Single cell analyses and machine learning define hematopoietic progenitor and HSC-like cells derived from human PSCs

American Society of Hematology
2021 L Street NW, Suite 900,
Washington, DC 20036
Phone: 202-776-0544 | Fax 202-776-0545
editorial@hematology.org

Single cell analyses and machine learning define hematopoietic progenitor and HSC-like cells derived from human PSCs.

Tracking no: BLD-2020-006229R1

Antonella Fidanza (University of Edinburgh, United Kingdom) Patrick Stumpf (University of Southampton, United Kingdom) Prakash Ramachandran (University of Edinburgh, United Kingdom) Sara Tamagno (University of Edinburgh, United Kingdom) Ann Babtie (Centre for Integrative Systems Biology and Bioinformatics, Department of Life Sciences, Imperial College London, United Kingdom) Martha Lopez-Yrigoyen (University of Edinburgh, United Kingdom) Alice Taylor (University of Edinburgh,) Jennifer Easterbrook (University of Edinburgh, United Kingdom) Beth Henderson (University of Edinburgh, United Kingdom) Richard Axton (University of Edinburgh, United Kingdom) Neil Henderson (University of Edinburgh, United Kingdom) Alexander Medvinsky (The University of Edinburgh, United Kingdom) Katrin Ottersbach (University of Edinburgh, United Kingdom) Nicola Romano (University of Edinburgh, United Kingdom) Lesley Forrester (University of Edinburgh, United Kingdom)

Abstract:
Haematopoietic stem and progenitor cells (HSPCs) develop through distinct waves at various anatomical sites during embryonic development. The in vitro differentiation of human pluripotent stem cells (hPSCs) is able to recapitulate some of these processes but it has proven difficult to generate functional haematopoietic stem cells (HSCs). To define the dynamics and heterogeneity of HSPCs that can be generated in vitro from hPSCs, we exploited single cell RNA sequencing (scRNAseq) in combination with single cell protein expression analysis. Bioinformatics analyses and functional validation defined the transcriptomes of naïve progenitors as well as erythroid, megakaryocyte and leukocyte-committed progenitors and we identified CD44, CD326, ICAM2/CD9 and CD18 as novel markers of these progenitors, respectively. Using an artificial neural network (ANN), that we trained on a scRNAseq derived from human fetal liver, we were able to identify a wide range of hPSCs-derived HPSC phenotypes, including a small group classified as HSCs. This transient HSC-like population reduced as differentiation proceeded and was completely missing in the dataset that had been generated using cells selected on the basis of CD43 expression. By comparing the single cell transcriptome of in vitro-generated HSC-like cells with those generated within the fetal liver we identified transcription factors and molecular pathways that can be targeted with the aim of improving HSC differentiation in vitro.

Conflict of interest: No COI declared

COI notes:

Preprint server: Yes; BioRxiv https://doi.org/10.1101/602565

Author contributions and disclosures: AF, designed and performed research, analyzed the data and wrote the manuscript. AF, PS, AB and NR performed bioinformatics analysis. PR, ST, MLY, AHT, JE, BH, RA performed research. LMF designed the experiment, analyzed data and wrote the manuscript. NH, AM, KO, and NR provided intellectual input and final approval of the manuscript.

Non-author contributions and disclosures: No;

Agreement to Share Publication-Related Data and Data Sharing Statement: We have created a webpage where the data can be freely browsed, plots can be generated and exported, and full datasets can be downloaded. The link is provided in the manuscript.

Clinical trial registration information (if any):

Single cell multimodal analyses and machine learning define haematopoietic progenitor and HSC-like cells derived in vitro from human pluripotent stem cells.

Antonella Fidanza[1]*, Patrick S Stumpf[2], Prakash Ramachandran[3], Sara Tamagno[1], Ann Babtie[4], Martha Lopez-Yrigoyen[1], A. Helen Taylor[1], Jennifer Easterbrook[1], Beth Henderson[3], Richard Axton[1], Neil C. Henderson[3], Alexander Medvinsky[1], Katrin Ottersbach[1], Nicola Romanò[5], Lesley M. Forrester[1]*.

1 - Centre for Regenerative Medicine, University of Edinburgh, Edinburgh, UK
2 - Joint Research Center for Computational Biomedicine, Uniklinik RWTH Aachen, Aachen, Germany
3 - Centre for Inflammation Research, University of Edinburgh, Edinburgh, UK
4 - Centre for Integrative Systems Biology and Bioinformatics, Department of Life Sciences, Imperial College London, London, UK
5 - Centre for Discovery Brain Sciences, University of Edinburgh, Edinburgh, UK

*correspondence to afidanza@ed.ac.uk and L.Forrester@ed.ac.uk

Key point 1 - Single-cell and CITE-seq profiling of human HSPCs derived in vitro from pluripotent stem cells browsable at http://188.166.158.65/scRNAseqHPC/

Key point 2 – Artificial Neural Network identifies HSC-like cells derived in vitro from hPSCs.

Abstract

Haematopoietic stem and progenitor cells (HSPCs) develop through distinct waves at various anatomical sites during embryonic development. The in vitro differentiation of human pluripotent stem cells (hPSCs) is able to recapitulate some of these processes, however, it has proven difficult to generate functional haematopoietic stem cells (HSCs). To define the dynamics and heterogeneity of HSPCs that can be generated in vitro from hPSCs, we exploited single cell RNA sequencing (scRNAseq) in combination with single cell protein expression analysis. Bioinformatics analyses and functional validation defined the transcriptomes of naïve progenitors as well as erythroid, megakaryocyte and leukocyte-committed progenitors and we identified CD44, CD326, ICAM2/CD9 and CD18

as markers of these progenitors, respectively. Using an artificial neural network (ANN), that we trained on a scRNAseq derived from human fetal liver, we were able to identify a wide range of hPSCs-derived HPSC phenotypes, including a small group classified as HSCs. This transient HSC-like population decreased as differentiation proceeded and was completely missing in the dataset that had been generated using cells selected on the basis of CD43 expression. By comparing the single cell transcriptome of in vitro-generated HSC-like cells with those generated within the fetal liver we identified transcription factors and molecular pathways that can be exploited in the future to improve the in vitro production of HSCs.

Introduction

Human pluripotent stem cells (hPSCs) can be differentiated in vitro into various haematopoietic cell types, providing both a model for basic research studies and a source of clinically relevant cells[1]. During embryonic development, two waves of restricted haematopoietic progenitors arise in the extraembryonic tissues of the yolk sac, before emergence of haematopoietic stem cells (HSCs) in the embryo proper[2]. In the mouse embryo, at E7.25 the first “primitive” wave gives rise to erythrocytes, megakaryocytes and macrophages[3,4], after at E8.25, the second wave, also known as the first “definitive” progenitors, consists of erythro-myeloid progenitors (EMPs) that can be distinguished from the primitive progenitors by their potential to generate granulocytes[5]. Furthermore, the monocytes that emerge from EMPs provide the embryo with tissue resident macrophage; the first life-long lasting population of immune cells[6–8]. Intraembryonic hematopoiesis is established during E10.5-E11.5 in the aorta-gonad-mesonephros (AGM) region with the emergence of HSCs, that will sustain the lifespan production of all blood lineages, also upon transplantation[9]. A number of studies have indicated that human haematopoietic development follows a comparable process[10–13] but for ethical reasons it has proven difficult to gain a clear insight into the lineage potential and hierarchical relationships between early human haematopoietic progenitors. The dynamic nature and the heterogeneity of haematopoietic progenitor populations that arise during development poses additional confounders to the identification of molecular mechanisms associated with their formation and function.

To gain insight into the transcriptome of developing human haematopoietic progenitors, we performed in-depth characterization of haematopoietic progenitors derived from

hPSCs. Single-cell expression profiles of hPSCs-derived haematopoietic cells have been reported but these previous studies either used a limited number of cells or used biased approaches in their isolation and so failed to depict their trajectory of differentiation [14–16]. This significantly impacted on the ability to resolve the complex heterogeneity of the progenitor pool, to identify the hierarchal relationship between subpopulations and to compare the transcriptome of hPSCs-derived progenitors to their in vivo counterparts. To address these issues, we generated scRNAseq data sets of human hPSCs-derived haematopoietic progenitors. Lineage trajectories predicted in silico were validated by functional assays of sorted cells and the specificity of our marker repertoire was confirmed using a CITE-seq[17] strategy. Furthermore, to annotate the hPSCs-derived progeny in an unbiased manner, we employed machine learning and trained an artificial neural network (ANN) to recognize the single-cell gene expression profiles of human fetal liver cells. This trained ANN was subsequently used to predict the identities of hPSCs-derived cells. The ANN thereby provides a mapping between in vivo and in vitro hematopoiesis and revealed a subset of hPSCs-derived cells that closely resembles HSCs in the foetal liver. Finally, by comparing that transcriptome of in vitro and in vivo-generated HSCs we identified genetic pathways that can be exploited to improve HSCs production in vitro from hPSCs.

Methods

Methods are available as supplementary methods.

Results

Single cell RNA sequencing of iPSCs-derived haematopoietic progenitor cells reveals the transcriptome of naïve and lineage committed progenitors.

To resolve the heterogeneity of in vitro generated hPSCs-derived haematopoietic progenitors we designed a minimal membrane marker strategy that allows to broadly isolate hPSCs-derived haematopoietic progenitors. This marker strategy was validated using two hPSC reporter lines, RUNX1C-GFP and KLF1-mCherry, CFU-C assays of sorted cell populations and gene expression profiling (Supplementary Figure S1). These data, together with previous reports[18–22], supported our rationale that the isolation of CD235a^{-}CD43^{+} cells would enrich for HSPCs and exclude cells derived from the primitive wave (Supplementary Figure S1). We anticipated that the CD235a^{-}CD43^{+} compartment

would also comprise the early stages of lineage commitment, capturing the downstream hierarchy of early human progenitors.

CD235a^{-}CD43^{+} suspension cells from two independent replicate cultures at day 13 of differentiation were sorted by FACS and subjected to scRNAseq and data analyses (Figure 1A). After quality control and clustering[23] we obtained the transcriptome of 11420 cells (Supplementary Figure 2A-C) belonging to eight clusters of cells (Figure 1B). Although the two replicates did not show obvious differences (Supplementary Figure 2C), any potential batch effect was regressed out prior to integration. We assigned cell identities based on the expression of known markers and identified additional markers from the dataset that were cluster specific (Figure 1C-D). Clusters containing more immature, uncommitted progenitors were identified by their expression of progenitor-associated genes such as *KIT* and *GATA2* and their lack of expression of genes associated with specific cell lineages and were thus annotated as naïve populations (Figure 1D, Supplementary Figure 2D). Clusters that displayed expression of lineage markers were annotated as lineage committed progenitors (Figure 1B-D, Supplementary Figure 2D), including clusters of cells committed towards the megakaryocyte (*GP9* and *PF4*), erythroid (*GYPA* and *KLF1*) and granulocyte (*AZU1* and *PRNT3*) lineages (Figure 1D). Markers for each of the cell clusters were identified by differential gene expression analysis, further supporting the identities assigned to each cluster (Figure 1C, Supplementary Table 1).

