A Unified Framework for Lineage Tracing and Trajectory Inference

Aden Forrow[1] and Geoffrey Schiebinger[2]

[1]Mathematical Institute, University of Oxford,
aden.forrow@maths.ox.ac.uk
[2]Department of Mathematics, University of British Columbia,
geoff@math.ubc.ca

July 31, 2020

Abstract

Analyzing the trajectories of cellular differentiation sheds light on key questions across biology, from how cell types are stabilized during embryonic development to how they destabilize with age or disease. New single cell measurement technologies offer the prospect of reconstructing these developmental trajectories from snapshots of cell state together with lineage trees inferred from continuously-induced mutations in heritable DNA-barcodes. However, current methods for reconstructing developmental trajectories are not designed to leverage the additional information contained in lineage trees. Inspired by these recent experimental advances, we present a novel framework for reconstructing developmental trajectories from snapshots of cell state combined with lineage trees. Our method learns from both kinds of information together using mathematical tools from graphical models and optimal transport. We find that lineage data helps disentangle complex state transitions with fewer measured time points, enabling increased accuracy for lower experimental cost. Moreover, integrating lineage tracing with trajectory inference in this way enables accurate reconstruction of developmental pathways that are impossible to recover with state-based methods alone.

Introduction

Understanding the genetic and epigenetic programs that control differentiation during development is a fundamental challenge, with broad impacts across biology and medicine. Single-cell measurement technologies like single-cell RNA-sequencing (scRNA-seq) [9, 12], single-cell ATAC-seq [2] and CRISPR-based lineage tracing [13, 16, 19] have opened new windows on these processes, but it remains challenging to analyze dynamic changes in cell state over time because the measurements are destructive: cells must be lysed before information about their state can be recovered, and so a cell's state can in general only be profiled at one point along its developmental trajectory.

In response, there has been a flurry of work on designing methods to infer developmental trajectories from static snapshots of cell state [5, 25, 31, 32], including our own

efforts [21]. While initial efforts have shed some light on important biological questions relating to embryonic development [1, 5], hematopoiesis, and induced pluripotent stem cell reprogramming [21], the field of trajectory inference is still in its infancy.

One of the most significant deficiencies of previous trajectory inference methods is that they are not designed to incorporate lineage trees. Technologies for reconstructing cellular lineage trees have seen tremendous recent advances, fueled by the CRISPR–Cas9 genome editing technology [4, 16, 19]. While developmental biologists have long used various methods to tag cells and trace the lineage of their descendants, newer approaches make it possible to recover more complex lineage relationships, including the full lineage tree of a population of cells [13, 16, 19]. These newer technologies employ CRISPR-Cas9 to continuously mutate an array of synthetic DNA barcodes, which are incorporated into the chromosomes so that they are inherited by daughter cells and can be further mutated over the course of development. By analyzing the pattern of mutations in the barcodes, one can reconstruct a lineage tree describing shared ancestry within a population of cells. Recent advances allow the DNA barcodes to be expressed as transcripts and recovered together with the rest of the transcriptome in scRNA-seq [16, 19]. This enables simultaneous collection of information on cell state and cell lineage, which provides an experimental solution to part of the trajectory inference problem. Note, however, that high-resolution lineage tracing does not obviate the need for trajectory inference because lineage tracing alone does not reveal the state of the ancestral cells. While the problems of reconstructing lineage trees and inferring trajectories have attracted substantial attention individually [17, 29], there is much to be gained from combining these two complementary perspectives [6].

Here, we propose an integrated mathematical framework for inferring developmental trajectories from snapshots of both cell lineage and cell state. Our framework, called LineageOT, is broadly applicable to lineage tracing time-courses, where populations of cells are profiled with both scRNA-seq and lineage tracing at various time points along a developmental process. As a proof of concept, we test our methodology on a time-course of *C. elegans* embryonic development, collected with scRNA-seq [15]. Because the lineage tree of *C. elegans* is known [26], we have an objective measure of performance. We find that our method significantly improves trajectory inference both on this data set and on simulated examples where algorithms without lineage information cannot completely recover the correct trajectories. Our results show a path towards realizing the substantial potential benefits of lineage tracing [6, 28] in applications across developmental biology.

Results

A unified framework for lineage tracing and trajectory inference

We develop a mathematical framework for analyzing scRNA-seq time-courses equipped with a lineage tree at each time point. We formulate the goal of trajectory inference in terms of recovering the embedding of these lineage trees, defined as follows. As a population of cells develops, each cell traces out trajectories in a high-dimensional vector space of cellular states (e.g. gene expression space). Cell divisions create branching paths, and the trajectories of related cells coincide up to the point when their ancestry diverges (Fig. 1a). For example, if all the cells share a common ancestor, then the trajectories will all originate from a common point. This collection of branching paths forms the embedded lineage tree for the population. Note the emphasis on 'embedded' — without

