Cell Stem Cell, Volume *19*

Supplemental Information

De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data

Dominic Grün, Mauro J. Muraro, Jean-Charles Boisset, Kay Wiebrands, Anna Lyubimova, Gitanjali Dharmadhikari, Maaike van den Born, Johan van Es, Erik Jansen, Hans Clevers, Eelco J.P. de Koning, and Alexander van Oudenaarden

SUPPLEMENTAL FIGURES

Figure S1. RaceID2 improves robustness of clustering. (Related to Figure 1)
(A) Gap statistitic (Tibshirani et al., 2001) computed with k-means clustering of the similarity matrix as in RaceID (left) and with k-medoids clustering using 1- pearson's correlation directly as clustering distance metric as in RaceID2 (right). (B) Jaccard's similarity computed by bootstrapping for k-means (upper panel) and k-medoids (lower panel) clustering with 5 clusters. K-medoids clustering shows higher reproducibility. (C) Criterion for the selection of the cluster number used for k-medoids clustering. If the change of the within-cluster dispersion (Tibshirani et al., 2001) upon increasing the cluster number ($k_{i+1} = k_i + 1$) is within the error of the average change upon further increase ($k_{i+2}, \ldots, k_{\max}$), k_i is chosen as input. The average change across cluster numbers $k_{i+2}, \ldots, k_{\max}$ and its error is computed from a linear regression. The within-cluster dispersion as a function of k is shown on the left. The right panel shows the change of the within-cluster dispersion as a function of k and the average dispersion for higher values of k with error bars (red). In both panels the selected cluster number is circled in blue. (D) Outliers identification by RaceID2 is the same as in RaceID. Shown is the number of outliers as a function of the p-value cutoff. The red line indicates the cutoff chosen for this work ($P<10^{-3}$).

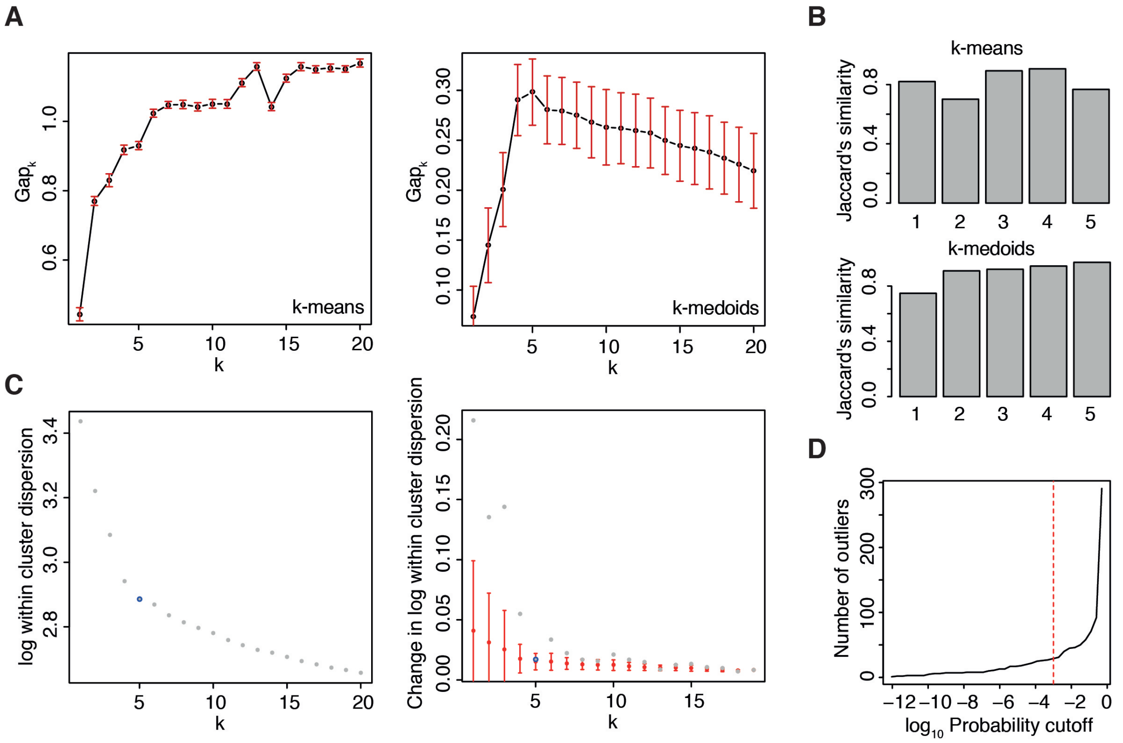

A
Gap$_k$
0.6
0.8
1.0
k-means
5
10
15
20
k
Gap$_k$
0.10
0.15
0.20
0.25
0.30
k-medoids
5
10
15
20
k
B
k-means
Jaccard's similarity
0.0
0.4
0.8
1
2
3
4
5
k-medoids
Jaccard's similarity
0.0
0.4
0.8
1
2
3
4
5
C
log within cluster dispersion
2.8
3.0
3.2
3.4
5
10
15
20
k
Change in log within cluster dispersion
0.00
0.05
0.10
0.15
0.20
5
10
15
k
D
Number of outliers
0
100
200
300
−12
−10
−8
−6
−4
−2
0
log$_{10}$ Probability cutoff

Figure S1

Figure S2. Lineage inference by StemID and comparison to an alternative method for the derivation of differentiation trajectories does not resolve secretory intestinal cells. (Related to Figure 2) (A) The heatmap shows the $\log_2$-ratio of the cell number assigned to each link between RaceID2 clusters and the expected number computed by a background model with randomized cell positions. Only significantly enriched or depleted links are highlighted. A $\log_2$-ratio of zero is assigned to all other links. (B) The heatmap shows the link score for each pair of clusters, reflecting how densely a link between clusters is populated with cells (see Experimental Procedures). Values close to one indicate dense coverage, while values close to zero indicate that cells are concentrated near the centers of the clusters connected by the link. A higher value reflects a higher likelihood that the link represents an actual differentiation trajectory. (C-E) The Monocle (Trapnell et al., 2014) algorithm was run on the single cell transcriptomes of the 5 days *Lgr5* lineage tracing data. (A) Minimum spanning tree computed by Monocle. Since 5 different cell types were observed in the data, Monocle was run with num_paths = 4. RaceID2 clusters were highlighted by numbers and colors used in Figure 1. (B) Expression of lineage markers (*Apoe1*: enterocytes; *Chga*: mature enteroendocrine cells; *Chgb*: early and mature enteroendocrine cells; *Clca3*: Goblet cells; *Clca4*: crypt bottom columnar cells; *Defa24*: Paneth cells) in cells assembled in pseudo-temporal order computed by Monocle. (C) Transcript counts (color legend) of mature lineage markers highlighted in the t-SNE map. RaceID2 clusters reliably discriminate different cell types (see Figure 2C). Monocle assigns stem, goblet, Paneth and enteroendocrine cells to one state and the inferred pseudo-temporal order does not reflect the published one shown in Figure 1A and inferred by StemID.

