Interpretable trajectory inference with single-cell Linear Adaptive Negative-binomial Expression (scLANE) testing

Authors: Jack R. Leary[1], and Rhonda Bacher[1,*]

Affiliations:

[1]Department of Biostatistics, College of Public Health and Health Professions, University of Florida, Gainesville, FL 32610, USA.

***Corresponding author:**

Rhonda Bacher, PhD, 2004 Mowry Rd, P.O. Box 117450 Gainesville, FL 326110. E-mail: rbacher@ufl.edu. TEL: (352) 294-5914. ORCID: 0000-0001-5787-476X

Abstract

The rapid proliferation of trajectory inference methods for single-cell RNA-seq data has allowed researchers to investigate complex biological processes by examining underlying gene expression dynamics. After estimating a latent cell ordering, statistical models are used to determine which genes exhibit changes in expression that are significantly associated with progression through the biological trajectory. While a few techniques for performing trajectory differential expression exist, most rely on the flexibility of generalized additive models in order to account for the inherent nonlinearity of changes in gene expression. As such, the results can be difficult to interpret, and biological conclusions often rest on subjective visual inspections of the most dynamic genes. To address this challenge, we propose scLANE testing, which is built around an interpretable generalized linear model and handles nonlinearity with basis splines chosen empirically for each gene. In addition, extensions to estimating equations and mixed models allow for reliable trajectory testing under complex experimental designs. After validating the accuracy of scLANE under several different simulation scenarios, we apply it to a set of diverse biological datasets and display its ability to provide novel biological information when used downstream of both pseudotime and RNA velocity estimation methods.

Introduction

In contrast to classical time-series sequencing experiments, single-cell RNA-sequencing (scRNA-seq) allows researchers to investigate progressive changes in gene expression without necessarily collecting samples across multiple timepoints[1–3]. Computational methods infer cellular trajectories assuming that snapshot experiments capture sufficient cellular heterogeneity representing progression through the biological process of interest. Trajectory inference can be broadly divided into two stages, with the first being the estimation of pseudotime - the latent cellular ordering that approximates progression through an underlying process whereby cells occupying initial states are assigned low pseudotime values, and vice versa for cells that have reached terminal states. Cells may also be divided into multiple lineages, in which the topology of the biological process separates into related but distinct subprocesses, as cellular orderings are not necessarily linear. After pseudotime values are assigned, trajectory differential expression methods seek to identify which genes exhibit dynamics that are significantly associated with pseudotime (**Fig. 1A-B**). Hypothesis tests of this form rest upon the assumption that identifying trajectory differentially expressed (DE) genes may shed light on how the biological process itself functions. Further downstream analysis then aims to compare gene dynamics between lineages or experimental conditions.

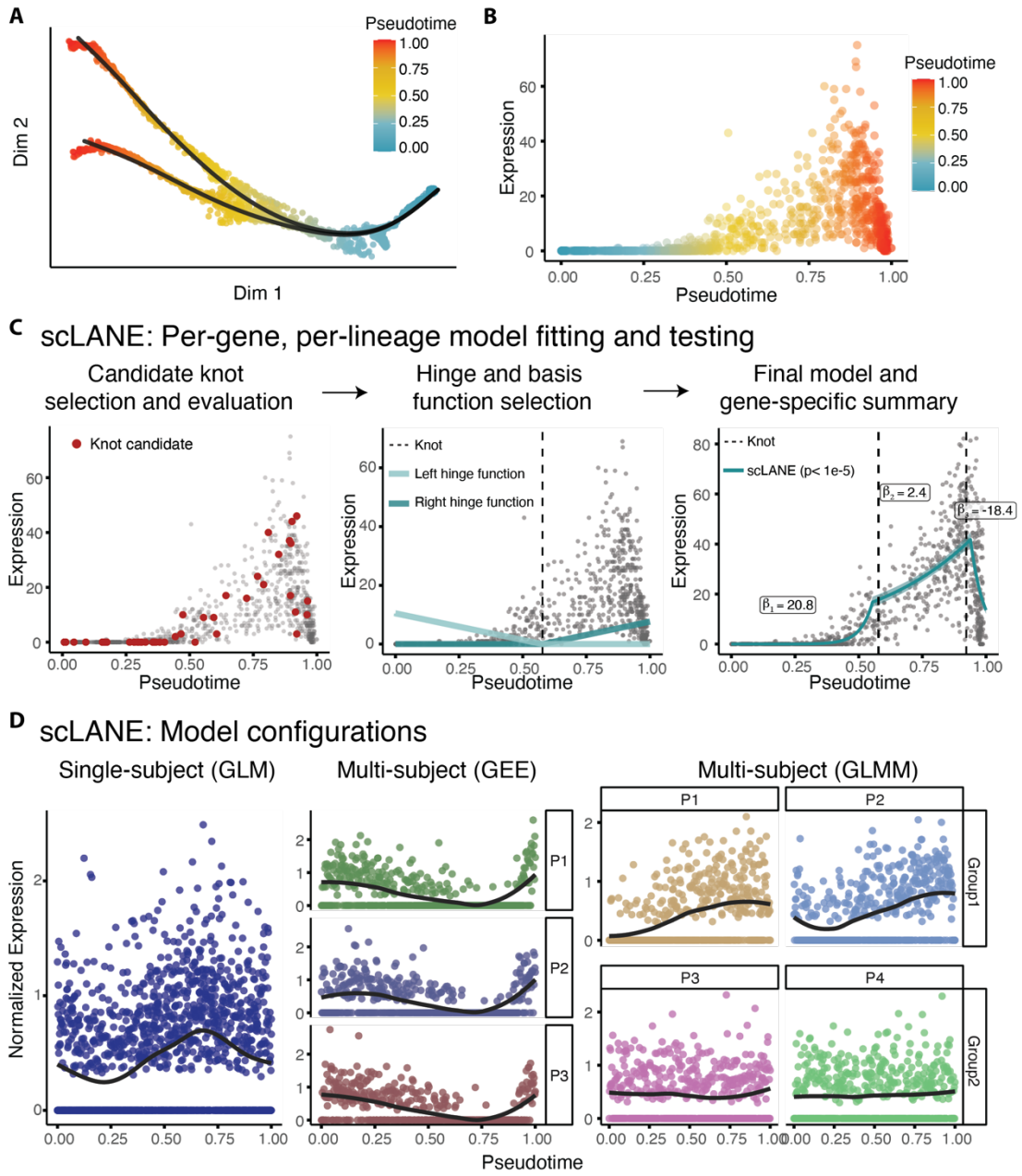

Figure 1: Overview of the scLANE trajectory differential expression workflow. (A) Trajectory (in gray) and pseudotime estimates in dimension-reduced space are shown for the zebrafish scRNA-seq dataset[4]. **(B)** Scatterplot of pseudotime versus gene expression for a single dynamic gene. **(C)**. Overview of the scLANE modeling procedure. First, candidate knots (highlighted in red) are chosen and scored. Next, left and right hinge functions at each knot are evaluated sequentially, and an optimal combination is selected. Gene expression is then modeled as a linear combination of hinge functions, and the final scLANE model is compared to a null model using a likelihood ratio test. Model outputs include fitted dynamics with 95% confidence intervals, knot locations, and coefficients for each segment. **(D)** Simulated gene expression dynamics demonstrating the use cases for the GLM, GEE, and

GLMM modes of scLANE, from left to right. Abbreviations: generalized linear model (GLM), generalized estimating equations (GEE), generalized linear mixed model (GLMM).

On the surface, methods that exist for analyzing traditional time-series RNA-seq experiments may seem appropriate. However, single-cell trajectory DE testing is complicated by a number of factors including nonlinearity, sparsity, and interpretability. Several single cell-specific methods have been proposed to address these issues[4–9]. For instance, tradeSeq[4] uses negative-binomial generalized additive models (GAMs) to model gene expression as a smooth function of pseudotime. Several forms of hypothesis testing are implemented, such as per- and cross-lineage differential expression. However, this approach does not account for population structure in datasets composed of cells from multiple subjects, which are becoming more common[10]. This concern was considered in the recent development of Lamian[5], which uses B-splines to model gene expression as a nonlinear function of pseudotime while including random effects per-subject through the estimation of a subject-level covariance matrix. While including random effects for each subject helps to control the false discovery rate, the model architecture suffers from the same difficulty of interpretation as methods built around GAMs[11,12]. In addition, Lamian doesn't provide subject-specific inference. Given that scRNA-seq analyses are often exploratory in nature, both researchers and clinical practitioners stand to benefit from statistics that offer a clearer biological meaning.

