

Lineage-informed factor analysis reveals heritable programs of single-cell gene expression

Stephen J. Staklinski¹ and Adam Siepel^{1,*}

¹Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

*Corresponding author: asiepel@cshl.edu

Abstract

Single-cell transcriptomics has transformed our ability to characterize cellular identity, but present-day gene expression profiles capture only a snapshot of a process that unfolds across cell division history. Recent single-cell lineage-tracing technologies make it possible to reconstruct cell division histories for thousands of cells, opening a window into how gene expression evolves. Yet observed gene expression is often redundant, with correlations among genes reflecting underlying latent biological programs and regulatory networks. To capture this structure, we introduce scPFA, a single-cell phylogenetic factor analysis framework that represents gene expression through a small set of latent factors that evolve along lineages under a phylogenetic prior, capturing correlated structure hidden in present-day observations alone. Simulations demonstrate accurate recovery of latent factors and covariance structure across conditions. Applied to developmental and cancer lineage-tracing datasets, the model uncovers biologically interpretable, lineage-associated expression programs. Together, these results demonstrate how lineage-informed factor analysis can reveal temporal biological structure hidden within high-dimensional single-cell data.

Introduction

Gene expression is a key way in which cellular identity, differentiation, and disease progression are measurable. A cell's transcriptional state reflects its type, its environment, its response to cell-cell interactions or signaling within that environment, and the regulatory programs active within it. Single-cell transcriptomics has made this variation visible at enormous scale [1,2], showing that tissues and tumors are not only collections of discrete cell types, but also mixtures of continuously varying molecular states [3–6]. This heterogeneity is central to many biological processes, from development and tissue homeostasis to cancer progression and therapeutic response [7–10]. Yet a transcriptional profile remains a snapshot. It describes the state of a cell at the moment it was measured, while the processes that produced that state unfolded through cell division, differentiation, migration, and regulatory change over time [5, 11, 12].

Much of recent single-cell analysis can be viewed as an effort to recover some of this missing temporal context [13]. Trajectory inference methods arrange cells along putative developmental paths, often defined

in terms of a pseudotime, by using similarities in their observed transcriptional profiles, whereas RNA velocity uses a biophysical model of local RNA splicing dynamics, based on nascent and mature transcript abundances, to infer the likely short-term direction of movement through transcriptional space [7, 11, 14–16]. These methods have provided powerful ways to reason about dynamic processes from static measurements, but they do not directly consider the ancestry of the cells being analyzed and they often rely on strong assumptions about the biological process being studied [12, 17, 18]. Explicit lineage-tracing efforts add a different kind of temporal information by capturing a record of cell division history through artificial barcodes, naturally occurring somatic mutations, or other markers which are claimed to be heritable [19–24]. When coupled with single-cell RNA sequencing, these technologies make it possible to observe a cell’s present-day gene expression together with a partial record of the history that produced it [12, 25].

Lineage-resolved transcriptomic data naturally invite models of gene expression on lineage trees. This follows from the covariance structure of the data, a point made early in comparative methods for phylogenetically structured data [26]. In organismal biology, stochastic processes defined along phylogenies have long been used to model the evolution of continuous quantitative traits, including gene expression, while accounting for shared evolutionary history [26–36]. Similar ideas are now being adapted to cell-lineage data to study transcriptional heritability and reconstruct changes in gene expression across time [37–40]. These approaches suggest that lineage information can reveal structure that is difficult to identify from present-day measurements alone.

One reason this hidden temporal structure may be biologically meaningful is that cellular phenotypes are rarely controlled by isolated genes acting alone. They are shaped by coordinated transcriptional programs, shared regulatory mechanisms, signaling pathways, and latent biological processes that affect many genes at once over time [41–44]. As a result, the field often assumes that the high-dimensional and noisy variation observed across thousands of genes often has lower-dimensional structure [45–48]. If these latent programs are themselves inherited, altered, or constrained along cell lineages, then treating genes independently may miss the level at which transcriptional heritability is well organized. The relevant object may be the inheritance of gene modules and the cellular states they define, not only the inheritance of individual genes [49, 50].

Single-gene models are therefore limited in their ability to represent the correlation structure among genes. One idea has been to use network-based module clustering analyses to identify heritable transcriptional signals carried by groups of genes [51]. This approach reveals gene modules, but it does not directly model how individual cells are represented by that heritable module structure. Answering that question would require post hoc integration with a separate method for learning latent representations of single cells [46] or a gene program scoring approach that assigns a score to each module in every cell [52]. Conversely, several methods learn lineage-informed embeddings or trajectories of cells, but do so without explicitly modeling the interpretable gene modules that give rise to that latent structure [53–58]. Similarly, active work has uncovered the heritability of predefined discrete cellular states along lineage trees, but without learning what gene modules those states represent within the same model [59–64]. Together, these approaches reflect substantial interest in lineage-informed gene expression structure and heritability, but each captures only one part of what appears to be a connected problem of discovering heritable gene

modules while also modeling the historical representation of cells in terms of those modules. This idea is related to recent work extending RNA velocity to consider gene regulatory networks [65], but our goal is to use cell-lineage information to learn a global representation of temporal gene expression structure rather than local transcriptional dynamics.

Factor analysis provides a natural language for this problem. Originally developed in the social sciences to explain correlations among measured variables using a smaller number of unobserved factors, factor analysis provides a rich general statistical framework for learning latent low-dimensional structure in multivariate data [66]. The key to factor analysis is that it not only learns the loading matrix defining the composition of underlying latent factors, but also learns a factor matrix of those factors that provides a latent representation for each individual in the observed data. Factor analysis is not new in cell biology. In single-cell genomics, factor analysis has been adapted in various ways for deconvolving bulk gene expression data into single-cell profiles [67]. A version of factor analysis has also been recently adapted for learning cellular ecosystems from single-cell transcriptomics data [68]. In parallel, recent work in evolutionary biology has begun to combine factor analysis of phenotypic traits with phylogenetic models to account for shared ancestral relationships [69, 70]. These efforts are part of a broader literature showing the value of learning low-dimensional structure in high-dimensional correlated traits evolving along phylogenies [71, 72]. Together, they suggest that factor analysis models are a natural fit for lineage-resolved single-cell gene expression, although, to the best of our knowledge, they have not yet been applied in this setting.

Here, we introduce scPFA (single-cell Phylogenetic Factor Analysis), a phylogenetic factor analysis framework for lineage-resolved single-cell gene expression. The model represents transcriptional variation through a smaller set of latent factors, fits stochastic processes for those factors along a cell-lineage tree, and learns the gene modules associated with each factor in a unified probabilistic framework. Using simulated and experimental lineage-tracing datasets, we work through several technical considerations and demonstrate that this approach can recover latent programs representing heritable modules of coordinated gene expression. In experimental data, the inferred factors capture biologically coherent programs that are reproducible across clones within the same system but distinct across different systems. These factors can also integrate modules identified as distinct by existing clustering methods into a single heritable axis of variation when they share similar evolutionary trajectories.

Results

A phylogenetic factor analysis model of heritable structure in single-cell gene expression

The goal of scPFA is to represent lineage-resolved gene expression at the level at which we expect much of the biology to be organized. We do not treat genes as independent, and we do not aim only to learn a minimally interpretable embedding of cells. Instead, we represent expression through interpretable latent transcriptional programs whose activities vary across the cell lineage. This framework is inspired by previ-

ous work on learning low-dimensional representations of single-cell gene expression [46] and on modeling the evolution of those representations along cell-lineage trees [54]. The distinction here is that the factors remain tied directly to genes, so that each latent axis can be interpreted as both a gene module and a heritable cellular state variable (**Fig. 1A**).

As a first formulation, we work with a log-transformed matrix of normalized expression values, $\mathbf{X} \in \mathbb{R}^{n \times p}$, for n cells and p genes, together with a cell-lineage tree T . We treat this matrix as a continuous trait matrix under a Gaussian observation model, and restrict the analysis to genes with evidence of phylogenetic signal. This does not model the unique observational properties of single-cell count data, as we have done in related prior work [40], but instead focuses the problem more generally on whether utilizing lineage information in a factor analysis framework helps recover interpretable low-dimensional structure in gene expression.

Observed expression is modeled as a noisy low-rank reconstruction, with $\mathbf{X} = \mathbf{Z} + \mathbf{E}$ and $\mathbf{Z} = \mathbf{F}\mathbf{L}$. Here $\mathbf{F} \in \mathbb{R}^{n \times k}$ is a cell-by-factor matrix, $\mathbf{L} \in \mathbb{R}^{k \times p}$ is a factor-by-gene loading matrix, and $\mathbf{E}$ represents residual noise, with $\mathbf{X}_{i \cdot} \sim \text{MVN}(\mathbf{Z}_{i \cdot}, \sigma_{\text{obs}}^2 \mathbf{I})$. The rows of $\mathbf{L}$ define gene modules, while the columns of $\mathbf{F}$ describe the activity of those modules across cells.

The lineage enters through a Brownian motion prior on the columns of $\mathbf{F}$. For each factor d , we place an $\text{MVN}(\mathbf{0}, \sigma_d^2 \mathbf{\Sigma})$ prior on $\mathbf{F}_{\cdot d}$, where $\mathbf{\Sigma}$ is the covariance matrix induced by the lineage tree and σ_d^2 is a factor-specific Brownian variance parameter learned from the data. This prior makes factor activities more similar among cells with more shared history, while allowing divergence to accumulate along longer or independent branches. The variance parameters determine the extent to which each latent factor varies across the lineage, allowing different transcriptional programs to evolve at different rates. More complex evolutionary models could be used, but Brownian motion provides a simple and interpretable baseline, avoids additional process parameters, and is consistent with cautionary work showing that parameter-rich alternatives such as Ornstein–Uhlenbeck models do not necessarily improve inference in practice [73].

The inferred quantities are $\mathbf{F}$, $\mathbf{L}$, the factor-specific Brownian variance parameters $\sigma_1^2, \dots, \sigma_k^2$, and σ_{obs}^2 (blue variables in **Fig. 1A**). The main biological output is $\mathbf{L}$, which defines gene modules associated with each latent factor (**Fig. 1B**). Factors can be annotated by ranking genes according to their positive and negative loadings and performing gene set enrichment analysis (**Fig. 1C**). The inferred matrix $\mathbf{F}$ can also be used to reconstruct each latent program across cellular history (**Fig. 1D**). As in any factor analysis model, the factorization has scale, sign, and permutation invariances. In the empirical work below on simulated data, we introduce the use of constrained $\mathbf{L}$ rows and post hoc rules as sufficient means for resolving these invariances.

