Inferring cell lineage trees and fate maps from lineage tracing data

Palash Sashittal

Department of Computer Science

Virginia Tech

Organismal Development

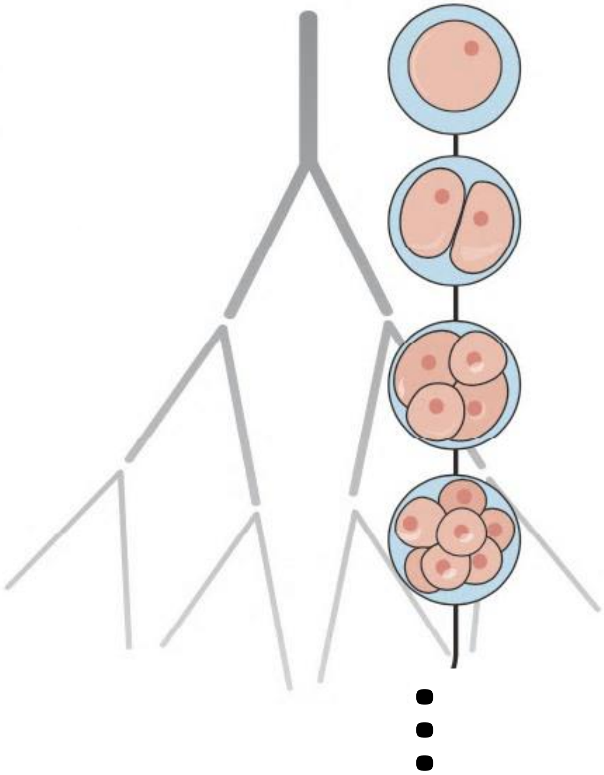
One cell
One cell type

Multiple cells
Multiple cell types

Embryogenesis

=

One cell

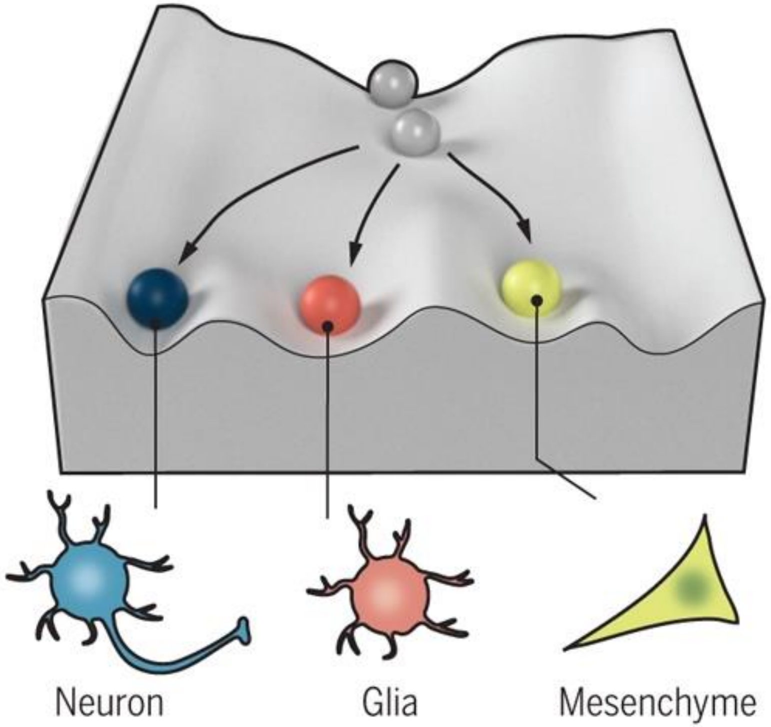


Multiple cells

Cell division

+

One cell type



Multiple cell types

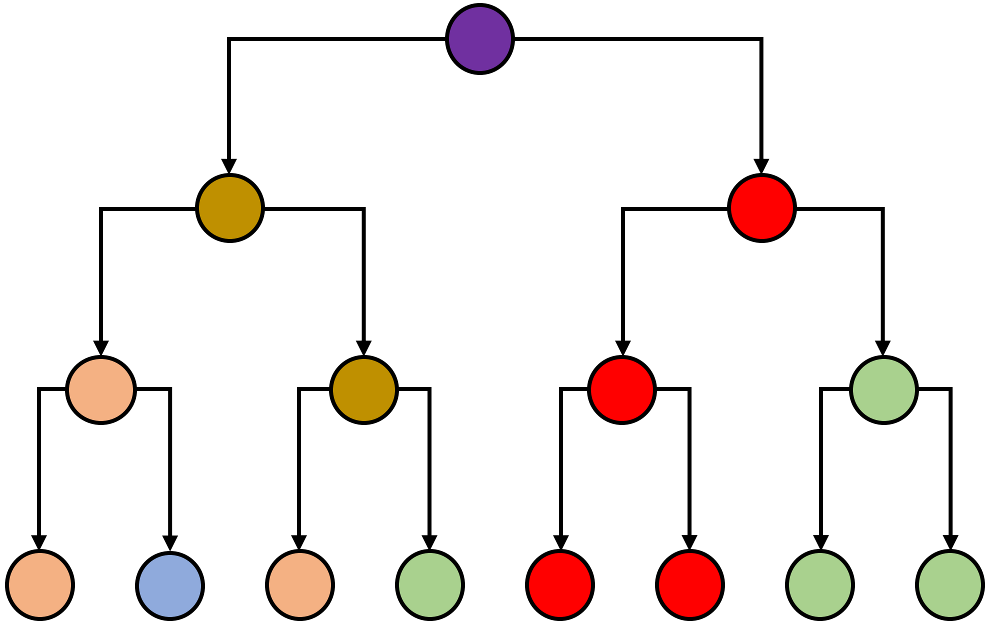
Cell differentiation

Organismal Development

One cell
One cell type

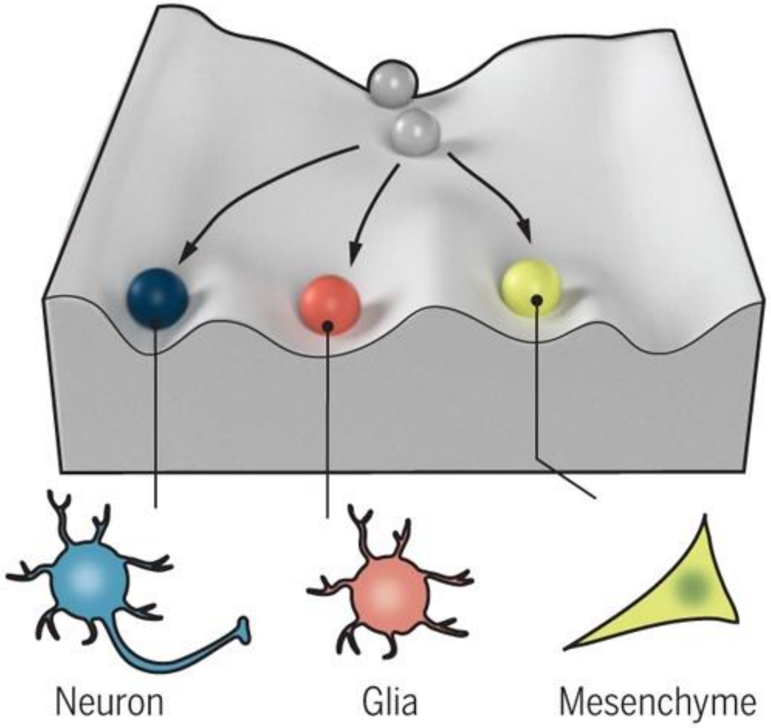
Multiple cells
Multiple cell types

Embryogenesis



Cell lineage tree T
Rooted tree with leaves
representing cells in the organism

One cell type



Multiple cell types

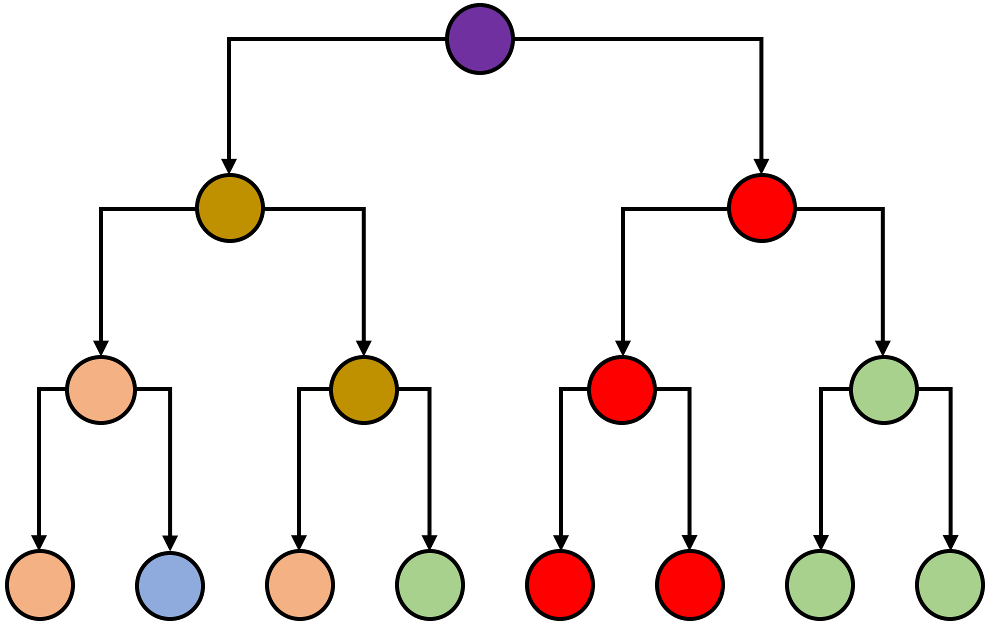
Cell differentiation

Organismal Development

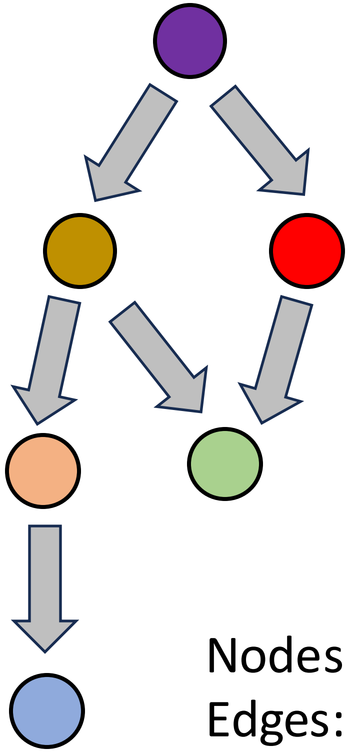
One cell
One cell type

Multiple cells
Multiple cell types

Embryogenesis



Cell lineage tree T
Rooted tree with leaves
representing cells in the organism

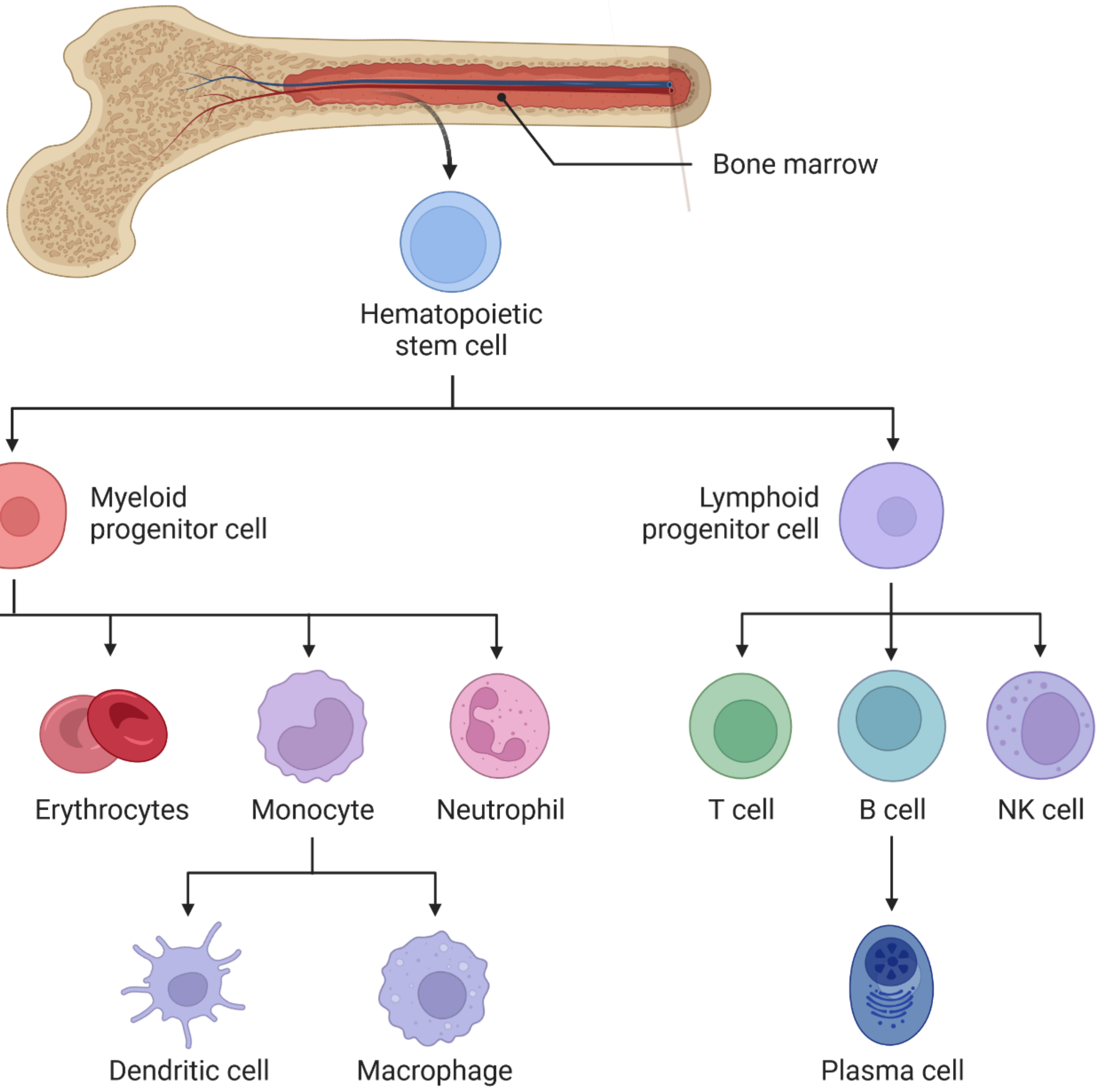


Nodes: cell types
Edges: transitions

Cell differentiation map F
Directed graph showing
cell type transitions

Cell division and differentiation is happening in your body every single day!

$\sim 5 \times 10^{11}$cells/day



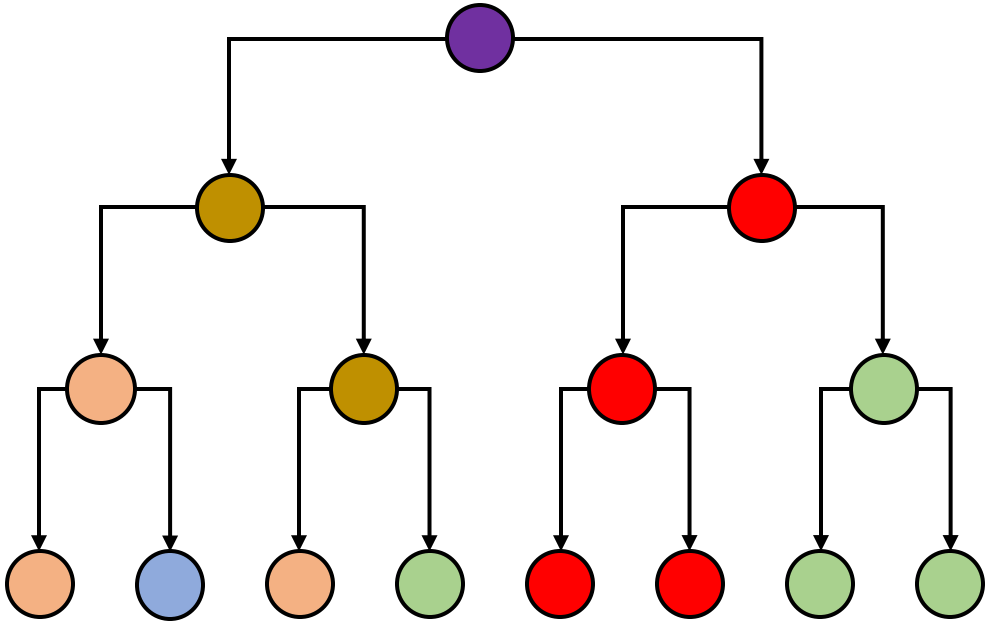
Human hematopoiesis differentiation map

Organismal Development

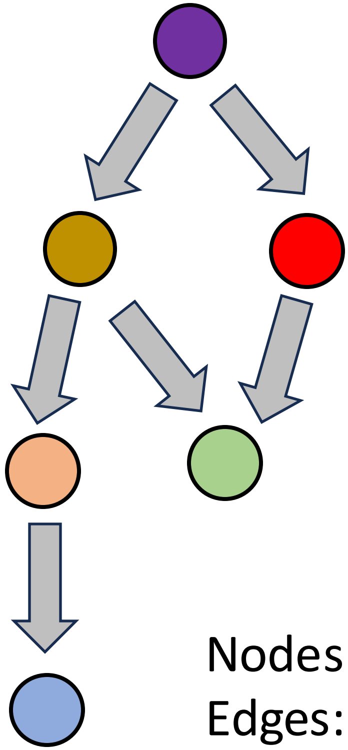
One cell
One cell type

Multiple cells
Multiple cell types

Embryogenesis



Cell lineage tree T

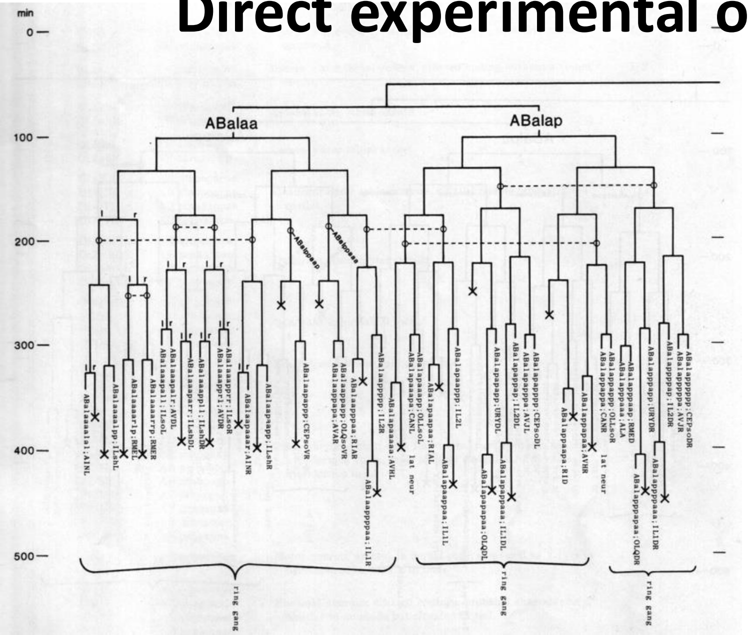


Nodes: cell types
Edges: transitions

Cell differentiation map F

Central problem in developmental biology
What is the history of cell division and differentiation during development?

