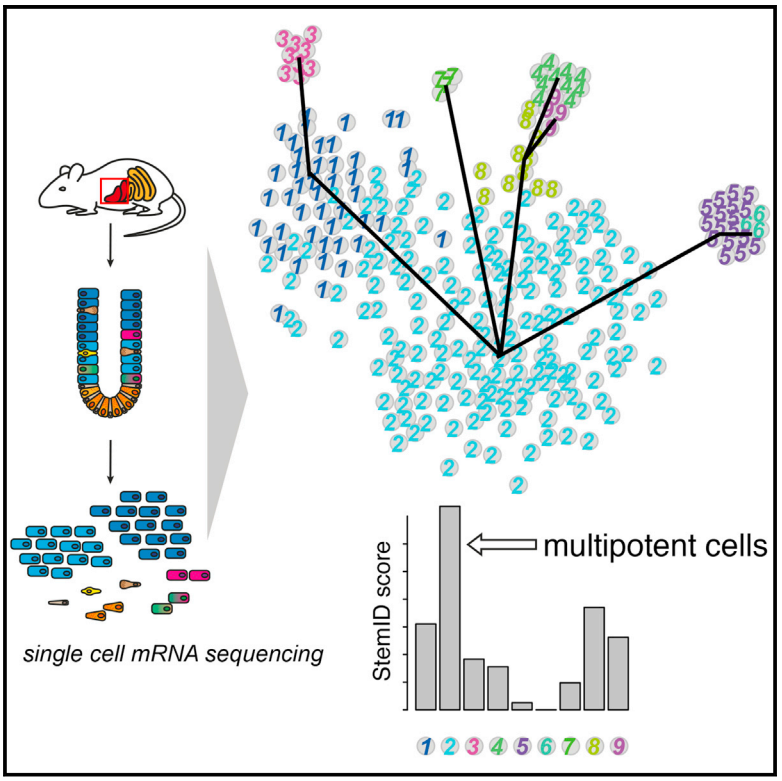
Resource

Cell Stem Cell

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

Graphical Abstract



Authors

Dominic Grün, Mauro J. Muraro, Jean-Charles Boisset, ..., Hans Clevers, Eelco J.P. de Koning, Alexander van Oudenaarden

Correspondence

gruen@ie-freibug.mpg.de (D.G.), a.vanoudenaarden@hubrecht.eu (A.v.O.)

In Brief

Grün et al. developed an algorithm, StemID, for the derivation of cell lineage trees and identification of stem cells from single-cell mRNA sequencing data. StemID successfully recovered known adult stem cell populations from the small intestine and bone marrow and was then used to predict a novel multipotent cell population in the human pancreas.

Highlights

- StemID infers the lineage tree and identifies stem cells from single-cell mRNA-seq data
- Direct links of stem cells to distinct sub-types reflect transcriptome plasticity
- The permissive stem cell transcriptome is characterized by high entropy
- StemID infers candidate multipotent cell populations in the human pancreas

Accession Numbers

GSE76408
GSE76983
GSE81076

Grün et al., 2016, Cell Stem Cell *19*, 266–277
August 4, 2016
http://dx.doi.org/10.1016/j.stem.2016.05.010

CellPress

Cell Stem Cell
Resource

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

Dominic Grün,[1,2,3,*] Mauro J. Muraro,[1,2] Jean-Charles Boisset,[1,2] Kay Wiebrands,[1,2] Anna Lyubimova,[1,2] Gitanjali Dharmadhikari,[1,2,4] Maaike van den Born,[1,2] Johan van Es,[1,2] Erik Jansen,[1,2] Hans Clevers,[1,2,5] Eelco J.P. de Koning,[1,2,4] and Alexander van Oudenaarden[1,2,*]

[1]Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences, 3584 CT Utrecht, the Netherlands
[2]Cancer Genomics Netherlands, University Medical Center Utrecht, 3584 CX Utrecht, the Netherlands
[3]Max Planck Institute of Immunobiology and Epigenetics, 79108 Freiburg, Germany
[4]Department of Medicine, Section of Nephrology and Section of Endocrinology, Leiden University Medical Center, 2333 ZA Leiden, the Netherlands
[5]Princess Maxima Center for Pediatric Oncology, 3508 AB Utrecht, the Netherlands
*Correspondence: gruen@ie-freibug.mpg.de (D.G.), a.vanoudenaarden@hubrecht.eu (A.v.O.)
http://dx.doi.org/10.1016/j.stem.2016.05.010

SUMMARY

Adult mitotic tissues like the intestine, skin, and blood undergo constant turnover throughout the life of an organism. Knowing the identity of the stem cell is crucial to understanding tissue homeostasis and its aberrations upon disease. Here we present a computational method for the derivation of a lineage tree from single-cell transcriptome data. By exploiting the tree topology and the transcriptome composition, we establish StemID, an algorithm for identifying stem cells among all detectable cell types within a population. We demonstrate that StemID recovers two known adult stem cell populations, *Lgr5*+ cells in the small intestine and hematopoietic stem cells in the bone marrow. We apply StemID to predict candidate multipotent cell populations in the human pancreas, a tissue with largely uncharacterized turnover dynamics. We hope that StemID will accelerate the search for novel stem cells by providing concrete markers for biological follow-up and validation.

INTRODUCTION

The identification of a stem cell in a tissue is a major challenge of pivotal importance. Being able to detect the stem cell population allows for powerful approaches to study cell differentiation dynamics by, for example, lineage tracing (Barker et al., 2007; Busch et al., 2015). Additionally, it provides a first step toward ex vivo propagation of primary stem cells in organoid cultures (Lancaster et al., 2013; Sato et al., 2009), important for applications in regenerative medicine. Moreover, stem cell populations relevant for disease progression, such as cancer stem cells, are promising targets for therapeutic intervention. Stem cells are typically rare, which makes their discovery by traditional population-based assays very difficult. For example, it took decades of dedicated research to define the population of hematopoietic stem cells (HSCs) (Eaves, 2015), but it remains an open question how much heterogeneity exists within this subpopulation of bone marrow cells (Wilson et al., 2015). Similarly, the discovery of intestinal stem cells (van der Flier and Clevers, 2009) took years of work, and heterogeneity within this compartment remains under debate (Buczacki et al., 2013).

The recent availability of single-cell mRNA sequencing methods allows profiling of healthy and diseased tissues with single-cell resolution (Grün et al., 2015; Jaitin et al., 2014; Macosko et al., 2015; Patel et al., 2014; Paul et al., 2015; Treutlein et al., 2014; Zeisel et al., 2015). The transcriptome of a cell can be interpreted as a fingerprint, revealing its identity. However, biological gene expression noise (Eldar and Elowitz, 2010; Raj and van Oudenaarden, 2008) and technical noise because of amplification of minute amounts of mRNA from a single cell (Brennecke et al., 2013; Grün et al., 2014) affects the readout and makes it a challenge to discriminate cell types based on their transcriptome. By sequencing large numbers of randomly sampled single cells from a tissue, it is now possible to compile a nearly complete inventory of cell types.

These inventories can be screened for cell types of particular interest, such as stem cells. An obvious strategy for the identification of the stem cell is the derivation of a lineage tree from single-cell sequencing data. However, transcriptomes of randomly sampled cells only represent a snapshot of the system, and temporal differentiation dynamics cannot be directly derived. However, if the system of interest comprises all differentiation stages, such as the intestinal epithelium or the bone marrow, then attempts can be made to infer a lineage tree by assembling single-cell transcriptomes in a pseudo-temporal order. Existing approaches assume a continuous temporal change of transcript levels to assemble differentiation trajectories (Bendall et al., 2014; Haghverdi et al., 2015; Trapnell et al., 2014), but resolving the correct tree topology remains a challenge.

Here we present a method to identify rare and abundant cell types of a system and use these cell type classifications to guide the inference of a lineage tree. We investigate the general properties characterizing the position of a cell type within the lineage tree and identify the number of branches and the transcriptome uniformity of a cell type as features correlating with the degree of pluripotency. We show that our approach successfully recovers the identity of the stem cell in the intestine and in the bone

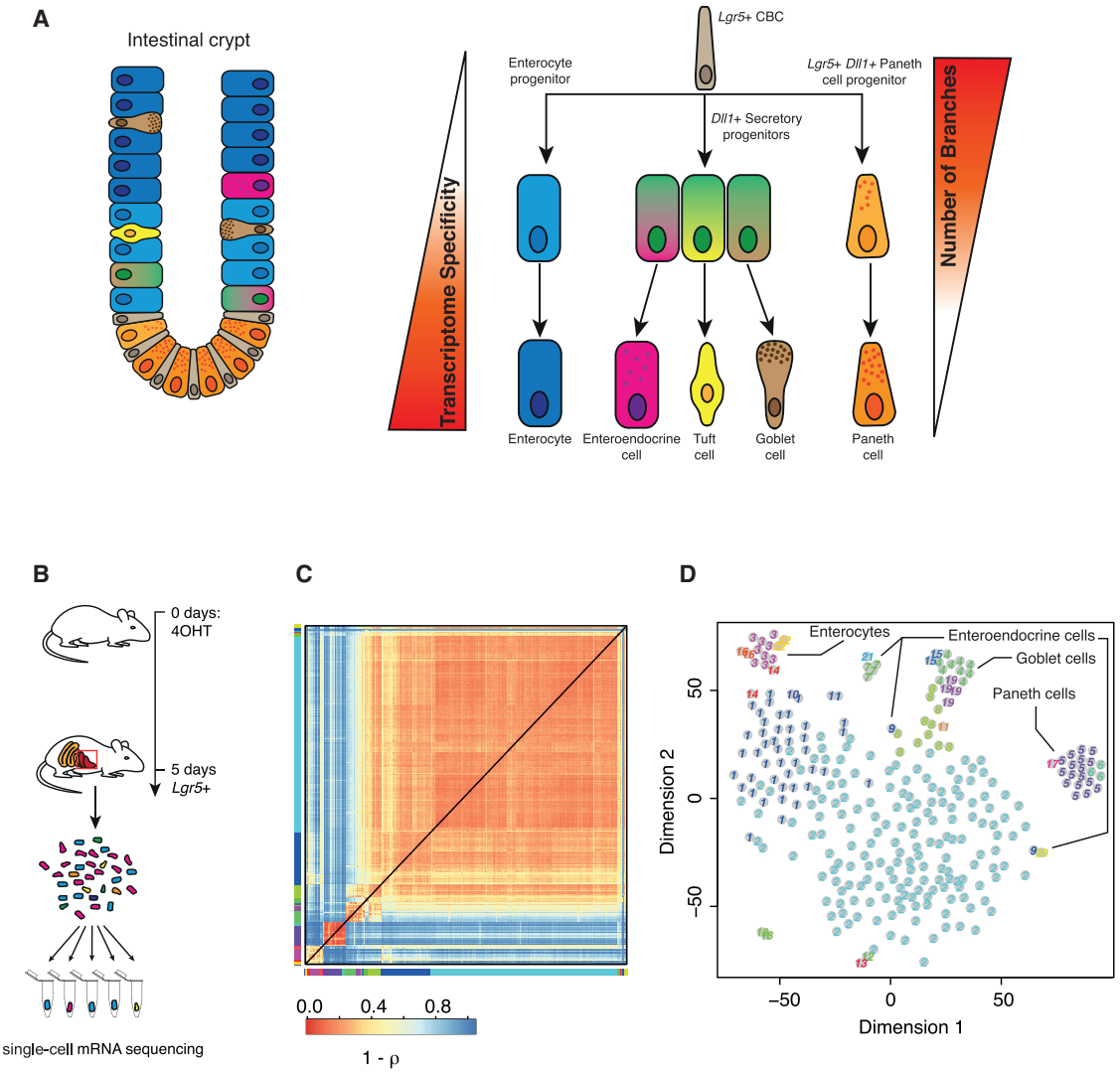


Figure 1. RaceID2 Recovers Intestinal Cell Types

(A) The intestinal epithelium is a well characterized differentiation system. Lgr5-positive stem cells give rise to secretory and absorptive precursors by WNT and NOTCH signaling that further differentiate into mature intestinal cell types.

(B) Summary of the lineage-tracing experiment performed to sequence single 5-day-old progeny of *Lgr5*-positive cells.

(C) Heatmap of cell-to-cell transcriptome distances measured by 1 – Pearson's correlation coefficient (ρ). RaceID2 clusters are color-coded along the boundaries.

(D) t-distributed stochastic neighbor embedding (t-SNE) map representation of transcriptome similarities between individual cells. The clusters identified in (C) are highlighted with different numbers and colors, and the corresponding intestinal cell types identified based on known marker genes are indicated.