The scGTM method[9] attempted to address the issue of interpretability by classifying dynamic genes into three broad patterns: monotone, hill-shaped, and valley-shaped. However, interpretation is restricted by the need to classify trends as belonging to the pre-defined categories. In addition, between-lineage comparisons are limited in terms of inference beyond

visual comparison[9]. To fill this methodological gap, we present scLANE testing, a negative-binomial generalized linear model (GLM) framework for modeling nonlinear relationships while accounting for correlation structures inherent to multi-sample scRNA-seq experiments. scLANE can be used downstream of any pseudotemporal ordering or RNA velocity estimation method, making it easy to add to existing analysis pipelines. Truncated power basis splines are chosen empirically per-gene, per-lineage, providing results that are specific to each gene's dynamics across each biological subprocess - an improvement on methods that use a common number of equidistant knots for all genes (**Fig. 1C**)[4,5,7,8]. The coefficients generated by scLANE carry the same multiplicative interpretation as any GLM, providing a quantitative measure and significance test of the relationship of pseudotime with gene expression over empirically selected pseudotime intervals from each lineage (**Fig. 1C**). By leveraging generalized estimating equations (GEEs) and generalized linear mixed models (GLMMs), scLANE is able to accommodate more complex experimental designs and provide both population-level and subject-specific effect sizes for hypothesis tests under a variety of different correlation structures (**Fig. 1D**). Lastly, scLANE easily handles multiple lineages and provides quantitative comparisons of gene dynamics across lineages. Our software package implementing scLANE in R is free, open source, and available for download at https://github.com/jr-leary7/scLANE.

Results

scLANE outperforms existing methods in simulation studies

To first validate scLANE's ability to accurately classify genes as dynamic or static, we generated a total of 764 simulated datasets under three different experimental design setups. For single-subject simulations, we compared the GLM mode of scLANE to tradeSeq[4]. Multi-subject

simulations were split into two categories: those with gene dynamics mostly conserved between subjects, and those with multiple heterogeneous subject groups. These two setups correspond to the use cases recommended for the GEE and GLMM modes of scLANE, respectively.

On single-subject simulated datasets, scLANE compares favorably to tradeSeq, exhibiting equal or higher performance on classification metrics such as balanced accuracy (**Fig. 2A**). Memory usage is also significantly lower for scLANE across all simulations regardless of the number of cells simulated (**Fig. 2B**). The largest datasets we simulated were composed of 5,000 cells; on those datasets scLANE used an average of just 0.33 gigabytes of memory. The difference between methods with respect to runtime was more complex; while scLANE processed more genes per minute on smaller datasets with less than 1,000 cells, tradeSeq was slightly faster on the largest datasets, with tradeSeq and scLANE processing a median 59 and 25 genes per minute, respectively (**Fig. 2C**). The GEE and GLMM modes of scLANE also performed well in terms of balanced accuracy, even on datasets with complex, subject-specific trajectory dynamics (**Fig. 2D-G**). As expected, datasets with larger numbers of cells resulted in better classification of truly dynamic genes. In addition, multi-subject datasets where trajectories were more conserved between subjects resulted in better GEE and GLMM mode performance (**Fig. 2E,G**).

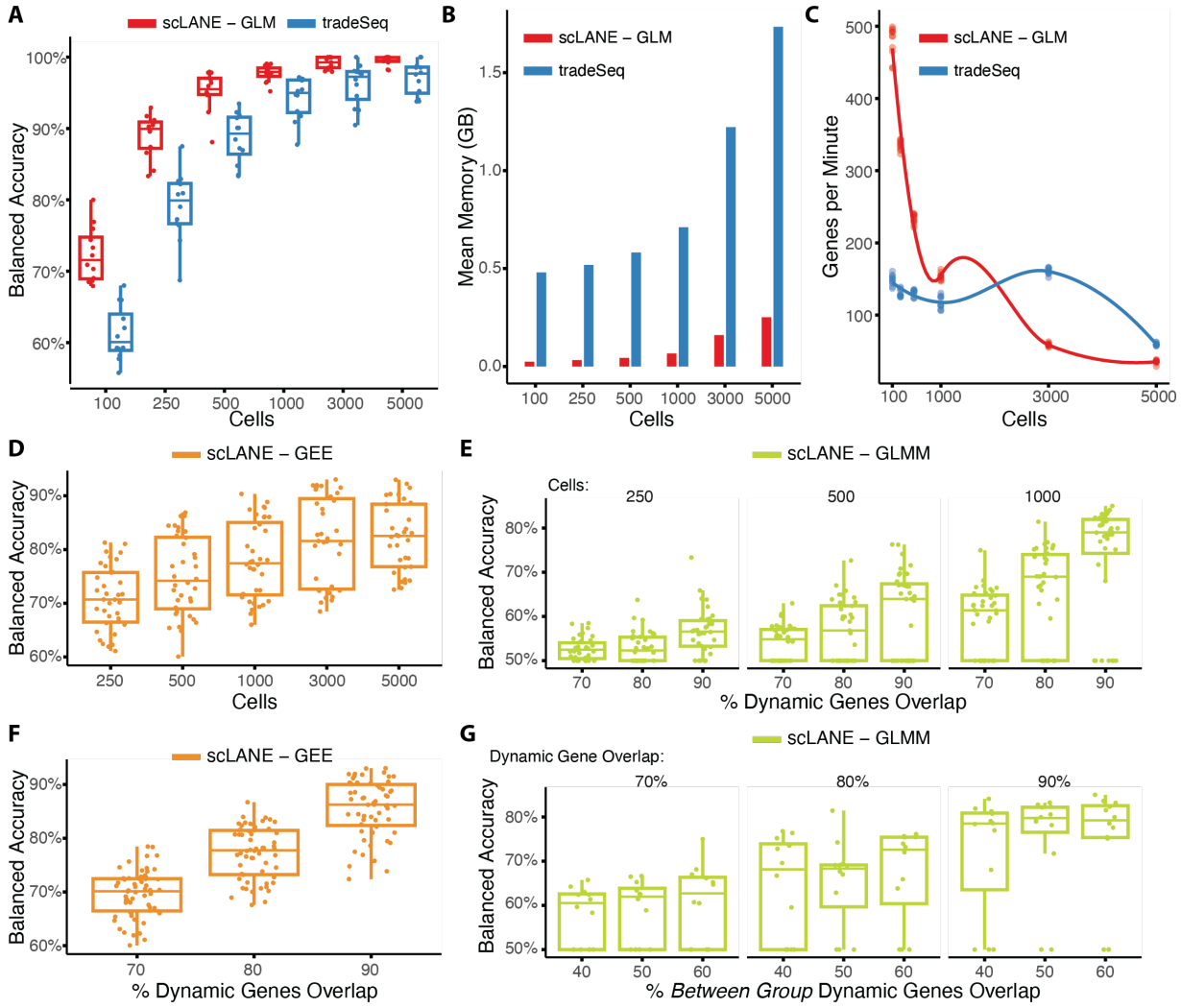

Figure 2: Simulation study results suggest that scLANE efficiently and accurately identifies truly-dynamic genes. (A) Boxplots of balanced accuracy split by reference dataset show that scLANE has consistently higher median balanced accuracy than tradeSeq. **(B)** Barplots displaying the mean memory usage in gigabytes for scLANE and tradeSeq split by simulated dataset size. scLANE uses much less memory than tradeSeq regardless of dataset size. **(C)** A scatterplot of dataset size versus runtime as measured by genes processed per minute for scLANE and tradeSeq. scLANE is faster on smaller datasets, while tradeSeq performs better on medium and large datasets, though as dataset scale increases the runtime differential gets smaller. **(D)** Boxplots of balanced accuracy split by the number of cells for the GEE mode. **(E)** Boxplots of balanced accuracy split by the number of cells (panels) and percent of dynamic genes shared between individuals (x-axis) for the GLMM mode. **(F)** Boxplots of balanced accuracy split by percent of dynamic genes shared between individuals for the GEE mode. **(G)** Boxplots of balanced accuracy split by percent of dynamic genes shared between individuals (panels) and percent of dynamic genes shared between groups (x-axis) for the GLMM mode.

Case studies: scLANE enables biological interpretations

Next, we demonstrate scLANE's ability to generate biological insight on three scRNA-seq datasets: murine pancreatic endocrine cell specification, zebrafish axial mesoderm development, and human fetal liver hematopoiesis. Processing details are described in Methods and in Supplementary Materials.

Transcription factor dynamics in murine pancreatic endocrinogenesis

We applied scLANE to a well-annotated dataset composed of 3,696 *Mus musculus* cells during pancreatic endocrinogenesis (**Fig. 3A; Suppl. Fig. 1**)[13]. While this dataset is widely used in trajectory method benchmarking and computational vignettes, the biological properties of the dataset are often left unexplored beyond recapitulating the original authors' analysis. This presented an opportunity to showcase the biological utility of scLANE. After estimating an RNA velocity-based latent time (**Methods**), scLANE identified a set of 2,478 trajectory DE genes using GLM mode on the spliced mRNA counts as the response (**Fig. 3B; Suppl. Fig. 1A&D**).

