nature biotechnology

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41587-019-0068-4

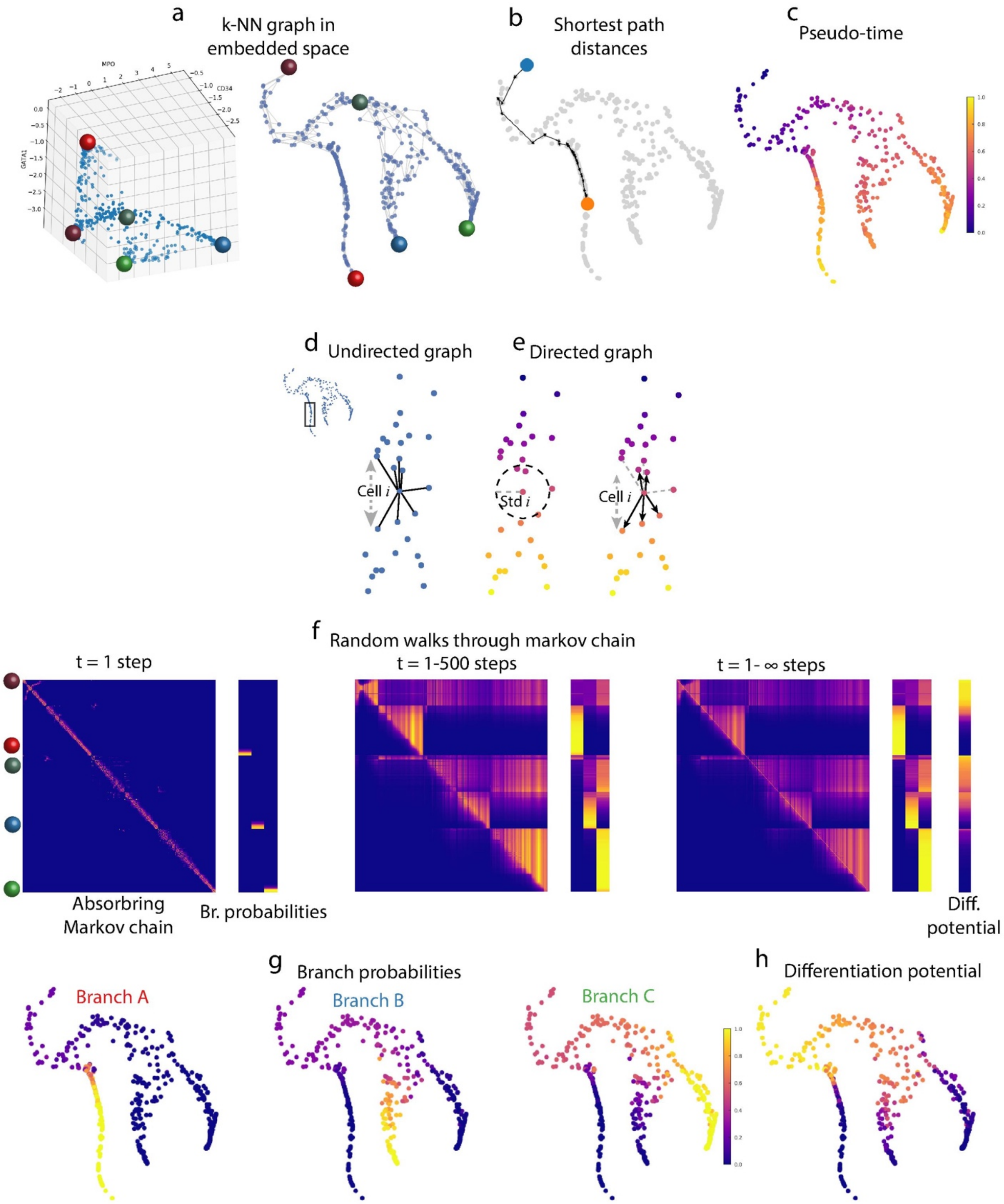
In the format provided by the authors and unedited.

Characterization of cell fate probabilities in single-cell data with Palantir

Manu Setty, Vaidotas Kiseliovas, Jacob Levine, Adam Gayoso, Linas Mazutis and Dana Pe'er*

Program for Computational and Systems Biology, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
*e-mail: peerd@mskcc.org

Supplementary figures



Supp. Fig. 1: Palantir algorithm outline

Illustration of steps in the Palantir algorithm, using the same data as Fig 1b.

(a) High dimensional representation of the data, each dot represents a cell plotted based on expression of CD34 (x-axis), MPO (y-axis) and GATA1 (z-axis) (left panel). Right panel: Same tSNE plot as Fig. 1b generated using the diffusion components of the cells in the left panel. Plots show the projection of 463 cells.

(b) Illustration of shortest path from blue cell to the orange cell.

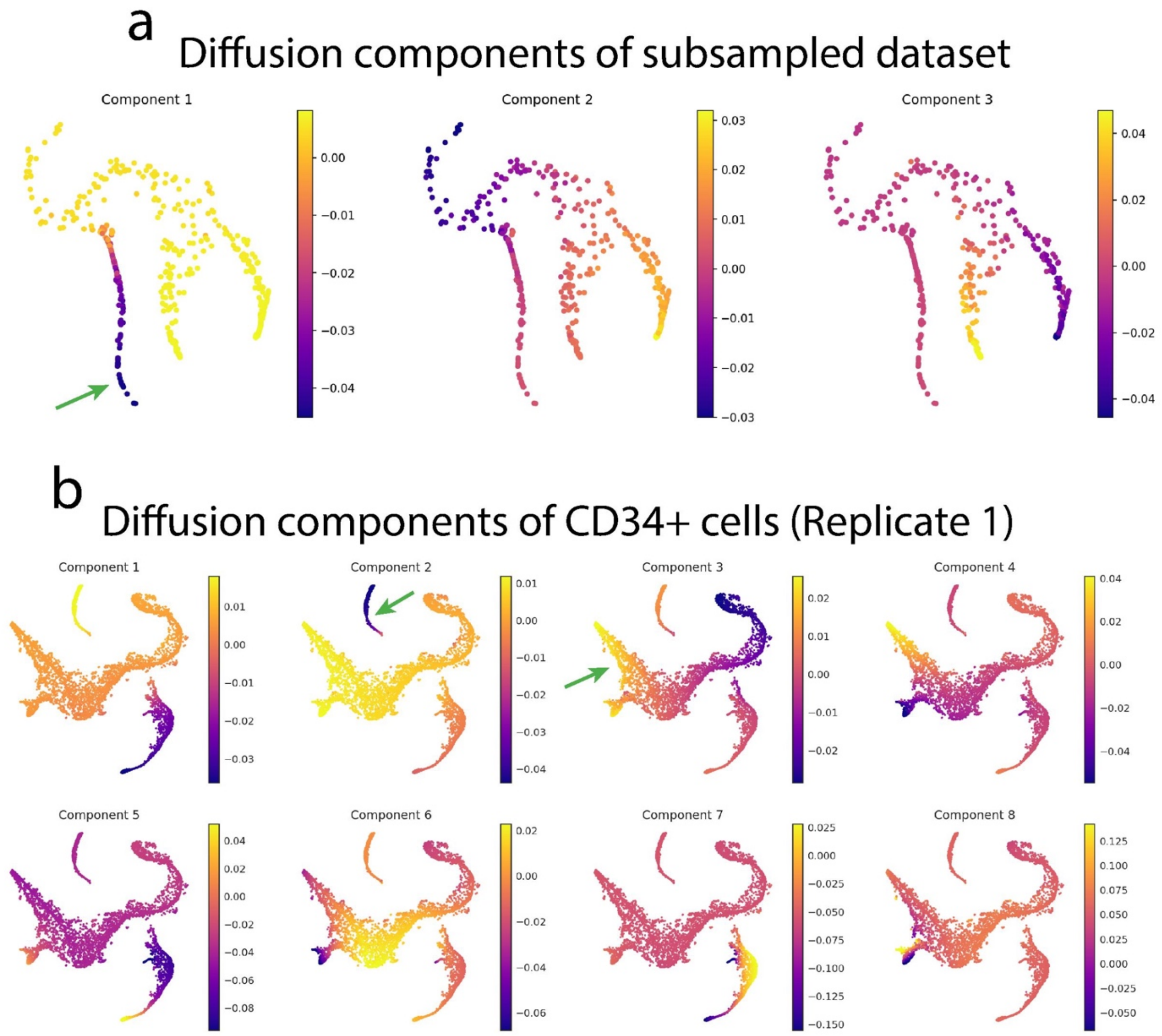
(c) Cells colored by Palantir pseudo-time.

(d-e) Illustration of Markov chain construction.

(d) Edges in the undirected graph can take cells both forward and back along pseudo-time. (e) The scaling factor associated with each cell (Equation 2) can be used as measure of uncertainty in the pseudo-time estimate (left panel). Edges that go backward beyond the pseudo-time uncertainty are pruned and the retained edges are converted to directed edges (right panel).

(f) Heatmaps showing the evolution of absorbing Markov chain and branch probabilities for random walks of different lengths. (Left panel: 1 step, middle panel: 1...500 steps and right panel: $1\ldots\infty$ steps). For each panel, rows and columns in the Markov chain heatmap represent all non-terminal cells ordered by Palantir pseudo-time. The value (i, j) represents the probability of cell i reaching cell j in the specified number of steps. Rows in the branch probabilities heatmap represent non-terminal cells and the columns represent the terminal states. The value (i, j) represents the probability of cell in non-terminal state i reaching terminal state j in the specified number of steps. The position of the individual cells highlighted in 1a are shown on the left.

(g - h) Cells colored by Palantir branch probabilities and differentiation potential.

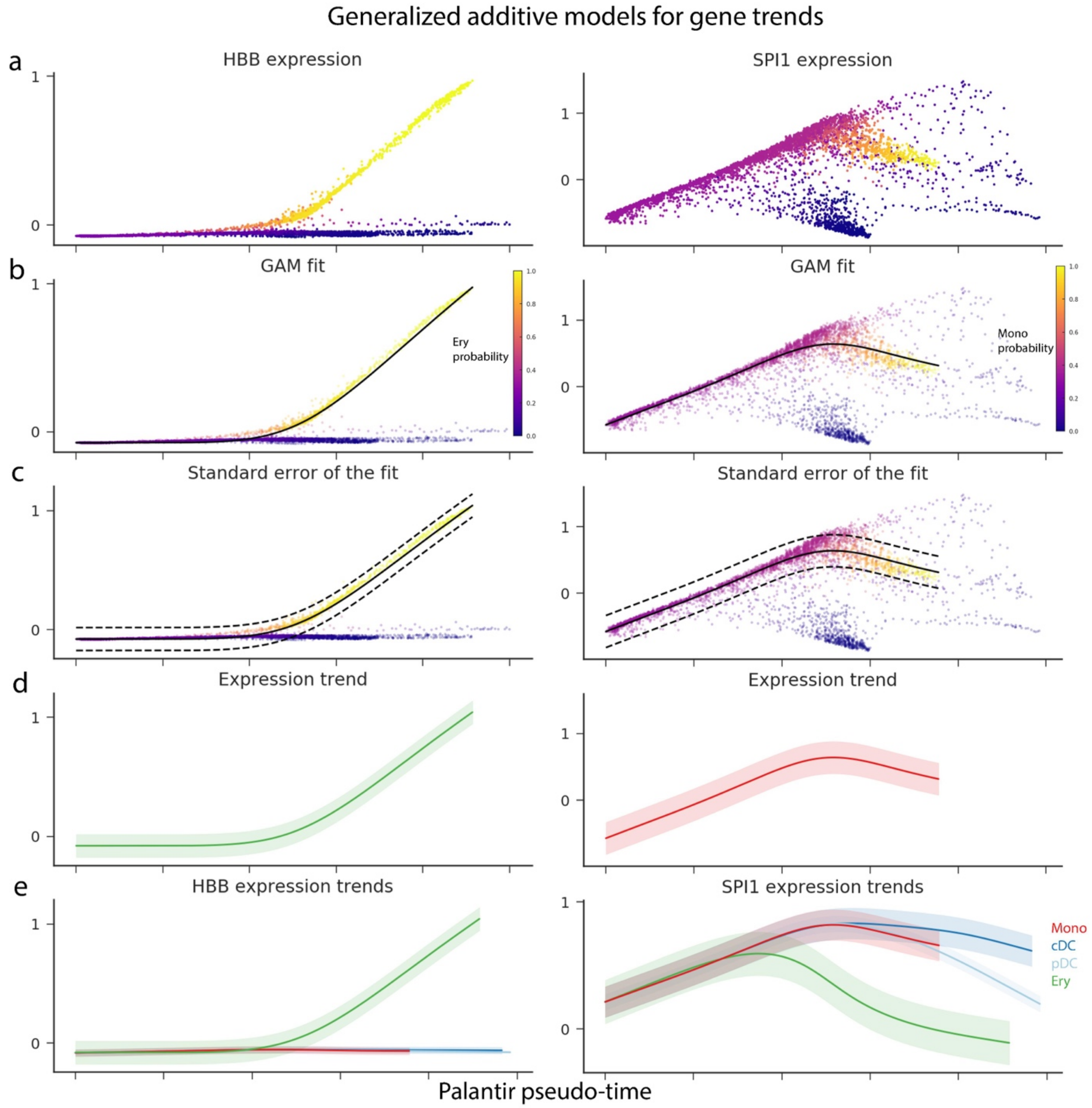


Supp. Fig. 2: Diffusion components are not sufficient to represent pseudo-time for all lineages

(a) tSNE plots of the subsampled dataset used in Fig. 1 and Supp.Fig. 1, colored by diffusion components. Plots show the projection of 463 cells.

(b) tSNE plots for CD34+ cells presented in figure 2, colored by diffusion components.

Green arrows indicate the lineages for which ordering can be determined using a single component, whereas the ordering of the remaining lineages requires two or more components. Plots show the projection of 5780 cells.



Supp. Fig. 3: Characterizing gene expression dynamics. The characterization of gene expression dynamics is illustrated with two examples: HBB, a gene expressed specifically in the erythroid lineage (left panels) and SPI1, a gene with higher expression in the monocytic lineage (right panels)

(a) Plots showing the MAGIC [1] imputed expression (y-axis) of HBB and SPI1 respectively along Palantir pseudo-time (x-axis). Cells are colored by the erythroid and monocyte branch probabilities respectively. Expression patterns are shown for 5780 cells.

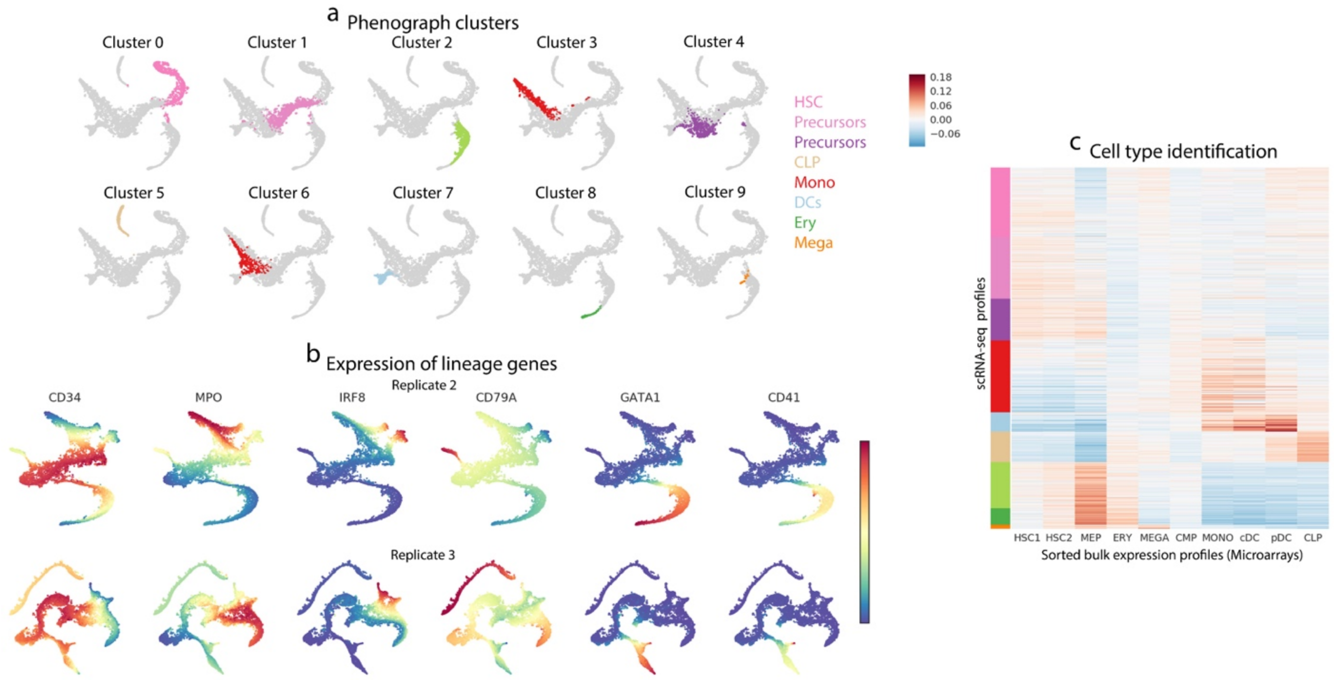
(b) Same as (a) with the trend fit computed using Generalized Additive Models (GAMs) [2] shown in black.

Each cell is weighted by the branch probability and thus no pre-selection of cells is necessary for computing trends along a particular lineage.

(c) Same as (b) with standard deviations of the fit shown in dotted lines.

(d) The expression trends are represented as a smooth fit with the standard deviation of the fit shown in a lighter shade.

(e) Gene expression trends for HBB and SPI1 for the erythroid and myeloid lineages. For any gene, trends can be computed across all lineages since Palantir determines a single pseudo-time across lineages. Solid line represents the fitted trend estimate and shaded region represents standard deviation.