Identifying phylogenetically informative genes prior to factor analysis

Matched lineage-tracing and transcriptomic data make it possible to ask which genes vary in a way that is consistent with shared lineage history. Because our model is intended to learn heritable gene modules, we first filtered to genes with evidence of phylogenetic signal. This is a simple approximation to a more general

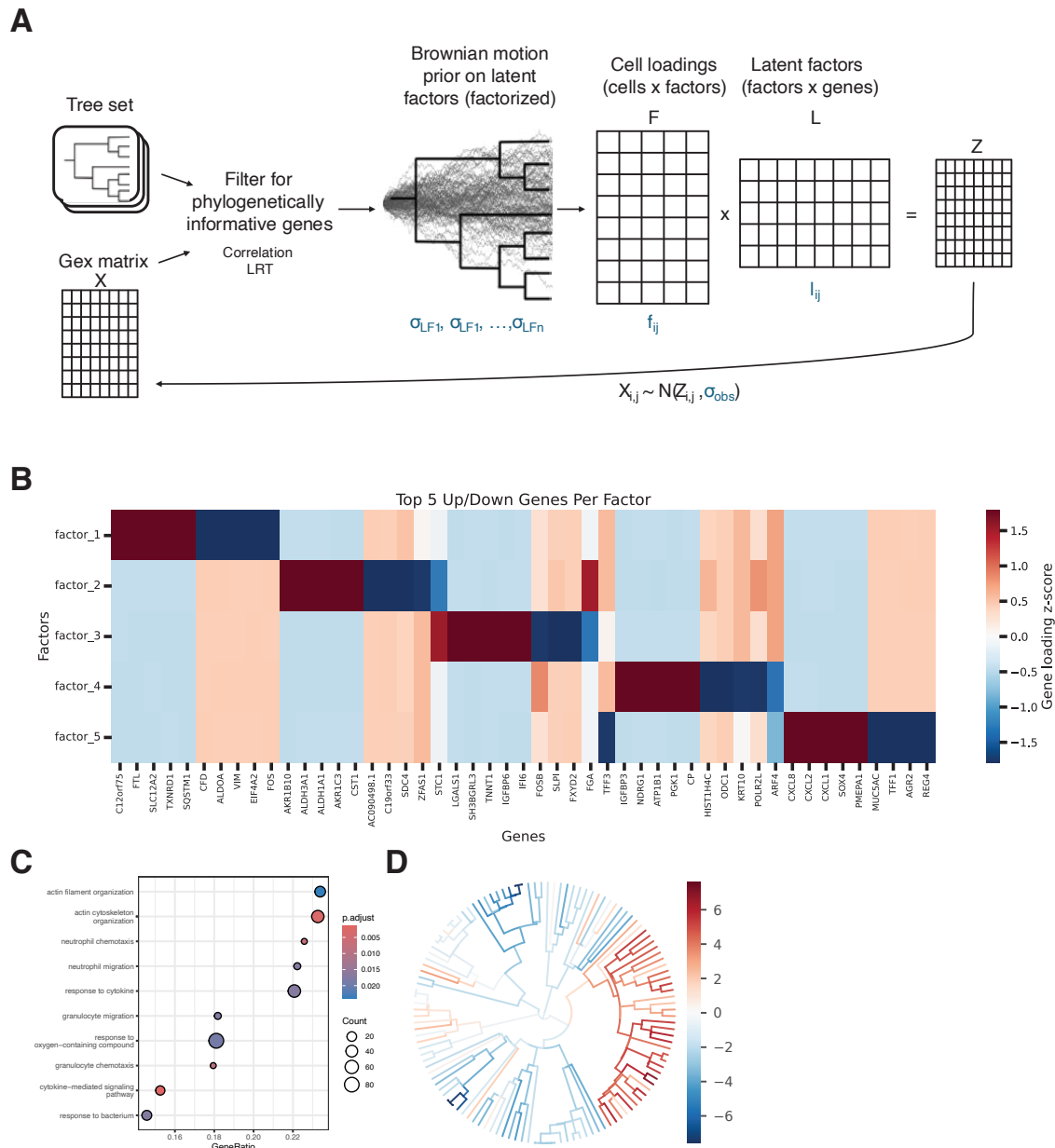


Figure 1: Overview of scPFA and representative visualizations.

(A) The scPFA model. Genes are first preprocessed to identify those with phylogenetic signal, and then latent factors are modeled on the tree using independent Brownian motion priors, yielding a cell-factor matrix F and factor-gene loading matrix L , whose product reconstructs gene expression. Blue variables denote model parameters inferred during fitting, whereas black variables are fixed. (B) Example factor-gene loading matrix L , shown only for the highest positive and negative gene loadings for each factor. (C) Gene set enrichment analysis of genes associated with one inferred factor, illustrating biological interpretation of latent gene programs. (D) Reconstruction of a latent factor across the lineage tree using the inferred cell-factor matrix F , enabling visualization of temporal dynamics along cellular histories.

model in which heritable and non-heritable components would be learned jointly.

We compared PATH, a recently proposed correlation-based filter [61], with two likelihood-ratio tests based on Brownian motion covariance models. The first LRT used an empirically estimated null distribution, whereas the second used a computationally efficient Pagel’s λ parameterization [74, 75] like that which we have used in prior work [40]. We also re-implemented PATH in C for convenience within our workflow, which gave numerically identical results to the original R implementation but ran faster.

In simulations with known heritable genes, PATH was precise but less sensitive than either LRT (**Fig. 2A**). The LRTs recovered nearly all simulated heritable genes, as expected under data generated from Brownian motion, but differed in runtime. The empirical-null LRT became expensive as dataset size increased, whereas the λ -LRT gave similar accuracy at much lower cost due to the analytical null distribution. We therefore find that it is best to use the λ -LRT for datasets up to approximately 1,000 cells and view PATH as a practical alternative for larger datasets, with the caveat that it may discard some informative genes.

PATH’s reduced recall motivated us to ask whether its performance depended on properties of the simulated lineage trees. The strongest pattern was for trees with very short branches, where nearby cells would be expected to have weak Brownian divergence relative to observational noise (**Supplementary Fig. S1**). Nonetheless, the impact on recall was modest, particularly on larger trees, and PATH’s precision remained high.

Constraining factor analysis invariances reproducibly recovers simulated latent structure

Factor analysis models are inherently non-identifiable. The same low-rank reconstruction can be written after rescaling, permuting, or changing the signs of the factors amongst other rotational invariances. In simulations, before resolving these invariances, individual columns of $\mathbf{F}$ and rows of $\mathbf{L}$ were not recovered consistently, even though the induced cell-cell and gene-gene covariance structures were recovered reasonably well (**Fig. 2B**). Post hoc alignment to the generating factors largely corrected this problem, confirming that much of the error reflected factor-basis ambiguity rather than failure to recover the latent structure (**Supplementary Fig. S2A**).

We therefore added simple constraints to make the fitted factors more reproducible (**Fig. 2C**). Rows of $\mathbf{L}$ were constrained to have unit ℓ_2 norm by rescaling them during inference, factors were ordered by decreasing Brownian variance post-inference, and signs were chosen deterministically from the sum of factor loadings post-inference. These choices were aimed at fixing scale, permutation, and sign without changing the model’s basic structure. An alternative would be to fix the Brownian variance parameters to 1, as in prior phylogenetic factor analysis work [69, 70]. In our setting, however, this gave poorer recovery than normalizing the rows of $\mathbf{L}$. This difference may reflect our maximum a posteriori optimization, where the factor-gene parameters are learned by stochastic gradient descent. Fixing the Brownian variances only changes the effective prior on $\mathbf{F}$, whereas row normalization provides a more direct constraint on the fitted loading matrix $\mathbf{L}$.

With these constraints in place, the fitted factors more closely matched the generating $\mathbf{F}$ and $\mathbf{L}$ across simulated dataset sizes, with moderate additional gains in recovery of the cell-cell and gene-gene covariance structure (**Fig. 2D**). It is also notable that the constraints seemed to substantially improve runtime. Post hoc alignment seemingly also had little left to improve, consistent with the main factorization ambiguities having been removed (**Supplementary Fig. S2B**). Representative simulations showed the same behavior at the level of the simulated $\mathbf{F}$, $\mathbf{L}$, $\mathbf{Z}$, and $\mathbf{X}$ matrices (**Supplementary Fig. S3A**) and in low-dimensional views of the latent and observed expression states (**Supplementary Fig. S3B**).

Recovery was generally stable across simulated dataset sizes, with the main exception occurring for the smallest trees paired with the largest gene sets. In the 500-cell, 5,000-gene setting, recovery was weaker (**Fig. 2D**), a pattern consistent with prior work showing that inference becomes more difficult when the number of traits exceeds the number of observations [72]. Interestingly, we did not observe catastrophic failure in this regime, just minor underperformance. We do not pursue this limitation further here because lineage-tracing datasets are moving toward larger trees, whereas the number of genes expressed by any cell is biologically constrained. Thus, the natural data regime for this problem is one with many more cells than genes, where the model seemingly performs well.

Regularization encourages distinct gene modules in real data

We next aimed to apply our model to two lineage-tracing datasets, one in a developmental biology setting [62] and the other in lung cancer [10]. Based on the simulation results above, we restricted these analyses to clones containing more than 1,000 cells, where reconstruction was generally strongest. This filtering left two clones from the developmental dataset and four clones from the lung cancer dataset.

We initially fit the model using the invariance constraints established in simulations. Although these constraints encouraged the factorization to be reproducible, the resulting factors often contained overlapping gene loadings that were difficult to interpret (**Fig. 3A**). In many respects, the goal of our model is not simply to reconstruct expression from a low-rank representation, but to identify relatively distinct and interpretable factors corresponding to biologically meaningful combinations of genes. We therefore sought to constrain two potential sources of factor overlap: insufficient sparsity within rows of $\mathbf{L}$ and similarity in the directions that different rows of $\mathbf{L}$ point through "gene space".

To address these sources of overlap, we implemented dataset-size-adjusted penalties consisting of an ℓ_1 penalty on each row of $\mathbf{L}$ and a cosine-similarity overlap penalty between rows. We evaluated their effects on numerical factor overlap using top-gene Jaccard similarity and row-wise Pearson correlation, on model fit using normalized log-likelihood, and on biological function overlap using shared Gene Ontology (GO) terms and the fraction of factors with significant GO enrichment (**Fig. 3B**).

The ℓ_1 penalty had little consistent effect on either numerical or biological overlap, other than substantially reducing model fit at high values. In contrast, the overlap penalty meaningfully reduced both top-gene overlap and correlation between rows of $\mathbf{L}$, while also tending to reduce overlap between their associated GO terms. We therefore selected an overlap penalty of $\lambda_C = 1$ and an ℓ_1 penalty of $\lambda_1 = 0.1$. The modest

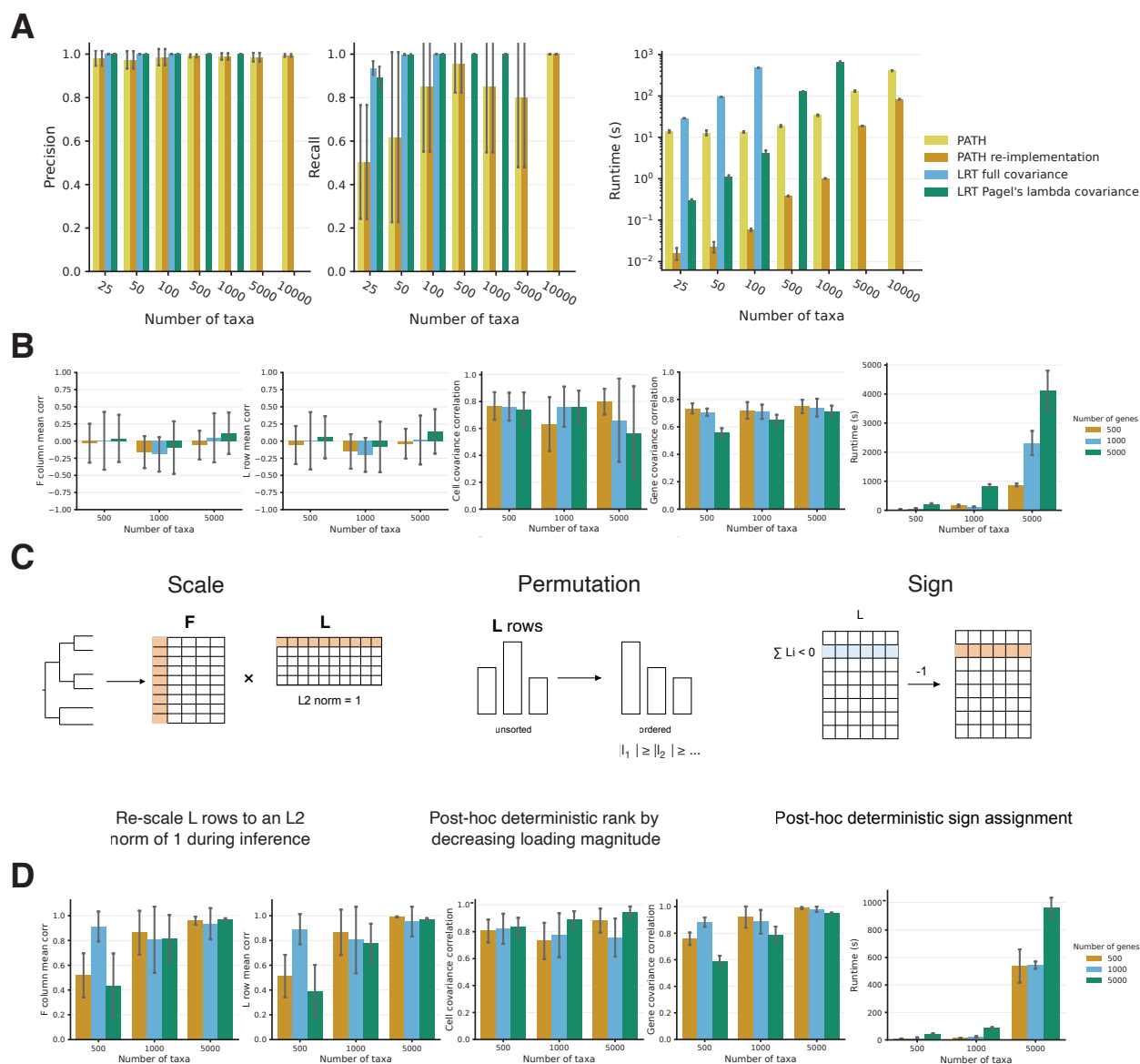


Figure 2: Simulation benchmarking and identifiability constraints for phylogenetic factor analysis.

