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

Rihao Qu[1,2,3*], Xiuyuan Cheng[4*], Esen Sefik[3], Sarah Platt[2,5], James Garritano[2,7], Ian Odell[3,5], Richard A. Flavell[3,6†], Peggy Myung[2,5†], Yuval Kluger[2,7†]

[1]Computational Biology & Bioinformatics Program, Yale University, New Haven, CT, USA

[2]Department of Pathology, Yale University School of Medicine, New Haven, CT, USA

[3]Department of Immunobiology, Yale University School of Medicine, New Haven, CT, USA

[4]Department of Mathematics, Duke University, Durham, NC, USA

[5]Department of Dermatology, Yale University School of Medicine, New Haven, CT, USA

[6]Howard Hughes Medical Institute, Yale University School of Medicine, New Haven, CT, USA

[7]Program of Applied Mathematics, Yale University, New Haven, CT, USA

*These authors contributed equally.

†These authors jointly supervised this work.

To whom correspondence should be addressed: yuval.kluger@yale.edu

Abstract

Single-cell RNA-sequencing has enabled researchers to study dynamic biological processes by inferring cell trajectories. Current cell trajectory methods attempt to utilize whole transcriptomic profiles to draw connections among cells. However, whole-transcriptome data incorporates multiple gene programs, some of which are nonspecific or irrelevant to the process of interest. This results in fuzzy cell trajectories that fail to reveal the true progression of the process studied. Here, we present GeneTrajectory, a new approach that deconvolves complex biological processes by identifying trajectories of genes rather than trajectories of cells. Specifically, the distances between gene expression distributions over the cell graph are computed using an optimal-transport metric to construct gene trajectories. GeneTrajectory unbiasedly extracts gene programs and their gene pseudotemporal order from whole-transcriptome data without relying on cell trajectory inferences. Using published and novel datasets, we demonstrate that GeneTrajectory identifies the gene dynamics of biological processes that cannot be effectively resolved by cell trajectory approaches.

Introduction

Single-cell transcriptomic technology has provided a valuable tool to study fundamental dynamic biological processes such as cell cycle[1], tissue/organ differentiation[2], or tumor-reactive T-cell

evolution[3]. In particular, cell trajectory inference approaches[4-9] have been widely used to identify key genes and their pseudotemporal order in an effort to delineate molecular events that govern biological processes. Briefly, these approaches use single-cell whole-transcriptome data to organize cells into lineages (as cell trajectories) and assign pseudotime to the cells within each lineage[10]. Genes associated with each cell trajectory are then ordered based on their changing expression patterns along the cell pseudotime. A major caveat to these methods is that many biological processes are complex and orchestrated by multiple gene programs, some of which co-occur in an intertwined manner (e.g., cell differentiation coupled with cell cycle[11]). These complex dynamic processes cannot be faithfully represented by a cell trajectory that considers all genes within each cell. Namely, two cells that are similar by one biological process but different by another process cannot be reliably resolved with current cell trajectory inference methods, which aggregate all genes within each cell to assess similarity. Besides, the accuracy of cell trajectory construction depends on whether the cell manifold in high-dimensional space can be effectively reduced to one-dimensional curves. When a cell manifold has an intrinsic dimension[12] ≥ 2, it will be infeasible to determine the unidimensional lineage of cells and thus inferring the pseudotemporal order in which genes are activated along each lineage is inherently unreliable.

To overcome these obstacles, we have developed a fundamentally new approach, GeneTrajectory, that detects trajectories of genes rather than trajectories of cells. Notably, GeneTrajectory automatically dissects out gene programs from whole transcriptome single-cell data without relying on an initial cell trajectory construction. Hence, it can be generally applied to a cell cloud with non-curvilinear geometric structure. Using this method, genes that sequentially contribute to a given biological process can be extracted and then organized into a gene trajectory. The ordering of genes along each gene trajectory implies how they temporally “come into play” during each underlying biological process. By overlaying the expression level of those genes on the cell embedding following their order, GeneTrajectory provides a “movie-like” perspective to visualize how different biological processes are coordinating and governing different cell populations. Further, by deconvolving co-occurring processes, each process (e.g., lineage differentiation) can be purely represented, excluding irrelevant biological effects from the other processes. For example, critical steps in cell commitment from multi/oligopotent progenitor states are frequently difficult to resolve using standard cell trajectory inference approaches since there are usually a mixture of multiple gene programs involved[13]. GeneTrajectory aims to “demix” them and track the molecular change of each specific biological process in an independent manner.

To construct gene trajectories that delineate the chronological dynamics of gene programs, GeneTrajectory quantifies the dissimilarity (i.e., the distance) between genes by leveraging optimal transport (OT) metrics. Specifically, in our framework, we view the expression profile of a gene as a distribution over cells and compute pairwise Wasserstein distances (as one classical type of OT metrics) between gene distributions over the cell graph to define gene-gene distances. Here, Wasserstein distance measures the minimal "effort" to transport one gene distribution over the cell graph until it is perfectly aligned with another gene distribution. In this construction, a pair of genes activated consecutively along a biological process are expected to have a small Wasserstein distance between their distributions, since they are likely to be expressed in the same region(s) of the cell graph and thus it takes negligible effort to shift one to the other. Finally, by employing a diffusion-based technique to the gene-gene Wasserstein distance matrix, we construct gene trajectories with each revealing the pseudotemporal gene dynamics for a specific biological process.

Here, we use a published human peripheral blood mononuclear cell (PBMC) scRNA-seq dataset (generated by the 10x Genomics Chromium system[14]) to show that GeneTrajectory identifies gene programs governing the maturation of two myeloid lineages. Moreover, we employ our approach on a novel mouse embryonic skin dataset to demonstrate that GeneTrajectory can resolve critical cell state transitions during the early-stage development of hair follicles[2, 15]. In this example, GeneTrajectory uniquely unravels a biological process in which cell cycle is coupled with cell fate commitment, which could not be resolved using standard cell trajectory or genetic lineage tracing methods. Lastly, we provide an exploratory example of identifying potential gene programs and pathways that regulate the fibrosis process[16] which is yet to be clearly studied.

Results

GeneTrajectory quantifies gene-gene dissimilarity using a cell-graph-based OT distance

A progressive dynamic biological process is usually governed by a finely-regulated gene cascade[17-19], in which genes are activated and deactivated in a temporal order along the process, dictating the transcriptomic changes of underlying cell states. Here, we present GeneTrajectory, which infers the "pseudotemporal distance" between genes over a biological process based on scRNA-seq data (Fig. 1a). The central goal of our approach is to unbiasedly delineate the underlying dynamics by constructing gene trajectories. Specifically, each gene trajectory captures a sequence of genes expressed along a biological process in a progressive fashion. In this

construction, a crucial question is what kind of metrics can rigorously quantify the distance between genes? For example, one of the simplest ways is to transpose the single-cell data matrix and consider genes as observations and cells as features, and then calculate a standard distance (e.g., Euclidean distance) between genes. However, viewing cells as features will lead to a poor gene geometry due to the imbalance in the abundance of different cell populations. Here, we quantify the distance of genes based on their expression distributions over a cell graph using optimal transport (OT) metrics (Fig. 1c).

