Machine learning predicts putative haematopoietic stem cells within large single-cell transcriptomics datasets

Fiona K. Hamey[a,*], Berthold Göttgens[a,**]

[a] *Wellcome - MRC Cambridge Stem Cell Institute and Department of Haematology, University of Cambridge, Jeffrey Cheah Biomedical Centre, Cambridge Biomedical Campus, Cambridge, CB2 0AW, UK*

Abstract

Haematopoietic stem cells (HSCs) are an essential source and reservoir for normal haematopoiesis, and their function is compromised in many blood disorders. HSC research has benefitted from the recent development of single-cell molecular profiling technologies, where single-cell RNA-sequencing (scRNA-seq) in particular has rapidly become an established method to profile HSCs and related haematopoietic populations. The classical definition of HSCs relies on transplantation assays, which have been used to validate HSC function for cell populations defined by flow cytometry. Flow cytometry information for single cells however is not available for many new high-throughput scRNA-seq methods, thus highlighting an urgent need for the establishment of alternative ways to pinpoint the likely HSCs within large scRNA-seq datasets. To address this, we tested a range of machine learning approaches and developed a tool, hscScore, to score single-cell transcriptomes from murine bone marrow based on their similarity to gene expression profiles of validated HSCs. We evaluated hscScore across scRNA-seq data from different laboratories, which allowed us to establish a robust method that functions across different technologies. To facilitate broad adoption of hscScore by the wider haematopoiesis community, we have made the trained model and example code freely available online. In summary, our method hscScore provides fast identification of mouse bone marrow HSCs from scRNA-seq measurements and represents a broadly useful tool for analysis of single-cell gene expression data.

Keywords: Haematopoiesis; Stem cell; Single-cell; RNA-sequencing; Bioinformatics

*Corresponding author: Dr Fiona K. Hamey, Email: fkh23@cam.ac.uk, Phone: +44-1223-336827, Address: Wellcome - MRC Cambridge Stem Cell Institute, Jeffrey Cheah Biomedical Centre, Cambridge Biomedical Campus, Cambridge, CB2 0AW, UK

**Corresponding author: Professor Berthold Göttgens, Email: bg200@cam.ac.uk, Phone: +44-1223-336829, Address: Wellcome - MRC Cambridge Stem Cell Institute, Jeffrey Cheah Biomedical Centre, Cambridge Biomedical Campus, Cambridge, CB2 0AW, UK

Introduction

It has been over 60 years since experiments first proved the existence of bone marrow cells capable of producing the whole blood system. In the following decades, multipotent haematopoietic stem cells (HSCs) have been the subject of many studies aimed at revealing the mechanisms controlling their function [1]. Strategies to isolate blood cells were developed following the invention of techniques to sort cells based on their expression of specific proteins. By isolating and transplanting different fractions of bone marrow, sorting strategies could be refined to enrich for populations passing the gold-standard stem cell assay of repopulation upon secondary transplantation into irradiated mice (for review, see [2]). Once HSCs could be isolated it became possible to measure molecular properties of these cells.

However, it is well-known that many of the surface marker-defined haematopoietic stem and progenitor (HSPC) populations are very heterogeneous in terms of both function and their molecular profiles [3, 4, 5]. The field of haematopoiesis has therefore been at the forefront of exploring single-cell technologies. In particular, many studies have used single-cell RNA-sequencing (scRNA-seq) to profile gene expression across haematopoietic populations [4, 6, 7, 8, 9, 10]. This has provided insights into processes such as differentiation, ageing and disease (for review, see [11]).

Initial scRNA-seq studies were limited in throughput due to the cost and difficulty of profiling large numbers of cells. However, newer technologies such as droplet-based scRNA-seq methods [12, 13, 14] are enabling generation of increasingly large datasets, with multiple studies capturing tens of thousands of cells from the blood system [8, 15, 16, 17]. This has many exciting implications for haematopoiesis research, yet these technologies bring their own challenges. Our best strategies for identifying HSCs rely on measurements of cell surface marker proteins [18, 19]. However, many scRNA-seq datasets do not incorporate these measurements. Even in those studies using technologies such as index sorting [20, 21] or CITE-seq [22] to link protein and gene expression, the identification of HSCs is still dependent on the choice of markers measured in the experiment. Therefore, identifying potentially rare populations of HSCs in single-cell data remains a challenge.

To address this, we decided to develop an approach that could be easily applied to scRNA-seq data with the aim of identifying transcriptional profiles belonging to HSCs. Using annotated data from a previous study of mouse HSPCs [18], we tested a range of machine learning methods to score single-cell transcriptomes based on their similarity to HSC gene expression, and identified a model performing well across data from a range of different laboratories and technologies. Along with this manuscript we provide freely available code and the trained model so that researchers can easily apply this tool to their own single-cell datasets.

Materials and methods

scRNA-seq datasets

Model training data. Models were trained on data from Wilson et al. [18]. In this study, 96 HSCs (Lin$^-$ c-Kit$^+$ Sca1$^+$ CD34$^-$ Flt3$^-$ CD48$^-$ CD150$^+$) from mouse bone marrow were profiled using the Smart-Seq2 protocol [23]. Cells were filtered to the same 92 cells that passed stringent quality control (QC) measures in the original publication. Wilson et al. used a classification approach to assign scores to each transcriptome representing its similarity to a population highly enriched for functional HSCs (Fig. S1A). Data were visualised using principal component analysis (PCA) coordinates from the original publication. Count data, HSC-scores, QC information and PCA coordinates can be downloaded from Zenodo (`https://zenodo.org/`, DOI: 10.5281/zenodo.3303783).

Index-sorted HSPC data. Data profiling 1,654 HSPCs were published in Nestorowa et al. [7]. These data were generated with the same Smart-Seq2 protocol as the training data. After QC, 798 Lin$^-$ c-Kit$^+$ Sca1$^-$, 701 Lin$^-$ c-Kit$^+$ Sca1$^+$ and 155 Lin$^-$ c-Kit$^+$ Sca1$^+$ CD34$^-$ Flk2$^-$ cells were retained, and the count data for these cells can be downloaded from Zenodo (DOI: 10.5281/zenodo.3303783). QC information can be obtained from the data website (`http://blood.stemcells.cam.ac.uk/single_cell_atlas.html`). Data were visualised using the diffusion map coordinates and cell type information downloaded from the same data website.

Dormant and active HSC data. This dataset was described in Cabezas-Wallscheid et al. [24]. scRNA-seq data were generated using the Fluidigm C1 microfluidics device to profile HSCs (Lin$^-$ c-Kit$^+$ Sca1$^+$ CD150$^+$ CD48$^-$ CD34$^-$) and the subset of these cells that were long-term label-retaining, described as dormant HSCs (dHSCs). Gene expression counts for these data were downloaded from ArrayExpress (E-MTAB-4547). For QC, cells with <50,000 mapped reads, <1,000 detected genes or >30% of reads mapping to External RNA Controls Consortium (ERCC) spike-ins were excluded, as in the original publication. For visualisation, expression data were filtered to the highly variable genes (HVGs) from the original publication (Supplementary Table 2 in [24]). Cells were normalised to have total counts equal to the median counts per cell and normalised counts were log(x+1) transformed with the *scanpy.preprocessing.log1p* function. A diffusion map was calculated on these log-transformed values using 30 neighbours and the 'gauss' method in the *scanpy.tools.diffmap function.*

Smart-Seq2 data of multipotent stem and progenitors. Data profiling LT-HSCs (Lin$^-$ c-Kit$^+$ Sca1$^+$ CD150$^+$ CD48$^-$), ST-HSCs (Lin$^-$ c-Kit$^+$ Sca1$^+$ CD150$^-$ CD48$^-$) and MPPs (Lin$^-$ c-Kit$^+$ Sca1$^+$ CD150$^-$ CD48$^+$) were described in Mann et al. [25]. Expression counts were downloaded from NCBI GEO (GSE100426).

