Integrative Inference of Spatially Resolved Cell Lineage Trees using LineageMap

Xinhai Pan[1], Yiru Chen[1], and Xiuwei Zhang[1,2]

[1] School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta GA 30332, USA
[2] Corresponding Author: xiuwei.zhang@gatech.edu

Abstract. Understanding the spatio-temporal processes of tissue growth, including how new cell types emerge and how cells form the tissue architecture, is a fundamental problem in biology. The emerging spatially resolved lineage tracing data, where three modalities, lineage barcodes, gene expression profiles, and spatial locations, are measured for each single cell, provides an unprecedented opportunity to understand these processes. Computational methods that take advantage of all three modalities to reconstruct cell lineage tree and ancestral cell states and locations are needed. We introduce LineageMap, a hybrid lineage inference algorithm that integrates the scalability of distance-based tree reconstruction methods with the flexibility of likelihood-based methods under a unified probabilistic framework. The input to LineageMap is spatially resolved lineage tracing data, where for each single cell, the gene expression, lineage barcode and spatial locations are available. LineageMap enables accurate, interpretable, and scalable inference of high-resolution lineage trees as well as locations of ancestral cells from the tri-modality single-cell data. Across simulated and experimental datasets, LineageMap consistently outperforms existing methods in the accuracy of reconstructed cell lineage trees, while revealing biologically coherent spatiotemporal trajectories. Our framework bridges molecular lineage tracing with spatial and transcriptomic information, advancing computational reconstruction of dynamic cellular ancestries in both time and space. LineageMap is available at: https://github.com/ZhangLabGT/LineageMap.

Keywords: Cell lineage inference · Spatial transcriptomics · Evolutionary modeling · Maximum likelihood · Spatio-temporal processes

1 Introduction

Understanding how cells divide, differentiate, and organize into tissues and organs is a central question in developmental biology. Lineage tracing technologies aim to reconstruct the *cell division history*, or lineage tree, of a tissue by recording heritable molecular marks that accumulate during successive cell divisions. Recent advances in sequencing-based lineage tracing,[22,27] particularly those leveraging CRISPR/Cas9 genome editing, have enabled joint measurement of lineage barcodes and transcriptomes across thousands of single cells.[28] These barcodes consist of engineered genomic sequences—lineage recorders—that undergo stochastic mutations at designated target sites over successive cell divisions,[1,4,24] that can be used to reconstruct detailed lineage trees to uncover latent developmental trajectories and division hierarchies that cannot be directly observed *in vivo*. In a standard lineage tree, each node represents a single cell, and each edge denotes a direct parent–daughter division event.

More recently, a new wave of spatiotemporal lineage tracing datasets are emerging,[11–13,29] which integrate lineage tracing with spatial transcriptomics technologies. This combination enables the simultaneous observation of cell lineage barcodes, expression and spatial locations at single-cell resolution, which respectively represent cell division, differentiation, and spatial organization. Such tri-modality datasets make it possible to reconstruct not only the ancestral relationships among cells, but also the spatial trajectories of cell divisions and migrations – giving rise to the concept of *spatially resolved lineage trees*. These results offer unprecedented opportunities to understand the coordination between lineage, spatial architecture and cell fate.

However, reconstructing spatially resolved lineage trees poses substantial computational challenges. Existing lineage reconstruction methods focus primarily on optimizing the topology of cell division trees from barcode mutations,[8,21] and some recent methods use barcode mutations and single cell gene expression data.[17,30] Without considering spatial information of cells, these methods do not take advantage of the mutual constraints between lineage, spatial proximity, and cell state continuity — key factors in developmental processes. To fully leverage these new datasets, there is a need for integrative frameworks that jointly infer lineage structure while modeling spatial coherence and state transitions.

Classical Neighbor Joining (NJ)[20] efficiently infers lineage tree topology based on pairwise distances between lineage barcodes. However, it assumes all taxa (cells) are distinct and that distances faithfully capture ancestral relationships. In real lineage barcode data, this assumption is violated due to dropout events and sparse mutation patterns, where many loci remain unedited or ambiguously sequenced, while *homoplasy*,[26] where identical or near-identical barcodes, arise independently in different cells. On the other hand, maximum-likelihood (ML)[6] or Bayesian tree inference methods[9,31] are becoming computationally intractable for the ever-increasing number of cells sequenced in lineage tracing experiments, as the number of possible unrooted trees increases in the order of $n!!$ for n taxa. Furthermore, local search algorithms used in likelihood optimization are prone to convergence toward local optima, particularly in high-dimensional barcode spaces.[7,14,15,25] These challenges render full ML inference impractical for modern lineage recording datasets comprising thousands of cells.

Besides these challenges, the *polytomy* problem[18,23] is also critical, where many cells share identical barcodes causing one internal node having degree greater than three in an unrooted tree, additional information sources such as spatial proximity or transcriptomic similarity can be integrated to enhance the distinguishability of cells and to refine local tree structures.

To this end, taking advantage of the cutting-edge spatially-resolved lineage tracing datasets, we developed LineageMap, a hierarchical hybrid framework that integrates the computational scalability of distance-based inference with the accuracy and flexibility of likelihood-based optimization. By coupling statistical modeling with multi-modal context, LineageMap enables accurate and interpretable reconstruction of high-resolution lineage trees from large-scale, dropout-prone single-cell datasets, paving the way for deeper insights into the spatial and temporal dynamics of development, regeneration, and disease progression. Our tests on both simulated and real datasets demonstrated that LineageMap outperforms state-of-the-art methods for lineage tree reconstruction.

2 Methods

2.1 Overview of LineageMap

In order to optmize computational cost and enhance robustness against technical noise in the barcode data, LineageMap first aggregates cells with highly similar lineage barcodes by applying Louvain clustering on a similarity graph derived from a dropout-aware weighted Hamming distance matrix on the barcodes, then obtains a cluster-level consensus barcode for each cluster, and applies Neighbor Joining to the consensus barcodes, which yields a backbone tree of cell clusters (Fig. 1).

