nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-25133-1 OPEN

LineageOT is a unified framework for lineage tracing and trajectory inference

Aden Forrow [1]✉ & Geoffrey Schiebinger [2]✉

Understanding the genetic and epigenetic programs that control differentiation during development is a fundamental challenge, with broad impacts across biology and medicine. Measurement technologies like single-cell RNA-sequencing and CRISPR-based lineage tracing have opened new windows on these processes, through computational trajectory inference and lineage reconstruction. While these two mathematical problems are deeply related, methods for trajectory inference are not typically designed to leverage information from lineage tracing and vice versa. Here, we present LineageOT, a unified framework for lineage tracing and trajectory inference. Specifically, we leverage mathematical tools from graphical models and optimal transport to reconstruct developmental trajectories from time courses with snapshots of both cell states and lineages. We find that lineage data helps disentangle complex state transitions with increased accuracy using fewer measured time points. Moreover, integrating lineage tracing with trajectory inference in this way could enable accurate reconstruction of developmental pathways that are impossible to recover with state-based methods alone.

[1] Mathematical Institute, University of Oxford, Oxford, UK. [2] Department of Mathematics, University of British Columbia, Vancouver, BC, Canada. ✉email: aden.forrow@maths.ox.ac.uk; geoff@math.ubc.ca

Analyzing the trajectories of cellular differentiation holds promise for key questions across biology, from how lineages diverge during embryonic development to how cell types destabilize with age or are perturbed in disease. Single-cell measurement technologies like single-cell RNA-sequencing (scRNA-seq)[1,2], single-cell ATAC-seq[3], and CRISPR-based lineage tracing[4–6] have opened new windows on these processes, but it remains challenging to analyze dynamic changes in cell state and cell lineage over time because the measurements are destructive: cells must be lysed before information about their state or lineage can be recovered. In response, there has been a flurry of work on designing methods to infer developmental trajectories from static snapshots of cell state[7–10], including our own efforts[11]. While initial efforts have shed some light on important biological questions relating to embryonic development[7,12], hematopoiesis[13], and induced pluripotent stem cell reprogramming[11], the field of trajectory inference is still in its infancy.

One of the most significant deficiencies of existing trajectory inference methods is that they are not designed to incorporate the rich information from lineage tracing. Technologies for reconstructing cellular lineage trees have seen tremendous recent advances, fueled by the CRISPR-Cas9 genome editing technology[5,6,14]. 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[4–6]. These 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

