cmgh
CELLULAR AND MOLECULAR GASTROENTEROLOGY AND HEPATOLOGY

REVIEW

Single-Cell Computational Strategies for Lineage Reconstruction in Tissue Systems

Charles A. Herring,[1,2] Bob Chen,[1,3] Eliot T. McKinley,[1,4] and Ken S. Lau[1,2,3]

[1]Epithelial Biology Center, [4]Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee; [2]Program in Chemical and Physical Biology, [3]Department of Cell and Developmental Biology, Vanderbilt University School of Medicine, Nashville, Tennessee

SUMMARY

Recent developments in single-cell technologies have stimulated growth in analysis techniques, in particular, computational tools for ordering cell states as a function of pseudotemporal progression. We provide a review of current algorithms and a generalized single-cell workflow tailored for trajectory analysis, with a focus on underlying assumptions and caveats.

Function at the organ level manifests itself from a heterogeneous collection of cell types. Cellular heterogeneity emerges from developmental processes by which multipotent progenitor cells make fate decisions and transition to specific cell types through intermediate cell states. Although genetic experimental strategies such as lineage tracing have provided insights into cell lineages, recent developments in single-cell technologies have greatly increased our ability to interrogate distinct cell types, as well as transitional cell states in tissue systems. From single-cell data that describe these intermediate cell states, computational tools have been developed to reconstruct cell-state transition trajectories that model cell developmental processes. These algorithms, although powerful, are still in their infancy, and attention must be paid to their strengths and weaknesses when they are used. Here, we review some of these tools, also referred to as *pseudotemporal ordering algorithms*, and their associated assumptions and caveats. We hope to provide a rational and generalizable workflow for single-cell trajectory analysis that is intuitive for experimental biologists. *(Cell Mol Gastroenterol Hepatol 2018;5:539–548; https://doi.org/10.1016/j.jcmgh.2018.01.023)*

Keywords: Trajectory; Pseudotime; Single-Cell Analysis; Differentiation; Cell State Transition; Stem Cells.

Cellular heterogeneity, defined by a diversity of co-occurring cell types in a tissue, is characteristic of practically every organ in the human body. The organs of the digestive system also comprise specialized cell populations that play important but diverse roles in absorption, secretion, and barrier function. For instance, distinct cell types of the pancreatic islet secrete different hormones, including insulin-secreting β cells, glucagon-secreting δ cells, and somatostatin-expressing δ cells.[1] Likewise, the small and large intestines exist in a dynamic equilibrium of heterogeneous stem, transitional, and differentiated cell populations, with the latter responsible for nutrient absorption, antimicrobial peptide secretion, and formation and maintenance of the mucus layer in the gut.[2] A fundamental question in developmental biology is the origin of cellular heterogeneity, which arises from a specification process initiated from multipotent cells. Recent developments in multiplex single-cell experimental tools have greatly facilitated the interrogation of individual cells; data on single cells then can be grouped into relevant cell populations. In digestive organ systems, populational analysis of single-cell data has been used for discovering previously unidentified β cell subpopulations in the pancreatic islet,[1] novel markers of intestinal tuft cells,[2,3] endocrine progenitor cell heterogeneity,[4] and signaling mechanisms between neighboring intestinal epithelial cells,[5] among others. Populational analysis using single-cell tools is a powerful approach for dissecting tissue-level heterogeneity, and has been reviewed extensively elsewhere.[6,7] Beyond defining cell populations, such as stem and differentiated cell types, single-cell experimental tools also can be used to characterize transitional intermediate cell states in various tissues and organoid systems.[8] Thus, it theoretically should be possible, using single-cell data, to trace terminal cell types through intermediate cell states back to their roots of differentiation in a series of progenitor–progeny relationships. Here, we review current computational tools by which a "virtual lineage trace," also known as a *pseudotemporal order*, can be extracted from multidimensional single-cell data.

Single-Cell Experimental Technologies to Interrogate Cell States From Tissues

The theoretical basis of pseudotemporal ordering is that asynchronous sampling from multiple time points over development[9] or snap-shot sampling at a single time point

Abbreviations used in this paper: MST, minimum spanning tree; PCA, principal component analysis; scRNA-seq, single-cell RNA-sequencing; t-SNE, t-distributed stochastic neighbor embedding.

Most current article

2352-345X
https://doi.org/10.1016/j.jcmgh.2018.01.023

of a continually renewing tissue (such as the intestine)[10] can result in a dense sampling of transitional states that can be aligned to reflect a time course of state transitions (Figure 1). A cell state is represented by the position of a cell in a data space defined by multiple molecular markers that describe the identity and behavior of cells (Figure 1). Ordering is conducted on the basis of similarity in cell states; dense sampling of these states is required to obtain a continuum of data by which the relationship between cell states can be inferred. Because transitional cell states often are rare compared with differentiated cells in tissue, it is required for single-cell technologies to be able to query a large volume of data points, as well as simultaneously measure multiple markers, to fully depict a continuum of cell states. Here, we briefly review common single-cell tools that can evaluate many cells in a multiplex fashion in the context of their classification into either suspension approaches or in situ approaches.

Suspension approaches involve cellular dissociation and then separate processing and analysis of individual cells, with the major caveat that the spatial context of the tissue is lost. Suspension approaches include protein-based techniques such as mass and multiparameter flow cytometry, and transcript-based techniques such as single-cell RNA-sequencing (scRNA-seq) and gene expression assays.[11] The advantage of these approaches is in their high-throughput capacity to produce data. Flow and mass cytometry can analyze hundreds of thousands of cells in a multiplex fashion (20–40 protein analytes per cell) on the order of minutes,[12] while scRNA-seq can quantify gene expression in an unbiased, genome-wide manner (thousands of gene analytes).[13] Multiple platforms of scRNA-seq exist, with variations in cell-containment strategies ranging from microwells[14–16] to liquid-oil emulsion droplets[17–19]; many of the current iterations can query up to thousands of cells.

