HiC  H3K27ac  Hi-ChIP  ATACseq  RNAseq  ChIPseq: FLI1, ERG, GATA2, RUNX1, TAL1, LYL1, LMO2, PU.1, CTCF, STAG2, H3K27ac, H3K4me3, H3K27me3
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 Subramanian1,*, Julie A.I. Thoms2,+, Yizhou Huang3, Paola Cornejo4, Forrest C. Koch5, Sebastien Jacquelin6, Sylvie Shen7, Emma Song7, Swapna Joshi1, Chris Brownlee8, Peter S. Woll9, Diego Chaon Fajardo3, Dominik Beck3, David J. Curtis10, Kenneth Yehson11, Vicki Antonenas11, Tracey O’Brien12, Annette Trickett7, Jason A. Powell13,14, Ian D. Lewis13, Stuart M. Pitson13, Maher K. Gandhi15, Steven W. Lane16, Fatemeh Vafaee5,17, Emily S. Wong4,5, Berthold Göttgens18, Hamid Alinejad Rokny10, Jason W.H Wong20,† and John E. Pimanda1,2,21,22†

1School of Clinical Medicine, University of New South Wales, Sydney, Australia
2School of Biomedical Sciences, University of New South Wales, Sydney, Australia
3Centre for Health Technologies and the School of Biomedical Engineering, University of Technology Sydney, Sydney, Australia
4Victor Chang Cardiac Research Institute, Sydney, Australia
5School of Biotechnology and Biomolecular Sciences, Faculty of Science, University of New South Wales, Sydney, Australia
6Macrophage Biology Laboratory, Mater Research, Brisbane, Australia
7Bone Marrow Transplant Lab, NSW Health Pathology, Prince of Wales Hospital, Randwick, NSW, Australia
8Mark Wainwright Analytical Centre, University of New South Wales, Sydney, Australia
9Center for Hematology and Regenerative Medicine, Department of Medicine (MedH), Karolinska Institutet, Huddinge, Sweden
10Australian Centre for Blood Diseases, Monash University, Melbourne, VIC, Australia
11Blood Transplant and Cell Therapies Laboratory, NSW Health Pathology, Westmead, NSW, Australia
12Sydney Children’s Hospital, Sydney, Australia
13Centre for Cancer Biology, SA Pathology, University of South Australia, Adelaide, Australia
14Adelaide Medical School, University of Adelaide, Adelaide, Australia
15Blood Cancer Research Group, Mater Research, University of Queensland, Brisbane, QLD, Australia
16Cancer Program, QIMR Berghofer Medical Research, Brisbane, Australia
17UNSW Data Science Hub, University of New South Wales, Sydney, Australia
18Wellcome - MRC Cambridge Stem Cell Institute, Cambridge, United Kingdom
19BioMedical Machine Learning Lab, Graduate School of Biomedical Engineering, University of New South Wales, Sydney, Australia
20School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong
21School of Medicine, SA Pathology, University of South Australia, Adelaide, Australia
22Centre for Cancer Biology, University of South Australia, Adelaide, Australia
23Haematology Department, Prince of Wales Hospital, Sydney, Australia

†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. 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, relatively pure populations which correspond to intermediate progenitor stages can be prospectively isolated based on cell surface markers.

Changes in cell identity across differentiation trajectories are directly related to altered transcriptional programs which are controlled by lineage-specific gene regulatory networks (GRNs). 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. 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. Further control is imposed by chromatin organization into topologically associated domains (TADs). Interactions between promoters and their CREs, mediated by chromatin loops and complexes of transcriptional regulators, modulate GRNs and therefore cell identity.

We previously showed that seven TFs (heptad: FLI1, ERG, GATA2, RUNX1, TAL1, LYL1, and LMO2), all known regulators of healthy and leukemic haematopoiesis, bind combinatorially in bulk CD34+ hematopoietic stem and progenitor cells (HSPCs) and leukemias. 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. 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](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](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 j.pimanda@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](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, and chromatin accessibility at heptad gene CREs is sufficient to predict blood cell identity. 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, 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/−23/−22/−15/−12) (Figure 3Ai, ii, Figure S6, Table S4. Some regions have been described in humans and/or in mice 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) 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 S6Bi, Ci, Di, 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 3Bi). 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 3Bi). 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. 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) \cite{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. 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). 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.