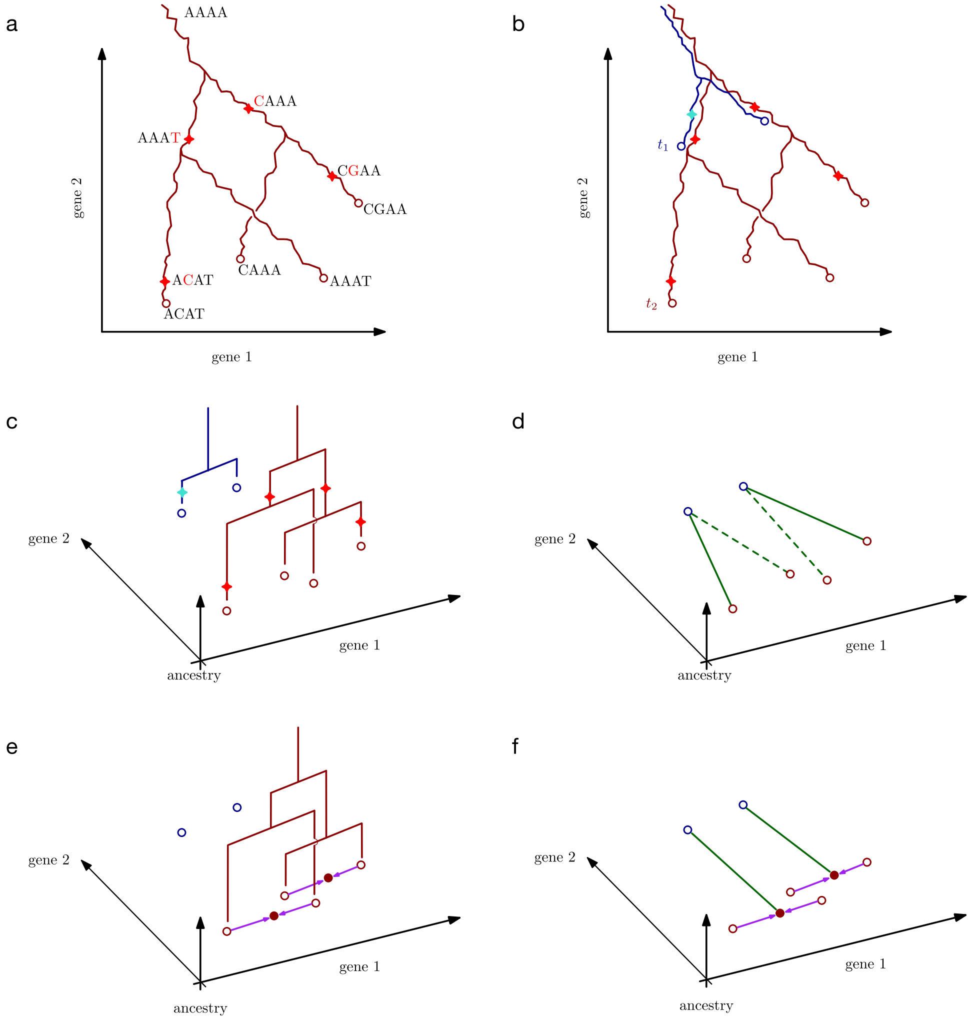

Fig. 1 Schematic of the LineageOT model and inference procedure. a A lineage tree embedded in two dimensional gene expression space. As cells change state over time, they trace out paths. Branches in the tree correspond to cell divisions, giving rise to four cells at the measurement time (red circles). Each cell has a barcode to track its lineage. Starting from the ancestral barcode sequence *AAAA*, mutations are indicated with a red star on the lineage tree and the change to the sequence is shown in red. **b** Embedded lineage trees from two independent realizations of the developmental process measured at times t_1 (blue) and t_2 (red). **c** The setup from (**b**) is shown in a 3d plot with lineage trees visualized in the vertical dimension. For each time point, we observe cell states (dots) and also the lineage tree, but not the lineage tree embedded in state space. **d** A purely state-based algorithm would fail to recover the correct trajectories in this example. Green lines connect cells at t_2 to their nearest neighbor at t_1. Dashed lines indicate erroneous connections. **e**, **f** The LineageOT procedure consists of two steps. **e** Adjust cells at time t_2 (purple arrows), based on lineage information. Cells with shared lineage are moved closer together, towards an estimate of ancestral state (solid dots). **f** Infer a coupling (green lines) connecting the adjusted cells from time t_2 (red) to cells from time t_1 (blue). This corrects the mistake made in (**d**).

together with the rest of the transcriptome in scRNA-seq[5,6]. Alternative methods use somatic mutations in mitochondria[15] to recover similar information without needing to introduce DNA barcodes.

Each of these technologies enables simultaneous collection of data on cell state and cell lineage. The two types of information are complementary: state measurements can be used for improved inference of lineage trees[16], while the lineage tree itself is intimately related to trajectories. However, high-resolution lineage tracing, even if informed by gene expression data as in[16], does not obviate the need for trajectory inference because the state of the ancestral cells remains unknown. While the problems of reconstructing lineage trees and inferring trajectories have attracted substantial attention individually[17,18], there is much to be gained from combining these two complementary perspectives[19].

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 (Fig. 1). As a proof of concept, we test our methodology on a time-course of *C. elegans* embryonic development (Figs. 2, 3), collected with scRNA-seq[20]. Because the lineage tree of *C. elegans* is known[21], we have an objective measure of performance. We find that our method significantly improves trajectory inference both on this dataset and on simulated examples where algorithms without lineage information cannot completely recover the correct trajectories (Fig. 4). Our results show a path towards realizing the substantial benefits of lineage tracing[19,22] 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 what we call the *embedded lineage tree* for the population. Note the emphasis on “embedded”—without 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, like those in 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. 1a). 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. 1b, c). 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?

Importantly, this cannot be solved in general from the topology of the lineage trees alone. Both biological variability and simple subsampling of cells in a tissue could cause the lineage tree at t_2 to be topologically distinct from an extension of the lineage tree at t_1, making it impossible to directly patch one tree onto the other in a biologically meaningful way. We sketch a hypothetical example with sampling in Fig. S1[23]. Instead, lineage information must be used together with gene expression data in a combined approach.

We have recently demonstrated[11] that a classical mathematical tool called optimal transport[24–26] can be applied to infer *state couplings* (Methods 1) from a scRNA-seq time-course, without any information about cell lineage. This method, called Waddington-OT, 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. Correctly-recovered couplings encode information about the probability of each differentiation pathway and the genes associated with fate specification.

Our present notion of an embedded lineage tree refines the notion of a coupling from[11]. While Waddington-OT 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 predicting the ancestry of the particular cells we observed. The distinction is significant in situations where cells can arrive at a particular state from different ancestral states (Methods 2). Lineage tracing helps resolve such 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.

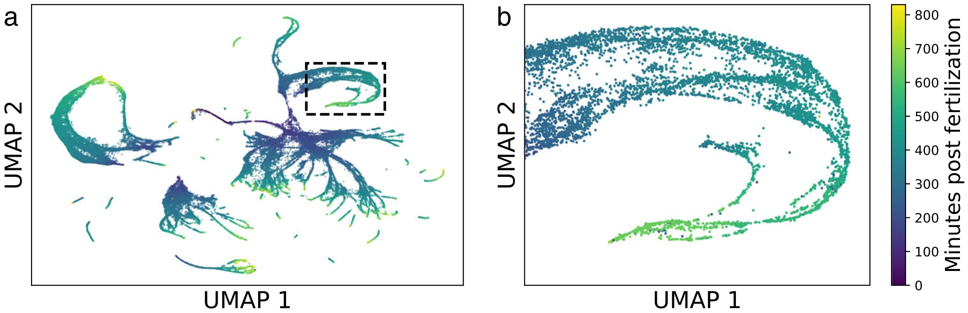

Fig. 2 Complex trajectories in *C. elegans* development. a UMAP of 81286 *C. elegans* cells from[20], using coordinates provided by Packer et al. Color indicates estimated time since fertilization following the colorbar in (**b**). **b** In the boxed region from (**a**), multiple developmental trajectories in the hypodermis converge to the same UMAP coordinates, suggesting a convergence in gene expression.