(A) Performance, in terms of precision (left), recall (middle), and runtime (right), of four methods for identifying phylogenetically informative genes across simulated datasets. Methods include the original PATH correlation-based approach, a reimplementation of PATH, and two likelihood-ratio test (LRT) approaches based on Brownian motion covariance models. (B) Recovery of latent factors before resolving model invariances. From left to right: mean Pearson correlation between inferred and true columns of F , mean Pearson correlation between inferred and true rows of L , cell-cell covariance correlation, gene-gene covariance correlation, and runtime across dataset sizes. (C) Graphical representation of constraints used to resolve scale, permutation, and sign invariances in the factorization. (D) Model performance after applying invariance constraints and refitting. Panels are shown in the same order as in (B), demonstrating improved recovery and interpretability of latent factors while maintaining accurate covariance reconstruction.

ℓ_1 penalty was retained to encourage some sparsity, although its empirical contribution appeared limited and we recognize that this may be partly redundant. Under these conditions, the fitted loading matrices contained substantially more distinct and interpretable factors (**Fig. 3C**).

For these tests, we fixed the number of factors to five across all six clones. However, the percentage of factors with significant GO enrichment was often below 100% (**Fig. 3B**), and the last factor in the representative loading matrix appeared to have relatively weak structure (**Fig. 3C**). These observations suggested that five factors may not be appropriate for every clone. Choosing dimensionality is a common challenge in similar tasks, and these results motivated us to consider it more carefully in the final model fits for each dataset.

Lineage-informed factors reveal conserved developmental transitions and clone-specific cell-cycle programs

The developmental lineage-tracing dataset was generated using a mouse embryonic stem cell model of the trunk that proceeded through four days of gastrulation followed by one day of neural tube differentiation [62]. We selected the two largest clones, trunk-like structures 1 and 2 (TLS1 and TLS2), each containing more than 1,000 cells. We first applied the correlation-based PATH filter to identify genes with phylogenetic signal. A large fraction of ribosomal and mitochondrial genes passed this filter (**Supplementary Fig. S4A**), consistent with prior work challenging the assumption that ribosomal genes are static housekeeping genes [50]. Retaining these genes and fitting the factor analysis model produced a single dominant ribosomal and mitochondrial factor that obscured structure in other modules, so we removed them from subsequent analyses.

Among the remaining genes, 8.4% in TLS1 and 15.8% in TLS2 showed phylogenetic signal (**Supplementary Fig. S4B**). Of the genes shared between samples, 5.8% showed signal in both, with an additional 3% specific to TLS1 and 11% specific to TLS2. Despite this limited gene-level overlap, the two sets were enriched for nearly identical GO biological processes, including pattern specification, embryonic organ development, morphogenesis, and stem cell differentiation (**Supplementary Fig. S4C**). Thus, different genes carried phylogenetic signal in the two clones, but they converged on the same developmental processes.

We next fit the phylogenetic factor analysis model to the filtered genes from each clone. To select the latent dimensionality k , we evaluated log-likelihood, significant GO enrichments, newly identified GO terms, and the number of factors explaining most of the variance under the Brownian motion prior across a range of values. Improvements in both clones plateaued near $k = 10$ (**Supplementary Fig. S5**). At larger values of k , the model collapsed excess factors by shrinking their Brownian variance to the specified floor. We therefore selected $k = 10$ for both clones.

Only four factors in TLS1 and six in TLS2 had Brownian variance above the floor in the final model fits (**Fig. 4A,C**). The factors had distinct top genes (**Fig. 4B,D**) and direction-specific GSEA enrichments within each clone (**Supplementary Fig. S6A**). Comparison of their gene loadings across clones identified three conserved programs (**Fig. 4E**). The strongest correspondence was between TLS1 factor 4 and TLS2 factor 2

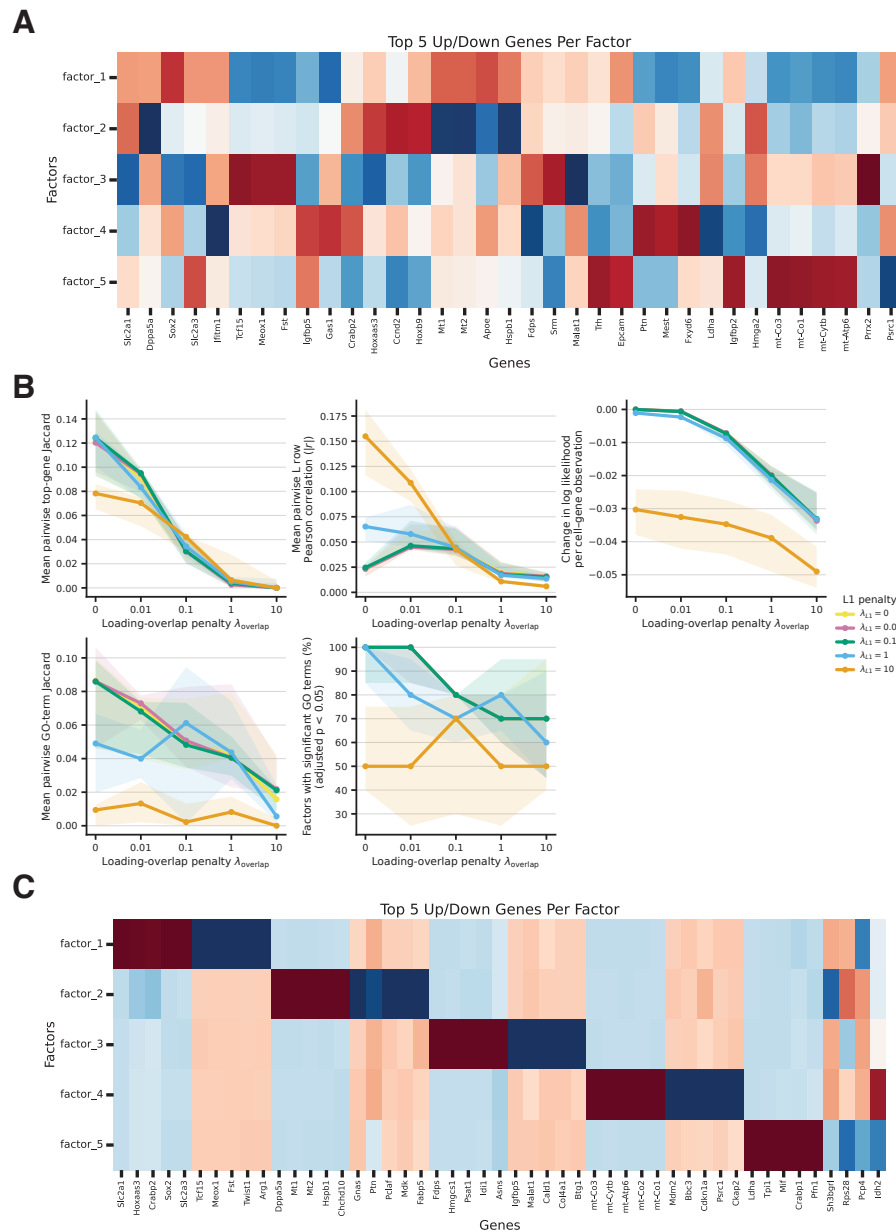


Figure 3: Regularization encourages distinct factors in real data.

(A) Factor-gene loading matrix $\mathbf{L}$ before constraining factor similarity, showing the five largest positive and negative gene loadings per factor. (B) Effects of varying row-wise ℓ_1 and cosine-similarity overlap penalties. Lines and shaded regions show the mean and interquartile range across datasets, respectively. Metrics are pairwise Jaccard similarity of the top 10 positive and negative genes per factor, ignoring sign (top left); pairwise Pearson correlation between rows of $\mathbf{L}$ (top middle); dataset-size-normalized log-likelihood (top right); pairwise Jaccard similarity of Gene Ontology (GO) terms (bottom left); and the percentage of factors with at least one significant GO term (bottom middle). (C) Loading matrix after applying the selected penalties of $\ell_1 \lambda_1 = 0.1$ and cosine-similarity overlap $\lambda_C = 1$, displayed as in (A).

($r = 0.93$). In both clones, one pole contained *Dppa5a*, *Chchd10*, and *Ckb*, whereas the opposing pole contained *Mdk*, *Gnas*, *Ldha*, *Fabp5*, and *Hmga2*. TLS1 factor 4 was positively enriched for leukemia inhibitory factor (LIF) response and negatively enriched for gliogenesis and spinal cord development. These results suggest a reproducible lineage-varying axis between a stem-cell-associated state and neural differentiation.

TLS1 factor 2 aligned with TLS2 factor 3 ($r = 0.71$). The corresponding poles in both factors contained *Crabp2*, *Ccnd2*, *Hoxaas3*, and posterior *Hox* genes, including *Hoxc8*, *Hoxb9*, and *Hoxb5os*. In TLS1 factor 2, this pole opposed epithelial-associated genes including *Epcam*, *Igfbp2*, *Krt8*, and *Emb*. In TLS2 factor 3, it opposed *Tcf15*, *Meox1*, *Fst*, *Foxd1*, and *Twist1*. Thus, the same posterior-patterning program was recovered in both clones but was coupled to different alternative transcriptional states.

TLS1 factor 3 corresponded inversely to TLS2 factor 1 ($r = -0.71$), reflecting an arbitrary reversal of factor orientation. In both cases, one pole was marked by *Ifitm1*, *Asns*, and *Psat1*, whereas the other contained *Peg3*, *Meg3*, *Prrx2*, and *Baspl*. The associated enrichments separated small-molecule and organophosphate metabolism from organ and tube morphogenesis, adhesion, motility, and vascular development. TLS1 factor 3 and TLS2 factor 1 therefore recovered the same coupling between biosynthetic activity and morphogenetic state despite their opposite signs.

The remaining active factors captured clone-specific variation. TLS1 factor 1 was marked on one side by *Tcf15*, *Meox1*, *Fst*, *Twist1*, and *Gas1* and on the other by *Slc2a3*, *Igfbp2*, *Bex1*, *Trh*, and *Epcam*. Its enrichments linked the first pole to pattern specification and muscle development and the second to carbohydrate transport. In TLS2, factor 4 was distinguished by the epithelial genes *Krt8*, *Krt18*, *Emb*, *Spink1*, and *Igfbp2*. TLS2 factors 5 and 6 instead resolved proliferation into two programs. Factor 5 contained mitotic genes including *Ube2c*, *Cenpf*, *Arl6ip1*, *Cks2*, and *Cdc20* and was strongly enriched for cell division and organelle fission. Factor 6 was marked by replication-associated histones together with *Rrm2* and *Pclaf* and was enriched for DNA replication, repair, and chromosome organization. The TLS2-specific signal therefore included an epithelial program and separately heritable mitotic and S-phase programs rather than a single generic cell-cycle axis.

