VOUS: VARIATIONAL ORNSTEIN-UHLENBECK STOCHASTICS LINKING SINGLE-CELL LINEAGE TRACING WITH DYNAMIC GENE EXPRESSION

Jiawei Xing, Stephen J. Staklinski & Adam C. Siepel
Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory
Cold Spring Harbor, NY 11724, USA
{xing,staklins,asiepel}@cshl.edu

ABSTRACT

Single-cell gene expression evolves dynamically along cell division histories. However, most existing single-cell methods treat cells as static snapshots, neglecting the rich information encoded in their underlying lineage structures. Recent advances in single-cell lineage tracing now enable the reconstruction of high-resolution lineage phylogenies, providing a natural framework pinpoint exactly when and where transcriptional changes occur. This capability is fundamental to decoding the dynamics of development, differentiation, and disease progression. To fully leverage this lineage information, we present VOUS (Variational Ornstein-Uhlenbeck Stochastics), a flexible probabilistic framework that models stochastic single-cell gene expression over inferred cell lineage trees. By grounding gene expression analysis in explicit cell lineage phylogenies with topology and branch lengths, VOUS enables the inference of continuous expression dynamics, despite the high sparsity and low coverage of sequencing data. We applied VOUS to scRNA-seq data from metastatic lung cancers, identifying gene programs associated with metastasis and potential therapeutic targets. By providing a rigorous foundation for modeling sparse count data on latent tree structures, VOUS establishes a generalizable framework that naturally extends to multi-gene programs, lineage uncertainty, and multi-modal integration, paving the way for a comprehensive atlas of single-cell stochastic dynamics.

1 INTRODUCTION

Understanding the precise history of cellular decision-making is fundamental to solving major challenges in biology, from deciphering the origins of chemoresistance in cancer metastasis to mapping the exact branching points of organ development. While single-cell transcriptomics has provided a high-resolution atlas of cell states, it captures only a static snapshot, often obscuring the historical context of dynamic cellular changes. Recent advances in CRISPR/Cas9-based barcoding technology now enable high-resolution single-cell lineage tracing (Wagner & Klein, 2020). Notably, these systems introduce heritable edits to barcodes continuously during cell growth *in vivo* (Salvador-Martínez et al., 2019). Some of these barcodes are transcribed, yielding simultaneous readouts of lineage and gene expression profiles (Raj et al., 2018). Unlike pseudotime trajectories inferred from transcriptomes alone, these genetically recorded trees provide an independent, ground-truth scaffold defined by explicit topology and branch lengths that reflect the actual cell division history. This offers a unique, yet largely unexploited, opportunity to study cell state changes within specific lineages, providing insights into cancer evolution and cell development (Chan et al., 2019; Quinn et al., 2021; Simeonov et al., 2021; Yang et al., 2022; Schiffman et al., 2024).

Following recent progress in lineage tracing, various computational methods have been developed to reconstruct high-resolution cell lineage trees from CRISPR barcodes (Jones et al., 2020; Chu et al., 2025; Staklinski et al., 2025; Siepel et al., 2025). However, integrating these phylogenies with single-cell gene expression remains a significant challenge. Most current approaches rely on statistical tests restricted to endpoint data measured at the leaves of the tree, neglecting the continuous

temporal dynamics embedded in the lineage history. To fully exploit the biological insights within these lineages, we require full generative models that treat gene expression as a dynamic trait evolving continuously from the root to the leaves, giving rise to the sparse, overdispersed read counts observed in single-cell sequencing. In contrast to testing differences at the leaves, this generative perspective enables the integration of evolutionary history, allowing for probabilistic inference of state changes and cellular dynamics.

Stochastic processes provide a natural framework for modeling lineage-based gene expression evolution (Brawand et al., 2011; Chen et al., 2019). In particular, Brownian motion captures the accumulation of small, random transcriptional fluctuations driven by a Wiener process as cells divide along the lineage tree, serving as a principled baseline model for neutral drift of gene expression:

$$dz_t = \sigma dW_t \tag{1}$$

However, Brownian motion alone is insufficient to capture complex biological processes, particularly when cell evolution is subject to adaptive selection driven by environmental signals, treatment pressures, or functional constraints. To address this, the Ornstein-Uhlenbeck (OU) process incorporates a mean-reversion property by modeling a biological restoring force that pulls expression toward an optimal state θ with strength α:

$$dz_t = \alpha(\theta - z_t)dt + \sigma dW_t \tag{2}$$

Based on these stochastic principles, here we introduce VOUS (Variational Ornstein-Uhlenbeck Stochastics), a flexible modeling framework for single-cell dynamics on lineage trees. Unlike most existing scRNA-seq methods that treat cells independently, VOUS explicitly models latent expression dynamics on lineage phylogenies, grounding inference in precise lineage structures provided by advanced lineage tracing technologies. In this work, we focus on a particularly challenging aspect of single-cell data: extreme sparsity, low read counts, and overdispersion. We address this by coupling tree-based OU dynamics with a Negative Binomial observation model and scalable variational inference. As a result, VOUS establishes a rigorous generative framework for lineage-based modeling, laying the foundation for future extensions, such as multi-gene regulation, lineage uncertainty, and multi-modal integration (Baysoy et al., 2023).

2 Related Work

2.1 Modeling expression read counts following the lineage

Inspired by previous tree-based models for species evolution, Pal et al. (2023) developed EvoGeneX that integrates stochastic models with additional Gaussian observation models at tree leaves. This was initially used to model tissue gene expression in Drosophila species, in which Gaussian observation models capture variance across biological replicates from each tissue. This work establishes a precedent for integrating observation models into tree-based stochastic frameworks.

More recently, Hirsch et al. (2025) applied EvoGeneX to single-cell data sampled from cancer clones. While this tree-based stochastic model exhibits higher specificity than differential expression analysis, several limitations hinder its application to general scRNA-seq datasets. For example, the study utilized a dataset with single cells sampled from individual clones. While this ensures that sequencing samples from each clone exhibit relatively similar gene expression patterns, typical tumors are highly heterogeneous, leading to greater variance across samples. This setting also restricts the phylogenetic resolution to the clonal level, whereas standard high-throughput scRNA-seq datasets typically contain $10^4 - 10^6$ cells, with hundreds to thousands of cells in each lineage tree. In addition, this study uses full-length scRNA-seq data from Smart-seq2, which provides a higher sequencing depth with fewer cells. Notably, the discrete count distribution converges to a Gaussian distribution with sufficient sequencing depth, justifying the use of a conjugated Gaussian as the observation model in EvoGeneX. However, typical scRNA-seq data from high-throughput sequencing exhibit high sparsity and skewness consistent with a Negative Binomial distribution, rendering the Gaussian assumption inappropriate (Sarkar & Stephens, 2021).

2.2 Smoothing read count variance across the lineage

During the preparation of this work, Stuart & McKenna (2025) presented SCOUT, another tree-based stochastic framework for single cells. Instead of using an explicit observation model for

single-cell read counts, SCOUT applies a smoothing preprocessing step to the expression data based on the phylogenetic distances on the lineage tree, yielding improved performance relative to a simple log-normal transformation.

