Miniaturized high-throughput version of CEL-Seq2 maintains high sensitivity and accuracy.

(a-e) Violin plots showing a comparison between manual CEL-Seq2 and the robotic version at different volume reductions for the distributions of (a) the number of transcripts per mESC, (b) the fraction of recovered ERCC\(^1\) spike-in RNA, (c) the number of genes per mESC, (d) Pearson’s correlation coefficient between measured and actual spike-in numbers, and (e) Pearson’s correlation coefficient between spike-in levels measured in distinct cells. In (a-e) violin plots represent the density distribution of the data. Overlaid boxplots show the median (white dot) and the interquartile range (box limits). The whiskers extend to the most extreme data point, within 1.5 times the interquartile range from the box. Outliers are indicated. The sample size in (a-e) 48 cells for each group from \(n=1\) experiments. (f, g) Dependence of sequencing efficiency on sequence composition. A regression was calculated of the average sequenced spike-in number on the actual spike-in number, setting the intercept to zero. The scatter plots show the dependence of the deviation of the measured spike-in level from the regression line, normalized by the average expression, on (f) percentage GC content and (g) sequence length. Data points for 96 ERCC\(^1\) spike-in sequences are shown in (f) and (g). Shown data are from one experiment.

Two different sorting schemes were used. In a first experiment with \( n = 2 \) mice (sorting scheme 1), we purified 384 Lineage- (Lin-) Kit<sup>hi</sup>Sca-1<sup>hi</sup> (LSK) cells with high surface expression of Kit and Sca-1 (encoded by Ly6a), a permissive sorting strategy to sample the pool of HSCs and multipotent progenitors. In addition, we sorted 768 lymphoid-primed multipotent progenitors (LMPP) as Flt3<sup>+</sup> LSK cells, and 384 Lin-Kit<sup>hi</sup>Sca-1<sup>lo</sup> as well as 192 Lin-Kit<sup>lo</sup>Sca-1<sup>lo</sup>/Flt3<sup>hi</sup> common lymphoid progenitors (CLP) in order to enrich for the lymphoid branch. We deliberately did not gate on Il7r to comprehensively sample CLPs<sup>hi</sup>. In order to capture more mature and...
potentially underrepresented progenitor states of all hematopoietic lineages we subsequently applied a second sorting strategy (sorting scheme 2). Here, we sequenced 1,152 cells from 16 adjacent windows spanning different ranges of Kit and Sca-1 expression and being either Flt3+ or Flt3−. Cells in (a) and (b) were sorted from n = 2 mice each in another independent experiment. Flow cytometry data are from one experiment with n = 1 mouse.


Supplementary Figure 3

Single-cell sequencing of index-sorted lymphoid progenitor populations.
(a) Boxplot of transcript count per cell distribution for LSK cells, LMPPs, and CLPs. The sample sizes are 370, 748, and 494 cells for LSK cells, LMPPs, and CLPs, sorted from two independent experiments with n = 2 mice. (b, c) Saturation analysis. (b) The number of UMIs detected in cells of the LMPP sample is shown as a function of the fraction of reads used for analysis. (c) The number of genes detected in cells of the LMPP sample is shown as a function of the fraction of reads used for analysis. (d) To benchmark the quality of our dataset, we compared Cd3+ LMPPs to mouse (and human) Cd3+ (CD34+) datasets generated with different sequencing technologies 1-3, including Smart-seq2 4, the commercial 10x GemCode technology and MARS-seq 5. Boxplot of UMI count distribution in Cd3+ (or CD34+) cells from different mouse (or human) hematopoietic progenitor datasets. Herman: Cd3+ LMPPs from this study (725 cells from n = 2 mice). Paul 2: Cd3+ cells from the common myeloid progenitor gate sequenced with MARS-seq (2,370 cells from n = 4 mice). Zheng (PB) 3: Cd3+ (mRNA-)positive cells from the CD34+ surface protein–positive peripheral blood population generated with 10x GemCode technology (3,213 cells from n = 1 donor). Zheng (BM) 3: Cd3+ cells from post-transplantation bone-marrow of an acute myeloid leukemia patient (AML027) generated with 10x GemCode technology (290 cells from n = 1 donor). Selection was always based on Cd34 (CD34) mRNA levels. (e) Boxplot of the number of detected genes per cell for the same samples as in (d) and a high-sensitivity dataset sequenced with a non-UMI based full-length coverage technology. Velten 1: CD34+ human bone marrow cells from individual 1 sequenced with Smart-seq2 (1,035 cells from n = 1 donor). In (a-e), bold line indicates the median and box limits represent interquartile range. The whiskers extend to the most extreme data point, within 1.5 times the interquartile range from the box. Outliers are indicated. (f-j) Shown are t-SNE maps highlighting fluorescence intensity measured by index-sorting to enable simultaneous quantification of the transcriptome and cell surface marker expression for (f) Kit, (g) Sca-1, (h) Flt3, (i) Il7r, and (j) Ly6d. In (f-j) 1,949 cells from n=4 animals are shown. (k) Barplot of Spearman’s correlation coefficient between surface protein expression quantified by fluorescence intensity and mRNA expression measured by single-cell RNA-seq for the same cell. The correlation coefficient was calculated for 1,949 cells from three independent experiments with n = 4 mice.