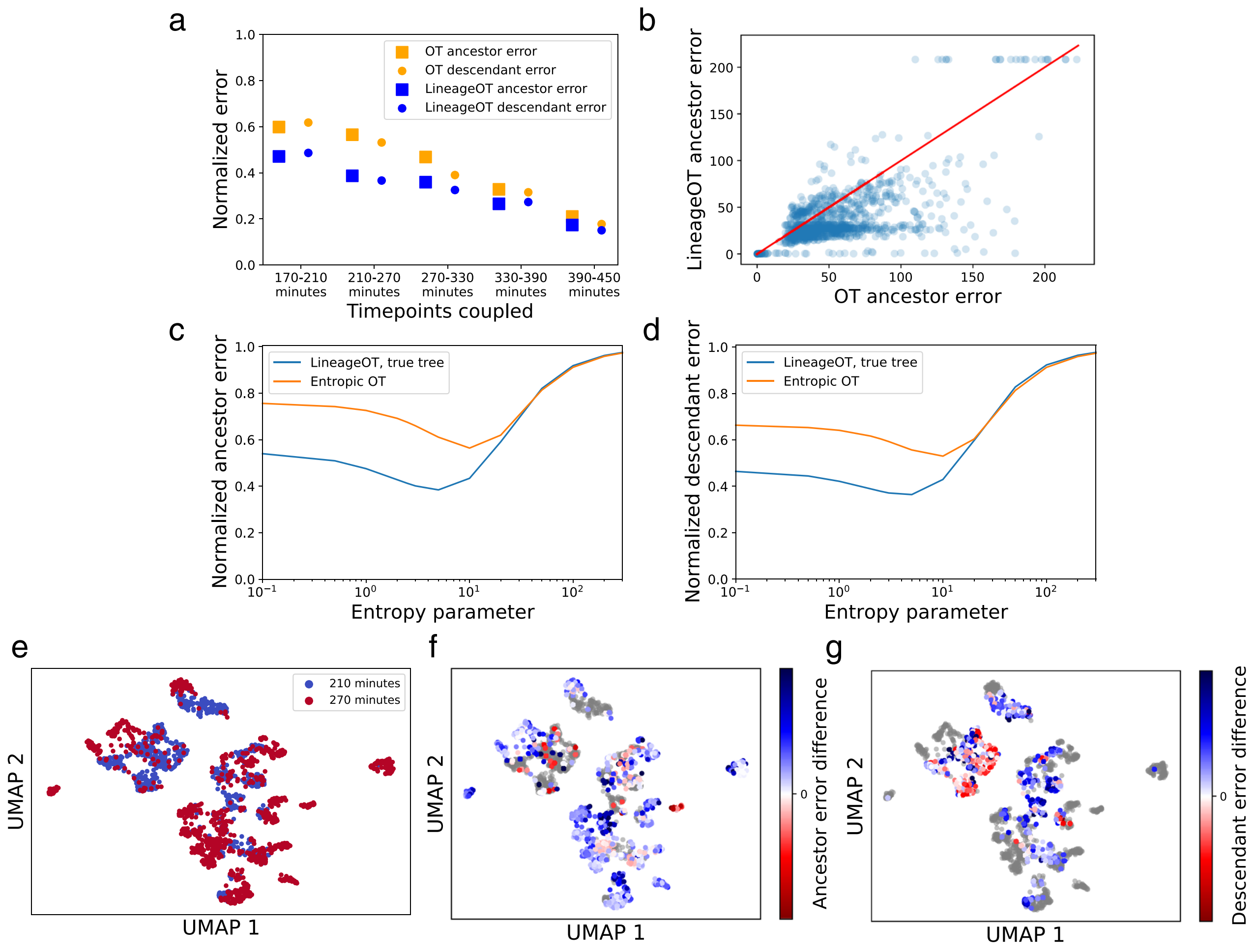

Fig. 3 When tested on lineage-labeled *C. elegans* data, LineageOT outperforms optimal transport with no lineage information. a Relative accuracy of optimal transport (OT) and LineageOT on the 5123 cells with complete lineage annotations. Errors were normalized by dividing by the error of the noninformative independent coupling. **b** The error in predicting ancestor states, like the error for predicting descendant states (Fig. S5[23]), is lower for most cells with LineageOT. Here each point represents one cell from the 270 min time point, which was coupled to the 210 min time point. The red line marks equal error for both methods. For each method in both (**a**, **b**) and (**f**, **g**), we chose the entropy parameters that gave the minimum error from parameter scans like those in (**c**, **d**). LineageOT consistently improves on Waddington-OT for reasonable values of the entropy parameter, both in ancestor error (**c**) and descendant error (**d**), shown here for the 210–270 min couplings. **e** UMAP visualization of the cells from the 210 (blue) and 270 min (red) time points. **f**, **g** Here, in the same UMAP, cells are colored by the ancestor (**f**) or descendant (**g**) error from Waddington-OT minus the same error from LineageOT. Blue indicates better performance by LineageOT, red better performance by Waddington-OT. The cells from 210 min and 270 min in (**f**) and (**g**), respectively, are gray, as the corresponding error metric does not apply to them.

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 trees can be reconstructed from mutations accumulated in DNA barcodes over the course of development (Fig. 1a), or some lineage information might be known in advance (as in *C. elegans* development). Lentiviral barcoding[13] would also provide usable, though low-resolution, lineage trees through tracing clones: cells from each clone would form leaves of a separate sub-tree, with an ancestral node at the time of barcoding. Here, we do not focus on how the unembedded lineage trees are estimated—our method assumes these are calculated separately, for example by neighbor joining[27], and given as input to the algorithm. However, 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.

1. We first leverage the lineage tree to adjust the positions of the cells at time t_2 (Fig. 1e).
2. We then connect them to their ancestors at time t_1 (Fig. 1f) using entropically regularized optimal transport.

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[28,29]. This gives a rigorous interpretation of our methodology, as we explain below.

The core problem involves a single pair of time points, t_1 and t_2, where we are given cells $\mathbf{x}_1, \ldots, \mathbf{x}_n$ sampled at time t_1, and cells $\mathbf{y}_1, \ldots, \mathbf{y}_m$ sampled at time t_2 together with an estimate of their lineage tree. We assume that these data are sampled from trajectories generated by diffusion plus drift through Waddington's landscape (i.e., a stochastic differential equation as described in Methods 2).

