Cell Fate Relationships Mapped by p-Creode Trajectory Analysis of Single-cell Data

By

Charles A Herring

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Chemical and Physical Biology

May 11, 2018

Nashville, Tennessee

Approved:

Vito Quaranta, M.D.

Ken S. Lau, Ph.D.

Gregor Neuert, Ph.D.

Erin C. Rericha, Ph.D.

John A. Capra, Ph.D.

To everyone culpable for the Great Recession, without your malfeasance I would likely still be roofing houses.

ACKNOWLEDGMENTS

Without the help and support of a number of people the work presented here would not have been possible. Firstly, my advisor, Ken Lau, for always being accessible and for being far more knowledge than his tenure track would let on. Since my work is computational and I was long ago banned from touching even a pipet, my fellow lab members and coauthors have been invaluable to the completion of this dissertation, including but not limited to Sun Wook Kim, Cherie Scurrah, Jie Ping, Joseph Roland, Jeffrey Franklin, Qi Liu, Michael Gerdes, Bob Chen, and Robert Coffey. I would particularly like to thank Amrita Banerjee, Joey Simmons, and Eliot McKinley for their contributions. Amrita performed the mouse experiments, Joey the mass cytometry experiments, and Eliot the MxIF experiments presented in this dissertation. I would also like to thank my committee Vito Quaranta, Gregor Neuert, Erin Rericha, and John Capra for all of their guidance throughout this process. I have always enjoyed my committee meetings.

I am grateful for the financial support from the Vanderbilt Stem Cell and Regenerative Developmental Biology Training Grant and the Ruth L. Kirschstein Predoctoral Individual National Research Service Award NIH F31 Fellowship. I would especially like to think Chris Wright and the selection committee for the Vanderbilt Stem Cell and Regenerative Developmental Biology Training Grant for taking a chance on a programmer.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

Introduction

The emergent behavior that is multi-cellular organ function arises from a heterogeneous collection of individual cells with distinct phenotypes and behaviors. For instance, the function of the small intestine is collectively prescribed by enterocytes that transport molecules across the intestinal barrier, goblet cells and Paneth cells that secrete barrier-promoting mucus and anti-microbial peptides, respectively, and enteroendocrine cells that release gastrointestinal hormones that aid digestion (van der Flier and Clevers, 2009). Understanding of organ function hinges on our understanding of cell state transitions leading to these functionally distinct cell types. Multipotent progenitor cells transition towards mature states through continuous, intermediary steps with increasingly restricted access to other cell states[1] (Waddington, 1957; Blanpain and Fuchs, 2014).

There are currently a variety of approaches for investigating transitional cell states. For instance, a stem cell can be identified by lineage tracing, a method whereby continuous generation and differentiation of cells from a labeled source results in permanently labeled organ units (Barker et al., 2007). Other seminal studies have determined the relationship between stem and differentiated cells by focusing on the effects of genetic and epigenetic perturbations on terminal cell states (Noah et al., 2011). While the behaviors of intermediate states such as progenitor cells remain to be fully elucidated, modern single-cell technologies have enabled the interrogation of transitional cell states that contain information regarding branching cell fate decisions across entire developmental continuums (Gerdes et al., 2013; Giesen et al., 2014; Klein et al., 2015; Grn et al., 2015; Treutlein et al., 2014; Paul et al., 2015; Simmons et al.,

[1] In the context of this dissertation, phenotype or cell type refers a population of cells previously characterized by common attributes (e.g., expression profile, morphology, etc.), while cell state refers to the condition of a single cell or small group of cells. A cell depending upon its condition state may or may not belong to a cell type, and cells within a cell type may have a range of cell states. Its important to note, inclusion or exclusion of cells within a particular cell type can be a subjective exercise. Therefore, care was taken to adhere to published normals (as referenced in the text) where cell type classifications were made.

2016). Despite having experimental tools to generate data at single-cell resolution, resolving cellular relationships from large volumes of data has remained a challenge.

Various computational approaches have been developed for tracking cell transition trajectories when time series data are available (Marco et al., 2014; Zunder et al., 2015). However, for adult and human tissues, *in vivo* cell transitions have to be inferred from data collected only as a snapshot in time. A major push in the field of single-cell biology is to enable data-driven arrangements of cell states into pseudo temporal trajectories and from these trajectories infer cell transitions. Mapping a snapshot of data into trajectories fall broadly into two categories: Minimum Spanning Tree (MST)-based (Trapnell et al., 2014; Shin et al., 2015; Ji and Ji, 2016; Anchang et al., 2016; Qiu et al., 2011) or non-linear data embedding (Haghverdi et al., 2015; Welch et al., 2016; Setty et al., 2016). (Zunder et al., 2015; Briggs et al., 2017; Qiu et al., 2017; Herring et al., 2018). MST based approaches were a necessary first approach for the field, serving as a guide for all subsequent approaches, but they suffer from known weaknesses. For instance, MST algorithms are widely known to not be robust with large datasets such that multiple distinct solutions are computed given the same dataset (Giecold et al., 2016). MST algorithms also tend to over-fit smaller datasets, producing topologies with superfluous branches (Zunder et al., 2015; Setty et al., 2016). While these tools have shown utility when applied to well-defined systems such as hematopoiesis, they do not provide a direct means to assess solutions for determining the correct topologies of less-defined systems. On the other hand, non-linear embedding algorithms, such as Diffusion Map, are sensitive to the distribution of data such that local resolution may be gained or lost, and thus are not universally applicable to all biological investigations and data types (Setty et al., 2016). While a large amount of effort has focused on visualization strategies (Zunder et al., 2015; Briggs et al., 2017), development of solutions to statistically assess computed results remains a work in progress. A class of algorithms developed by Dana Pe'er's group using supervised-random walk over a cell network produces robust results and can be statistically scored (Bendall et al., 2014; Setty et al., 2016). A more recent advance, named Wishbone, can identify bifurcation points in a trajectory, but is limited to cases with a known single branch point (Setty et al., 2016). There is currently no experimentally robust algorithm using biological replica, that generates cell transition hierarchies *de novo* to map multiple branching decisions in a statistically testable way.

Tuft cells, also known as brush or caveolated cells, in the gut are a rare population of chemosensory cells that remains poorly understood (Gerbe et al., 2016). They originate from epithelial stem cells (Gerbe et al., 2011), and express taste receptors such as α-gustducin (Hfer et al., 1996) and TRPM5 (Bezenon et al., 2008, 2007), which implicate their function in chemoreception similar to lingual taste cells. Recently, a number of important studies have demonstrated their role in immune responses against helminth infection by establishing an IL25-IL13 circuit with innate lymphoid cells type 2 (ILC2s) (Gerbe et al., 2016; von Moltke et al., 2015; Howitt et al., 2016). Thus, understanding the development of tuft cells is important in intestinal disease contexts. Tuft cells are commonly thought to be specified from the secretory lineage (Gerbe et al., 2011) along with goblet, Paneth, and enteroendocrine cells (VanDussen et al., 2012), although their origins have recently been disputed (Bjerknes et al., 2012).

In this dissertation, I present p-Creode, a novel algorithm built from first principles capable of deriving multi-branching transition pseudo temporal trajectories with an unique method to statistically score resulting trajectories. The term creode was coined by C.H. Waddington, combining the Greek words for necessary and path (Waddington, 1957). I validated p-Creode using a variety of single-cell platforms and biological tissues ranging in differentiation complexity. To demonstrate utility, p-Creode was applied to mouse small intestine and colon to clarify the lineage origin of tuft cells. Along with experimental validation, p-Creode revealed tuft cells may be specified outside the Atoh1-dependent secretory lineage in the small intestine, but are regulated by Atoh1 in the colon. These findings highlight important physiological differences between the small intestine and the colon, which directly impact the development and function of tuft cells in these two anatomically distinct regions.

In the following sections, I provide an in depth background on single-cell technologies and existing trajectory mapping algorithms, with particular focus on strengths and weaknesses of each. Contained within the background is an overview of common analyses that can be combined with trajectory analysis. A rigorous validation of p-Creode is presented next, demonstrating utility and robustness versus existing algorithms (Bendall et al., 2011; Setty et al., 2016; Qiu et al., 2017). Following validation, a detailed overview of p-Creode is given which includes an independent validation of p-Creode's scoring metric. p-Creode scoring metric is then applied to data from independents datasets, demonstrating sensitivity when comparing trajectories across datasets at the biological and technical replicate level. Lastly, I give an overview of experimental details utilized in this dissertation and conclude with remarks on the current and future relevance of p-Creode within the field.

CHAPTER 2

Background

2.1. Single-cell Experimental Technologies

The theoretical basis of pseudo temporal ordering is that asynchronous sampling from multiple time points over development (Marco et al., 2014) or snap-shot sampling at a single time point of a continually renewing tissue (such as the intestine) (Qiu et al., 2011; Trapnell et al., 2014) can result in a dense sampling of transitional states that can be aligned to reflect a time course of state transitions (Figure 2.1). A cell state is represented by the position of a cell in a data space defined by multiple molecular markers that describe cellular identity and behavior (Figure 2.1). Ordering of these cell states is conducted on the basis of similarity in marker expression and dense sampling of these states is required to obtain a continuum of data by which relationship between cell states can be inferred. Because transitional cell states are often rare compared with differentiated cells in tissue, it is required that single-cell technologies be able 1) to measure multiple markers to characterize a state, and 2) to query a large volume of data points in order to fully depict a data continuum. Below is a brief review of 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 multi-parameter flow cytometry, and transcript-based techniques such as single-cell RNA-sequencing (scRNA-seq) and gene expression assays (Kim et al., 2016). 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 (Bendall et al., 2011), while scRNA-seq can quantify gene expression in an unbiased, genome-wide manner (thousands of gene analytes) (Grn et al., 2015; Jiang et al., 2016). Multiple platforms of scRNA-seq exist, with variations in cell containment strategies ranging from microwells (Jaitin et al., 2014; Treutlein et al., 2014; Ziegenhain et al., 2017) to liquid-oil emulsion droplets (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017) and many of these current versions are able to query up to thousands of cells from a single sample.

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 the certain niches. Current multiplex imaging technologies for proteins can be classified either as mass-based or iterative. Mass-based imaging approaches, including imaging mass cytometry (Giesen et al., 2014) and multiplex ion beam imaging (Angelo et al., 2014) rely on metal-tagged antibodies coupled to mass spectrometry, while iterative approaches, including MxIF (Gerdes et al., 2013), CycIF (Lin et al., 2015), and others (Remark et al., 2016; Riordan et al., 2015; Zrazhevskiy and Gao, 2013) rely on cycles of staining, imaging, and de-staining to enable multiplexity. Microscopy approaches can also query thousands of cells if whole tissues are imaged at the appropriate resolution, although the time for acquisition of such datasets can be substantial on a per sample basis (McKinley et al., 2017).

The choice of suspension or *in situ* techniques is highly dependent on the experimental question being sought and can oftentimes be complementary. Suspension approaches are much higher throughput in terms of the number of cells and analytes analyzed, while *in situ* techniques can provide spatial resolution. 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.

2.2. Feature Selection

Feature selection is a pre-processing 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, due to the processing and amplification of small amounts of nucleic acids (Ziegenhain et al., 2017) and the biological phenomenon of bursting transcription (Chubb et al., 2006). The effects of noise are compounded in multi-dimensional space in a phenomenon known as the curse of dimensionality (Bittner, 1962), 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 the rest of the uninformative or noisy features. The feature selection step is implicitly performed in candidate-based approaches, such as CyTOF and multiplex microscopy, since 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 scTDA and SLICE incorporate annotated gene sets from gene ontology resources such as PANTHER or DAVID to select features in a semi-supervised fashion (Guo et al., 2017; Rizvi et al., 2017). For studies with minimal or unreliable prior knowledge, completely unsupervised methods that leverage general gene expression patterns may be used.

Unsupervised feature selection methods vary in complexity and assumptions made about data. For example, a commonly used method in analyzing scRNA-seq data involves identifying transcriptomic features with highly variable expression across the entire dataset of single cells. Here, the assumption is that different cell types express genes at different levels, therefore genes with high variability are more adept at characterizing cell states. 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 (Brennecke et al., 2013). In a way, this method is analogous to Principal Component Analysis (PCA) in selecting dimensions with the highest variances (Hotelling, 1933). Due to reliance on variance, gene expression of cell types with a large number of cells may dominate the gene selection process, leaving genes that characterize rare populations under represented in the final gene set. Other aspects that can influence this approach include biological and technical noise. For instance, biological noise produced by over abundance of cell cycle genes can create more variation among dividing and non-dividing cells than between different cell types (Barron and Li, 2016). It should be noted, biological noise from highly expressed ribosomal RNA (rRNA) do not have to be considered when scRNA-seq protocols use an oligo-dT based primer, which selects for polyA-positive mRNA and not polyA-negative rRNA (Fang and Akinci-Tolun, 2016). Technical variation can potentially exceed meaningful biological variation, and filtering methods can be confounded by the simultaneous occurrence of these two sources of variation, leading to selection of genes that more represent technical variation rather than biological significance (Hicks et al., 2017). A major source of technical variation stems from dropouts, where a gene present in a cell fails to be read and is reported as having an expression level of zero. Possible technical reasons for a dropout to occur includes mRNA degradation after cell lysis, variability in amplification efficiency, and dilution of cell libraries or sequencing depth (Lun et al., 2016; Vallejos et al., 2017). Despite these potential drawbacks, 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 cell relationships.

Other methods have been developed that focus on different patterns of gene expression to identify relevant features. The Trapnell group developed dpFeature, a method that selects differentially expressed genes between cell populations described by unsupervised clustering as biologically meaningful for downstream trajectory analysis (Qiu et al., 2017). Clusters of cells automatically identified are representative of distinct cell states, and differentially expressed genes represent likely regulators of these states. However, datasets that depict transitions are generally continuously distributed and do not form distinct clusters. Clustering in these cases is based on arbitrary cutoffs, and thus, how dpFeature performs on these types of datasets remains to be tested.

To handle continuous data distributions, Welch *et al.* developed a metric called neighborhood variance. By implementing a K-nearest neighbors graph approach with each cell represented as a node (Welch et al., 2016), this method defines a neighborhood of locally varying cell states. Variance of a feature is analyzed over each defined neighborhood and compared to 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.* developed a Bayesian approach for identifying subsets of gene expression patterns over three adjacent cell states that is useful for defining lineage relationships (Furchtgott et al., 2017). These feature selection methods highlight unique expression patterns locally, over adjacent cell states, and downplay the significance of globally varying genes. More refined gene expression patterns perhaps can be identified in the future for more sophisticated feature selection.

2.3. t-Distributed Stochastic Neighbor Embedding (t-SNE) - 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 2.1). Thus, multiple methods, such as principal component analysis (PCA), have been developed to represent high dimensional data in a lower-dimensional space while best retaining the underlying relationships amongst data points in the original data space (Hotelling, 1933). 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 and the transitional process aligns with the selected axes. In practice however, all dimensionality reduction techniques result in information loss, as 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. Because PCA uses variance for its axes selection criteria, it is similar to variation based feature selection methods, and, therefore, susceptible to biological and technical noise. Although it should be noted, there are variations of PCA that are less susceptible to noisy outliers, namely independent component analysis (ICA) (Hyvrinen et al., 2001), but ICA will still not preform well on datasets overrun with noise. Therefore, these optimization strategies may not retain the relevant variables for depicting state transitions in 2- or 3-dimensional space. One of the primary objectives of trajectory analysis techniques is thus to find and retain the necessary information from a multi-dimensional data space relevant for mapping transitory relationships in a different data space.

t-SNE, a non-linear dimensionality reduction approach (Maaten and Hinton, 2008) has emerged as a popular and powerful technique for the analysis of single-cell data generated by a wide variety of experimental platforms (Camp et al., 2017; Lavin et al., 2017; See et al., 2017; Yu et al., 2016). t-SNE focuses on preserving the local structure while deemphasizing 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 displayed in 2 or 3 dimensions. While useful for defining divergent cell populations, the prospect for using t-SNE for trajectory analysis remains undefined. As t-SNE is a stochastic algorithm emphasizing local data structure, the membership of each t-SNE-defined cluster is robust while the positions of the clusters are randomized in every run of the same data (Wattenberg et al., 2016). Of note, the relative distances and positions between t-SNE-defined clusters may not be meaningful and should be carefully evaluated. 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 litmus test for suitability of trajectory analysis; data that form distinct clusters may not be suitable, whereas a dataset contained in large continuous data cloud is likely a good candidate for trajectory analysis (Figure 2.2). t-SNE can serve as a gating strategy prior to trajectory analysis to identify cells that are related in same lineage continuum for further analysis, as opposed to those that are no longer in a state of transition (Figure 2.2). This step is crucial, as most trajectory alignment algorithms will try to establish relationships between all cells in the input data even if transitional cell states do not biologically exist between them. Essentially, the algorithms may be confounded by the absence of transitional cell states and infer a lineage among potentially developmentally unrelated cell types.

2.4. Established Algorithms for Trajectory Reconstruction

Trajectory analysis algorithms can be generally categorized into two groups, minimum spanning tree (MST)-based approaches and non-linear embedding approaches. A MST is an acyclic graph with all the nodes connected in such a way to minimize the edge weight, which in many cases, represent 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 2D. Multiple MST algorithms (e.g., SPADE, Monocle1, TSCAN, Waterfall) exist and they differ by their applications on datasets generated by different experimental platforms, as well as the type and degree of clustering of data that occurs prior to MST construction (Ji and Ji, 2016; Qiu et al., 2011; Shin et al., 2015; Trapnell et al., 2014). MST represent the first algorithms that attempted 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 iterations on the same dataset result in multiple, seemingly random solutions (Giecold et al., 2016). MST al-

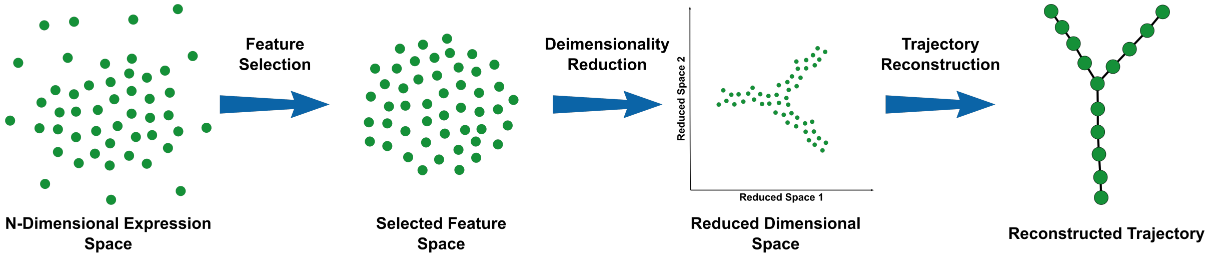


Figure 2.1: General workflow of trajectory analysis algorithms. Beginning with data in multi-dimensional space, feature selection is first performed to include relevant analytes and exclude noise. From the selected feature set, dimension reduction is applied to best emphasize the part of the data most relevant to cell-state transitions. Trajectories are then reconstructed in this reduced space and analyzed as pseudotime courses.

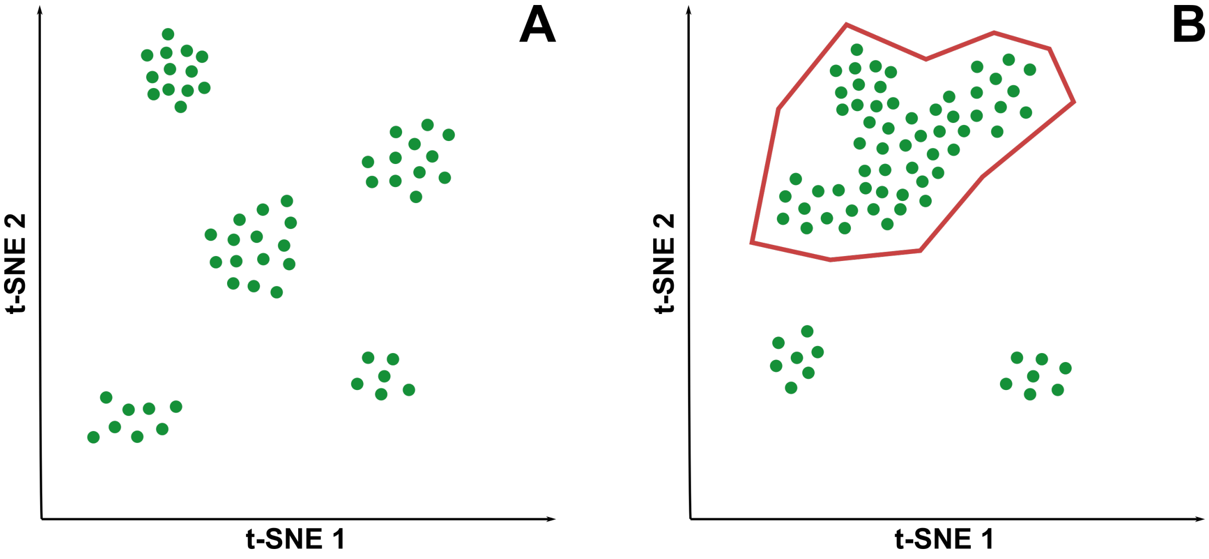


Figure 2.2: Example t-SNE plots of population versus continuous data (A) Example t-SNE plot of data not well suited for trajectory analysis, characterized by distinct populations of cells (B) Example t-SNE plot with both distinct cell populations and a continuous data cloud of cell states. Red outline signifies gating of data cloud from stable cell types that would be used for trajectory analysis

gorithms also tend to overfit smaller datasets, producing topologies with superfluous branches (Setty et al., 2016; Zunder et al., 2015). Thus, MST-based tools have shown utility only in well-defined systems such as hematopoiesis, where a previously determined correct solution can be selected from an ensemble of solutions that includes 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 (Trapnell et al., 2014). Other approaches such as ECLAIR take a cohort of MSTs generated from the same dataset and attempt to extract a consensus tree from the most common connections (Giecold et al., 2016). 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, non-linear embedding, incorporates non-linear dimensionality reduction techniques to deconvolute difficult to interpret high-dimensional data into more approachable 2-3-dimensional representation. Unlike PCA, which assumes linear combinations of features can approximate the original data, non-linear embedding assumes the data cloud in mathematical space lies on a non-linear manifold, which is a mathematical topological space (sphere, torus, etc.) that preserves the distances of points in close proximity. t-SNE is one such non-linear embedding approach, but different classes of algorithms have different assumptions regarding the nature, distribution, and shape of the data cloud. Unlike t-SNE that non-linearly 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 (Coifman *et al.*, 2005) which is used in various algorithms such as Wishbone (Haghverdi et al., 2015; Setty et al., 2016), local linear embedding used in SLICER (Welch et al., 2016), and multi-dimensional scaling and mapper applied in scTDA (Rizvi et al., 2017). Adoption of non-linear 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, non-linear 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, multi-branching trajectories to be robustly depicted.

2.5. 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 (Zunder et al., 2015), SPRING (Briggs et al., 2017), is a graph visualization strategy where a densely connected network in multidimensional space is redistributed in a lower dimensional space (e.g., in 2D) by considering edges as weighted springs and using physical laws to simulate the equilibrium position of nodes as an energy minimization problem. Algorithms can be differentiated by whether or not cells are clustered or whether and what type of prior dimension reduction has been performed. Force-directed layout resolves the problem of stochasticity of MST algorithms by using multiple connections to guide the layout. Yet, the interconnectedness of the graphs makes it difficult to analyze cellular transitions outside of visualization, given that all cell states in the graph will be connected to multiple other cell states. A significant advantage, however, is the possibility of representing non-acyclic structures, such as loops that occur in cell cycle state transitions (Briggs et al., 2017; Rizvi et al., 2017). Another new algorithm, Monocle2 (Qiu et al., 2017), utilizes a process called reverse graph embedding to construct pseudo temporal trajectories in an unsupervised fashion (Figure 2.3). Monocle2 is currently the most widely used next generation algorithm for trajectory analysis capable of producing multi-branching trees. In principle, Monocle2 iteratively embeds data points, in a process similar to k-means clustering, into multiple principal curves (Hastie and Stuetzle, 1989). Instead of learning clusters of cells, Monocle2 learns multiple principal curves connecting into a spanning tree that reflects a transitional hierarchy (Figure 2.3). As with other techniques, Monocle2 works best with expert guidance, as multiple parameters that significantly affect the output must be specified. These parameters tune the fit of the principal curves in mathematical space. For example, including 2 (default) or 10 (arbitrarily set) principal components greatly altered the number of cell lineages that can be identified (Figure 3.20). Monocle2 results have been demonstrated to be robust on multiple runs and different parameters on singly-bifurcating trajectories.

2.6. Downstream Analysis of Reconstructed Trajectories

Once trajectories are generated by reconstruction algorithms, there are a 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 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 (Seita and Weissman, 2010), whereas a shallow, star shaped topology can be interpreted as pre-patterning where individuals from a seemingly homogeneous pool of progenitor cells (identified by RNA or protein) are already fated towards cell types (Notta et al., 2016; Paul et al., 2015) 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 (Han et al., 2004; Shen-Orr et al., 2002; Yook et al., 2004). Second, a common analysis is to plot and visualize relative changes in analyte expression values over a pseudo-time course (Marco et al., 2014; Setty et al., 2016). This type of analysis can be done over separate branches to identify mechanisms of maturation (Trapnell et al., 2014)

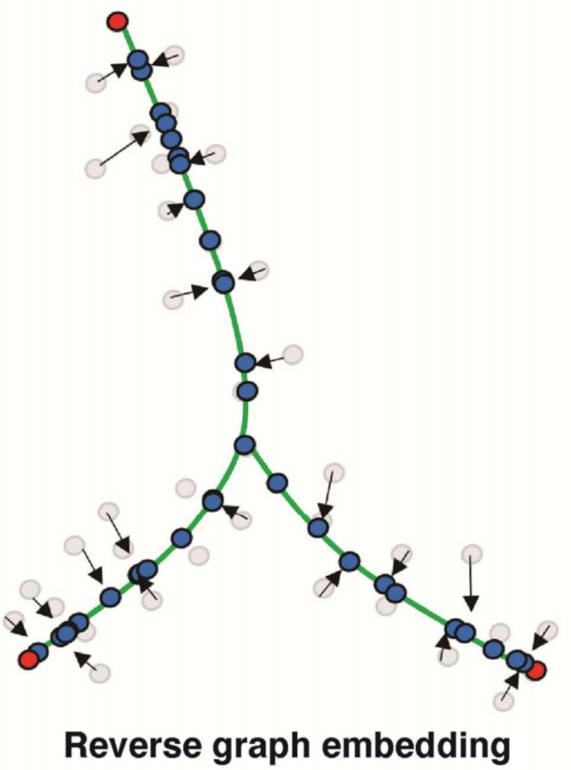


Figure 2.3: Reverse graph embedding. Monocle2 embeds the data cloud into a graph composed of principal curves. Then projects (arrows) individual cells onto principle curves.

or over branch points to reveal mechanisms of cell-fate decisions where a cell must choose between two or more unique differentiation routes (Qiu et al., 2017). Manifold alignment algorithms, such as MATCHER, facilitate integrated comparisons between different trajectories (different routes/ different data types depicting the same route, etc.) with different cell state and temporal units (Welch et al., 2017). Third, differentially expressed gene analysis along trajectories can be performed. In this case, however, instead of looking at genes differentially expressed between two conditions, one would group genes together on the basis that they exhibit similar dynamics over a pseudo-time course (e.g., transient versus sustained expression). The hypothesis is that genes 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 (Matsumoto and Kiryu, 2016; Moignard et al., 2015; Rizvi et al., 2017; Shin et al., 2015).

2.7. 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. Here, I have listed a few points of consideration when applying these methods. It is important to note that these considerations apply equally to p-Creode.

- Pseudotime currently has no real correspondence to real time. The number of cell states that recapitulates a trajectory reflects the number of cells sampled and not the length of transition time. A longer branch more likely reflects 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 reduce 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 more detailed and less susceptible to effects from outliers for common cell types than rare ones. For instance, a 0.1% occurrence of a rare cell type in a 4000-cell data set will only be represented by 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. There are several experimental factors that should be considered. For trajectory analysis algorithms, the key assumption is that transitioning cell states are represented by collected cells. Thus, whether tissue is being harvested during embryonic development versus adult will greatly affect the interpretation of results. For instance, pancreatic islet development is completed by E16.5. Thus, an adult pancreatic dataset 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 (Giecold et al., 2016). More recently, next generation approaches that leverage mutational scars, such as those induced by CRISPR, have been developed for accurately determine if individual cells belong to the same lineage in the classical, parent-child sense (Frieda et al., 2017; McKenna et al., 2016). These approaches can potentially be combined with single-cell approaches to combine cell-state transitional information with parent-child lineage data.

- Often when information is referred to in the context of single-cell analysis it is generally understood to mean measured quantities from the transcriptome or proteome. How this information is obtained varies platform to platform. Each of single-cell platform collects information in particular ways, which can affect downstream analysis by altering the distribution of data points in analyte space. Therefore, any algorithm benchmarked on one platform may not perform equally well on another platform.

CHAPTER 3

Validation of Algorithm

3.1. p-Creode Maps Synthetic Single-cell Data into the Correct Trajectories

p-Creode aims to identify consensus routes from relatively noisy single-cell data. Conceptually, p-Creode determines the geometric shape of a collection of dense data points (i.e., a data cloud) in order to reveal the underlying structure of transitional routes. Similar to other pseudo-progression analyses, this algorithm assumes a continuous transition process, but a dense dataset (on the order of thousands) is required to capture switch-like transitions. We created a synthetic, single-cell dataset in 2-dimensional space that recapitulates a multi-branch hierarchy (5 end-states, 3 branch-points) with realistic noise (Figure 3.1). We developed and applied p-Creode on this visually and analytically tractable dataset to identify the 5 end-states and 3 branch-points.

The p-Creode algorithm consists of 6 steps (Figure 3.1, see Chapter 4 for detailed overview of algorithm):

1. A density-dependent down-sampling is performed to empirically normalize the representation of rare cell states with overrepresented ones (Qiu et al., 2011).

2. A density based k-nearest neighbor graph is constructed using each data point representing a single cell as a node in the graph. The d-kNN graph is constructed by connecting each node with its k nearest neighbors, ranked by Euclidean distance, where k is determined by the relative density of cells in the neighborhood of each node.