Trajectory analyses reveal the hierarchy of in vitro derived haematopoietic progenitors.

To study the hierarchical relationship between cell populations, we performed trajectory analysis using different methods including diffusion analysis[24] using Seurat R package[23] and pseudotemporal ordering, using Monocle R package[25] and Partition-based graph abstraction (PAGA)[26]. Diffusion analysis identified a central core from which three distinct trajectories emerged (Figure 2A). The central core corresponded to cells that we had annotated as naïve progenitors (Figure 2A-B). Branches comprised cells expressing genes associated with specific lineages, annotated as Erythroid (Ery)-, Megakaryocyte (Mega)- and Granulocyte (Granulo)-committed lineages. Comparable trajectories were observed using pseudotemporal ordering with PAGA and Monocle (Figure 2C-D). The PAGA analysis showed that the naïve cells were highly connected to erythroid, megakaryocyte

and granulocyte committed cells (Figure 2C). Pseudotime reconstruction of the hierarchy showed that cells annotated as naïve 1 were located at the top of the hierarchy and appeared to progress to naïve 2 cells before entering branches containing lineage committed cells (Figure 2D-E). Lineage commitment was also inferred from the expression of lineage-associated transcription factors that were filtered from the cluster specific marker genes according to their GO annotation (Figure 2F, Supplementary Figure 2E). For example, erythroid committed clusters demonstrated expression of both *KLF1* and *MYC*, with the latter decreasing in Ery 2 compared to Ery1, in keeping with their position within the hierarchy (Figure 2E, F). Within the megakaryocyte-committed clusters 1 and 2 we observed the expression of *GATA1*, *TAL1* and *FLI1* a cocktail of genes recently used for hPSCs forward programming to megakaryocytes (Figure 2F)[27]. Granulocyte-committed cells were represented by a separate branch and demonstrated the expression of *CEBP-D*, *CEBP-B*, *CEBP-A* and *CEBP-E* (Figure 2F). We then focused our attention on the transcription factors expressed by the naïve progenitor clusters and noted a high level of expression of LMO4 and ID2, as well as GATA2 which is known to be expressed in HSPCs (Figure 2F). We then inferred their role in the gene network using a network inference algorithm (Partial Information Decomposition and Context, PIDC)[28,29]. Single cell transcriptomic data offers the potential to observe dependencies between the expression profiles of pairs of genes, that if co-regulated, are expected to change in a coordinated fashion. Genes with highest statistical dependencies are connected by edges that altogether define the network[28,29]. Many of the transcription factors previously detected to be highly expressed within the naïve cell populations such as *ID2, ID4 and LMO4*, occupy nodes within this large network (Supplementary Figure 3A-B). This strategy corroborates the importance of the identified transcription factors as functional elements within the single cell gene network.

CD44 membrane expression marks human clonogenic haematopoietic progenitors.

To experimentally validate the results of our trajectory analyses experimentally, we set out to assess the haematopoietic potential of the naïve progenitor populations. We defined a prospective sorting strategy using genes encoding the cell surface markers *CD33*, *CD44*, and *ITGB2* (also known as CD18) that were enriched within the naïve progenitors' clusters (Figure 3A). CD33 was expressed by both naïve 1 and naïve 2 progenitors whereas CD44 and CD18 expression appeared higher in the naïve 1 population (Figure 3A). We

fractionated CD235a$^-$CD43$^+$CD33$^+$ cells and identified subpopulations as naïve 1A (CD44$^+$CD18$^-$), naïve 1B (CD44$^+$CD18$^+$) and naïve 2 (CD44$^-$CD18$^-$) (Figure 3B). Trajectory analysis predicted that naïve 1 cells were at the top of the hierarchy and gave rise to the naïve 2 cells prior to lineage commitment (Figure 2D-E). To test this in silico prediction, we used a chimeric co-culture system using the Zeiss Green (ZsG) reporter (Figure 3C). This approach allowed us to sort, for example, ZsG-labelled naïve 1 cells, then track their ZsG progeny after being placed back in the complex differentiation environment. We synchronously differentiated the ZsG-iPSC line, constitutively expressing the fluorescent reporter[30], and the parental iPSC line. To verify the progressions of naïve 1 to naïve 2 and, naïve 2 to lineage committed cells, we sorted naïve 1 (CD33$^+$CD44$^-$CD18$^-$) or naïve 2 (CD33$^+$CD44$^+$CD18$^{-/+}$) cells from ZsG-iPSCs at day 10 and co-cultured these with the synchronized differentiating parental cells for a further 3 days. As predicted from the trajectory analysis, the naïve 1 cell population was able to generate ZsG-expressing naïve 2 cells. We also noted that the naïve 1 cells retained their immunophenotype, indicating some self-renewal capacity (Figure 3D). Interestingly, naïve 2 cells demonstrated some potential to acquire CD44 and CD18, markers of naïve 1 cells (Figure 3D), suggesting fluidity between these states. As predicted by our trajectory analyses (Figure 2D-E), naïve 2 cells acquired the ability to generate more mature cells including erythroid cells (CD235a$^+$), megakaryocytes (CD41$^+$) and macrophages (25F9$^+$) (Supplementary Figure 3C). We compared the colony forming capacity of naïve 1 and 2 progenitors present at day 10 and day 13. When plated in clonogenic CFU-C assays, CD44$^+$ naïve 1 cells formed CFU-C colonies but virtually no colonies were generated by naïve 2 cells at either time point (Figure 3E-F). These data support the proposed hierarchy and indicate that CD44 expression alone resolves colony forming cells. Our chimeric co-culture system was able therefore to assess the lineage output that could not be assessed solely by CFU-C assays. We observed that naïve progenitors expressed high levels of ID genes (Figure 2F), and that they were identified as nodes within the gene network (Supplementary Figure 3A). As ID genes are targets of BMP signaling, we predicted that these naïve cells would be responsive to BMP stimulation. We added BMP4 to differentiation culture from day 10, when both naïve 1 and 2 were present and then assessed the proportion of these cells 3 days later, In presence of BMP4, we observed a 25% and 59% expansion of naïve 1 and 2 cells respectively (Supplementary Figure 3E). In this experiment we used both hESCs and hiPSCs and showed that naïve progenitors are present at a comparable frequency in both

hESCs and hiPSCs (Supplementary Figure 3D), and that naive progenitors derived from both lines responded to BMP stimulation in a comparable manner (Supplementary Figure 3D-E). Thus this experiment not only identified an important functional signaling pathway but also confirmed that the markers we used to define naïve progenitors, and their response to BMP signaling, are not PSC line specific.

To assess whether the naïve cell populations identified using our unique sorting strategy showed features of definitive haematopoietic progenitors, we assessed the expression of the RUNX1C-GFP reporter. We observed RUNX1C-GFP expression in both cell types, with a higher proportion of RUNX1C^{+} cells in the naïve 1 compared to naïve 2 population (Figure 3G). Definitive HSPCs are generated via endothelial to hematopoietic transition (EHT) during embryonic development[31,32] so, we would expect comparable hPSCs-derived progenitors to have hallmarks of their endothelial origin. Here we demonstrate that naïve CD44^{+} cells generated in vitro from hPSCs co-expressed CD34 and the endothelial marker CD144 (also known as VeCad) with expression being higher at day 10, when the majority of naïve progenitors were present (Supplementary Figure 3L). This endothelial signature of naïve progenitors, together with their lineage potential reflects their definitive features. To confirm that CD44 expression was associated with HSPCs in vivo we demonstrated its co-localization with CD45 and CD144 in the mouse yolk sac and AGM region (Supplementary figure 3F-J). At E10.5 in the yolk sac, CD44 was expressed on endothelial cells in a bimodal fashion, with vessels expressing low and high levels, the latter being associated with bright clusters of haematopoietic cells (Supplementary figure 3G). By flow cytometry, we observed that by E11, all CD45^{+} cells and a proportion of CD144^{+} cells were within the CD44^{high} population (Supplementary figure 3F). Within the embryo proper, CD44 was expressed on the membrane of endothelial cells within the dorsal aorta, whereas venous endothelial layers were CD44 negative (Supplementary figure 3H-I). CD44 was also co-expressed with CD45^{+} in the AGM region (Supplementary figure 3H-J). Furthermore, expression of LMO4 in CD44^{+} cells within the AGM region is in keeping with its high level of expression in naïve hPSCs-derived HSPCs (Figure 2F) and supports its identification as a novel haematopoietic transcription factor. These data suggest that CD44 is expressed on haemogenic endothelial cells and it is retained on emerging haematopoietic cells in vivo, similar to what we have observed during the in vitro differentiation of human progenitors (Supplementary Figure 3I).

Identification of membrane markers of lineage committed progenitors

We next selected membrane markers that we predicted could be used for the isolation of lineage committed progenitors. Erythroid-primed clusters 1 and 2 both showed expression of *MYC* (Figure 2F) and *EPCAM* (also known as CD326) (Supplementary Figure 4A), indicative of early committed erythroid cells[33,34]. We confirmed that CD326 was expressed in the majority of CD235a^{+} cells at day 13 of iPSC differentiation but interestingly, we noted a small number of CD326^{+}CD235a^{-} (Supplementary Figure 4B), suggesting that CD326 might be marking commitment to the erythroid lineage prior to CD235a acquisition. To test this, we assessed the expression dynamics of these markers during the erythroid differentiation of umbilical cord blood CD34^{+} (UCB34^{+}) cells. At day 10 of differentiation, CD326 was expressed in CD235a$^{-/low}$ cells but not in CD235a^{high} cells, the latter corresponding to more mature erythroid cells (Supplementary Figure 4B). CD326 was not expressed in cells at day 18 of the differentiation protocol (when the majority of cells are mature CD235a^{+} cells) nor in the mature erythrocytes found in adult peripheral blood (Supplementary Figure 4B). Taken together these data suggest that CD326 marks early erythroid progenitors in both hiPSC, fetal and adult derived cells. We also noted the expression of *HBG1*, *HBG2*, *HBA1*, and *HBA2*, subunits of fetal hemoglobin, indicative of erythroid cells derived from definitive hematopoiesis (Supplementary Figure 4C).