See also Figure S1.

marrow, two systems with a well described stem cell population. We then use our method to predict multipotent cell populations in the adult human pancreas.

RESULTS

Robust Identification of Mouse Intestinal Cell Types by RaceID2

To develop a robust approach for the inference of differentiation trajectories, we used a previously published dataset from a lineage tracing experiment comprising the progeny of *Lgr5*-positive mouse intestinal stem cells (Grün et al., 2015). This system is ideal for testing the inference of differentiation dynamics because the lineage tree is already well characterized (Figure 1A). The continuously self-renewing intestinal epithelium is arranged in crypts and villi, with a small number of *Lgr5*+ stem cells, also known as crypt base columnar cells (CBCs), residing near the crypt bottom. These CBCs give rise to rapidly proliferating transit-amplifying (TA) cells that migrate upward along the crypt-villus axis and develop into the terminally differentiated cell types (Barker, 2014; van der Flier and Clevers, 2009). Although absorptive enterocytes constitute the most abundant

cell type, the secretory lineage comprises rare cells, such as mucus-producing goblet cells, hormone-secreting enteroendocrine cells, and antimicrobial Paneth cells. Labeled cells were collected 5 days after label induction using an *Lgr5*-CreERT2 construct and a Rosa26-YFP reporter with a loxP-flanked transcriptional roadblock (Figure 1B).

We first improved the robustness of the initial clustering step of our previously developed RaceID algorithm (Grün et al., 2015) by replacing the k-means clustering with k-medoids clustering (Figure S1). Second, we noticed that the previously used gap statistic (Tibshirani et al., 2001) was not ideal for determining the cluster number. Although increasing the number of clusters in many cases leads to a growing gap statistic, the decrease of the within-cluster dispersion (Tibshirani et al., 2001) saturates quickly. A further increase of the cluster number, therefore, reduces cluster reproducibility. In RaceID2, we thus determine the cluster number by identifying the saturation point of the within-cluster dispersion. Together, these two changes lead to a more robust initial clustering of RaceID2 (Experimental Procedures; Figure S1).

For the intestinal lineage tracing data (Experimental Procedures), RaceID2 recovered a larger group of *Lgr5*+ stem cells (cluster 2) and early progeny (clusters 1 and 8) as well as the major mature cell types; i.e., enterocytes (cluster 3), goblet (clusters 4 and 19), Paneth (clusters 5 and 6), and enteroendocrine cells (cluster 7) (Figures 1C and 1D). These cell types could be unambiguously assigned based on the cluster-specific upregulation of marker genes inferred by RaceID2 (Table S1).

Inference of the Lineage Tree with Guided Topology

One of the major challenges for the inference of differentiation pathways in a system with multiple cell lineages is the determination of branching points. To overcome this problem, we predefined the topology of the lineage tree by allowing differentiation trajectories linking each pair of clusters. A putative differentiation trajectory links the medoids of two clusters, and the ensemble of all inter-cluster links defines the possible topology of the lineage tree. To minimize the effect of technical noise and, at the same time, the computational burden, we first reduce the dimensionality of the input space requiring maximal conservation of all point-to-point distances. In a second step, we assign each cell to its most likely position on a single inter-cluster link. To find this position, the vector connecting the medoid of a cluster to one of its cells is projected onto the links between the medoid of this and all remaining clusters, and the cell is assigned to the link with the longest projection after normalizing the length of each link to one. The projection also defines the most likely position of the cell on the link (Figure 2A), reflecting its differentiation state (Experimental Procedures). If this strategy is applied to the intestinal data, then only a subset of links is populated (Figure 2B). To determine links that are more highly populated than expected by chance and are therefore candidates for actual differentiation trajectories, we computed an enrichment p value based on comparison with a background distribution with randomized cell positions (Figure 2B; Figure S2A). Furthermore, we reasoned that the coverage of a link by cells indicates how likely it is that this link represents an actual differentiation trajectory and not only biased perturbations driving the transcriptome of a given cluster preferentially toward the transcriptome of another cluster without leading to actual differentiation events. We defined a link score as one minus the maximum difference between the positions of each pair of neighboring cells on the link after normalizing the length of each link to one (Figure S2B). If this score is close to one, then the link is densely covered with cells with only small gaps in between. If the link score is close to zero, the cell density is only concentrated near the cluster centers connected by this link. A detailed description of the algorithm is given in the Experimental Procedures. The computationally inferred intestinal lineage tree is consistent with the known lineage tree (Figure 1A). Secretory cell types (clusters 4, 5, 6, and 7) populate individual branches emanating from the central *Lgr5*+ cluster, and absorptive enterocytes (cluster 3) differentiate from the same group via a more abundant group of TA cells (cluster 1).

We compared the inferred lineage tree to the tree predicted by Monocle (Trapnell et al., 2014), a recent method for the derivation of branched lineage trees that does not rely on a predefined tree topology, and found that Monocle could not resolve the different branches of secretory cells (Figure S2).

High Connectivity and High Transcriptome Entropy Reveals the Identity of the Stem Cell

Next we attempted to predict the stem cell identity from the lineage tree. Our working definition of a stem cell for this purpose purely relies on multipotency. More precisely, we try to identify, from the lineage tree, the cell population with the highest degree of multipotency. We noticed that different cell types showed a variable number of populated links to other clusters. The link score is reflected by the thickness of the line in our graphical representation (Figure 2B). We also show links with a low link score because they are informative about the associated cell state. For example, a cell type with many low-scoring links can fluctuate toward a diversity of fate biases, whereas cell types with only a few links are much more canalized. These two scenarios reflect a more promiscuous transcriptome, such as expected for stem cells, versus a more confined transcriptome, as expected for a mature cell type. In our data, cluster 2, which contains cells positive for *Lgr5* and other established stem cell markers (*Ascl2* and *Clca4*) (Figure 2C), was the most highly connected cluster. Another putative property of stem cells is the tendency to exhibit a more uniform composition of the transcriptome in comparison with differentiated cells. Mature cell types frequently express a small number of genes at very high levels, crucial for cell type-specific functions. The transcriptome of Paneth cells, for instance, is dominated by high numbers of lysozymes and other host defense genes. The uniformity of the transcriptome is reflected by Shannon's entropy (Shannon, 1948), and this concept has previously been applied to study cellular differentiation (Anavy et al., 2014; Banerji et al., 2013; Piras et al., 2014) (Experimental Procedures). We anticipate that the transcriptome of a multipotent cell type is more uniform in each individual cell. In addition, multiple state biases could coexist within this population that can give rise to diverse mature cell types upon external stimuli, or stochastically, leading to high entropy (Banerji et al., 2013; Ridden et al., 2015). For the intestinal lineage tracing data, both Paneth and goblet cells had clearly reduced entropy compared with *Lgr5*-positive cells, whereas the entropy of enterocytes and enteroendocrine cells was comparable with stem cells (Figure 2D). We found that, for all analyzed datasets (see below), the number of links discriminates better between

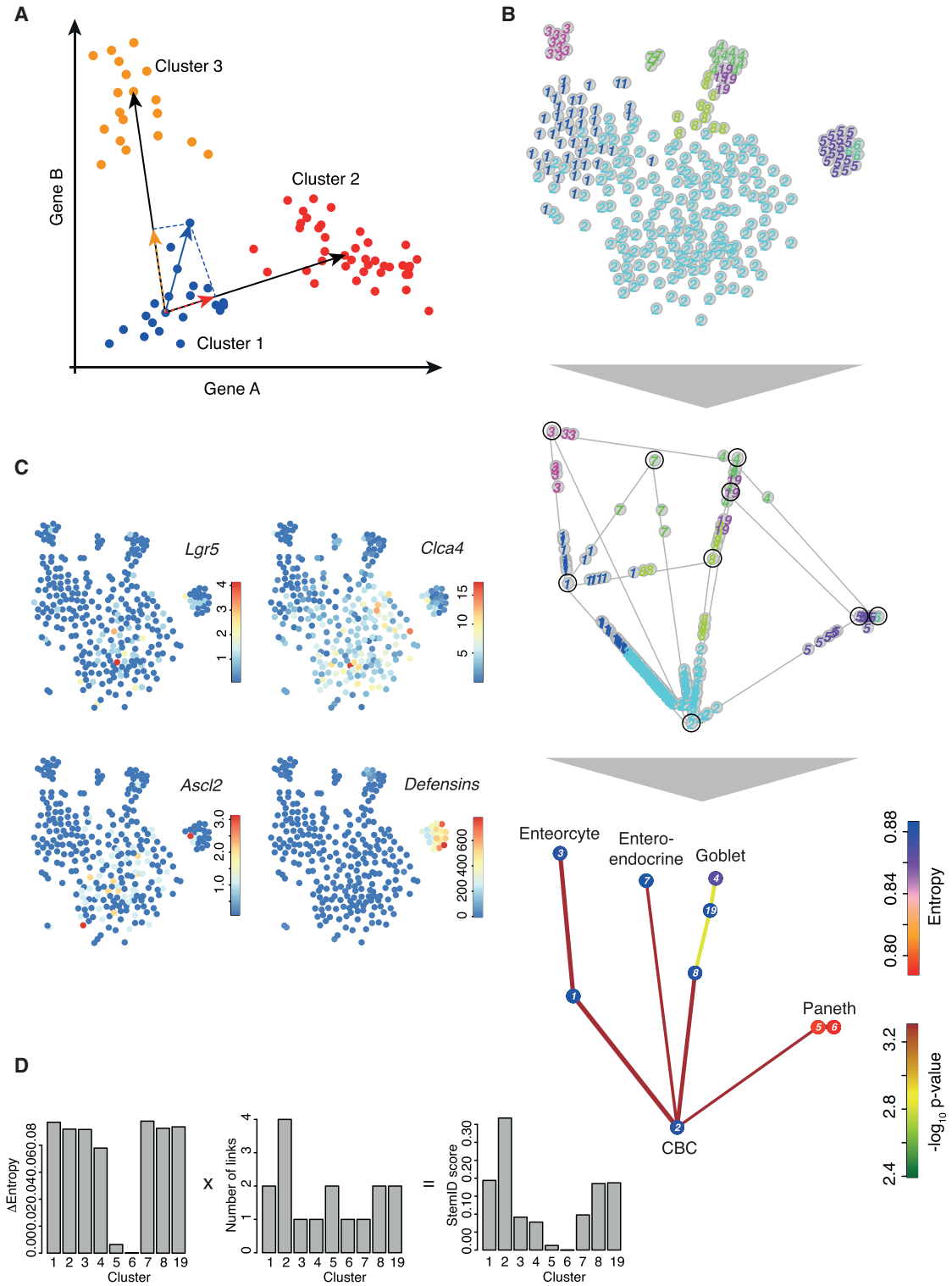


Figure 2. Lineage Tree Inference for Intestinal Stem Cell Progeny

(A) Schematic of the method used to infer differentiation trajectories (see main text and Experimental Procedures).