3. The graph attribute, closeness, is used to identify stable end-states (stem or differentiated cell types). Closeness is a graph measure of node centrality, defined as the inverse sum of its pairwise graph distances to all other nodes in the graph, where graph distance is the summed stepwise weighted edge distance between two nodes. In p-Creode, nodes

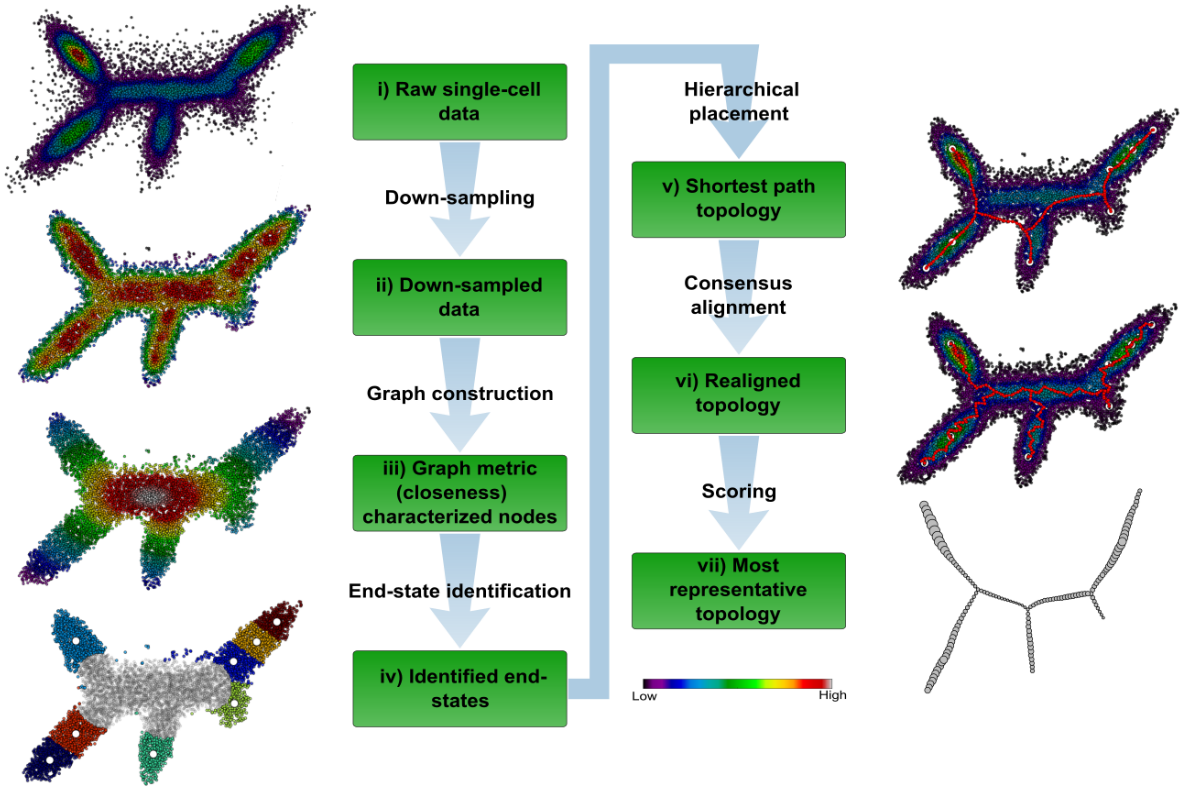


Figure 3.1: Overview of p-Creode Algorithm for Analyzing Single-Cell Data. (i) Synthetic dataset representing single cells in two-dimensional expression space with five end states and three branch points. Overlay represents density of cells. (ii) Density-normalized representation of the original dataset from downsampling. Overlay represents the density after downsampling. (iii) Density-based k-nearest neighbor (d-kNN) network constructed from downsampled data. Overlay represents the graph measure of closeness centrality derived from the d-kNN network, which is a surrogate for cell state (low, end state; high, transition state). (iv) End states identified by K-means clustering and silhouette scoring of cells with low closeness values (¡mean). The number of end-state clusters is doubled to allow for rare cell types. End-state clusters are colored, and open circles represent the centroid per cluster. (v) Topology constructed with a hierarchical placement strategy of cells on path nodes between end states (red), which allows for the placement of data points along an ancestral continuum. Overlay represents the original density of cells. (vi) Aligned topology (red) with maximal consensus though iterative assignment and repositioning of path nodes using neighborhood cell densities. (vii) Representative topology extracted using p-Creode scoring from an ensemble of *N* topologies. Node size in the output graph represents the original density of cells.

with high closeness represent relatively unstable transition states (see Section 4.3 for details), while nodes with low closeness represent relatively stable end-states. End-states are identified by K-means clustering of nodes with low closeness. The optimal cluster number (K), as determined by scoring, is then doubled to allow representation of rare cell types.

4. A weighted shortest path topology is constructed with a novel hierarchical placement strategy of nodes within the paths between end-states (Figure 4.1). Unlike hierarchical clustering where all data points appear on the leaves of a dendrogram, this method places data points on branches to allow depiction of ancestral relationships.

5. Positions of nodes included in the shortest path topology are then aligned through iterative assignment and repositioning using all node positions, which determines consensus routes between end-states from all available data.

6. Leveraging stochasticity generated by resampling, we obtain an ensemble of *N* topologies that is scored by a GromovHausdorff-inspired metric to depict the relative robustness of the data in supporting the computed topology (Figure 4.4). This newly developed metric has general applicability outside of p-Creode to assess graphs with simultaneously differing node positions and connections. The validity of the p-Creode metric is demonstrated in Figure 4.5. The topology most representative of the ensemble is selected based on this metric.

3.2. p-Creode Produces Accurate and Precise Trajectories of Hematopoiesis

We applied p-Creode to publicly available mass cytometry data generated from normal human bone marrow (Bendall et al., 2011), as the well-defined process of hematopoiesis usually serves as a first litmus test for pseudo-progression analysis algorithms. The mass cytometry dataset is composed of approximately 240,000 cells evaluated by a reagent panel of 13 cell-surface markers that describes hematopoietic differentiation (CD45, CD45RA, CD19, CD11b, CD4, CD8, CD34, CD20, CD33, CD123, CD38, CD90, CD3). t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten et al., 2011) and manual gating generated several groupings of well-known cell types (Figure 3.2). A large portion of cells remained unidentified, as previously described (Amir et al., 2013), which we considered to be potential transitional cell states. Application of p-Creode on this dataset ($N = 100$) generated topologies that correctly delineated the known hematopoietic differentiation hierarchy (Miyazaki et al., 2014), with HSCs giving rise to myeloid and lymphoid progenitor trajectories, lymphoid cells transitioning into CD3+ T cells and CD19+ B cells, and CD3+ T cells further branching into CD4+ helper T cells and CD8+ cytotoxic T cells (Figure 3.2, 3.3). We then compared our results to a MST-based algorithm, SPADE, which grouped similar cells into populations, but the cell transitions inferred by the MST connecting these populations were inconsistent with known hematopoietic differentiation (Figure 3.2C, 3.4A). For example, MST-defined hierarchies placed CD8+ T cells as a transitional cell type in the B cell trajectory (Figure 3.2C - left), and, in another instance, T cells (lymphocytes) and myeloid cells shared a common trajectory (Figure 3.2C - right). These results demonstrate the utility of p-Creode in generating accurate results for inferring cellular hierarchies from single-cell data. To assess the precision of the p-Creode algorithm, we evaluated whether resampled runs generate quantitatively similar topologies. Hematopoietic differentiation in the normal bone marrow is a relatively well-regulated process, and thus we expect to extract a conserved topology with multiple runs on the same dataset. We leveraged data down-sampling to produce resampled runs on both SPADE and p-Creode (Figure 3.2C-D, 3.4). In general, p-Creode generated topologies consisting of 133 nodes, and we compared its performance with SPADE, first also with 133 nodes (clusters), and then with 200 nodes corresponding to the default SPADE setting at Cytobank.org and the published number of nodes used in the original analysis (Bendall et al., 2011). Resampled topologies generated by p-Creode were notably more similar than those produced by SPADE (Figure 3.2C-D, 3.4). To quantitatively and statistically assess this similarity, we developed a GromovHausdorff-inspired scoring metric (called p-Creode score) that captures the dissimilarity of one topology to another given changing node positions and connections (Figure 4.4, 4.5). At either node number, p-Creode topologies had both lower mean p-Creode scores and lower variances compared to SPADE (Figure 3.2E), demonstrating p-Creode to be a more robust method that generates a lower number of outlier topologies. While p-Creode demonstrated lower variance, a larger than expected range of variance was observed with both algorithms. The source of this variance is believed to stem from technical noise, given that this was one of the first mass cytometry datasets to ever be published on this scale. These results demonstrate the ability of p-Creode to derive well-established cell transition relationships from single-cell data in an unsupervised and robust manner.

3.3. p-Creode Analysis of Differentiating Thymic T Cells Reveals Protein Expression Dynamics

Wishbone is a novel non-linear embedding algorithm for analyzing single bifurcation events, given a user-defined starting cell (Setty et al., 2016). We aimed to compare results generated by p-Creode, which is capable of producing multi-branching hierarchies, to Wishbone, which is designed to analyze protein dynamics over a process containing a single branch point. Wishbone requires prior knowledge of an existing, single branch point, and thus, it differs from p-Creode, which is an unsupervised tool. In the original Wishbone manuscript, mass cytometry datasets of mouse thymus were used for reconstructing T cell development where lymphoid progenitor cells bifurcate into CD4+ helper T cells and CD8+ cytotoxic T cells. These datasets were also used for benchmarking Wishbone to existing trajectory reconstruction algorithms, such as Diffusion Map, SCUBA, and Monocle (Haghverdi et al., 2015; Marco et al., 2014; Trapnell et al., 2014). In line with these comparisons, we applied p-Creode to analyze the same mouse thymus mass cytometry datasets from the Wishbone manuscript (Figure 3.5, 3.6), using as input the same 14 cell-surface markers (CD27, CD4, CD5, CD127, CD44, CD69, CD117, CD62L, CD24, CD3, CD8, CD25, TCRb, CD90). p-Creode generated topologies that reflected canonical T cell development with CD4-/CD8- double negative (DN) (1-4) cell states

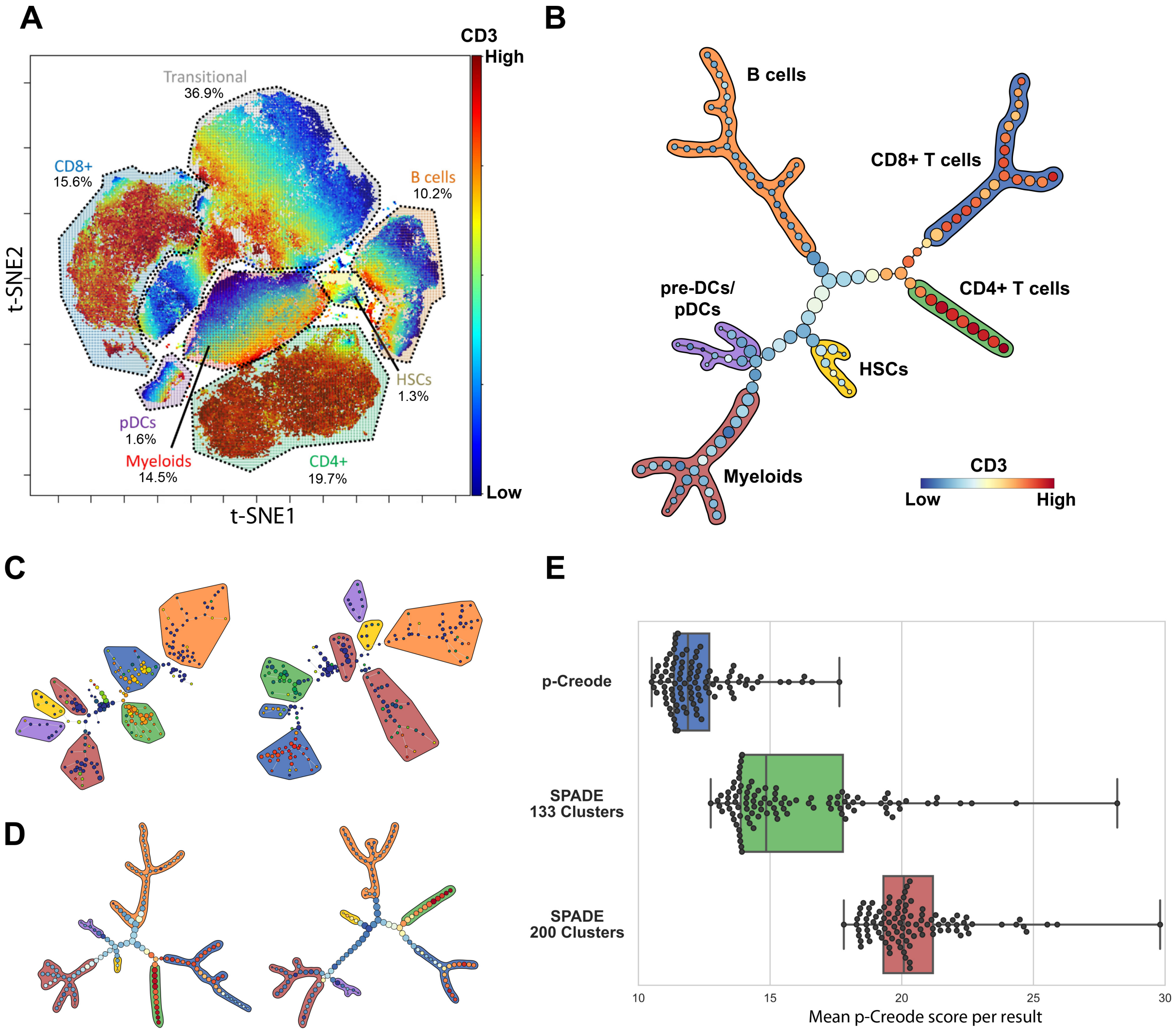


Figure 3.2: p-Creode analysis of single-cell mass cytometry data identifies the hematopoietic differentiation hierarchy. (A) t-SNE analysis of a 13-marker panel mass cytometry dataset from Bendall *et al*. Cell types, as defined by clusters on the t-SNE map, were manually annotated. Overlay represents CD3 levels. (B) p-Creode analysis of the same dataset in A. The most representative graph over $N = 100$ runs, as defined by the graph with the minimum p-Creode scoring metric when compared to all graphs in the analysis, is represented. Colored outlines indicate cell types defined in A, and overlay indicates CD3 levels. (C) Two random runs of the same dataset in A using SPADE (200 nodes), with the same color scheme for cell types and overlay in B. (D) Two random runs ($N = 100$) of the same dataset in A using p-Creode, with the same color scheme for cell types and overlay in B. (E) Comparison of the robustness of p-Creode, SPADE run with 133 nodes, and SPADE run with 200 nodes. Each data point represents the mean p-Creode score calculated for each resulting graph ($N = 100$). Boxes show the quartiles while whiskers show the minimum and maximum scores.

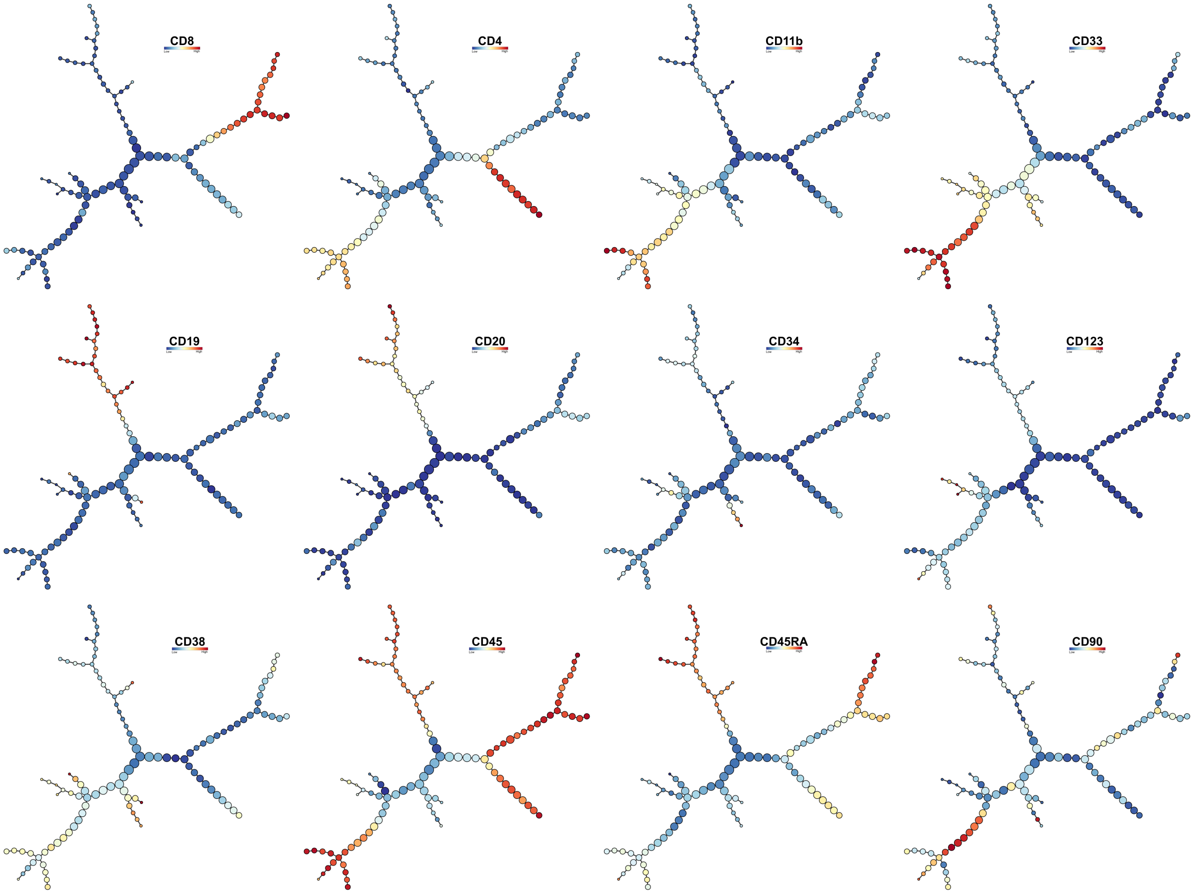


Figure 3.3: Overlay of different protein markers on p-Creode of hematopoietic cell specification in the bone marrow. Overlays represent ArcSinh-scaled mass cytometry data from Bendall *et al*.

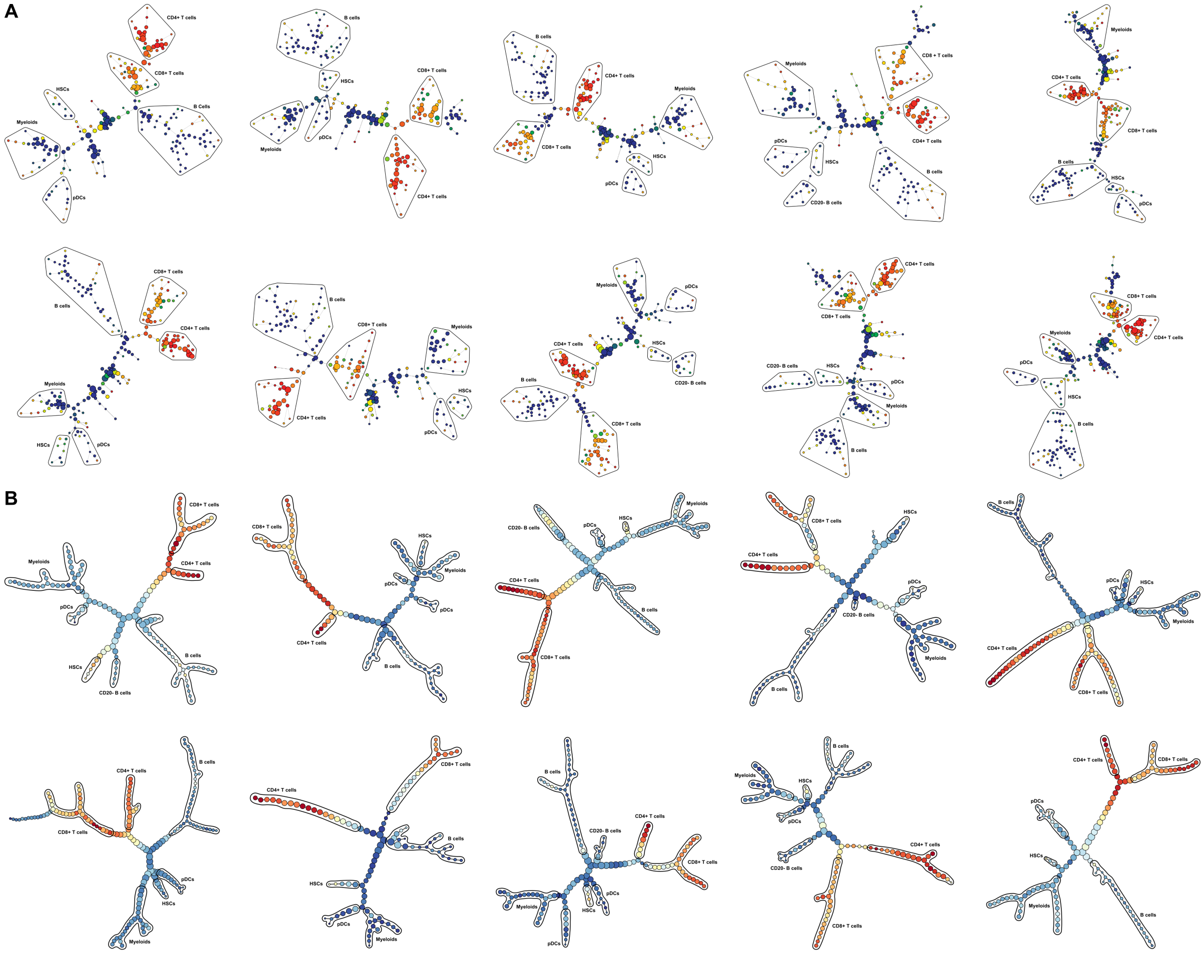


Figure 3.4: Robust computational results generated by p-Creode compared to minimum-spanning tree. (A) Ten randomly selected resampled runs of SPADE (200 nodes) of the hematopoietic dataset. (B) Ten randomly selected resampled runs of p-Creode of the same dataset. Overlay represents CD3 level.

progressing thru CD4+/CD8+ double positive (DP) cell states before finally bifurcating into the CD4+ and CD8+ single positive (SP) cell types (Figure 3.5A-C, 3.6) (Koch and Radtke, 2011), a result that was also obtained by Wishbone. However, unlike Wishbone, p-Creode does not require user knowledge of a single branch point in the hierarchy, thus allowing for the discovery of uncharacterized branches. Based on this feature, p-Creode identified a T cell population branching from the DN to DP transition area (Figure 3.5A, 3.6A), which was also identified by t-SNE (Figure 3.5D population X), but not by Wishbone. This population, marked by the expression of CD24 and CD27, may represent progenitors developing towards gamma delta T cells (Fahl et al., 2014; Ribot et al., 2009).

We next focused on the dynamics of marker expression as a function of cell specification. In this example, p-Creode generated the correct topology for cell specification and also identified the macroscopic dynamics of marker fluctuation (Figure 3.7A-B). However, due to the increased sparseness of cell states in multi-dimensional space, the resolution of marker dynamics, as controlled by the number of cell states within a trajectory, was lower. Diffusion Map is a non-linear embedding tool to fold multi-dimensional data into elongated and compressed shapes (Coifman et al., 2005). Wishbone relies on Diffusion Map as a preprocessing step, and thus we hypothesized that the decreased number of dimensions resulting from this step would lead to a higher resolution of protein marker dynamics by reducing data sparseness. Without additional downstream analysis, Diffusion Map itself was able to capture the branching structure of differentiation, as well as the associated protein expression dynamics (Figure 3.5E), as previously noted for another Diffusion Map-based algorithm (Haghverdi et al., 2015). Application of p-Creode downstream of Diffusion Map enhanced the resolution of protein marker dynamics to a level comparable to Wishbone (Figure 3.5B-C, 3.6B-C) (Setty et al., 2016). We then tested whether Diffusion Map by itself can be used to analyze more complex, multi-branching hierarchies. Application of Diffusion Map to the Bendall *et al.* hematopoietic dataset resulted in convoluted trajectories with illogical transitions between the major differentiated cell types (Figure 3.7C), such as a direct transition between CD4+ T

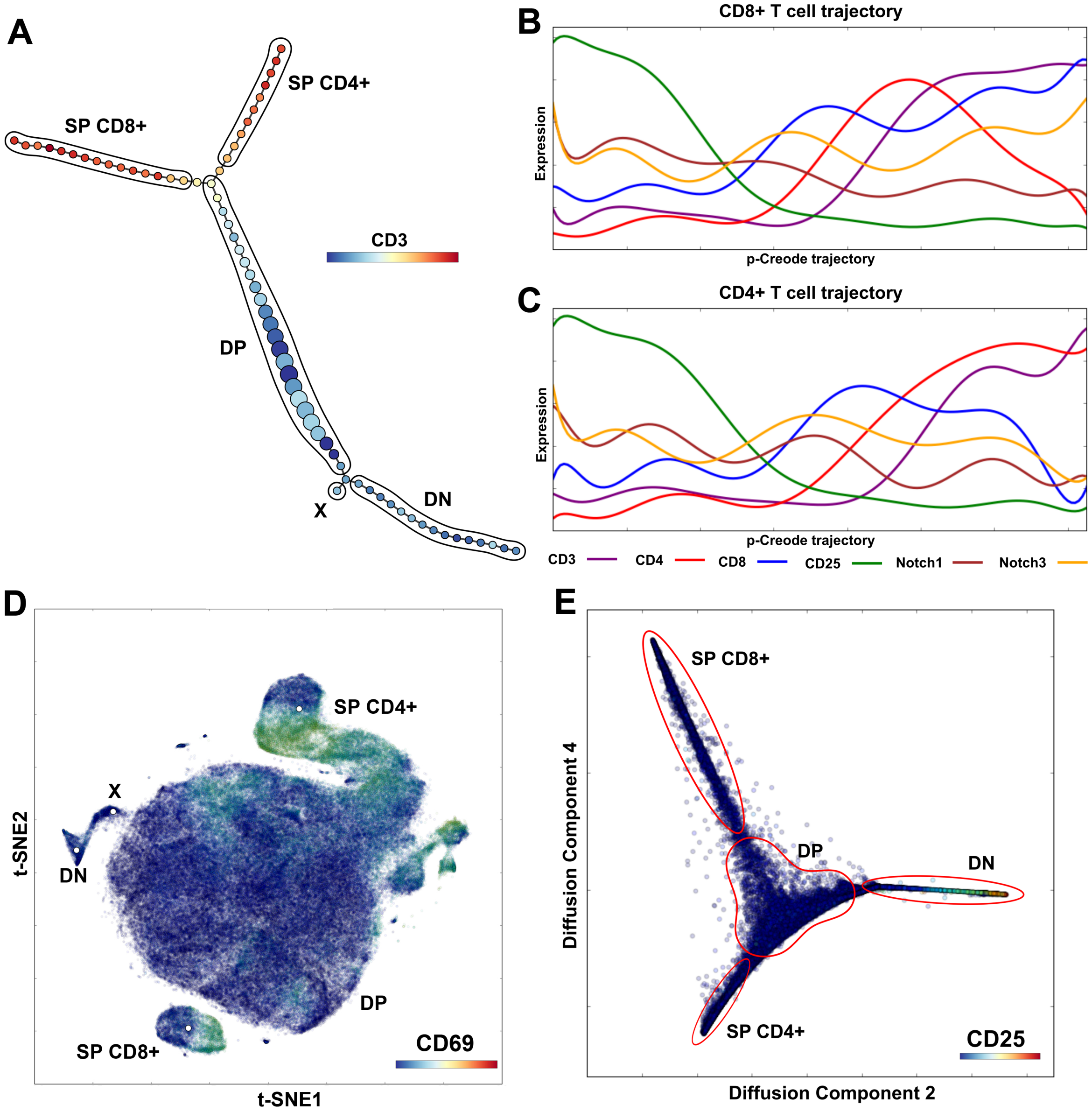


Figure 3.5: p-Creode analysis of single-cell mass cytometry data generates topologies that reflect thymic T cell development. (A) p-Creode analysis of the first replicate 14-marker mass cytometry dataset from Setty *et al.* with PCA preprocessing, representative of $N = 100$ runs. Cell populations were manually labeled. Overlay represents CD3 levels. (B, C) Marker trends along p-Creode trajectories with Diffusion Map preprocessing for CD8+ SP (B) and CD4+ SP (C) trajectories. Trends are similar to results obtained by Wishbone analysis and consistent with established stages of T-cell differentiation. (D) t-SNE analysis of the Setty *et al.* dataset with manual annotation of clusters, including population X identified by p-Creode. (E) Diffusion Map of the dataset depicting T cell maturation.

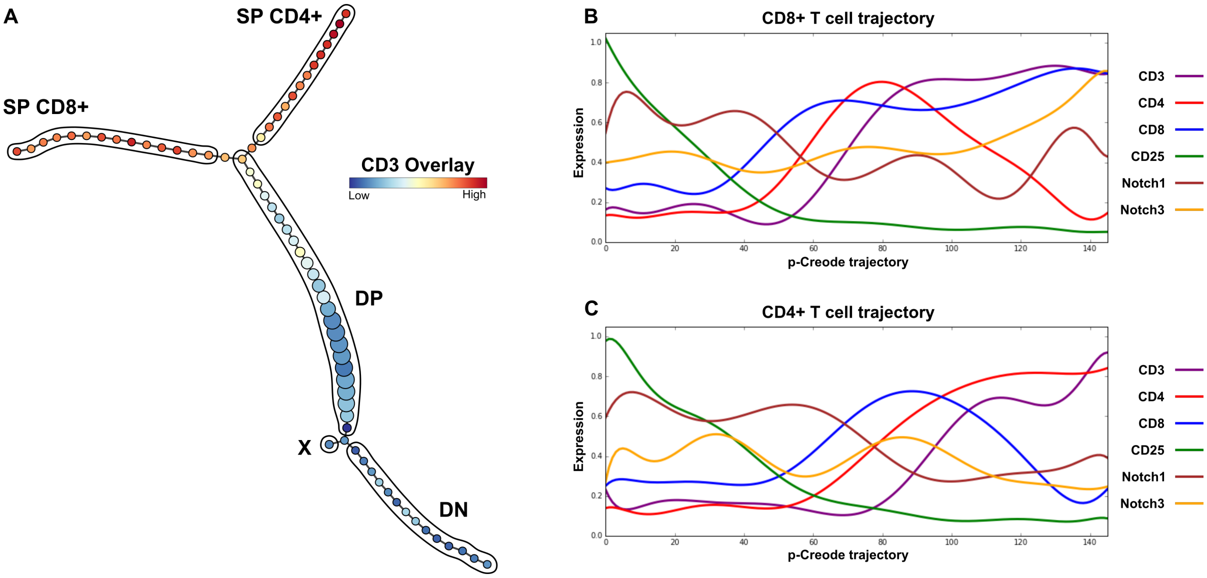


Figure 3.6: p-Creode analysis of single-cell mass cytometry data depicting thymic T cell development. (A) p-Creode analysis of the second replicate 14-marker mass cytometry dataset from Setty *et al*. with PCA preprocessing, representative of $N = 100$ runs. Cell populations were manually labeled. Overlay represents CD3 levels. (B, C) Marker trends along p-Creode trajectories with Diffusion Map preprocessing for CD8+ SP (B) and CD4+ SP (C) trajectories. Trends are similar to results obtained by Wishbone analysis and consistent with known stages of T-cell differentiation.

cells and myeloid cells. These results suggest that Diffusion Maps may not scale well with data depicting multi-branching transitions, while p-Creode is capable of analyzing such data.

3.4. p-Creode Analysis of MxIF Data Generates Robust Topologies Depicting Intestinal Cell Specification

An unresolved issue in single-cell data analysis is the applicability of various algorithms across experimental platforms, such as flow-based or imaging-based methods, that generate data with different distributions. Therefore, we applied p-Creode to derive biological insights from data generated on a different technological platform, MxIF, to analyze intestinal cell transition relationships using single-cell data. MxIF is an iterative fluorescence staining procedure that dramatically increases the number of protein analytes that can be analyzed in a single tissue section (Gerdes et al., 2013). We applied MxIF to generate single-cell data depicting cell specification at homeostasis of the murine intestinal and colonic epithelia, which are continuously renewing tissues fueled by a stem cell-driven process (van der Flier and Clevers, 2009). Similar to hematopoiesis in the bone marrow, transitioning cell states (known as transit-amplifying or TA cells in the gut) are present at any snapshot in time, but they are poorly characterized and lack specific markers (Buczacki et al., 2013; van Es et al., 2012; Tetteh et al., 2016). We used MxIF with a 18-marker panel that broadly covers the stem-to-differentiated cell spectrum (Hopx, PCNA, Lgr5(GFP), Sox9, Survivin, CK20, Chromogranin A, DCLK1, Lysozyme, Muc2, p-EGFR(Y1068), Ki67, Villin, β-Catenin, NaKATPase, pan-Cytokeratin-PCK26, CD44v6, S6), with the assumption that multiple marker combinations can delineate transitioning cell states.

