Models and Algorithms for Cancer Evolution

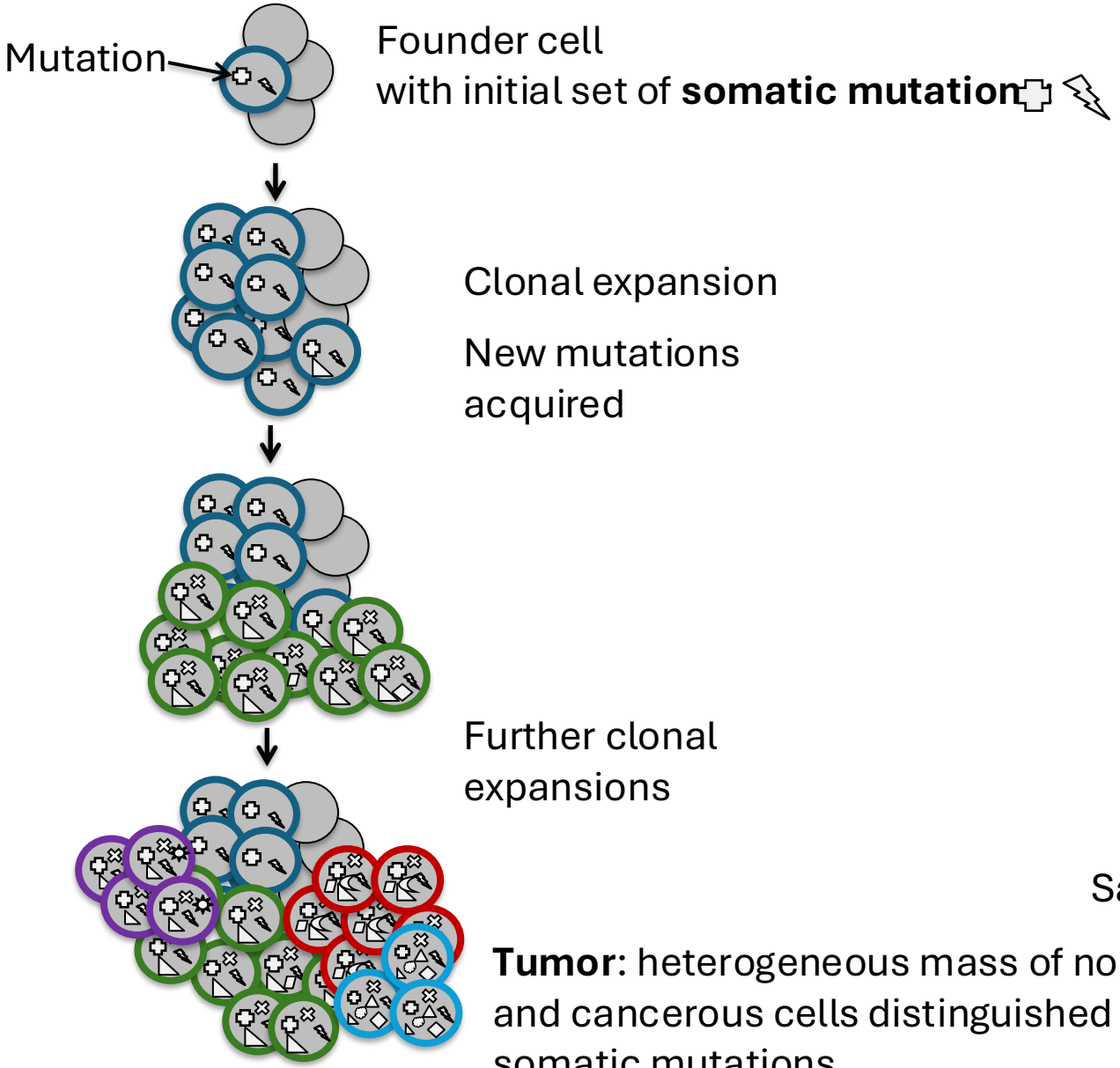
Ben Raphael
Department of Computer Science

PRINCETON UNIVERSITY

IPAM: Mathematics of Cancer

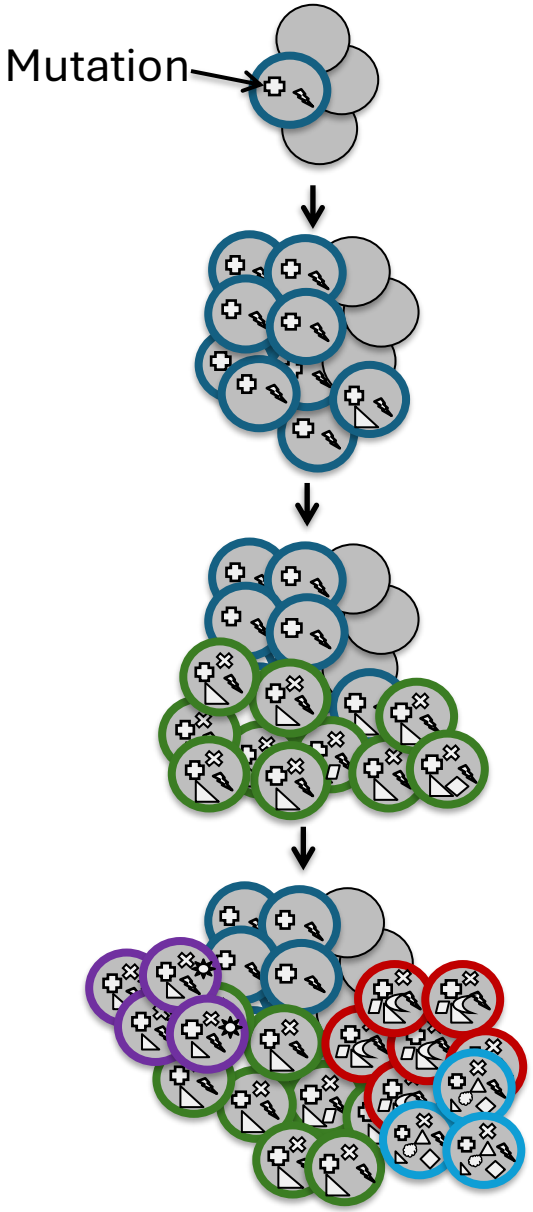
February 24,2026

Cancer is an evolutionary process



Tumor: heterogeneous mass of normal and cancerous cells distinguished by somatic mutations

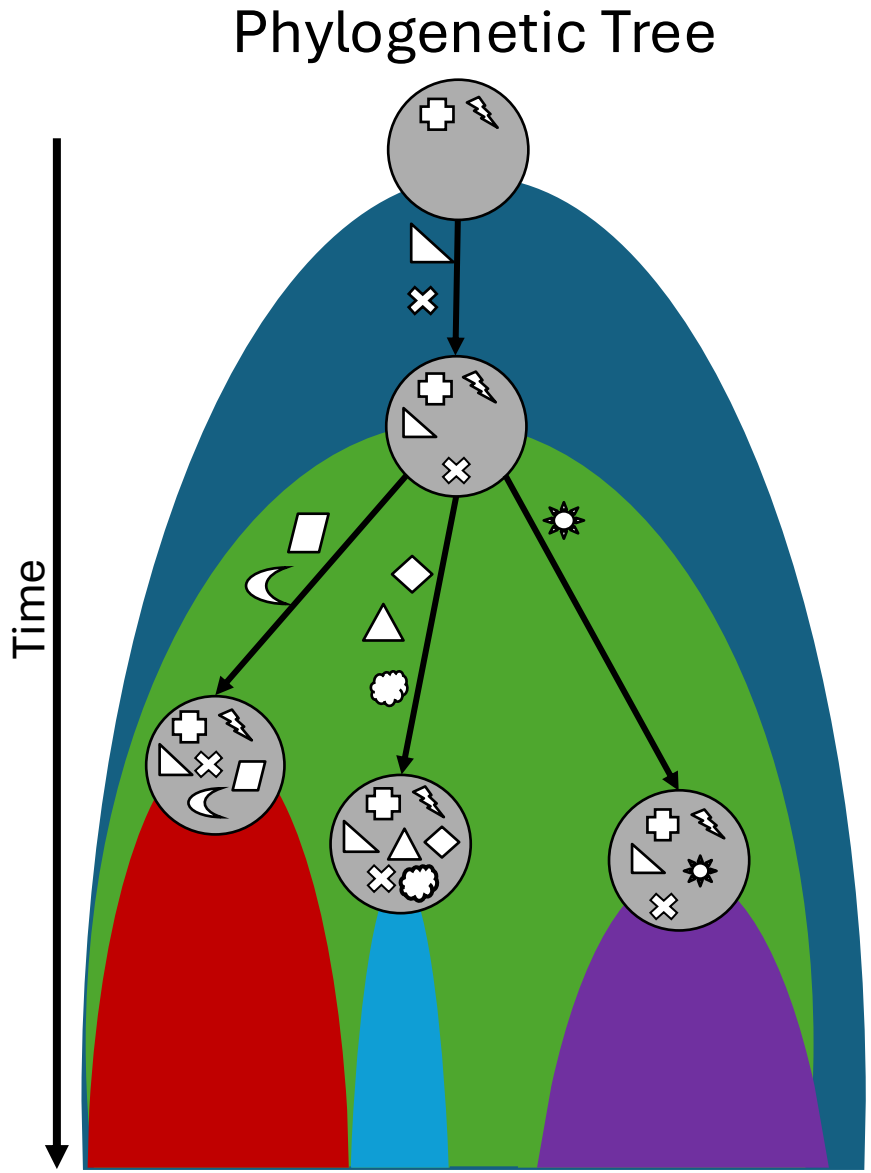
Cancer evolution is described by phylogenetic tree



Vertices =
cells or clones

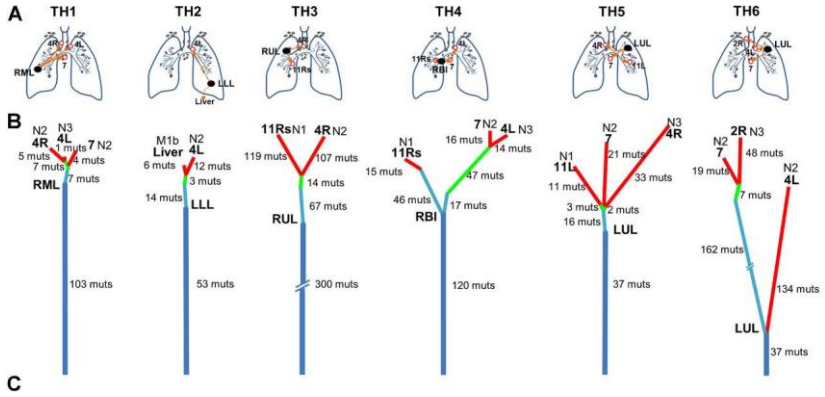
Clone = group of cells with shared set of mutations

Edges =
ancestral relationships



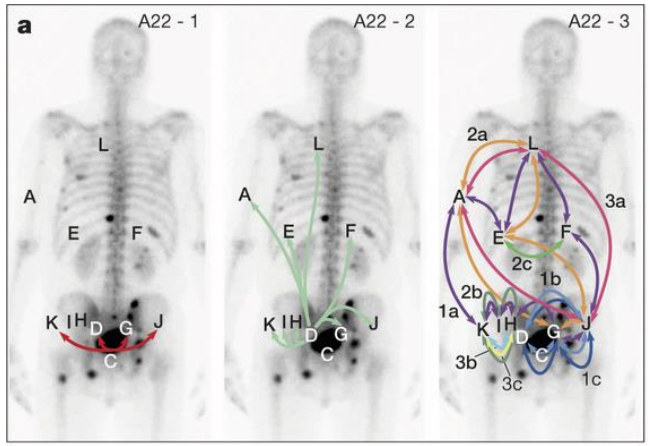
Tumor Evolution is Key to Understanding and Treating Cancer

Recognize common patterns of tumor evolution



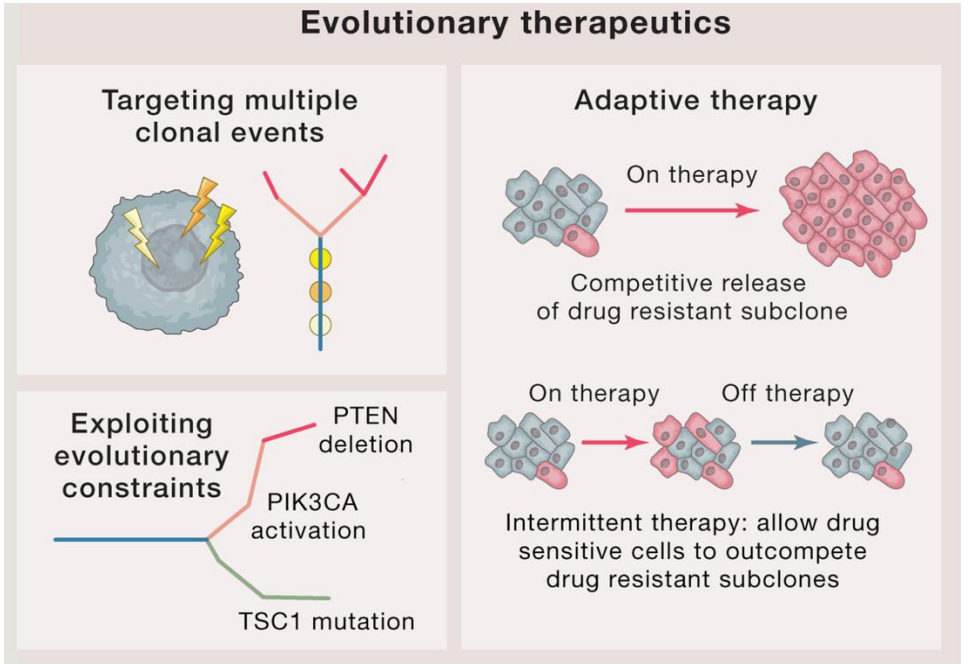
[Um et al., *Cancer Research* 2016]

Understand metastatic development



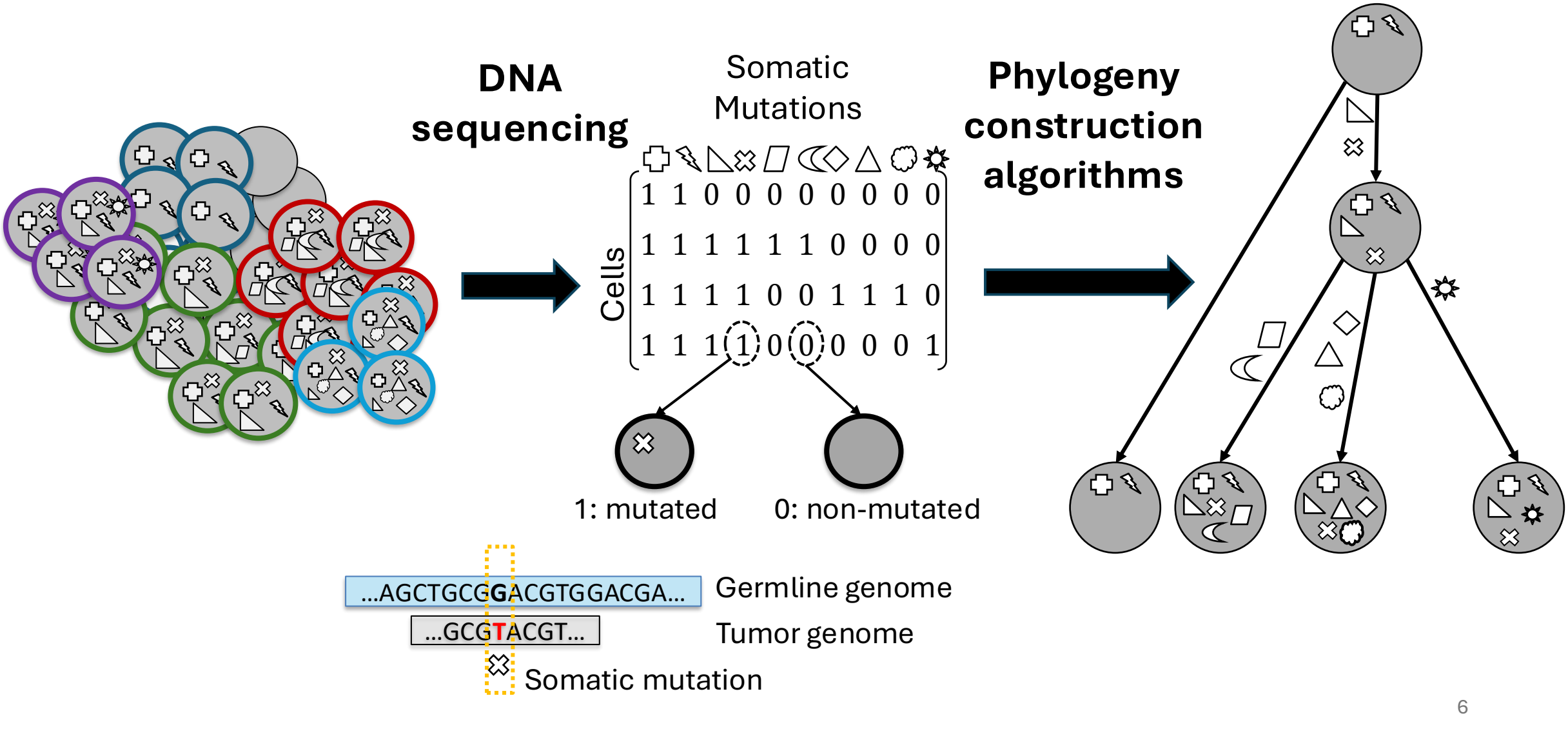
[Gundem et al., *Nature* 2015)

Identify targets for treatment



[McGranahan et al., *Cell* 2017]

Tumor sample → Tumor phylogeny



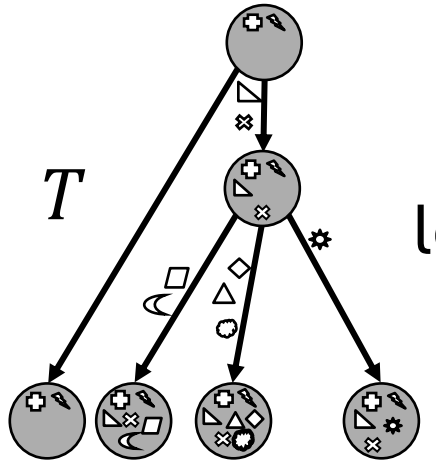
Phylogenetic Tree Inference

Given a character matrix M, what is the "most likely" tree T that generated M?

