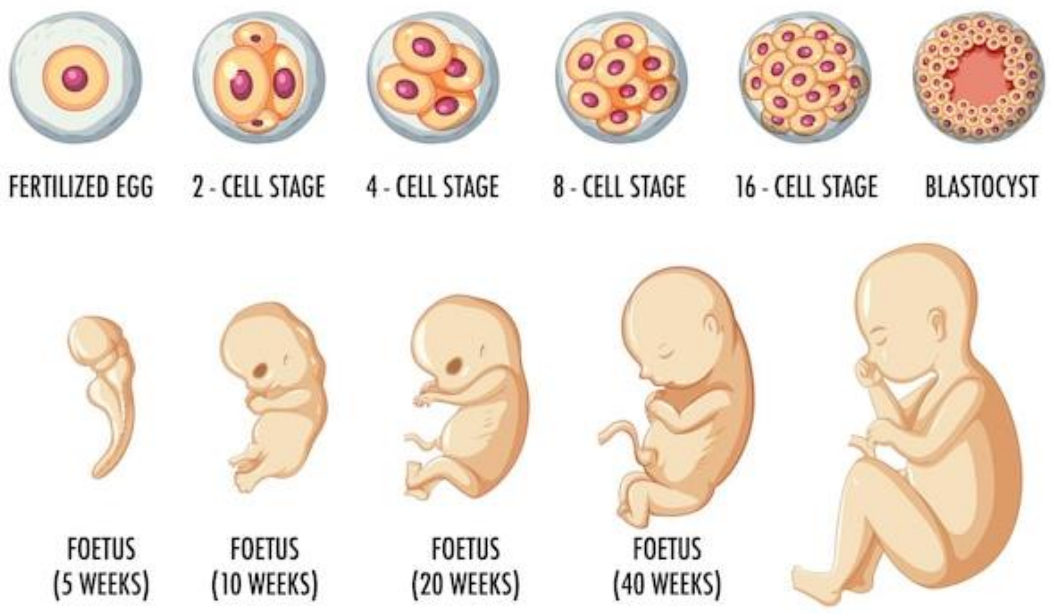
# Startle: a star homoplasy approach for CRISPR-Cas9 lineage tracing

Palash Sashittal*, Henri Schmidt*, Michelle Chan, Ben Raphael

PRINCETON UNIVERSITY

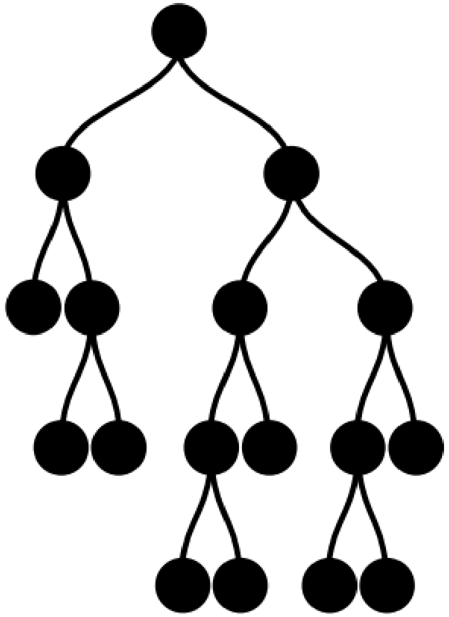
# Biological developmental processes

1 cell



Embryonic development

$\approx 26 \times 10^9$ cells



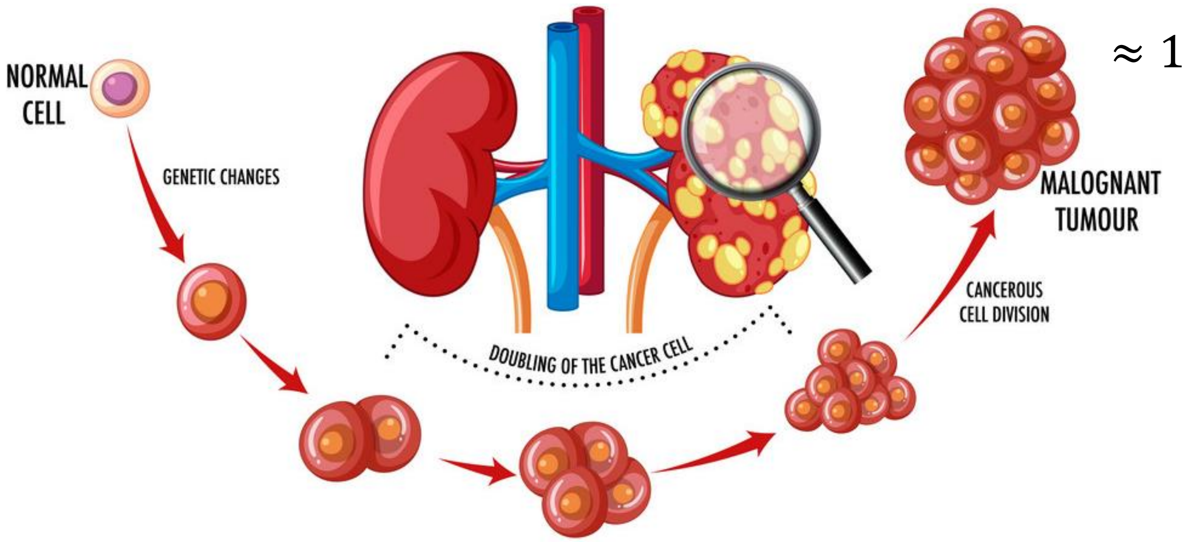
Cell lineage tree

What is the history of cell divisions during the developmental process?

Figures from vectorstock.com and freepik.com

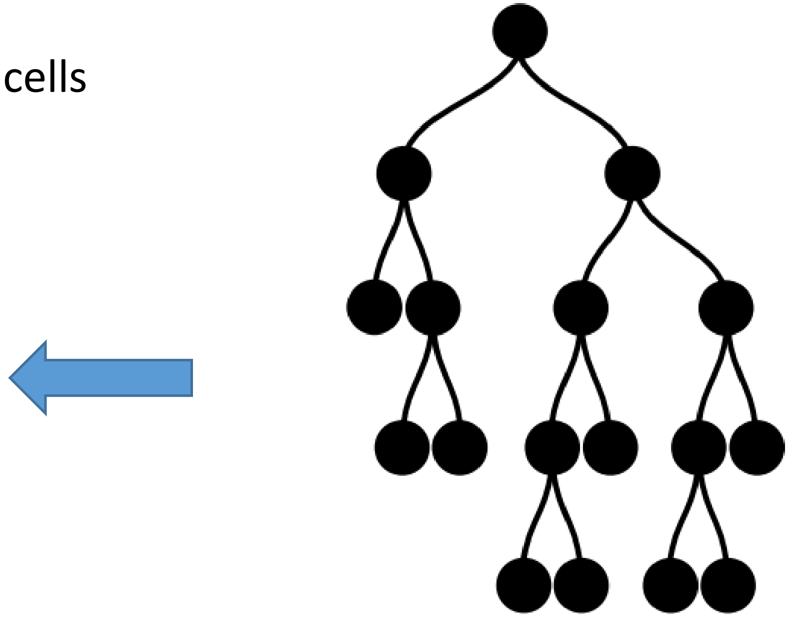
# Biological developmental processes

1 cell



$\approx 10^6$ cells

Cancer development



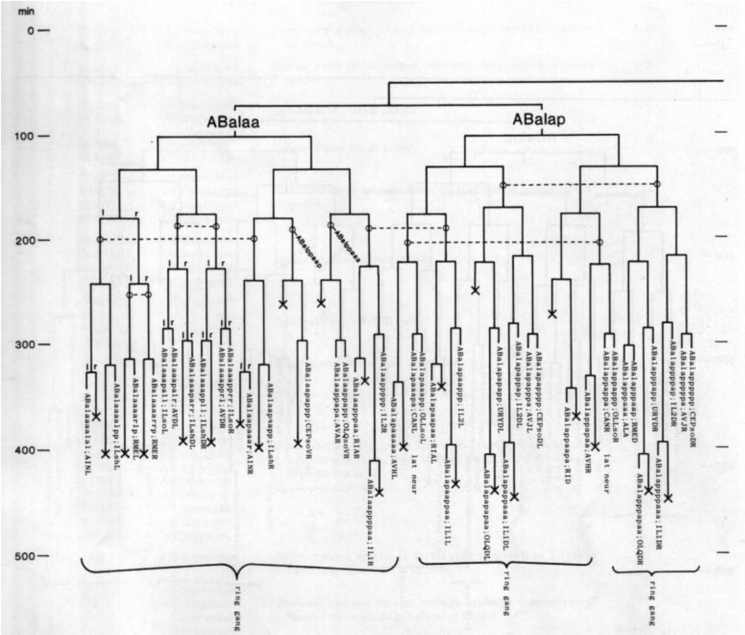
Cell lineage tree

What is the history of cell divisions during the developmental process?

Figures from vectorstock.com and freepik.com

# Lineage Tracing: introduction and motivation

**Direct experimental observation**



**2002 Nobel Prize in Physiology or Medicine**
**Sydney Brenner**, **H. Robert Horvitz** and **John E. Sulston**

*Caenorhabditis elegans* $= 959$ cells

Every cell division and developmental fate of every cell has been mapped

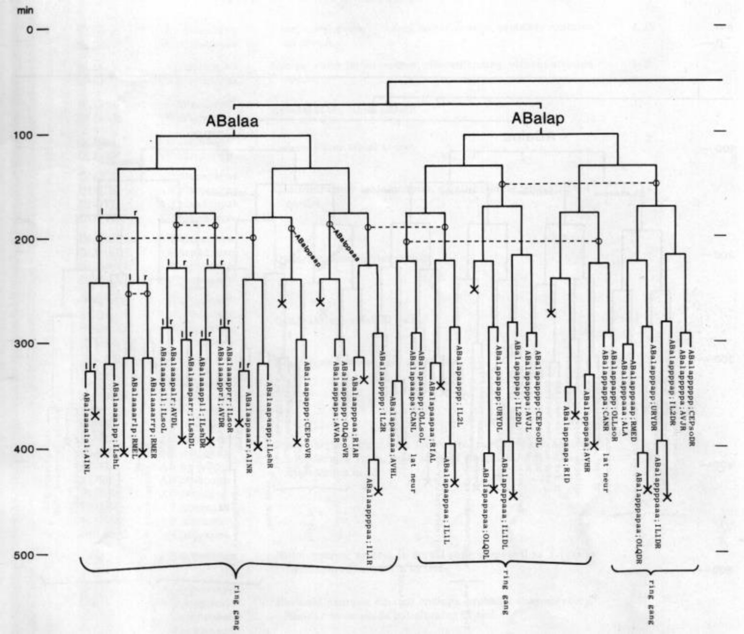
- Identification of progenitor cells

- Discovery and characterization of key genes controlling programmed cell death and organ development