Mature cell types can be identified with canonical markers, such as Muc2 marking goblet cells, DCLK1 marking tuft cells, Villin marking enterocytes, Chromogranin A marking enteroendocrine cells, Lysozyme marking Paneth cells. Combinations of p-EGFR, Hopx, and Sox9 marked distinct TA cells (Figure 3.8A-B). More importantly, the spatial resolution afforded by MxIF allowed the direct visualization of transitioning cells above the bottom of the

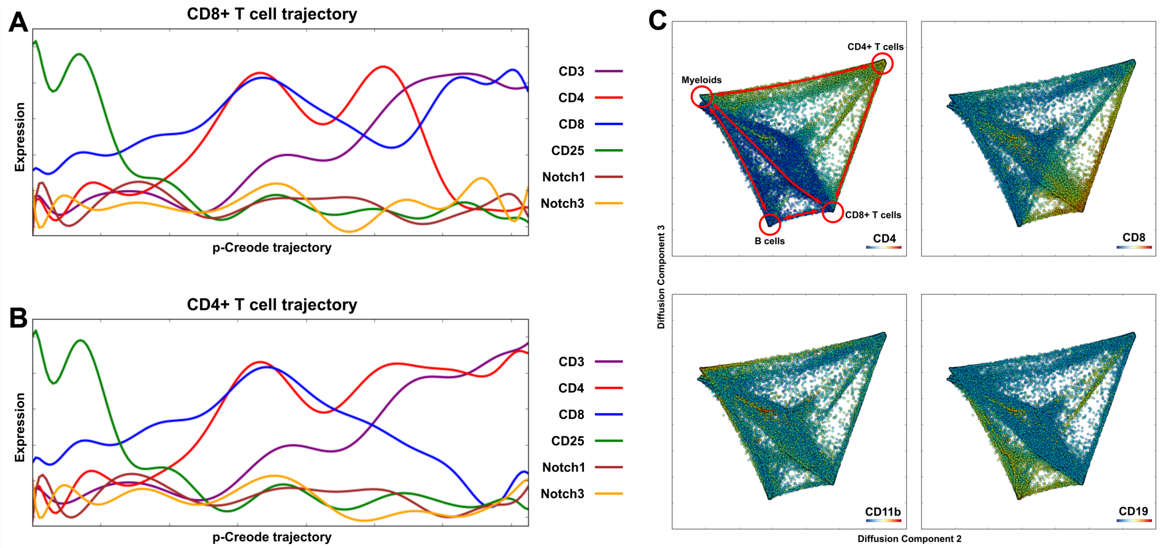


Figure 3.7: Macroscopic dynamics of cell states in branching trajectories. (A, B) Marker trends along p-Creode trajectories with PCA preprocessing for CD8+ SP (A) and CD4+ SP (B) trajectories. Due to sparseness of cell states in multi-dimensional space, the resolution of marker dynamics compared to Wishbone is lower. (C) Diffusion Map of the mass cytometry dataset from Bendall et al. depicting hematopoiesis. Illogical transitions between the major differentiated cell types, such as a direct transition between CD4+ T cells and myeloid cells, are depicted by red arrows. Cell types were manually annotated. Overlays represent relative protein expression.

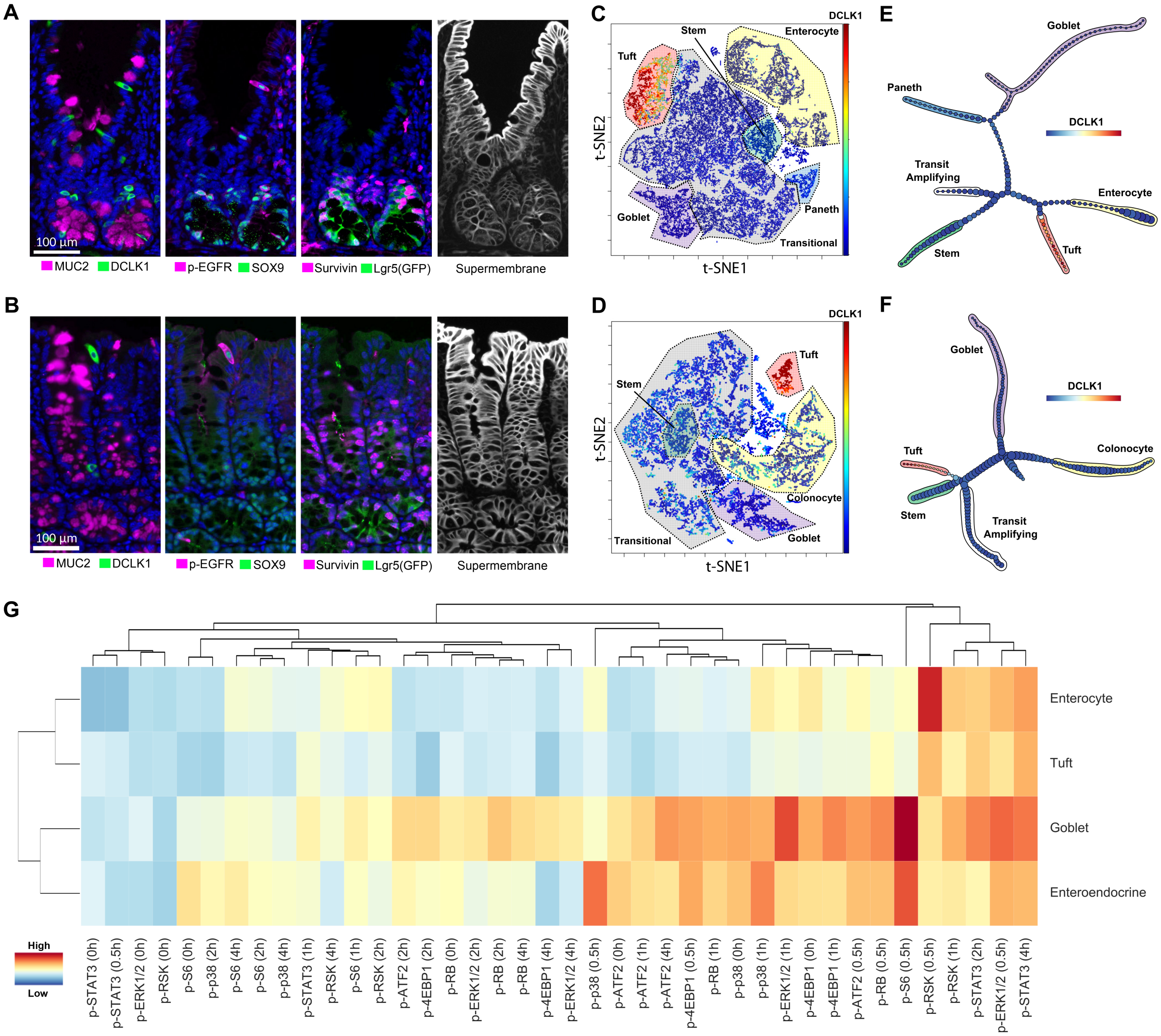


Figure 3.8: p-Creode analysis of single-cell multiplex immunofluorescence (MxIF) data reveals an alternate origin for tuft cells in small intestine versus colon. (A, B) Raw MxIF images where quantitative single-cell data are derived by extracting segmented cell objects using a combined, supermembrane mask. Example staining for differentiated, transit-amplifying (TA), and stem cell markers in the small intestinal (A) and the colonic epithelium (B). (C, D) t-SNE analysis on 19-marker MxIF datasets of the small intestinal (C) and the colonic epithelium (D). Cell types, as defined by clusters on the t-SNE map, were manually annotated. Overlay represents DCLK1 levels. (E, F) p-Creode analysis of datasets in E and F with the most representative graphs over $N = 100$ runs, for small intestine (E) and colon (F). Overlay represents DCLK1 levels. (G) Hierarchical clustering of major epithelial cell types by their response to *in vivo* stimulation by TNF. Clustering on all normalized signals (indicated by heat map) measured by DISSECT-CyTOF (See Figure 3.16).

crypt. For example, Lgr5(GFP) from a reporter mouse marked thin, wedge-shaped stem cells (crypt-based columnar cells or CBCs) (Barker et al., 2007) intercalating Paneth and Paneth-like cells at the crypt base, while Survivin marked CBCs and also transitioning cells in the mid-crypt (Figure 3.8A-B). A full depiction of all markers throughout the crypt-luminal axis of the small intestine and colon is presented in Figures 3.9 and 3.10, respectively. Object segmentation with a super-membrane mask (β-Catenin, NaKATPase, PCK26, CD44v6), preprocessing to remove non-cells, and quantification of single cells were performed as previously described (McKinley et al., 2017). We also applied an additional filter to remove data points (cells) that were gated negative for all markers and, therefore, uninformative to the analysis. Overall, data from 39,000 and 17,000 individual cells acquired from the small intestine and colon, respectively, were analyzed. t-SNE and manual gating applied to the small intestine and colon datasets revealed several groupings of well-known intestinal epithelial cell types, as well as a large portion of unidentified, potentially transitioning cell states in both tissues (Figure 3.8C-D). p-Creode analysis of these datasets with $N = 100$ resampled runs generated topologies with the same terminal cell types identified by t-SNE analysis (Figure 3.8E-F). Furthermore, the topologies connecting these terminal cell types through transitional cells largely resembled the known differentiation hierarchy of the small intestinal and colonic epithelium (Kim et al., 2014). At $N = 100$ runs, robust results were obtained with most of the individual runs generating similar topologies (Figure 3.11 and 3.12). In the small intestine, Lgr5(GFP)+ stem cells were depicted to transit through cell states with variable expression of Survivin, Ki67, PCNA, Sox9, p-EGFR, and Hopx in cells largely residing outside the stem cell zone, as indicated by imaging (Figure 3.13, 3.14). The topology implied a decision between secretory and absorptive lineages in this transitioning zone with secretory progenitors biased towards Hopx and Sox9 (Paneth and goblet progenitors), and absorptive progenitors biased towards proliferative markers (Ki67 and Survivin) (Figure 3.13, 3.14). This bias is supported by studies of Notch activation and inhibition, which controls secretory versus absorptive cell specification associated with proliferation (Fre et al., 2005; Tsai et al., 2014). Secretory progenitors further branched into goblet and

Paneth cells in the intestine, which are known to share a common origin.

Two possible abnormalities were identified from these topologies. First, Chromogranin A+ enteroendocrine cells were not identified, stemming from the extreme rarity of this cell type in our dataset ($< 0.2\%$), which makes them indistinguishable from technical noise in downsampling. The rarity of Chromogranin A+ endocrine cells is supported by a recent study using a Chromogranin A-GFP reporter mouse (Engelstoft et al., 2015). Tuft cells, also relatively rare, make up approximately 1% of all epithelial cells in our datasets and thus were differentiated from noise. Second, cycling cells (Ki67/PCNA+) were identified as an end-state with its own branch, although the location of the branch in the topology was correct (in the TA population close to stem cells) (Figure 3.8E-F). Appearance of additional branches can result from using markers denoting cells in other states (such as in the cell cycle) distinct from the process of interest (cell specification). When we eliminated proliferative markers from the analysis, p-Creode was able to align TA cells, which express Survivin, Ki67, and PCNA, into the correct transitional trajectory between stem cells and differentiated cells (Figure 3.15). Thus, markers selected for the analysis of a specific cell-transition process must be considered since a complex biological system engages multiple processes simultaneously. Overall, p-Creode analysis on single-cell MxIF data was able to generate cell-transition topologies of the gut that are supported by the literature.

3.5. Tuft Cells are Specified Outside the Atoh1-dependent Secretory Lineage in the Small Intestine in Contrast to the Colon

Tuft cells are luminal-sensing epithelial cells recognized as a secretory cell type akin to goblet and Paneth cells. In the p-Creode analysis, tuft cells in the small intestine appeared distinct from the secretory lineage consisting of goblet and Paneth cells, and instead shared a common trajectory with enterocytes (Figure 3.8E). In the colon, however, tuft cells exhibited an alternative trajectory close to stem cells (Figure 3.8F). These results suggest alternate routes for tuft cell development between the small intestine and the colon. To determine if tuft cells in the

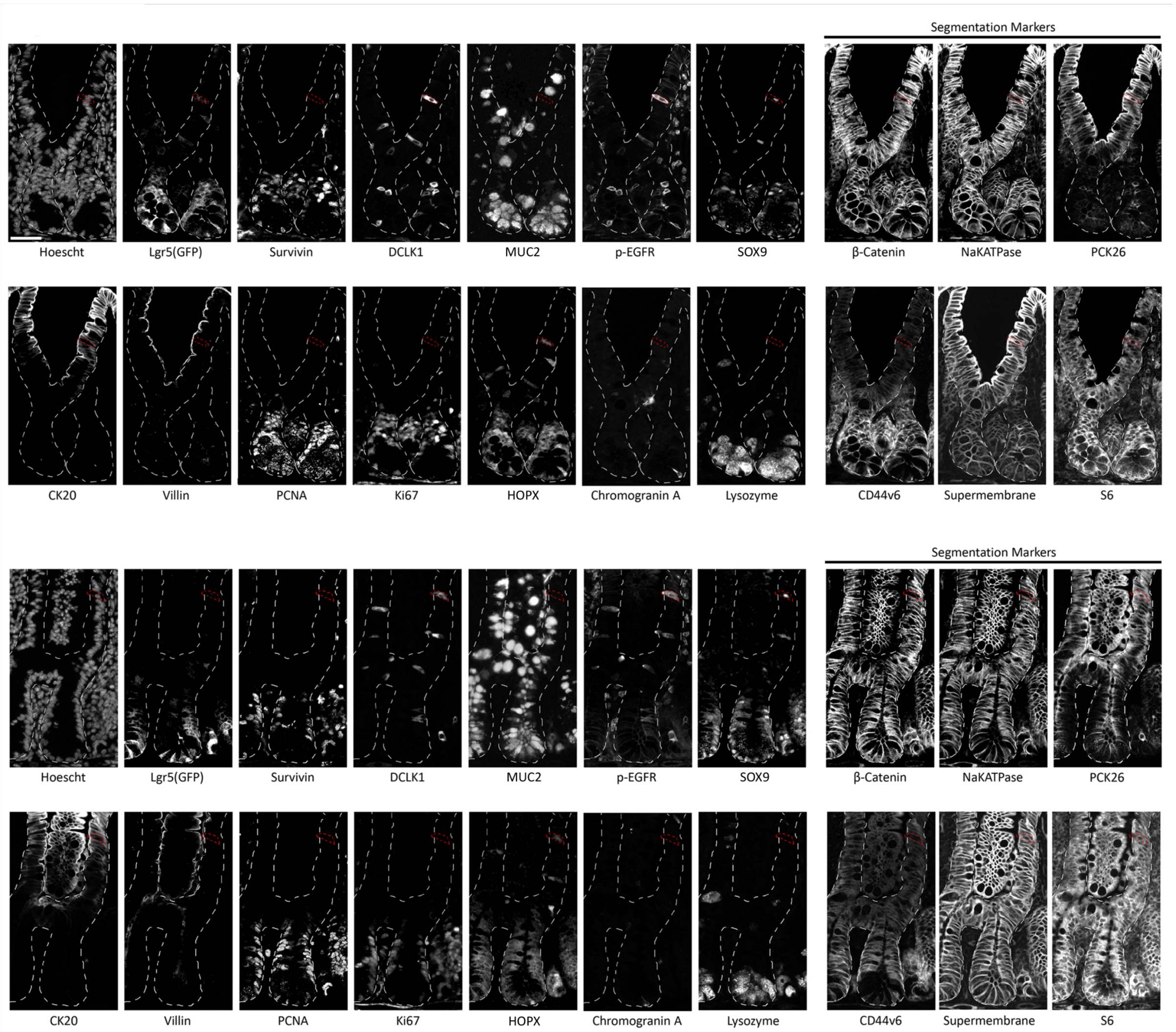


Figure 3.9: Example image data of MxIF from mouse small intestine. Two examples of 19-plex imaging of small intestinal epithelium using MxIF. Dotted lines outline the epithelium within the crypt-villus axis, and red outlines differentiated tuft cells.

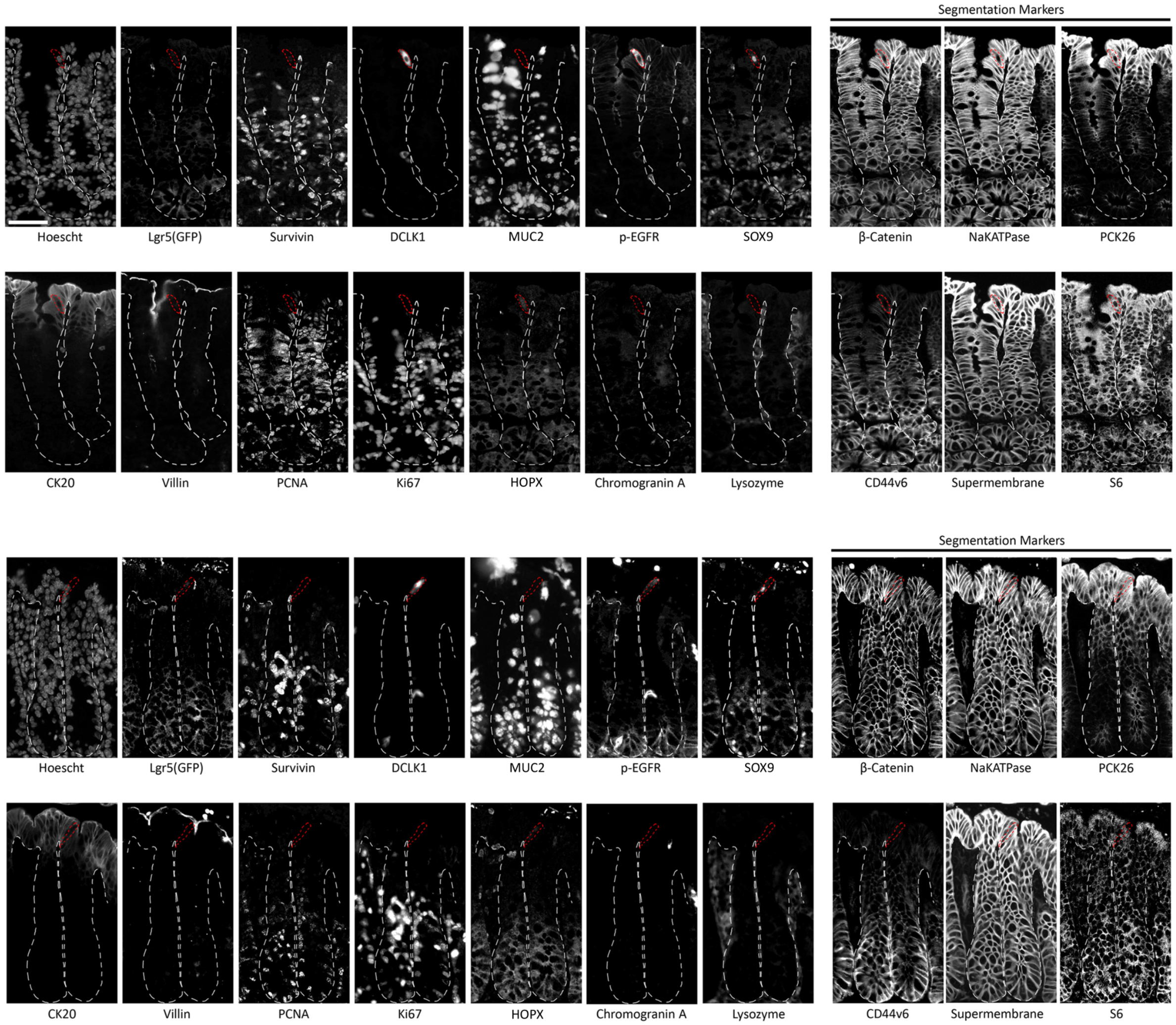


Figure 3.10: Example image data of MxIF from mouse colon. Two examples of 19-plex imaging of colonic epithelium using MxIF. Dotted lines outline the epithelium within the crypt-surface axis, and red outlines differentiated tuft cells.

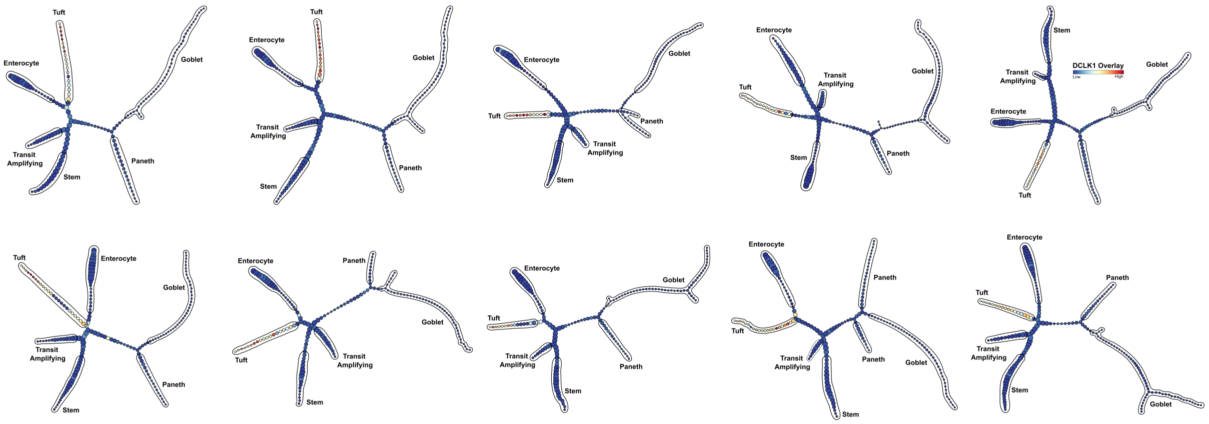


Figure 3.11: Robust sampling of p-Creode results from the gut epithelium. Ten resampled runs of p-Creode on MxIF datasets from small intestine. Overlay represents DCLK1 levels.

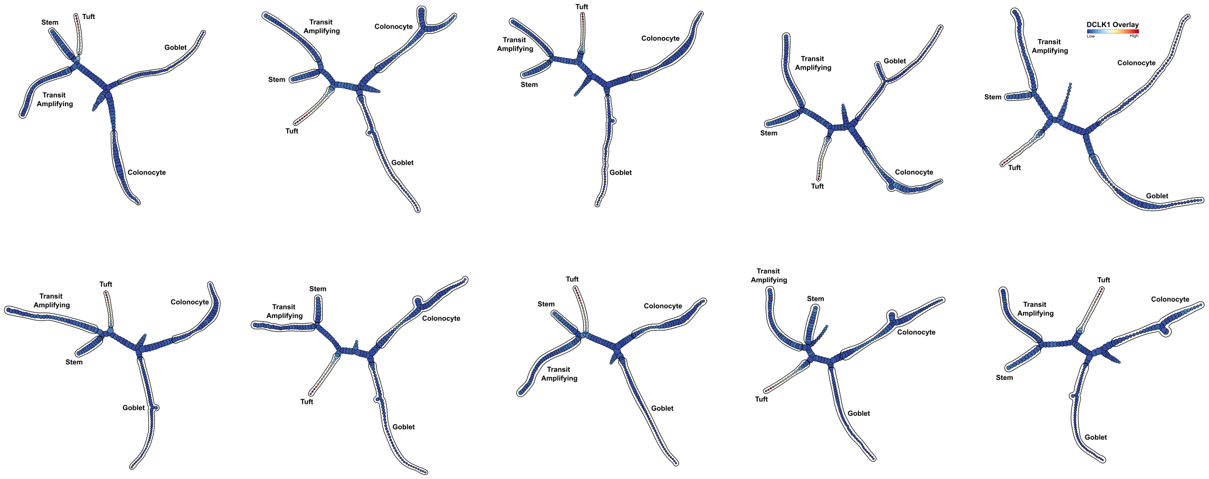


Figure 3.12: Robust sampling of p-Creode results from the gut epithelium. Ten resampled runs of p-Creode on MxIF datasets from colon. Overlay represents DCLK1 levels.

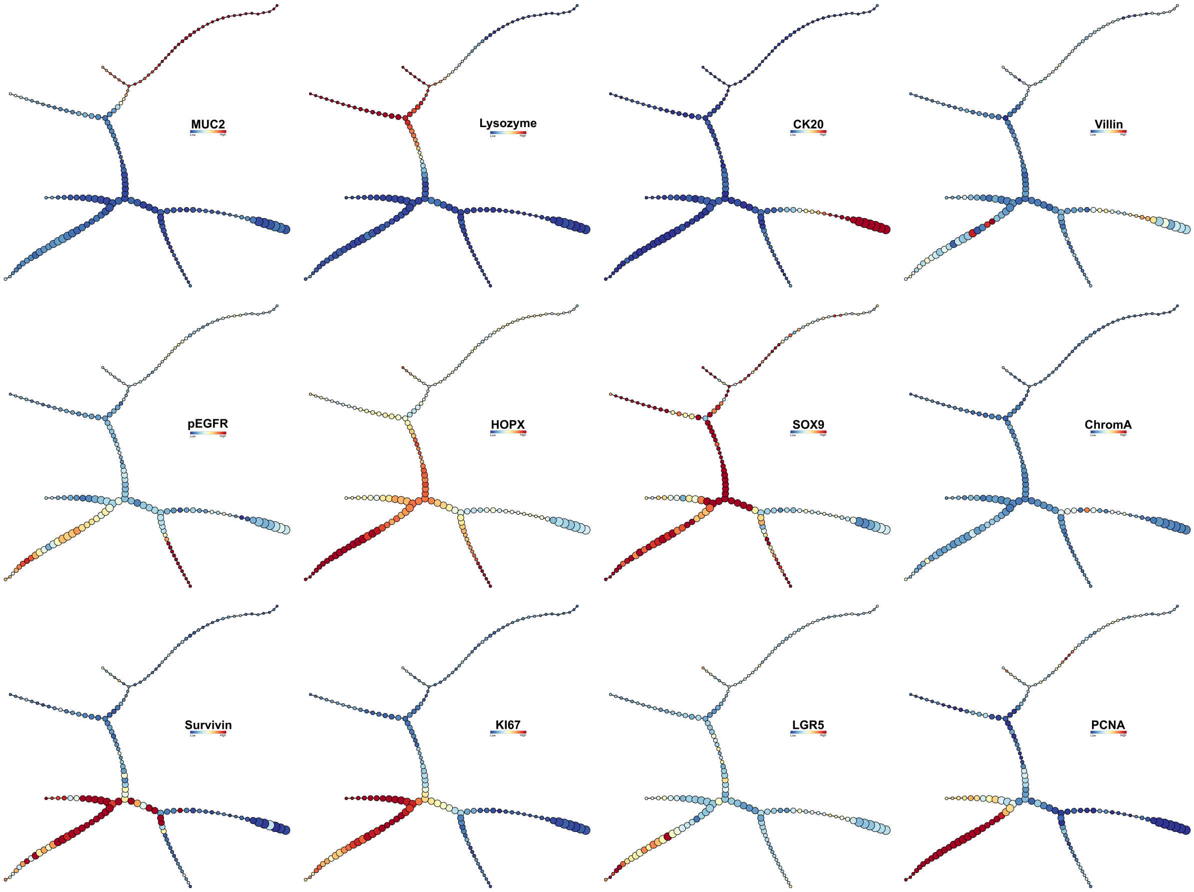


Figure 3.13: Overlay of different protein markers on p-Creode of epithelial cell specification in mouse small intestine. Overlays represent relative protein expression by MxIF.

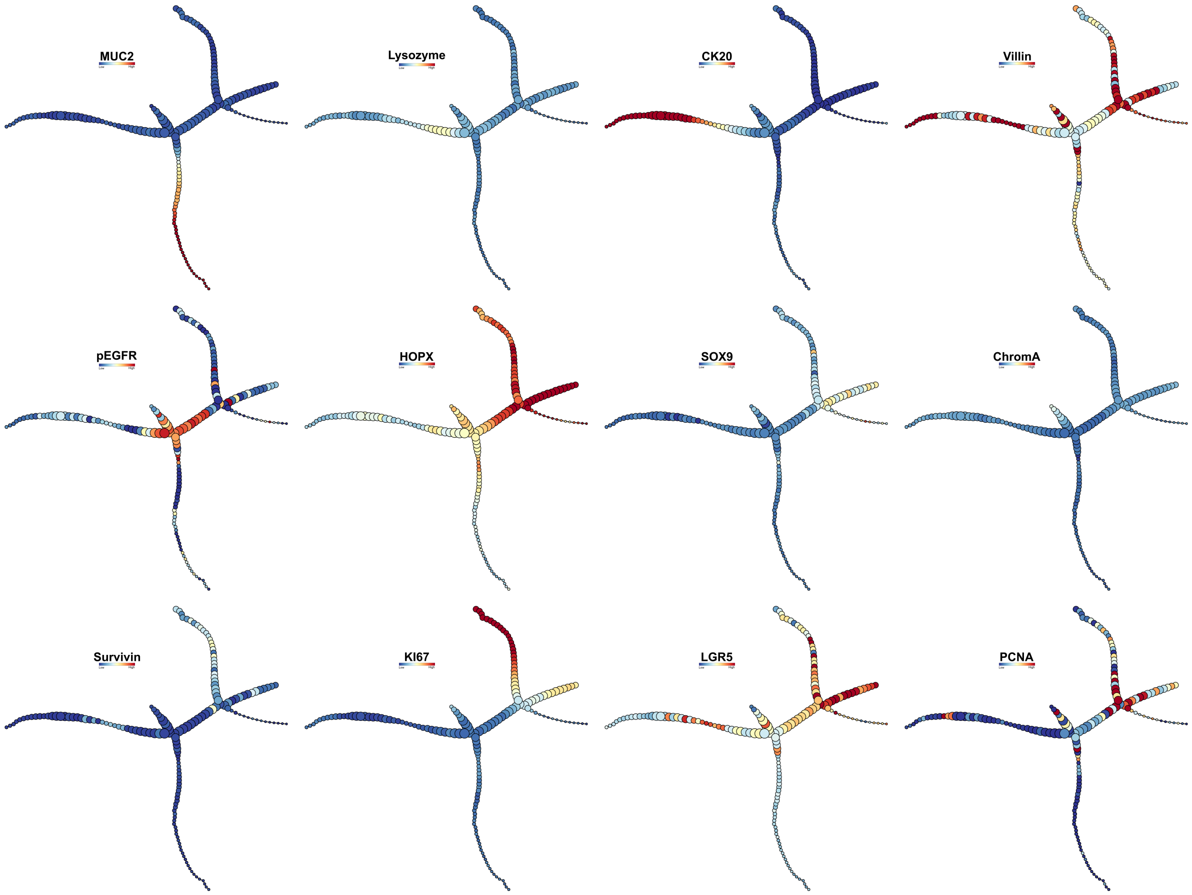


Figure 3.14: Overlay of different protein markers on p-Creode of epithelial cell specification in mouse colon. Overlays represent relative protein expression by MxIF.

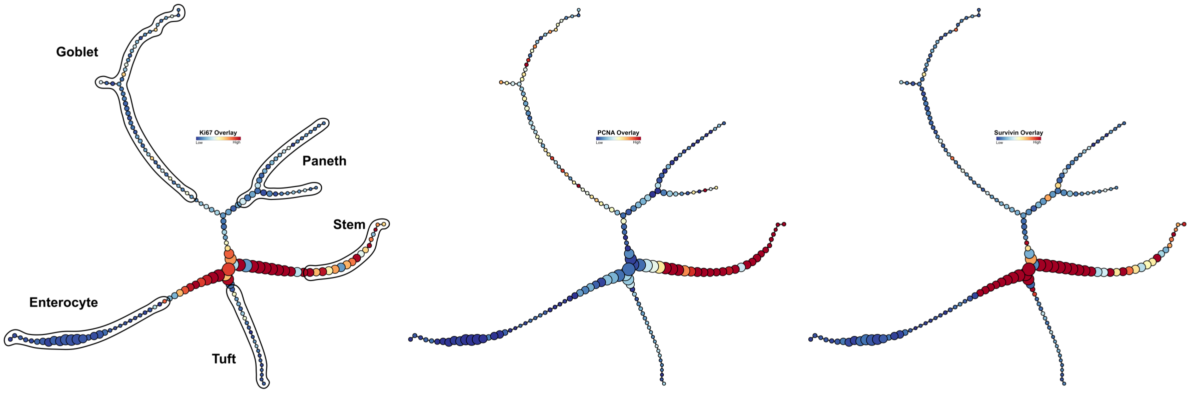


Figure 3.15: p-Creode analysis of small intestinal cell trajectories without proliferative markers. p-Creode analysis constructed without proliferative markers but overlaid with such markers and Survivin (a TA cell marker).

small intestine behave more similarly to secretory or absorptive cells, we evaluated epithelial cell type-specific responses to tumor necrosis factor (TNF) stimulation. Using the DISSECT approach (Simmons et al., 2015), intestinal epithelial tissues were collected over specific time points over a four-hour time course after systemic administration of TNF, disaggregated, evaluated by mass cytometry, and data were gated into different villus cell populations (Figure 3.16A). From these populations, 8 signaling proteins previously determined to respond to TNF were measured. As previously shown (Simmons et al., 2016), TNF elicited stronger signaling responses in secretory cells compared to enterocytes (p-S6, p-ATF2, p-RB, p-p38, p-4EBP1, p-ERK1/2) (Figure 3.16B). Tuft cells shared low signaling amplitudes with enterocytes, as well as similar transient p-ERK and p-RSK dynamics (Figure 3.16B). Summarizing these observations, we used hierarchical clustering on all signaling parameters to determine similarities among cell types. Secretory cells clustered together, as expected, whereas enterocytes and tuft cells clustered together in contrast to their established lineages (Figure 3.8G). These results demonstrate that the signaling behaviors of small intestinal tuft cells over multiple pathways do not resemble secretory cells, consistent with p-Creode results of their origins.

