Probabilistic reconstruction of cellular differentiation trees from single-cell RNA-seq data

Anonymous Author(s)
Affiliation
Address
email

Abstract

Until recently, transcriptomics was limited to bulk RNA sequencing, obscuring the underlying expression patterns of individual cells in favor of a global average. Thanks to technological advances, we can now profile gene expression across thousands or millions of individual cells in parallel. However, this new type of data poses complex computational challenges for disentangling biological variation from technical variation, uncovering novel cell types, and reconstructing trajectories of differentiating cells from data that are noisy, heterogenous, and sparse. Here, we develop a full generative model for probabilistically reconstructing trees of cellular differentiation from single-cell RNA-seq data. Specifically, we extend the framework of the classical Dirichlet diffusion tree (DDT) to simultaneously infer branch topology and latent cell state along diffusive trajectories over the full tree. Finally, we demonstrate that Markov chain Monte Carlo inference with our augmented DDT model can recover latent trajectories from simulated single-cell transcriptomes.

1 Introduction

Many problems in biology invoke the question of how to describe and measure cell state. One particularly informative measure is *gene expression*, i.e., the amount of messenger RNA (mRNA) corresponding to each gene. Recent techniques offer unprecedented insight into cellular gene expression by facilitating massively-parallel quantification of RNA molecules at single-cell resolution (*single-cell RNA sequencing*, or scRNA-seq) [1–3]. However, the resulting data are noisy and zero-inflated, confounding traditional analysis techniques [4, 5]. In this work, we employ a Bayesian approach to directly model sources of uncertainty in single-cell transcriptomic data and infer interpretable, probabilistic insight into cell state.

The particular biological phenomenon we study is *cellular differentiation*, the process by which a less specialized progenitor (e.g., a stem cell) gives rise to cells with more specialized function [9]. Differentiation is ubiquitous to multicellular life, occuring during development – when a zygote rapidly divides to form a complex organism [6] – and over the course of adulthood – e.g., in humans, the blood immune system undergoes continuous regeneration (through hematopoiesis) [7] and the gut lining is entirely replenished by new cells on the order of days [8]. The process of cellular differentiation can be represented as a tree whose branches designate the incremental progression of more general cells into various mature cell types [9]. However, many fundamental questions remain. How do identical progenitors reliably give rise to a suite of branching cell fates? How do the dynamics of gene expression change over time and across lineages? What is the molecular program by which orchestrated changes in co-expression lead cells down one path or another?

To address these questions, we seek to infer the latent tree of cellular differentiation and the genes that drive its topology from scRNA-seq measurements, which provide a noisy snapshot of cell state. Since

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.

the assays used to measure gene expression are destructive to cells, we cannot follow a single cell along its trajectory through time, and must therefore infer this trajectory by sampling many individual cells [4, 5, 10]. Further complicating analysis, any given sample (e.g., of hematopoietic cells) is a non-uniform draw of unlabeled time points from the underlying tree of differentiation [10].

In this work, we begin by reviewing previous approaches to reconstructing cell trajectories and Bayesian inference on trees in Section 2. In Section 3, we develop a new generative model for scRNA-seq data arising from cells undergoing a dynamic, bifurcating differentiation process. In Section 4, we describe our inference algorithm, which arises naturally by applying Bayes' theorem to invert the generative model into a method for sampling from its stationary distribution. Finally, in Section 5, we present initial experiments that demonstrate the ability of our proposed techniques to recover structure from simulated single-cell transcriptomes.

Ultimately, we aim to apply our novel model to scRNA-seq datasets of human hematopoietic and mouse gut epithelial cells. This analysis will uncover how changes in cell state are driven by systems-level transcriptional "programs" – networks of co-regulated genes whose expression is orchestrated by complex interactions among a handful of transcription factors. Further, our work will enable experimental interrogation of model predictions [11], by observing changes in gene expression and tree topology in response to targeted knockdown of genes implicated in particular trajectories.

2 Background

Many methods exist to infer lineage relationships from single-cell transcriptomic data [10]. However, most require suitable normalization or dimensionality reduction beforehand, assume a fixed number of branch points (often zero or one), retain no notion of uncertainty, or do not infer differentially expressed genes in conjunction with reconstructing the tree [12–21].

