A statistical framework for differential pseudotime analysis with multiple single-cell RNA-seq samples

Hongkai Ji, Ph.D.

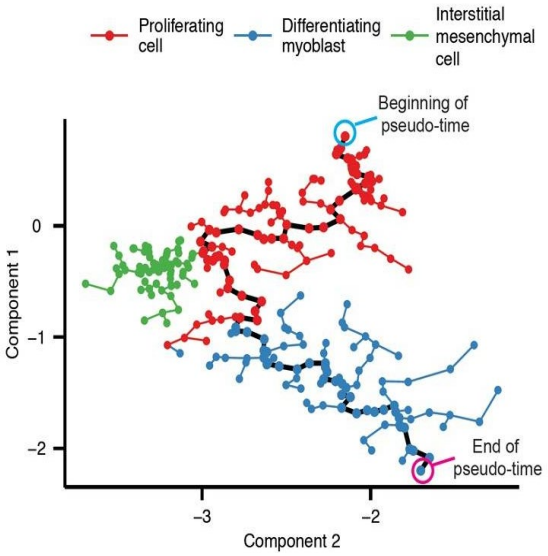
Email: hji@jhu.edu

Department of Biostatistics

Johns Hopkins Bloomberg School of Public Health

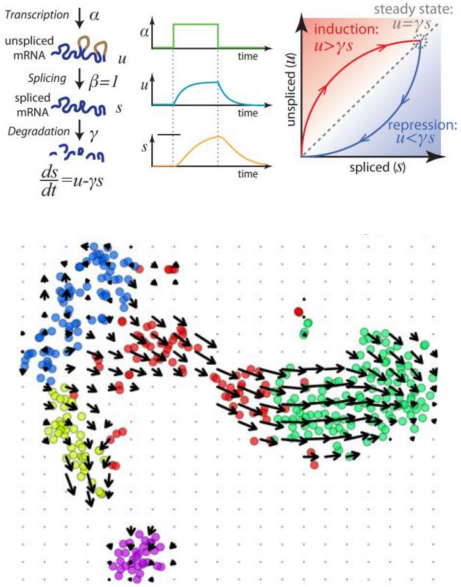
Reconstructing temporal cellular processes using single-cell data

Pseudotime/Trajectory



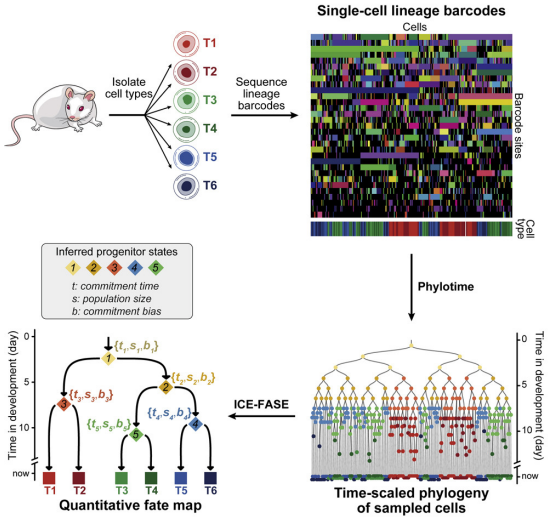
Trapnell et al., Nat Biotechnol. 2014, 32:381-6

RNA Velocity



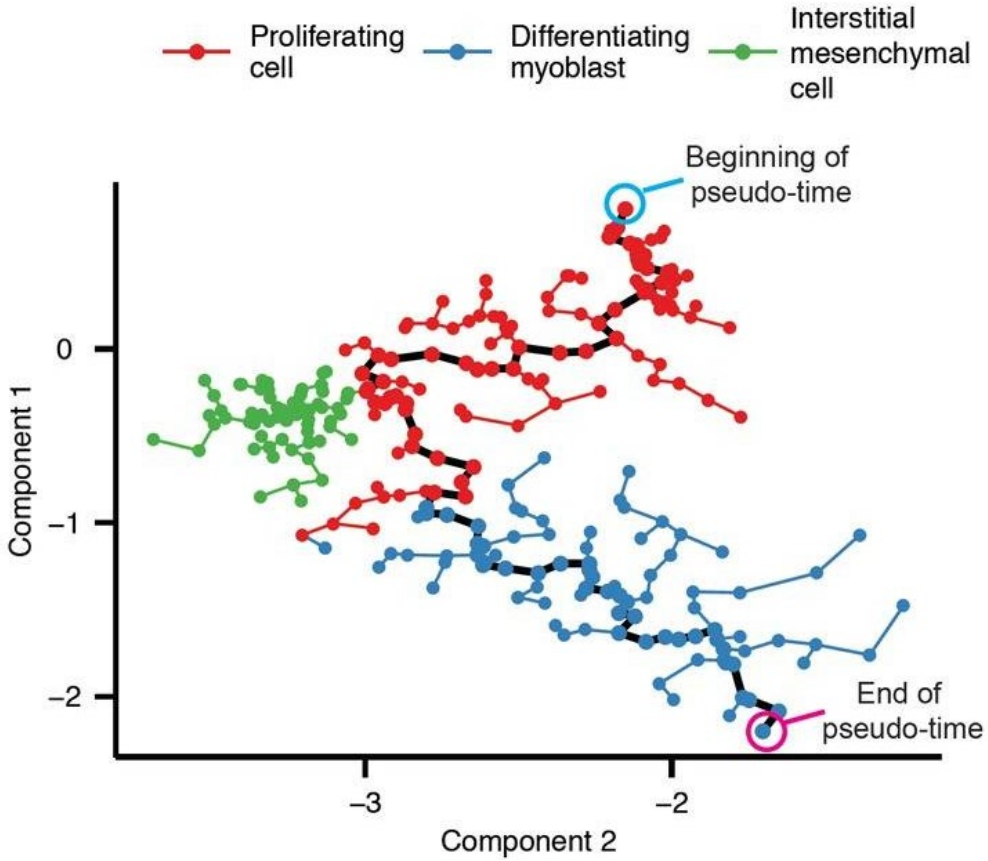
La Manno et al. Nature. 2018 560(7719):494-498