Previously, OT metrics (e.g., Wasserstein distance) have been applied in a wide range of scenarios in single-cell analysis, including 1) defining a distance measure between cells[20, 21] or cell populations[22], 2) constructing cell trajectories[23, 24], 3) spatial reconstruction of single-cell transcriptome profiles[25, 26], and 4) multi-omics data integration[27]. In these works, the dissimilarity was quantified either between a pair of cells, or between a pair of cell populations. In our work, we distinctively define the graph-based Wasserstein distance between pairs of genes to study their underlying pseudotemporal dynamics. Specifically, we normalize the expression of a gene into a probabilistic distribution over cells and then compute the Wasserstein distances between gene distributions in the cell graph (Fig. 1c). Here, the cell graph is constructed in a way that provides a representation of cells which preserves the cell manifold structure in the high-dimensional space (Fig. 1b). In this construction, the graph-based Wasserstein distance between pairwise gene distributions has the following characteristics: 1) It takes into account the geometry of cells, i.e., it assigns a higher cost to transport a point mass from one cell to a distant cell as compared to its adjacent neighbors. 2) It is defined by the minimal cost of transporting one gene distribution to the other. 3) It prevents the transport across the ambient cell space which is often a problematic issue when using spatial distance measures (e.g., the Euclidean distance in the cell space). Namely, when cell geometry presents a low-dimensional manifold structure, the transport should always proceed along the cell manifold instead of taking a shortcut through empty regions where there are no cells. This is consistent with the biological intuition that a smoothly transitioning process is expected to associate with a coherent cell manifold instead of disconnected cell partitions[28].

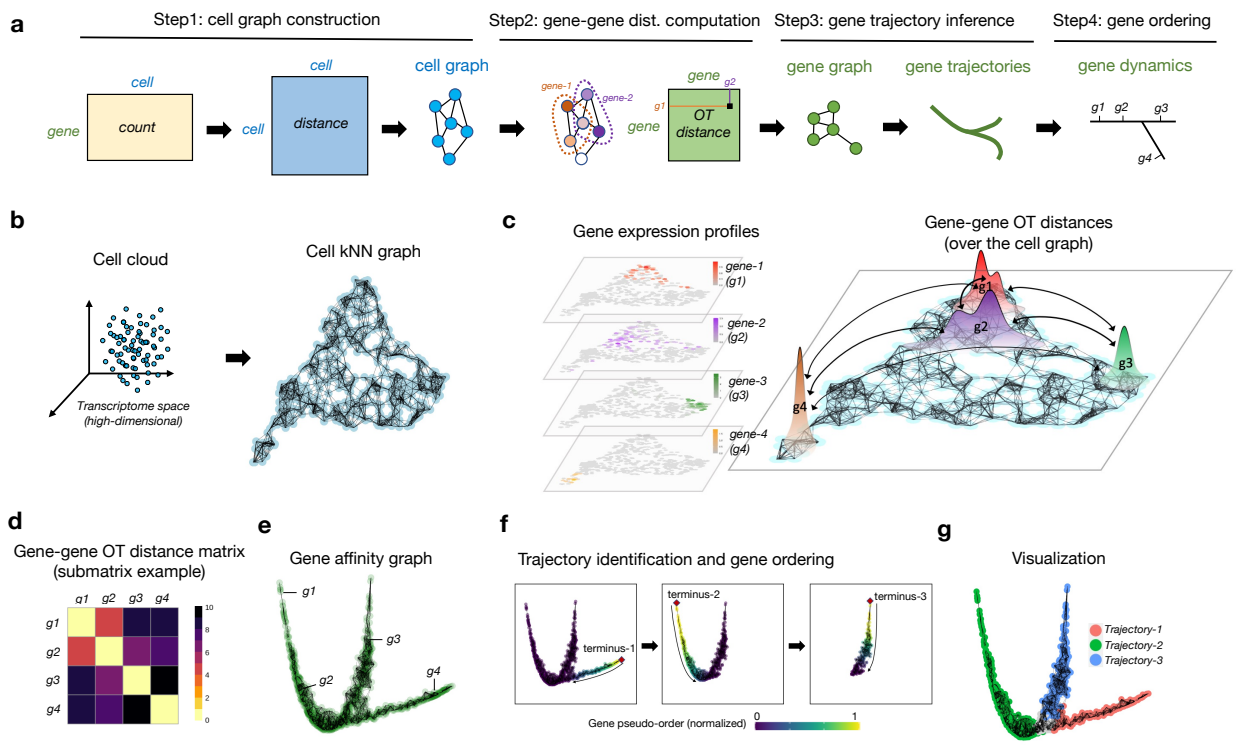

Fig.1 | Overview of GeneTrajectory. a, Schematic of the major workflow of GeneTrajectory. **b**, Cell kNN graph construction. **c**, Measuring graph-based OT (Wasserstein) distances between paired gene distributions (four representative genes are shown) over the cell graph. Gene distributions are defined by their normalized expression levels over cells. **d**, OT (Wasserstein) distance heatmap for genes *g1-g4* in **c**. **e**, Gene graph construction based on gene-gene affinities (transformed from gene-gene Wasserstein distances). **f**, Sequential trajectory identification using a diffusion-based strategy. The initial node (terminus-1) is defined by the gene with the largest distance from the origin in the Diffusion Map embedding. We then employ a random-walk procedure to select the other genes that belong to the trajectory terminated at terminus-1. To identify the second trajectory, we remove the genes assigned to the first trajectory and repeat the steps above. This is done iteratively until a predefined number of trajectories are extracted. **g**, Diffusion Map visualization of gene trajectories.

In our approach, the computation of gene-gene Wasserstein distances is based on the following two steps:

- *Construct a cell-cell graph.* As an initial step, we learn a reduced-dimensional cell embedding that can capture and represent the cell manifold structure in the original high-dimensional space. Here, we advise users to apply Diffusion Map[29] for the dimension reduction of cells. Specifically, we calculate the pairwise Euclidean distances among cells in the transcriptome space and use a Gaussian kernel to transform distances into affinities. Diffusion Map takes cell-cell affinities as the input to generate a list of eigenvectors which encode the manifold

information of the cell cloud. Next, we construct a k-nearest-neighbor (kNN) graph of cells based on their relative distances in the cell embedding (Fig. 1b). In this graph, two cells are connected only if one of them resides in the k-sized neighborhood of the other. Essentially, this establishes a cell-cell connectivity map that serves as the "roadmap" for transporting gene distributions in the next step. Here, 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. This graph distance will be used to define the cost of transporting a point mass between cells u and v in the next step.

- *Compute gene-gene Wasserstein distances over the cell-cell graph.* We model the expression level of genes as discrete distributions on the cell graph. Specifically, we divide the original expression level of a given gene in each cell by the sum of its expression level in all cells. We then define the distance between two gene distributions by the graph-based Wasserstein-p distance (W_p-distance, $1 \leq p < \infty$) (Fig. 1c,d). Accordingly, the transport cost between cell u and v is defined as $C_{uv} = (d_G)_{uv}^p$. Here, p is a user-defined parameter, and $p = 1$ refers to the well-known Earth-Mover distance. Algorithmic details are described in *Methods*.

In practice, computing the Wasserstein distance between all pairwise gene distributions can be computationally expensive. For example, when gene expression profiles are very noisy (e.g., a high sequencing dropout rate results in fuzzy gene distributions), we advise users to perform data denoising (e.g., using ALRA[30]) prior to Wasserstein distance computation. Another example is when the cell graph is large, the time cost for finding the optimal transport solution increases exponentially. In our framework, we have designed two strategies to accelerate the computation based on 1) cell graph coarse-graining, and 2) gene graph sparsification (See details in *Methods*). Briefly, cell graph coarse-graining aims to reduce the cell number by aggregating nearest cells into "meta-cells". Gene graph sparsification aims to skip the computation for two gene distributions if they are very far away from each other at a coarse-grained level, as they are unlikely to participate in the same biological process. We note that while coarse-graining the cell graph to a crude scale can make it extremely fast for computation, it may lose accuracy and compromise the resolution. Hence users should judiciously choose the level of coarse graining based on the capacity of their computing resources.