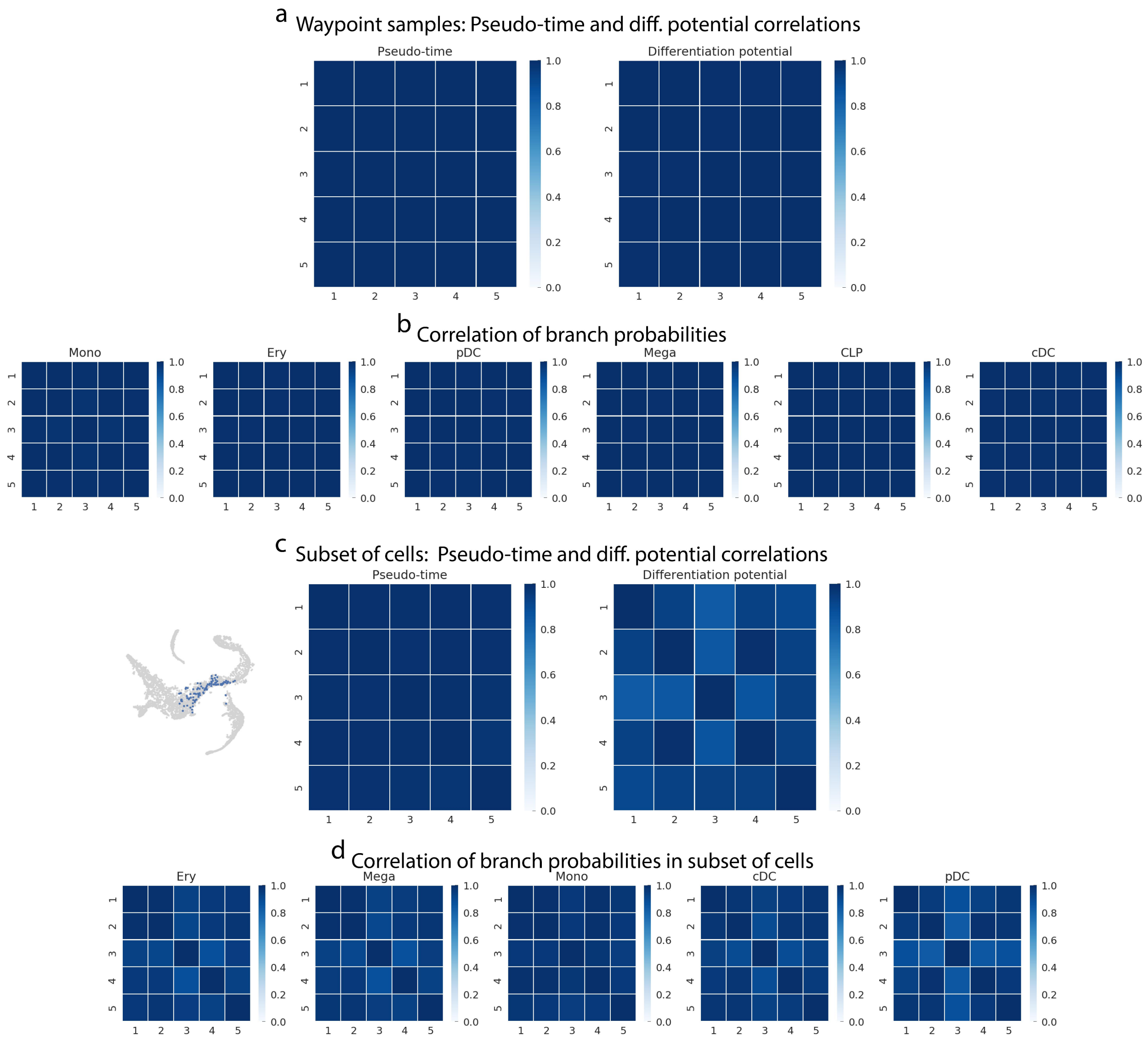


Supp. Fig. 4: Cell lineages identified in CD34+ human bone marrow cells

(a) Replicate 1 CD34+ cells from human bone marrow colored by Phenograph clusters using the scheme presented in Fig. 2b. 5780 cells are shown on the tSNE map.

(b) Replicate 2 and 3 cells, colored by expression of lineage characteristic genes (Fig. 2f) demonstrating that the spectrum of lineages identified from CD34+ bone marrow cells is consistent across three independent human donors. 6501 and 12046 cells are shown on the tSNE map for replicates 2 and 3 respectively.

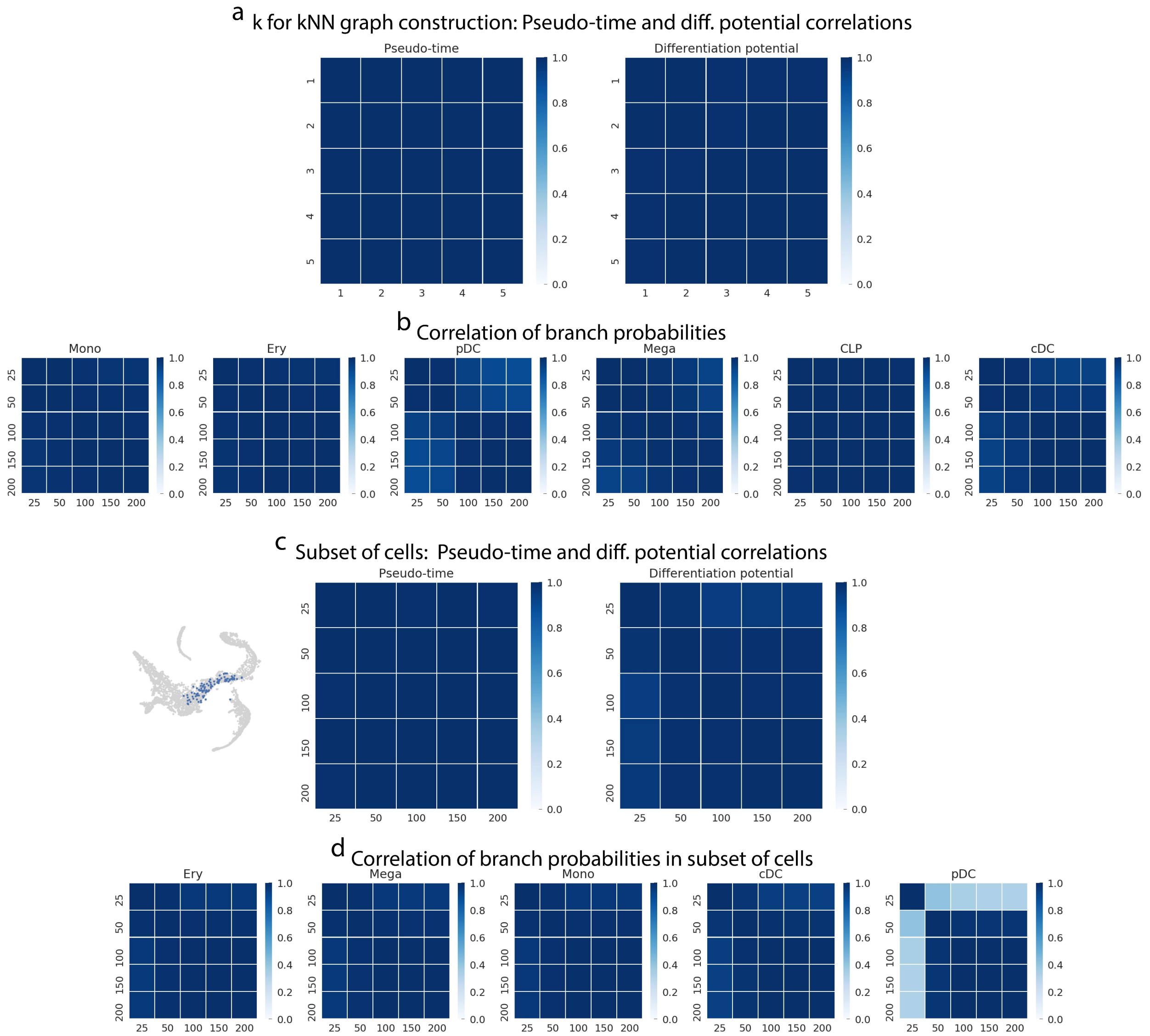
(c) Heatmap of the correlation between bulk sorted expression profiles generated using microarrays and scRNA-seq profiles (Replicate 1). Rows represents single cell and columns represent bulk samples. Cells are ordered as in Fig. 2a. Median expression profiles from cell clusters (a) were correlated with bulk expression profiles to annotate clusters with cell types using Pearson correlation. Heatmap shows the correlation of 5780 cells with 10 averaged bulk expression profiles.



Supp. Fig. 5: Palantir results are robust to different waypoint samplings

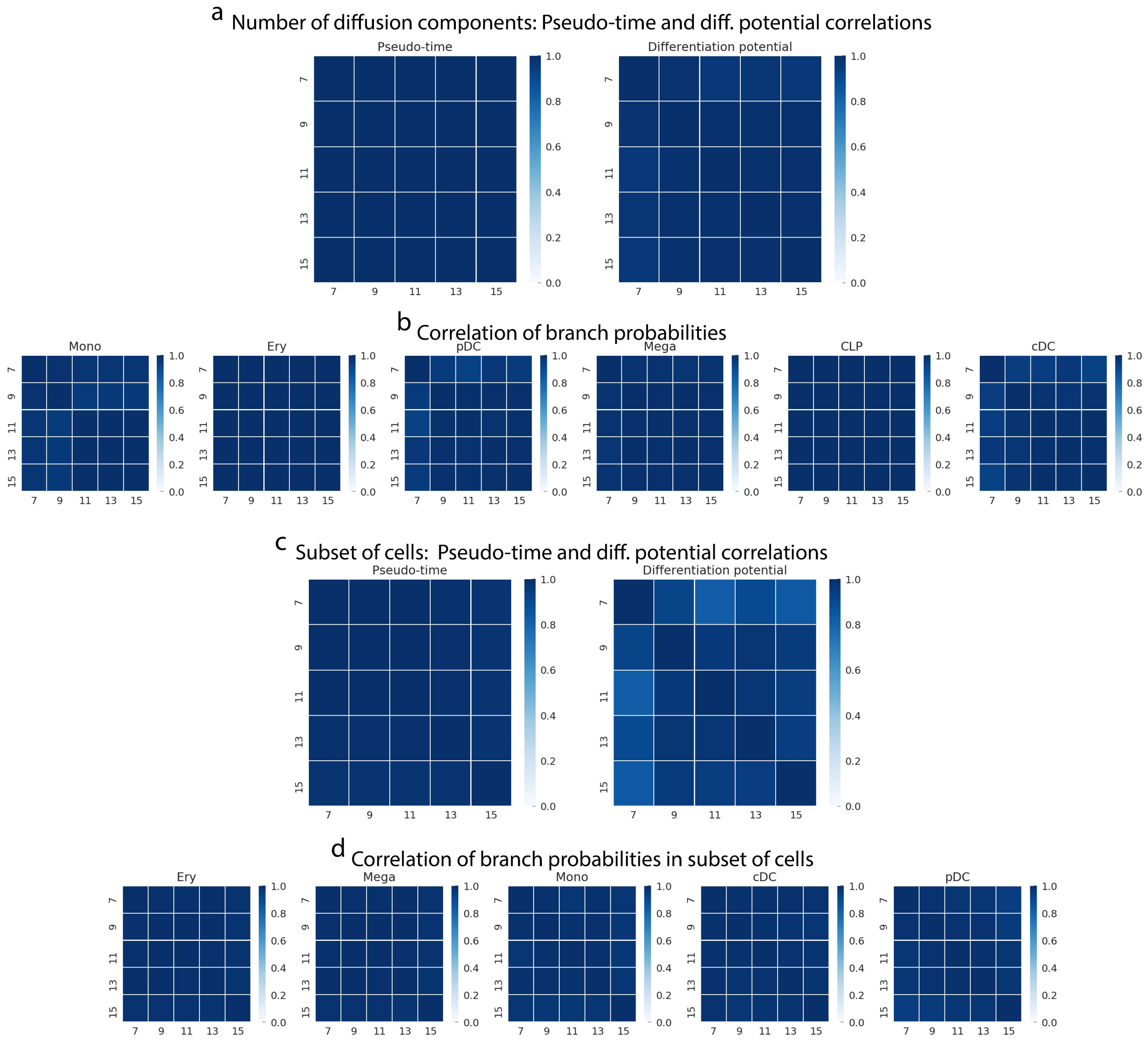
Palantir robustness was measured by testing a range of different parameters using replicate 1 as the test set. The results between two different runs were compared by determining the Pearson correlation between pseudo-time, differentiation potential and branch probabilities. Each heatmap represents the correlation of either pseudo-time, DP or BP between a pair of Palantir runs.

(a) Pearson correlations of pseudo-time orderings and differentiation potentials for different waypoint samplings for all cells. (b) Correlations for branch probabilities for all cells. Correlations were computed using 5780 cells. (c) Left: a subsample of cells from the middle of the differentiation trajectory. (c-d) Same as (a-b), for the subset of cells showing (c). Correlations were computed using the 100 subsampled cells.



Supp. Fig. 6: Palantir results are robust to k, the number of neighbors for kNN graph construction

(a-d) Same as Supp. Fig. 5



a Number of diffusion components: Pseudo-time and diff. potential correlations

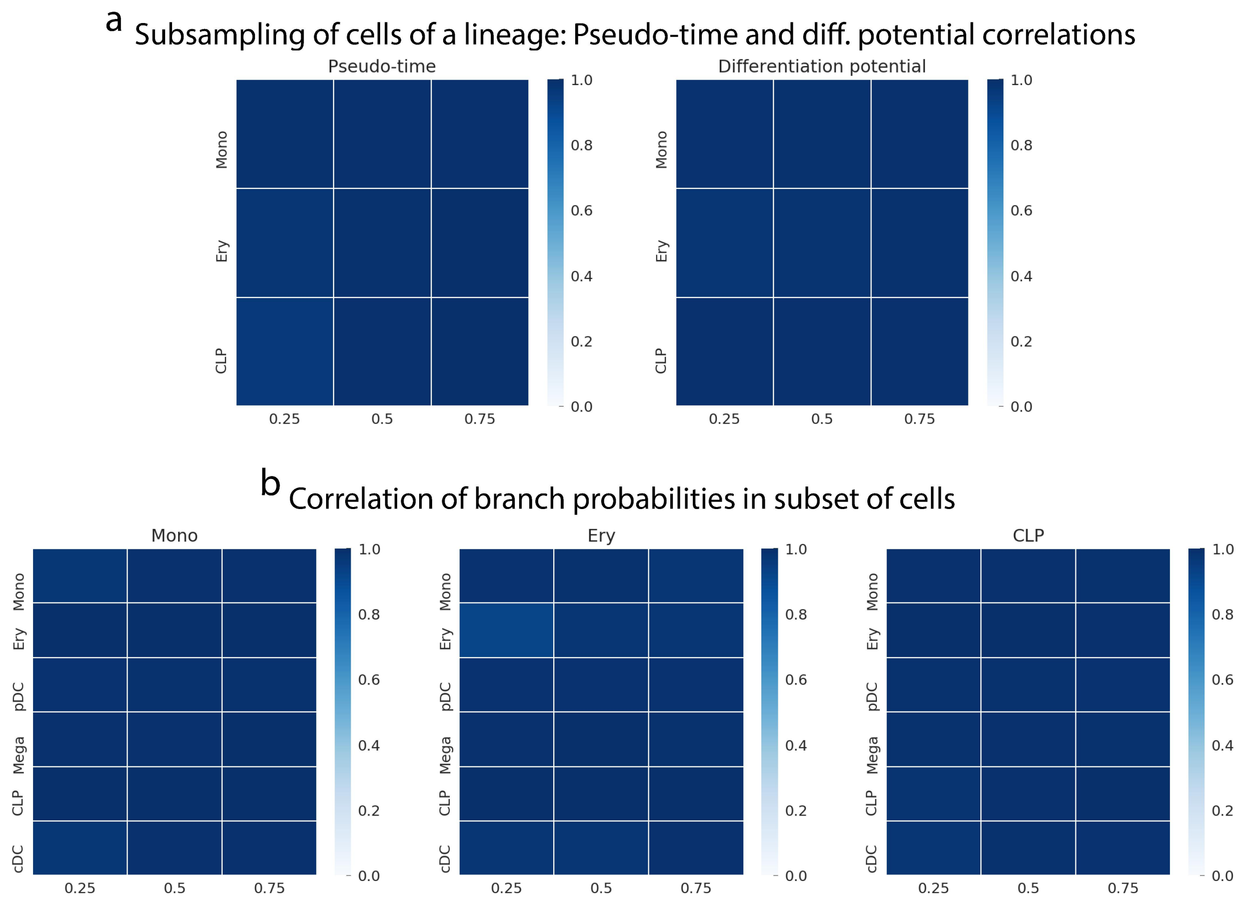
b Correlation of branch probabilities

c Subset of cells: Pseudo-time and diff. potential correlations

d Correlation of branch probabilities in subset of cells

Supp. Fig. 7: Palantir results are robust to number of diffusion components

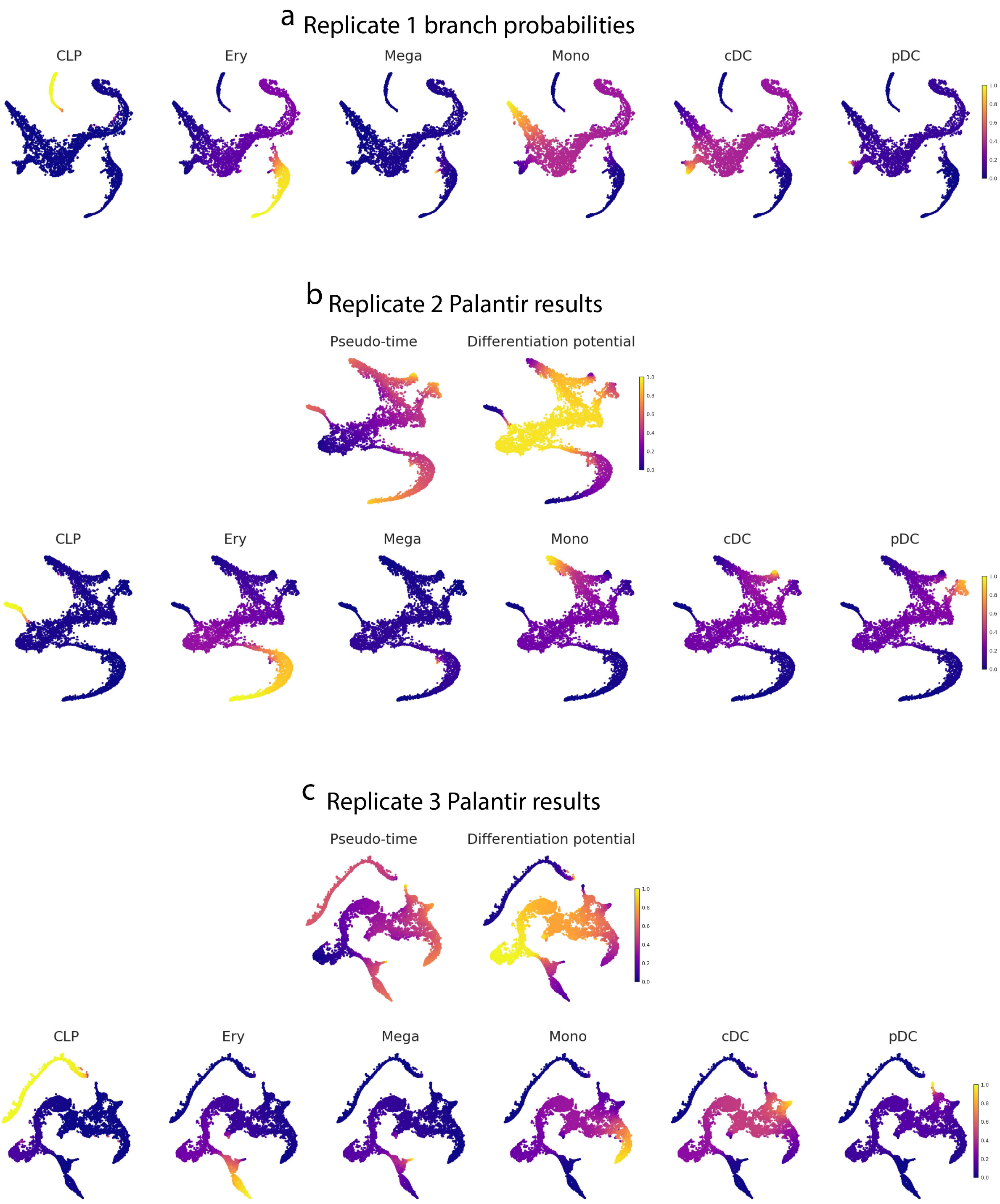
(a-d) Same as Supp. Fig. 5


a Subsampling of cells of a lineage: Pseudo-time and diff. potential correlations

b Correlation of branch probabilities in subset of cells

Supp. Fig. 8: Palantir results are robust to subsampling of cells in different lineages