(B) Outline of the method visualized in the t-SNE-embedded space. All RaceID2 clusters with more than two cells (top) are connected by links, and, for each cell, the link with the maximum projection is determined as shown in (A). Only populated links are shown (center). Cluster centers are circled in black. Significant links are inferred by comparison with the background distribution with randomized cell positions (Experimental Procedures). Only significant links are

(legend continued on next page)

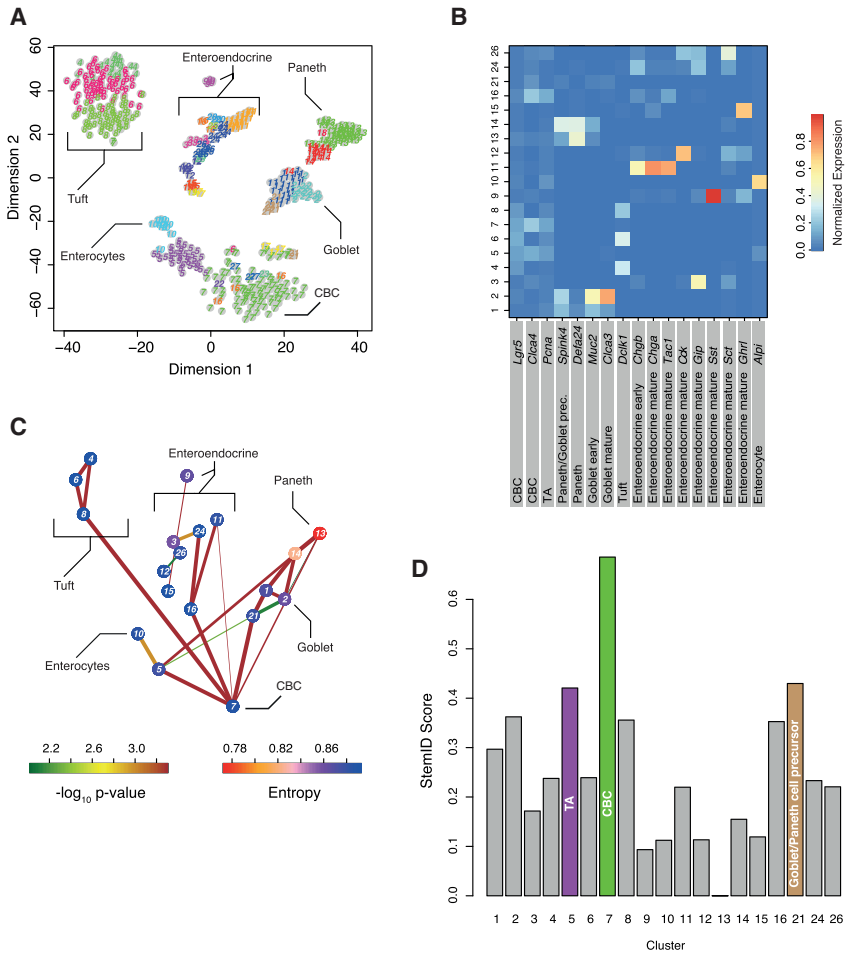


Figure 3. StemID Identifies Stem Cells in Complex with Non-random Mixtures of Intestinal Cells

(A) t-SNE map of transcriptome similarities of intestinal cells from a variety of single-cell mRNA sequencing experiments (main text and Figure S3). RaceID2 clusters are highlighted with different numbers and colors. Cell types identified based on marker gene expression are shown.

(B) Heatmap showing the average expression of known cell type markers across all clusters with more than five cells. For each gene, the sum of expression values over all clusters is normalized to one.

(C) Inferred intestinal lineage tree. Only significant links are shown ($p < 0.01$). 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 (Experimental Procedures).

(D) Barplot of StemID scores for intestinal clusters.

In (B)–(D), only clusters with more than five cells were analyzed. See also Figures S3, S6, and S7.

multipotent and differentiated cells when rescaled by the entropy. Therefore, the simplest score that performs well in discriminating multipotent cells from the remaining cell types was a product of the median entropy (after subtracting the minimal entropy observed in the system) and the number of links of a cluster (Experimental Procedures). This score exhibits a clear maximum for cluster 2 comprising the *Lgr5*+ stem cells (Figure 2D). We named our algorithm StemID for the lineage tree inference and the derivation of this score.

StemID Recovers Intestinal Stem Cells in a Complex Dataset with Non-random Cell Type Frequencies

Next we wanted to test whether StemID could identify *Lgr5*+ cells in a larger and more complex dataset comprising intestinal cells of various independent experiments conducted in our lab. In this dataset, we combined 3 weeks and 8 weeks of *Lgr5* lineage tracing data. A subset of those was enriched in secretory cells by fluorescence-activated cell sorting (FACS) on CD24 (van Es et al., 2012; Figure S3). For both time points, we also sorted non-traced CD24+ control cells (Experimental Procedures; Figure S3). RaceID2 revealed the known intestinal cell types within this dataset based on cluster-specific expression of known cell type marker genes and subdivided these into stages of differentiation or maturation (Figures 3A and 3B; Figure S3A). A full list of differentially expressed genes for each cluster is given in Table S2. For example, intestinal stem cells in cluster 7, marked by high expression of *Lgr5* and *Clca4* (Figure 3B), were connected directly to all secretory branches, whereas TA cells (cluster 5) primarily give rise to enterocytes (cluster 10) (Figure 3C; Figures S3C and S3D). Interestingly, we observed two distinct differentiation trajectories for Paneth cells (clusters 13 and 14), one via a *Dll1*-positive common precursor of Paneth and goblet cells (cluster 1) and another one directly connecting stem cells (cluster 7) or TA cells in cluster 5, marked by upregulation of the cell-cycle gene *Pcna*, directly to the mature Paneth cell clusters. Both the *Dll1*-dependent (van Es et al., 2012) and the direct route (Farin et al., 2014; Sawada et al., 1991), which was observed after ablation of Paneth cells, have been described. The recovery of alternative differentiation pathways demonstrates the power of our guided lineage inference. We were not able to recapitulate

shown ($p < 0.01$). 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 (Experimental Procedures).

(C) Transcript counts (color legend) of the intestinal stem cell markers *Lgr5*, *Clca4*, and *Ascl2* are highlighted in the t-SNE map. Expression of these genes is restricted to cluster 2 and clusters 5 and 6. Clusters 5 and 6 comprise Paneth cells, which were shown to co-express *Lgr5* (Grün et al., 2015). Accumulated transcript counts across all Defensin genes, which are markers of Paneth cells, are shown at the bottom right.

(D) Barplot of StemID scores for all clusters. The median transcriptome entropy of each cell type was computed across all cells in a cluster (left). The lowest entropy across all cell types was subtracted for each cell types because absolute differences were only small. This Δentropy was multiplied by the number of significant links for each cluster (center), yielding the StemID score (right).

See also Figure S2.

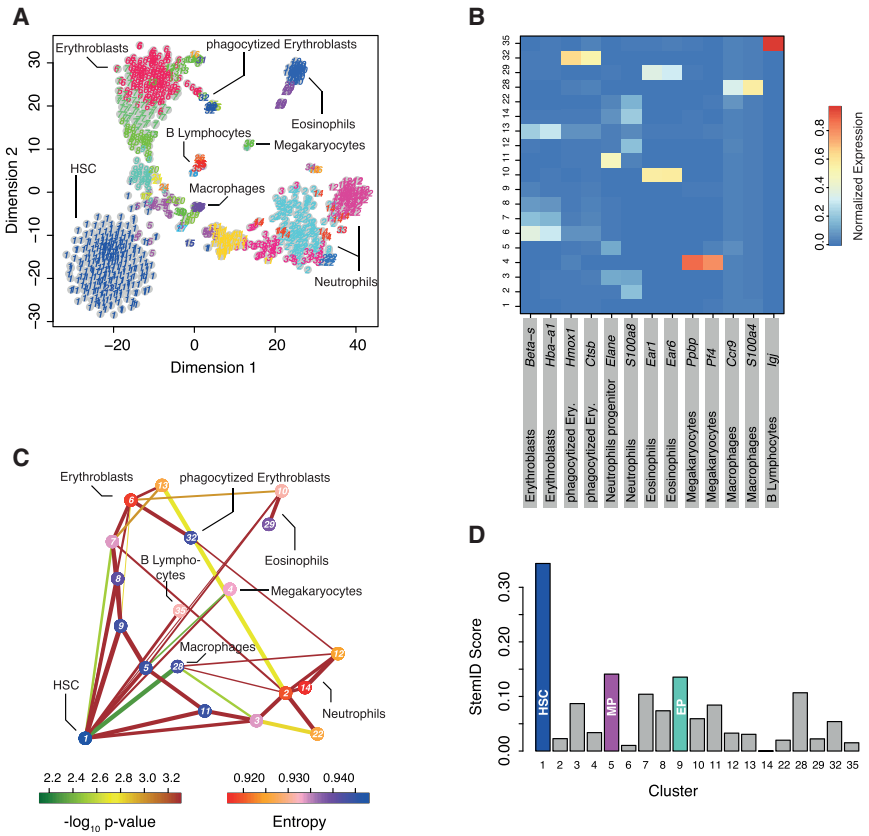


Figure 4. StemID Identifies Hematopoietic Stem Cells in Non-random Mixtures of Bone Marrow Cells

(A) t-SNE map of transcriptome similarities of hematopoietic cells sampled from physically interacting doublets or multiplets (main text and Figure S4). RaceID2 clusters are highlighted with different numbers and colors. Cell types identified based on marker gene expression are shown.

(B) Heatmap showing the average expression of known cell type markers across all clusters with more than five cells. For each gene, the sum of expression values over all clusters is normalized to one.

(C) Inferred hematopoietic lineage tree. Only significant links are shown ($p < 0.01$). 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 (Experimental Procedures).

(D) Barplot of StemID scores for hematopoietic clusters. MP, myeloid progenitor; EP, erythroblast progenitor.

See also Figures S4, S6, and S7.

this finding with a minimum spanning tree-based alternative approach (Figure S3E).

We then computed the StemID score and found that the *Lgr5+/Clca4+* cells (cluster 7) exhibit the highest score (Figure 3D). The second highest score was observed for cluster 21, which represents a common progenitor to Paneth and goblet cells. The TA cells in cluster 5, which our lineage inference identifies as progenitors with an enterocyte fate bias, acquire the third-highest StemID score.

Noticeably, Paneth cells in cluster 13 and mature goblet cells in cluster 2 show the same connectivity as the stem and progenitor cells in clusters 7, 5, and 21, but rescaling by entropy helps correctly assign a mature state to these cells (Figure S3F). In conclusion, StemID could identify intestinal stem cells and distinguish progenitor populations from more mature intestinal cell types.

StemID Recovers Hematopoietic Stem Cells within a Non-random Sample of Bone Marrow Cells

To test the performance of StemID in a different biological system, we applied the algorithm to single-cell sequencing data of mouse bone marrow cells. These cells were selected based on physical interactions between doublets or larger groups of cells and are thus not sampled randomly from all cell types in the bone marrow. This dataset was complemented with $Kit^+Sca\text{-}1^+Lin^-CD48^-CD150^+$ HSCs (Kiel et al., 2005) sorted from the bone marrow (Experimental Procedures; Figure S5B). Cell types identified by RaceID2 were dominated by the myeloid lineage and comprised HSCs, erythroblasts, megakaryocytes, two groups of granulocytes (neutrophils and eosinophils), macrophages, a small group of B lymphocytes, and several clusters representing progenitor stages of the myeloid lineage (Figures 4A and 4B; Figure S6A). A full list of differentially expressed genes for each cluster is shown in Table S3. Cluster 1 comprises almost exclusively sorted HSCs (Figure S4B). The inferred lineage tree (Figure 4C; Figures S6C and S6D) indicates that HSCs differentiate into multipotent progenitor cells (cluster 5) but are also directly linked to mature lineages. HSCs and multipotent progenitors are both linked to megakaryocytes (cluster 4), eosinophils (clusters 10 and 29), macrophages (cluster 28), and two branches covering a spectrum of progenitor and mature states of the neutrophil (clusters 11, 3, 2, 14, 12, and 22) and erythroid lineage (clusters 9, 8, 7, 6, and 13), respectively. The B lymphocytes are only directly linked to the HSCs, suggesting that cluster 5 represents a myeloid progenitor population, and no lymphoid progenitors were present in our sample. The inferred lineage tree is therefore consistent with the existence of a common myeloid progenitor population giving rise to erythrocytes, megakaryocytes, granulocytes, and macrophages (Orkin and Zon, 2008). StemID determines the highest score for cluster 1 and, therefore, correctly recovers HSCs among all cell types in the mixture (Figure 4D; Figure S6). The second-highest score discriminates the multipotent myeloid progenitors (cluster 5) from the remaining cell types, and the third-highest score was assigned to the earliest progenitor of the erythroblast lineage. Therefore, the level of multipotency also correlates with the StemID score of bone marrow-derived cells.

