STEM CELLS

STEM CELL TECHNOLOGY: EPIGENETICS, GENOMICS, PROTEOMICS, AND METABONOMICS

Cross-Tissue Identification of Somatic Stem and Progenitor Cells Using a Single-Cell RNA-Sequencing Derived Gene Signature

PETRA C. SCHWALIE,[a] PALOMA ORDÓÑEZ-MORÁN,[b] JOERG HUELSKEN,[b] BART DEPLANCKE[a]

Key Words. Genomics • Progenitor cells • Self-renewal • Somatic stem cells • Stem/progenitor cells

[a]Laboratory of Systems Biology and Genetics, Institute of Bioengineering and Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland; [b]ISREC (Swiss Institute for Experimental Cancer Research), School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL)

Correspondence: Bart Deplancke, Ph.D., Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL) and Swiss Institute of Bioinformatics, Station 19, CH-1015 Lausanne, Switzerland. Telephone: +41216931821; e-mail: bart.deplancke@epfl.ch

Received April 5, 2017; accepted for publication October 2, 2017; first published online in STEM CELLS *EXPRESS* October 16, 2017.

http://dx.doi.org/10.1002/stem.2719

ABSTRACT

A long-standing question in biology is whether multipotent somatic stem and progenitor cells (SSPCs) feature molecular properties that could guide their system-independent identification. Population-based transcriptomic studies have so far not been able to provide a definite answer, given the rarity and heterogeneous nature of these cells. Here, we exploited the resolving power of single-cell RNA-sequencing to develop a computational model that is able to accurately distinguish SSPCs from differentiated cells across tissues. The resulting classifier is based on the combined expression of 23 genes including known players in multipotency, proliferation, and tumorigenesis, as well as novel ones, such as *Lcp1* and *Vgll4* that we functionally validate in intestinal organoids. We show how this approach enables the identification of stem-like cells in still ambiguous systems such as the pancreas and the epidermis as well as the exploration of lineage commitment hierarchies, thus facilitating the study of biological processes such as cellular differentiation, tissue regeneration, and cancer. STEM CELLS *2017; 00:000–000*

SIGNIFICANCE STATEMENT

This novel transcriptomics-based approach exploits the increased molecular resolution provided by single-cell RNA-sequencing to accurately identify somatic stem and progenitor cells (SSPCs) across a wide range of tissues. The developed classifier uniquely combines expression information of only few genes, suggesting that SSPCs share specific molecular properties across tissues. The method provides a valuable resource to identify SSPCs and order them across differentiation stages in yet poorly characterized systems. Among its downstream applications we envision, besides the detection of novel stem-like cells, the estimation of the fraction of SSPCs in a heterogeneous sample or, in an expanded version, the detection of highly malignant cells in tumors.

INTRODUCTION

The identification and molecular characterization of multipotent somatic stem and progenitor cells (SSPCs) is of fundamental interest for understanding development, homeostasis, and regeneration of complex multicellular organs. In addition, the study of cells with multipotent capacity provides novel regenerative therapy opportunities [1] as well as new insights into disease mechanisms or treatments [2]. While hematopoietic, neural, epidermal, and gastrointestinal stem cells have already been phenotypically well-characterized, it has proven very challenging to identify SSPCs and establish their hierarchy in a broad range of other systems, including the lung, kidney, mesenchyme, heart, liver, and pancreas [3–5]. This is largely because such SSPCs tend to constitute small and heterogeneous populations that reside in tissues of complex composition and lack universal markers, which leads to technological or experimental limitations such as population-level molecular measurements and tedious in vivo functional assays [6].

Recent advances in single-cell genomics have revolutionized our ability to reveal the composition of individual tissues or specific developmental patterns [7–10]. To determine which single-cell RNA-sequencing (scRNA-seq)-assessed cells are stem-like, studies tend to cluster or arrange them along a differentiation time-line (termed "pseudo-time") and attribute labels to groups based on marker gene expression [8–11]. This is an iterative and highly variable process, as it is sensitive to particular algorithmic (clustering, graph walking etc.) as well as biological (markers, cell stages) choices. The recently developed StemID alleviates these issues by streamlining the procedure, but it is only applicable to cells across multiple differentiation stages and not to homogeneous single-cell populations [12]. For such cases, one is obliged to leverage on the knowledge accumulated in other systems to assign functionality. To formalize such an approach, we developed a cross-tissue SSPC identification method.

A wide range of efforts have already been directed toward delineating a stem cell (SC) molecular signature (referred to as "stemness" signature) [13–19]. For example, over 100 stemness-related resources, including curated gene sets, computationally derived signatures, and transcription factor (TF) targets have been collected and made centrally available through the online platform StemChecker [20]. Importantly, while there is substantial overlap between the individual resources, there is by no means a consensus on the molecular program that underlies the core properties of SSPCs [20] or a universally accepted marker gene set that facilitates their identification [6]. This may be because SSPCs may only be captured through a combination of system-specific features rather than single marker genes [4, 16, 21]. At the same time, while scRNA-seq studies revealed cell hierarchies and novel cell types within systems of interest [7–10, 22–24], cross-study analyses have so far only focused on cell cycle stage identification [25]. In this study, we aimed to bridge this gap by exploring how well multipotent SSPCs could be identified and distinguished from differentiated cells based on scRNA-seq data. We demonstrate here that a lasso logistic regression-based model is able to accurately identify various SSPCs by using a unique combination of genes that together reach good cross-system performance.

Materials and Methods

In our approach, we carefully considered (a) the quality/diversity of the training data, (b) the genes included as features in the training, (c) the algorithm used, and (d) the normalization applied prior to training, exploring several options in order to obtain a robust model.

Datasets

We used the following data, also listed in Supporting Information Table S1: (1) as positive SSPCs, (a) as train and test data, we included quiescent and activated neural stem cells assessed after injury (qNSCs and aNSCs, d1) [10], macrophage dendritic cell progenitors (MDPs, d2) [22], hematopoietic stem cells (HSCs, d3, and d4) [26, 27] and (b) as independent test data quiescent and activated NSCs (qNSCs and aNSCs, d9) [28], mesenchymal stem cells (MSCs, d10) [29], short- and long-term HSCs (stHSCs and ltHSCs, d11–12) [30], intestinal stem cells (ISCs, d13) [9]; (2) as negative data (non-SSPCs), (a) as train and test data, we included astrocytes (STR and CTX, d5) [10], T cells (d6) [31], AT2 lung epithelial cells (AT2 cells, d7) [8], and adult non-stem neuronal cells (neuro, d8) [23] and (b) as independent test data non-stem intestinal cells (REG4+ and LGR5-, d14–15) [9], lymphoblastoid cell lines (LCLs, d16) [32], adult non-stem neuronal cells (neuro, d17) [33] as well as adult hepatocytes and endothelial cells isolated from human liver (d33) [34]. To test predictions on cells with a gradually decreasing level of stemness, we used: (a) oligodendrocytes (Oligo, d18), neuroblasts (NB, d19), neural transit amplifying cells (TAPs, d20), activated and quiescent NSCs (aNSCs and qNSCs, d21–22); (b) megakaryocyte progenitor cells (CD150- MkPCs, d23); (c) dendritic cell precursors (PreDCs, d24) and common dendritic cell progenitors (CDPs, d25) [22]; (d) an unsorted mixture of cells progressively loosing progenitor status from embryonic day 14 over 16 to 18 (E14, E16, and E18, d26–27) [8]. To keep the fraction of positive and negative cells balanced, we selected only a (random) subset of cells among those that were part of large datasets: d7, d8, and d11–12 [23, 30, 31]. Finally, we used dissociated mouse epidermis cells [35] as well as human pancreas cells [12, 36, 37] for generating novel predictions.

Data Normalization

We used the expression estimates provided by the individual studies, which were publicly available through Gene Expression Omnibus [38], irrespective of the underlying experimental and computational methods that they were based on (Supporting Information Table S1). For developing the stem and progenitor cell identification model, due to the highly heterogeneous nature of scRNA-seq, exacerbated by both experimental and data processing methodological differences among the studies as well as variability across distinct biological sources, we did not use these raw expression estimates. Rather, we applied rank-based normalization and reduced the dynamic range of the data by using quantiles instead of ranks or expression estimates. We ran the lasso logistic regression model described below on a range of ranked data (for non-zero data points: 10, 20, 30, 50, 100, 500, 1,000, and 1,500 quantiles, with zero as an additional lowest rank in each set; plus the full ranked matrix, Supporting Information Fig. S2A), finding that differences in the binomial deviances were minimal. We further used ventiles (20-quantiles), which showed minimal average values, for all reported results.

To generate the t-SNE maps [39] with the R library Rtsne, we used log-normalized expression estimates of all genes having maximal log expression across all cells in the datasets $\geq$ c, where c was calculated for each dataset to equal the third quartile of log expression estimates of a reference gene set (housekeeping genes and ERCC spike-ins when available) expressed in $\leq$ 50% of all cells. Cells in which a high fraction ($\geq$ third quartile + 2*standard deviation across all cells) of reference genes were not measured (=0) were removed from the analysis in a first step.