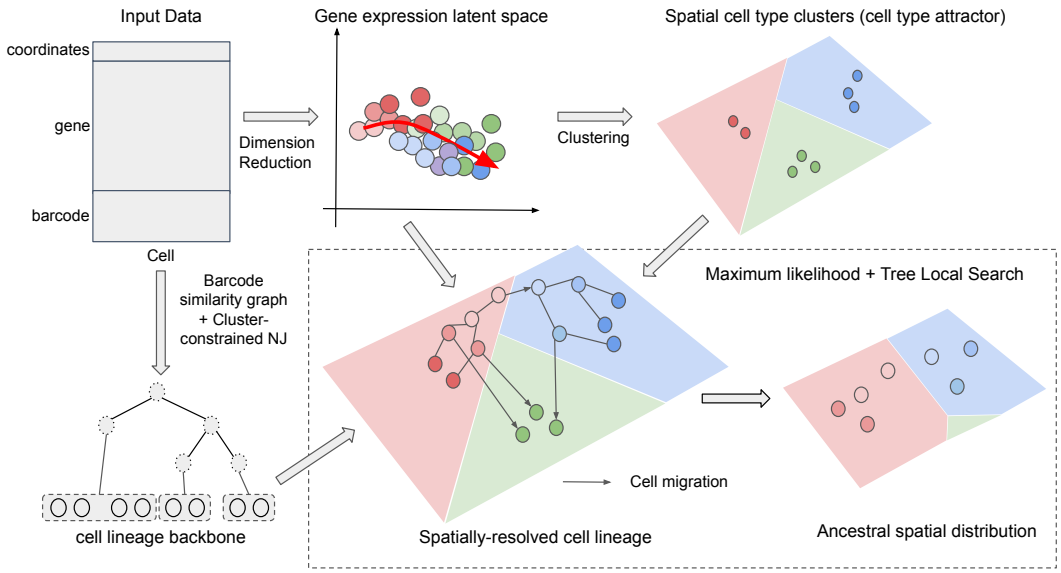

Fig. 1. Overview of LineageMap algorithm. LineageMap integrates spatial coordinates, gene expression, and DNA barcodes to infer spatially resolved cell lineages. Gene expression is embedded and clustered to define cell-state groups, while a barcode-similarity graph with cluster-constrained Neighbor Joining provides an initial tree backbone. Using a maximum-likelihood framework with local tree search, LineageMap refines this backbone by incorporating spatial constraints and cluster structure to produce a spatially resolved lineage and its inferred ancestral spatial distribution.

LineageMap then performs maximum likelihood optimization within each cluster to determine the tree topology and branch lengths for these cells using multi-modal information, including barcode states, spatial coordinates, and cell states (Fig. 1). By modeling processes of barcode mutations, cell type differentiation and cells' spatial localization during cell divisions, the likelihood function couples these relevant constraints, thus the three modalities together are expected to lead to more accurate reconstructed cell lineage trees with higher resolution, compared to methods using only the barcode or the barcode and gene expression data.

2.2 Cell lineage tree

We define a rooted, directed, bifurcating cell lineage tree $T = (V, E)$ with leaf set corresponding to the observed cells, internal nodes representing unobserved ancestors, and branch lengths $\{t_e\}_{e \in E}$ proportional to elapsed time. An edge e between two nodes represents a parent-daughter cell relationship. Let $\mathbf{b}_v$, $\mathbf{s}_v$ and $\mathbf{z}_v$ denote the (latent) barcode, spatial coordinates, and cell state of node v.

2.3 Spatially-resolved lineage tree inference problem

Let n be the number of observed single cells. For each cell $i \in \{1, \ldots, n\}$ we observe:

- a lineage barcode $\mathbf{b}_i = (b_{i1}, \ldots, b_{iL})$ across L target sites,
- a spatial coordinate $\mathbf{s}_i \in \mathbb{R}^d$ ($d = 2$ in most of the cases, and $d = 3$ if 3-D spatial locations are used),
- and a categorical *cell state* label $\mathbf{z}_i \in \{1, \ldots, C\}$.

We seek a rooted tree $T = (V, E)$ that best explains the observed tri-modality data through spatio-temporal cell division and localization processes. We will first introduce our method to detect clones of cells and to build the backbone lineage. Then, we will formulate the generative models for cells' lineage barcodes, spatial locations, and cell states, under a unified Bayesian framework, to derive the likelihood function to optimize.

2.4 Detection of lineage clones and tree backbone inference

To address the lineage polytomy problem, we first identify coherent *lineage clones*—groups of cells sharing highly similar or identical barcodes—using a community detection approach on the barcode similarity graph. Specifically, we apply Louvain clustering with a tunable resolution threshold to partition cells into candidate clonal groups. Each clone represents a local subtree that can be independently refined before integrating into a global lineage topology. The threshold parameter should be tuned to the dropout level in the data, and the detailed description of determining the threshold value can be found in Supplementary Material.

We then infer a coarse-grained *backbone tree* representing the relationships among clones. We adapt the standard NJ method to reconstruct this tree while ensuring cells in the same clone to form a subtree. We name this NJ method the "cluster-constrained NJ" method. The resulting tree is a backbone tree that provides the overall topology while allowing fine-grained optimization within the clone.

Finally, each clone subtree is inferred in parallel using a stochastic optimization framework that maximizes a likelihood function, that jointly models the evolution of barcode, cell states and cell locations. After local refinement, all subtrees are merged back into the backbone tree by replacing clone tips on the backbone with their corresponding optimized subtrees.

2.5 Generative models of spatio-temporal lineage trees

Barcode evolution model

We model the evolution of CRISPR-Cas9 lineage barcodes as an irreversible mutational continuous-time Markov chain (CTMC) process on a finite set of editing outcomes. Each barcode consists of L independent target sites ("cut sites"), indexed by $l = 1, \ldots, L$. For cell i, the observed barcode is $\mathbf{b}_i = (b_{i1}, \ldots, b_{iL})$, where $b_{il} \in \mathcal{A}_l$, and $\mathcal{A}_l = \{0, a_1, \ldots, a_{K_l}\}$ denotes the character space of locus l. Here, 0 corresponds to the unedited state, while $a_1, \ldots, a_{K_l}$ correspond to distinct irreparable mutations ("scars") generated by Cas9-induced indels. Additionally, an extra symbol $\varnothing$ is included to represent dropout or missing observations.

The underlying CTMC starts in state 0 at the root cell. A character in the 0 state of the parent barcode mutates at a rate of $0 < \lambda < 1$. When a character in the 0 state mutates, it acquires a new state from the set $\mathcal{A}_l = \{0, a_1, \ldots, a_{K_l}\}$, where the state a_i is acquired with probability $P(a_i)$. This way, the CTMC is parameterized by $\lambda > 0$ and the state probabilities $q_{a_i} \geq 0$. We denote $\theta_b = (\lambda, q_{a_1}, q_{a_2}, \ldots, q_{a_{K_l}})$.

Spatial evolutionary model of cell division and differentiation

For a daughter cell v that divided from a parent cell u, the spatial location of v, s_v, follows a diffusion model along the tree,

$$\mathbf{s}_v \mid \mathbf{s}_u, t_e \sim \mathcal{N}\big(\mathbf{s}_u + t_e\, \mathbf{d}(\mathbf{s}_u, \mathbf{z}_u),\, \sigma_s^2 t_{uv} I_d\big),$$

where $\mathbf{d}(\cdot)$ is a possible drift field (e.g., movement along tissue axes or chemotactic gradients), and σ_s^2 is spatial diffusivity. We allow $\mathbf{d}$ to depend on the cell state $\mathbf{z}_u$, which means cells in different states migrate differently. t_{uv} is the edge length of edge (u, v). I_d is an $d \times d$ identity matrix, where $d = 2$.

Brownian motion model of cell states

We model the cell state of node v, $\mathbf{z_v}$, as a multivariate trait evolving via a Brownian motion process, depending on the cell state of its parent, $\mathbf{z_u}$:

$$\mathbf{z}_v \mid \mathbf{z}_u, t_e \sim \mathcal{N}\big(\mathbf{m}(\mathbf{z}_u, t_e),\, \Sigma_z(t_e)\big),$$

where $\mathbf{m}(\mathbf{z}_u, t_{uv}) = \mathbf{z}_u + t_e\, \beta(\mathbf{z}_u)$ can include drift toward attractor states (lineage-specific differentiation directions).

