nature
cell biology

Human haematopoietic stem cell lineage commitment is a continuous process

Lars Velten[1,10], Simon F. Haas[2,3,4,10], Simon Raffel[2,4,5,10], Sandra Blaszkiewicz[2,3], Saiful Islam[6], Bianca P. Hennig[1], Christoph Hirche[2,3], Christoph Lutz[5], Eike C. Buss[5], Daniel Nowak[7], Tobias Boch[7], Wolf-Karsten Hofmann[7], Anthony D. Ho[5], Wolfgang Huber[1], Andreas Trumpp[2,4,8,11,12], Marieke A. G. Essers[2,3,11,12] and Lars M. Steinmetz[1,6,9,11,12]

Blood formation is believed to occur through stepwise progression of haematopoietic stem cells (HSCs) following a tree-like hierarchy of oligo-, bi- and unipotent progenitors. However, this model is based on the analysis of predefined flow-sorted cell populations. Here we integrated flow cytometric, transcriptomic and functional data at single-cell resolution to quantitatively map early differentiation of human HSCs towards lineage commitment. During homeostasis, individual HSCs gradually acquire lineage biases along multiple directions without passing through discrete hierarchically organized progenitor populations. Instead, unilineage-restricted cells emerge directly from a 'continuum of low-primed undifferentiated haematopoietic stem and progenitor cells' (CLOUD-HSPCs). Distinct gene expression modules operate in a combinatorial manner to control stemness, early lineage priming and the subsequent progression into all major branches of haematopoiesis. These data reveal a continuous landscape of human steady-state haematopoiesis downstream of HSCs and provide a basis for the understanding of haematopoietic malignancies.

All mature blood and immune cells are thought to derive from self-renewing and multipotent HSCs. According to the current model, initiation of differentiation is associated with the loss of self-renewal and generation of discrete multipotent, oligopotent and subsequently unipotent progenitor cell stages[1,2]. These lineage-restricted progenitors are thought to be generated in a stepwise manner by several subsequent binary branching decisions leading to the classical hierarchical tree-like model of haematopoiesis[1–6]. However, this model is mainly based on analyses of fluorescence-activated cell sorting (FACS)-purified cell populations. Even if followed up by single-cell assays[3,4,7], such analyses derive average properties of predefined cell populations and thereby miss both quantitative changes within gates as well as transition states falling between often subjectively set gates.

Moreover, the lineage contribution associated with each population is typically determined by assays such as colony formation or transplantation. While these assays read out lineage potential, the actual cell fate during homeostasis *in vivo* may be different[8,9]. Depending on the assays and markers used, partly conflicting branching points and hierarchies have been proposed[10–14].

Recent studies based on novel single-cell approaches have challenged more fundamental aspects of this classical model. For instance, unipotent progenitors can derive directly from HSCs without proceeding through oligopotent progenitors[14,15] and lineage commitment was observed in progenitors proposed to be oligopotent[7,10,16]. However, many of these studies focused on more differentiated compartments[7,10,16] or used predefined subpopulations to investigate single-cell heterogeneity[7,17], impeding the characterization of transitions between cell stages. Therefore, it remains unclear how individual HSCs enter lineage commitment during homeostasis *in vivo*. To establish a comprehensive model of haematopoiesis that can reconcile previous findings, a combined view of transcriptomic and functional changes along the developmental progression of individual cells is required. Here we developed an approach that integrates the reconstruction of developmental trajectories[18,19] with the quantitative linkage between transcriptomic and functional single-cell data[17] and thus provides a

[1]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, 69117 Heidelberg, Germany. [2]Heidelberg Institute for Stem Cell Technology and Experimental Medicine (HI-STEM gGmbH), 69120 Heidelberg, Germany. [3]Division of Stem Cells and Cancer, Haematopoietic Stem Cells and Stress Group, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. [4]Division of Stem Cells and Cancer and DKFZ-ZMBH Alliance, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. [5]Department of Internal Medicine V, University of Heidelberg, 69120 Heidelberg, Germany. [6]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA. [7]Department of Hematology and Oncology, Medical Faculty Mannheim, University of Heidelberg, 68167 Mannheim, Germany. [8]German Cancer Consortium (DKTK), 69120 Heidelberg, Germany. [9]Stanford Genome Technology Center, Palo Alto, California 94304, USA. [10]These authors contributed equally to this work. [11]These authors jointly supervised this work.
[12]Correspondence should be addressed to A.T., M.A.G.E. or L.M.S. (e-mail: a.trumpp@dkfz-heidelberg.de or m.essers@dkfz-heidelberg.de or larsms@embl.de)

Received 21 December 2016; accepted 20 February 2017; published online 20 March 2017; DOI: 10.1038/ncb3493

detailed view on lineage commitment of individual haematopoietic stem and progenitor cells (HSPCs) into all major branches of human haematopoiesis.

RESULTS

Healthy human bone marrow cells were labelled with a panel of up to 11 FACS surface markers commonly used to characterize human HSPCs[5,6] (see Methods and Supplementary Table 1). All HSPCs, defined by the absence of lineage markers (Supplementary Table 1) and expression of CD34 (Lin$^-$CD34$^+$), were individually sorted and enriched for immature cells (see Methods). The surface marker fluorescence intensities of all markers were recorded to retrospectively reconstruct immunophenotypes (CD10, CD38, CD45RA, CD90, CD135, and depending on the experiment CD2, CD7, CD49f, CD71, CD130, FCER1A, ITGA5 and KEL, Supplementary Fig. 1a). Such index-sorted HSPCs derived from the bone marrow of two healthy individuals were subjected to RNA-seq analysis ('index-omics', 1,034 and 379 single cells; see Supplementary Fig. 1b for the distribution of cells within classically defined gates[5,6] and Supplementary Fig. 2 for quality metrics of single-cell RNA-seq) to determine their transcriptomes or individually cultured *ex vivo* ('index-culture', 2,038 single cells) to quantify megakaryocytic, erythroid and myeloid lineage potential. Subsequently, the functional and transcriptomic data sets were integrated by regression models using commonly indexed surface marker expression to identify the molecular and cellular events associated with the differentiation of human HSCs at the single-cell level (Fig. 1). To make this data type accessible, we developed indeXplorer, a web-based platform that combines features of FACS software (for example, custom gating) with tools for single-cell transcriptomics data analysis (for example, differential expression analysis, clustering, principal component analysis) in a single graphical user interface (Supplementary Fig. 3 and http://steinmetzlab.embl.de/shiny/indexplorer/?launch=yes).

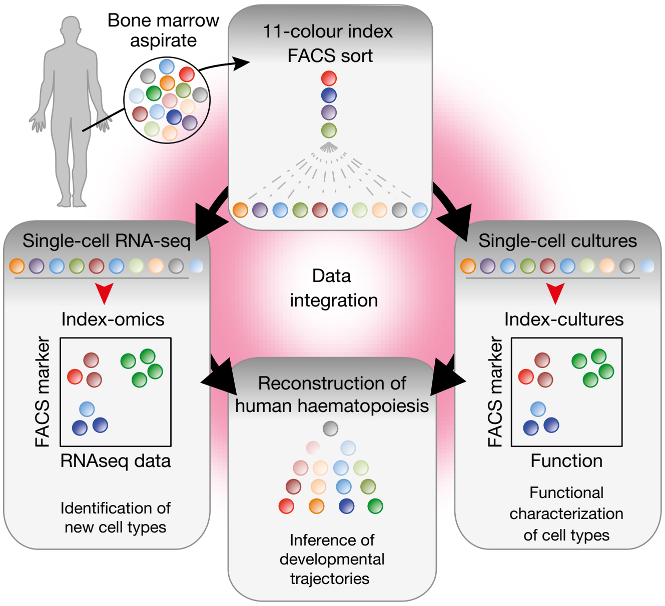

Bone marrow aspirate
11-colour index FACS sort
Single-cell RNA-seq
Index-omics
FACS marker
RNAseq data
Identification of new cell types
Data integration
Reconstruction of human haematopoiesis
Inference of developmental trajectories
Single-cell cultures
Index-cultures
FACS marker
Function
Functional characterization of cell types

Figure 1 Experimental strategy. Adult human HSPCs were stained with antibodies against up to 11 surface markers and individually sorted for either single-cell RNA-seq or single-cell cultures. Data from the two experiments were then integrated on the basis of surface marker expression to reconstruct developmental trajectories of haematopoiesis.

Early haematopoiesis is a continuous process

HSCs and their immediate progeny, such as multipotent progenitors (MPPs) or multilymphoid progenitors (MLPs), are located in the Lin$^-$CD34$^+$CD38$^-$ compartment, whereas more differentiated progenitors reside in the Lin$^-$CD34$^+$CD38$^+$ compartment[5,7]. Global gene expression analysis of single cells within these two compartments revealed fundamentally different transcriptomic structures. In both individuals, the Lin$^-$CD34$^+$CD38$^+$ progenitors could be separated into clusters corresponding to distinct progenitor cell types of all major branches of haematopoiesis (Fig. 2a and see below). In contrast, clustering within the Lin$^-$CD34$^+$CD38$^-$ compartment was largely unstable, as demonstrated by cluster stability analysis (Supplementary Fig. 4a), the absence of clusters according to Gap statistics (Supplementary Fig. 4b), and a recently published algorithm for the clustering of single cells[20] (Supplementary Fig. 4c). A simulated series of random steps from an individual cell to one of its nearest neighbours (see Methods) revealed that the majority of Lin$^-$CD34$^+$CD38$^-$ cells were highly interconnected, contrasting the disconnected cell types from the Lin$^-$CD34$^+$CD38$^+$ compartment (Fig. 2b). Unsupervised visualization of all individual cells irrespective of FACS markers by t-SNE confirmed that Lin$^-$CD34$^+$CD38$^-$ cells formed a single continuously connected entity. In contrast, Lin$^-$CD34$^+$CD38$^+$ cells emerged into locally clustered cell populations, with the exception of some phenotypic common myeloid progenitors (CMPs) and CD10$^+$ MLPs, suggesting that the classification based on differential CD38 expression is excellent, but not absolute (Fig. 2c).

Notably, the absence of hierarchical structures in the primitive Lin$^-$CD34$^+$CD38$^-$ compartment was due to the gradual nature of differences between cells in that compartment, and not due to insufficient data quality or a lack of transcriptomic heterogeneity: a principal component analysis of Lin$^-$CD34$^+$CD38$^-$ cells resolved more than 10 distinct, variable biological processes in this compartment, such as cell cycle activation and lineage priming (Supplementary Fig. 4d–f). These processes are tightly correlated to surface marker expression (Supplementary Fig. 4g).

Collectively, these observations are incompatible with the classical model of early haematopoiesis, which assumes a hierarchical tree-like structure of discrete progenitors downstream of HSCs. In contrast, our data suggest that HSCs and their immediate progeny are initially part of a continuum of low-primed undifferentiated ('CLOUD')-HSPCs within the Lin$^-$CD34$^+$CD38$^-$ compartment (see also below). Discrete populations are established only when differentiation has progressed to the level of restricted progenitors typically associated with the upregulation of CD38.

Lineage-restriction downstream of the HSPC continuum

To characterize the discrete populations in the Lin$^-$CD34$^+$CD38$^+$ compartment, we performed gene expression and cell surface marker analyses as well as functional validations at the single-cell level. Our analyses revealed that these populations correspond to lineage-restricted progenitors of all major branches of bone marrow haematopoiesis, including B-cell progenitors of distinct stages, megakaryocyte/erythrocyte committed progenitors (ME, Ery, Mk),

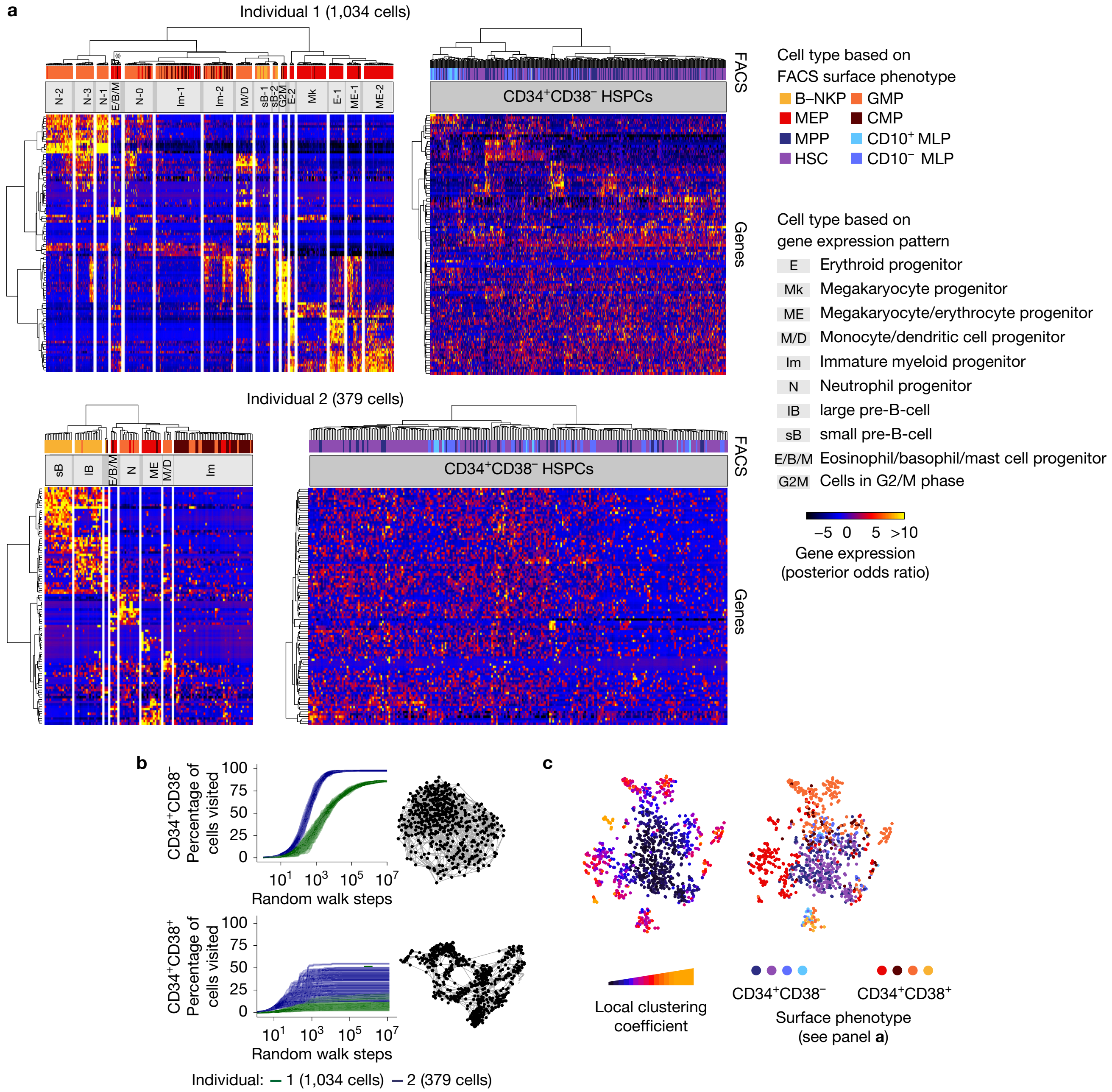


Figure 2 A stem and progenitor cell continuum precedes the establishment of discrete lineages at the CD34$^+$CD38$^+$ stage. (**a**) Hierarchical clustering of Lin$^-$CD34$^+$CD38$^-$ (individual 1: 467 cells, individual 2: 261 cells) and Lin$^-$CD34$^+$CD38$^+$ (individual1: 567 cells, individual 2: 118 cells) compartments for both individuals. Clustering was performed on the most variable 1,000 genes of each population. The most variable 100 genes are displayed in the heatmap. The asterisk indicates that 3 putative eosinophil/basophil/mast cell progenitor subclusters of <5 cells were merged. Cells labelled G2M showed high expression of genes indicative for G2/M phase of the cell cycle and likely clustered together based on their cell cycle state rather than cell-type-specific gene expression. (**b**) Random walk analysis of Lin$^-$CD34$^+$CD38$^-$ and Lin$^-$CD34$^+$CD38$^+$ compartments for both individuals. One hundred random walks, that is, series of random steps from one cell to any of its five nearest neighbours in correlation distance space, were simulated and the number of cells reached was evaluated in relation to the total number of cells. Five-nearest-neighbour networks are depicted on the right. (**c**) t-SNE visualization of all cells (individual 1) highlighting the degree to which cells are associated with local clusters (left panel, see also Methods) and the immunophenotype (right panel).