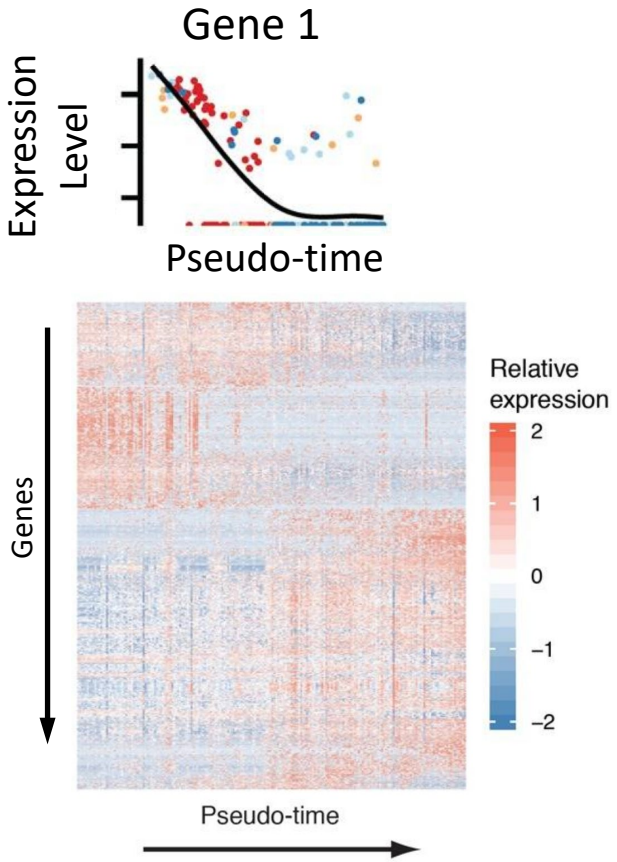
Lineage Barcode



Fang et al. Cell. 2022 185(24):4604-4620.e32

Pseudotime (trajectory) analysis of single-cell genomic data





Trapnell et al., Nat Biotechnol. 2014, 32:381-6

A long list of trajectory analysis methods

nature biotechnology

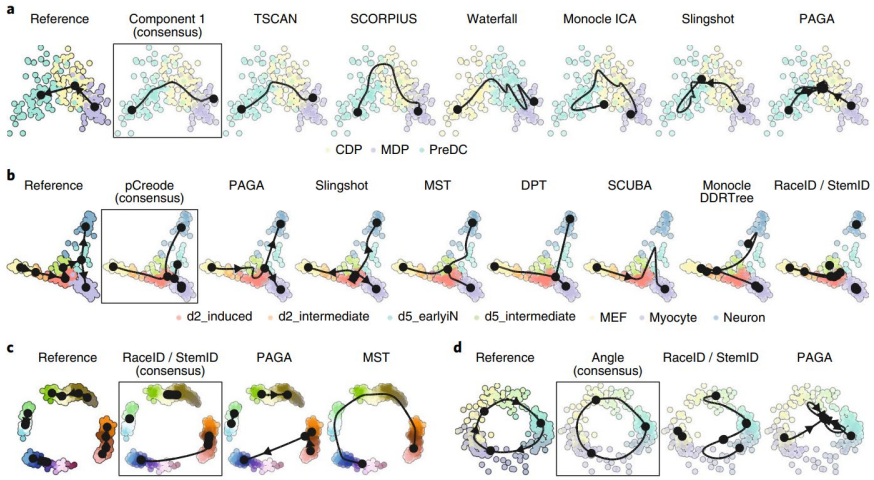
ARTICLES

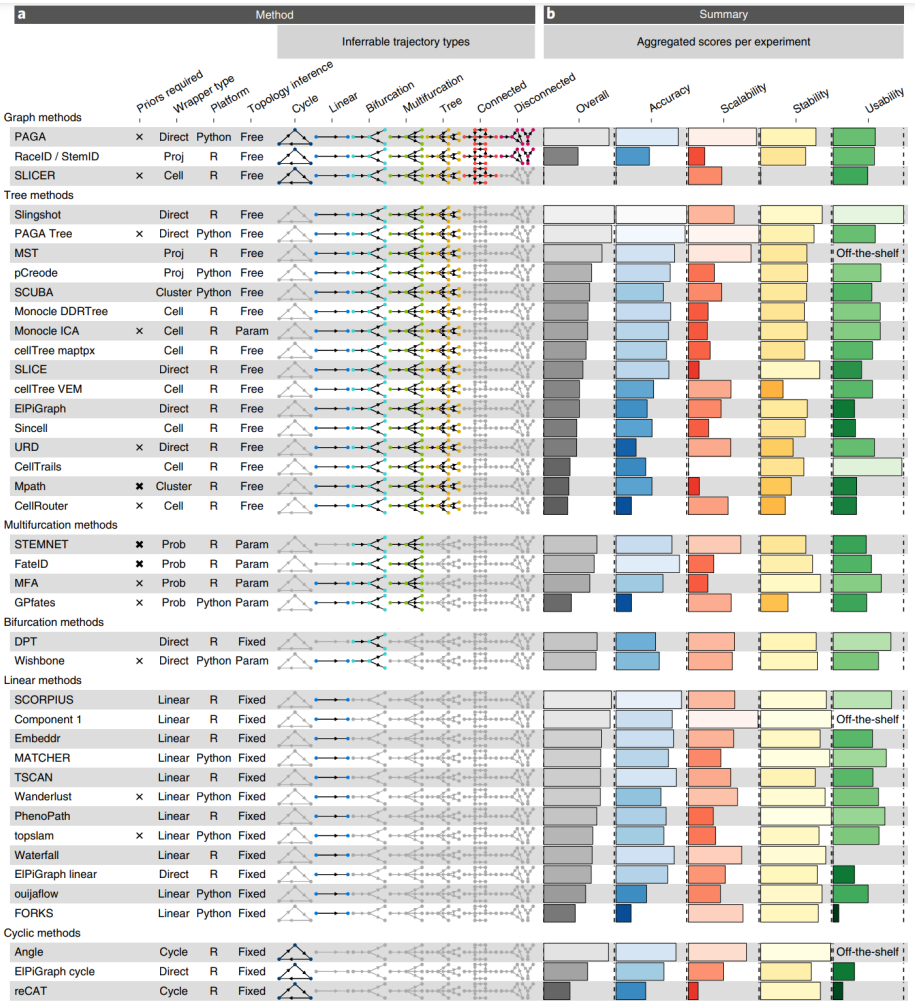
https://doi.org/10.1038/s41587-019-0071-9

A comparison of single-cell trajectory inference methods

Wouter Saelens[1,2,6], Robrecht Cannoodt[1,3,4,6], Helena Todorov[1,2,5] and Yvan Saeys[1,2*]

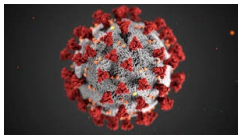
Trajectory inference approaches analyze genome-wide omics data from thousands of single cells and computationally inf the order of these cells along developmental trajectories. Although more than 70 trajectory inference tools have already be developed, it is challenging to compare their performance because the input they require and output models they produce va substantially. Here, we benchmark 45 of these methods on 110 real and 229 synthetic datasets for cellular ordering, topolog scalability and usability. Our results highlight the complementarity of existing tools, and that the choice of method should depe mostly on the dataset dimensions and trajectory topology. Based on these results, we develop a set of guidelines to help use select the best method for their dataset. Our freely available data and evaluation pipeline (https://benchmark.dynverse.or will aid in the development of improved tools designed to analyze increasingly large and complex single-cell datasets.

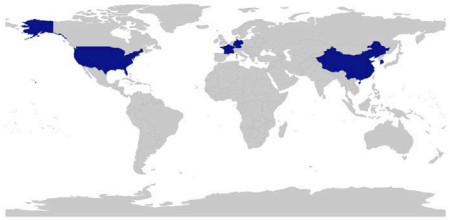


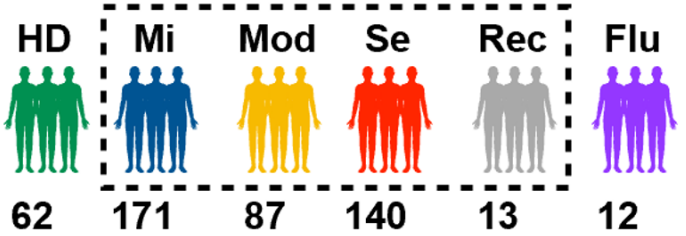


However, few of the existing methods tackle trajectory differential analysis across conditions with multiple samples per condition, while such studies become increasingly common.

Example 1: COVID-19 Single-cell RNA-seq

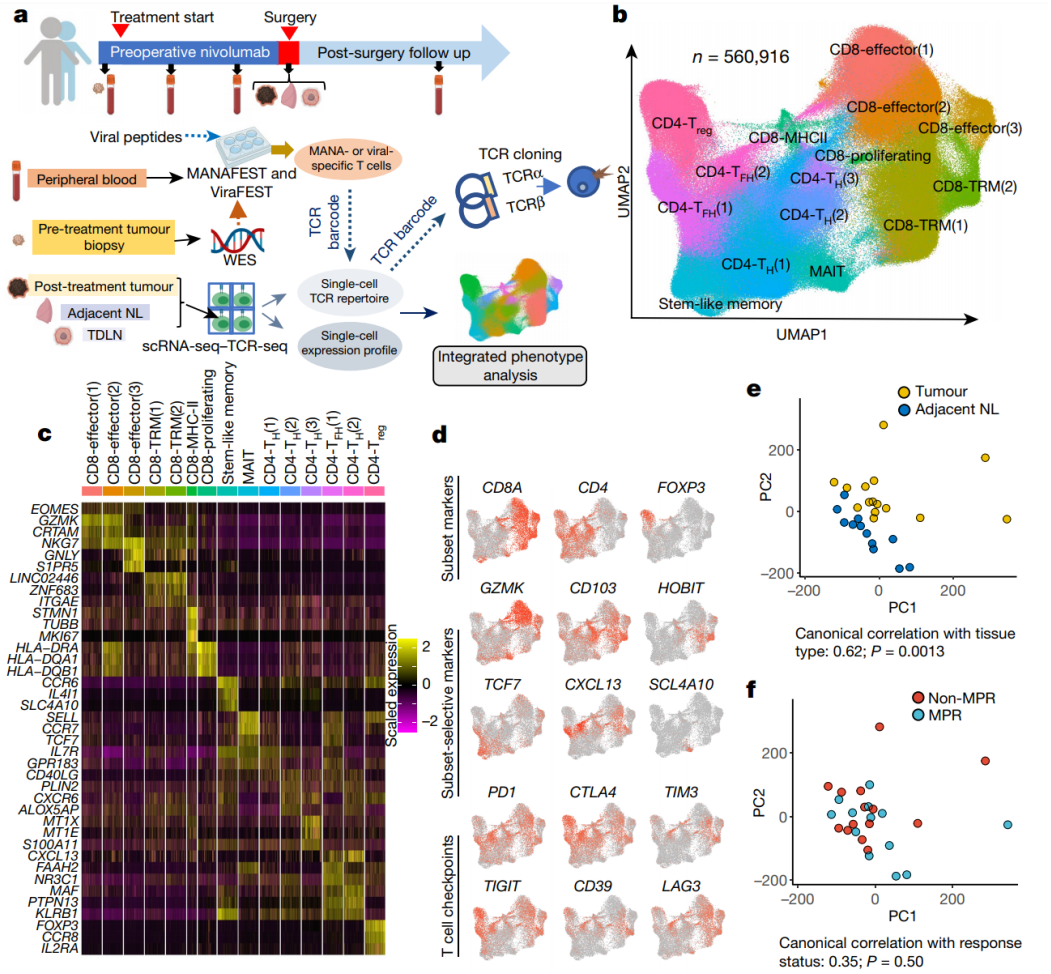






Study	Sample Number	Subject Number	Cohort Number	Sample Disease Status	Location
Wilk et al., Nat Med., 2020	14	13	1	Healthy Donor; COVID-19 Moderate, Severe	USA
Wen et al., Cell Discov., 2020	15	15	1	Healthy Donor; COVID-19 Recovered	China
Lee et al., Sci Immunol., 2020	20	17	1	Healthy Donor; COVID-19 Mild, Severe; Influenza	Korea
Guo et al., Nat Commun., 2020	5	2	1	Healthy Donor; COVID-19 Severe, Recovered	China
Yu et al., Cell Res., 2020	9	9	1	Healthy Donor; COVID-19 Mild	China
Arunachalam et al., Science, 2020	12	12	1	Healthy Donor; COVID-19 Moderate, Severe	USA
Schulte-Schrepping et al., Cell, 2020	101	52	2	Healthy Donor; COVID-19 Mild, Severe	Germany
Silvin et al., Cell, 2020	9	6	1	Healthy Donor; COVID-19 Mild, Severe	France
Su et al. Cell, 2020	270	145	1	Healthy Donor; COVID-19 Mild, Moderate, Severe	USA
Zhu et al. Immunity, 2020	23	10	1	Healthy Donor; COVID-19 Mild, Severe; Influenza	China
Mudd et al. Sci. Adv.,2020	7	7	1	Healthy Donor; COVID-19 Severe; Influenza	USA
Total	**485**	**288**	**12**		

Example 2: Tumor infiltrating lymphocytes in immunotherapy treated lung cancer patients



Caushi et al. Nature, 596:126-132 (2021)

Limitations of existing methods

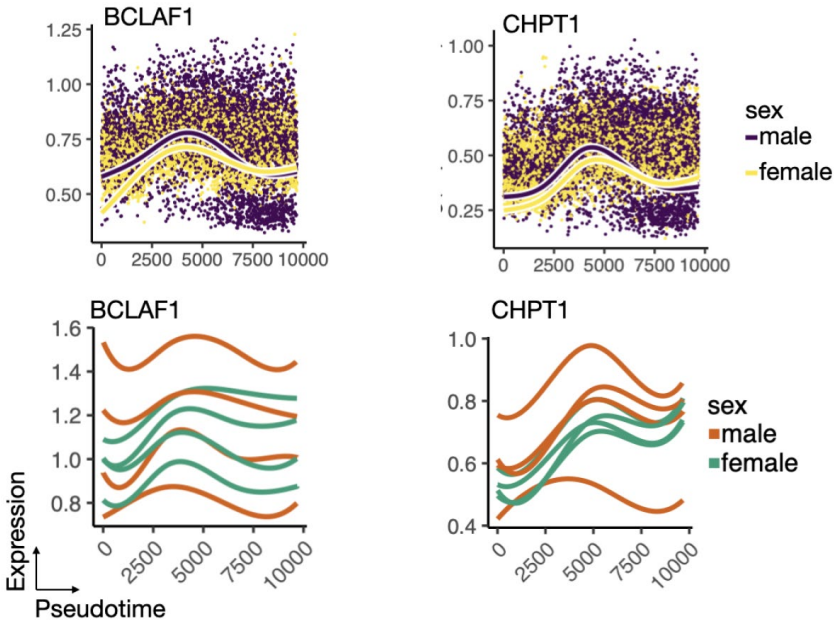
- Monocle, TSCAN, Slingshot, tradeSeq, etc.: Do not analyze DE across conditions.