2.6 Joint likelihood and posterior

The observed data is $\mathcal{D} = \{\mathbf{b}_i, \mathbf{s}_i, \mathbf{z}_i\}_{i=1}^n$, which represent the lineage barcodes, spatial locations and cell states of leaf cells. The joint likelihood given $(T, \{t_e\}, \theta)$ factorizes, with latent node variables marginalized:

$$\mathcal{L}(T, \theta) = \prod_{l=1}^{L} p\big(\{b_{il}\}_{i=1}^n \mid T, \theta_b\big) \;\times\; p\big(\{\mathbf{z}_i\}_{i=1}^n \mid T, \theta_z\big) \;\times\; p\big(\{\mathbf{s}_i\}_{i=1}^n \mid T, \theta_s\big). \tag{1}$$

The three terms that multiply are respectively the likelihood of lineage barcodes, cell states and cell locations on a lineage tree T. These three terms are detailed in following sections. The posterior probability then becomes:

$$p(T, \theta \mid \mathcal{D}) \propto p(T)\, p(\theta)\, \mathcal{L}(T, \theta),$$

where $p(T)$ is the prior for the tree candidate T and $p(\theta)$ includes priors for the latent parameters, including barcode mutation rate, state probabilities, and spatial drift parameters.

2.7 Barcode likelihood

Let $T = (V, E)$ be a rooted tree with root r, internal nodes V_{int}, and leaves corresponding to the n observed cells. Each leaf i has an observed barcode $\mathbf{b}_i = (b_{i1}, \ldots, b_{iL})$ across L independent loci. For locus l, the finite state space is $\mathcal{A}_l$ as defined earlier. Denote by t_{uv} the length of edge $(u \to v) \in E$.

Barcode state transition model. For locus l, edits occur irreversibly from the unedited state. That is, once a locus is edited, its state is locked and will not mutate again. Let λ_{lk} be the rate of editing from un-mutated state 0 to a mutated state k, with total mutation rate $\lambda_l = \sum_{k=1}^{K_l} \lambda_{lk}$. On an edge of length t, following Camin-Sokal Parsimony,[2] we can derive the transition probabilities between barcode states: Thus, for edge $(u \to v)$ of length t_{uv},

$$P_{uv}^{(l)}[0 \to 0] = e^{-\lambda_l t_{uv}}, \quad P_{uv}^{(l)}[0 \to k] = \tfrac{\lambda_{lk}}{\lambda_l}\big(1 - e^{-\lambda_l t_{uv}}\big), \quad P_{uv}^{(l)}[k \to k] = 1.$$

A matrix form of the transition probabilities is included in Supplementary Material.

Given the transition probabilities, we use a bottom-up approach to recursively calculate the barcode likelihood on the whole tree. Define the likelihood of a subtree rooted at node v at locus l as $\ell_v^{(l)}(a) :=$ Pr(observations in subtree rooted at $v \mid v$ has character a) .

If v is a leaf node, denote the observation by $o_{vl} \in \mathcal{A}_l \cup \{\varnothing\}$, where $\varnothing$ denotes dropout. In case of dropout events in the data, $\ell_v^{(l)}(a) = \Pr(o_{vl} \mid \text{true mutation} = a)$; If there is no dropout, $\ell_v^{(l)}(a) = \mathbf{1}\{o_{vl} = a\}$.

If v is an internal node with children $\mathrm{ch}(v)$, we have:

$$\ell_v^{(l)}(k) = \prod_{c \in \mathrm{ch}(v)} \ell_c^{(l)}(k), \qquad k \geq 1,$$

$$\ell_v^{(l)}(0) = \prod_{c \in \mathrm{ch}(v)} \Big[e^{-\lambda_l t_{vc}}\, \ell_c^{(l)}(0) + (1 - e^{-\lambda_l t_{vc}}) \sum_{k=1}^{K_l} \tfrac{\lambda_{lk}}{\lambda_l}\, \ell_c^{(l)}(k)\Big].$$

When we get to the root r, the likelihood $\ell_r^{(l)}(k)$ becomes the likelihood of the whole tree for locus l. With prior $\pi_r^{(l)}$ on the root, the likelihood on locus l can be calculated as $p(\text{data}_l \mid T) = \sum_{a \in \mathcal{A}_l} \pi_r^{(l)}(a)\, \ell_r^{(l)}(a)$; If the root is fixed to the unedited state, such as in the CRISPR/Cas9 model, the per-locus likelihood is $p(\text{data at locus } l \mid T) = \ell_r^{(l)}(0)$.

Assuming loci evolve independently, the total likelihood is

$$\mathcal{L}(T) = \prod_{l=1}^{L} p(\text{data}_l \mid T), \qquad \log \mathcal{L}(T) = \sum_{l=1}^{L} \log p(\text{data}_l \mid T).$$

2.8 Likelihood of state and space evolution

We next consider the joint evolution of discrete cell states and continuous spatial locations of cells that evolve following $T = (V, E)$ with root r and edge lengths $\{t_{uv}\}$. Leaves $i = 1, \ldots, n$ have observed cell state z_i and continuous spatial coordinate $s_i \in \mathbb{R}^d$. The joint likelihood of cell states and locations of leaf cells is

$$\mathcal{L}(T, \theta) = \Pr(\{z_i, s_i\}_{i=1}^n \mid T, \theta),$$

computed by marginalizing over unobserved internal node variables. When deriving this likelihood, we use CTMC to model cell state evolution and Gaussian message passing to model cells' spatial location evolution.

Discrete state evolution using CTMC. Cell states $z \in \{1, \ldots, C\}$ evolve along the tree as a continuous-time Markov chain with rate matrix R. For edge $(u \to v)$ of length t_{uv}, we define the transition probability matrix M, where

$$M_{uv} = \exp(R t_{uv}), \qquad M_{uv}[a \to b] = \Pr(z_v = b \mid z_u = a).$$

We then use a bottom-up recursive procedure to calculate the cell state likelihood on the tree, similar to what we do with the barcode likelihood. The recursion aggregates subtree likelihoods by summing over child state transitions and multiplying across independent branches. We define $\ell_v(b)$, as the likelihood of a subtree rooted at node v. When v is a leaf, we have $\ell_v(b) = e_v(b)$, where $e_v(b) = \Pr(o_v \mid z_v = b)$ if there may be dropouts, and $\ell_v(b) = \mathbf{1}\{o_v = b\}$ in dropout-free data.