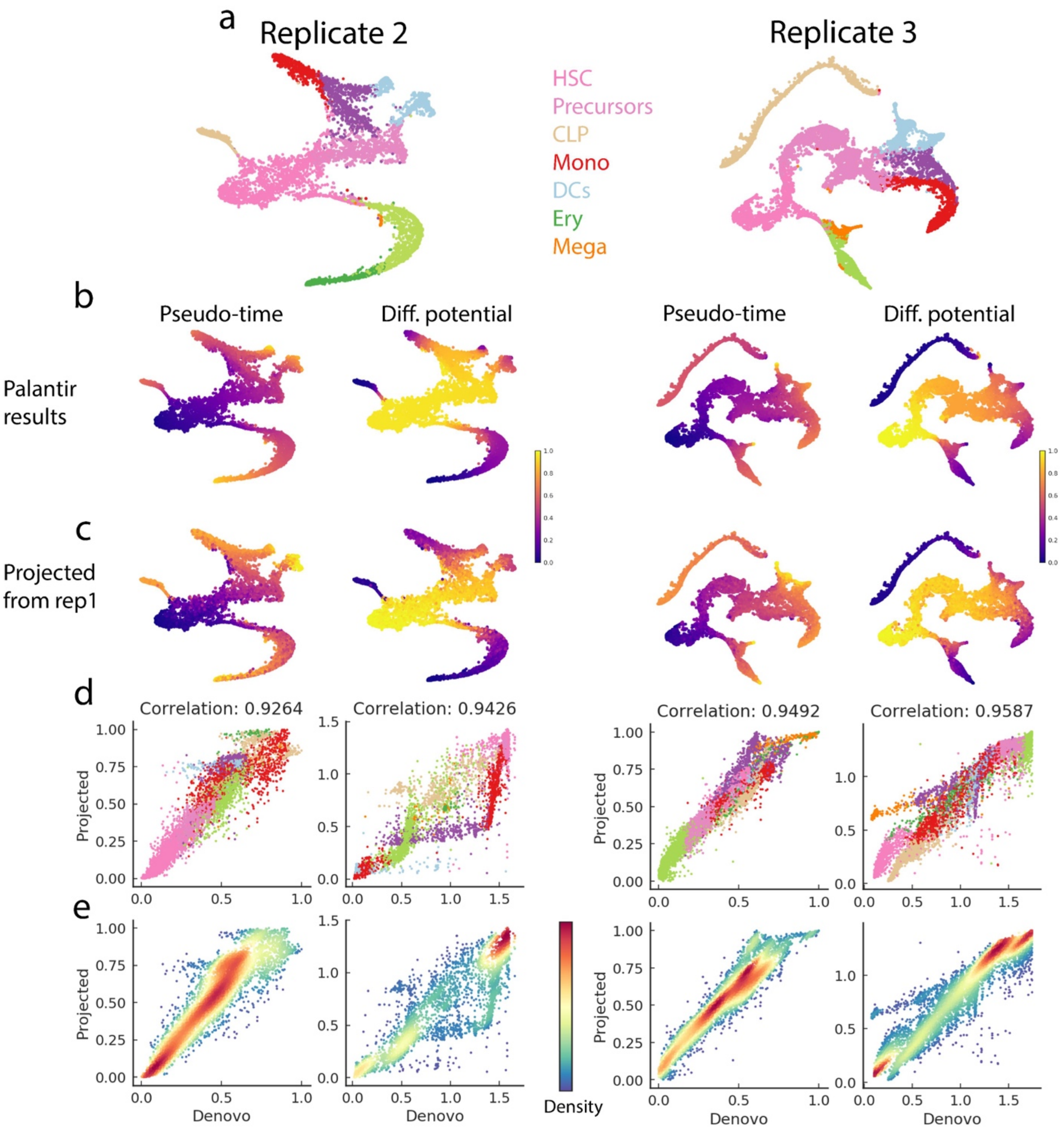
Pearson correlation of pseudo-time, DP (a) and branch probabilities (b) by sub-sampling the specified proportion of cells from the respective lineage. Correlations were computed using 1445, 2890 and 4335 cells representing 25%, 50% and 75% of the total cells.


a
Replicate 1 branch probabilities
CLP
Ery
Mega
Mono
cDC
pDC
1.0
0.8
0.6
0.4
0.2
0.0
b
Replicate 2 Palantir results
Pseudo-time
Differentiation potential
1.0
0.8
0.6
0.4
0.2
0.0
CLP
Ery
Mega
Mono
cDC
pDC
1.0
0.8
0.6
0.4
0.2
0.0
c
Replicate 3 Palantir results
Pseudo-time
Differentiation potential
1.0
0.8
0.6
0.4
0.2
0.0
CLP
Ery
Mega
Mono
cDC
pDC
1.0
0.8
0.6
0.4
0.2
0.0

Supp. Fig. 9: Palantir is reproducible across human bone marrow replicates

(a) Cells plotted on tSNE based on diffusion components and colored by Palantir branch probabilities for Replicate 1. 5780 cells are shown on the tSNE map.

(b-c) Cells plotted on tSNE based on diffusion components and colored by Palantir results: pseudo-time, differentiation potential and branch probabilities for replicates 2 and 3. 6501 and 12046 cells are shown on the tSNE maps for replicates 2 and 3 respectively.


a
Replicate 2
Replicate 3
HSC
Precursors
CLP
Mono
DCs
Ery
Mega
b
Pseudo-time
Diff. potential
Pseudo-time
Diff. potential
Palantir results
c
Projected from rep1
1.0
0.8
0.6
0.4
0.2
0.0
d
Correlation: 0.9264
Correlation: 0.9426
Correlation: 0.9492
Correlation: 0.9587
Projected
1.00
0.75
0.50
0.25
0.00
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
e
Denovo
Density

Supp. Fig. 10: Reproducibility of Palantir pseudo-time and differentiation potential across replicates

(a) tSNE plots highlighting the different cell populations in replicate 2 (left) and replicate 3 (right). Cells are colored by Phenograph clusters and colors were chosen to maintain consistency with replicate 1 (Fig. 2b). Replicates 2 and 3 contain 6501 and 12046 cells respectively.

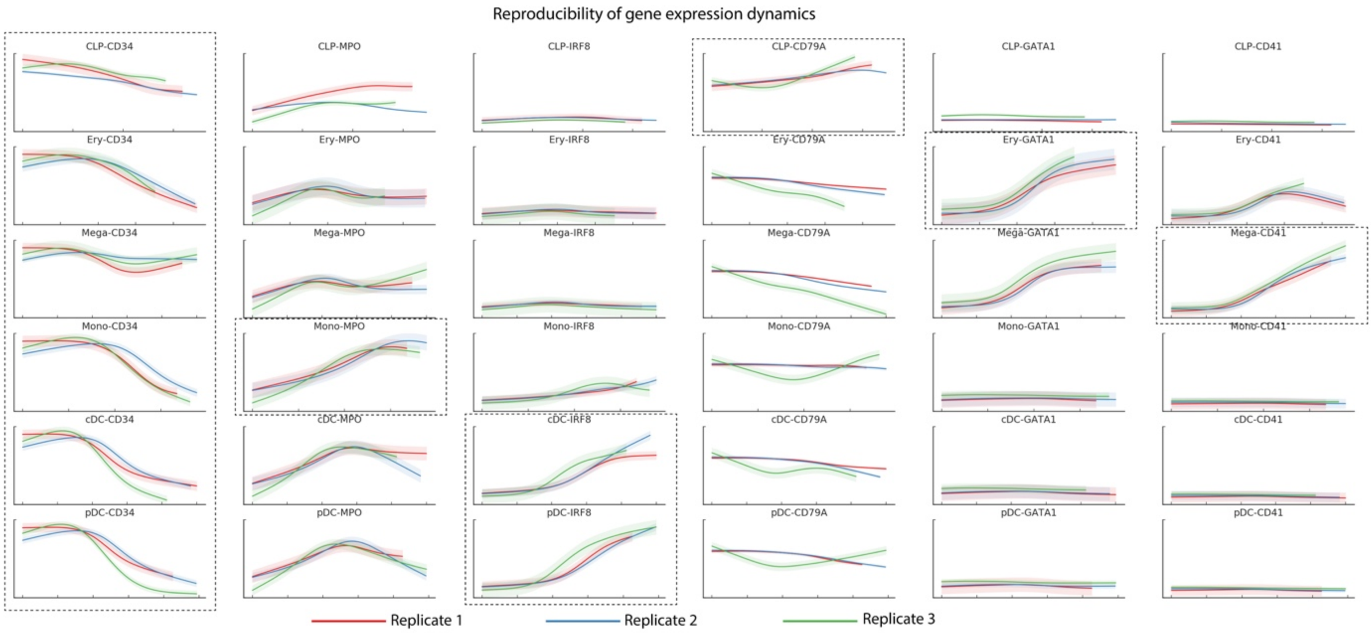
(b) *De-novo* Palantir results for replicates 2 and 3 generated using one of the HSCs as the start cell.

The reproducibility of Palantir results was further tested by projecting Palantir results of one replicate to a second replicate and comparing the projected results with those generated *de-novo* using the second replicate.

(c) Replicate 1 Palantir results projected onto the replicates 2 and 3. The projections were determined mutually nearest neighbors between replicate 1 and replicate 2 (or 3), see methods. The pseudo-time and differentiation potential of cells in replicate 2 (or 3) were computed as weighted average of the pseudo-time and differentiation potential respectively of replicate 1 mutually nearest neighbors.

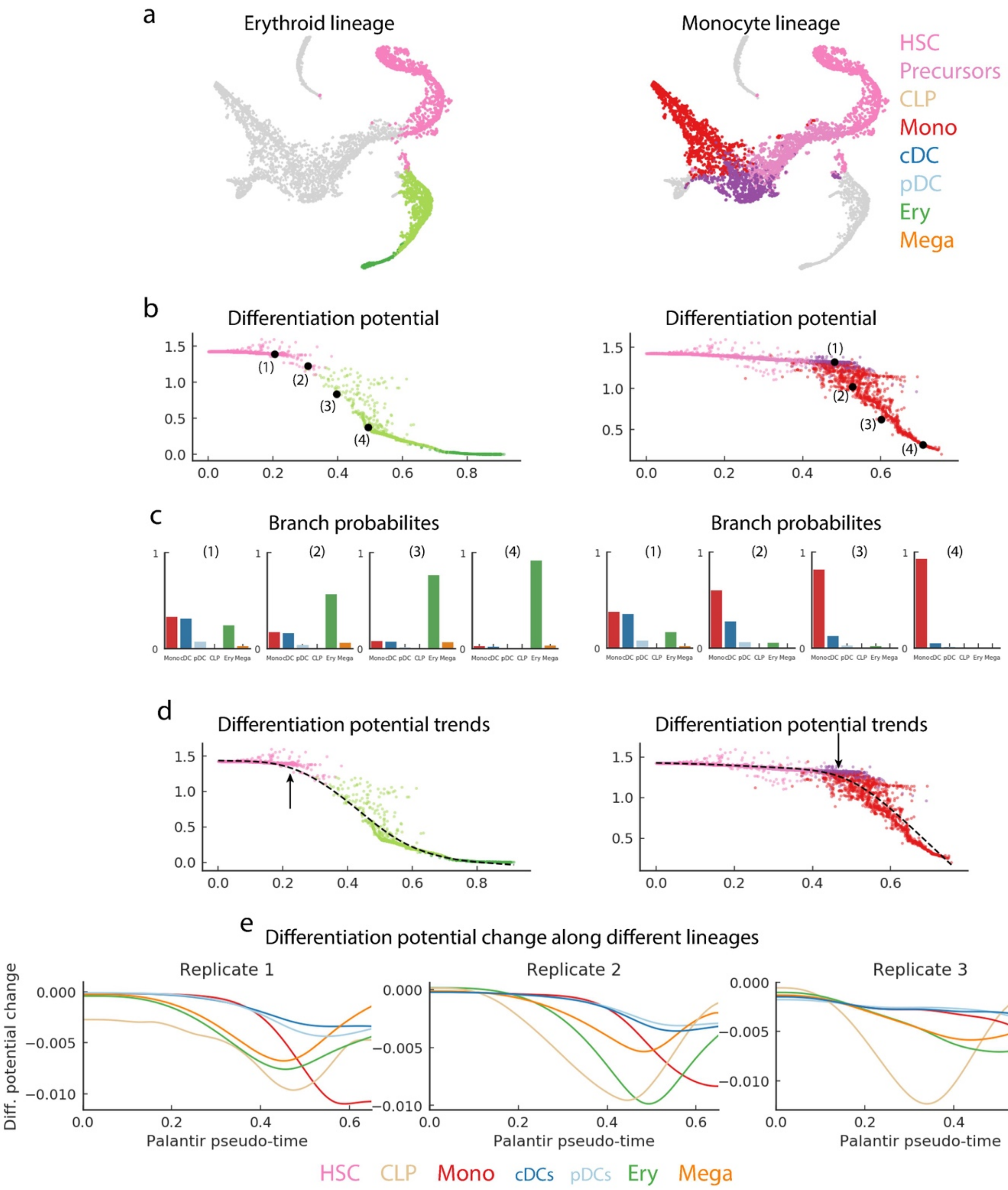
(d) Plots showing the Pearson correlation between de-novo and projected Palantir results. The cells are colored by clusters as in (a).

(e) Same as d, with cells colored by density.



Supp. Fig. 11: Reproducibility of gene expression dynamics across replicates

Plots showing reproducibility of expression trends of key lineage marker genes across the three replicates. The relevant lineages for each gene are bounded by dotted rectangles. Solid line represents the fitted trend estimate and shaded region represents standard deviation


a
Erythroid lineage
Monocyte lineage
HSC
Precursors
CLP
Mono
cDC
pDC
Ery
Mega
b
Differentiation potential
Differentiation potential
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
(1)
(2)
(3)
(4)
c
Branch probabilites
Branch probabilites
(1)
(2)
(3)
(4)
1
0
Mono
cDC
pDC
CLP
Ery
Mega
d
Differentiation potential trends
Differentiation potential trends
e
Differentiation potential change along different lineages
Replicate 1
Replicate 2
Replicate 3
Diff. potential change
0.000
−0.005
−0.010
Palantir pseudo-time
HSC
CLP
Mono
cDCs
pDCs
Ery
Mega

Supp. Fig. 12: Illustration of Differentiation potential.

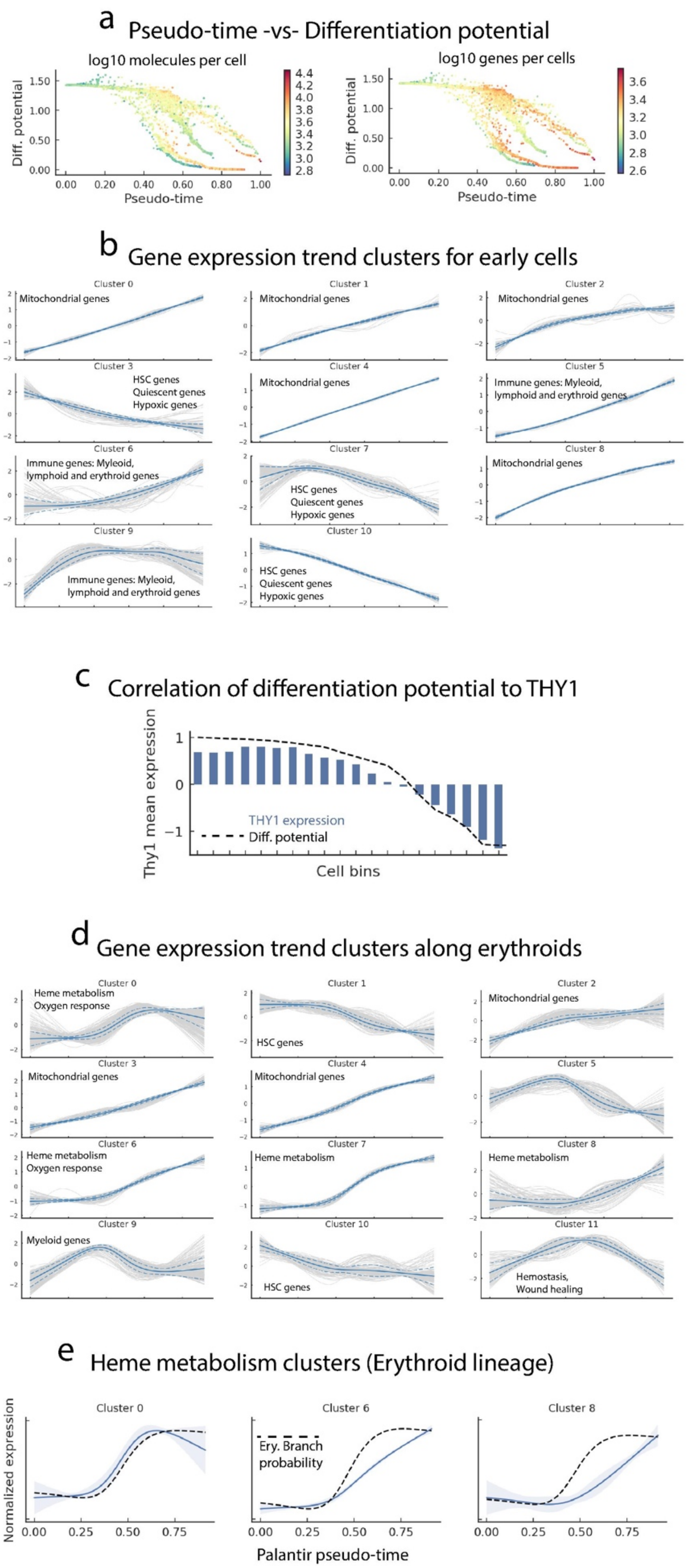
(a) tSNE plots highlight the cells of the erythroid lineage (left) and monocyte lineage (right). Cells are colored by Phenograph clusters as in Fig. 2b. 2125 and 3921 cells belong to the erythroid monocyte lineage respectively.

(b) Plots of DP (y-axis) along pseudo-time (x-axis). Each dot is a cell, color coded by the cluster. Representative cells are highlighted and numbered for each lineage.

(c) Histogram representation of the branch probabilities of the cells highlighted in (b), bars are colored coded by clusters in Fig. 2b. As cells commit towards a particular lineage, they also lose the ability to differentiate to other lineages. This is reflected in the gradual increase in probability of reaching the corresponding terminal state accompanied by a decrease in probability of reaching all other lineages.

(d) Same as (b), with trend plot of the DP along the corresponding lineages shown in black. The position along the ordering with the first substantial drop in DP corresponds to lineage specification and this downward trend continues until the cells are committed to the lineage and have completely lost the ability to differentiate to any other lineages.