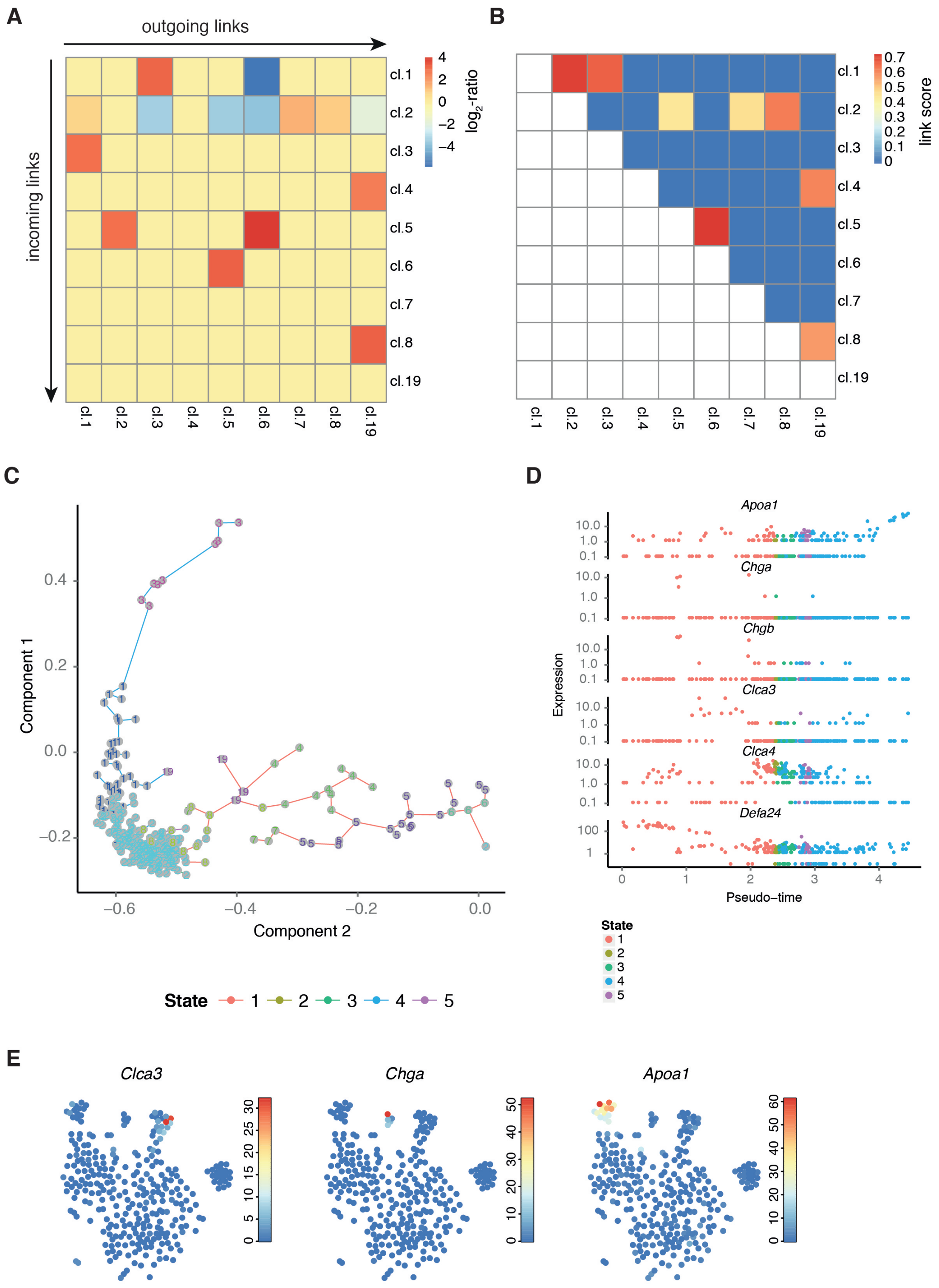

A
outgoing links
incoming links
cl.1
cl.2
cl.3
cl.4
cl.5
cl.6
cl.7
cl.8
cl.19
4
2
0
−2
−4
log$_2$-ratio
B
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
link score
C
Component 1
Component 2
0.4
0.2
0.0
−0.2
−0.6
−0.4
−0.2
0.0
State
1
2
3
4
5
D
Apoa1
Chga
Chgb
Clca3
Clca4
Defa24
Expression
10.0
1.0
0.1
100
1
Pseudo−time
0
1
2
3
4
State
1
2
3
4
5
E
Clca3
Chga
Apoa1
0
5
10
15
20
25
30
40
50
60

Figure S2

Figure S3. StemID identifies stem cells in a complex intestinal dataset. (Related to Figure 3)

We ran RaceID2 and StemID on a dataset combining single mouse intestinal cell transcriptome data from a variety of experiments conducted in our lab, comprising *Cd24*-positive secretory cells, 3 weeks old progeny of *Lgr5*-positive cells and a sub-population of those positive for *Cd24*, and 8 weeks old *Cd24*-positive progeny of Lgr5-positive cells. (A) Heatmap of cell-to-cell transcriptome distances measured by 1 – Pearson's correlation (ρ) coefficient. RaceID2 cluster are color coded along the boundaries. (B) t-SNE map representation of transcriptome similarities between individual cells. Different experiments are highlighted with different colors and symbols. (C) The heatmap shows the $\log_2$-ratio of the cell number assigned to each link between RaceID2 clusters and the expected number computed by a background model with randomized cell positions. Only significantly enriched or depleted links are highlighted. A $\log_2$-ratio of zero is assigned to all other links. (D) The heatmap shows the link score for each pair of clusters, reflecting how densely a link between clusters is populated with cells (see Experimental Procedures). Values close to one indicate dense coverage, while values close to zero indicate that cells are concentrated near the centers of the clusters connected by the link. A higher value reflects a higher likelihood that the link represents an actual differentiation trajectory. (E) t-SNE map showing the projections of all cells as computed in a high dimensional space (see Experimental Procedures) in the embedded two-dimensional space. The black solid line indicates a minimum spanning tree connecting the cluster centers, which was computed based on the distances between cluster centers. The minimum spanning tree recovers the main differentiation trajectories, but does not identify a number of alternative trajectories revealed by the projection-based approach. (F) Barplot of the number of links (upper panel) and the Δentropy (lower panel). A comparison to the StemID score (Figure 3C) shows that neither of these quantities alone could rank the cell types by pluripotency with the same specificity as the StemID score.