However, SCOUT relies on several restrictive assumptions. First, it assumes that phylogenetically adjacent cells share similar transcriptional profiles. However, cells separated by short phylogenetic distances can exhibit divergent states, whereas the smoothing risks conflating distinct cell states and obscuring rapid expression shifts. Moreover, it assumes Gaussian distributions for preprocessed read counts, without explicitly modeling the empirical count distribution. Given the sparsity and overdispersion of scRNA-seq data, gene expression in single cells does not always follow these assumptions, underscoring the necessity for a more rigorous observation model.

2.3 Reconstructing ancestral cell states on the lineage

In addition to the above approaches that model and identify gene expression evolution based on cell lineage, other lineage-based methods focus on gene expression correlations (DeTomaso & Yosef, 2021), trajectory inference (Forrow & Schiebinger, 2021), and ancestral reconstruction (Ouardini et al., 2021). Notably, TreeVAE reconstructs ancestral cell states using a variational autoencoder (VAE). It models the latent cell states as Brownian motion along the tree and approximates the latent posterior from the observation model using mean-field variational inference, offering a precedent for applying variational inference to phylogenetic stochastic modeling. However, reconstructing cell states from lineage history is effective primarily when gene expression is tightly coupled to the tree structure, whereas expression patterns are often intrinsically stochastic and are highly affected by environmental noise, and ancestral signatures may have faded from contemporary observations.

Given the above limitations of current models, here we present VOUS, which employs a Negative Binomial observation model through variational inference, avoiding Gaussian assumptions or heuristic preprocessing. We focus on hypothesis testing for lineage-specific gene expression changes rather than the reconstruction of specific ancestral states. This modeling framework provides a strong foundation for lineage-based single-cell analyses with applications to critical biological questions.

3 Methods

3.1 Model overview

VOUS models single-cell gene expression in three parts (Figure 1):

i. A Gaussian prior modeled by an OU process along the lineage tree $\mathcal{T}$:

$$\boldsymbol{z} \sim \mathrm{OU}(\mathcal{T}, \alpha, \sigma, \boldsymbol{\theta}) \tag{3}$$

ii. A softplus transformation mapping latent states $\boldsymbol{z}$ to expected expression values $\boldsymbol{\lambda}$ at the leaves:

$$\boldsymbol{\lambda} = \zeta(\boldsymbol{z}) = \log(1 + \exp(\boldsymbol{z})) \tag{4}$$

iii. An observation model generating overdispersed single-cell gene expression read counts $\boldsymbol{x}$ following a Negative Binomial likelihood:

$$\boldsymbol{x} \sim \mathrm{NegativeBinomial}(\boldsymbol{\lambda}, r) \tag{5}$$

Finally, the log-likelihood of the observed scRNA-seq read counts is approximated by the evidence lower bound (ELBO) using mean-field variational inference at the leaves:

$$\log p(\boldsymbol{x}) = \mathcal{L}_q + D_{\mathrm{KL}}(q \| p) \geq \mathcal{L}_q \tag{6}$$

3.2 Modeling stochastic gene expression along lineage trees

We model the evolution of latent cell states along a cell lineage tree using the OU stochastic process. Given a rooted lineage tree $\mathcal{T}$ inferred from cellular markers such as CRISPR barcodes, each leaf i of $\mathcal{T}$ represents a cell $c_i \in \mathcal{C}$ with specific meta information n_i. Notably, the total number of nodes exceeds the number of leaves due to the presence of ancestral cells at the internal nodes of $\mathcal{T}$. Therefore, the total cells $\mathcal{C} = \mathcal{C}_I \cup \mathcal{C}_L$, where $\mathcal{C}_I$ and $\mathcal{C}_L$ are cells at internal nodes and leaves of $\mathcal{T}$,

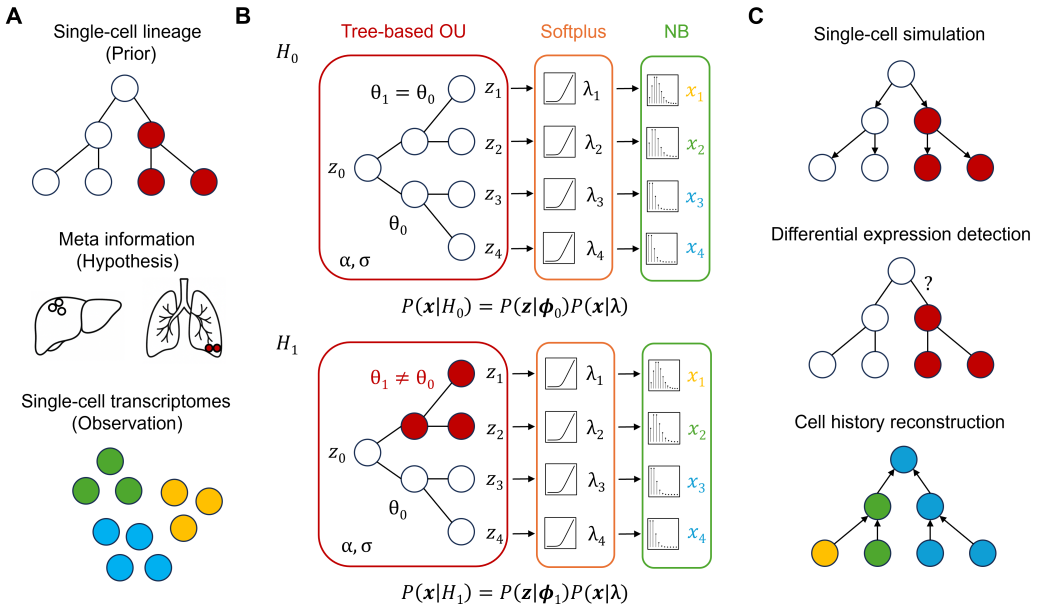

Figure 1: Overview of VOUS. (A) Input data for the model. (B) Model design with hypothesis testing. Variable names are discussed in the main text. (C) Applications of the model.

respectively. If tumor cells in $\mathcal{C}_L$ are labeled by tissues from which they are sampled, the ancestral cell labels for $\mathcal{C}_I$ can be inferred using the Fitch-Hartigan algorithm retrospectively on $\mathcal{T}$ (Jones et al., 2020). Cells that are located along adjacent tree branches and have the same labels n form a regime τ. Cells from τ share a common $\theta_n \in \boldsymbol{\theta}$, which represents the optimal expression level of tumor cells in tissue n. For simplicity, here we consider a uniform variance σ and selective strength α across $\mathcal{T}$ in the OU model, and assign a free θ_n to each tissue n to model tissue-specific tumor gene expressions (Sanghvi et al., 2024).