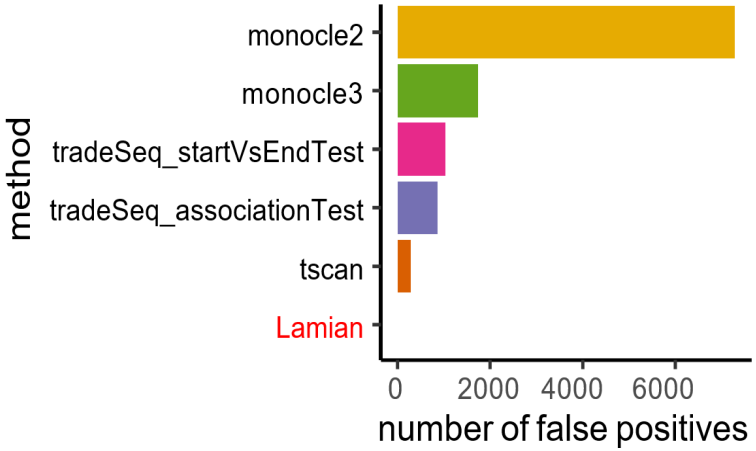
- Phenopath (Campbell & Yau Nat. communications 9:2442, 2018): Linear expression change along pseudotime, cannot handle arbitrary DE as non-linear functions of pseudotime, no separation of cell and sample variance

- Condiments (Hector Roux de Bézieux et al. bioRxiv 2021.03.09.433671): One sample per condition, not optimal for multiple-sample analyses

Limitations of existing methods

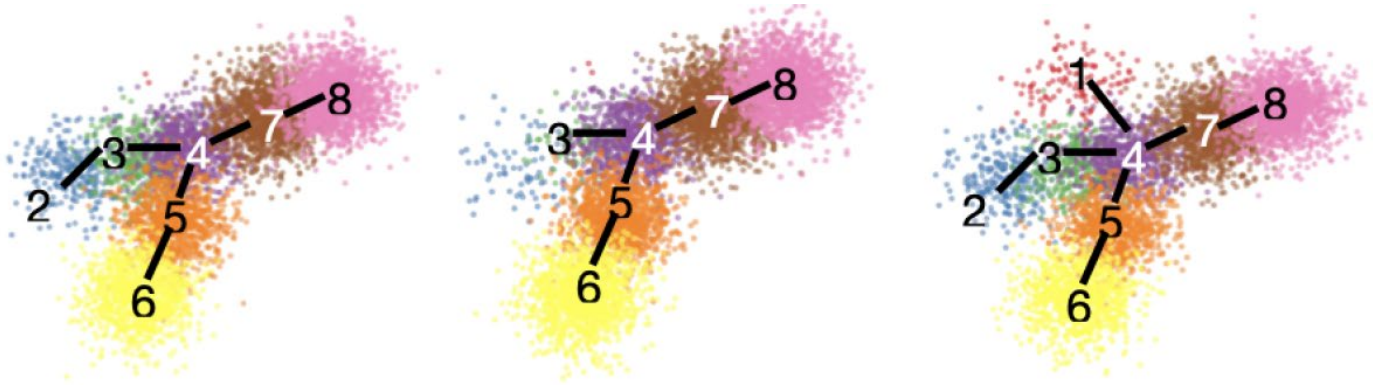
- Ignoring sample-level variability will create false positives (sometimes a lot) in a null dataset without differential signals.





Limitations of existing methods

- Few methods account for uncertainty and variability of the trajectory topology
  - PseudotimeDE (Song & Li, Genome Biology, 2021 22:124) – do not consider multiple samples



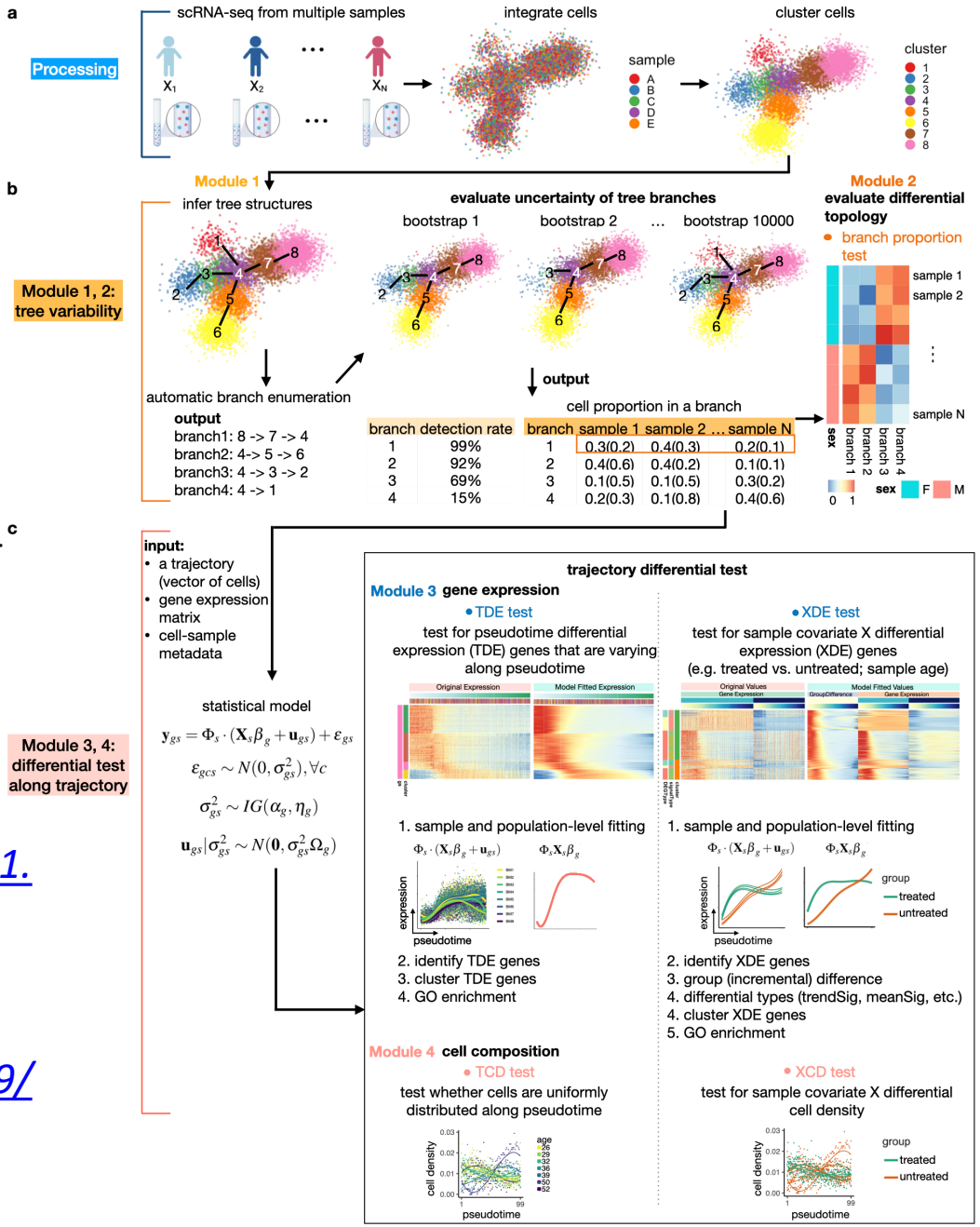
- Changes may occur in gene expression or cell abundance, but not all methods consider both

Lamian

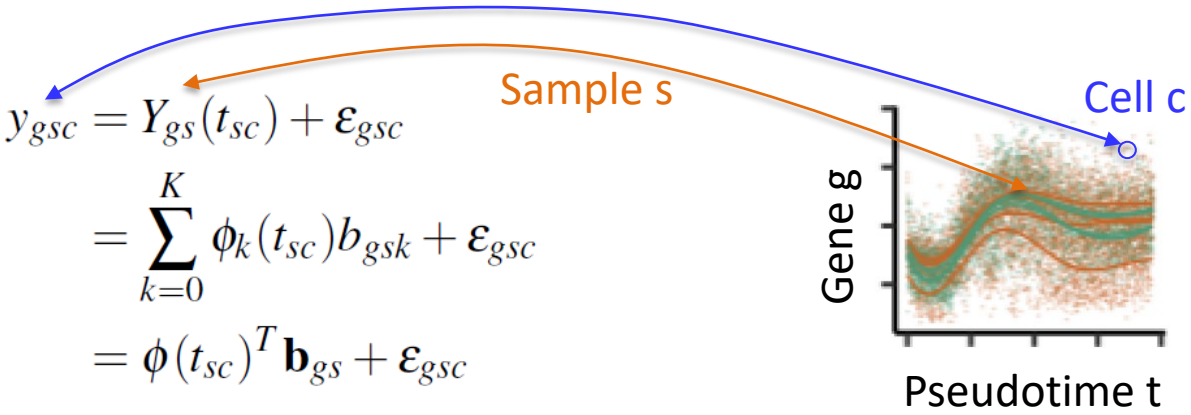
A statistical framework for differential pseudotime analysis with multiple single-cell RNA-seq samples (under revision)

Hou et al. bioRxiv 2021.07.10.451910; doi: https://doi.org/10.1101/2021.07.10.451910

Software: https://github.com/Winnie09/Lamian



Lamian model



$$y_{gsc} = Y_{gs}(t_{sc}) + \varepsilon_{gsc}$$
$$= \sum_{k=0}^{K} \phi_k(t_{sc}) b_{gsk} + \varepsilon_{gsc}$$
$$= \phi(t_{sc})^T \mathbf{b}_{gs} + \varepsilon_{gsc}$$

