bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Integrative Single-cell RNA-Seq and ATAC-Seq Analysis of Human Foetal Liver and Bone Marrow Haematopoiesis

Anna Maria Ranzoni[1, 2, 3, *], Andrea Tangherloni[1, 2, 3, *], Ivan Berest[4], Simone Giovanni Riva[1, 2, 3], Brynelle Myers[2, 3], Paulina M. Strzelecka[1, 3, 5], Jiarui Xu[1, 2, 3], Elisa Panada[2, 3], Irina Mohorianu[2], Judith B. Zaugg[4], Ana Cvejic[1, 2, 3, ł]

[1] University of Cambridge, Department of Haematology, Cambridge, CB2 0AW, UK

[2] Wellcome Trust – Medical Research Council Cambridge Stem Cell Institute, Cambridge, CB2 0AW, UK

[3] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK

[4] European Molecular Biology Laboratory, Structural and Computational Biology Unit, Meyerhofstrasse 1, 69115, Heidelberg, Germany

[5] Current address: Charité Universitätsmedizin Berlin, 13353, Berlin, Germany

[*] Equal contribution

[ł] Corresponding author:
Ana Cvejic, as889@cam.ac.uk

Abstract

Regulation of human foetal haematopoiesis remains poorly defined. Here, we applied single-cell (sc)RNA-Seq and scATAC-Seq analysis to over 8,000 human immunophenotypic blood cells from liver and bone marrow from 18 foetuses. We inferred their differentiation trajectory and identified three highly proliferative oligopotent progenitor populations downstream from haematopoietic stem cell/multipotent progenitors (HSC/MPPs). Along this trajectory, we observed opposing patterns of chromatin accessibility and differentiation that coincided with dynamic changes in activity of distinct lineage-specific transcription factors. Integrative analysis of chromatin accessibility and gene expression revealed extensive epigenetic but not transcriptional priming of HSC/MPPs prior to their lineage commitment. Finally, we refined and functionally validated the sorting strategy for the HSC/MPPs and achieved 90% enrichment. Our study provides a useful framework for future investigation of human foetal haematopoiesis in the context of blood pathologies and regenerative medicine.

Keywords

Foetal haematopoiesis, haematopoietic stem cells, scRNA-Seq, scATAC-Seq, bone marrow, foetal liver.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

During embryonic development, haematopoietic stem cells (HSCs) need to rapidly differentiate into mature blood cells. Our current knowledge of foetal haematopoietic stem and progenitor cells (HSPCs) has been mainly advanced by murine and *in vitro* model systems. It has been demonstrated that foetal haematopoiesis consists of several, separate waves of specification, migration, and differentiation of rare HSCs at distinct organs during development (Ivanovs et al., 2017). In humans, definitive haematopoiesis starts with the appearance of HSCs within haematopoietic clusters, in the dorsal aorta, at 27 days post-conception. These definitive HSCs first colonise the foetal liver at 4 post-conceptional weeks (pcw) where they expand in numbers. At 10.5 pcw, the haematopoietic site shifts once more to the cavities of bones (i.e., bone marrow), where adult haematopoiesis is established permanently. The first HSCs that seed the bone marrow are thought to continue to rapidly increase in numbers before undergoing a dramatic change in their proliferative and differentiation properties to accommodate the need for high production of differentiated progeny (Mikkola and Orkin, 2006).

Historically, differentiation processes in the haematopoietic system have been depicted as a series of intermediate steps, defined by panels of cell surface markers (i.e., cluster of differentiation, CD). In this model, often represented as a “haematopoietic tree”, HSCs give rise to increasingly lineage-restricted cell types, eventually leading to mature blood cells (Akashi et al., 1999) (Weissman, 2000). This paradigm shifted in the last five years with several studies reporting the transcriptomes of thousands of single haematopoietic cells, isolated by cell surface markers, both in the mouse model and in adult humans (Paul et al., 2015) (Velten et al., 2017). These reports showed that progenitor populations, previously thought to be homogeneous, are actually very heterogeneous on the transcriptional level.

The mechanisms underlying early fate decisions in HSCs are largely unknown. It has been postulated that the stochastic expression of lineage-specific transcription factors (TFs) above the noise threshold can “lock” a cell into a distinct cell fate (Graf and Enver, 2009). In line with this, co-expression of genes associated with antagonistic lineages, including key TFs, have been observed in multipotent haematopoietic cells, albeit at low levels (Hu et al., 1997) (Miyamoto et al., 2002). This points towards the presence of sub-populations of cells within the multipotent compartment that are permissive for opposing cell fates prior to their lineage commitment, a phenomenon referred to as priming (Nimmo et al., 2015). More recently, single-cell RNA sequencing (scRNA-Seq) of human HSPCs introduced a different concept of priming. Studies of adult bone marrow and foetal liver haematopoiesis identified sub-populations of haematopoietic stem cells and multipotent progenitors (HSC/MPPs) with a coordinated

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

expression of marker genes, specific for distinct uni-lineage differentiation programmes, that gradually increased along all differentiation branches (Velten et al., 2017) (Popescu et al., 2019). In addition, there are some indications that lineage priming in the HSC compartment might be happening not only on the transcriptional but also at the epigenetic level (Nimmo et al., 2015). Data from single-cell Assay for Transposase Accessible Chromatin sequencing (scATAC-Seq) of phenotypic HSPCs from the adult human bone marrow show that phenotypic multipotent progenitors have variations in chromatin accessibility consistent with a bias towards erythroid and lymphoid lineages (Buenrostro et al., 2018).

Here we performed an integrative analysis of scRNA-Seq and scATAC-Seq of more than 8,000 immunophenotypic HSPCs, from 17-22 pcw human foetal liver, femur, and hip to define transcriptional and epigenetic changes during blood differentiation. We explored lineage priming at the transcriptional and chromatin level in HSC/MPPs and refined the sorting strategy for the isolation of a highly enriched HSC/MPP population.

Results

Single-cell transcriptome of the haematopoietic compartment in human foetal liver and bone marrow

The traditional methods of identifying HSPCs rely on the expression of combinations of cell-surface markers. Using sets of well-defined antibodies (CD markers) and fluorescence-activated cell sorting (FACS), we isolated committed (Lin- [CD3, CD8, CD11b, CD14, CD19, and CD56], CD34+ CD38+) and non-committed (Lin- CD34+ CD38-) blood progenitors as well as phenotypic HSPC populations from matched (i.e., from the same individual) foetal livers, femur, and hip (iliac) bones, between 17 and 22 pcw (Figure 1A). Specifically, we sorted HSCs, multipotent progenitors (MPPs), common myeloid progenitors (CMPs), megakaryocyte-erythroid progenitors (MEPs), granulocyte-monocyte progenitors (GMPs), and common lymphoid progenitors (CLPs). In addition, we sorted T cells, NK cells, innate lymphoid cells (ILCs), monocytes, dendritic cells, mast cells, basophils, neutrophils, eosinophils, erythroid progenitors, erythrocytes, immature megakaryocytes (MKs), mature MKs, progenitor B cells (pro-B cells), precursor B cells (pre-B cells), mature B cells, and endothelial cells (Supplementary Table 1, Supplementary Figure 1).

Single cells from 15 foetuses were index-sorted into 96-well plates and processed for scRNA-Seq using the SmartSeq2 protocol (Picelli et al., 2014) (Figure 1A).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

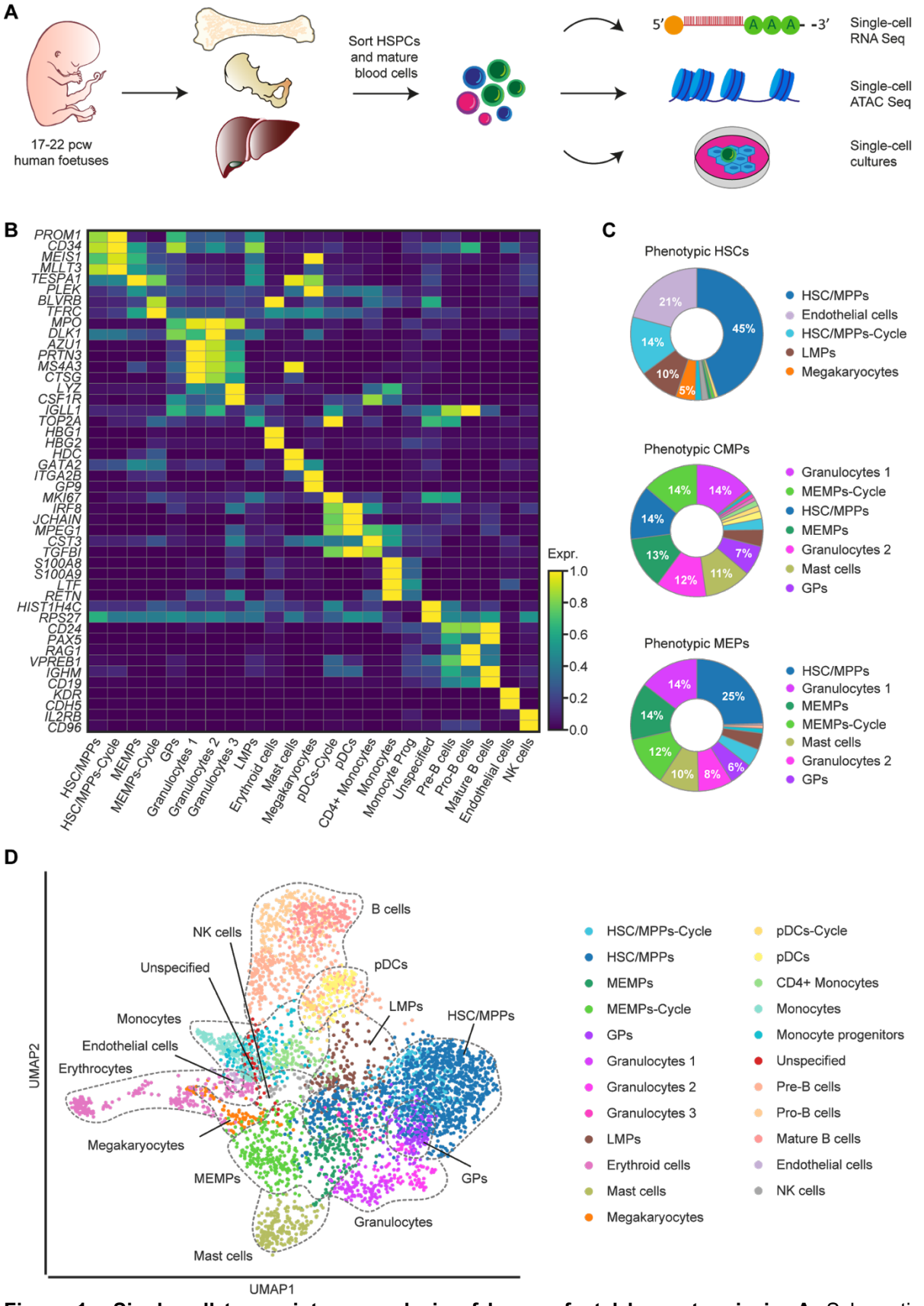


Figure 1 - Single-cell transcriptome analysis of human foetal haematopoiesis. A. Schematic overview of the experimental workflow. From each foetus (age 17-22 pcw), phenotypically defined HSPCs and mature blood cells were sorted from bone marrow (femur and hip) and liver and processed

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

for scRNA-Seq (n=15), scATAC-Seq (n=3), as well as for single-cell *in vitro* differentiation assays (n=4). **B.** Heatmap of the mean expression value of two manually selected marker genes for each cell type. The expression of the genes is standardised between 0 and 1. For each gene, the minimum value is subtracted and the result is divided by the maximum. The standardised expression level is indicated by colour intensity. **C.** Donut plots showing the percentage of transcriptionally defined (i.e., manually curated) cell populations in each of the phenotypically defined stem and progenitor populations. The colours correspond to the identified cell types. **D.** UMAP visualisation of haematopoietic cells from liver and bone marrow coloured by cell type. HSC/MPPs-Cycle - cycling haematopoietic stem cells/multipotent progenitors; HSC/MPPs - haematopoietic stem cells/multipotent progenitors; MEMPs - megakaryocyte-erythroid-mast progenitors; MEMPs-Cycle - cycling megakaryocyte-erythroid-mast progenitors; GPs - granulocytic progenitors; LMPs - lympho-myeloid progenitors; pDCs-Cycle - cycling plasmacytoid dendritic cells; pDCs - plasmacytoid dendritic cells.

Overall, 4,504 cells passed quality control (QC) with an average of ~3,600 genes per cell (Supplementary Figure 2A-C). To exclude technical batch effects, we merged the datasets using autoencoders (AEs) and applied the batch balanced *k* nearest neighbours (BBKNN) approach (Polański et al., 2020) to the latent space (Tangherloni et al., 2019) (Supplementary Figure 2M). We applied the graph-based Leiden clustering algorithm (Traag et al., 2019) to the batch corrected neighbourhood graph. Based on differential expression (DE) analysis and top 20 marker genes (Figure 1B) ranked on the significance of standardised expression, we manually annotated 23 distinct populations. Within the haematopoietic progenitor compartment, we annotated clusters as HSC/MPPs, HSC/MPPs-Cycle, lympho-myeloid progenitors (LMPs), megakaryocyte-erythroid-mast progenitors (MEMPs), MEMPs-Cycle, granulocytic progenitors (GPs), as well as numerous mature blood cell types as shown in the Uniform Manifold Approximation and Projection (UMAP) space (Becht et al., 2018) (Figure 1D).

Of the mature blood cell types, we identified clear transcriptional signatures of erythroid cells (expressing *HBG1*, *HBA1*, *GYPA*, and *ALAS2*), megakaryocytes (expressing *FLI1*, *ITGA2B*, and *GP9*), monocyte progenitors and monocytes (expressing *SPI1*, *MPEG1*, and *FCN1*), CD4+ monocytes, mast cells (expressing *CD63*, *CPA3*, and *HDC*), plasmacytoid dendritic cells (pDCs - expressing *IL3RA*, *IRF8*, *MPEG1*, and *JCHAIN*), with an additional cluster of highly cycling pDCs (expressing pDC and proliferation markers, e.g., *MKI67*) and granulocytes 1, 2, and 3 (expressing *AZU1*, *MPO*, and *PRTN3*). While granulocytes were present in our dataset, we could not clearly distinguish neutrophils, basophils, and eosinophils due to the mixed expression signatures. In the lymphoid compartment, we identified NK cells (expressing

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CD3D, *IL2RB*, and *CD96*) and B cells (expressing *CD79A* and *CD79B*). B cell lineage included pro-B, which showed expression of *IGLL1*, *DNTT*, and *RAG1*, and pre-B, expressing high levels of *CD79B*, *VPREB1*, and *CD24.* Finally, we identified a cluster of mature B cells, expressing high levels of *IGHM* and *IGHD* and decreased levels of *IGLL1*, compared to pro/pre B clusters. We did not detect any T cells or ILCs in the liver or in the femur in spite of sorting phenotypic T cells and ILCs using broad cell surface markers for these populations. By using a Deep Neural Network (DNN) (LeCun et al., 2015) and the top 30 marker genes for each cluster, we were able to correctly classify the cells to the prospective clusters with 90.23% accuracy, confirming that our manual annotation of clusters separated well the distinct cell types/states (see Methods, Supplementary Figure 3A).

The index sorting approach allowed us to compare the extent to which the phenotypic identity of cell populations (as defined by CD markers) matched their transcriptional state, i.e., our manually curated clusters. Single-cell analysis revealed substantial transcriptional heterogeneity within all immunophenotypically-defined stem and progenitor populations, with some phenotypic progenitor populations such as HSCs, MPPs, CMPs, GMPs, MEPs, and CLPs being comprised of more than ten different transcriptionally-defined populations. (Figure 1C, Supplementary Figure 3B). This observation is in agreement with recent research showing a high level of heterogeneity of the progenitor compartment of human cord blood (Knapp et al., 2018). Taken together, our comparative analysis has unequivocally shown that currently used cell-surface markers are a poor predictor of the transcriptional state of human foetal haematopoietic progenitors.