Supplementary Figure 4

Expression domains of lineage marker genes.

Shown are t-SNE maps highlighting log₂-transformed normalized transcript levels of **Kit** and **Ly6a** (multipotency), **Flt3** and **Il7r** (lymphoid lineage), **Rag1** and **Ebf1** (B cell lineage), **Irf8** and **Siglech** (plasmacytoid dendritic cells), **Cd79a** (preB cells), **Cd74** and **Itgax** (conventional dendritic cells), **Mpo** (granulocyte/monocyte lineage), **Elane** (neutrophils), **Car2**, **Gata1** and **Hbb-bs** (erythrocyte lineage), **Thy1** (innate lymphoid cells), **Ncr1** (natural killer cells), **Icos** and **Gata3** (innate lymphoid helper cell type 2 lineage), **Cd3d** (NKT cells), and **Pf4** (megakaryocyte lineage). All t-SNE maps show 1,949 cells from three independent experiments with n = 4 mice.
Supplementary Figure 5

Benchmarking of RaceID3.

RaceID3 shows superior performance in recovering domains of marker gene expression in comparison to RCA\(^1\), SC3\(^2\), and Seurat\(^3\), and ICGS\(^4\). RaceID3 was run with random forests-based reclassification (rf) and without. All clustering methods except for ICGS were
run with different parameters to change sensitivity and obtain different cluster numbers (see Online methods). By this strategy, overlapping ranges of cluster numbers were obtained for each method. Only ICGS does not have a parameter to allow adjustment of the sensitivity. Shown is the maximum log₂-transformed fold-enrichment (left panel) and the entropy of the distribution of average mean expressions (right panel) of a given lineage marker gene across all clusters detected as a function of cluster number. In comparison to all other tested methods, the fold-enrichment of the RaceID3 predictions is substantially higher for most lineage markers and of similar magnitude to the best performing methods for the other ones. At the same time, the entropy as a function of cluster number is consistently lower for most marker genes and follows a similar trend for the remaining ones, when comparing RaceID3 to the other methods. In conclusion, the benchmarking demonstrates that RaceID3 optimizes the overlap of known marker gene expression domains with predicted cell types.

Supplementary Figure 6

FateID analysis of mouse hematopoietic progenitors.

(a, b) The fate bias, corresponding to the probability of a cell to be assigned to a given lineage, is color-coded in the t-SNE map. The fate bias predicted by FateID (left) and STEMNET (middle) along with log₂-transformed aggregated normalized expression of two
lineage markers. Fate bias and marker gene expression is shown for (a) the conventional dendritic cell, and (b) the innate lymphoid lineage. In (a) and (b) data for 1,802 cells from n=4 animals are shown. (c) Scatterplot of normalized expression levels of Kit and Mpo. The predicted neutrophil fate bias is color-coded. Fate bias increases with Mpo expression and is inversely related to the level of Kit. (d) Scatterplot of normalized expression levels of Kit and Tcf4. The predicted pDC fate bias is color-coded. Fate bias increases with Tcf4 expression and is inversely related to the level of Kit. (e) Scatterplot of normalized expression levels of Kit and Ebf1. The predicted B cell fate bias is color-coded. Fate bias increases with Ebf1 expression and is inversely related to the level of Kit. (f) Scatterplot of normalized expression levels of Kit and Car2. The predicted erythrocyte fate bias is color-coded. Fate bias increases with Car2 expression and is inversely related to the level of Kit. (g) Scatterplot of normalized expression levels of Kit and Cd74. The predicted cDC fate bias is color-coded. Fate bias increases with Cd74 expression and is inversely related to the level of Kit. (h) Scatterplot of normalized expression levels of Kit and Tcf7. The predicted NK/NKT/ILC2 fate bias is color-coded. Fate bias increases with Tcf7 expression and is inversely related to the level of Kit. (i-n) Scatterplots comparing fate bias predicted by FateID and STEMNET for (i) the neutrophil, (j) the pDC, (k) the B cell, (l) the erythrocyte, (m) the cDC, and (n) the NK/NKT/ILC lineage. Although the predictions are overall correlated, STEMNET predicts rather uniform levels across a large fraction of the multipotent cell population, suggesting that each cell is multipotent, while FateID predictions sample the full range of possible values for all lineages, comprising cells with zero fate probability, intermediate values, or more substantial bias towards a given lineage. The higher resolution of FateID is potentially explained by the dynamic composition of the training set, which comprises more mature cells during the first iterations, while it includes earlier differentiation stages at later iterations to classify more naïve cells. In contrast to STEMNET, this strategy avoids classifying naïve cells with the help of genes expressed only at late stages of differentiation. In all panels data for 1,802 cells from three independent experiments with n = 4 mice are shown.