The high connectivity of cluster 1 provides evidence for early fate biases already in HSCs. Moreover, the high entropy of HSCs reflects a more uniform transcriptome in individual cells of this population. The entropy distribution across all cells in this cluster is shifted in comparison with all other groups

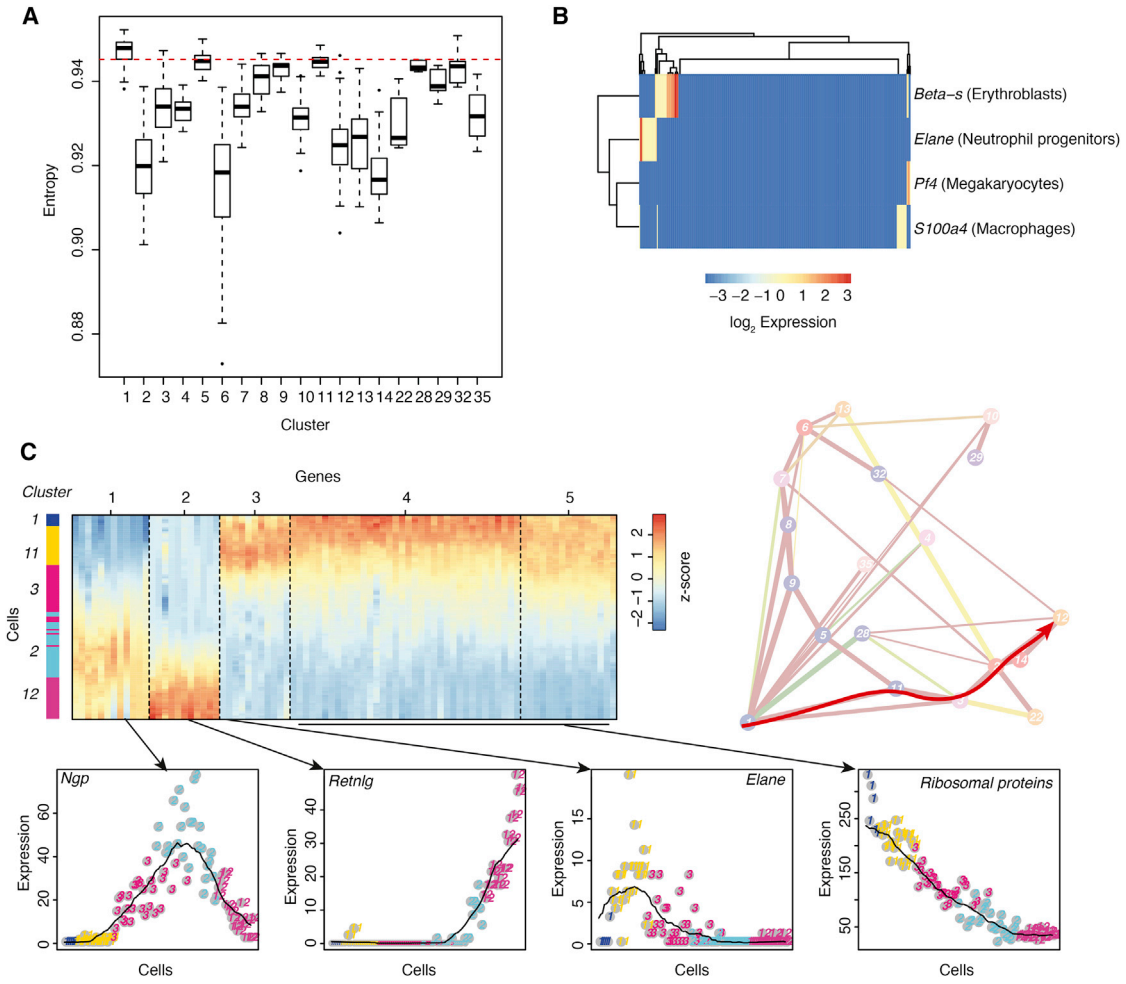


Figure 5. The Multipotency of HSCs Is Reflected by High Transcriptome Entropy

(A) Boxplot of the transcriptome entropy for all RaceID2-derived bone marrow cell types with more than five cells. The boundaries of the box represent the 25% and 75% quantiles, the thick line corresponds to the median, and whiskers extend to the 5% and 95% quantiles. The broken red line indicated the 25% quantile for HSCs (cluster 1).

(B) Two-dimensional clustering of lineage markers in all HSCs (cluster 1). The heatmap shows logarithmic expression.

(C) Self-organizing map (SOM) of *Z*-score-transformed, pseudo-temporal expression profiles along the neutrophil differentiation trajectory (clusters 1, 11, 3, 2, and 12), indicated by the red arrow superimposed on the lineage tree (Experimental Procedures). The pseudo-temporal order was inferred from the projection coordinates of all cells. The color-coding on the left indicates the cluster of origin. The SOM identified five different modules of co-regulated genes. Examples are shown at the bottom. The clusters of origin are indicated as colors and numbers. The black line represents a moving average (window size 25).

In (A)–(C), only clusters with more than five cells were analyzed.

(Figure 5A). In general, the inter-cluster variability substantially exceeds the intra-cluster variability. The narrow entropy distribution of cluster 1 also rules out a strong dependence on the cell cycle. However, we also observed that 54 of the 276 HSCs (20%) show distinct fate biases, revealed by low expression of lineage-specific marker genes (Figure 5B), a finding that is consistent with a recent report based on lineage tracing (Perié et al., 2015). Because the sensitivity of single-cell sequencing is limited, this number is almost certainly an underestimation. We note that most HSCs (112 of 276) are assigned to the link with the multipotent progenitor (cluster 5). We cannot address whether the observed fate bias persists during differentiation or whether stochastic switching between distinct cell fates occurs during differentiation. Our observation is also consistent with a recent single-cell transcriptome analysis showing an unexpected heterogeneity of myeloid progenitor cell populations and suggests the existence of an early cell fate bias (Paul et al., 2015). We observe very similar sets of marker genes, as found in this study, but our lineage inference permits an analysis of the temporal dynamics of gene expression. As an example, we extracted all cells from the neutrophil branch (clusters 1, 11, 3, 2, and 12) in pseudo-temporal order derived from the projection coordinates and clustered temporal expression profiles by using self-organizing maps (Experimental Procedures). A *Z*-score of gene expression values along this trajectory reveals that the RaceID2 clusters represent sets of cells with common modules of co-expressed genes and that gene expression within these modules changes smoothly over time (Figure 5C). Although

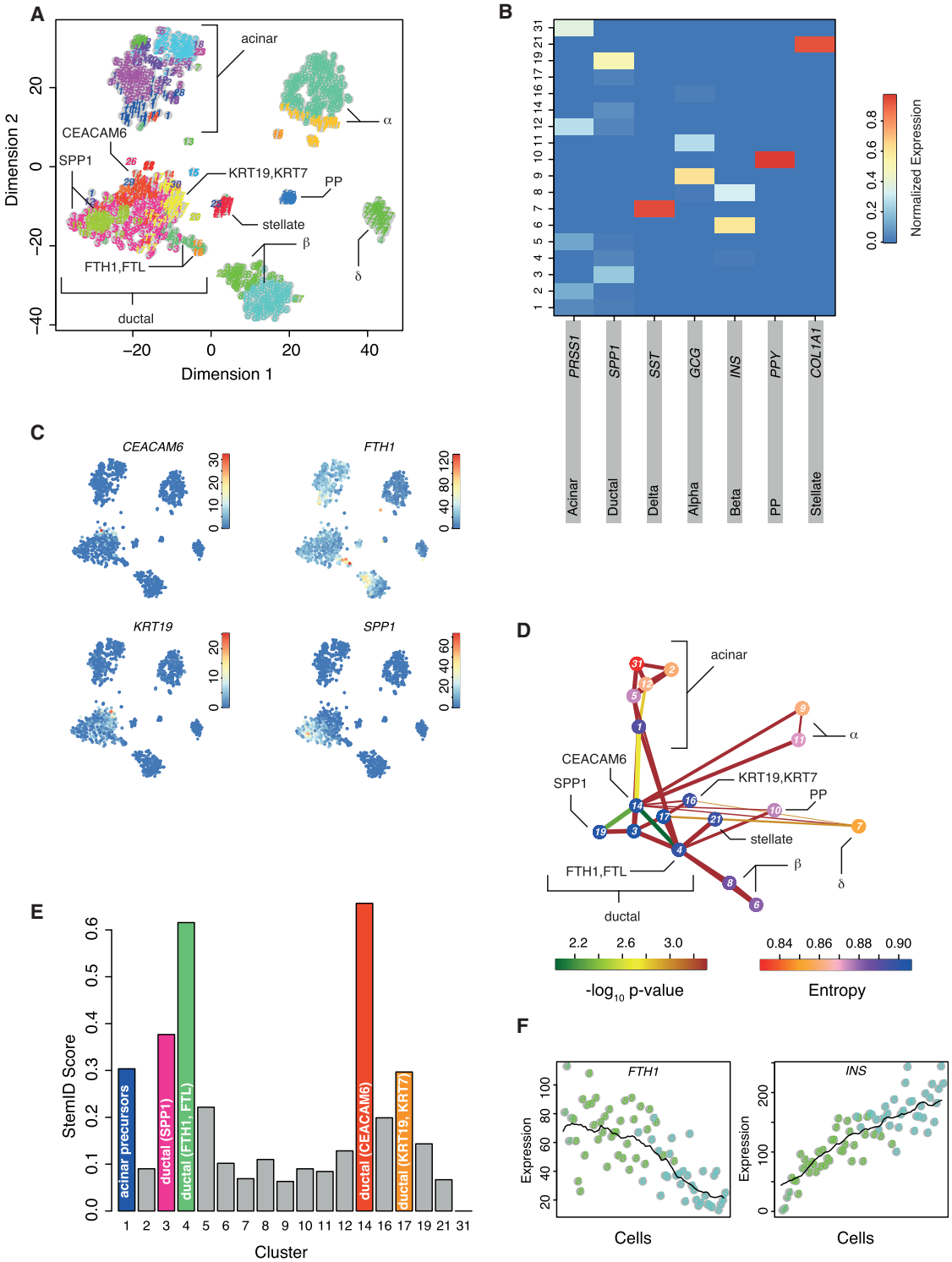


Figure 6. StemID Predicts Human Pancreatic Pluripotent Cells

(A) t-SNE map of transcriptome similarities of human pancreatic cells. RaceID2 clusters are highlighted with different numbers and colors. Cell types identified based on marker gene expression are shown. For ductal cells, marker genes of sub-populations are shown.

(B) Heatmap showing the average expression of known cell type markers across all clusters with more than five cells. For each gene, the sum of expression values over all clusters is normalized to one.

(C) Transcript counts (color legend) of the ductal sub-type markers *CEACAM6*, *FTH1*, *KRT19*, and *SPP1* are highlighted in the t-SNE map.

(D) Inferred pancreatic lineage tree. Only significant links are shown ($p < 0.01$). 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 (Experimental Procedures).

(legend continued on next page)

ribosomal protein-encoding genes and other components of the translational machinery slowly decline during differentiation, other genes are transiently switched on in progenitor populations (e.g., *Elane*) or immature neutrophils (e.g., *Ngp*) or only upregulated in mature cells (e.g., *Retnlg*).

Finally, we note that the identification of the HSC population by StemID is robust to changing the contribution of this population to the mixed sample. For example, when only ten HSCs are randomly selected and all others are discarded from the dataset, StemID still assigns the highest score to the small HSC cluster (data not shown).

In summary, StemID could successfully identify the stem cell type in a complex mixture of cells isolated from bone marrow. The inferred lineage tree recovered known trajectories but suggested an early cell fate bias present already in HSCs.

StemID Predicts Multipotent Ductal Cell Populations among Human Adult Pancreatic Cells

After having demonstrated that StemID can robustly identify the stem cell population in two distinct biological systems, we applied the algorithm to predict multipotent cell populations in a less characterized system: the human pancreas. The pancreas consists of acinar cells that produce the digestive enzymes, ductal cells secreting bicarbonate to neutralize stomach acidity, and hormone-producing endocrine cells that regulate hormone metabolism (Jennings et al., 2015). It is unclear which multipotent cells maintain pancreatic homeostasis and can give rise to different mature cell types during regeneration upon injury. Although early studies have suggested that, in humans, these cell populations could reside within the exocrine compartment or that dedifferentiation of exocrine cells could give rise to endocrine cells (Bonner-Weir et al., 2000; Puri et al., 2015), the identity of multipotent cell populations is still unclear (Jiang and Morahan, 2014). We sequenced pancreatic cells from human donors (Experimental Procedures), and application of RaceID2 revealed all major cell types, including different subpopulations of acinar and ductal cells; hormone-producing α, β, δ, and pancreatic polypeptide producing (PP) cells; and stellate cells (Figures 6A and 6B; Figures S5A and S5B). A full list of differentially expressed genes for each cluster is shown in Table S4. In particular, we discovered novel subpopulations of ductal cells. In one of these groups (cluster 14), the cell surface glycoprotein *CEACAM6* was significantly upregulated ($p < 0.01$; Experimental Procedures), whereas components of the ferritin protein (*FTH1*, *FTL*), which is the major intracellular iron storage protein, were significantly upregulated ($p < 0.01$; Experimental Procedures) in the other group (cluster 4) (Figure 6C).