Three clusters with megakaryocyte and platelet signatures (Mega-primed 1, 2 and 3) were predicted by expression of *ITGA2B* (CD41), *GP9*, *PF4* (Figure 1C-D and Supplementary Table 1). ICAM2 was expressed at higher level in cluster Mega-primed 3 (Supplementary Figure 4D), as for CD9, known to increase along megakaryocytes differentiation[35]. ICAM2 and CD9 co-expression was confirmed by flow cytometry (Supplementary Figure 4D). We observed a population of CD41^{+}CD9^{+}ICAM2^{+} cells, with around 85% of the CD41^{+}CD42a^{+} (Supplementary Figure 4E), that did not detect polyploidy, supporting their immature status (Supplementary Figure 4F-G).

Granulocyte-committed clusters were predicted by expression of markers such as *MPO*, *AZU1*, *RNASE2* and *ITGB2* which encodes the membrane marker CD18, subunit of different leukocytes marker such as CD11a-d, Mac-1and LFA-1 (Figure 1C, Supplementary Table 1). We sorted CD235a^{-}CD43^{+}CD33^{+}CD44^{-}CD18^{+} cells and confirmed the phenotype of granulocytes and monocytes based on their nuclear morphology (Supplementary Figure 4H). Further clustering revealed three sub-clusters corresponding to eosinophil, neutrophils and monocytes lineages (Supplementary Figure

4I-L). Noteworthy, *RUNX3* expression was specifically associated with the monocyte subcluster (Supplementary Figure 4J) previously reported to be important for zebrafish stem cells and macrophages[36], and to be expressed by mouse tissue resident macrophages of the skin[37].

In summary, we showed that naïve progenitors give rise also to committed progenitors characterized by features of granulocytes and monocyte, cell types that emerge exclusively in the definitive waves[5].

CITE-seq analysis of human iPSC-derived haematopoietic progenitors.

To further study the temporal emergence of the progenitor populations and their associated markers, we carried out CITE-seq analysis whereby single cell membrane marker expression can be directly correlated with the relative transcriptome[17]. To ensure that we sampled even the rarest progenitor cell population we extended the CITE-seq analysis to adherent cells and included an earlier time point (day 10) in addition to day 13. Again, to exclude primitive erythroid cells, we selected CD235a-negative suspension cells but, in this experiment, we included and enriched for $CD43^-$ cells that had been excluded in our initial study. (Supplementary figure 5B). We expected early progenitors to express CD31 and to potentially remain part of hematopoietic clusters within the adherent fraction of the culture and so we FAC-sorted the adherent cells into $CD31^-$ and $CD31^+$ fractions. Cells were labeled with oligonucleotide tagged antibody specific for the membrane markers that we identified in our initial experiment (ADT_CD18, ADT_CD33, ADT_CD41, ADT_CD44, ADT_CD102, ADT_CD326; ADT: Antibody-Derived Tag) as well as other markers of endothelial and early committed hematopoietic cells (ADT_CD144) and of macrophages (ADT_CD163). To test the specificity of the membrane marker repertoire previously identified on the suspension cells, we subset the two libraries corresponding to suspension cells collected at day 10 and 13 (Figure 4, Supplementary Figure 5B-C). After multidimension reduction and clustering analysis we identified a naïve progenitor population (Figure 4A), comparable to our first sequencing experiment (Figure 2A). These naïve progenitors exhibited erythroid (Ery), megakaryocyte (Mega), and granulocyte and monocytes (Gra-Mo) lineage potential, with increased lineage commitment at day 13 compared to day 10 (Figure 4B); in line with the expression pattern of genes associated with naïve and committed stages in these days (Supplementary Figure 5D). Analysis of the ADTs showed that each marker was expressed in the expected cluster (Figure 4C) thus

298 supporting them as markers for defined progenitors. To further explore the power of the
299 ADT approach, we performed multidimension reduction using ADTs as the only input
300 dimensions and proved that ADT data alone identified remarkably similar clusters (Figure
301 4D-E), strongly correlated with the clusters derived from the entire transcriptome (Figure
302 4F). Taken together, the CITE-seq approach confirms that the markers identified from our
303 first scRNAseq analysis define the hierarchy of human developmental hematopoiesis in
304 vitro with high specificity.
305
306 **Comparison of in vitro generated progenitors with in vivo produced cells.**
307 The use of human PSCs as a renewable source of hematopoietic cell types faces major
308 challenges relating to, for example, the inefficient repopulation capacity of progenitor cells
309 and the incomplete maturation of differentiated cell types. To identify the underlying
310 molecular basis associated with these deficiencies in hPSC-derived cells, we compared
311 our dataset to a human fetal liver dataset which contains the complete hematopoietic
312 hierarchy from long-term reconstituting HSCs to differentiated cell types.
313 To assess how hPSCs-derived naïve and lineage-committed progenitors compared to their
314 equivalent counterpart generated in vivo, we assessed the expression of selected genes
315 identified to distinguish the various cell types detected in the human fetal liver[38] (Figure
316 4G). An initial analysis of marker genes of lineage commitment in the developing embryo
317 revealed that these markers are remarkably powerful for discriminating the equivalent in
318 vitro cell types identified in our in vitro study (Figure 4G, Supplementary Table 1).
319 Interestingly, *SPINK2,* a newly reported marker of fetal HSC/MPP[38], was also expressed
320 specifically by our naïve progenitor cells (Figure 4G), together with CD34 (Supplementary
321 figure 3L). These specific similarities observed between in vitro and in vivo developing
322 hematopoietic progenitor cells led us to investigate in a more comprehensive manner the
323 phenotype of cell types that are produced in vitro and how well these in vitro derived cells
324 reflect the corresponding cell types during in vivo development. Therefore, we used the
325 same published human fetal liver scRNAseq data as a reference, firstly, to identify in vitro
326 derived cells with gene expression signatures of human fetal liver hematopoietic cells and,
327 secondly, to quantify the similarity to their corresponding transcriptomes. To address the
328 first question, we employed machine learning to transfer labels from the fetal liver
329 reference data to our in vitro-derived blood cells (Figure 5A). This approach enabled a
330 much broader and unbiased identification of cell types compared to inference based purely

331 on marker genes. We followed our recently developed strategy[39] and trained an artificial
332 neural network (ANN)[39] to recognize single-cell gene expression profiles of human foetal
333 liver cells that were sampled at a time in development at which the liver is the main site of
334 blood cell formation[38]. Briefly, this ANN is trained using the expression data of 3,479
335 genes and 145,725 cells from fetal liver as an input[38]. From these labelled data, the ANN
336 learns to predict, from which of the 28 different fetal liver cell types a particular gene
337 expression pattern originates. Once trained, the ANN is given previously unseen test data
338 from in vitro derived cells as an input in order to annotate these data with human fetal liver
339 cell labels. Since this approach considers 3,479 genes, it enabled a more comprehensive
340 identification of cell types based on similarities in global gene expression patterns rather
341 than specific marker genes.
342 The ANN was able to identify cell types within the source domain (the fetal liver data) with
343 high accuracy as shown by the performance metrics obtained from 5-fold cross-validation
344 (Supplementary figure 6A-B). The trained ANN was subsequently applied to the target
345 domain (in vitro) to test if the hPSCs-derived cells were similar to those present in the
346 foetal liver, in which case the label of that specific in vivo cell would be transferred. The
347 ANN was able to assign labels to 92% of in vitro produced cells into various cell types
348 present in vivo (Supplementary figure 6 C-D), most notably, a small population was
349 labeled as HSC/MPP. This indicates that the global gene expression pattern of a subset of
350 the in vitro derived cells is very similar to HSC/MPPs from the in vivo reference data in
351 fetal liver. To quantify precisely how similar these in vitro derived HSC/MPPs are to their in
352 vivo counterparts, we calculated the average pairwise Euclidean distance between
353 HSC/MPPs, using the human fetal liver as a reference. This analysis indicates that fetal
354 liver HSC/MPPs are, on average, only marginally more similar to one another as they are
355 to iPSC derived HSC/MPPs (Supplementary Figure 8A). In summary, this analysis
356 indicates that the in vitro derived HSC/MPPs closely, yet not perfectly, reflect the gene
357 expression patterns of their in vivo counterparts. Using the ANN we also observed that the
358 relative abundance of the predicted HSC/MPP population decreased with time by day 13
359 (Figure 5B), whereas, the relative abundance of committed cells increased over this time
360 as expected (Supplementary Figure 6E). When we applied the same ANN strategy to our
361 first data set, that was generated from day 13 progenitors that were selected on the basis
362 of CD43 expression, no HSC/MPP were detected (Figure 5C). This is consistent with our
363 observation that this transient HSC/MPP population is present in higher numbers earlier at

day 10, when they are almost equally distributed in the adherent CD31+ and suspension CD235a- compartment (Supplementary Figure 6F). We looked for marker genes that defines this predicted HSC/MPP cell population in vitro and looked specifically for membrane markers according to their GO annotation (Supplementary Table 1). Together with expected markers such as *CD34*, *CD44* and *CD33*, we also detected *CD132*, *CD52*, *CD180* and *IL3RA* and many others that will allow to design a prospective sorting strategy to isolate this specific population. We then subset the in vivo and in vitro HSC/MPP and integrated the two datasets (Figure 5D). The integrated data allowed for direct comparison of their transcriptome and identified 54 differentially expressed genes (Supplementary Table 1), all of which were lower in HSC/MPP produced in vitro compared to those generated in vivo. GO analysis of these genes identified enrichment for KEGG signaling pathways such as NOD-like receptor, IL-17, NF-Kappa B and HIF-1 (Supplementary Table 1). We also identified 6 genes encoding transcription factors: *EGR1*, *ZFP36L1*, *NR4A1*, *FOS*, *JUN* and *JUNB* (Figure 5E). Interestingly, the EGR1 binding site was enriched, amongst others, in the upstream region of the differentially expressed genes (Figure 5F), suggesting an important regulatory role of EGR1.

We also compared the predicted HSC/MPP derived from hPSCs to hematopoietic progenitors isolated from different sited of hematopoiesis in the developing embryo including to fetal liver HSC/MPP[38], yolk sac MPP[38] that were collected at Carnegie stages 5 to 14, and AGM[40] sorted progenitors (CD34+CD45+CD235a-) collected at Carnegie stage 15, around the time of early HSC emergence (Supplementary Figure 7A-B). Whole transcriptome comparison, followed by KEGG pathway analysis, showed that in vitro HSC/MPP cells are marked by genes associated with oxidative phosphorylation (Supplementary Table 1), indicating metabolic differences between in vitro and in vivo produced progenitors. Hypoxic conditions characterize mammalian embryo develepoment[41], and more specifically the development of the hematopoietic system, where hypoxia has been detected in hematopoietic clusters in the AGM region, and in the fetal liver [42]. The hematopoietic progenitors derived from hPSCs were instead differentiated in normoxic conditions which could explain their different metabolic profile. The fetal liver cells were marked by HLA genes and consequently KEGG pathways associated with antigen presentation and T-cell development (Supplementary Table 1). The AGM dataset displayed high expression levels of genes associated with Notch pathway, such as *HES1*, *NOTCH1*, *NOTCH2*, *JAG1* and *JAG2*. This is in line with the

developmental stage at which they were collected when the Notch pathway is orchestrating the HSC emerge[43]. Within the markers of yolk sac progenitors, we detected genes related to early hematopoietic development. *FRZB*, mesodermal cell marker, and *HBE1*, marker of primitive hematopoiesis, were listed in the top 10 differentially expressed genes: this underlines the early developmental features of yolk sac progenitors. Finally, we noted also that SPINK1 was identified as marker for YS progenitors. While SPINK2, identified here and by others as marker of progenitor cells[38,40] was expressed by progenitors from all the tissues, SPINK1 was detected exclusively in the YS (Supplementary Figure 7X), suggesting that this gene could discriminate extraembryonic from intraembryonic hematopoiesis.