Classifier genes depend on differentiation stage.

(a-e) The heatmaps show genes with a random forests importance measure >0.02 and a ratio between the absolute importance and its standard deviation >2 for at least a single iteration. Iterations are depicted on the x-axis with the first iteration to the left and the final iteration to the right. Early iterations correspond to more mature stages while late iterations correspond to more naïve stages. A hierarchical clustering dendogram is indicated on the left margin. Heatmaps are shown for (a) the neutrophil, (b) the pDC, (c) the erythrocyte, (d) the cDC, and (e) the NK/NKT/ILC lineage. (f) Heatmap of Spearman’s correlation coefficient between the fate bias predicted by FateID and cell surface marker expression. An elastic-net regularized linear regression by the normal family as used by Velten et al. confirms the same trends, but the correlation-based analysis better discriminates the sub-population corresponding to distinct lineages. In (a-f) data derived from 1,802 cells from three independent experiments with n = 4 mice are shown.

Supplementary Figure 8

FateID identifies progenitor stages of the pDC lineage.

(a) Self-organizing map (SOM) of z-score-transformed pseudo-temporal expression profiles along the pDC developmental trajectory derived from the t-SNE map in Fig. 3a. Example profiles are shown for four genes dynamically expressed during pDC differentiation. The black line indicates a local regression. The SOM has been computed for 711 cells with predicted pDC fate bias >0.15 derived from three independent experiments with n = 4 mice. (b) Shown are t-SNE maps highlighting normalized transcript levels of $\text{Il}7r$, $\text{Cd}34$, $\text{Cs}f1r$ and $\text{Ly}6d$ for 1,802 cells from three independent experiments with n = 4 mice.

Nature Methods: doi:10.1038/nmeth.4662
FateID reveals fate bias in myeloid progenitors.

(a) A t-SNE map based on transcriptome similarity highlighting the origin of each cell is shown. The published dataset\(^1\) comprises common myeloid progenitors (CMP), Irf8-GFP^+MHC-II^+ dendritic cell progenitors, Flt3^+Csf1r^* monocyte progenitors, and Cd41^*.
megakaryocyte progenitors sorted from the CMP gate. (b) Shown is a t-SNE map highlighting clusters of cells with similar transcriptomes derived by RaceID3. Clusters 17, 1, 11, 5, and 8 were used as FateID target clusters for the erythrocyte, megakaryocyte, dendritic, monocyte, and granulocyte lineage. (c-g) The fate bias, corresponding to the probability of a cell to be assigned to a given lineage, is color-coded in the t-SNE map. The fate bias predicted by FateID (left) and STEMNET\textsuperscript{2} (middle) is shown along with log\textsubscript{2}-transformed normalized expression of a lineage marker. Fate bias and marker gene expression is shown for (c) the megakaryocyte, (d) the dendritic, (e) the granulocyte, (f) the monocyte, and (g) the erythrocyte lineage. (h) Barplot comparing Spearman’s correlation coefficient between the expression levels of early lineage markers and fate bias computed by FateID and STEMNET. Error bars correspond to standard errors of Fisher’s z-transformed correlation values calculated across all cells after removal of target clusters (1,927 cells from n=4 animals). P-values were derived from the difference of z-scores divided by the standard error assuming a standard normal distribution using William’s test (*P < 0.05, **P < 0.001). (i) Depicted is a t-SNE map highlighting expression of Flt3, showing that Flt3 does not discriminate between the monocyte and the dendritic cell lineage. (j-n) Shown are scatterplots comparing fate bias predicted by FateID and STEMNET for (j) the megakaryocyte, (k) the monocyte, (l) the granulocyte, (m) the dendritic cell, and (n) the erythrocyte lineage. Although the predictions are overall correlated, STEMNET predicts more uniform levels across a larger fraction of the multipotent cell population. In (a-g) and (i-n) for 2,370 cells from four independent experiments with n = 4 mice are shown.

