Gene Trajectory Inference for Single-cell Data by Optimal Transport Metrics

Rihao Qu[1,2,3*], Xiuyuan Cheng[4*], Peggy Myung[2,5], Esen Sefik[3], Richard Flavell[3,6], Yuval Kluger[2,7†]

[1]Computational Biology & Bioinformatics Program, Yale University, New Haven, CT, USA
[2]Department of Pathology, Yale University, New Haven, CT, USA
[3]Department of Immunobiology, Yale University, New Haven, CT, USA
[4]Department of Mathematics, Duke University, Durham, NC, USA
[5]Department of Dermatology, Yale University, New Haven, CT, USA
[6]Howard Hughes Medical Institute, Yale University, New Haven, CT, USA
[7]Program of Applied Mathematics, Yale University, New Haven, CT, USA

Abstract

Single-cell sequencing technologies have enabled researchers to study dynamical biological processes, such as cell type differentiation, by constructing single-cell trajectories. Trajectories connecting the transcriptomes of groups of cells are typically inferred from cell manifolds embedded in high-dimensional space. To study the mechanisms underlying a biological process, genes whose expression patterns change smoothly along cell trajectories are extracted. However, when a mixture of biological processes coexist in the system, they often cannot be effectively de-mixed using standard cell trajectory approaches. This poses a major challenge to identifying genes that mediate each process and delineating their temporal order. Here, we present a new approach, `GeneTrajectory`, to unravel gene trajectories associated with distinct biological processes. To infer gene trajectories, we first compute a cell-cell graph that preserves the manifold structure of the cells. Then, we construct a gene-gene graph where the affinities between genes are based on the Wasserstein distances between their distributions on the cell graph. Notably, this method does not require any prior knowledge of the cell lineages or gene players involved. We use simulated and two real-world scRNA-seq datasets to demonstrate that `GeneTrajectory` reveals the gene dynamics underlying complex biological processes that cannot be resolved by standard cell trajectory inference.

1 Introduction

Advances in single-cell transcriptomics research allow investigators to study fundamental dynamical biological processes, e.g, cell cycle [15], embryonic tissue/organ differentiation [19], or T cell development in tumor microenvironment [29]. In particular, finding key genes and the order in which they come into play during biological processes is important. Many computational methods have been developed to infer cell trajectories [23, 6, 27, 25, 9], which organize cells into lineages and assign pseudotime to the cells within each lineage. Genes that are associated with each cell trajectory can be then identified and ordered based on their changing patterns along the cell pseudotime [4]. However, if a mixture of biological processes coexist in the system, the cell manifolds in high-dimensional space cannot be effectively reduced to

*These authors contributed equally.
†Corresponding author. Email: yuval.kluger@yale.edu

locally-linear trajectories, and hence it prevents practitioners to reliably determine the pseudotime of cells and the corresponding pseudo-temporal order in which genes are activated along various lineages. Here, we develop `GeneTrajectory`, a new method to detect gene trajectories without any prior knowledge of the cell lineages or gene players involved. This method allows to identify for each process a trajectory that organizes relevant genes in a pseudo-temporal manner, unravelling the mechanisms underlying complex biological systems.

In order to construct the gene trajectory and recover the pseudotime of genes, it is essential to infer the geometric relationships between genes. One way to address this problem is to model each gene as a distribution (where the distribution weight on each cell is proportional to gene expression level) over cell space and then measure dissimilarity between gene pairs as divergence between these distributions. The perspective of viewing each gene as a distribution over the space of cells has been previously considered [26]. In [26], Vandenbon et al. aimed to identify the genes whose distributions are localized in the cell space using an estimated Kullback–Leibler divergence between gene distributions. In our work, we view genes as distributions over a cell graph (a graph of cell-cell affinities which encodes the cell space geometry) and focus on the gene dynamic aspects: we use the pairwise optimal transport (OT) metric between gene distributions to construct a low-dimensional embedding of the genes. In this construction, pairs of genes activated consecutively along a biological process are expected to have a small OT distance between their distributions. Equipped with OT distances between gene pairs, we construct a gene graph and then use Diffusion Map [7] to extract gene trajectories. Each trajectory is associated with a specific biological process and reveals the pseudo-temporal order in which genes are activated.

Previously, OT metrics have been applied in a wide range of scenarios in single cell analysis, including 1) defining a similarity measure between cells [13, 2] or cell populations [18], 2) constructing cell trajectories [20, 28], 3) spatial reconstruction of single-cell transcriptome profiles [5, 16], and 4) multi-omics data integration [10]. In these works, dissimilarities were computed using Wasserstein distances between pairs of cell distributions, or between the normalized gene expression profiles of cell pairs. In our work, we first construct a low-dimensional representation of cells which preserves the manifold structure of cell geometry in the original high dimensional space. We then associate each gene to a gene distribution on the cell graph, which is the normalized gene expression over the cells. Subsequently, we compute the Wasserstein distances between the gene distributions. For clarity, we note that if the Wasserstein distance between two gene distributions is small, these two genes have similar (normalized) expression profiles across the cell graph. Because the Wasserstein distance by definition incorporates the distance between cells when defining a distance between genes, our method can be viewed as leveraging the geometric information of the cells to infer the geometry of the genes.

In practice, the computation of Wasserstein distance between high dimensional distributions can be computationally expensive, therefore fast algorithms need to be implemented. Previous works [22, 8, 24, 17] have proposed several ways to improve the computation efficiency based on certain assumptions. In our workflow, we devise two strategies (optional to users) to achieve this goal, cf. Section 2.2. Efficient computation is crucial for applying our method to datasets at the scale of single-cell data.