Inference of differentiation trajectories during foetal haematopoiesis

Next, we used a Force-Directed Graph drawing algorithm, ForceAtlas2, to infer the differentiation trajectory of haematopoietic cells during human foetal development (Jacomy et al., 2014). We initialised a ForceAtlas2 layout with Partition-based Approximate Graph Abstraction (PAGA) coordinates from our annotated cell types (Wolf et al., 2019). This initialisation generated an interpretable single-cell embedding that is faithful to the global topology. The obtained global topology revealed HSC/MPPs at the tip of the trajectory (Figure 2A, Supplementary Figure 4). HSC/MPPs showed high expression of *MLLT3*, a crucial regulator of human HSC maintenance (Calvanese et al., 2019), *HLF*, a TF involved in preserving quiescence in HSCs (Komorowska et al., 2017) and *MEIS1*, a TF involved in limiting oxidative stress in HSCs, which is necessary for quiescence (Unnisa et al., 2012) (Wang et al., 2018) (Fig 2C). Cells in this cluster also expressed high levels of surface markers of HSPCs such as *CD34*, *CD52* (Morisot et al., 2006), *SELL* (Ivanovs et al., 2017), and *PROM1* (de Wynter et al., 1998*)* (Saha et al., 2020).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

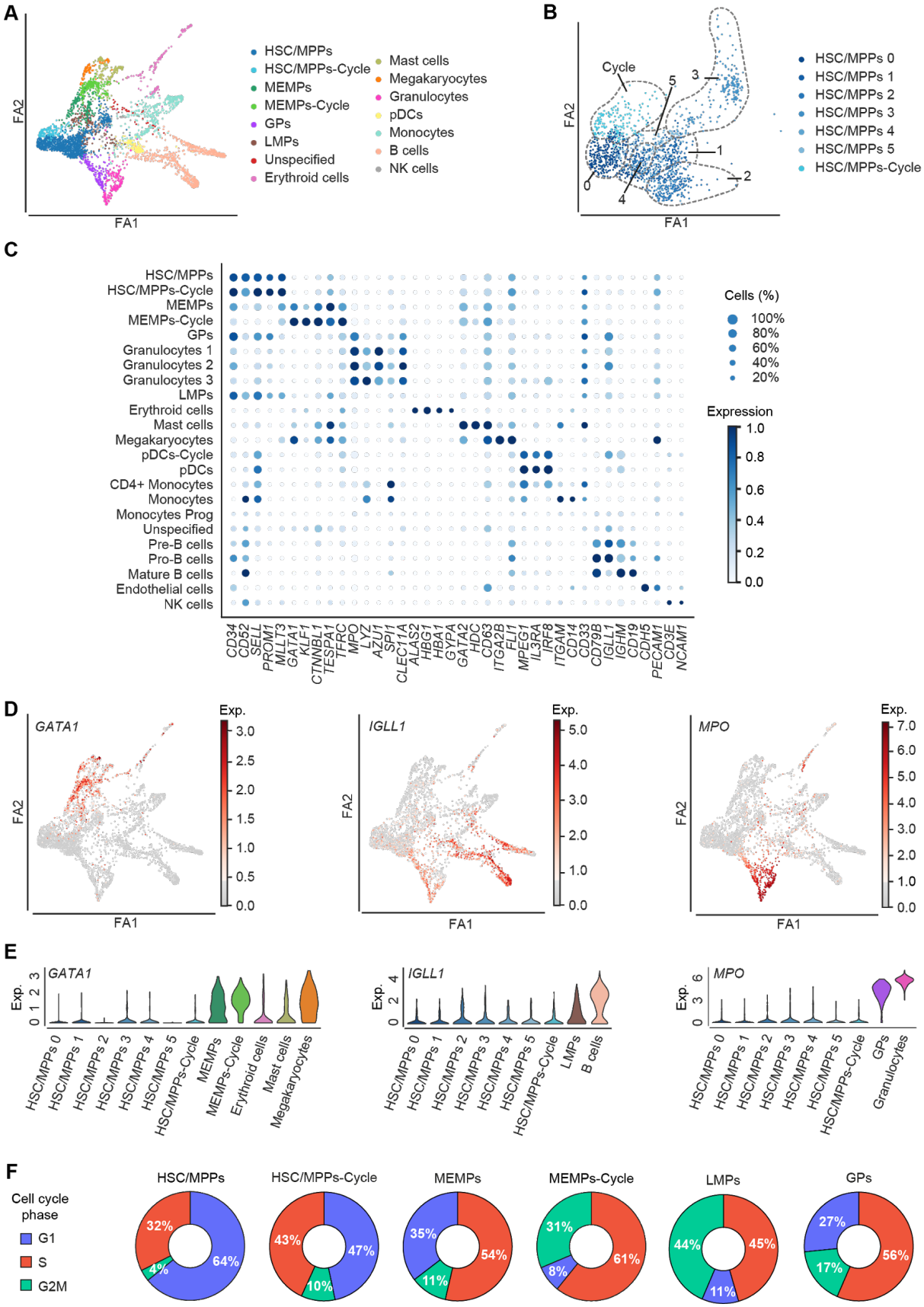


Figure 2 - Differentiation trajectory of human foetal haematopoietic cells. A. FDG visualisation of the differentiation trajectory of haematopoietic cells from Figure 1D. **B.** FDG visualisation of the

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

HSC/MPP sub-populations. **C.** Dot plot of the standardised expression of 38 manually selected marker genes in the identified cell types. For each gene, the minimum value of its expression is subtracted and the result is divided by the maximum value of its expression. The dot size indicates the percentage of cells that express the gene of interest within each cell type. **D.** FDG visualisation of the *log*-normalised gene expression of *GATA1*, *IGLL1*, and *MPO* along the differentiation trajectory. **E.** Violin plots showing the *log*-normalised median gene expression of *GATA1*, *IGLL1*, and *MPO* in the HSC/MPP sub-populations and more differentiated haematopoietic cell. **F.** Donut plots showing the percentages of cells in G1, S, and G2M in HSC/MPPs, HSC/MPPs-Cycle, MEMPs, MEMPs-Cycle, LMPs, and GPs. Force-Directed Graph - FDG; ForceAtlas2 - FA2; HSC/MPPs-Cycle - cycling haematopoietic stem cells/multipotent progenitors; HSC/MPPs - haematopoietic stem cells/multipotent progenitors; MEMPs - megakaryocyte-erythroid-mast progenitors; MEMPs-Cycle - cycling megakaryocyte-erythroid-mast progenitors; GPs - granulocytic progenitors; LMPs - lympho-myeloid progenitors; pDCs-Cycle - cycling plasmacytoid dendritic cells; pDCs - plasmacytoid dendritic cells.

We next examined whether HSC/MPP cells simultaneously primed several different lineage-affiliated programs of gene activity. While HSC/MPPs sporadically expressed lymphoid, myeloid, or megakaryocyte-erythroid differentiation genes, we did not observe consistent expression of antagonistic lineage-affiliated genes in individual cells. In addition, after further sub-clustering the HSC/MPPs, there was no evident consolidation of lineage-affiliated transcriptional programs in any of the sub-populations (Figure 2B, D, and E, Supplementary Figure 4). Our scRNA-Seq data, thus, do not support recently reported transcriptional lineage priming in the foetal HSC/MPP compartment (Popescu et al., 2019) and suggest that, transcriptionally, our HSC/MPP cluster represents a highly immature population of cells.

DE analysis between HSC/MPPs-Cycle and HSC/MPPs revealed upregulation of genes involved in cell cycle regulation (*FOS*, *PTP4A1*, *MCL1,* and *PKN2*) in HSC/MPPs-Cycle (Supplementary Figure 5A) thus confirming that they are indeed a population of cycling stem and multipotent cells. In line with this, cell cycle analysis confirmed that ~36% of HSC/MPPs were cycling compared to ~53% of HSC/MPPs-Cycle (Figure 2F). The HSC/MPPs-Cycle had an increased expression of genes involved in glycolysis, a feature commonly found in proliferating cells (Ito and Suda, 2014) (Supplementary Figure 5G). However, there were no other transcriptional differences between HSC/MPPs and HSC/MPP-Cycle, excluding the presence of transcriptional priming in the HSC/MPP-Cycle cluster.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Downstream of HSC/MPPs, we identified three distinct, highly proliferative, oligopotent progenitor populations. MEMPs connected HSC/MPPs with megakaryocytes, erythroid, and mast cells. In line with this, differentially regulated genes in the HSC/MPPs transition to MEMPs included megakaryocyte/erythroid/mast cells lineage-specific genes such as *GATA1*, *PLEK*, *KLF1*, *HDC, and MS4A3*, (Figure 2D-E, Supplementary Figure 5B). Presence of MEMPs in our dataset is consistent with studies in mouse models proposing a common trajectory between erythroid, megakaryocytic, and mast cell lineages (Franco et al., 2010). This concept was more recently supported by a study in human foetal liver showing a shared progenitor of megakaryocyte, erythroid, and mast cells (Popescu et al., 2019). In addition, we identified a proliferative population of MEMPs-Cycle of which ~92% were in the G2M/S phase compared to 65% of MEMPs (Figure 2F). MEMPs-Cycle population further upregulated erythroid-specific genes such as *KLF1*, *BLVRB*, *SMIM1*, and *TFRC* compared to MEMPs suggesting their gradual commitment towards erythroid lineage (Figure 2C-E, Supplementary Figure 5C).

GPs connected the HSC/MPP cluster with granulocyte clusters. Cells in this cluster differentially expressed myeloid lineage-specific genes (e.g., *LYZ*, *AZU1*, and *MPO*) compared to HSC/MMPs (Figure 2D and 2E, Supplementary Figure 5) and were highly cycling, with 73% of cells in the G2M/S phase (Figure 2F). Finally, our data pointed towards the existence of a common progenitor population for B cells, monocytes, pDCs, and NK cells, here annotated as lymphoid-myeloid progenitors (LMPs). Cells in this cluster expressed genes specific to those lineages, including *IGLL1*, *HMGB1* and *CD79B* (lymphoid) (Figure 2D and 2E) and upregulated lymphoid genes such as *CD81*, *IGLL1,* and *HMGN2* compared to HSC/MPP cluster (Supplementary Figure 5). Again, this was a highly proliferative population of cells with ~89% of cells being in the G2M/S phase (Figure 2F). Our findings support previous studies on early lymphoid commitment in human cord blood, both *in vitro* and *in vivo*, which identified a shared lineage progenitor between lymphoid, NK, B, and T cells, monocytes, and dendritic cells (Doulatov et al., 2010) (Collin et al., 2011). Interestingly, the LMP cluster had higher expression of MPP-related genes such as *SPINK2*, *CD52,* and *SELL* compared to MEMPs, suggesting that these progenitors represent a more immature population compared to MEMPs (Supplementary Figure 5).

Previous research showed that, contrary to adult blood progenitors that are mainly unilineage, foetal liver blood progenitors maintain multilineage potential (Notta et al., 2016). Our data are consistent with this observation and point towards the existence of three oligopotent progenitor populations downstream of the HSC/MPP compartments: MEMPs giving rise to erythroid,

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

megakaryocytes, and mast cells, GPs differentiating into granulocytes and LMPs generating lymphoid, monocytes and dendritic cells.

scATAC-Seq of foetal non-committed progenitors (CD34+ CD38-)

Detection of low abundant transcripts, such as TFs, might be difficult in scRNA-Seq data due to technical limitations of the approach, leading to false negatives (so-called drop-outs). The activity of these TFs can be inferred, however, from chromatin accessibility, emphasizing the importance of approaches integrating scRNA-Seq and scATAC-Seq data. In addition, chromatin accessibility at regulatory regions might precede gene activity and thus have predictive value for future transcription of a gene. Therefore, to further investigate the regulatory events in the very immature cell populations, we examined the single-cell chromatin accessibility landscape (using scATAC-Seq) of human foetal Lin- CD34+ CD38- cells (see Methods). We sequenced 4,001 cells from the liver and femur of three foetuses, 18, 20, and 21 pcw. (see Methods). Based on our scRNA-Seq data, we expected that 90% of captured cells would be associated with one of the six populations: HSC/MPPs, HSC/MPPs-Cycle, MEMPs, MEMPs-Cycle, GPs, and LMPs, with HSC/MPPs(Cycle) constituting the majority (Supplementary Figure 3B).

To capture peaks that are present in less abundant cell types such as MEMPs, MEMPs-Cycle, GPs, and LMPs, we employed an iterative peak-calling approach. We first defined open chromatin regions by pooling all the data and calling peaks in the pooled samples. Following dimensionality reduction with Diffusion maps (Haghverdi et al., 2015) and clustering using the Louvain community detection algorithm (Blondel et al., 2008), we performed a second round of peak calling in the clusters with more than 50 cells. Out of the initial ~474,000 reads, after preprocessing steps, (Supplementary Figure 2D-F), on average we detected ~32,400 fragments per cell and 56% of those mapped to peaks (Supplementary Figure 2G-H-K). Following filtering steps (Supplementary Figure 2I, J, L), 3,611 cells passed QC with 152,282 distinct peaks.

Motif accessibility dynamics along the inferred differentiation trajectories

In order to merge samples and remove the batch effects, we applied Harmony (Korsunsky et al., 2019) on the first 50 Latent Semantic Indexing (LSI) components, excluding the first one because it was highly correlated to the sequencing depth (Supplementary Figure 2N-O). By using a shared nearest neighbour (SNN) modularity optimization based clustering algorithm, we obtained seven distinct clusters of differentially accessible peaks (Figure 3A).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

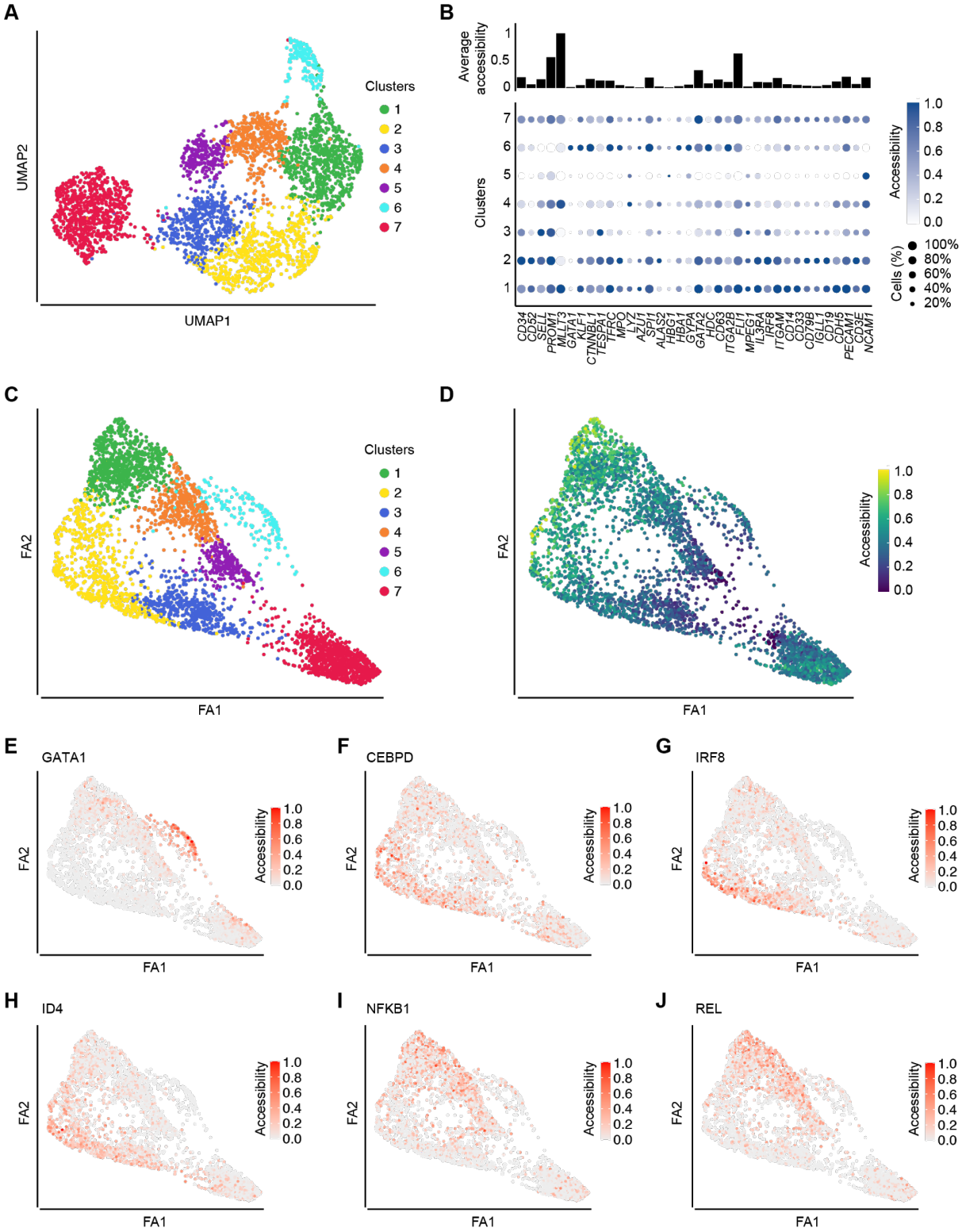


Figure 3 - Single-cell chromatin accessibility analysis of human foetal haematopoiesis. A. UMAP visualisation of scATAC-Seq dataset (n=3,611 nuclei from CD34+ CD38- cells from liver and bone marrow) coloured by cluster. **B.** (Top) Bar plot showing the average accessibility of 36 selected marker genes from our scRNA-Seq data considering all cells. (Bottom) Dot plot of the standardised accessibility of the marker genes (gene body ± 3 kb) in each of the seven clusters. For each gene, the minimum value of its accessibility is subtracted and the result is divided by the maximum value of its accessibility.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The dot size indicates the percentage of cells in each cluster in which the gene of interest is accessible. **C.** FDG visualisation of the differentiation trajectory of haematopoietic cells from **(A)**. **D.** FDG visualisation of the min-max normalised global chromatin accessibility (i.e., number of fragments in peaks per cell). **(E-J)** FDG visualisation of the min-max normalised TF motif accessibility along the differentiation trajectory. **E.** GATA1. **F.** CEBPD. **G.** IR8. **H.** ID4. **I.** NFKB1. **J.** REL. Force-Directed Graph - FDG; ForceAtlas2 - FA2.