Figure adapted from Sulston et al., 1983, Developmental Biology; McKenna et al., 2019, Development

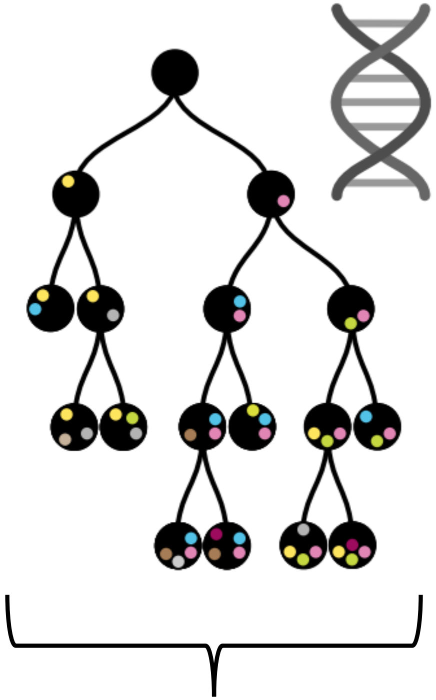
# Lineage Tracing: introduction and motivation

**Direct experimental observation**



**2002 Nobel Prize in Physiology or Medicine**
**Sydney Brenner**, **H. Robert Horvitz** and **John E. Sulston**

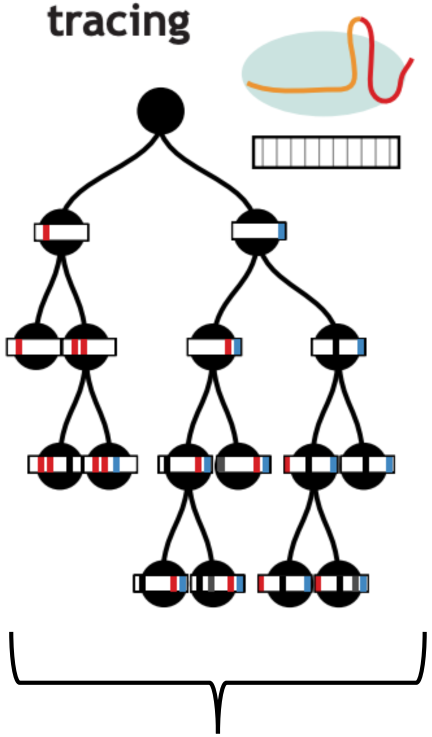
**Somatic mutations**



Retrospective lineage tracing

Carlson et al. 2012, Nat. Methods; Behjati et al. 2014, Nature; Lodato et al. Science, 2015 and many more

**Dynamic lineage tracing**

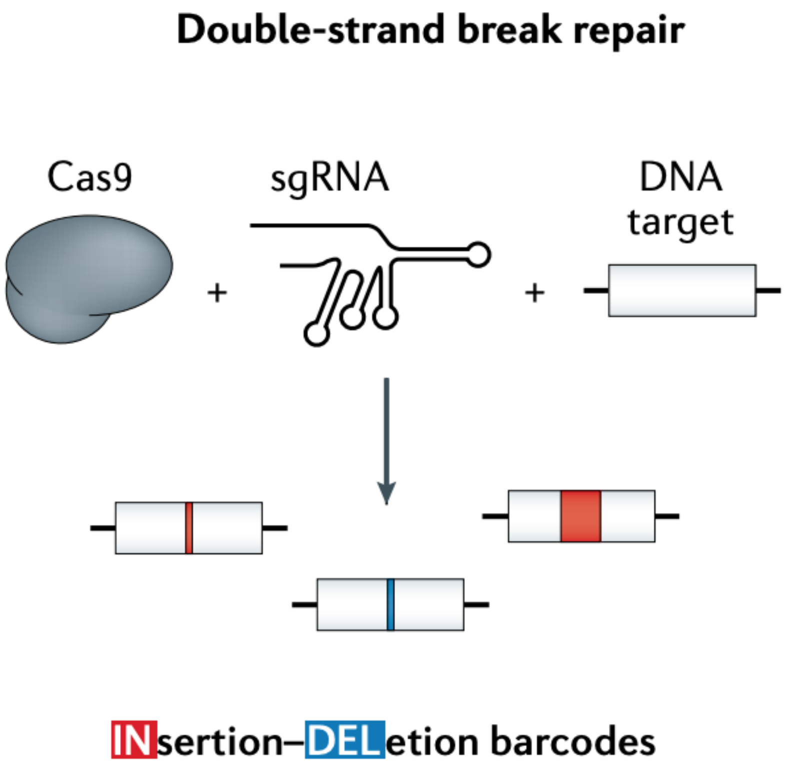


Introspective lineage tracing

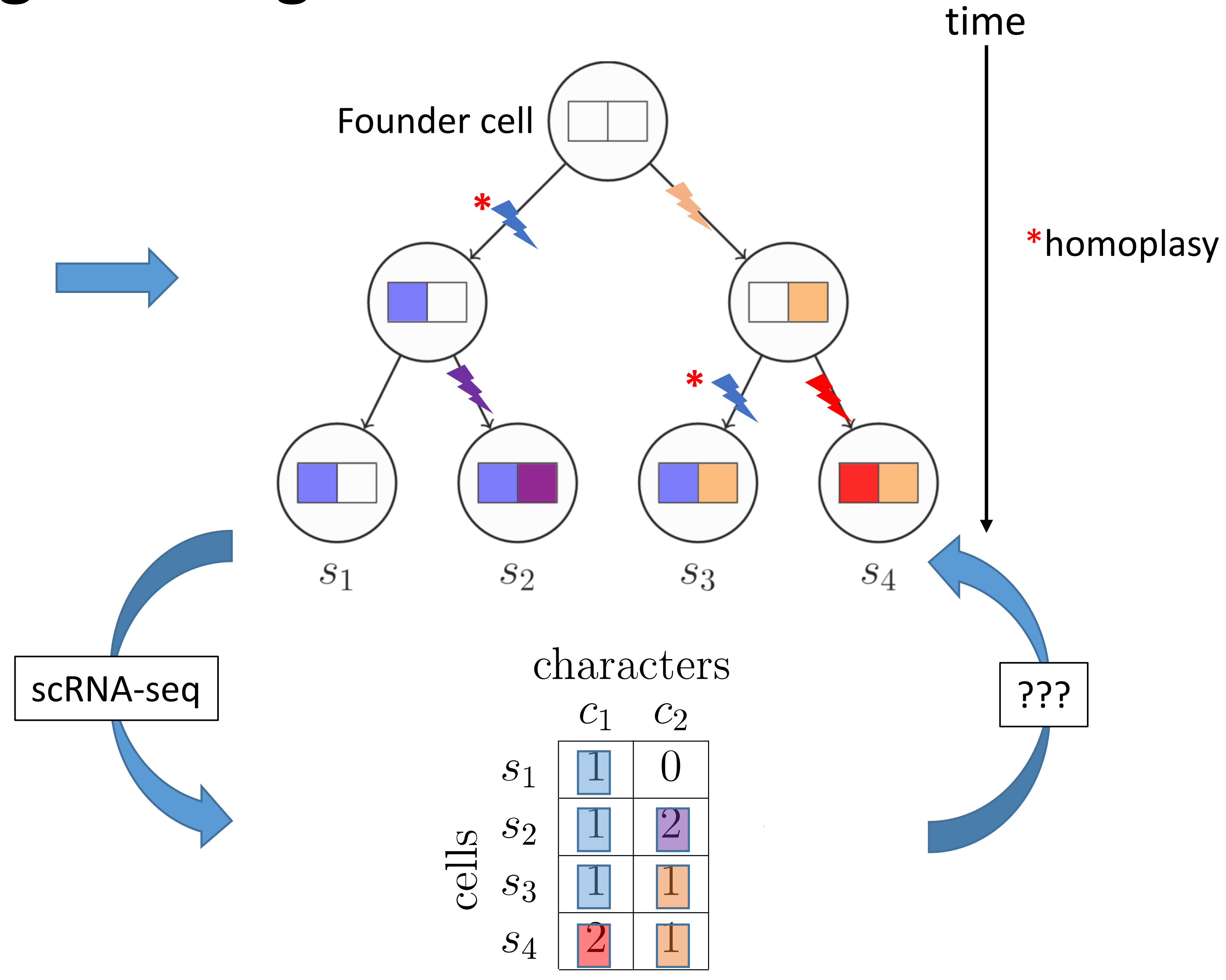
McKenna et al. 2016, Science; Alemany et al. 2018, Nature; Chan et al. Nature, 2019 and many more

Figure adapted from Sulston et al., 1983, Developmental Biology; McKenna et al., 2019, Development

# CRISPR-Cas9 based lineage tracing



- ✓ Heritable
- ✓ Irreversible
- ✓ Non-modifiable



| cells | $c_1$ | $c_2$ |
|---|---|---|
| $s_1$ | 1 | 0 |
| $s_2$ | 1 | 2 |
| $s_3$ | 1 | 1 |
| $s_4$ | 2 | 1 |