Model Training and Testing

We trained lasso logistic regression (*cv.glmnet* ≪ nfolds = 10 type.measure="deviance" lower.limit = 0 family="binomial" alpha = 1≫), elastic-net logistic regression (*cv.glmnet* ≪ nfolds = 10 type.measure="deviance" lower.limit = 0 family-"binomial" alpha = 0.5≫), and random forest (*randomForest* ≪importance = T proximity = T ntree = 1000≫) models on two thirds of d1–8 and tested them on the remaining one third as well as on the full d10–28 (Supporting Information Fig. S2B, S2C; Table S1) datasets, using the R libraries glmnet [40] and randomForest [41]. We used a set of 4,528 genes commonly annotated (but not necessarily expressed) in all train and test data (d1-d28) (All.g; Supporting Information Table S3) as starting features. We define commonly annotated genes as human-mouse orthologous genes with shared gene symbol across species and for which expression information (even if equal to 0) is present across all used datasets. For comparison, we also trained: (a) lasso logistic regression models (*cv.glmnet* ≪ nfolds = 10 type.measure="deviance" lower.limit = 0 familiy="binomial" alpha = 1≫) using as features only genes part of previous population-derived stemness signatures—PluriNet [42] and Wong et al. [43] (Lasso.LogReg, l1.PluriNet.g, and l1.Wong.g); (b) logistic regression models (glm ≪family="binomial") using as features only genes part of previous population-derived stemness signatures—PluriNet [42] and Wong et al. [43] or retained from the full set of 4,528 genes after l1 (lasso) regularization (LogReg, PluriNet.g, Wong.g, and SSPCI.g); (3) a logistic regression model (*glm* ≪family="binomial"≫) using as features the predicted probabilities for a cell to be in G1 or G2/M phase of the cell cycle [25] (Fig. 2A, 2C, Supporting Information Fig. S2B, S2C and Table S3). We also calculated the average correlation between a cell's gene expression and the average expression of known positive SSPCs (using log-normalized expression estimates of the 4,528 shared genes and Spearman's rank correlation coefficient), exploring how well this measure can be used to classify cells (Fig. 2A, 2C). ROCR curves were generated to visualize the performance of all distinct trained models on both the test and the independent test datasets using the ROCR R package (*performance* ≪"tpr","fpr"≫) [44].

The final reported model (referred to as SSPCI), was trained using lasso logistic regression and cross-validation, as implemented in the R library glmnet [40] (*cv.glmnet* ≪ nfolds = 10 type.measure="deviance" lower.limit = 0 familiy="binomial" alpha = 1 ≫). We generated predictions on new data using the function *predict* with the parameters ≪ pred.type = response≫ for obtaining fitted probabilities of class 1, referred to as "stemness probabilities." We used ≪pred.type="class" s = lambda.1se ≫ for binary predictions (1 = SSPC, 0 = non-SSPC), corresponding to the largest value of lambda for which the error is within one standard error of the minimum. Stemness probabilities were displayed using the function *beanplot* [45] (≪ what = c(1,1,1,0) overallline = "median" log=""bw="nrd0" ≫). The model was trained on two thirds of the data by 10-fold-cross-validation and then tested on the remaining one third of the data (not used for parameterization) as well as the completely independent test data. Feature selection was automatically performed by the lasso regularization, which sets the coefficients of a large number of genes to zero. The intersection of features retained >10% of the time (>10 across 100 repetitions) was further used as final feature set, corresponding to 23 genes in the final model. By constraining the model with ≪lower.limit = 0≫, we only obtained positive coefficients, more easily biologically interpretable as genes associated with SSPCs. We refer to the probability of being classified as positive as the "probability (*p*) of stemness" and to the cells predicted as positives (i.e., $p \geq .5$) as "somatic stem and progenitor cells" or "SSPCs" throughout the manuscript.

Gene Sets

One constraint of the initial training gene set is that it should ideally contain universally expressed orthologous genes across mouse and human, which are likely to be present in a new prediction dataset. The intersection of all our train and test data (d1-d28, irrespective of their expression level) resulted in 4,528 genes (All.g; Supporting Information Table S3), the majority of which we assume is likely to be universally measurable and thus present in future datasets. We assessed the robustness of the obtained gene signature to variations in the initial starting gene set by altering the test data (removing one or two datasets, and thus starting with 4.841, 6.803, and 7.317 genes, respectively), or by subsampling (10 repetitions) from the largest (7.317 genes) initial gene set. The vast majority (19 of 23) of signature genes were also retained in ≥80% of these alternative starting sets (Supporting Information Fig. S5C, Table S4). Moreover, among the top 10 signature genes ordered by logistic regression coefficient (and thus relative importance), 9 were retained in >80% of all tested models. Together, these results strongly suggest that the gene signature used by the classifier is robust.

Furthermore, we used genes part of previous population-derived stemness signatures—PluriNet [42] and Wong et al. [43] (PluriNet.g and Wong.g) to test how informative they are in separating scRNA-seq assessed SSPCs from non-SSPCs, as well as gene sets previously specifically associated with ISC function (1: previously used ISC marker genes, as listed on Wikipedia on 03.12.2015, ISC.g1 and 2: Munoz et al. ISC signature genes, ISC.g2 [46]), genes annotated with the "cell cycle" GO Term (GO:0007049) (Cyc.g), genes annotated with various differentiation and stemness-related terms, including "differentiation," development," "stem cell maintenance," "stem cell development," stem cell division," "stem cell commitment" ("GO:0030154," "GO:0048468," "GO:0019827," "GO:0048864," "GO:0017145," "GO:0072089," "GO:0048865," "GO:0048863") (Diff.g), and genes annotated with the "Metabolism" GO Term (GO:0008152) (Metab.g) to validate the biological relevance of the reported stemness probability (Supporting Information Table S3). We calculated the Spearman's rank correlation coefficient (rho) of the stemness probability with the expression level of genes in these sets. We found that the correlations obtained for the two ISC-specific gene sets were significantly higher than those obtained for the background or for all other gene sets ($p \leq .005$ for any of the comparisons, Wilcoxon rank sum test), suggesting that our approach indeed uncovered genes that are functionally relevant for stemness.

To identify stemness-specific genes in the intestinal organoids [9] (Supporting Information Table S3, ISC.g3), in the human pancreas [12, 36] (Supporting Information Table S2), and in the mouse epidermis [35] (Supporting Information Table S2) based on stemness probability, we took a two-step approach: (a) we only included genes expressed (>0 read counts) in ≥60% of the predicted positives and ≤20% of the predicted negatives and (b) we only included genes with a log expression estimate highly correlating (Pearson's *r*, FDR 0.05) with the stemness *p*. Finally, we analyzed the properties of genes retained by SSPCI (SSPCI.g) using StemChecker as well as gene ontology (Supporting Information Table S4).

We also used a list of "housekeeping" genes, expected to be universally expressed at similar levels across diverse scRNA-seq data based on their presence (>50% of cells) and high (FDR 0.05) Gaussian rank correlation with ERCC spike-ins in ≥2 datasets (Supporting Information Table S3).

Other Methods

Comparisons of stemness probabilities and expression values were performed using one-sided Wilcoxon rank sum tests and the significance of overlaps was assessed using one-sided Fisher's exact tests. The gene ontology enrichment was performed using the topGO library, the ≪elimCount≫ method and a *p* value cutoff of .001. Overlaps with population-derived signature gene sets were assessed using the webserver StemChecker [20]. Cell cycle stages were predicted using Cyclone with default parameters and based on known markers [25]. All analyses were performed using R version 3.1.1 and Bioconductor version 3.0.

For the intestinal organoid analyses, statistical computations were performed using GraphPad Prism6. Experimental data are presented as mean ± standard deviation. Statistical significance was assessed by two-tailed unpaired Student's *t* test; the data were considered not significant (ns) for $p > .05$.

Mouse Model

Animal experiments were performed in accordance with protocols approved by the "Service de la Consommation et des Affaires Vétérinaires" of Canton Vaud, Switzerland. APC^{min} mice were described previously [47]. This strain was backcrossed onto C57BL/6 for at least 10 generations.

Isolation of Intestinal Tissue and Organoid Culture

Organoid cultures were established from total intestinal crypt preparations of APC^{min} mice as described previously [48]. Freshly isolated small intestines were incised along their length and villi were removed by scraping. Then, the adenomas were isolated and after several washes, the tissue was incubated in PBS/EDTA (Sigma-Aldrich, St. Louis, MO, US) (2 mM), pH8 for 5 minutes at 4°C. Gentle shaking removed remaining villi, and intestinal tissue was subsequently incubated in phosphate-buffered saline (PBS)/EDTA (2 mM), pH8 for 30 minutes at 4°C. Upon dissociation, samples were passed through a 70 µm filter and washed four times in cold Advanced Dulbecco's Modified Eagle's Medium (DMEM)/F12, (Gibco, NY, US) media. Afterward crypt cells were embedded in Matrigel (BD Biosciences, San Jose, CA, US) and plated in 400 µl of organoid media (Advanced DMEM/F12 with B27, N2, and N-acetylcysteine; containing growth factors EGF 50 ng/ml (Invitrogen, NY), R-Spondin 1 µg/ml, Noggin 100 µg/ml) into 24-well plates at a concentration of 500 crypts per well. Growth factors were added every other day and the entire medium was changed every 4 days. Secondary organoid culture was achieved by removing the organoids from Matrigel followed by mechanical dissociation with a glass pipette, gentle centrifugation (800 rpm), and then transferred to fresh Matrigel. Serial passaging (passage 1, passage 3, and passage 4) of shRNA-expressing organoids was compared to control organoids.

Plasmids, shRNAs, and Protein Production

Lcp1 (5′ TTGAAGAGATCGTCGGTGTTGG 3′) and *Vgll4* (5′ TCACTGCTGTTCTTAGTCAGGG 3′) were suppressed by shRNAs with the indicated antisense oligos in a MIR30 backbone expressed from a lentiviral vector (pSM2, Openbiosystems). The expression plasmid for R-spondin1 was a kind gift from Calvin Kuo (Standford University, U.S.) and this and Noggin proteins were produced as published [49] using proteinG or Ni-NTA purification, respectively.

