Cytopath: Simulation based inference of differentiation trajectories from RNA velocity fields

R. Gupta[1,‡] , D. Cerletti[3,4,‡] , G. Gut[3] , A. Oxenius[4], M. Claassen[1,2*]

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.

‡shared contribution
Corresponding author: manfred.claassen@med.uni-tuebingen.de

Abstract

Unsupervised trajectory inference from single cell RNA sequencing data bears the potential to systematically reconstruct complex differentiation processes but remains challenging notwithstanding the many available solutions. In general, trajectory and pseudotime inference methods have so far suffered from the ambiguity of the static single-cell transcriptome snapshots lacking a measure of directionality of transcriptional activity.

We report Cytopath, a method for trajectory inference that takes advantage of transcriptional activity information from RNA velocity of single cells to perform trajectory inference. Cytopath performs this task by defining a Markov chain model, simulation of an ensemble of possible differentiation trajectories and subsequent statistical analysis to extract the topological and molecular characteristics of the studied differentiation process. We demonstrate the capability of Cytopath to reconstruct differentiation trajectories with varying bifurcated, circular, convergent and mixed topology studied in single-snapshot as well as time-series single-cell RNA sequencing experiments. We demonstrate superior and enabling capability to reconstruct differentiation trajectories in comparison to state-of-the art trajectory inference approaches.

Trajectory inference constitutes a frequent step in interpreting single-cell RNA sequencing studies. We expect Cytopath to enable researchers to tap the directionality information present in single-cell RNA sequencing data to achieve trajectory inference for possibly complex lineages at an unprecedented precision and resolution.

Keywords Single cell RNA sequencing · RNA velocity · Trajectory inference · Simulation based inference

1 Introduction

Biological processes such as neuronal cell type differentiation [Manno et al., 2018], pancreatic endocrinogenesis [Byrnes et al., 2018], immune response to viral infection [Cerletti et al., 2020] and cell cycle [Mahdessian et al., 2021] are defined by a temporal sequence 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. Furthermore, biological processes are often asynchronously triggered. The differentiation processes give rise to state sequences with varying topologies including bi-, multi-furcation, cycles 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 (scRNAseq) has gained in wide spread use due to the broad applicability of sequencing technology. While these measurements are information rich, their analysis and interpretation is challenged by high-dimensionality, low sequencing depth, measurement noise and destructive nature only yielding snapshots of the whole process.

Different computational approaches have been proposed to reconstruct cell state sequences of biological processes from scRNAseq data. This reconstruction involves a computational inference step, also known as pseudotime ordering. This procedure aims at ordering the measured molecular profiles consistent with the state sequence of the underlying process and is typically guided by the assumption that phenotypic similarity reflects temporal proximity. Therefore similarity

of gene expression profiles is utilized to estimate adjacency in a graph of cell state transitions. Various statistical approaches have been developed to address this task with varying amount of success depending on the size and topology of the data. The interpretation of pseudotime ordering for branched processes is particularly challenging and requires an additional trajectory inference step. A comprehensive review by [Saelens et al., 2019] on the performance of various methods on multiple topologies suggests, Slingshot [Street et al., 2018] has the best general performance on datasets with multiple branching lineages, and further highlights Monocle3 [Trapnell et al., 2014], as a popular runner-up for pseudotime inference based on a lineage tree.

Static expression profiles are ambiguous with respect to the directionality of potential cell state transitions. This ambiguity constitutes a major limitation of pseudotime ordering and trajectory inference, and specifically precludes data driven assignment of start and endpoints without previous knowledge about the process, as well as resolving complex, i.e. cyclical or convergent process topologies. Recently, it became possible to estimate transcriptional activity from scRNAseq data via RNA velocity analysis [Manno et al., 2018] consequently enabling the inference of likely transitions between different cell states in a data-driven fashion, ultimately opening the possibility to mitigate the limitations of afore mentioned reconstruction approaches.

So far no functional trajectory inference approach incorporates such directionality information to estimate pseudotime or construct trajectories. In this work we aim at closing this gap by presenting Cytopath, a simulation based trajectory inference approach that accounts for RNA velocity. We show that Cytopath infers accurate and robust cell state trajectories of known biological process with linear, circular, bifurcated and convergent topologies from scRNAseq datasets. Trajectory inference with Cytopath is superior to state-of-the art trajectory inference approaches such as Slingshot and addresses the major limitations of such methods.

2 Results

In this section we present an overview of the Cytopath approach and its trajectory inference performance assessed on five scRNAseq datasets consisting of cellular differentiation processes with various topologies. We show that Cytopath successfully recovers known biological processes in all the datasets with high accuracy. Furthermore we compare the performance of Cytopath to widely used trajectory inference tools, Slingshot and Monocle3 as well as two prototype approaches accounting for RNA velocity information, i.e. VeTra [Weng et al., 2021] and Cellpath [Zhang and Zhang, 2020]. We demonstrate that the shortcomings of these tools are successfully addressed by Cytopath and that our velocity based approach enables the resolution of so far intractable topologies.

Simulation based trajectory inference with Cytopath

Trajectory inference with Cytopath follows the RNA velocity analysis of a scRNAseq dataset, and is specifically based on the resulting cell-to-cell transition probability matrix. While we base our analysis on an RNA velocity analysis, in principle any cell-cell transition probability matrix can be used as input for trajectory inference. The objective of trajectory inference with Cytopath is to estimate trajectories from root to terminal cell states, defining the start and end points of the studied differentiation process, respectively. Root and terminal states can either be derived from a Markov random-walk model utilizing the transition probability matrix itself, as described in [Manno et al., 2018], or can be supplied by the user based on suitable prior knowledge.

The trajectory inference process is divided into four steps [Figure 1 and 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 of unique cell state sequences reaching the terminal states have been generated [Figure 1B-C].

The generated cell state sequences are individual simulations of the biological process from the root to the terminal state. Due to the stochastic nature of the sampling process, the cell state sequences cannot be considered as 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. The procedure aligns simulations to a common differentiation coordinate such that cell states from any simulation at a particular differentiation coordinate (pseudotime) represent similar cell states [Figure 1C-D].

In the third step, consensus 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 as the consensus state of the trajectory at the particular step. Alternatively, trajectories can be anchored to observed cell states in the data by choosing the cell state closest to the aforementioned average value.

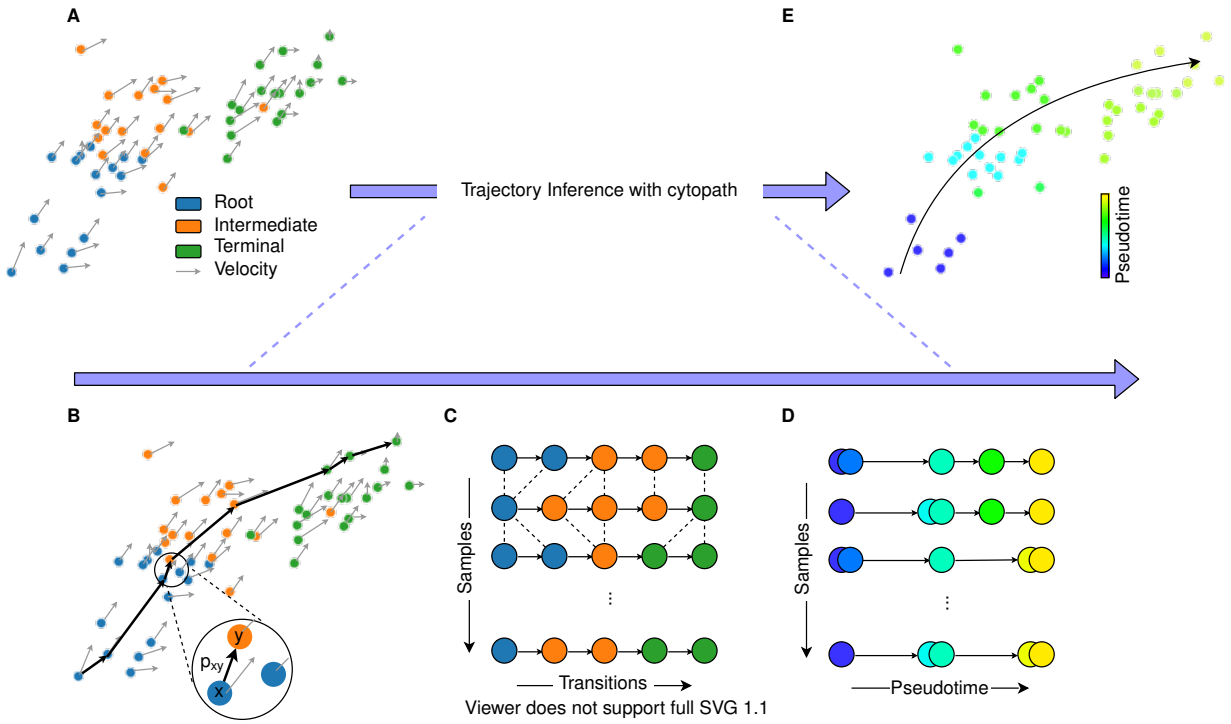

Figure 1: Cytopath overview. (A) Input data for trajectory inference consisting of single cell gene expression profiles, RNA velocity profiles, root and terminal state annotation. (B) Simulation of the differentiation process by sampling a markov chain is performed based on the cell to cell transition probabilities. Sampling is initialized on cells from a RNA velocity derived root state. (C) Simulation is performed for a defined number of steps. Transition steps are aligned using Dynamic Time warping. (D) After alignment cells at each transition step represent the same consensus state. (E) Cells along the inferred trajectory are assigned the aligned transition-step value as the pseudotime.

In the final step, cell groups are assigned to each step of the inferred trajectory. Assignment is based on an alignment score evaluating for each cell both similarity of its static as well as the velocity profile with those of the consensus states. The cell level score is used to estimate position within the trajectory i.e. pseudotime as well as the relative association of a cell state to possibly multiple branches of a differentiation processes with complex topology.

Reconstruction of neuronal differentiation in the developing mouse hippocampus

We assessed the capability to reconstruct developmental processes with multiple branching. To this end, we applied Cytopath and afore mentioned baseline approaches 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. RNA velocity was used to identify the root and terminal regions [Manno et al., 2018].

Data from the original publication including t-SNE projection, RNA velocity and transition probability matrix were utilized for trajectory inference with Cytopath. 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 [Figure 2B]. While we also supply the known root and terminal states as supervision to Slingshot and Monocle3 (accepts root states only), we could not do this for VeTra and Cellpath since these approaches do not allow inclusion of this supervision. The dataset comprises five terminal regions and a common root state. The topology of the data is multifurcating with development branches arising directly from the root state, namely astrocytes and oligodendrocyte precursors (OPC), but also as branches from intermediate states, namely neuroblast and Cajal-Retzius differentiation.

Trajectories and pseudotime estimated by each trajectory inference method are shown in Figure 2C-E. Cytopath estimates the correct number of trajectories to each terminal state while Slingshot produces a spurious trajectory despite having the terminal states as supervision. Monocle3 fails to produce a connected trajectory, producing a disjoint graph and therefore unable to estimate pseudotime for a large portion of the dataset.

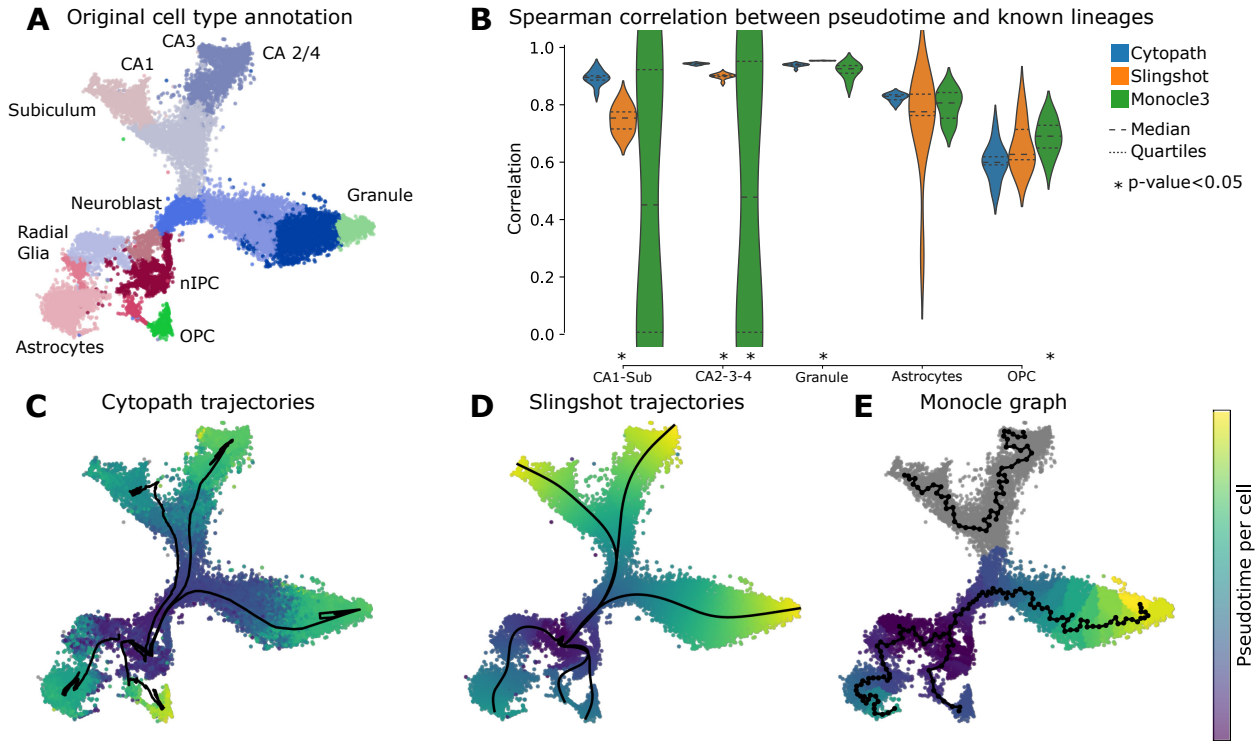

Figure 2: Reconstruction of neuronal differentiation in the developing mouse hippocampus. (A) t-SNE projection of the dendate gyrus scRNAseq dataset annotated with stages of neuronal differentiation. (B) Spearman correlation of pseudotime values assigned by each method to the known ordering of cell types per trajectory. Significantly different medians for Slinsghot and Monocle compared to Cytopath are marked with an asterix. (C-E) Trajectories inferred and mean pseudotime per cell by Cytopath (C), Slingshot (D) Monocle3 (E).