This study profiled cells from young (8-12 weeks) and old (20-24 months) mice, and in stimulated (LPS treated) and unstimulated conditions. For testing the hscScore method only unstimulated cells were used. QC was performed by removing cells with fewer than 2,000 detected genes. For visualisation, HVGs were identified using the *scanpy.preprocessing.filter_genes_dispersion* function with default settings and data were normalised and log-transformed as above. PCA was calculated on the log-transformed counts.

Droplet-based c-Kit$^+$ cells. Transcriptomes for 22,993 Lin$^-$ c-Kit$^+$ Sca1$^+$ and 21,809 Lin$^-$ c-Kit$^+$ transcriptomes were generated using the 10x genomics [12] droplet-based sequencing method and described in Dahlin et al. [15]. Data can be downloaded from `https://gottgens-lab.stemcells.cam.ac.uk/adultHSPC10X/` and NCBI GEO (GSE107727). Lin$^-$ c-Kit$^+$ cells from W^{41}/W^{41} mouse bone marrow were profiled similarly with data available from the same online resources. Data were visualised using the force-directed graph co-ordinates calculated for the original publication.

Droplet-based multipotent progenitors. Rodriguez-Fraticelli et al. [26] describe the generation of inDrops [13] scRNA-seq data from mouse bone marrow for each of the LT-HSC (Lin$^-$ c-Kit$^+$ Sca1$^+$ Flt3$^-$ CD150$^+$ CD48$^-$), ST-HSC (Lin$^-$ c-Kit$^+$ Sca1$^+$ Flt3$^-$ CD150$^-$ CD48$^-$), MPP2 (Lin$^-$ c-Kit$^+$ Sca1$^+$ Flt3$^-$ CD150$^+$ CD48$^+$), MPP3 (Lin$^-$ c-Kit$^+$ Sca1$^+$ Flt3$^-$ CD150$^-$ CD48$^+$) and MPP4 (Lin$^-$ c-Kit$^+$ Sca1$^+$ Flt3$^+$ CD48$^+$) fractions. Processed count matrices were downloaded from NCBI GEO (GSE90742) and QC was performed by excluding any cells with fewer than 1,000 detected genes. For visualisation, PCA was calculated as above and then UMAP [27] coordinates calculated using the *scanpy.tools.umap* function with default parameters.

Data pre-processing

Before input into the model, count data were processed by gene filtering and normalisation. The gene filtering retained genes in one of three sets: 1) all protein-coding genes, 2) HVGs or 3) MolO and NoMO gene sets. For option 1, only non-mitochondrial genes annotated as 'protein_coding' in the Ensembl version 81 annotation [28] were retained. For option 2, HVGs were calculated on normalised counts of all protein-coding genes (normalised using the *scanpy.preprocessing.normalize_total* function with default parameters). These normalised counts were log(x+1)-transformed and HVGs identified with the *scanpy.preprocessing.highly_variable_genes* function with default parameters. Raw count data were filtered to this set of HVGs for input into the model. Option 3 retained the genes from Wilson et al. Supplementary Table 3 annotated as either MolO or NoMO genes [18]. These genes were those with significant correlation with the HSC-score assigned to each cell (adjusted P-value < 0.1, Benjamini-Hochberg correction for multiple testing).

After feature selection, count data were normalised on the selected genes using one of two alternatives: 1) rank normalisation or 2) total count normalisation. For rank normalisation, expression in each cell

was replaced by a vector representing the expression values ranked within that cell. Genes with identical counts were replaced with their average rank. For option 2, normalisation was performed with the *scanpy.preprocessing.normalize_total* function to normalise each cell to have the same summed counts. This number of counts was set to be the median number of counts for the Wilson et al. data across the gene set of choice. Total count-normalised data were then log(x+1)-transformed.

Model training

To identify optimal parameters for each type of model, a search over parameters was performed using the *sklearn.GridSearchCV* function with 5-fold cross validation. Parameters explored for each model can be found in Supplementary Table S2. Before training, 25% of the data were held back as a test set and the remaining 75% were scaled using the *sklearn.StandardScaler* function and then (optionally) PCA-transformed. The optimal parameters identified by the grid search are shown in Supplementary Table S3, along with the model R^2 scores for each cross-validation fold, the mean and standard deviation of these scores, and the score of the trained model on the unseen test data. After the optimal parameters were obtained the models were retrained on the whole dataset using these parameters.

Plotting

Plotting was performed in python using either *scanpy* [29], *seaborn* or *matplotlib* functions.

Clustering and cell cycle scoring

Leiden clustering [30] was performed using the *scanpy.tl.leiden* function with resolution equal to either 1.0 for lower resolution clustering or 1.5 for higher resolution clustering. Before clustering, data from Nestorowa et al. were normalised using the *scanpy.preprocessing.normalize_total* function, log(x+1)-transformed and then HVGs identified with the *scanpy.preprocessing.highly_variable_genes* function. PCA was calculated on the HVG values and the top 8 principal components used for input to the clustering. Cell cycle scoring was performed by using the *scanpy.tl.score_genes_cell_cycle* function with S phase and G2/M phase genes downloaded from Macosko et al. [14].

Code availability

Scripts for identifying model parameters and producing plots in this manuscript are hosted on GitHub (`https://github.com/fionahamey/hscScore`). The trained model can be downloaded from Zenodo (DOI: 10.5281/zenodo.3332150). An example notebook of applying the model to new data is also hosted on GitHub.

Software versions

Versions of all software used can be found in the supplementary text.

Results

Linked stem cell function and gene expression data can be used to train models to identify HSCs

As our aim was to identify HSCs we first required data where it was already known which transcriptomes belonged to these cells. This annotation could be done using surface marker expression, but even the purest HSC strategies still only contain up to 70% functional stem cells [18]. Therefore, we chose a dataset of HSCs that were profiled as part of a study where these cells were annotated with an HSC-score based on their gene expression [18]. This score represented each cell's transcriptional similarity to a highly homogeneous population of HSCs (Fig. 1A, S1A). In this previous work, cells profiled using scRNA-seq were index-sorted to measure 11 flow cytometry parameters. To establish a link between the HSC-score and the functional output of a stem cell, single-cell transplantation assays were performed where the same 11 flow cytometry parameters were recorded for each of the transplanted cells. Based on these shared parameters dimensionality reduction was used to show that the repopulating HSCs in the single-cell transplantation experiments possessed similar surface marker profiles to the high HSC-score cells. Therefore, this study established the correlation between having a high HSC-score and giving a positive read-out in a transplantation assay designed to test for stem cell function [18]. Here, our aim was to use these scored transcriptomes to train models to predict the HSC-score of cells from new datasets (Fig. 1B). To find the most suitable type of model for this prediction, we trained a number of different machine learning methods (linear regression, random forest regression, nearest neighbours regression, support vector regression and multi-layer perceptron (MLP) regression) and scored the performance of each method on a test subset of the data (Fig. S1B). Model parameters were fitted using a grid search approach with 5-fold cross validation and then models were tested on unseen test to assess their accuracy in predicting the HSC-score.