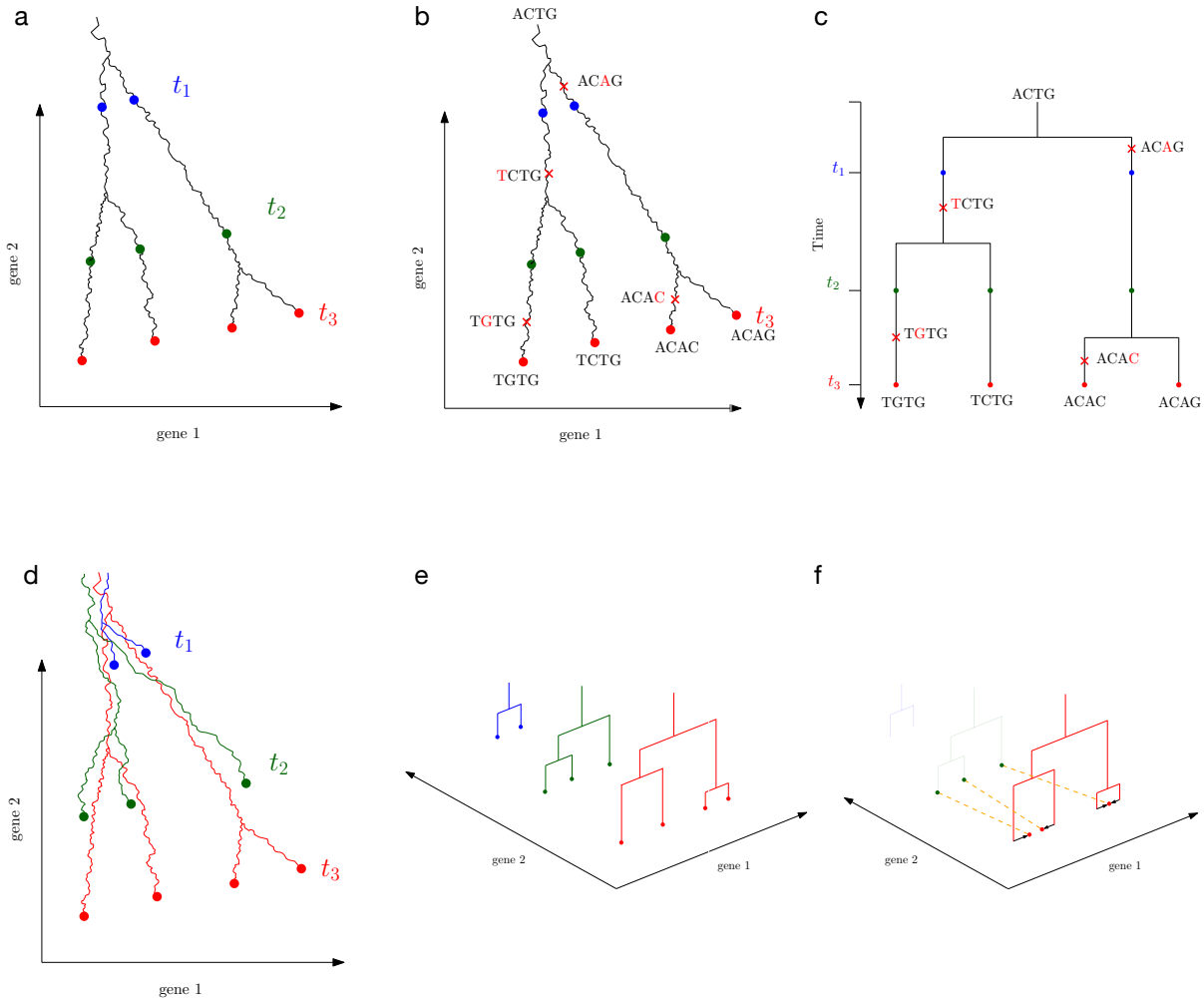

Figure 1: Schematic of the LineageOT model and inference procedure. (a) A lineage tree embedded in two dimensional gene expression space. As cells change state over time, they trace out paths. Branches in the tree correspond to cell divisions, giving rise to four cells at time t_3 (red dots). Cell states from two ancestral periods are also highlighted: three ancestors at time t_2 (green dots), and two ancestors at time t_1 (blue dots). (b) A barcode sequence is mutated over time and mutations are shown on the embedded lineage tree from (a). Starting from the ancestral barcode sequence *ACTG*, mutations are indicated with a red $\times$ on the lineage tree and the change to the sequence is shown in red. (c) The lineage tree of the population is shown with straight black lines. Note that this tree is not embedded in gene expression space. The vertical axis represents time and cell divisions correspond to bifurcations in the tree. Red $\times$s indicate mutation events. (d) Embedded lineage trees from multiple independent realizations of the developmental process. In a scRNA-seq time-course we measure just the leaves of each tree, without observing the lineage. (e) A scRNA-lineage time-course with three time points (red, green, and blue). For each time point, we observe cell states (dots) and also the lineage tree. The lineage tree is visualized in the vertical dimension, separate from the two pictured dimensions of gene expression space. (f) The LineageOT procedure consists of two steps. (1) Adjust cells at the later time, here t_3 (black arrows). The red dots show the adjusted estimates of ancestral state, based on lineage information. Note that cells with shared lineage are moved closer together. (2) Infer a coupling (dashed lines) connecting the cells from time t_2 (green) to cells from time t_3 (red).

this modifier, the term 'lineage tree' refers to the coordinate-free tree structure, where all information about the embedded state of each ancestral node is lost (Fig. 1c).

Single cell measurement technologies allow us to sample from a population and measure cell states together with barcodes that enable recovery of the lineage tree any point in time (Fig. 1b,c). However, because the measurements are destructive, we cannot directly chart the embedded lineage tree at multiple time points. One can, however, leverage the reproducibility of development and collect samples from separate populations at different time points (Fig. 1d,e). For example, one can prepare two independent populations of cells and collect samples from the first population at time t_1 and samples from the second population at time t_2. The key question is then: *which cell from the first sample would have given rise to each cell from the second sample, if these were two views of the same population*?

We have recently demonstrated [21] that a classical mathematical tool called optimal transport [8, 14, 27] can be applied to infer 'state couplings' from a scRNA-seq time-course, without any information about cell lineage. This method, called WaddingtonOT, connects cells sampled at time t_1 to their putative descendants at time t_2 by minimizing the total distance traveled by all cells. It also includes entropic regularization with a tunable regularization parameter to model the inherent stochasticity in developmental trajectories and allows for variable rates of growth across cells by adjusting the distributions at times t_1 and t_2 based on estimates of division rates. The inferred connections approximate the frequency of transitions between regions of cell-state space, i.e. the couplings of the developmental process.

Our present notion of an embedded lineage tree refines the notion of a coupling from [21]. Where WaddingtonOT aims for a state coupling describing all possible ancestries of a hypothetical cell with a given state, our embedded lineage tree gives rise to a lineage-resolved coupling. The difference is significant in situations where cells can arrive at a particular state from different ancestral states (Methods 1). Lineage tracing helps resolve these ambiguities: without lineage tracing, we must assume that cells with similar states have similar ancestral states; with lineage tracing, we instead assume that cells with similar *lineage* have similar ancestral states.

We apply optimal transport to recover lineage couplings, considered as approximations to embedded lineage trees, from scRNA-seq time-courses equipped with an unembedded lineage tree at each time point. We refer to these datasets as scRNA-lineage time-courses. In practice, the unembedded tree can be reconstructed from mutations accumulated in DNA barcodes over the course of development (Fig. 1b,c), or some lineage information might be known in advance (e.g. in *C. elegans* development). We do not focus on how the unembedded lineage tree is estimated, but we do demonstrate in simulations below that our method is robust to errors in the estimated lineage tree.

Our method applies two key steps to recover the lineage coupling spanning a pair of time points t_1, t_2. We first leverage the lineage tree to adjust the positions of the cells at time t_2 before connecting them to their ancestors at time t_1 using entropically regularized optimal transport (Fig. 1f). The adjustment in the first step can be interpreted as sharing information between closely related cells in order to construct a rough initial estimate of the ancestral states at the earlier time t_1. The rationale behind the second step is based on Schrödinger's discovery that entropically regularized optimal transport gives the maximum likelihood coupling of diffusing particles [11, 22]. Assuming that the trajectories are generated by diffusion plus drift through Waddington's landscape (Methods 2), our estimates of ancestral states from the first step are approximately normally distributed,

as we explain below. Therefore our procedure gives an approximate maximum likelihood estimate of the lineage coupling.