Then for internal node v with state a and children $\text{ch}(v)$, we have:

$$\ell_v(a) = \prod_{c \in \text{ch}(v)} \Big(\sum_{b=1}^{C} M_{vc}[a \to b]\, \ell_c(b) \Big).$$

Finally, for root node i with root prior π_r, we have

$$p(\{o_i\} \mid T, R) = \sum_{a=1}^{C} \pi_r(a)\, \ell_r(a)$$

Spatial coordinate evolution with Gaussian message passing. Let $s \in \mathbb{R}^d$ denote continuous spatial coordinates. Along an edge $(u \to v)$ of length t, we use a Brownian motion model for the shift of daughter cells from the parent cell:

$$s_v \mid s_u \sim \mathcal{N}(s_u,\ \Sigma_e(t)), \qquad \Sigma_e(t) = \sigma^2 t\, I_d \quad \text{where } \sigma \text{ controls the step size.}$$

Message passing. Suppose child c contributes message $p(\text{data in subtree } c \mid s_c) \propto \mathcal{N}(s_c; m_c, S_c)$. Integrating over s_c yields a Gaussian in s_v, we have:

$$\int \mathcal{N}(s_c; s_v, \Sigma_e(t_{vc}))\, \mathcal{N}(s_c; m_c, S_c)\, ds_c = \mathcal{N}(s_v; m_c, S_c + \Sigma_e(t_{vc})).$$

Thus each child contributes a Gaussian distribution in s_v, and their product is still a Gaussian distribution.

Information form update. For child c, define

$$S_{c\to v} = S_c + \Sigma_e(t_{vc}), \quad \Lambda_{c\to v} = S_{c\to v}^{-1}, \quad \eta_{c\to v} = \Lambda_{c\to v} m_c.$$

Aggregating across children, we have:

$$\Lambda_v = \sum_{c\in \mathrm{ch}(v)} \Lambda_{c\to v}, \qquad \eta_v = \sum_{c\in \mathrm{ch}(v)} \eta_{c\to v}, \qquad S_v = \Lambda_v^{-1},\ m_v = S_v \eta_v.$$

Likelihood at the root. If the root has prior $s_r \sim \mathcal{N}(\mu_0, \Sigma_0)$, then with message $\mathcal{N}(s_r; m_r, S_r)$ from the subtree,

$$p(\text{data}) = \mathcal{N}(m_r; \mu_0, S_r + \Sigma_0).$$

2.9 State-dependent spatial OU model with lineage-specific attractors

In order to model cell migration, we extend the Brownian motion model to Ornstein–Uhlenbeck(OU) model (Details about the extension derivations can be found in Supplementary Material). Let $T = (V, E)$ be a rooted tree with branch $(u \to v)$ of length $t_{uv} > 0$. Each node v has a d-dimensional spatial coordinates $\mathbf{s}_v \in \mathbb{R}^d$ and a cell state $z_v \in \{1, ..., \mathcal{C}\}$ (observed at leaves, latent at internals). Dynamics along an edge are governed by the *parent* state z_u.

For each state $z \in \{1, ..., \mathcal{C}\}$, we posit a scalar OU rate $\alpha_s > 0$, a scalar diffusion $\sigma_s^2 > 0$ (isotropic covariance), and an attractor $\boldsymbol{\mu}_s \in \mathbb{R}^d$. The OU transition on edge $(u \to v)$ is

$$\mathbf{s}_v \mid \mathbf{s}_u, z_u = s, t \ \sim\ \mathcal{N}\Big(A_s(t)\,\mathbf{s}_u + \mathbf{b}_s(t),\ S_{e,s}(t)\Big),$$

with

$$A_s(t) = e^{-\alpha_s t}\, I_q, \qquad \mathbf{b}_s(t) = \big(1 - e^{-\alpha_s t}\big)\,\boldsymbol{\mu}_s, \qquad S_{e,s}(t) = \frac{\sigma_s^2}{2\alpha_s}\big(1 - e^{-2\alpha_s t}\big)\, I_q.$$

For $t = 0$ use the continuous limits $A_s(0) = I_q$, $\mathbf{b}_s(0) = \mathbf{0}$, $S_{e,s}(0) = 0$.

Upward message from child c to parent u. Suppose the child's upward message is Gaussian $\mathcal{N}(\mathbf{m}_c, S_c)$. Let $W_s(t) \equiv \big(S_c + S_{e,s}(t)\big)^{-1}$. With $A = A_s(t)$ and $\mathbf{b} = \mathbf{b}_s(t)$, the child contribution on $\mathbf{s}_u$ is Gaussian with information parameters

$$\Lambda_{c\to u} \ =\ A^\top W_s(t)\, A \ =\ e^{-2\alpha_s t}\, W_s(t), \qquad \boldsymbol{\eta}_{c\to u} \ =\ A^\top W_s(t)\,\big(\mathbf{m}_c - \mathbf{b}\big).$$

At a parent u with children $\mathrm{ch}(u)$, combine by summation

$$\Lambda_u = \sum_{c\in \mathrm{ch}(u)} \Lambda_{c\to u}, \qquad \boldsymbol{\eta}_u = \sum_{c\in \mathrm{ch}(u)} \boldsymbol{\eta}_{c\to u}, \qquad S_u = \Lambda_u^{-1},\ \mathbf{m}_u = S_u \boldsymbol{\eta}_u.$$

At the root r, combine with a Gaussian prior $\mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$ in the usual way:

$$\Lambda_r^\star = S_r^{-1} + \Sigma_0^{-1}, \quad \boldsymbol{\eta}_r^\star = S_r^{-1}\mathbf{m}_r + \Sigma_0^{-1}\boldsymbol{\mu}_0, \quad S_r^\star = (\Lambda_r^\star)^{-1},\ \mathbf{m}_r^\star = S_r^\star \boldsymbol{\eta}_r^\star.$$

Marginal spatial log-likelihood (accumulated). During the upward pass, add

$$\log \mathcal{L}_{\text{spatial}} \ \mathrel{+}=\ \sum_{(u\to c)\in E} \log \mathcal{N}\big(\mathbf{m}_c;\ A_s(t)\mathbf{m}_u + \mathbf{b}_s(t),\ S_c + S_{e,s}(t)\big) \ +\ \log \mathcal{N}\big(\mathbf{m}_r;\ \boldsymbol{\mu}_0,\ S_r + \Sigma_0\big).$$

The posterior means $\{\mathbf{m}_u\}_{u\in V}$ provide the inferred internal coordinates.

2.10 Construct the full lineage tree and whole tree refinement

The joint likelihood in Eq. 1 is used to optimize subtrees for each cell clone, using a local search algorithm. We adopted random Subtree Swapping (rSS), which is a derivative of the widely used random Subtree Pruning and Regrafting (rSPR).[5] We randomly select two nodes on the current tree and prune the subtrees attached to the specific nodes. Then, we regraft either one of the pruned subtrees to the location of the other subtree. The advantage of rSS is that with the initialization of a binary tree, this operation will not disrupt the binary property throughout the searching process.[17]

With the inferred tree backbone and the optimized subtrees, LineageMap builds the initial full lineage by re-attaching the subtrees onto the tree backbone at the leaves. In the end, we run an additional whole tree likelihood evaluation to infer ancestral state and spatial coordinates.