2 Approach

2.1 Step 1. Cell graph construction and computation of intrinsic distances between cells

In our workflow (Figure 1), the essential question is how to appropriately define the transport cost from one cell to another which can be used to compute the Wasserstein distance between gene distributions on the cell graph (see Section 2.2). We reckon that, when cell geometry presents low-dimensional manifold structure, the optimal transport should be always done across the cell manifold instead of taking a shortcut through empty regions in the ambient space where there are no cells. Hence, we need to build a cell graph

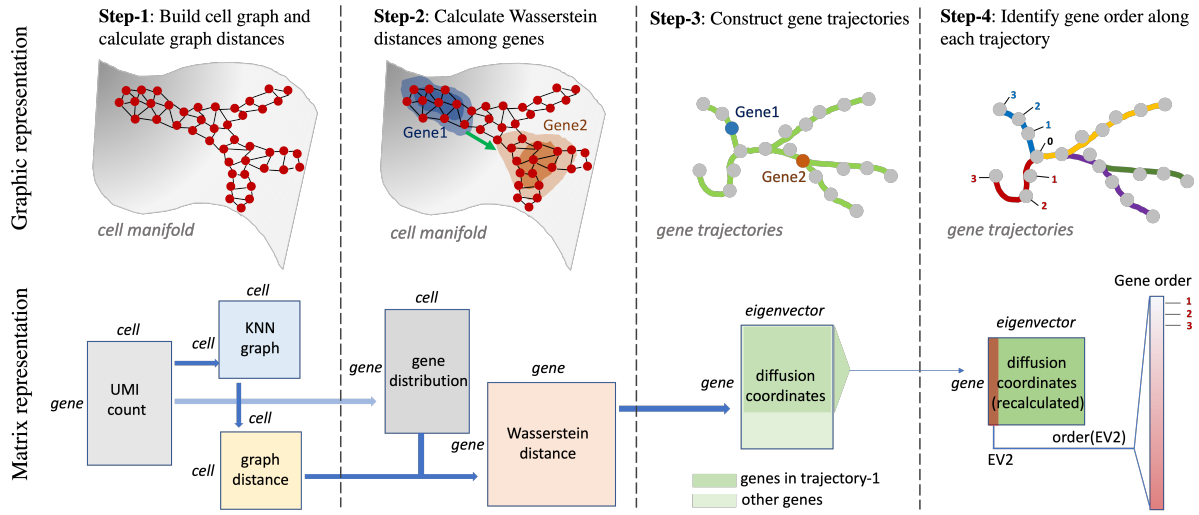

Figure 1: The major workflow of our approach. **Step 1**. After preprocessing the original cell-by-gene count matrix, cell kNN graph is constructed based on a low-dimensional representation of cells, e.g., PCA or the diffusion coordinates by Diffusion Map. As the input to Step 2, graph distances between cells are computed to define the costs of transporting a gene distribution on the cell graph. **Step 2**. Pairwise gene-gene Wasserstein distances between pre-selected genes are computed, which quantifies the minimum cost of transporting the distribution of a given gene into the distribution of another gene. **Step 3**. Gene embedding is constructed by computing Diffusion Map using the gene-gene Wasserstein distance matrix. Gene trajectories are then identified in a sequential manner. For each trajectory, the gene with the largest distance from the origin of the Diffusion Map embedding is selected as a starting point. We then use a graph diffusion-based approach to determine the genes that belong to the trajectory with the terminus given by that starting-point gene. **Step 4**. For each trajectory, we recompute a subgraph involving only genes within that trajectory. By using Diffusion Map to construct the diffusion embedding, those genes are then ordered by sorting the values of the first nontrivial eigenvector (EV2).

that can capture the cell manifold structure. There are multiple options to obtain a dimension reduced representation of cells for constructing such a graph. For example, one can use PCA or Diffusion Map for the dimension reduction of cells, and then construct the kNN cell graph based on their low-dimensional coordinates. For a given pair of cells u and v, we search for the shortest path connecting them in the kNN cell graph and denote its length as the graph distance $(d_G)_{uv}$ between cells u and v. The graph distance constructed in this way reveals manifold geodesic distance, which is the intrinsic distance between pairs of cells, in the limit of large number of cells, as has been previously shown in the manifold learning literature [1, 3].

2.2 Step 2. Computation of Wasserstein distance between genes

We model the expression level of different genes as a discrete distribution on the cell graph, and define the distance between two genes by the Wasserstein-p distance (W_p-distance) on the graph, $1 \leq p < \infty$. The W_p-distance between genes i and j can be computed by solving a discrete optimal transport mapping over the cell graph, with a transport cost between cell u and v defined as $C_{uv} = (d_G)^p_{uv}$, see details in the supplementary information (SI) A.1.

To overcome the burden of computing gene-gene W_p-distances, we have designed two strategies as described below to accelerate the computation. While Strategy 1) can be made extremely fast with a crude coarse-graining, it may lose accuracy, which is remedied in Strategy 2).

1) We coarse-grain the cell graph by aggregating cells using K-means clustering algorithm. Specifically,

cells are partitioned into K subsets, where cells in the same subset are then aggregated into a "meta-cell" positioned at the centroid of the subset in the diffusion embedding space. Each gene is now formulated as a distribution on the coarse-grained meta-cell graph.

2) We sparsify the gene affinity graph by setting to zero the entries where the pair-wise Wasserstein distance between genes is greater than a threshold (the resulting affinity is exponentially small and thus contributes negligibly to the gene affinity graph). The distant pairs are identified using the approximate Wasserstein distance on a coarse-grained cell graph (Strategy 1) which allows fast computation. The Wasserstein distances for the remaining (un-sparsified) gene pairs are then computed on the fine-grained cell graph.

More details can be found in SI A.2 and A.3.

2.3 Step 3. Gene trajectory construction and trajectory identification

After we obtain the gene Wasserstein distance matrix, we convert it into an affinity matrix (using a locally adaptive Gaussian kernel) and use Diffusion Map to get a low-dimensional embedding of genes. We then identify gene trajectories in a sequential manner. Specifically, to retrieve the first trajectory we use a procedure of selecting the starting point on the gene graph followed by a diffusion on the gene graph from this initial node:

- *Selection of the initial node.* We denote by $M(g_i)$ the spectral coordinates of gene i computed by Diffusion Map, retaining the top d nontrivial eigenvectors. We choose the gene with the largest embedding norm $||M(g_i)||_2$ as the starting point of a diffusion on the gene graph.
- *Propagation on the gene graph.* The diffusion is performed by propagating a point mass from the starting point in the gene graph, where the transition matrix of the random walk is constructed using $M(g_i)$ the spectral coordinates of genes. We run the diffusion up to T times in the gene graph and then group the genes with probability weights larger than a given threshold as members of that trajectory. T is a tunable parameter which we select by a simple heuristic.