Finally, we compared the fitted factors with the mutually exclusive gene modules that can be recovered through network-based clustering using a method named Hotspot [51]. Several factors had direct Hotspot counterparts (**Fig. 4F**). TLS1 factors 4, 2, and 1 aligned most strongly with Hotspot modules M1, M2, and M3, respectively, while TLS2 factors 4, 2, 1, and 3 aligned with M5, M1, M2, and M3. However, individual factors could also collapse multiple Hotspot modules. For example, TLS1 factor 4 was positively associated with M1 and negatively associated with M4, placing two mutually exclusive modules on opposing sides of the conserved stem-cell-to-differentiation axis. These modules exhibited coordinated changes along the lineage (**Fig. 4G**), supporting their representation as components of a single lineage-varying program. Thus, Hotspot recovered many of the same local gene associations, while the factor model organized them into signed developmental axes and separately identified the TLS2 mitotic and DNA-replication programs.

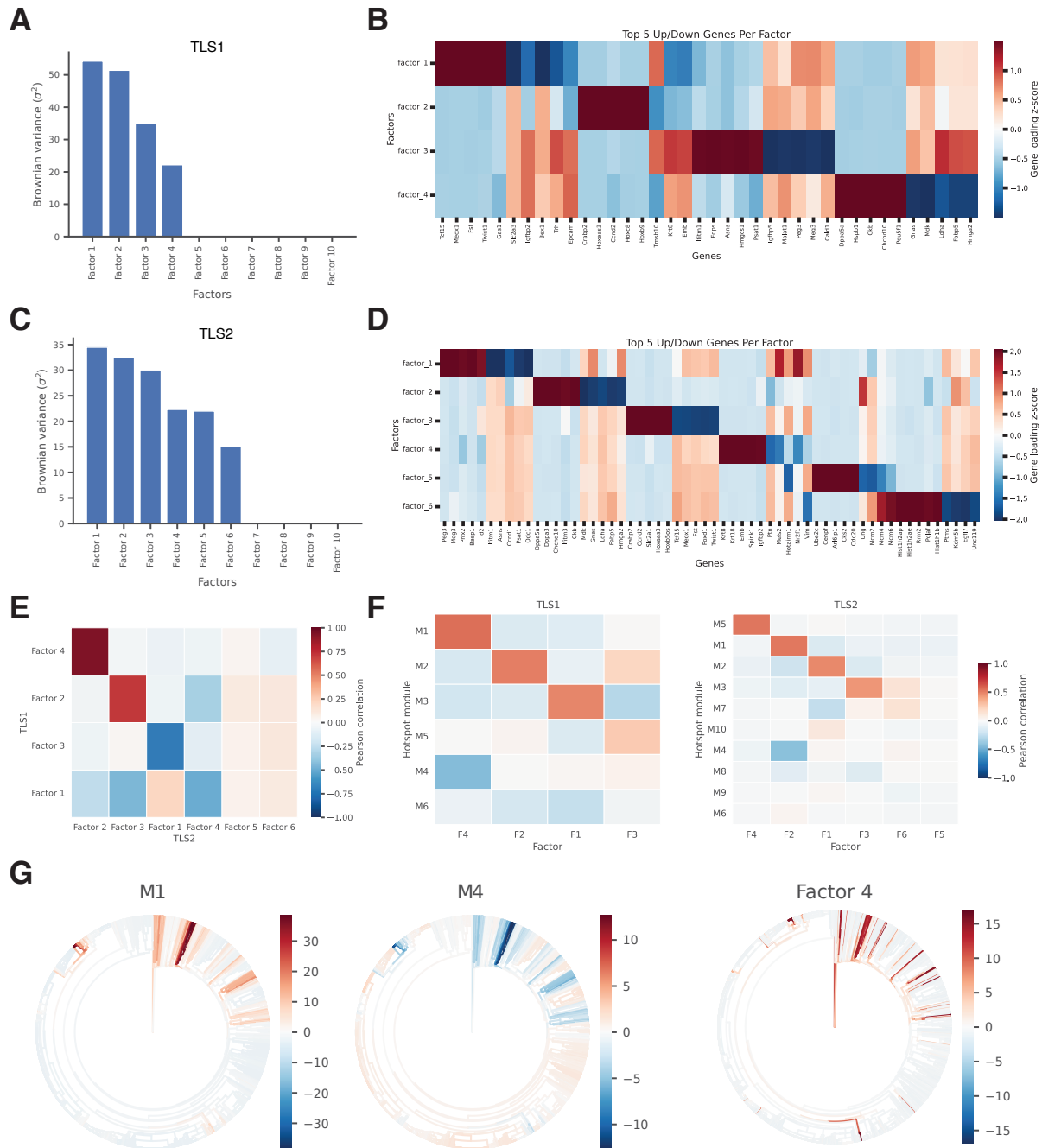


Figure 4: Phylogenetic factors recover developmental gene expression programs.

(A) Fitted Brownian variance (σ^2) across 10 TLS1 factors. (B) Standardized loadings for the five largest positive and negative genes per active TLS1 factor. (C–D) Brownian variances and gene loadings for TLS2, displayed as in (A–B). (E) Pearson correlations between active TLS1 and TLS2 factor-loading profiles. (F) Pearson correlations between factor-loading profiles and Hotspot modules in TLS1 (left) and TLS2 (right). (G) Representative collapse of TLS1 Hotspot modules M1 and M4 into factor 4, with module and factor scores projected along the lineage tree.

Lineage-informed factors reveal shared and clone-specific programs in lung cancer progression

The lung cancer dataset of interest contained four clones with more than 1,000 cells, ranging from ~1,000 cells in CP4 to ~10,000 cells in CP1, resulting from a KRAS-mutant lung adenocarcinoma cell line (A549) xenograft model that progressed to metastasis over 54 days [10]. As in the developmental dataset, ribosomal and mitochondrial genes frequently showed phylogenetic signal (**Supplementary Fig. S7A**) and contributed strongly to a single dominant factor when retained in downstream factor analysis. We therefore removed these genes before subsequent analyses.

Among the remaining genes, the fraction with phylogenetic signal varied from 1.4% in CP3 to 9.2% in CP1 (**Supplementary Fig. S7B**). Of the 16,773 genes detected in all four clones, 2,075 showed phylogenetic signal in only one clone, whereas only 26 showed phylogenetic signal in all four. The associated GO enrichments also differed across clones. CP1 was dominated by oxidative phosphorylation, wound healing, and cell migration; CP2 by apoptosis, reactive oxygen species metabolism, and detoxification; and CP4 by immune responses and detoxification. CP3 had no significant GO enrichment among its smaller filtered gene set (**Supplementary Fig. S7C**). Thus, the cancer clones shared some broad stress-response processes but differed substantially in both the number and identity of genes carrying phylogenetic signal.

We next fit models across a range of latent dimensions using the same selection criteria as for the developmental data (**Supplementary Fig. S8**). CP1 and CP2 were the two largest clones and retained the most genes with phylogenetic signal, whereas CP3 retained few genes, providing limited support for fitting a multifactor model. We therefore focused subsequent factor analysis on CP1 and CP2. For both clones, Brownian variance was concentrated in relatively few factors, and excess factors were commonly shrunk to the variance floor. We selected $k = 10$ as a common upper bound. Four factors in each clone retained Brownian variance above the floor in the final model fits (**Fig. 5A,C**). These factors had distinct top genes (**Fig. 5B,D**) and direction-specific GSEA enrichments within each clone (**Supplementary Fig. S9**).

Three of the four active factors in CP1 had clear counterparts in CP2 (**Fig. 5E**). The strongest correspondence connected CP1 factor 1 with CP2 factor 1 ($r = 0.89$). Both factors had a pole marked by *KRT18*, *GSTP1*, *CYBA*, and *PRDX4*, consistent with an epithelial oxidative-stress program. CP1 factor 3 aligned with CP2 factor 3 ($r = 0.88$) and was marked in both clones by *CXCL8*, *CXCL2*, *CXCL1*, *NFKBIA*, and *PMEPA1*. The opposing poles contained secretory epithelial genes including *MUC5AC*, *AGR2*, and *TFF3*. Thus, both clones contained a closely matched lineage-varying program separating an inflammatory chemokine state from a secretory epithelial state. A third correspondence linked CP1 factor 2 and CP2 factor 2 ($r = 0.68$). Both were distinguished by *AKR1B10*, *ALDH3A1*, *KRT81*, and related detoxification genes. In contrast, CP1 and CP2 factors 4 showed little similarity ($r = 0.11$). CP1 factor 4 separated *AKR1C2*, *IGFBP4*, *ALDH1A1*, and *SAT1* from *TFF1*, *REG4*, and *LGALS4* and was enriched for morphogenesis and organ development. CP2 factor 4 instead had a pole marked by *CEACAM6*, *C12orf75*, *REG4*, and *CYP11B1* and was enriched more broadly for signaling and cell communication. The two largest clones therefore shared inflammatory, oxidative-stress, and detoxification programs but differed in an additional lineage-varying program.

The fitted factors were again concordant with several Hotspot modules in the lung cancer dataset (**Fig. 5F**). CP1 factor 3 and CP2 factor 3 aligned most strongly with Hotspot modules M1 and M2, respectively, while the detoxification-associated factor 2 aligned with CP2 modules M3 and M5. As in the developmental data, individual factors also connected multiple mutually exclusive Hotspot modules. CP1 factor 4 captured coordinated variation represented separately by modules M3, M4, and M5, whose scores changed together along the lineage tree (**Fig. 5G**). These results support the recovered gene-level programs while showing how the signed factors organize multiple co-expression modules into shared and clone-specific axes of lineage-associated variation.

Lineage-associated programs are distinct across developmental and lung cancer model systems

We next asked whether the factors learned in the developmental and lung cancer datasets represented similar biological processes. Because the two datasets were generated using cells from different species and retained different phylogenetically informative genes, we compared each active developmental factor with each active cancer factor using the Jaccard similarity of their significantly enriched GO biological process terms (**Fig. 6**). Most cross-context comparisons showed no overlap, with 61 of 80 factor pairs having a Jaccard similarity of zero, and even the strongest similarity was only 0.18. Thus, most learned lineage-associated factors appear to be specific to the biological system in which they were inferred, rather than reflecting a small set of generic heritable expression programs.

The few detectable similarities largely involved broad processes that are known to be engaged in both development and cancer. TLS1 factor 2 overlapped most strongly with CP1 and CP2 factor 3 (Jaccard similarities of 0.18 and 0.17, respectively), sharing terms related to transcriptional and biosynthetic regulation, cell differentiation, neurogenesis, and organ development. TLS2 factor 1 also overlapped with CP1 and CP2 factor 3 (0.16 and 0.13), with shared processes including cell migration and adhesion, cytoskeletal organization, wound response, cell differentiation, morphogenesis, and vascular development. Weaker correspondences linked developmental factors to cancer programs through neural, epithelial, tube, and circulatory-system morphogenesis, while TLS2 factors 5 and 6 shared only general cell-cycle terms with cancer factor 3. These limited overlaps suggest that the common signal is concentrated in general regulatory, proliferative, and morphogenetic processes, whereas the more specific composition of most lineage-varying programs remains context dependent.

An analogous comparison of GO enrichments for Hotspot modules gave a similar result. Of the 240 developmental–cancer module pairs, 141 shared no significant terms, and the strongest Jaccard similarity was only 0.13 (**Supplementary Fig. S10**). Thus, the cross-system functional distinctness was also evident among modules learned by an independent lineage-aware network-based clustering method.