(e) Plots showing the DP along pseudo-time for all the lineages, for the three replicates. The positions of significant DP changes are staggered along the pseudo-time indicating a hierarchical commitment of HSCs to different lineages. Cells first commit towards CLP (beige), followed by erythroid and megakaryocytic lineages (green, orange) and finally the myeloid lineages: monocyte (red) and DCs (blues).



a Pseudo-time -vs- Differentiation potential

b Gene expression trend clusters for early cells

c Correlation of differentiation potential to THY1

d Gene expression trend clusters along erythroids

e Heme metabolism clusters (Erythroid lineage)

Supp. Fig. 13: Clustering of gene expression trends in early and erythroid cells

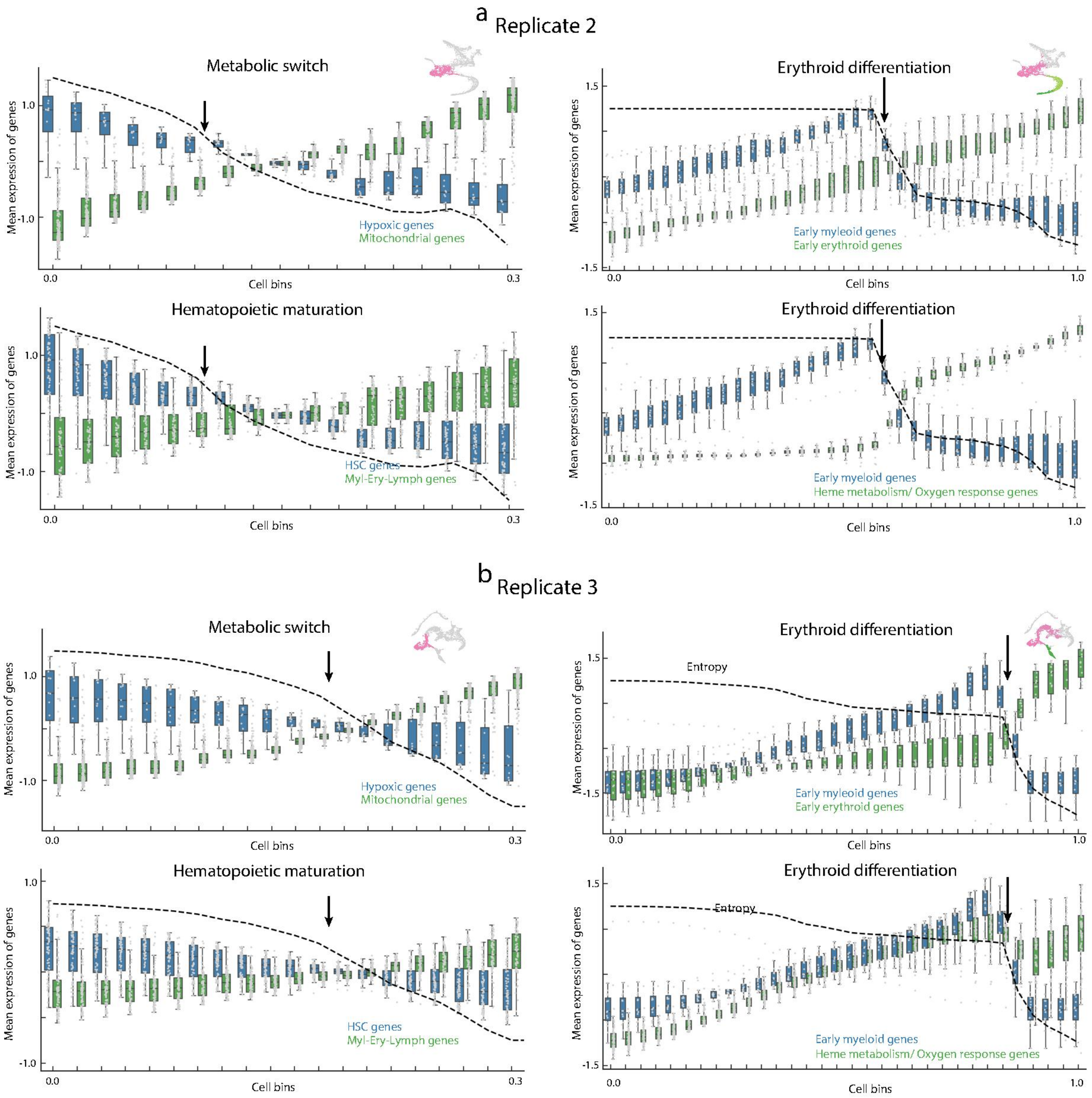
(a) Plot showing the comparison of pseudo-time ordering and differentiation potential with cells colored by the number of molecules (left) and number of genes (right) detected in each cell.

(b) Plots showing cluster results of expression trends of genes that are significantly high or low in early hematopoietic clusters (0 and 1 in Supp. Fig. 4a). Each grey line represents the expression trend of a particular gene. Solid blue line represents the mean expression trend of the cluster and dotted blues lines represent the standard deviation. Each panel is labeled with enriched gene ontology terms.

(c) Similar to Fig. 3b with the bar plot representing the mean expression of Thy1 in the bins.

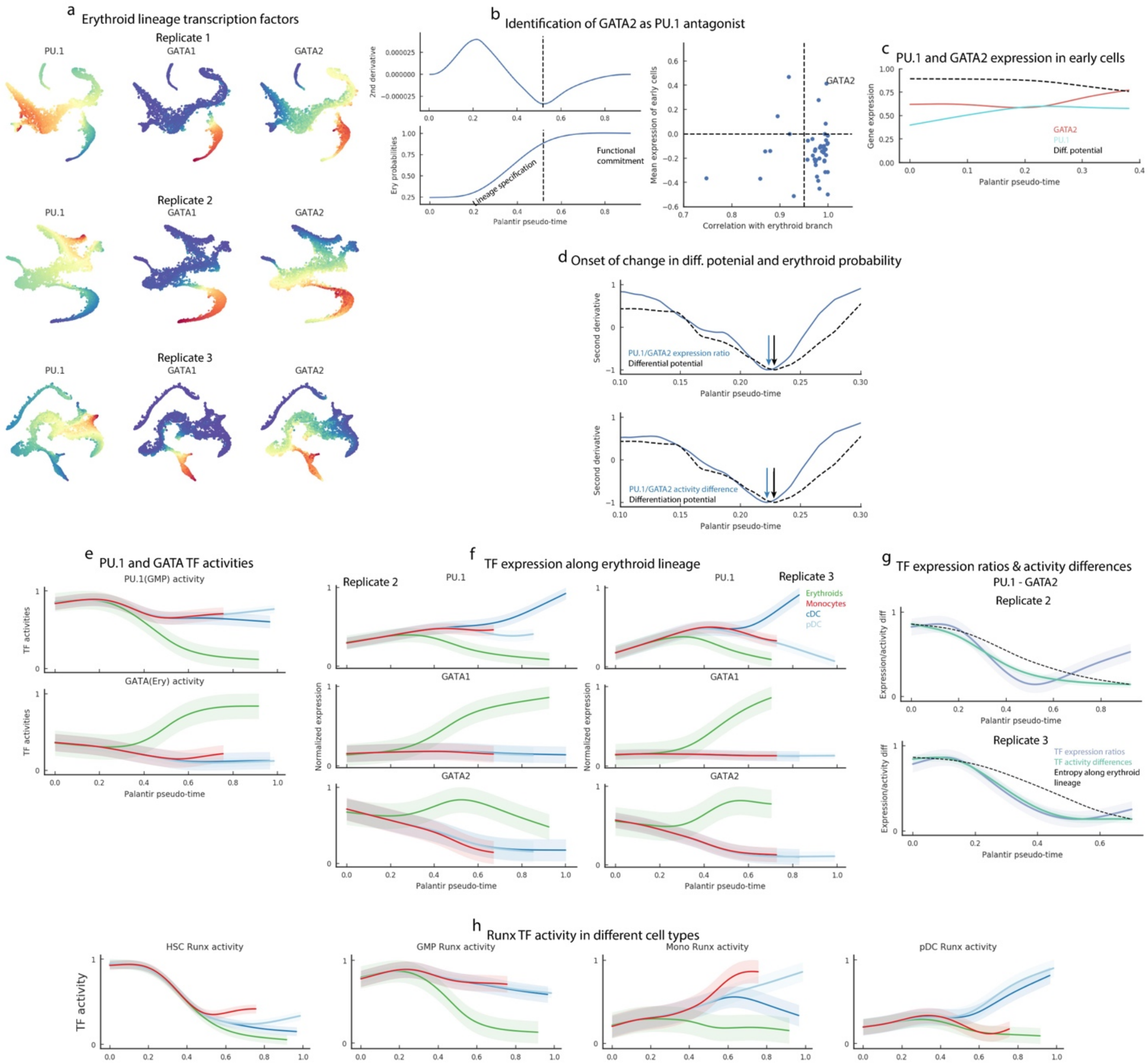
(d) Same as a - for genes significantly high or low in the clusters that correspond to the erythroid lineage (clusters 0, 1, 2 and 8 in Supp. Fig. 4a).

(e) Average gene expression trends of heme metabolism clusters (0, 6 and 8 in b). Erythroid branch probability shown in black. Cluster 0 genes correlates the most with erythroid branch probability and are enriched with key erythroid TFs. Cluster 6 genes include KLF3. Cluster 8 contains genes such as HBB, that confer functional identity to erythroid cells. Solid line represents the fitted trend estimate and shaded region represents standard deviation



Supp. Fig. 14: Palantir differentiation potential identifies landmarks of hematopoietic differentiation (Replication of results in Fig. 3)

Plots showing the reproducibility of results in Fig. 3 in replicates 2 and 3. Genes identified using replicate 1 were used for this analysis. Boxplots represent the mean and 1.5 s.t.ds.



Supp. Fig. 15: Identification of GATA2 as an agonist of PU.1 in driving erythroid specification

(a) tSNE plots showing the expression of PU.1, GATA1 and GATA2 in the three replicates. 5780 cells are shown in the tSNE map.

(b) Top left: Second derivative of erythroid branch probability trend. Dotted black represents the inflection point, i.e., the point of maximum change. Bottom left: Erythroid branch probability trend along Palantir pseudo-time.

Right panel: Scatter-plot of TF expression trends with erythroid branch trends during the lineage specification phase (x-axis) and average TF expression in the early cells (y-axis). Each dot represents a TF. GATA2 (labeled) is a clear outlier.

(c) Gene expression trends of PU.1 and GATA2 in early cells. Black line represents differentiation potential.

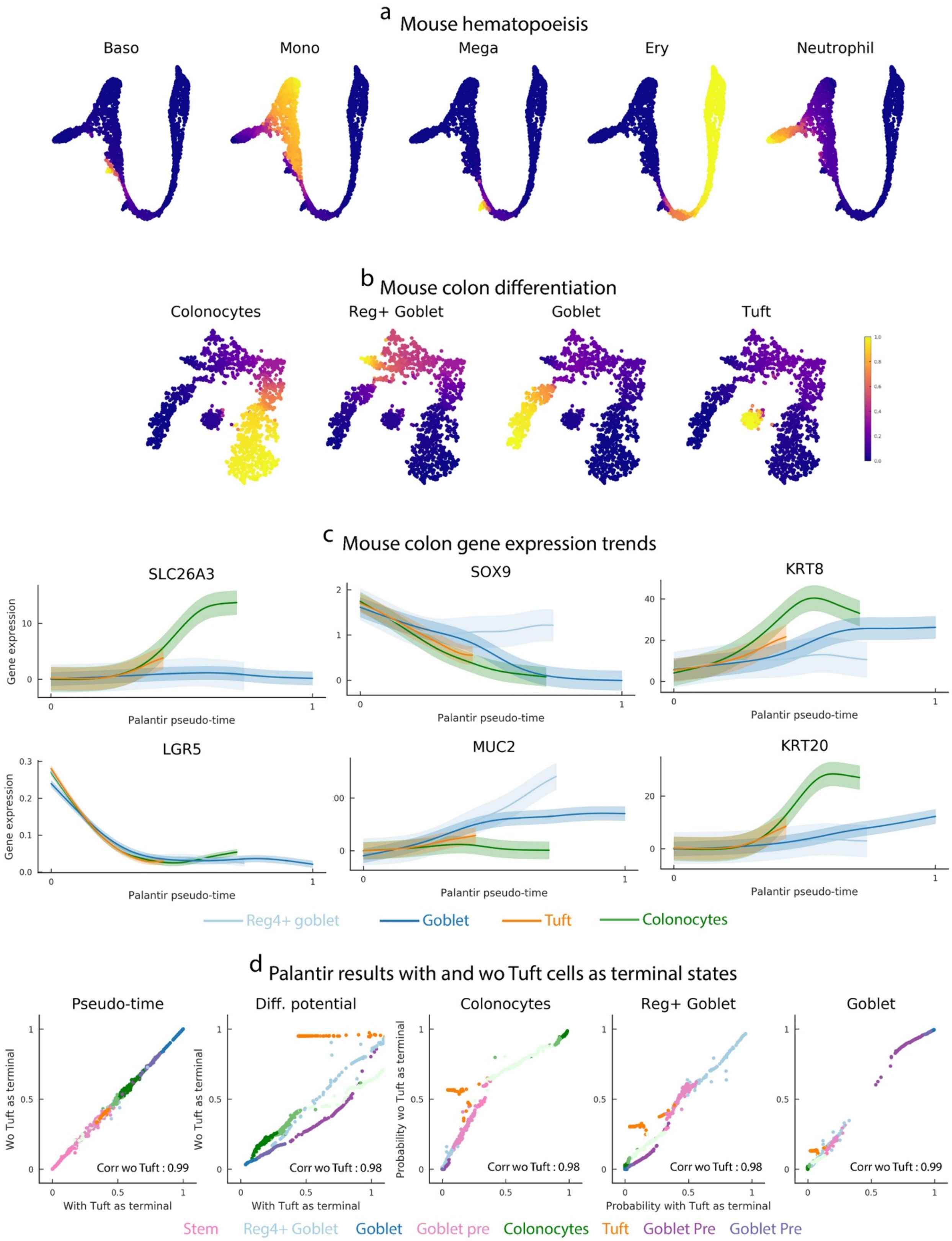
(d) Top panel: Plot showing the second derivatives and the inflection points of the PU.1/GATA2 expression ratio (in blue) and the differentiation potential (in black). The change in expression ratio (blue arrow) precedes the change in differentiation potential (black arrow) indicating that PU.1/GATA2 ratio is predictive of lineage commitment. Bottom panel: Same as the top panel but showing the TF activity differences between PU.1 and GATA2.

(e) PU.1 and GATA1 activity trends along erythroid and myeloid lineages. PU.1 activities were determined using bulk GMP cells and GATA activities using bulk erythroid cells. PU.1 activity, is similar to its expression, showing an increasing trend in the myeloid lineages and GATA activity shows an upregulation specifically in the erythroid lineage. Solid line represents the fitted trend estimate and shaded region represents standard deviation

(f) Same as Fig. 4a, for replicates 2 and 3

(g) Same as Fig. 4c, for replicates 2 and 3

(h) Plots showing the Runx activity trends determined using Runx targets in different sorted populations. The activity is high specifically in the cell type from which the targets were derived highlighting the cell - type specificity of TF targets and the inferred TF activity.


a
Mouse hematopoeisis
Baso
Mono
Mega
Ery
Neutrophil
b
Mouse colon differentiation
Colonocytes
Reg+ Goblet
Goblet
Tuft
1.0
0.8
0.6
0.4
0.2
0.0
c
Mouse colon gene expression trends
SLC26A3
Gene expression
10
0
0
1
Palantir pseudo-time
SOX9
2
1
0
0
1
Palantir pseudo-time
KRT8
40
20
0
0
1
Palantir pseudo-time
LGR5
Gene expression
0.3
0.2
0.1
0.0
0
1
Palantir pseudo-time
MUC2
00
0
0
1
Palantir pseudo-time
KRT20
20
0
0
1
Palantir pseudo-time
Reg4+ goblet
Goblet
Tuft
Colonocytes
d
Palantir results with and wo Tuft cells as terminal states
Pseudo-time
Wo Tuft as terminal
1
0.5
0
Corr wo Tuft : 0.99
0
0.5
1
With Tuft as terminal
Diff. potential
Wo Tuft as terminal
1
0.5
0
Corr wo Tuft : 0.98
0
0.5
1
With Tuft as terminal
Colonocytes
Probability wo Tuft as terminal
1
0.5
0
Corr wo Tuft : 0.98
0
0.5
1
Reg+ Goblet
1
0.5
0
Corr wo Tuft : 0.98
0
0.5
1
Probability with Tuft as terminal
Goblet
1
0.5
0
Corr wo Tuft : 0.99
0
0.5
1
Stem
Reg4+ Goblet
Goblet
Goblet pre
Colonocytes
Tuft
Goblet Pre
Goblet Pre

Supp. Fig. 16: Palantir generalizes to mouse hematopoiesis and colon differentiation

(a-b) tSNE plots, with cells colored by Palantir branch probabilities for mouse hematopoiesis [3] and colon differentiation datasets [4] respectively. tSNE maps show 2700 and 1811 cells respectively for the two datasets.

(c) Gene expression trends for the mouse colon data: Palantir results recapitulate the known behavior of key genes along different lineages. Solid line represents the fitted trend estimate and shaded region represents standard deviation

(d) Plots showing the correlation between Palantir results when Tuft cells (orange) were included (x-axis) or excluded as a terminal state (y-axis). Cells are color coded based on clusters in Fig. 5a. Correlation was computed using Pearson correlation

A

Inputs and outputs of trajectory detection algorithms

Algorithm	Specify start state	Specify no. of branchings	Clusters as input	Automatic detection of terminal states	Unified ordering to compare lineages	Probability of branch membership	Differentiation potential	Topological structure
Palantir	✗			✓	✓	✓	✓	
PAGA	✗				✓			✓
Slingshot	✗		✗	✓				
FateID	✗		✗			✓		
DPT	✗	✗			✓			
Monocle 2	✗				✓			

B

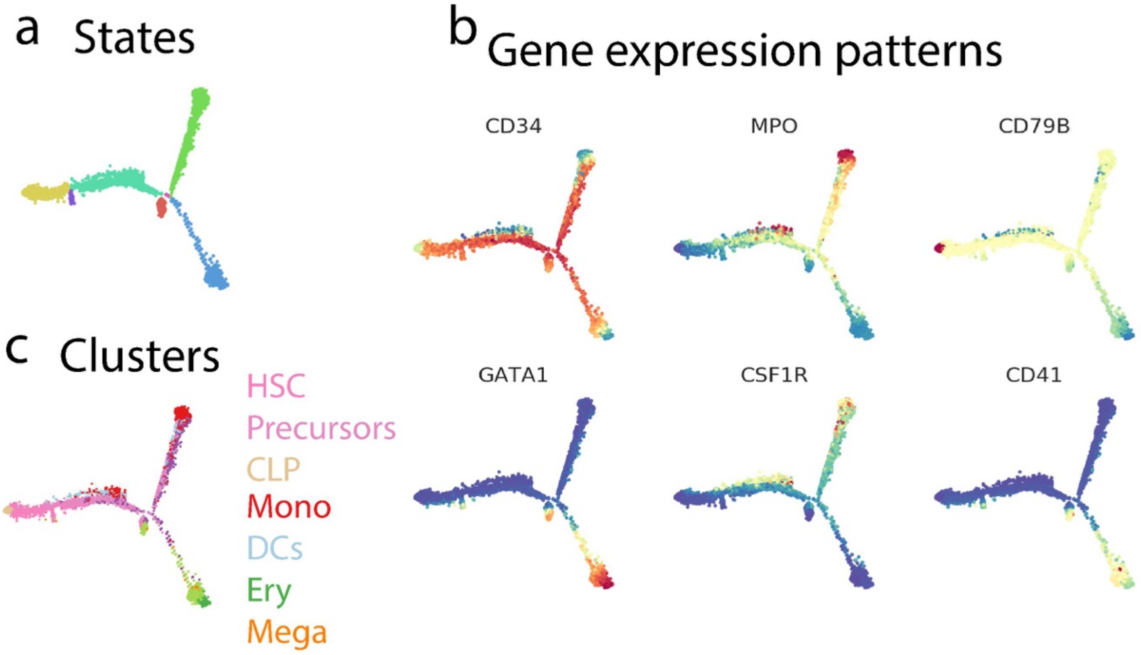
Performance of trajectory detection algorithms in CD34+ human bone marrow data

Algorithm	DC distinction	Megakaryocyte lineage	Gene expression trends						
			CD34	MPO (Mono)	CD79B (CLP)	GATA1 (Ery)	CSF1R (pDC)	CSF1R (cDC)	CD41 (Mega)
Citation									
Palantir	✓	✓	✓	✓	✓	✓	✓	✓	✓
PAGA			✓	✓	✓	✓	✓		
Slingshot			✓	✓	✓	✓	✓		
FateID (RaceID clustering)									
FateID (Palantir processing)		✓	✓	✓	✓				
DPT		✓	✓		✓	✓			*
Monocle 2									

Supp. Fig. 17: Metrics for evaluating trajectory detection algorithms

(a) Metrics used to evaluate the degree of *a priori* biological information required by the algorithm and the nature of outputs generated by the algorithm.