After the first gene trajectory is identified, we remove its genes from the gene graph, and repeat the procedure on the remaining genes. This algorithm retrieves a sequence of gene trajectories until a predefined number of trajectories are identified.

2.4 Step 4. Gene ordering along each trajectory and visualization

To determine the gene ordering along a given gene trajectory, we first extract the corresponding sub-matrix of gene-by-gene Wasserstein distances as computed in Section 2.2. We then re-compute the Diffusion Map on the sub-graph only to obtain the spectral embedding of genes in that trajectory. The first non-trivial eigenvector of the Diffusion Map embedding provides an ordering of the genes within this trajectory. Assuming genes in the trajectory approximately lies on a one-dimensional sub-manifold, the first non-trivial eigenvector reveals the ordering of genes according to the spectral convergence theory of Diffusion Map [7, 21].

To better visualize each gene trajectory, we select two or three spectral coordinates computed on the full gene-gene graph which are most variant across the entries associated with that trajectory. Details are explained in SI A.4 and results are shown in Figure A.1 and Figure A.2.

3 Experiments

In this section, we present experimental results by applying the proposed workflow on simulated datasets (designed to mimic representative biological processes, cf. Section 3.1) and two real scRNA-seq datasets

(cf. Sections 3.2 and 3.3). The implementation of our algorithm is released at `https://github.com/KlugerLab/GeneTrajectory`.

3.1 Simulation experiments with various types of cell graph topology

Here we apply our workflow to three simulation examples with different types of cell graph topology (Figure 2). The examples are inspired by representative biological processes such as cell differentiation, cell cycle and their combination.

To simulate the distributions of all genes in each of the three examples, we first construct the random-walk matrix on the cell graphs and then generate gene distributions using graph diffusion. More details of the simulation procedure can be found in SI A.5. The locations of the cell graph nodes (in 2D for Examples 1 and 2, and in 3D for Example 3) are shown in the second column of Figure 2. The gene distributions are constructed to be local on the graph in the first two examples, and non-local in the third example. Specifically, in Example 3, two classes of gene distributions are generated: The first class of distributions are localized in the x-y section of the cylinder and spread through z-axis, and the second class of distributions are localized in the z-coordinate and spread through the x-y plane on the cylinder. The generated gene distributions on cell graphs are shown in the third column of Figure 2.

After the simulated data are prepared, we applied our workflow as described in Section 2 to construct the gene trajectories. As shown in each of the first two rows of Figure 2, the gene trajectory structure found in Example 1 and 2 respectively resembles the cell trajectory structure, since gene distributions are localized on the cell graph. Each of the identified gene trajectories also recovers the simulated order of genes on the cell manifold. In Example 3, we identified two gene trajectories. The blue ring-shaped trajectory mimics the cell cycle process and the red curve trajectory mimics differentiation process. In this example, the gene trajectories have a different topology which is not present in the cell manifold, because gene distributions are not localized on the cell graph. The results show that our method can successfully capture the ordering of genes in all three examples where in each example the cell graph presents a different topology. In particular, as shown by Example 3, the method can detect gene trajectories corresponding to two distinct simulated biological processes, which one cannot disentangle from the cell graph itself.

3.2 Maturation processes of various cell types in hippocampus dentate neurogenesis

In a previous study [12], Hochgerner et al. collected mouse hippocampus dentate cells from different stages of development, including perinatal, juvenile, and adult stages. They inferred four different developmental cell trajectories corresponding to the maturation process of granule cells, pyramidal cells, astrocytes, and oligodendrocytes. In this example, we extracted the cells involved in these maturation processes (Figure 3A) from the original dataset and constructed the gene trajectories using `GeneTrajectory` (Figure 3B). We identified four different gene trajectories, as visualized using several leading nontrivial nontrivial Diffusion Map eigenvectors (Figure 3B). The order of gene along each trajectory is visualized in Figure 3C.

To detect which cell populations are involved along each gene trajectory, we constructed a movie-like series of frames on the cell embedding (Figure 3D). Each frame is colored according to the average expression level of genes within the respective bin, where bins are ordered successively along the gene trajectory (e.g., bin-1 comprises genes starting at one end of the trajectory, and genes in bin-5 are close to its other end). This displays each trajectory is associated with a specific biological process. Careful examination of the genes and their order along trajectory-1 link this trajectory to astrocyte maturation. *Hes5* precedes *Glul, Aldoc, Aqp4* along the gene trajectory. This is consistent with the prior knowledge that *Hes5* is activated in the process of astrocyte differentiation while *Glul, Aldoc, Aqp4* are characteristic of mature astrocytes [14]. Trajectory-2 and 4 correspond to the maturation process from neuroblasts to granule cells and pyramidal cells respectively. Along trajectory-4, *Sox11* and *Calb2* are activated in the beginning,

