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

Miriam Shiffman[1,2,5], **Will Stephenson**[2], **Geoffrey Schiebinger**[3,5], **Trevor Campbell**[2,3], **Jonathan Huggins**[2], **Aviv Regev**[4,5], **Tamara Broderick**[2,3]

[1] Computational & Systems Biology, [2] CSAIL, [3] IDSS, [4] Biology @ [5] BROAD INSTITUTE

abstract

- interested in understanding ***cellular differentiation*** — how one cell type gives rise to others
- **but** must infer this unobserved, inherently dynamic process from static, noisy snapshots (e.g. scRNA-seq)
- ***Bayesian nonparametric*** approach provides flexible tree densities & principled uncertainties

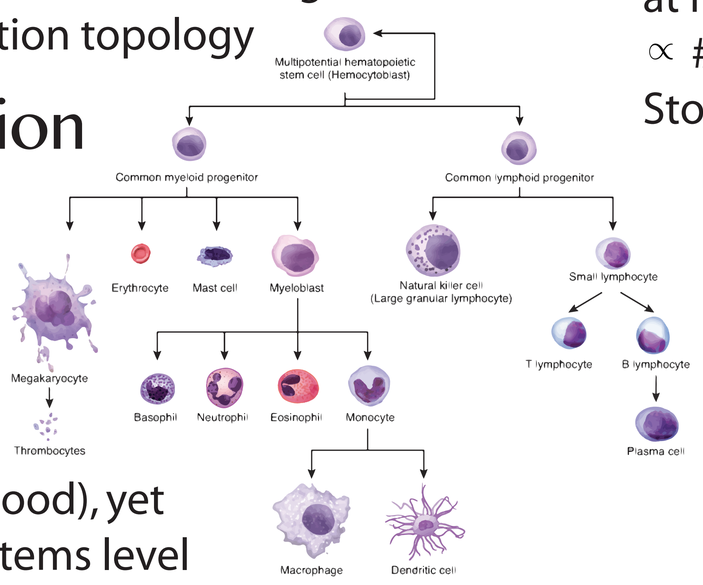
Here, **we develop a full generative model and inference** (novel MCMC sampler) to:

1) directly model **sources of uncertainty** in single-cell transcriptomic data

2) infer **interpretable, probabilistic** insight into cell state & differentiation topology

biological motivation

Question: how does a less-specialized progenitor (stem cell) reliably give rise to many cell fates?



- ubiquitous to multicellular life (development + adulthood), yet not well-understood at systems level
  ↳ implications for **disease + basic science**
- insight into stable differentiation landscape
  $\Longleftrightarrow$ insight into **master regulatory wiring** / molecular program

Challenge #1: reconstruct dynamic process from many static snapshots

Challenge #2: do not directly observe cell state, but rather a noisy proxy — e.g. transcriptome (genes recently transcribed into mRNA)

model desiderata

Goal: reconstruct continuous differentiation process (where latent cell states form a tree)

- learn tree **topology** of differentiation
- learn where **cells** fall along tree
- learn how **genes** drive branching, how expression changes along branches

Given:

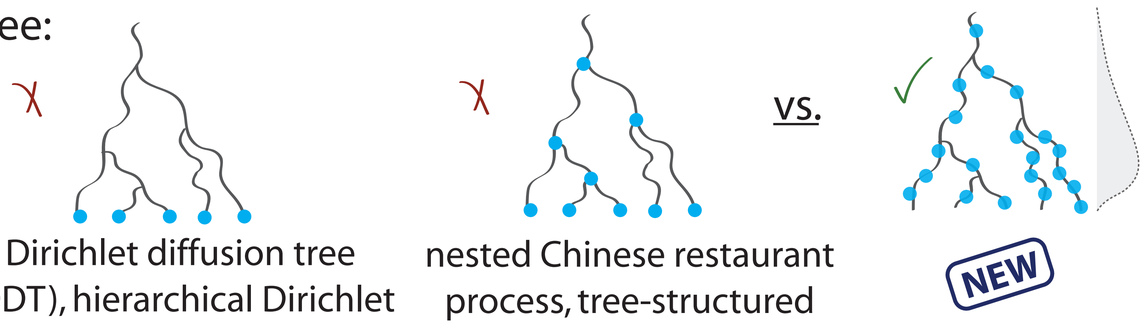
$$X = \underset{\text{cells } (10^3 - 10^5)}{\Big[\quad\Big]}_{\text{genes } \in \mathbb{N}^{20,000}}$$

mRNA molecule counts (sparse!)

- holistic inference of **count** data
- coherent **uncertainties**

Bayesian tree models

Probabilistic tree models allow for explicit **quantification of uncertainty** & flexible **binary branch topologies**, but existing models do not allow data to live arbitrarily along tree:



Dirichlet diffusion tree (DDT), hierarchical Dirichlet process, Kingman's coalescent

nested Chinese restaurant process, tree-structured stick-breaking

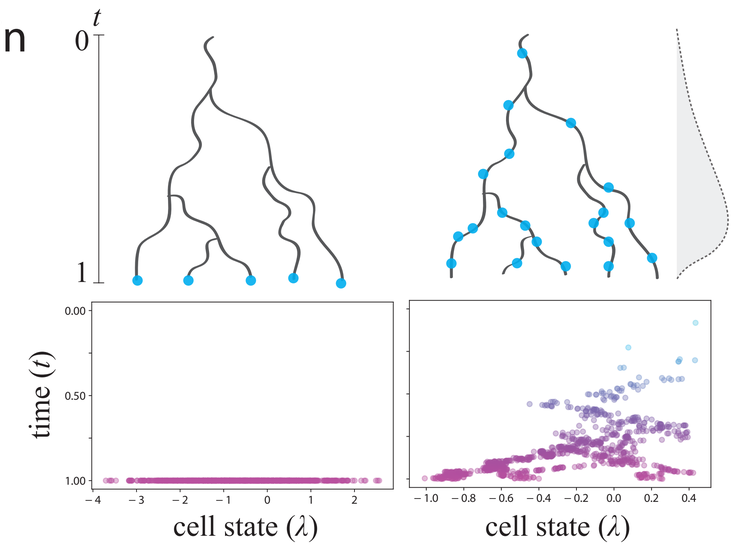
vs.

NEW

our augmented Dirichlet diffusion tree model

Recall: to perform inference, we begin by writing out the generative model, then use Bayes' rule to invert it.

1) **Generate tree** (node states + times) according to *K*-leaf DDT
   - Gaussian diffusion b/w nodes
   - branch times via hazard process
   - regularize depth with prior over *K*



classical DDT (L) vs. our model (R), top: cartoon of tree and data points; bottom: 1-d latent states simulated from each model (colored by time)

2) **Generate cells along tree**: each cell traverses the tree, starting at root & choosing branches w.p. $\propto$ # of cells down each path. Stop at time drawn from continuous distribution over (0,1).

For each gene:

i) draw latent cell state (λ) according to Brownian bridge defined by latent states of neighbors

ii) draw mRNA transcript count,

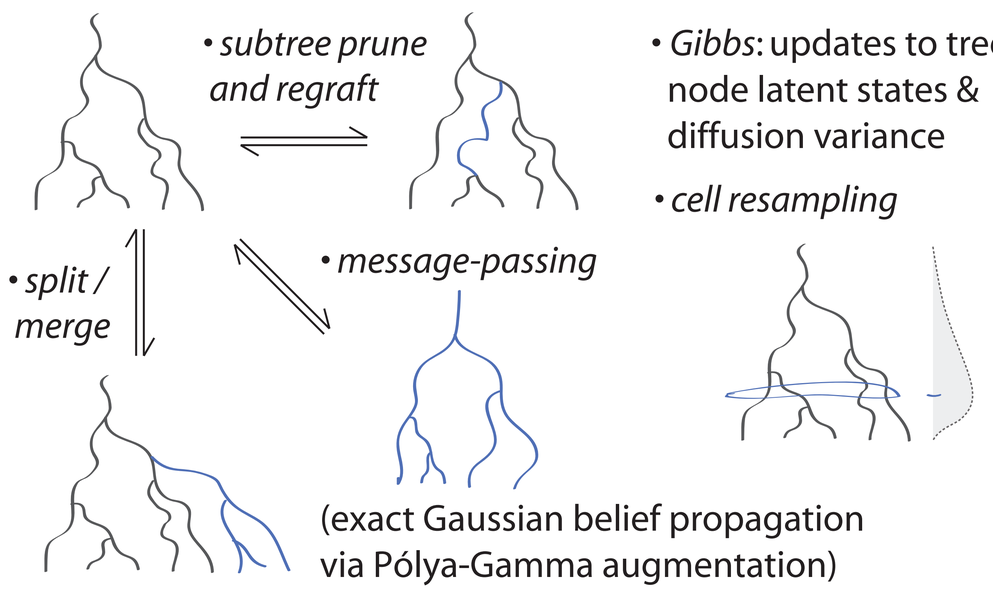
$$x \sim \text{Binom}\left(N_{\text{UMI}},\ 1 - e^{-q\,h(\lambda)}\right)$$

from summing over nonzero Poissons

- N_{UMI}: max countable transcripts/gene
- q: hyperparam related to dropout
- h: positive link, $h : \mathbb{R} \to \mathbb{R}_{\geq 0}$

inference

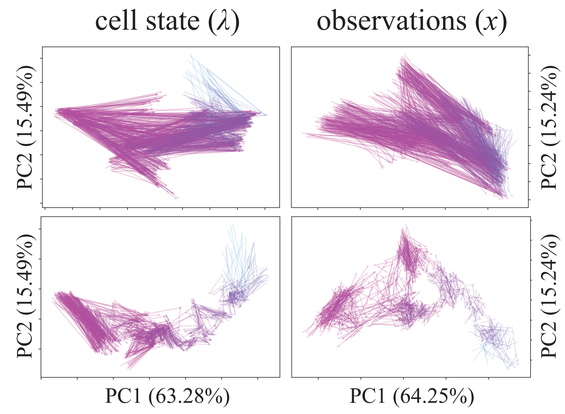
As usual, exact inference intractable. Instead: approximate inference via novel **Markov chain Monte Carlo** sampler...



(exact Gaussian belief propagation via Pólya-Gamma augmentation)

experiments

Partial **recovery of latent structure** from simulated data, including recovery of the true # of leaves.



convergence to ground truth demonstrated by shortening of arrow from true to inferred value per "cell," from initialized tree (top) to max-likelihood sampled (bottom)

future / ongoing

- beyond toy data
- tree metrics
- cell fate determinism vs. stochasticity
- test predictions by experimental perturbation

contact: shiffman@{mit.edu, broadinstitute.org} www.mit.edu/~shiffman @meereve