In contrast, we develop a *Bayesian* model of differentiation that is not fragile to preprocessing methods, since we directly model gene expression counts, and yields interpretable results for differential expression across lineages. Our approach provides a generative means of evaluating and simulating from the model, as well as a principled way of accounting for technical factors like zero-inflation due to gene dropout. Further, we leverage *Bayesian nonparametrics* to flexibly learn trees of unbounded width and depth, with no requirement for *a priori* knowledge of the exact number of cell fates.

Previous approaches to Bayesian modeling of latent tree structures either assume data are generated only at the leaves (e.g. Dirichlet [22] or Pitman Yor [23] diffusion trees, hierarchical Dirichlet process [24]) or at the nodes of the tree (nested Chinese restaurant process [25], tree-structured stick breaking process [26]). In contrast, we seek to model data arising from a more challenging regime in which observations are generated continuously over the entire tree, from root to leaves.

3 Generative Bayesian framework for cellular differentiation

3.1 Observation model for single-cell RNA-seq

Consider a single cell containing a set of mRNA transcripts corresponding to each gene that is currently expressed. We assume that, for a particular cell type, the discrete count M of transcripts of a particular gene g has a Poisson distribution with rate $\lambda^{(g)}$: $M \sim \mathsf{Poiss}(\lambda^{(g)})$.

Throughout the workflow for droplet-based sequencing [1–3], the current state of the art for single-cell transcriptomes, there are several processes known to affect accurate observation of M [4, 5, 27] (Appendix A). First, after each cell is captured by a droplet containing a single bead, transcripts must bind to barcoded DNA primers coating the bead. We assume each mRNA molecule hybridizes with probability p_h to a primer with an i.i.d. uniform primer-specific barcode (unique molecular identifier, or UMI) [1]. Next, transcripts must be reverse transcribed and amplified through Polymerase Chain Reaction (PCR); we assume probability p_d of successful amplification per round. After R rounds of PCR, molecules hybridize to the flow cell for sequencing with some probability; call this p'_h. Since the original quantity M was Poisson distributed, we can use the thinning property and the marking property to show that the number attached to each unique UMI and ultimately sequenced is

$$M_1, \ldots, M_{N_{\text{UMI}}} \overset{\text{i.i.d.}}{\sim} \mathsf{Poiss}\left(\frac{(1+p_d)^R\, p'_h\, p_h\, \lambda^{(g)}}{N_{\text{UMI}}}\right), \tag{1}$$

where N_{UMI} is the number of unique UMIs, with $N_{\text{UMI}} = 4^{10}$ for a 10 basepair UMI. Finally, mRNA counts per gene are quantified by aligning partial transcripts to a reference genome, resulting in an overall count $x^{(g)}$ for this particular gene, with $x^{(g)} = \sum_{i=1}^{N_{\text{UMI}}} \mathbb{1}\left[M_i > 0\right]$. The distribution of $x^{(g)}$ has a closed-form expression:

$$x^{(g)} \sim \text{Binom}\left(N_{\text{UMI}}\ ,\ \ 1 - e^{-q\lambda^{(g)}}\right) \qquad \text{with} \qquad q := \frac{(1+p_d)^R\, p'_h\, p_h}{N_{\text{UMI}}}, \tag{2}$$

where q is a hyperparameter accounting for gene dropout. Because of dropout and the fact that most genes are turned off at any given time, the expression profile for cell c, x_c, is a sparse vector of digital counts in roughly $\mathbb{N}^{20,000}$ (for human cells) [4, 5, 27].

3.2 Augmented Dirichlet diffusion trees for inferring cell trajectories

We model observed expression profiles as arising from an underlying branching process. In particular, we model the hidden abstraction of cellular developmental state, $\lambda \in \mathbb{R}^G$, as draws from a latent tree, and replace the original Poisson rate parameter for gene expression with $h(\lambda^{(g)})$ for some link function $h : \mathbb{R} \to \mathbb{R}_{\geq 0}$. Next, we describe in more detail the generation of $\lambda = [\lambda^{(1)}, ..., \lambda^{(G)}]$, corresponding to genes $1, \ldots, G$, for each cell.