A factor to consider when applying suspension approaches, especially on organs of the digestive system, is the perturbation imposed on cells when they are disaggregated from tissue. Cells of the hematopoietic system exist either as single-cell suspensions or in loosely connected tissues, which are readily amenable to single-cell analysis.[12] For intestinal cells, specifically for those in the lamina propria, protocols have been developed such that the correct numbers and types of cells can be retrieved for single-cell analysis, providing critical insights into biological and disease processes.[20] For epithelial tissues that are tightly connected, additional factors must be considered so as to not introduce technical artifacts during the single-cell dissociation process.[21] Disaggregation for Intracellular Signaling of Single Epithelial Cells from Tissue was developed as a fixation approach for preserving the intact state of epithelial cells for single-cell signaling analysis using mass and flow cytometry.[5] Disaggregation for Intracellular Signaling of Single Epithelial Cells from Tissue can be applied to formalin-fixed paraffin-embedded tissues, for instance, to observe signaling state alterations in human colorectal cancer specimens.[22] On the scRNA-seq side, Adam et al[23] adapted psychrophilic proteases for single-cell dissociation in the cold, which drastically reduces artifacts and maintains native cell states. Adaptation of a similar strategy to fixed[24,25] or frozen tissues[26] may enable scRNA-seq of preserved cell states. For cells that cannot be dissociated without compromising integrity, such as neurons with long and fragile axonal processes, single nucleus profiling of fresh and preserved tissues is a viable strategy to obtain a glimpse of cell state.[27–29] It should be noted, however, that transcriptomes obtained from the nucleus may be drastically different from those obtained from the entire cell. These efforts highlight recent developments into suspension approaches to enable high-throughput evaluation of native cell states for characterizing cellular heterogeneity and developmental events.

Unlike suspension approaches, in situ imaging techniques allow cells and their niche components to be analyzed in their native spatial context. Because of the lack of tissue dissociation, communication mechanisms between niche cells and epithelial cells can be directly visualized and quantified. Recent advances have improved the multiplex capabilities of microscopy approaches, enabling detection and quantification of dozens of markers leading to accurate identification of cell types that reside within certain niches. Current multiplex imaging technologies for proteins can be classified either as mass-based or iterative. Mass-based imaging approaches,

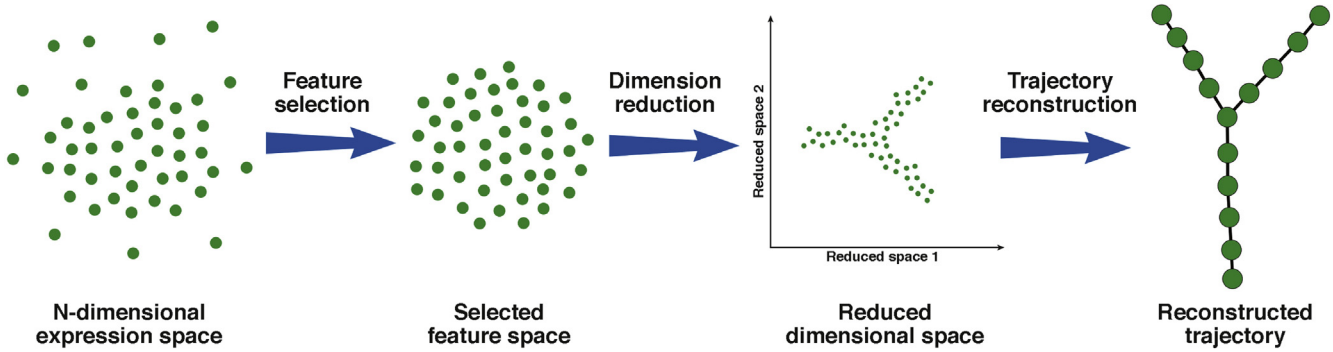

Figure 1. General workflow of trajectory analysis algorithms. Beginning with data in multidimensional space, feature selection is first performed to include relevant analytes and exclude noise. From the selected feature set, dimension reduction is applied to best emphasize the part of the data most relevant to cell-state transitions. Trajectories then are reconstructed in this reduced space and analyzed as pseudotime courses.

including imaging mass cytometry[30] and multiplex ion beam imaging,[31] rely on metal-tagged antibodies coupled to mass spectrometry, while iterative approaches, including Multiplex Immunofluorescence,[32] Cyclic Immunofluorescence,[33] and others,[34–36] rely on cycles of staining, imaging, and destaining to enable multiplexity. Iterative approaches also are available for in situ gene expression analysis, including iterative RNA fluorescence in situ hybridization[37,38] and in situ sequencing.[39] More recent work also has combined the barcoding capabilities of nucleic acids with protein antibody staining in an approach called *nucleic acid exchange* to achieve higher levels of efficiency and multiplexity.[40] Microscopy approaches also can query thousands of cells if whole tissues are imaged at the appropriate resolution, although the time for acquisition of such data sets can be large on a per-sample basis.[3] Currently, multiplex approaches are developed for 2-dimensional imaging, but future efforts may combine tissue clearing[41–43] along with intravital techniques[44] to enable 3-dimensional imaging of cells in real time. Although a variety of techniques can generate intricate multiplex images of intact tissue, challenges in the automatic identification of objects hinder quantitative analysis of spatial relationships among cells and niche components. Although these tools are in their infancy, in situ multiplex approaches hold the promise for understanding cell-to-environment interactions in the context of cell-state transitions.

The choice of suspension or in situ techniques is highly dependent on the experimental question being sought and oftentimes can be complementary. Suspension approaches are much higher throughput in terms of the number of cells and analytes analyzed, whereas in situ techniques can afford spatial resolution. We have previously coupled the 2 classes of tools, using suspension-based signaling analysis and in situ microscopy to define neighbor cell signaling mechanisms.[5] An integrative strategy of using suspension-based analysis to deeply profile cell populations and in situ approaches to define spatial relationships between identified populations is one of many powerful strategies for delineating functionally meaningful relationships in tissue systems.

Feature Selection: A Preprocessing Step for Trajectory Analysis of scRNA-Seq Data

Multiplex cytometry and scRNA-seq techniques both attempt to capture extremely complex cell states in the form of high-dimensional data, in proteomic or transcriptomic spaces, respectively. scRNA-seq is known to produce noisy data on a per-feature basis, especially for lowly expressed genes, owing to the processing and amplification of small amounts of nucleic acids[16] and the biological phenomenon of bursting transcription.[45] The effects of noise are compounded in multidimensional space in a phenomenon known as the *curse of dimensionality*,[46] which greatly affects downstream trajectory analysis when using the full ensemble of features. A way to mitigate this effect is to select and analyze only a subset of the most important features that maximally captures the phenomenon of interest, while ignoring uninformative or noisy features. The feature selection step is implicitly performed in candidate-based approaches, such as Cytometry Time-of-Flight and multiplex microscopy, because the user is picking the most important markers to measure. How to pick informative features while eliminating uninformative ones from genome-scale scRNA-seq experiments is still an active area of research.