$$\phi(t) = [\phi_0(t), \phi_1(t), \ldots, \phi_K(t)]^T$$

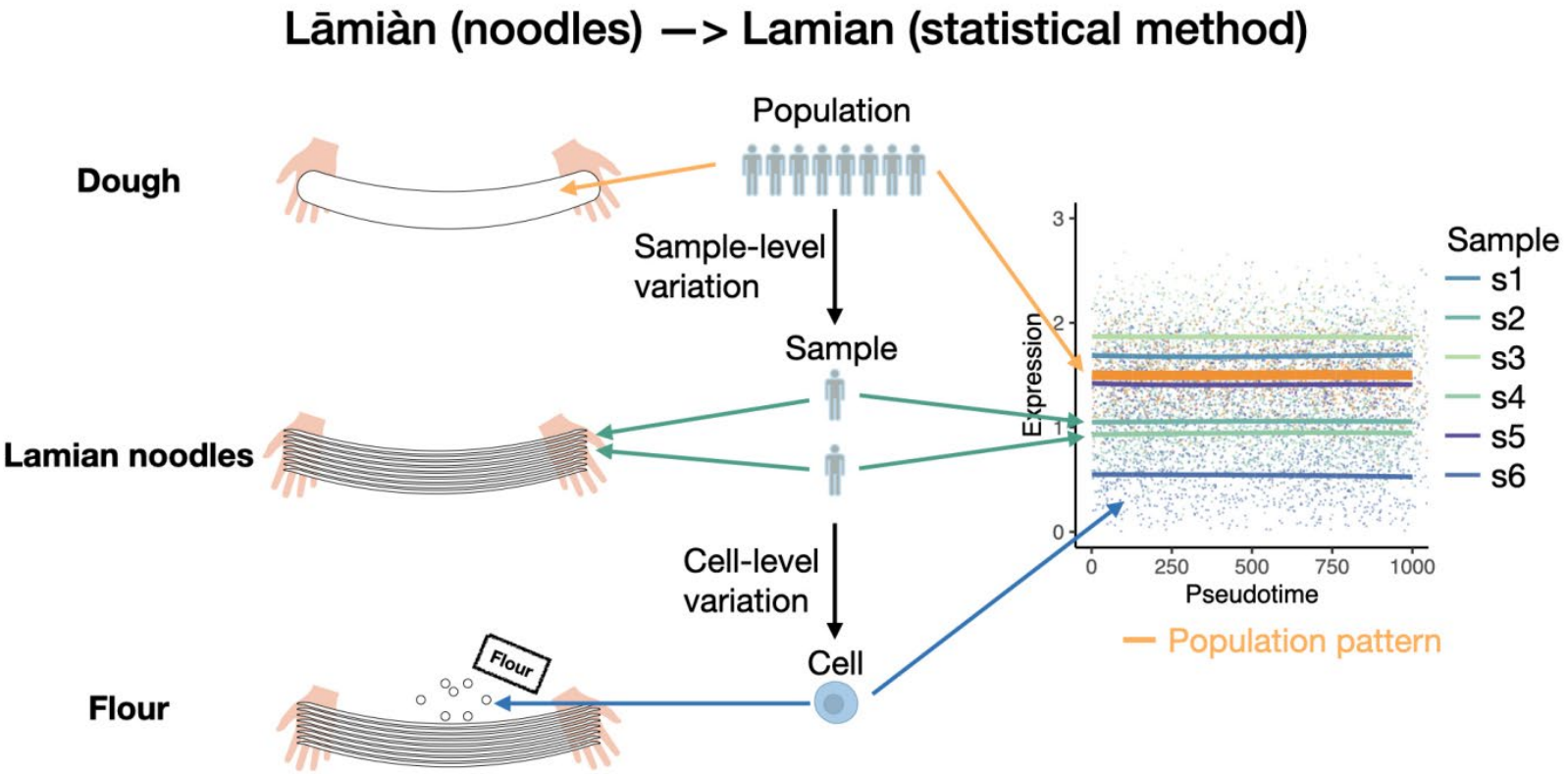
$$\varepsilon_{gsc} \sim N(0, \sigma_{gs}^2)$$

$$\mathbf{b}_{gs} = \begin{bmatrix} b_{gs0} \\ b_{gs1} \\ \vdots \\ b_{gsK} \end{bmatrix} = \begin{bmatrix} \beta_{g00} & \beta_{g01} & \cdots & \beta_{g0V} \\ \beta_{g10} & \beta_{g11} & \cdots & \beta_{g1V} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{gK0} & \beta_{gK1} & \cdots & \beta_{gKV} \end{bmatrix} \begin{bmatrix} 1 \\ x_{s1} \\ \vdots \\ x_{sV} \end{bmatrix} + \begin{bmatrix} u_{gs0} \\ u_{gs1} \\ \vdots \\ u_{gsK} \end{bmatrix} = \mathbf{B}_g \mathbf{x}_s + \mathbf{u}_{gs}$$

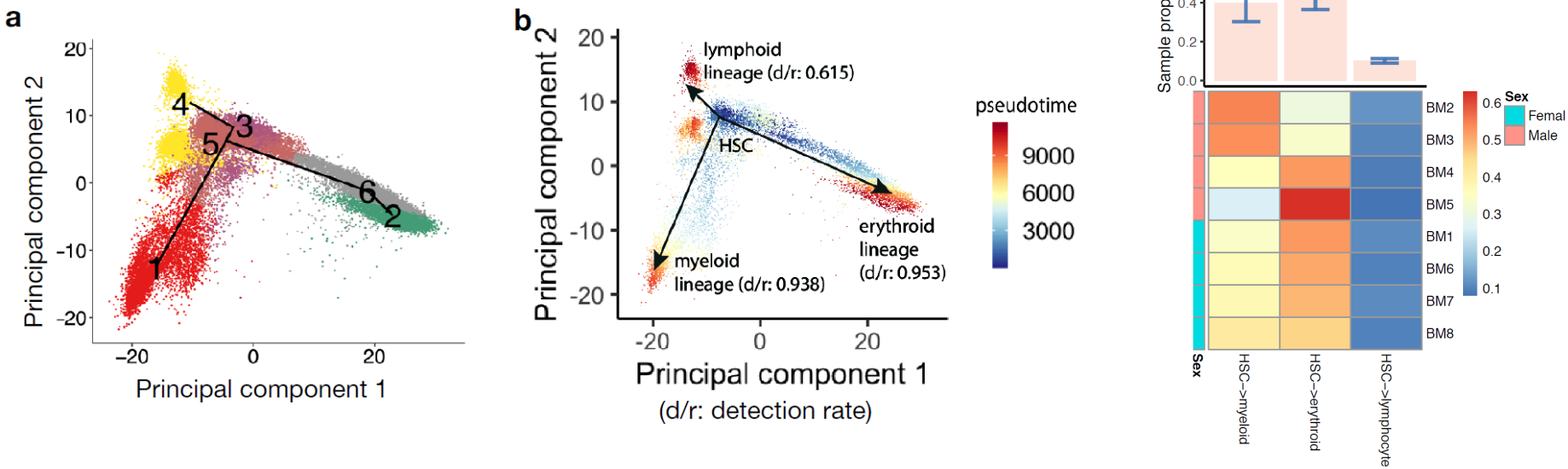
$$\mathbf{u}_{gs} \sim N(\mathbf{0}, \sigma_{gs}^2 \Omega_g)$$