(b) Performance of the different algorithms to recover key lineages and gene expression dynamics of key lineage markers in human hematopoiesis.

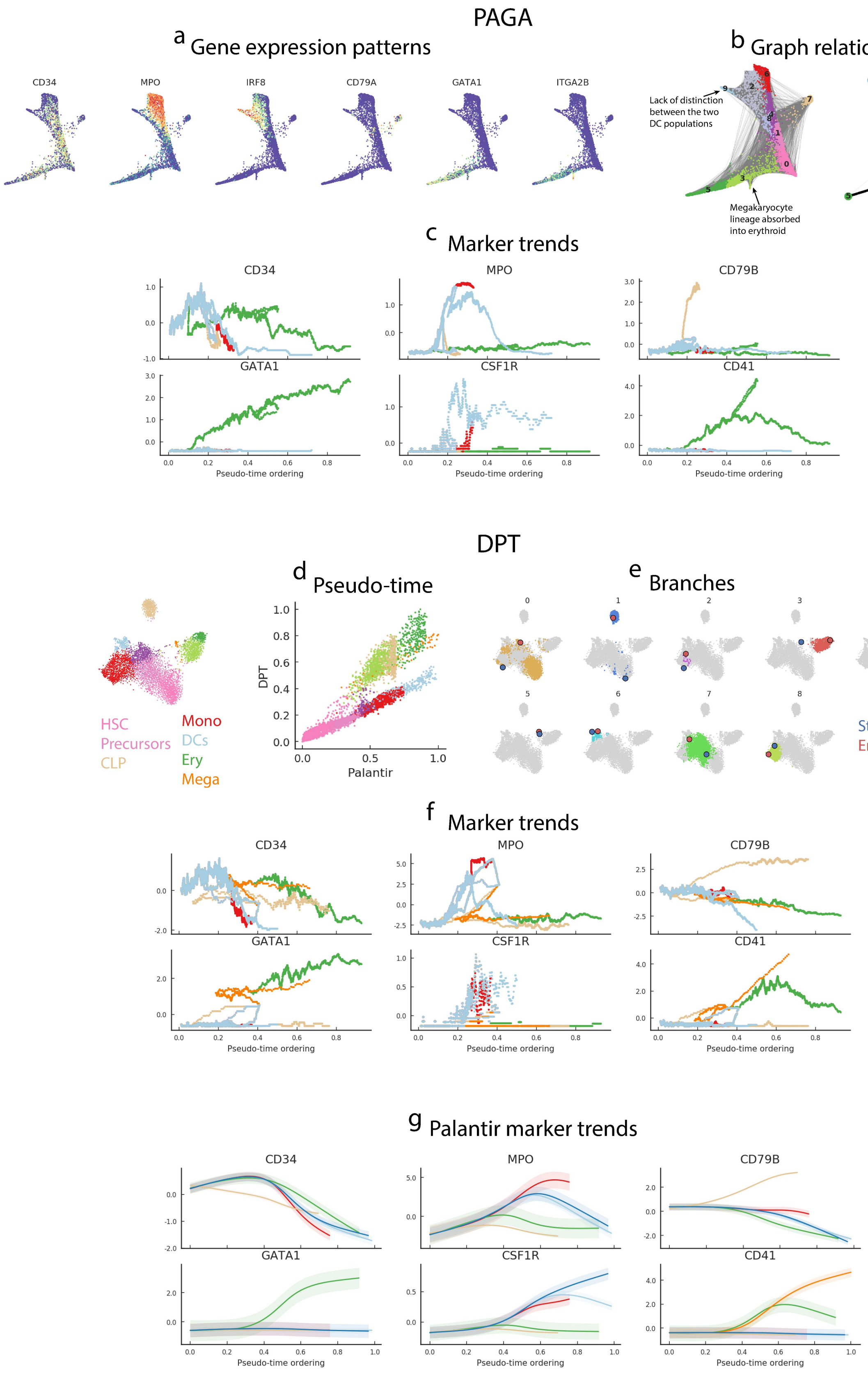


Supp. Fig. 18: Monocle 2 applied to human hematopoiesis

(a) Monocle2 generated low dimensional embedding of the data with cells colored by one of 6 Monocle 2 identified states. 5780 cells are shown in the tSNE map.

(b) Gene expression patterns of key lineage markers. MAGIC imputed data was used for plotting

(c) Same as (a) with cells colored by Phenograph clusters (Supp. Fig. 4a).


PAGA
a Gene expression patterns
CD34
MPO
IRF8
CD79A
GATA1
ITGA2B
b Graph relationships
Lack of distinction between the two DC populations
Megakaryocyte lineage absorbed into erythroid
9
2
6
7
8
1
0
3
5
c Marker trends
CD34
MPO
CD79B
GATA1
CSF1R
CD41
Pseudo-time ordering
DPT
d Pseudo-time
HSC
Precursors
CLP
Mono
DCs
Ery
Mega
DPT
Palantir
e Branches
0
1
2
3
4
5
6
7
8
Start
End
f Marker trends
CD34
MPO
CD79B
GATA1
CSF1R
CD41
Pseudo-time ordering
g Palantir marker trends
CD34
MPO
CD79B
GATA1
CSF1R
CD41
Pseudo-time ordering

Supp. Fig. 19: PAGA and DPT applied to human hematopoiesis

(a) PAGA representation of the data highlighting key marker genes. 5780 cells are shown in PAGA generated projection.

(b) Left: Connectivity among the hematopoietic cells as inferred by PAGA. Each dot is a cell color coded to represent lineages in Fig. 2b. Grey lines represent edges between two cells.

Right: PAGA abstract clusters color coded to resemble Phenograph clusters in Fig. 2b. Thickness of the connection between clusters represents the strength of connectivity. PAGA lacks distinction between the two DC lineages and embeds megakaryocyte lineage to be part of erythroid lineage.

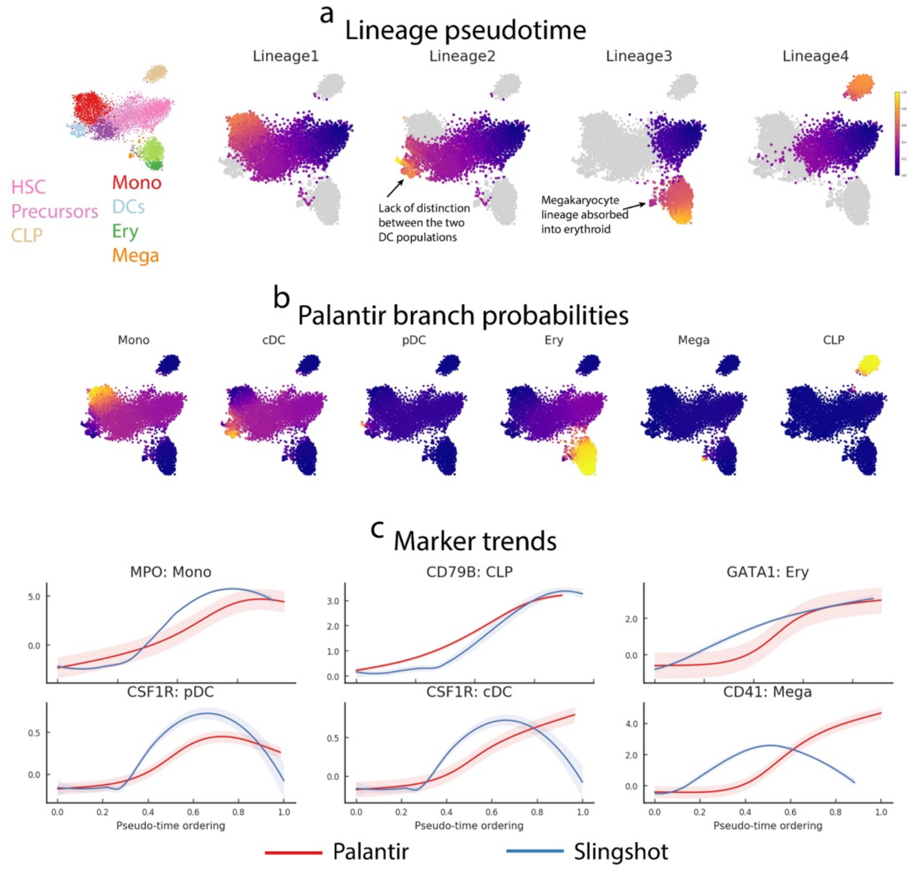
(c) Gene expression trends of key lineage markers as determined by PAGA. The trends are computed using a sliding window, making the estimates highly sensitive to noise in the data.

(d) Comparison of Palantir and DPT pseudo-time ordering of cells. Cells are colored by clusters (Supp. Fig. 4a). 5780 cells are shown in the tSNE map.

(e) Plots highlighting the different branches identified by DPT. The start and end cells of each branch are colored by blue and red respectively.

(f) Same as (c) - gene expression trends determined by DPT

(g) Palantir gene expression trends for the same genes for reference. Solid line represents the fitted trend estimate and shaded region represents standard deviation.

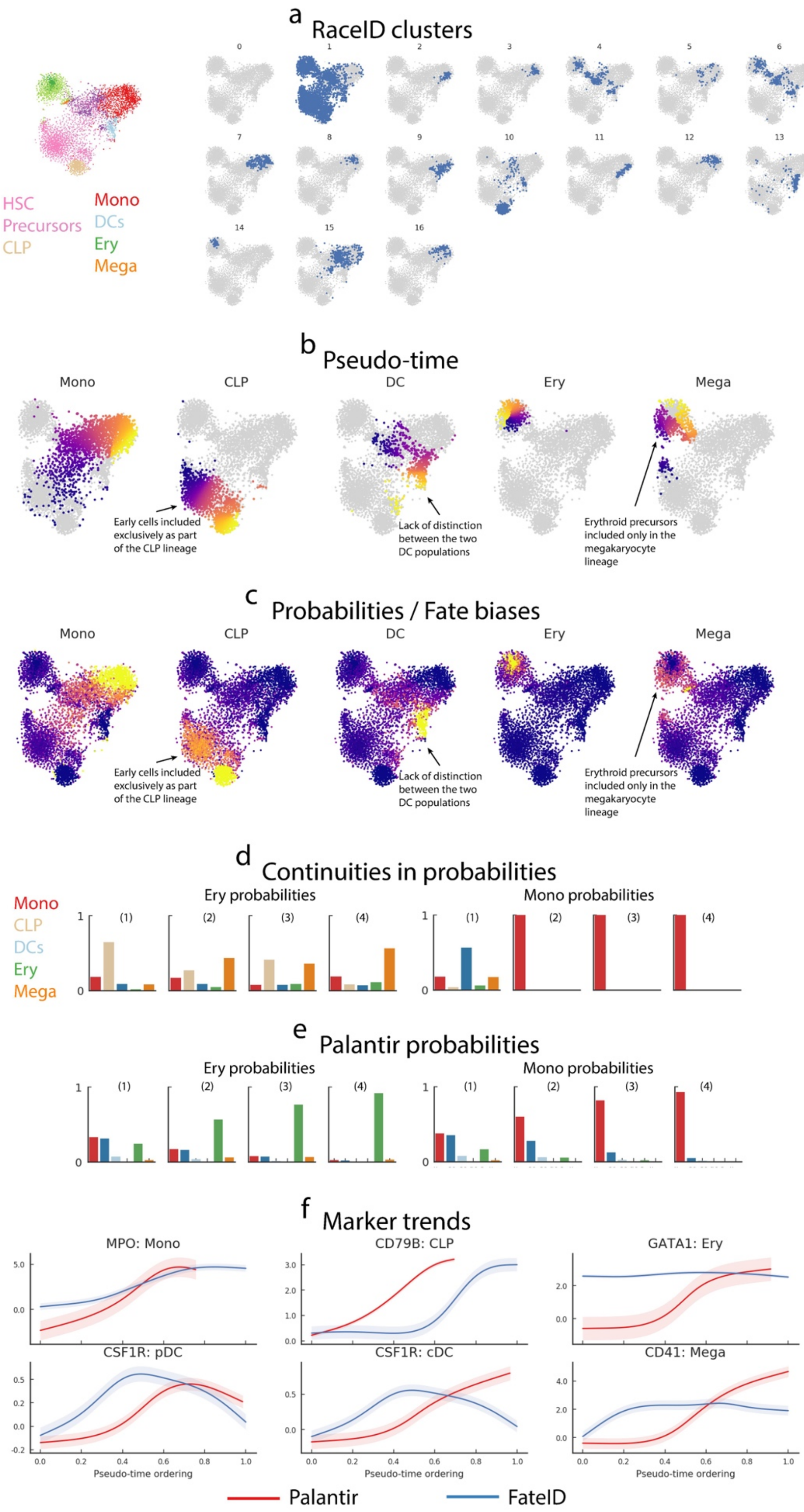


Supp. Fig. 20: Slingshot applied to human hematopoiesis

(a) Cells colored by Slingshot pseudo-time for the four different lineages Slingshot identified. Slingshot results lacks distinction between the two DC lineages and embeds megakaryocyte lineage to be part of erythroid lineage. 5780 cells are shown in the tSNE map.

(b) Palantir branch probabilities for reference

(c) Gene expression trends determined by Slingshot (blue) and Palantir (red) with Slingshot trends showing (1) unexpected marginal downregulation of CD79B at the beginning of CLP lineage. (2) unexpectedly high upregulation of CD41 along the erythroid lineage since the megakaryocytes are included as part of the erythroid lineage and (3) lack of distinction between CSF1R dynamics in the two DC lineages since they were embedded as part of the same lineage. Solid line represents the fitted trend estimate and shaded region represents standard deviation.


a
RaceID clusters
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
HSC
Precursors
CLP
Mono
DCs
Ery
Mega
b
Pseudo-time
Mono
CLP
DC
Ery
Mega
Early cells included exclusively as part of the CLP lineage
Lack of distinction between the two DC populations
Erythroid precursors included only in the megakaryocyte lineage
c
Probabilities / Fate biases
Mono
CLP
DC
Ery
Mega
Early cells included exclusively as part of the CLP lineage
Lack of distinction between the two DC populations
Erythroid precursors included only in the megakaryocyte lineage
d
Continuities in probabilities
Mono
CLP
DCs
Ery
Mega
Ery probabilities
Mono probabilities
1
0
(1)
(2)
(3)
(4)
e
Palantir probabilities
Ery probabilities
Mono probabilities
1
0
(1)
(2)
(3)
(4)
f
Marker trends
MPO: Mono
CD79B: CLP
GATA1: Ery
5.0
0.0
3.0
2.0
1.0
0.0
2.0
0.0
CSF1R: pDC
CSF1R: cDC
CD41: Mega
0.5
0.2
0.0
-0.2
0.5
0.0
4.0
2.0
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Pseudo-time ordering
Pseudo-time ordering
Pseudo-time ordering
Palantir
FateID

Supp. Fig. 21: FateID applied to human hematopoiesis

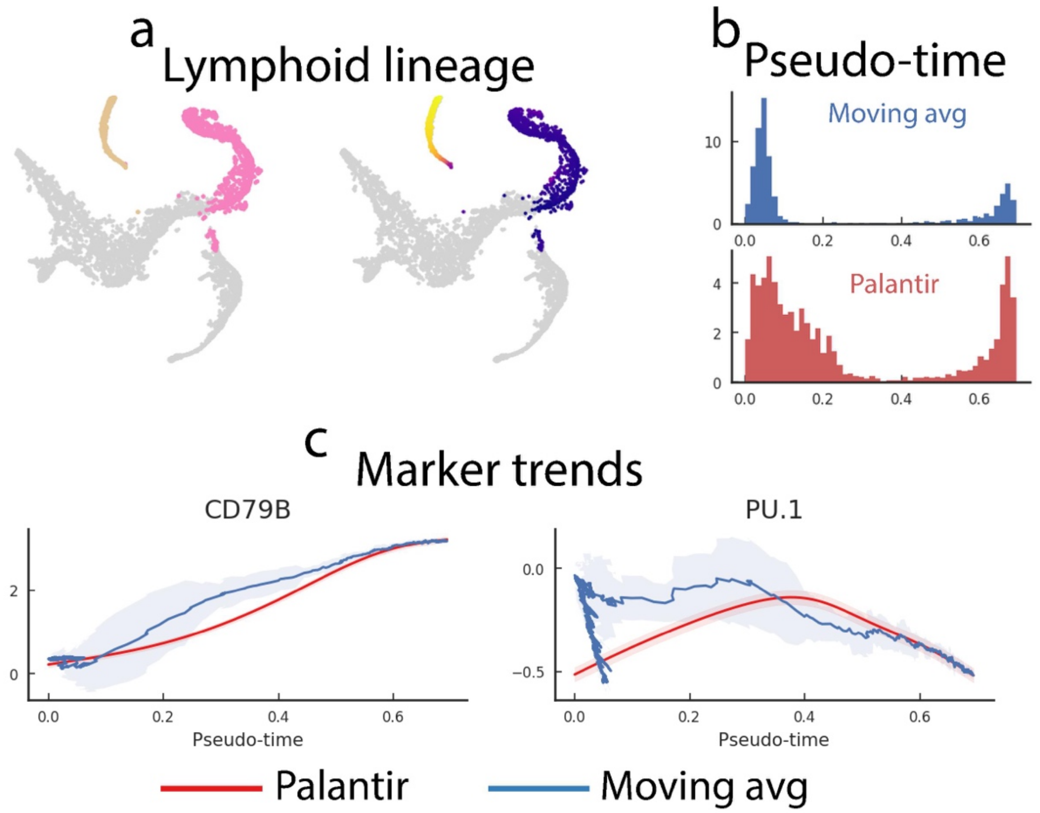
(a) FateID recommends using RaceID for clustering of the data. RaceID results in over clustering of data and does not represent coherent cell types or states. 5780 cells are shown in the tSNE map.

(b-c) FateID results generated using Palantir recommended preprocessing. FateID includes all the early cells exclusively as part of the CLP lineage, is largely incorrect for most cell lineage probabilities and is moreover unable to identify the low frequency lineages.

(d) FateID fate biases for cells along the myeloid and erythroid lineages (Supp. Fig. 9), which fails to get correct erythroid probabilities and smooth transition into myeloid lineages, highlighting the loss in resolution to identify continuities in fate choice commitment

(e) Palantir probabilities for the same cells for reference

(f) Gene expression dynamics of key lineage genes as determined by FateID (blue), in comparison to Palantir (red). The gene expression dynamics are heavily influenced by exclusive inclusion of early cells in the CLP lineage: (i) GATA1 does not show the expected upregulation in the erythroid lineage, (ii) MPO, CEBPG show a high basal level of expression at the earliest stages of ordering. Solid line represents the fitted trend estimate and shaded region represents standard deviation.