2.11 Simulating synthetic datasets with SpaTedSim

To quantitatively evaluate LineageMap and other methods under controlled and biologically realistic conditions, we developed SpaTedSim (Spatio-Temporal dynamics Simulation of single cells), a simulator that jointly models cell division, cell state shifts, and cellular movement in 2-D space (details in Supplementary Material and Supplementary Fig. 1). SpaTedSim directly simulates how the cell population grows from a single cell to observed cells with various cell types. With the modeling of the cell state tree, cell lineage tree, and spatial movements of cells, SpaTedSim generates realistic, paired gene expression, lineage barcodes, and spatial coordinates, and with the ground truth, we can look into the spatial distribution of cell types and lineages and uncover the spatio-temporal dynamics of cells. SpaTedSim consists of two major steps: 1. Simulating cells' lineage barcode and gene expressions on the cell division tree; 2. Simulating cells' spatial coordinates based on the cell division tree.

3 Results

3.1 LineageMap outperforms existing methods on synthetic datasets

SpaTedSim provides complete ground-truth lineage trees and spatial coordinates for every cell, enabling quantitative assessment of reconstruction accuracy. Given the observed cells' spatial locations (Fig. 2a), LineageMap can infer spatially-resolved lineage tree and infer ancestral spatial locations (Fig. 2b). We can compare the inferred lineage tree with the ground truth and benchmark against other methods.

We benchmarked LineageMap against state-of-the-art methods: LinRace,[17] which uses both lineage barcodes and gene expression; Cassiopeia,[10] a parsimony-based heuristic; Startle,[21] a technique based on star homoplasy evolutionary model, and an NJ baseline on the weighted Hamming distances. We tested the methods across multiple simulation regimes that vary in three hyperparameters: the number of cells, target sites, and dropout rates (Fig. 2c-f). These conditions capture the major factors of difficulty in barcode-based lineage reconstruction. Evaluation metrics include Robinson–Foulds (RF) distance,[19] Nye similarity,[16] and path-length correlation, capturing complementary aspects of tree topology and branch-length structure (details on the metrics are in Supplementary Material).

In the benchmark setting with 1024 cells, 128 characters and 20% dropouts (Fig. 2c), LineageMap achieves the best overall performance, having the best RF distance and Nye Similarity, and second to Startle for path length correlation. To comprehensively compare the performances of the methods under various conditions, in Fig. 2d-f, we further compare the methods with a smaller number of cells (128 cells), a smaller number of target sites (16 target sites), or higher dropout rate (up to 80%).

When the number of cells is reduced, most methods improve due to the reduction in tree complexity. Nevertheless, LineageMap maintains leading performance across all three metrics, with LinRace being the second best. Surprisingly, Startle fails to recover meaningful topology, producing near-random topologies revealed by RF distance and Nye Similarity.

With only 16 target sites, the mutation signal becomes extremely sparse, creating a challenging inference scenario. As expected, all methods show reduced accuracy, but the performance gap widens. LineageMap

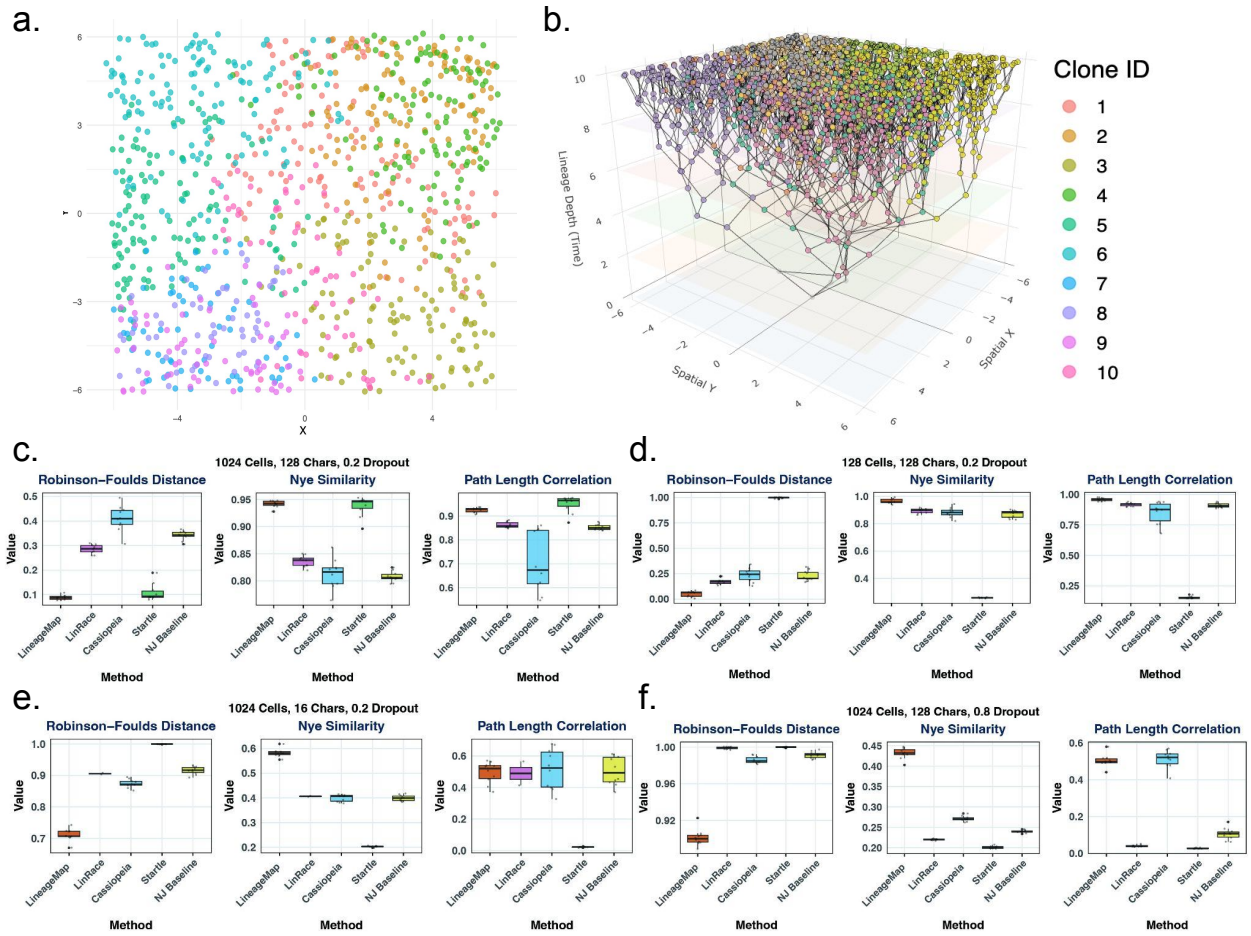

Fig. 2. (a–b) Visualization of the default simulation dataset (replicate 1).(a) 2D scatter plot of cell spatial coordinates (X–Y), colored by Clone ID (major sub-lineages from hierarchical clustering on the reconstructed tree; average linkage on cophenetic distances, k=10), showing clear spatial segregation.(b) Plotly 3D visualization of the reconstructed lineage tree, with X–Y denoting spatial position, Z denoting evolutionary depth, and colors indicating clone ID.(c–f) Benchmark results for LineageMap, NJ, Cassiopeia (two modes), and Startle across simulated conditions. LineageMap consistently achieves the best or near-best performance in RF distance (lower is better), Nye similarity, and path-length correlation (higher is better). Panels show results for (c) 1024 cells, 128 characters, 0.2 dropout; (d) 128 cells, 128 characters, 0.2 dropout; (e) 1024 cells, 16 characters, 0.2 dropout; and (f) 1024 cells, 128 characters, 0.8 dropout. Under high dropout (f), LineageMap shows markedly superior robustness.

again achieves the best RF and Nye scores, recovering substantially more topological structure than any baseline. For path length correlation, LinRace, Cassiopeia and NJ remains relatively similar performances as LineageMap. These results highlight that LineageMap is uniquely effective in low-information settings, preserving stable reconstruction accuracy even when mutations are scarce.

