Machine learning models of cell differentiation processes with single-cell transcriptomic measurements

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Revant Gupta
aus Allahabad, Indien

Tübingen
2023

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	25.09.2023
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Manfred Claassen
2. Berichterstatter:	Prof. Dr. Stephan Ossowski

Summary

Dynamic biological phenomena such as the development of immunity due to vaccination or the division of a single zygote into the ∼37 trillion cells in an adult human are triggered and driven by bio-molecular interactions. The bio-molecular species involved in these interactions are categorised based on their molecular properties and physiological function. Typically, the abundance or characteristics of only a single category of molecular species are measured in experimental protocols, and the data generated is noisy, biased and incomplete.

Due to the limitations of measurement technology, computational models cannot represent bio-molecular interactions in full mechanistic detail and have to be restricted to operational definitions of complex biological phenomena. Despite these constraints, computational models tailored to the idiosyncracies of data generated by various technologies enable the identification of bio-molecular species and interactions relevant to particular biological processes.

A cell is composed of various bio-molecular species such as nucleic acids, proteins, metabolites etc. The entire bio-molecular composition of a cell is known as a cell-state. mRNA are polymeric bio-molecules whose sequence encodes information for the production of proteins. While proteins are ultimately responsible for the execution of cellular functions, mRNA can be measured much more comprehensively with single-cell RNA sequencing technology. mRNA sequences corresponding to different protein segments are called transcripts, and the relative abundance of the various transcripts indicates the functional properties of the cell. Therefore, the cell-state can be approximated as a vector of mRNA transcript abundance.

The change of the cell-state over the course of a biological process is called differentiation. This thesis presents three models of cell differentiation and their application for different scRNAseq. experimental protocols and discovery goals. The first two models are based on the simulation of cell differentiation with Markov chains. The first model provides a generally applicable trajectory inference approach to model differentiation in any biological system with no topological constraints. The second model utilises simulations to model differentiation as a latent state-space process and is used to cluster cells based on transcriptional activity in order to identify transitional cell-states. The third model is based on ordinal logistic regression and is used to identify transcripts whose expression varies along a specified ordinal axis, even in data with other prominent sources of variation.

Zusammenfassung

Dynamische biologische Phänomene wie die Entwicklung von Immunität durch eine Impfung oder die Teilung einer einzigen Zygote in die 37 Billionen Zellen eines erwachsenen Menschen werden durch biomolekulare Wechselwirkungen ausgelöst und vorangetrieben. Die an diesen Wechselwirkungen beteiligten biomolekularen Spezies werden anhand ihrer molekularen Eigenschaften und physiologischen Funktion kategorisiert. In der Regel werden in Versuchsprotokollen nur die Häufigkeit oder die Eigenschaften einer einzigen Kategorie von Molekülarten gemessen, und die dabei gewonnenen Daten sind verrauscht, verzerrt und unvollständig.

Aufgrund der Beschränkungen der Messtechnik können Computermodelle die biomolekularen Wechselwirkungen nicht in allen mechanistischen Details darstellen und müssen sich auf operative Definitionen komplexer biologischer Phänomene beschränken. Trotz dieser Einschränkungen ermöglichen Computermodelle, die auf die Besonderheiten der mit verschiedenen Technologien erzeugten Daten zugeschnitten sind, die Identifizierung von biomolekularen Spezies und Wechselwirkungen, die für bestimmte biologische Prozesse relevant sind.

Eine Zelle besteht aus verschiedenen biomolekularen Spezies wie Nukleinsäuren, Proteinen, Metaboliten usw. Die gesamte biomolekulare Zusammensetzung einer Zelle wird als Zellzustand bezeichnet. mRNA sind polymere Biomoleküle, deren Sequenz Informationen für die Herstellung von Proteinen kodiert. Während Proteine letztlich für die Ausführung zellulärer Funktionen verantwortlich sind, kann mRNA mit der Einzelzell-RNA-Sequenzierungstechnologie viel umfassender gemessen werden. mRNA-Sequenzen, die verschiedenen Proteinsegmenten entsprechen, werden als Transkripte bezeichnet, und die relative Häufigkeit der verschiedenen Transkripte gibt Aufschluss über die funktionellen Eigenschaften der Zelle. Daher kann der Zellzustand als ein Vektor der mRNA-Transkript-Häufigkeit angenähert werden.

Die Veränderung des Zellzustands im Verlauf eines biologischen Prozesses wird als Differenzierung bezeichnet. In dieser Arbeit werden drei Modelle der Zelldifferenzierung und ihre Anwendung für verschiedene scRNAseq.-Experimentierprotokolle und Forschungsziele vorgestellt. Die ersten beiden Modelle basieren auf der Simulation der Zelldifferenzierung mit Markov-Ketten. Das erste Modell bietet einen allgemein anwendbaren Ansatz zur Trajektorieninferenz, um die Differenzierung in jedem biologischen System ohne topologische Einschränkungen zu modellieren. Das zweite Modell nutzt Simulationen, um die

Differenzierung als Prozess in einem latenten Zustandsraum zu modellieren, und wird verwendet, um Zellen auf der Grundlage der Transkriptionsaktivität zu gruppieren, um Übergangszellzustände zu identifizieren. Das dritte Modell basiert auf einer ordinalen logistischen Regression und wird verwendet, um Transkripte zu identifizieren, deren Expression entlang einer bestimmten ordinalen Achse variiert, selbst in Daten mit anderen auffälligen Variationsquellen.

Acknowledgements

The solitary years spent in service of my doctoral research have been the most challenging of my life, and I would not have been able to complete this journey without the support of the people that matter the most.

I express my deepest gratitude to the pillars of my life, my parents, whose love, encouragement, and sacrifices have been the cornerstone of my pursuit of knowledge.

To my beloved wife, your unwavering patience, understanding, and constant motivation have been my driving force. Your belief in my dreams gave me the strength to persevere through long nights and demanding days. Your love has been my refuge, and I am profoundly grateful for your presence in my life.

To my colleagues, my friends, your camaraderie, discussions, and shared experiences have enriched my academic journey and made the hardest years more manageable.

I extend my sincere appreciation to my supervisor, Manfred, who has constantly stood by me throughout our convoluted academic path. I am deeply grateful for the opportunity to pursue my ambitions under your guidance.

At the close of this endeavour, I have come to realize that hardships are not stumbling blocks but stepping stones that propel us toward growth and self-discovery. Each obstacle has provided an opportunity for me to refine my skills, enhance my resilience, and evolve into a stronger individual. May the bonds we share continue to inspire and uplift me as I embark on new chapters in my academic and personal journey.

With heartfelt thanks,

Revant

Contents

List of Publications

1. **R. Gupta**, D. Cerletti, G. Gut, A. Oxenius, and M. Claassen, *Simulation based inference of differentiation trajectories from RNA velocity fields*, Cell Reports Methods, vol. 2, p. 100359, Dec. 2022

   Contribution: I conceptualised, developed & implemented the model, performed the analysis, wrote the manuscript and produced the figures.

2. D. Cerletti, I. Sandu, **R. Gupta**, A. Oxenius, and M. Claassen, *Fate trajectories of CD8+ T cells in chronic LCMV infection*, bioRxiv, Dec. 2020

   Contribution: I performed the trajectory analysis and produced relevant plots & text for the manuscript.

3. **R. Gupta** and M. Claassen, *Factorial state space modelling for kinetic clustering and lineage inference*, bioRxiv, Aug. 2023

   Contribution: I conceptualised, developed & implemented the model, performed the analysis, wrote the manuscript and produced the figures.

4. W. Macnair, **R. Gupta**, and M. Claassen, *psupertime: supervised pseudotime analysis for time-series single-cell RNA-seq data*, Bioinformatics, vol. 38, pp. i290–i298, June 2022

   Contribution: I performed the comparative analysis of the model with other methods, run-time analysis with respect to the size of the dataset and produced relevant plots & text for the manuscript.

5. J. Bues, M. Biočanin, J. Pezoldt, R. Dainese, A. Chrisnandy, S. Rezakhani, W. Saelens, V. Gardeux, **R. Gupta**, R. Sarkis, J. Russeil, Y. Saeys, E. Amstad, M. Claassen, M. P. Lutolf, and B. Deplancke, *Deterministic scRNA-seq captures variation in intestinal crypt and organoid composition*, Nature Methods, vol. 19, pp. 323–330, Feb. 2022

   Contribution: I performed the psupertime analysis and produced relevant text for the manuscript.

Abbreviations

DNA	DioxyriboNucleic acid
RNA	RiboNucleic acid
mRNA	messenger RNA
scRNAseq.	single-cell RNA sequencing
MST	Minimum Spanning Tree
UMAP	Uniform Manifold Approximation and Projection
PCA	Principal Component Analysis
LCMV	Lymphocytic ChorioMeningitis Virus
TCR	T-Cell Receptor

1. Introduction

1.1 Tissue development and organisation

Complex, multi-cellular life forms rely on the coordinated activity of hundreds of specialised cell-types[1]. In the early days of cell biology, after the invention of the compound microscope and subsequent discovery of cells as the atomic unit of biological organisation, cells were characterised based on morphology, tissue localisation and sub-cellular organisation. The discovery of the molecular, biochemical basis of biological activity ushered in a new understanding of a cell-type. The current view includes bio-molecular composition and the corresponding impact on biological function as criteria for cell-type categorisation [2].

Cell differentiation is the process by which cells change in bio-molecular composition and, therefore, in biological functionality. The change in biological functionality leads to the development of specialised cell-types that multi-cellular organisms are composed of. For example, the development of the human body from a single zygote is a cell differentiation process. The process by which naive T-cells are activated and acquire cytotoxic functions in response to a viral infection is also an example. Even cell division can be framed as a differentiation process considering cell-cycle stages analogous to cell-types. In general, cell differentiation involves the emergence of cells with distinct functional characteristics, or cell-types, from cells of a different type [3].

The concept of discrete cell-types defined based on tissue functions, morphology, and bio-molecular composition is challenged by considerable cell-to-cell variation within cell-types. This variation can be attributed to individual genetics, the influence of the external environment and the prior cellular micro-environment. In the context of differentiation, such variation suggests that cell-types arise from gradual changes in bio-molecular composition rather than discrete shifts. A cell-state is the exact bio-molecular composition of a cell without reference to its functional identity or other cells. The ontological relationship between cell-types defined on common functional or phenotypic properties and cell-states defined on bio-molecular composition has not been fully resolved [2].

[1] Cell ontology of the human cell atlas already categorises ∼1900 cell-types and their relationships [1].

1.2 Molecular measurements in single-cell biology

The central dogma of molecular biology describes the hierarchical relationship of bio-molecular species within a living system [4]. The study of different bio-molecular species is comprehensively organised into various *omics* domains. Genomics is the study of the genetic code (DNA), which is common to all cells in an organism. Transcriptomics is the study of the transcriptional products of the genome, which are RNA molecules. The biological activity of a cell is regulated by the differential expression of genes leading to differences in function between cell-types. messenger RNA (mRNA) are transcriptional products of gene expression that encode information for the synthesis of proteins. Proteins are molecular *machines* that execute physiological processes. Proteomics involves the study of proteins, their structure, function, and interactions. Epigenomics is the study of non-structural modifications of the genetic code that regulate gene expression, such as DNA methylation and histone modifications [5].

Genomic and transcriptomic measurement technologies have improved dramatically in scale, mainly due to the super-exponential decrease in sequencing costs and the development of microfluidics. These technologies are commonly called high throughput sequencing technologies [5]. The advent of single-cell resolution measurements has enabled highly resolved investigations of cell-state changes in the context of differentiation. In particular, single-cell RNA sequencing (scRNAseq.) technologies measure the expression of the entire set of genes/transcripts at cellular resolution, providing a broad and detailed view of the transcriptional composition of a cell. In contrast, single-cell proteomic measurements are restricted in the number of distinct species that can be measured and single-cell epigenomic measurements are highly sparse. Therefore, single-cell transcriptomic measurements via RNA sequencing have been widely adopted to study biological systems at single-cell resolution [6]. scRNAseq. measurements also have additional properties that make them particularly suited to modelling differentiation processes that will be detailed in the forthcoming sections.

1.3 Conceptual models of cell differentiation

C. H. Waddington's epigenetic landscape model has been very influential in conceptualising cell differentiation and continues to be widely used. The epigenetic landscape represents transitions toward specialised cell-types as a slope with valleys and ridges. In this analogy, cells act as marbles released from the top of the slope. The marbles will roll down and, despite variation in initial positioning, be canalised into the valleys. As the marbles roll down, the number of potential

paths they could take reduces, reflecting the increasing specialisation of cells and an irreversible commitment. The marbles come to rest in positions with no downward inclination, representing the terminal cell-types beyond which cells no longer differentiate [7].

While not entirely correct[2], the landscape analogy provides a useful abstraction of the regulatory mechanisms that drive cell-type transitions. The biological activity of a cell is the result of a complex biochemical interaction; however, experimental bio-molecular techniques do not produce comprehensive measurements of the bio-molecular composition of cell-states and are typically restricted to partial measurements of a single bio-molecular species. Therefore, the abstraction[3] of molecular mechanisms is essential in modelling differentiation processes from experimental data[4]. For example, the change in expression of transcripts is regulated by a class of proteins called transcription factors. Transcription factors are themselves the products of corresponding transcripts. In general, the effect of the expression of a transcript on another is indirect and mediated via intermediate bio-molecules like proteins. However, a model of cell differentiation based on scRNAseq. data can only represent the interaction between transcripts as direct interactions [11]. Furthermore, information such as the epigenetic state of the genome, which governs the allele variant expressed, cannot be represented at all.

1.4 Differentiation models with single-cell transcriptomics

In the context of transcriptomics, a cell-state is the amount of expression of all genes/transcripts in a cell. Differentiation processes are interpreted as sustained changes in gene/transcript expression. Change in expression leads to a change in functional properties of the cell, thus giving rise to a different cell-type. In contrast to the description of cell differentiation as transitions between cell-types, models developed on single-cell measurements typically consider transitions between cell-states. Interpretation of differentiation in terms of cell-type transitions, along with the development of models of cell-state transitions, presents unresolved challenges.

[2]Although Waddington's landscape would appear to suggest an analogy to the change in potential energy of marbles rolling down a hill, several mechanisms exist by which cells *lower* in the landscape can transition into cells *upper* in the landscape or transition laterally *over ridges*. The de-differentiation of terminal cells into stem-like cancerous cells and cellular reprogramming of skin cells into induced pluripotent stem cells are two such examples [8][9].

[3]Conceptual abstraction being the process of selecting aspects of a concept relevant to a specific purpose.

[4]Experimental protocols that allow for the measurement of multiple *omics* modalities are increasingly common. Such data-sets are expected to enable the inference of biologically causal mechanisms, thus reducing the need for abstraction [10].

A common approach is to cluster cell-states and identify clusters of cell-states as functional cell-types. Subsequently, a model of cell-type transitions is developed, and variability within clusters is discarded as biological or technical stochasticity [12]. Alternatively, differentiation can be modelled with greater resolution. In the latter formulation, change in expression is gradual and measured cell-states may represent transitional states that are possibly intermediate in both expression and function. For differentiation processes with multiple outcomes, an incremental commitment towards terminal fates is suggested [13][14].

scRNAseq. measurements read out the number of molecules of each gene/transcript in a single cell. The process of gene expression is influenced by individual genetics, the epigenetic state of the genome and external conditions. Models of cell differentiation seek to distinguish sources of variation in cell-states that represent a change in biological activity and functionality. It is implicitly assumed that not all variation in cell-states is physiologically meaningful. Conversely, whether experimental techniques capture the entire spectrum of cell-states in a differentiation process is unknown.

Descriptive models of differentiation processes characterise cell-state transitions along the differentiation axis and can be used to discover relevant genes in an associative fashion. It is relatively difficult to model the data-generating process even with modelling approaches agnostic to mechanistic accuracy. For example, whether interpolating measured cell-states or generative models [15][16] would produce physiologically viable cell-states is unclear. Thus far, no universal functional mapping between expression and biological activity has been established. Furthermore, experimental investigation of predicted or simulated cell-states remains challenging due to the absence of experimental tools capable of inducing exact expression profiles[5].

Conceptually, scRNAseq. measurements enable the development of highly resolved models of expression dynamics. In practice, measurement noise, biological stochasticity and the phenomenon of dropout[6] lead to a trade-off between resolution and confidence. In addition, technical variability between data from different experiments is high and is referred to as a batch-effect[7].

[5]Perturbation screens provide a degree of control over gene expression and can be used to investigate the biology of induced cell-states however, these interventions affect expression levels coarsely [17].

[6]single-cell RNA sequencing protocols begin with the isolation of single-cells, followed by the capture of individual mRNA molecules. Lastly, a data matrix is entered with the number of molecules of each gene/transcript for each cell. This matrix is very sparse, and the current understanding of this phenomenon is that most missing values represent a technical inability to capture mRNA molecules rather than an absence of transcription. The sparsity precludes partitioning of cells on a restricted selection of well-characterised genes.

[7]Batch-effects are a combined effect of stochastic biological differences between samples, dif-

scRNAseq measurements can be conducted using either a time-series protocol or single samples called snapshots. In time-series experiments, biological samples are collected at regular intervals over a period of time. The expression dynamics over time, possibly for multiple cell-types, can be modelled with such data. The activation of differentiation in biological systems is often asynchronous; therefore, individual samples in both protocols will measure a mixture of cell-states at various stages of the differentiation process.

Models of cell differentiation, commonly called lineage inference or trajectory inference, aim to order the measured cell-states along an estimated differentiation coordinate known as pseudotime representing the continuum of differentiation. Models developed on time-series data can additionally utilise temporal labels associated with each sample as side information, prior knowledge, or for supervised learning. Time-series data also enables modelling changes in different cell-types' frequency as differentiation progresses. If multiple differentiation processes are active in the biological sample being measured, then in addition to pseudotime, cell-states are assigned to co-occurring lineages, each corresponding to one terminal fate [18].

1.5 Transcriptional dynamics-based differentiation models

The life-cycle of mRNA molecules comprises three steps, first is the transcription of the DNA sequence to produce unspliced mRNA transcripts that contain both exonic and intronic sequences. In the second step, unspliced transcripts are spliced to create a mature spliced transcript that can be read by a ribosome. Finally, mRNA transcripts are degraded by ribonuclease back into individual nucleotides.

scRNAseq. protocols are designed to measure spliced transcripts; however, through several mechanisms, unspliced transcripts are also measured[8]. The detection of both mature, spliced mRNA and nascent, unspliced mRNA has been used to develop models that make a local prediction of a cell's future state. These models are commonly known as RNA velocity [20]. RNA velocity models produce estimates of each gene's rate of change of expression by fitting parameters for the transcription-splicing process described above.

Several challenges of this modelling approach have been outlined, primarily on measurement biases due to unintentional capture of unspliced transcripts and the over-simplification of biological processes [21]. Despite these issues[9], RNA velocity

ferences in scRNAseq. protocol, technical variability due to different sequencing runs and differences in library preparation.

[8]The mechanisms by which unspliced transcripts are measured, and the proportion of such measurements vary depending on the scRNAseq. protocol [19].

[9]RNA velocity modelling is an active topic of computational and experimental research.

models have been useful in modelling cell differentiation with scRNAseq. data.

Models that do not utilise RNA velocity rely on distance measures to arrange cells into a pseudo-temporal ordering. The lack of directed cell-state transitions limits the ability to model the directionality of repeating or convergent differentiation patterns. Furthermore, the overall direction of differentiation has to be inferred from the biological context, an approach that can be challenging to apply to under-characterised biological systems. RNA velocity-based models consider the rate of change of expression, therefore, can infer pseudo-temporal ordering reflecting the temporal distance between cell-states rather than approximate it with distance in gene expression. This leads to a pseudotime that is a better representation of differentiation dynamics.

Protocols specifically designed to capture this signal combined with more sophisticated modelling of the transcription-splicing process have been published since the original publication by La Manno et. al. in 2018 [22][23].

2. Objectives

The ultimate goal of differentiation modelling is to identify transitional relationships between cell-states and cell-types and aid in the identification of relevant bio-molecular interactions. scRNAseq. measurements performed for the study of various biological conditions under different experimental protocols differ in the availability of prior biological knowledge, data from other modalities and contextual information. This necessitates the development of models of differentiation that address specific research goals. I developed and applied differentiation models during my doctoral research, particularly utilising RNA velocity.

Trajectory inference methods either cannot model non-tree-like differentiation patterns that require revisiting cell-states or use specialised approaches only applicable to cyclical patterns. Since such models do not have a signal for directed cell-state transitions, they do not discriminate between cell-states with similar expression but different transitions. Typically, biological context is required to assign the overall directionality of the modelled lineage. I developed a model that utilises RNA velocity to overcome the limitations of conventional methods. The model performs trajectory inference with no assumptions about the topology of the process and is generally applicable, including cyclical and convergent patterns. The model allows data-driven identification of differentiation processes and therefore finds application in under-characterised biological systems [Manuscript 1].

Due to the high variability and dropout, the functional identity of cell-states is typically established by clustering cell-states and then identifying clusters as cell-types. While aiding interpretability, the exclusive assignment of cell-states to cell-types may hinder the discovery of transitional cell-states. Clustering cell-states on only expression does not utilise dynamical information from RNA velocity. I developed a latent state-space model that models differentiation dynamics in a discrete latent state-space. The model can perform kinetic *soft* clustering of cells which aids the discovery of transitional cell-states while also producing highly resolved estimates of pseudotime and cell-fate.[Manuscript 3].

Prior information on the ordering of cell-types or the sequential order of scRNAseq. data from multiple samples, if utilised as input, can enable the characterisation of differentiation processes even if there are large technical or unrelated biological sources of variation. I benchmarked and applied a regularised ordinal logistic regression-based modelling approach for estimating pseudotime and identifying genes in scRNAseq. data with ordinal labels [Manuscript 4].

3. Results

The following sections summarise the manuscripts included in this thesis. Each section covers one differentiation model and consists of the primary motivation for developing the model, the model specification and a description of analyses of scRNAseq. data with the model to demonstrate key utilities.

3.1 Simulation-based trajectory inference

RNA velocity models estimate the rate of change of gene expression of cell-states measured in a scRNAseq. experiment. This estimated *velocity* can be used to predict the future state of a cell; however, since velocity cannot be assumed to be constant over time, only local predictions can be made. Assuming that the measured cell-states are representative of cell-states that arise during the differentiation process, the differentiation process can be simulated as a Markov chain of cell-state transitions [20]. The probability of transitioning from cell c_a to cell c_b, p_{ab} is estimated as following,

$$p_{ab} \propto e^{\cos(\theta_{ab})} \tag{3.1}$$

where θ_{ab} is the angle between the difference of gene expression vectors c_a, and c_b, and the velocity vector of c_a.

Manuscript 1 reports Cytopath, a data-driven trajectory inference method that utilises properties of RNA velocity to infer pseudotime, trajectories and cell-fate probabilities with no assumptions on the topology of the differentiation process. The method aims to improve upon prior trajectory inference methods, particularly for non-tree-like topologies like cycles, convergence and interlaced processes, by incorporating the directional signal from RNA velocity.

3.1.1 Cytopath: Modelling differentiation with Markov chains

Cytopath estimates trajectories from root to terminal cell-states of a differentiation process [Figure 3.1A]. The method involves four steps. First, multiple simulations of the differentiation processes are produced by Markov sampling of cell-state sequences initialised at pre-defined root states [Figure 3.1B.1].

The cell-state c_{ij} at step i of the simulation is selected randomly according to the transition probability matrix T from the nearest neighbours of c_{i-1}. Let F be

the cumulative probability distribution of T. A value κ is sampled from a uniform distribution over $[0, 1)$ and,

$$c_i = \arg\min_{c \in \mathcal{C}} (F(\tfrac{c}{c_{i-1}}) - \kappa) \ni F(\tfrac{c}{c_{i-1}}) \geq \kappa \tag{3.2}$$

Second, simulations with a common terminal state are aligned using Dynamic Time Warping to establish a common differentiation coordinate (pseudotime) [Figure 3.1B.2]. Third, consensus expression states are estimated by averaging cell-states at each step of the aligned sequence to obtain the trajectory. Last, cell-states are then assigned to the trajectory; for a cell with neighbours K, its alignment score to step i of a trajectory is calculated as,

$$\xi_i^f = \frac{1}{|K|} \sum_k^K \cos(\eta_k^f) \cdot \exp(\gamma_k) \tag{3.3}$$

where η is the cosine angle between the section of the trajectory and all possible transition partners $k \in K$ of the cell. γ is the cosine similarity between the velocity vector of the cell with the distance vector between the cell and its neighbours [Figure 3.1B.3]. Following the assignment, pseudotime and cell-fate scores are estimated per cell [Figure 3.1C.2-3].

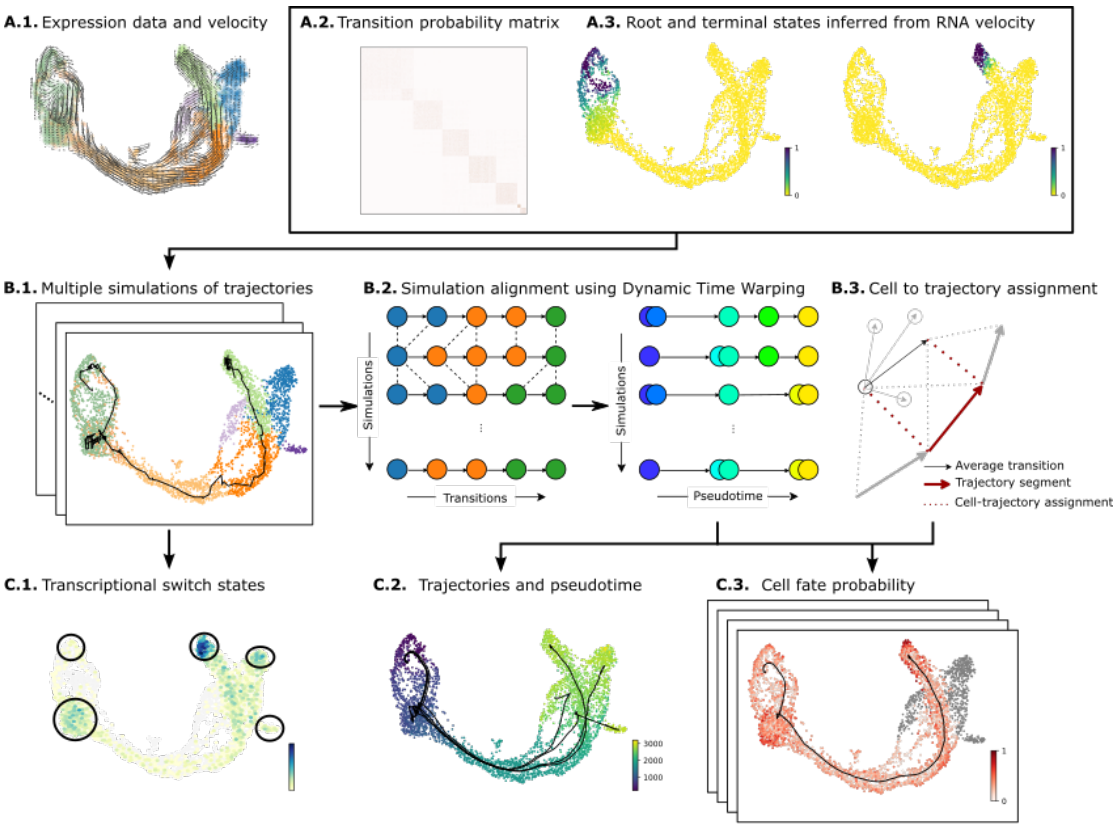


Figure 3.1: **Cytopath overview.** **(A)** Inputs for Cytopath trajectory inference subsequent to a RNA velocity analysis. (A.1.) Single-cell gene expression profiles, RNA velocity profiles, (A.2.) Transition probability matrix, (A.3.) Root and terminal state annotation. (shown here: inferred using RNA velocity) **(B)** Steps performed during Cytopath inference. (B.1.) Simulations of the differentiation process generated by sampling a Markov chain based on the cell-to-cell transition probabilities. Sampling is initialized on cells annotated as root states. (B.2.) Simulations are performed for a fixed number of steps that are automatically selected using the properties of the transition probability matrix. Simulations are aligned using Dynamic Time Warping. After alignment, cells at each transition step represent the same consensus state. (B.3.) Cells along the inferred trajectory are assigned to multiple trajectory segments based on the alignment of their average transition vector (with respect to neighbours) and the trajectory segment. **(C)** Outputs from Cytopath trajectory inference. (C.1.) The frequency of simulations terminating at each cell highlights regions of switch in transcriptional programs as well as terminal regions. (C.2.) Trajectories are inferred independently for each terminal region. The trajectories are composed of multiple segments. The pseudotime of a cell is estimated as the weighted average segment rank of all the segments it aligns with. (C.3.) Differential alignment scores to multiple trajectories are used to estimate the cell-fate probability with respect to the terminal regions.

Figure reproduced from figure 1 of manuscript 1.

3.1.2 Modelling non-tree-like differentiation topologies

Modelling of tree-like differentiation topologies can be accomplished by assuming that a greater difference in gene expression between cell-states corresponds to a greater difference in progress along the differentiation axis. Cell-states with divergent gene expression trends may be grouped into lineages representing parallel differentiation processes in the sample. RNA velocity-enabled trajectory inference may better estimate branch points in such data. However, methods that assume the differentiation topology can be modelled as a Minimum Spanning Tree (MST) are conceptually suitable.

However, modelling convergent differentiation topologies requires an increase in the similarity of gene expression profiles of cell-states to correspond to the progress of the differentiation process. Cyclical processes require modelling oscillating patterns of gene expression. Conceptually, Cytopath enables modelling such topologies by relying on directed transitions estimated using RNA velocity, that are asymmetric between pairs of cell states. Therefore, sequences of cell-state transitions sampled from the transition probability matrix can *navigate* through gene expression space that may appear isotropic.

Cell cycle reconstruction

Cytopath's utility in modelling differentiation processes with a cycle followed by further linear differentiation was demonstrated with an analysis of scRNAseq. data of cells undergoing cell-cycling [24]. Cell-cycle stages were experimentally annotated based on the fluorescence intensity of Green and Red fluorescent protein tagged proteins. Cell cycle stage annotations were only used to validate the analysis. Trajectory inference with Cytopath modelled an unbroken trajectory starting in the G1 stage (root states inferred with RNA velocity) through the intermediate stages back to the G1 stage and further into the G1-checkpoint stage [Figure 3.2C].

Cells in the G1 stage can be partitioned into two groups based on pseudotime inferred by Cytopath [Figure 3.2G]. The expression of markers associated with cell cycle is significantly higher in early-pseudotime G1 cells than in those committed towards the G1 checkpoint phase and, accordingly, are associated with higher pseudotime [Figure 3.2H]. The partitioning of cells in the G1 stage and the difference in marker expression can be observed as two separate bands of G1 cells (blue) in the radial plot [Figure 3.2A]. Comparative analysis with non-velocity-based methods highlighted issues arising from the inability to incorporate RNA velocity (see Manuscript 1).

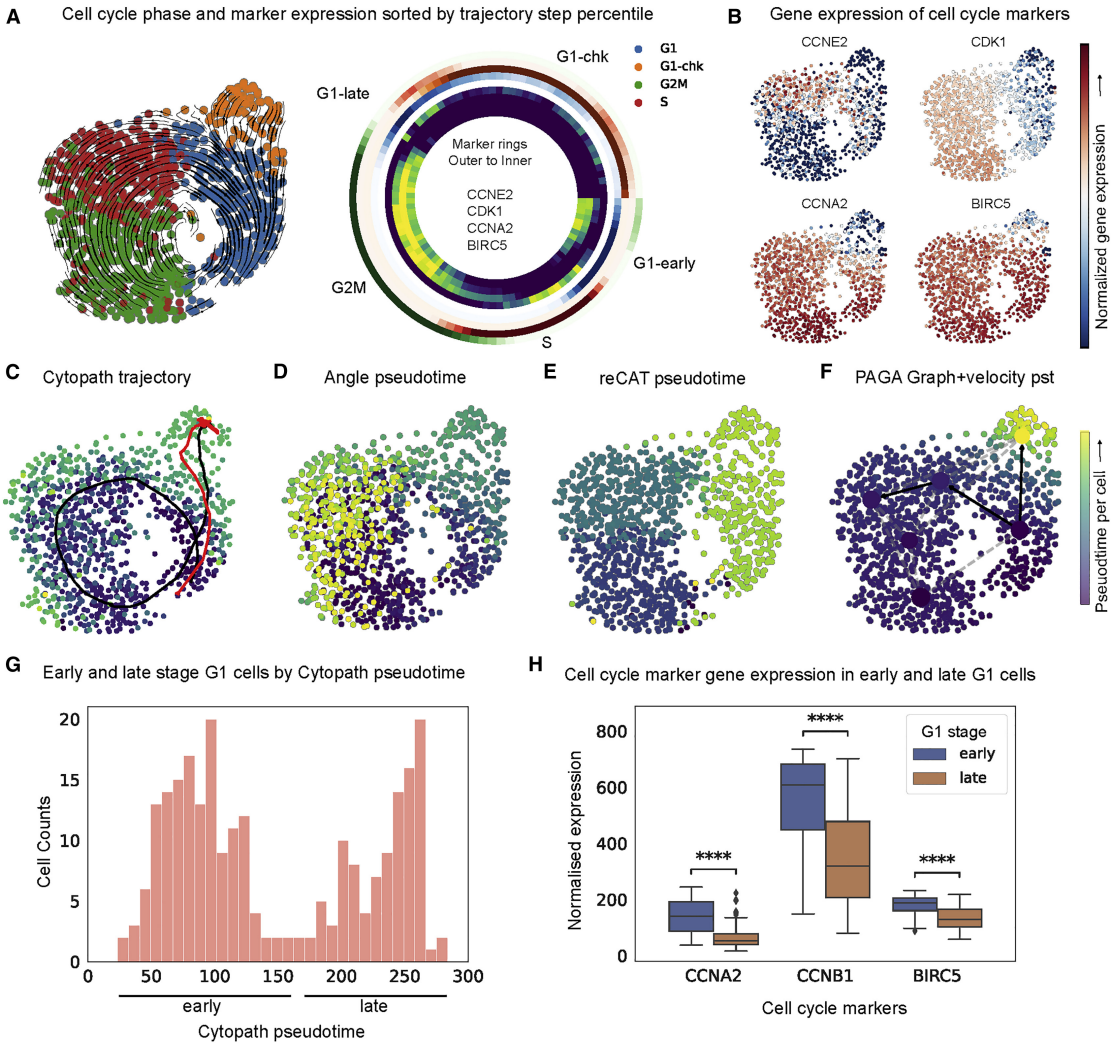


Figure 3.2: **Reconstruction of cell cycle in U2OS cell line.** (A) RNA velocity stream plot overlayed on the UMAP projection, annotated with the cell cycle phase adapted from Mahdessian et al. [24] Considering all cell-to-trajectory alignments binned into percentiles, the radial heatmap shows cell cycle phase fraction (outer set of rings) and marker expression (inner set of rings) sorted by trajectory step. The directionality of the radial heatmap is clockwise, with the origin at zero degrees (B) The separation of the G1 phase into G1 and G1-chk was performed based on marker expression of cell cycle genes. (C–F) Trajectories inferred and pseudotime per cell by (C) Cytopath, (D) Angle, (E) ReCAT, (F) PAGA and velocity pseudotime (vpt). (G) Distribution of Cytopath pseudotime for cells in the G1 cluster. (H) Normalized expression of cells classified as early and late G1 cells (blue/orange, respectively). Significance was estimated by an independent t-test for each marker.

Figure reproduced from figure 3 of manuscript 1

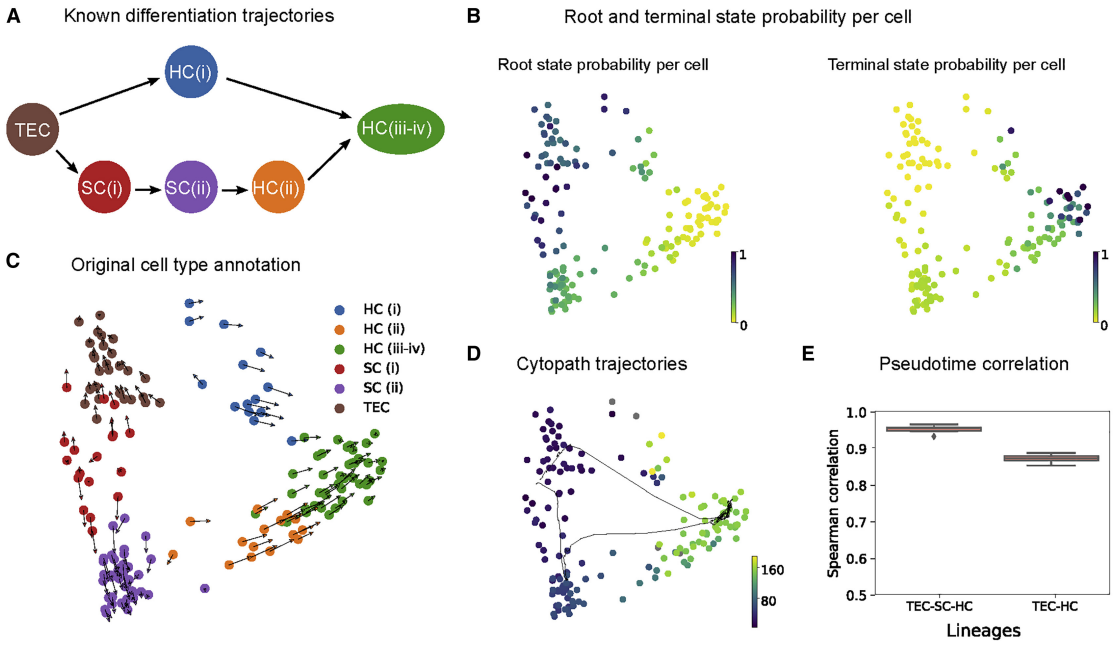


Figure 3.3: **Reconstruction of convergent differentiation in developing neonatal mouse inner ear.** (A) Known differentiation trajectories from [25]. (B) Probability estimated based on RNA velocity of a cell being a root and terminal states, respectively. (C) RNA velocity overlayed on the PCA projection of neonatal mouse inner ear data annotated with stages of differentiation. (D) Inferred trajectories and mean pseudotime by Cytopath. (E) Spearman correlations between known lineage ordering of cell types and pseudotime inferred by Cytopath (10 runs).

Figure reproduced from figure 5 of manuscript 1

Reconstruction of convergent differentiation in developing neonatal mouse inner ear

The development of hair cells (HCs) in the sensory epithelium of the utricle originates from transitional epithelial cells (TECs) via support cells (SC). A secondary differentiation path from TECs to HCs and a transitional zone where cells can easily switch fate results in two convergent differentiation trajectories [25] [Figure 3.3A].

Root and terminal state probability estimation using RNA velocity was used to select root and endpoints. A PCA projection of the data was generated, as indicated in the original study. Trajectory inference with Cytopath reproduced the two differentiation trajectories in a completely data-driven manner. The correlation between known cell type ordering and pseudotime estimated by Cytopath is robust for either lineage [Figure 3.3D-E].

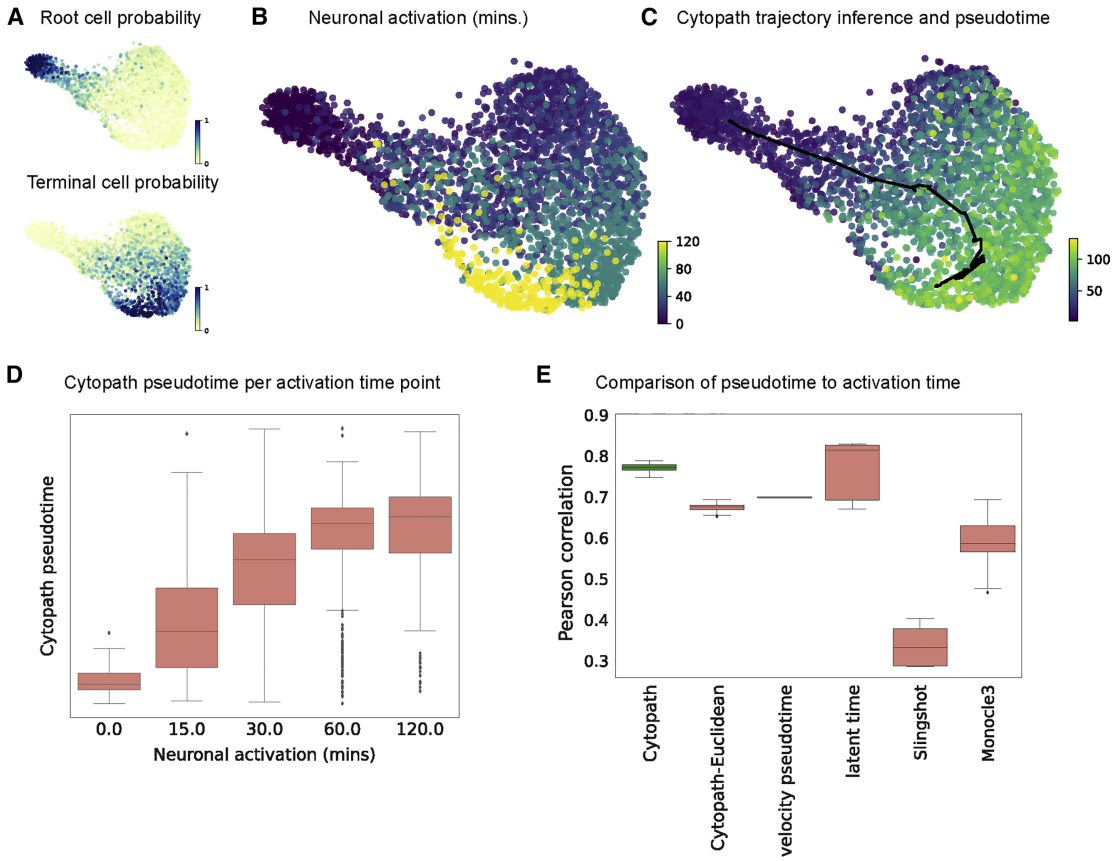


Figure 3.4: **Reconstruction of convergent differentiation in developing neonatal mouse inner ear.** (A) Root and terminal state probability inferred using RNA velocity. (B) UMAP projection annotated with the duration of stimulation for each cell. (C) UMAP projection annotated with trajectory and pseudotime inferred by Cytopath. (D) Cytopath pseudotime per cell with respect to stimulation duration. Note the monotonic relationship between median pseudotime and stimulation duration. (E) Pearson correlation between pseudotime inferred by Cytopath, non-velocity-based pseudotime estimated using Cytopath trajectory inference (Cytopath-Euclidean) and baseline methods.

Figure reproduced from figure 6 of manuscript 1

3.1.3 Approximating the real rate of differentiation

RNA velocity is an estimate of the rate of change of expression. Therefore, pseudotime estimated based on RNA velocity can order cell-states along a coordinate that represents progress along the differentiation process, unlike prior methods that only measure the difference in expression values and do not account for the difference in the rate of change of expression.

Pseudotime inference was performed for scRNAseq. data from mouse cortical neurons stimulated for various durations within the range of 0-120 minutes.

Asynchronous activation of differentiation typically precludes the use of time-series labels as a proxy for differentiation progress in individual cells. Therefore, only gene expression of activity-regulated genes was considered for the analysis, and thus the duration of stimulation would indicate differentiation progress [22].

Pseudotime inferred by Cytopath had a monotonic relationship with stimulation time [Figure 3.4D]. Comparative analysis with pseudotime estimated using both velocity and non-velocity-based methods revealed higher Pearson (linear) correlations between pseudotime and stimulation time for velocity-based methods [Figure 3.4E].

The impact of RNA velocity-based cell-to-trajectory assignment with Cytopath was assessed by computing a pseudotime (Cytopath-Euclidean pseudotime) using non-velocity Euclidean distance-based cell assignment to the trajectory inferred by Cytopath. Cytopath-Euclidean pseudotime had a lower correlation with stimulation time but outperformed non-velocity-based methods.

3.1.4 Fate trajectories of CD8+ T cells in chronic LCMV infection

CD8+ T cells are cytotoxic cells essential for the clearance of viral infections. Under conditions of chronic infection, these cells undergo a process of *exhaustion* during which they lose the ability to clear the viral load but persist in a senescent-like state. This state is characterised by reduced effector function and the expression of co-inhibitory receptors due to persistent T-cell receptor (TCR) stimulation. Manuscript 2 reports a study on CD8 T cell differentiation towards either terminally-exhausted or memory-like-exhausted CD8+ T cell subsets. Trajectory inference with Cytopath [Manuscript 1] suggested an early commitment of CD8+ T cells towards either the terminally-exhausted or the memory-like-exhausted CD8+ T cell fate.

Comprehensive multi-sample scRNAseq. time-series data captured from the induction of infection to terminal exhaustion in CD8+ T cells was produced to investigate lineage relationships between the cell-types that arise during the process. scRNAseq. data was produced post-infection corresponding to early phases (day 1-4), peak phase (day 7), contraction phase (day 14) and late phase (day 21) [Figure 3.5A]. Root and terminal cell states were inferred data-driven using RNA velocity. Subsequently, trajectory inference was performed using Cytopath [Manuscript 1] to model the terminally-exhausted and memory-like lineages.

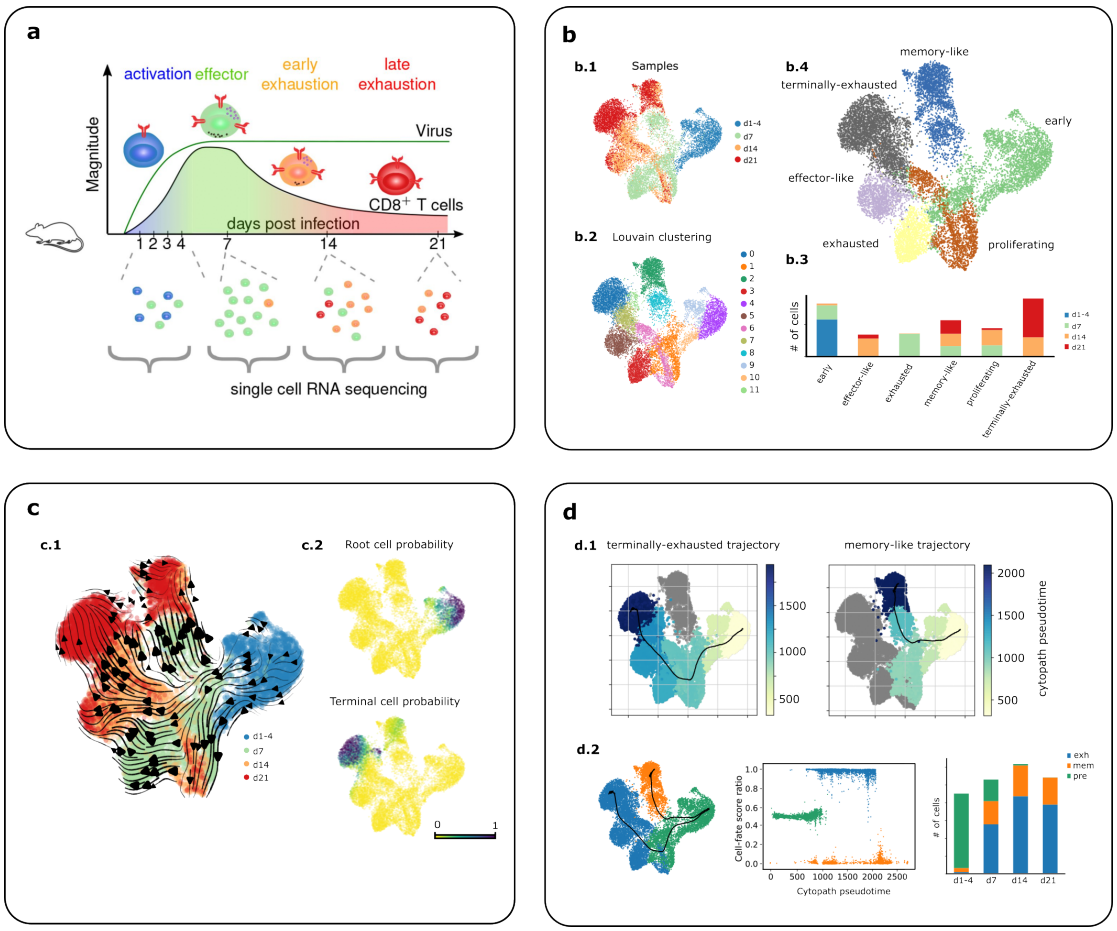


Figure 3.5: **Overview of CD8+ T cell trajectory inference with Cytopath.** (A) Transgenic P14 CD8 T cells were sampled in a time-series protocol under chronic infection with LCMV. The samples were acquired from four phases of the infection activation (day 1-4), effector (day 7), early exhaustion (day 14) and late exhaustion (d21), and scRNAseq was performed using the 10x Genomics platform. Figure reproduced from figure 1 of manuscript 2. (B) UMAP projection of scRNAseq. data. (b.1) Color indicates the time-point after infection, at which cells were isolated for scRNAseq. (b.2) Louvain cluster assignment (b.3) Cell composition of the phenotypic clusters by sample time-point. (b.4) phenotypic cluster annotation based on marker and differentially expressed genes. Figure adapted from figure 2 of manuscript 2. (C) RNA velocity analysis of scRNAseq. data. (c.1) Stream plot visualising transitions between cells inferred from RNA velocity (c.2) The stationary distribution of the backward and the forward transition matrix, respectively, indicate root and terminal cell-state. Figure adapted from figure 4 of manuscript 2. (D) Trajectory inference with Cytopath. (d.1) Trajectory path and pseudotime for the terminally-exhausted and memory-like cell-types projected on UMAP. (d.2) Cell-fate commitment of cells towards either lineage. Figure adapted from figures 5 and figure 6 of manuscript 2.

Inference of CD8+ T cell exhaustion lineages

Cell states were assigned cell-type labels based on the over-expression of canonical marker genes [Figure 3.5B.4]. Root cell probability was high in a subset of early-phase cells, and terminal state probability was high in cells assigned to the terminally-exhausted and memory-like-exhausted groupings [Figure 3.5C.2]. Trajectory inference for either terminal cell-type was performed. Early phase cells were assigned to both trajectories, and subsequently, the two trajectories diverged, with cells being assigned exclusively to one of the two trajectories [Figure 3.5D].

Several marker genes were found to have differential expression patterns along the two trajectories. *Slamf6*, *Ccr6*, *Tnfsf8*, *Xcl1* and *Cxcl10* were expressed at higher levels in the memory-like trajectory. *Slamf6*, *Ccr6*, *Tnfsf8* gradually increased in expression along the trajectory. A pattern of increasing expression of *Cxcr6*, *Ccl5* and *Nkg7* was observed in the terminally-exhausted trajectory. *Ifngr1* and *Lgals3* were transiently up-regulated in the exhausted trajectory exclusively but subsequently decreased in expression.

Early phase cell-fate commitment in CD8+ T cell exhaustion

Cells were probabilistically assigned to lineages by computing an assignment score per cell per lineage. The lineages were considered to have branched when cells were only assigned to a single lineage. In the early phase, the cell-states appeared to have equal assignment scores on average but rapidly branch between steps 800-1200 [Figure 3.5D.2]. Most cell-states from the sample corresponding to day 7 were already exclusively committed to either lineage. Cell-states with the lowest pseudotime appeared to be uncommitted. This suggested that the commitment may occur between days 5-6 after infection.

Cell-states in the early phase could be separated into three categories based on cell-fate scoring with phenotypes indicative of a pre-committed state, a committed state towards the terminally-exhausted endpoint or the memory-like endpoint. A linear classifier identified genes predictive of these categories. CXCR6 and TCF1 were chosen as the candidates for sorting the branches into pre-committed (CXCR6- TCF1-), memory-like (CXCR6- TCF1+) and exhausted (CXCR6+ TCF1-) cells.

Validation of cell-fate commitment via adoptive transfer experiments

Cells committed towards the terminally-exhausted lineage, memory-like lineage and uncommitted cells were isolated based on CXCR6 and TCF at day 5 post-LCMV infection in a follow-up experiment. The isolated cells were transferred to infection-matched hosts. At day 12 from the start of the experiment, the CD8+

T cell population arising from the transferred cells was extracted [Figure 3.6A-B]. Cells predicted to be committed to the terminally-exhausted lineage gave rise to terminally-exhausted cells, and cells predicted to be uncommitted gave rise to both terminally-exhausted and memory-like cells. Cells predicted to be committed towards the memory-like lineage gave rise to both terminally-exhausted and memory-like-exhausted cells [Figure 3.6C].

Memory-like exhausted CD8+ T cells retain the capacity for proliferation and recall response [26]. In a separate set of adoptive transfer experiments, the three categories of cells committed towards the terminally-exhausted lineage, memory-like lineage and uncommitted cells were transferred into infection-matched hosts infected with an escape mutant of the LCMV virus. The mutant virus did not activate the TCR receptors of the transferred cells. In the absence of TCR stimulation, the cells pre-disposed towards the memory-like lineage only give rise to memory-like cells at day 12, suggesting that in the absence of further stimulation, the bifurcation model of differentiation is correct [Figure 3.6D].

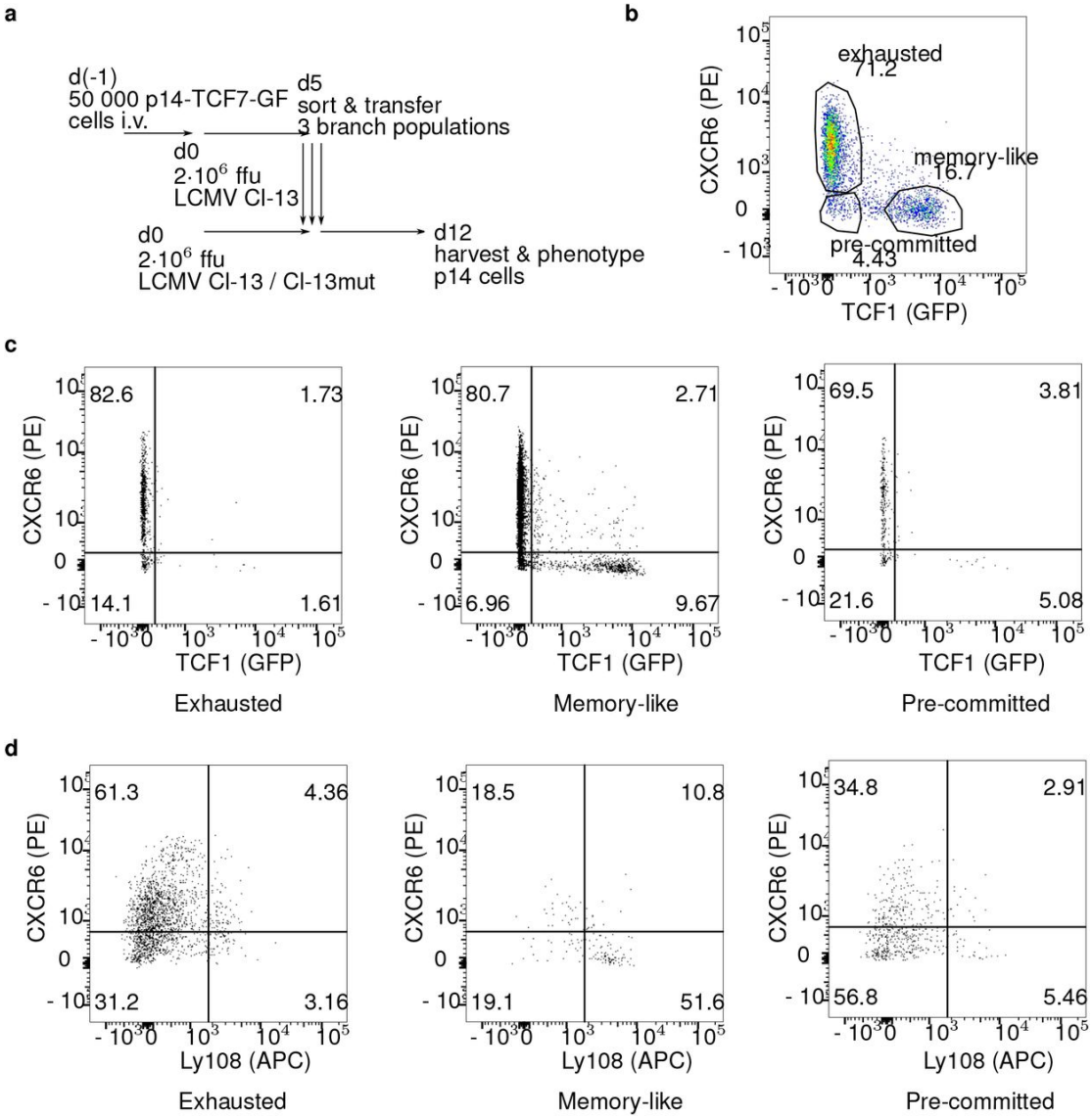


Figure 3.6: **Adoptive transfer of fate-biased cell populations into infection-matched hosts.** (A) The three P14 lineage populations were isolated at 5-day post-infection (dpi) from high-dose Clone-13 infected mice (that had been transferred with naive P14 cells prior to infection) and transferred into infection-matched hosts, and their phenotype was assessed 7 days post-transfer. (B) Flow cytometry gates used to sort cells committed towards terminally-exhausted lineage, memory-like lineage and uncommitted cells. (C) Phenotype of the recovered P14 cells at 12 dpi from spleens after high-dose Clone-13 infection and transfer of either exhausted, memory-like or pre-committed cell populations isolated at 5 dpi. Cells are gated on P14 cells. (D) Phenotype of the recovered cells at 12 dpi from spleens after transfer into hosts infected with Clone-13 P14 escape mutant. Naïve P14 cells were first transferred into naive C57BL6 mice, followed by Clone-13 infection. At 5 dpi, terminally-exhausted, memory-like and uncommitted populations were sorted and adoptively transferred into infection-matched hosts with Clone-13 escape mutant. Recovered P14 cells are shown.

Figure reproduced from figure 8 of manuscript 2.

3.2 Latent state-space process model of differentiation

scRNAseq. data measures individual cell-states, but due to the high amount of variation as well as the phenomenon of dropout, functional interpretation of cell-states is difficult. Therefore, the clustering of cell-states is an essential step in interpreting scRNAseq. data from a biological functions perspective. The identification of clusters of cell-states as cell-types forms the basis for the interpretation of associative analysis such as differential expression testing. Differentiation processes are often described in the literature as a hierarchy of transitions between functional cell-types. However, the increase in interpretability has an associated decrease in resolution.

Clustering implicitly assumes that variation within a cluster does not reflect changes in the functional identity assigned to the cluster. This assumption is invalid for scRNAseq. data of differentiation processes, where sustained and gradual changes in gene expression lead to transitions between functional identities. Simplified, *coarse-grained* models of cell-type transitions are restricted to modelling the likelihood of transition between cell-types and do not model the dynamics of these transitions [12]. Manuscript 3 reports a factorial latent state-space model that infers dynamics in a smaller latent space vis-a-vis the number of measured cell-states. The model is used to perform kinetic clustering of cell-states and identify transitional cell-states.

3.2.1 Kinetic clustering and lineage inference

Model input

For the set of measured cell-states (observed states) $O = \{o^{(1)}, \ldots, o^{(n)}\}$, an initial probability vector $Y_0 = \{P(o^{(1)} \mid i = 0), \ldots, P(o^{(n)} \mid i = 0)\}$, is estimated using the process for obtaining root state probabilities (aforementioned).

The transition probability matrix $\mathbf{T}$ over observed states is used to simulate the differentiation process as a sequence of probability vectors $\mathbf{Y}$. The simulation is performed as,

$$Y_i = Y_{i-1} \cdot \mathbf{T} = Y_0 \cdot \mathbf{T}^i \tag{3.4}$$

The simulation is considered to have converged if $Y_i = Y_{i-1}$, i.e. when the simulation reaches the stationary state of the Markov chain with the transition matrix $\mathbf{T}$. The latent dynamic model considers the simulated process,

$$\mathbf{Y} = \{Y_i\} = \{\{P(o^{(1)} \mid i), \ldots, P(o^{(n)} \mid i)\}\} \qquad \forall i = 0, 1, \ldots, I \tag{3.5}$$

as input. In the following text, P_o is used to indicate a probability vector over states O such as $Y_i = P_o(o \mid i)$. It is assumed that the observed states are emissions from a lower dimensional latent state-space, and the dynamics in the observed state-space are caused due to dynamics in the latent space.