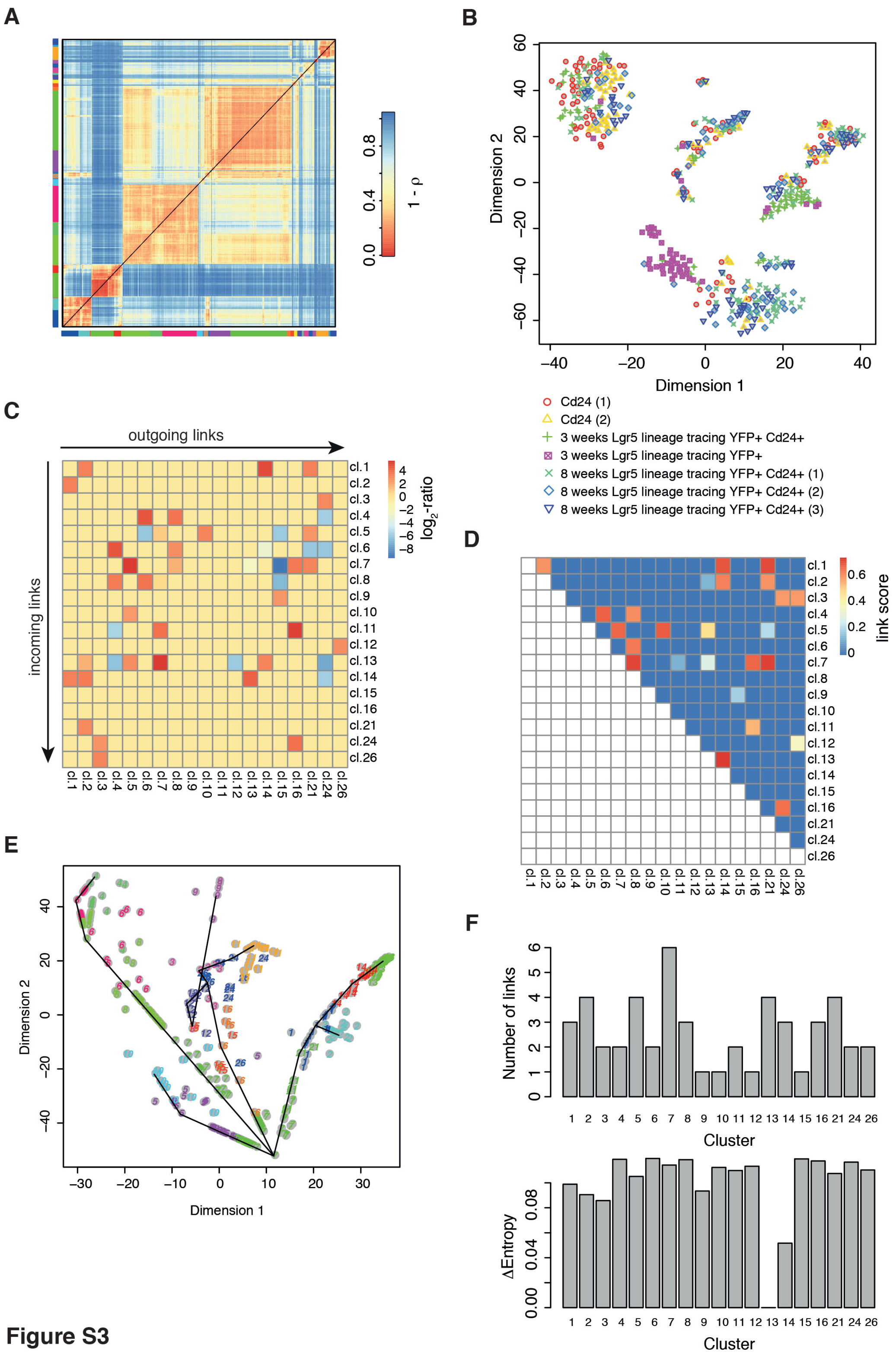
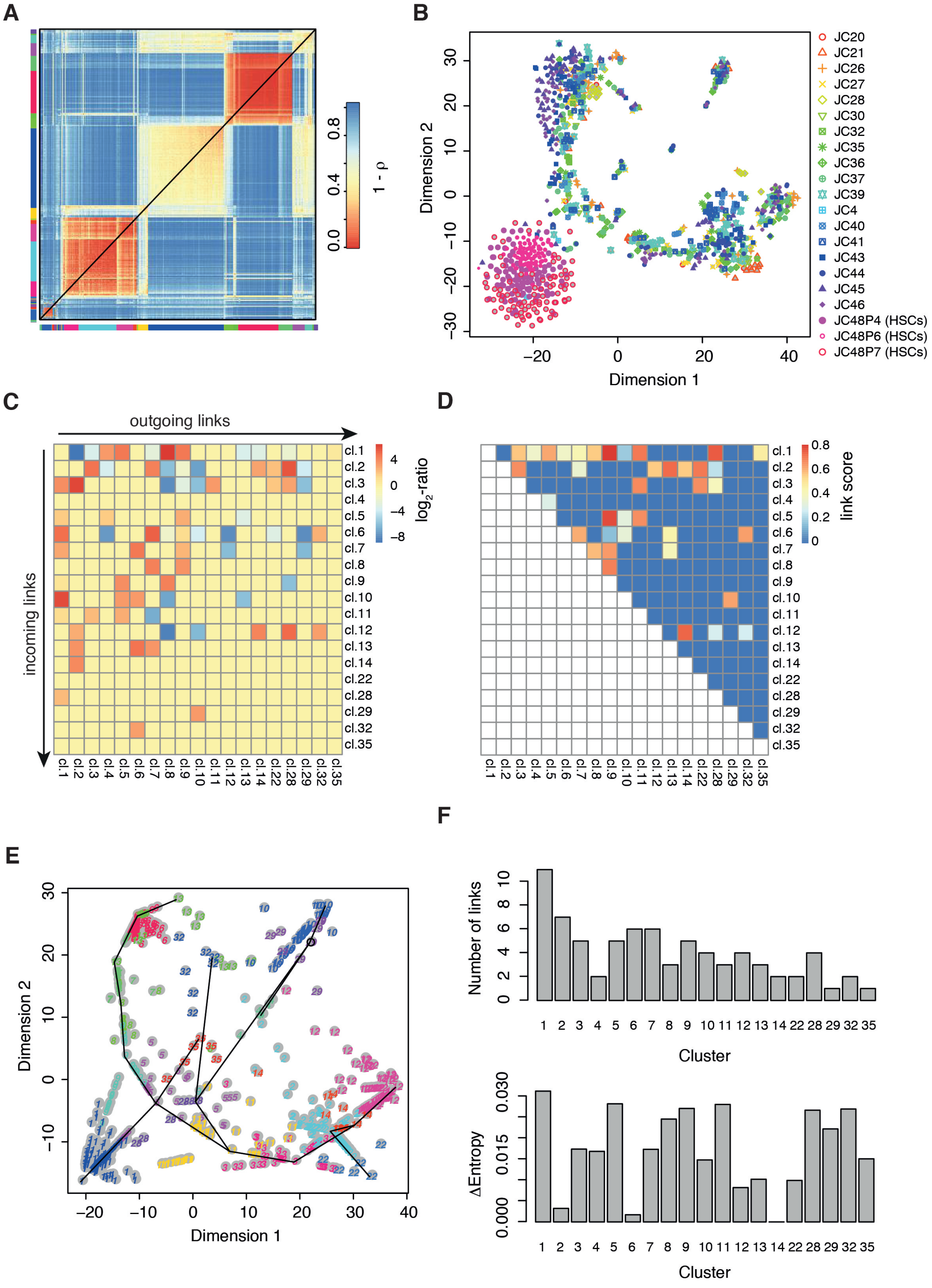

A
1 - ρ
0.0
0.4
0.8
B
Dimension 2
Dimension 1
60
40
20
0
−20
−40
−60
−40
−20
0
20
40
○ Cd24 (1)
△ Cd24 (2)
+ 3 weeks Lgr5 lineage tracing YFP+ Cd24+
⊠ 3 weeks Lgr5 lineage tracing YFP+
× 8 weeks Lgr5 lineage tracing YFP+ Cd24+ (1)
◇ 8 weeks Lgr5 lineage tracing YFP+ Cd24+ (2)
▽ 8 weeks Lgr5 lineage tracing YFP+ Cd24+ (3)
C
outgoing links
incoming links
log2-ratio
4
2
0
−2
−4
−6
−8
cl.1
cl.2
cl.3
cl.4
cl.5
cl.6
cl.7
cl.8
cl.9
cl.10
cl.11
cl.12
cl.13
cl.14
cl.15
cl.16
cl.21
cl.24
cl.26
D
link score
0.6
0.4
0.2
0
cl.1
cl.2
cl.3
cl.4
cl.5
cl.6
cl.7
cl.8
cl.9
cl.10
cl.11
cl.12
cl.13
cl.14
cl.15
cl.16
cl.21
cl.24
cl.26
E
Dimension 2
Dimension 1
40
20
0
−20
−40
−30
−20
−10
0
10
20
30
F
Number of links
0
1
2
3
4
5
6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 21 24 26
Cluster
ΔEntropy
0.00
0.04
0.08
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 21 24 26
Cluster

Figure S3

Figure S4. StemID identifies hematopoietic stem cells in single cells sequenced from the bone marrow. (Related to Figure 4)

We ran RaceID2 and StemID on a single cell sequencing dataset comprising mouse bone marrow cells manually isolated from interacting doublets or multiplets of cells and Kit$^+$ Sca-1$^+$ Lin$^-$ CD48$^-$ CD150$^+$ hematopoietic stem cells (HSCs). (A) Heatmap of cell-to-cell transcriptome distances measured by 1 – Pearson's correlation (ρ) coefficient. RaceID2 cluster are color coded along the boundaries. (B) t-SNE map representation of transcriptome similarities between individual cells. Different experiments are highlighted with different colors and symbols. (C) The heatmap shows the $\log_2$-ratio of the cell number assigned to each link between RaceID2 clusters and the expected number computed by a background model with randomized cell positions. Only significantly enriched or depleted links are highlighted. A $\log_2$-ratio of zero is assigned to all other links. (D) The heatmap shows the link score for each pair of clusters, reflecting how densely a link between clusters is populated with cells (see Experimental Procedures). Values close to one indicate dense coverage, while values close to zero indicate that cells are concentrated near the centers of the clusters connected by the link. A higher value reflects a higher likelihood that the link represents an actual differentiation trajectory. (E) t-SNE map showing the projections of all cells as computed in a high dimensional space (see Experimental Procedures) in the embedded two-dimensional space. The black solid line indicates a minimum spanning tree connecting the cluster centers, which was computed based on the distances between cluster centers. The minimum spanning tree recovers the main differentiation trajectories, but does not identify a number of alternative trajectories revealed by the projection-based approach. (F) Barplot of the number of links (upper panel) and the Δentropy (lower panel). A comparison to the StemID score (Figure 4C) shows that neither of these quantities alone could rank the cell types by pluripotency with the same specificity as the StemID score.