The inferred lineage tree assigns a central position to the ductal cells (Figure 6D; Figures S7C–S7E). Distinct subtypes of ductal cells appear to give rise to different endocrine sub-types and acinar cells. Although differentiation trajectories link cluster 4 to acinar, PP, and β cells, cluster 14 is linked to α and δ cells. Consistently, clusters 4 and 14 acquire the highest StemID score, indicating the highest level of multipotency among the cell types detected in this system (Figure 6E; Figure S7F). The following ranks of the StemID score were occupied by other ductal sub-types and precursor cells that give rise to two substates of acinar cells. Interestingly, cluster 4 also directly connects to stellate cells. Upon injury, these cells can switch to an activated state and migrate to the injured location to participate in tissue repair (Omary et al., 2007).

To collect further evidence that cluster 4 is an endocrine progenitor cell, we plotted the expression of the cluster 4 marker *FTH1* and the β cell marker insulin (*INS*) in single cells residing on the differentiation trajectory connecting these two cell types. Cells were ordered by their projection coordinate. The genes exhibited smooth, anti-correlated gradients suggestive of a continuous transition between these two cell types (Figure 6F). To independently validate this observation, we performed antibody staining against insulin and FTL in human pancreatic tissue sections. We were able to detect individual cells co-expressing insulin and FTL within ductal structures, confirming the existence of cluster 4 cells (Figure 7A). Co-staining of glucagon revealed that these cells specifically produce insulin and not glucagon (Figure 7B), as suggested by our analysis (Figure 6C). Our results indicate that the ferritin-positive sub-population of ductal cells might differentiate into mature β cells.

DISCUSSION

In this study, we present an approach to identify stem cells using single-cell transcriptomics data. Because the physiological state of a cell is an approximate reflection of its transcriptome, it is a reasonable assumption that cell types can be discriminated based on their transcriptome. However, determining the stem cell identity among all rare cell types discovered also requires the derivation of a lineage tree.

To address this task, we combined cell type identification by RaceID2 with a tree reconstruction by guided topology. We first introduce an improved version of our previous RaceID algorithm (Grün et al., 2015) with a more robust initial clustering step. The replacement of k-means by k-medoids leads to increased robustness of clustering for all datasets analyzed in the paper. For the complex intestinal dataset (Figure 3), the fraction of clusters with Jaccard's similarity of > 0.7 is 40% for k-means versus 73% for k-medoids. The corresponding fractions are 58% versus 83% for the bone marrow data and 40% versus 90% for the pancreas data.

To infer differentiation trajectories, we assign every cell onto a specific link between its cluster of origin and another cluster based on the longest projection of the vector connecting the cluster center with the cell position onto these links. This adequately reflects how much a cell has moved from the most representative cell state in the same cluster (the medoid) toward another cell identity (or vice versa). If significantly more cells reside on a link than expected by chance, this provides strong evidence that cells of the cluster of origin exhibit a pronounced transcriptome bias toward another cell fate. In addition, if a

(E) Barplot of StemID scores for pancreatic clusters.

(F) Pseudo-temporal expression profiles for *INS* and *FTH1*. The transcript count is plotted for cells on the link, connecting clusters 4, 8, and 6. Cells are ordered by the projection coordinate.

In (B), (D), and (E), only clusters with more than five cells were analyzed. See also Figure S5.

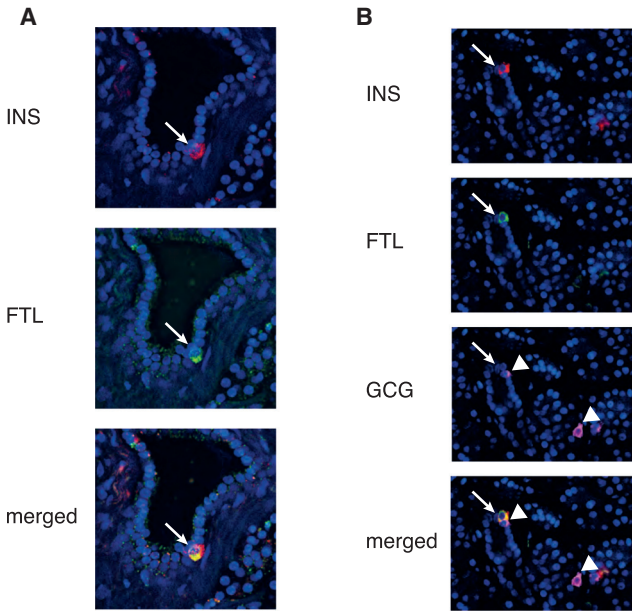


Figure 7. Validation of Putative Endocrine Precursor Cells in Ductal Subpopulations by Antibody Staining

(A and B) Antibody staining for INS and FTH1 in human pancreatic showing a single cell positive for INS and FTH1 residing in the lining of the duct (arrow). (B) Antibody staining for INS, FTH1, and GCG in human pancreatic tissue. Shown is a single cell positive for INS and FTH1 residing in the lining of the duct (arrow) next to a GCG-expressing cell (arrowhead). Another GCG-expressing cell is found nearby (arrowhead). Both GCG-expressing cells are FTH1-negative.

continuum of cell states covers a given link, as evidenced by a high link score, then this link represents a strong candidate for an actual differentiation trajectory. Significant links with reduced link scores, on the other hand, indicate plasticity of the connected cell types in a sense that the transcriptome of a cell type can, to some extent, fluctuate toward another fate.

The quality of our lineage inference is supported by the recovery of known differentiation trajectories in the intestinal epithelium and the bone marrow. Remarkably, we recovered a rare alternative differentiation pathway where *Lgr5*+ cells differentiate directly into Paneth cells without intermediate *Dll1*+ progenitors (Farin et al., 2014; Sawada et al., 1991). We could also show, for the intestinal and the bone marrow data, that StemID infers a lineage tree with substantially higher resolution in comparison with methods published previously (Haghverdi et al., 2015; Trapnell et al., 2014; Figure S6).

The derived lineage tree for the bone marrow suggested that, in contrast to the classical view of dichotomous differentiation via a hierarchy of increasingly restricted progenitor populations (Giebel and Punzel, 2008), a cell fate bias already exists at stages as early as the HSC stage (Figure 5B). This observation is consistent with a recent single-cell transcriptome analysis revealing heterogeneity of the common myeloid progenitor cell population, indicating early fate bias (Paul et al., 2015). Moreover, direct generation of progenitors restricted to the myeloid fate from mouse HSCs has been described in the past (Yamamoto et al., 2013), and the existence of unipotent cells within human HSCs (Notta et al., 2016) and classically defined mouse multipotent progenitor populations was shown recently (Perié et al., 2015).

For both model systems, the StemID score, which quantifies very general properties of a cell type (i.e., the number of links and the entropy of the transcriptome), ranks RaceID2-predicted cell types by their level of multipotency. *Lgr5*+ CBCs and sorted HSCs acquire the highest score among all cell types of the intestine and bone marrow, respectively, demonstrating the performance of our algorithm. We could further demonstrate the performance of StemID on two previously published datasets (Figure S7) for cells from developing lung epithelium (Treutlein et al., 2014) and differentiating human radial glial cells (Pollen et al., 2015).

Potential problems for the StemID algorithm arise in the absence of intermediate progenitors or the occurrence of unrelated cell types. In the absence of intermediate progenitors, StemID infers a link to a more multipotent population. For example, B lymphocytes in the bone marrow dataset are directly linked to HSCs. It is known that a spectrum of progenitors will reside on this trajectory, and, as we have observed for the other lineages, an early fate bias toward lymphocytes could exist in HSCs. In the absence of intermediate progenitors, a link to a more multipotent population reflects all information on the lineage relationship that can be extracted from the data. If the stem cell itself is missing from the sample, StemID will identify the cell type with the highest level of multipotency. The presence of unrelated cell types in the mixture could lead to false positive links. However, because the feature space is high-dimensional, it is likely that none of the links between an unrelated cell type and the remaining lineage tree will be significantly populated. We also argue that links of mature cell types to related progenitor or stem cell populations were identified with high specificity (oftentimes only a single link in line with previous findings was detected). This makes the occurrence of significant links between unrelated cell types unlikely.

Finally, we used StemID to screen human adult pancreatic cells for multipotent cell populations. It is unclear which adult pancreatic cell types can give rise to the different mature pancreatic lineages during normal tissue turnover or regeneration. Although initial evidence suggested that multipotent cells within the ductal compartment could differentiate into endocrine cells both in humans and mice (Jiang and Morahan, 2014), subsequent lineage-tracing experiments produced contradictory results. Although mouse lineage tracing of carbonic anhydrase II (Ca2)-positive ductal cells revealed that these cells give rise to β cells upon injury (Bonner-Weir et al., 2008), lineage tracing of *Sox9*-, *Muc1*-, or *Hnf1β*-positive cells could not confirm this finding (Furuyama et al., 2011; Kopinke and Murtaugh, 2010; Kopp et al., 2011; Solar et al., 2009). Using StemID, we were able to predict distinct sub-populations of ductal cells with varying differentiation potential. Although ductal cells marked by high levels of CEACAM6 are predicted to differentiate into α, δ, and PP cells, another sub-population expressing high levels of the ferritin complex primarily appears to give rise to β cells and acinar cells. We note that the latter sub-population does not express any of the markers used in previous lineage-tracing experiments, but we caution that expression of these genes might be too low to be reliably detected by single-cell mRNA sequencing.

We further remark that β cell differentiation in the adult pancreas might not be conserved between human and mouse.

We provide the well documented R source code for RaceID2 and the StemID algorithm at https://github.com/dgrun/StemID. We hope that StemID will be useful for a better understanding of differentiation dynamics in a variety of systems.

EXPERIMENTAL PROCEDURES

Lineage-Tracing Experiments

For lineage-tracing experiments, we injected 0.4 mg tamoxifen into 3-month-old Lgr5-CreERT2 C57Bl6/J mice bred to Rosa26LSL-YFP reporter mice.

Isolation of Crypts from Mouse Small Intestine

Crypts were isolated from mice as described previously (Sato et al., 2009). See the Supplemental Experimental Procedures for more details.

Human Islet Isolation, Dispersion, and Sorting

Pancreatic cadaveric tissue was procured from a multiorgan donor program and only used when the pancreas could not be used for clinical pancreas or islet transplantation, according to national laws, and when 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. (1988). See the Supplemental Experimental Procedures for details regarding culturing and cell sorting.

Immunofluorescence

Pancreatic tissue samples were fixed overnight in 4% formaldehyde (Klinipath), stored in 70% ethanol, and subsequently embedded in paraffin. After deparaffinization and rehydration in xylene and ethanol, respectively, antigen retrieval was performed in citric buffer for 20 min. Sections were blocked with 2% normal donkey serum and 1% lamb serum in PBS. Primary antibodies were rabbit anti-Ftl (ab69090), mouse anti-glucagon (ab10988), and guinea pig anti-insulin (ab7842). Alexa Fluor-conjugated secondary antibodies against rabbit, mouse, and guinea pig immunoglobulin G (IgG) (Life Technologies; A11008, A10037, and A21450) were used at a dilution of 1:200. Nuclear counterstaining was done by embedding with DAPI Vectashield (Vector Laboratories, H-1500). Imaging was performed on a Leica SP8 confocal microscope using hybrid detectors.

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 of the Royal Netherlands Academy of Arts and Sciences 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). See the Supplemental Experimental Procedures for details regarding single cell isolation.

Single-Cell Sequencing Library Preparation

The protocol was carried out as described previously (Grün et al., 2015). See the Supplemental Experimental Procedures for a detailed description.

Quantification of Transcript Abundance

Read mapping and quantification were done as described previously (Grün et al., 2015). See the Supplemental Experimental Procedures for a detailed description.

RaceID2 and StemID

A brief overview is given in the Results. The algorithm and follow-up analyses are described in full detail in the Supplemental Experimental Procedures.

ACCESSION NUMBERS

The accession numbers for the RNA sequencing datasets reported in this paper are GEO: GSE76408, GSE76983, and GSE81076.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.stem.2016.05.010.

AUTHOR CONTRIBUTIONS

D.G. and A.v.O. conceived the study. D.G. developed the algorithm and performed all computational analyses. Single-cell sequencing of pancreatic cells and antibody staining were performed by M.J.M with the help of G.D. Single-cell sequencing of intestinal cells was performed by K.W. with the help of A.L., J.v.E., and M.v.d.B. Single-cell sequencing of bone marrow cells was performed by J.C.B. E.J. helped with antibody staining. D.G. wrote the manuscript, and all authors read and edited the manuscript. A.v.O. supervised D.G., M.J.M., K.W., and J.C.B. and the project itself. E.J.P.d.K. supervised G.D. and E.J. H.C. supervised M.v.d.B. and J.v.E.

ACKNOWLEDGMENTS

This work was supported by European Research Council Advanced Grant ERC-AdG 294325-GeneNoiseControl, a Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) Vici award, the DON Foundation, and the Dutch Diabetes Research Foundation.