The trajectory inference procedure was performed ten times for each method. Spearman correlation between the pseudotime estimated by each method and the true lineage ordering of cell identities shown indicates that pseudotime estimated by Cytopath outperforms other methods in terms of recovering the expected ordering of cells for a majority of the lineages. Furthermore, the correlation distribution for Cytopath per lineage has much lower variance compared to Slingshot and Monocle3 and thus the trajectory inference is more reliable. [Figure 2B].

Both VeTra and Cellpath infer erroneous trajectories that initiate at terminal or intermediate states [Figure S15A]. Both methods also exclude a large number of cells from the trajectory inference process. Since these methods initialise trajectories in intermediate states, most cells are only assigned to one or two trajectories while Cytopath correctly assigns root cells to all trajectories and only terminal states correspond to a single trajectory [Figure S16A].

Cell cycle reconstruction

We hypothesized that the ability to infer repeating patterns during differentiation, likely differentiates RNA velocity based trajectory inference from other methods based on similarity of expression. To assess this hypothesis. we compared the reconstruction of the cell cycle in a dataset comprising 1067 U2OS cells generated using the SMART-Seq2 protocol. [Mahdessian et al., 2021].

Within cells annotated in the original publication as G1, root and terminal cells were selected based on probabilities estimated using *scvelo* [Bergen et al., 2020]. The terminal region was annotated as G0 based on the relatively lower expression of cell cycle marker genes *Ccne2*, *Cdk1*, *Ccna2*, *Birc5* [Figure 3E-F]. Of the three methods only 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 G0 stage [Figure 3B-D]. 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.

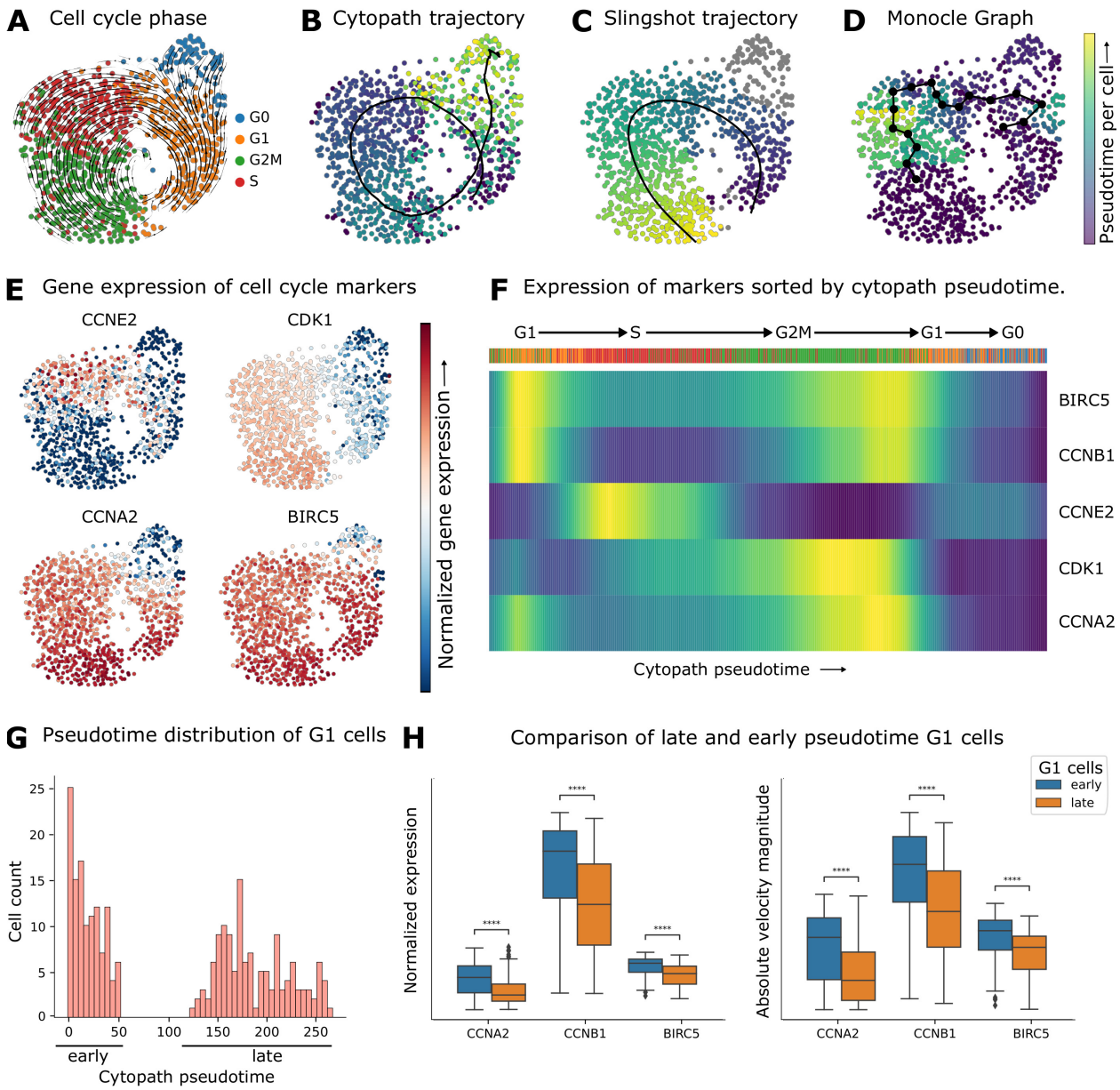

Figure 3: Reconstruction of cell cycle in U2OS cell line. (A) RNA velocity stream plot overlayed on the UMAP projection, annotated with cell cycle phase adapted from [Mahdessian et al., 2021]. The separation of G1 phase into G1 and G0 was performed on the basis of marker expression shown in this figure panel. (B-D) Trajectories inferred and pseudotime per cell by Cytopath (B), Slingshot (C) Monocle3 (D). (E-F) Normalised gene expression of (E) cell cycle marker genes, (F) sorted by Cytopath pseudotime. The colourbar represents the ordering of cell cycle phase by Cytopath pseudotime. (G) Distribution of Cytopath pseudotime for cells in the G1 cluster. (H) Normalised expression and absolute velocity magnitude of cells classified as early and late G1 cells (blue/orange respectively). Significance was estimated by an independent t-test for each marker.

In contrast, the pseudotime inferred by Slingshot and Monocle are inconsistent with marker expression as well as their own trajectories not ending in the end point region provided as input and instead ending prematurely in intermediate cell cycle stages [Figure 3E-F].

Furthermore, we observe that cells in the G1 cluster can be divided into two groups on the basis of pseudotime inferred by Cytopath [Figure3G]. Both expression and velocity of markers associated with cell cycle are significantly higher in

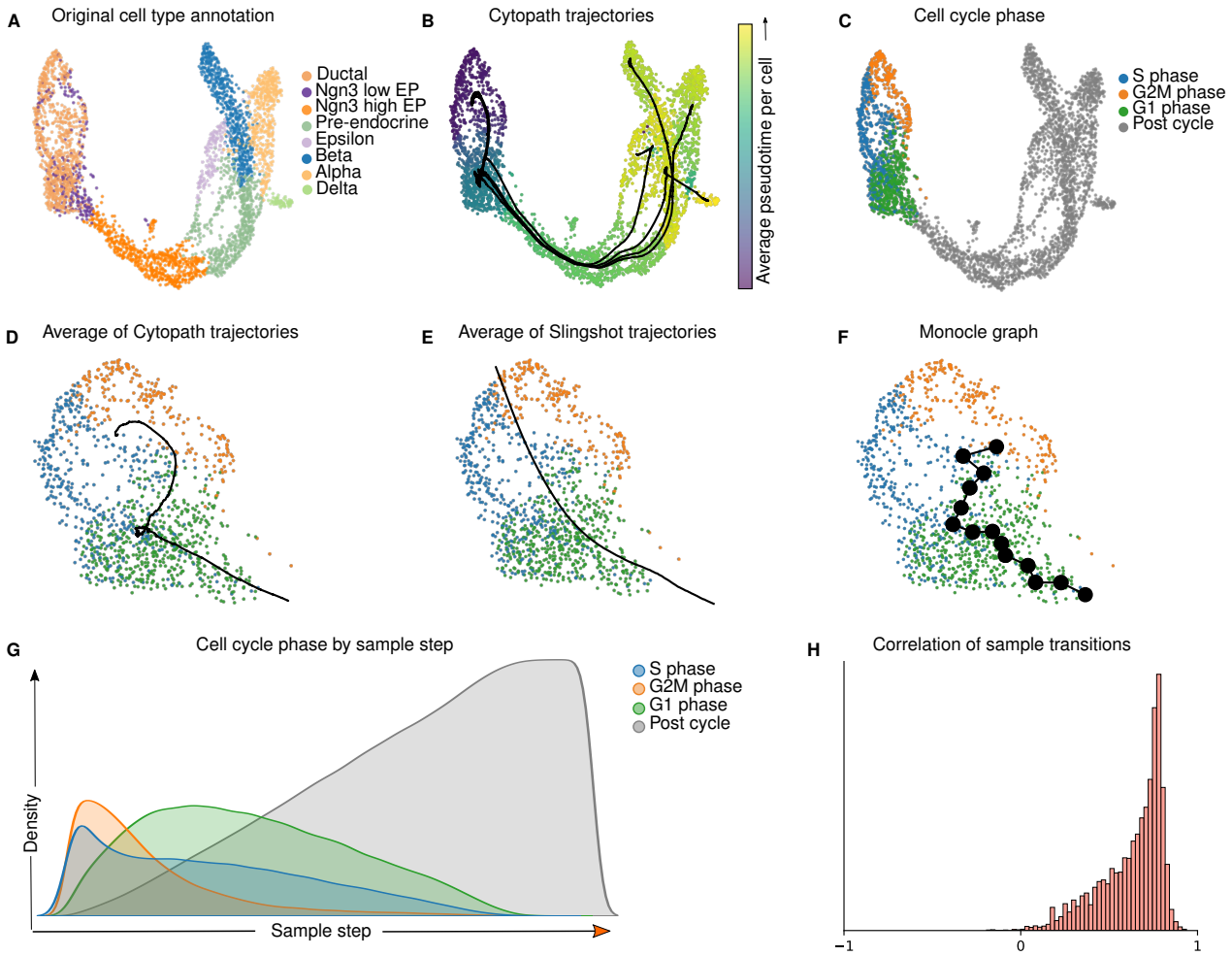

Figure 4: Reconstruction of interlaced cell cycling and bifurcated differentiation in pancreatic endocrinogenesis. (A) UMAP projection of pancreas scRNAseq data annotated with stages of differentiation. (B) Trajectories inferred by Cytopath and mean pseudotime per cell. (C) Estimated cell cycle phase assignment. (D-F) Cell cycle phase annotation and mean trajectory in the root region by Cytopath (D), Slingshot (E) and Monocle3 (F). (G) Cell cycle phase as a function of transition step of individual Cytopath simulations. (H) Spearman correlation of cell cycle phase with transition step of individual Cytopath simulations

the early pseudotime G1 cells than for those destined to move into G0 phase and accordingly associated with higher pseudotime [Figure 3H].

Cytopath and Slingshot both infer a second linear trajectory. This is not unexpected as apart from the circular route, a direct connection of the root to the terminal region is also possible [Figure S2].

Similar to Slingshot, VeTra does not include cells corresponding to G0 phase in the trajectory inferred by it. The pseudotime appears to be inconsistent with the root and terminal state probabilities. Cellpath also infers an erroneous trajectory that does not capture the cyclical process both originating and terminating in intermediate states [Figure S15B].

Reconstruction of interlaced cell cycling and bifurcated differentiation in pancreatic endocrinogenesis

We further 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 [Byrnes et al., 2018]. Preprocessing, RNA velocity and the transition probability matrix estimation were performed with *scvelo* [Bergen et al., 2020].