To explore the chromatin accessibility profiles across the seven clusters, we examined the accessibility of selected marker genes from our scRNA-Seq data (Figure 3B). We observed higher accessibility of marker genes associated with stem cells (e.g., *MLLT3*, *PROM1*, *FLI1*, and *GATA2*) and lower accessibility of genes associated with distinct lineages (e.g., *MPO*, *ALAS2*, *MPEG1*, and *CD19*), keeping in line with the undifferentiated nature of sorted cells. Interestingly, we observed a clear separation of clusters in terms of their overall accessibility of marker genes, with clusters 1, 2, 4, and 7 being more accessible and clusters 3 and 5 being less accessible. Cluster 6 had a mixed signature (Figure 3B).

Extensively open chromatin in multipotent cells has been previously associated with a permissive state to which multiple programmes of gene regulation may be applied upon differentiation and is considered important for the maintenance of pluripotency (Gaspar-Maia et al., 2011). To further investigate if there were global dynamic changes in accessibility patterns associated with the differentiation of foetal HSC/MPPs, we inferred differentiation pseudotime from our scATAC-Seq data using the same approach as with scRNA-Seq described above. Briefly, we built a Force-directed Graph from our seven scATAC-Seq clusters by initialising a ForceAtlas2 layout with PAGA coordinates. The generated trajectory revealed two branches with a clear trend between chromatin accessibility and differentiation in each branch (Figure 3C-D). We observed the highest accessibility in clusters 1, 2, and 4 that gradually decreased towards the tips of the two branches (i.e., clusters 3 on one side, and 6 on the other). This result is compatible with the notion that clusters 1, 2, and 4 represent HSC/MPP population.

Control of gene expression is a dynamic process that involves both the cell-type-specific expression of TFs and the establishment of an accessible chromatin state that permits binding of TFs to a defined motif. Thus, to assess regulatory programs that are active in HSPCs, we used chromVAR (Schep et al., 2017) to calculate the most variable accessible TF sequence motifs in different clusters and examine their activity along the differentiation trajectory. We

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

observed profound differences in the accessibility of important haematopoietic TF motifs such as GATA1, TAL1, CEBPD, IRF8, NFKB1/2, REI, RELB, TFE2, HTF4, HXB7/8/9, *and* ID4 (Figure 3E-J, Supplementary Figure 6).

GATA1 motif activity (Figure 3E) and gene-body accessibility (Figure 3B) were enriched in cluster 6. GATA1 is known to be an important regulator of erythroid, megakaryocytic, and mast cells differentiation (Katsumura et al., 2017) and was exclusively expressed in the MEMP cluster, in our scRNA-Seq dataset. Thus, cluster 6 most likely represents the MEMP population. Interestingly, in cluster 6, compared to clusters 2 and 3, we detected opposing patterns of motif accessibility for the two different TAL1 binding sites (TAL1.0.A and TAL1.1.A, respectively) (Supplementary Figure 6A,B). Substantial changes in occupancy by TAL1 during differentiation have been observed, which are dependent on its binding partners (Wu et al., 2014). It has been previously reported that TAL1.0.A was co-occupied by TAL1 and GATA1 (Kassouf et al., 2010) whereas TAL1.1.A by TAL1 and TCF3 (Hsu et al., 1994). Our analysis, thus, revealed that the two different TAL1 binding motifs are active in distinct haematopoietic progenitor populations during foetal haematopoiesis (Supplementary Figure 6).

Clusters 2 and 3 also showed increased activity of CEBPD and IRF8, crucial for myeloid and dendritic cell differentiation and of ID4 and HTF4, which are involved in the establishment of the lymphoid lineage (Miyazaki et al., 2017) (Figure 3F-H and Supplementary Figure 6F). This points towards a common initial trajectory between the myeloid and lymphoid fate, consistent with our observations in scRNA-Seq data. Cluster 1 and 4 were characterised by a high level of activity of TFs of the NF-kB pathway (i.e., NF-kB1, NF-kB2, REL, and RELB), (Figure 3I,J and Supplementary Figure 6C and E), known to be involved in the regulation of HSCs maintenance and self-renewal (Zhao et al., 2012) (Espín-Palazón and Traver, 2016).

Integrating scRNA-Seq and scATAC-Seq data

Next, we wanted to map the cells from our scATAC-Seq data to specific cell types. Since, currently, no chromatin accessibility maps are available for human foetal HSPCs, we chose a strategy to integrate our scRNA-Seq and scATAC-Seq by mapping cells based on their gene body accessibility. We used a recently developed method which identifies pairwise correspondences (termed "anchors") between single cells across two different types of datasets, and their transformation into the shared space (Stuart et al., 2019). This approach allowed us to transfer scRNA-Seq derived annotations, learned by a classifier, onto scATAC-Seq data (see Methods).

We trained the classifier on CD34+ CD38- cells from the scRNA-Seq experiment using the six most abundant cell types (see Methods). Overall, ~57% of scATAC-Seq cells were assigned

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

to the HSC/MPP cluster, ~18% to HSC/MPPs-Cycle, ~5% to MEMPs, ~7% to MEMPs-Cycle, ~7% to GPs, and ~3% to LMPs. Cells with the prediction score lower than 40% were labelled as unclassified (~5%) (Figure 4A).

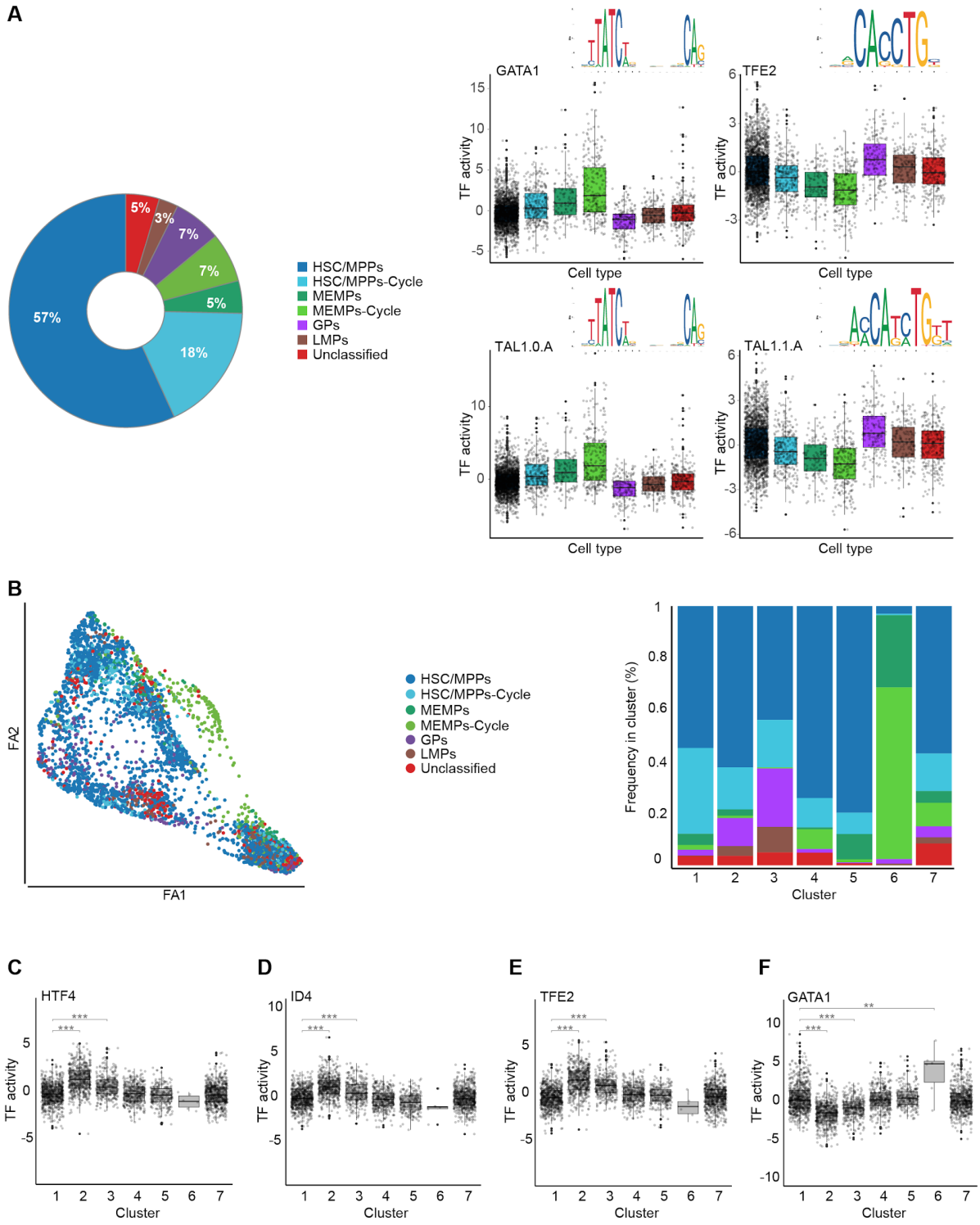


Figure 4 - Integration of scRNA-Seq and scATAC-Seq data. A. (Left) Donut plots showing the percentage of scATAC-Seq cells automatically assigned to different cell types. (Right) Boxplot showing

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the accessibility of GATA1, TFE2, TAL1.0.A, and TAL1.1.A motifs in the annotated cell types. On the top right of each boxplot, the TF sequence logos from JASPAR database similar to the analysed motifs are shown. **B.** (Left) FDG visualisation of the differentiation trajectory of scATAC-Seq cells coloured by their annotation. (Right) Barplot showing the percentage of cells within each cluster assigned to the annotated cell types. (C-D) Boxplots showing the accessibility of lineage-specific TF motifs in HSC/MPPs across the seven clusters. **C.** HTF4 ($p\text{-value}_{1,2}<2\times10^{-16}$; $p\text{-value}_{1,3}<2\times10^{-16}$). **D.** ID4 ($p\text{-value}_{1,2}<2\times10^{-16}$; $p\text{-value}_{1,3}=10^{-13}$). **E.** TFE2 ($p\text{-value}_{1,2}<2\times10^{-16}$; $p\text{-value}_{1,3}<2\times10^{-16}$). **F.** GATA1 ($p\text{-value}_{1,2}<2\times10^{-16}$; $p\text{-value}_{1,3}<2\times10^{-16}$; $p\text{-value}_{1,6}=0.0071$). p-value<0.001 (***); 0.001<p-value<0.01 (**); 0.01<p-value<0.05 (*); p-value≥0.05 (ns). Force-Directed Graph - FDG; ForceAtlas2 - FA2.

The frequency of assigned cell types in the scATAC-Seq data set was highly concordant with the ones from scRNA-Seq data (Supplementary Figure 3B) suggesting that overall the two modalities i.e. chromatin accessibility and transcriptome are correlated. To validate the cell type assignment of scATAC-Seq cells, we examined the accessibility of selected lineage specific TF motifs in each of the annotated cell types (Figure 4A). In line with the predicted annotations, the GATA1 motif showed the highest accessibility in MEMPs and MEMPs-Cycle, whereas TEF2 (known to play a role in myeloid and lymphoid differentiation, (Miyamoto et al., 2002)) was mostly active in GPs and LMPs. Confirming our earlier observation, two distinct TAL1 motifs had anticorrelated accessibility. TAL1.0.A was preferentially active in MEMPs and MEMPs-Cycle, while TAL1.1.A in GPs and LMPs (Figure 4A).

The Force Atlas representation of the classified scATAC-Seq cells revealed, however, considerable intermixing of different cell types across the trajectory with enrichment of MEMPs/MEMPs-Cycle in cluster 6 and to a lesser extent of GPs and LMPs in clusters 2 and 3 (Figure 4B). HSC/MPPs(Cycle) were distributed across all seven clusters but were most abundant in clusters 1, 4, and 5. This wide distribution of HSC/MPPs(Cycle) across multiple clusters within scATAC-Seq data suggested that, even though chromatin accessibility and the transcriptional state of foetal HSC/MPPs are correlated, there is extensive chromatin priming in the HSC/MPP population that results in their heterogeneity. To explore this further, we compared accessibility of selected lineage specific TF motifs in HSC/MPPs across the seven clusters (Figure 4C-F). We observed a low level of activity of all examined TFs in cluster 1 followed by a statistically significant increase of HTF4, ID4, and TFE2, and decrease of GATA1, in HSC/MPPs in clusters 2 and 3. GATA1 activity, however, increased in HSC/MPPs in cluster 6. Taken together, our data suggest that, within the transcriptionally homogeneous population of HSC/MPPs, there are significant differences in the activity of specific TFs that

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

may precede gene expression and mark initial priming of HSC/MPPs prior to their commitment to the specific lineage.

Validation of HSC/MPP identity and their differentiation capacity

Given the observed limitation of commonly used sorting markers to isolate pure progenitor populations, we devised a new FACS sorting strategy for HSC/MPPs based on the top 20 marker genes for this cluster in our scRNA-Seq dataset. The refined panel for HSC/MPPs included Lin- CD34+ CD38- CD52+ CD62L+ CD133+ (CD-REF from now on, Figure 5A).

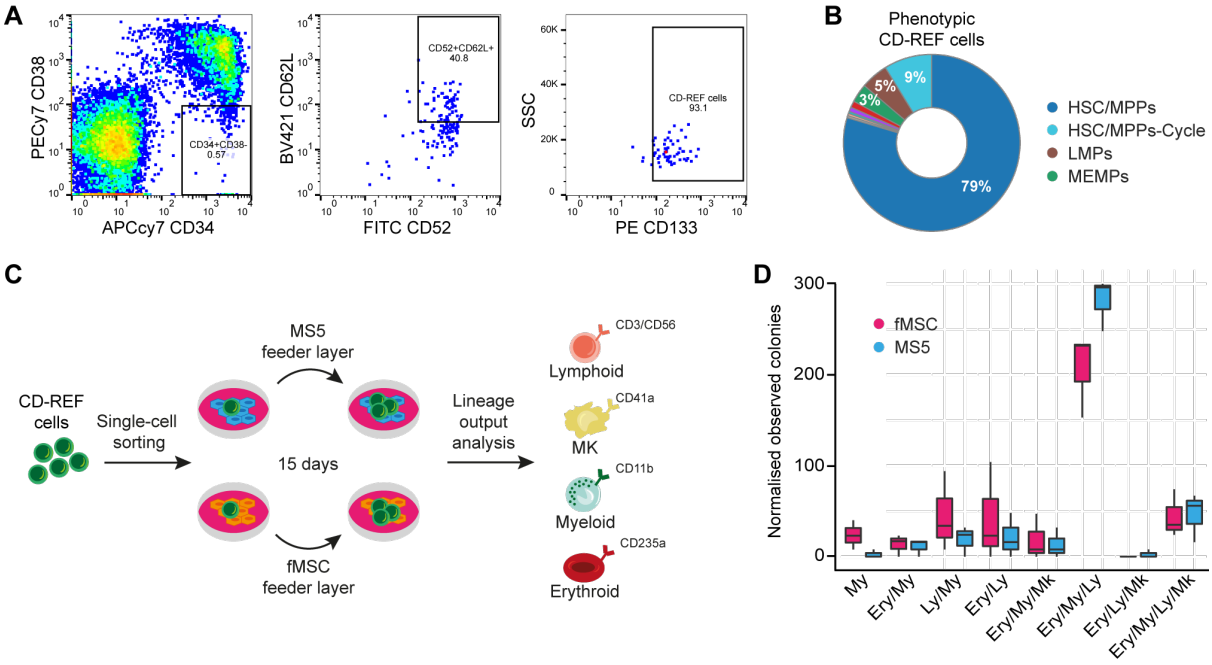


Figure 5 - Refining the sorting strategy to isolate foetal HSC/MPPs. A. Novel FACS panel (CD-REF panel) designed to increase the purity of the sorted HSC/MPP population. After excluding debris, doublets, and Lin+ cells, CD34+ CD38- CD52+ CD62L+ CD133+ were sorted. **B.** Donut plots showing the percentage of transcriptionally defined (i.e., manually curated) cell populations in the phenotypically defined CD-REF population. The colours correspond to the identified cell types. **C.** Schematic overview of the single-cell *in vitro* differentiation assay. Single CD-REF cells were sorted in liquid culture with either a mouse stromal cell line (MS5) or a human foetal primary feeder layer (fMSC). After 15 days of culture, lineage output was assessed by the expression of lineage markers CD41a (megakaryocytic), CD235a (erythroid), CD3/CD56 (lymphoid), and CD11b (myeloid) by flow cytometry. **D.** Normalised number of colonies observed *in vitro* that was assigned to a particular combination of cell lineages for the two feeder layers, n=3. My - myeloid, Ery - erythroid, Ly - lymphoid, Mk - megakaryocytic.

We FACS sorted cells from the femur of four foetuses using CD-REF panel and profiled them again by scRNA-Seq and single-cell *in vitro* differentiation assays. CD-REF cells on average

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

accounted for 40% (± 13%, *n*=4) of Lin- CD34+ CD38- cells in the femur, based on FACS analysis. The scRNA-Seq analysis of cells sorted with the refined panel showed that ~88% of CD-REF cells labelled HSC/MPP and HSC/MPP-Cycle clusters combined (Figure 5B). Thus, we achieved a significant enrichment of the HSC/MPP compartment compared to commonly used CD panels for HSCs (Lin- CD34+ CD38- CD45RA- CD90+ CD49f+) and MPPs (Lin- CD34+ CD38- CD90- CD45RA- CD49f- CD10- CD7-) where ~59% and ~73% respectively of sorted cells had a transcriptional signature of our most immature cell population (Figure 1C, Supplementary Figure 3B).

