# A Single Cell Resolution Map of Mouse Haematopoietic Stem and Progenitor Cell Differentiation

Running title: Single Cell Map of HSPC Differentiation

Sonia Nestorowa[1*], Fiona K. Hamey[1*], Blanca Pijuan Sala[1], Evangelia Diamanti[1], Mairi Shepherd[1], Elisa Laurenti[1], Nicola K. Wilson[1#], David G. Kent[1#], Berthold Göttgens[1#]

1: Department of Haematology and Wellcome Trust and MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge UK

*: Equal contribution

#: Corresponding authors

B. Gottgens; E-mail: bg200@cam.ac.uk, Tel. +44-1223-336829, FAX +44-1223-762670

D. Kent; E-mail: dgk23@cam.ac.uk, Tel. +44-1223-3362155, FAX +44-1223-762670

N. Wilson; E-mail: nkw22@cam.ac.uk, Tel. +44-1223-336822, FAX +44-1223-762670

Word count:

Abstract: 247

Main Text: 4162

No of Figures: 6

No of Tables: 0

No of references: 55

**Key Points** (<140 characters)

1) An expression map of HSPC differentiation from single cell RNA-Seq of 1,656 HSPCs provides new insights into blood stem cell differentiation

2) A user-friendly webresource provides access to single cell gene expression profiles for the wider research community

## Abstract

Maintenance of the blood system requires balanced cell fate decisions of hematopoietic stem and progenitor cells (HSPCs). Since cell fate choices are executed at the level of individual cells, new single cell profiling technologies offer exciting possibilities to map the dynamic molecular changes underlying HSPC differentiation. Here we have used single cell RNA-Seq to profile over 1,600 single HSPCs, where deep sequencing has enabled detection of an average of 6,558 protein-coding genes per cell. Index sorting, in combination with broad sorting gates, allowed us to retrospectively assign cells to 12 commonly sorted HSPC phenotypes while also capturing intermediate cells typically excluded by conventional gating. We further show that independently generated single cell datasets can be projected onto the single cell resolution expression map to directly compare data from multiple groups and to build and refine new hypotheses. Reconstruction of differentiation trajectories reveals dynamic expression changes associated with early lymphoid, erythroid-megakaryocytic and granulocyte-macrophage differentiation. The latter two trajectories were characterized by common upregulation of cell cycle and oxidative phosphorylation transcriptional programs. Using external spike-in controls, we estimate absolute mRNA levels per cell, showing for the first time that despite a general reduction in total mRNA, a subset of genes shows higher expression levels in immature stem cells consistent with active maintenance of the stem cell state. Finally, we report the development of an intuitive web interface as a new community resource, to permit visualization of gene expression in HSPCs at single cell resolution for any gene of choice.

## Introduction

Haematopoietic stem cells (HSCs) sit at the apex of a differentiation hierarchy that produces the full spectrum of mature blood cells via intermediate progenitor stages. For almost three decades, researchers have developed protocols for the prospective isolation of increasingly refined hematopoietic stem and progenitor cell (HSPC) populations, reaching purities of more than 50% for long-term repopulating HSCs[1–5]. While these approaches have provided many significant advances, none of the populations purified to date is comprised of a single homogeneous cell type, and the purification protocols necessitate the use of restrictive gates to maximise population purity, thus excluding potential "transitional" cells located outside of these gates.

It has long been recognised that a mechanistic understanding of differentiation processes requires detailed knowledge of the changes in gene expression that accompany and/or drive the progression from one cellular state to the next. Conventional bulk expression profiling of heterogeneous populations captures average expression states that may not be representative of any single cell. Recently developed single cell profiling techniques are able to resolve population heterogeneity[6,7], and profile "transitional" cells when scaled up to large cell numbers[8]. Full flow cytometry phenotypes can be recorded using index sorting[9] to link single cell gene expression profiles with single cell function[10]. Single cell profiling also enables reconstruction of regulatory network models[11–13] and inference of differentiation trajectories[8,14].

Web interfaces providing access to comprehensive transcriptomic resources have been instrumental in supporting research into the molecular mechanisms of normal and malignant haematopoiesis[15–20]. However, there is as yet no comparable resource or web interface for single HSPC transcriptome data. Here we present 1,656 single HSPC transcriptomes, analysed by scRNA-seq with broad gates, deep sequencing, and index sorting to retrospectively identify populations by surface marker expression. The resulting single-cell resolution gene expression landscape has been incorporated into a freely accessible online resource that can be utilized to visualize HSC to progenitor transitions, highlight putative lineage branching points, and identify lineage-specific transcriptional programs.

## Methods

### Single cell RNA-Seq

HSPCs were collected from the bone marrow of 10 female 12-week-old C57BL6 mice over two consecutive days, with cells from 4 mice pooled together and one mouse analysed separately each day. The bone marrow was lineage depleted using the EasySep™ Mouse Hematopoietic Progenitor Cell Enrichment Kit (StemCellTechnologies). The following antibodies were used: anti-EPCR-PE (Clone: RMEPCR1560, StemCellTechnologies, 60038PE), antiCD48-PB (Clone: HM481, Biolegend, 103418), antiLin-BV510 (StemCellTechnologies, 19856), antiCD150-PE/Cy7 (Clone: TC15012F12.2, Biolegend, 115914), antiCD16/32-Alexa647 (Clone: 93, Biolegend, 101314), antiCKit-APC/Cy7 (Clone: 2B8, Biolegend, 105856), antiFlk2-PE/Cy5 (Clone: A2F10, eBioscience, 115914), antiCD34-FITC (Clone: RAM34, BD Pharmingen, 553733) and DAPI. scRNA-seq analysis was performed as described previously[10,21]. Single cells were individually sorted by FACS into wells of a 96-well PCR plate containing lysis buffer. The Illumina Nextera XT DNA preparation kit was used to prepare libraries. Pooled libraries were sequenced using the Illumina HiSeq 2500 system and re-sequenced using the Illumina HiSeq 4000 system (single-end 125bp reads). Reads were aligned using G-SNAP[22] and the mapped reads were assigned to Ensembl genes (release 81)[23] by HTSeq[24].

To pass quality control, cells were required to have at least 200,000 reads mapping to nuclear genes, at least 4,000 genes detected, less than 10% of mapped reads mapping to mitochondrial genes and less than 50% of mapped reads mapping to the ERCC spike-ins (Life Technologies, 4456740) (supplementary Figure S1). Reads were normalised following the method of Lun et al.[25] using an initial clustering step to group cells with similar expression patterns. ERCC spike-ins were used to estimate the level of technical variance as described by Brennecke et al[26].Variable genes were defined as having a squared coefficient of variation exceeding technical noise, with 4773 genes passing this threshold (supplementary Figure S2B).

Raw data has been uploaded to NCBI GEO (accession number GSE81682). Data were normalised in R (https://www.r-project.org), using *flowCore* to extract and compensate the data and *ComBat* from the *sva* package to normalise the data. Thresholds for each population were assigned retrospectively based on published literature[27–30] and comparison with normalised index data with FlowJo (Treestar). E-SLAM cells were gated as EPCR$^+$CD48$^-$ CD150$^+$ as CD45 was not available in the index data. The gates were set in two ways: covering all cells (broad gating) or leaving unclassified cells in between populations to ensure that the gates did not contain any overlap (narrow gating).

**Computational analysis**

All computational analysis was performed in the R programming environment (https://www.r-project.org). Hierarchical clustering was performed using the *hclust* function, with distance (1 - Spearman's correlation)/2 and average linkage. Discrete clusters were identified using *cutreeDynamic* (*dynamicTreeCut* package), with the hybrid method and minimum cluster size = 10. The deepSplit parameter was set to 1, resulting in 4 broad clusters. For each cluster, gene expression was compared between cells in the cluster and the rest of the dataset. Genes expressed ($\log_2$ expression value > 4) in at least half of the cells in a cluster were tested for differential expression using a Wilcoxon rank sum test with

Benjamini-Hochberg correction. Genes with false discovery rate < 0.001 were ranked by fold change and the 10 genes with highest fold change for each cluster are displayed in Figure 1B.

Dimensionality reduction was performed on $\log_2$-transformed expression data for the 4773 variable genes using the diffusion map method[31] (*destiny* package[32]) with cosine distance and Gaussian kernel width = 0.16. Three-dimensional plots were produced using the *scatter3D* function from the *plot3D* package, and the *dm.predict* function was used to project external data. Due to high cell numbers, data of Kowalczyk et al.[33] were randomly sampled to obtain 50 cells from each condition (cell type, condition and strain) for clearer visualisation.

The three-dimensional diffusion map embedding was used to identify a start cell (within the E-SLAM population) and end cells for each of the 3 lineages (E, GM and L). Identifying broad branches between start and end cells was done by finding cells centred around shortest paths in the diffusion map, following the procedure of Ocone et al.[13]. To identify genes up/downregulated with trajectories, cells were ordered in pseudotime, and gene expression smoothed by calculating the mean for a sliding window of size 20. Spearman's correlation between smoothed pseudotime and expression values was calculated for each gene, genes with absolute correlation > 0.5 were identified, and clustered using hierarchical clustering with average linkage on Spearman's correlation.