We now describe these two steps in detail. The core problem involves a single pair of time points, t_1 and t_2, where we are given cells $x_1, \ldots, x_n$ sampled at time t_1, and cells $y_1, \ldots, y_m$ sampled at time t_2 together with an estimate of their lineage tree. Because diffusion dominates drift on short time-scales, we can infer the ancestral state of y_i at time t_1 by assuming the dynamics are driven by pure diffusion. However, conditional on the lineage tree, the cells are not diffusing independently. Intuitively, cells with similar lineage should diffuse back towards one another to reach a common ancestral state. The difference in cell state across each of the edges of the lineage tree is given by an independent Gaussian random variable with variance proportional to the time-span along the edge (Methods 4). This implies that the ancestral state at time t_1 for each y_i is normally distributed with mean and variance that can be calculated from the lineage tree (Methods 4). Because the ancestral states of each y_i are normally distributed, optimal transport will give the maximum likelihood matching to the observed ancestors $x_1, \ldots, x_n$, when we use an entropy parameter proportional to the inferred variance of ancestral states (Methods 4). This matching, or lineage-resolved coupling, summarizes our knowledge of the ancestral states of cells from t_2 and the hypothetical descendant states of cells from t_1, providing a window onto the embedded lineage tree of each time point.

LineageOT outperforms WaddingtonOT on a lineage-resolved time-course of *C. elegans* embryonic development

We sought to test our method by applying it to a scRNA-lineage time-course. While CRISPR-based lineage tracing [3, 23] offers tremendous potential for generating scRNA-lineage time-courses, this type of dataset has not yet been published. We reasoned, however, that we could create a scRNA-lineage time-course from an ordinary, non-barcoded, scRNA-seq time-course of *C. elegans* embryonic development [15], because the lineage tree is entirely known [26]. Packer et. al sampled 86,024 cells with 10X from loosely synchronized embryos spanning the first 800 minutes of *C. elegans* embryonic development. Because the precise timing of each embryo is not known, they estimated the developmental time of each cell by correlating gene expression levels with data from a previous bulk RNA time-course [7, 15]. They then divided the cells into groups with similar estimated developmental times. We treat these groups of cells as discrete time points along a scRNA-seq time-course, using the end of each group's time interval as the group's time of sampling.

To obtain the scRNA-lineage time-course required for LineageOT, we needed to incorporate lineage information at each time point. However, lineage annotations are missing from 46% of the cells in the Packer et al. dataset. Moreover, the lineage of many of the annotated cells is not completely specified: some symmetric lineages are not distinguished (e.g. cells whose true lineage is ABprp or ABplp are all labeled as ABpxp). We explored three different strategies to get around this problem of incomplete lineage information. We first simply filtered out all cells with imperfect lineage annotation. This leaves us with only 5123 cells but with no ambiguity in the lineage tree. Second, we filtered out only cells completely lacking lineage annotation. For cells with incomplete annotations, we imputed a precise lineage label by randomly selecting from the options consistent with the partial annotation. Third, we restricted attention to the well-annotated ABpxp sublineage, which contains 7087 cells (entirely distinct from the 5123 cells above), and we

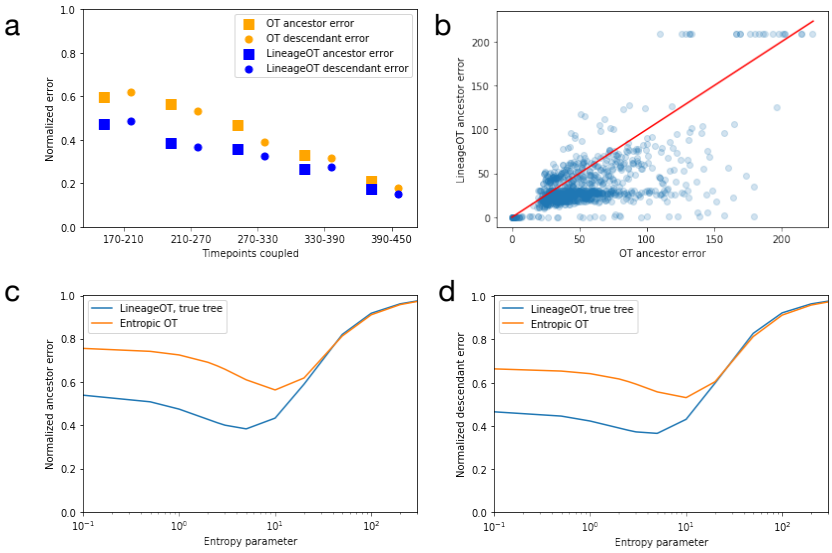

Figure 2: When tested on lineage-labeled *C. elegans* data, LineageOT outperforms optimal transport with no lineage information. (a) Relative accuracy of optimal transport and LineageOT on the 5123 cells with complete lineage annotations. Errors were normalized by dividing by the error of the noninformative independent coupling. (b) The error in predicting ancestor states, like the error for predicting descendant states (Fig. S3), is lower for most cells with LineageOT. Here each point represents one cell from the 270 minute time point, which was coupled to the 210 minute time point. The red line marks equal error for both methods. For each method in both (a) and (b), we chose the entropy parameters that gave the minimum error from parameter scans like those in (c) and (d). LineageOT consistently improves on WaddingtonOT for reasonable values of the entropy parameter, both in ancestor error (c) and descendant error (d), shown here for the 210-270 minute couplings.

treated the lineages ABprp and ABplp as if they were identical. For each approach, we also removed a small number of cells ($< 5\%$) whose assigned sampling time was before their birth time according to the reference lineage tree. These three strategies yield three scRNA-lineage datasets which we analyze separately. The results we describe below are broadly similar for each of the three strategies (Fig. 2, Fig. S1, Fig. S2).

With each strategy, we applied both WaddingtonOT [21] and LineageOT to infer developmental trajectories and compare their performance. We provide both methods with ground-truth growth rates (Methods 7), and compute state couplings and lineage couplings connecting each pair of time points. The input cell states are the first 50 coordinates from principal components analysis of the 46,159 cells with partial lineage annotations, corrected for background counts as in [15]. For LineageOT, the input lineage trees come from mapping cell lineage annotations onto the known *C. elegans* lineage tree.

We compare each fitted coupling to a ground-truth lineage coupling computed directly from the lineage-annotated data. This ground truth is constructed by connecting each early cell to all late cells labeled as being its descendants. Note that creating a coupling in this way would not be possible in other organisms without cell annotations from a known, invariant lineage tree. While previous work [17] has measured the success of trajectory inference by reducing to discrete branching representations, we directly check whether the predicted ancestors and descendants are similar in state to the true ancestors

and descendants, respectively (Methods 6). These are two separate error metrics: the ancestor prediction error and the descendant prediction error.