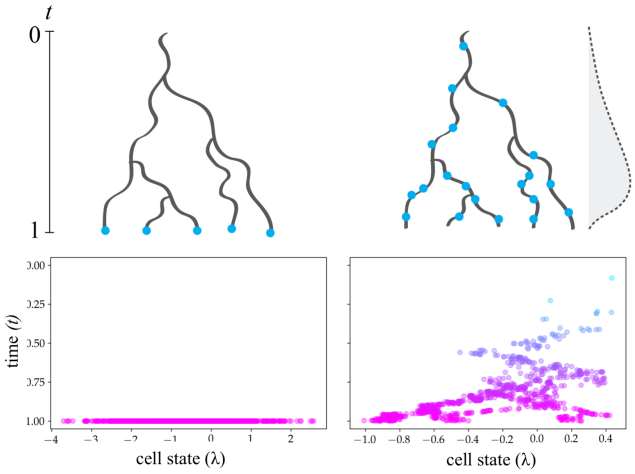

Figure 1: Classical DDT (*left*) vs. our augmented DDT (*right*). *Upper*: cartoon of tree and data points. *Upper right inset*: A prior Beta$(5, 1)$ distribution over cell pseudotimes in our augmented DDT. *Lower*: Cellular developmental states simulated from each model (colored by time).

Dirichlet diffusion trees (DDTs) provide a nonparametric model for a latent branching process, including tree topology as well as locations along the branches (Appendix B). As classically formulated, the DDT serves to flexibly model densities for data generated at the leaves of a binary tree, with Gaussian diffusion (Brownian motion) between nodes [22] (see Figure 1, left). We extend the DDT model to generate latent values (cell states) according to a continuous-time distribution over the entire tree (Figure 1, right).

A draw from the distribution describing a K-leaf DDT yields a set of locations for both internal and leaf nodes, as well as a means of sampling all branch locations. In practice, we do not use the full continuum of branch locations (gray lines, Figure 1) but focus on the locations of each cell (blue dots, Figure 1, right). Therefore, in our algorithm, we instantiate the set of tree node locations, τ, and we additionally instantiate only those branch locations that correspond to cells. We can think of each location, or point, in the tree as a pair comprising the rate $\lambda \in \mathbb{R}^G$ (the horizontal axis in Figure 1) and the pseudotime t (the vertical axis in Figure 1). Thus, cell state λ_c can be seen as a projection of its overall location in the tree, (λ_c, t_c).

In our case, the number of leaves (cell fates) is no longer fixed to the number of data points as in a classical DDT, so we place an appropriate prior to regularize the depth of the tree, e.g. $K \sim 1 + \text{Poiss}(K_0)$. Given a set of node locations τ (times and rates) drawn according to the distribution describing a classical DDT for some initial setting of K (Appendix B), each cell c diffuses down this tree (as detailed in Appendix C) until a random time point. Its location in the tree at this time yields a latent rate $\lambda_c = [\lambda_c^{(1)}, \ldots, \lambda_c^{(G)}]$ for the cell. Finally, we sample its gene expression profile $x_c = [x_c^{(1)}, \ldots, x_c^{(G)}]$ according to the observation model, $x_c^{(g)} \mid h(\lambda_c^{(g)})$, as in Section 3.1. Here, we make the simplifying assumption that genes are expressed independently conditioned on the latent hierarchical structure of differentiation.

4 Inference

Having specified our generative model, we can derive a Markov chain Monte Carlo sampler to approximate the Bayesian posterior for the model parameters. In particular, we aim to recover the tree topology and cell rates and pseudotimes. We derive Metropolis-Hastings proposals for efficient mixing over trees in order to sample from our augmented DDT model. In brief, our proposals include:

(i) *Cell resampling* to propose new pseudotimes and rates for all cells, conditioned on τ. This is akin to a Gaussian mixture model over extant branches at a given time slice t_c.