Using a select subset of genes for training produces the most accurate models

Before training any models it was first necessary to define a pipeline for processing any scRNA-seq dataset given as input to the model. In particular, it was important to choose analysis steps that would allow comparison of data across different experiments. Although scRNA-seq can measure thousands of genes per cell, the majority of genes detected across a dataset have very noisy expression. To avoid obscuring biological variation in the data, often only a set of so-called highly variable genes (HVGs) that exceed a certain level of variance are used for analysis [31]. To explore the effect of gene set choice we decided to test models on

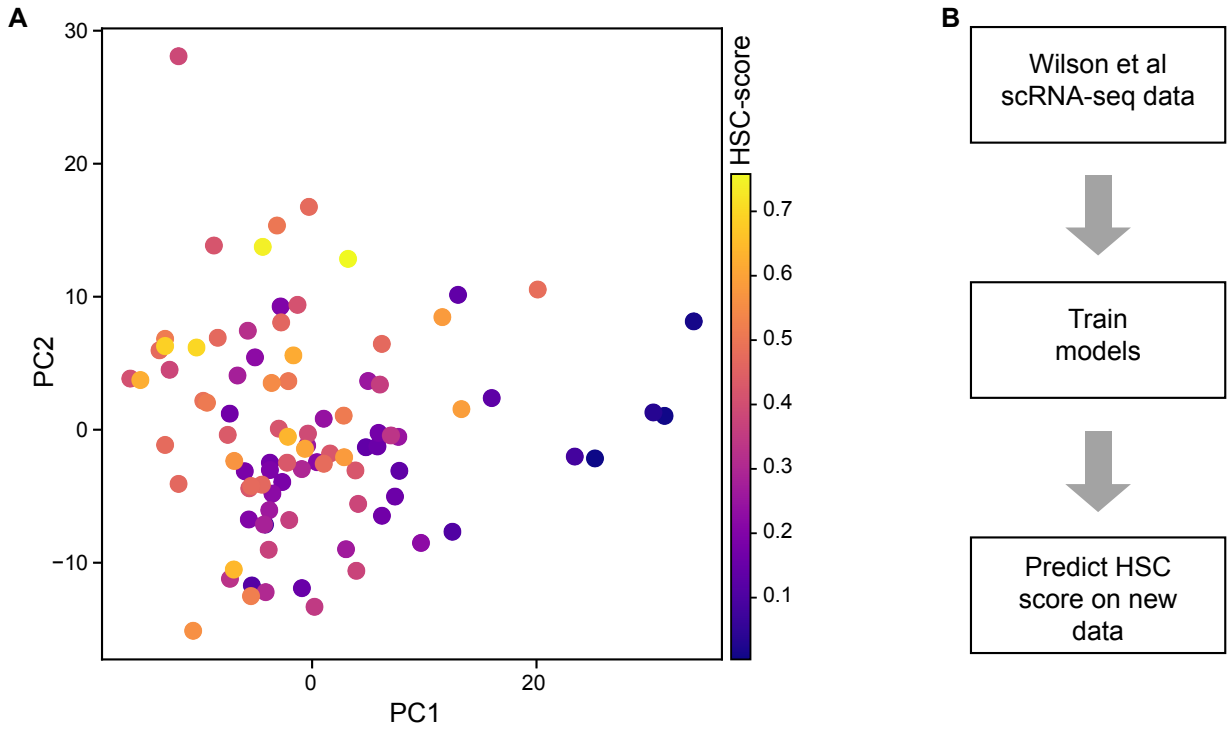

Figure 1: **Predicting HSC identity in single-cell gene expression datasets.** (A) Data from Wilson et al. [18] was used as training data for models predicting HSC identity in scRNA-seq datasets. In this study, 92 transcriptomes of HSCs were assigned a value, the HSC-score, where a higher HSC-score represents greater similarity to transcriptional profiles of functionally validated HSCs. (B) Outline of the training process for building the HSC prediction tool.

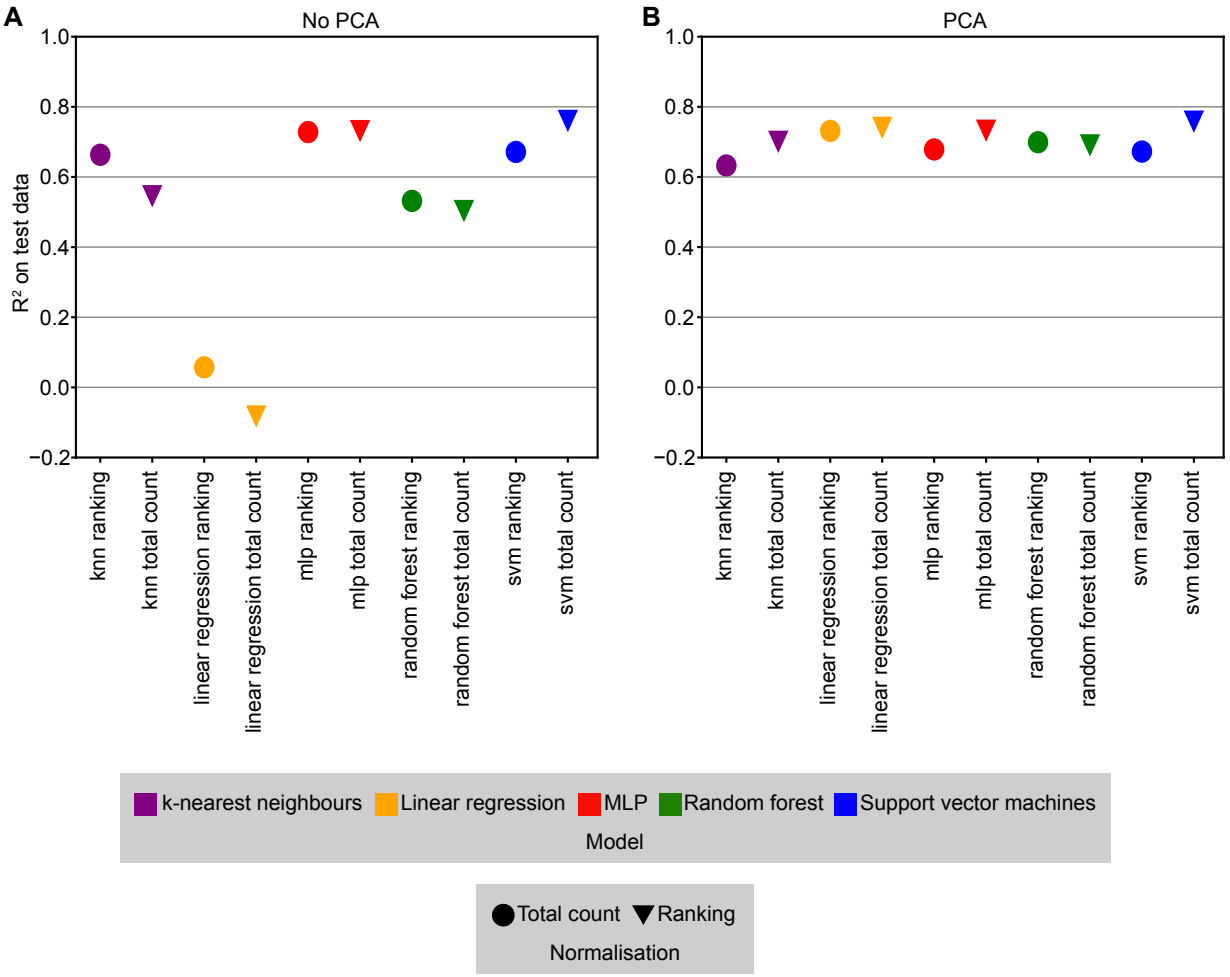

Figure 2: **Trained models can predict HSC-score on unseen test data.** (A, B) R^2 score of predicted compared to actual HSC-score on test subset of data for models trained with best identified parameters. Shape indicates normalisation, and colour the type of method. Results are shown for models trained on raw counts (A) or PCA-transformed counts (B).