Model specification

With latent states $S = \{s^{(1)}, \ldots, s^{(m)}\}$ and analogous to the simulation over observed states, we describe the dynamics over latent states as

$$\mathbf{Q} = \{Q_i\} = \{\{P(s^{(1)} \mid i), \ldots, P(s^{(m)} \mid i)\}\} \qquad \forall i = 0, 1, \ldots, I \tag{3.6}$$

Let $\mathbf{H}$ be the transition probability matrix over latent states S, then corresponding to the simulation (Eq. (3.4)), a Markov chain in the latent space has the form,

$$Q_i = Q_{i-1} \cdot \mathbf{H} = Q_0 \cdot \mathbf{H}^i \tag{3.7}$$

With the assumption of constant emission probabilities of observed states over the latent process $P(o \mid s, i) = P(o \mid s)$, we express Y_i as

$$Y_i = \sum_{s \in S} P_o(o, s \mid i) = \sum_{s \in S} P_o(o \mid s) P(s \mid i) \tag{3.8}$$

and due to Eq. (3.7):

$$Y_i = \sum_{s \in S} P_o(o \mid s) \cdot Q_i = \sum_{s \in S} P_o(o \mid s) \cdot (Q_0 \cdot \mathbf{H}^i) \tag{3.9}$$

Lineages L are modelled as independent Markov chains in the latent space. Furthermore, restricting the lineages to a common latent state-space $P(o \mid s, l) = P(o \mid s)$,

$$Y_i = \sum_{l \in L} P(l) \sum_{s \in S} P_o(o \mid s) \cdot (Q_0^{(l)} \cdot \mathbf{H}^{(\mathbf{l})}) \tag{3.10}$$

where $\mathbf{H}^{(\mathbf{l})}$ is the latent state transition probability matrix for lineage $l \in L$ and,

$$\mathbf{Q}^{(\mathbf{l})} = \{Q_i^{(l)}\} = \{\{P(s^{(1)} \mid i, l), \ldots, P(s^{(m)} \mid i, l)\}\} \qquad \forall i = 0, 1, \ldots, I \tag{3.11}$$

Model training

The trainable parameters of the model are the conditional latent state transition probability matrices $\mathbf{H}$,

$$\mathbf{H_{ij}^{(l)}} = P(s_i \mid s_j, l) \qquad \forall s \in S, l \in L \tag{3.12}$$

the emission probabilities $\mathbf{E}$,

$$E^{(s)} = P_o(o \mid s) \qquad \forall s \in S \tag{3.13}$$

the lineage weights W,

$$W = P(l) \qquad \forall l \in L \tag{3.14}$$

and the initial latent state probabilities $\mathbf{Q_0}$,

$$Q_0^{(l)} = P(s \mid i = 0, l) \qquad \forall s \in S, l \in L \tag{3.15}$$

Let $\hat{\mathbf{Y}}$ be the model estimate of $\mathbf{Y}$. The estimated sequence $\hat{\mathbf{Y}}$ is obtained as,

$$\hat{Y}_i = \sum_L \sum_S W Q_i^{(l)} \mathbf{E} \tag{3.16}$$

The parameters of the model are optimized by minimising the element-wise Kullback–Leibler (KL) divergence using gradient descent.

$$\mathrm{KL}(\hat{\mathbf{Y}}, \mathbf{Y}) = \sum_{i \in I} \sum_{o \in O} Y_i^{(o)} \log(\tfrac{Y_i^{(o)}}{\hat{Y}_i^{(o)}}) \tag{3.17}$$

Model output

The emission probabilities are a *soft* assignment of cells to latent states. The approach to clustering cell-states considering the transitional nature of cells during differentiation is termed kinetic clustering. The model enables the representation of intermediate cell-states and gradual divergence of branching events while simultaneously aiding the interpretation of differentiation processes at the level of functional cell-type identities. The condition probability of latent states with respect to observed states O is used to assign cluster memberships,

$$\arg\max_{s \in S} P(S = s \mid o) \tag{3.18}$$

The transition entropy of a cell is the sum of the entropy of the joint probability of a cell and each latent state.

$$-\sum_{s \in S} P(o, s) \cdot \ln(P(o, s)) \tag{3.19}$$

The model accommodates multiple concurrent differentiation processes by considering them as independent sequences of latent state transitions. Each lineage $l \in L$ is a sequence of transitions in a common state space S. The trajectories of lineages in latent state space are represented as sequences of most probable latent states at each step i,

$$\{\arg\max_{s \in S} P(S = s \mid l, i)\} \qquad \forall i = 1, \ldots, I \tag{3.20}$$

3.2.2 Transitional-cell identification in developing human forebrain

The developing human forebrain dataset consists of the glutamatergic neuronal lineage in human embryonic cells. The process follows a linear differentiation path from Radial-glia (progenitor) cells via a neuroblast (intermediate) population that is locked into the neuron (mature) fate [20]. The intermediate neuroblasts are highly motile cells that migrate to target brain regions before terminal differentiation [27] [28].

Root and terminal states inferred with RNA velocity correspond to the Radial-glia and mature neurons, respectively, as has been previously reported [Figure 3.7A.3-4] [20]. Kinetic clustering partitioned the data into two clusters [Figure 3.7B.1]. Static clusters computed using the Leiden algorithm overlap exclusively with one of the kinetic clusters except Leiden cluster 3, which appears to be split between the two kinetic clusters [Figure 3.7B.3].

Pseudotime estimated using RNA velocity has high variance in Leiden cluster 3, suggesting that this cluster may contain transitional cells [Figure 3.7C.1]. Transitional cells were identified as cells with high transitional entropy [Figure 3.7C.2-3]. Marker genes for neuroblasts were enriched in the set of genes positively correlated with transitional entropy [Figure 3.7D.1] [29].

Cells high in the expression of *EOMES* [20] and of *NHLH1* [30], canonical markers for neuroblast cells, are spread across multiple Leiden clusters. Cells expressing marker genes overlap with cells identified as transitional [Figure 3.7D.2]. This analysis concluded that transitional entropy is a useful criterion for selecting transitional cells.

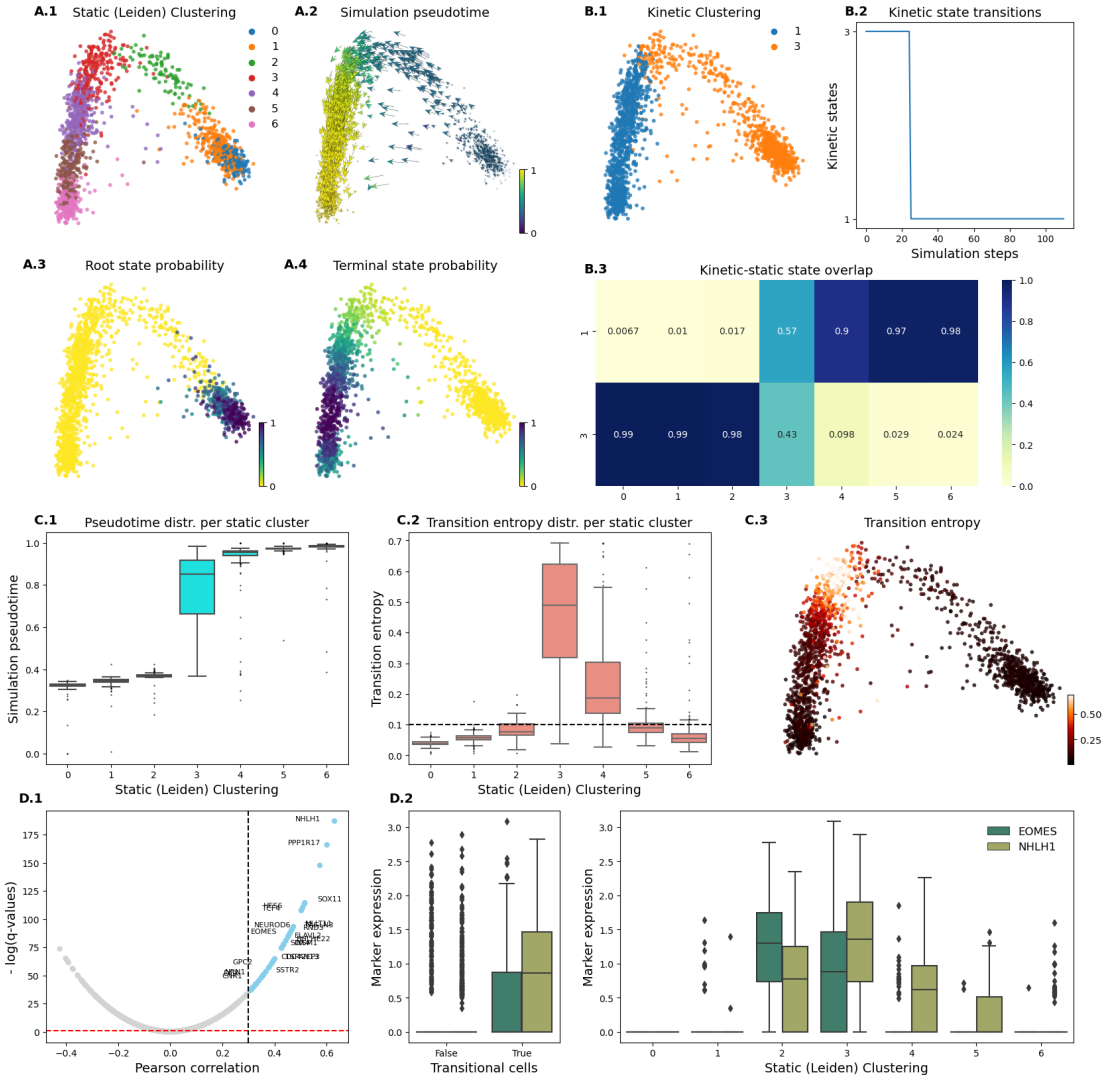


Figure 3.7: **Identifying transitional cells in developing human forebrain.** **(A)** Outputs of standard workflow scRNAseq. and RNA velocity analysis annotated on the first two principal components. (A.1.) Leiden clustering of cell states, (A.2.) RNA velocity vectors and estimated pseudotime. The pseudotime of a cell is calculated as the mean step weighted by the probability of observing a cell at each step. (A.3.) Root and (A.4.) Terminal cell states inferred using RNA velocity. **(B)** Outputs of our latent state space model. (B.1.) Kinetic clustering of cell states which is the most probable latent state per cell. (B.2.) Most probable latent state at each simulation step. (B.3.) Ratio of overlapping cells in each static (Leiden) and kinetic cluster. **(C)** Identification of transitional cell states. (C.1.) Pseudotime distribution of cells in each static cluster. (C.2.) Transition entropy of cells; computed as the entropy of the joint probability of a cell and each latent state; distribution over static clusters. The red line is the threshold to discriminate transitional cells from the rest. (C.3.) Transition entropy of cells. **(D)** Biological identification of transitional cells as neuroblasts. (D.1.) Pearson correlation between gene expression and transitional entropy of cells. (D.2.) Marker genes' (*EOMES, NHLH1*) expression distribution in transitional cells vs rest and for each Leiden cluster.

Figure reproduced from figure 1 of manuscript 3

3.3 Modelling scRNAseq. data with ordinal labels

scRNAseq. measurements performed in a time-series protocol can utilise the temporal labels as additional information for modelling differentiation. Ordinal labels may also be used to incorporate a prior hypothesis in the modelling of both single[1] and multi-sample dataset[2]. Manuscript 4 reports psupertime, a regression-based approach to infer pseudotime and identify genes relevant to the differentiation process using ordinal labels as supervision. An application of psupertime on scRNAseq. data is presented in manuscript 5

Unsupervised pseudotime inference methods are primarily designed to model cell-state and cell-type transitions without reference to the sample identity. The sample-to-sample variation in scRNAseq. measurements referred to as batch effects present challenges in clustering and identifying similar cell-states across batches. In this context, supervised analysis with ordinal labels enables the characterisation of differentiation processes in single-cell RNA-seq data with sequential labels, even in the presence of substantial unrelated variation.

psupertime is a supervised pseudotime approach based on regularised ordinal regression modelling. psupertime identifies a concise set of genes that exhibit coherent variations over the samples' sequential ordering. It assigns pseudotime values to individual cells based on a linear combination of these genes, approximating the specified order of the sequential labels. The model can also be used as a classifier for estimating labels in new data. Benchmarking psupertime against unsupervised pseudotime models suggests that psupertime performs better at identifying time-varying genes and at pseudotime estimation in time-series scRNAseq. protocols.

3.3.1 Regularised ordinal regression for pseudotime inference

psupertime is an L1 regularised proportional odds ordinal logistic regression model. Normalised, log-transformed and z-scored expression data is used to predict the ordinal labels as specified below. A concise set of genes characterising the differentiation process is identified by selecting genes with non-zero coefficients. The linear combination of gene expression produces a predicted position of the cell that may lie between ordinal labels and is interpreted as the pseudotime.

Ordinal logistic regression extends the concept of logistic regression to ordinal response variables. The model can be interpreted as simultaneous logistic

[1] For example, an ordering of cell-type labels may be used to incorporate prior knowledge of cell-type transitions in a single-sample scRNAseq. analysis.

[2] An example of non-temporal labels could be disease status. Samples collected from healthy and diseased individuals could be ordered based on disease severity.

regressions with the same coefficients for input features. Each regression predicts a different binary response variable derived from the ordinal labels. For the proportional odds formulation of ordinal logistic regression, the models estimate $\log(P(Y \geq j)/P(Y < j))$ where j is each of the ordinal categories. The proportional odds formulation will fit larger coefficients for features that vary across the entire range of ordinal labels. Alternatively, for the continuation ratio formulation, the models estimate $\log(P(Y = j)/P(Y < j))$ and thus, features that vary within a subset of the ordinal labels will be selected.

For input data $X \in R^{n \times p}$ and $y \in N^n$ ordinal labels, the ordinal logistic regression has the following cumulative distribution function,

$$P(y_i \leq j \mid X_i) = \phi(\theta_j - \beta^T X_i) = \frac{1}{1+(\beta^T X_i - \theta_j)} \tag{3.21}$$

X_i and y_i are the ith input and label respectively. j is the index of the ordinal categories. β are the coefficients and θ_j are the thresholds between labels. ϕ is the logit link function. The unregularised likelihood is,

$$L(\beta, \theta \mid y, X) = \prod_{i=1}^{N} (\phi(\theta_{y_i} - \beta^T X_i) - \phi(\theta_{y_{i-1}} - \beta^T X_i)) \tag{3.22}$$

With the L1 penalty, we obtain the optimal values of β and θ by minimizing the following regularised loss function,

$$\arg\min_{\beta\theta}(\lambda \sum_{k=1}^{p} |\beta_p| - \log(L(\beta, \theta \mid y, X))) \tag{3.23}$$

The regularisation strength λ is selected via cross-validation.

3.3.2 Performance benchmarks against unsupervised methods

Cell-level pseudotime orderings identified by psupertime were compared with those from three alternative, unsupervised pseudotime techniques [Figure 3.8]. As a simple and interpretable baseline, projection of scRNAseq. on the first PCA component was the first comparison. Monocle2 with start cell selection was the second comparison [31]. Slingshot, which permits the selection of both start and end cell states and, therefore can be considered a semi-supervised approach, was the third comparison [14]. The methods were chosen for being commonly used and high performing in a comparison of trajectory inference methods [18]. Tempora, a method that performs supervised analysis with temporal labels but produces predictions for cell-types rather than cell-states, was also compared [32].

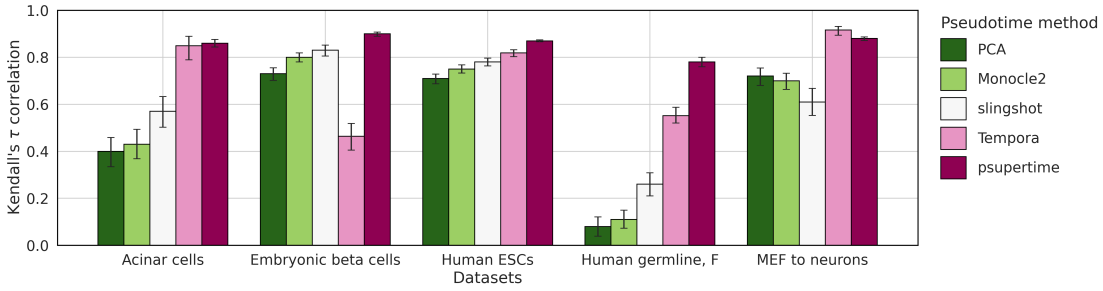


Figure 3.8: Absolute Kendall's τ correlation coefficient between label sequences (treated as sets of integers 1,...J) and calculated pseudotimes. Error bars show a 95% confidence interval over 1000 bootstraps. Figure adapted from figure 2E of manuscript 4.

3.3.3 Marker discovery in intestinal organoids

Manuscript 5 reports a droplet-based scRNAseq. protocol designed for efficient cell capture for samples with a low (<1000) number of cells. The technology was validated by performing scRNAseq. measurements of intestinal organoids. Organoids mimic adult intestinal mucosa on key geometric, architectural and cellular hallmarks and are generated from intestinal stem cells through a stochastic self-organisation process on 3D matrices. Even under identical in-vitro conditions, organoids exhibiting diverse morphologies may form.

Intestinal organoids were developed from cultured cells, differentiation was allowed to occur at day 2, and daily samples were collected from days 3-6 (S0-3) from the initiation of the protocol. The sampled organoids were selected in a biased manner to maximise diversity in size and morphology. psupertime analysis was performed to discover differentiation-associated genes in each organoid. Apart from known markers, the analysis discovered additional genes expressed in subsets of the organoids, such as Gastric inhibitory polypeptide (*Gip*), Zymogen granule protein 16 (*Zg16*), Vanin 1 (*Vnn1*) and Defensin alpha 24 (*Defa24*)

4. Discussion

The scientific understanding of biological systems can be expressed as conceptual models such as the central dogma of molecular biology or Waddington's landscape. Statistical analysis and simulation further require mathematical formalism to be applied to the conceptual model. In theory, each bio-molecular entity and its interactions can be parameterised; however, the complexity of biological interactions and, more importantly, the limitations of experimental technology make this an untenable proposition. Therefore, computational modelling of biological systems works with operational definitions, i.e. definitions of biological phenomena based on what can be measured. For example, the effect of a gene's expression on another gene depends on several intermediate biological mechanisms, but these relationships are considered direct effects in modelling with scRNAseq. data. Since operational limitations make the development of comprehensive, *universal* models infeasible, the development of differentiation models suited to particular research goals is required.

The statistical characteristics of experimental data inform the choice in the class of models and assumptions regarding the distribution of random variables. Prior biological knowledge and contextual information must also be incorporated into the model. The information can be used to constrain the structure of the model in terms of the relationship between variables or as constraints on optimisation. For example, scRNAseq. data is sparse; therefore, zero-inflated distributions are typically used to model gene expression. Neighbourhood graphs between cell-states are constructed based on top principal components since genes' expression are highly correlated. Depending on the pre-processing, the distributions used can be discrete or continuous, and each of these assumptions implicitly impacts the conceptualisation of the biological phenomenon under study. The models presented in this dissertation were developed to produce outputs that enable further investigation of biological mechanisms with scRNAseq. data derived from different experimental protocols or biological systems.

The concept of RNA velocity has had a profound impact on differentiation modelling. Prior methods estimated the ordering of cells with only gene expression profiles. However, the similarity of gene expression does not necessarily imply a common differentiation state. Cytopath [Manuscript 1] utilises directionality imparted by RNA velocity to enable the inference of cyclical differentiation patterns where cell-states entering or exiting stages of the cycle will have similar expression. Conversely, for convergent patterns, dissimilarity of expression does not

indicate divergence. Several conventional trajectory inference methods fit a Minimum Spanning Tree to model differentiation and are thus conceptually unsuitable for modelling cyclical and convergent patterns. However, since no topological constraints are placed on trajectory inference with Cytopath, convergent patterns can be modelled. Even for bifurcating differentiation processes, RNA velocity enables more sensitive detection of divergence by complementing differences in expression with divergence in transitions, an example being the detection of early heterogeneity in CD8+ T cell differentiation presented in Manuscript 2.

scRNAseq. allows the study of variation in gene expression between and within cell-types. Molecular phenotyping of cell-types has conventionally been based on the presence of molecular products corresponding to a few characteristic marker genes. In contrast, the feature space of scRNAseq. data is comprehensive with respect to genes and, therefore, can offer a more resolved view of cell-states. While comprehensive, scRNAseq. measurements suffer from dropout; therefore, individual genes' expression cannot be used to partition or compare cell-states. Clustering-based approaches have become standard in grouping cell-states and identifying cell clusters as functional cell-types. While the functional identity of cell clusters can be established relatively easily, there may be low frequency cell-states with distinct, divergent or transitional expression that are also functionally relevant—identifying and characterising these cell-states with scRNAseq. data continues to be challenging. Using a latent state-space model to model the transcriptional dynamics of cell-state, in manuscript 3, a joint approach to trajectory inference and clustering enabled the identification of transitional cell-states.

scRNAseq. measurements are not absolute quantification of gene expression and are biased with respect to genes. Large technical variation between experiments remains a challenge for comparative data analysis from multiple scRNAseq. libraries. Therefore, models of differentiation are generally restricted to the measured cell-state space, even if gene expression is parameterised using continuous and unbounded distributions. The purpose of lineage inference models is the construction of pseudo-temporal ordering of measured cell-states and quantifying relative association with co-occurring lineages.

While RNA velocity may appear to be an approach that could enable the estimation of intermediate cell-states based on a vector field approach, there are several outstanding challenges [21]. RNA velocity models operate with very simplified assumptions on the splicing dynamics. The models assume constant rates of expression, splicing and degradation. Furthermore, the parameters are estimated and scaled independently for each gene and therefore, the gene-wise velocity components are not on a common scale. The models are also not discrete and consider expression to be continuous. Thus, while RNA velocity enables local predictions

of a cell's future state, the predictions, in general, cannot be expected to be biologically plausible cell-states. And so, even RNA velocity-based trajectory inference remains restricted to the measured cell-state space.

Computational biology has been essential in producing testable experimental hypotheses from experimental data that is often noisy, biased and incomplete. Advances in experimental technologies are usually accompanied by model development to utilise the data effectively. Typically, computational modelling with scRNAseq. data has been used to discover relevant genes, define cell-types and perform pathway enrichment. The analysis is usually validated in subsequent experiments, and a conceptual mechanistic model is developed to explain the findings. The models presented in this dissertation work within this paradigm.

While biological interactions are complex, representation of biological processes could be possible with sufficient experimental data and appropriate models. Deep learning has emerged as a popular modelling approach in this context. For example, autoencoder-based factorisation models have been used to learn embeddings with latent nodes representing interacting gene modules such as regulatory programs or molecular pathways [33]. Supervised models have been used to map cell-states to functional outcomes such as drug response. Generative models have been used to sample cell-states corresponding to counterfactual scenarios [34]. In the context of differentiation, deep generative models have been used to sample intermediate cell-states [15].

Experimental advances in perturbation screening have enabled high throughput screens over thousands of gene perturbations in a single experiment, and this data could contribute to the development of generative models that can be used to sample cell-states corresponding to single and combination of perturbations. However, the higher fitting power of deep learning models typically results in lower interpretability; therefore, extracting knowledge from these models is challenging. Furthermore, so far, such models have struggled with generating cell-states corresponding to out-of-distribution scenarios suggesting that the models do not learn generalisable rules of biological organisation. A possible reason for the lack of generalisation could be the limited information in scRNAseq. measurements relative to the complexity of the biological system.

Apart from gene expression, single-cell measurements can be performed for other data modalities such as chromatin accessibility, surface protein markers or spatial orientation. Multi-modal single-cell measurements with various combinations are increasingly being applied instead of simple scRNAseq [35]. The simultaneous measurement of multiple bio-molecular species in a single cell as well as interventional data, may enable the inference of biological interactions underlying the measurements and thus reduce the need for abstraction.

Mechanistic models structurally encode prior biological knowledge and are an expression of a hypothesis on the working of a biological system. The parameter space of a mechanistic model is interpretable by design and more suited to the discovery of general rules of biological interactions than statistical associations or *black-box* deep learning models. I presented two approaches for modelling differentiation based on the simulation of the biological process with relatively simple parameterisations. For example, the parameters of the latent state-space dynamic model presented in manuscript 3 are not identifiable with scRNAseq. data, however, additional information on the chromatin accessibility of regulators may allow the latent states to be identified as regulatory regimes of the biological process.

More generally, in the context of differentiation processes, state-space modelling is a useful formalism to develop mechanistic models that can encode the hierarchical flow of information in biological systems. The conceptual model can flexibly be parameterised based on the characteristics of the data measured for each modality. Simulation of biological systems with mechanistic models can expand computational modelling from associative analyses to the inference of mechanisms. Advances in likelihood-free inference have increased the viability of such an approach to complex biological systems [36]. In combination with the increasing availability of high throughput perturbation screening and multi-modal data, new modelling approaches focused on in-silico experimentation and causal inference may become possible.

References

[1] D. Osumi-Sutherland, C. Xu, M. Keays, A. P. Levine, P. V. Kharchenko, A. Regev, E. Lein, and S. A. Teichmann, "Cell type ontologies of the human cell atlas," *Nature Cell Biology*, vol. 23, pp. 1129–1135, Nov. 2021.

[2] "What is your conceptual definition of "cell type" in the context of a mature organism?," *Cell Systems*, vol. 4, pp. 255–259, Mar. 2017.

[3] S. Iwanami and S. Iwami, "Quantitative immunology by data analysis using mathematical models," in *Encyclopedia of Bioinformatics and Computational Biology* (S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, eds.), pp. 984–992, Oxford: Academic Press, 2019.

[4] F. CRICK, "Central dogma of molecular biology," *Nature*, vol. 227, pp. 561–563, Aug. 1970.

[5] X. Dai and L. Shen, "Advances and trends in omics technology development," *Frontiers in Medicine*, vol. 9, July 2022.

[6] D. Jovic, X. Liang, H. Zeng, L. Lin, F. Xu, and Y. Luo, "Single-cell RNA sequencing technologies and applications: A brief overview," *Clinical and Translational Medicine*, vol. 12, Mar. 2022.

[7] C. Waddington, *The Strategy of the Genes*. Routledge, Apr. 2014.

[8] D. Friedmann-Morvinski and I. M. Verma, "Dedifferentiation and reprogramming: origins of cancer stem cells," *EMBO reports*, vol. 15, pp. 244–253, Feb. 2014.

[9] K. Takahashi and S. Yamanaka, "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors," *Cell*, vol. 126, pp. 663–676, Aug. 2006.

[10] M. A. Dimitriu, I. Lazar-Contes, M. Roszkowski, and I. M. Mansuy, "Single-cell multiomics techniques: From conception to applications," *Frontiers in Cell and Developmental Biology*, vol. 10, Mar. 2022.

[11] G. Iacono, R. Massoni-Badosa, and H. Heyn, "Single-cell transcriptomics unveils gene regulatory network plasticity," *Genome Biology*, vol. 20, June 2019.

[12] F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, and F. J. Theis, "PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells," *Genome Biology*, vol. 20, Mar. 2019.

[13] L. Haghverdi, F. Buettner, and F. J. Theis, "Diffusion maps for high-dimensional single-cell analysis of differentiation data," *Bioinformatics*, vol. 31, pp. 2989–2998, May 2015.

[14] K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit, "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics," *BMC Genomics*, vol. 19, June 2018.

[15] G. H. T. Yeo, S. D. Saksena, and D. K. Gifford, "Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions," *Nature Communications*, vol. 12, May 2021.

[16] J. Ding and A. Regev, "Deep generative model embedding of single-cell RNA-seq profiles on hyperspheres and hyperbolic spaces," *Nature Communications*, vol. 12, May 2021.

[17] A. Dixit, O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman, and A. Regev, "Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens," *Cell*, vol. 167, pp. 1853–1866.e17, Dec. 2016.

[18] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys, "A comparison of single-cell trajectory inference methods," *Nature Biotechnology*, vol. 37, pp. 547–554, Apr. 2019.

[19] "Interpreting Intronic and Antisense Reads in 10x Genomics Single Cell Gene Expression Data," Aug. 2021.

[20] G. L. Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, and P. V. Kharchenko, "RNA velocity of single cells," *Nature*, vol. 560, pp. 494–498, aug 2018.

[21] V. Bergen, R. A. Soldatov, P. V. Kharchenko, and F. J. Theis, "RNA velocity—current challenges and future perspectives," *Molecular Systems Biology*, vol. 17, Aug. 2021.

[22] Q. Qiu, P. Hu, X. Qiu, K. W. Govek, P. G. Cámara, and H. Wu, "Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq," *Nature Methods*, vol. 17, pp. 991–1001, Aug. 2020.

[23] S. Li, P. Zhang, W. Chen, L. Ye, K. W. Brannan, N.-T. Le, J. ichi Abe, J. P. Cooke, and G. Wang, "A relay velocity model infers cell-dependent RNA velocity," *Nature Biotechnology*, Apr. 2023.

[24] D. Mahdessian, A. J. Cesnik, C. Gnann, F. Danielsson, L. Stenström, M. Arif, C. Zhang, T. Le, F. Johansson, R. Schutten, A. Bäckström, U. Axelsson, P. Thul, N. H. Cho, O. Carja, M. Uhlén, A. Mardinoglu, C. Stadler, C. Lindskog, B. Ayoglu, M. D. Leonetti, F. Pontén, D. P. Sullivan, and E. Lundberg, "Spatiotemporal dissection of the cell cycle with single-cell proteogenomics," *Nature*, vol. 590, pp. 649–654, Feb. 2021.

[25] J. C. Burns, M. C. Kelly, M. Hoa, R. J. Morell, and M. W. Kelley, "Single-cell RNA-seq resolves cellular complexity in sensory organs from the neonatal inner ear," *Nature Communications*, vol. 6, Oct. 2015.

[26] N. Budimir, G. D. Thomas, J. S. Dolina, and S. Salek-Ardakani, "Reversing t-cell exhaustion in cancer: Lessons learned from PD-1/PD-l1 immune checkpoint blockade," *Cancer Immunology Research*, vol. 10, pp. 146–153, Feb. 2022.

[27] R. D. Hodge, R. J. Kahoud, and R. F. Hevner, "Transcriptional control of glutamatergic differentiation during adult neurogenesis," *Cellular and Molecular Life Sciences*, vol. 69, pp. 2125–2134, Jan. 2012.

[28] C. Bressan and A. Saghatelyan, "Intrinsic mechanisms regulating neuronal migration in the postnatal brain," *Frontiers in Cellular Neuroscience*, vol. 14, Jan. 2021.

[29] G. L. Manno, D. Gyllborg, S. Codeluppi, K. Nishimura, C. Salto, A. Zeisel, L. E. Borm, S. R. Stott, E. M. Toledo, J. C. Villaescusa, P. Lönnerberg, J. Ryge, R. A. Barker, E. Arenas, and S. Linnarsson, "Molecular diversity of midbrain development in mouse, human, and stem cells," *Cell*, vol. 167, pp. 566–580.e19, Oct. 2016.

[30] E. Braun, M. Danan-Gotthold, L. E. Borm, E. Vinsland, K. W. Lee, P. Lönnerberg, L. Hu, X. Li, X. He, Ž. Andrusivová, J. Lundeberg, E. Arenas, R. A. Barker, E. Sundström, and S. Linnarsson, "Comprehensive cell atlas of the first-trimester developing human brain," Oct. 2022.

[31] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, and C. Trapnell, "Reversed graph embedding resolves complex single-cell trajectories," *Nature Methods*, vol. 14, pp. 979–982, Aug. 2017.

[32] T. N. Tran and G. D. Bader, "Tempora: Cell trajectory inference using time-series single-cell RNA sequencing data," *PLOS Computational Biology*, vol. 16, p. e1008205, Sept. 2020.

[33] M. Lotfollahi, S. Rybakov, K. Hrovatin, S. Hediyeh-zadeh, C. Talavera-López, A. V. Misharin, and F. J. Theis, "Biologically informed deep learning to query gene programs in single-cell atlases," *Nature Cell Biology*, Feb. 2023.

[34] M. Lotfollahi, A. K. Susmelj, C. D. Donno, Y. Ji, I. L. Ibarra, F. A. Wolf, N. Yakubova, F. J. Theis, and D. Lopez-Paz, "Compositional perturbation autoencoder for single-cell response modeling," Apr. 2021.

[35] A. Baysoy, Z. Bai, R. Satija, and R. Fan, "The technological landscape and applications of single-cell multi-omics," *Nature Reviews Molecular Cell Biology*, June 2023.

[36] K. Cranmer, J. Brehmer, and G. Louppe, "The frontier of simulation-based inference," *Proceedings of the National Academy of Sciences*, vol. 117, pp. 30055–30062, May 2020.

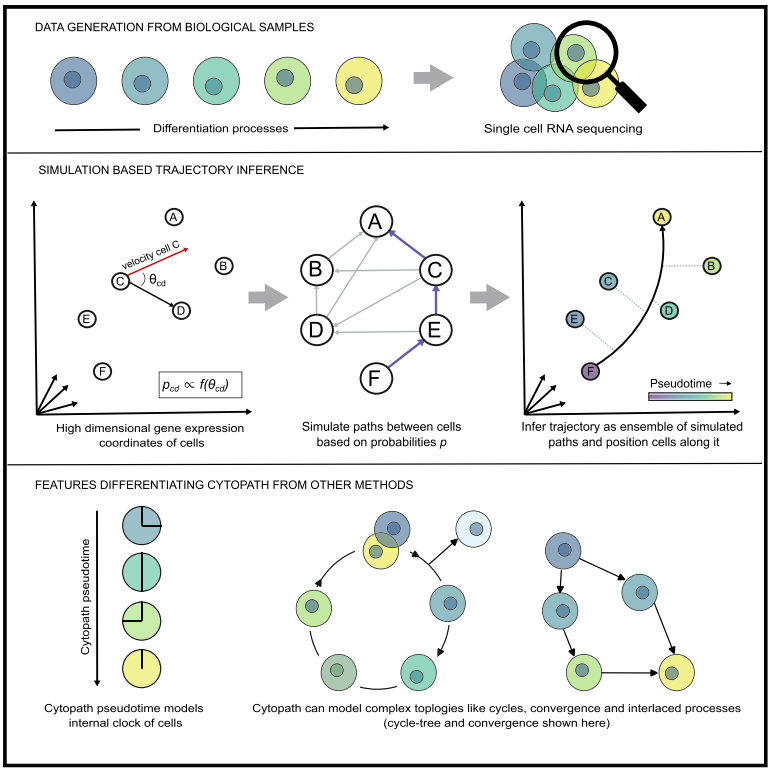
Manuscripts

Cell Reports
Methods

Article

Simulation-based inference of differentiation trajectories from RNA velocity fields

Graphical abstract



Highlights

- Trajectory inference leveraging RNA velocity with no topological constraints
- Cytopath models complex behavior like cycles, convergences, and interlaced processes
- Pseudotime ordering inferred by Cytopath approximates rate of change

Authors

Revant Gupta, Dario Cerletti, Gilles Gut, Annette Oxenius, Manfred Claassen

Correspondence

manfred.claassen@med.uni-tuebingen.de

In brief

Gupta et al. develop a method to model differentiation processes with single-cell RNA sequencing data that has the capacity to model complex behaviors like cycling and convergence as well as co-occurring combinations of multiple processes.

Gupta et al., 2022, Cell Reports Methods *2*, 100359
December 19, 2022
https://doi.org/10.1016/j.crmeth.2022.100359

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

Simulation-based inference of differentiation trajectories from RNA velocity fields

Revant Gupta,[1,2] Dario Cerletti,[3,4] Gilles Gut,[3] Annette Oxenius,[4] and Manfred Claassen[1,2,5,*]
[1]Internal Medicine I, University Hospital Tübingen, Faculty of Medicine, University of Tübingen, Tübingen, Germany
[2]Department of Computer Science, University of Tübingen, Tübingen, Germany
[3]Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland
[4]Institute of Microbiology, ETH Zurich, Zurich, Switzerland
[5]Lead contact
*Correspondence: manfred.claassen@med.uni-tuebingen.de
https://doi.org/10.1016/j.crmeth.2022.100359

MOTIVATION Trajectory inference from single-cell RNA sequencing data has the potential to systematically reconstruct complex differentiation processes, but inferring trajectories that accurately model the biological characteristics of varied processes continues to be a challenge, regardless of the many available solutions. In general, trajectory and pseudotime inference methods have so far suffered from the ambiguity of static single-cell transcriptome snapshots lacking a concept of directionality and rate of transcriptional activity.

SUMMARY

We report Cytopath, a method for trajectory inference that takes advantage of transcriptional activity information from the RNA velocity of single cells to perform trajectory inference. Cytopath performs this task by defining a Markov chain model, simulating an ensemble of possible differentiation trajectories, and constructing a consensus trajectory. We show that Cytopath can recapitulate the topological and molecular characteristics of the differentiation process under study. In our analysis, we include differentiation trajectories with varying bifurcated, circular, convergent, and mixed topologies studied in single-snapshot as well as time-series single-cell RNA sequencing experiments. We demonstrate the capability to reconstruct differentiation trajectories, assess the association of RNA velocity-based pseudotime with actually elapsed process time, and identify drawbacks in current state-of-the art trajectory inference approaches.

INTRODUCTION

Biological processes such as cell type differentiation,[1–3] immune response,[4] or cell division[5] can be conceptualized as temporal sequences of coordinated, phenotypic state changes in the context of possibly heterogeneous cell populations. Such phenotypic states can be characterized by, e.g., epigenetic, transcriptional, and proteomic cell profiles. These differentiation processes are often triggered asynchronously. The differentiation processes give rise to state sequences with varying topologies, including bifurcating, multi-furcating, cyclical, and convergent trajectories.

This situation requires single-cell approaches to measure and ultimately investigate these processes. The repertoire of suitable technologies to monitor different types of molecular profiles has increased dramatically over the last years. In particular, single-cell RNA sequencing (scRNA-seq) has gained widespread use because of the broad applicability of sequencing technology. Although these measurements are information rich, their analysis and interpretation are challenged by high dimensionality, low sequencing depth, measurement noise, and its destructive nature, only yielding snapshots of the whole process.

Different computational approaches have been proposed to model differentiation processes from scRNA-seq data, specifically covering the tasks of pseudotime estimation, trajectory inference, or cell fate prediction. These tasks are related but typically require different approaches (Table S1). The goal of cell fate prediction is to determine the terminal differentiation state (fate) of any cell, possibly already early in the differentiation process. Such methods generate a score or probability per cell with respect to terminal differentiation states.[6,7]

Pseudotime estimation addresses the task of ordering observed cells into a sequence of cell states traversed by a differentiation process. Typically, the estimated pseudotime values

are interpreted as temporal ordering, not capturing the pace of differentiation. It has been suggested that RNA velocity-based pseudotime has the potential to overcome this limitation.[8] Although pseudotime estimation might constitute sufficient characterization of a linear differentiation process, the description of complex processes with more involved topologies, such as bifurcations, requires an additional step of trajectory inference. Trajectory inference methods seek to infer a representative sequence of states that characterizes the possibly multiple differentiation processes in branching or convergent differentiation.[9–12]

Typical trajectory inference methods are guided by the assumption that phenotypic similarity reflects temporal proximity. However, static expression profiles are ambiguous with respect to the directionality of potential cell state transitions. This ambiguity is a major limitation of pseudotime ordering and trajectory inference and specifically precludes data-driven assignment of root and terminal states without previous knowledge about the process as well as resolving complex (i.e., cyclical[5] or convergent[3]) process topologies. It has now become possible to estimate transcriptional activity from scRNA-seq data via RNA velocity analysis,[1] enabling inference of likely transitions between different cell states in a data-driven fashion, ultimately opening the possibility to mitigate the limitations of the aforementioned reconstruction approaches.

In this work we present Cytopath, a simulation-based trajectory inference approach that takes advantage of RNA velocity. We demonstrate that Cytopath infers accurate and robust cell state trajectories of known differentiation processes with linear, circular, bifurcated, tree-like, and convergent topologies from scRNA-seq datasets. We show that Cytopath has the potential to model interlaced processes with different topologies as well as detect regions of transcriptional program switching. We also assess the ability of pseudotime estimated by Cytopath to represent the biological process time, also referred to as the "internal clock" of a cell. Trajectory inference with Cytopath addresses the limitations of state-of-the-art trajectory inference approaches as well as recently developed RNA velocity-based methods.[11,12]

RESULTS

Here, we present an overview of Cytopath and its trajectory inference performance, assessed on six scRNA-seq datasets consisting of cellular differentiation processes with various topologies.[1–5,13] We compare the performance of Cytopath with the best trajectory inference models for each topology:[14] Slingshot[9] for tree-like topology, Angle[14] for cell cycle, and partition-based graph abstraction (PAGA; directionality enabled by velocity pseudotime)[8,15] for graph models. We also include a comparison with Monocle3[16] as well as two approaches accounting for RNA velocity information: VeTra[11] and Cellpath.[12]

Simulation-based trajectory inference with Cytopath

Trajectory inference with Cytopath is performed downstream of the RNA velocity analysis of an scRNA-seq dataset and is specifically based on the resulting cell-to-cell transition probability matrix. The transition probability matrix considers each cell to be a node in a graph, and each node is assumed to represent a possible state of the differentiation process under study. The entries of this matrix are the probabilities of transitioning from a given state to any other state represented in the graph.[1,8] Although we base our analysis on an RNA velocity analysis, in principle, any cell-to-cell transition probability matrix can be used as input for trajectory inference (Figure 1A.2).

The objective of trajectory inference with Cytopath is to estimate trajectories from root to terminal cell states, which correspond to the origin and terminus of the differentiation process under study. Root and terminal states can be derived from a Markov random-walk model utilizing the transition probability matrix itself (Figure 1A.3), as described by La Manno et al.,[1] or can be supplied by the user based on suitable prior knowledge.

The trajectory inference process is divided into four steps (Figure 1B; STAR Methods). In the first step, Markov sampling of consecutive cell state transitions is performed based on the probabilities of the transition matrix, resulting in an ensemble of simulated cell state sequences. Sampling is initialized at predefined root states and performed for a fixed number of steps until a sufficient number (auto-selected with default settings) of unique cell state sequences terminating within clusters containing the terminal states have been generated (Figure 1B.1). A precomputed clustering can be provided to Cytopath to determine terminal regions; otherwise, a clustering is computed internally using Louvain.

The generated cell state sequences are individual simulations of the differentiation process from root to terminal state. Because of the stochastic nature of the sampling process, the cell state sequences cannot be considered aligned with respect to the cell states at each transition step. Consequently, in the second step, simulations that terminate at a common terminal state are aligned using dynamic time warping, an algorithm for comparing and aligning temporal sequences with a common root and terminus but possibly different rates of progression. The procedure aligns simulations with a common differentiation coordinate so that cell states from any simulation at a particular differentiation coordinate (pseudotime) represent similar cell states (Figure 1B.2).

Third, consensus expression states across the steps of the aligned simulations are estimated, giving rise to the reported trajectory. Cell states at every step of the ensemble of aligned simulations are averaged, and the average value is considered the consensus state of the trajectory at the particular step (Figure 1C.2). Alternatively, trajectories can be anchored to observed cell states by choosing the cell state closest to the aforementioned average value. Subsequently, the coordinates of the trajectory with respect to the expression space as well as any lower dimensional embeddings, such as UMAP or t-SNE, are calculated.

In the final step, cells are assigned to each step of the inferred trajectory. Assignment is based on an alignment score that evaluates for each cell the similarity of its static as well as the velocity profile with each trajectory step. For efficiency, this alignment score evaluation is restricted to cells in the neighborhood around each trajectory step. However, the user can optionally compute alignment of every cell for every step of every trajectory. The cell level score is used to estimate position in the trajectory (i.e., pseudotime) as well as the relative association of a cell state to

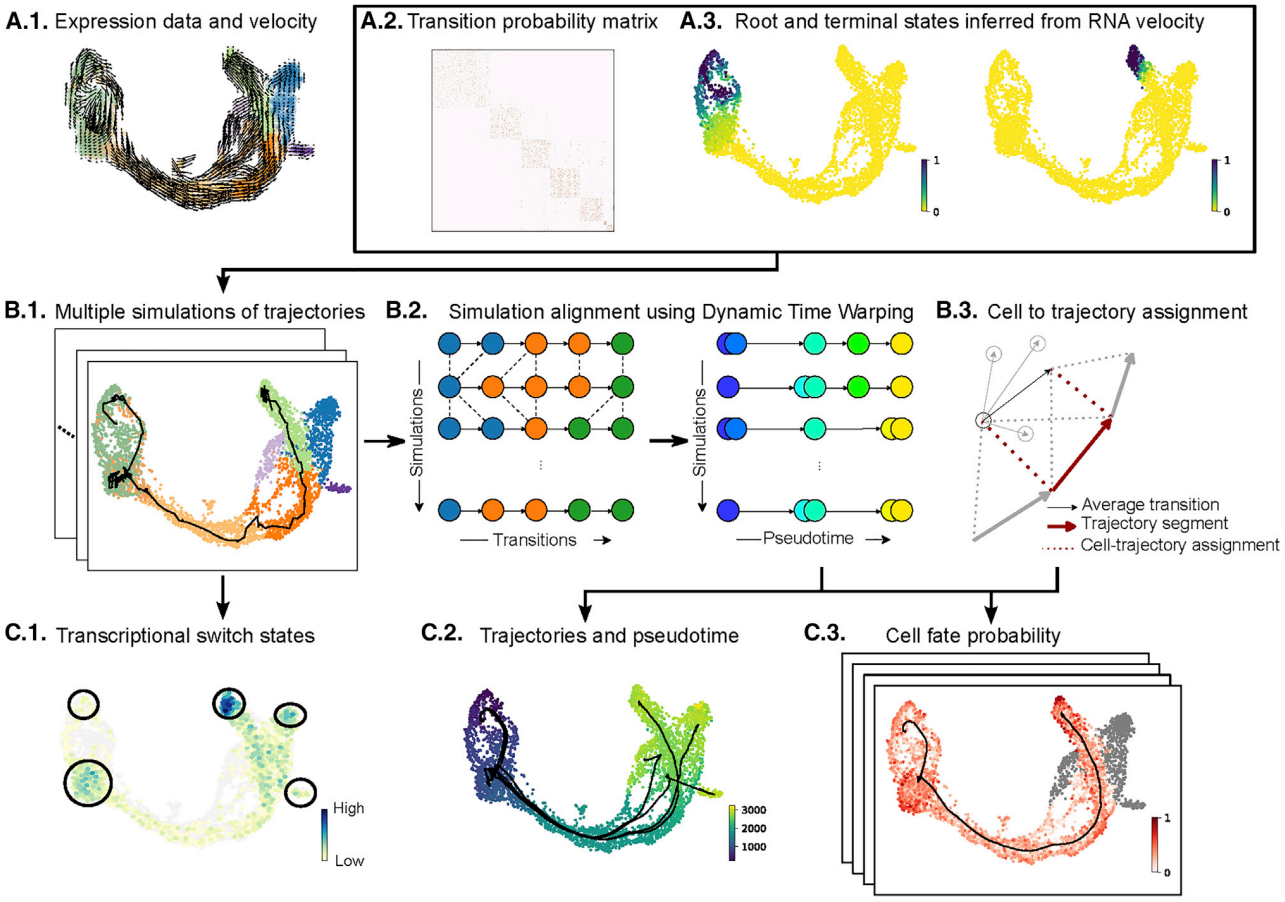


Figure 1. Cytopath overview

(A) Inputs for Cytopath trajectory inference subsequent to an RNA velocity analysis (shown here: inferred using RNA velocity).

(A.1.) Single-cell gene expression profiles and RNA velocity profiles.

(A.2.) Transition probability matrix.

(A.3.) Root and terminal state annotation.

(B) Steps performed during Cytopath inference.

(B.1.) Simulations of the differentiation process generated by sampling a Markov chain based on the cell-to-cell transition probabilities. Sampling is initialized on cells annotated as root states.

(B.2.) Simulations are performed for a fixed number of steps that are automatically selected using the properties of the scRNA-seq dataset. Transition steps are aligned using dynamic time warping. After alignment, cells at each transition step represent the same consensus state.

(B.3.) Cells along the inferred trajectory are assigned to multiple trajectory segments based on the alignment of their average transition vector (with respect to neighbors) and the trajectory segment.

(C) Outputs from Cytopath trajectory inference.

(C.1.) The frequency of simulations terminating at each cell highlights regions of switch in transcriptional programs as well as terminal regions.

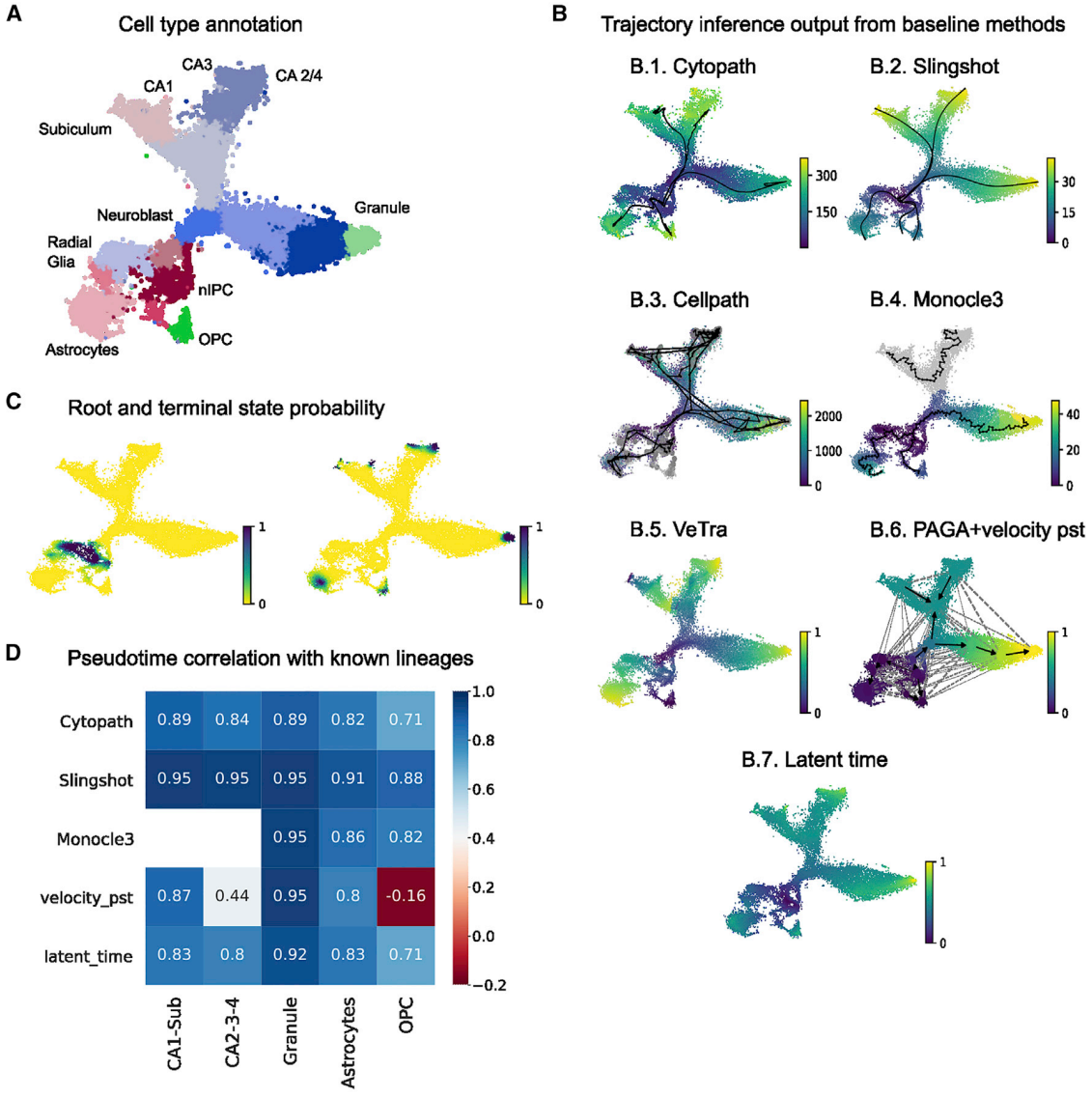
(C.2.) Trajectories are inferred independently for each terminal region. The trajectories are composed of multiple segments. The pseudotime of a cell is estimated as the weighted average segment rank of all segments with which it aligns.

(C.3.) Differential alignment scores to multiple trajectories are used to estimate the cell fate probability with respect to the terminal regions.

possibly multiple branches of a differentiation processes with complex topology (i.e., cell fate) (Figures 1C.2 and C.3).

Reconstruction of neuronal differentiation in the developing mouse hippocampus

We assessed the capability to reconstruct developmental processes with multiple branching, which is a frequent topology for scRNA-seq datasets generated from experiments studying differentiation processes. To this end, we applied Cytopath and baseline methods to the developing mouse hippocampus dataset, which was first used to demonstrate RNA velocity of single cells. This dataset is composed of 18,140 cells. The dataset comprises five terminal regions and a common root state. The topology of the data is multi-furcating, with development branches arising directly from the root state (astrocytes and oligodendrocyte precursors [OPCs]) but also as branches from intermediate states (neuroblast and Cajal-Retzius [CA] differentiation).



(legend on next page)

We recreated the analysis outputs, including RNA velocity and the transition probability matrix, as indicated in the original publication, using scripts made available by the authors. RNA velocity was used to estimate the root and terminal state probabilities.[1]

Spearman correlation of inferred pseudotime with the cell type identities and their ordering reported in the initial study (Figure 2A) were used for trajectory inference performance assessment (Figures 2D and 2E). Cytopath was run using default parameters. It internally selects root and terminal states based on the root and terminal state probabilities estimated in the prior velocity analysis. We also supply the known root and terminal states as supervision to Slingshot and Monocle3 (which accepts root states only) to get the best performance from these methods. This could not be done for VeTra and Cellpath because these approaches do not allow inclusion of this supervision. We also assessed the performance of PAGA with velocity-pseudotime-based directionality and scvelo latent time. The latter two are not trajectory inference methods; PAGA only generates a coarse graph of cluster connectivities, and latent time is only a pseudotime that does not compute lineage association of cells. However, as RNA velocity-based methods that have a partial overlap with Cytopath's core functionality, this comparison is of interest to the community.

Trajectories and pseudotime estimated by each method are shown in Figure 2B. Cytopath estimates a single trajectory to each terminal state as expected from known biology. The Spearman correlation between pseudotime inferred by Cytopath for each trajectory to known ordering of cell types is high (Figure 2D) and robust across multiple runs (Figure 2E). Slingshot also produces trajectories to each terminal state with high correlation but generated one or more spurious trajectories in each run. Although the median Spearman correlation of trajectories inferred by Slingshot is high, it appears to have high variability in its performance, with substantially lower correlation for some runs. This may be due to projection artifacts in the embedding generated in those instances. Monocle3 fails to produce a connected trajectory, producing a disjoint graph, and is therefore unable to estimate pseudotime for a large portion of the dataset. Monocle3 pseudotime, velocity pseudotime, and latent time are inferred per cell, and, unlike other methods presented here, do not partition cells into trajectories. We used known cell type ordering to select cells relevant to each lineage to perform the correlation analysis.

The velocity-pseudotime method appears to compute a global pseudotime that is incompatible with known differentiation of lineages in this dataset. Consequently, the directionality of cluster transitions for PAGA appears to be reversed for CA1-Sub and CA2-3-4 lineages and is unclear for OPC and astrocyte lineages. The latent time method appears to have a similar correlation profile to known lineages as Cytopath (Figure 2D).

VeTra and Cellpath infer erroneous trajectories that initiate at terminal or intermediate states. Cellpath generates a large number of trajectories far exceeding the number of known lineages. This is a pattern that is consistent across several datasets; thus, a quantitative comparison as performed for other methods presented here is not feasible (Figure S1). Both methods also exclude a large number of cells from the trajectory inference process. Because VeTra and Cellpath initialize trajectories in intermediate states, cell assignment by these methods does not correspond to a pattern expected for a hierarchical branching structure (Figures S2A.3 and A.4).

Cell cycle reconstruction

We hypothesized that the ability to infer repeating patterns during differentiation likely differentiates RNA velocity-based trajectory inference from other methods that are based on similarity of expression. First, revisiting the root state implies that inferring the overall direction of the trajectory is not trivial. Second, cells at the origin are a mix of late- and early-stage states that are co-located in expression space but can be expected to have differing velocities. To assess this hypothesis, we compared the reconstruction of the cell cycle in a dataset comprising 1,067 U2OS cells generated using the SMART-Seq2 protocol.[5]

Based on the comparatively low expression of the cell cycle marker genes *Ccne2*, *Cdk1*, *Ccna2*, and *Birc5* (Figures 3A and 3B), we annotated a portion of G0-stage cells (cluster 5) as a G1 checkpoint (Figure S4A). The cell cycle phase annotation per cell from Mahdessian et al.[5] was determined using the fluorescence intensity of GFP-tagged GMNN (530 nm) and RFP-tagged CDT1 (585 nm). Therefore, the association between cell cycle phase and expression levels of markers is not in phase, unlike computational cell cycle phase prediction (Figures S4B and S4C). We use phase annotations only to validate the trajectory reconstruction but not for inference. Root and terminal states were selected based on probabilities estimated using scvelo.[8]

Cytopath generates a full circular trajectory without interruptions from the cells in G1 stage through the intermediate stages back to the G1 stage and further into the cells in the G1 checkpoint stage (Figure 3C). The lower expression of relevant cell cycle markers in the terminal region of the trajectory inferred by Cytopath indicates that the pseudotime inferred by Cytopath is a valid representation of the temporal process. Cytopath infers a second linear trajectory, indicated in red (Figure 3C). This is not unexpected because, apart from the circular route, a direct

Figure 2. Reconstruction of neuronal differentiation in the developing mouse hippocampus

(A) t-SNE projection of the dentate gyrus scRNA-seq dataset annotated with stages of neuronal differentiation.

(B) Trajectory and/or pseudotime inference using (B.1.) Cytopath, (B.2.) Slingshot, (B.3.) Cellpath, (B.4.) Monocle3, (B.5.) VeTra, (B.6.) PAGA + velocity pseudotime (vpt), and (B.7.) scvelo latent time.

(C) Root and terminal state probability used by Cytopath to select root and terminal regions.

(D) Spearman correlation of pseudotime inferred by each method with known ordering of cell types for each lineage.

(E) Methods were run 10 times to assess the effect of stochasticity in inference (Cytopath) and stochastic estimation of the UMAP embedding (Slingshot and Monocle3). Monocle3 produced disconnected graphs in two of 10 runs corresponding to the CA1-Sub and CA2-3-4 lineages, and the correlation value could not be calculated for these two runs.

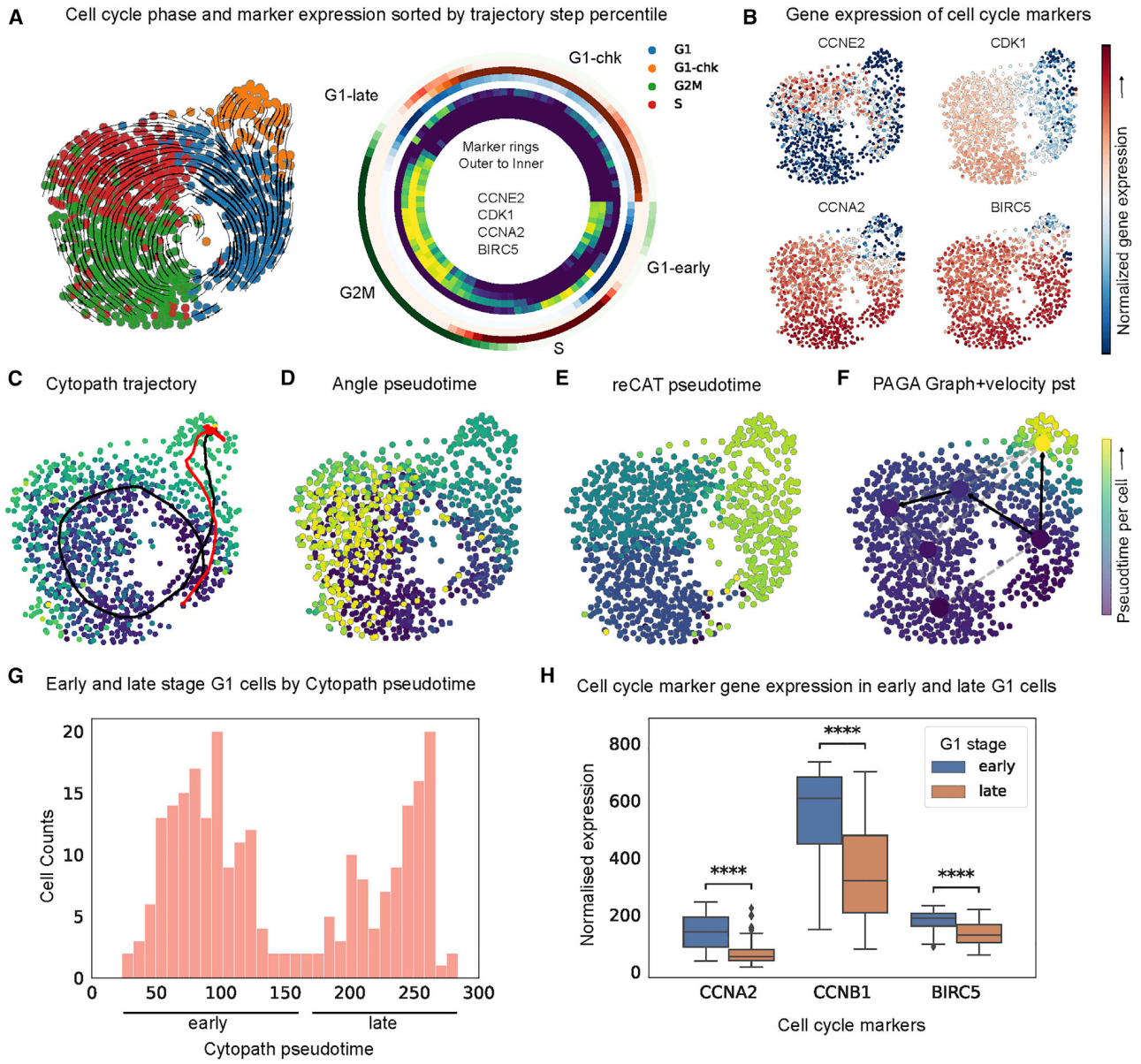


Figure 3. Reconstruction of the cell cycle in the U2OS cell line

(A) RNA velocity stream plot overlayed on the UMAP projection, annotated with the cell cycle phase adapted from Mahdessian et al.[5] Considering all cell-to-trajectory alignments binned into percentiles, the radial heatmap shows cell cycle phase fraction (outer set of rings) and marker expression (inner set of rings) sorted by trajectory step. The directionality of the radial heatmap is clockwise, with the origin at zero degrees.