Cell Culture and Selection of shRNA-TurboRFP Gene Suppression Cells by Flow Cytometry

The mouse colon carcinoma-derived cells CT26 were cultured in DMEM 10% fetal bovine serum (FBS, Invitrogen, NY, US). The shLcp1-TurboRFP, shVgll4-TurboRFP, and their respective controls were generated by lentiviral transduction. To this end, we generated a lentiviral vector driving TurboRFP expression that enables identification of positive cells by fluorescence-activated cell sorting (FACS). TurboRFP^{+} cells were sorted from a population of transduced cells by flow cytometry (FACSAria II; Beckton Dickinson (BD) Biosciences, Franklin Lakes, NJ, USA, or MoFlo; Dako, Hamburg, Germany).

Lentiviral Production and Transduction

Lentiviruses were generated in 293T cells using third generation lentivirus packaging vectors, and virus particles were concentrated by ultracentrifugation for 2 hours at 22,000 rpm. Lentiviral titration was performed in 293T cells. Cell lines were infected with lentiviruses overnight at 37°C. Colon carcinoma cell lines (CT26) and murine intestinal organoids were infected by lentiviral transduction as published [50]. Lentiviral infection was performed after 5 minutes Trypsin-EDTA (37°C) organoid dissociation. After several washes with 5%FBS-PBS, the cells were spun with the lentiviral particles at 1,300 rpm for 20 minutes and incubated with organoid media for 2 hours at 37°C. The pellet of single cells was then embedded in Matrigel and fresh organoid medium was added. Single cells gave rise to organoids after 10 days of culture. Images and qRT-PCR analyses were performed between day 13 and 17. Serial passaging was then performed at the following time points: P1 day 18, P2 day 28, P3 day 35, and P4 day 40.

Real-Time-qPCR

RNA was prepared using the mini or micro RNA kit (Qiagen, Germany) from organoids. cDNAs was synthesized using Superscript-II reverse transcriptase (Invitrogen, NY, US) and oligodT priming. qPCR was performed in a StepOnePlus thermocycler (Applied Biosystems Thermo Fisher, NY, US) using the Power SYBR green PCR Master Mix (Applied Biosystems, Thermo Fisher, NY, US) and the specific primers are listed below. Relative gene expression was determined by the comparative C_T method.

Murine gene	Forward primer	Reverse primer
Anpep	CCTGTAGCAAAGATGTGTGGATT	GGATGAGATTGGCAAAGGAGAAA
Lgr5	CTCCACACTTCGGACTCAACAG	AACCAAGCTAAATGCACCGAAT
MKi67	GATGGAAGCATTGTGAGAACCA	CCTGCTCTTCCACAGATTCAAG
Tff3	CTTTGACTCCAGTATCCCAAATG	TGGCTGTGAGGTCTTTATTCTTC
Vdr	AGGGACGTATCTTCAAACTCCA	AACGCATGATCAGCAAGAAGTA
Wwp1	ATGATGGCCAGTCTTCAAAAGT	GACATCCTACCTGAAAGCAACC
shLcp1	AACAAAGCCCTGGAGAATGAC	TGTTGATCGTTCTCTCGTCAA
shVgll4	CCACCTGTACGCATCTCTCC	GCCTGTGTCACTGCTGTTCTTA

Results

Accurate Identification of SSPCs Across Biological Systems

We collected diverse publicly available scRNA-seq data across biological systems and germ layers (Fig. 1, Supporting Information Fig. S1; Table S1, and Methods), using only two-thirds of the cells from each dataset (138 positives consisting of mostly stem cells but also early progenitors, including neural and hematopoietic stem cells as well as MDPs, and 170 negatives consisting of differentiated cells such as neuronal cells, T cells, and lung epithelial cells) for 10-fold-cross validation-

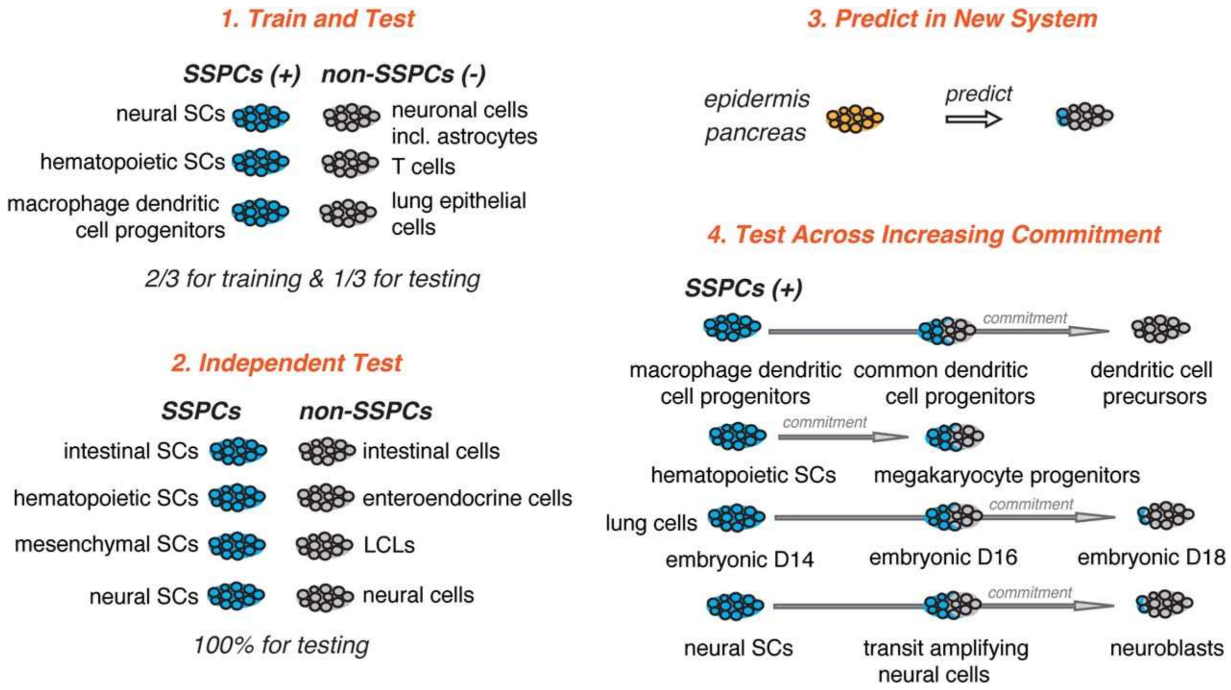


Figure 1. Classification strategy. Datasets used for training and testing of the models, see Supporting Information Table S1 for full description. Abbreviations: SSPCs, somatic stem and progenitor cells; SCs, stem cells; LCLs, lymphoblastoid cell lines.

based feature (gene) selection and parameterization. We separately retained a test set (one-third, 69 positives and 85 negatives) completely untouched during training and also used completely independent (467 positives and 554 negatives, including distinct studies and germ layers) test data (Fig. 1, Supporting Information Table S1 and Methods). We note that for simplicity, we refer to the probability of being classified as positive as the "probability (p) of stemness" and to the cells predicted as positives (i.e., $p \geq .5$) as "somatic stem and progenitor cells" or "SSPCs" throughout the article.

We evaluated a broad range of models on both test and independent test data and found that de novo lasso logistic regression (referred to as somatic stem and progenitor cell identifier, SSPCI, hereafter) showed best overall performance, followed by de novo elastic-net logistic regression and de novo randomForest (Fig. 2A–2D, Supporting Information Fig. S2B, S2C and Methods). While average correlation-based and logistic-regression models trained on population-derived stemness gene sets (both with or without regularization) performed only marginally inferior to SSPCI on the test samples (Fig. 2A), they showed worse performance on the independent test samples (Fig. 2C). Remarkably, SSPCI identified 89% of all SSPCs and 97% of all non-SSPCs as such (Fig. 2B, 2D, Supporting Information Table S1).

Importantly, SSPCI was also able to identify SSPCs in completely independent biological systems such as the mouse intestine [9], corresponding to a germ layer not originally covered in the training data (Fig. 2D, 2E, Supporting Information Fig. S2D). In particular, LGR5+ ISCs were attributed significantly higher stemness probabilities than the LGR5- cells (Fig. 2E, $p = 10^{-32}$, Wilcoxon rank sum test) and only 17 (of a total of 229) LGR5-cells were classified as positives. All of the latter were transcriptionally similar to the LGR5+ population, as visualized in a 2D t-SNE projection [39] (Fig. 2E, Supporting Information Fig. S2D). These cells likely represent highly proliferative and multipotent transit amplifying cells and are also partially characterized by high expression of *Hopx* and *Gnl3*, other commonly used stem cell markers (Fig. 2E, Supporting Information Fig. S2D) [46, 51, 52]. It is thus possible that some of these misclassification events do not represent false positive predictions, as these multipotent cells may genuinely be very close to ISCs and may thus also be categorized as SSPCs [5]. As an additional control, we tested the classifier on human liver data and confirmed that neither adult hepatocytes nor endothelial cells were classified as positives (Supporting Information Fig. S2E, S2F).

We next asked whether the cell cycle stage of the cells influences the performance of the classifier, given that proliferation and the ability to self-renew by cell division are intrinsically linked to stemness. To do so, we used a recently developed scRNA-seq-based method to attribute each cell to one specific cell cycle stage (G1, G2/M, S or if the assignment is ambiguous, na) [25]. This analysis revealed that across all our datasets, SSPCs did not show an association with a specific cell cycle stage (Fig. 2E, Supporting Information Fig. S2G, S2H). Furthermore, we found that a linear regression model trained using only the probabilities of cells being in one of the three cell cycle stages as features performs very poorly in identifying SSPCs (Supporting Information Fig. S2B, S2C). Thus, we conclude that SSPCs cannot be identified solely based on their cell cycle stage.