Direct experimental observations



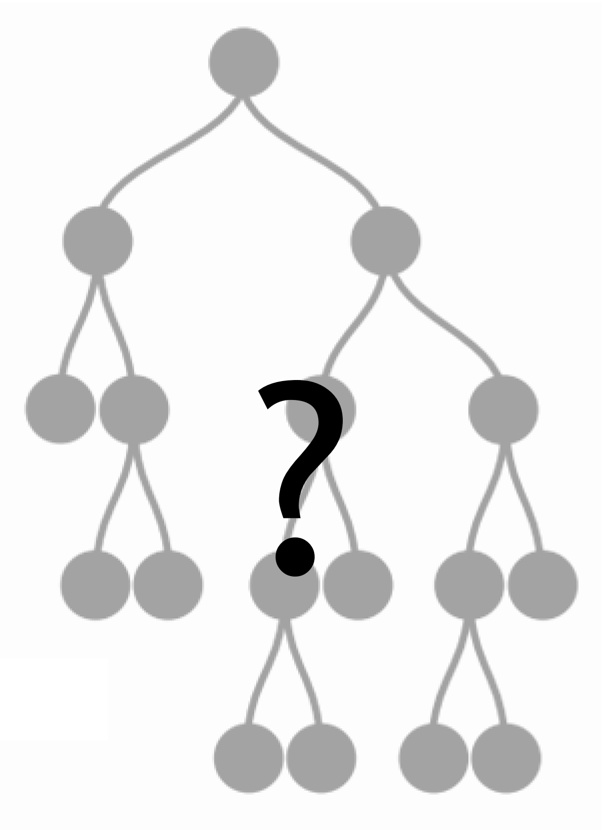
Caenorhabditis elegans

959 cells

Cell division history and differentiation of **every** cell has been mapped!

2002 Nobel Prize in Physiology or Medicine S. Brenner, H. Horvitz and **J. Sulston**

"for their discoveries concerning genetic regulation of organ development and programmed cell death"

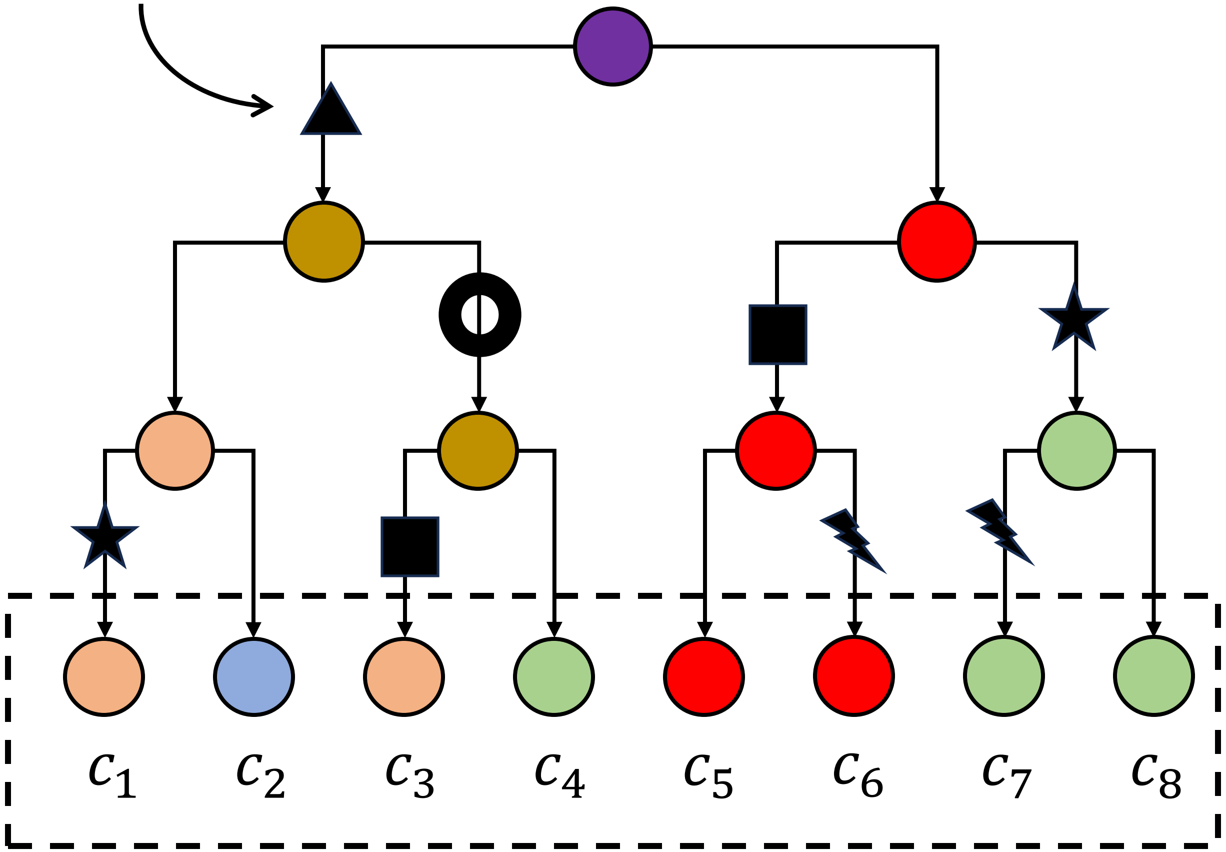


$\approx 26 \times 10^9$ cells

What is the history of cell division and differentiation during mammalian development?

The Era of Lineage Tracing Technologies

Artificial mutations introduced using **genome editing** tools such as **CRISPR-Cas9**



Cell lineage tree T

Single-cell sequencing

Measurement of mutations and cell types of leaves of the tree

	▲	○	★	■	ϟ	Cell type
c_1	1	0	1	0	0	🟧
c_2	1	0	0	0	0	🟦
c_3	1	1	0	1	0	🟧
c_4	1	1	0	0	0	🟩
c_5	0	0	0	1	0	🟥
c_6	0	0	0	1	1	🟥
c_7	0	0	1	0	1	🟩
c_8	0	0	1	0	0	🟩

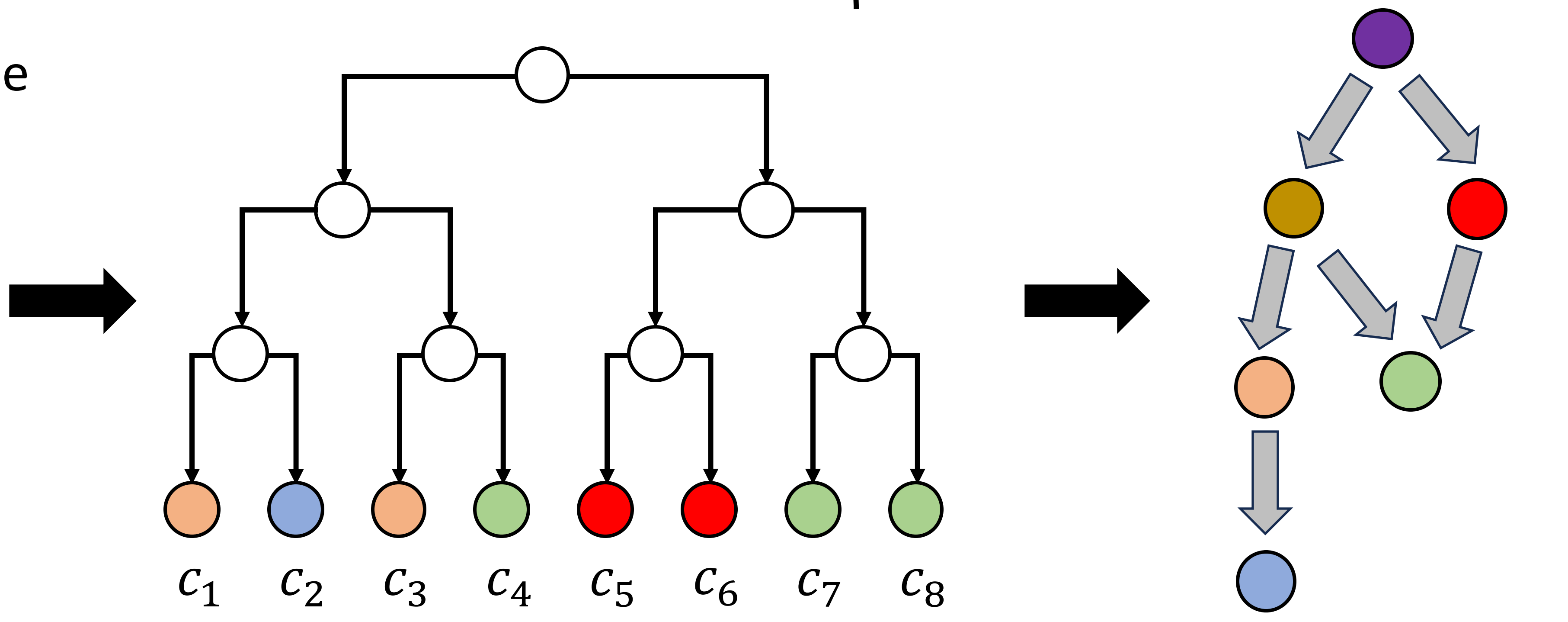
Character matrix M

Lineage tracing data

Problem 1: Cell lineage tracing

$$\begin{array}{c|ccccc|c}
 & \blacktriangle & \circledcirc & \bigstar & \blacksquare & \text{⚡} & \text{Cell type} \\
c_1 & 1 & 0 & 1 & 0 & 0 & \text{🟧} \\
c_2 & 1 & 0 & 0 & 0 & 0 & \text{🟦} \\
c_3 & 1 & 1 & 0 & 1 & 0 & \text{🟧} \\
c_4 & 1 & 1 & 0 & 0 & 0 & \text{🟩} \\
c_5 & 0 & 0 & 0 & 1 & 0 & \text{🟥} \\
c_6 & 0 & 0 & 0 & 1 & 1 & \text{🟥} \\
c_7 & 0 & 0 & 1 & 0 & 1 & \text{🟩} \\
c_8 & 0 & 0 & 1 & 0 & 0 & \text{🟩}
\end{array}$$

Character matrix M



Leaf labeled cell lineage tree T

Differentiation map F

Problem 2: Cell differentiation mapping

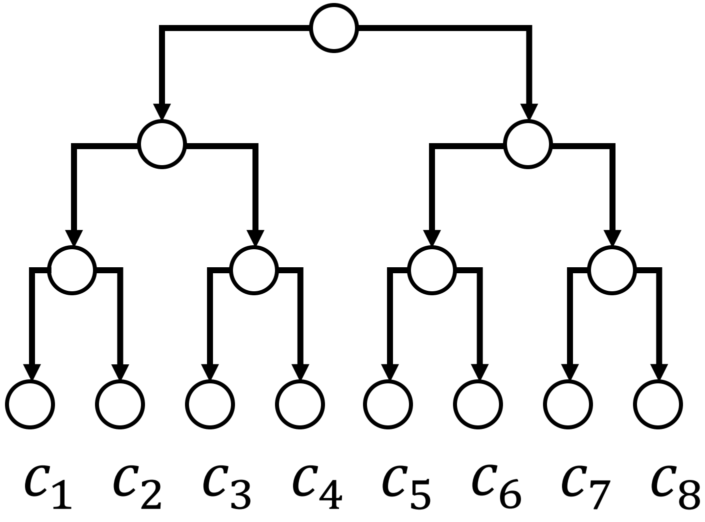
(1) Cell lineage tracing

- Star homoplasy model for CRISPR-Cas9 mutations
- **Startle** infers more accurate cell lineage trees than competing methods

Sashittal*, Schmidt* et al., *Cell Systems,* 2023
Also accepted at RECOMB 2023

$$\begin{array}{c|ccccc} & \blacktriangle & \bigcirc & \bigstar & \blacksquare & \text{⚡} \\ \hline c_1 & 1 & 0 & 1 & 0 & 0 \\ c_2 & 1 & 0 & 0 & 0 & 0 \\ c_3 & 1 & 1 & 0 & 1 & 0 \\ c_4 & 1 & 1 & 0 & 0 & 0 \\ c_5 & 0 & 0 & 0 & 1 & 0 \\ c_6 & 0 & 0 & 0 & 1 & 1 \\ c_7 & 0 & 0 & 1 & 0 & 1 \\ c_8 & 0 & 0 & 1 & 0 & 0 \end{array}$$

Character matrix M

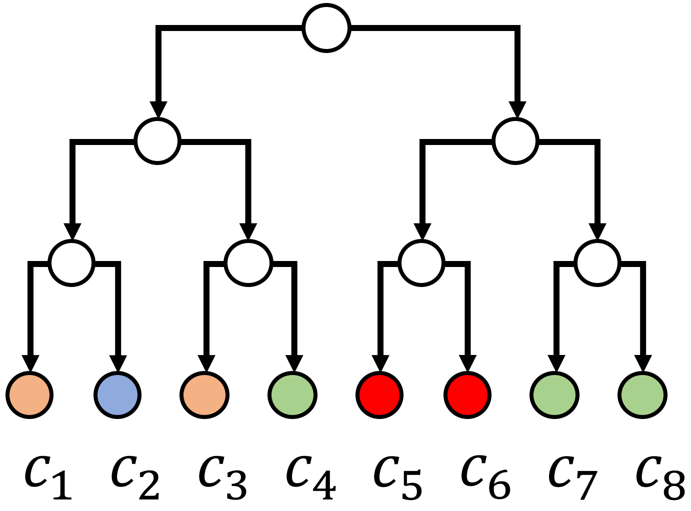


cell lineage tree T

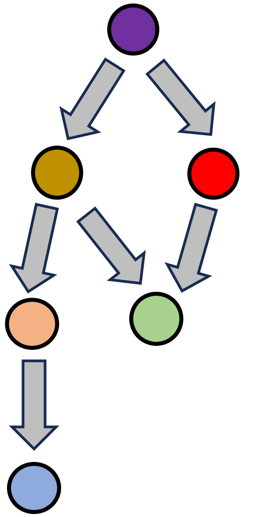
(2) Cell differentiation mapping

- Formalized the problem of inferring cell differentiation maps from lineage tracing data
- **Carta** balances the trade-off between the complexity and fit of the differentiation map