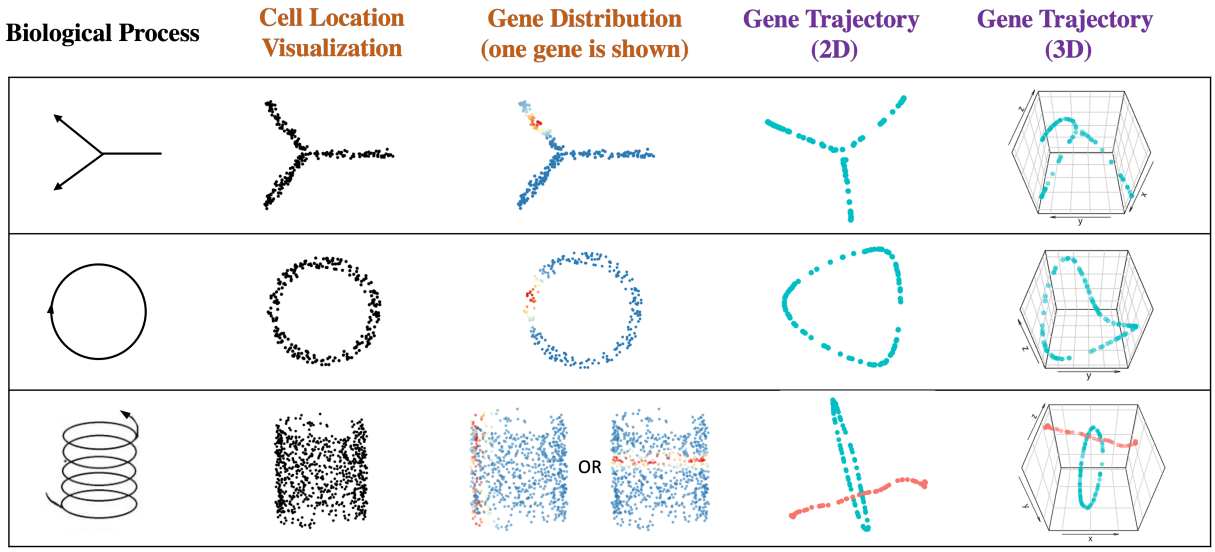

Figure 2: Simulation-1: A bifurcated tree-structured cell trajectory in a two-dimensional space. Tree-structured trajectories are commonly seen in the biological systems where multiple cell types are differentiated from the same pool of progenitors. **Simulation-2**: A ring-shaped cell trajectory in a two-dimensional space, which is typically representative of the cell cycle process. **Simulation-3**: A cylinder-structured cell manifold in a three-dimensional space. A mixture of two biological processes is represented, which includes cell cycle (x-y plane) and cell differentiation (z-axis). This can be seen when cell proliferation and differentiation is coupled with each other, which is a common phenomenon in the early stages of embryonic development. (Each dot in the visualization of cell location and gene distribution represents a cell, while each dot in the gene trajectory panels (2D or 3d) represents a gene.)

prior to the expression of the terminal pyramidal marker, *Neurod6*. Trajectory-3 delineates the maturation process of oligodendrocytes, where the transition from oligodendrocyte precursor cells (OPC) to newly formed oligodendrocytes (NFOL) and mature oligodendrocytes (MOL) is captured. `GeneTrajectory` not only identifies gene players that contribute to each biological process but also captures a step-by-step view of the developmental programs in maturation of astrocytes, neuroblasts, granule and pyramidal cells.

3.3 Comparative gene trajectory analysis of dermal condensate genesis during the early mouse embryonic development

At the early stage of mouse embryonic development, dermal condensates (DC) emerge around day-14.5 (E14.5) which marks an essential step in the formation of hair follicles. Previous studies [11, 19] have shown that an intricate regulation of morphogen signaling is necessary to guide the differentiation of DC cells. In this example, the single-cell RNA-seq data was retrieved from a published dataset [19], including the E14.5 dermal cell populations from a wildtype mouse embryo and a paired epidermal β-catenin knockout mouse embryo. The knockout of β-catenin causes the depletion of placodes in the epidermis and specifically blocks the downstream differentiation process of DC (Figure 4A, D). We applied our approach to construct the gene trajectories of the wildtype and knock-out samples respectively (Figure 4B-C, 4E-F) and compared the differences in matched gene trajectories (Figure 4H).

Notably, in contrast to Example 3.2, application of `GeneTrajectory` to the wildtype sample unraveled three prominent gene trajectories that correspond to the differentiation of known dermal cell types (lower dermis (LD) and DC) as well as to cell cycle (Figure 4C). For each of these gene trajectories we identified several established markers associated with the respective process. For example, Wnt-signaling pathway