De Novo Identification of SSPCs in scRNA-Seq Dissected Tissues

We subsequently tested the power of our approach to resolve still ambiguous systems that were recently dissected by scRNA-seq: the human pancreas [12, 36, 37] and the mouse epidermis [35]. In the pancreas, we found that virtually all

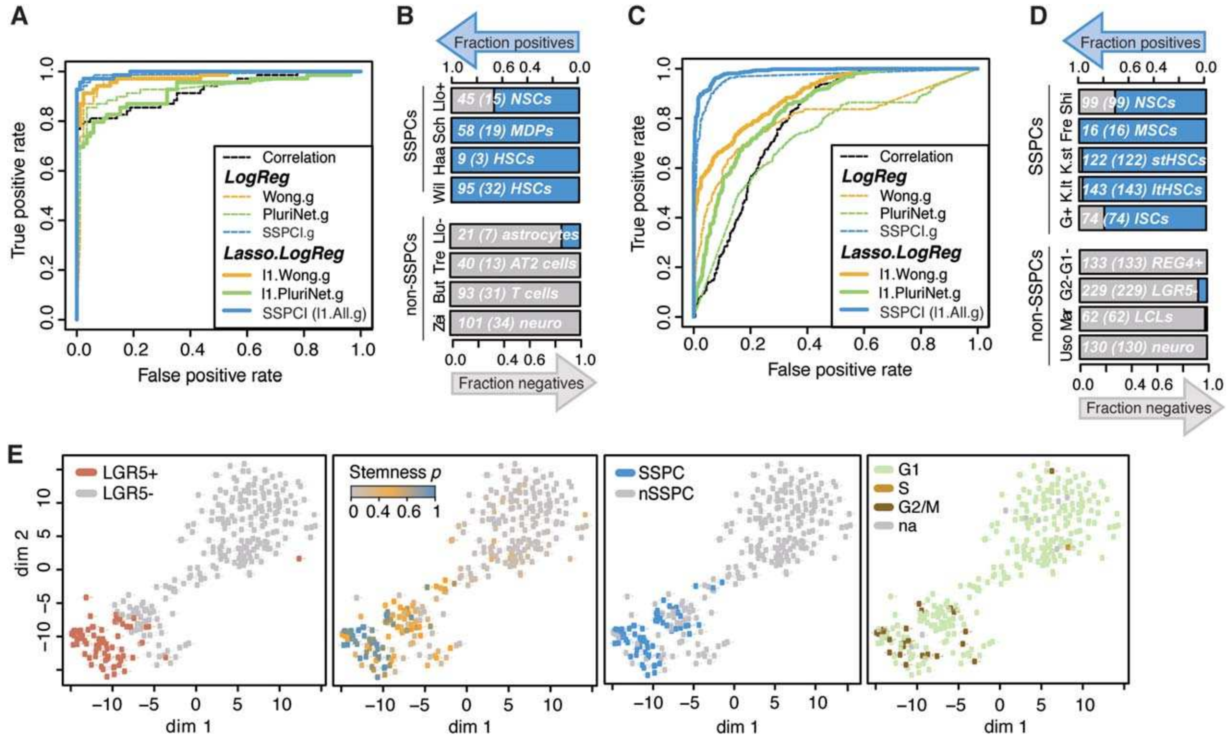


Figure 2. Accurate classification of somatic stem and progenitor cells. Receiver operating characteristic (true positive rate vs. false positive rate across all cutoffs) plots for distinct classifiers as assessed on all *test* **(A)** and *independent test* (distinct experiments or studies, **(C)** data, including SSPCI, the best performing model (de novo lasso-logistic-regression, blue, continuous line), models trained on population-derived stemness gene sets (yellow, Wong et al. [43] and green, PluriNetwork [42]) as well as the correlation with gene expression-based classification (black) (*Methods*). (A, C) The fraction of cells predicted positive (blue) and negative (gray) for each of the *test* **(B)** and *independent test* **(D)** datasets by SSPCI. The total number of cells as well as those used for testing only (in brackets) per each dataset is indicated. Supporting Information Table S1 contains information on the full dataset, including acronym legends. **(E)** t-SNE maps of intestinal epithelial ex vivo and organoid-derived LGR5+ (classically considered SCs, red) and LGR5- (assumed to be non-SCs, gray) cells [9]. Predicted stemness probabilities (gradient gray-orange-blue), binary classification outcome (blue: SSPC, gray: non-SSPC) by SSPCI, as well as predicted cell cycle stage [25] (light green: G1, light brown: S, dark brown: G2/M, gray: not assigned-na) are indicated for all cells. Abbreviations: ltHSCs, long-term hematopoietic stem cells; MDPs, macrophage dendritic cell progenitors; NSCs, neural stem cells; (n)SSPC, (negative) somatic stem and progenitor cell; stHSCs, short-term hematopoietic stem cells; MSCs, mesenchymal stem cells; ISCs, intestinal stem cells.

tested endocrine cells (alpha, beta, gamma, and delta) had very low stemness probabilities and were classified as negatives (Fig. 3A–3E). This result is in line with the notion that proliferation of differentiated cells is the main mechanism active in the islets [53, 54]. In contrast to endocrine cells, a substantial (10%–60%) percentage of ductal, acinar, and most notably stellate cells showed high stemness probabilities and were predicted positive (Fig. 3A–3E). Importantly, these results were highly consistent across three independent studies (Fig. 3A–3E) [12, 36, 37], and individuals [36] (Supporting Information Fig. S3A, S3B), demonstrating the robustness of our approach to laboratory-induced variability. Our findings are compatible with the increased clonogenic capacity and phenotypic plasticity of these three cell types compared to endocrine cells [55]. They also support the notion that stellate cells are major contributors to pancreatic maintenance and repair [56, 57], as our prediction suggests that the stellate cell fraction contains most SSPCs in this system. Interestingly, both disease (type I and II diabetes, respectively), and age seems to alter the SSPC proportion in the pancreas, with a higher fraction of stem-like cells in affected adults as well as healthy children compared to healthy adults (Fig. 3E). This is consistent with the observation of increased de-differentiation of alpha and beta cells in diabetes patients compared to adult controls [37].

A second system that we examined is the mouse epidermis. Using scRNA-seq of thousands of epidermal cells, our predictor revealed almost twice as many SSPCs among basal cells compared to all other cells (Fig. 3F, left), and interestingly, no difference in the fraction of predicted SSPCs between basal cells expressing at least one of the epidermal stem cell markers *Cd34, Lgr5, Lgr6, Gli1, Lrig1, Krt14,* and those not expressing any (Fig. 3F, right). Surprisingly, when examining which subclass of basal cells is predicted to contain most SSPCs, we found that it is the interfollicular epidermis (IFE), with 7%–13% of stem-like IFE basal cells compared to only ~4% of bulge cells (Fig. 3G). We thus focused on the IFE area and examined where high scoring cells are positioned on the previously estimated differentiation trajectory [35]. We found that these cells are indeed largely located at the root of the trajectory (Fig. 3H). Consistently, our predicted stemness probability was significantly negatively correlated to previously estimated differentiation pseudotime values (Spearman's rho = -0.417, $p < 10^{-5}$). In sum, our analysis supports the notion that there is no specific subcluster of SSPCs in the epidermis [58], as this system appears characterized by