To model the latent cell state z at each node along $\mathcal{T}$, we start with the initial cell c_0 at the root. By default, we assume that c_0 has reached the optimal expression θ_0 of the initial regime. Alternatively, a Gaussian prior $\mathcal{N}(z_0; \theta_0, s^2)$ can be used for c_0, where s^2 depends on the observed read counts. If $c_i \in \mathcal{C}$ is the direct child of c_0, and the branch connecting c_0 and c_i has a length of t_i, then the latent state of c_i is sampled from a Gaussian according to OU$(\alpha, \boldsymbol{\theta}, \sigma^2)$ from Equation 2:

$$z_i | z_0 \sim \mathcal{N}\left(\theta_0, \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha t_i})\right) \tag{7}$$

where the mean is unchanged, and the variance is approaching the ratio of σ^2 and 2α, indicating a stabilizing selection around θ_0. The same sampling approach can be repeated as long as the cells are still within the same regime. If a cell $c_j \in \mathcal{C}$ falls into a different regime than its immediate parent $c_i \in \mathcal{C}_I$, and the branch connecting c_i and c_j is of length t_j, then the latent state of c_j is sampled as:

$$z_j | z_i \sim \mathcal{N}\left(\theta_1 + (z_i - \theta_1)e^{-\alpha t_j}, \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha t_j})\right) \tag{8}$$

where the mean depends on the parent state z_i and is driven towards the new optimum θ_1 at the new regime, indicating an adaptive selection. The same sampling approach can be repeated along $\mathcal{T}$ until reaching leaves $\mathcal{C}_L$. As a result, the latent cell states at $\mathcal{C}_L$ follow a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{z}; \boldsymbol{W\theta}, \boldsymbol{\Sigma})$, where the mean is the weighted sum of $\boldsymbol{\theta}$ according to regimes along $\mathcal{T}$, and the covariance matrix $\boldsymbol{\Sigma}$ is defined by the location of the most recent common ancestor (MRCA) on $\mathcal{T}$ between each pair of cells in $\mathcal{C}_L$. This gives the OU likelihood for latent states at leaves:

$$P(\boldsymbol{z} \mid \mathcal{T}, \alpha, \sigma^2, \boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^n \det \boldsymbol{\Sigma}}} \exp\left[-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{W\theta})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{z} - \boldsymbol{W\theta})\right] \tag{9}$$

Specifically, the weight matrix $\boldsymbol{W}$ determines the latent state z_i at each leaf $c_i \in \mathcal{C}_L$ by weighting the optimal expression $\theta_n \in \boldsymbol{\theta}$ of each regime τ from the root c_0 to the leaf c_i along $\mathcal{T}$. Each row of $\boldsymbol{W}$ corresponds to a leaf c_i, and each column corresponds to a θ_n. For each element $W_{i,n}$, two sets of binary indicators determine the contribution of θ_n to c_i. The root activator γ_n indicates whether θ_n is active at the root regime, with $\gamma_n = 1$ if θ_n is active and 0 otherwise. t defines the total branch lengths from the root c_0 to leaf c_i, and $e^{-\alpha t}\gamma_n$ describes the exponential decay of the root value z_0. The regime activator $\beta^{\tau}_{i,n}$ indicates whether θ_n is active for c_i in regime τ, with $\beta^{\tau}_{i,n} = 1$ if θ_n is active and 0 otherwise. t^{τ}_i defines the total branch lengths from the root c_0 to the last cell of the regime τ along the path of c_0 to c_i. The difference of exponential decays after τ and $\tau - 1$ describes the mean-reversion towards θ_n within the regime τ:

$$W_{i,n} = e^{-\alpha t}\gamma_n + \sum_{\tau} \left(e^{-\alpha(t - t^{\tau}_i)} - e^{-\alpha(t - t^{\tau-1}_i)} \right) \beta^{\tau}_{i,n} \tag{10}$$

The covariance matrix $\boldsymbol{\Sigma}$ models the correlation structure between leaves of $\mathcal{T}$, which accounts for the shared variance accumulated up to the MRCA and the subsequent independent drift. Each element of $\Sigma_{i,j}$ defines a pair of leaves $c_i, c_j \in \mathcal{C}_L$, with t_i as the total sum of branch lengths from the root c_0 to leaf c_i, and s_{ij} as the total shared branch lengths between c_i and c_j (root to MRCA):

$$\Sigma_{ij} = \frac{\sigma^2}{2\alpha} e^{-\alpha(t_i + t_j - 2s_{ij})}(1 - e^{-2\alpha s_{ij}}) \tag{11}$$

3.3 Modeling single-cell read counts with observation models

The latent cell states $\boldsymbol{z}$ from the tree-based OU process are transformed to $\boldsymbol{\lambda} \in (0, \infty)$ by a softplus function (Equation 4). To model single-cell read counts in scRNA-seq, we present two observation models. The Poisson observation model describes the discrete, sparse read counts across L leaves:

$$P(\boldsymbol{x} \mid \boldsymbol{\lambda}) = \prod_{i=1}^{L} \frac{\lambda_i^{x_i} e^{-\lambda_i}}{x_i!} \tag{12}$$

By default, we use the Negative Binomial model, which better represents single-cell overdispersion with an additional parameter r, with a smaller r corresponding to a larger variance $\lambda + \frac{\lambda^2}{r} > \lambda$. For simplicity, we assume a uniform r across all leaves:

$$P(\boldsymbol{x} \mid \boldsymbol{\lambda}, r) = \prod_{i=1}^{L} \frac{\Gamma(x_i + r)}{\Gamma(r)\, x_i!} \left(\frac{\lambda_i}{r + \lambda_i} \right)^{x_i} \left(\frac{r}{r + \lambda_i} \right)^{r} \tag{13}$$

3.4 Approximating full likelihood with variational inference

The likelihood of observed gene expression read counts given the full VOUS model is:

$$P(\boldsymbol{x} \mid \boldsymbol{\phi}, r) = \int P(\boldsymbol{x} \mid \boldsymbol{\lambda}, r) P(\boldsymbol{z} \mid \boldsymbol{\phi}) d\boldsymbol{z} \tag{14}$$

where $\boldsymbol{\phi} = (\mathcal{T}, \alpha, \sigma, \boldsymbol{\theta})$ contains the tree and the OU parameters. However, this integral is intractable due to unknown latent values $\boldsymbol{z}$ at leaves. Therefore, we applied mean-field variational inference to approximate the latent posterior with a factorized Gaussian distribution:

$$q(\boldsymbol{z}) = \prod_{c \in \mathcal{C}} q_c(z_c) = \prod_{c \in \mathcal{C}} \mathcal{N}(z_c \mid \mu_{qc}, \sigma^2_{qc}) \tag{15}$$

where $\boldsymbol{\mu}_q$ and $\boldsymbol{\sigma}_q$ are chosen to minimize the Kullback–Leibler (KL) divergence to the true posterior:

$$D_{\mathrm{KL}}\big(q(\boldsymbol{z}) \| p_{\boldsymbol{\phi},r}(\boldsymbol{z} \mid \boldsymbol{x})\big) = \mathbb{E}_q\Big[\log q(\boldsymbol{z}) - \log p_{\boldsymbol{\phi},r}(\boldsymbol{z} \mid \boldsymbol{x}) \Big] \tag{16}$$

