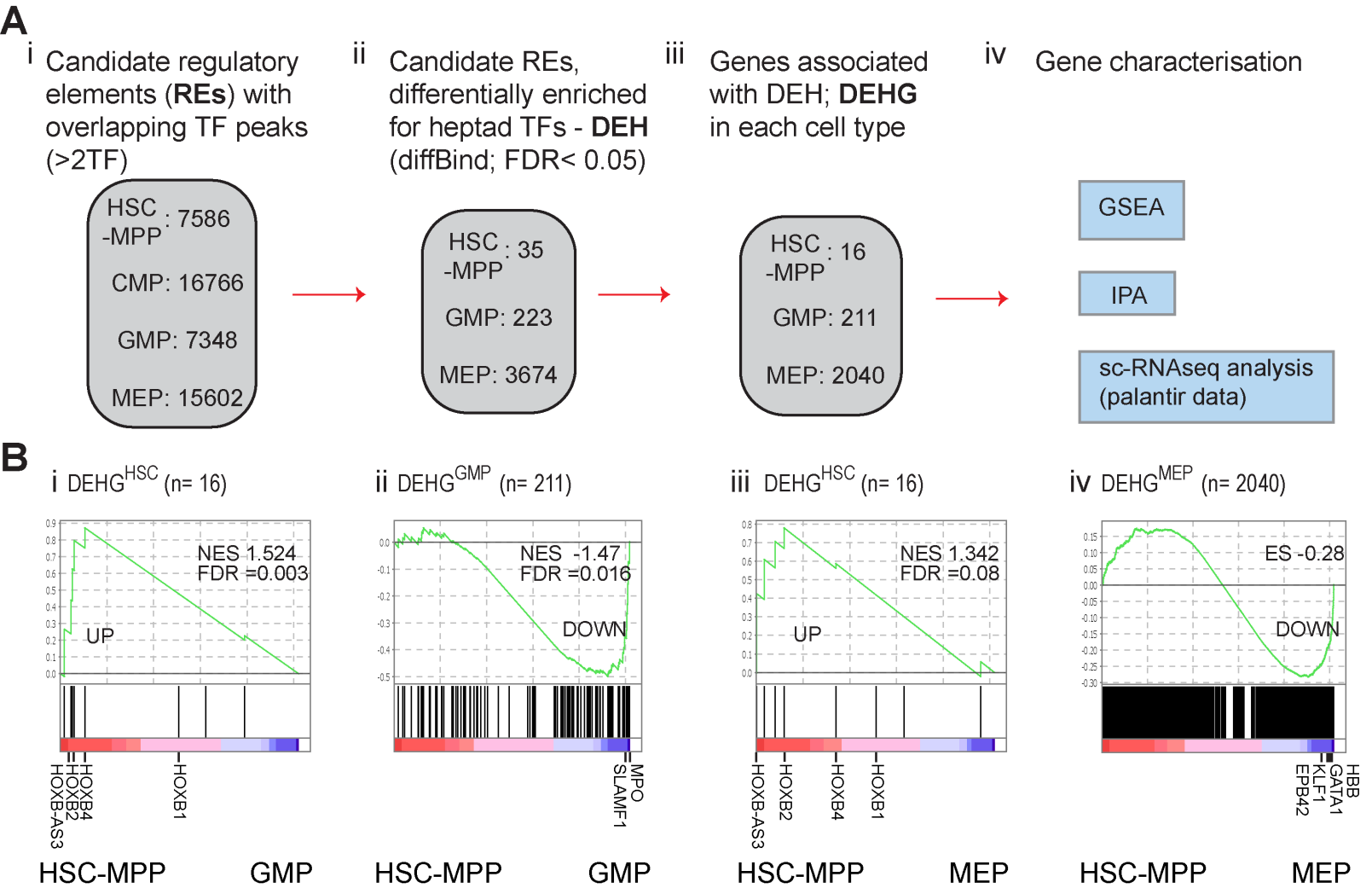

A
i Candidate regulatory elements (REs) with overlapping TF peaks (>2TF)
HSC -MPP: 7586
CMP: 16766
GMP: 7348
MEP: 15602
ii Candidate REs, differentially enriched for heptad TFs - DEH (diffBind; FDR< 0.05)
HSC -MPP: 35
GMP: 223
MEP: 3674
iii Genes associated with DEH; DEHG in each cell type
HSC -MPP: 16
GMP: 211
MEP: 2040
iv Gene characterisation
GSEA
IPA
sc-RNAseq analysis (palantir data)
B
i DEHG^HSC (n= 16)
NES 1.524
FDR =0.003
UP
HOXB-AS3
HOXB2
HOXB4
HOXB1
HSC-MPP
GMP
ii DEHG^GMP (n= 211)
NES -1.47
FDR =0.016
DOWN
SLAMF1
MPO
HSC-MPP
GMP
iii DEHG^HSC (n= 16)
NES 1.342
FDR =0.08
UP
HOXB-AS3
HOXB2
HOXB4
HOXB1
HSC-MPP
MEP
iv DEHG^MEP (n= 2040)
ES -0.28
DOWN
HBB
GATA1
KLF1
EPB42
HSC-MPP
MEP
C
i
HSC
Precursors
Ery precursors
Ery
Mega
Mono precursors
Mono
DC
CLP
ii DEHG^HSC
iii DEHG^GMP
iv DEHG^MEP

Figure 4

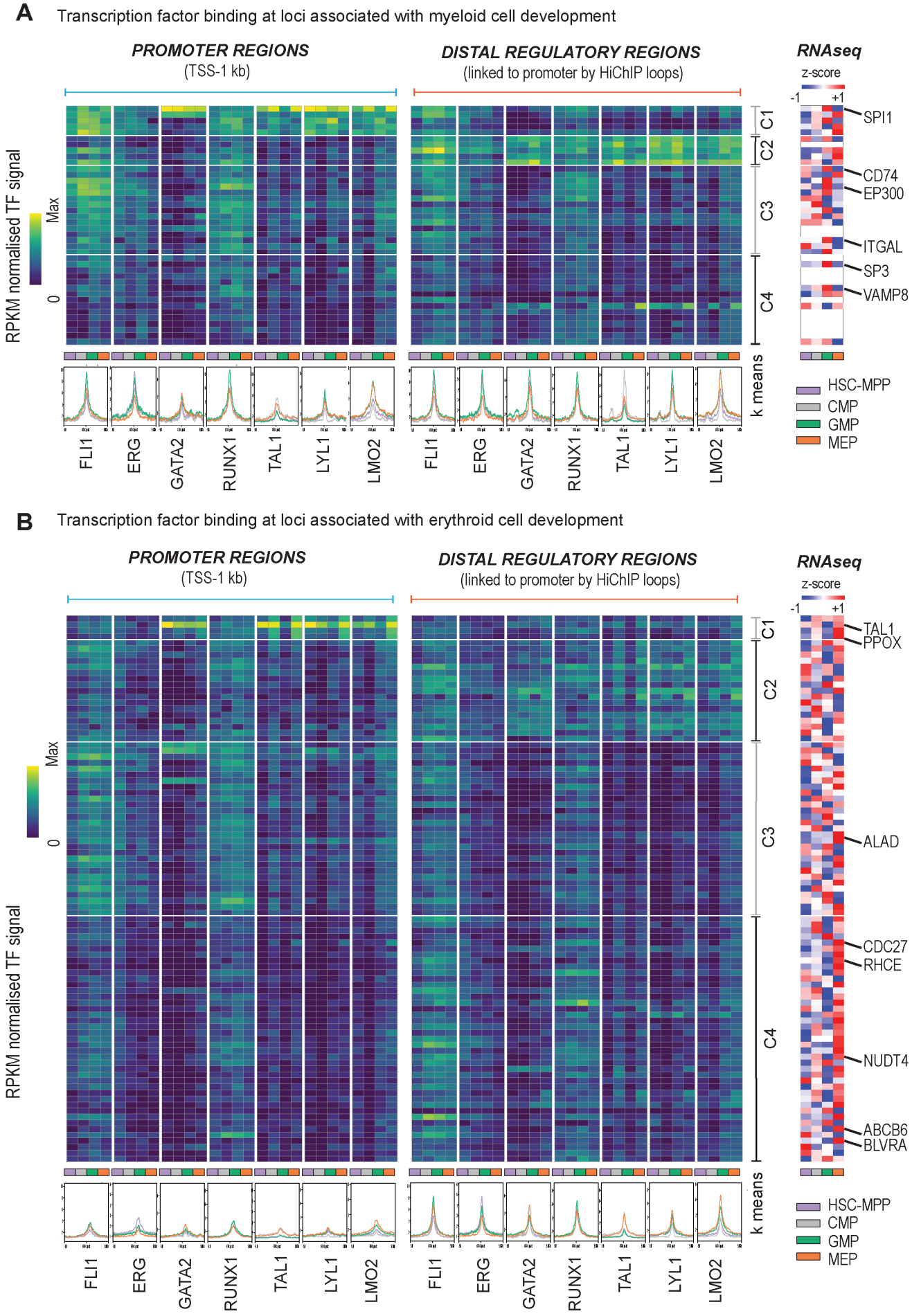

A
Transcription factor binding at loci associated with myeloid cell development
PROMOTER REGIONS
(TSS-1 kb)
DISTAL REGULATORY REGIONS
(linked to promoter by HiChIP loops)
RNAseq
z-score
-1
+1
RPKM normalised TF signal
Max
0
C1
C2
C3
C4
k means
SPI1
CD74
EP300
ITGAL
SP3
VAMP8
HSC-MPP
CMP
GMP
MEP
FLI1
ERG
GATA2
RUNX1
TAL1
LYL1
LMO2
B
Transcription factor binding at loci associated with erythroid cell development
PROMOTER REGIONS
(TSS-1 kb)
DISTAL REGULATORY REGIONS
(linked to promoter by HiChIP loops)
RNAseq
z-score
-1
+1
RPKM normalised TF signal
Max
0
C1
C2
C3
C4
k means
TAL1
PPOX
ALAD
CDC27
RHCE
NUDT4
ABCB6
BLVRA
HSC-MPP
CMP
GMP
MEP
FLI1
ERG
GATA2
RUNX1
TAL1
LYL1
LMO2