three different gene sets: all protein-coding genes, HVGs, and the set of "MolO" and "NoMO" genes defined by Wilson et al. [18] (Table S1). Wilson et al. correlated the expression of all genes with the HSC-score within their scRNA-seq data, and denoted genes with significant positive correlation with the HSC-score as "MolO" genes and those with significant negative correlation as "NoMO" genes. Further details of these three different gene lists used for training can be found in the Materials and methods. As well as the choice of gene set, we also chose to test different data normalisation methods, similar to work aimed at predicting cell cycle state based on gene expression [32].

There are many different normalisation approaches that have been applied to single-cell data, yet we needed one that would yield comparable results across multiple datasets. This requirement excluded many of the

more sophisticated methods that share information across a sample to perform normalisation [33, 34]. We tested both total count normalisation and a ranking normalisation method (see Materials and methods). Finally, we also tried training models on PCA-transformed data, reasoning that projecting new data into the PCA-space of the training data could help to relate datasets from different technologies. Inspection of models trained across these combinations of pre-processing variables showed that the best performing models were all trained using the MolO and NoMO genes (Fig. S2). In general, models trained on the PCA-transformed data performed better on unseen data (Fig. 2), although some models trained on untransformed counts were still amongst the highest scoring (Fig. 2, S2).

HSCs are successfully identified in a broad dataset of blood stem and progenitors

After assessing the performance of the models on test data held back from the original dataset, we next applied the highest scoring models to an alternative dataset containing over 1,600 HSPCs from mouse bone marrow using the same scRNA-seq protocol as the training data [7]. As this protocol was a plate-based method, cells were index sorted hence single-cell transcriptomes could be retrospectively assigned to one of 10 different phenotypic cell types (Fig. 3A). This dataset contained 38 cells from the highly specific ESLAM (Lin^- c-Kit^+ $Sca1^+$ $EPCR^+$ $CD48^-$ $CD150^+$) HSC population [19] as well as more mature progenitor cells, allowing our models to be tested on a broader population than the training data. Diffusion map dimensionality reduction [35, 36] showed separation of HSCs from cells differentiating into erythroid, lymphoid and myeloid lineages. For the majority of high scoring models, high HSC-scores were localised to the top of the diffusion map in the region occupied by the ESLAM cells (Fig. 3B, S3A). HSC scores were significantly higher in the ESLAM population when compared to other phenotypic cell types for a number of the models (Fig. 3C, S3B, Wilcoxon rank-sum test, P-values in figure). Overall, the MLP model with total count normalisation and no PCA-transformation gave the best distribution of HSC-scores across the dataset, with high scoring cells largely restricted to the ESLAM population. The score across all other populations was low, meaning this model was specifically highlighting the stem cells. As this combination of parameters mostly clearly highlighted the ESLAM cells that are enriched for functional HSCs in dimensionality reduction and violin plots, we therefore chose to carry this model forward for testing across a wider range of experiments and denote this prediction pipeline as hscScore.

One of the most widely-used steps in the analysis of single-cell data is the application of clustering algorithms. Comparison of hscScore with a graph-based clustering approach [30] showed that whilst clustering could identify a broad stem cell region, it nevertheless struggled to separate out the highest HSC-score cells even with increased clustering resolution (Fig. S4A,B). Clustering is also limited as it assigns cells into discrete groups, whereas haematopoietic differentiation may be better defined by a continuous representation [1]. Next, as there is known to be a link between cell cycle activity and repopulation capability of HSCs, we

decided to compare scoring cells with hscScore to scoring cells by their expression of cell cycle genes [14]. In keeping with the reported quiescent nature of functional HSCs [37], we found a correlation between HSC-score and cell cycle score with the group of cells most transcriptionally similar to HSCs having very low expression of cell cycle genes (Fig. S4C). This inverse relation between the HSC-score and cell cycle activity again supports the ability of hscScore to identify the stem cell population.

hscScore locates HSCs in single-cell datasets produced by different technologies

To test the model performance on data generated from a different laboratory and using an alternative scRNA-seq technology we decided to investigate data from work by Cabezas-Wallscheid et al. [24]. In this study, the authors profiled dormant HSCs (dHSCs), a subset of HSCs that show long-term label retention in label-retaining assays. Previous work had shown that these dHSCs were enriched for repopulation potential, and therefore represent a subset of HSCs containing a higher proportion of functional stem cells. 146 dHSCs and 170 HSCs were profiled using microfluidics scRNA-seq technology (Fig. 4Ai). Diffusion map dimensionality reduction shows a progression from dHSCs to other cells within the HSC gate, which in the original study are shown to represent more "active" HSCs primed for cell cycle entry. Applying hscScore to these data showed significantly higher ($P=1.1 \times 10^{-19}$, Wilcoxon rank-sum test) scores in the dHSCs compared to the overall HSC population (Fig. 4Aii, iii). We also tested our model on an additional dataset containing long-term HSC (LT-HSC), short-term HSC (ST-HSC) and multipotent progenitor (MPP) populations [25] (Fig. 4Bi). Again, highest scores were seen in the LT-HSC population, with lowest scores in the MPPs (Fig. 4Bii, iii, S4D).

Next, we wanted to see if our method would also work for higher throughput single-cell gene expression methods such as droplet-based scRNA-seq. These approaches capture much higher numbers of cells but at least until now have had much lower sequencing depth. Additionally, many existing HSPC droplet-based scRNA-seq datasets do not have surface marker information for cells that would allow phenotypic populations to be identified. Application of hscScore to droplet-based data profiling Lin^- $c\text{-}Kit^+$ mouse bone marrow cells [15] identified highest scoring cells in a specific region of the diffusion map (Fig. 4Ci). Inspection of HSC marker genes *Procr* [38] and *Hoxb5* [39] showed overlap between high HSC-score and expression of these genes (Fig. 4Cii, iii). To examine another lower sequencing depth method, we calculated HSC-scores for LT-HSC, ST-HSC and MPP cells profiled using alternative droplet-based method [26], and again the highest scores were found in the LT-HSCs (Fig. S5A).