Replacing the true posterior with Bayes theorem:

$$p_{\boldsymbol{\phi},r}(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{p_{\boldsymbol{\phi}}(\boldsymbol{z}) p_r(\boldsymbol{x} \mid \boldsymbol{z})}{p(\boldsymbol{x})} \tag{17}$$

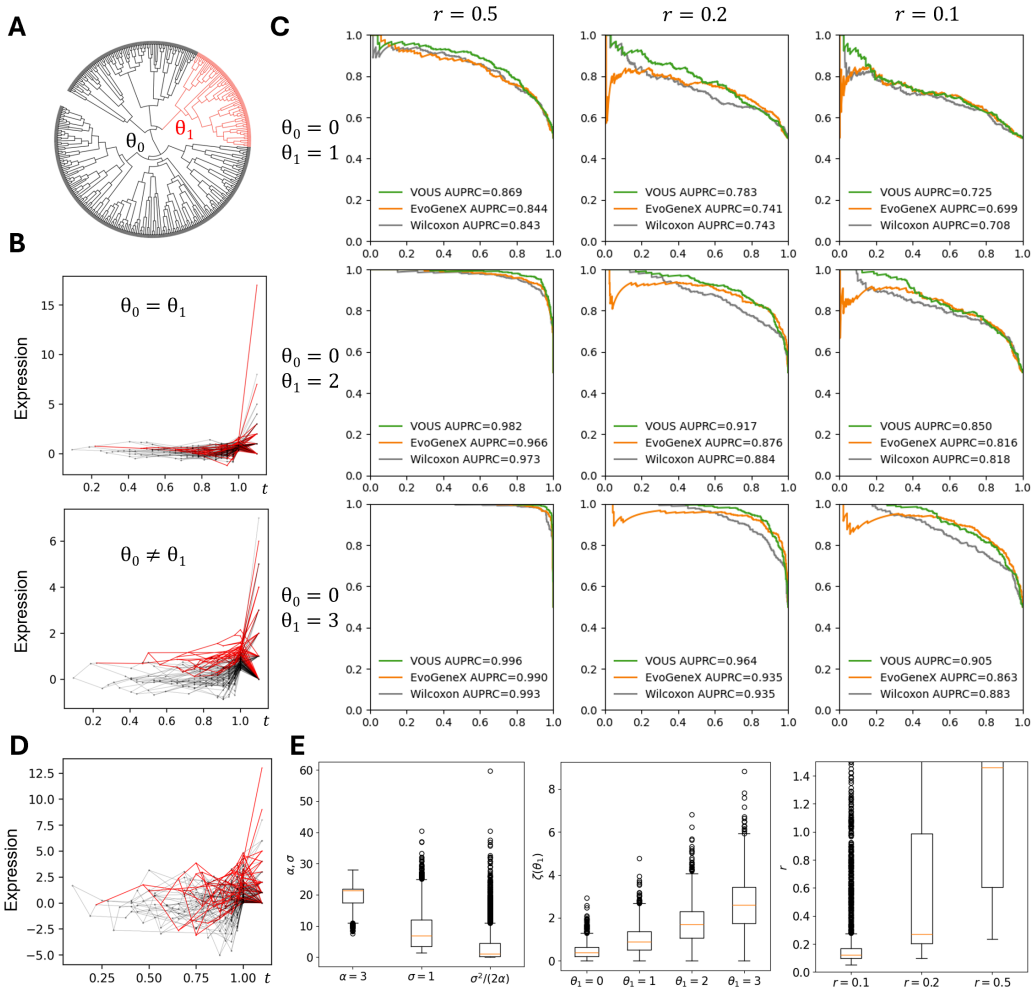

Figure 2: Simulations and validations. (A) Lineage simulation. Red shows metastatic lineage. (B) Examples of negative and positive gene expression simulations ($\alpha = 3, \sigma = 1, r = 0.5, \theta_0 = 0, \theta_1 = 0$ or 1). t, normalized branch length along trees. (C) Precision-recall curves of methods across simulations. (D) Simulation of $\theta_0 = 0, \theta_1 = 1$ using parameters inferred by VOUS ($\alpha = 18, \sigma = 11, r = 3$). (E) All parameters inferred by VOUS.

gives the ELBO for Equation 6:

$$\mathcal{L}_{\boldsymbol{\phi},r}(q) = \mathbb{E}_q[\log p_{\boldsymbol{\phi}}(\boldsymbol{z})] + \mathbb{E}_q[\log p_r(\boldsymbol{x} \mid \boldsymbol{\lambda})] - \mathbb{E}_q[\log q(\boldsymbol{z})] \tag{18}$$

We use this ELBO to approximate the lower bound of the log-likelihood given a model, and update the model parameters to maximize the log-likelihood of the observed scRNA-seq data. By fitting two hypothetical models with a uniform regime or distinct regimes (Figure 1B), we perform a likelihood ratio test for each gene to identify differentially expressed genes in the selected regime:

$$\lambda_{LR} = 2\big[\log p(\boldsymbol{x} \mid \hat{\phi}_1, \hat{r}_1) - \log p(\boldsymbol{x} | \hat{\phi}_0, \hat{r}_0)\big] \tag{19}$$

4 Results

4.1 Validating VOUS on simulation data

Tools that jointly simulate single-cell lineage and gene expression profiles typically lack precise control over expression shifts across specific regimes on the tree (Zhang et al., 2019; Pan et al.,

2022). Therefore, we simulated a lineage tree from an agent-based cancer cell model, and generated differential and non-differential gene expression patterns using VOUS with various parameters.

First, we simulated cancer cell growth and metastasis using the agent-based TTP model adapted from MACHINA (El-Kebir et al., 2018). We sampled 30% of the cells from each simulated clone, yielding a sample cell lineage with 197 cells at the leaves and a single metastasis that forms a new regime (Figure 2A).

Then, we followed the VOUS framework for gene expression simulation, including an OU model along the tree, a softplus transformation to map latent states to positive rates at the leaves, and a Negative Binomial model to generate read counts in single cells. While previous studies typically sweep α and σ to evaluate the model performance (Hirsch et al., 2025; Stuart & McKenna, 2025), here we selected moderate parameters considering the tree length ($\alpha = 3, \sigma = 1$), and focused on a grid search of $\boldsymbol{\theta}$ and r. $\boldsymbol{\theta}$ is biologically meaningful for tissue-specific expression levels, and r is particularly indicative of overdispersion in scRNA-seq, with a smaller r indicating a larger variance. Using these parameters, we simulated 500 negative genes following a uniform regime with θ_0 across the tree, and another 500 positive genes with a different $\theta_1 \neq \theta_0$ in the metastatic regime (Figure 2B). We tested current models for detecting differentially expressed genes, including VOUS, EvoGeneX (Hirsch et al., 2025), and the tree-free Wilcoxon rank-sum test.