Mutations / Characters

$$M = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

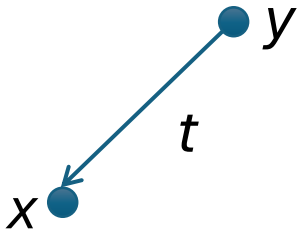
Cells



Branch lengths $\boldsymbol{b}$

Key Assumptions:

1. Each mutation (character) evolves **independently**: $\Pr[\, M \mid T, \boldsymbol{b}] = \Pi_i \Pr[\, M_i \mid T, \boldsymbol{b}]$
2. Evolution of each mutation (character) follows a **continuous-time Markov process**



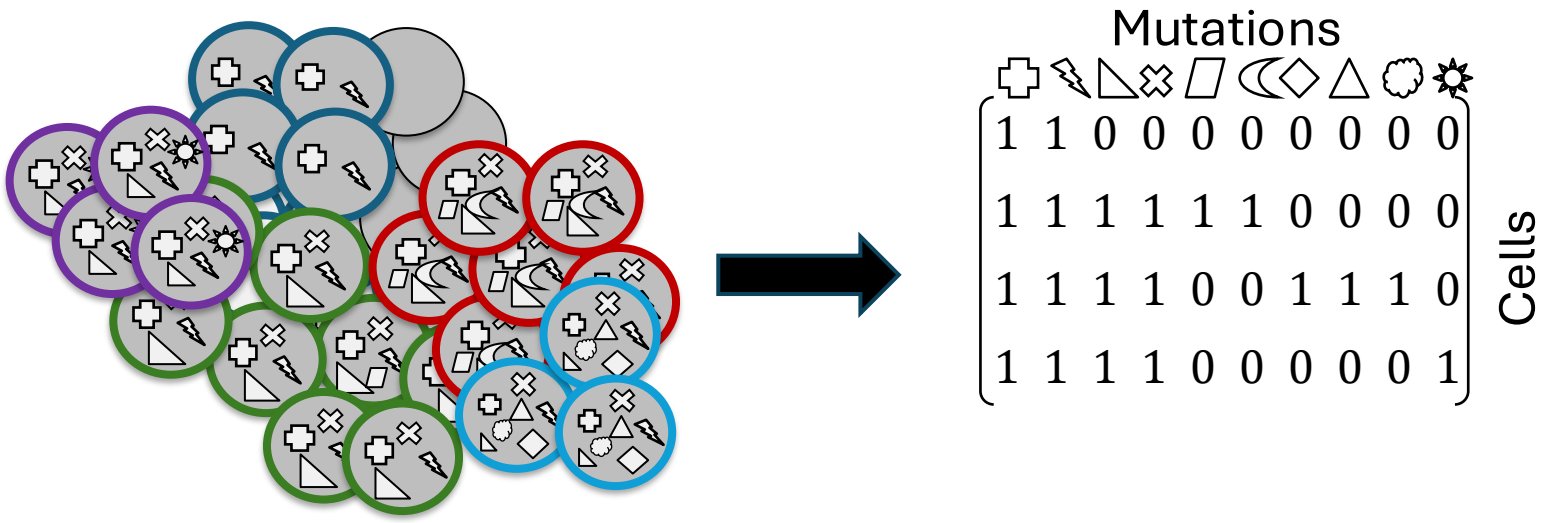
$\Pr[\, x \mid y, t]$ = probability that y mutates to x in time $t = e^{Qt}$
where Q is rate matrix

Calculation $\Pr[\, M_i \mid T, \boldsymbol{b}]$ $\operatorname{argmax}_{\boldsymbol{b}} \Pr[\, M_i \mid T, \boldsymbol{b}]$ $\operatorname{argmax}_{T,\boldsymbol{b},} \Pr[\, M_i \mid T, \boldsymbol{b}]$

EASY HARD

Major challenges in cancer evolution

1. Technological limitations in measuring somatic mutations
   - “Species”/taxa ≡ single cell
   - Whole-genome single-cell DNA sequencing remains technically difficult!

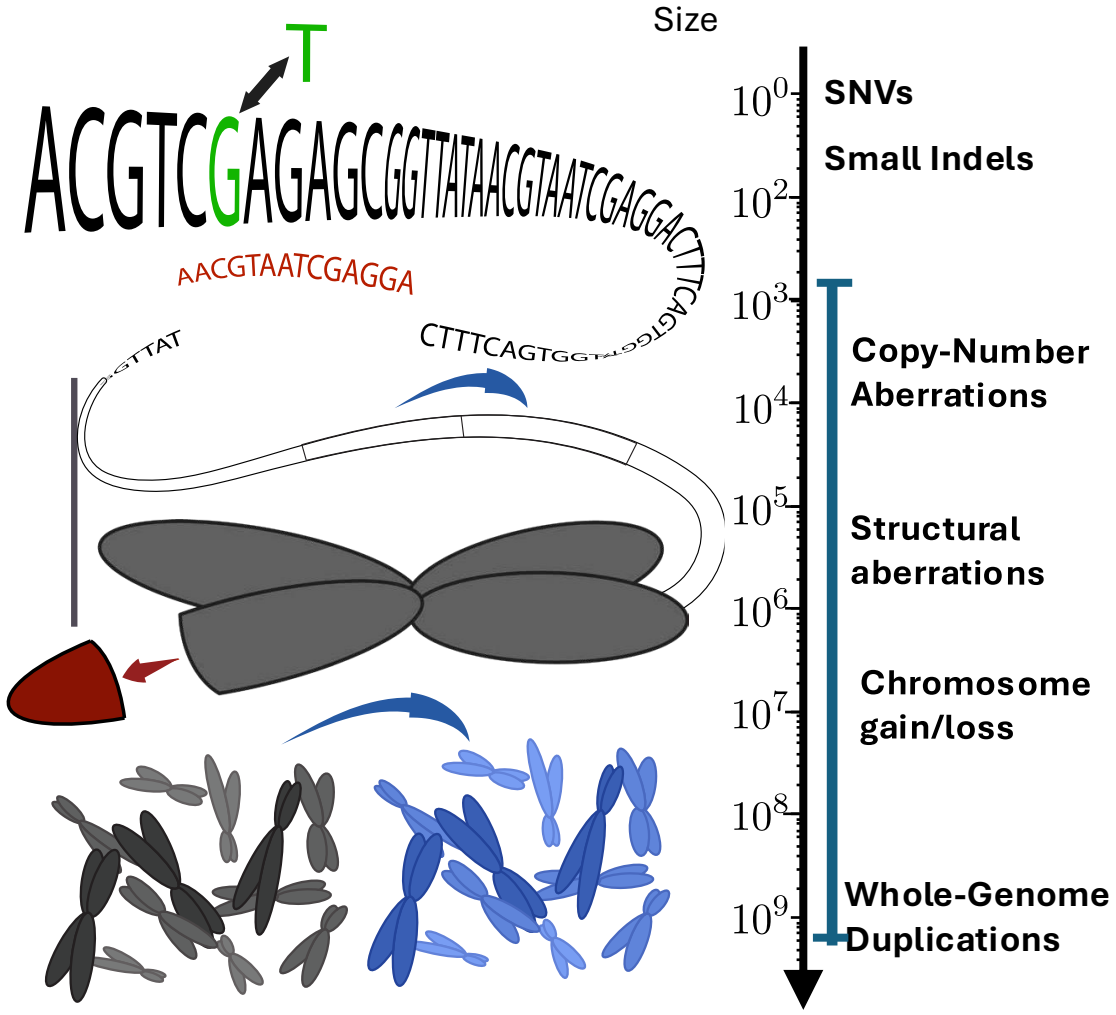


→ Constructing phylogeny under uncertainty in mutations

Major challenges in cancer evolution

2. Cancer genomes are **complex**: somatic mutations occur over ***all*** genomic scales

→ Specialized evolutionary models/distances

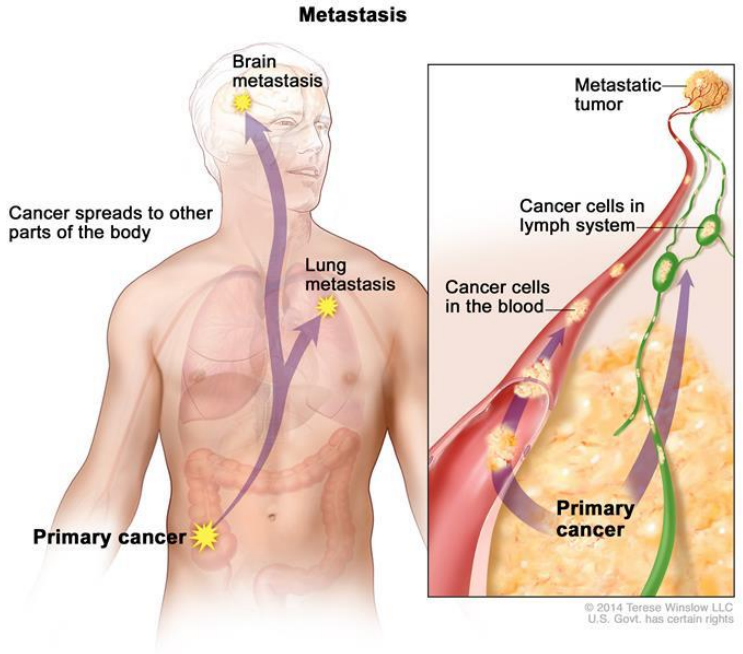


Challenges in cancer phylogeny

3. Phylogeography and ancestral reconstruction

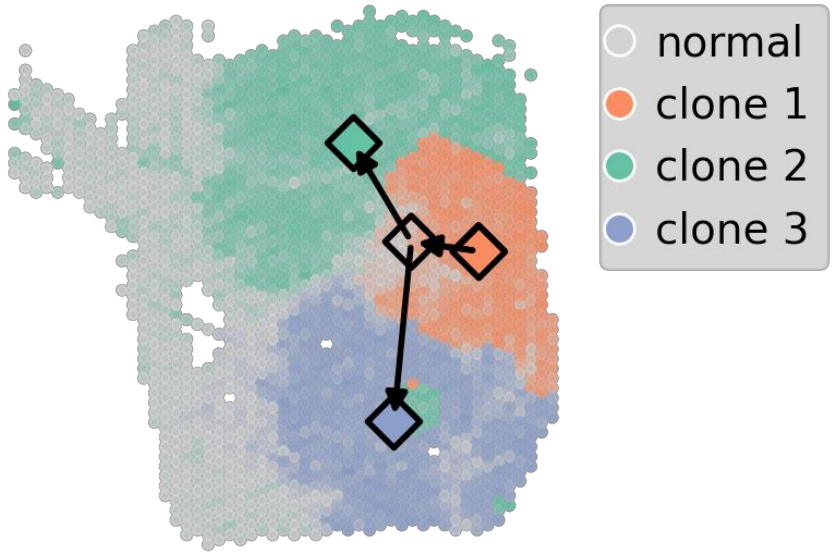
Metastasis

migration between
distant anatomical sites



Tumor growth

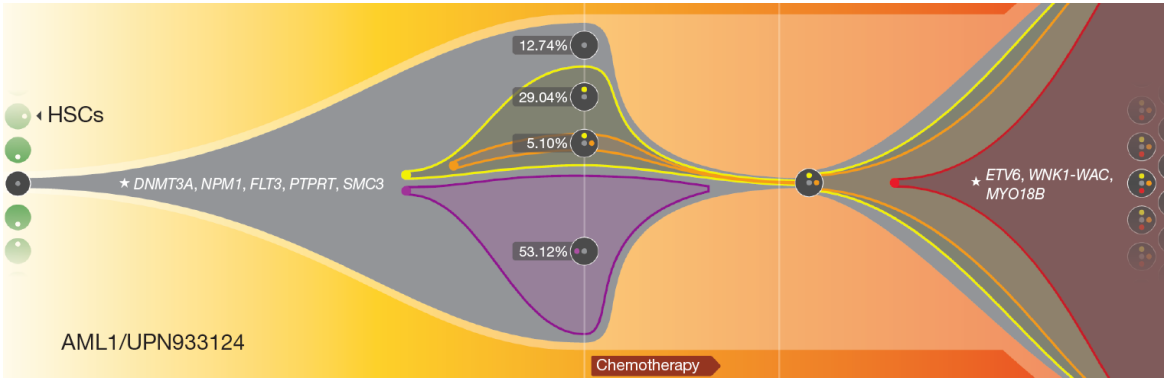
Local migration



Ma et al. *Nature Methods* 2024

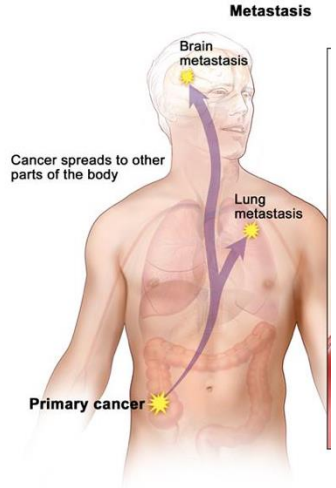
Outline

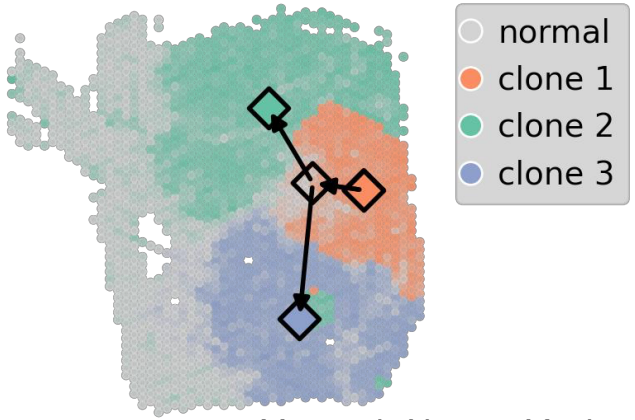
1. Constructing phylogenies from bulk and single-cell cancer sequencing