Finally, we compared lineage committed cells identified by the ANN, in our in vitro dataset, to their in vivo counterpart from fetal liver, to identify genes that can be used as targets to improve the production in vitro of differentiated blood cell types. We listed the differentially expressed genes between in vitro and fetal liver cells and identified the transcription factors within the list (Supplementary Figure 8, Supplementary Table 1). Particularly interesting, late erythroid cells in vitro show high level of PCLG2, phospholipase C gamma 2, able to control intracellular calcium via production of IP3, inositol triphosphate. Intracellular calcium peaks just before enucleation, prior nuclei extrusion in the orthochromatic erythroblasts[44]. Erythroid cells derived from hPSCs are characterized by a general inefficient enucleation[45,46], independently or their primitive or definitive origin (Supplementary Figure 1J) and this could be related to their intracellular calcium control.

In summary, we have identified a rare population of HSC/MPP-like cells in vitro that emerge early during differentiation of hPSCs and that display broadly similar gene expression patterns when compared to HSCs in development. However subtle differences are also apparent and a more detailed study of these differences could explain the known deficiencies of PSC-derived cells and ultimately be exploited to improve their therapeutic use. Our novel approach combines scRNAseq and machine learning to help identify candidate genes that may improve the production of HSCs and mature lineage cells from pluripotent stem cells in vitro, by closely recapitulating in vivo hematopoiesis.

Discussion

We described the single cell transcriptome and membrane markers of naïve hematopoietic progenitors and their lineage committed descendants derived in vitro from human pluripotent stem cells. The repertoire of membrane markers proved to be remarkably accurate in capturing the different states prior to and after lineage commitment.

We identified a population of naïve progenitors situated at the top of the differentiation hierarchy, marked by CD44, a protein involved in the hematopoietic transition of the hemogenic endothelium in the mouse AGM region[47]. We validated their lineage potential employing a chimeric culture system, where isolated naïve progenitors, marked by Zeiss-Green expression, demonstrated overlapping lineage output to that predicted in silico.

We also observed that progenitors are capable of moving between the naïve states, as well as progressing into committed states. This is in keeping with many other scRNAseq and proteomic studies that have reported a continuum of cell states as opposed to sequential discrete cell types hierarchies[48–51]. In line with a recent murine study[47], we have shown that CD44 is expressed in naïve hPSCs-derived progenitors and here we demonstrated that both human and mouse progenitors also express LMO4, a LIM-domain protein[52]. Recent scRNAseq detected LMO4 in both human granulocyte progenitors in the bone marrow[48] and adult mouse HSC[53], but its associated proteins have not been identified. We also reported high levels of *ID* genes within the progenitors, target genes of BMP signaling known to be involved in HSC emergence[54–56]. IDs, like LMOs proteins, do not present DNA binding domain and rather act through binding of other proteins in complexes also involved in HSPC development[57] and erythropoiesis[58] . Overexpression of ID2 in human HSC from cord blood has been reported to enhance their functional stemness in vivo[59], supporting the idea that this class of proteins might maintain the progenitor status and thus might be useful in alternative programming strategies of hPSCs.

The use of scRNAseq on vast numbers of cells allows to detect even the rarest cell population and we considered that it might enable the detection of rare HSC-like cells in differentiating hPSCs cultures. We showed to hPSCs-derived cells showed a remarkable specific expression pattern of marker genes identified in the human embryo, for example, *SPINK2*, a novel marker of human fetal liver HSC and MPP, marked also our naïve progenitors. By using machine learning we identify specific cell types sampled in vivo and detected a small and transient population of HSC-like cells that, when compared to their in vivo counterpart from fetal liver, showed only small transcriptional differences. Previous reports described the hematopoietic progenitors obtained with the differentiation employed

463 in this work as intraembryonic-like[18], using T-cells lineage as hallmark of definitive
464 hematopoiesis. However, yolk sac shows T-cell potential prior to HSC emergence[60,61],
465 thus limiting the use of T-cell assay alone as discriminative of the corresponding
466 developmental wave. Our machine learning approach and the detection of HSC-like cells
467 strongly supports the intraembryonic identities of the hematopoietic cells differentiated in
468 vitro and provide an alternative and multifactorial approach to address questions regarding
469 the similarities to developmental counterparts. The unbiased and comprehensive
470 comparison used in this study allowed us to pinpoint differentially expressed genes
471 between in vitro-derived and in vivo HSCs can now be exploited to improve production of
472 HSC in vitro. Our analysis indicates in vitro HSC-like cells do not express CD43,
473 comparable to mouse Pro-HSC prior their maturation into functional definitive HSC[62]. This
474 could suggest that the widely acknowledged inability of hPSCs-derived progenitors to
475 reconstitute the hematopoietic system, could be due to their immature phenotype and the
476 lack of appropriate culture conditions for HSCs maturation and maintenance. In addition,
477 the identification of the HSC-like population using our machine learning approach, which is
478 based on high similarity in the gene expression profiles, could suggest that the molecular
479 basis of the functional deficiency of this in vitro derived population could reside at a post-
480 transcriptional level. Thus, future experiment will be required to assess whether a further
481 ad-hoc maturation step of sorted HSC-like cells would achieve reconstitution.
482 When we compared the hematopoietic progenitors developed in the human embryo
483 throughout gestations together with those derived in vitro, we found that while SPINK2
484 was expressed by all progenitors, *SPINK1* was exclusively detected in cells from the yolk
485 sac. SPINK1 is able to bind to EGFR and induce epithelial to mesenchymal transition in
486 cancer cells[63,64], a process similar to the endothelial to hematopoietic transition, where the
487 role of SPINK1 remains largely unexplored. In summary, we propose here SPINK1 as a
488 possible marker for primitive hematopoiesis which could be an extremely useful genes to
489 trace the cells that colonize the embryo from the yolk sac.
490 The differentiation protocol used in this study is well defined and serum-free and is one of
491 the most commonly used protocols used by many laboratories. We also showed that our
492 markers are able to identify functionally similar progenitors in different cell lines. Thus, our
493 browsable datasets and the findings of our study will be of interest to many in the field of
494 hematopoiesis and will allow to test how the frequency of this populations vary in response
495 to different cytokines conditions. In addition, the increasing availability of large scRNAseq

dataset of human tissue makes our pipeline applicable to the analyses of other systems where the hPSCs differentiation aims to produce adult-like cells for clinical application. In this way cell types differentiated in vitro can now be annotated in an unbiased manner that does not rely on few known markers and allows the identification of transcriptional discrepancies between cell types produced in vitro and their in vivo counterparts.
In conclusion, our browsable dataset provides a comprehensive transcriptional characterization of in vitro derived hematopoietic progenitors. This work defines the makeup of the progenitor populations that give rise to immune cells, such as macrophages and granulocytes, as well as HSC-like cells, which holds great promise for their therapeutically application.

Acknowledgment

The work was funded by Wellcome Trust (Grant No. 102610), MRC Innovate UK (Grant No. 102853), BBSRC (Grant No. S002219/1). AF received a Carnegie Incentive Grant (Grant No. RIG008218). AB received BBSRC Future Leaders Fellowship (Grant reference BB/N011597/1). Sequencing was carried out by Edinburgh Genomics, The University of Edinburgh. Edinburgh Genomics is partly supported through core grants from NERC (R8/H10/56), MRC (MR/K001744/1) and BBSRC (BB/J004243/1). We thank Professor Ben Macarthur for suggesting the application of machine learning to our study; Andrew Elefanty for sharing the RUNX1C-GFP cell line; Fiona Rossi, Claire Cryer, Bindi Heer and Andrea Corsinotti from the Flow Facility as well as Bertand Verney and Matthieu Vermeren from the imaging facility. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/).

Author Contribution

AF, designed and performed research, analyzed the data and wrote the manuscript. AF, PS, AB and NR performed bioinformatics analysis. PR, ST, MLY, AHT, JE, BH, RA performed research. LMF designed the experiment, analyzed data and wrote the manuscript. NH, AM, KO, and NR provided intellectual input and final approval of the manuscript.

Declaration of interest

Authors declare no competing interests.

References

1. Vo LT, Daley GQ. De novo generation of HSCs from somatic and pluripotent stem cell sources. *Blood*. 2015;125(17):2641–8.
2. Palis J. Hematopoietic stem cell-independent hematopoiesis: emergence of erythroid, megakaryocyte, and myeloid potential in the mammalian embryo. *FEBS Lett.* 2016;590(22):3965–3974.
3. Tober J, Koniski A, McGrath KE, et al. The megakaryocyte lineage originates from hemangioblast precursors and is an integral component both of primitive and of definitive hematopoiesis. *Blood*. 2007;109(4):1433–41.
4. Palis J, Robertson S, Kennedy M, Wall C, Keller G. Development of erythroid and myeloid progenitors in the yolk sac and embryo proper of the mouse. *Development*. 1999;126(22):5073–84.
5. McGrath KE, Frame JM, Fegan KH, et al. Distinct Sources of Hematopoietic Progenitors Emerge before HSCs and Provide Functional Blood Cells in the Mammalian Embryo. *Cell Rep.* 2015;
6. Mass E, Ballesteros I, Farlik M, et al. Specification of tissue-resident macrophages during organogenesis. *Science (80-.).* 2016;353:6304.
7. Schulz C, Gomez Perdiguero E, Chorro L, et al. A lineage of myeloid cells independent of Myb and hematopoietic stem cells. *Science (80-.).* 2012;336(6077):86–90.
8. Stremmel C, Schuchert R, Wagner F, et al. Yolk sac macrophage progenitors traffic to the embryo during defined stages of development. *Nat. Commun.* 2018;9(1):75.
9. Medvinsky A, Dzierzak E. Definitive hematopoiesis is autonomously initiated by the AGM region. *Cell*. 1996;86(6):897–906.
10. Ivanovs A, Rybtsov S, Welch L, et al. Highly potent human hematopoietic stem cells first emerge in the intraembryonic aorta-gonad-mesonephros region. *J. Exp. Med.* 2011;208(12):2417–2427.
11. Easterbrook J, Fidanza A, Forrester LM. Concise review: Programming human pluripotent stem cells into blood. *Br. J. Haematol.* 2016;173(5):.
12. Ivanovs A, Rybtsov S, Anderson RA, Turner ML, Medvinsky A. Identification of the niche and phenotype of the first human hematopoietic stem cells. *Stem Cell Reports*. 2014;2(4):449–456.
13. Tavian M, Hallais MF, Péault B. Emergence of intraembryonic hematopoietic precursors in the pre-liver human embryo. *Development*. 1999;126(4):793–803.
14. Guibentif C, Rönn RE, Böiers C, et al. Single-Cell Analysis Identifies Distinct Stages of Human Endothelial-to-Hematopoietic Transition. *Cell Rep.* 2017;19(1):10–19.
15. Angelos MG, Abrahante JE, Blum RH, Kaufman DS. Single Cell Resolution of Human Hematoendothelial Cells Defines Transcriptional Signatures of Hemogenic Endothelium. *Stem Cells*. 2018;36(2):206–217.
16. Han X, Chen H, Huang D, et al. Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell RNA-sequencing. *Genome Biol.* 2018;19(1):1–19.
17. Stoeckius M, Hafemeister C, Stephenson W, et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods*. 2017;14(9):865–868.
18. Sturgeon CM, Ditadi A, Awong G, Kennedy M, Keller G. Wnt signaling controls the specification of definitive and primitive hematopoiesis from human pluripotent stem