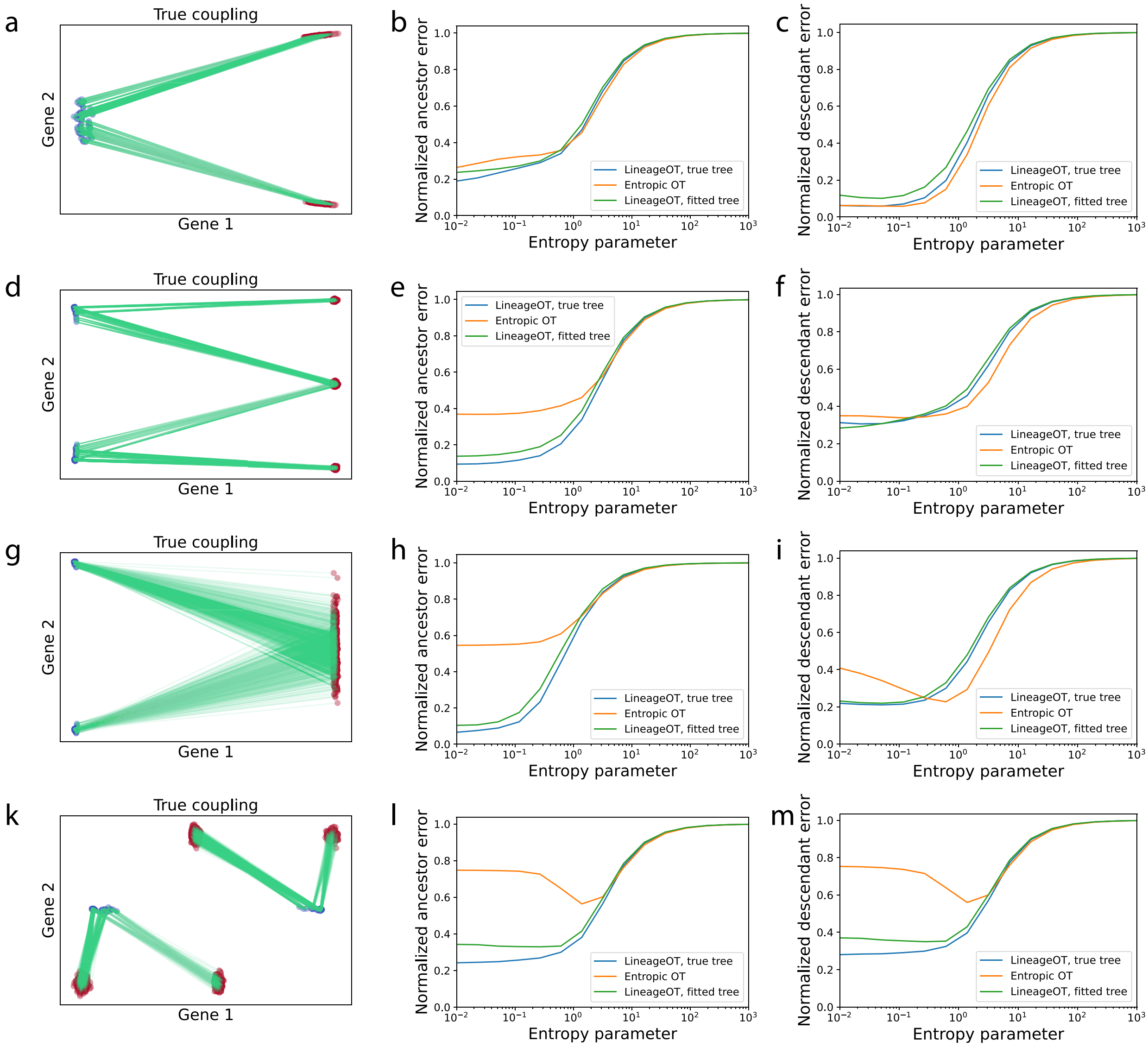

Fig. 4 LineageOT matches the performance of Waddington-OT for simple trajectories and exceeds it for complex trajectories. a–c For a simple bifurcation, optimal transport alone works well and adding lineage information makes little difference. **a** We simulated a cluster of cells at an early time point splitting into two clusters at a later time point. Green lines connect ancestors in blue to descendants in red in (**a**, **d**, **g**, **k**). The ancestor errors (**b**) and descendant errors (**c**) are similar for optimal transport (OT, orange) and LineageOT (blue) with any entropy parameter, even when LineageOT is given an imperfect tree fitted to simulated barcodes (green). **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 that each split; later, two of the resulting clusters merge together. Using LineageOT reduces error substantially for ancestor prediction (**e**) and slightly for descendant prediction (**f**). **g–i** The improvement due to lineage information when trajectories converge does not require nearby unconverged clusters. Here we see qualitatively similar improvement for two early clusters whose distributions of descendant cells almost entirely overlap. **k–m** With sufficient time between samples, clusters of cells may move closer to early time point cells that are not their ancestors. **k** In this simulation, after two early clusters each split, two of the late clusters are closer to non-ancestral cells than to their true ancestors. Optimal transport couples clusters incorrectly, leading to high error for predicting both ancestors (**l**), and descendants (**m**). LineageOT corrects the errors in this example by averaging with other clusters that are mapped correctly.

Since diffusion dominates drift on short time scales, we can estimate the ancestral state of $\mathbf{y}_i$ 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 toward 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 $\mathbf{y}_i$ is normally distributed with mean and variance that can be calculated from the lineage tree (Methods 4). Because the ancestral states of each $\mathbf{y}_i$ are normally distributed, optimal transport will give the maximum-likelihood matching to the observed ancestors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, 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.

As described in the original Waddington-OT paper[11], couplings of the kind fit by LineageOT contain trajectory information covering the entire period between the earliest and latest sampling times. Given a long time course, each method fits a coupling between each consecutive pair of time points t_i and t_{i+1}; concatenating this sequence of pairwise couplings gives connections across longer periods. In between measured time points, trajectories can be estimated with geodesic interpolation. Moreover, in cases where the lineage tree at t_2 is a direct extension of the lineage tree at t_1, the correct coupling provides precisely the matching required to patch the tree topologies.

Subsequent analyses on the cell-cell couplings will reveal aspects of the underlying biological processes, such as transcription factors associated with any transition or the timing of fate specification. These are the same set of questions that other trajectory inference methods attempt to answer; the key question in comparing approaches is their relative accuracy. For LineageOT, the foundation for reliable downstream analyses is accurate recovery of couplings between two time points, which we demonstrate in the following sections.

LineageOT outperforms Waddington-OT 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[30,31] 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[20] because the lineage tree is entirely known[21]. The known lineage tree allows us to create a ground-truth reference to directly evaluate the accuracy of fitted trajectories, making this a uniquely appropriate dataset for testing LineageOT.