(B) The separation of G1 phase into G1 and G1-chk was performed on the basis of marker expression of cell cycle genes.

(C–F) Trajectories inferred and pseudotime per cell by (C) Cytopath, (D) Angle, (E) ReCAT, (F) PAGA and vpt.

(G) Distribution of Cytopath pseudotime for cells in the G1 cluster.

(H) Normalized expression of cells classified as early and late G1 cells (blue/orange, respectively). Significance was estimated by an independent t test for each marker.

connection of the root to the terminal region is also a possible outcome.

Because Slingshot and Monocle3 assume a tree-like topology, they are inherently unsuited to model cyclical trajectories. To compare Cytopath with non-velocity-based baseline methods, we selected Angle and ReCAT, methods intended to model the cell cycle in scRNA-seq data. In contrast to trajectories inferred by Cytopath, the pseudotime inferred by Angle and ReCAT is inconsistent with marker expression and cell cycle phase annotation. Both methods fail to correctly model the cell cycle, possibly because of the presence of G1 checkpoint cells that are not actively participating in the cycling process.

Although pseudotime inferred by Angle represents a partially correct sequence of clusters, the G1 checkpoint cluster is not distinguished from the G1 cluster. S phase is incorrectly identified as the terminal state. ReCAT successfully identifies the G1 checkpoint as the terminal state but detects an incorrect sequence of cell cycle phases (Figures 3D and 3E).

With respect to the second part of our hypothesis, we observe that cells in the G1 cluster can be divided into two groups on the basis of pseudotime inferred by Cytopath (Figure 3G). Expression of markers associated with cell cycle is significantly higher in early-pseudotime G1 cells than in those destined to move into G1 checkpoint phase and, accordingly, associated with higher pseudotime (Figure 3H). After trajectory inference, cells are assigned to trajectories using the alignment procedure shown in Figure 1B.3. We sort these cell-to-trajectory alignments by trajectory step percentile and then compute cell cycle phase frequency and average marker expression. Partitioning of G1 cells into early and late stage as well as the difference in marker expression can be clearly observed as two separate bands of G1 (blue) in the radial plot (Figure 3A).

We assessed PAGA with directionality inferred using velocity pseudotime (Figure 3F). PAGA failed to estimate an unbroken sequence of cluster transitions with a default threshold (connections in black), and when the entire connectivity graph is considered (all connections), there are several spurious connections. Although the underlying velocity pseudotime is positively correlated with cell cycle phase, G1 cells are not partitioned into early- and late-stage states. Latent time also correctly models the sequence of cell cycle phases and identifies G1 checkpoint phase as the terminus. However, similar to velocity pseudotime, G1-phase cells are not partitioned into early- and late-stage states (Figure S3B.3).

Finally, VeTra and Cellpath, which are RNA velocity-enabled trajectory inference methods, fail to correctly reconstruct the cell cycle. The pseudotime inferred by VeTra appears to be inconsistent with the root and terminal state probabilities. Both methods infer erroneous trajectories that do not capture the cyclical process, both originating and terminating in intermediate states. Both methods do not assign large number of cells to any trajectory (Figures S1B.3, S1B.4, S2B.3, and S1B.4).

Reconstruction of interlaced cell cycling and bifurcated differentiation in pancreatic endocrinogenesis

We next assessed trajectory inference performance for processes with multiple interlaced non-trivial topologies. To this end, we considered a dataset studying pancreatic endocrinogenesis with lineages to four terminal states (alpha, beta, gamma, and delta cells) and dominant cell cycling at the onset of differentiation.[2,8] Pre-processing, RNA velocity, and transition probability matrix estimation were performed with scvelo[8] using parameters indicated in the notebook associated with this dataset.

Cell type annotation from Bastidas-Ponce et al.[2] and Bergen et al.[8] was used to provide terminal state supervision to Cytopath (Figure 4A), whereas root states were inferred using RNA velocity. The inferred terminal state probability only identifies the beta terminal state (Figure S5A3). If the other terminal states are not manually specified, then this exclusively data-driven approach would report only the trajectory corresponding to the beta lineage. However, Cytopath can be used to generate undirected simulations not constrained to terminate at a fixed terminal region (STAR Methods). For each cell, the frequency of simulations terminating at that state can be used to discover regions of transcriptional state switching. Using this approach, two more terminal states (alpha and delta) could be recovered (Figure 4C). However, the trajectory to the epsilon terminal state could only be constructed by explicitly providing it as a terminal state. The set of trajectories estimated by Cytopath corresponding to the four terminal cell types captures the expected differentiation events of endocrinogenesis (Figure 4B).

The trajectories visualized on the UMAP indicate a potentially cycling structure in early-stage (root-region) cells. To investigate this, we initialized simulations at random cells in the dataset (Figure 4C). We observed an enrichment of terminal states of these undirected simulations in Louvain cluster 0 (Figure S4D). We propose that this observation suggests that this differentiation process is structured in two stages, a cycling and a commitment stage, with the cells in Louvain cluster 0 corresponding to a region of transcriptional switch away from cell cycling. The following inquiries aim to identify evidence for this hypothesis.

Cell cycle scoring of cells in the root region was performed and clearly revealed distinct cell cycle states (Figure 4D). This interpretation is also supported by the differential expression of cell cycle marker genes in the root region (Figure 4E). The trajectory inferred by Cytopath from the ensemble of 8,000 simulations appears to recapitulate the circular structure of the cell cycle (Figure 4G1). Spearman correlation of cell cycle phase with the transition steps of each simulation indicates faithful recapitulation of the cell cycle stages at the single-simulation level (Figure 4F).

The simulation-based approach of Cytopath ensures that, even in the absence of explicit supervision, cyclic transcriptional patterns are reconstructed faithfully. In contrast, possibly because of the absence of RNA velocity information, the designated root states appear to be isotropic for conventional trajectory inference approaches like Slingshot; therefore, they are unable to capture structured transcriptional heterogeneity in this region (Figure 4G.2).

We also show the trajectory estimation with respect to the full pancreatic endocrinogenesis process. Slingshot and Monocle3 produce spurious or too few trajectories, respectively, when provided with all root and endpoints. VeTra reports a spurious trajectory that terminates at the ductal stage, whereas trajectories to beta and alpha are initialized in intermediate or terminal cell states. VeTra and Cellpath exclude a large number of cells from the trajectory inference process (Figures S1C, S2C, and S3C).

Reconstruction of convergent differentiation in the developing neonatal mouse inner ear

Burns et al.[3] have shown that the development of hair cells (HCs) in the sensory epithelium of the utricle originates from transitional epithelial cells (TECs) via support cells (SC). This study also demonstrated a secondary differentiation path from TECs to HCs and put forward the existence of a transitional zone where cells can easily switch fate, resulting in two convergent

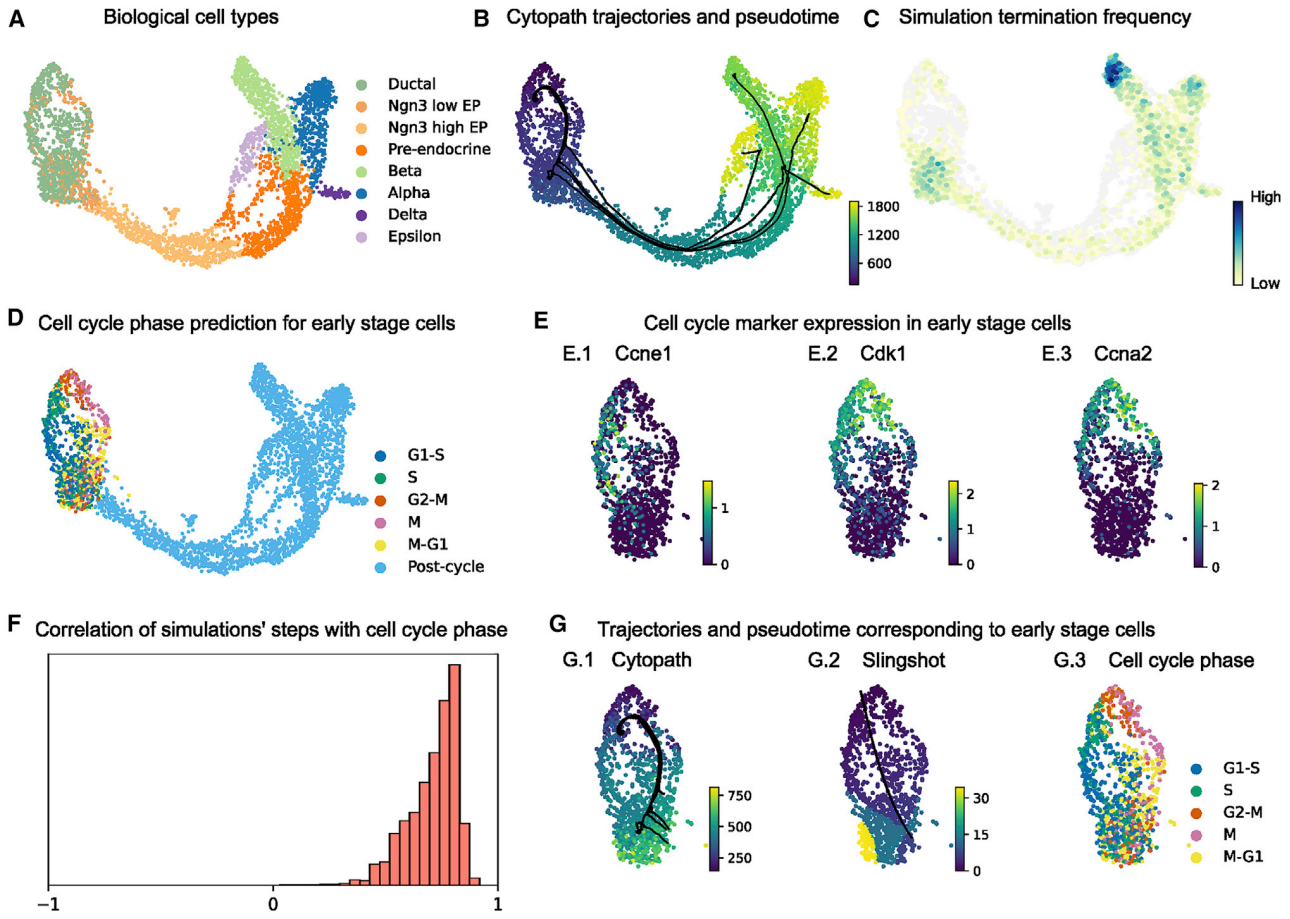


Figure 4. Reconstruction of interlaced cell cycling and bifurcated differentiation in pancreatic endocrinogenesis

(A) UMAP projection of pancreas scRNA-seq data annotated with stages of differentiation.

(B) Trajectories inferred by Cytopath and mean pseudotime per cell.

(C) Log terminal state frequency per cell of undirected simulations initialized at randomly chosen cells.

(D) Computational cell cycle phase prediction.

(E) Cell cycle marker gene expression in the early stage (Louvain clusters 9, 0, and 4) (Figure S4D).

(F) Spearman correlation of cell cycle phase with the transition step of individual Cytopath simulations.

(G) Trajectories and pseudotime inferred by (G.1) Cytopath and (G.2) Slingshot in the root region and (G.3) cell cycle phase annotation.

differentiation trajectories[3] (Figure 5A). This dataset presents a challenge for typical trajectory inference methods, given its relatively small size of 157 cells as well as a convergent differentiation topology that violates the topology-related assumptions of several methods.

Root and terminal state probability estimation using RNA velocity was used to select root and endpoints. Principal-component analysis (PCA) projection of the data was generated as indicated in the original study. Cytopath successfully models the two differentiation trajectories demonstrated in the study. The correlation between known cell type ordering and pseudotime inferred by Cytopath is robust for either lineage (Figures 5D and 5E).

Slingshot, VeTra, and Cellpath generate spurious trajectories that terminate at intermediate states (Figures S1D.2–S1D.4). None of these methods infer the convergent process. PAGA does not return an unbroken chain of cluster transitions with the default threshold (Figure S3D.4) Monocle3 requires a UMAP embedding; therefore, it was not benchmarked in this analysis.

Cytopath pseudotime inference approximates the internal clock of cells

The difference between two expression states is sufficient to order the cells with respect to progression (difference in expression profiles), but without information about the rate of change of gene expression at any state, the pace of differentiation (i.e., the difference in expression profile relative to the internal clock) cannot be inferred. RNA velocity-based trajectory inference and pseudotime inference have the potential to resolve this drawback because RNA velocity provides an approximation of the rate of change of gene expression for each cell.

Single-cell metabolically labeled new RNA tagging sequencing (scNT-seq) was developed as a means to experimentally measure the age of cells undergoing active transcription. To validate their

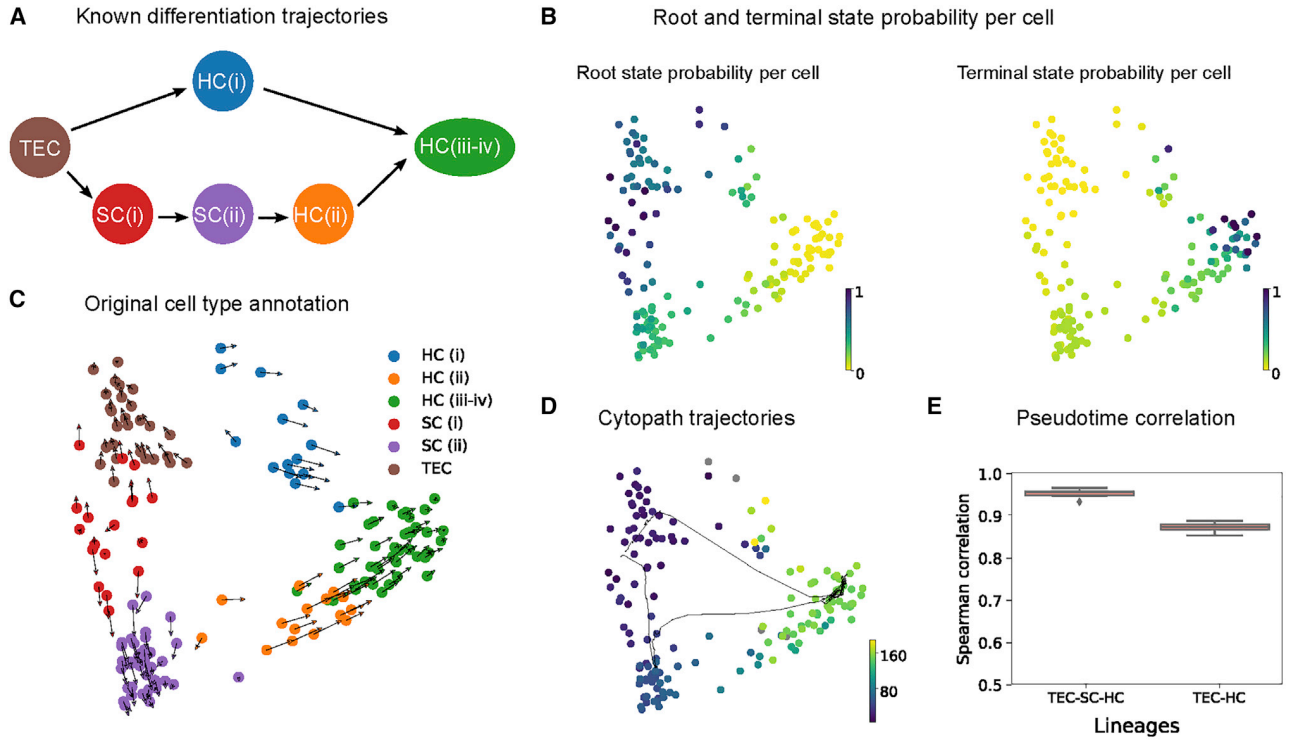


Figure 5. Reconstruction of convergent differentiation in the developing neonatal mouse inner ear

(A) Known differentiation trajectories from Burns et al.[3]

(B) Probability estimated based on RNA velocity of a cell being a root state and terminal state, respectively.

(C) RNA velocity overlayed on the PCA projection of neonatal mouse inner ear data, annotated with stages of differentiation.

(D) Inferred trajectories and mean pseudotime by Cytopath.

(E) Spearman correlations between known lineage ordering of cell types and pseudotime inferred by Cytopath (10 runs).

method, Qiu et al.[13] generated a dataset of mouse cortical neurons stimulated for durations ranging from 0–120 min (Figure 6B). The authors also identified a set of activity-regulated genes (ARGs) whose expression can be directly linked to the duration of stimulation. Unlike typical time-series scRNA-seq datasets, where the asynchronous expression of cells implies that the experimental time is decoupled from the internal clock, in the setting described above, the duration of stimulation is a representation of the biological process (internal clock) time with respect to the ARGs. We performed RNA velocity analysis followed by trajectory inference considering only ARG expression (Figures 6A and 6C). Pseudotime inferred by Cytopath has a monotonic relationship with stimulation time and high Pearson (linear) correlation (Figures 6C and 6D). To assess the specific relevance of RNA velocity in inferring a pseudotime that better approximates the internal clock, we computed a non-velocity-based pseudotime using the trajectory inferred previously by Cytopath (Cytopath-Euclidean pseudotime). This non-velocity pseudotime has lower median correlation over 10 runs than Cytopath pseudotime. Other velocity-based pseudotime estimates also have relatively higher correlation compared with non-velocity-based methods (Figure 6E).

In response to stimulation, neuronal cells undergo a relatively fast polarization phase and subsequently slowly return to a depolarized state that is similar to the root state in terms of expression but not rate of change of expression; i.e., RNA velocity. Cytopath-Euclidean pseudotime does not have a monotonic relationship with stimulation duration and places the 120-min group at a lower pseudotime. We observed the same pattern with Slingshot and Monocle3. However, this may partly be due to poor trajectory inference as well as non-velocity-based pseudotime inference. Surprisingly, latent time and velocity pseudotime, which are also RNA velocity-based pseudotime methods, also showed a similar pattern of lower pseudotime associated with the 120-min group of cells (see notebooks).

In the absence of an experimental measure of process time, it is difficult to conclusively explore the association of pseudotime with process time in other datasets presented in this paper. However, if we assume that RNA velocity is a good approximation of the transcriptional rate of change, then we find that pseudotime inferred using Cytopath outperforms non-velocity-based methods at approximating the real rate of change of transcription. The similarity of a cell's velocity to each of its neighbors indicates the pace of coherent change of transcription in a region of transcriptional space. To quantify this property we define velocity cohesiveness (STAR Methods). High velocity cohesiveness indicates that the cell is present in a region of coherent and therefore rapid transcriptional change because the cell has

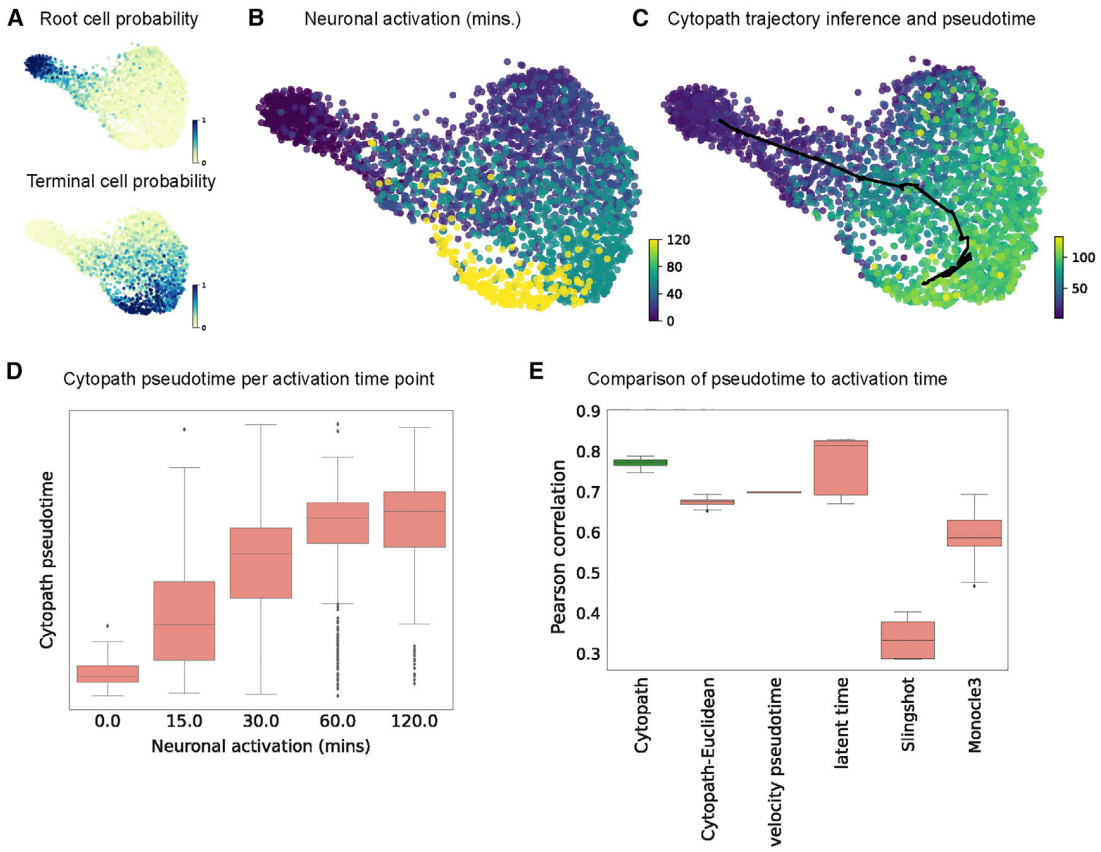


Figure 6. Reconstruction of ARG expression trajectory in mouse cortical neurons

(A) Root and terminal state probability inferred using RNA velocity.

(B) UMAP projection annotated with duration of stimulation for each cell.

(C) UMAP projection annotated with trajectory and pseudotime inferred by Cytopath.

(D) Cytopath pseudotime per cell with respect to stimulation duration. Note the monotonic relationship between median pseudotime and stimulation duration.

(E) Pearson correlation between pseudotime inferred by Cytopath, non-velocity-based pseudotime inferred using Cytopath trajectory inference (Cytopath-Euclidean), and baseline methods.

a high transition probability to similarly oriented transition partners. Conversely, low velocity cohesiveness indicates that the rate of transcriptional change is low and that the cell transitions to its neighbors are less coherently directed. Because simulations generated by Cytopath are based on the aforementioned transitions, we expect that the pseudotime estimated by Cytopath better reflects the rate of real transcriptional change compared with a non-RNA velocity-based pseudotime that, by design, is forced to assume a uniform rate of transcription.

We tested this hypothesis by comparing pseudotime estimated using Slingshot and Cytopath for the pancreatic endocrinogenesis dataset. For each lineage, we estimated the relative rate of change of Slingshot pseudotime with respect to Cytopath pseudotime per cell. The high positive correlation between velocity cohesiveness and velocity magnitude indicates that, in regions with lower velocity magnitude, Slingshot has a lower rate of change in pseudotime compared with Cytopath and vice versa (Figures S4G and S4H). We define the simulation step density of a cell as the number of unique simulation steps visiting this cell. The negative correlation between simulation density per cell and velocity cohesiveness indicates an enrichment of transitions in regions of slower transcriptional change and vice versa (Figure S4I). The overall trajectory inferred from these simulations assigns a larger range of pseudotime values to regions with lower velocity cohesion because smaller changes in expression are associated with relatively larger passage of time compared with regions of higher velocity cohesiveness.

Reconstruction of bifurcating differentiation of CD8$^+$ T cells from scRNA-seq time series data

We assessed the performance of Cytopath on an scRNA-seq time series dataset from CD8 T cell differentiation in chronic lymphocytic choriomenengitis virus [LCMV] infection.[4] In this infection model system, CD8 T cells differentiate from early-activated cells into exhausted and memory-like cells over a period of 3 weeks. Samples were collected at eight experimental time points after infection with LCMV to cover all stages of the process and were sequenced in four batches (Figure 7A). Although these samples are heterogeneous snapshots of a spectrum of differentiation states at a particular time point, they provide an

CellPress
OPEN ACCESS

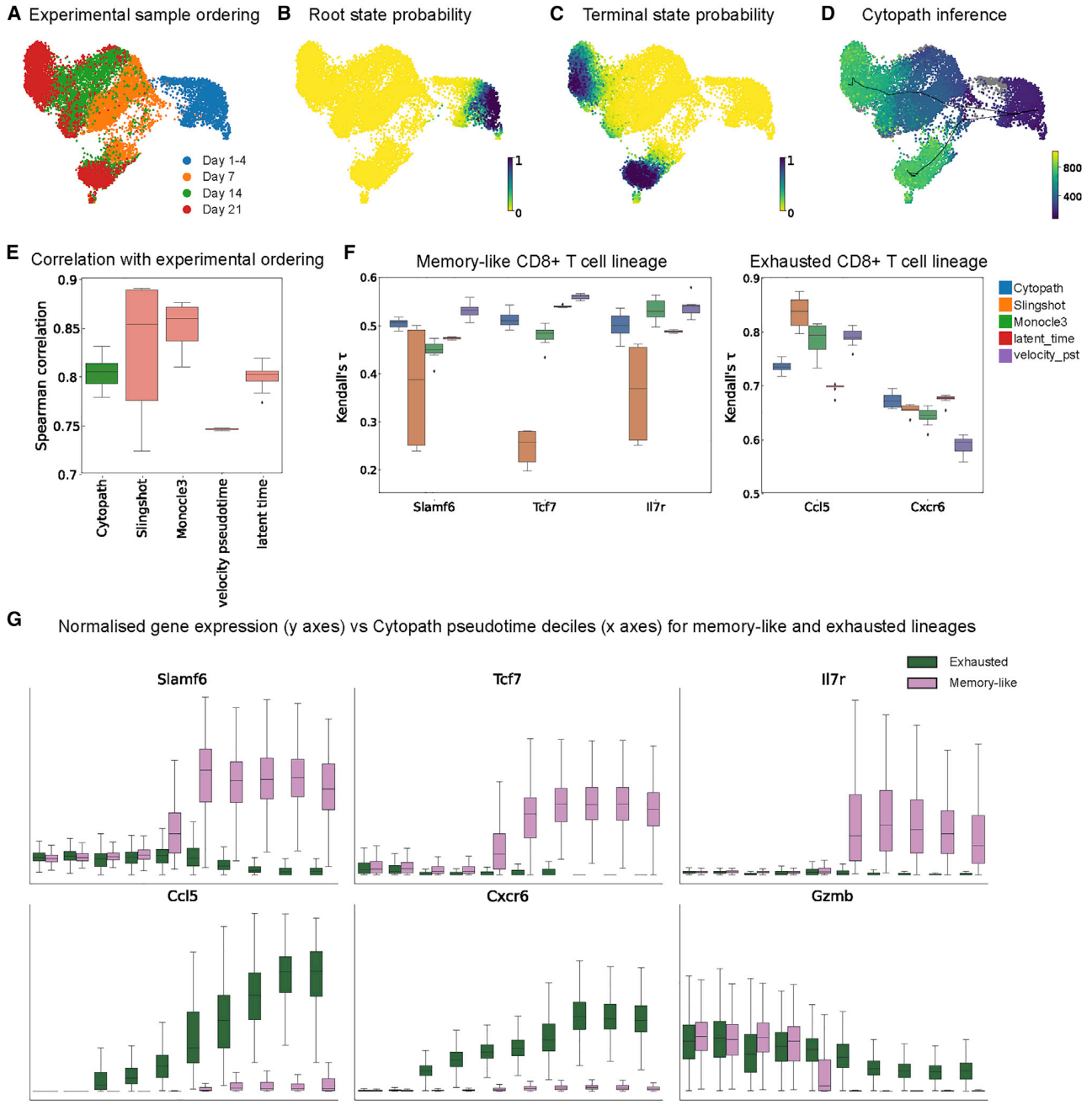


Figure 7. Reconstruction of bifurcating development of CD8⁺ T cells from scRNA-seq time series

(A) Ordering of samples with respect to time of LCMV inoculation.

(B and C) Probability estimated based on RNA velocity of the cell being (B) a root state (C) a terminal state. (D) Trajectories inferred and mean pseudotime per cell by Cytopath.

(E) Correlation of pseudotime estimated by each method, with markers relevant to memory-like CD8⁺ T cell differentiation and exhausted CD8⁺ T cell differentiation, respectively.

(F) Kendall's tau correlation between marker expression and Cytopath pseudotime.

(G) Normalized expression of key marker genes in each Cytopath pseudotime decile per lineage. Top row (*Slamf6*, *Tcf7*, and *Il7r*) markers are expected to be expressed in the memory-like lineage. *Ccl5* and *Cxcr6* are expected to be expressed only in the exhausted lineage. *Gzmb* is not expected to be differentially expressed between the two lineages.

approximate development coordinate. Starting from a population of cells at an early-activated state, differentiation leads into the two distinct terminal states within 5 days of LCMV infection. This differentiation process is characterized by strong transcriptional changes and expression of different surface genes.

We identified the root and terminal states of the process using scvelo[8] (Figures 7B and 7C). The terminal states were validated by expression levels of known marker genes.[4] The exhausted terminal state showed high expression in co-inhibitory markers like CD39 (*Entpd1*), CD160 (*Cd160*), and PD-1 (*Pdcd1*). The memory-like terminal state had high expression of TCF1 (*Tcf7*) and interleukin-7R (IL-7R; *Il7r*). Trajectory inference with Cytopath resulted in two trajectories that led from a shared starting region to the two expected terminal regions. The two trajectories overlapped in the beginning of the process but then sharply diverged at a branchpoint (Figure 7D).

Comparing the pseudotime estimates from Cytopath with the discrete experimental time labels from the samples showed high agreement of the two. The experimental time labels, corresponding roughly to the developmental coordinate, were ordered correctly and with high Spearman correlation between pseudotime and time labels (Figure 7E). Unlike analysis of the neuronal activation dataset's ARGs, the experimental time in this setting is not a precise representation of the internal clock of every cell. This is due to asynchronous activation of expression arising from cell-to-cell variation of antigen exposure times. Therefore, we only expect a correct ordering of experimental time labels with respect to median pseudotime per experimental sample and not a perfect correlation of pseudotime and experimental time label per cell.

Cytopath suggests a bifurcating trajectory model with trajectories originating at biologically relevant root states and terminating at either of the expected terminal states (Figure 7D). Slingshot also inferred trajectories to either terminus but generated a third spurious trajectory in the early-stage cell group that cannot be matched to any expected infection-induced differentiation trajectory (Figure S1F.2). VeTra and Cellpath infer multiple erroneous trajectories that do not initiate at the root state and estimate pseudotime that does not correspond to the differentiation process at all (Figures S1F.3, S1F.4, S2F.3, and S1F.4). Monocle3 reconstructs the global structure of the data but includes additional loops and branches in the exhaustion branch (Figure S3F.2).

We also assessed the correlation of pseudotime estimates with canonical gene expression markers of the memory-like (*Slamf6*, *Tcf7*, and *Il7r*) and exhaustion branch (*Ccl5* and *Cxcr8*). We observe that pseudotime inferred by Cytopath is highly correlated with lineage-specific markers (Figure 7F).

We then tested the validity of the average trajectories of Cytopath by the expression profiles of known lineage marker genes in the differentiation process. The chemokine receptor CXCR6 has been shown to mark exhausted T cells in chronic LCMV infection.[17] The average expression of *Cxcr6* increases in the trajectory toward the exhausted cluster just prior to divergence of the two branches (Figure 7G), indicating that the paths are indeed governed by the exhaustion process. Conversely, T cell factor 1 (TCF1) and expression of its gene *Tcf7* are established markers of memory-like cells.[18] Expression of this gene was increased in memory-like cells just after the cells started to diverge after the bifurcation point. Toward the memory-like terminal state at late time points, *Tcf7* expression is exclusive to the memory-like population. An additional observation is the high expression of *Gzmb* early in both branches that drops off toward later time points (Figure 7G). The expression of *Gzmb* is a shared feature of both branches and known to decrease in both branches as the infection progress and expression is low toward late timepoints.[19]

Cytopath is able to reconstruct biologically relevant differentiation trajectories from a long-term time series dataset in a more accurate and reproducible manner than widely used tools. We identified correct differentiation branches of CD8 T cells in chronic infection, demonstrated by correct ordering of the experimental time labels and expression levels of branch-specific gene expression markers. For this system, several phenotypic populations and characteristic markers have been described before, but the connecting differentiation trajectories of those populations are a subject of ongoing research.[20–23] These studies provide evidence of branching in the development process, and only recently, in conjunction with simulation-based trajectory inference, has it been possible to resolve this event in more detail.[4]

DISCUSSION

Trajectory inference is a challenging task since scRNA seq data is noisy and - until recently - has been evaluated to achieve only static expression profile snapshots. Trajectory inference tools typically operate in low-dimensional embeddings, especially two-dimensional projections such as UMAP and t-SNE, possibly obfuscating complex trajectory topologies such as multifurcations and cycles because of more dominant sources of variation. Inclusion of directional transcriptional activity estimates from RNA velocity analyses is expected to achieve more precise and sensitive trajectory inference. With Cytopath, we present an approach that takes advantage of this information.

The transition probability matrix used to generate simulations is computed from high-dimensional gene expression and velocity profiles. Because Cytopath is based on transitions that use the full expression and velocity profiles of cells, it is less prone to projection artifacts distorting expression profile similarity. This approach specifically considers likely and discards unlikely transitions and therefore is able to identify, for instance, cyclic trajectories in an apparently diffusely populated and isotropic region of expression space (Figures 3 and 4). These hidden transcriptional patterns are made apparent by the simulation-based approach without any explicit supervision. Non-RNA velocity-based methods struggle to discriminate between cells corresponding to different stages or branches of cyclical and convergent processes, respectively, because the cells appear to be co-located in expression space. However, even RNA velocity-based methods[6] do not readily present this information to the user, even when the pseudotemporal ordering or cell fate scoring estimated by these tools captures these patterns.

Cytopath analysis requires specification of the root and terminal regions. This requirement is met easily when the cellular process is sufficiently well characterized up to the level of *a priori*

definition of these regions. However, even when this is not the case, Cytopath can detect and utilize tentative root and terminal states from the cell-to-cell transition probability matrix using scvelo.[8] We found pseudotime inferred by Cytopath to be robust to root and terminal probability thresholds within a range of [0.85, 1] (data not shown). Simulations generated using Cytopath can also aid identification of terminal cell types. Compared with the absorbing Markov process-based inference of terminal states implemented in velocyto and scvelo, this approach appears to highlight more terminal states in the pancreatic dataset (Figure S5B3). Intermediate quasi-stationary cell states induced, for instance, by a switch of transcriptional programs appear to be highlighted by this procedure, as indicated by the switch from the cell cycle to islet cell differentiation in the pancreatic endocrinogenesis study.

Although Cytopath is primarily a trajectory inference tool, we leverage the alignment-based association of cells to inferred trajectories to generate additional results that overlap in functionality with other RNA velocity-based tools. For instance, Cytopath can also be used to predict cell fate. Differentiation potential of cells estimated as the entropy of cell fate probability across the terminal states can be used to investigate branching events (Figure S6).

Other RNA velocity-based trajectory inference tools, VeTra and Cellpath, appear to perform significantly worse than non-velocity-based methods used as benchmarks in this study (Figures S1, S2, and S3). We assume that the reason for the difficulty to recapitulate trajectories could be their built-in lack to guide trajectory inference by separately providing root and terminal states. This appears to be a contrived problem because biological knowledge regarding the identity and role of cells is typically available or, as discussed before, can be estimated separately. Ignoring this information seems to make trajectory inference unnecessarily difficult and could be the reason why Cytopath, as well as Slingshot and Monocle3, perform better than VeTra and Cellpath. Regarding trajectory inference with Cytopath, we find that automatic selection of root and terminal states tends to match biologically relevant root and terminal states, with the strong exception of the pancreatic endocrinogenesis dataset. In general, we recommend that root and terminal state selection should be done by synthesizing all available sources of information, including application-relevant cell type marker expression profiles, analytically derived probabilities based on RNA velocity, and simulations using Cytopath.

PAGA[15] is another popular tool that can include RNA velocity to infer directed connectivity between clusters of an scRNA-seq dataset. Velocity pseudotime allows directed edges to be inferred using PAGA, but an unbroken sequence of connections is not guaranteed (Figures S3A.4–S3F.4). The coarse graph approach has a few disadvantages compared with trajectory inference methods. Dedicated trajectory inference methods such as Cytopath, Slingshot, and Monocle3 can model gradual divergence of lineages. Cell fate scoring estimated by Cytopath constitutes fuzzy assignment of cells to multiple lineages. These methods also support relatively diffuse regions of branching. In contrast, PAGA considers cells in a cluster to be homogeneous with respect to lineage assignment; therefore, branching can only be defined at the cell cluster level.

Addition of RNA velocity is expected to allow pseudotime inference that is a better representation of the internal clock of the cell that corresponds to the pace of differentiation. We show that pseudotime inferred by Cytopath has a monotonic relationship with the process time. We show three points of evidence in this study. The first is the ability to partition cells in the G1 phase of the cell cycling dataset into late- and early-stage cells. From the perspective of gene expression profiles, these cells are co-clustered and appear as a single cell type. However, the RNA velocity-based cell-to-trajectory alignment procedure implemented in Cytopath assigns these cells to an early trajectory step corresponding to G1-S phase transition or a late stage indicating G1-to-G1 checkpoint transition. The biological relevance of this partitioning can be validated by the significant difference in gene expression of a selected set of cell cycle marker genes (Figure 3A). Second, we investigate in more detail the correlation of pseudotime inferred by Cytopath with stimulation duration for the neuronal activation dataset. By restricting the analysis to ARGs whose expression is triggered in response to stimulation and, hence, synchronized per cell, we can consider the experimental time ordering to be coupled to the process time in this dataset (Figure 6). Third, we examine the relationship between velocity magnitude and rate of change of pseudotime. Intuitively, regions with high velocity magnitude are expected to have relatively larger change of expression with respect to the internal clock of cells and vice versa. We show that this relationship is better modeled by Cytopath than non-velocity-based methods (Figures S4G–S4I). Aforementioned results suggest that pseudotime estimated by Cytopath is an improvement on approximating of real rate of change of gene expression.

Cytopath considers generic properties of scRNA-seq datasets, such as the total number of cells and number of inferred root and terminal states, to initialize the hyperparameters of the trajectory inference process. This selection is done with the objective of computational efficiency as well as robust detection of trajectories. All analyses presented in this study utilized the default automatic hyperparameter selection approach, but users still have the option of performing manual tuning.

We expect simulation-based trajectory inference approaches like Cytopath to enable sufficiently precise and unambiguous trajectory inference to achieve testable hypotheses to identify drivers and derive mathematical models of complex differentiation processes.

Limitations of the study

In certain datasets, RNA velocity estimation could be unrepresentative of the true transcriptional dynamics. These issues arise from the simplifying assumptions of time-invariant rates of transcription, splicing, and degradation as well as the assumption of each gene operating under a single regulatory regime. Although these issues typically lead to erroneous inference of dynamics for only a few genes, it is possible for the overall process to be incorrectly modeled when the dynamics of a high proportion of genes are modeled incorrectly. Particular scenarios where RNA velocity estimation fails to recapitulate known dynamics have been explored by Bergen et al.[24]

In the context of trajectory inference with Cytopath, the overall structure and directionality is inferred using the transition

probability matrix, which, in turn, represents the aggregate behavior of RNA velocity of all genes included in the analysis. We do not expect that trajectory inference with Cytopath is meaningful when the underlying RNA velocity estimation itself is faulty.

Specifically, incorrect RNA velocity estimation has an effect on inference of root and terminal cell states (Figure S7). For the erythroid gastrulation dataset, a spurious set of root points is inferred toward the terminus of the expected process, and the endpoints are spuriously inferred in the middle of the expected ordering of cell types (Figure S7B). With the parameterization used by Bergen et al.,[24] we generated a simulated dataset (scvelo simulation function) consisting of features with time-dependent degradation rates (Figures S7E and S7F) and observed a similarly spurious inference of root and terminal states. RNA velocity approaches are typically further challenged when the dataset is composed of mature, terminally differentiated cell types. In a dataset composed of cells representing hepatocyte zonation, we observe a clear directionality inferred using RNA velocity even though no expression dynamics are expected. The root cell probability estimation appears to be fuzzy, and the root cell probability is distributed across the dataset with no discernible pattern (Figures S7C and S7D). We do not recommend performing trajectory inference with Cytopath when the root and terminal cell states are not clearly identifiable. Although not applicable in every scenario when independent sources of information regarding the biological identity of cells are available, such as expression of validated expression markers, we recommend that users verify the plausibility of root and terminal states before proceeding with trajectory inference.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Trajectory inference with cytopath
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Evaluating dynamical properties of cytopath pseudotime
  - Comparison of trajectory inference approaches
  - RNA velocity analysis
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2022.100359.

ACKNOWLEDGMENTS

We thank Jan Schleicher and Sebastian Bischoff for critical feedback. R.G. was supported by Zentren für Personalisierte Medizin (ZPM) Innovationsthemen and DFG EXC number 2064/1 – project number 390727645.

AUTHOR CONTRIBUTIONS

Conceptualization, R.G. and M.C.; methodology, R.G., G.G., and M.C.; software, R.G. and G.G.; formal analysis, R.G.; investigation, R.G., D.C., and G.G.; resources, D.C. and A.O.; data curation, R.G.; writing – original draft, R.G. and D.C.; writing – review & editing, R.G. and M.C.; visualization, R.G.; supervision, M.C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 20, 2022
Revised: October 2, 2022
Accepted: November 11, 2022
Published: December 19, 2022

REFERENCES

1. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498. https://doi.org/10.1038/s41586-018-0414-6.
2. Bastidas-Ponce, A., Tritschler, S., Dony, L., Scheibner, K., Tarquis-Medina, M., Salinno, C., Schirge, S., Burtscher, I., Böttcher, A., Theis, F., et al. (2019). Massive single-cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. Development *146*, dev. 173849. https://doi.org/10.1242/dev.173849.
3. Burns, J.C., Kelly, M.C., Hoa, M., Morell, R.J., and Kelley, M.W. (2015). Single-cell RNA-Seq resolves cellular complexity in sensory organs from the neonatal inner ear. Nat. Commun. *6*, 8557. https://doi.org/10.1038/ncomms9557.
4. Cerletti, D., Sandu, I., Gupta, R., Oxenius, A., and Claassen, M. (2020). Fate trajectories of CD8+ T cells in chronic LCMV infection. Preprint at bioRxiv. https://doi.org/10.1101/2020.12.22.423929.
5. Mahdessian, D., Cesnik, A.J., Gnann, C., Danielsson, F., Stenström, L., Arif, M., Zhang, C., Le, T., Johansson, F., Schutten, R., et al. (2021). Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. Nature *590*, 649–654. https://doi.org/10.1038/s41586-021-03232-9.
6. Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., Schiller, H.B., et al. (2022). CellRank for directed single-cell fate mapping. Nat. Methods *19*, 159–170. https://doi.org/10.1038/s41592-021-01346-6.
7. Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L., and Pe'Er, D. (2019). Characterization of cell fate probabilities in single-cell data with Palantir. Nat. Biotechnol. *37*, 451–460. https://doi.org/10.1038/s41587-019-0068-4.
8. Bergen, V., Lange, M., Peidli, S., Wolf, F.A., and Theis, F.J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. Nat. Biotechnol. *38*, 1408–1414. https://doi.org/10.1038/s41587-020-0591-3.
9. Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genom. *19*, 477. https://doi.org/10.1186/s12864-018-4772-0.
10. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*, 381–386. https://doi.org/10.1038/nbt.2859.
11. Weng, G., Kim, J., and Won, K.J. (2021). VeTra: a tool for trajectory inference based on RNA velocity. Bioinformatics *37*, 3509–3513. https://doi.org/10.1093/bioinformatics/btab364.

12. Zhang, Z., and Zhang, X. (2021). Inference of high-resolution trajectories in single-cell RNA-seq data by using RNA velocity. Cell Rep. Methods *1*, 100095. https://doi.org/10.1016/j.crmeth.2021.100095.

13. Qiu, Q., Hu, P., Qiu, X., Govek, K.W., Cámara, P.G., and Wu, H. (2020). Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. Nat. Methods *17*, 991–1001. https://doi.org/10.1038/s41592-020-0935-4.

14. Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nat. Biotechnol. *37*, 547–554. https://doi.org/10.1038/s41587-019-0071-9.

15. Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F.J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. *20*, 59. https://doi.org/10.1186/s13059-019-1663-x.

16. Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496–502. https://doi.org/10.1038/s41586-019-0969-x.

17. Sandu, I., Cerletti, D., Oetiker, N., Borsa, M., Wagen, F., Spadafora, I., Welten, S.P.M., Stolz, U., Oxenius, A., and Claassen, M. (2020). Landscape of exhausted virus-specific CD8∼T cells in chronic LCMV infection. Cell Rep. *32*, 108078. https://doi.org/10.1016/j.celrep.2020.108078.

18. Utzschneider, D.T., Charmoy, M., Chennupati, V., Pousse, L., Ferreira, D.P., Calderon-Copete, S., Danilo, M., Alfei, F., Hofmann, M., Wieland, D., et al. (2016). T cell factor 1-expressing memory-like CD8 T cells sustain the immune response to chronic viral infections. Immunity *45*, 415–427. https://doi.org/10.1016/j.immuni.2016.07.021.

19. Wherry, E.J., Ha, S.-J., Kaech, S.M., Haining, W.N., Sarkar, S., Kalia, V., Subramaniam, S., Blattman, J.N., Barber, D.L., and Ahmed, R. (2007). Molecular signature of CD8+ T cell exhaustion during chronic viral infection. Immunity *27*, 670–684. https://doi.org/10.1016/j.immuni.2007.09.006.

20. Chen, Z., Ji, Z., Ngiow, S.F., Manne, S., Cai, Z., Huang, A.C., Johnson, J., Staupe, R.P., Bengsch, B., Xu, C., et al. (2019). TCF-1-Centered transcriptional network drives an effector versus exhausted CD8∼T cell-fate decision. Immunity *51*, 840–855.e5. https://doi.org/10.1016/j.immuni.2019.09.013.

21. Zander, R., Schauder, D., Xin, G., Nguyen, C., Wu, X., Zajac, A., and Cui, W. (2019). CD4+ T cell help is required for the formation of a cytolytic CD8+ T cell subset that protects against chronic infection and cancer. Immunity *51*, 1028–1042.e4. https://doi.org/10.1016/j.immuni.2019.10.009.

22. Yao, C., Sun, H.-W., Lacey, N.E., Ji, Y., Moseman, E.A., Shih, H.-Y., Heuston, E.F., Kirby, M., Anderson, S., Cheng, J., et al. (2019). Single-cell RNA-seq reveals TOX as a key regulator of CD8+ T cell persistence in chronic infection. Nat. Immunol. *20*, 890–901. https://doi.org/10.1038/s41590-019-0403-4.

23. Raju, S., Xia, Y., Daniel, B., Yost, K.E., Bradshaw, E., Tonc, E., Verbaro, D.J., Satpathy, A.T., and Egawa, T. (2020). Latent Plasticity of Effector-like Exhausted CD8 T cells contributes to memory responses. Preprint at bioRxiv. https://doi.org/10.1101/2020.02.22.960278.

24. Bergen, V., Soldatov, R.A., Kharchenko, P.V., and Theis, F.J. (2021). RNA velocity—current challenges and future perspectives. Mol. Syst. Biol. *17*, e10282. https://doi.org/10.15252/msb.202110282.

25. Hochgerner, H., Zeisel, A., Lönnerberg, P., and Linnarsson, S. (2018). Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. Nat. Neurosci. *21*, 290–299. https://doi.org/10.1038/s41593-017-0056-2.

26. Pijuan-Sala, B., Griffiths, J.A., Guibentif, C., et al. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. Nature *566*, 490–495. https://doi.org/10.1038/s41586-019-0933-9.

27. MacParland, S.A., Liu, J.C., Ma, X.-Z., Innes, B.T., Bartczak, A.M., Gage, B.K., Manuel, J., Khuu, N., Echeverri, J., Linares, I., et al. (2018). Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. Nat. Commun. *9*, 4383. https://doi.org/10.1038/s41467-018-06318-7.

28. R Core Team (2021). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing). https://www.R-project.org/.

29. Salvador, S., and Chan, P. (2007). FastDTW: Toward accurate dynamic time warping in linear time and space. In Intelligent Data Analysis (IOS Press), pp. 561–580.

30. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

31. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

32. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Dentate Gyrus	Hochgerner et al.[25]	GEO:GSE95753
Cell cycle	Mahdessian et al.[5]	GEO:GSE146773
Pancreatic endocrinogenesis	Bastidas-Ponce et al.[2]	GEO:GSE132188
Neonatal mouse inner ear	Burns et al.[3]	GEO:GSE71982
CD8 development	Cerletti et al.[4]	ENA:PRJEB43201
Neuronal activation	Qiu et al.[13]	GEO:GSE141851
Erythroid gastrulation	Pijuan-Sala et al.[26]	ArrayExpress:E-MTAB-6970
Hepatocyte zonation	MacParland et al.[27]	GEO:GSE115469
Software and algorithms		
Cytopath v0.1.8.post1	This manuscript	https://github.com/aron0093/cytopath https://doi.org/10.5281/zenodo.7278035
Slingshot v2.2.0	Street et al.[9]	https://github.com/kstreet13/slingshot
Monocle3 v1.0.0	Cao et al.[16]	https://github.com/cole-trapnell-lab/monocle3
VeTra commit 63e638c1d60c9faa46f74 e8ce5a226c8d9f5c40e	Weng et al.[11]	https://github.com/wgzgithub/VeTra
Cellpath v0.2.dev0	Zhang et al.[12]	https://github.com/PeterZZQ/CellPath
Cellrank v1.5.1	Lange et al.[6]	https://github.com/theislab/cellrank
scvelo v0.2.4	Bergen et al.[24]	https://github.com/theislab/scvelo
dynverse/ti_angle:v0.9.9.02	Saelens et al.[14]	https://github.com/dynverse/dynmethods/blob/master/R/ti_angle.R
dynverse/ti_recat:v0.9.9.01	Saelens et al.[14]	https://github.com/dynverse/dynmethods/blob/master/R/ti_recat.R
Python v3.8.6	Python Software Foundation.	https://www.python.org/
R v4.1.2	R Core Team.[28]	https://www.r-project.org/
Other		
Dentate Gyrus	La Manno et al.[1]	https://github.com/velocyto-team/velocyto-notebooks/blob/master/python/DentateGyrus.ipynb
Cell cycle	Mahdessian et al.[5]	https://github.com/CellProfiling/SingleCellProteogenomics/
Pancreatic endocrinogenesis	Bergen et al.[24]	https://scvelo.readthedocs.io/scvelo.datasets.pancreas/#scvelo.datasets.pancreas
Erythroid gastrulation	Bergen et al.[24]	https://scvelo.readthedocs.io/scvelo.datasets.gastrulation_erythroid/#scvelo.datasets.gastrulation_erythroide
Hepatocyte zonation	Pijuan-Sala et al.[26]	https://github.com/BaderLab/HumanLiver
Simulation	Bergen et al.[24]	https://scvelo.readthedocs.io/scvelo.datasets.simulation/#scvelo.datasets.simulation

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Manfred Claassen (manfred.claassen@med.uni-tuebingen.de).

Materials availability

This study did not generate new unique reagents.

CellPress
OPEN ACCESS

Data and code availability

- Raw data for all datasets analyzed in this study are accessible on public repositories. The accession numbers are listed in the key resources table. Links to processed data, if available, have also been provided.
- Cytopath has been implemented as a python package and can be found at the following GitHub repository (https://github.com/aron0093/cytopath) and at https://doi.org/10.5281/zenodo.7278035.
- Cytopath is also available for installation via PyPI using the command 'pip install cytopath'.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Trajectory inference with cytopath

Cell clustering

Any grouping of cell states, such as clustering from widely used community detection algorithms or cell type annotations generated by domain experts can be provided as input to Cytopath.

For the set of all cells C and the set of all clusters S, the clustering is $f^s : C \rightarrow S$ where $|S| < |C|$.

By default Cytopath will perform clustering of cells using Louvain via *scvelo*.[8]

Stationary state selection

Root (P_r) and terminal (P_t) state probabilities are used to determine the stationary states as described previously.[1,8] By default, a threshold of 0.99 is used.

The set of root cells C^r is defined as $\{c \in C : P_r(c) \geq 0.99\}$. Similarly, the set of all terminal cells C^t is defined as $\{c \in C : P_t(c) \geq 0.99\}$. Terminal regions S^t are defined as $\{f^s(c) \in S : c \in C^t\}$.

When prior biological knowledge regarding the data exists, users can also manually specify root and terminal cell states or designate entire clusters as root or terminal regions.

Simulations

Simulations are initialized at random cell states selected uniformly within the defined root cells and consist of a fixed number of cell to cell transitions.

At each step, a single transition from the current cell state is realised based on the cell to cell transition probability matrix P . Each row of the matrix contains the probability of transition from the current state (row index) to another cell in the dataset (column index). The cell state c_{ij} at step i of simulation j is selected randomly according to P from the nearest neighbors of $c_{(i-1)j}$.

Let F be the cummulative probability distribution of P . A value κ is sampled from a uniform distribution over [0, 1) and,

$$c_{ij} = argmin_{c \in \mathcal{C}} \left(F\left(c \,/\, c_{(i-1)j}\right) - \kappa \right) \ni F\left(c \,/\, c_{(i-1)j}\right) \geq \kappa$$

Technical parameter selection

Under default settings, the number of simulation steps and minimum number of simulations to be generated are automatically adjusted based on the size of the dataset and number of terminal regions. The purpose of this adjustment is to make the sampling process computationally efficient. The scaling parameters were determined by empirical testing.

The number of simulations steps i_{max} is initialised as $[5 * log_{10}(|C|)]$ This represents an increase of five simulation steps per order of magnitude increase in the size of the dataset. The minimum number of unique simulations to be generated per terminal region m is selected as $[500 * log_{10}(|C|)]$.

Subsequently the number of simulation steps and number of simulations to be sampled are adjusted during the sampling process in an iterative fashion based on the proportion of simulations that terminate at terminal regions in the previous iteration, until the minimum number of unique simulations per terminal region have been generated.

Let J be the set of all simulations generated in an iteration and J^t be the set of simulations terminating in terminal regions. If $|J^t| \leq 0.1 * |J|$ then the number of simulation steps is doubled.

Let J^t_{lag} be the set of simulations terminating at the terminal region with least number of simulations, $\min_{s \in S^t} |J^t_s|$. If $\left|J^t_{lag}\right| < 0.6 * m$ then the number of simulation steps are incremented by $i_{max} * (m \,/ max(i_{max}, \left|J^t_{lag}\right|))$.

If the two conditions above are met and the minimum number of unique simulations per terminal region are not obtained for any terminal region then more samples are generated until $\left|J^t_{lag}\right| = = m$.

Trajectory inference

Simulations that terminate within terminal regions are considered for trajectory inference. Trajectory inference is performed by first clustering the simulations and then aligning them using Dynamic Time Warping, which is an algorithm that allows alignment of temporal sequences with a common origin and terminus that possibly have different rates of progression.

For any two simulations $A = \{c_{0a} \ldots c_{ia}\}$ and $B = \{c_{0b} \ldots c_{ib}\}$, the Euclidean Hausdorff distance $H(A,B)$ is defined as,

$$H(A,B) = \max\left(\max_{c_a \in A}\left(\min_{c_b \in B} d(c_a, c_b) \right), \max_{c_b \in B}\left(\min_{c_a \in A} d(c_b, c_a) \right) \right)$$

where d is Euclidean distance. Simulations terminating within a single terminal region J_s^t for $s \in S^t$, are clustered using Louvain based on Euclidean Hausdorff distance.

Each cluster of simulations is then aligned in a greedy-pairwise fashion using the *fastdtw* python package to generate a single ensemble sequence per cluster which we refer to as a sub-trajectory.[29]

The mean value of coordinates of cells at each step of the aligned simulations is the coordinate of the (sub-) trajectory at that step. Subsequently, a second round of clustering and alignment is performed, using the sub-trajectories from the first round to produce trajectories that are reported by Cytopath. By default the number of expected trajectories per terminal region is not specified allowing from unbiased inference of multiple trajectories per terminal region. However, users have the option to manually enforce the number of trajectories to infer per terminal region.

Identification of compositional clusters

For each trajectory compositional clusters are identified from the set of clusters provided in the step *cell clustering*. For each step i of the trajectory, its neighborhood M_i in PCA space is recovered with a K-dimensional tree search.[30] Cell clusters with a representation larger than a threshold frequency (Default:0.3) for at least one M_i are considered compositional clusters of that trajectory.

Alignment score

After the trajectory coordinates have been inferred, the cell-to-trajectory association is computed. Trajectories inferred by Cytopath are segmented and cells in the compositional clusters of a trajectory are aligned to the segments of the trajectory.

For a cell with neighbors K, its alignment score to step i of a trajectory is the maximum of two scores. The score with respect to the trajectory segment b from steps $i-1$ to i, ξ_i^b is calculated as,

$$\xi_i^b = \frac{1}{|K|}\sum_k^K cos\left(\eta_k^b\right) \cdot exp\left(\gamma_k\right)$$

where η is the cosine angle between the section of the trajectory and all possible transition partners k of the cell. γ is the cosine similarity between the velocity vector of the cell with the distance vector between the cell and its neighbors.

The score with respect to the trajectory segment f from steps i to $i+1$, ξ_i^f is calculated similarly,

$$\xi_i^f = \frac{1}{|K|}\sum_k^K cos\left(\eta_k^f\right) \cdot exp\left(\gamma_k\right)$$

The alignment score is an extension of the transition probability estimation implemented in *scvelo* by weighting each transition of a cell with respect to its alignment to a trajectory. The score is used for assessing the position of a cell with respect to a trajectory (pseudotime) and subsequently to compute a fate score with respect to multiple lineages. The alignment score p_i of the cell with respect to step i is $max(\xi_i^b, \xi_i^f)$.

For each cell the final alignment score p with respect to a particular trajectory is an average of its alignment scores to multiple step segments of the trajectory. By default, it is the mean, however other averages can also be used.

Cell fate score

For each cell its alignment score, relative to multiple trajectories is the cell fate score. Cell fate score f_{traj} for cell with respect to a trajectory is

$$f_{traj} = \frac{p_{traj}}{\sum_{trajectories} p_{traj}}$$

Differentiation potential

For each cell the entropy of its cell fate distribution over all terminal states of the dataset was estimated using *scipy.stats.entropy* function. The values were scaled to range [0, 1] over the dataset.

Pseudotime estimation

For each trajectory cells that have an alignment score greater than zero and also belong to a compositional cluster of the trajectory are assigned a pseudotime value with respect to the trajectory. Since a cell can align to multiple steps within a trajectory, the mean step value of a cell weighted by alignment score is taken as its pseudotime value, for each trajectory. Optionally, other averages can also be used.