We also asked how our method compared to a naïve approach of simply averaging MolO gene expression across cells, as we had previously found this to be useful to highlight the HSC population [15]. Whilst we confirmed that this approach of averaging the expression of a specific gene set gave higher averages in the HSCs, these differences were not as clear as the hscScore model results. Instead, the average expression

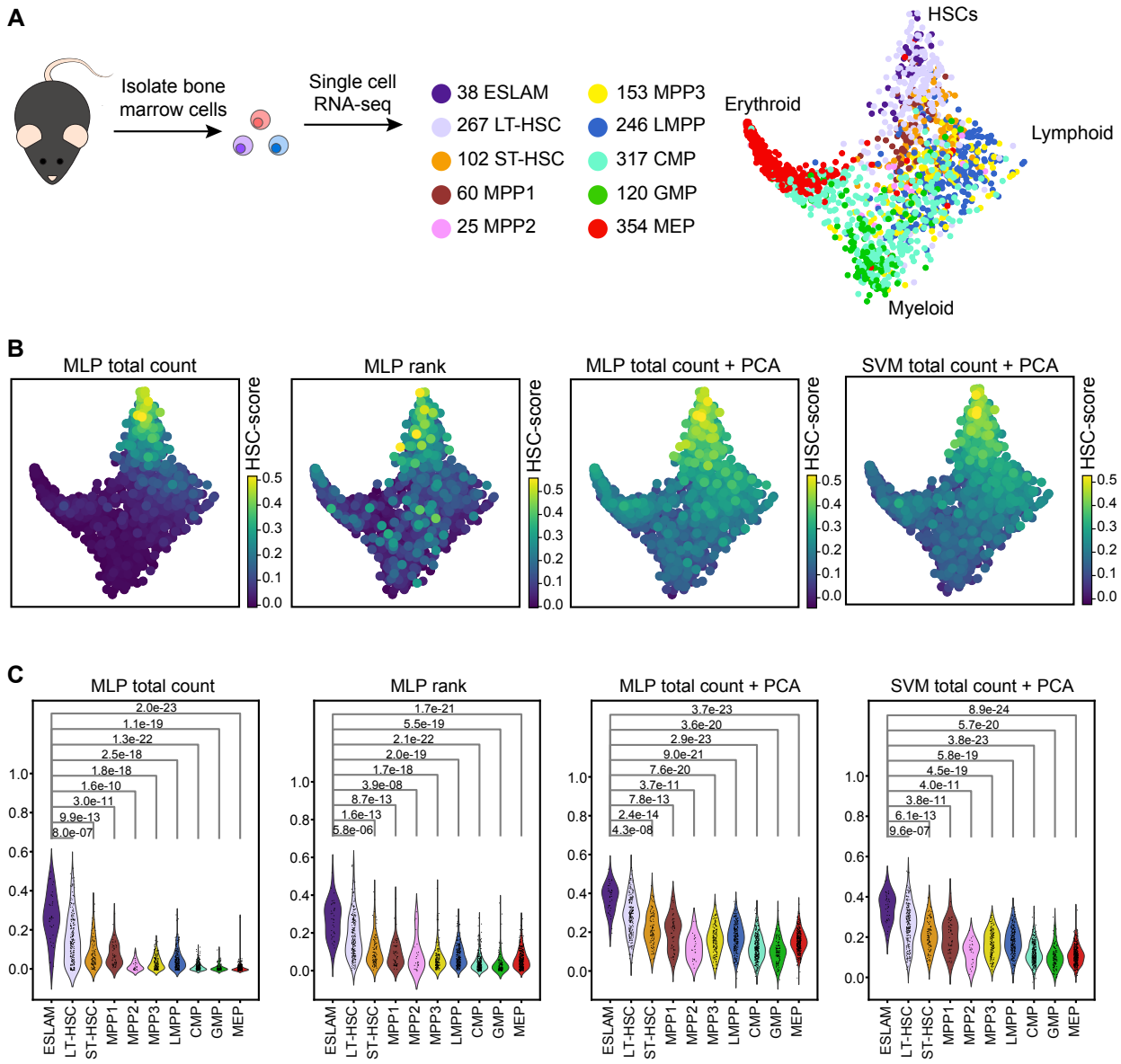

Figure 3: **Top-performing models can identify HSCs in alternative dataset profiling haematopoietic stem and progenitor cells.** (A) Schematic of experiment from Nestorowa et al. [7] showing the number of cells for each surface marker-defined cell type in the scRNA-seq dataset. Diffusion map dimensionality reduction is coloured by surface marker cell type. (B) Diffusion map coloured by the predicted HSC score from the top-performing models. Additional plots are shown in Fig. S3. Highest scores are seen in the region corresponding to phenotypic stem cells. (C) Violin plots showing distribution of scores across surface marker-defined phenotypes. P-values indicate significance of pairwise tests between scores of each population in comparison to scores of ESLAM population, Wilcoxon rank-sum test. Additional plots are shown in Fig. S3. ESLAM, EPCR$^+$ subset of HSCs; LT-HSC, long-term HSC; ST-HSC, short-term HSC; MPP, multipotent progenitor; LMPP, lymphoid-primed multipotent progenitor; CMP, common myeloid progenitor; GMP, granulocyte-macrophage progenitor; MEP, megakaryocyte-erythroid progenitor; MLP, multi-layer perceptron; SVM, support vector machine.