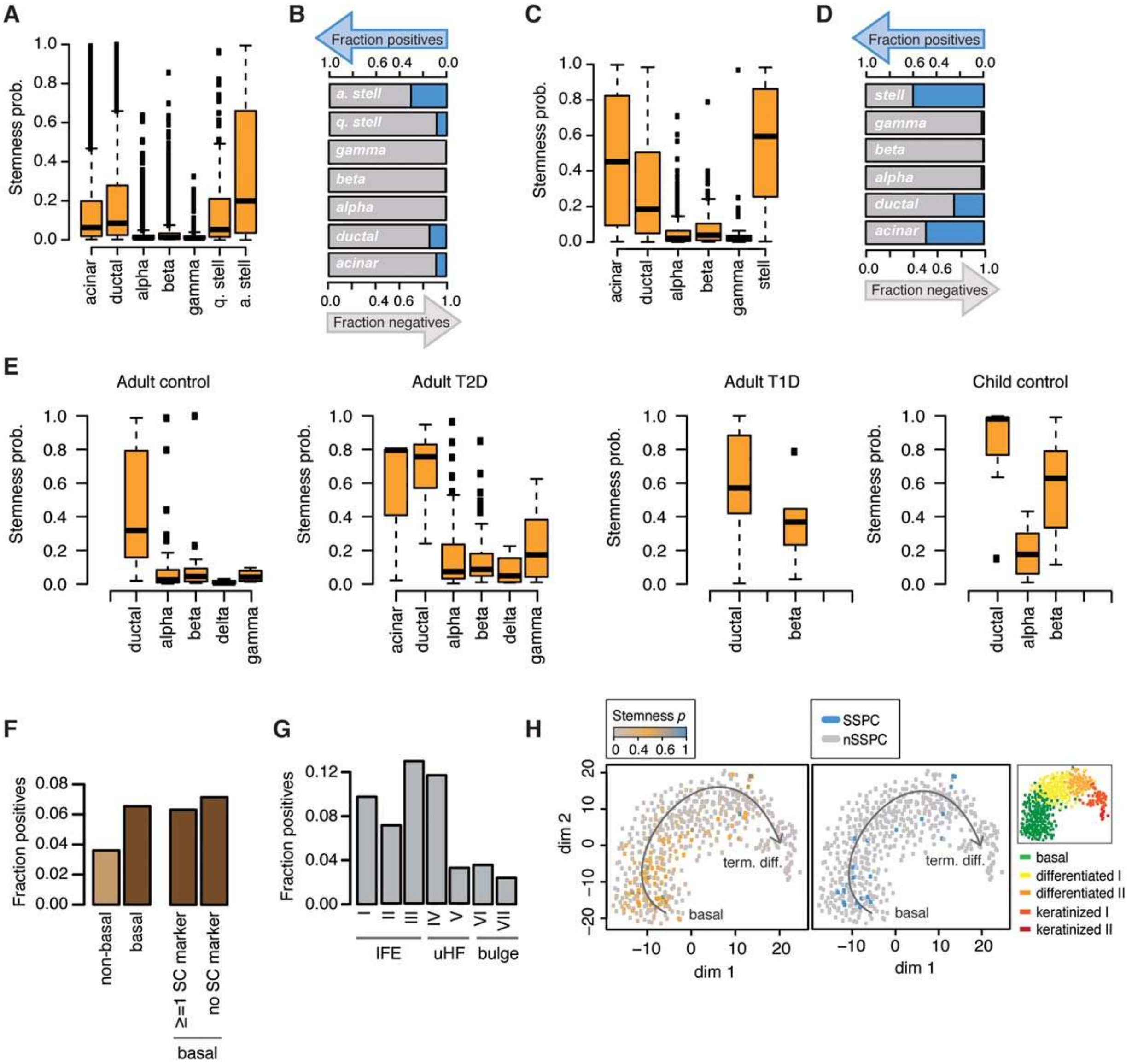


Figure 3. Predicted stem-like cells in human pancreas and mouse epidermis. **(A, C, E)**: The probability of a positive prediction, referred to as "stemness" probability, for endocrine (alpha, beta, gamma, and delta) and exocrine (acinar, ductal, activated/quiescent stellate) cells using three distinct datasets (A: Baron et al. 2016 [36], C: Gruen et al. 2016 [12], and E: Wang et al. 2016 [37]). **(B, D)**: The fraction of cells that were predicted positive (blue) and negative (gray) for each of the cell types in A, C. **(F)**: The fraction of epidermal cells predicted positive among the categories: nonbasal, basal, basal with expression of at least one epidermal stem cell marker (≥1 SC marker), or basal with no expression of any epidermal stem cell markers (no SC marker). **(G)**: The fraction of epidermal basal cells predicted positive among seven distinct previously determined clusters [35]. The anatomical location of cells in each cluster is indicated (IFE–inter-follicular epidermis, uHF–upper hair follicle). **(H)**: t-SNE maps of cells in the IFE, corresponding to a temporal progression from most undifferentiated (basal, green), to terminally differentiated (term. diff., in red) cells [35]. Predicted stemness probabilities (gradient gray to blue) and binary classification outcome (blue: SSPC, gray: non-SSPC) are depicted for all cells. Abbreviations: (n)SSPC, (negative) somatic stem and progenitor cell; SC, stem cell

interspersed, preponderantly basal cells with stem-like expression patterns. We found that many of the latter cells are located in the IFE and the upper hair follicle region, as also suggested recently [35].

Finally, for all predicted SSPCs in both systems, we highlighted their associated genes (Supporting Information Table S2; Fig. S3C, S3D), which can guide follow-up experiments aimed at their isolation and characterization or validate previously reported associations. For example, it has been recently suggested that the TF STMN1 may act as a marker for progenitor-like acinar cells [59]. Our data supports this association given that its expression is highly correlated with acinar SSPCs in both pancreas datasets (Supporting Information Table S2; Fig. S3E, S3F).

Toward a Quantitative Measure of Multipotency

Prediction results on the intestinal, pancreas, and epidermal data suggested that intermediate stemness probabilities may reveal cells in transition stages between multipotent SSPCs and fully differentiated cells (Fig. 2E, 3A–3E, Supporting Information Fig. S2D). To test the capacity of our classier to distinguish among stem-like and differentiated cells, we analyzed

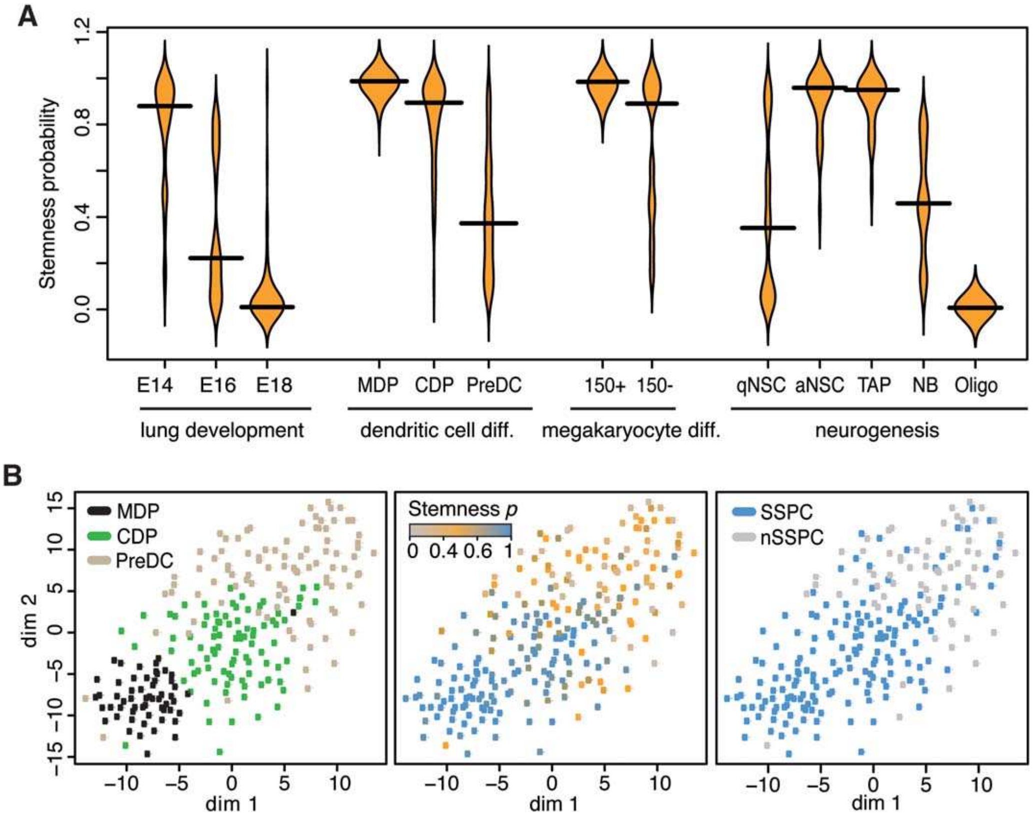


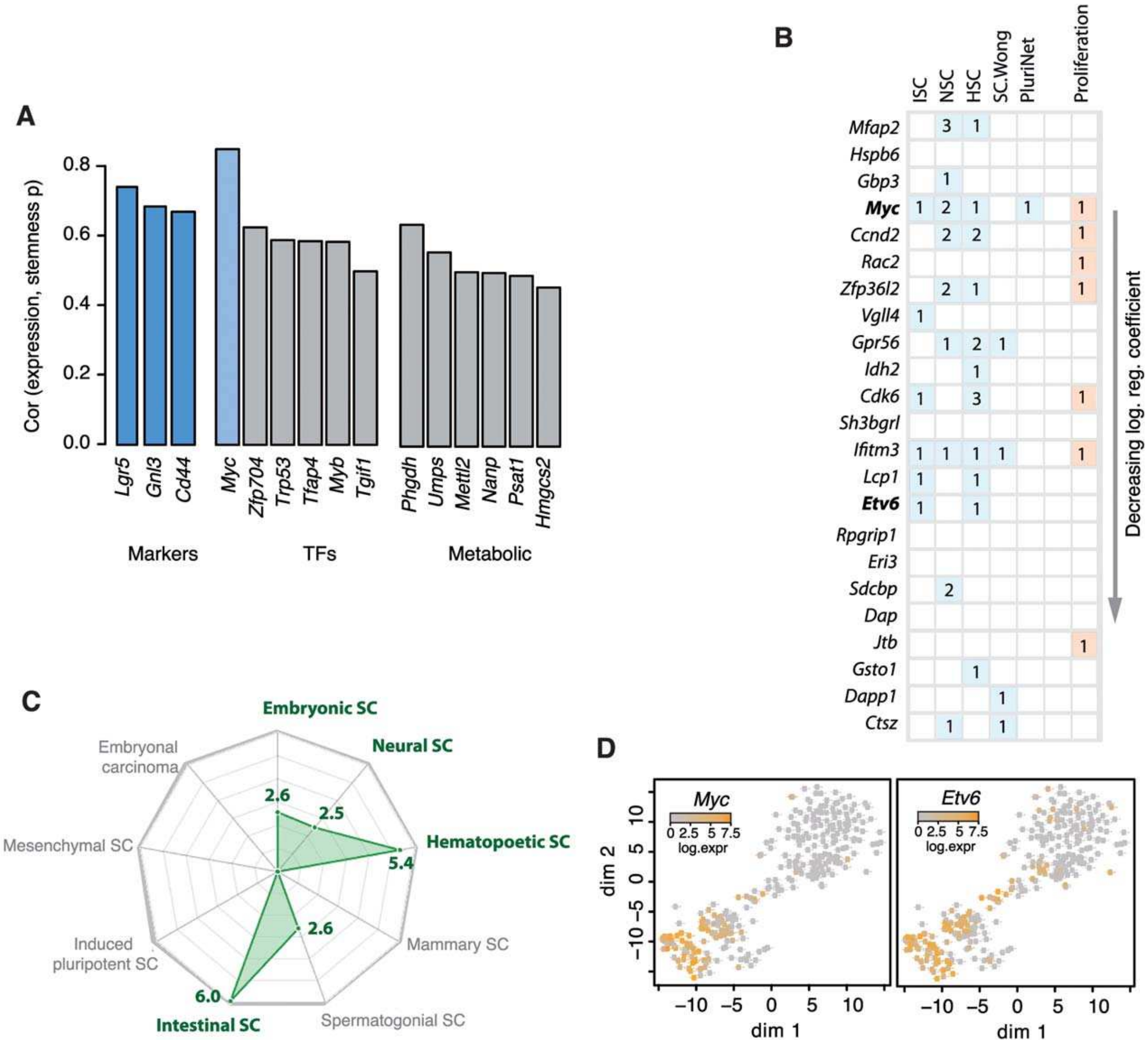
Figure 4. The predicted stemness probability decreases as cells become increasingly differentiated. **(A)**: The probability of a positive prediction, referred to as "stemness" probability, for four groups of cells showing various degrees of lineage commitment and differentiation. E14, MDP, 150+, and NSC (a–activated and q–quiescent) represent the least and E18, PreDCs, 150- and Oligo the most committed/differentiated cell type for each group. Supporting Information Table S1 contains information on the full dataset, including acronym legends. **(B)**: t-SNE map of the complete dendritic cell differentiation dataset [22], including MDP (black), CDP (green), and preDC (light brown). Predicted stemness probabilities (gradient gray to blue) and binary classification outcome (blue: SSPC, gray: non-SSPC) are depicted for all cells. Abbreviations: MDP, macrophage dendritic cell progenitor; CDP, common dendritic cell progenitor; PreDC, dendritic cell precursor; NSC, neural stem cell; TAP transit amplifying cell; NB neuroblast; Oligo, oligodendrocyte; (n)SSPC, (negative) somatic stem and progenitor cell.