Character matrix

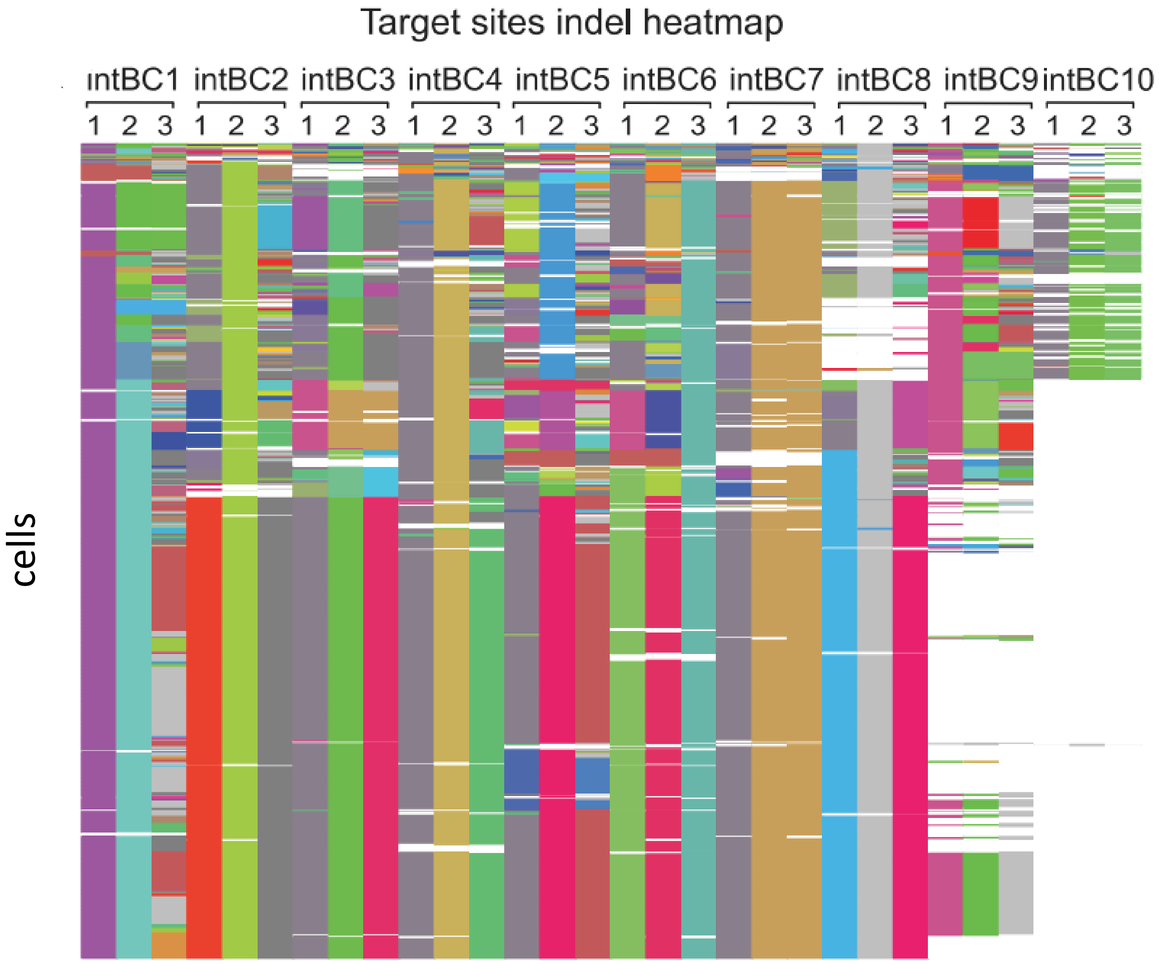
# CRISPR-Cas9 based lineage tracing data

**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)

Standard phylogenetic methods not suited for this data

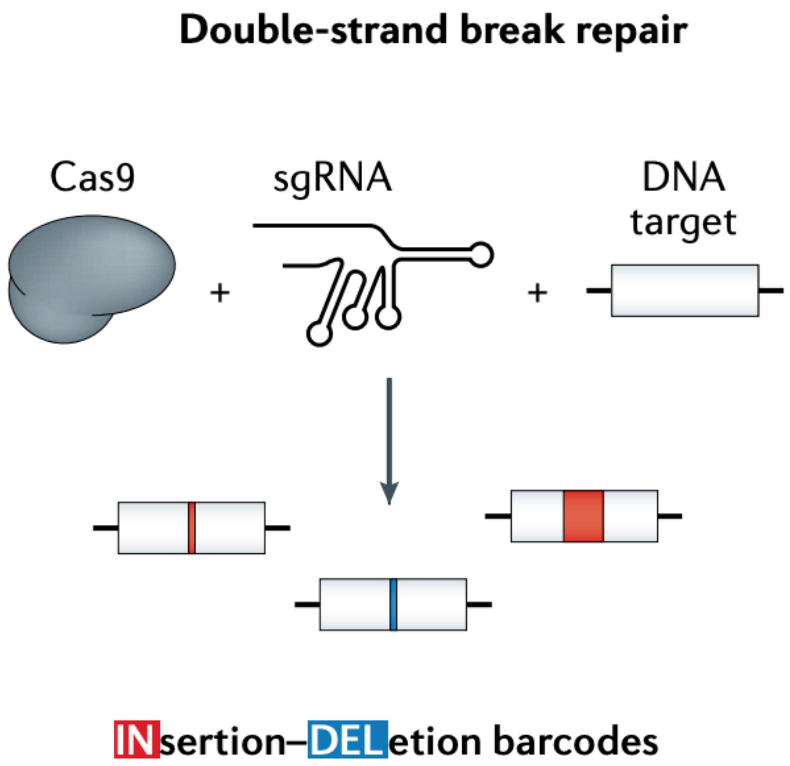
Specialized methods have been introduced and benchmarked in a DREAM challenge (Gong et al., 2021, Cell Systems)



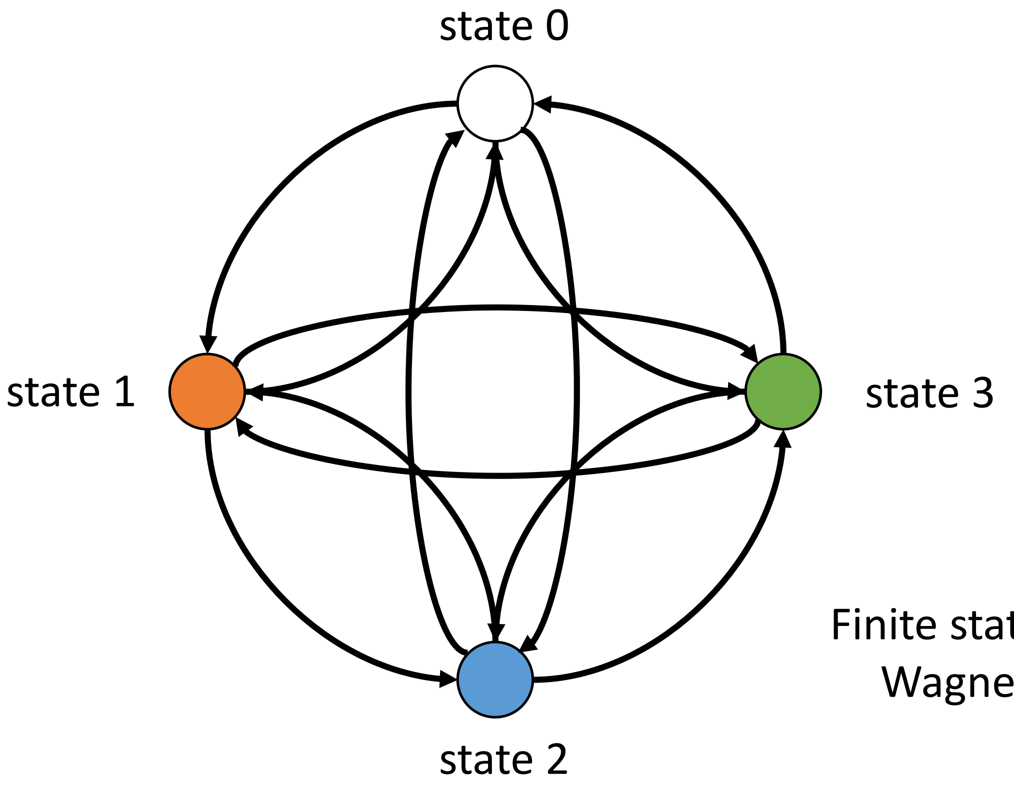
What is an appropriate evolutionary model that captures the characteristic features of CRISPR-Cas9 mutations?

Figure adapted from Yang et al., 2022, Cell

# Evolutionary models for CRISPR-Cas9 based lineage tracing



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

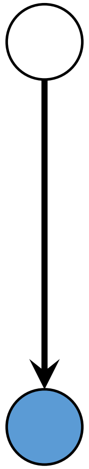


State transition graph

(Swofford et al., 1992)

Finite states model
Wagner, 1961

# Specialized evolutionary models for lineage tracing

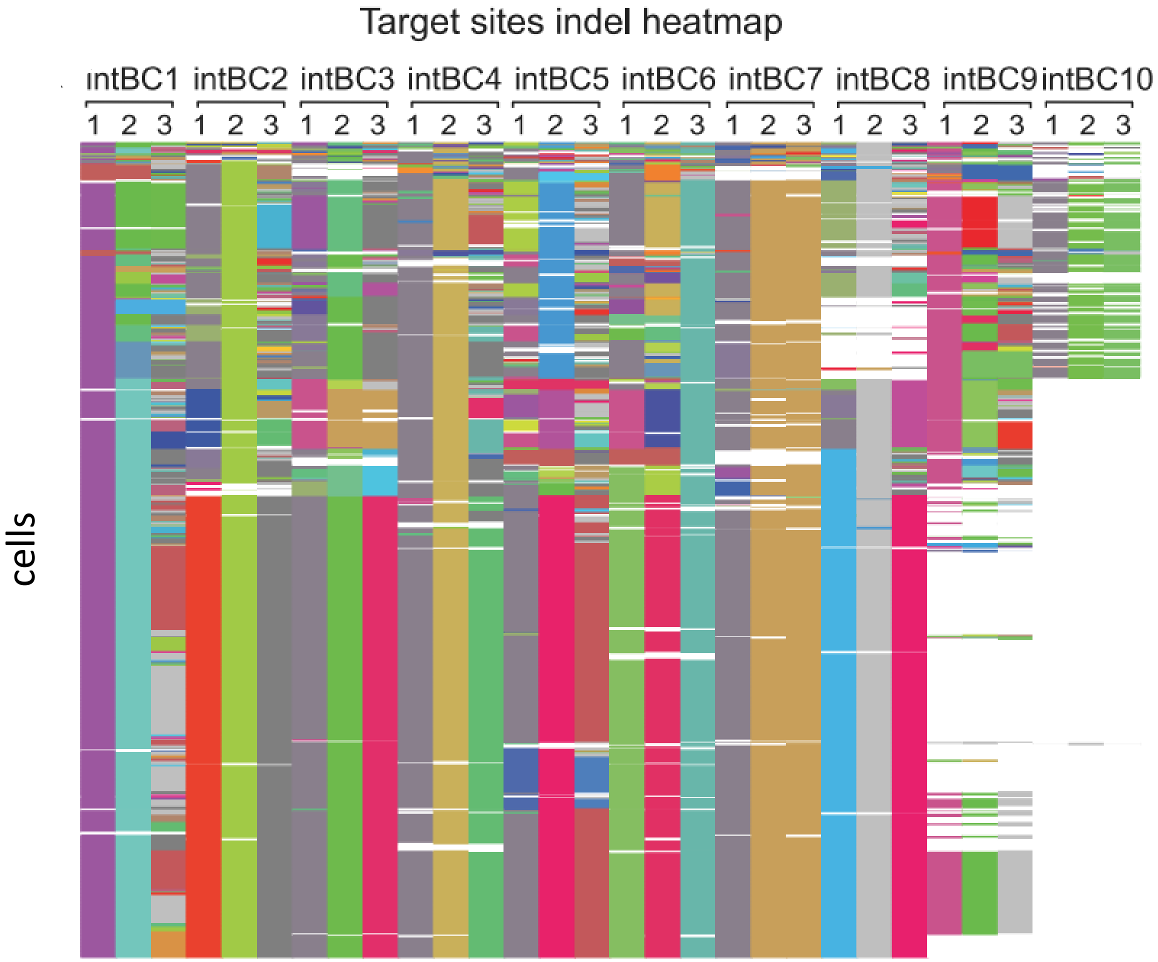


Two-state
Camin-Sokal model

(Camin et al., 1965, Evolution)