$$\sigma_{gs}^2 \sim IG(\alpha_g, \eta_g)$$

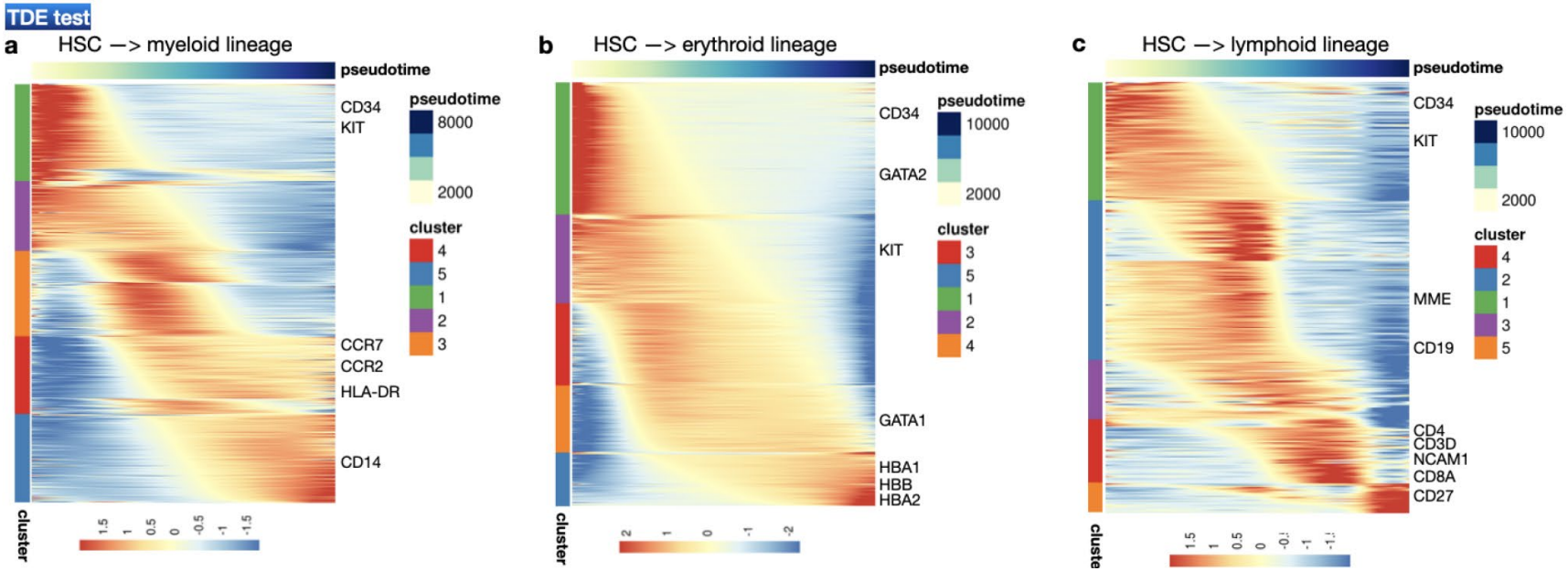
Lamian



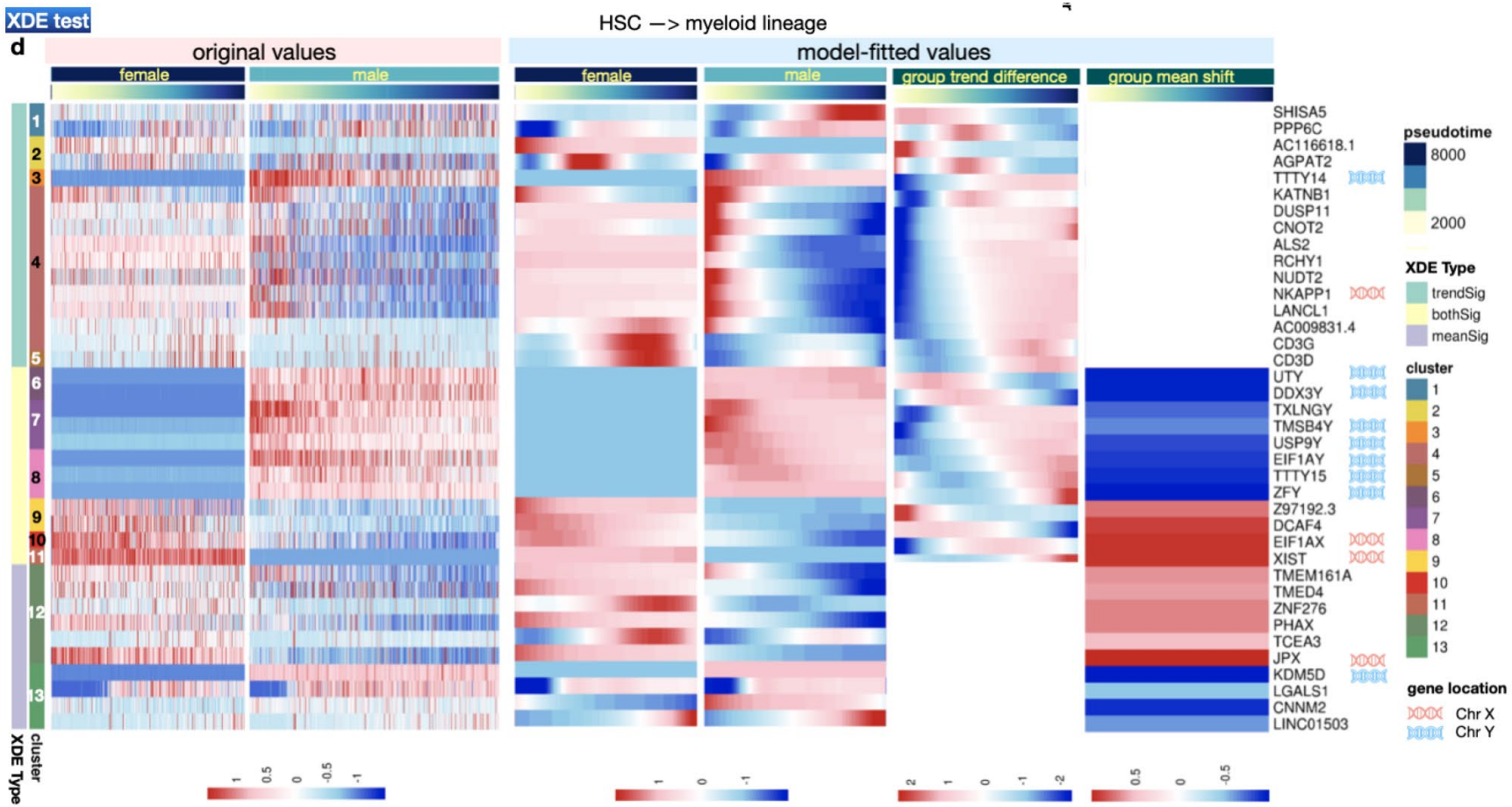
Lamian supports assessment of uncertainty and changes of topology of pseudotemporal trajectories



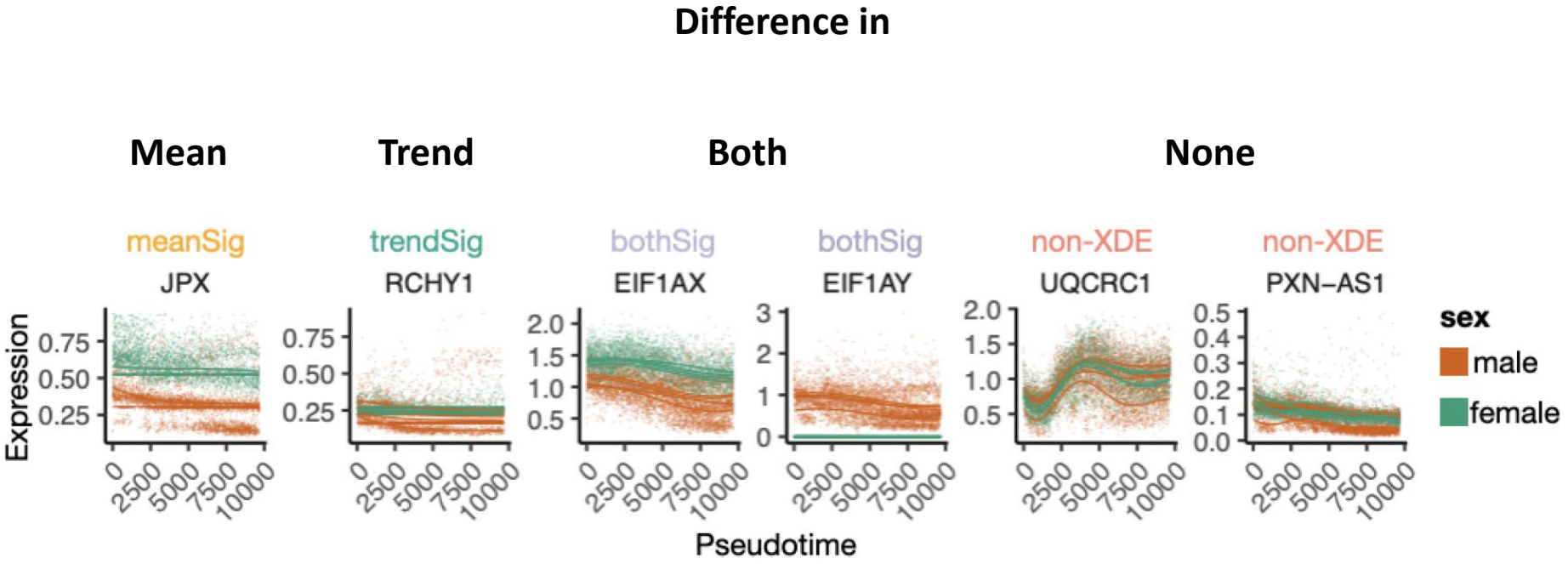
Lamian supports differential gene expression analysis along pseudotime (TDE)



Lamian supports differential gene expression analysis across conditions (XDE)



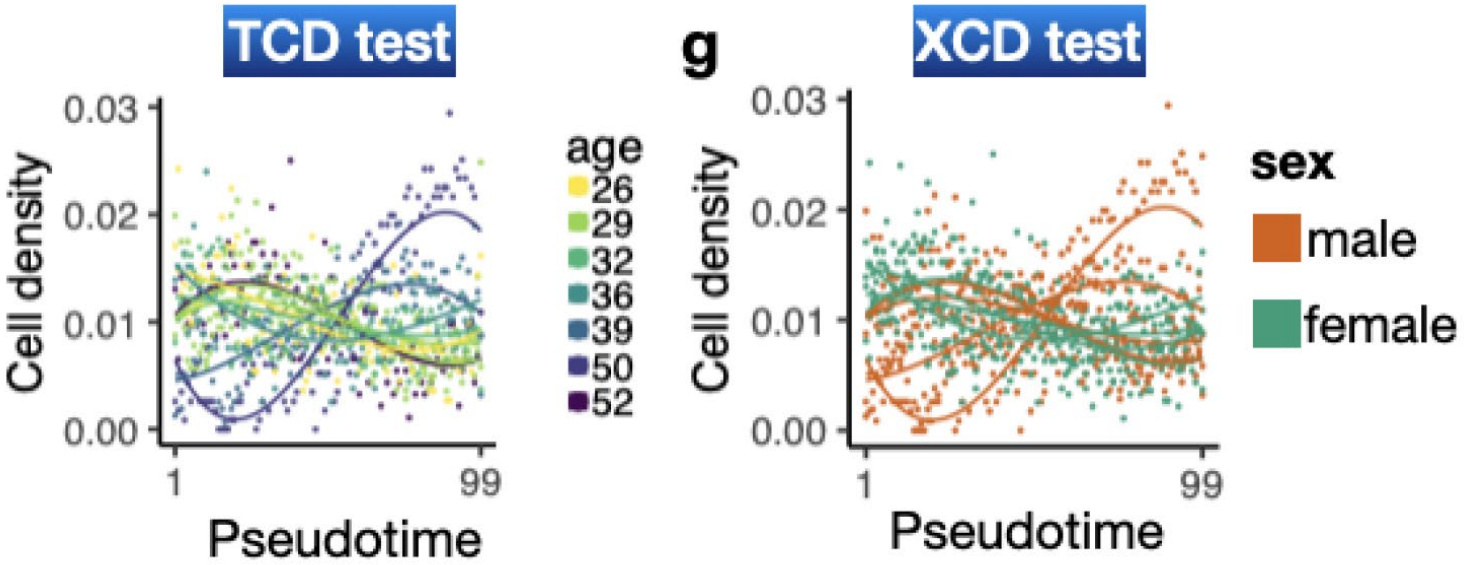
Lamian classifies XDE genes



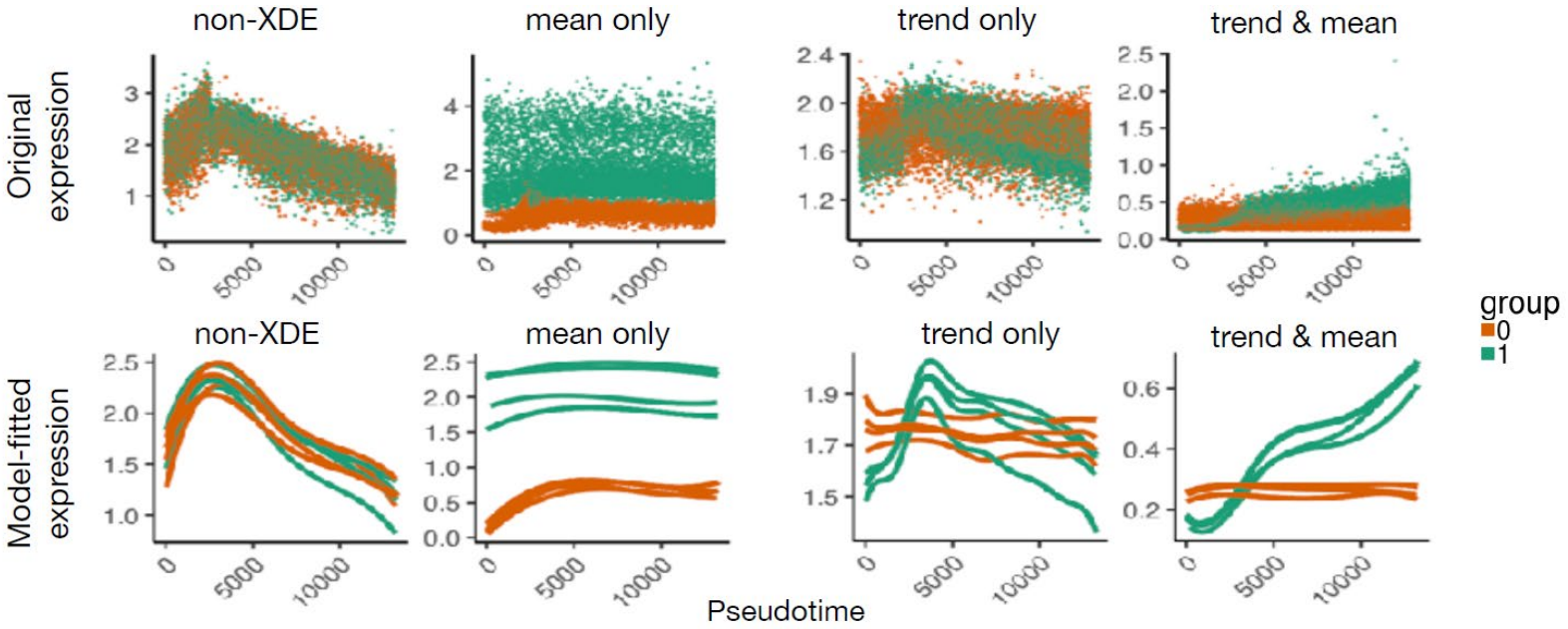
- M_0: $\beta_{g.v} = [\beta_{g0v}, \beta_{g1v}, \ldots, \beta_{gKv}]^T = \mathbf{0}$.
- M_1: $\beta_{g.v} \neq \mathbf{0}$ and $\beta_{g0v} = \beta_{g1v} = \ldots = \beta_{gKv} = c$.
- M_2: $\beta_{g.v} \neq \mathbf{0}$.