(ii) *Subtree prune and regraft*, as originally formulated for DDTs [22], detaches a random subtree and draws its new parent from the prior (resampling cells on the affected interval).

(iii) *Split/merge* to propose growing or pruning subtrees, changing the dimension of the tree by one or more leaves (and resampling all affected cells).

(iv) *Gibbs updates* to latent node rates (conditioned on their immediate neighbors) and diffusion parameter σ (with a conjugate Inverse Gamma prior).

(v) *Message passing* on trees to perform exact inference on node and cell rates, given their neighbors. Here, we are working toward leveraging Pólya-gamma (PG) augmentation [28] (with an appropriate choice of link function h) to reparameterize Binomial observations as Gaussian likelihoods, conditioned on Gaussian latent variates and auxiliary PG variates.

5 Initial results

In preliminary experiments, we simulated single-cell data in order to check our ability to infer latent parameters against known ground truth. Initial results demonstrate that we can recover latent structure using a limited version of the sampler. We obtained the best results by initializing with $K_{\text{init}} \gg K_{\text{true}}$, such that the initial tree spans a diverse subspace of cell outcomes and the sampler reshapes and prunes the low-likelihood branches.

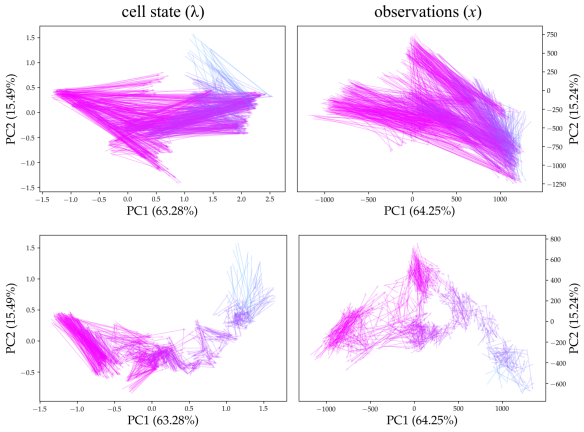

Figure 2: Shifts in λ (*left*) and x (*right*), based on the first two principal components. Arrows, each corresponding to a single "cell," are colored by true time and point from true to inferred value, following PCA of each set of plotted points. *Upper*: initialized tree; *lower*: maximum a posteriori tree.

We simulated data by sampling cell times $t_c \sim \text{Beta}(5, 1)$ and drawing latent locations $\lambda_c \in \mathbb{R}^{10}$ from the augmented DDT model with 5 leaves and concentration $\alpha = 1$. We simulated expression profiles $x_c^{(g)} \mid h(\lambda_c^{(g)})$ with link $h(\lambda^{(g)}) = e^{\lambda^{(g)}}$.

We initialized the sampler with a tree drawn from a prior over DDTs with 25 leaves and the same concentration parameter, with initial cell times randomly sampled from the same time distribution. Following a burn-in of 500 steps and 2500 additional iterations of MCMC, we examined sampled trees and assessed convergence based on trace plots. The maximum a posteriori (MAP) tree[1] recovered the true number of leaves ("cell fates"), although node times and locations were not identical. The latter discrepancy motivates future work toward developing tree metrics based on node-matching, including comparing trees of differing depths. Further, examination of per-cell shifts in the true and inferred tree locations and simulated observations demonstrates partial recovery of ground-truth values. Specifically, we observe that the distances between true and inferred values per cell (visualized following PCA) greatly shrink from initialization to MAP tree, indicating that we are approaching ground truth (Figure 2).

Ultimately, we are interested in the stochasticity of lineage fate specification – whether cells "commit" to (are probabilistically inclined toward) particular fates prior to branching – and in reconstructing the master regulatory programs and fitness landscapes that govern large-scale changes in cell state.

[1] More precisely, the sampled tree with highest probability.

References

[1] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll. "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets." *Cell* 161.5 (2015), pp. 1202–1214.

[2] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner. "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells." *Cell* 161.5 (2015), pp. 1187–1201.

[3] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas. "Massively parallel digital transcriptional profiling of single cells." *Nature Communications* 8 (2017).