To assess the differentiation potential and robustness of the lineage output of CD-REF cells we sorted individual cells on either mouse MS5 feeder layer or on a more physiologically relevant, primary human foetal mesenchymal stem cells (fMSCs) (Figure 5C, see Methods section). After two weeks, 80% of cells sorted on MS5 and 85% of cells sorted on human foetal fMSCs generated colonies. In total, we analysed 302 colonies for their size and lineage output (erythroid-Ery, myeloid-My, megakaryocytic-Mk, lymphoid-Ly) using FACS (see Methods and Supplementary Figure 7A). Based on the median proportions of the uni-, bi-, tri-, and quadri-lineage colonies calculated from three experiments, our FACS analysis revealed that 12% of colonies on MS5 and 6% on fMSCs were quadri-lineage, 72% and 50% were tri-lineage, 14% and 36% were bilineage, and 10% and 8% were unilineage cells (solely giving rise to myeloid colonies), respectively (Supplementary File 1). Our finding that CD-REF cells indeed have multipotent potential is in line with our scRNAseq data and thus confirms they represent a highly enriched population of HSC/MPP.

We next compared the quantitative differences in the number of colonies assigned to each of the combinations of cell types, size of each colony (i.e., number of cells), and the relative contribution of each cell type to the colony, on MS5 *vs* fMSCs. We developed a novel mathematical pipeline based on Chi-squared (ChiSq) and Fisher's exact tests, applied on scaled frequency distributions to evaluate the stability of the distribution per cell combination and the individual contribution of each cell type to the global significance.

First, all frequency distributions were scaled for comparability. Using as an expected distribution a Random Uniform distribution on the number of colonies assigned to a particular combination of cell types, we observed a marginally increased presence of EryLyMkMy - and a significant increase in the number of EryMyLy colonies in MS5 compared to fMSCs but no significant differences in the colony sizes. In contrast, My and LyMy colonies were more frequent on fMSCs compared to MS5, albeit smaller, under a standard two-sided, unequal variance T-test (Figure 5D and Supplementary Figure 7B). There was no significant difference

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

in the proportion of cell types that were assigned to each of the cell type combinations except for EryMyLy (Supplementary Figure 7C). Overall our comparative analysis of lineage output and colony size of bone marrow derived CD-REF cells on MS5 and fMSCs confirmed their robust multipotent lineage output that is largely unaffected by the feeder layer.

Comparative analysis of HSC/MPP cells from different haematopoietic organs

Cells in the HSC/MPP cluster originated from the liver, femur, and hip. This provided a unique opportunity to assess potential qualitative and quantitative differences in the HSC/MPP population that originated from foetal liver or bone marrow. We first applied Fisher's exact test on the number of liver and femur cells in the different cell cycle states to determine whether there are non-random associations between the cycle state and the organ of origin (see Methods for further details). Interestingly, there was a statistically significant difference (p-value=4.25×10^{-9}) in the cell cycle state of cells in the HSC/MPP cluster between femur and liver (Figure 6A). Cells in the femur were predominantly in G1 (~70% of cells) compared to the same population in the liver (~52%) (Figure 6B). These data suggest that HSC/MPPs become more quiescent as they migrate from the liver to the bone marrow during the second trimester of human development. In line with this, HSC/MPP cells were significantly less frequent in femur compared to the liver (Figure 6D), as confirmed by Fisher's exact test on the total number of liver and femur cells (Figure 6E). This is in agreement with the increased proportion of phenotypic non-committed progenitors (CD34+ CD38-) found in the liver compared to the bone marrow (Figure 6C).

In order to evaluate if there is a statistically significant difference in the number of expressed genes between HSC/MPPs collected from the liver and femur, we used both the Kolmogorov–Smirnov (KS) and Mann–Whitney–Wilcoxon (MWW) test. We applied a subsampling strategy to downsample the cluster with more cells and balance the two distributions (see Methods). KS and MWW revealed a statistically significant decrease in the number of expressed genes in HSC/MPPs in the femur compared to the liver (Figure 6F-G). Differential expression analysis revealed that HSC/MPPs in the femur up-regulate genes involved in nucleosome assembly and chromatin silencing, such as *HIST1H1E*, *HIST1H2BN*, possibly marking their entry into quiescence (Figure 6H). This shift of HSC/MPPs from highly proliferative to quiescent as they migrate from foetal liver to bone marrow signifies the role of niche in modulation of HSC/MPPs behaviour.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

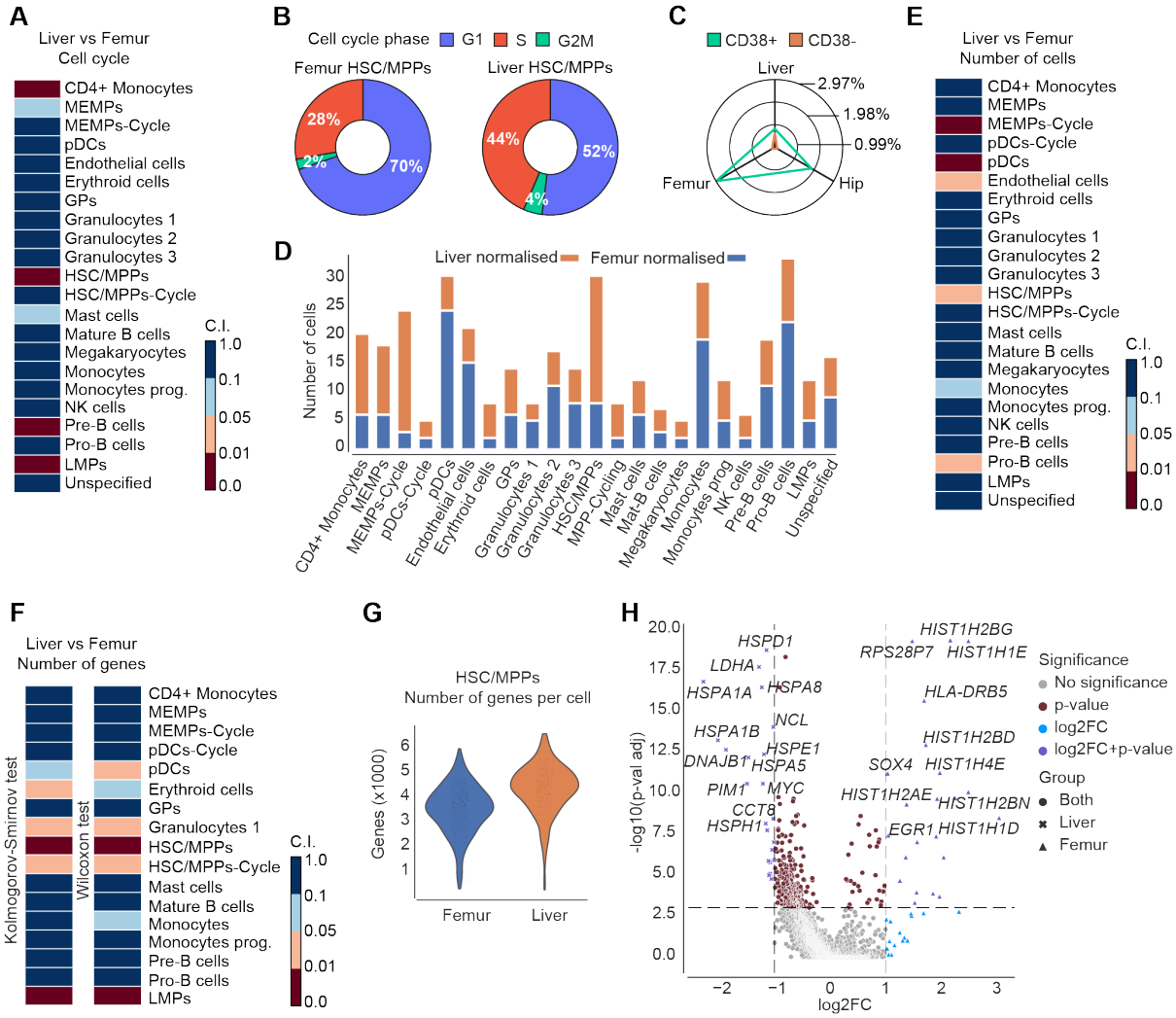


Figure 6 - Statistically significant differences between femur and liver cells across cell types. A. Heatmap showing the confidence interval of Fisher's exact test on the normalised number of different haematopoietic cell types sorted from liver and femur in G2M/S compared to G1. **B.** Donut plots displaying the percentage of cells in G1, G2M, and S in HSC/MPPs sorted from femur or liver. **C.** Radar chart representing the proportion of CD34+ CD38- and CD34+ CD38+ cells of total live cells present in the liver and bone marrow (femur and hip) n=15. **D.** Bar plot of the normalised distributions of the number of cells in each cell type sorted from liver or femur. **E.** Heatmap showing the confidence interval of Fisher's exact test on the normalised number of cells in each cell type collected from liver or femur **F.** Heatmaps depicting the confidence interval of the KS (left) and MWW (right) test on the number of expressed genes in each cell type collected from femur or liver cells. Notice that all the confidence intervals are split into 4 subintervals (i.e., [0, 0.01] strong statistically significant difference; (0.01, 0.05] statistically significant difference; (0.05, 0.1] marginal statistically significant difference; (0.1, 1] no statistically significant difference). **G.** Violin plot of the number of expressed genes in HSC/MPPs collected from the femur (blue) or liver (orange). **H.** Volcano plot showing DEGs in HSC/MPPs collected from femur or liver cells. The x-axis shows the log_2 fold-change (magnitude of change), while the y-axis shows the $-log_{10}$ adjusted p-value (statistical significance). We used the Wilcoxon rank-sum with the

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Benjamini-Hochberg correction. Colours represent the significance of the genes, both in terms of p-value and log_2 fold-change.

Discussion

Here, we present an integrative analysis of the single-cell transcriptome and chromatin accessibility of human foetal HSPCs. Our strategy involved plate-based sorting of well-defined immunophenotypic HSPCs from matched foetal liver and bone marrow. This approach enabled us to go beyond cataloguing heterogeneity of cellular states during foetal haematopoiesis and to: i) examine the extent to which phenotypic markers used over the last decade coincided with the true nature of the sorted foetal blood populations, ii) refine the sorting strategy for HSC/MPPs, iii) identify cell-cycle and gene expression differences between HSC/MPPs from foetal liver and bone marrow, iv) infer HSCPs differentiation trajectory, and v) explore lineage priming within the HCS/MPP population.

In doing so, we observed a striking level of heterogeneity in all immunophenotypic HSPCs, with more than ten transcriptionally-defined cell populations identified in each of the progenitor populations. Although this is consistent with previous studies of human adult and cord blood haematopoiesis (Knapp et al., 2018), it further emphasized the need for refining the sorting strategy for human foetal HSPCs. Our CD-REF panel achieved nearly 90% enrichment of HSC/MPPs, which we validated using single-cell *in vitro* differentiation assays and scRNASeq. CD-REF cells comprised 40% of all CD34+ CD38- cells in the foetal bone marrow with the majority of HSC/MPPs not cycling. The shift from highly proliferative state to quiescence coincided with the migration of HSC/MPPs from the foetal liver to bone marrow suggesting an important role of niche in modulation of HSC/MPPs behaviour. This is remarkably different from previous studies in mice, where extensive proliferation of HSPCs in the bone marrow continued up to three weeks after birth (Bowie et al., 2006).

Downstream of the HSC/MPPs, we identified three highly proliferative oligopotent progenitor populations (MEMPs, LMPs, and GPs). Integrative scRNA-Seq and scATAC-Seq analysis of HSC/MPPs and all main progenitor populations revealed correlation between chromatin accessibility and gene expression but also pointed out that, within transcriptionally homogeneous HSC/MPPs, there are multiple subpopulations that differed in their overall chromatin accessibility as well as lineage-specific TF activity. This indicates that within the HSC/MPP population, regulatory programmes permissive for different fates are being primed on the chromatin level, prior to their commitment to a specific lineage. The higher coordination of transcription and chromatin accessibility only occurred along the commitment of HSC/MPPs

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

towards MEMPs, implying a hierarchy of different levels of commitment in the foetal progenitor compartment with MEMP population being the most committed (compared to LMPs and GPs).

Overall, our study has provided a high resolution transcriptional and chromatin accessibility map of foetal HSPCs from liver and bone marrow that will be essential for further exploration of HSC/MPPs in the context of blood pathologies and for the purpose of regenerative medicine.

Methods

Ethics and Tissue acquisition

Human foetal bone and liver samples were obtained from 33 foetuses aged 17-22 pcw, following termination of pregnancy and informed written consent. The human foetal material was provided by the Joint MRC/Wellcome Trust (Grant MR/R006237/1) Human Developmental Biology Resource (http://www.hdbr.org), in accordance with ethical approval by the NHS Research Health Authority, REC Ref: 18/LO/0822.

Tissue processing

Tissues were kept in cold DMEM medium (Invitrogen) until dissection and processed on the same day of collection. Single-cell suspensions were generated from matched foetal liver and bone tissues after rinsing them with cold PBS (Gibco). Liver samples were passed through a 70 µm strain into a falcon tube prefilled with cold PBS. Bone marrow from long bones was isolated by flushing cold PBS into the diaphysis and collected into a falcon tube. Bone marrow from hip bone was collected by dissecting the bone with a sterile scalpel and flushing cold PBS in the marrow cavity into a falcon tube. The suspension obtained from long bones and hip bones was then passed through a 70 µm strain into a new falcon tube. Cells were then centrifuged for 5 minutes at 300 g, 4°C and the pellet was resuspended into the RBC lysis buffer (eBioscience) for 2 minutes at room temperature, after which 20 ml of cold PBS were added to stop the lysis reaction. RBC step was not performed when sorting erythroid cells. Live cell enrichment was performed using MACS columns (Miltenyi Biotec - 130-090-101) following the manufacturer's instructions. When sorting CD34+ or CD45+ cells, column enrichment was performed using MACS columns (Miltenyi Biotec - 130-046-702 and 130-045-801 respectively for CD34+ and CD45+ cells), following the manufacturer's instructions.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fluorescence-activated cell sorting

Cells were stained with antibody cocktails (Supplementary Table 3) in a total volume of 100 µl 5% FBS (Gibco) in PBS for 30 minutes at 4°C, centrifuged for 5 minutes at 300 g, 4°C, resuspended in a final volume of 500 µl of 5% FBS in PBS and subsequently filtered into polypropylene FACS tubes (ThermoFisher). For scRNA-Seq experiments, single cells were index sorted using a BD Influx Sorter into wells of 96-well plates (4titude) prefilled with 2 µl of lysis buffer consisting of 0.2% Triton X-100 (Sigma) and 1 U/µl RNAse inhibitor (Life Technologies) in nuclease-free water (Invitrogen). For scATAC-Seq experiments, 5,000 - 20,000 cells were sorted using a BD Influx machine into 1.5 ml tubes (Eppendorf). Following bulk tagmentation with Tn5 (Chen et al., 2018), single nuclei were index sorted in wells of 384-well plates (Eppendorf) prefilled with 2 µl of lysis buffer consisting of 0.2% SDS, 20 µg/ml proteinase K (Ambion), 50 mM Tris-HCl (Gibco) and 50mM NaCl (Sigma) in nuclease-free water.

Library preparation

The Smart-Seq2 method (Picelli et al., 2014) was used for library preparation for the scRNA-Seq experiments, with some modifications as described in (Macaulay et al., 2016). The quality of libraries was evaluated with Bioanalyzer (Agilent). Good-quality libraries were subsequently quantified with KAPA Library Quantification Kit (Roche) and submitted for sequencing. Library preparation for the scATAC-Seq experiments was performed using a recently described method (Chen et al., 2018). Library traces were evaluated using Bioanalyzer.

Sequencing

Libraries for scRNA-Seq experiments were multiplexed using Nextera Index sets A, B, C, and D (v.2, Illumina) and sequenced on HiSeq4000 and NovaSeq6000 (Illumina) in pair-end mode, with an interquartile range (IQR) of 697,427 uniquely mapped reads (average: 666,632; standard deviation: 557,274). Libraries for scATAC-Seq experiments were sequenced on HiSeq4000 in pair-end mode, with a mean read count of 473,886 and IQR 341,210.

Upstream analysis: alignment and quantification of scRNA-Seq data

Smart-Seq2 sample demultiplex fastq files were quality checked, aligned and quantified by using the scRNA-Seq pipeline. This pipeline is based on STAR with default parameters (v.2.5.4a) (Dobin et al., 2013) index and annotation from the Ensembl release 91 of the GRCh38 human reference genome. Transcript and gene counts were quantified using the option *quantMode GeneCounts* provided by STAR. Since we used different sets of well-

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

defined antibodies to isolate different cell types, we applied specific thresholds for each sample to filter out both the cells and genes (Supplementary Table 2). We detected on average 3,642 genes per cell (IQR: 2,239; standard deviation: 1,621).

