ITMO UNIVERSITY

CT Lab
ITMO UNIVERSITY

Pseudotime cell trajectories in scRNA-seq

Konstantin Zaitsev, ITMO University

Masters program, Feb 12th, 2020

scRNA-seq is a static snapshot

- But we still would like to understand some dynamics (possible time events)
- Pseudotime, differentiation, polarization and other stuff

Language

- Differentiation is transcriptional change
  (not necessarily change of the cells type)
- Lets first assume we have a trajectory/transition:
  cell type A differentiate to cell type B and in our dataset, we have cell from both A and B (and maybe something from between A and B)

Pseudotime

- If we have a transition A -> B and we assume all the cells to “participate” in this transition
- Pseudotime is “How transcriptionally different is your cell from the start of the transition? How mush progress did the cell make?”
- Once we know can order cells by pseudotime we can identify important transition-related genes

nature biotechnology

LETTERS

ITMO UNIVERSITY

The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells

Cole Trapnell[1,2,6], Davide Cacchiarelli[1–3,6], Jonna Grimsby[2], Prapti Pokharel[2], Shuqiang Li[4], Michael Morse[1,2], Niall J Lennon[2], Kenneth J Livak[4], Tarjei S Mikkelsen[1–3] & John L Rinn[1,2,5]

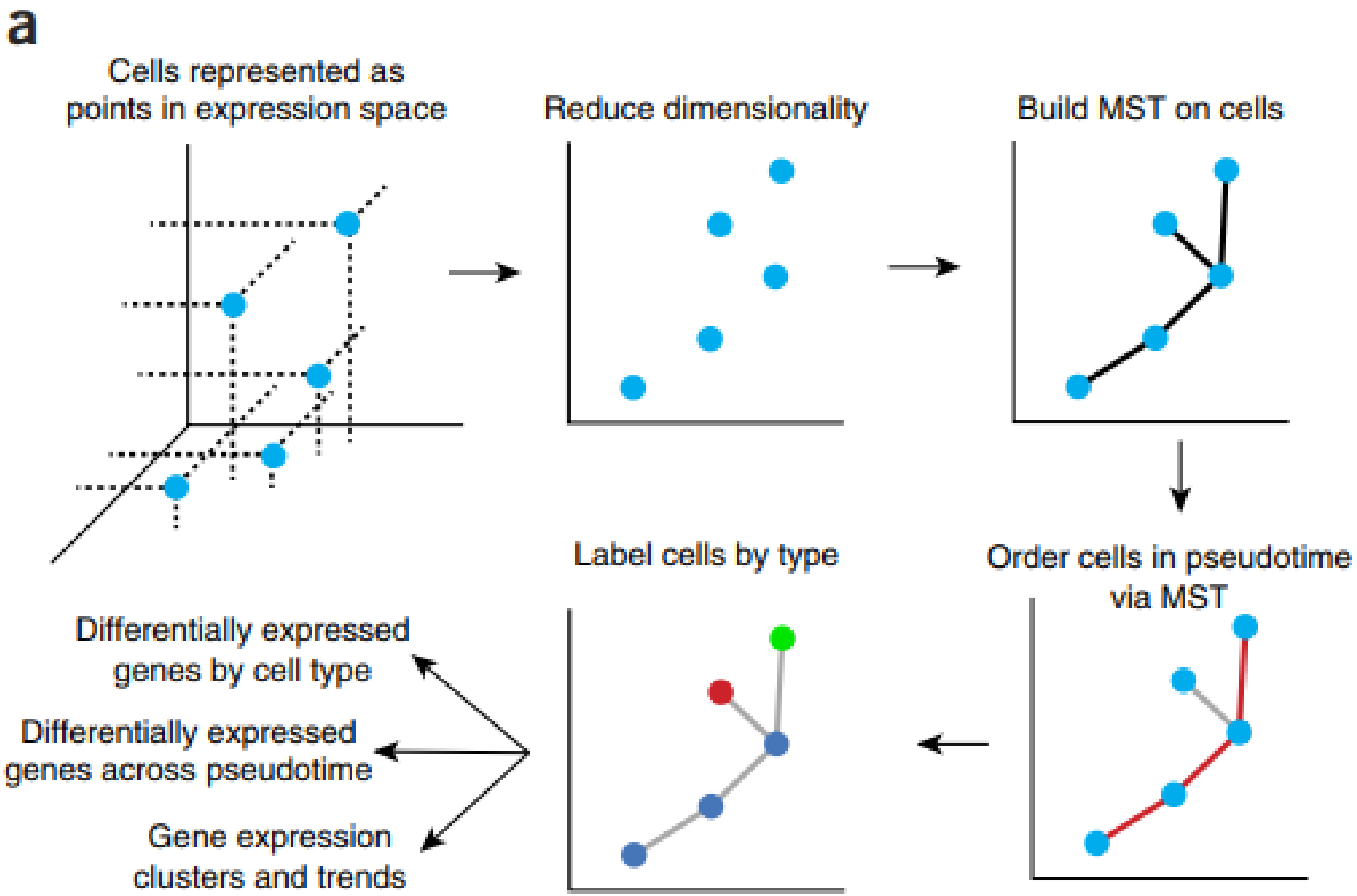
Defining the transcriptional dynamics of a temporal process such as cell differentiation is challenging owing to the high variability in gene expression between individual cells. Time-series gene expression analyses of bulk cells have difficulty distinguishing early and late phases of a transcriptional cascade or identifying rare subpopulations of cells, and single-cell proteomic methods rely on a priori knowledge of key distinguishing markers[1]. Here we describe Monocle, an unsupervised algorithm that increases the temporal resolution of transcriptome dynamics using single-cell RNA-Seq data collected at multiple time points. Applied to the differentiation of primary human myoblasts, Monocle revealed switch-like changes in expression of key regulatory factors, sequential waves of gene regulation, and expression of regulators that were not known to act in differentiation. We validated some of these predicted regulators in a loss-of function screen. Monocle can in principle be used to recover single-cell gene expression

Such averaging artifacts can make factors that are correlated appear to be uncorrelated or even make positively correlated factors appear negatively correlated. As a population of cells captured at the same time may include many distinct intermediate differentiation states, considering only its average properties would mask trends occurring across individual cells. Solving this problem by experimental synchronization of cells or by stringent isolation of precursors at distinct stages is challenging and can sharply alter differentiation kinetics.