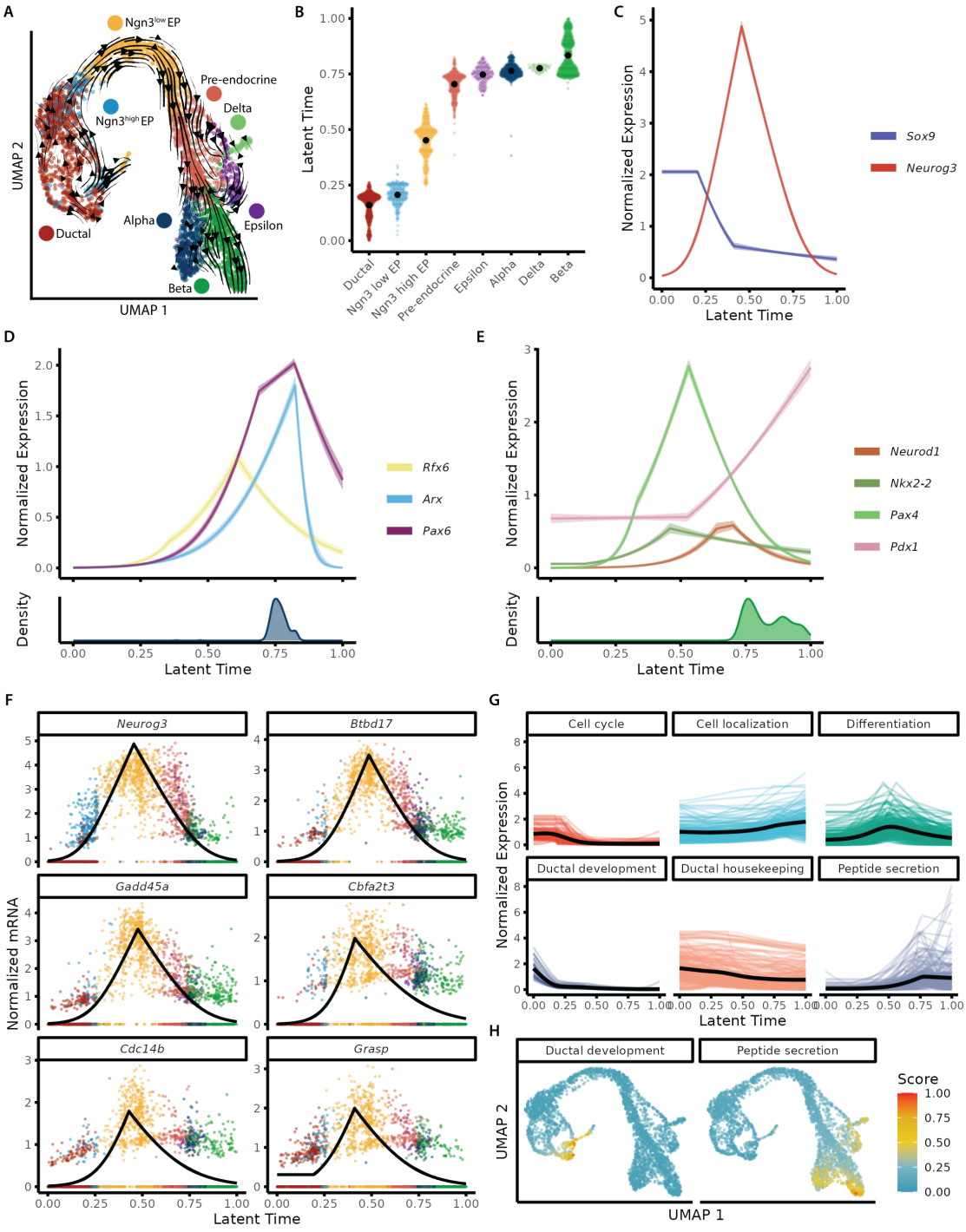

Figure 3: scLANE identifies biologically meaningful and specific gene programs regulating endocrinogenesis. (A) Streamline plot of differentiation direction vectors projected onto the UMAP embedding. **(B)** Beeswarm plot displaying the distribution of latent time estimates within each celltype. The mean value per-celltype is displayed in black. **(C)** scLANE fitted dynamics of transcription factors that act as master regulators of endocrine differentiation. **(D)** Fitted dynamics of genes regulating alpha cell development across latent

time, with the distribution of alpha cells in latent time shown below. **(E)** Fitted dynamics of genes regulating beta cell development as in **(D)**. **(F)** Fitted dynamics of the five nearest-neighbors of *Neurog3* across latent time. **(G)** Fitted dynamics of annotated gene programs across latent time. The overall trend per-program is shown in black via a LOESS fit. **(H)** Annotated gene program module scores for the ductal development and peptide production gene programs displayed on the UMAP embedding.

The results displayed in Figure 3 showcase scLANE's ability to recapitulate extant biological knowledge while adding interpretability. It is well-known that the transcription factor neurogenin-3 (*Neurog3*) is a primary regulator of pancreatic endocrine cell differentiation and maturation [13–15]. Itself enhanced by *Sox9* expression, *Neurog3* expression guides specification of alpha and beta cells during the pre-endocrine stage[14,16]. We see this in the dynamics of both genes, where *Sox9* is active at the beginning of latent time, whereas *Neurog3* is upregulated until midway through the trajectory (knot = 0.455) before decreasing (slope = -13.812) (**Fig. 3C**). Commitment to the alpha cell phenotype is partially driven by expression of the transcription factors *Arx*, *Rfx6*, and *Pax6*, all of which can be seen to experience peak expression near (in latent time) to where the alpha cell population is located (**Fig. 3D**). *Arx* is a target of *Neurog3*, which scLANE identifies correctly by placing a peak expression knot later in latent time (knot = 0.824) for *Arx* than for *Neurog3* (**Fig. 3C**). In addition, upregulation of *Arx* suppresses expression of *Pax4* and vice versa, which can be seen in the dynamics of *Arx* where extremely swift downregulation (slope = -43.042) occurs in latent time (knot = 0.824) just prior to the appearance of a large group of beta cells (mean latent time = 0.833) (**Fig. 3D-E**).

scLANE fitted dynamics also help to explain beta cell development. *Pax4* expression is induced by *Neurog3* and is essential for commitment to the beta cell fate[14]. The transcription factor NK homeobox 2 (*Nkx2-2*) (peaking at knot=0.459) activates *Neurod1*, upregulation of which helps to produce beta cells, as shown by its peak (knot = 0.702) just prior to beta cell emergence[14]. Finally, the transcription factor *Pdx1* directly targets the insulin gene *Ins*, insulin being the

primary peptide hormone secreted by pancreatic beta cells, exhibits a knot in latent time prior to the appearance of any beta cells (knot = 0.526), after which expression switches from being static to increasing (**Fig. 3E**). This is reflected in the estimated coefficients before (0.076) and after (5.642) the knot[14,17].

As the regulation of alpha versus beta cell fate specification by competing expression of *Arx* and *Pax4* is not completely understood, we sought to determine genes with similar dynamics to *Neurog3*, which activates both transcription factors[14]. After embedding the gene dynamics in a *k*-nearest neighbors graph, we identified the five genes most similar to *Neurog3*, all of which had peak expression around the same location in latent time (mean knot=0.442) and similar increasing trends prior to it (mean slope = 12.816) (**Fig. 3F**). One of the top five nearest neighbors was the cell cycle inhibitor *Gadd45a*, previously identified as upregulated in endocrine progenitors and associated with endocrine cell fate specification[13]. However, the most-similar gene, *Btbd17*, has no known association with endocrine cell fate determination, although it was previously shown to be differentially expressed between endocrine progenitor cell types[16]. In addition, *Btbd17* is a validated target of *Neurog3* in the human pancreas[18]. Finally, *Cbfa2t3* has not been previously shown to be involved in endocrinogenesis, but is a well-validated transcriptional corepressor during hematopoiesis, and also promotes myeloid differentiation during acute myeloid leukemia[19–21].

Lastly, we sought to identify gene programs that were interpretable biologically. After clustering the scLANE gene dynamics, we used pathway analysis to examine the gene ontology biological process (GO:BP) terms that each cluster of genes was significantly enriched for (**Fig. 3G-H**). We annotated the six programs, several of which were especially relevant to endocrinogenesis (**Fig.**

159 **3G; Suppl. Fig. 1E**). These programs highlight the biologically-relevant downstream analysis
160 utilities offered by scLANE in addition to interpretable gene dynamics.

161 **Multiple lineages: Zebrafish musculoskeletal system development**

162 We next applied scLANE to a zebrafish embryogenesis dataset having a multi-lineage topology,
163 with the goal of modeling differences between two branching lineages of axial mesoderm
164 development[22]. As shown in Figure 4, we obtained two lineages matching the authors' original
165 celltype annotations by estimating pseudotime with Slingshot[22,23] (**Fig. 4A-C**). We generated
166 trajectory DE gene sets for both the notochord and prechordal plate lineages by running scLANE
167 with the default settings in GLM mode. Building lineage-specific models for the top 3,000 highly
168 variable genes and 2,341 cells with 8 threads took 5.05 minutes on a high-performance
169 computation node allocated 64 gigabytes of memory. By examining the distribution of scLANE-
170 identified knots for each lineage, we discovered that the prechordal plate lineage exhibits a group
171 of knots earlier in pseudotime (approx. knot mean = 0.193) than the corresponding cluster in the
172 notochord lineage (approx. knot mean = 0.251) (**Fig. 4D**). This could indicate that fate
173 specification for the prechordal plate cells occurs earlier, since the transcriptional switches were
174 located prior to the branching of the two lineages.