In all our tests, LineageOT has consistently lower error for both ancestor and descendant prediction at reasonable levels of entropy (Fig. 2a, Fig S1, Fig S2). LineageOT systematically predicts better for the majority of cells (Fig. 2b). The degree of improvement depends on the choice of entropic regularization parameter and the strategy for getting complete lineage annotations (Fig. 2c-d, Fig. S1, Fig. S2), but there is no entropy choice for which LineageOT performs significantly worse.

Lineage-based trajectory inference outperforms state-based trajectory inference on complex trajectories

We next explored the performance of LineageOT on simulated data, with the goal of characterizing some of the settings where lineage-based trajectory inference can significantly outperform state-based trajectory inference. We found that lineage information is most helpful in resolving *convergent* trajectories, where similar cells arise from different ancestral states. Moreover, we found that LineageOT is robust to imperfections in the lineage tree. Below we present three simulations illustrating these concepts.

In each simulation, we generate an embedded lineage tree by allowing an initial population of cells to follow a vector field with diffusion and also to divide. Each cell has a lineage barcode that randomly mutates and is inherited by the cell's descendants. We sample populations of cells at two time points, compute couplings with WaddingtonOT and LineageOT, and compare to the ground truth coupling from the simulation, using the ancestor and descendant prediction errors we described above. We also test the robustness of LineageOT by giving the algorithm either a lineage tree constructed from the simulated barcodes (Methods 5) or the ground-truth lineage tree.

Simulation 1 Our first example is a simple bifurcation of a single progenitor cell type into two descendant cell types (Fig. 3a). This is one of the simplest trajectory structures to recover and one where ordinary state-based inference already does well. Given a sufficiently accurate tree, LineageOT performs marginally better at ancestor prediction (Fig. 3b) and marginally worse at descendant prediction (Fig. 3c). In hindsight, this is not surprising. The lineage tree, rather than providing substantial new information, just reaffirms the natural assumption that cells in the same cluster are a bit more closely related.

Simulation 2 Inferring whether a single differentiated cell type came from multiple lineages is a common problem [24] and one of the standard goals of lineage tracing methods [28]. These convergent trajectories are difficult for state-based trajectory inference, which cannot distinguish the different ancestries of cells with similar measured states. Here we simulate two clusters that each split; after the split, two of the resulting clusters merge together (Fig. 3d). Now lineage information is important: LineageOT can separate cells in the convergent cluster by ancestry, while state-based methods cannot. Incorporating lineage information leads to substantially better prediction of ancestors than purely state-based optimal transport (Fig. 3e), without undermining descendant prediction (Fig. 3f).

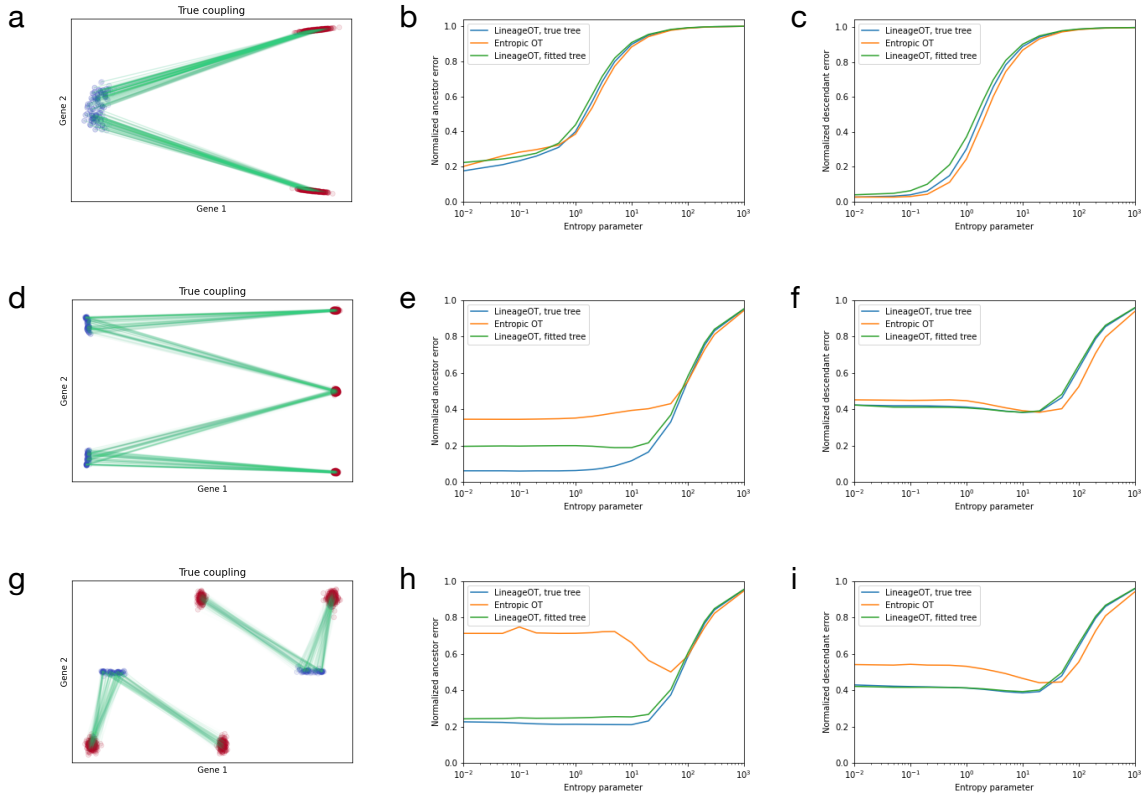

Figure 3: LineageOT matches the performance of WaddingtonOT for simple trajectories and exceeds it for complex trajectories. (a-c) For a simple bifurcation, optimal transport alone works well and adding lineage information makes little difference. (a) We simulated a cluster of cells at an early time point (blue) splitting into two clusters at a later time point (red). Green lines connect ancestors to descendants in (a), (d), and (e). The ancestor errors (b) and descendant errors (c) are similar for optimal transport and LineageOT with any entropy parameter, even when LineageOT is given an imperfect tree fitted to simulated barcodes. (d-f) For a convergent trajectory, LineageOT significantly improves ancestor prediction with no loss of accuracy in descendant prediction, even with an imperfectly fitted lineage tree. (d) Here we simulated two early clusters (blue) that each split; later, two of the resulting clusters (red) merge together. Using LineageOT reduces error substantially for ancestor prediction (e) and slightly for descendant prediction (f). (g-i) With sufficient time between samples, clusters of cells may move closer to early time point cells that are not their ancestors. (g) We again start with two early clusters (blue) that each split. In this simulation, two of the late clusters (red) are closer to non-ancestral cells than to their true ancestors. Optimal transport couples clusters incorrectly, leading to high error for predicting both ancestors (h) and descendants (i). LineageOT corrects the errors in this example by averaging with other clusters that are mapped correctly.