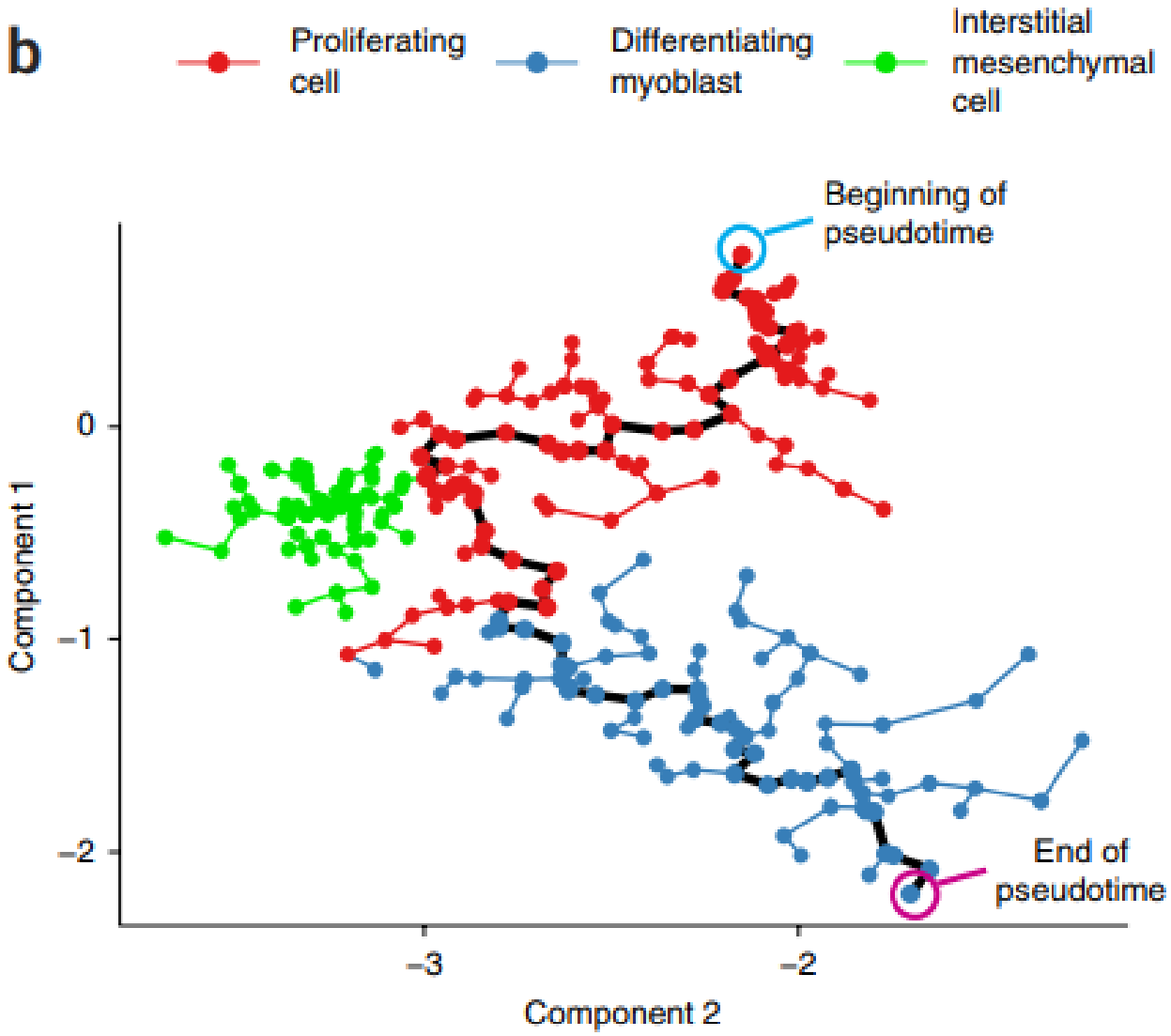
Computational analysis of gene expression data could help define biological progression between cellular states and reveal regulatory modules of genes that co-vary in expression across individual cells[9]. Previous analyses have used approaches from computational geometry[10,11] to order bulk cell populations from time-series microarray experiments by progress through a biological process independently of when the samples were collected. The recently developed SPD algorithm can resolve progression along multiple lineages arising from a progenitor cell type using supervised machine learning[12]. However,

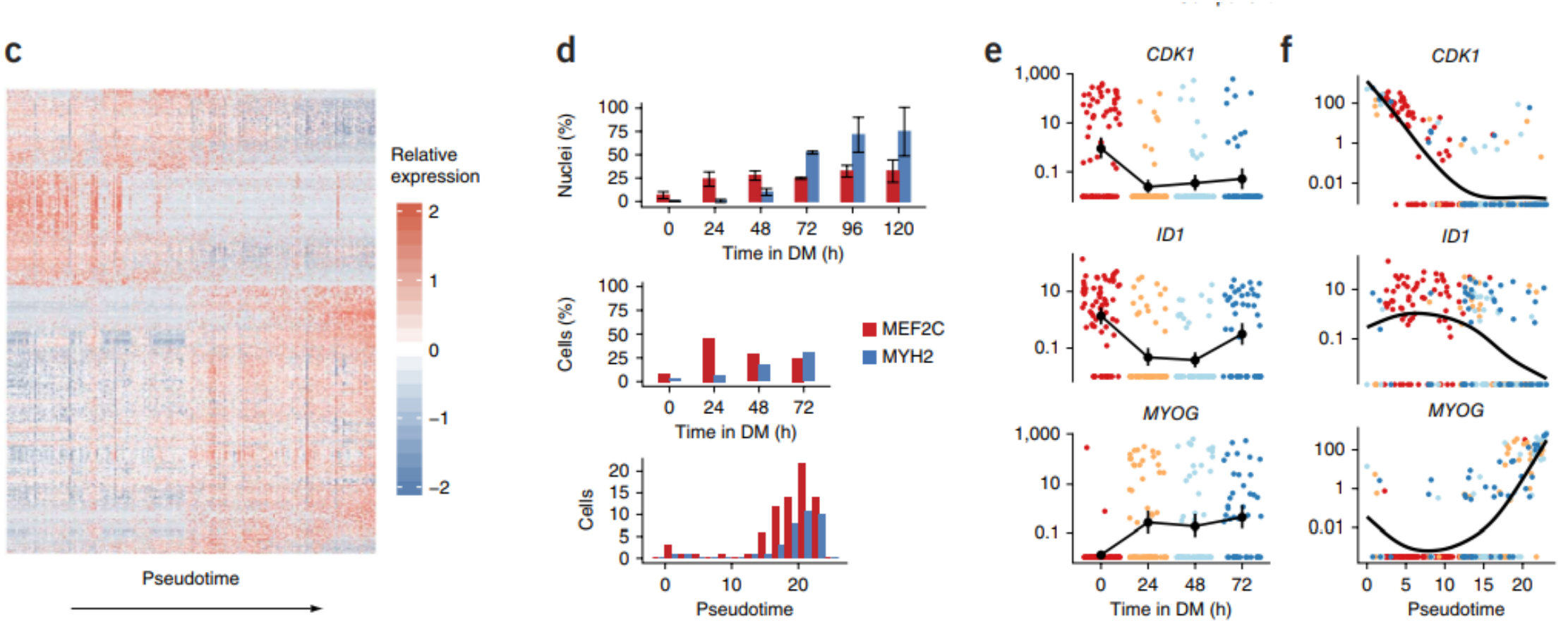
© 2014 Nature America, Inc. All rights reserved.

- This paper is about Monocle v1 2014


a
Cells represented as points in expression space
Reduce dimensionality
Build MST on cells
Order cells in pseudotime via MST
Label cells by type
Differentially expressed genes by cell type
Differentially expressed genes across pseudotime
Gene expression clusters and trends

of coarse kinetic trends[16]. We expanded primary human skeletal muscle myoblasts (HSMM) under high-mitogen conditions (GM) and induced differentiation by switching to low-serum medium (DM). We captured between 49 and 77 cells at each of four time points after the switch to DM using the Fluidigm C_1 microfluidic system. RNA from each cell was isolated and used to construct a single mRNA-Seq library per cell, which was then sequenced to a depth of ~4 million reads per library, resulting in a complete gene expression profile for each cell (**Fig. 1a** and **Supplementary Fig. 1**).




c
Relative expression
2
1
0
−1
−2
Pseudotime
d
Nuclei (%)
100
75
50
25
0
0 24 48 72 96 120
Time in DM (h)
Cells (%)
100
75
50
25
0
0 24 48 72
Time in DM (h)
MEF2C
MYH2
Cells
20
15
10
5
0
0 10 20
Pseudotime
e
CDK1
1,000
10
0.1
ID1
10
0.1
MYOG
1,000
10
0.1
0 24 48 72
Time in DM (h)
f
CDK1
100
1
0.01
ID1
10
0.1
MYOG
100
1
0.01
0 5 10 15 20
Pseudotime

ilmilarly, children of Q nodes might need to be emitted in reverse order to nake smooth transitions in the final ordering of cells. Magwene *et al.*[10] do not ddress this situation, preferring to emit the PQ tree itself. Monocle always mits an ordering of cells, and thus it exhaustively searches orderings encoded y the PQ tree to find one that obeys its constraints and minimizes the total listance traveled by the resulting polygonal reconstruction in the embedding eometry $\mathbb{R}^d$, beginning at one end of the diameter path of the full MST. While his might result in superpolynomial running time and memory, in practice vith real data, this procedure takes only a few seconds and small amount of AM on a laptop because the number of cells in P nodes is typically very small elative to the cells in Q nodes.