To further validate p-Creode-generated results, we selectively ablated Atoh1, a master transcription factor that regulates the secretory lineage in the intestinal epithelium (VanDussen and Samuelson, 2010). We used the Lrig1CreERT2/+ driver to induce excision of the Atoh1 floxed allele in intestinal epithelial stem and progenitor cells (Powell et al., 2012), generating Lrig1CreERT2/+;Atoh1flox/flox mice. Tamoxifen administration in adult mice resulted in complete ablation of CLCA1+ goblet and Lysozyme+ Paneth cells in the small intestine and CLCA1+ goblet cells in the colon (Figure 3.17A-D, 3.18A-B). In contrast to previous findings (Gerbe et al., 2011), tuft cells, as marked by DCLK1, increased in the small intestine, rather than being suppressed (Figures 3.17A-B, 3.17E, 3.18A-C). These DCLK1 cells are bona fide tuft cells and not stem-like cells, as evidenced by their villus localization, candle-like tufted morphologies, and multi-marker protein signature (McKinley et al., 2017) (Figures 3.17B, 3.17E, 3.18B).

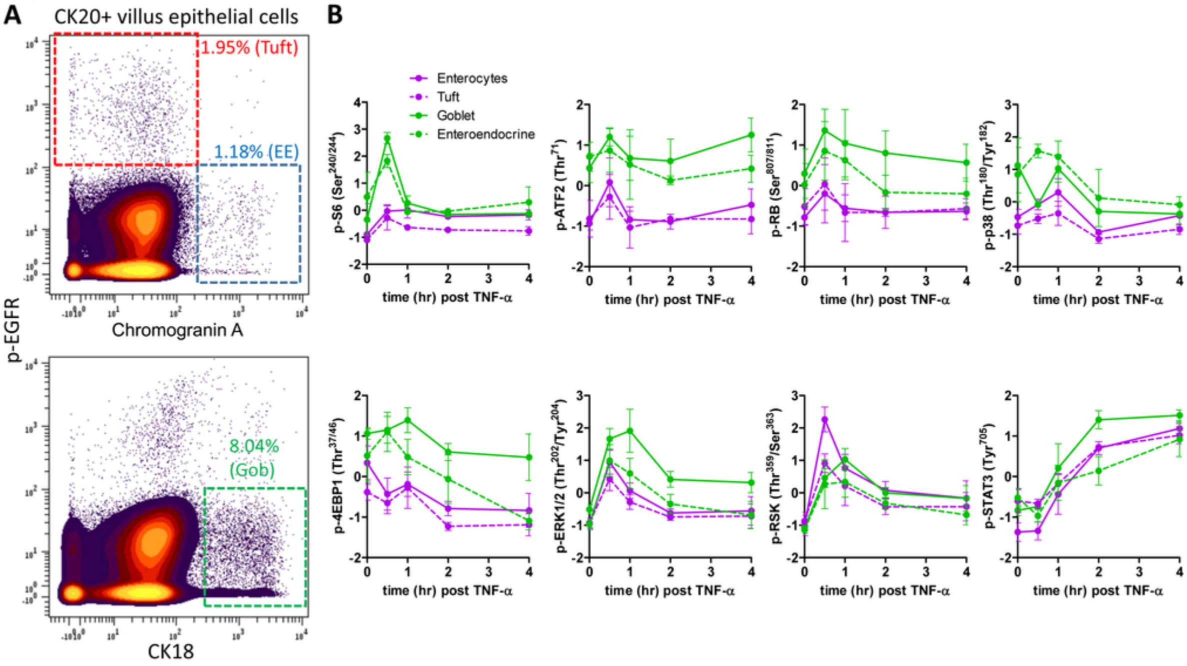


Figure 3.16: Tuft cells respond to exogenous stimulus in a different way compared to other secretory epithelial cells. (A) Gating scheme for differentiated (CK20+ villus) tuft, enteroendocrince (EE) and goblet (Gob) cells in the small intestinal epithelium from DISSECT-CyTOF data. (B) Epithelial cell type-specific time courses of different (8) signaling proteins in response to *in vivo* TNF stimulation. Cell types were gated as described in A, and multiplex data were collected from DISSECT-CyTOF. Error bars represent SEM from n=3 animals. Data scales are Z-score values derived from mean centering and variance scaling of each time course experiment after ArcSinh scaling.

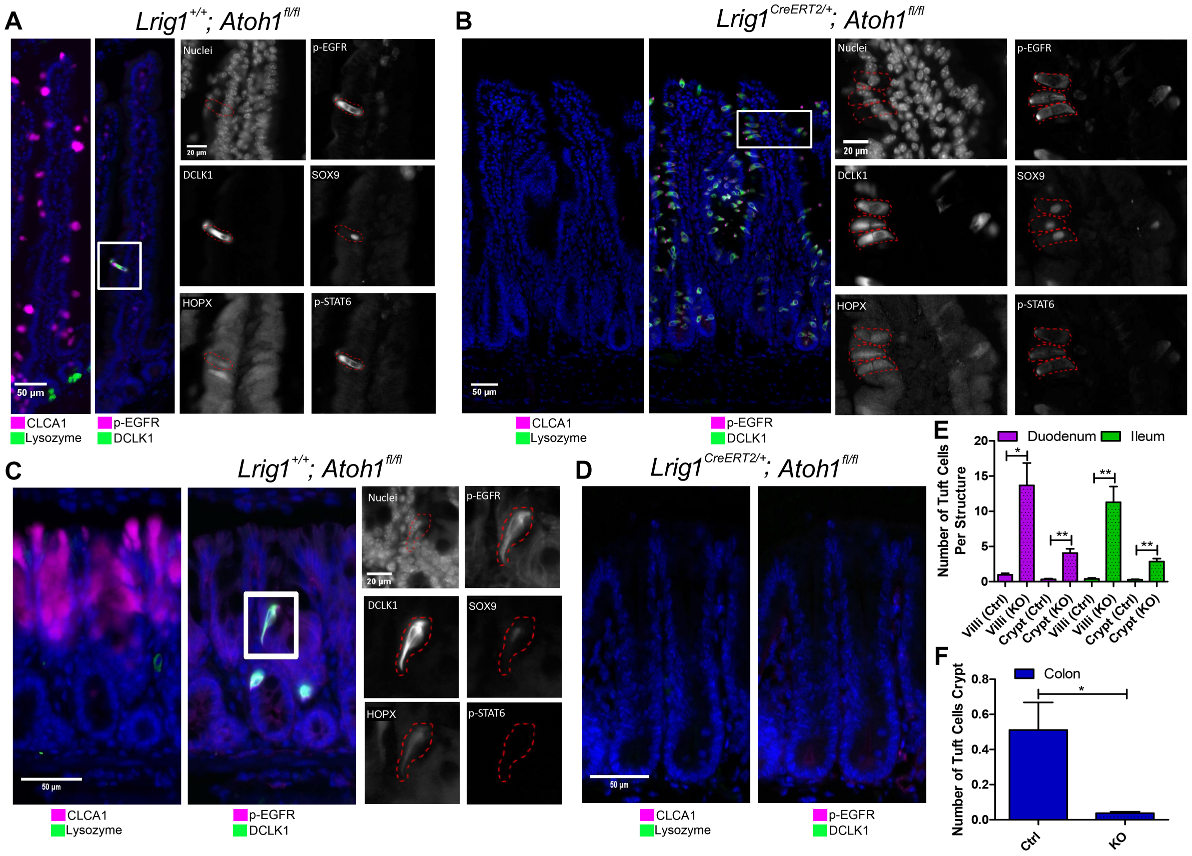


Figure 3.17: Tuft cells have alternative specification requirements in small intestine versus the colon. (A) Control (Lrig1+/+;Atoh1fl/fl + tamoxifen) and (B) epithelial-specific Atoh1 ablated (Lrig1CreERT2/+;Atoh1fl/fl + tamoxifen) duodenum, with acute ablation of Atoh1 at 8 weeks of age and analysis performed 2 weeks later. Analysis of Paneth (Lysozyme+), goblet (CLCA1+), and tuft (DCLK1+; p-EGFR+) cells. Inset represents a multi-marker tuft cell signature of cells on the villi with certain markers (p-STAT6, p-EGFR) demonstrating an apical tuft staining pattern. (C) Control and (D) epithelial-specific Atoh1 ablated colon, analyzed the same way as A,B. (E,F) Quantitative analysis of DCLK1+ cells from images per crypt or villus in the small intestine (E) and colon (F). Error bars represent SEM from $n = 3$ animals. **$P < 0.01$, *$P < 0.05$

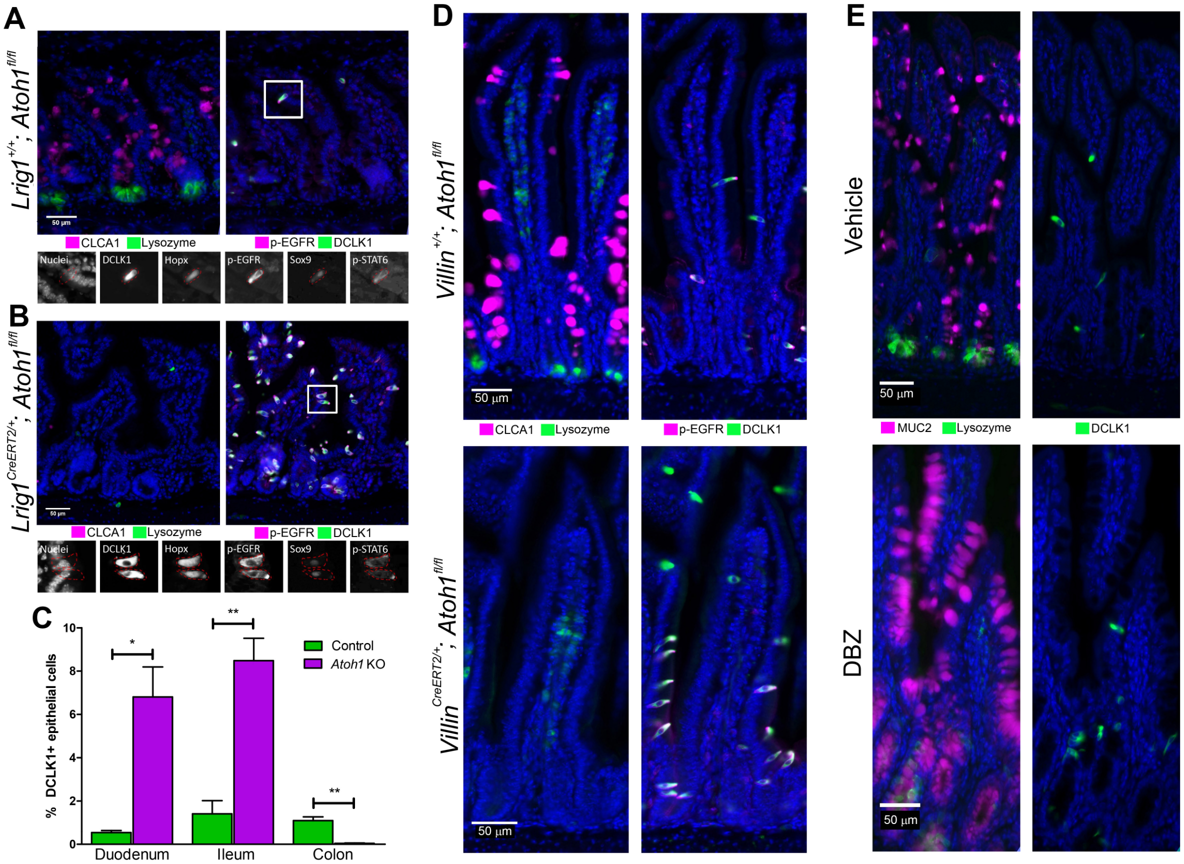


Figure 3.18: Tuft cells specification as a function of Atoh1. (A) Control (Lrig1+/+;Atoh1fl/fl + tamoxifen) and (B) epithelial-specific Atoh1 ablated (Lrig1CreERT2/+;Atoh1fl/fl + tamoxifen) ileum, with acute ablation of Atoh1 at 8 weeks of age and analysis performed two weeks later. Analysis of Paneth (Lysozyme+), goblet (CLCA1+), and tuft (DCLK1+; p-EGFR+) cells. Inset represents a multi-marker tuft cell signature of cells on the villi with certain markers (p-STAT6, p-EGFR) demonstrating an apical tuft staining pattern. (C) Automated image analysis of tuft cell percentage by DCLK1, image segmentation, and processing. Approximately 100,000 cells analyzed for each sample over entire Swiss rolls. Error bars represent SEM from n=3 animals. $**P < 0.01$,$*P < 0.05$. (D) Representative image of control (Villin+/+;Atoh1fl/fl) and epithelial-specific Atoh1-ablated (VillinCreERT2/+;Atoh1fl/fl) duodenum. Analysis of Paneth (Lysozyme+), goblet (CLCA1+), and tuft (DCLK1+; p-EGFR+) cells. (E) Vehicle and DBZ-treated duodenum. Analysis of Paneth (Lysozyme+), goblet (Muc2+), and tuft (DCLK1+) cells.

Because previous work has used a VillinCreERT2/+ driver to induce recombination, we repeated our experiment using VillinCreERT2/+;Atoh1flox/flox tissue, which again resulted in the increase of DCLK1+ cells (Fig. 3.18D). In contrast, dibenzazepine (DBZ), a γ-secretase inhibitor known to inhibit Notch signaling, resulted in complete conversion of the epithelium into secretory cells, yet showed only a slight increase in numbers of tuft cells (Fig. 3.18E) (VanDussen et al., 2012). Since Atoh1 is the most proximal inducer of intestinal secretory progenitors (Buczacki et al., 2013; Kim et al., 2014, 2016; Li et al., 2016; Shroyer et al., 2005), these results again suggest that tuft cells do not descend from the established secretory lineage in the small intestine. Contrary to the small intestine, tuft cells in the colon, marked by DCLK1 expression, were absent when Atoh1 was ablated, responding to Atoh1 loss in a similar fashion to CLCA1+ goblet cells (Figures 3.17C-D, 3.17F). This result suggested that colonic tuft cell specification was indeed controlled by the master secretory cell transcription factor Atoh1, whereas this was not the case in the small intestine. These experiments corroborated our p-Creode assessment of tuft cell specification differences between the small intestine and colon.

3.6. Application of p-Creode on Published scRNA-seq Data Recapitulates Complex Differentiation Processes

In contrast to candidate-based approaches such as mass cytometry and MxIF, scRNA-seq enables unbiased characterization of cells using thousands of transcript analytes. We assessed the performance of p-Creode on scRNA-seq data, using two publicly available datasets that describe (1) lung alveolar epithelial cell differentiation (Treutlein et al., 2014), and (2) blood cell differentiation from myeloid progenitors (Paul et al., 2015). These two datasets were generated using different analytical strategies. The former used Fluidigm C1 to analyze hundreds of data points by FKPM, while the latter used a plate-based MARS-seq platform to analyze thousands of data points by transcript counting. The variety of technology used allowed for the evaluation of the general applicability of p-Creode to scRNA-seq data. For the alveolar differentiation analysis, we combined separate datasets collected at E14.5, E16.5, E18.5 and

P107 (AT2), which covers the progression of development from bipotential progenitors (BPs) to alveolar type 1 and type 2 cells. Data were processed with a neighborhood variance gene selection procedure as described previously (Welch et al., 2016) (see Section 2.2 and Section 6.6). p-Creode analysis, modified for sparse datasets (see Section 4.8), resulted in a characteristic T-shaped topology with a single branch point, depicting the differentiation of BP (Sox11) into type 1 (Pdpn, Ager) and type 2 (Lyz2) cells, consistent with previous analyses (Figure 3.19A-B) (Treutlein et al., 2014; Welch et al., 2016). Importantly, the timing of differentiation from E14.5 to P107 was recapitulated on differentiation trajectories constructed solely from expression data (Figure 3.19A). While p-Creodes design was not intended to run on small datasets ($<$ 500 cells), these results demonstrate that it is possible to adapt this algorithm to perform adequately on sparse scRNA-seq data.

In contrast, the myeloid progenitor dataset contains a large number of data points and thus, it is well-suited for p-Creode analysis. After data processing as above, p-Creode analysis resulted in highly reproducible, multi-branched trajectories (Figure 3.21). In the original paper, multiple cell populations primed for megakaryocyte, erythrocyte, monocyte, and granulocyte development were identified (Figure 3.19C - inset) (Paul et al., 2015). However, previous lineage reconstruction algorithms were not able to place these sub-branches, and only identified the two major branches, the megakaryocyte-erythrocyte (ME) and granulocyte-monocyte (GM) branches, arising from the Cd34+ common myeloid progenitor (CMP) cells (Campbell and Yau, 2017; Setty et al., 2016; Qiu et al., 2017). For example, in Qiu *et al.* Monocle2 was benchmarked (see section 2.4 and Figure 2.3 for more details) against Wishbone, SLICER, and Monocle1 using the Paul *et al.* dataset. While Monocle2 and Wishbone both produced accurate results, trajectory resolution was limited to the two major ME and GM branches (Figure 3.20A) (Qiu et al., 2017). In an attempt to increase the resolution of Monocle2 the number of principle components used in preprocessing was increased from 2 (Figure 3.20A) to 10 principle components (Figure 3.20B). While resolution was gained, resulting trajectory inaccurately placed the megakaryocyte lineage withing GMP branch (Figure 3.20B - Red Arrow). p-Creode, be-

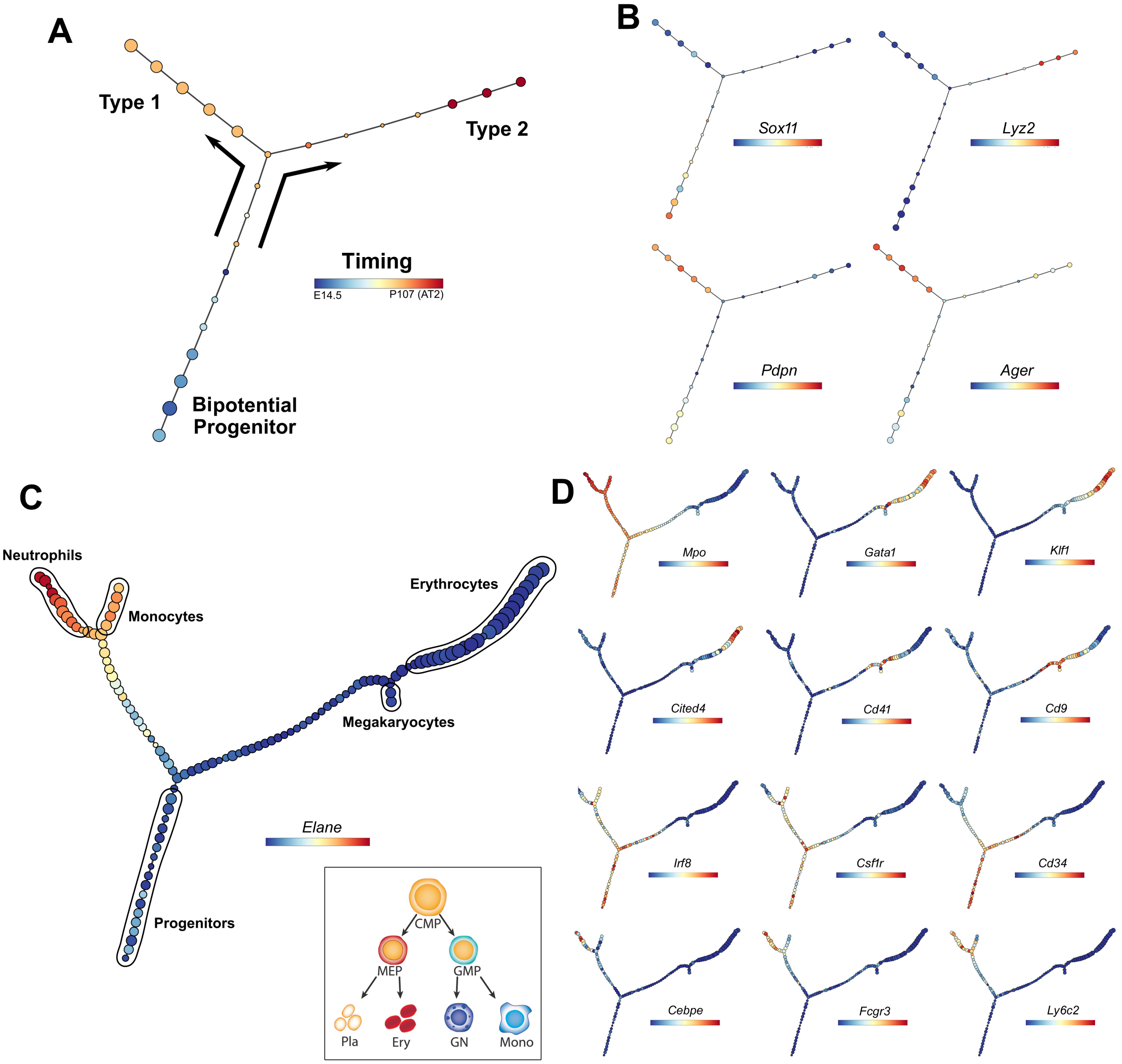
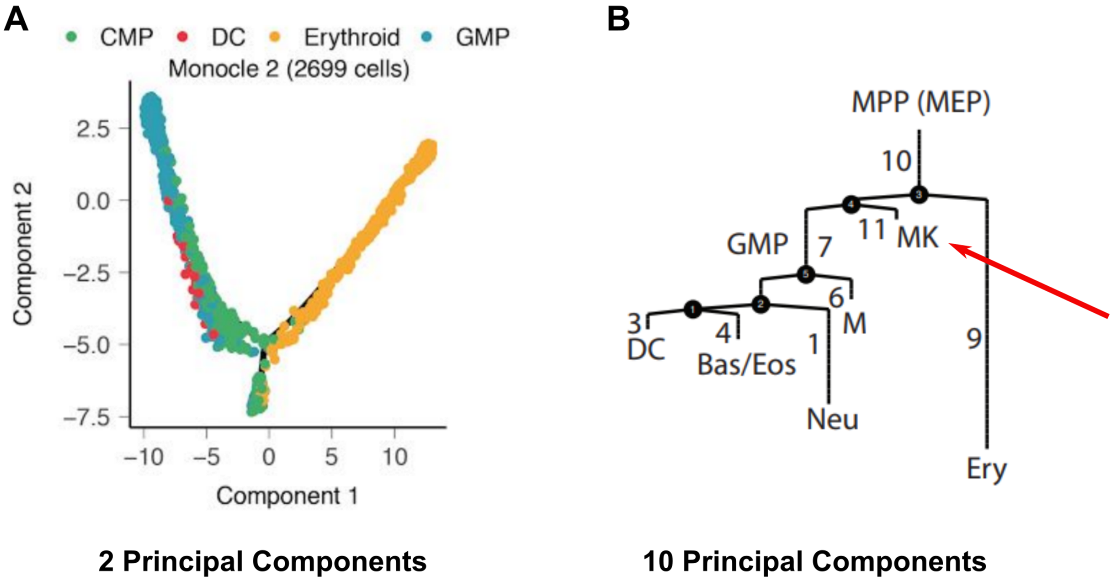


Figure 3.19: Application of p-Creode on published scRNA-seq data reveals multi-branching topologies. (A) p-Creode analysis of the scRNA-seq dataset generated from alveolar cells by Treutlein *et al*. Cells collected over multiple developmental time points were mixed and analyzed together. Overlay represents developmental time that was recovered. (B) Overlay of selected transcripts depicting alveolar cell differentiation on the p-Creode topology generated in A. (C) p-Creode analysis of the scRNA-seq dataset generated from myeloid progenitor cells by Paul *et al*., most representative graphs over $N = 100$ runs. Overlay represents Elane transcript levels. Inset represents an accepted model of myeloid differentiation. (D) Overlay of selected transcripts depicting myeloid cell differentiation on the p-Creode topology generated in (C). Overlays represent ArcSinh-scaled gene expression data.



2 Principal Components

10 Principal Components

Figure 3.20: Monocle2 derived trajectories of myeloid differentiation. (A) Representation of bifurcation of common myeloid progenitors (CMP) into either erythroid cells or granulocyte-macrophage progenitors (GMP), with trajectories inferred by Monocle2 using 2 principle components ((Qiu et al., 2017) - Supplementary Figure 9A). (B) Skeleton of the trajectory learned by Monocle2 describing the lineage relationships learned with Monocle2 using 10 principle components. The numerical labels correspond to the ”State” label of each segment of the tree. Labels where Qiu *et al.* interpretation of corresponding cell type (inferred based on comparison with classifications made in the original study (Paul et al., 2015). MPP (MEP): multipotent pluripotent progenitor or myeloid and erythroid progenitors; MK: megakaryocyte; GMP: granulocyte and monocyte progenitor; DC: dendritic cell; Neu/Eos: neutrophil or eosinophil, Bas: basophil, M: monocyte; Ery: erythrocyte ((Qiu et al., 2017) - Supplementary Figure 16A). Red arrow points to incorrect placement of megakaryocyte cells as part of the GMP lineage. In both figures data first published in Paul *et al.* was used as input to Monocle2.

cause of its ability to map multi-branching trajectories, generated a more complex and accurate topology with further sub-branches arising from major the ME (Gata1, Car2), and GM (Elane, Mpo) branches, including the neutrophil (Cebpe, FcgR3), monocyte (Irf8, Csf1r), erythrocyte (Klf1, Cited4) and megakaryocyte (Cd41, Cd9) sub-branches (Figure 3.19C-D). In line with the down-sampling issue of distinguishing rare cells from noise, eosinophils and basophils, which makes up 0.3% and 0.8% of the cells, respectively, were not identified as end-states. These results highlight the ability of p-Creode for generating multi-branching trajectories from high resolution scRNA-seq data.

3.7. p-Creode Application on scRNA-seq Data Generated from Mouse Colon Reveals Additional Cell Transition Relationships

We then generated scRNA-seq data on the mouse colon with the inDrop platform, which uses droplet-based encapsulation in conjunction with a barcoding strategy analogous to MARS-seq to query thousands of cells (Klein et al., 2015). Using an epithelial enrichment procedure (Sato et al., 2011), single cells were isolated with at least 85% viability (Leelatian et al., 2017b). After additional viability enrichment (see Section 6.4) to $>$ 99% viability, approximately 1900 and 700 colonic cells from two replicates were encapsulated and sequenced. After sequence mapping, barcode deconvolution, and filtering by reads (Klein et al., 2015), 2402 (92%) colonic cells with an average of 49,680 reads per cell were recovered. In line with previous results with inDrop, the doublet rate appeared close to 0% (Figure 3.22A). We then performed t-SNE analysis and observed that the data from the two replicates were largely interspersed within each other, signifying minimal batch effects (Figure 3.22B). t-SNE analysis on these data revealed the presence of progenitor cells, secretory cells, absorptive cells, and immune cells identified by lineage-specific markers (Figure 3.22C-D, 3.23). Immune cells, presumably intraepithelial lymphocytes, were gated out such that only epithelial cells were further analyzed. p-Creode analysis on colonic scRNA-seq data revealed a characteristic cell transition pattern with a stem/progenitor branch (Lgr5, Lrig1, Sox9), an absorptive colonocyte branch (Slc26a3, Car1) and a secretory goblet cell branch (Muc2, Clca1) (Figure 3.22E-F). Progenitor to differentiated cell relationships can be clearly delineated with the pan-differentiation marker Krt20 (CK20). Unlike MxIF which is candidate-based, scRNA-seq afforded additional details regarding cell trajectories. For instance, a Reg4+ goblet cell branch can be seen arising from the secretory lineage (Figure 3.22F). Reg4+ goblet cells were recently identified as deep crypt secretory cells that exhibit niche roles in the colon analogous to Paneth cells in the small intestine (Rothenberg et al., 2012; Sasaki et al., 2016). Similar to Paneth cells, they share a trajectory with goblet cells in the colon in our analysis. These cells appear to arise from a Sox9+ progenitor, and Sox9 is a known transcription factor required for Paneth cell differentiation (Mori-Akiyama et al., 2007). In addition, Atoh1, the master transcription factor for the secretory lineage, was also mapped to secretory cell progenitors.

While p-Creode has the potential to contribute the ongoing debate on the existence of multiple reserve stem cell populations or whether reserve stem cells are dedifferentiated committed cells (Buczacki et al., 2013; Li et al., 2016; Yan et al., 2017), our limited dataset does not allow us to reach a definitive conclusion. Because of the over representation of committed cell states, the resolution required to depict the more nuanced relationships among rare populations of reserve stem cells (approximately 5 cells in a set of $>$ 2000 cells) was lacking. To refine these relationships, it will be necessary to enrich these populations prior to encapsulation in a more targeted analysis. Similar to stem cells, tuft cells were also underrepresented in our dataset (Figure 3.22B). The cells we identify as tuft cells expressed markers Dclk1 and Nrgn (Middelhoff et al., 2017), and also Il25 (data not shown), a cytokine recently identified to be expressed in tuft cells to modulate type 2 immune responses (Gerbe et al., 2016; von Moltke et al., 2015) (Figure 3.22F). Similar to analysis derived from MxIF data, both t-SNE and p-Creode analysis placed the tuft cell lineage close to the stem cell lineage in the colon (Figure 3.22B, E). These results reveal the global structure of cell-state transitions from unbiased scRNA-seq data from the colonic epithelium.

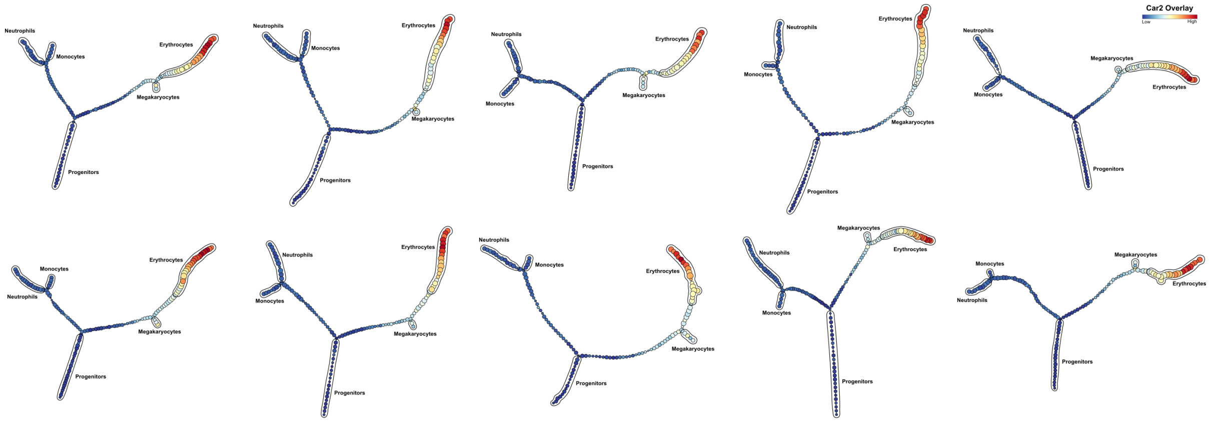


Figure 3.21: Robust sampling of p-Creode results depicting myeloid differentiation. Cells collected over multiple developmental time points were mixed and analyzed together. Ten resampled runs of p-Creode on scRNA-seq from Paul *et al*. depicting a robust myeloid differentiation process. Overlay represents Car2 transcript levels.