Ding et al. *Nature* 2012

2. Migration: Metastasis and Spatial tumor evolution

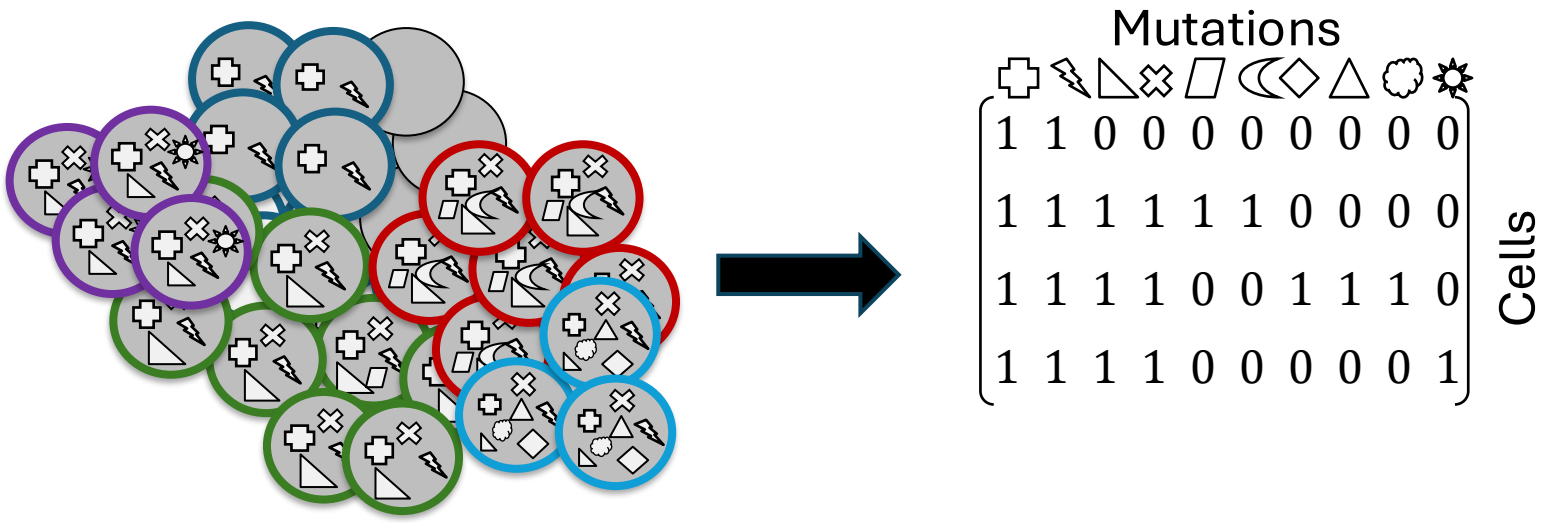




Ma et al. *Nature Methods* 2024

Major challenges in cancer evolution

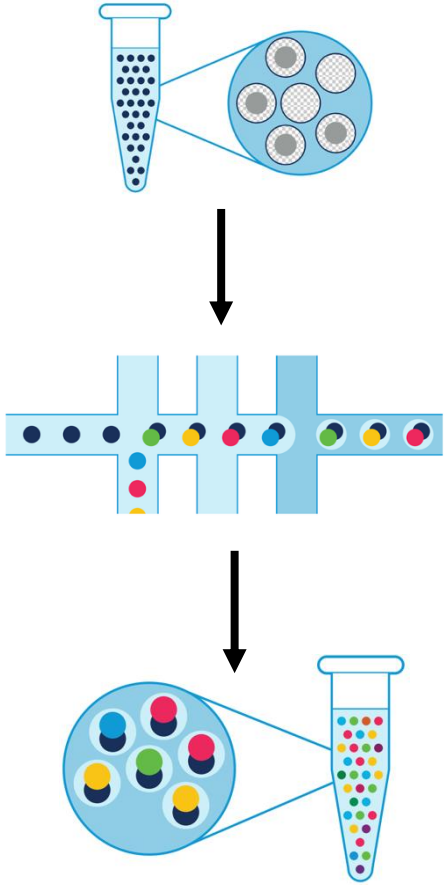
1. Technological limitations in measuring somatic mutations
   - “Species”/taxa ≡ single cell
   - Whole-genome single-cell DNA sequencing remains technically difficult!



→ Constructing phylogeny under uncertainty in mutations

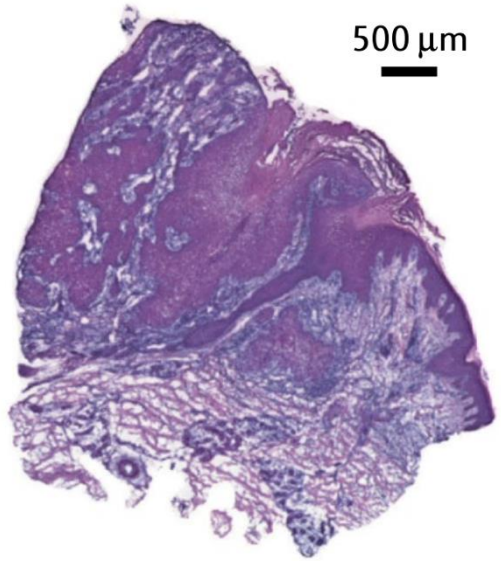
DNA/RNA sequencing technologies

Single-cell



Barcoded **cells** for DNA/RNA sequencing

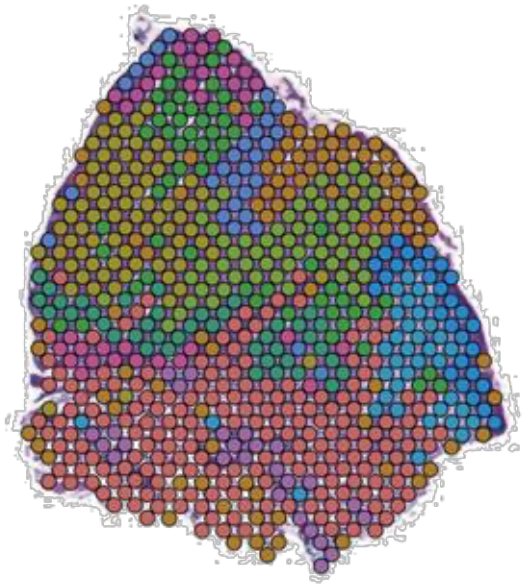
Bulk



Tissue Sample

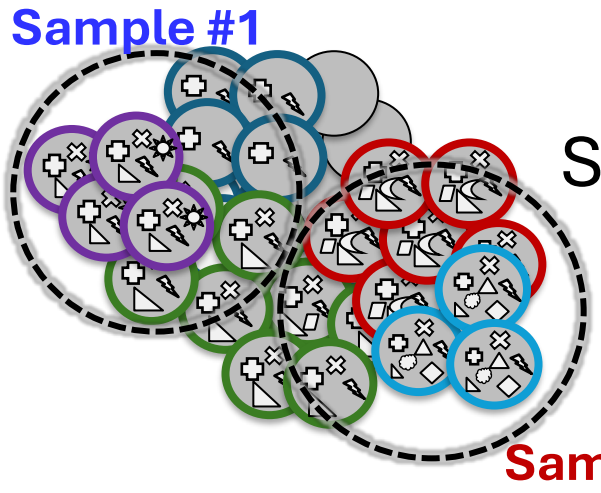
Single **aggregate** measurement

Spatial

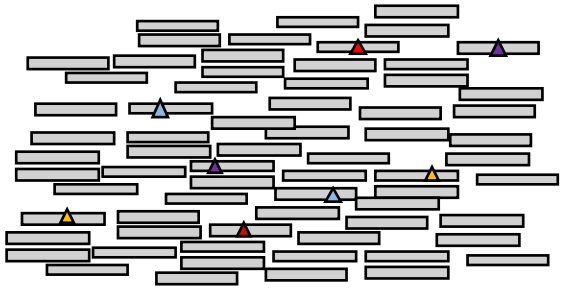


Barcoded **spots** in tissue slice for DNA/RNA sequencing*

DNA Sequencing of Bulk Tumor Samples



Bulk DNA Sequencing



Billions of short DNA sequences **(reads)** from all sequenced cells

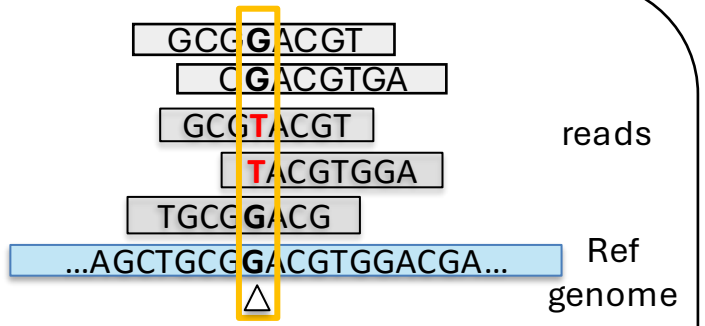
Sample is thousands – millions of cells

Mutations

Samples

$$\begin{bmatrix} 1 & 1 & .5 & .5 & .2 & .2 & 0 & 0 & 0 & .3 \\ 1 & 1 & .7 & .7 & .3 & .3 & .4 & .4 & .4 & 0 \end{bmatrix}$$

Mutation Frequency Matrix F

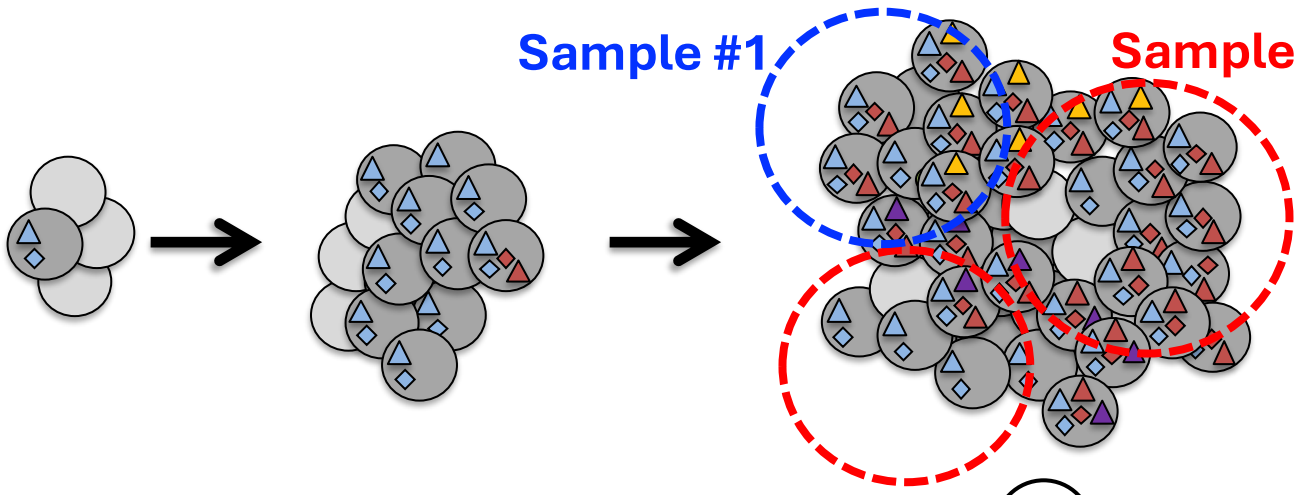


$$f(\Delta) = \frac{\#\text{ reads with } \Delta}{\#\text{ reads}}$$

$$\propto \%\text{ cells with } \Delta$$

Variant allele frequency

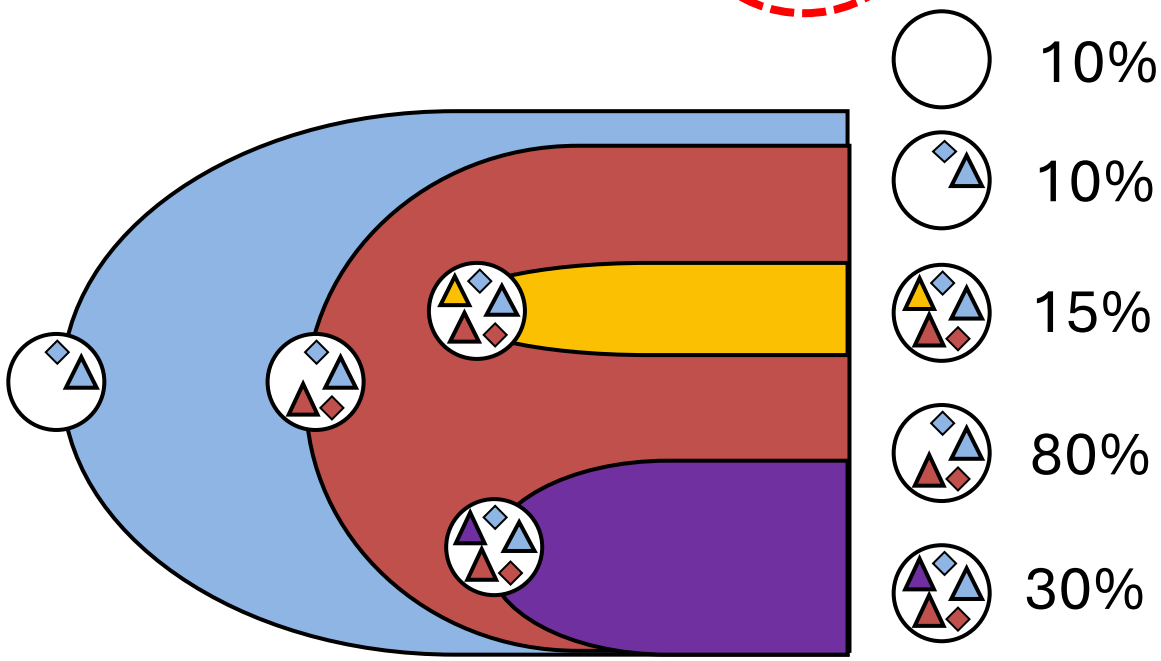
Phylogenetic Deconvolution



Bulk tumor sequencing

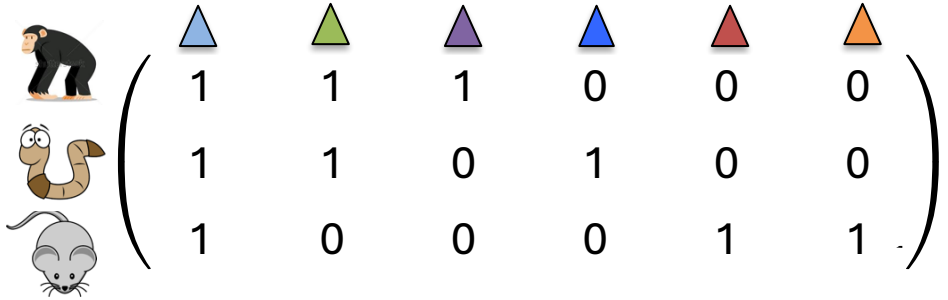
Mutation	Estimated Frequency (Sample 1)
◆	90%
▲	90%
▲	80%
◆	80%
▲	30%
▲	15%

???