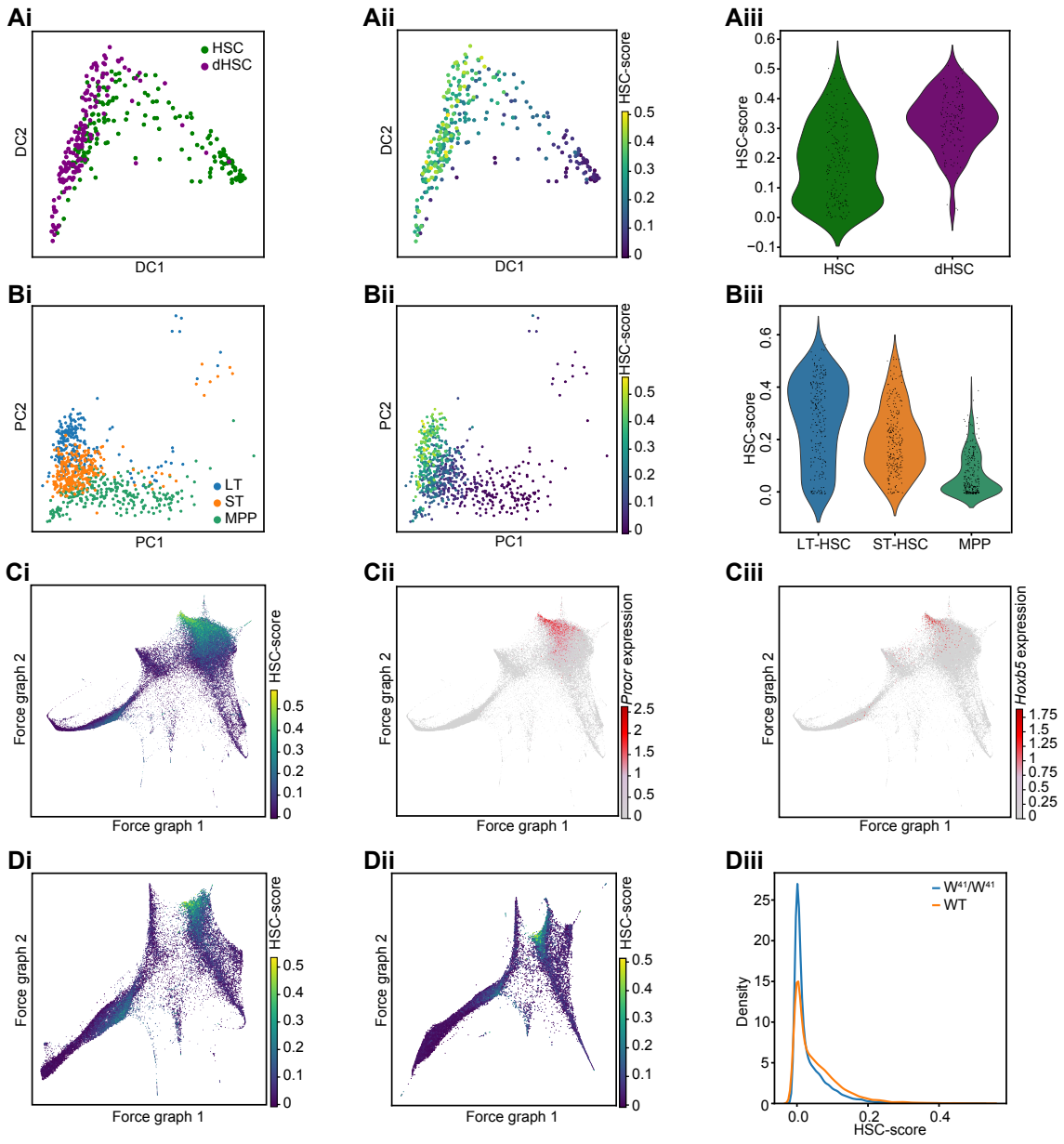

Figure 4: **HSCs can be identified in datasets generated using different technologies.** (A) Model performance on 316 HSCs from Cabezas-Wallscheid et al. [24]. Diffusion maps show data coloured by cell sorting gate (i) and by predicted hscScore (ii). dHSC, dormant HSC. (iii) Violin plot shows HSC-score distribution over the dHSC and HSC gates. (B) Model applied to 718 SMART-Seq2 scRNA-seq profiles of stem and progenitor cells from Mann et al. [25]. PCA plots show the cell type (i) and predicted HSC-score (ii). (iii) The violin plot shows the score distribution across LT-HSC, ST-HSC and MPP cells. (C) Application of top-performing model to droplet-based scRNA-seq data of 44,802 Lin$^-$ c-Kit$^+$ bone marrow cells from Dahlin et al. [15]. Data are visualised using a force-directed graph coloured by predicted HSC-score (i). Expression of HSC marker genes *Procr* and *Hoxb5* are shown in panels (ii) and (iii), respectively. (D) Force-directed graph of Lin$^-$ c-Kit$^+$ bone marrow cells from wild-type (i) and W^{41}/W^{41} (ii) mouse bone marrow coloured by predicted HSC-score. (iii) Distribution of HSC-score across the wild-type and W^{41}/W^{41} datasets.

showed more of a gradient across HSPC populations (Fig. S5B-F), making it more challenging to clearly distinguish the HSCs with this approach.

hscScore distribution is in keeping with lower proportion of stem cells in bone marrow of Kit mutant mouse

Finally, we applied our scoring method to previously published droplet-based scRNA-seq data from W^{41}/W^{41} mouse bone marrow [15]. The W^{41}/W^{41} mutation leads to reduced c-Kit signalling activity and these mice have a lower proportion of stem cells [40, 41]. We wanted to see if our approach could both detect stem cells in the mutant background and identify their shift in numbers. Dimensionality reduction on both wild-type and W^{41}/W^{41} Lin$^-$ c-Kit$^+$ cells showed very similar appearances and localisation of the cells with high HSC-scores, verifying that this tool can be applied to these data from perturbed haematopoiesis (Fig. 4Di, ii). The distribution of the HSC-score across the whole dataset showed the W^{41}/W^{41} population had overall lower scores, in keeping with the reduction of HSCs within this mutant model (Fig. 4Diii). The wild-type Lin$^-$ c-Kit$^+$ population is expected to contain around 1% HSCs so we calculated the 99th percentile of the wild-type Lin$^-$ c-Kit$^+$ HSC-score. Only 0.56% of W^{41}/W^{41} HSCs had a predicted score above this same threshold. This was in spite of the numerical range of scores being similar across these datasets (-7.8×10^{-3} – 0.51 for W^{41}/W^{41} and -8.3×10^{-3} – 0.53 for wild-type cells). Together this shows that the hscScore method gives results in keeping with the reduced frequency of stem cells in the W^{41}/W^{41} mouse model.

Discussion

A rapidly growing number of studies use single-cell gene expression profiling to investigate the molecular state of blood stem and progenitor cells. One of the challenges when working with this type of data is to reliably identify the transcriptomes belonging to rare cell types. This is particularly relevant for those cell types conventionally defined by expression of specific cell surface marker proteins as many scRNA-seq datasets do not contain information about protein expression. In this work we trained and tested a range of predictive machine learning models to develop a tool to score single-cell gene expression profiles for their transcriptional similarity to a functionally pure population of HSCs.

It is well-established that integrating or comparing scRNA-seq data from different sources can be difficult due to so-called batch-effects arising from factors such as different experimental techniques [42, 43]. We therefore tested our method across a number of datasets and identified a pipeline that performed well across scRNA-seq platforms with different sequencing depths. Optimal model performance was found when training on a small set of genes highly correlated with the HSC-score. We chose to include genes with both positive and negative associations to provide as much information as possible to distinguish between “good” and “bad”

stem cells. The inclusion of these negatively correlated genes as well as the fact that the hscScore model can learn specific weights for each gene offers benefits over simply averaging the expression of a geneset. The flexibility in the MLP framework also allows varying weights across genes, meaning that there are different combinations of gene expression enabling a cell to get a high HSC-score.

We made efforts to ensure that our approach can be easily applied by other researchers, providing both the trained model and example code online. We envisage the hscScore method to be an easy step in the analysis of murine bone marrow scRNA-seq samples, enabling fast and reliable identification of HSCs in a dataset. When the expected frequency of stem cells in a sample is known, it could be used to select a threshold for classifying cells based on their HSC-score, although this information will not be available for all datasets. In these cases, hscScore can still be used to reveal the most likely stem cells instead of being used for strict classification. Our hscScore approach also has the potential to be used as part of a pipeline for refining stem cell sorting strategies by identifying any genes that encode for surface marker proteins and have expression levels correlated with the HSC-score. With high quality cell state annotation this approach could be applied to other systems. In particular, this would be worth exploring in systems where there are linked functional data and expression data, for example through the expression of shared surface marker profiles. Of special interest to haematopoiesis it would be interesting to try and extend this approach to identifying human HSCs, as a number of markers differ between human and mouse HSCs.

An exciting potential application of the hscScore method will be to compare data across different conditions, including genetic perturbations such as the W^{41}/W^{41} mouse model explored here. A number of blood disorders affect stem cell behaviour, and in such situations surface marker expression is commonly disrupted, making it unreliable to identify HSCs using conventional strategies. In particular, there are several mouse models where an increase in the number of phenotypic HSCs but a decrease in the number of functional HSCs has been described. Where this decreased functionality is linked to transcriptional changes then a lower frequency of stem cells should be seen with hscScore. Being able to robustly identify HSCs within scRNA-seq data could therefore provide important new insights into disrupted haematopoiesis in these situations.

In summary, the hscScore model provides a fast and simple approach for identification of HSCs within scRNA-seq datasets from mouse bone marrow. This should provide a broadly useful tool for analysis of single-cell gene expression data, which we hope will be adopted widely by the community.