[4] A. Wagner, A. Regev, and N. Yosef. "Revealing the vectors of cellular identity with single-cell genomics." *Nature Biotechnology* 34.11 (2016), pp. 1145–1160.

[5] O. Stegle, S. A. Teichmann, and J. C. Marioni. "Computational and analytical challenges in single-cell transcriptomics." *Nature Reviews Genetics* 16.3 (2015), pp. 133–145.

[6] L.-F. Chu, N. Leng, J. Zhang, Z. Hou, D. Mamott, D. T. Vereide, J. Choi, C. Kendziorski, R. Stewart, and J. A. Thomson. "Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm." *Genome Biology* 17.1 (2016), p. 173.

[7] S. H. Orkin and L. I. Zon. "Hematopoiesis: An Evolving Paradigm for Stem Cell Biology." *Cell* 132.4 (2008), pp. 631–644.

[8] N. Barker. "Adult intestinal stem cells: critical drivers of epithelial homeostasis and regeneration." *Nature Reviews Molecular Cell Biology* 15.1 (2014), pp. 19–33.

[9] S. F. Gilbert and M. J. F. Barresi. *Developmental Biology, 11th edition*. Sinauer, 2016.

[10] R. Cannoodt, W. Saelens, and Y. Saeys. "Computational methods for trajectory inference from single-cell transcriptomics." *European Journal of Immunology* 46.11 (2016), pp. 2496–2506.

[11] A. Dixit, O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman, and A. Regev. "Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens." *Cell* 167.7 (2016), pp. 1853–1866.

[12] L. Haghverdi, M. Buttner, F. A. Wolf, F. Buettner, and F. J. Theis. "Diffusion pseudotime robustly reconstructs lineage branching." *Nature Methods* 13 (2016), pp. 845–848.

[13] M. Setty, M. D. Tadmor, S. Reich-Zeliger, O. Angel, T. M. Salame, P. Kathail, K. Choi, S. Bendall, N. Friedman, and D. Pe'er. "Wishbone identifies bifurcating developmental trajectories from single-cell data." *Nature Biotechnology* 34 (2016), pp. 637–645.

[14] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells." *Nature Biotechnology* 32 (2014), pp. 381–386.

[15] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. Pliner, and C. Trapnell. "Reversed graph embedding resolves complex single-cell trajectories." *Nature Methods* 14.10 (2017), pp. 979–982.

[16] K. R. Moon, D. van Dijk, Z. Wang, W. Chen, M. J. Hirn, R. R. Coifman, N. B. Ivanova, G. Wolf, and S. Krishnaswamy. "PHATE: A Dimensionality Reduction Method for Visualizing Trajectory Structures in High-Dimensional Biological Data." *bioRxiv* (2017).

[17] A. Boukouvalas, J. Hensman, and M. Rattray. "BGP: Branched Gaussian processes for identifying gene-specific branching dynamics in single cell data." *bioRxiv* (2017).

[18] J. D. Welch, A. J. Hartemink, and J. F. Prins. "SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data." *Genome Biology* 17 (2016).

[19] E. Marco, R. L. Karp, G. Guo, P. Robson, A. H. Hart, L. Trippa, and G.-C. Yuan. "Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape." *Proceedings of the National Academy of Sciences* 111.52 (2014), E5643–E5650.

[20] K. R. Campbell and C. Yau. "Probabilistic modeling of bifurcations in single-cell gene expression data using a Bayesian mixture of factor analyzers." *Wellcome Open Research* 2 (2017).

[21] T. Lönnberg, V. Svensson, K. R. James, D. Fernandez-Ruiz, I. Sebina, R. Montandon, M. S. F. Soon, L. G. Fogg, A. S. Nair, U. N. Liligeto, M. J. T. Stubbington, L.-H. Ly, F. O. Bagger, M. Zwiessele, N. D. Lawrence, F. Souza-Fonseca-Guimaraes, P. T. Bunn, C. R. Engwerda, W. R. Heath, O. Billker, O. Stegle, A. Haque, and S. A. Teichmann. "Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves Th1/Tfh fate bifurcation in malaria." *Science Immunology* 2.9 (2017).