Figure 5

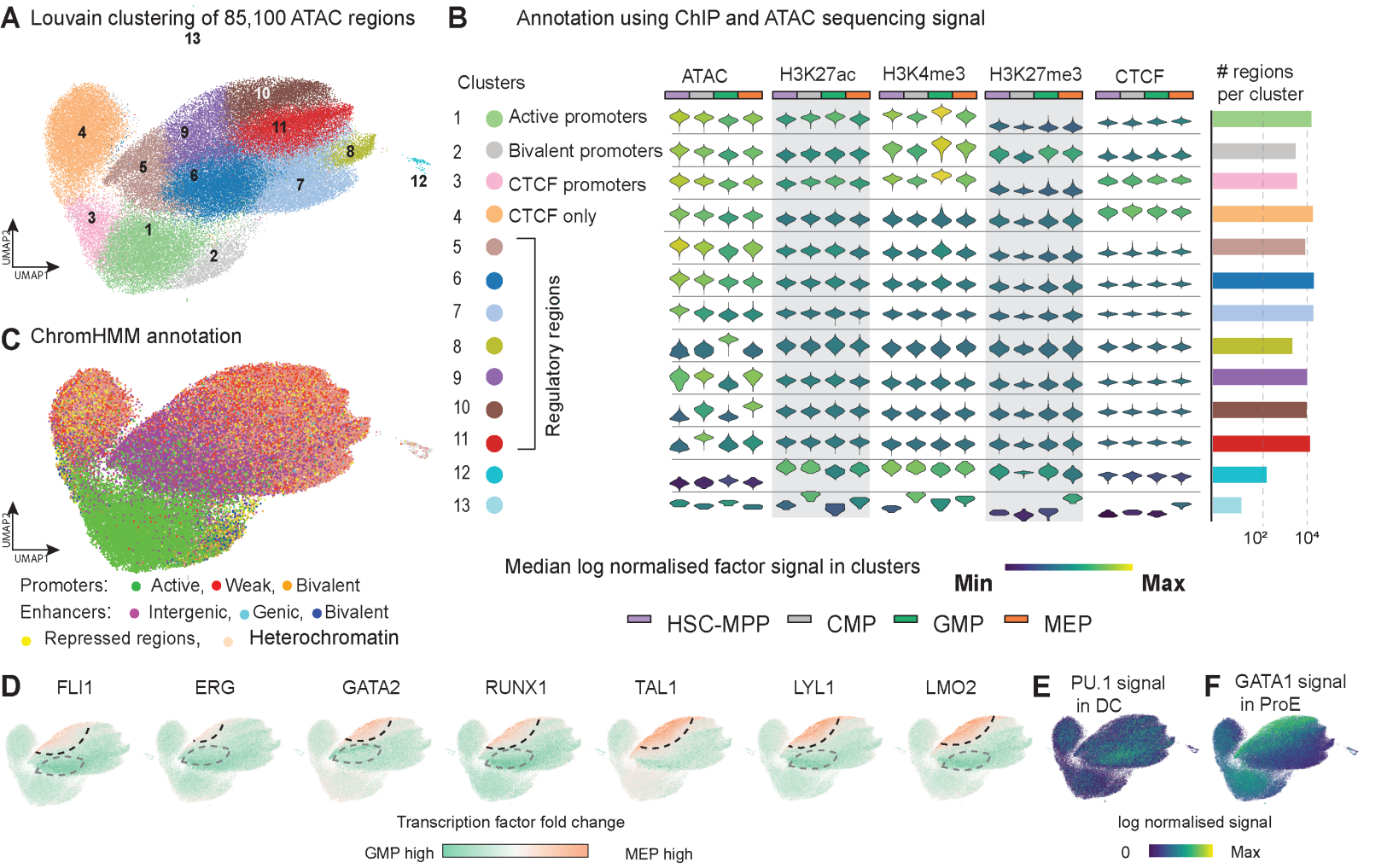

A Louvain clustering of 85,100 ATAC regions
13
10
4
9
11
8
5
6
7
12
3
1
2
UMAP2
UMAP1
B Annotation using ChIP and ATAC sequencing signal
Clusters
1 Active promoters
2 Bivalent promoters
3 CTCF promoters
4 CTCF only
5
6
7
8
9
10
11
12
13
Regulatory regions
ATAC
H3K27ac
H3K4me3
H3K27me3
CTCF
regions per cluster
10²
10⁴
C ChromHMM annotation
Promoters: Active, Weak, Bivalent
Enhancers: Intergenic, Genic, Bivalent
Repressed regions, Heterochromatin
Median log normalised factor signal in clusters
Min
Max
HSC-MPP
CMP
GMP
MEP
D
FLI1
ERG
GATA2
RUNX1
TAL1
LYL1
LMO2
Transcription factor fold change
GMP high
MEP high
E PU.1 signal in DC
F GATA1 signal in ProE
log normalised signal
0
Max

Figure 6

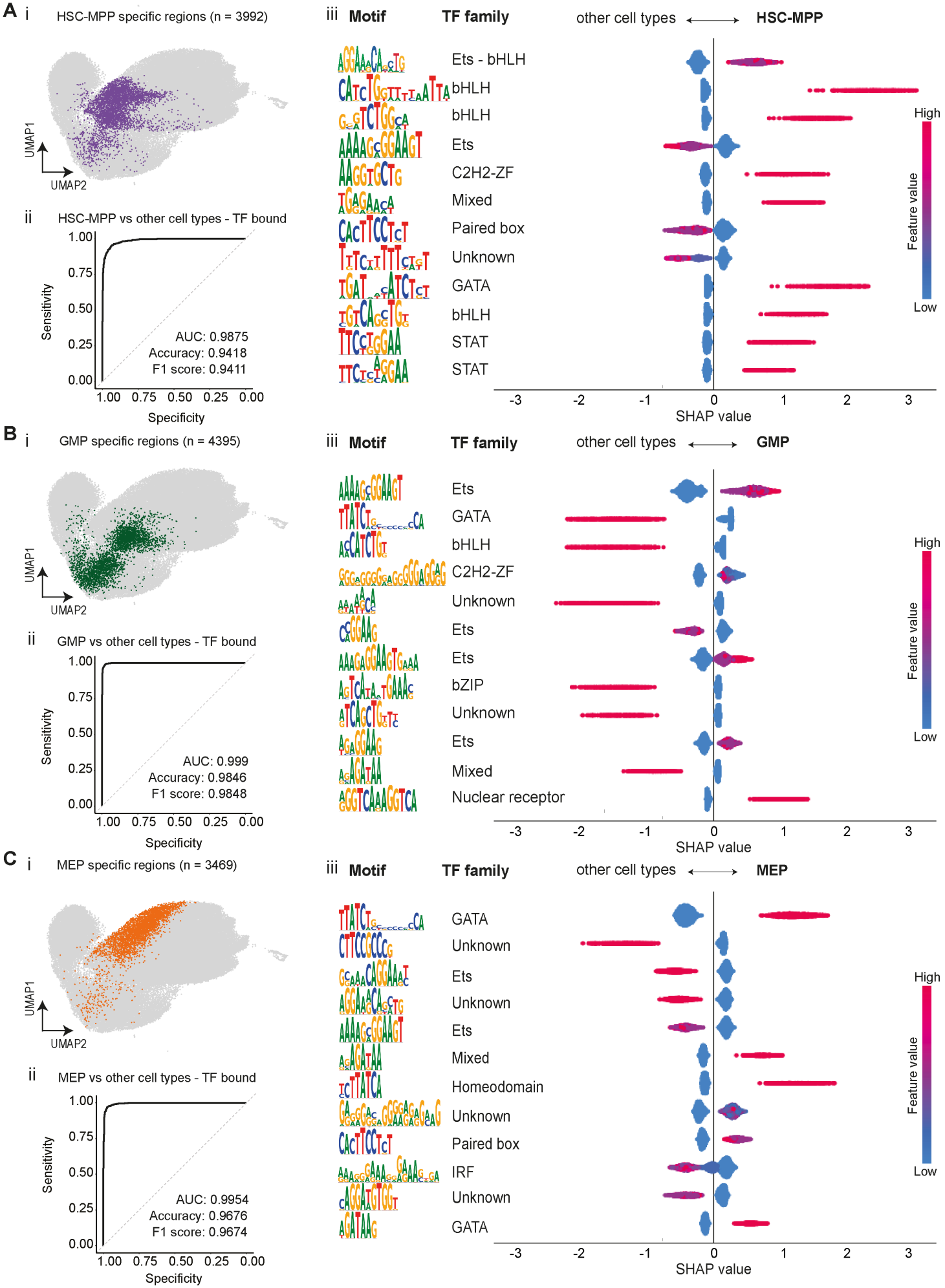

A
i
HSC-MPP specific regions (n = 3992)
UMAP1
UMAP2
ii
HSC-MPP vs other cell types - TF bound
Sensitivity
Specificity
1.00
0.75
0.50
0.25
0.00
AUC: 0.9875
Accuracy: 0.9418
F1 score: 0.9411
iii
Motif
TF family
other cell types
HSC-MPP
Ets - bHLH
bHLH
bHLH
Ets
C2H2-ZF
Mixed
Paired box
Unknown
GATA
bHLH
STAT
STAT
High
Feature value
Low
-3
-2
-1
0
1
2
3
SHAP value
B
i
GMP specific regions (n = 4395)
UMAP1
UMAP2
ii
GMP vs other cell types - TF bound
Sensitivity
Specificity
AUC: 0.999
Accuracy: 0.9846
F1 score: 0.9848
iii
Motif
TF family
other cell types
GMP
Ets
GATA
bHLH
C2H2-ZF
Unknown
Ets
Ets
bZIP
Unknown
Ets
Mixed
Nuclear receptor
High
Feature value
Low
SHAP value
C
i
MEP specific regions (n = 3469)
UMAP1
UMAP2
ii
MEP vs other cell types - TF bound
Sensitivity
Specificity
AUC: 0.9954
Accuracy: 0.9676
F1 score: 0.9674
iii
Motif
TF family
other cell types
MEP
GATA
Unknown
Ets
Unknown
Ets
Mixed
Homeodomain
Unknown
Paired box
IRF
Unknown
GATA
High
Feature value
Low
SHAP value

Figure 7

Subramanian et al Supplemental Tables and Figures, and Methods

List of supplemental Tables:

Table S1. Donor information

Table S2. ChIP data mapping summary

Table S3. Significant HiChIP loops

Table S4. List of heptad *cis* Regulatory Elements (CREs)

Table S5. Gene lists used in Figure 4 and Figure 5

Table S6. Complete statistics for Supplemental Figure 9

Table S7. Model performance

Table S8. Motifs

Table S9. Key reagents used in this work

Table S10. Indexing primers

Supplemental Figure Titles and Legends

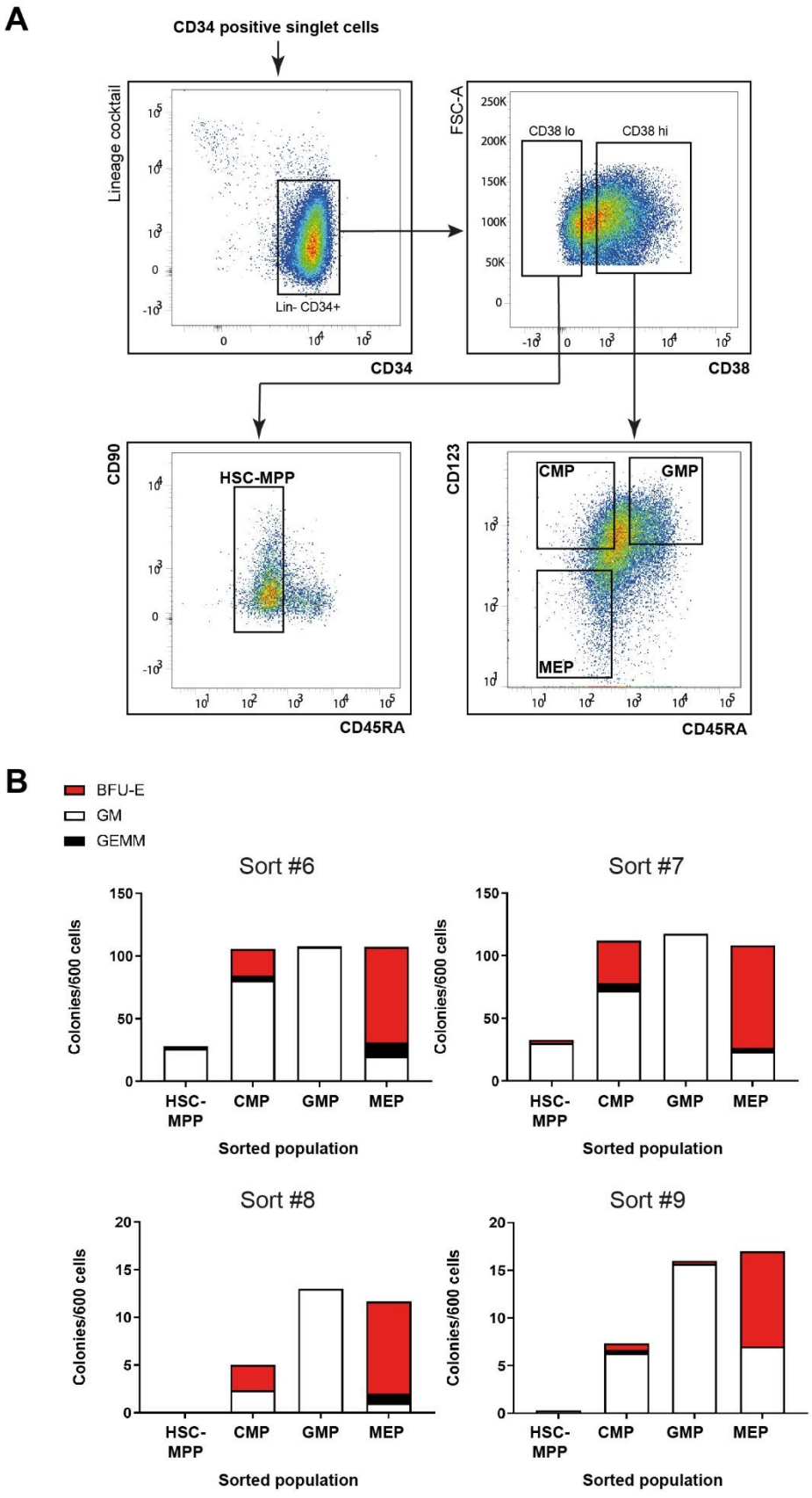

A
CD34 positive singlet cells
Lineage cocktail
Lin- CD34+
CD34
FSC-A
CD38 lo
CD38 hi
CD38
CD90
HSC-MPP
CD45RA
CD123
CMP
GMP
MEP
CD45RA
B
BFU-E
GM
GEMM
Sort #6
Sort #7
Sort #8
Sort #9
Colonies/600 cells
HSC-MPP
CMP
GMP
MEP
Sorted population

Figure S1

Figure S1. ***Workflow for deriving HSPC fractions and verifying their identity***. **A)** Gating strategy and representative sort gates for isolation of HSC-MPP, CMP, GMP, and MEP. **B)** Representative CFU assays using sorted HSC-MPP, CMP, GMP, and MEP populations.