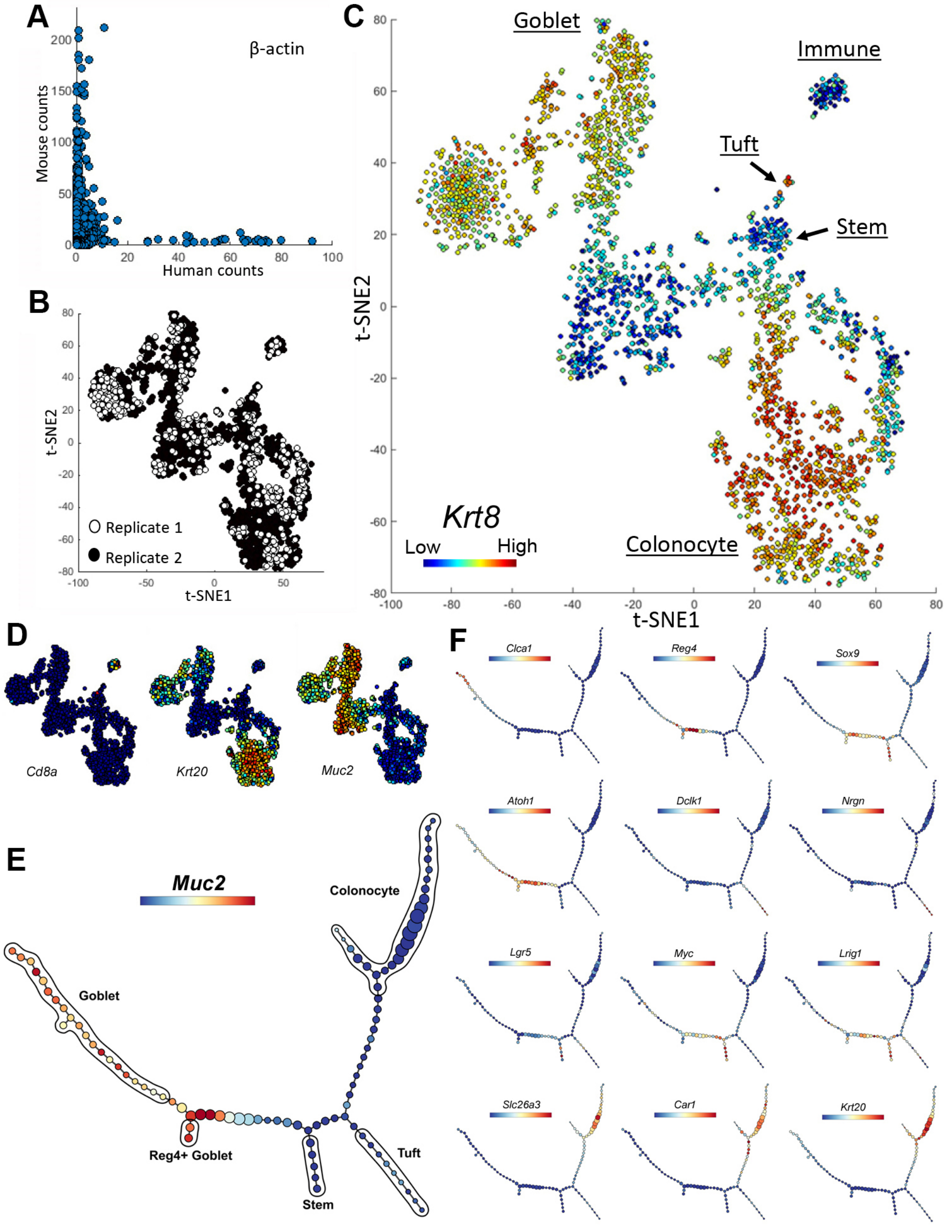


Figure 3.22: inDrop scRNA-seq reveals the development of Reg4+ secretory cells in the murine colon. (A) Human versus mouse -actin transcript count by mapping to human and mouse reference genomes, respectively. Each data point represents a single cell. (B) t-SNE analysis of scRNA-seq data demonstrating the absence of segregation of data points from 2 replicates. (C) t-SNE analysis of murine colonic cells using scRNA-seq data. Cell types, as defined by clusters corresponding to specific cell type markers on the t-SNE map, were manually annotated. Overlay represents Krt8 transcript levels. (D) Overlay of selected transcripts depicting colonic epithelial differentiation and cell type markers on the t-SNE map generated in C. (E) p-Creode analysis of scRNA-seq data generated by inDrop from colonic epithelial cells, most representative graph over $N = 100$ runs. Overlay represents Muc2 transcript levels. (F) Overlay of selected transcripts depicting colonic epithelial cell differentiation on the p-Creode topology generated in E. Overlays represent ArcSinh-scaled gene expression data.

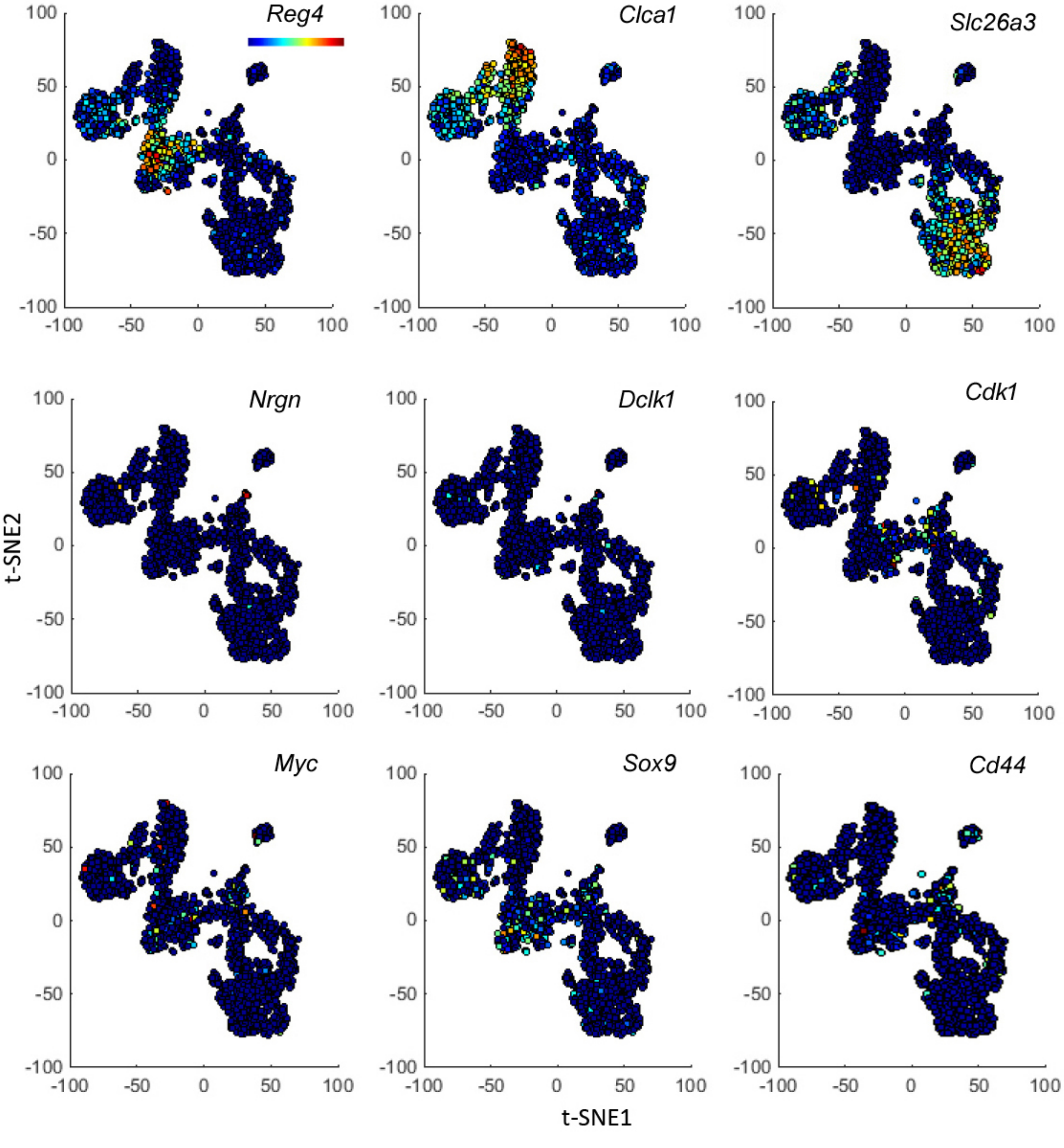


Figure 3.23: scRNA-seq data generated from the colonic epithelium by inDrop. Overlay of selected transcripts depicting colonic epithelial differentiation and cell type markers on a t-SNE map generated from inDrop scRNA-seq data. Overlays represent ArcSinh-scaled gene expression data.

CHAPTER 4

Detailed Algorithmic Overview

The purpose of p-Creode is take inherently noisy single cell data and reveal the robust, underlying structure under such data with *n* cells in *N* dimensional analyte space. The inherent technical variabilities generated by single-cell approaches conceal this structure to varying degrees. This is dependent on the process of study and the technology applied. Each of p-Creode algorithms 6 steps is geared towards managing this issue: i) Down-sampling, ii) Graph construction, iii) End-state identification, iv) Topology reconstruction, v) Consensus alignment, vi) Scoring.

4.1. Down-sampling

As a first step p-Creode performs a density-dependent down-sampling to normalize the representation of rare and overrepresented cell states. Our approach to down-sampling is similar to the approach outlined in (Qiu et al., 2011). To start, a radius (rad) must be calculated to serve as a limiting factor in calculating the local density (LD_i) of each data point (or cell) *i*. *rad* for each dataset is determined by taking the product of a user defined scaling factor (*SF*) and the median minimum distance (*MMD*), where *MMD* is the median Euclidean distance of 5,000 randomly-selected data points (or all data points if sum total is less than 5,000) to their closest neighbors in the complete dataset. The LD_i for each cell is then calculated by counting the number of data points contained in an *N* dimensional sphere defined by the *rad*. Next, a probability (P_i) of inclusion is established for each data point based on LD_i and its relationship to the user provided target density (*TD*) and outlier density (*OD*). If a cell has LD_i lower than or equal to *OD*, the cell is considered noise and not selected. If LD_i is higher than *OD* and less than or equal to *TD* then the cell is automatically selected for inclusion. Finally, if the LD_i is greater than *TD*, P_i is equal to LD_i divided by *TD*:

$$P_i = \begin{cases} \mathrm{P_i} = 0 & \text{if } \mathrm{LD_i} \leq \mathrm{OD} \\ \mathrm{P_i} = 1 & \text{if } \mathrm{OD}¡\mathrm{LD_i} \leq \mathrm{TD} \\ \mathrm{P_i} = \frac{LD_\mathrm{i}}{TD} & \text{if } \mathrm{LD_i} > \mathrm{TD} \end{cases} \tag{4.1}$$

Three user defined input variables are required for performing the down-sampling procedure. The *SF* is used to scale rad, effectively shrinking or growing the volume of the sphere surrounding each data point. Care must be taken to insure that the selected *SF* parameter sufficiently maps the different pockets of density that represent each cell type. Due to biological and technical differences that exist among individual datasets, the distribution of data points will differ between each, causing the value *SF* to differ as well. When available, we used prior knowledge to aid in SF determination. For instance, in a bone marrow dataset (Bendall et al., 2011), we knew that myeloids, CD4+ T, cells and CD8+ T cells were in abundance, which allowed us to check our density calculations by ensuring that each population was adequately represented after down-sampling. When prior knowledge was not available, a histogram of cellular densities was used to visualize the range and abundances of densities, where *SF* parameters producing a large range of populated densities were preferred over values producing a low range of densities or a large range of densities with regions of near-zero densities. The next user-defined parameters, *OD* and *TD*, are responsible for adjusting the probabilities of inclusion, and like *SF*, will vary dataset to dataset. The *OD* is essentially a noise threshold for inclusion; anything with a density below *OD* is considered noise and uninformative. The last user-defined parameter, *TD*, was largely used to control the size of the downsampled dataset. In our case, we selected values of *TD* to produce down-sampled datasets that contained 14,000 cells to allow efficient computational processing.

4.2. Graph Construction

As has been previously reported (Bendall et al., 2014; Setty et al., 2016), spurious relations (edges) between cells (nodes) that are farther apart in development space but closer in analyte data space (short circuits) is a serious issue when attempting to order cellular trajectories. To combat this issue and as noted before (Setty et al., 2016), we leveraged graph distances derived from undirected k-nearest neighbor (kNN) graphs between cells instead of depending on simple distances in analyte data space depicting expression similarities. This process ensures that relationships are specified in regions of data space occupied by data points, instead of short-circuiting between points simply due to lowest pairwise distances. In p-Creode, we have made additions to typical kNN graph construction by incorporating density into k selection and into how edge weights are calculated. In place of a typical kNN, we use a density based k-nearest neighbor (d-kNN). Similar to a traditional kNN graph, the d-kNN graph is built by connecting each node in the down-sampled dataset with its k nearest neighbors, ranked by Euclidean distance, but the value of k in a d-kNN graph is set according to the density of that node, as calculated before being down-sampled. The value of k ranges from 10, for the densest nodes, to 2, for the least dense. These values were chosen as a balance between increasing the risk of short circuits on the high end and limiting the amount of unconnected graph components on the low end. To complete graph construction, isolated graph components are connected by closest inter-node distances, in order to ensure that the final d-kNN graph is one fully connected graph. After graph construction, edges are weighted with a cellular similarity component in the form of Euclidean distance and a cellular transitional likelihood in the form of a density weight. More specifically, weighted edge ($E_{i,j}$) connecting nodes *i* and *j* is defined as the product of the Euclidean distance (*dist*) between the two nodes and a density weight (*dens*):

$$E_{i,j} = dist(i,j) \times dens(i,j) \tag{4.2}$$

where *dens(i,j)* is defined as:

$$dens(i,j) = 1.0 - minimum(nLD_i, nLD_j) + 0.1 \tag{4.3}$$

Here, nLD_i refers to the densities calculated before down-sampling, but normalized (min=0, max=1) over all down-sampled data points to limit the effects of density outliers on the normalization process. A constant of 0.1 is added to insure no edge has a weight equal to zero.

4.3. End-state Identification

A requirement in creating lineage trajectories in some algorithms is the identification of end-states from which to start and end trajectory construction (Trapnell et al., 2014; Setty et al., 2016). This process can be performed manually, but is made harder when multiple end-states in multiple dimensions exist. Another strategy to identify cell states is by clustering, but this is inadequate for continuous data, when data points are artificially forced into clusters even if no identifiable clusters exist. To identify end-states automatically in multi-dimensional data space, we use a graph-theoretical metric from our d-kNN graphs, called closeness centrality, to separate end-states from transition states. Closeness (cls_i) measures the inverse mean graph distance from a node i to all other nodes in a graph, where the graph distance (*gdist*) is the shortest weighted edge path connecting node i to node j as determined by Dijkstras algorithm (Dijkstra, 1959).

$$cls_i = \frac{n}{\sum_j^n gdist(i,j)} \tag{4.4}$$

Nodes with high closeness are more interconnected within a graph and nodes with low closeness are less interconnected. In our graphs, end-states are characterized by low values of closeness, representing geometric fringes in data space similar to (Korem et al., 2015). These data points are less interconnected, having connections with other cell states in a single or few geometric directions. Transitional cell states, on the other hand, are more interconnected in the d-kNN graph from being completely flanked by other transiting cell states, thus, having high closeness. Nodes with low closeness (end-states with less than mean closeness over all cells) are clustered using K-means clustering. The number of end-states is determined by silhouette scoring (Rousseeuw, 1987) over varying values of K, where K is the number of clusters. The optimal K is then doubled to allow for the possibility of underrepresented cell states. From each cluster, a most representative node is selected by finding the data point in the dataset closest to the centroid of the cluster.

4.4. Topology Reconstruction using Hierarchical Placement

Next, p-Creode constructs a representative topology from the identified end-states and selected transition states with a hierarchical placement strategy. Unlike hierarchical clustering where all data points appear on the leaves of a dendrogram, this method places data points on branches to allow depiction of ancestral relationships. p-Creode starts by connecting the two closest end-states (Figure 4.1 step 1 - orange nodes), as measured by graph distance through the d-kNN graph. The connection consists of transition states selected along the shortest path, now defined as path nodes, connecting the end-states. The now connected end-states and any connecting path nodes form a graph component (Figure 4.1 step 2 green nodes) that is added back to the pool of possible connections for the remaining end-states. Connections between nodes in the same graph component are disregarded but all other connections are possible whether they be between an end-state and another end-state (connected or unconnected) (Figure 4.1 - Step 2-3), a connected path node from one component to another connected path node from another component (Figure 4.1 - Step 4-5), or an end-state and a connected path node (Figure 4.1 - Step 5-6). Following these rules the remaining end-states are iteratively connected until all the end-states form one complete graph containing only end-states and path-connecting nodes. Specifically, bifurcation points will be contained with the set of transition states selected to be path nodes.

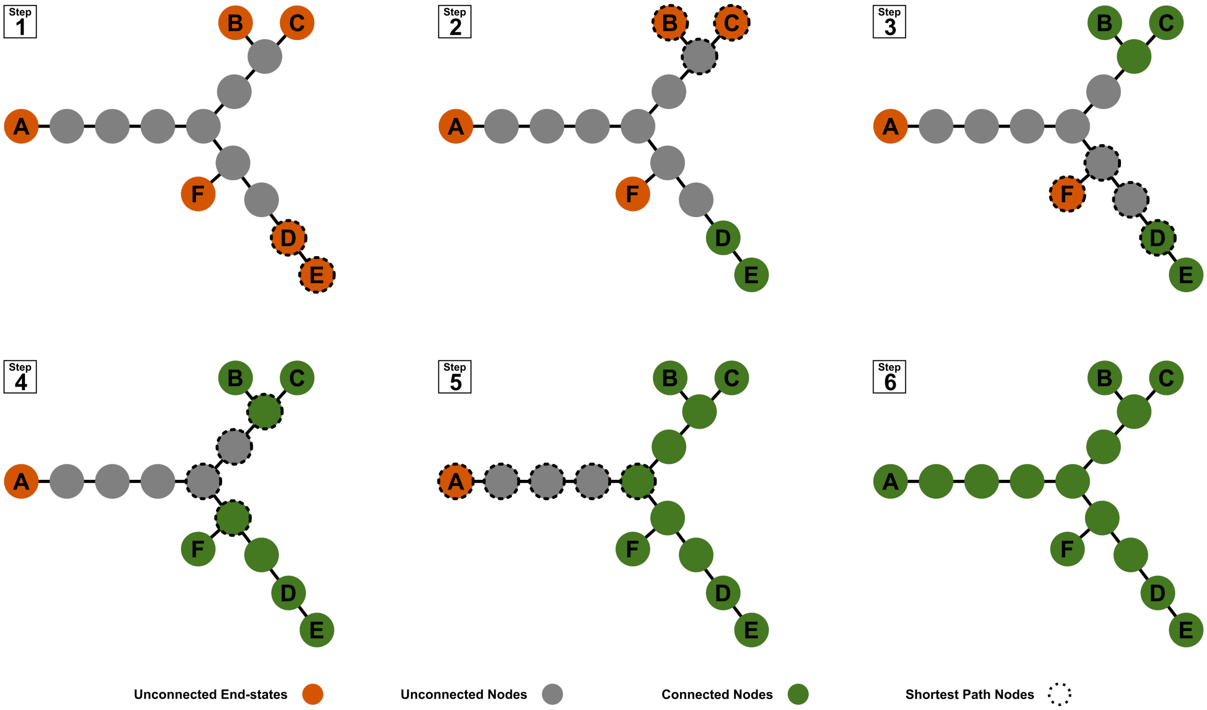


Figure 4.1: Overview of p-Creode's hierarchical placement strategy (Step 1) p-Creode starts by connecting the two closest end-states (orange nodes) measured by graph distance. Connected end-states and any connecting path nodes (dash outlined nodes) form a graph component (green nodes) that is added back to the pool as a single connectable entity. The graph is built hierarchically by connecting close entities: an end-state to another end-state (connected or unconnected) (Step 2-3), a connected path node from one component to another connected path node from another component (Step 4), or an end-state to a connected path node from a component (Step 5). The algorithm ends when all the end-states form a complete graph containing only end-states and path-connecting nodes (Step 6).

4.5. Consensus Alignment

As an additional step to alleviate short circuiting due to noise and sparse data, we add a consensus alignment step that reassigns the locations of path nodes in the constructed topology in a way that more accurately reflects the paths observed in the data. The consensus alignment starts by taking each of the path nodes from the constructed topology in step iv and iteratively assigns random data points from the original dataset (1000 at a time minus noisy data points) to one of the path nodes based on Euclidean distance, while updating the position of the path node by taking the median between assigned data points and the path node. This process functions similar to a relaxation, where the relative abundance of the data points serves as a potential energy function, pulling path nodes into basins. After reassignment, the new path nodes are re-connected using our hierarchical placement strategy, creating a new multi-branching topology. Following this step, lineages leading to nodes not identified as end-states are removed and a final topology is produced.

Because consensus alignment employs a sequence of cells for mapping the most consensus routes, we systematically tested whether randomizing this sequence will alter the final results generated by p-Creode. Starting from the same graph generated by hierarchical placement, we ran consensus alignment 100 times with a randomized sequence of cells each time. The results were then evaluated by clustering on p-Creode scoring to determine whether this type of randomization will result in different clusters of graph topologies. Having a single cluster of graphs with few outliers reflects that the sequence by which consensus alignment is performed has minimal effect on p-Creode results.

As a detailed illustration, we used the Setty *et al.* thymic dataset as an example. Performing the above procedure resulted in 96% of the topologies placed in the major cluster with 4% as outliers (Figure 4.2A). The outlier topologies consisted of graphs without the gamma delta T cell population (X), which for the most part, has been gated out of the dataset aside from a small precursor remnant (Figure 4.2B). The major branches consisting of progenitor, CD4+, and

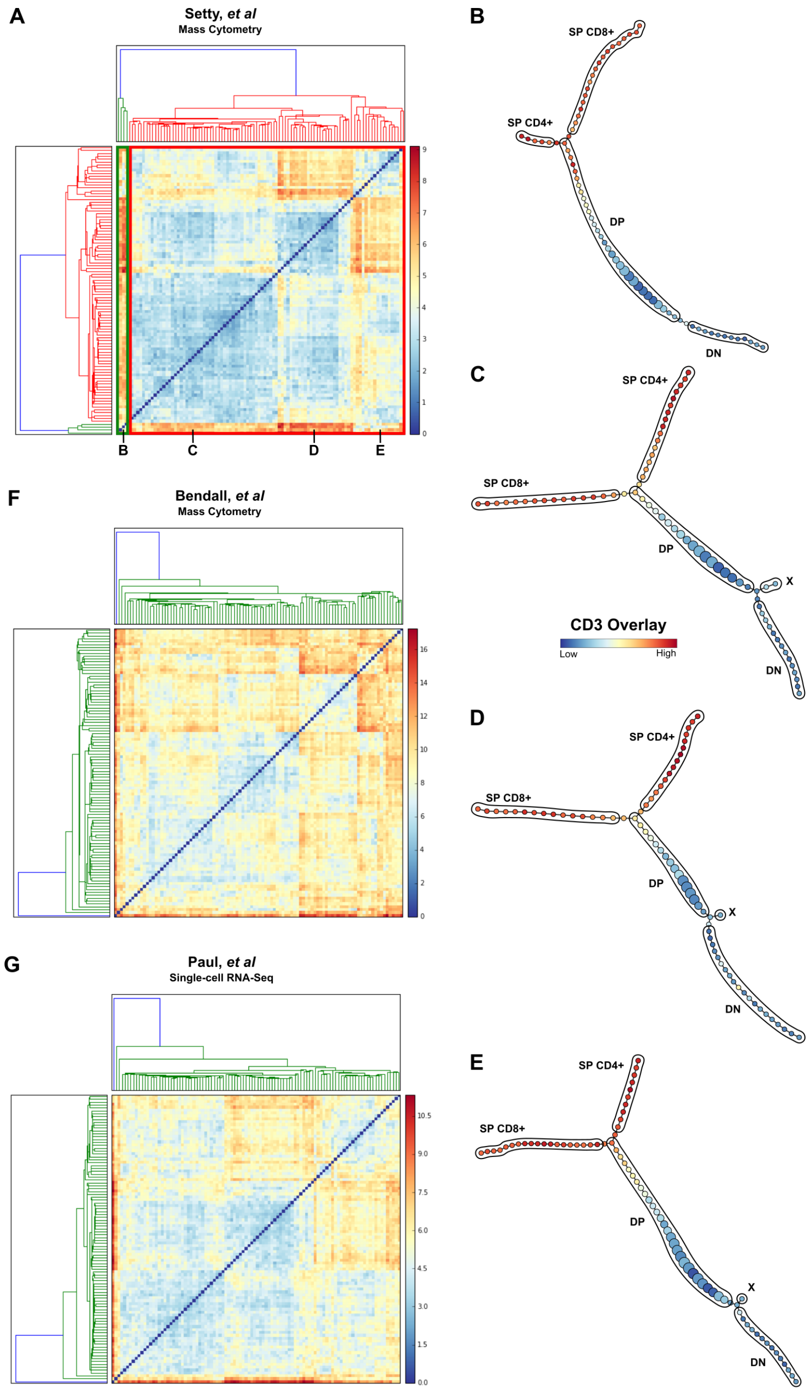


Figure 4.2: The effects of randomizing the sequence of cells used in consensus alignment. (A) 100 randomly initiated runs of consensus alignment starting from the same graph from hierarchical placement. Setty *et al.* data used. Final graphs from p-Creode were clustered based on p-Creode scoring (heat map) to evaluate whether randomized consensus alignment runs produce largely the same graph. (B) Final p-Creode topology of an outlier graph in A. (C, D, E) Final p-Creode topology of graphs sampled from different sub-clusters of the major graph population in A. (F,G) 100 randomly initiated runs of consensus alignment starting from the same graph from hierarchical placement performed as in A, but with different datasets. Bendall *et al.* and Paul *et al.* data used, respectively.

CD8+ T cells remained intact. Graphs sampled from different subclusters of the main cluster consisted of very similar topologies (Figure 4.2C-E). We also performed the same procedure on the Bendall *et al.* bone marrow mass cytometry dataset and Paul *et al.* myeloid scRNA-seq dataset, with similar results (Figure 4.2F-G). These results demonstrate that the order by which cells are used for consensus alignment has minimal effect on the final outcome of p-Creode.

4.6. Graph Scoring Overview

Due to random down-sampling, multiple independent runs will produce an ensemble of N final topologies. We leveraged an ensemble approach to account for the possibility of alternative routes reflected in the data. For example, in dysregulated processes such as cancer, we envision that increased plasticity leads to multiple alternative routes of cell transitions that reflect real biological differences. For well-controlled processes in healthy tissue though, we expect p-Creode to mostly produce a single cell transition trajectory from the data (with outliers due to technical noise). In this report, we are dealing with the latter and not the former, and representation of dysregulated transition processes will be addressed in future studies. To standardize our ensemble approach, we used $N = 100$ runs. The reasons behind this number are as of follows:

1. If we use a small number for N (e.g., N=10), alternate topologies may appear only once, which may not allow us to distinguish between technical noise versus biological variations.

2. For $N > 100$, we begin to see a diminishing rate of return. That is, we do not get added value in seeing different classes of graphs while the computation costs increase. We performed an additional study varying N (Paul *et al.* dataset), and we observed that changing N (from 10 to 1000) in this well-controlled process has minimal effect on the final p-Creode output (Figure 4.3).

It appears that p-Creode is robust to changing N values given that we are analyzing home-

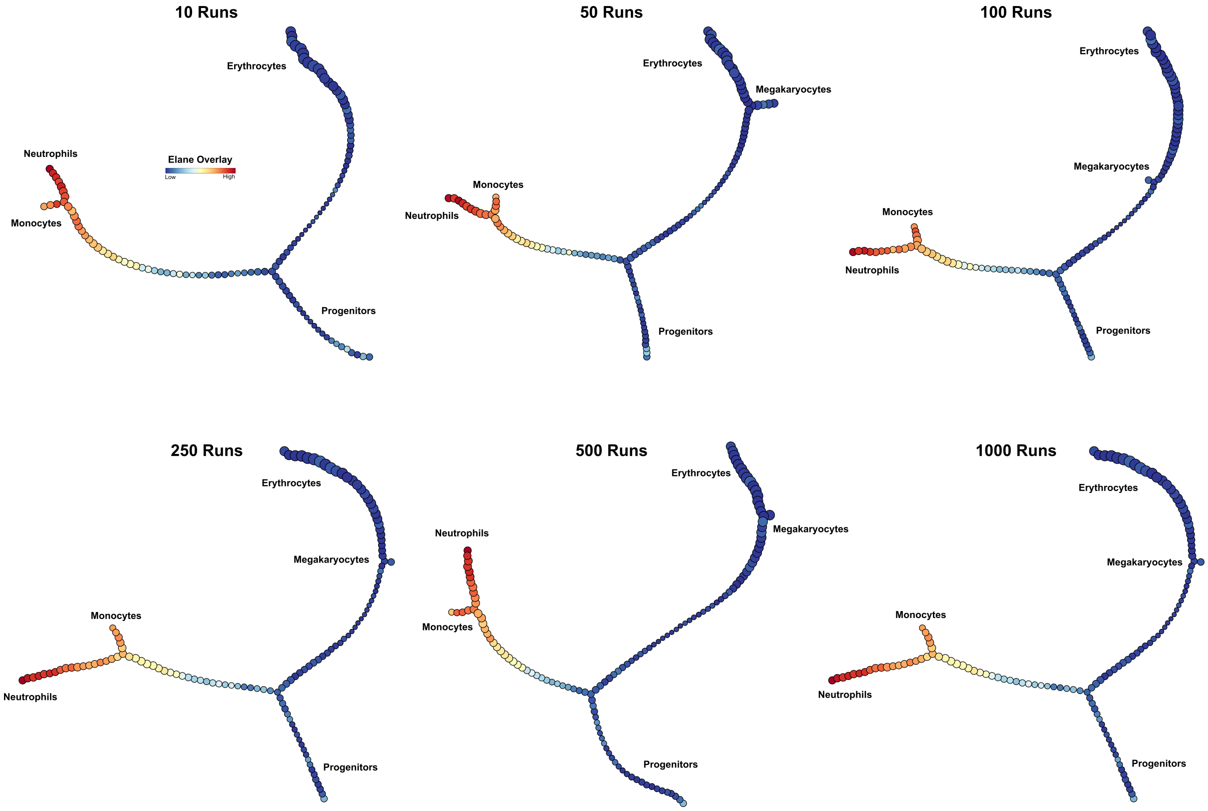


Figure 4.3: The effects of running *N* number of iterations of p-Creode on resampled data. Paul *et al.* dataset used for demonstrating the robustness of p-Creode to number of iterations used for selecting a representative trajectory. *N*=10, 50, 100, 250, 500, 1000.

ostatic processes. Having $N = 100$ allows for an ensemble approach that captures different alternative topologies by providing a high enough N to enable statistically significant clustering of these alternative graphs, if they exist. This ensemble aspect will be useful for modeling dysregulated processes in future investigations.

To come to a consensus on a final topology and to measure the robustness of the generated topologies, we developed a scoring metric (called p-Creode score) to compare the dissimilarity between topologies – a non-trivial problem given that each computed graph has different edges and nodes. Each topology can be viewed as a metric space where a distance matrix can be defined by the graph distance between all nodes. This definition allows us to leverage a version of the Gromov-Hausdorff distance, previously used to compare the distance between metric spaces (Edwards, 1975). While inspired by the Gromov-Hausdorff distance, our approach is tailored to the task of measuring the dissimilarity of graph ensembles generated from a common pool of data points. Specifically, our metric is uniquely built to comparing graphs generated by p-Creode and other similar lineage reconstruction algorithms, being that they are acyclic, undirected, and asymmetric in architecture.