Downstream analysis of scRNA-Seq data: quality control, batch correction, clustering, marker genes, and annotation

We performed the downstream analysis of scRNA-Seq using the Python (v.3.6.9) package SCANPY (v.1.4.5.1) (Wolf et al., 2018). Our pipeline included: 1) a QC step (number of identified counts and number of expressed genes using the *filter_cells* function, and fraction of mitochondrial genes), 2) removing the genes expressed in less than 10 cells (*filter_genes* function), 3) data normalisation (*normalize_per_cell* function with scaling factor 10,000 and *log1p* function), 4) detection of the top 1,000 highly variable genes (HVGs) (*highly_variable_genes* function, in which the HVGs were selected separately within each batch and then merged, where each batch corresponds to a specific sample), 5) scaling of the features to unit variance and zero mean (*scale* function with max_value equal to 10), 6) application of scAEspy on the HVG space by considering the raw expression (i.e., counts) (Tangherloni et al., 2019), 7) batch correction by sample applying BBKNN algorithm (v.1.3.6, *bbkkn* function with use_faiss equal to false, approx equal to false and the Euclidean distance) to the latent space (16 components) generated by the used AE, 8) Leiden algorithm (*leiden* function with resolution equal to 2.2) applied to the neighbourhood graph generated by BBKNN. The 27 obtained clusters were manually annotated by considering the merged data using well-known cell type specific genes and the Differentially Expressed Genes (DEGs). DEGs were computed by using *rank_genes_groups* function (Wilcoxon rank-sum with adjusted p-values for multiple testing with the Bonferroni correction), which compares each cluster to the union of the rest of the clusters. The clusters that either did not express specific cell type genes or expressed marker genes of different cell types had been iteratively subclustered. Specifically, we applied the Leiden algorithm (*leiden* function with resolution equal to 0.5) to subcluster Endothelial cells, obtaining four distinct clusters: the first two clusters have been annotated as Monocytes 2, the third as NK cells and the fourth as Endothelial cells. Finally, we used the Leiden algorithm (*leiden* function with resolution equal to 0.5) to cluster the Unspecified cluster getting four clusters. We merged three clusters with the HSC/MPP cluster while one was annotated as Unspecified.

Dimensionality reduction and trajectory analysis

After the detection of the first 1,000 HVGs, we applied scAEspy to HVG space by setting alpha and lambda equal to 0 and 2, respectively, in order to obtain the Gaussian Mixture Maximum

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Mean Discrepancy Variational AE (GMMMDVAE) (Tangherloni et al., 2019). We run GMMMDVAE for 100 epochs with a batch size equal to 100, one hidden layer of 64 neurons, a latent space of 16 neurons, 15 Gaussian distributions, learnable prior distribution, constrained Poisson loss function, and sigmoid activation function. Then, we applied BBKNN to the latent space (16 components) to generate the neighbourhood graph by identifying top neighbours of each cell in each batch separately. We applied UMAP (v.0.3.10, SCANPY *umap* function with random_state equal to 8 and n_components equal to 3) to the obtained neighbourhood graph. PAGA and Force-Directed Graph (FDG) were applied to infer the development trajectories. We removed the endothelial cells and recalculated the neighbourhood graph (*neighbors* function with n_neighbors equal to 30) on the latent space (16 components) to exploit the data before batch correction (Luecken and Theis, 2019). We computed the PAGA graph (*paga* function with model equal to v1.2) and the ForceAtlas2 (FA2) using PAGA-initialization (*draw_graph* function, which exploits the FA2 class from fa2 (v.0.3.5) Python package, using the HSC/MPP cluster as root and maxiter equal to 1,000).

Differential expression analysis

Following cluster annotation, we performed biologically-relevant pairwise DE tests between pairs of clusters to identify DEGs and to examine the quantitative changes in the expression levels between the clusters. Specifically, we tested MPPs against MPPs-Cycle, MEMPs against MEMPs-Cycle, MPPs against MEMPs, MPPs against LMPs, MPPs against GPs, and MEMPs against LMPs. In order to cope with the unbalanced distributions between two groups of cells, due to the different number of cells in each cluster, we used the following subsampling strategy. Given two groups of cells, the biggest group was randomly subsampled taking a number of cells equal to the number of cells composing the smallest group. For each gene, a two-sided T-test for the means of two independent samples (i.e., biggest group and subsampled one) was applied. We used the *ttest_ind* function (equal_var equal to false) provided by the Python SciPy (Virtanen et al., 2020) package (v.1.4.1). Since we did not assume that the two groups have identical variances, the Welch's t-test was automatically applied. Then, we calculated the median of the p-values of these T-tests. We applied this subsampling strategy 1001 times and calculated the median of the medians to select the subset of the biggest group to run the DE analysis.

For a given subset of cells from the biggest group and the smallest one, we calculated the DEGs by applying the *rank_genes_groups* function (Wilcoxon rank-sum with adjusted p-values for multiple testing with the Benjamini-Hochberg correction). Then, we filtered out the obtained DEGs by using the *filter_rank_genes_groups* function (min_in_group_fraction equal to 0.3 and max_out_group_fraction equal to 1, so that a gene is expressed in at least 30% of

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the cells in both the tested groups; min_fold_change equal to 0). Following the aforementioned workflow, we compared cells from liver and femur from the same cluster. Finally, we analysed HSC/MPPs and HSC/MPPs-Cycle to see which genes contributed to the observed difference between cells from femur and liver.

Cell type classification

We trained both a Random Forest classifier (Pedregosa et al., 2011) and a DNN to predict the cell types by considering the top 5, 10, 20, 30, 50, and 100 marker genes for each cluster using the *log*-normalised expression. Since some marker genes are shared among the clusters, we considered them only once to avoid duplicated columns in the feature matrices. We merged the following clusters: HSC/MPPs and HSC/MPPs-Cycle as HSC/MPPs, MEMPs and MEMPs-Cycle as MEMPs, Granulocytes 1, Granulocytes 2, and Granulocytes 3 as Granulocytes; pDCs and pDCs-Cycle as pDCs; CD4+ Monocytes, Monocytes, and Monocyte Prog as Monocytes; Pre-B cells, Pro-B cells, and Mature B cells as B cells. Thus, we obtained 14 distinct clusters.

We used the RandomForestClassifier (n_estimators equal to 100 and Gini criterion) provided by Scikit-learn (Pedregosa et al., 2011) (v.0.21.2). We developed the DNN by using Keras[1] (v.2.2.4) with Tensorflow (Abadi et al., 2016) (v.1.12.0) as backend. The network is composed of 2 dense hidden layers of 64 and 32 neurons, respectively. We added a dropout (50%) layer before the first layer as well as a dropout (30%) layer before the second layer. We trained the DNN for 1,000 epochs using the Adam optimiser (Kingma and Ba, 2014) by minimising the categorical cross-entropy loss function. We also set an early stopping with 100 epochs as patience to avoid overfitting.

We applied a stratified 10 fold cross-validation (Scikit-learn *StratifiedKFold* function) resampling procedure to evaluate both the Random Forest and DNN. The Random Forest achieved the best result when the top 30 marker genes per cluster were used (mean accuracy equal to 88.43% and standard deviation equal to 1.03%;), while the DNN considering the top 30 (mean accuracy equal to 90.03% standard deviation equal to 1.14%) and 50 marker genes per cluster (90.28% and standard deviation equal to 1.28%).

As a further test, we evaluated the ability of our DNN to generalise on unseen data. We split the dataset into a train set (80%) and a test set (20%) (Scikit-learn *train_test_split* function with test_size equal to 0.2). We then divided the train set into a train set (85%) and a validation set (15%, (*train_test_split* function with test_size equal to 0.15). We trained our DNN with the

[1] Chollet *et al.*: https://keras.io

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

train set, validating it using the validation set. When we took into account the top 30 marker genes, we achieved an accuracy equal to 90.39% on the validation set. When considering the top 50 marker genes the accuracy was 90.76%. Finally, we predicted the labels of the test set by obtaining an accuracy equal to 90.23% (30 marker genes) and 89.23% (50 marker genes).

Upstream analysis and quality control of scATAC-Seq data

We performed the upstream analysis using the samtools (Li et al., 2009) (v1.9), bedtools (Quinlan, 2014) (v2.27.1), Picard tools[2] (v2.9.0) and BWA (Li and Durbin, 2009) (v0.7.17). First we aligned fastq files to the GRCh38 reference genome (average 473,886 reads per cell), followed by marking duplicates with *MarkDuplicates* function from Picard tools and removing duplicates using samtools *view* with -F 1804 parameter per each cell. Overall with average duplicates rate 77% we obtained 91,554 reads per cell after removing duplicates. Next, we transformed bam files to bed files using *bamtobed* bedtools function in bedpe mode and kept only fragments that are not bigger then 1000 bp using custom script. We called peaks (for the clusters with more than 50 cells) using the SnapATAC approach (Fang et al., 2019) with macs2 (Zhang et al., 2008) parameters "--nomodel --shift 100 --ext 200 --qval 5e-2 -B" and obtained 152,283 peaks. Importantly, for the downstream analysis in R we binarized counts per cell using Signac[3] *BinarizeCounts* function, resulting in 32,217 fragments per cell on average.

The downstream analysis was done in R 3.6.1 applying Seurat (Butler et al., 2018) (Stuart et al., 2019) (v3.1.4), Signac (v0.2.4), chromVAR (Schep et al., 2017) (v.1.8) and Harmony (Korsunsky et al., 2019) (v1.0). The pipeline included a QC step (duplicates removal, number of fragments, fragments per peak, fraction of reads mapping to blacklist regions, nucleosome signal, and transcriptional start site (TSS) enrichment), application of LSI dimensionality reduction to the three samples independently (*RunTFIDF* function with method equal to 2, *FindTopFeatures* function setting min.cutoff to q0, and *RunSVD* function using the peaks as assay), batch correction by sample, lane, and organ applying Harmony on the first 50 LSI components, excluding the first one, (*RunHarmony* function setting assay.use to peaks, max.iter.harmony to 20, max.iter.cluster to 200, sigma to 0.25, and theta to 2, 4, 4 in order to weight more the batch related to samples). TF activities on the ATAC-seq data were calculated using the Signac implementation of chromVAR using the *RunChromVAR* function taking as tested motifs dataset from HOCOMOCO (Kulakovskiy et al., 2018) v11 human TF binding models database (769 TFs).

[2] Broad Institute: http://broadinstitute.github.io/picard/
[3] Stuart *et al.*: https://github.com/timoast/signac/

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Dimensionality reduction of scATAC-Seq data and trajectory inference

We applied the UMAP algorithm to the first 50 LSI components corrected by Harmony (*RunUMAP* function with umap.method equal to uwot and n.neighbors equal to 10, *FindNeighbors* function setting annoy.metric to cosine). We identified seven distinct clusters by using the Seurat function *FindClusters* (resolution equal to 0.5). We inferred the development trajectories by applying PAGA and FDG. We recalculated the neighbourhood graph using the SCANPY *neighbors* functions (n_neighbors equal to 30) on the 50 LSI components corrected by Harmony. We computed the PAGA graph (SCANPY *paga* function with model equal to v1.0) and used it to initialise the FA2 algorithm (SCANPY *draw_graph* function using cluster 1 as root and maxiter equal to 1,000).

Integration of scRNA-Seq and scATAC-Seq datasets

We integrated scRNA-Seq and scATAC-Seq data using a recently developed method by Stuart *et al.* (Stuart et al., 2019). Namely, we used our scRNA-Seq data as reference dataset to train the classifier and automatically assign a cell type to each scATAC-Seq cell. The training of the classifier was performed using 511 CD34+ CD34- cells from our scRNA-Seq experiment. In order to have a suitable number of cells for each cell type to train the classifier, we considered scRNA-Seq clusters with at least 20 cells (i.e., HSC/MPPs, HSC/MPPs-Cycle, MEMPs, MEMPs-Cycle, GPs, and LMPs). We generated a gene expression matrix from our scATAC-Seq data set by assigning each peak to the gene by considering the genome coordinates of the gene body ± 3 kb. We applied the Seurat function *FindTransferAnchors* (query.assay equal to RNA_promoter, features equal to the counts of the RNA_promoter, and k.anchor equal to 6) on the Canonical Correlation Analysis (CCA) space because it was more suitable, compared to the LSI space, for capturing the shared feature correlation structure between scRNA-Seq and scATAC-Seq data. We assigned the cell types to the scATAC-Seq cells by applying the Seurat *TransferData* on the first 50 LSI components corrected by Harmony considering the calculated anchors (refdata equal to the six scRNA-Seq clusters). In order to avoid assignments based on a low score, all cells with the prediction score lower than 40% (the value of a uniform distribution of six clusters is 16,67%) were labelled as unknown.

Isolation of human foetal MSCs

Human primary fMSCs were isolated from the femur of a 19 pcw sample following an established protocol used for mouse bones (Perpétuo et al., 2019). Briefly, the bone was rinsed in PBS and the bone epiphyses cut with a scalpel. The bone marrow was flushed with 50 ml PBS, centrifuged at 300 g for 5 minutes, resuspended in alphaMEM medium (Thermo

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fisher Scientific) supplemented with 2 mM L-glutamine (Thermo Fisher Scientific), 100 U/ml penicillin/streptomycin (Thermo Fisher Scientific) and 10% fetal bovine serum (Sigma) at a concentration of $5x10^6$ cells/ml and cultured at 37° at 5% CO_2. After 24 hours, floating cells were removed by washing twice with PBS and medium was changed twice a week until the culture was 70% confluent. Cells were cryopreserved until use.

Single-cell in vitro culture

Single Lin- CD34+ CD38- CD62L+ CD52+ CD133+ cells, isolated from the foetal bone marrow of four different foetuses (20-22 pcw), were index-sorted into 96-well plates seeded with fMSCs or MS5 (obtained by DSMZ) and supplemented with cytokines as previously described (Velten et al., 2017). Cells were cultured for 15 days at 37° at 5% CO2. At the end of the culture, colonies were filtered to exclude feeder layer cells, and their lineage output was assessed by the expression of CD41a (megakaryocytic-Mk), CD235a (erythroid-Ery), CD3/CD56 (lymphoid-Ly), and CD11b (myeloid-My) by flow cytometry using a BD LSR-Fortessa analyser. Colonies were considered positive for a lineage if ≥ 30 cells were detected in the relative gate.

Statistical analysis of *in-vitro* differentiation assays

The analysis of the colonies comprises of three stages, all considering both the qualitative and quantitative aspects: [1] probability of observing the distribution of cells under a random uniform hypothesis, [2] stability of the observed distributions across replicates and between feeder layers, [3] contribution of particular a cell type to the observed stability.

[1] To calculate the probability of observing the number of cells in one category we used ChiSq and Fisher's exact tests. The random uniform hypothesis for the test is to assign equal probabilities for one (Ery, Ly, Mk, and My), two (EryLy, EryMk, EryMy, LyMk, LyMy, MkMy), three (EryLyMk, EryLyMy, EryMKMy, LyMkMy), and four (EryLyMkMy) lineage combinations and equal probabilities within each subset; namely equal probabilities for each entry for the single-cell annotation, ditto for each distinct entry (i.e., a combination of cells {X, Y} is equivalent to {Y, X}) for the two cell combinations, ditto for each distinct three cell combination and one entry for all four cell combination. The observed distributions are scaled to the same total as the expected. The expected vs observed distributions are compared using a standard ChiSq test (default parameters). Because the ChiSq significance output is weighted, averaged result of several factors, in addition, for each combination of cells, we calculate a Fisher's Exact test to assess whether, individually, the combination of cells matches the expected

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

probability or not. -1 is assigned by default to the output of the Fisher's exact test, indicating an impossible combination of cells (i.e., not obtained in culture).
[2] To assess the stability (within each replicate) of distributions across the wells we applied the ChiSq and Fisher's exact tests on the scaled frequency distributions per colony. For this evaluation, the number of pseudo-replicates corresponds to the number of colonies grown from single cells, i.e., for the MS5 layer 50 colonies (rep1, rep2), 48 colonies (rep3), for the fMSC layer 50 colonies (rep1), 34 colonies (rep2), 70 colonies (rep3). All pairwise comparisons between colony-distributions for each replicate are performed. The distributions of p values, split per cell identity, are represented using standard boxplots (the IQR corresponds to 25%-50%-75%, and the whiskers extend to 5% and 95% respectively). [3] The contribution of a particular cell type is evaluated using Fisher's exact tests. For each pairwise comparison and for each cell type the 2x2 contingency table is created on the scaled frequency of the cell type and the complement to 1 for the specific cell type. The results are summarised using boxplots, split per cell type and annotation.

The variation between biological replicates (3 MS5 and 3 fMSCs) was assessed using two-sided, unequal variance t-tests on the scaled distributions, per comparable replicates, i.e., replicate 2 and 3 were matched between MS5 and fMSCs, replicate 1 comprised only of cells grown on the MS5 feeder layer and replicate 4 comprised only of cells grown on the fMSC feeder layer.

Qualitative and quantitative characterisation of statistically significant differences across cell types

In order to assess qualitative and quantitative differences between the haematopoietic cells collected from the liver and femur, we implemented different statistical tests. For each cluster, we calculated if there is a statistically significant difference in the number of cells (Test 1), the number of expressed genes per cell (Test 2), and the cell cycle state of blood cells collected from liver and femur (Test 3).