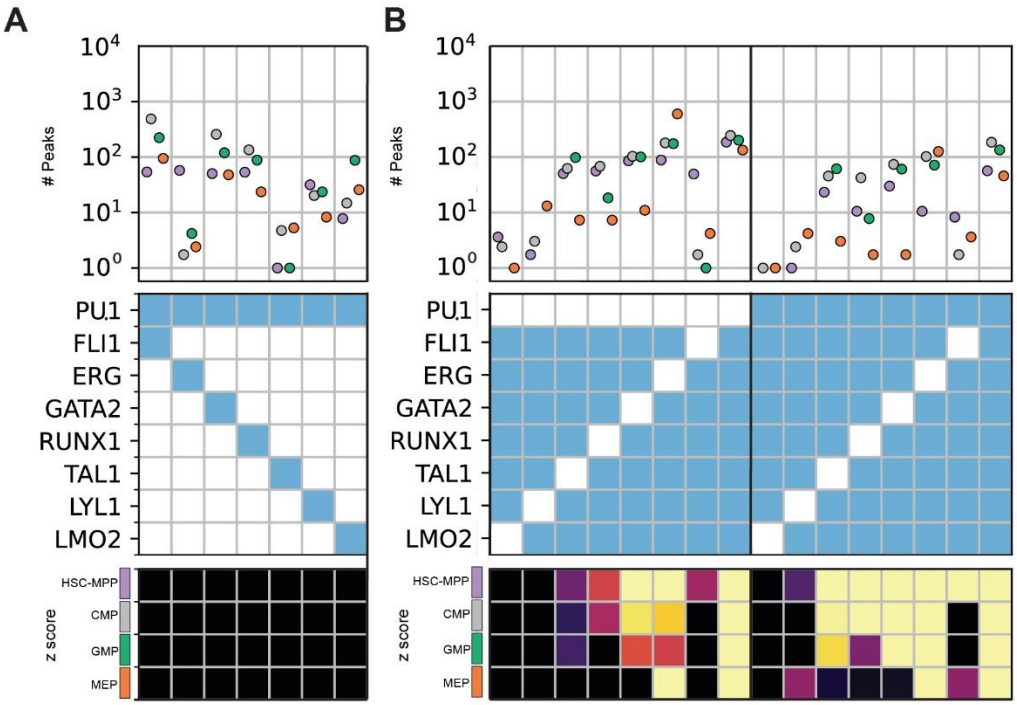


Figure S2

Figure S2. ***Combinatorial binding of heptad transcription factors and PU.1.*** **(A-B)** composite graphs with three components; *(upper)* number of combinatorial binding peaks identified in the four cell types, for *(middle)* combinations of heptad factors and PU.1 and *(lower)* heatmap showing z-scores for the combinations presented above. **A)** 2 TF combinations including PU.1. **B)** 6, 7, 8 TF combinations of heptad TFs without and with PU.1.

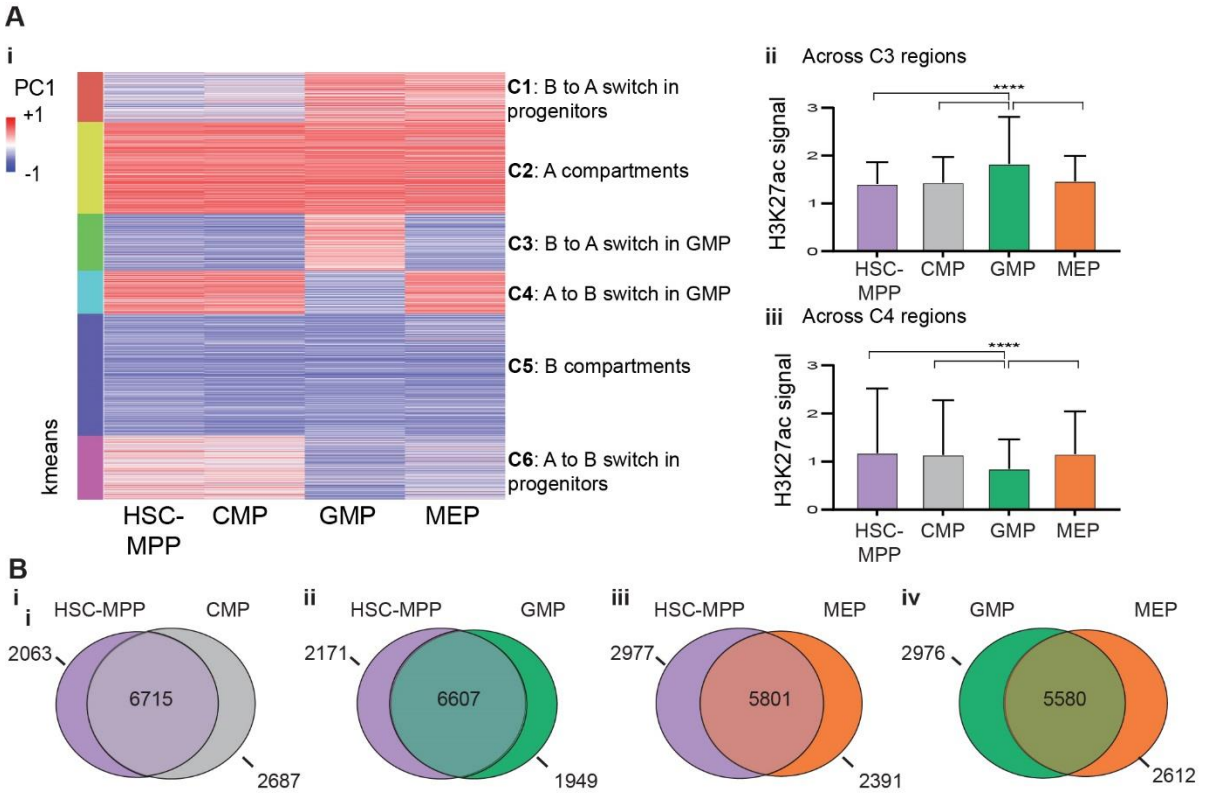


Figure S3

Figure S3. ***Compartment switching and conserved TADs along blood stem cell differentiation***. **A)** i) A *k*-means clustered heatmap of the PC1 values showing compartmental switches taking place between the four HSPC populations. Average H3K27ac signal in ii) Cluster C3 regions, and in iii) Cluster C4 regions among the four cell types. Significance scores were calculated using pairwise t tests ($p < 0.0001$). **B)** Pairwise comparisons of topological domain (TAD) boundaries between i) HSC-MPP and CMP, ii) HSC-MPP and GMP, iii) HSC-MPP and MEP, and iv) GMP and MEP (domains identified using HOMER).

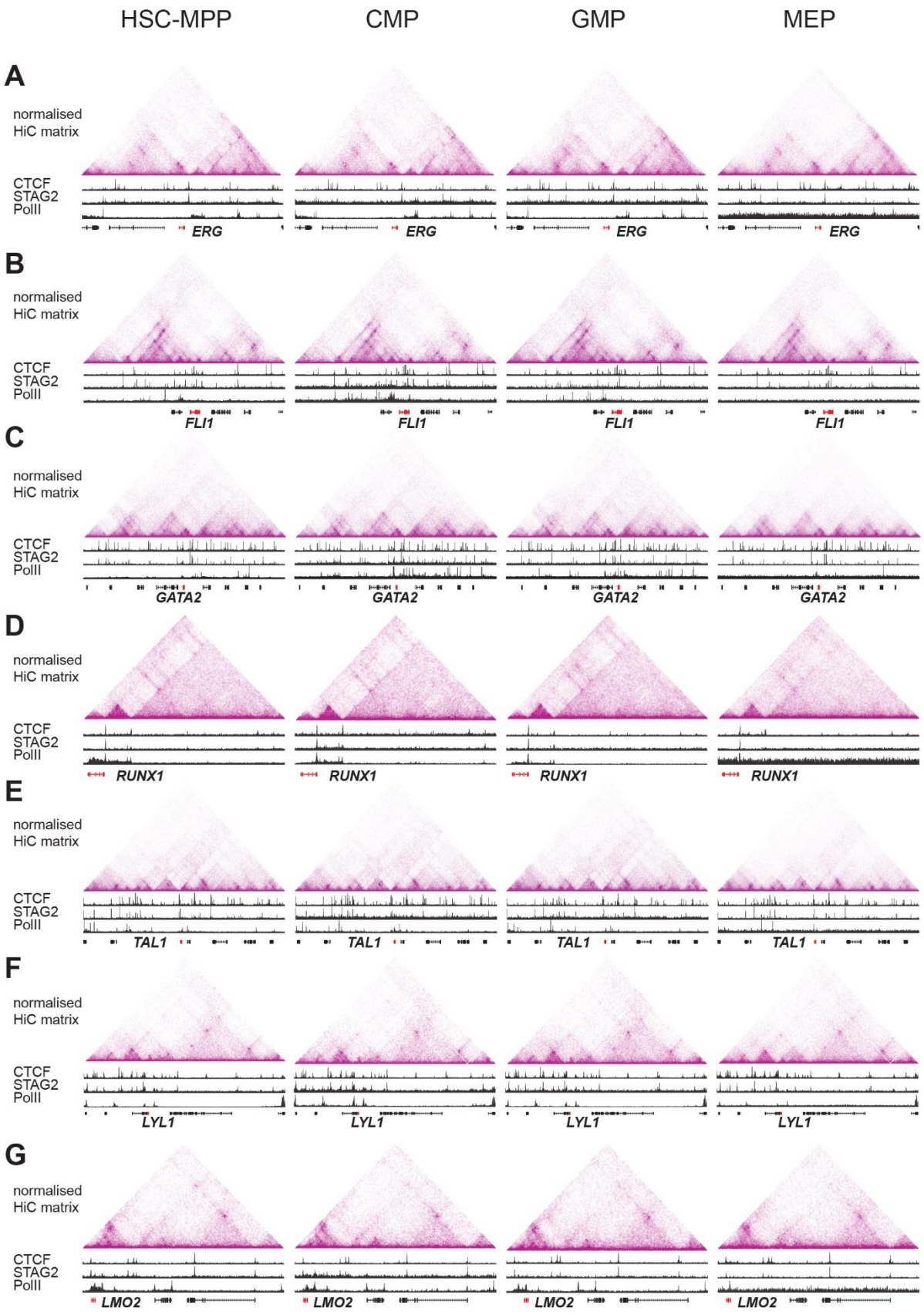

HSC-MPP
CMP
GMP
MEP
A
normalised HiC matrix
CTCF
STAG2
PolII
ERG
B
FLI1
C
GATA2
D
RUNX1
E
TAL1
F
LYL1
G
LMO2

Figure S4. ***The genome architecture at heptad gene regulatory loci is conserved across HSPC subsets.*** Normalised HiC contact matrices at 10 kb resolution located at individual heptad genes' regulatory loci – **A)** *FLI1* locus (GRCh38 chr11:128511084-128978507), **B)** *ERG* locus (GRCh38 chr21:37370238-39198738), **C)** *GATA2* locus (GRCh38 chr3:128262936-128761435), **D)** *RUNX1* locus (GRCh38 chr21:34758869-36011624), **E)** *TAL1* locus (GRCh38 chr1:47168881-47340728), **F)** *LYL1* locus (GRCh38 chr19:12787014-13852204), and **G)** *LMO2* locus (GRCh38 chr11:33831641-34445745), in HSC-MPP, CMP, GMP, and MEP respectively. Accompanying each triangular plot are ChIP-seq tracks showing normalized CTCF, STAG2, and PolII signal.