Received: February 24, 2016
Revised: April 4, 2016
Accepted: May 12, 2016
Published: June 23, 2016

REFERENCES

Anavy, L., Levin, M., Khair, S., Nakanishi, N., Fernandez-Valverde, S.L., Degnan, B.M., and Yanai, I. (2014). BLIND ordering of large-scale transcriptomic developmental timecourses. Development *141*, 1161–1166.

Banerji, C.R.S., Miranda-Saavedra, D., Severini, S., Widschwendter, M., Enver, T., Zhou, J.X., and Teschendorff, A.E. (2013). Cellular network entropy as the energy potential in Waddington's differentiation landscape. Sci. Rep. *3*, 3039.

Barker, N. (2014). Adult intestinal stem cells: critical drivers of epithelial homeostasis and regeneration. Nat. Rev. Mol. Cell Biol. *15*, 19–33.

Barker, N., van Es, J.H., Kuipers, J., Kujala, P., van den Born, M., Cozijnsen, M., Haegebarth, A., Korving, J., Begthel, H., Peters, P.J., and Clevers, H. (2007). Identification of stem cells in small intestine and colon by marker gene Lgr5. Nature *449*, 1003–1007.

Bendall, S.C., Davis, K.L., Amir, A.D., Tadmor, M.D., Simonds, E.F., Chen, T.J., Shenfeld, D.K., Nolan, G.P., and Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell *157*, 714–725.

Bonner-Weir, S., Taneja, M., Weir, G.C., Tatarkiewicz, K., Song, K.H., Sharma, A., and O'Neil, J.J. (2000). In vitro cultivation of human islets from expanded ductal tissue. Proc. Natl. Acad. Sci. USA *97*, 7999–8004.

Bonner-Weir, S., Inada, A., Yatoh, S., Li, W.-C., Aye, T., Toschi, E., and Sharma, A. (2008). Transdifferentiation of pancreatic ductal cells to endocrine beta-cells. Biochem. Soc. Trans. *36*, 353–356.

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., and Heisler, M.G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods *10*, 1093–1095.

Buczacki, S.J.A., Zecchini, H.I., Nicholson, A.M., Russell, R., Vermeulen, L., Kemp, R., and Winton, D.J. (2013). Intestinal label-retaining cells are secretory precursors expressing Lgr5. Nature *495*, 65–69.

Busch, K., Klapproth, K., Barile, M., Flossdorf, M., Holland-Letz, T., Schlenner, S.M., Reth, M., Höfer, T., and Rodewald, H.-R. (2015). Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. Nature *518*, 542–546.

Eaves, C.J. (2015). Hematopoietic stem cells: concepts, definitions, and the new reality. Blood *125*, 2605–2613.

Eldar, A., and Elowitz, M.B. (2010). Functional roles for noise in genetic circuits. Nature *467*, 167–173.

Farin, H.F., Karthaus, W.R., Kujala, P., Rakhshandehroo, M., Schwank, G., Vries, R.G.J., Kalkhoven, E., Nieuwenhuis, E.E.S., and Clevers, H. (2014). Paneth cell extrusion and release of antimicrobial products is directly controlled by immune cell-derived IFN-γ. J. Exp. Med. *211*, 1393–1405.

Furuyama, K., Kawaguchi, Y., Akiyama, H., Horiguchi, M., Kodama, S., Kuhara, T., Hosokawa, S., Elbahrawy, A., Soeda, T., Koizumi, M., et al. (2011). Continuous cell supply from a Sox9-expressing progenitor zone in adult liver, exocrine pancreas and intestine. Nat. Genet. *43*, 34–41.

Giebel, B., and Punzel, M. (2008). Lineage development of hematopoietic stem and progenitor cells. Biol. Chem. *389*, 813–824.

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.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., and Amit, I. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science *343*, 776–779.

Jennings, R.E., Berry, A.A., Strutt, J.P., Gerrard, D.T., and Hanley, N.A. (2015). Human pancreas development. Development *142*, 3126–3137.

Jiang, F.-X., and Morahan, G. (2014). Pancreatic stem cells remain unresolved. Stem Cells Dev. *23*, 2803–2812.

Kiel, M.J., Yilmaz, O.H., Iwashita, T., Yilmaz, O.H., Terhorst, C., and Morrison, S.J. (2005). SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. Cell *121*, 1109–1121.

Kopinke, D., and Murtaugh, L.C. (2010). Exocrine-to-endocrine differentiation is detectable only prior to birth in the uninjured mouse pancreas. BMC Dev. Biol. *10*, 38.

Kopp, J.L., Dubois, C.L., Schaffer, A.E., Hao, E., Shih, H.P., Seymour, P.A., Ma, J., and Sander, M. (2011). Sox9+ ductal cells are multipotent progenitors throughout development but do not produce new endocrine cells in the normal or injured adult pancreas. Development *138*, 653–665.

Lancaster, M.A., Renner, M., Martin, C.-A., Wenzel, D., Bicknell, L.S., Hurles, M.E., Homfray, T., Penninger, J.M., Jackson, A.P., and Knoblich, J.A. (2013). Cerebral organoids model human brain development and microcephaly. Nature *501*, 373–379.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Notta, F., Zandi, S., Takayama, N., Dobson, S., Gan, O.I., Wilson, G., Kaufmann, K.B., McLeod, J., Laurenti, E., Dunant, C.F., et al. (2016). Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. Science *351*, 126–127.

Omary, M.B., Lugea, A., Lowe, A.W., and Pandol, S.J. (2007). The pancreatic stellate cell: a star on the rise in pancreatic diseases. J. Clin. Invest. *117*, 50–59.

Orkin, S.H., and Zon, L.I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. Cell *132*, 631–644.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396–1401.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., et al. (2015). Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell *163*, 1663–1677.

Perié, L., Duffy, K.R., Kok, L., de Boer, R.J., and Schumacher, T.N. (2015). The Branching Point in Erythro-Myeloid Differentiation. Cell *163*, 1655–1662.

Piras, V., Tomita, M., and Selvarajoo, K. (2014). Transcriptome-wide variability in single embryonic development cells. Sci. Rep. *4*, 7137.

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.

Puri, S., Folias, A.E., and Hebrok, M. (2015). Plasticity and dedifferentiation within the pancreas: development, homeostasis, and disease. Cell Stem Cell *16*, 18–31.

Raj, A., and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. Cell *135*, 216–226.

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.

Ridden, S.J., Chang, H.H., Zygalakis, K.C., and MacArthur, B.D. (2015). Entropy, Ergodicity, and Stem Cell Multipotency. Phys. Rev. Lett. *115*, 208103.

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., and Clevers, H. (2009). Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. Nature *459*, 262–265.

Sawada, M., Takahashi, K., Sawada, S., and Midorikawa, O. (1991). Selective killing of Paneth cells by intravenous administration of dithizone in rats. Int. J. Exp. Pathol. *72*, 407–421.

Shannon, C.E. (1948). A Mathematical Theory of Communication. Bell Syst. Tech. J. *27*, 379–423, 623–656.

Solar, M., Cardalda, C., Houbracken, I., Martín, M., Maestro, M.A., De Medts, N., Xu, X., Grau, V., Heimberg, H., Bouwens, L., and Ferrer, J. (2009). Pancreatic exocrine duct cells give rise to insulin-producing beta cells during embryogenesis but not after birth. Dev. Cell *17*, 849–860.

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. Stat. 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.

van der Flier, L.G., and Clevers, H. (2009). Stem cells, self-renewal, and differentiation in the intestinal epithelium. Annu. Rev. Physiol. *71*, 241–260.

van Es, J.H., Sato, T., van de Wetering, M., Lyubimova, A., Nee, A.N.Y., Gregorieff, A., Sasaki, N., Zeinstra, L., van den Born, M., Korving, J., et al. (2012). Dll1+ secretory progenitor cells revert to stem cells upon crypt damage. Nat. Cell Biol. *14*, 1099–1104.

Wilson, N.K., Kent, D.G., Buettner, F., Shehata, M., Macaulay, I.C., Calero-Nieto, F.J., Sánchez Castillo, M., Oedekoven, C.A., Diamanti, E., Schulte, R., et al. (2015). Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. Cell Stem Cell *16*, 712–724.

Yamamoto, R., Morita, Y., Ooehara, J., Hamanaka, S., Onodera, M., Rudolph, K.L., Ema, H., and Nakauchi, H. (2013). Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells. Cell *154*, 1112–1126.

Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142.

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}, …, $k_{\max}$), k_i is chosen as input. The average change across cluster numbers k_{i+2}, …, $k_{\max}$ 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 ($P<10^{-3}$).

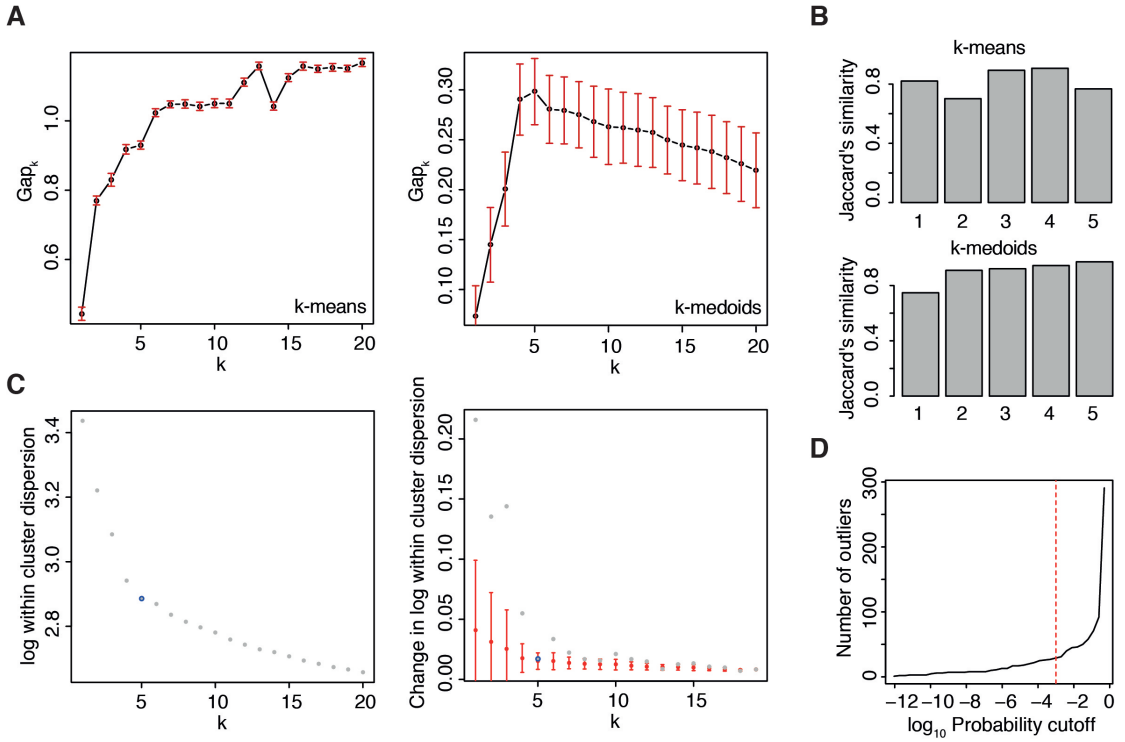

A
B
C
D
k-means
k-medoids
Gap$_k$
k
Jaccard's similarity
log within cluster dispersion
Change in log within cluster dispersion
Number of outliers
log$_{10}$ Probability cutoff

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.