[22] R. M. Neal. "Density modeling and clustering using Dirichlet diffusion trees." *Bayesian Statistics* 7 (2003), pp. 619–629.

[23] D. A. Knowles and Z. Ghahramani. "Pitman Yor Diffusion Trees for Bayesian Hierarchical Clustering." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.2 (2015), pp. 271–289.

[24] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. "Hierarchical Dirichlet processes." *Journal of the American Statistical Association* 101.476 (2006), pp. 1566–1581.

[25] D. M. Blei, T. L. Griffiths, and M. I. Jordan. "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies." *Journal of the ACM* 57.2 (2010).

[26] Z. Ghahramani, M. I. Jordan, and R. P. Adams. "Tree-Structured Stick Breaking for Hierarchical Data." *Advances in Neural Information Processing Systems 23*. 2010, pp. 19–27.

[27] A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, and S. A. Teichmann. "The Technology and Biology of Single-Cell RNA Sequencing." *Molecular Cell* 58.4 (2015), pp. 610–620.

[28] N. G. Polson, J. G. Scott, and J. Windle. "Bayesian inference for logistic models using Pólya-Gamma latent variables." *Journal of the American Statistical Association* 108.504 (2013), pp. 1339–1349.

[29] I. C. Macaulay, V. Svensson, C. Labalette, L. Ferreira, F. Hamey, T. Voet, S. A. Teichmann, and A. Cvejic. "Single-Cell RNA-Sequencing Reveals a Continuous Spectrum of Differentiation in Hematopoietic Cells." *Cell* 14 (2016), pp. 966–977.

A Observation model for single-cell RNA-seq

Consider a single cell containing a set of mRNA transcripts corresponding to each gene that is currently expressed. We assume that, for a particular cell type, the discrete count M of transcripts of a particular gene g has a Poisson distribution with rate $\lambda^{(g)}$,

$$M \sim \mathsf{Poiss}(\lambda^{(g)}). \tag{3}$$

For droplet-based methods, the current state-of-the-art for scRNA-seq, cells are flowed through a microfluidic device such that each cell is captured by a droplet of fluid containing a single microbead covered in a large number (roughly 10^8) of barcoded DNA primers [1–3, 27]. Each primer contains a PCR handle, bead-specific barcode, and primer-specific barcode (unique molecular identifier, or UMI). Following cell lysis, mRNA transcripts hybridize to these randomized primers. Under the assumption that each droplet contains a single microbead and a single cell, the bead-specific barcode acts as a cell-specific barcode and the UMI acts as a transcript-specific barcode [1]. As there are vastly more primers on the microbead than mRNA levels in the cell (at most roughly 10^6), we assume that each transcript hybridizes with probability p_h to a primer with an i.i.d. uniform UMI. Since the original quantity M was Poisson distributed, we can use the thinning property and the marking property to show that the number attached to each unique UMI is

$$M_1, \ldots, M_{N_{\text{UMI}}} \overset{\text{i.i.d.}}{\sim} \mathsf{Poiss}\left(\frac{p_h \lambda^{(g)}}{N_{\text{UMI}}}\right). \tag{4}$$

Ideally, each $M_i \in \{0, 1\}$; if $M_i > 1$, we will underestimate the number of copies of mRNA for a given gene. However, this caveat is decreasingly important as N_{UMI} increases, and effectively disappears when $N_{\text{UMI}} \gg \lambda$ (i.e. the number of unique UMIs greatly exceeds the number of mRNA transcripts for each gene). Assuming a 10 basepair UMI, this process enables digital quantification of mRNA molecules up to 4^{10} transcripts per gene [1].