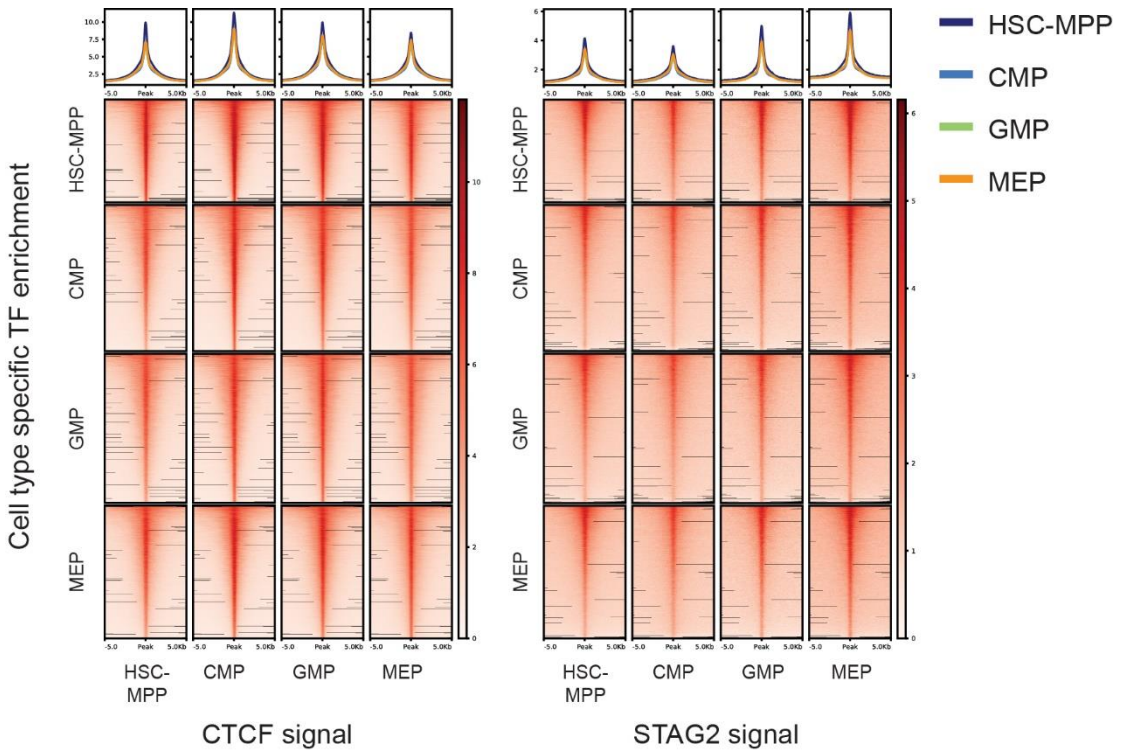


Figure S5

Figure S5. ***CTCF and STAG2 occupancy across cell types.*** Density plots showing CTCF (*left*) and STAG2 (*right*) signal in each cell type at CTCF peak regions (identified by macs2; p <1e-5) in HSC-MPP, CMP, GMP, MEP. CTCF and STAG2 occupancy at CTCF-bound sites shows minimal variation between cell types.

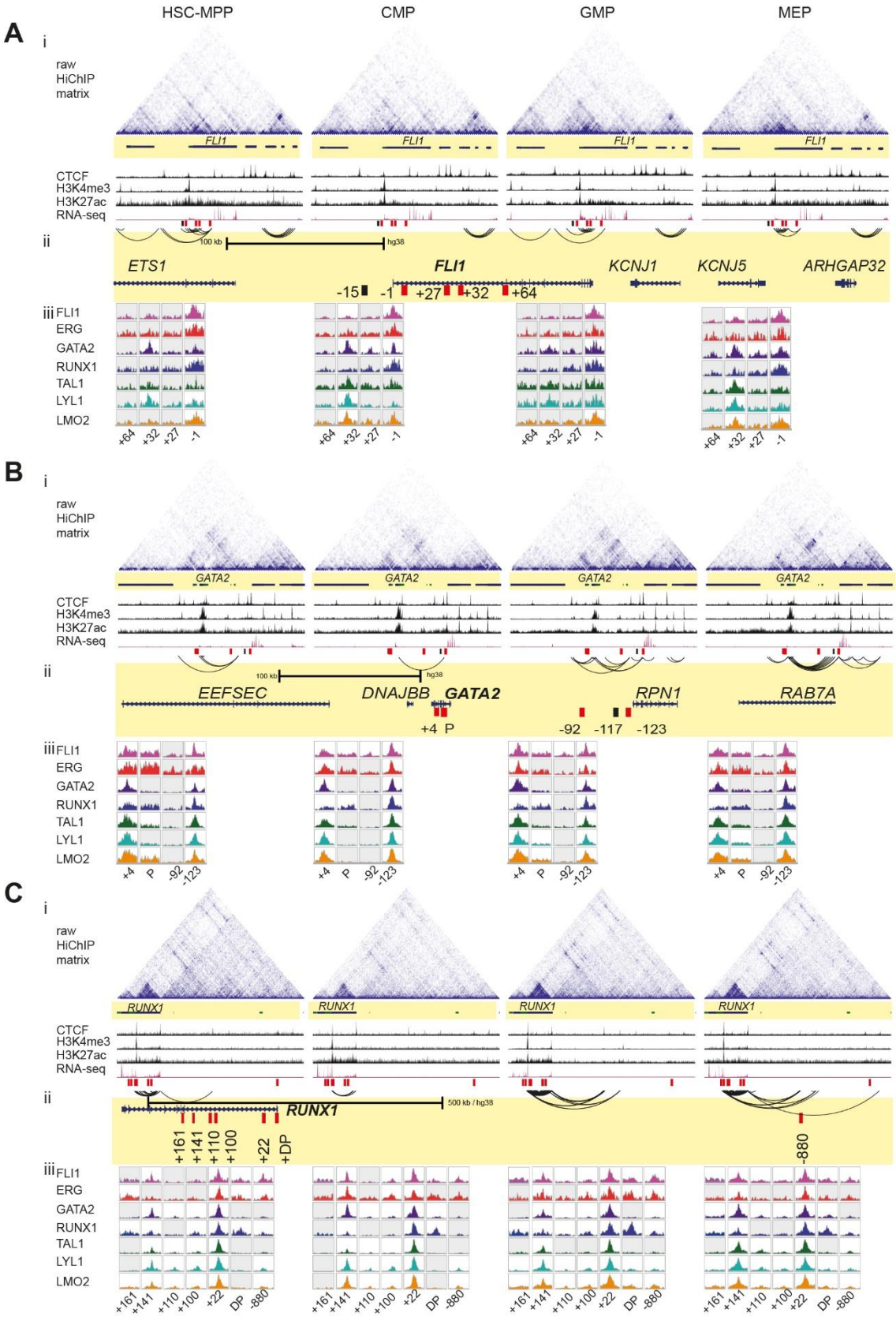

A
B
C
HSC-MPP
CMP
GMP
MEP
raw HiChIP matrix
CTCF
H3K4me3
H3K27ac
RNA-seq
FLI1
GATA2
RUNX1
ETS1
KCNJ1
KCNJ5
ARHGAP32
EEFSEC
DNAJBB
RPN1
RAB7A
100 kb
hg38
500 kb / hg38
-15 -1 +27 +32 +64
+4 P -92 -117 -123
+161 +141 +110 +100 +22 +DP -880
ERG
TAL1
LYL1
LMO2

Figure S6

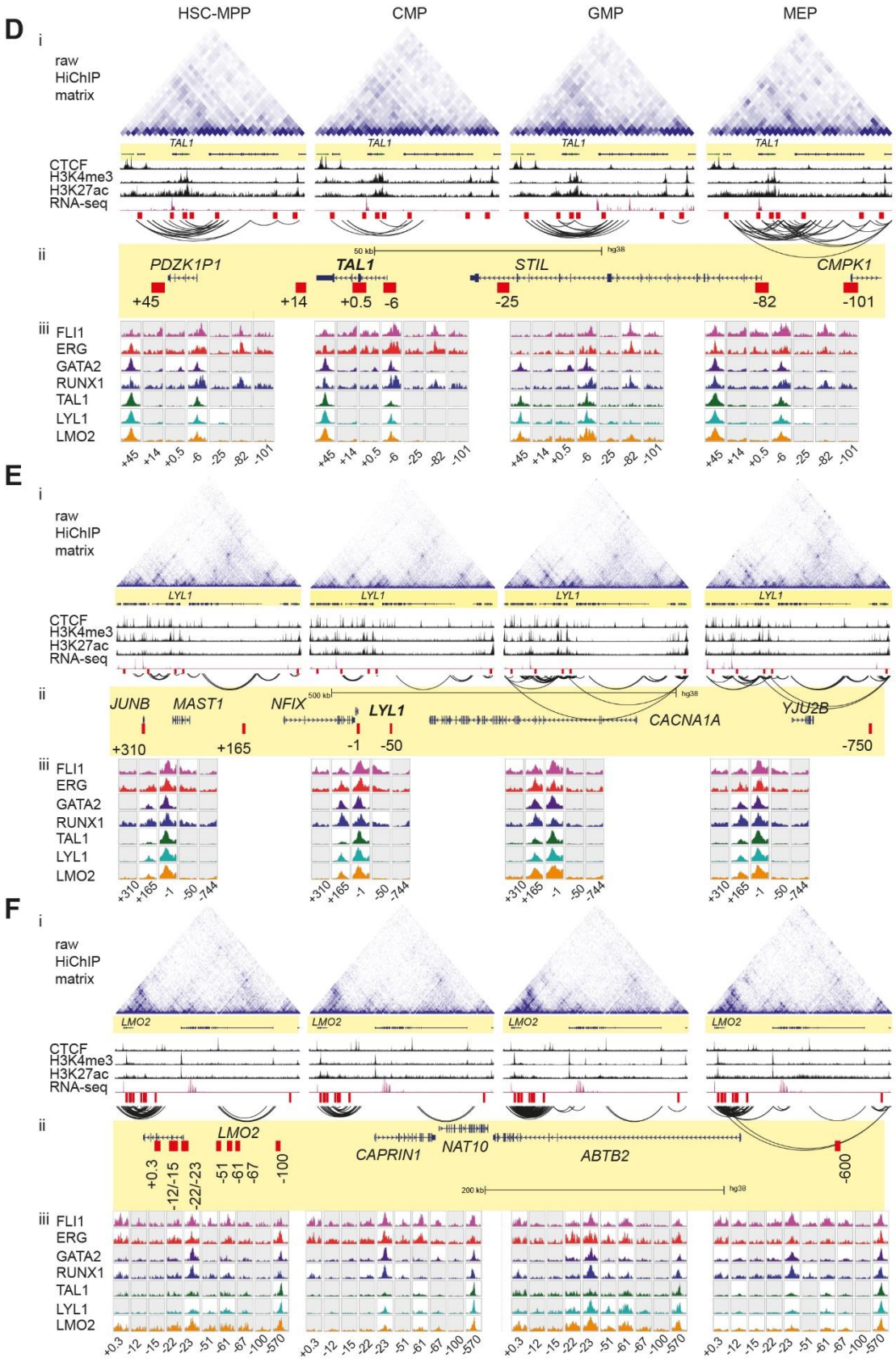

D
HSC-MPP
CMP
GMP
MEP
i
raw
HiChIP
matrix
TAL1
CTCF
H3K4me3
H3K27ac
RNA-seq
ii
PDZK1P1
TAL1
STIL
CMPK1
+45
+14
+0.5
-6
-25
-82
-101
iii
FLI1
ERG
GATA2
RUNX1
TAL1
LYL1
LMO2
E
LYL1
JUNB
MAST1
NFIX
CACNA1A
YJU2B
+310
+165
-1
-50
-750
F
LMO2
CAPRIN1
NAT10
ABTB2
+0.3
-12/-15
-22/-23
-51
-61
-67
-100
-600