neutrophil-primed progenitors (Neutro), monocyte/dendritic cell (Mono/DC) progenitors, and eosinophil/basophil/mast cell progenitors (Eo/Baso/Mast), as well as immature myeloid progenitors (Fig. 3a and Supplementary Table 2). Importantly, populations cluster by cell type and not by individual in a cross-individual comparison (Fig. 3b). The comparison of the surface marker expression of these populations to the commonly applied gating scheme[5] using our indexed data set showed that immunophenotypically defined oligopotent progenitor populations (megakaryocyte–erythroid progenitors, MEPs; granulocyte–monocyte progenitors, GMPs; B-cell–NK-cell

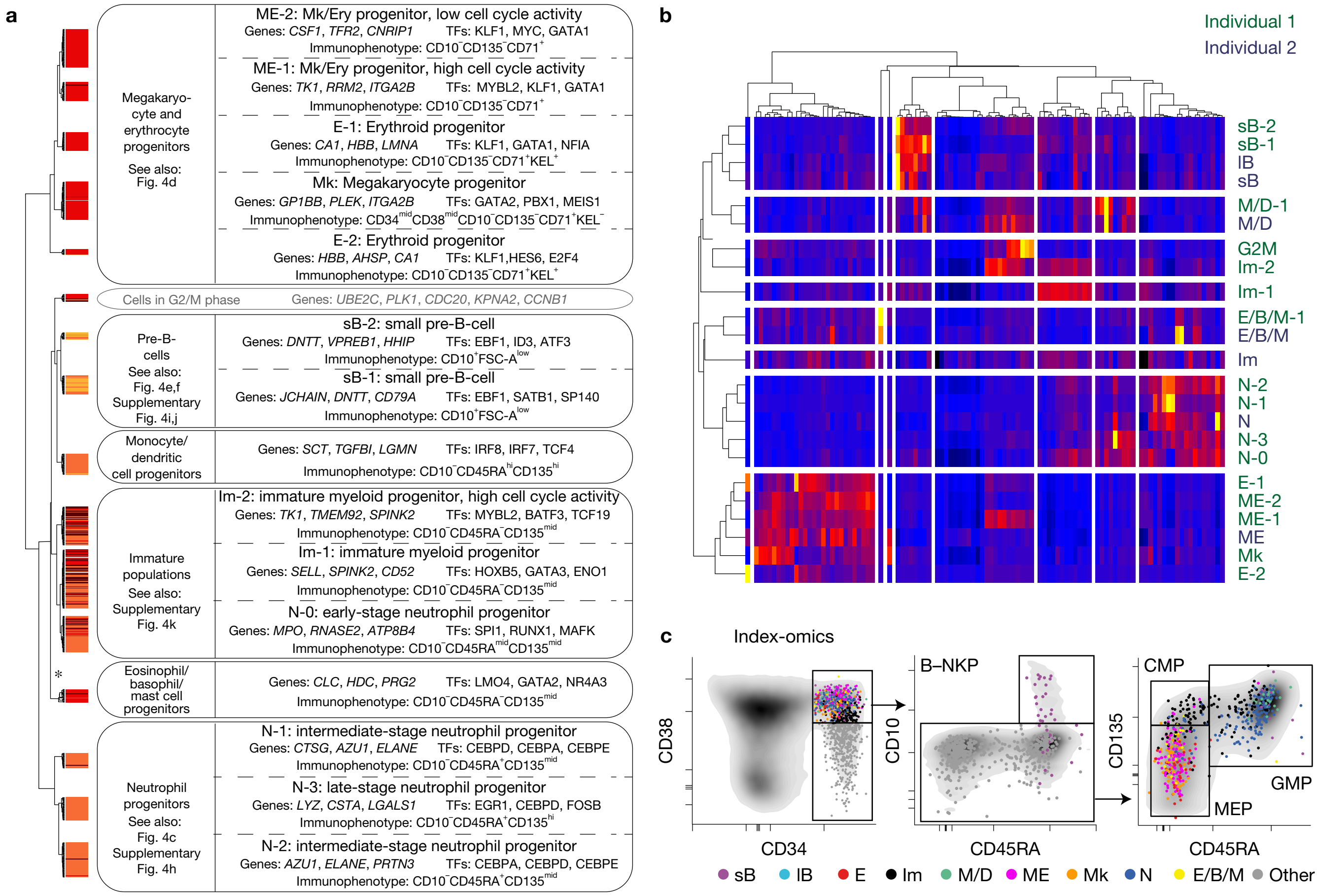


Figure 3 The Lin$^-$CD34$^+$CD38$^+$ compartment consists of distinct lineage-restricted progenitors. (**a**) Overview of putative cell types in individual 1 (see panel **b** for a comparison between individuals). Classes obtained from hierarchical clustering of the Lin$^-$CD34$^+$CD38$^+$ compartment (Fig. 2a) were assigned to putative cell types based on analyses of gene and surface marker expression. The asterisk indicates that 3 putative eosinophil/basophil/mast cell progenitor subclusters of <5 cells were merged for this analyses. TFs, transcription factors. (**b**) Averaged gene expression profiles for cell types from both individuals defined in Fig. 2a were clustered on the basis of the 1,000 most variable genes. Only the most variable 100 genes are shown in the heatmap. (**c**) Index-omics display of Lin$^-$CD34$^+$CD38$^+$ progenitors. Sequenced single Lin$^-$CD34$^+$CD38$^+$ cells were arranged according to their cell surface marker expression in classical FACS gating strategies to identify B- and NK-cell progenitors (B–NKPs), megakaryocytic–erythroid progenitors (MEPs), common myeloid progenitors (CMPs) and granulocyte–monocyte progenitors (GMPs). Cells were colour-coded on the basis of their cell type identity from Fig. 3a.

progenitors, B–NKPs) were mainly comprised of cell types with unilineage-specific gene expression profiles (Fig. 3c) and functional unipotency (Fig. 4a,b).

Cells within the classic GMP compartment were separated into several neutrophil-primed progenitors (N-0 to N-3), as well as into monocyte/dendritic cell progenitors (Mono/DC). The distinct neutrophil-primed progenitors probably represent progenitors at different developmental stages and granule composition (Fig. 4c and Supplementary Fig. 4h)[21,22]. Immunophenotypically, all neutrophil-primed progenitors express the surface markers CD135 and CD45RA, which are progressively upregulated during maturation (Fig. 4c). In contrast to neutrophil-primed progenitors, Eo/Baso/Mast progenitors did not fall into the classical GMP gate but displayed a Lin$^-$CD34$^+$CD38$^+$CD10$^-$CD45RA$^-$CD135$^{\text{mid}}$ immunophenotype (Fig. 3c), and expressed transcription factors important for early MEP commitment (GATA2 and TAL1) supporting a recent study suggesting that granulocyte subtypes might derive from distinct haematopoietic lineages[12].

The MEP gate consisted of megakaryocytic (Mk) progenitors expressing typical Mk genes, of erythroid-committed (E-1, E-2) progenitors of distinct developmental stages, differing in haemoglobin and GATA1 expression, as well as of subpopulations showing combined expression of megakaryocytic and erythroid genes (M/E). Our single-cell transcriptome data suggested CD71 (TRFC) and the red blood cell antigen KEL to be highly indicative for erythroid fate, which was confirmed by single-cell culture assays using CD71 and KEL as indexing antibodies (Fig. 4d).

For individual 2, two CD10$^+$ B-cell progenitor clusters (small pre-B-cells, sB and large pre-B-cells, lB) were observed. sB was characterized by high CD9 messenger RNA expression, high CD10 surface expression and small cell size (forward scatter (FSC)), whereas lB showed high expression of interleukin-7 receptor (IL7RA) mRNA, intermediate CD10 surface levels, expression of cell-cycle-related genes and large cell size (Fig. 4e and Supplementary Fig. 4i and Supplementary Table 2). This suggests that sB corresponds to small pre-B-cells, and lB to large pre-B-cells, progenitor populations that

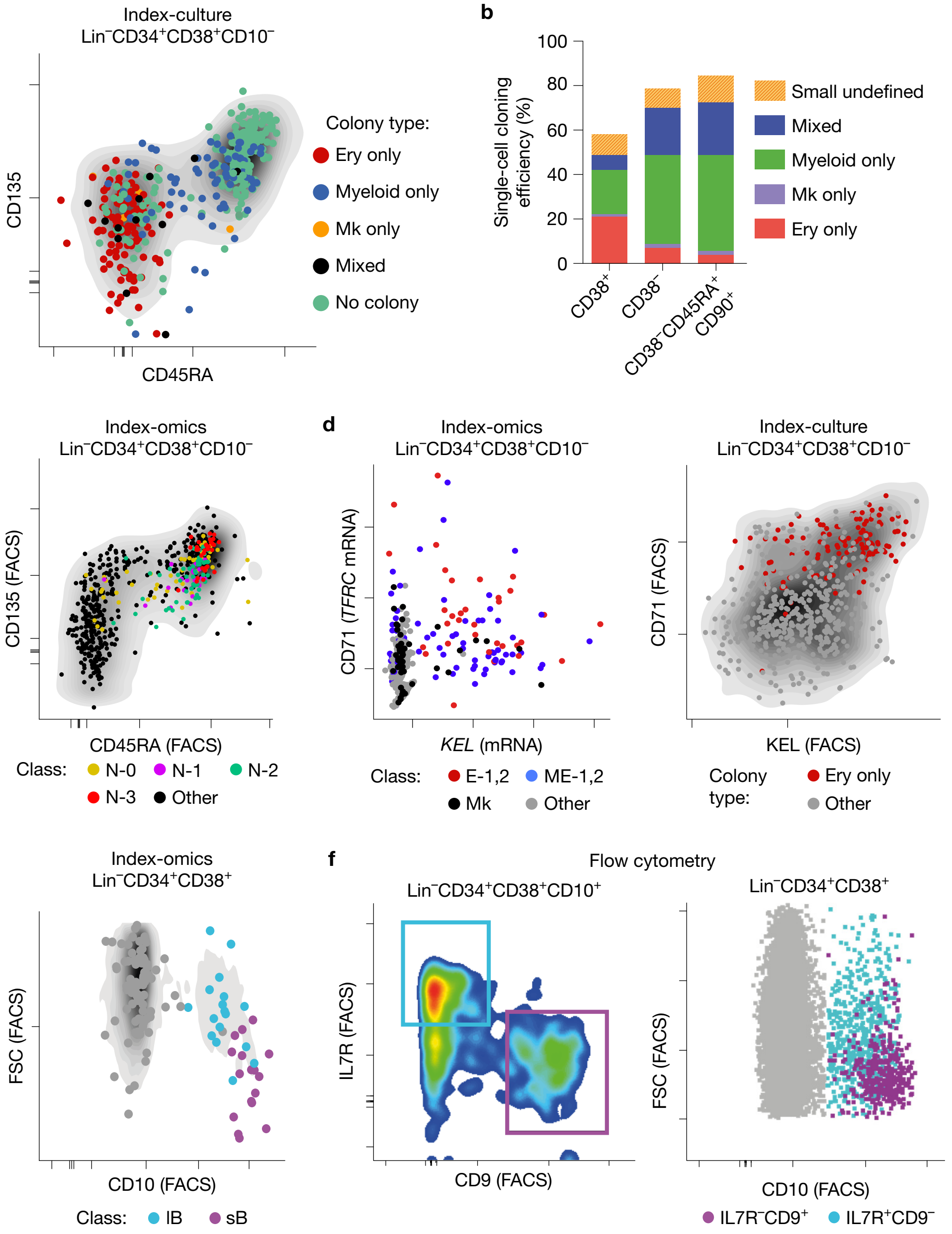


Figure 4 Characterization of Lin$^-$CD34$^+$CD38$^+$ lineage-restricted progenitors. (**a**) Index-culture display of Lin$^-$CD34$^+$CD38$^+$CD10$^-$ HSPCs. Single HSPCs were cultured for 3 weeks and the resulting colony type was plotted in relation to CD45RA and CD135. (**b**) Single cells from the *ex vivo* culture assay were scored as unipotent (gave rise to one lineage) or mixed (gave rise to more than one lineage). (**c**) Neutrophil-primed subpopulations in relation to CD45RA and CD135 surface marker expression. (**d**) Megakaryocytic/erythroid-primed subpopulations in relation to *TFRC* (CD71) mRNA and *KEL* mRNA expression (left panel) and erythroid colony output in relation to CD71 and KEL surface marker expression (right panel). (**e**) Pre-B-cell subpopulations from individual 2 in relation to CD10 surface expression and forward scatter (FSC). (**f**) Prospective isolation of B-cell subpopulations sB and IB using classical flow cytometry. FACS markers for IL7R and CD9 permit the separation of two populations with FSC/CD10 profiles corresponding to sB and IB, as suggested from gene expression data.

have been well characterized in the murine system, but to a lesser extent in the human system[23]. To validate and prospectively isolate large pre-B-cells and small pre-B-cells, we used IL7RA and CD9 FACS markers, which allowed us to recapitulate the levels of CD10 surface expression, cell size and cell cycle activity as predicted from the index-omics data (Fig. 4f and Supplementary Fig. 4j). In contrast to individual 2, in individual 1, only small pre-B-cells were observed (Fig. 3b).

For both individuals, we also observed CD38-positive HSPCs with a gene expression profile of rather immature cells (Im) (Fig. 3a). These clustered globally with the Lin$^-$CD34$^+$CD38$^-$ compartment in t-SNE analyses, and expressed lower levels of CD38 (Supplementary Fig. 4k). Most of these cells displayed an immunophenotype typical for CMPs (Lin$^-$CD34$^+$CD38$^+$CD45RA$^-$CD135$^+$); however, the composition of the cell types present in the CMP gate depends strongly on the exact gating strategy applied (see below, Supplementary Fig. 5h, i).

On the basis of these analyses, we provide markers and gating strategies for the prospective isolation of several of these newly defined populations using standard flow cytometry (Figs 3 and 4).

Developmental trajectories of early human haematopoiesis

To obtain a detailed view on the transition from stem cells to lineage-restricted progenitors in the continuous HSPC landscape, we developed STEMNET, a new dimensionality reduction algorithm. STEMNET identifies genes specific to the six $\mathrm{Lin}^{-}\mathrm{CD34}^{+}\mathrm{CD38}^{+}$ restricted progenitor populations defined above (Neutro, Eo/Baso/Mast, B-cell, Mono/DC, Ery and Mk; see Supplementary Table 3 for a list of genes used by STEMNET) and then computes the probability that each primitive ('CLOUD') HSPC can be assigned to any of these classes. STEMNET thereby places the six developmental endpoints on the corners of a simplex. This resulted in the arrangement of the least-primed HSCs, such as $\mathrm{CD49f}^{+}$ HSCs, to the centre, and the remaining HSPCs localizing in between according to their degree of priming (Fig. 5a, and see Supplementary Fig. 5a,b for individual 2). To describe the position of each cell we computed the predominant direction of priming d as the developmental endpoint closest to the cell and the degree of lineage priming S^{rel} as the (Kullback–Leibler) distance from the least-primed cell.

This analysis suggests that HSCs located in the centre of the 'CLOUD' gradually acquired continuous lineage priming into either of the major branches. While lympho/myeloid and megakaryocytic/erythroid priming formed major points of attraction, a clear separation into single lineages was not present at this stage (Fig. 5a). In contrast, lineages were clearly separated at the level of $\mathrm{Lin}^{-}\mathrm{CD34}^{+}\mathrm{CD38}^{+}$ progenitors, without further sub-branching in this compartment (Fig. 5a, see Supplementary Fig. 5c for CD38 expression). Importantly, these results are not due to limitations of the bioinformatics method, as STEMNET is able to detect both subsequent branching points and discrete intermediate populations on simulated data (Supplementary Fig. 6a–d). Moreover, applying diffusion pseudotime (DPT), a different recently published method for the inference of developmental trajectories[24] to our data confirmed the absence of subsequent binary branch points and the direct lineage commitment from CLOUD-HSPCs along continuous trajectories (Supplementary Fig. 6e).

Within the differentiation continuum, STEMNET analysis located previously defined immunophenotypic populations according to their known lineage potential[5] (Fig. 5b, see Supplementary Fig. 5b for individual 2). For example, GMPs were distributed to the neutrophil and monocytic/dendritic cell branches while MEPs were located to the megakaryocytic and erythroid branches (notice that the localization of CMPs critically depends on the exact CD38 and CD135 gating strategy, Supplementary Fig. 5h,i). In contrast, immunophenotypic MLPs were located close to the separation of lymphoid, neutrophil and monocytic/dendritic cell lineages (Fig. 5b and Supplementary Fig. 5b), with individual cells already primed towards specific lineages, in line with frequent functional commitment to single lineages in mouse LMPPs[15]. Together, these analyses suggest that developmental stages immediately downstream of HSCs such as MLPs and MPPs do not represent discrete cell types located at defined branching points, but should rather be considered as transitory states within the HSPC continuum with higher probability for commitment to particular lineages.