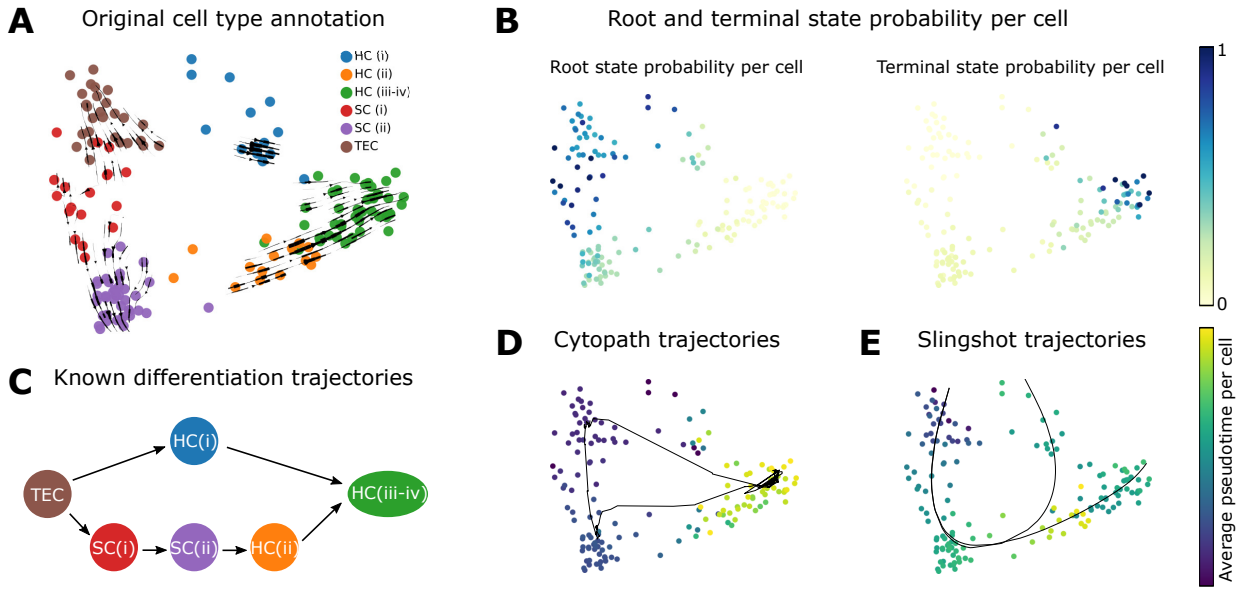

Figure 5: Reconstruction of convergent differentiation in developing neonatal mouse inner ear. (A) RNA velocity stream plot overlayed on the PCA projection of neonatal mouse inner ear data annotated with stages of differentiation. (B) Probability estimated based on RNA velocity of the cell being a root state and terminal state respectively. (C) Known differentiation trajectories from [Burns et al., 2015]. (D-E) Inferred trajectories and mean pseudotime per cell by (D) Cytopath and (E) Slingshot.

Biological cell type annotation from [Byrnes et al., 2018] was used to provide root and terminal state supervision to Cytopath [Figure 4A]. Trajectories estimated by Cytopath to each terminal state capture the expected differentiation events of endocrinogenesis [Figure 4B] [Byrnes et al., 2018].

Cell cycle scoring of cells in the root region was performed using *scanpy* [Wolf et al., 2018]. This clearly revealed distinct cell cycle states [Figure 4C]. The structure of the mean trajectory estimated by the baseline approaches in the root region of this dataset does not adhere to the cyclic structure of the cell cycle stages [4D-F]. In contrast, the trajectory inferred by Cytopath from the ensemble of 8000 simulations recapitulates the circular structure of the cell cycle. We further assessed the frequency of cell cycle phase as a function of the transition step of individual simulations comprising the final trajectory [Figure 4G]. Spearman correlation of cell cycle phase with the transition steps of each simulation indicate faithful recapitulation of the cell cycle stages at the single-simulation level [Figure 4H].

The simulation based approach of Cytopath ensures that even in the absence of explicit supervision, cyclic transcriptional patterns are faithfully reconstructed. In contrast, due to the absence of RNA velocity information, the designated root states appear to be isotropic for conventional trajectory inference approaches like Slingshot and Monocle3 and therefore are unable to capture structured transcriptional heterogeneity within this region.

We also show the trajectory estimation with respect to the full pancreatic endocrinogenesis process. Slingshot and Monocle3 produce spurious or to few trajectories respectively [Figure S1].

VeTra reports a spurious trajectory that terminates at the Ductal stage, while trajectories to Beta and Alpha are initialised in intermediate or terminal cell states. Cellpath also initialises trajectories to alpha in intermediate states. Both methods erroneously include the delta cells in trajectories to other terminal cell states. Both methods exclude a large number of cells from the trajectory inference process [Figure S15C,S16C].

Reconstruction of convergent differentiation in developing neonatal mouse inner ear

Burns et. al. have shown that the development of hair cells (HC) within the sensory epithelium of the urticle originates from transitional epithelial cells (TEC) via support cells (SC). This study also reveals a secondary differentiation path that from TEC to HC and put forward the existence of a transitional zone where cells can easily switch fate resulting in two convergent differentiation trajectories [Burns et al., 2015].

Root and terminal state probability estimation using RNA velocity recovered the same root and terminal states described in the study. Cells belonging to the TEC cluster were chosen as the root and HC (iii-iv) as the terminus. As with all other analyses of datasets presented here, we chose to work in the same embedding as the original study (PCA).

Cytopath successfully recovers the two differentiation trajectories demonstrated in the study while Slingshot is unable to do so. In particular it generated a spurious branch including HC(i) as end point and missed the convergent path TEC-HC(i)-HC(iii-iv) [Figure 5D-E].

VeTra reports two non-overlapping trajectories, one terminates prematurely at SC(ii) while the second in erroneously initialised at HC(i). Cellpath reports two partially overlapping trajectories that simultaneously cover both convergent branches, with one trajectory initialised in intermediate states. Since Cytopath correctly infers the two convergent branches, we observe a trajectory count of two for the root and terminal states while intermediate states receive a count of one [Figure S16D]. Monocle3 requires a UMAP embedding and therefore it was not benchmarked in this analysis.

Reconstruction of bifurcating differentiation of CD8+ T cells from scRNA seq time series data.

We assessed the performance of Cytopath on a scRNAseq time-series dataset from CD8 T cell differentiation in chronic LCMV infection [Cerletti et al., 2020]. In this infection model system CD8 T cells differentiate from early activated cells into *exhausted* and *memory-like cells* over a period of three weeks. Samples were collected at eight experimental time points after infection with LCMV to cover all stages of the process [Figure 6A]. Although these samples are heterogeneous snapshots of a spectrum of differentiation states at a particular time point, they provide an approximate development coordinate. Starting from a population of cells at an early activated state, differentiation leads into the two distinct terminal states within five days of the LCMV infection. This differentiation process is characterized by strong transcriptional changes and expression of different surface genes.

We identified the root and end points of the process using *scvelo* [Bergen et al., 2020] [Figure 6B-C]. The endpoints were validated by expression levels of known marker genes [Cerletti et al., 2020]. The *exhausted* endpoint showed high expression in co-inhibitory markers like CD39 (*Entpd1*), CD160 (*Cd160*) and PD-1 (*Pdcd1*). The *memory-like* endpoint had high expression of TCF1 (*Tcf7*) and IL-7R (*Il7r*). Applying Cytopath to our data resulted in two trajectories leading from a shared starting region to the two expected endpoints. The two trajectories strongly overlapped in the beginning of the process but then sharply diverged at a branching point [Figure 6D].

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 very little overlap between the time-points. Despite the non-trivial structure of the UMAP projection, cells ordered correctly. The relative ordering is respected and the two terminal regions have the highest pseudotime. This illustrates that the average trajectory produced by Cytopath respects the experimental time ordering of the cells [Figure 6D]. Slingshot inferred trajectories to either terminus but generated a third spurious trajectory within the early stage cell group that cannot be matched to any expected infection induced differentiation trajectory [Figure 6E]. Monocle3 reconstructs the global structure of the data, but includes additional loops and branches within the exhaustion branch [Figure 6F]. VeTra and Cellpath both infer two non overlapping trajectories. Both estimate, one trajectory initialised at the root but terminating at the branch point of the exhausted and memory-like lineage. The second is erroneously initialised in the proliferative cell cluster [Figure S15E,S16E].

Further, we assessed the correlation of pseudotime estimates with canonical gene expression markers of each the memory-like (*Slamf6*, *Tcf7*, *Il7r*) and exhaustion branch (*Ccl5*, *Cxcr8*) [Figure 6G-H]. We observe that Cytopath achieves competitive to superior absolute performance and significantly lower variance according to this measure.

We further 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 [Sandu et al., 2020]. The average expression of *Cxcr6* increases in the trajectory towards the exhausted cluster right before the divergence of the two branches [Figure 6H], indicating that the paths are indeed governed by the exhaustion process. Conversely T cell Factor 1 (TCF1) and the expression of its gene *Tcf7* is an established marker of memory-like cells [Utzschneider et al., 2016]. Expression of this gene was increased in memory-like cells just after the cells started to diverge after the bifurcation point. Towards the memory-like endpoint at late time-points, *Tcf7* expression is exclusive to the memory-like population. An additional observation is the high expression of *Gzmb* early both branches that drops off towards later timepoints [Figure 6H]. The expression of *Gzmb* is a shared feature of both branches and known to decrease in both branches as the infection progresses and expression is low towards late timepoints [Wherry et al., 2007].

In summary, 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

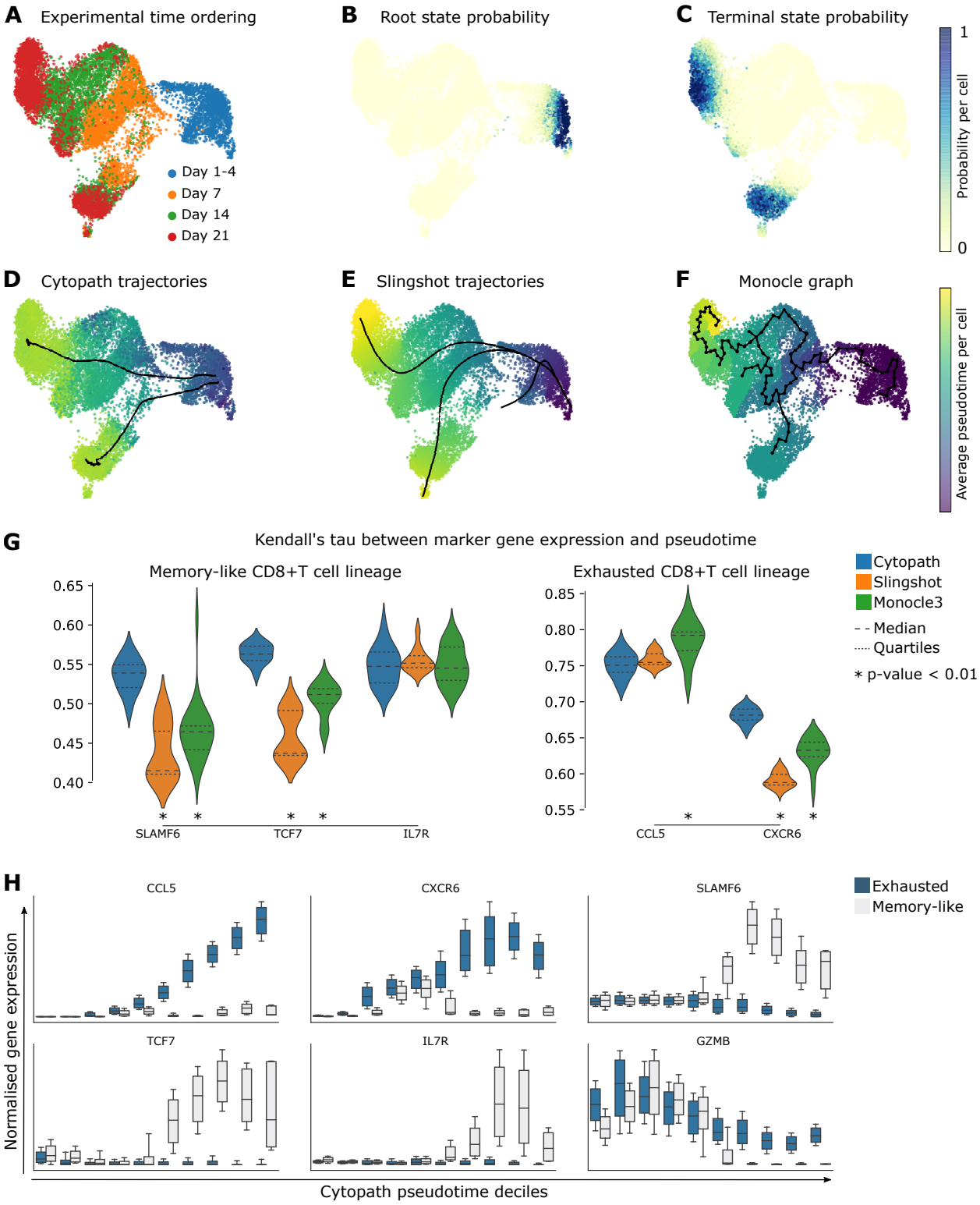

Figure 6: Reconstruction of bifurcating development of CD8+ T cells from scRNA seq time series. (A) Ordering of samples with respect to time of LCMV innoculation. (B-C) Probability estimated based on RNA velocity of the cell being (B) a root state (C) a termimal state (D-F) Trajectories inferred and mean pseudotime per cell by Cytopath (D), Slingshot (E) Monocle3 (F). (G) Correlation of pseudotime estimated by each method with marker relevant to the Memory-like CD8+ T cell differentiation and Exhausted CD8+ T cell differentiation respectively. Significantly different medians for Slinsghot and Monocle compared to Cytopath are marked with an asterix. (H) Normalized expression of key marker genes in each average Cytopath pseudotime decile.