Gene set enrichment analysis was performed in Enrichr[34]. Results with adjusted p-value <0.05 (using Benjamini-Hochberg correction for multiple testing) were considered significant. Full tables of results can be found in the supplementary material. Cell cycle genes were downloaded from Reactome (http://www.reactome.org/ (25/04/16)). Cell cycle category was inferred using a recently described method[35]. To estimate absolute gene expression, external ERCC spike-ins were used to normalise reads within each plate by calculating spike-in size factors using function *computeSpikeFactors* from the *scran* package, before normalising cells with these size factors. To account for batch effect differences in ERCC concentration between lanes (supplementary Figure S5) we applied *ComBat* from the *SVA* package, using the sorting gate (HSPC/Prog/LT-HSC) as an adjustment variable. Estimates of the total RNA content were calculated by summing absolute normalised counts per cell. Significance of differences in RNA content and FSC-H between cell types was calculated using a one-way ANOVA test. To identify genes downregulated in pseudotime in absolute terms the previously obtained downregulated lists (found using relative gene expression values) were filtered to remove any genes that did not have > 2-fold absolute expression change between the first 10% cells in a pseudotime trajectory and the final 10%.

## Results

### An atlas of single cell HSPC expression profiles

Single cell resolution RNA-Seq of embryonic stem and muscle progenitor cell differentiation has demonstrated that differentiation likely occurs as a near-continuous process, with gradual changes in gene expression as cells traverse the transcriptional landscape[14,36]. To comprehensively sample cells across the entire spectrum of the mouse HSPC transcriptional landscape, we isolated single cells using two broad sorting gates based on c-Kit and Sca1 expression, encompassing long-term HSCs (LT-HSCs), lymphoid multipotent progenitors (LMPPs) and multipotent progenitors (MPPs) in one gate, called the HSPC gate, and megakaryocyte-erythrocyte progenitors (MEPs), common myeloid progenitors (CMP), and granulocyte-monocyte progenitors (GMPs) in the second gate, called the Progenitor/Prog gate (Figure 1A). As LT-HSCs (Lin$^-$ c-Kit$^+$ Sca1$^+$ CD34$^-$ Flk2$^-$) are much less frequent than other populations in the HSPC gate, additional LT-HSCs were also sorted. Cells were retrospectively categorised into specific HSPC populations[27,28] using index-sorting data[10]. Each cell was also stained with three additional antibodies against CD150, CD48 and EPCR to retrospectively assign cells to other commonly used sorting schemes for populations such as E-SLAM (CD48$^-$ CD150$^+$ CD45$^+$ EPCR$^+$)[3] or MPP subpopulations[27,29].

Single cells were processed for RNA-Seq as described[21] with 156 HSCs, 701 HSPCs and 799 Progenitors passing stringent quality control parameters (see methods). Technical noise analysis[26] revealed 4,773 genes with expression variability exceeding technical noise. Unsupervised clustering partitioned the 1,656 cells into 4 major clusters (Figure 1B). Cluster 1 is mostly made up of LT-HSCs and is represented by genes such as *Procr (*EPCR*)* and *Trpc6*. Clusters 2 and 3 are both composed of all investigated cell types, and share expression of many of the representative genes, but are differentiated by higher expression of a number of genes including *Ccl9, Clec12a* and *Tyrobp* in Cluster 3. Cluster 4 is mainly composed of MEPs and is characterised by expression of genes such as *alpha hemoglobin* (*Hba-a1*) and *Smim1*. This analysis suggests that the transcriptomes of 1,656 single HSPCs presented here provide new opportunities to explore the transcriptional landscape of early HSC differentiation at single cell resolution.

**Visualising gene expression along the continuum of HSPC differentiation**

Diffusion maps have recently emerged as a dimensionality-reduction procedure particularly suited to displaying continuous differentiation processes from single cell snapshot data[11,31,37]. When applied to the 1,656 cells profiled here (Figure 2A), an intuitive graphical representation of the early process of HSPC differentiation emerges. The diffusion map can be coloured based on the previously identified clusters (Figure 2B), revealing that Clusters 1 (purple), 3 (gold) and 4 (pink) form separate branches of the diffusion map, and Cluster 2 (turquoise) encompasses cells between the three branches. Expression levels of individual genes can be plotted in the diffusion map to reveal their expression profiles across the HSPC transcriptional landscape (Figure 2C). *Gata1* expression is concentrated in Cluster 4, consistent with it being made up of mostly MEPs. *Procr* and *Mpl* expression is seen mainly in Cluster 1, made up of LT-HSCs. Of note, the recently reported LT-HSC markers *Hoxb5*, *Fgd5* and *Ctnaal1/alpha-catulin*[38–40] all showed predominant expression in Cluster 1.

Visualisation of surface marker expression from the normalised index data marked coherent territories within the diffusion map consistent with a robust separation of HSCs and more mature progenitors (Figure 2D). These results illustrate how the diffusion map representation of our dataset is a powerful way of interrogating the gene expression of any gene across the transcriptional landscape of HSPC differentiation. We therefore developed a user-friendly website (http://blood.stemcells.cam.ac.uk/single_cell_atlas.html) where users can explore the three-dimensional structure of the diffusion map graph as well as visualise expression profiles for any gene of interest, and surface marker expression. Of note, alternative dimensionality reduction methods such as principal component analysis showed similar relationships between the clusters (see supplementary Figure S4). This novel dataset and accompanying online resource permits interrogation of individual genes and surface markers at single cell resolution and can be broadly applied to a range of applications including full integration of other single cell datasets.

**The single cell transcriptional landscape illustrates the nature of HSPC populations and cellular phenotypes**

The relationships between different surface-marker-defined HSPC populations remain an area of active debate. Having used a uniform panel of nine surface markers for index sorting, cells were retrospectively assigned to 12 distinct HSPC phenotypes and displayed in the diffusion map (Figure 3A). With the exception of the CMP population which has been described as functionally heterogeneous[41], all other populations occupied defined territories. The original paper describing MEPs showed that GMPs are more common than MEPs[42]; however, they performed partial lineage depletion which differs from the conditions used in this study, thus influencing the ratios of GMPs/MEPs/CMPs isolated. Importantly, while lineage depletion can be variable, retrospective back-gating places the cells accurately. The three populations containing LT-HSCs overlapped as expected, with additional substantial overlaps between MPP3 and LMPP, and potential progressions such as a putative journey from E-SLAM via ST-HSC and LMPP to GMP.

The diffusion map protocol has recently been developed to permit projection of new data into the coordinates of an existing diffusion map[32], which allowed us to interrogate cellular phenotypes of other recently published single cell datasets. Projection of young and old HSCs in C57BL/6 and DBA/2 mouse strains[43] and Vwf-EGFP mice[33] showed that both young and old HSCs cluster together with LT-HSCs from our dataset, with old HSCs forming a tighter cluster suggestive of a more homogenous population. This analysis therefore not only demonstrates that our large expression atlas permits robust comparisons between single cell datasets generated in different labs, but also reveals a consistent phenotypic change of old HSCs in both studies, where old stem cells are more concentrated in what seems to be the core “HSC territory” of the diffusion map.

**Mapping differentiation trajectories from the single cell expression landscape**

Having established that single cells in the diffusion map are arranged in a pattern consistent with known lineage relationships, we next identified three differentiation trajectories (see methods) starting each time with E-SLAM HSCs and ending with erythroid (E), granulocyte macrophage (GM) and lymphoid (L) progenitors respectively (see Figure 4A). Based on gene expression profiles, each cell within a differentiation trajectory is given a pseudotime timestamp, and can therefore be arranged in a pseudotemporal ordering (see methods). Visualisation of surface marker expression from the index data revealed dynamic profiles consistent with known expression patterns, thus validating the pseudotemporal ordering (Figure 4B). This analysis also showed that the E trajectory traverses through a significant proportion of cells co-expressing CD150 and CD48, whereas the proportion of cells with that surface marker phenotype is much smaller for the GM and L trajectories.

We next identified genes showing statistically significant positive or negative correlation with the pseudotemporal ordering (Figure 4C). Gene set enrichment analysis (Figure 4D) showed enrichments consistent with the respective trajectories such as tetrapyrrole biosynthesis for E upregulated genes and neutrophil-mediated immunity for GM upregulated genes.  This analysis also revealed a major contribution of cell cycle associated genes to both the E and GM upregulated genes. The three differentiation trajectories mapped out here are therefore consistent with current knowledge of early haematopoiesis, suggesting that the pseudotime reconstruction will provide a powerful means to chart the dynamic processes that underlie early HSPC differentiation at single cell resolution.

**Single cell resolution analysis of cell cycle activation during HSPC differentiation**

Having identified cell cycle as the most highly enriched term for the genes upregulated along both the E and GM trajectories, we next took advantage of a recently reported predictor for allocating individual cells to G0/G1, S and G2/M cell cycle categories based on their single cell transcriptomes[35]. The distribution of single cells across these three cell cycle categories was in good agreement with the enrichment of cell cycle terms in the genes upregulated along the E and GM trajectories (Figure 5A,B). The analysis also demonstrated that large scale transitioning of cells to S and G2/M phase occurs after the divergence of the L trajectory from the E and GM trajectories, thus suggesting that transition to rapid cell cycling is secondary to transcriptional diversification.

Since terms associated with cell cycle had dominated the gene set enrichment analysis for the E and GM trajectories described in Figure 4, we next intersected the E and GM upregulated genes with a curated set of 405 cell cycle associated genes. The filtered E-only and GM-only gene sets showed strong enrichment for terms associated with their known biological functions, such as porphyrin biosynthesis for heme production (E-only) and defense response to other organisms (GM-only) respectively (Figure 5C). Of note, the cell cycle-filtered genes upregulated in both the E and GM trajectories showed strong enrichment for terms associated with mitochondrial ATP production, consistent with previous reports that HSCs primarily use glycolysis[44–46], but switch to mitochondrial oxidative phosphorylation to meet the rapidly increasing energy demands for differentiation[47].