Figure S6

Figure S6. ***H3K27ac HiChIP identifies cell-type specific interactions between heptad gene promoters and potential regulatory regions.*** **A)** i) Raw HiChIP contact matrix, CTCF, H3K4me3, H3K27ac, IgG, RNA-seq, and significant H3K27ac HiChIP interactions (FDR ≤ 0.01) at the *FLI1* locus (GRCh38 chr11:128511084-128978507). ii) Magnified view of the *FLI1* locus, with potential regulators looping to the promoter shown in red and potential regulators engaged in indirect regulatory activities shown in black. iii) FLI1, ERG, GATA2, RUNX1, TAL1, LYL1, and LMO2 peaks at the regulatory regions defined in ii, are shown. **B)** i) Raw HiChIP contact matrix, CTCF, H3K4me3, H3K27ac, IgG, RNA-seq, and significant H3K27ac HiChIP interactions (FDR ≤ 0.01) at the *GATA2* locus (chr3:128,262,936-128,761,435). ii) Magnified view of the *GATA2* locus, with potential regulators looping to the promoter shown in red and potential regulators engaged in indirect regulatory activities shown in black. iii) FLI1, ERG, GATA2, RUNX1, TAL1, LYL1, and LMO2 peaks at the regulatory regions defined in ii, are shown. **C)** i) Raw HiChIP contact matrix, CTCF, H3K4me3, H3K27ac, IgG, RNA-seq, and significant H3K27ac HiChIP interactions (FDR ≤ 0.01) at the *RUNX1* locus (GRCh38 chr21:34758869-36011624). ii) Magnified view of the *RUNX11* locus, with potential regulators looping to the promoter shown in red. iii) FLI1, ERG, GATA2, RUNX1, TAL1, LYL1, and LMO2 peaks at the regulatory regions defined in ii, are shown. **D)** i) Raw HiChIP contact matrix, CTCF, H3K4me3, H3K27ac, IgG, RNA-seq, and significant H3K27ac HiChIP interactions (FDR ≤ 0.01) at the *TAL11* locus (GRCh38 chr1:47,168,881-47,340,728). ii) Magnified view of the *TAL11* locus, with potential regulators looping to the promoter shown in red. iii) FLI1, ERG, GATA2, RUNX1, TAL1, LYL1, and LMO2 peaks at the regulatory regions defined in ii, are shown. **E)** i) Raw HiChIP contact matrix, CTCF, H3K4me3, H3K27ac, IgG, RNA-seq, and significant H3K27ac HiChIP interactions (FDR ≤ 0.01) at the *LYL1* locus (GRCh38 chr19:12787014-13852204). ii) Magnified view of the *LYL1* locus, with potential regulators looping to the promoter shown in red. iii) FLI1, ERG, GATA2,

RUNX1, TAL1, LYL1, and LMO2 peaks at the regulatory regions defined in ii, are shown. **F)** i) Raw HiChIP contact matrix, CTCF, H3K4me3, H3K27ac, IgG, RNA-seq, and significant H3K27ac HiChIP interactions (FDR ≤ 0.01) at the *LMO2* (GRCh38 chr11:33831641-34445745). ii) Magnified view of the *LMO2* locus, with potential regulators looping to the promoter shown in red. iii) FLI1, ERG, GATA2, RUNX1, TAL1, LYL1, and LMO2 peaks at the regulatory regions defined in ii, are shown.

Only those HiChIP interactions where both interacting ends were found at the given locus are shown. In addition the ChIP-seq peaks shown are RPKM-normalised and white boxes indicate presence of a computationally called ChIP-seq peak at the specific region.

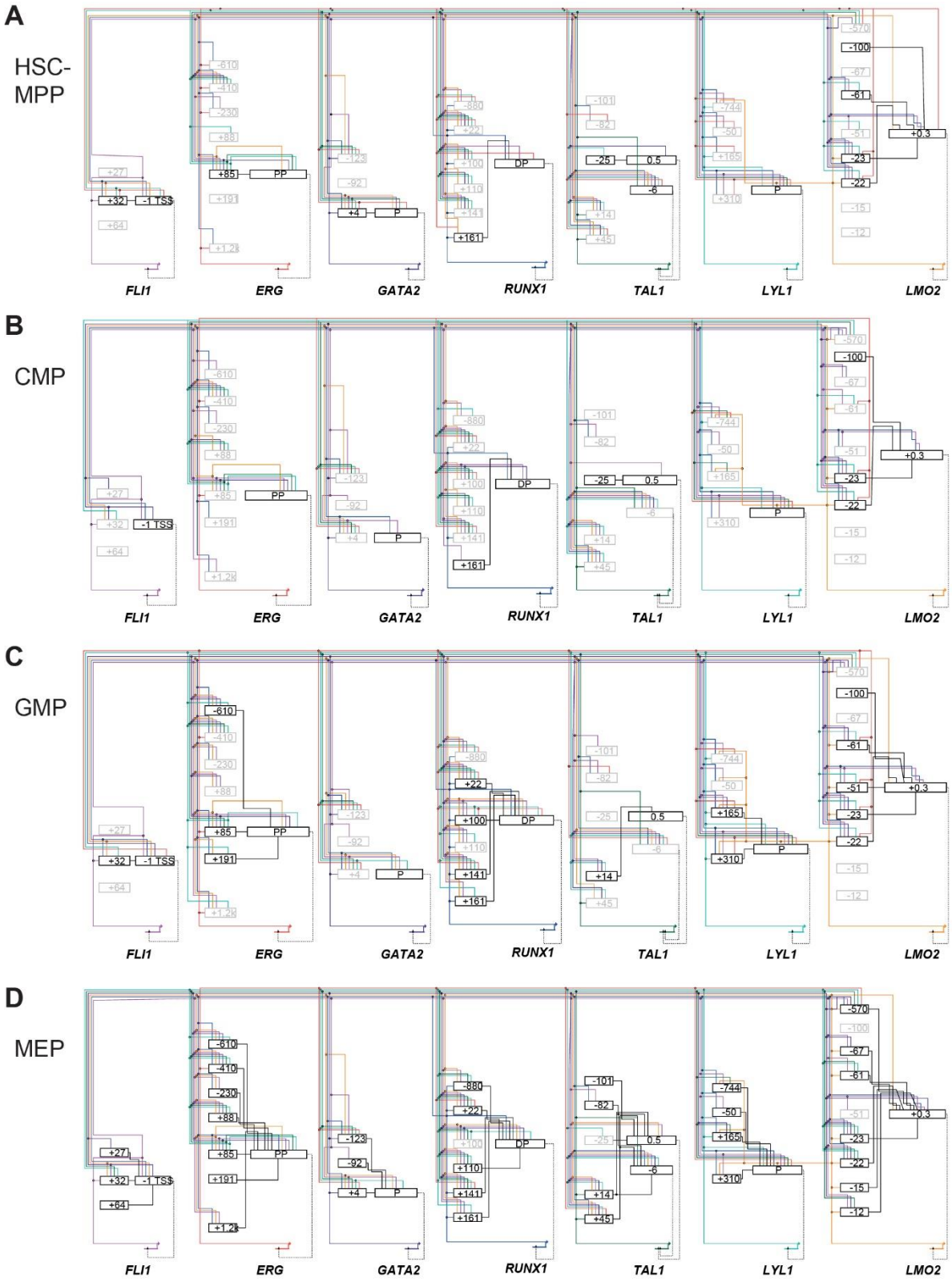

A
HSC-MPP
B
CMP
C
GMP
D
MEP
FLI1
ERG
GATA2
RUNX1
TAL1
LYL1
LMO2
-1 TSS
PP
P
DP
0.5

Figure S7

Figure S7. ***Gene regulatory network maps of the heptad genes in HSPC subsets*****.** Heptad GRNs in **A)** HSC-MPP, **B)** CMP, **C)** GMP, and **D)** MEP, constructed using BioTapestry software. Boxes in bold show active regulators, and their interaction with respective promoters marked with solid black lines. Solid coloured lines indicate heptad factors binding to regulatory regions (FLI1-pink, ERG-red, GATA2-purple, RUNX1-dark blue, TAL1-green, LYL1-aqua, and LMO2-orange), while dashed lines link regulatory sub-circuits to individual genes.

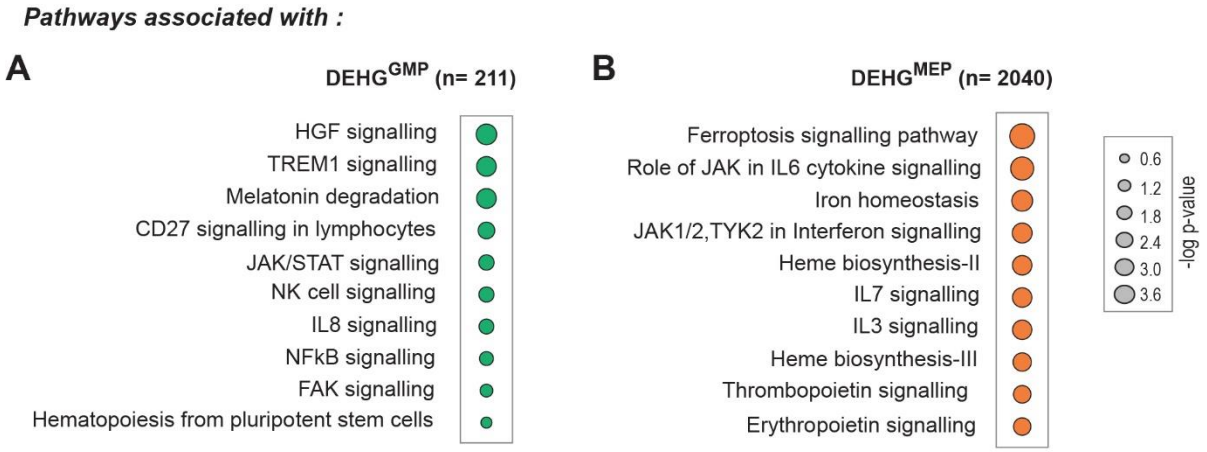


Figure S8

Figure S8. ***The role of heptad transcription factors in regulating lineage-specific gene expression.*** Ingenuity pathway analysis performed in DEHGs: **A)** DEHG^{GMP}, and **B)** DEHG^{MEP}.