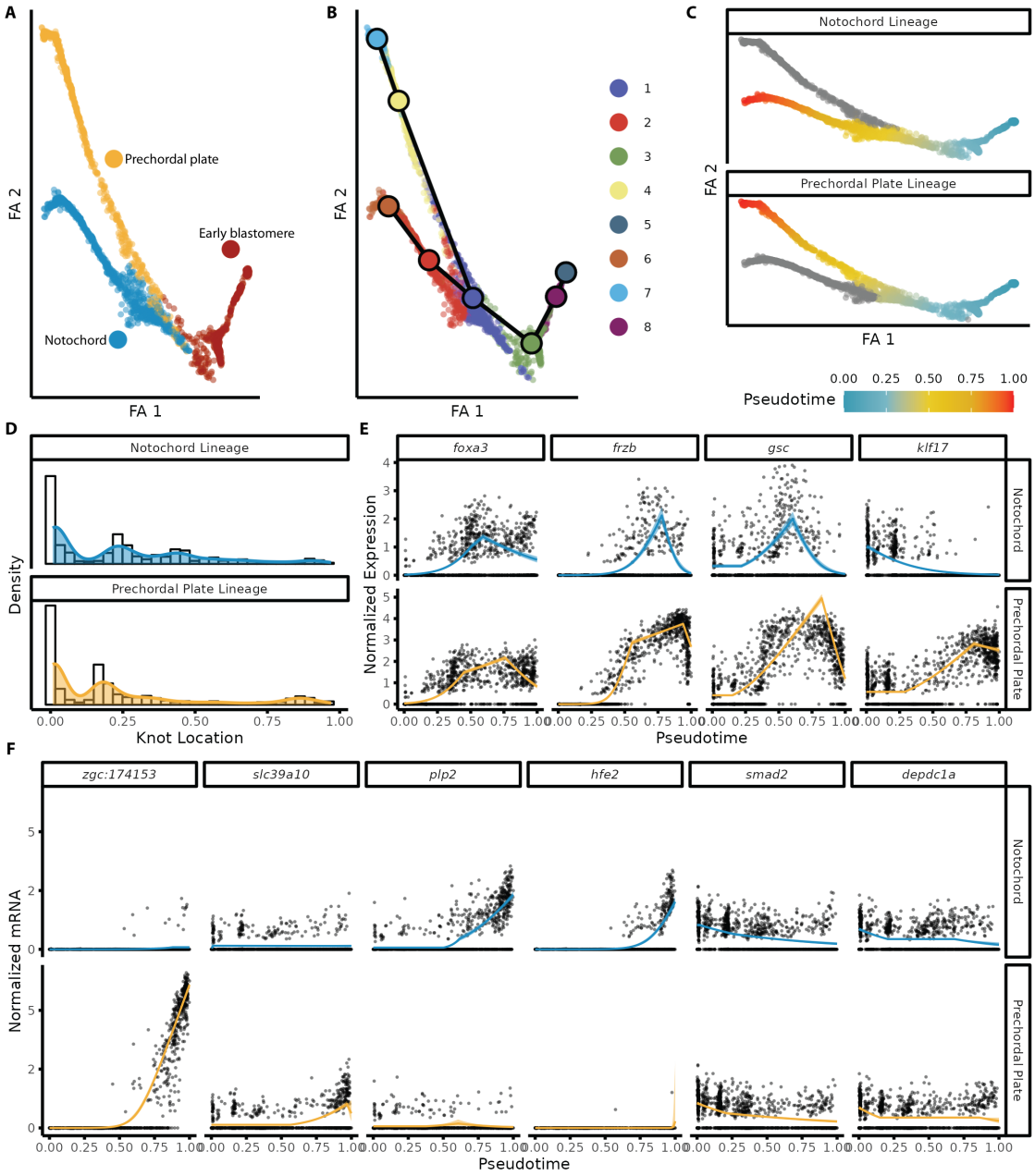

Figure 4: scLANE provides a finer-resolution characterization of zebrafish embryogenesis. (A) Force-directed graph embedding of zebrafish axial mesoderm cells colored by celltype. **(B)** The minimum spanning tree estimated by Slingshot prior to pseudotime estimation with cells colored by cluster. **(C)** Pseudotime estimates for the notochord and prechordal plate lineages. Pseudotime has been min-max scaled within each lineage. **(D)** Empirical density estimates of the distribution of knots chosen by scLANE across each lineage. **(E)** Fitted gene dynamics from scLANE for four transcription factors relevant to prechordal plate development. The four genes are significantly dynamic across the notochord lineage as well. **(F)** Fitted gene dynamics for selected genes of interest that have trends that are

either unique to the prechordal plate lineage, unique to the notochord lineage, or shared across both.

The original authors identified four transcription factors of interest exhibiting significant changes across the prechordal plate lineage: *foxa3*, *frzb*, *gsc*, and *klf17*. Of these, only *foxa3* was originally designated as trajectory DE over the notochord lineage as well[22]. In contrast, scLANE designated all four genes as significantly trajectory DE across both lineages (**Fig. 4E**). Genes *frzb*, *gsc*, and *klf17* exhibited concave trends across both lineages, and for all three genes their peak expression occurred much earlier in the notochord lineage than the prechordal plate lineage (mean difference = 0.21). Gene expression of *klf17* initially decreased for both lineages, however it continued to gradually decrease in the notochord lineage (slope = -4.564), while shifting and increasing along the prechordal plate lineage (slope = 5.761) (**Fig. 4E**).

We next identified genes that were uniquely dynamic across each of the two lineages, as well as genes that exhibited markedly similar dynamics across lineages by examining the ratios of the lineage-specific test statistics (**Fig. 4F**). Genes *zgc:174153* and *slc39a10* exhibited dynamics specific to the prechordal plate lineage, however neither gene has an annotated role in axial development; *zgc:174153* is associated with immune response, and *slc39a10* with the epithelial to mesenchymal transition[24]. Genes such as these without well-defined functionalities in specific biological processes like axial development showcases the ability of scLANE to provide novel, lineage-specific and gene-specific biological information. Two genes significantly trajectory DE in the notochord lineage but not the prechordal plate lineage were *hfe2* and *plp2*. While *hfe2* is a known biomarker for notochord cells, the function of *plp2* is less clear[24]. Its dynamic expression during notochord development indicates that it might play a regulatory role unique to the differentiation of notochord cells. Lastly, the genes *smad2* and *depdc1a* had highly similar dynamics across both lineages and are each involved in the development of anatomical structures

197 such as axis elongation[24]. This more general functionality aligns with the congruence of the
198 genes' dynamics across both lineages.

199 **Scaling up: Monocyte development in human fetal liver cells**

200 Next, in order to showcase the efficiency and scalability of scLANE we performed trajectory DE
201 testing on a human fetal liver hematopoiesis dataset consisting of 38,464 cells from 14 fetal
202 samples[25]. After preprocessing and pseudotime estimation, we obtained a single lineage
203 beginning with hematopoietic stem cells and ending in Kupffer cells, with several intermediate
204 monocyte and macrophage phenotypes (**Fig. 5A**). Since human monocytes are relatively well-
205 annotated and their development is largely understood, this dataset served as a confirmatory
206 example of scLANE's ability to generate valid biological insights at scale. The underlying
207 trajectory structure was composed of a single lineage, upon which we ran scLANE using the
208 GLM mode due to the size of the dataset and identified 1,404 trajectory DE genes with highly
209 heterogeneous functionality. The original publication focused on the differences in expression of
210 dynamically-regulated genes between tissues, specifically the dynamics of the canonical
211 monocyte marker genes *S100A8/A9/A12* and *FCGR1A/2A*[25]. ScLANE recapitulated the
212 trajectory dynamics of these marker genes **(Suppl. Fig. 3F),** however, since their importance to
213 myeloid development and various types of immune response is well-known, we mainly focused
214 on a set of transcription factors known to be regulators of commitment to the macrophage cell
215 fate (**Fig. 5B**).

216 Among the most highly dynamic transcription factors identified by scLANE and not discussed in
217 the original publication were *EGR1*, *KLF4*, *NR4A2,* and *KLF2*. Early-growth response (*EGR1*) is
218 an essential regulator of macrophage differentiation, both promoting expression of macrophage

genes and suppressing expression of neutrophil-lineage genes, as supported by its static expression until the monocyte precursor stage (knot = 0.280)[26–29]. Similarly, the Kruppel-like factor 4 (*KLF4*) has a well-understood role in monocyte differentiation and macrophage activation[30,31]. The other two transcription factors: nuclear receptor 4A subfamily 2 (*NR4A2*) and Kruppel-like factor 2 (*KLF2)* do not have an annotated roles in differentiation into the macrophage phenotype, despite having similar dynamics to *EGR1* and *KLF4*. *NR4A2* has been shown in mice to regulate neuronal differentiation and be expressed in monocytes[32,33], and *KLF2* regulates both myeloid immune response and quiescence of CD4+ and CD8+ T-cells[34,35]. All four transcription factors showed steady increases in expression (mean slope = 4.15) over pseudotime until almost exactly the same point (0.892, 0.897, 0.898, and 0.894, respectively), after which expression decreased rapidly (mean slope = -16.89). This indicates that *NR4A2* and *KLF2* might be necessary for commitment to the macrophage cell fate in addition to *EGR1* and *KLF4* or they may potentiate the activity of other regulators.