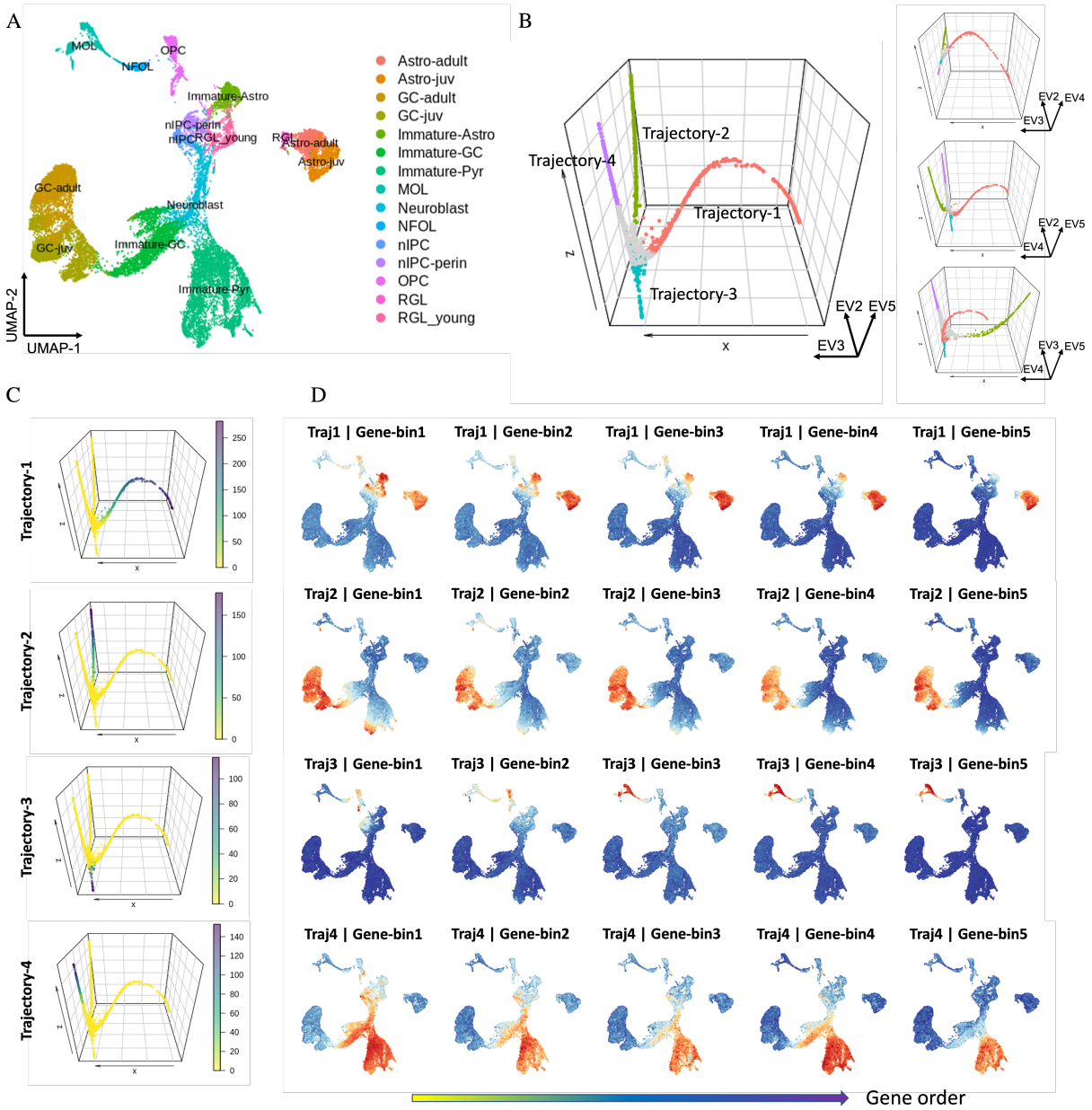

Figure 3: Gene trajectory construction and dissection of hippocampus dentate neurogenesis. **A)** UMAP embedding of 15 pre-annotated cell types that form the landscape of hippocampus dentate neurogenesis. **B)** Gene trajectories visualized by the combination of leading nontrivial Diffusion Map eigenvectors. Four gene trajectories are identified with each corresponding to a distinct hippocampus dentate celltype maturation process. **C)** Gene order along each trajectory is indicated by the color bar. (The gene at the terminus of each trajectory has the largest value.) **D)** Cell UMAP embedding is colored by the average gene expression of genes within the respective bin, where bins are ordered successively along the gene trajectory. These movie-like frames reveal different transport patterns of gene distributions.

genes (e.g., *Lef1*, *Dkk1*) and SHH-signaling pathway genes (e.g., *Ptch1*, *Gli1*) are present in the DC gene trajectory and are known to be essential for DC differentiation. Further, the order in which the genes appear along each trajectory was concordant with the relative stage in which they are known to be activated along each process, including the cell cycle (e.g. the beginning of the cell cycle trajectory is enriched for genes associated with S-phase, while the distal portion of the trajectory is enriched for genes associated with G2/M phase).

Similar with the previous example in Section 3.2, inspecting the movies of the cell embedding (for each gene trajectory) reveal a continuous progression (Figure 4G), which would be hard or impossible to infer using standard cell trajectory inference methods. However, gene trajectories in this section reveal a different and coherent topology when compared with cell trajectories. This phenomenon is clearly illustrated in the third simulation in Section 3.1. It implies that if a cell population is involved in multiple concurrent biological processes (e.g., cell cycle and differentiation), it is often infeasible to disentangle these processes by applying cell trajectory inference methods. Here, `GeneTrajectory` dissects out gene trajectories that correspond to different pure biological processes even when the transcriptomes of the cell population represent mixtures of coexisting biological processes.

We applied the same workflow in the β-catenin knockout sample to construct the gene trajectories for the same set of genes as used in the wildtype analysis. We observed a similar layout as compared with the wildtype gene trajectories (Figure 4B, E). After extracting the order of genes in each trajectory (Figure 4F), we compared the consistency in the gene order between the matched wildtype and knock-out gene trajectories (Figure 4H). Notably, while there is a high consistency of gene ordering in the cell cycle trajectory and the LD trajectory, we observed a significant distortion of the gene order in the DC trajectory of the knock-out, which suggests the genetic defects caused by β-catenin knockout specifically disrupt the differentiation process of DC.

4 Discussion

We developed `GeneTrajectory`, a method for detecting gene trajectories where each trajectory comprises of genes organized in pseudo-temporal order which characterize the transcriptional dynamics of a unique biological process. The standard approach for unraveling dynamical processes in single cell RNA sequencing studies is based on inference of cell trajectories. However, cells participate in concurrent biological processes and application of cell trajectory inference tools often fail to detect coherent trajectories. In the current paper we take a different approach by considering optimal-transport based gene-gene metrics, which naturally leverage the underlying continuity of the cell geometry to reveal the coherent successive relations among the genes when they exist. In contrast to cell trajectory inference tools, `GeneTrajectory` is designed to de-mix concurrent biological processes at the gene space and output for each of these processes a coherent gene trajectory. This is not only useful for the task of inferring gene dynamics of different processes which typically involves challenging experiments, but also for differential analysis of gene dynamics between experimental conditions (e.g, before and after treatment, wildtype vs. gene knock-out) which can be performed by inspecting the changes of the gene trajectory corresponding to each biological process.

In the current manuscript, we demonstrate the utility of `GeneTrajectory` to unravel gene dynamics using scRNA-seq data. Our method can be generalized to other single-cell modalities, including but not limited to scATAC-seq, spatial transcriptomics, etc. The data analysis approach can also be extended in several directions. For one thing, we currently adopt the common strategy by focusing on top highly variable genes in our analysis. However, genes with high variances may not necessarily exhibit significant local enrichment patterns, for instance, *Xist* in combined male and female samples is expected to have a high variance of expression level but it is not biological interesting to incorporate in gene trajectory analysis. As a result, our approach can potentially benefit from an improved algorithm for screening localized genes.