Packer et al. sampled 86,024 cells with 10× from loosely synchronized embryos spanning the first 800 min of *C. elegans* embryonic development. As visible in a UMAP embedding (Fig. 2a), the differentiation process is complex. In addition to the many branchings, the gene expression of distinct transient cell types converges for several tissues, including the hypodermis (Figs. 2b, S8 of[20]) and IL1/IL2 neurons (Fig. 4a of[20]). Such convergences cause difficulties for state-based trajectory inference, because cells with similar measured state have different histories.

Because the precise timing of each embryo is not known, Packer et al. estimated the developmental time of each cell by correlating gene expression levels with data from a previous bulk RNA time course[20,32]. They then divided the cells into groups with similar estimated developmental times. We treat the six groups of cells between 130 and 450 min post fertilization 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. Using both known marker genes and UMAP trajectories, Packer et al. annotated their dataset with cells' location on the reference *C. elegans* lineage. The resulting tree can be used in LineageOT in the same way as a tree recovered from lineage tracing, with the potentially important difference that lineage labels are ultimately based on state measurements. Because lineage information is most helpful where it is not redundant with cell state data, that difference most likely biases this evaluation against LineageOT. For example, convergent trajectories that LineageOT might separate with lineage tracing independent of cell state may not be distinguishable.

While the majority of cells (54%) in the Packer et al. dataset are annotated with lineage information, 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 restricted attention to the well-annotated ABpxp sublineage, which contains 7087 cells (entirely distinct from the 5123 cells above), and we treated the lineages ABprp and ABplp as if they were identical. Third, 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. 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 (visualized in Fig. S2[23]) which we analyze separately. The results we describe below are broadly similar for each of the three strategies (Figs. 3, S3, S4 in[23]).

With each strategy, we applied both Waddington-OT[11] 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 normalized, log-transformed counts for the 46,159 cells with partial lineage annotations, corrected for background counts as in[20]. For LineageOT, the input lineage trees come from mapping cell lineage annotations onto the known *C. elegans* lineage tree. This additional information, given to LineageOT and not Waddington-OT, is precisely the information now measurable with lineage tracing.

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. 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[18] 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. As an alternative, performance can be evaluated by the probability ancestor-descendant pairs from the ground truth are correctly linked. Though LineageOT does improve on Waddington-OT by that metric (Fig. S6[23]), we prefer the ancestor and descendant error metrics because they do not require assuming the lineage trees from the two time points match.

In all our tests, LineageOT has consistently lower error for both ancestor and descendant prediction at reasonable levels of entropy (Figs. 3a, S3, S4 in [23]), including after concatenating couplings across more than two time points (Fig. S7[23]). LineageOT systematically predicts better for the majority of cells (Fig. 3b). The degree of improvement depends on the choice of entropic regularization parameter and the strategy for getting complete lineage annotations (Figs. 3c, d, S3, S4 in[23]), but there is no entropy choice for which LineageOT performs significantly worse. The increased accuracy comes from effectively using the information in the lineage tree.

Lineage-informed 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 four 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 (Methods 8). 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 Waddington-OT 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) a lineage tree constructed from the simulated barcodes using a heuristic algorithm called neighbor-joining[27] (Methods 5) or (b) the ground-truth lineage tree. For comparison, in the supplement we present the results of applying PAGA[10]a well-regarded trajectory inference method that only uses state information, to our simulated data.

Our first example, Simulation 1, is a simple bifurcation of a single progenitor cell type into two descendant cell types (Fig. 4a). 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. 4b) and marginally worse at descendant prediction (Fig. 4c). 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.

Inferring whether a single differentiated cell type came from multiple lineages is a common problem[33] and one of the standard goals of lineage tracing methods[22]. These convergent trajectories are difficult for state-based trajectory inference, which cannot distinguish the different ancestries of cells with similar measured states. In Simulation 2, we simulate two clusters that each split; after the split, two of the resulting clusters merge together (Fig. 4d). 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. 4e), without undermining descendant prediction (Fig. 4f).

We included two unmerged clusters at the late time in Simulation 2 to illustrate how lineage information can resolve ambiguity: cells whose ancestry is unclear from expression alone should be coupled similarly to their close relatives in the lineage tree with unambiguous ancestry. The existence of separate unconverged clusters is not, however, necessary for separating cells by ancestry. In Simulation 3, we show two clusters that converge to a single final cell type (Fig. 4g). The distributions of descendants from each early cluster overlap too much for state-based methods like Waddington-OT to accurately infer ancestors. Despite the overlap, the descendant distributions remain sufficiently distinct for LineageOT to have nearly perfect assignment of late cells to early clusters, and thereby low ancestor error (Fig. 4h) with no loss of accuracy in descendant prediction (Fig. 4i).

Our fourth example illustrates that lineage information can go beyond resolving ambiguity and even correct mistakes from state-based inference. For Simulation 4, 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. 4k). Optimal transport fails 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, as shown for PAGA in Fig. S8[23]. 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. 4l–m).

In Simulation 4, increasing the temporal resolution by sampling the system in between the two time points could allow optimal transport or other state-based methods to accurately describe the trajectories. In real biological systems with this type of curled trajectory, lineage tracing may limit the need for many expensively-sampled time points, though it does add other costs like integrating the barcode editing technology into the genome. For convergent trajectories like Simulation 2 or 3, adding more time points without lineage information would not be enough: in that setting, lineage tracing is necessary for correct trajectory inference.

Discussion

Analyzing the trajectories cells traverse during differentiation is crucial for understanding development and for harnessing the potential of stem cell therapies. However, general-purpose techniques for directly measuring trajectories of cellular differentiation remain elusive. In certain biological contexts, such as hematopoeisis[13] where cells are grown in suspension, daughter cells can be split and measured at different time points. These direct connections between time points allow for improved trajectory inference[34]. Such specialized techniques, however, are not applicable to systems where cells are adherent, because splitting daughter cells would perturb the developmental process. Trajectory inference from independent snapshots remains the most promising approach for understanding the genetic and epigenetic forces driving development in diverse biological contexts.