While undergoing lineage commitment only very few cells acquired a transcriptomic state of dual-lineage priming (Supplementary Fig. 5d,e), in accordance with a recent single-cell transcriptomic study on mouse GMPs[20]. However, our analyses suggest that a direct transition from a primed multi-lineage towards a unilineage transcriptomic state represents the main route of lineage commitment, whereas dual-lineage states (such as $\mathrm{Gfi1}^{+}\mathrm{Irf8}^{+}$ GMPs, Supplementary Fig. 5f) exist, but represent rare exceptions. Importantly, both transcriptomic and functional (Supplementary Fig. 5g) lineage combinations of bipotent cells were not restricted to the combinations predicted by the classical model, conflicting with a strictly ordered hierarchy of branching events. Along these lines, co-expression of opposing pairs of transcription factors, such as IRF8 and PU.1 (SPI1) that have been thought to establish an oligopotent state, occurred at much lower frequency than previously expected (see Fig. 8a(viii,xi))[25].

Transcriptomic priming mediates lineage commitment

Single-cell RNA-seq protocols require cell lysis and therefore prohibit subsequent functional interrogation of the same single cell. However, the use of indexed FACS surface markers common to both single-cell *ex vivo* culture data and single-cell RNA-seq data allowed us to quantitatively link the amount and direction of transcriptomic priming to functional properties such as lineage potential and proliferative capacity. For example, the STEMNET-predicted dominant direction of transcriptional priming into the lympho/myeloid versus the megakaryocytic/erythroid direction was strongly correlated to the surface marker expression of CD135 and CD45RA (Fig. 6a(i,ii)), which could be used to qualitatively predict the predominant cell type in colonies of our single-cell cultures (note that lymphoid progenitors do not grow in these conditions, and that myeloid sublineages are not resolved) (Fig. 6a(ii)). Utilizing all recorded surface markers for linear models on the single-cell RNA-seq data allowed us to quantitatively predict the dominant cell type present in the single-cell cultures for the $\mathrm{Lin}^{-}\mathrm{CD34}^{+}\mathrm{CD38}^{+}$ ($P = 3.7 \times 10^{-23}$) and the $\mathrm{Lin}^{-}\mathrm{CD34}^{+}\mathrm{CD38}^{-}$ compartment ($P = 3.7 \times 10^{-22}$, Fig. 6a(iii) and Supplementary Fig. 7a for the full specification of regression models). Moreover, predicting erythroid and megakaryocytic priming individually revealed that the amount of lineage-specific priming was linked to functional lineage commitment (Fig. 6b,c and Supplementary Fig. 7b,c). However, colonies derived from Mk-primed cells were frequently dominated by other cell types due to their lower proliferative capacity *ex vivo* (Supplementary Fig. 7b). STEMNET further predicted $\mathrm{Lin}^{-}\mathrm{CD34}^{+}\mathrm{CD38}^{-}\mathrm{CD45RA}^{-}\mathrm{CD90}^{-}\mathrm{CD135}^{-}$ cells to be primed towards megakaryocytic differentiation (Fig. 6d, left panel). To functionally validate this prediction *in vivo*, we FACS-sorted these cells, transplanted them into sublethally irradiated NSG mice and quantified their lineage output 14 days post transplantation. As predicted, these cells, which we termed Mk-primed MPPs, predominantly generated thrombocytes if compared with MLPs and HSCs (Fig. 6d, right panel). Together, these analyses revealed that transcriptomic priming is linked to the restriction of lineage potential at an early stage *in vitro* and *in vivo*.

We next estimated the degree of transcriptomic lineage priming S^{rel} for individual cells from the culture experiments (Fig. 7a,b). As expected, committed progenitors with a high degree of inferred

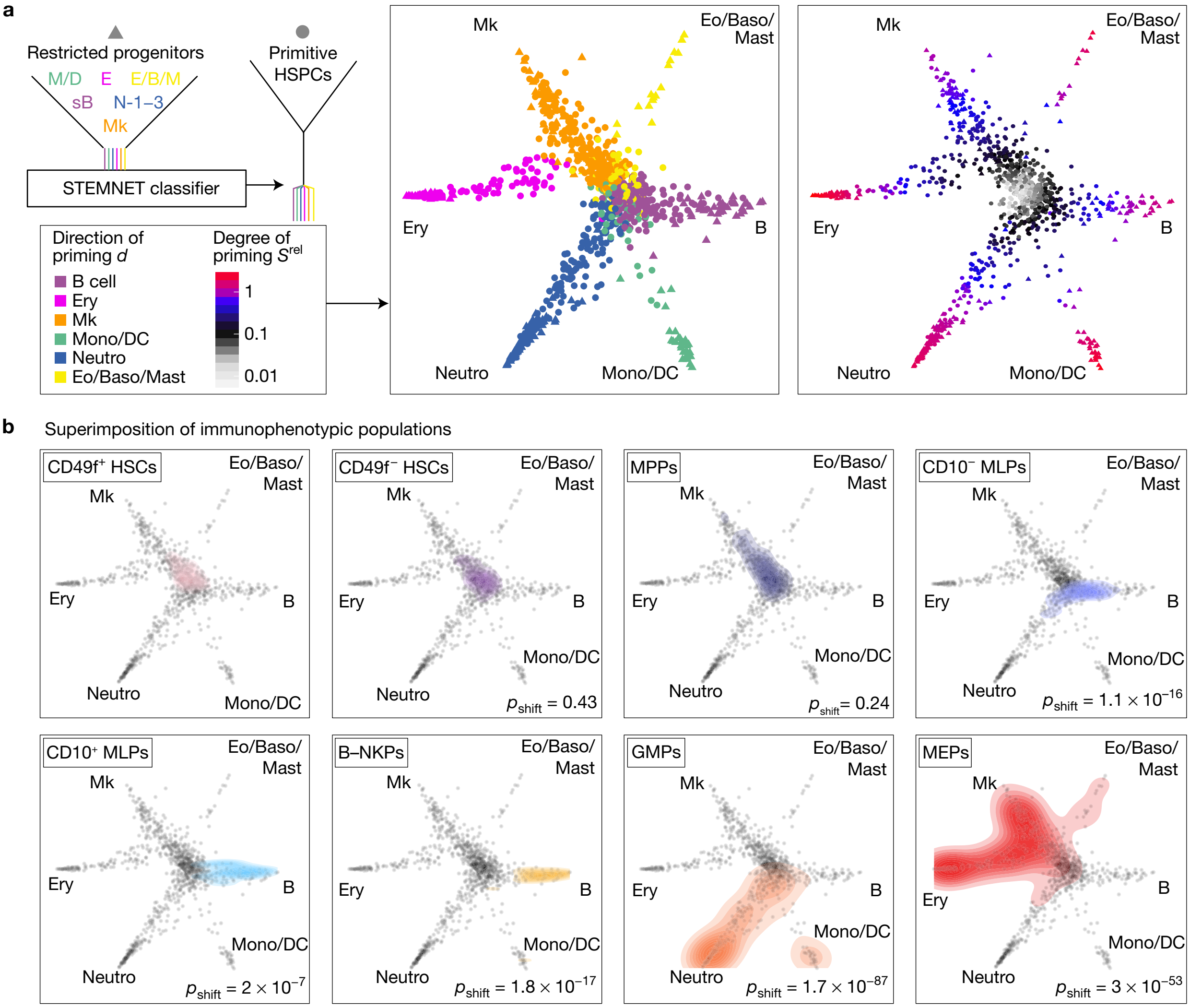


Figure 5 Visualization of the HSPC continuum. (**a**) The similarity of every cell to each of the progenitor classes was computed by STEMNET (see Methods), projected on a unit circle, and used to quantify the degree and direction of transcriptomic priming. Data from individual 1 are shown; for individual 2 see Supplementary Fig. 5a,b. (**b**) Immunophenotypic populations[5,6] were highlighted on the HSPC continuum. p_{shift} indicates P values calculated by kernel-density-based tests comparing each population with CD49f$^+$ HSCs. For CMPs, see Supplementary Fig. 5h,i. For CD49f$^+$ HSCs, $n=101$ single cells; CD49f$^-$HSCs, $n=117$; MPPs, $n=176$; CD10-MLPs, $n=52$; CD10$^+$MLPs, $n=16$; B–NKPs, $n=26$; GMPs, $n=244$; MEPs, $n=231$.

transcriptomic lineage priming formed small colonies (Fig. 7a) of a single-cell type (Fig. 7b). In contrast, primitive HSPCs (low inferred S^{rel}) frequently displayed multi- or bilineage potential (Fig. 7b) and generated much larger colonies (Fig. 7a). However, not all of the primitive HSPCs displayed multipotency, but frequently appeared to be lineage-restricted while typically retaining a high proliferative capacity comparable to their multipotent counterparts (Fig. 7c). These data suggest that proliferative capacity and lineage potency are not obligatorily linked.

To investigate the ability of cells with various amounts of priming to switch lineage potential, we cultured HSPCs in the absence and presence of erythropoietin (EPO). Progenitors that formed exclusively erythroid colonies in the presence of EPO were unable to give rise to alternative lineages in the absence of EPO (Fig. 7d). Moreover, we cultured single HSPCs for one week, split the colonies into four and determined the lineage outcome of the daughter colonies two weeks later. In line with the predictions of our model, the degree of transcriptomic priming was anticorrelated to the propensity of cells to generate daughters with variable lineage composition (Supplementary Fig. 7d,e). Together, these results support the hypothesis that early lineage priming of primitive HSPCs coincides with a loss of functional plasticity.

Molecular processes underlying HSC commitment

To characterize stemness, early lineage priming and transcriptional cell type manifestation on the molecular level, we identified co-expressed gene modules whose activities were associated with the direction and/or the degree of priming. We visualized the activity of these gene modules on the differentiation landscape established above (Fig. 8a(i)) and along the progression from HSCs to each of the six lineages (Fig. 8b and Supplementary Fig. 8a,b and Supplementary Table 4 for a complete list). Importantly, data from both individuals

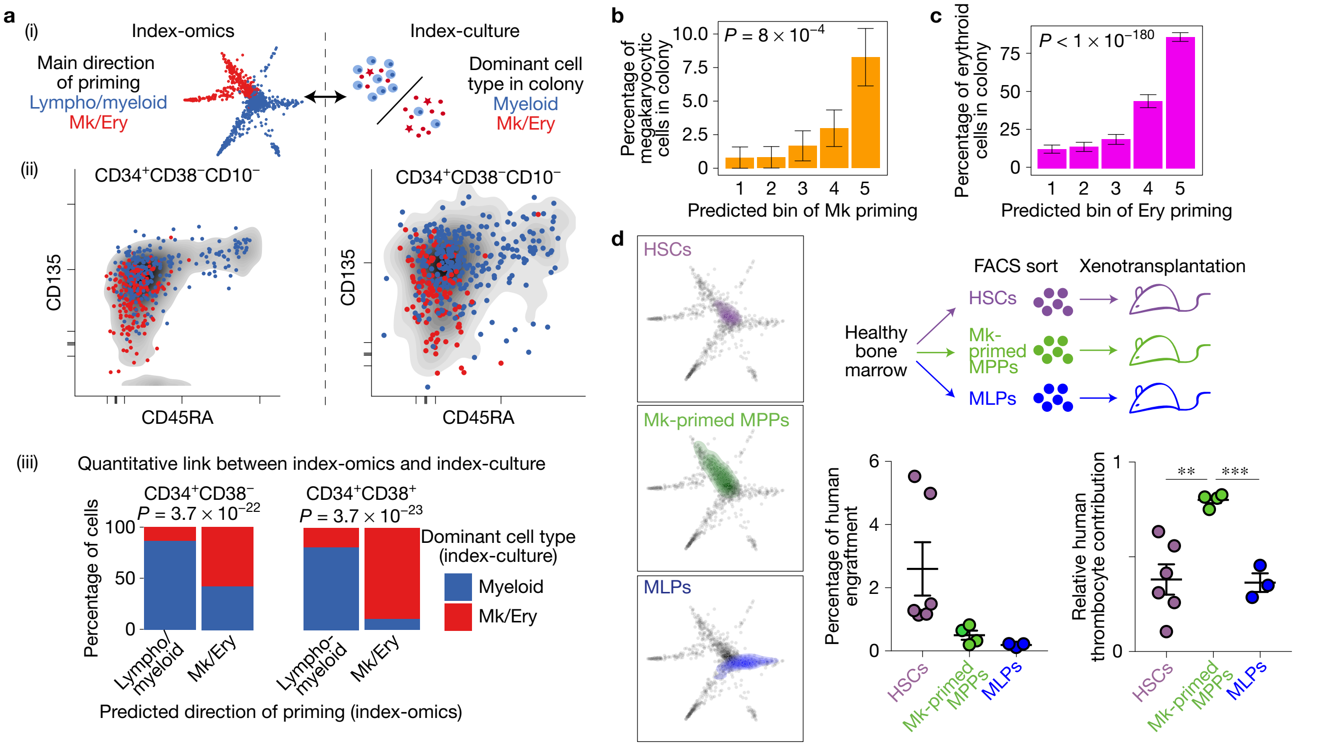


Figure 6 The direction of transcriptomic priming is quantitatively linked to functional lineage potential. (**a**) Comparison of the predominant direction of priming *d* (lympho/myeloid versus megakaryocyte/erythroid) obtained from single-cell transcriptomics to the dominant cell type observed in colonies from single-cell culture. (i) Illustration. (ii) Qualitative comparison of the two quantities with respect to CD45RA and CD135 surface marker expression. (iii) Quantitative link. The most likely dominant direction of priming was estimated for each founder cell from index-culture based on regression models constructed on all surface markers and compared with the observed colony composition (see Supplementary Fig. 7a). *P* values are from a Fisher test with $n = 434$ cells (left panel) and $n = 193$ cells (right panel). (**b**) Comparison between inferred amount of transcriptomic Mk priming and the percentage of CD41$^+$ Mk cells per colony. Errors bars denote s.e.m. *P* value is from a Pearson product moment correlation test with $n = 627$ single cells that formed colonies. See also Supplementary Fig. 7c. (**c**) Comparison between inferred amount of transcriptomic erythroid priming and the percentage of CD235$^+$ erythroid cells per colony. See also Supplementary Fig. 7c. Errors bars denote s.e.m. *P* value is from a Pearson product moment correlation test with $n = 627$ single cells that formed colonies. (**d**) Xenotransplantation validating a Mk-primed MPP population identified by STEMNET. HSCs, MLPs and a population of putatively Mk-primed MPPs (Lin$^-$CD34$^+$CD38$^-$CD45RA$^-$CD90$^-$CD135$^-$) were sorted, transplanted into immunocompromised mice and chimaerism of human lympho/myeloid cells (CD45$^+$), thrombocytes and erythrocytes were determined 2 weeks post transplantation. Experimental set-up (top right panel), localization of populations in STEMNET (left panels), and human engraftment (bottom right panels, error bars denote s.e.m.) are indicated. Relative contribution of thrombocytes was significantly higher in Mk-primed MPPs compared with HSCs ($P = 0.0031$) and MLPs ($P = 0.0002$, two-tailed unpaired *t*-test, $n = 6$ HSCs, $n = 4$ Mk-primed MPPs, $n = 3$ MLPs). Asterisks indicate level of significance as follows: $^{**}P < 0.01$; $^{***}P < 0.001$.

yielded highly comparable results (Supplementary Fig. 8). To gain additional information about biological processes associated with HSC differentiation, we determined the mean expression of genes for each gene ontology (GO) term, and selected representative examples that changed significantly during early lineage priming (Fig. 8c). Together, these analyses provide insights into the global molecular and cell biological processes HSCs encounter while undergoing continuous lineage priming, unilineage commitment and subsequent differentiation.

The least-primed state was characterized by expression of the *HOXA3/PRDM16/HOXB6* module[26–28] (Fig. 8a(ii),b and Supplementary Table 4) and associated with typical stem cell properties such as cell cycle quiescence, low expression of the entire gene expression machinery, low total RNA content (measured by mRNA reads per *in vitro* spike in RNA read), low cellular respiration[29], low CD38 and high CD90 surface expression[5] (Fig. 8c). The expression of the *HLF/ZFP36L2* module (which also contains the transcription factors *MECOM/EVI1*, *HFL*, *GATA3*) was highest in immature HSCs, but present in the entire 'CLOUD' (Fig. 8a(iii),b and Supplementary Table 4)[30–32].