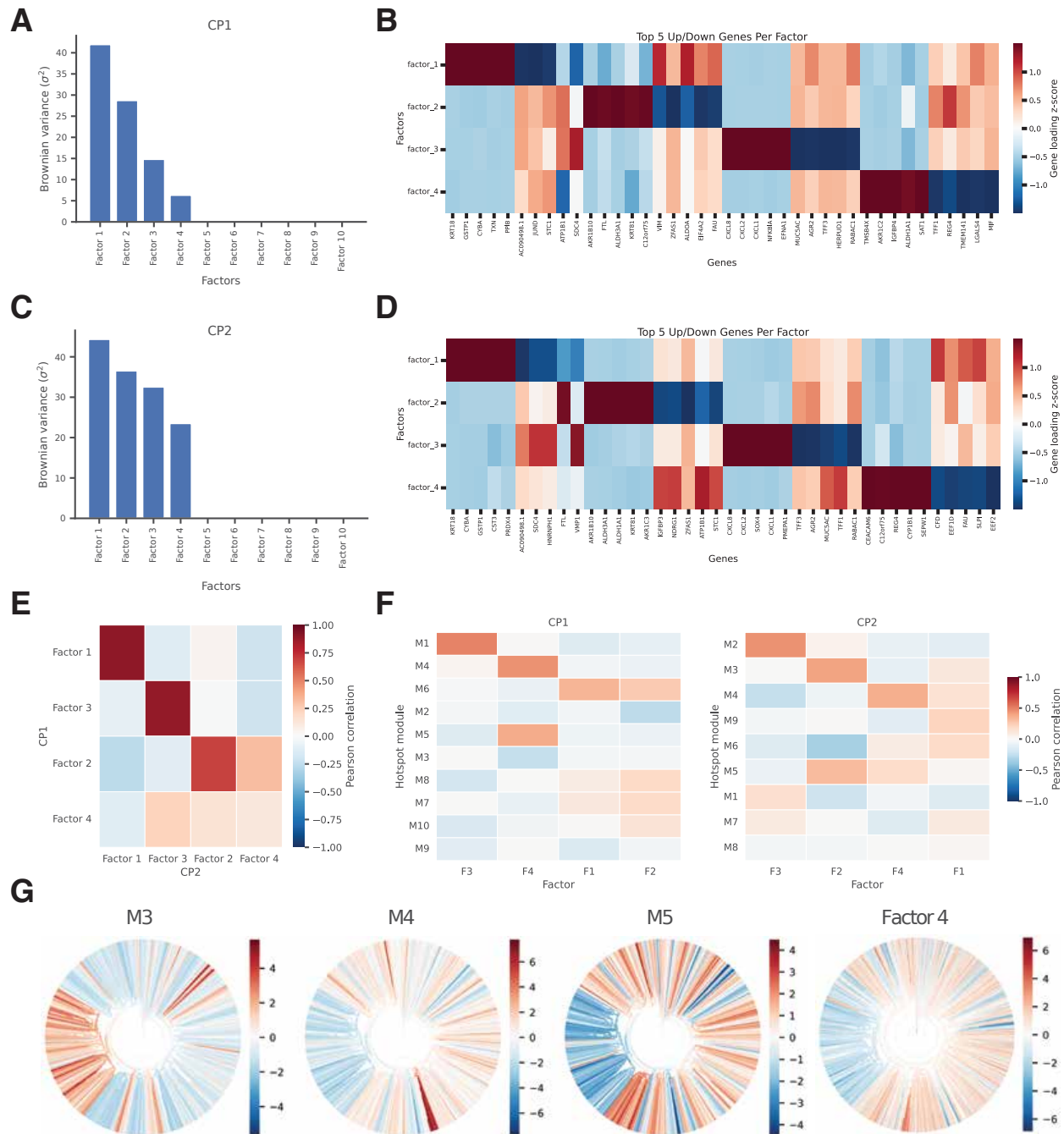


Figure 5: Phylogenetic factors recover gene expression programs in lung cancer.

(A) Fitted Brownian variance (σ^2) across 10 CP1 factors. (B) Standardized loadings for the five largest positive and negative genes per active CP1 factor. (C–D) Brownian variances and gene loadings for CP2, displayed as in (A–B). (E) Pearson correlations between active CP1 and CP2 factor-loading profiles. (F) Pearson correlations between factor-loading profiles and Hotspot modules in CP1 (left) and CP2 (right). (G) Representative collapse of CP1 Hotspot modules M3, M4, and M5 into factor 4, with module and factor scores projected along the lineage tree.

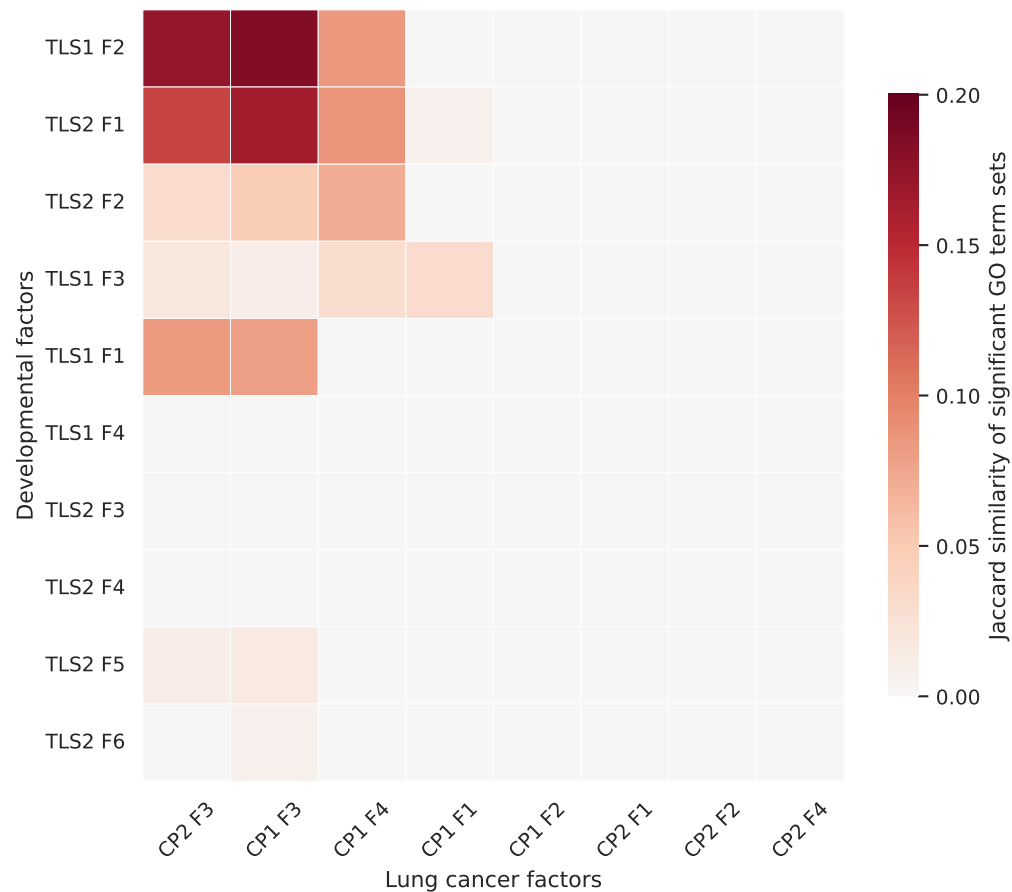


Figure 6: Phylogenetic factors are functionally distinct across developmental and cancer systems.

Jaccard similarity between sets of significantly enriched Gene Ontology biological process terms for each active developmental factor (rows) and active lung cancer factor (columns). Factors are ordered around the strongest one-to-one matches.

Discussion

We introduced scPFA, a phylogenetic factor analysis model that jointly learns lineage-varying representations of cells and the gene modules that define them. Results from fitting the model to simulated data showed that it could recover latent factors and their implied covariance structure, while additional regularization produced more distinct and interpretable factors in real data. Applied to developmental and lung cancer lineage-tracing datasets, the model identified a small number of factors with substantial Brownian variance, biologically coherent gene loadings, and significant functional enrichments that aligned well with expectations for each respective model system. Several factors were reproducible across clones and aligned with modules identified by Hotspot, while others captured clone-specific programs or combined multiple Hotspot modules into a single signed axis of lineage-associated variation.

Our gene-filtering analyses identified a small percentage of genes with detectable phylogenetic signal across both model systems, consistent with evidence for heritable transcriptional memory in single cells [49]. However, most genes showed no such signal, suggesting that gene-level heritability across the entire lineage tree is not pervasive under our chosen gene-filtering model and significance thresholds. Ribosomal and mitochondrial genes were notable exceptions, frequently exhibiting strong phylogenetic signal. This result agrees with recent evidence that many broadly expressed genes, including ribosomal genes that are often considered to be housekeeping genes, may retain notable lineage-dependent expression variation [50]. Interestingly, the lung cancer clones differed more strongly in the number and identity of genes with phylogenetic signal, consistent with a more constrained set of transitions in the directed developmental system and a broader range of heritable transcriptional outcomes in cancer. It is worth noting, however, that the present data do not distinguish cellular plasticity from differences in clone history, environment, or sampling, so further work will be required.

The biological composition of our learned heritable factors reflected the different settings represented by the two datasets well. In the developmental system, three programs were conserved across clones, including axes associated with stem-cell maintenance and neural differentiation, posterior patterning, and the coupling of morphogenesis to biosynthetic state. The lung cancer clones shared several programs centered on oxidative stress, inflammatory chemokine expression, secretory epithelial state, and detoxification.

Direct comparison across the two biological settings further showed that most inferred programs were context specific. More than three quarters of developmental–cancer factor pairs shared no significantly enriched GO terms, and the strongest overlaps remained modest. The exceptions involved broad transcriptional and biosynthetic regulation, differentiation, morphogenesis, migration and adhesion, vascular development, and cell-cycle control. Thus, a small core of general cellular programs may recur across systems, but most differ, or perhaps are reorganized into different heritable modules, between embryonic development and lung cancer progression.

More broadly, these results provide a proof of concept for using an explicit evolutionary model to discover heritable gene modules rather than testing genes independently or defining modules through clustering in a way that is constrained by the lineage. This formulation connects the composition of a program, its ac-

tivity in individual cells, and its variation through cellular history within one joint probabilistic framework. It also creates a basis for formal hypothesis testing through comparisons of nested models. For example, Brownian motion could be compared with Ornstein–Uhlenbeck models of stabilizing regulation, models with shifts in optima or rates along particular branches, or models in which different factors respond to different lineage or environmental regimes. Related model comparisons have been useful for testing gene-expression dynamics in our prior work [40]. Extending them to latent factors would allow such hypotheses to be tested at the level of coordinated transcriptional programs. The Brownian model used here therefore serves as a simple baseline upon which richer biological process models can be built.

It is also worth noting that the inferred factors imply a lower-dimensional gene-expression space in which movement along cellular lineages is naturally reconstructed, yielding lineage-aware analogs of trajectory inference or Waddington landscapes that we did not fully explore here [76]. Here, we focused our efforts on recovering the gene modules themselves. However, the fitted factor analysis model provides much richer information and could be immediately used for a variety of other tasks including the reconstruction of these differentiation landscapes and the study of discrete versus continuous dynamics of cell-state transitions.

The current implementation also has several limitations that motivate future work. First, it treats normalized log expression as a continuous Gaussian trait rather than modeling single-cell counts and their mean–variance relationship directly. Second, filtering genes before factor analysis helps computational tractability but conditions the model on a fixed set of genes with detectable phylogenetic signal. A joint model of heritable and non-heritable components could propagate uncertainty in this classification and recover weak signals that are apparent only at the module level. The present continuous model also does not naturally represent genes with binary or strongly zero-inflated expression, which could be incorporated through mixed observation models. Finally, the lineage tree was treated as known in the analyses presented here. Our code is set up to accept a collection of candidate trees and marginalize the Brownian factor priors across them, but we used a single tree per clone to keep this initial analysis simple. Evaluating that functionality on posterior samples of lineage trees will be important because our prior work has confirmed that uncertainty in reconstructed histories can meaningfully affect downstream biological conclusions [77]. Such extensions could propagate tree, expression, and module uncertainty together, although fully Bayesian phylogenetic factorization may require additional approximations to remain scalable to large single-cell datasets [69, 70]. These are all relatively direct extensions of the current work because they retain a fixed factor-loading matrix across the lineage. A more ambitious model could allow module composition itself to evolve as genes enter, leave, or change their contributions to a program. This could reveal how regulatory programs are reorganized during development and disease progression.