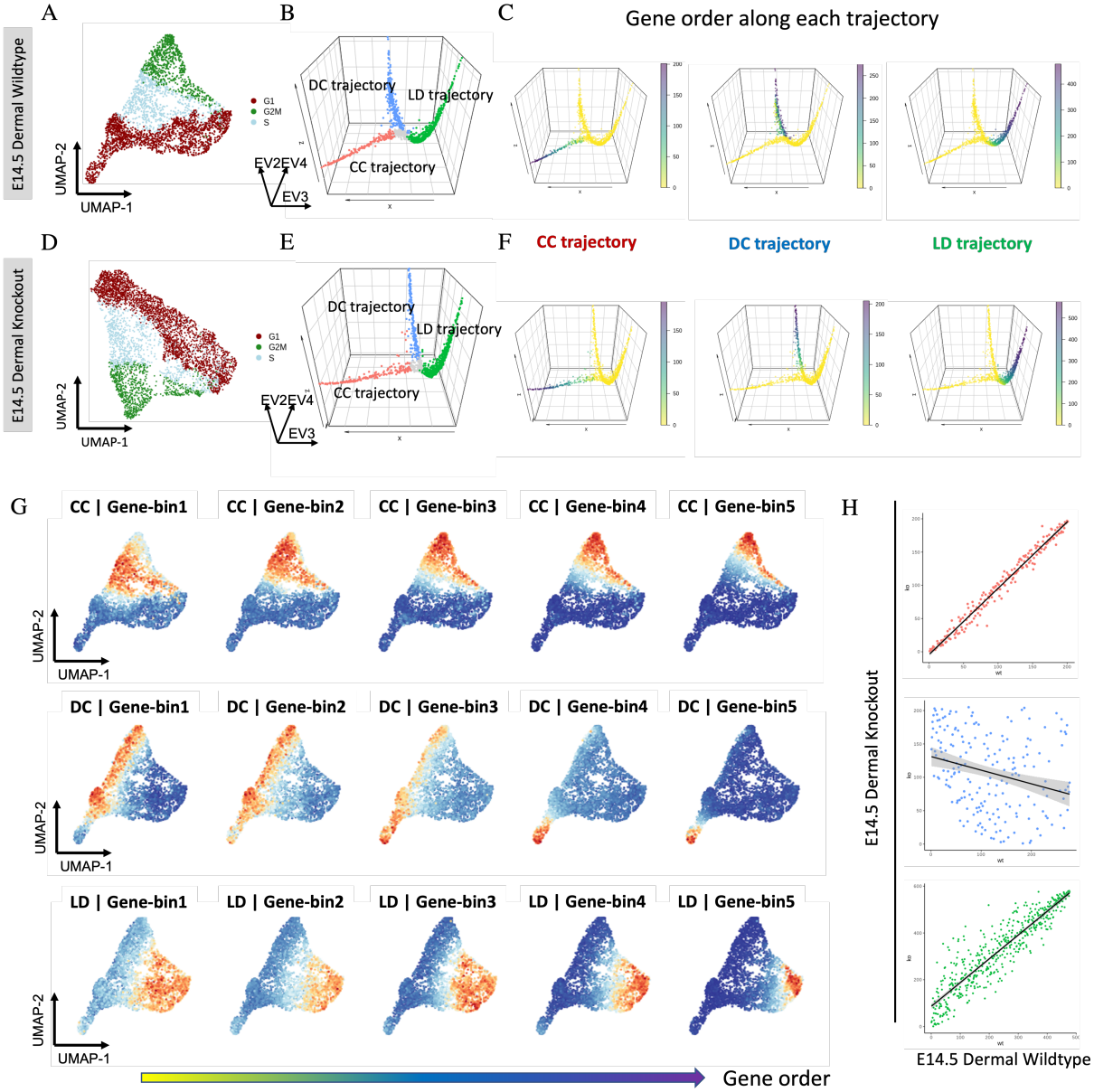

Figure 4: Comparative analysis of gene trajectories of dermal condensate genesis between the wildtype and the β-catenin knockout at day-14.5 (E14.5) of mouse embryonic development. **A, D)** UMAP embedding of dermal cell population collected at E14.5 from a wildtype mouse embryo and a paired β-catenin knockout mouse embryo, colored by cell cycle phase annotation. **B, E)** Gene trajectories visualized by the top three nontrival Diffusion Map eigenvectors for the wildtype and the β-catenin knockout respectively. Three trajectories are identified in both scenarios, with each trajectory corresponding to cell cycle, lower dermis differentiation, or DC differentiation. **C, F)** Gene order along each trajectory is indicated by the color bar (as noted in Figure 3). **G)** Cell UMAP embedding (wildtype) is colored by the average gene expression of genes within the respective bin, where bins are ordered successively along the gene trajectory (as noted in Figure 3). **H)** Comparing the ordering of genes between the wildtype and the β-catenin knockout on the gene trajectories that correspond to the same biological process.

References

[1] Mukund Balasubramanian and Eric L Schwartz. The isomap algorithm and topological stability. *Science*, 295(5552):7–7, 2002.

[2] Riccardo Bellazzi, Andrea Codegoni, Stefano Gualandi, Giovanna Nicora, and Eleonora Vercesi. The gene mover's distance: Single-cell similarity via optimal transport. *arXiv preprint arXiv:2102.01218*, 2021.

[3] Mira Bernstein, Vin De Silva, John C Langford, and Joshua B Tenenbaum. Graph approximations to geodesics on embedded manifolds. Technical report, Department of Psychology, Stanford University, 2000.

[4] Kieran R Campbell and Christopher Yau. A descriptive marker gene approach to single-cell pseudotime inference. *Bioinformatics*, 35(1):28–35, 2019.

[5] Zixuan Cang and Qing Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature communications*, 11(1):1–13, 2020.

[6] Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J Steemers, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019.

[7] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[8] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[9] Louise Deconinck, Robrecht Cannoodt, Wouter Saelens, Bart Deplancke, and Yvan Saeys. Recent advances in trajectory inference from single-cell omics data. *Current Opinion in Systems Biology*, 27:100344, 2021.

[10] Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. Scot: Single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology*, 29(1):3–18, 2022.

[11] Khusali Gupta, Jonathan Levinsohn, George Linderman, Demeng Chen, Thomas Yang Sun, Danni Dong, M Mark Taketo, Marcus Bosenberg, Yuval Kluger, Keith Choate, et al. Single-cell analysis reveals a hair follicle dermal niche molecular differentiation trajectory that begins prior to morphogenesis. *Developmental cell*, 48(1):17–31, 2019.

[12] Hannah Hochgerner, Amit Zeisel, Peter Lönnerberg, and Sten Linnarsson. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell rna sequencing. *Nature neuroscience*, 21(2):290–299, 2018.

[13] Geert-Jan Huizing, Gabriel Peyré, and Laura Cantini. Optimal transport improves cell–cell similarity inference in single-cell omics data. *Bioinformatics*, 38(8):2169–2177, 2022.