FateID reveals fate bias of intestinal epithelial progenitors.

(a) Shown is a t-SNE map highlighting clusters of cells with similar transcriptomes derived by RaceID3 on single-cell RNA-seq data of intestinal epithelial cells. (b) Heatmap of log2-transformed averaged normalized expression across clusters. The cluster number and
color are indicated on the right. Only clusters with >3 cells were included. A hierarchical clustering dendogram is shown on the right margin. (c-g) The fate bias, corresponding to the probability of a cell to be assigned to a given lineage, is color-coded in the t-SNE map. The fate bias predicted by FateID (left) and STEMNET\textsuperscript{2} (middle) is shown along with log\_2-transformed aggregated normalized expression of two lineage markers. Fate bias and marker gene expression is shown for the (c) Paneth cell, (d) the goblet cell, (e) the enteroendocrine, (f) the enterocyte, and (g) the tuft cell lineage. In (a-g) data for 505 cells from n=3 animals are shown. (h) Barplot comparing Spearman’s correlation coefficient between the expression levels of early lineage markers and fate bias computed by FateID and STEMNET. Error bars correspond to standard errors of Fisher’s z-transformed correlation values calculated across all cells after removal of target clusters (303 cells from three independent experiments with n = 3 mice). P-values were derived from the difference of z-scores divided by the standard error assuming a standard normal distribution using William’s test (**P < 0.001, ***P < 0.001). (i-l) Shown is a t-SNE maps highlighting log\_2-transformed normalized transcript levels of (i) Neurog3, (j) Neurod1, (k) Muc2, and (l) Clca4. (m-q) Scatterplots comparing fate bias predicted by FateID and STEMNET for (m) the Paneth cell, (n) the goblet cell, (o) the enteroendocrine cell, (p) the enterocyte cell, and (q) the tuft cell lineage. Although the predictions are overall correlated, STEMNET predicts more uniform levels across a larger fraction of the multipotent cell population. In (a-g) and (i-q) data for 505 cells from three independent experiments with n = 3 mice are shown.

Supplementary Figure 11

Monocle 2 identifies major branches.

(a) Lineage tree inferred by Monocle 2 using reverse graph embedding. See Online methods for details. Cell types identified based on marker genes are highlighted. (b) Same as (a) but clusters inferred by Monocle 2 are highlighted. (c) Shown are Monocle 2 derived lineage trees highlighting log₂-transformed normalized transcript levels of Kit and Ly6a (multipotency), Dntt (lymphoid lineage), Ebf1 (B cell lineage), Siglech (plasmacytoid dendritic cells), Cd74 (conventional dendritic cells), Elane (neutrophils), Gata1 (erythrocye lineage) and Thy1 (innate lymphoid cells). Monocle 2, which is an established method for the derivation of multi-branched lineage trees failed to resolve Thy1-positive innate lymphocyte progenitors from B cell progenitors and distributed pDC progenitors across several branches. Furthermore, the fixed assignment of a cell to a branch does not predict residual bias to one or more alternative lineages. A probabilistic view on the process of cell fate decision is more appropriate for capturing transitions between cell states biased towards distinct fates, and quantification of co-existing multi-lineage bias is not achieved by available algorithms for the prediction of lineage trees. Monocle 2 results are shown for 1,802 cells from three independent experiments with n = 4 mice.

Supplementary Figure 12

Predicting the fate bias of human hematopoietic progenitors with FateID.