Monocle thus emits an ordering of the cells that relies on the MST to 'sketch' he basic shape of the polygonal reconstruction and uses the PQ tree to handle

This paper is about
Monocle v2
2017

Reversed graph embedding resolves complex single-cell trajectories

Xiaojie Qiu[1,2], Qi Mao[3], Ying Tang[4], Li Wang[5], Raghav Chawla[2], Hannah A Pliner[2] & Cole Trapnell[1,2]

Single-cell trajectories can unveil how gene regulation governs cell fate decisions. However, learning the structure of complex trajectories with multiple branches remains a challenging computational problem. We present Monocle 2, an algorithm that uses reversed graph embedding to describe multiple fate decisions in a fully unsupervised manner. We applied Monocle 2 to two studies of blood development and found that mutations

Monocle 2 begins by identifying genes that define a biological process using an unsupervised procedure we term 'dpFeature'. The procedure works by selecting the genes that are differentially expressed between clusters of cells identified with *t*-distributed stochastic neighbor embedding (tSNE) dimension reduction followed by density-peak clustering[7]. When applied to four different data sets[1,8–10] most of the genes returned by dpFeature were also recovered by a semisupervised selection method guided by aspects of the experimental design and were highly enriched for relevant Gene Ontology terms, confirming that dpFeature is a powerful and general approach for unsupervised feature selection (**Supplementary Figs. 1–3**).

To develop a pseudotime-trajectory-reconstruction algorithm that does not require cell fate or branch numbers as input, we used RGE[5,6], a machine-learning technique to learn a parsimonious 'principal graph'. Informally, a principal graph is like a principal curve[11] that passes through the 'middle' of a data set but is allowed to have branches[12]. However, learning a principal graph that describes a population of scRNA-seq profiles is very challenging because each expressed gene adds an additional dimension to the gene expression space, and learning geometry is dramatically

ger Nature. All rights reserved.

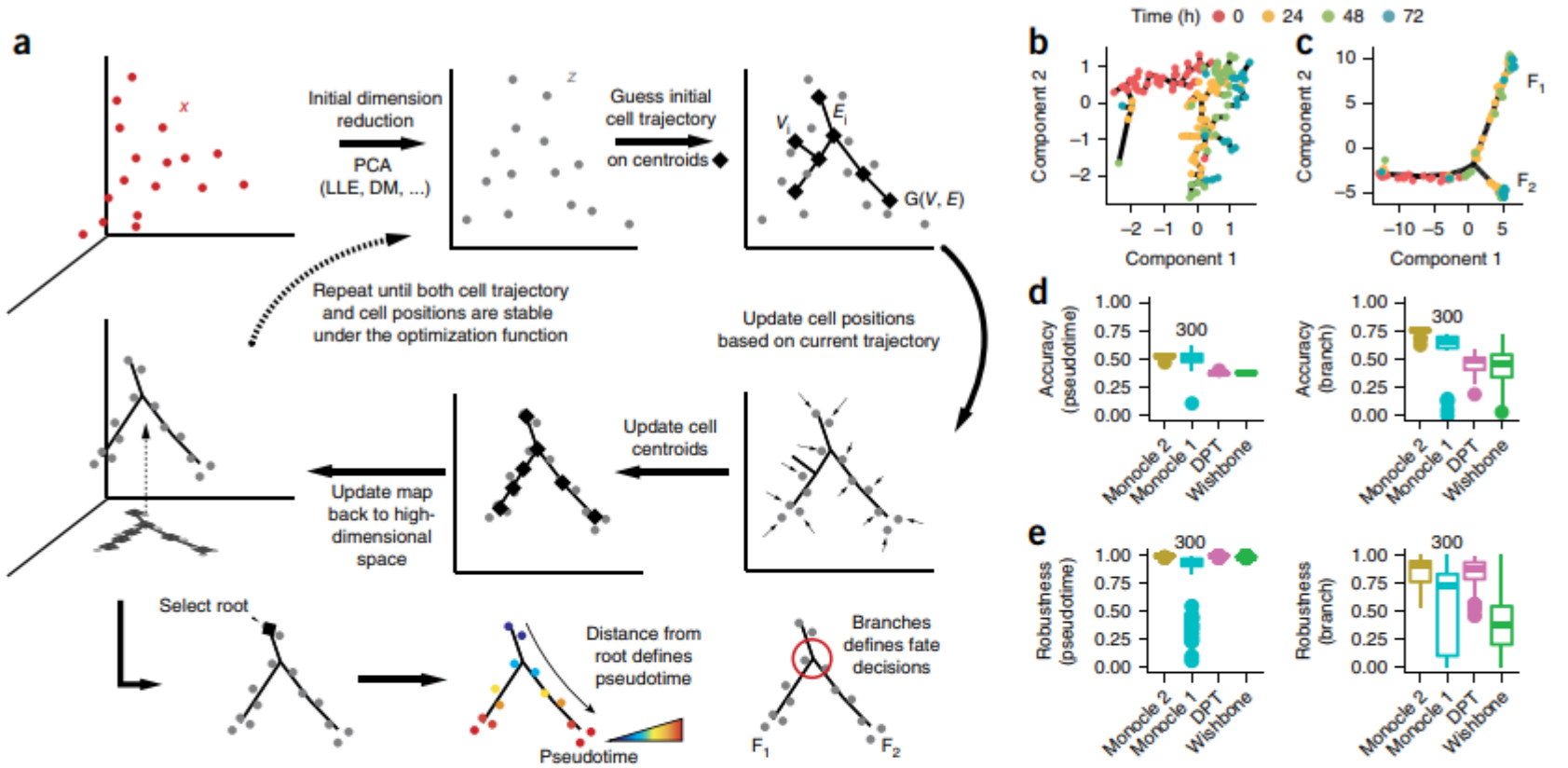


Figure 1 | Monocle 2 discovers a cryptic alternative outcome in myoblast differentiation. (**a**) Monocle 2 automatically learns single-cell trajectories and branch points by reversed graph embedding (Online Methods). Each cell is represented as a point in high-dimensional space (x), where each dimension corresponds to the expression level of an ordering gene. Data are projected onto a lower-dimensional space (z) by dimension-reduction methods such as PCA, and Monocle 2 constructs a spanning tree on a set of centroids (diamonds) chosen automatically using k-means clustering. Cells are then shifted toward the nearest tree vertex, vertex positions are updated to 'fit' cells, a new spanning tree is learned, and the process is iterated until the tree and cells converge. The user then selects a tip as the 'root', each cell's pseudotime is calculated as its geodesic distance along the tree to the root, and its branch is automatically assigned based on the principal graph. (**b**) Differentiating human skeletal myoblasts projected onto the first two components from an ICA by Monocle 1. Black segments indicate cells connected in a minimum spanning tree. (**c**) Cells from **b** projected into a two-dimensional space by Monocle 2 using DDRTree. Black segments indicate the graph learned as illustrated in **a**. The components are distinct from those shown in **b**. Trajectory outcomes are indicated as F_1 and F_2. (**d**,**e**) Accuracy (**d**) and consistency (**e**) of pseudotime calculation (left) or branch assignments (right) from each algorithm under repeated subsampling of 80% of the cells on the Paul *et al.* data set[10]. For **d**, a marker-based ordering scheme was used as ground truth (Online Methods). For **e**, all pairwise downsamplings were used to calculate the Pearson's rho and adjusted Rand index (ARI). For benchmarking, Monocle 2, DPT, and Wishbone used the full data set, whereas Monocle 1 used only a random downsample of 300 cells.

Street *et al. BMC Genomics* (2018) 19:477
https://doi.org/10.1186/s12864-018-4772-0

BMC Genomics

METHODOLOGY ARTICLE Open Access

Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics

Kelly Street[1,8], Davide Risso[2], Russell B. Fletcher[3], Diya Das[3,9], John Ngai[3,6,7], Nir Yosef[4,8], Elizabeth Purdom[5,8] and Sandrine Dudoit[1,5,8,9*]

Abstract

Background: Single-cell transcriptomics allows researchers to investigate complex communities of heterogeneous cells. It can be applied to stem cells and their descendants in order to chart the progression from multipotent progenitors to fully differentiated cells. While a variety of statistical and computational methods have been proposed for inferring cell lineages, the problem of accurately characterizing multiple branching lineages remains difficult to solve.