577 cells. *Nat. Biotechnol.* 2014;32(6):554–561.
578 19. Vodyanik MA, Thomson JA, Slukvin II, Dulac C, Péault B. Leukosialin (CD43)
579 defines hematopoietic progenitors in human embryonic stem cell differentiation
580 cultures. *Blood*. 2006;108(6):2095–105.
581 20. Garcia-Alegria E, Menegatti S, Fadlullah MZH, et al. Early Human Hemogenic
582 Endothelium Generates Primitive and Definitive Hematopoiesis In Vitro. *Stem Cell*
583 *Reports*. 2018;11(5):1061–1074.
584 21. Ng ES, Azzola L, Bruveris FF, et al. Differentiation of human embryonic stem cells to
585 HOXA+ hemogenic vasculature that resembles the aorta-gonad-mesonephros. *Nat.*
586 *Biotechnol.* 2016;34(11):1168–1179.
587 22. Sroczynska P, Lancrin C, Kouskoff V, Lacaud G. The differential activities of Runx1
588 promoters define milestones during embryonic hematopoiesis. *Blood*.
589 2009;114(26):5279–89.
590 23. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell
591 transcriptomic data across different conditions, technologies, and species. *Nat.*
592 *Biotechnol.* 2018;36(5):411–420.
593 24. Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime
594 robustly reconstructs lineage branching. *Nat. Methods*. 2016;13(10):845–848.
595 25. Qiu X, Mao Q, Tang Y, et al. Reversed graph embedding resolves complex single-
596 cell trajectories. *Nat. Methods*. 2017;14(10):979–982.
597 26. Wolf FA, Hamey FK, Plass M, et al. PAGA: graph abstraction reconciles clustering
598 with trajectory inference through a topology preserving map of single cells. *Genome*
599 *Biol.* 2019;20(1):1–9.
600 27. Moreau T, Evans AL, Vasquez L, et al. Large-scale production of megakaryocytes
601 from human pluripotent stem cells by chemically defined forward programming. *Nat.*
602 *Commun.* 2016;7(1):11208.
603 28. Stumpf PS, Smith RCG, Lenz M, et al. Stem Cell Differentiation as a Non-Markov
604 Stochastic Process. *Cell Syst.* 2017;5(3):268-282.e7.
605 29. Chan TE, Stumpf MPH, Babtie AC. Gene Regulatory Network Inference from Single-
606 Cell Data Using Multivariate Information Measures. *Cell Syst.* 2017;5(3):251-267.e3.
607 30. Lopez-Yrigoyen M, Fidanza A, Cassetta L, et al. A human iPSC line capable of
608 differentiating into functional macrophages expressing ZsGreen: A tool for the study
609 and in vivo tracking of therapeutic cells. *Philos. Trans. R. Soc. B Biol. Sci.*
610 2018;373(1750):.
611 31. Jaffredo T, Gautier R, Eichmann A, Dieterlen-Lièvre F. Intraaortic hemopoietic cells
612 are derived from endothelial cells during ontogeny. *Development*.
613 1998;125(22):4575–83.
614 32. Zovein AC, Hofmann JJ, Lynch M, et al. Fate Tracing Reveals the Endothelial Origin
615 of Hematopoietic Stem Cells. *Cell Stem Cell*. 2008;3(6):625–636.
616 33. Jayapal SR, Lee KL, Ji P, et al. Down-regulation of Myc is essential for terminal
617 erythroid maturation. *J. Biol. Chem.* 2010;285(51):40252–65.
618 34. Lammers R, Giesert C, Grünebach F, et al. Monoclonal antibody 9C4 recognizes
619 epithelial cellular adhesion molecule, a cell surface antigen expressed in early steps
620 of erythropoiesis. *Exp. Hematol.* 2002;30(6):537–545.
621 35. Clay D, Rubinstein E, Mishal Z, et al. CD9 and megakaryocyte differentiation. *Blood*.
622 2001;97(7):1982–1989.
623 36. Kalev-Zylinska ML, Horsfield JA, Flores MVC, et al. Runx3 is required for
624 hematopoietic development in zebrafish. *Dev. Dyn.* 2003;228(3):323–336.
625 37. Fainaru O, Woolf E, Lotem J, et al. Runx3 regulates mouse TGF-b-mediated

626 dendritic cell function and its absence results in airway inflammation. *EMBO J.*
627 2004;23:969–979.
628 38. Popescu DM, Botting RA, Stephenson E, et al. Decoding human fetal liver
629 haematopoiesis. *Nature*. 2019;574(7778):365–371.
630 39. Stumpf PS, Du D, Imanishi H, et al. Mapping biology from mouse to man using
631 transfer learning. *bioRxiv*. 2019;
632 40. Zeng Y, He J, Bai Z, et al. Tracing the first hematopoietic stem cell generation in
633 human embryo by single-cell RNA sequencing. *Cell Res.* 2019;29(11):881–894.
634 41. Dunwoodie SL. The Role of Hypoxia in Development of the Mammalian Embryo.
635 *Dev. Cell*. 2009;17(6):755–773.
636 42. Imanirad P, Solaimani Kartalaei P, Crisan M, et al. HIF1α is a regulator of
637 hematopoietic progenitor and stem cell development in hypoxic sites of the mouse
638 embryo. *Stem Cell Res.* 2014;12(1):24–35.
639 43. Bigas A, Espinosa L. Hematopoietic stem cells: To be or Notch to be. *Blood*.
640 2012;119(14):3226–3235.
641 44. Wölwer CB, Pase LB, Russell SM, Humbert PO. Calcium signaling is required for
642 erythroid enucleation. *PLoS One*. 2016;11(1):.
643 45. Lopez-Yrigoyen M, Yang C-T, Fidanza A, et al. Genetic programming of
644 macrophages generates an in vitro model for the human erythroid island niche. *Nat.*
645 *Commun.* 2019;10(1):881.
646 46. Yang C-T, Ma R, Axton RA, et al. Activation of KLF1 Enhances the Differentiation
647 and Maturation of Red Blood Cells from Human Pluripotent Stem Cells. *Stem Cells*.
648 2017;35(4):886–897.
649 47. Oatley M, Bölükbası ÖV, Svensson V, et al. Single-cell transcriptomics identifies
650 CD44 as a marker and regulator of endothelial to haematopoietic transition. *Nat.*
651 *Commun.* 2020;11(1):.
652 48. Paul F, Arkin Y, Giladi A, et al. Transcriptional Heterogeneity and Lineage
653 Commitment in Myeloid Progenitors. *Cell*. 2015;163(7):1663–1677.
654 49. Velten L, Haas SF, Raffel S, et al. Human haematopoietic stem cell lineage
655 commitment is a continuous process. *Nat. Cell Biol.* 2017;19(4), 271–281.
656 50. Rodriguez-Fraticelli AE, Wolock SL, Weinreb CS, et al. Clonal analysis of lineage
657 fate in native haematopoiesis. *Nature*. 2018;553(7687):212–216.
658 51. Knapp DJHF, Hammond CA, Wang F, et al. A topological view of human CD34+ cell
659 state trajectories from integrated single-cell output and proteomic data. *Blood*.
660 2019;133(9):927-939.
661 52. Grutz G, Forster A, Rabbitts TH. Identification of the LMO4 gene encoding an
662 interaction partner of the LIM-binding protein LDB1/NLI1: a candidate for
663 displacement by LMO proteins in T cell acute leukaemia. *Oncogene*.
664 1998;17(21):2799–2803.
665 53. Lai S, Huang W, Xu Y, et al. Cell Discovery Comparative transcriptomic analysis of
666 hematopoietic system between human and mouse by Microwell-seq. *Cell Discov.*
667 2018;4:34.
668 54. Souilhol C, Gonneau C, Lendinez JG, et al. Inductive interactions mediated by
669 interplay of asymmetric signalling underlie development of adult haematopoietic
670 stem cells. *Nat. Commun.* 2016;(2016): 1-13.
671 55. Crisan M, Kartalaei PS, Vink CS, et al. BMP signalling differentially regulates distinct
672 haematopoietic stem cell types. *Nat. Commun.* 2015;6(1):8040.
673 56. McGarvey AC, Rybtsov S, Souilhol C, et al. A molecular roadmap of the AGM region
674 reveals BMPER as a novel regulator of HSC maturation. *J. Exp. Med.*

2017;214(12):3731–3751.
57. Wilson NK, Foster SD, Wang X, et al. Combinatorial Transcriptional Control In Blood Stem/Progenitor Cells: Genome-wide Analysis of Ten Major Transcriptional Regulators. *Cell Stem Cell*. 2010;7(4):532–544.
58. Wadman IA, Osada H, Grütz GG, et al. The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *EMBO J.* 1997;16(11):3145–57.
59. van Galen P, Kreso A, Wienholds E, et al. Reduced Lymphoid Lineage Priming Promotes Human Hematopoietic Stem Cell Expansion. *Cell Stem Cell*. 2014;14(1):94–106.
60. Yoshimoto M, Porayette P, Glosson NL, et al. Autonomous murine T-cell progenitor production in the extra-embryonic yolk sac before HSC emergence. *Blood*. 2012;119(24):5706–14.
61. Gentek R, Ghigo C, Hoeffel G, et al. Epidermal γδ T cells originate from yolk sac hematopoiesis and clonally self-renew in the adult. *J. Exp. Med.* 2018;215(12):2994–3005.
62. Rybtsov S, Batsivari A, Bilotkach K, et al. Tracing the origin of the HSC hierarchy reveals an SCF-dependent, IL-3-independent CD43- embryonic precursor. *Stem Cell Reports*. 2014;3(3):489–501.
63. Wang C, Wang L, Su B, et al. Serine protease inhibitor Kazal type 1 promotes epithelial-mesenchymal transition through EGFR signaling pathway in prostate cancer. *Prostate*. 2014;74(7):689–701.
64. Chen F, Long Q, Fu D, et al. Targeting SPINK1 in the damaged tumour microenvironment alleviates therapeutic resistance. *Nat. Commun.* 2018;9(1):1–19.