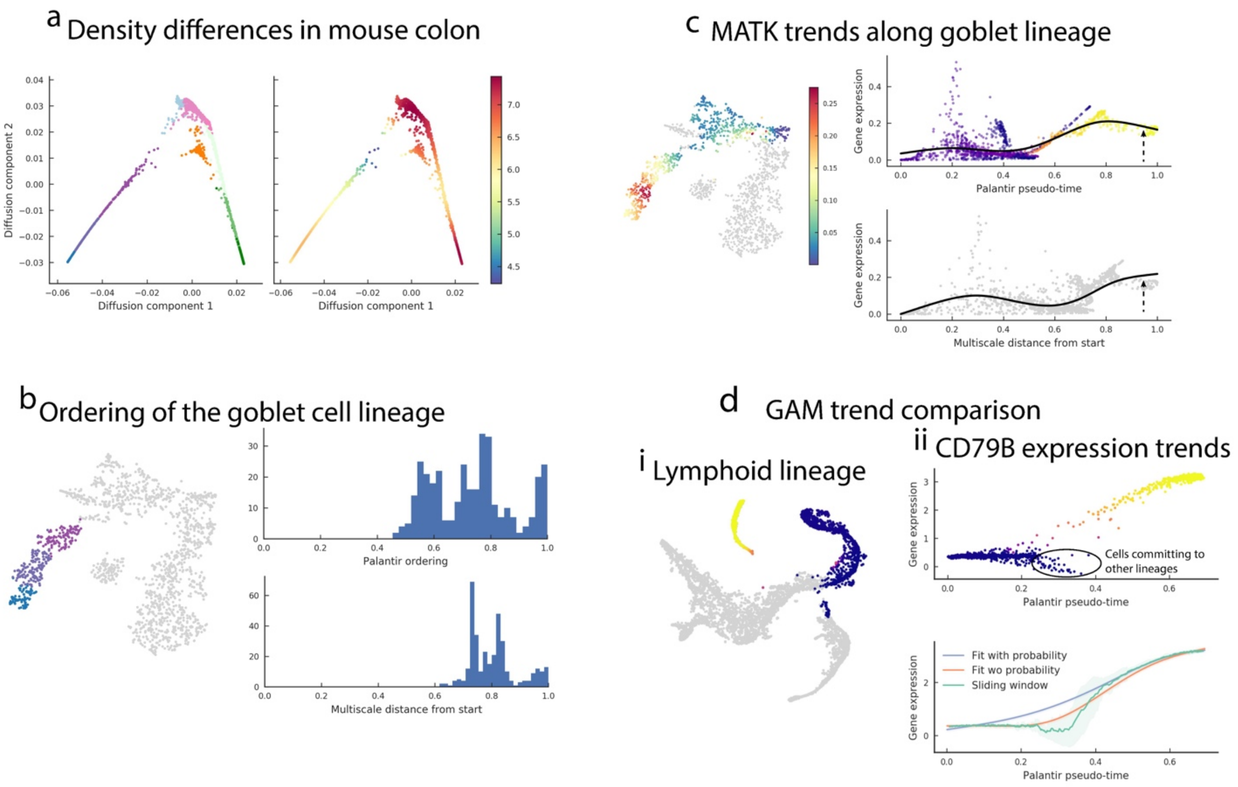


Supp. Fig. 22: Diffusion components for charting lineage dynamics

(a) (left) tSNE plots highlight the cells along lymphoid lineage and the pseudo-time of these cells as determined by projection along the 2nd diffusion component (Supp. Fig. 2b). 1606 cells are part of the lymphoid lineage.

(b) Histograms showing the density of distribution of cells along diffusion component (blue) or Palantir pseudo-time (red).

(c) Gene expression trends determined by sliding windows (blue) and Palantir (red). Solid line represents the fitted trend estimate and shaded region represents standard deviation.



Supp. Fig. 23: Palantir features key to identifying accurate identification of pseudo-time ordering and branch probabilities

(a) Projection of the mouse colon data (Fig. 5) along the first two diffusion components. Each dot it a cell and is color coded by clusters in Fig. 5a (left panel) and density (right panel). The projection of 1811 cells are shown.

(b) (i) tSNE map highlighting cells of the goblet lineage. Cells are color coded by clusters in Fig. 5a. 350 cells belong to the goblet lineage.

(ii) Histograms showing the distribution of goblet cells along the pseudo-time derived using shortest path distances (top panel) and multi scale distances (bottom panel). Distribution of cells using multiscale distance leads to loss in resolution with increased local concentration of goblet cells at the end of the pseudo-time.

(c) (i) tSNE plot showing the expression of MATK along the cells that contribute to the goblet cell lineage (left panel).

(ii) Plots showing MATK expression along pseudo-time where each dot is a cell.

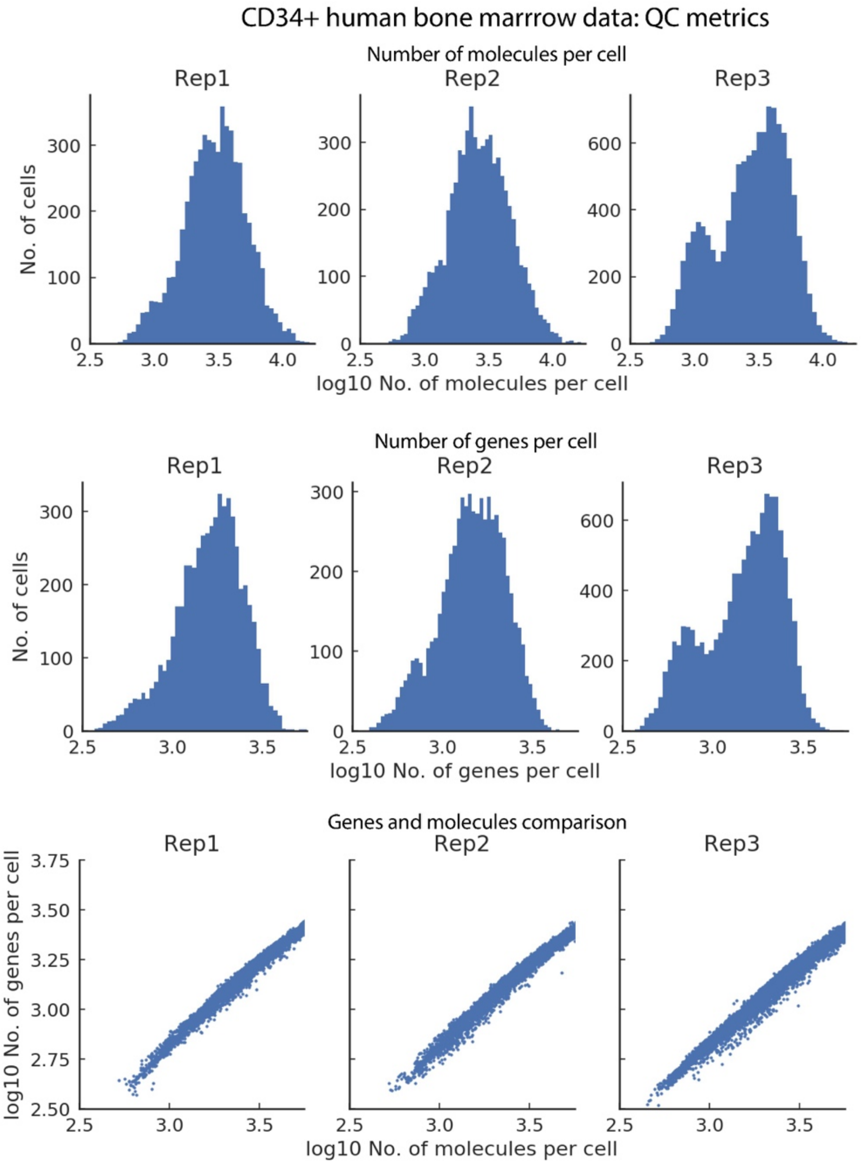
Top: Cells colored by goblet cell probability. Black line represents the MATK trend obtained by GAM fit using goblet cell probabilities as weights for cells.

Bottom: MATK trend with GAMs fitted on cells ordered by pseudo-time generated using multi scale distances. The loss of resolution in pseudo-time results in even GAMs miss trends such as MATK downregulation towards the end of goblet lineage.

(d) (i) tSNE plot highlighting the cells along the lymphoid lineage (clusters 0 and 5 in Supp. Fig, 4a) colored by lymphoid branch probability from the human hematopoiesis dataset.

(ii) Top: Plot showing expression of CD79B along Palantir pseudo-time. Each dot is a cell and is colored by lymphoid branch probability. Cells that are committing towards other lineages are highlighted.

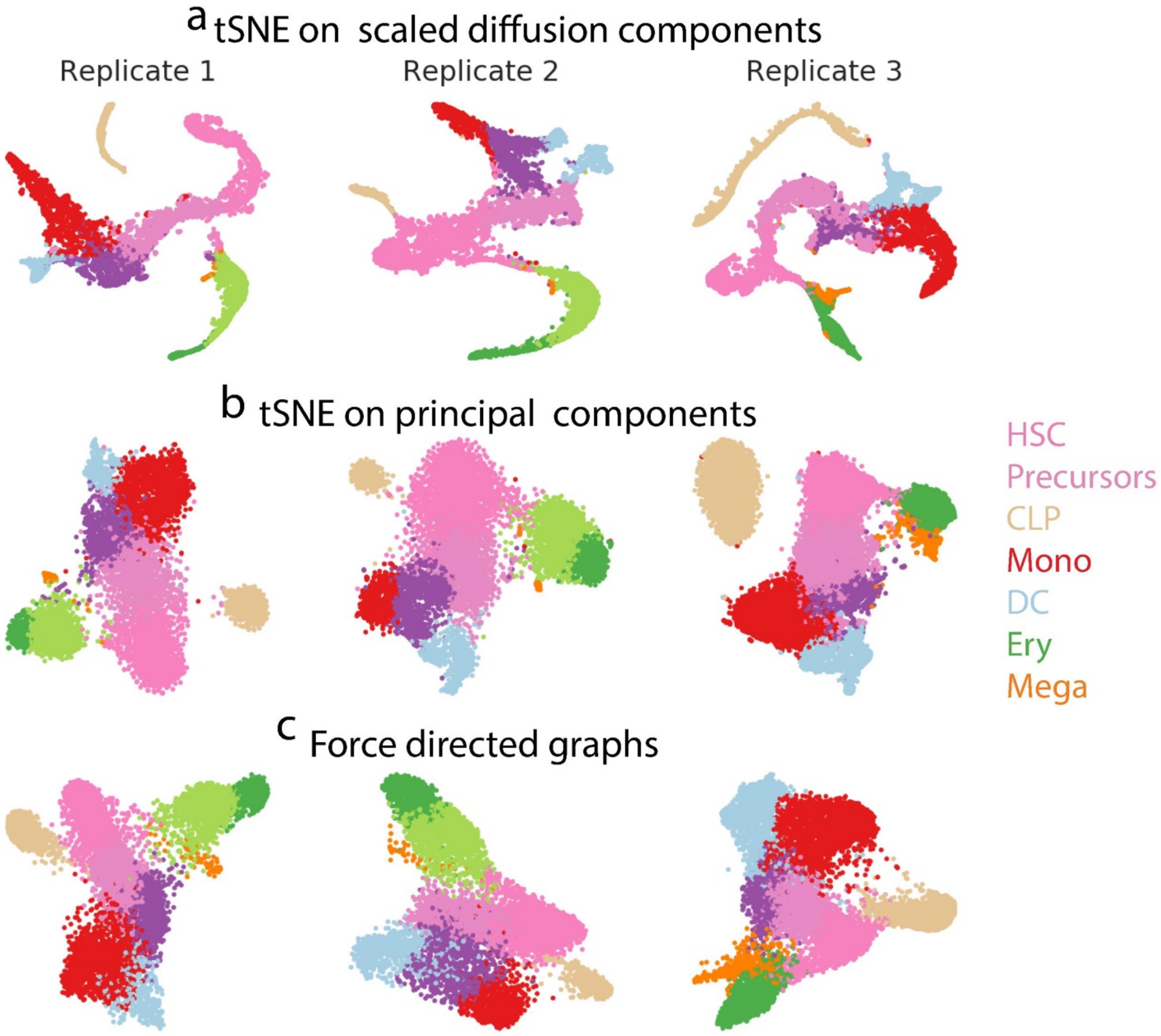
Bottom: Gene expression trends computed using sliding windows (green), GAMs fit without using Palantir probabilities (orange) and GAMs fit using Palantir probabilities (blue). Sliding window and GAMs fit without using branch probabilities are heavily influenced by cells committing to other lineages leading to incorrect trend estimates.



Supp. Fig. 24: QC metrics of cells post filtering

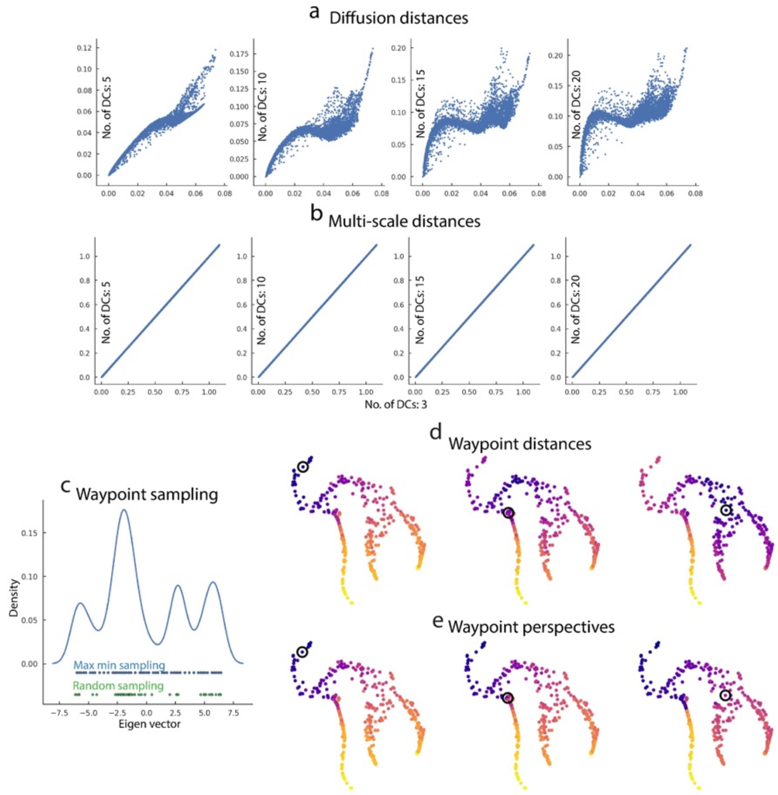
(a-b) Histograms of the number of molecules and cells detected per cell (in log10).

(c) Plots comparing the number of molecules and cells detected per cell.



Supp. Fig. 25: Comparison of data visualizations

tSNE maps generated using (a) scaled diffusion components and (b) principal components. (c) Force directed graphs for the same cells. Cells are color coded by clusters in Fig. 2b. Replicates 1-3 contain 5780, 6501 and 12046 cells respectively.



Supp. Fig. 26: Multi-scale distances, waypoint sampling and perspectives

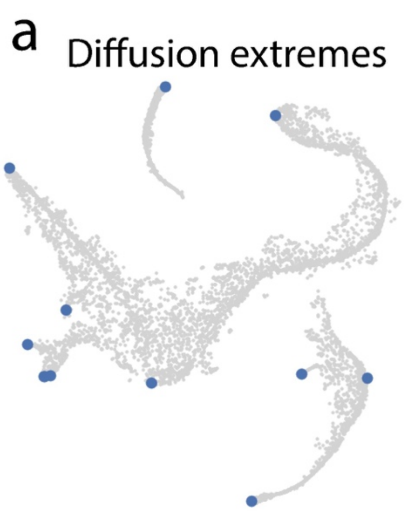
(a) Plots comparing the diffusion distances from an early HSC cell when different number of components are used.

(b) Same as (a), with distances computed using multi-scale distance.

(c) Plot showing the variable density of cells along a particular diffusion component. Random sampling of waypoints samples cells from high density regions while ignoring of low density and high variability (green dots). Max-min sampling however samples cells along the entire spectrum of the diffusion component and generates a more representative sample of cells (blue dots).

(d) Shortest path distances from a subset of waypoints.

(d-e) Cells from subsampled dataset (Fig. 1b) colored by shortest path distances and perspectives from the highlighted waypoints.



Supp. Fig. 27: Identification of terminal states

(a-b) Diffusion map boundaries (11 cells) and the identified terminal states (6 cells) for replicate 1 of the human hematopoiesis data.

Supplementary Notes

Supplementary Note 1: The Palantir algorithm

Constructing a nearest neighbor graph representing the phenotypic manifold

Palantir first constructs a nearest-neighbor graph representing the phenotypic manifold, where each cell is connected to its most similar cells. Key to the success of this approach is that the resulting graph neighbors consist of cells in similar developmental states and that longer paths correspond to developmental trajectories. Given the extensive degree of sparsity and noise in scRNA-seq, finding nearest neighbors in the raw data using a simple similarity metric is likely to accumulate spurious connections and obscure the structure we are seeking.

To construct the neighbor graph based on robust trends in the data, Palantir uses diffusion maps [5], which project the data onto a low dimensional manifold that approximates the differentiation landscape. Diffusion maps have been previously used to study differentiation in single cell data [6, 7] and are particularly adept at capturing differentiation. Diffusion maps generate a low-dimensional embedding by approximating all possible paths via random walks through the graph, which effectively capture the major axes of variation in the data (Supp. Fig. 1).

The first step in constructing diffusion maps is to define a measure of similarity between cells. Following [1], we use an adaptive (width) Gaussian kernel to convert distances into affinities, so that similarity between two cells decreases exponentially with their distance. Typically, an isotropic or non-adaptive Gaussian kernel is used to measure the similarity with an inherent assumption the density of the data is uniform along the trajectory. However previous single cell studies have shown that while differentiation trajectories are continuous, they are punctuated by large changes in densities [6, 8] possibly representing meta-stable states. A non-adaptive kernel would be strongly biased by the densest regions. The adaptive kernel [1]corrects for the densities by using the distance to the l^{th} nearest neighbor as a scaling factor, thus equalizing the effective number of neighbors for each cell.

Formally, given a dataset, $\boldsymbol{X} \in \boldsymbol{R}^{N\times M}$, with N cells and M genes, a k-nearest neighbor graph, $\mathrm{G_X} \in \mathbb{R}^{N\times N}$ is constructed using the Euclidean distance. The distances are converted to affinities using the adaptive kernel as defined below.

The scaling factor of cell i is determined by

$$\sigma_i = Distance\ to\ l^{th}\ neighbor\ (l < k) \tag{1}$$

Given this, the similarity measure between two cells i and j is given by

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{1}{\left|2\pi(\sigma_i + \sigma_j)\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\frac{(\boldsymbol{x_i} - \boldsymbol{x_j})(\boldsymbol{x_i} - \boldsymbol{x_j})^T}{\sigma_i + \sigma_j}\right) \tag{2}$$

Where $\boldsymbol{x_i}$ is the vector of gene expression for cell i. Thus, the above adaptive anisotropic kernel is used to define an affinity matrix, $\mathbf{K} \in \mathbb{R}^{N\times N}$ from the data. We then compute the Laplacian of the affinity matrix $\mathbf{K}$ to derive the diffusion operator $\mathbf{T} \in \mathbb{R}^{N\times N}$, where T_{ij} represents the probability of reaching cell j from cell i in one step. The Eigenvectors of the diffusion operator $\mathbf{T}$ are termed diffusion components and these represent major axes of variation in the underlying manifold from which the data was sampled. The top diffusion components (Eigenvectors of $\mathbf{T}$)

define a non-linear low dimensional embedding that approximates the phenotypic manifold of the data (Supp. Fig. 2) [5].