According to the areas under precision-recall curves (AUPRC), all methods perform the best under large expression shifts ($\theta_0 = 0, \theta_1 = 3$) and low overdispersions ($r = 0.5$), when the expression signals are strong and clear (Figure 2C). Although improvements are modest, VOUS better distinguishes true differentially expressed genes from random drift under a small r, reflecting an extremely high variance in single-cell read counts. Specifically, both VOUS and EvoGeneX work well with a lower classification threshold, suggesting that the integration of cell lineages helps detect the true expression changes. However, EvoGeneX exhibits a performance drop at the high classification threshold, indicating that the overdispersion of single-cell read counts violates its Gaussian observation assumption, resulting in false positive genes due to high sequencing variances.

Finally, we tested whether VOUS can infer model parameters that capture cellular dynamics. Overall, we observed inflated estimates of α and σ from VOUS across most simulation conditions (Figure 2E), indicating the need for an L2 regularization. However, the ratio $\sigma^2/(2\alpha)$ remains small, and the estimated θs agree well with the ground truth, suggesting that the given data can be explained by different parameter combinations. We further re-simulated $\theta_0 = 0, \theta_1 = 1$ expression using median parameters inferred by VOUS ($\alpha = 18, \sigma = 11, r = 3$) from the $\theta_1 = 1, r = 0.5$ simulation condition (Figure 2D). Although using larger parameter values, this re-simulation yielded a similar read count distribution, suggesting that VOUS makes reasonable inferences on model parameters and cellular dynamics given the observed data. Intriguingly, r is sometimes overestimated, reflecting a lower variance in the Negative Binomial model (Figure 2E). In this case, the data variance may instead be described by the inflated σ of the OU process, indicating non-identifiability or "cross-talk" between model parameters and the complexity of systems dynamics.

4.2 Applying VOUS to metastatic lung cancer

Previous studies have provided a few scRNA-seq datasets with paired cancer cell lineages from CRISPR barcodes (Quinn et al., 2021; Simeonov et al., 2021; Yang et al., 2022). Here we apply VOUS to one of the largest datasets of metastatic lung cancer from Yang et al. (2022). This dataset comprises 9 tumor clones with metastatic events, which are critical for our differential gene expression analysis. We selected the largest metastatic clone as an example, which contains 20,992 cells from three tissues: lung (primary tumor), liver (metastatic tumor), and soft tissue (metastatic tumor). Notably, only 950 cells from this clone have unique barcodes, whereas more than 90% of cells share the same barcodes, rendering their precise lineage relationships unresolvable. Therefore, we preprocessed the cells by selecting a single representative cell per unique barcode, based on a custom score computed for transcriptome qualities.

As a result, we collected 1,453 cells with unique CRISPR barcodes and high-quality transcriptional profiles. We reconstructed a maximum likelihood cell lineage tree for these cells using LAML (Chu et al., 2025), which provides more reliable tree topology and branch lengths compared to the parsimony tree reported in the original paper. The original study selected 4,000 highly variable genes from the scRNA-seq dataset. We removed genes present in fewer than 1% of representative cells

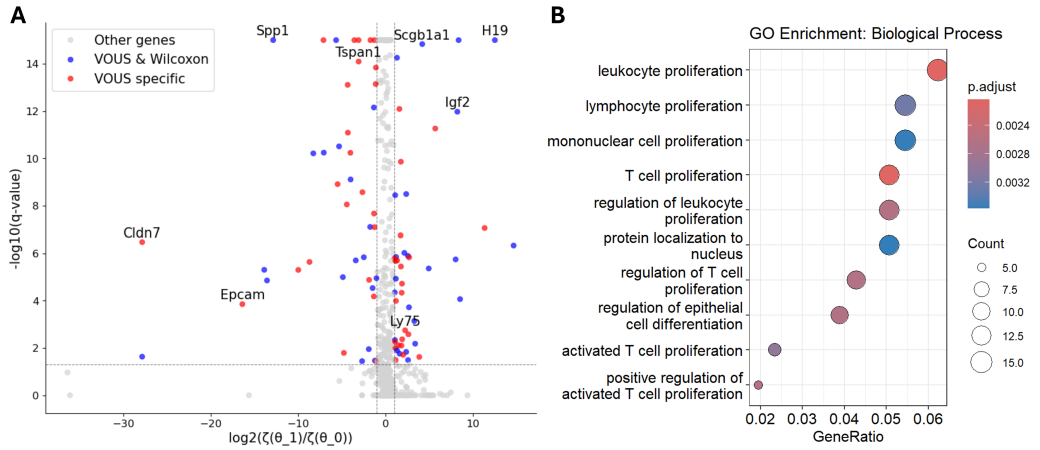

Figure 3: VOUS results on metastatic lung cancer. (A) Differentially expressed genes in metastatic tumors identified by VOUS. (B) GO enrichment analysis of significant genes.

and provided 691 genes as final candidates. We also provided a scaling factor for each cell based on the library size, thereby reducing technical bias arising from sequencing depth.

We applied VOUS to this curated dataset and tested for differentially expressed genes in metastatic tumors relative to the primary tumor. Three genes, *Malat1*, *Sftpc*, and *S100a6*, failed to converge in VOUS, possibly due to high expression and variance. Among the remaining 688 genes, 284 were detected with significant differential expression in the metastatic tumors (q-values < 0.05), of which 129 were upregulated and 155 were downregulated. 91 significant genes have log2 fold changes larger than 1 using the softplus transformed θ inferred from VOUS (Figure 3A). The top significant genes include markers of epithelial-mesenchymal transition (EMT), such as *Spp1* that remodels the immune microenvironment to favor tumor seeding (Xie et al., 2024), *H19* and *Igf2* regulated by the imprinting center (Matouk et al., 2015), and *Scgb1a1* associated with other cancer types (Wang et al., 2024). Strikingly, more than half of these genes were not detected by the Wilcoxon rank-sum test, including *Epcam* and *Cldn7* for cell adhesion (Huang et al., 2018; Philip et al., 2014), *Tspan1* for tissue invasion (Zhou et al., 2023), and *Ly75* for immune response (Bigioni et al., 2017), many of which are validated therapeutic targets.

Finally, we conducted GO enrichment analysis on all significant genes identified by VOUS (Figure 3B). In particular, the term "regulation of epithelial cell differentiation" is enriched among these genes, supporting EMT for tumor metastasis. This demonstrates that VOUS provides biological insights into cancer metastasis by integrating the lineage information and modeling single-cell stochastics. This framework is broadly applicable to other scRNA-seq datasets with paired cell lineages.

5 Conclusion

In this study, we developed VOUS, a probabilistic framework that redefines the analysis of single-cell gene expression by explicitly grounding it in cellular history. While lineage tracing technologies have successfully reconstructed the timing and topology of cell divisions, computational modeling has largely lagged behind, often reducing these rich phylogenetic histories to simple endpoint comparisons. VOUS bridges this gap by modeling gene expression not as a static state, but as a continuous stochastic process evolving along the lineage tree. By coupling a latent OU process with a Negative Binomial observation model, we provide a modeling framework that captures the evolutionary forces driving cellular dynamics while respecting the discrete, sparse, and overdispersed nature of single-cell data. By leveraging variational inference and GPU-accelerated optimization via PyTorch, VOUS scales up to large datasets with thousands of genes and deep lineage trees.