Gene trajectory construction and gene pseudotemporal ordering

The gene-gene Wasserstein distance captures the pseudotemporal relations of genes in the sense that if two genes are activated consecutively along a biological process, their distributions are expected to have a significant overlap in the cell graph and thus have a small Wasserstein distance between each other. Here, to visualize the global geometry of all genes, we convert pairwise gene-gene Wasserstein distances into gene-gene affinities (using a Gaussian kernel) and employ Diffusion Map to get a low-dimensional representation of genes. When the dynamics of a set of genes underlying smooth-transitioning biological processes is captured, viewing the gene embedding by a combination of leading Diffusion Map eigenvectors delineates trajectories of genes (Fig. 1e). Each trajectory is linked with a specific gene program that dictates the underlying biological process.

In our approach, the extraction of gene trajectories is performed in a sequential manner (Fig. 1f). To identify the first trajectory, we search for the gene that has the largest distance from the origin of Diffusion Map embedding, which serves as the terminus of the first gene trajectory. To retrieve the other genes along the first trajectory, we take that terminus gene as the starting point of a diffusion process. Specifically, we assign a unit point mass to that gene and then diffuse the mass to the other genes - the diffusion process is a random walk on the gene graph where the probability of transition from one gene to another is higher if they are closer under the gene-gene Wasserstein distance. As the probability mass propagates along the gene trajectory from its terminus, the trajectory can be retrieved by a thresholding procedure (see details in Methods). After retrieving genes for the first trajectory, we remove those genes from the pool of all genes and iterate same procedure on the remaining genes, until a predefined number of gene trajectories are identified (Fig. 1f-g) (See details in *Methods*).

To order the genes along a given trajectory, we retain only these genes to re-compute a Diffusion Map embedding based on their pairwise gene-gene Wasserstein distances. The outputted first non-trivial eigenvector of the new Diffusion Map embedding provides an intrinsic ordering of the genes along that trajectory. This is guaranteed by the theorem that if points approximately lie on a one-dimensional curve, the ordering of their first non-trivial eigenvector coordinates reflects their ordering along the curve, according to the spectral convergence theory of Diffusion Map[31].

To examine how the gene order along a given gene trajectory is reflected over the cell graph, we can track how these genes are expressed across different regions in the cell embedding. Specifically, we first group genes along each gene trajectory into successive bins. For each gene bin, we generate a cell embedding “snapshot” by coloring each cell based on the proportion of

genes from that bin that are expressed in the cell. Here, genes in the first and last bins are usually expressed in distinct cell populations which correspond to the initial and the terminal states in the underlying biological process. Genes in the middle bin are expected to be expressed in the transient cell populations linking between the precursor population and terminal population. Assembling all "snapshots" in a successive manner provides a movie-like progression reflecting the gene program over the cell embedding, which reveals how the underlying biological process proceeds in the landscape of cells.

Evaluating performance based on gene dynamics simulation

Assuming that a progressive biological process is temporally dictated by a sequence of genes, we simulated several artificial scRNA-seq datasets with a variety of gene dynamics by modeling the change of gene expression over time (Supplementary Fig. 2a,b). Specifically, for a gene involved in a given biological process, we simulate its expected expression level $\lambda(t)$ as a function of time t. For clarity, we note that *t* represents the pseudotime of a biological process, linked with the cell state (e.g., differentiation status) rather than the actual time (e.g., specific day of a developmental process). Here, we use multiple parameters to account for the heterogeneity of gene expression profiles in single-cell data, including the variation of duration time and expression intensities (See details in *Methods*). For each cell state at t along a biological process, we apply a Poisson sampling to generate the observed expression level of each gene by taking $\lambda(t)$ as the mean of Poisson distribution. Notably, as the ground truth for downstream analyses, we know both the pseudotime of each cell in the corresponding biological process as well as the temporal order of genes that dictate each process.

Here, we first simulated 1) a cycling biological process in which the change of gene expression shows a periodical pattern over time (Fig. 2a, Supplementary Fig. 2c); and 2) a biological process with a branching point where it diverges into two different lineages and each lineage succeeds with distinct gene dynamics (Fig. 2b, Supplementary Fig. 2d). The first process is inspired by the cell cycle, while the second process is inspired by the lineage differentiation process. As expected, inspection of the gene trajectories in these two simulation examples reveals similar layouts with their cell embeddings (Fig. 2a,b). The ordering of genes along each gene trajectory shows a high concordance with the ground truth.