Sashittal*, Zhang* et al., *Nature Methods,* 2025
Also accepted at RECOMB 2025

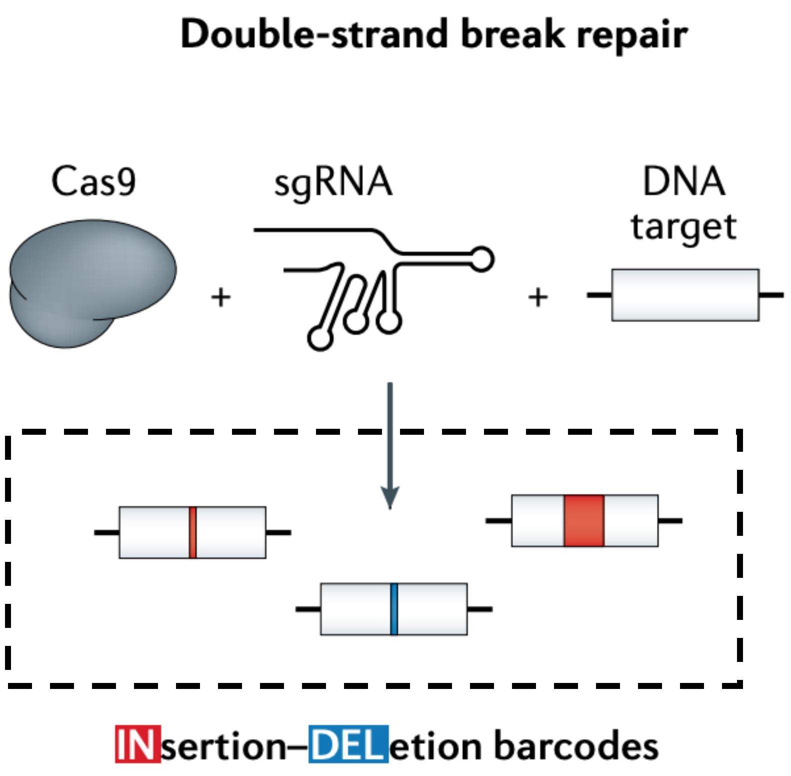


cell lineage tree T

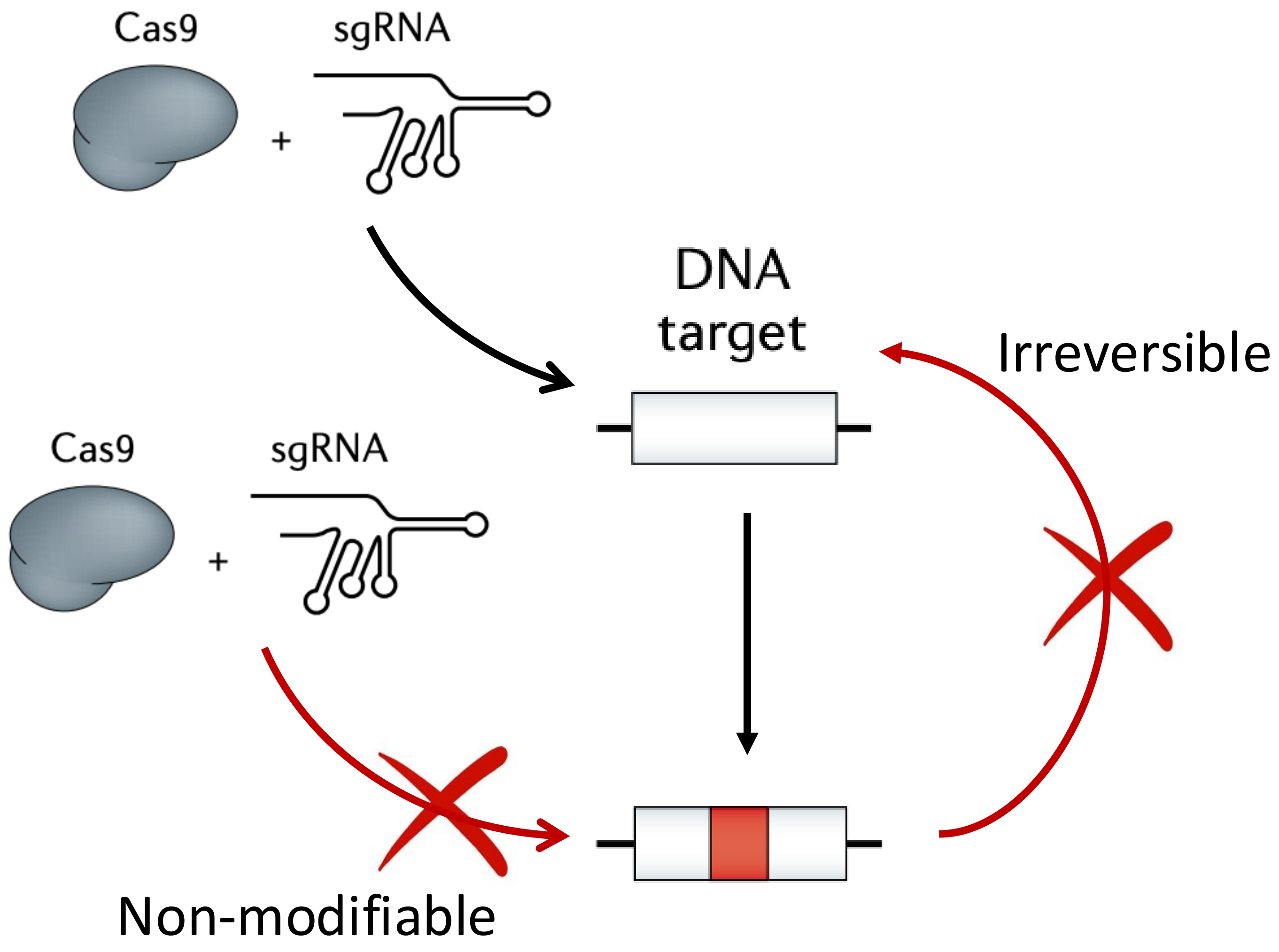


Differentiation map F

CRISPR-Cas9-based lineage tracing

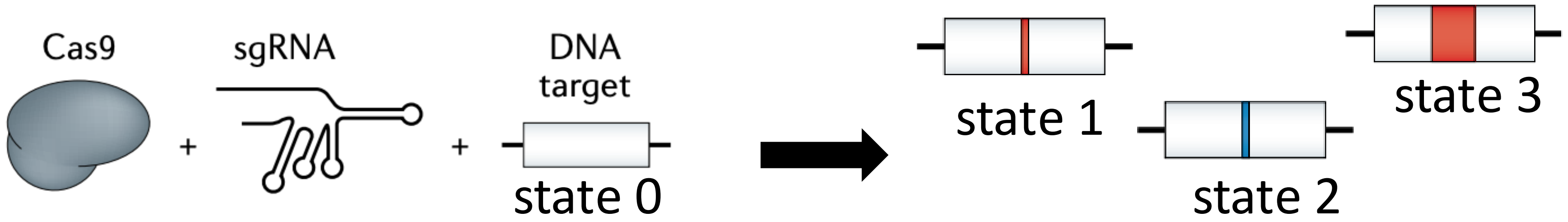


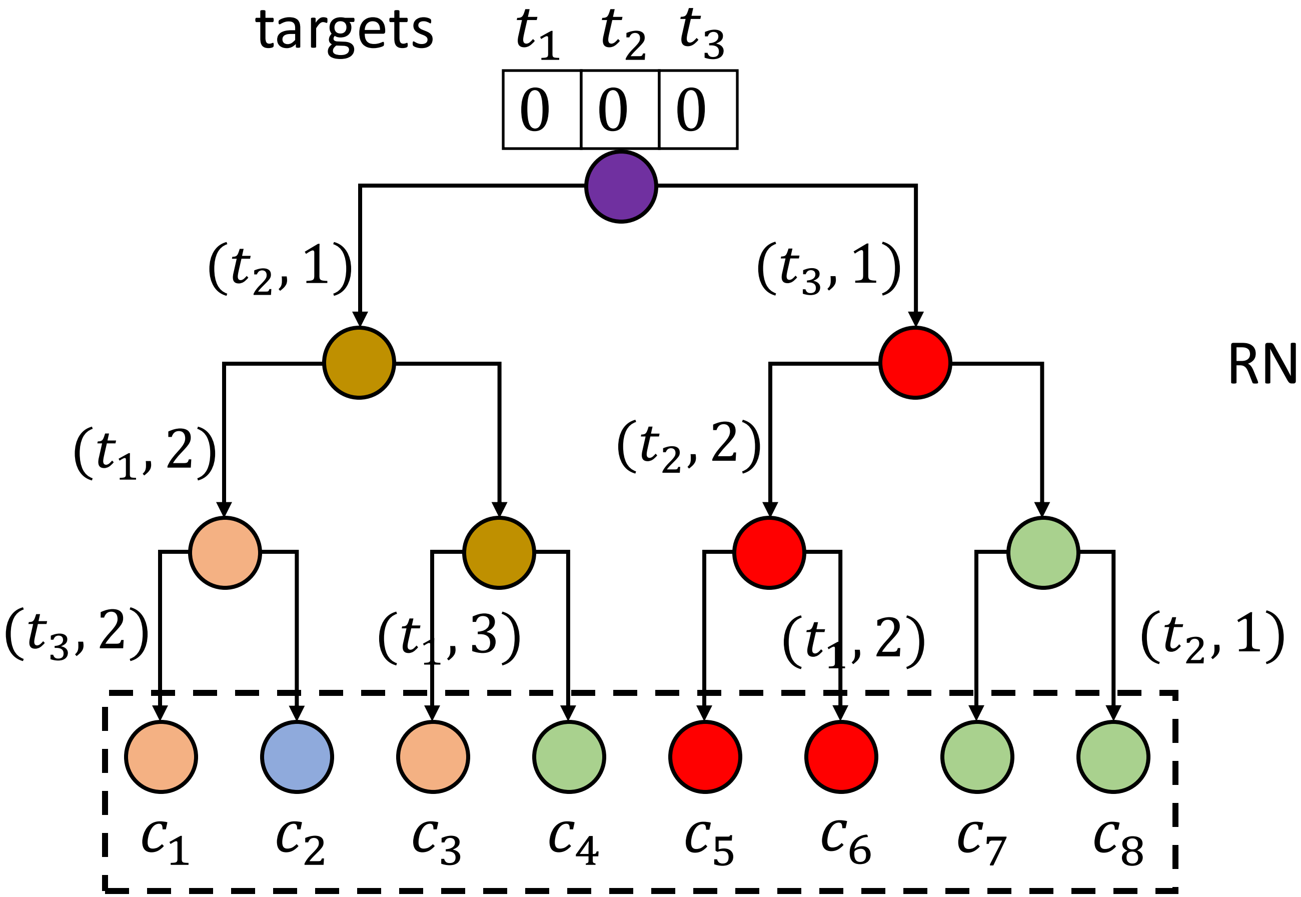
- ✓ Irreversible
- ✓ Non-modifiable
- ✓ Multi-state



Chan et al. Nature 2019

CRISPR-Cas9-based lineage tracing





Single-cell RNA sequencing

$$\begin{array}{c|c} & \begin{matrix} t_1 & t_2 & t_3 \end{matrix} \\ \begin{matrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \\ c_7 \\ c_8 \end{matrix} & \begin{pmatrix} 2 & 1 & 2 \\ 2 & 1 & 0 \\ 3 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \\ 2 & 2 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \end{array}$$

Cell type
🟧
🟦
🟧
🟩
🟥
🟥
🟩
🟩

CRISPR-Cas9-based lineage tracing

Challenges in real data

- Large number (50 to 100) of states (indels) for each character (target site)
- Large number (100s to 1000s) of cells
- Many missing entries (white) in the character matrix (around 20% dropout)

What is the model for the evolution of CRISPR-Cas9 induced mutations?

Specialized methods have been introduced and benchmarked in a DREAM challenge (Gong et al., 2021, Cell Systems)

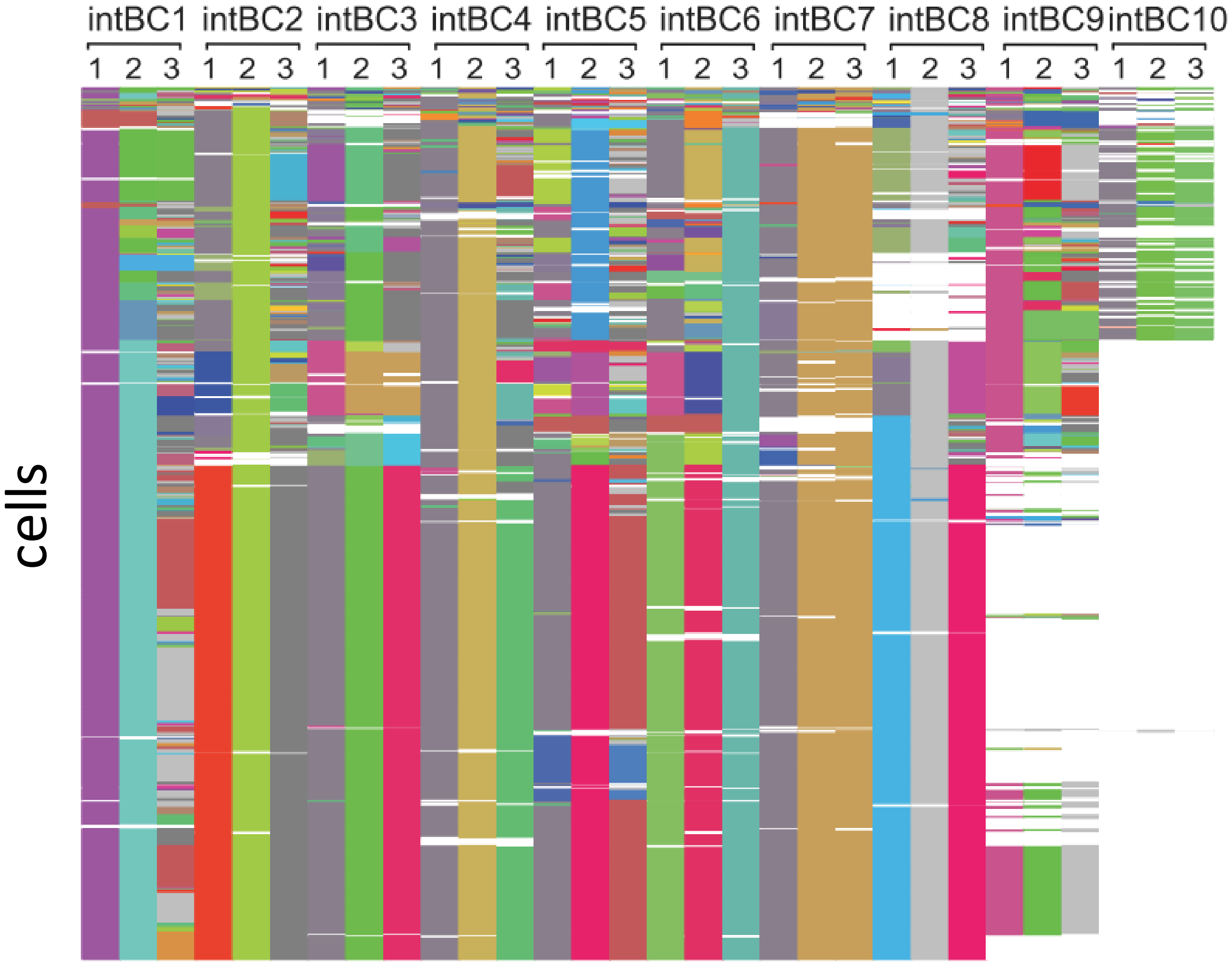
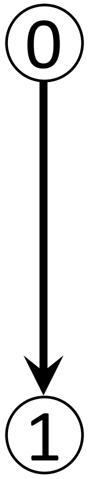


Figure adapted from Yang et al., 2022, Cell

Specialized models for CRISPR-Cas9-based lineage tracing

Two-state
Camin-Sokal model

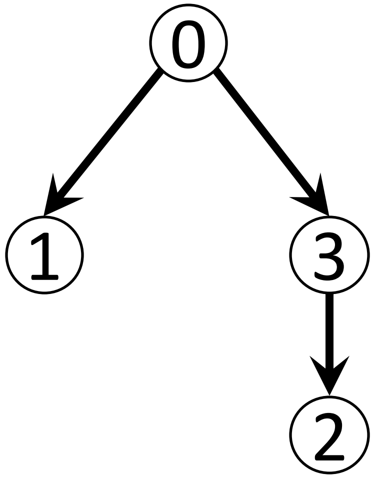


Camin et al., 1965

- ✗ Multi-state
- ✓ Irreversible
- *✓ Non-modifiable

McKenna et al., *Science* (2016)
Raj et al., *Nature Biotechnology* (2018)

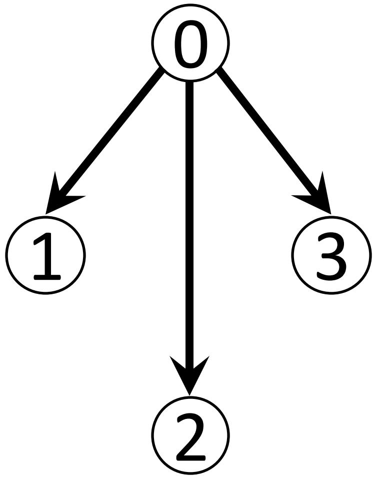
Multi-state
Camin-Sokal model



Felsenstein et al., 2004

- ✓ Multi-state
- ✓ Irreversible
- ✗ Non-modifiable

Multi-state
Star homoplasy model



Sashittal et al., 2023

- ✓ Multi-state
- ✓ Irreversible
- ✓ Non-modifiable

Star homoplasy tree inference problem statement

Star Homoplasy Problem [**Sashittal et al., 2023**]
Given character matrix M and mutation weights w, find star homoplasy phylogeny T for M that minimizes parsimony score $W(T)$.

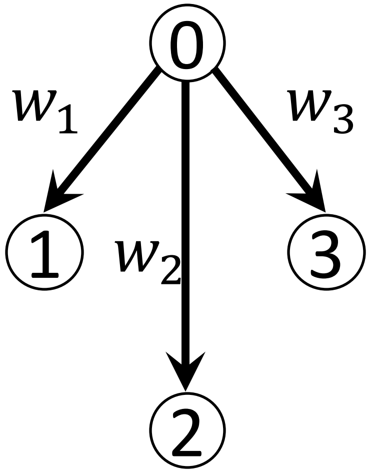
Theorem [**Sashittal et al., 2023**]
Star homoplasy problem is NP-hard, even when the number k of homoplasies is fixed and $k \geq 4$.

*Reduction from *Cubic Vertex Cover Problem*

Input

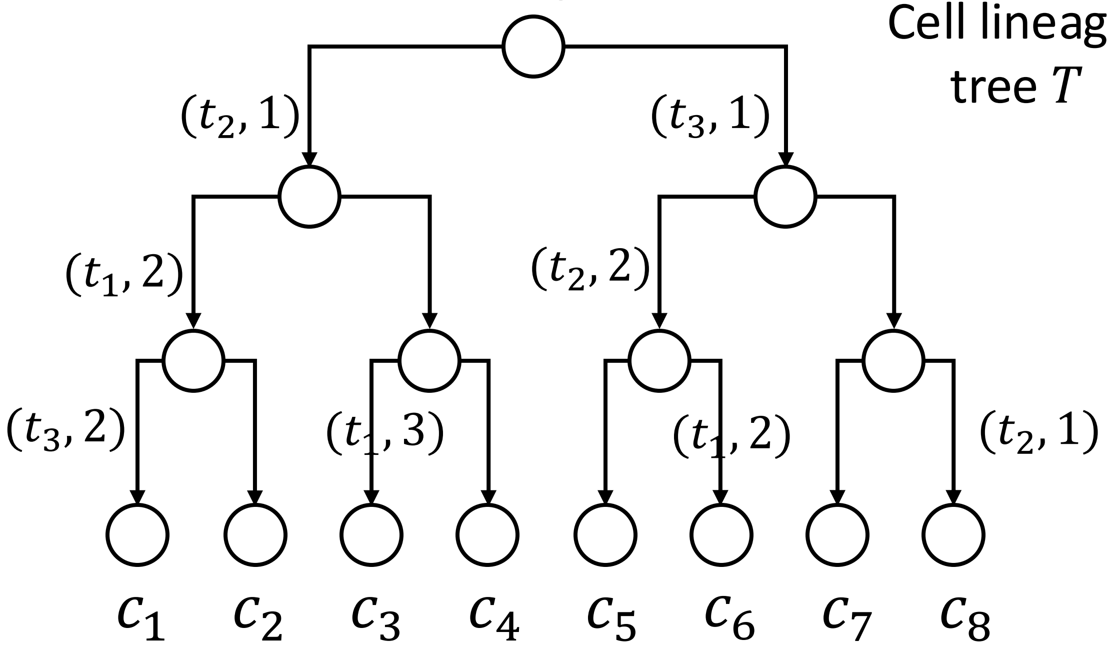
$$\begin{array}{c|ccc} & t_1 & t_2 & t_3 \\ \hline c_1 & 2 & 1 & 2 \\ c_2 & 2 & 1 & 0 \\ c_3 & 3 & 1 & 0 \\ c_4 & 0 & 1 & 0 \\ c_5 & 0 & 2 & 1 \\ c_6 & 2 & 2 & 1 \\ c_7 & 0 & 0 & 1 \\ c_8 & 0 & 1 & 1 \end{array}$$

Character matrix M



Weights indicating probability of mutation

Output



Cell lineage tree T

$$W(T) = 3w_1 + 4w_2 + w_3$$

Startle performs hill climbing in the space of trees

Search through tree space using NNI moves

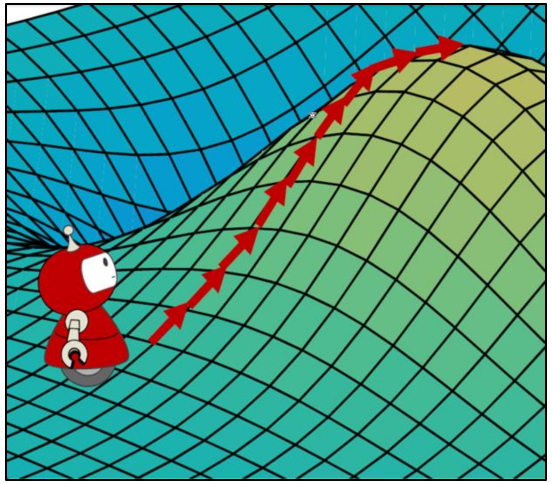
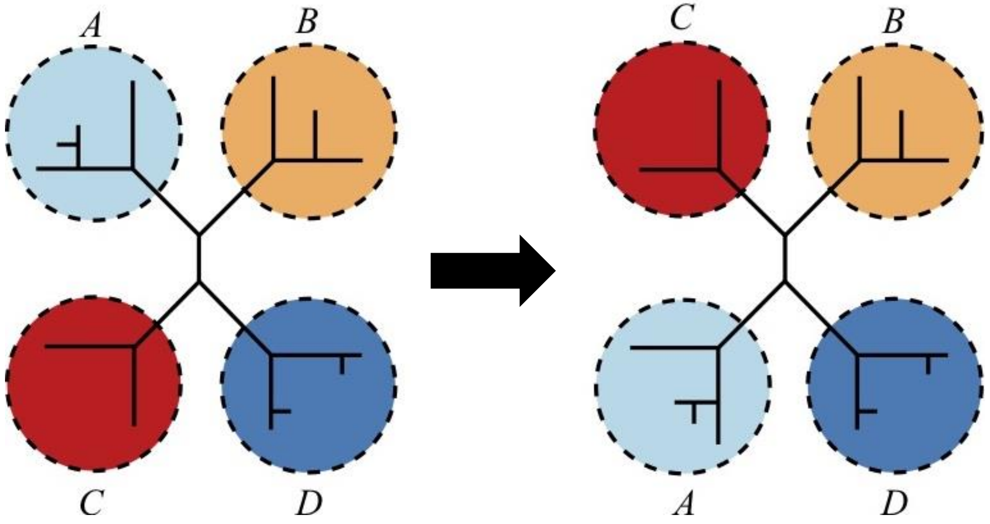


Figure from mathworks.com



Nearest neighbor interchange (NNI)

How do we evaluate a given tree T?