Intriguingly, stem cells also expressed genes from the earliest priming modules from both the lympho/myeloid (*FLT3/SATB1* module) and the megakaryocyte/erythrocyte (*GATA2/NFE2* module)[33] lineages in a non-exclusive manner (Fig. 8a(iv–v)). These data suggest that the first transcriptional priming events into the predominantly lympho/myeloid or the megakaryocyte/erythrocyte direction are already present in most primitive HSCs, coinciding with the occurrence of first functional lineage biases already at this stage (Figs 6a,b, 7a S^{rel} bin 1 and 2). A number of additional gene modules were activated in a combinatorial fashion between lineages, similar to previous observations from bulk RNA-seq[34] (Fig. 8 and Supplementary Fig. 8a and Supplementary Table 4).

Following acquisition of lineage priming, HSCs upregulate their gene expression machinery, mRNA and protein biosynthesis, and

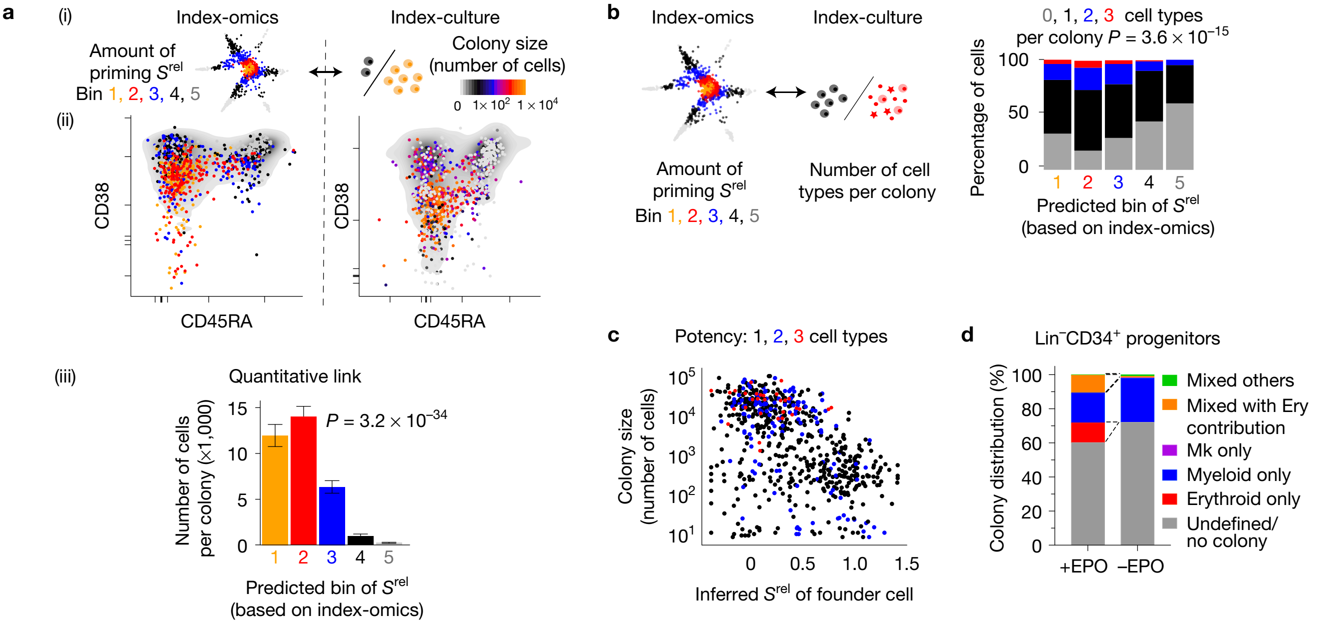


Figure 7 The degree of transcriptomic priming is quantitatively linked to multipotency and proliferative capacity. (**a**) Comparison between the inferred amount of transcriptomic priming S^{rel} of the founder cell and the resulting colony size (cell number). (i) Illustration. (ii) Qualitative link. (iii) Quantitative link. Errors bars denote s.e.m. P value is from a Pearson product moment correlation test with $n = 1{,}031$ single cells. (**b**) Comparison between the inferred amount of priming S^{rel} of the founder cell and the number of cell types in the colony. P value is from a Pearson product moment correlation test with $n = 1{,}031$ single cells. (**c**) Inferred transcriptomic degree of priming S^{rel} (x axis) in relation to the colony size (y axis) and the number of cell types per colony (colour code). (**d**) Distribution of colony types in relation to the presence or absence of erythropoietin (EPO) in the culture medium.

respiration[29,35], while cell cycle activity increases only marginally (Fig. 8c). At this stage, cells start to express lineage-specific gene modules, for example the *SPI1*/*GFI1* module for the neutrophil lineage (Fig. 8a(viii)) or the *IRF1*/*CASP1* module[33] for the B-cell lineage (Fig. 8a(vi)). Other modules active at this stage, however, are shared between lineages; for example, the *TAL1*/*HFS1* module is shared between the erythroid and the megakaryocytic lineage, whereas the *EAF2*/*KLF4* module is shared between the neutrophil and the monocyte lineage. This coincides with the observation that most progenitors at this stage display narrow restriction in their developmental potential, whereas some progenitor cells remain oligopotent[15] (Fig. 7b, S^{rel} bin 3).

Manifestation of lineage-specific differentiation is accomplished by activation of gene modules such as the *CEBPA*/*CEPBD* module for the neutrophil lineage, the *EBF1*/*ID3* module for the B-cell lineage, the *IRF8* module for the monocytic/dendritic lineage, the *GPI1BB*/*PBX1* module for the megakaryocytic lineage and the *GATA1*/*KLF1* module for the erythroid lineage[33,36,37] (Fig. 8a(x–xv),b). In all cases, this step is accompanied by cell cycle activation, CD38 surface marker upregulation (Fig. 8c) and unipotency (Fig. 7b, S^{rel} bin 4 and 5).

Together, our data suggest that HSCs are characterized by the expression of specific stem cell modules in combination with early, probably antagonizing priming modules. During the continuous priming and differentiation process the stem cell modules and certain (but not all) early priming modules already expressed in HSCs are turned off, while specific lineage modules become reinforced to drive differentiation towards lineage commitment and manifestation (Fig. 8a,b). Transcription factors from upstream modules may trigger expression of downstream modules, as in the case of *GATA2*, *TAL1* and *GATA1*[33]. In contrast, transcription factors from mutually exclusive downstream modules may inhibit each other; for example, *IRF8* is known to repress *CEBPA*[38]. Such inhibitory interactions may render oligopotent progenitors unstable[7,10,15], and thus less abundant than previously anticipated (Fig. 7b). In contrast, in cells with a low amount of priming, expression levels of mutually exclusive modules are sufficiently small to allow uni-, oligo- or multipotency.

DISCUSSION

In summary, we provide a global view of the early human haematopoiesis during homeostasis. Our data set combines both information on the lineage potential of HSCs (index-culture) and insights into the unperturbed lineage commitment of HSCs during human haematopoiesis (reconstruction of developmental trajectories from static single-cell expression data), where lineage tracing approaches[8,9] are not possible. Here, we rely on single-cell culture data and xenotransplantation for functional validation, which unlike gene expression or cellular barcoding measure developmental potential, not fate.

Our results are incompatible with fundamental aspects of the differentiation-tree model, in which HSCs are required to pass through discrete and definable intermediate progenitor cell stages by subsequent binary cell fate decisions made on branching points. Instead, we propose that early haematopoiesis is represented by a cellular continuum of low-primed undifferentiated (CLOUD)-HSPCs. This HSPC continuum contains phenotypic MPPs and MLPs, which do not constitute discrete progenitor cell types, but rather transitory states. CLOUD-HSPCs gradually acquire transcriptomic lineage priming in a combination of multiple directions, with

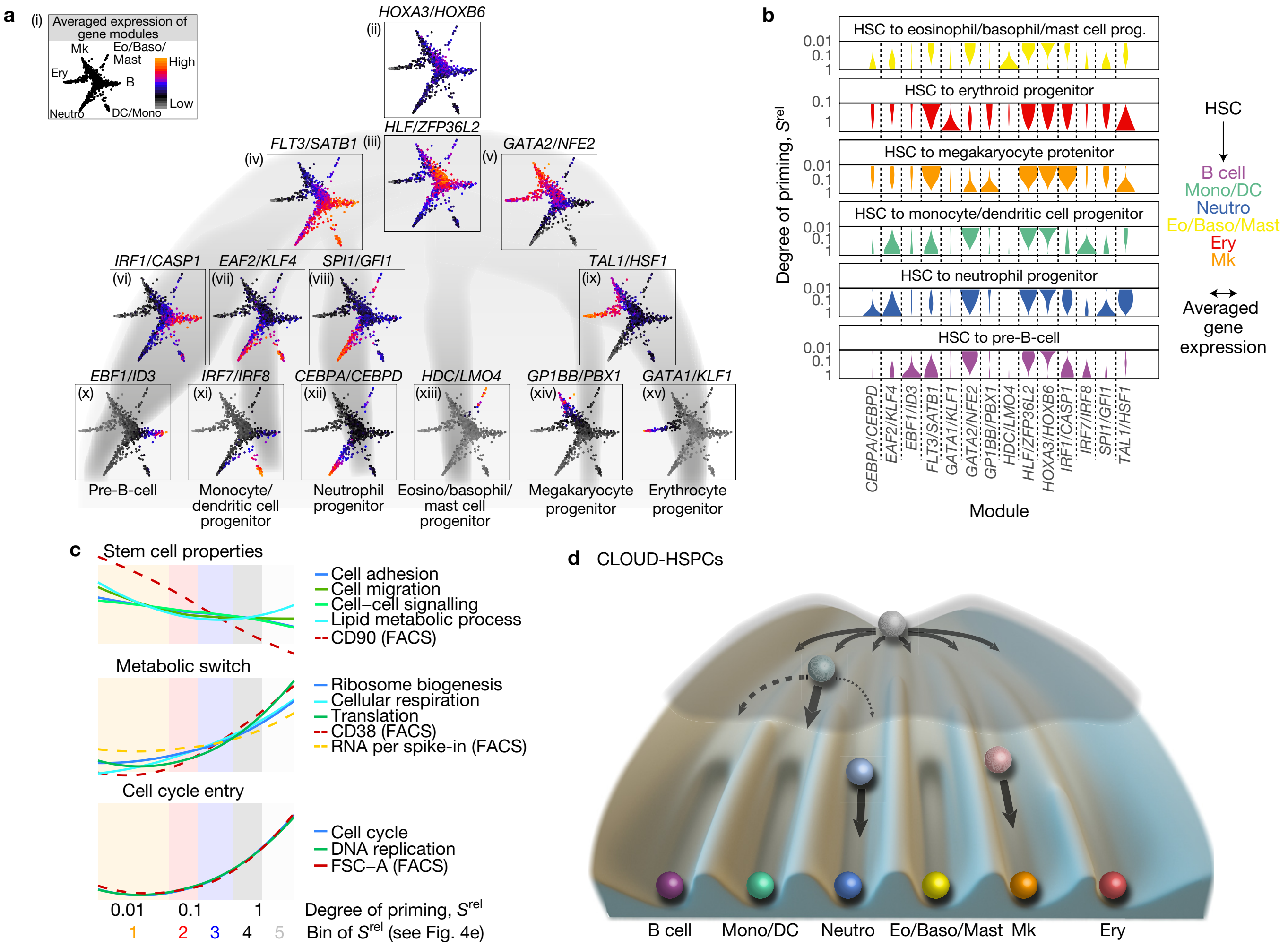


Figure 8 Lineage commitment is a layered multi-step process. (**a**,**b**) Activity of gene modules associated with developmental progression of HSPCs. Genes depending on the degree and/or direction of priming were identified and clustered into modules displaying similar expression patterns (see Methods). Averaged gene expression of selected modules from individual 1 was highlighted in the HSPC differentiation continuum (**a**) or smoothened and plotted against the degree of lineage-specific priming (**b**). For a complete list of modules and individual 2, see Supplementary Fig. 8 and Supplementary Table 4. (**c**) Gene ontology and FACS marker changes along the early priming of HSPCs ($S^{rel} < 0.4$). During later stages of priming, GO activity and FACS marker expression additionally depend on the direction of priming (not shown). (**d**) Graphical summary of a continuum-based model of bone marrow haematopoiesis. Due to the interactions of gene regulatory networks, some cell states and transitions are more likely than others, represented by a lower elevation within a Waddington landscape. During early lineage commitment, small barriers between lineages arise early, thereby creating lineage biases in HSCs. At the progenitor stage these barriers are already more pronounced, making the oligopotent stage less likely. Note that T- and NK-cell development predominantly occurs outside the bone marrow[42].

some cell state transitions and lineage combinations more likely to occur than others. Distinct lineages emerge directly from CLOUD-HSPCs, earlier than previously anticipated and without passing through a series of discrete, stable progenitors. Our data suggest a multidimensional molecular and cellular landscape of steady-state human haematopoiesis defined by a continuous flow of differentiation and emergence of lineage trajectories independent of each other. This landscape can be visualized by using the classical Waddington's landscape as a blueprint[39–41], which more appropriately reflects the continuous nature of haematopoiesis than a 'cell type tree' (Fig. 8d). Haematopoietic stem cells reside in a flat valley at the top. Barriers separating individual lineages emerge early and deepen gradually, illustrating the acquisition of lineage biases driven by small differences in gene expression of early fate mediators. When barriers become insurmountable, cell type manifestation and lineage commitment are established.

While our study provides detailed insight into lineage commitment from HSCs into all branches of human bone marrow haematopoiesis, it does not cover lineage decisions occurring further downstream or outside the bone marrow, such as T-cell development. Given the low frequency of eosinophil/basophil/mast cell and monocyte/dendritic cell progenitors within the CD34$^+$ bone marrow compartment, our study cannot fully resolve the separation and maturation of these lineages.

Together, our data determine a comprehensive continuum-based model of early human haematopoiesis, which will probably have important implications for the aetiology of haematologic disorders and which may serve as a paradigm for other adult stem cell systems. □

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of this paper.

Note: Supplementary Information is available in the online version of the paper

ACKNOWLEDGEMENTS

We thank C. Drumm for help with 3D graphics, K. Hexel, S. Schmitt, C. Felbinger and M. Eich from the DKFZ flow cytometry facility for flow cytometry support, the EMBL Genomics Core Facility for sequencing and R. Aiyar, A. Jones, M. Milsom and all members of HI-STEM and the Steinmetz group for helpful discussions on the manuscript as well as T. Schroeder and D. Löffler for initial discussions. This work was supported by the SFB873 funded by the Deutsche Forschungsgemeinschaft (DFG) (to C.L., M.A.G.E. and A.T.), the Dietmar Hopp Foundation (to M.A.G.E. and A.T.) and the US National Institutes of Health (P01 HG000205 to L.M.S.).