One intuitive method for feature selection is a supervised approach that only includes genes of interest. For instance, candidate genes can be selected from a differentially expressed gene set from a bulk RNA-seq experiment that uses a time course or genetic perturbation experimental design. Pipelines such as Single-cell Topological Data Analysis and Single Cell Lineage Inference Using Cell Expression Similarity and Entropy incorporate annotated gene sets from gene ontology resources such as Protein ANalysis THrough Evolutionary Relationships or the Database for Annotation, Visualization and Integrated Discovery to select features in a semi-supervised fashion.[47,48] For studies with minimal or unreliable prior knowledge, completely unsupervised methods that leverage general gene expression patterns may be used.

Different unsupervised feature selection methods vary in their assumptions as well as complexity. For example, a commonly used method in analyzing scRNA-seq data involves identifying transcriptomic features with highly variable expression across the entire data set of single cells. Here, the assumption is that variance in gene expression between cells corresponds to meaningful gene regulation. This method calculates the variance of each gene across all data points (cells), and filters the features to capture only those with the highest variances.[49] In a way, this method is analogous to principal component analysis (PCA) in selecting the dimensions with the highest variances.[50] Technical variation can potentially exceed meaningful biological variation, and filtering methods can be confounded by the simultaneous occurrence of these 2 sources of variation. However, because of their computational tractability, variance ranking methods can provide a quick evaluation of data quality by enumerating the number of biologically relevant genes returned, which can be collected to potentially reveal both known and unknown cellular relationships.

More sophisticated methods based on different patterns of gene expression have been developed to identify biologically relevant features. Qui et al[51] developed dpFeature, a method that selects differentially expressed genes between cell populations described by unsupervised clustering for downstream trajectory analysis. Clusters of cells automatically identified are representative of distinct cell states, and differentially expressed genes represent likely regulators of these states. However, data sets that depict transitions are generally continuously distributed and do not form distinct clusters. Clustering in these cases are based on arbitrary cut-off values, and, thus, how dpFeature performs on these types of data sets remains to be tested.

To handle continuous data distributions, Welch et al[52] developed a metric called *neighborhood variance*. Implementing a K-nearest neighbors graph approach with each

cell represented as a node, this method defines neighborhoods of locally varying cell states. Variance of a feature is analyzed over each defined neighborhood and compared with the global variance of that feature, with a threshold of selection for downstream analysis. Selected features exhibit small local variance with gradual and monotonic changes, consistent with progressively transitioning cell states. In addition, Furchtgott et al[53] developed a Bayesian approach for identifying subsets of gene expression patterns over 3 cell states that are useful for defining lineage relationships. These feature selection methods use unique patterns of gene expression present in single-cell data sets to filter out genes whose variances are either owing to noise or are irrelevant to the phenomenon of interest. More refined gene expression patterns perhaps can be identified in the future for more sophisticated feature selection.

t-Distributed Stochastic Neighbor Embedding: A Technique for Cell Population Analysis

A challenge of the analysis of highly multiplexed single-cell data is the inherent difficulty of visualizing high-dimensional data spaces (Figure 1). Thus, multiple methods, such as PCA, have been developed to represent high-dimensional data in a lower-dimensional space while best retaining the underlying relationships among data points in the original data space.[50] In principle, cell-state transition relationships, based on a continuum of similar states, can be visualized in 2- or 3-dimensional space given the correct information within the data is retained. In practice, however, all dimensionality reduction techniques result in information loss because some parts of the data are discarded for lower-dimension representations. For instance, PCA represents high-dimensional data with linear combinations of variables with the highest variances while discarding low variance variables as "noise." This optimization strategy may not have retained the relevant variables for depicting state transitions. One of the primary objectives of many trajectory analysis techniques is thus to find and retain the necessary information from a multidimensional data space relevant for mapping transitory relationships in a different data space.

t-Distributed stochastic neighbor embedding (t-SNE), a nonlinear dimensionality reduction approach,[54] has emerged as a popular and powerful technique for the analysis of single-cell data generated by a wide variety of experimental platforms.[55–58] t-SNE focuses on preserving the local structure while de-emphasizing the global structure of high-dimensional data, resulting in similar data points clustering together in an unsupervised manner. Because t-SNE allows user definition of the number of axes for analysis, cell populations can be unbiasedly shown in 2 or 3 dimensions. Although useful for defining divergent cell populations, the prospect for using t-SNE for trajectory analysis remains undefined. Because t-SNE is a stochastic algorithm emphasizing local data structure, the membership of each t-SNE–defined cluster is robust whereas the positions of the clusters are randomized in every run of the same data.[59] Of note, the relative distances and positions between t-SNE–defined clusters may not be meaningful and should be evaluated carefully. Thus, using t-SNE to establish relationships between cell populations to model transition from one cell population to another (such as from a stem cell population to a differentiated cell population) may not be appropriate. Nevertheless, t-SNE can be used as a gating strategy before trajectory analysis to identify cells that are related in the same lineage continuum for further analysis, as opposed to those that are in separate lineages. This step is crucial because most trajectory alignment algorithms (noted later) will try to establish relationships between all cells in the input data, even though such relationships do not exist biologically.

Established Algorithms for Trajectory Reconstruction

Trajectory analysis algorithms generally can be categorized into 2 groups, minimum spanning tree (MST)-based approaches and nonlinear embedding approaches. A MST is an acyclic graph with all the nodes connected in such a way to minimize the total edge weight, which in many cases represents the distance in data space between nodes. The idea is that nodes of the MST, which represent cells or clusters of cells, and their connections approximate the geometric shape of the data cloud when laid out in 2 dimensions. Multiple MST algorithms (eg, Spanning-tree Progression Analysis of Density-normalized Events, Monocle1, Tools for Single Cell ANalysis, Waterfall) exist and they differ by their applications on different experimental platforms, and the type and degree of clustering of data that occurs before MST construction.[10,60–62] MSTs represent the first algorithms that attempt to map transition trajectories from single-cell data. In addition to the general problem with clustering continuous data, MST-based algorithms are well known to be unstable, such that multiple applications on the same data set result in multiple, seemingly random solutions.[63,64] MST algorithms also tend to overfit smaller data sets, producing topologies with superfluous branches.[65,66] Thus, MST-based tools have shown utility mostly in well-defined systems such as hematopoiesis, in which a previously determined correct solution can be selected from an ensemble of solutions that include incorrect ones. Some MST-based algorithms developed strategies to mitigate some of these issues. For instance, Monocle1 allows the user to set a parameter to limit the number of branches present in the final graph, but this parameter requires prior knowledge as to how many independent differentiated cell types are present, which may not be known in less-defined systems.[10] Other approaches such as Ensemble Cell Lineage Analysis with Improved Robustness take a cohort of MSTs generated from the same data set and attempt to extract a consensus tree from the most common connections.[63] However, given the general instability of MSTs, the common connections may only generate the most rudimentary topology that may or may not provide new biological insights. Thus, the field has adopted other algorithms that are more robust and provide consistent results when applied to the same data.