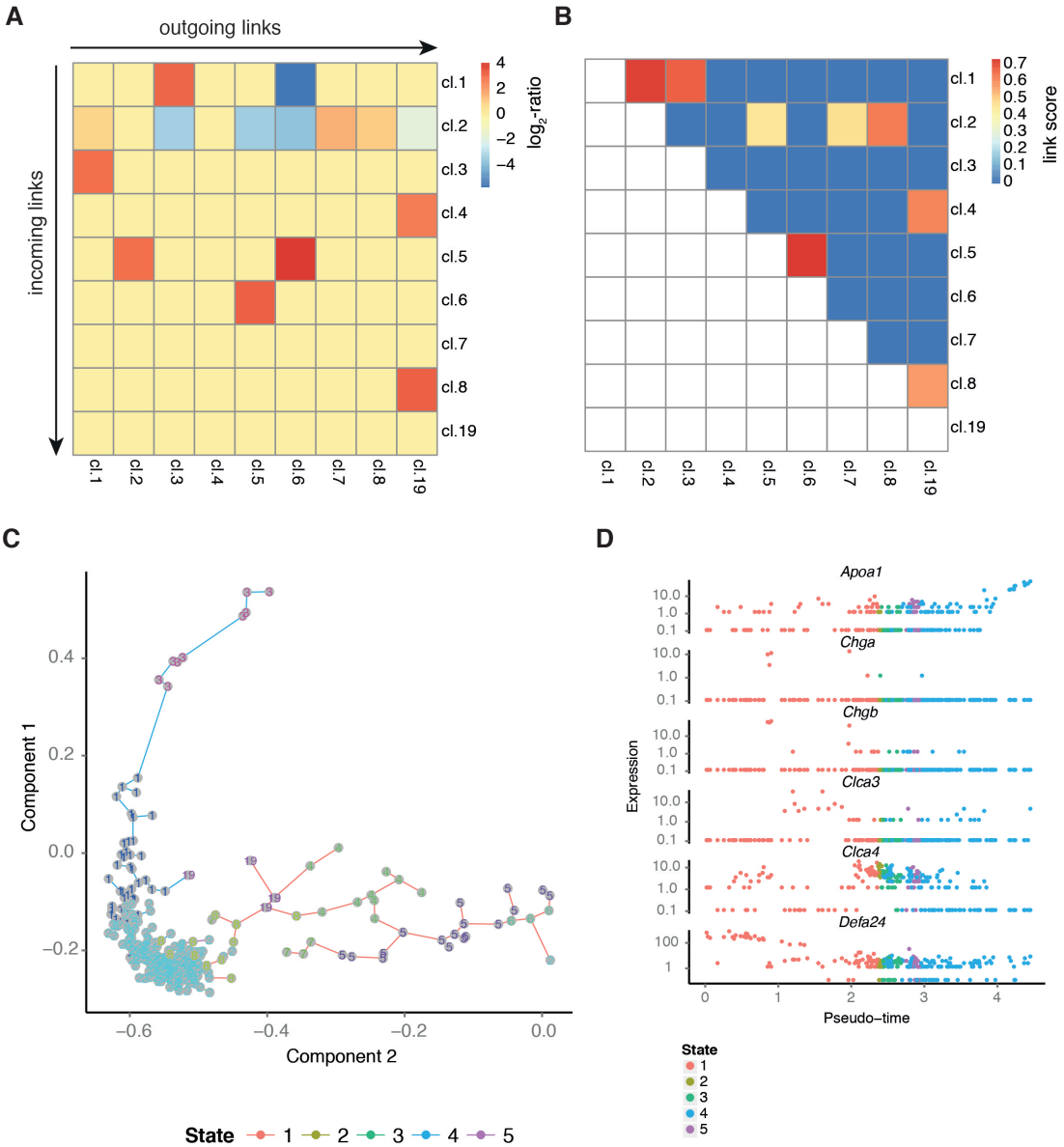

A
outgoing links
incoming links
log2-ratio
cl.1
cl.2
cl.3
cl.4
cl.5
cl.6
cl.7
cl.8
cl.19
B
link score
C
Component 1
Component 2
State 1 2 3 4 5
D
Apoa1
Chga
Chgb
Clca3
Clca4
Defa24
Expression
Pseudo−time
State
E
Clca3
Chga
Apoa1

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.

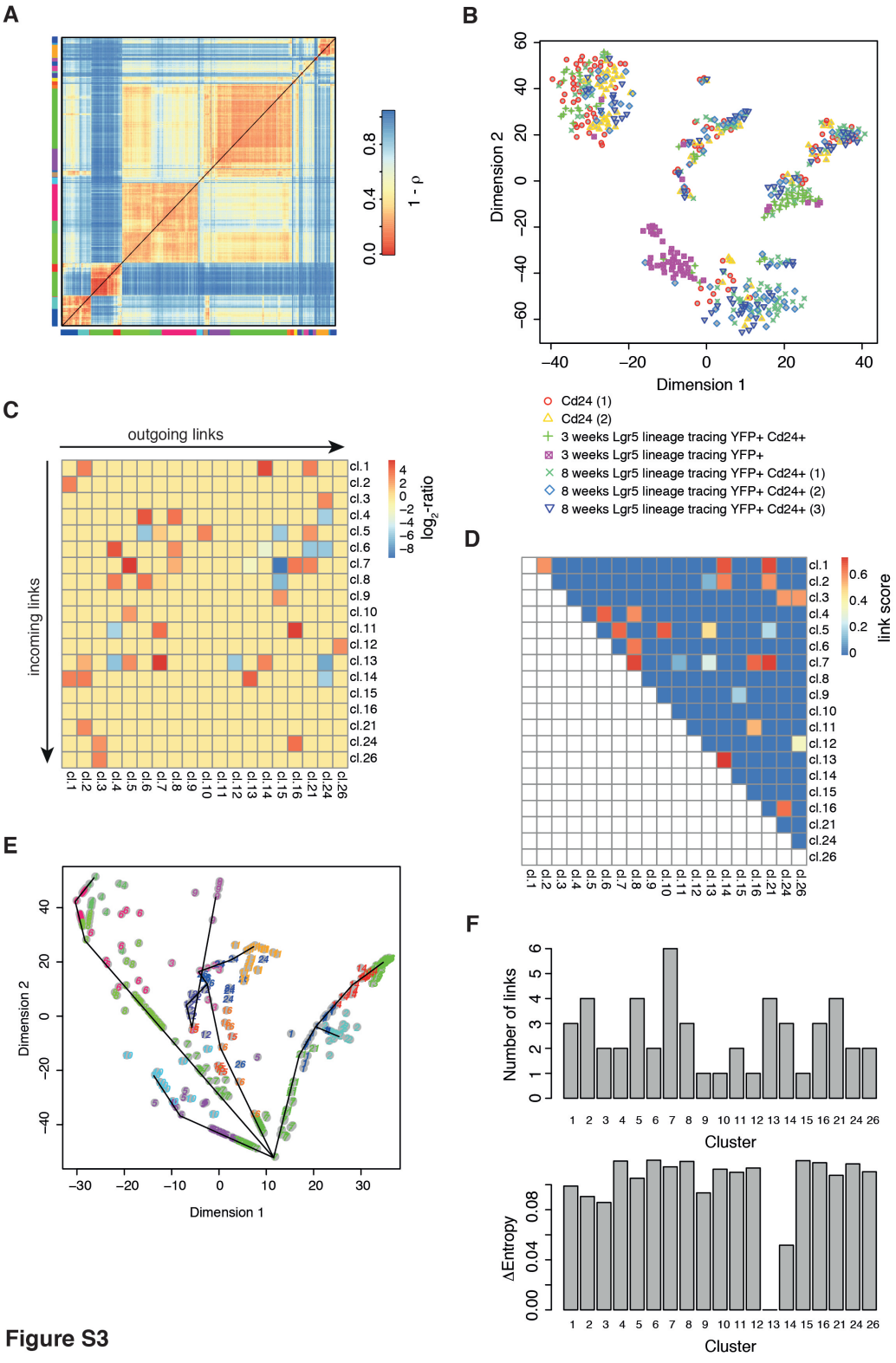

A
1 - ρ
0.0
0.4
0.8
B
Dimension 1
Dimension 2
Cd24 (1)
Cd24 (2)
3 weeks Lgr5 lineage tracing YFP+ Cd24+
3 weeks Lgr5 lineage tracing YFP+
8 weeks Lgr5 lineage tracing YFP+ Cd24+ (1)
8 weeks Lgr5 lineage tracing YFP+ Cd24+ (2)
8 weeks Lgr5 lineage tracing YFP+ Cd24+ (3)
C
outgoing links
incoming links
log2-ratio
D
link score
E
Dimension 1
Dimension 2
F
Number of links
Cluster
ΔEntropy
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^+ $Sca\text{-}1^+$ 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.

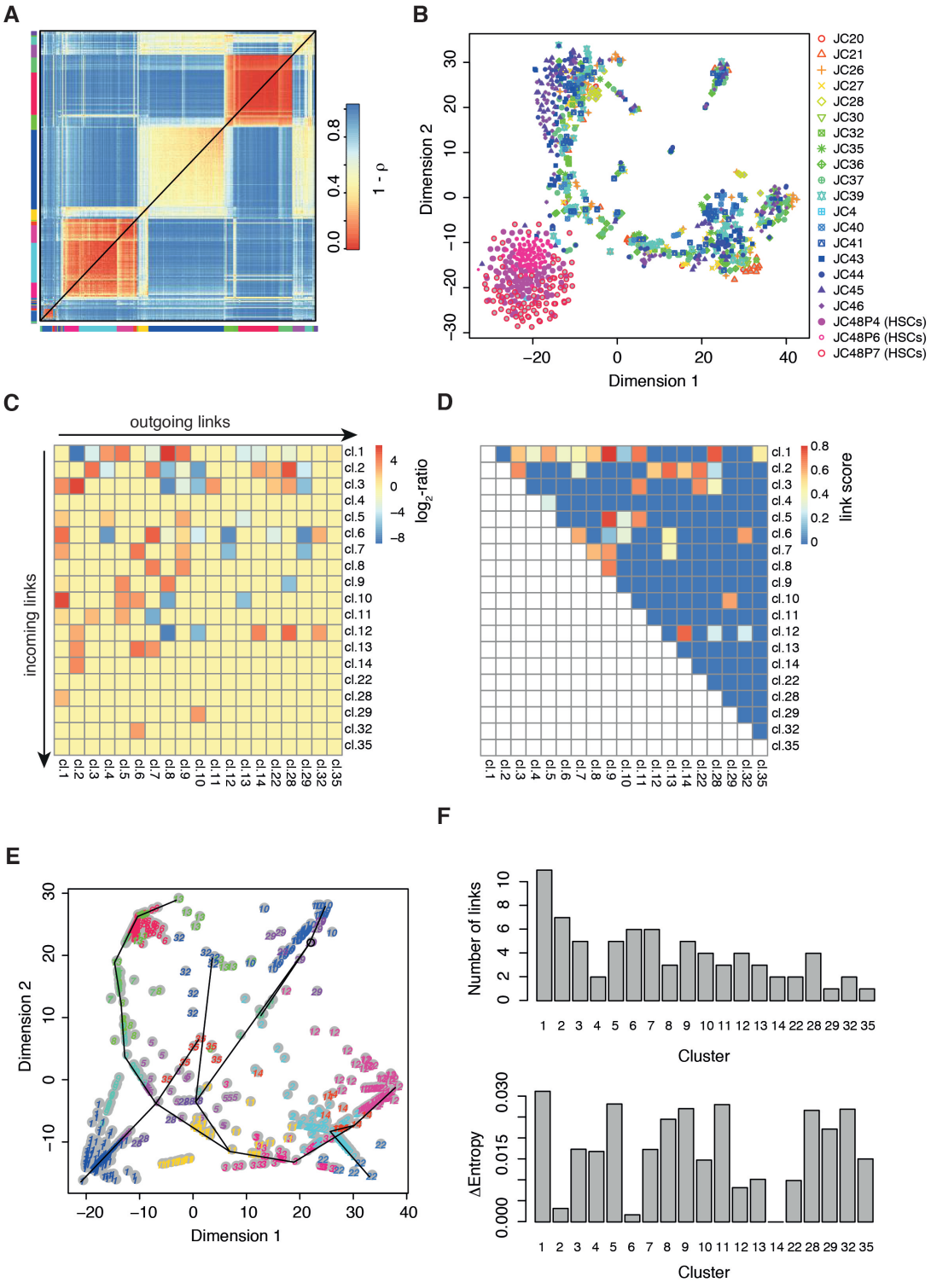

A
1 - ρ
0.0
0.4
0.8
B
Dimension 1
Dimension 2
JC20
JC21
JC26
JC27
JC28
JC30
JC32
JC35
JC36
JC37
JC39
JC4
JC40
JC41
JC43
JC44
JC45
JC46
JC48P4 (HSCs)
JC48P6 (HSCs)
JC48P7 (HSCs)
C
outgoing links
incoming links
log2-ratio
D
link score
E
Dimension 1
Dimension 2
F
Number of links
Cluster
ΔEntropy
Cluster

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.