QUANTIFICATION AND STATISTICAL ANALYSIS

Evaluating dynamical properties of cytopath pseudotime

Terminal state identification using undirected simulations

For each dataset, five thousand simulations were initialized at randomly chosen cells from the dataset and sampling was performed for 30 steps. The log of total count of simulations, from all the clusters, terminating at each cell, rescaled to range [0,1] is reported as propensity for constituting either a terminal state or an intermediate state representing a switch in transcriptional programme.

Cytopath-Euclidean pseudotime

Cytopath-Euclidean pseudotime was computed by considering all the cell-to-trajectory alignments inferred by Cytopath. The cell was assigned to the trajectory segment with minimum Euclidean distance between the cell and the segment in PCA space.

Velocity cohesiveness

Cosine similarity of the distance vector between a cell to its transition partners, and the cell's velocity vector, per cell is stored as the velocity graph. Velocity cohesiveness of cells are the mean values of each row of this matrix.

Pseudotime comparison

Cells were ordered by Cytopath pseudotime. For each cell a window of 50 cells centered on the cell was considered. For each window a linear model was fit with respect to Slingshot and Cytopath pseudotime using *scipy.stats.linregress* function. Rate of change of Slingshot pseudotime vs Cytopath pseudotime was assessed by estimating the slope.

Simulation step density

Simulation step density per cell is the log of the count of simulation steps that are a transition to the cell.

Runtime analysis

process_time function from the *time* package was used to record the time taken for each trajectory inference procedure.

Comparison of trajectory inference approaches

Parameter settings

Default settings were used for all methods except VeTra and Cellpath. Recommended settings based on information published by the authors were used for VeTra and Cellpath. Cytopath was run using default parameters for all datasets. The same root and terminal states used for trajectory inference with Cytopath were provided to other methods. Figure S5 shows the root and terminal state probabilities estimated using *scvelo* for each dataset.

Pseudotime comparison

Spearman correlation of the pseudotime values generated by each method with the cell type cluster ordering for each biological lineage was used to compare the performance of the methods. Kendall's tau was used to assess the correlation of marker expression with the estimated pseudotime. For each dataset and method, analysed with the correlation analysis described above, the analysis was performed on the dataset with ten independent initialisations of the entire processing pipeline.

RNA velocity analysis

Pre-processed data was used wherever available. Subsequent analysis was performed with *scvelo* using standard workflow.

Dentate gyrus

Prepossessing was performed as indicated by the authors of the original study using code published by La Manno et al.[1]

Neonatal mouse inner ear

Raw sequencing data was downloaded from the NCBI GEO database under accession code GSE71982. Quality control including read filtering and adaptor trimming was performed using *fastp*.[31] Reads were aligned to the GRCm39 mouse genome assembly using *STAR version=2.7.8a-2021- 04-2*.[32] Spliced and unspliced counts were estimated using the *velocyto run-smartseq2* command following the recommendation of the developers.

CD8 development

Read counts were realigned and sorted for spliced and unspliced counts using the analysis pipeline from velocyto.[1] Other contaminating cell types were removed from the dataset based on outliers in diffusion components.[4]

Hepatocyte zonation

Data corresponding to clusters Hep 1, 2 and 4 from patient 3 was used to perform the analysis.[27]

Read counts were realigned and sorted for spliced and unspliced counts using the analysis pipeline from velocyto.[1]

ADDITIONAL RESOURCES

Trajectory inference analysis with Cytopath including pre-processing and velocity analysis for each dataset presented in this paper can be found at https://github.com/aron0093/cytopath-notebooks. Download links for *anndata* objects for each dataset are also available in the corresponding notebook.

Documentation including installation instructions can be obtained at https://cytopath.readthedocs.io/.

FATE TRAJECTORIES OF CD8$^+$ T CELLS IN CHRONIC LCMV INFECTION

A PREPRINT

December 22, 2020

Dario Cerletti[1,2], Ioana Sandu[1,2], Revant Gupta[3], Annette Oxenius[2,&,*] and Manfred Claassen[3,&,*]

[1]Institute of Molecular Systems Biology, ETH Zurich, Otto-Stern-Weg 3, 8093 Zürich, Switzerland

[2]Institute of Microbiology, ETH Zurich, Vladimir-Prelog-Weg 4, 8093 Zürich, Switzerland

[3]Internal Medicine I, University Hospital Tübingen, Faculty of Medicine, University of Tübingen, Otfried-Müller-Straße 10, 72076 Tübingen, Germany

[&]shared contribution

[*]Corresponding authors: oxenius@micro.biol.ethz.ch, Manfred.Claassen@med.uni-tuebingen.de

ABSTRACT

In chronic infections CD8$^+$ T cells acquire a state termed "exhaustion" which is characterized by impaired effector functions and expression of co-inhibitory receptors as response to continuous TCR stimulation. Recently, the pool of exhausted T cells has been shown to harbor multiple functionally distinct populations with memory-like and effector-like features, though differentiation and lineage relations between these are unclear. In this work we present a comprehensive scRNAseq time-series analysis from beginning of infection to established exhaustion in CD8 T cells. We apply lineage inference using informed cell transitions derived from RNA velocity to identify differential start and end states and connections between them. We identify a branch region early during chronic infection where pre-committed cells separate into an exhausted and a memory-like lineage and discovered molecular markers demarcating this branch event. Adoptive transfer experiments confirmed fate-commitment of cells only after this branch point. We additionally linked the progression along developmental lineages to antigenic TCR stimulation.

Keywords LCMV · chronic infection · single cell · transcriptome · trajectory · computational

1 Introduction

Viral infections with human immunodeficiency virus (HIV), hepatitis C virus (HCV) and in mice with lymphocytic choriomenengitis virus can result in chronic infection with ongoing viral replication and high antigenic load for weeks or months. This continuous exposure to antigen drives $CD8^+$ T cells into a functionally distinct phenotype, termed exhaustion [1]. This state is characterized by functional, transcriptional and epigenetic changes that result in expression of co-inhibitory receptors such as PD-1, LAG-3, 2B4 and CD160, decreased secretion of cytokines like INFγ and TNFα as well as reduced proliferation and survival. Acquisition of this exhausted phenotype is a continuous and gradual process driven by excessive TCR stimulation [2].

In this context of chronic antigen exposure, CD8 T cells undergo a differentiation program that differs markedly from the one observed during acute resolved infection. Previous studies have analyzed and inferred differentiation trajectories of virus-specific CD8 T cells using bulk or single cell transcriptomic profiling in various systems, including the model of chronic LCMV infection [3, 4, 5]. Most of these studies have inferred differentiation trajectories based on bulk or single cell analyses performed at one or two time points during chronic LCMV infection [4, 5].

Asynchronicity in this process as well as different micro-environments that $CD8^+$ T cells experience result in a heterogeneous population of cells at a given time point of the infection. One sub-population of virus-specific T cells acquires a phenotype that shares properties with memory T cells from acute infection and has been linked to the expression and activity of T cell Factor 1 (TCF1) [3] [6]. In contrast to terminally exhausted or effector T cells, these cells retain proliferative activity and have better survival in the infected host [4]. It is not yet fully understood how and when these different cell states arise during the course of the infection and which intermediate cell states precede these.

Recent advances in sequencing technologies have made it possible to profile cells genome wide on the transcriptional level using single-cell RNA sequencing (scRNAseq). This technology allows capturing the transcriptional heterogeneity of multiple cell populations and to computationally infer sequences of cell states traversed during dynamic processes such as T cell differentiation in chronic infections. When analyzing scRNAseq datasets, cells are treated as points in transcriptome space based on their expression profile. Dimensionality reduction techniques like t-SNE [7] and UMAP [8] construct two-dimensional representations for analysis and interpretation of the high dimensional single-cell expression data. Pseudotime and lineage inference methods aim at constructing likely transitions between cell states [9].

Recent studies aimed at reconstructing cell state sequences of $CD8^+$ T cell differentiation in chronic LCMV infection [4], [5], [10]. They discovered multiple phenotypic subsets, namely memory-like, terminally exhausted and effector-like cells and investigated likely transitions between these subsets. However, these studies lack temporal resolution to reliably infer trajectories and to identify potential branching events in the differentiation process. Samples were either generated from different infection settings at single time-points, or at far spaced time-points. Further, applied trajectory inference methods infer pseudotime and lineages based on similarity of transcripts and lack taking advantage of all the information present in the scRNA seq data. Directionality information is now – in principle – available for trajectory inference via RNA velocity analysis. RNA velocity [11] considers additional information about the ratio of un-spliced to spliced mRNA in transcript data, which serves as a measure to determine the stage (early, intermediate, late) of individual gene expressions and allows to predict the future expression state and hence to better infer the directionality towards their neighbors in the high-dimensional transcriptional space. So far no study leveraged RNA velocity in order to include this information to disambiguate the results from conventional trajectory inference.

In this work we conducted scRNAseq measurements at multiple time-points ranging from the beginning of chronic LCMV infection until manifestation of exhaustion three weeks after infection. This level of time resolution allows more detailed identification of cell states and their differentiation. We further included information from RNA velocity analysis to perform simulation based trajectory inference of differentation events leading to the different terminal $CD8^+$ T cell states observed in chronic LCMV infection. This is the first attempt to make use of RNA velocity to produce informed differentiation trajectories that connect the different cell states. This analysis allowed us to construct faithful

lineage trajectories towards the two endpoints of differentiation, namely a terminally exhausted and a TCF1$^+$ cell population. We identified a potential branching point in the initially shared trajectories and validated our findings using adoptive transfer experiments of cells arising before or after the branching point. We confirmed that cells before the branch point gave rise to both exhausted and TCF1$^+$ cells, whereas exhausted cells after the branch point maintained their phenotype. Additionally, we demonstrated that TCF1$^+$ cells largely retained their phenotype in absence of antigen stimulation, corroborating the end-point differentiation characteristics of this population. However, if exposed to antigen stimulus, the TCF1$^+$ population has the ability to differentiate into terminally exhausted cells, in line with previous adoptive transfer experiments.

2 Results

We first investigated the differentiation landscape of CD8 T cells, followed by RNA velocity analysis to reveal developmental endpoints. Afterwards we identified two branching trajectories towards memory-like and exhausted cell states, respectively. We validated commitment to the branches using adaptive transfer experiments and additionally highlight the importance of antigen stimulation during development.

2.1 Differentation landscape of CD8 T cells during chronic LCMV infection

We acquired single cell transcriptomic data from multiple time points during chronic infection, covering the very early phases (day 1-4), peak phase (day 7), contraction phase (day 14) and late phase (day 21) (Fig. 1), with the aim to capture an increased spectrum of the transcriptional landscape during the course of the infection that would allow a time-resolved analysis of single cell heterogeneity and possibly more accurate inference of differentiation trajectories of virus-specific CD8 T cells. To this end, T cell receptor (TCR) transgenic (tg) LCMV gp33-41-specific CD8 T cells (P14 cells) were adoptively transferred into naïve C57BL/6 mice, followed by infection with LCMV clone 13 (Cl13). Activated and expanded P14 cells were isolated at the above indicated time points and subjected to single cell RNAseq (scRNAseq) analysis using the 10x Genomics platform.

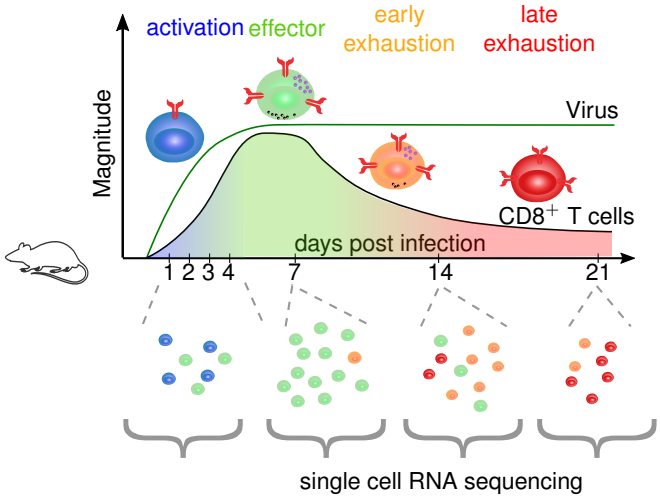


Figure 1: Transgenic P14 CD8 T cells were sampled longitudinally during infection. The samples were acquired from four phases of the infection activation (day 1-4), effector (day 7), early exhaustion (day 14) and late exhaustion (d21) and scRNAseq was performed using the 10x Genomics platform.

For exploratory analysis of the transcriptome data of the multiple time points, we applied commonly used filtering and scaling of the raw data [12] and applied principal component analysis to reduce noisy signals. The resulting multidimensional data was projected into two dimensions using UMAP [8].

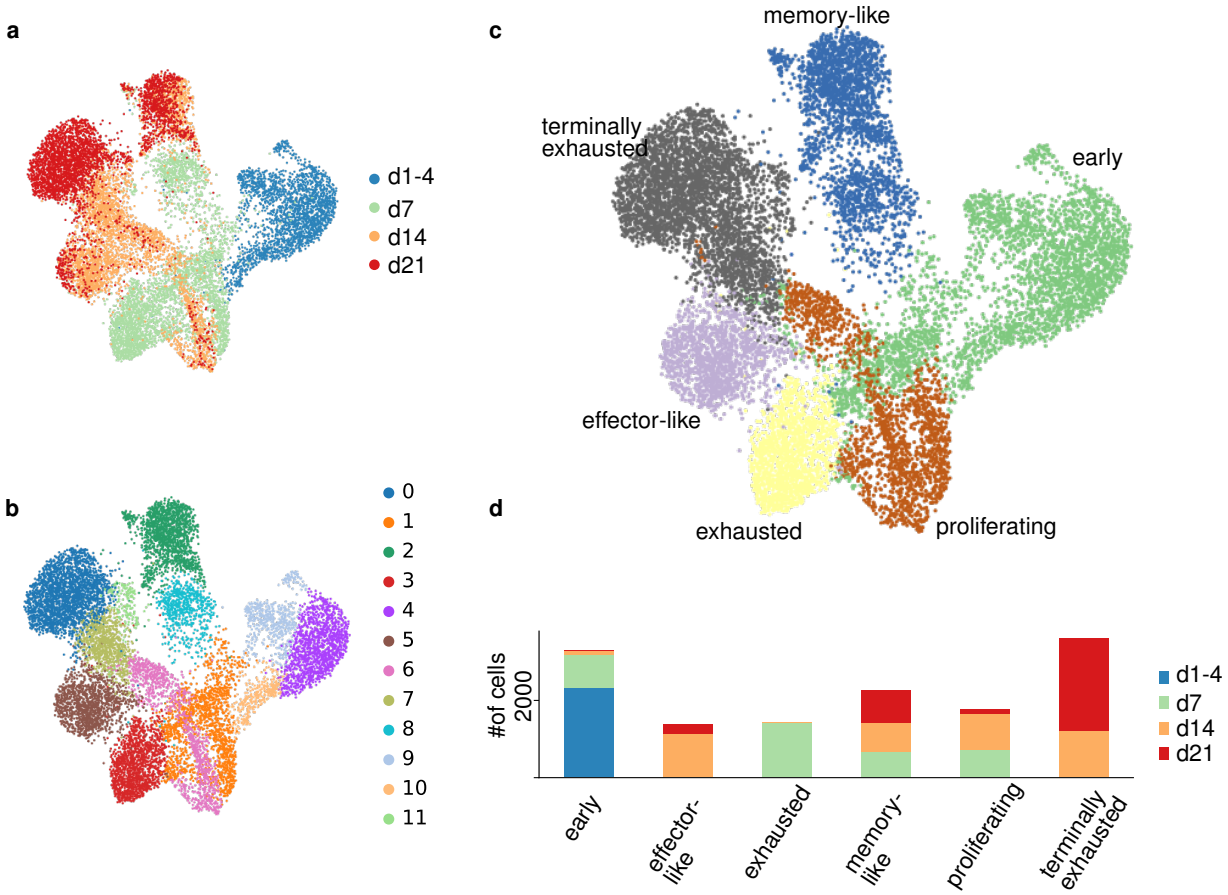


Figure 2: UMAP projections of the filtered and normalized transcript counts are shown. (a) Color indicates the time-point after infection, at which cells were isolated for scRNAseq (b) louvain cluster assignment based on the first 50 PCs (c) phenotypic cluster annotation based on previous marker genes and differentially expressed genes [4] (d) Cell composition of the phenotypic clusters by sample time-point.

The UMAP projection revealed a continuous structure of the data emerging from the time-resolved samples, supporting a continuous developmental process (Fig. 2a). Louvain clustering [13] defined 11 distinct clusters across all samples (Fig. 2b,d, Supp. Fig. S3). Based on previously described markers for the exhausted subsets, such as CD160, CX3CR1 and TCF1 [3, 1, 4], we further aggregated these cluster into 6 phenotypic groups (Fig. 2c). Activated cells from day 1 to 4 post infection (dpi) clustered at one peripheral region in the data, termed in the following **early group** (green in Fig. 2c). Differential expression revealed that the early group presented expression patterns of proliferation as well as of exhaustion, indicated by expression of *Mki67*, *Cdca3* but also *Cd160*.

Conversely, two distinct populations from the latest time point at 21 dpi clustered at the other extremes of the spectrum. Of these "late" endpoints, one population showed high expression of a number of inhibitory receptors, including *PD-1 (Pdcd1)*, *CD39 (Entpd1)*, *LAG-3 (Lag3)* and *CD160 (Cd160)* (Fig. 3a, b), indicating a terminally exhausted phenotype [14]. This **terminally exhausted group** (grey in Fig 2c) was composed of clusters from d7 and 14 and had the highest expression in co-inhibitory receptors and additionally showed high expression of the transcription factor *EOMES*. The other end-point populations showed high expression of the transcription factor *Tcf7* (Fig. 3a, b), the memory-marker *Il7r* as well as *Slamf6*, revealing this cluster as the previously described memory-like population [3]. This **memory-like group** (blue in Fig 2c) was composed of two clusters from day 7, 14 and 21, all having high expression of *Tcf7*.

Cells from 7 and 14 dpi were situated in between the 14 dpi and 21 dpi samples, with the 7 dpi samples being similar to cells from early time-points on one end of the spectrum, but also connecting to already splitting trajectories into exhausted and memory-like populations on dpi 14. At day 7 was one cluster identified that presented clear signatures of exhausted CD8 T cells but retained some expression of *Gzmb* but also apoptotic genes like *Anxa1*, we termed this the **exhauted group** (yellow in Fig. 2c).

At 14 dpi, some of the cells were still connected to the 7dpi cell states but a considerable fraction of cells had already further differentiated towards the two endpoints of 21dpi, in particular towards the memory-like endpoint. One cluster expressing *Cx3cr1* exclusively was termed **effector-like group** (purple in Fig. 2c). Differential expression analysis (Fig. 3a, Supp. Fig. S3) revealed higher expression levels of *Cx3cr1* and additionally killer lectin receptor genes (*Klre1*, *Klra3*). These effector-like cells were only present in samples from day 14 and 21.

We also identified a strong cell cycle component in the two clusters presenting high expression levels of e.g. *Mki67* (cluster 1 & 6) (Fig. 3b). Further, we calculated scores for cell cycle and cell division genes based on the three different cell cycle stages G1 phase, S phase and G2/M phase (Supp. Fig. S4). We observed that this **proliferating group** (brown in Fig. 2c) presented high scores for G2/M phase and was composed of cells from the 7 and 14 dpi and to a lesser degree from the 21 dpi time-point.

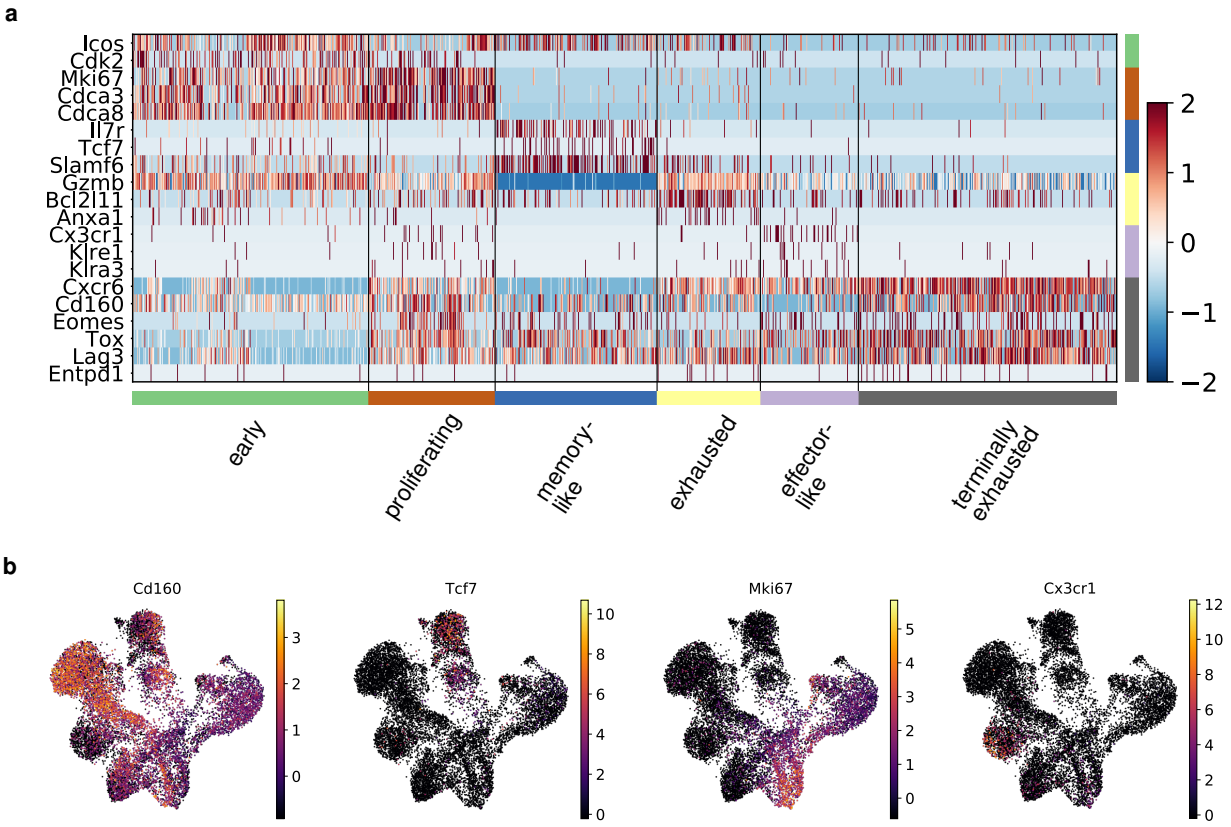


Figure 3: (a) Heatmap of normalized gene expression for a selection of group specific genes. Columns are individual cells arranged by phenotypic groups. The genes in the rows are grouped according to their phenotypic assignment. (b) UMAP projection of expression pattern of identified group specific genes for the terminally exhausted (Cd160), memory-like (Tcf7), proliferating (Mki67) and effector-like group (Cx3cr1)

2.2 RNA velocity analysis reveals developmental end points

Having mapped serial single cell transcriptomes into a continuous landscape of cellular states, we next aimed at inferring informed differentiation trajectories into this landscape. To this end we leveraged RNA velocity [11] and applied it to our longitudinal data set, revealing a vector field demarcating likely transitions between all cell states in the dataset. Applied to our dataset, this analysis clearly revealed a transition flow from early activated cells at 1-4 dpi towards early (dpi 7) and late exhausted cells (dpi 14 & 21) (Fig. 4a).

Using the calculated cell transitions, we defined a Markov process with cells as the states and transition probability estimates from RNA velocity (see Method section). Computing the equilibrium distribution of the forward Markov process corresponded biologically to the most differentiated phenotypes as the cells transitioned to more differentiated states until they had acquired their final transcriptional state and did not differentiate further. Conversely, we inverted the transition probabilities and computed the equilibrium of the backward Markov process. Cell states thereby transitioned to their most likely previous transcriptional state resulting in the most undifferentiated state, having the highest probability.

This allowed us to assign cells that are most likely at the start and at the end of the differentiation process. The highest probably of a start region was at the edge region of the early group (Fig. 4b). This seemed plausible, as we would expect that the differentiation process started at the edge of the earliest sample (dpi 1-4), where there are no preceding cell states. This region showed gene signatures indicating strong DNA synthesis and cell cycle activity, that conferred an activated phenotype (Supp. Fig. S6).

The highest probability for end points was in regions from dpi 21 in the terminally exhausted group (0.5 on average in cluster, maximum 1.0). In this region many signaling related genes had changed expression, which could be a result of co-inhibitory receptor signaling (Supp. Fig. S6). Additionally, there was a local maximum in end point probability in the memory-like group from dpi 21 (0.1 on average, maximum 0.3). Differentially expressed genes in this region comprised typical genes of the memory-like signature, namely *Il7r* and *Tcf7* (Supp. Fig. S6). Both the terminally exhausted cells as well as the memory-like cells seemed to comprise an endpoint of differentiation state.

We assessed and confirmed the robustness of the velocity fields by confirming the practical equivalence of start and endpoint estimates across 574 parameter variants (Supp Fig. S5).

RNA velocity analysis additionally indicated that the process from the start to the end-points is gradual, since there were a multitude of intermediate transcriptional states between the two extremes. We observed also transitions between all these intermediate states (Fig. 4a).

2.3 Simulation based inference reveals trajectories towards exhausted and memory-like phenotypes

Based on the high-dimensional vector field resulting from the RNA velocity analysis, we calculated a transition matrix that contains likely future states for each cell. This transition matrix we used to simulate differentiation trajectories from cells in the start region. We aimed at understanding the developmental paths that an activated $CD8^+$ T cell could follow to acquire the two differentiated end-point phenotypes. The probability of a cell moving from one transcriptional state to another can be approximated by the transition probabilities from RNA velocity. We used the calculated root cells (Fig. 4b) as starting points for stochastic simulations. Each differentiation step in the transition matrix was simulated according to the transition probability until one of the previously defined end stages (Fig. 4b) was reached. This sequence of steps approximated one possible path of differentiation for each cell (Supp. Fig. S7). We simulated 2000 trajectories per endpoint to sample the whole spectrum of possible differentiation trajectories. We observed a strong disbalance in preference for the endpoints. Only about 1% of the simulated sequences ended up in the memory-like cluster, whereas the remaining ones differentiated into the exhausted endpoint. We expected the ratio to be shifted towards the exhausted phenotype but not to this extent, since we measured around 10% memory-like cells at day 21. We performed more simulations towards the memory-like endpoint to balance the number of trajectories. All the obtained

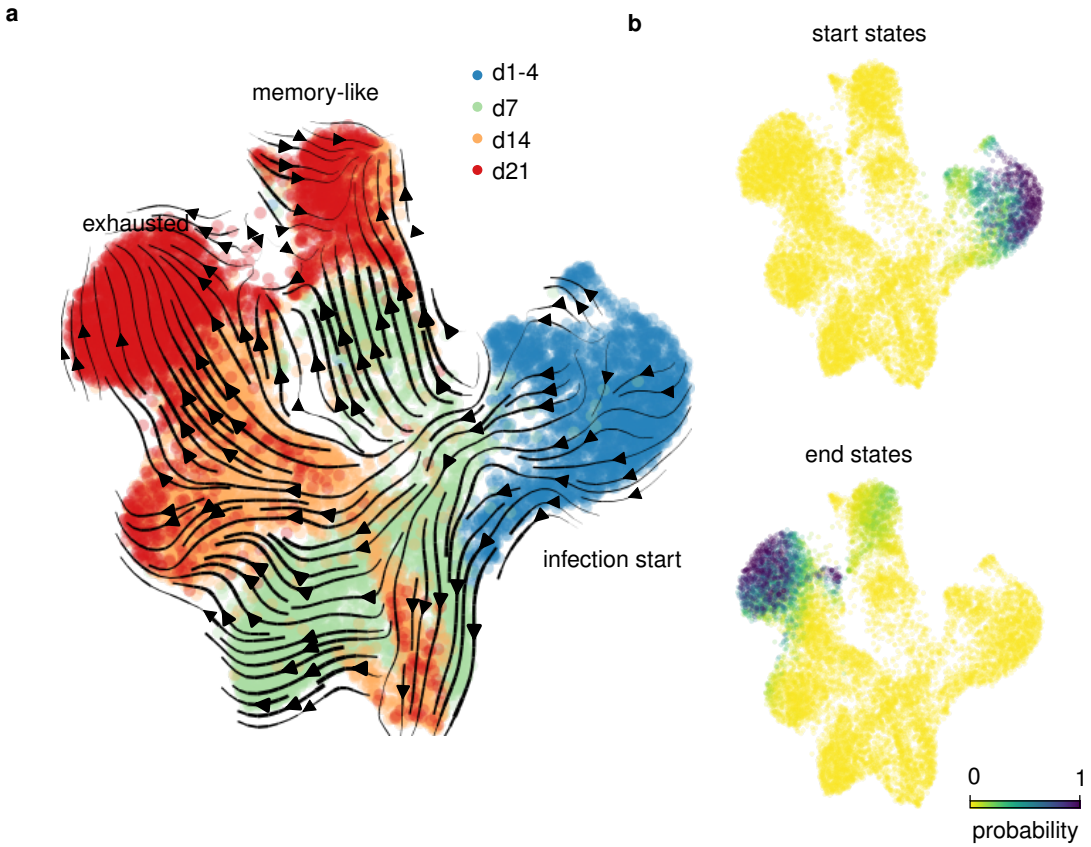


Figure 4: (a) stream plot visualizing likely transitions between cells inferred from RNA velocity
(b) The stationary distribution of the backward and the forward transition matrix, respectively, indicate start and end cell states.

trajectories were then aligned using dynamic time warping and clustered to generate average trajectories. Each cell was assigned to the nearest average trajectory according to the alignment score (Method section 4), calculated from cosine distance of its RNA velocity with the direction of the trajectory (Fig. 6a). This resulted in a temporal ordering of the cells in conjunction with a score to which trajectory it belongs. The detailed procedure of Cytopath is described in [15]).

Our analysis revealed two main trajectories, one towards the exhausted endpoint the other to the memory-like endpoint (Fig. 5). Both trajectories shared the same cell populations up to a region that was composed of cells obtained from around 4 dpi. From thereon the trajectories started to diverge into the two phenotypic branches. The differentiation trajectory towards the exhausted path included the cell population with high cell cycle activity. The memory-like trajectory did not seem to pass the region of high proliferation, but diverged earlier and transitioned towards the memory-like endpoint.

Multiple genes were found to be differentially expressed between the two trajectories (Supp. Fig. S8). In the memory-like trajectory *Slamf6*, *Ccr6*, *Tnfsf8*, *Xcl1* and *Cxcl10* were expressed at higher levels. Many of them showing gradually increasing expression towards the differentiated endpoint (*Slamf6*, *Ccr6*, *Tnfsf8*). *Xcl1* was highest at the start of the trajectory and later decreased but was still maintained at much higher levels than on the exhausted branch. The exhausted branch showed increasingly higher expression of *Cxcr6*, *Ccl5* and *Nkg7* as the trajectory progressed towards the end point. The two genes *Ifngr1* and *Lgals3* were transiently upregulated in the exhausted trajectory exclusively, but decreased towards the end point (Supp. Fig. S8).

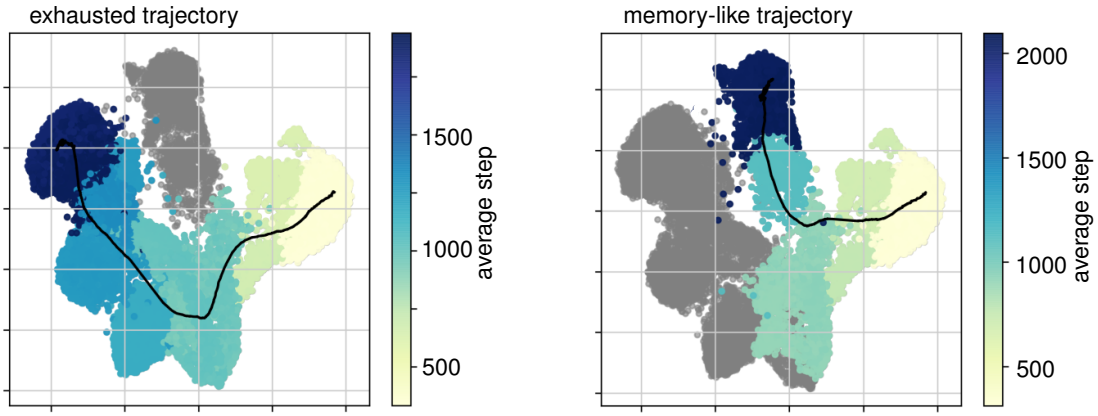


Figure 5: Lineage trajectories after simulations towards the two endpoint populations using cytopath. The average simulation steps to arrive at a cell are color coded per cluster. The coordinates of the average trajectory after alignment is depicted as black line. The shown average trajectories are based on 2000 simulations per endpoint.

2.4 Branching point of differentiation towards the end points

Based on the simulated mean trajectories we wanted to identify a branching point at which the trajectories of cells moving towards the exhausted and the memory-like endpoints would diverge. This branching region would demarcate the point after which a cell would be committed to only one endpoint.

We therefore computed the ratio of the alignment score for each cell to the two average trajectories (Fig 6a). We observed that in the unstructured early region all cells had equal scores for both the exhausted and the memory-like fate. However, between average simulation step 800 and 1200, the cells started to be uniquely assigned to only one of the two trajectories. Interestingly, we observed a region along the differentiation trajectories where some cells aligned clearly to the exhausted trajectory, some to the memory-like trajectory and some to both. It seemed that this region was where bifurcation took place. Using a threshold on the alignment score we assigned all cells to either the exhausted branch (blue), memory-like branch (orange) or pre-committed branch (green, Fig. 6).

To determine which biological time-point would correspond to the identified branch region, we investigated the branch composition of the four samples (Fig. 6c). The earliest samples were almost exclusively composed of pre-committed cells, whereas the day 7 sample contained already a large fraction of lineage-committed cells. We reasoned that bifurcation must take place between day 5 and 6 after infection.

Differential gene expression between the branching region and its immediately adjacent committed branches did not reveal any clear transcriptional signatures, that would precede or succeed bifurcation. However the trajectory assignment clearly implied that after the branching region, cells were fully committed to their lineage. Since the alignment score also considers the velocity direction of each cell, alignment to only one trajectory indicates that differentiation will take place along this path. The abrupt increase in the alignment score after the branch region suggests that cells beyond the bifurcation do exhibit coherent and significant velocity away from the bifurcation and this process is unlikely to be reversible.

2.5 Identification of marker genes predictive for different developmental fates

To further demarcate markers that would specify the branching point, we searched for genes that were characteristic for this bifurcation - either being expressed at divergent levels before, at, or after the branching point. We trained a classifier to predict the assigned branch label from the transcriptional profile of each cell (see Method section). If a gene was expressed in one branch but not the other, it was considered relevant for the prediction. To validate the

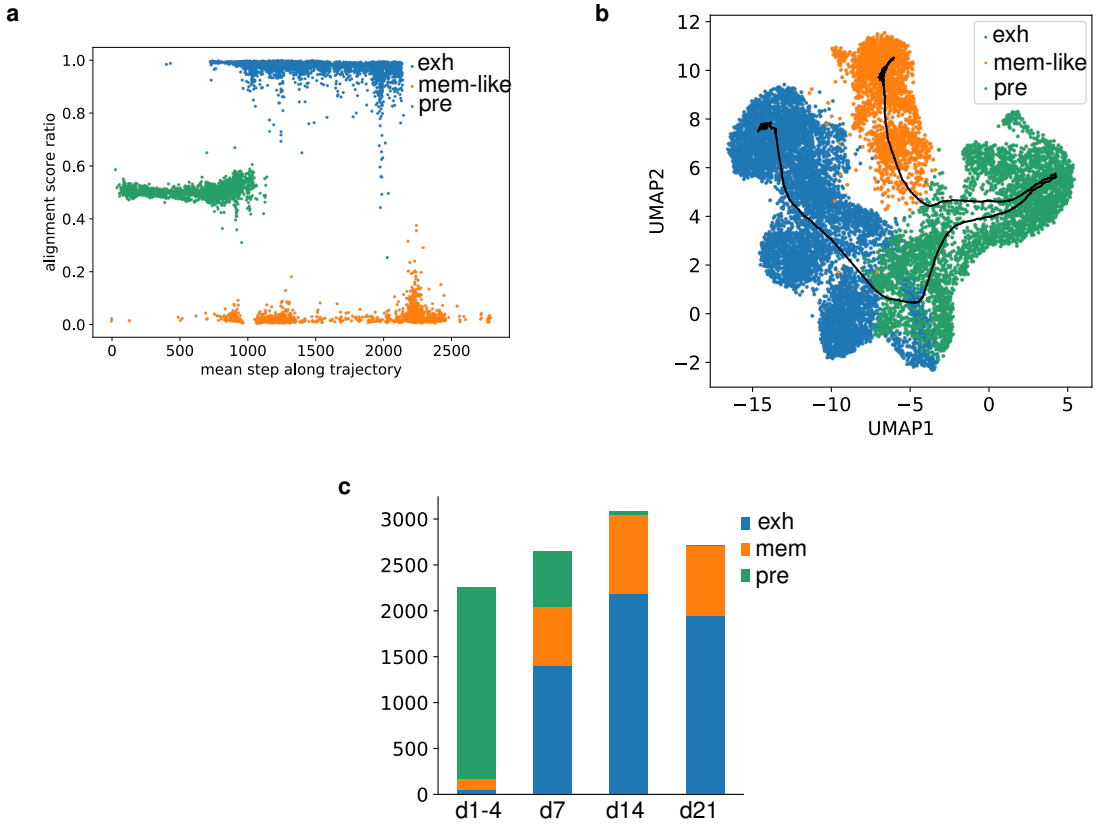


Figure 6: Cells are aligned to the branches of the two fates exhausted (exh), memory-like (mem) or the pre-commited branch (pre).
(a) The two average trajectories (black) towards the exhausted (blue) and the memory-like fate (orange) with an pre-commited shared part (green).
(b) Alignment score ratio along the cytopath simulation steps. High values indicated preferential alignment with the exhaustion trajectory, low values alignment with the memory-like trajectory.
(c) sample time-point composition with respect to branch assignment.

identified branching markers later on a protein level using flow cytometry, we restricted the transcriptional input to genes transcribing proteins with validated antibodies for staining. The result of this analysis variant allowed us to sort and experimentally analyze the differentiation potential of pre-committed and committed cell states in later validation experiments.

The classifier identified 12 genes that were most relevant to distinguish the three branches (Fig. 7). These included already described markers of the memory-like population, such as *Il7r* (IL7R), *Tcf7* (TCF1) and *Slamf6* (Ly108) [3, 4], but revealed also potential new candidates *Icos* (CD278), *Ly6e* (SCA-2) and *Itgb1* (CD29) that are highly expressed in cells from the memory-like branch. Markers relevant for the exhausted branch contained *Cxcr6* (CXCR6), *Ifngr1* (IFNGR1) and *Cd3g* (CD3G) but also *Selplg* (CD162), of which only CXCR6 has been linked previously to exhaustion [16]. The pre-committed branch showed high expression of *Gzmb* (Granzyme B) and Mif (MIF) both of which were expressed at lower levels in the other lineages.

Predicting the branch labels using only these markers resulted in good prediction accuracy (0.85). Additionally, using only these genes as input to UMAP showed a good separation into the three branches (Supp. Fig. S9).

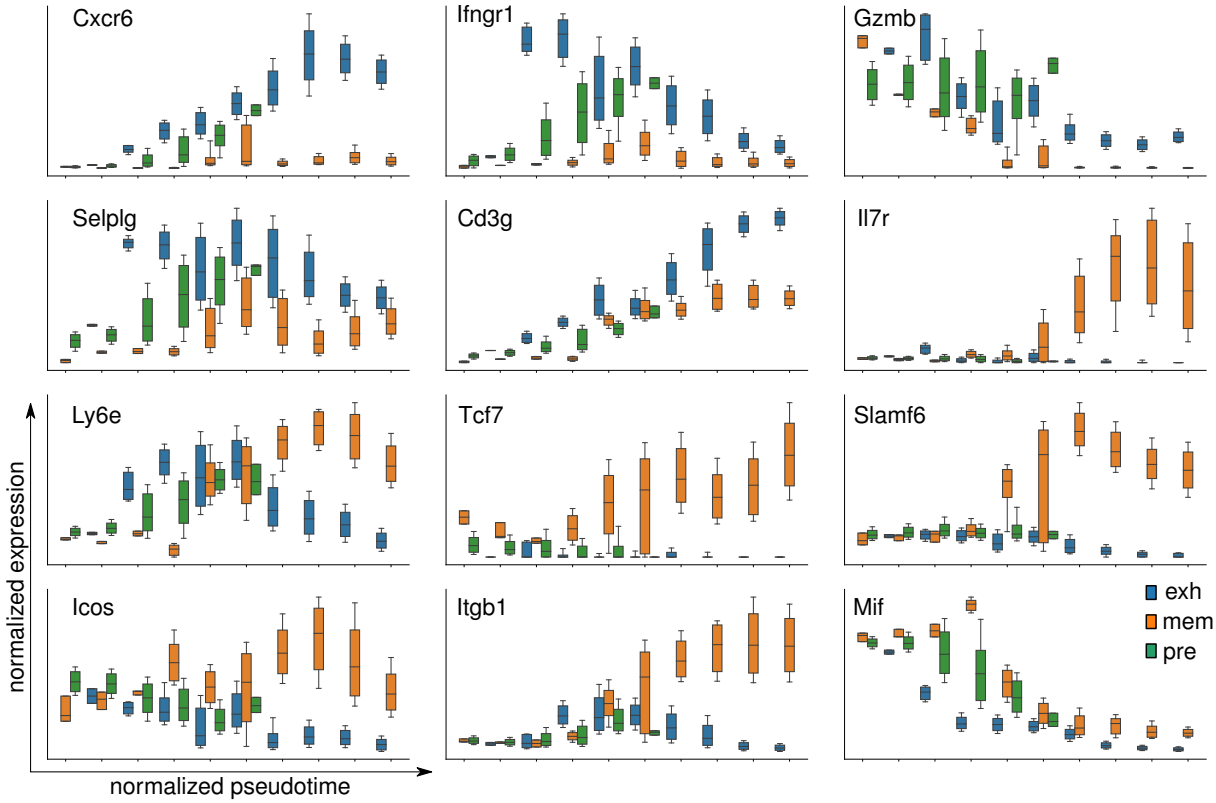


Figure 7: The 12 genes identified by the classifier to predict the three branch labels. Boxplot of normalized expression are shown along normalized pseudotime from cytopath. Color indicates the branch the cells were assigned to based on the Cytopath alignment score, exhausted (exh), memory-like (mem) or the pre-commited branch (pre).

2.6 Experimental transfer of cells with a presumably pre-committed or committed phenotype show distinct differentiation potential

To experimentally validate our branch classification, we set out to identify and later sort LCMV-specific CD8 T cells with phenotypes indicative of a pre-committed state, a committed state towards the exhausted endpoint or the memory-like endpoint. For validation of phenotypic markers classifying these three cell states, we tested all 12 identified markers for their ability to discriminate early populations to then test their fate potential. Specifically, we transferred naïve P14 cells into naïve C57BL/6 mice, followed by chronic LCMV infection. 5 days post infection (when representatives of all three populations of interest had formed, i.e. the uncommitted cells and the committed exhausted and memory-like cells), P14 cells were analyzed according to the markers identified using the classifier (data not shown). We first identified protein markers that showed high variance at the branching time-point. We determined CXCR6 and TCF1 as the prime candidates for sorting the branches into pre-committed (CXCR6$^-$ TCF1$^-$), memory-like (CXCR6$^-$ TCF1$^+$) and exhausted (CXCR6$^+$ TCF1$^-$) cells.

Having identified markers that allowed to distinguish between uncommitted and committed cells into the exhausted and memory-like branch, we used these markers to isolate the respective populations 5 days into chronic LCMV infection. Specifically, we transferred naïve P14 T cells expressing GFP under the TCF1 promoter into C57BL/6 mice and infected them with high dose LCMV Clone-13 (Fig. 8a). At dpi 5 we sorted P14 cells from the three branches according to expression of CXCR6 and TCF1 (detected by GFP) and transferred them into infection-matched hosts (Fig. 8b). At one week after transfer (at dpi 12 from the initial infection), we analyzed the progeny of cells originating from the three branches in the spleen (Fig. 8c). We observed that cells recovered after transfer of exhausted cells into

infection-matched recipients retained their exhausted phenotype. Cells recovered after transfer of the memory-like branch exhibited phenotypes of both exhausted and memory-like cells, confirming previous results of differentiation from memory-like into exhausted cells [4] but contradicting our finding of a memory-like endpoint. Recovered cells after transfer of pre-committed cells exhibited both a memory-like or an exhausted phenotype, confirming their differentiation potential into both memory-like and exhausted cells. However, there was a strong bias towards differentiation along the exhaustion branch, which might be explained by much more extensive proliferation of these cells compared to memory-like cells.

2.7 Differentiation transitions are driven by antigenic TCR stimulation

Since our RNA velocity and trajectory analysis had revealed the memory-like cells as an end-point of differentiation, we speculated that their unexpected differentiation into terminally exhausted cell states after adoptive transfer is triggered by external cues that drastically changed their state. We investigated the possibility of TCR stimulation as such a cue by transferring cells belonging to the three branches (isolated from mice at 5 dpi following adoptive transfer of P14 cells and induction of chronic infection with Clone-13) into hosts infected with a LCMV Clone-13 strain that expresses a variant of the gp33 peptide that is not recognized by P14 cells (Fig. 8a).

We recovered cells from spleen and lymph nodes and analyzed their phenotype based on CXCR6 and Ly108 expression (Fig. 8d). The recovered cells after transfer of the exhausted branch again retained their exhausted phenotype. Surprisingly, we recovered much fewer cells with an exhausted phenotype after transfer of memory-like cells and a major fraction retained their memory-like phenotype, indicating that further differentation was largely halted in the absence of antigen. The transfer of pre-committed cells resulted in recovery of cells with a largely undifferentiated phenotype of neither terminally exhausted nor memory-like, largely retaining their pre-committed state. These results clearly pointed towards TCR stimulation being a major driver of differentiation during chronic infection for both the pre-committed state and for further differentiation of the memory-like state into fully exhausted cells.

3 Discussion

We analyzed differentiation trajectories of virus-specific $CD8^+$ T cells during chronic LCMV infection using scRNAseq time-series data from four different time points covering activation, peak, contraction and late phase of the response. The time-resolved traversal of the transcriptional landscape revealed a continuous and bifurcating process, with early activated cells at the beginning and both terminally exhausted as well as memory-like cells at the end of this process. We observed cells with an exhausted phenotype and an effector-like cell population as transient states during early and late stages of this process, respectively. We further observed for all time points a population of cells with transcriptional profiles of proliferation.

Applying RNA velocity analysis to our single-cell transcriptional data allowed us to estimate transitions between the cell states during the progressing immune response. We computed most likely end and start regions and identified two major differentiation paths leading to the exhausted population and the memory-like population respectively. At the early time-points until about day 5, the two average trajectories were nearly indistinguishable, but then diverged increasingly towards their respective endpoints.

We identified a branching region in the early stages of infection before day 5 post infection. Before this branching point, pre-committed cells would still have the potential to differentiate into both the exhausted and the memory-like phenotype, whereas cells that had passed the branching point would be destined to differentiate into the endpoint they had committed to.

Although it is conceivable that factors that were not revealed by transcriptional analysis might be involved to already pre-determine cell fate during early activation [17] and regulate differentiation, this is not evident on a transcriptional and protein expression level, where branching of the two main fates manifested itself around day 5 post infection.

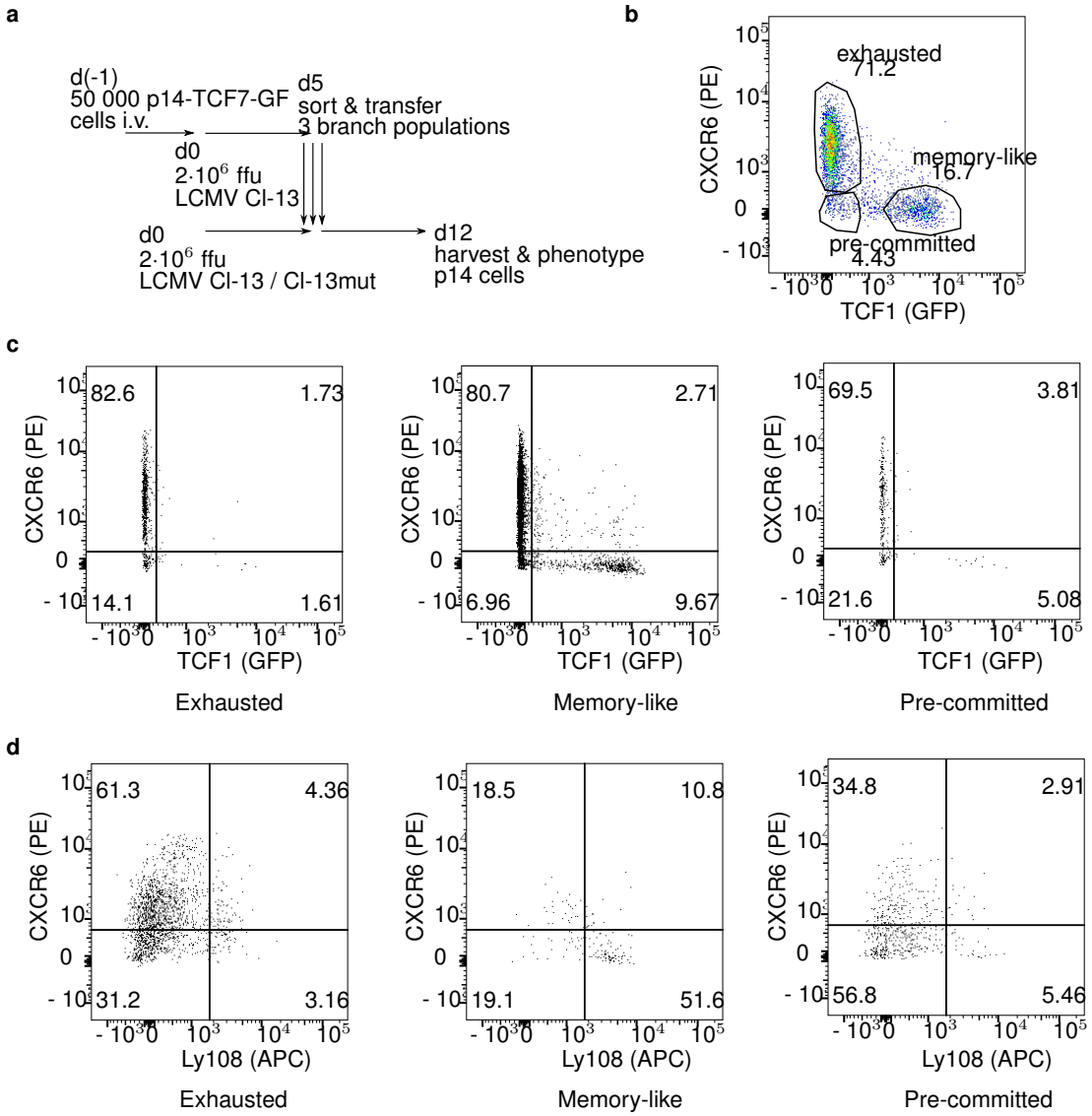


Figure 8: (a) The three P14 branch populations were isolated at 5 dpi from high-dose Clone-13 infected mice (that had been transferred with naive P14 cells prior to infection) and transferred into infection matched hosts and their phenotype was assessed 7 days post transfer.
(b) flow cytometry gates used to sort exhausted, memory-like and pre-committed branch
(c) phenotype of the recovered P14 cells at 12 dpi from spleens after high-dose Clone-13 infection and transfer of either exhausted, memory-like or pre-committed cell populations isolated at 5 dpi. Cells are gated on P14 cells.
(d) phenotype of the recovered cells at 12 dpi from spleens after transfer into hosts infected with Clone-13 P14 escape mutant. Naïve P14 cells were first transferred into naive C57BL6 mice, followed by Clone-13 infection. At 5 dpi, exhausted, memory-like and pre-committed populations were sorted and adoptively transferred into infection matched hosts with Clone-13 escape mutant. Recovered P14 cells are shown.

We derived a small set of gene markers that separate cells into the pre-committed (TCF1^{neg} CXCR6^{neg}), exhausted (TCF1^{neg} CXCR6^{hi}) and memory-like branch (TCF1^{hi} CXCR6^{neg}) by a classification model. Adoptive transfer of cells sorted according to these markers, and thereby likely belonging to the three branches, at day five post chronic

LCMV infection into infection-matched new hosts, confirmed the plasticity of pre-committed cells to acquire both exhausted as well as memory-like phenotype. Conversely, progeny from committed exhausted cells largely retained their phenotypes. Although, our velocity and trajectory analysis suggested the memory-like cells to represent a developmental end-point, we still recovered both memory-like and exhausted cells after transfer of cells from the memory-branch. These transitions might be rare (and fast) events in the integral setting of an infected mouse, and thus not be represented in the scRNAseq data. Additionally, i.v. adoptive transfer of memory-like cells into circulation might expose them to different antigenic burden compared to their natural niches, thereby accelerating a differentation process.

Previously published work used scRNAseq to study CD8 T cell differentiation during chronic infection, isolating virus specific T cells at different time-points [4, 5, 10, 18]. They made use of dimensionality reduction and computational tools for trajectory inference. Although, these studies used fewer time-points and the lineage inference included one time-point only. All these studies consistently described the memory-like and the terminally exhausted cell state, which we also identified. Some studies [5, 18] additionally described an effector-like CX3CR1^{+} population arising late during the infection, which we also found in our late samples from 14 & 21 dpi.

Two studies [4, 5] investigated plasticity and differentiation of the memory-like cells computationally by lineage inference on a single time-point and through adoptive transfer experiments, concluding that memory-like cells partially maintain their phenotype and can give rise terminally exhausted and effector-like cells. Our lineage inference across multiple time-points suggests that the exhausted and the memory-like lineages are separate. We could not exclude, that we missed certain cellular states in our data set, since we only studied CD8 cells isolated from the spleen of infected animals. We were unable study developmental stages that are spatially restricted to lymph nodes or specific anatomical regions within secondary lymphoid organs, or specific differentiation processes that are restricted to non-lymphoid organs [16]. Additionally, if cell state transitions are rare or fast, it is unlikely that we would capture them in our snapshot analysis.

Our adoptive transfer experiments of memory-like cells revealed extensive transitions to terminally exhausted states, which our lineage inference did not detect. However, our transfer experiments into Clone-13 P14 escape mutant infected hosts suggested that theses transitions were strictly dependent on antigenic TCR stimulation. This could imply that we did not observe these transitions in our scRNAseq data because memory-like cells receive little TCR stimulation from their microenvironment in a natural setting. Isolation of memory-like cells by removing them from their niche and transferring them via intravenous injections could expose them to excessive antigen and trigger differentiation towards terminally-exhausted cells.

Chen et al. [4] studied early bifurcation events towards either an effector state or TCF^{+} precursor state. However, the population they termed “effector” cells already expressed many co-inhibitory receptors like our exhausted group from 7 dpi. They found this early effector cells to be very short-lived and disappear between 8 and 12 dpi. Although our data suggested a similar bifurcation into effector and memory-like lineage, our exhausted trajectory placed the early exhausted cells as an intermediate state on the differentiation towards terminally exhausted states. Since Chen et al. used KLRG1 to identify their effector population, the disappearance of these cells could be explained by down-regulation of *Klrg1* during differentiation towards terminal exhaustion.

The velocity based endpoint analysis did not reveal either the early exhausted or the effector-like states to compose stable end-points, but that all those states differentiated into terminally exhausted cells. The velocity transitions did show some flow out of these populations though, which could indicate migration out of the tissue or apoptosis of these cells, although we did not find strong apoptotic signatures in our data. Considering, that apoptotic cells are cleared very fast by the phagocytes, clearance of apoptotic cells might be too fast to capture.

Our classification analysis of the branch point revealed a set of genes that discriminates between the three branches. Even though, our validation experiments confirmed that CXCR6 and TCF1 expression patterns capture the differentiation potential of CD8 T cells early during the infection, other identified genes might still be relevant in shaping this bifurcation. Both genes *Ifngr1* and *Selplg* are transiently upregulated around the birfucation point, which could

imply some influence on cell fate. Considering, that withdrawing T cells from antigen stimulation practically halted differentiation of pre-committed cells and maintained their state, points towards a significant role of TCR stimulation and signaling in this bifurcation and decision process.

This work provides additional insights into the differentiation process of CD8 T cells using a combined approach of scRNAseq analysis, computational trajectory inference and adoptive transfer experiments. Our study revealed an early bifurcation event, that shaped the differentiation fate during the course of a chronic infection and additionally highlights TCR stimulation as a significant driver of this differentiation.

4 Material & Methods

Infections and cell isolation

Mice Wild-type male C57BL/6J mice were purchased from Janvier Elevage. Nr4a1-GFP mice expressing GFP under the control of the NUR77 promoter [19], P14 transgenic (CD45.1) mice expressing a TCR specific for LCMV peptide gp33–41 [20] and TCF1-GFP mice expressing GFP under the control of the Tcf7 promoter [3] were housed and bred under specific pathogen–free conditions at the ETH Phenomics Center Hönggerberg. All mice used in experiments were between 6–16 weeks of age. P14-Nr4a1-GFP mice were generated by crossing Nr4a1-GFP mice to P14 mice. P14-TCF1-GFP mice were generated by crossing TCF1-GFP mice to P14 mice. All animal experiments were conducted according to the Swiss federal regulations and were approved by the Cantonal Veterinary Office of Zürich (Animal experimentation permissions 147/2014, 115/2017).

Virus LCMV clone 13 [21] was propagated on baby hamster kidney 21 cells. LCMV clone 13 P14 escape mutant [2] was propagated on MC57G cells. Viral titers of virus stocks were determined as described previously [22].

Infection 10^4 transgenic cells (P14, P14-TCF1-GFP or P14-Nr4a1-GFP) were adoptively transferred 1 day prior LCMV clone 13 intravenous (IV) infection with 2 x 10^6 ffu/mouse. For isolation at 1, 2, 3, 4 days post-infection, 10^5 P14-Nr4a1-GFP cells were transferred.

Cell isolation from tissues After 1, 2, 3, 4, 7, 14 and 21 days of chronic infection, mice were sacrificed with carbon dioxide and organs (spleen, lymph nodes) were isolated. Spleens and lymph nodes were mashed through 70 µm filters with a syringe (1 mL) plunger. Cell suspensions were filtered (70 µm) and treated with ammonium-chloride-potassium buffer (150 mM NH4Cl, 10 nM KHCO3, 0.1 mM EDTA in water) to lyse erythrocytes for 5 min at room temperature.

Cell sorting Spleen samples were depleted of CD4 and B cells by incubating splenocyte suspensions in enrichment buffer (PBS, 1%FCS, 2 mM EDTA) with biotinylated α-CD4 and α-B220 antibodies at room temperature for 20 min, followed by incubation with streptavidin-conjugated beads (Mojo, Biolegend) (4%) for 5 min at room temperature. Cells were then placed on a magnetic separator (StemCell) for 10 min at room temperature, followed by collection of supernatant. For scRNAseq samples, cell suspensions of spleens isolated from five mice were pooled in samples from day 7, 14 and 21 post infection and from three mice for samples from early timepoints at day 1, 2, 3 and 4 order to ensure the samples were representative of a population. All samples from day 1 to 4 were pooled for sorting and sequencing due to the low frequency of P14 cells in these samples. Enriched samples from the spleen or cell suspensions from lymph nodes were stained with α-CD8-PerCP, α-CD45.1-APC and fixable lifedead marker to sort live P14 cells (ARIA cell sorter, BD Biosciences).

Single-cell RNA sequencing & analysis

Sorted P14 cells from different time-points were washed and resuspended in 0.04% BSA. The single cell sequencing was performed at the Functional Genomics Center Zurich. The cell lysis and RNA capture was performed according

to the 10XGenomics protocol (Single Cell 3' v2 chemistry). The cDNA libraries were generated according to the manufacturer's protocol (Illumina) and further sequenced (paired-end) with NovaSeq technology (Illumina). The transcripts were mapped with 10Xgenomics CellRanger pipeline (version 2.0.2).

Pre-processing & Normalization Read counts were realigned and sorted for spliced and unspliced counts using the analysis pipeline from velocyto [11]. Contaminating other cell types were removed from the dataset based on outliers in diffusion components. Reads were filtered and normalized according to the Zheng recipe [12] of the scanpy analysis pipeline [23] retaining 5000 highly variable genes. Louvain clustering and UMAP projection were computed using standard parameters, using the first 50 principle components.