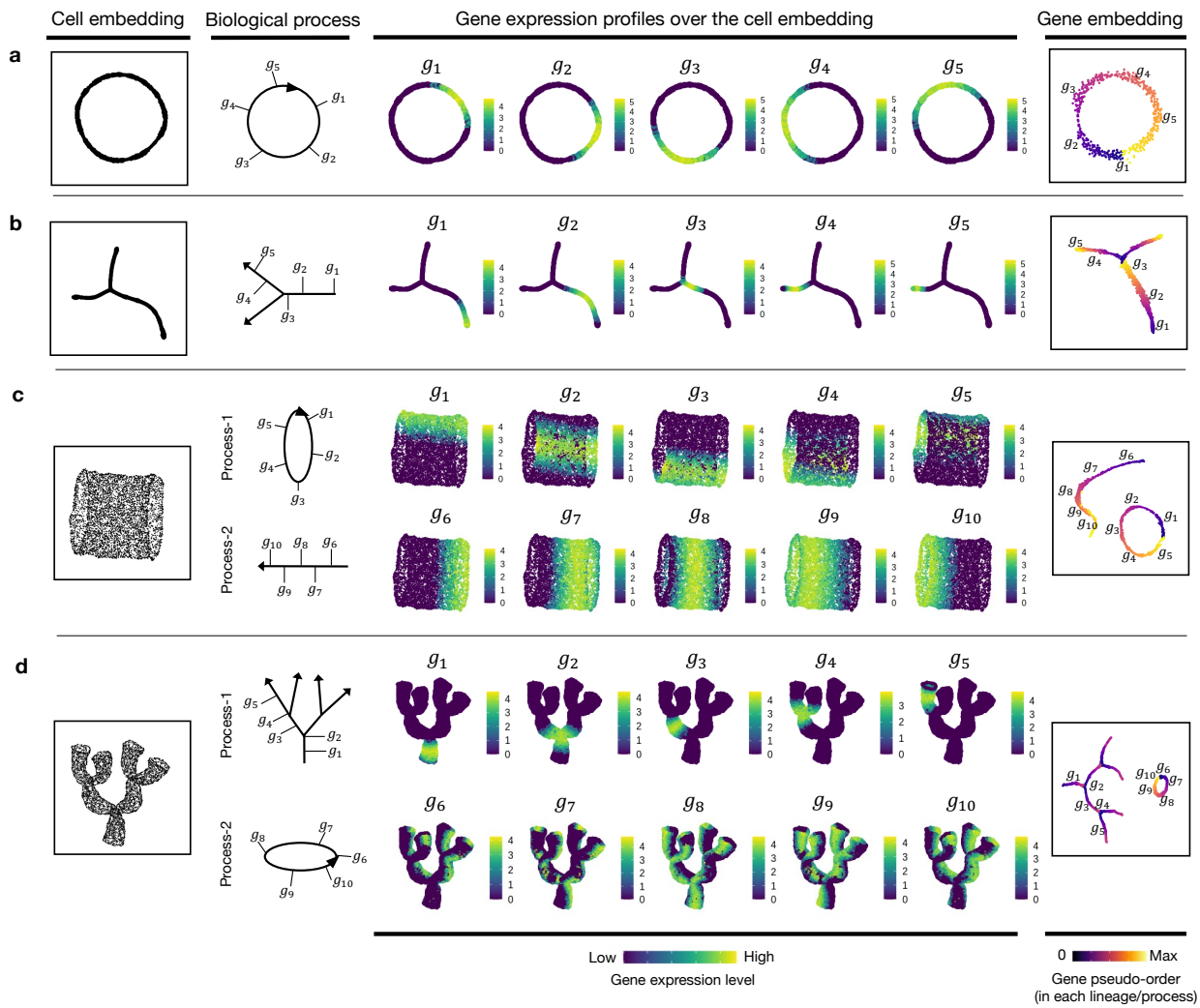

Fig.2 | GeneTrajectory performance evaluation based on simulation experiments. a, Simulation of a cycling process (cell cycle). The cell embedding and gene embedding showcase the same topology which has a ring-shaped structure. **b**, Simulation of a differentiation process with two lineages. The cell embedding and gene embedding showcase the same topology which has a bifurcating-tree structure. **c**, Simulation of a linear differentiation process coupled with cell cycle. The cell embedding and gene embedding showcase distinct topologies. Cells are organized along a cylinder-shaped manifold which has an intrinsic dimension of two. Genes that contribute to the two processes are deconvolved and organized along a ring-shaped trajectory and a linear trajectory. **d**, Simulation of a multi-level lineage differentiation process coupled with cell cycle. The cell embedding and gene embedding showcase distinct topologies. Cells are organized along a coral-shaped manifold which has an intrinsic dimension of two. Genes that contribute to the two processes are deconvolved and organized along a ring-shaped trajectory and a multilayered-tree-structured trajectory. [**a** and **b** are visualized by t-SNE; **c** and **d** are visualized by UMAP. The first column shows the embedding of cells; The second column delineates the progressive dynamics of the simulated biological process with five genes selected along each process; The 3rd-7th columns show the expression of selected genes in the cell embedding following their pseudotemporal order; The 8th column displays the embedding of genes, colored by the ground truth of gene pseudo-order.]

We next created two more complex examples, each simulating a mixture of two concurrent biological processes (Fig. 2c,d, Supplementary Fig. 2e,f). Specifically, one process mimics cell differentiation (linear or branched in a multilayered fashion), and the other mimics the cell cycle. In other words, each cell state in these two examples is determined by two hidden variables – a pseudotime along the differentiation process and a pseudotime in the cell cycle. Here, we assume that these two processes are unrelated to each other. Namely, for any cell state, its pseudotimes in these two processes are independent. For each process, we simulated an exclusive set of genes with distinct dynamic characteristics (Supplementary Fig. 2e,f) (See details in *Methods*), generating a cell manifold with a cylinder-shaped or a coral-shaped structure (Fig. 2c,d). In both examples, our approach deconvolves the original mixture of two biological processes into two gene trajectories representing a (linear or tree-like) differentiation process and a (circular) cell cycle process. Along each trajectory, genes are ordered in high concordance with the ground truth. We note that, since cell trajectory inference assumes the cell cloud can be reduced into 1D curves, it is infeasible to construct reliable cell trajectories in these two examples in which the cell manifold has an intrinsic dimension of two. Taken together, these simulation experiments demonstrate two unique advantages of GeneTrajectory over cell trajectory inference approaches: 1) it allows deconvolving a mixture of biological processes that take place simultaneously in the same group of cells, which cannot be deconvolved by cell trajectory construction; and 2) it makes no assumption on the intrinsic dimension of cell manifold, which overcomes the limitation in current cell trajectory inference approaches.

GeneTrajectory resolves gene dynamics underlying myeloid lineage differentiation

In this example, we applied GeneTrajectory to infer the pseudotemporal gene dynamics underlying the differentiation process of two major myeloid lineages. Here, human myeloid cells were extracted from a public 10x PBMC dataset in which we identified four cell types based on the expression of canonical markers (Fig. 3a, Supplementary Fig. 3b-c). Specifically, it comprises CD14+ monocytes (HLA-DR low and high), CD16+ monocytes, and CD14- dendritic cells (type-2). Visualizing the cell embedding by UMAP depicts a continuum of cell states underlying myeloid lineage genesis, which includes monocyte maturation and dendritic cell differentiation. As a well-defined process both *in vivo* and *in vitro*, human monocyte maturation involves upregulation of CD16 on a subset of CD14+ classical monocytes[32]. Specifically, CD14+ monocytes first transition into an intermediate subset of CD14+CD16+ monocytes, and then differentiate into CD16+ non-conventional monocytes which mediate non-overlapping effector functions. While human CD14+

monocytes specialize in phagocytosis and production of reactive oxygen species and inflammatory cytokines, CD16+ non-conventional monocytes are poor phagocytes and patrol the endothelium of blood vessels after adoptive transfer[33].

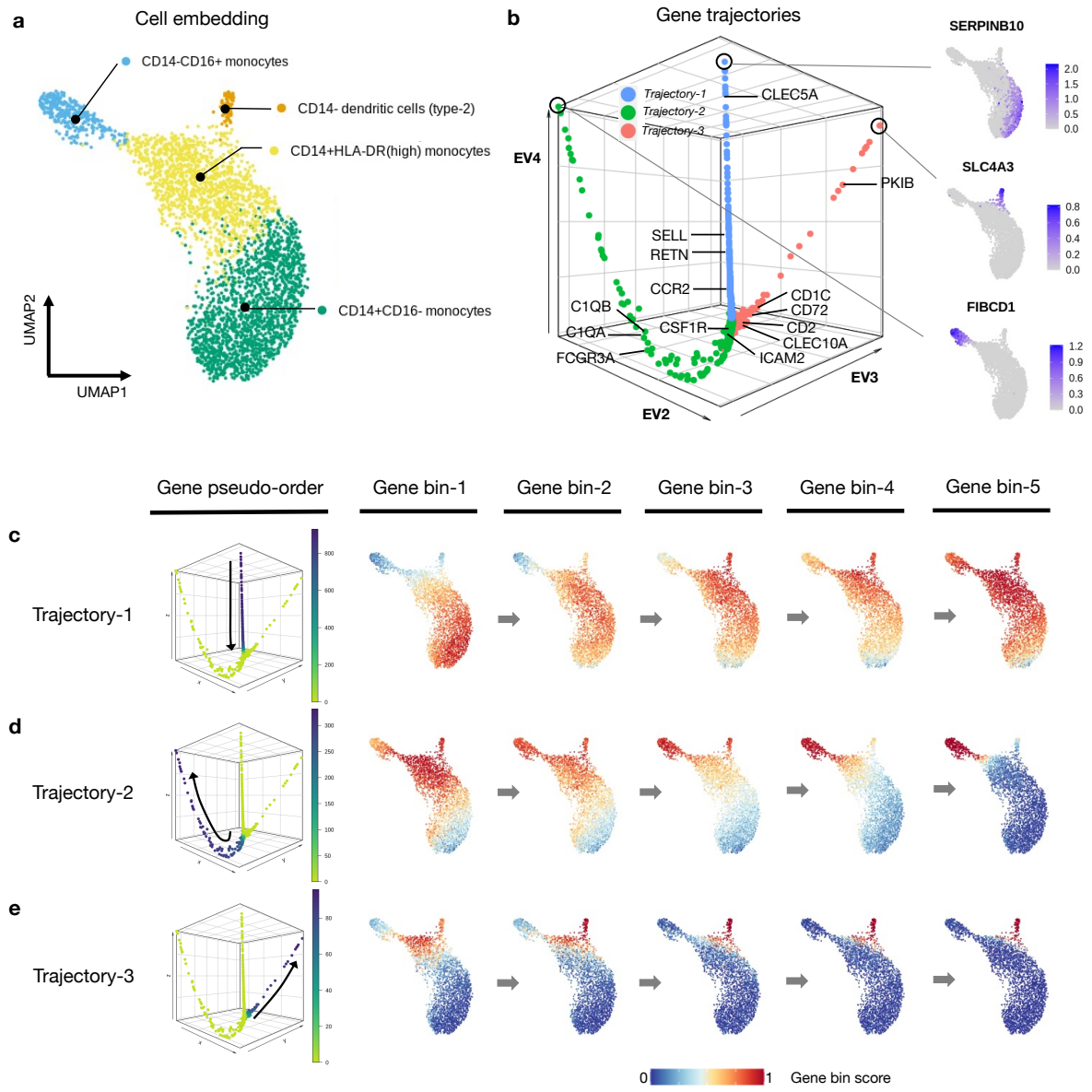

Fig.3 | Gene trajectory inference on a myeloid scRNA-seq dataset. a, UMAP of myeloid cell population colored by cell types. **b**, DM embedding of the gene graph based on gene-gene Wasserstein distances (visualized using the three leading non-trivial eigenvectors). Three prominent gene trajectories are identified. Expression profiles of the genes at the terminus of these trajectories are visualized, each indicating a distinct myeloid cell state. **c-e**, "Gene movies": highlighting the gene expression activities along each gene trajectory over the cell embedding. Genes along each trajectory are ordered and then split into five equal-sized bins. Gene bin score is defined by the proportion of genes (from each bin) expressed in each cell.