Together, our results demonstrate the promise of phylogenetic factor analysis for extracting interpretable, lineage-associated structure from high-dimensional single-cell data. By connecting gene modules, cellular states, and their shared histories in a unified framework, this approach provides both a practical tool for studying transcriptional heritability and a flexible foundation for increasingly realistic models of cellular evolution.

Methods

Phylogenetic factor model specification

Tree tips and expression profiles were restricted to their shared cells, and each tree was rescaled to unit height so that Brownian variance parameters are fit on a similar scale across datasets. Before phylogenetic gene filtering and factor-model fitting, counts were normalized to a total of 10,000 per cell to control for library size, transformed as $\log(1+x)$ to stabilize their variance and make a Gaussian working approximation more reasonable, and centered across cells for each gene to allow modeling Brownian motion around a mean of 0.

For n cells, p retained genes, and k latent factors, we modeled the resulting expression matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ as

$$\mathbf{X} = \mathbf{F}\mathbf{L} + \mathbf{E}, \quad (1)$$

$$E_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_{\text{obs}}^2), \quad (2)$$

$$\mathbf{F}_{\cdot d} \sim \text{MVN}(\mathbf{0}, \sigma_d^2 \mathbf{\Sigma}), \quad d = 1, \dots, k. \quad (3)$$

Here, $\mathbf{F} \in \mathbb{R}^{n \times k}$ contains factor activities across cells, $\mathbf{L} \in \mathbb{R}^{k \times p}$ contains their gene loadings, and $\mathbf{\Sigma}$ is the Brownian covariance induced by the lineage tree, with Σ_{ij} equal to the time shared by cells i and j from the root to their most recent common ancestor. Each factor has its own Brownian variance σ_d^2 , whereas σ_{obs}^2 describes residual expression variation shared across genes and cells. We treated the factors as independent Brownian processes and did not estimate covariance between them, providing a simpler baseline in which lineage dependence is modeled within, but not between, factors.

Inference balances reconstructing the observed expression matrix with $\mathbf{F}\mathbf{L}$ and requiring each column of $\mathbf{F}$ to vary across cells in a manner consistent with Brownian motion on the lineage tree. The observation variance controls unexplained expression noise, while each Brownian variance controls the amount of lineage-associated variation assigned to its factor. The observation log-likelihood is $\log p(\mathbf{X} \mid \mathbf{F}, \mathbf{L}, \sigma_{\text{obs}}^2)$. The Brownian log prior is $\log p(\mathbf{F} \mid T, \sigma_1^2, \dots, \sigma_k^2)$. Their sum is the unpenalized joint log density used for maximum a posteriori inference of the latent factor activities:

$$\mathcal{L}_0 = \frac{np}{2} \log(2\pi\sigma_{\text{obs}}^2) + \frac{\|\mathbf{X} - \mathbf{F}\mathbf{L}\|_F^2}{2\sigma_{\text{obs}}^2} \quad (4)$$

$$+ \frac{1}{2} \sum_{d=1}^k \left[n \log(2\pi) + \log |\sigma_d^2 \mathbf{\Sigma}| + \mathbf{F}_{\cdot d}^\top (\sigma_d^2 \mathbf{\Sigma})^{-1} \mathbf{F}_{\cdot d} \right]. \quad (5)$$

We minimized the penalized objective $\mathcal{L} = \mathcal{L}_0 + P(\mathbf{L})$, where $P(\mathbf{L})$ contains the loading penalties described below. This is equivalent to maximum a posteriori estimation of the latent factor activities under

their Brownian priors. Brownian densities and gradients were evaluated by Gaussian pruning along the tree, avoiding repeated construction or inversion of dense covariance matrices.

Loading entries were initialized from a standard normal distribution and normalized by row. A provisional $\mathbf{F} = \mathbf{X}\mathbf{L}^\top$ was used only to initialize the Brownian variance parameters. For each provisional factor, its empirical variance across cells was used to choose an initial σ_d^2 that produced the same expected variance among the tree tips. Before each objective evaluation, including the first, $\mathbf{F}$ was set to its exact conditional optimum given the current loadings and variance parameters. This update diagonalized Σ once and solved a $k \times k$ linear system for each tree eigenmode. Each entry of the loading matrix $\mathbf{L}$ and the logarithms of the variance parameters were optimized as free parameters using Adam [78] with a learning rate of 0.01 and adaptive gradient-norm clipping. Brownian variances were bounded below by 10^{-6} , and additional loading constraints were enforced as described below.

Optimization proceeded for at least 500 iterations. Convergence was assessed from the mean objective over successive 50-iteration windows and declared when the relative improvement was no greater than 10^{-6} . The parameter state with the lowest observed objective was retained. The factor dimension k must be specified and is then fixed within each fit.

Lineage-based gene expression simulation

We simulated ultrametric cell-lineage trees using the Cassiopeia (v2.1.0) [79] birth–death fitness simulator, with birth and death rates of 0.075 and 0.005, respectively, and rescaled each tree to unit height. Gene expression and latent factors were then generated forward under the same Brownian model used for inference. For a tree with covariance matrix Σ , each latent factor was drawn as

$$\mathbf{F}_{\cdot d} \sim \text{MVN}(\mathbf{0}, \sigma_d^2 \Sigma). \quad (6)$$

To evaluate phylogenetic gene filtering, we simulated 10 trees at each of 25, 50, 100, 500, 1,000, 5,000, and 10,000 cells. For every tree, we generated 1,000 heritable genes under Brownian motion and 1,000 non-heritable genes using identity covariance, with both sets having variance 0.25 at the tips. These known classes were used to calculate the precision and recall of each filtering method.

To evaluate factor-model recovery, we simulated 10 datasets for each combination of 500, 1,000, or 5,000 cells and 500, 1,000, or 5,000 genes. Each dataset contained five factors with Brownian variances (20, 10, 5, 2, 1). Loading entries were drawn independently from a standard normal distribution and each loading row was rescaled to unit Euclidean norm. We generated the observed expression matrix as

$$\mathbf{X} = \mathbf{F}\mathbf{L} + \mathbf{E}, \quad E_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 0.1), \quad (7)$$

and fit the model directly to these simulated matrices without phylogenetic gene filtering, allowing the

inferred and generating factors, loadings, covariance structures, and objective components to be compared directly.

Statistical tests for phylogenetic gene filtering

For the provided tree, we calculated the Brownian covariance matrix Σ , whose entry Σ_{ij} is the shared root-to-ancestor time for cells i and j . Let $\mathbf{y}_g$ denote the centered expression of gene g across n cells.

For the autocorrelation test, we followed the approach introduced in PATH (v1.0) [61] and confirmed that our implementation was numerically identical to the original. Pairwise patristic distances were calculated as $d_{ij} = \Sigma_{ii} + \Sigma_{jj} - 2\Sigma_{ij}$ and converted to weights $w_{ij} = 1/d_{ij}$ for $i \neq j$, with $w_{ii} = 0$. The full weight matrix was then normalized so that $\sum_{ij} w_{ij} = 1$. Phylogenetic autocorrelation was measured using Moran’s statistic,

$$I_g = \frac{n \mathbf{y}_g^\top \mathbf{W} \mathbf{y}_g}{\mathbf{y}_g^\top \mathbf{y}_g}. \quad (8)$$

We compared the observed I_g with its expected value under random assignment of expression values across cells, $E[I_g] = -1/(n-1)$. We divided this difference by its analytic standard deviation, as defined in PATH, to obtain a z score and calculated a two-sided normal P value.

The full Brownian likelihood-ratio test compared an independent Gaussian null model, $\mathbf{y}_g \sim \text{MVN}(\mu_g \mathbf{1}, \sigma_g^2 \mathbf{I})$, with the Brownian alternative $\mathbf{y}_g \sim \text{MVN}(\mu_g \mathbf{1}, \sigma_g^2 \Sigma)$. Under each model, the mean was estimated by generalized least squares and the variance by its closed-form maximum-likelihood estimate, thereby profiling both parameters without numerical optimization. The test statistic was

$$D_g = 2(\ell_{\text{BM},g} - \ell_{0,g}). \quad (9)$$

Because these covariance models are not nested through a single fitted parameter, we generated 1,000 independent Gaussian datasets under the gene-specific null model and calculated the P value as the fraction producing a statistic at least as large as D_g .

The Pagel’s λ likelihood-ratio test provided a nested version of this comparison, following its original phylogenetic formulation [74, 75] and our previous application to lineage-traced expression data [40]. For each gene, we formed $\Sigma(\lambda)$ by retaining the diagonal of Σ and multiplying its off-diagonal entries by λ , then maximized the profiled Gaussian likelihood over $0 \leq \lambda \leq 1$ using bounded Brent optimization, with both boundary values also evaluated explicitly. Thus, $\lambda = 0$ represents phylogenetically independent expression and increasing λ restores the covariance expected from the lineage tree. We calculated $D_g = 2\{\ell_g(\hat{\lambda}) - \ell_g(0)\}$ and obtained its P value from the $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ boundary-null distribution. For all three tests, P values were adjusted across genes using the Benjamini–Hochberg procedure, and genes with $q \leq 0.05$ were retained. The likelihood-ratio tests additionally required $D_g > 0$.

Factor-model identifiability constraints

The factorization $\mathbf{X} \approx \mathbf{F}\mathbf{L}$ is unchanged if a column $\mathbf{F}_{\cdot d}$ is multiplied by a constant and the corresponding loading row $\mathbf{L}_{d\cdot}$ is divided by the same constant. Paired sign changes and factor permutations are likewise equivalent. We fixed the scale of each factor by constraining every row of the loading matrix to have unit Euclidean norm,

$$\|\mathbf{L}_{d\cdot}\|_2 = 1. \quad (10)$$

During optimization, loading gradients were projected onto the tangent space of this constraint and loading rows were renormalized after each update. The corresponding column of $\mathbf{F}$ therefore contained the factor's fitted scale and variation across cells.

After fitting, factors were ordered by decreasing Brownian variance σ_d^2 . Because simultaneously multiplying $\mathbf{F}_{\cdot d}$ and $\mathbf{L}_{d\cdot}$ by -1 leaves their product unchanged, we also oriented each factor so that $\sum_j L_{dj} \geq 0$, reversing both its scores and loadings when necessary. These conventions do not change the fitted expression matrix, but provide consistent factor scales, labels, and directions for reporting and comparison.

Factor-loading regularization

We regularized the factor-loading matrix $\mathbf{L}$ by adding ℓ_1 and loading-overlap penalties to the model objective. For n cells, p genes, and k factors, the combined penalty was

$$P(\mathbf{L}) = \lambda_1 \frac{n}{k} \sum_{d=1}^k \sum_{j=1}^p |L_{dj}| + \lambda_C \frac{np}{\binom{k}{2}} \sum_{a < b} \frac{(\mathbf{u}_a^\top \mathbf{u}_b)^2}{\|\mathbf{u}_a\|_2^2 \|\mathbf{u}_b\|_2^2}, \quad (\mathbf{u}_d)_j = \sqrt{L_{dj}^2 + 10^{-8}}. \quad (11)$$

The ℓ_1 term favors factors defined by a sparse subset of genes. The second term is the squared cosine similarity between the absolute loading magnitudes of each factor pair and therefore discourages different factors from repeatedly using the same genes, regardless of loading sign. The scaling factors account for the size of the expression matrix and the number of factors or factor pairs in the objective.