The second class of algorithms, nonlinear embedding, incorporates nonlinear dimensionality reduction techniques to deconvolute difficult-to-interpret, high-dimensional data into more approachable 2- to 3-dimensional representation. Unlike PCA, which assumes linear combinations of features can approximate the original data, nonlinear embedding assumes the data cloud in mathematical space lies on a nonlinear manifold, which is a mathematical topologic space (sphere, torus, and so forth) that preserves the distances of points in close proximity. t-SNE is one such nonlinear embedding approach, but different classes of algorithms have different assumptions regarding the nature, distribution, and shape of the data cloud. Unlike t-SNE, which nonlinearly transforms data into distinct clusters, trajectory analysis on continuous data aims for embedding of data into elongated and compressed shapes to capture major structures and progressive trends in the data. Multiple such embedding approaches have been adopted for single-cell data analysis, including Diffusion maps[67] used in various algorithms such as Wishbone,[65,68] local linear embedding used in Selective Locally Linear Inference of Cellular Expression Relationships,[52] and multidimensional scaling and mapper in scTDA.[47] Adoption of nonlinear embedding algorithms, which were not originally designed for biological data, has accessibility issues with biologists. Specifically, the parameters for tuning these algorithms are mathematical in nature, but can have dramatic effects in shrinking or expanding the data such that local resolution may be gained or lost. Thus, nonlinear embedding algorithms are largely used for depicting simple topologies that can be described by the largest variation in the data most insensitive to parameter changes. One of the major goals of newer algorithms is for complex, multibranching trajectories to be depicted robustly.

The Next Generation of Algorithms to Reconstruct Cell-State Transition Trajectories

Next-generation algorithms that do not fall within the MST or nonlinear embedding categories have been developed recently. Force-directed layout, such as FLOW-MAP[66] and SPRING,[69] are a graph visualization strategy in which a densely connected network in multidimensional space is redistributed in a lower-dimensional space (eg, in 2D) by considering edges as weighted springs and using physical laws to simulate the equilibrium position of nodes as an energy minimization problem. Whether cells are clustered or whether and what type of prior dimension reduction has been performed differentiates these algorithms. Force-directed layout resolves the problem of stochasticity of MST algorithms by using multiple connections to guide the layout. However, the interconnectedness of the graphs makes it difficult to analyze cellular transitions outside in addition to visualization, given that all cell states in the graph will be connected to multiple other cell states. A significant advantage, however, is the possibility to represent nonacyclic structures, such as loops that occur in cell-cycle state transitions.[47,69]

Another new algorithm, Monocle2,[51] uses a process called *reverse graph embedding* to construct pseudotemporal trajectories in an unsupervised fashion. Monocle2 is currently the most widely used next-generation algorithm for trajectory analysis capable of producing multibranching trees. In principle, Monocle2 iteratively embeds data points, in a process similar to k-means clustering, into multiple principal curves.[70] Instead of learning clusters of cells, Monocle2 learns multiple principal curves connecting into a spanning tree that reflects a transitional hierarchy (Figure 2*A*). As with other techniques, Monocle2 works best

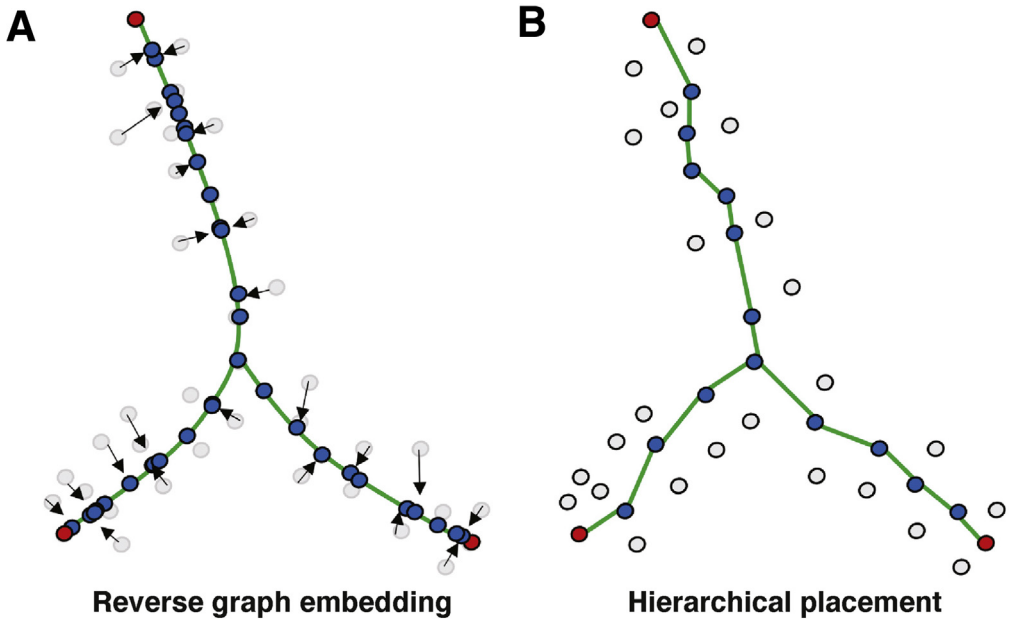

Figure 2. New approaches for trajectory analysis from single-cell data. (*A*) Monocle2 embeds the data cloud into a graph composed of principal curves. (*B*) p-Creode learns the most likely path through the data cloud as a function of density and shape. *Arrows* represent data embedding into the graph.

with expert guidance because multiple parameters that significantly affect the output must be specified. These parameters tune the fit of the principal curves in mathematical space. An example of how user input can alter interpretations is the fact that including 2 (default) or 10 principal components greatly altered the number of cell lineages that can be identified. Monocle2 results have been shown to be robust on multiple runs and different parameters on singly-bifurcating trajectories.