[14] Michael Lattke, Robert Goldstone, James K Ellis, Stefan Boeing, Jerónimo Jurado-Arjona, Nicolás Marichal, James I MacRae, Benedikt Berninger, and Francois Guillemot. Extensive transcriptional and chromatin changes underlie astrocyte maturation in vivo and in culture. *Nature communications*, 12(1):1–18, 2021.

[15] Diana Mahdessian, Anthony J Cesnik, Christian Gnann, Frida Danielsson, Lovisa Stenström, Muhammad Arif, Cheng Zhang, Trang Le, Fredric Johansson, Rutger Shutten, et al. Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. *Nature*, 590(7847):649–654, 2021.

[16] Noa Moriel, Enes Senel, Nir Friedman, Nikolaus Rajewsky, Nikos Karaiskos, and Mor Nitzan. Novosparc: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nature Protocols*, 16(9):4177–4200, 2021.

[17] Jonas Mueller, Tommi Jaakkola, and David Gifford. Modeling persistent trends in distributions. *Journal of the American Statistical Association*, 113(523):1296–1310, 2018.

[18] Darya Y Orlova, Noah Zimmerman, Stephen Meehan, Connor Meehan, Jeffrey Waters, Eliver EB Ghosn, Alexander Filatenkov, Gleb A Kolyagin, Yael Gernez, Shanel Tsuda, et al. Earth mover's distance (emd): a true metric for comparing biomarker expression levels in cell populations. *PloS one*, 11(3):e0151859, 2016.

[19] Rihao Qu, Khusali Gupta, Danni Dong, Yiqun Jiang, Boris Landa, Charles Saez, Gwendolyn Strickland, Jonathan Levinsohn, Pei-lun Weng, M Mark Taketo, et al. Decomposing a deterministic path to mesenchymal niche formation by two intersecting morphogen gradients. *Developmental Cell*, 57(8):1053–1067, 2022.

[20] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[21] Amit Singer. From graph to manifold laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.

[22] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (ToG)*, 34(4):1–11, 2015.

[23] Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19(1):1–16, 2018.

[24] Alexander Y Tong, Guillaume Huguet, Amine Natik, Kincaid MacDonald, Manik Kuchroo, Ronald Coifman, Guy Wolf, and Smita Krishnaswamy. Diffusion earth mover's distance and distribution embeddings. In *International Conference on Machine Learning*, pages 10336–10346. PMLR, 2021.

[25] Koen Van den Berge, Hector Roux de Bézieux, Kelly Street, Wouter Saelens, Robrecht Cannoodt, Yvan Saeys, Sandrine Dudoit, and Lieven Clement. Trajectory-based differential expression analysis for single-cell sequencing data. *Nature communications*, 11(1):1–13, 2020.

[26] Alexis Vandenbon and Diego Diez. A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data. *Nature communications*, 11(1):1–10, 2020.

[27] F Alexander Wolf, Fiona K Hamey, Mireya Plass, Jordi Solana, Joakim S Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J Theis. Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology*, 20(1):1–9, 2019.

[28] Stephen Zhang, Anton Afanassiev, Laura Greenstreet, Tetsuya Matsumoto, and Geoffrey Schiebinger. Optimal transport analysis reveals trajectories in steady-state systems. *PLoS computational biology*, 17(12):e1009466, 2021.

Table A.1: List of notations

u, v	Index of cells
i, j	Index of genes
m	Original number of cells
n	Original number of genes
m'	Reduced number of cells after coarse-graining
$\delta^{(p)}(\rho, \rho')$	Wasserstein-p distance between distribtuions ρ and ρ'
$(d_G)(u, v)$	Graph distance between cell u and v
C	Transport cost matrix on the cell graph. $C_{u,v}$ represents the cost of transport between cell u and v
C'	Transport cost matrix on the coarse-grained cell graph
A	kNN membership matrix for the cell graph. $A_{u,a} = 1/\|a\|$ if and only if the cell u belongs to the ath subset where $\|a\|$ represents the number of cells in the subset a, otherwise $A[u, a] = 0$
E	Diffusion embedding of cells
E'	Diffusion embedding of cells after coarse-graining

[29] Liangtao Zheng, Shishang Qin, Wen Si, Anqiang Wang, Baocai Xing, Ranran Gao, Xianwen Ren, Li Wang, Xiaojiang Wu, Ji Zhang, et al. Pan-cancer single-cell landscape of tumor-infiltrating t cells. *Science*, 374(6574):abe6474, 2021.

A Supplementary Information

A.1 W_p-distance formulation and computation

We denote the W_p-distance by $\delta^{(p)}$ and set up some notations: For a graph consisting of m nodes $V = [m]$, a *graph distribution* is a non-negative vector $\rho \in \mathbb{R}_+^m$ which is summed to be one and the distribution assigns measure $\rho(u)$ to node u. The graph is equipped with a graph ground distance $d_G(u, v)$ for $u, v \in V$. The graph distance d_G is used to specify the cost of the optimal transport, that is, the cost matrix C is defined as $C_{u,v}^{(p)} = d_G(u, v)^p$. In this work, we use shortest path distance on a kNN graph as d_G, while the computational method also allows other options of d_G or even letting the cost matrix take a more general form. For two graph distributions ρ and ρ', the W_p-distance is defined as

$$\delta^{(p)}(\rho, \rho') = \min_{F \in \Pi_{\rho,\rho'}} \langle F, C^{(p)} \rangle^{1/p}, \tag{A.1}$$

where $\Pi_{\rho,\rho'} = \{F,\ F_{u,v} \geq 0, \sum_v F_{u,v} = \rho(u) \text{ for all } u,\ \sum_u F_{u,v} = \rho'(v) \text{ for all } v\}$ denotes the set of transport plan F which pushes from the source distribution ρ to the target distribution ρ'.

To compute gene-gene OT distances, let $\rho_i \in \mathbb{R}_+^m$ be the distribution of the i-th gene g_i on the cell graph G (that is, $\rho_i(u) = g_i(u)/\sum_v g_i(v)$). We compute the W_p-distance $\delta_{ij}^{(p)} = \delta^{(p)}(\rho_i, \rho_j)$ between the i-th and j-th gene and use that as the dissimilarity measure between genes. In practice, the minimization in (A.1) can be solved by linear programming, which is computationally prohibitive on large cell graph and between all the pairs of genes. We accelerate the computation by two additional techniques introduced in Section 2.2, with details in SI A.2 and A.3.