Acknowledgements

We thank Nicola Wilson and David Kent for helpful discussions on this project. Research in B.G.'s laboratory is supported by Bloodwise, CRUK and the infrastructure from Wellcome and the MRC to the Wellcome

– MRC Cambridge Stem Cell Institute. F.K.H. is funded by a MRC Physical Biology of Stem Cells PhD studentship and by part of a Wellcome Strategic Award (105031/D/14/Z) awarded to W. Reik, S. Teichmann, J. Nichols, B.D. Simons, T. Voet, S. Srinivas, L. Vallier, B.G. and J.C. Marioni.

References

[1] E. Laurenti, B. Göttgens, From haematopoietic stem cells to complex differentiation landscapes, Nature 553 (7689) (2018) 418. doi:10.1038/nature25022.

[2] A. Mayle, M. Luo, M. Jeong, M. A. Goodell, Flow cytometry analysis of murine hematopoietic stem cells, Cytometry 83A (1) (2013) 27–37. doi:10.1002/cyto.a.22093.

[3] D. Karamitros, B. Stoilova, Z. Aboukhalil, F. Hamey, A. Reinisch, M. Samitsch, L. Quek, G. Otto, E. Repapi, J. Doondeea, B. Usukhbayar, J. Calvo, S. Taylor, N. Goardon, E. Six, F. Pflumio, C. Porcher, R. Majeti, B. Göttgens, P. Vyas, Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells, Nature Immunology 19 (1) (2018) 85–97. doi:10.1038/s41590-017-0001-2.

[4] F. Paul, Y. Arkin, A. Giladi, D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, D. Winter, D. Lara-Astiaso, M. Gury, A. Weiner, E. David, N. Cohen, F. K. B. Lauridsen, S. Haas, A. Schlitzer, A. Mildner, F. Ginhoux, S. Jung, A. Trumpp, B. T. Porse, A. Tanay, I. Amit, Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors, Cell 163 (7) (2015) 1663–1677. doi:10.1016/J.CELL.2015.11.013.

[5] L. Perié, K. R. Duffy, L. Kok, R. J. de Boer, T. N. Schumacher, The Branching Point in Erythro-Myeloid Differentiation, Cell 163 (7) (2015) 1655–1662. doi:10.1016/j.cell.2015.11.059.

[6] B. K. Tusi, S. L. Wolock, C. Weinreb, Y. Hwang, D. Hidalgo, R. Zilionis, A. Waisman, J. R. Huh, A. M. Klein, M. Socolovsky, Population snapshots predict early haematopoietic and erythroid hierarchies, Nature 555 (7694) (2018) 54–60. doi:10.1038/nature25741.

[7] S. Nestorowa, F. K. Hamey, B. Pijuan Sala, E. Diamanti, M. Shepherd, E. Laurenti, N. K. Wilson, D. G. Kent, B. Göttgens, A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation, Blood 128 (8) (2016) e20–e31. doi:10.1182/blood-2016-05-716480.

[8] A. Giladi, F. Paul, Y. Herzog, Y. Lubling, A. Weiner, I. Yofe, D. Jaitin, N. Cabezas-Wallscheid, R. Dress, F. Ginhoux, A. Trumpp, A. Tanay, I. Amit, Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis, Nature Cell Biology 20 (7) (2018) 836–846. doi:10.1038/s41556-018-0121-4.

[9] A. Olsson, M. Venkatasubramanian, V. K. Chaudhri, B. J. Aronow, N. Salomonis, H. Singh, H. L. Grimes, Single-cell analysis of mixed-lineage states leading to a binary cell fate choice, Nature 537 (7622) (2016) 698–702. doi:10.1038/nature19348.

[10] L. Velten, S. F. Haas, S. Raffel, S. Blaszkiewicz, S. Islam, B. P. Hennig, C. Hirche, C. Lutz, E. C. Buss, D. Nowak, T. Boch, W.-K. Hofmann, A. D. Ho, W. Huber, A. Trumpp, M. A. G. Essers, L. M. Steinmetz, Human haematopoietic stem cell lineage commitment is a continuous process, Nature Cell Biology 19 (4) (2017) 271–281. doi:10.1038/ncb3493.

[11] S. Watcham, I. Kucinski, B. Göttgens, New insights into hematopoietic differentiation landscapes from single-cell RNA sequencing, Blood 133 (13) (2019) 1415–1426. doi:10.1182/blood-2018-08-835355.

[12] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, J. H. Bielas, A. K. Shalek, Q. F. Wills, D. A. Jaitin, A. A. Pollen, E. Z. Macosko, A. M. Klein, G. X. Zheng, V. M. Narasimhan,

Y. Mostovoy, B. J. Hindson, P. Brennecke, G. Sherlock, L. J. P. van der Maaten, G. E. Hinton, F. Borrego, M. Masilamani, A. I. Marusina, X. Tang, J. E. Coligan, P. G. Chu, D. A. Arber, A. Schiopu, O. S. Cotoi, M. A. Turman, T. Yabe, C. McSherry, F. H. Bach, J. P. Houchins, E. Lubberts, S. Ronchetti, Y. Y. Lin, A. M. Greer, A. N. Harman, A. P. Patel, I. Tirosh, M. C. Lee, K. T. Kim, J. W. Vardiman, J. F. Zhong, A. Dobin, M. Stephens, Q. Liu, N. Novershtern, M. Bonora, C. Schinke, Massively parallel digital transcriptional profiling of single cells, Nature Communications 8 (2017) 14049. doi:10.1038/ncomms14049.

[13] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, M. W. Kirschner, Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells, Cell 161 (5) (2015) 1187–1201. doi:10.1016/J.CELL.2015.04.044.

[14] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, S. A. McCarroll, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets, Cell 161 (5) (2015) 1202–1214. doi:10.1016/j.cell.2015.05.002.

[15] J. S. Dahlin, F. K. Hamey, B. Pijuan-Sala, M. Shepherd, W. W. Y. Lau, S. Nestorowa, C. Weinreb, S. Wolock, R. Hannah, E. Diamanti, D. G. Kent, B. Göttgens, N. K. Wilson, A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in kit mutant mice., Blood 131 (21) (2018) e1–e11. doi:10.1182/blood-2017-12-821413.

[16] S. Zheng, E. Papalexi, A. Butler, W. Stephenson, R. Satija, Molecular transitions in early progenitors during human cord blood hematopoiesis, Molecular Systems Biology 14 (3) (2018) e8041. doi:10.15252/msb.20178041.

[17] D. Pellin, M. Loperfido, C. Baricordi, S. L. Wolock, A. Montepeloso, O. K. Weinberg, A. Biffi, A. M. Klein, L. Biasco, A comprehensive single cell transcriptional landscape of human hematopoietic progenitors, Nature Communications 10 (1) (2019) 2395. doi:10.1038/s41467-019-10291-0.

[18] N. K. Wilson, D. G. Kent, F. Buettner, M. Shehata, I. C. Macaulay, F. J. Calero-Nieto, M. Sánchez Castillo, C. A. Oedekoven, E. Diamanti, R. Schulte, C. P. Ponting, T. Voet, C. Caldas, J. Stingl, A. R. Green, F. J. Theis, B. Göttgens, Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations, Cell Stem Cell (2015) 712–724doi:10.1016/j.stem.2015.04.004.