Although most algorithms aim to produce one output representation of cell-state transition processes, few evaluate the quality of such output by its statistical support by data. In many cases, the output of an algorithm is solely evaluated based on its fit to a known differentiation hierarchy, which raises the possibility of overfitting. Although cross-validation and bootstrapping methods are useful methods of evaluation, the difficulty lies in the current inability to compare overall topologic structures of graph outputs with both differing nodes and edges, which are produced over multiple different runs on the same data set. The p-Creode algorithm[64] is unique in this respect by leveraging an ensemble of N resampled topologies to lessen the effects of overfitting. p-Creode uses a unique hierarchical placement strategy for generating cell-state transition trajectories from end states identified in an unsupervised manner (Figure 2*B*). Instead of placing data points on leaves on a dendrogram as in hierarchical clustering, hierarchical placement allowed tiered assignment of data points as ancestor-descendent relationships. Multiple resampled runs then are evaluated by a graph dissimilarity metric called the p-Creode score to identify the number of different classes of topologies as well as the most representative topology from the ensemble. The parameters required to run p-Creode also are designed to be robust and accessible to nonexperts, which can be tuned according to how the data cloud visually appears. p-Creode also has been shown to generate robust and accurate results on complex multibranching trajectories even with noisy data. Despite these positives, p-Creode reliance on a downsampling preprocessing step may pose a problem for the automatic identification of rare cells, which cannot be distinguished from noise at the current time. Rare cell detection from relatively noisy single-cell data is a necessary and important area of development for all types of single-cell data analysis, and we anticipate rapid advances in this field.[13,71]

Downstream Analysis of Reconstructed Trajectories

Once trajectories are generated by various reconstruction algorithms, there are a substantial number of methods to extract biological insight, many of which are borrowed from bulk analyses such as RNA-seq. We will mention a few of the most common and insightful here. First, the topology of a cell-state transition trajectory may indicate when and where developmental decisions are made. For instance, a deep hierarchical topology may reflect a process by which a series of branching cell-fate decisions are made through identifiable progenitor states,[72] whereas a shallow, star-shaped topology can be interpreted as prepatterning, in which individuals from a seemingly homogeneous pool of progenitor cells (identified by RNA or protein) are already fated toward cell types[73,74] by mechanisms not evaluated (such as epigenetics). The analysis of network topologies can be formalized by graph theory, such as those used for identifying motifs, degree distribution, and transience of hubs from protein–protein interaction networks.[75–77] Second, a common analysis is to plot and visualize relative changes in analyte expression values over a pseudotime course.[9,65] This type of analysis can be performed over separate branches to identify mechanisms of maturation[10] or over branch points to show mechanisms of cell-fate decisions in which a cell must choose between 2 or more unique differentiation routes.[51] Manifold alignment algorithms, such as Manifold Alignment to CHaracterize Experimental Relationships, facilitate integrated comparisons between different trajectories (different routes/different data types depicting the same route, and so forth) with different cell state and temporal units.[78] Third, differentially expressed gene analysis along trajectories can be performed. In this case, however, instead of looking at genes differentially expressed between 2 conditions, one would group genes together on the basis that they show similar dynamics over a pseudotime course (eg, transient vs sustained expression). The hypothesis is that genes that are expressed in a correlated fashion may share common biological functions. As such, higher-level meta-analyses such as gene ontology enrichment, gene set enrichment, transcription factor–gene correlation analysis, and mathematical logic modeling have been used for constructing regulatory networks and models that are postulated to directly control cell decision making and/or progression.[47,62,79,80]

Notes on Using and Evaluating Trajectory Reconstruction Algorithms

As outlined in the previous sections, there are multiple algorithmic options for reconstructing trajectories from single-cell data. We leave you with a few points of considerations when applying these methods.

- Many algorithms are developed by showing that existing algorithms do not perform well on a synthetic data set or a newly generated data set, thus motivating the development of a new algorithm. Overfitting is a point of consideration when the data set used for building the algorithm also is used to show its effectiveness. The effectiveness of the algorithm also should be shown on existing data sets that generated well-behaved results by previous algorithms.
- Pseudotime currently has no real correspondence to real time. The number of cell states that recapitulates a trajectory can reflect the frequency of a transition event or the rate of transition. For instance, a longer branch can reflect a lineage that produces many cells compared with a shorter one.
- The distribution of the input data matters. Tissue-level data sets, which are expected to contain multiple

cellular phenotypes, usually are distributed with common and rare cell subsets. The power of droplet-based scRNA-seq approaches lies in their ability to query thousands of cells, and thus reduces the need for flow-sorting enrichment of rare cell populations for analysis. Uncommon cells can be extracted computationally after the data have been collected. However, results undoubtedly will be better analyzed for common cell types than rare ones. For instance, a 0.1% representation of a rare cell type even in a 4000-cell data set will be represented by only 4 data points in the data set. Although down-sampling and other strategies can be applied to normalize the distribution of data post hoc, a better strategy would be to tackle this issue during data collection. Enrichment experimental strategies for target populations, and/or methods to remove overly abundant or uninteresting cells may be considered depending on the biological question and the cell type of study.

- All computational modeling approaches are hypothesis-generating tools that require assumptions to be fulfilled and results to be validated. For trajectory analysis algorithms, the key assumption is that transitioning cell states are represented within the collected data. Thus, whether tissue is being harvested during embryonic development vs adult will greatly affect the interpretation of results. For instance, pancreatic islet development is completed by embryonic day 16.5. Thus, an adult pancreatic data set collected at homeostasis will contain very few transitioning cells and will be unsuitable for trajectory analysis. Furthermore, results generated in silico should always be confirmed experimentally by methods such as conventional lineage tracing or lineage perturbation experiments.[64] More recently, next-generation approaches that leverage mutational scars, such as those induced by Clustered Regularly Interspaced Short Palindromic Repeats, have been developed for accurately determining if individual cells belong to the same lineage in the classic, parent-child sense.[81,82] These approaches can potentially be integrated with single-cell approaches to combine cell-state transitional information with parent-child lineage data.

Future development of trajectory analysis algorithms will probably improve scalability to meet the demands of even higher throughput technologies, adopt approaches to reduce the impact of overfitting, and ideally be more user-friendly to nonexpert biologists, either by being completely unsupervised or by incorporating more intuitive methods and visualization for the parameter tuning process.