Data acquisition, preprocessing, and VINE lineage reconstruction

We obtained paired single-cell RNA sequencing and CRISPR lineage-recording data for the TLS1 and TLS2 mouse trunk-like structures from GSE220949 [62], and for the CP1–CP4 lineage groups from the mouse 5k lung-cancer xenograft experiment in GSE161363 [10]. For the developmental data, allele tables were converted to integer character matrices using Cassiopeia, after which characters and cells containing only unedited or missing states were removed. For the cancer data, we used the deposited allele-threshold character matrices and encoded missing states as -1 . In both datasets, the 10x count matrix was restricted

and reordered to the cells represented in the corresponding character matrix, and genes with no counts among the retained cells per clone were removed.

Lineage trees were inferred separately for each sample or clone using VINE (v0.3.5) [80] with its CRISPR model. We required at least 400 variational iterations and assessed convergence over a window of 100 iterations. Each run produced 100 approximate posterior tree samples. The first sampled tree from each run was used in the present analyses.

Gene set enrichment and Gene Ontology analysis procedures

Gene Ontology (GO) and gene set enrichment analyses were performed with clusterProfiler (v4.10.0) [81] in R (v4.3.3). Gene symbols were mapped to Entrez identifiers with `bitr` using `org.Mm.eg.db` or `org.Hs.eg.db` (v3.18.0) for the mouse and human datasets, respectively. To characterize the binary gene-filtering results, genes passing and not passing the phylogenetic filter were tested separately for over-representation of GO Biological Process terms using `enrichGO`, with all mapped genes tested in that sample used as the background. To characterize the continuous factor loadings, mapped genes were ranked from the most positive to the most negative loading for each factor and tested for GO Biological Process enrichment using `gseGO`. Zero and nonfinite loadings were excluded, duplicate Entrez identifiers were represented by the loading with the largest absolute value, and the sign of the normalized enrichment score identified enrichment toward the positive or negative loading direction.

Latent-factor ancestral-state reconstruction

For each fitted factor d , we reconstructed its value at every internal node conditional on the fitted tip scores $\mathbf{F}_d$. Under the Brownian model, information about a node can be represented by a Gaussian summary (μ, v) , where μ is the estimated state and v is its uncertainty. Transporting this summary across a branch of length t leaves its mean unchanged and increases its variance from v to $v + \sigma_d^2 t$. Two independent summaries (μ_1, v_1) and (μ_2, v_2) for the same node were combined by precision weighting,

$$v = (v_1^{-1} + v_2^{-1})^{-1}, \quad \mu = v \left(\frac{\mu_1}{v_1} + \frac{\mu_2}{v_2} \right). \quad (12)$$

Zero variance denotes a fixed state and was handled as the limiting case of this calculation.

We applied this calculation in two tree traversals. First, a postorder traversal summarized the information below each node. Each tip began with the fixed state $(F_{id}, 0)$. The summaries from its two children were transported across their respective branches and combined at the parent. Second, a preorder traversal summarized the complementary information arriving from above each node. This traversal began at the root with the Brownian prior $(0, \sigma_d^2 t_{\text{root}})$. The summary passed to a child was obtained by combining the summary from above its parent with the transported summary from its sibling, and then transporting the result across the branch to the child. Finally, the descendant and complementary summaries were combined

at each internal node. The resulting mean $\hat{F}_{id}$ is therefore conditional on all fitted tip states, while tip values remain equal to their fitted factor scores.

Hotspot gene-module inference

We identified lineage-associated gene modules with Hotspot (v1.1.3) [51], following the parameter settings in its lineage-data tutorial (https://hotspot.readthedocs.io/en/latest/Lineage_Tutorial.html). For each clone, we retained cells present in both the count matrix and lineage tree, pruned the tree to those cells while preserving branch lengths, and retained genes detected in at least 10 cells. Because raw counts were available, we used Hotspot’s depth-adjusted negative-binomial model (danb) with each cell’s total UMI count as its scaling factor. We constructed an unweighted lineage-neighborhood graph with 30 neighbors. We first calculated gene-level autocorrelation and retained genes with $\text{FDR} < 0.05$. Pairwise local correlations were then calculated among all retained genes. Modules were obtained by agglomerative clustering with a minimum module size of 50 genes, an FDR threshold of 0.05 for merging branches, and `core_only=True` so that genes with ambiguous module membership were left unassigned. Per-cell module scores were calculated using `hs.calculate_module_scores()`. For visualization along the lineage tree, the score at each internal node was defined as the mean score among its descendant cells.

Pearson correlation analysis of factor loadings and Hotspot modules

We compared fitted factors across clones using their complete gene-loading profiles rather than only their top genes. Factors with Brownian variance $\sigma_d^2 > 10^{-5}$ were retained as active, and the two loading matrices were restricted to the intersection of genes present in both fits. For factors a and b , with loadings L_{ag} and L_{bg} over the shared gene set G , we calculated

$$r_{ab} = \frac{\sum_{g \in G} (L_{ag} - \bar{L}_a)(L_{bg} - \bar{L}_b)}{\sqrt{\sum_{g \in G} (L_{ag} - \bar{L}_a)^2} \sqrt{\sum_{g \in G} (L_{bg} - \bar{L}_b)^2}}. \quad (13)$$

No additional scaling or imputation was applied before calculating these correlations.

To compare factors with Hotspot, we used the intersection of genes in the factor-loading matrix and the full set of genes tested for Hotspot autocorrelation. For each Hotspot module m , we defined a binary vector H_{mg} equal to one when gene g was assigned to m and zero when it was assigned to another module or remained unassigned. Pearson correlation between H_m and each active factor-loading vector L_a then measured whether genes assigned to a module tended to have consistently positive or negative loadings on that factor. This is equivalent to a point-biserial correlation. For visualization only, rows and columns of each correlation matrix were ordered using a maximum-weight one-to-one assignment based on absolute correlation. This ordering did not alter the reported correlation values.

Cross-context comparison of factor or module biological functions

To compare factors between the developmental and lung cancer datasets, we collected GO biological process terms with GSEA-adjusted $P \leq 0.05$ across both loading directions. For a developmental factor a and cancer factor b , with significant term sets S_a and S_b , we calculated the unsigned Jaccard similarity

$$J_{ab} = \frac{|S_a \cap S_b|}{|S_a \cup S_b|}. \quad (14)$$

Pairs with no terms in their union were assigned a similarity of zero. For visualization only, rows and columns were ordered using a maximum-weight one-to-one assignment based on Jaccard similarity, which did not alter the similarity values.

We applied the same procedure to the existing GO over-representation results for Hotspot modules. Modules with at least one GO biological process term having adjusted $P \leq 0.05$ were retained, and Jaccard similarities were calculated between the significant term sets of each developmental and cancer module.

Software implementation

The phylogenetic signal gene filtering models, phylogenetic factor analysis model, and simulation tool were all implemented in C (C99). Source code is available at <https://github.com/StephenStaklinski/scPFA> under the BSD 3-Clause License. Any additional pipelines and scripts needed to reproduce the analyses presented here are available at https://github.com/StephenStaklinski/scPFA_analyses.

Funding

Funding for this work was provided by NIH National Institute of General Medical Sciences Grant R35-GM127070, NCI Grants R01-CA272466 and 5P30CA045508, Starr Cancer Consortium Grant I16-0060, a National Science Foundation Graduate Research Fellowship (S.J.S.), a Starr Centennial Scholarship (S.J.S.), and the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

Conflict of Interest

The authors declare no competing interests.

Acknowledgments

We thank other members of the community at Cold Spring Harbor Laboratory for helpful feedback, particularly Bruce Stillman, David M. McCandlish, and Hannah V. Meyer.

References

- [1] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–630, 2013.
- [2] Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome Research*, 25(10):1491–1498, 2015.
- [3] Anoop P. Patel, Itay Tirosh, John J. Trombetta, Alex K. Shalek, Shawn M. Gillespie, Hiroaki Wakimoto, Daniel P. Cahill, Brian V. Nahed, William T. Curry, Robert L. Martuza, David N. Louis, Orit Rozenblatt-Rosen, Mario L. Suvà, Aviv Regev, and Bradley E. Bernstein. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.
- [4] Itay Tirosh, Benjamin Izar, Sanjay M. Prakadan, Marc H. Wadsworth, Daniel Treacy, John J. Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, Mohammad Fallahi-Sichani, Ken Dutton-Regester, Jia-Ren Lin, Ofir Cohen, Parin Shah, Diana Lu, Alex S. Genshaft, Travis K. Hughes, Carly G. K. Ziegler, Samuel W. Kazer, Aleth Gaillard, Kellie E. Kolb, Alexandra-Chloé Villani, Cory M. Johannessen, Aleksandr Y. Andreev, Eliezer M. Van Allen, Monica Bertagnolli, Peter K. Sorger, Ryan J. Sullivan, Keith T. Flaherty, Dennie T. Frederick, Judit Jané-Valbuena, Charles H. Yoon, Orit Rozenblatt-Rosen, Alex K. Shalek, Aviv Regev, and Levi A. Garraway. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016.
- [5] Caleb Weinreb, Alejo Rodriguez-Fraticelli, Fernando D. Camargo, and Allon M. Klein. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, 367(6479):eaaw3381, 2020.
- [6] 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.e9, 2021.
- [7] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, and John L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, 2014.
- [8] Jeffrey A. Farrell, Yiqun Wang, Samantha J. Riesenfeld, Karthik Shekhar, Aviv Regev, and Alexander F. Schier. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360(6392):eaar3131, 2018.

- [9] Charissa Kim, Ruli Gao, Emi Sei, Rachel Brandt, Johan Hartman, Thomas Hatschek, Nicola Crosetto, Theodoros Foukakis, and Nicholas E. Navin. Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. *Cell*, 173(4):879–893.e13, 2018.
- [10] J. J. Quinn, M. G. Jones, R. A. Okimoto, S. Nanjo, M. M. Chan, N. Yosef, T. G. Bivona, and J. S. Weissman. Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science*, 371(6532), Feb 2021.
- [11] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriiti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V. Kharchenko. Rna velocity of single cells. *Nature*, 560(7719):494–498, 2018.
- [12] 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.
- [13] Lennart Kester and Alexander van Oudenaarden. Single-cell transcriptomics meets lineage tracing. *Cell Stem Cell*, 23(2):166–179, 2018.
- [14] Laleh Haghverdi, Maren Büttner, F. Alexander Wolf, Florian Buettner, and Fabian J. Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10):845–848, 2016.
- [15] Kelly Street, Davide Risso, Russell B. Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, 19(1):477, 2018.
- [16] Volker Bergen, Marius Lange, Stefan Peidli, F. Alexander Wolf, and Fabian J. Theis. Generalizing rna velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38(12):1408–1414, 2020.
- [17] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, 2019.
- [18] Daniel E. Wagner and Allon M. Klein. Lineage tracing meets single-cell omics: opportunities and challenges. *Nature Reviews Genetics*, 21(7):410–427, 2020.
- [19] Aaron McKenna, Gregory M. Findlay, James A. Gagnon, Marshall S. Horwitz, Alexander F. Schier, and Jay Shendure. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, 353(6298):aaf7907, 2016.
- [20] Leif S. Ludwig, Caleb A. Lareau, Jacob C. Ulirsch, Elena Christian, Christoph Muus, Lauren H. Li, Karin Pelka, Will Ge, Yaara Oren, Alison Brack, Travis Law, Christopher Rodman, Jonathan H. Chen, Genevieve M. Boland, Nir Hacohen, Orit Rozenblatt-Rosen, Martin J. Aryee, Jason D. Buenrostro, Aviv Regev, and Vijay G. Sankaran. Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell*, 176(6):1325–1339.e22, 2019.