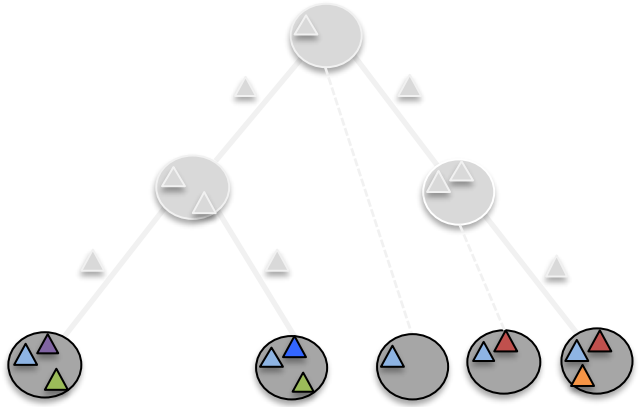
Recover **evolutionary history** of tumor from **mixture**

Traditional Phylogeny



$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Observe: Leaves of *unknown* tree



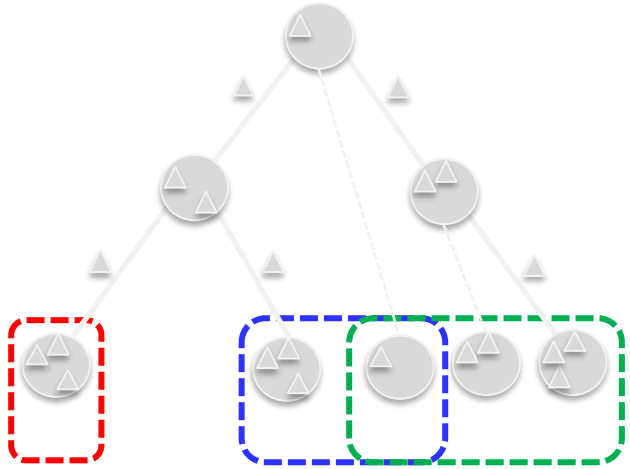
Mixture Phylogeny

(bulk sequencing)

$$\begin{pmatrix} 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

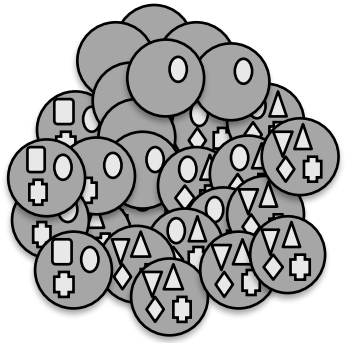
Frequency Matrix ***F***

Observe: Mixture of *unknown* leaves of *unknown* tree in *unknown* proportions



Single-cell DNA sequencing has high rates of missing data and errors

Tumor Sample



Extract single cells

Whole-genome amplification

Single-cell DNA sequencing

Mutations

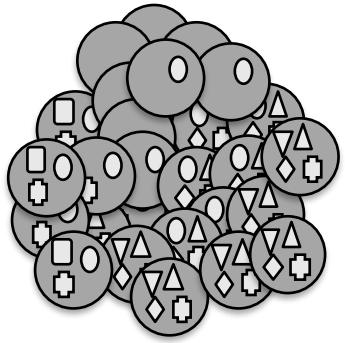
Cells

1	0	0	0	0	0	0
1	1	0	0	0	0	1
1	1	0	0	0	1	0
1	1	0	0	1	1	0
1	1	1	0	1	1	0
1	1	1	0	1	1	0
1	1	0	0	1	0	0
1	1	0	0	1	0	0

[Navin, *Genome Biology* 2014]

Single-cell DNA sequencing has high rates of missing data and errors

Tumor Sample



Extract single cells

Whole-genome amplification

Single-cell DNA sequencing

Mutations

Cells

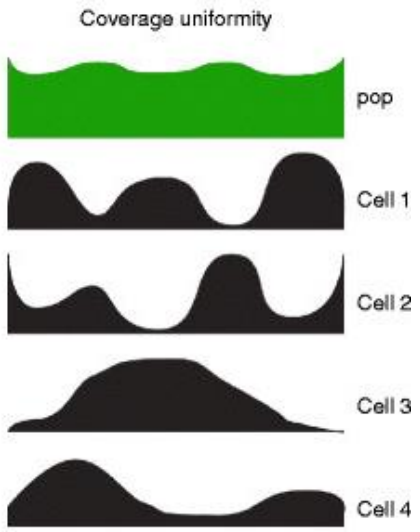
1	0	?	1	0	?	0	
0	1	?	0	0	0	1	
1	0	0	?	0	0	?	
?	1	0	?	1	1	0	
1	0	1	0	?	1	0	
?	1	1	?	1	?	?	
1	0	0	0	0	0	?	
0	1	?	0	0	0	?	

[Navin, *Genome Biology* 2014]

False positives/negatives

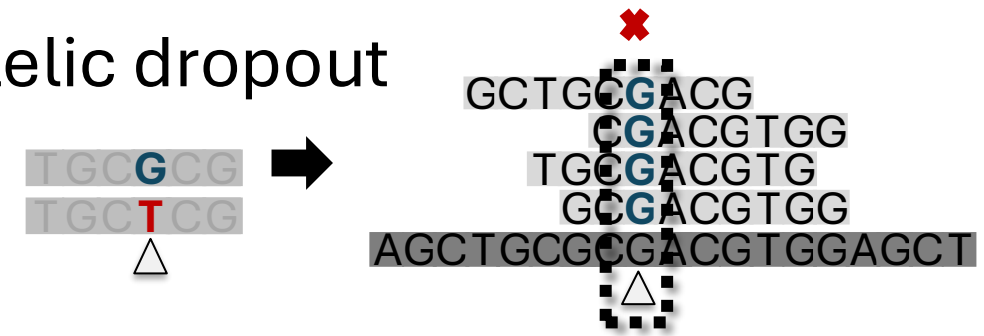
? = missing data

Errors in SNV detection in scDNA-seq

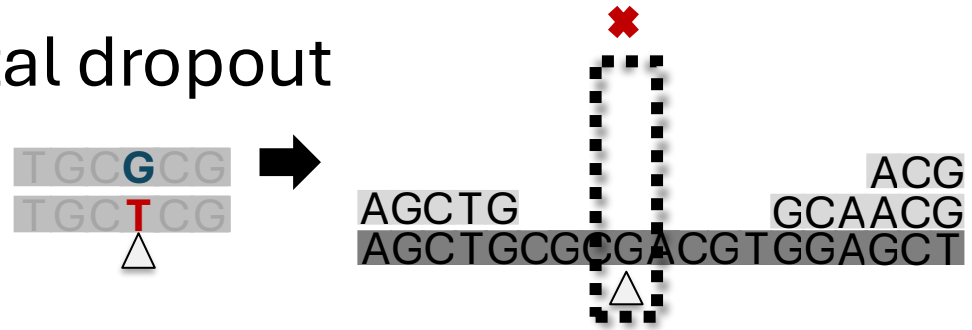


(Navin, *Genome Biology* 2014)

Allelic dropout



Total dropout



False Negatives

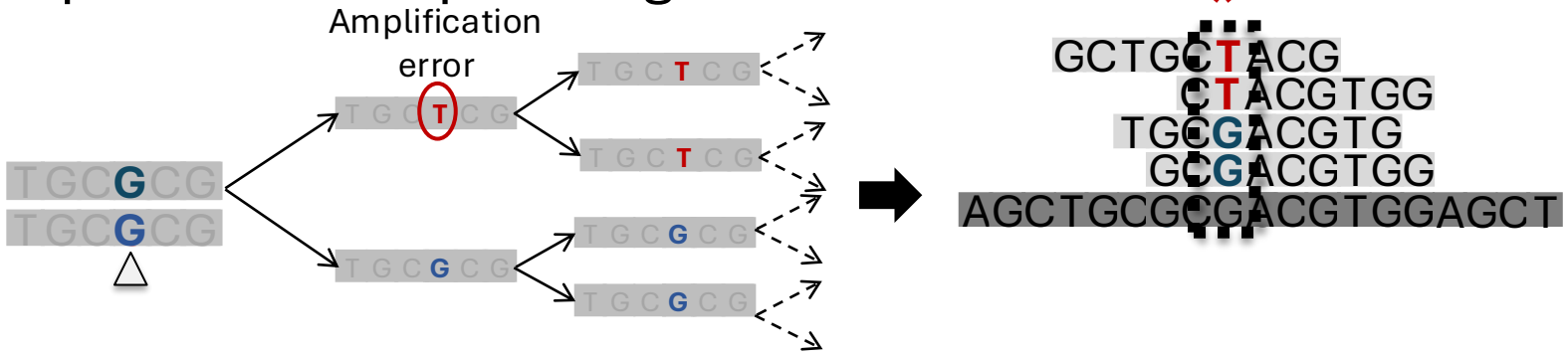
Missing Data

Mutations

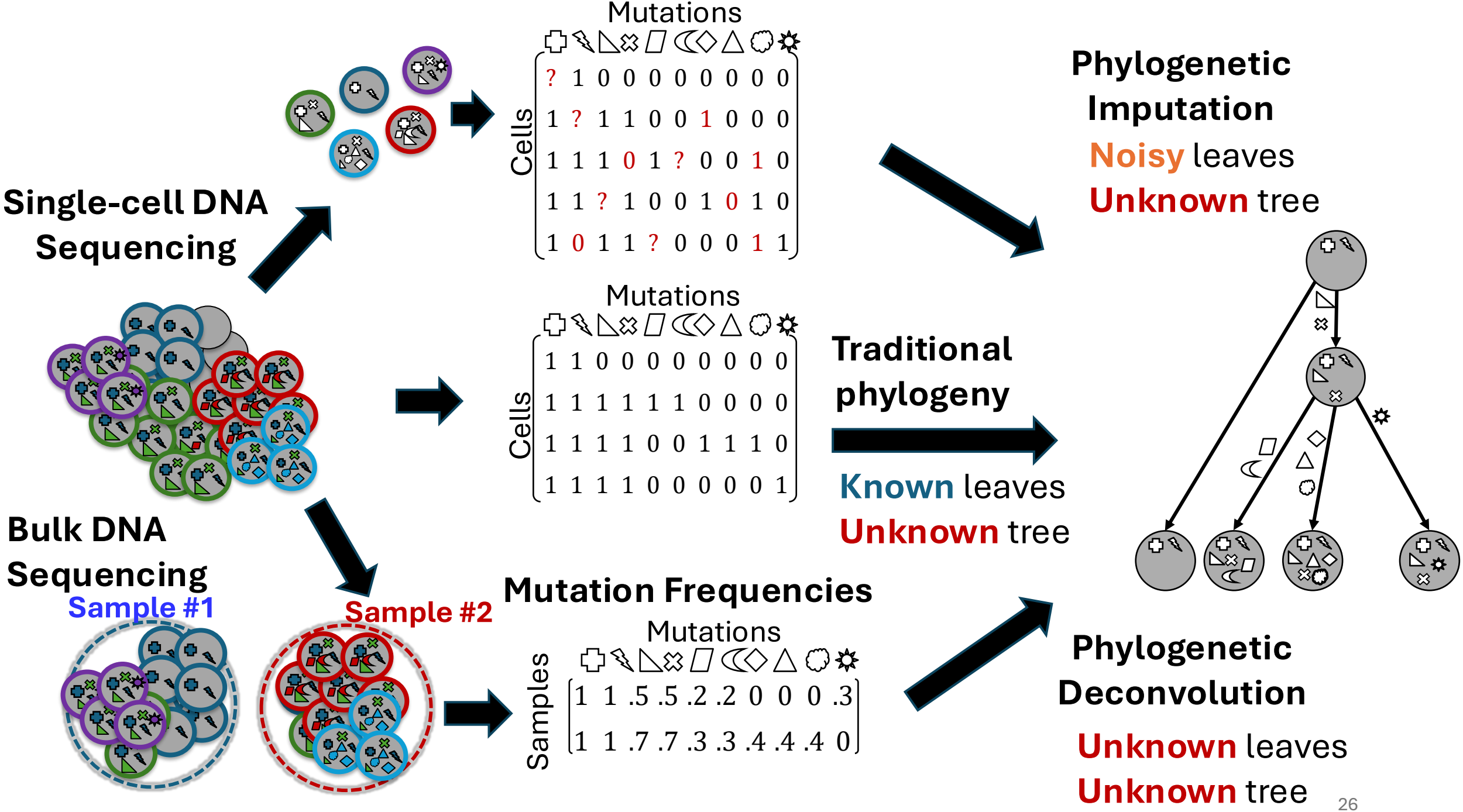
Cells

?	1	0	0	0	0	0	0	0	0
1	?	1	1	0	0	1	0	0	0
1	1	1	0	1	?	0	0	1	0
1	1	?	1	0	0	1	0	1	0
1	0	1	1	?	0	0	0	1	1

Amplification/sequencing errors



False Positives


Single-cell DNA Sequencing
Bulk DNA Sequencing
Sample #1
Sample #2
Mutations
Cells
? 1 0 0 0 0 0 0 0 0
1 ? 1 1 0 0 1 0 0 0
1 1 1 0 1 ? 0 0 1 0
1 1 ? 1 0 0 1 0 1 0
1 0 1 1 ? 0 0 0 1 1
Mutations
Cells
1 1 0 0 0 0 0 0 0 0
1 1 1 1 1 1 0 0 0 0
1 1 1 1 0 0 1 1 1 0
1 1 1 1 0 0 0 0 0 1
Mutation Frequencies
Mutations
Samples
1 1 .5 .5 .2 .2 0 0 0 .3
1 1 .7 .7 .3 .3 .4 .4 .4 0
Phylogenetic Imputation
Noisy leaves
Unknown tree
Traditional phylogeny
Known leaves
Unknown tree
Phylogenetic Deconvolution
Unknown leaves
Unknown tree

How to solve **difficult phylogenetic imputation** and **phylogenetic deconvolution** problems?

Use a simple evolutionary model: **perfect phylogeny**

Mutations

Cells

?	1	0	0	0	0	0	0	0	0
1	?	1	1	0	0	1	0	0	0
1	1	1	0	1	?	0	0	1	0
1	1	?	1	0	0	1	0	1	0
1	0	1	1	?	0	0	0	1	1

Phylogenetic Imputation

Noisy leaves

Unknown tree

Mutations

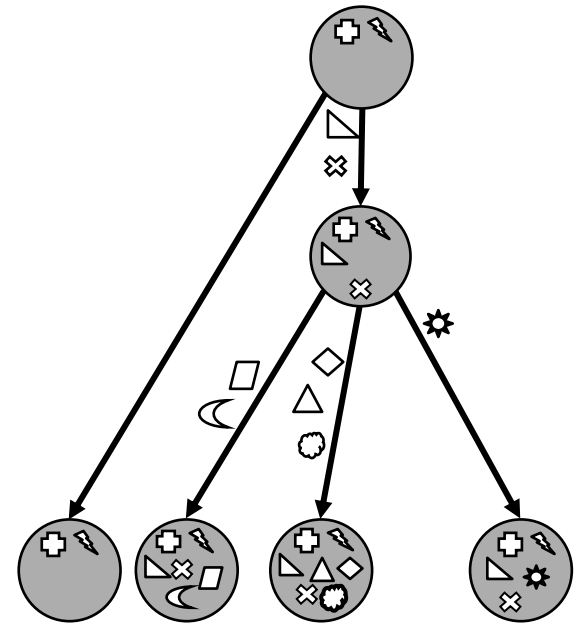
Cells

1	1	0	0	0	0	0	0	0	0
1	1	1	1	1	1	0	0	0	0
1	1	1	1	0	0	1	1	1	0
1	1	1	1	0	0	0	0	0	1

Traditional phylogeny

Known leaves

Unknown tree



Mutation Frequencies

Mutations

Samples

1	1	.5	.5	.2	.2	0	0	0	.3
1	1	.7	.7	.3	.3	.4	.4	.4	0

Phylogenetic Deconvolution

Unknown leaves

Unknown tree

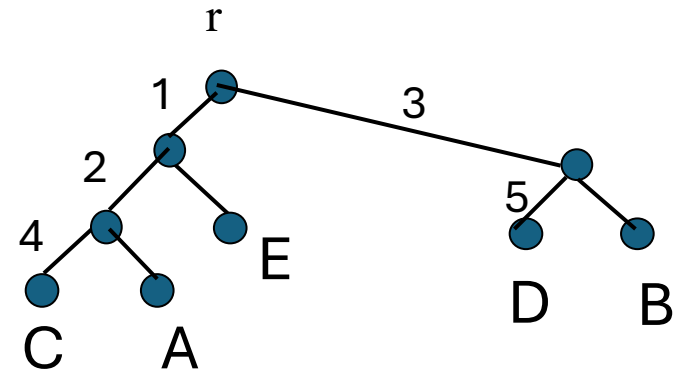
Perfect Phylogeny

Rooted, Binary Character with **all zero ancestor**

1. Evolutionary model
   - Binary characters {0,1}
   - Each character changes state ***at most once**** in evolutionary history (**no homoplasy**!).
2. Rooted binary tree *T* with:
   - root = (0, 0, 0, 0, 0) [germline genome]
   - Every character labels exactly one edge in *T* (edge where character changes 0→ 1)

characters

species	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	0
C	1	1	0	1	0
D	0	0	1	0	1
E	1	0	0	0	0



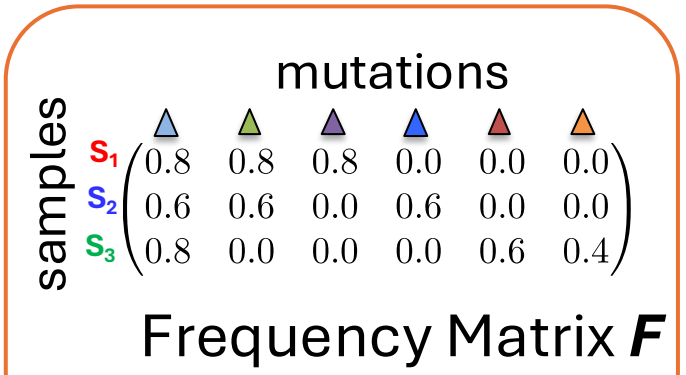
Theorem: Given $n \times m$ *binary matrix* $\boldsymbol{M}$*, can construct phylogeny T (if exists) in* $O(mn)$ *time*