References

1. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM, Melton DA, Yanai I. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. Cell Syst 2016;3:346–360.
2. Haber AL, Biton M, Rogel N, Herbst RH, Shekhar K, Smillie C, Burgin G, Delorey TM, Howitt MR, Katz Y, Tirosh I, Beyaz S, Dionne D, Zhang M, Raychowdhury R, Garrett WS, Rozenblatt-Rosen O, Shi HN, Yilmaz O, Xavier RJ, Regev A. A single-cell survey of the small intestinal epithelium. Nature 2017;551:333–339.
3. McKinley ET, Sui Y, Al-Kofahi Y, Millis BA, Tyska MJ, Roland JT, Santamaria-Pang A, Ohland CL, Jobin C, Franklin JL, Lau KS, Gerdes MJ, Coffey J. Optimized multiplex immunofluorescence single-cell analysis reveals tuft cell heterogeneity. JCI Insight 2017;2:11.
4. Yan KS, Gevaert O, Zheng GXY, Anchang B, Probert CS, Larkin KA, Davies PS, Cheng Z, Kaddis JS, Han A, Roelf K, Calderon RI, Cynn E, Hu X, Mandleywala K, Wilhelmy J, Grimes SM, Corney DC, Boutet SC, Terry JM, Belgrader P, Ziraldo SB, Mikkelsen TS, Wang F, von Furstenberg RJ, Smith NR, Chandrakesan P, May R, Chrissy MAS, Jain R, Cartwright CA, Niland JC, Hong Y-K, Carrington J, Breault DT, Epstein J, Houchen CW, Lynch JP, Martin MG, Plevritis SK, Curtis C, Ji HP, Li L, Henning SJ, Wong MH, Kuo CJ. Intestinal enteroendocrine lineage cells possess homeostatic and injury-inducible stem cell activity. Cell Stem Cell 2017;21:78–90.
5. Simmons AJ, Banerjee A, McKinley ET, Scurrah CR, Herring CA, Gewin LS, Masuzaki R, Karp SJ, Franklin JL, Gerdes MJ, Irish JM, Coffey RJ, Lau KS. Cytometry-based single-cell analysis of intact epithelial signaling reveals MAPK activation divergent from TNF-α-induced apoptosis in vivo. Mol Syst Biol 2015;11:835.
6. Mair F, Hartmann FJ, Mrdjen D, Tosevski V, Krieg C, Becher B. The end of gating? An introduction to automated analysis of high dimensional cytometry data. Eur J Immunol 2016;46:34–43.
7. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med 2017;9:75.
8. Lyons J, Herring CA, Banerjee A, Simmons AJ, Lau KS. Multiscale analysis of the murine intestine for modeling human diseases. Integr Biol (Camb) 2015;7:740–757.
9. Marco E, Karp RL, Guo G, Robson P, Hart AH, Trippa L, Yuan G-C. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. Proc Natl Acad Sci U S A 2014;111:E5643–E5650.
10. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 2014;32:381–386.
11. Kim T, Saadatpour A, Guo G, Saxena M, Cavazza A, Desai N, Jadhav U, Jiang L, Rivera MN, Orkin SH, Yuan G, Shivdasani RA. Single-cell transcript profiles reveal multilineage priming in early progenitors derived from Lgr5(+) intestinal stem cells. Cell Rep 2016;16:2053–2060.
12. Bendall SC, Simonds EF, Qiu P, Amir ED, Krutzik PO, Finck R, Bruggner RV, Melamed R, Trejo A, Ornatsky OI, Balderas RS, Plevritis SK, Sachs K, Pe'er D, Tanner SD, Nolan GP. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science 2011;332:687–696.

13. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature 2015;525:251–255.
14. Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow M, Quake SR. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 2014;509:371–375.
15. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, Amit I. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science 2014;343:776–779.
16. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, Leonhardt H, Heyn H, Hellmann I, Enard W. Comparative analysis of single-cell RNA sequencing methods. Mol Cell 2017;65:631–643.
17. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 2015;161:1187–1201.
18. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 2015;161:1202–1214.
19. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, Bielas JH. Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017;8:14049.
20. Chuang L-S, Villaverde N, Hui KY, Mortha A, Rahman A, Levine AP, Haritunians T, Evelyn Ng SM, Zhang W, Hsu N-Y, Facey J-A, Luong T, Fernandez-Hernandez H, Li D, Rivas M, Schiff ER, Gusev A, Schumm LP, Bowen BM, Sharma Y, Ning K, Remark R, Gnjatic S, Legnani P, George J, Sands BE, Stempak JM, Datta LW, Lipka S, Katz S, Cheifetz AS, Barzilai N, Pontikos N, Abraham C, Dubinsky MJ, Targan S, Taylor K, Rotter JI, Scherl EJ, Desnick RJ, Abreu MT, Zhao H, Atzmon G, Pe'er I, Kugathasan S, Hakonarson H, McCauley JL, Lencz T, Darvasi A, Plagnol V, Silverberg MS, Muise AM, Brant SR, Daly MJ, Segal AW, Duerr RH, Merad M, McGovern DPB, Peter I, Cho JH. A frameshift in CSF2RB predominant among Ashkenazi Jews increases risk for Crohn's disease and reduces monocyte signaling via GM-CSF. Gastroenterology 2016;151:710–723.
21. Simmons AJ, Lau KS. Deciphering tumor heterogeneity from FFPE tissues: its promise and challenges. Mol Cell Oncol 2016;4:e1260191.
22. Simmons AJ, Scurrah CR, McKinley ET, Herring CA, Irish JM, Washington MK, Coffey RJ, Lau KS. Impaired coordination between signaling pathways is revealed in human colorectal cancer using single-cell mass cytometry of archival tissue blocks. Sci Signal 2016;9:rs11.
23. Adam M, Potter AS, Potter SS. Psychrophilic proteases dramatically reduce single-cell RNA-seq artifacts: a molecular atlas of kidney development. Development 2017;144:3625–3632.
24. Alles J, Karaiskos N, Praktiknjo SD, Grosswendt S, Wahle P, Ruffault P-L, Ayoub S, Schreyer L, Boltengagen A, Birchmeier C, Zinzen R, Kocks C, Rajewsky N. Cell fixation and preservation for droplet-based single-cell transcriptomics. BMC Biol 2017;15:44.
25. Thomsen ER, Mich JK, Yao Z, Hodge RD, Doyle AM, Jang S, Shehata SI, Nelson AM, Shapovalova NV, Levi BP, Ramanathan S. Fixed single-cell transcriptomic characterization of human radial glial diversity. Nat Methods 2016;13:87–93.
26. Guillaumet-Adkins A, Rodríguez-Esteban G, Mereu E, Mendez-Lago M, Jaitin DA, Villanueva A, Vidal A, Martinez-Marti A, Felip E, Vivancos A, Keren-Shaul H, Heath S, Gut M, Amit I, Gut I, Heyn H. Single-cell transcriptome conservation in cryopreserved cells and tissues. Genome Biol 2017;18:45.
27. Hu P, Fabyanic E, Kwon DY, Tang S, Zhou Z, Wu H. Dissecting cell-type composition and activity-dependent transcriptional state in mammalian brains by massively parallel single-nucleus RNA-seq. Mol Cell 2017;68:1006–1015.
28. Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, Choudhury SR, Aguet F, Gelfand E, Ardlie K, Weitz DA, Rozenblatt-Rosen O, Zhang F, Regev A. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods 2017;14:955–958.
29. Habib N, Li Y, Heidenreich M, Swiech L, Avraham-Davidi I, Trombetta JJ, Hession C, Zhang F, Regev A. Div-seq: single-nucleus RNA-seq reveals dynamics of rare adult newborn neurons. Science 2016;353:925–928.
30. Giesen C, Wang HO, Schapiro D, Zivanovic N, Jacobs A, Hattendorf B, Schüffler PJ, Grolimund D, Buhmann JM, Brandt S, Varga Z, Wild PJ, Günther D, Bodenmiller B. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. Nat Methods 2014;11:417–422.
31. Angelo M, Bendall SC, Finck R, Hale MB, Hitzman C, Borowsky AD, Levenson RM, Lowe JB, Liu SD, Zhao S, Natkunam Y, Nolan GP. Multiplexed ion beam imaging of human breast tumors. Nat Med 2014;20:436–442.
32. Gerdes MJ, Sevinsky CJ, Sood A, Adak S, Bello MO, Bordwell A, Can A, Corwin A, Dinn S, Filkins RJ, Hollman D, Kamath V, Kaanumalle S, Kenny K, Larsen M, Lazare M, Li Q, Lowes C, McCulloch CC, McDonough E, Montalto MC, Pang Z, Rittscher J, Santamaria-Pang A, Sarachan BD, Seel ML, Seppo A, Shaikh K, Sui Y, Zhang J, Ginty F. Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. Proc Natl Acad Sci U S A 2013;110:11982–11987.
33. Lin J-R, Fallahi-Sichani M, Sorger PK. Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. Nat Commun 2015;6:8390.