A
1 - ρ
0.0
0.4
0.8
B
Dimension 1
Dimension 2
−20
0
20
40
−30
−20
−10
0
10
20
30
JC20
JC21
JC26
JC27
JC28
JC30
JC32
JC35
JC36
JC37
JC39
JC4
JC40
JC41
JC43
JC44
JC45
JC46
JC48P4 (HSCs)
JC48P6 (HSCs)
JC48P7 (HSCs)
C
outgoing links
incoming links
log$_2$-ratio
4
0
−4
−8
cl.1
cl.2
cl.3
cl.4
cl.5
cl.6
cl.7
cl.8
cl.9
cl.10
cl.11
cl.12
cl.13
cl.14
cl.22
cl.28
cl.29
cl.32
cl.35
D
link score
0.8
0.6
0.4
0.2
0
cl.1
cl.2
cl.3
cl.4
cl.5
cl.6
cl.7
cl.8
cl.9
cl.10
cl.11
cl.12
cl.13
cl.14
cl.22
cl.28
cl.29
cl.32
cl.35
E
Dimension 1
Dimension 2
−20
−10
0
10
20
30
40
−10
0
10
20
30
F
Number of links
0
2
4
6
8
10
Cluster
1 2 3 4 5 6 7 8 9 10 11 12 13 14 22 28 29 32 35
ΔEntropy
0.000
0.015
0.030
Cluster
1 2 3 4 5 6 7 8 9 10 11 12 13 14 22 28 29 32 35

Figure S4

Figure S5. StemID predicts pluripotent cells in random mixtures of human pancreatic cells. (Related to Figure 6)

We ran RaceID2 and StemID on a single cell sequencing dataset comprising single human pancreatic cells isolated form five different donor (D2, D3, D7, D10, D17). Different enrichment strategies were applied to collect random mixture, endocrine and exocrine cells, or subsets of those. (A) Heatmap of cell-to-cell transcriptome distances measured by 1 – Pearson's correlation coefficient (ρ). RaceID2 cluster are color coded along the boundaries. (B) t-SNE map representation of transcriptome similarities between individual cells. Different experiments are highlighted with different colors and symbols. (C) The heatmap shows the $\log_2$-ratio of the cell number assigned to each link between RaceID2 clusters and the expected number computed by a background model with randomized cell positions. Only significantly enriched or depleted links are highlighted. A $\log_2$-ratio of zero is assigned to all other links. (D) The heatmap shows the link score for each pair of clusters, reflecting how densely a link between clusters is populated with cells (see Experimental Procedures). Values close to one indicate dense coverage, while values close to zero indicate that cells are concentrated near the centers of the clusters connected by the link. A higher value reflects a higher likelihood that the link represents an actual differentiation trajectory. (E) t-SNE map showing the projections of all cells as computed in a high dimensional space (see Experimental Procedures) in the embedded two-dimensional space. The black solid line indicates a minimum spanning tree connecting the cluster centers, which was computed based on the distances between cluster centers. The minimum spanning tree recovers the main differentiation trajectories, but does not identify a number of alternative trajectories revealed by the projection-based approach. (F) Barplot of the number of links (upper panel) and the Δentropy (lower panel). A comparison to the StemID score (Figure 6C) shows that neither of these quantities alone could rank the cell types by pluripotency with the same specificity as the StemID score.

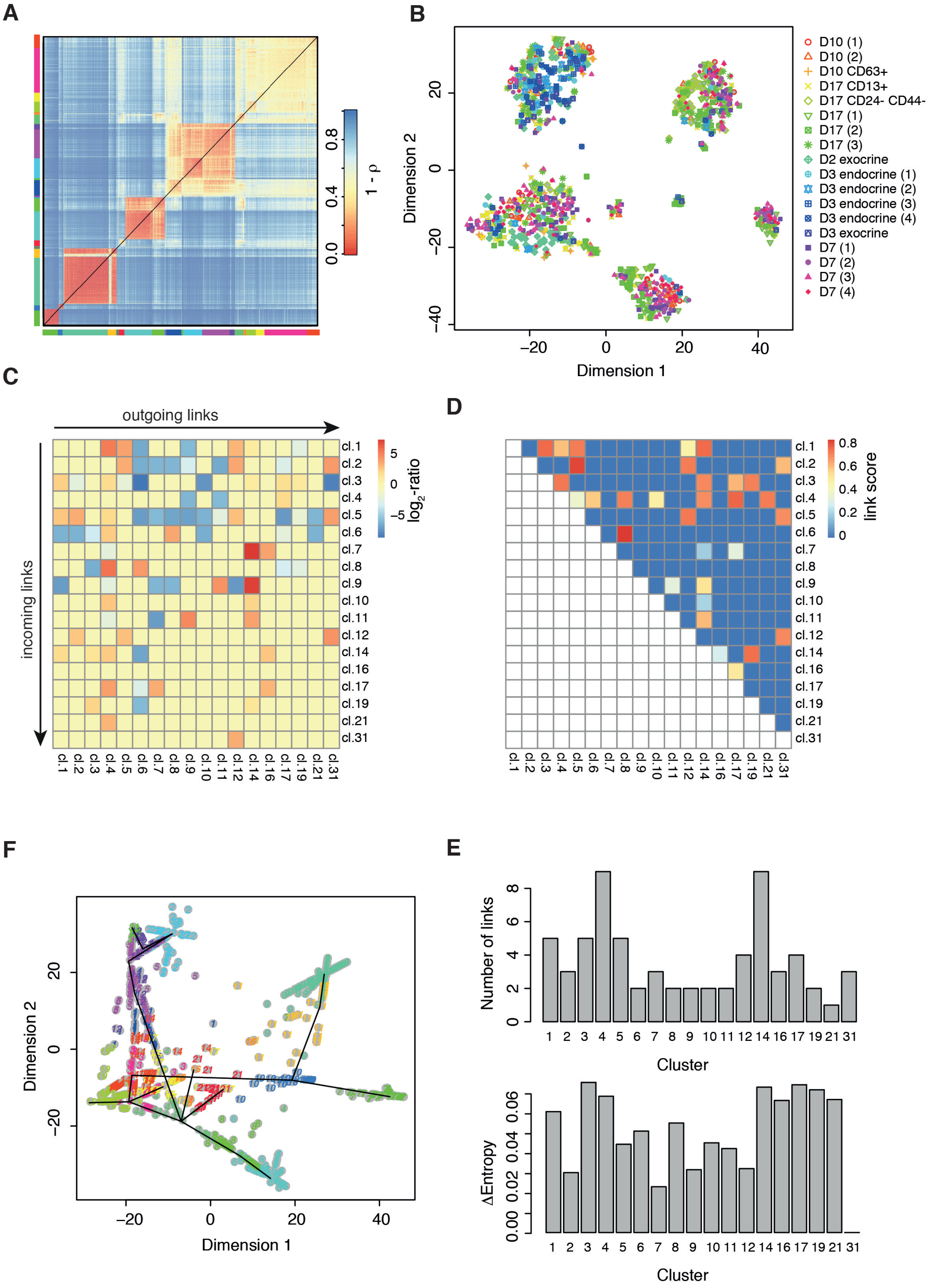

A
1 - ρ
0.0
0.4
0.8
B
Dimension 1
Dimension 2
−40
−20
0
20
40
D10 (1)
D10 (2)
D10 CD63+
D17 CD13+
D17 CD24- CD44-
D17 (1)
D17 (2)
D17 (3)
D2 exocrine
D3 endocrine (1)
D3 endocrine (2)
D3 endocrine (3)
D3 endocrine (4)
D3 exocrine
D7 (1)
D7 (2)
D7 (3)
D7 (4)
C
outgoing links
incoming links
$\log_2$-ratio
5
0
−5
cl.1
cl.2
cl.3
cl.4
cl.5
cl.6
cl.7
cl.8
cl.9
cl.10
cl.11
cl.12
cl.14
cl.16
cl.17
cl.19
cl.21
cl.31
D
link score
0
0.2
0.4
0.6
0.8
F
Dimension 1
Dimension 2
−20
0
20
40
E
Number of links
Cluster
ΔEntropy
0.00
0.02
0.04
0.06
0
2
4
6
8
1 2 3 4 5 6 7 8 9 10 11 12 14 16 17 19 21 31