Results: We introduce Slingshot, a novel method for inferring cell lineages and pseudotimes from single-cell gene expression data. In previously published datasets, Slingshot correctly identifies the biological signal for one to three branching trajectories. Additionally, our simulation study shows that Slingshot infers more accurate pseudotimes than other leading methods.

Conclusions: Slingshot is a uniquely robust and flexible tool which combines the highly stable techniques necessary for noisy single-cell data with the ability to identify multiple trajectories. Accurate lineage inference is a critical step in the identification of dynamic temporal gene expression.

Keywords: RNA-Seq, Single cell, Lineage inference, Pseudotime inference

Table 1 Summary of existing lineage and pseudotime inference methods

	Dimensionality reduction	Cluster based	Graph	Pseudotime calculation	Branching	Supervision
Diffusion Pseudotime	Diffusion maps	No	Weighted k-NN graph on cells	Transition probabilities over arbitrary length random walks	Yes	Starting cell
Embeddr	Laplacian eigenmaps	No	N/A	Principal curve, orthogonal projection	No	Path direction[1], subsetting[2]
Monocle	ICA	No	MST on cells	Diameter path, PQ trees	Yes[3]	Path direction[1], number of lineages
Monocle 2	Reversed graph embedding	No	Principal graph on cells	Distance to root	Yes	Starting cluster
TSCAN	PCA	Yes	MST on clusters	Cluster centers, orthogonal projection	Yes	Starting cluster
Waterfall	PCA	Yes	MST on clusters	Cluster centers, orthogonal projection	Yes[4]	Path direction[1]
Wishbone	Diffusion maps	No	Ensemble of k-NN graphs on cells	Distance refinement by waypoints	Yes[5]	Starting cell
Slingshot	Any	Yes	MST on clusters	Simultaneous principal curves, orthogonal projection	Yes	Starting cluster, end clusters (optional)

[1] Some methods infer a single path or backbone and rely on the user to assign its directionality
[2] Methods that do not detect branching events require manually subsetting the data down to a single lineage
[3] Monocle does not detect the number of branching events, the number of lineages must be supplied by the user
[4] Waterfall detects branching events, but requires subsetting to a single lineage for pseudotime calculation
[5] Wishbone can only detect a single branching event (two lineages)

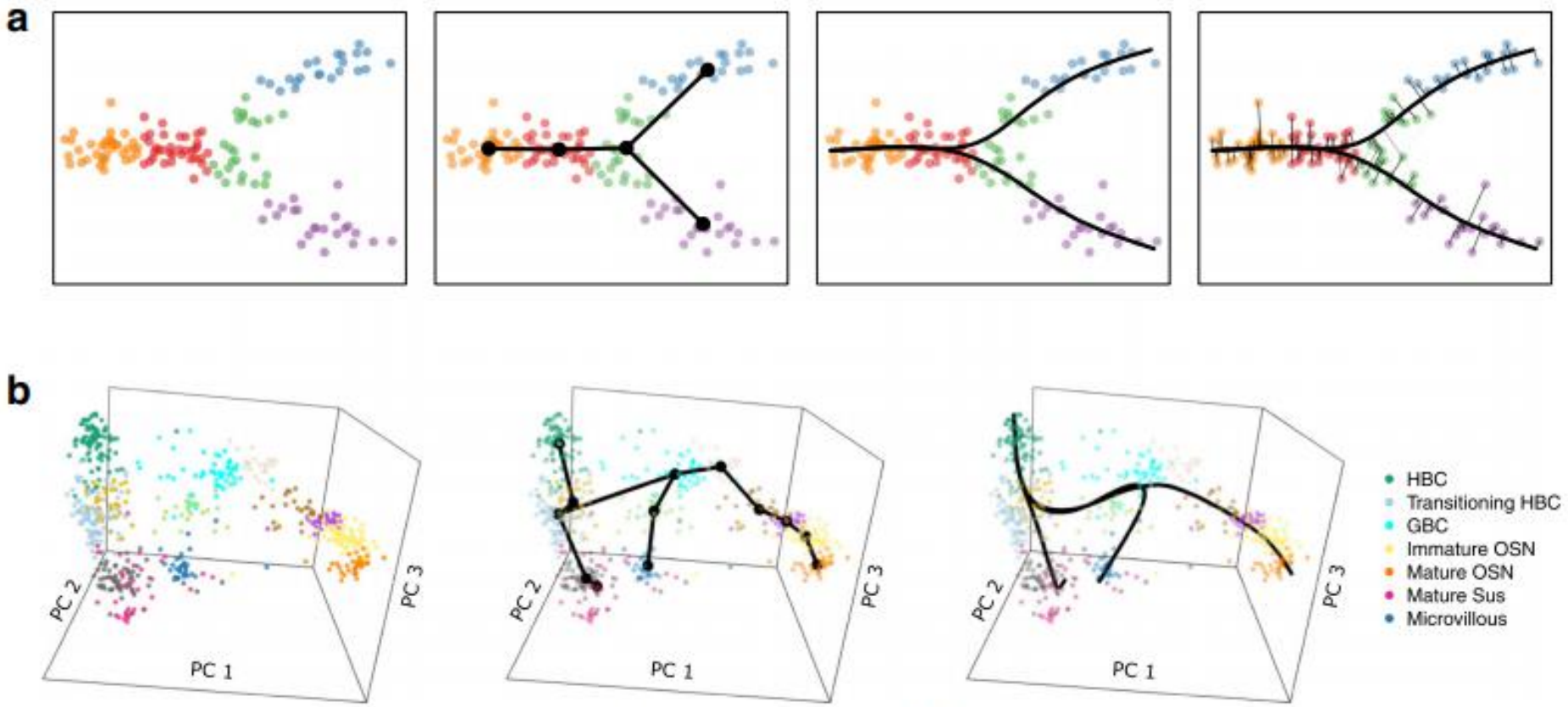


Fig. 1 Schematics of Slingshot's main steps. The main steps for Slingshot are shown for: Panel (**a**) a simple simulated two-lineage two-dimensional dataset and Panel (**b**) the single-cell RNA-Seq olfactory epithelium three-lineage dataset of [26] (see Results and discussion for details on dataset and its analysis). Step 0: Slingshot starts from clustered data in a low-dimensional space (cluster labels indicated by color). For Panel (**b**), the plot shows the top three principal components, but Slingshot was run on the top five. Step 1: A minimum spanning tree is constructed on the clusters to determine the number and rough shape of lineages. For Panel (**b**), we impose some constraints on the MST based on known biology. Step 2: Simultaneous principal curves are used to obtain smooth representations of each lineage. Step 3: Pseudotime values are obtained by orthogonal projection onto the curves (only shown for Panel (**a**))