- [21] Jin Xu, Kevin Nuno, Ulrike M. Litzenger, Yanyan Qi, M. Ryan Corces, Ravindra Majeti, and Howard Y. Chang. Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial dna. *eLife*, 8:e45105, 2019.
- [22] Ryan N. Serio, Armin Scheben, Billy Lu, Domenic V. Gargiulo, Lucrezia Patruno, Caroline L. Buckholtz, Ryan J. Chaffee, Megan C. Jibilian, Steven G. Persaud, Stephen J. Staklinski, Rebecca Hassett, Lise M. Brault, Daniele Ramazzotti, Christopher E. Barbieri, Adam C. Siepel, and Dawid G. Nowak. Clonal lineage tracing with somatic delivery of recordable barcodes reveals migration histories of metastatic prostate cancer. *Cancer Discovery*, 14(10):1990–2009, 2024.
- [23] Mengyang Chen, Ruijiang Fu, Yiqian Chen, Li Li, and Shou-Wen Wang. High-resolution, noninvasive single-cell lineage tracing in mice and humans based on dna methylation epimutations. *Nature Methods*, 22(3):488–498, 2025.
- [24] Tamara Prieto, Dennis J. Yuan, John Zinno, Clayton Hughes, Nicholas Midler, Sheng Kao, Jani Huuh-tanen, Ramya Raviram, Foteini Fotopoulou, Neil Ruthen, Srinivas Rajagopalan, Joshua S. Schiffman, Andrew R. D’Avino, Sang-Ho Yoon, Jesus Sotelo, Nathaniel D. Omans, Noelle Wheeler, Alejandro Garces, Barun Pradhan, Alexandre Pellan Cheng, Nicolas Robine, Catherine Potenski, Katharine God-frey, Nobuyuki Kakiuchi, Akira Yokoyama, Seishi Ogawa, Julian Abrams, Ivan Raimondi, and Dan A. Landau. Large-scale single-cell phylogenetic mapping of clonal evolution in the human aging esoph-agus. *bioRxiv*, 2025. Preprint.
- [25] Michelle M. Chan, Zachary D. Smith, Stefanie Grosswendt, Helene Kretzmer, Thomas M. Norman, Britt Adamson, Marco Jost, Jeffrey J. Quinn, Dian Yang, Matthew G. Jones, Alex Khodaverdian, Nir Yosef, Alexander Meissner, and Jonathan S. Weissman. Molecular recording of mammalian embryo-genesis. *Nature*, 570(7759):77–82, 2019.
- [26] Joseph Felsenstein. Phylogenies and the comparative method. *The American Naturalist*, 125(1):1–15, 1985.
- [27] L. L. Cavalli-Sforza and A. W. F. Edwards. Phylogenetic analysis: models and estimation proce-dures. *American Journal of Human Genetics*, 19(3 Pt 1):233–257, 1967. PMID: 6026583; PMCID: PMC1706274.
- [28] Joseph Felsenstein. Maximum-likelihood estimation of evolutionary trees from continuous char-acters. *American Journal of Human Genetics*, 25(5):471–492, 1973. PMID: 4741844; PMCID: PMC1762641.
- [29] Russell Lande. Natural selection and random genetic drift in phenotypic evolution. *Evolution*, 30(2):314–334, 1976.
- [30] Thomas F. Hansen. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5):1341–1351, 1997.
- [31] Marguerite A. Butler and Aaron A. King. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. *The American Naturalist*, 164(6):683–695, 2004.

- [32] Brian C. O’Meara, Cécile Ané, Michael J. Sanderson, and Peter C. Wainwright. Testing for different rates of continuous trait evolution using likelihood. *Evolution*, 60(5):922–933, 2006.
- [33] Trevor Bedford and Daniel L. Hartl. Optimization of gene expression by natural selection. *Proceedings of the National Academy of Sciences*, 106(4):1133–1138, 2009.
- [34] David Brawand, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harri-gan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, Frank W. Albert, Ul-rich Zeller, Philipp Khaitovich, Frank Grützner, Sven Bergmann, Rasmus Nielsen, Svante Pääbo, and Henrik Kaessmann. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348, 2011.
- [35] 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.
- [36] 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), 2023.
- [37] Hirotaka Matsumoto and Hisanori Kiryu. Scoup: a probabilistic model based on the ornstein–uhlenbeck process to analyze single-cell expression data during differentiation. *BMC Bioinformatics*, 17(1):232, 2016.
- [38] M. G. Hirsch, Soumitra Pal, Farid Rashidi Mehrabadi, Salem Malikic, Charli Gruen, Antonella Sas-sano, Eva Pérez-Guijarro, Glenn Merlino, S. Cenk Sahinalp, Erin K. Molloy, Chi-Ping Day, and Teresa M. Przytycka. Stochastic modeling of single-cell gene expression adaptation reveals non-genomic contribution to evolution of tumor subclones. *Cell Systems*, 16(1), 2025.
- [39] Hannah Stuart and Aaron McKenna. Scout: Ornstein–uhlenbeck modelling of gene expression evolu-tion on single-cell lineage trees. *bioRxiv*, 2025. Preprint.
- [40] Jiawei Xing, Stephen J. Staklinski, Zhihan Liu, Dawid Nowak, and Adam Siepel. Lineage-aware stochastic modeling reveals gene-expression dynamics in development and disease. *bioRxiv*, 2026. Preprint.
- [41] Xiling Wen, Stefanie Fuhrman, George S. Michaels, Daniel B. Carr, Susan Smith, Jeffery L. Barker, and Roland Somogyi. Large-scale temporal gene expression mapping of central nervous system de-velopment. *Proceedings of the National Academy of Sciences*, 95(1):334–339, 1998.
- [42] Saeed Tavazoie, Jason D. Hughes, Michael J. Campbell, Raymond J. Cho, and George M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22(3):281–285, 1999.
- [43] Eran Segal, Michael Shapira, Aviv Regev, Dana Pe’er, David Botstein, Daphne Koller, and Nir Fried-man. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166–176, 2003.

- [44] Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007.
- [45] Orly Alter, Patrick O. Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000.
- [46] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.
- [47] Xiaotong Fu, Dominik Klein, Egor Antipov, Alessandro Palma, Alejandro Tejada-Lapueta, Mojtaba Bahrami, Louis B. Kümmerle, Manuel Lubetzki, Francesco Paolo Casale, Malte D. Luecken, and Fabian J. Theis. Benchmarking gene expression reconstruction from single-cell latent representations. *bioRxiv*, 2026. Preprint.
- [48] K. Wang, R. Meng, Z. Hu, and D. Zhou. Towards a quantitative understanding of cellular dynamics via lineage tracing inference. *CSIAM Trans Life Sci*, 2(3):495–521, 2026.
- [49] Sydney M. Shaffer, Benjamin L. Emert, Raúl A. Reyes Hueros, Christopher Cote, Guillaume Harnage, Dylan L. Schaff, Ann E. Sizemore, Rohit Gupte, Eduardo Torre, Abhyudai Singh, Danielle S. Bassett, and Arjun Raj. Memory sequencing reveals heritable single-cell gene expression programs associated with distinct cellular behaviors. *Cell*, 182(4):947–959.e17, 2020.
- [50] Yongjie Lin, Xiangru Chen, Long Wu, Yutong Zhou, and Yihan Lin. Widespread transcriptional memory shapes heritable states and functional heterogeneity in cancer and stem cells. *Cell Systems*, 17(7):101621, 2026.
- [51] David DeTomaso and Nir Yosef. Hotspot identifies informative gene modules across modalities of single-cell genomics. *Cell Systems*, 12(5):446–456.e9, 2021.
- [52] William N. Colgan, Luke W. Koblan, JoAnne Villagrana, Tien-Chi Jason Hou, Minming Wang, Gokul Gowri, Whitney Chandler, Leonardo A. Sepúlveda, Didar Ciftci, Karina Smolyar, Alicia Young, Lars Wittler, Styliani Markoulaki, Kyle Loh, Xiaowei Zhuang, Nir Yosef, Zachary D. Smith, and Jonathan S. Weissman. Comprehensive lineage tracing maps the landscape of cell fate decisions in mouse embryogenesis. *bioRxiv*, 2026. Preprint.
- [53] Aden Forrow and Geoffrey Schiebinger. Lineageot is a unified framework for lineage tracing and trajectory inference. *Nature Communications*, 12(1):4940, 2021.
- [54] 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*, 2021. Preprint.
- [55] Shou-Wen Wang, Michael J. Herriges, Kilian Hurley, Darrell N. Kotton, and Allon M. Klein. Cospar identifies early cell fate biases from single-cell transcriptomic and lineage information. *Nature Biotechnology*, 40(7):1066–1074, 2022.

- [56] Kun Wang, Liangzhen Hou, Xin Wang, Xiangwei Zhai, Zhaolian Lu, Zhike Zi, Weiwei Zhai, Xionglei He, Christina Curtis, Da Zhou, and Zheng Hu. Phylovelo enhances transcriptomic velocity field mapping using monotonically expressed genes. *Nature Biotechnology*, 42(5):778–789, 2024.
- [57] Hannah M. Schlüter and Caroline Uhler. Integrating representation learning, permutation, and optimization to detect lineage-related gene expression patterns. *Nature Communications*, 16(1):1062, 2025.
- [58] Yuhong Wen, Jiankang Xiong, Fuzhou Gong, Liang Ma, and Lin Wan. Decoupling lineage and intrinsic information in single-cell lineage tracing data with deep disentangled representation learning. *bioRxiv*, 2026. Preprint.
- [59] Weixiang Fang, Claire M. Bell, Abel Sapirstein, Soichiro Asami, Kathleen Leeper, Donald J. Zack, Hongkai Ji, and Reza Kalhor. Quantitative fate mapping: A general framework for analyzing progenitor state dynamics via retrospective lineage barcoding. *Cell*, 185(24):4604–4620.e32, 2022.
- [60] Dian Yang, Matthew G. Jones, Santiago Naranjo, William M. Rideout, III, Kyung Hoi (Joseph) Min, Raymond Ho, Wei Wu, Joseph M. Replogle, Jennifer L. Page, Jeffrey J. Quinn, Felix Horns, Xiaojie Qiu, Michael Z. Chen, William A. Freed-Pastor, Christopher S. McGinnis, David M. Patterson, Zev J. Gartner, Eric D. Chow, Trevor G. Bivona, Michelle M. Chan, Nir Yosef, Tyler Jacks, and Jonathan S. Weissman. Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution. *Cell*, 185(11):1905–1923.e25, 2022.
- [61] Joshua S. Schiffman, Andrew R. D’Avino, Tamara Prieto, Yakun Pang, Yilin Fan, Srinivas Rajagopalan, Catherine Potenski, Toshiro Hara, Mario L. Suvà, Charles Gawad, and Dan A. Landau. Defining heritability, plasticity, and transition dynamics of cellular phenotypes in somatic evolution. *Nature Genetics*, 56(10):2174–2184, 2024.
- [62] Adriano Bolondi, Benjamin K. Law, Helene Kretzmer, Seher Ipek Gassaloglu, René Buschow, Christina Riemenschneider, Dian Yang, Maria Walther, Jesse V. Veenliet, Alexander Meissner, Zachary D. Smith, and Michelle M. Chan. Reconstructing axial progenitor field dynamics in mouse stem cell-derived embryoids. *Developmental Cell*, 59(12):1489–1505.e14, 2024.
- [63] Palash Sashittal, Richard Y. Zhang, Benjamin K. Law, Henri Schmidt, Alexander Strzalkowski, Adriano Bolondi, Michelle M. Chan, and Benjamin J. Raphael. Inferring cell differentiation maps from lineage tracing data. *Nature Methods*, 23(3):532–541, 2026.
- [64] William Howard-Snyder, Richard Zhang, Henri Schmidt, Michelle Chan, and Benjamin J. Raphael. Inferring cell differentiation dynamics with unobserved progenitors. *bioRxiv*, 2025. Preprint.
- [65] Weixu Wang, Zhiyuan Hu, Philipp Weiler, Sarah Mayes, Marius Lange, Daniel M. Fountain, Julianna O. Haug, Jingye Wang, Zhengyuan Xue, Tatjana Sauka-Spengler, and Fabian J. Theis. Regvelo: Gene-regulatory-informed dynamics of single cells. *Cell*, 189(12):3773–3800.e44, 2026.
- [66] D. N. Lawley and A. E. Maxwell. Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12(3):209–229, 1962.

- [81] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287, 2012.