We next investigated how hydrogen ion transmembrane transport gene and cell cycle gene expression changes through pseudotime (Figure 5D). In the GM trajectory, expression increases after cells enter the GM/E trajectory, with highest expression achieved once the cells enter the GM only trajectory. For the E trajectory, expression already increases before cells leave the GM/E/L trajectory and continues to increase as cells transition into the E trajectory. As expected from the gene set enrichment analysis (Figure 4D), there is no substantial increase of both hydrogen ion transmembrane transport and cell cycle genes along the L trajectory.

**Identification of genes downregulated in absolute terms during HSC differentiation**

The relative quiescence and low metabolic activity of HSCs might be reflected in low amounts of total mRNA per cell. However, conventional bulk microarray or RNA-Seq analysis is geared towards identifying relative expression differences only. Single cell profiling on the other hand can be used to estimate absolute differences in total mRNA content. To estimate total mRNA content per cell, we used external spike-in controls, sorted single cells from HSPC, Progenitor and LT-HSC gates into all twenty 96-well plates in a predetermined layout, and sequenced each plate on a single lane so that consistent differences between the amounts of reads between cell types would become detectable (Figure 6A). Estimation of absolute mRNA content per cell revealed a gradual increase in average mRNA content from E-SLAM HSCs to LMPPs to GMPs to MEPs (Figure 6B-C) (cells assigned to populations based on index sorting data; see Figure 3). Of note, forward scatter is recognised as a correlate to cell size, and showed a similar, but not identical, pattern (Figure 6D), thus suggesting that mRNA content per cell is related, but not completely coupled, to cell size during early HSC differentiation.

We next used the spike-in based normalisation to investigate whether genes identified as downregulated in Figure 4 were indeed downregulated in real terms, e.g. fewer mRNA molecules per single cell. Importantly, conventional analysis would not have been able to distinguish this absolute downregulation from relative downregulation. In a situation where there is an increase of total amount of RNA per cell, as our spike-in based analysis shows for HSC differentiation, a given gene might appear to be downregulated in the relative expression analysis while it actually stays the same in absolute terms while a large fraction of the transcriptome is upregulated. However, the majority of downregulated genes from Figure 4 were downregulated in absolute terms along the E and GM trajectories (109/112 for E and 55/56 for GM), thus highlighting a subset of genes actively expressed in HSCs despite their quiescent and metabolically less-active state (see supplementary table). Gene set enrichment analysis showed enrichment for terms associated with megakaryocytes, although on closer inspection this corresponded to genes such as *Mpl* and *Procr*, known to be highly expressed in HSCs. Only 18 genes were specifically downregulated in the GM trajectory, thus precluding the identification of any statistically significant gene set overlaps. Terms enriched with the E downregulated genes corresponded to genes associated with the immune response. Taken together, these data demonstrate that single cell analysis allows estimation of total mRNA amounts per cell in the various HSPC compartments, thus allowing identification of genes that are, in real terms, more highly expressed in HSCs than the various downstream progenitors such as GMP and MEP.

## Discussion

Here we have taken advantage of recent advances in molecular profiling technologies to provide a single cell resolution expression atlas of early blood stem cell differentiation, which (i) overcomes several shortcomings of population-based bulk expression profiling, (ii) provides new insights into the diversification of transcriptional programs during HSC differentiation, and (iii) represents a powerful new resource for the haematopoiesis research community facilitated through the development of a new user-friendly website.

Previous bulk transcriptome analyses have made several important contributions to enhancing our understanding of HSPCs including the identification of new candidate regulators[48] and complex patterns of co-ordinately expressed gene sets[16]. Comprehensive single cell transcriptome data provide opportunities not readily available with conventional population-average data. For example, absolute differences in mRNA levels can be estimated for cells belonging to distinct differentiation stages. The quiescent nature[27] and low metabolic activity[46,47] of HSCs might have been taken to imply that the HSC state is characterised by a general low level of transcription, in line also with the well-documented low activity of Myc in HSCs[49–51]. Our data confirm this hypothesis in some respect by demonstrating that HSCs consistently contain less mRNA per cell than E and GM cells. Nevertheless, there exists a subset of genes with higher expression in absolute terms in HSCs, suggesting that some genes might contribute to actively maintaining the stem cell state.

The ability to project external single cell transcriptional data onto our single cell transcriptome atlas offers an attractive method of hypothesis generation. We projected data from two different laboratories and two different mouse strains[33,43], which all gave similar results, thus underscoring the robustness of this approach. When compared with HSCs from young mice, HSCs from old mice were more confined to the HSC territory of the diffusion map, suggesting that HSCs from old mice represent a more molecularly homogeneous population, with fewer cells already engaged in a differentiation trajectory. Of note, this observation was not reported in the two original publications, presumably because they lacked the extensive landscape of single HSPCs transcriptional states as a comparator. Interestingly however, conventional expression profiling of HSCs from old mice when coupled with epigenetic analysis had already suggested that in old HSCs the transcriptomic and epigenetic landscape promotes HSC self-renewal at the expense of differentiation[52]. Future exploitations of the single cell atlas as a comparator are likely to include the analysis of single cell transcriptomes from mouse models, including inducible mouse models of leukaemia.

When gene expression states are measured using thousands of genes, progression of a cell through a differentiation program can be thought of as a journey through a transcriptional landscape. This study captures 1,656 single cell gene expression snapshots of the HSPC transcriptional landscape, which provides several important insights. For example, dimensionality reduction methods such as diffusion maps represent a useful way to visualise and interpret datasets of over 8 million data points (e.g. 1,656 cells x 4,773 heterogeneously expressed genes). This is supported by the observation that previously defined HSPC populations form coherent groupings on the diffusion map with one major exception (CMPs), which have recently been described as highly heterogeneous[41,53].

Furthermore, while the arrangement of cells in the diffusion map is consistent with known developmental progressions (e.g. LT-HSC to ST-HSC to LMPP to GMP), there is substantial intermingling within transition zones. Some cells sorted for example as LMPPs will therefore be virtually identical at the transcriptome level to cells sorted as ST-HSCs. Moreover, for other transitions such as LMPP to GMP, conventional gating fails to capture a substantial number of cells in the transition zone. Of note, molecular characterisation of such “transition cells” may be particularly important to advance our understanding cellular differentiation.

A number of methods have been developed to reconstruct differentiation trajectories from single cell expression data[8,14]. Given the likely plasticity of immature cells, we opted for developing broad trajectories where a given cell at any moment in time would have the option of making sideway movements rather than just finding the shortest path between the two endpoints. It is remarkable therefore that even with these relatively broad trajectories, the three journeys reconstructed here already diverge within the part of the diffusion map occupied mostly by the ST-HSC population. While this observation is at odds with the more traditional view of the haematopoietic lineage tree[54], it is consistent with recent analysis of both mouse and human cell fate diversification[41,53,55]. Importantly, we now provide for the first time a reconstruction of the likely dynamics of expression changes during these early stages of HSPC fate diversification.

An important consideration with single cell RNA-Seq is to strike a balance between the number of cells profiled and the sequencing depth achieved for each cell. We opted for substantial sequencing depth detecting on average 6,558 protein-coding genes per cell. Emerging droplet sequencing technology facilitates increased throughput[36], but current methods do not afford ways of recording surface marker expression analogous to the index sorting employed here. Moreover, studies published so far have opted for much lower sequencing depth to keep overall costs manageable. This however makes it impossible to develop an online resource such as the one reported here, which can be used to display the expression profile for any gene of interest. Substantial sequencing depth is also required if single cell data are to be exploited for the discovery of molecular mechanisms that may drive cellular differentiation and diversification. The dataset and analysis reported here should be well placed to serve this function for the wider haematopoiesis research community.

## Acknowledgments

We thank Reiner Schulte, Chiara Cossetti, and Michal Maj at the CIMR Flow Cytometry Core for their help with cell sorting, Dean Pask and Tina Hamilton for technical assistance, and Wajid Jawaid, Paul Sumption and Chee Lim for help with setting up the website. Work in the author's laboratory is supported by grants from Bloodwise, Cancer Research UK, Biotechnology and Biological Sciences Research Council, Leukemia Lymphoma Society, the National Institute for Health Research Cambridge Biomedical Research Centre and core support grants by the Wellcome Trust to the Cambridge Institute for Medical Research and Wellcome Trust-MRC Cambridge Stem Cell Institute. S.N. and F.K.H. are recipients of Medical Research Council PhD Studentships. DGK is the recipient of a Bennett Fellowship from Bloodwise, and EL the recipient of a Sir Henry Dale Fellowship from the Wellcome Trust.

## Authorship Contributions

SN, MS, NKW and DGK performed experiments, FKH analysed single cell sequencing data, FKH and BPS analysed index data, ED mapped sequencing data, EL, NKW, DGK and BG designed and supervised the study, SN, FKH, EL, NKW, DGK and BG wrote the paper.

## Conflict of Interest Disclosures

The authors confirm that there are no conflicts of interest to declare.