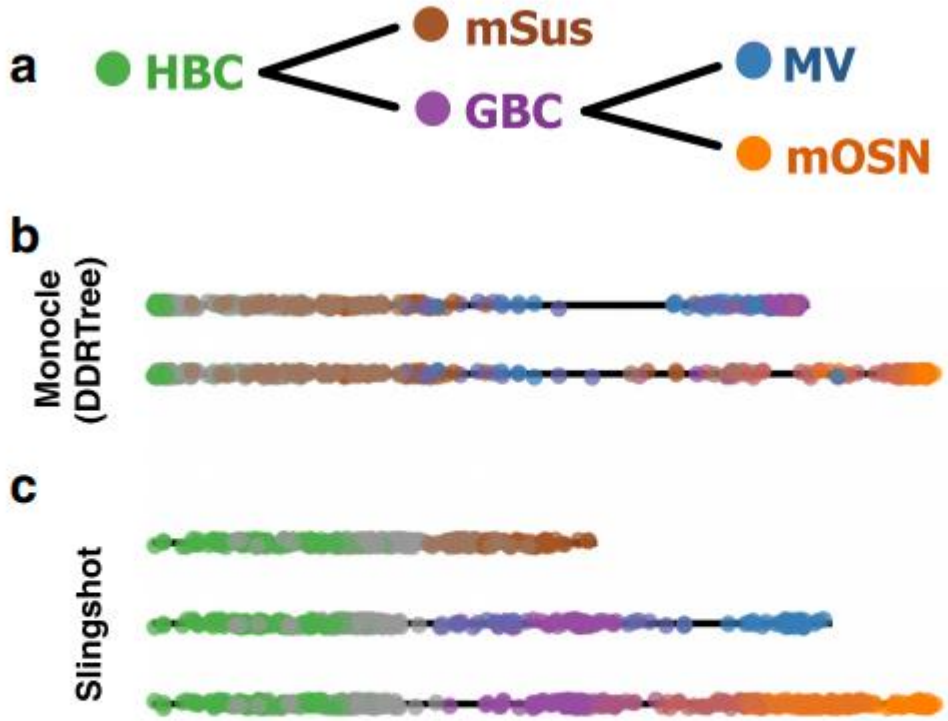


Fig. 3 Multiple lineage inference: OE dataset. Pseudotime variables for each lineage inferred by Slingshot and Monocle 2 on the three-lineage OE dataset of [26]. Panel (**a**): Known biological relationships between cell types. Panel (**b**): For Monocle 2, we used the DDRTree algorithm to obtain a two-dimensional (or five-dimensional, see Additional file 1: Figure S3d) representation of the data and selected the starting state based on the highest percentage of cells from the HBC cluster. Panel (**c**): For Slingshot, we used the top five PCs and clustered cells by RSEC, as in the original article. The HBC cluster was specified as the origin and the mSus cluster as an endpoint; other endpoints were identified without supervision

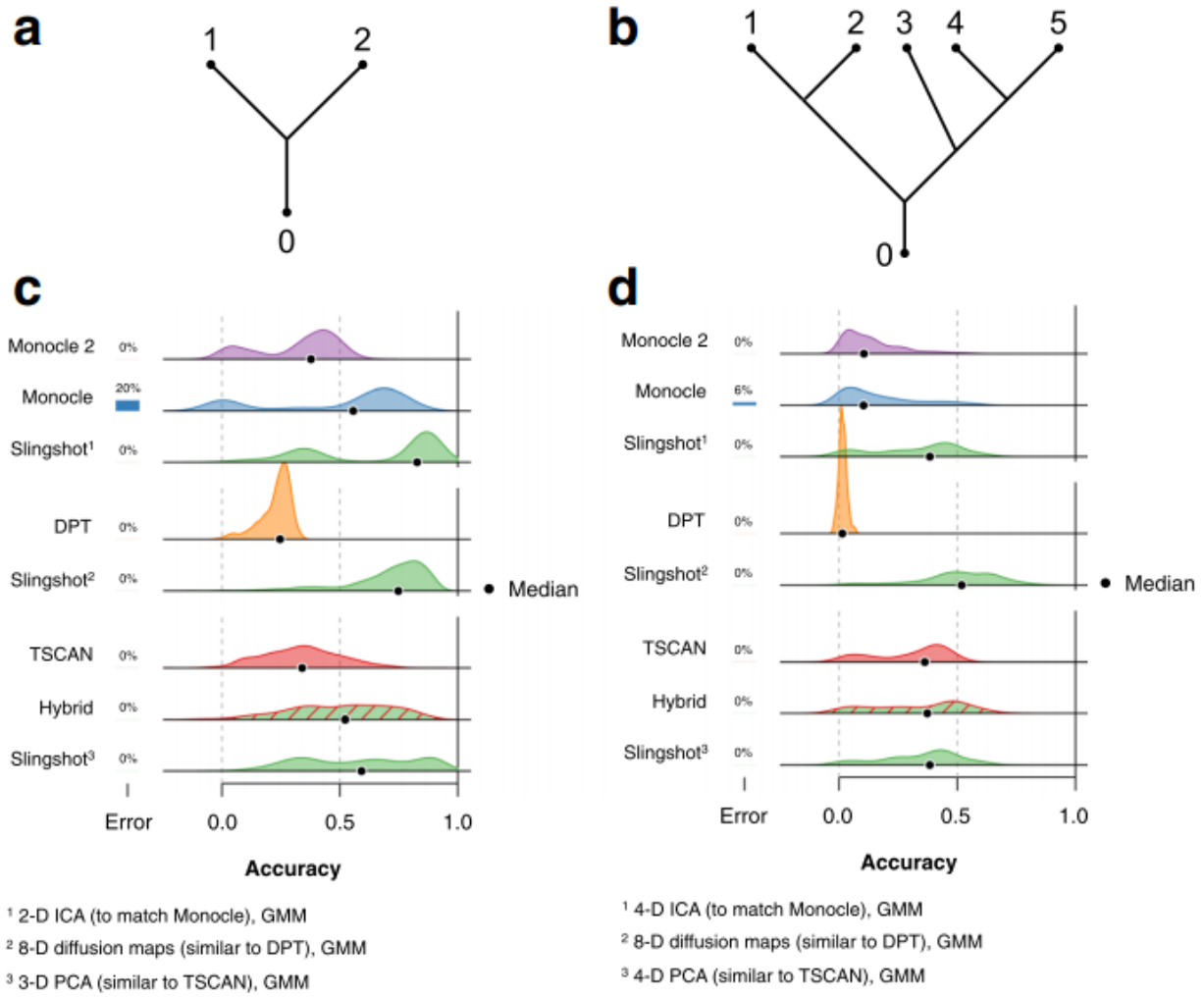


Fig. 4 Comparison of accuracy scores for lineage and pseudotime inference methods: Simulated datasets. Gaussian kernel density plots of accuracy scores show how five lineage inference methods performed on a series of simulated datasets with two different topologies: Panels (**a**,**c**) two lineages and Panels (**b**,**d**) five lineages. In both settings, the simulated data contained variable numbers of cells and levels of noise. Bars to the left of each density plot represent the percentage of datasets on which a method returned an error. Errors are treated as 0 values for calculating the median score, but are not included in the density estimates. Monocle, Monocle 2, DPT, and TSCAN were implemented in several ways and these densities represent the best results obtained by each method. Slingshot was implemented with various dimensionality reduction techniques, chosen to match the best-case settings of the other methods and with clusters assigned by Gaussian mixture modeling (GMM). See Simulation study for the definition of accuracy scores based on Kendall's rank correlation coefficient and Additional file 1 for details on simulation scenarios

RNA velocity

- Spliced RNA in the cell is a static snapshot
- Unspliced RNA in the cell is an “intention” of the cell

RNA abundance is a powerful indicator of the state of individual cells. Single-cell RNA sequencing can reveal RNA abundance with high quantitative accuracy, sensitivity and throughput[1]. However, this approach captures only a static snapshot at a point in time, posing a challenge for the analysis of time-resolved phenomena such as embryogenesis or tissue regeneration. Here we show that RNA velocity—the time derivative of the gene expression state—can be directly estimated by distinguishing between unspliced and spliced mRNAs in common single-cell RNA sequencing protocols. RNA velocity is a high-dimensional vector that predicts the future state of individual cells on a timescale of hours. We validate its accuracy in the neural crest lineage, demonstrate its use on multiple published datasets and technical platforms, reveal the branching lineage tree of the developing mouse hippocampus, and examine the kinetics of transcription in human embryonic brain. We expect RNA velocity to greatly aid the analysis of developmental lineages and cellular dynamics, particularly in humans.