Figure S5

Figure S6. StemID provides novel information in comparison to published methods. (Related to Figure 3 and 4)

For the complex intestinal data set (Fig. 3) and the bone marrow data (Fig. 4) we derived a lineage tree with two previously published methods. On the one hand we used Monocle (Trapnell et al., 2014), which constructs a minimum spanning tree connecting all cells based on transcriptome similarity, and on the other hand we applied a recent method based on diffusion maps (Haghverdi et al., 2015). Results of Monocle and diffusion maps are shown in (A) and (B) for the intestinal data and in (C) and (D) for the bone marrow data. For the intestinal data (A, B) both methods reveal major branches (Paneth/goblet cells, tuft cells, enterocytes, compare to Figure 3 for colors and cluster labels). However, the small clusters of different enteroendocrine cells could not be assembled onto a branched tree by any method. Moreover, none of the methods reveals that Paneth and goblet cells have a common precursor, but rather place mature *Clca3* expressing goblet cells on the same branch with mature Paneth cells. Monocle does not recover the relation between TA cells and mature enterocytes. Crucially, none of these methods provides a cell type inference and a prediction of the stem cell identity. For both methods, it is not apparent from the topology that cluster 7 represents the stem cell identity.

For the bone marrow data (C, D) both methods recover the major branches of neutrophils and erythroblasts, but intermingle the low frequency cell types with myeloid precursors.

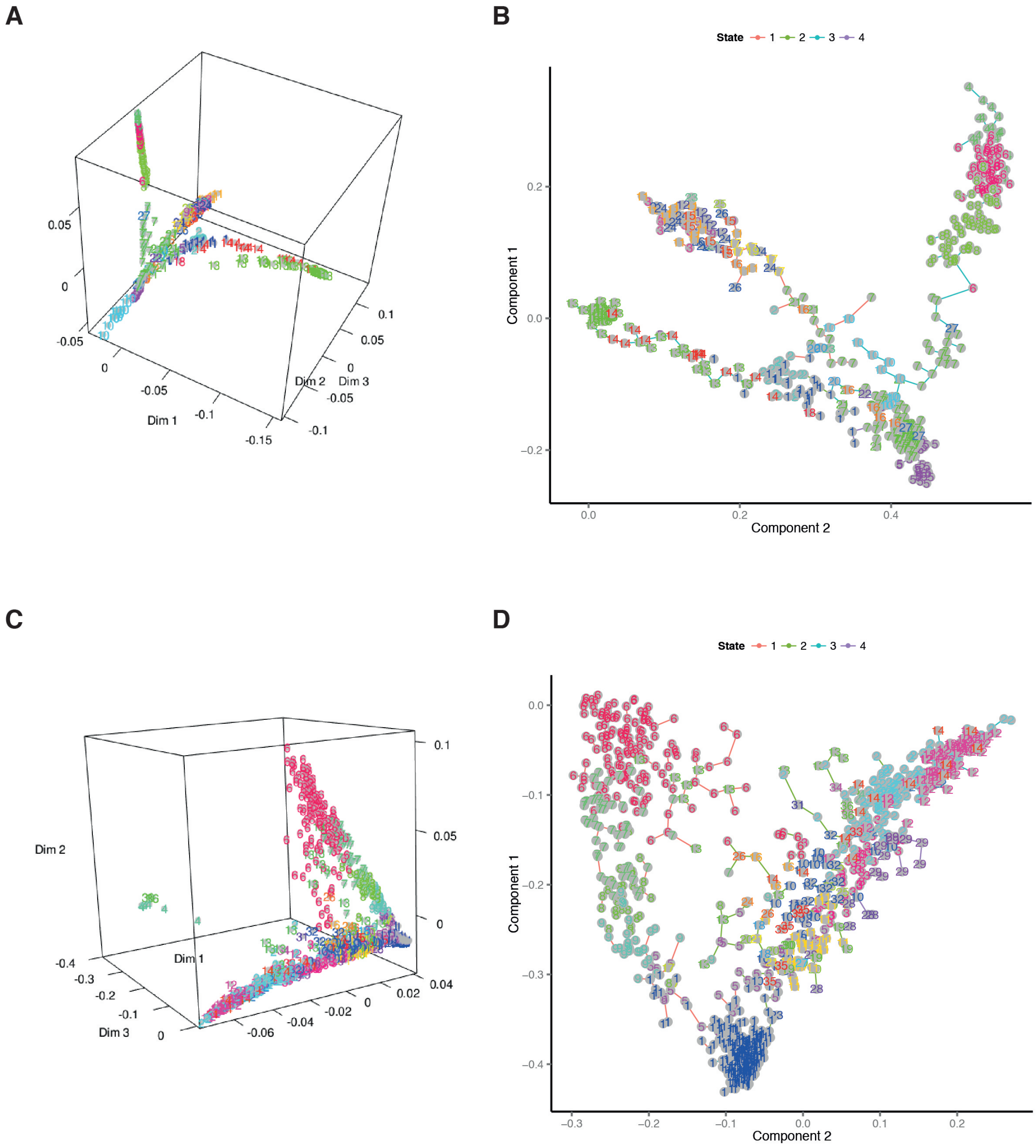

A
0.05
0
-0.05
0
-0.05
Dim 1
-0.1
-0.15
Dim 2
Dim 3
0.1
0.05
0
-0.05
-0.1
B
State
1
2
3
4
Component 1
0.2
0.0
−0.2
0.0
0.2
0.4
Component 2
C
Dim 2
Dim 1
Dim 3
-0.4
-0.3
-0.2
-0.1
0
0.1
0.05
0
0.04
0.02
0
-0.02
-0.04
-0.06
D
State
1
2
3
4
Component 1
0.0
−0.1
−0.2
−0.3
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
Component 2

Figure S6

Figure S7. StemID predicts the stem cell identity for previously published data sets. (Related to Figure 3 and 4)

To test StemID on additional published datasets we searched the literature for single cell profiling of stem cell differentiation systems. We could not find suitable unique molecular identifier (UMI) based data and therefore applied StemID to read based data for the developing lung epithelium (Treutlein et al., 2014) and for developing radial glia cells (Pollen et al., 2015). Although our algorithm was not designed for read based quantification, StemID could infer correct lineage trees and correctly predict the stem cell identity in both systems. (A-D) StemID on 80 cells extracted from mouse lung epithelium at E18.5 (Treutlein et al., 2014). (A) t-SNE map showing the major populations inferred by RaceID2. Clusters are highlighted with different numbers and colors. Alveolar type 1 (AT1) and bipotential progenitors (BP) clustered together (cluster 1). Since our outlier identification is designed for UMI based quantification these subtypes remained unresolved. The other major groups correspond to Clara cells and alveolar type 2 (AT2) cells. (B) Expression of population specific markers (Treutlein et al., 2014) was highlighted in t-SNE maps on a logarithmic ($\log_2$) scale (color legend). (C) Inferred intestinal lineage tree. Only significant links are shown (P<0.05). The color of the link indicates the $-\log_{10}$p-value. The color of the vertices indicates the entropy. Cells are shown in the background as grey dots. A black circle indicates a significant projection component. From these cells an additional link between cluster 1 and clusters 3 and 4 can be recognized, which is marginally significant (P~0.06). (D) Barplot of StemID scores. The BP/AT1 cluster acquires the highest StemID score. With the additional marginal link the difference between cluster 1 and the other clusters would be even larger. (E-H) StemID on 393 cells from the ventricular and subventricular zone of the human cortex at gestational week 16-18 (Pollen et al., 2015). (E) t-SNE map showing the major populations inferred by RaceID2. Clusters are highlighted with different numbers and colors. Clusters 1,2 and 3 represent radial glia cells while 4,6,7,8 represent intermediate progenitors and mature neurons. (F) t-SNE map highlighting expression of radial glia markers (PAX6, PTPRZ1) and an early neuronal marker (NEUROD1) on a logarithmic ($\log_2$) scale (color legend). Up-regulation of PTPRZ1 identifies cluster 3 as outer and cluster 1 and 2 as ventricular radial glia (RG) cells. (C) Inferred cortical lineage tree. Only significant links are shown (P<0.05). The color of the link indicates the $-\log_{10}$p-value. The color of the vertices indicates the entropy. The thickness indicates the link score reflecting how densely a link is covered with cells (see Experimental procedure). The tree links the RG sub-types to the mature neurons (cluster 4 and 8) via a NEUROD1 expressing progenitor population (D)

Barplot of StemID scores. The highest score was correctly assigned to outer RG cells, which have been shown to express self-renewal pathways (as opposed to ventricular RG cells) and differentiate into various neural and glial cell types.