Simulation 3 Our third example illustrates that lineage information can go beyond resolving ambiguity and even correct mistakes from state-based inference. We consider two clusters that split so that two of the late-time clusters end up closer to early cells that are not their ancestors (Fig. 3g). Optimal transport fails dramatically in this case, mapping entire clusters to the wrong set of ancestors. The failure is not due to any mistake in the algorithm: any method that uses only state information could not correctly infer the trajectory from this data. LineageOT, on the other hand, can use the shared ancestry to match clusters correctly, leading to significantly better prediction of both ancestors and descendants (Fig. 3h-i).

Optimal transport finds the shortest possible path between the initial and final distributions; mathematically, this means following the shortest geodesic according to the optimal transport metric. Here, that is a mistake for WaddingtonOT, but not because the true trajectory is far from a geodesic. Locally, the true trajectory does minimize the distance travelled by cells. Globally, however, the geodesic followed by the true trajectory is not the shortest path between its endpoints. We can see this cleanly by plotting the true and inferred couplings with a smaller number of cells (Fig. 4a-c). We computed interpolated distributions from each coupling, found the pairwise optimal transport distances between interpolants, and visualized the resulting distance matrix with multidimensional scaling (Fig. 4d, Methods 9). The resulting paths are approximately straight lines, distorted because there cannot be two straight lines of different lengths between the same pair of points in Euclidean space. For this example, therefore, increasing the temporal resolution by sampling the system in between the two time points we present could allow optimal transport or other state-based methods to accurately describe the trajectories, albeit with greater experimental cost. However, note that this is not the case for the convergent trajectory in Simulation 2: there, adding more time points without lineage information would not help characterize cells' ancestry.

Discussion

Analyzing the trajectories cells traverse during differentiation is crucial for understanding development and harnessing the potential of stem cell therapies. However, general-purpose techniques for directly measuring differentiation trajectories have remained elusive, except in limited biological contexts, such as hematopoeisis [30] and other systems grown in suspension. Trajectory inference therefore remains the most promising general-purpose approach for understanding the genetic and epigenetic forces driving development.

We develop a unified mathematical framework for inferring developmental trajectories from scRNA-seq time-courses equipped with a lineage tree at each time-point. Lineage tracing techniques are progressing from early demonstrations of the technology [3, 23] through elaboration of the potential value of the data [28] and on towards future widespread use. We envision that scRNA-lineage time-courses will soon replace traditional scRNA-seq time-courses, because adding lineage information enables a far more powerful form of trajectory inference.

We demonstrate that LineageOT dramatically outperforms WaddingtonOT on a time-course of *C. elegans* development (Fig. 2, Fig. S1, Fig. S2), and we illustrate through simulation that LineageOT can accurately recover complex trajectory structures that are impossible to recover from measurements of cell state alone (Fig. 3d-i). Lineage trees are

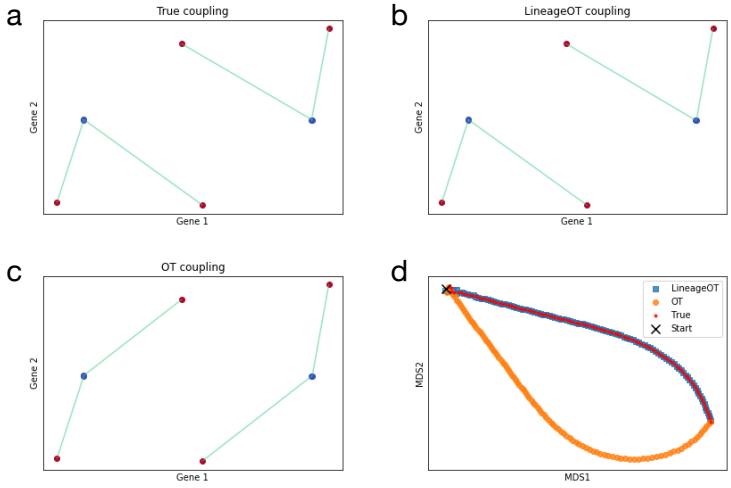

Figure 4: When the assumption that descendant states are closest to their ancestors is violated, WaddingtonOT makes clear mistakes that can be corrected with lineage information. Here the cell dynamics are the same as in Fig. 3g except for a lower division rate which leads to a smaller number of cells. The true coupling (a) matches the LineageOT coupling (b) exactly, while WaddingtonOT (c) mismatches the cells in the center. Early cells are shown in blue and late cells in red. (d) Visualization of distributions interpolated using each coupling. Each point is a distribution in between times t_1 and t_2 embedded in two dimensions using multidimensional scaling (Methods 9); the black $\times$ marks the initial distribution at time t_1, corresponding to the blue cells in (a-c). The interpolants for each coupling would follow a straight line if visualized individually; the curvature is a distortion required for a picture in two dimensional Euclidean space.

particularly helpful for untangling *convergent* trajectories, where cells arrive at a particular state from multiple ancestries. This occurs, for example, in the development of the lymphatic endothelium [24] and in macrophage development. While finer temporal resolution might allow state-based trajectory inference to succeed in some of these examples, LineageOT can achieve higher accuracy with fewer time points (Fig. 4d). Therefore our methodology has the potential to dramatically reduce the experimental cost of single cell trajectory studies of development.

Our algorithm is derived from a flexible mathematical framework that can be adapted to include future methodological advances. Most immediately, novel methods for inferring a lineage tree from any kind of experiment, or from prior knowledge, can be used directly in the LineageOT pipeline. To leverage this to its fullest extent, one could incorporate an explicit quantification of uncertainty in the lineage tree. Furthermore, there could be significant advantages to simultaneously inferring the lineage tree together with the trajectories, rather than first fitting the tree and subsequently recovering a coupling. Finally, it might be possible to incorporate additional information, beyond cell state and cell lineage. For example, measurements of RNA velocity [10] could be incorporated into our framework of estimating ancestor or descendant states and then coupling across time points. As with LineageOT, the resulting algorithm would apply optimal transport with a modified cost function.

All of these improvements would build on the key observation that lineage tracing allows us to share information across closely related cells. State-based trajectory inference relies exclusively on the assumption that each descendant considered individually should be close in state to its ancestor. As we have demonstrated, expanding that assumption to consider related cells together allows for more powerful trajectory inference that can recover more complicated trajectories without relying on the restrictive assumption that cells with similar states having similar ancestry. LineageOT analyses of future cell state and lineage time-courses collected with current technologies will provide a new, more accurate window on the intricate processes of development.

Acknowledgements

The authors are grateful to, among many others, Ruth Baker for helpful discussions and comments. AF is supported by the Royal Commission for the Exhibition of 1851. GS is supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, an NFRF Exploration Grant, and an NSERC Discovery Grant.