Here, by leveraging GeneTrajectory, we identified three gene trajectories with each representing a specific aspect of myeloid lineage differentiation process (Fig. 3b). Viewing the "gene movies" of Trajectory-1 illustrates that a subset of CD14+ monocytes start a differentiation cascade and gradually shift towards CD16+ monocytes, which suggests Trajectory-1 captures the gene dynamics underlying the early stage of monocyte maturation (Fig. 3c). Notably, *CLEC5A*, *RETN*, *CCR2*, and *SELL* (CD62L) are known to be associated with the initial CD14+ monocyte cellular state[32] and are highlighted as part of Trajectory-1 (Fig. 3b). Subsequently, the ordering of genes that define Trajectory-2 provides a pseudotemporal view on the later stage of CD16+ monocyte differentiation (Fig. 3d). This process is primarily driven in response to cytokine *CSF1* (MCSF) and requires *CSF1R*[34]. As expected, *CSF1R* is one of the first genes along Trajectory-2. While ordered after *CSF1R*, *ICAM2* is known to be constitutively expressed in CD16+ monocytes and is necessary for their patrolling ability across endothelium of blood vessels[35, 36]. Coming towards the end, *C1QA*, *C1QB*[37], *FCGR3A*, markers broadly expressed by fully differentiated CD16+ monocytes are identified. In addition, we retrieved a third gene trajectory (Trajectory-3) which marks the differentiation of type-2 dendritic cells as a distinct myeloid lineage (Fig. 3e). Myeloid type-2 dendritic cells have two subsets - CD14+ and CD14-. Specifically, the CD14+ subset shares overlapping features with CD14+ monocytes, whereas the CD14- subset is delineated here as corresponding with a separate gene trajectory[38]. In contrast to CD14-CD16+ monocytes, these CD14- dendritic cells differentiate in response to GMCSF and *IL4*, in line with expression of *CCR5*, *CD2*, *CLEC10A*, *CD72*, *CD1C*, *PKIB*[39] (Fig. 3b, Supplementary Fig. 3a).

In summary, the construction of gene trajectories here reveals that our approach can automatically extract gene programs associated with different processes (or phases) of myeloid lineage differentiation. The overall consistency between the relative order of well-studied marker genes (along each trajectory) and our prior knowledge indicates a great potential of using GeneTrajectory to infer how genes are temporally involved in biological processes. This can provide valuable hints for designing experiments to further uncover the underlying molecular mechanisms.

GeneTrajectory deconvolves coupled cell cycle and differentiation during dermal condensate genesis

During mouse embryonic skin development, dermal condensates (DCs) are the first morphologically distinguishable dermal subpopulations to form and are essential for hair follicle

formation. DCs emerge rapidly in a patterned array across the dermis around embryonic day 14.5 (E14.5) and are composed of densely clustered quiescent cells (Fig. 4b). Previous studies have shown that an intricate regulation of morphogen signals dictates the differentiation of DC cells[2, 15], including Wnt/β-catenin signaling, which plays a critical role in DC competency and specification. However, the role of Wnt/β-catenin signaling in DC differentiation remains unclear, as the process of DC genesis occurs with few intermediate states. Moreover, there has been a long-standing debate regarding how seemingly randomly proliferating dermal progenitors transition into quiescent differentiated DC cells. In this example, we collected skin from E14.5 wildtype and paired *K14Cre;Wntless*$^{fl/fl}$ (Wls) mutant embryos for single-cell RNA sequencing (Fig. 4a). The genetic defect in the Wls mutant results in attenuated dermal Wnt signaling and a lack of DCs and hair follicles (Fig. 4c). Here, we applied our approach to construct gene trajectories for the combined dermal cell populations in these two conditions. We associated each gene trajectory with a specific biological process and examined the differences between the wildtype and mutant dermal populations.

Visualizing cells on UMAP reveals a continuum of cell states composed of lower dermal cells (*Dkk2*+) and Wnt-activated upper dermal cells (*Dkk1*+ or *Lef1*+) which include DC cells (Fig. 4d, Supplementary Fig. 4a). As expected, DC cells, which express *Sox2*, are completely absent in the mutant sample. Notably, there are significant differences in the pre-DC cell population between the two conditions, which suggests the genetic defect interrupts DC differentiation process in an early stage prior to DC commitment. To compare the two conditions, we applied GeneTrajectory to the combined dermal cell populations and extracted three prominent gene trajectories that correspond to the differentiation of lower dermis (LD), dermal condensate (DC), as well as cell cycle (CC) (Fig. 4e). For each of these gene trajectories, we identified several established markers associated with the respective process. Specifically, Wnt signaling pathway genes (e.g., *Lef1*, *Dkk1*) and SHH-signaling pathway genes (e.g., *Ptch1*, *Gli1*), two morphogenetic signals recently shown to be necessary and sufficient for DC differentiation[2], are present in the DC gene trajectory. Interestingly, similar to the third simulation example, the upper dermal cell embedding integrates a mixture of biological processes (CC and DC differentiation) that co-occur within the same cell population. By employing GeneTrajectory, each biological process can be now separated from the other and independently examined.