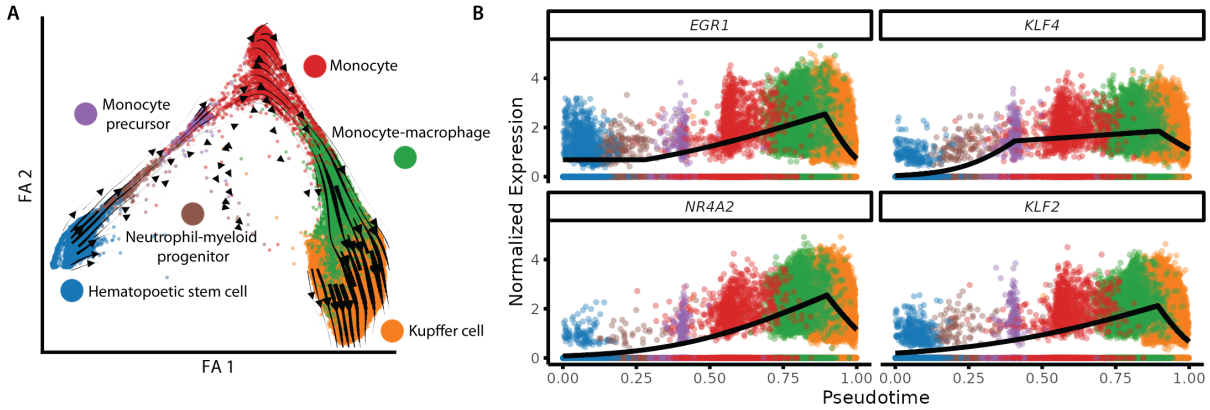

Figure 5: scLANE reveals biologically-meaningful transcription factors. (A) UMAP embedding of cells undergoing myelopoiesis. Differentiation directions are overlaid as streamlines. **(B)** Gene dynamics scatterplots for four transcription factors of interest. The fit from scLANE is overlaid in black.

233 **Discussion**

234 A core difficulty in the utilization of scRNA-seq experiments is the complexity of interpreting
235 downstream analysis results. For trajectory analyses, methods too often rely on visual inspection
236 of trends and subjective assessments of whether genes are upregulated, downregulated, or static
237 at any given point in pseudotime. Results are often used to confirm pre-existing notions of
238 marker gene behavior. To fully extract new knowledge from single cell data, it is imperative that
239 downstream analysis methods carry a high degree of interpretability in addition to providing
240 quantitative rather than qualitative results. We developed scLANE to address this by providing
241 linear regression-based effect sizes with simple, multiplicative interpretations that any scientist
242 can use to quantify the degree to which genes change during a biological process. Our
243 framework is modular and can be easily extended to accommodate complex experimental
244 designs while remaining user-friendly.

245 Several challenges remain, in particular with respect to computational efficiency. scRNA-seq
246 datasets have dramatically increased in size during the past five years, and modeling atlas-scale
247 trajectory dynamics is still a time-consuming and resource-heavy endeavor[10,36]. While we
248 employed several memory conservation and approximation techniques when developing
249 scLANE, fitting dynamics for larger datasets such as the fetal liver hematopoiesis dataset -
250 composed of 38,464 cells - took around 5 hours on a compute cluster even when using four
251 threads and 96 gigabytes of memory to test the top 2,000 highly variable genes. As such,
252 reducing computation time is an active area of research, and should be considered a primary
253 concern when improving single cell methods.

254 In applying our method to a wide variety of developmental datasets, both with respect to species
255 and experimental design, we demonstrated scLANE's flexibility and ease of relevant biological

interpretations. Our extensive simulation study provided a set of benchmarks that point toward scLANE's ability to accurately and precisely classify genes as dynamic or static under a variety of pseudotime structures, while providing quantitative outputs for each gene's dynamics. In applying our method to three scRNA-seq case studies, we have displayed scLANE's capacity to not only recapitulate known biological information by estimating gene trajectory dynamics, but also in identifying gene-programs and enabling pathway analysis. In our analysis of the endocrinogenesis dataset, scLANE identified genes with shared dynamics to that of the endocrinogenesis activator and master regulator *Neurog3,* revealing under-explored candidate genes that warrant more in-depth study. The analysis of a multi-lineage zebrafish dataset using scLANE revealed its ability to characterize lineage-specific transcriptional switches controlling cell fate specification. Finally, our investigation of fetal liver hematopoiesis demonstrated scLANE's capability of handling large datasets and generating dynamic gene programs involved in the differentiation of myeloid cells, the understanding of which is relevant for the study of immune responses.

Methods

Modeling expression over pseudotime

The scLANE method is an extension of the MARGE modeling framework, itself an extension of multivariate adaptive regression splines (MARS) to generalized estimating equations (GEEs)[37,38]. While multiple types of models are supported, the testing procedure is modular and each modeling mode is conceptually similar to the others. The only required inputs are a matrix of unnormalized expression counts, a matrix of pseudotime values with cells as rows and columns as lineages, a vector of library size factors, and a vector of subject IDs if cells were taken from multiple subjects[39]. In order to reduce runtime, model fitting is parallelized across genes using the foreach and doParallel R packages[40,41]. Memory usage is kept to a minimum by storing the counts as a read-only temporary file[42]. The basic mode-agnostic steps in scLANE testing as performed in the **testDynamic()** function are shown in **Algorithm 1**.

Algorithm 1 scLANE testing

---

for genes = 1 to n_g **do**

  for lineages = 1 to n_l **do**

    Identify optimal pseudotime knots and basis functions

    Fit negative binomial model using appropriate optimization routine

    Fit intercept-only, null model

    Compare fitted and null models with appropriate hypothesis test

    Estimate p-value

end for

end for

Collate coefficients, fitted values, and test results for all models

Adjust *p*-values for multiple hypothesis testing

The first step in scLANE testing is the selection of optimal knots across lineage-specific pseudotime, which is performed - with some slight modifications - as previously described[37]. Briefly, from the set of distinct pseudotime values a smaller number of candidate knots is generated by removing points close to the boundaries and requiring a minimum distance between neighboring knots. This is followed by subsampling if the number of candidate knots exceeds a preset maximum, 50 by default. Next, a generalized cross validation (GCV) criterion is used in a forward selection process to select optimal knots and basis functions up to a specified maximum number of covariates, after which a Wald information criterion (WIC) is used to prune the model during a backward selection procedure[37,43]. The final model is then fit using the desired mode, and a test of differential expression over the specified lineage is performed. For the GLM and GLMM frameworks a likelihood ratio test (LRT) is used, whereas for the GEE mode a global Wald test based on the sandwich variance-covariance matrix is used[44–47]. For both types of hypothesis test, the statistic is assumed to be asymptotically Chi-squared distributed with degrees of freedom varying based on the mode used. After all models have been fit, *p*-values are adjusted using Holm's correction, which is uniformly more powerful than the commonly-used Bonferroni correction[48].

Each truncated power basis function can take one of two directions, left and right. By default, all values of the power parameter p are set to 1 for the sake of interpretability; truncated power basis functions of this form as displayed in Equation 1 are known as hinge functions[38].

$$
\begin{aligned}
B_{\text{left}}(x) &= \begin{cases} (k-x)^p & k-x>0 \\ 0 & k-x \leq 0 \end{cases} \\
B_{\text{right}}(x) &= \begin{cases} (x-k)^p & x-k>0 \\ 0 & x-k \leq 0 \end{cases}
\end{aligned} \tag{1}
$$

After identifying the fully-specified set of basis functions, coefficients for each are estimated via iteratively-reweighted least squares[46,49]. We refer to this approach as the GLM mode, which fits a negative-binomial model with the canonical log link function using the MASS R package[50]. The mean on the link scale is thus the sum of the p chosen basis functions weighted by their respective estimated coefficients plus the log of an offset as shown in Equation 2. By default we formulate the offset by dividing the sequencing depth of each cell by 10,000 to match the counts-per-10k normalization used by Seurat, though the user can easily specify their own offset[51].

$$
\begin{aligned}
\mu &= ne^{\beta \mathbf{X}} \\
\eta &= \log(\mu) \\
&= \log(n) + \beta_0 + \sum_{j=1}^{p} \beta_j B_j(x) \\
\mathrm{Var}(Y|\mathbf{X}) &= V(\mu)\phi
\end{aligned} \tag{2}
$$

For studies with multiple subjects in which gene dynamics can reasonably be assumed to be conserved across subjects we developed an extension to the GLM mode based on estimating equations. While the MARGE method extended the classical MARS algorithm to the multi-subject domain, it did not support negative-binomial GEEs[37]. The architecture of the GEE mode

is similar to that of the GLM mode, with the estimating function shown in Equation 3 being optimized via Fisher scoring[37,45,52,53]. The mean μ is defined with an offset as above in Equation 2, and an estimate of the overdispersion parameter θ is obtained using the method of moments as implemented in the MASS package[50]. Finally, model fitting is performed via the geeM R package, an efficient implementation of estimating equations that supports the negative-binomial distribution with a known overdispersion parameter[52]. Sandwich variance estimators are used to provide standard errors for the estimated coefficients[44].