With a high dropout rate as 80%, most mutations are masked which creates another complex inference scenario. LineageMap remains the top performer by a large margin, with the lowest RF distance, highest Nye similarity and the best path length correlation alongside Cassiopeia. However, Cassiopeia, alongside with other methods, fall short in terms of inferring the overall topology. One thing to mention is that we observe a severe performance collapse of Startle from the default settings to the varied settings, and we think that such inconsistency might be due to some fatal error in the program or the model's unable to handle different

barcode qualities. We also performed additional analysis on varying dropout rates (0, 0.2, 0.4, 0.6, 0.8) to characterize the trend of the performances when the barcode data gets increasingly corrupted (Supplementary Fig. 2). Across three metrics, LineageMap is the most consistent and best performing method, especially when the dropout rate are higher, presenting the best robustness (≥ 0.4).

Across all four benchmarking conditions—varying cell count, barcode complexity, and dropout - LineageMap consistently achieves the best or near-best performance on every metric, with notably higher stability across replicates. The method performs particularly strongly in the most challenging scenarios (low target-site count and high dropout), where competing tree-reconstruction algorithms frequently fail to recover meaningful structure. These results show that LineageMap is a robust and accurate approach for lineage reconstruction across a wide range of experimental settings.

3.2 LineageMap reconstructs spatially-resolved lineage tree of baseMEMOIR mESC cell population

baseMEMOIR[3] is a method that combines CRISPR base editing with in situ imaging of engineered "recording" arrays and sequential hybridization readout, allowing high-resolution recording of lineage barcodes in spatial context. Because baseMEMOIR provides spatially-indexed barcode data, it offers a direct testbed for benchmarking the ability of our algorithm (LineageMap) to resolve complex lineages in space, to test polytomy resolution, and to validate spatial-state coupling in inferred trees.

Given the observed lineage barcode, cell state, and spatial location of single-cells on grown mESC dishes, we use LineageMap to infer the lineage tree and the ancestral spatial locations at the same time (Fig. 3a). This spatially resolved lineage tree can illustrate how the population is grown from a single mESC cell at the root to the observed leaf cells. If we compress all generations into a 2D visualization (Fig. 2b-c), we can see that LineageMap can preserve the spatial clustering of each cell type, with naive mESC cells in the middle and formative mESC cells around the boundary of the population. Results on other samples can be found in the supplementary Fig. 3-4.

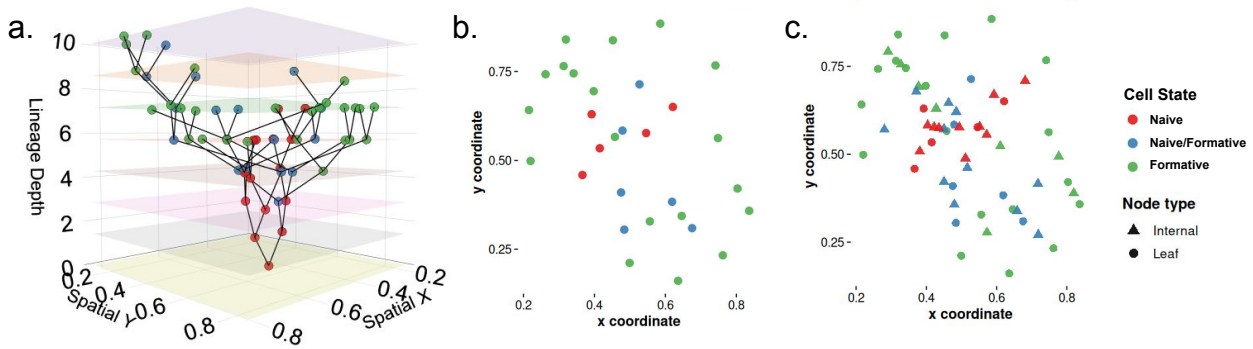

Fig. 3. LineageMap infers spatially-resolved lineage on baseMEMOIR mESC data (colony 2). (a) A ploty 3D visualization of the reconstructed lineage tree projected into spatial and temporal dimensions. X and Y axes represent the spatial coordinates of the cells, while Z axis represents the evolutionary depth from the root, and cells are colored with cell types. (b) 2D visualization of spatial coordinates of observed leaf cells, colored by cell states. (c) 2D visualization of spatial coordinates of both observed leaf cells, and inferred ancestral cells, colored by cell states. The cell states and spatial coordinates of ancestral cells are inferred using LineageMap's likelihood function.

We can also perform quantitative analysis to demonstrate the accuracy of the inferred lineage trees by comparing them to the consensus tree using BEAST2 analysis, as described in the original paper. From the benchmark results in Table 1, we can observe that LineageMap outperforms other methods in lineage tree

accuracy with the three metrics we described before. This result shows the advantage of utilizing complementary spatial and gene expression information to resolve polytomies in the lineage trees.

Table 1. Benchmarking reconstruction accuracy across colonies for all methods and metrics.

	Robinson–Foulds (lower better)					Nye Similarity (higher better)					Path-Length Corr. (higher better)				
Colony	**NJ**	**LM**	**LinRace**	**Cass.**	**Startle**	**NJ**	**LM**	**LinRace**	**Cass.**	**Startle**	**NJ**	**LM**	**LinRace**	**Cass.**	**Startle**
2	0.815	**0.185**	0.815	0.889	0.852	0.528	**0.872**	0.528	0.423	0.497	0.516	**0.939**	0.516	0.435	0.524
4	0.600	**0.467**	0.600	0.933	0.800	0.656	**0.705**	0.656	0.451	0.578	0.589	**0.750**	0.549	0.591	0.743
5	0.583	**0.333**	0.583	0.833	0.750	0.654	**0.810**	0.654	0.500	0.545	**0.923**	0.888	0.895	0.643	0.511

4 Discussion

In this paper, we developed a new computational method, LineageMap, to infer spatially-resolved lineage trees from jointly profiled lineage barcode, gene expression and spatial coordinates data. This design specifically addresses core challenges in lineage reconstruction, including utlizing spatial location modality, high dropout rates, sparse mutation profiles, homoplasy, and barcode-induced polytomies.