For this analysis, single cell RNA-seq data generated by Smart-seq2 for individual 1 from Velten et al.¹ were used. (a) t-SNE map showing clusters of cells with similar transcriptomes derived by RaceID3. (b) Heatmap of log2-transformed averaged normalized
expression of known marker genes across clusters. The cluster number and color are indicated on the right. Only clusters with >3 cells were included. A hierarchical clustering dendogram is shown on the right margin. (c) Correlation of predicted fate bias and cell surface marker expression measured by index-sorting for FateID (left) and STEMNET (right). Surface expression of CD135 (encoded by FLT3) and CD45RA discriminates progenitors of neutrophils, monocytes, pDCs, and lymphocytes, on the one hand, and eosinophils/basophils/mast cells, erythrocytes and megakaryocytes, on the other hand. Surface expression of the two markers is positively correlated to fate bias towards the former group of lineages and inversely correlated to fate bias towards the latter group of lineages. FateID predictions show a more pronounced difference between the two groups. (d) Shown are t-SNE maps highlighting log-transformed fluorescence intensity measured by flow cytometry (index-sorting) for CD135 (top) and CD35RA (bottom). (e) The fate bias predicted by FateID, corresponding to the probability of a cell to be assigned to a given lineage, is color-coded in the t-SNE map. The fate bias is shown for the B cell lineage (left) and the pDC lineage (right). The black circle marks a population of cells with enhanced fate bias towards pDCs or B cells. (f) Shown are t-SNE maps highlighting log-transformed normalized transcript expression of the B cell lineage marker VPREB1 (left) and the pDC lineage marker IRF8 (right). The expression domains of these markers overlap with the predicted domain of fate bias towards the respective lineage and lymphoid progenitors with enhanced bias towards the pDC lineage exhibit co-expression of the two markers (black circle). (g) The fate bias predicted by STEMNET shown for the B cell lineage (left) and the pDC lineage (right). Black circle: see (f). (h) Scatterplots comparing fate bias predicted by FateID and STEMNET for the pDC lineage (left) and the B cell lineage (right). Although the predictions are overall correlated (Spearman’s correlation coefficient is 0.74 for the B cell lineage and 0.67 for the pDC lineage), STEMNET predicts more uniform level across a larger fraction of progenitors. All panels show data from 1,035 cells sequenced from n = 1 donor.

Supplementary Figure 13

Sorting strategy for *in vitro* differentiation of B cell and pDC progenitors.

(a) Sorting strategy for B cell versus plasmacytoid dendritic cell biased progenitors and staining of cultured progenitors. Cells were gated using SSC-A versus FSC-A (not shown). Only single cells were considered further on using SSC-W versus SSC-H gating.
followed by FSC-W versus FSC-H gating (not shown). Thereupon only lineage negative (Lin−) cells were considered by exclusion of cells positive for the lineage markers TER-119, B220, Cd11b, Gr-1, SiglecH, Cd19 and Cd3ε (FITC). The next gating included only Flt3 positive (Flt3+) cells. The Flt3+ cells were plotted using Kit (BV510) versus Sca-1 (BV650) and cells with intermediate expression of Kit and Sca-1 were considered (common lymphoid progenitors, CLP). Finally CLP cells were plotted using IL-7R (BV421) and CD34 (PE). For the culture experiments and for single cell sequencing stringent gates were used to sort Cd34-IL7r+, Cd34+Il7r+ and Cd34+Il7r− cells. In addition, lymphoid-primed multipotent progenitors (LMPP) were sorted for culturing and for single cell sequencing. Fluorescence minus one controls were used for CD34 (PE) to set the sorting gates appropriately (not shown). The threshold for Il7r-positive cells was set according to the unstained control (not shown). (b) Exemplary flow cytometry plots of progenitors from one mouse after 7 days of culturing in either B cell medium or plasmacytoid dendritic cell (pDC) medium. The surface marker expression of SiglecH and Cd19 was assessed to check for lineage commitment towards the pDC or B cell lineage. Independent experiments were performed for n = 5 animals. Data for flow cytometry plot (a) is from one experiment with n = 3 mice, whereas data for flow cytometry plot (b) is derived from one experiment with n = 1 mouse.
(a) t-SNE map with a principal curve fitted to all cells within the progenitor clusters (1,2,3,7). (b) Heatmap of z-score transformed pseudo-temporal expression profiles of a number of multipotency and lineage marker genes. Cells were ordered along the principal curve in (a) and profiles were smoothened by a local regression. The bottom panel depicts a local regression of the fate bias using the same temporal ordering. Successive expression of multipotency markers Kit, Ly6a, Ifitm1, Cd34, Cd48, and Flt3 is consistent with the ordering of multipotent progenitor (MPP) stages MPP1 to MPP4 as previously inferred by bulk measurements\(^1,2\). Lineage markers comprise Gata2, Car2, Gata1, Pl4 for the erythrocyte/megakaryocyte lineage, Cebpa, Csf1r, Mpo for the granulocyte and monocye lineage, Itgax for the conventional dendritic cell lineage, and Irf8, Tcf4 for the plasmacytoid dendritic cell lineage, and Il7r, Rag1, Ebf1, Dntt for the B cell lineage. (c) Pictorial representation of the derived lineage tree. In (a) and (c) data are shown for 1,802 cells from 3 independent experiments with \(n = 4\) mice. In (b) profiles are shown for 1,416 cells from three independent experiments with \(n = 4\) mice. (d) Hematopoietic lineage tree inferred by StemID2. Only significant links are shown (\(P < 0.01\)). The color of the link indicates the \(-\log_{10}\) value. The color of the vertices indicates the entropy. The thickness indicates the link score reflecting how densely a link is covered with cells. The lineage tree is consistent with independently derived fate bias estimates: erythrocytes (cluster 9) branch off from cluster 1, while cluster 2 and 3 give rise to granulocytes/monocytes (cluster 17), and cluster 2 and 7 comprise progenitors of pDCs (cluster 12).
and B cells (cluster 10). (e) Barplot of StemID2 scores for hematopoietic clusters. Cluster 1, which shows highest expression of HSC markers, such as Ifitm1, receives the highest StemID2 score. In (a) and (b), only clusters with >5 cells were included and a link score cut-off of 0.5 was applied. The StemID2 in (d) and (e) computation has been performed on 1,802 cells from three independent experiments with n = 4 mice.