G. Estabrook, C. Johnson, and F. McMorris (1976); Gusfield (1991)

*** Infinite sites assumption**

Perfect Phylogeny Mixture → **Matrix factorization**

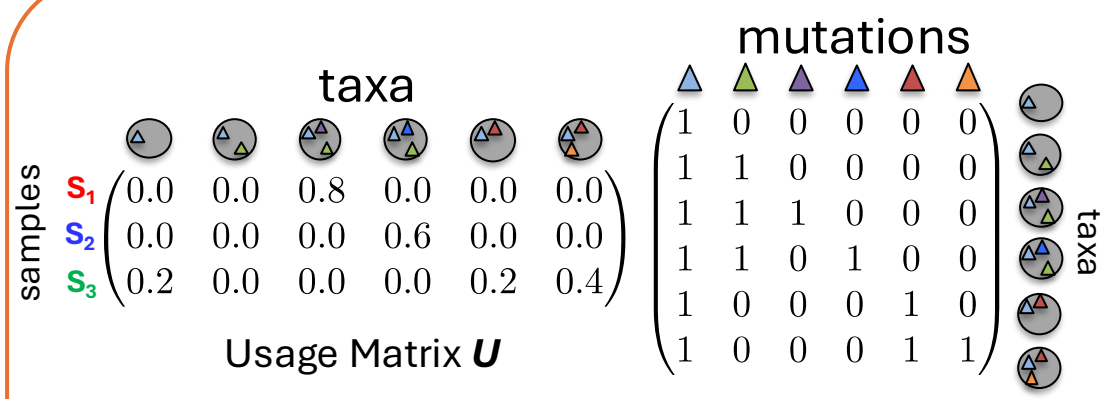
Observed



Frequency Matrix ***F***

=

Unobserved



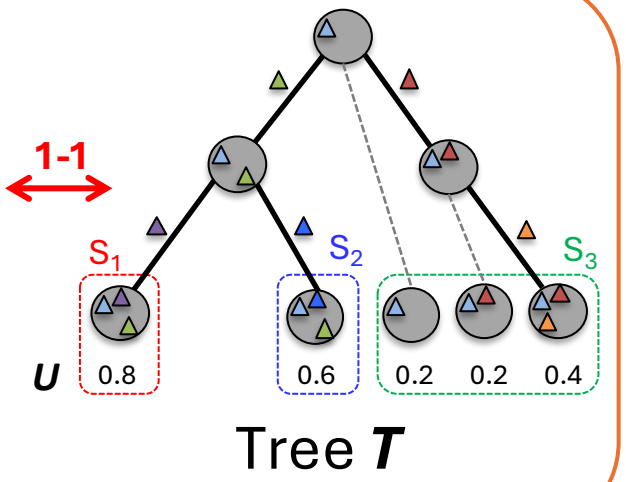
Usage Matrix ***U***

$0 \leq u_{ij} \leq 1$

$\sum_j u_{ij} \leq 1$

Matrix ***B***

Perfect phylogeny matrix



Tree ***T***

Tools

single-sample:

[Nik-Zanial et al. *Cell* 2012]
TrAp [Strino *et al.*, 2013]
Rec-BTP [Hajirasouliha *et al.*, 2014]
CITUP-single [Malikic *et al.*, 2016]
…

multi-sample:

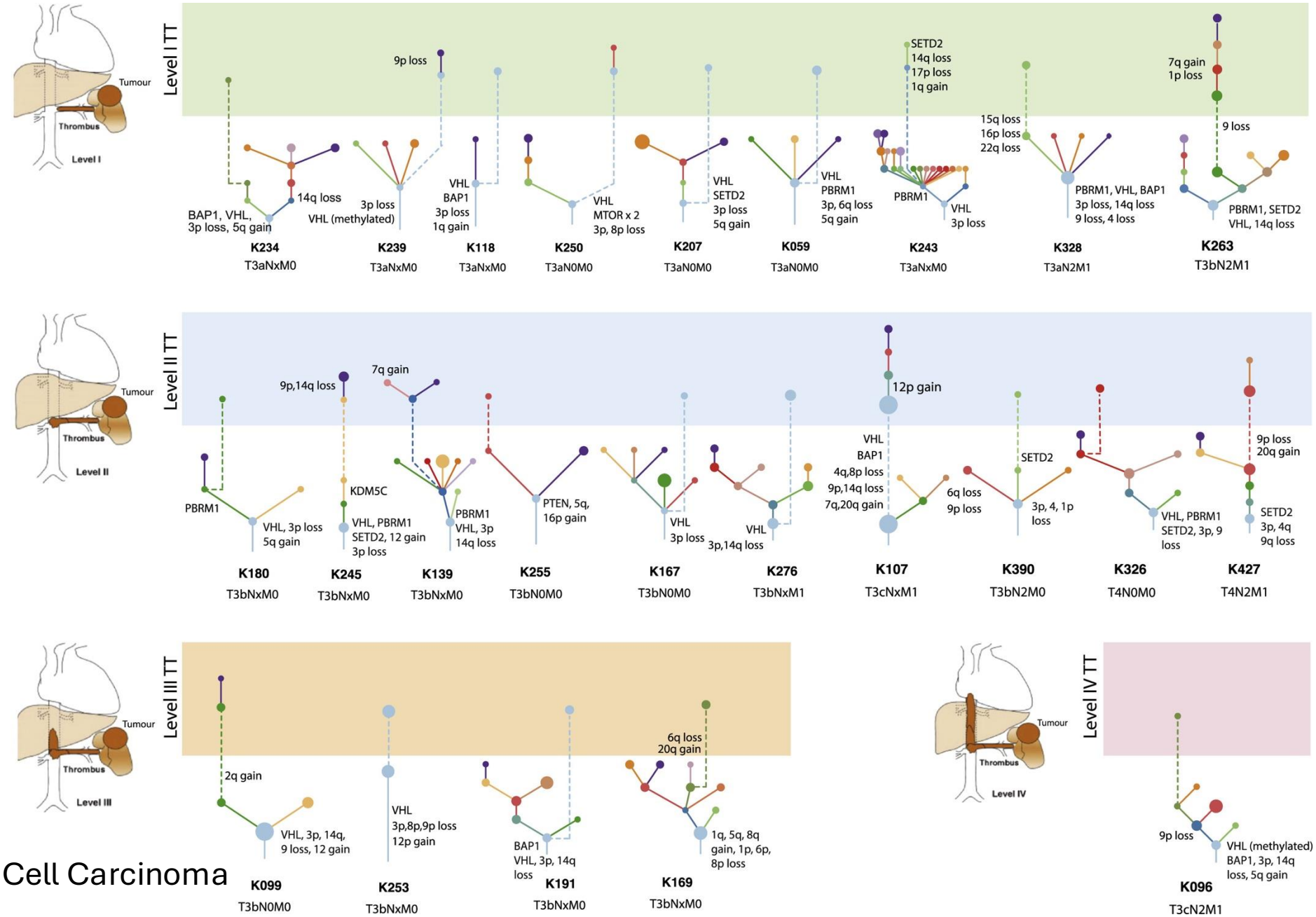
PhyloSub [Jiao *et al.*, 2014]
Clomial [Zare *et al.*, 2014]
Binary *F* [Hajirasouliha *et al.*, 2014]
PhyloWGS [Deshwar, et al. 2015]
AncesTree [El-Kebir et al. 2015]
SCHISM [Niknafs *et al.*, 2015]
CITUP [Malikic *et al.*, 2015]
BitPhylogeny [Yuan *et al.*, 2015]
LICHeE [Popic *et al.*, 2015]
Canopy [Jiang et al., 2016]
PASTRI [Satas and R. (2017)]
…

Mohammed El-Kebir

Proposition. T is a solution if and only if T is a spanning tree of graph $G(F)$ built from F that satisfies a linear constraint (“*sum condition*”)

NP hard problem. Solve via ILP

El Kebir, et al. [*Bioinformatics/ISMB* 2015].



TRACERx Study of Renal Cell Carcinoma

[Turajlic, et al. *Cell* 2018]

Single cell: Phylogeny reconstruction as **matrix correction**

M'

Mutations

Cells

1	0	0	0	0	**1**	0
0	1	0	0	0	0	1
1	1	**?**	0	0	1	0
1	**0**	0	0	1	1	**?**
0	1	**0**	0	1	**?**	0
1	1	1	0	1	1	0
?	**0**	0	0	**0**	0	0
1	1	**0**	0	1	0	0

M

Mutations

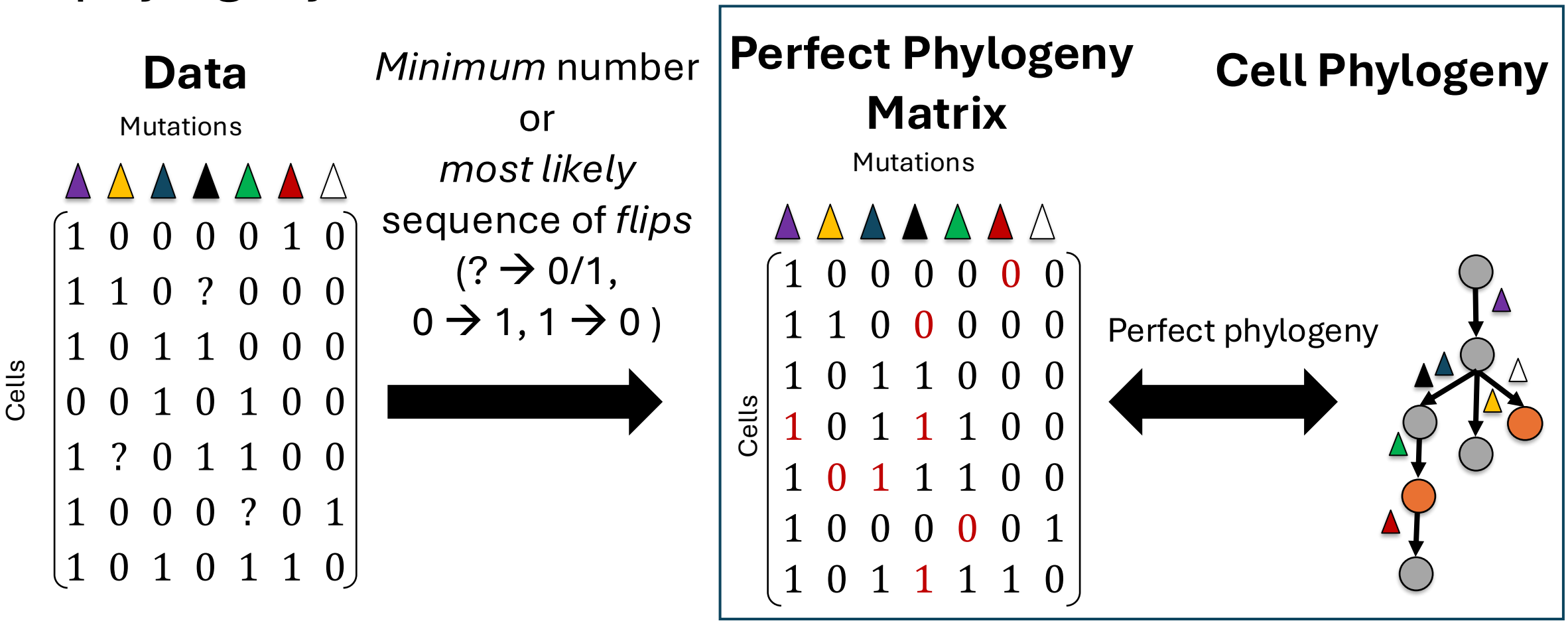
Cells

1	0	0	0	0	**0**	0
1	1	0	0	0	0	1
1	1	**0**	0	0	1	0
1	**1**	0	0	1	1	**0**
1	1	**1**	0	1	**1**	0
1	1	1	0	1	1	0
1	1	0	0	**1**	0	0
1	1	**0**	0	1	0	0

Cells = leaves of phylogenetic tree

Idea: Find perfect phylogeny matrix ***M*** that is "closest" to measured mutation matrix ***M'***

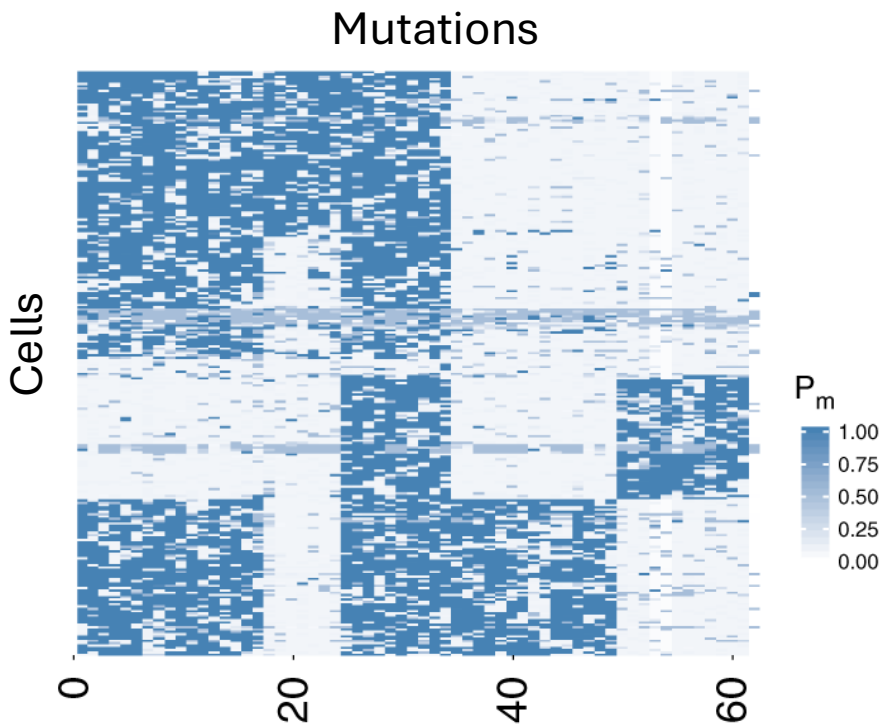
Phylogenetic imputation and error correction under perfect phylogeny model



- *Minimum*: NP-hard problem [Chen et al. 2006]
- *Probabilistic*: Compute $\Pr[T \mid Data]$ [OncoNEM (Ross et al. 2016)] by *marginalizing* over cell assignments and using MCMC [SCITE (Jahn et al. 2016), Sciϕ (Singer et al. 2018), …]

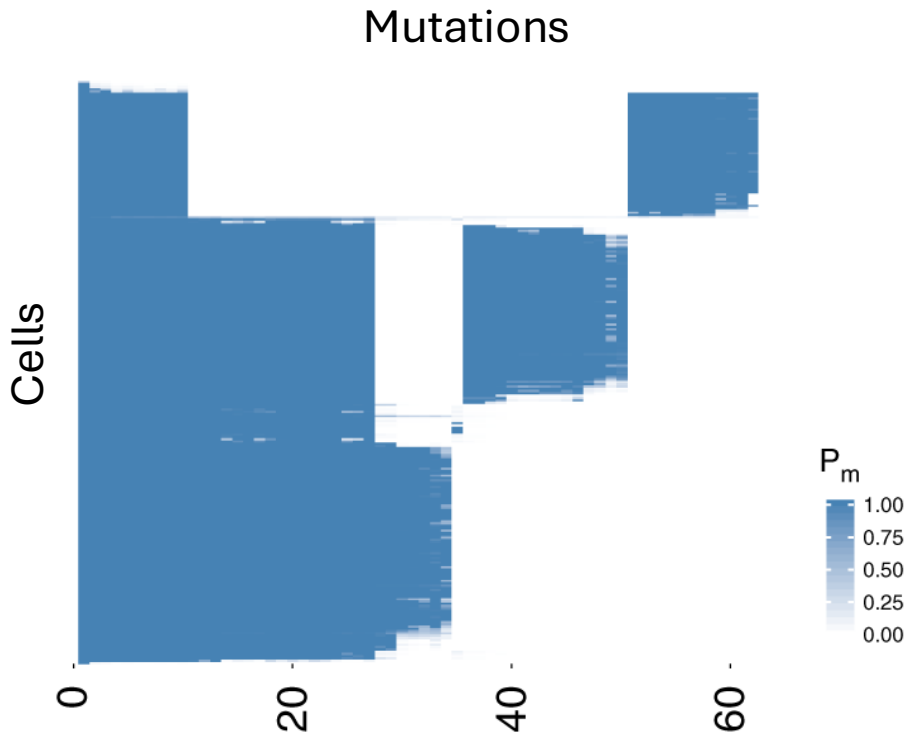
Single-cell sequencing of acute lymphoblastic leukemia (ALL)