Our approach is composed of two scoring components, a graph distance component and a topological component, to capture changes in the position of and connection between nodes, respectively. More formally, we wish to compare two graphs A and B defined as a pair of sets (*V,E*), where V refers to a set of nodes and E is a set of weighted edges. If graph A contains x number of nodes and graph B y number of nodes. Then the nodes sets are defined as V_A = $\{a_1, a_2,., a_x\}$ and V_B = $\{b_1, b_2,., b_y\}$ respectively. The p-Creode score ($Score_{AB}$) of comparing graph A to graph B is the sum of the graph distance component (GD_{AB}) and the topological component (TP_{AB}),

$$Score_{AB} = GD_{AB} + TP_{AB} \tag{4.5}$$

GD_{AB} is the per comparison average difference in pairwise Euclidean weighted graph distance (*gdist*) plus a transformation distance or

$$GD_{AB} = \binom{x}{2}^{-1} \sum_{\{a,\acute{a}\} \in V_a} \delta_{AB}(a,\acute{a}) \tag{4.6}$$

where

$$\delta_{AB}(a,\acute{a}) = abs(gdist_A(a,\acute{a}) - [trans_{AB}(a) + trans_{AB}(\acute{a}) + gdist_B(T_B(a),T_B(\acute{a})]) \tag{4.7}$$

trans$_{AB}$*(a)* is the Euclidean distance between node *a* in graph A and its closest neighboring node in graph B or *T*$_B$*(a)*,

$$trans_{AB}(a) = dist(a, T_B(a)) \tag{4.8}$$

The *TP*$_{AB}$ is the per comparison average difference in summed degree of path nodes with degree over 2 (*pd*), where the degree of a node in an undirected graph is equal to the number of edges connected to it, and any node with a degree over 2 signifies a branch point in the topology. More specifically,

$$TP_{AB} = \sum_{\{a,\acute{a}\} \in V_a} \beta(a,\acute{a}) \tag{4.9}$$

where

$$\beta(a,\acute{a}) = abs([pd_A(a,\acute{a}) - (2 \times |pd_A(a,\acute{a})|]) - [pd_B(T_B(a),T_B(\acute{a}) - (2 \times |pd_B(T_B(a),T_B(\acute{a}))|]) \tag{4.10}$$

and $|pd|$ is the number of nodes in the shortest path with degree greater than two. Since p-Creode scoring function is not inherently symmetric ($Score_{AB} \neq Score_{BA}$) the maximum of the comparisons is taken,

$$pScore(A,B) = max(Score_{AB}, Score_{BA}) \tag{4.11}$$

For a given set of computed graphs, a score matrix is created by comparing each graph to all other graphs. The most representative topology is the graph with the lowest overall mean p-Creode score. Figure 4.4 demonstrates a single comparison on a toy graph. Two graphs are constructed from nodes selected from a common pool of 8 data points in 2-dimensional analyte space (Figure 4.4A, B). These two graphs are comprised of different numbers and identities of nodes with different connections. For the single comparison, we are comparing the difference in the graph relationship between nodes 1 and 8. For the distance component of the metric, the graph distance between nodes 1 and 8 in graph A is calculated as before, but since neither 1 or 8 are in graph B a node transformation is performed by finding nodes in graph A that are closest in Euclidean distance to 1 and 8 in graph B (nodes 2 and 7, respectively) (Figure 4.4C, D). The transformation constitutes a penalty for subsequent calculations (Figure 4.4 - dashed lines). The distance between nodes 1 and 8 in graph B is calculated by summing the total transformation penalty and the graph distance between nodes 2 and 7 in graph B. For the topological component, the number of branches points (Figure 4.4C, D - red centered nodes) are counted along shortest path length between nodes being compared (Figure 4.4C, D - dashed outlined nodes). The difference between these counts is the topological component.

There are two applications of the p-Creode score to the overall algorithm. First, over an ensemble of *N* iterations, we can use the p-Creode score to gauge the similarity of each graph to each other in a run, with a low score describing a highly robust result (i.e., getting the same or similar graphs over multiple runs). This would provide a glimpse on how regulated a process is in terms of biological variation or how noisy the data are in terms of technical variation. Secondly, we use the p-Creode score to select the most representative graph for visualization purposes. If the data describes a homeostatic process that is well-regulated producing one major population of graphs via p-Creode, the graph with the lowest average score (most similar to all other graphs) is chosen to be visualized. We envision that the p-Creode score has ap-

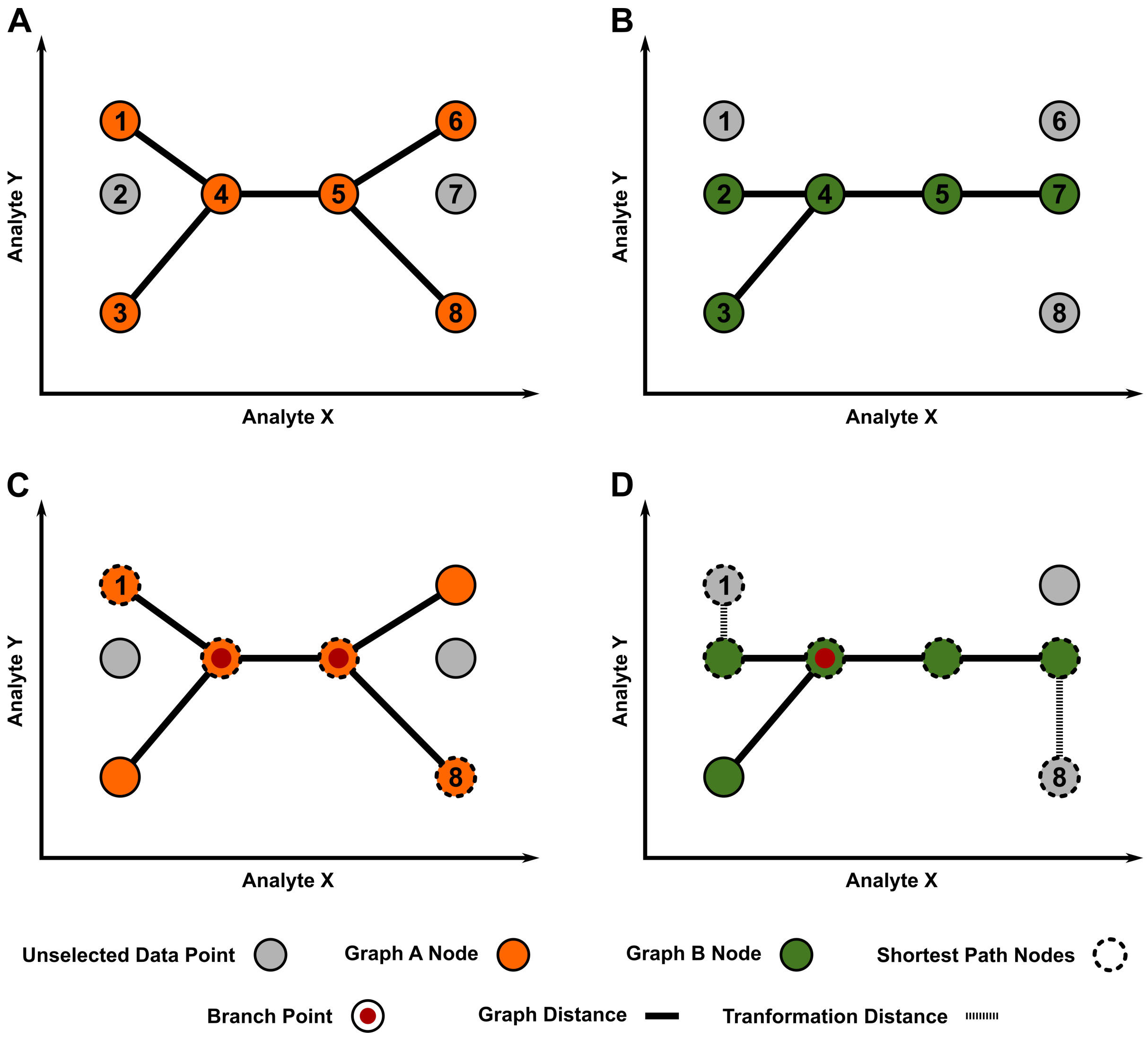


Figure 4.4: p-Creode scoring on a single pairwise comparison. Two graphs (A,B) constructed from nodes selected from a common pool of eight data points in two-dimensional analyte space. Graphs are comprised of different numbers and identities of nodes with different edge connections. (C,D) Comparison of the difference in the graph relationship between nodes 1 and 8. For the distance component of the metric, the graph distance between nodes 1 and 8 in graph A is calculated (C), but since neither 1 or 8 are in graph B a node transformation is performed by finding nodes in graph A that are closest in Euclidean distance to 1 and 8 in graph B (nodes 2 and 7, respectively). The transformation constitutes a penalty for subsequent calculations (D dashed lines). The distance between nodes 1 and 8 in graph B is calculated by summing the total transformation penalty and the graph distance between nodes 2 and 7 in graph B. For the topological component, the number of branches points (C,D - red centered nodes) are counted along the path between nodes being compared (C,D - dashed outlined nodes). The difference between these counts is the topological component.

plication beyond the p-Creode algorithm for comparing graph structures in other data science problems. However, because this scoring metric is new, we sought to verify its utility in graph comparisons independent of the p-Creode algorithm, as detailed below.

4.7. Independent Validation of the p-Creode Score

To empirically demonstrate the validity of the p-Creode score, we compared and contrasted its utility with other common metrics in a toy graph optimization problem representing a large random cohort of graph comparisons. We ran this stochastic optimization problem 100 times to sample the performance of this metric over a search space of over 10200 graphs (Cayley) to obtain a comprehensive variety of graph comparison situations. We compared the performance of our metric to two other common metrics, the Euclidean distance between matching (nearest) nodes (positional scoring) and the difference in average path lengths between graphs (path scoring). These two metrics capture aspects of the changing positions of and connections (graph structure) between nodes. To describe the optimization problem briefly, the goal is to match a starting graph to a target graph using a progressive search function guided by a metric (either the p-Creode score or the common metrics). Our search space is composed of all graphs (10200) that can possibly be constructed by 100 nodes (points) distributed in 2 dimensional space. The starting graph is chosen randomly and can comprise any number of nodes connected in any way into a single graph, on which an iteration of perturbation and scoring is performed. The most similar graph according to scoring using the p-Creode score, positional scoring, or path scoring is kept after each iteration until convergence is reached. Examples of this problem are shown in Figure 4.5 A and B, with a simple and complex initial graph, respectively. The details of this algorithm is described in the pseudocode below. The p-Creode score was the only scoring metric used that was able to consistently return the target graph with matching nodes and connections. As expected, position scoring by Euclidean distance returned graphs with mostly matching nodes but with edges randomly oriented. Path scoring by average path length returned graphs with rudimentary structure and node positions not matching the target

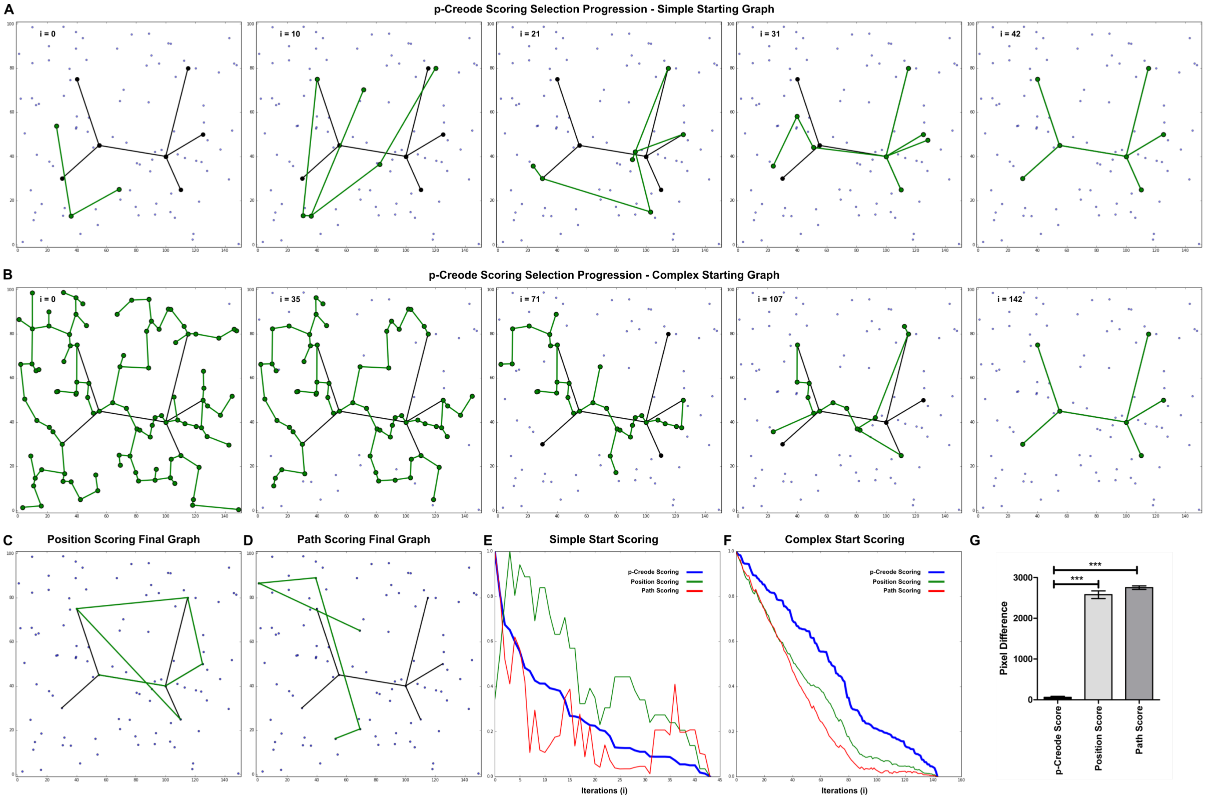


Figure 4.5: Comparison of position, path, and p-Creode scoring via a graph optimization problem. (A,B) Iteration of graph solutions towards final solution (the target graph) guided by p-Creode scoring during the optimization procedure with simple (A) and complex (B) initial graphs. Graphs may contain any of the 100 data points distributed in two-dimensional space. (C,D) Final solutions of graphs guided by position scoring (C) and path scoring (D). (E,F) Tracking of the other scoring metrics during optimization guided by p-Creode scoring for a simple (E) or complex (F) initial graph. (G) Pixel-by-pixel difference between the final solutions (derived from p-Creode, position, and path scoring) and the target graph calculated by processing images of graphs. Error bars represent SEM from $n = 100$ runs. $****P < 0.0001$.

graph. When guided by p-Creode scoring, tracking of the other scoring metrics during iterative optimization shows a similar downward trend, reflecting increasing similarity (over both positional and structural aspects of graph comparisons) as the target graph (Figure 4.5E, F). Because the data points exist in two-dimensional space, it is possible to visually compare the graphs. As an unbiased way to visually compare graphs, we used image processing to quantify graphical differences between the selected graphs and the target graph (Figure 4.5G). Over 100 simulations using the p-Creode score, position score, and path score were conducted and the selected graphs from each were compared to the target graph by image processing. Similar to our qualitative assessment, graphical comparisons revealed that only when guided by p-Creode scoring does the final selected graphs closely resemble the target graph.

4.8. p-Creode Modifications for Sparse Datasets

p-Creode was originally designed to run on large datasets consisting of thousands of data points in order to ensure that transition states required for connecting a graph are adequately represented. For instance, for samples where terminal states are overrepresented and transition states are underrepresented, end-state selection, hierarchical placement, and consensus alignment would fail since there will be an infinite number of ways to connect cell states. Down-sampling equalizes the distribution of data points across all cell states to ensure adequate representation. With that said, p-Creode can theoretically run on small datasets (hundreds of cells) where data points are originally well-distributed across cell states. This is the case with alveolar dataset. The following modifications were made to p-Creode for running sparse datasets:

1. No noise was removed, so all data points were viewed as informative.
2. We did not double the number of clusters identified as end-states due to the sparseness of the dataset. Performing this procedure in such a small dataset will artificially split cell populations.

Table 4.1: Pseudo-code for scoring optimization routine

```
1. $bestgraph =$ randomly initialized starting graph
2. $bestscore =$ starting graph score (scored by position, path, or p-Creode)
3. $unsuccessfulattempts = 0$
4. for(100,000 attempts)
    • randomly perturb best graph by selecting from the following
        (a) add a new node (or datapoint) to the graph
        (b) delete a node from the graph
        (c) swap a node ID (or position in dimensional space) with ID of node already in
            the graph
        (d) swap a node ID with ID of node not currently in graph
    • $runscore = mutatedgraphscore$
    • $unsuccessfulattempts += 1$
    • if($runscore < bestscore$):
        – $bestscore = runscore$
        – $bestgraph =$perturbed graph
        – $unsuccessfulattempts = 0$
        – if($bestscore == 0$):
            (a) break
    • elseif($unsuccessfulattempts >= 2,000$):
        – $bestscore = runscore$
        – $bestgraph =$perturbed graph
        – $unsuccessfulattempts = 0$
```

3. The closeness threshold was raised from 0 to 1, given the sparseness of transitional cell states over inflates the closeness values of denser end-states.

CHAPTER 5

Comparison of Trajectories Across Datasets

5.1. Introduction

Techniques designed for single-cell analysis have largely focused on identification and characterization of cell populations within a dataset (Bacher and Kendziorski, 2016; Mair et al., 2016; Haque et al., 2017; Diggins et al., 2018; Yalcin et al., 2016; Amir et al., 2013; Levine et al., 2015). These applications have not been designed to quantify the similarity of cell types but to provide a means of evaluating significant differences. The difference being, significance testing is an absolute declaration of difference, whereas quantification refers to a measure of similarity on a sliding scale that is also capable of proving significance. More recently, algorithms have came online capable of quantifying similarity of cellular phenotypes within a dataset as well as across datasets (Crow et al., 2018; Orlova et al., 2016). One such algorithm is MetaNeighbor (Crow et al., 2018) which aims to quantify how well transcriptomic cell subtypes replicate across studies using a cross-validation scoring scheme to assess cell-type identity. Another approach (Orlova et al., 2016) uses the Wasserstein metric or, as it is better known, earth mover's distance (EMD). EMD is used in statistics to measure the distance between two probability distributions over a certain space. Orlova *et al* applied EMD across datasets to quantify similarity of two cell populations by measuring the distance between expression distributions. These two algorithms have shown to preform well in quantifying phenotypic similarities (Crow et al., 2018; Orlova et al., 2016), but it is unclear how these algorithms perform on more heterogeneous cell populations, like those found in continuous data cloud of transitioning cell states (Figure 2.2). While we have shown in previous sections that p-Creode scoring is capable of characterizing (Figure 4.2) and quantifying similarities (Figure 3.2E) among an ensemble of trajectories produced from the same dataset, it is unclear if p-Creode scoring can characterize differences across datasets between biological phenomena that presents as a continuous or heterogeneous spectrum of cell states. Since p-Creode scoring is able to compare graphs containing different edges and nodes, we hypothesize that p-Creode scoring is also capable of comparing the similarity of trajectories derived from continuous data clouds across datasets. In the following section we apply p-Creode scoring as a proof of concept to technical and biological replicates produced by scRNA-seq.

5.2. Trajectory Comparison Across Replicates

To test if p-Creode scoring is sensitive enough for across tissue comparisons, we first generated scRNA-seq data for a series of replicates from two tissues in flux, the adult mouse colon and mouse pancreas at embryonic day 14.5 (E14.5), with the inDrop platform (Klein et al., 2015) (see section 3.7 and 6.4 for additional processing details). While adult colonic tissue, as a part of homeostasis, is in a state of constant renewal given the high turnover rate of the intestinal epithelium (Creamer et al., 1961), adult pancreatic tissue has a much lower turnover rate (Burke et al., 2007) and, therefore, would not be a good candidate for trajectory analysis. Given the low turnover rate of adult pancreas, we chose tissue still under development at E14.5 which is characterized by active growth, branching, and cellular differentiation (Murtaugh and Melton, 2003). From the colon, 5 datasets were produced, comprising of 3 biological (bio1-3) and 3 technical (tech1-3) replicates; the 3 technical replicates were produced from the third biological replicate. From the pancreas, 2 biological replicates (pan1-2) were generated. Cell counts for the replicates ranged from 440 to close to 1600, see Table 5.1 for cell count breakdown. Next, we performed t-SNE analysis on the pancreatic replicates combined separately with the biological (Figure 5.1 - Left) and technical (Figure 5.1 - Right) replicates from the colon. In both t-SNE plots separation was observed between tissues, as expected, while all tissue specific replicates were largely interspersed, signifying minimal batch effects.

We then ran p-Creode on each replicate $N = 100$ producing a total of 700 trajectories. Graphs were generated from datasets independently processed using the neighborhood variance gene selection routine (Welch et al., 2016)(see Section 2.2 and 6.5 for more details).

Table 5.1: Cell counts for biological and technical replicates

Replicate	Cell Count
tech1	1137
tech2	1517
tesh3	440
bio2	1593
bio3	1288
pan1	329
pan2	614

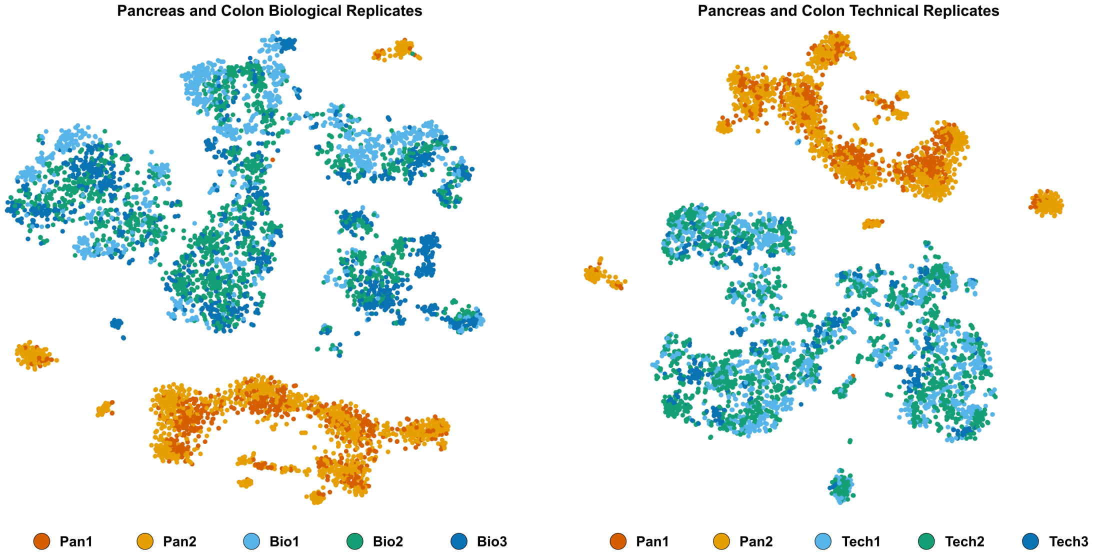


Figure 5.1: t-SNE plots of replicates from mouse pancreas and colon. t-SNE analysis of scRNA-seq data of pancreas with biological replicates from colon (Left) and pancreas with colonic technical replicates (Right) shows the absence of segregation of data points among all replicates and segregation between tissues.

Because gene selection was processed separately, the selected gene sets differ replicate-to-replicate with varying degrees of overlap. Due to differing gene sets, each dataset was first converted into a common data space, where all cells are described by a shared gene set. Initially, we sought to apply the same feature selection routine (Welch et al., 2016) applied independently to a combined dataset consisting of cells from all replicates, but worried that the larger datasets would exert more influence on the selection process and skew the results, possibly affecting downstream scoring. Another option considered, sampling the larger datasets to the lowest common size denominator, was rejected because of information loss incurred by removing data points. As a result, we decided to convert the replicates into a common gene space by the following process:

1. An initial gene list was composed, consisting of all unique genes selected as a result of independent gene selection (Welch et al., 2016).

2. The initial gene list was then reduced by removing all genes not present across all replicates; genes with low counts across all cells in a dataset are removed prior to gene selection.

3. PCA was preformed on the combined dataset with the reduced gene list from step 2. The top 3 components were selected and used as input for p-Creode scoring.

Once converted, p-Creode scoring was applied and results were grouped by hierarchical clustering. Clustering produced two distinct clusters, one cluster representing pancreatic tissue (Figure 5.2 - Green) and the other colonic tissue (Figure 5.2 - Red). These results suggest p-Creode scoring is sensitive to differences in trajectories across tissues.

Next, we wanted to determine whether p-Creode scoring is capable of detecting differences between technical and biological replicates from the same tissue. We expect trajectories from technical replicates, generated from the same mouse, should be more similar than biological replicates, generated from different mice (Figure 5.3 - Left). Using the same 5 colonic replicates described above, t-SNE analysis was preformed to check for any batch bias among

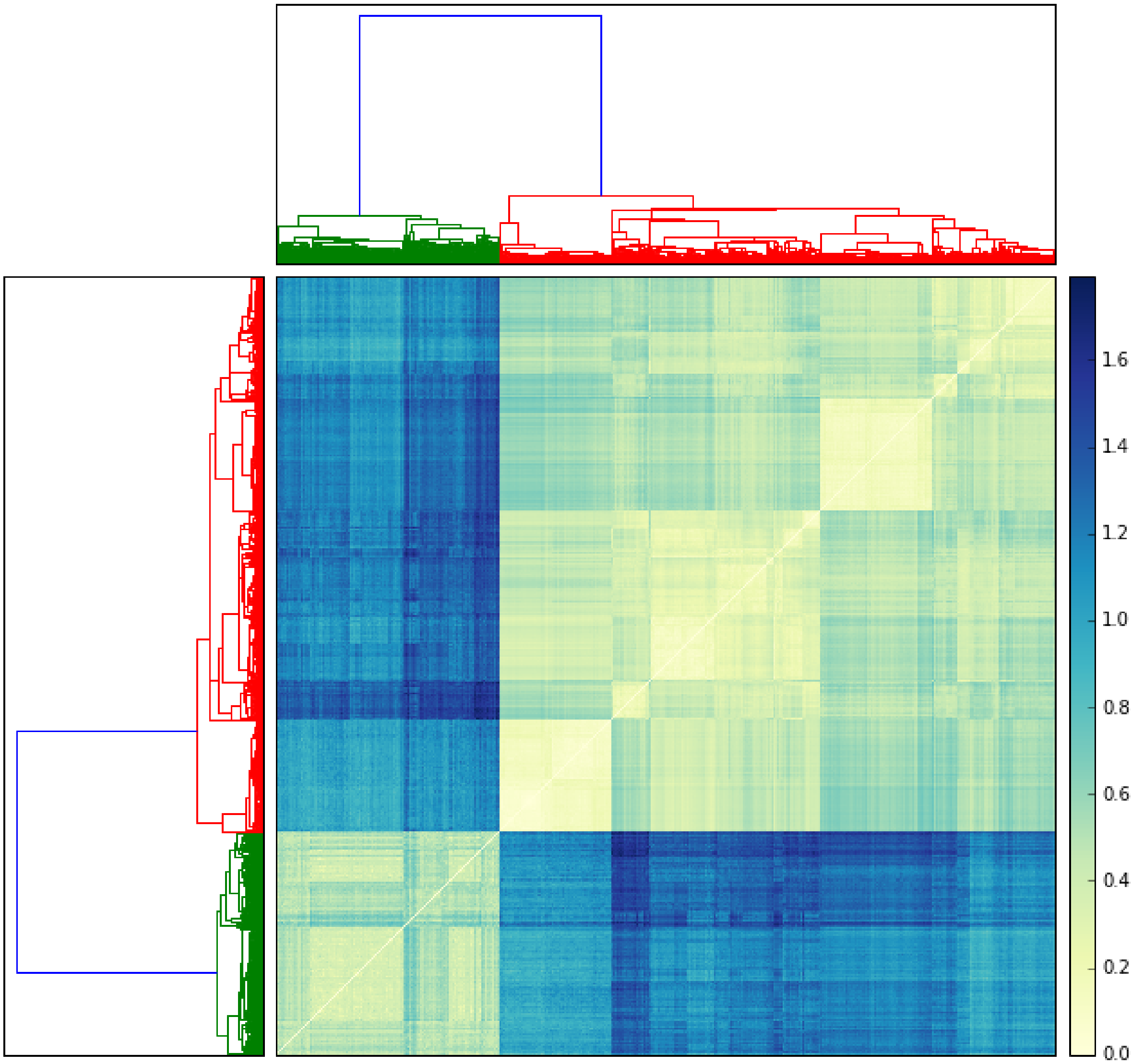


Figure 5.2: Clustering of p-Creode scores from across tissue trajectories. Hierarchical clustering (heat map) of trajectories derived from colonic (Red) and pancreatic (Green) tissues.

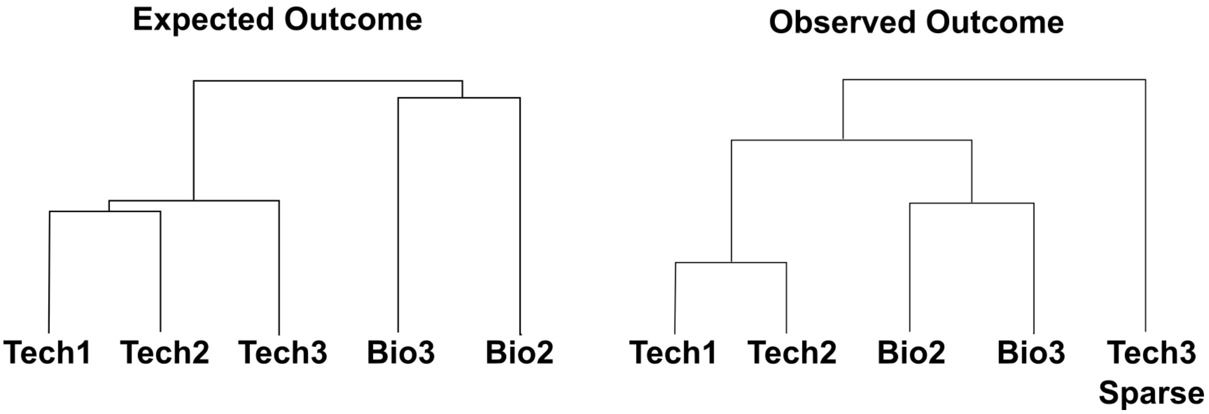


Figure 5.3: Expected versus observed outcomes across replicates. (Left) Expected clustering of colonic replicates, technical replicates from same mouse are more similar than biological replicates from different mice. (Right) Observed clustering results where the most sparse dataset (tech3) was an outlier to both technical and biological replicates.

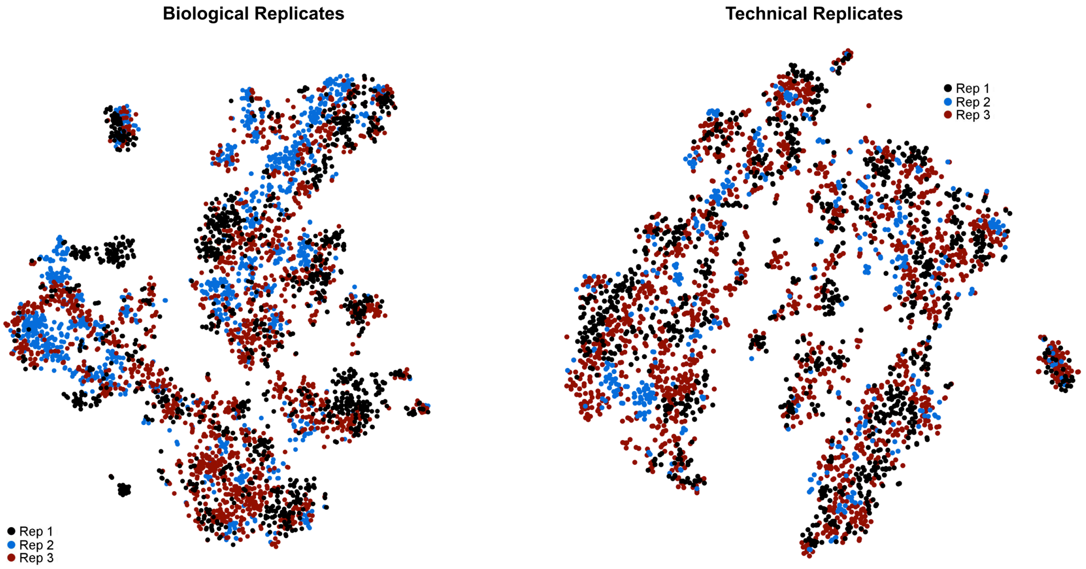


Figure 5.4: t-SNE plots of colonic biological and technical replicates. t-SNE analysis of scRNA-seq data biological replicates (Left) and technical replicates (Right) shows the absence of segregation of data points among all replicates.

biological and technical replicates. t-SNE plot displayed interspersion within each group of replicates, suggesting minimal batch effects (Figure 5.4). The procedure listed above was employed to convert the 5 datasets into a shared gene space. As a trial run, p-Creode scoring was then applied to 10 graphs selected at random from the 100 generated for each replicate. Hierarchical clustering of trajectory scores suggested, in contrast to the expected outcome, tech3 was an outlier to all other replicates (Figure 5.5 and Figure 5.3 - Right). Examination of representative graphs (Figure 5.6) reveals a common trajectory hierarchy among all replicates with the same characteristic cell transition pattern observed in Figure 3.21E-F. A stem/progenitor branch giving rise to an absorptive colonocyte branch and a secretory goblet cell branch (see section 3.7 for more details). The most noticeable difference between the tech3 trajectory and the other trajectories is the sparsity of tech3's graph nodes. Sparseness of graph nodes in a p-Creode trajectory or lineage is governed by the number of intermediate cell states in between end-states being mapped. Therefore, the sparsity of tech3's trajectory stems from its relatively low cell count, 440, compared to all others at > 1100 (Table 5.1). To test the effects of sparsity on trajectory comparison, p-Creode was applied to a simulated sparse dataset generated by selecting at random 440 cells from tech1 (sparse tech1). If sparsity of the dataset is impacting trajectory scoring, sparse tech1 should also be an outlier, despite being composed of data points from tech1. Ten trajectories generated from sparse tech1 were then compared to ten random trajectories from all other replicates. Clustering of graphs (not shown) showed three clusters with tech3 and sparse tech1 occupying one cluster each and all other replicates occupying the last. An investigation into the underlying cause of bias against sparse graphs revealed an unknown additional penalty being accrued during the scoring algorithm's node transformation routine. When a node is not contained in both graphs being compared, a node transformation is performed by finding a node in the first graph that is closest in Euclidean distance to the second. In this case, when graphs being compared were not derived from the same dataset, every node comparison requires such a transformation (Figure 4.4). The additional penalty stems from the node-to-node transformation requirement, where two very similar trajectories can be scored higher if their nodes are not sufficiently close to one another (Figure 5.7A). This issue is amplified with sparse graphs, where the lack of nodes increases the likelihood of long transformation distances (Figure 5.7C). To correct this bias, the scoring routine was updated to allow for transformations into the edges as well as nodes in the opposing graph (Figure 5.7B). After p-Creode scoring algorithm was updated, we applied it again to the 5 colonic replicates $N = 100$. Clustering results mirror the expected outcome (Figure 5.3 - Left) with technical replicates clustered together (Figure 5.8 - Green) and biological replicates each occupying a single cluster (5.8 - Red and Blue). These results suggest that, with the scoring improvement, p-Creode is sensitive across replicates from the same tissue.