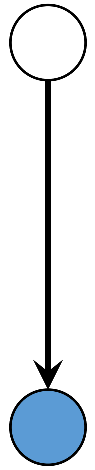
- ✗ Multi-state
- ✓ Irreversible
- *✓ Non-modifiable

McKenna et al., 2016, Science
Raj et al., 2018, Nature Biotechnology



Lineage tracing data from Yang et al., 2022, Cell

# Specialized evolutionary models for lineage tracing



Two-state
Camin-Sokal model

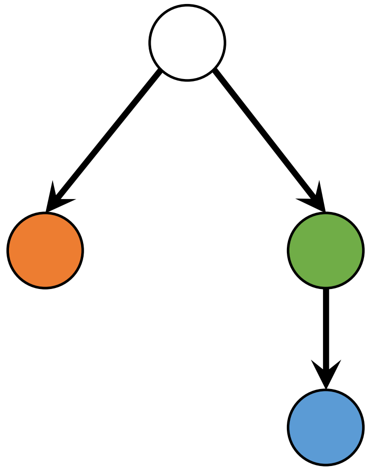
(Camin et al., 1965, Evolution)

✗ Multi-state

✓ Irreversible

*✓ Non-modifiable

McKenna et al., 2016, Science
Raj et al., 2018, Nature Biotechnology



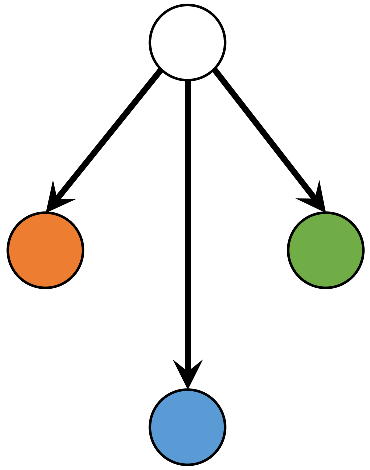
Multi-state
Camin-Sokal model

(Felsentein et al., 2004)

✓ Multi-state

✓ Irreversible

✗ Non-modifiable



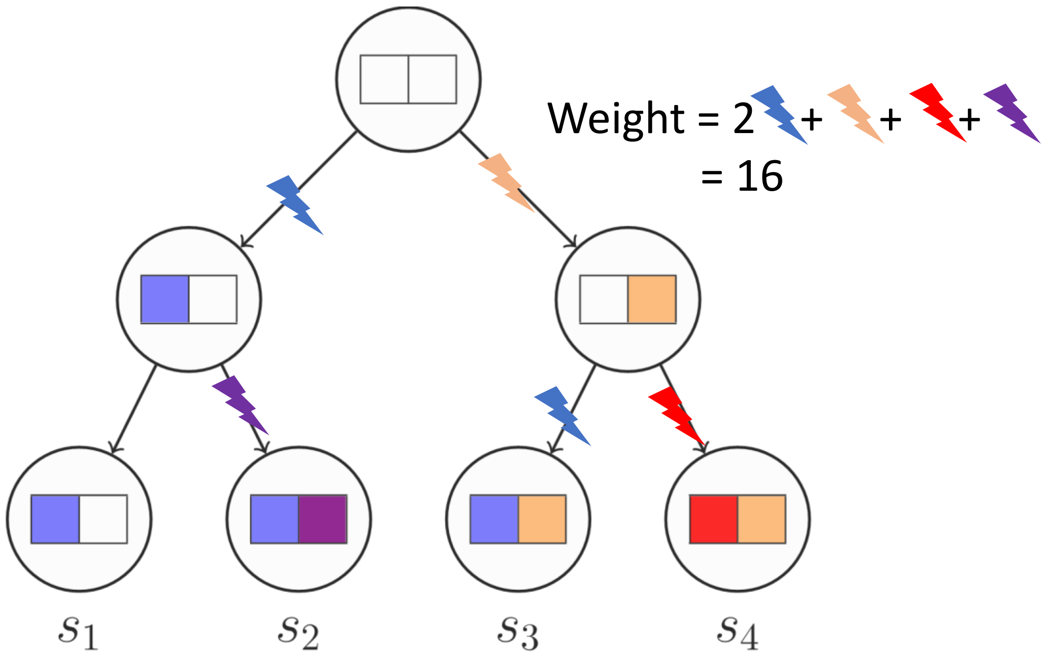
Multi-state
**Star homoplasy model**
(This work)

✓ Multi-state

✓ Irreversible

✓ Non-modifiable

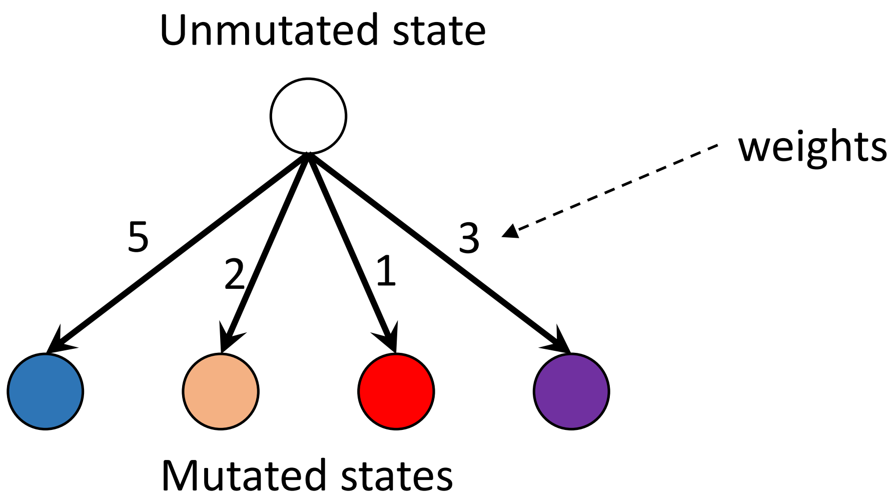
# Startle*: maximum parsimony for star homoplasy model



characters

| cells | $c_1$ | $c_2$ |
|---|---|---|
| $s_1$ | 1 | 0 |
| $s_2$ | 1 | 2 |
| $s_3$ | 1 | 1 |
| $s_4$ | 2 | 1 |

Character matrix

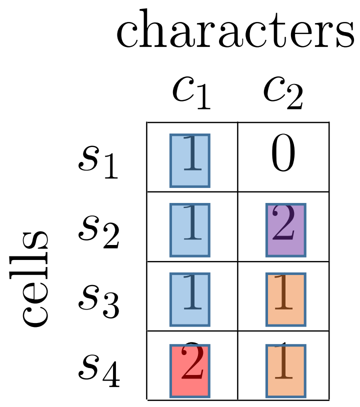


**Star homoplasy model**:

- Each character can **change state at most once** in a lineage (a path from root to leaf)
- Characters evolve **independently** (standard assumption)

*Star tree lineage exploration: **maximum parsimony** methods using the star homoplasy model

# Maximum parsimony problem for the star homoplasy model

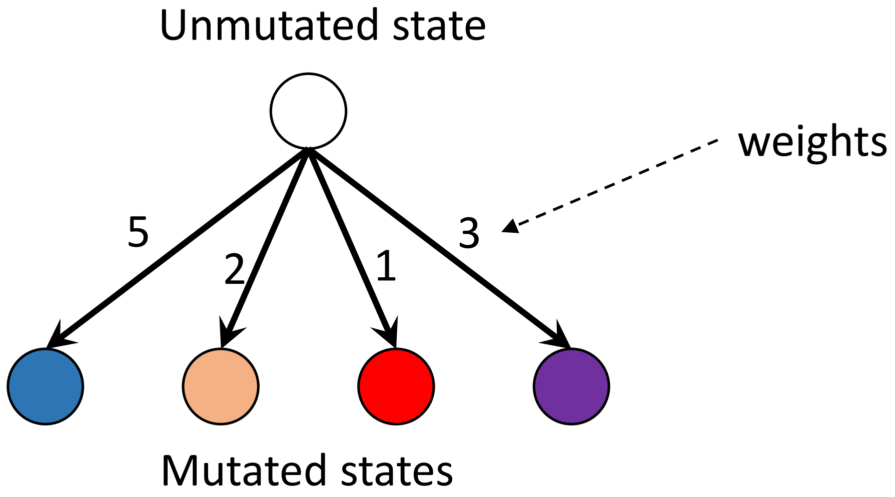


characters

cells

|  | $c_1$ | $c_2$ |
|---|---|---|
| $s_1$ | 1 | 0 |
| $s_2$ | 1 | 2 |
| $s_3$ | 1 | 1 |
| $s_4$ | 2 | 1 |

Character matrix



**Star homoplasy model**:

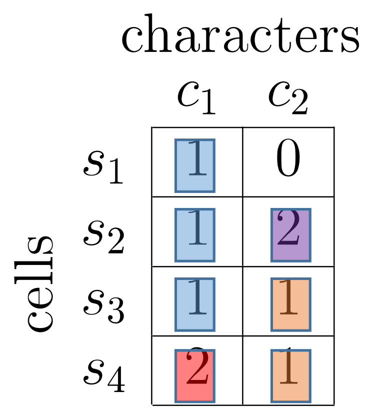
- Each character can **change state at most once** in a lineage (a path from root to leaf)
- Characters evolve **independently** (standard assumption)

**Input**: Character matrix and mutation weights.

**Problem**: Find the star homoplasy phylogeny such such that the total weight is minimized.

**Startle-NNI**: nearest neighbor interchanges to perform hill climbing in tree space and find the most parsimony star homoplasy phylogeny