LineageMap introduces two main advances: (i) barcode-based community detection combined with a cluster-constrained Neighbor Joining strategy to generate an efficient and robust global topology, and (ii) a unified likelihood model that integrates irreversible barcode mutations, spatial evolution, and cell-state transitions to refine local subtrees and infer ancestral states and locations.

Results on synthetic datasets demonstrate that LineageMap achieves consistent improvements across all major evaluation metrics, particularly in regimes where barcode quality decreases due to factors such as dropouts. Competing methods often show degraded or unstable performance under these challenging conditions, whereas LineageMap maintains both accuracy and robustness—indicating that spatial and transcriptomic information provide meaningful complementary signal for resolving ambiguous lineage relationships.

Analysis of the baseMEMOIR dataset further shows the utility of LineageMap in real experimental settings. The method reconstructs lineage topologies that better agree with consensus phylogenies while inferring spatially coherent ancestral positions and state transitions. This joint inference reveals the correspondence between lineage structure and spatial organization, a property that is inaccessible to barcode-only methods.

Current limitations include simplified spatial models that may not fully capture tissue-specific migration patterns. Future extensions incorporating richer spatial priors, accelerated likelihood evaluation, or continuous latent-state models may expand applicability and scalability.

Overall, LineageMap provides a scalable and statistically principled solution for reconstructing spatially resolved lineage trees from emerging tri-modality single-cell datasets. Although publicly available spatially resolved lineage tracing datasets are currently not widely available, given the increasing number of publications on such data,[11–13,29] methods that integrate lineage, spatial, and cellular-state information will play an essential role in high-resolution analysis of developmental and tissue dynamics in the coming future.

5 DATA AVAILABILITY

LineageMap is available at: https://github.com/ZhangLabGT/LineageMap.
SpaTedSim is a simulator for paired lineage barcode,spatial location and gene expression data available at: https://github.com/ZhangLabGT/SpaTedSim.
Cassiopeia is an end-to-end pipeline for single-cell lineage tracing experiments available at: https://github.com/YosefLab/Cassiopeia.
Startle is a set methods for lineage tree reconstruction by infering the most parsimonious tree under the star homoplasy evolutionary model available at: https://github.com/raphael-group/startle.
LinRace is a R package for lineage tree reconstruction, which can integrate the lineage barcode and gene expression data available at: https://github.com/ZhangLabGT/LinRace

6 FUNDING AND ACKNOWLEDGMENT

This work was supported by National Institutes of Health grant R35GM143070.

References

1. Alemany, A., Florescu, M., Baron, C.S., Peterson-Maduro, J., van Oudenaarden, A.: Whole-organism clone tracing using single-cell sequencing. Nature **556**(7699), 108–112 (Apr 2018). https://doi.org/10.1038/nature25969, https://doi.org/10.1038/nature25969
2. Camin, J.H., Sokal, R.R.: A method for deducing branching sequences in phylogeny. Evolution **19**(3), 311–326 (Sep 1965). https://doi.org/10.1111/j.1558-5646.1965.tb01722.x, http://dx.doi.org/10.1111/j.1558-5646.1965.tb01722.x
3. Chadly, D.M., Frieda, K.L., Gui, C., Klock, L., Tran, M., Sui, M.Y., Takei, Y., Bouckaert, R., Lois, C., Cai, L., Elowitz, M.B.: Reconstructing cell histories in space with image-readable base editor recording. bioRxiv (Jan 2024). https://doi.org/10.1101/2024.01.03.573434, http://dx.doi.org/10.1101/2024.01.03.573434
4. Chan, M.M., Smith, Z.D., Grosswendt, S., Kretzmer, H., Norman, T.M., Adamson, B., Jost, M., Quinn, J.J., Yang, D., Jones, M.G., Khodaverdian, A., Yosef, N., Meissner, A., Weissman, J.S.: Molecular recording of mammalian embryogenesis. Nature **570**(7759), 77–82 (Jun 2019). https://doi.org/10.1038/s41586-019-1184-5, https://doi.org/10.1038/s41586-019-1184-5
5. Evans, S.N., Winter, A.: Subtree prune and regraft: A reversible real tree-valued markov process. The Annals of Probability **34**(3) (May 2006). https://doi.org/10.1214/009117906000000034, http://dx.doi.org/10.1214/009117906000000034
6. Felsenstein, J.: Evolutionary trees from dna sequences: A maximum likelihood approach. Journal of Molecular Evolution **17**(6), 368–376 (Nov 1981). https://doi.org/10.1007/bf01734359, http://dx.doi.org/10.1007/BF01734359
7. Fu, W.Y.: Accelerated high-dimensional global optimization: A particle swarm optimizer incorporating homogeneous learning and autophagy mechanisms. Information Sciences **648**, 119573 (Nov 2023). https://doi.org/10.1016/j.ins.2023.119573, http://dx.doi.org/10.1016/j.ins.2023.119573
8. Gong, W., Granados, A.A., Hu, J., Jones, M.G., Raz, O., Salvador-Martínez, I., Zhang, H., Chow, K.H.K., Kwak, I.Y., Retkute, R., Prusokas, A., Prusokas, A., Khodaverdian, A., Zhang, R., Rao, S., Wang, R., Rennert, P., Saipradeep, V.G., Sivadasan, N., Rao, A., Joseph, T., Srinivasan, R., Peng, J., Han, L., Shang, X., Garry, D.J., Yu, T., Chung, V., Mason, M., Liu, Z., Guan, Y., Yosef, N., Shendure, J., Telford, M.J., Shapiro, E., Elowitz, M.B., Meyer, P.: Benchmarked approaches for reconstruction of in vitro cell lineages and in silico models of c. elegans and m. musculus developmental trees. Cell Systems **12**(8), 810–826.e4 (Aug 2021). https://doi.org/10.1016/j.cels.2021.05.008, https://doi.org/10.1016/j.cels.2021.05.008
9. Heaps, S.E., Nye, T.M., Boys, R.J., Williams, T.A., Embley, T.M.: Bayesian modelling of compositional heterogeneity in molecular phylogenetics. Statistical Applications in Genetics and Molecular Biology **13**(5), 589–609 (Aug 2014). https://doi.org/10.1515/sagmb-2013-0077, http://dx.doi.org/10.1515/sagmb-2013-0077
10. Jones, M.G., Khodaverdian, A., Quinn, J.J., Chan, M.M., Hussmann, J.A., Wang, R., Xu, C., Weissman, J.S., Yosef, N.: Inference of single-cell phylogenies from lineage tracing data using cassiopeia. Genome Biology **21**(1) (Apr 2020). https://doi.org/10.1186/s13059-020-02000-8, https://doi.org/10.1186/s13059-020-02000-8
11. Jones, M.G., Sun, D., Min, K.H.J., Colgan, W.N., Wang, H., Török, T., Cardoso, E.C., Tian, L., Weir, J.A., Chen, V.Z., Koblan, L.W., Yost, K.E., Mathey-Andrews, N., D'Souza, E., Russell, A.J., Stickels, R.R., Balderrama, K.S., Rideout, W.M., Dai, M., Marrero, G., Kumar, V., Saqi, A., Herzberg, B., Izar, B., Chang, H.Y., Lee, J.H., Jacks, T., Chen, F., Weissman, J.S., Yosef, N., Yang, D.: Spatiotemporal lineage tracing reveals the dynamic spatial architecture of tumour growth and metastasis. BioRxiv (Oct 2024). https://doi.org/10.1101/2024.10.21.619529, http://dx.doi.org/10.1101/2024.10.21.619529
12. Koblan, L.W., Yost, K.E., Zheng, P., Colgan, W.N., Jones, M.G., Yang, D., Kumar, A., Sandhu, J., Schnell, A., Sun, D., Ergen, C., Saunders, R.A., Zhuang, X., Allen, W.E., Yosef, N., Weissman, J.S.: High-resolution spatial mapping of cell state and lineage dynamics in vivo with PEtracer. Science **390**(6770), eadx3800 (Oct 2025)
13. Li, K.R., Yu, P.L., Zheng, Q.Q., Wang, X., Fang, X., Li, L.C., Xu, C.R.: Spatiotemporal and genetic cell lineage tracing of endodermal organogenesis at single-cell resolution. Cell **188**(3), 796–813.e24 (Feb 2025). https://doi.org/10.1016/j.cell.2024.12.012, http://dx.doi.org/10.1016/j.cell.2024.12.012
14. Liu, C., Zhou, X., Li, Y., Hittinger, C.T., Pan, R., Huang, J., Chen, X.x., Rokas, A., Chen, Y., Shen, X.X.: The influence of the number of tree searches on maximum likelihood inference in phylogenomics. Systematic Biology **73**(5), 807–822 (Jun 2024). https://doi.org/10.1093/sysbio/syae031, http://dx.doi.org/10.1093/sysbio/syae031
15. Money, D., Whelan, S.: Characterizing the phylogenetic tree-search problem. Systematic Biology **61**(2), 228 (Mar 2012). https://doi.org/10.1093/sysbio/syr097, http://dx.doi.org/10.1093/sysbio/syr097
16. Nye, T.M., Lio, P., Gilks, W.R.: A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. Bioinformatics **22**(1), 117–119 (Oct 2005). https://doi.org/10.1093/bioinformatics/bti720, https://doi.org/10.1093/bioinformatics/bti720