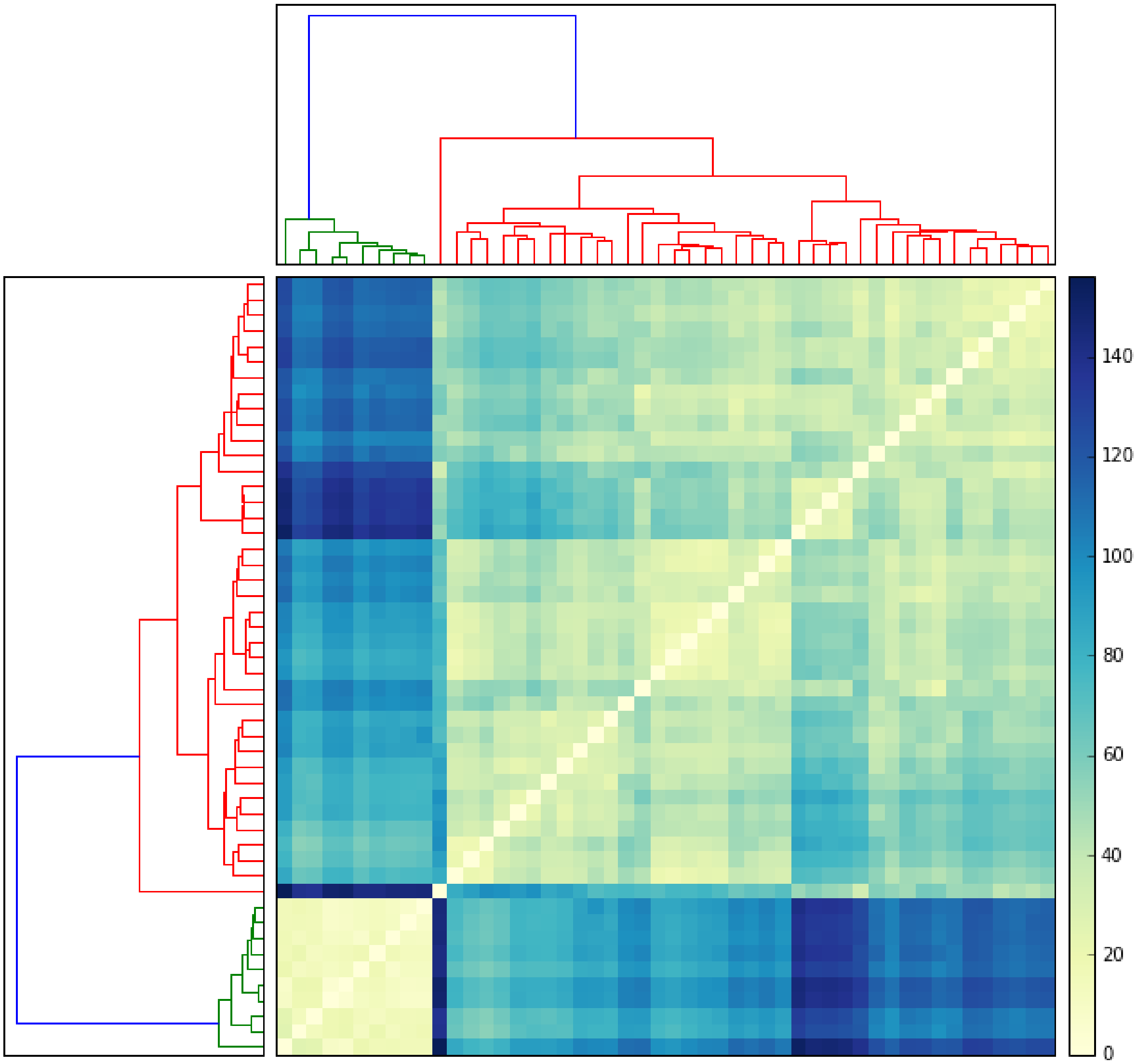


Figure 5.5: Clustering of p-Creode scores from across replicate trajectories. Hierarchical clustering (heat map) of trajectories suggests sparse tech3 dataset (Green) is an outlier to all other replicates (Red)

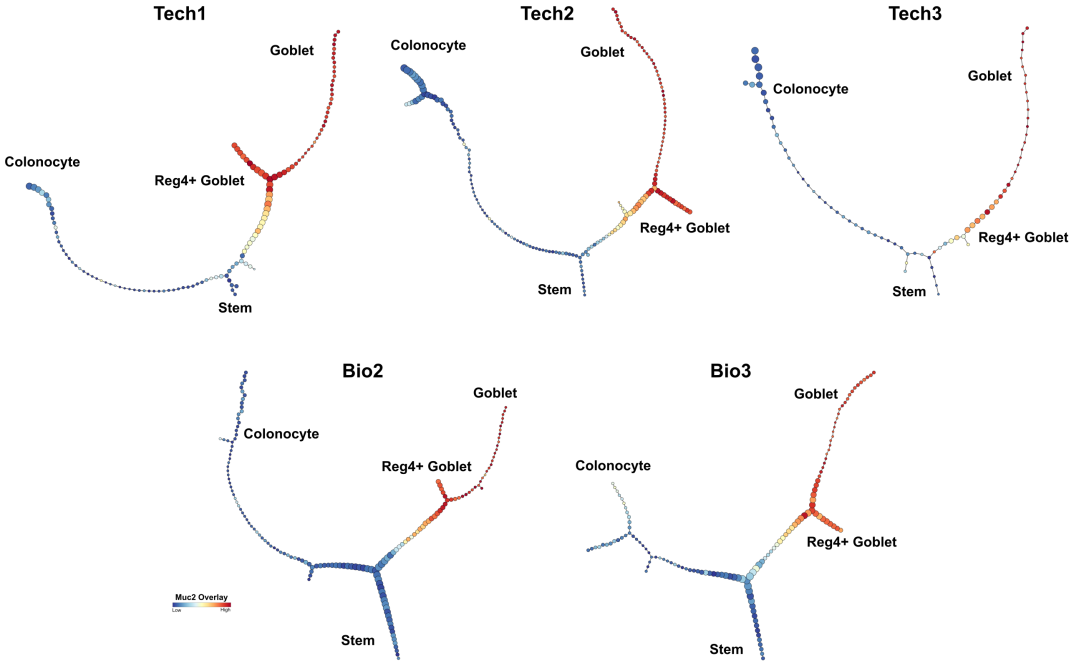


Figure 5.6: Representative p-Creode trajectories from colonic replicates. p-Creode analysis of scRNA-seq data generated by inDrop from colonic tissue replicates, most representative graph over $N = 100$ runs. Overlay represents Muc2 transcript levels. Tech3 trajectory resembles other trajectory replicates but noticeably contains fewer graph nodes.

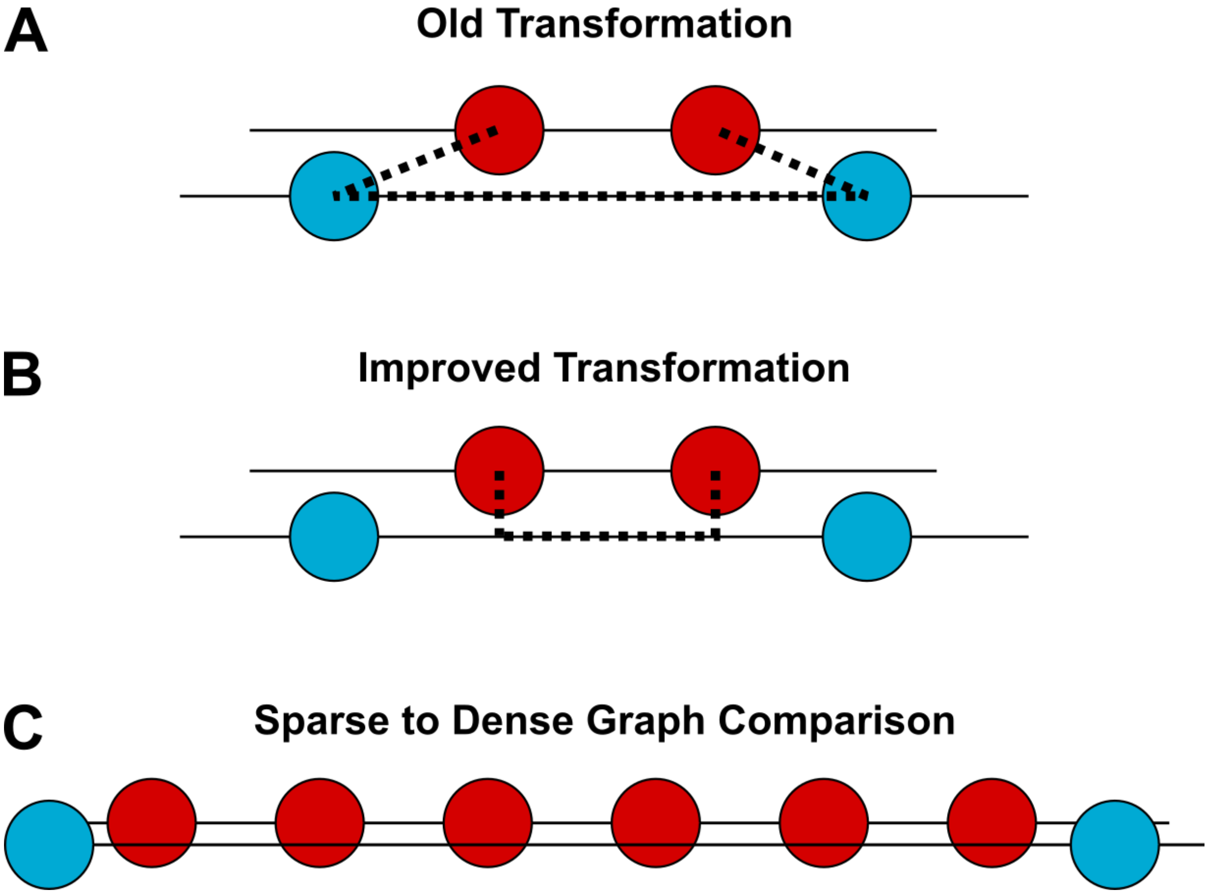


Figure 5.7: Old and new scoring transformation examples for across dataset comparisons. (A) Old scoring transformation where the node to node requirement unnecessarily penalized similar trajectories. (B) Improved transformation routine, where nodes can be transformed into edges as well as nodes in opposing graph. (C) Node to node requirement is amplified when a dense graph is being compared to a sparse graph, due to the likelihood of increased transformation distances.

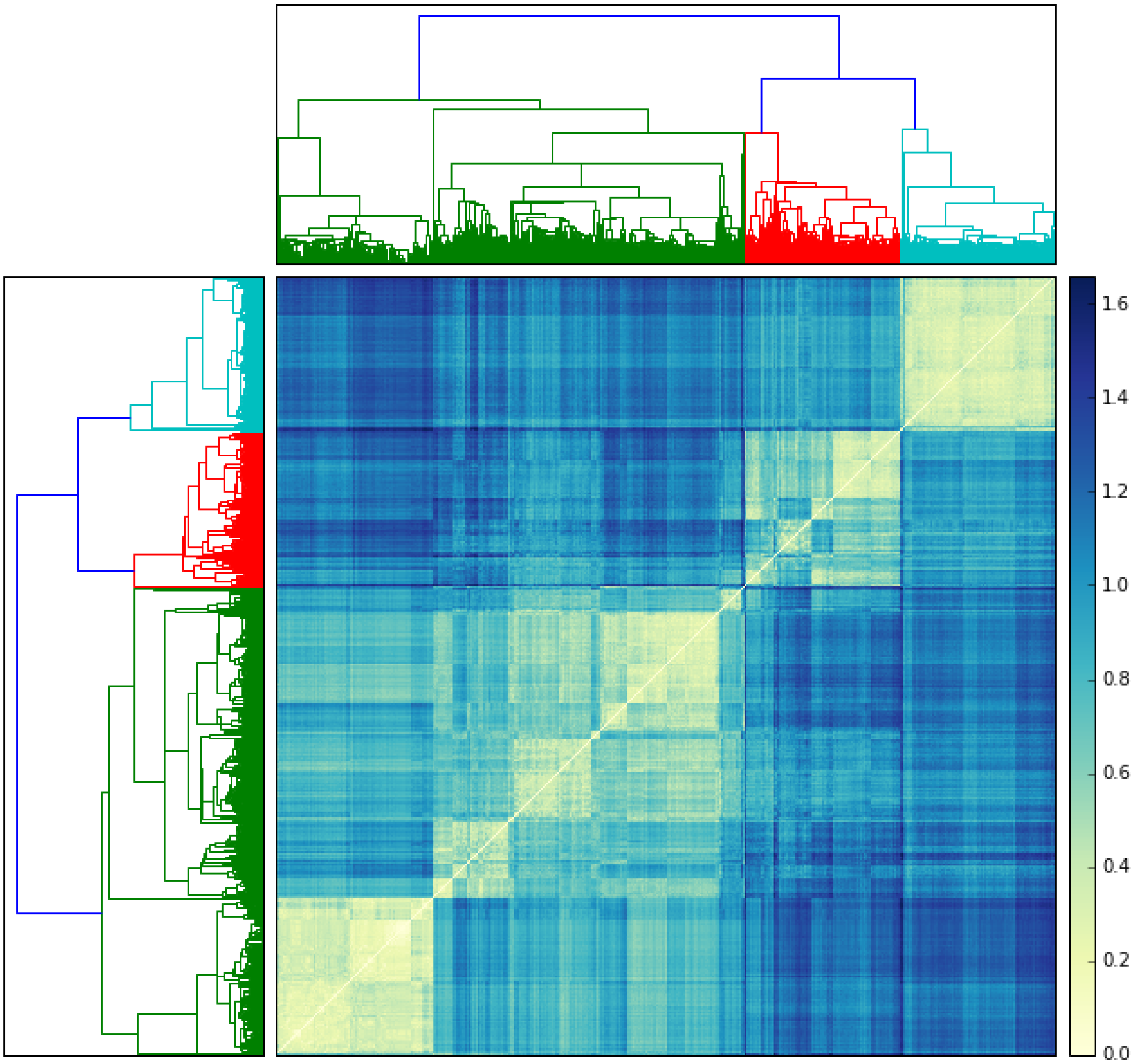


Figure 5.8: Clustering with improved p-Creode scoring from across colonic replicates. Clustering of p-Creode scores from across replicate trajectories. Hierarchical clustering (heat map) of trajectories, technical replicates (Green) and biological replicates (Red and Blue).

CHAPTER 6

Experimental Details and Methods

6.1. Mouse Experiments

Animal experiments were performed under protocols approved by the Vanderbilt University Animal Care and Use Committee and in accordance with NIH guidelines. Mice were stimulated with TNF- as a time course, and their duodena (proximal small intestine) were collected for analysis as previously described (Lau et al., 2012). For DISSECT, a previously published protocol was used (Simmons et al., 2015). For FFPE embedding for MxIF imaging, tissues were fixed in 4% formaldehyde for 24 hours and then were subjected to standard embedding procedures. For Cre-induced recombination experiments, 2 mg of tamoxifen (Sigma) was administered intraperitoneally at 2 months of age for 4 consecutive days, and animals were sacrificed and their tissues harvested 14 days after the first injection. Both Cre and wildtype mice were administered tamoxifen to control for its effects. Lrig1CreERT2 and Atoh1fl strains were purchased from the Jackson Laboratory in a C57BL/6 background. VillinCreERT2 and DBZ experiments were performed as previously described (Kim et al., 2014).

6.2. Mass Cytometry Analysis

Mass cytometry was performed on a Fluidigm-DVS CyTOF1 instrument with elemental calibration bead spike-ins (Finck et al., 2013). Cells were gated using intercalator (Iridium) following established procedures to identify intact single cells and eliminate cell doublets and clusters from analysis (Simmons et al., 2015). Single cells were then analyzed for intensity of multiple antibody conjugates.

6.3. MxIF Analysis

FFPE tissues were sectioned at 4 m and processed using standard immunohistological and antigen-retrieval techniques. MxIF was performed by using a sequential staining and fluorescence-inactivation protocol as previously described (Gerdes et al., 2013). Imaging was performed on an Olympus X81 inverted microscope with a motorized stage and acquired at 20x magnification. Antibody staining was performed overnight at 4C. At each round, images were computationally registered, and corrected for illumination and autofluorescence. Processed images were then segmented using a multi-marker supermembrane mask, and individual cells were quantified, as described (Gerdes et al., 2013). Partial and poorly segmented cells were removed. The mean, standard deviation, median, and maximum staining intensity for each protein was quantified with respect to the whole cell, cell membrane, cytoplasm, and nucleus, as well as cell location, area, and shape. Image processing was performed on the Amazon Cloud through the KNIME parallel architecture.

6.4. Single-cell RNA-sequencing

Colonic epithelium was enriched by incubating and shaking colonic tissues in a 2mM EDTA/EGTA chelation buffer, as previously described (Sato et al., 2011). The epithelium was then dissociated into single cells with a collagenase/DNAse enzyme cocktail (2mg/ml Collagenase I, 2.5mg/ml DNAse1) in a modified protocol that maintains high cell viability (Leelatian et al., 2017a). Cell viability was determined by counting Trypan Blue positive cells. The cell suspension was further enriched with a MACS dead cell removal kit (Miltenyi) prior to encapsulation, and the density of cells were calculated by counting. Before encapsulation 10% human K562 cells were spiked into the suspension to evaluate the doublet rate. Single cells were encapsulated and barcoded using the inDrop platform (1CellBio) with an in vitro transcription library preparation protocol (Klein et al., 2015). The number of cells encapsulated was calculated by the density of cells arriving at the device multiplied by the duration of encapsulation. After library preparation, the samples were sequenced using Nextseq 500 (Illumina) using a 150bp paired-end sequencing kit in a customized sequencing run (50 cycles read 2, 6 for the index read, rest for read 1). The replicates were multiplexed in a single sequencing run. After sequencing, reads were filtered, sorted by their barcode of origin and aligned to the reference transcriptome using inDrops pipeline (https://github.com/indrops/indrops). Mapped reads were quantified into UMI-filtered counts per gene, and barcodes that correspond to cells were retrieved based on previously established methods (Klein et al., 2015).

6.5. Data Analysis

t-SNE analysis was performed using the Barnes-Huts algorithm in Python (van der Maaten et al., 2011). Unpaired t-tests were performed using Prism (Graphpad). p-Creode was written in Python. Hierarchical clustering was performed in MATLAB (Mathworks). Villus and crypt cell quantification was performed with custom scripts in ImageJ with manual annotation using multiple fields of view per animal. For fully automated analysis, segmentation using nuclear markers (DAPI, Sox9, IL33, etc.) and membrane markers (PCK26, -Catenin, etc.) was used to generate epithelial nuclear masks and count total epithelial cells. p-Creode was written in Python and analysis was performed on an Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz with 32GB of RAM with Ubuntu 16.04.3 LTS. Down-sampled datasets with up to 14,000 data points can be analyzed with p-Creode with each independent iteration taking between 3-4 minutes each.

6.6. scRNA-seq Data Analysis

For the alveolar data obtained using the Fluidigm C1 system, previous criteria of selecting cells with $>=$ 1000 genes detected at $>=$ 1 FPKM with control ERCC $>$ 0 were used (Welch et al., 2016). For the myeloid data obtained using MARS-seq, previous criteria of filtering out cells with $<$ 250 molecules was used (Setty et al., 2016). For both MARS-seq and inDrop data, which consisted of raw transcript count, mitochondrial genes and genes where the maximal counts are one (noise) were filtered out, resulting in approximately 15,000 genes remaining. Transcript counts for each gene were normalized to the total transcript count per cell multiplied by the median total transcript count across all cells, as previously described (Setty et al., 2016). Data generated for all 3 approaches C1, MARS-seq, and inDrop were then ArcSinh normalized to stabilize the variance with a cofactor of 5, noting that the outcomes were not sensitive to the cofactor being used. From these normalized data table, the same select gene procedure was applied as previously described (https://github.com/jw156605/SLICER) (Welch et al., 2016). Briefly, the procedure selects monotonically increasing genes using a neighborhood variance approach. These data were then analyzed by p-Creode.

CHAPTER 7

Discussion and Future Relevance of p-Creode

7.1. Discussion and Conclusion

Presented in this dissertation is a new single-cell data analysis platform, called p-Creode, for the unsupervised mapping of multi-branching topologies from high-dimensional, single-cell data. Importantly, a metric for scoring graph structures comprised of both changing nodes and edges was derived to statistically evaluate the reproducibility of computed results and to compare trajectories across datasets. To my knowledge, this metric is the first of its kind in the field of graph theory, and can be applied to a variety of graphs such as signal transduction networks or phylogenetic trees.

p-Creode's performance was assessed against current MST-based, non-linear embedding-based, and next generation algorithms and we applied p-Creode to a variety of datasets from mass cytometry, MxIF, and scRNA-seq, both publicly available and generated by our group. Specifically important was the ability of p-Creode to generate multi-branching trajectories in each of these cases to recapitulate the complexity of cell-state transitions, which is a significant step forward in the field of single-cell biology. It is also important to note previous algorithms have primarily been designed and validated on a single platform. For example, Monocle2 and many other have been designed for scRNA seq while SPADE was designed for mass cytometry. An exception to this rule is Wishbone, which has been validated on both mass cytometry and scRNA seq. This is of importance due to the differences in data distributions produced by each technological platform.

Our analysis uncovered alternative routes of tuft cell ontogeny between the small intestine and the colon. Tuft cells were originally found to be specified in the secretory lineage (Gerbe et al., 2011), but their origins have since been contested (Bjerknes et al., 2012; Westphalen et al., 2014). Both our computational and experimental analyses indicate an Atoh1-

independent, and possibly, non-secretory cell origin of tuft cells in the small intestine, and an alternative origin of tuft cells in the colon. These observations support recent speculations by Gerbe and Jay regarding the potential functional differences among tuft cells at different anatomical sites (Gerbe and Jay, 2016), as well as our previous observations of different tuft cell distributions between the small intestine and the colon (McKinley et al., 2017). The discrepancies in phenotypes among studies and organ systems may arise due to the secondary effects of the microbiome. It has been shown that tuft cells can be regulated by luminal parasites, such as (Gerbe et al., 2016; Howitt et al., 2016; von Moltke et al., 2015), and commensal bacteria (McKinley et al., 2017). As such, knockout of Atoh1 ablates microbiome-regulating goblet and Paneth cells, which can subsequently affect tuft cells as a secondary effect. It should be noted that the small intestine and colon are characterized by large differences in microbial content and load, and we observe differential dependence of Atoh1 on tuft cell development between the two regions. A recent study also suggested that tuft cells may share a common progenitor with subsets of enteroendocrine cells in the small intestine (Yan et al., 2017). Because of the importance of the microbiome in various ailments, modulating luminal-sensing tuft cells may be important in controlling allergic and inflammatory diseases.

One of the features of p-Creode is its use of an ensemble approach that allows for delineation of alternative routes of specification. Statistical ensemble representation of p-Creode can allow greater depth of analysis, including the assessment of the level of regulation of a transitional process and representation of loops. Another key feature of p-Creode is the ability to manage large numbers of cells, which facilitates the tracking of less continuous, switch-like processes. A number of newer algorithms, such as SLICE (Guo et al., 2017), are largely designed for sparse scRNA-seq datasets, and they produce rudimentary trajectories (with single branch points) that may not scale to more complex multi-branching processes. However, scRNA-seq technologies that can query thousands of cells are emerging (Klein et al., 2015; Macosko et al., 2015), and the datasets generated, such as those here, will require scalable algorithms such as p-Creode.

A specific point of consideration is the selection of markers or principal components to include in the analysis; these should be related to the cell transition process of interest. Unrefined selection of markers will result in ectopic identification of terminal states that depict unrelated cellular behaviors, for instance, cell cycle states in a differentiation hierarchy. The selection of markers is especially crucial for candidate-based approaches such as MxIF and CyTOF, since larger emphasis is placed on each marker due to the relatively small number of markers evaluated. This problem may be somewhat alleviated in large-scale scRNA-seq studies, where cell state transitions are driven by massive epigenetic changes reflected in gene expression programs, and genes with coordinated changes can be selected without bias, as we have done in this study.

Another issue with single-cell analysis algorithms is the discrimination of rare cell populations from technical noise. Even in well-controlled single-cell experiments, misidentified data points (doublets/mis-segmented cells) exist and will appear as sparse data points with unique profiles distributed similarly to rare cells. p-Creode limits the effects of technical noise by 1) removing outliers at down-sampling, 2) consensus alignment of pathway nodes, and 3) selecting representative topologies using dissimilarity scoring. Improvements in rare cell detection can be achieved by having less noisy data and by developing better down-sampling algorithms that can distinguish technical noise from rare cells. These approaches may leverage current strategies for rare cell detection, such as raceID (Grn et al., 2015) and GiniClust (Jiang et al., 2016).

p-Creode analysis on single-cell, tissue-level data generates hypotheses regarding cellular transitions. Specifically, by being capable of comparing trajectories across datasets, p-Creode can be used to provide insights as to how the structures of transitional topologies change upon external perturbations such as in disease or wound repair. Our scoring metric provides a rigorous way to quantify the probabilistic nature of cell transitions where we expect a diverse ensemble of computed topologies in more stochastic transition processes. One could envision applying p-Creode to tumor samples to measure how regulated the cell transitional process is, by comparing trajectories produced from the same dataset, and to measure how similar the transitional process is tumor to tumor, by comparing trajectories derived from separate datasets. Overall, we believe broad advances in single-cell data analysis, such as p-Creode, may have significant potential in a range of biomedical applications.

7.2. Future Relevance of p-Creode

Trajectory analysis is a competitive field with new algorithms being published regularly, pushing the innovation of the field forward. SPADE was the first multi-branch mapping algorithm (Bendall et al., 2011), demonstrating to the single-cell community the potential of trajectory mapping and inspiring development of new approaches. Following SPADE, algorithms like Wanderlust (Bendall et al., 2014) and Wishbone (Setty et al., 2016) were developed that traded mapping of complex trajectories for increases in robustness and analysis of analyte dynamics. More recent algorithms like Moncole2 (Qiu et al., 2017; Zunder et al., 2015; Briggs et al., 2017) have returned the focus of the field to mapping complex trajectories more robustly. Given the state of the field and p-Creode's ability to robustly map multi-branching topologies, I believe p-Creode currently ranks among the top performing algorithms, but given the rate of development in the field, this ranking could be fleeting. With that being said, I think there are attributes of p-Creode that make it notable. The sensitivity of p-Creode to detect alternate routes of tuft cell development between two similar tissues like the small intestine and colon was very surprising, demonstrating p-Creode's potential utility when comparing other similar tissues (e.g., wild-type tissue versus perturbed tissue). p-Creode's scoring is unique to the field allowing for an ensemble approach and the ability to compare trajectories across datasets. An ensemble approach prevents overfitting, something I believe gives an advantage when processing noisy data, particularly mass cytometry and multiplex imaging datasets. Quantifying trajectories across datasets has not been developed outside of p-Creode, and, in my opinion, could be, if any, p-Creode's lasting contribution by spurring new approaches.

Like any field, predicting the direction of innovation long term is difficult, but there are certain paths I believe are probable short term. For instance, tuning of an algorithm's parameters can be difficult for users not familiar with the mathematical details. Therefore, it is likely, future algorithms will contain more approachable parameters, making the tuning process more interpretable for users, or will optimize the tuning process, removing the need for user input altogether. This was a point of focus during development of p-Creode, where more approachable density based parameters were simplified from more complex parameters. Further development of p-Creode includes optimizing 2 of 3 input parameters to ease user burden. Due to the subjectiveness of identifying outliers from what is thought to be a rare cell type, the noise threshold parameter would be difficult to optimize; it should be noted, outlier removal is not an issue unique to p-Creode. Another possible direction of the field is to use multiple algorithms applying them according to their strengths, similar to how different clustering algorithms are currently utilized (Xu and Wunsch, 2010). For example, Wishbone might be used for trajectories with single bifurcations, and Monocle2 or p-Creode used for multi-branching topologies depending on if the data is noisy or not, Monocle2 for the former and p-Creode for the latter. Multiple algorithms could also be used in conjunction to independently verify results. With the current rate of development in the trajectory analysis field, the fate of these predictions, like p-Creode's fate, is likely to be determined (right or wrong) sooner rather than later.

BIBLIOGRAPHY

E.-A. D. Amir, K. L. Davis, M. D. Tadmor, E. F. Simonds, J. H. Levine, S. C. Bendall, D. K. Shenfeld, S. Krishnaswamy, G. P. Nolan, and D. Pe'er. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.*, 31(6):545–552, June 2013.

B. Anchang, T. D. P. Hart, S. C. Bendall, P. Qiu, Z. Bjornson, M. Linderman, G. P. Nolan, and S. K. Plevritis. Visualization and cellular hierarchy inference of single-cell data using SPADE. *Nat. Protoc.*, 11(7):1264–1279, July 2016.

M. Angelo, S. C. Bendall, R. Finck, M. B. Hale, C. Hitzman, A. D. Borowsky, R. M. Levenson, J. B. Lowe, S. D. Liu, S. Zhao, Y. Natkunam, and G. P. Nolan. Multiplexed ion beam imaging of human breast tumors. *Nat. Med.*, 20(4):436–442, Apr. 2014.

R. Bacher and C. Kendziorski. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology*, 17(1), Dec. 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0927-y. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0927-y.

N. Barker, J. H. van Es, J. Kuipers, P. Kujala, M. van den Born, M. Cozijnsen, A. Haegebarth, J. Korving, H. Begthel, P. J. Peters, and H. Clevers. Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature*, 449(7165):1003–1007, Oct. 2007.

M. Barron and J. Li. Identifying and removing the cell-cycle effect from single-cell RNA-Sequencing data. *Scientific Reports*, 6(1), Dec. 2016. ISSN 2045-2322. doi: 10.1038/srep33892. URL http://www.nature.com/articles/srep33892.