Small Star Homoplasy Problem [Sashittal et al., 2023]
Given a tree T for character matrix M and mutation weights w, find the minimum parsimony score $W(T)$.

Theorem [Sashittal et al., 2023]:
Small Star Homoplasy problem can be solved using dynamic programing in $O(nm)$ time.

Theorem [Sashittal et al., 2023]:
We can compute parsimony scores $W(T')$ for all $O(n)$ trees T' in the NNI neighborhood of a tree T in $O(nmd)$ time, where d is the average depth of T.

*Naïve implementation will take $O(n^2m)$ time

Star tree lineage estimator (**Startle**)

Henri Schmidt

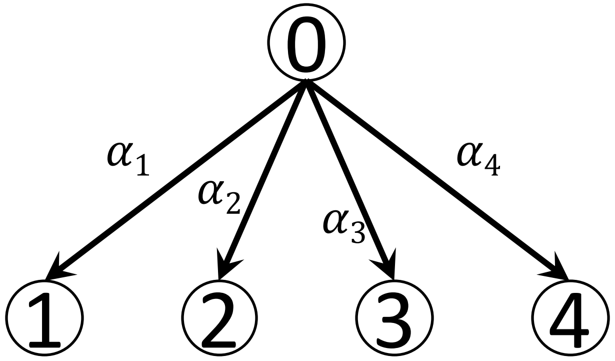
Benjamin Raphael

Character matrix

		characters	
		c_1	c_2
cells	s_1	1	0
	s_2	1	2
	s_3	1	1
	s_4	2	1

Star homoplasy model

Unmutated state



Mutated states

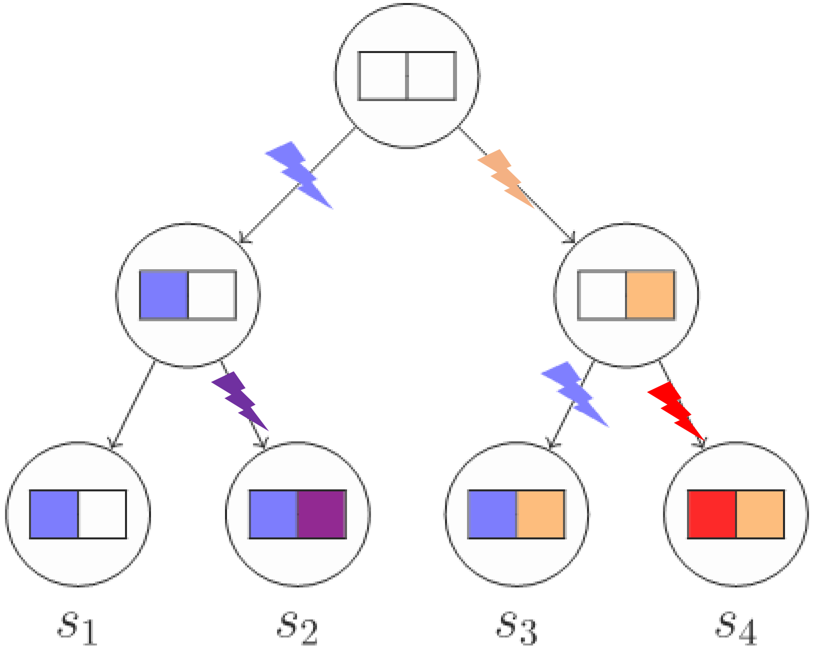
Startle

Tree search using nearest neighbor interchange (NNI) moves

+

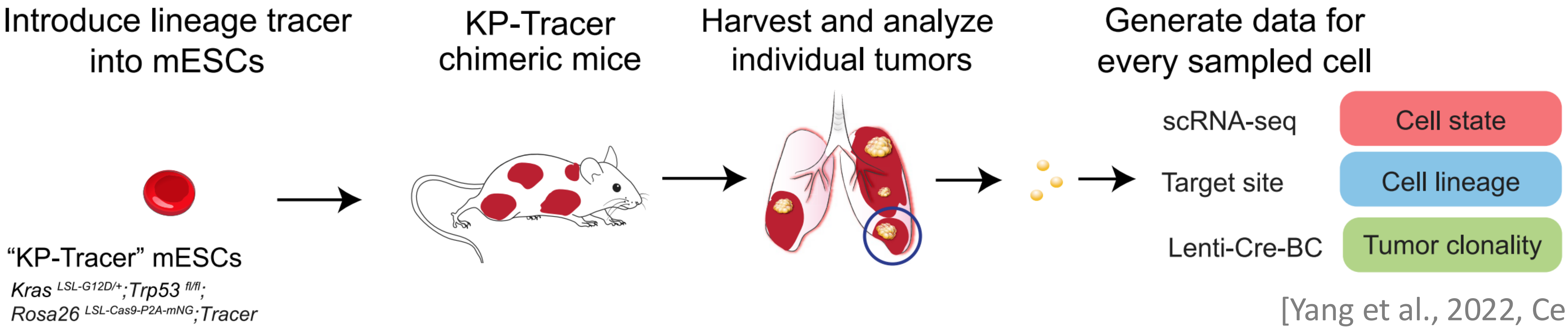
ILP for bounded-homoplasy version

Maximum parsimony star homoplasy phylogeny



Sashittal*, Schmidt*, et al.
RECOMB 2023; Cell Systems 2023

Mouse metastatic lung adenocarcinoma data



Introduce lineage tracer into mESCs

KP-Tracer chimeric mice

Harvest and analyze individual tumors

Generate data for every sampled cell

"KP-Tracer" mESCs

$\mathit{Kras}^{LSL\text{-}G12D/+}$;$\mathit{Trp53}^{fl/fl}$;
$\mathit{Rosa26}^{LSL\text{-}Cas9\text{-}P2A\text{-}mNG}$;*Tracer*

scRNA-seq	Cell state
Target site	Cell lineage
Lenti-Cre-BC	Tumor clonality

[Yang et al., 2022, Cell]

Largest dataset in the study
(3724_NT_T1_All):

n = 21108 cells across 5 tumors

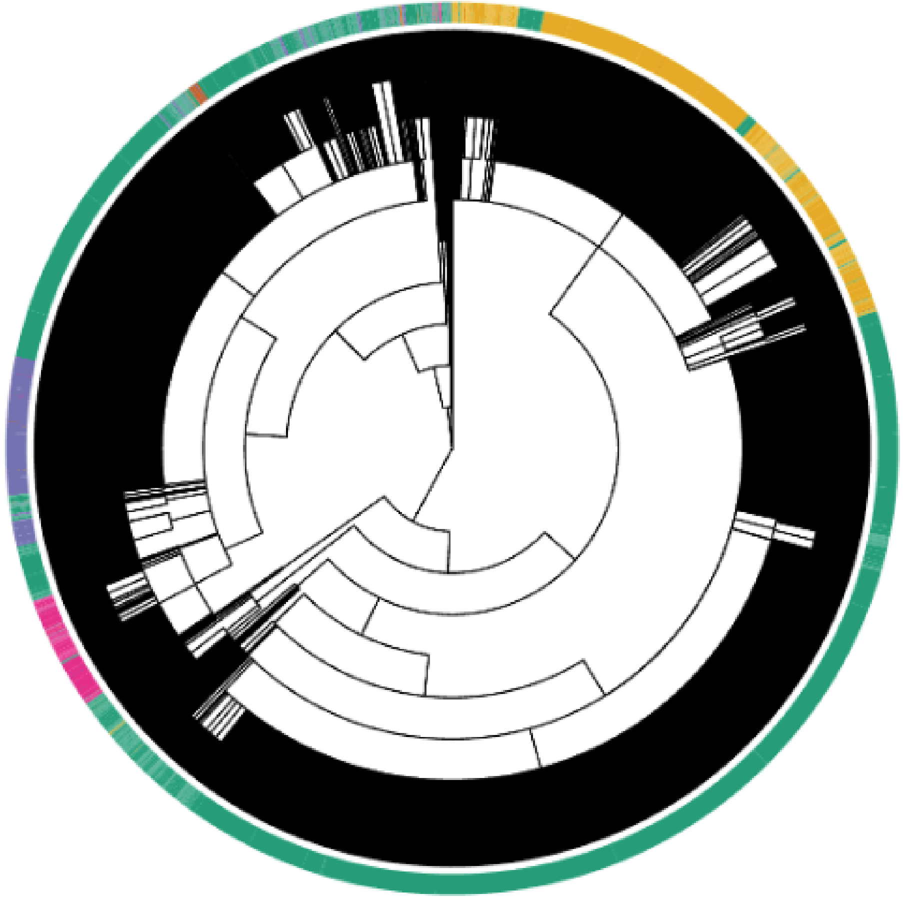
What is the cell lineage tree for these cancer cells?

Tumor	**# of cells**
Lung	14852
Soft tissue	3891
Liver met 1	90
Liver met 2	1512
Liver met 3	863

Startle produces more parsimonious trees

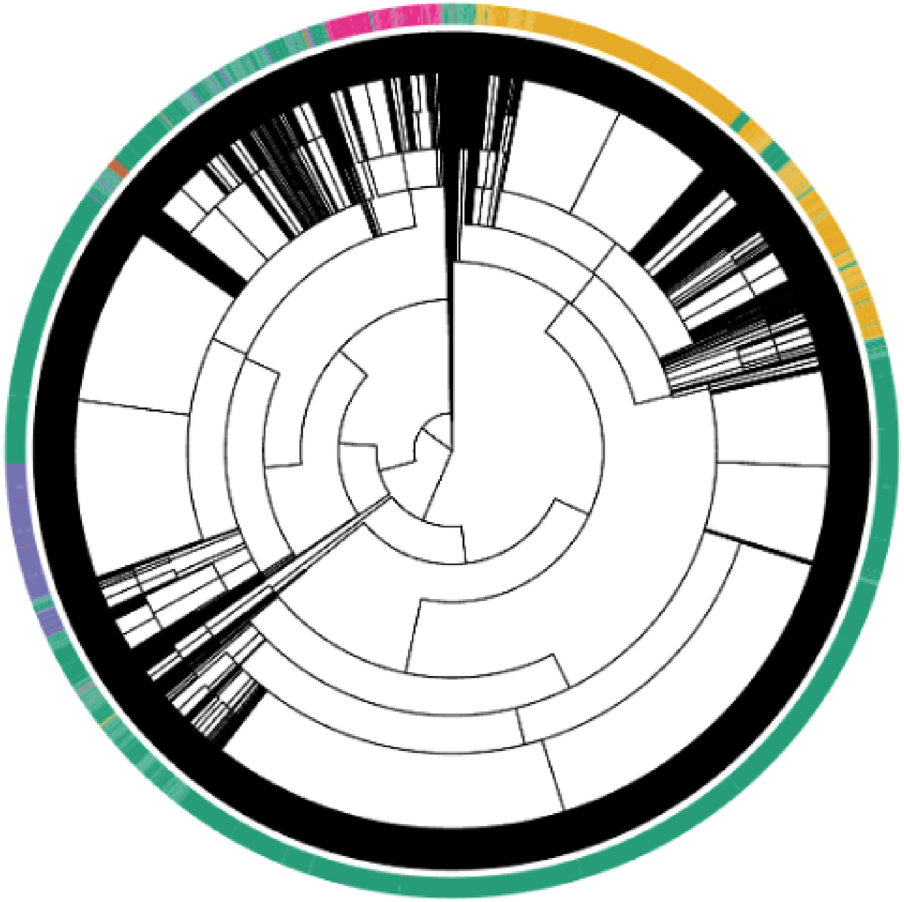
Published phylogeny

Cassiopeia [Jones et al., 2021, *Genome Biology*]



Parsimony Score = 4827.43

Startle phylogeny



Parsimony Score = **4715.5**

Anatomical sites (cells)

🟩 Primary tumor (14852)

🟧 Liver met. 1 (90)

🟪 Liver met. 2 (1512)

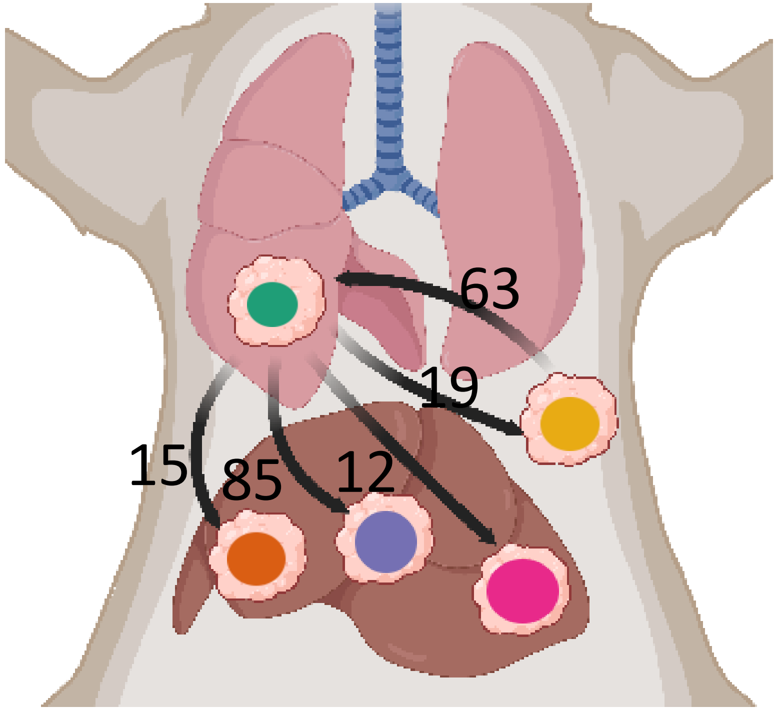
🟥 Liver met. 3 (863)

🟨 Soft tissue met. (3891)

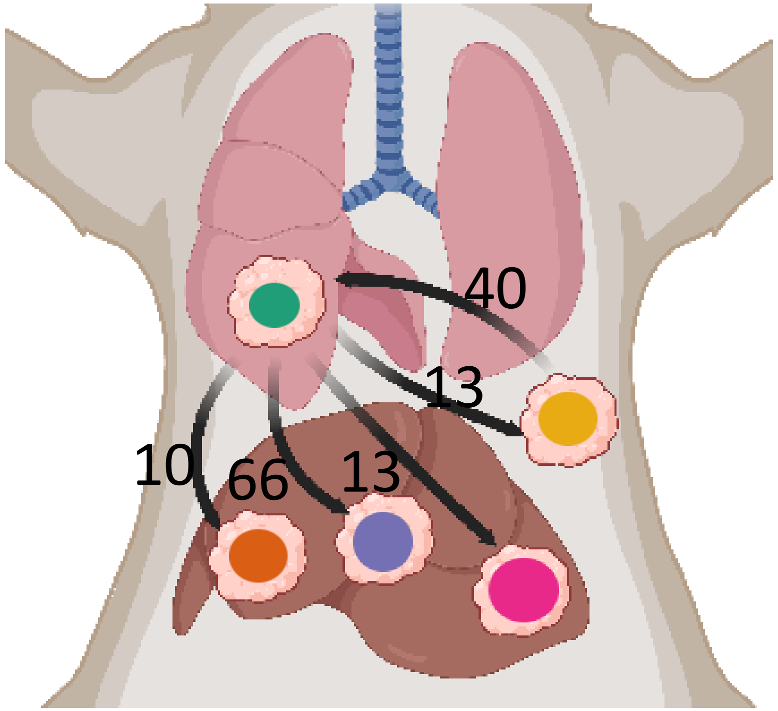
Total cells: 21108

Startle trees have fewer migrations between anatomical sites

Inferred* migrations from published tree



Inferred* migrations from Startle tree



Anatomical sites (cells)

- 🟩 Primary tumor (14852)
- 🟧 Liver met. 1 (90)
- 🟪 Liver met. 2 (1512)
- 🟥 Liver met. 3 (863)
- 🟨 Soft tissue met. (3891)