FateID recovers bipotent progenitor of monocytes and neutrophils.

(a) t-SNE map of the clusters identified by RaceID3 analysis on the murine hematopoietic transcriptome data from Olsson et al. The cell types described in the original study were recovered and are indicated next to the clusters. (b) t-SNE maps highlighting normalized expression of lineage-specific markers. (c) FateID fate bias predictions for all lineages. (d) Scatterplot for the comparison of fate bias towards the monocyte and the neutrophil lineage. Aggregated marker gene expression is highlighted for monocytes (left: Irf8, Csf1r, Ly86), for neutrophils (middle: Gfi1, Cebpe, S100a8) and the bi-potential progenitor (right: Ctsg, Elane, Mpo). The markers are taken from Figure 1c of Olsson et al. The FateID analysis reveals that lineage-specific marker gene expression coincides with uni-lineage bias, while progenitors with similar fate biases for monocytes and neutrophils express markers, which are expressed in both of the lineages as well as low levels of the lineage-specific markers. This is consistent with the interpretation of Olsson et al., that these cells represent bi-potent progenitors. We note that the bi-potency of this transitional state was validated by in vitro differentiation assays in the original study. All panels show data of 382 cells sequenced from n = 3 mice.

Supplementary Note 1

Comparing the performances of mCEL-Seq2 and manual CEL-Seq2
We tested a series of volume reductions to compare the performance of mCEL-Seq2 to the original, manual version of the CEL-Seq2. We sequenced 48 mouse embryonic stem cells (mESCs) each with 5-, 7- and 10-fold volume reduction using mCEL-Seq2 as well as the original, manual CEL-Seq2 protocol. Compared to manual CEL-Seq2 a slight decrease of 20% in the median total number of transcripts per cell was observed. Further volume reduction led to more substantial losses in sensitivity, mainly due to an increasing fraction of cells with low sequencing efficiency (Supplementary Fig. 1a). A minor decrease in sensitivity upon 5-fold reduction was also apparent from a drop of 11% in the median sequencing efficiency $\beta$ of external ERCC spike-in RNA$^2$ (Supplementary Fig. 1b). Consistent with previous findings, mCEL-seq2 yielded 2.7-fold higher sensitivity ($\beta=9\%$) compared to the original CEL-seq protocol$^3$, for which we had determined a sensitivity of 3.4% based on spike-in RNA$^4$. Since our previous quantification of the CEL-Seq sensitivity using single-molecule FISH yielded an estimate of 12.5%$^4$, we predict the true sensitivity of mCEL-Seq2 to be $\sim$34%. Further volume reduction, however, led to an increased number of samples with low spike-in sequencing efficiency, reflected by a bimodal distribution (Supplementary Fig. 1b). The median number of genes quantified per cell showed a similar behavior with a slight drop of 7% upon 5-fold reduction (Supplementary Fig. 1c). Next, we verified that the accuracy as quantified by Pearson’s correlation between the known log-transformed spike-in numbers per sample and the sequenced number of transcripts for each spike-in species remained high (median $\sim 0.90$) with an increasing number of outliers upon 7- and 10-fold reduction (Supplementary Fig. 1d). For our analysis, we decided to operate mCEL-Seq2 with a 5-fold volume reduction to maintain high accuracy and sensitivity.

We note that the sequenced numbers of spike-ins exhibit a much higher cell-to-cell correlation, with a median of 0.99 (Supplementary Fig. 1e). We verified that the lower correlation to the actual spike-in numbers is due to a dependence of the sequencing efficiency on the GC-content of the spike-ins: high GC-content led to lower sequencing efficiency (Supplementary Fig. 1f). The length of the spike-ins did not have any effect (Supplementary Fig. 1g).