## References

1. Beerman I, Bhattacharya D, Zandi S, et al. Functionally distinct hematopoietic stem cells modulate hematopoietic lineage potential during aging by a mechanism of clonal expansion. *Proc Natl Acad Sci U S A*. 2010;107(12):5465-5470.
2. Challen GA, Boles NC, Chambers SM, Goodell MA. Distinct hematopoietic stem cell

subtypes are differentially regulated by TGF-beta1. *Cell Stem Cell*. 2010;6(3):265-278.

3. Kent DG, Copley MR, Benz C, et al. Prospective isolation and molecular characterization of hematopoietic stem cells with durable self-renewal potential. *Blood*. 2009;113(25):6342-6350.

4. Kiel MJ, Yilmaz OH, Iwashita T, Yilmaz OH, Terhorst C, Morrison SJ. SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell*. 2005;121(7):1109-1121.

5. Morita Y, Ema H, Nakauchi H. Heterogeneity and hierarchy within the most primitive hematopoietic stem cell compartment. *J Exp Med*. 2010;207(6):1173-1182.

6. Mahata B, Zhang X, Kolodziejczyk AA, et al. Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. *Cell Rep*. 2014;7(4):1130-1142.

7. Jaitin DA, Kenigsberg E, Keren-Shaul H, et al. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science (80- )*. 2014;343(February):776-779.

8. Bendall SC, Davis KL, Amir ED, et al. Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. *Cell*. 2014;157(3):714-725. doi:10.1016/j.cell.2014.04.005.

9. Osborne GW. Recent advances in flow cytometric cell sorting. *Methods Cell Biol*. 2011;102:533-556. doi:10.1016/B978-0-12-374912-3.00021-3.

10. Wilson NK, Kent DG, Buettner F, et al. Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. *Cell Stem Cell*. 2015;16(6):712-724. doi:10.1016/j.stem.2015.04.004.

11. Moignard V, Woodhouse S, Haghverdi L, et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat Biotechnol*. 2015;33(3):269-276. doi:10.1038/nbt.3154.

12. Schütte J, Wang H, Antoniou S, et al. An experimentally validated network of nine

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

13. Ocone A, Haghverdi L, Mueller NS, Theis FJ. Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics*. 2015;31(12):i89-i96.

14. Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*. 2014;32(4):381-386.

15. Bagger FO, Sasivarevic D, Sohi SH, et al. BloodSpot: a database of gene expression profiles and transcriptional programs for healthy and malignant haematopoiesis. *Nucleic Acids Res*. October 2015:gkv1101.

16. Novershtern N, Subramanian A, Lawton LN, et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*. 2011;144(2):296-309.

17. Seita J, Sahoo D, Rossi DJ, et al. Gene Expression Commons: an open platform for absolute gene expression profiling. *PLoS One*. 2012;7(7):e40321.

18. Watkins NA, Gusnanto A, de Bono B, et al. A HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood*. 2009;113(19):e1-e9.

19. Chambers SM, Shaw CA, Gatza C, Fisk CJ, Donehower LA, Goodell MA. Aging hematopoietic stem cells decline in function and exhibit epigenetic dysregulation. *PLoS Biol*. 2007;5(8):e201.

20. Hebestreit K, Gröttrup S, Emden D, et al. Leukemia gene atlas--a public platform for integrative exploration of genome-wide molecular data. *PLoS One*. 2012;7(6):e39148.

21. Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc*. 2014;9(1):171-181. doi:10.1038/nprot.2014.006.

22. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in

short reads. *Bioinformatics*. 2010;26(7):873-881.

23. Flicek P, Amode MR, Barrell D, et al. Ensembl 2014. *Nucleic Acids Res*. 2014;42(Database issue):D749-D755.

24. Anders S, Pyl PT, Huber W. HTSeq - A Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2014;31(2):166-169.

25. L. Lun AT, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol*. 2016;17(1):75.

26. Brennecke P, Anders S, Kim JK, et al. Accounting for technical noise in single-cell RNA-seq experiments. 2013;10(11). doi:10.1038/nmeth.2645.

27. Wilson A, Laurenti E, Oser G, et al. Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell*. 2008;135(6):1118-1129. doi:10.1016/j.cell.2008.10.048.

28. Pronk CJH, Rossi DJ, Månsson R, et al. Elucidation of the phenotypic, functional, and molecular topography of a myeloerythroid progenitor cell hierarchy. *Cell Stem Cell*. 2007;1(4):428-442.

29. Pietras EM, Reynaud D, Kang Y-A, et al. Functionally Distinct Subsets of Lineage-Biased Multipotent Progenitors Control Blood Production in Normal and Regenerative Conditions. *Cell Stem Cell*. 2015;17(1):35-46.

30. Cabezas-Wallscheid N, Klimmeck D, Hansson J, et al. Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. *Cell Stem Cell*. 2014;15(4):507-522.

31. Haghverdi L, Buettner F, Theis FJ. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*. 2015:1-10.

32. Angerer P, Haghverdi L, Büttner M, Theis FJ, Marr C, Buettner F. destiny - diffusion maps for large-scale single-cell data in R. *Bioinformatics*. 2015;32(December 2015):btv715 - . doi:10.1093/bioinformatics/btv715.

33. Kowalczyk MS, Tirosh I, Heckl D, et al. Single cell RNA-seq reveals changes in cell

cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res*. 2015:gr.192237.115. doi:10.1101/gr.192237.115.

34. Chen EY, Tan CM, Kou Y, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. 2013;14:128.

35. Scialdone A, Natarajan KN, Saraiva LR, et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*. 2015;85:54-61. doi:10.1016/j.ymeth.2015.06.021.

36. Klein AM, Mazutis L, Akartuna I, et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*. 2015;161(5):1187-1201.

37. Coifman RR, Lafon S, Lee  a B, et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc Natl Acad Sci U S A*. 2005;102(21):7426-7431. doi:10.1073/pnas.0500896102.

38. Chen JY, Miyanishi M, Wang SK, et al. Hoxb5 marks long-term haematopoietic stem cells and reveals a homogenous perivascular niche. *Nature*. 2016;530(7589):223-227.

39. Acar M, Kocherlakota KS, Murphy MM, et al. Deep imaging of bone marrow shows non-dividing stem cells are mainly perisinusoidal. *Nature*. 2015;526(7571):126-130.

40. Gazit R, Mandal PK, Ebina W, et al. Fgd5 identifies hematopoietic stem cells in the murine bone marrow. *J Exp Med*. 2014;211(7):1315-1331.

41. Paul F, Arkin Y, Giladi A, et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*. 2015;163(7):1663-1677. doi:10.1016/j.cell.2015.11.013.

42. Akashi K, Traver D, Miyamoto T, Weissman IL. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature*. 2000;404(6774):193-197. doi:10.1038/35004599.

43. Grover A, Sanjuan-Pla A, Thongjuea S, et al. Single-cell RNA sequencing reveals molecular and functional platelet bias of aged haematopoietic stem cells. *Nat Commun*. 2016;7:11075.

44. Takubo K, Nagamatsu G, Kobayashi CI, et al. Regulation of glycolysis by Pdk functions as a metabolic checkpoint for cell cycle quiescence in hematopoietic stem cells. *Cell Stem Cell*. 2013;12(1):49-61.

45. Simsek T, Kocabas F, Zheng J, et al. The distinct metabolic profile of hematopoietic stem cells reflects their location in a hypoxic niche. *Cell Stem Cell*. 2010;7(3):380-390.

46. Suda T, Takubo K, Semenza GL. Metabolic regulation of hematopoietic stem cells in the hypoxic niche. *Cell Stem Cell*. 2011;9(4):298-310.

47. Yu W-M, Liu X, Shen J, et al. Metabolic regulation by the mitochondrial phosphatase PTPMT1 is required for hematopoietic stem cell differentiation. *Cell Stem Cell*. 2013;12(1):62-74.

48. Chambers SM, Boles NC, Lin K-YK, et al. Hematopoietic fingerprints: an expression database of stem cells and their progeny. *Cell Stem Cell*. 2007;1(5):578-591.

49. Wilson A, Murphy MJ, Oskarsson T, et al. c-Myc controls the balance between hematopoietic stem cell self-renewal and differentiation. *Genes Dev*. 2004;18(22):2747-2763.

50. Guo Y, Niu C, Breslin P, et al. c-Myc-mediated control of cell fate in megakaryocyte-erythrocyte progenitors. *Blood*. 2009;114(10):2097-2106.

51. Laurenti E, Varnum-Finney B, Wilson A, et al. Hematopoietic stem cell function and survival depend on c-Myc and N-Myc activity. *Cell Stem Cell*. 2008;3(6):611-624.

52. Sun D, Luo M, Jeong M, et al. Epigenomic profiling of young and aged HSCs reveals concerted changes during aging that reinforce self-renewal. *Cell Stem Cell*. 2014;14(5):673-688.

53. Perie L, Duffy KR, Kok L, Boer RJ De, Schumacher TN. The Branching Point in Erythro-Myeloid Differentiation. *Cell*. 2015;163:1655-1662. doi:10.1016/j.cell.2015.11.059.

54. Kondo M, Wagers AJ, Manz MG, et al. Biology of hematopoietic stem cells and progenitors: implications for clinical application. *Annu Rev Immunol*. 2003;21:759-

806.

55. Notta F, Zandi S, Takayama N, et al. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science*. 2015;351(6269):aab2116. doi:10.1126/science.aab2116.