Figure Legends

Figure 1 - Single cell transcriptome analysis reveals clusters of naïve and lineage-committed haematopoietic progenitors.

(A) Schematic of the single cell RNA sequencing experiment where iPSCs (SFCi55) were differentiated in vitro (IVD) for 13 days (Supplementary Figure1A), $CD235a^{-}CD43^{+}$ suspension cells were isolated by flow cytometry and subjected to 10x genomics sequencing platform. **(B)** tSNE visualization of 11,420 cells divided into 8 clusters including clusters defined by gene expression as naïve (naïve 1 and 2), and others that expressed genes associated with erythroid (Ery1 and 2), megakaryocyte (Mega 1,2 and 3) and granulocyte (granulo) lineages. **(C)** Heatmap showing expression of the top 10 marker genes for each cluster (colors for each cluster as in Figure 1B). **(D)** Gene expression levels of marker genes associated with different progenitor cell types that were identified by clustering, visualized on tSNE.

Figure 2 - Trajectory analyses support naïve progenitor identity and their progression to lineage committed progenitors.

(A) Diffusion plot displays the naïve progenitors in the core region of the plot from where the three direction of commitment originates, the arrows indicates the commitment directions. **(B)** Representation of each cluster on the diffusion plot. **(C)** PAGA analysis show that naïve cluster are connected to the lineage committed cells. Each node contains

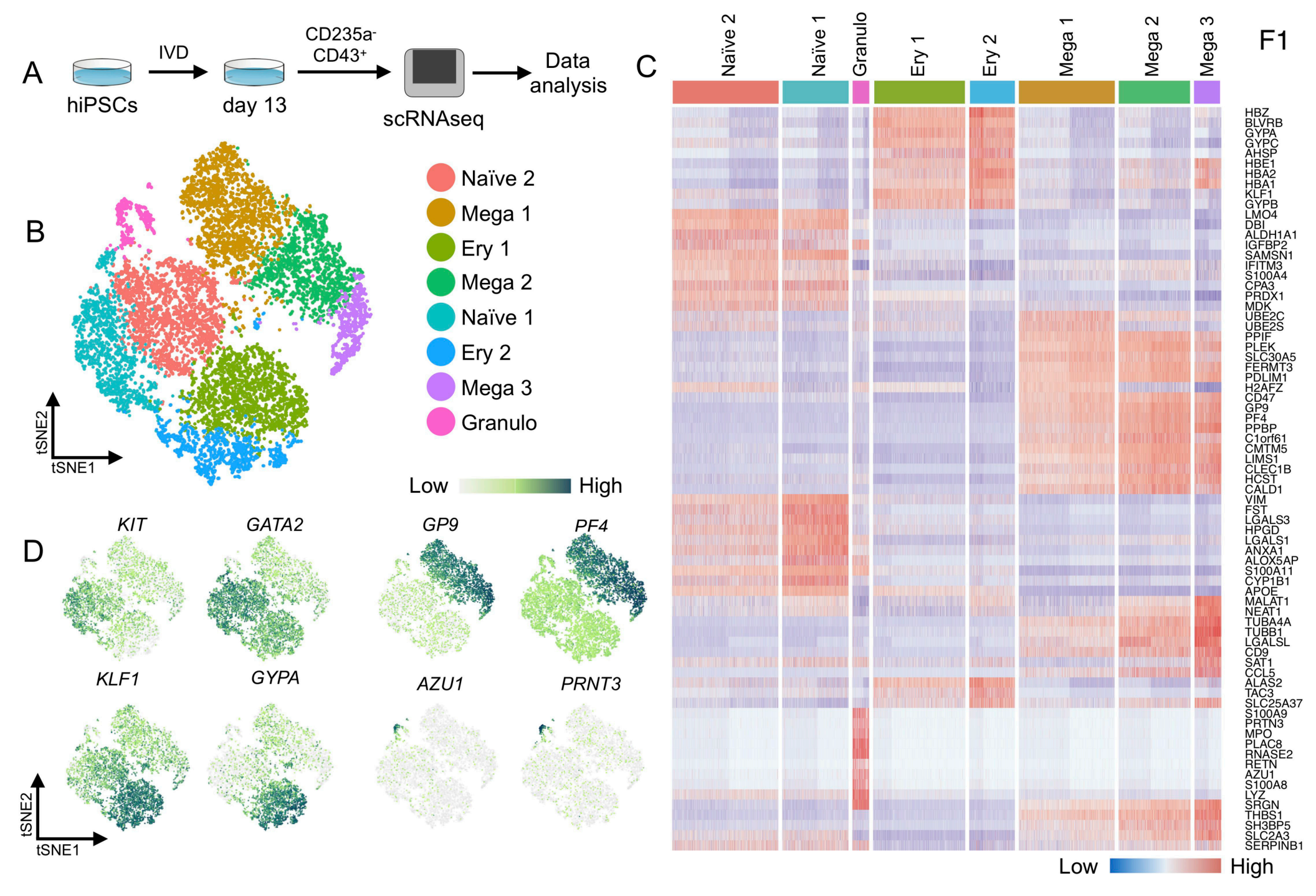
a pie chart showing the proportion of cells for each cluster. Colors indicate cluster identities. **(D)** Monocle trajectory analyses demonstrates a similar pattern to that obtained from the diffusion plot shown in A with naïve progenitors at the top of the hierarchy, with progression toward committed. **(E)** Monocle trajectory visualizing each cluster individually. **(F)** Expression levels of the marker genes coding for transcription factors associated with each cluster, bars color indicates the cluster.

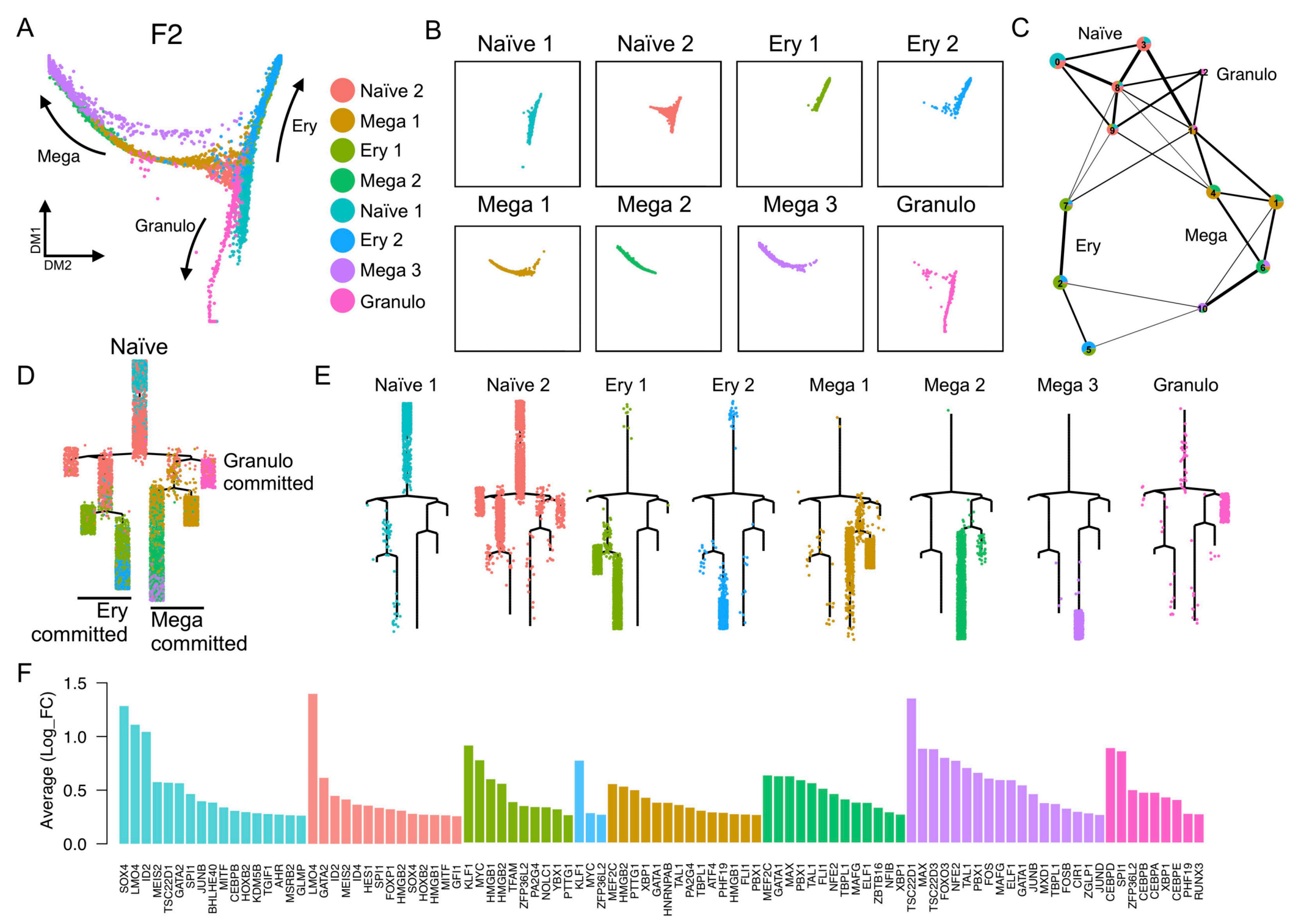
Figure 3 - CD44 identifies clonogenic hematopoietic progenitors.
(A) Expression levels of genes encoding cell surface markers, *CD33*, *CD44* and *CD18* that were associated with the naïve progenitor clusters. **(B)** Scatter plot of flow cytometry profile of naïve 1A, 1B and 2 cells at day 13 of differentiation (hiPSCs-SFCi55). Cells are gated on CD235a^{-}CD43^{+}CD33^{+}. **(C)** Schematic of the chimeric culture system using the ZsGreen reporter to trace cells during the differentiation process. ZsGreen and parental line (SFCi55) were differentiated in a synchronous manner, at day 10 naïve 1 and naïve 2 cells are sorted and co-cultured with the parental line differentiation. Co-culture is then analyzed at day 13. **(D)** Representative flow cytometry profile of the day 13 naïve descendants' cells after sorting at day 10 and chimeric co-culturing of naïve 1 (teal) and naïve 2 (pink) cells. Contribution of naïve 1, in teal, and naïve 2, in pink, to the naïve 1A, 1B and 2 compartment (n=6, multinomial logistic regression, *p<0.05, **p<0.01, ***p<0.005). **(E)** CFU-C analyses of FAC-sorted naïve 1 and naïve 2 cells from day 10 (n=3, paired t-Test p=0.0753) (hiPSCs-SFCi55). **(F)** CFU-C analyses of FAC-sorted naïve 1A, 1B and 2 cells from day 13 (n=9, Holm-Sidak's test, p<=0.001) (hiPSCs-SFCi55). **(G)** RUNX1-GFP expression in naïve 1 and naïve 2 at both day 10 and day 13 (n=12, paired t test; * p<0.05, ** p<0.01).