A trained person can process up to eight 384-well plates with mCEL-Seq2 in a single day, i.e. $\sim$3,000 cells. A 5-fold volume reduction implies an almost 5-fold reduction of library preparation costs down to a level comparable with library preparation costs of MARS-Seq$^5$, a published plate-based high-throughput scRNA-seq method. Since CEL-Seq2 outperforms similar methods in terms of sensitivity and accuracy$^6$ our mCEL-Seq2 protocol represents an ideal approach for plate-based sensitive scRNA-seq. Compatibility of the scRNA-seq protocol with cell sorting is essential for our purposes, since rare, transitional states would be underrepresented in a population of randomly extracted cells and pre-sorting of rare cells frequently does not yield enough input material for more cost-efficient high-throughput approaches based on nanoliter droplets$^7,8$. 

Nature Methods: doi:10.1038/nmeth.4662
**Supplementary Note 2**

**RaceID3 detects cell types with high specificity**
To benchmark the performance of RaceID3 we compared its specificity and sensitivity to four other methods, i.e. Seurat$^8$, SC3$^9$, and RCA$^{10}$, and ICGS$^{11}$. Seurat is an established algorithm, which successfully discriminates cell types in big data sets comprising thousands of cells at relatively low sequencing depths as generated by nanoliter droplet based sequencing$^{7,8}$. SC3 and RCA are two recently published methods, which exhibit excellent performance in comparison to a number of published algorithms$^{9,10}$. ICGS successfully resolved hematopoietic progenitor states in a previous analysis$^{11}$. For benchmarking of RaceID3 we utilize prior knowledge on established marker genes for distinct hematopoietic lineages present in our sorted populations. We selected two markers for each of these lineages (Fig. 1d) and assume that the expression domain of these markers represents a ground truth delineating the associated cell type. We then computed for every cluster the fold-enrichment of each marker, defined by the ratio of mean expression within a given cluster and the mean expression across all other clusters. We then extracted for each marker the maximum fold enrichment obtained in any of the clusters. High fold-enrichment indicates that the expression domain can be discriminated from background. To assess specificity, we computed the entropy of the relative mean expression across all clusters. As input probability we used the relative mean expression of a cluster, obtained by dividing the mean expression of a cluster by the sum of the mean expressions across all clusters. Low entropy represents concentration of the gene expression domain in a single or very few clusters, while high expression indicates more uniform expression across clusters. For this comparison, RaceID3 was run with or without random forests-based reclassification. While reclassification reduces the number of outlier clusters, facilitating the interpretation of the identified cell types and states, it also results in a decrease in sensitivity and specificity as revealed by our analysis. We ran each clustering method with varying cluster number to obtain results for overlapping ranges of cluster numbers between the methods (Supplementary Fig. 5 and Online methods). While a high cluster number favors high fold-enrichment, a low cluster number in general achieves lower entropy values. This analysis revealed that RaceID3, with or without reclassification, achieves consistently higher fold-enrichment and/or lower entropy for the majority of all cell type-specific markers (Supplementary Fig. 5).
**Supplementary Note 3**

**FateID quantifies fate bias of common myeloid progenitors**

We tested FateID on recently published single cell transcriptome data of common myeloid progenitors (CMP)\(^2\). In this study, myeloid progenitors were sorted as LIN\(^-\) Kit\(^+\) Sca-1\(^-\) cells. Utilizing cell surface marker expression of specific sub-populations, progenitors of the megakaryocyte, monocyte, and dendritic cell lineage were enriched as Cd41\(^+\), Flt3\(^+\)Csfr\(^+\), and Irf8\(^+\)MHC-II\(^+\) populations, respectively (Supplementary Fig. 9a).

We performed RaceID3 analysis of this dataset, and could recover distinct clusters of cells with markers of granulocytes, erythroblasts, megakaryocytes, dendritic cells, and monocyte precursors (Supplementary Fig. 9b). FateID analysis was performed with target clusters 1, 5, 8, 11, and 17, for the megakaryocyte, monocyte, neutrophil, dendritic cell and erythrocyte cluster on the basis of expression of Pf4, Csfr, Elane, Cd74 and Hba-a2, respectively. The subpopulations with lineage-specific fate bias overlap well with the sorted Cd41\(^+\), Irf8\(^+\)MHC-II\(^+\) or Flt3\(^+\)Csfr\(^+\) populations of megakaryocytes, dendritic progenitors or monocytes, respectively, and with populations delineated by early lineage markers (Supplementary Fig. 9c-g). STEMNET inferred overall comparable fate bias (Supplementary Fig. 9h), but failed to discriminate Irf8\(^+\)MHC-II\(^+\) dendritic progenitors from Flt3\(^+\)Csfr\(^+\) monocyte progenitors. Although STEMNET suggests that Flt3 expression correlates with monocyte fate bias, its expression domain does not discriminate between the monocyte and the dendritic cell lineage (Supplementary Fig. 9i). Moreover, STEMNET exhibits a bi-modal distribution of fate bias with few cells of intermediate fate bias, while FateID infers a more continuous distribution of fate biases (Supplementary Fig. 9j-n).
Supplementary Note 4