## Figure Legends

**Figure 1. Generating linked transcriptional and surface marker profiles for over 1,600 single HSPCs.**

(A) Schematic of the sorting strategy used paired with index sorting data. Bone marrow cells were stained with 9 antibodies against various cell surface markers in order to isolate HSPCs (Lin$^-$ c-Kit$^+$ Sca1$^+$) and Progenitors (Lin$^-$ c-Kit$^+$ Sca1$^-$). Almost all cells in the Flk2-CD34 gate and the CD16/32-Flk2 gate were collected for HSPCs and Progenitors, respectively, within broad, all-encompassing gates. In addition, LT-HSCs were (Lin$^-$ c-Kit$^+$ Sca1$^+$ CD34$^-$ Flk2$^-$) collected separately to ensure adequate numbers were collected. Each cell population retrospectively identified is shown in the table, colours and names remain consistent throughout the text. Letters indicate populations in the flow cytometry diagrams. (B) Unsupervised hierarchical clustering of gene expression data for all cells. Clustering was performed using all 4,773 variable genes except Ly6a/Sca-1 to avoid bias in clustering. The cells split into four major clusters (Cluster 1 - purple; Cluster 2 - turquoise; Cluster 3 - gold; Cluster 4 - pink). The top 10 genes enriched in each cluster are displayed in the heatmap, showing gene expression on a log$_2$ scale from blue to orange (low to high). The clusters were also compared by cell type composition, following both broad and narrow gating strategies. Broad gating involved the classification of all cells into a cell type category, whereas narrow gating included only cells that are more likely to fit the predefined HSPC classification, gated around the greatest density of cells within the population gating strategy. Cell type is coloured based on the scheme used in Figure 1A. Grey cells in the narrow gating strategy represent cells unassigned to any population.

**Figure 2. Multidimensional analysis can be used to visualise gene expression across HSPC differentiation.**

(A) Schematic explaining how diffusion maps are used as a dimensionality-reduction procedure. (B) Diffusion map of all cells coloured based on previously defined clusters (Cluster 1 - purple; Cluster 2 - turquoise; Cluster 3 - gold; Cluster 4 - pink). Diffusion components 1, 2 and 3 are shown. (C) Diffusion map of all cells coloured according to the expression of selected genes. The genes were chosen based on published literature or were identified computationally as highly expressed in specific cell populations. The colour corresponds to a $\log_2$ scale of expression ranging between 0 and the maximum value for each gene. (D) Diffusion map of all cells coloured by surface marker expression from the normalised index data. The majority of these markers were used for cell selection, with the exception of CD48, CD150 and EPCR. The colour corresponds to a linear scale of expression ranging between the minimum and maximum value for each marker.

**Figure 3. The single cell HSPC transcriptional landscape can be used to visualise HSPC populations and their relationships.**

(A) Diffusion map of all cells coloured based on cell population using narrow gating. All populations were identified retrospectively using the index sorting data. Populations are identified using normalised index data. The cells of interest for each population are coloured purple and enlarged for easier visibility. (B) Diffusion map of all cells with projection of data from recently published datasets. Data collected by Kowalczyk et al. (C57BL/6, DBA/2) and Grover et al. (Vwf-EGFP) is displayed. Both groups collected HSCs from mice 2-3 months (orange) and 20-25 months (blue) old. HSCs were defined as Lin$^-$ c-Kit$^+$ Sca1$^+$ CD150$^+$ CD48$^-$.

**Figure 4. Pseudotime analysis reveals trends in surface marker and gene expression for differentiation trajectories.**

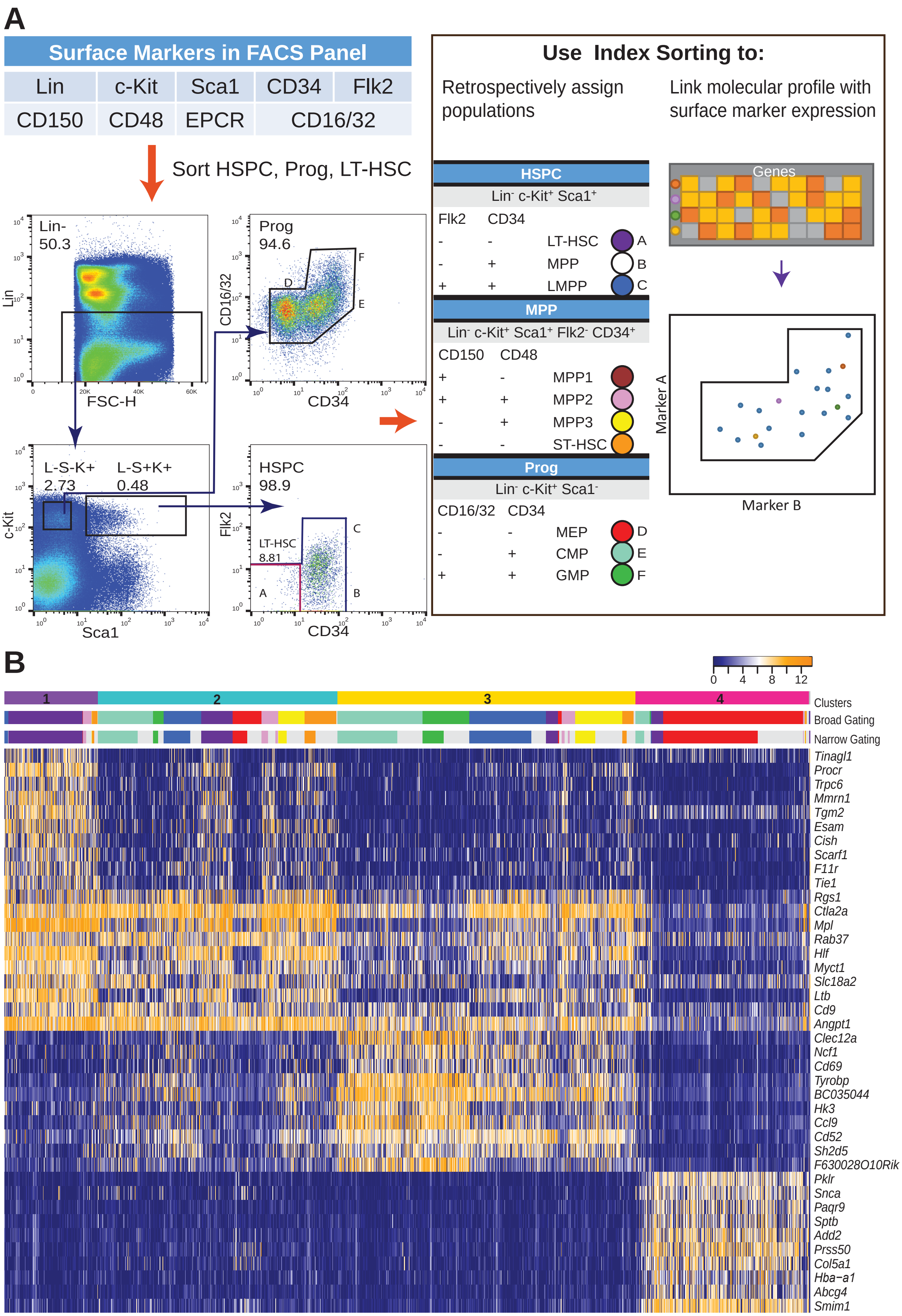
(A) Diffusion map coloured by pseudotime trajectories to erythroid (E), granulocyte macrophage (GM) and lymphoid (L) fates. Each trajectory starts from a HSC (blue) and ends with a progenitor (red). (B) Changes in surface marker expression and FSC-H through pseudotime for each of the three trajectories, obtained from the normalised index data. For each trajectory it is possible to see what cell types are passed through to reach the final cell fate. (C) Normalised expression of genes positively (up) or negatively (down) correlated with the pseudotemporal ordering for each trajectory. Mean normalised expression is plotted with standard deviation. (D) Most significant relevant terms from gene set enrichment analysis for all the trajectories, performed in Enrichr. Terms with an adjusted p-value <0.05 (using Benjamini-Hochberg correction for multiple testing) were considered significant. The full tables of results can be found in the supplemental data.

**Figure 5. Analysis of cell cycle activation during HSPC differentiation at single-cell resolution.**

(A) Diffusion map of all cells coloured by computationally-assigned cell cycle category. There is no assignment for $G_0$ separately due to limitations of the method. (B) Proportion of E-SLAMs, LMPPs, GMPs and MEPs in each of the cell cycle categories. The cell types displayed are based on the narrow gating strategy. (C) Gene set enrichment analysis was performed for the three trajectories after the removal of cell cycle genes. The most relevant significant terms for genes positively correlated with pseudotime analysis are shown. Terms with an adjusted p-value <0.05 (using Benjamini-Hochberg correction for multiple testing) were considered significant. The full tables of results can be found in the supplemental data. (D) Average expression of hydrogen ion transmembrane transport genes and cell cycle genes across pseudotime. Each gene was normalised across the median of all 3 trajectories for plotting. The average expression is coloured by trajectory and means are shown with standard deviation.