AUTHOR CONTRIBUTIONS

S.F.H., S.R., L.V., S.B. and C.H. performed the experiments. L.V. analysed the data, with conceptual input from S.F.H., S.R., L.M.S., M.A.G.E. and A.T., and analytical advice from W.H. S.I. and B.P.H. optimized genomics methods. C.L., E.C.B., D.N., T.B., W.-K.H. and A.D.H. obtained bone marrow aspirates. L.V., S.F.H., S.R., M.A.G.E., L.M.S. and A.T. jointly conceived and designed the study, and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://dx.doi.org/10.1038/ncb3493

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Chao, M. P., Seita, J. & Weissman, I. L. Establishment of a normal hematopoietic and leukemia stem cell hierarchy. *Cold Spring Harb. Symp. Quant. Biol.* **73,** 439–449 (2008).
2. Morrison, S., Uchida, N. & Weissman, I. The biology of hematopoietic stem cells. *Annu. Rev. Cell Dev. Biol.* **11,** 35–71 (1995).
3. Kondo, M., Weissman, I. L. & Akashi, K. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell* **91,** 661–672 (1997).
4. Akashi, K., Traver, D., Miyamoto, T. & Weissman, I. L. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature* **404,** 193–197 (2000).
5. Doulatov, S. *et al.* Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development. *Nat. Immunol.* **11,** 585–593 (2010).
6. Notta, F. *et al.* Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. *Science* **333,** 218–221 (2011).
7. Notta, F. *et al.* Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351,** aab2116 (2016).
8. Sun, J. *et al.* Clonal dynamics of native haematopoiesis. *Nature* **514,** 322–327 (2014).
9. Busch, K. *et al.* Fundamental properties of unperturbed haematopoiesis from stem cells *in vivo*. *Nature* **518,** 542–546 (2015).
10. Perié, L., Duffy, K. R., Kok, L., de Boer, R. J. & Schumacher, T. N. The branching point in erythro-myeloid differentiation. *Cell* **163,** 1655–1662 (2015).
11. Haas, S. *et al.* Inflammation-induced emergency megakaryopoiesis driven by hematopoietic stem cell-like megakaryocyte progenitors. *Cell Stem Cell* **17,** 422–434 (2015).
12. Görgens, A. *et al.* Revision of the human hematopoietic tree: granulocyte subtypes derive from distinct hematopoietic lineages. *Cell Rep.* **3,** 1539–1552 (2013).
13. Adolfsson, J. *et al.* Identification of Flt3^{+} lympho-myeloid stem cells lacking erythro-megakaryocytic potential. *Cell* **121,** 295–306 (2005).
14. Yamamoto, R. *et al.* Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells. *Cell* **154,** 1112–1126 (2013).
15. Naik, S. H. *et al.* Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature* **496,** 229–232 (2013).
16. Paul, F. *et al.* Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163,** 1663–1677 (2015).
17. Wilson, N. K. *et al.* Combined single-cell functional and gene expression analysis resolves heterogeneity within stem cell populations. *Cell Stem Cell* **16,** 712–724 (2015).
18. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32,** 381–386 (2014).
19. Shin, J. *et al.* Single-cell RNA-Seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell* **17,** 360–372 (2015).
20. Olsson, A. Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature* **537,** 698–702 (2016).
21. Theilgaard-Monch, K. The transcriptional program of terminal granulocytic differentiation. *Blood* **105,** 1785–1796 (2005).
22. Borregaard, N. Neutrophils, from marrow to microbes. *Immunity* **33,** 657–670 (2010).
23. Clark, M. R., Mandal, M., Ochiai, K. & Singh, H. Orchestrating B cell lymphopoiesis through interplay of IL-7 receptor and pre-B cell receptor signalling. *Nat. Rev. Immunol.* **14,** 69–80 (2013).
24. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13,** 845–848 (2016).
25. Hoppe, P. *et al.* Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. *Nature* **535,** 299–302 (2016).
26. Fischbach, N. A. *et al.* HOXB6 overexpression in murine bone marrow immortalizes a myelomonocytic precursor *in vitro* and causes hematopoietic stem cell expansion and acute myeloid leukemia *in vivo*. *Blood* **105,** 1456–1466 (2005).
27. Iacovino, M. *et al.* HoxA3 is an apical regulator of haemogenic endothelium. *Nat. Cell Biol.* **13,** 72–78 (2011).
28. Chuikov, S., Levi, B. P., Smith, M. L. & Morrison, S. J. Prdm16 promotes stem cell maintenance in multiple tissues, partly by regulating oxidative stress. *Nat. Cell Biol.* **12,** 999–1006 (2010).
29. Ito, K. & Suda, T. Metabolic requirements for the maintenance of self-renewing stem cells. *Nat. Rev. Mol. Cell Biol.* **15,** 243–256 (2014).
30. Shojaei, F. *et al.* Hierarchical and ontogenic positions serve to define the molecular basis of human hematopoietic stem cell behavior. *Dev. Cell* **8,** 651–663 (2005).
31. Kataoka, K. *et al.* Evi1 is essential for hematopoietic stem cell self-renewal, and its expression marks hematopoietic cells with long-term multilineage repopulating activity. *J. Exp. Med.* **208,** 2403–2416 (2011).
32. Frelin, C. *et al.* GATA-3 regulates the self-renewal of long-term hematopoietic stem cells. *Nat. Immunol.* **14,** 1037–1044 (2013).
33. Hattangadi, S. M., Wong, P., Zhang, L., Flygare, J. & Lodish, H. F. From stem cell to red cell: regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications. *Blood* **118,** 6258–6269 (2011).
34. Novershtern, N. *et al.* Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144,** 296–309 (2011).
35. Signer, R. A. J., Magee, J. A., Salic, A. & Morrison, S. J. Haematopoietic stem cells require a highly regulated protein synthesis rate. *Nature* **509,** 49–54 (2014).
36. Friedman, A. D. Transcriptional control of granulocyte and monocyte development. *Oncogene* **26,** 6816–6828 (2007).
37. Hystad, M. E. *et al.* Characterization of early stages of human B cell development by gene expression profiling. *J. Immunol.* **179,** 3662–3671 (2007).
38. Kurotaki, D. *et al.* IRF8 inhibits C/EBPα activity to restrain mononuclear phagocyte progenitors from differentiating into neutrophils. *Nat. Commun.* **5,** 4978 (2014).
39. Waddington, C. H. *The Strategy of the Genes* (Routledge, 1957).
40. Brock, A., Chang, H. & Huang, S. Non-genetic heterogeneity—a mutation-independent driving force for the somatic evolution of tumours. *Nat. Rev. Genet.* **10,** 336–342 (2009).
41. Huang, S. Non-genetic heterogeneity of cells in development: more than just noise. *Development* **136,** 3853–3862 (2009).
42. Freud, A. G. & Caligiuri, M. A. Human natural killer cell development. *Immunol. Rev.* **214,** 56–72 (2006).

METHODS

Bone marrow aspirations. Bone marrow aspirates from healthy individuals between 25 and 39 years of age were obtained at the University clinics in Heidelberg and Mannheim after written informed consent. The use of human samples for RNA-seq and functional studies was approved by the local ethics committees in accordance with the Declaration of Helsinki. Bone marrow mononuclear cells were isolated by gradient centrifugation using Histopaque-1077 (Sigma).

Flow cytometry. Bone marrow mononuclear cells were stained with surface markers for 30 min on ice according to standard protocols. For FACS sorting, BD FACS Aria II/III or Fusion flow cytometers (BD Bioscience) equipped with 405 nm, 488 nm, 561 nm and 633 nm (Aria)/642 nm (Fusion) lasers were used. For flow cytometric analyses, LSRII and LSRFortessa flow cytometers (BD Biosciences) equipped with 350 nm, 405 nm, 488 nm, 561 nm and 640 nm lasers were used. For Ki67-Hoechst cell cycle analysis, surface staining was performed as described previously[43]. Subsequently, cells were fixed and permeabilized using cytofix–cytoperm buffer (BD Bioscience), and incubated with Ki67 antibody overnight at 4 °C. Cells were stained with 2 μg ml^{-1} Hoechst 33342 (Invitrogen) and analysed. Data were analysed using FlowJo (TreeStar), indeXplorer or R.

Single-cell liquid cultures ('index-cultures'). Fresh human bone marrow mononuclear cells were stained as described above with fluorescence-labelled antibodies against CD2, CD34, CD38, CD45RA, CD71, CD90, CD130, CD135, CD238 (KEL), FcϵRI and a lineage cocktail consisting of CD4, CD8, CD11b, CD14, CD19, CD20, CD56, CD235a and CD10. Single Lin$^-$CD34$^+$CD38$^+$CD10$^-$ and Lin$^-$CD34$^+$CD38$^-$CD10$^-$HSPCs were sorted into ultralow attachment 96-well plates (Corning) containing 100 μl StemSpan SFEM media (Stem Cell Technologies), L-glutamine (100 ng m^{-1}), penicillin/streptomycin (100 ng ml^{-1}) and the following human cytokines: SCF (20 ng ml^{-1}, Peprotech), Flt3-L (20 ng ml^{-1}, Peprotech), TPO (50 ng ml^{-1}, Peprotech), IL-3 (20 ng ml^{-1}, Peprotech), IL-6 (20 ng ml^{-1}, Peprotech), G-CSF (20 ng ml^{-1}, Peprotech), IL-5 (20 ng ml^{-1}, Peprotech), M-CSF (20 ng ml^{-1}, Peprotech), GM-CSF (20 ng ml^{-1}, Peprotech) and EPO (4 U m^{-1}, R&D). For the experiment displayed in Fig. 7d, Epo was left out from the medium. Note that the CD38$^+$ and CD38$^-$ gates were set to touch (see also Supplementary Fig. 1a).

Fluorescence intensities were recorded for every channel for each sorted cell and used to retrospectively reconstruct immunophenotypic populations. Cells were cultured for 21 days at 5% CO$_2$ and 37 °C. To characterize clonal progeny, colonies were imaged by microscopy and subsequently analysed for CD15, CD33, CD41a and CD235a expression by flow cytometry. Note that under these conditions, only myeloid (CD33), erythroid (CD235a) and megakaryocytic (CD41a) colonies are efficiently generated. Colonies were judged on the basis of their visual morphology and expression of surface markers. Colony size and lineage output were based on flow cytometry and confirmed by microscopy. A colony was determined to be positive for a particular lineage if $\geq$10 cells of the respective cell type were detected.

For the 'split-in-four' experiment (Supplementary Fig. 7d,e), colonies were evaluated 7 days after seeding of single cells and colonies with more than 50 cells were equally split into 4 wells and cultured for an additional 14 days before colony size and lineage output were scored.

Mouse experiments. NSG mice were bred and housed under specific pathogen-free conditions at the central animal facility of the German Cancer Research Center. All animal experiments were approved by the Regierungspräsidium Karlsruhe under Tierversuchsantrag numbers G108/12 and G210/12.

A total of 17,000 FACS-sorted HSCs (Lin$^-$CD34$^+$CD38$^-$CD90$^+$CD45RA$^-$), MLPs (Lin$^-$CD34$^+$CD38$^-$CD45RA$^+$) or Mk-primed MPPs (Lin$^-$CD34$^+$CD38$^-$CD90$^-$CD135$^-$) from healthy bone marrow were injected into the femoral bone marrow cavity of female mice at 15 weeks of age that had been sublethally irradiated (200 cGy) 24 h before injection.

Two weeks after xenotransplantation, lineage-specific human engraftment in the injected femur was evaluated by flow cytometry using anti-human-CD45-PE, anti-human-CD235a-APC and anti-human-CD41a-FITC antibodies.

Single-cell transcriptome sequencing ('index-omics'). A 25-year-old male donor (individual 1) and a 29-year-old female donor (individual 2) were selected for single-cell RNA-seq. Fresh bone marrow mononuclear cells were stained as described above with fluorescence-labelled antibodies against CD34, CD38, CD45RA, CD90, CD49f, CD135, CD10, CD7 and a lineage cocktail consisting of CD4, CD8, CD11b, CD14, CD19, CD20, CD56 and CD235a. Fluorescence intensities were recorded for every channel for each sorted cell and used to reconstruct immunophenotypic populations subsequently.

While the frequently used smart-seq2 protocol[44] failed to amplify transcriptomes from bone marrow-derived human HSPCs, both the QUARTZ-seq protocol[45] and a modified smart-seq2 protocol (see below) yielded good-quality cDNA (Supplementary Fig. 2a). To avoid method-specific biases, data were generated using both QUARTZ-seq (individual 2) and smart-seq2.HSC (individual 1), and all findings were systematically compared between individuals (Figs 2 and 3b and Supplementary Figs 4a,b, 5a,b and 8c).

For individual 1, eight plates of Lin$^-$CD34$^+$CD38$^-$ and six plates of Lin$^-$CD34$^+$CD38$^+$ HSPCs were sorted and whole transcriptome amplification was performed using the smart-seq2 protocol[44], but using 5 μl of a modified RT buffer containing 1× SMART First Strand Buffer (Clontech), 1 mM dithiothreitol (Clontech), 1 μM template switching oligo (Exiqon), 10 U μl^{-1} SMARTScribe (Clontech) and 1 U μl^{-1} RNASin plus (Promega). ERCC spike-ins were included at a final dilution of 1:1,000,000. Libraries were constructed using a home-made Tn5 transposase (based on ref. 46). Note that the CD38$^+$ and CD38$^-$ gates were set to touch (see also Supplementary Fig. 1a).

For individual 2, eight plates of Lin$^-$CD34$^+$CD38$^-$, one plate of Lin$^-$CD34$^+$CD38$^-$CD90$^+$CD45RA$^-$ and four plates of Lin$^-$CD34$^+$CD38$^+$ HSPCs were sorted and whole transcriptome amplification was performed using the QUARTZ-Seq protocol[45]. ERCC spike-ins were included into the lysis buffer at a final dilution of 1:2,000,000. Libraries were constructed using Nextera Tn5 (Illumina) following the protocol provided, but using 1/4 of all volumes. Libraries were then sequenced on an Illumina HiSeq 2500 platform.

Raw data processing and quality control. Reads were demultiplexed and, where applicable, the remaining poly-A tail of the mRNA was trimmed off. Reads were then aligned to the *Homo sapiens* genome (build 37.68, also containing the ERCC spike in sequences) using GSNAP[47], with the expected paired-end length set to 400 bp and the allowable deviation from the expected paired-end length set to 100 bp. Reads overlapping uniquely with mRNA genes were counted using HTSeq[48]. As a first filtering step, we retained all cells in which we observed more than 750 genes at a minimum of 10 reads each, and a total of at least 150,000 reads. We removed all genes from the data set that were not observed by at least 10 reads in at least 5 cells. Statistics on these filtering steps are displayed in Supplementary Fig. 2.

We then fitted error models[49] to the readcount data (see also below). In 35 cells of individual 2 and 1 cell of individual 1, we observed an extreme overdispersion of the genes classified as non-dropout events. These cells were removed. In individual 1, we further excluded 13 cells with an abnormal CD38$^-$CD90$^{\mathrm{high}}$ immunophenotype (Supplementary Fig. 1a). These cells were clear outliers also with regard to gene expression, as they mostly expressed genes associated with various types of mature immune cell (not shown).

Data normalization using posterior odds ratio. We designed a normalization method to address the following two challenges: single-cell transcriptomics has large technical variability; and human haematopoietic stem and progenitor cells largely differ in RNA content (Supplementary Fig. 2h).

While lowly expressed genes are sometimes observed in cells with high total RNA content, they are almost never seen in cells with low total RNA content (Supplementary Fig. 2i). As this effect is the same for all genes of low expression level, it will induce some correlation structure on the data. In our data set, the first principal component was correlated to the library size and mRNA content, which may dominate over the effects of developmental transitions (Supplementary Fig. 2j, panel i). Normalization through division by total library size or harmonic mean estimator does not resolve this issue, as lowly expressed genes are still unobserved (zero) in cells of low mRNA content (Supplementary Fig. 2i,j panel ii). We and others have therefore used hierarchical models that assume that molecule counts are created by sampling from the true amount of mRNA molecules with cell-specific sampling efficiencies[50,51]. To adapt these approaches to the case where no molecular barcodes were used, we here use the error model of ref. 49, which describes the posterior probability of a gene expression level x in a cell $\mathbf{c}$ as

$$p(\mathbf{x}|r_{\mathrm{c}}, \Omega_{\mathrm{c}}) = p_{\mathrm{d}}(\mathbf{x})\mathbf{p}_{\mathrm{Poisson}}(\mathbf{x}) + (1 - \mathbf{p}_{\mathrm{d}}(\mathbf{x}))\mathbf{p}_{\mathrm{NB}}(x|\mathbf{r}_{\mathrm{c}})$$

where $\mathbf{p}_{\mathrm{d}}$ is the probability of a dropout event at gene expression $\mathbf{x}$, $\mathbf{p}_{\mathrm{NB}}$ is the probability of observing $\mathbf{r}_{\mathrm{c}}$ reads in the case of no dropout and $\mathbf{p}_{\mathrm{Poisson}}(\mathbf{x})$ is the probability of observing $\mathbf{r}_{\mathrm{c}}$ spurious reads in the case of a dropout. Ω_{c} is a vector of cell-specific and numerically optimized parameters: the slope and intercept of $\mathbf{p}_{\mathrm{d}}$ as a function of $\mathbf{r}_{\mathrm{c}}$; the slope and intercept of $\mathbf{x}$ as a function of $\mathbf{r}_{\mathrm{c}}$; the dispersion of the negative binomial distribution $\mathbf{p}_{\mathrm{NB}}(\mathbf{x}|\mathbf{r}_{\mathrm{c}})$; and the background frequency λ of the Poisson distribution, which was fixed to 0.1.

The maximum posterior average expression across all cells is then given by

$$\mu = \underset{x}{\arg\max} \prod_{c} p(x|r_{\mathrm{c}}, \Omega_{\mathrm{c}})$$

While the mean of $\prod_{\mathrm{c}} p(x|r_{\mathrm{c}}, \Omega_{\mathrm{c}})$ describes the expression magnitude of a gene in a given cell, its spread describes the uncertainty due to technical noise. To obtain a single number that weighs expression magnitude by confidence level, we compute a posterior odds ratio (POR):

$$\mathrm{POR} = \log_2 \frac{\int_{\mu}^{\infty} p(x|r_c, \Omega_c)\mathrm{d}x}{\int_{-\infty}^{\mu} p(x|r_c, \Omega_c)\mathrm{d}x}$$

POR can be interpreted as the evidence (in bits) that a specific gene in a specific cell is expressed more highly (or lowly) than in the average cell. The use of POR scores in principal component analysis solved the problems associated with the above-mentioned normalization strategies (Supplementary Fig. 2j panel iii). POR scores were used as the measure of gene expression for all analyses.