[19] D. G. Kent, M. R. Copley, C. Benz, S. Wöhrer, B. J. Dykstra, E. Ma, J. Cheyne, Y. Zhao, M. B. Bowie, Y. Zhao, M. Gasparetto, A. Delaney, C. Smith, M. Marra, C. J. Eaves, Prospective isolation and molecular characterization of hematopoietic stem cells with durable self-renewal potential., Blood 113 (25) (2009) 6342–50. doi:10.1182/blood-2008-12-192054.

[20] G. W. Osborne, Recent Advances in Flow Cytometric Cell Sorting, in: Methods in Cell Biology, 2011, Ch. 21, pp. 533–556. doi:10.1016/B978-0-12-374912-3.00021-3.

[21] R. Schulte, N. K. Wilson, J. C. M. Prick, C. Cossetti, M. K. Maj, B. Göttgens, D. G. Kent, Index sorting resolves heterogeneous murine hematopoietic stem cell populations., Experimental Hematology 43 (9) (2015) 803–11. doi:10.1016/j.exphem.2015.05.006.

[22] M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, P. Smibert, Simultaneous epitope and transcriptome measurement in single cells, Nature Methods 14 (9) (2017) 865–868. doi:10.1038/nmeth.4380.

[23] S. Picelli, O. R. Faridani, Å. K. Björklund, G. Winberg, S. Sagasser, R. Sandberg, Full-length RNA-seq from single cells using Smart-seq2, Nature Protocols 9 (1) (2014) 171–181. doi:10.1038/nprot.2014.006.

[24] N. Cabezas-Wallscheid, F. Buettner, P. Sommerkamp, D. Klimmeck, L. Ladel, F. B. Thalheimer, D. Pastor-Flores, L. P. Roma, S. Renders, P. Zeisberger, A. Przybylla, K. Schönberger, R. Scognamiglio, S. Altamura, C. M. Florian, M. Fawaz, D. Vonficht, M. Tesio, P. Collier, D. Pavlinic, H. Geiger, T. Schroeder, V. Benes, T. P. Dick, M. A. Rieger, O. Stegle, A. Trumpp, Vitamin A-Retinoic Acid Signaling Regulates Hematopoietic Stem Cell Dormancy, Cell 169 (5) (2017) 807–823.e19. doi:10.1016/j.cell.2017.04.018.

[25] M. Mann, A. Mehta, C. G. de Boer, M. S. Kowalczyk, K. Lee, P. Haldeman, N. Rogel, A. R. Knecht, D. Farouq, A. Regev, D. Baltimore, Heterogeneous Responses of Hematopoietic Stem Cells to Inflammatory Stimuli Are Altered with Age, Cell Reports 25 (11) (2018) 2992–3005.e5. doi:10.1016/j.celrep.2018.11.056.

[26] A. E. Rodriguez-Fraticelli, S. L. Wolock, C. S. Weinreb, R. Panero, S. H. Patel, M. Jankovic, J. Sun, R. A. Calogero, A. M. Klein, F. D. Camargo, Clonal analysis of lineage fate in native haematopoiesis., Nature 553 (7687) (2018) 212–216. doi:10.1038/nature25168.

[27] L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv (2018). arXiv:1802.03426.

[28] D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Girón, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates, P. Flicek, Ensembl 2018, Nucleic Acids Research 46 (D1) (2018) D754–D761. doi:10.1093/nar/gkx1098.

[29] F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis, Genome Biology 19 (1) (2018) 15. doi:10.1186/s13059-017-1382-0.

[30] V. A. Traag, L. Waltman, N. J. van Eck, From louvain to leiden: guaranteeing well-connected communities, Scientific Reports 9 (1) (2019) 5233. doi:10.1038/s41598-019-41695-z.

[31] P. Brennecke, S. Anders, J. K. Kim, A. A. Kołodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni, M. G. Heisler, Accounting for technical noise in single-cell RNA-seq experiments, Nature Methods 10 (11) (2013) 1093–1095. doi:10.1038/nmeth.2645.

[32] A. Scialdone, K. N. Natarajan, L. R. Saraiva, V. Proserpio, S. A. Teichmann, O. Stegle, J. C. Marioni, F. Buettner, Computational assignment of cell-cycle stage from single-cell transcriptome data., Methods 85 (2015) 54–61. doi:10.1016/j.ymeth.2015.06.021.

[33] C. A. Vallejos, J. C. Marioni, S. Richardson, BASiCS: Bayesian Analysis of Single-Cell Sequencing Data, PLOS Computational Biology 11 (6) (2015) e1004333. doi:10.1371/journal.pcbi.1004333.

[34] A. T. L. Lun, K. Bach, J. C. Marioni, Pooling across cells to normalize single-cell RNA sequencing data with many zero counts, Genome Biology 17 (1) (2016) 75. doi:10.1186/s13059-016-0947-7.

[35] R. R. Coifman, S. Lafon, a. B. Lee, M. Maggioni, B. Nadler, F. Warner, S. W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps., Proc Natl Acad Sci U S A 102 (21) (2005) 7426–7431. doi:10.1073/pnas.0500896102.

[36] L. Haghverdi, F. Buettner, F. J. Theis, Diffusion maps for high-dimensional single-cell analysis of differentiation data, Bioinformatics 31 (18) (2015) 2989–2998.

[37] A. Wilson, E. Laurenti, G. Oser, R. C. van der Wath, W. Blanco-Bose, M. Jaworski, S. Offner, C. F. Dunant, L. Eshkind, E. Bockamp, P. Lió, H. R. MacDonald, A. Trumpp, Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair, Cell 135 (6) (2008) 1118 – 1129. doi:https://doi.org/10.1016/j.cell.2008.10.048.

[38] A. B. Balazs, A. J. Fabian, C. T. Esmon, R. C. Mulligan, Endothelial protein C receptor (CD201) explicitly identifies hematopoietic stem cells in murine bone marrow, Blood 107 (6) (2006) 2317–2321. doi:10.1182/blood-2005-06-2249.

[39] J. Y. Chen, M. Miyanishi, S. K. Wang, S. Yamazaki, R. Sinha, K. S. Kao, J. Seita, D. Sahoo, H. Nakauchi, I. L. Weissman, Hoxb5 marks long-term haematopoietic stem cells and reveals a homogenous perivascular niche., Nature 530 (7589) (2016) 223–7. doi:10.1038/nature16943.

[40] K. Nocka, J. C. Tan, E. Chiu, T. Y. Chu, P. Ray, P. Traktman, P. Besmer, Molecular bases of dominant negative and loss of function mutations at the murine c-kit/white spotting locus: W37, Wv, W41 and W., The EMBO journal 9 (6) (1990) 1805–13.

[41] Y. Sharma, C. M. Astle, D. E. Harrison, Heterozygous Kit Mutants with Little or No Apparent Anemia Exhibit Large Defects in Overall Hematopoietic Stem Cell Function, Experimental Hematology 35 (2) (2007) 214.e1–214.e9. doi:10.1016/J.EXPHEM.2006.10.001.

[42] L. Haghverdi, A. T. L. Lun, M. D. Morgan, J. C. Marioni, Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors, Nature Biotechnology 36 (5) (2018) 421–427. doi:10.1038/nbt.4091.

[43] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species, Nature Biotechnology 36 (5) (2018) 411–420. doi:10.1038/nbt.4096.