**Figure 6. Single cell analysis can be used to estimate absolute differences in total mRNA content across cell types.**

(A) Schematic explanation of how plate composition and ERCC spike-ins are used to estimate absolute RNA levels. The plate organisation for this study included cells from multiple sorting gates (HSPC, Prog, LT-HSCs) and each well contained ERCC spike-ins. The sequencing depth varies across lanes and cell types, therefore ERCC spike-ins are used to normalise across cell types within a lane, in which the spike-in content becomes level within a lane but cell mRNA content may still vary. After this step, RNA content can be normalised across lanes. (B) Diffusion map of all cells coloured by RNA content. Estimates of total RNA content were calculated by summing the absolute normalised counts per cell. The scale ranges from blue to green to yellow to red with increasing RNA content. (C) Sum of normalised counts for E-SLAMs, LMPPs, GMPs and MEPs, coloured by the scheme used in Figure 1A. Significance in differences in RNA content between cell types was calculated using a one-way ANOVA test (* p<0.01, ** p<0.001, ***p<0.0001) (D) FSC-H for E-SLAMs, LMPPs, GMPs and MEPs, coloured by the scheme used in Figure 1A. FSC-H is used as an indicator of cell size. Significance in differences in FSC-H between cell types was calculated using a one-way ANOVA test (* p<0.01, ** p<0.001, ***p<0.0001) (E) Most relevant