$$\begin{aligned} U(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^{\mathbf{T}} \left(\mathbf{A}_i^{1/2}\mathbf{R}(\alpha)\mathbf{A}_i^{1/2}\right)^{-1} \frac{Y_i - \mu_i}{\phi} \\ \mathbf{R}(\alpha) &= \mathrm{Corr}(Y_i) \\ \mathbf{A}_i &= \mathrm{diag}\big(V(\mu_{i1}), \dots, V(\mu_{it})\big) \end{aligned} \tag{3}$$

While multiple correlation structures are supported, the default is the autoregressive (AR1) structure, which is preferred when observations that are closer in time can be assumed to be more highly correlated[53,54]. In our case, since pseudotime is mostly driven by transcriptomic similarity it follows that observations near one another in pseudotime will have similar measurements of expression. The estimator of the AR1 correlation parameter α for an autoregressive process with order k is given by Equation 4. The null model is fit using the same correlation structure as the full model, and a Wald test provides a statistic quantifying the significance of the full model[44,52].

$$\begin{aligned} \hat{r}_{it} &= \frac{y_{it} - \hat{\mu}_{it}}{\sqrt{V(\hat{\mu}_{it})}} \\ \hat{\boldsymbol{\alpha}} &= \hat{\phi}^{-1} \sum_{i=1}^{n} \left(\frac{\sum_{t=1}^{n_i-0} \hat{r}_{i,t}\, \hat{r}_{i,t+0}}{n_i}, \dots, \frac{\sum_{t=1}^{n_i-k} \hat{r}_{i,t}\, \hat{r}_{i,t+k}}{n_i} \right) \\ \mathrm{Corr}(y_{it}, y_{is}) &= \alpha^{|t-s|} \end{aligned} \tag{4}$$

Lastly, when subjects are taken from multiple groups or it can be assumed that gene dynamics will differ in meaningful ways between subjects, we turn to mixed models in order to estimate population-averaged and subject-specific trends. The GLMM mode relies on the glmmTMB R package to fit the underlying models, as it supports the negative-binomial distribution[55]. For each subject a small number of optimal knots is selected as described previously, after which the basis functions are recomputed for all subjects simultaneously. Pruning is performed to ensure that the chosen per-subject basis functions are not highly collinear. The remaining basis functions are then used to fit a GLMM with a random intercept for each subject in addition to per-subject random slopes on each basis function[55,56]. As in the GLM and GEE mode, a per-cell offset is included. The general form of the GLMM framework is shown in Equation 5, with **X** being the matrix of basis functions and **Z** the matrix of random effects.

$$\begin{aligned} \mu &= e^{n\mathbf{X\beta}+\mathbf{Z\gamma}} \\ \eta &= \log(\mu) \\ &= n\mathbf{X\beta} + \mathbf{Z\gamma} \end{aligned} \qquad (5)$$

Quasi-Newton's method optimization is performed via maximum likelihood instead of restricted maximum likelihood (REML) such that a likelihood ratio test may be performed between the fitted and null models[54,55,57]. In this mode, the null model takes a slightly different form; a random intercept is included for each subject in order to account for differing levels of expression across subjects. As with the GLM and GEE modes, a Chi-squared asymptotic distribution is used to estimate *p*-values, with the degrees freedom being equal to the difference in the number of estimated parameters between models[54,58]. The varying null model structures are shown together at **Table 1**.

Mode	Null Model Structure	Fitting Method
GLM	Intercept-only NB generalized linear model	MASS::glm.nb()
GEE	Intercept-only NB estimating equations with desired correlation structure	geeM::geem()
GLMM	NB generalized linear mixed model with a random intercept per subject	glmmTMB::glmmTMB()

Table 1: Structures of scLANE's null model for each mode.

368 **Data**

369 **Simulated datasets**

370 In order to obtain unbiased estimates of trajectory DE method performance we simulated 72
371 single-subject and 240 multi-subject UMI-based scRNA-seq datasets under several different
372 parameter regimes via the Scaffold R package[59]. The datasets fall into three broad categories:
373 cells from a single subject, cells from multiple subjects with gene dynamics largely conserved
374 between subjects, and cells from multiple subjects with heterogeneous gene dynamics. These
375 three regimes represent the use cases for the GLM, GEE, and GLMM modes implemented in
376 scLANE. Under each regime, we generated datasets using three different reference datasets; the
377 first composed of healthy human pancreas cells, the second of human embryonic neurons, and
378 the last of healthy murine pancreatic cells during endocrinogenesis[13,60,61]. Prior to simulation,
379 each reference dataset was filtered to ensure that all genes were present in at least three cells. The
380 simulation method implemented in the Scaffold package ensured that regardless of which
381 reference was used, all simulated datasets reflected the sparsity and overdispersion inherent to
382 scRNA-seq count data[1,59,62,63]. For single-subject datasets we varied the number of cells

generated along with the percentage of truly dynamic genes. In addition to those parameters, for multi-subject datasets we also varied the sample allocation between subjects (balanced vs. unbalanced), and to what degree gene dynamics were conserved at the population and group levels.

Each simulated dataset was then preprocessed with a typical workflow built with the Scater and Scran R packages[39,64,65]. Briefly, counts were depth-normalized then variance-stabilized using a log1p-transform, after which 2,000 HVGs were selected and used as input to PCA[66,67]. The top 30 PCs were then used as an initialization for UMAP in order to embed the cells in two dimensions for visualization purposes[68,69]. Finally, the cells were clustered using the Louvain algorithm after estimation of a shared nearest-neighbor (SNN) graph in PCA space using the cosine distance[70–72]. Reproducibility was ensured and stochasticity controlled through the implementation of the above as a pipeline using the Targets R package[73].

Pancreatic endocrinogenesis dataset

The raw spliced and unspliced mRNA counts were obtained using the download utility in the scVelo Python library[74]. We filtered the genes to only include those with at least 20 counts total between the spliced and unspliced assays, then selected 3,000 HVGs using the Seurat method of variance estimation[51,74]. The counts were then log1p-transformed and scaled prior to PCA, after which 30 components were retained and used to create an SNN graph with parameter $k = 20$ and the cosine distance metric[75]. Smoothed first-order moments of the spliced and unspliced counts were estimated using the set of nearest-neighbors identified previously[74]. The SNN graph was sorted into clusters via the Leiden algorithm with resolution parameter $r = 0.3$ [76]. Finally,

404 UMAP and force-directed graph embeddings were generated in two dimensions using the SNN
405 graph estimated previously[68,77].

406 The scVelo Python library was used to estimate RNA velocities with the dynamical model, after
407 which the velocity vectors were projected onto the two-dimensional UMAP embedding and
408 visualized[74]. In addition, when computing the velocity graph we estimated an uncertainty value
409 for each cosine similarity between velocity vectors and true changes in gene expression between
410 cell states. These uncertainty estimates were propagated forward using Monte Carlo sampling to
411 the estimation of a cell-cell transition probability matrix using the velocity kernel in CellRank[78].
412 Next, CytoTRACE differentiation potential scores and directions were estimated per-cell using
413 the CytoTRACE kernel in CellRank, and the two kernels were combined with equal
414 weighting[78,79]. The combined kernel's transition probability matrix was projected onto the
415 UMAP embedding and visualized.

416 Next, macrostates were generated with the GPCCA estimator and initial and terminal cell states
417 set based on prior biological knowledge[78,80]. Absorption probabilities and corresponding
418 absorption times for each terminal cell fate were estimated, and lineage drivers identified by
419 computing Spearman correlations between gene expression and cell fate probability. Lastly, a
420 gene-shared latent time ordering was computed using the probabilities of each cell being an
421 initial or terminal state as priors[74,78].

422 The latent time ordering was then used as input to scLANE along with a per-cell offset, with
423 spliced mRNA counts serving as the response variable. The GLM framework was used to run
424 scLANE, as all cells in the dataset were taken from the same batch, timepoint E15.5. Genes
425 classified as dynamic under default settings were then clustered using the Leiden algorithm on a
426 KNN graph and embedded in 2D space via UMAP[68,76]. Gene clusters were tested for pathway

427 enrichment using the gprofiler2 R package, and the clusters were annotated based on pathway
428 enrichment, per-cell module scores for each set of genes, and biological knowledge concerning
429 the most-significant TDE genes in each cluster[51,81]. The module scores for four gene programs of
430 interest - cell cycle, beta and delta cell development, endocrine cell maturation, and alpha and
431 epsilon cell development were visualized in UMAP space. Lastly, to determine other possible
432 regulators of endocrine cell fate specification we identified the set of nearest neighbors for
433 *Neurog3* and visualized their dynamics across pseudotime.