- *XDE test*: the null model M_0 is compared with the alternative model M_2. Rejecting M_0 implies XDE.
- *Mean test*: M_0 and M_1 are compared. Rejecting M_0 implies mean shift.
- *Trend test*: M_1 and M_2 are compared. Rejecting M_1 implies trend difference.

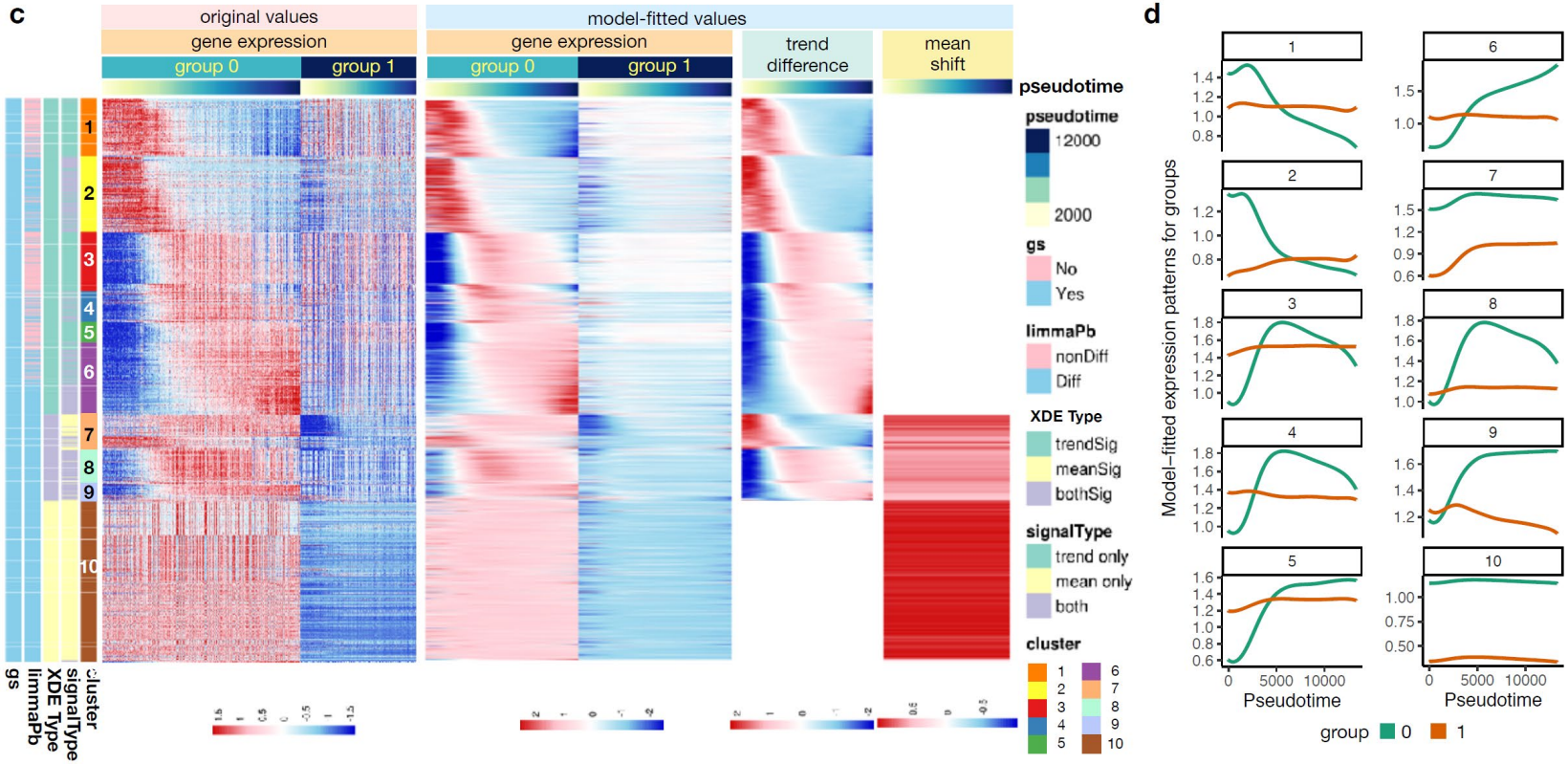
Lamian supports differential cell abundance analysis