significant terms from gene enrichment expression analysis on genes downregulated in absolute terms in E, GM and E&GM trajectories. The numbers of genes showing downregulation along pseudotime in absolute terms is displayed in the Venn diagram. Terms with an adjusted p-value <0.05 (using Benjamini-Hochberg correction for multiple testing) were considered significant. The full tables of results can be found in the supplemental data.


A
Surface Markers in FACS Panel
Lin
c-Kit
Sca1
CD34
Flk2
CD150
CD48
EPCR
CD16/32
Sort HSPC, Prog, LT-HSC
Lin-
50.3
Lin
FSC-H
Prog
94.6
D
F
E
CD16/32
CD34
L-S-K+
2.73
L-S+K+
0.48
c-Kit
Sca1
HSPC
98.9
LT-HSC
8.81
A
B
C
Flk2
CD34
Use Index Sorting to:
Retrospectively assign populations
Link molecular profile with surface marker expression
HSPC
Lin- c-Kit+ Sca1+
Flk2 CD34
- - LT-HSC A
- + MPP B
+ + LMPP C
MPP
Lin- c-Kit+ Sca1+ Flk2- CD34+
CD150 CD48
+ - MPP1
+ + MPP2
- + MPP3
- - ST-HSC
Prog
Lin- c-Kit+ Sca1-
CD16/32 CD34
- - MEP D
- + CMP E
+ + GMP F
Genes
Marker A
Marker B
B
0
4
8
12
1
2
3
4
Clusters
Broad Gating
Narrow Gating
Tinagl1
Procr
Trpc6
Mmrn1
Tgm2
Esam
Cish
Scarf1
F11r
Tie1
Rgs1
Ctla2a
Mpl
Rab37
Hlf
Myct1
Slc18a2
Ltb
Cd9
Angpt1
Clec12a
Ncf1
Cd69
Tyrobp
BC035044
Hk3
Ccl9
Cd52
Sh2d5
F630028O10Rik
Pklr
Snca
Paqr9
Sptb
Add2
Prss50
Col5a1
Hba−a1
Abcg4
Smim1

# Figure 1

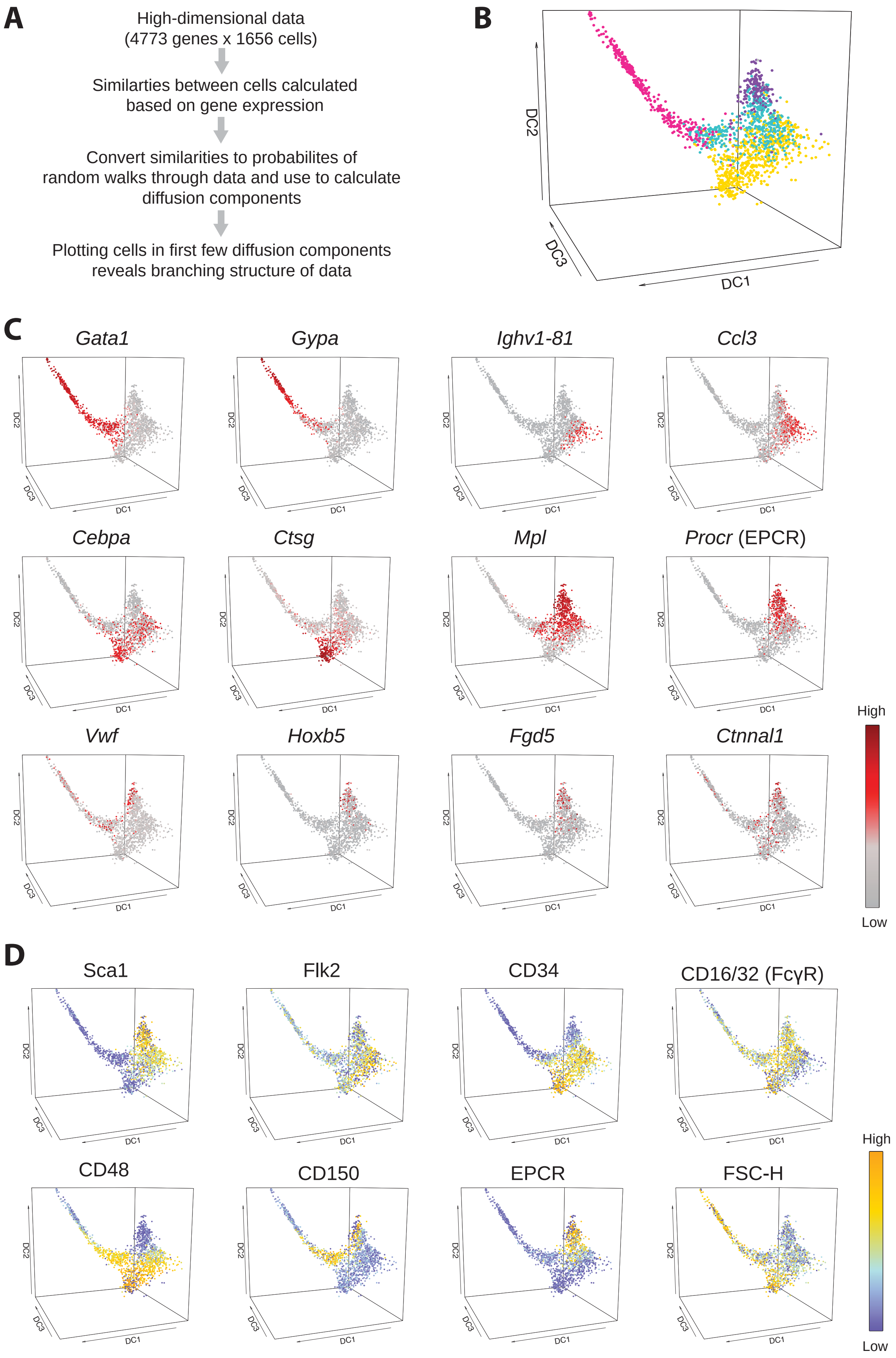
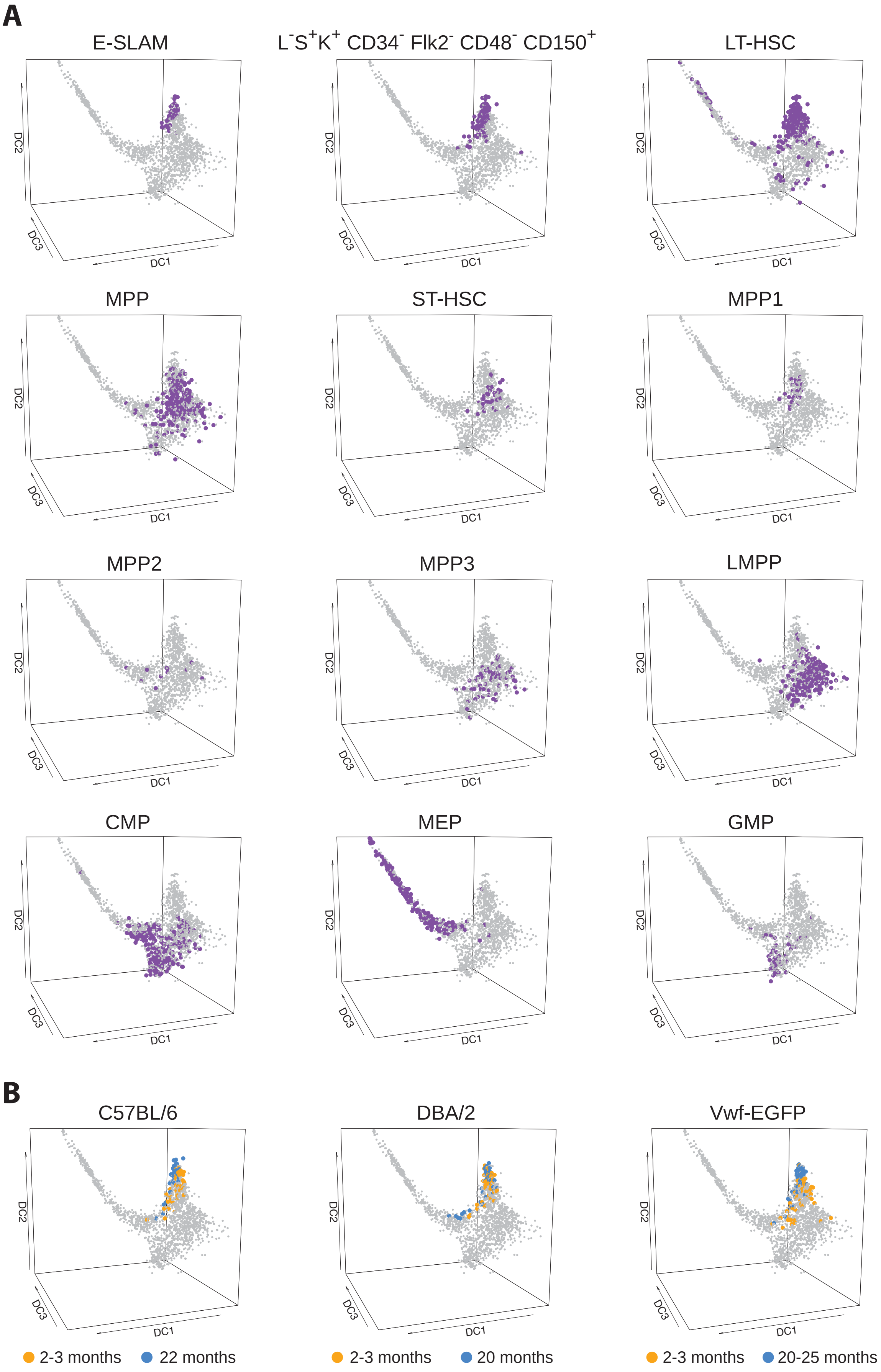
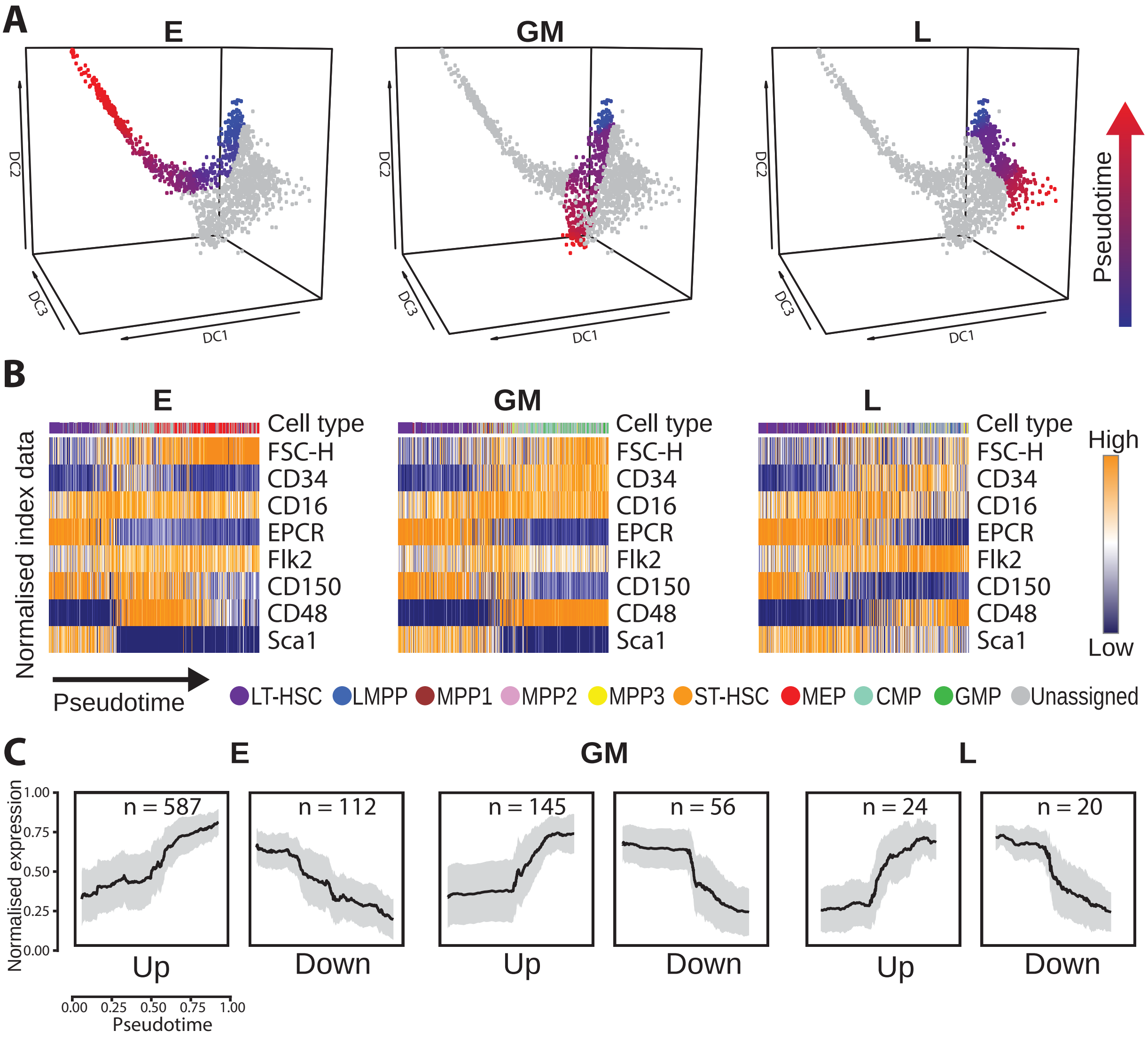