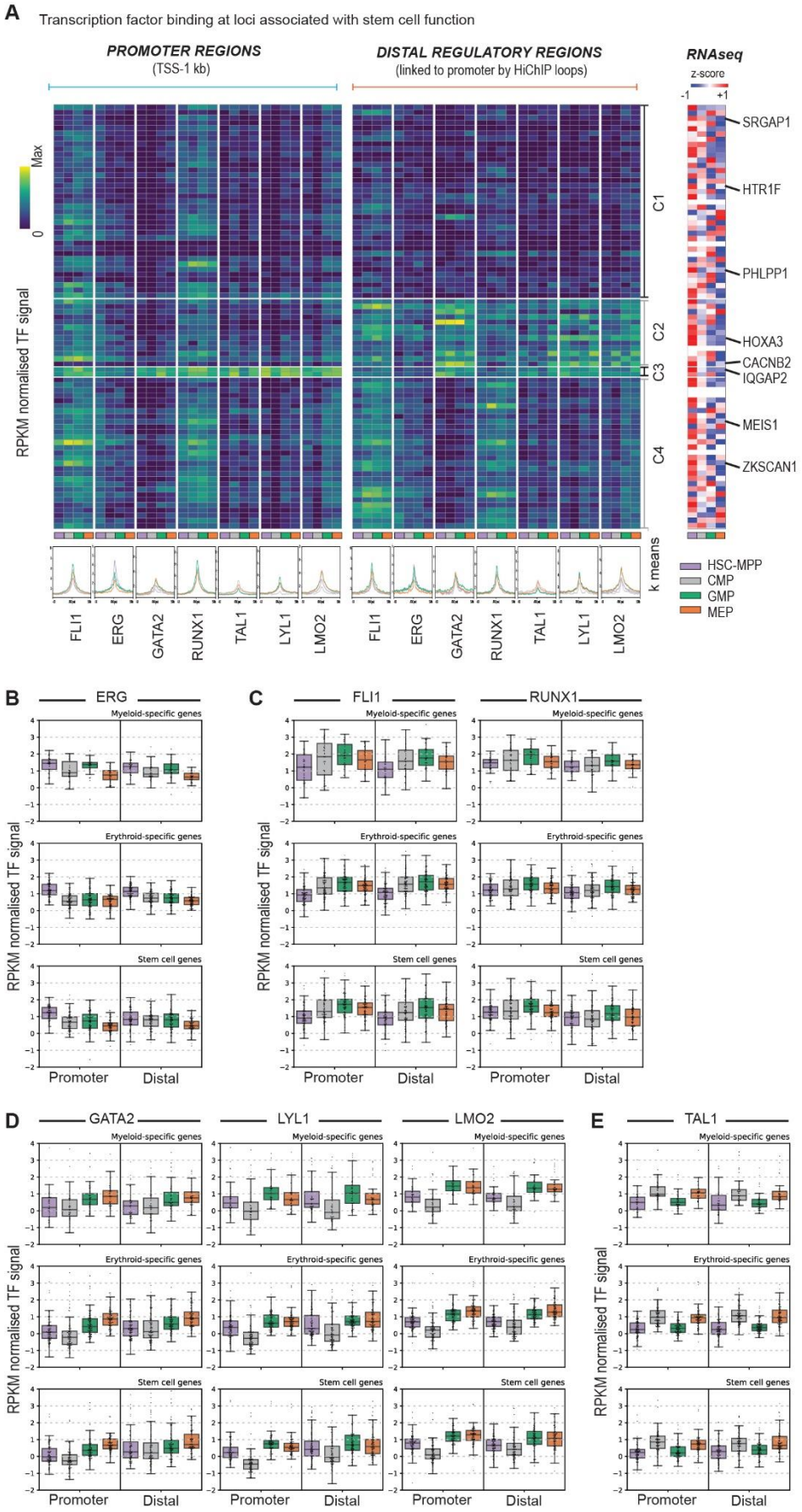

A
Transcription factor binding at loci associated with stem cell function
PROMOTER REGIONS
(TSS-1 kb)
DISTAL REGULATORY REGIONS
(linked to promoter by HiChIP loops)
RNAseq
z-score
-1
+1
Max
0
RPKM normalised TF signal
C1
C2
C3
C4
SRGAP1
HTR1F
PHLPP1
HOXA3
CACNB2
IQGAP2
MEIS1
ZKSCAN1
k means
HSC-MPP
CMP
GMP
MEP
FLI1
ERG
GATA2
RUNX1
TAL1
LYL1
LMO2
B
ERG
C
FLI1
RUNX1
D
GATA2
LYL1
LMO2
E
TAL1
Myeloid-specific genes
Erythroid-specific genes
Stem cell genes
Promoter
Distal

Figure S9

Figure S9. ***Cell-type specific binding patterns of heptad factors identified across stem cell specific genes***. **A)** Genes associated with stem cell function. *Left:* k-means clustered heatmaps of TF binding intensity at promoters and distal regulatory regions. Profile plots show normalised signal for each TF in each cell type at the regions depicted in the heatmap. *Right:* z-score normalised heatmaps of RNA-seq counts (GSE75384) for the corresponding gene in each cell type. **B-E)** Normalised TF signal at all promoters and distal regulatory regions for myeloid, erythroid, and stem cell genes. P-values for all pairwise comparisons (paired t-test) are shown in Table S4. **B)** Boxplots showing normalised ERG signal at promoters and distal regulatory regions of myeloid, erythroid, and stem cell genes. **C)** Boxplots showing normalised FLI1 and RUNX1 signal at promoters and distal regulatory regions of myeloid, erythroid, and stem cell genes. **D)** Boxplots showing normalised GATA2, LYL1, and LMO2 signal at promoters and distal regulatory regions of myeloid, erythroid, and stem cell genes. **E)** Boxplots showing normalised TAL1 signal at promoters and distal regulatory regions of myeloid, erythroid, and stem cell genes.

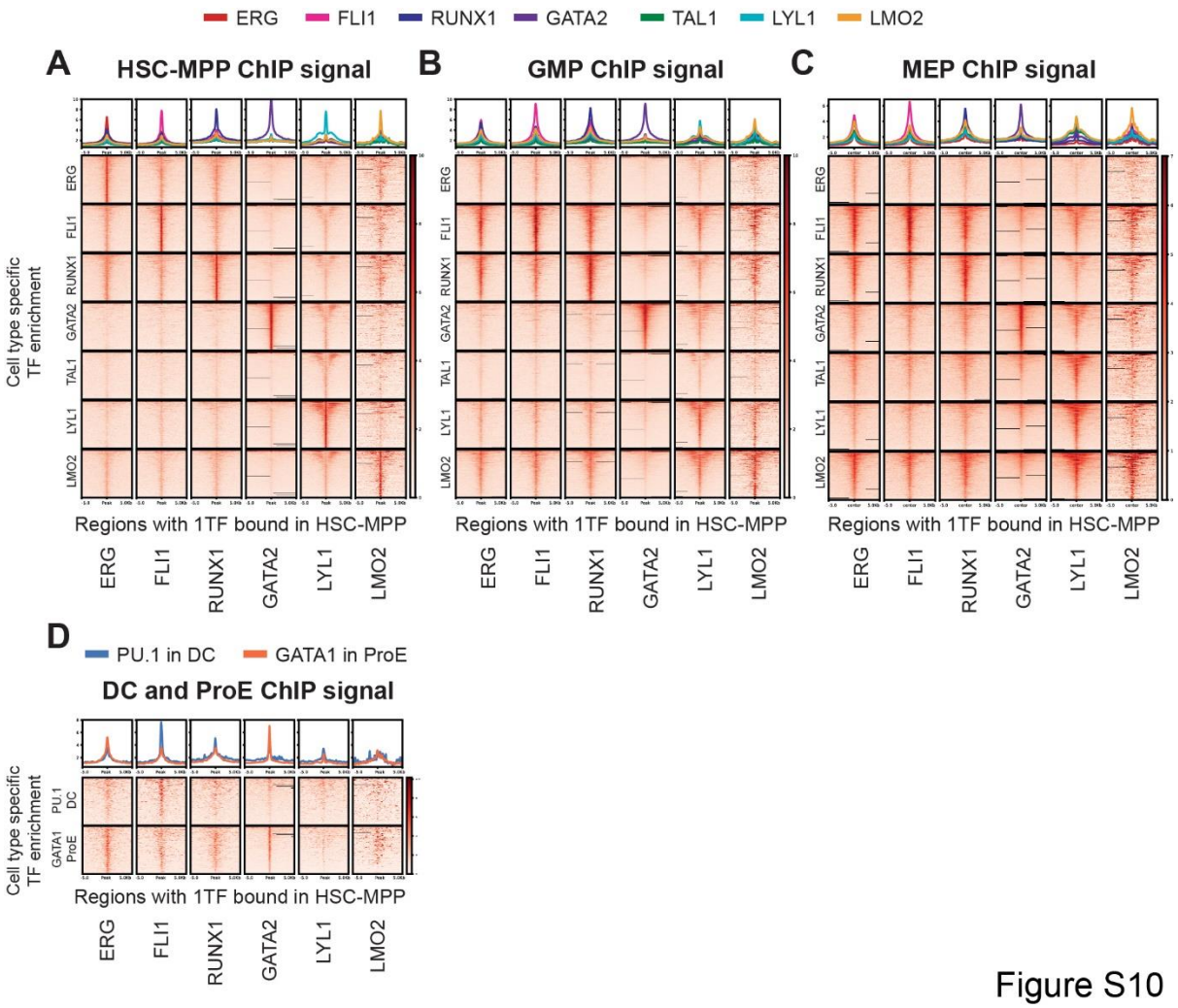


Figure S10

Figure S10. ***Sites bound by single heptad factors in HSC-MPPs are commonly bound by additional heptad factors in committed progenitors and mature cells.*** **A-C)** Density plots showing enrichment of TFs in **A)** HSC-MPPs **B)** GMPs **C)** MEPs at regions called as peaks for one heptad factor in HSC-MPP. TAL1 had only 54 peaks in HSC-MPPs which precluded analysis using this method. Columns show peak regions that are unique to the indicated heptad factor in HSC-MPP, rows show signal density of the indicated TF across those regions. **D)** Density plots showing enrichment of PU.1 in dendritic cells (DC) (GSE58864) and GATA1 in proerythroblasts (ProE) (GSE36985) at regions called as peaks for one heptad factor in HSC-MPP.

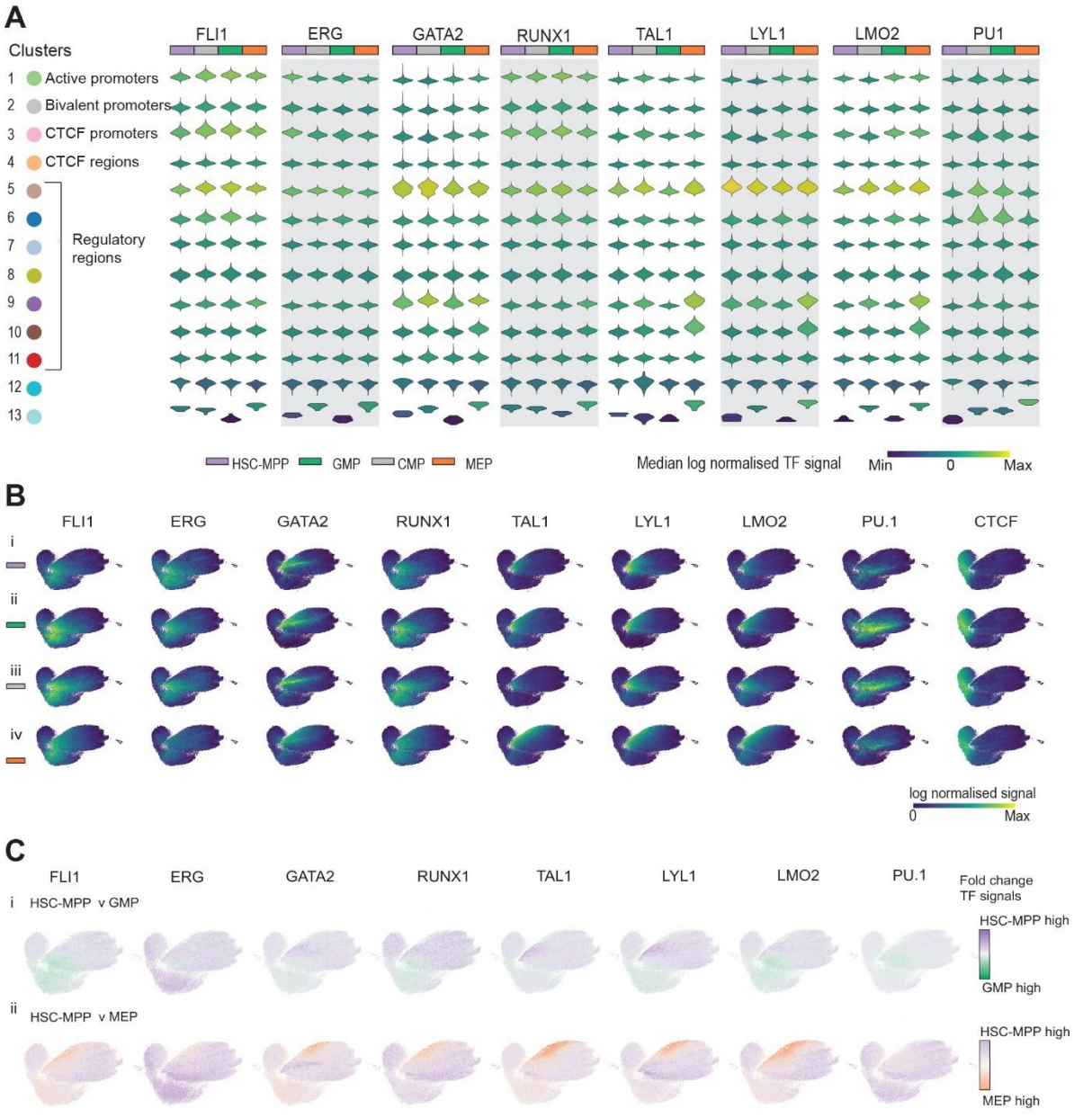


Figure S11

Figure S11. ***Transcription factor signal enrichment at ATAC regions in HSPCs.*** **A)** Individual violin plots show normalized transcription factor signals for FLI1, ERG, GATA2, RUNX1, TAL1, LYL1, and LMO2, in each of the 13 derived clusters for the four cell types studied. **B)** UMAPs showing heptad-factor-, PU.1-, and CTCF-normalized signals at accessible regions in i) HSC-MPP, ii) CMP, iii) GMP, and iv) MEP. **C)** UMAPs colored based on log2

fold change of binding of heptad transcription factors and PU.1 in pairwise comparisons between i) HSC-MPP and GMP, and ii) HSC-MPP and MEP.