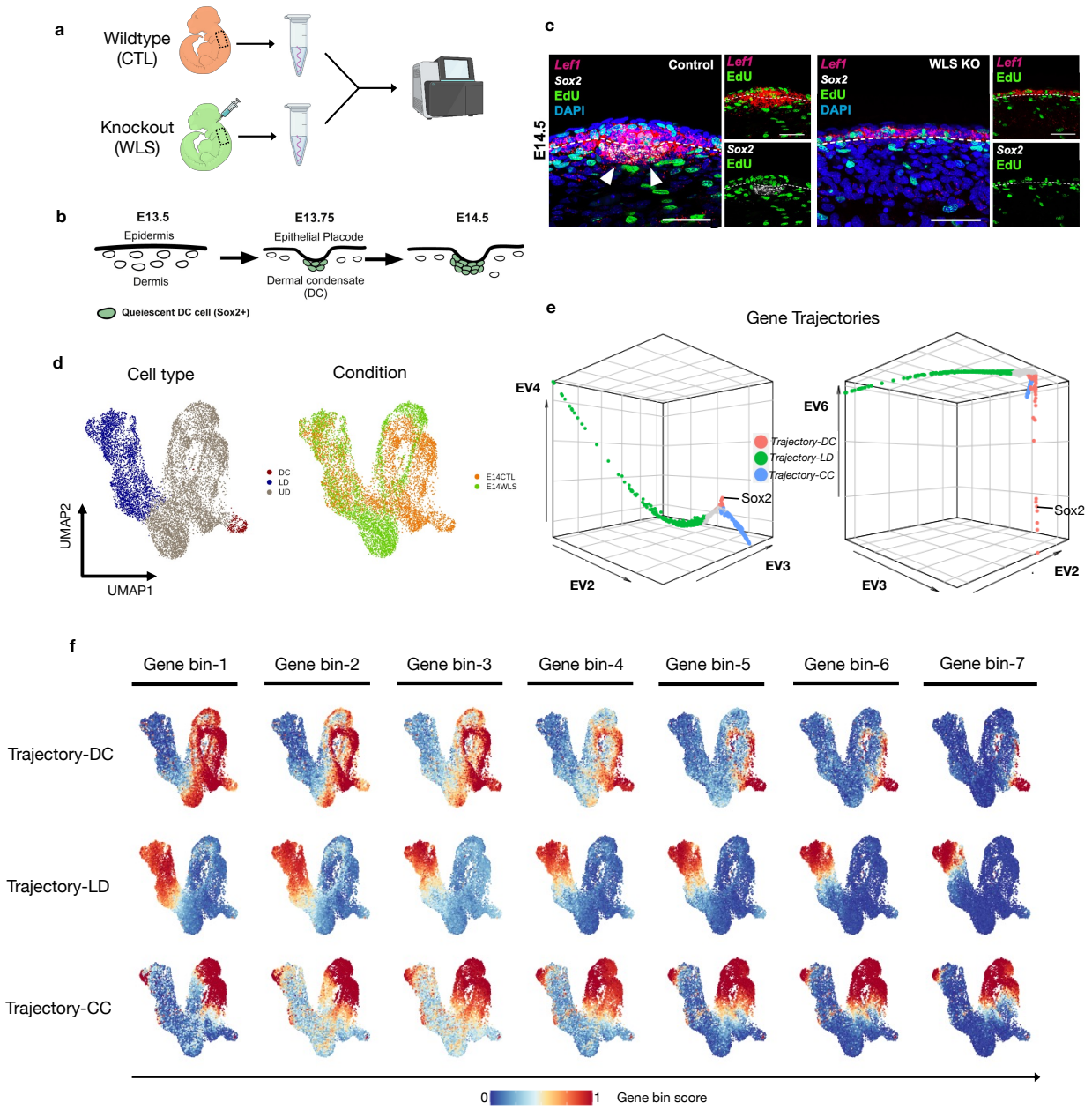

Fig.4 | GeneTrajectory deconvolves two mixed processes during dermal condensate (DC) genesis. a, Experimental design of extracting skin tissue from a pair of wildtype and Wntless (Wls) knockout embryos at day E14.5 for single-cell RNA sequencing. **b**, Cartoon illustrating the normal development process of dermal condensate (DC) from day E13.5 to E14.5. **c**, Fluorescent in situ hybridization (FISH) images (scale bar = 50μm) showing the spatial distribution of Lef1, Sox2, EdU nucleotide, and DAPI in the upper dermis of wildtype and Wls knockout. (EdU is a nucleotide that is incorporated by cells in the S-phase of cell cycle.) **d**, UMAP of cells colored by cell types and by condition. **e**, DM embedding of the gene graph to visualize three identified gene trajectories (two different combinations of leading non-trivial eigenvectors are displayed). **f**, "Gene movies" delineating the dynamics of each process (including DC differentiation, LD differentiation, cell cycle), in which genes along each trajectory are split into 7 equal-sized bins. Gene bin score is defined by the proportion of genes (from each bin) expressed in each cell.

Next, we split the genes along each gene trajectory into seven equal-sized bins and inspected the "gene movies" on the cell embedding (Fig. 4f). Remarkably, when viewing the "gene movies" for the CC and DC gene trajectories together, GeneTrajectory reveals that DC progenitors are actively proliferating throughout all stages and eventually exit the cell cycle at the terminus of DC differentiation. In other words, it implies DC is formed as an intermediate progeny of proliferative progenitors in the upper dermis. However, due to the existence of cell cycle effects, some of those direct progenitors are distant from the terminally differentiated DC cells in the original transcriptome space, which hinders cell trajectory inference from assigning a coherent pseudotime for them. Here, GeneTrajectory is able to extract a DC differentiation gene program without being conflated by cell cycle effects and resolve its pseudotemporal gene dynamics in an independent manner.

GeneTrajectory uncovers how attenuation of Wnt signaling disrupts the gene cascade

To examine how attenuated Wnt signaling affects the DC differentiation gene program, we compared the "gene movies" for the DC gene trajectory between the wildtype and Wls mutant. This visualization shows that mutant cells fall behind in the DC differentiation process, losing expression of later-stage markers (Fig. 5a-c). Notably, we did not detect a significant difference between the mutant and control with respect to the other two gene trajectories (Supplementary Fig. 5a,b), indicating the defect is specific to DC differentiation.

We next used GeneTrajectory to unbiasedly stratify the pool of progenitors distributed across different stages of DC differentiation by taking the following steps. Based on the consideration that genes in each bin act as markers indicative of a specific stage during DC differentiation, we first identified cells that express more than half of the genes in the last (seventh) bin, which represent cells in the final stage of differentiation (Stage-7). Subsequently, among the remaining cells, we identified the cells that express more than half of the genes in the sixth bin as progenitors in Stage-6. This procedure was repeated iteratively until all seven gene bins were associated with their matched cell populations (Fig. 5a-d).

By comparing the composition of progenitors in different stages between the wildtype and Wls mutant, we found that the lag in transition by mutant cells begins between Stage-4 and Stage-5, when key markers in Wnt (e.g., *Lef1*) and SHH (e.g., *Gli1*, *Ptch1*) signaling pathways start to emerge (Fig. 5a-c). Notably, the average expression level of Wnt target genes is uniformly lower in the mutant than in the wildtype condition (Supplementary Fig. 5c,e), while the proportion of

cells in G1 phase of the cell cycle is higher in the mutant across all stages (Fig. 5e). Consistent with this, the rate of EdU nucleotide incorporation is significantly lower in the Wls mutant (Fig. 4c, Supplementary Fig. 5d). These data suggest that higher levels of Wnt signaling are necessary to maintain a normal rate of cell proliferation across the DC differentiation process until DC progenitors exit the cell cycle at the Stage-7. Here, GeneTrajectory inference demonstrates the coupling between DC differentiation and cell cycle. It implies that a cell's DC differentiation status is associated with the number of times it has divided, which is a phenomenon that would be difficult to capture using standard cell trajectory inference or biological experiments.