Original mutation matrix
255 single cells from ALL



Sciϕ
[Singer et al. 2018]

Perfect phylogeny mutation matrix



**Note*: Cells and mutations rearranged from left figure

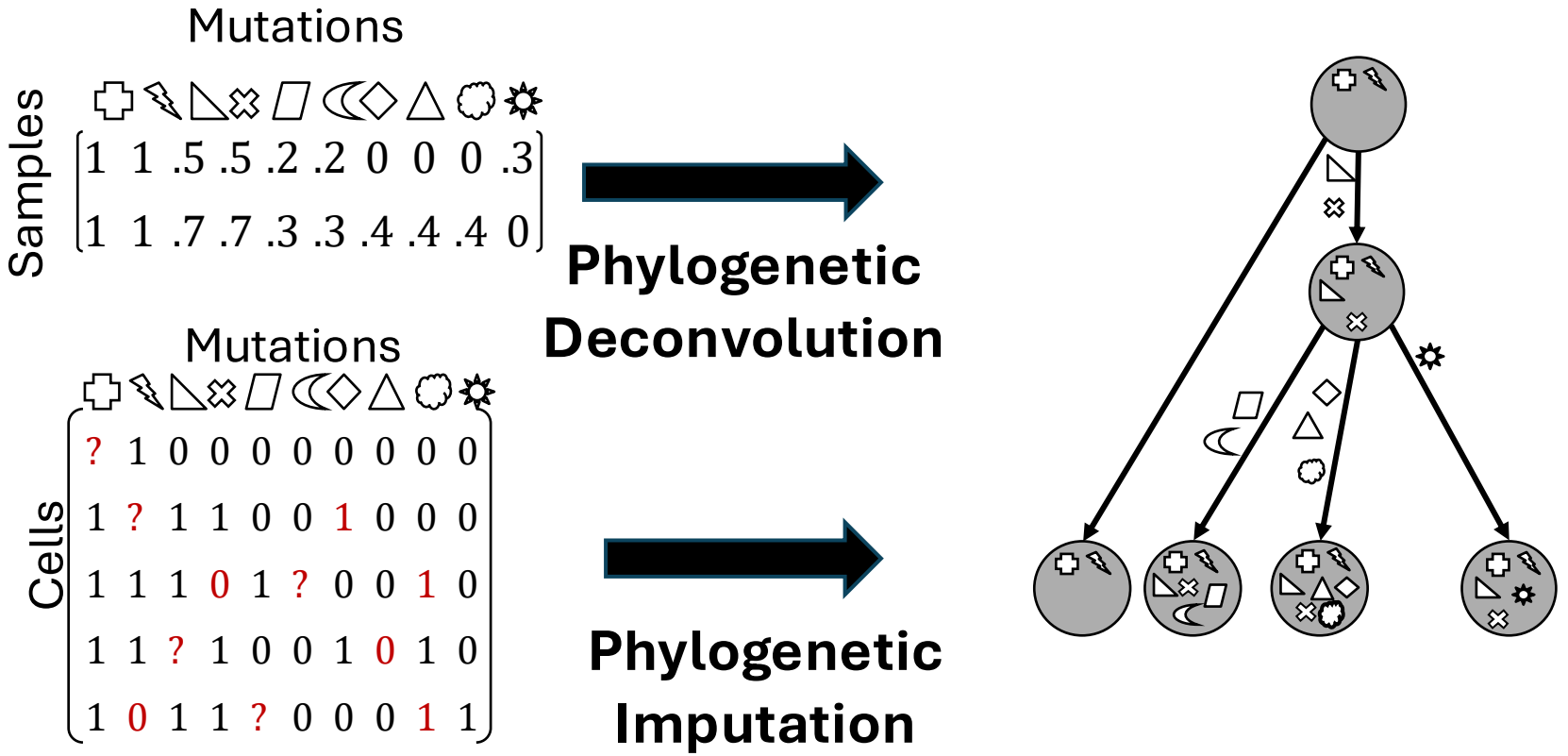
[Figure: Modified from Singer et al. 2018]

Phylogenetic imputation with other models

- Dollo
  - SPhyR [El-Kebir 2018] and SASC [Ciccolella et al. 2021]
  - ConDOR [Sashittal and R.. 2023]
- Finite states
  - SiFit [Zafar, et al. 2017]
  - CellPhy [Kozlov, …, Posada, 2022]
  - …

Summary

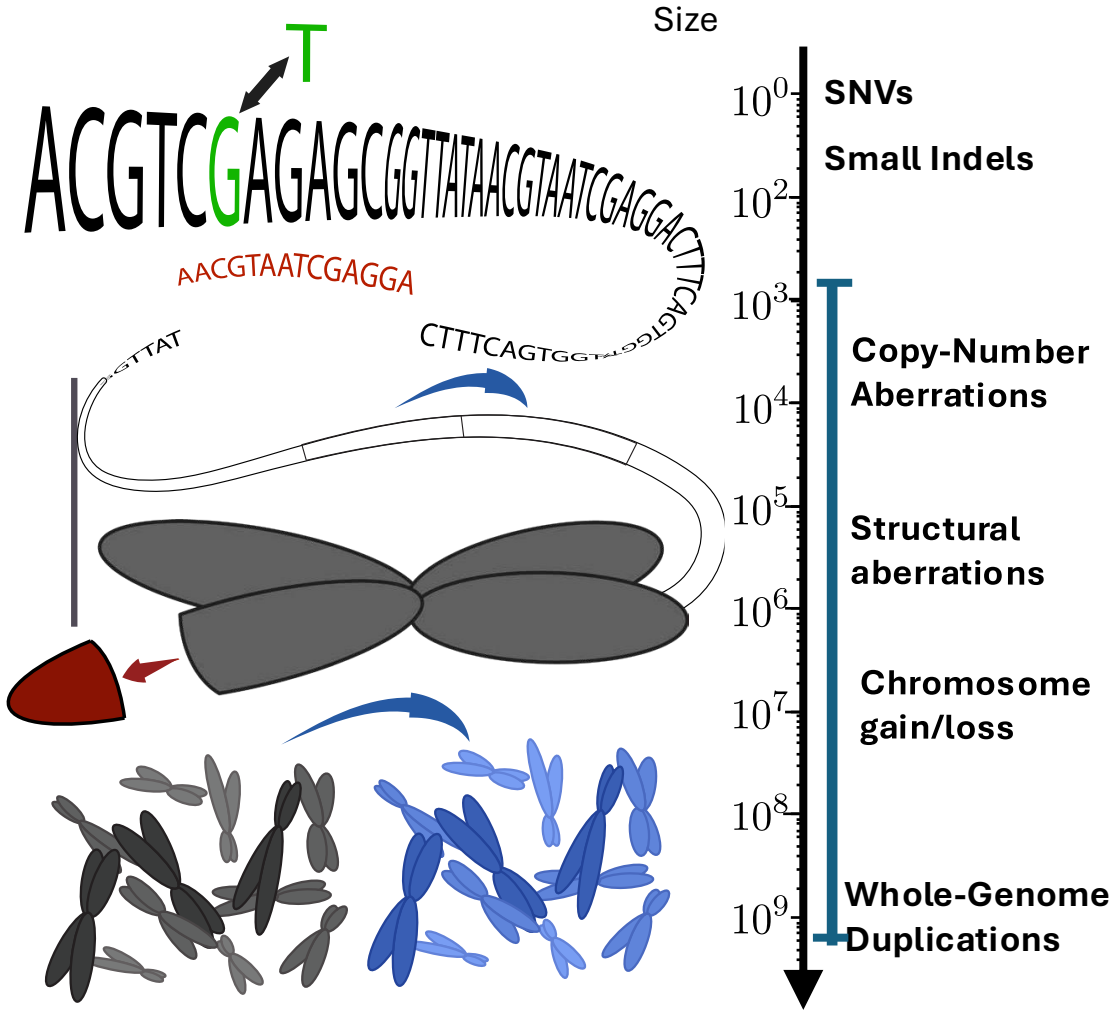
Perfect phylogeny model gives **constraints** for solving *underdetermined* **phylogenetic deconvolution** and **phylogenetic imputation** problems in cancer evolution.



Major challenges in cancer evolution

2. Cancer genomes are **complex**: somatic mutations occur over ***all*** genomic scales

→ Specialized evolutionary models/distances



Copy number aberrations and copy number profiles

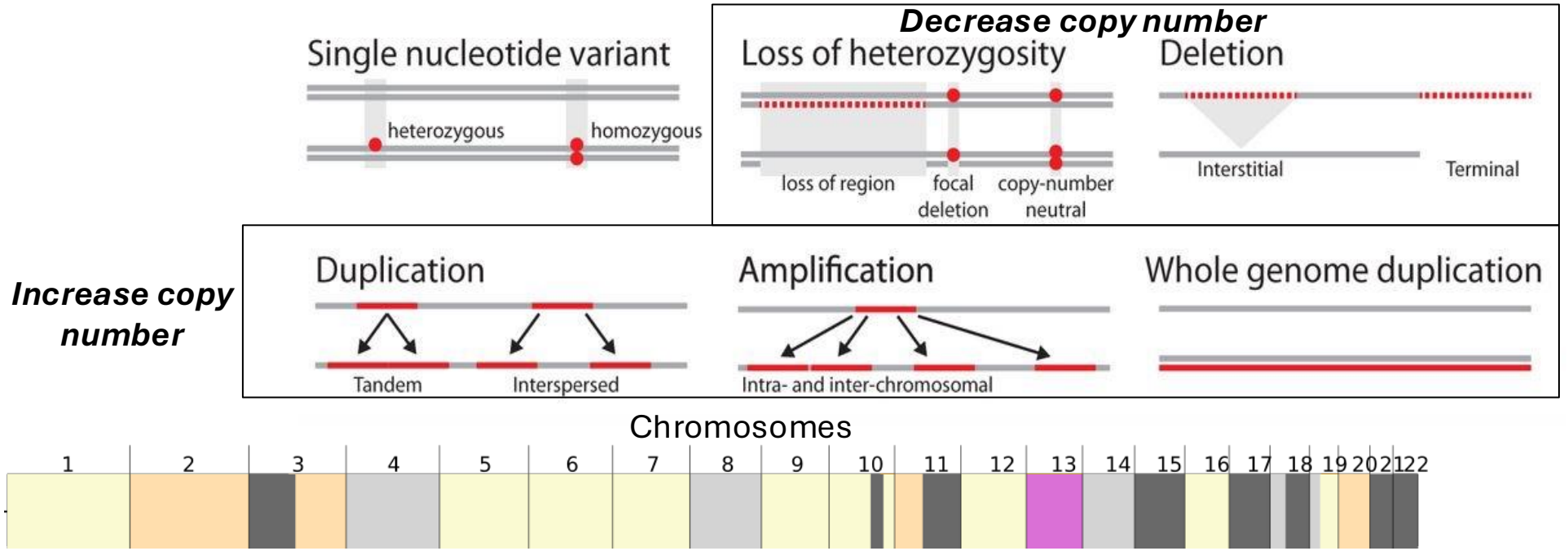
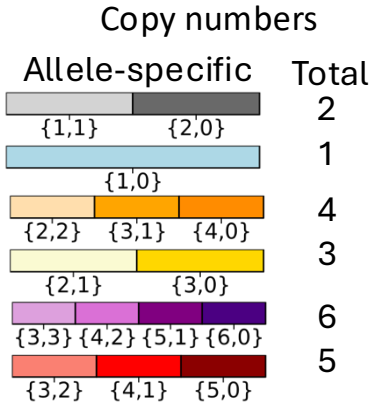


Figure: Beerenwinkel,, et al. *Systematic biology* (2015)

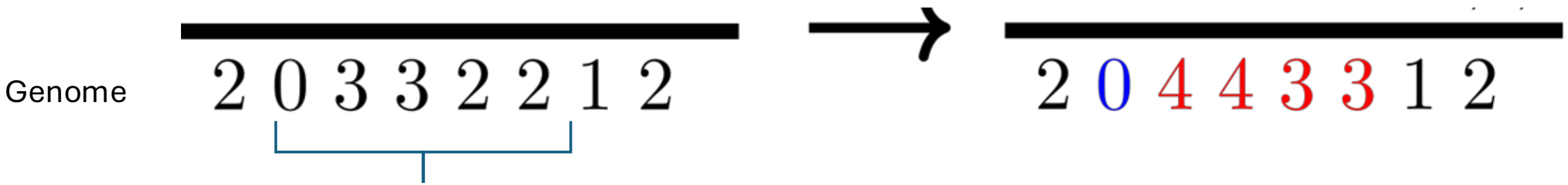


Copy number profile: a tuple $P = [a_1, a_2, \ldots, a_n]$ of *non-negative* integers where a_i = number of copies of segment i.

Allele-specific copy number profile a tuple $P = [(a_1, b_1), (a_2, b_2), \ldots (a_n, b_n)]$ of pairs of non-negative integers where a_i (resp. b_i) are the number of copies of segment i from one parent (resp. other parent)

Phylogenies with copy number aberrations

1. Copy number profile is altered by copy number aberrations (CNAs)



Duplication increases copies of *non-zero* entries in P

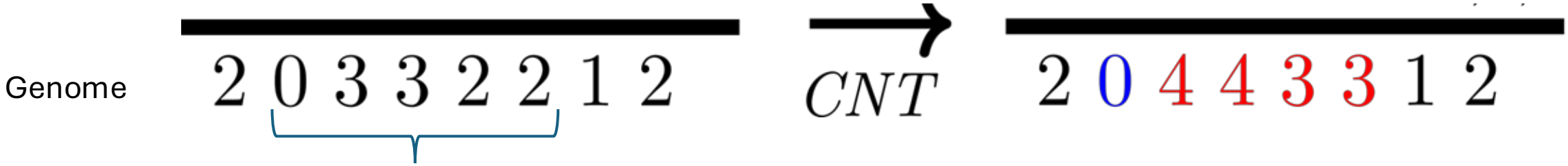
Entries in P do **not** evolve independently!

Breaks fundamental assumptions of most phylogenetic models

Phylogenies with copy number aberrations

1. Copy number profile is altered by copy number aberrations (CNAs)

Copy number transformation (CNT) model



$C_{2,6,+1}$ Transformations increase/decrease counts of *segments*

Given copy number profiles P and P', find a shortest sequence of CNTs $C_{i_1,j_1,s_1}, \ldots, C_{i_k,j_k,s_k}$ such that $C_{i_1,j_1,s_1} \circ \cdots \circ C_{i_k,j_k,s_k}\, P = P'$

Ron Zeira

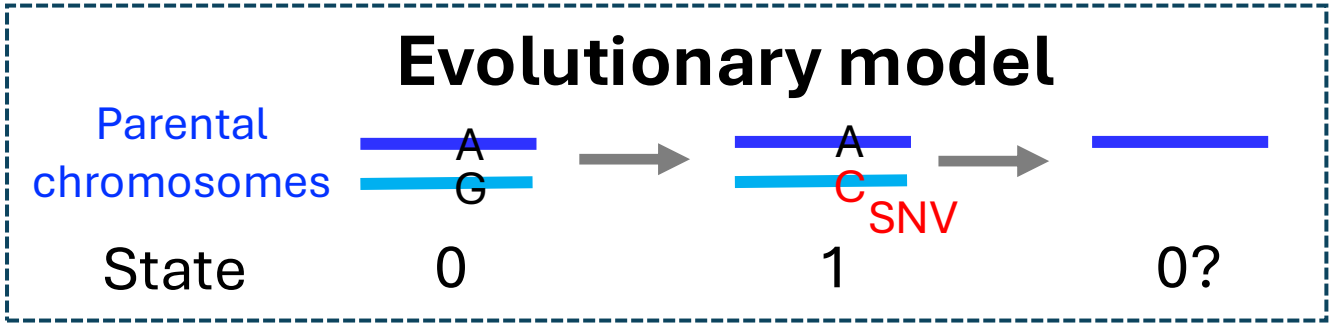
Mohammed El-Kebir

[Schwarz,, et al. *PLoS CB* (2014) ; Zeira et al. (2017, 2020}, El-Kebir, et al. (2017)]

- CNTs may overlap
- Every intermediate profile must be a valid copy number profile (non-negative integers)

Phylogenies with copy number aberrations

2. Incorporate CNAs with single-nucleotide variants (SNVs)



Deletions remove SNVs

Idea:

Relax perfect phylogeny model to allow *loss of mutations* (1 → 0 transitions)

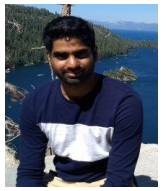
k-Dollo model: ≤ 1 gain (0→ 1), $\leq k$ loss (1→ 0) on T