Clustering. For hierarchical clustering, we selected the 1,000 most variable genes of each population. We then used Ward linkage on Euclidean distances. Gap statistics was computed on the same hierarchical clustering function using the R package cluster. Random walk analysis[52] was performed by constructing a 5-nearest-neighbour graph on correlation distances, initializing at a random node, and then simulating a series of random steps on the 5-connected graph. The local clustering coefficient of a node in such a graph quantifies the extent to which the neighbours of two connected cells are themselves connected to each other. It was computed using the transitivity function of the igraph package[53].

STEMNET. *Basic set-up.* To identify processes associated with the transition of HSCs to progenitor cell types, we sought a lower-dimensional representation of the HSPC data that reflects lineage priming. We therefore trained an elastic-net regularized generalized linear model (GLMNET) of the multinomial family on the most mature populations (N1-3, EBM, MD, spB1/2, E1/2 and Mk from Fig. 2a for individual 1, or lpB, EBM, N, ME and MD for individual 2), using class membership as the response variable. During this step, a number of population-specific genes was identified (Supplementary Table 3). The classifier then used the expression of these genes in all cells to estimate the probability $\mathbf{p}_{ij}$ that a cell i belongs to class j. From these probabilities, we compute the Kullback–Leibler distance from the average HSPC, which can be interpreted as the amount of lineage information a given cell has acquired:

$$S_i^{\mathrm{rel}} = \sum_{j=1}^{6} p_{ij} \log \frac{p_{ij}}{\bar{p}_j}$$

where $\bar{\mathbf{p}}_j$ is the average probability of a cell to belong to class j. We further assign each cell a predominant direction of priming as

$$d_i = \arg\max_j \frac{p_{ij}}{\bar{p}_j}$$

For displaying the six-dimensional vector $\mathbf{p}_i$ in two dimensions, the developmental endpoints are arranged on the edge of a circle and all cells are placed in between. Each endpoint $\mathbf{k}$ is assigned with an angle α_k. The class probabilities $\mathbf{p}_{ik}$ are then transformed to Cartesian coordinates by

$$x_i = \sum_k p_{ik} \cos\alpha_k$$

and

$$y_i = \sum_k p_{ik} \sin\alpha_k$$

To find the optimal arrangement of the developmental endpoints on the circle, lineages with common precursor stages are placed next to each other. The proximity between lineages l and k is computed by

$$D_{kl} = \sum_i p_{il} \times p_{ik}$$

All arrangements are tested and the arrangement with the highest proximity is chosen. This approach is based on a method termed 'circular a posteriori projection'[51].

Data simulation. To test the ability of the STEMNET method to uncover binary branching events and discrete subpopulations, we quantitatively specified alternative models of cell fate specification and reshuffled our original data according to these models (Supplementary Fig. 6). In particular, we assumed that each cell is located on a binary tree, where nodes represent branching points and edges between nodes represent developmental trajectories. Each node $\mathbf{V}_i$ is specified by a tuple ($\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{p}_1$, $\mathbf{p}_2$, $\mathbf{h}$) with $\mathbf{E}_{1,2}$ pointing to the left and right child, $\mathbf{p}_{1,2}$ giving the probability that a cell adapts the fate associated with the left and right child ($\mathbf{p}_1 + \mathbf{p}_2 = 1$), and $\mathbf{h} \in (0,1)$ giving the height of the node (for developmental endpoints, $\mathbf{h} = 1$, and for the root, $\mathbf{h} = 0$). A cell is then defined by the tuple ($\mathbf{h}$,$\mathbf{E}$), where $\mathbf{E}$ points to the next node downstream of the cell.

For the scenario depicted in Supplementary Fig. 6a, cells were generated by randomly drawing values $\mathbf{h}$ from a Beta distribution with parameters (2,3). $\mathbf{E}$ was assigned by moving down a distance of $\mathbf{h}$ from the root and randomly choosing a branch according to $\mathbf{p}_{1,2}$ at each node that was passed. For the scenario depicted in Supplementary Fig. 6d cells were then scattered around the nearest node assuming an average distance of 0.01. The developmental distance $\mathbf{D}(\mathbf{c}_i, \mathbf{V}_j)$ between a cell $\mathbf{c}_i$ and a node $\mathbf{V}_j$ is then computed by traversing through the tree and summing all distances $\mathbf{h}$ that are passed along the way. For example, the distance between two developmental endpoints that diverge at a node with $\mathbf{h} = 0.6$ is 0.8. To generate synthetic data from these cell fate specification models, we extracted the coefficients of the STEMNET classifier (Supplementary Table 3), and for each developmental endpoint j compiled lists of genes with nonzero coefficient. Gene expression values for these genes were then reordered across cells i to follow the developmental distance $\mathbf{D}(\mathbf{c}_i, \mathbf{V}_j)$ (that is, assuming that gene expression of lineage-specific genes was entirely determined by developmental distance, Supplementary Fig. 6a). Alternatively, gene expression values were randomly reshuffled such that the correlation between developmental distance from $\mathbf{V}_j$ and gene expression equals the empirically observed correlation between gene expression and $\mathbf{p}_j$ from the STEMNET classifier (Supplementary Fig. 6b–d).

Quantitative link between single-cell transcriptomics and single-cell culture. To quantitatively link single-cell transcriptomic properties (such as the amount or direction of priming) to single-cell functional properties, we made use of FACS markers used in both experiments. In particular, for each transcriptomic property, we constructed a regression model with logicle transformed flow cytometry markers as explanatory variables and the property as a response variable. To achieve greater robustness than in standard linear regression, we applied GLMNET models of the normal family for this task, and used tenfold cross-validation to determine the regularization parameter λ. The regression coefficients of these models are shown in Supplementary Fig. 7a together with the R^2 these models achieve in tenfold cross-validation if applied to the single-cell transcriptomic data. We then applied these classifiers to logicle transformed flow cytometry data from the single-cell culture experiment to estimate the magnitude of single-cell transcriptomic properties in that experiment. To further improve the classifier, we also included rank-transformed mRNA expression levels of *TFRC (CD71)* and *KEL* in the training data, and rank-transformed flow cytometry data of CD71 and KEL in the single-cell culture experiment.

Identification of gene clusters associated with lineage priming. We then identified genes whose expression depends on S^{rel}, $\mathbf{d}$, or both, by separately fitting four different linear models to the expression data of each gene. The first model describes gene expression as a function of the predominant direction $\mathbf{d}$, which is a categorial variable. It best fits to genes that are up- or downregulated early during developmental progression in a certain direction and stay unchanged until the end. The second model describes gene expression as a function of a third-degree polynomial through $\log_{10} S^{\mathrm{rel}}$. It best fits to genes that are up- or downregulated at a specific stage of developmental progression, independent of the developmental direction. The third model describes gene expression as a function of $\mathbf{d}$, a third-degree polynomial through $\log_{10} S^{\mathrm{rel}}$ and the interaction of $\mathbf{d}$ and $\log_{10} S^{\mathrm{rel}}$. It best fits to genes that are up- or downregulated at a specific stage of development in a specific direction. The fourth model describes gene expression as a constant. It best fits to genes that do not change systematically during acquisition of lineage fate. For each gene, we identified the optimal model by comparing the models' Bayesian Information Criteria (BIC). For each class of genes (dependent on $\log_{10} S^{\mathrm{rel}}$, $\mathbf{d}$ or both) separately, we identified subgroups of genes that display similar dependencies on $\log_{10} S^{\mathrm{rel}}$ and $\mathbf{d}$ by performing hierarchical clustering using correlation distance and complete linkage on the fitted values from the preferred model.

Statistics and reproducibility. Single-cell RNA-seq was performed on two different individuals. Totals of 1,034 (for I1) and 379 cells (for I2) were included into the study. Single-cell culture was performed for 2,038 cells. As indicated in the figure legends, P values are computed from the Pearson product moment correlation test, kernel-density-based global two-sample comparison test or two-tailed unpaired t-test.

For animal experiments, no statistical method was used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to animal allocation during experiments and outcome assessment.

Code availability. Most analyses were performed in indeXplorer, a custom-made software for the analysis of single-cell index-sorting/transcriptomic data sets. indeXplorer was written in R and relies on the package shiny; code is available from https://git.embl.de/velten/indeXplorer.

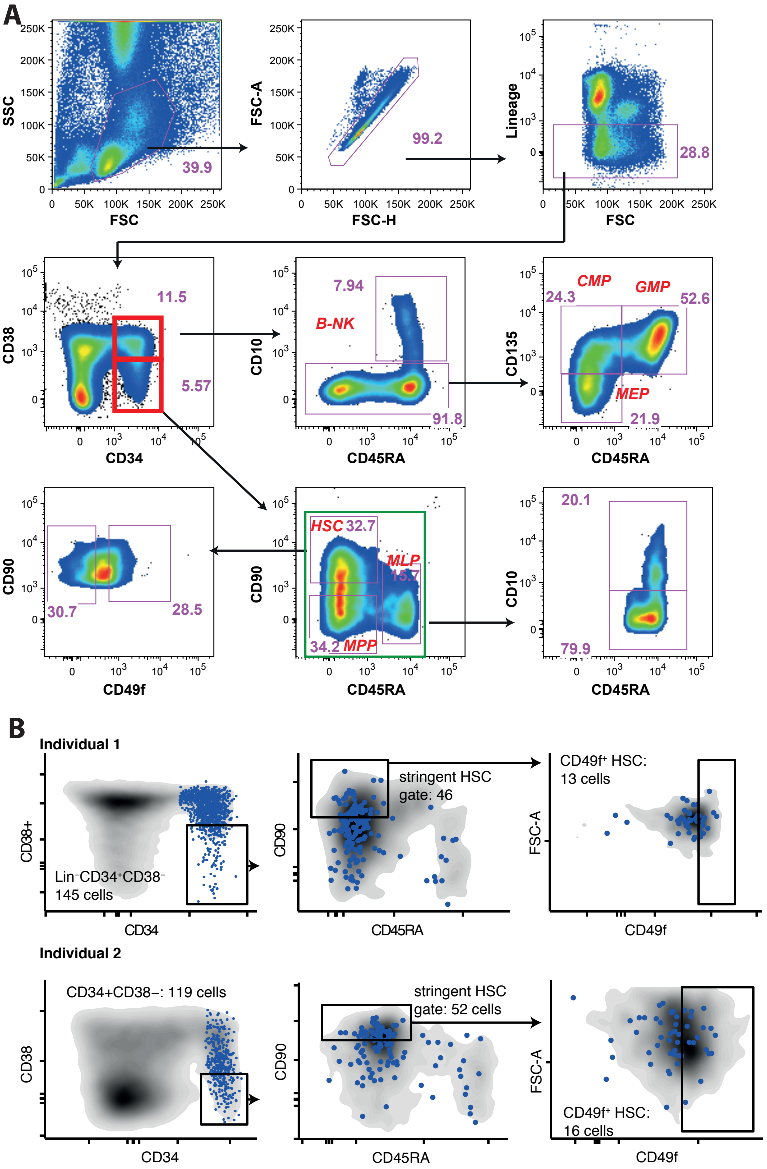
For analyses that were not performed in indeXplorer directly, we provide an R package containing all code at https://git.embl.de/velten/STEMNET.

Data availability. RNA-seq data that support the findings of this study have been deposited in the Gene Expression Omnibus (GEO) under accession code GSE75478. Processed data are available at http://steinmetzlab.embl.de/shiny/indexplorer/?launch=yes for browsing. All other data supporting the findings of this study are available from the corresponding author on reasonable request.

43. Essers, M. A. G. *et al.* IFNα activates dormant haematopoietic stem cells *in vivo*. *Nature* **458,** 904–908 (2009).
44. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10,** 1096–1098 (2013).
45. Sasagawa, Y. *et al.* Quartz-Seq: a highly reproducible and sensitive single-cell RNA-Seq reveals non-genetic gene expression heterogeneity. *Genome Biol.* **14,** R31 (2013).
46. Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24,** 2033–2040 (2014).
47. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26,** 873–881 (2010).
48. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).
49. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11,** 740–742 (2014).
50. Velten, L. *et al.* Single-cell polyadenylation site mapping reveals 3' isoform choice variability. *Mol. Syst. Biol.* **11,** 812 (2015).
51. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343,** 776–779 (2014).
52. Van Dongen, S. Graph clustering via a discrete uncoupling process. *SIAM J. Matrix Anal. Appl.* **30,** 121–141 (2008).
53. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **Complex Systems,** 1695 (2006).

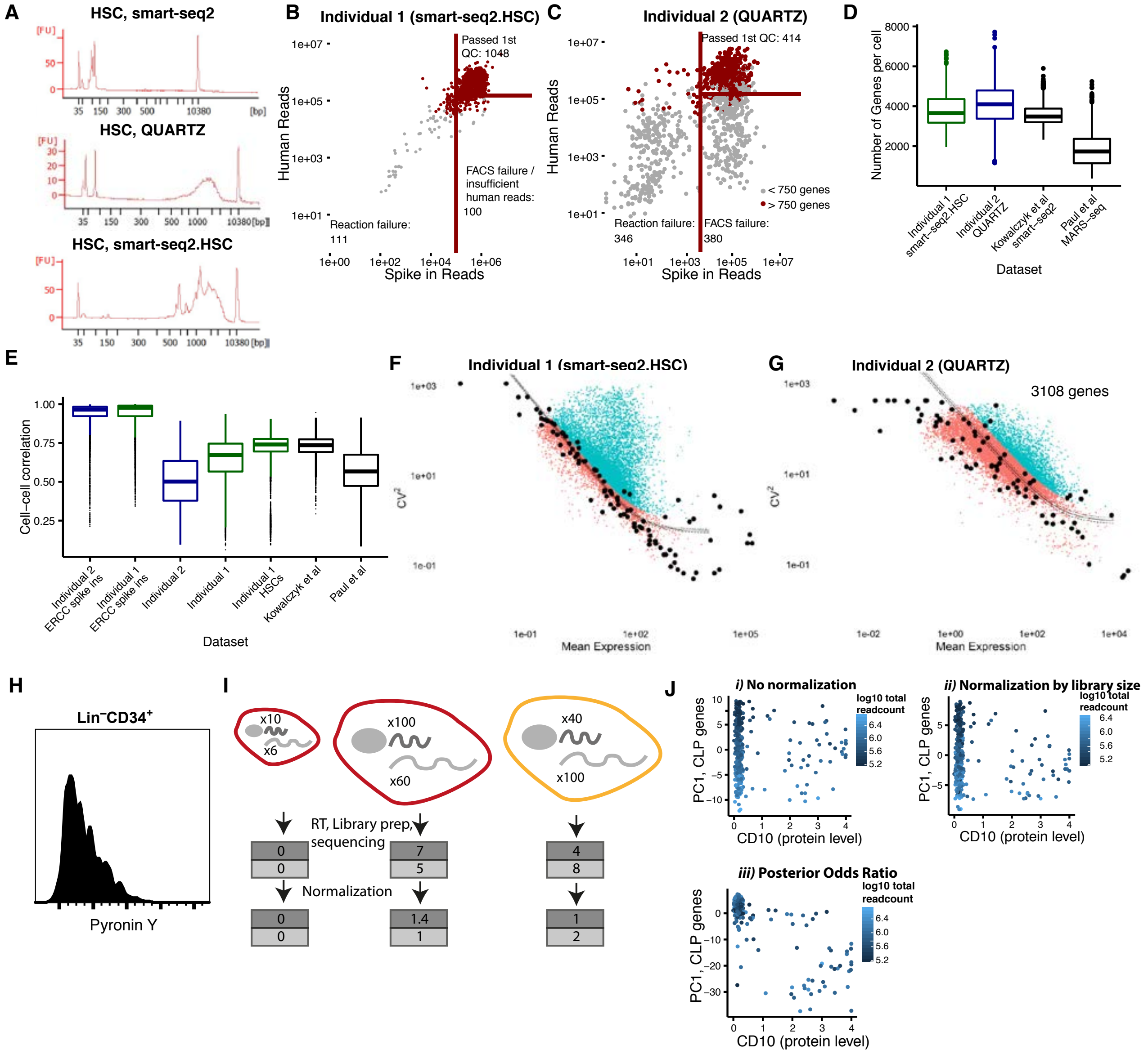
DOI: 10.1038/ncb3493

In the format provided by the authors and unedited.