# Bounded homoplasy version: k-star homoplasy model

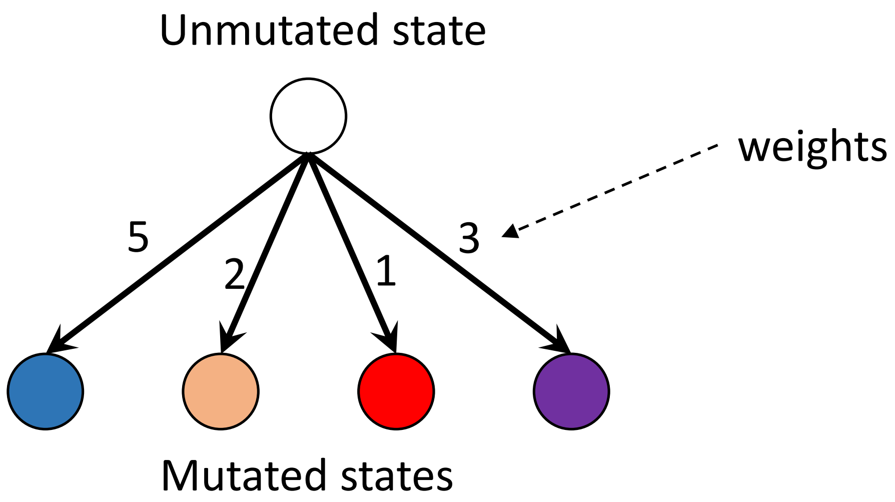


characters

cells

$$\begin{array}{c|cc} & c_1 & c_2 \\ \hline s_1 & 1 & 0 \\ s_2 & 1 & 2 \\ s_3 & 1 & 1 \\ s_4 & 2 & 1 \end{array}$$

Character matrix



**Input**: Character matrix and mutation weights.

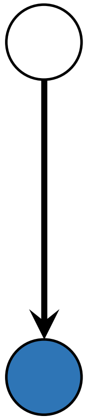
**Problem**: Find the k-star homoplasy phylogeny such that the total weight is minimized.

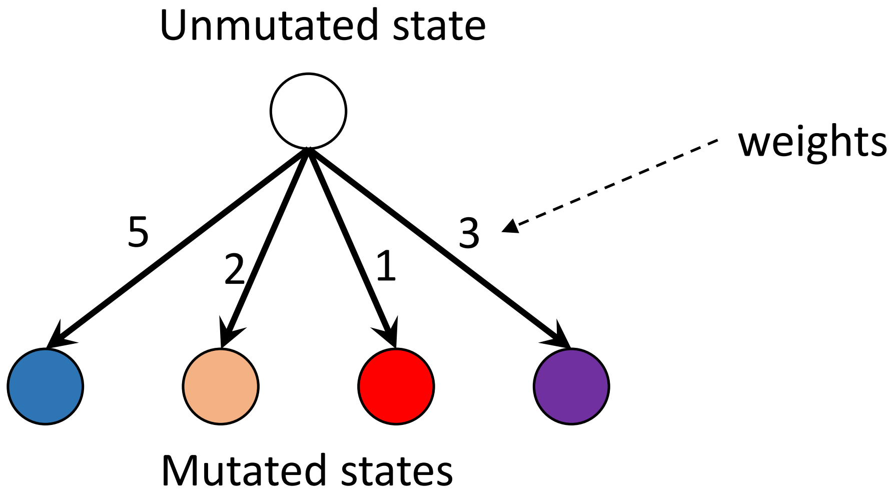
**k-Star homoplasy model**:

- Each character can **change state at most once** in a lineage (a path from root to leaf)
- Characters evolve **independently** (standard assumption)
- Each mutation can occur at most k times in the phylogeny

Characterize all character matrices that admit a k-star homoplasy phylogeny by leveraging a connection between **k-star homoplasy** and **two-state perfect phylogeny** models

# Bounded homoplasy version: k-star homoplasy model





**Two-state perfect phylogeny model**:

- Each character can **change state at most once** in the phylogeny
- Characters evolve **independently** (standard assumption)

Kimura, 1969, Genetics
Gusfield, 1991, Networks

**k-Star homoplasy model**:

- Each character can **change state at most once** in a lineage (a path from root to leaf)
- Characters evolve **independently** (standard assumption)
- Each mutation can occur at most k times in the phylogeny

Characterize all character matrices that admit a k-star homoplasy phylogeny by leveraging a connection between **k-star homoplasy** and **two-state perfect phylogeny** models

# Startle-ILP algorithm for k-star homoplasy phylogeny inference

characters

cells

|  | $c_1$ | $c_2$ |
|---|---|---|
| $s_1$ | 1 | 0 |
| $s_2$ | 1 | 2 |
| $s_3$ | 1 | 1 |
| $s_4$ | 2 | 1 |

? →

|  | $(c_1,1,0)$ | $(c_1,1,1)$ | $(c_1,2,0)$ | $(c_1,2,1)$ | $(c_2,1,0)$ | $(c_2,1,1)$ | $(c_2,2,0)$ | $(c_2,2,1)$ |
|---|---|---|---|---|---|---|---|---|
| $s_1$ | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $s_2$ | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| $s_3$ | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| $s_4$ | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |

Character matrix

$n \times m$

k-star binarization of the character matrix

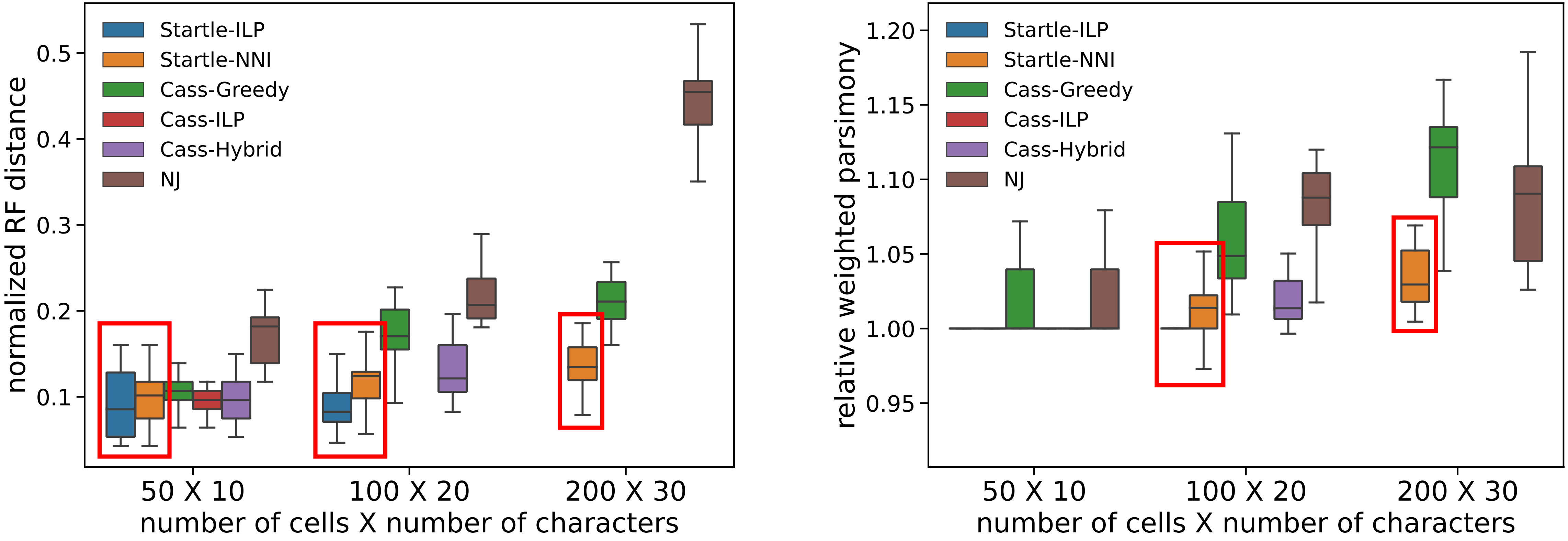
$n \times mrk$

A character matrix $A$ admits a **k-star homoplasy phylogeny** if and only if there exists a **k-star binarization** of $A$ that admits a **two-state perfect phylogeny**

**Startle-ILP**: We formulate a MILP to find the most parsimonious k-star homoplasy phylogeny from lineage tracing data

# Startle outperforms existing methods on simulated data

Simulations with dropout rate of 15%



Cassiopeia*: parsimony-based method (Jones et al. Genome Biology, 2020)
Neighbor Joining: distance-based method (Saitou et al. MBE, 1987)

* One of the top performing methods in the DREAM challenge (Gong et al., 2021, Cell Systems)

# Mouse metastatic lung adenocarcinoma data

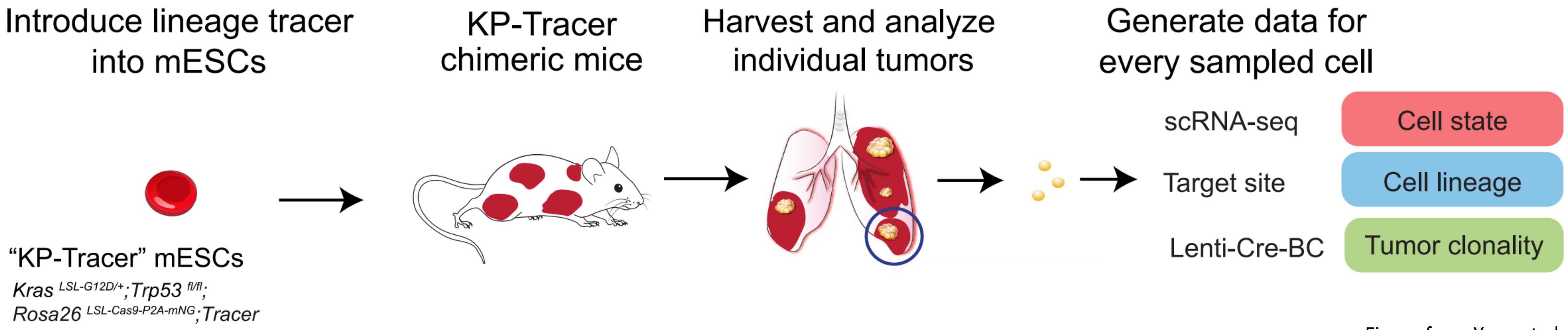

Introduce lineage tracer into mESCs
KP-Tracer chimeric mice
Harvest and analyze individual tumors
Generate data for every sampled cell
“KP-Tracer” mESCs
*Kras* $^{LSL\text{-}G12D/+}$;*Trp53* $^{fl/fl}$;
*Rosa26* $^{LSL\text{-}Cas9\text{-}P2A\text{-}mNG}$;*Tracer*
scRNA-seq
Cell state
Target site
Cell lineage
Lenti-Cre-BC
Tumor clonality