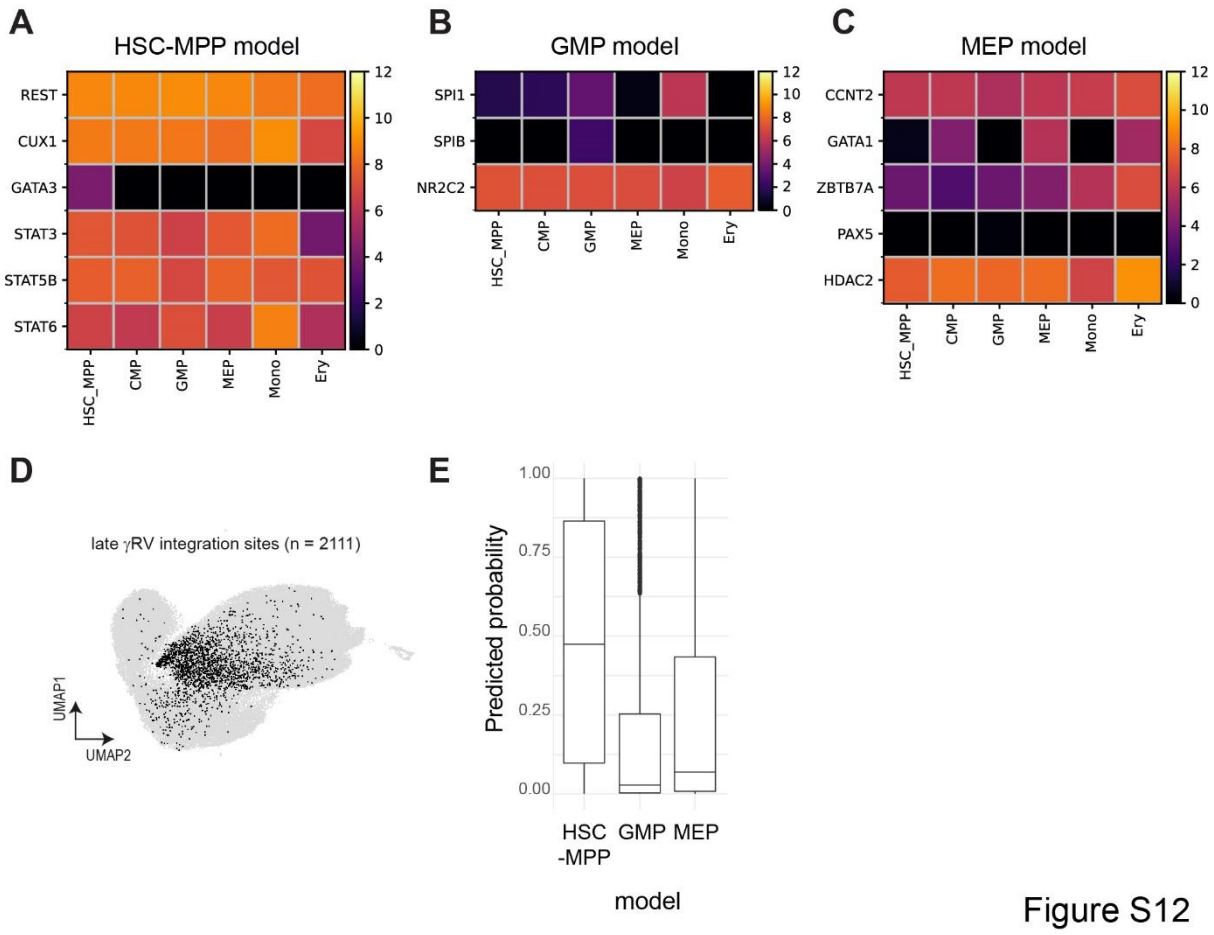


Figure S12. ***Cell-type specificity of regulatory elements is encoded in the underlying motif composition***. **(A-C)** Heatmaps showing cpm-normalised RNA-seq counts (GSE75384) for selected genes in HSC-MPPs, CMPs, GMPs, MEPs, monocytes (Mono), erythroblasts (Ery). Genes shown correspond to motifs with positive SHAP values (Figure 7) for **A)** HSC-MPP model, **B)** GMP model, **C)** MEP model. Only genes expressed in at least one cell type are shown. **D)** UMAP representation of ATAC-seq regions in $CD34^+$ cells (grey) with regions corresponding to late γ retroviral integration sites (γRV-IS, n = 2111) colored in black. Late γRV-IS integration sites correspond to functionally defined regulatory regions in long term repopulating human HSCs [1]. **E)** Graphs showing predicted probabilities that γRV-IS are HSC-MPP-, GMP-, or MEP-specific for n = 1674 γRV-IS. Sites that overlapped the model training data were excluded from prediction analysis.

Supplemental Methods

Antibodies, key reagents, and software

A complete list of key reagents is available in Table S9.

Biological samples

Mobilized peripheral blood samples were collected with patient consent in accordance with the Declaration of Helsinki, and used with institutional ethics approval ref:08/190 from South Eastern Sydney Local Health District, NSW, Australia.

Isolation of $CD34^+$ cells from cryopreserved apheresis packs

Cryopreserved cells were thawed, diluted 1:4 in 2.5% dextran/12.5% human albumin in 0.9% saline, centrifuged 200g/15 min/room temperature (RT) and resuspended in phosphate-buffered saline (DPBS) containing 10% fetal bovine serum (FBS). Cells were underlaid with lymphoprep, centrifuged 800g/30 min/RT, and mononuclear cells (MNCs) collected from the interface and washed with DPBS. MNCs were resuspended in ice cold CliniMACS buffer supplemented with 0.5% human albumin or AutoMACS running buffer then labelled with anti-CD34 microbeads according to manufacturer's instructions (Miltenyi Biotec). $CD34^+$ cells were enriched using either a CliniMACS Plus (Miltenyi Biotec) using standard clinical parameters or an AutoMACS (Miltenyi Biotec) using the program posseld2.

Labelling and sorting of $CD34^+$ cells.

$CD34^+$ cells were resuspended in FACS buffer (5% FBS/1mM EDTA in DPBS)

containing 1/10 diluted FcBlock at a concentration of 10^7 cells/100 µL and stained on ice for 30 min with a cocktail of antibodies (CD34 subset markers – CD38, CD38, CD123, CD45RA, CD90: Lineage markers (LIN) – CD2, CD3, CD4, CD7, CD8a, CD10, CD11b, CD14, CD19, CD20, CD56, GlyA/GPA/CD235ab). Cells were sorted using a BD FACS ARIA II into the following populations: HSC-MPP (LIN^-, $CD34^+$, $CD38^{lo}$, $CD45RA^-$), CMP (LIN^-, $CD34^+$, $CD38^+$, $CD45RA^-$, $CD123^+$), GMP (LIN^-, $CD34^+$, $CD38^+$, $CD45RA^+$, $CD123^+$), MEP (LIN^-, $CD34^+$, $CD38^+$, $CD45RA^-$, $CD123^-$) (Figure S1A). Population gates were set using fluorescence minus one controls, and phenotypic purity checks were performed on collected cell fractions.

Functional validation of cell purity was carried out for a subset of experiments. Sorted cells were resuspended 1% methylcellulose supplemented with cytokines as described [2], then plated in triplicate at 500 cells per dish and incubated in a humidified 37°C incubator with 5% CO2 for 14 days. Three major types of colonies were counted: erythroid-lineage (BFU-E) colonies, myeloid-lineage (GM) colonies, and colonies with mixed-potential (GEMM) (Figure S1B).

Crosslinking and preparation of nuclei

Sorted cells were incubated in freshly prepared 1% formaldehyde in DPBS for 10 min at RT. Crosslinking was quenched by adding glycine to a final concentration of 0.125 M and incubating for 5 min at RT. Subsequent steps were performed at 4°C with cold buffers. Cells were washed then resuspended in cell lysis buffer (10mM Tris-Cl pH 8.0, 10mM NaCl, 0.2% Tergitol, supplemented with 1µg/mL leupeptin, 1mM Pefabloc SC, 10mM sodium butyrate) and incubated on ice for 10 minutes. Nuclei were centrifuged 1450g/10 min/4°C then snap frozen and stored at -80°C for later use.

Chromatin Immunoprecipitation (ChIP)

ChIP for H3K27ac and H3K4me3 was carried out essentially as described [3]. Nuclei ($2–5 \times 10^6$/IP) were resuspended in 0.65mL nuclei lysis buffer (50mM Tris-Cl pH 8.0, 10mM EDTA, 1% SDS, protease inhibitors), incubated on ice for 10 minutes, with 0.4mL IP dilution buffer (20mM Tris-Cl pH 8.0, 2mM EDTA, 150mM NaCl, 1% Triton X-100, 0.01% SDS) and sonicated for 10 cycles in a Bioruptor Pico™ sonicator (Diagenode). Cleared supernatants were further diluted with 2.2mL IP dilution buffer, precleared with rabbit IgG then incubated overnight at 4°C with 5–10 µg of antibody. Antibody-chromatin complexes were recovered using protein G-agarose beads (Roche). After washing, immunoprecipitated DNA was eluted from beads, crosslinks reversed, and DNA purified using phenol-chloroform-isoamyl alcohol. ChIP libraries were prepared by a commercial supplier (Novogene). Donor cells used in each experiment are listed in Table S1.

ChIPmentation

ChIPmentation (CM) was carried out as described [4] with minor modifications. Biological triplicate experiments were performed for TFs except where noted otherwise (Table S2). Five million nuclei were resuspended in 100µL sonication buffer (10mM Tris pH 8.0, 2mM EDTA, 0.25% SDS), sonicated for 10 cycles in a Bioruptor Pico™ sonicator (Diagenode) and diluted 1:1.5 in equilibrium buffer (10mM Tris pH 8.0, 233mM NaCl, 1.66% Triton X-100, 0.166% sodium deoxycholate, 1mM EDTA). Cleared supernatants were incubated overnight at 4°C with 2µg of antibody, and antibody-chromatin complexes recovered using protein A/G magnetic beads. To improve signal to noise ratio for ERG CM only, we used a modified pull down protocol with two major differences: 1) nuclei lysates

were not cleared by centrifugation after sonication, and 2) anti-ERG antibody was pre-conjugated to magnetic protein A/G beads, and then incubated with nuclear lysate overnight at 4^{o}C to recover antibody-chromatin complexes.

After extensive washing, bead-bound complexes were resuspended in tagmentation mixture (25µL reaction containing 1 µL of enzyme in 1X buffer, Illumina) and incubated at 37°C for 25minutes. Crosslinking was reversed and DNA purified using a MinElute PCR Purification Kit (Qiagen). Barcoding/adapter primers (Table S10) and KAPA™ HiFi HotStart Ready Mix (Roche) were used to amplify libraries; the number of PCR cycles used was empirically determined for each reaction. Amplified DNA was purified then size selected using AMPure XP beads (Beckman) and sequenced using a standard Illumina 2 x 150bp PE pipeline (Novogene).

HiC and HiChIP

Duplicate HiC and HiChIP libraries for each cell type were generated using the Arima Genomics HiC+ kit (Arima cat#A101020). Nuclei were lysed and chromatin digested with a restriction enzyme cocktail prior to end-filling with biotinylated nucleotides and ligation of proximal ends. For HiChIP, ligated fragments were then immunoprecipitated with the H3K27ac antibody. Biotinylated fragments were enriched and sheared prior to library preparation which was performed using Accel NGS 2S Plus DNA Library kit (Swift Biosciences).

Bioinformatic processing

Analyses were run using default parameters for each tool unless otherwise indicated. Bigwig files were visualized using the UCSC browser [5]. Reads were aligned to the GRCh38

genome accession GCF_000001405.26 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/). Heatmaps were plotted using seaborn unless otherwise indicated.

Processing ATAC sequencing data

SRA files were downloaded from GEO (GSE74912), converted to fastq format, and aligned to the GRCh38 genome. Mitochondrial and duplicate reads were removed (respectively) , reads shifted to account for Tn5-mediated excision, peaks called, and bigwig files generated (pipeline: SRA tools - BWA - samtools - picard MarkDuplicates - alignmentSieve - macs2 with a minimal threshold p-value of $1x10^{-5}$ - deeptools bamCoverage)[6,7]. To obtain a composite of all accessible regions in HSPCs, ATAC peaks from HSC-MPP, CMP, GMP, and MEP were merged using bedtools [8].