Pseudo-time ordering of cells

Once the manifold is constructed, the next step is to infer a pseudo-time for all cells in the data. The computed pseudo-time does not represent a single trajectory, but rather assigns each cell their relative distance from a starting cell, regardless of their lineage or terminal fates. Typically, diffusion maps have been used to characterize pseudo-time ordering of cells, constructing separate trajectories for each major axes of variation, based on individual diffusion components (DCs) [7, 9, 10]. While a single DC can sometimes offer a reasonable approximation of an ordering towards a specific fate, we often observe many to many relationships between DCs and ordering leading to each terminal fate (Supp. Fig. 2). Therefore, Palantir takes multiple DCs into account when computing the pseudo-time of cells.

The embedded space (diffusion map) is used as an approximation of the differentiation landscape. Palantir uses Euclidean distance at multiple scales or multi-scale distance (elaborated below) in this embedded space to construct a more reliable nearest-neighbor graph, $\mathrm{G_E} \in \mathbb{R}^{N \times N}$ that filters out much of the noise in the original neighbor graph $\mathrm{G_X}$ (Supp. Fig 1a). Then pseudo-time is determined using shortest path distances in the graph $\mathrm{G_E}$ (Supp. Fig. 1b), since shortest path lengths better approximate the geodesic distances in the manifold [11].

The extremes of the diffusion components determine the boundaries of the phenotypic space and the start cell is defined as the boundary cell closest to the user defined starting point. Then, pseudo-time is initialized as the shortest path distances from this start cell. A shortcoming of shortest path distance is that it tends to accumulate noise with increasing distances [6, 8], thus, similarly to [6, 8], waypoints are used to refine the pseudo-time, defining the ordering based on a weighted vote of waypoints. Waypoints act as guides: the waypoint closest to the cell gets the highest vote in determining the position of the cell along pseudo-time. Thus, the success of Palantir requires that all regions of the manifold are well covered by waypoints that can guide the positioning of cells in their respective regions.

Our previous pseudo-time algorithms [6, 8] used a random sample of cells as waypoints. However, random sampling does not successfully cover the landscape in complex datasets, with multiple branches and the variable densities of cells along different lineages. Therefore, Palantir uses max-min sampling [12], an iterative procedure to choose waypoints that spread over and represent the entire manifold, rather than representing only the regions of high density.

In summary, the positioning is initialized based on shortest path distances from the start cell and is iteratively refined using the waypoints to fine tune the distances of the cells within the region of each waypoint. A weighted average across all waypoints is used to ensure the computation of a consistent global structure. Convergence of this procedure defines a final pseudo-time ordering of cells (Supp. Fig. 1c). Below we provide more detail:

Measuring distances between cells using multi-scale distance

Let the manifold be represented by $\mathbf{E} \in \mathbb{R}^{N \times L}$ where L is the dimension of the embedding with $L < M$. The dimension L of the embedding is chosen using an Eigen gap among the top

Eigenvectors. Let $\lambda_1, \lambda_2, \ldots \lambda_L$ be the corresponding Eigenvalues associated with diffusion components that define the manifold.

Given this, the distance between cells i and, j known as the diffusion distance, is defined by

$$DD_t(e_i, e_j)^2 = \sum_{l=1}^{L} \lambda_l^{2t} \left(e_i{}^{(l)} - e_j{}^{(l)}\right)^2 \tag{3}$$

where t is the number of steps through the graph and $e_i{}^{(l)}$ is the embedding of cell i along diffusion component l. Different stages of differentiation happen at different rates and occur at different densities in the population, thus a single t is unsuitable across the entire population. To avoid setting a particular t, in a similar manner to [7], we use multi-scale distance that accounts for all scales:

$$MS(e_i, e_j)^2 = \sum_{t=1}^{\infty} \sum_{l=1}^{L} \lambda_l^{2t} \left(e_i{}^{(l)} - e_j{}^{(l)}\right)^2 \tag{4}$$

By definition, $1 > \lambda_1 > \lambda_2 > \ldots > \lambda_L > 0$, thus Equation (4) can be rewritten as

$$MS(e_i, e_j)^2 = \sum_{l=1}^{L} \left(\frac{\lambda_l}{1 - \lambda_l}\right)^2 \left(e_i{}^{(l)} - e_j{}^{(l)}\right)^2 \tag{5}$$

The use of multi-scale distance avoids the selection of an additional parameter (t) and also renders the distance robust to different choices of L (Supp. Fig. 26a-b), robust to outlier cells and density differences.

Max-min Waypoint sampling

Max-min sampling is an iterative procedure, where at each iteration, the chosen waypoint maximizes the minimum distance to the set of current waypoints [12], thus covering a new region of the manifold. Palantir uses max-min sampling along each diffusion component to sample waypoints.

Let $\mathbf{E}^{(l)}$ be the l^{th} diffusion component. Max-min sampling is initialized with a randomly sampled cell from the diffusion component: $WS^{\,(l)} = Random(N,1)$. Distances along the component to the current waypoint set are computed for all the cells

$$wd_{ij} = \left(e_i^{(l)} - e_j^{(l)}\right)^2 \forall\, j \in WS^{\,(l)} \tag{6}$$

For each cell i, minimum of the current waypoint distances is computed

$$md_i = min\left(wd_{ij}\right) |\, j \in WS^{\,(l)} \tag{7}$$

The cell with the maximum of these minimum distances is added to the waypoint set

$$WS^{\,(l)} = \bigcup\left(WS^{\,(l)},\ \operatorname{argmin}(\mathbf{md})\right) \tag{8}$$

This procedure is repeated until the desired number of waypoints is sampled along the component and then repeated for all components. Union of the waypoints sampled along all diffusion components represents the final waypoint set, WS. An example of waypoint sampling along a component is shown in Supp. Fig. 26c.

Iterative pseudo-time computation

Palantir begins with designating a start cell based on a user defined starting point. It is assumed that the starting cell would reside at the boundary of the manifold, that is a cell that projects onto an extreme endpoint along of one of the diffusion components. First, the set of boundary cells is determined using

$$C = \bigcup_{l=1}^{\mathrm{L}} \left(\operatorname{argmin} \mathbf{E}^{(l)}, \operatorname{argmax} \mathbf{E}^{(l)}\right) \tag{9}$$

The extreme cell closest to the user input early cell s is then used as the start of the pseudo-time, s'.

$$s' = \operatorname*{argmin}_{i \in C} MS(e_s, e_i) \tag{10}$$

The pseudo-time, $\tau_i^{(0)}$, is initialized as the shortest path distances from the start cell s'. Shortest path distances are computed from each of the waypoints to all cells (Supp. Fig. 26d). These distances are then aligned to the start cell distances to compute waypoint perspectives (Supp. Fig. 26e). The pseudo-time is then updated as the weighted average of the different waypoint perspectives, ensuring that the pseudo-time of a cell is most strongly influenced by the waypoints closest to it, while maintaining a consistent global structure.

Formally, let D_{wi} be the shortest path distance of cell i from to waypoint w. The perspective of a cell i relative to waypoint w is the distance of from early cell s'. is computed as

$$V_{wi} := \begin{cases} \tau_w^{(0)} + D_{wi} & \text{if } \tau_i^{(0)} > \tau_w^{(0)} \\ \tau_w^{(0)} - D_{wi} & \text{otherwise} \end{cases} \tag{11}$$

Note that the perspective of the early cell s' is the initial ordering $\boldsymbol{\tau}^{(\mathbf{0})}$ itself.

The weighted average of waypoint perspectives is used to refine the pseudo-time, using an exponential weighting scheme where the weight is inversely proportional to the distance between the waypoint and the cell. The weights are determined as follows

$$W_{wi} = exp\left(\frac{-D_{wi}^2}{\sigma}\right) \Big/ \sum_{k=1:N} exp\left(\frac{-D_{wk}^2}{\sigma}\right) \tag{12}$$

where σ is the standard deviation of distance matrix $\mathbf{D}$. This defines the weight matrix $\mathbf{W} \in R^{nW \times N}$**.** The weighted average is then calculated by

$$\tau_i^{(1)} = \sum_{w \in WS} V_{wi} * W_{wi} \tag{13}$$

Note that the waypoints themselves are also cells and thus their relative distance to the start cell is modified and updated by this procedure. The updated ordering is then iteratively refined until convergence to obtain a final pseudo-time, $\boldsymbol{\tau}$ (Supp. Fig. 1c).

Inferring the Terminal Fates and differentiation potential

Modeling Differentiation as an Absorbing Markov Chain

Consider the neighbor-graph spanning the waypoints, $\mathrm{G'}_{\mathrm{E}} \in \mathrm{G}_{\mathrm{E}}$. Differentiation is modeled as a stochastic process, implemented as a Markov chain, where a cell reaches one or more terminal states through a series of steps in the manifold (Fig. 1b), based on the assumption that paths in the neighbor-graph G_{E} correspond to possible differentiation paths. However, differentiation is a directed process, from a less differentiated to a more differentiated state, whereas $\mathrm{G'}_{\mathrm{E}}$ is an undirected graph.

The inferred pseudo-time $\boldsymbol{\tau}$ provides directionality that can be used to orient neighboring edges in $\mathrm{G'}_{\mathrm{E}}$, thus allowing construction of a directed graph for the Markov chain. A naïve approach would prune all edges that violate the pseudo-time order to prevent de-differentiation paths. However, there is uncertainty in the pseudo-time estimate of the cells and therefore we use the estimated scaling factor for each cell in Equation 1 as a measure of the uncertainty in the pseudo-time estimate. Specifically, an undirected edge between cell i and its neighbor cell j is converted to a directed edge from cell i to cell j if $\tau_i < \tau_j$. The edge between cell i to cell j is pruned if $\tau_i > \tau_j$ and the distance between the two cells exceeds the scaling factor of cell i determined using Equation 1 (Supp. Fig. 1d-h).

Formally, undirected graph in the manifold, G_{E} is converted to directed graph, $\mathrm{G}_{\mathrm{D}} \in \mathbb{R}^{N \times N}$using

$$G_{D_{ij}} = \begin{cases} G'_{E_{ij}}\, if\, \tau_i < \tau_j \\ G'_{E_{ij}} if\, \tau_i > \tau_j and\, \tau_i - \tau_j < \sigma_i \\ 0\, if\, \tau_i > \tau_j and\, \tau_i - \tau_j > \sigma_i \end{cases} \qquad (14)$$

These distances are then converted to transition probabilities to construct the Markov chain. First, distances are transformed to an affinity matrix $\mathbf{Z} \in \mathbb{R}^{nW \times nW}$ using the kernel defined in

Equation (2) where nW is the number of waypoints. These affinities can be converted to probability matrix by dividing each affinity by the degree of the node in $\mathbf{Z}$ representing that cell.

$$P_{ij} = \frac{Z_{ij}}{\sum_k Z_{ik}} \tag{15}$$

The transition probability matrix $\mathbf{P}$ represents the Markov chain of the manifold, where P_{ij} represents the probability of reaching a cell in state j from a cell in state i in one step. As a first degree of approximation, our approach assumes that this probability of transition corresponds to the degree of cell state similarity between i and j. While development is a closely regulated process, at these very close distances, stochastic molecular processes of degradation and transcription likely play a significant role in. At longer distances, the regulatory processes driving development are implicitly encoded in the defined structure of the manifold graph $\mathrm{G_D}$. That is, the probability of reaching a cell in state j from *a more distinct cell* in state i is computed over the course of *many steps* and will be high if *many paths* connect them, i.e., there is high density of *observed* intermediary cell states between them.

By definition terminal states are not expected to differentiate further, thus to ensure that the random walks terminate when a terminal state is reached, all outgoing edges are removed from terminal states. Terminal states can be externally defined based on prior knowledge or can be computationally derived directly from the Markov chain using no additional knowledge, as we describe below. Given a set of terminal states TS, we convert the Markov chain $\mathbf{P}$ into an absorbing Markov chain $\mathbf{A}$ by setting the terminal states as absorbing states i.e., a state with no outgoing edges.

$$A_{ij} = 0 | \; i \in TS; j = 1..nW \qquad (16)$$

Identifying Terminal States

The graph structure and its associated Markov chain can be used to infer the terminal states directly from the data, using only the initial starting point as prior information. In the Markov chain $\mathbf{P}$, random walks tend to move in the direction of the terminal states. Since a pseudo-time underlies the Markov chain, we expect random walks to converge into the terminal states at the boundaries of the manifold. If the graph construction were perfect, we expect that these terminal states have no outgoing edges and thus be absorbing states. However, the chain was constructed with implicit uncertainty (e.g. the backward edges within the range of the scaling factor) and is therefore imperfect. Nevertheless, as the random walks are directed towards the terminal states, the steady state distribution of the Markov chain is expected to impart high probabilities to terminal states and states proximal to them as opposed to the intermediate states. Thus, Palantir identifies terminal states as extrema of diffusion components (boundary cells, C), that are also outliers in the steady state distribution of the Markov chain (Supp. Fig. 27).

The stationary distribution is the probability distribution over the states of the Markov chain that remains invariant as time progresses, i.e. the steady state distribution. Formally, if $\boldsymbol{\pi}$ represents the stationary distribution, then $\boldsymbol{\pi} = \mathbf{P} * \boldsymbol{\pi}.$ The first *left* Eigen vector of the Markov chain $\mathbf{P}$

represents the stationary distribution and is thus easy to compute. The outliers in this distribution can be identified using the Gaussian percent point function (i.e., inverse of the cumulative distribution function) using the median absolute deviation of the stationary distribution as the scale. Median absolute deviation is a robust measure of variance in univariate data [13]. Let $\boldsymbol{\pi}$ represent the stationary distribution. The median absolute deviation is computed as

$$sc = \mathrm{Median}\big(\pi_i - \mathrm{Median}(\boldsymbol{\pi}\,)\big) \tag{17}$$

The outliers are identified as

$$TS^{cands} = \{i|\pi_i > \mathrm{gaussian\_ppf}(0.9999, \mathrm{Median}(\boldsymbol{\pi}\,), sc)\} \tag{18}$$

This threshold robustly identifies the different terminal states across the different data sets. The set of states in TS^{cands} that are also diffusion component extremes are chosen as the terminal states of the system (Supp. Fig. 27, Fig 1c).

$$TS = \bigcap\big(TS^{cands}, C\big) \tag{19}$$

Cell fate/differentiation potential characterization

Random walks through the Markov chain between intermediate and terminal states can be used to compute the probability of a cell starting at an intermediate state reaching the corresponding terminal state. For each cell, we wish to calculate its branch probability vector $\mathbf{B}_{\boldsymbol{i}}$, denoting the probabilities it might reach each of b absorbing terminal states. An advantage in modeling differentiation as an absorbing Markov chain is that the branch probabilities can be computed as follows:

The absorbing Markov chain $\mathbf{A}$ can be represented as

$$\mathbf{A} = \begin{bmatrix} \mathbf{Q} & \mathbf{R} \\ 0 & \mathbf{I} \end{bmatrix} \tag{20}$$

Where $\mathbf{Q}$ is a $(nW - b) \times (nW - b)$ matrix of transition probabilities between intermediate states, $\mathbf{R}$ is a $(nW - b) \times b$ matrix of probabilities between intermediate states and terminal states and $\mathbf{I}$ is a $b \times b$ identity matrix.

Next the fundamental matrix $\mathbf{F}$ is computed using

$$\mathbf{F} = (\mathbf{I} - \mathbf{Q})^{-\mathbf{1}} \tag{21}$$

F_{ij} represents the probability of reaching intermediate state j from another intermediate state i in $1, 2, \ldots \infty$ steps

The fundamental matrix is then used to compute the differentiation probabilities

$$\mathbf{B} = \mathbf{F} * \mathbf{R} \tag{22}$$

where B_{ij} represents the probability of cell in intermediate state i reaching the terminal state j in $1, 2, \ldots \infty$ steps. $\mathbf{B}_i$ is a multinomial probability distribution such that $\sum_j B_{ij} = 1$. The branch probabilities of terminal states are set as follows.

$$B_{ij} = \begin{cases} 1 \; if \; i == j \\ 0 \; if \; i <> j \end{cases} \tag{23}$$

The waypoint branch probabilities are projected onto all the cells using weighting scheme defined in Equation (13)

$$B_{ij} = \sum_{w \in WS} B_{wj} * W_{wi} \tag{24}$$

Finally, we define the differentiation potential of each state to be the entropy of the branch probability vector $\mathbf{B}_i$ and this captures the degree of uncertainty in final terminal state (Fig. 1d).

Differentiation potential is a quantitative measure of the cell fate plasticity and represents the potential set of terminal states that a cell in an intermediate state can reach. Greater the entropy, higher is the number of terminal states the can potentially be reached by the cell in a particular state. As a result, the cells at the beginning of the pseudo-time are associated with the highest differentiation potential (entropy) (Fig. d(1)) whereas cells close to terminal states have the lowest differentiation potential (Fig. d(3, 7)). Crucially, differentiation potential captures the continuity in cell fate determination (Fig. 1d(2, 4-6)) and is a better representation of the differentiation processes as opposed to well-defined branch points. In summary, Palantir characterizes the continuity in both cell state and cell fate by modeling differentiation as a stochastic process.

Supplementary Note 2: Gene expression trends along lineages

Palantir's pseudo-time represents an ordering over all cells across all lineages and provides the position of all the cells relative to the start cell. In addition, Palantir branch probabilities represent the probability of a cell, in any state, to reaching each of the terminal states. Therefore, Palantir's ordering and branch probabilities represent a unified framework that enables computation and comparison of gene expression trends across the different lineages. This framework is used to compute gene expression trends for each lineage as follows: rather than segmenting the cells that belong to each lineage, the trend is computed using *all* the cells, each weighted by its probability to belong to that particular lineage. Cells that are not committed to a particular lineage can provide input to multiple lineages, whereas low probabilities naturally exclude cells that belong to unrelated lineages.

We take two approaches to improve the robustness and resolution of the computed trends: MAGIC [1] to impute missing values and generalized additive models [2] to determine robust trends. Gene expression trends are computed using MAGIC imputed [1] data to prevent dropouts from adversely affecting the trends. MAGIC imputes missing values for each cell based on cells that are most similar to it by using the covariate relationships between genes. Sliding window approaches on the other hand, average expression over many cells in a univariate manner, regardless of other genes. MAGIC, like Palantir is also based on diffusion maps and we use the same diffusion operator for both MAGIC and Palantir. We note that imputed data is only used to compute the expression trends after Palantir pseudo-time and branch probabilities have been computed using the non-imputed data.

Sliding window approaches are sensitive to density differences even with imputed data (Supp. Fig. 23d). We therefore used generalized additive models (GAMs) to determine gene expression trends along each lineage (Supp. Fig. 3), increasing robustness and rendering trends less sensitive to changes in cell density along the lineage. The gene expression trend for gene g and branch b is fit using

$$y_{gi} = \beta_o + f(\tau_i)\, for\ i\ \in B_{ib} > 0 \tag{25}$$

where y_{gi} is the expression of gene g in cell i and $\tau_i$is the pseudotime ordering of cell i. Cubic splines are used as the smoothing functions since they are effective in capturing non-linear relationships [2].