four distinct systems featuring cells with increasing level of commitment: (a) lung development [8], (b) dendritic cell development [22], (c) megakaryocyte differentiation [26] and neurogenesis [10] (Fig. 4A, Supporting Information Table S1). In all of these datasets, we observed a progressive decrease in stemness probabilities across time/differentiation (Fig. 4A, Supporting Information Fig. S4A; Table S1). For instance, while the vast majority of multipotent MDPs and CDPs were predicted as positives, the stemness probabilities assigned to CDPs were significantly lower ($p = 10^{-9}$, Wilcoxon rank sum test) and further, the majority (68%) of dendritic cell precursors (PreDCs) were identified as negatives (Fig. 4B, Supporting Information Fig. S4A). These results suggest that the stemness probability provided by our classifier can be used as a quantitative indicator of a cell's progression along a specific differentiation path, independently of the cell's particular cell cycle stage at the time of measurement (Supporting Information Fig. S4B, S4C).

Having established that our derived stemness probability is a robust indicator of a cell's differentiation status, we next asked whether it could be used to functionally associate genes with stemness. The large number of stem and non-stem cells present in the intestinal organoid data [9] enabled us to explore this question in detail. Among genes that showed significant expression correlation with stemness probability and that were largely confined to cells predicted as positives (Methods), we identified the bona fide stemness marker LGR5 as well as other surface-associated stem cell markers such as GNL3 and CD44 (Fig. 5A). A broad range of TFs were also among these genes as well as several metabolic factors such as PHGDH, which is part of the known intestinal SC (ISC) signature [46] (Fig. 5A). In total, we found 98 genes that were ISC-associated (ISC.g3, Supporting Information Table S3), representing known (26% overlapped with an ISC stemness signature) and putatively novel players in ISC function. Gene set enrichment analysis also revealed terms previously associated with stem cell function, including retinoic acid signaling, DNA replication, cell cycle, and proliferation (Supporting Information Fig. S5A).

Signature Genes Are Required for a Stem-Like Phenotype

Given the robust performance of our classifier across biological systems, we reasoned that some of the genes that were retained in our model (Methods and Supporting Information Table S4) are likely to be more generally required for SSPC function. Using the resource collection "StemChecker" [20], we thus asked which of these 23 genes (Fig. 5B) had previously been implicated in stem cell biology. We found that 70% of them



	ISC	NSC	HSC	SC.Wong	PluriNet		Proliferation
Mfap2		3	1				
Hspb6							
Gbp3		1					
Myc	1	2	1		1		1
Ccnd2		2	2				1
Rac2							1
Zfp36l2		2	1				1
Vgll4	1						
Gpr56		1	2	1			
Idh2			1				
Cdk6	1		3				1
Sh3bgrl							
Ifitm3	1	1	1	1			1
Lcp1	1		1				
Etv6	1		1				
Rpgrip1							
Eri3							
Sdcbp		2					
Dap							
Jtb							1
Gsto1			1				
Dapp1				1			
Ctsz		1		1			

Figure 5. Somatic stem and progenitor cell (SSPC)-specific and discriminative genes include both known and novel players in SSPC biology. **(A)**: Correlation (Pearson's *r*) (Cor) between the stemness probabilities (stemness *p*) assigned to intestinal organoid cells [9] and expression of the stem cell markers *Lgr5*, *Gnl3, and Cd44* as well as top ($\geq$0.4) correlated TFs and metabolism-related genes. **(B)**: All genes (sorted in decreasing order of their logistic regression coefficients) retained in the final classifier, and their overlap with population-derived ISC, NSC, HSC, somatic stem cell (SC.Wong), and pluripotency (PluriNet) signatures, as well as their previous implication in proliferation/cell cycle [20]. Numbers correspond to the number of signature sets each gene is part of. **(C)**: Overlap (–10*log_*p*) between the 23 genes retained in the final classifier and population-derived stemness signatures, as visualized by StemChecker [20]. **(D)**: t-SNE map of intestinal epithelial ex vivo and organoid-derived LGR5+ (classically considered as SCs) and LGR5- (assumed to be non-SCs) cells [9], depicting expression of the TFs *Myc* and *Etv6* (gray to orange gradient, orange highest expression). Abbreviations: HSC, hematopoietic stem cell; ISC, intestinal stem cell; NSC, neural stem cell; TF, transcription factor.

were part of at least one population-derived stemness gene set, most often characteristic of neural, hematopoietic, or intestinal SCs (Bonferroni adjusted $p \leq .03$) (Fig. 5B, 5C, Supporting Information Table S4). Thus, while the majority of our classifier genes are individually part of population-derived stemness gene sets, they have not been previously used collectively to discriminate SSPCs from differentiated cells.

Interestingly, 12 out of the 23 classifier genes showed significantly higher expression in ISCs versus all other cells (Supporting Information Fig. S5B; Table S4). This is despite the fact that no ISCs, or any other endodermal SSPCs, have been used for training the model. Among genes with higher expression in ISCs were those that have already been implicated in ISC biology such as *Myc* and *Ccnd2* [60, 61] (Fig. 5D, Supporting Information Fig. S5B). Others have only been partially linked to cell differentiation, and have so far not been implicated in ISC nor SSPC function in general. While a role in ISC function is not an absolute requirement for classifier genes with higher expression in ISCs, it is clearly intuitive. Therefore, we focused on two such genes, *Vgll4* and *Lcp1,* and examined the impact of their shRNA-mediated gene suppression on intestinal organoid growth. We used intestinal organoids from APC^{min} mouse precancerous adenomas (Fig. 6A), since the upregulation of the Wnt signaling pathway induces activation and expansion of LGR5+ ISCs and facilitates the analysis of stem cell phenotypes [62, 63]. We expect that the results obtained from this model can be extrapolated to wild-type ISCs as shown previously [63].