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 which 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\textsuperscript{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\textsuperscript{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\textsuperscript{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. 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. 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, 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. 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 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 (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 targeting 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 \textit{GATA1} enhancers can drive restricted expression of \textit{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\textsuperscript{75,8}, often as a result of specific translocation events including RUNX1-RUNX1T1\textsuperscript{76,77} and inv(3)\textsuperscript{65,64}. Recent sequencing studies found that small structural variants (SVs) can mediate enhancer hijacking and subsequent oncogenic transformation\textsuperscript{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 heterogeneous\textsuperscript{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\textsuperscript{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

1 Conceptualization: SS, JAIT, JWHW, JEP
2 Methodology: SS, JAIT, YH, SJ, SSh, ES, CB, PSW, DCF, DB, AT, SWL
3 Investigation: SS, JAIT, YH, SSh, ES, SJ, DJC, KY, VA, TOB, JAP, IL, SMP, MKG
4 Formal analysis: SS, JAIT, PC, FCK, FV, EW, BG, HAR, JWHW, JEP
5 Supervision: JAIT, JWHW, JEP
6 Writing - original draft: SS, JAIT, JEP
7 Writing – review and editing: JAIT, JEP
8 Funding acquisition: JEP
CONFLICTS of INTEREST

The authors declare no relevant competing interests.
REFERENCES


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


**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](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 (SHaply Additive exPlanations) 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.
Figure 2
Figure 3
A

i. Candidate regulatory elements (REs) with overlapping TF peaks (>2TF)

- HSC: 7586
- CMP: 16766
- GMP: 7348
- MEP: 15602

ii. Candidate REs, differentially enriched for heptad TFs - DEH (diffBind; FDR< 0.05)

- HSC: 35
- CMP: 
- GMP: 223
- MEP: 3674

iii. Genes associated with DEH; DEHG in each cell type

- HSC: 16
- CMP: 
- 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

- Barcodes

- HOXB4, HOXB3, HOXB1

ii. DEHG^{GMP} (n= 211)

- NES -1.47
- FDR = 0.016

- DOWN

- Barcodes

- HOXB2, HOXB1

iii. DEHG^{HSC} (n= 16)

- NES 1.342
- FDR = 0.08

- UP

- Barcodes

- HOXB4, HOXB1

iv. DEHG^{MEP} (n= 2040)

- ES -0.28

- Barcodes

- HBB, CNTA1, EPB42

C

i. DEHG^{HSC}

- HSC
- Precursors
- Ery precursors
- Ery
- Mega
- Mono precursors
- Mono
- DC
- CLP

ii. DEHG^{GMP}

- HSC
- Precursors
- Ery precursors
- Ery
- Mega
- Mono precursors
- Mono
- DC
- CLP

iii. DEHG^{MEP}

- HSC
- Precursors
- Ery precursors
- Ery
- Mega
- Mono precursors
- Mono
- DC
- CLP

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  -1

SPI1  CD74  EP300
ITGAL  SP3  VAMP8

-1  -1

HSC-MPP  CMP  GMP  MEP

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  -1

TAL1  PPOX
ALAD  CDC27  RHCE

-1  -1

HSC-MPP  CMP  GMP  MEP

Figure 5
Figure 6
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 \textit{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
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. 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. **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).
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. *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.
Figure S6
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 *RUNX1* 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 *TAL1* locus (GRCh38 chr1:47,168,881-47,340,728). ii) Magnified view of the *TAL1* 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.
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. *The role of heptad transcription factors in regulating lineage-specific gene expression.* Ingenuity pathway analysis performed in DEHGs: A) DEHG\textsuperscript{GMP}, and B) DEHG\textsuperscript{Mep}.
Transcription factor binding at loci associated with stem cell function

A. 

**PROMOTER REGIONS**
(TSS -1 kb)

**DISTAL REGULATORY REGIONS**
(linked to promoter by HICMP loops)

RNaseq

<table>
<thead>
<tr>
<th>Species</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRAP1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HTRIF</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHLP1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H0X3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CADN82</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IQGAP2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEIS1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZKSCAN1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B. ERG

C. FLI1

D. GATA2

E. TAL1

17
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. 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. *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 posselld2.

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<sup>-</sup>, CD34<sup>+</sup>, CD38<sup>lo</sup>, CD45RA<sup>-</sup>), CMP (LIN<sup>-</sup>, CD34<sup>+</sup>, CD38<sup>+</sup>, CD45RA<sup>-</sup>, CD123<sup>+</sup>), GMP (LIN<sup>-</sup>, CD34<sup>+</sup>, CD38<sup>+</sup>, CD45RA<sup>+</sup>, CD123<sup>+</sup>), MEP (LIN<sup>-</sup>, CD34<sup>+</sup>, CD38<sup>+</sup>, CD45RA<sup>-</sup>, CD123<sup>-</sup>) (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<sup>2</sup>, 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 × 10⁶/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°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 25 minutes. 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

**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 x 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)\(^\text{17}\) and the WashU browser was used for loop visualization\(^\text{18}\). 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\(^\text{19}\) 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\(^\text{20}\). 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 log2 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 genesets25 (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 - HALLMARKHEME_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 log2-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 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. 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. 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 log2 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. 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, 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