Looking forward, VOUS establishes a robust modeling foundation that naturally extends to future applications. For instance, VOUS supports joint inference across multiple lineage trees or

pre-defined gene sets, amplifying statistical power by integrating information across sparse, noisy sequencing data. Importance sampling can be implemented to move beyond variational approximations toward rigorous Bayes factor computation. Joint inference of lineage topology and gene expression allows transcriptomic data to refine lineage reconstruction (Zafar et al., 2020; Pan et al., 2023). In addition, VOUS is extensible to other modalities, such as the stochastic modeling of epigenetic landscapes and hypothesis testing across distinct spatial regions, opening new avenues for multi-omics integration. Ultimately, by shifting the analytical paradigm from static snapshots to dynamic evolutionary models, VOUS offers a powerful new lens for investigating fundamental biological questions, from defining the precise developmental moments of cell fate commitment to untangling the evolutionary drivers of cancer metastasis and therapeutic resistance.

ACKNOWLEDGMENTS

This work was supported by US National Institutes of Health (NIH) National Institute of General Medical Sciences Grant R35-GM127070 and National Cancer Institute (NCI) Grants R01-CA272466 and 5P30CA045508, as well as Starr Cancer Consortium Grant I16-0060, a National Science Foundation Graduate Research Fellowship (to S.J.S.), a Starr Centennial Scholarship (to S.J.S.), and the Simons Center for Quantitative Biology at CSHL. We thank everyone in the Siepel Lab and Dr. Zhihan Liu from CSHL for helpful discussions.

REFERENCES

Alev Baysoy, Zhiliang Bai, Rahul Satija, and Rong Fan. The technological landscape and applications of single-cell multi-omics. *Nature Reviews Molecular Cell Biology*, 24(10):695–713, 2023.

Mario Bigioni, Giuseppe Merlino, Cristina Bernadó Morales, Rossana Bugianesi, Attilio Crea, Rosanna Manno, Joaquin Arribas, Rachel Dusek, Nickolas Attanasio, Keith Wilson, et al. Men1309, a novel antibody drug conjugate (adc) targeting ly75 antigen, induces complete responses in several xenografts of solid tumors. *Cancer Research*, 77(13_Supplement):2630–2630, 2017.

David Brawand, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, et al. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348, 2011.

Michelle M Chan, Zachary D Smith, Stefanie Grosswendt, Helene Kretzmer, Thomas M Norman, Britt Adamson, Marco Jost, Jeffrey J Quinn, Dian Yang, Matthew G Jones, et al. Molecular recording of mammalian embryogenesis. *Nature*, 570(7759):77–82, 2019.

Jenny Chen, Ross Swofford, Jeremy Johnson, Beryl B Cummings, Noga Rogel, Kerstin Lindblad-Toh, Wilfried Haerty, Federica Di Palma, and Aviv Regev. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome research*, 29(1):53–63, 2019.

Gillian Chu, Uyen Mai, Henri Schmidt, and Benjamin J Raphael. Maximum likelihood inference of time-scaled cell lineage trees with mixed-type missing data using laml. *Genome biology*, 26(1): 189, 2025.

David DeTomaso and Nir Yosef. Hotspot identifies informative gene modules across modalities of single-cell genomics. *Cell systems*, 12(5):446–456, 2021.

Mohammed El-Kebir, Gryte Satas, and Benjamin J Raphael. Inferring parsimonious migration histories for metastatic cancers. *Nature genetics*, 50(5):718–726, 2018.

Aden Forrow and Geoffrey Schiebinger. Lineageot is a unified framework for lineage tracing and trajectory inference. *Nature communications*, 12(1):4940, 2021.

MG Hirsch, Soumitra Pal, Farid Rashidi Mehrabadi, Salem Malikic, Charli Gruen, Antonella Sassano, Eva Pérez-Guijarro, Glenn Merlino, S Cenk Sahinalp, Erin K Molloy, et al. Stochastic modeling of single-cell gene expression adaptation reveals non-genomic contribution to evolution of tumor subclones. *Cell Systems*, 16(1), 2025.

Li Huang, Yanhong Yang, Fei Yang, Shaomin Liu, Ziqin Zhu, Zili Lei, and Jiao Guo. Functions of epcam in physiological processes and diseases. *International journal of molecular medicine*, 42 (4):1771–1785, 2018.

Matthew G Jones, Alex Khodaverdian, Jeffrey J Quinn, Michelle M Chan, Jeffrey A Hussmann, Robert Wang, Chenling Xu, Jonathan S Weissman, and Nir Yosef. Inference of single-cell phylogenies from lineage tracing data using cassiopeia. *Genome biology*, 21(1):92, 2020.

Imad J Matouk, David Halle, Michal Gilon, and Abraham Hochberg. The non-coding rnas of the h19-igf2 imprinted loci: a focus on biological roles and therapeutic potential in lung cancer. *Journal of translational medicine*, 13(1):113, 2015.

Khalil Ouardini, Romain Lopez, Matthew G Jones, Sebastian Prillo, Richard Zhang, Michael I Jordan, and Nir Yosef. Reconstructing unobserved cellular states from paired single-cell lineage tracing and transcriptomics data. *bioRxiv*, pp. 2021–05, 2021.

Soumitra Pal, Brian Oliver, and Teresa M Przytycka. Stochastic modeling of gene expression evolution uncovers tissue-and sex-specific properties of expression evolution in the drosophila genus. *Journal of Computational Biology*, 30(1):21–40, 2023.

Xinhai Pan, Hechen Li, and Xiuwei Zhang. Tedsim: temporal dynamics simulation of single-cell rna sequencing data and cell division history. *Nucleic acids research*, 50(8):4272–4288, 2022.

Xinhai Pan, Hechen Li, Pranav Putta, and Xiuwei Zhang. Linrace: cell division history reconstruction of single cells using paired lineage barcode and gene expression data. *Nature Communications*, 14(1):8388, 2023.

Rahel Philip, Sarah Heiler, Wei Mu, Markus W Büchler, Margot Zöller, and Florian Thuma. Claudin-7 promotes the epithelial–mesenchymal transition in human colorectal cancer. *Oncotarget*, 6(4):2046, 2014.

Jeffrey J Quinn, Matthew G Jones, Ross A Okimoto, Shigeki Nanjo, Michelle M Chan, Nir Yosef, Trever G Bivona, and Jonathan S Weissman. Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science*, 371(6532):eabc1944, 2021.

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

Irepan Salvador-Martínez, Marco Grillo, Michalis Averof, and Maximilian J Telford. Is it possible to reconstruct an accurate cell lineage using crispr recorders? *Elife*, 8:e40292, 2019.