We present a general-purpose method for inferring developmental trajectories from scRNA-seq time courses equipped with lineage information each time point. Lineage tracing techniques are progressing from early demonstrations of the technology[30,31] through elaboration of the potential value of the data[22] and on toward 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 outperforms Waddington-OT on a time-course of *C. elegans* development (Figs. 3, S3, S4 in[23]), 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. 4d–i). Effectively using the lineage information experimentally accessible within each time point intuitively ought to improve inference accuracy across time points. LineageOT realizes that implicit potential.

Lineage trees are 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[33], macrophage development from embryonic or monocyte-derived progenitors[35], mouse gut endoderm[36], and several tissues in *C. elegans* ([20], Figs. 4a, S8, S17). 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. The couplings we recover enable a direct, rigorous approach to answer biological questions about ancestor-descendant relation in developmental processes and predict regulators that govern those transitions, as demonstrated by[11].

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[37] 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 more accurate window on the intricate processes of development.

Methods

Developmental stochastic processes and state couplings. A *developmental stochastic process* is a mathematical description of a population of cells developing over time, where a single cell is represented by a point in a high-dimensional vector space $\mathcal{X}$ of cellular states and a population of cells is represented by a probability distribution $\mathbb{P}$ on $\mathcal{X}$. In deriving our method, we do not assume any particular state space: a vector $\mathbf{x} \in \mathcal{X}$ could contain a cell's raw gene expression, coordinates from principal components analysis of expression as used in our *C. elegans* case study, or other state measurements like chromatin accessibility.

When we profile the population with scRNA-seq, we model the resulting data as a set of random samples from $\mathbb{P}$. 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* (defined here) and *lineage couplings* (defined in the next section). 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 $(\mathbf{x}, \mathbf{y})$, with $\mathbf{x}$ and $\mathbf{y}$ corresponding to cells alive at t_1 and t_2 respectively. Conditioning on cell-state $\mathbf{x}$ at time t_1 gives a distribution over possible descendant states $\mathbf{y}$ 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.

Forward and backward lineage couplings. 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 $\mathbf{y}_j$. 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 of specific sampled cells rather than of states. We name this a *lineage-resolved coupling* (or "lineage coupling" for short).

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 state measurements are similar.

Above, we introduced a developmental stochastic process from the perspective of time-dependent probability distributions $\mathbb{P}_t$ connected by state couplings. From a complementary perspective, we can consider the time evolution of individual cells rather than distributions.

Each cell state at time t is a point $\mathbf{x}(t)$ in the state space $\mathcal{X}$. Over time, cells follow some true path through $\mathcal{X}$ according to a stochastic differential equation combining diffusion and drift:

$$\mathrm{d}\,\mathbf{X}_t = \mathbf{v}(\mathbf{X}_t)\,\mathrm{d}t + \sqrt{2D}\mathrm{d}\,\mathbf{B}_t \tag{1}$$

where $\mathbf{v}$ denotes a velocity field and $\mathbf{B}_t$ denotes standard Brownian motion scaled by the diffusion constant D. On short time scales, diffusion, which is $O(\sqrt{\mathrm{d}\,t})$, dominates drift, which is $O(\mathrm{d}t)$.

In this model, an experiment samples a set of cell paths $\{\mathbf{x}_i(t)\}$ 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 $\mathbf{x}_i = \mathbf{x}_i(t_1)$ for the one measurement time t_1. In a time-course experiment, in addition to measuring a set of cell states $\{\mathbf{x}_i\}$ at time t_1 we also measure the states $\{\mathbf{y}_j\}$ of a second set of identically prepared cells 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; for a lineage coupling, these are pairs of cells (i, j) rather than pairs of states $(\mathbf{x}, \mathbf{y})$. When $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_j\}$ are discrete sets, as they are here, γ is a matrix whose entries sum to 1. The true lineage coupling connects each cell j from the late-time point to its unobserved ancestor at the early time point. Though that true coupling is experimentally inaccessible, we can attempt to couple early or late cells respectively to cells similar in state to their hypothetical descendants or true ancestors.

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 $\{\mathbf{y}_j\}$, 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 $\{\mathbf{x}_i\}$, is approximately the true distribution of the ancestors of cell j. For one cell, that true ancestor distribution will be a single point mass.

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 $\{\mathbf{x}_i\}$ and conditioned to end up with the distribution of $\{\mathbf{y}_j\}$. If the likelihood of a cell at $\mathbf{x}$ ending at $\mathbf{y}$ is $p(\mathbf{y}|\mathbf{x}) = e^{-\frac{c(\mathbf{x},\mathbf{y})}{\epsilon}}$, maximizing the log-likelihood $\log(p(\mathbf{y}|\mathbf{x}, \gamma))$ leads to

$$\gamma^{ML} = \underset{\gamma}{\arg\min} \sum_{ij} \gamma_{ij} c(\mathbf{x}_i, \mathbf{y}_j) - \epsilon H(\gamma) \tag{2}$$

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 $c(\mathbf{x}, \mathbf{y})$ and entropy parameter ϵ.

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

$$c(\mathbf{x}_i, \mathbf{y}_j) = \| \mathbf{x}_i(t_1) - \mathbf{y}_j(t_1)\|^2 \tag{3}$$

and entropy parameter

$$\epsilon = D(t_2 - t_1). \tag{4}$$

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 cost for optimal transport.