RNA velocity RNA velocity uses the relative abundance in reads of un-spliced to spliced mRNA to infer the future state of a particular cell, with a high ratio of un-spliced / spliced mRNA being indicative of recent gene activation, a balanced ratio of un-spliced / spliced mRNA being indicative of gene expression equilibrium, and a low rate of un-spliced / spliced mRNA being indicative for terminating gene expression. Integrating the expression levels of the corresponding genes in neighboring cells allows computing likely transitions between different cellular states in our data set, revealing a vector field demarcating likely transitions.

Scvelo [24] was used to estimate RNA velocity and infer transition probabilities between cells. The transition probabilities were used to construct a Markov process. Inference of RNA velocity relies an a set of assumptions that can be adjusted through several parameters before analysis. The type of pre-processing used, the number of principle components used as well as the neighborhood size for imputation may influence the resulting transition matrix. We conducted an extensive assessment of a many parameter combinations to validate the stability of our RNA velocity analysis. Comparing the equilibrium distributions across 574 parameter sets revealed that the global structure of the transition matrix was quite stable. The parameter set we have chosen for further analysis produced results found in an overwhelming majority of tested sets (Supp. Fig. S5).

We estimated gene moments using neighborhood connectivities using 50 principle components, 30 neighbors with the UMAP method. Velocity was inferred using "stochastic" mode.

Cytopath Cytopath is RNA velocity based lineage inference tool [15]. We applied scvelo's terminal states routine to compute equilibrium distributions of the forward and backward Markov process, excluding self transitions. Regions with terminal state probability higher than 0.3 were identified and the louvain clusters corresponding to these regions used as start and endpoints. Markov simulations were initialized at the start points and simulated for a maximum of 2000 steps or until they reached the endpoint. We simulated 2000 trajectories from random cell states in the starting region. All simulated paths were aligned to average trajectories from startpoint to each endpoint using dynamic time warping. Neighboring cells (2000 nearest neighbors) were aligned to the trajectories using an alignment score, which was computed based on distance and cosine distance between the cell's velocity and the direction of the trajectory. Cells that aligned to only one trajectory were assigned to the exhausted or memory-like branch, respectively. Cells at the beginning of the infection that aligned to both trajectories were assigned to the pre-commited branch.

All cells were assigned exhausted, memory-like or pre-commited fate according to their alignment score to the trajectories. Gene expression profiles of 650 genes coding for antibody stainable proteins were then used to predict these labels using L1-penalized Logistic Regression. We used cross validation to identify the optimal L1-penalty that would give a reasonably small number of genes but still good prediction accuracy at C=0.1. The resulting prediction using 12 proteins still classified most cells correctly (accuracy: 0.85). These proteins were then stained on P14 cells from chronic infection at d5 to sort the branches and adoptively transfer them into infection matched hosts.

Adoptive transfer experiments

After 5 days of chronic infection, CD8 T cell enriched samples from the spleen or cell suspensions from lymph nodes were stained with α-CD8-PerCP, α-CD45.1-APC/FITC and α-Cxcr6-PE and α-Ly108-APC to sort P14 cells into the exhausted, memory-like and pre-commited populations. (ARIA cell sorter, BD Biosciences).

Sorted cells from exhausted (10^6 cells), memory-like ($2x10^5$ cells) and pre-committed ($5x10^4$ cells) populations were transferred via intravenous (IV) injection into infection matched hosts infected with either Clone-13 or Clone-13 P14esc mutant. Cells were recovered from from spleens of these mice 12 days post infection prior to phenotypic characterization.

Flow cytometry

Surface staining was performed at room temperature for 30 minutes in FACS buffer (2% FCS, 1% EDTA in PBS). LIVE/DEAD™ Fixable Near-IR Dead (Thermo Fisher) was used to discriminate alive from dead cells. Fluorophore-conjugated antibodies used for flow cytometry were purchased from BioLegend (Lucerna Chem AG, Luzern, Switzerland) (α-CD45.1 BV711 A20; α-CD45.1 APC A20; α-CD8 PerCP 53-6.7; α-CD8 BV395 53-6.7; α-PD-1 PE-BV605 29F.1A12; α-Cxcr6 PE SA051D1; α-Ly108 APC 330-AJ; α-CD8 BV395 53-6.7). Data was acquired LSR II Fortessa using Diva software (BD Biosciences, Allschwil, Switzerland) and analyzed in FlowJo (BD Biosciences, Allschwil, Switzerland). Gating and plotting was done using FlowJo (BD Bioscience, Allschwil, Switzerland).

Aknowledgement

We thank Prof. Roman Spörri for providing the Nr4a1-GFP mice. We thank the Pinschewer laboratory through the European Virus Archive (University of Basel) for providing the Clone-13 P14 escape mutant viral strain. We thank Franziska Wagen and Nathalie Oetiker for great technical support. We are grateful for the constructive input of the members of the Claassen, Oxenius, Joller and Sallusto Group during discussions and group meetings. Funding: This work was support by the ETH Zürich (grant no. ETH-39 14-2 to MC and AO) and Novartis.

Contributions

D.C., M.C., A.O. designed the experiments; D.C., I.S. carried out the experiments, D.C., R.G., M.C., A.O. analyzed the experiments; D.C., M.C., A.O. wrote the manuscript.

Data Availability

Sequencing Data is available on request from the authors.

References

[1] E. J. Wherry, S.-J. Ha, S. M. Kaech, W. N. Haining, S. Sarkar, V. Kalia, S. Subramaniam, J. N. Blattman, D. L. Barber, and R. Ahmed, "Molecular signature of cd8+ t cell exhaustion during chronic viral infection.," *Immunity*, vol. 27, pp. 670–684, oct 2007.

[2] D. T. Utzschneider, F. Alfei, P. Roelli, D. Barras, V. Chennupati, S. Darbre, M. Delorenzi, D. D. Pinschewer, and D. Zehn, "High antigen levels induce an exhausted phenotype in a chronic infection without impairing t cell expansion and survival," *The Journal of Experimental Medicine*, vol. 213, pp. jem.20150598–jem.20150598, jul 2016.

[3] D. T. Utzschneider, M. Charmoy, V. Chennupati, L. Pousse, D. P. Ferreira, S. Calderon-Copete, M. Danilo, F. Alfei, M. Hofmann, D. Wieland, S. Pradervand, R. Thimme, D. Zehn, and W. Held, "T cell factor 1-expressing memory-like cd8+ t cells sustain the immune response to chronic viral infections," *Immunity*, vol. 45, no. 2, pp. 415–427, 2016.

[4] Z. Chen, Z. Ji, S. F. Ngiow, S. Manne, Z. Cai, A. C. Huang, J. Johnson, R. P. Staupe, B. Bengsch, C. Xu, S. Yu, M. Kurachi, R. S. Herati, L. A. Vella, A. E. Baxter, J. E. Wu, O. Khan, J.-C. Beltra, J. R. Giles, E. Stelekati, L. M. McLane, C. W. Lau, X. Yang, S. L. Berger, G. Vahedi, H. Ji, and E. J. Wherry, "TCF-1-centered transcriptional network drives an effector versus exhausted CD8 t cell-fate decision," *Immunity*, oct 2019.

[5] R. Zander, D. Schauder, G. Xin, C. Nguyen, X. Wu, A. Zajac, and W. Cui, "CD4+ t cell help is required for the formation of a cytolytic CD8+ t cell subset that protects against chronic infection and cancer," *Immunity*, vol. 51, pp. 1028–1042.e4, dec 2019.

[6] B. C. Miller, D. R. Sen, R. A. Abosy, K. Bi, Y. V. Virkud, M. W. LaFleur, K. B. Yates, A. Lako, K. Felt, G. S. Naik, M. Manos, E. Gjini, J. R. Kuchroo, J. J. Ishizuka, J. L. Collier, G. K. Griffin, S. Maleri, D. E. Comstock, S. A. Weiss, F. D. Brown, A. Panda, M. D. Zimmer, R. T. Manguso, F. S. Hodi, S. J. Rodig, A. H. Sharpe, and W. N. Haining, "Subsets of exhausted CD8+ t cells differentially mediate tumor control and respond to checkpoint blockade," *Nature Immunology*, vol. 20, pp. 326–336, feb 2019.

[7] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. nov, pp. 2579–2605, 2008. Pagination: 27.

[8] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using UMAP," *Nature Biotechnology*, vol. 37, pp. 38–44, dec 2018.

[9] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys, "A comparison of single-cell trajectory inference methods," *Nature Biotechnology*, apr 2019.

[10] C. Yao, H.-W. Sun, N. E. Lacey, Y. Ji, E. A. Moseman, H.-Y. Shih, E. F. Heuston, M. Kirby, S. Anderson, J. Cheng, O. Khan, R. Handon, J. Reilley, J. Fioravanti, J. Hu, S. Gossa, E. J. Wherry, L. Gattinoni, D. B. McGavern, J. J. O'Shea, P. L. Schwartzberg, and T. Wu, "Single-cell RNA-seq reveals TOX as a key regulator of CD8+ t cell persistence in chronic infection," *Nature Immunology*, vol. 20, pp. 890–901, jun 2019.

[11] G. L. Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, and P. V. Kharchenko, "RNA velocity of single cells," *Nature*, vol. 560, pp. 494–498, aug 2018.

[12] 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," *Nature Communications*, vol. 8, jan 2017.

[13] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008, oct 2008.

[14] S. D. Blackburn, H. Shin, W. N. Haining, T. Zou, C. J. Workman, A. Polley, M. R. Betts, G. J. Freeman, D. A. A. Vignali, and E. J. Wherry, "Coregulation of CD8+ t cell exhaustion by multiple inhibitory receptors during chronic viral infection," *Nature Immunology*, vol. 10, pp. 29–37, nov 2008.

[15] D. Gupta & Cerletti, R. Gupta, G. Gut, and C. Manfred, "*cytopath:* simulation based inference of differentiation trajectories from rna velocity fields," *Unpublished*, 2020.

[16] I. Sandu, D. Cerletti, N. Oetiker, M. Borsa, F. Wagen, I. Spadafora, S. P. Welten, U. Stolz, A. Oxenius, and M. Claassen, "Landscape of exhausted virus-specific CD8 t cells in chronic LCMV infection," *Cell Reports*, vol. 32, p. 108078, aug 2020.

[17] J.-C. Beltra, S. Manne, M. S. Abdel-Hakeem, M. Kurachi, J. R. Giles, Z. Chen, V. Casella, S. F. Ngiow, O. Khan, Y. J. Huang, P. Yan, K. Nzingha, W. Xu, R. K. Amaravadi, X. Xu, G. C. Karakousis, T. C. Mitchell, L. M. Schuchter, A. C. Huang, and E. J. Wherry, "Developmental relationships of four exhausted CD8+ t cell subsets reveals underlying transcriptional and epigenetic landscape control mechanisms," *Immunity*, vol. 52, pp. 825–841.e8, may 2020.

[18] S. Raju, Y. Xia, B. Daniel, K. E. Yost, E. Bradshaw, E. Tonc, D. J. Verbaro, A. T. Satpathy, and T. Egawa, "Latent plasticity of effector-like exhausted CD8 t cells contributes to memory responses," *BiorXive*, feb 2020.

[19] A. E. Moran, K. L. Holzapfel, Y. Xing, N. R. Cunningham, J. S. Maltzman, J. Punt, and K. A. Hogquist, "T cell receptor signal strength in tregand iNKT cell development demonstrated by a novel fluorescent reporter mouse," *The Journal of Experimental Medicine*, vol. 208, pp. 1279–1289, may 2011.

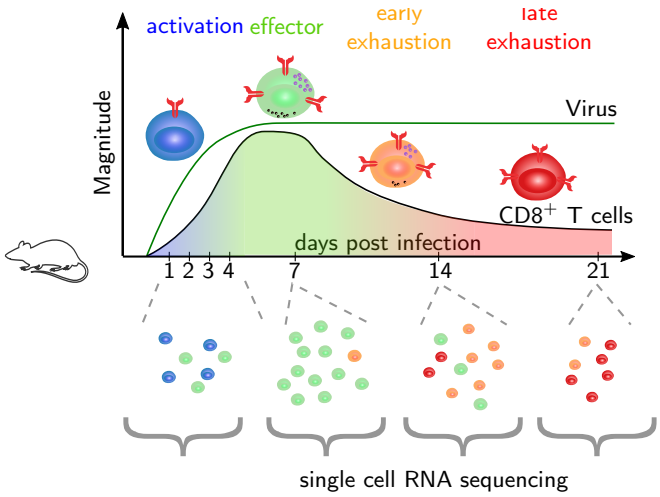
[20] H. Pircher, K. Bürki, R. Lang, H. Hengartner, and R. M. Zinkernagel, "Tolerance induction in double specific t-cell receptor transgenic mice varies with antigen," *Nature*, vol. 342, pp. 559–561, nov 1989.

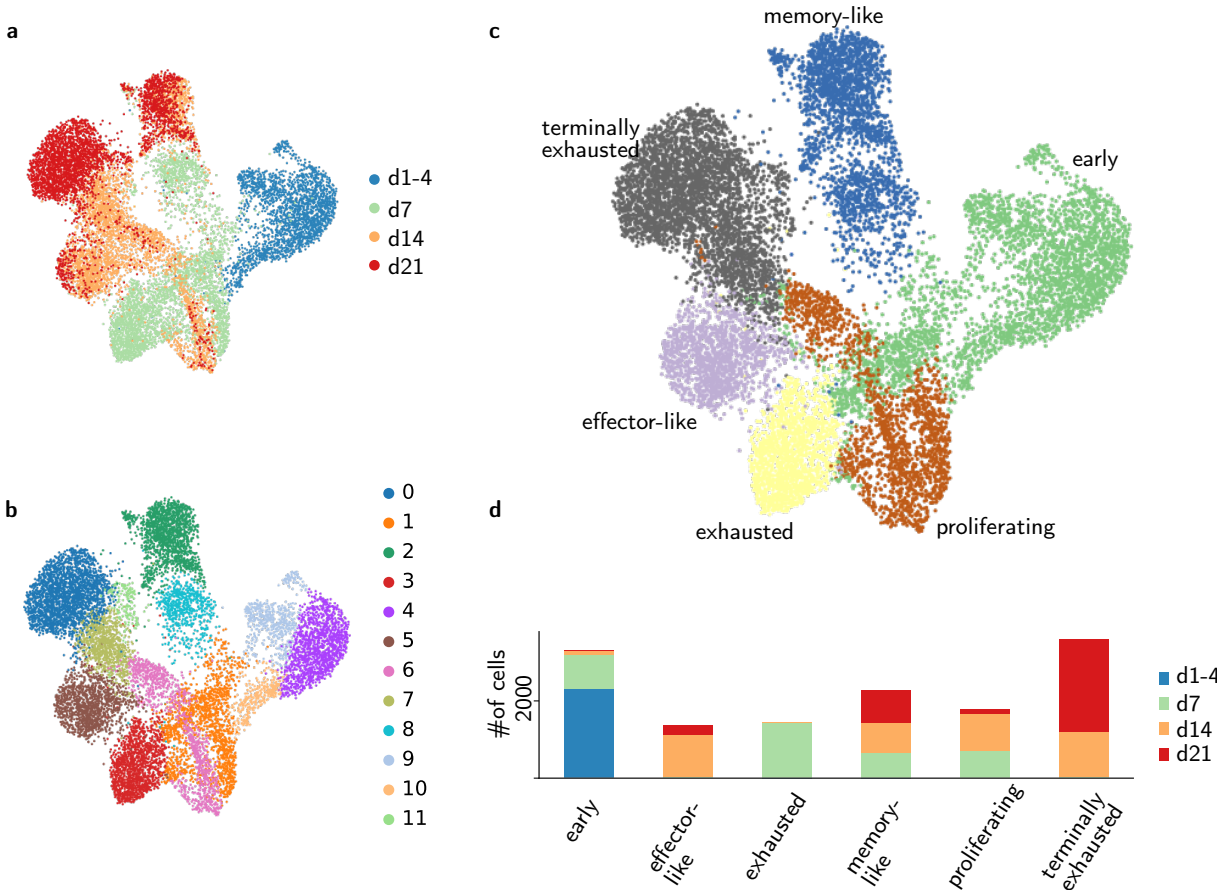
[21] R. Ahmed, A. Salmi, L. D. Butler, J. M. Chiller, and M. B. Oldstone, "Selection of genetic variants of lymphocytic choriomeningitis virus in spleens of persistently infected mice. role in suppression of cytotoxic t lymphocyte response and viral persistence.," *Journal of Experimental Medicine*, vol. 160, pp. 521–540, aug 1984.

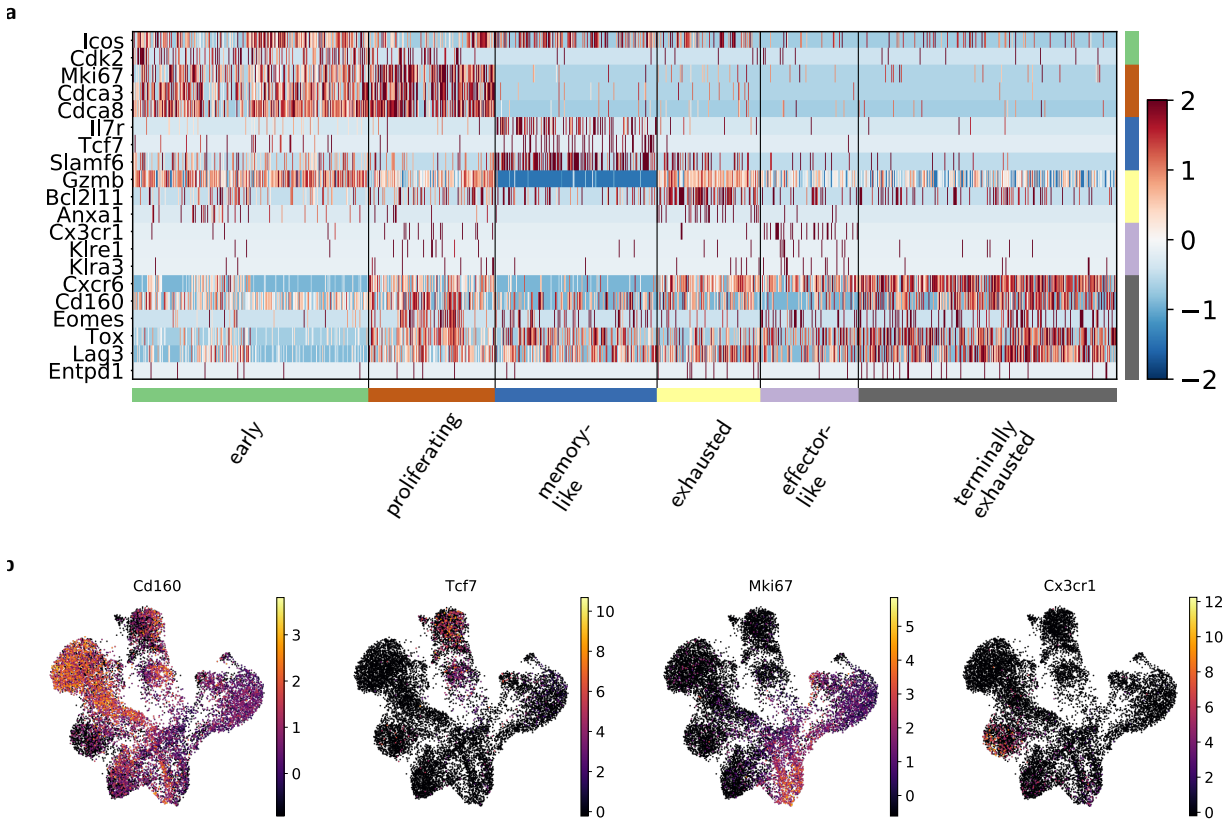
[22] M. Battegay, S. Cooper, A. Althage, J. Bänziger, H. Hengartner, and R. M. Zinkernagel, "Quantification of lymphocytic choriomeningitis virus with an immunological focus assay in 24- or 96-well plates," *Journal of Virological Methods*, vol. 33, pp. 191–198, jun 1991.

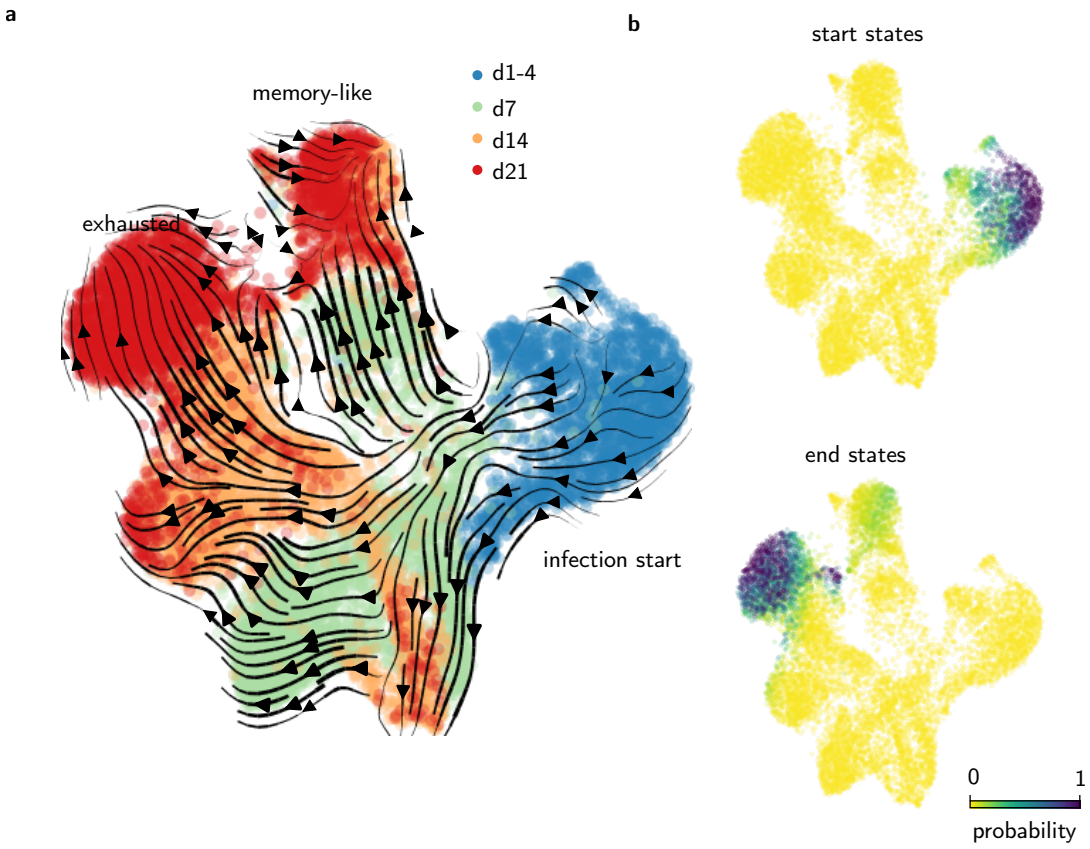
[23] F. A. Wolf, P. Angerer, and F. J. Theis, "SCANPY: large-scale single-cell gene expression data analysis," *Genome Biology*, vol. 19, feb 2018.

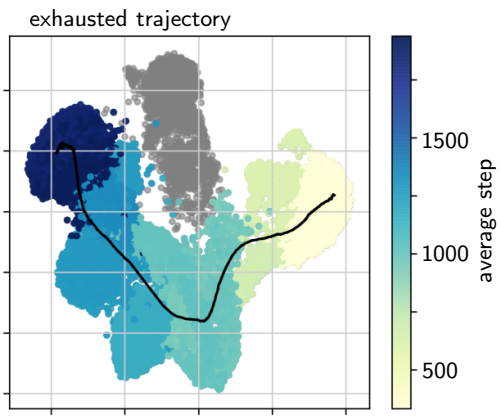
[24] V. Bergen, M. Lange, S. Peidli, F. A. Wolf, and F. J. Theis, "Generalizing RNA velocity to transient cell states through dynamical modeling," *Nature Biotechnology*, aug 2020.

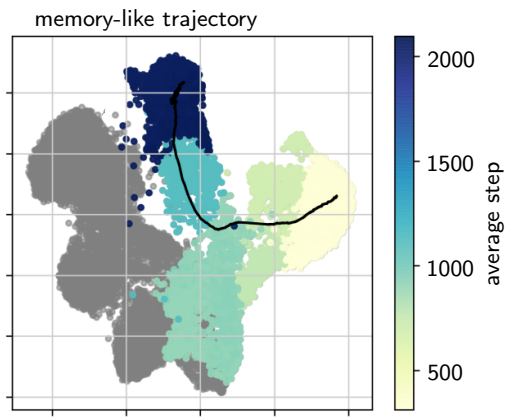

activation
effector
early exhaustion
late exhaustion
Magnitude
Virus
CD8+ T cells
days post infection
1 2 3 4
7
14
21
single cell RNA sequencing

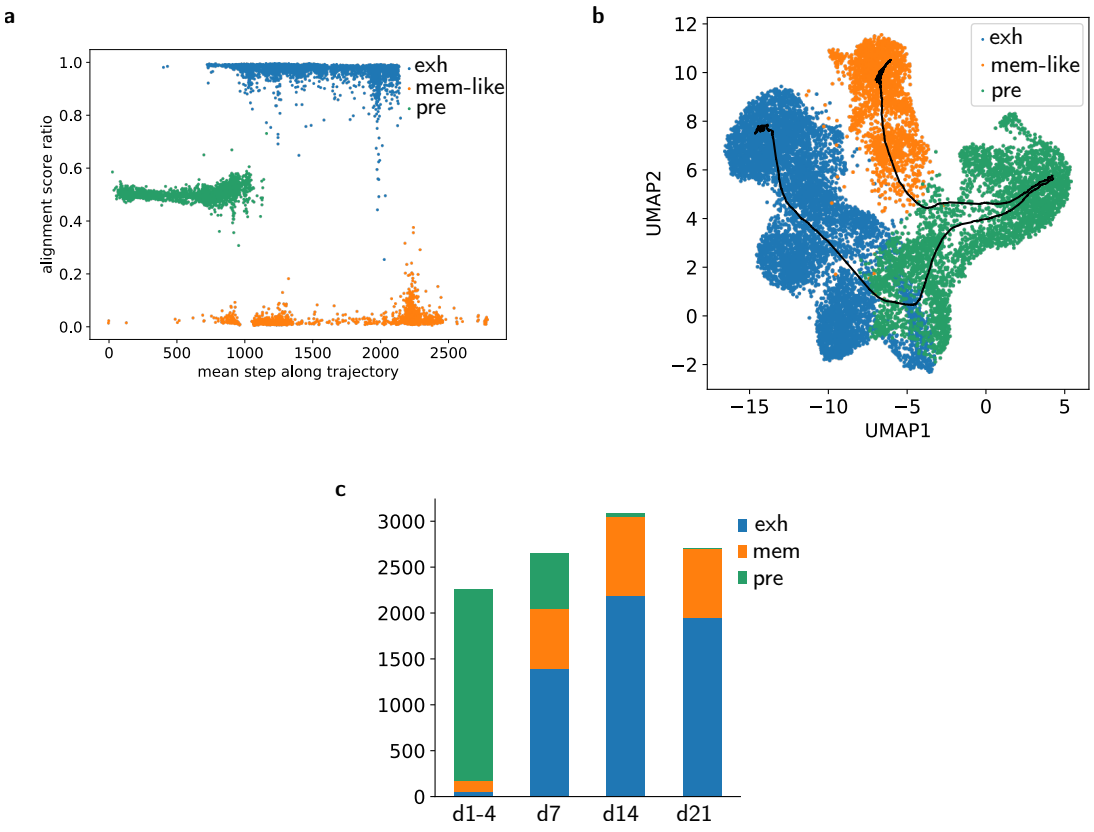

a
d1-4
d7
d14
d21
b
0
1
2
3
4
5
6
7
8
9
10
11
c
memory-like
terminally exhausted
early
effector-like
exhausted
proliferating
d
#of cells
2000
early
effector-like
exhausted
memory-like
proliferating
terminally exhausted
d1-4
d7
d14
d21

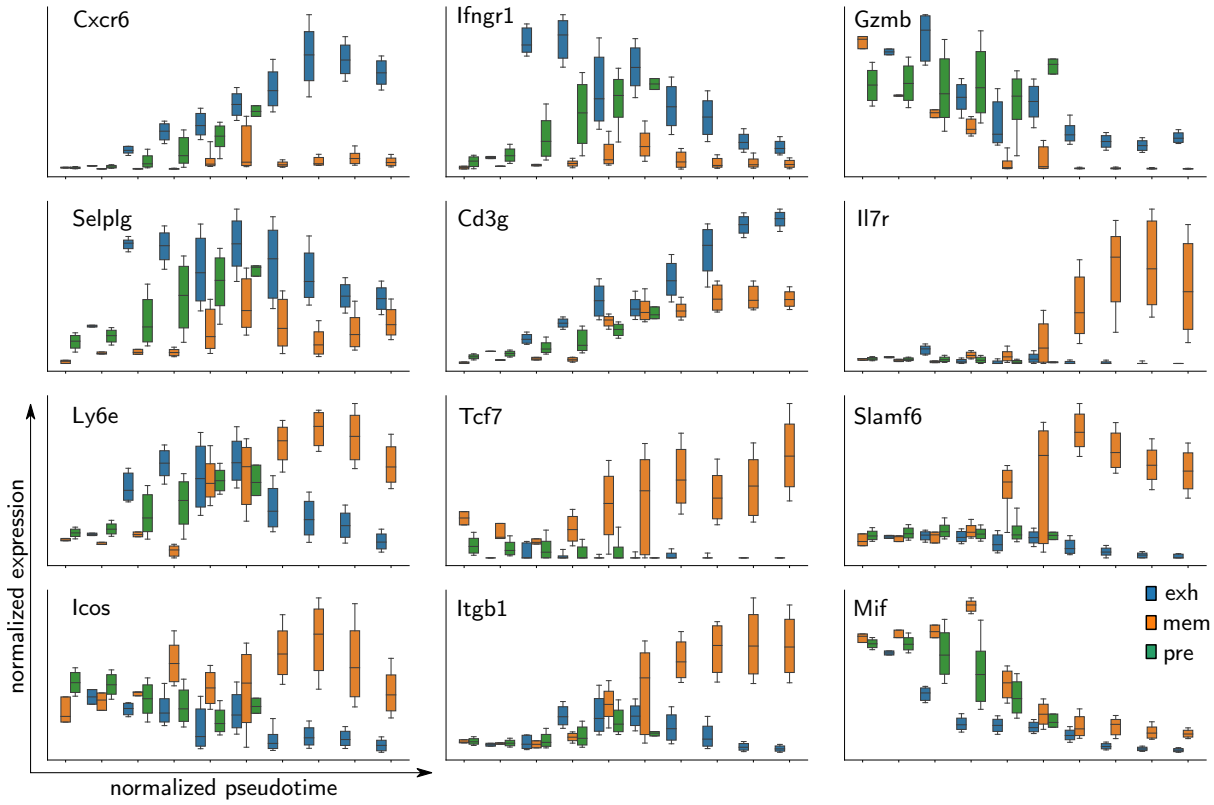

a
Icos
Cdk2
Mki67
Cdca3
Cdca8
Il7r
Tcf7
Slamf6
Gzmb
Bcl2l11
Anxa1
Cx3cr1
Klre1
Klra3
Cxcr6
Cd160
Eomes
Tox
Lag3
Entpd1
2
1
0
−1
−2
early
proliferating
memory-like
exhausted
effector-like
terminally exhausted
b
Cd160
Tcf7
Mki67
Cx3cr1

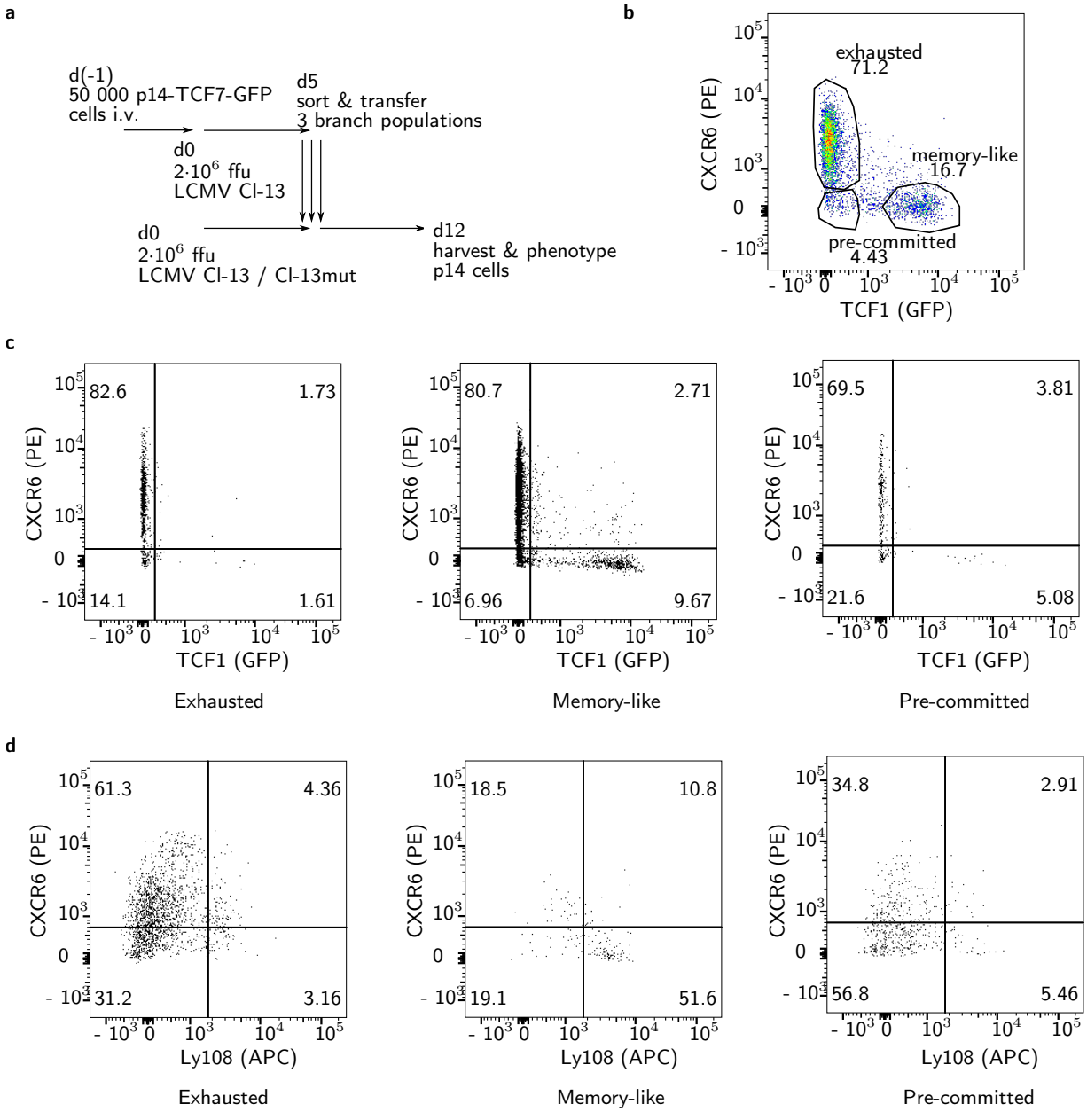

a
memory-like
exhausted
infection start
d1-4
d7
d14
d21
b
start states
end states
0
1
probability


exhausted trajectory
1500
1000
500
average step


memory-like trajectory
2000
1500
1000
500
average step


a
b
c
exh
mem-like
pre
alignment score ratio
mean step along trajectory
UMAP1
UMAP2
mem
d1-4
d7
d14
d21


Cxcr6
Ifngr1
Gzmb
Selplg
Cd3g
Il7r
Ly6e
Tcf7
Slamf6
Icos
Itgb1
Mif
exh
mem
pre
normalized expression
normalized pseudotime


a
d(-1)
50 000 p14-TCF7-GFP
cells i.v.
d0
2·10^6 ffu
LCMV Cl-13
d5
sort & transfer
3 branch populations
d0
2·10^6 ffu
LCMV Cl-13 / Cl-13mut
d12
harvest & phenotype
p14 cells
b
exhausted
71.2
memory-like
16.7
pre-committed
4.43
CXCR6 (PE)
TCF1 (GFP)
c
82.6
1.73
14.1
1.61
80.7
2.71
6.96
9.67
69.5
3.81
21.6
5.08
CXCR6 (PE)
TCF1 (GFP)
Exhausted
Memory-like
Pre-committed
d
61.3
4.36
31.2
3.16
18.5
10.8
19.1
51.6
34.8
2.91
56.8
5.46
CXCR6 (PE)
Ly108 (APC)
Exhausted
Memory-like
Pre-committed

Factorial state-space modelling for kinetic clustering and lineage inference

R. Gupta[1,2] , M. Claassen[1,2*]

1 University Hospital Tübingen, Faculty of Medicine, University of Tübingen, Tübingen, Germany.
2 Department of Computer Science, University of Tübingen, Tübingen, Germany.

*Corresponding author: manfred.claassen@med.uni-tuebingen.de

Abstract

Single-cell RNA sequencing (scRNAseq) protocols measure the abundance of expressed transcripts for single cells. Gene expression profiles of cells (cell-states) represent the functional properties of the cell and are used to cluster cell-states that have a common functional identity (cell-type). Standard clustering methods for scRNAseq data perform *hard* clustering based on KNN graphs. This approach implicitly assumes that variation among cell-states within a cluster does not correspond to changes in functional properties. Differentiation is a directed process of transitions between cell-types via gradual changes in cell-states over the course of the process. We propose a latent state-space Markov model that utilises cell-state transitions derived from RNA velocity to model differentiation as a sequence of latent state transitions and to perform *soft* kinetic clustering of cell-states that accommodates the transitional nature of cells in a differentiation process. We applied this model to the differentiation of Radial-glia cells into mature neurons and demonstrate the utility of our method in discriminating between functional and transitional cell-states.

Keywords single-cell RNA sequencing · Hidden Markov models · Lineage inference · Kinetic clustering

Introduction

Differentiation processes are typically represented as hierarchical transitions between functional cell-types. In contrast, it is widely assumed that scRNAseq data capture incremental shifts of gene expression along the differentiation process. Gene expression profiles of single cells measured using scRNAseq contain information about the functional properties of cell-states, and a differentiation model of incremental change of cell-states i.e. change in gene expression, is not consistent with the former model of discrete cell-type transitions. The latter model of differentiation allows for the existence of transitional cell-states that may be intermediate in both expression and function [1].

In standard scRNAseq analysis workflow, a k-nearest neighbours (KNN) graph of cell-states is used to cluster cells using community detection algorithms like Louvain or Leiden [2][3][4][5]. Clusters of cell-states are identified as cell-types based on marker gene expression and differential expression analysis. Grouping cells in this manner aids interpretability; however, since cell-states within a cluster are all assigned the same label, variation within the clusters is implicitly discarded as uninteresting noise in further analyses.

For scRNAseq data from differentiation processes, lineage inference models seek to impose a transitional relationship between cell-states. The models compute pseudotime, a score representing progress along the differentiation axis and also partition cell-states between multiple co-occurring lineages. Pseudotime estimation can model incremental shifts in gene expression, but due to the high sparsity of scRNAseq data, inferring the functional properties of individual cell-states from gene expression is challenging. Therefore, the clustering of cell-states remains important for interpretation.

Prior work has attempted to develop models of cell-type transitions, mainly by building minimum spanning trees among cell-state clusters [6][7] or by aggregating the connectivity between cell-states for each cluster [8]. The underlying clustering itself is not informed by cell-state transitions since a KNN graph is undirected and symmetric; in contrast, the directed signal obtained from RNA velocity enables the estimation of transition probabilities between cell-states. This information can be represented as a directed and asymmetric graph [9].

We introduce a latent state-space model based on cell-state transitions that enable the clustering of cells based on their transition dynamics. We refer to this form of clustering as kinetic clustering. Cell state transitions are assumed to be observed emissions from dynamics in a smaller latent state-space. The model allows for the probabilistic *soft* assignment of cells towards latent states and can be used to identify transitional cell-states. The dynamics of the differentiation process are captured by transitions between latent states. Multiple lineages can be further modelled with additional factors representing parallel sequences of latent state transitions [10].

Method

Model Input

RNA velocity of single cells can be used to estimate transition probabilities among cell-states [9][11]. For the set of measured cell-states (observed states) $O = \{o^{(1)}, \ldots, o^{(n)}\}$ and an initial probability vector $Y_0 = \{P(o^{(1)} \mid i = 0), \ldots, P(o^{(n)} \mid i = 0)\}$, the transition probability matrix $\mathbf{T}$ over observed states is used to simulate the differentiation process as a sequence of probability vectors $\mathbf{Y}$. The simulation is performed as,

$$Y_i = Y_{i-1} \cdot \mathbf{T} = Y_0 \cdot \mathbf{T}^i \tag{1}$$

The simulation is considered to have converged if $Y_i = Y_{i-1}$, i.e. when the simulation reaches the stationary state of the Markov chain with the transition matrix $\mathbf{T}$. The latent dynamic model considers the simulated process,

$$\mathbf{Y} = \{Y_i\} = \{\{P(o^{(1)} \mid i), \ldots, P(o^{(n)} \mid i)\}\} \qquad \forall i = 0, 1, \ldots, I \tag{2}$$

as input. In the following text, P_o is used to indicate a probability vector over states O such as $Y_i = P_o(o \mid i)$.

Model Specification

With latent states $S = \{s^{(1)}, \ldots, s^{(m)}\}$ and analogous to the simulation over observed states, we describe the dynamics over latent states as

$$\mathbf{Q} = \{Q_i\} = \{\{P(s^{(1)} \mid i), \ldots, P(s^{(m)} \mid i)\}\} \qquad \forall i = 0, 1, \ldots, I \tag{3}$$

Let $\mathbf{H}$ be the transition probability matrix over latent states S, then corresponding to the simulation (Eq. (1)), a Markov chain in the latent space has the form,

$$Q_i = Q_{i-1} \cdot \mathbf{H} = Q_0 \cdot \mathbf{H}^i \tag{4}$$

With the assumption of constant emission probabilities of observed states over the latent process $P(o \mid s, i) = P(o \mid s)$, we express Y_i as

$$Y_i = \sum_{s \in S} P_o(o, s \mid i) = \sum_{s \in S} P_o(o \mid s) P(s \mid i) \tag{5}$$

and due to Eq. (4):

$$Y_i = \sum_{s \in S} P_o(o \mid s) \cdot Q_i = \sum_{s \in S} P_o(o \mid s) \cdot (Q_0 \cdot \mathbf{H}^i) \tag{6}$$

Lineages L are modelled as independent Markov chains in the latent space. Furthermore, restricting the lineages to a common latent state-space $P(o \mid s, l) = P(o \mid s)$,

$$Y_i = \sum_{l \in L} P(l) \sum_{s \in S} P_o(o \mid s) \cdot (Q_0^{(l)} \cdot \mathbf{H}^{(\mathbf{l})}) \tag{7}$$

where $\mathbf{H}^{(\mathbf{l})}$ is the latent state transition probability matrix for lineage $l \in L$ and,

$$\mathbf{Q}^{(\mathbf{l})} = \{Q_i^{(l)}\} = \{\{P(s^{(1)} \mid i, l), \ldots, P(s^{(m)} \mid i, l)\}\} \qquad \forall i = 0, 1, \ldots, I \tag{8}$$

Model Training

The trainable parameters of the model are the conditional latent state transition probability matrices $\mathbf{H}$,

$$\mathbf{H}_{\mathbf{ij}}^{(\mathbf{l})} = P(s_i \mid s_j, l) \qquad \forall s \in S, l \in L \tag{9}$$

the emission probabilities $\mathbf{E}$,

$$E^{(s)} = P_o(o \mid s) \qquad \forall s \in S \tag{10}$$

the lineage weights W,

$$W = P(l) \qquad \forall l \in L \tag{11}$$

and the initial latent state probabilities $\mathbf{Q_0}$,

$$Q_0^{(l)} = P(s \mid i = 0, l) \qquad \forall s \in S, l \in L \tag{12}$$

Let $\hat{\mathbf{Y}}$ be the model estimate of $\mathbf{Y}$. The estimated sequence $\hat{\mathbf{Y}}$ is obtained as,

$$\hat{Y}_i = \sum_L \sum_S W Q_i^{(l)} \mathbf{E} \tag{13}$$

The parameters of the model are optimized by minimising the element-wise Kullback–Leibler (KL) divergence using gradient descent.

$$\mathrm{KL}(\hat{\mathbf{Y}}, \mathbf{Y}) = \sum_{i \in I} \sum_{o \in O} Y_i^{(o)} \log(\tfrac{Y_i^{(o)}}{\hat{Y}_i^{(o)}}) \tag{14}$$

The training is regularised for sparsity in the latent state transition matrix with the addition of element-wise KL divergence between the diagonal value of the latent transition matrix and a vector of ones to the loss.

$$\mathrm{KL}(\mathrm{diag}(\sum_L W\mathbf{H}), \mathbf{1}_s) \tag{15}$$

In order to obtain non-redundant latent states, in lieu of model selection, the Jensen-Shannon divergence of the conditional probability vectors of observed states given latent states is minimised.

$$\mathrm{JSD}(P_o(o \mid s^{(1)}), \ldots, P_o(o \mid s^{(m)})) = \sum_{o \in O} \tfrac{1}{|S|} \cdot \sum_{s \in S} \mathrm{KL}(P_o(o \mid s), \rho_S) \tag{16}$$

where,

$$\rho_S = \tfrac{1}{|S|} \cdot \sum_{s \in S} P_o(o \mid s) \tag{17}$$

Model Output

The pseudotime of any cell $o \in O$ is estimated as the mean step of a cell weighted by the probability of observing a cell at each step i,

$$\frac{\sum_{i \in I} P(o|i) \cdot i}{\sum_{i \in I} P(o|i)} \tag{18}$$

The conditional probability of latent states with respect to observed states (cells) is used to assign kinetic cluster memberships,

$$\arg\max_{s \in S} P(S = s \mid o) \tag{19}$$

The transition entropy of a cell is the sum of the entropy of the joint probability of a cell and each latent state,

$$-\sum_{s \in S} P(o, s) \cdot \ln(P(o, s)) \tag{20}$$

Each lineage $l \in L$ is a sequence of transitions in a common state space S. The trajectories of lineages in latent state space are represented as sequences of most probable latent states at each step i,

$$\{\arg\max_{s \in S} P(S = s \mid l, i)\} \qquad \forall i = 1, \ldots, I \tag{21}$$

Data availability

Developing human forebrain

Data was downloaded from *scvelo v0.2.5*

Code availability

Notebooks

All datasets were processed using standard scRNAseq and RNA velocity workflow. Details of the analyses can be found in the following notebooks. https://github.com/aron0093/cy2path_notebooks.

Implementation

The model was implemented using PyTorch and Python code for the project can be found at the following repository https://github.com/aron0093/cy2path.

Results

Factorial state space modelling for differentiation processes

The observed state-space is considered to be discrete and composed of all observed cell-states. The differentiation process is conceptualized as the evolving probability distribution over the observed states and is modelled by simulating the RNA velocity-derived transition probability matrix of cell-states. The purpose of the latent state-space model is to create an interpretable summary of the simulation.

Under the latent state-space model, the differentiation process is the transitions between discrete latent states with probabilistic emissions of observed states. The model is parameterised with a transition probability matrix over latent states and emission probabilities of observed states for each latent state. Analogous to the simulation over observed states, the dynamics over latent states are learnt by minimizing the divergence between the simulation and the estimate from the latent state-space model.

Dynamics over the latent states are more interpretable since the number of latent states is much lower than the observed states. The probabilistic assignment of cells towards latent states is referred to as kinetic clustering. Kinetic clustering of cells is based on state transitions unlike clustering based solely on gene expression profiles. Kinetic clusters group cell-states that arise together during the differentiation process. Lineages are modelled as transitions between latent states and are also informative of the relative persistence of latent states. Multiple co-occurring lineages are modelled as independent Markov chains in latent state-space and can be considered independent components of the observed differentiation dynamics.

Identifying transitional cells in developing human forebrain

The developing human forebrain dataset consists of the glutamatergic neuronal lineage in human embryonic cells. The process follows a linear differentiation path from Radial-glia (progenitor) cells via a neuroblast (intermediate)

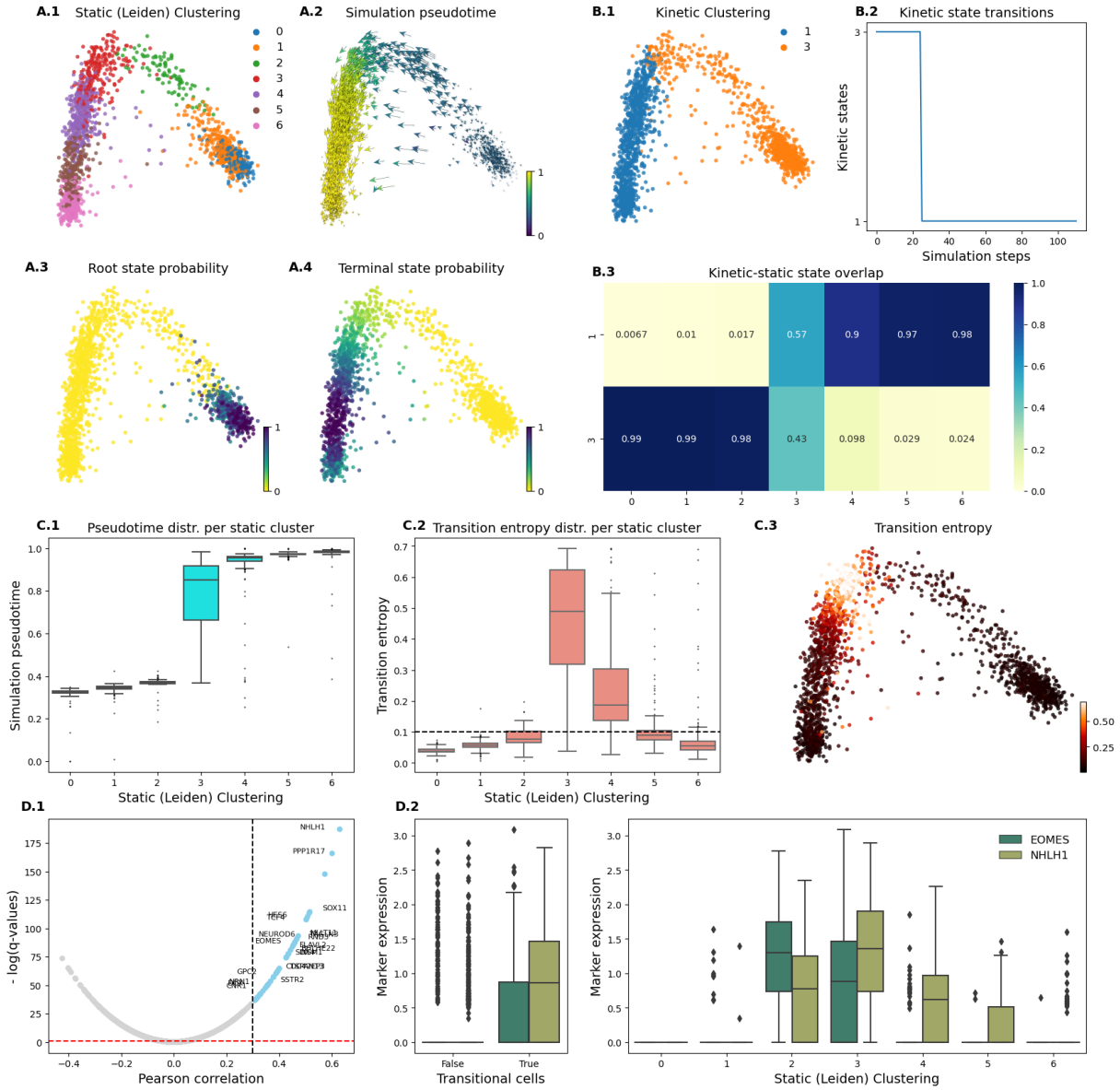


Figure 1: Identifying transitional cells in developing human forebrain. **(A)** Outputs of standard workflow scRNAseq and RNA velocity analysis annotated on the first two principal components. (A.1.) Leiden clustering of cell states, (A.2.) RNA velocity vectors and estimated pseudotime. The pseudotime of a cell is calculated as the mean step weighted by the probability of observing a cell at each step. (A.3.) Root and (A.4.) Terminal cell states inferred using RNA velocity. **(B)** Outputs of our latent state space model. (B.1.) Kinetic clustering of cell states which is the most probable latent state per cell. (B.2.) Most probable latent state at each simulation step. (B.3.) Ratio of overlapping cells in each static (Leiden) and kinetic cluster. **(C)** Identification of transitional cell states. (C.1.) Pseudotime distribution of cells in each static cluster. (C.2.) Transition entropy of cells; computed as the entropy of the joint probability of a cell and each latent state; distribution over static clusters. The red line is the threshold to discriminate transitional cells from the rest. (C.3.) Transition entropy of cells. **(D)** Biological identification of transitional cells as neuroblasts. (D.1.) Pearson correlation between gene expression and transitional entropy of cells. (D.2.) Marker genes' (*EOMES, NHLH1*) expression distribution in transitional cells vs rest and for each Leiden cluster.

population that is locked into the neuron (mature) fate [9]. The intermediate neuroblasts are highly motile cells that migrate to target brain regions before terminal differentiation [12] [13].

Root and terminal states inferred with RNA velocity correspond to the Radial-glia and mature neurons, respectively, as has been previously reported [Figure 1A.3-4] [9]. The model was fit using default parameters [Notebooks] . Kinetic clustering partitioned the data into two clusters [Figure 1B.1]. Static clusters computed using the Leiden algorithm overlap exclusively with one of the kinetic clusters except Leiden cluster 3, which appears to be split between the two kinetic clusters [Figure 1B.3].

Pseudotime estimated using RNA velocity has high variance in Leiden cluster 3, suggesting that this cluster may contain transitional cells [Figure 1C.1]. Transitional cells were identified as cells with high transitional entropy [Figure 1C.2-3]. Marker genes for neuroblasts were enriched in the set of genes positively correlated with transitional entropy [Figure 1D.1] [14].

Cells high in the expression of *EOMES* [9] and of *NHLH1* [15], canonical markers for neuroblast cells, are spread across multiple Leiden clusters. Cells expressing marker genes overlap with cells identified as transitional [Figure 1D.2]. This analysis concludes that transitional entropy is a useful criterion for selecting transitional cells.

Discussion

Differentiation processes are generally represented as a sequence of transitions between cell-types. Intermediate cell-types represent distinct, physiological stages of the process. Clustering of cell-states based on gene expression profiles is an essential step in the study of both terminally differentiated and differentiating cells. Groups of cells obtained in this manner are identified as canonical cell-types by marker identification and functional analysis. In contrast, lineage inference approaches utilise the highly resolved measurement of cell-states with scRNAseq, to model cell-state transitions as gradual processes and not as discrete transitions between cell clusters [16][11]. These methods infer differentiation coordinates for individual cells in the form of pseudotime and cell fate probability.

Inference of the identity and function of individual cell-states is challenging due to dropout of genes measured in scRNAseq and high biological stochasticity between similar cell-states. Dimensionality reduction is a necessary step to compensate for missing measurements by exploiting correlation in genes' expression. While dimensionality reduction reduces the number of features, clustering is an analogous process of reducing the number of states. Cells within a cluster are assigned the same label and subsequent analysis such as differentiation expression testing implicity assumes variation between these cell-states to be uninformative variation. The reduction of states aids interpretabiltiy and discovery of functional associations between genes.

For scRNAseq data from differentiation processes, unlike terminally differentiated cells, some variation within clusters corresponds to the differentiation process itself. A model of transitions between clusters, while interpretable, cannot faithfully represent an incremental process as well as gradual divergence of lineages. Information on transitional cells, the relative time of transitions and the persistence of intermediate states is lost in such a representation [8][6][7].

The asynchronous differentiation of cells is the fundamental basis for constructing models of differentiation processes from scRNAseq data. It can therefore be expected that biological samples collected at different time points will have different distributions of cell-states. Therefore, we propose an approach that models the differentiation process as an evolving probability distribution over observed cell-states. Such an approach allows for the simultaneous persistence of cell-states in different stages of the process while also capturing the sequence of transitions as well as the relative temporal coordinate. RNA velocity of single cells has enabled the estimation of asymmetric transitions between cell-states and several methods have used these transitions for lineage inference [9][11]. In prior work, we demonstrated the utility of a Markov simulation-based lineage inference approach that exploits emergent properties not discernable via analytical formulations [17]. While simulations over cell-states have the desired properties for our modelling approach to differentiation processes, a simulation over several thousand cell-states is not interpretable.

Therefore, we introduce a latent state-space model where we consider the simulation over cell-states to be driven by a latent Markov process over a much smaller number of latent states. The inferred latent process has higher interpretability while retaining the attributes of our approach to differentiation processes. Kinetic clustering of cells is based on state transitions, unlike clustering based on only expression profiles. In the analysis of the human forebrain dataset, we demonstrate that kinetic clustering groups cells with temporal similarity and the utility of this approach in the identification of transitional cells. The model has been implemented using pyTorch and can make use of GPU parallelisation. The code relies on standard packages in the field and can easily be incorporated into RNA velocity-based trajectory inference workflows.

References

[1] "What is your conceptual definition of "cell type" in the context of a mature organism?," *Cell Systems*, vol. 4, pp. 255–259, Mar. 2017.

[2] F. A. Wolf, P. Angerer, and F. J. Theis, "SCANPY: large-scale single-cell gene expression data analysis," *Genome Biology*, vol. 19, Feb. 2018.

[3] Y. Hao, S. Hao, E. Andersen-Nissen, W. M. M. III, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zagar, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. B. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert, and R. Satija, "Integrated analysis of multimodal single-cell data," *Cell*, 2021.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008, Oct. 2008.

[5] V. A. Traag, L. Waltman, and N. J. van Eck, "From louvain to leiden: guaranteeing well-connected communities," *Scientific Reports*, vol. 9, Mar. 2019.

[6] K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit, "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics," *BMC Genomics*, vol. 19, June 2018.

[7] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, and C. Trapnell, "Reversed graph embedding resolves complex single-cell trajectories," *Nature Methods*, vol. 14, pp. 979–982, Aug. 2017.

[8] F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, and F. J. Theis, "PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells," *Genome Biology*, vol. 20, Mar. 2019.

[9] G. L. Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, and P. V. Kharchenko, "RNA velocity of single cells," *Nature*, vol. 560, pp. 494–498, aug 2018.

[10] Z. Ghahramani and M. I. Jordan *Machine Learning*, vol. 29, no. 2/3, pp. 245–273, 1997.

[11] V. Bergen, M. Lange, S. Peidli, F. A. Wolf, and F. J. Theis, "Generalizing RNA velocity to transient cell states through dynamical modeling," *Nature Biotechnology*, vol. 38, pp. 1408–1414, Aug. 2020.

[12] R. D. Hodge, R. J. Kahoud, and R. F. Hevner, "Transcriptional control of glutamatergic differentiation during adult neurogenesis," *Cellular and Molecular Life Sciences*, vol. 69, pp. 2125–2134, Jan. 2012.

[13] C. Bressan and A. Saghatelyan, "Intrinsic mechanisms regulating neuronal migration in the postnatal brain," *Frontiers in Cellular Neuroscience*, vol. 14, Jan. 2021.

[14] G. L. Manno, D. Gyllborg, S. Codeluppi, K. Nishimura, C. Salto, A. Zeisel, L. E. Borm, S. R. Stott, E. M. Toledo, J. C. Villaescusa, P. Lönnerberg, J. Ryge, R. A. Barker, E. Arenas, and S. Linnarsson, "Molecular diversity of midbrain development in mouse, human, and stem cells," *Cell*, vol. 167, pp. 566–580.e19, Oct. 2016.

[15] E. Braun, M. Danan-Gotthold, L. E. Borm, E. Vinsland, K. W. Lee, P. Lönnerberg, L. Hu, X. Li, X. He, Ž. Andrusivová, J. Lundeberg, E. Arenas, R. A. Barker, E. Sundström, and S. Linnarsson, "Comprehensive cell atlas of the first-trimester developing human brain," Oct. 2022.

[16] L. Haghverdi, F. Buettner, and F. J. Theis, "Diffusion maps for high-dimensional single-cell analysis of differentiation data," *Bioinformatics*, vol. 31, pp. 2989–2998, May 2015.

[17] R. Gupta, D. Cerletti, G. Gut, A. Oxenius, and M. Claassen, "Simulation-based inference of differentiation trajectories from RNA velocity fields," *Cell Reports Methods*, vol. 2, p. 100359, Dec. 2022.