Processing ChIP sequencing data

ChIP fastq files were analysed as single end data. fastq files were processed to remove adapter sequences (cutadapt)[9], trimmed to 100 bp (Trim Galore), then aligned to GRCh38 (BWA). Reads were sorted (samtools) and duplicates removed (picard MarkDuplicates), then bam files from replicates merged (samtools merge). IgG bam files from the four cell types were merged to generate a single IgG control. Peaks were called (macs2 with a minimal threshold p-value of 1×10^{-5} and using the IgG track as the control), and RPKM-normalized bigwig files generated and plotted (deeptools bamCoverage, plotHeatmap) [10]. Composite plots showing ChIPseq signal at specific genomic regions were plotted in python as previously described [11].

Motif enrichment analysis

Motif enrichment analysis was performed using the FIMO tool from the MEME analysis suite [12] using ETS, GATA, RUNX, and E-Box motifs sourced from JASPAR [13] as a position weight matrix.

Analysis of combinatorial binding

Genomic locations with occupancy of multiple heptad TFs were identified by intersecting ChIP peak coordinates (bedtools intersect). To assess the significance of each combination we performed a bootstrapping analysis essentially as previously described [14]. Briefly, we applied a permutation test to address significance of combinatorial binding events between the seven transcription factors for all 119 possible binding patterns. We chose the merged ATAC peak set of 85,117 peaks to estimate the background distribution of combinatorial binding events. The standardized z-score metric was used to express the deviation of the combinatorial binding events in high confidence peaks from the expected mean (normalized by the standard deviation) of the background distribution.

Analysis of HiC and HiChIP data

HiChIP and HiC fastq files were processed and mapped to GRCh38, then PCR duplicates removed and contact matrices generated from the merged valid-pairs files (HiC-Pro hicpro2juicebox) [15]. Contact matrices (.hic files) were visualized using juicebox [16]. HOMER was used to identify compartments and TADs from balanced HiC-Pro contact matrices. The first principal component (PC1) was generated using runHiCpca.pl at 50 kb

bins. In addition, H3K4me3 and H3K27me3 bed files from the respective cell types were used to assign accurate compartment labelling. To further identify TADs, findTADsAndLoops.pl function was used to generate TAD calls for each replicate separately – and merge2Dbed.pl to generate a union of TADs identified in each replicate. HiChIP contact matrices were used to generate interaction pairs at 5 kb resolution (MaxHiC)[17] and the WashU browser was used for loop visualization [18]. Most interactions spanned distances >10 kb.

High confidence interactions (FDR ≤ 0.01) were used to generate a final list of promoter–regulator interactions. To map promoter-regulator interactions at heptad gene loci we identified HiChIP fragments which overlapped known promoters. Distal fragments that were linked to these promoters were intersected with ATAC peaks from the relevant populations to precisely map the contact region within 5kb HiCHIP fragments (Table S3). Contact regions were named according to their linear genomic distance upstream (-) or downstream (+) from the transcriptional start site (TSS).

Visualization of gene regulatory networks (GRNs)

GRNs were visualized with BioTapestry [19] using ChIPseq peak calls and HiChIP-derived promoter–regulatory links to construct the network maps.

Identifying differentially bound regions

Candidate regulatory elements (REs) were defined as regions displaying combinatorial binding of heptad factors with a positive z-score, indicating that the combination is observed at higher frequency than expected by random chance. DiffBind was used to identify regions showing a significantly higher (FDR ≤ 0.05) combinatorial TF signal in one cell type compared to all others [20]. Two criteria were used for linking differentially

bound REs with genes: the presence of a RE at a gene promoter, or within a HiChIP fragment that was in turn linked to a gene promoter. Gene lists derived using this method were subsequently used for GSEA, IPA, and single cell RNAseq analysis (Figure 4).

Analysis of bulk RNAseq data

Fastq files and count tables were downloaded from GEO (GSE75384) and fastq files aligned to GRCh38 (STAR) [21]. edgeR was used to normalise the count table and calculate $\log_2$ CPM values [22] then derive a z-score of RNA expression.

Analysis of single cell RNAseq data

SCANPY [23] was used to process existing single-cell RNA sequencing data [24]. The SCANPY score_genes tool was used to generate a score for our gene sets, which was ultimately plotted on the original tSNE map generated by those authors.

TF occupancy at specific gene regulatory regions

Lists of genes associated with stem, myeloid, or erythroid function (Table S5B) were compiled from MSigDB genesets[25] (stem cell function – EPPERT_HSC_R[26], IVANOVA_HEMATOPOIESIS_STEM_CELL[27]: myeloid cell development - BIOCARTA_MONOCYTE_PATHWAY, BIOCARTA_GRANULOCYTES_PATHWAY, GOBP_GRANULOCYTE_DIFFERENTIATION, GOBP_GRANULOCYTE_MIGRATION, GOBP_GRANULOCYTE_ACTIVATION: erythroid cell development - HALLMARK_HEME_METABOLISM[28], GNATENKO_PLATELET_SIGNATURE[29],

BIOCARTA_PLATELETAPP_PATHWAY). Curated lists may not comprehensively catalogue every lineage specific gene. For this analysis promoter regions for each gene were defined as the ATAC peak occurring up to 10 kb upstream of the TSS. Distal fragments that were linked to these promoters in HiChIP datasets were intersected with ATAC peaks from the relevant population to define the distal regulatory element. Gene promoters lacking a looped distal region were excluded from this analysis. To determine the heptad TF signal for each distal region, we added together the average signal across each region from log_2-normalised bigwig tracks from each TF (bigWigAverageOverBed). If a looped distal regulator contained more than one ATAC peak, the TF signal from each peak region was averaged. We performed *k*-means clustering on the derived data using SciPy [30] and plotted the resulting heatmaps with seaborn.

Clustering analysis

The merged set of ATAC peaks which represent open chromatin regions across HSPCs were annotated using ChIP (heptad, PU.1, CTCF, ATAC, H3K27ac, H3K4me3, and H3K27me3) and ATAC signal from each individual cell type to create a dataframe of 85,100 rows and 52 columns, and the regions clustered with SCANPY using the Louvain method [23,31]. Gene associations for each ATAC region were predicted with the Genomic Regions of Enrichment Analysis Tool (GREAT) using the basal plus extension method with default parameters [32,33]. To identify regulatory regions preferentially used in specific cell types we compared TF signal at each region in HSC-MPP versus GMP, HSC-MPP versus MEP, and GMP versus MEP. We classified regions with log_2 fold change > 2 for each heptad factor in a cell type as cell-specific-regions. Regions identified by this method were subsequently used for machine learning models (Figure 7).

Machine learning analysis to predict cell type

We trained models using the R package XGBoost [34]. Briefly, we read a table of motif counts across the individual cell-specific-regions and took 70% of the peaks at random as the training set. We removed motifs with low variability and retained the remaining 30% of peaks as the test set. During the training, a series of decision trees were created such that a "loss function" was reduced (binary logistic in our case), to minimize cell type prediction error. Post training, prediction was performed on the test set. SHapley Additive exPlanation (SHAP) scores were calculated for every motif and peak used in the training set to indicate their respective contribution to the classification. A positive SHAP score for any given motif indicates that the presence of that motif in a region increases the probability that region belongs to the target cell type while a negative score indicates that the presence of that motif in a region increases the probability that region belongs to the background set (i.e., one of the other cell types). We then ranked motifs according to their importance by adding the absolute SHAP scores for every motif. To identify the direction of enrichment, we calculated the mean number of counts of every motif in the peaks that come from the target cell type or from the background peaks separately; if the mean was higher in the peaks from the target cell type those motifs were indicated as enriched in the target cell type and vice versa for the background set.

Analysis of γ retroviral integration site data

Late γ retroviral integration cluster genomic coordinates [1], representing viral integration sites (and thus potentially active regulatory elements) in long term repopulating human HSCs, were first overlapped with our merged ATAC peaks to generate genomic regions to use with our machine learning models (late γRV-IS, n = 2111). Of these, 395 (18.7%) corresponded to HSC-MPP-specific ATAC regions identified in our analysis; conversely 9.8% of our HSC-

MPP-specific sites corresponded to late γRV-IS. We next filtered out any regions that were part of the training sets for our models, leaving 1674 regions representing functionally validated regulatory regions that are active in hematopoietic stem cells with the capacity to engraft humans in a gene therapy setting (late γRV-IS). Late γRV-IS regions were scored against each machine learning model, and the predicted probabilities for each model plotted.

Data availability

A UCSC browser session for visualisation of chromatin occupancy and looping data is provided at http://genome.ucsc.edu/s/PimandaLab/Heptad_Regulome. We also provide a web tool for data exploration (http://unsw-data-analytics.shinyapps.io/CD34_Heptad_Regulome). Raw and processed sequencing files have been uploaded to GEO with accession # GSE231486.

References

1. Wunsche P, Eckert ESP, Holland-Letz T, et al. Mapping Active Gene-Regulatory Regions in Human Repopulating Long-Term HSCs. *Cell Stem Cell*. 2018;23(1):132-146 e139.

2. Schuller CE, Jankowski K, Mackenzie KL. Telomere length of cord blood-derived CD34(+) progenitors predicts erythroid proliferative potential. *Leukemia*. 2007;21(5):983-991.

3. Diffner E, Beck D, Gudgin E, et al. Activity of a heptad of transcription factors is associated with stem cell programs and clinical outcome in acute myeloid leukemia. *Blood*. 2013;121(12):2289-2300.

4. Schmidl C, Rendeiro AF, Sheffield NC, Bock C. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat Methods*. 2015;12(10):963-965.

5. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res*. 2002;12(6):996-1006.

6. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.

7. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.

8. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-842.

9. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal*. 2011;17(1):10-12.

10. Ramirez F, Ryan DP, Gruning B, et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*. 2016;44(W1):W160-165.

11. Thoms JAI, Truong P, Subramanian S, et al. Disruption of a GATA2-TAL1-ERG regulatory circuit promotes erythroid transition in healthy and leukemic stem cells. *Blood*. 2021;138(16):1441-1455.

12. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics*. 2011;27(7):1017-1018.

13. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*. 2022;50(D1):D165-D173.

14. Beck D, Thoms JA, Perera D, et al. Genome-wide analysis of transcriptional regulators in human HSPCs reveals a densely interconnected network of coding and noncoding genes. *Blood*. 2013;122(14):e12-22.

15. Servant N, Varoquaux N, Lajoie BR, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol*. 2015;16:259.

16. Durand NC, Shamim MS, Machol I, et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst*. 2016;3(1):95-98.

17. Alinejad-Rokny H, Ghavami Modegh R, Rabiee HR, et al. MaxHiC: A robust background correction model to identify biologically relevant chromatin interactions in Hi-C and capture Hi-C experiments. *PLoS Comput Biol*. 2022;18(6):e1010241.

18. Li D, Hsu S, Purushotham D, Sears RL, Wang T. WashU Epigenome Browser update 2019. *Nucleic Acids Res*. 2019;47(W1):W158-W165.

19. Longabaugh WJ. BioTapestry: a tool to visualize the dynamic properties of gene regulatory networks. *Methods Mol Biol*. 2012;786:359-394.

20. Stark R, Brown G. DiffBind: differential binding analysis of ChIP-Seq peak data 2011.

21. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21.

22. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139-140.

23. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;19(1):15.

24. Setty M, Kiseliovas V, Levine J, Gayoso A, Mazutis L, Pe'er D. Characterization of cell fate probabilities in single-cell data with Palantir. *Nat Biotechnol*. 2019;37(4):451-460.

25. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739-1740.

26. Eppert K, Takenaka K, Lechman ER, et al. Stem cell gene expression programs influence clinical outcome in human leukemia. *Nat Med*. 2011;17(9):1086-1093.

27. Ivanova NB, Dimos JT, Schaniel C, Hackney JA, Moore KA, Lemischka IR. A stem cell molecular signature. *Science*. 2002;298(5593):601-604.

28. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*. 2015;1(6):417-425.

29. Gnatenko DV, Dunn JJ, McCorkle SR, Weissmann D, Perrotta PL, Bahou WF. Transcript profiling of human platelets using microarray and serial analysis of gene expression. *Blood*. 2003;101(6):2285-2293.

30. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261-272.

31. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*. 2008;2008(10):P10008.

32. McLean CY, Bristor D, Hiller M, et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol*. 2010;28(5):495-501.

33. Tanigawa Y, Dyer ES, Bejerano G. WhichTF is functionally important in your open chromatin data? *PLoS Comput Biol*. 2022;18(8):e1010378.

34. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA; 2016.