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 had been described before, but the connecting differentiation trajectories of those populations are a subject of ongoing research [Chen et al., 2019, Zander et al., 2019, Yao et al., 2019, Raju et al., 2020]. These studies provide evidence for branching in the development process, and only recently in conjunction with simulation based trajectory inference it was possible to resolve this event in more detail [Cerletti et al., 2020].

3 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. 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. We demonstrate superior capability to reconstruct complex differentiation hierarchies from diverse scRNAseq datasets with respect to number of cells, differentiation topologies, type of embedding and sequencing technology.

Transitions between cells are inferred on the basis of the high dimensional expression and velocity profiles. Since Cytopath is based upon transitions that use the full expression and velocity profiles of cells, it is less prone to projection artifacts distorting expression profile similarity. In addition, this approach specifically considers likely and discards unlikely transitions, and thereby is able to identify for instance cyclic trajectories in an apparently diffusely populated and isotropic region of expression space [Figure 3, 4]. Furthermore, 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 since the cells appear to be co-located in expression space. However RNA velocity based pseudotime and cell fate estimation based on analytical analysis of Markov chain properties, performed by *scvelo* and *CellRank* [Lange et al., 2020] also does not readily present the user this information even if the pseudotemporal ordering estimated by these tools captures these patterns. In subsequent work, Cytopath simulations could be the basis for learning mechanistic differentiation models and gene regulatory networks.

While 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. In particular, Cytopath can be used to predict cell fate [Figure S6-S8].

Assessment of confidence in the inferred trajectories is important for their interpretation. Cytopath estimates the uncertainty of its pseudotime at the single-cell level, as well as its trajectories. To this end we take advantage of the alignment of cells to possibly multiple segments of inferred trajectories. With this output it is possible to estimate the uncertainty in the pseudotime value and therefore the temporal ordering of each cell [Figure S9-S11]. Furthermore by considering the distribution of alignment scores at each step of the trajectory it is possible to visualise the relative confidence of the inferred trajectory at each step [Figure S12-S14]. In this context, Cytopath complements the functionality of tools such as CellRank. Cell to cell transition matrices estimated using CellRank that account for uncertainty in RNA velocity vectors can be used to initialise the trajectory inference procedure.

Partition-based graph abstraction (PAGA) [Wolf et al., 2019] is yet another popular tool that can include RNA velocity to infer directed connectivity between clusters of a scRNA dataset. While the current implementation of PAGA requires an user-defined threshold to separate genuine and spurious connections, Cytopath's compositional clusters for each trajectory offer an alternative automatic estimate of cluster orderings and can be used to generate a coarse graph of cluster connectivity.

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 if this is not the case, Cytopath can detect and utilize tentative root and end points from the cell to cell transition probability matrix using *scvelo* [Bergen et al., 2020]. Furthermore, simulations generated using Cytopath can also aid in the identification of terminal cell types. Compared to the absorbing Markov process based inference of terminal states implemented in *velocyto* and *scvelo*, this approach appears to highlight all the terminal states in the pancreatic dataset [Figure S5]. Intermediate cells of interest that represent a shift in transcriptional process, that is cell cycle to islet cell differentiation also appear to be highlighted. While this is not a core functionality of the presented method, it enables its application to less characterized datasets.

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 [Figure S15-S16]. We assume that the reason for the difficulty to recapitulate trajectories could be the in-built lack of root and terminal state supervision in these methods. This appears to be a contrived problem since 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.

Cytopath's runtime scales with the number of simulation steps and not the number of cells [Figure S3A]. Additional flexibility in selecting parameters for RNA velocity estimation such as neighborhood size and in the trajectory inference process itself allows for moderate runtime even for relatively large datasets [Figure S3B].

In summary 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.

4 Methods

Trajectory inference with Cytopath

Cell clustering. Any grouping of cell states reflecting biological similarity can be provided as input to Cytopath. By default Cytopath will perform clustering of cells using louvain [Bergen et al., 2020].

Simulations. Simulations are initialized at random cell states selected uniformly within the defined root state cluster 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. The cell state c_{ij} at step *i* of simulation *j* is selected randomly according to P from the from the k nearest neighbors of $c_{(i-1)j}$.

Trajectory inference. Simulations that terminate within defined terminal state clusters are considered for trajectory inference. Simulations are generated until a predefined number of unique simulations have been returned. Simulations are clustered using louvain based on euclidean hausdorff distance. Simulations within clusters are aligned in a pairwise fashion using the *fastdtw* python package to achieve a subtrajectory. [Salvador and Chan, 2007] The mean value of the expression profile of cells at each step of the aligned simulations is the coordinate of the trajectory at that step. Subsequently, a second round of clustering and alignment is performed to produce trajectories that are reported by Cytopath. By default the number of expected trajectories per terminal region is not specified however it can be provided as an option. For the analysis presented above, this parameter was not specified and instead the number of reported trajectories was inferred using louvain clustering of subtrajectories.

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 a fixed size set M_i of neighboring cells are recovered with a K-dimensional tree search [Virtanen et al., 2020]. Cell clusters with cell representation larger than a user specified threshold frequency of ν for at least one M_i are considered a compositional cluster of the trajectory.

Alignment score. Alignment score of a cell with neighbors *K* to step *i* is the maximum of two scores. The score with respect to the trajectory segment from *i-1* to *i*, ξ_i is calculated as follows,

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

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 it's neighbors. The score with respect to the trajectory segment from *i* to *i+1*, τ_i is calculated similarly, The alignment score is a 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 and subsequently to compute a fate score with respect to multiple lineages.

$$\tau_i = \frac{1}{K} \sum_k cos(\eta_k^f) \cdot exp(\gamma_k) \tag{2}$$

The alignment score of the cell with respect to step *i*, d_i is

$$d_i = max(\xi_i, \tau_i) \tag{3}$$

For each cell the final alignment score with respect to a particular trajectory is an average of it's 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 the sum over trajectories normalised, relative alignment score is the cell fate score. Cell fate score f_{ij} for cell *i* with respect to trajectory *j* is

$$f_{ij} = \frac{d_{ij}}{\sum_j d_{ij}} \tag{4}$$

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.

Terminal state identification using undirected simulations

A clustering for the scRNAseq dataset is assumed as input, e.g. from a cell type annotation procedure. Each cluster was successively chosen as the origin for two thousand unique simulations with no defined terminal cluster. The total count of simulations, from all the clusters, terminating at each cell is reported as propensity for constituting a terminal state.

Parameter settings for comparative methods

Default settings were used for Slingshot and Monocle3. Recommended settings were used for VeTra and Cellpath. All supported inputs were provided including number of expected trajectories for VeTra and Cellpath.

Datasets

Dendate Gyrus. The dataset was obtained and processed using the following notebook (https://github.com/velocyto-team/velocyto-notebooks/blob/master/python/DentateGyrus.ipynb) [Manno et al., 2018].

Biological cell type annotation and RNA velocity derived root states were used to supervise trajectory inference for Cytopath, Slingshot and Monocle3 (root states only). Inference was peformed in the original t-SNE projection for Slingshot and Monocle3. For Cytopath, in addition to the projection the original RNA velocity field was used to infer cell to cell transitions.

Cell cycle The loom file containing spliced and unspliced counts was obtained from the following github repository (https://github.com/CellProfiling/SingleCellProteogenomics/) [Mahdessian et al., 2021]

RNA velocity was estimated using the stochastic model. Genes with less than ten combined spliced and unspliced counts were discarded. Expression was normalized by total count per cell. Top 5000 variable genes were selected. First twenty PCA components were used to calculate the neighborhood graph with 30 nearest neigbhours based on connectivity. Louvain clusters containing 99th percentile or above root and terminal state probability cells were provided as root and terminal clusters to Cytopath and Slingshot. Inference was performed in the UMAP projection for all methods.

Pancreatic endocrinogenesis. The data was downloaded and processed using *scvelo*. RNA velocity was estimated using the stochastic model. Genes with less than ten combined spliced and unspliced counts were discarded. Expression was normalized by total count per cell. First 30 PCA components were used to calculate the neighborhood graph with 30 nearest neigbhours based on connectivities. Louvain clustering was performed using this neighborhood graph. Biological cell type annotation was used to supervise trajectory inference for all methods. Inference was performed in the UMAP projection for all methods. Cell cycle scoring was performed using *scanpy* [Wolf et al., 2018].

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 peformed using *fastp* [Chen et al., 2018]. Reads were aligned to the GRCm39 mouse genome assembl using *STAR version=2.7.8a-2021-04-2* [Dobin et al., 2012]. Spliced and unspliced counts were estimated using the *velocyto run-smartseq2* command following the reccomendation of the developers.

RNA velocity was estimated using the stochastic model. Genes with less than ten combined spliced and unspliced counts were discarded. Expression was normalized by total count per cell. Top 1000 variable genes were selected. First five PCA components were used to calculate the neighborhood graph with 30 nearest neigbhours based on distance.

Biological cell type annotation was used to supervise trajectory inference for Cytopath and Slingshot. Inference was performed in the PCA projection for all methods.

CD8 development. Transgenic P14 CD8 T cells were sampled longitudinally during chronic infection with LCMV Clone-13 infection. The samples were acquired from four phases of the infection, namely activation (day 1-4), effector (day 7), early exhaustion (day 14) and late exhaustion (d21) and scRNAseq was performed using the 10x Genomics platform.

Read counts were realigned and sorted for spliced and unspliced counts using the analysis pipeline from velocyto [Manno et al., 2018]. Contaminating other cell types were removed from the dataset based on outliers in diffusion components. [Cerletti et al., 2020] RNA velocity was estimated using the stochastic model. Genes with less than ten combined spliced and unspliced counts were discarded. Expression was normalized by total count per cell. Top 3000 variable genes were selected. First 30 PCA components were used to calculate the neighborhood graph with 30 nearest neigbhours based on connectivities. Louvain clustering was performed using this neighborhood graph. Inference was performed in the UMAP projection for all methods.

Comparison of trajectory inference approaches

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 three methods. Kendall's tau was used to assess the correlation of marker expression with the estimated pseudotime.

Mood's median test was used to assess whether the median correlation differed significantly for Slingshot or Monocle with respect to Cytopath.

Availability

Cytopath has been implemented as a python package and can be found at the following GitHub repository (https://github.com/aron0093/cytopath).

References

[Bergen et al., 2020] Bergen, V., Lange, M., Peidli, S., Wolf, F. A., and Theis, F. J. (2020). Generalizing rna velocity to transient cell states through dynamical modeling. *Nature Biotechnology*.

[Burns et al., 2015] 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. *Nature Communications*, 6(1).

[Byrnes et al., 2018] Byrnes, L. E., Wong, D. M., Subramaniam, M., Meyer, N. P., Gilchrist, C. L., Knox, S. M., Tward, A. D., Ye, C. J., and Sneddon, J. B. (2018). Lineage dynamics of murine pancreatic development at single-cell resolution. *Nature Communications*, 9(1).

[Cerletti et al., 2020] Cerletti, D., Sandu, I., Gupta, R., Oxenius, A., and Claassen, M. (2020). Fate trajectories of cd8+ t cells in chronic lcmv infection. *bioRxiv*.

[Chen et al., 2018] Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics*, 34(17):i884–i890.

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

[Dobin et al., 2012] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2012). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21.

[Lange et al., 2020] Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., Schiller, H. B., Pe'er, D., and Theis, F. J. (2020). CellRank for directed single-cell fate mapping. *BiorXiv*.

[Mahdessian et al., 2021] Mahdessian, D., Cesnik, A. J., Gnann, C., Danielsson, F., Stenström, L., Arif, M., Zhang, C., Le, T., Johansson, F., Shutten, R., and et al. (2021). Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. *Nature*, 590(7847):649–654.

[Manno et al., 2018] Manno, G. L., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R.,

Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S., and Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, 560(7719):494–498.

[Raju et al., 2020] 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. *BiorXive*.

[Saelens et al., 2019] Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology*.

[Salvador and Chan, 2007] Salvador, S. and Chan, P. (2007). Fastdtw: Toward accurate dynamic time warping in linear time and space. Armament and Technical Products.

[Sandu et al., 2020] Sandu, I., Cerletti, D., Oetiker, N., Borsa, M., Wagen, F., Spadafora, I., Welten, S. P., Stolz, U., Oxenius, A., and Claassen, M. (2020). Landscape of exhausted virus-specific CD8 T cells in chronic LCMV infection. *Cell Reports*, 32(8):108078.

[Street et al., 2018] 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 Genomics*, 19(1).

[Trapnell et al., 2014] 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. *Nature Biotechnology*, 32(4):381–386.