ITMO UNIVERSITY

RNA velocity

LETTER

https://doi.org/10.1038/s41586-018-0414-6

RNA velocity of single cells

Gioele La Manno[1,2], Ruslan Soldatov[3], Amit Zeisel[1,2], Emelie Braun[1,2], Hannah Hochgerner[1,2], Viktor Petukhov[3,4], Katja Lidschreiber[5], Maria E. Kastriti[6], Peter Lönnerberg[1,2], Alessandro Furlan[1], Jean Fan[3], Lars E. Borm[1,2], Zehua Liu[3], David van Bruggen[1], Jimin Guo[3], Xiaoling He[7], Roger Barker[7], Erik Sundström[8], Gonçalo Castelo-Branco[1], Patrick Cramer[5,9], Igor Adameyko[6], Sten Linnarsson[1,2]* & Peter V. Kharchenko[3,10]*

RNA abundance is a powerful indicator of the state of individual cells. Single-cell RNA sequencing can reveal RNA abundance with high quantitative accuracy, sensitivity and throughput[1]. However, this approach captures only a static snapshot at a point in time, posing a challenge for the analysis of time-resolved phenomena such as embryogenesis or tissue regeneration. Here we show that RNA velocity—the time derivative of the gene expression state—can be directly estimated by distinguishing between unspliced and spliced mRNAs in common single-cell RNA sequencing protocols. RNA velocity is a high-dimensional vector that predicts the future state of individual cells on a timescale of hours. We validate its accuracy in the neural crest lineage, demonstrate its use on multiple published datasets and technical platforms, reveal the branching lineage tree of the developing mouse hippocampus, and examine the kinetics of transcription in human embryonic brain. We expect RNA velocity to greatly aid the analysis of developmental lineages and cellular dynamics, particularly in humans.

for transcriptional dynamics[2], in which the first time derivative of the spliced mRNA abundance (RNA velocity) is determined by the balance between production of spliced mRNA from unspliced mRNA, and the mRNA degradation (Fig. 1b and Supplementary Note 1). Under such a model, steady states are approached asymptotically when the rate of transcription α is constant, with the steady-state abundances of spliced (s) and unspliced (u) molecules determined by α, and constrained to a fixed-slope relationship where $u = \gamma s$ (Supplementary Note 2 Section 1). The equilibrium slope γ combines degradation and splicing rates, capturing gene-specific regulatory properties, the ratio of intronic and exonic lengths, and the number of internal priming sites. Using a recently published compendium of mouse tissues[11], we found that the steady-state behaviour of most genes across a wide range of cell types was consistent with a single fixed slope γ (Extended Data Fig. 3a–c). However, 11% of genes showed distinct slopes in different subsets of tissues (Extended Data Fig. 3d, e), suggesting tissue-specific alternative splicing (Extended Data Fig. 3f) or degradation rates.

- RNA velocity
- 2018, Nature

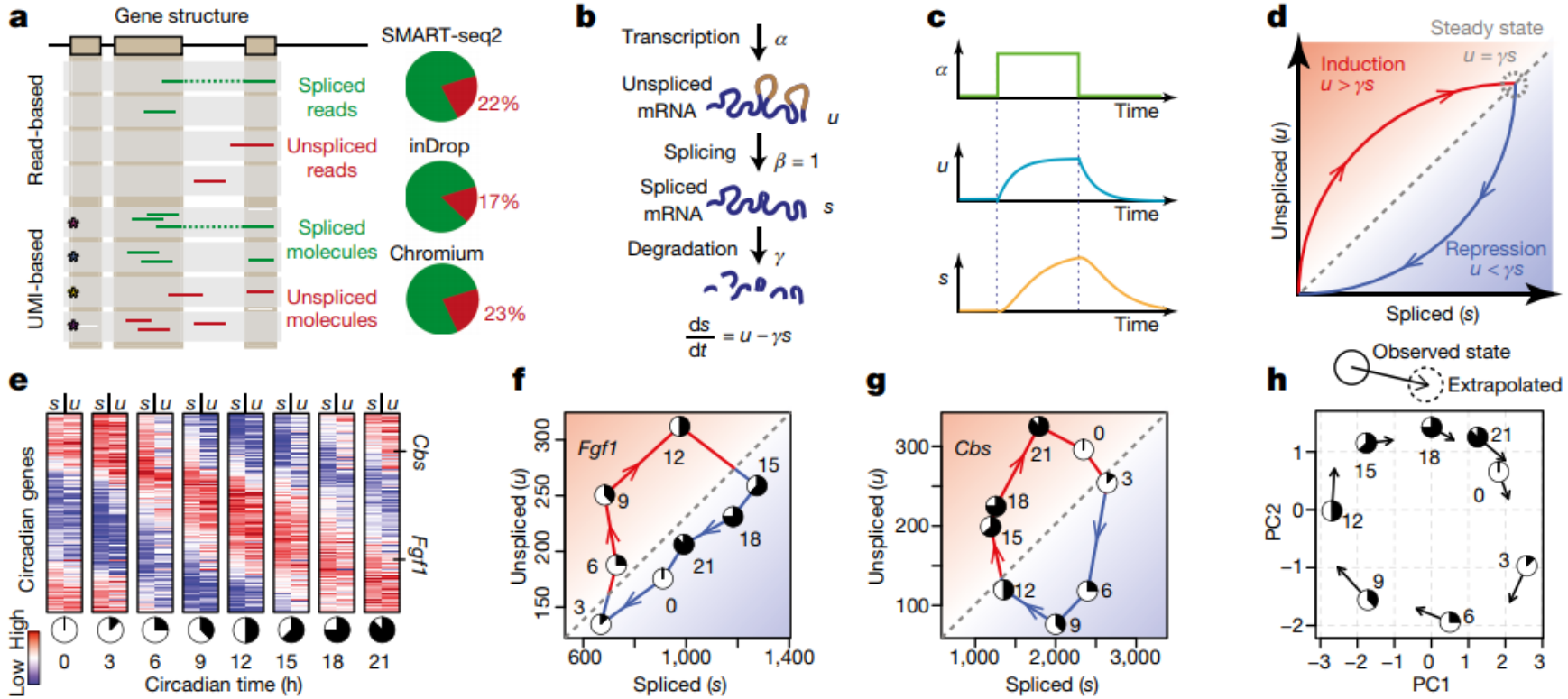


Fig. 1 | Balance between unspliced and spliced mRNAs is predictive of cellular state progression. a, Spliced and unspliced counts are estimated by separately counting reads that incorporate the intronic sequence. Multiple reads associated with a given molecule are grouped (boxes with asterisks) for unique molecular identifier (UMI)-based protocols. Pie charts show typical fractions of unspliced molecules. **b**, Model of transcriptional dynamics, capturing transcription (α), splicing (β) and degradation (γ) rates involved in production of unspliced (u) and spliced (s) mRNA products. **c**, Solution of the model in **b** as a function of time, showing unspliced and spliced mRNA dynamics in response to step changes in α. **d**, Phase portrait showing the same solution shown in **c** (solid curves). Steady states for different values of transcription rates α fall on the diagonal given by slope γ (dashed line). Levels of unspliced mRNA above or below this proportion indicate increasing (red shading) or decreasing (blue shading) expression of a gene, respectively. **e**, Abundance of spliced (s) and unspliced (u) mRNAs for circadian-associated genes in the mouse liver over a 24-h time course[12]. The unspliced mRNAs are predictive of spliced mRNA at the next time point. **f**, **g**, Phase portraits observed for a pair of circadian-driven genes: *Fgf1* (**f**) and *Cbs* (**g**). The circadian time of each point is shown using a clock symbol (corresponding to those in **e**). The dashed diagonal line shows the steady-state relationship, as predicted by γ fit. **h**, Change in expression state at a future time t, as predicted by the model, is shown in the space of the first two principal components (PCs), recapitulating the progression along the circadian cycle. Each circle shows the observed expression state, with the arrow pointing to the position of the future state, extrapolated from velocity estimates.