Figure from Yang et al., 2022, Cell

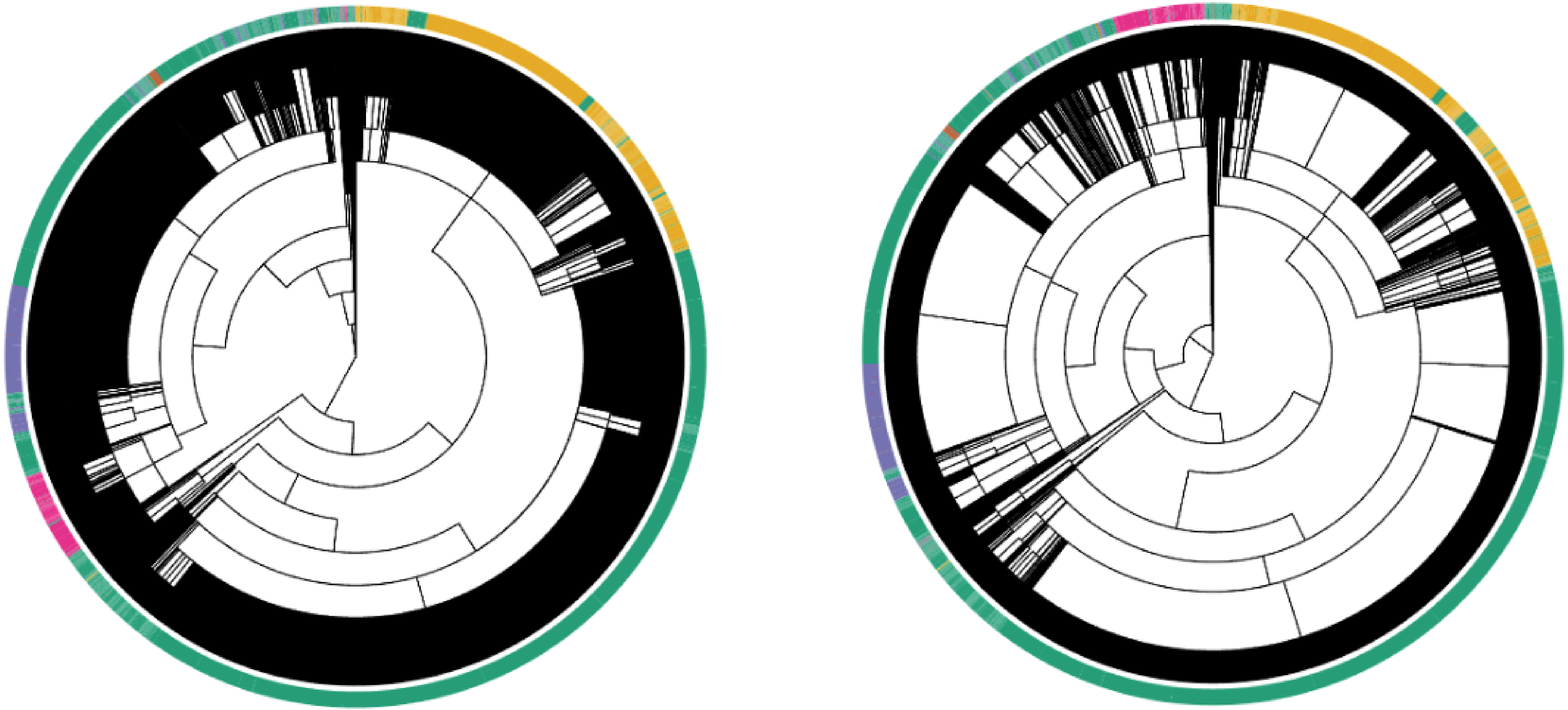
The authors used Cassiopeia (Jones et al., 2021, Genome Biology) to build lineage trees which were then used to study

- Clonal fitness and expansion
- Plasticity of tumor cells
- Migration patterns during metastasis

Largest dataset in the study (3724_NT_T1_All):

- Total cells : 21108
  - Primary (lung) tumor : 14852
  - Soft tissue metastasis tumor : 3891
  - Liver metastasis tumor 1: 90
  - Liver metastasis tumor 2: 1512
  - Liver metastasis tumor 3: 863

# Startle trees are more parsimonious than published results



Published phylogeny

Startle phylogeny

Weight = 4827.43

Weight = **4715.5**

Anatomical sites (cells)

🟩 Primary tumor (14852)

🟧 Liver met. 1 (90)

🟪 Liver met. 2 (1512)

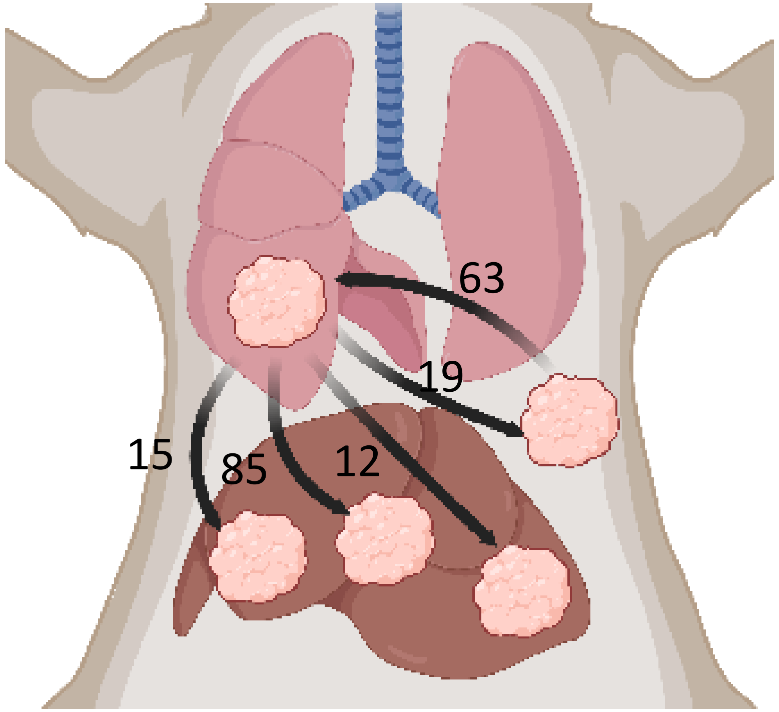
🟥 Liver met. 3 (863)

🟨 Soft tissue met. (3891)

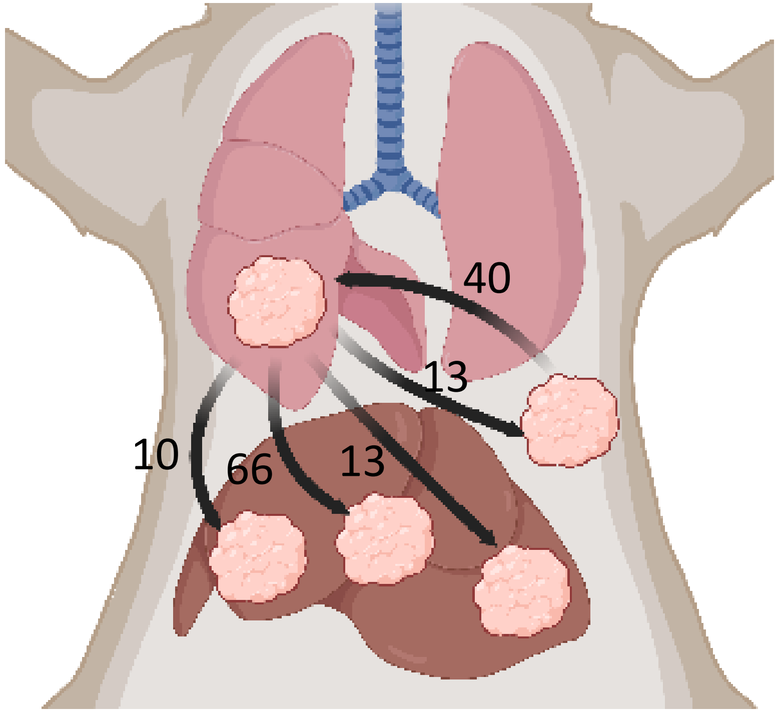
Total cells: 21108

# Startle trees have fewer migrations between anatomical sites

Migrations inferred* from published tree



Migrations inferred* from Startle tree

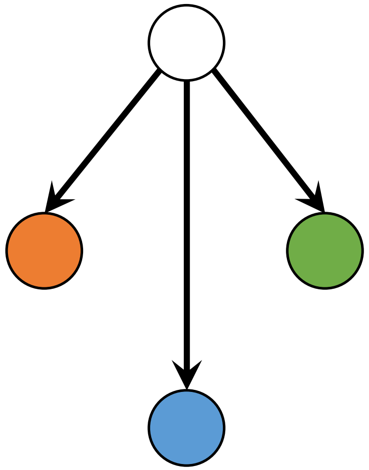


Startle tree infers the same migration pattern but with far fewer migration events compared to published results

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

# Conclusion

- We propose the **star homoplasy model** for the evolution of CRISPR-Cas9 induced mutations
- We derive a correspondence between the **k-star homoplasy model** and the **two-state perfect phylogeny**
- We developed **Startle-ILP** and **Startle-NNI** for inference of most parsimonious star homoplasy phylogenies from lineage tracing data



Multi-state
Star homoplasy model

✓ Multi-state

✓ Irreversible

✓ Non-modifiable

Paper

Code

https://github.com/raphael-group/startle

# Acknowledgments

**Raphael Group:**

**Dr. Ben Raphael**

Dr. Cong Ma

Dr. Metin Balaban

Dr. Uyen Mai

Dr. Hirak Sarkar

Dr. Brian Arnold

Uthsav Chitra

Alexander Strzalkowski

Sereno Lopez-Darwin

Ahmed Shuaibi

Xinhao Liu

Maya Gupta

Akhil Jakatdar

Gillian Chu

Clover Zheng

**Henri Schmidt**

**Dr. Michelle Chan**

NIH NATIONAL CANCER INSTITUTE

# Backup

# Startle-ILP algorithm for k-star homoplasy phylogeny inference

characters

| cells | $c_1$ | $c_2$ |
|---|---|---|
| $s_1$ | 1 | 0 |
| $s_2$ | 1 | 2 |
| $s_3$ | 1 | 1 |
| $s_4$ | 2 | 1 |

Character matrix

$n \times m$

?