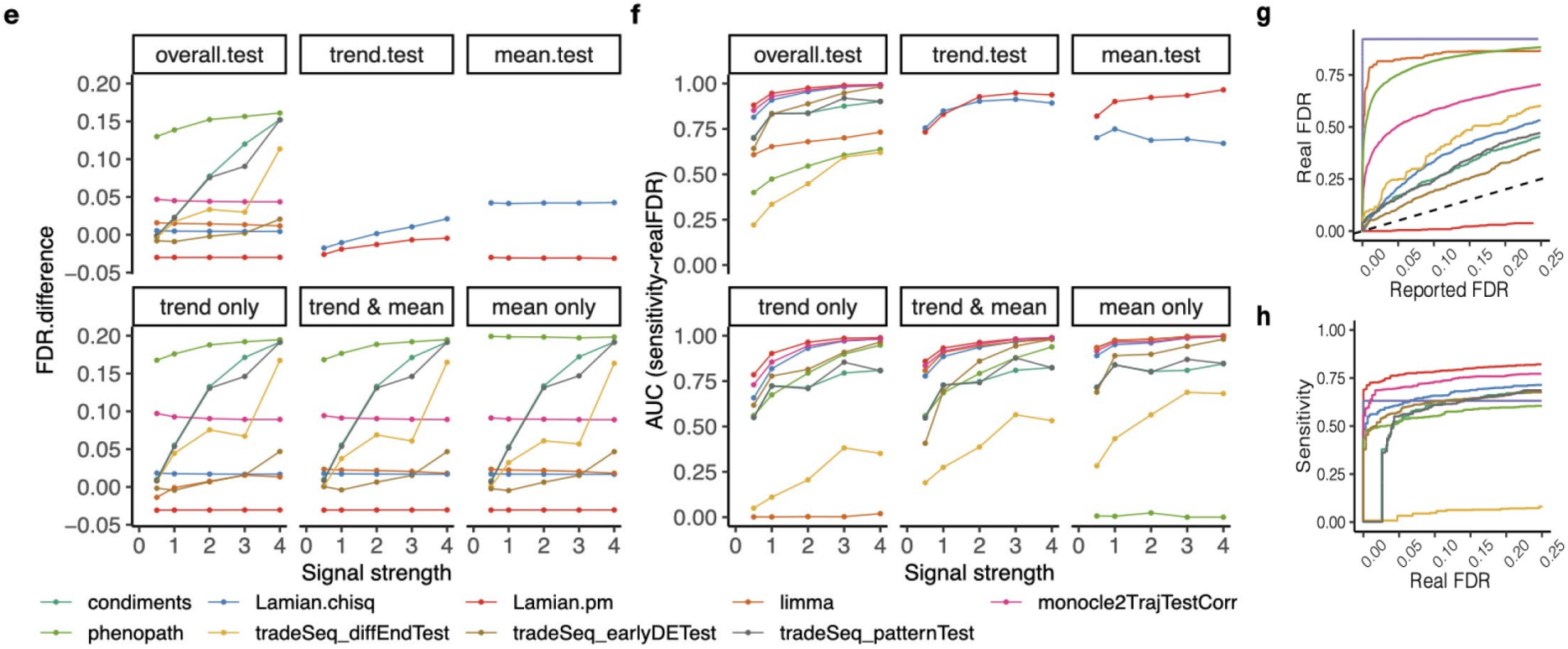
Benchmark simulation



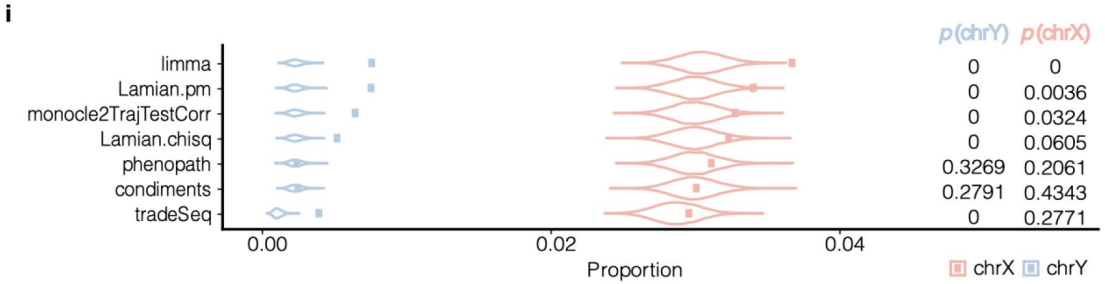
Benchmark simulation

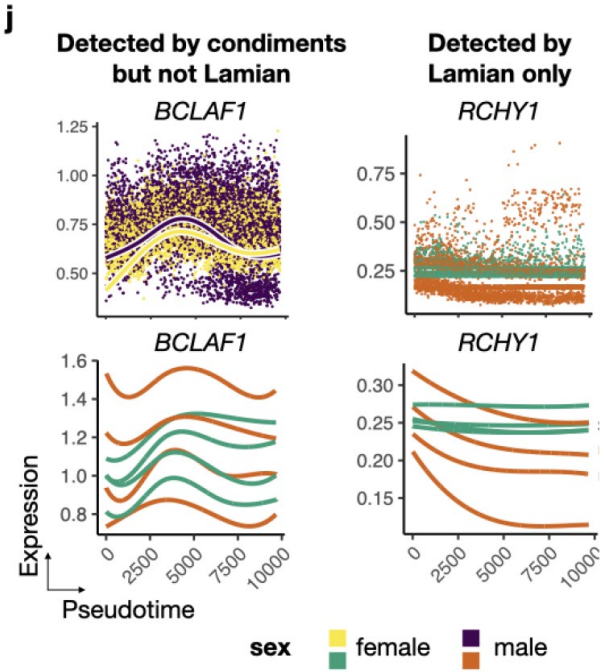


Method comparisons - XDE

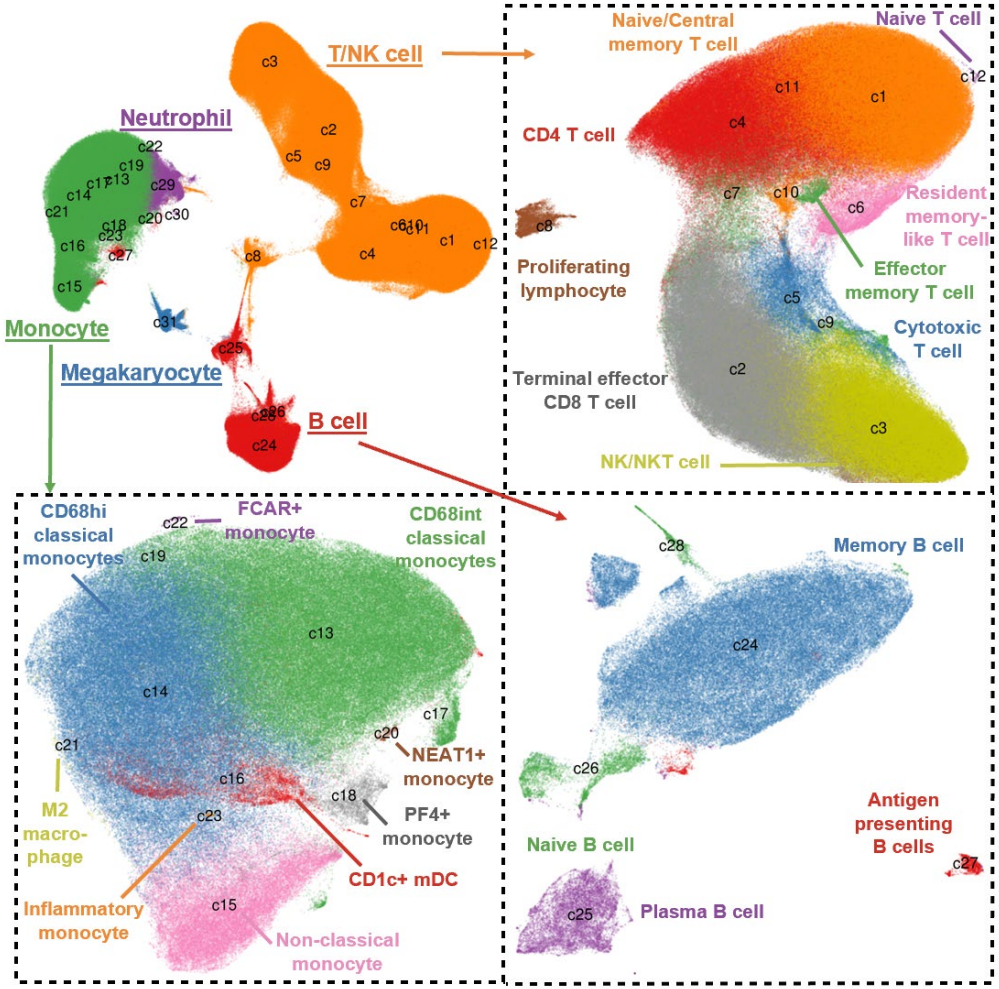


Method comparisons – sex difference in bone marrow samples



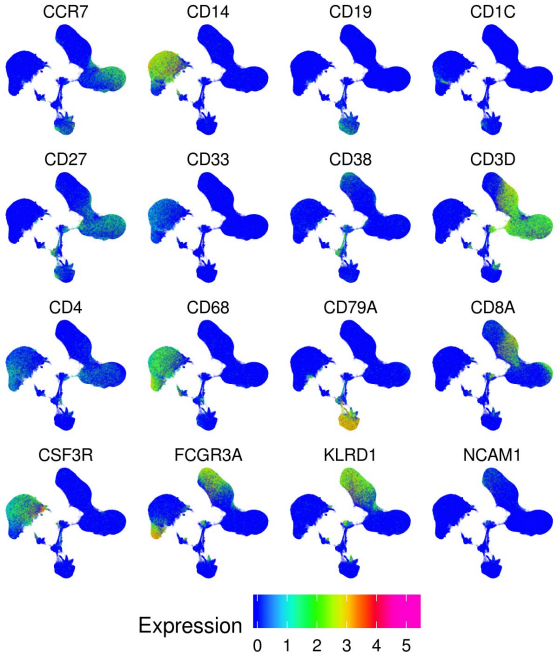


Example 1: COVID-19 scRNA-seq analysis



31 cell clusters from **5 meta-cell categories:**

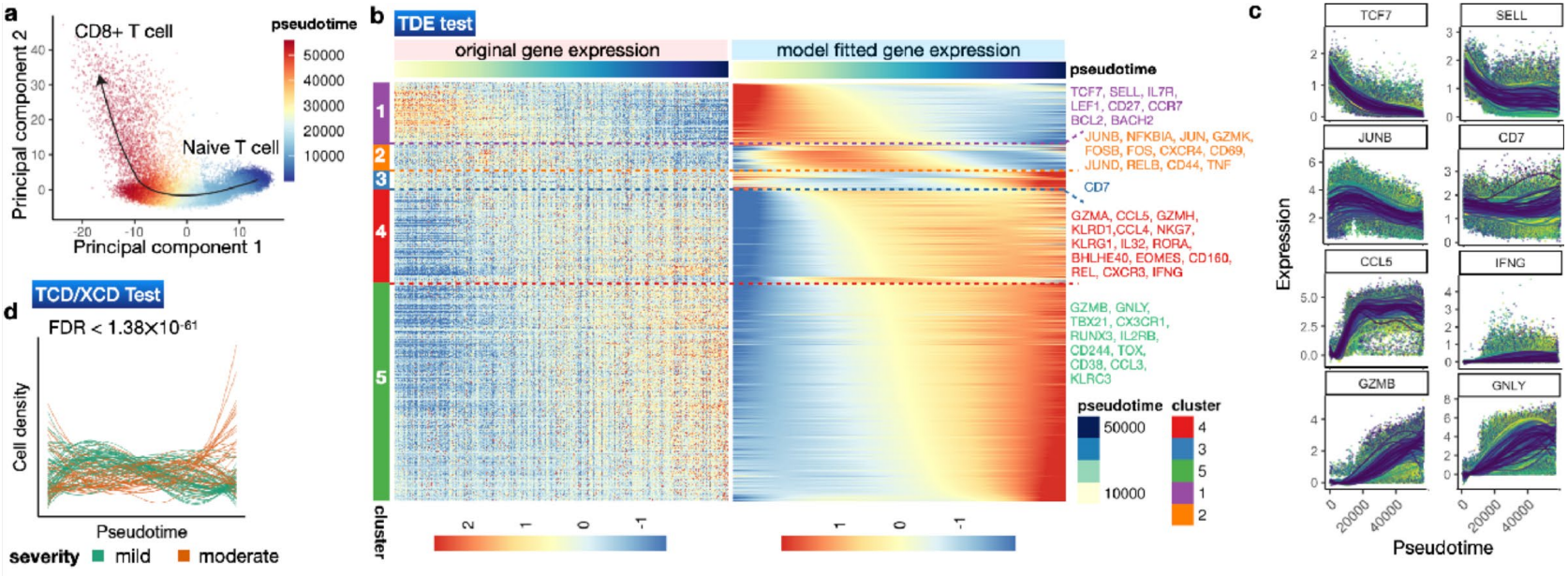
- T and natural killer (NK) cells (c1-c12)
- Monocytes (c13-c23)
- B cells (c24-c28)
- Neutrophils (c29-c30)
- Megakaryocytes (c31)



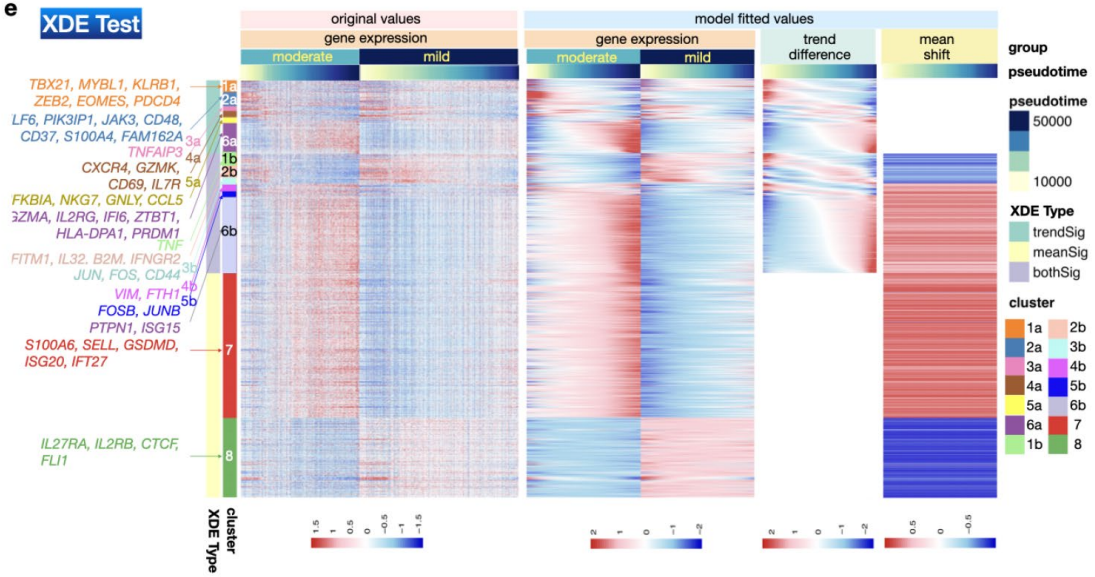
Lamian analysis of CD8+ T cell activation in COVID-19 patients

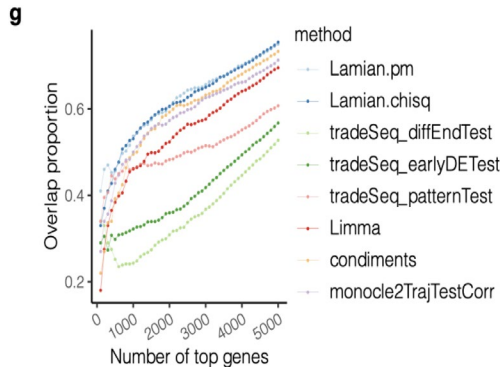
- How does disease severity change the cellular programs?