Gryte Satas

...but **only when there is a deletion** in copy number profile

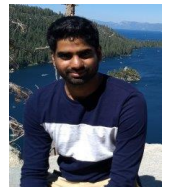
Loss supported phylogeny [Satas et al. *Cell Systems* 2020]

Constrained k-Dollo model [Sashittal et al. *Genome Biology* 2023]



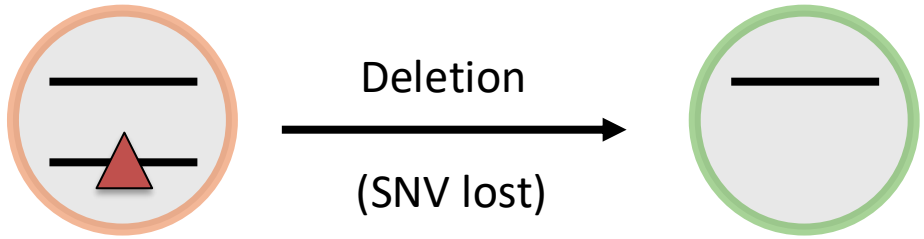
Palash Sashittal

Constrained Dollo Reconstruction (**ConDoR**)



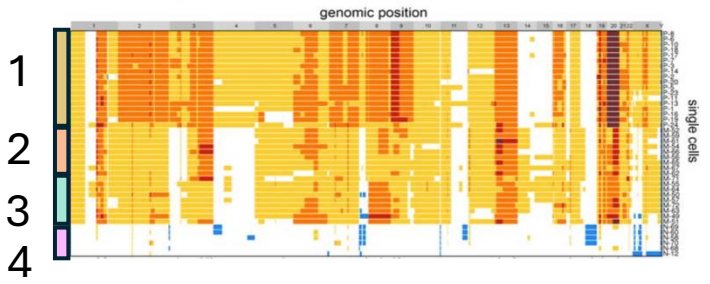
Palash Sashittal

Constrain losses to occur between clusters/clones



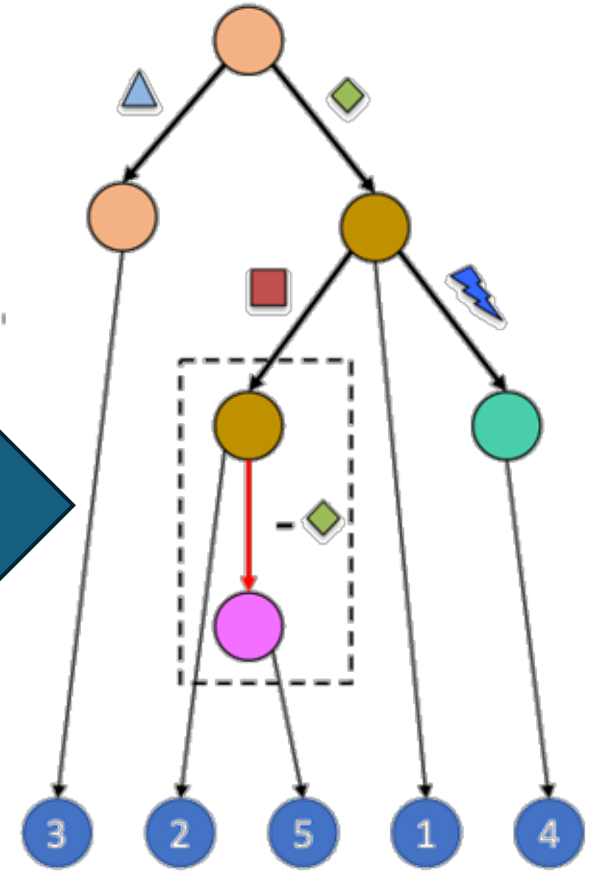
DNA sequencing

1. Identify SNVs
2. Cluster cells by copy number profiles



cells	mutations				mutations			
1	5	10	0	1	50	12	34	42
2	0	23	25	3	0	38	40	22
3	10	3	1	0	15	24	53	11
4	0	32	0	19	71	54	42	23
5	0	0	13	0	2	33	29	40
	Variant read counts				Total read counts			

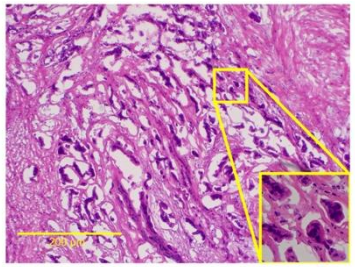
CONDOR



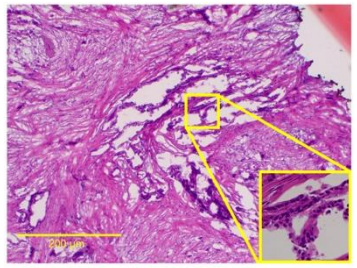
Constrained k-Dollo Phylogeny

[Sashittal et al. *Genome Biology* 2023]

ConDoR on single-cell pancreatic cancer

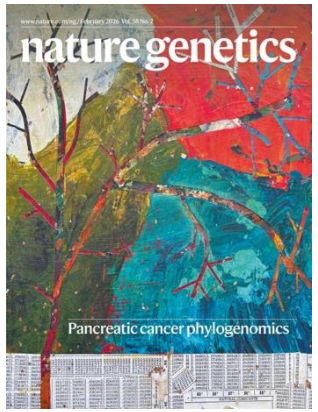


region S1
(987 cells)



region S2
(1166 cells)

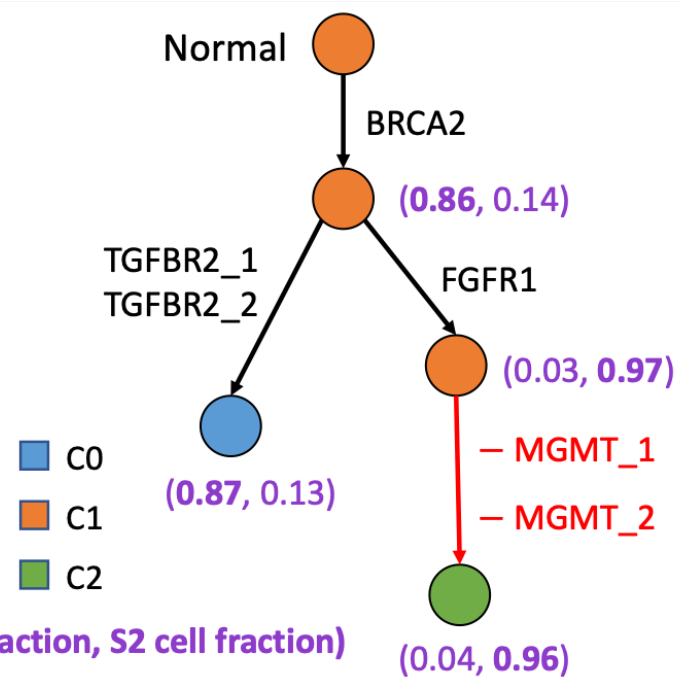
Targeted DNA sequencing
MissionBio Tapestri

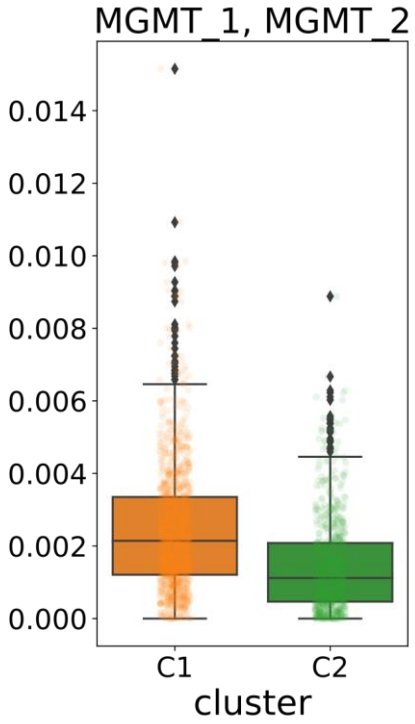


Zhang, Sashittal et al.
Nature Genetics
February 2026

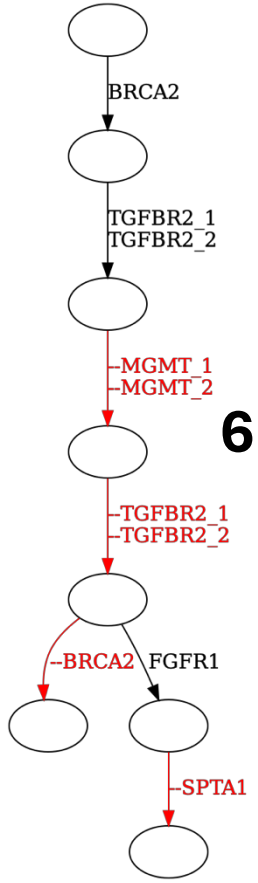
Memorial Sloan Kettering Cancer Center
Dr. Christine Iacobuzio-Donahue
Haochen Zhang

ConDoR
(Constrained *k*-Dollo)





SPhyR
(*k*-Dollo)



El-Kebir 2018,
Bioinformatics

6 mutation losses

[Sashittal et al. *Genome Biology* 2023]

Summary

Cancer evolution presents variations of classical phylogenetic problems

1. Measurement challenges → **Phylogenetic deconvolution** and **phylogenetic imputation** problems
2. Specialized models for complex mutations: **copy number aberrations**

Future Directions

1. From phylogenetic to population genetic models: selection, fitness, sampling
2. Whole-genome models of cancer evolution: whole-genome duplications, ecDNA, etc.
3. Epigenetic evolution

Models and Algorithms for Cancer Evolution

Part 2

Ben Raphael

Department of Computer Science

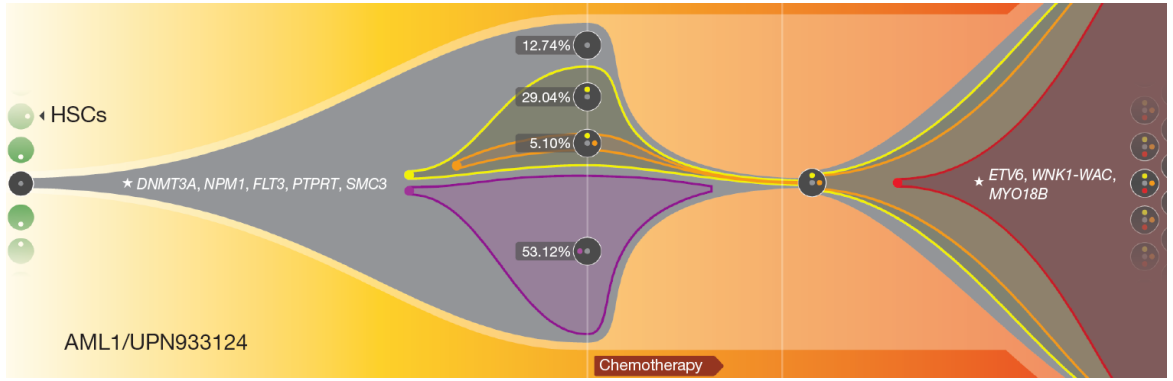
PRINCETON UNIVERSITY

IPAM: Mathematics of Cancer

February 24,2026

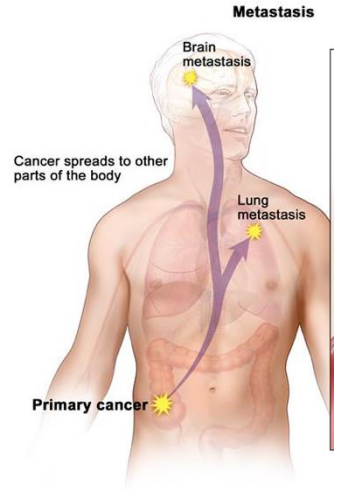
Outline

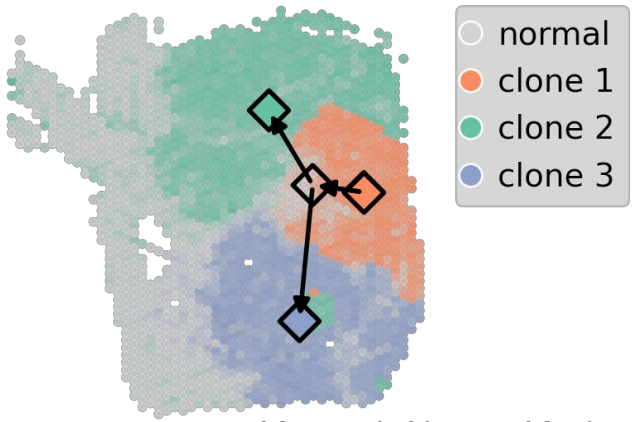
1. Constructing phylogenies from bulk and single-cell cancer sequencing



Ding et al. *Nature* 2012

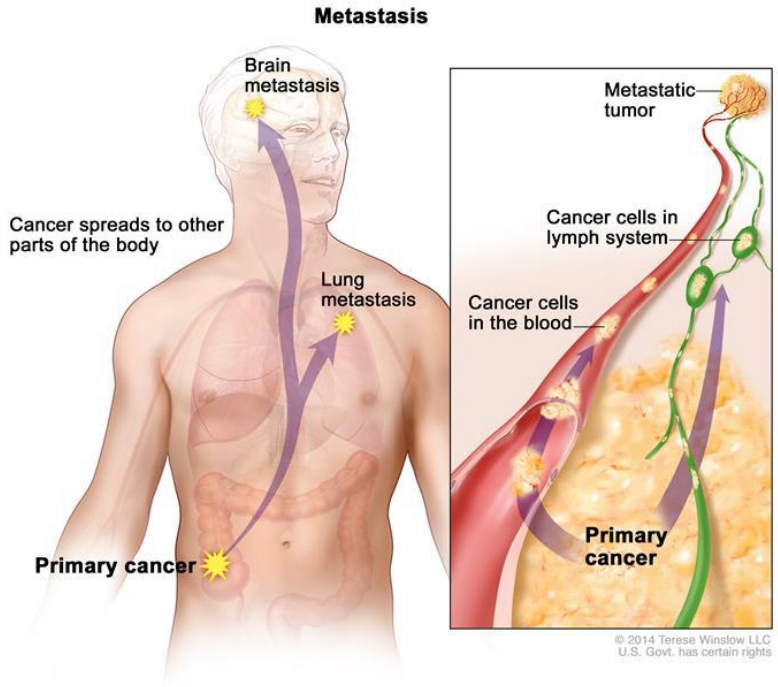
2. **Migration: Metastasis and Spatial tumor evolution**

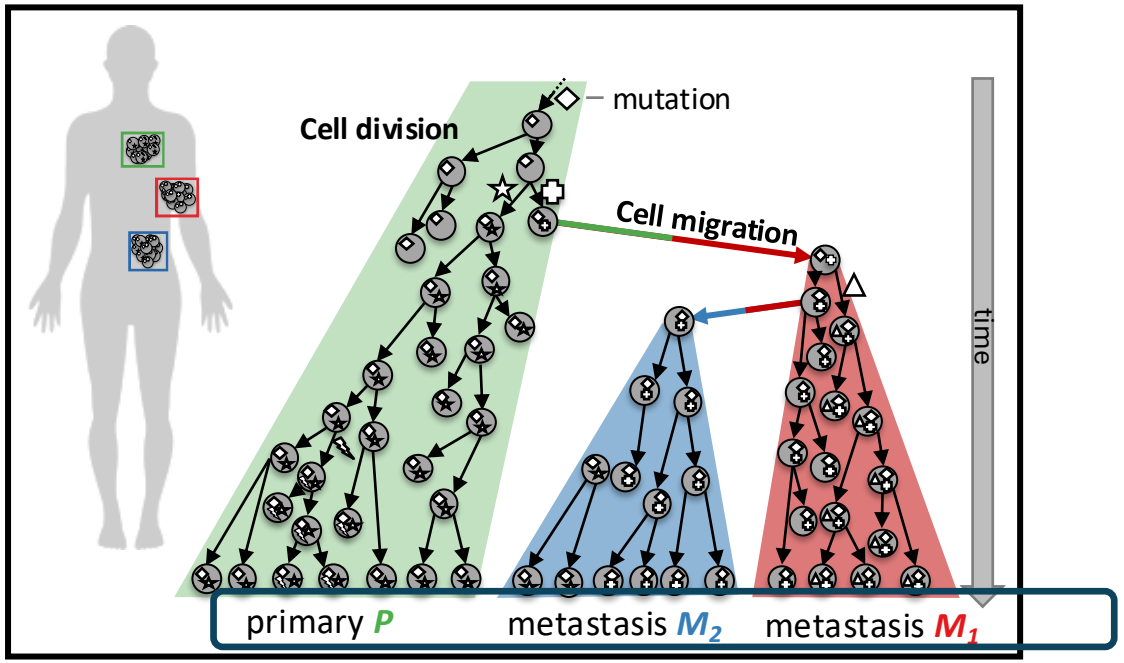




Ma et al. *Nature Methods* 2024

When and how do cancer cells metastasize to seed tumors at distant locations?