Neel Sanghvi, Camilo Calvo-Alcañiz, Padma S Rajagopal, Stefano Scalera, Valeria Canu, Sanju Sinha, Fiorella Schischlik, Kun Wang, Sanna Madan, Eldad Shulman, et al. Charting the transcriptomic landscape of primary and metastatic cancers in relation to their origin and target normal tissues. *Science Advances*, 10(49):eadn0220, 2024.

Abhishek Sarkar and Matthew Stephens. Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis. *Nature genetics*, 53(6):770–777, 2021.

Joshua S Schiffman, Andrew R D'Avino, Tamara Prieto, Yakun Pang, Yilin Fan, Srinivas Rajagopalan, Catherine Potenski, Toshiro Hara, Mario L Suvà, Charles Gawad, et al. Defining heritability, plasticity, and transition dynamics of cellular phenotypes in somatic evolution. *Nature genetics*, 56(10):2174–2184, 2024.

Adam Siepel, Rebecca Hassett, and Stephen J Staklinski. Variational inference with node embeddings (vine) for scalable bayesian phylogenetics. *bioRxiv*, pp. 2025–12, 2025.

Kamen P Simeonov, China N Byrns, Megan L Clark, Robert J Norgard, Beth Martin, Ben Z Stanger, Jay Shendure, Aaron McKenna, and Christopher J Lengner. Single-cell lineage tracing of metastatic cancer reveals selection of hybrid emt states. *Cancer cell*, 39(8):1150–1162, 2021.

Stephen J Staklinski, Armin Scheben, Lise Brault, Rebecca Hassett, Ryan Serio, Jiawei Xing, Dawid G Nowak, and Adam Siepel. Bayesian inference of tissue-migration histories in metastatic cancer from cell-lineage tracing data. *bioRxiv*, pp. 2025–09, 2025.

Hannah Stuart and Aaron McKenna. Scout: Ornstein–uhlenbeck modelling of gene expression evolution on single-cell lineage trees. *bioRxiv*, pp. 2025–11, 2025.

Daniel E Wagner and Allon M Klein. Lineage tracing meets single-cell omics: opportunities and challenges. *Nature Reviews Genetics*, 21(7):410–427, 2020.

Jing Wang, Qianqian Xu, Jiangbo Yu, Aotian Xu, Lizheng Yu, Zhenggang Chen, Yang Cao, Rongtao Yuan, and Zhongjie Yu. Scgb1a1 as a novel biomarker and promising therapeutic target for the management of hnscc. *Oncology Letters*, 28(5):527, 2024.

Sun-Zhe Xie, Lu-Yu Yang, Ran Wei, Xiao-Tian Shen, Jun-Jie Pan, Shi-Zhe Yu, Chen Zhang, Hao Xu, Jian-Feng Xu, Xin Zheng, et al. Targeting spp1-orchestrated neutrophil extracellular traps-dominant pre-metastatic niche reduced hcc lung metastasis. *Experimental hematology & oncology*, 13(1):111, 2024.

Dian Yang, Matthew G Jones, Santiago Naranjo, William M Rideout, Kyung Hoi Joseph Min, Raymond Ho, Wei Wu, Joseph M Replogle, Jennifer L Page, Jeffrey J Quinn, et al. Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution. *Cell*, 185(11):1905–1923, 2022.

Hamim Zafar, Chieh Lin, and Ziv Bar-Joseph. Single-cell lineage tracing by integrating crispr-cas9 mutations with transcriptomic data. *Nature communications*, 11(1):3055, 2020.

Xiuwei Zhang, Chenling Xu, and Nir Yosef. Simulating multiple faceted variability in single cell rna sequencing. *Nature communications*, 10(1):2611, 2019.

Zhihang Zhou, Zihan Yang, Li Zhou, Mengsu Yang, and Song He. The versatile roles of testrapanins in cancer from intracellular signaling to cell–cell communication: cell membrane proteins without ligands. *Cell & Bioscience*, 13(1):59, 2023.

A OBJECTIVE FUNCTION

From Equation 18, the ELBO is composed of expectations of the OU log-likelihood and the observation log-likelihood with respect to $q(\boldsymbol{z})$, as well as the entropy of $q(\boldsymbol{z})$.

A.1 KARUSH-KUHN-TUCKER CONDITION

From Equation 9, the log-likelihood of the latent expression $\boldsymbol{z}$ given the OU model is:

$$\log p_{\boldsymbol{\phi}}(\boldsymbol{z}) = -\frac{1}{2}(\boldsymbol{z} - \boldsymbol{W\theta})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{z} - \boldsymbol{W\theta}) - \frac{1}{2}\log\det\boldsymbol{\Sigma} + \text{const.} \tag{20}$$

where const. denotes terms independent of the model parameters and $\boldsymbol{\phi} = (\mathcal{T}, \alpha, \sigma, \boldsymbol{\theta})$ contains the tree and the OU parameters. Therefore, the expectation of the OU log-likelihood with respect to $q(\boldsymbol{z})$ is:

$$\mathbb{E}_q[\log p_{\boldsymbol{\phi}}(\boldsymbol{z})] = -\frac{1}{2}\left[(\boldsymbol{\mu}_q - \boldsymbol{W\theta})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_q - \boldsymbol{W\theta}) + \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathrm{diag}(\boldsymbol{\sigma}_q))\right] - \frac{1}{2}\log\det\boldsymbol{\Sigma} + \text{const.} \tag{21}$$

which contains an additional trace term arising from the expectation of the quadratic form. To reduce the number of free parameters and improve calculation efficiency, we apply the Karush-Kuhn-Tucker (KKT) conditions adapted from Pal et al. (2023). When $\mathcal{L}_q$ is maximized at an optimal solution ($\hat{\alpha}$, $\hat{\sigma}$, $\hat{\boldsymbol{\theta}}$), the partial derivatives of $\mathcal{L}_q$ with respect to these parameters should be 0. By setting $\partial\mathcal{L}_q/\partial\boldsymbol{\theta} = 0$, we estimate $\hat{\boldsymbol{\theta}}$ as:

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{W}^\top \tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W})^{-1}\boldsymbol{W}^\top \tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}_q \tag{22}$$

where $\tilde{\boldsymbol{\Sigma}} = \frac{1}{\sigma^2}\boldsymbol{\Sigma}$. By setting $\partial \mathcal{L}_q / \partial \sigma = 0$, we estimate $\hat{\sigma}^2$ as:

$$\hat{\sigma}^2 = \frac{1}{n}\left[(\boldsymbol{\mu}_q - \boldsymbol{W}\hat{\boldsymbol{\theta}})^\top \tilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_q - \boldsymbol{W}\hat{\boldsymbol{\theta}}) + \mathrm{Tr}(\tilde{\boldsymbol{\Sigma}}^{-1}\mathrm{diag}(\boldsymbol{\sigma}_q))\right] \tag{23}$$

By substituting Equation 23 into Equation 21, the new expectation becomes:

$$\mathbb{E}_q[\log p_{\boldsymbol{\phi}}(\boldsymbol{z})] = -\frac{1}{2}(\log\det\tilde{\boldsymbol{\Sigma}} + n\log\hat{\sigma}^2) + \mathrm{const.} \tag{24}$$

which essentially depends on $q(\boldsymbol{z})$ and α.