434 **Zebrafish dataset**

435 The zebrafish dataset was downloaded via a utility included with the CellRank Python
436 package[22,78]. The original force-directed graph embedding was retained, and after filtering out
437 cells with null embedding values and genes expressed in less than 5 cells the dataset comprised
438 15,363 genes across 2,341 cells[77]. The cells were taken from 12 timepoints between 3.3 and 12
439 hours post-fertilization[4]. After counts-per-10k depth normalization and log1p-transformation,
440 3,000 HVGs were selected and scaled using Seurat[51]. PCA was used to reduce dimensionality,
441 after which an SNN graph with 20 nearest-neighbors was computed on the first 30 PCs using the
442 cosine distance. Lastly, cells were partitioned into 8 clusters via the Leiden algorithm with
443 resolution set to 0.5[76]. Pseudotime was estimated for two lineages with Slingshot[23] based on the
444 first two force-directed graph components, with cluster 5 set as the root node based on its being
445 composed of early blastomere cells from the earliest timepoint.

446 scLANE testing was performed for each lineage with default settings on the top 3,000 most-
447 variable genes with mean normalized expression greater than 0.1, and *p*-values were adjusted
448 with Holm's correction[48]. The fitted values for each gene within each lineage were treated as

449 smoothed expression estimates and were then dimension-reduced to 20 components with PCA.
450 Genes were clustered using the Leiden algorithm after being sorted into an SNN graph using the
451 cosine distance in PCA space, and embedded into 2D UMAP space using the same[68,76].

452 **Fetal liver hematopoesis dataset**

453 After downloading the hematopoesis dataset from the Human Developmental Cell Atlas portal,
454 we subset the cells to just the six celltypes relevant to the Kupffer cell lineage using the authors'
455 original annotations[25,82]. All preprocessing steps were performed using the Scanpy Python
456 library[75]. After selecting 4,000 HVGs, we trained an scVI model for 250 epochs with two layers
457 and a negative-binomial likelihood to integrate samples across the fetal ID from which cells were
458 taken[83]. Gene dispersions were estimated within each celltype. The scVI model provided a 20-
459 dimensional latent representation of the cells which was then used as a drop-in replacement for
460 the usual PCA embedding. Using Scanpy, a nearest-neighbors graph was estimated in scANVI
461 space with $k = 20$ and the cosine distance[75]. Smoothed expression estimates were generated
462 across each cell's nearest neighbors using the scVelo Python library[74]. Clustering was performed
463 with the Leiden algorithm with the resolution parameter set to 0.2[76]. The nearest neighbors graph
464 was also used as the initialization for a force-directed graph embedding computed using the
465 ForceAtlas2 algorithm[77]. Diffusion map and partitioned graph abstraction (PAGA) embeddings
466 were computed using the default Scanpy settings[75,84,85]. After estimating pseudotime over a
467 single lineage with Slingshot[23], TDE testing was performed with scLANE's GLM mode on the
468 top 2,000 HVGs using 4 threads.

469 The fitted gene dynamics were dimension-reduced using PCA and then clustered using the
470 Leiden algorithm on an SNN graph with the resolution parameter set to 0.4[76]. A gene-level

471 UMAP embedding was generated using the PCA embedding as initialization[68,69]. The gene
472 clusters were then used to estimate pathway enrichment using the gProfiler2 R package[81]. The
473 gene clusters were manually annotated by examining both the overall most significantly-enriched
474 pathways as well as the pathways that were uniquely enriched in each cluster. Per-cell scores for
475 each gene program were estimated using the **AddModuleScore()** function from the Seurat R
476 package and then min-max normalized within each gene program[51].

477 In order to validate the biological conclusions generated by scLANE we also analyzed the
478 hematopoesis dataset using CellRank[78]. Using the smoothed expression estimates we fit a
479 CytoTRACE kernel, which provided a cell-cell transition matrix as well as per-cell CytoTRACE
480 pseudotime and differential potential estimates[79]. We also fit a pseudotime kernel using the
481 estimated cellular ordering from Slingshot[23], then generated a weighted combination of the two
482 kernels with CytoTRACE and Slingshot pseudotime weighted at 0.2 and 0.8, respectively. With
483 the combined kernel we used the GPCCA estimator to compute a Schur decomposition of the
484 cell-cell transition matrix, determine initial and terminal cell macrostates, compute cell fate
485 probabilities, and identify lineage driver genes[78,80]. Based on biological knowledge we
486 designated the monocyte and Kupffer cell clusters as terminal states and the HSC cluster as an
487 initial state.

488 **Declarations**

489 **Author contributions**

490 R.B. and J.R.L. conceived the method and study. J.R.L. and R.B. implemented the software
491 package. R.B. provided statistical guidance and methodological feedback. J.R.L. designed the

simulation experiments, performed real dataset analyses, and interpreted results. J.R.L. and R.B. wrote and edited the manuscript.

Acknowledgements

We thank Dr. Aaron Molstad for his thoughts on the mixed models implementation, and Xiaoru Dong for her thoughtful commentary throughout method development.

Funding

This work is supported by NIH grant R35GM146895 to R.B.

Data and code availability

An R software library implementing the method as described can be downloaded from https://github.com/jr-leary7/scLANE. The pipeline used to generate simulated datasets and benchmark methods is available at https://github.com/jr-leary7/scLANE-Simulations. The code used to generate all results is available at https://github.com/jr-leary7/scLANE-Paper.

Conflicts of interest

The authors have no conflicts of interest to declare.

514 **References**

515 1. Bacher, R. & Kendziorski, C. Design and computational analysis of single-cell RNA-
516 sequencing experiments. *Genome Biol* **17**, 63 (2016).

517 2. Cannoodt, R., Saelens, W. & Saeys, Y. Computational methods for trajectory inference
518 from single-cell transcriptomics. *Eur. J. Immunol.* **46**, 2496–2506 (2016).

519 3. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**,
520 1491–1498 (2015).

521 4. Van den Berge, K. *et al.* Trajectory-based differential expression analysis for single-cell
522 sequencing data. *Nat Commun* **11**, 1201 (2020).

523 5. Hou, W. *et al.* A statistical framework for differential pseudotime analysis with multiple
524 single-cell RNA-seq samples. *Nat Commun* **14**, 7286 (2023).

525 6. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis.
526 *Nature* **566**, 496–502 (2019).

527 7. Song, D. & Li, J. J. PseudotimeDE: inference of differential gene expression along cell
528 pseudotime with well-calibrated p-values from single-cell RNA sequencing data. *Genome Biol*
529 **22**, 124 (2021).

530 8. Ji, Z. & Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-
531 seq analysis. *Nucleic Acids Res* **44**, e117 (2016).

532 9. Cui, E. H., Song, D., Wong, W. K. & Li, J. J. Single-cell generalized trend model
533 (scGTM): a flexible and interpretable model of gene expression trend along cell pseudotime.
534 *Bioinformatics* **38**, 3927–3934 (2022).

535 10. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol*
536 **21**, 31 (2020).

537 11. Hastie, T. & Tibshirani, R. *Generalized additive models*. (Chapman & Hall/CRC, 1990).

538 12. Rigby, R. A. & Stasinopoulos, D. M. Generalized additive models for location, scale and
539 shape. *J Royal Statistical Soc C* **54**, 507–554 (2005).

540 13. Bastidas-Ponce, A. *et al.* Massive single-cell mRNA profiling reveals a detailed roadmap
541 for pancreatic endocrinogenesis. *Development* dev.173849 (2019) doi:10.1242/dev.173849.

542 14. Bastidas-Ponce, A., Scheibner, K., Lickert, H. & Bakhti, M. Cellular and molecular
543 mechanisms coordinating pancreas development. *Development* **144**, 2873–2888 (2017).

544 15. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).

545 16. Scavuzzo, M. A. *et al.* Endocrine lineage biases arise in temporally distinct endocrine
546 progenitors during pancreatic morphogenesis. *Nat Commun* **9**, 3356 (2018).

547 17. Boland, B. B., Rhodes, C. J. & Grimsby, J. S. The dynamic plasticity of insulin

548 production in β-cells. *Molecular Metabolism* **6**, 958–973 (2017).

549 18. Schreiber, V. *et al.* Extensive NEUROG3 occupancy in the human pancreatic endocrine
550 gene regulatory network. *Molecular Metabolism* **53**, 101313 (2021).

551 19. Steinauer, N. *et al.* Myeloid translocation gene CBFA2T3 directs a relapse gene program
552 and determines patient-specific outcomes in AML. *Blood Advances* **3**, 1379–1393 (2019).

553 20. Steinauer, N., Guo, C. & Zhang, J. The transcriptional corepressor CBFA2T3 inhibits all-
554 trans-retinoic acid–induced myeloid gene expression and differentiation in acute myeloid
555 leukemia. *Journal of Biological Chemistry* **295**, 8887–8900 (2020).

556 21. Jakobczyk, H. *et al.* Reduction of RUNX1 transcription factor activity by a CBFA2T3-
557 mimicking peptide: application to B cell precursor acute lymphoblastic leukemia. *J Hematol*
558 *Oncol* **14**, 47 (2021).

559 22. Farrell, J. A. *et al.* Single-cell reconstruction of developmental trajectories during
560 zebrafish embryogenesis. *Science* **360**, eaar3131 (2018).