Following reverse transcription of the bound mRNA to complementary DNA (cDNA), we use Polymerase Chain Reaction (PCR) to exponentially amplify the cDNA library [27]. Assume each molecule has some probability p_d of successfully replicating for each round of PCR. Again using the Poisson marking/thinning property, after R rounds of PCR we have

$$M'_1, \ldots, M'_{N_{\text{UMI}}} \overset{\text{i.i.d.}}{\sim} \mathsf{Poiss}\left(\frac{(1+p_d)^R\, p_h\, \lambda^{(g)}}{N_{\text{UMI}}}\right). \tag{5}$$

The library is then loaded onto a flow cell for sequencing; we assume that each transcript hybridizes to the lawn of oligonucleotides with some probability p'_h, yielding

$$M''_1, \ldots, M''_{N_{\text{UMI}}} \overset{\text{i.i.d.}}{\sim} \mathsf{Poiss}\left(\frac{(1+p_d)^R\, p'_h\, p_h\, \lambda^{(g)}}{N_{\text{UMI}}}\right). \tag{6}$$

Finally, following sequencing, mRNA counts per gene are quantified by aligning partial transcripts to a reference genome, with basic error correction to eliminate singletons and account for sequencing error [1, 3, 27], resulting in an overall count for this particular gene,

$$x^{(g)} = \sum_{i=1}^{N_{\text{UMI}}} \mathbb{1}\left[M''_i > 0\right]. \tag{7}$$

The distribution of $x^{(g)}$ has a closed-form expression:

$$x^{(g)} \sim \mathsf{Binom}\left(N_{\text{UMI}},\ \ 1 - e^{-q\lambda^{(g)}}\right) \tag{8}$$

$$q := \frac{(1+p_d)^R\, p'_h\, p_h}{N_{\text{UMI}}}, \tag{9}$$

where q is a hyperparameter accounting for gene dropout.

Because of dropout and the fact that most genes are turned off at any given time, the expression profile for cell c, x_c, is a sparse vector of digital molecular counts in roughly $\mathbb{N}^{20{,}000}$ (for human cells) [4, 5, 27].

B Dirichlet diffusion trees

The Dirichlet diffusion tree (DDT) model provides a family of priors of over infinitely exchangeable data that derive from a latent binary branching process. As classically formulated, the DDT model generalizes Dirichlet process mixture models for data that are hierarchical, such that data points are generated at the leaves and internal nodes correspond to hierarchical clusters [22].

Consider a draw from the distribution over K-leaf DDTs. If marginalized over the paths between nodes, the sampled DDT consists of a set of locations (internal or leaf nodes) and pseudotimes,

$$\tau = \{(\lambda_r, t_r)\}_{r=1}^{2K-1}. \tag{10}$$

Here, we use r to index the locations of tree nodes (internal nodes and leaves), as opposed to c, which we later use to index the locations of cells along the tree.

Tree topology is generated by iteratively simulating paths of particles according to a Gaussian diffusion process (i.e. Brownian motion). Specifically, a particle that has reached $X(t)$ at time $t \in (0, 1)$ will diffuse to $X(t+dt) = X(t) + \mathcal{N}(0,\ \sigma_0^2 \mathrm{I} \cdot dt)$ after an infinitesimal amount of time dt, for some σ_0^2 that governs diffusion (which can be learned). Integrated over a discrete time interval Δt, then, $X(t+\Delta t) \sim \mathcal{N}\left(X(t),\ \sigma_0^2 \mathrm{I}(\Delta t)\right)$ [22].

Following Neal [22], we assume a branching rate of $a(t) = \alpha/(1-t)$, where α is a smoothness parameter related to whether branches are concentrated toward the root or the leaves. Let

$$A(t) \triangleq \int_0^t a(u)\,\mathrm{d}u = -\alpha \log(1-t); \tag{11}$$

this is the cumulative branching function. If a particle is on a leg of the tree bookended by times $[t_a, t_b]$, and there are m particles that have traversed this path, the probability of branching at some time $t \in (t_a, t_b)$ is

$$B_{t_a}(t) \triangleq P(\text{branch in } [t_a, t]) = 1 - e^{(A(t_a) - A(t))/m} = 1 - \left(\frac{1-t}{1-t_a}\right)^{\alpha/m}. \tag{12}$$

To see when/if a new particle branches on an existing leg of the tree $[t_a, t_b]$, we calculate t_r by the inverse CDF method. If $t_r > t_b$, then the particle does not create a new branch on this leg and instead follows one of the existing branches at time t_b.