References

[1] James A. Briggs, Caleb Weinreb, Daniel E. Wagner, Sean Megason, Leonid Peshkin, Marc W. Kirschner, and Allon M. Klein. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science*, 360(6392), 2018.

[2] Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218, 2013.

[3] Michelle M. Chan, Zachary D. Smith, Stefanie Grosswendt, Helene Kretzmer, Thomas M. Norman, Britt Adamson, Marco Jost, Jeffrey J. Quinn, Dian Yang, Matthew G. Jones, Alex Khodaverdian, Nir Yosef, Alexander Meissner, and Jonathan S. Weissman. Molecular recording of mammalian embryogenesis. *Nature*, 2019.

[4] Wei Cong, Yun Shi, Yanqing Qi, Jinyun Wu, Ling Gong, and Miao He. Viral approaches to study the mammalian brain: Lineage tracing, circuit dissection and therapeutic applications. *Journal of Neuroscience Methods*, 335:108629, 2020.

[5] Jeffrey A. Farrell, Yiqun Wang, Samantha J. Riesenfeld, Karthik Shekhar, Aviv Regev, and Alexander F. Schier. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360(6392), 2018.

[6] Russell B Fletcher, Diya Das, and John Ngai. Creating lineage trajectory maps via integration of single-cell RNA-sequencing and lineage tracing. *Bioessays*, 40(8):e1800056, 2018.

[7] Tamar Hashimshony, Martin Feder, Michal Levin, Brian K. Hall, and Itai Yanai. Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer. *Nature*, 519(7542):219–222, 2015.

[8] Leonid Kantorovich. On the translocation of masses. *C.R. (Doklady) Acad. Sci. URSS (N.S.)*, 1942.

[9] Allon M. Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A. Weitz, and Marc W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[10] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, 2018.

[11] C. Leonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems - Series A*, 2014.

[12] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[13] Aaron McKenna, Gregory M. Findlay, James A. Gagnon, Marshall S. Horwitz, Alexander F. Schier, and Jay Shendure. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, 353(6298), 2016.

[14] G. Monge. Mémoire sur la théorie des déblais et de remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[15] Jonathan S. Packer, Qin Zhu, Chau Huynh, Priya Sivaramakrishnan, Elicia Preston, Hannah Dueck, Derek Stefanik, Kai Tan, Cole Trapnell, Junhyong Kim, Robert H. Waterston, and John I. Murray. A lineage-resolved molecular atlas of C. Elegans embryogenesis at single-cell resolution. *Science*, 365(6459), 2019.

[16] Bushra Raj, Daniel E. Wagner, Aaron McKenna, Shristi Pandey, Allon M. Klein, Jay Shendure, James A. Gagnon, and Alexander F. Schier. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nature Biotechnology*, 36(5):442–450, 2018.

[17] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, 2019.

[18] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

[19] Rina C. Sakata, Soh Ishiguro, Hideto Mori, Mamoru Tanaka, Kenji Tatsuno, Hiroki Ueda, Shogo Yamamoto, Motoaki Seki, Nanami Masuyama, Keiji Nishida, Hiroshi Nishimasu, Kazuharu Arakawa, Akihiko Kondo, Osamu Nureki, Masaru Tomita, Hiroyuki Aburatani, and Nozomu Yachie. Base editors for simultaneous introduction of C-to-T and A-to-G mutations. *Nature Biotechnology*, 38(July), 2020.

[20] Irepan Salvador-Martínez, Marco Grillo, Michalis Averof, and Maximilian J Telford. Is it possible to reconstruct an accurate cell lineage using CRISPR recorders? *eLife*, 8, 2019.

[21] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. Optimal-transport analysis of single-cell gene Eexpression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.e22, 2019.

[22] Erwin Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. *Ann. Inst. H. Poincaré*, 2:269–310, 1932.

[23] Bastiaan Spanjaard, Bo Hu, Nina Mitic, Pedro Olivares-Chauvet, Sharan Janjuha, Nikolay Ninov, and Jan Philipp Junker. Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nature Biotechnology*, 36(5):469–473, 2018.

[24] Oliver A. Stone and Didier Y.R. Stainier. Paraxial Mesoderm Is the Major Source of Lymphatic Endothelium. *Developmental Cell*, pages 1–9, 2019.

[25] 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):477, 2018.

[26] J. E. Sulston, E. Schierenberg, J. G. White, and J. N. Thomson. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Developmental Biology*, 100(1):64–119, 1983.

[27] Cédric Villani. *Optimal Transport, Old and New*. Springer-Verlag, Berlin, 2009.

[28] Daniel E. Wagner and Allon M. Klein. Lineage tracing meets single-cell omics: opportunities and challenges. *Nature Reviews Genetics*, 21(July), 2020.

[29] Caleb Weinreb and Allon M Klein. Lineage reconstruction from clonal correlations. *Proc. Nat. Acad. Sci. USA*, 2020, 2020.

[30] Caleb Weinreb, Alejo Rodriguez-Fraticelli, Fernando D. Camargo, and Allon M. Klein. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, 3381(February), 2020.

[31] Caleb Weinreb, Samuel Wolock, Betsabeh K. Tusi, Merav Socolovsky, and Allon M. Klein. Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences*, 115(10):E2467–E2476, 2018.

[32] 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.

Methods

1 State couplings and lineage couplings

A *developmental stochastic process* is a mathematical representation of a population of cells developing over time, where a single cell is represented by a point in a high-dimensional vector space of cellular states (e.g. gene expression space), and a population of cells is represented by a probability distribution on this state space. When we profile the population with scRNA-seq, we model the resulting data as a set of random samples from this probability distribution. In the context of development, a time-varying distribution $\mathbb{P}_t$ represents the cells alive at time t, and the data from a scRNA-seq time-course consists of samples from $\mathbb{P}_t$ collected at various times $t_1, t_2, \ldots, t_N$. The crucial point is that the random samples from different time points are *independent* in the probabilistic sense, because each time point is typically collected from a separate biological sample.

This brings us to the second key concept of a developmental stochastic process: the notion of a coupling connecting a pair of time points. We distinguish between two kinds of couplings: *state couplings* and *lineage couplings*. Intuitively, the state coupling connecting time t_1 to t_2 specifies relationships between ancestral states at t_1 and descendant states at t_2. Mathematically, it is a joint probability distribution over pairs of cell states (x, y), with x and y corresponding to cells alive at t_1 and t_2 respectively. Conditioning on cell state x at time t_1 gives a distribution over possible descendant states y at time t_2. In other words, while $\mathbb{P}_t$ simply describes the states of cells that exist at each time point, the state couplings specify the trajectories that give rise to the changes we observe in the population. The state couplings contain information lost in a scRNA-seq time-course: the measurements are destructive so we cannot simultaneously measure the state of a cell and the state of its ancestors or descendants.