For (B-D) and (F-H) only clusters with >5 cells were analyzed.

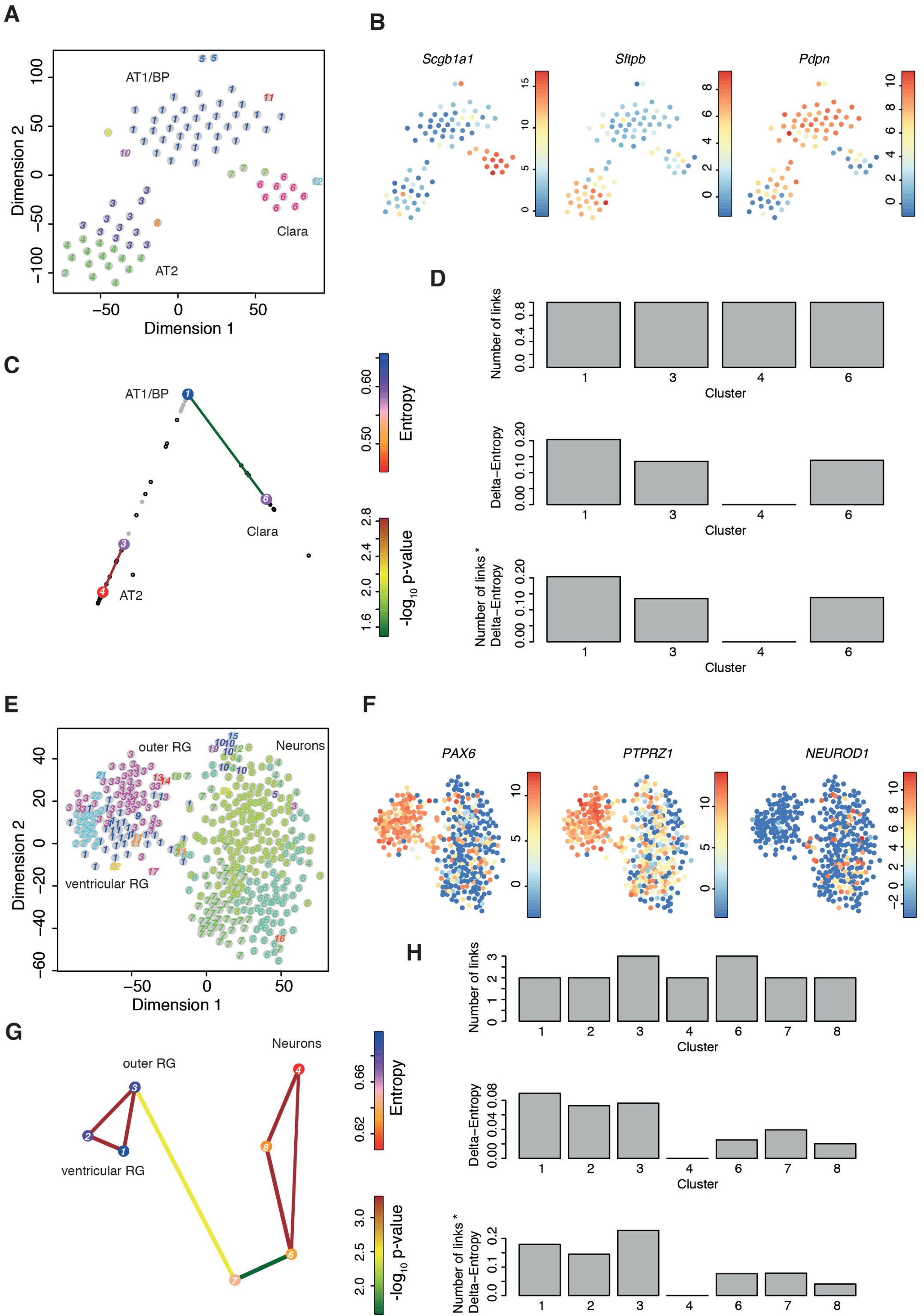

A
B
C
D
E
F
G
H
AT1/BP
Clara
AT2
Dimension 1
Dimension 2
Scgb1a1
Sftpb
Pdpn
Entropy
-log$_{10}$ p-value
Number of links
Delta-Entropy
Number of links * Delta-Entropy
Cluster
outer RG
Neurons
ventricular RG
PAX6
PTPRZ1
NEUROD1

Figure S7

SUPPLEMENTAL TABLE LEGENDS

Table S1. Differentially regulated genes within cell clusters derived for the 5 days *Lgr5* lineage tracing data. (Related to Figure 1)

For each cluster, the first column contains the gene identifier, composed of the official gene symbol and the chromosome separated by a double underscore. The second and third columns contain the average expression across all cells not in the cluster and across cells within the cluster, respectively, normalized to the median expression within the cluster. The third column indicates the fold change and the last column shows the p-value for the observed fold change (see Experimental Procedures).

Table S2. Differentially regulated genes within cell clusters derived for the complex intestinal data. (Related to Figure 3)

For each cluster, the first column contains the gene identifier, composed of the official gene symbol and the chromosome separated by a double underscore. The second and third columns contain the average expression across all cells not in the cluster and across cells within the cluster, respectively, normalized to the median expression within the cluster. The third column indicates the fold change and the last column shows the p-value for the observed fold change (see Experimental Procedures).

Table S3. Differentially regulated genes within cell clusters derived for the bone marrow data. (Related to Figure 4)

For each cluster, the first column contains the gene identifier, composed of the official gene symbol and the chromosome separated by a double underscore. The second and third columns contain the average expression across all cells not in the cluster and across cells within the cluster, respectively, normalized to the median expression within the cluster. The third column indicates the fold change and the last column shows the p-value for the observed fold change (see Experimental Procedures).

Table S4. Differentially regulated genes within cell clusters derived for the pancreatic data. (Related to Figure 6)

For each cluster, the first column contains the gene identifier, composed of the official gene symbol and the chromosome separated by a double underscore. The second and third columns contain the average expression across all cells not in the cluster and across cells within the cluster, respectively, normalized to the median expression within the cluster. The third column indicates the fold change and the last column shows the p-value for the observed fold change (see Experimental Procedures).

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Isolation of crypts from mouse small intestine

Crypts were isolated from mice as described previously (Sato et al., 2009). Briefly, the whole of the small intestine was dissected, flushed with cold Ca^{++} and Mg^{++}-free PBS and cut to 4 – 5 cm pieces for convenience. Intestines were cut open longitudinally and villi were scraped off with a glass slide. Intestine fragments were washed twice with cold Ca^{++} and Mg^{++}-free PBS, then incubated with 5 mM EDTA in PBS at 4°C for 30 minutes, with gentle agitation. Crypts were released by vigorous shaking of the tissue fragments, pelleted by centrifugation (200g at 4°C for 5 minutes), washed once with cold PBS and once with Advanced DMEM/F12 medium (Life Technologies) and pelleted by centrifugation. Crypts were washed once with DMEM and resuspended in DMEM containing 2mg/mL Trypsin (Sigma) and 2000U/mL DNaseI (Sigma) and incubated 30 minutes at RT, pipetting up and down the crypts every 5 minutes. Single cells were peletted by centrifuging (400g at 4C for 5 minutes). Single cells were resuspended in DMEM containing 4000U/mL DNaseI and strain through a 40 uM mesh into a FACS tube. Viable cells were gated by negative DAPI staining. CD24 antibody (Life Technologies) was added 1:200.