Total cells: 21108

*MACHINA: El-Kebir et al., 2018, *Nature Genetics*

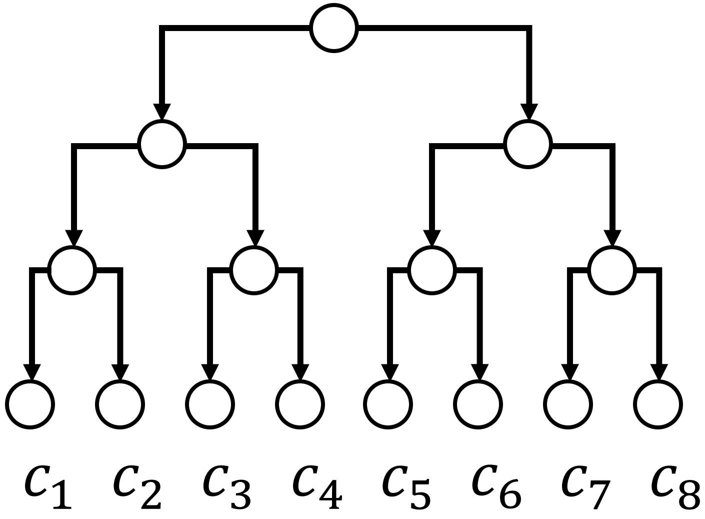
(1) Cell lineage tracing

- Star homoplasy model for CRISPR-Cas9 mutations
- **Startle** infers more accurate cell lineage trees than competing methods

Sashittal*, Schmidt* et al., *Cell Systems,* 2023
Also accepted at RECOMB 2023

$$\begin{array}{c} \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \\ c_7 \\ c_8 \end{array} \begin{array}{c} \begin{array}{ccccc} \blacktriangle & \circ & \star & \blacksquare & \text{⚡} \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

Character matrix M

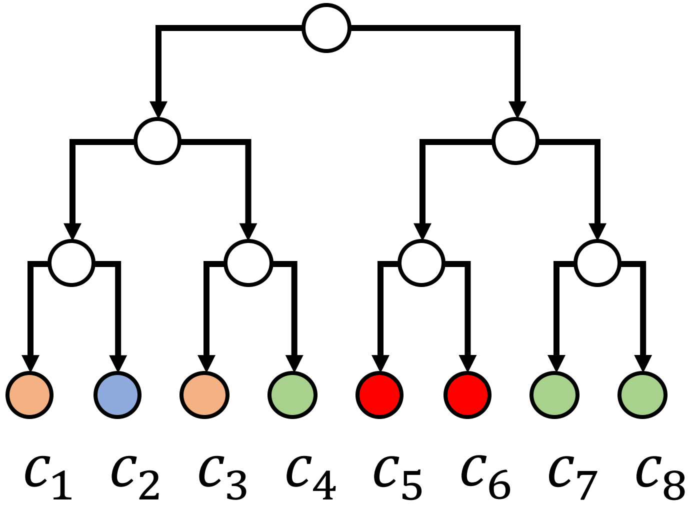


cell lineage tree T

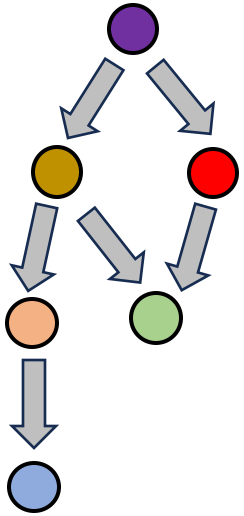
(2) Cell differentiation mapping

- Formalized the problem of inferring cell differentiation maps from lineage tracing data
- **Carta** balances the trade-off between the complexity and fit of the differentiation map

Sashittal*, Zhang* et al., *Nature Methods,* 2025
Also accepted at RECOMB 2025

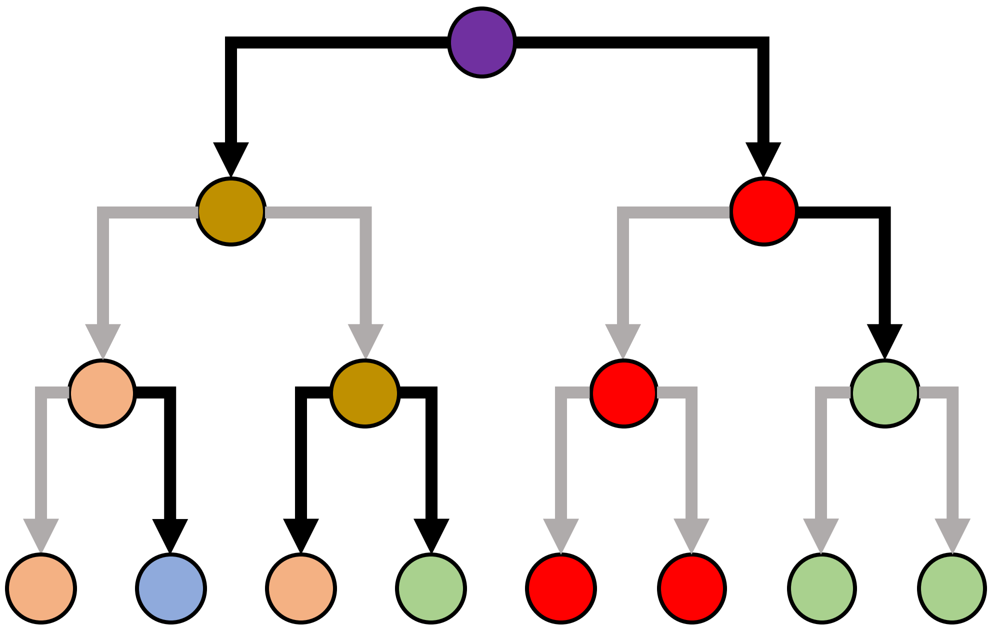


cell lineage tree T

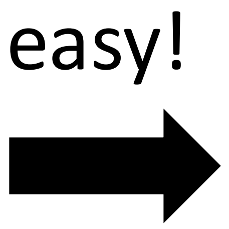


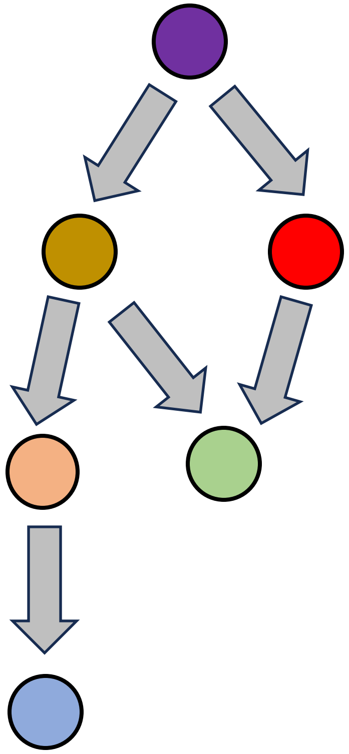
Differentiation map F

Ancestral cell types reveal the differentiation map



Cell lineage tree T
with all cell types known



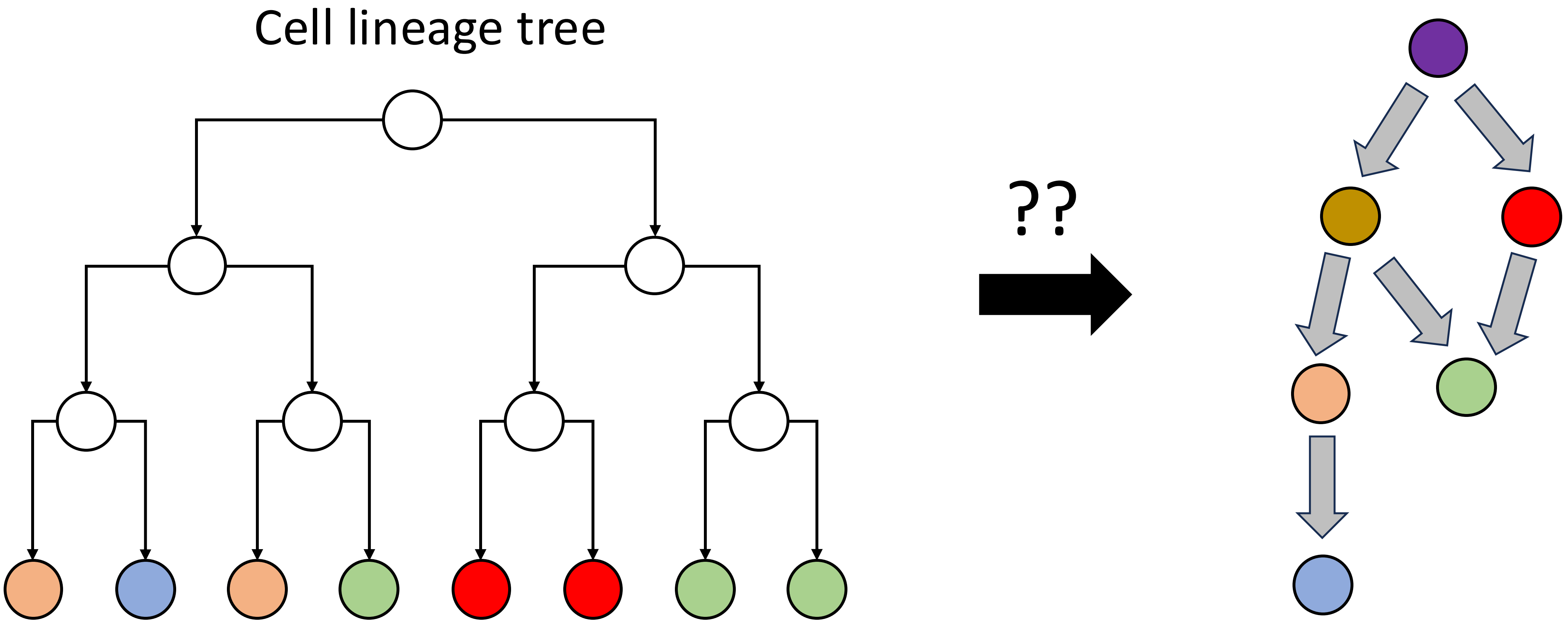


Cell differentiation map F

Given ancestral cell types, we can trivially get:

1. Cell types in the differentiation process
2. Transitions between cell types

Key challenges in cell differentiation mapping



Key challenges in inferring the type of ancestral cells

1. Which progenitors are not observed at present time?
2. Which of the observed cell types are progenitors?

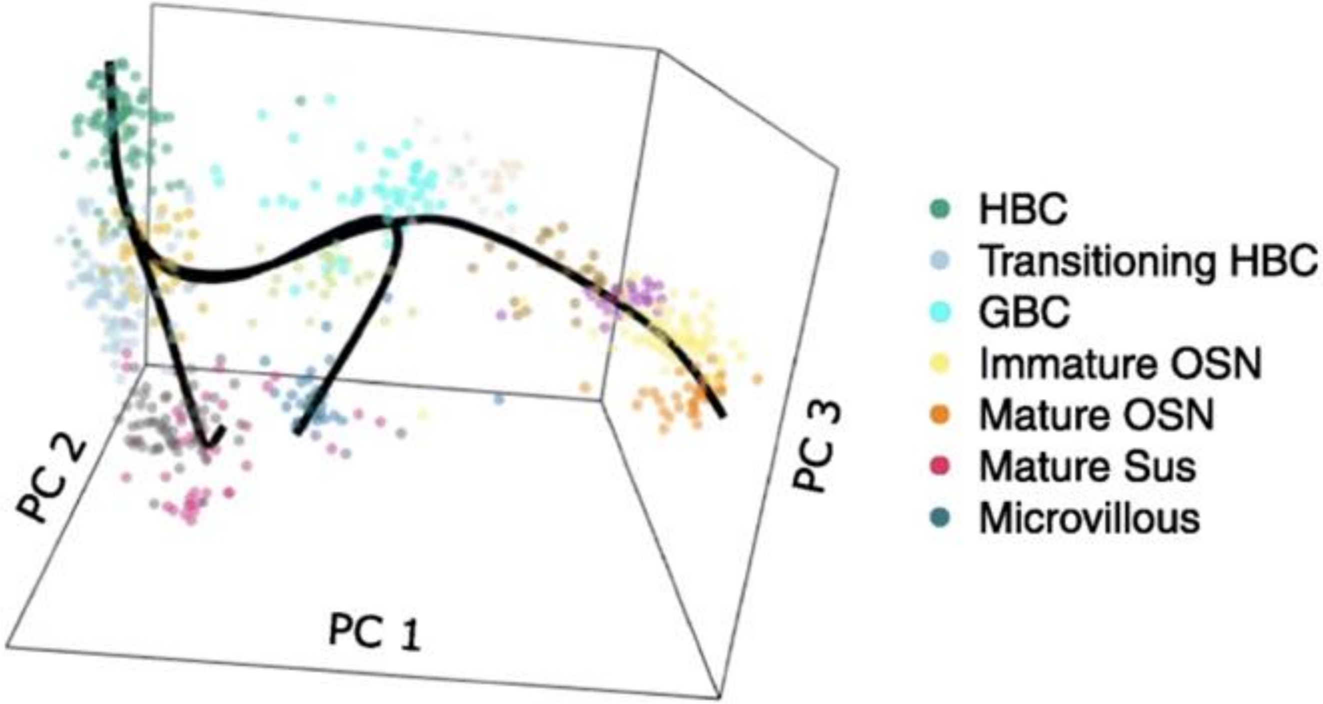
Unobserved progenitors 🟣 🟡

Observed progenitors 🟠 🔴

Terminal cell types 🔵 🟢

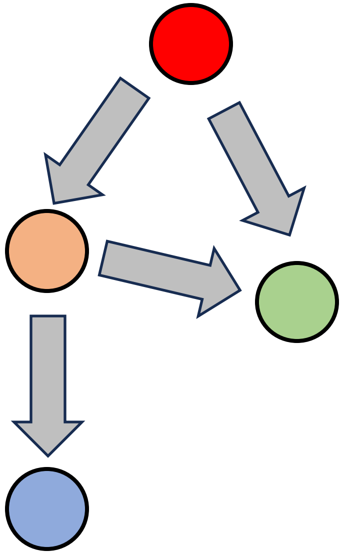
Cell differentiation mapping

scRNA-seq data from one or more timepoints
(with or without lineage information)



Principal curves or ridge estimation

All progenitors
are observed



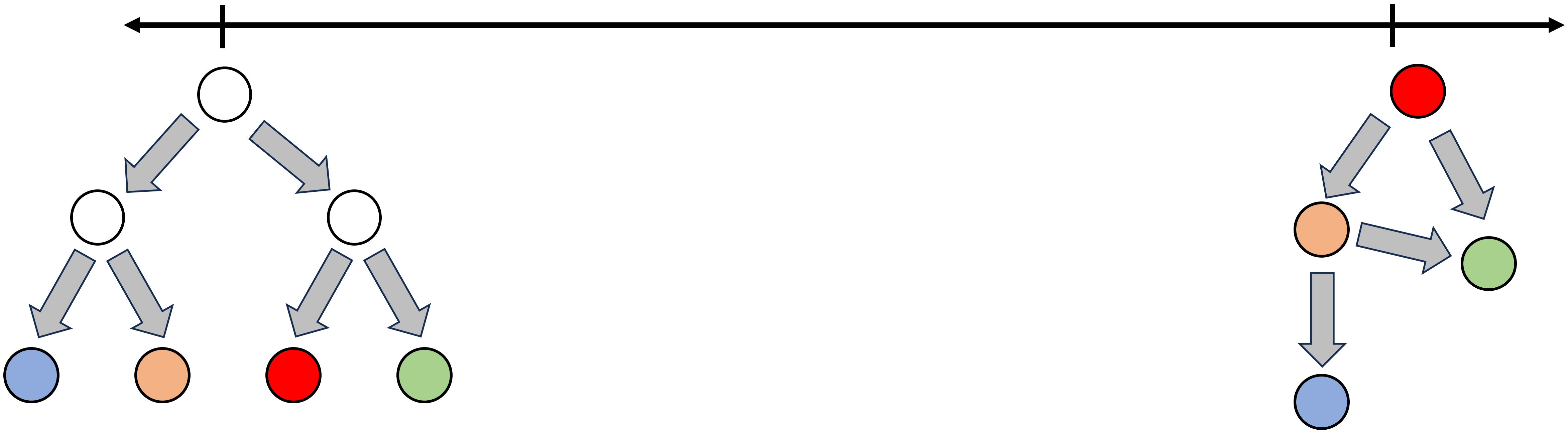
Trajectory inference methods

Trapnell et al., 2014, Nat. Biotech.; Haghverdi et al., 2016, Nat. Methods; Manno et al., 2018, Nature; Qiu et al., 2017a, Nat. Methods; Setty et al., 2016, Nat. Biotech and many more ….