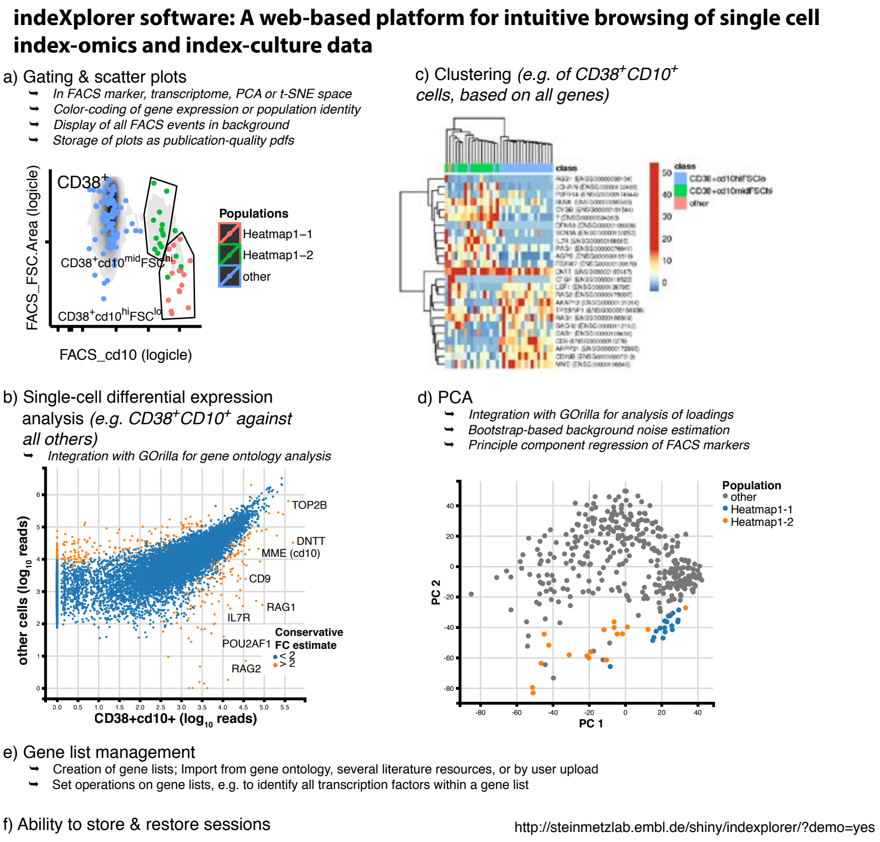


Population (according to Doulatov et al 2010)	Number of cells, individual 1	Number of cells, individual 2
HSC	46	52
MPP	43	22
MLP	10	11
B-NK	14	33
GMP	203	12
MEP	147	12
CMP	94	13
Cells between gates	**477 (46%)**	**224 (60%)**

Supplementary Figure 1 Flow cytometric display of the setup used for index-omics. a, Sorting was performed exclusively on the gates highlighted in red in the CD34 versus CD38 panel and all surface markers were indexed. Cells outside the green gate in the CD45RA versus CD90 panel were excluded retrospectively as they represented mature immune cells (not shown). Percentage of cells within each gate is indicated. Data from individual 1 is shown. b, Distribution of cells subjected to single-cell RNA-Seq within classically defined gates[1,2]



Supplementary Figure 2 Quality metrics of single-cell RNA-Seq. a, Bioanalyzer traces of amplified cDNA generated from single human HSPCs with the default smart-seq2 protocol (upper panel), QUARTZ-Seq (middle panel, applied to individual 2) and a modified version of smart-seq2 (lower panel, applied to individual 1, see methods). b, c, Filtering of cells based on total read counts and number of genes expressed. The use of the modified smart-seq2 protocol (b) strongly decreased the dropout rate compared to the QUARTZ-Seq protocol (c). The large dropout rate in the QUARTZ-Seq protocol was due to the small volume used. d, e, The number of genes per cell (d) and cell-cell correlation (e) for the two individuals compared to two other recent single-cell RNA-Seq data sets from the haematology field[3,4]. Box plots display median bar, first–third quantile box and 5th–95th percentile whiskers. n=379 cells individual 1, n=1034 cells individual 2, n=218 cells Individual 1, HSCs; n=2730 cells Paul et al., n=1058 cells Kowalczyk et al. f, g, The mean read count and variance of spike-ins (large black dots) and genes (small dots) were compared in order to identify genes whose biological noise exceeded technical variability (cyan dots)[5]. h, The total RNA content of $\text{Lin}^-\text{CD34}^+$cells varies widely. i, Cartoon describing the hypothetical effect of large variations in RNA amount in homogeneous populations. Two cells from the same population (red) display identical RNA concentrations for two genes, but differ in RNA amount by 10-fold. A third cell from a different population expresses the two sample genes at a different ratio but absolute high number. Following sequencing, the genes are more likely to be lost in the smaller cell, which cannot be reverted by normalization. j, PCA performed on lymphoid (CLP) specific genes[6] should clearly separate cells expressing the lymphoid surface marker CD10. However, without normalization cells are only arranged by read count (*i*). Standard normalization using a harmonic mean estimator of library size does not solve the problem (*ii*). Following normalization by Posterior Odds Ratios (POR, see Online Methods) a PCA performed on CLP specific genes clearly separates CD10^+ and CD10^- cells (*iii*).



indeXplorer software: A web-based platform for intuitive browsing of single cell index-omics and index-culture data

a) Gating & scatter plots
- *In FACS marker, transcriptome, PCA or t-SNE space*
- *Color-coding of gene expression or population identity*
- *Display of all FACS events in background*
- *Storage of plots as publication-quality pdfs*

c) Clustering *(e.g. of CD38+CD10+ cells, based on all genes)*

b) Single-cell differential expression analysis *(e.g. CD38+CD10+ against all others)*
- *Integration with GOrilla for gene ontology analysis*

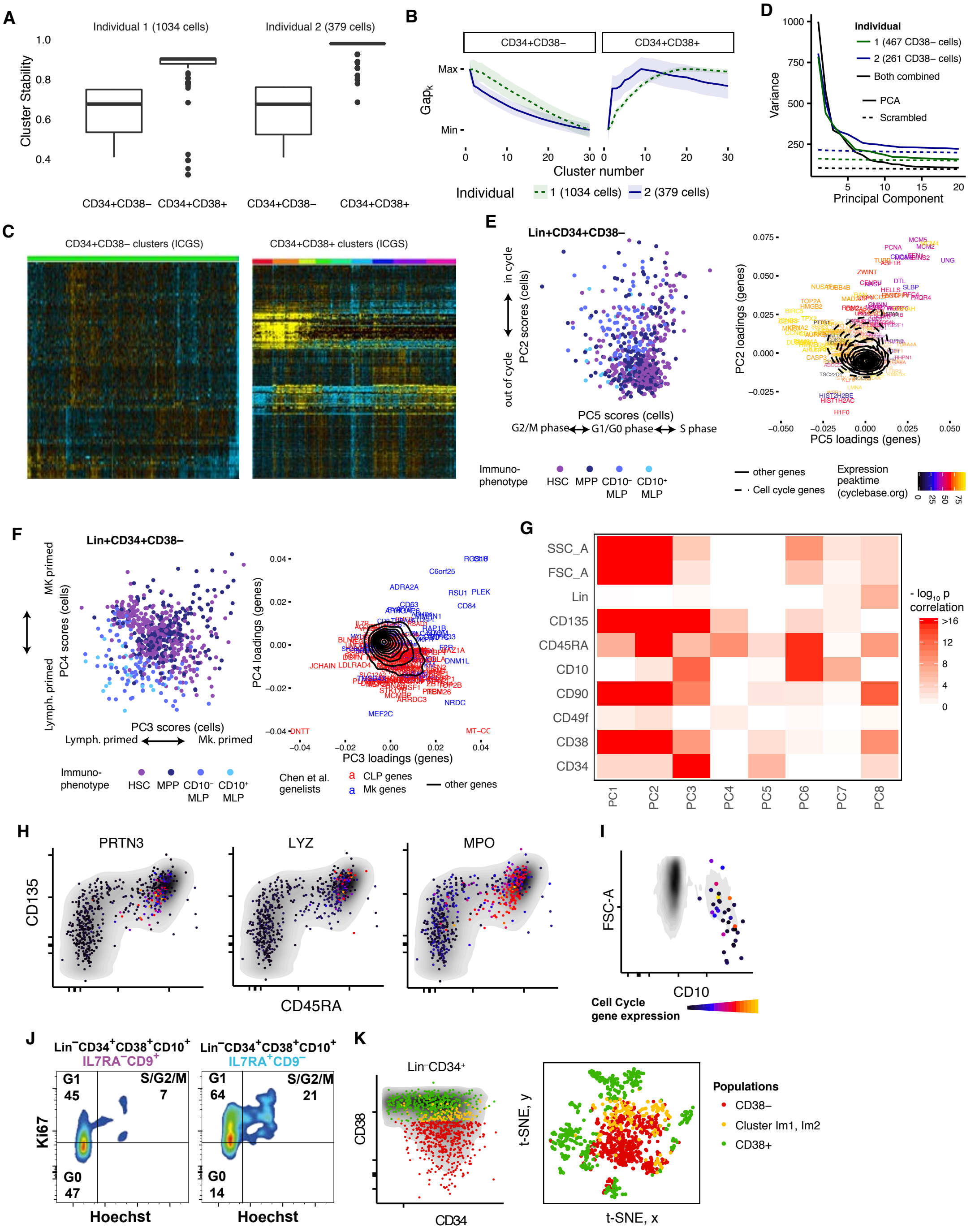
d) PCA
- *Integration with GOrilla for analysis of loadings*
- *Bootstrap-based background noise estimation*
- *Principle component regression of FACS markers*

e) Gene list management
- Creation of gene lists; Import from gene ontology, several literature resources, or by user upload
- Set operations on gene lists, e.g. to identify all transcription factors within a gene list

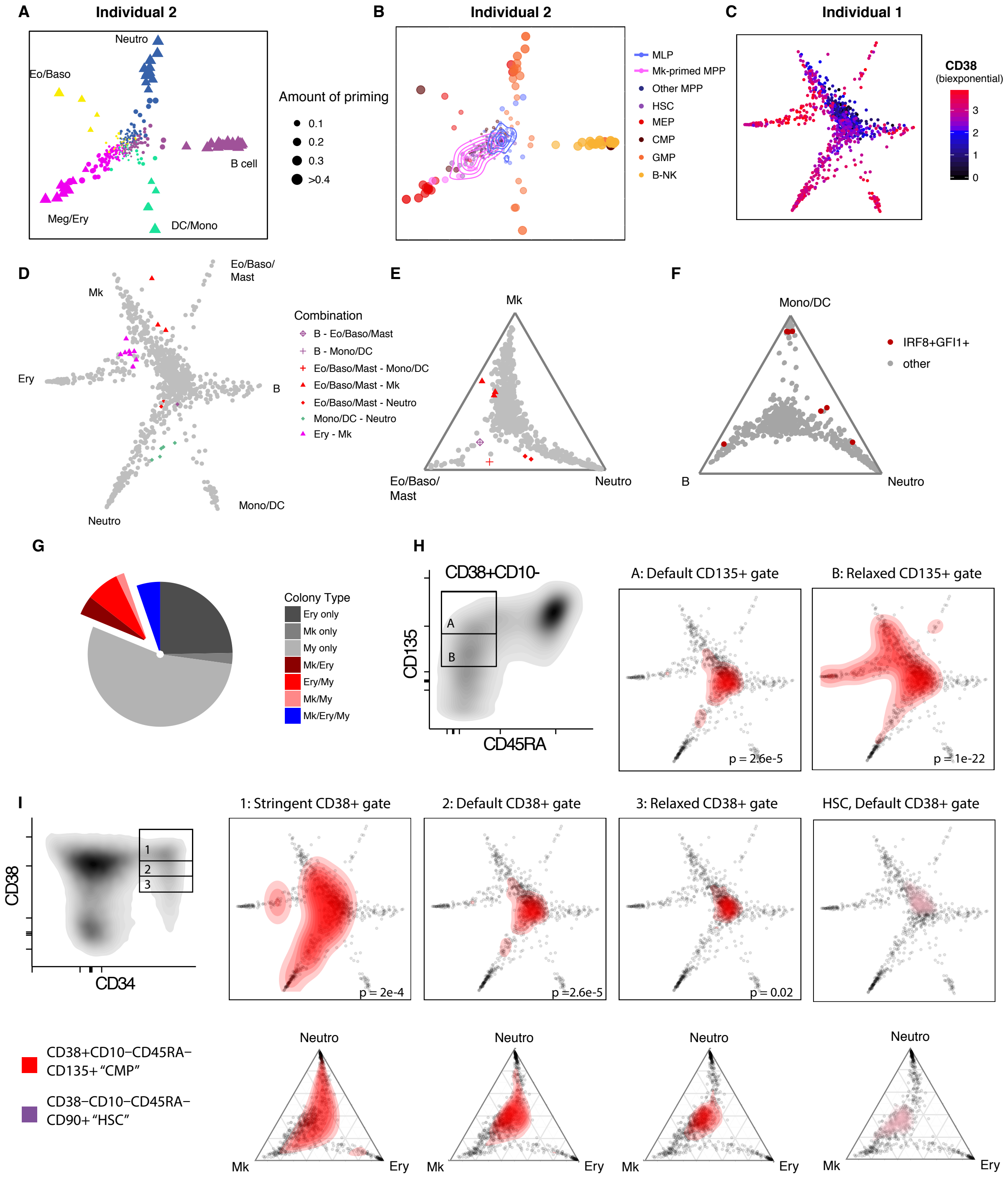
f) Ability to store & restore sessions

http://steinmetzlab.embl.de/shiny/indexplorer/?demo=yes

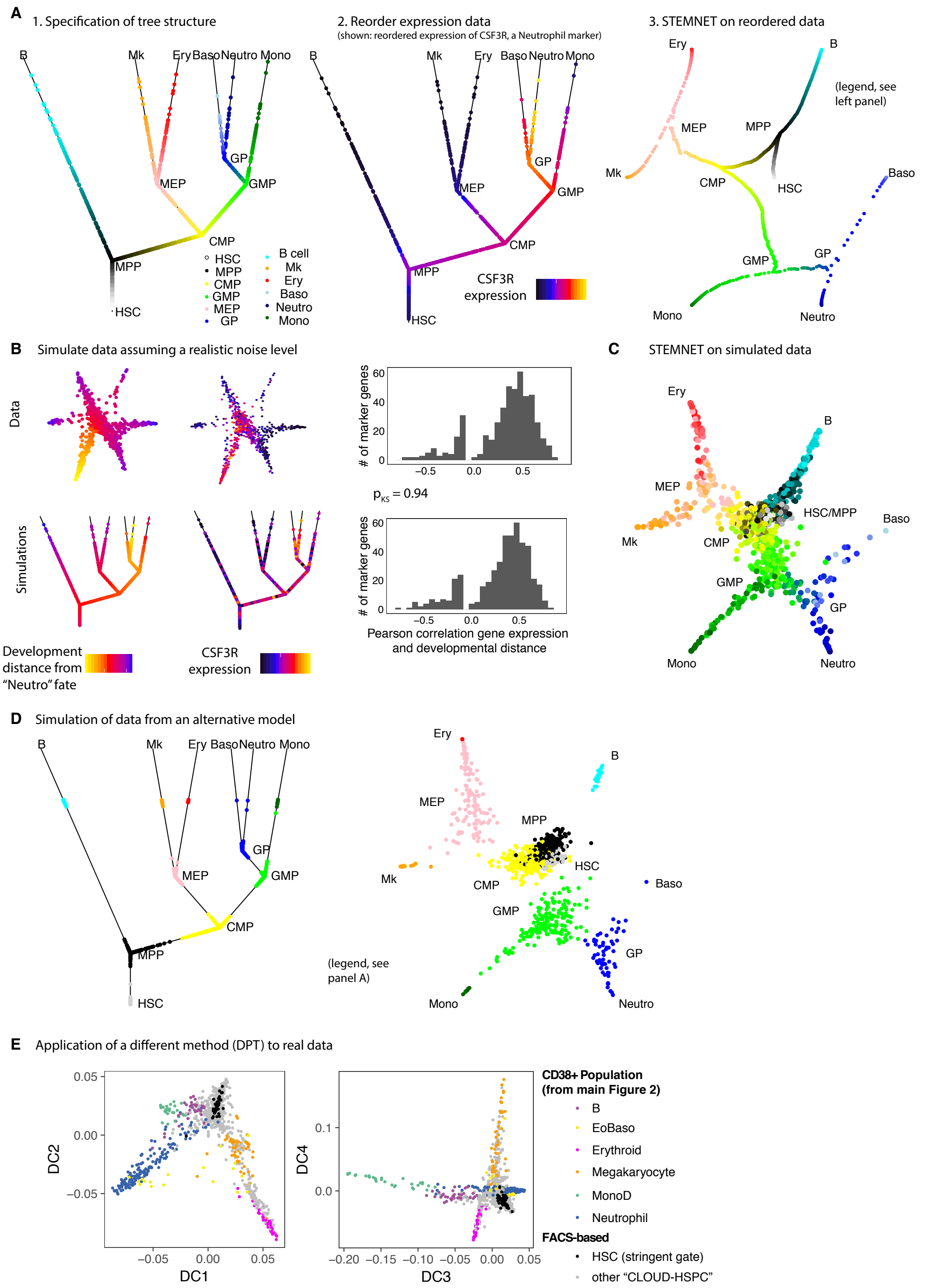
Supplementary Figure 3 *indeXplorer*, a web-based GUI for exploring single-cell index-omics and index-culture data. *indeXplorer* combines the capabilities of a FACS software with tools for the analysis of single cell transcriptomics data in a single graphical user interface. FACS and transcriptomics modules are tightly linked, allowing for example the display of gene expression or transcriptomic clusters on FACS scatter plots (a), differential expression testing of arbitrarily gated populations (b), as well as hierarchical clustering (c) and principal component analysis (d). *indeXplorer* further provides tools for gene list management, allows the user to download plots as publication-quality pdfs, and to store & restore sessions. On http://steinmetzlab.embl.de/shiny/indexplorer/?demo=yes we provide a short interactive introduction into the use of *indeXplorer*.