Human islet isolation, dispersion and sorting

Pancreatic cadaveric tissue was procured from a multiorgan donor program and only used if the pancreas could not be used for clinical pancreas or islet transplantation, according to national laws, and if research consent was present. Human islet isolations were performed in the islet isolation facility of the Leiden University Medical Center according to a modified protocol originally described by Ricordi et al. (Ricordi et al., 1988). Islets were cultured in CMRL 1066 medium (5.5 mM glucose) (Mediatech) supplemented with 10% human serum, 20 μg/ml ciprofloxacin, 50 μg/ml gentamycin, 2 mM L-glutamin, 0.25 μg/ml fungizone, 10 mM HEPES and 1.2 mg/ml nicotinamide for 3-6 days. Islets were maintained in culture at 37°C in a 5% CO_2 humidified atmosphere. Medium was refreshed the day after isolation and every 2-3 days thereafter until cell sorting.

For cell sorting cultured Islets were briefly washed in cold PBS to remove any residual medium. The islet pellet was then suspended in 1 ml of Accutase per 5000 islet equivalents and incubated at 37 degrees with gentle intermittent shaking for 8-10 minutes until the islets were dispersed into single cells. The digestion process was stopped using an excess volume of cold RPMI medium containing 10% FCS. The dispersed tissue was washed briefly with cold PBS followed by filtering through a sieve to get rid of any debris and undigested material. To assess the viability of the cells, Propidium iodide (PI) or DAPI was added to the suspension of cells. The tissue was stored on ice until sorting using a FACSAria II (BD biosciences). Cells were sorted into 96-well skirted qPCR plates (Greiner) in a mix of TRIzol reagent (Ambion) and 1:250.000.000 ERCC spike-in mix (Ambion; 4456740) and immediately frozen to -80˚C.

Preparation of mouse hematopoietic cells

We used C57Bl/6 female or male mice, from 23 to 52 weeks, bred in our facility. Experimental procedures were approved by the Dier Experimenten Commissie (DEC) of the KNAW, and performed according to the guidelines. Bone marrow was isolated from femur and tibia by flushing Hank's Balanced Salt Solution (HBSS, Invitrogen) without calcium or magnesium, supplemented with 1% heat-inactivated Fetal Calf Serum (FCS, Sigma). Bone marrow was then mildly dissociated by a few pipetting up-and-down. Small interacting structures were selected by visual inspection under a dissection stereomicroscope (Leica) and transferred by mouth pipetting to a microscope (Zeiss) equipped with micromanipulators (Narishige). These structures could be doublets, triplets, etc. or slightly bigger units composed of around 10 to 20 cells. In the case of small structures, the cells were manually pulled apart, without enzymatic dissociation, with the help of two pulled needles. For the bigger units, small structures were first sequentially trimmed off the unit, with the help of the dissection needles, and then single-cell dissected as described previously. The single-cells were mouth pipetted directly into eppendorf tubes containing 100 µL of TRIzol (Life technologies), 0,02 µL of 1:50.000 ERCC Spike-in RNA (Ambion), and 0,2 µL of GlycoBlue (Ambion). Tubes were immediately frozen on dry ice. The pipette used for mouth pipetting was always washed in between pipetting with HBSS 1% FCS.

CEL-seq library preparation

The protocol was carried out as described previously (Grün et al., 2015). Briefly, single cells were processed using the previously described CEL-seq technique (Hashimshony et al., 2012), with several modifications. A 4bp random barcode as unique molecular identifier (UMI) was added to the primer in between the cell specific barcode and the poly T stretch. Dried RNA, prepared from single cells by TRIZOL extraction method, was resuspended in primer solution, denatured at 70°C for 2 minutes and quickly chilled, after which the first strand synthesis mix was added.

Libraries were sequenced on an Illumina HighSeq 2500 using 50 bp paired end sequencing.

Quantification of transcript abundance

Paired end reads obtained by CEL-seq were aligned to the transcriptome using bwa (Li and Durbin, 2010) (version 0.6.2-r126) with default parameters. The transcriptome contained all RefSeq gene models based on the mouse genome release mm10 downloaded from the UCSC genome browser (Meyer et al., 2013) and contained 31,109 isoforms derived from 23,480 gene loci. All isoforms of the same gene were merged to a single gene locus. The right mate of each read pair was mapped to the ensemble of all gene loci and to the set of 92 ERCC spike-ins (Baker et al., 2005) in sense direction. Reads mapping to multiple loci were discarded. The left read contains the barcode information: the first eight bases correspond to the cell specific barcode followed by 4 bases representing the unique molecular identifier. The remainder of the left read contains a polyT stretch followed by few (<15) transcript-derived bases. The left read was not used for quantification. For each cell barcode we counted the number of unique molecular identifiers for every transcript and aggregated this number across all transcripts derived from the same gene locus. Based on binomial statistics we converted the number of observed unique molecular identifiers into transcript counts (Grün et al., 2014).

RaceID2

The RaceID2 algorithm incorporates a number of improvements of the previously published RaceID algorithm (Grün et al., 2015). To safeguard against technical artifacts only down-sampling is used for data normalization. For initial clustering the k-medoids algorithm is used instead of k-means, since it leads to more robust clustering results. K-medoids clustering is directly done with the correlation based distance metric $d_{i,j}=1-\rho_{\mathrm{i,j}}$, where $\rho_{\mathrm{i,j}}$ is Pearson's correlation coefficient between the transcript count vectors of cell i and j.

RaceID2 also utilizes a more robust approach to determine the initial number of clusters used as input for k-medoids. The cluster number is inferred based on the saturation of the average within-cluster dispersion. In this approach the number of clusters is the minimal number k_i such that the change of the within-cluster dispersion upon further increase of the cluster number $k_{i+1}= k_i + 1$ is equal, within the estimated error interval, to the average change upon further increase of the cluster number quantified by a linear regression across k_{i+2}, …, $k_{\mathrm{max.}}$. In other words, the cluster number is determined such that adding more clusters only leads to a linear decrease of the within cluster-dispersion.

For better visualization using the t-SNE map, the t-SNE algorithm is initialized with positions in the embedded space as determined by classical multidimensional scaling.

To derive significantly up- or down-regulated genes for each cluster the same strategy as in RaceID is applied, but gene expression is compared between all cells in a cluster and the remaining cells not in this cluster, as opposed to comparing to all cells.

The R-code of RaceID2 with extensive documentation is available for download at https://github.com/dgrun/StemID.

StemID

StemID is an algorithm based on RaceID2 for the inference of differentiation trajectories and the prediction of the stem cell identity. As an initial step, the algorithm embeds the space of transcript counts for each gene, in which every cell can be represented, into a lower dimensional space in order to maintain only the number of dimensions necessary to represent all point-to-point distances. For the Euclidean metric, only n-1 dimensions are necessary to embed n data points from a high dimensional space ($>n$ dimensions) with exactly conserved distances. For a correlation-based metric as used by RaceID2 this is not true. Here, we embed into $k<n-1$ dimensions, with k being the number of positive eigenvalues of the squared double-centered distance matrix. The distance $d_{i,j}$ between cells i and j is defined as $d_{i,j} = 1 - \rho_{i,j}$, where $\rho_{i,j}$ equals Pearson's correlation coefficient of the transcriptome of these cells. The embedding is computed in R using the function cmdscale.

For the derivation of differentiation trajectories the medoid m_i of cluster i is connected to the medoids m_j of all other clusters j ($j = 1, \ldots, i-1, i+1, \ldots, N$) in the embedded space. Subsequently, for each cell k in cluster i the vector $z_{i,k} = y_{i,k} - m_i$ connecting its position $y_{i,k}$ to m_i is projected onto each link $l_{i,j} = m_j - m_i$ between cluster i and j ($j = 1, \ldots, i-1, i+1, \ldots, N$, i. e. the component of this vector (anti-)parallel to each connection is calculated. Projections $p_{k,i,j}$ correspond to the cosine of the angle $\alpha_{k,i,j}$ between $z_{i,k}$ and $l_{i,j}$ times the length of $l_{i,j}$ and are computed based on the dot product of the two vectors:

$$p_{k,i,j} = \left|z_{i,k}\right| \cdot \cos\alpha_{k,i,j} = \frac{z_{i,k} \cdot l_{i,j}}{\left|l_{i,j}\right|}$$

If the vector component is anti-parallel to a link it will be negative. The respective cell is then assigned to the connection with the longest projection using the coordinate computed from the projection. This procedure is repeated for every cell in each cluster. To determine connections with significantly more assigned cells than expected by chance, the computation is repeated after randomizing the cell positions in the embedded space. Randomization is performed by sampling new cell positions from a uniform interval with boundaries given by the real data for each embedded dimension. Cluster centers are kept constant to maintain the topology of the configuration.