[Utzschneider et al., 2016] Utzschneider, D. T., Charmoy, M., Chennupati, V., Pousse, L., Ferreira, D. P., Calderon-Copete, S., Danilo, M., Alfei, F., Hofmann, M., Wieland, D., Pradervand, S., Thimme, R., Zehn, D., and Held, W. (2016). T cell factor 1-expressing memory-like cd8 t cells sustain the immune response to chronic viral infections. *Immunity*, 45(2):415–427.

[Virtanen et al., 2020] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

[Weng et al., 2021] Weng, G., Kim, J., and Won, K. J. (2021). Vetra: a tool for trajectory inference based on rna velocity. *Bioinformatics*.

[Wherry et al., 2007] 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(4):670–684.

[Wolf et al., 2018] Wolf, F. A., Angerer, P., and Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1).

[Wolf et al., 2019] 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 Biology*, 20(1).

[Yao et al., 2019] 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., Khan, O., Handon, R., Reilley, J., Fioravanti, J., Hu, J., Gossa, S., Wherry, E. J., Gattinoni, L., McGavern, D. B., O'Shea, J. J., Schwartzberg, P. L., and Wu, T. (2019). Single-cell RNA-seq reveals TOX as a key regulator of CD8+ t cell persistence in chronic infection. *Nature Immunology*, 20(7):890–901.

[Zander et al., 2019] 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(6):1028–1042.e4.

[Zhang and Zhang, 2020] Zhang, Z. and Zhang, X. (2020). Inference of high-resolution trajectories in single cell rna-seq data from rna velocity.

Supplementary Figures

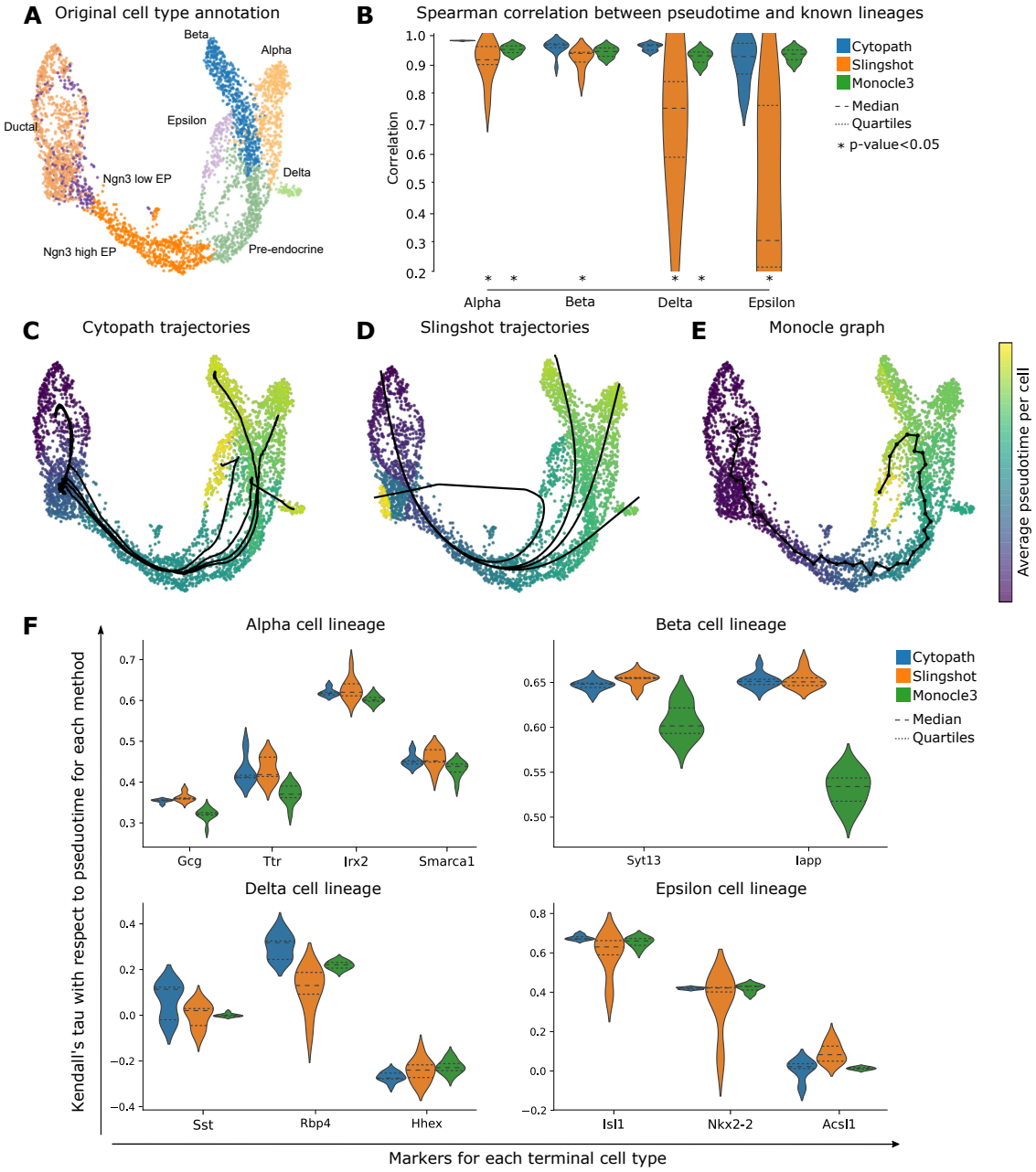

Figure S1: Comparison of methods for the reconstruction of bifurcated differentiation in pancreatic endocrinogenesis. (A) UMAP projection of the pancreatic endocrinogenesis dataset annotated with stages of differentiation. (B) Spearman correlation of pseudotime values assigned by each method to the known ordering of cell types per trajectory. Significantly different medians for Slinsghot and Monocle compared to Cytopath are marked with an asterix. (C-E) Trajectories inferred and mean pseudotime per cell by Cytopath (C), Slingshot (D) Monocle3 (E). (F) Correlation of pseudotime estimated by each method with markersrelevant to the trajectories inferred for each terminal cell type.

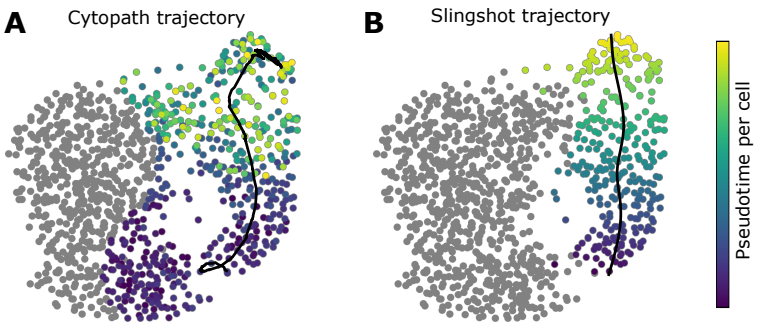

Figure S2: Secondary trajectories inferred by Cytopath and Slingshot on the cell cycle dataset.

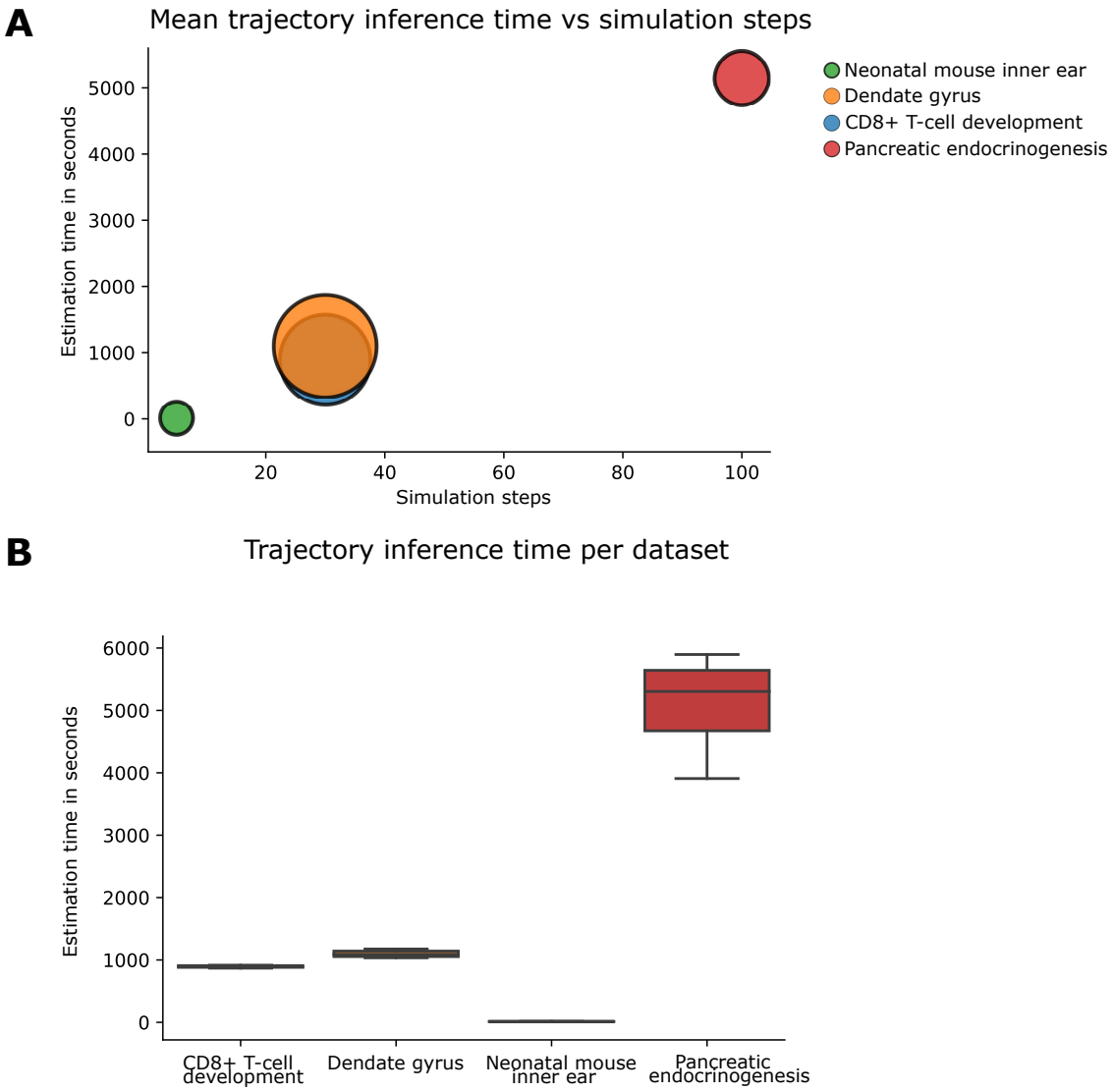

Figure S3: Dependence of estimation time on number of simulation steps and size of the dataeset.

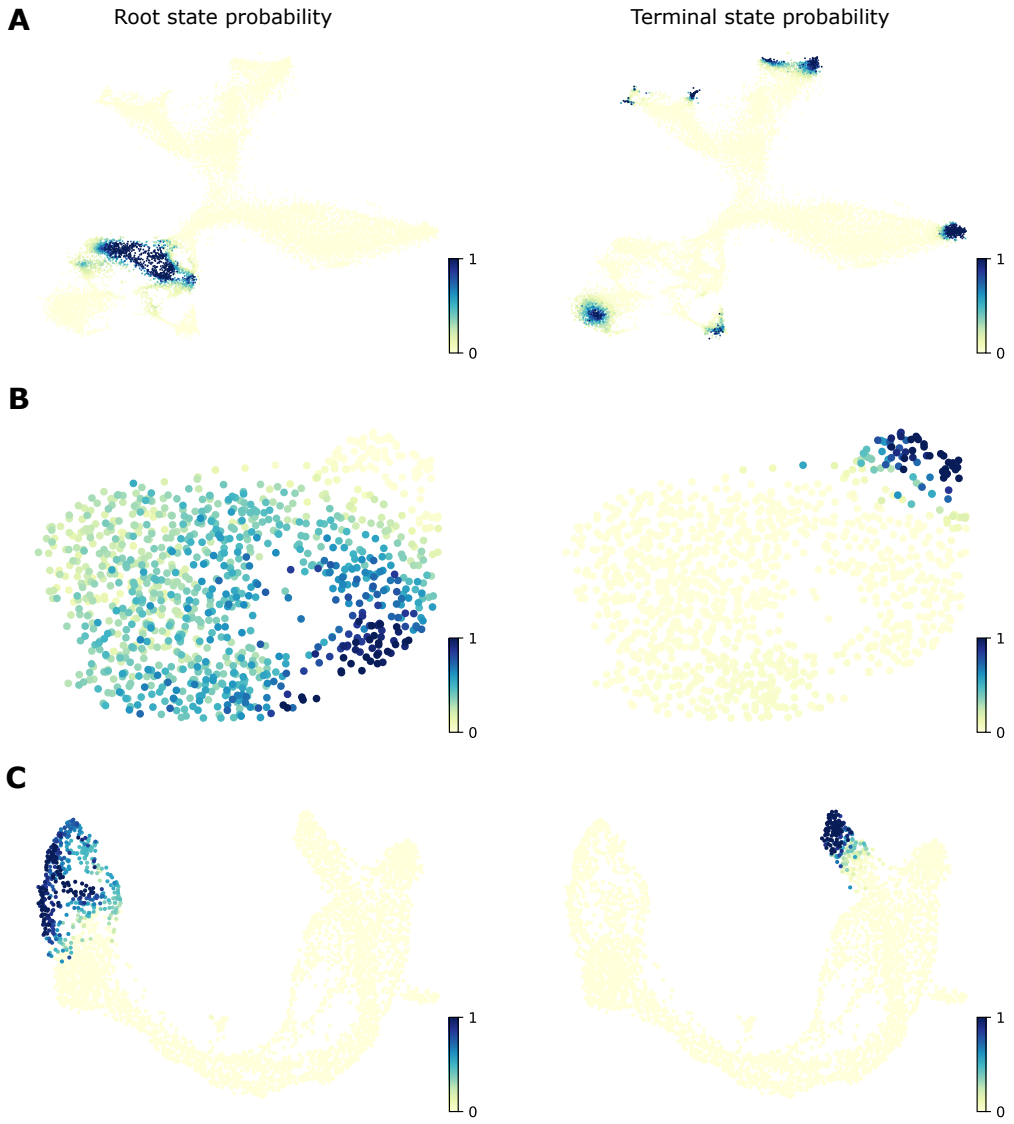

Figure S4: Root and terminal state probability estimation based on RNA velocity (A-C) Root and terminal state probability of cells belonging to (A) Dendate gyrus, (B) cell cycle, (C) pancreatic endocrinogenesis datasets.