Cell differentiation mapping

None of the progenitors
are observed

All progenitors
are observed



Distance-based heuristics
to infer tree-structured
differentiation maps

Chan et al., 2019, Nature.; Yang et al., 2022, Cell; Kahlor et al., 2022, Cell

Trajectory inference methods

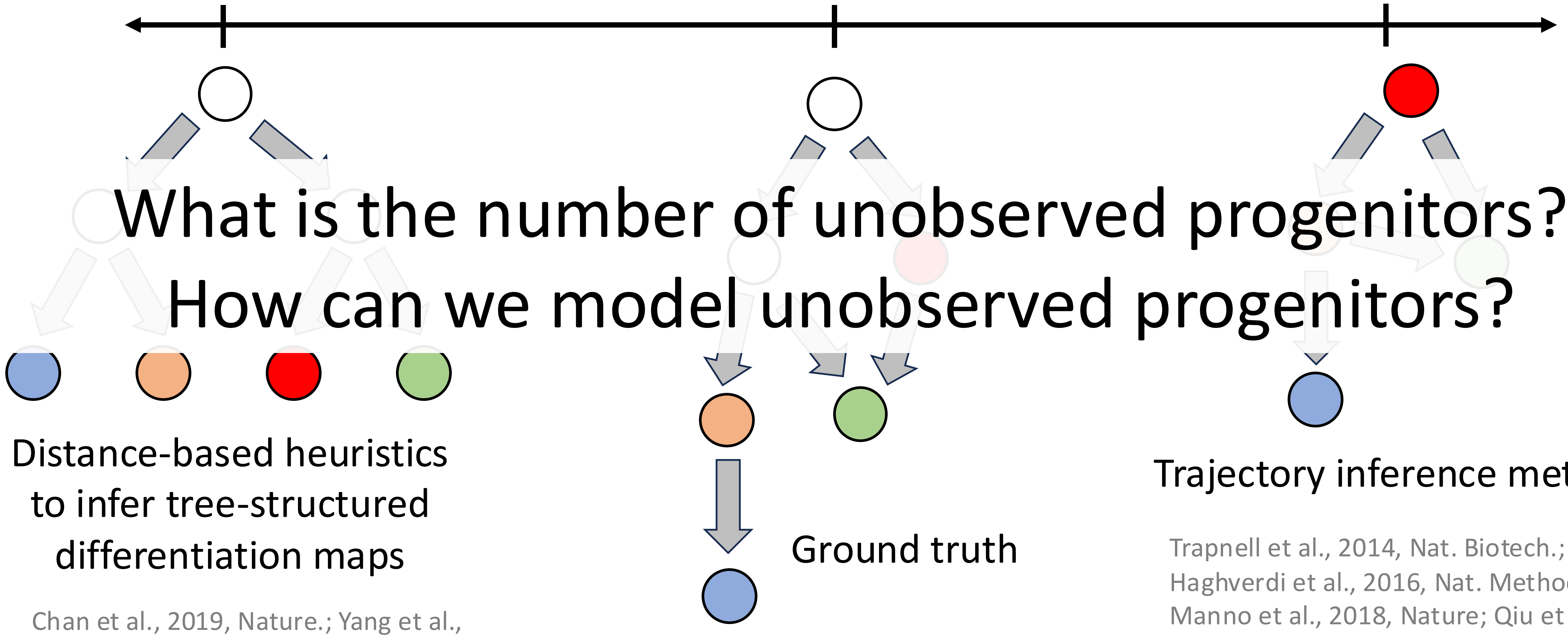
Trapnell et al., 2014, Nat. Biotech.; Haghverdi et al., 2016, Nat. Methods; Manno et al., 2018, Nature; Qiu et al., 2017a, Nat. Methods; Setty et al., 2016, Nat. Biotech and many more ....

Cell differentiation mapping

None of the progenitors are observed

Early progenitors are not observed
Late progenitors are observed

All progenitors are observed



What is the number of unobserved progenitors?
How can we model unobserved progenitors?

Distance-based heuristics to infer tree-structured differentiation maps

Chan et al., 2019, Nature.; Yang et al., 2022, Cell; Kahlor et al., 2022, Cell

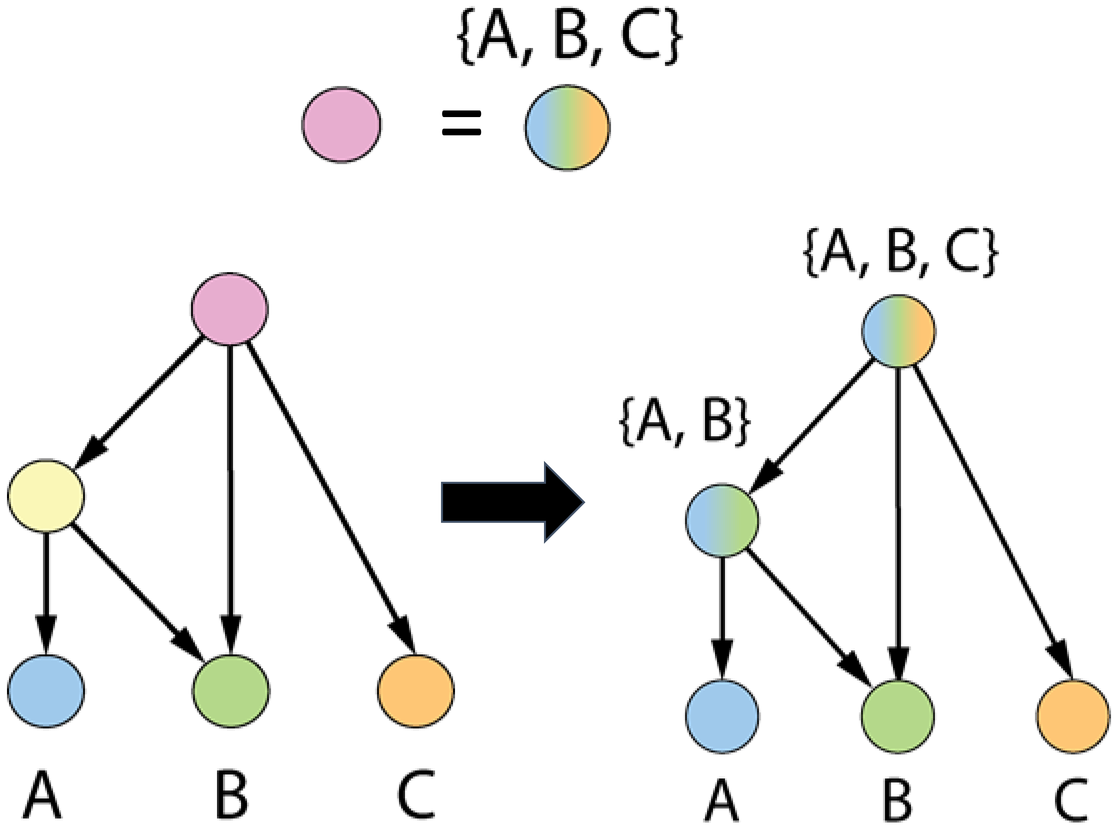
Ground truth

Trajectory inference methods

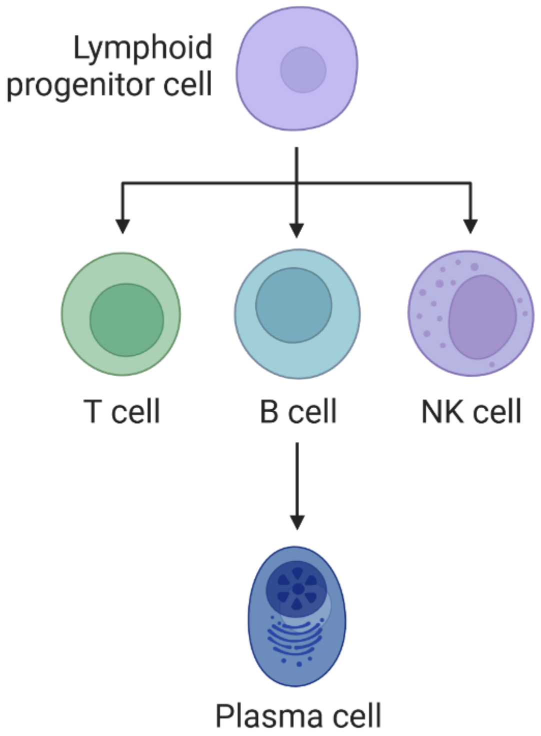
Trapnell et al., 2014, Nat. Biotech.; Haghverdi et al., 2016, Nat. Methods; Manno et al., 2018, Nature; Qiu et al., 2017a, Nat. Methods; Setty et al., 2016, Nat. Biotech and many more ….

Modeling unobserved progenitors: **Potency Set**

Definition: *potency* set *S* = {cell types that their descendants can differentiate into}

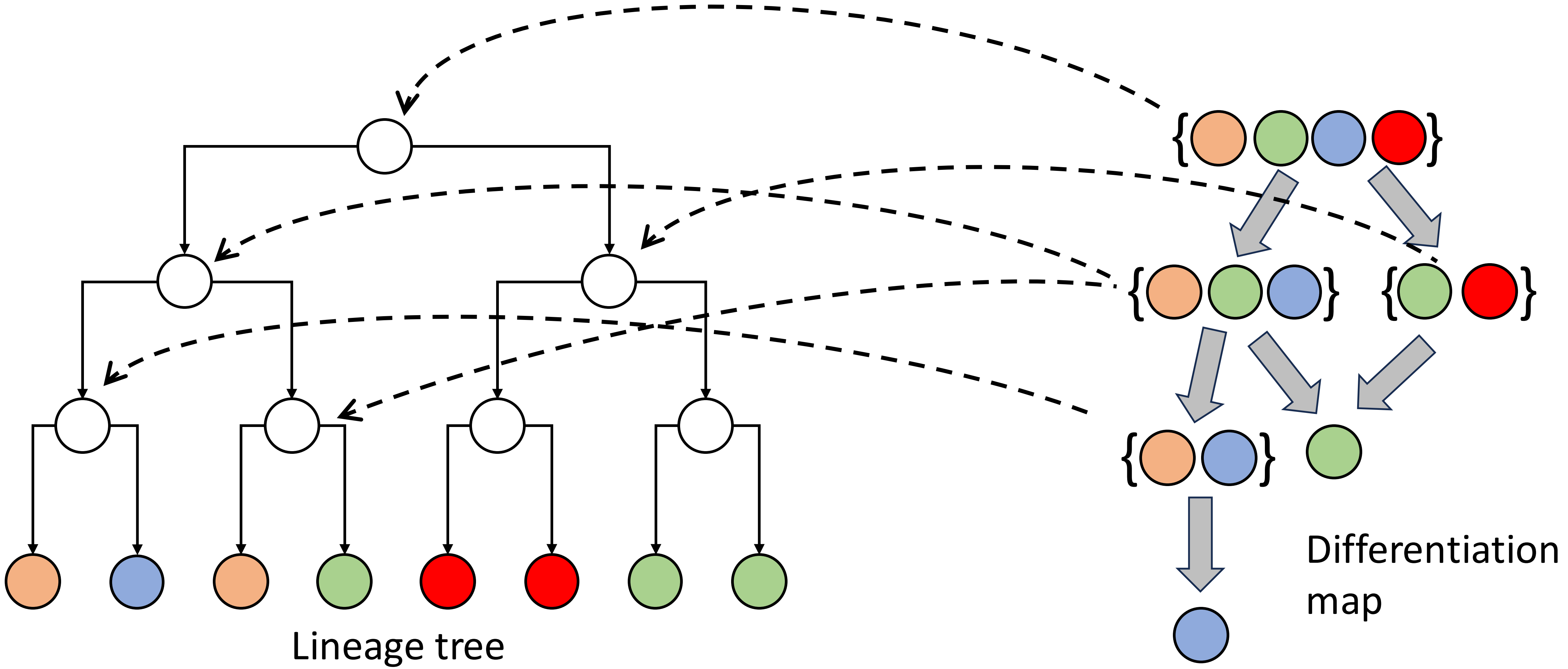


Formalizes how developmental biologists describe progenitors



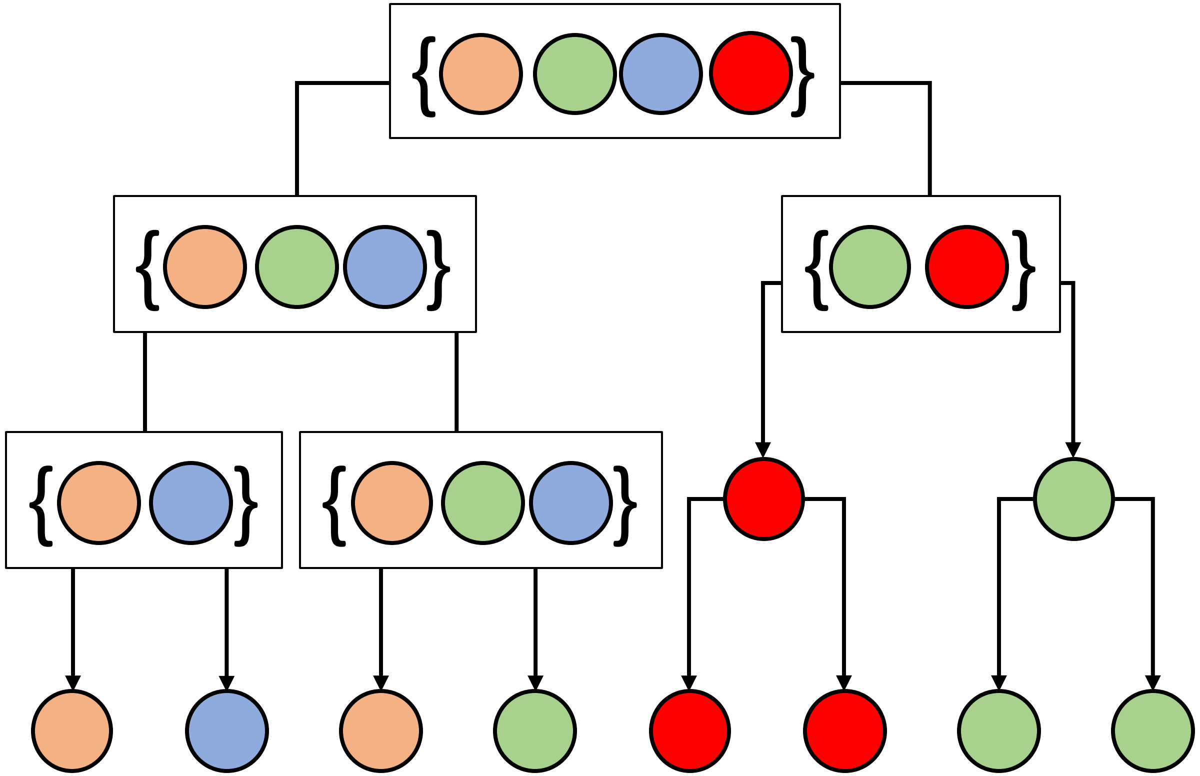
Lymphoid progenitor cells differentiates into lymphoid cells

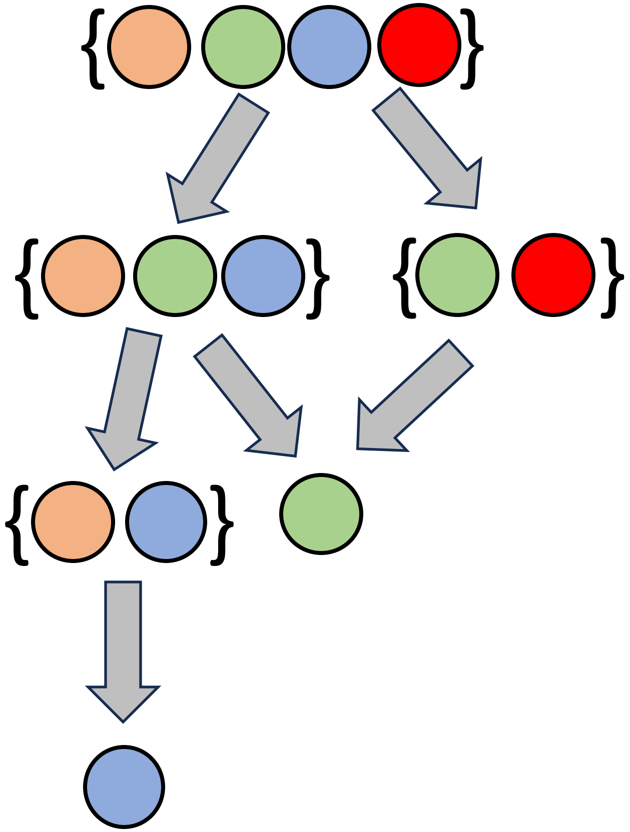
Cell differentiation map labels ancestors in cell lineage tree



How well does the cell differentiation map fit the data?

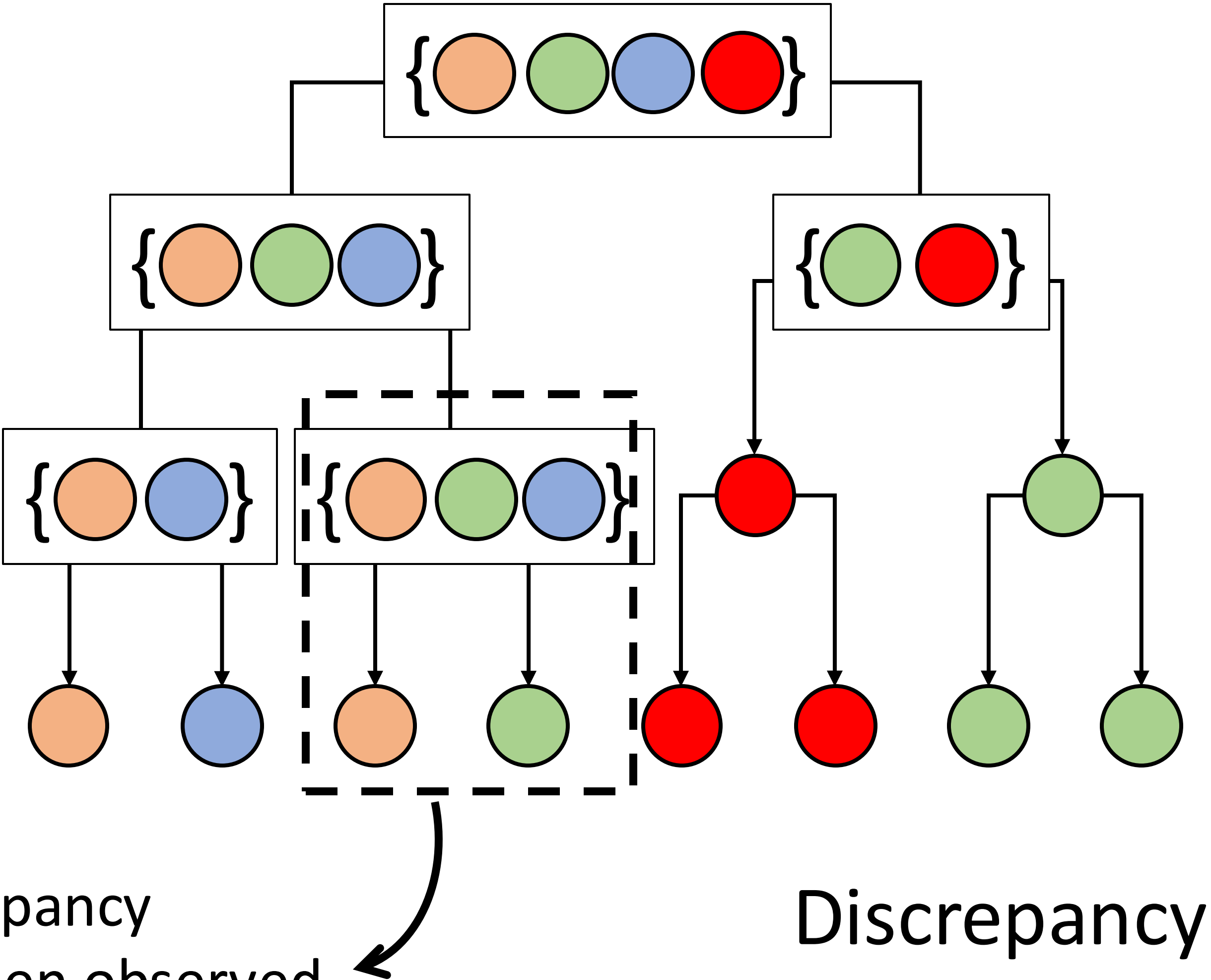
Cell differentiation map labels ancestors in cell lineage tree



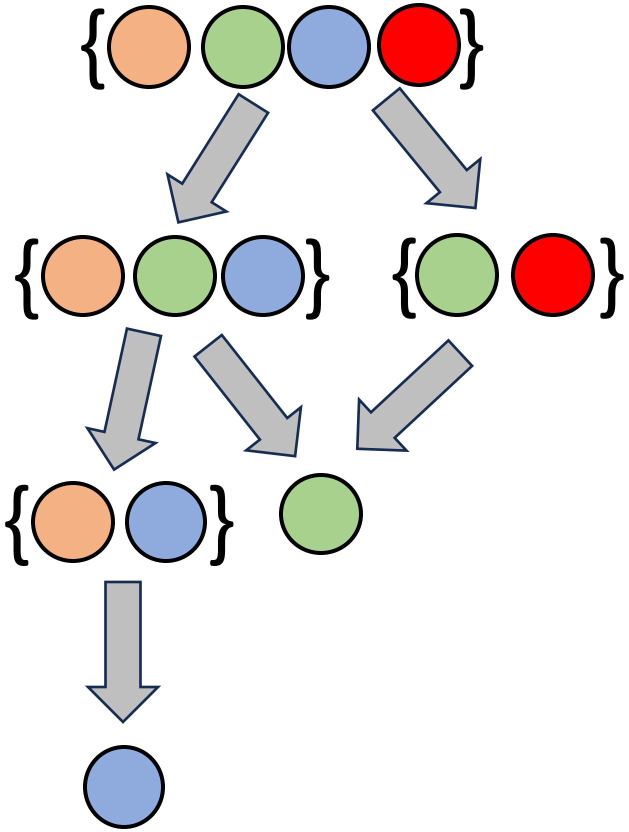


What is the mapping that best fits the data?