A.2 Regularization

In experiments, we noticed that α tends to diverge without strictly improving the ELBO. Intuitively, by considering Equation 11, $\log\det\tilde{\boldsymbol{\Sigma}}$ from Equation 24 can be approximated as $-n\log 2\alpha$. By considering Equations 23 and 11, $n\log\hat{\sigma}^2$ from Equation 24 can be approximated as $n\log(2\alpha/n) = n\log 2\alpha + \mathrm{const}$. As a result, the two terms approximately cancel out, resulting in a flat loss landscape that promotes overfitting. Therefore, we add an L2 regularization for α. This is equivalent to a Gaussian prior $\mathcal{N} \sim (\alpha; 0, \beta)$, where β is controllable by the users.

A.3 Approximation

With the Poisson observation model from Equation 12, the ELBO becomes:

$$\mathcal{L}_q = -\frac{1}{2}(\log\det\tilde{\boldsymbol{\Sigma}} + n\log\hat{\sigma}^2) + \sum_{i=1}^{L}\Big[x_i\mathbb{E}_q[\log\lambda_i] - \mathbb{E}_q[\lambda_i]\Big] + \frac{1}{2}\sum_{i=1}^{L}\log\sigma_{qi}^2 + \mathrm{const.} \tag{25}$$

Similarly, the Negative Binomial observation model from Equation 13 gives the ELBO as:

$$\begin{aligned}\mathcal{L}_q = &-\frac{1}{2}(\log\det\tilde{\boldsymbol{\Sigma}} + n\log\hat{\sigma}^2) + \sum_{i=1}^{L}\Big[\log\Gamma(x_i+r) - \log\Gamma(r) + r\log r + x_i\mathbb{E}_q[\log\lambda_i] \\ &- (x_i+r)\mathbb{E}_q[\log(r+\lambda_i)]\Big] + \frac{1}{2}\sum_{i=1}^{L}\log\sigma_{qi}^2 + \mathrm{const.}\end{aligned} \tag{26}$$

where the expression $\lambda \in (0, \inf)$ was transformed from the latent z. We tested two transformations: exponential $\lambda = e^z$ and softplus $\lambda = \zeta(z) = \log(1+e^z)$. Although the exponential transformation simplifies analytic calculation, we observe a better performance with softplus due to a relatively linear relationship for large z.

To approximate $\mathbb{E}_q[\lambda] = \mathbb{E}_q[\zeta(z)]$, $\mathbb{E}_q[\log\lambda] = \mathbb{E}_q[\log\zeta(z)]$, and $\mathbb{E}_q[\log(r+\lambda)] = \mathbb{E}_q[\log(r+\zeta(z))]$, we use Monte Carlo simulation by default. Alternatively, we tested Taylor series expansion. For $\mathbb{E}_q[\zeta(z)]$:

$$\mathbb{E}_q[\zeta(z)] \approx \begin{cases}\zeta(\mu_q) + \frac{\sigma_q^2}{2}\sigma(\mu_q)(1-\sigma(\mu_q)) & \text{if } \mu_q < 5, \\ \mu_q & \text{if } \mu_q \geq 5\end{cases} \tag{27}$$

where $\sigma(\mu_q) = (1+e^{-\mu_q})^{-1}$ denotes the sigmoid function. For $\mathbb{E}_q[\log\zeta(z)]$:

$$\mathbb{E}_q[\log\zeta(z)] \approx \begin{cases}(1-\omega)\mathcal{T}_0 + \omega\mu_q & \text{if } \mu_q < 2, \\ \log(\mu_q + e^{-\mu_q}) - \frac{(1-e^{-\mu_q})^2}{2(\mu_q+e^{-\mu_q})^2}\sigma_q^2 & \text{if } 2 \leq \mu_q < 10, \\ \log\mu_q & \text{if } \mu_q \geq 10\end{cases} \tag{28}$$

where $\omega = (1+e^{2(\mu_q+2)})^{-1}$ is a gating factor, and $\mathcal{T}_0$ is the Taylor expansion around $z=0$:

$$\mathcal{T}_0 = \log(\log 2) + \frac{\mu_q}{2\log 2} + \frac{(\log 2 - 1)(\mu_q^2 + \sigma_q^2)}{8(\log 2)^2}. \tag{29}$$

B MODEL IMPLEMENTATION

We implemented the model using PyTorch, leveraging vectorized optimization to process multiple genes in parallel batches, which significantly accelerates computation on GPU hardware.

We use the PyTorch Adam optimizer for gene-wise optimization. To assess model convergence, we compute the relative decrease in loss for each gene within a window during iterations. Genes with loss decreases within the tolerance are frozen and excluded from the subsequent optimization iterations until all the genes are converged or iterations are above the maximum. Parameters such as the learning rate, initial values, window size, tolerance, and maximal iterations are modifiable by users. Loss trajectories during optimization can be monitored via Weights & Biases.

Furthermore, we implement an EM algorithm for model optimization. In the E step, the model parameters are fixed, whereas the $q(\boldsymbol{z})$ means $\boldsymbol{\mu}_q$ and variances $\boldsymbol{\sigma}_q^2$ from variational inference are optimized. In the M step, the variational parameters are fixed, and the model parameters from OU and Negative Binomial are optimized. By default, we optimize all parameters simultaneously, which makes the convergence more efficient.

For likelihood ratio test, we test significance based on a χ^2 distribution. The p-values are adjusted by false discovery rate using the Benjamini-Hochberg procedure. We also provide options to simulate empirical null distributions using inferred model parameters from the null hypothesis. These simulations can be performed either using gene-specific model parameters, or using sampled model parameters from all genes. However, we did not observe obvious improvements using empirical null distributions.

Finally, we extend the model to include multiple input files. For example, multiple tree files from the same tumor and multiple gene expression files within the same gene set can be combined to an average likelihood by the model for information integration.

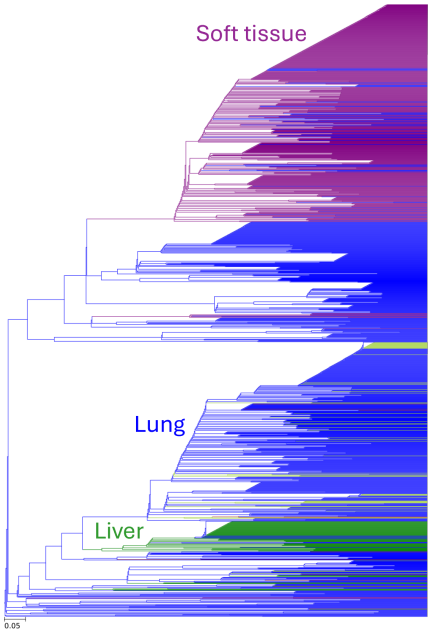

Figure S1: Lineage tree of the largest clone in the Yang et al. (2022) dataset built with LAML. Colors show tissue labels.