The pseudo-time is then divided into 500 equally sized bins and the smooth trend is derived by using the fit from Equation (24) to predict the expression of the gene at each bin (Supp. Fig. 3). The standard deviations of expression along each bin is determined by the standard deviation of the residuals of the fit and is computed as follows

$$SE\left(\widehat{y_p}\right) = \sqrt{1 + \frac{1}{n} + \frac{\left(\tau_p - \bar{\tau}\right)^2}{\sum_{j=1}^{n}\left(\tau_j - \bar{\tau}\right)^2}} \tag{26}$$

where $\widehat{y_p}$ is the predicted expression at bin p and $\bar{\tau}$ is the average pseudo-time across all cells. The computation and plotting of gene expression trends are demonstrated here:

http://nbviewer.jupyter.org/github/dpeerlab/Palantir/blob/master/notebooks/Palantir_sample_notebook.ipynb. Gene expression trends were computed using the R gam[2] package with default parameters.

Supplementary Note 3: Clustering of gene expression trends

Genes were selected based on significant differential expression as determined by MAST (FDR corrected p-value < 1e-2 and absolute log fold change > 1.25). Genes that were significantly high or low in stem and precursor cell clusters (0 or 1: 2176 genes) (Supp. Fig. 4a) were used for analysis in Fig. 3b and genes that were significantly high or low in early cell and erythroid cell clusters (0, 1, 2 or 8: 3322 genes) (Supp. Fig. 4a) were used analysis in Fig. 3c-d. Gene expression trends were z-transformed to put them on the same scale and clustered using Phenograph [14] (Supp. Fig. 13). A high-value of k (150) was used to avoid over-fragmentation of the gene trend clusters. The within cluster sum-of-squares each trend cluster was significantly lower for clusters derived from Phenograph when compared to trend matching techniques such as dynamic time warping. Gene ontology analysis was performed to annotate each cluster, measuring enrichment using the hypergeometric test. The following gene sets from Molecular Signature Database (MSigDB) (http://software.broadinstitute.org/gsea/msigdb/index.jsp) were tested: (a) c5 GO biological process gene set, (b) H hallmark gene sets and (c) c2 canonical pathway gene sets.

In order to compare the change of differentiation potential relative to gene expression changes, cells were divided into equal sized bins along the Palantir pseudo-time ordering. Mean expression of genes from the relevant clusters were determined to generate the histograms in Fig 3. The point of maximal differentiation potential change was determined using the second derivative. Stem and precursor cells (clusters 0 and 1) were used for analysis in Fig. 3b whereas these cells along the erythroid lineage (clusters 0, 2 and 8) were used for analysis in Fig. 3c-e. Similar analysis was performed for replicates 2 and 3 using the genes that were differential in replicate 1 (Supp. Fig. 14).

The computation and plotting of gene expression trend clustering for a particular lineage is shown here : http://nbviewer.jupyter.org/github/dpeerlab/Palantir/blob/master/notebooks/Palantir_sample_notebook.ipynb

Supplementary Note 4: GATA2 identification

We downloaded the list of human TFs from AnimalTFDb [15]. TFs that were significantly high in one of the erythroid clusters (clusters 2 or 8 - Supp. Fig. 4a) were annotated as erythroid TFs. We reason that a TF that potentially plays a role in lineage specification should correlate with the branch probability during the specification phase i.e., when the branch probability begins to increase along pseudo-time. We used the second derivative of the erythroid probability trend to approximate the point along the pseudo-time where there is a switch from lineage specification to functional commitment, since the second derivative indicates the point of maximal change in the trend (Supp. Fig. 15b).

We computed the correlation between TF expression trend and the erythroid probability trend for each erythroid TF defined above (Supp. Fig. 15b). To avoid down-weighting TFs that are potentially downregulated following commitment, during the functional specification phase, we computed the correlation until the point of maximal differentiation potential change (and not along the entire pseudo-time). Only TFs with both a high correlation and with sufficiently high expression levels in early cells were considered candidate erythroid specifiers. The comparison of branch probability correlation and mean expression in stem cell cluster (cluster 0 - Supp. Fig. 4a) shows that GATA2 is a clear outlier (Supp. Fig. 15b). GATA2 is also the only factor with high correlation and high progenitor cell expression for which a motif was identified in bulk ATAC-seq data.

Supplementary Note 5: Single cell transcription factor activities

Bulk ATAC-seq data processing

Bulk ATAC-seq data was downloaded from GEO (GSE75384) Reads were aligned to hg38 genome using bowtie2. PCR duplicates were removed using samtools, rmdup. Reads with fragment size < 150 bp, representing TF binding events [16] were retained for downstream analysis. After size selection, reads were pooled from all cell types and replicates. Only the first read from the pair was used for peak calling since a single transposase nick is sufficient proof for exposed chromatin [16]. Peak calling was performed using macs2 with a permissive p-value threshold of 1e-5 and with the parameter “nomodel” turned on to prevent shifting of positive and negative reads towards each other [17]. IDR [18] was then used to identify reproducible peaks for each cell type: IDR was performed on each pair of available replicates and a peak was assigned to a cell type if the peak passed IDR < 0.1 in at least 50% of the replicate comparisons.

Motif discovery

SeqGL [17] was run separately for each cell type with default parameters using the reproducible peaks for the respective cell type. SeqGL outputs a predicted sequence affinity for each TF, peak pair. The sequence affinity represents a quantitative measure of the k-mer sequence preferences: a higher value represents a greater chance that the TF binds at the genomic location spanned by the ATAC-seq peak.

Single cell TF activity

ATAC-seq peaks were assigned to the gene with the nearest transcription start site, which is a reasonable approximation of enhancer target assignment in the absence of chromosome interaction data [19]. Sequence affinities for all TF-gene pairs were determined by aggregating the affinities across all peaks assigned to gene. Recent studies have shown that these affinities correlate strongly with expression change of the targets indicating that the sequence affinities approximate the regulatory effect of a TF on its target [19]. Therefore, correlation between target expression and predicted TF sequence affinity was used as the TF activity for each cell (Fig. 4b). The activities were determined separately for promoter peaks (peaks within 2kb of the transcription start site) and enhancer peaks (peaks at distance > 2kb of the transcription start site). As a demonstration of the importance of cell type context for determining TF targets, Supp. Fig. 15h shows TF activity trends for Runx in different cell types. The targets of Runx, a transcriptional activator, show higher expression in the corresponding cell type, demonstrating the accuracy of computing TF activities using correlation between target expression and predicted sequence affinities

Supplementary Note 6: Performance of competing methods on the CD34+ marrow data

Palantir summary

Palantir was run using one of the CD34+ cells as the start. Palantir automatically determined all the different lineages and assigned for each cell, a probability of reaching the different terminal states. Palantir also provides a unified pseudo-time ordering to enable comparison of gene expression trends across lineages. The trends were as expected, based on ground truth derived from prior publications.

Monocle 2

Monocle 2 uses a reverse graph embedding which simultaneously learns a principal graph that approximates the low dimensional manifold and projection of cells onto this graph to reconstruct single cell trajectories.

Monocle 2 was run with default parameters for UMI counts specified in the Monocle 2 vignette to embed data into two dimensions. Monocle2 identified six distinct states in the data, but we could not attribute specific cell types to any of these states based on expression of marker genes (Supp. Fig. 18). Moreover, key canonical markers for progenitors (CD34), myeloid (MPO, IRF8) and B-cell lineages (CD79B) are spread across all projections with no trend or coherence (Supp. Fig. 18). Thus, Monocole2 failed to correctly compute pseudo-time, identify terminal fates and generate expression trends on this data. We note that in the original publication, Monocle2 was demonstrated on a small dataset, with well-distinguished, sorted populations, rather than a complex differentiating system.

Partition based Graph Abstraction (PAGA)

PAGA aims to reconcile clustering and trajectory inference and is particularly adept to scaling to large number of cells. PAGA generates an abstracted graph representing the differentiation structure underlying the data. The gene expression trends are fit by computing a pseudo time ordering for each lineage separately using diffusion pseudo time (DPT) and then a moving average along the resulting pseudo-time of cells. PAGA was run using default preprocessing steps outlined in https://github.com/theislab/paga/blob/master/blood/paul15/paul15.ipynb with log transformation of the normalized data.

PAGA succeeds in recovering the different hematopoietic lineages and their relationships (Supp. Fig. 19a-b) for the larger populations. However, PAGA embeds the megakaryocyte lineage cells into the erythroid cell lineage and is unable to distinguish between the two DC lineages (Supp. Fig. 19b). The abstracted graph constructed by PAGA represents the topological structure of the lineage decision process and the strong interconnectivity among clusters representing the intermediate states provides further evidence for lack of well-defined bifurcations in human hematopoiesis (Supp. Fig. 19b).

PAGA generates a unified pseudo-time ordering of cells and enables comparison of gene expression trends across lineages. PAGA uses a sliding window to infer gene expression trends and requires a manual specification of clusters that contribute to each particular lineage. PAGA's gene expression trends for the key markers are shown in Supp. Fig. 19c). While the patterns of CD34, MPO, CD79B and GATA1 are qualitatively consistent with their known behavior, the sliding window approach leads to a loss in resolution and makes the trend estimates very noisy (Supp. Fig. 19c). The trends for CSF1R and CD41 do not reflect known biology, since DC lineages are not distinguished and megakaryocyte cells are included as part of the erythroid lineage (Supp. Fig. 19c).

Diffusion Pseudotime (DPT)

Diffusion pseudotime (DPT) was primarily designed for estimating pseudo-time ordering of cells using diffusion maps. DPT also uses a heuristic based on Kendall Tau's correlation to identify the branches in the data by using the start cell and the number of branchings as input. DPT imposes tree-like structure to model the differentiation process, representing each bifurcation as a discrete point. We applied DPT using the scanpy implementation, following similar preprocessing steps to Palantir. We used the same start cell as Palantir and used 3 as the number of branchings, since this value gave us the best results (additional branches created splits that do not correspond to known lineages).

Since both Palantir and DPT are based on diffusion maps, the pseudo-time ordering of the cells are correlated (Supp. Fig. 19d). DPT was able to identify most of the lineages except for the distinction between DC lineages (Supp. Fig. 19e). DPT also suffers from a loss of resolution in characterizing gene expression trends, although the qualitative patterns for gene expression changes are correctly identified for all markers except for the CSF1R gene, since the two DC populations are not separated (Supp. Fig. 19f).

Slingshot

Slingshot takes as input a clustering and low dimensional embedding of the data. Slingshot first determines a minimum spanning tree through the clusters to identify the overall branch structure of the data. Slingshot then fits principle curves for each branch/ lineage and uses orthogonal projections against these principle curves to determine the pseudo-time ordering. The curves are fit separately for each lineage and hence Slingshot does not provide for comparison of gene expression dynamics across lineages. Finally, Slingshot uses GAMs with loess fits to determine gene expression trends. Slingshot does not make explicit recommendations for dimensionality reduction and clustering algorithms, both required as input. Therefore, to maximize similarity to Palantir, we applied Slingshot to the hematopoiesis data using diffusion maps and Phenograph clusters as input.

Slingshot identifies four lineages from the data: (monocyte, lymphoid, erythroid and DC) (Supp. Fig. 20a). Similar to PAGA, Slingshot fails to distinguish between the two DC clusters (since they are clustered together) and embeds the megakaryocyte population to be a stage along erythroid lineage, even though these are clustered separately (Supp. Fig. 20a, Supp. Fig 4a). Slingshot accurately recovers the gene expression trends of MPO and GATA1 in myeloid and erythroid lineages respectively (Supp. Fig. 20c). While the CD79B upregulation in CLP is also identified, there is an unexpected downregulation at the beginning of the CLP ordering since the cells committing towards the myeloid lineages are included as part of the lymphoid lineage (Supp. Fig. 20c). Since the gene expression trends for the two DC lineages are identical, we cannot distinguish between expression dynamics of key DC TFs such as CSF1R (Supp. Fig. 20c) and CD41 dynamics in erythroid and megakaryocyte lineages cannot be characterized since megakaryocytes are absorbed into the erythroid lineage (Supp. Fig. 20c).

FateID

FateID aims to compute fate biases of each cell, towards each of the pre-specified terminal states. FateID determines these probabilities by using a random forest classifier applied successively for cells from the terminals to the start of the trajectories. The cells that belong to a particular lineage are determined based on the identified fate biases. Pseudo-time ordering is determined separately for each lineage by fitting principal curves through low dimensional embedding such as tSNE. As such, FateID does not allow for a comparison of gene expression trends across different lineages. FateID requires pre-specification of clustering and pre-specification of the terminal states.

FateID recommends using RaceID to cluster the data. However, RaceID led to over clustering of the data and did not yield coherent clusters representing the different lineages (17 clusters, none matching known lineages or cell types, Supp. Fig. 21a) and thus running FateID as recommended by its authors, fails to generate a coherent map of hematopoiesis. Therefore, to maximize similarity, we used Palantir's preprocessing, clusters and terminal fates as inputs to FateID to maximize similarity to Palantir and this indeed resulted in better representative results of human hematopoiesis (Supp. Fig. 21). However, there were a number of serious issues with the derived fate biases: (i) all of the early cells (20% of the cells) are included exclusively as part of the lymphoid lineage which renders cell fate probabilities for these cells incorrect and affects the computation of gene expression trends of different lineages, We note that FateID was originally developed to study the lymphoid lineages and hence the design and modeling choices might over-fit towards this lineage, (ii) all later precursors included exclusively as part of the myeloid lineages (17% of cells), providing no real precursors for erythroid and megakaryocytic lineages, which are earliest to branch off, (iii) lack of distinction between the DC lineages since they are not separate clusters, and (iii) erythroid precursors included exclusively as part of the megakaryocytic lineage (Supp. Fig. 21b-c). Since FateID also generates probabilities or fate biases, we next compared the changing probabilities in cells committing towards the erythroid and monocyte lineage (Supp. Fig. 12). We do not observe any consistent change in probabilities in the erythroid lineage (Supp. Fig. 21d - left panel) whereas the commitment towards the monocytic lineage is rather abrupt (Supp. Fig. 21d - right panel). Thus, the cell fate probabilities are largely incorrect for a large fraction of the cells and do not follow the correct hierarchy for hematopoiesis.

We next compared the gene expression trends along individual lineages since FateID does not generate a unified ordering for comparison of trends across lineages (Supp. Fig. 21f). The inclusion of all the early cells specifically in the lymphoid lineage results in correct identification of CD79B trend in the lymphoid lineage but leads to a number of issues in the gene expression trends: (i) GATA1 does not show the expected upregulation in the erythroid lineage, (ii) MPO shows a high basal level of expression at the earliest stages of ordering, (iii) CD41 does not show the expected upregulation in megakaryocyte lineage since cells of the erythroid lineage are included as part of this lineage (Supp. Fig. 21f). In summary, FateID performs poorly on this data.

Diffusion maps

Diffusion maps are widely applied for pseudo-time ordering of cells by projecting cells along *individual* components to determine pseudo-time for a lineage. Gene expression trends are then estimated by using a sliding window approach along these projections [9, 10]. There are three key limitations to this approach: First, projection of cells onto a *single* diffusion component does not always generate an accurate ordering of cells along a lineage. Second, diffusion maps generate a projection of *all* cells along each component and therefore segmentation of the data (e.g. based on clustering) is necessary to determine gene expression dynamics. Finally, projection of cells along different components does not allow for a *direct comparison* of dynamics between two different lineages.

In particular, our data demonstrate that only the monocytic and lymphoid lineages can be unambiguously explained by a single diffusion component; all other lineages require multiple components to accurately determine pseudo-time (Supp. Fig. 2b). Bearing these limits in mind, we used projection of lymphoid lineage cells to characterize the effect of using individual diffusion components for determining pseudo-time and gene expression trends (Supp. Fig. 22a). Projections along this component amplify the already-significant density differences in the data (Supp. Fig. 22b). Sliding window approaches are particularly sensitive to density differences and as a result, the expression trend estimates are not reliable. In this particular case, loss of resolution in the sliding window approach prevents an accurate characterization of key TFs such as PU.1 (Supp. Fig. 22b), which has been shown to play a key role in lymphoid specification[20].

Population Balance Analysis (PBA)

A recently published approach, population balance analysis (PBA) [21, 22] presents a framework to characterize differentiation using spectral graph theory to solve a system of differential equations representing the dynamics of maturation along a lineage. In practice, this translates to using Markov chains for characterizing differentiation, providing further support for this approach to model differentiation. We note that PBA requires extensive use of prior knowledge to infer the proliferation and loss rates for each state (cell) in the system, which form the fundamental basis for the Markov chain construction. In the particular case of mouse hematopoiesis, the rates for the different lineages were estimated separately using data from multiple fate mapping studies [21]. In addition, PBA requires explicit specification of the terminal states in the system *a priori*. We could not apply PBA to human hematopoiesis owing to paucity of such fate mapping studies in human. In contrast, Palantir requires only specification of an early cell and can automatically construct the Markov chain and determine the set of terminal states based on single cell RNA-seq measurements alone. This data-driven approach to characterizing differentiating systems is a key strength of Palantir's utility and applicability to model tissue systems without established lineages.

References

1. van Dijk, D. et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716-729 e727 (2018).
2. Hastie, T.J. & Tibshirani, R.J. Generalized Additive Models. . (Chapman & Hall/CRC, 1990).
3. Paul, F. et al. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **163**, 1663-1677 (2015).
4. Herring, C.A. et al. Unsupervised Trajectory Analysis of Single-Cell RNA-Seq and Imaging Data Reveals Alternative Tuft Cell Origins in the Gut. *Cell Syst* **6**, 37-51 e39 (2018).
5. Coifman, R.R. et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc Natl Acad Sci U S A* **102**, 7426-7431 (2005).
6. Setty, M. et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature biotechnology* **34**, 637-645 (2016).
7. Haghverdi, L., Buttner, M., Wolf, F.A., Buettner, F. & Theis, F.J. Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods* **13**, 845-848 (2016).
8. Bendall, S.C. et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714-725 (2014).
9. Haber, A.L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333-339 (2017).
10. Ibarra-Soria, X. et al. Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation. *Nature cell biology* **20**, 127-134 (2018).
11. Tenenbaum, J.B., de Silva, V. & Langford, J.C. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319-2323 (2000).
12. de Silva, V. & Tenenbaum, J.B. (ed. S. University) (2004).
13. Leys, C., Ley, C., Klein, O., Bernard, P. & Licata, L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology* **49** (2013).
14. Levine, J.H. et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184-197 (2015).
15. Zhang, H.M. et al. AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res* **43**, D76-81 (2015).
16. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213-1218 (2013).
17. Setty, M. & Leslie, C.S. SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. *PLoS Comput Biol* **11**, e1004271 (2015).
18. Li, Q., Brown, J.B., Huang, H. & Bickle, P.J. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics* **5**, 1752-1779 (2011).
19. Gonzalez, A.J., Setty, M. & Leslie, C.S. Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. *Nature genetics* **47**, 1249-1259 (2015).
20. Rosenbauer, F. & Tenen, D.G. Transcription factors in myeloid development: balancing differentiation with transformation. *Nat Rev Immunol* **7**, 105-117 (2007).
21. Rodriguez-Fraticelli, A.E. et al. Clonal analysis of lineage fate in native haematopoiesis. *Nature* **553**, 212-216 (2018).

22. Weinreb, C., Wolock, S., Tusi, B.K., Socolovsky, M. & Klein, A.M. Fundamental limits on dynamic inference from single-cell snapshots. *Proc Natl Acad Sci U S A* **115**, E2467-E2476 (2018).