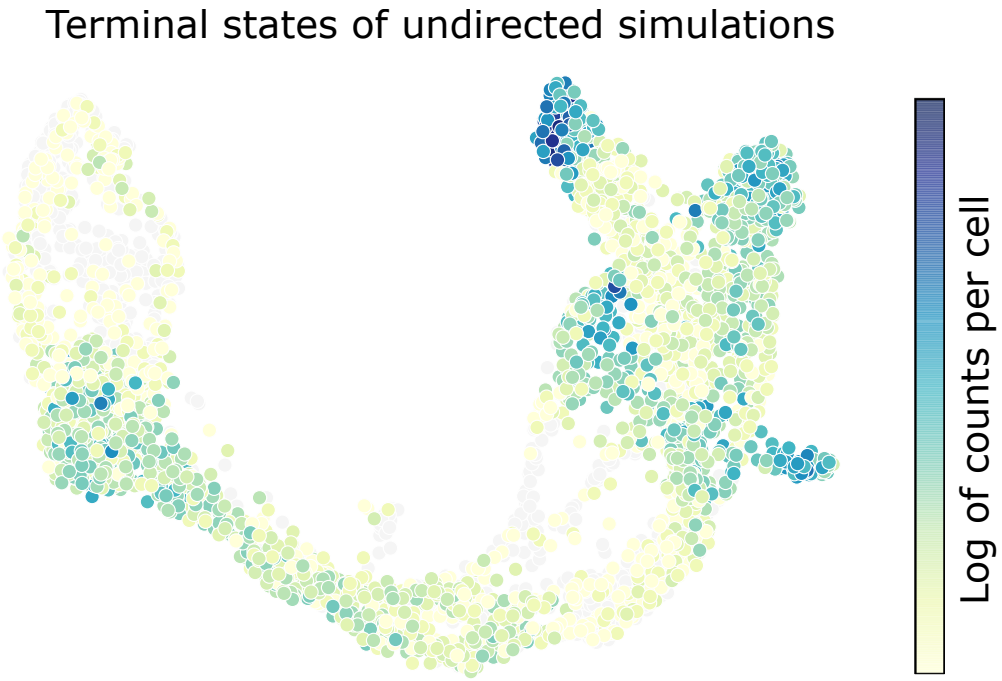

Figure S5: Terminal cell states of undirected simulations initiated at each cell type. Two thousand simulations were initiated with each cell type cluster successively considered as root. For each cell the total count of simulations (from any root) terminating at that cell were counted. The log of these counts has been annotated on the UMAP above.

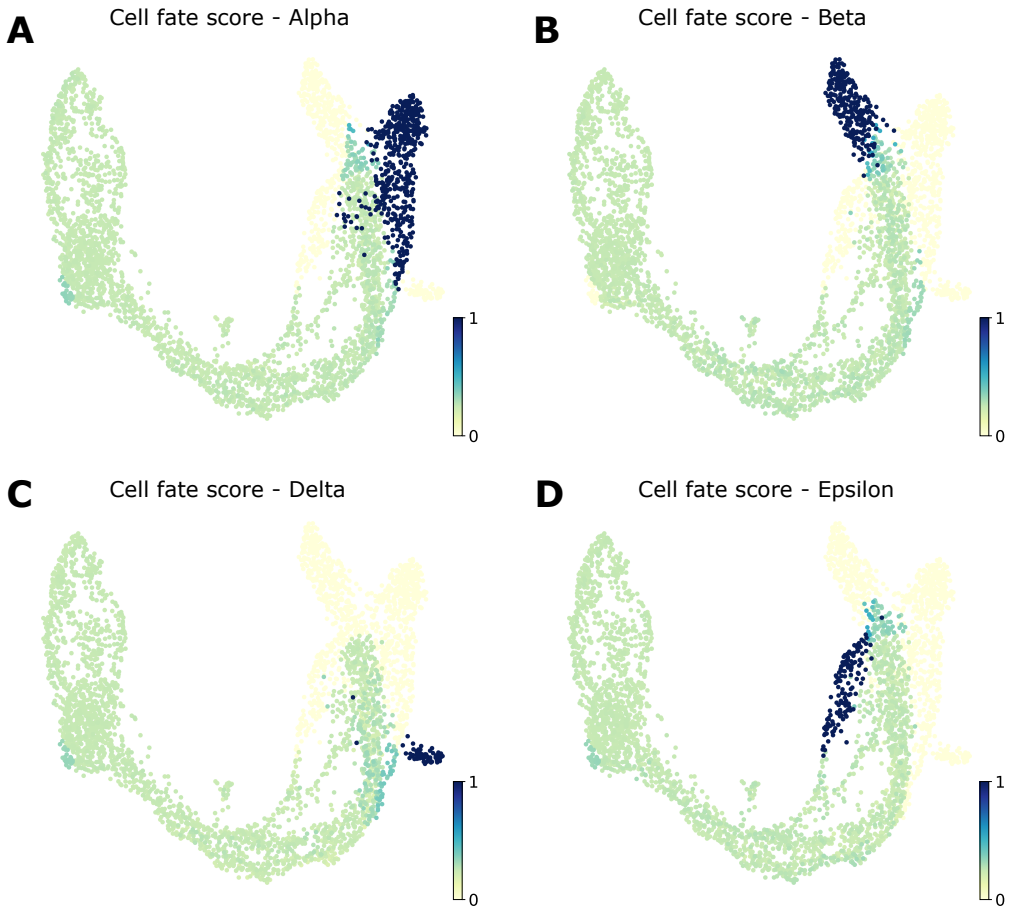

Figure S6: Cytopath estimation of cell fate based on relative alignment of cells. (A-D) Cell fate score per cell of (A) Alpha, (B) Beta, (C) Delta and (D) Epsilon cell type.

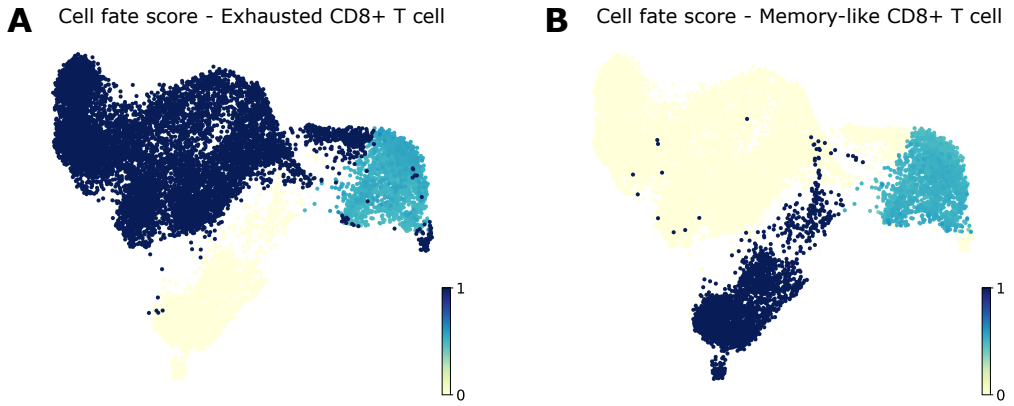

Figure S7: Cytopath estimation of cell fate based on relative alignment of cells. (A-B) Cell fate score per cell of (A) Exhausted CD8+ T cells, (B) Memory-like CD8+ T cells.

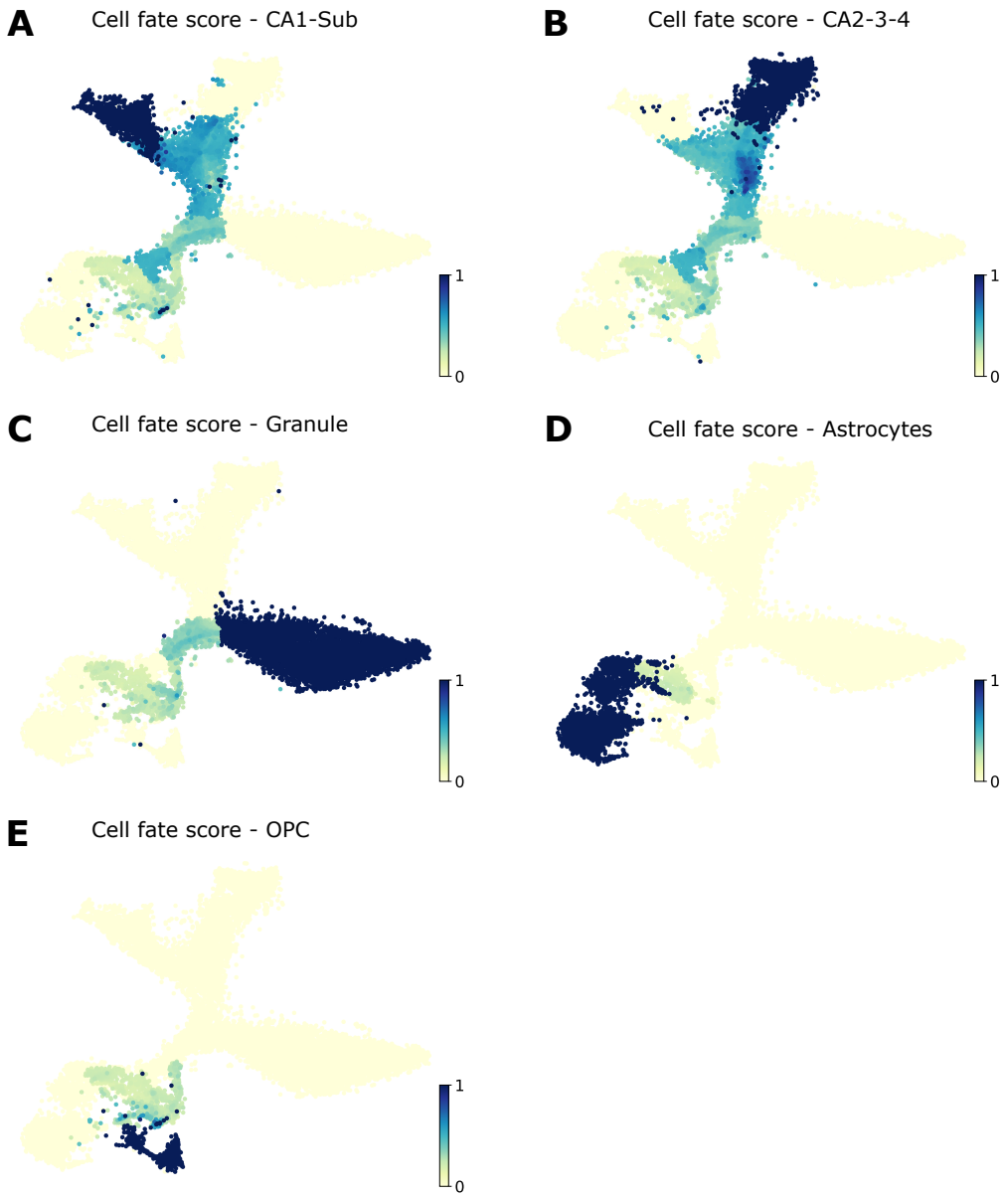

Figure S8: Cytopath estimation of cell fate based on relative alignment of cells. (A-E) Cell fate score per cell of (A) CA1-Subculum, (B) CA2-3-4, (C) Granulocytes, (D) Astrocytes and (E) OPC cell type.