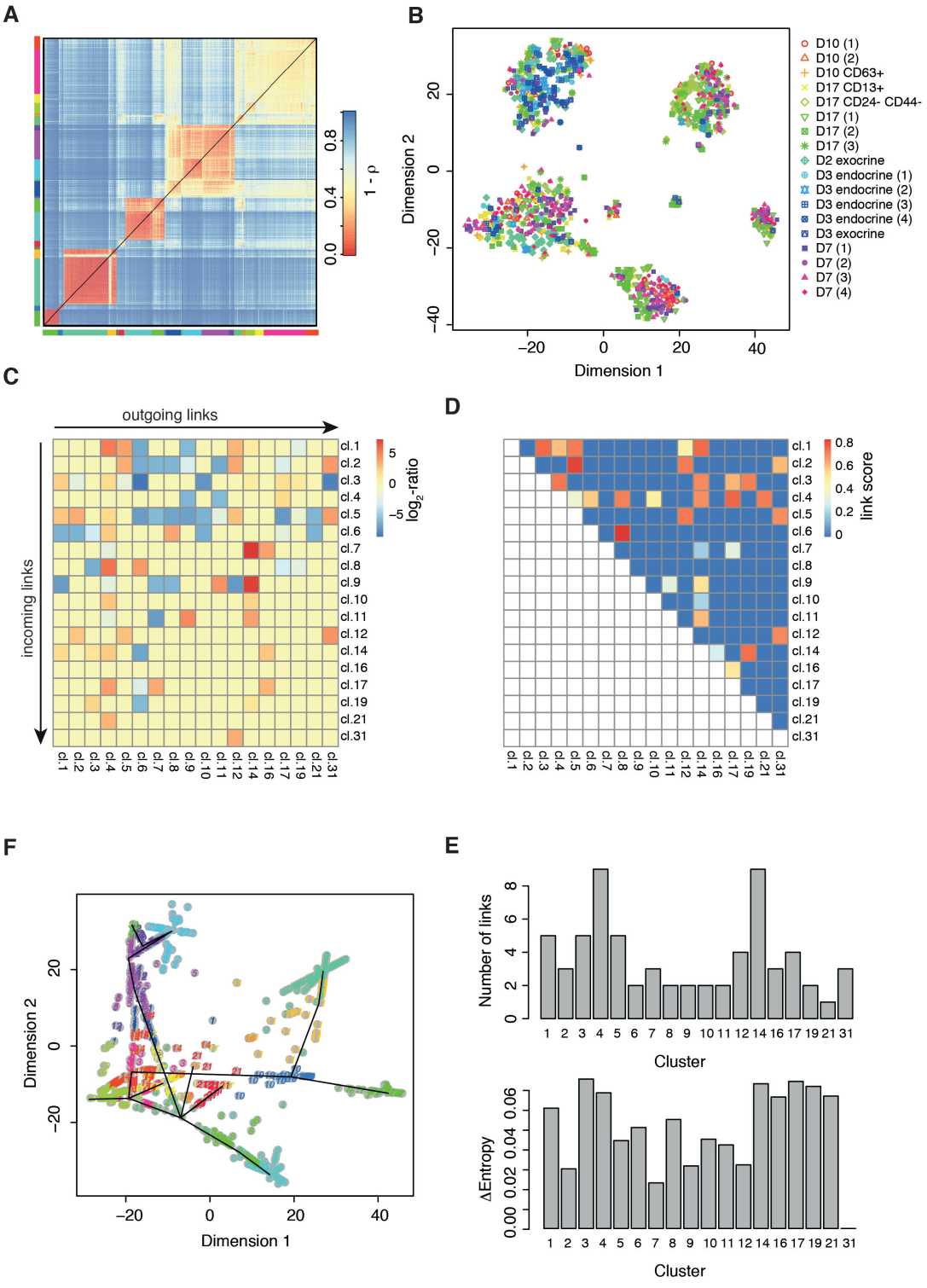

A
1 - ρ
0.0
0.4
0.8
B
Dimension 1
Dimension 2
D10 (1)
D10 (2)
D10 CD63+
D17 CD13+
D17 CD24- CD44-
D17 (1)
D17 (2)
D17 (3)
D2 exocrine
D3 endocrine (1)
D3 endocrine (2)
D3 endocrine (3)
D3 endocrine (4)
D3 exocrine
D7 (1)
D7 (2)
D7 (3)
D7 (4)
C
outgoing links
incoming links
log2-ratio
D
link score
F
Dimension 1
Dimension 2
E
Number of links
Cluster
ΔEntropy
Cluster

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.

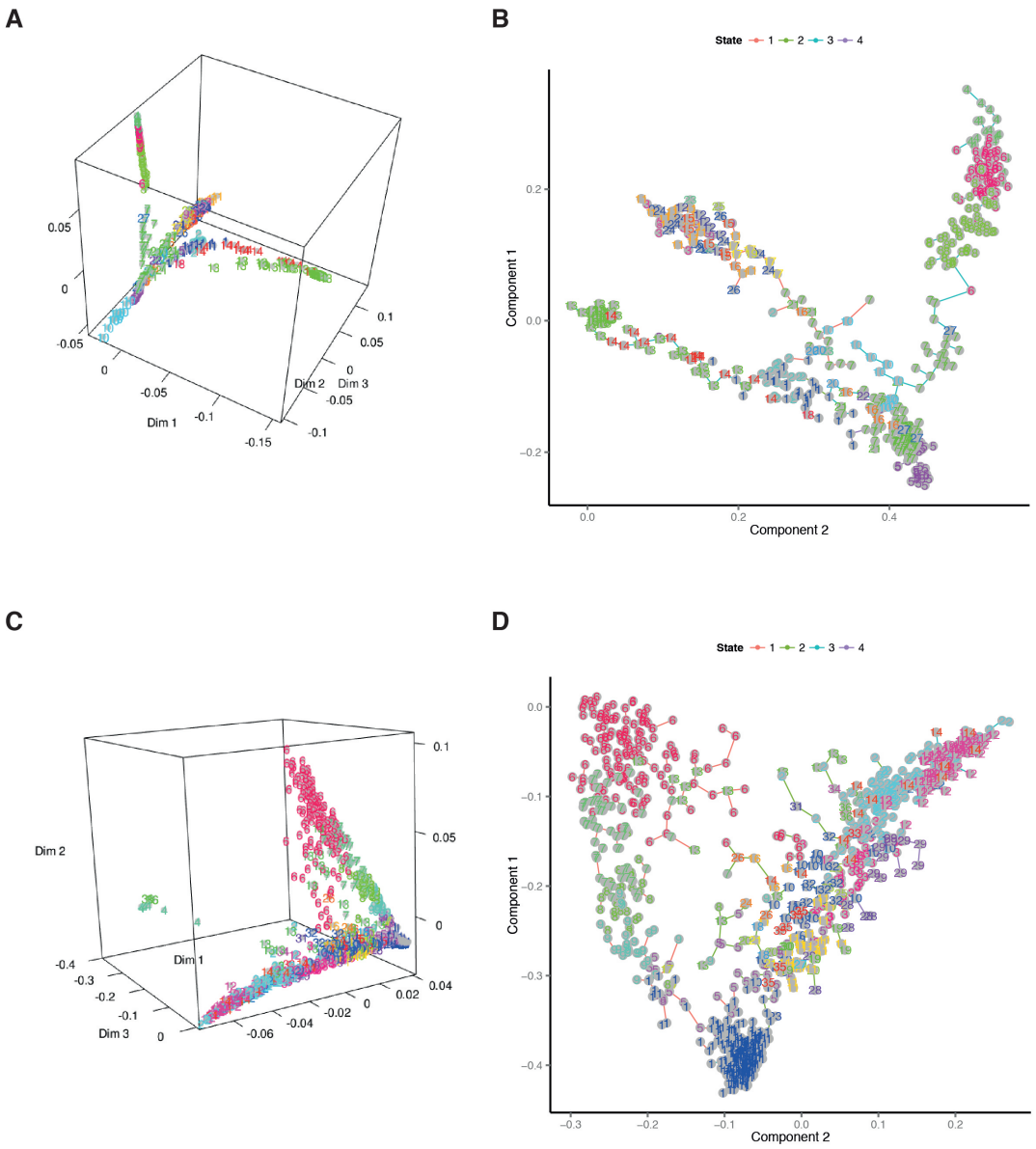


Figure S6

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.

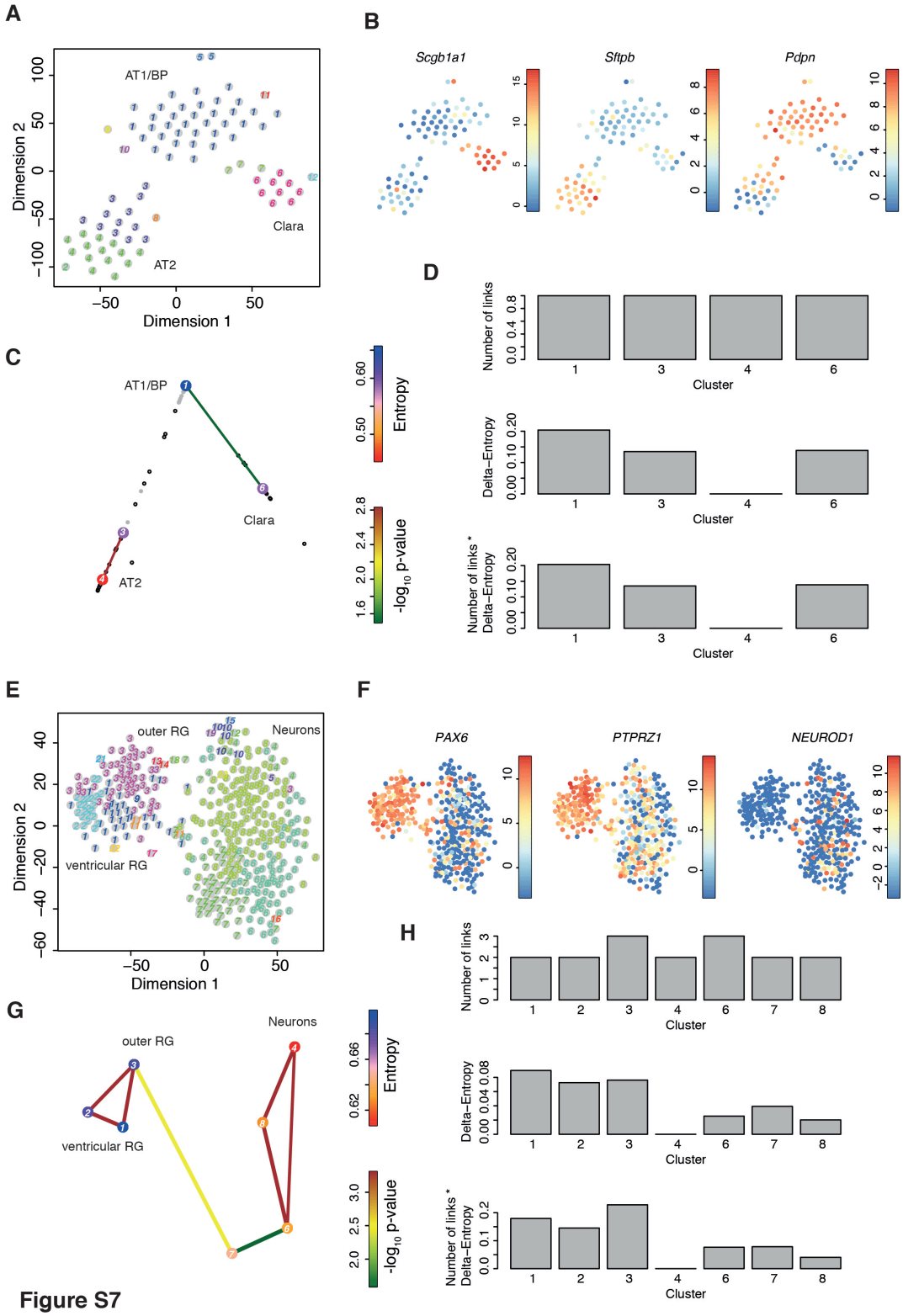

A
AT1/BP
Clara
AT2
Dimension 1
Dimension 2
B
Scgb1a1
Sftpb
Pdpn
C
AT1/BP
Clara
AT2
Entropy
-log10 p-value
D
Number of links
Delta−Entropy
Number of links * Delta−Entropy
Cluster
E
outer RG
Neurons
ventricular RG
Dimension 1
Dimension 2
F
PAX6
PTPRZ1
NEUROD1
G
outer RG
Neurons
ventricular RG
Entropy
-log10 p-value
H
Number of links
Delta−Entropy
Number of links * Delta−Entropy
Cluster

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 $d_{ij}=1-\rho_{i,j}$, where $\rho_{i,j}$ 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}, …, $k_{max.}$. 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^+ $Sca\text{-}1^+$ 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* pre-

ribosomal 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.