Test 1. Since we used different gates to sort cells and we sorted a different number of cells in each experiment, we first normalised the number of cells from liver and femur. We selected only the matched gates (i.e., the gates where we sorted haematopoietic cells from both liver and femur). Then, we selected cells from the liver (or femur) from each gate in each of the clusters. For each cluster, we normalised the number of cells inside the cluster in the range [0, 100] by dividing the number of cells for the total number of cells of the gate in order to obtain a number of cells equal to 100. Next, for each cluster, we calculated the median of the cells in the liver (or femur) among the different gates. In order to evaluate if there is a

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

statistically significant difference between the number of cells in the liver and femur considering all the clusters, we applied the ChiSq test by normalising the distributions (i.e, the median of the gates of each cluster) of the cells from liver and femur among the clusters. We applied the *chi2_contingency* function provided by the Python SciPy. Since the obtained p-value is equal to 1.02×10^{-4}, we applied Fisher's exact test (SciPy *fisher_exact* function) to each cluster to find which clusters contributed to the difference.

Test 2. In this test, we evaluated the number of expressed genes between cells collected from femur and liver. In order to remove possible technical effects for each cell, we divided the number of expressed genes by the number of reads uniquely mapped against the reference genome. For each cluster, we applied both the KS test (SciPy *ks_2samp* function) and the MWW test (SciPy *mannwhitneyu* function). Since the number of cells from femur and liver is very different in any given cluster (giving rise to unbalanced distributions) we used a subsampling strategy similar to that used for the DE analysis. We randomly subsampled the biggest group 1,001 times taking a number of cells equal to the number of cells composing the smallest group. We applied the KS (and MWW) test comparing the smallest group to the subsampled ones obtaining a distribution of p-values. Finally, we calculated the median of this distribution of p-values to evaluate if there is a statistically significant difference between the number of expressed genes in the cells from the liver and femur. Note that we excluded the clusters where the number of the cells from femur or liver was lower than 20.

Test 3. For each cluster, we compared G2M/S and G1 states by normalising the number of cells from the liver and femur in the two states. We applied Fisher's exact test (SciPy *fisher_exact* function) to each cluster to find a possible statistically significant difference between the number of cells in G2M/S and G1 states in the liver and femur.

Acknowledgements

The authors would like to thank the WTSI Cytometry Core Facility for their help with single-cell index sorting and the WTSI DNA pipelines and the CRUK Cambridge Institute Genomics Core Facility for their contribution in sequencing the data. We would also like to thank the CellGen bioinformatics team for their help with pre-processing of scRNA-Seq data and Jana Eliasova for her precious support and help with the illustrations. Finally, we would like to thank the Human Developmental Biology Resource (HDBR) for providing samples.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Funding

The study was supported by European Research Council project 677501 – ZF_Blood (to A.C. and A.M.R), EMBO small grant (to A.C.) and a core support grant from the Wellcome Trust and MRC to the Wellcome Trust – Medical Research Council Cambridge Stem Cell Institute.

Authors contributions

A.C. and A.M.R. conceived the study; A.M.R. performed all the experiments with help from B.M., P.M.S., J.X., and E.P.; A.T. carried out the computational analysis of the scRNA-Seq data; S.G.R. carried out parts of the computational analysis, under the supervision of A.T.; I.B. performed the analysis of the scATAC-Seq data and the integrative analysis of scATAC-Seq and scRNA-Seq, under the supervision of J.B.Z.; I.M. carried out the statistical analysis of *in-vitro* experiments; A.C., A.M.R, and A.T. designed the figures and wrote the manuscript with inputs from the other authors. All authors approved the final version of the manuscript.

Data availability

The raw RNA-Seq data (i.e., fastq files) and cell assignment are deposited at ArrayExpress with accession code TBA, while the raw ATAC-Seq data (i.e., fastq files) and cell assignment are deposited at ArrayExpress with accession code TBA.

Code availability

All scripts, functions, and Jupyter Notebook developed for this study are freely available on GitLab: https://gitlab.com/cvejic-group/integrative-scrna-scatac-human-foetal. The repository also contains the gene expression and fragment matrices.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. ArXiv.

Akashi, K., Kondo, M., Cheshier, S., Shizuru, J., Gandy, K., Domen, J., Mebius, R., Traver, D., and Weissman, I.L. (1999). Lymphoid development from stem cells and the common lymphocyte progenitors. Cold Spring Harb. Symp. Quant. Biol. *64*, 1–12.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. *2008*, P10008.

Bowie, M.B., McKnight, K.D., Kent, D.G., McCaffrey, L., Hoodless, P.A., and Eaves, C.J. (2006). Hematopoietic stem cells proliferate until after birth and show a reversible phase-specific engraftment defect. J. Clin. Invest. *116*, 2808–2816.

Buenrostro, J.D., Corces, M.R., Lareau, C.A., Wu, B., Schep, A.N., Aryee, M.J., Majeti, R., Chang, H.Y., and Greenleaf, W.J. (2018). Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. Cell *173*, 1535-1548.e16.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420.

Calvanese, V., Nguyen, A.T., Bolan, T.J., Vavilina, A., Su, T., Lee, L.K., Wang, Y., Lay, F.D., Magnusson, M., Crooks, G.M., et al. (2019). MLLT3 governs human haematopoietic stem-cell self-renewal and engraftment. Nature *576*, 281–286.

Chen, X., Miragaia, R.J., Natarajan, K.N., and Teichmann, S.A. (2018). A rapid and robust method for single cell chromatin accessibility profiling. Nat. Commun. *9*, 5345.

Collin, M., Bigley, V., Haniffa, M., and Hambleton, S. (2011). Human dendritic cell deficiency: the missing ID? Nat. Rev. Immunol. *11*, 575–583.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Doulatov, S., Notta, F., Eppert, K., Nguyen, L.T., Ohashi, P.S., and Dick, J.E. (2010). Revised

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development. Nat. Immunol. *11*, 585–593.

Espín-Palazón, R., and Traver, D. (2016). The NF-κB family: Key players during embryonic development and HSC emergence. Exp. Hematol. *44*, 519–527.

Fang, R., Preissl, S., Hou, X., Lucero, J., Wang, X., Motamedi, A., Shiau, A.K., Mukamel, E.A., Zhang, Y., Behrens, M.M., et al. (2019). Fast and Accurate Clustering of Single Cell Epigenomes Reveals *Cis* -Regulatory Elements in Rare Cell Types. BioRxiv.

Franco, C.B., Chen, C.-C., Drukker, M., Weissman, I.L., and Galli, S.J. (2010). Distinguishing mast cell and granulocyte differentiation at the single-cell level. Cell Stem Cell *6*, 361–368.

Gaspar-Maia, A., Alajem, A., Meshorer, E., and Ramalho-Santos, M. (2011). Open chromatin in pluripotency and reprogramming. Nat. Rev. Mol. Cell Biol. *12*, 36–47.

Haghverdi, L., Buettner, F., and Theis, F.J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics *31*, 2989–2998.

Hsu, H.L., Huang, L., Tsan, J.T., Funk, W., Wright, W.E., Hu, J.S., Kingston, R.E., and Baer, R. (1994). Preferred sequences for DNA recognition by the TAL1 helix-loop-helix proteins. Mol. Cell. Biol. *14*, 1256–1265.

Hu, M., Krause, D., Greaves, M., Sharkis, S., Dexter, M., Heyworth, C., and Enver, T. (1997). Multilineage gene expression precedes commitment in the hemopoietic system. Genes Dev. *11*, 774–785.

Ito, K., and Suda, T. (2014). Metabolic requirements for the maintenance of self-renewing stem cells. Nat. Rev. Mol. Cell Biol. *15*, 243–256.

Ivanovs, A., Rybtsov, S., Ng, E.S., Stanley, E.G., Elefanty, A.G., and Medvinsky, A. (2017). Human haematopoietic stem cell development: from the embryo to the dish. Development *144*, 2323–2337.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS ONE *9*, e98679.

Kassouf, M.T., Hughes, J.R., Taylor, S., McGowan, S.J., Soneji, S., Green, A.L., Vyas, P., and Porcher, C. (2010). Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. Genome Res. *20*, 1064–1083.

Katsumura, K.R., Bresnick, E.H., and GATA Factor Mechanisms Group (2017). The GATA factor revolution in hematology. Blood *129*, 2092–2102.

Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. ArXiv.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Knapp, D.J.H.F., Hammond, C.A., Hui, T., van Loenhout, M.T.J., Wang, F., Aghaeepour, N., Miller, P.H., Moksa, M., Rabu, G.M., Beer, P.A., et al. (2018). Single-cell analysis identifies a CD33+ subset of human cord blood cells with high regenerative potential. Nat. Cell Biol. *20*, 710–720.

Komorowska, K., Doyle, A., Wahlestedt, M., Subramaniam, A., Debnath, S., Chen, J., Soneji, S., Van Handel, B., Mikkola, H.K.A., Miharada, K., et al. (2017). Hepatic Leukemia Factor Maintains Quiescence of Hematopoietic Stem Cells and Protects the Stem Cell Pool during Regeneration. Cell Rep. *21*, 3514–3523.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296.

Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., et al. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. *46*, D252–D259.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. *15*, e8746.

Macaulay, I.C., Svensson, V., Labalette, C., Ferreira, L., Hamey, F., Voet, T., Teichmann, S.A., and Cvejic, A. (2016). Single-Cell RNA-Sequencing Reveals a Continuous Spectrum of Differentiation in Hematopoietic Cells. Cell Rep. *14*, 966–977.

Mikkola, H.K.A., and Orkin, S.H. (2006). The journey of developing hematopoietic stem cells. Development *133*, 3733–3744.

Miyamoto, T., Iwasaki, H., Reizis, B., Ye, M., Graf, T., Weissman, I.L., and Akashi, K. (2002). Myeloid or lymphoid promiscuity as a critical step in hematopoietic lineage commitment. Dev. Cell *3*, 137–147.

Miyazaki, M., Miyazaki, K., Chen, K., Jin, Y., Turner, J., Moore, A.J., Saito, R., Yoshida, K., Ogawa, S., Rodewald, H.-R., et al. (2017). The E-Id Protein Axis Specifies Adaptive Lymphoid

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Cell Identity and Suppresses Thymic Innate Lymphoid Cell Development. Immunity *46*, 818-834.e4.

Morisot, S., Georgantas, R.W., and Civin, C.I. (2006). 345. Hematopoietic Stem-Progenitor Cells Express CD52 mRNA and Membrane Protein. Mol. Ther. *13*, S131–S132.

Nimmo, R.A., May, G.E., and Enver, T. (2015). Primed and ready: understanding lineage commitment through single cell analysis. Trends Cell Biol. *25*, 459–467.

Notta, F., Zandi, S., Takayama, N., Dobson, S., Gan, O.I., Wilson, G., Kaufmann, K.B., McLeod, J., Laurenti, E., Dunant, C.F., et al. (2016). Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. Science *351*, aab2116.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., et al. (2015). Transcriptional heterogeneity and lineage commitment in myeloid progenitors. Cell *163*, 1663–1677.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J Mach Learn Res *12*, 2825–2830.

Perpétuo, I.P., Bourne, L.E., and Orriss, I.R. (2019). Isolation and generation of osteoblasts. Methods Mol. Biol. *1914*, 21–38.

Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181.

Polański, K., Young, M.D., Miao, Z., Meyer, K.B., Teichmann, S.A., and Park, J.-E. (2020). BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics *36*, 964–965.

Popescu, D.-M., Botting, R.A., Stephenson, E., Green, K., Webb, S., Jardine, L., Calderbank, E.F., Polanski, K., Goh, I., Efremova, M., et al. (2019). Decoding human fetal liver haematopoiesis. Nature *574*, 365–371.

Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr. Protoc. Bioinformatics *47*, 11.12.1-34.

Saha, S.K., Islam, S.M.R., Kwak, K.-S., Rahman, M.S., and Cho, S.-G. (2020). PROM1 and PROM2 expression differentially modulates clinical prognosis of cancer: a multiomics analysis. Cancer Gene Ther. *27*, 147–167.

Schep, A.N., Wu, B., Buenrostro, J.D., and Greenleaf, W.J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods *14*, 975–978.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y.,

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888-1902.e21.

Tangherloni, A., Ricciuti, F., Besozzi, D., Lio, P., and Cvejic, A. (2019). scAEspy: a unifying tool based on autoencoders for the analysis of single-cell RNA sequencing data. BioRxiv.

Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. *9*, 5233.

Unnisa, Z., Clark, J.P., Roychoudhury, J., Thomas, E., Tessarollo, L., Copeland, N.G., Jenkins, N.A., Grimes, H.L., and Kumar, A.R. (2012). Meis1 preserves hematopoietic stem cells in mice by limiting oxidative stress. Blood *120*, 4973–4981.

Velten, L., Haas, S.F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B.P., Hirche, C., Lutz, C., Buss, E.C., Nowak, D., et al. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. Nat. Cell Biol. *19*, 271–281.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.

Wang, H., Liu, C., Liu, X., Wang, M., Wu, D., Gao, J., Su, P., Nakahata, T., Zhou, W., Xu, Y., et al. (2018). MEIS1 regulates hemogenic endothelial generation, megakaryopoiesis, and thrombopoiesis in human pluripotent stem cells by targeting TAL1 and FLI1. Stem Cell Reports *10*, 447–460.

Weissman, I.L. (2000). Stem cells: units of development, units of regeneration, and units in evolution. Cell *100*, 157–168.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.

Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F.J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. *20*, 59.

Wu, W., Morrissey, C.S., Keller, C.A., Mishra, T., Pimkin, M., Blobel, G.A., Weiss, M.J., and Hardison, R.C. (2014). Dynamic shifts in occupancy by TAL1 are guided by GATA factors and drive large-scale reprogramming of gene expression during hematopoiesis. Genome Res. *24*, 1945–1962.

de Wynter, E.A., Buck, D., Hart, C., Heywood, R., Coutinho, L.H., Clayton, A., Rafferty, J.A., Burt, D., Guenechea, G., Bueren, J.A., et al. (1998). CD34+AC133+ cells isolated from cord blood are highly enriched in long-term culture-initiating cells, NOD/SCID-repopulating cells

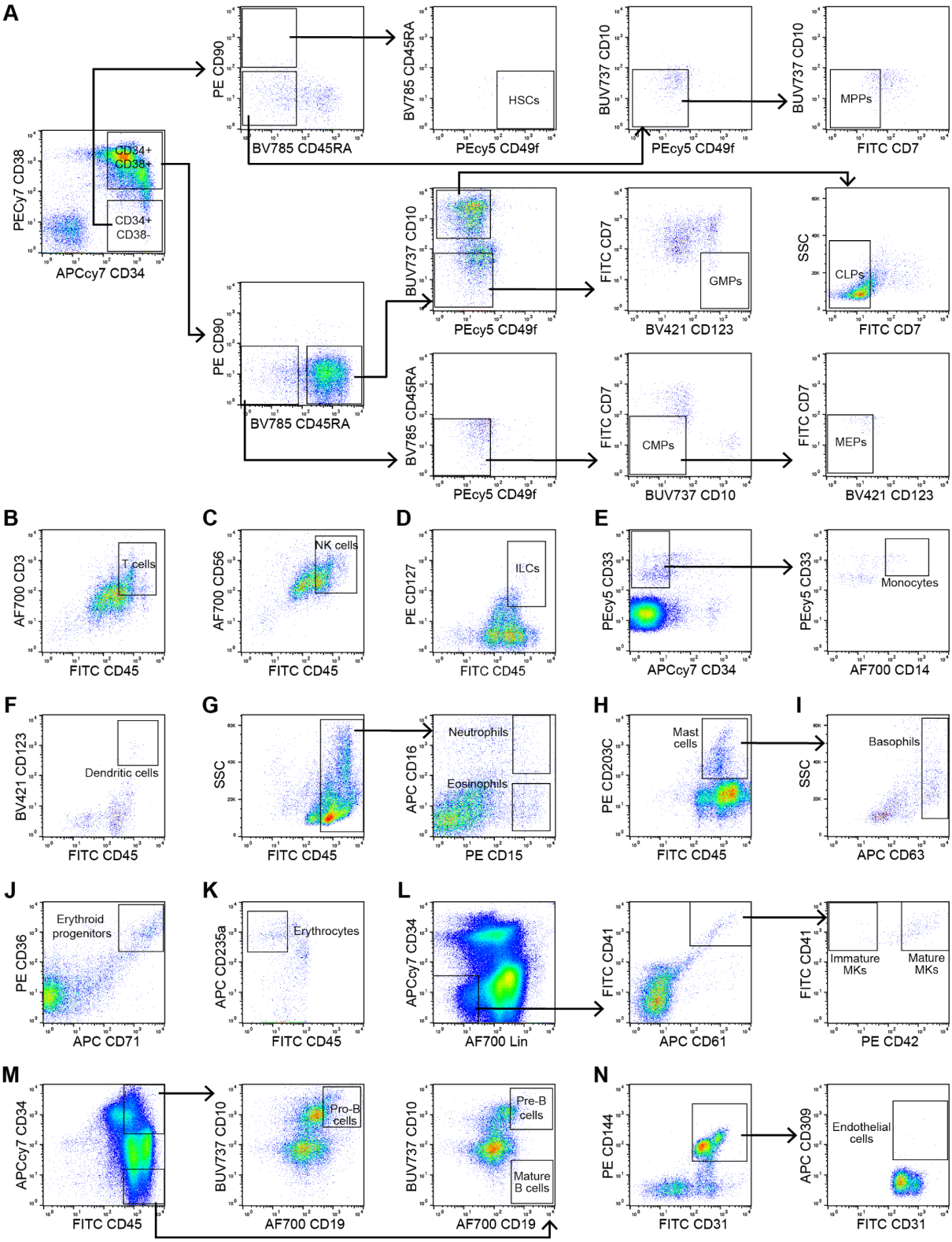
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and dendritic cell progenitors. Stem Cells *16*, 387–396.