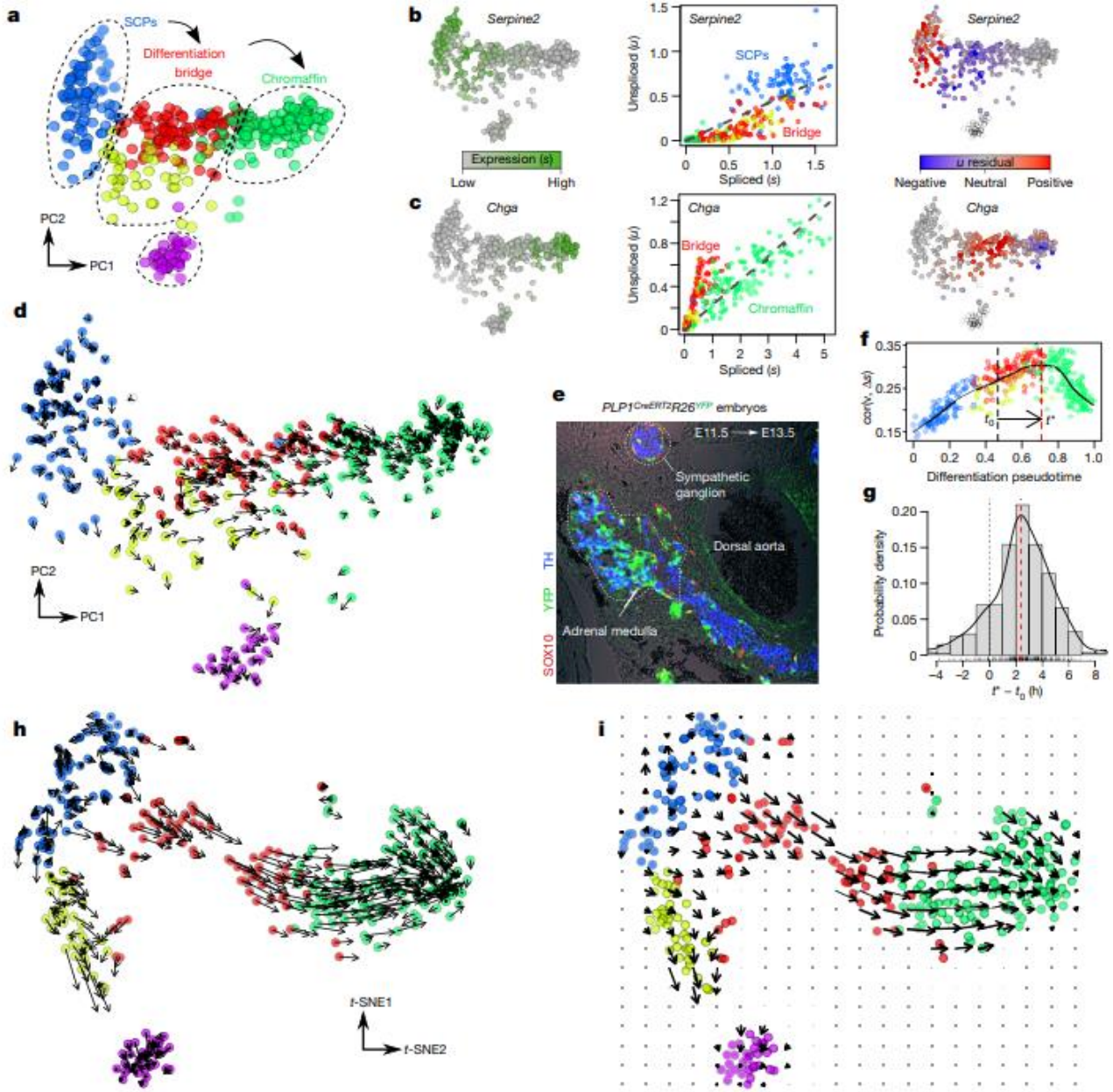


Fig. 2 | RNA velocity recapitulates dynamics of chromaffin cell differentiation. a, PCA projection showing major subpopulations of Schwann cell precursors (SCPs) differentiating into chromaffin cells in embryonic day (E)12.5 mice ($n = 385$ cells). **b**, **c**, Expression pattern (left), unspliced–spliced phase portraits (centre, cells coloured according to ($n = 3$ replicates). YFP labels Htr3a$^+$ bridge cells; TH marks chromaffin cells; TH$^+$YFP$^+$ marks chromaffin cells that are freshly differentiated from the bridge population. **f**, Extrapolation distance along the chromaffin differentiation trajectory is estimated for a single cell at pseudotime t_0, on the basis of the correlation (y axis) between the velocity v and