S. C. Bendall, E. F. Simonds, P. Qiu, E.-A. D. Amir, P. O. Krutzik, R. Finck, R. V. Bruggner, R. Melamed, A. Trejo, O. I. Ornatsky, R. S. Balderas, S. K. Plevritis, K. Sachs, D. Pe'er,

S. D. Tanner, and G. P. Nolan. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030):687–696, May 2011.

S. C. Bendall, K. L. Davis, E.-A. D. Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, and P. Dana. Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. *Cell*, 157(3):714–725, 2014.

C. Bezenon, J. le Coutre, and S. Damak. Taste-signaling proteins are coexpressed in solitary intestinal epithelial cells. *Chem. Senses*, 32(1):41–49, Jan. 2007.

C. Bezenon, A. Frholz, F. Raymond, R. Mansourian, S. Mtairon, J. Le Coutre, and S. Damak. Murine intestinal cells expressing Trpm5 are mostly brush cells and express markers of neuronal and inflammatory cells. *J. Comp. Neurol.*, 509(5):514–525, Aug. 2008.

L. Bittner. R. Bellman, Adaptive Control Processes. A Guided Tour. XVI + 255 S. Princeton, N. J., 1961. Princeton University Press. Preis geb. $ 6.50. *ZAMM Z. Angew. Math. Mech.*, 42(7-8):364–365, 1962.

M. Bjerknes, C. Khandanpour, T. Mry, T. Fujiyama, M. Hoshino, T. J. Klisch, Q. Ding, L. Gan, J. Wang, M. G. Martn, and H. Cheng. Origin of the brush cell lineage in the mouse intestinal epithelium. *Dev. Biol.*, 362(2):194–218, Feb. 2012.

C. Blanpain and E. Fuchs. Stem cell plasticity. Plasticity of epithelial stem cells in tissue regeneration. *Science (New York, N.Y.)*, 344(6189):1242281, June 2014. ISSN 1095-9203. doi: 10.1126/science.1242281.

P. Brennecke, S. Anders, J. K. Kim, A. A. Ko\lodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni, and M. G. Heisler. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods*, 10(11):1093–1095, Nov. 2013.

J. A. Briggs, V. C. Li, S. Lee, C. J. Woolf, A. Klein, and M. W. Kirschner. Mouse embryonic stem cells can differentiate via multiple paths to the same state. *Elife*, 6, Oct. 2017.

S. J. A. Buczacki, H. I. Zecchini, A. M. Nicholson, R. Russell, L. Vermeulen, R. Kemp, and D. J. Winton. Intestinal label-retaining cells are secretory precursors expressing Lgr5. *Nature*, 495(7439):65–69, Mar. 2013.

Z. Burke, S. Thowfeequ, M. Peran, and D. Tosh. Stem cells in the adult pancreas and liver. *Biochemical Journal*, 404(2):169–178, June 2007. ISSN 0264-6021, 1470-8728. doi: 10.1042/BJ20070167. URL http://biochemj.org/lookup/doi/10.1042/BJ20070167.

J. G. Camp, K. Sekine, T. Gerber, H. Loeffler-Wirth, H. Binder, M. Gac, S. Kanton, J. Kageyama, G. Damm, D. Seehofer, L. Belicova, M. Bickle, R. Barsacchi, R. Okuda, E. Yoshizawa, M. Kimura, H. Ayabe, H. Taniguchi, T. Takebe, and B. Treutlein. Multilineage communication regulates human liver bud development from pluripotency. *Nature*, 546 (7659):533–538, June 2017.

K. R. Campbell and C. Yau. Probabilistic modeling of bifurcations in single-cell gene expression data using a Bayesian mixture of factor analyzers. *Wellcome Open Res*, 2:19, Mar. 2017.

A. Cayley. A theorem on trees. In *The Collected Mathematical Papers*, pages 26–28.

J. R. Chubb, T. Trcek, S. M. Shenoy, and R. H. Singer. Transcriptional pulsing of a developmental gene. *Curr. Biol.*, 16(10):1018–1025, May 2006.

R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl. Acad. Sci. U. S. A.*, 102(21):7426–7431, May 2005.

B. Creamer, R. G. Shorter, and J. Bamforth. The turnover and shedding of epithelial cells. *Gut*, 2(2):110–116, 1961. ISSN 0017-5749. doi: 10.1136/gut.2.2.110. URL http://gut.bmj.com/content/2/2/110.

M. Crow, A. Paul, S. Ballouz, Z. J. Huang, and J. Gillis. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nature Communications*, 9(1), Dec. 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03282-0. URL http://www.nature.com/articles/s41467-018-03282-0.

K. E. Diggins, J. S. Gandelman, C. E. Roe, and J. M. Irish. Generating Quantitative Cell Identity Labels with Marker Enrichment Modeling (MEM). *Current Protocols in Cytometry*, 83:10.21.1–10.21.28, Jan. 2018. ISSN 1934-9300. doi: 10.1002/cpcy.34.

E. W. Dijkstra. A note on two problems in connexion with graphs. *Numer. Math.*, 1(1):269–271, 1959.

D. A. Edwards. The Structure of Superspace. In *Studies in Topology*, pages 121–133. 1975.

M. S. Engelstoft, M. L. Lund, K. V. Grunddal, K. L. Egerod, S. Osborne-Lawrence, S. S. Poulsen, J. M. Zigman, and T. W. Schwartz. Research Resource: A Chromogranin A Reporter for Serotonin and Histamine Secreting Enteroendocrine Cells. *Mol. Endocrinol.*, 29 (11):1658–1671, Nov. 2015.

S. P. Fahl, F. Coffey, and D. L. Wiest. Origins of T Cell Effector Subsets: A Riddle Wrapped in an Enigma. *The Journal of Immunology*, 193(9):4289–4294, 2014.

N. Fang and R. Akinci-Tolun. Depletion of Ribosomal RNA Sequences fromSingle-CellRNA-SequencingLibrary. *Current Protocols in Molecular Biology*, 115:7.27.1–7.27.20, July 2016. ISSN 1934-3647. doi: 10.1002/cpmb.11.

R. Finck, E. F. Simonds, A. Jager, S. Krishnaswamy, K. Sachs, W. Fantl, D. Pe'er, G. P. Nolan, and S. C. Bendall. Normalization of mass cytometry data with bead standards. *Cytometry A*, 83(5):483–494, May 2013.

S. Fre, M. Huyghe, P. Mourikis, S. Robine, D. Louvard, and S. Artavanis-Tsakonas. Notch signals control the fate of immature progenitor cells in the intestine. *Nature*, 435(7044): 964–968, June 2005.

K. L. Frieda, J. M. Linton, S. Hormoz, J. Choi, K.-H. K. Chow, Z. S. Singer, M. W. Budde, M. B. Elowitz, and L. Cai. Synthetic recording and in situ readout of lineage information in single cells. *Nature*, 541(7635):107–111, Jan. 2017.

L. A. Furchtgott, S. Melton, V. Menon, and S. Ramanathan. Discovering sparse transcription factor codes for cell states and state transitions during development. *Elife*, 6, Mar. 2017.

F. Gerbe and P. Jay. Intestinal tuft cells: epithelial sentinels linking luminal cues to the immune system. *Mucosal Immunol.*, 9(6):1353–1359, Nov. 2016.

F. Gerbe, J. H. van Es, L. Makrini, B. Brulin, G. Mellitzer, S. Robine, B. Romagnolo, N. F. Shroyer, J.-F. Bourgaux, C. Pignodel, H. Clevers, and P. Jay. Distinct ATOH1 and Neurog3 requirements define tuft cells as a new secretory cell type in the intestinal epithelium. *J. Cell Biol.*, 192(5):767–780, Mar. 2011.

F. Gerbe, E. Sidot, D. J. Smyth, M. Ohmoto, I. Matsumoto, V. Dardalhon, P. Cesses, L. Garnier, M. Pouzolles, B. Brulin, M. Bruschi, Y. Harcus, V. S. Zimmermann, N. Taylor, R. M. Maizels, and P. Jay. Intestinal epithelial tuft cells initiate type 2 mucosal immunity to helminth parasites. *Nature*, 529(7585):226–230, Jan. 2016.

M. J. Gerdes, C. J. Sevinsky, A. Sood, S. Adak, M. O. Bello, A. Bordwell, A. Can, A. Corwin, S. Dinn, R. J. Filkins, D. Hollman, V. Kamath, S. Kaanumalle, K. Kenny, M. Larsen, M. Lazare, Q. Li, C. Lowes, C. C. McCulloch, E. McDonough, M. C. Montalto, Z. Pang, J. Rittscher, A. Santamaria-Pang, B. D. Sarachan, M. L. Seel, A. Seppo, K. Shaikh, Y. Sui, J. Zhang, and F. Ginty. Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proc. Natl. Acad. Sci. U. S. A.*, 110(29):11982–11987, July 2013.

G. Giecold, E. Marco, S. P. Garcia, L. Trippa, and G.-C. Yuan. Robust lineage reconstruction from high-dimensional single-cell data. *Nucleic Acids Res.*, 44(14):e122, Aug. 2016.

C. Giesen, H. A. O. Wang, D. Schapiro, N. Zivanovic, A. Jacobs, B. Hattendorf, P. J. Schffler, D. Grolimund, J. M. Buhmann, S. Brandt, Z. Varga, P. J. Wild, D. Gnther, and B. Bodenmiller. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods*, 11(4):417–422, Apr. 2014.

D. Grn, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, and A. van Oudenaarden. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255, Sept. 2015.

M. Guo, E. L. Bao, M. Wagner, J. A. Whitsett, and Y. Xu. SLICE: determining cell differentiation and lineage based on single cell entropy. *Nucleic Acids Res.*, 45(7):e54, Apr. 2017.

L. Haghverdi, F. Buettner, and F. J. Theis. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31(18):2989–2998, Sept. 2015.

J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88–93, July 2004.

A. Haque, J. Engel, S. A. Teichmann, and T. Lnnberg. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.*, 9(1):75, Aug. 2017.

T. Hastie and W. Stuetzle. Principal Curves. *J. Am. Stat. Assoc.*, 84(406):502–516, June 1989.

C. A. Herring, A. Banerjee, E. T. McKinley, A. J. Simmons, J. Ping, J. T. Roland, J. L. Franklin, Q. Liu, M. J. Gerdes, R. J. Coffey, and K. S. Lau. Unsupervised Trajectory Analysis of

Single-Cell RNA-Seq and Imaging Data Reveals Alternative Tuft Cell Origins in the Gut. *Cell Syst*, 6(1):37–51.e9, Jan. 2018.

S. C. Hicks, F. W. Townes, M. Teng, and R. A. Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics (Oxford, England)*, Nov. 2017. ISSN 1468-4357. doi: 10.1093/biostatistics/kxx053.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, 24(6):417–441, 1933.

M. R. Howitt, S. Lavoie, M. Michaud, A. M. Blum, S. V. Tran, J. V. Weinstock, C. A. Gallini, K. Redding, R. F. Margolskee, L. C. Osborne, D. Artis, and W. S. Garrett. Tuft cells, taste-chemosensory cells, orchestrate parasite type 2 immunity in the gut. *Science*, 351(6279): 1329–1333, Mar. 2016.

A. Hyvrinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc., New York, USA, May 2001. ISBN 978-0-471-40540-5 978-0-471-22131-9. doi: 10.1002/0471221317. URL http://doi.wiley.com/10.1002/0471221317.

D. Hfer, B. Pschel, and D. Drenckhahn. Taste receptor-like cells in the rat gut identified by expression of alpha-gustducin. *Proc. Natl. Acad. Sci. U. S. A.*, 93(13):6631–6634, June 1996.

D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, and I. Amit. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779, Feb. 2014.

Z. Ji and H. Ji. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.*, 44(13):e117, July 2016.

L. Jiang, H. Chen, L. Pinello, and G.-C. Yuan. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biol.*, 17(1):144, July 2016.

T.-H. Kim, F. Li, I. Ferreiro-Neira, L.-L. Ho, A. Luyten, K. Nalapareddy, H. Long, M. Verzi, and R. A. Shivdasani. Broadly permissive intestinal chromatin underlies lateral inhibition and cell plasticity. *Nature*, 506(7489):511–515, Feb. 2014.

T.-H. Kim, A. Saadatpour, G. Guo, M. Saxena, A. Cavazza, N. Desai, U. Jadhav, L. Jiang, M. N. Rivera, S. H. Orkin, G.-C. Yuan, and R. A. Shivdasani. Single-Cell Transcript Profiles Reveal Multilineage Priming in Early Progenitors Derived from Lgr5(+) Intestinal Stem Cells. *Cell Rep.*, 16(8):2053–2060, Aug. 2016.

A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, May 2015.

U. Koch and F. Radtke. Mechanisms of T cell development and transformation. *Annu. Rev. Cell Dev. Biol.*, 27:539–562, July 2011.

Y. Korem, P. Szekely, Y. Hart, H. Sheftel, J. Hausser, A. Mayo, M. E. Rothenberg, T. Kalisky, and U. Alon. Geometry of the Gene Expression Space of Individual Cells. *PLoS Comput. Biol.*, 11(7):e1004224, July 2015.

K. S. Lau, V. Cortez-Retamozo, S. R. Philips, M. J. Pittet, D. A. Lauffenburger, and K. M. Haigis. Multi-scale in vivo systems analysis reveals the influence of immune cells on TNF--induced apoptosis in the intestinal epithelium. *PLoS Biol.*, 10(9):e1001393, Sept. 2012.

Y. Lavin, S. Kobayashi, A. Leader, E.-A. D. Amir, N. Elefant, C. Bigenwald, R. Remark, R. Sweeney, C. D. Becker, J. H. Levine, K. Meinhof, A. Chow, S. Kim-Shulze, A. Wolf, C. Medaglia, H. Li, J. A. Rytlewski, R. O. Emerson, A. Solovyov, B. D. Greenbaum, C. Sanders, M. Vignali, M. B. Beasley, R. Flores, S. Gnjatic, D. Pe'er, A. Rahman, I. Amit, and M. Merad. Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses. *Cell*, 169(4):750–765.e17, May 2017.

N. Leelatian, D. B. Doxie, A. R. Greenplate, B. C. Mobley, J. M. Lehman, J. Sinnaeve, R. M. Kauffman, J. A. Werkhaven, A. M. Mistry, K. D. Weaver, R. C. Thompson, P. P. Massion, M. A. Hooks, M. C. Kelley, L. B. Chambless, R. A. Ihrie, and J. M. Irish. Single Cell Analysis of Human Tissues and Solid Tumors with Mass Cytometry. *Cytometry B Clin. Cytom.*, July 2017a.

N. Leelatian, D. B. Doxie, A. R. Greenplate, J. Sinnaeve, R. A. Ihrie, and J. M. Irish. Preparing Viable Single Cells from Human Tissue and Tumors for Cytomic Analysis. *Curr. Protoc. Mol. Biol.*, 118:25C.1.1–25C.1.23, Apr. 2017b.

J. H. Levine, E. F. Simonds, S. C. Bendall, K. L. Davis, E.-a. D. Amir, M. D. Tadmor, O. Litvin, H. G. Fienberg, A. Jager, E. R. Zunder, R. Finck, A. L. Gedman, I. Radtke, J. R. Downing, D. Pe'er, and G. P. Nolan. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, 162(1):184–197, July 2015. ISSN 1097-4172. doi: 10.1016/j.cell.2015.05.047.

N. Li, A. Nakauka-Ddamba, J. Tobias, S. T. Jensen, and C. J. Lengner. Mouse Label-Retaining Cells Are Molecularly and Functionally Distinct From Reserve Intestinal Stem Cells. *Gastroenterology*, 151(2):298–310.e7, Aug. 2016.

J.-R. Lin, M. Fallahi-Sichani, and P. K. Sorger. Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. *Nat. Commun.*, 6:8390, Sept. 2015.

A. T. L. Lun, K. Bach, and J. C. Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17:75, Apr. 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0947-7.

L. v. d. Maaten and G. Hinton. Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, 9(Nov): 2579–2605, 2008.

E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, May 2015.

F. Mair, F. J. Hartmann, D. Mrdjen, V. Tosevski, C. Krieg, and B. Becher. The end of gating? An introduction to automated analysis of high dimensional cytometry data. *Eur. J. Immunol.*, 46(1):34–43, Jan. 2016.

E. Marco, R. L. Karp, G. Guo, P. Robson, A. H. Hart, L. Trippa, and G.-C. Yuan. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc. Natl. Acad. Sci. U. S. A.*, 111(52):E5643–50, Dec. 2014.

H. Matsumoto and H. Kiryu. SCOUP: a probabilistic model based on the Ornstein-Uhlenbeck process to analyze single-cell expression data during differentiation. *BMC Bioinformatics*, 17(1):232, June 2016.

A. McKenna, G. M. Findlay, J. A. Gagnon, M. S. Horwitz, A. F. Schier, and J. Shendure. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, 353(6298):aaf7907, July 2016.

E. T. McKinley, Y. Sui, Y. Al-Kofahi, B. A. Millis, M. J. Tyska, J. T. Roland, A. Santamaria-Pang, C. L. Ohland, C. Jobin, J. L. Franklin, K. S. Lau, M. J. Gerdes, and R. J. Coffey. Optimized multiplex immunofluorescence single-cell analysis reveals tuft cell heterogeneity. *JCI Insight*, 2(11), June 2017.

M. Middelhoff, C. B. Westphalen, Y. Hayakawa, K. S. Yan, M. D. Gershon, T. C. Wang, and M. Quante. Dclk1-expressing tuft cells: Critical modulators of the intestinal niche? *Am. J. Physiol. Gastrointest. Liver Physiol.*, page ajpgi.00073.2017, July 2017.

K. Miyazaki, M. Miyazaki, and C. Murre. The establishment of B versus T cell identity. *Trends Immunol.*, 35(5):205–210, May 2014.

V. Moignard, S. Woodhouse, L. Haghverdi, A. J. Lilly, Y. Tanaka, A. C. Wilkinson, F. Buettner, I. C. Macaulay, W. Jawaid, E. Diamanti, S.-I. Nishikawa, N. Piterman, V. Kouskoff, F. J. Theis, J. Fisher, and B. Gttgens. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.*, 33(3):269–276, Mar. 2015.

Y. Mori-Akiyama, M. van den Born, J. H. van Es, S. R. Hamilton, H. P. Adams, J. Zhang, H. Clevers, and B. de Crombrugghe. SOX9 is required for the differentiation of paneth cells in the intestinal epithelium. *Gastroenterology*, 133(2):539–546, Aug. 2007.

L. C. Murtaugh and D. A. Melton. Genes, signals, and lineages in pancreas development. *Annual Review of Cell and Developmental Biology*, 19:71–89, 2003. ISSN 1081-0706. doi: 10.1146/annurev.cellbio.19.111301.144752.

T. K. Noah, B. Donahue, and N. F. Shroyer. Intestinal development and differentiation. *Exp. Cell Res.*, 317(19):2702–2710, Nov. 2011.

F. Notta, S. Zandi, N. Takayama, S. Dobson, O. I. Gan, G. Wilson, K. B. Kaufmann, J. McLeod, E. Laurenti, C. F. Dunant, J. D. McPherson, L. D. Stein, Y. Dror, and J. E. Dick. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science*, 351(6269):aab2116, Jan. 2016.

D. Y. Orlova, N. Zimmerman, S. Meehan, C. Meehan, J. Waters, E. E. B. Ghosn, A. Filatenkov, G. A. Kolyagin, Y. Gernez, S. Tsuda, W. Moore, R. B. Moss, L. A. Herzenberg, and G. Walther. Earth Movers Distance (EMD): A True Metric for Comparing Biomarker Expression Levels in Cell Populations. *PLOS ONE*, 11(3):1–14, 2016. doi: 10.1371/journal.pone.0151859. URL https://doi.org/10.1371/journal.pone.0151859.

F. Paul, Y. Arkin, A. Giladi, D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, D. Winter, D. Lara-Astiaso, M. Gury, A. Weiner, E. David, N. Cohen, F. K. B. Lauridsen, S. Haas, A. Schlitzer,

A. Mildner, F. Ginhoux, S. Jung, A. Trumpp, B. T. Porse, A. Tanay, and I. Amit. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell*, 163(7): 1663–1677, Dec. 2015.

A. E. Powell, Y. Wang, Y. Li, E. J. Poulin, A. L. Means, M. K. Washington, J. N. Higginbotham, A. Juchheim, N. Prasad, S. E. Levy, Y. Guo, Y. Shyr, B. J. Aronow, K. M. Haigis, J. L. Franklin, and R. J. Coffey. The pan-ErbB negative regulator Lrig1 is an intestinal stem cell marker that functions as a tumor suppressor. *Cell*, 149(1):146–158, Mar. 2012.

P. Qiu, E. F. Simonds, S. C. Bendall, K. D. Gibbs, Jr, R. V. Bruggner, M. D. Linderman, K. Sachs, G. P. Nolan, and S. K. Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.*, 29(10):886–891, Oct. 2011.

X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, and C. Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods*, 14(10):979–982, Oct. 2017.

R. Remark, T. Merghoub, N. Grabe, G. Litjens, D. Damotte, J. D. Wolchok, M. Merad, and S. Gnjatic. In-depth tissue profiling using multiplexed immunohistochemical consecutive staining on single slide. *Sci Immunol*, 1(1):aaf6925, July 2016.

J. C. Ribot, A. deBarros, D. J. Pang, J. F. Neves, V. Peperzak, S. J. Roberts, M. Girardi, J. Borst, A. C. Hayday, D. J. Pennington, and B. Silva-Santos. CD27 is a thymic determinant of the balance between interferon-- and interleukin 17producing T cell subsets. *Nat. Immunol.*, 10 (4):427–436, 2009.

D. P. Riordan, S. Varma, R. B. West, and P. O. Brown. Automated Analysis and Classification of Histological Tissue Features by Multi-Dimensional Microscopic Molecular Profiling. *PLoS One*, 10(7):e0128975, July 2015.

A. H. Rizvi, P. G. Camara, E. K. Kandror, T. J. Roberts, I. Schieren, T. Maniatis, and

R. Rabadan. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nat. Biotechnol.*, 35(6):551–560, June 2017.

M. E. Rothenberg, Y. Nusse, T. Kalisky, J. J. Lee, P. Dalerba, F. Scheeren, N. Lobo, S. Kulkarni, S. Sim, D. Qian, P. A. Beachy, P. J. Pasricha, S. R. Quake, and M. F. Clarke. Identification of a cKit(+) colonic crypt base secretory cell that supports Lgr5(+) stem cells in mice. *Gastroenterology*, 142(5):1195–1205.e6, May 2012.

P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65, 1987.

N. Sasaki, N. Sachs, K. Wiebrands, S. I. J. Ellenbroek, A. Fumagalli, A. Lyubimova, H. Begthel, M. van den Born, J. H. van Es, W. R. Karthaus, V. S. W. Li, C. Lpez-Iglesias, P. J. Peters, J. van Rheenen, A. van Oudenaarden, and H. Clevers. Reg4+ deep crypt secretory cells function as epithelial niche for Lgr5+ stem cells in colon. *Proc. Natl. Acad. Sci. U. S. A.*, 113(37):E5399–407, Sept. 2016.

T. Sato, J. H. van Es, H. J. Snippert, D. E. Stange, R. G. Vries, M. van den Born, N. Barker, N. F. Shroyer, M. van de Wetering, and H. Clevers. Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. *Nature*, 469(7330):415–418, Jan. 2011.

P. See, C.-A. Dutertre, J. Chen, P. Gnther, N. McGovern, S. E. Irac, M. Gunawan, M. Beyer, K. Hndler, K. Duan, H. R. B. Sumatoh, N. Ruffin, M. Jouve, E. Gea-Mallorqu, R. C. M. Hennekam, T. Lim, C. C. Yip, M. Wen, B. Malleret, I. Low, N. B. Shadan, C. F. S. Fen, A. Tay, J. Lum, F. Zolezzi, A. Larbi, M. Poidinger, J. K. Y. Chan, Q. Chen, L. Rnia, M. Haniffa, P. Benaroch, A. Schlitzer, J. L. Schultze, E. W. Newell, and F. Ginhoux. Mapping the human DC lineage through the integration of high-dimensional techniques. *Science*, 356 (6342), June 2017.

J. Seita and I. L. Weissman. Hematopoietic stem cell: self-renewal versus differentiation. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, 2(6):640–653, Nov. 2010.

M. Setty, M. D. Tadmor, S. Reich-Zeliger, O. Angel, T. M. Salame, P. Kathail, K. Choi, S. Bendall, N. Friedman, and D. Pe'er. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.*, 34(6):637–645, June 2016.

S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.*, 31(1):64–68, May 2002.

J. Shin, D. A. Berg, Y. Zhu, J. Y. Shin, J. Song, M. A. Bonaguidi, G. Enikolopov, D. W. Nauen, K. M. Christian, G.-L. Ming, and H. Song. Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. *Cell Stem Cell*, 17(3):360–372, Sept. 2015.

N. F. Shroyer, D. Wallis, K. J. T. Venken, H. J. Bellen, and H. Y. Zoghbi. Gfi1 functions downstream of Math1 to control intestinal secretory cell subtype allocation and differentiation. *Genes Dev.*, 19(20):2412–2417, Oct. 2005.

A. J. Simmons, A. Banerjee, E. T. McKinley, C. R. Scurrah, C. A. Herring, L. S. Gewin, R. Masuzaki, S. J. Karp, J. L. Franklin, M. J. Gerdes, J. M. Irish, R. J. Coffey, and K. S. Lau. Cytometry-based single-cell analysis of intact epithelial signaling reveals MAPK activation divergent from TNF--induced apoptosis in vivo. *Mol. Syst. Biol.*, 11(10):835, Oct. 2015.

A. J. Simmons, C. R. Scurrah, E. T. McKinley, C. A. Herring, J. M. Irish, M. K. Washington, R. J. Coffey, and K. S. Lau. Impaired coordination between signaling pathways is revealed in human colorectal cancer using single-cell mass cytometry of archival tissue blocks. *Sci. Signal.*, 9(449):rs11, Oct. 2016.

P. W. Tetteh, O. Basak, H. F. Farin, K. Wiebrands, K. Kretzschmar, H. Begthel, M. van den Born, J. Korving, F. de Sauvage, J. H. van Es, A. van Oudenaarden, and H. Clevers. Replacement of Lost Lgr5-Positive Stem Cells through Plasticity of Their Enterocyte-Lineage Daughters. *Cell Stem Cell*, 18(2):203–213, Feb. 2016.

C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, 32(4):381–386, Apr. 2014.

B. Treutlein, D. G. Brownfield, A. R. Wu, N. F. Neff, G. L. Mantalas, F. H. Espinoza, T. J. Desai, M. A. Krasnow, and S. R. Quake. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, 509(7500):371–375, May 2014.

Y.-H. Tsai, K. L. VanDussen, E. T. Sawey, A. W. Wade, C. Kasper, S. Rakshit, R. G. Bhatt, A. Stoeck, I. Maillard, H. C. Crawford, L. C. Samuelson, and P. J. Dempsey. ADAM10 regulates Notch function in intestinal stem cells of mice. *Gastroenterology*, 147(4):822–834.e13, Oct. 2014.

C. A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, and J. C. Marioni. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nature Methods*, 14(6):565–571, June 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4292.

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

L. van der Maaten, L. van der Maaten, and G. Hinton. Visualizing non-metric similarities in multiple maps. *Mach. Learn.*, 87(1):33–55, 2011.

J. H. van Es, T. Sato, M. van de Wetering, A. Lyubimova, A. N. Y. Nee, A. Gregorieff, N. Sasaki, L. Zeinstra, M. van den Born, J. Korving, A. C. M. Martens, N. Barker, A. van Oudenaarden, and H. Clevers. Dll1+ secretory progenitor cells revert to stem cells upon crypt damage. *Nat. Cell Biol.*, 14(10):1099–1104, Oct. 2012.

K. L. VanDussen and L. C. Samuelson. Mouse atonal homolog 1 directs intestinal progenitors to secretory cell rather than absorptive cell fate. *Dev. Biol.*, 346(2):215–223, Oct. 2010.

K. L. VanDussen, A. J. Carulli, T. M. Keeley, S. R. Patel, B. J. Puthoff, S. T. Magness, I. T. Tran, I. Maillard, C. Siebel, \. Kolterud, A. S. Grosse, D. L. Gumucio, S. A. Ernst, Y.-H. Tsai, P. J. Dempsey, and L. C. Samuelson. Notch signaling modulates proliferation and differentiation of intestinal crypt base columnar stem cells. *Development*, 139(3):488–497, Feb. 2012.

J. von Moltke, J. Ming, L. Hong-Erh, and R. M. Locksley. Tuft-cell-derived IL-25 regulates an intestinal ILC2epithelial response circuit. *Nature*, 529(7585):221–225, 2015.

C. H. Waddington. *The Strategy of the Genes, a Discussion of Some Aspects of Theoretical Biology, by C. H. Waddington,... With an Appendix [Some Physico-chemical Aspects of Biological Organisation] by H. Kacser,..* 1957.

M. Wattenberg, F. Vigas, and I. Johnson. How to Use t-SNE Effectively. *Distill*, 1(10), Oct. 2016.

J. D. Welch, A. J. Hartemink, and J. F. Prins. SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biol.*, 17(1):106, May 2016.

J. D. Welch, A. J. Hartemink, and J. F. Prins. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol.*, 18 (1):138, July 2017.

C. B. Westphalen, S. Asfaha, Y. Hayakawa, Y. Takemoto, D. J. Lukin, A. H. Nuber, A. Brandtner, W. Setlik, H. Remotti, A. Muley, X. Chen, R. May, C. W. Houchen, J. G. Fox, M. D. Gershon, M. Quante, and T. C. Wang. Long-lived intestinal tuft cells serve as colon cancer-initiating cells. *J. Clin. Invest.*, 124(3):1283–1295, Mar. 2014.

R. Xu and D. C. Wunsch. Clustering algorithms in biomedical research: a review. *IEEE reviews in biomedical engineering*, 3:120–154, 2010. ISSN 1941-1189. doi: 10.1109/RBME.2010.2083647.

D. Yalcin, Z. M. Hakguder, and H. H. Otu. Bioinformatics approaches to single-cell analysis in developmental biology. *Molecular Human Reproduction*, 22(3):182–192, Mar. 2016. ISSN 1460-2407. doi: 10.1093/molehr/gav050.

K. S. Yan, O. Gevaert, G. X. Y. Zheng, B. Anchang, C. S. Probert, K. A. Larkin, P. S. Davies, Z.-F. Cheng, J. S. Kaddis, A. Han, K. Roelf, R. I. Calderon, E. Cynn, X. Hu, K. Mandleywala, J. Wilhelmy, S. M. Grimes, D. C. Corney, S. C. Boutet, J. M. Terry, P. Belgrader, S. B. Ziraldo, T. S. Mikkelsen, F. Wang, R. J. von Furstenberg, N. R. Smith, P. Chandrakesan, R. May, M. A. S. Chrissy, R. Jain, C. A. Cartwright, J. C. Niland, Y.-K. Hong, J. Carrington, D. T. Breault, J. Epstein, C. W. Houchen, J. P. Lynch, M. G. Martin, S. K. Plevritis, C. Curtis, H. P. Ji, L. Li, S. J. Henning, M. H. Wong, and C. J. Kuo. Intestinal Enteroendocrine Lineage Cells Possess Homeostatic and Injury-Inducible Stem Cell Activity. *Cell Stem Cell*, 21(1): 78–90.e6, July 2017.

S.-H. Yook, Z. N. Oltvai, and A.-L. Barabsi. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942, Apr. 2004.

Y. Yu, J. C. H. Tsang, C. Wang, S. Clare, J. Wang, X. Chen, C. Brandt, L. Kane, L. S. Campos, L. Lu, G. T. Belz, A. N. J. McKenzie, S. A. Teichmann, G. Dougan, and P. Liu. Single-cell RNA-seq identifies a PD-1ilc progenitor and defines its development pathway. *Nature*, 539 (7627):102–106, Nov. 2016.

G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8:14049, Jan. 2017.

C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, and W. Enard. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell*, 65(4):631–643.e4, Feb. 2017.

P. Zrazhevskiy and X. Gao. Quantum dot imaging platform for single-cell molecular profiling. *Nat. Commun.*, 4:1619, 2013.

E. R. Zunder, E. Lujan, Y. Goltsev, M. Wernig, and G. P. Nolan. A continuous molecular roadmap to iPSC reprogramming through progression analysis of single-cell mass cytometry. *Cell Stem Cell*, 16(3):323–337, Mar. 2015.