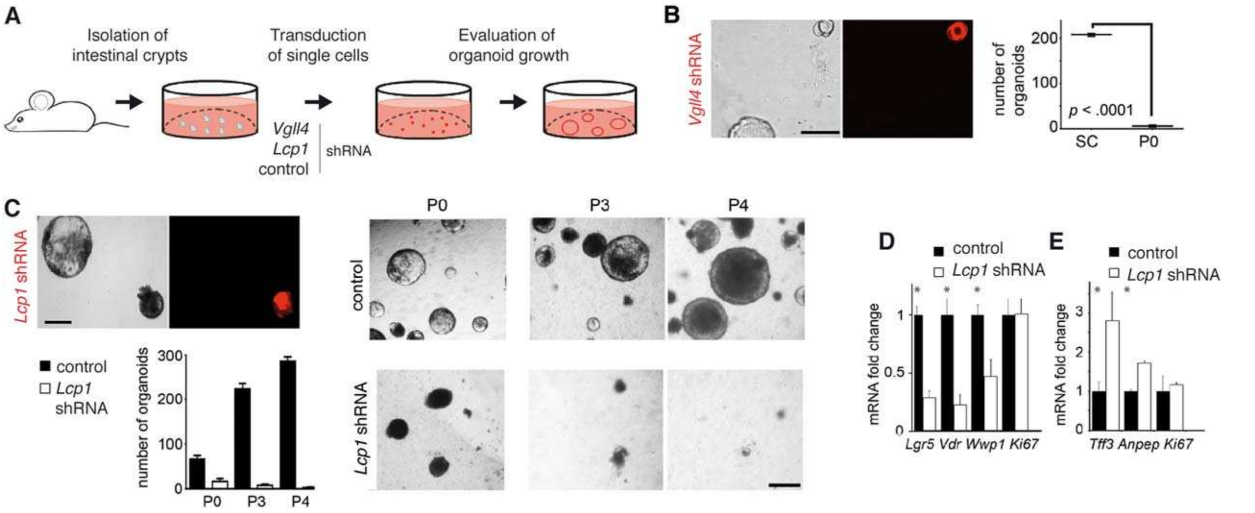


Figure 6. Effect of shRNA-mediated suppression of SSPC-discriminating genes in murine organoids. **(A)**: Scheme of the isolation of intestinal organoids from APC^{min} mice and the transduction with a lentiviral vector allowing shRNA-mediated gene suppression of *Lcp1, Vgll4*, or a control. **(B, C)**: Organoid development is arrested upon *Vgll4* and *Lcp1* shRNA-mediated gene suppression at passage 0. Right (B), quantification of SC and RFP$^+$ organoids expressing the shRNA-*Vgll4*. Images show representative organoids, both RFP$^+$ (and thus expressing the shRNA) and RFP$^-$ (which function as control) (bright field, left and fluorescence image, right). Scale bars: 200 μm. (C) Serial passaging (P0 to P4) of intestinal organoids. Images show that the morphology of shRNA-*Lcp1* expressing organoids was compromised compared to control organoids in the third (P3) and fourth (P4) passage. Scale bar, 200 μm. Bottom left, quantification of RFP$^+$ organoids expressing the shRNA-*Lcp1* over different passages. **(D, E)**: Quantitative Reverse Transcription Polymerase Chain Reaction analyses of gene expression changes upon shRNA-mediated gene suppression of *Lcp1* in intestinal organoids showing downregulation of the stem cell markers *Lgr5* and *Vdr* as well as the potential tumour oncogene *Wwp1* at passage 0 (D), and upregulation of the differentiation markers *Tff3* and *Anpep* at passage 4. (E) The proliferation marker *Ki67* was not significantly altered (D, E). *Gapdh* was used for normalization; *, $p < .05$. Abbreviation: SC, single cell.

We first demonstrated successful shRNA-mediated gene suppression of the two genes in the murine CT26 colon carcinoma cell line and intestinal organoids (Fig. S6A). While shRNA-mediated gene suppression in CT26 cells did not affect cell phenotype or proliferation per se (data not shown), we observed a striking reduction in the number and size of intestinal organoids in the gene suppression samples (Fig. 6B, 6C, Supporting Information Fig. S6B–S6D). Strikingly, organoid growth after *Vgll4* gene suppression was completely impaired, such that organoids could not be passaged and RNA could not be collected for further experiments. For *Lcp1*, the gene suppression effects were observed both at initial and subsequent passages (Fig. 6C, Supporting Information Fig. S6B–S6D). Importantly, we also found that canonical stemness marker genes such as *Lgr5*, *Vdr*, and the potential oncogene *Wwp1* [64], were significantly downregulated at passage 0 when *Lcp1* was suppressed, while *Ki67*, a proliferation marker, showed no significant difference (Fig. 6D, 6E). Further passaging of these organoids (passage 3 and 4), showed that these shRNA-mediated gene suppressions increased differentiation markers such as *Anpep* (absorptive lineage) and *Tff3* (secretory lineage).

Together, these results demonstrate for the first time the requirement of two classifier genes, *Lcp1* and *Vgll4*, in APC^{min} intestinal organoid growth and propagation. However, follow-up studies will be required to consolidate the role of these genes in stem cell function in general throughout many tissues.

Discussion

Despite substantial progress in understanding SSPC function and their essential role in maintaining and repairing mammalian tissues, it is currently unknown how many kinds of distinct SSPCs exist, which tissues harbor them and what their molecular and functional similarities and differences are [4, 5]. Recent technological developments in the field of scRNA-seq now facilitate analyses of unprecedented molecular resolution, which enabled us to develop a universally applicable method for the de novo identification of self-renewing and multipotent cells. The timeliness of our study is demonstrated by the fact that numerous SSPC and non-SSPC scRNA-seq datasets have only very recently become available, enabling for the first time cross-system analyses. Nevertheless, we acknowledge that we were still constrained in the choice of training and test sets, leading to overrepresentation of certain lineages (hematopoietic) and germ layers (mesoderm) in our analyses as well as the lack of distinction between true stem cells and multipotent progenitor cells. Ideally, all germ layers and SSPC types (e.g., quiescent vs. active) should be equally well represented in both training and test sets to derive the best classifier. Therefore, when more datasets will become available, one can envisage updating and further scaling up the binary classification approach that we used here to a multi-class system. The underlying aim would be to increase the discriminative power of our approach by distinguishing activated and quiescent somatic SCs, multipotent and unipotent progenitors, for instance. This will in turn allow the exploration of specific transcriptional properties for each of the individual stages as well as detailed delineation of commonalities and differences between SSPCs.

A further limitation of our approach is the relative simplicity of the used data normalization (quantiles), which dampens the large dynamic expression range that is a powerful feature of RNA-seq. We took this route given the large technical and biological noise that is currently characterizing scRNA-seq data, likely exacerbated by working across experimental and data analysis protocols [65]. We anticipate that, with the maturation of experimental and computational scRNA-seq techniques, the actual distribution of expression estimates will be taken into account, such that the full dynamic range of this powerful method can be exploited.

Despite these limitations, we showed here that our SSPC identification method can be applied to previously completely unseen scRNA-seq datasets, across different technologies (e.g., CEL-seq and SMART-seq) [66, 67] and experimental systems, including highly proliferating (e.g., ISCs) and quiescent (long-term HSCs) SSPCs. Consequently, we envision several downstream applications for our classifier. First, it may allow the detection of stem-like cells in underexplored, heterogeneous tissues, with the aim of identifying specifically expressed marker genes to enable downstream phenotypic analyses, as shown here for the mouse epidermis and the human pancreas. Indeed, while the markers that we propose in these two systems will require further validation, we also demonstrated the feasibility of marker detection given the de novo identification of the established markers LGR5 and GNL3 in intestinal epithelial organoids [62, 68]. Second, such a classifier may support estimating the fraction of stem-like cells in a heterogeneous mix of cells. This may provide novel insights into tissue organization, as we also illustrated here for the epidermis and pancreas, or into a tissue's or cell fraction's regenerative capacity, for example over the course of aging [30], or to inform cell transplantation experiments [69]. A third application is the detection of highly malignant cells in tumors. That scRNA-seq data can be used in the context of tumor heterogeneity to reveal stemness-exhibiting cells has in fact recently been demonstrated in glioma samples [70]. While these latter applications go beyond the scope of the current study, we envisage that they will constitute very fruitful research areas in the future.

Conclusion

We developed a sc RNA-seq-based classifier that accurately detects SSPCs across tissues, aiding in the resolution of still poorly characterized systems. In addition, we demonstrate that this classifier can provide a quantitative measure of a cell's progression along a differentiation path. The method's good cross-system performance supports the notion that SSPCs can be recognized based on shared, but not necessarily identical molecular properties. Specifically, genes with high discriminative power have putative roles in SSPC biology, of which many are known such as those of *Myc* and *Ccnd2*, but others are novel such as *Vgll4* and *Lcp1*'s function in the intestinal system.

Availability

We provide the trained model as well as the two best-performing alternative models (elastic-net logistic regression and random forest), training and test data as well as an R script to produce the predictions discussed in this manuscript and use on novel data. Available at https://github.com/DeplanckeLab/SSPCI. Predictions can also be obtained and visualized through the web-based platform ASAP at https://asap.epfl.ch/[71].

Acknowledgments

We thank Robert Baldwin for constructive discussions as well as Vincent Gardeux, Angela Goncalves, Ernest Turro, Rachana Pradhan, Maroun Bou Sleiman, Andre Faure, Paul Flicek, and Benita Wolf for helpful suggestions and carefully reading of the manuscript. We also thank all authors that dully deposited and annotated the data associated with their publications, and in particular Maria Kasper, Simon Joost, Dominic Gruen, Klaus Kaestner, Dana Tzfati, and Jonathan Schug for providing crucial details and feedback on the original data and the results of our analyses. This research was supported by the Human Frontier Science Program LT001032/2013 (to PCS), by the Swiss National Science Foundation Grant (31003A_162735), and by institutional support from the Swiss Federal Institute of Technology in Lausanne (EPFL).

Author Contributions

P.C.S.: conception and design, data analysis and interpretation, manuscript writing, conducted all scRNA-seq related analyses; P.O.M.: performed all experiments, analyzed all data on intestinal organoids; J.H.: supervised the intestinal organoids experiments, data evaluation and financial support; B.D.: conception and design, data analysis and interpretation, manuscript writing, financial support; P.C.S., P.O.M., J.H., and B.D.: read and approved the final manuscript.

Disclosure of Potential Conflicts of Interest

The authors indicated no potential conflicts of interest.

References

1 Mimeault M, Batra SK. Recent progress on tissue-resident adult stem cell biology and their therapeutic implications. Stem Cell Rev 2008;4:27–49.

2 Wang A, Chen L, Li C et al. Heterogeneity in cancer stem cells. Cancer Lett 2015; 357:63–68.

3 Eckfeldt CE, Mendenhall EM, Verfaillie CM. The molecular repertoire of the "almighty" stem cell. Nat Rev Mol Cell Biol 2005;6:726–737.

4 Clevers H. What is an adult stem cell? Science 2015;350:1319–1320.

5 Tajbakhsh S. Stem cell: What's in a name? Nat Reports Stem Cells 2009. 10.1038/stemcells.2009.90

6 Hoppe PS, Coutu DL, Schroeder T. Single-cell technologies sharpen up mammalian stem cell research. Nat Cell Biol 2014;16:919–927.

7 Saliba A-E, Westermann AJ, Gorski SA et al. Single-cell RNA-seq: Advances and future challenges. Nucleic Acids Res 2014;42: 8845–8860.

8 Treutlein B, Brownfield DG, Wu AR et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 2014;509:371–375.

9 Grün D, Lyubimova A, Kester L et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature 2015;525: 251–255.

10 Llorens-Bobadilla E, Zhao S, Baser A et al. Single-cell transcriptomics reveals a population of dormant neural stem cells that become activated upon brain injury. Cell Stem Cell 2015;17:329–340.