Bioinformatics, 38, 2022, i290–i298
https://doi.org/10.1093/bioinformatics/btac227
ISCB/ISMB 2022

psupertime: supervised pseudotime analysis for time-series single-cell RNA-seq data

Will Macnair[1], Revant Gupta[2] and Manfred Claassen[2,3,*]

[1]Institute of Molecular Systems Biology, Department of Biology, ETH Zurich, Zurich 8093, Switzerland, [2]Inner Medicine I, Faculty of Medicine, University of Tübingen, University Hospital Tübingen, 72074, Germany and [3]Department of Computer Science, University of Tübingen, Tübingen 72074, Germany

*To whom correspondence should be addressed.

Abstract

Motivation: Improvements in single-cell RNA-seq technologies mean that studies measuring multiple experimental conditions, such as time series, have become more common. At present, few computational methods exist to infer time series-specific transcriptome changes, and such studies have therefore typically used unsupervised pseudotime methods. While these methods identify cell subpopulations and the transitions between them, they are not appropriate for identifying the genes that vary coherently along the time series. In addition, the orderings they estimate are based only on the major sources of variation in the data, which may not correspond to the processes related to the time labels.

Results: We introduce psupertime, a supervised pseudotime approach based on a regression model, which explicitly uses time-series labels as input. It identifies genes that vary coherently along a time series, in addition to pseudotime values for individual cells, and a classifier that can be used to estimate labels for new data with unknown or differing labels. We show that psupertime outperforms benchmark classifiers in terms of identifying time-varying genes and provides better individual cell orderings than popular unsupervised pseudotime techniques. psupertime is applicable to any single-cell RNA-seq dataset with sequential labels (e.g. principally time series but also drug dosage and disease progression), derived from either experimental design and provides a fast, interpretable tool for targeted identification of genes varying along with specific biological processes.

Availability and implementation: R package available at github.com/wmacnair/psupertime and code for results reproduction at github.com/wmacnair/psupplementary.

Contact: manfred.claassen@med.uni-tuebingen.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Single-cell RNA-sequencing studies have been used to define the transcriptional changes in biological time series, including embryonic development (Petropoulos *et al.*, 2016), response to stimulus (Treutlein *et al.*, 2016), differentiation (Bendall *et al.*, 2014) and ageing (Enge *et al.*, 2017). Such studies are based on single-cell RNA-seq measurements over a sequence of experimental labels of the successive timepoints. These data are typically analysed using unsupervised pseudotime techniques to extract the corresponding temporal sequence of transcriptomic states. These approaches use similarities between cells to computationally order them along trajectories, allowing researchers to identify high-level cell subpopulations and the transitions between them. However, unsupervised methods are not designed to identify genes associated with a process unfolding over time. In addition, they assume that the major driver of variation in the data is most indicative of the time series-induced cell orderings. This means that where the changes along the time series are subtle, or where there are strong additional sources of variation, the orderings they identify may not be those associated with the time series (Saelens *et al.*, 2019). Only recently, approaches have been published to derive or refine pseudotime with time-series information (Shao *et al.*, 2021; Tran and Bader, 2020). To further address this methodological gap, we introduce a supervised pseudotime technique, psupertime, which explicitly uses time-series labels as input (Fig. 1A). psupertime is based on penalized ordinal regression (Fig. 1B), a statistical technique used where data have categorical labels that follow a sequence. psupertime produces three outputs. Firstly, it learns a small, interpretable set of genes that vary coherently over the time series. Secondly, a linear combination of these genes assigns a pseudotime value to each cell, which approximately recapitulates the ordering specified by the sequence of labels. Thirdly, it can be used to classify new data according to the process labelled in the data used for training. These outputs allow for

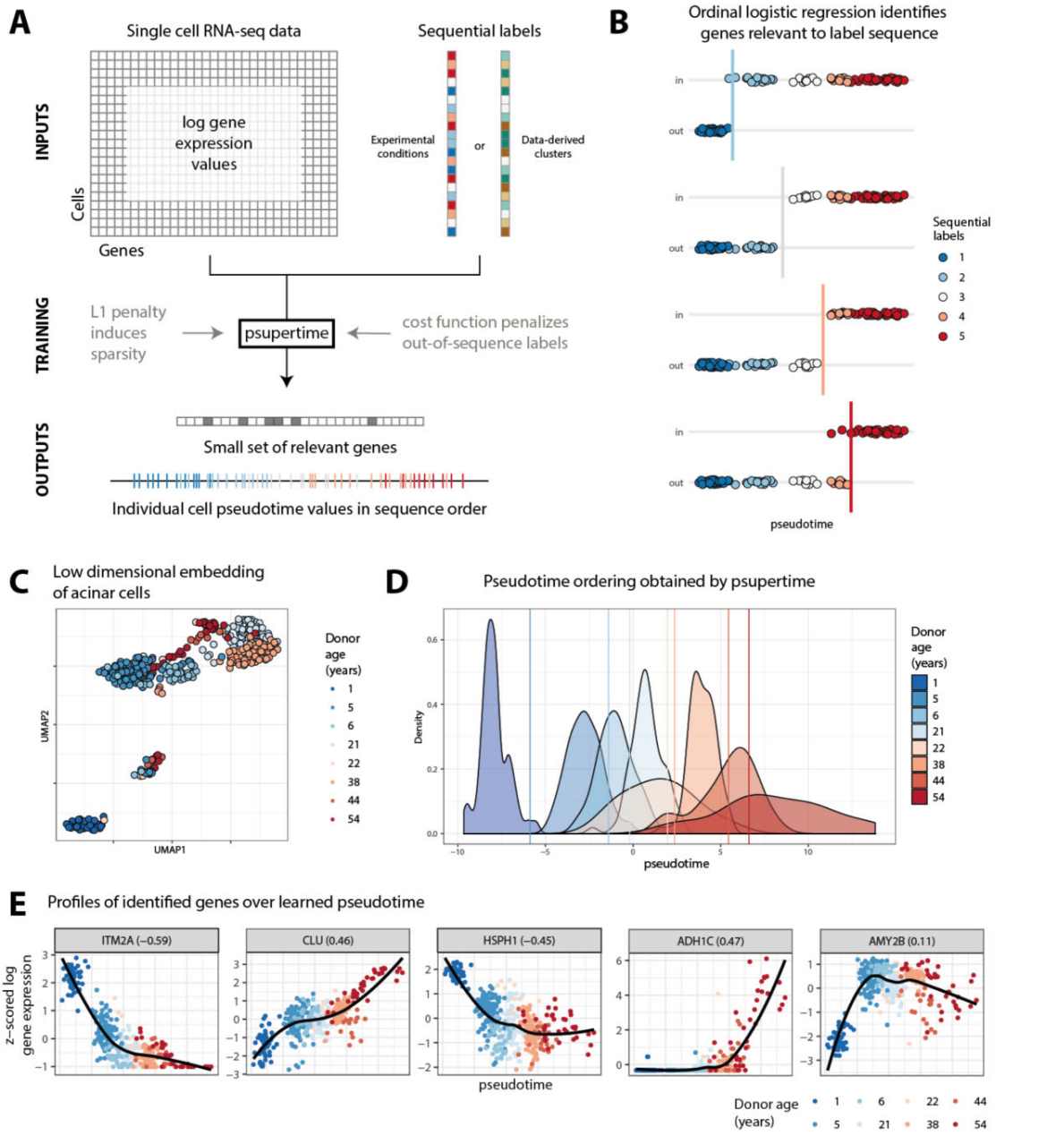


Fig. 1. (**A**) Inputs to psupertime are single-cell RNA-seq data, where the cells have sequential labels associated with them. psupertime then identifies a sparse set of ordering coefficients for the genes. Multiplying the gene expression values by this vector of coefficients gives pseudotime values for each cell, which place the labels approximately in sequence. (**B**) Cartoon of statistical model used by psupertime, including thresholds between labels. Where there is a sequence of K condition labels, psupertime learns $K-1$ simultaneous (i.e. sharing coefficients) logistic regressions, each seeking to separate labels $1 \dots k-1$ (out) from $k \dots K$ (in). (**C**) Dimensionality reduction of 411 human acinar cell data with ages ranging from 1 to 54 (Enge *et al.*, 2017). Representations in two dimensions via non-linear dimensionality reduction technique UMAP. Colours indicate donor age. (**D**) Distributions of donor ages for acinar cells over the pseudotime learned psupertime. Vertical lines indicate thresholds learned by psupertime distinguishing between earlier and later sets of labels; colour corresponds to the next later label. (**E**) Expression values of selected genes (five with largest absolute coefficients; see Supplementary Fig. S2 for 20 largest). The x-axis is psupertime value learned for each cell; y-axis is z-scored $\log_2$ gene expression values. Gene labels also show the Kendall's τ correlation between sequential labels (treated as a sequence of integers $1, \dots, K$) and gene expression

targeted characterization of processes for any single-cell RNA-seq data where sequential labels are available (such as time, disease progression or unidimensional spatial measurements), despite substantial variation not associated with the process of interest. Full details of the method are given in Section 2.

We demonstrate psupertime on a dataset comprising 411 acinar cells from the pancreas, from eight human donors with ages from 1 to 54 years (Enge *et al.*, 2017). Acinar cells perform the exocrine function of the pancreas, producing enzymes for the digestive system. This dataset was selected because each set of cells was obtained from different donors, resulting in significant variation in the dataset unrelated to donor age (Fig. 1C). Despite this variation, psupertime finds a cell-level ordering which respects the age progression, while separating the labels from each other (Fig. 1D). We show that the

performance of psupertime is robust, including perturbations in labels (see Supplementary Results S1).

2 Materials and methods

2.1 Overview of psupertime methodology

psupertime requires two inputs: (i) a matrix of log read counts from single-cell RNA-seq, where columns correspond to genes and rows correspond to cells; and (ii) a set of labels for the cells, with a defined sequence for the labels (e.g. a set of cells could have labels *day1*, *day3*, *day1*, *day2*, *day3*). (Note that not all cells need to be labelled: psupertime can also be run on a labelled subset.) psupertime then identifies a set of ordering coefficients, β_i, one for each gene (Fig. 1A). Multiplication by this vector of coefficients converts the matrix of log gene expression values into pseudotime values for each individual cell. The set of pseudotime values recapitulates the known label sequence (so the cells with labels *day1* will on average have lower pseudotime values than those labelled *day2* and so on). The vector of coefficients is *sparse*, in the sense that many of the values are zero; these therefore have no influence on the ordering of the cells. Genes with non-zero coefficients are therefore identified by psupertime as relevant to the process which generated the sequential labels.

Suppose the sequence of condition labels we have is $1, \ldots, K$. Intuitively, psupertime learns a weighted average of gene expression values that separates the cells with label 1 from the cells with labels $2, \ldots, K$, at the same time as separating 1, 2 from $3, \ldots, K$, and 1, 2, 3 from $4, \ldots, K$ and so on (Fig. 1B). This can be thought of as solving K—1 simultaneous logistic regression problems and is termed *ordinal logistic regression* (McCullagh, 1980).

As described so far, psupertime can be thought of as minimizing a cost, where the cost is the error in the resulting ordering. To make the results more interpretable, we would like psupertime to use a small set of genes for prediction. To do this, we add a cost for each coefficient β_i used, so that psupertime is minimizing $error + \lambda \sum_i |\beta_i|$; approaches like this are termed *regularization*, and in this case *L1 regularization*. The parameter λ controls the balance between minimizing error, and minimizing the 'coefficient cost'. The method for implementing this approach is based on the R package glmnetcr, which we have extended with an additional statistical model.

The results of this procedure are: (i) a small and therefore interpretable set of genes with non-zero coefficients; (ii) a pseudotime value for each individual cell, obtained by multiplying the log gene expression values by the vector of coefficients; and (iii) a set of values along the pseudotime axis indicating the thresholds between successive sequential labels (these can then be used for classification of new samples). Where the data do not have condition labels, psupertime can be combined with unsupervised clustering to identify relevant processes (see Supplementary Results S3).

2.2 Pre-processing of data

To restrict the analysis to relevant genes and denoise the data, psupertime first applies pre-processing to the log transcripts per million values. Specifically, psupertime first restricts to highly variable genes, as defined in the scran package in R, i.e. genes that show above the expected variance relative to genes with similar mean expression (Lun *et al.*, 2016). Genes that are only expressed in a small number of cells (the default is 1%) are excluded. psupertime implement data denoising and dropout correction by calculating correlations between the log expression values across all selected genes for each pair of cells, using the correlations to identify the 10 nearest neighbours for each cell and replacing the value for a given cell by the mean value over these neighbours. Finally, the resulting log-count values for each gene are scaled to have mean zero and standard deviation one.

2.3 Penalized ordinal logistic regression

psupertime applies cross-validated regularized ordinal logistic regression to the processed data, using the labels as the sequence. Ordinal logistic regression is an extension of binary logistic regression to an outcome variable with more than two labels, where the labels have a known or hypothesized sequence. The likelihood for ordinal logistic regression is defined by multiple simultaneous logistic regressions, where each one models the probability of a given observation having an earlier or later label, with the definition of 'early'/'late' differing across the simultaneous regressions (Fig. 1B). The same linear combination of input variables is used across all individual logistic regressions. This specific model of ordinal logistic regression, in which the simultaneous logistic regressions each seek to separate labels $1 \ldots k$ from labels $k+1 \ldots K$, is termed *proportional odds*. (A commonly used alternative is the *continuation ratio* model, where the regressions seek to separate labels $1 \ldots k$ from label $k+1$ alone. This is also implemented as an option in psupertime.)

In the case where the number of input variables is high relative to number of observations and may include many uninformative variables, as is common in single-cell RNA-seq, it can be helpful to introduce sparsity (i.e. to increase the number of zero coefficients). psupertime uses *L1 regularization* to do this. Our approach is based on that in the R package glmnetcr (Archer and Williams, 2012), which reformulates the data and associated likelihood functions into one single regression model, to take advantage of the fast performance of the glmnet package (Friedman *et al.*, 2010). The model originally implemented in glmnetcr is the continuation ratio likelihood; we have extended this approach to implement the proportional odds likelihood, as this model is more appropriate for assessing an entire biological process. Under the proportional odds assumption, the two categories are: categories j and higher, and categories lower than j; the regression therefore estimates $log(P(Y >= j)/P(Y < j))$. Under the continuation ratio assumption, the two categories are: j, and categories lower than j; here, the regression estimates $log(P(Y = j)/P(Y < j))$. Intuitively, the proportional odds framework models an observation's global progression along the ordinal values, while the continuation ratio framework models the probability of proceeding to the next ordinal value. For most of the examples that we have seen, such as studying development or ageing, the proportional odds framework is appropriate. However, the continuation ratio framework may be appropriate in some cases, for example in disease progression, or evolutionary processes. Given input data $X \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{N}^n$ condition labels (which for simplicity we assume are integers), this results in the following cumulative distribution function for ordinal logistic regression:

$$P(y_i \leq j | X_i) = \phi(\theta_j - \beta^T X_i) = \frac{1}{1 + \exp(\beta^T X_i - \theta_j)}.$$

Here, X_i and y_i are the vector and integer corresponding to the ith observation and label respectively, j indicates one of the possible condition labels, β is the vector of coefficients and $\{\theta_j\}$ are the thresholds between labels. ϕ is the logit link function, which transforms the linear combination of predictors into a probability. Note that the probability given here is cumulative and that to calculate the probability of an individual label, we have to calculate the difference between successive labels. This results in the following *unpenalized* likelihood:

$$L(\beta, \theta | y, X) = \prod_{i=1}^{N} (\phi(\theta_{y_i} - \beta^T X_i) - \phi(\theta_{y_i - 1} - \beta^T X_i)),$$

where y_i is the label of observation i. Including the L1 penalty, for a given value of λ, we obtain the optimal values of β and θ by maximizing the following penalized objective function:

$$\underset{\beta, \theta}{\text{argmax}} (\log L(\beta, \theta | y, X) - \lambda \sum_{i=1}^{p} |\beta_i|).$$

psupertime uses cross-validation (with 5 folds as default) to identify

the optimal level of L1 regularization: the optimal λ is the value with the highest mean score over all held-out folds (either accuracy or cross-entropy may be selected as the score; the default is cross-entropy). To increase sparsity, we use the highest value of λ with mean training score within one standard error of the optimal λ, rather than take the optimal λ itself [following Friedman *et al.* (2010)]. The model is then retrained using all training data, with this value of λ, to obtain the best-fitting model.

Where psupertime is used to classify completely new data (e.g. from a different experiment), to make the predictions more robust, the cross-validation should take data structure into account (e.g. selecting entire samples to be left out, rather than cells selected at random).

2.4 Psupertime outputs

The psupertime procedure results in a set of coefficients for all input genes (many of which will be zero) that can be used to project each cell onto a pseudotime axis, and a set of cut-offs indicating the thresholds between successive sequential labels (Fig. 1D). These can be analysed in various useful ways.

The small, interpretable set of genes reported to have non-zero coefficients permits both validation that the procedure has been successful (by observation of genes known to be relevant to the process) and discovery of new relevant genes. The magnitude of a coefficient is a measure of the contribution of this gene to the cell ordering. More precisely, for a gene i with coefficient β_i, each unit increase in log transcript abundance multiplies the odds ratio between earlier and later labels by e^{β_i}. Where β_i is small, a Taylor expansion indicates this is approximately equal to a linear increase by a factor of β_i.

The thresholds indicate the points along the psupertime axis at which the probability of label membership is equal for labels before the cut-off, and after the cut-off. The distances between thresholds, namely the size of transcriptional difference between successive labels, are not assumed to be constant and are learned by psupertime. Distances between thresholds therefore indicate dissimilarity between adjacent labels, and thresholds which are close together suggest labels which are transcriptionally difficult to distinguish.

The learned geneset can also be used as input to dimensionality reduction algorithms such as t-SNE or UMAP; this is discussed in more detail in Supplementary Results S4.

Rather than learning a pseudotime for one fixed set of input points, psupertime learns a function from transcript abundances to the pseudotime. It can therefore be trained on one set of labels and applied to new data with unknown or different labels: any data with overlapping gene measurements can be assessed with regard to the learned process. Furthermore, psupertime can be learned on two different datasets, with different labels, and then each applied to the other dataset: the sequential labels from one dataset allow coefficients relevant to that sequence to be learned, which can then be used to predict these labels for the second dataset. See Supplementary Figure S23 for more discussion.

2.5 Simulations of single-cell RNA-seq data

psupertime is principally useful because it can identify genes which vary over the course of time-series labels. To test this capability, we simulated single-cell RNA-seq data to include three types of gene profiles, defined in terms of their mean expression: mean varying as a time series; sample-specific variation in the mean; and constant mean expression. All genes have biological and technical noise around this mean. This mimics the likely experimental setup, in which the expression at each timepoint is composed of both processes related to the time series, and unrelated variability particular to that sample, e.g. where the samples are derived from different mice.

Our simulation procedure was as follows: (i) calculate relevant statistics from a selected reference dataset, composed of multiple labels, (ii) randomly sample latent time values for each cell, around a common mean for the cell's label, (iii) randomly assign one of the three gene profile types to each gene and randomly sample some parameters for each gene and (iv) sample counts for each cell and gene based on the combination of cell- and gene-level parameters. We discuss each of these steps in turn.

As a reference dataset, we used 575 mouse embryonic beta cells (Qiu *et al.*, 2017a), restricted to 2666 highly variable genes by the procedure described in Lun *et al.*, where the cells were labelled with seven distinct time labels. The statistics used were library size for each cell (i.e. the total number of reads observed) and the mean μ_g and dispersion ρ_g for each gene g (calculated using edgeR; Robinson *et al.*, 2010), assuming a negative binomial distribution. In each simulation, the library sizes of cells were randomly permuted, and the number of cells allocated to each label was randomly permuted.

To sample the latent time values for label i, l_i, we assumed an exponential distribution of time until the next timepoint. The first time point label has mean value 0, then the time to each subsequent timepoint is drawn from an exponential distribution with rate 0.5 (i.e. mean time difference of 2): $l_i|l_{i-1} \sim \text{Exp}(0.5) + l_{i-1}$. To allow for cell-to-cell variability, we then add Gaussian noise to the values for each cell c, with mean 0 and standard deviation 1: $t_c|l_i \sim N(l_i, 1)$. This results in a latent time value for each cell. We then scale these values to have minimum 0 and maximum 1.

The three possible types of gene expression profile that we defined were: time series; label-specific; and non-specific. Each gene follows one of these profiles. Each gene has dispersion and base mean expression defined by the reference dataset. The gene expression profiles were simulated as follows:

- Time-series genes have expression which changes with respect to the latent time values for each cell, where the log fold change relative to the mean follows a logistic curve. This curve is defined by three values: t_0, the curve's midpoint; k, half the derivative of the curve at that midpoint; and L, the asymptotic maximum value of the curve. The log mean expression of this gene in a cell with latent time value t_c is therefore $\log(\mu_g) + L * \text{logistic}((t_c - t_0) * k)$. For each gene, we sampled t_0 from a uniform distribution over $[0, 1]$; k from a log10-normal distribution with mean 1 and standard deviation 1; and L from a gamma distribution with shape 4 and rate 2.
- Label-specific gene profiles are defined by two parameters: the sample in which they show differential expression, and the log fold change in that sample relative to the mean. For each gene, we uniformly at random select a label, and sample the log fold change from a gamma distribution with shape 4 and rate 2.
- Genes with non-specific expression are defined by the dispersion and base mean identified from the reference dataset, and have no difference in distribution across labels.

Each simulation has a defined set of proportions for each type of gene profile, $(p_{\text{ts}}, p_{\text{label}}, p_{\text{non}})$. Each gene is randomly assigned one of the types according to these probabilities.

We now have all the parameters required to sample counts for each combination of cell and gene. The gene-level parameters define, via the combination of base mean expression and possibly also a log fold change relative to the base mean, the mean expression for a given gene, plus its dispersion. The cell-level parameters define the library size for each cell, which is used to scale the base mean. For each cell and gene combination, we sample from the defined negative binomial distribution.

2.6 Simulations of single-cell RNA-seq data with cell types

To simulate time-series data comprising multiple cell types, we used fluorescence-activated cell-sorted stem cells at different stages of differentiation (Koh *et al.*, 2016), which had previously been used for benchmarking (Duò *et al.*, 2018). We assumed that genes had the following four profile types: global time-series, cell-type time series, batch effect and non-specific genes. Global time-series genes have

Table 1. Details of datasets used in benchmarking of pseudotime cell orderings

Dataset name	Source	Accession	Labels used	No. of labels	No. of cells	No. of highly varying genes
Acinar cells	Enge *et al.* (2017)	GSE81547	Donor age	8	411	827
Human germline, F	Li *et al.* (2017)	GSE86146	Age (weeks)	12	992	1081
Embryonic beta cells	Qiu *et al.* (2017a)	GSE87375	Developmental stage	7	575	2666
Human ESCs	Petropoulos *et al.* (2016)	E-MTAB-3929	Embryonic day	5	1529	2876
MEF to neurons	Treutlein *et al.* (2016)	GSE67310	Days since induction	5	315	1698
Colon cells	Herring *et al.* (2017)	GSE102698	User-selected clusters	4, 5	1894	1515
iPSCs	Schiebinger *et al.* (2019)	GSE106340	Days during reprogramming	11	3600	731

the same timing and effect size on gene expression for all cell types; cell-type time-series genes vary over time in all cell types, but the timing and strength of effect are variable between cell types. Specifically, we sampled the parameters defining the response to time series in exactly the same way as for our previous simulations (see Section 2.5); however, for globally varying genes, these parameters were constant across cell types, while for cell-type varying genes, these parameters were sampled independently for each cell type. After simulating count data, we applied psupertime to each cell type individually, and to all cell types grouped together.

2.7 Benchmarking of time-series gene identification against classifiers

psupertime is a classifier that identifies a small subset of relevant features. We therefore compared it to alternative classification methods, which also produce variable importance measures.

Multinomial regression is a simple baseline approach to classification (Venables and Ripley, 2002). For each label, a linear logistic regression is performed to distinguish label from non-label, resulting in k coefficients for each gene. To identify relevant features, for each gene, we calculate the sum of squares of the k coefficients; where a gene is relevant for classifying many labels, or strongly relevant for one label, it will have a large combined weight.

Random forest is a widely used classification algorithm that is known to have good performance in many circumstances (Caruana *et al.*, 2008). One of the outputs produced by the algorithm is 'importance', which is the (relative) mean increase in error when a given variable is permuted. This can be used to identify which genes are most critical to classification performance.

To assess performance, we simulated single-cell RNA-seq data (as described in Section 2.5), assuming that mean gene expression followed one of three possible profiles: time-series, label-specific or a constant mean across labels. We varied the proportions of these types of gene, so that the proportion of genes following time-series profiles, p_{ts}, was 0.1, 0.3, 0.5, 0.7 or 0.9. The proportion of genes following label-specific profiles, p_{label}, was between 0.1 and $1 - p_{\text{ts}}$. The proportion of non-specific genes, p_{non}, accounted for the remainder of genes.

For each triplet of distinct $(p_{\text{ts}}, p_{\text{label}}, p_{\text{non}})$ values, we did 20 simulations starting from 20 different random seeds. For a given simulation, applying psupertime and the benchmark methods resulted in an 'importance' value for each gene. We used this variable to predict time-series-specific genes, and calculated precision-recall curves for each classifier on the basis of how successfully these values identified the true time-series genes.

We note that due to different combinations of randomly selected parameters, some time-series genes in the simulations will be easier to detect and some more difficult. For example, a gene with low absolute log fold change value L, and high dispersion ρ, will have a poor signal-to-noise ratio for the detection of time-series trends. This puts biologically realistic limits on the best performance possible for any algorithm, as for some genes any time-series trends will be obscured by transcriptional variability. For this reason, and also because psupertime is intended to identify a small set of genes, we have restricted our analysis to values of recall between 0% and 10%.

2.8 Benchmarking of cell orderings against pseudotime methods

Both psupertime and unsupervised pseudotime techniques produce a cell ordering, which may or may not correlate with the label ordering. We compared psupertime against unsupervised pseudotime methods, on five datasets with time-series labels (Table 1). We first performed common pre-processing and identification of relevant genes for each dataset, to identify either highly variable genes, or genes showing high correlation with the label sequence. See Supplementary Results S2 for further discussion.

To identify highly variable genes, we followed the procedure described by Lun *et al.*, using an false discovery rate (FDR) cut-off of 10% and biological variability cut-off of 0.5 [see Lun *et al.* (2016) for details of these parameters]. To identify genes showing high correlation with the labels, we calculated the Spearman's correlation coefficient between sequential labels converted into integers, and log gene expression value. Genes with absolute correlation >0.2 were selected.

For principle component analysis (PCA), we calculated the first principal component of the log counts and used this as the pseudotime. Calculation of Monocle2 uses the following default settings: genes with mean expression < 0.1 or expressed in <10 cells filtered out; negbinomial expression family used; dimensionality reduction method *DDRTree*; root state selected as the state with the highest number of cells from the first label; function orderCells used to extract the ordering.

Calculation of slingshot uses the following default settings: first 10 PCA components used as dimensionality reduction; clustering via Gaussian mixture model clustering using the R package mclust, number of clusters selected by Bayesian information criterion; root and leaf clusters selected as the clusters with highest number of cells from the earliest and latest labels, respectively; lineage selected for pseudotime is path from root to leaf cluster. *Note:* For cells very distant from the selected path, slingshot does not give a pseudotime value. For these cells, we assigned the mean pseudotime value over those that slingshot did calculate. Calculation of psupertime used default settings, as described in Section 2.

We tested the extent to which each pseudotime method could correctly order the cells by calculating measures of correlation between the learned pseudotime, and the sequential labels. Kendall's τ considers pairs of points and calculates the proportion of pairs in which the rank ordering within the pair is the same across both possible rankings.

To identify genes with high correlation with the sequential condition labels (Supplementary Table S1), we treated the sequential labels as the set of integers $1, \ldots, K$, calculated the Spearman correlation coefficient with the gene expression. Genes were selected that showed absolute correlation of >0.2 with the sequential labels (few genes showed high correlation with the sequential labels; this low

cut-off was used to ensure that a sufficient number of genes was selected).

2.9 Identification of relevant biological processes

To identify biological processes associated with the condition labels, psupertime first clusters all genes selected for training (e.g. the default highly variable genes), using the R package fastcluster, using five clusters by default. These are ordered by correlation of the mean expression values with the learned pseudotime, i.e. approximately into genes that are up- or down-regulated along the course of the labelled process. psupertime then uses topGO to identify biological processes enriched in each cluster, relative to the remaining clusters; enriched GO terms are calculated using algorithm = 'weight' and statistic = 'fisher' (Alexa and Rahnenführer, 2009).

3 Results

psupertime produces as output a set of ordering coefficients, one for each gene, most of which are zero (i.e. the coefficient vector is 'sparse'). A non-zero ordering coefficient indicates that a gene was relevant to the label sequence. This balances the requirement for predictive accuracy against that for a small and therefore interpretable set of genes. For example, applied to the acinar cells, psupertime used 82 of the 827 highly variable genes to attain a test accuracy of 83% over the eight possible labels (Supplementary Fig. S1). Many of the genes identified via their absolute coefficient values are already known to be relevant to the ageing of pancreatic cells (see expression profiles shown in Fig. 1E, Supplementary Figs. S2 and S3). For example, clusterin (*CLU*) plays an essential role in pancreas regeneration and is expressed in chronic pancreatitis (Lee *et al.*, 2011; Xie *et al.*, 2002); α-amylase (*AMY2B*) is a characteristic gene for mature acinar cells, encoding a digestive enzyme (Omichi and Hase, 1993). In addition, psupertime suggests candidates for further study: *ITM2A* has the highest absolute gene coefficient and is highly differentially regulated in a model of chronic pancreatitis, but has not been investigated in acinar cells (Ulmasov *et al.*, 2013). The genes identified by psupertime were not discussed in the source manuscript, and, importantly, would not be found by naively calculating correlations between the sequential labels and gene expression (see Supplementary Results S2).

GO term enrichment analysis provides further support for the validity of the cell ordering identified by psupertime. We clustered the expression profiles of the highly variable genes and identified GO terms characteristic of each cluster (see Section 2). This procedure identified genes related to digestion as being up-regulated in early ages ('proteolysis' and 'digestion' enriched in cluster 1), and terms related to ageing later in the process ('negative regulation of cell proliferation' and 'positive regulation of apoptotic process' enriched in cluster 5; see Supplementary Figs. S4 and S5). This analysis confirms that the cell ordering learned by psupertime is plausible.

To compare psupertime to other classifiers, we simulated single-cell RNA-seq data to contain genes that vary over time, and also genes with other profiles (see Section 2.5). We compared psupertime's performance against two benchmark classification methods, which also identify relevant features: multinomial regression, as a simple baseline approach to classification (Venables and Ripley, 2002) and a popular classification algorithm that performs well under many circumstances (Caruana *et al.*, 2008). Both classifiers give measures of importance for each variable (see Section 2); we used these to determine how well the classifiers identified time-series genes. We found that the coefficients identified by psupertime identify time-series genes more precisely than the benchmark classifiers (Fig. 2D, Supplementary Fig. S6). In addition, psupertime is able to recapitulate the true latent time values of the cells (Supplementary Fig. S7). The other classifiers assume no structure across the labels and identify any gene which is helpful for distinguishing one label from another; this results in them also identifying genes with sample-specific rather than time-varying variation. The model for psupertime assumes and therefore identifies genes that vary coherently over the timepoint labels.

Unsupervised projection techniques are commonly applied to analyse time-series single-cell RNA-seq data. We therefore compared the cell-level orderings identified by psupertime with those from three alternative, unsupervised pseudotime techniques: projection onto the first PCA component, as a simple, interpretable baseline; Monocle 2 (Qiu *et al.*, 2017b), which is widely used, shown to perform well in a benchmark study (Saelens *et al.*, 2019) and permits the selection of a starting point; and slingshot (Street *et al.*, 2018), which was also shown to perform well (Saelens *et al.*, 2019) and allows both the start and end point of a trajectory to be selected (it is therefore semi-supervised). Applied to the acinar cells, low-dimensional embeddings of the data (including PCA) indicate that while donor-specific factors account for much of the variation, very little transcriptional variation is related to age (Fig. 2A and B; Supplementary Fig. S8). Acinar cell orderings identified by the benchmark methods are not consistent with the known label sequence (Fig. 2C and E). In contrast, the one-dimensional projection learned by psupertime (Fig. 2C) successfully orders the cells by donor age (Kendall's τ correlation coefficient 0.86, which quantifies the concordance between two orderings), while providing a sparse interpretable gene signature related to age.

In addition to the acinar cells, we compared psupertime to the three alternative methods on four further datasets, as specified in Table 1. The correlation of the orderings from the benchmark methods with the labels varies considerably depending on the dataset (Supplementary Table S2), and in particular, depending on the extent of variation unrelated to the labels (Supplementary Fig. S8): both Monocle 2 and PCA show Kendall's τ values of 0.12 or below for the human germline dataset (Li *et al.*, 2017; Supplementary Fig. S9), in comparison to values of at least 0.71 for the human embryonic stem cells (ESCs) dataset (Petropoulos *et al.*, 2016; Supplementary Fig. S10). In all datasets considered, the cell ordering given by psupertime has a higher correlation with the known label sequence than the other pseudotime methods (Fig. 2E). The pseudotime methods used for comparison do not use the timepoint label as input, so it is not surprising that psupertime is better able to recapitulate the label orderings. However, considering that unsupervised methods are frequently used to analyse time series and other ordered data, this comparison is relevant for users. Where genes and processes associated with time labels are the primary interest, our analysis shows that unsupervised techniques alone are not appropriate (see also Supplementary Results S6).

Many datasets comprise samples composed of multiple distinct cell types. In a time-series experiment, this could in principle make it more difficult for psupertime to identify relevant genes: time-related signal for one cell type could be diluted when cell types are analysed together. To test this, we generated synthetic time-series data from multiple cell types, modelling genes that varied over time both globally, and individually within cell types (see Section 2.6). We found that psupertime is best able to identify globally varying genes when applied to all cell types together, and best able to identify cell-type-specific genes via application to each cell type individually (see Supplementary Results S5). In addition, we applied psupertime to a biological dataset comprising multiple distinct trajectories leading to different cell fates, specifically reprogramming mouse embryonic fibroblast cells (MEFs) to induced pluripotent stem cells (iPSCs; Schiebinger *et al.*, 2019). We identified two clear branches (Supplementary Fig. S20): one branch corresponding to reprogramming from MEFs to iPSCs, and one to reprogramming from MEFs to stromal cells. We then applied psupertime three times: to the entire dataset; to the iPSC branch; and to the stromal branch. In each case, we trained psupertime using the experimental days as labels. psupertime identified relevant genes for the global process (e.g. *Dppa5a*; Lee *et al.*, 2014), for reprogramming to iPSCs (e.g. *Cd24a*; Shakiba *et al.*, 2015) and for reprogramming to non-pluripotent cells (e.g. *Xist*; Minkovsky *et al.*, 2012; Supplementary Fig. S21). Taken together, these results show that sensible use of psupertime can identify both globally varying and cell-type-specific time-

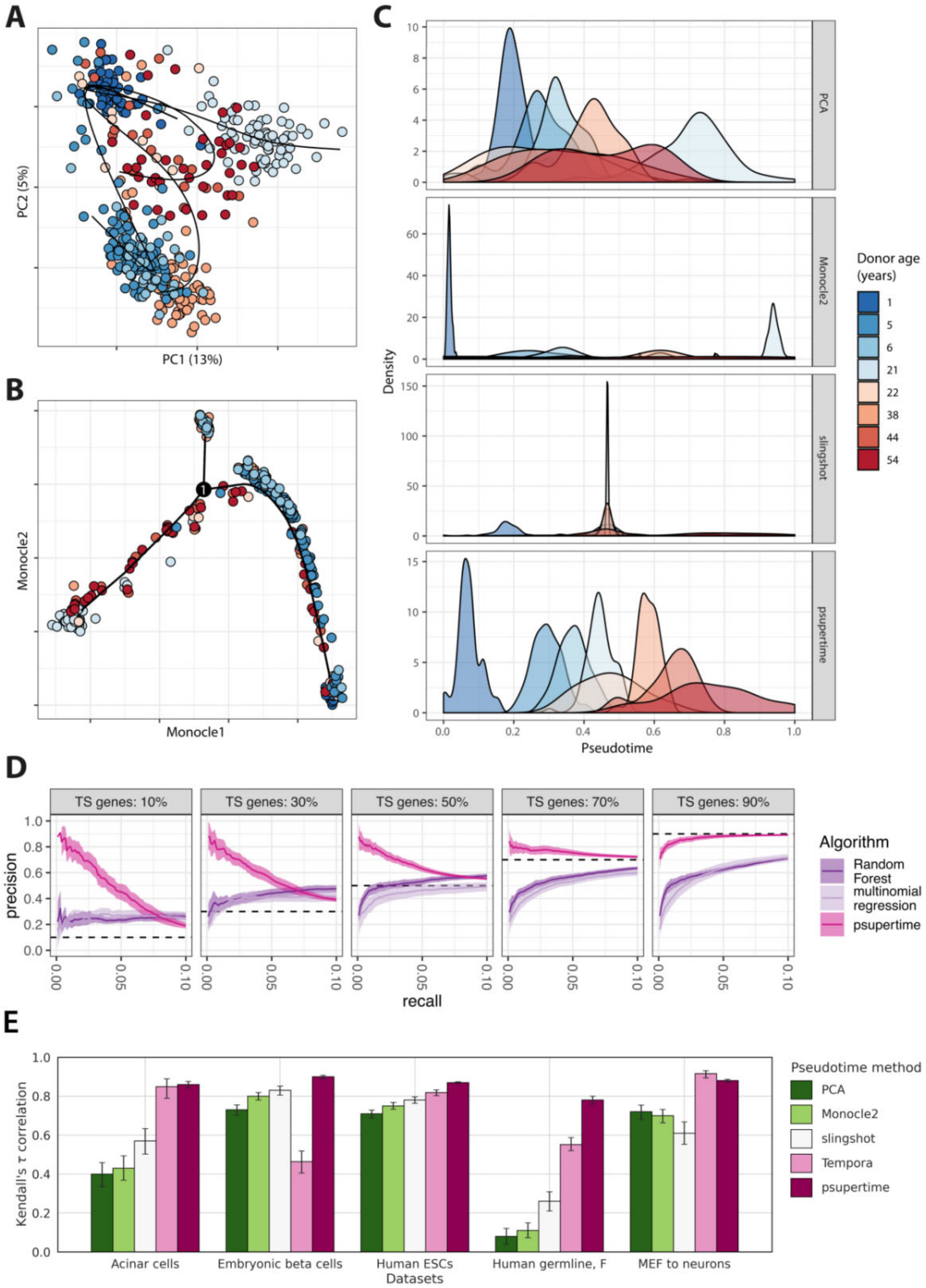


Fig. 2. Performance of psupertime against benchmark methods. See Section 2.8 for details of data processing and use of benchmark methods. All results for (**A–C**) based on 411 aging human acinar cell data with ages ranging from 1 to 54 (Enge *et al.*, 2017), using 827 highly variable genes. Colours indicate donor age. (A) Projection of acinar cells into first two principal components (% of variance explained shown). Curves learned by slingshot shown (note that here we show the projection of these curves into the first two principal components). (B) Projection of acinar cells into dimensionality reduction calculated by Monocle 2, annotated with pseudotime learned by Monocle 2 (Qiu *et al.*, 2017b). (C) Results of benchmark pseudotime methods applied to acinar data. For each method, the x-axis is a one-dimensional representation for each cell (see Section 2.8), scaled to $[0, 1]$ and given the direction with the highest positive correlation with the label sequence. The y-axis is density of the distributions for each label used as input, as calculated by the function geom_density in the R package ggplot2. (**D**) Performance of psupertime and benchmark classifiers in identifying simulated time-series genes. Precision-recall curves based on identification of time-series genes via variable importance measures for each method (see Section 2.7). Line and area show mean and ± 2 standard error, respectively, over 20 simulations. Recall is limited to range 0–10%. Panels correspond to simulations with different proportions of time-series (TS) genes; all panels include 10% batch effect genes which are sample-specific. (**E**) Absolute Kendall's τ correlation coefficient between label sequences (treated as sets of integers $1, \ldots K$) and calculated pseudotimes. Error bars show 95% confidence interval over 1000 bootstraps, calculated with boot package in R. For Tempora, this calculation was performed using scipy package in python. Datasets are specified in Table 1

Table 2 psupertime performance and timings on comparison datasets

Dataset name	Accuracy (%)	Time taken (s)	Sparsity (%)
Acinar cells	75.7 ± 1.1	5.5 ± 0.41	90.6 ± 1.8
Human germline, F	43.4 ± 1.5	25 ± 0.73	80.4 ± 4.9
Embryonic beta cells	78.5 ± 0.9	19 ± 0.67	96.4 ± 0.4
Human ESCs	97.6 ± 0.2	35 ± 0.80	90.0 ± 1.1
MEF to neurons	89.6 ± 1.7	4.7 ± 0.062	96.6 ± 0.4

Note: Mean and standard deviation of psupertime accuracy, timing and sparsity calculated over 10 random seeds.

varying genes. (Details of both analyses are given in Supplementary Results S5.)

Typical workflows for single-cell RNA-seq data first restrict to highly variable genes. If the data are instead first restricted to genes that correlate strongly with the sequential labels, the relative performance of the benchmark methods might improve. Despite the selection of genes that correlate with the labels, psupertime consistently outperforms unsupervised methods in terms of identifying individual cell orderings (Supplementary Results S2). This illustrates that the genes identified by psupertime as most relevant to the process are not necessarily those with highest correlation; for example, genes with expression profiles like *AMY2B* in Figure 1E show a non-linear, step-like expression profile, which results in a correlation of 0.11 with the condition labels. Despite low correlation, such genes were nonetheless found to be useful for cell ordering and suggest that psupertime discovers meaningful non-linear structure in the data.

The time taken for psupertime to run varies over the five test datasets from 4 s for a dataset with ≈300 cells, to 32 s for one with ≈1500 cells (Table 2). We empirically observe a linear runtime dependency for the dataset size in terms of number of cells (∼5 min/10k cells). While maintaining classification accuracies of between 43% and 98%, psupertime uses a small set of genes: for example, for a classification accuracy of 76% on 10% of the acinar cells held out for testing, psupertime uses 10% of the input genes (Table 2). psupertime is based on a form of penalized linear regression. We show that the ordinal logistic model, rather than a linear model based on regarding the sequential labels as integers, is both the natural and the best-performing model for this problem (see Supplementary Results S2).

4 Discussion

The number of studies using single-cell RNA-seq is increasing exponentially (Saelens *et al.*, 2019), and many of these include time-series labels. psupertime is explicitly designed to take advantage of such a setting, complementing unsupervised pseudotime techniques. The presence of time-series labels allows a simple, regression-based model to identify relevant cell orderings; here, the more sophisticated pseudotime approaches required for unlabelled data identify the principal variation in the data, rather than that associated with the labels. The potential asynchrony of dynamic processes is expected to affect classification performance. Specifically, we expect the misclassification rate to increase with stronger asynchrony. While poor classification performance can have other causes than asynchrony, we recommend to consider asynchrony as a possible cause for poor psupertime classification performance and to resort to other dedicated tools/experiments to investigate possible asynchrony. psupertime uses L1 regularization to obtain a small set of reported genes. However, this may result in exclusion of other relevant genes: where there are multiple highly correlated genes that are predictive of the sequential labels, L1 regularization will tend to result in only one of these genes being reported, and produce zero coefficients for other correlated genes. This issue can be addressed by calculating the psupertime ordering, and reviewing all genes that have high correlations with the genes identified by psupertime. Alternatively, a simple extension to psupertime would allow training with a combination of L1 and L2 penalties (the elastic net), resulting in a compromise between sparsity and prediction performance. psupertime could possibly benefit from alternative normalization techniques, such as regularized negative binomial regression resulting in Pearson residuals (Butler *et al.*, 2018), as well as combination with RNA velocity-based pseudotime (Bergen *et al.*, 2020). psupertime is applicable to any experimental design with sequential labels, most obviously time series but also to biological questions regarding drug dose–response, and disease progression. psupertime could further be used in situations without experimental labels by combining with unsupervised techniques (see Supplementary Results S3) or to align new data to orderings learned from alternative processes or separate lineage branches (see Supplementary Results S5 and Figs. S19–S22). More broadly, we have used it to improve dimensionality reduction (see Supplementary Results S4) and are developing extensions including to additional single-cell technologies such as mass cytometry (see Supplementary Results S6). This demonstrates the potential of ordinal regression models for further methodological developments. psupertime has wide applicability and will enable quick and effective identification of the genes and profiles relevant to state sequences of biological processes in single-cell RNA-sequencing data. We have developed an R package available for download at github.com/wmacnair/psupertime.

Financial Support: none declared.

Conflict of Interest: none declared.

References

Alexa,A. and Rahnenführer,J. (2009) Gene set enrichment analysis with topGO. *Bioconductor Improv.* **27, 1–26**.

Archer,K.J. and Williams,A.A.A. (2012) L1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Stat. Med.* **31**, 1464–1474.

Bendall,S.C. *et al.* (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725.

Bergen,V. *et al.* (2020) Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414.

Butler,A. *et al.* (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420.

Caruana,R. *et al.* (2008). An empirical evaluation of supervised learning in high dimensions. In: *Proceedings of the 25th International Conference on Machine Learning*. ICML '08, ACM, New York, NY, pp. 96–103.

Duò,A. *et al.* (2018) A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res* **7**, 1141.

Enge,M. *et al.* (2017) Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. *Cell* **171**, 321–330.e14.

Friedman,J. *et al.* (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22.

Herring,C.A. *et al.* (2018) Unsupervised trajectory analysis of single-cell RNA-seq and imaging data reveals alternative tuft cell origins in the gut. Cell Syst. , **6**, 37–51.e9. https://doi.org/10.1016/j.cels.2017.10.012.

Koh,P.W. *et al.* (2016) An atlas of transcriptional, chromatin accessibility, and surface marker changes in human mesoderm development. *Sci. Data* **3**, 160109.

Lee,D.-S. *et al.* (2014) An epigenomic roadmap to induced pluripotency reveals DNA methylation as a reprogramming modulator. *Nat. Commun.* **5**, 5619.

Lee,S. *et al.* (2011) Essential role of clusterin in pancreas regeneration. *Dev. Dyn.* **240**, 605–615.

Li,L. *et al.* (2017) Single-cell RNA-Seq analysis maps development of human germline cells and gonadal niche interactions. *Cell Stem Cell* **20**, 858–873.e4.

Lun,A.T.L. *et al.* (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. *F1000Res* **5**, 2122.

McCullagh,P. (1980) Regression models for ordinal data. *J. R. Stat. Soc. Series B Stat. Methodol.* **42**, 109–142.

Minkovsky,A. *et al.* (2012) Concise review: pluripotency and the transcriptional inactivation of the female mammalian X chromosome. *Stem Cells* **30**, 48–54.

Omichi,K. and Hase,S. (1993) Identification of the characteristic amino-acid sequence for human α-amylase encoded by the AMY2B gene. *Biochim. Biophys. Acta* **1203**, 224–229.

Petropoulos,S. *et al.* (2016) Single-cell RNA-Seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* **165**, 1012–1026.

Qiu,W.-L. *et al.* (2017a) Deciphering pancreatic islet β cell and α cell maturation pathways and characteristic features at the single-cell level. *Cell Metab.* **25**, 1194–1205.e4.

Qiu,X. *et al.* (2017b) Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982.

Robinson,M.D. *et al.* (2010) Edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140.

Saelens,W. *et al.* (2019) A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554.

Schiebinger,G. *et al.* (2019) Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 928–943.e22.

Shakiba,N. *et al.* (2015) CD24 tracks divergent pluripotent states in mouse and human cells. *Nat. Commun.* **6**, 7329.

Shao,L. *et al.* (2021) Identify differential genes and cell subclusters from time-series scRNA-seq data using scTITANS. *Comput. Struct. Biotechnol. J.* **19**, 4132–4141.

Street,K. *et al.* (2018) Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477.

Tran,T.N. and Bader,G.D. (2020) Tempora: cell trajectory inference using time-series single-cell RNA sequencing data. *PLoS Comput. Biol.* **16**, e1008205.

Treutlein,B. *et al.* (2016) Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature* **534**, 391–395.

Ulmasov,B. *et al.* (2013) Differences in the degree of cerulein-induced chronic pancreatitis in C57BL/6 mouse substrains lead to new insights in identification of potential risk factors in the development of chronic pancreatitis. *Am. J. Pathol.* **183**, 692–708.

Venables,W.N. and Ripley,B.D. (2002). *Modern Applied Statistics with S*, **4**th edn. Springer, New York, NY. ISBN 0-387-95457-0.

Xie,M.-J. *et al.* (2002) Expression of clusterin in human pancreatic cancer. *Pancreas* **25**, 234–238.

nature methods

ARTICLES

https://doi.org/10.1038/s41592-021-01391-1

Deterministic scRNA-seq captures variation in intestinal crypt and organoid composition

Johannes Bues[1,2,10], Marjan Biočanin[1,2,10], Joern Pezoldt[1,2,10], Riccardo Dainese[1,2], Antonius Chrisnandy[3], Saba Rezakhani[3], Wouter Saelens[1,2,4,5], Vincent Gardeux[1,2], Revant Gupta[6], Rita Sarkis[1], Julie Russeil[1,2], Yvan Saeys[4,5], Esther Amstad[7], Manfred Claassen[6,8], Matthias P. Lutolf[3,9] and Bart Deplancke[1,2] ✉

Single-cell RNA sequencing (scRNA-seq) approaches have transformed our ability to resolve cellular properties across systems, but are currently tailored toward large cell inputs (>1,000 cells). This renders them inefficient and costly when processing small, individual tissue samples, a problem that tends to be resolved by loading bulk samples, yielding confounded mosaic cell population read-outs. Here, we developed a deterministic, mRNA-capture bead and cell co-encapsulation dropleting system, DisCo, aimed at processing low-input samples (<500 cells). We demonstrate that DisCo enables precise particle and cell positioning and droplet sorting control through combined machine-vision and multilayer microfluidics, enabling continuous processing of low-input single-cell suspensions at high capture efficiency (>70%) and at speeds up to 350 cells per hour. To underscore DisCo's unique capabilities, we analyzed 31 individual intestinal organoids at varying developmental stages. This revealed extensive organoid heterogeneity, identifying distinct subtypes including a regenerative fetal-like *Ly6a*$^{+}$ stem cell population that persists as symmetrical cysts, or spheroids, even under differentiation conditions, and an uncharacterized 'gobloid' subtype consisting predominantly of precursor and mature (*Muc2*$^{+}$) goblet cells. To complement this dataset and to demonstrate DisCo's capacity to process low-input, in vivo-derived tissues, we also analyzed individual mouse intestinal crypts. This revealed the existence of crypts with a compositional similarity to spheroids, which consisted predominantly of regenerative stem cells, suggesting the existence of regenerating crypts in the homeostatic intestine. These findings demonstrate the unique power of DisCo in providing high-resolution snapshots of cellular heterogeneity in small, individual tissues.

Single-cell RNA sequencing (scRNA-seq)[1] induced a paradigm shift in biomedical sciences, given that it enables the dissection of cellular heterogeneity by high-dimensional data. Recent technological developments, particularly for cell capture and reaction compartmentalization[2–6], have led to a substantial increase in experimental throughput, enabling massive mapping efforts such as the fly, mouse and human cell-atlas studies[5,7–9]. However, given that the majority of scRNA-seq methods rely on stochastic cell capture, entailing large sample inputs, efficient processing of small samples (<1,000 cells) remains challenging. The three main reasons for this are, first, the high fixed run costs, which lead to a large expense per cell at low inputs. To reduce scRNA-seq costs, cell hashing approaches[10–12] were developed that enable sample multiplexing. However, these are not applicable to small input samples due to high cell losses during the mandatory cell washing procedures. The second reason is the requirement of minimum cell inputs. For example, fluorescence-activated cell sorting (FACS)-based or 10X Chromium systems require minimum cell inputs ranging between 10,000 and 500 cells, respectively[13,14]. The third reason is the reduced effectiveness at low inputs resulting from limited cell capture efficiencies or cell size-selective biases[15] when processing small heterogeneous samples. To illustrate these limitations, we summarize the performance of various scRNA-seq technologies on low-input samples in Table 1. Consequently, small samples involving, for instance, zebrafish embryos[16], organisms such as *Caenorhabditis elegans*[17], or intestinal organoids[18–20] are still pooled to obtain cell numbers that are compatible with stochastic microfluidic or well-based technologies, or processed as whole individual tissues[21]. This particularly hampers research on emergent and self-organizing multicellular systems, such as organoids, which are heterogeneous and small at critical development stages.

In this study we develop a deterministic, mRNA-capture bead and cell co-encapsulation dropleting system (DisCo) for low cell input scRNA-seq. DisCo depends on machine-vision to actively detect cells and coordinate their capture in droplets, allowing for continuous operation and enabling free per-run scaling and serial processing of samples. We demonstrate that DisCo can efficiently process samples containing only a few hundred cells, a sample type that tends to fall outside the scope of current cell processing platforms (Table 1). To illustrate DisCo's unique capabilities, we explored the heterogeneous, early development of individual intestinal organoids at the single cell level. In total, we processed 31 single organoids using DisCo at four developmental time points after symmetry breaking, and identified striking differences in cell type

[1]Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [2]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland. [3]Laboratory for Stem Cell Bioengineering, Institute of Bioengineering, School of Life Sciences École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [4]Data Mining and Modelling for Biomedicine, VIB Center for Inflammation Research, Ghent, Belgium. [5]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium. [6]Internal Medicine I, University Hospital Tübingen, Faculty of Medicine, University of Tübingen, Tübingen, Germany. [7]Soft Materials Laboratory, Institute of Materials, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [8]Department of Computer Science, University of Tübingen, Tübingen, Germany. [9]Roche Institute for Translational Bioengineering (ITB), Roche Pharma Research and Early Development, Basel, Switzerland. [10]These authors contributed equally: J. Bues, M. Biočanin, J. Pezoldt. ✉e-mail: bart.deplancke@epfl.ch

Table 1 | Comparison of the performance of established scRNA-seq platform technologies with the DisCo system

Approach	Droplets (stochastic)			FACS & plate based		Traps	Microwells		Droplets (deterministic)[c]
Technology	10X Chromium	inDrop	Drop-seq	Smart-seq2	Cel-seq2	Fluidigm C1	iCell8	Seq-well	DisCo (this study)
Minimum input	500 (HT)/100 (LT)	1,000	50,000	10,000	10,000	<50	1,600	400	<50
Efficiency	45%[a]/30%[a]	25%[a]	2.3%[a]	-	-	30–45%[a]	43%[b]	30%[a]	75%[a]
US$ per cell (100 output cells)	$20/$5.9	$2.1	$6	$10.6	$3.6	$29 (96 cells)	$5	$2.2	$1
US$ per cell (100 input cells)	$44.4/$19.8	$8.4	$260.9	-	-	$62.2	$11.6	$7.5	$1.3
Additional remarks or limitations	Multiplexing possible but requires multiple washing procedures[10,11] and thus substantial efficiency losses expected.			Fluorescent labeling necessary, expensive to scale up (automation).		Size-selective properties[15,40].	High initial acquisition cost.		

[a]Efficiency estimates including cell capture efficiency. [b]Efficiency estimates excluding cell capture efficiency. Performance metrics were derived from the literature. Noteworthy, as for lack of consensus experiments, efficiency metrics represent different values as elaborated in [a] and [b]. Furthermore, the cost per cell is calculated for 100 cells (output: 100 single cells that are successfully processed, thus not incorporating platform-specific processing inefficiencies; input: a sample of 100 total cells that are processed on the respective system, hence considering platform-specific processing inefficiencies) to match the sample size used for the DisCo experiments. (References and the calculation of metrics are detailed in the Methods section.) Performance metrics calculated for the DisCo system in this study.

composition between individual organoids. Among these organoid subtypes, we detected spheroids that are composed of regenerative fetal-like stem cells as well as a rare, novel subtype that is predominantly composed of precursor and mature goblet cells, which is why we named this subtype 'gobloid'. Finally, we used DisCo to uncover cellular variation between individual intestinal crypts, providing evidence for this technology's capacity to also process low cell input, in vivo-derived tissues.

Results

Engineering and operational features of DisCo. The main engineering and operational features of our DisCo device are summarized in Fig. 1 and Extended Data Fig. 1. The important points are the implementation of Quake-style microvalves[22] to facilitate flow control during operation; the development of a machine-vision-based approach utilizing subsequent image subtraction for blob detection (Extended Data Fig. 1b); the induction of deterministic particle displacement patterns (Extended Data Fig. 1c) to move particles into the target region of interest with 95.9% of particles placed in an ~200-µm-wide region (Extended Data Fig. 1d); and the precise pressurization of the dropleting valve to enable accurate control of droplet volume (Extended Data Fig. 1e and Supplementary Video 1). With all components operating in tight orchestration, we were able to generate monodisperse emulsions with high co-encapsulation purity (Fig. 1e and Supplementary Video 2).

DisCo efficiently captures cells from low-input samples. As a first benchmarking experiment, we set out to determine the cell capture efficiency and the bead and cell co-encapsulation purity of DisCo. We found that on average, 91.4% of all droplets contain a cell and a bead, and only 1.7% contain an independent cell doublet (Fig. 1f). Overall, the system provided high cell capture efficiencies of 90% at around 200 cells per hour for a cell concentration of 2 cells per µl (Fig. 1g). At a higher cell concentration of 20 cells per µl, the processing speed could be increased to 350 cells per hour, although with a decreased capture efficiency of approximately 75%. Next, we benchmarked the performance of DisCo for scRNA-seq. With drastically reduced bead amounts contained in the generated sample emulsion, we used our previously developed and characterized chip-based complementary DNA generation protocol[23]. Initially, as a library quality measure, we performed a species mixing experiment of human embryonic kidney (HEK) 293T and murine immortalized brown pre-adipocyte (IBA) cells. We observed clear species separation (Fig. 1h) and an increased read-utilization rate compared with conventional Drop-seq experiments (Extended Data Fig. 2a). In addition, we were able to improve the detectable number of transcripts per cell as compared with published Drop-seq datasets on HEK 293T cells[2,23] (Fig. 1i) by exploiting the uniquely low number of beads in DisCo samples (<500), which enabled us to identify and merge closely related barcodes (described in Methods) without compromising single cell purity (Extended Data Fig. 2b,c). Given that DisCo requires a longer time period to process cells (for example, compared with the 10X Chromium instrument), we also assessed time-dependent effects on the quality of the single cell data by analyzing HEK 293T cells that were loaded on our system (at 22 °C (room temperature)) for 0–20, 20–40 and 40–60 min. Furthermore, we sampled cells that were stored for 120 and 180 min on ice. Based on cell stress metrics such as mitochondrial read content and heat shock protein expression as well as further gene expression analyses, we found artifacts only for cell suspensions that were stored for extended periods (>2 h) of time (Extended Data Fig. 2d,e).

Given that DisCo actively controls fluid flow on the microfluidic device and is capable of efficiently processing cells from the first cell on, we hypothesized that the system should provide reliable performance on small samples of <100 cells. To determine the overall cell capture efficiency of DisCo, we quantified the number of input cells using impedance measurements, which enabled accurate counting of the number of input cells as validated by microscopy (Extended Data Fig. 2f). We processed cell numbers of between 50 and 200 cells, of which 74.9% (s.d. ±10.7%) of input cells had more than 500 unique molecular identifiers (UMIs) per cell (Fig. 1j). To contextualize these performance metrics, we performed similar experiments involving 38, 125 and 215 HEK 293T cells on DisCo's closest competitor system for low cell input samples: the Fluidigm C1 platform (Table 1). We chose the 96-trap chip given that, according to the user manual, it is the more suitable chip for low cell inputs. We found that the Fluidigm C1 system achieves absolute processing efficiencies between 30% and 45% (Extended Data Fig. 2g), matching the performance listed by the manufacturer for the 215-cell condition. Overall, these results, together with reported data, indicate that the DisCo approach outperforms other technologies that are capable of analyzing low input cell samples in terms of processing efficiency.