A.2 Coarse graining

We coarse-grain the cell graph by aggregating m cells into m' "pseudo-cells" using K-means clustering algorithm. Let A be the m-by-m' membership matrix where $A[u, a] = 1/|a|$ if and only if the cell u belongs

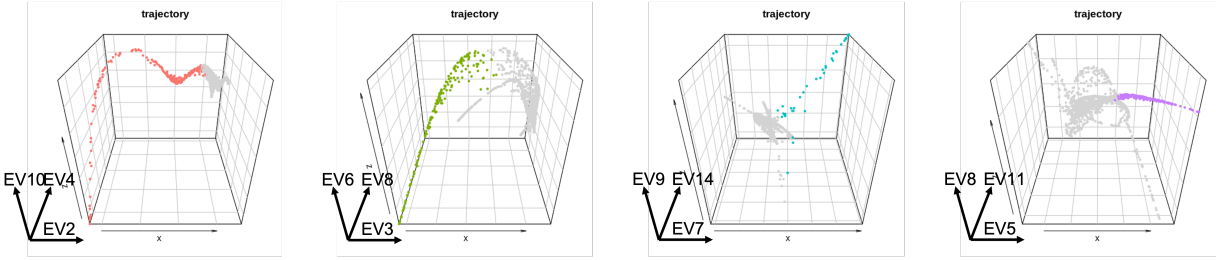

Figure A.1: Eigenvectors highlighting each gene trajectory in hippocampus dentate neurogenesis.

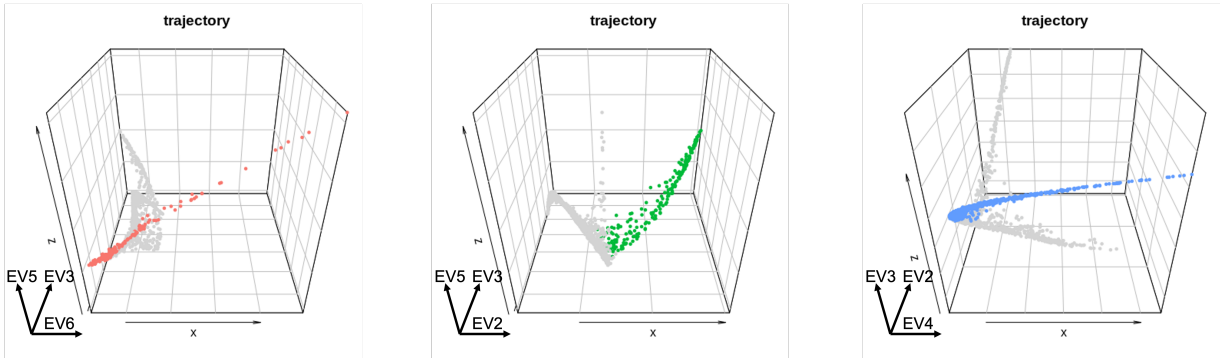

Figure A.2: Eigenvectors highlighting each gene trajectory in dermal condensate genesis.

to the ath subset where $|a|$ represents the number of cells in the subset a, otherwise $A[u, a] = 0$, then the updated transport cost matrix C' on the coarse-grained cell graph is defined by $A^T CA$. The aggregated cell embedding matrix E' is defined by $A^T E$.

A.3 Graph sparsification

This is formulated in the following way: if we want to construct the gene Wasserstein distance matrix on a cell graph of an original size m, we firstly coarse-grain the m cells into m' "meta-cells" using the procedure described above where m' is the size that can be quickly handled. Based on the gene-by-gene Wasserstein distance matrix constructed on m' "meta-cells", we identify the αK nearest neighbors for each gene (where α is the predefined parameter and K is the neighborhood size to construct the local adaptive kernel used for computing the Diffusion Map). Returning to the computation on the original cell graph, we then only compute the Wasserstein distance between a pair of genes if at least one of them is included in the other's αK nearest neighbors. Practically, this can reduce the running time by a factor about $\alpha K/m'$ compared with computing Wasserstein distances for all pairs of genes.

A.4 Selection of eigenvectors for trajectory visualization

We denote by $M(g_i)$ the diffusion embedding of genes (truncated to the top d nontrivial eigenvectors) with each row i corresponding to gene i and each column l representing the lth nontrivial eigenvector. For each $l \in [1, d]$, we compute the score $\sum_{i \in S} M_l(g_i)^2$, where S is the set of genes that belong to a given trajectory. We select top three eigenvectors with the largest scores to visualize gene trajectories in which the given trajectory can be best visually displayed.

A.5 Details of simulated gene distributions

In generating the gene distributions in Section 3.1, the cell-graph random walk matrix P is constructed as $P = D(W)^{-1}W$, where $D(W)$ is the degree matrix of W ($D(W)$ is a diagonal matrix where $D(W)_{ii} = \sum_j W_{ij}$), and W_{ij} is set to be the Gaussian-kernel affinity between cell locations. The kernel affinity is computed with a locally adaptive bandwidth: the bandwidth of each cell set equal to the graph distance to its k-nearest neighbor, while k=30 in Examples 1 and 2, and k=100 in Example 3.

The way of simulating gene distributions on the cell graph are as follows:

In Examples 1 and 2, we randomly sampled 100 cells as the origins of graph diffusion, and then propagate on the cell graph by multiplying matrix P to the distribution vector of a delta mass at the origin cell. This produces 100 localized distributions which we use as gene distributions.

In Example 3, we generate 100 genes of class (i) and (ii) respectively. For class (i), we use the x and y coordinates of the cell locations only to generate localized distributions same as before. This produces gene distributions that are localized on the x-y section of the cylinder. Class (i) genes represent the genes associated with cell cycle. For class (ii), we similarly use the z coordinate only to generate localized distributions, which produces gene distributions that are localized along the z axis. Class (ii) genes represent genes involved in the differentiation process.