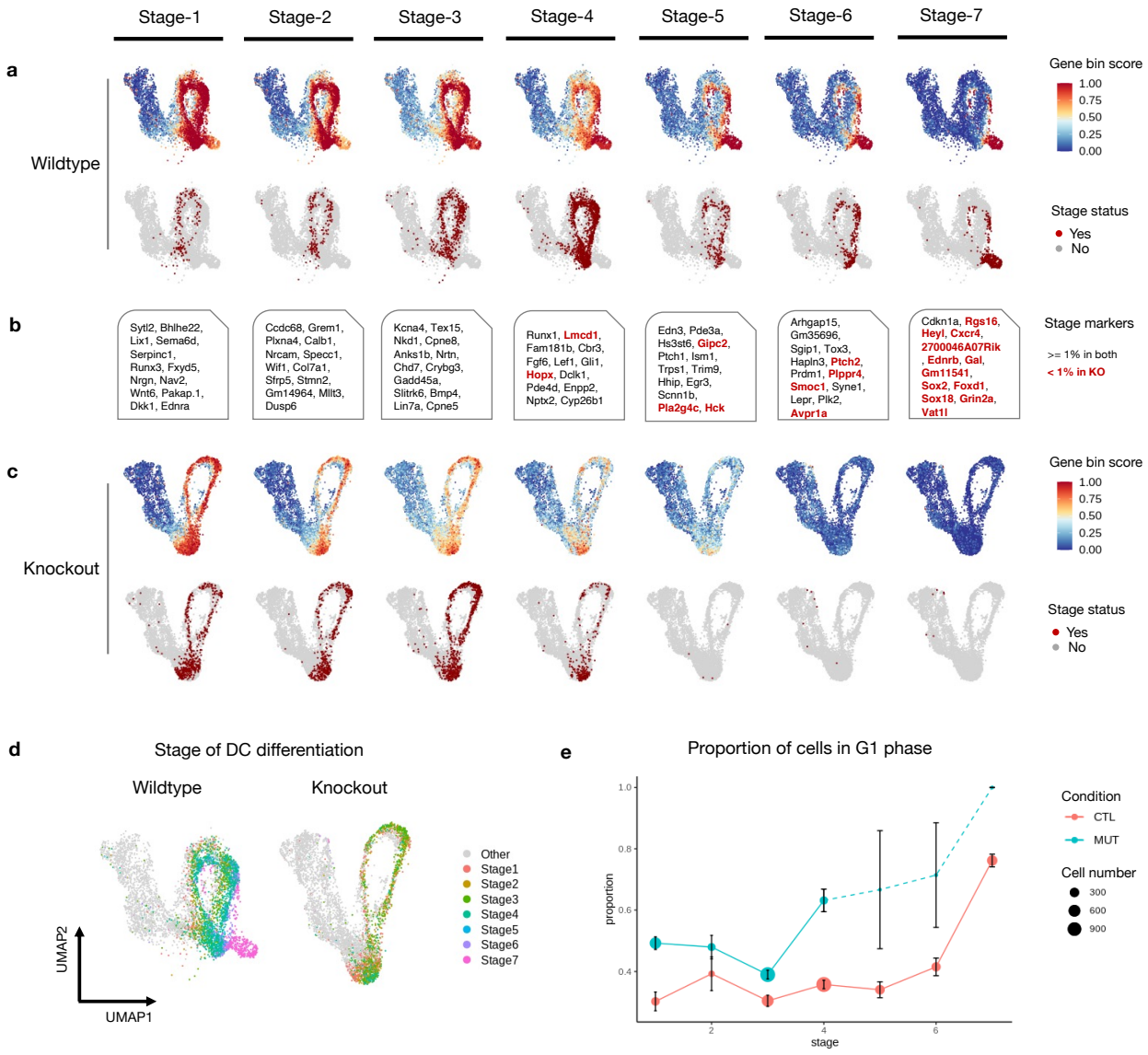

Fig.5 | Comparison of gene dynamics between the wildtype and Wls mutant. a, "Gene movies" of the DC gene trajectory over the wildtype cell embedding. Cells involved in the DC differentiation process are stratified into seven different stages. **b**, Lists of genes in different bins

along the DC gene trajectory. **c**, "Gene movies" of the DC gene trajectory over the Wls KO cell embedding. Cells involved in the DC differentiation process are stratified into seven different stages. **d**, UMAP visualizing cells at different stages, separated by condition. **e**, Change in G1 proportion across different stages in DC differentiation. G1 proportion is uniformly higher in the Wls KO than in the wildtype. Error bars: mean $\pm$ SE. Dashed lines: the number of cells in the subsequent stage ≤ 10.

GeneTrajectory infers gene pathways regulating the fibrosis process

Fibrosis is a dynamic biological process induced by chronic inflammation in which fibroblasts differentiate into myofibroblasts that produce excess extracellular matrix and cause tissue dysfunction. Genetic markers of myofibroblasts across tissues have been highlighted by recent scRNA-seq studies[40-42]. However, the dynamic process of fibroblast to myofibroblast differentiation is poorly understood. The prototypic human fibrotic disease is the autoimmune disease scleroderma [systemic sclerosis (SSc)], which causes fibrosis of the skin and internal organs. To infer the gene program underlying fibroblast differentiation in SSc, we applied GeneTrajectory to a publicly available scRNA-seq dataset generated from healthy and SSc skin[16] .

In this example, we first extracted fibroblasts from the original cell pool based on their expression of *PDGFRA*. Applying GeneTrajectory on the combined healthy and SSc fibroblast populations (Supplementary Fig. 6a) identified three gene trajectories (Supplementary Fig. 6b). Viewing the "gene movies" (Supplementary Fig. 6c) shows that each gene trajectory corresponds to a transition process occurring in a distinct fibroblast subset, marked by expression of *SFRP1*, *SFRP2*, and *APOE*[43, 44] (Supplementary Fig. 6d). Given that *SFRP2*-expressing fibroblasts were previously identified as myofibroblast precursors[16], we specifically focused on the genes along Trajectory-2 for a detailed analysis. We used DAVID[45] to perform functional enrichment analysis of the Trajectory-2 genes, which identified significant enrichment of canonical WNT signaling and positive regulation of angiogenesis. Genes involved in additional identified pathways include interferon (IFN) stimulated genes, innate inflammation related genes, and myofibroblast markers. The order of genes involved in these pathways along Trajectory-2 suggests a sequence of signaling responses that occur during myofibroblast differentiation. Specifically, these observations imply that most WNT and IFN-stimulated genes were initially upregulated, followed by a positive regulation of angiogenesis and innate inflammation, and culminating in myofibroblast markers (Supplementary Fig. 6e).

Discussion

We developed GeneTrajectory, an approach for constructing gene trajectories where each trajectory comprises genes organized in a pseudotemporal order that characterizes the transcriptional dynamics of a specific biological process. The standard pseudotime analysis approach currently used in scRNA-seq studies is based on inference of cell trajectories. However, when cells participate in concurrent biological processes, application of cell trajectory inference methods often fails to generate biologically coherent trajectories. In GeneTrajectory, we take a different perspective by considering optimal-transport-based gene-gene dissimilarity metrics, which naturally leverage the underlying geometry of cell-cell graph to reveal the coherent relations among genes that are involved in progressive biological processes. Distinctively, GeneTrajectory allows deconvolving a mixture of biological processes and output for each of them a unique gene trajectory. This is not only useful for unraveling gene dynamics of complex processes which are typically challenging to resolve by experiments, but also for comparative analysis of gene programs between experimental conditions (e.g., before and after treatment, wildtype vs. gene knock-out). Specifically, one can inspect the changes in the gene trajectory to examine how each biological process is differentially altered by treatment effects or genetic perturbations.

Moreover, it is worthy of highlight that GeneTrajectory performs gene-level pseudotime analysis in a more unbiased way than existing cell-trajectory-based methods, since it utilizes the geometric information of the whole cell-cell graph instead of ordering genes based on their covarying patterns along the prebuilt one-dimensional ordering of cells. Theoretically, a gene trajectory is established if GeneTrajectory detects a specific “path” in the cell graph along which a sequence of gene distributions shift between each other in a “relay-race” manner. Here, one common caveat (which also exists in cell trajectory inference) is GeneTrajectory cannot automatically infer the directionality of transition. Namely, genes located at the terminus of a trajectory can represent either an initial stage or a terminal stage of the corresponding process. In the end, we should combine the inference with prior knowledge to determine the directionality of gene ordering along each trajectory.

Besides, the data analysis framework in GeneTrajectory can also be extended in several directions. For example, we currently adopted a common pre-filtering strategy of focusing on top highly variable genes in our analysis. However, genes with high variances may not be involved in any biological process, e.g., Xist in combined male and female samples is expected to have a high variance of expression level but is not biological interesting to incorporate into analysis. As a result, GeneTrajectory can potentially benefit from an innovative algorithm for screening genes

localized in the cell graph, the ones that are more likely to play important roles in a specific biological process than randomly distributed genes. In the current work, we demonstrated 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. Specifically, we anticipate GeneTrajectory can be applied to resolve biological processes using dual modalities at the same time. For instance, we can quantify the pairwise distances between the distributions of gene expression and chromatin accessibility, which facilitates understanding the interplay between epigenetic dynamics and transcriptomic dynamics underlying a complex biological process.

Methods

A detailed description of GeneTrajectory algorithm as well as details for biological experiments & data analyses can be found in the supplementary LaTeX-formatted PDF file.