DisCo on individual intestinal organoids resolves cell types. As a real-world application, we used DisCo to explore the developmental

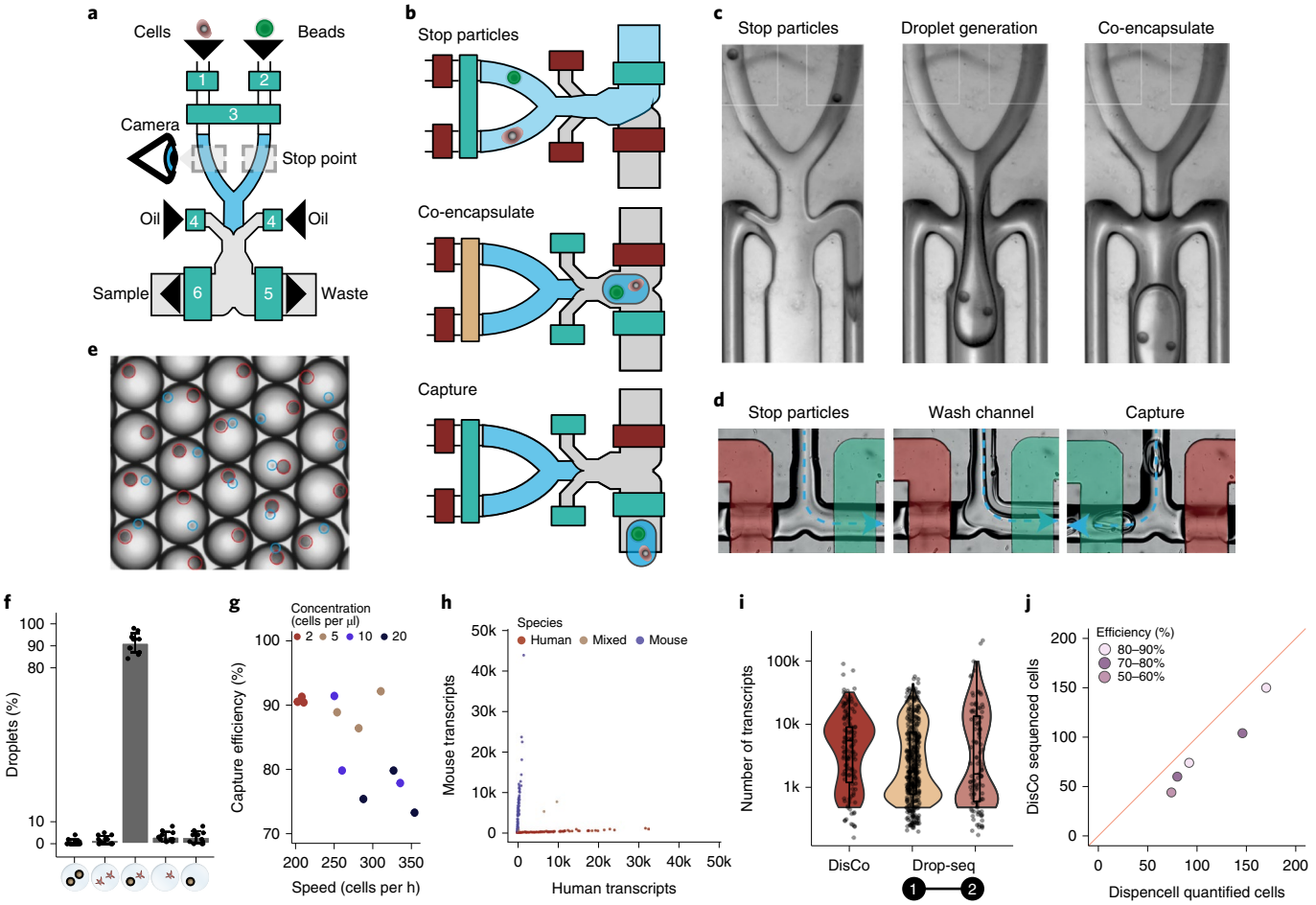


Fig. 1 | Overview and critical feature assessment of the DisCo system. a, Schematic diagram of the DisCo microfluidic device, which contains three inlet channels for cells, beads and oil (shown twice for illustration purposes); two outlets for waste and sample liquids, and several Quake-style microvalves (green boxes): 1, cell valve; 2, bead; 3, dropleting; 4, oil; 5, waste; 6, sample. Particles are detected by a camera and are placed at the Stop point. **b**, DisCo co-encapsulation process on the DisCo device (red, closed; green, open; light brown, dropleting pressure (partially closed)). **c**, The co-encapsulation process of two beads as observed on-chip. Dyed liquids were used to examine the liquid interface of the carrier liquids. Channel sections with white squares are 100 µm wide. **d**, The droplet capture process as observed on-chip. Valves are highlighted according to their actuation state (red, closed; green, open). **e**, Image of DisCo droplet contents. Cells (blue circles) and beads (red circles) were co-encapsulated and the captured droplets were imaged. Mean bead-size is approximately 30 µm. **f**, Droplet occupancy of DisCo-processed cells and beads (total encapsulations, $n=1{,}203$). Bars represent the mean, and error bars represent ±s.d. **g**, Cell capture efficiency and speed for varying cell concentrations (2–20 cells per µl, total encapsulations, $n=1{,}203$). **h**, DisCo scRNA-seq species separation experiment. HEK 293T and murine IBA cells were processed with the DisCo workflow for scRNA-seq, the barcodes merged and the species separation visualized as a Barnyard plot. **i**, Comparison of detected transcripts (UMIs) per cell of conventional Drop-seq experiments. UMIs per cell from HEK 293T data for conventional Drop-seq experiments (1, from ref. [23]; 2, from ref. [2]) are compared with the HEK 293T DisCo data. Drop-seq datasets were down-sampled to a similar sequencing depth. Box plot elements showing UMI counts per cell represent the following values: center line, median; box limits, upper and lower quartiles; whiskers, 1.5-fold the interquartile range; points, UMIs per cell. **j**, Total cell processing efficiency of DisCo at low cell inputs. Input cells (HEK 293T) ranging from 74 to 170 were quantified by impedance measurement. Subsequently, all cells were processed with DisCo, sequenced and quality filtered (>500 UMIs). The red line represents 100% efficiency, and samples were colored according to the recovery efficiency after sequencing.

heterogeneity of intestinal organoids[24]. These polarized epithelial tissues are generated by intestinal stem cells in three-dimensional matrices through a stochastic self-organization process, and mimic key geometric, architectural and cellular hallmarks of the adult intestinal mucosa (for example, a striking crypt–villus-like axis)[24]. When grown from single stem cells, organoids of very different morphologies form under seemingly identical in vitro conditions (Fig. 2a and overview image in Extended Data Fig. 3a). Pooled tissue scRNA-seq data have shed light on the in vivo-like cell type composition of these organoids[18–20,25], but cannot resolve inter-organoid heterogeneity. Critical for organoid development is an early symmetry breaking event at day 2 (16–32-cell stage) that is triggered by cell-to-cell variability and which results in the generation of the first Paneth cell that is responsible for crypt formation[18]. Here, we were particularly interested in examining the emergence of heterogeneity between individual organoids subsequent to the symmetry breaking time point. To do so, we isolated single cells that were positive for LGR5 (leucine-rich repeat-containing G-protein-coupled receptor 5) by FACS, and maintained them in a stem cell state using a mixture of CHIR99021 and valproic acid (CV)[26]. On day 3 of culture, CV was removed to induce differentiation. In total, we sampled 31 single intestinal organoids across four time points (days 3–6, annotated as S0–S3) (Fig. 2a). These organoids were selected in a biased manner based on differences in morphology (for example,

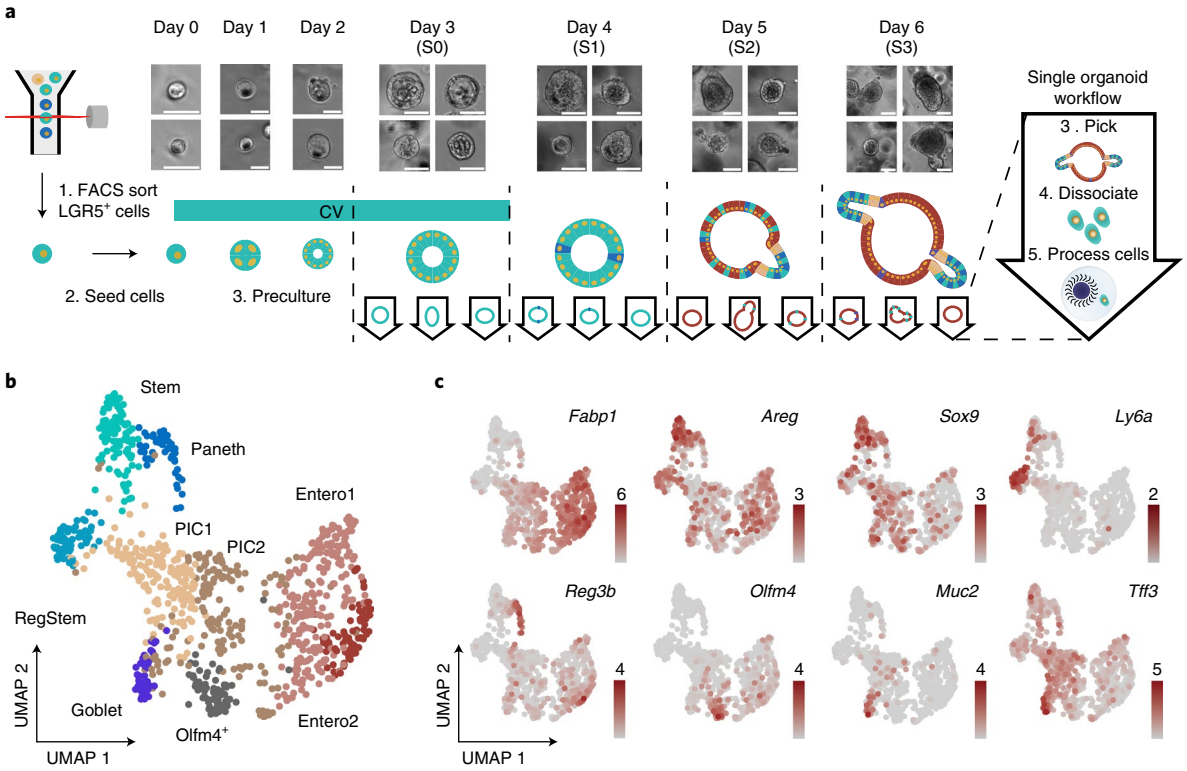


Fig. 2 | Use of DisCo to map intestinal organoid cell identities. a, Overview of the experimental design for using DisCo to process individual organoids. Single LGR5+ intestinal stem cells were isolated via FACS and precultured for 3 days under stem cell maintenance conditions (day 0 seeding, ENR CV days 0–3). On day 3, CV was removed from the culture, and organoids differentiated under ENR conditions for up to 3 days. For each day during development (S0–S3), individual organoids were isolated, dissociated and processed on the DisCo platform. Representative bright-field imaging examples of individual organoids for each day are shown on top. Scale bars, 25 µm (days 0–2); 50 µm (days 3–6). Organoids were collected in three replicates. **b**, UMAP embedding of all sequenced cells. All 945 processed cells from 31 organoids were clustered with k-means clustering, after which clusters were annotated according to specific marker gene expression. Entero1/2, enterocyte cluster 1/2; PIC1/2, potential intermediate cell cluster 1/2[30]; RegStem, regenerative stem cells; Stem, stem cells. **c**, UMAP-based visualization of the expression of specific markers that were used for cluster annotation.

size variation and cystic versus non-cystic morphologies). As a quality control, we correlated the number of encapsulated cells with the number of retrieved barcodes, which was in approximate accordance (Extended Data Fig. 3b; for an overview of the number of sequenced cells per organoid see Supplementary Table 1). The even distribution of the number of reads mapping to ribosomal protein-coding genes and the observed low expression of heat shock protein-coding genes thereby indicated that most cells were not affected by dissociation and on-chip processing (Extended Data Fig. 3c).

We first jointly visualized all 945 cells passing the quality thresholds using Uniform Manifold Approximation and Projection (UMAP). We found that the data were consistent with previously published pooled organoid scRNA-seq read-outs[19,25] given that it showed expected cell types including *Fabp1*-expressing enterocytes, *Muc2*-expressing goblet cells, *Reg3b*-positive Paneth cells and *Olfm4*-expressing stem cells (Fig. 2b,c). In addition, a rare subset of cells had *ChgA* and *ChgB* expression, indicating the expected presence of enteroendocrine cells (Extended Data Fig. 3d). Noteworthy, we found that batch effects are correctable given that neither batch-based nor cell quality-driven clustering was observed after correction (Extended Data Fig. 3c,e). To further validate that batch effects between individual organoids can be corrected, we generated an independent dataset of an additional nine individual organoids (Extended Data Fig. 4a). One of these nine organoids was split into two independent samples and processed with a 60 min time delay in between. We found that the two halves of the split organoid were overlapping in the denominator UMAP (Extended Data Fig. 4b), indicating that batch effects between individual organoids with extended storage times are indeed correctable. These findings support the cell type-resolving power of our DisCo platform (Fig. 2c, extensive heatmap in Extended Data Fig. 5a and list in Supplementary Table 2).

In addition to the expected cell types, we observed a distinct cluster marked by high expression of stem cell antigen 1 (*Sca1* or *Ly6a*), *Anxa1* and *Clu* (Extended Data Fig. 3d), as well as increased YAP1 target gene expression (Extended Data Fig. 5b), suggesting that these cells are most likely regenerative fetal-like stem cells[27–29]. The two remaining clusters did not show a striking marker gene signature, but probably represent stem cells and potential intermediate cells (PICs)[30], given their occurrence at early developmental time points (Fig. 3a). To further leverage the temporal component in the DisCo data, we used slingshot trajectory analysis[31] to infer lineage relationships between cell types and to identify genes that may be of particular significance for waypoints along differentiation (Fig. 3b). Beyond the previously utilized marker genes for cell type annotation, for example *Reg3b* and *Reg3g* for Paneth cells, additional markers that were validated in previous studies[32] were identified, such as *Agr2* and *Spink4*, and *Fcgbp* for goblet cells (Fig. 3c). Overall, this suggests that the meta-data produced with our DisCo platform align with and expand prior knowledge.

DisCo uncovers heterogeneity in intestinal organoids. Intriguingly, we observed the maintained presence of the *Ly6a*+ stem cell population at S0, S1 and S3. Given that cells with similar

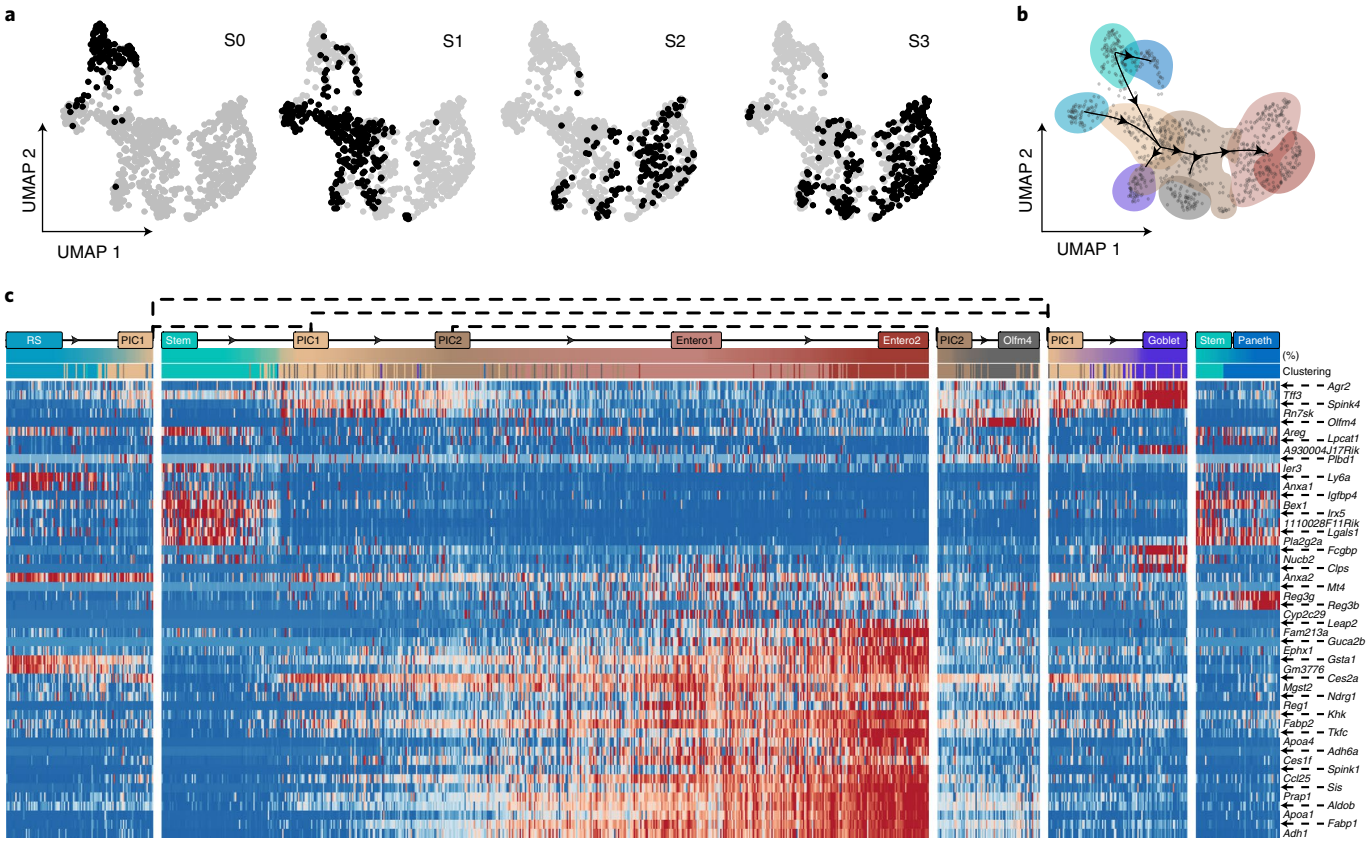


Fig. 3 | Use of DisCo to map intestinal organoid cell identity along development. a, Temporal occurrence of cells. Cells are highlighted on the UMAP embedding according to sampling time point (S0–S3). **b**, Developmental trajectory based on the cluster annotation and the sampling time point derived by slingshot[31]. Cells were annotated in accordance with this clustering. **c**, Heatmap of differentially expressed genes along the waypoints of the trajectory. Waypoints are annotated in accordance with the cell clustering in **b**. RS, regenerative stem cells.

expression signatures were previously described under alternate culture conditions as belonging to a distinct organoid subtype termed "spheroids"[33], we next aimed to verify the presence of such spheroids in the sampled organoids. To do so, we stratified the cells according to the individual organoids from which they were derived by mapping this information onto the reference scaffold (Fig. 4a). We observed strong heterogeneity within single organoids, showing that *Ly6a*$^+$ cells were indeed present in a distinct subset of organoids, predominantly composed of these cells (Fig. 4a S1a, S3e). Furthermore, images obtained prior to dissociation showed that *Ly6a*$^+$ cell-containing organoids (Fig. 4a S3e) had a larger, cystic-like structure (Extended Data Fig. 6a). To confirm the presence of *Ly6a*$^+$ organoids in the cultures, we used an RNA fluorescence in situ hybridization assay, RNAscope (Fig. 4b, with controls in Extended Data Fig. 6b), to localize *Ly6a*, *Muc2* and *Fabp1* expression in organoid sections. These analyses showed canonical budding organoids, containing few *Muc2*$^+$ goblet cells and *Fabp1*$^+$ enterocytes, and *Ly6a*-expressing cells in spherical organoids that did not contain differentiated cell types such as enterocytes or goblet cells. The presence of *Ly6a*$^+$ cells during the first day of sampling suggested that these cells constitute a second, *Lgr5*-independent stem cell population in the organoid culture, as further supported by the trajectory analysis (Fig. 3b,c). To test this, we sorted and differentiated LGR5$^+$LY6A$^-$ cells (3.3% compared with 24.5% of LGR5$^-$LY6A$^+$ and only 0.4% of double positive cells, Fig. 4c), showing that both LGR5$^+$LY6A$^-$ and LGR5$^-$LY6A$^+$ cells can give rise to organoids of similar morphological heterogeneity (Fig. 4d). These results indicate that LGR5$^-$LY6A$^+$ cells have full stem cell potential, similar to that of previously described fetal-like stem cells[33]. Furthermore, the fact that LGR5$^-$LY6A$^+$ cells did not show a propensity towards spheroid formation suggests that environmental conditions (for example, variation in matrix stiffness) rather than the initial cell state dictate the formation of spheroids.

Besides the *Ly6a*$^+$ cell-enriched organoids, the data suggested the presence of additional organoid subtypes in the per-organoid mappings (Fig. 4a). The two most striking additional subtypes were three organoids that contained mostly enterocytes (Fig. 4a S2c, S3a, S3d), and two that consisted predominantly of immature and mature goblet cells (Fig. 4a S1b and especially S2f). The identity of the observed subtypes was further substantiated when visualizing the cell type abundance per organoid (Fig. 4e) and marker gene expression in individual organoids (Extended Data Fig. 7a). Similar to the spheroids, both subtypes had aberrant morphologies, tending to be small and round, as compared with canonical organoids bearing a crypt–villus axis (for example, S3c, Extended Data Fig. 6a). To detect more subtle molecular differences, we used psupertime[34] to identify genes that are dynamically expressed during the development of individual organoids. This analysis showed additional genes that are expressed in subsets of organoids, such as Gastric inhibitory polypeptide (*Gip*), Zymogen granule protein 16 (*Zg16*), Vanin 1 (*Vnn1*) and Defensin alpha 24 (*Defa24*) (Extended Data Fig. 7b).

Although organoids dominated by enterocytes have been previously described as enterocysts[18], organoids displaying goblet cell hyperplasia, here termed 'gobloids', were unknown, to our knowledge. To validate the existence of the uncovered organoid subtypes, we used RNAscope to localize the expression of enterocyte (*Fabp1*) and goblet cell (*Muc2*) markers (Fig. 4f, with controls in Extended

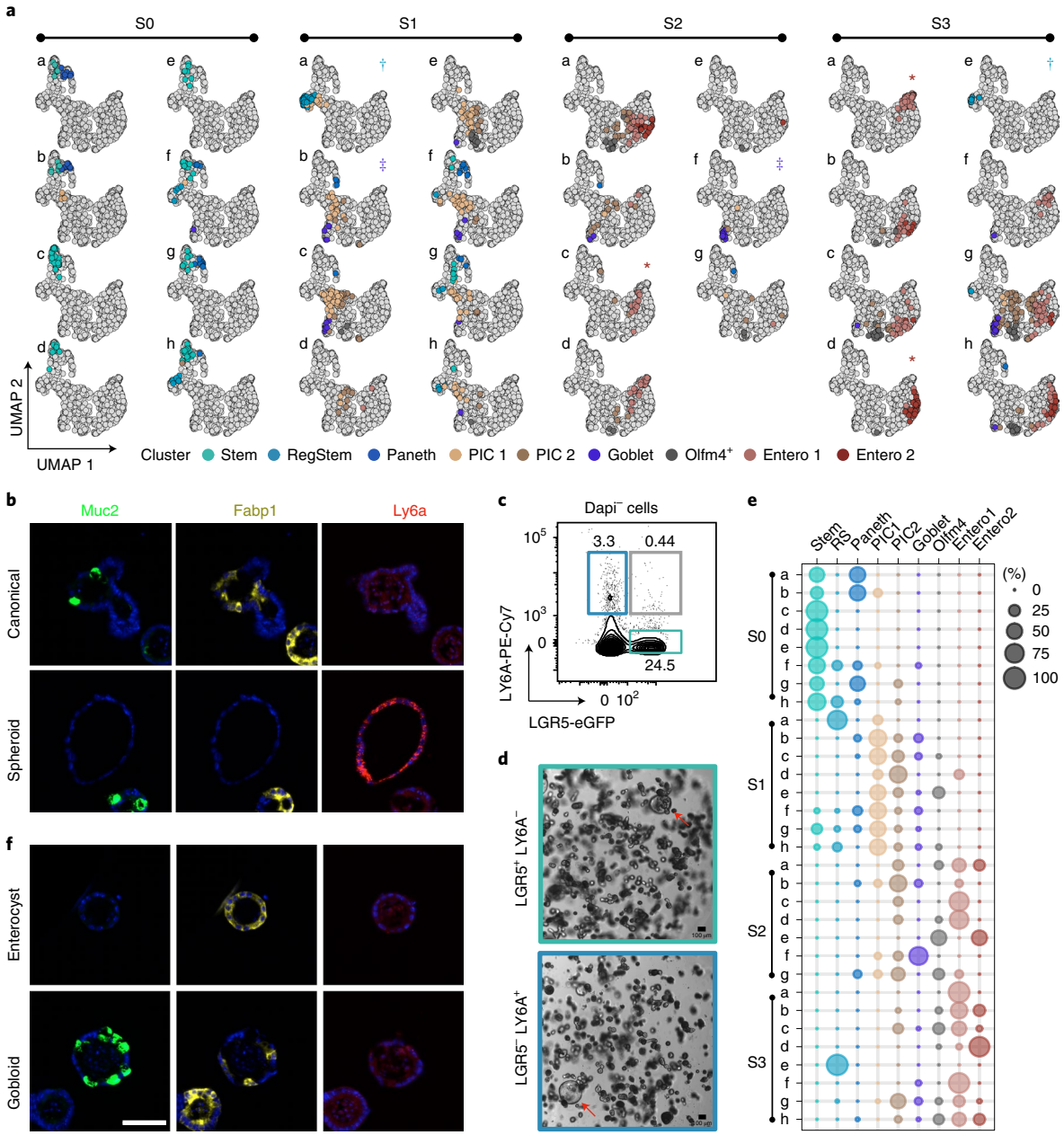


Fig. 4 | Cell type distribution and marker gene expression across individual intestinal organoids. a, Projection of cell types onto 31 individual organoids. Cells per single organoid were colored according to their global clustering and highlighted on the UMAP embedding of all sequenced cells. Projections are grouped according to their sampling time. Manually classified organoids were annotated with the following symbols: *, enterocysts; †, spheroids; ‡, gobloids. **b**, In situ RNA detection of *Ly6a*, *Fabp1* and *Muc2* expression. A representative canonical and *Ly6a*-expressing organoid is displayed. Scale bar (in **f**), 50 µm. Data of one out of two replicates are shown. **c**, Surface LY6A and LGR5-eGFP expression under ENR CV conditions. The dot plot shows LGR5-eGFP and LY6A expression in organoid-derived single-cell suspensions. The numbers indicate frequency (%). **d**, Culturing outcomes of LGR5$^+$ cells and LY6A$^+$ cells. Single LGR5$^+$LY6A$^-$ and LGR5$^-$LY6A$^+$ cells were isolated by FACS and seeded in Matrigel. Cells were cultured as shown in Fig. 2a and imaged using bright-field microscopy at S3. Red arrows point to spheroid morphologies. Scale bars, 100 µm. Data of one out of three replicates are shown. **e**, Dot plot of the distribution of annotated cell types per organoid. Dot size represents the percentage of cells associated with each cluster per organoid. **f**, In situ RNA detection of *Fabp1* and *Muc2* expression. Representative images of the enterocyst and gobloid subtypes are shown. Scale bar, 50 µm.

Data Fig. 6b). In agreement with our data and prior research, we detected organoids that exclusively contained *Fabp1*$^+$ cells, most likely representing enterocysts. Most importantly, we were able to identify organoids that contained a high number of *Muc2*$^+$ goblet cells, confirming the existence of gobloids.

DisCo reveals compositional differences among intestinal crypts. Finally, to complement the intestinal organoid data, we set out to analyze individual crypts that were isolated from the small intestine of adult C57BL/6J mice. However, we found that the dissociation of these crypts into single cells was more challenging than that of in vitro grown organoids, achieving efficiencies of only up to 20% with elevated multiplet rates (Supplementary Table 3 and Methods). In total, we analyzed 21 individual crypts involving 372 cells at a comparable cell recovery efficiency as for organoids (Extended Data Fig. 8a and Supplementary Table 1 for the number of sequenced

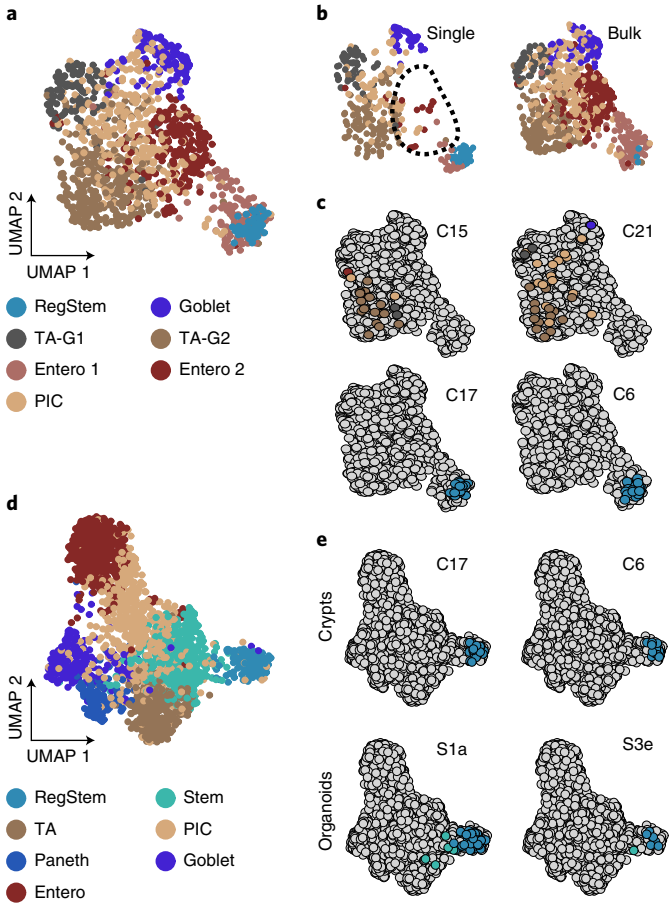


Fig. 5 | Cell type distribution across individual intestinal crypts. a, UMAP embedding of all cells collected from bulk and individual crypts. All 775 processed cells from bulk crypts and 372 cells from individual crypts were clustered with k-means clustering, after which clusters were annotated according to marker gene expression. **b**, UMAP representation of cells derived from bulk or individual crypts. The dotted line highlights the enterocyte cluster. **c**, UMAP from **a** superimposed with cells from exemplary single crypts. **d**, UMAP embedding of all sequenced cells obtained from intestinal organoids (Fig. 2) and crypts. All 2,244 processed cells were clustered with k-means clustering, after which clusters were annotated according to marker gene expression. **e**, UMAP from **d** stratified by exemplary single crypts and organoids that are largely composed of regenerative stem cells. Entero, enterocytes; G1, G1/S cell cycle phase; G2, G2/M cell cycle phase; PIC, potential intermediate cells; RegStem, regenerative stem cells; TA, transit amplifying cells.

cells). Next, we also used DisCo to generate a reference map of 775 cells derived from pooled crypts (bulk), which we integrated with the individual crypt cells to resolve their composition. This enabled us to identify distinct groups of cells: clusters marked by the expression of the cell cycle genes *Orc6* and *Top2a*, suggesting that these represent transit amplifying cells in the G1/S phase and G2/M phase, respectively. We identified two enterocyte clusters, marked by *Fabp1* and *Apoa1* expression and a goblet cell cluster marked by *Muc2* expression (Fig. 5a and Extended Data Fig. 8b). Most of these cell types were observed in bulk as well as in individual crypt samples, except for enterocytes, which were mainly detected in the bulk proportion, probably reflecting contamination by residual villi (Fig. 5b). Globally, the data overlapped with previously reported single cell data from bulk crypts[25], except for the lack of rare enteroendocrine cells and tuft cells, which had an expected abundance of only 1% in bulk crypt isolates[25], and Paneth cells. However, we were able to identify the latter independent from clustering, namely by their gene expression signature (Extended Data Fig. 8c).

Next to the expected cell types, we observed an additional cluster marked by the expression of *Clu* and *Anxa1*, which are established markers of regenerative, or revival stem cells[28]. Interestingly, we found three crypts that contained only these regenerative stem cells and that (providing an accurate compositional representation after dissociation) are thus depleted of other intestinal cell types (Fig. 5c and all crypts in Extended Data Fig. 8d). Given that this observation aligned with our intestinal organoid (spheroid)-based findings, we integrated our crypt data with the previously generated organoid data to explore whether spheroids and crypts contain similar regenerative stem cells. This integration yielded a common dataset of 2,244 cells (Fig. 5d and Extended Data Fig. 9) with overlapping 'regenerative stem cell' clusters, suggesting that this cell state can be recovered in both intestinal crypts and organoids, and that thus spheroids and regenerating crypts are compositionally similar (Fig. 5e and all crypts and organoids in Extended Data Fig. 10). Although caution is warranted when interpreting these results given the encountered dissociation issues, the findings indicate that some organoid heterogeneity recapitulates in vivo tissue heterogeneity, but also that crypts that predominantly contain regenerative stem cells are present in the homeostatic intestine. Altogether, the crypt data support DisCo's capacity to profile in vivo-derived small, individual tissues, rendering the dissociation efficiency, and no longer the processing efficiency, as the overall limiting factor.

Discussion

A key feature of our DisCo approach is the ability to deterministically control the cell capture process. Despite lowering the throughput compared with stochastic droplet systems[2,3], our approach provides the advantage of being able to process low cell input samples at high efficiency and at a strongly decreased per-cell cost (Table 1). Thus, we believe that DisCo is filling an important gap in the scRNA-seq toolbox. Another critical feature of DisCo is the use of machine-vision to obtain full control of the entire bead and cell co-encapsulation process, enabling the correct assembly of most droplets and virtually eradicating confounding factors that arise due to failed co-encapsulation[35,36]. In concept, DisCo is thus fundamentally different to passive particle pairing approaches such as traps[37–39] and, compared with these technologies, offers the advantage of requiring simpler and reusable chips without cell or particle size and shape selection biases[15,40]. This renders the DisCo approach universally applicable to any particle co-encapsulation application[41,42], for example, cell–cell encapsulations, with the only limiting factor being particle visibility.

To demonstrate DisCo's capacity to process small tissues and systems that were so far difficult to access experimentally, we have analyzed the cell heterogeneity of chemosensory organs from *Drosophila melanogaster* larvae[43] and, as shown here, single intestinal organoids and crypts. It is thereby worth noting that, based on our handling of distinct tissues, we found that not DisCo itself, but instead cell dissociation, has become the efficiency-limiting factor (see Methods), a well-recognized challenge in the field[44,45].

scRNA-seq of individual organoids led us to uncover organoid subtypes of aberrant cell type distribution that were previously not resolved with pooled organoid scRNA-seq[18,19,25]. Of particular interest among the identified organoid subtypes is one that we termed 'gobloid' given that it predominantly consists of immature and mature goblet cells. Another subtype contained predominantly cells that were strikingly similar to previously described fetal-like stem cells or revival stem cells that occur during intestinal regeneration[27–29]. This subtype, previously described under alternate culture conditions as spheroid-type organoids[20,33,46], was identified here under standard organoid differentiation conditions, indicating that these organoids are capable of maintaining their unique

state. Moreover, we found that these LGR5$^-$ LY6A$^+$ cells readily give rise to canonical organoids, indicating that they are capable of providing a pool of multipotent stem cells. Interestingly, in our proof-of-principle single intestinal crypt DisCo dataset, we identified crypts that largely consisted of cells with a similar regenerative gene expression signature. Although crypts with these properties have been previously described upon injury, for example, by irradiation[28], our data suggest that such regenerating crypts are also present in the homeostatic intestine.

Here, we demonstrate that our DisCo analysis of individual intestinal organoids and crypts is a powerful approach to explore in vitro and in vivo tissue heterogeneity, and to provide new insights into how this heterogeneity arises. As well as catalyzing research on other tissues or systems of interest, we believe that the technology and findings of this study will contribute to future research on (intestinal) organoid development and thus aid the engineering of more robust organoid systems. Furthermore, we believe that the utility of the approach described here, extends to research on all developing multicellular organisms, and, coupled with lineage tracing[47], might offer an entirely new perspective on interindividual variation. Finally, we expect this approach to be applicable to rare, small clinical samples to gain detailed insights into disease-related cellular heterogeneity and dynamics.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-021-01391-1.

Received: 2 February 2021; Accepted: 22 December 2021;
Published online: 14 February 2022

References

1. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
2. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
3. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
4. Gierahn, T. M. et al. RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).
5. Han, X. et al. Mapping the mouse cell atlas by Microwell-Seq. *Cell* **172**, 1091–1107 (2018).
6. Rosenberg, A. B. et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
7. Li, H. et al. Fly cell atlas: a single-cell transcriptomic atlas of the adult fruit fly. Preprint at *bioRxiv* https://doi.org/10.1101/2021.07.04.451050 (2021).
8. Tabula Muris Consortium Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
9. Han, X. et al. Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
10. Stoeckius, M. et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
11. McGinnis, C. S. et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* **16**, 619–626 (2019).
12. Gehring, J., Hwee Park, J., Chen, S., Thomson, M. & Pachter, L. Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins. *Nat. Biotechnol.* **38**, 35–38 (2020).
13. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 1–14 (2018).
14. 10X Genomics. *Chromium Single Cell 3′ Reagent Kits User Guide (v3 Chemistry)* (CG000183 Rev C) (2018).
15. DeLaughter, D. M. The use of the Fluidigm C1 for RNA expression analyses of single cells. *Curr. Protoc. Mol. Biol.* **122**, e55 (2018).
16. Wagner, D. E. et al. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
17. Packer, J. S. et al. A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science* **365**, eaax1971 (2019).
18. Serra, D. et al. Self-organization and symmetry breaking in intestinal organoid development. *Nature* **562**, 66–72 (2019).
19. Grün, D. et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).
20. Lukonin, I. et al. Phenotypic landscape of intestinal organoid regeneration. *Nature* **586**, 275–280 (2020).
21. Tirier, S. M. et al. Pheno-seq: linking visual features and gene expression in 3D cell culture systems. *Sci. Rep.* **9**, 12367 (2019).
22. Unger, M. A., Chou, H. P., Thorsen, T., Scherer, A. & Quake, S. R. Monolithic microfabricated valves and pumps by multilayer soft lithography. *Science* **288**, 113–116 (2000).
23. Biocanin, M., Bues, J., Dainese, R., Amstad, E. & Deplancke, B. Simplified Drop-seq workflow with minimized bead loss using a bead capture and processing microfluidic chip. *Lab Chip* **19**, 1610–1620 (2019).
24. Sato, T. et al. Single Lgr5 stem cells build crypt–villus structures in vitro without a mesenchymal niche. *Nature* **459**, 262–265 (2009).
25. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
26. Yin, X. et al. Niche-independent high-purity cultures of Lgr5+ intestinal stem cells and their progeny. *Nat. Methods* **11**, 106–112 (2014).
27. Gregorieff, A., Liu, Y., Inanlou, M. R., Khomchuk, Y. & Wrana, J. L. Yap-dependent reprogramming of Lgr5+ stem cells drives intestinal regeneration and cancer. *Nature* **526**, 715–718 (2015).
28. Ayyaz, A. et al. Single-cell transcriptomes of the regenerating intestine reveal a revival stem cell. *Nature* **569**, 121–125 (2019).
29. Roulis, M. et al. Paracrine orchestration of intestinal tumorigenesis by a mesenchymal niche. *Nature* **580**, 524–529 (2020).
30. Battich, N. et al. Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies. *Science* **367**, 1151–1156 (2020).
31. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
32. Birchenough, G., Johansson, M., Gustafsson, J., Bergstrom, J. & Hansson, G. C. New developments in goblet cell mucus secretion and function. *Mucosal Immunol.* **8**, 712–719 (2015).
33. Yui, S. et al. YAP/TAZ-dependent reprogramming of colonic epithelium links ECM remodeling to tissue regeneration. *Cell Stem Cell* **22**, 35–49 (2018).
34. Macnair, W. & Claassen, M. psupertime: supervised pseudotime inference for single cell RNA-seq data with sequential labels. Preprint at *bioRxiv* https://doi.org/10.1101/622001 (2019).
35. Lareau, C. A., Ma, S., Duarte, F. M. & Buenrostro, J. D. Inference and effects of barcode multiplets in droplet-based single-cell assays. *Nat. Commun.* **11**, 866 (2020).
36. Lun, A. T. L. et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).
37. Chung, M., Núñez, D., Cai, D. & Kurabayashi, K. Deterministic droplet-based co-encapsulation and pairing of microparticles via active sorting and downstream merging. *Lab Chip* **17**, 3664–3671 (2017).
38. Cheng, Y. H. et al. Hydro-Seq enables contamination-free high-throughput single-cell RNA-sequencing for circulating tumor cells. *Nat. Commun.* **10**, 2163 (2019).
39. Zhang, M. et al. Highly parallel and efficient single cell mRNA sequencing with paired picoliter chambers. *Nat. Commun.* **11**, 2118 (2020).
40. Pollen, A. A. et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32**, 1053–1058 (2014).
41. Dura, B. et al. Profiling lymphocyte interactions at the single-cell level by microfluidic cell pairing. *Nat. Commun.* **6**, 5940 (2015).
42. Gérard, A. et al. High-throughput single-cell activity-based screening and sequencing of antibodies using droplet microfluidics. *Nat. Biotechnol.* **38**, 715–721 (2020).
43. Maier, G. L. et al. Multimodal and multisensory coding in the Drosophila larval peripheral gustatory center. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.21.109959 (2020).
44. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9**, 75 (2017).
45. Denisenko, E. et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* **21**, 130 (2020).
46. Mustata, R. C. et al. Identification of Lgr5-independent spheroid-generating progenitors of the mouse fetal intestinal epithelium. *Cell Rep.* **5**, 421–432 (2013).
47. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).

Methods

Animal experimentation. Intestinal organoids were generated from mouse small intestinal stem cells isolated from 5–10-week-old heterozygous, male/female, Lgr5-eGFP-IRES-CreERT2 mice (Jackson Laboratory). These experiments were approved by the Service de la Consummation et des Affaires Vétérinaires in Epalinges, Switzerland (license number 2681.0). Intestinal crypts were isolated from 7-week-old male C57BL/6J mice. Crypt isolation was conducted under the license VD3406e granted by the local authorities: Direction Générale de l'Agriculture, de la Viticulture et des Affaires Vétérinaires in Epalinges, Switzerland. Mice were housed at a temperature of 22 °C ± 2 °C, relative humidity of 55% ± 10%, and a 12 h–12 h light–dark cycle (light at 7:00 and dark at 19:00).

System comparison metrics. Performance metrics for Table 1 were calculated in the following ways.

Minimum cell input estimates. The minimum cell input values were derived from the following sources: for 10X Chromium HT/LT[48], the lowest cell input number given in the 10X Chromium manual (HT, CG000183 Rev C; LT, CG000399 Rev B); for inDrop[3], the lowest numbers mentioned in the 1CellBio manual (Single Cell Encapsulation Protocol, version 2.4); for Drop-seq[2], the lowest numbers used in ref. [49] (it is probable that lower cell numbers can be processed, but Drop-seq has been suggested to be used "When the sample is abundant"[49]); for FACS-based methods[50,51], the input limits as described in ref. [13]; for Fluidigm C1 (ref. [40]), the lowest cell input number used for benchmarking experiments given in the present study; for Wafergen iCell8 (ref. [52]), the lowest cell numbers derived from the iCell manual (CELL8 Single-Cell ProtocolD07-000025 Rev. C; according to the manual, 80 µl of a suspension of 0.02 cells per nl are prepared for dispensing); for Seq-well[4], the lowest cell number used for capture in ref. [4]; and for DisCo, the lowest cell number processed in this study.

Efficiency estimates. Efficiency estimates were derived from varying sources and represent different efficiencies. The efficiencies for 10X Chromium HT, inDrop and Drop-seq were derived from ref. [49] from quantified cellular inputs (>1,000 cells) and sequenced cells passing the quality thresholds. Given that these efficiencies stem from experiments that were performed with optimized cell inputs, we can assume lower efficiencies when processing low cell inputs (<1,000). For the 10X Chromium LT kit, efficiencies were derived from the user manual CG000399 Rev B. The efficiency for the Fluidigm C1 system was determined in the present study (Extended Data Fig. 1l). For the Wafergen iCell8 system, an efficiency estimate was derived from ref. [53] and represents the efficiency of conversion from captured to sequenced cells passing the quality thresholds, thus it does not include cell capture inefficiency. The efficiency for Seq-well was derived from ref. [4] at an input of 400 cells and represents an inferred efficiency from quantified cell input to sequenced cells passing the quality threshold. Specifically, the library conversion efficiency, that is, the percentage of captured cells identified in the sequencing data passing the quality threshold, was calculated based on the species mixing experiment involving 10,000 input cells. The library conversion efficiency, in combination with capture efficiencies at 400 cells, was used to determine the efficiency at low cell numbers. Hence, this is inferred from quantified cellular inputs to sequenced cells passing the quality thresholds. The efficiencies for DisCo were derived in this study and represent mean efficiencies for low cell inputs (50–200), from quantified cell input to sequenced cells passing the quality thresholds.

Cost per cell estimates. Two cost estimates are listed for 100 cells: the cost for 100 cells not considering system efficiencies (US$ per cell, 100 output cells), and the cost for 100 input cells considering the listed efficiencies (US$ per cell, 100 input cells). The run costs for Smart-seq2, Cel-seq2, inDrop, Drop-seq and Seq-well were derived from Supplementary Table 8 in ref. [54]. The run costs for 10X Chromium HT, Fluidigm C1 (96) and Wafergen iCell8 were derived from table 2 in ref. [53]. The costs for 10X Chromium LT were derived from the 10X price list of the École Polytechnique Fédérale de Lausanne (EPFL) Gene Expression Core Facility (GECF). For the Wafergen iCell8 it was assumed that eight samples (one per dispensing nozzle) can be processed on one chip in parallel, thus decreasing the cost by a factor of 8. The DisCo cost estimate includes reagents for library generation, that is, the cost for beads, oil, reverse transcription reaction, exonuclease treatments, polymerase chain reaction (PCR), and library preparation (Nextera XT).

DisCo experimental procedure. The experimental setup and all necessary steps are described in the step-by-step protocol: https://www.epfl.ch/labs/deplanckelab/wp-content/uploads/2021/09/DisCo_protocol.pdf

Machine-vision software. The software for cell detection and coordination was implemented in C++. Camera images were obtained with the XiApi library (version 4.15). Images were processed in real time using the OpenCV computer vision library (version 3.4). A schematic visualization of the particle detection algorithm is given in Extended Data Fig. 1b. In brief, a detection region of interest (ROI) was extracted by cropping, after which a Gaussian blur was applied to the resulting image. Two subsequent images were subtracted, and the resulting image converted to a binary image by intensity thresholding. The binary image was dilated to fill potential holes. Finally, contours were detected using the findContours function, and classified for area and circularity. Upon particle detection, the particles were properly positioned by valve oscillation and monitoring of the ROI at the target zone (Extended Data Fig. 1c). Once two particles were positioned in their respective target zones, particles were co-ejected by pressurization of the dropleting valve, and the droplet was sheared by actuation of the oil valve.

Microfluidic chip design and fabrication. The design of the microfluidic chip for deterministic co-encapsulation is presented in Extended Data Fig. 1a and the computer-aided design (CAD) files are available in Supplementary Data 1. Chips were designed using Tanner L-Edit CAD software (Mentor, version 2016.2). The 5-inch chromium masks were exposed in a VPG200 laser writer (Heidelberg instruments) for both the control and flow layers. Masks were developed using an HMR 900 mask processor (Hamatech). For the control layer, a thick SU8 photoresist layer was deposited with an LSM-200 spin coater (Sawatec), exposed on an MJB4 single side mask aligner (SÜSS MicroTec), and manually developed. The SU8 processing steps were carried out according to the manufacturer's instructions for the 3010 series (Microchem). For the flow layer, wafers were produced using the AZ40XT (Microchem) positive photoresist on the ACS200 coating and developing system (Gen3, SÜSS MicroTec). Developed master-wafers were reflowed for 45–75 s at 120 °C on a hotplate until the channels appeared round under an inspection microscope. The control layer master-wafers were used as molds for polydimethylsiloxane (PDMS) chips after passivation with 1% silane dissolved in hydrofluoroether (HFE). For the flow layer, master-wafers were used to generate replica molds for chip production. To this end, the primary replica mold was obtained by mixing PDMS with curing agent at a ratio of 10:1 using a centrifugal mixer (Thinky), degassing for 15 min, and curing for 60 min at 80 °C. The PDMS-based primary replica mold was then sylanized and subsequently used to obtain secondary replica molds for PDMS flow layer production. The PDMS flow layer was fabricated using PDMS and curing agent at a ratio of 5:1, degassed and cured at 80 °C for 30 min. The control layer was fabricated by spin coating the PDMS and curing agent at a ratio of 20:1 on the flow layer wafer at 650 r.p.m. for 35 s with a 15 s ramp time, followed by baking at 80 °C for 30 min. Cured PDMS was then cut from the flow layer secondary replica mold, and flow layer inlet holes were punched with a 0.5 mm diameter biopsy punch. The two PDMS layers were manually aligned and bonded at 80 °C for at least 60 min. Assembled and cured PDMS chips were cut from the molds, and the control layer inlet holes were punched. Finally, chips were oxygen plasma activated (45 s at ~500 mTorr O_2) and bonded to a surface-activated glass slide followed by incubation at 80 °C for at least 2 h. Materials and reagents are listed in the Material and reagent list, point 1.

Mammalian cell culture handling for the species mixing experiment. For benchmarking the DisCo platform, HEK 293T cells (American Type Culture Collection (ATCC) cat. no. SD-3515) and IBA cells (provided by C. Wolfrum's laboratory, ETH Zürich) were used. Cells were cultured to 90% confluency in Glutamax DMEM supplemented with FBS and penicillin–streptomycin. Prior to use, the cells were washed with PBS, dissociated with Trypsin-EDTA, washed with cell wash buffer and counted with Trypan blue live–dead stain using a Countess cell counter (Invitrogen). Cells were mixed in a 1:1 ratio, adjusted to 20 cells per µl, resuspended in cell loading buffer, and finally loaded on the DisCo chip. Material and reagents are listed in the Material and reagent list, point 11.

Droplet content and co-encapsulation performance quantification. As for conventional DisCo runs, experiments were set up with Chemgen beads and varying concentrations of HEK 293T cells. Approximately 100 co-encapsulations were performed and recorded per condition. The recorded video data were manually reviewed and the droplet contents, and passing cells were counted (Fig. 1f).

Benchmarking DisCo efficiency using the DISPENCELL platform. To benchmark single-cell recovery efficiencies throughout the complete DisCo workflow, we quantified HEK 293T cells (ATCC cat. no. SD-3515) using the DISPENCELL pipetting robot (SEED Biosciences SA). Prior to use, HEK 293T cells were diluted to 20 cells per µl. Cells were loaded into the DISPENCELL tip and then dispensed directly into a Prot/Elec gel loading tip containing cell loading buffer. Cells were then processed with DisCo and the libraries prepared as described above.

Benchmarking the Fluidigm C1. To benchmark the cell recovery efficiency of the Fluidigm C1 platform, HEK 293T cells (ATCC cat. no. SD-3515) were diluted with Suspension Reagent (Fluidigm) to reach approximately 10, 20 and 40 cells per µl. The obtained suspensions were generated separately from the same stock and then quantified in triplicate using microscopy by examining a volume of 2 × 2.5 µl of the suspension between two coverslips. Counts of all triplicates were averaged to determine the cell input for 5 µl Cell Mix, and the same volume was subsequently loaded on the C1 integrated fluidic circuit (IFC). The experiments on the C1 machine were performed according to the SMART-Seq v4 Ultra Low Input RNA

Kit for the Fluidigm C1 System, IFCs User Manual (Clontech Laboratories) using 10–17 µm 96-trap C1 IFC OpenApp chips. The protocol was run on SMART-Seq version 4 (1861x/1862x/1863x) programs on the C1 machine. To verify successful loading and cell trapping, traps were examined using a Cell xCellence (Olympus) microscope. Final cDNA was quantified using the PicoGreen double-stranded (ds) DNA assay and then fragmented and tagged (that is, tagmented) using the Nextera XT library preparation kit according to the manufacturer's instructions.

Material and reagent list for Fluidigm C1 benchmarking. For single-cell chip loading and priming, the C1 Single-Cell mRNA Seq HT Reagent kit version 2 (Fluidigm, cat. no. SKU 101-3473) was used as well as the 10–17 µm 96-trap C1 IFC (Fluidigm, cat. no. SKU 100-8134). For cDNA generation, a SMART-Seq v4 Ultra Low Input RNA Kit (Clontech Laboratories, cat. no. 634888) was used. cDNA was quantified using the PicoGreen high-sensitivity dsDNA assay (Invitrogen, cat. no. P11496) and then libraries were generated using Nextera XT (Illumina, cat. no. FC-131-1096) and the Nextera XT index kit set A (Illumina, cat. no. FC-131-2001).

Temporal batch effect experiment. A single-cell suspension of HEK 293T (ATCC cat. no. SD-3515) was loaded on the system as described above, and the remaining volume stored on ice. After 20 min, generated droplets were evacuated from the system and sequencing libraries prepared. A new cell loading tip was inserted into the sample outlet port and the experimental run was resumed. The previous steps were repeated after 40 and 60 min. After 120 min the system was loaded with cells stored on ice and the cells were captured for approximately 20 min. Subsequently, droplets were evacuated from the system, cDNA was generated, and sequencing libraries prepared as described above. The former steps were repeated for cells stored for 180 min on ice.

The material and reagent list for the temporal batch effect experiment is given in the section 'Mammalian cell culture handling for the species mixing experiment'.

Mouse intestinal organoid culture and handling. The isolation of stem cells harboring an enhanced green fluorescent protein expressed from the *Lgr5* locus (Lgr5-eGFP stem cells) from 5- to 10-week-old Lgr5-eGFP-IRES-CreERT2 mice (Jackson Laboratory) and initial culture were performed as previously described[55]. For the developmental time-course experiments, organoids were dissociated to single cells, live LGR5$^+$-eGFP cells were isolated using a FACS ARIA II (BD Biosciences) and were embedded in Matrigel. For sorting of LY6A$^+$ LGR5$^-$ and LY6A$^-$ LGR5$^+$, cells were surface stained with anti-mouse Anti-Sca1/Ly6A/E (0.2 µg per 1 million cells; Biolegend, cat. no. 122514) for 30 min in culture medium at 4 °C. Subsequently, cells were washed by centrifugation in a 4 °C cooled centrifuge for 5 min at 774 ×*g*, and resuspended in organoid culture medium. Live–dead discrimination was carried out using 4,6-diamidino-2-phenylindole (DAPI) added shortly before sorting. The gating strategy is shown in Supplementary Fig. 1.

After Matrigel polymerization, cells were cultured in ENR CV medium (epidermal growth factor (E), Noggin mouse (mNoggin; N), R-spondin (R) and CV) supplemented with the ROCK inhibitor, thiazovivin. Growth factors (E, N, R, C, V) were replenished after 2 days of culture. At day 3 of culture, a full medium change was performed for the differentiation growth medium (ENR only). At day 5, growth factors (E, N, R) were replenished. Organoids were sampled at day 3 (S0) prior to the medium change, at day 4 (S1), at day 5 (S2) and at day 6 (S3).

Single organoids were collected by dissolving Matrigel with ice-cold Cell Recovery Solution for approximately 5 min while carefully pipetting up and down with a 1,000 µl pipette. Subsequently, single organoids were isolated by hand-picking, after which they were transferred to a Nunc microwell culture plate with single-organoid dissociation mix. Single organoids were dissociated by combining trituration using siliconized pipette tips every 5 min and incubation at 37 °C for 15 min. Following dissociation, cell suspensions were diluted in cell loading buffer in the loading tip connected to the DisCo chip. Materials and reagents are listed in the Material and reagent list, points 12–16.

Intestinal organoids were cultured in Matrigel (Corning, cat. no. 356230) with organoid base medium (described in point 13) supplemented with ENR (and CV where indicated) and ROCK inhibitor (where indicated; Sigma, cat. no. Y0503). Organoid base medium was prepared using DMEM/F12 (Gibco, cat. no. 11320033), 100 mM HEPES (Gibco, cat. no. 15630056), 100 U per ml penicillin–streptomycin (Gibco, cat. no. 15140122), 1 µM B27 supplement (Gibco, cat. no. 17504-044), 1 µM N2 supplement (Gibco, cat. no. 17502001) and 1 µM *N*-acetyl-L-cysteine (Sigma, cat. no. A9165). ENR medium was prepared using base medium (as above), 50 ng per ml epidermal growth factor (E; LifeTechnologies, cat. no. PMG8043), 100 ng per ml mNoggin (N; produced in-house) and 1 µg per ml R-spondin (R; produced in-house). ENR CV medium was prepared with the addition of 3 µM CHIR99021 (C; CalBiochem) and 3 mM valproic acid (V; Sigma, cat. no. P4543) to the ENR medium. Single-organoid, single-cell dissociation mix was prepared using PBS (Gibco, cat. no. 14190-094), 10 mg per ml *Bacillus licheniformis* protease (Sigma, cat. no. P5380), 5 mM EDTA (Sigma, cat. no. 03690), 5 mM EGTA (BioWorld, cat. no. 40520008-1), 10 µg per ml DNase I (Roche, cat. no. 11 284 932 001) and 0.68X Accutase (Sigma, cat. no. A6964) in a total volume of 20 µl per reaction. For single-organoid dissociation, Nunc MicroWell plates (cat. no. 438733) and siliconized p10 pipette tips (VWR, cat. no. 53509-134) were used.

Split organoid experiment. Organoids for the split organoid experiment were cultured in ENR medium as previously described[20]. Single organoids, derived from days 2–6 after crypt splitting, were isolated from Matrigel as described above. Subsequently, single organoids were isolated by hand-picking into a 384-well plate containing single-organoid dissociation mix. As before, single organoids were dissociated by combining triaturation using non-filter pipette tips every 5 min and incubation at 37 °C for 15 min in a 100 µl volume. Finally, the dissociation mix was diluted with cell loading buffer. The single-cell suspension of one organoid was split into two separate samples and introduced subsequently on the system.

Material and reagent list for the split organoid experiment. Intestinal organoids were cultured in Matrigel (Corning, cat. no. 356230) with organoid base medium supplemented with ENR (not containing CV). Organoid base medium was prepared using DMEM/F12 (Gibco, cat. no. 11320033), 100 mM HEPES (Gibco, cat. no. 15630056), 100 U per ml penicillin–streptomycin (Gibco, cat. no. 15140122), 1 µM B27 supplement (Gibco, cat. no. 17504-044), 1 µM N2 supplement (Gibco, cat. no. 17502001) and 1 µM *N*-acetyl-L-cysteine (Sigma, cat. no. A9165). ENR medium was prepared using base medium (as above), 50 ng per ml epidermal growth factor (E; LifeTechnologies, cat. no. PMG8043), 100 ng per ml mNoggin (N; produced in-house) and 1 µg per ml R-spondin (R; produced in-house).

Single-cell isolation from the small mouse intestine. Crypts were isolated from the small intestines of single 7-week-old male C57BL/6J mice following the protocol in ref. [56]. In brief, the small intestine of a single mouse was isolated and then washed both on the inside and outside with ice-cold PBS. The small intestine was cut open longitudinally and washed again with PBS. The intestine was then digested non-enzymatically for 3 min in PBS, EDTA and dithiothreitol (DTT). Next, the tissue was cut into small pieces and transferred into a 50 ml Falcon tube containing 20 ml ice-cold PBS. The PBS solution containing the tissue pieces was gently triturated 10 times using a 10 ml pipette. After tissue fragments had sedimented, the supernatant was removed and the process was repeated three more times until the supernatant was clear. Next, the supernatant was removed, PBS and EDTA were added and the sample incubated for 30 min at 4 °C on a rocking plate. After incubation, tissue fragments were left for sedimentation (up to 5 min), then the supernatant was removed. Subsequently, tissue fragments were triturated with ice-cold PBS by pipetting up and down. After large tissue fragments had sedimented (up to 5 min), the supernatant containing crypts was collected as fraction 1 (F1). Fraction collection was then repeated four times (F2–F5), followed by trituration with ice-cold PBS, while each fraction was stored separately. Each fraction was inspected for cell debris and villus contamination.

For single-cell bulk sample preparation, crypts from F2 or F3 were spun down at 600 ×*g* for 10 min (brake 5). Following centrifugation the supernatant was removed and the cells were enzymatically dissociated for 1 min at 37 °C. Cells were then washed twice in PBS containing 0.01% BSA and strained twice using a Flowmi 40 µm strainer to minimize the amount of multiplets. Cell suspensions were diluted in cell loading buffer and loaded on the DisCo chip.

For single-cell isolation from single crypts, crypts from F3 were transferred to FBS-coated 6-well plates. Subsequently, single crypts were isolated by hand-picking, after which they were transferred to a Nunc microwell culture plate containing single crypt dissociation mix. Single crypts were dissociated by combining trituration using non-filter pipette tips every 5 min and incubation at 37 °C for a total of 15 min. As also noted in the Results section, obtaining a true single-cell suspension proved highly challenging, despite testing several dissociation buffer compositions (Supplementary Table 3), given that many cells were lost or were only partially recovered in multiplets or clumps. Following dissociation, cell suspensions were diluted in cell loading buffer and loaded on the DisCo chip.