34. Zrazhevskiy P, Gao X. Quantum dot imaging platform for single-cell molecular profiling. Nat Commun 2013; 4:1619.
35. Remark R, Merghoub T, Grabe N, Litjens G, Damotte D, Wolchok JD, Merad M, Gnjatic S. In-depth tissue profiling using multiplexed immunohistochemical consecutive staining on single slide. Sci Immunol 2016;1:aaf6925.
36. Riordan DP, Varma S, West RB, Brown PO. Automated analysis and classification of histological tissue features by multi-dimensional microscopic molecular profiling. PLoS One 2015;10:e0128975.
37. Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L. Single-cell in situ RNA profiling by sequential hybridization. Nat Methods 2014;11:360–361.
38. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 2015;348:aaa6090.
39. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso S, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. Science 2014;343:1360–1363.
40. Jungmann R, Avendaño MS, Woehrstein JB, Dai M, Shih WM, Yin P. Multiplexed 3D cellular super-resolution imaging with DNA-PAINT and Exchange-PAINT. Nat Methods 2014;11:313–318.
41. Yang B, Treweek JB, Kulkarni RP, Deverman BE, Chen CK, Lubeck E, Shah S, Cai L, Gradinaru V. Single-cell phenotyping within transparent intact tissue through whole-body clearing. Cell 2014;158:945–958.
42. Sylwestrak EL, Rajasethupathy P, Wright MA, Jaffe A, Deisseroth K. Multiplexed intact-tissue transcriptional analysis at cellular resolution. Cell 2016;164:792–804.
43. Murray E, Cho JH, Goodwin D, Ku T, Swaney J, Kim SY, Choi H, Park YG, Park JY, Hubbert A, McCue M, Vassallo S, Bakh N, Frosch MP, Wedeen VJ, Seung HS, Chung K. Simple, scalable proteomic imaging for high-dimensional profiling of intact systems. Cell 2015; 163:1500–1514.
44. Ritsma L, Ellenbroek SIJ, Zomer A, Snippert HJ, de Sauvage FJ, Simons BD, Clevers H, van Rheenen J. Intestinal crypt homeostasis revealed at single-stem-cell level by in vivo live imaging. Nature 2014;507:362–365.
45. Chubb JR, Trcek T, Shenoy SM, Singer RH. Transcriptional pulsing of a developmental gene. Curr Biol 2006; 16:1018–1025.
46. Bellman RE. Adaptive control processes: a guided tour. Princeton, NJ: Princeton University Press, 1961.
47. Rizvi AH, Camara PG, Kandror EK, Roberts TJ, Schieren I, Maniatis T, Rabadan R. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. Nat Biotechnol 2017; 35:551–560.
48. Guo M, Bao EL, Wagner M, Whitsett JA, Xu Y. SLICE: determining cell differentiation and lineage based on single cell entropy. Nucleic Acids Res 2017;45:e54.
49. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, Heisler MG. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods 2013; 10:1093–1095.
50. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol 1933; 24:417–441.
51. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods 2017;14:979–982.
52. Welch JD, Hartemink AJ, Prins JF. SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. Genome Biol 2016;17:106.
53. Furchtgott LA, Melton S, Menon V, Ramanathan S. Discovering sparse transcription factor codes for cell states and state transitions during development. Elife 2017;6:e20488.
54. Van Der Maaten LJP, Hinton GE. Visualizing high-dimensional data using t-sne. J Mach Learn Res 2008; 9:2579–2605.
55. Camp JG, Sekine K, Gerber T, Loeffler-Wirth H, Binder H, Gac M, Kanton S, Kageyama J, Damm G, Seehofer D, Belicova L, Bickle M, Barsacchi R, Okuda R, Yoshizawa E, Kimura M, Ayabe H, Taniguchi H, Takebe T, Treutlein B. Multilineage communication regulates human liver bud development from pluripotency. Nature 2017;546:533–538.
56. Yu Y, Tsang JCHH, Wang C, Clare S, Wang J, Chen X, Brandt C, Kane L, Campos LS, Lu L, Belz GT, McKenzie ANJ, Teichmann SA, Dougan G, Liu P. Single-cell RNA-seq identifies a PD-1hi ILC progenitor and defines its developmental pathway. Nature 2016; 539:102–106.
57. See P, Dutertre C-A, Chen J, Günther P, McGovern N, Irac SE, Gunawan M, Beyer M, Händler K, Duan K, Sumatoh HR, Ruffin N, Jouve M, Gea-Mallorquí E, Hennekam RCM, Lim T, Yip CC, Wen M, Malleret B, Low I, Shadan NB, Fen CFS, Tay A, Lum J, Zolezzi F, Larbi A, Poidinger M, Chan JKY, Chen Q, Rénia L, Haniffa M, Benaroch P, Schlitzer A, Schultze JL, Newell EW, Ginhoux F. Mapping the human DC lineage through the integration of high-dimensional techniques. Science 2017;356:eaag3009.
58. Lavin Y, Kobayashi S, Leader A, Amir ED, Elefant N, Bigenwald C, Remark R, Sweeney R, Becker CD, Levine JH, Meinhof K, Chow A, Kim-Shulze S, Wolf A, Medaglia C, Li H, Rytlewski JA, Emerson RO, Solovyov A, Greenbaum BD, Sanders C, Vignali M, Beasley MB, Flores R, Gnjatic S, Pe'er D, Rahman A, Amit I, Merad M. Innate immune landscape in early lung adenocarcinoma by paired single-cell analyses. Cell 2017;169:750–765.
59. Wattenberg M, Viegas F, Johnson I. How to use t-SNE effectively. Distill 2016. https://doi.org/10.23915/distill.00002.
60. Qiu P, Simonds EF, Bendall SC, Gibbs KD, Bruggner RV, Linderman MD, Sachs K, Nolan GP, Plevritis SK. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat Biotechnol 2011; 29:886–891.