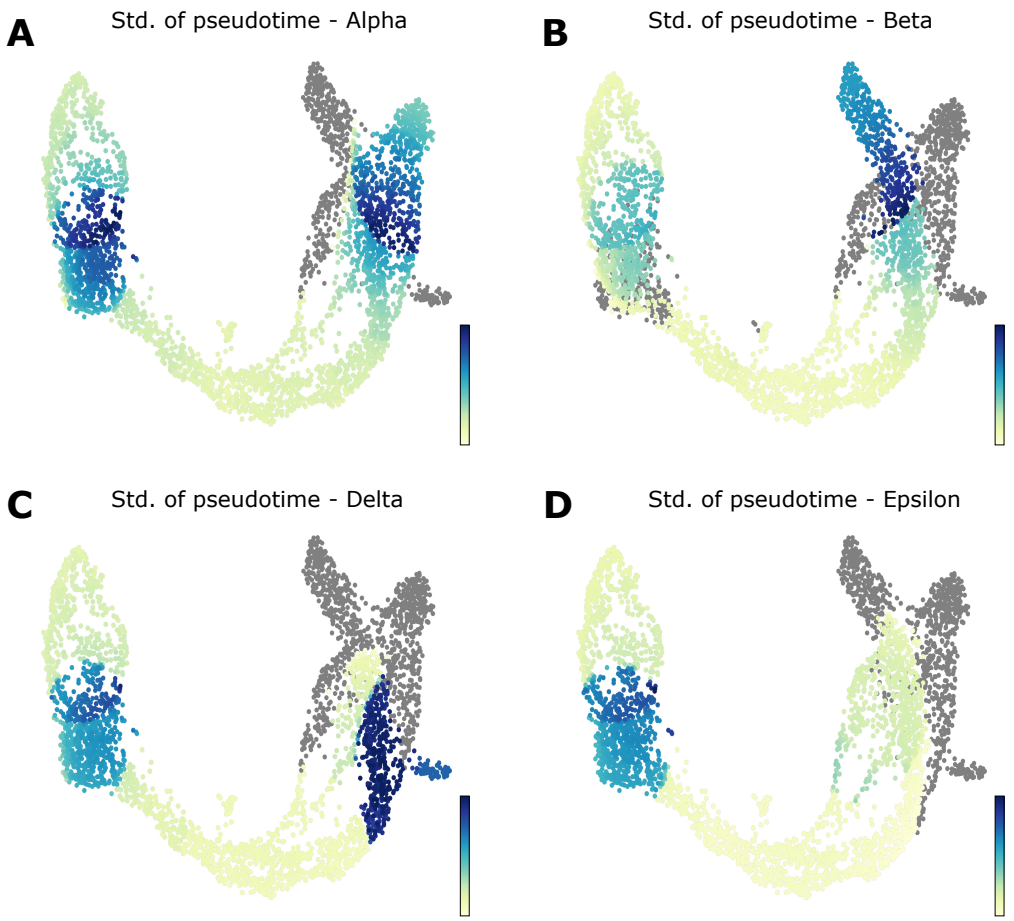

Figure S9: Standard deviation of the distribution of pseudotime values which arises from alignment of cells to multiple segments of each trajectory. (A-D) Standard deviation of pseudotime per cell of (A) Alpha, (B) Beta, (C) Delta and (D) Epsilon cell type.

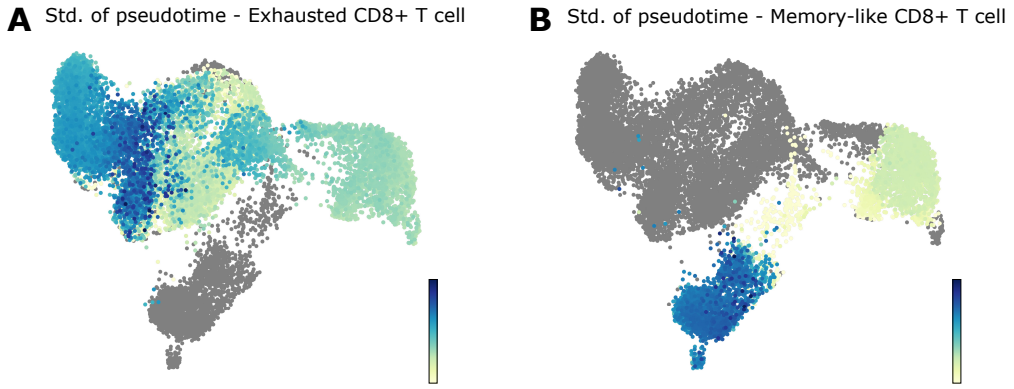

Figure S10: Standard deviation of the distribution of pseudotime values which arises from alignment of cells to multiple segments of each trajectory (A-B) Standard deviation of pseudotime per cell of (A) Exhausted CD8+ T cells, (B) Memory-like CD8+ T cells.

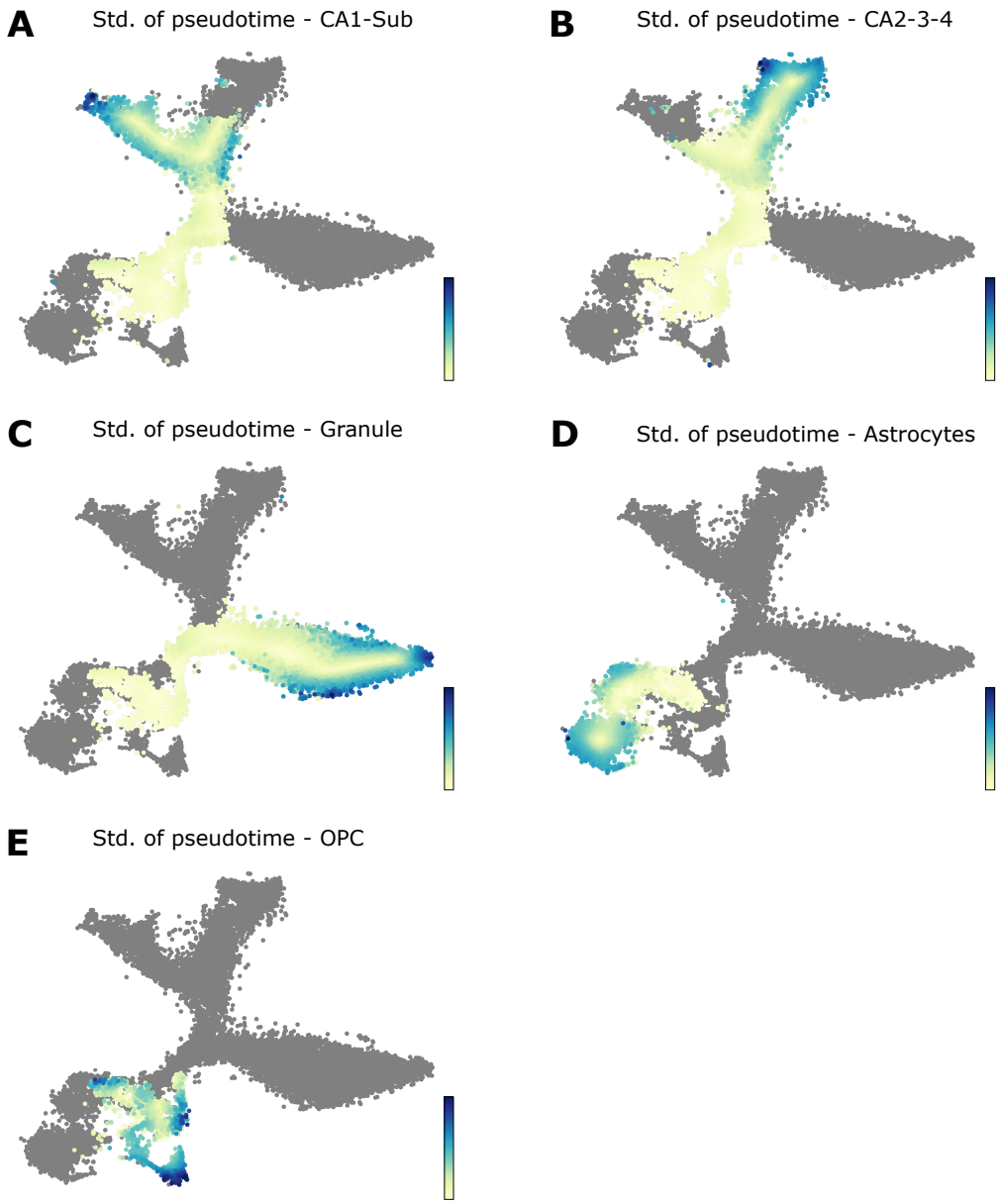

Figure S11: Standard deviation of the distribution of pseudotime values which arises from alignment of cells to multiple segments of each trajectory (A-E) Standard deviation of pseudotime per cell of (A) CA1-Subculum, (B) CA2-3-4, (C) Granulocytes, (D) Astrocytes and (E) OPC cell type.

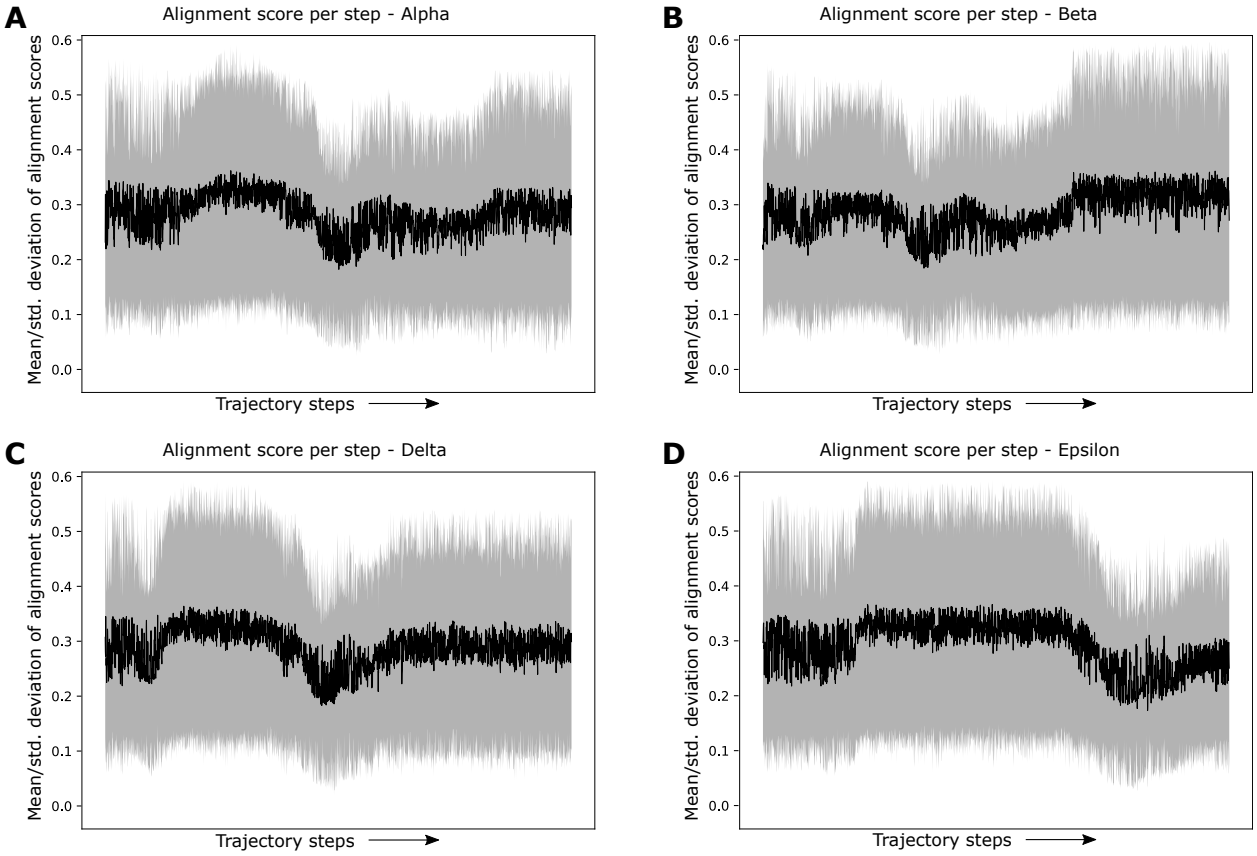

Figure S12: Mean and standard deviation of alignment score per trajectory step. (A-D) Alignment score per step of (A) Alpha, (B) Beta, (C) Delta and (D) Epsilon cell type.

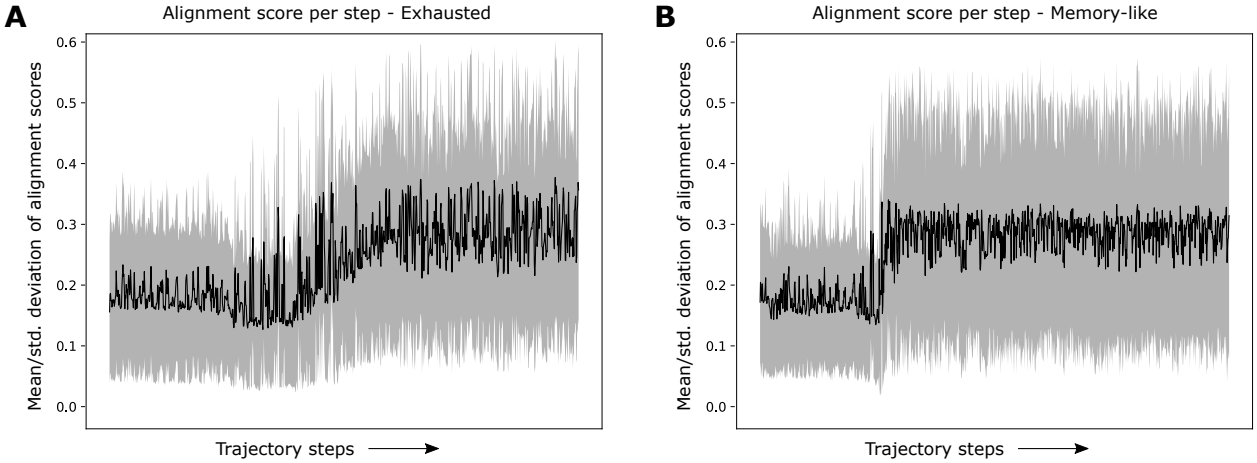

Figure S13: Mean and standard deviation of alignment score per trajectory step. (A-B) Alignment score per step of (A) Exhausted CD8+ T cells, (B) Memory-like CD8+ T cells.