Outgoing and incoming links are distinguished for the p-value calculation, i. e. for each cluster it is computed how many of its cells are assigned to each link to another cluster. The distribution of expected cells on each outgoing link is sampled by repeating the randomization procedure 2,000 times. A p-value for each link is derived as the quantile of this distribution corresponding to the actual number of cells on the link. In general, a cluster can have an enriched outgoing link, which is at the same time a depleted incoming link. We consider a link significantly enriched if this is true for either the outgoing or the incoming link.

To compute a p-value, the sampling is repeated sufficiently often. For instance, if a p-value threshold of *P*<0.01 is chosen to assign significance to a link, the randomization is repeated 2,000 times to calculate the 1%-quantile with sufficient confidence. For lower p-values the number of randomizations needs to be increased.
The ensemble of significant connections can be interpreted as a predicted lineage tree comprising all commonly used differentiation trajectories of a system. The projection of a cell onto a trajectory reflects its differentiation progress measured by pseudo-time and can be used to infer pseudo-temporal ordering of cells on a trajectory defined by a connected set of links.
To assess the confidence of a particular link, a link score is computed that reflects its coverage by cells. This score is defined by the maximum difference between two neighboring cell positions after rescaling the link length to one. Values close to zero reflect coverage only near the connected cluster centers, while values close to one indicate uniform link coverage.

To predict the stem cell identity the algorithm also takes into account the transcriptome entropy of each cell. The entropy E_j of cell j is computed as

$$E_j = \sum_{i=1}^{N} p_{i,j} \log_N p_{i,j} ,$$

where $p_{i,j} = n_{i,j}/N$ and $n_{i,j}$ equals the number of transcripts of gene i in cell j. N equals the total number of transcripts in each cell, which is the same for all cells due to the downsampling (or median-normalization) performed by RaceID2. Next, the median delta-entropy ΔE_k is computed for each cluster k, defines as

$$\Delta E_k \equiv \text{median}_{j\in k}\left(E_j\right) - \min_l\left(\text{median}_{j\in l}\left(E_j\right)\right).$$

To predict the stem cell identity, StemID computes a score for each cluster k given by

$$s_k = l_k \cdot \Delta E_k\,,$$

where l_k denotes the number of significant links of cluster k.

The R-code of RaceID2 and StemID with extensive documentation is available for download at https://github.com/dgrun/StemID.

Datasets and parameter settings

We used previously published mouse intestinal *Lgr5*+ 5 days lineage tracing data (Grün et al., 2015) for the data presented in Figure 1 and 2. Before filtering, this dataset comprises 432 cells with a median number of 5,469 sequenced transcripts per cell. RaceID2 analysis was performed with parameters mintotal=3000, maxexpr=500 and default parameters otherwise. StemID was run with cthr=2, i. e. only clusters with >2 cells are included in the lineage analysis. Very small clusters are considered uninformative for this analysis. The dataset presented in Figure 3 comprises randomly isolated mouse intestinal *Cd24*+ cells (enrichment for secretory cells) and a set of *Cd24*+ cells from an *Lgr5*+ 8-weeks lineage tracing experiment. Additionally, cells from an *Lgr5*+ 3-weeks lineage tracing experiment were included, a subset of which were also *Cd24*+ (see Figure S3B). In total, a median number of 6,208 transcripts were sequenced in 672 cells for this dataset. RaceID2 was run with the same parameters as the first dataset and cln=5, an adjusted cluster number suggested by our saturation criterion. StemID was run with cthr=5, since substantially more cells were available compared to the first dataset. The hematopoietic data presented in Figure 4 comprise mouse bone marrow cells and sorted Kit$^+$ Sca-1$^+$ Lin$^-$ CD48$^-$ CD150$^+$ HSCs (Figure S4B). Prior to filtering this dataset was composed of 2,104 cells with a median number of 938 transcripts per cell. One library was removed from the original data, since cells clustered separately from the remaining cells. Moreover, we noticed the presence of cells with high expression of *Kcnq1ot1*, which we had observed as a non-cell type specific marker of subsets of cells in all datasets analyze. We hypothesize that these have either been exposed to stress during isolation affecting the transcriptome and therefore discarded cells with 10 or more *Kcnq1ot1* transcripts. We then removed *Kcnq1ot1* and the *Rn45s* preribosomal RNA from our pool of reference transcripts which both confounded the cell type identification. The hematopoietic transcriptome data required more pruning, because the overall sensitivity was substantially lower than for the intestinal datasets. To account for the reduced sensitivity, RaceID2 was run with parameters mintotal=900, minexpr=3, maxexpr=500 and default parameters otherwise. StemID was run with cthr=5.

The human pancreatic dataset comprises material from five donors (D3, D7, D10, D17), obtained with or without specific enrichment of cell types (see Figure S5B). In total, 1,728 cells were sequenced with a median number of 4,885 transcripts per cell. RaceID2 was run with parameters mintotal=2000, minexpr=4, probthr=10^{-5} and default parameters otherwise. We adjusted the probability threshold, since heterogeneity was increased due to cells included from different patients. RaceID2 clusters marked by up-regulation of *Kcnq1ot1* were removed before subsequent analysis.

Inference of co-expressed gene modules

To identify modules of co-expressed genes along a specific differentiation trajectory (defined as a succession of links) all cells assigned to these links assembled in pseudo-temporal order based on their projection coordinate. Next, all genes that are not present with at least three transcripts in at least a single cell are discarded from the sub-sequent analysis. Next, a running mean is computed along the differentiation trajectory with a window-size of 25 cells. The pseudo-temporal gene expression profiles of all genes are sub-sequently z-score transformed and topologically ordered by computing a one-dimensional self-organizing map (SOM) with 1,000 nodes. Due to the large number of nodes relative to the number of clustered profiles, similar profiles are assigned to the same node. Only nodes with more than 5 assigned profiles are retained for visualization of co-expressed gene modules.

SUPPLEMENTAL REFERENCES

Baker, S.C., Bauer, S.R., Beyer, R.P., Brenton, J.D., Bromley, B., Burrill, J., Causton, H., Conley, M.P., Elespuru, R., Fero, M., et al. (2005). The External RNA Controls Consortium: a progress report. Nat. Methods *2*, 731–734.

Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nat. Methods *11*, 637–640.

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature *525*, 251–255.

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

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. *2*, 666–673.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., et al. (2013). The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. *41*, D64–D69.

Pollen, A.A., Nowakowski, T.J., Chen, J., Retallack, H., Sandoval-Espinosa, C., Nicholas, C.R., Shuga, J., Liu, S.J., Oldham, M.C., Diaz, A., et al. (2015). Molecular Identity of Human Outer Radial Glia during Cortical Development. Cell *163*, 55–67.

Ricordi, C., Lacy, P.E., Finke, E.H., Olack, B.J., and Scharp, D.W. (1988). Automated method for isolation of human pancreatic islets. Diabetes *37*, 413–420.

Sato, T., Vries, R.G., Snippert, H.J., van de Wetering, M., Barker, N., Stange, D.E., van Es, J.H., Abo, A., Kujala, P., Peters, P.J., et al. (2009). Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. Nature *459*, 262–265.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B (Statistical Methodol. *63*, 411–423.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*, 381–386.

Treutlein, B., Brownfield, D.G., Wu, A.R., Neff, N.F., Mantalas, G.L., Espinoza, F.H., Desai, T.J., Krasnow, M.A., and Quake, S.R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature *509*, 371–375.