Cell differentiation map labels ancestors in cell lineage tree



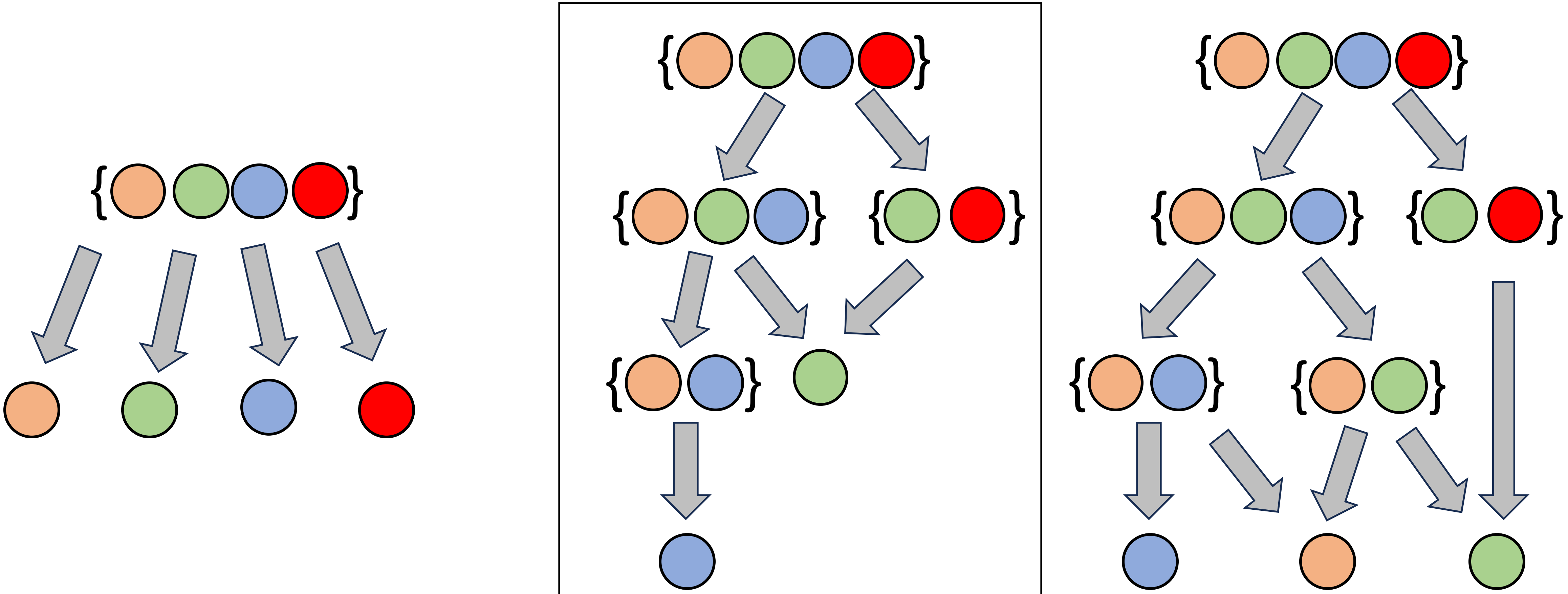
Discrepancy between observed potency and labeling



Discrepancy $D = 1$

#Progenitors $k = 4$

Characterization of progenitors and cell differentiation map



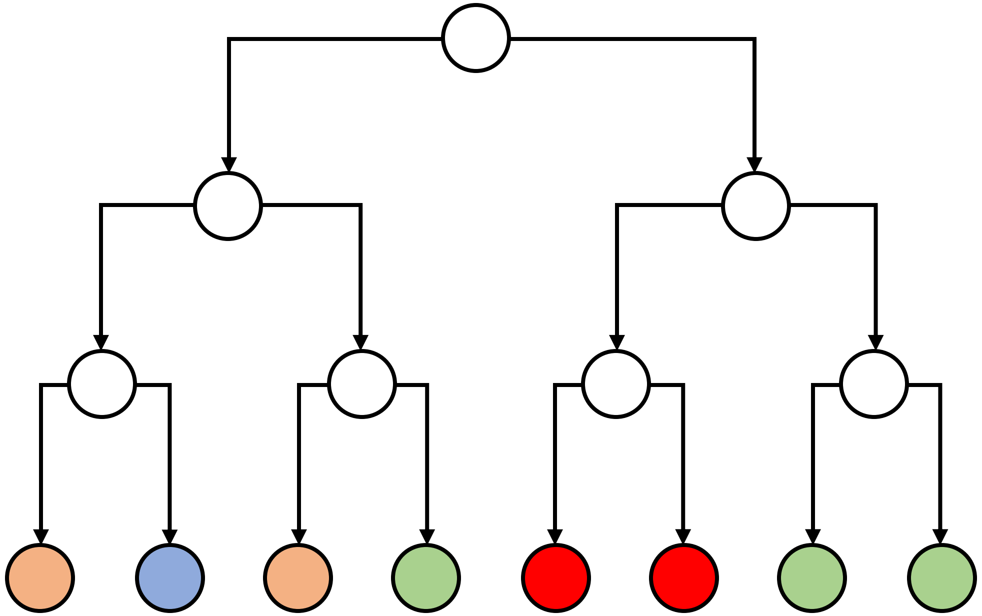
Discrepancy $D = 7$
#Progenitors $k = 1$

Discrepancy $D = 1$
#Progenitors $k = 4$

Discrepancy $D = 0$
#Progenitors $k = 5$

Cell differentiation mapping problem

Input



Leaf labeled cell lineage tree T

n cells, m cell types

Typically, $n \gg m$

Cell Differentiation Mapping (CDM) [**Sashittal et al., 2025**]
Given a leaf labeled cell lineage tree T and integer k, find a cell differentiation map F with k progenitors that minimizes discrepancy $D(T, F)$.

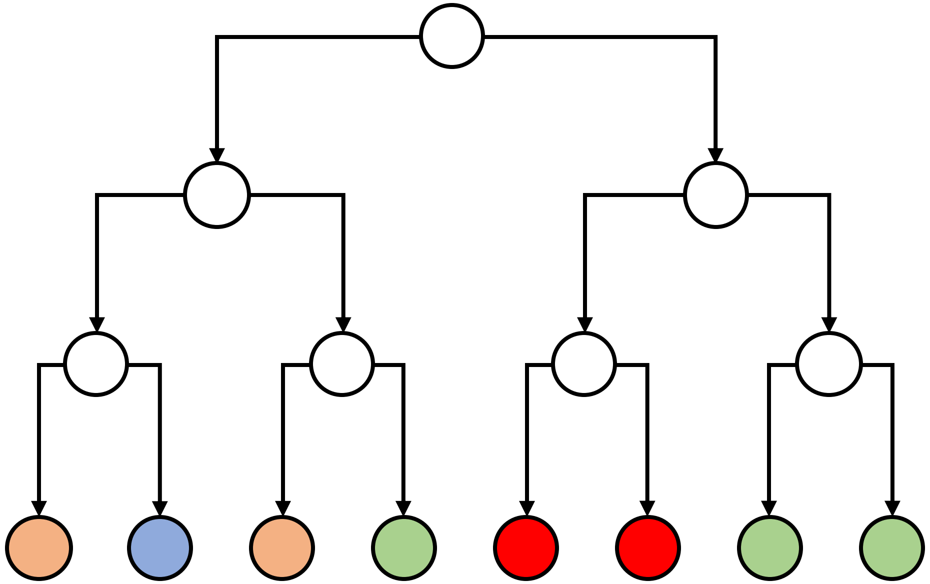
Theorem [Sashittal et al., 2025]:
Decision version of CDM Problem is NP-hard.

Theorem [Sashittal et al., 2025]:
Counting sets of k progenitors with minimum discrepancy is #P-hard

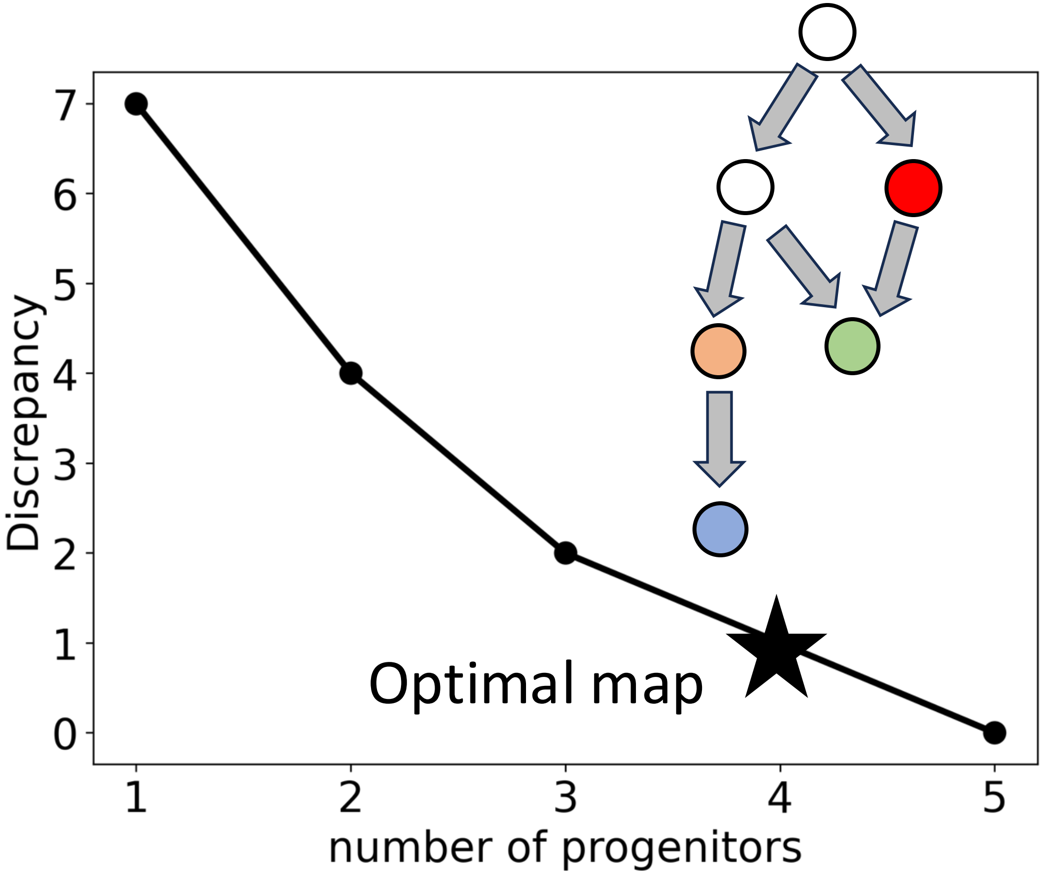
Reduction from *Vertex Cover Problem*

Theorem [Sashittal et al., 2025]:
Cell differentiation tree problem is fixed parameter tractable (FPT) in the number m of cell types.

CARTA reveals the trade-off between discrepancy and the number of progenitors



Leaf labeled cell lineage tree(s)



Richard Zhang

Michelle Chan

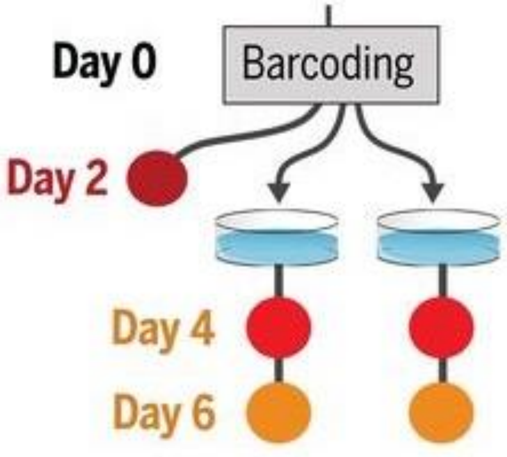
Benjamin Raphael

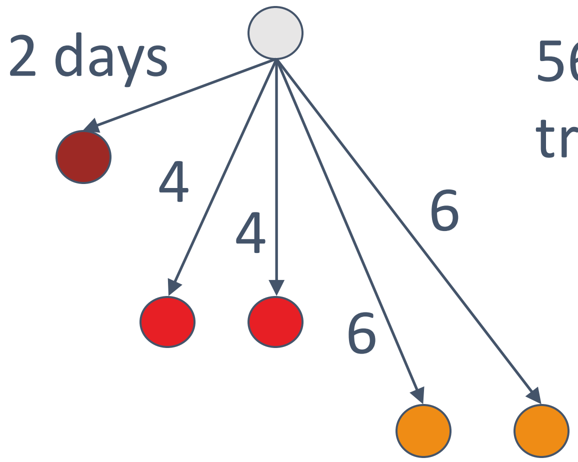
We provide a systematic way to test the number of progenitors in the cell differentiation map

Sashittal*, Zhang*, et al. RECOMB 2025; Nature Methods 2025

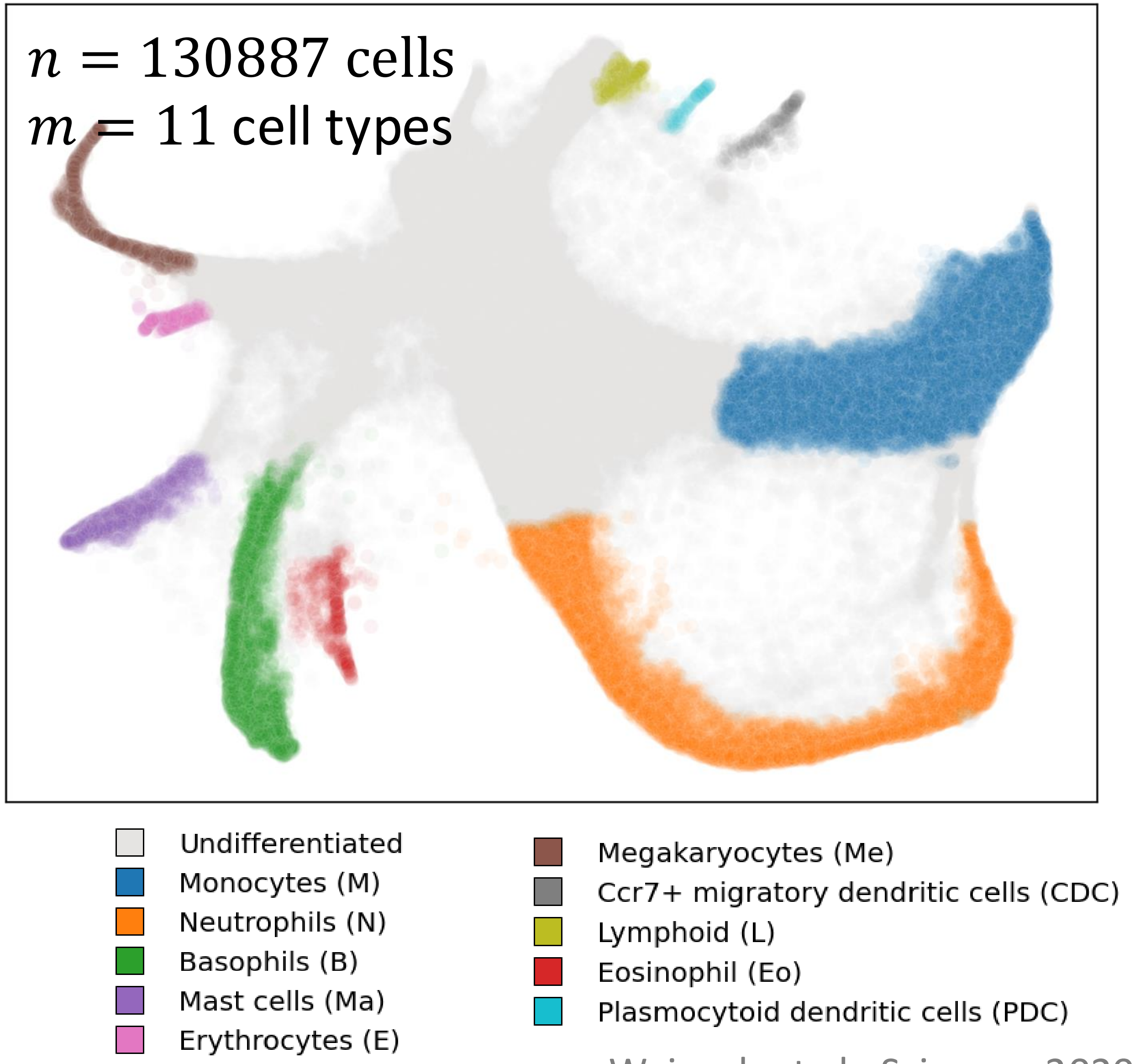
Mapping differentiation in mouse hematopoiesis