561 23. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell
562 transcriptomics. *BMC Genomics* **19**, 477 (2018).

563 24. Bradford, Y. M. *et al.* Zebrafish information network, the knowledgebase for *Danio rerio*
564 research. *Genetics* **220**, iyac016 (2022).

565 25. Popescu, D.-M. *et al.* Decoding human fetal liver haematopoiesis. *Nature* **574**, 365–371
566 (2019).

567 26. Trizzino, M. *et al.* EGR1 is a gatekeeper of inflammatory enhancers in human
568 macrophages. *Sci. Adv.* **7**, eaaz8836 (2021).

569 27. Krishnaraju, K., Hoffman, B. & Liebermann, D. A. Early growth response gene 1
570 stimulates development of hematopoietic progenitor cells along the macrophage lineage at the
571 expense of the granulocyte and erythroid lineages. *Blood* **97**, 1298–1305 (2001).

572 28. Nguyen, H. Q., Hoffman-Liebermann, B. & Liebermann, D. A. The zinc finger
573 transcription factor Egr-1 is essential for and restricts differentiation along the macrophage
574 lineage. *Cell* **72**, 197–209 (1993).

575 29. Laslo, P. *et al.* Multilineage Transcriptional Priming and Determination of Alternate
576 Hematopoietic Cell Fates. *Cell* **126**, 755–766 (2006).

577 30. Feinberg, M. W. *et al.* The Kruppel-like factor KLF4 is a critical regulator of monocyte
578 differentiation. *EMBO J* **26**, 4138–4148 (2007).

579 31. Feinberg, M. W. *et al.* Kruppel-like Factor 4 Is a Mediator of Proinflammatory Signaling
580 in Macrophages. *Journal of Biological Chemistry* **280**, 38247–38258 (2005).

581 32. Boulet, S., Le Corre, L., Odagiu, L. & Labrecque, N. Role of NR4A family members in
582 myeloid cells and leukemia. *Current Research in Immunology* **3**, 23–36 (2022).

583 33. Safe, S. *et al.* Nuclear receptor 4A (NR4A) family – orphans no more. *The Journal of*
584 *Steroid Biochemistry and Molecular Biology* **157**, 48–60 (2016).

585 34. Mahabeleshwar, G. H. *et al.* The Myeloid Transcription Factor KLF2 Regulates the Host
586 Response to Polymicrobial Infection and Endotoxic Shock. *Immunity* **34**, 715–728 (2011).

587 35. Cao, Z., Sun, X., Icli, B., Wara, A. K. & Feinberg, M. W. Role of Krüppel-like factors in
588 leukocyte development, function, and disease. *Blood* **116**, 4404–4414 (2010).

589 36. Angerer, P. *et al.* Single cells make big data: New challenges and opportunities in
590 transcriptomics. *Current Opinion in Systems Biology* **4**, 85–91 (2017).

591 37. Stoklosa, J. & Warton, D. I. A Generalized Estimating Equation Approach to
592 Multivariate Adaptive Regression Splines. *Journal of Computational and Graphical Statistics*
593 **27**, 245–253 (2018).

594 38. Friedman, J. H. Multivariate Adaptive Regression Splines. *Ann. Statist.* **19**, (1991).

595 39. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing,
596 quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*
597 (2017).

598 40. Daniel, F., Ooi, H., Calaway, R. & Weston, S. foreach: Provides Foreach Looping
599 Construct. (2022).

600 41. Weston, S. doParallel: Foreach Parallel Adaptor for the 'parallel' Package. (2022).

601 42. Privé, F., Aschard, H., Ziyatdinov, A. & Blum, M. G. B. Efficient analysis of large-scale
602 genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics* **34**, 2781–2787
603 (2018).

604 43. Golub, G. H., Heath, M. & Wahba, G. Generalized Cross-Validation as a Method for
605 Choosing a Good Ridge Parameter. *Technometrics* **21**, 215–223 (1979).

606 44. Fay, M. P. & Graubard, B. I. Small-sample adjustments for Wald-type tests using
607 sandwich estimators. *Biometrics* **57**, 1198–1206 (2001).

608 45. Liang, K.-Y. & Zeger, S. L. Longitudinal data analysis using generalized linear models.
609 *Biometrika* **73**, 13–22 (1986).

610 46. Nelder, J. A. & Wedderburn, R. W. M. Generalized Linear Models. *Journal of the Royal*
611 *Statistical Society. Series A (General)* **135**, 370 (1972).

612 47. Wang, M. Generalized Estimating Equations in Longitudinal Data Analysis: A Review
613 and Recent Developments. *Advances in Statistics* **2014**, 1–11 (2014).

614 48. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian*
615 *Journal of Statistics* **6**, 65–70 (1979).

616 49. Marschner, I. C. glm2: Fitting Generalized Linear Models with Convergence Problems.
617 *The R Journal* **3**, 12–15 (2011).

50. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. (Springer, 2002).

51. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

52. McDaniel, L. S., Henderson, N. C. & Rathouz, P. J. Fast Pure R Implementation of GEE: Application of the Matrix Package. *R J* **5**, 181–187 (2013).

53. Hardin, J. W. & Hilbe, J. M. *Generalized Estimating Equations*. (Chapman and Hall/CRC, 2012). doi:10.1201/b13880.

54. Faraway, J. J. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, Second Edition*. (Chapman and Hall/CRC, 2016). doi:10.1201/9781315382722.

55. Brooks, M., E. *et al.* glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal* **9**, 378 (2017).

56. Breslow, N. E. & Clayton, D. G. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* **88**, 9 (1993).

57. Dennis, J. E., Gay, D. M. & Walsh, R. E. An Adaptive Nonlinear Least-Squares Algorithm. *ACM Trans. Math. Softw.* **7**, 348–368 (1981).

58. McCulloch, C. E. & Searle, S. R. *Generalized, linear, and mixed models*. (John Wiley & Sons, 2001).

59. Bacher, R. *et al.* Enhancing biological signals and detection rates in single-cell RNA-seq experiments with cDNA library equalization. *Nucleic Acids Research* (2021).

60. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst* **3**, 346-360.e4 (2016).

61. La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* **167**, 566-580.e19 (2016).

62. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nat Biotechnol* **38**, 147–150 (2020).

63. Choudhary, S. & Satija, R. Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biol* **23**, 27 (2022).

64. Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat Methods* **17**, 137–145 (2020).

65. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* **5**, 2122 (2016).

66. Pearson, K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**, 559–572 (1901).

652 67. Hotelling, H. Analysis of a complex of statistical variables into principal components.
653 *Journal of Educational Psychology* **24**, 417–441 (1933).

654 68. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and
655 Projection for Dimension Reduction. (2020).

656 69. Melville, J. uwot: The Uniform Manifold Approximation and Projection (UMAP)
657 Method for Dimensionality Reduction. (2022).

658 70. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of
659 communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).

660 71. Aggarwal, C. C., Hinneburg, A. & Keim, D. A. On the Surprising Behavior of Distance
661 Metrics in High Dimensional Space. in *Database Theory — ICDT 2001* (eds. Van Den Bussche,
662 J. & Vianu, V.) vol. 1973 420–434 (Springer Berlin Heidelberg, 2001).

663 72. Kim, T. *et al.* Impact of similarity metrics on single-cell RNA-seq data clustering.
664 *Briefings in Bioinformatics* **20**, 2316–2326 (2019).

665 73. Landau, W. The targets R package: a dynamic Make-like function-oriented pipeline
666 toolkit for reproducibility and high-performance computing. *JOSS* **6**, 2959 (2021).

667 74. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity
668 to transient cell states through dynamical modeling. *Nat Biotechnol* **38**, 1408–1414 (2020).

669 75. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression
670 data analysis. *Genome Biol* **19**, 15 (2018).

671 76. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-
672 connected communities. *Sci Rep* **9**, 5233 (2019).

673 77. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a Continuous Graph
674 Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS*
675 *ONE* **9**, e98679 (2014).

676 78. Weiler, P., Lange, M., Klein, M., Pe'er, D. & Theis, F. J. *Unified fate mapping in*
677 *multiview single-cell data*. https://doi.org/10.1101/2023.07.19.549685 (2023).

678 79. Gulati, G. S. *et al.* Single-cell transcriptional diversity is a hallmark of developmental
679 potential. *Science* **367**, 405–411 (2020).

680 80. Reuter, B., Fackeldey, K. & Weber, M. Generalized Markov modeling of nonreversible
681 molecular kinetics. *The Journal of Chemical Physics* **150**, 174103 (2019).

682 81. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and
683 conversions of gene lists (2019 update). *Nucleic Acids Research* **47**, W191–W198 (2019).

684 82. Regev, A. *et al.* The Human Cell Atlas White Paper. (2018).

685 83. Gayoso, A. *et al.* A Python library for probabilistic analysis of single-cell omics data. *Nat*
686 *Biotechnol* **40**, 163–166 (2022).

687 84. Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference
688 through a topology preserving map of single cells. *Genome Biol* **20**, 59 (2019).

689 85. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-
690 cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).

691

692