11 Trapnell C, Cacchiarelli D, Grimsby J et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 2014;32:381–386.

12 Grün D, Muraro M, Boisset J-C et al. De novo prediction of stem cell identity using single-cell transcriptome data. Cell Stem Cell 2016;19:266–277.

13 Ramalho-Santos M, Yoon S, Matsuzaki Y et al. "Stemness": Transcriptional profiling of embryonic and adult stem cells. Science 2002;298:597–600.

14 Ivanova NB, Dimos JT, Schaniel C et al. A stem cell molecular signature. Science 2002;298:601–604.

15 Evsikov AV, Solter D. Comment on "'Stemness': Transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature". Science 2003;302: 393.

16 Zipori D. The nature of stem cells: State rather than entity. Nat Rev Genet 2004;5: 873–878.

17 Müller F-J, Schuldt BM, Williams R et al. A bioinformatic assay for pluripotency in human cells. Nat Methods 2011;8:315–317.

18 Scheubert L, Schmidt R, Repsilber D et al. Learning biomarkers of pluripotent stem cells in mouse. DNA Res 2011;18:233–251.

19 Noh M, Smith JL, Huh YH et al. A resource for discovering specific and universal biomarkers for distributed stem cells. PLoS One 2011;6:e22077.

20 Pinto JP, Kalathur RK, Oliveira DV et al. StemChecker: A web-based tool to discover and explore stemness signatures in gene sets. Nucleic Acids Res 2015;43:W72–W77.

21 Leychkis Y, Munzer SR, Richardson JL. What is stemness?. Stud Hist Philos Biol Biomed Sci 2009;40:312–320.

22 Schlitzer A, Sivakamasundari V, Chen J et al. Identification of cDC1- and cDC2-committed DC progenitors reveals early lineage priming at the common DC progenitor stage in the bone marrow. Nat Immunol 2015;16:718–728.

23 Zeisel A, Munoz-Manchado AB, Codeluppi S et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 2015;347:1138–1142.

24 Liu N, Liu L, Pan X. Single-cell analysis of the transcriptome and its application in the characterization of stem cells and early embryos. Cell Mol Life Sci 2014;71:2707–2715.

25 Scialdone A, Natarajan KN, Saraiva LR et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. Methods 2015;85:54–61.

26 Haas S, Hansson J, Klimmeck D et al. Inflammation-induced emergency megakaryopoiesis driven by hematopoietic stem cell-like megakaryocyte progenitors. Cell Stem Cell 2015;17:422–434.

27 Wilson NK, Kent DG, Buettner F et al. Combined single-cell functional and gene expression analysis resolves heterogeneity within stem cell populations. Cell Stem Cell 2015;16:712–724.

28 Shin J, Berg DA, Zhu Y et al. Single-cell RNA-Seq with waterfall reveals molecular cascades underlying adult neurogenesis. Cell Stem Cell 2015;17:360–372.

29 Freeman BT, Kouris NA, Ogle BM. Tracking fusion of human mesenchymal stem cells after transplantation to the heart. Stem Cells Translational Medicine 2015;4:685–694.

30 Kowalczyk MS, Tirosh I, Heckl D et al. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res 2015;25:1860–1872.

31 Buettner F, Natarajan KN, Casale FP et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol 2015;33:155–160.

32 Marinov GK, Williams BA, McCue K et al. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. Genome Res 2014;24:496–510.

33 Usoskin D, Furlan A, Islam S et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat Neurosci 2015;18:145–153.

34 Camp JG, Sekine K, Gerber T et al. Multilineage communication regulates human liver bud development from pluripotency. Nature 2017;546:533–538.

35 Joost S, Zeisel A, Jacob T et al. Single-cell transcriptomics reveals that differentiation and spatial signatures shape epidermal and hair follicle heterogeneity. Cell Syst 2016;3:221–237.e9.

36 Baron M, Veres A, Wolock S et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. Cell Syst 2016;3: 346–360.e4.

37 Wang YJ, Schug J, Won K-J et al. Single cell transcriptomics of the human endocrine pancreas. Diabetes 2016;65:3028–3038.

38 Barrett T, Wilhite SE, Ledoux P et al. NCBI GEO: Archive for functional genomics data sets–update. Nucleic Acids Res 2013;41: D991–D995.

39 Van Der Maaten LJP, Hinton GE. Visualizing data using t-SNE. J Mach Learn Res 2008; 9:2579–2605.

40 Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33: 1–22.

41 Breiman and Cutler's random forests for classification and regression. Available at https://cran.r-project.org/web/packages/randomForest/randomForest.pdf. Accessed June 26, 2017.

42 Müller F-J, Laurent LC, Kostka D et al. Regulatory networks define phenotypic classes of human stem cell lines. Nature 2008; 455:401–405.

43 Wong DJ, Liu H, Ridky TW et al. Module map of stem cell genes guides creation of epithelial cancer stem cells. Cell Stem Cell 2008;2:333–344.

44 Sing T, Sander O, Beerenwinkel N et al. ROCR: Visualizing the performance of scoring classifiers. R Packag Version 2005;21:3940–3941.

45 Kampstra P. Beanplot: A boxplot alternative for visual comparison of distributions. J Stat Softw 2008;28:1–9.

46 Muñoz J, Stange DE, Schepers AG et al. The Lgr5 intestinal stem cell signature: Robust expression of proposed quiescent "+4" cell markers. Embo J 2012;31:3079–3091.

47 Shibata H, Toyama K, Shioya H et al. Rapid colorectal adenoma formation initiated by conditional targeting of the Apc gene. Science 1997;278:120–123.

48 Sato T, Vries RG, Snippert HJ et al. Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. Nature 2009;459:262–265.

49 Ootani A, Li X, Sangiorgi E et al. Sustained in vitro intestinal epithelial culture within a Wnt-dependent stem cell niche. Nat Med 2009;15:701–706.

50 Koo B-K, Stange DE, Sato T et al. Controlled gene expression in primary Lgr5 organoid cultures. Nat Methods 2012;9:81–83.

51 Takeda N, Jain R, LeBoeuf MR et al. Interconversion between intestinal stem cell populations in distinct niches. Science 2011; 334:1420–1424.

52 Tsai RYL. Turning a new page on nucleostemin and self-renewal. J Cell Sci 2014;127: 3885–3891.

53 Dor Y, Brown J, Martinez OI et al. Adult pancreatic beta-cells are formed by self-duplication rather than stem-cell differentiation. Nature 2004;429:41–46.

54 Kopp JL, Grompe M, Sander M. Stem cells versus plasticity in liver and pancreas regeneration. Nat Cell Biol 2016;18:238–245.

55 Puri S, Folias AE, Hebrok M. Plasticity and dedifferentiation within the pancreas: Development, homeostasis, and disease. Cell Stem Cell 2015;16:18–31.

56 Apte MV, Pirola RC Wilson JS. Pancreatic stellate cells: A starring role in normal and diseased pancreas. Frontiers in Physiology 2012;28:344.

57 Mato E, Lucas M, Petriz J et al. Identification of a pancreatic stellate cell population with properties of progenitor cells: New role for stellate cells in the pancreas. Biochem J 2009;421:181–191.

58 Schepeler T, Page ME, Jensen KB. Heterogeneity and plasticity of epidermal stem cells. Development 2014;141:2559–2567.

59 Wollny D, Zhao S, Everlien I et al. Single-cell analysis uncovers clonal acinar cell heterogeneity in the adult pancreas. Dev Cell 2016;39:289–301.

60 Muncan V, Sansom OJ, Tertoolen L et al. Rapid loss of intestinal crypts upon conditional deletion of the Wnt/Tcf-4 target gene c-Myc. Mol Cell Biol 2006;26:8418–8426.

61 Cole AM, Myant K, Reed KR et al. Cyclin D2 - Cyclin-dependent kinase 4/6 is required for efficient proliferation and tumorigenesis following Apc loss. Cancer Res 2010;70:8149–8158.

62 Barker N, van Es JH, Kuipers J et al. Identification of stem cells in small intestine and colon by marker gene Lgr5. Nature 2007; 449:1003–1007.

63 Ordonez-Moran P, Dafflon C, Imajo M et al. HOXA5 counteracts stem cell traits by inhibiting Wnt signaling in colorectal cancer. Cancer Cell 2015;28:815–829.

64 Zhang L, Wu Z, Ma Z et al. WWP1 as a potential tumor oncogene regulates PTEN-Akt signaling pathway in human gastric carcinoma. Tumor Biol 2015;36:787–798.

65 Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet 2015;16:133–145.

66 Hashimshony T, Wagner F, Sher N et al. CEL-Seq: Single-cell RNA-Seq by multiplexed linear amplification. Cell Rep 2012;2:666–673.

67 Ramsköld D, Luo S, Wang Y-C et al. Full-length mRNA-Seq from single-cell levels of

RNA and individual circulating tumor cells. Nat Biotechnol 2012;30:777–782.

68 Powell AE, Wang Y, Li Y et al. The pan-ErbB negative regulator lrig1 is an intestinal stem cell marker that functions as a tumor suppressor. Cell 2012;149: 146–158.

69 Huh YH, Noh M, Burden FR et al. Sparse feature selection identifies H2A.Z as a novel, pattern-specific biomarker for asymmetrically self-renewing distributed stem cells. Stem Cell Res 2015;14:144–154.

70 Patel AP, Tirosh I, Trombetta JJ et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 2014;344:1396–1401.

71 Gardeux V, David FPA, Shajkofci A et al. ASAP: A web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. Bioinformatics 2017;33:3123–3125.

See www.StemCells.com for supporting information available online.