A
High-dimensional data
(4773 genes x 1656 cells)
Similarties between cells calculated
based on gene expression
Convert similarities to probabilites of
random walks through data and use to calculate
diffusion components
Plotting cells in first few diffusion components
reveals branching structure of data
B
DC2
DC3
DC1
C
Gata1
Gypa
Ighv1-81
Ccl3
Cebpa
Ctsg
Mpl
Procr (EPCR)
Vwf
Hoxb5
Fgd5
Ctnnal1
High
Low
D
Sca1
Flk2
CD34
CD16/32 (FcγR)
CD48
CD150
EPCR
FSC-H
High
Low

**Figure 2**


A
E-SLAM
$L^-S^+K^+$ CD34$^-$ Flk2$^-$ CD48$^-$ CD150$^+$
LT-HSC
MPP
ST-HSC
MPP1
MPP2
MPP3
LMPP
CMP
MEP
GMP
B
C57BL/6
DBA/2
Vwf-EGFP
DC1
DC2
DC3
● 2-3 months ● 22 months
● 2-3 months ● 20 months
● 2-3 months ● 20-25 months

# Figure 3



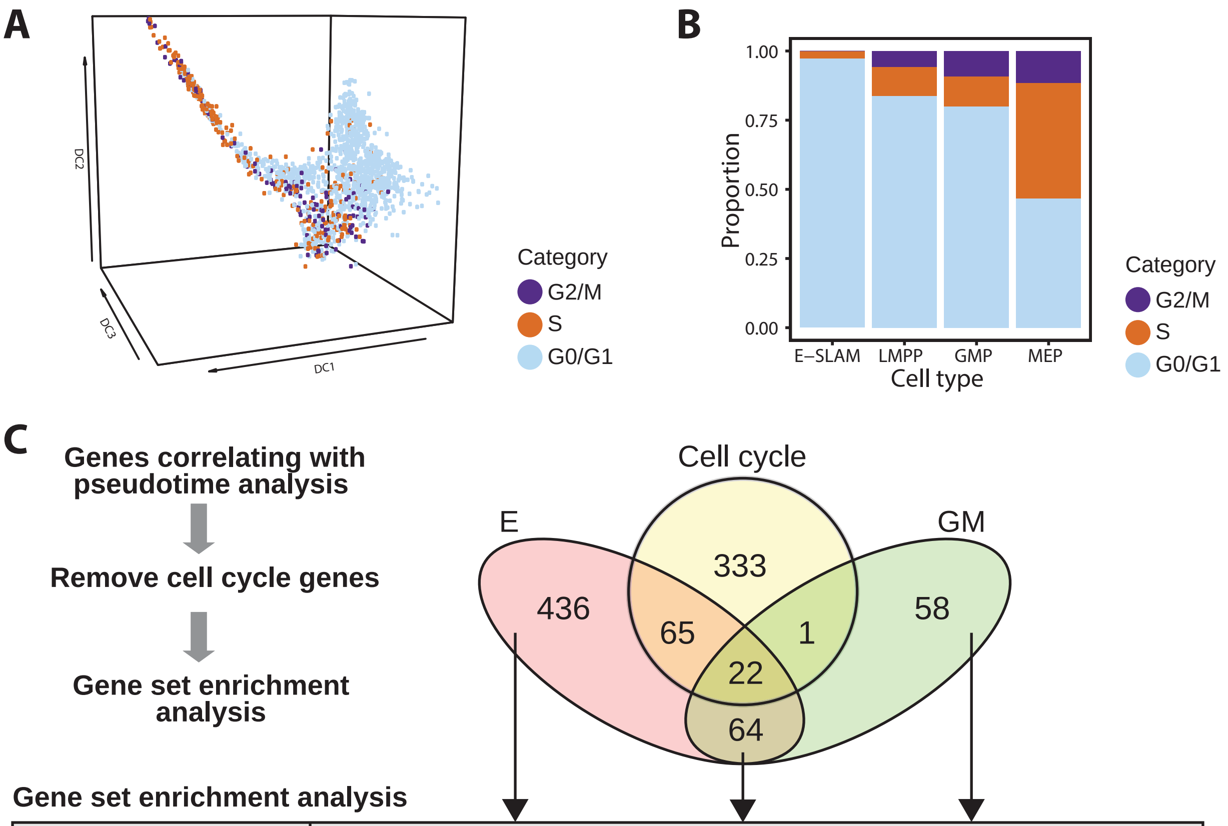
## D

**Gene set enrichment analysis**

| Category | E up | E down | GM up | GM down | L up | L down |
|---|---|---|---|---|---|---|
| **Biological processes** | Mitotic cell cycle<br>0<br>Tetrapyrrole biosynthetic process<br>$8.0 \times 10^{-7}$ | No significant terms | Mitotic cell cycle<br>$1.01 \times 10^{-9}$<br>Neutrophil mediated immunity<br>0.012 | Coagulation<br>0.017 | Regulation of lymphocyte differentiation<br>0.063 | Haemostasis<br>0.0091 |
| **Molecular function** | ATP binding<br>$3.0 \times 10^{-11}$ | Leukocyte activation<br>0.0012 | Cytochrome-c oxidase activity<br>0.010 | Guanyl nucleotide binding<br>0.035 | No significant terms | No significant terms |
| **MGI mammalian phenotype** | Abnormal haematopoietic system<br>$9.0 \times 10^{-8}$ | Abnormal immune system<br>$1.8 \times 10^{-7}$ | Abnormal immune system<br>$7.2 \times 10^{-7}$ | Abnormal haematopoietic system<br>0.0028 | Abnormal immune system<br>0.021 | Abnormal homeostasis<br>0.0082 |
| **Reactome (Pathways)** | Mitotic cell cycle<br>0 | Haemostasis<br>$2.4 \times 10^{-6}$ | Mitotic cell cycle<br>$4.2 \times 10^{-8}$ | Cytokine signaling in immune system<br>0.012 | No significant terms | Haemostasis<br>0.00081 |
| **Cell types (Mouse gene atlas)** | Megakaryocyte erythrocyte progenitor<br>$1.7 \times 10^{-21}$ | Mast cells<br>0.0028 | Granulocyte monocyte progenitor<br>$1.3 \times 10^{-6}$ | Stem cells HSC<br>0.0032 | Thymocyte DP CD4+CD8+<br>0.049 | Mast cells<br>$3.5 \times 10^{-6}$ |

**Terms shown along with adjusted *P* values (Benjamini-Hochberg method for correction for multiple hypotheses testing)**

# Figure 4



Genes correlating with pseudotime analysis

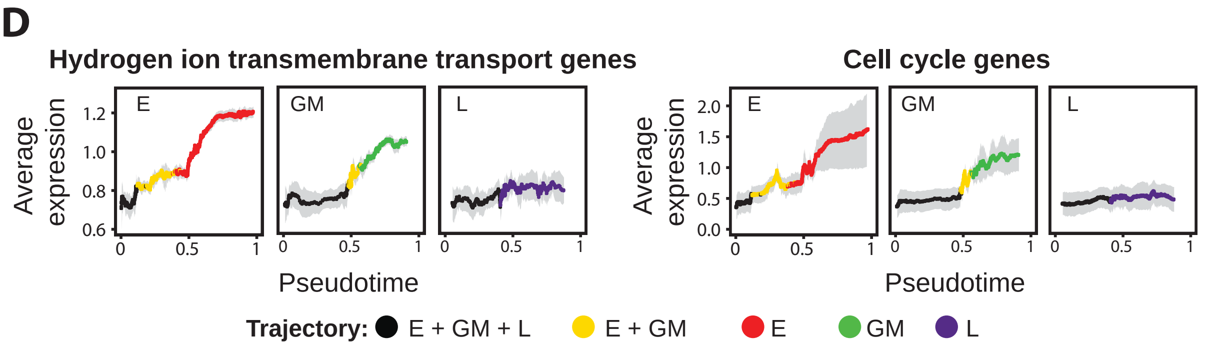
Remove cell cycle genes

Gene set enrichment analysis

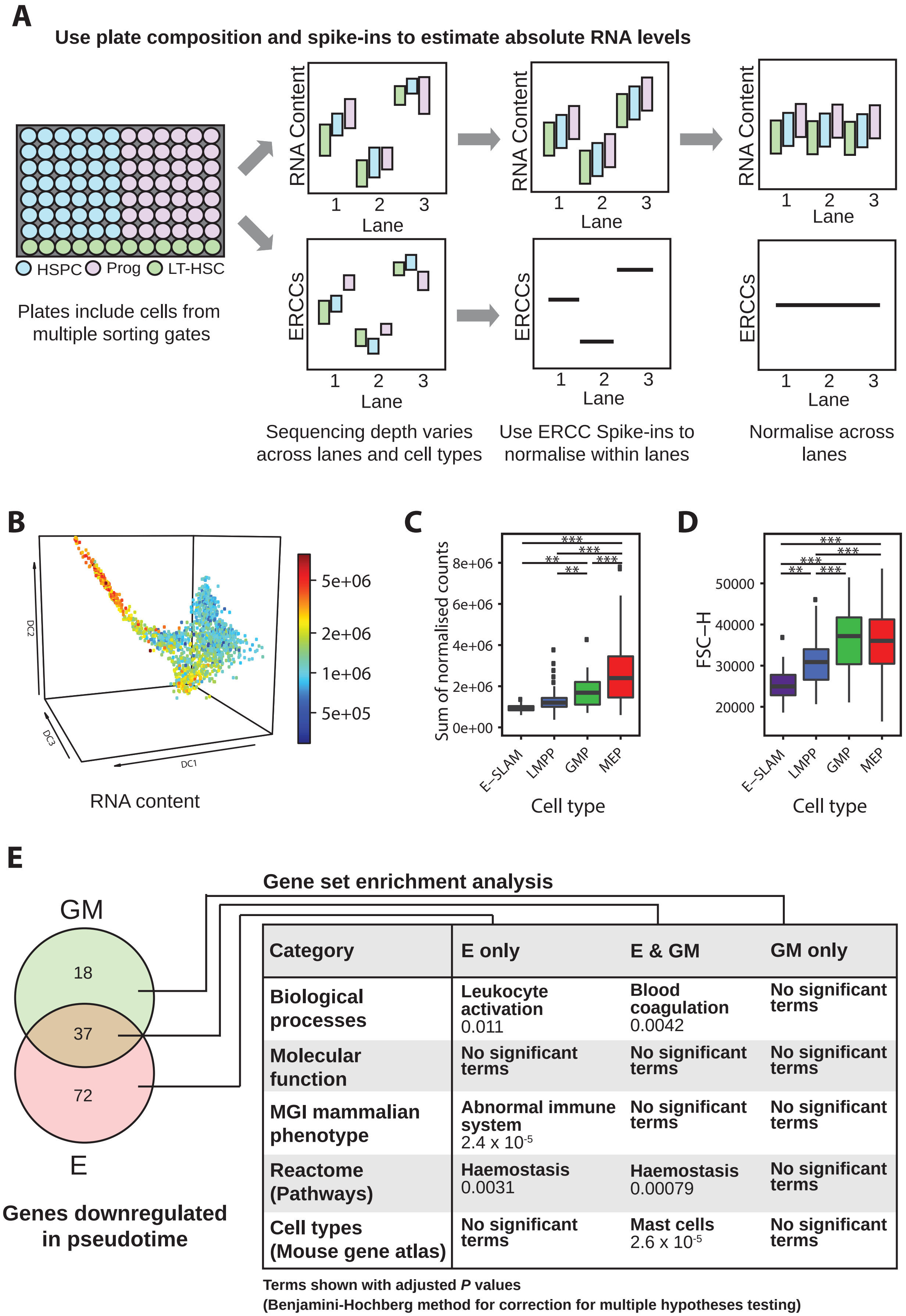
Gene set enrichment analysis

| Category | E only | E & GM | GM only |
|---|---|---|---|
| Biological processes | Porphyrin-containing compound biosynthetic process<br>$8.7 \times 10^{-7}$ | Hydrogen ion transmembrane transport<br>$1.6 \times 10^{-4}$ | Defense response to other organism<br>$4.9 \times 10^{-4}$ |
| Molecular function | ATP binding<br>$5.8 \times 10^{-6}$ | Hydrogen ion transmembrane transporter activity<br>$2.5 \times 10^{-5}$ | Serine-type peptidase activity<br>0.013 |
| MGI mammalian phenotype | Abnormal haematopoietic system<br>$2.0 \times 10^{-4}$ | Abnormal prenatal growth/weight<br>0.0042 | Abnormal immune system<br>$4.3 \times 10^{-6}$ |
| Reactome (Pathways) | Cytosolic tRNA aminoacylation<br>$1.1 \times 10^{-6}$ | Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins<br>$7.3 \times 10^{-4}$ | Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell<br>0.029 |
| Cell types (Mouse gene atlas) | Megakaryocyte erythrocyte progenitor<br>$3.2 \times 10^{-16}$ | No significant terms | Granulocyte monocyte progenitor<br>$1.7 \times 10^{-5}$ |

Terms shown with adjusted *P* values (Benjamini-Hochberg method for correction for multiple hypotheses testing)



# Figure 5



| Category | E only | E & GM | GM only |
|---|---|---|---|
| Biological processes | Leukocyte activation 0.011 | Blood coagulation 0.0042 | No significant terms |
| Molecular function | No significant terms | No significant terms | No significant terms |
| MGI mammalian phenotype | Abnormal immune system 2.4 x 10$^{-5}$ | No significant terms | No significant terms |
| Reactome (Pathways) | Haemostasis 0.0031 | Haemostasis 0.00079 | No significant terms |
| Cell types (Mouse gene atlas) | No significant terms | Mast cells 2.6 x 10$^{-5}$ | No significant terms |

Terms shown with adjusted *P* values
(Benjamini-Hochberg method for correction for multiple hypotheses testing)

**Figure 6**