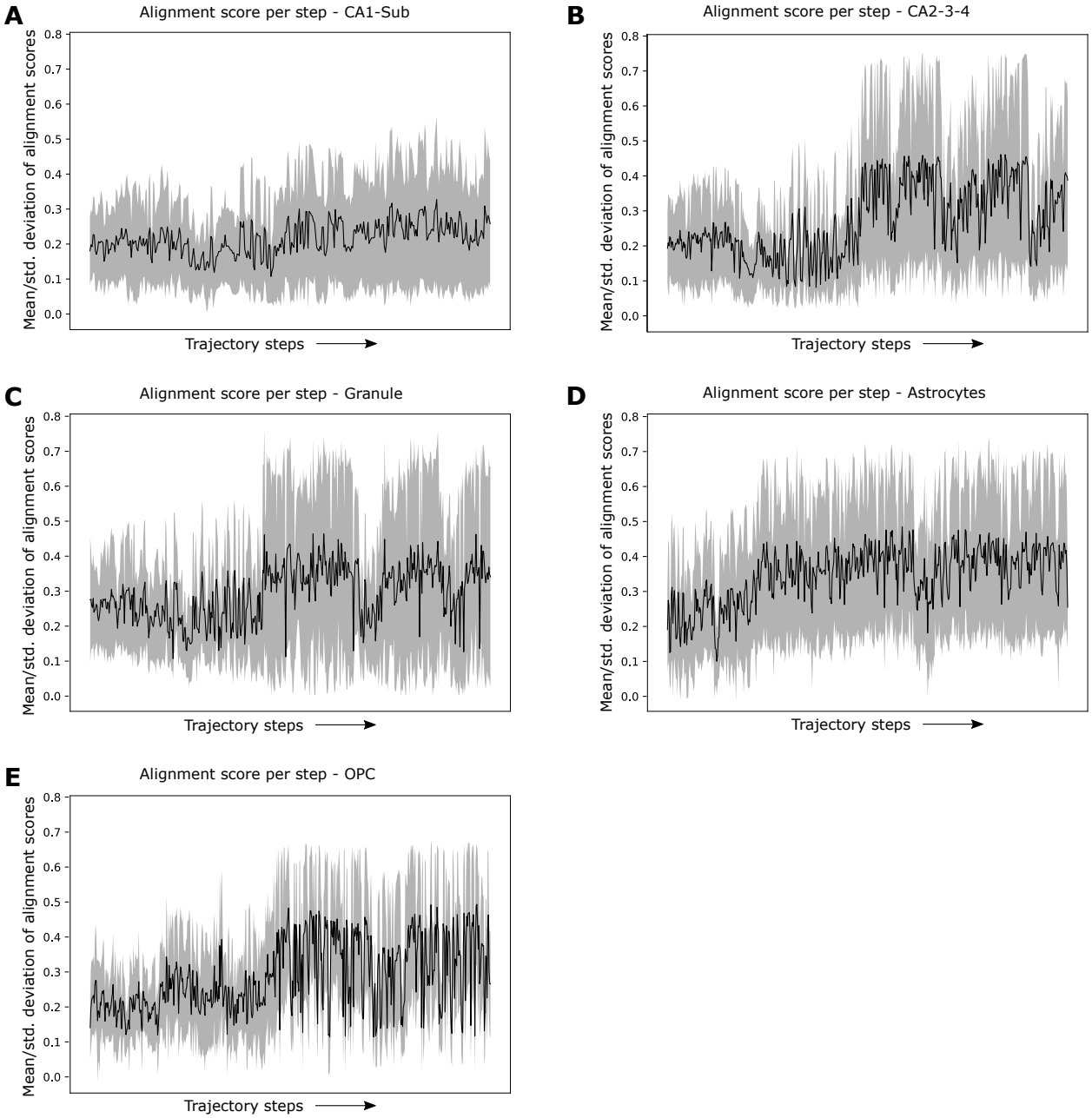

Figure S14: Mean and standard deviation of alignment score per trajectory step. (A-E) Alignment score per step of (A) CA1-Subculum, (B) CA2-3-4, (C) Granulocytes, (D) Astrocytes and (E) OPC cell type.

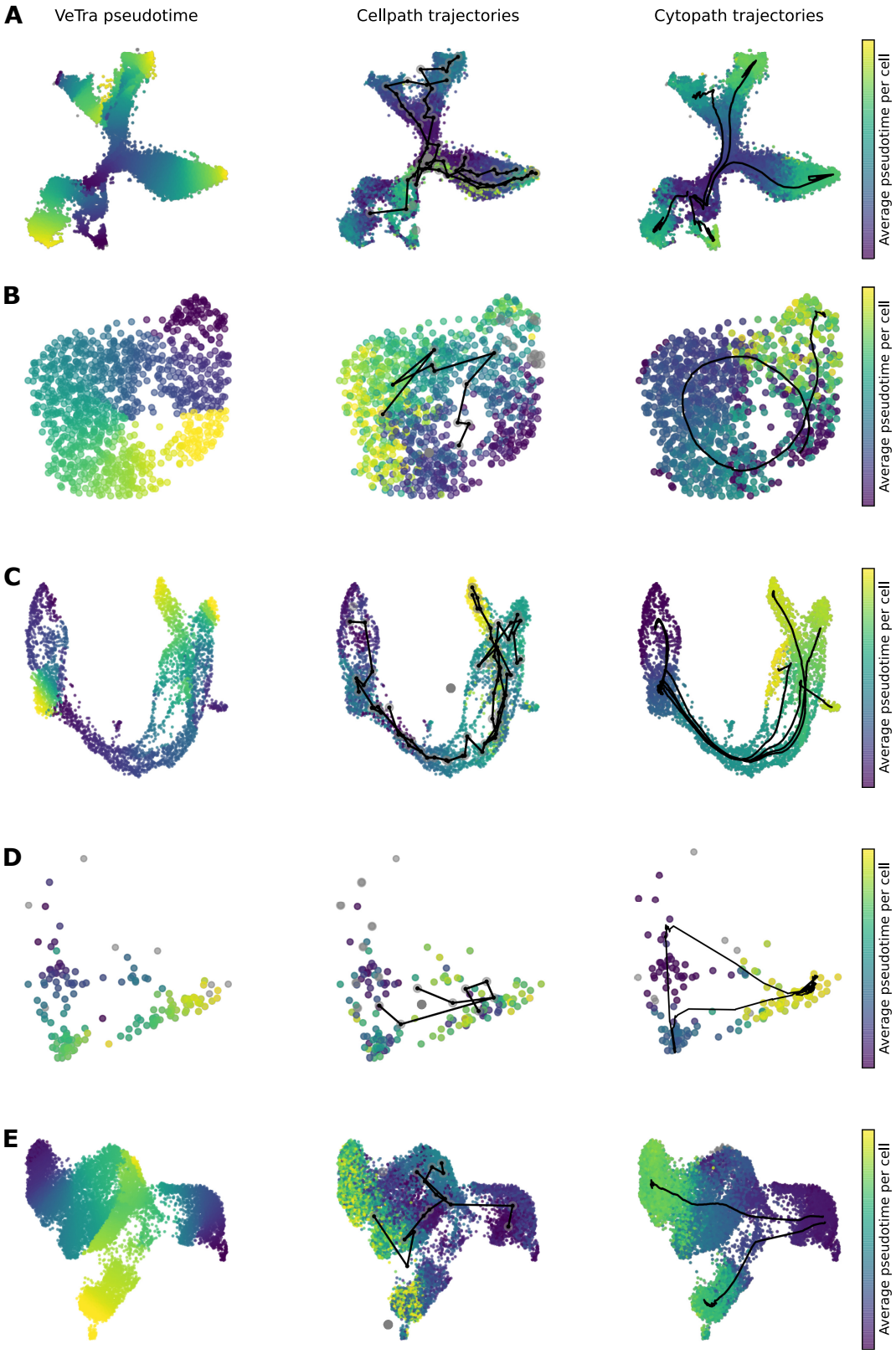

Figure S15: Average pseudotime and trajectories inferred by RNA velocity based trajectory inference methods. (A-E) Average pseudotime by VeTra, Cellpath and Cytopath for (A) Dendate Gyrus, (B) cell cycle, (C) pancreatic endocrinogenesis, (D) mouse inner ear and (E) CD8+ T cell exhaustion datasets. Expected behaviour is that all terminal cell states have the highest pseudotime and root states have the lowest pseudotime. Erroneous trajectories initialised at terminal or intermediate states will lead to a patch like appearance.

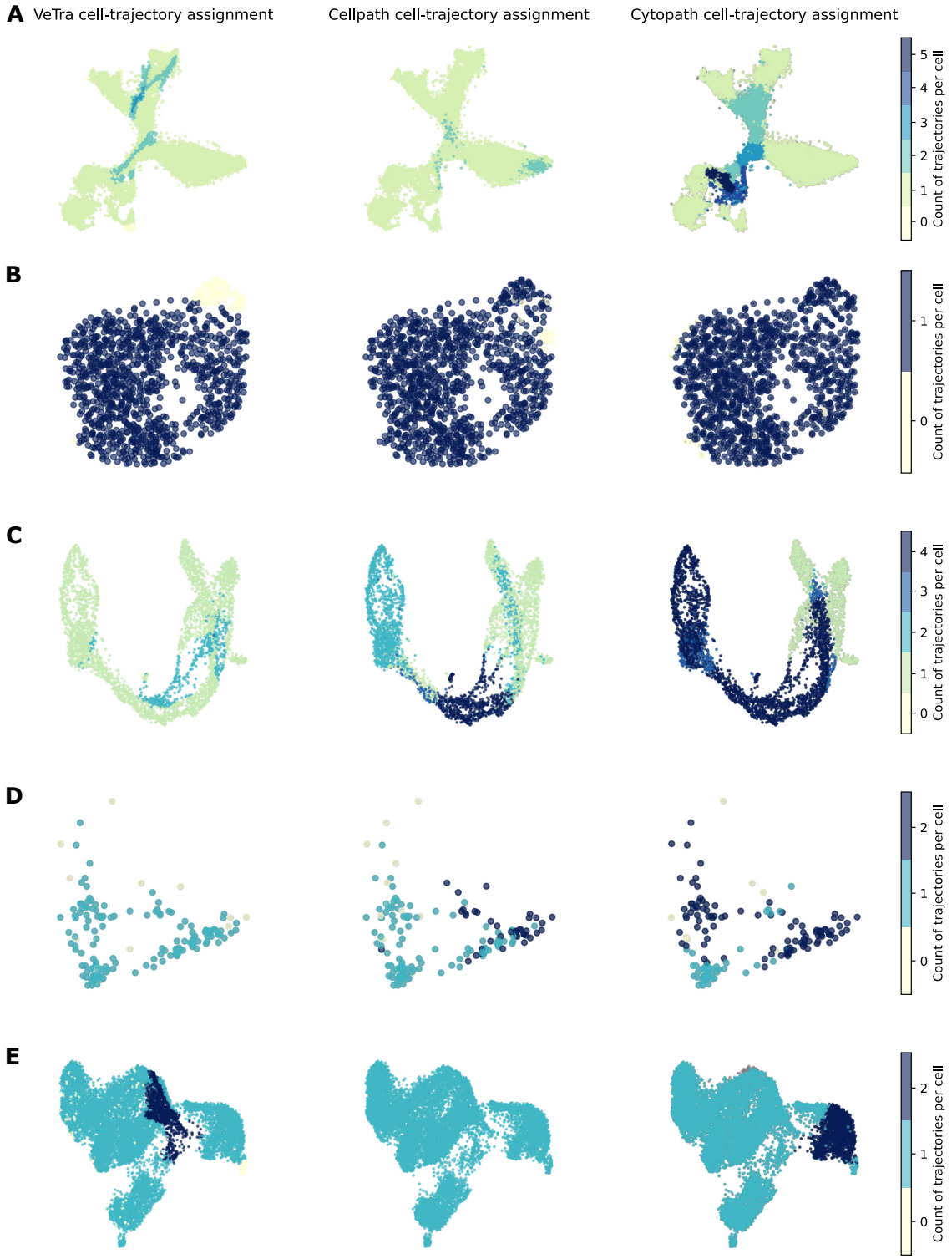

Figure S16: Count of trajectories to which each cell is assigned by RNA velocity based trajectory inference methods. (A-E) Count of trajectories to which each cell is assigned by VeTra, Cellpath and Cytopath for (A) Dendate Gyrus, (B) cell cycle, (C) pancreatic endocrinogenesis, (D) mouse inner ear and (E) CD8+ T cell exhaustion datasets. Expected behaviour is that typically terminal state cells are assigned to only one trajectory in a multi-terminal state dataset. Cells at the root are typically assigned to all trajectories with each branching event lowering the count. Zero count indicates that the cell was missed by the trajectory inference process entirely.