Overall, to create a DDT with K particles:

1. Set the root of the tree to some origin $(\mu_0, 0)$, corresponding to a typical value for the data – here, we leverage prior knowledge about average expression profiles for stem cells. Draw the first particle's leaf location as $\lambda_{\ell_1} \sim \mathcal{N}(\mu_0,\ \sigma_0^2 \mathrm{I})$.
2. For $k = 2, \ldots, K$:
   a) Find the next branch point and time, (λ_b, t_b), (e.g. for $k = 3$, this will be where the second particle diverged) and find the number of particles m that have taken this path. Draw $u \sim \mathrm{Unif}(0, 1)$ and compute a proposed branching time,
   $$t_r = B_{t_a}^{-1}(u) = 1 - e^{\log(1-t_a) + \frac{m}{\alpha}\log(1-u)}.$$
   b) If $t_r < t_b$, branch at time t_r. Sample the branching location according to the Brownian bridge defined by its Markov blanket, i.e. nodes a and b bookending the start and end of this leg,
   $$\lambda_r \sim \mathcal{N}\left(\lambda_a + \frac{t_r - t_a}{t_b - t_a}(\lambda_b - \lambda_a)\ ,\ \ (t_r - t_a)\left(1 - \frac{t_r - t_a}{t_b - t_a}\right)\sigma_0^2 \mathrm{I}\right). \tag{13}$$
   This equation derives from the properties of Brownian bridges. A Brownian bridge on the interval $[0, T]$, starting at $X_0 = 0$, and ending at $X_T = 0$ has $X_t \sim \mathcal{N}(0,\ t(1-t/T)\sigma_0^2 \mathrm{I})$. We want a bridge between λ_a and λ_b for times $t_r \in (t_a, t_b)$, or, equivalently, $t_r - t_a \in (0, t_b - t_a)$. Rewriting the time interval this way yields the variance in Eq. (13), and the mean comes from interpolating between λ_a and λ_b over time. Record this branch point (λ_r, t_r) and sample its final leaf location from $\mathcal{N}(\lambda_r, (1-t_r)\,\sigma_0^2 \mathrm{I})$. Move on to the next particle.
   c) If $t_r > t_b$, do not branch off of this leg. Instead, pick one of the two branches at time t_b with probability equal to m_i/m, where $m_i,\ i \in \{1, 2\}$ is the number of particles that previously chose that branch. Go back to step a).

C Model for gene expression rates conditional on a tree

In order to model cells as arising from a continuous-time distribution over a Dirichlet diffusion tree, we draw $\tau \mid K \sim$ DDT and sample each cell c as follows:

1. Draw $t_c \sim F(\cdot)$, where F is some distribution over $[0, 1]$ that represents our belief of how cells are distributed over the tree – e.g. for hematopoiesis we expect most cells to be near the leaves [29], so might choose something like $\mathrm{Beta}(5, 1)$.
2. Over all branches that exist at time t_c, pick one according to the fraction of particles beneath that branch. That is, if there are n_r particles beneath branch r and K particles total, pick branch r with probability n_r/K. This is equivalent to running a new particle through the tree with zero probability of creating a new branch and stopping at time t_c.

3. Find the points on the chosen branch (nodes or cells) that have $t_a < t_c < t_b$, with no other points in between. Let λ_a and λ_b be the latent locations of these points. Then, sample λ_c according to the Brownian bridge defined by its Markov blanket (nodes a and b), as in Eq. (13):

$$\lambda_c \sim \mathcal{N}\left(\lambda_a + \frac{t_c - t_a}{t_b - t_a}(\lambda_b - \lambda_a)\ ,\ \ (t_c - t_a)\left(1 - \frac{t_c - t_a}{t_b - t_a}\right)\sigma_0^2\,\mathrm{I}\right). \tag{14}$$

4. Finally, for each gene $g \in \{1, \ldots, G\}$, sample gene expression level

$$x_c^{(g)} \sim \mathsf{Binom}\left(N_{\text{UMI}}\ ,\ \ 1 - e^{-q\,h\left(\lambda_c^{(g)}\right)}\right) \tag{15}$$

as in Eq. (2), for positive link function h.