FateID reveals fate bias of intestinal epithelial cells
To test FateID on a different cell type, we analyzed fate bias for a published dataset of intestinal epithelial cells\(^\text{13}\). These data comprise Lgr5\(^+\) intestinal stem cells and all mature cell types including absorptive enterocytes and rare secretory cells, i.e. Paneth cells with antimicrobial and niche function, mucus-producing goblet cells, hormone-secreting entero-endocrine cells, and tuft cells sensing the lumen\(^\text{14}\). Secretory cells are over-represented in this dataset due to enrichment of Cd24\(^+\) cells. RaceID3 recovers the previously identified sub-types (Supplementary Fig. 10a,b). To infer fate bias of the five main branches irrespective of segregation into sub-types, we ran FateID with target clusters inferred by the algorithm given a set of lineage specific markers (Lyz1 for Paneth cells, Clca3 for goblet cells, Chgb for entero-endocrine cells, Dclk1 for tuft cells, Alpi for enterocytes). The predicted fate bias recapitulates domains of progenitor marker gene expression and delineates the origin of distinct lineages (Supplementary Fig. 10c-h). For example, Neurog3\(^+\) and Neurod1\(^+\) positive progenitors of enteroendocrine cells are discriminated from Muc2\(^+\) cells of the goblet and Paneth cell progenitors by their fate bias (Supplementary Fig. 10i-k). Moreover, the fact that most Lgr5\(^+\) and Clca4\(^+\) cells are biased towards enterocytes (Supplementary Fig. 10l) is not unexpected, given that this cell type outnumbers all other lineages by far. Although STEMNET infers comparable results to FateID (Supplementary Fig. 10h), the dynamic range of STEMNET is very small and an almost uniform base level of 10% to 20% fate bias is predicted for all cells that do not belong to the branch with elevated fate bias (Supplementary Fig. 10m-q). In conclusion, FateID successfully quantifies fate bias across different cell types and achieves better resolution than STEMNET.
Supplementary Note 5

FateID suggests a common progenitor population of B cells and pDCs in human

In order to compare the mouse hematopoietic progenitor populations to their counterpart in human, we re-analyzed data from a recent study\(^1\) (Supplementary Fig. 12a,b). As one of the core findings, which was extensively validated by in vivo and in vitro differentiation experiments, surface expression of CD135 (encoded by FLT3) and CD45RA discriminated lymphoid/myeloid- from megakaryocyte/erythrocyte-primed progenitors. FateID analysis confirmed this segregation, but showed a more pronounced separation of the two groups (Supplementary Fig. 12c). Moreover, the correlation of the predicted FateID bias to surface marker expression leads to co-clustering of eosinophil/basophil/mast cell progenitors with megakaryocyte/erythrocyte progenitors. This finding, which is not apparent from the STEMNET predictions, is supported by co-expression of common transcription factors (GATA2 and TAL1) and consistent with a recently described early bifurcation into Gata1-positive erythrocyte/megakaryocyte/eosinophil/mast cell progenitors and Gata1-negative monocytes/neutrophils/lymphocytes in the murine system\(^16\).

Moreover, RaceID3 could discriminate two transcriptionally similar lymphoid progenitor clusters, i.e. cluster 3 and 8, which specifically up-regulated genes of the B cell and pDC lineages, respectively, such as EBF1 and IRF8. Both clusters co-express lymphoid genes, such as DNTT and IL7R, akin to the common progenitor cluster observed in the mouse data. Within this population, FateID discriminated cells with more pronounced bias towards either B cells or pDCs, depending on expression of IRF8 (Supplementary Fig. 12d-f). Finally, FateID predicted increased B cell bias within a more naïve population, coinciding with low expression of the preB cell receptor component VPREB1. This remained undetected by STEMNET, which showed a more uniform fate bias in early progenitors (Supplementary Fig. 12g,h).

In summary, FateID analysis supports the existence of a common differentiation pathway of B cells and pDCs also in human.
REFERENCES