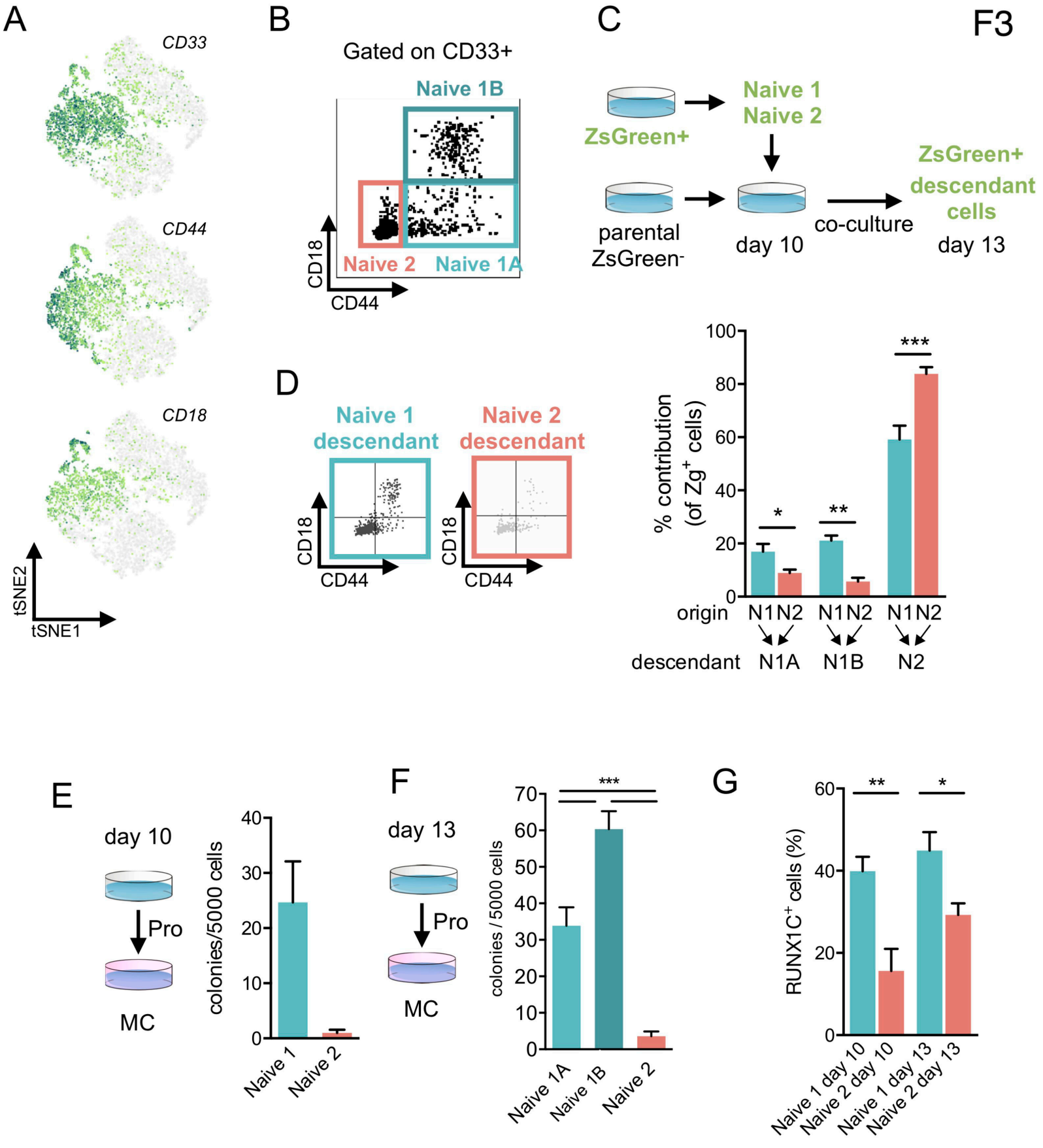
Figure 4 - CITE-seq analyses confirm markers for naïve and lineage-committed progenitor cells.
(A) tSNE visualization of the CITE-seq analysis of day 10 and day 13 CD235a- suspension cells; reduction and clustering were performed using only transcriptomic data (hiPSC-SFCi55). **(B)** tSNE visualization of the libraries obtained from CD235a^{-} suspension cells collected at day 10 (pink) and day 13 (teal) showing lineage commitment direction. **(C)** Single cell protein expression level of the membrane markers associated with the different cell types. Data are visualized on tSNE (ADT = antibody derived tags). **(D)** tSNE plot and annotation of clustering obtained from analysis derived from ADT data alone. **(E)** Visualization of the libraries obtained from CD235a^{-} suspension cells projected on the tSNE obtained from ADT data in D, cell progression shows lineage commitment trajectory. Cells are colored according to the day of collection (day 10 in pink and day 13 in teal). **(F)** Confusion matrix of clustering obtained from complete transcriptomic data (RNA) and that obtained from ADT data alone. Color of each box indicates the % of cells classified in each RNA versus ADT cluster. **(G)** Gene expression levels of human foetal gene marker genes in our naïve and lineage committed progenitors.

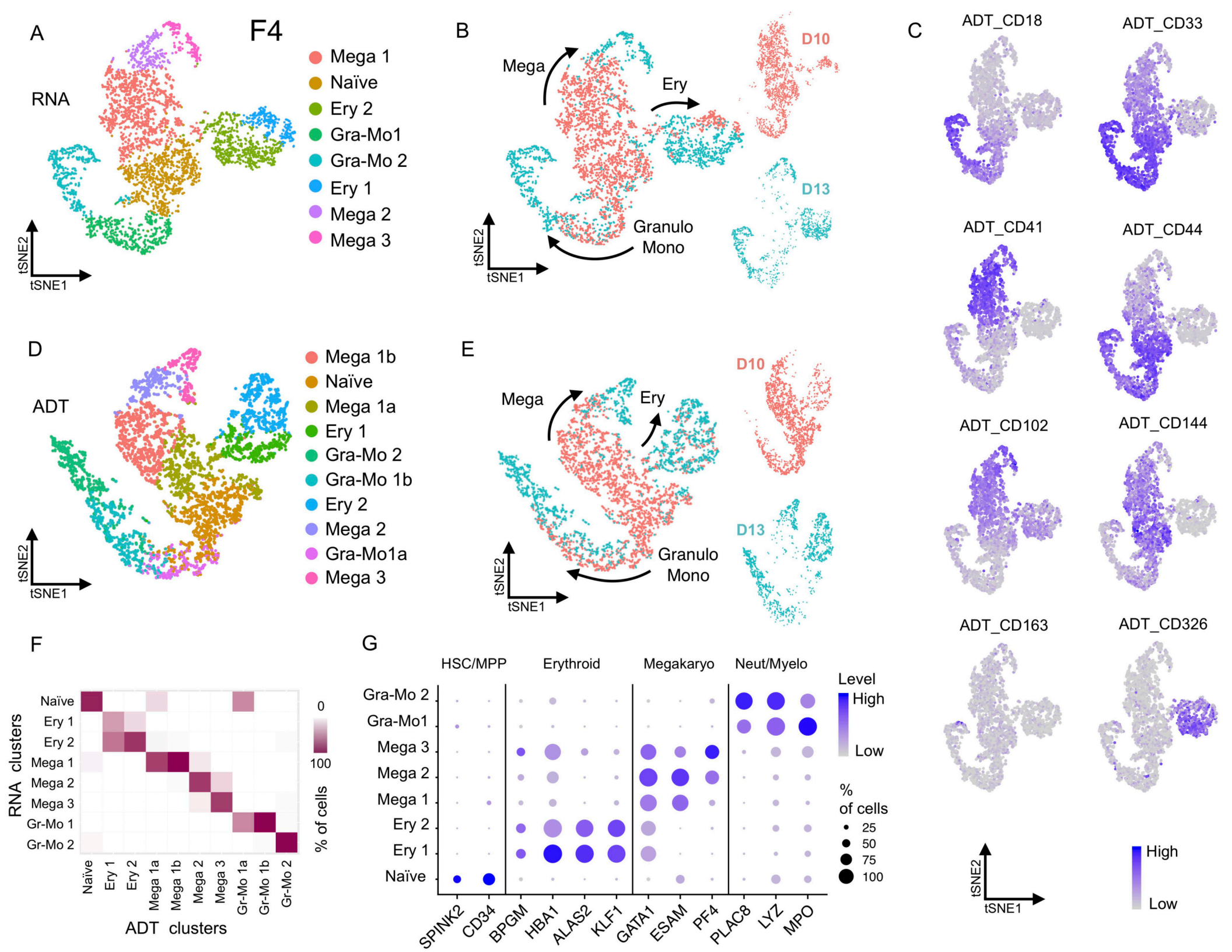
Figure 5 - Artificial neural network identifies HSC-like cells in iPSC derived hematopoietic cells.