61. Ji Z, Ji H. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Res 2016;44:e117.
62. Shin J, Berg DA, Zhu Y, Shin JY, Song J, Bonaguidi MA, Enikolopov G, Nauen DW, Christian KM, Ming GL, Song H. Single-cell RNA-seq with waterfall reveals molecular cascades underlying adult neurogenesis. Cell Stem Cell 2015;17:360–372.
63. Giecold G, Marco E, Garcia SP, Trippa L, Yuan G-C. Robust lineage reconstruction from high-dimensional single-cell data. Nucleic Acids Res 2016;44:e122.
64. Herring CA, Banerjee A, McKinley ET, Simmons AJ, Ping J, Roland JT, Franklin JL, Liu Q, Gerdes MJ, Coffey RJ, Lau KS. Unsupervised trajectory analysis of single-cell RNA-seq and imaging data reveals alternative tuft cell origins in the gut. Cell Syst 2018;6:37–51.
65. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, Choi K, Bendall S, Friedman N, Pe'er D. Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat Biotechnol 2016;34:637–645.
66. Zunder ER, Lujan E, Goltsev Y, Wernig M, Nolan GP. A continuous molecular roadmap to iPSC reprogramming through progression analysis of single-cell mass cytometry. Cell Stem Cell 2015;16:323–337.
67. Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, Zucker SW. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc Natl Acad Sci U S A 2005;102:7426–7431.
68. Haghverdi L, Buettner F, Theis FJ. Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics 2015;31:2989–2998.
69. Briggs JA, Li VC, Lee S, Woolf CJ, Klein A, Kirschner MW. Mouse embryonic stem cells can differentiate via multiple paths to the same state. Elife 2017;6:e26945.
70. Hastie T, Stuetzle W. Principal curves. J Am Stat Assoc 1989;84:502–516.
71. Jiang L, Chen H, Pinello L, Yuan G-C. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol 2016;17:144.
72. Seita J, Weissman IL. Hematopoietic stem cell: self-renewal versus differentiation. Wiley Interdiscip Rev Syst Biol Med 2010;2:640–653.
73. Paul F, Arkin Y, Giladi A, Jaitin DA, Kenigsberg E, Keren-Shaul H, Winter D, Lara-Astiaso D, Gury M, Weiner A, David E, Cohen N, Lauridsen FKB, Haas S, Schlitzer A, Mildner A, Ginhoux F, Jung S, Trumpp A, Porse BT, Tanay A, Amit I. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. Cell 2015;163:1663–1677.
74. Notta F, Zandi S, Takayama N, Dobson S, Gan OI, Wilson G, Kaufmann KB, McLeod J, Laurenti E, Dunant CF, McPherson JD, Stein LD, Dror Y, Dick JE. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. Science 2016;351:aab2116.
75. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet 2002;31:64–68.
76. Han J-DJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJM, Cusick ME, Roth FP, Vidal M. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 2004;430:88–93.
77. Yook S-H, Oltvai ZN, Barabási A-L. Functional and topological characterization of protein interaction networks. Proteomics 2004;4:928–942.
78. Welch JD, Hartemink AJ, Prins JF. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. Genome Biol 2017;18:138.
79. Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, Buettner F, Macaulay IC, Jawaid W, Diamanti E, Nishikawa S-I, Piterman N, Kouskoff V, Theis FJ, Fisher J, Göttgens B. Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat Biotechnol 2015;33:269–276.
80. Matsumoto H, Kiryu H. SCOUP: a probabilistic model based on the Ornstein-Uhlenbeck process to analyze single-cell expression data during differentiation. BMC Bioinformatics 2016;17:232.
81. Frieda KL, Linton JM, Hormoz S, Choi J, Chow K-HK, Singer ZS, Budde MW, Elowitz MB, Cai L. Synthetic recording and in situ readout of lineage information in single cells. Nature 2017;541:107–111.
82. McKenna A, Findlay GM, Gagnon JA, Horwitz MS, Schier AF, Shendure J. Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science 2016;353:aaf7907.

Received January 2, 2018. Accepted January 31, 2018.

Correspondence
Address correspondence to: Ken S. Lau, PhD, Epithelial Biology Center, Vanderbilt University Medical Center, 2213 Garland Avenue, 10475 MRB IV, Nashville, Tennessee 37232-0441. e-mail: ken.s.lau@vanderbilt.edu; fax: (615) 343-1591.

Acknowledgments
The authors would like to thank the Vanderbilt Epithelial Biology Center for helpful discussions.

Author contributions
Charles A. Herring and Ken S. Lau wrote the majority of the manuscript, with specific sections contributed by Bob Chen and Eliot T. McKinley.

Conflicts of interest
The authors disclose no conflicts.

Funding
Funded by training grant (NICHD) T32HD007502 and predoctoral grant (NIGMS) F31GM120940 from the National Institutes of Health (C.A.H.); by training grant (NCI) R25CA092043 (E.T.M.); and grant (NIDDK) R01DK103831, and pilot project grants (NIDDK) P30DK058404 and (NCI) P50CA095103 (K.S.L.).