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

Zhao, C., Xiu, Y., Ashton, J., Xing, L., Morita, Y., Jordan, C.T., and Boyce, B.F. (2012). Noncanonical NF-κB signaling regulates hematopoietic stem cell self-renewal and microenvironment interactions. Stem Cells *30*, 709–718.

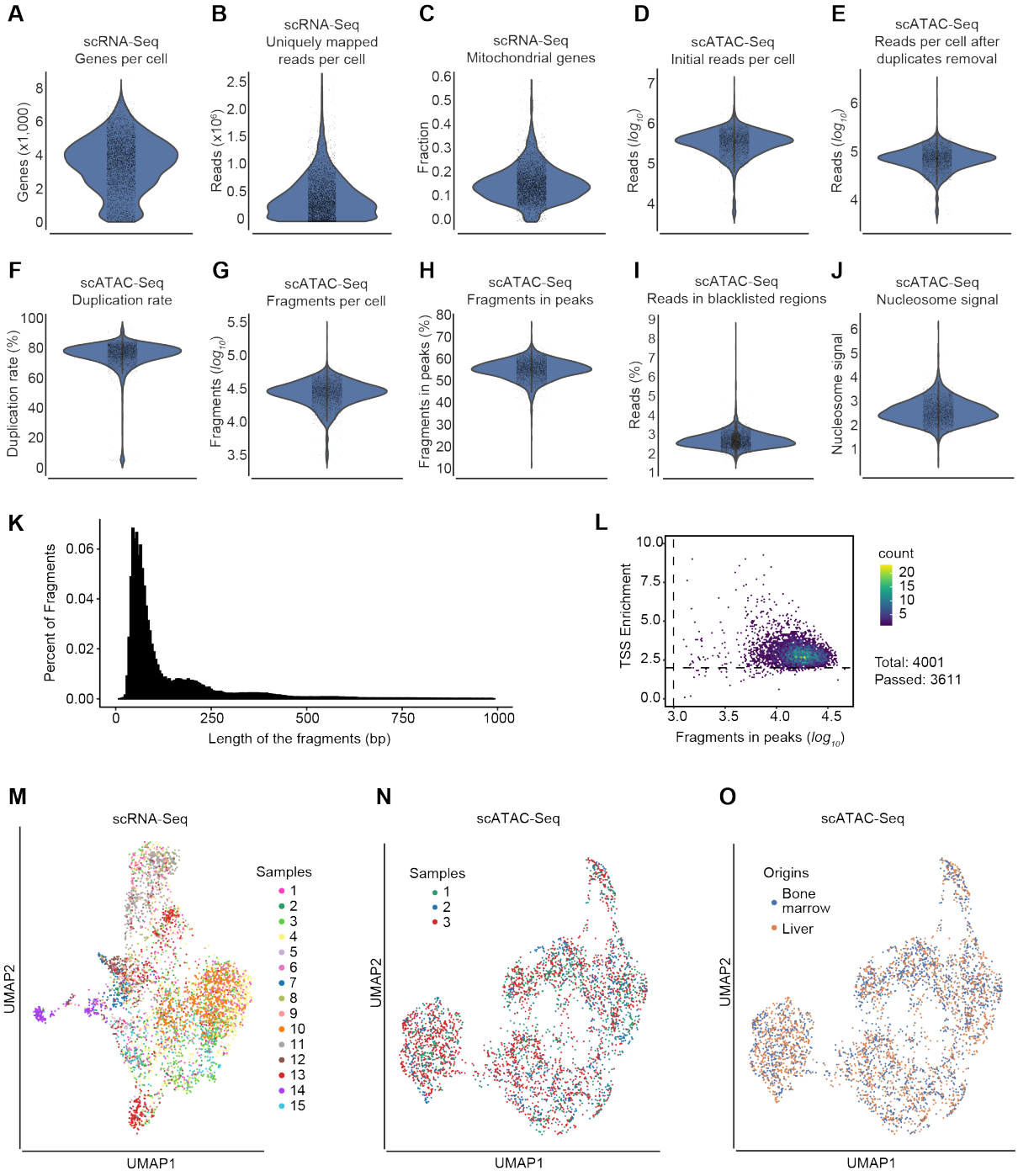
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary figure 1 - Sorting panels. FACS sorting panel and gating strategy for the isolation of phenotypically defined cell types. **A.** Committed and non-committed haematopoietic progenitors, HSCs, MPPs, CMPs, GMPs, MEPs, and CLPs. **B.** T cells. **C.** NK cells. **D.** ILCs. **E.** Monocytes. **F.** Dendritic cells. **G.** Neutrophils and eosinophils. **H.** Mast cells. **I.** Basophils. **J.** Erythroid progenitors. **K.** Erythrocytes. **L.** Immature and mature MKs. **M.** Pro-B cells, pre-B cells, and mature B cells. **N.** Endothelial cells. HSCs - haematopoietic stem cells, MPPs - multipotent progenitors, CMPs - common myeloid progenitors, GMPs - granulocyte-monocyte progenitors, MEPs - megakaryocyte-erythroid progenitors and CLPs - common lymphoid progenitors, NK cells - natural killer cells, ILCs - innate lymphoid cells, MKs - megakaryocytes.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

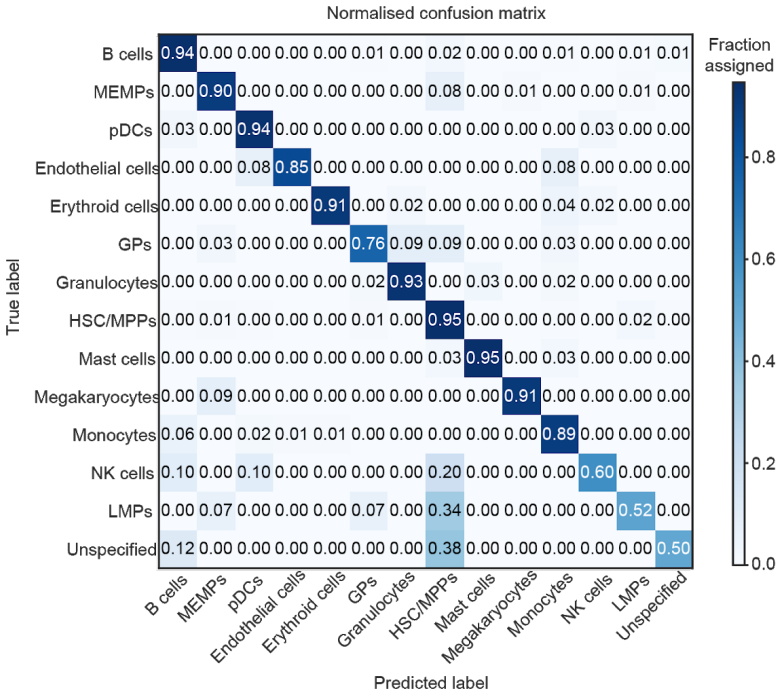


bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

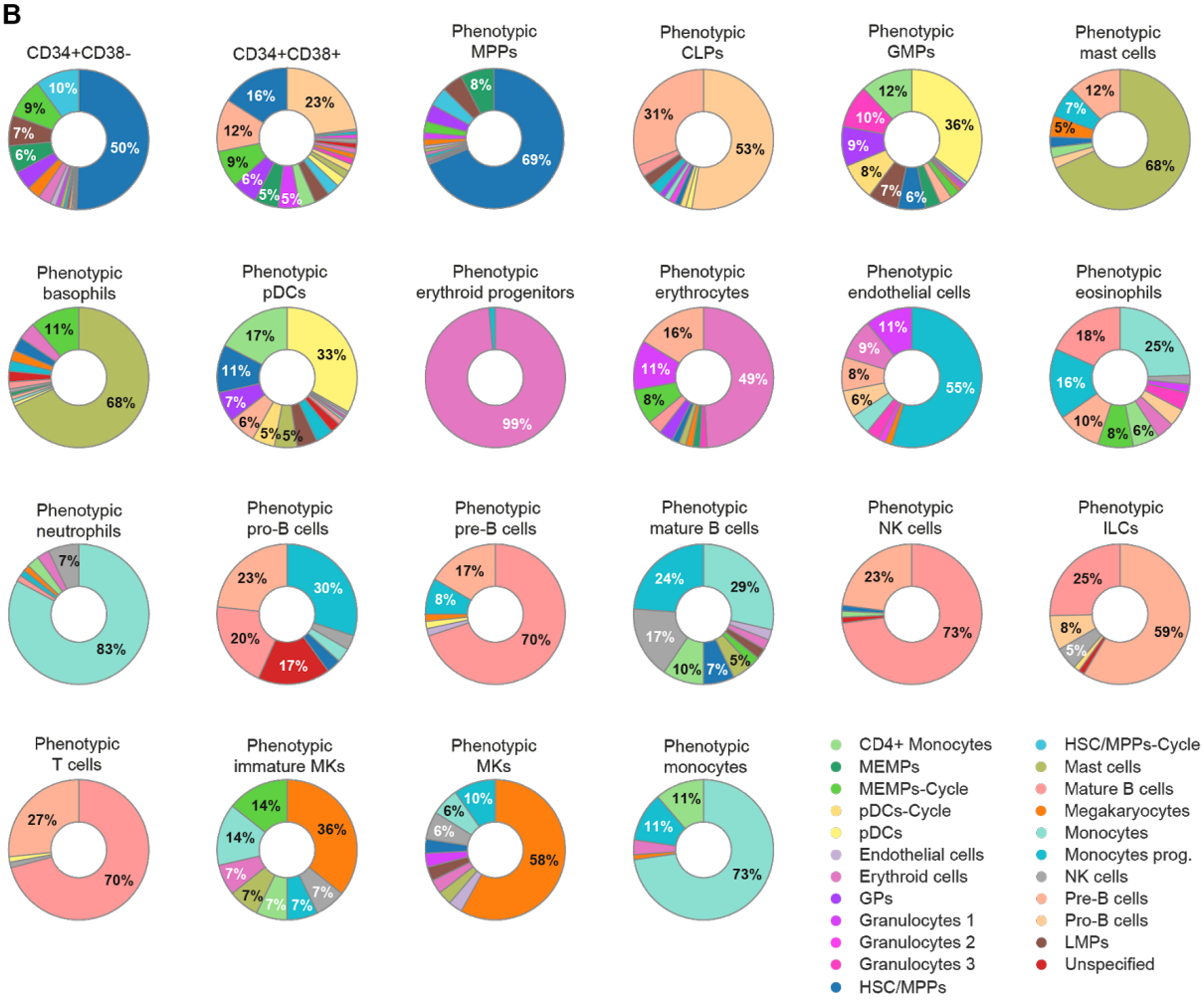
Supplementary figure 2 - Quality control and batch effects correction in scRNA-Seq and scATAC-Seq data. A. Violin plots showing the number of expressed genes per cell in scRNA-Seq data. **B.** Violin plots showing the number of uniquely mapped reads against the reference genome per cell in scRNA-Seq data. **C.** Violin plots showing the fraction of mitochondrial genes compared to all genes per cell in scRNA-Seq data. **D.** Violin plots showing the number of reads per cell, prior to duplicates removal, in scATAC-Seq data. The y-axis is in log_{10} scale. **E.** Violin plots showing the number of reads per cell after duplicates removal in scATAC-Seq data. The y-axis is in log_{10} scale. **F.** Violin plots showing the duplicate rate in scATAC-Seq data. **G.** Violin plots showing the number of fragments per cell in scATAC-Seq data. The y-axis is in log_{10} scale. **H.** Violin plots showing the percentage of fragments per peak in scATAC-Seq data. **I.** Violin plots showing the percentage of reads mapping to the blacklist regions in scATAC-Seq data. **J.** Violin plots showing the nucleosome signal per cell in scATAC-Seq data. **K.** Histogram showing the length of the fragments in terms of base pairs (200 bins). **L.** Scatterplot showing the fragments in peaks with respect to TSS enrichment. The colour intensity represents the number of counts. The x-axis is in log_{10} scale. **M.** UMAP visualization of the scRNA-Seq samples (n=15) after the batch effect correction with BBKNN. Each colour represents a different sample. **N.** UMAP visualization of the scATAC-Seq samples (n=3) after the batch effect correction with Harmony. Each colour represents a different sample. **O.** UMAP visualization of the scATAC-Seq bone marrow (blue) and liver (orange) CD34+ CD38- cells after the batch effect correction with Harmony.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A



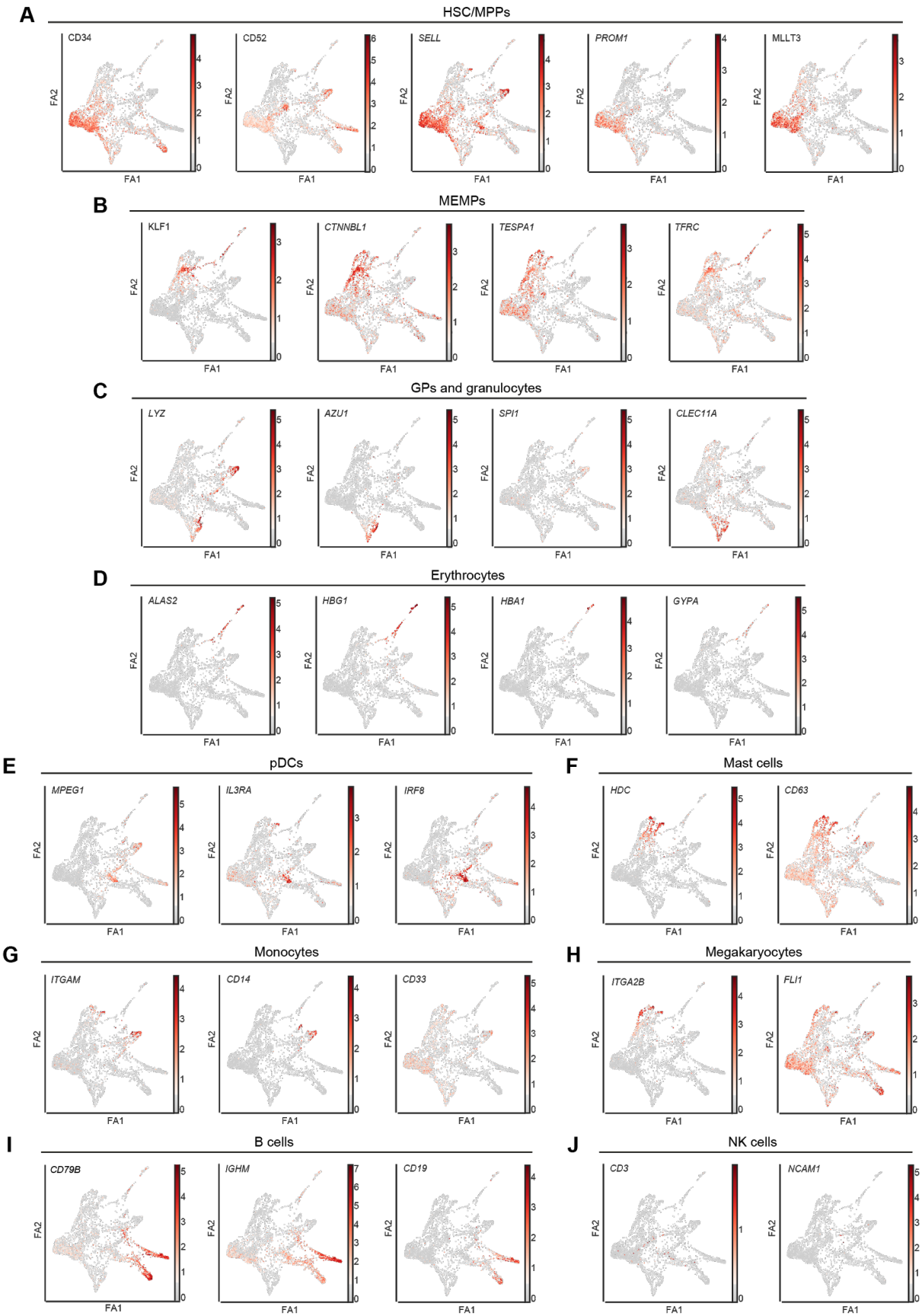
B



bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary figure 3 - Validation of the cell type assignment and transcriptional heterogeneity of phenotypically defined cell populations. A. Confusion matrix showing the cell type assignment achieved by the DNN on the test set (901 cells), considering the top 30 marker genes per cell type (14 distinct cell types). The colour intensity represents the fraction of the assigned cells per cell type. **B.** Donut plots showing the percentage of transcriptionally defined (i.e., manually curated) cell populations in each of the phenotypically defined populations (Expanded from Figure 1C). Each colour represents a different cell type. HSC/MPPs-Cycle - cycling haematopoietic stem cells/multipotent progenitors; HSC/MPPs - haematopoietic stem cells/multipotent progenitors; MEMPs - megakaryocyte-erythroid-mast progenitors; MEMPs-Cycle - cycling megakaryocyte-erythroid-mast progenitors; GPs - granulocytic progenitors; LMPs - lympho-myeloid progenitors; pDCs-Cycle - cycling plasmacytoid dendritic cells; pDCs - plasmacytoid dendritic cells.