Ancestor inference with lineage information. A complete lineage tree $\mathcal{T}$ for $\{\mathbf{y}_j\}$ encodes the time t_{j_1, j_2} of the most recent common ancestor of each pair of cells $\{\mathbf{y}_{j_1}, \mathbf{y}_{j_2}\}$. In terms of paths $\mathbf{y}_{j_1}(t)$ and $\mathbf{y}_{j_2}(t)$ of the cells, a common ancestor at time t_{j_1, j_2} implies that

$$\forall t \leq t_{j_1, j_2}, \quad \mathbf{y}_{j_1}(t) = \mathbf{y}_{j_2}(t). \tag{5}$$

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

$$\gamma^{lineage} = \underset{\gamma}{\arg\max}\ \log(p(\mathbf{x}|\{\mathbf{y}\}, \gamma, \mathcal{T})). \tag{6}$$

Like Waddington-OT, LineageOT is derived from the diffusive model where the differences in cell state over time are Gaussian. In that model, conditional on the connectivity of $\mathcal{T}$ the cell-state values at the nodes (i.e., the common ancestors of the leaves) are sampled from a Gaussian graphical model on $\mathcal{T}$. We can then additionally condition on the observed values $\{\mathbf{y}_j(t_2)\}$ to find the posterior density $p(\mathbf{y}_j(t_1)|\{\mathbf{y}_k(t_2)\}, \mathcal{T})$. This density will be Gaussian with each mean $\bar{\mathbf{y}}_j(t_1)$ equal to a weighted average of the values $\{\mathbf{y}_j(t_2)\}$. We then use an entropically regularized optimal transport coupling between $\{\mathbf{y}_j(t_1)\}$ and $\{\mathbf{x}_i(t_1)\}$ to approximate the backwards coupling γ^B.

Specifically, LineageOT implements the following procedure:

1. Fit a lineage tree estimate $\hat{\mathcal{T}}$ for $\{\mathbf{y}_j(t_2)\}$ including the estimated time of division of each most recent common ancestor, for example via neighbor joining on CRISPR barcodes.
2. Add nodes $\{\mathbf{y}_j(t_1)\}$ 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 $\mathbf{y}_0(t_2)$. The difference in state of other nodes of $\hat{\mathcal{T}}$ with respect to this reference (i.e., $\mathbf{y}_v - \mathbf{y}_0(t_2)$) is assumed to be normally distributed with mean zero; the precision matrix has entries
$$\Lambda_{uv} = \frac{1}{D|t_u - t_v|}\mathbf{1}\left[(u, v) \in \hat{\mathcal{T}}\right]. \tag{7}$$
4. Condition on the values $\mathbf{y}_j(t_2)$ for the set $\mathcal{O}$ of observed nodes. The conditional means for $\mathbf{y}_v - \mathbf{y}_0(t_2)$ in the set $\mathcal{U} = \mathcal{O}^c$ of unobserved nodes can then be found using the appropriately truncated precision matrix: for each gene g, with $\mathbf{1}$ the vector of $|\mathcal{O}|$ ones,
$$\mu^g_{\mathcal{U}} = \Lambda^{-1}_{\mathcal{U}\mathcal{U}}\Lambda_{\mathcal{U}\mathcal{O}}(\mathbf{y}^g_{\mathcal{O}} - y^g_0\mathbf{1}). \tag{8}$$
5. Compute the entropic optimal transport coupling between $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_j\}$ with cost
$$c(i, j) = \frac{\| \mathbf{x}_i - \mu_{\mathbf{y}_j(t_1)} \|^2}{\sigma^2_{\mathbf{y}_j(t_1)}}, \tag{9}$$
where $\mu_{\mathbf{y}_j(t_1)}$ and $\sigma^2_{\mathbf{y}_j(t_1)}$ 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.

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[27]. Neighbor joining starts from pairwise lineage distance estimates, which can be estimated in CRISPR-based barcoding approaches using the Hamming distances between observed barcodes[30]. The fitted tree will not be perfect, and indeed simulations with currently plausible experimental parameters find significant errors in the inferred tree topology[38]. 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 significantly improves the estimated lineage distances (Fig. S9[23]).

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 $\{\mathbf{x}_i\}$ as the mean squared optimal transport distance between $\gamma_{i,:}$ and $\gamma^*_{i,:}$ considered as distributions over $\{\mathbf{y}_j\}$:

$$\mathcal{L}^D(\gamma) = \sum_i W^2_2\left(\gamma^*_{i,:}, \gamma_{i,:}\right) \tag{10}$$

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 $\{\mathbf{y}_j\}$ as the mean squared optimal transport distance between $\gamma_{:,j}$ and $\gamma^*_{:,j}$ considered as distributions over $\{\mathbf{x}_i\}$:

$$\mathcal{L}^A(\gamma) = \sum_j W^2_2\left(\gamma^*_{:,j}, \gamma_{:,j}\right). \tag{11}$$

***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 $\gamma^*_{ij} = (|\{\mathbf{x}_i\}|n_{d,i})^{-1}$ if i is an ancestor of j and $\gamma^*_{ij} = 0$ otherwise, where $n_{d,i}$ is the number of descendants of cell i in $\{j\}$. This forward coupling has a uniform marginal over $\{\mathbf{x}_i\}$ but not over $\{\mathbf{y}_j\}$. For simplicity, rather than using soft marginal constraints with estimated growth rates as Waddington-OT does, we use the true marginals of γ^* for all fitted couplings. Knowledge of the true marginals should help Waddington-OT and LineageOT approximately equally without significantly affecting the comparison between them.

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 nearly constant, with normally distributed variability with a small variance (so that all sampled cell lifetimes are positive). Changing this variance in cell division times or setting it to 0 does not significantly affect our results. Meaningful dynamics occur in either two dimensions (for Simulations 1, 2, and 3) or three dimensions (for Simulation 4). The vector field is always constant in the first dimension, making x_1 a proxy for time since the start of the simulation. In the remaining nontrivial one or two dimensions, we set the vector field to be the negative gradient of a potential with minima at locations we would like to have clusters, with the minima changing with x_1. We simulate in three dimensions in all cases; for Simulations 1, 2, and 3, in the third dimension cells diffuse with no mean velocity. Thus, for example, the simple bifurcation of Simulation 1 follows the flow field

$$\mathbf{v}(x) = \begin{pmatrix} v_1(x) \\ v_2(x) \\ v_3(x) \end{pmatrix} = \begin{pmatrix} 1 \\ -x^3_2 + x_1x_2 \\ 0 \end{pmatrix}. \tag{12}$$

Initially, with $x_1 < 0$, $x_2 = 0$ is the only stable value of x_2; later, with $x_1 > 0$, there are two stable states $x_2 = \pm\sqrt{x_1}$. For the remaining simulations, which involve more complex piecewise-smooth flow fields, we refer readers to our code: https://github.com/aforr/LineageOT.