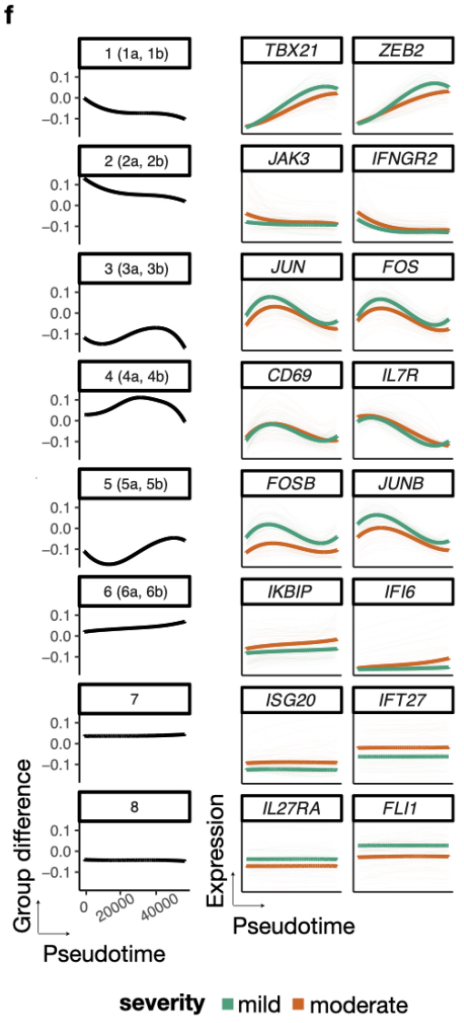
66 mild vs. 48 moderate COVID samples, 55,953 cells



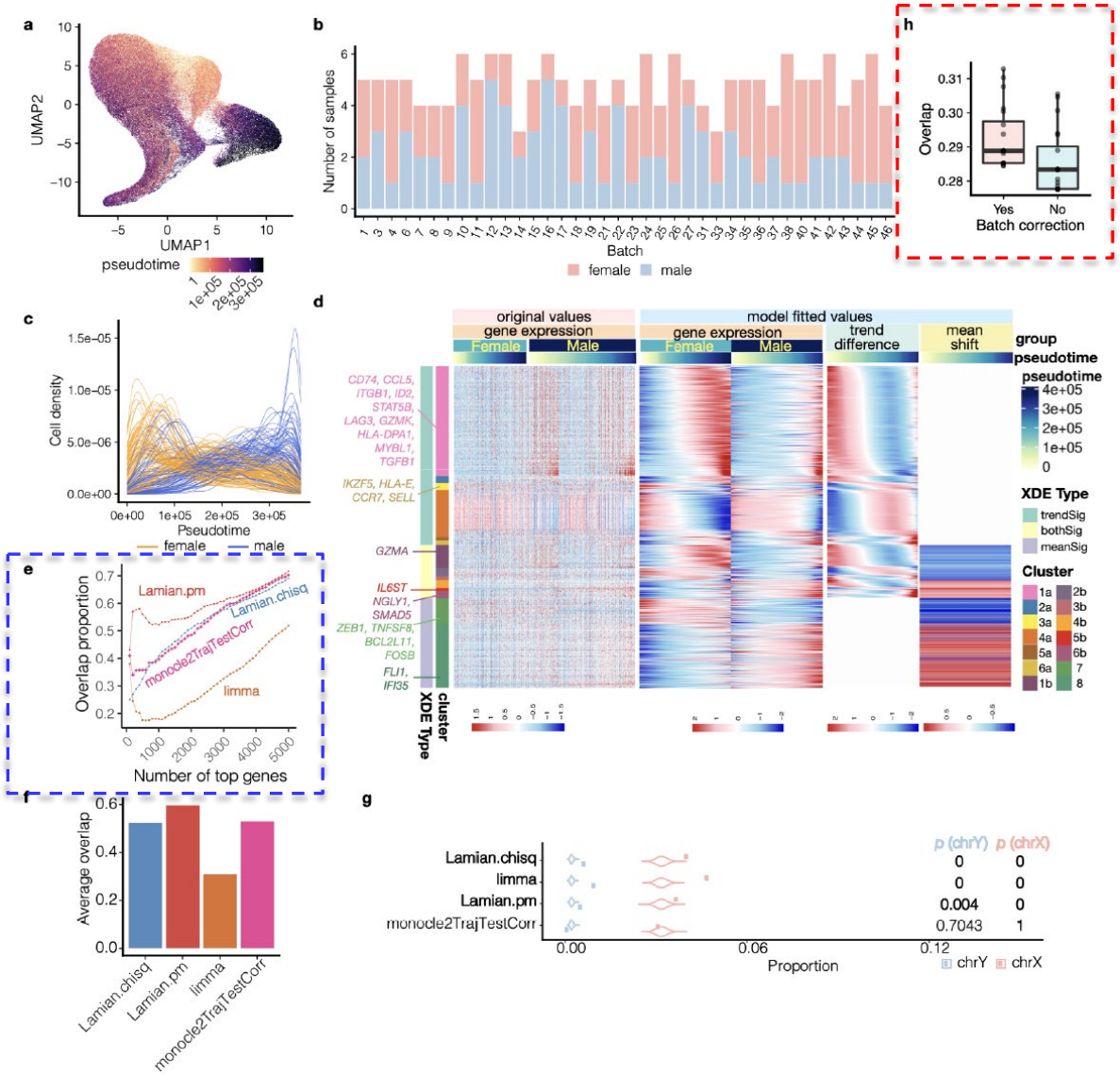
Lamian analysis of CD8+ T cell activation in COVID-19 patients







Example 2: Sex difference in tuberculosis(TB)



337,191 memory T cells from **184** donors (100 females and 84 males) in a tuberculosis (TB) cohort

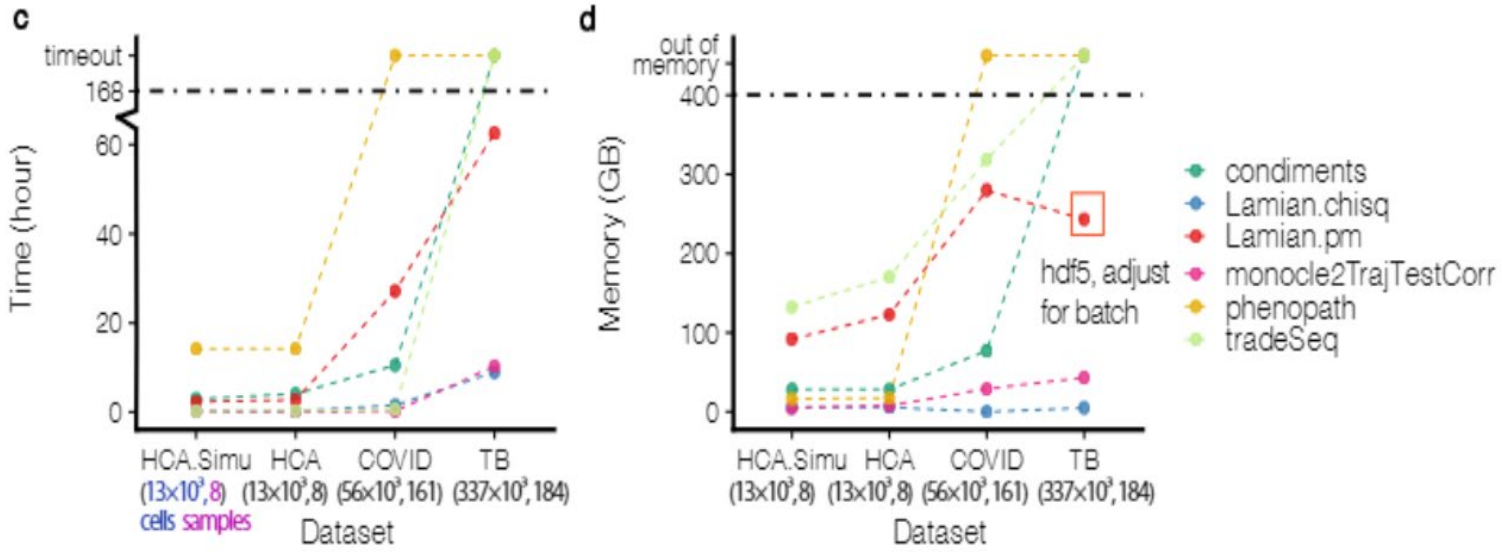
Computational efficiency

Computational Time (Hour)

			condiments	Lamian.chisq	Lamian.pm	monocle2TrajTestCorr	phenopath	tradeSeq
	NumberOfSamples	NumberOfCells						
HCA.Simu	8	13k	2.961	0.228	2.344	0.1496	14.1616	0.2448
HCA	8	13k	4.0775	0.233	2.70877778	0.072333333	14.17666667	0.30933333
COVID	161	56k	10.497	1.529	27.12725	0.187	NA	0.578
TB	184	337k	NA	8.926	62.605	10.275	NA	NA

Memory (GB)

			condiments	Lamian.chisq	Lamian.pm	monocle2TrajTestCorr	phenopath	tradeSeq
	NumberOfSamples	NumberOfCells						
HCA.Simu	8	13k	28.2649872	5.406388	91.9243463	4.8552312	16.2346992	132.309859
HCA	8	13k	28.103318	5.692792	122.82056	7.924004	17.42598667	170.492868
COVID	161	56k	77.20384	4.002128	279.96936	28.895608	NA	318.775632
TB	184	337k	NA	4.89568	243.04102	43.28332	NA	NA



Summary

- Lamian provides a solution to differential trajectory analysis with multi-sample single-cell RNA-seq data

  Open source software: https://github.com/Winnie09/Lamian

- It provides a comprehensive pipeline for assessing topology uncertainty, differential topology, differential gene expression and cell abundance along pseudotime and across covariates

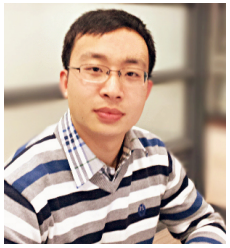
- By accounting for sample-level variability, Lamian properly controls false discovery rate and offers higher sensitivity

- Future extensions
  - Multi-sample trajectory analysis for other single-cell data types such as single-cell ATAC-seq
  - Reconstruct gene regulatory programs through multi-omic trajectory analysis

Acknowledgments

Johns Hopkins Bloomberg School of Public Health

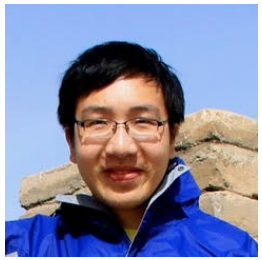
Wenpin Hou
(current: Columbia Univ)



Zhicheng Ji
(current: Duke Univ)

Stephanie Hicks

University of Pennsylvania



Zeyu Chen

E. John Wherry

Funding

NIH R01HG009518, R01HG010889

Questions?

Thank You!