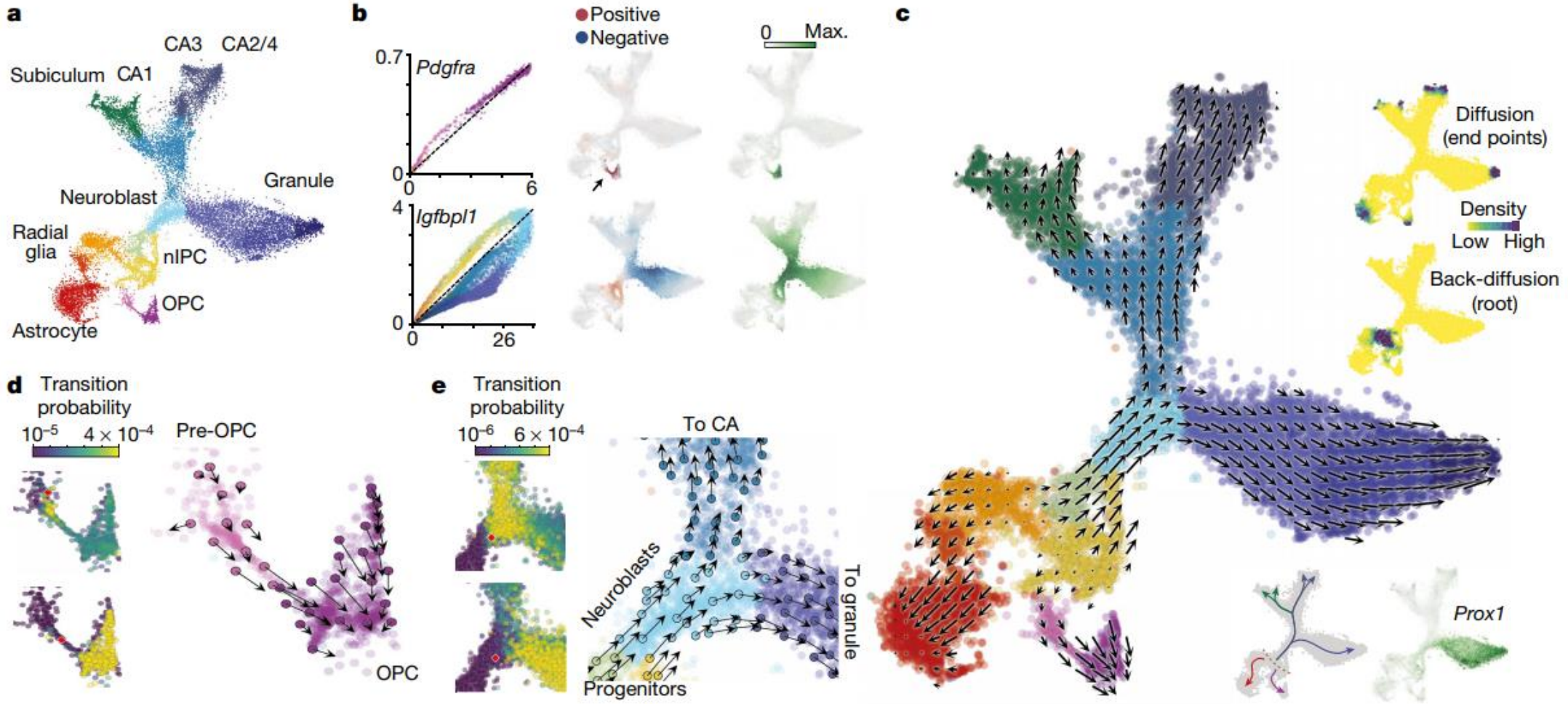


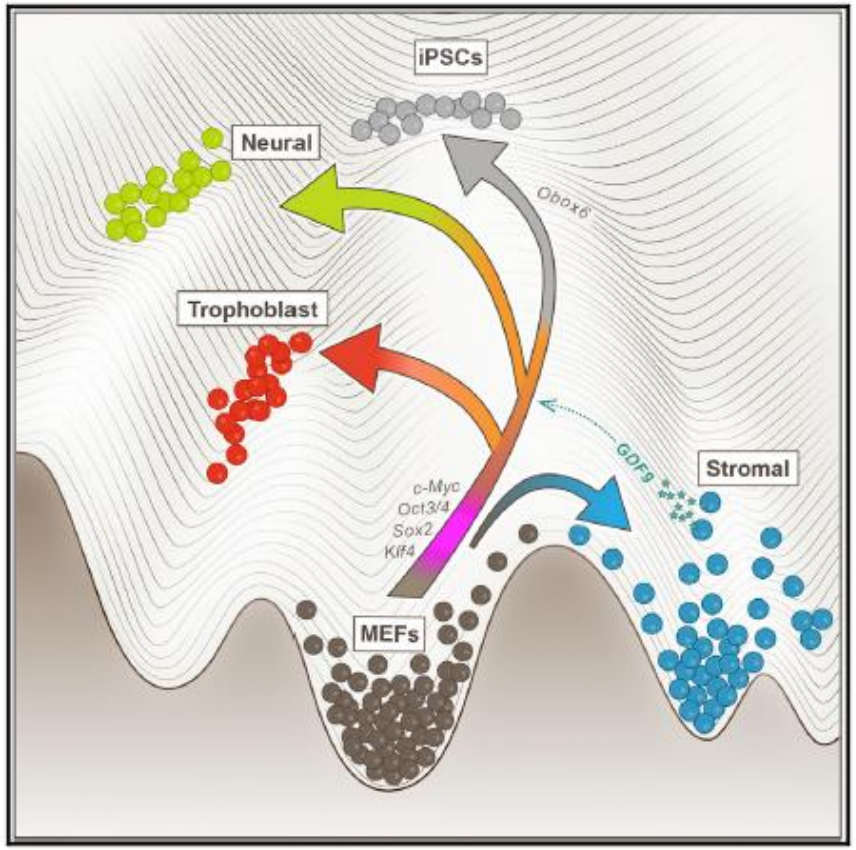
Fig. 3 | RNA velocity field describes fate decisions of major neural lineages in the hippocampus. a, *t*-SNE plot of the developing mouse hippocampus cells ($n = 18{,}213$ cells), showing major transient and mature subpopulations. **b**, Phase portraits (left, coloured as in **a**), unspliced residuals (middle) and spliced expression (right) are shown for two regulated genes. *k*-nearest neighbour (*k*NN) cell pooling was used. **c**, Velocity field projected onto the *t*-SNE plot. Arrows show the local average velocity evaluated on a regular grid. Top right inset, differentiation endpoints as high density regions on the manifold after forward Markov process with velocity-based transition probabilities; the root of the branching tree is identified simulating the process in the reverse direction. Bottom right inset, summary schematic of the RNA velocity field, and expression of the transcription factor *Prox1*. **d**, Commitment to oligodendrocyte fate. Left, visualization of single-step transition probabilities from two starting cells (red) to neighbouring cells. Right, velocities of a sampled subset of cells shown on the *t*-SNE plot in **c**. **e**, Fate decision of neuroblasts. Left, visualization of single-step transition probabilities from two starting cells (red) to neighbouring cells. Right, velocities of a sampled subset of cells shown on the *t*-SNE plot in **c**.

Resource

Cell

Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming

Graphical Abstract



Authors

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, ..., Rudolf Jaenisch, Aviv Regev, Eric S. Lander

Correspondence

jianshu@broadinstitute.org (J.S.), aregev@broadinstitute.org (A.R.), lander@broadinstitute.org (E.S.L.)

In Brief

Application of a new analytical approach to examine developmental trajectories of single cells offers insight into how paracrine interactions shape reprogramming.

Highlights

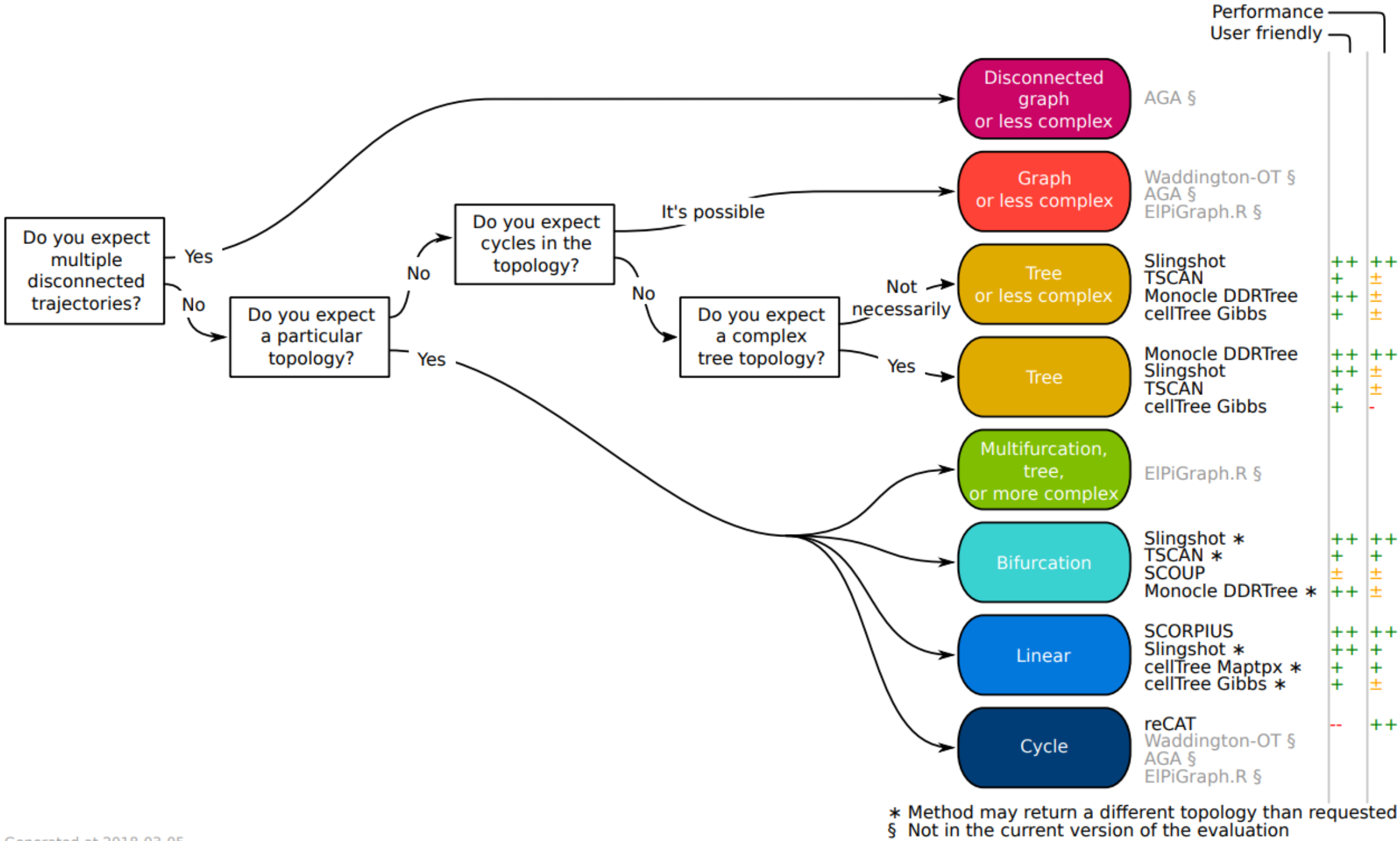
- Optimal transport
- 2019, cell

What is there

A lot of algorithms already developed:

- Diffusion maps / pseudotime
- Slingshot
- Optimal transport
- RNA velocity

- https://www.biorxiv.org/content/biorxiv/early/2018/03/05/276907.full.pdf
- This is the best current review about the topic



Generated at 2018-03-05

Questions?