Material and reagent list for single-cell isolation from mouse intestinal crypts. For small intestine washing, PBS (Gibco, cat. no. 14190-094) was used. For the non-enzymatic dissociation of small intestinal pieces, PBS (Gibco, cat. no. 14190-094), 3 mM EDTA (Sigma, cat. no. 3690) and 0.5 mM DTT (Applichem, cat. no. A2948,0005) were used. Intestinal pieces were incubated in PBS (Gibco, cat. no. 14190-094) and 2 mM EDTA (Sigma, cat. no. 3690), followed by fraction collection in PBS (Gibco, cat. no. 14190-094). Bulk single-cell crypt preparations from the small intestine were prepared using PBS (Gibco, cat. no. 14190-094), 1X TrypLE select (Gibco, cat. no. A1217701) and 10 mg per ml DNase I (Roche, cat. no. 11 284 932 001) in a total volume of 500 µl per reaction. Cells were then washed using PBS (Gibco, cat. no. 14190-094) and 0.01% BSA (Sigma, cat. no. B8667), and strained using a Flowmi 40 µm strainer (Sigma, cat. no. BAH136800040-50EA) into a 500 µl final volume. Single cells were dissociated from single crypts using PBS (Gibco, cat. no. 14190-094), 20 mg per ml *B. licheniformis* protease (Sigma, cat. no. P5380), 10 mM EDTA (Sigma, cat. no. 03690), 10 mM EGTA (BioWorld, cat. no. 40520008-1), 20 µg per ml DNase I (Roche, cat. no. 11 284 932 001) and 0.6X Accutase (Thermo Fisher, cat. no. A1110501) in a total volume of 20 µl per reaction. For single crypt dissociation, Nunc MicroWell plates (cat. no. 438733) and non-filtered 10 µl pipette tips (VWR, cat. no. 53509-134) were used.

RNA fluorescence in situ hybridization (RNAscope) on intestinal organoids. For the RNAscope assay, organoids in Matrigel were fixed in paraformaldehyde

(PFA) at 4 °C overnight. The next day, organoids were washed with PBS and embedded in histogel. Histogel blocks were subsequently infiltrated with paraffin using a standard histological procedure (VIP6, Sakura). RNAscope Multiplex Fluorescent V2 assay was performed according to the manufacturer's protocol on 4 µm paraffin sections, hybridized with the probes Mm-Ly6a-C2, Mm-Fabp1-C1, Mm-Muc2-C2, Mm-PpiB-C2 positive control and Duplex negative control at 40 °C for 2 h and revealed with TSA Opal650 for the C1 channel and TSA Opal570 for the C2 channel. Tissues were counterstained with DAPI and mounted with Prolong Diamond Antifade Mountant. Slides were imaged on an Olympus VS120 whole slide scanner (Olympus). The resulting images were converted to the TIFF file format using the Fiji (version 1.52p) plugin BIOP VSI Reader (version 7). ROIs were extracted using a custom Python (version 2.7.15) script and the PIL library (version 6.2.2). Brightness of the extracted ROIs was adjusted in Fiji: images of one target were loaded, stacked, and the brightness adjusted for the whole stack using the setMinAndMax() function. Finally, images were unstacked, merged with other channels and exported as PNG files. Materials and reagents are listed in the Material and reagent list, points 17, 18.

Sequencing, analysis, barcode correction. The data analysis was performed using the Drop-seq tools package (version 2.3.0, https://github.com/broadinstitute/Drop-seq/releases/tag/v2.3.0)[2,57] on the EPFL Scientific IT and Application Support (SCITAS) High Performance Computing (HPC) platform. All data pre-processing steps were done according to the Drop-seq tools manual, except for the DetectBeadSubstitutionErrors function, which was replaced by the barcode merging strategy described below. After trimming and sequence tagging, reads were aligned to the human (hg38), mouse (GRCm38), or mixed reference genomes[2] (GSE63269), depending on the origin of the cellular input material, using STAR (version 2.7.0.e)[58]. Following alignment, BAM files were processed to obtain initial read-count matrices (RCMs) per sample (note: DGE summary files were used for the experiments in Fig. 1h,i). Cell barcodes were prefiltered at >35 UMIs (for the species mixing experiment, the sum of 35 UMIs for both species was used as a prefiltering criterion). Graphs were built by identifying barcodes connected by Levenshtein distance 1. For each graph, the barcode containing the highest number of UMIs was identified as the central barcode. The graphs were pruned (barcodes removed) at a Levenshtein distance >2 to the central barcode, and the remaining barcodes in the graph were merged. The script for barcode merging is available in Supplementary Software 1.

For cell recovery efficiency experiments using the DISPENCELL platform (Fig. 1j) and for Drop-seq comparison experiments (Fig. 1i), barcodes encompassing at least 500 UMIs were compiled into the RCMs. Additionally, for the Drop-seq comparison experiments, the processed BAM files were down-sampled to the same read depth using samtools (version 1.9) (http://www.htslib.org/doc/samtools.html).

Time course organoid kinetic analysis. RCMs were further processed via R (version 3.6.2) using Seurat (version 3.1.1) and uwot (version 0.1.3)[59]. For each individual organoid-RCM, cells with >800 features and <7.5% mitochondrial reads were retained in the analysis. The time course kinetics of organoids were processed in three independent experiments, which were considered as three individual batches. The three independent experiments were merged using FindIntegrationAnchors(list(experimental_batches), anchor.features = 80, dims = 1:12, k.filter = 200, k.anchor = 8) and IntegrateData(). Data were scaled and the principal components (PCs) computed using default settings. UMAP dimensional reduction via RunUMAP() and FindNeighbors() was performed using the first 12 principal component analysis (PCA) dimensions as input features. FindClusters() was computed at a resolution of 0.75.

The intestinal organoids for the split organoid experiment were processed in four independent experiments, which were considered as four individual batches, each encompassing at least two independent single intestinal organoids. The four independent experiments were merged using FindIntegrationAnchors(list(experimental_batches), anchor.features = 120, dims = 1:10, k.filter = 100, k.anchor = 12) and IntegrateData(). Data were scaled and the PCs computed using default settings. UMAP dimensional reduction via RunUMAP() and FindNeighbors() was performed using the first 14 PCA dimensions as input features. FindClusters() was computed at a resolution of 0.9.

The intestinal crypts were processed in five independent experiments, which were considered as five individual batches each encompassing single intestinal crypts and pooled (bulk) samples. The five independent experiments were merged using FindIntegrationAnchors(list(experimental_batches), anchor.features = 150, dims = 1:10, k.filter = 150, k.anchor = 10) and IntegrateData(). Data were scaled and the PCs computed using default settings. UMAP dimensional reduction via RunUMAP() and FindNeighbors() was performed using the first 15 PCA dimensions as input features. FindClusters() was computed at a resolution of 0.8.

Combined intestinal crypts and organoids were processed as eight independent batches. These eight batches were merged using FindIntegrationAnchors(list(experimental_batches), anchor.features = 150, dims = 1:15, k.filter = 150, k.anchor = 10) and IntegrateData(). Data were scaled and the PCs computed using default settings. UMAP dimensional reduction via RunUMAP() and FindNeighbors() was performed using the first 15 PCA dimensions as input features. FindClusters() was computed at a resolution of 0.9. Merging retained the global grouping of the data but introduced minor annotation discrepancies in similar clusters between the individual and merged datasets. For example, cells that were annotated TA-G1 in the crypt data (Extended Data Fig. 8d) were annotated as stem cells in the merged data (Extended Data Fig. 10).

Merged data were visualized using the Seurat intrinsic functions VlnPlot(), FeaturePlot(), DotPlot() and DimPlot(). Differentially expressed genes per cluster were identified using FindAllMarkers() using default parameters. The Seurat object is accessible via GSE148093. Cumulative Z-scores were calculated based on the scaled expression per cell across the defined gene signatures[18,27]. Pie chart, bubble plot and bar graph visualizations were carried out with ggplot2. The analysis script is available in Supplementary Software 1.

C1 HEK library processing and analysis. Libraries were sequenced on a NextSeq500 sequencer (Illumina) in paired-end run format (read 1, 16 bp; read 2, 59 bp) with an average of 3×10^6 reads per library. The read quality of sequenced libraries was evaluated with FastQC. Sequencing reads were aligned to the reference human genome assembly GRCh38.90 using STAR[58]. Reads aligned to annotated genes were quantified with htseq-count[60].

Slingshot analysis. The trajectories were constructed using the Slingshot wrapper implemented in the dyno package (https://github.com/dynverse/dyno)[61]. The method was provided with the first five dimensions of a multi-dimensional scaling as dimensionality reduction, the clustering as described earlier, and the stem cell cluster as the starting cell population. All other parameters were left as the default settings. Genes that change along the trajectory were ranked using the calculate_overall_feature_importance function from the dynfeature package (version 1.0, https://github.com/dynverse/dynfeature), and the top 50 differentially expressed genes were selected. The dynplot package (version 1.1, https://github.com/dynverse/dynplot) was used to plot the trajectory within a scatterplot and heatmap.

Psupertime analysis. Cell labels and sample-day labels were extracted from the merged and batch-corrected meta-data of the Seurat object to run psupertime, a method of identifying genes relevant to biological processes using cell-level temporal labels to build an L1 regularized ordinal logistic regression model[34]. Sample-day labels indicating the experimental temporal order were used to conduct a psupertime analysis on batch-corrected and normalized gene expression data of cells, with selected cell type labels. The analysis was performed including all genes and encompassing a 10-fold cross-validation using default settings. Genes with coefficients (beta-values) greater than zero were considered relevant for the temporal expression dynamics. Expression of relevant genes was plotted per organoid per cell.

Material and reagent list for all experiments. Material information is listed in the following format: material name (vendor, ordering number). Reagent information is listed in the following format: reagent name (final concentration in the solution, vendor, order number).

1. For microfluidic device fabrication, SU8 3010 (Microchem) negative photoresist, AZ40XT (Microchem) positive photoresist, HFE-7500 (3 M, Novec 297730-93-9), Trichloro(1H,1H,2H,2H-perfluorooctyl) silane (1%, Aldrich, 448931) and biopsy punchers (Darwin microfluidics, KPUNCH05) were used.
2. For microfluidic device handling Prot/Elec 200 µl gel loading tips (Biorad, 223-9915), dH_2O (Invitrogen, 10977035), Tygon tubing (Cole-Parmer, GZ-06420-02), beads (Chemgenes, lot 051917, Macosko-2011-10), droplet generation oil (Biorad, 186-4006), murine RNase inhibitor (100 U, NEB, M0314L) were used. Cell wash buffer was prepared using PBS (1X, Gibco, 14190-094) and BSA (0.01%, Sigma, B8667). Cell loading buffer was prepared using PBS (1X, Gibco, 14190-094), Optiprep (6%, Sigma, D1556), and BSA (0.01%, Sigma, B8667). Lysis buffer was prepared from Optiprep (28%, Sigma, D1556), Sarkosyl (2.2%, Sigma, L7414), EDTA (20 mM, Sigma, 3690), Tris (100 mM, Sigma, T2944) and DTT (50 mM, Applichem, A2948,0005).
3. For sample washing prior to reverse transcription, SSC (6X, Sigma, S6639) and dH_2O (Invitrogen 10977035) were used.
4. For the reverse transcription (RT) reaction, dH_2O (Invitrogen, 10977035), Ficoll PM-400 (4%, Sigma, F5415), dNTPs (1 mM, Thermo, R0193), murine RNase inhibitor (100 U, NEB, M0314L), Maxima H Minus Reverse Transcriptase (500 U, Thermo Scientific, EP0753) and Template Switching Oligo (AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG, 2.5 µM, IDT) were used in a total volume of 50 µl per reaction.
5. For the exonuclease I reaction, exonuclease I (100 U, NEB, M0293L) and exonuclease buffer were used in a total volume of 50 µl per reaction.
6. For cDNA amplification, Kapa HiFi Hot start ready mix 2X (Roche, KK2602), dH_2O (Invitrogen, 10977035) and SMART PCR primer (AAGCAGTGGTATCAACGCAGAGT, 0.8 µM, IDT) used in a total volume of 50 µl per reaction. CleanPCR magnetic beads (0.6X ratio, GC biotech, CPCR-0050), Fragment Analyzer (Agilent, DNF-474-0500 kit) and the Qubit High Sensitivity kit (Invitrogen, Q33231) were used for cDNA purification and quantification.
7. For library preparation, Tn5 was produced in-house. To stop tagmentation, SDS was used (0.2%, Sigma, 71736). For library amplification the Kapa HiFi

kit with dNTPs (Roche, KK2102), P5 SMART PCR (AATGATACGGCGAC CACCGAGATCTACACGCCTGTCCGCGGAAGCAGTGGTATCAA CGCAGAGT*A*C, 0.3 µM, IDT), custom Nextera oligos[62] (0.3 µM, IDT) and dH_2O (Invitrogen, 10977035) were used. Libraries were purified and quantified using CleanPCR magnetic beads (0.6X ratio, GC biotech, CPCR-0050), Fragment Analyzer (Agilent, DNF-474-0500 kit) and Qubit High Sensitivity kit (Invitrogen, Q33231).

8. TE-TW wash buffer was prepared in dH_2O (Invitrogen, 10977035) using Tris (10 mM, Sigma T2944), EDTA (1 mM, Sigma, 3690), and Tween 20 (0.01%, Sigma, P9416).
9. TE-SDS wash buffer was prepared in dH_2O (Invitrogen, 10977035) using Tris (10 mM, Sigma, T2944), EDTA (1 mM, Sigma, 03690), and SDS (0.5%, Sigma, 71736).
10. Tris wash buffer was prepared in dH_2O (Invitrogen, 10977035) using Tris (10 mM, Sigma, T2944).
11. For mammalian cell culture dissociation and counting, Trypsin-EDTA (Gibco, 25200056) and trypan blue were used (0.4%, Thermo Fisher Scientific, T10282). Cell culture medium was prepared using DMEM Glutamax (Gibco, 10565018), FBS (10%, Gibco, 10270106) and penicillin–streptomycin (100 U per ml, Gibco, 15140122). Cell wash and cell loading buffers were prepared as described above.
12. Intestinal organoids were cultured in Matrigel (Corning, 356230) with organoid base medium (described in point 13) supplemented with ENR (and CV where indicated) and ROCK inhibitor (where indicated, Sigma, Y0503). Matrigel was dissolved with Cell Recovery Solution (Corning, 354253).
13. Organoid base medium was prepared using DMEM/F12 (Gibco, 11320033), HEPES (100 mM, Gibco, 15630056), penicillin–streptomycin (100 U per ml, Gibco, 15140122), B27 supplement (1 µM, Gibco, 17504-044), N2 supplement (1 µM, Gibco, 17502001) and *N*-acetyl-L-cysteine (1 µM, Sigma, A9165).
14. ENR medium was prepared using base medium (as above), EGF (E, 50 ng per ml, LifeTechnologies, PMG8043), mNoggin (N, 100 ng per ml, produced in-house) and R-spondin (R, 1 µg per ml, produced in-house).
15. ENR CV medium was prepared with the addition of CHIR (C, 3 µM, CalBiochem, CHIR99021) and valproic acid (V, 3 mM, Sigma P4543) to ENR medium.
16. Single-organoid single-cell dissociation mix was prepared using PBS (Gibco, 14190-094), *B. licheniformis* protease (10 mg per ml, Sigma P5380), EDTA (5 mM, Sigma 03690), EGTA (5 mM, BioWorld, 40520008-1), DNase I (10 µg per ml, Roche 11 284 932 001) and Accutase (0.68X, Sigma, A6964) in a total volume of 20 µl per reaction. For single-organoid dissociation, Nunc MicroWell plates (Nunc, 438733) and siliconized p10 pipette tips (VWR, 53509-134) were used.
17. For intestinal organoid preparation for RNAscope, cold Cell Recovery Solution (Corning, 354253), Histogel (Thermo Scientific, HG-4000-012) and paraformaldehyde (4%, PFA, Electron Microscopy Sciences, 15714) were used.
18. For the RNAscope assay, organoids were stained using RNAscope Multiplex Fluorescent V2 assay (ACD Bio-Techne, 323110), Ly6a probe (ACD Bio-Techne, 427571-C2), Fabp1 probe (ACD Bio-Techne, 562831), Muc2 probe (ACD Bio-Techne, 315451-C2), PpiB probe (ACD Bio-Techne, 313911-C2), Duplex negative control (ACD Bio-Techne, 320751), TSA Opal650 (Perkin Elmer, FP1496001KT), TSA Opal570 (Perkin Elmer, FP1488001KT) and Prolong Diamond Antifade Mountant (Thermo Fisher, P36965).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The GEO (Gene Expression Omnibus) accession number for scRNA-seq data reported in this paper is GSE148093. The raw data and count matrices for Fig. 1h and Extended Data Fig. 2c are stored under the access code GSM4454017. The raw data and count matrices for Fig. 1i and Extended Data Fig. 2a are available under the access code GSM4454017. The raw data and count matrices for Fig. 1j are stored under the access codes GSM4454012–GSM4454016. The raw data and count matrices for Extended Data Fig. 2e,f are stored under the access codes GSM5567775–GSM5567779. The raw data and count matrices for Extended Data Fig. 2g are stored under the access codes GSM5567571–GSM5567730. The raw data and count matrices for Extended Data Fig. 4 are stored under the access codes GSM5567845–GSM5567854. The raw data for intestinal organoids embedded in Figs. 2 and 3, Extended Data Figs. 3–5, Figs. 4a,e and 5d,e and Extended Data Figs. 7, 9 and 10 are stored under access codes GSM4453981–GSM4454011. The raw data and count matrices for intestinal crypts embedded in Fig. 5 and Extended Data Figs. 8–10 are stored under the access codes GSM5567818–GSM5567844. Additionally, dataset GSM1544799 and data from ref. [23] (https://doi.org/10.1039/C9LC00014C, data available on request) were used for Fig. 1i and Extended Data Fig. 2a. In this study the following reference genomes were used: hg38 (GCF_000001405.26), mm10 (GCF_000001635.20) and mixed reference genome (GSE63269) of hg19 combined with mm10.

Code availability

This technology has been developed as an open source platform, therefore all required information for its implementation is publicly available. The source code for the machine-vision software is available on github (https://github.com/DeplanckeLab/DisCo_source) and the barcode merging script is supplied as Supplementary Software 1.

References

48. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2016).
49. Zhang, X. et al. Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems. *Mol. Cell* **73**, 130–142 (2019).
50. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
51. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
52. Goldstein, L. D. et al. Massively parallel nanowell-based single-cell gene expression profiling. *BMC Genomics* **18**, 519 (2017).
53. Wang, Y. et al. Comparative analysis of commercially available single-cell RNA sequencing platforms for their performance in complex human tissues. Preprint at *bioRxiv* https://doi.org/10.1101/541433 (2019).
54. Ding, J. et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).
55. Gjorevski, N. et al. Designer matrices for intestinal stem cell and organoid culture. *Nature* **539**, 560–564 (2016).
56. Bas, T. & Augenlicht, L. H. Real time analysis of metabolic profile in ex vivo mouse intestinal crypt organoid cultures. *J. Vis. Exp.* **93**, e52026 (2014).
57. Macosko, E., Goldman, M. & McCarroll, S. *Drop-Seq Laboratory Protocol version 3.1*. http://mccarrolllab.org/download/905/ (2015).
58. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
59. Stuart, T. et al. Comprehensive integration of single-dell data. *Cell* **177**, 1888–1902 (2019).
60. Anders, S., Pyl, P. T. & Huber, W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
61. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
62. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

Acknowledgements

The authors thank the former members of the Deplancke laboratory (EPFL) for their support: W. Chen and P.C. Schwalie for constructive discussions, and V. Braman for help in establishing the intestinal organoid culture. Furthermore, the authors thank G. Sorrentino from K. Schoonjans' laboratory (EPFL) for valuable advice and support during organoid culture establishment; L. Aeberli and G. Muller from SEED Biosciences for cell sorting support; and the EPFL CMi, GECF, BIOP, FCCF, Histology Core Facility, SCITAS, and UNIL VITAL-IT for device fabrication, sequencing, imaging, sorting, histology, and computational support, respectively. The authors also thank J. Sordet-Dessimoz from the Histology Core Facility for her support with the RNAscope assay. This research was supported by the Swiss National Science Foundation Grant (IZLIZ3_156815) and a Precision Health and Related Technologies (PHRT-502) grant (to B.D.), the Swiss National Science Foundation SPARK initiative (CRSK-3_190627) and the EuroTech PostDoc Programme co-funded by the European Commission under its framework program Horizon 2020 (754462, to J.P.), as well as by the EPFL SV Interdisciplinary PhD Funding Program (to B.D. and E.A.). Y.S. is an ISAC Marylou Ingram scholar.

Author contributions

B.D., J.B. and M.B. designed the study. B.D., J.B., M.B. and J.P. wrote the manuscript. J.B. and R.D. designed and fabricated the microfluidic chips. J.B. developed the machine-vision integration for DisCo. J.B. and M.B. benchmarked the system and performed all single-cell RNA-seq experiments. J.P, J.B., M.B., W.S., V.G. and R.G. performed data analysis related to single organoid and crypt scRNA-seq experiments. J.B., S.R. and M.B. performed all organoid and cell culture assays. J.B., A.C., J.R. and R.S. performed all imaging assays. M.B. and J.R. isolated intestinal crypts, and M.B. picked and dissociated crypts. Y.S. supervised data analysis aspects and reviewed the manuscript. E.A. provided critical comments regarding microfluidic chip design and fabrication. M.C. provided critical comments on intestinal organoid scRNA-seq data analysis. M.P.L. provided critical comments regarding intestinal organoid scRNA-seq data and the design of critical confirmation experiments. All authors read, discussed and approved the final manuscript.

Competing interests

B.D., J.B., M.B. and R.D. have filed a patent application for the deterministic co-encapsulation system (patent no. US20190240664A1). All other authors have no competing interests.

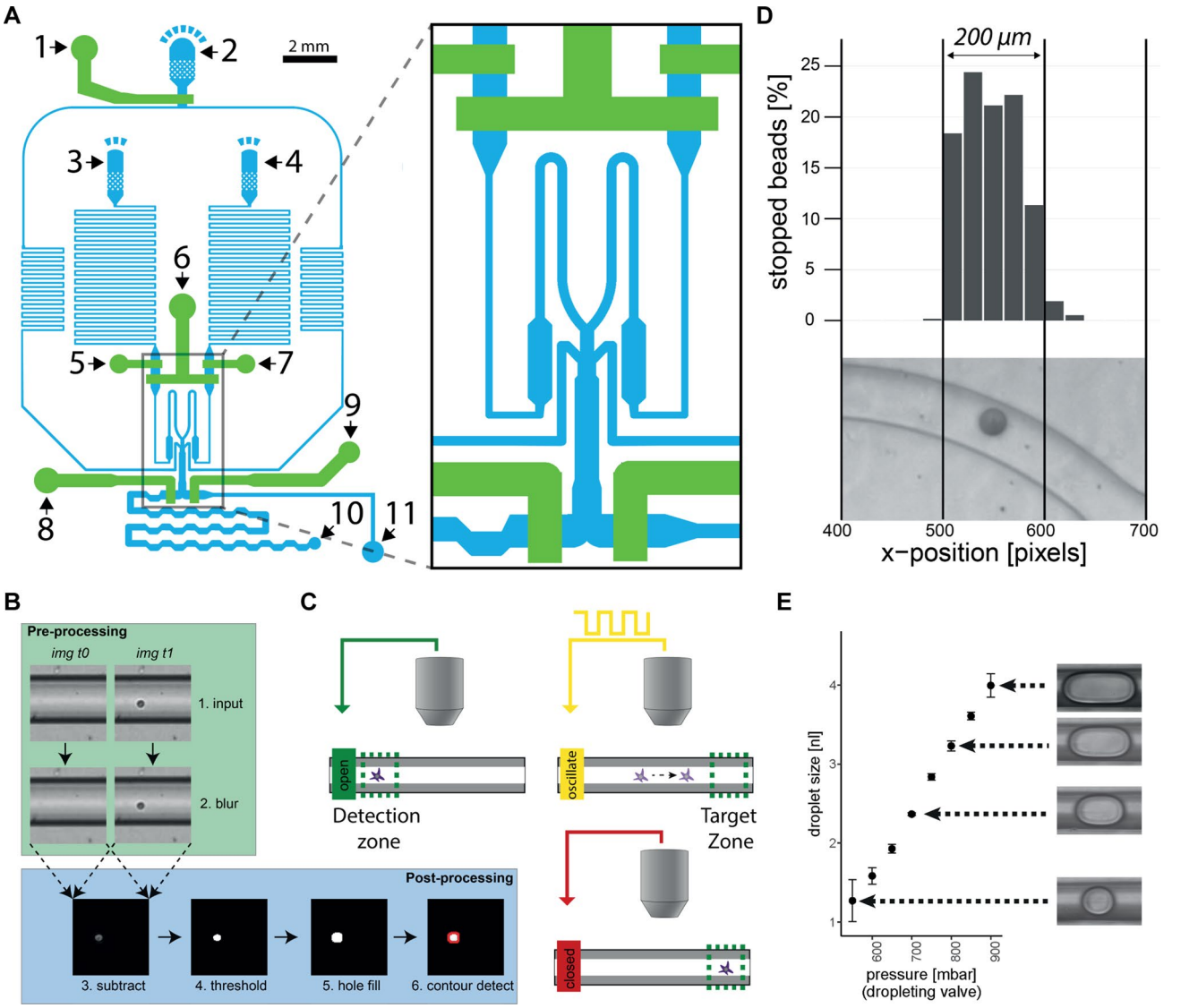
Additional information

Extended data are available for this paper at https://doi.org/10.1038/s41592-021-01391-1.

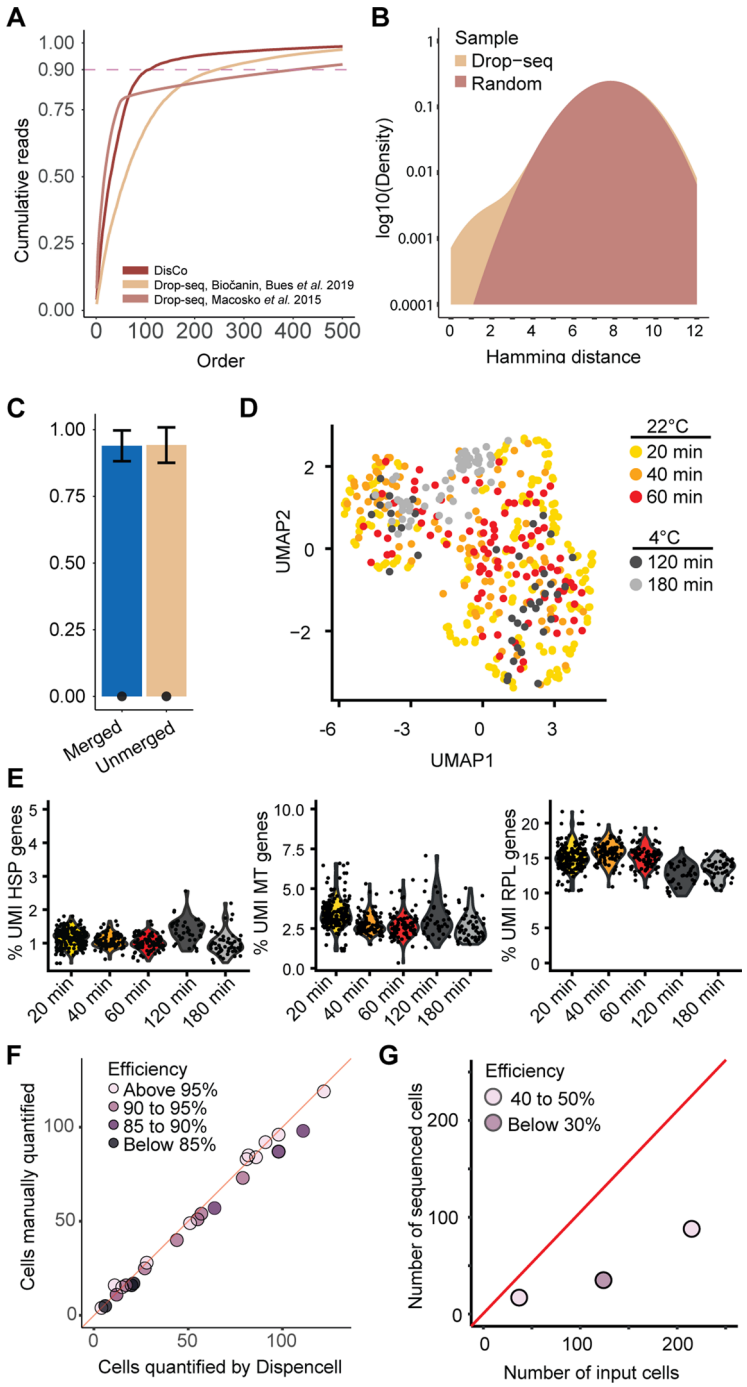
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-021-01391-1.

Correspondence and requests for materials should be addressed to Bart Deplancke.

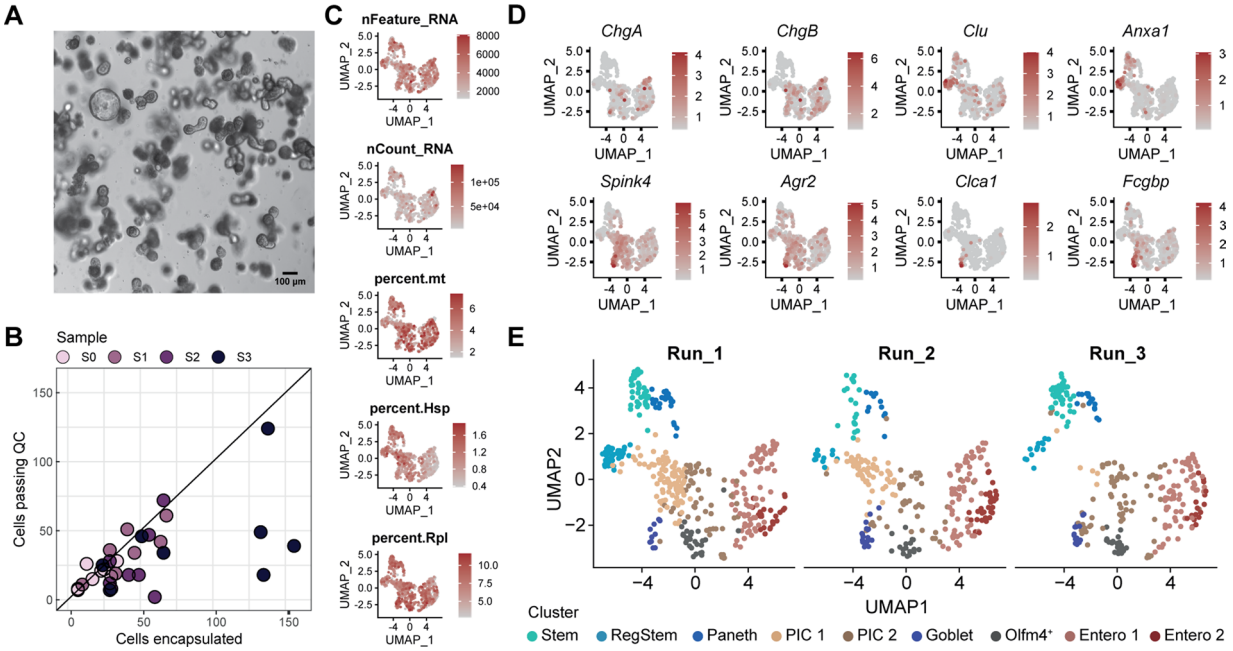
Peer review information *Nature Methods* thanks Jarrett Camp and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Lei Tang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.



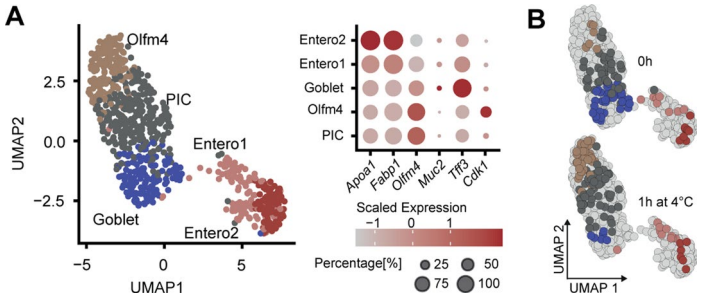
Extended Data Fig. 1 | DisCo device features and their performance. (**a**) Schematic of the DisCo device design (blue: flow layer, green: control layer). 1. oil valve, 2. oil inlet, 3. cell inlet, 4. bead inlet, 5. cell valve, 6. dropleting valve, 7. bead valve, 8. sample valve, 9. waste valve, 10. sample outlet, 11. waste outlet. (**b**) Visualization of real-time image processing for particle detection. (**c**) Particle positioning by valve oscillation. Approaching particles are detected in the detection zone. Once a particle is detected, the channel valve is oscillated to induce discrete movements of particles. Oscillation is terminated once correct placement of a particle is achieved. (**d**) Stopping accuracy in a defined window. Beads (n = 744) were positioned using valve oscillation, their position was manually determined within the stopping area. Scale was approximated from channel width. (**e**) Volume-defined droplet on-demand generation by valve pressurization. Droplets (n = 68, ~8 per condition, 1 experiment) were produced by pressurizing the dropleting valve at different pressures. Size was determined by imaging the dropleting process. Volumes were calculated from the imaging data based on droplet length and channel geometry. Thus, they should be considered an approximation. Points represent mean values, error bars +/-SD. The channel width of displayed images is 250 µm.



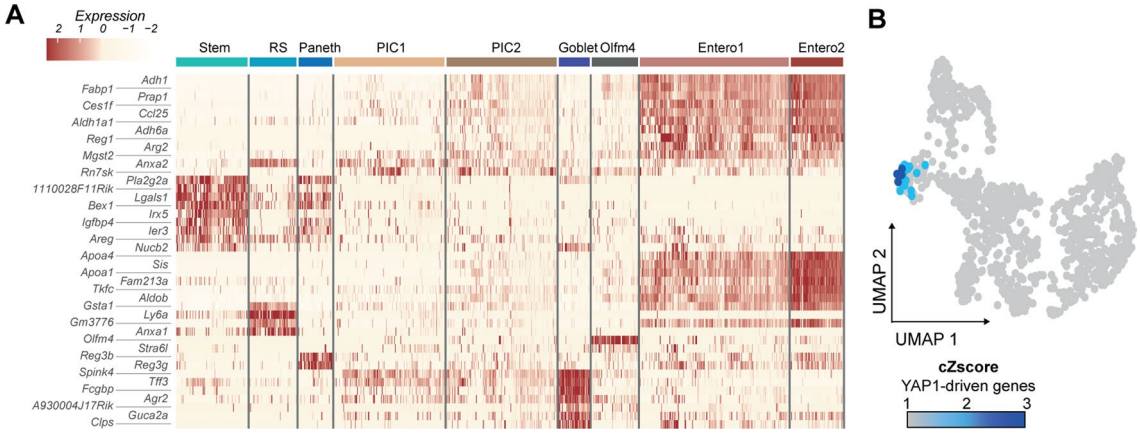
Extended Data Fig. 2 | Quality assessment of DisCo scRNA-seq data. (**a**) Cumulative reads per barcode (n = 500) for DisCo and two Drop-seq experiments[2,23]. (**b**) Hamming distances between all 12 nt barcodes of a Drop-seq experiment and generated 12 nt random barcode sequences representing the probability density for each set of barcodes. (**c**) Species purity (bars) and doublet ratio (dots) for unmerged (n = 949) and merged barcodes (n = 274). Data represent mean values, error bars standard deviation. (**d-e**) HEK 293T cells were processed with DisCo at 22 °C after 20, 40 or 60 min or stored on ice for 120 or 180 min and subsequently processed. (**d**) UMAP embedding of all profiled HEK 293T cells from the five sampling time points, color-coded by sampling time. (**e**) Violin plots showing the percentage of UMIs per cell of heat-shock-protein (HSP), mitochondrial protein-coding (MT), or ribosomal protein-coding (RPL) genes. (**f**) Correlation of the number of manually counted cells by fluorescence microscopy and the number of cells quantified by the DISPENCELL platform. (**g**) A quantified number of HEK 293T cells was processed with the Fluidigm C1 system. Processing efficiency was calculated as the percentage of cells retrieved from the sequencing data respective to the quantified number of input cells. The red line represents 100% efficiency, and samples were colored according to the recovery efficiency after sequencing.



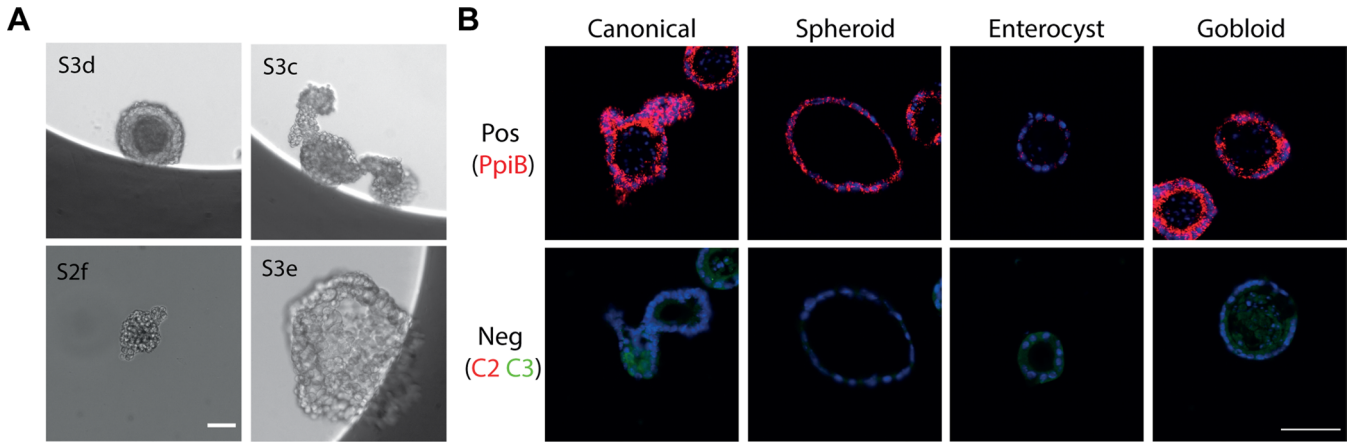
Extended Data Fig. 3 | DisCo performance on individual intestinal organoids and data analysis. (a) Representative bright-field image of a differentiated organoid culture from single LGR5⁺ cells, as performed for experiments shown in Fig. 2. (**b**) Correlation of encapsulated cells on-chip with the number of cells detected after sequencing (cells passing QC, filtered above 800 genes/cell). (**c**) UMAP embedding colored by the number of detected genes (nFeature) per cell, the number of detected UMIs (nCount) per cell, the percentage of mitochondrial (mt) reads per cell, and the percentage of reads mapping to genes coding for respectively heat-shock proteins (Hsp), and ribosomal proteins (Rpl) per cell. (**d**) UMAP embedding colored by expression of selected marker genes (*Clu*, *Anxa1*, *Spink4*, *ChgB*, *ChgA*, *Agr2*, *Clca1*, and *Fcgbp*). (**e**) UMAP embedding for each of the three independent experimental batches colored by cluster annotation.



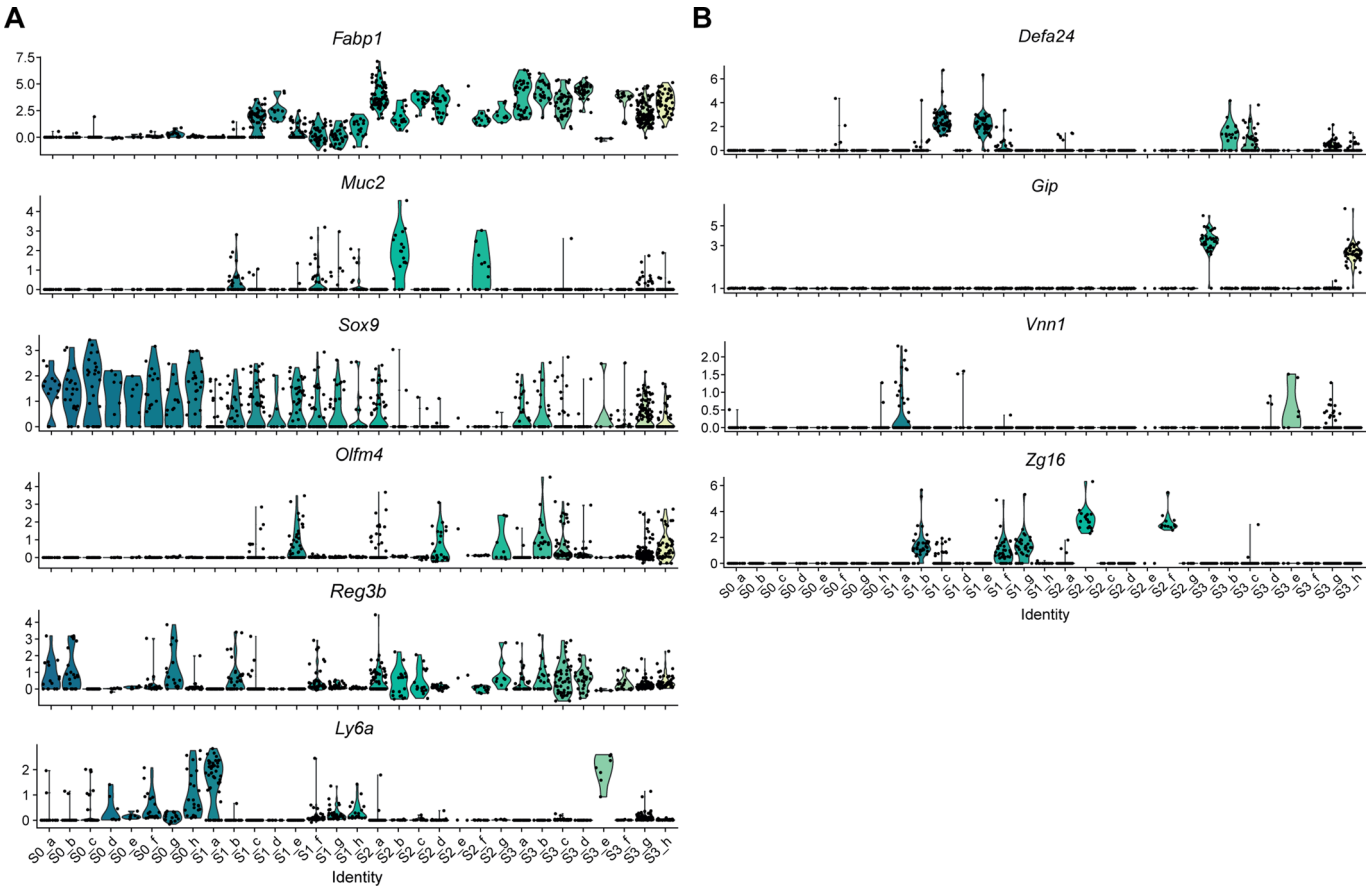
Extended Data Fig. 4 | Batch effect assessment on organoids by split organoid experiment. (**a**) UMAP embedding of cells collected from nine additional individual organoids (under maintenance conditions) for the purpose of evaluating batch effects. *Left:* All 748 processed cells clustered with k-means clustering, after which clusters were annotated according to marker gene expression. *Right:* Expression dot plot of selected marker genes. (**b**) Projection of cells (colored by cell type) derived from one organoid that was split into two independent samples (split organoid) on the reference UMAP shown in **a**). Organoid 'S2_2' was split into two batches, which were processed subsequently, with a one-hour delay, during which the second batch was stored at 4 °C.



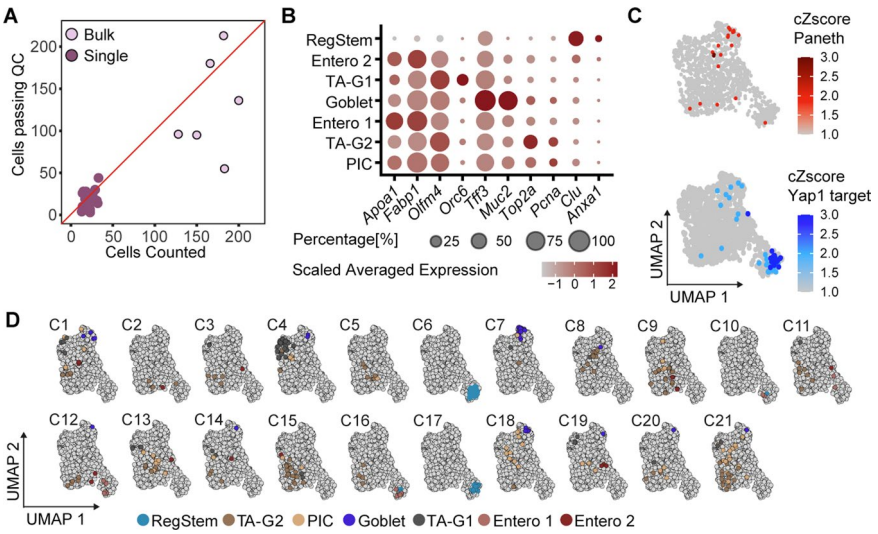
Extended Data Fig. 5 | Marker genes and YAP1 target gene activity of intestinal organoid cells. (**a**) Heatmap of top DE genes per annotated cluster. (**b**) YAP1 target gene activity on a UMAP embedding. The expression of genes that are positively regulated by YAP1[27] was calculated as the cumulative Z-score and projected on the UMAP embedding of all sequenced cells.



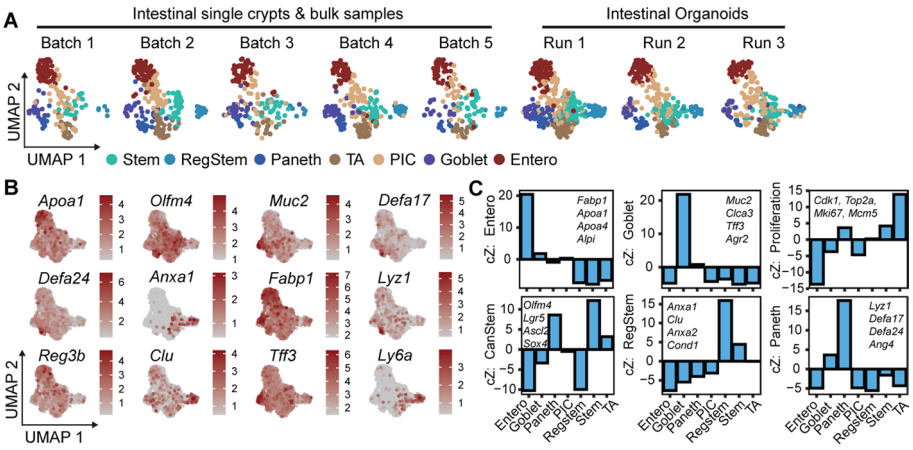
Extended Data Fig. 6 | Control images of DisCo- and RNAscope-processed organoids. (**a**) Selected organoids from Fig. 4, imaged in microwell plates before dissociation to single cells. Scale bar: 50 µm. (**b**) RNAscope controls for organoids shown in Fig. 4b,f. Positive control (*PpiB*), and negative control (Duplex negative). Scale bar: 50 µm.



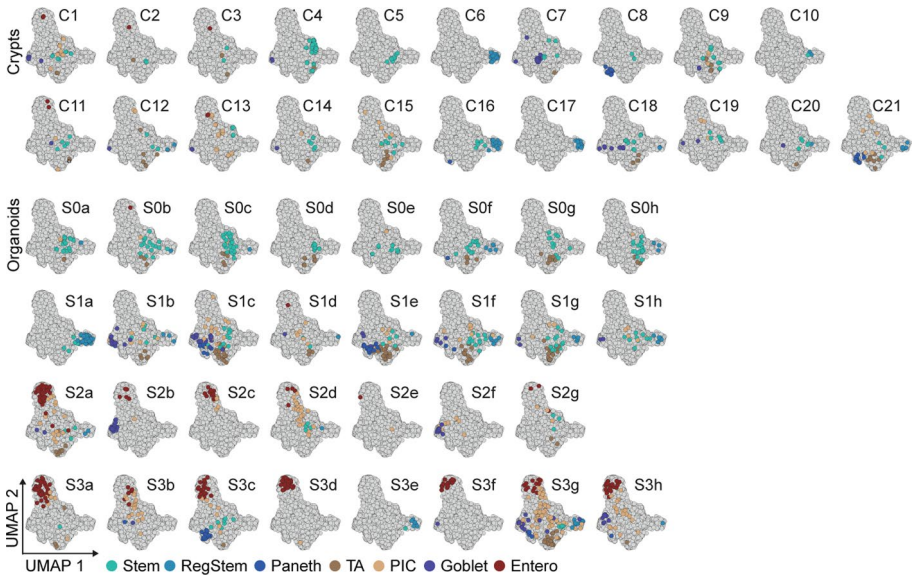
Extended Data Fig. 7 | Marker gene expression per individual organoid. (**a**) Violin plots showing marker gene expression (*Fabp1*, *Muc2*, *Sox9*, *Olfm4*, *Reg3b*, *Ly6a*) per organoid. (**b**) Violin plots showing the expression of selected genes (*Defa24, Gip, Vnn1, Zg16*) identified via psupertime analysis per individual organoid.



Extended Data Fig. 8 | DisCo performance on individual intestinal crypts and data analysis. (**a**) Processing efficiency of DisCo for individual and bulk intestinal crypts. All cells processed with DisCo were manually counted during the experiment, and compared to cell numbers after quality filtering (>500 UMIs). The red line represents 100% efficiency, and samples are colored according to sample type. (**b**) Expression dot plot of marker genes for clusters shown in Fig. 5a. (**c**) Gene activity represented as the cumulative Z-score and projected on the UMAP embedding of all sequenced cells using the expression of *Top*: Paneth cell-associated genes encompassing *Lyz1*, *Defa17*, *Defa24* and *Ang4* and *Bottom*: genes that are positively regulated by YAP1[27]. (**d**) Projection of cell types onto the reference UMAP of cells derived from the 21 individual crypts. Cells per single crypt were colored according to their global clustering and highlighted on the UMAP embedding of all sequenced cells. Enterocytes (Entero), PIC (Potential intermediate cells), RegStem, (Regenerative Stem), TA (Transit amplifying cells; G1: G1/S and G2: G2/M cell cycle phase).



Extended Data Fig. 9 | Analysis of combined intestinal organoid and crypt data. (**a**) Combined UMAP embedding (as shown in Fig. 5d) stratified by the five individual batches of intestinal crypt samples and the three independent experimental batches of intestinal organoid differentiation samples, collectively embedded and colored by cluster annotation. (**b**) UMAP-based visualization of the expression of specific markers that were used for cluster annotation. (**c**) Bar graph depicting the cumulative Z-score of the expression of genes that are indicated within the respective bar graph. *CanStem*: canonical stem cell, *RegStem*: regenerative stem cell.



Extended Data Fig. 10 | Individual organoids and crypts mapped onto the denominator UMAP. Projection of cell types onto the reference UMAP of the *ex vivo* cell preparation for the 21 individual intestinal crypts and bulk samples embedded together with the 31 individual intestinal organoids. Cells per single crypt or organoid are colored according to their global clustering and highlighted on the UMAP embedding of all sequenced cells.

nature research

Corresponding author(s): Bart Deplancke

Last updated by author(s): Nov 17, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Microfluidic devices were designed using Tanner L-Edit CAD software (Mentor, v 2016.2). For the custom machine-vision code (available on github: https://github.com/DeplanckeLab/DisCo_source) the Open Source Computer Vision Library OpenCV version 3.4, and the proprietary XiApi library version 4.15. were used. RNA-scope data was acquired on an Olympus VS120 whole slide scanner using Olympus OlyVIA software.
Data analysis	The following software was used for single cell RNA-sequencing data analysis: Sequencing reads were aligned with STAR aligner version 2.7.0.e. Aligned reads were processed using samtools (version 1.9) and the Drop-seq tools package (version 2.3.0, Macosko et al., Cell 2015). Barcodes were merged as described in the Material and Methods section using a custom R-script (available upon request). Read count matrices were analyzed using the Seurat R toolkit for single cell genomics version 3.1.1, and uwot (version 0.1.3). For slingshot analysis the Dyno package (https://github.com/dynverse/dyno) was used, version 1.0 for identification of genes that change along the trajectory, and version 1.1 for data visualization. For psupertime analysis the psupertime package (Macnair & Claassen, biorxiv 2019) was utilized. Flow cytometry data analysis was performed using FlowJo software version 10.6 (Tree Star). RNA-scope images were extracted and processed in Fiji version 1.52p, utilizing the BIOP VSI Reader plugin version 7. ROIs were extracted from the images using a custom script (available upon request) written in Python version 2.7.15 using the Python Imaging Library PIL version 6.2.2.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The GEO accession number for scRNA-seq data reported in this paper is GSE148093. The raw data and count matrices for Figure 1H and Extended Data Figure 2C are stored under the access code GSM4454017. The raw data and count matrices for Figure 1I and Extended Data Figure 2A are available under the access code GSM4454017. The raw data and count matrices for Figure 1J are stored under the access codes GSM4454012 - GSM4454016. The raw data and count matrices for Extended Data Figure 2D/E are stored under the access code GSM5567775 - GSM5567779. The raw data and count matrices for Extended Data Figure 2G are stored under the access codes GSM5567571 - GSM5567730. The raw data and count matrices for Extended Data Figure 4 are stored under the access codes GSM5567845 - GSM5567854. The raw data for intestinal organoids embedded in Figure 2/3, Extended Data Figure 3/5, Figure 4A/E and 5D/5E, Extended Data Figure 7A/B, Extended Data Figure 9A-C, and Extended Data Figure 10 are stored under access codes GSM4453981- GSM4454011. The raw data and count matrices for intestinal crypts embedded in Figure 5A-E, Extended Data Figure 8, Extended Data Figure 9, and Extended Data Figure 10 are stored under the access codes GSM5567818 - GSM5567844. Additionally, dataset GSM1544799 and data from Biocanin & Bues et al. LoC 2019 (doi: 10.1039/C9LC00014C, data available on request) were used for Figure 1I and Extended Data Figure 2A.

Within this study the following reference genomes were used: hg38 (GCF_000001405.26), mm10 (GCF_000001635.20) and mixed reference genome (GSE63269) of hg19 combined with mm10.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Encapsulation performance measurements shown in Figure 1F/H were performed for 12 samples, each comprising approximately 100 encapsulation events at various cell concentrations. scRNA-seq benchmarking experiments shown in Figure 1H/I were performed once, as is standard for this type of experiment. Quantification of total cell processing efficiencies of DisCo (Figure 1J) was performed on five independent samples containing varying numbers of cells. Capture efficiency comparison experiments shown in Extended Data Figure 2G were performed on three samples. scRNA-seq of individual organoids (Figure 2, 3, 4A/E) was performed on 31 organoids and a total of 945 cells. This number was sufficient to resolve cell-types and individual organoids. As stated in the results section, biased selection and limited number of organoids did not allow us to draw conclusions on subtype prevalence. For scRNA-seq of individual crypts (Figure 5) we collected 21 crypts summing a total of 372 cells. As the cell recovery rate from crypts was low, we additionally collected five samples of 775 cells derived from bulk crypts. These experimental aspects are disclosed in the results section. Cells from bulk and individual crypts combined allowed us to resolve cell types and composition of individual crypts. We did not perform calculations to determine sample size. Sample sized utilized was sufficient to support the results of our study.
Data exclusions	Cells that did not reach described quality criteria were removed from the dataset. This was done for all scRNA-seq datasets displayed in this manuscript. Exclusion thresholds are described in the Methods section.
Replication	The species mixing experiment (shown in Figure 1, Extended Data Figure 2) was executed once. The HEK processing kinetic (shown in Extended Data Figure 2D&E) was executed once. The C1 processing efficiency experiment (shown in Extended Data Figure 2G) was executed once. The intestinal organoid batch processing experiment (shown in Extended Data Figure 4) was executed once. RNA-scope (Figure 4, Extended Data Figure 6) were performed in two replicates, of which one is displayed in the Figures. The scRNA-seq data of intestinal organoids (shown in Figure 2/3/4/5, and Extended Data Figure 3/5/7/9/10) was acquired in three independence experimental runs, and is thus considered a triplicate. The scRNA-seq data of intestinal crypts (shown in Figure 5, and Extended Data Figure 8/9/10) was acquired in five independent experimental runs, and is thus considered a quintuplicate. FACS sorting experiments were done in triplicates.
Randomization	Experiments were not randomized, as all samples underwent the same experimental treatment.
Blinding	Investigators were not blinded in this study, as all samples underwent the same experimental treatment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used: For flow cytometry:
PE-Cy7 anti-mouse Ly-6A/E (Sca1) monoclonal antibody (Biolegend 122514)

Validation: Validation of antibodies was performed by the vendors: "Each lot of this antibody is quality control tested by immunofluorescent staining with flow cytometric analysis.". No additional validation steps were performed.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s): HEK 293T: ATCC CRL-11268

IBA cells, provided by Christian Wolfrum's lab at ETHZ: Klein, J., Fasshauer, M., Klein, H. H., Benito, M. & Kahn, C. R. Novel adipocyte lines from brown fat: a model system for the study of differentiation, energy metabolism, and insulin action. BioEssays 24, 382–388 (2002).

Lgr5-GFP instestinal stem cells, isolated as described in: Gjorevski, N. et al. Designer matrices for intestinal stem cell and organoid culture. Nature 539, 560–564 (2016).

Authentication: Cell lines were not authenticated.

Mycoplasma contamination: Cells were not tested for mycoplasma contamination.

Commonly misidentified lines (See ICLAC register): No commonly misidentified cell lines were used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals: Intestinal organoids were generated from mouse small intestinal stem cells isolated from 5–10 weeks old heterozygous, male/female, Lgr5-eGFP-IRES-CreERT2 mice (Jackson Laboratory). Isolation of intestinal crypts were isolated from 7-week old male C57BL/6J mice. Mice were kept at temperatures of 22°C +/- 2°C, relative humidity of 55% +/- 10%, and a dark/light cycle of 12h/12h (starting light at 7am and dark at 7pm).

Wild animals: No wild animals were used in this study.

Field-collected samples: No field-collected samples were used in this study.

Ethics oversight: Intestinal organoid generation experiments were approved by the "Service de la consummation et des affaires vétérinaires", Epalinges, Switzerland, license number 2681.0. Crypt isolation was conducted under the license VD3406e granted by the local authorities: "Direction générale de l'agriculture, de la viticulture et des affaires vétérinaires", Epalinges, Switzerland.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Organoid dissociation to single cells: Matrigel was depolymerized by removing culture medium and adding ice cold Cell Recovery Solution to a well, and subsequent incubation at 4°C for 30 minutes. Next, the suspension was pipetted up and down approximately 20 times with a 200 ul pipette to dissociate the remaining solid Matrigel and shear the organoids. The organoid containing suspension was collected in a FBS treated tube, and centrifuged in a 4°C cooled centrifuge at 95 x g for 5 minutes. The supernatant was carefully aspirated, and the organoid pellet resuspended in TrypLE containing 10 µg/mL DNaseI, and ROCK inhibitor. The suspension was pipetted up and down 20 times with a 1000 ul pipette, incubated for 5 minutes at 37°C, and this process repeated once. Next, the suspension was pipetted up and down 20 times, and centrifuged in a 4°C cooled centrifuge for 5 minutes at 774 x g. Finally, the cells were resuspended in organoid culture medium containing ROCK inhibitor, and strained using a 35 um cell strainer. FACS staining: Surface staining with anti-mouse Anti-Sca1/Ly6A/E (0.2 ug per 1 million cells) was performed for 30minutes in culture medium at 4°C (this step was not performed for the organoid scRNA-seq time course experiments as cells were only sorted based on GFP signal). Subsequently, cells were washed by centrifugation in a 4°C cooled centrifuge for 5 minutes at 774 x g, and resuspension in organoid culture medium. Live/dead discrimination was carried out using DAPI added shortly before sorting.
Instrument	FACS was performed on a a FACS ARIA II (BD).
Software	Flow cytometry data analysis was performed using FlowJo software version 10.6 (Tree Star).
Cell population abundance	Cell population abundances are depicted in Figure 3.
Gating strategy	Cells were gated in FSC vs. SSC as to exclude debris. A Figure of the gating strategy is provided in the following file: "BuesBiocaninPezoldt_2021_FACSgating.pdf".

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.