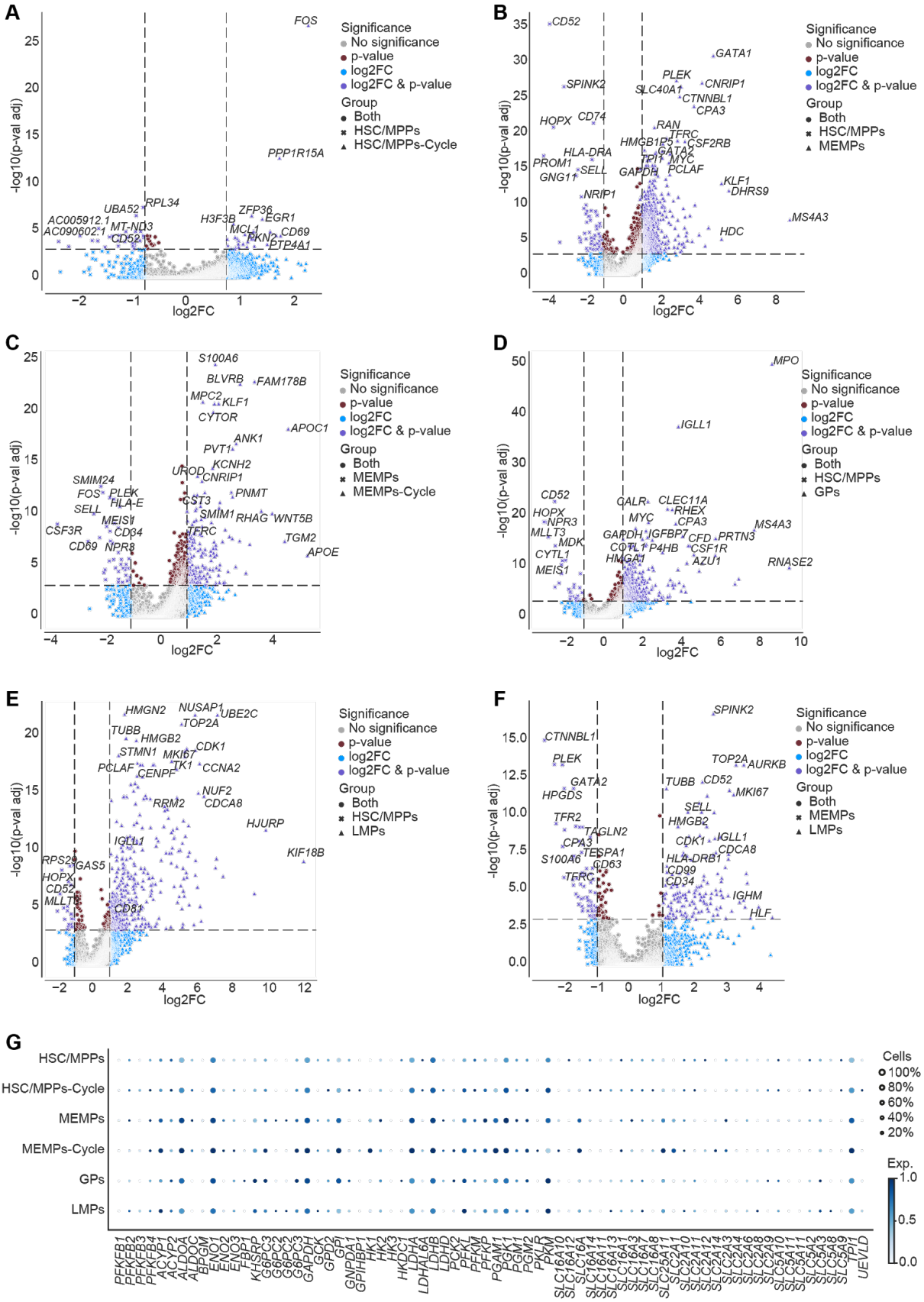
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary figure 4 - Expression of top marker genes along the differentiation trajectory. (A-J) FDG visualisation of the *log*-normalised gene expression of marker genes along the differentiation trajectory. **A**. HSC/MPPs (*CD34*, *CD52*, *SELL*, *PROM1*, and *MLLT3).* **B.** MEMPs (*KLF1*, *CTNNBL1*, *TESPA1*, *and TFRC*). **C.** GPs and granulocytes (*LYZ*, *AZU1*, *SPI1*, and *CLEC11A*). **D.** Erythrocytes (*ALAS2*, *HBG1*, *HBA1*, and *GYPA*). **E.** pDCs (*MPEG1*, *IL3RA*, and *IRF8*). **F.** Mast cells (*HDC* and *CD63*). **G.** Monocytes (*ITGAM*, *CD14*, and *CD33*). **H.** Megakaryocytes (*ITGA2B* and *FLI1*). **I.** B cells (*CD79B*, *IGHM*, and *CD19*). **J.** NK cells (*CD3* and *NCAM1*). Force-Directed Graph - FDG; ForceAtlas2 - FA2; HSC/MPPs - haematopoietic stem cells/multipotent progenitors; MEMPs - megakaryocyte-erythroid-mast progenitors; GPs - granulocytic progenitors; pDCs-Cycle - cycling plasmacytoid dendritic cells.

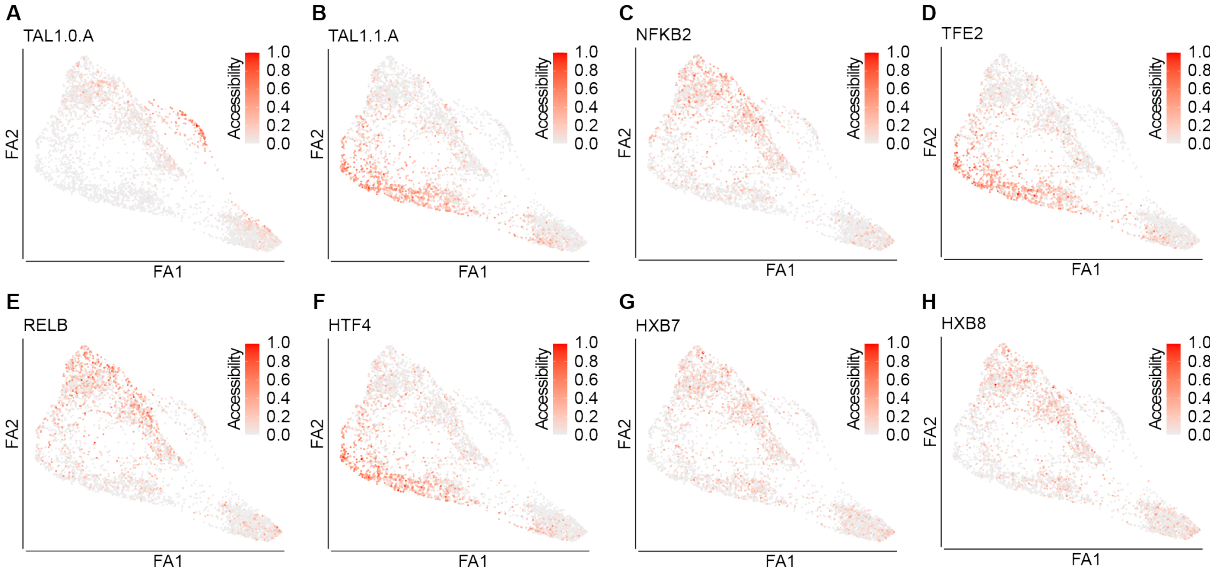
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

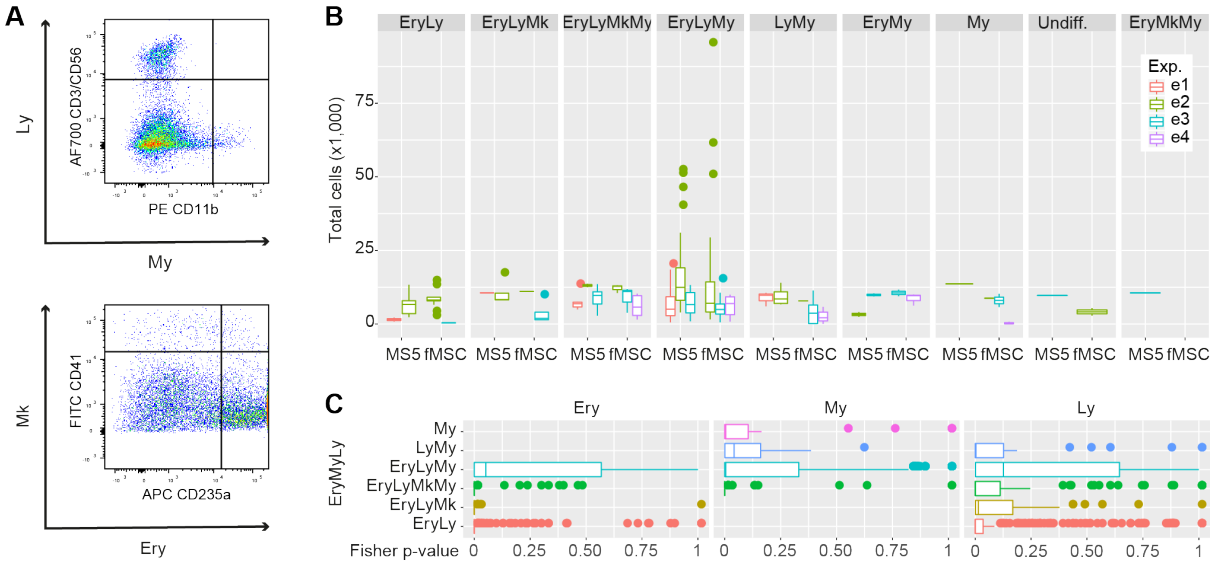
Supplementary figure 5 - Differential expression analysis of the progenitor compartment. (A-F) Volcano plot showing DEGs between two cell types of interest. **A.** HSC/MPPs and HSC/MPPs-Cycle. **B.** HSC/MPPs and MEMPs. **C.** MEMPs and MEMPs-Cycle. **D.** HSC/MPPs and GPs. **E.** HSC/MPPs and LMPs. **F.** MEMPs and LMPs. The x-axes show the log_2 fold-change (magnitude of change), while the y-axes show the $-log_{10}$ adjusted p-value (statistical significance). We used the Wilcoxon rank-sum with the Benjamini-Hochberg correction. Colours represent the significance of the genes, both in terms of p-value and log_2 fold-change. **G.** Dot plot of the expression of metabolic genes involved in glycolysis in the identified progenitor compartment. The expression of the genes is standardised between 0 and 1. For each gene, the minimum value is subtracted and the result is divided by the maximum. The spot size indicates the percentage of cells that express the gene of interest within each cell type. The colour intensity represents the standardised expression level. HSC/MPPs-Cycle - cycling haematopoietic stem cells/multipotent progenitors; HSC/MPPs - haematopoietic stem cells/multipotent progenitors; MEMPs - megakaryocyte-erythroid-mast progenitors; MEMPs-Cycle - cycling megakaryocyte-erythroid-mast progenitors; GPs - granulocytic progenitors; LMPs - lympho-myeloid progenitors.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Supplementary figure 6 - Motif accessibility of selected transcription factors along the differentiation trajectory. (A-H) FDG visualisation of the min-max normalised TF motif accessibility along the differentiation trajectory**. A.** TAL1.O.A. **B.** TAL1.O.A. **C.** NFKB2. **D.** TFE2. **E.** RELB. **F.** HTF4. **G.** HXB7. **H.** HXB8. Force-Directed Graph - FDG; ForceAtlas2 - FA2.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Supplementary figure 7 - Statistical analysis of lineage output. A. Flow cytometry gating strategy used to assess the lineage output of single-cell derived colonies. **B.** Box plot showing differences in the number of cells per cell combination between the two feeder layers. **C.** Representative box plot showing Fisher's exact tests to assess the contribution of particular cell types (Ery, Ly, My) to the variability of EryLyMy trilineage colonies.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Table 1 - Cell-surface markers used to isolate cell types

Phenotypic cell type	Cell-surface markers
HSCs	Lin- CD34+ CD38- CD45RA- CD90+ CD49f+/-
MPPs	Lin- CD34+ CD38- CD90- CD45RA- CD49f- CD10- CD7-
CMPs	Lin- CD34+ CD38+ CD90- CD45RA- CD49f- CD10- CD7-
MEPs	Lin- CD34+ CD38+ CD90- CD45RA- CD49f- CD10- CD7- CD123-
GMPs	Lin- CD34+ CD38+ CD90- CD45RA+ CD49f- CD10- CD7- CD123+/-
CLPs	Lin-, CD34+ CD38+ CD90- CD45RA+ CD49f - CD10+ CD7-
T cells	CD45+ CD3+
NK cells	CD45+ CD56+
ILCs	Lin+/- CD45+ CD127+
Monocytes	CD34- CD33+ CD14+
Dendritic cells	Lin- CD45+ CD123+
Mast cells	CD45+ CD203c+
Basophils	CD45+ CD203c+ CD63+
Neutrophils	CD45+ CD15+ CD16+
Eosinophils	CD45+ CD15+ CD16-
Erythroid progenitors	CD36+ CD71+
Erythrocytes	CD235a+
Immature MKs	Lin- CD34- CD41+ CD61+ CD42-
Mature MKs	Lin- CD34- CD41+ CD61+ CD42+
Pro-B cells	CD45+ CD34+ CD19+ CD10+
Pre-B cells	CD45+ CD34- CD19+ CD10+
Mature B cells	CD45+ CD34- CD19+ CD10-
Endothelial cells	CD31+ CD144+ CD309+

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary table 2 - scRNA-Seq samples

No.	Gates	Min #counts	Max #counts	Min #genes	Max #genes	Max %mito
1	Non-committed progenitors Committed progenitors	10,000	1,750,000	200	7,500	40
2	Non-committed progenitors Committed progenitors HSCs	10,000	2,000,000	200	7,500	40
3	Non-committed progenitors Committed progenitors HSCs MPPs CMPs GMPs MEPs	10,000	1,500,000	200	7,500	40
4	Non-committed progenitors Committed progenitors HSCs MPPs CMPs GMPs MEPs	10,000	2,750,000	200	9,000	40
5	Immature MKs Mature MKs	10,000	1,750,000	200	9,000	40
6	Non-committed progenitors Committed progenitors HSCs MPPs CMPs GMPs MEPs	10,000	1,500,000	350	7,500	40
7	Bone marrow cells	400	1,750,000	200	7,500	60
8	Endothelial cells Pro-B cells Pre-B cells Mature B cells	400	1,500,000	200	7,500	60
9	CLPs GMPs Monocytes	1,000	2,000,000	200	7,500	40
10	CD-Ref cells	1,000	550,000	200	6,000	40
11	ILCs NK cells T cells	2,000	1,500,000	200	10,000	40

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

12	Neutrophils Eosinophils Erythrocytes	2,000	1,500,000	200	6,000	40
13	Dendritic cells Mast cells Basophils	2,000	1,500,000	200	7,500	40
14	Erythroid progenitors	500	3,000,000	100	4,500	40
15	CD62L- CD52- CD114- CD125- CD117+	2,000	1,000,000	1,200	7,000	40

Supplementary Table 3 - Antibodies

Antibody	**Clone**	**Manufacturer**
CD3 AF700	OKT3	Biolegend
CD8 AF700	SK1	Biolegend
CD11b AF700	ICRF44	Biolegend
CD14 AF700	61D3	Thermo Fisher
CD19 AF700	HIB19	Biolegend
CD56 AF700	B159	BD Biosciences
CD34 APC-Cy7	581	Biolegend
CD38 PE-Cy7	HB7	BD Biosciences
CD45RA BV785	HI100	Biolegend
CD90 PE	5E10	BD Biosciences
CD49f PE-Cy5	GoH3	BD Biosciences
CD10 BUV737	HI10a	BD Biosciences
CD7 FITC	MT701	BD Biosciences
CD123 BV421	9F5	BD Biosciences
CD45 FITC	HI30	Thermo Fisher
CD127 PE	A019D5	Biolegend
CD33 PE-Cy5	WM53	Biolegend
CD203c PE	NP4D6	Biolegend
CD63 APC	H5C6	Biolegend
CD15 PE	HI98	Thermo Fisher

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.080259; this version posted May 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CD16 APC	3G8	Biolegend
CD36 PE	5271	Biolegend
CD71APC	CY1G4	Biolegend
CD235a APC	GAR2	BD Biosciences
CD41 FITC	HIP8	BD Biosciences
CD42 PE	HIP1	BD Biosciences
CD61 APC	Y251	Dako
CD31 FITC	WM59	Biolegend
CD144 PE	BV9	Biolegend
CD309 APC	7D46	Biolegend
CD11b PE	ICRF44	Biolegend