17. Pan, X., Li, H., Putta, P., Zhang, X.: Linrace: cell division history reconstruction of single cells using paired lineage barcode and gene expression data. Nature Communications **14**(1) (Dec 2023). https://doi.org/10.1038/s41467-023-44173-3, http://dx.doi.org/10.1038/s41467-023-44173-3
18. Purvis, A., Garland, T.: Polytomies in comparative analyses of continuous characters. Systematic Biology **42**(4), 569–575 (Dec 1993). https://doi.org/10.1093/sysbio/42.4.569, http://dx.doi.org/10.1093/sysbio/42.4.569
19. Robinson, D., Foulds, L.: Comparison of phylogenetic trees. Mathematical Biosciences **53**(1-2), 131–147 (Feb 1981). https://doi.org/10.1016/0025-5564(81)90043-2, https://doi.org/10.1016/0025-5564(81)90043-2
20. Saitou, N., Nei, N.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution (Jul 1987). https://doi.org/10.1093/oxfordjournals.molbev.a040454, http://dx.doi.org/10.1093/oxfordjournals.molbev.a040454
21. Sashittal, P., Schmidt, H., Chan, M., Raphael, B.J.: Startle: A star homoplasy approach for crispr-cas9 lineage tracing. Cell Systems **14**(12), 1113–1121.e9 (Dec 2023). https://doi.org/10.1016/j.cels.2023.11.005, http://dx.doi.org/10.1016/j.cels.2023.11.005
22. Short, S., García-Tejera, R., Schumacher, L.J., Coutu, D.L.: Next generation lineage tracing and its applications to unravel development. npj Systems Biology and Applications **11**(1) (Jun 2025). https://doi.org/10.1038/s41540-025-00542-w, http://dx.doi.org/10.1038/s41540-025-00542-w
23. Slowinski, J.B.: Molecular polytomies. Molecular Phylogenetics and Evolution **19**(1), 114–120 (Apr 2001). https://doi.org/10.1006/mpev.2000.0897, http://dx.doi.org/10.1006/mpev.2000.0897
24. Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., Junker, J.P.: Simultaneous lineage tracing and cell-type identification using crispr–cas9-induced genetic scars. Nature Biotechnology **36**(5), 469–473 (May 2018). https://doi.org/10.1038/nbt.4124, https://doi.org/10.1038/nbt.4124
25. Togkousidis, A., Kozlov, O.M., Haag, J., Höhler, D., Stamatakis, A.: Adaptive raxml-ng: Accelerating phylogenetic inference under maximum likelihood using dataset difficulty. Molecular Biology and Evolution **40**(10) (Oct 2023). https://doi.org/10.1093/molbev/msad227, http://dx.doi.org/10.1093/molbev/msad227
26. Torres-Montúfar, A., Borsch, T., Ochoterena, H.: When homoplasy is not homoplasy: Dissecting trait evolution by contrasting composite and reductive coding. Systematic Biology **67**(3), 543–551 (Jul 2017). https://doi.org/10.1093/sysbio/syx053, http://dx.doi.org/10.1093/sysbio/syx053
27. VanHorn, S., Morris, S.A.: Next-generation lineage tracing and fate mapping to interrogate development. Developmental Cell **56**(1), 7–21 (Jan 2021). https://doi.org/10.1016/j.devcel.2020.10.021, http://dx.doi.org/10.1016/j.devcel.2020.10.021
28. Wagner, D.E., Klein, A.M.: Lineage tracing meets single-cell omics: opportunities and challenges. Nature Reviews Genetics **21**(7), 410–427 (Jul 2020). https://doi.org/10.1038/s41576-020-0223-2, https://doi.org/10.1038/s41576-020-0223-2
29. Yang, J., Hou, L., Wang, X., Zhang, N., Bian, Y., Lu, Z., Chen, Y., Xie, D., Fang, Y., Wang, K., Wan, R., Jin, Y., Chen, Y., Cai, X., On Lee, L.T., Hu, Z., Ji, H.: Spatiotemporal lineage mapping of tumor immune escape with etracer (Aug 2025). https://doi.org/10.1101/2025.08.06.668639, http://dx.doi.org/10.1101/2025.08.06.668639
30. Zafar, H., Lin, C., Bar-Joseph, Z.: Single-cell lineage tracing by integrating CRISPR-cas9 mutations with transcriptomic data. Nature Communications **11**(1) (Jun 2020). https://doi.org/10.1038/s41467-020-16821-5, https://doi.org/10.1038/s41467-020-16821-5
31. Zou, Y., Zhang, Z., Zeng, Y., Hu, H., Hao, Y., Huang, S., Li, B.: Common methods for phylogenetic tree construction and their implementation in r. Bioengineering **11**(5), 480 (May 2024). https://doi.org/10.3390/bioengineering11050480, http://dx.doi.org/10.3390/bioengineering11050480