| | $(c_1, 1, 0)$ | $(c_1, 1, 1)$ | $(c_1, 2, 0)$ | $(c_2, 1, 0)$ | $(c_2, 2, 0)$ |
|---|---|---|---|---|---|
| $s_1$ | 1 | 0 | 0 | 0 | 0 |
| $s_2$ | 1 | 0 | 0 | 0 | 1 |
| $s_3$ | 0 | 1 | 0 | 1 | 0 |
| $s_4$ | 0 | 0 | 1 | 1 | 0 |

k-star binarization of the character matrix

$n \times mrk$

necessary

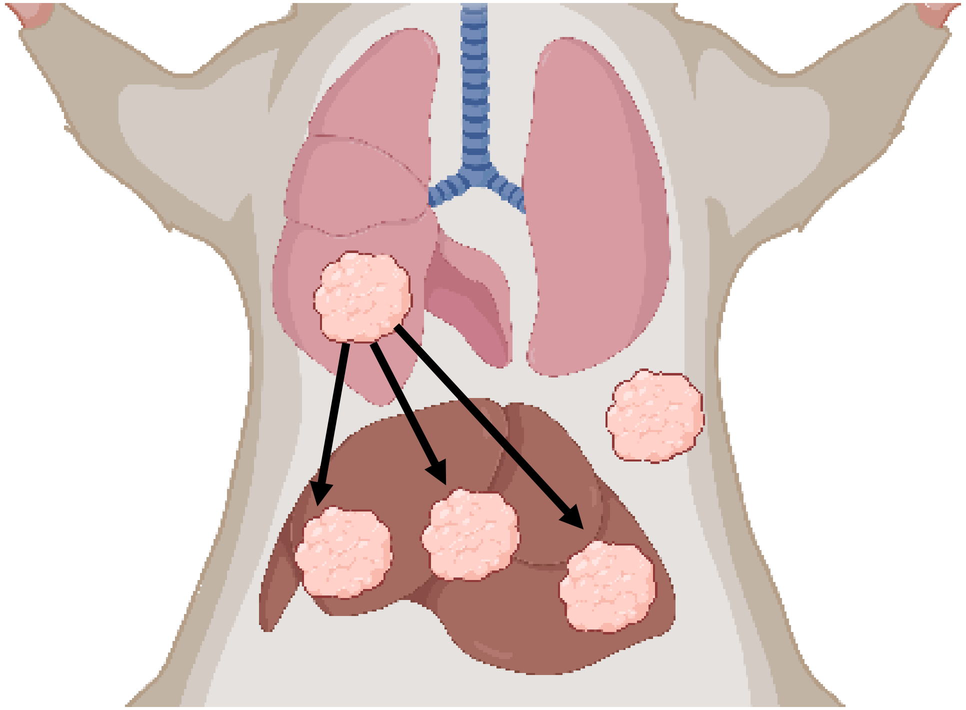
sufficient

A character matrix $A$ admits a k-star homoplasy phylogeny if and only if there exists a k-star binarization of $A$ that admits a two-state perfect phylogeny

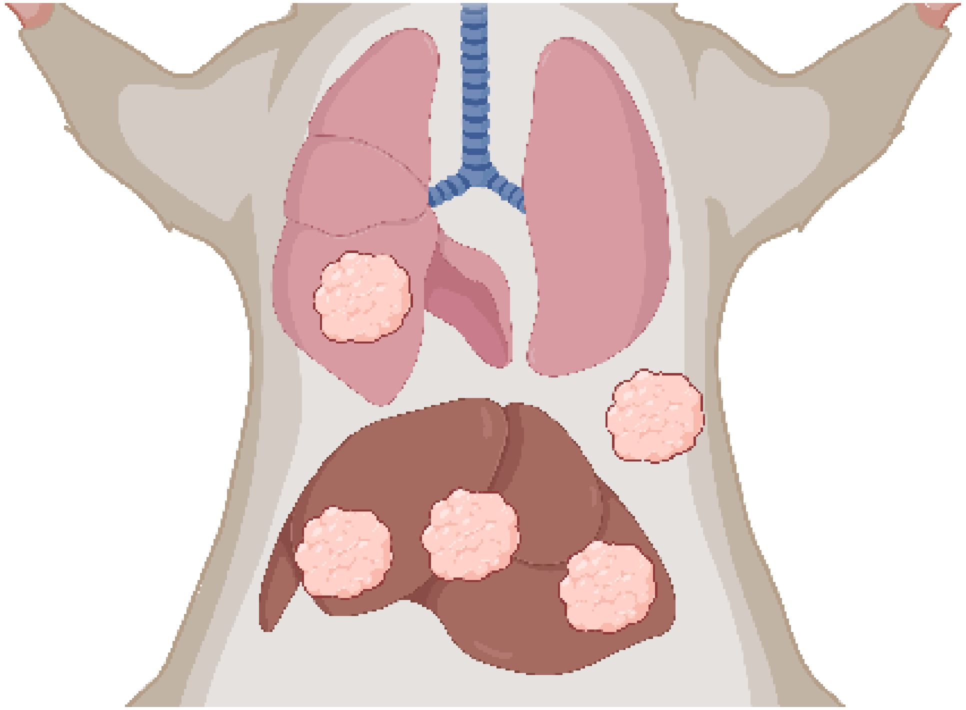
**Startle-ILP**: We formulate a MILP to find the most parsimonious k-star homoplasy phylogeny from lineage tracing data

# Startle supports more parsimonious than published results

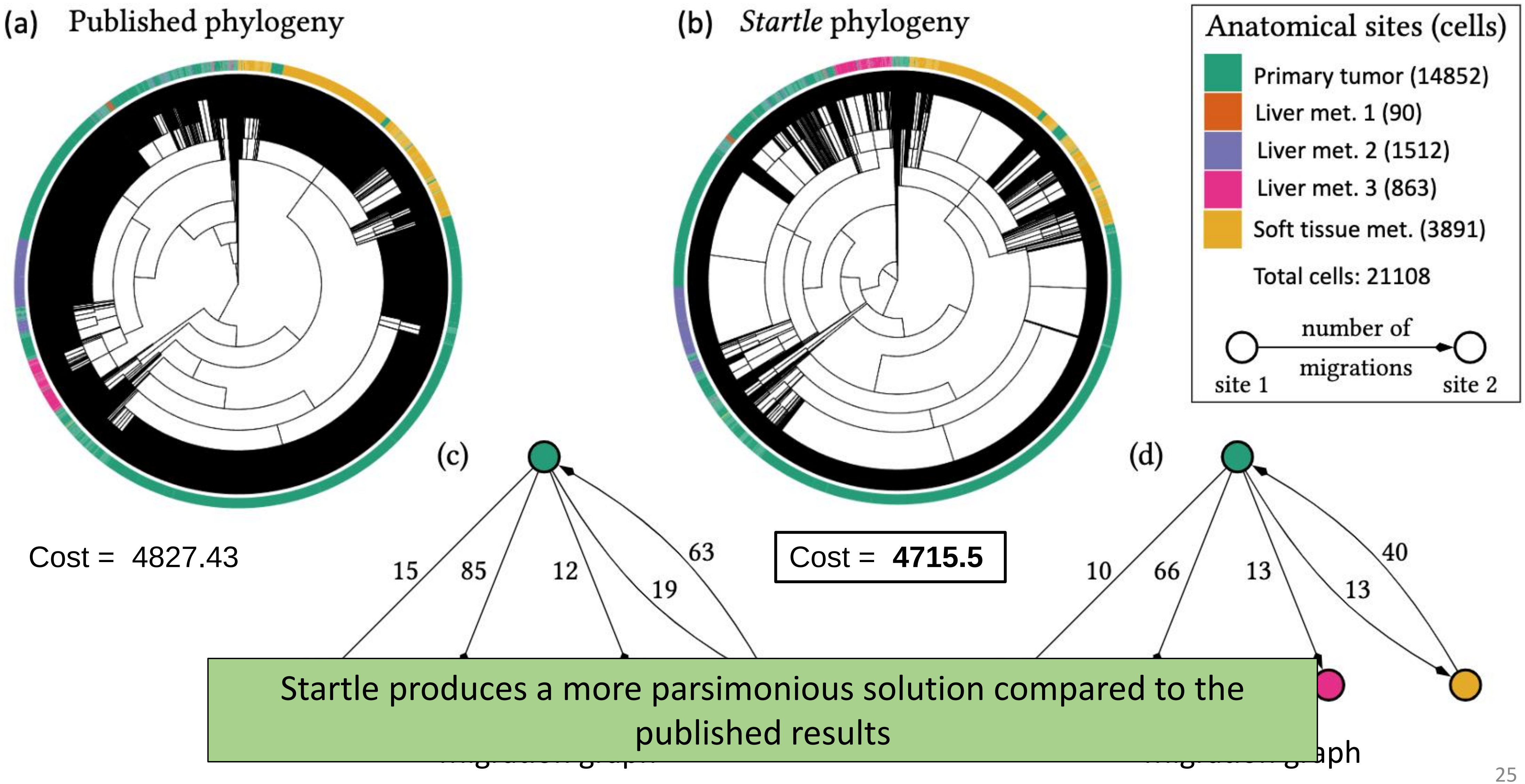
Migration graph from published tree



Migration graph from Startle tree



# Startle trees are more parsimonious than published results



Cost = 4827.43

Cost = **4715.5**

Startle produces a more parsimonious solution compared to the published results

# Startle-ILP algorithm for k-star homoplasy phylogeny inference

characters

cells

$$\begin{array}{c|c|c|} & c_1 & c_2 \\ \hline s_1 & \boxed{1} & 0 \\ \hline s_2 & \boxed{1} & \boxed{2} \\ \hline s_3 & \boxed{1} & \boxed{1} \\ \hline s_4 & \boxed{2} & \boxed{1} \\ \hline \end{array} \xrightarrow{?} \begin{array}{c|c|c|c|c|c|c|c|c|} & (c_1,1,0) & (c_1,1,1) & (c_1,2,0) & (c_1,2,1) & (c_2,1,0) & (c_2,1,1) & (c_2,2,0) & (c_2,2,1) \\ \hline s_1 & \boxed{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline s_2 & \boxed{1} & 0 & 0 & 0 & 0 & 0 & \boxed{1} & 0 \\ \hline s_3 & 0 & \boxed{1} & 0 & 0 & \boxed{1} & 0 & 0 & 0 \\ \hline s_4 & 0 & 0 & \boxed{1} & 0 & \boxed{1} & 0 & 0 & 0 \\ \hline \end{array}$$

Character matrix

$n \times m$

k-star binarization of the character matrix

$n \times mrk$

A character matrix $A$ admits a **k-star homoplasy phylogeny** if and only if there exists a **k-star binarization** of $A$ that admits a **two-state perfect phylogeny**