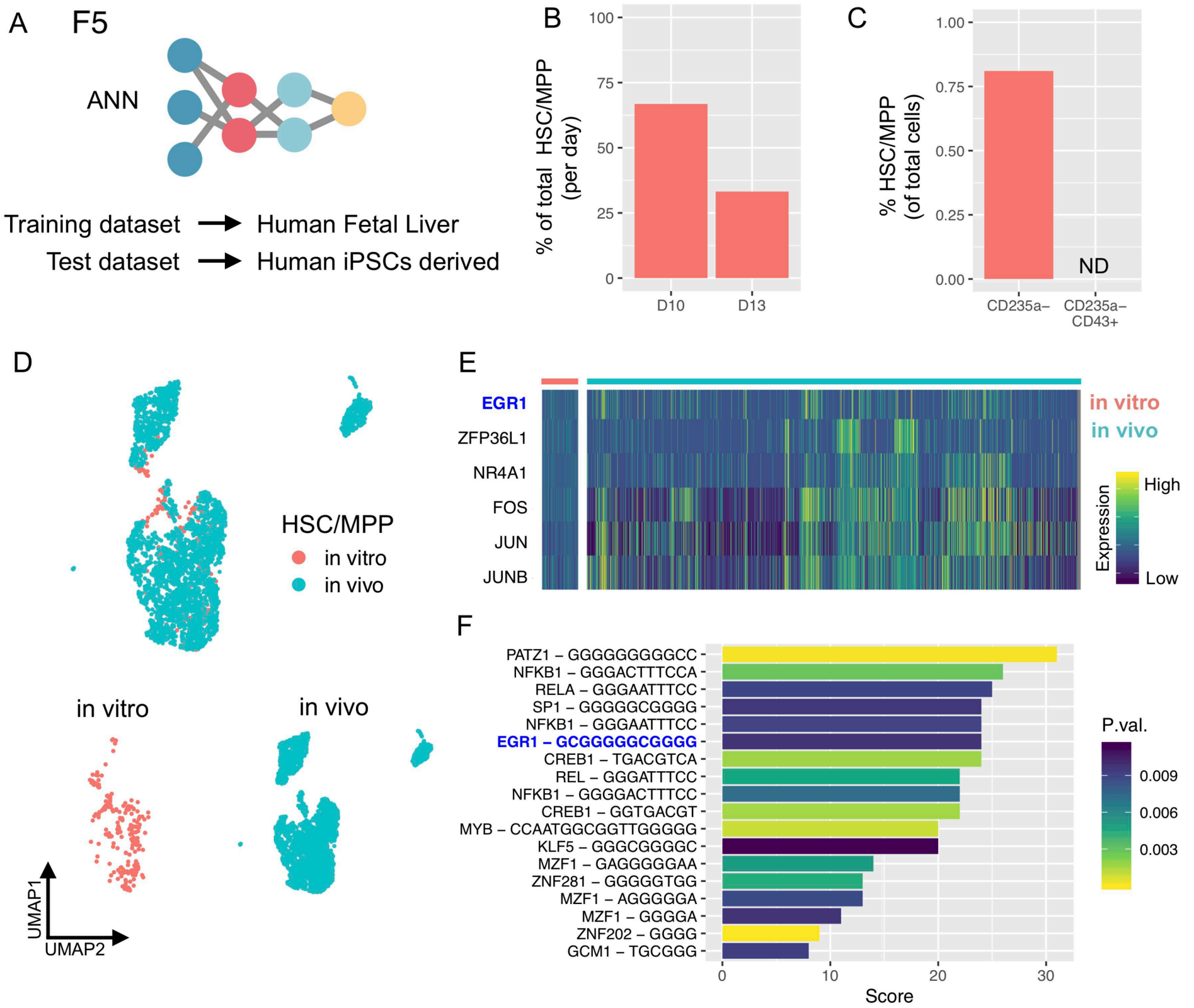
((A) Schematic of the artificial neural network (ANN) architecture for label-transfer. An input layer (3479 units), two fully connected hidden layers (64 and 32 units) and a 28-unit softmax layer corresponding to cell types in the source domain (human foetal liver scRNAseq data) used for training. Classification of cell types in the target domain of human iPSC-derived single cell transcriptomes (test data). **(B)** Proportion of cells labelled HSC/MPPs at day 10 or day 13 of hiPSC differentiation in vitro. **(C)** Proportion of in vitro derived CD235a^{-} progenitors and CD235a^{-}CD43^{+} cells labelled HSC/MPP by the ANN (ND = not detected). **(D)** UMAP visualization of the integrated dataset containing in vivo derived (blue) and in vitro annotated (pink) HSC/MPPs. **(E)** Heatmap of differentially expressed genes coding for transcription factors obtained from the comparison of in vivo and in vitro derived HSC/MPPs. **(F)** Transcription factor binding motifs enriched upstream of the 54 genes identified as differentially expressed between HSC/MPP generated in vitro and in vivo.


A
hiPSCs
IVD
day 13
CD235a-
CD43+
scRNAseq
Data analysis
B
tSNE2
tSNE1
Naïve 2
Mega 1
Ery 1
Mega 2
Naïve 1
Ery 2
Mega 3
Granulo
Low
High
D
KIT
GATA2
GP9
PF4
KLF1
GYPA
AZU1
PRNT3
tSNE2
tSNE1
C
Naïve 2
Naïve 1
Granulo
Ery 1
Ery 2
Mega 1
Mega 2
Mega 3
F1
HBZ
BLVRB
GYPA
GYPC
AHSP
HBE1
HBA2
HBA1
KLF1
GYPB
LMO4
DBI
ALDH1A1
IGFBP2
SAMSN1
IFITM3
S100A4
CPA3
PRDX1
MDK
UBE2C
UBE2S
PPIF
PLEK
SLC30A5
FERMT3
PDLIM1
H2AFZ
CD47
GP9
PF4
PPBP
C1orf61
CMTM5
LIMS1
CLEC1B
HCST
CALD1
VIM
FST
LGALS3
HPGD
LGALS1
ANXA1
ALOX5AP
S100A11
CYP1B1
APOE
MALAT1
NEAT1
TUBA4A
TUBB1
LGALSL
CD9
SAT1
CCL5
ALAS2
TAC3
SLC25A37
S100A9
PRTN3
MPO
PLAC8
RNASE2
RETN
AZU1
S100A8
LYZ
SRGN
THBS1
SH3BP5
SLC2A3
SERPINB1
Low
High


A
F2
Mega
Ery
Granulo
DM1
DM2
Naïve 2
Mega 1
Ery 1
Mega 2
Naïve 1
Ery 2
Mega 3
Granulo
B
Naïve 1
Naïve 2
Ery 1
Ery 2
Mega 1
Mega 2
Mega 3
Granulo
C
Naïve
Granulo
Ery
Mega
D
Naïve
Granulo committed
Ery committed
Mega committed
E
Naïve 1
Naïve 2
Ery 1
Ery 2
Mega 1
Mega 2
Mega 3
Granulo
F
Average (Log_FC)
1.5
1.0
0.5
0.0
SOX4 LMO4 ID2 MEIS2 TSC22D1 GATA2 SPI1 JUNB BHLHE40 MITF CEBPB HOXB2 KDM5B TGIF1 AHR MSRB2 GLMP
LMO4 GATA2 ID2 MEIS2 ID4 HES1 SPI1 FOXP1 HMGB2 SOX4 HOXB2 HMGB1 MITF GFI1
KLF1 MYC HMGB1 HMGB2 TFAM ZFP36L2 PA2G4 NOLC1 YBX1 PTTG1
KLF1 MYC ZFP36L2 MEF2C HMGB2 PTTG1 XBP1 GATA1 HNRNPAB TAL1 PA2G4 TBPL1 ATF4 PHF19 HMGB1 FLI1 PBX1
MEF2C GATA1 MAX PBX1 TAL1 FLI1 NFE2 TBPL1 MAFG ELF1 ZBTB16 NFIB XBP1
TSC22D1 MAX TSC22D3 FOXO3 NFE2 TAL1 PBX1 FOS MAFG ELF1 GATA1 JUNB MXD1 TBPL1 FOSB CIR1 ZGLP1 JUND
CEBPD SPI1 ZFP36L2 CEBPB CEBPA XBP1 CEBPE PHF19 RUNX3


A
CD33
CD44
CD18
tSNE2
tSNE1
B
Gated on CD33+
Naive 1B
Naive 2
Naive 1A
CD18
CD44
C
ZsGreen+
Naive 1
Naive 2
parental
ZsGreen-
day 10
co-culture
ZsGreen+
descendant
cells
day 13
D
Naive 1
descendant
Naive 2
descendant
CD18
CD44
% contribution
(of Zg+ cells)
origin
descendant
N1A
N1B
N2
E
day 10
Pro
MC
colonies/5000 cells
Naive 1
Naive 2
F
day 13
Pro
MC
colonies / 5000 cells
Naive 1A
Naive 1B
Naive 2
G
RUNX1C+ cells (%)
Naive 1 day 10
Naive 2 day 10
Naive 1 day 13
Naive 2 day 13


A
F4
RNA
Mega 1
Naïve
Ery 2
Gra-Mo1
Gra-Mo 2
Ery 1
Mega 2
Mega 3
tSNE2
tSNE1
B
Mega
Ery
D10
D13
Granulo
Mono
tSNE2
tSNE1
C
ADT_CD18
ADT_CD33
ADT_CD41
ADT_CD44
ADT_CD102
ADT_CD144
ADT_CD163
ADT_CD326
tSNE2
tSNE1
High
Low
D
ADT
Mega 1b
Naïve
Mega 1a
Ery 1
Gra-Mo 2
Gra-Mo 1b
Ery 2
Mega 2
Gra-Mo1a
Mega 3
tSNE2
tSNE1
E
Mega
Ery
D10
D13
Granulo
Mono
tSNE2
tSNE1
F
RNA clusters
Naïve
Ery 1
Ery 2
Mega 1
Mega 2
Mega 3
Gr-Mo 1
Gr-Mo 2
ADT clusters
Naïve
Ery 1
Ery 2
Mega 1a
Mega 1b
Mega 2
Mega 3
Gr-Mo 1a
Gr-Mo 1b
Gr-Mo 2
0
100
% of cells
G
HSC/MPP
Erythroid
Megakaryo
Neut/Myelo
Gra-Mo 2
Gra-Mo1
Mega 3
Mega 2
Mega 1
Ery 2
Ery 1
Naïve
SPINK2
CD34
BPGM
HBA1
ALAS2
KLF1
GATA1
ESAM
PF4
PLAC8
LYZ
MPO
Level
High
Low
% of cells
25
50
75
100


A
F5
ANN
Training dataset → Human Fetal Liver
Test dataset → Human iPSCs derived
B
% of total HSC/MPP (per day)
100
75
50
25
0
D10
D13
C
% HSC/MPP (of total cells)
1.00
0.75
0.50
0.25
0.00
ND
CD235a−
CD235a− CD43+
D
HSC/MPP
in vitro
in vivo
in vitro
in vivo
UMAP1
UMAP2
E
EGR1
ZFP36L1
NR4A1
FOS
JUN
JUNB
in vitro
in vivo
Expression
High
Low
F
PATZ1 – GGGGGGGGGCC
NFKB1 – GGGACTTTCCA
RELA – GGGAATTTCC
SP1 – GGGGGCGGGG
NFKB1 – GGGAATTTCC
EGR1 – GCGGGGGCGGGG
CREB1 – TGACGTCA
REL – GGGATTTCC
NFKB1 – GGGGACTTTCC
CREB1 – GGTGACGT
MYB – CCAATGGCGGTTGGGGG
KLF5 – GGGCGGGGC
MZF1 – GAGGGGGAA
ZNF281 – GGGGGTGG
MZF1 – AGGGGGA
MZF1 – GGGGA
ZNF202 – GGGG
GCM1 – TGCGGG
0
10
20
30
Score
P.val.
0.009
0.006
0.003