Even state couplings, however, still omit some of the information from a specific experiment or realization of the stochastic process. A cell j at time t_2 has a true history, which may differ from the average history of cells with state equal to y_j. Lineage information makes it possible to recover the history of j in particular. The history can again be described by a coupling, this time thought of as a coupling on cells rather than on states. We refer to this coupling as a *lineage coupling*.

One reason the distinction matters is that our descriptions of cell state are incomplete. Gene expression profiles, for example, are only one easily measured part of the cell state. Cells with similar current gene expression but different history could in principle differ in other aspects of their current state. Investigating that possibility requires separating the cells by ancestry even when their current states are similar.

2 Stochastic differential equation model

We consider a cell at time t to be a point $x(t)$ in some high-dimensional space $\mathcal{X}$ such as gene expression space. Over time, cells follow some true path through $\mathcal{X}$ according to a stochastic differential equation combining diffusion and drift:

$$dX_t = v(X_t)dt + \sqrt{2D}dB_t \tag{1}$$

where v denotes a velocity field and B_t denotes standard Brownian motion. Note that diffusion dominates drift on short time-scales because diffusion is $O(\sqrt{dt})$ and drift is $O(dt)$.

In this setting, we can model an experiment as sampling a set of cell paths $\{x_i\}$ from a distribution $\mathcal{P}$ over the space of paths $[0,1] \to \mathcal{X}$. Importantly, these paths are not observed in full; we only see $x(t_1)$ for the one measurement time t_1. In a time-course experiment, in addition to measuring a set of cells $\{x_i\}$ at time t_1 we also measure a second set of identically prepared cells $\{y_j\}$ at time t_2.

We then want to couple the early and late distributions in order to trace cells forward and backward in time. As described above, a coupling γ is a joint distribution over pairs (x_i, y_j). When $\{x_i\}$ and $\{y_j\}$ are discrete sets, as they are here, γ is a matrix whose entries sum to 1.

The forward and backward questions are in principle different for lineage couplings. We could seek either a coupling γ^F such that $\gamma^F_{i,:}$, considered as a distribution on the $\{y_j\}$, is approximately the true distribution of the descendants of cell i; or we could seek a coupling γ^B such that $\gamma^B_{:,j}$, considered as a distribution on the $\{x_i\}$, is approximately the true distribution of the ancestors of cell j. For one cell, that true ancestor distribution will be a single point mass.

3 Optimal transport as maximum likelihood estimate

For both the forwards and backwards problems, entropic optimal transport can be understood as the maximum likelihood coupling between an infinite population of cells started with the distribution of $\{x_i\}$ and conditioned to end up with the distribution of $\{y_j\}$. If the likelihood of a cell at x ending at y is $p(y|x) = e^{-\frac{c(x,y)}{\epsilon}}$, maximizing the log-likelihood $\log(p(y|x,\gamma))$ leads to

$$\gamma^{ML} = \arg\min_{\gamma} \sum_{ij} \gamma_{ij} c(x_i, y_j) - \epsilon H(\gamma)$$

where $H(\gamma) = -\sum_{ij} \gamma_{ij} \log(\gamma_{ij})$ is the entropy of γ. This is precisely the objective function for optimal transport with cost $c(x,y)$ and entropy parameter ϵ.

If the times t_1 and t_2 are sufficiently close together, the dynamics of x between t_1 and t_2 are approximately purely diffusive, so that $x(t_2) - x(t_1) \sim \mathcal{N}(0, D(t_2 - t_1))$. This then translates to a quadratic optimal transport cost

$$c(x_i, y_j) = \|x_i(t_1) - y_j(t_1)\|^2$$

and entropy parameter

$$\epsilon = D(t_2 - t_1).$$

Note that because the likelihood is symmetric there is no difference between estimating forwards and estimating backwards. Other assumptions about the dynamics of the cells, such as might come from RNA velocity, could be incorporated here. Our goal with LineageOT is to find an appropriate replacement for the likelihood using the lineage information and use that as a new cost for optimal transport.

4 Ancestor inference with lineage information

A complete lineage tree $\mathcal{T}$ for $\{y_j\}$ encodes the time t_{j_1,j_2} of the most recent common ancestor of each pair of cells $\{y_{j_1}, y_{j_2}\}$. In terms of paths $y_{j_1}(t)$ and $y_{j_2}(t)$ of the cells, a common ancestor at time t_{j_1,j_2} implies that

$$\forall t \le t_{j_1,j_2}, \quad y_{j_1}(t) = y_{j_2}(t). \tag{2}$$

This gives no direct information about the unknown $\{y_j(t_1)\}$; instead, it tells us something about the correlations among $\{y_j(t_1)\}$. For LineageOT, we follow the same maximum-likelihood derivation that leads to entropic optimal transport but replace the distribution of x conditional on y_j with the distribution of x conditional on the full sample $\{y_j\}$ and the lineage tree $\mathcal{T}$:

$$\gamma^{lineage} = \arg\max_{\gamma} \log(p(x|\{y\}, \gamma, \mathcal{T})). \tag{3}$$

In the diffusive model where the differences in gene expression over time are Gaussian, if we are given the tree $\mathcal{T}$ for $\{y_j\}$ then the expression values at the nodes (i.e., the common ancestors) are sampled from a Gaussian graphical model on the tree. We can then condition on the observed values $\{y_j(t_2)\}$ to find the posterior density $p(y_j(t_1)|\{y_k(t_2)\}, \mathcal{T})$. This density will be Gaussian with each mean $\bar{y}_j(t_1)$ equal to a weighted average of the values $\{y_j(t_2)\}$. We then use an entropically regularized optimal transport coupling between $\{\bar{y}_j(t_1)\}$ and $\{x_i(t_1)\}$ to approximate the backwards coupling γ^B.

Specifically, LineageOT implements the following procedure:

1. Fit a lineage tree estimate $\hat{\mathcal{T}}$ for $\{y_j(t_2)\}$ including the estimated time of division of each most recent common ancestor, for example via neighbor-joining on CRISPR barcodes.

2. Add nodes $\{y_j(t_1)\}$ for the ancestor of each time t_2 cell at time t_1 to $\hat{\mathcal{T}}$. Some cells may share an ancestor here.