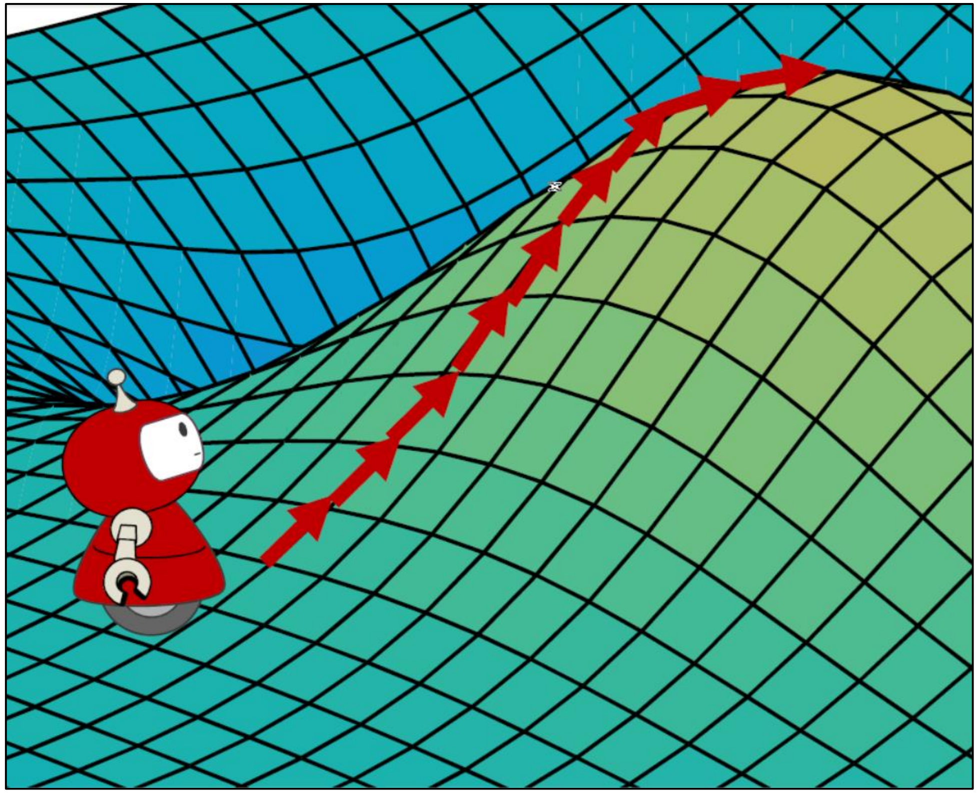
**Startle-ILP**: We formulate a MILP to find the most parsimonious k-star homoplasy phylogeny from lineage tracing data

# Startle-NNI algorithm for star homoplasy phylogeny inference

Hill climbing in the tree space using nearest neighbor interchange (NNI) moves.

Naïve implementation will take O($n^2$m) to compute score of all trees in the 1-move neighborhood of a given tree

Evaluating a tree topology is an instance of the **small parsimony problem**

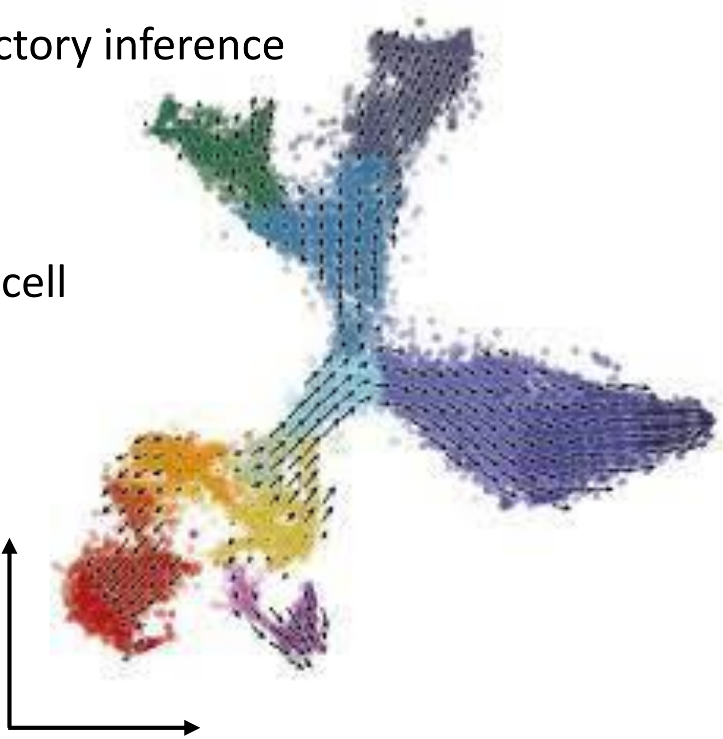


**Startle-NNI**: We use dynamic programing to compute the scores in O(nmd), where d is the average depth of the given tree.

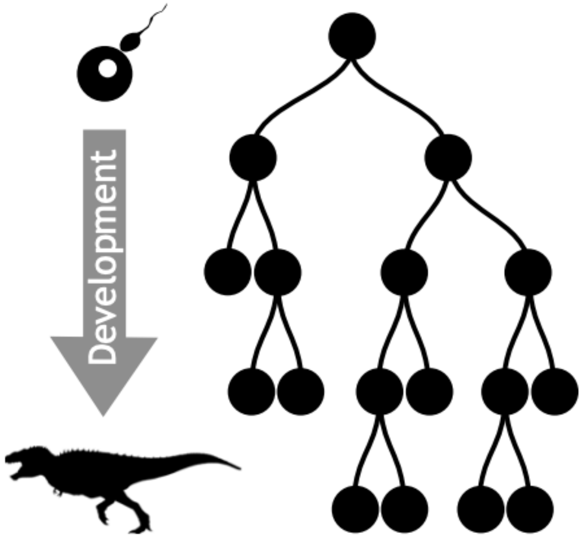
Figures from mathworks.com

# Lineage Tracing: introduction and motivation

Trajectory inference



Lineage tracing

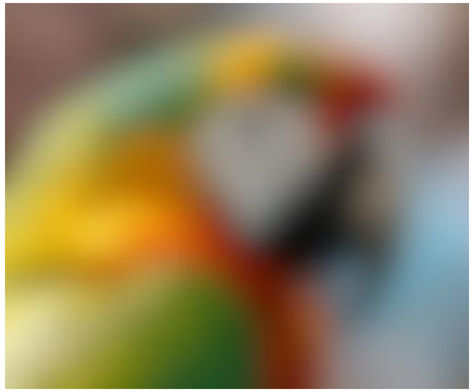


Description of *average* cell dynamics and cell state relationships

Description of *individual* cell dynamics and lineage relationships

Trapnell et al., 2014, Nat. Biotech.
Wolf et al., 2019, Genome Research
Haghverdi et al., 2016, Nat. Methods
Ji et al., 2016, Nucleic Acid Res.
Welch et al., 2018, Genome Biology
Manno et al., 2018, Nature
Qiu et al., 2017, Nat. Methods
Setty et al., 2016, Nat. Biotech.

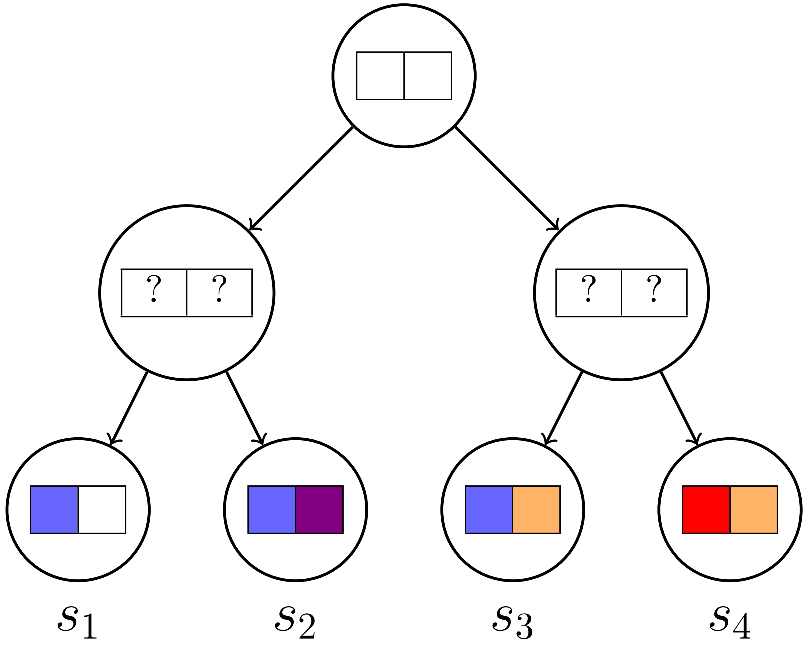
.... and many more



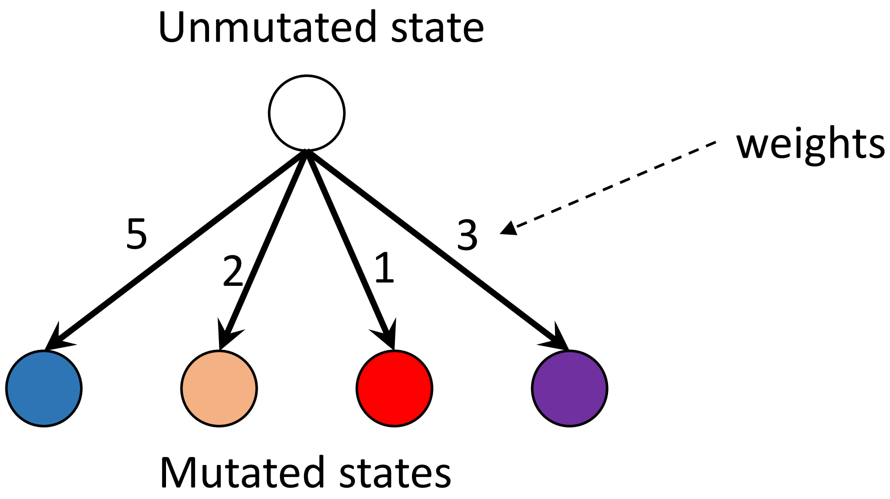
How can we perform lineage tracing?

Manno et al., 2018, McKenna et al., 2019, Development; rare-gallery.com

# Small parsimony problem under the star homoplasy model



**Input**: Leaf labeled phylogeny and mutation weights.



**Star homoplasy model**:

- Each character can **change state at most once** in a lineage (a path from root to leaf)
- Characters evolve **independently** (standard assumption)

**Problem**: Find the labeling of the internal vertices such that the total weight is minimized.

**Solution**: solved in linear time using a dynamic program.
Now we can score a given phylogeny!