Each cell has a lineage barcode that randomly mutates and is inherited by the cell's descendants. The global mutation rate r is set so that the expected proportion of sites in a barcode of length ℓ that are unmutated at the time of sampling t_2, equal to $e^{-\frac{rt_2}{\ell}}$, is close to 0.5 and so relatively far from both 0 and 1. The rate is then neither so slow that little lineage information is recorded so fast that barcodes are saturated before the sampling time. This choice was inspired by similar numbers in experimental data from[30].

For each vector field, we simulate a single embedded lineage tree measured at two time points and compute the couplings inferred by Waddington-OT, 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*.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The *C. elegans* data is available on GEO with accession code GSE126954[39].

Code availability

A Python package implementing LineageOT and all simulations is available at https://github.com/aforr/LineageOT with documentation at https://lineageot.readthedocs.io[40].

Received: 7 October 2020; Accepted: 8 July 2021;
Published online: 16 August 2021

References

1. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
2. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
3. Buenrostro, J. D. et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
4. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, **353**, aaf7907-1–aaf7907-11 (2016).
5. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).
6. Sakata, R. C. et al. Base editors for simultaneous introduction of C-to-T and A-to-G mutations. *Nat. Biotechnol.* **38**, 865–869 (2020).
7. Farrell, J. A. et al. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, eaar3131 (2018).
8. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genom.* **19**, 477 (2018).
9. Weinreb, C. et al. Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl Acad. Sci.* **115**, E2467–E2476 (2018).

10. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 1–9 (2019).
11. Schiebinger, G. et al. Optimal-transport analysis of single-cell gene Eexpression identifies developmental trajectories in reprogramming. *Cell* **176**, 928–943.e22 (2019).
12. Briggs, J. A. et al. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, eaar5780 (2018).
13. Weinreb, C. et al. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **360**, eaaw3381 (2020).
14. Cong, W. et al. Viral approaches to study the mammalian brain: Lineage tracing, circuit dissection and therapeutic applications. *J. Neurosci. Methods* **335**, 108629 (2020).
15. Ludwig, L. S. et al. Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* **176**, 1325–1339.e22 (2019).
16. Zafar, H. et al. Single-cell lineage tracing by integrating CRISPR-Cas9 mutations with transcriptomic data. *Nat. Commun.* **11**, 3055 (2020).
17. Weinreb, C. & Klein, A. M. Lineage reconstruction from clonal correlations. *Proc. Nat. Acad. Sci. U.S.A.* **117**, 17041–17048 (2020).
18. Saelens, W. et al. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
19. Fletcher, R. B. et al. Creating lineage trajectory maps via integration of single-cell RNA-sequencing and lineage tracing. *Bioessays* **40**, e1800056 (2018).
20. Packer, J. S. et al. A lineage-resolved molecular atlas of C. Elegans embryogenesis at single-cell resolution. *Science* **365**, eaax1971 (2019).
21. Sulston, J. E. et al. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev. Biol.* **100**, 64–119 (1983).
22. Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* **21**, 410–427 (2020).
23. Forrow, A. & Schiebinger, G. Supplementary Material: LineageOT is a Unified Framework for Lineage Tracing and Trajectory Inference. https://doi.org/10.1101/2020.07.31.231621 (2021).
24. Kantorovich, L. On the translocation of masses. *C.R. (Doklady) Acad. Sci. URSS (N.S.)* (1942).
25. Monge, G. Mémoire sur la théorie des déblais et de remblais. *Histoire de l'Académie Royale des Sciences de Paris* (1781).
26. Villani, C. *Optimal Transport, Old and New*. (Springer-Verlag, 2009).
27. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
28. Leonard, C. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems—Series A* **34**, 1533–1574 (2014).
29. Schrödinger, E. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. *Ann. Inst. H. Poincaré* **2**, 269–310 (1932).
30. Chan, M. M. et al. Molecular recording of mammalian embryogenesis. *Nature* **570**, 77–82 (2019).
31. Spanjaard, B. et al. Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).
32. Hashimshony, T. et al. Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer. *Nature* **519**, 219–222 (2015).
33. Stone, O. A. & Stainier, D. Y. R. Paraxial mesoderm is the major source of lymphatic endothelium. *Dev. Cell* **50**, 1–9 (2019).
34. Prasad, N. et al. Optimal transport using GANs for lineage tracing. *arXiv* Preprint at https://arxiv.org/abs/2007.12098 (2020).
35. Varol, C. et al. Macrophages: development and tissue specialization. *Annu. Rev. Immunol.* **33**, 643–675 (2015).
36. Nowotschin, S. et al. The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature* **569**, 361–367 (2019).
37. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
38. Salvador-Martínez, I. et al. Is it possible to reconstruct an accurate cell lineage using CRISPR recorders? *eLife*, **8**, e40292 (2019).
39. Packer, J. S. et al. A lineage-resolved molecular atlas of C. Elegans embryogenesis at single-cell resolution, published data. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126954 (2019).
40. Forrow, A. & Schiebinger, G. LineageOT is a Unified Framework for Lineage Tracing and Trajectory Inference. https://github.com/aforr/LineageOT, https://doi.org/10.5281/zenodo.5018867 (2021).

Acknowledgements

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

Author contributions

A.F. and G.S. conceived the project. A.F. designed the method with input from G.S. A.F. wrote the code. A.F. and G.S wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-25133-1.

Correspondence and requests for materials should be addressed to A.F. or G.S.

Peer review information *Nature Communications* thanks Laleh Haghverdi, Wouter Saelens and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.