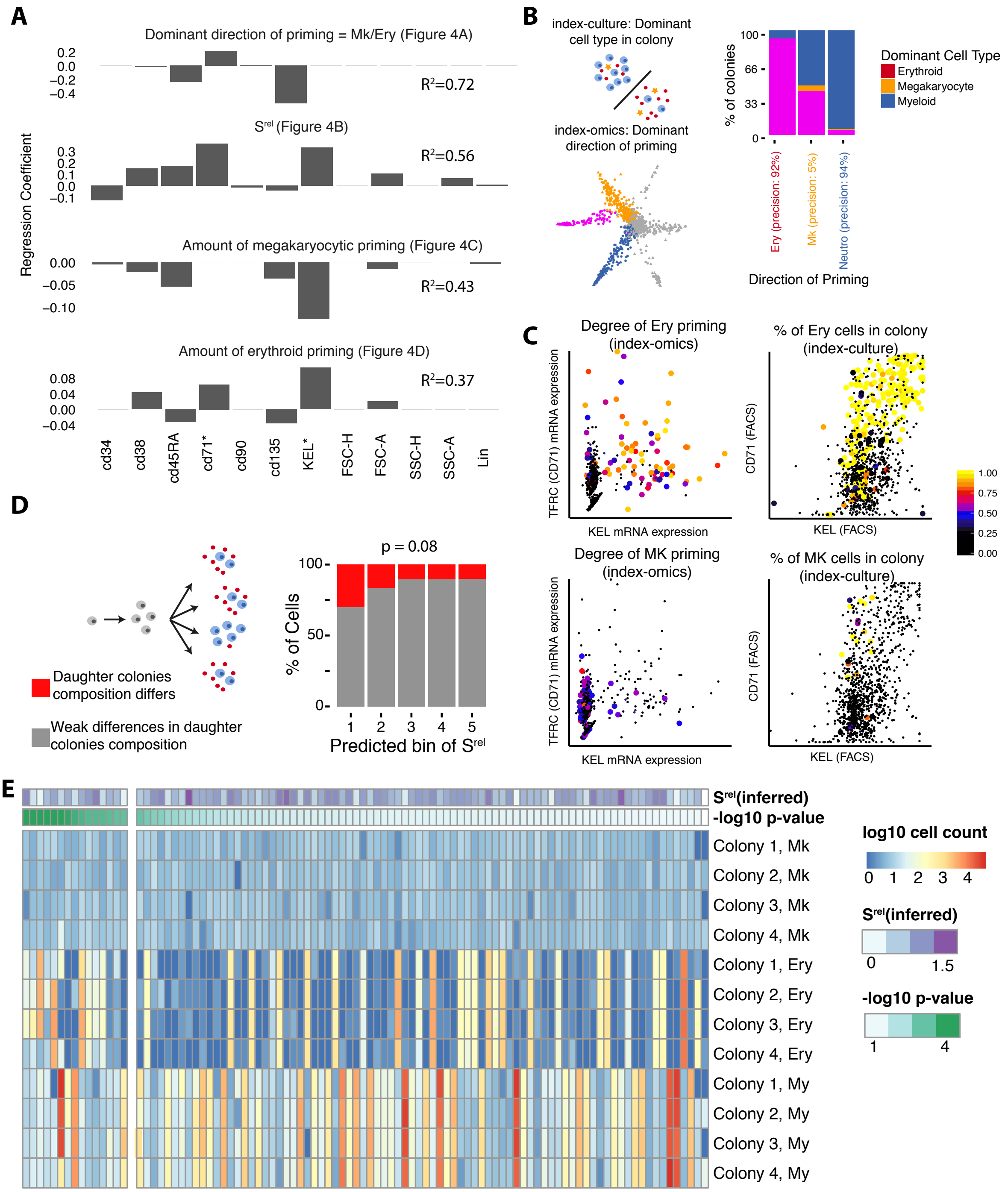
Supplementary Figure 4 Unsupervised analyses of single-cell transcriptomics. a, Cluster stability analysis[7] of the Lin$^-$CD34$^+$CD38$^-$ and Lin$^-$CD34$^+$CD38$^+$ populations. For n=500 repetitions, 66% of cells were randomly selected, clustering was performed and a consensus clustering was computed. The probability that clusterings obtained from random subsets of the data agree with the consensus is plotted on the y axis (box plot with median bar, first–third quantile box and 5th–95th percentile whiskers). b, Gap-statistic (Gap_k) of Lin$^-$CD34$^+$CD38$^-$ and Lin$^-$CD34$^+$CD38$^+$ compartments. A maximum of Gap_k indicates the statistically optimal cluster number[8]. c, clustering obtained using ICGS[9]. 4 outlier cells in the Lin$^-$CD34$^+$CD38$^-$ compartment (left panel, blue bar) were characterized by a lower number of genes detected, but no coherent differences in gene expression (not shown). d-f, Transcriptomic heterogeneity in the Lin$^-$CD34$^+$CD38$^-$ compartment. d, >10 principal components in Lin-CD34$^+$CD38$^-$ exceed noise. e, Principal components 2 and 5 of a PCA performed on combined data from both individuals. Loadings of all genes with annotated cell-cycle phase dependent gene expression patterns[10] are shown in the right panel. Cell cycle associated genes are shifted compared to other genes on PC2 and arranged by peak time of gene expression on PC5. Scores of all Lin$^-$CD34$^+$CD38$^-$ cells are shown in the left panel. f, Principal components 3 and 4. Loadings of all genes annotated as CD38$^+$CD10$^+$ "CLP" or CD41$^+$CD42$^+$GP6$^+$ "Mk" specific[6] are shown, demonstrating that PC3 and PC4 correlate with lymphoid versus megakaryocytic priming. Scores of all Lin$^-$CD34$^+$CD38$^-$ cells are shown in the left panel. g. Principal components of Lin$^-$CD34$^+$CD38$^-$ cells are significantly correlated to surface marker expression. Data from individual 1 are shown. h, Expression of neutrophil marker genes in relation to CD45RA and CD135. See also Main Fig. 4c. i, Expression of cell cycle genes suggests that the CD10$^{\text{mid}}$FSC-A$^{\text{high}}$ population is more actively cycling. j, Ki67-Hoechst cell cycle analyses of IL7R-CD9$^+$ and IL7R$^+$CD9$^-$ populations, corresponding to sB and lB respectively. k, Cells from the transcriptomic *Im* cluster have intermediate CD38 expression and group with Lin$^-$CD34$^+$CD38$^-$ HSPCs in t-SNE analysis.



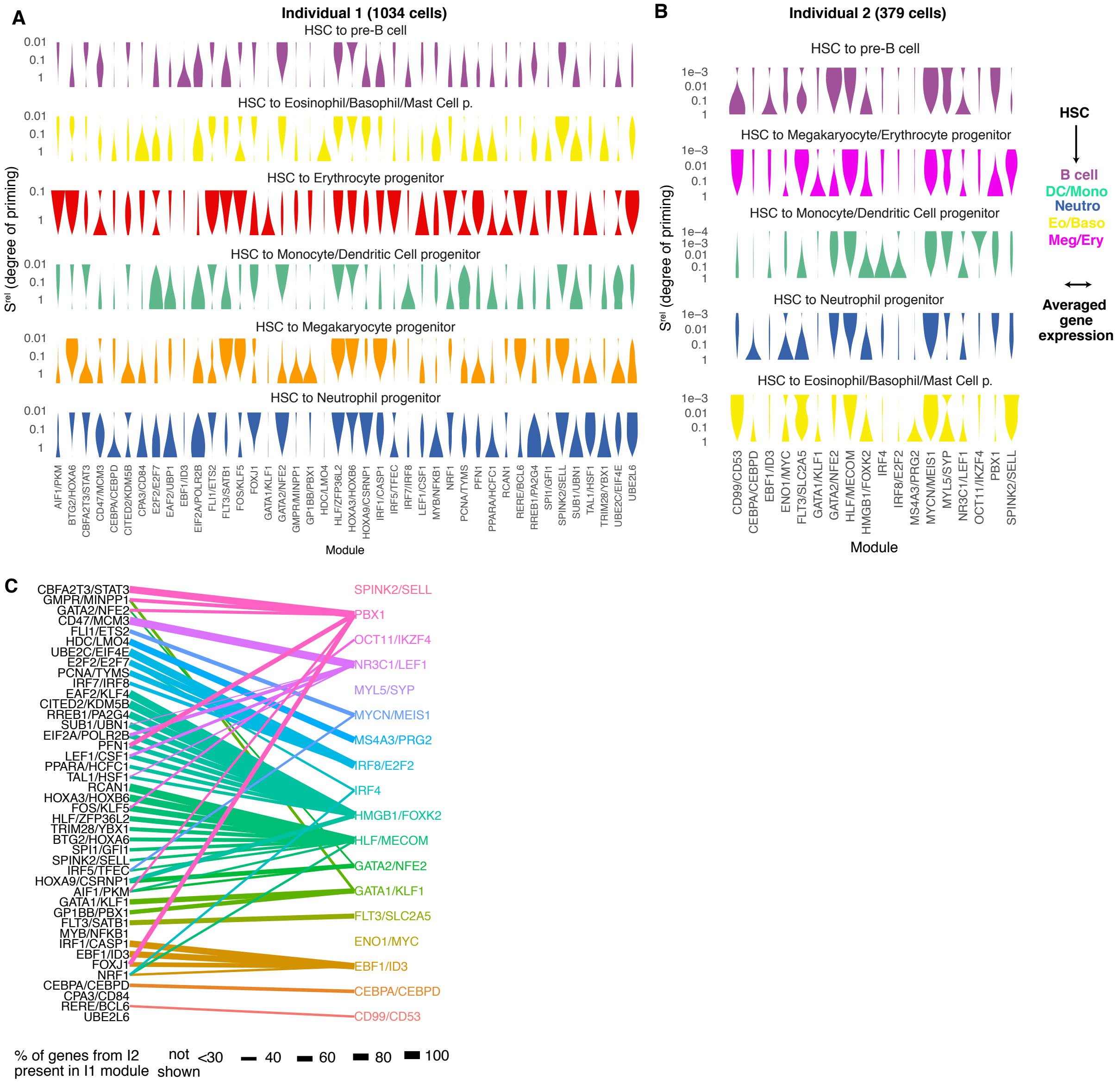
Supplementary Figure 5 Analyses using STEMNET. a, The similarity of every cell to each of the progenitor classes was computed by STEMNET (see methods), projected on a unit circle, and used to quantify the degree and direction of transcriptomic priming. Data from individual 2 is shown. b, immunophenotypes highlighted on the STEMNET plot for individual 2. c, CD38 surface marker expression highlighted on the STEMNET plot for individual 1. d, e, Dual lineage primed cells, defined as cells with more than 25% priming in two directions, were highlighted on the STEMNET plot (d) or in a ternary plot depicting only priming in the Mk, Neutro, and Eo/Baso/Mast directions (e). f, Rare $IRF8^{+}GFI1^{+}$ progenitors[9] are not a typical intermediate stage between granulocytes and monocytes but appear displaced from developmental trajectories or are fully primed towards individual lineages. g, Distribution of colony types observed in the index-culture experiment. Functionally bipotent cells are highlighted. h, i, The transcriptomic lineage priming of immunophenotypic CMPs depends strongly on the gating strategy. Cells from the CMP gate ($Lin^{-}CD34^{+}CD38^{+}CD45RA^{-}CD135^{+}$) were highlighted on the STEMNET plot (upper panels) or as ternary plots (lower panels). The effect of variations in the CD135 (h) and CD38 (i) gates are shown. P-values were calculated by kernel-density based tests comparing each population to $CD49f^{+}$ HSCs. For $CD49f^{+}$ HSCs, n=101 single cells; CMPs, default gate, n=64; CMPs, relaxed CD38 gate, n=164; CMPs, stringent CD38 gate, n=24; CMPs, relaxed CD135 gate, n=180.



Supplementary Figure 6 Simulation of data from alternative models of cell fate specification. a, To demonstrate the ability of STEMNET to identify subsequent binary branching events, we assumed a scenario where cells locate on developmental trajectories between universally defined branching points (left panel, see also methods). For each gene used by STEMNET as a marker specific to a given developmental endpoint, we reordered the expression data to parallel the developmental distance from that endpoint. The middle panel depicts exemplary the reordered expression of CSF3R, a neutrophil marker. Finally, we apply STEMNET to the reordered data set (right panel). b, To simulate data using a more realistic noise level, we estimated the correlation between developmental distance and gene expression from the data for each gene (upper panels). We then reshuffled the expression values such that the correlation between marker gene expression and (simulated) developmental distance approximates the correlation estimated from the data (lower panels). c, STEMNET on reshuffled data. d, To simulate a scenario where HSCs pass through discrete progenitor cell types, cells were placed near branching points, data was simulated as described for panel (b), and STEMNET was applied to the reshuffled data. e. Projection of single cell expression data into diffusion map space[11].



Supplementary Figure 7 The quantitative link between index-omics and index-culture. a, Regression models used to estimate transcriptomic quantities from FACS surface marker expression. Model coefficients and the fraction of variance explained in a 10-fold cross validation scheme (R^2) are shown. For genes marked with an asterisk, regression models were constructed on mRNA expression and applied to FACS surface marker expression. b, Linkage of the exact predicted direction of transcriptomic priming (for the cell types with robust colony forming abilities; Neutro, Ery, Mk) to the actual cell type composition of the *ex vivo* colonies. Illustration (left panel) and quantitative linkage (right panel) are shown. The exact direction of transcriptomic priming was estimated for each founder cell from index-culture based on regression models constructed on all surface markers and compared to the observed colony composition. c, CD71 and KEL FACS marker and mRNA expression in relation to the degree of transcriptomic Ery/Mk priming and the percentage of Ery/Mk cells in the colony. d, e, As an additional experimental measure of developmental plasticity, we cultured single HSPCs for 1 week, split the colony in four and determined the lineage outcome of the daughter colonies two weeks later. For several colonies, the lineage output varied significantly across daughters (e, p-values are from a chi-square test for independence). These colonies tended to derive from developmentally more primitive cells (d). p-value is from a Pearson product moment correlation test with n=96 split-in-four experiments.



Supplementary Figure 8 Gene modules affected by the onset of lineage. a, b, Averaged gene expression of all gene modules, including those omitted from the main figure, was smoothened and plotted against the degree of lineage-specific priming. Data is shown for individual 1 (a) and 2 (b). c, Comparison between gene modules from individual 1 and 2. For each module from individual 1, the overlap with each module from individual 2 is shown. Due to the higher number of cells analysed, gene modules from individual 2 split up into multiple modules from individual 1, while modules from individual 1 overlap only with a single module from individual 2. Only genes discovered in both individuals were included in this analysis.

Supplementary Table Legends

Supplementary Table 1 List of antibodies used.

Supplementary Table 2 Genes overexpressed by cell populations in the Lin$^-$CD34$^+$CD38$^+$ compartment.

Supplementary Table 3 STEMNET models.

Supplementary Table 4 Gene modules with dependence on direction and degree of priming.

References

1. Doulatov, S. *et a*l. Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development. *Nat. Immunol.* **11,** 585–93 (2010).
2. Notta, F. *et al*. Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. *Science* **333,** 218–21 (2011).
3. Paul, F. *et al.* Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163,** 1663–1677 (2015).
4. Kowalczyk, M. S. *et al.* Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.* **25,** 1860–1872 (2015).
5. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10,** 1093–5 (2013).
6. Chen, L. *et al.* Transcriptional diversity during lineage commitment of human blood progenitors. *Science (80-.).* **345,** 1251033–1251033 (2014).
7. Ohnishi, Y. *et al.* Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. *Nat. Cell Biol.* **16,** 27–37 (2014).
8. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **63,** 411–423 (2001).
9. Olsson, A. *et al.* Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature* **537**, 698–702 (2016).
10. Santos, A., Wernersson, R. & Jensen, L. J. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Res.* **43,** D1140–D1144 (2014).
11. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).