Data availability

The human peripheral blood mononuclear cell (PBMC) scRNA-seq dataset is available at https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k_v3. The novel mouse embryonic skin dataset generated and analyzed in this study will be deposited in the Gene Expression Omnibus (GEO) before publication. The human skin (healthy and diseased) dataset is available at GEO accession GSE138669. All other data are available from the corresponding authors upon reasonable request.

Code availability

An initial version of R implementation of GeneTrajectory and scripts used for data analyses in this work will be available soon on GitHub.

References

1. Mahdessian, D. et al. Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. *Nature* **590**, 649-654 (2021).

2. Qu, R. et al. Decomposing a deterministic path to mesenchymal niche formation by two intersecting morphogen gradients. *Developmental Cell* **57**, 1053-1067. e1055 (2022).
3. Zheng, L. et al. Pan-cancer single-cell landscape of tumor-infiltrating T cells. *Science* **374**, abe6474 (2021).
4. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics* **19**, 1-16 (2018).
5. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502 (2019).
6. Wolf, F.A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology* **20**, 1-9 (2019).
7. Van den Berge, K. et al. Trajectory-based differential expression analysis for single-cell sequencing data. *Nature communications* **11**, 1-13 (2020).
8. Deconinck, L., Cannoodt, R., Saelens, W., Deplancke, B. & Saeys, Y. Recent advances in trajectory inference from single-cell omics data. *Current Opinion in Systems Biology* **27**, 100344 (2021).
9. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nature biotechnology* **37**, 547-554 (2019).
10. Campbell, K.R. & Yau, C. A descriptive marker gene approach to single-cell pseudotime inference. *Bioinformatics* **35**, 28-35 (2019).
11. Ruijtenberg, S. & van den Heuvel, S. Coordinating cell proliferation and differentiation: Antagonism between cell cycle regulators and cell type-specific gene expression. *Cell cycle* **15**, 196-212 (2016).
12. Facco, E., d'Errico, M., Rodriguez, A. & Laio, A. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports* **7**, 1-8 (2017).
13. Kucinski, I. et al. Interactions between lineage-associated transcription factors govern haematopoietic progenitor states. *The EMBO journal* **39**, e104983 (2020).
14. Zheng, G.X. et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 1-12 (2017).
15. Gupta, K. et al. Single-cell analysis reveals a hair follicle dermal niche molecular differentiation trajectory that begins prior to morphogenesis. *Developmental cell* **48**, 17-31. e16 (2019).
16. Tabib, T. et al. Myofibroblast transcriptome indicates SFRP2hi fibroblast progenitors in systemic sclerosis skin. *Nature communications* **12**, 1-13 (2021).
17. Zhu, H., Zhao, S.D., Ray, A., Zhang, Y. & Li, X. A comprehensive temporal patterning gene network in Drosophila medulla neuroblasts revealed by single-cell RNA sequencing. *Nature communications* **13**, 1-19 (2022).
18. Sood, P. et al. Modular, cascade-like transcriptional program of regeneration in Stentor. *Elife* **11**, e80778 (2022).
19. Li, J. et al. Systematic reconstruction of molecular cascades regulating GP development using single-cell RNA-seq. *Cell reports* **15**, 1467-1480 (2016).
20. Huizing, G.-J., Peyré, G. & Cantini, L. Optimal transport improves cell–cell similarity inference in single-cell omics data. *Bioinformatics* **38**, 2169-2177 (2022).
21. Bellazzi, R., Codegoni, A., Gualandi, S., Nicora, G. & Vercesi, E. The Gene Mover's Distance: Single-cell similarity via Optimal Transport. *arXiv preprint arXiv:2102.01218* (2021).
22. Orlova, D.Y. et al. Earth mover's distance (EMD): a true metric for comparing biomarker expression levels in cell populations. *PloS one* **11**, e0151859 (2016).

23. Schiebinger, G. et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 928-943. e922 (2019).
24. Zhang, S., Afanassiev, A., Greenstreet, L., Matsumoto, T. & Schiebinger, G. Optimal transport analysis reveals trajectories in steady-state systems. *PLoS computational biology* **17**, e1009466 (2021).
25. Cang, Z. & Nie, Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature communications* **11**, 1-13 (2020).
26. Moriel, N. et al. NovoSpaRc: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nature Protocols* **16**, 4177-4200 (2021).
27. Demetci, P., Santorella, R., Sandstede, B., Noble, W.S. & Singh, R. SCOT: Single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology* **29**, 3-18 (2022).
28. Zhou, P., Wang, S., Li, T. & Nie, Q. Dissecting transition cells from single-cell transcriptome data through multiscale stochastic dynamics. *Nature communications* **12**, 1-15 (2021).
29. Coifman, R.R. & Lafon, S. Diffusion maps. *Applied and computational harmonic analysis* **21**, 5-30 (2006).
30. Linderman, G.C., Zhao, J. & Kluger, Y. Zero-preserving imputation of scRNA-seq data using low-rank approximation. *BioRxiv*, 397588 (2018).
31. Singer, A. From graph to manifold Laplacian: The convergence rate. *Applied and Computational Harmonic Analysis* **21**, 128-134 (2006).
32. Patel, A.A. et al. The fate and lifespan of human monocyte subsets in steady state and systemic inflammation. *Journal of Experimental Medicine* **214**, 1913-1923 (2017).
33. van de Veerdonk, F.L. & Netea, M.G. Diversity: a hallmark of monocyte society. *Immunity* **33**, 289-291 (2010).
34. Chitu, V. & Stanley, E.R. Colony-stimulating factor-1 in immunity and inflammation. *Current opinion in immunology* **18**, 39-48 (2006).
35. Ancuta, P. et al. Transcriptional profiling reveals developmental relationship and distinct biological functions of CD16+ and CD16-monocyte subsets. *BMC genomics* **10**, 1-19 (2009).
36. Imhof, B.A. & Dunon, D. Leukocyte migration and adhesion. *Advances in immunology* **58**, 345-416 (1995).
37. Ghebrehiwet, B., Hosszu, K.K., Valentino, A., Ji, Y. & Peerschke, E.I. Monocyte expressed macromolecular C1 and C1q receptors as molecular sensors of danger: implications in SLE. *Frontiers in Immunology* **5**, 278 (2014).
38. Heger, L. et al. Subsets of CD1c+ DCs: Dendritic Cell Versus Monocyte Lineage. *Frontiers in Immunology* **11**, 559166 (2020).
39. Higashi, N. et al. The macrophage C-type lectin specific for galactose/N-acetylgalactosamine is an endocytic receptor expressed on monocyte-derived immature dendritic cells. *Journal of Biological Chemistry* **277**, 20686-20693 (2002).
40. Gur, C. et al. LGR5 expressing skin fibroblasts define a major cellular hub perturbed in scleroderma. *Cell* **185**, 1373-1388. e1320 (2022).
41. Buechler, M.B. et al. Cross-tissue organization of the fibroblast lineage. *Nature* **593**, 575-579 (2021).
42. Tsukui, T. et al. Collagen-producing lung cell atlas identifies multiple subsets with distinct localization and relevance to fibrosis. *Nature communications* **11**, 1-16 (2020).

43. Ascensión, A.M., Fuertes-Alvarez, S., Ibanez-Sole, O., Izeta, A. & Araúzo-Bravo, M.J. Human dermal fibroblast subpopulations are conserved across single-cell RNA sequencing studies. *Journal of Investigative Dermatology* **141**, 1735-1744. e1735 (2021).
44. Plikus, M.V. et al. Fibroblasts: Origins, definitions, and functions in health and disease. *Cell* **184**, 3852-3872 (2021).
45. Huang, D.W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**, 44-57 (2009).

Acknowledgement

The authors thank Ronald Coifman, Junchen Yang, Manolis Roulis, and Jay Stanley for fruitful discussions. Y.K. and X.C. acknowledge support by NIH grant R01GM131642. Y.K. also acknowledges support by NIH grant UM1DA051410, U54AG076043, P50CA121974, and U01DA053628.