3. Pick a reference cell $y_0(t_2)$. The difference in expression of other nodes of $\hat{\mathcal{T}}$ with respect to this reference (i.e., $y_v - y_0(t_2)$) are assumed to be normally distributed with mean zero; the precision matrix has entries

$$\Lambda_{uv} = \frac{1}{D|t_u - t_v|}\mathbf{1}\left[(u, v) \in \hat{\mathcal{T}}\right].$$

4. Condition on the values $y_j(t_2)$ for the set $\mathcal{O}$ of observed nodes. The conditional means for $y_v - y_0(t_2)$ in the set $\mathcal{U} = \mathcal{O}^c$ of unobserved cells can then be found using the appropriately truncated precision matrix:

$$\mu_{\mathcal{U}} = \Lambda_{\mathcal{U}\mathcal{U}}^{-1}\Lambda_{\mathcal{U}\mathcal{O}}(y_{\mathcal{O}} - y_0).$$

5. Compute the entropic optimal transport between $\{x_i\}$ and $\{y_j\}$ with cost

$$c(i, j) = \frac{(x_i - \mu_{y_j(t_1)})^2}{\sigma^2_{y_j(t_1)}},$$

where $\mu_{y_j(t_1)}$ and $\sigma^2_{y_j(t_1)}$ are the conditional mean and variance respectively for the ancestor of each t_2 cell at time t_1.

In practice, despite being designed for ancestor prediction rather than descendant prediction, LineageOT outperforms entropic optimal transport on both tasks for all but the simplest trajectories.

5 Fitting a lineage tree

To apply LineageOT, we need to infer a lineage tree that will define the structural equation model. We do not optimize this step, instead relying on a heuristic algorithm called neighbor joining [18]. Neighbor joining starts from pairwise lineage distance estimates, which can be estimated in CRISPR-based barcoding approaches using the Hamming distances between observed barcodes [3]. The fitted tree will not be perfect, and indeed simulations with currently plausible experimental parameters find significant errors in the inferred tree topology [20]. As our own simulations demonstrate, however, an imperfectly inferred tree can still substantially improve trajectory inference. Moreover, the source of the tree does not matter: a lineage tree based on detailed prior biological knowledge, as is available for *C. elegans*, can be used directly in LineageOT.

For LineageOT, we need not only the tree topology but also the time elapsed along each edge of the tree. The raw lineage distances computed from Hamming distances, however, give very noisy estimates of the edge times. We therefore correct the distances using the fact that all cells were sampled at the same time; this means that all leaves of the tree must have the same total distance to the root. Minimizing the mean squared error to the Hamming distance estimates subject to this constraint is a quadratic program that can be solved with standard convex optimization techniques and dramatically improves the estimated lineage distances (Fig. S4).

6 Error metrics

While we only produce one estimated coupling for both ancestor and descendant prediction, we separate out the two questions in evaluation. Given a true coupling γ^*, we define the *descendant prediction error* $\mathcal{L}^D(\gamma)$ for a fitted coupling γ with the same marginal over $\{x_i\}$ as the mean squared optimal transport distance between $\gamma_{i,:}$ and $\gamma^*_{i,:}$ considered as distributions over $\{y_j\}$:

$$\mathcal{L}^D(\gamma) = \sum_i W_2^2\left(\gamma^*_{i,:}, \gamma_{i,:}\right) \tag{4}$$

where $W_2(\mu, \nu)$ denotes the optimal transport distance between distributions μ and ν with quadratic cost, also called the Wasserstein-2 distance. Symmetrically, we define the *ancestor prediction error* $\mathcal{L}^A(\gamma)$ for a fitted coupling γ with the same marginal over $\{y_j\}$ as the mean squared optimal transport distance between $\gamma_{:,j}$ and $\gamma^*_{:,j}$ considered as distributions over $\{x_i\}$:

$$\mathcal{L}^A(\gamma) = \sum_j W_2^2\left(\gamma^*_{:,j}, \gamma_{:,j}\right). \tag{5}$$

7 *C. elegans* ground truth and growth rates

Our ground truth coupling γ^* for the *C. elegans* time-course is the forward coupling based on the lineage labels: we set $\gamma^*_{ij} = (|\{x_i\}|n_{d,i})^{-1}$, where $n_{d,i}$ is the number of descendants of cell x_i in $\{y_j\}$. This forward coupling has a uniform marginal over $\{x_i\}$ but not over $\{y_j\}$. For simplicity, rather than using soft marginal constraints with estimated growth rates as WaddingtonOT does, we use the true marginals of γ^* for all fitted couplings. Knowledge of the true marginals should help WaddingtonOT and LineageOT approximately equally without significantly affecting the comparison between them.

8 Simulations

For our simulations, we construct a vector field to recreate a biologically plausible trajectory structure. Cells follow the vector field with diffusion and occasional cell division; the time between cell divisions is normally distributed with variance sufficiently small that all sampled cell lifetimes are positive. Each cell has a lineage barcode that randomly mutates and is inherited by the cell's descendants. For each vector field, we simulate a single embedded lineage tree measured at two time points and compute the couplings inferred by WaddingtonOT, LineageOT given the true lineage tree, and LineageOT given a lineage tree fitted to the simulated barcodes. Because the simulated division rates are uniform across cells, we set the marginals for each fitted coupling to be uniform rather than inputting the true marginals as we did for the *C. elegans* evaluations. The fitted couplings are compared to the true coupling with the same ancestor and descendant prediction errors we used for *C. elegans*.

9 Multidimensional scaling of interpolated distributions

Each coupling γ implicitly defines a family of distributions interpolating between the distribution of $\{x_i\}$ and $\{y_j\}$. Given a time t between t_1 and t_2 and a sample (x_i, y_j) from γ, let

$$z_{ij}(t) = x_i \frac{t_2 - t}{t_2 - t_1} + y_j \frac{t - t_1}{t_2 - t_1}. \tag{6}$$

The distribution $Z_\gamma(t)$ of $\{z_{ij}(t)\}$ continuously changes from the distribution of $\{x_i\}$ at $t = t_1$ to the distribution of $\{y_j\}$ at $t = t_2$. If γ accurately captures the true cell dynamics, the path these interpolants $Z_\gamma(t)$ follow in the space of distributions on $\mathcal{X}$ will approximately match the path of the unobserved true distributions. We sought to compare the interpolated distributions visually by embedding the paths in two dimensions.

To create a clean visualization, we simulated the double bifurcation from Fig. 3g-i with a low division rate so there were four cells at each time point. We computed interpolated distributions at 100 intermediate time points for each of the WaddingtonOT, LineageOT, and ground-truth couplings and then found the optimal transport distances between all pairs of interpolated distributions. Using that distance matrix, we embed the distributions in two dimensions using multidimensional scaling, an algorithm that attempts to place points in Euclidean space so that pairwise distances are preserved. Because the space of distributions with the optimal transport metric is both high-dimensional and non-Euclidean, the embedding necessarily has some distortion.

The interpolants from an optimal transport coupling follow the shortest geodesic between the initial and final distributions. In the example of Fig. 4, LineageOT, rather than deviating from a geodesic locally, approximately follows the correct longer geodesic between the two distributions.