Mouse hematopoietic progenitor cells





5624 star lineage trees

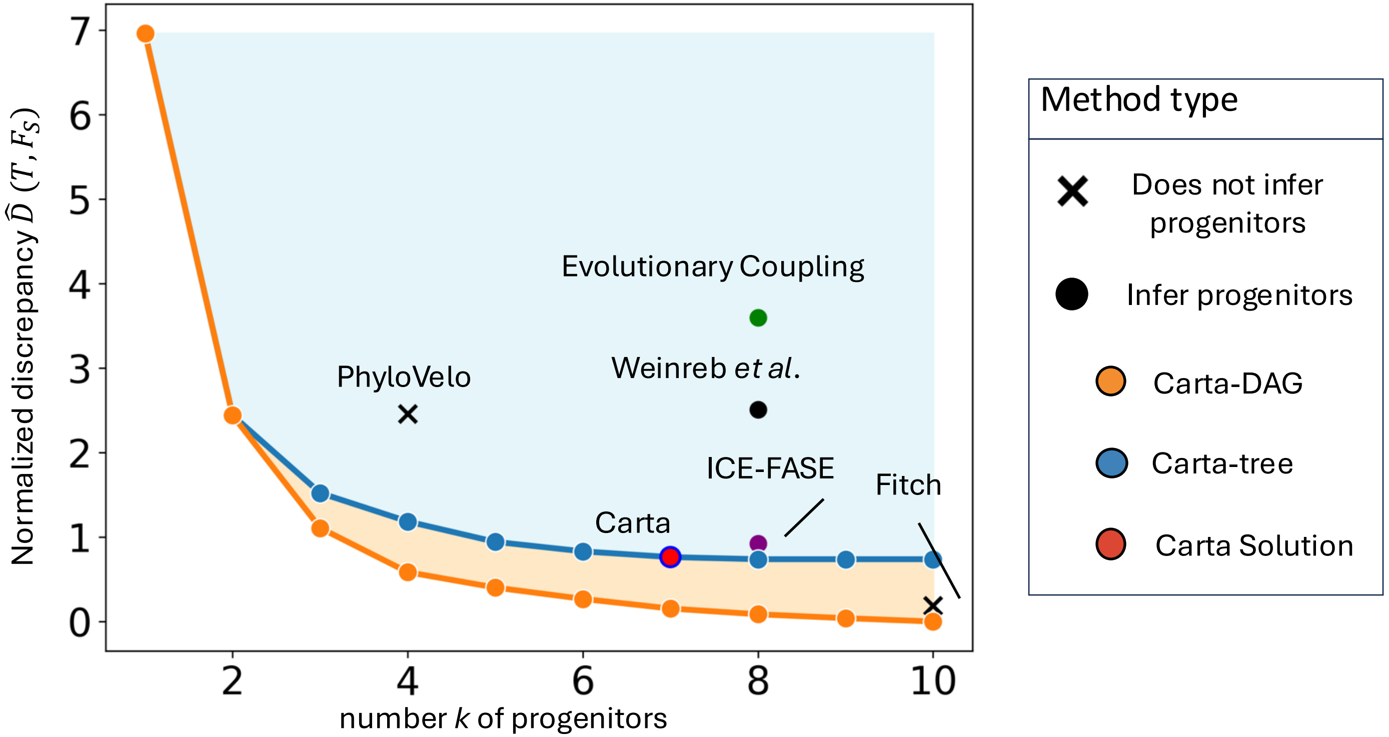


$n = 130887$ cells

$m = 11$ cell types

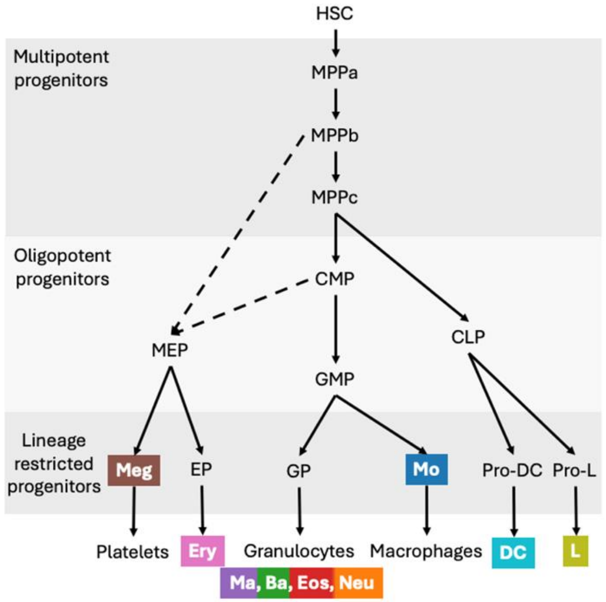
Weinreb et al., Science, 2020

Carta obtains more accurate cell differentiation map



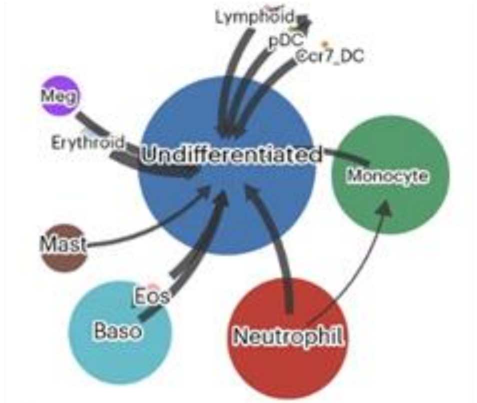
Hematopoietic differentiation map

(Seita and Weissman, 2010)



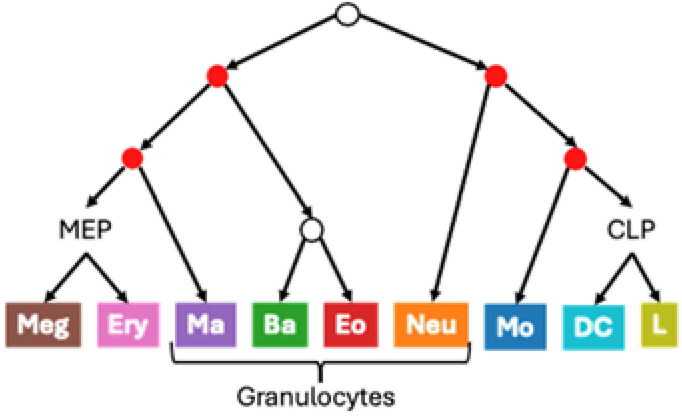
PhyloVelo

[Wang et al., Nature Biotech. 2023]

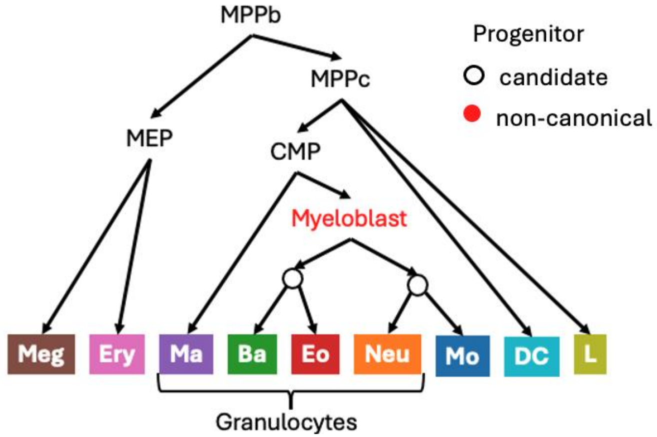


Weinreb et al.

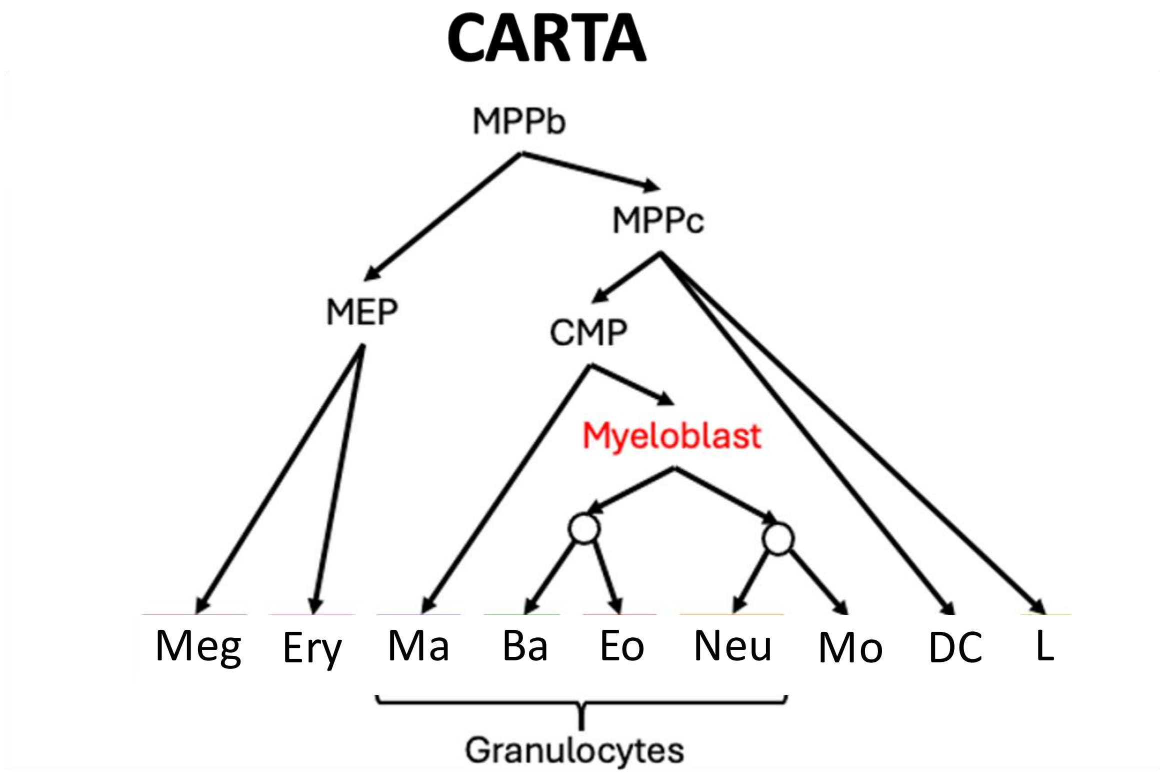
[Science 2020]

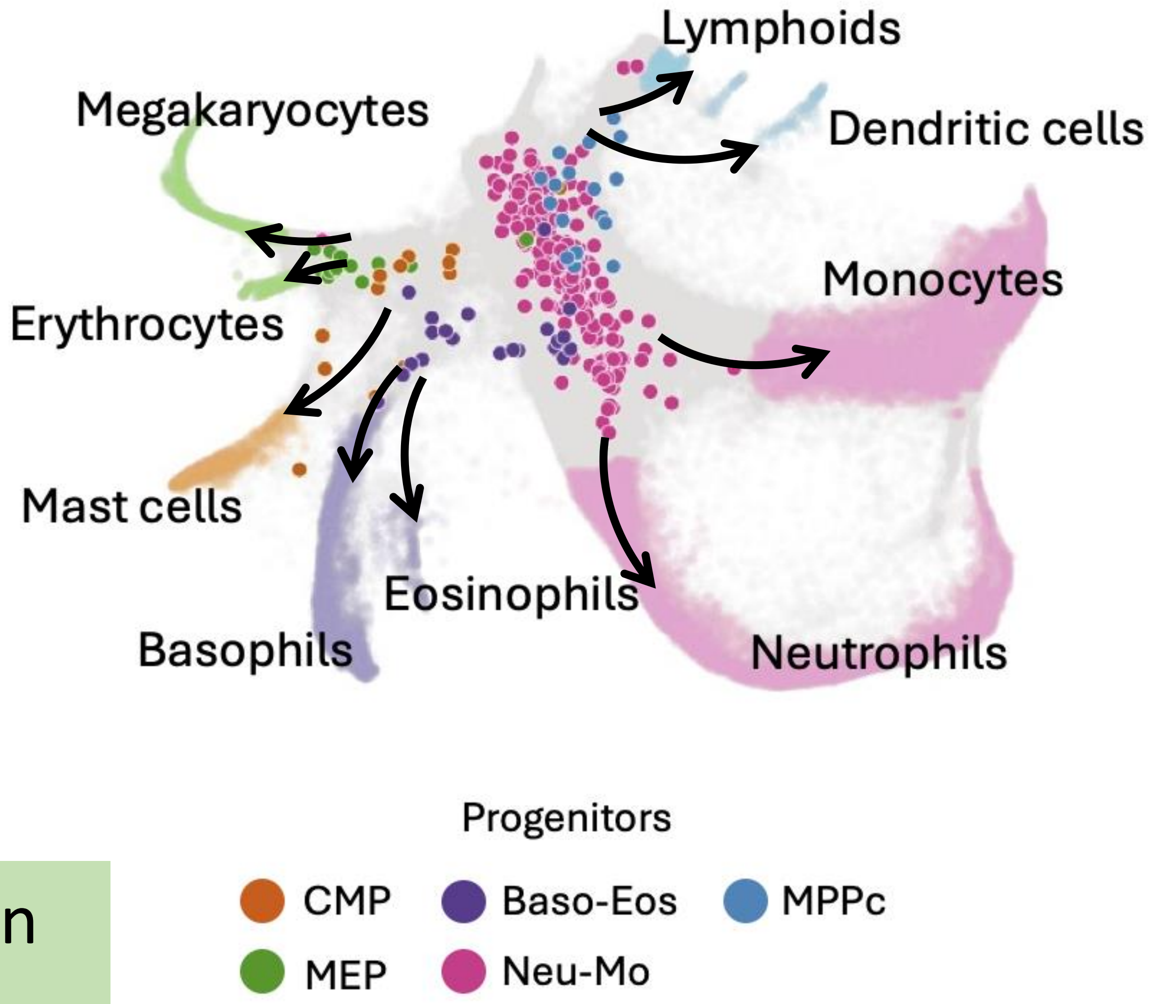


CARTA



Carta predicted cell fates align with gene expression





Good agreement of gene expression with potency inferred by CARTA

(1) Cell lineage tracing using Startle

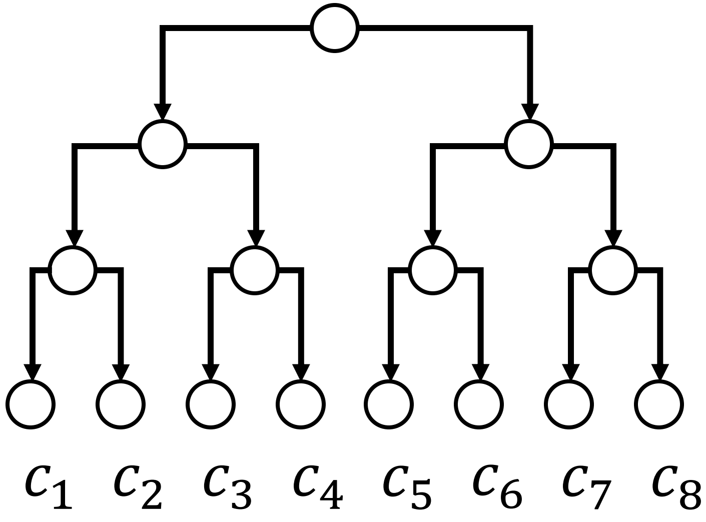
Paper

Code

Sashittal*, Schmidt* et al., *Cell Systems,* 2023
Also accepted at RECOMB 2023

$$\begin{array}{c|ccccc} & \blacktriangle & \bigcirc & \bigstar & \blacksquare & \text{⚡} \\ \hline c_1 & 1 & 0 & 1 & 0 & 0 \\ c_2 & 1 & 0 & 0 & 0 & 0 \\ c_3 & 1 & 1 & 0 & 1 & 0 \\ c_4 & 1 & 1 & 0 & 0 & 0 \\ c_5 & 0 & 0 & 0 & 1 & 0 \\ c_6 & 0 & 0 & 0 & 1 & 1 \\ c_7 & 0 & 0 & 1 & 0 & 1 \\ c_8 & 0 & 0 & 1 & 0 & 0 \end{array}$$

Character matrix M



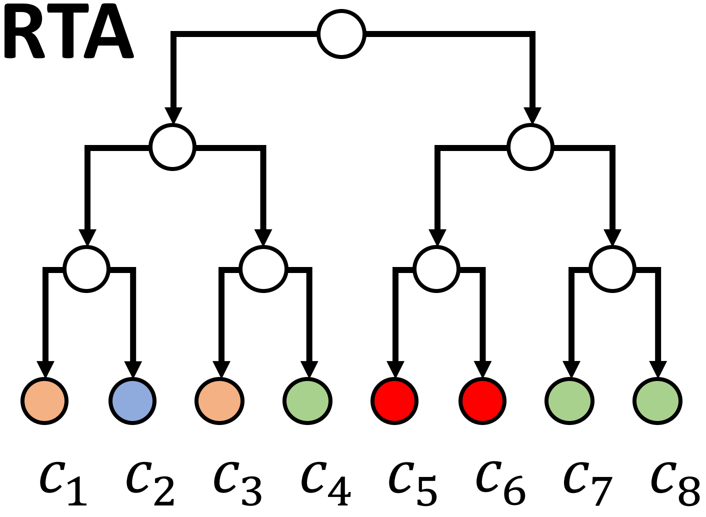
cell lineage tree T

(2) Differentiation mapping using CARTA

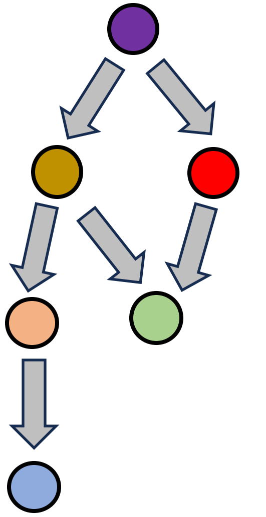
Paper

Code

Sashittal*, Zhang* et al., *Nature Methods,* 2025
Also accepted at RECOMB 2025



cell lineage tree T



Differentiation map F

BACKUP