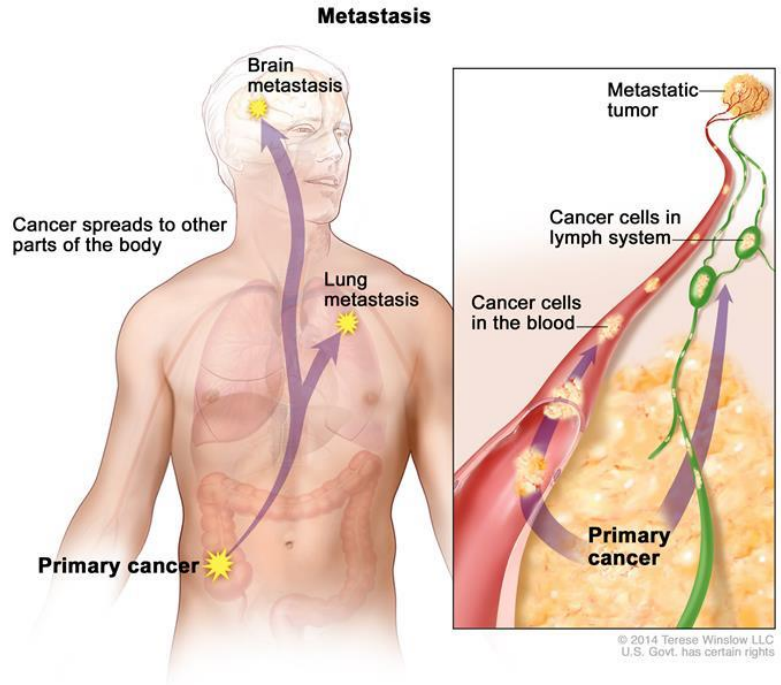


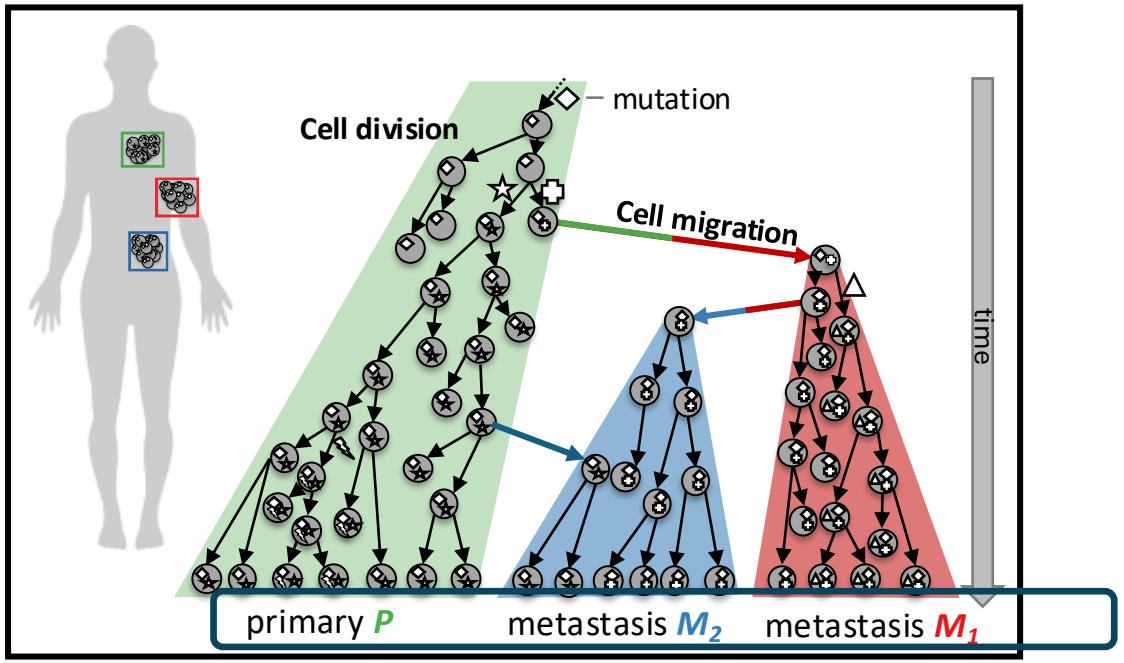


For solid tumors, metastasis is responsible for 90% of deaths

[Chaffer and Weinberg, *Science* 2011]

When and how do cancer cells metastasize to seed tumors at distant locations?





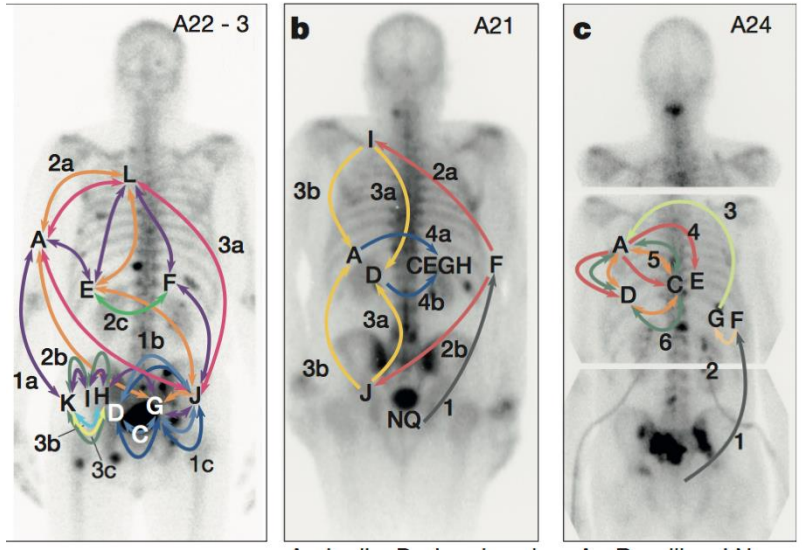
Does a metastasis arise from a single migration (***monoclonal***) or multiple migrations (***polyclonal***)?

Anatomical locations of ancestral cells are unknown → Infer the ***past***

Polyclonal/Multiclonal/Re- Seeding of Metastases?

Prostate Cancer

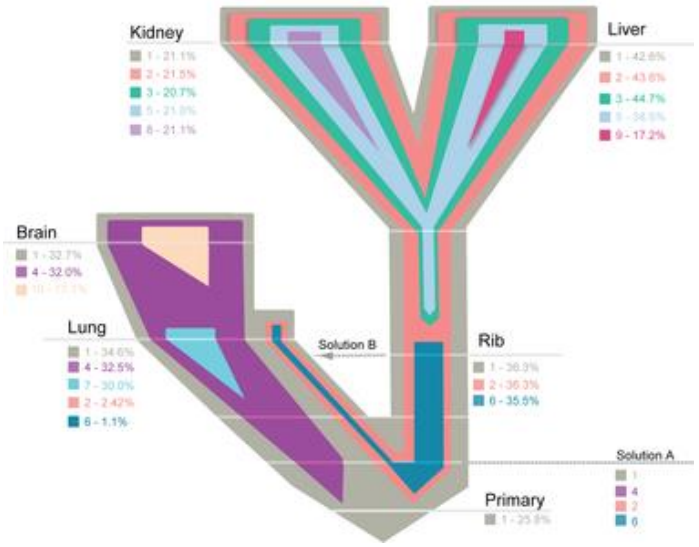
[Gundem et al. (2015) *Nature*]



*"..in all five patients with **poly-clonal seeding**, subclones...were found to have **re-seeded** multiple sites."*

Breast Cancer

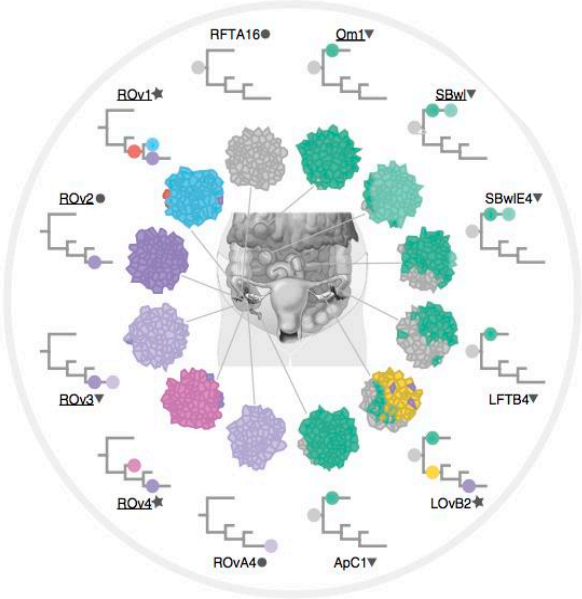
[Hoadley KA, et al. (2016) *PLOS Medicine*]



*"**multiclonal seeding** from the primary tumor to the metastases can occur..."*

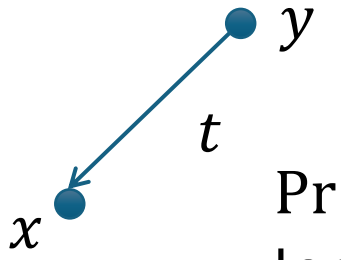
Ovarian Cancer

[McPherson et al. (2016) *Nature Genetics*]



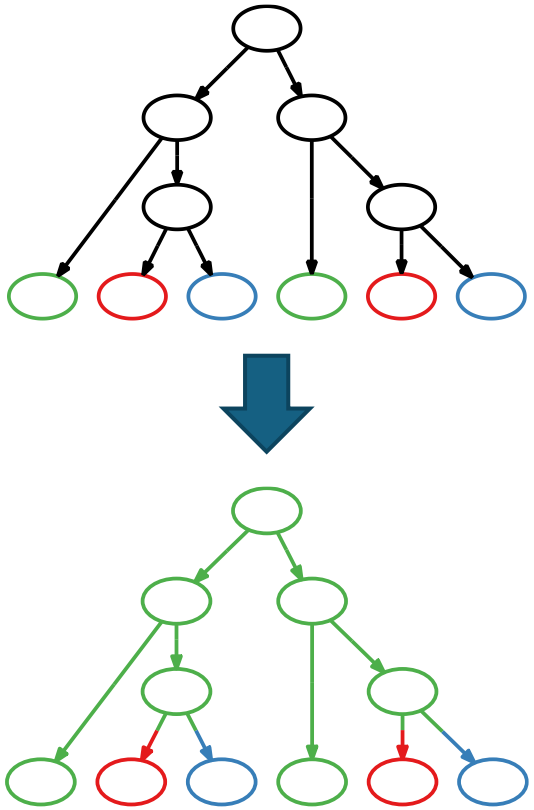
*"multiple samples composed of divergent lineages is concordant with **polyclonal reseeding** or **polyclonal migration**..."*

Maximum Likelihood Migration Inference



Anatomical location is a single phylogenetic character (color)

$\Pr[\, x \mid y, t]$ = probability that cell migrates from location y to x in time t: $= e^{Qt}$ where Q is rate matrix



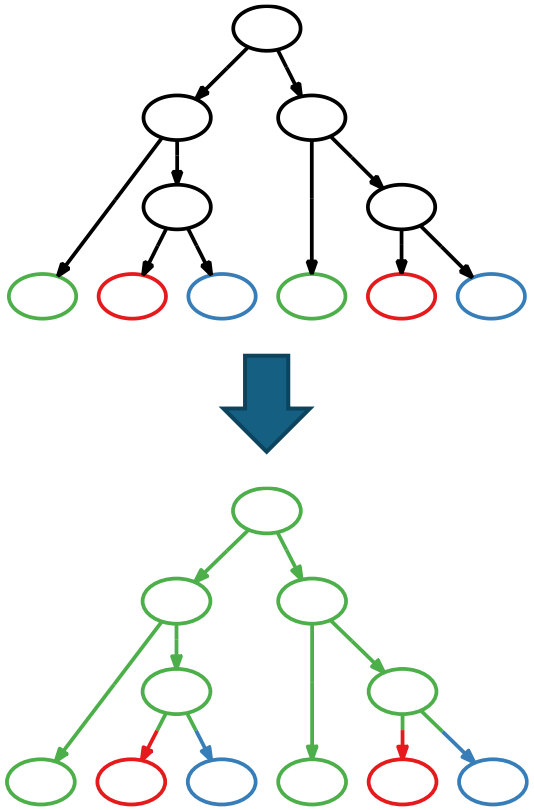
Most likely ancestral reconstruction : Given a leaf-labeled cell/clone tree $T = (V, E, \boldsymbol{b})$, find a label $l(v)$ for each internal vertex that ***maximizes*** $\sum_{(u,v)\in E} \log \Pr[l(v)|l(u), b(e)]$

Simplification:

$$s(x \mid y) = -\log \Pr[x|y, b(e)] = \begin{cases} 0, \text{if } x = y \\ 1, \text{otherwise} \end{cases}$$

Most Parsimonious History Problem

Given a leaf-labeled cell/clone tree ***T***, label each vertex with an anatomical location (color) that ***minimizes* number μ** of migrations.



Formally: Compute $\mu^* = \min_{\ell} [\mu[\ell]]$, where $\mu[\ell]$ = #migrations for vertex labeling ℓ

Special case of ***small parsimony problem***: solve via Sankoff-Rousseau recurrence

$M_v(a)$ = number of migrations at subtree rooted at v, if v is labeled with location a

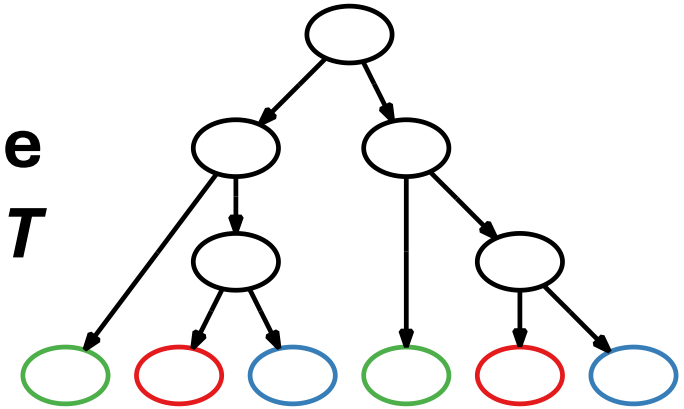
$$M_v(a) = \sum_{w \in C(v)} \min_b (M_w(b) + s(b \mid a))$$

where $C(v)$ = children of v

Maximum Parsimony for migrations [Slatkin and Maddison *Genetics* (1989)]

Ancestral Labeling → Migration history → Migration graph

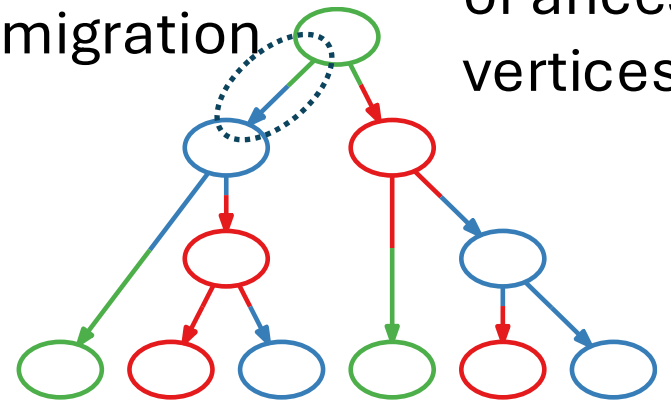
Clone Tree *T*



Maximum likelihood
Maximum parsimony

migration

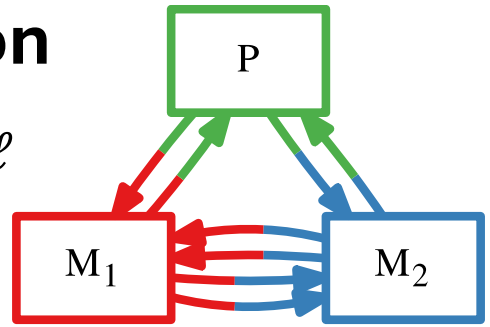
Labeling $\ell(T)$ of ancestral vertices



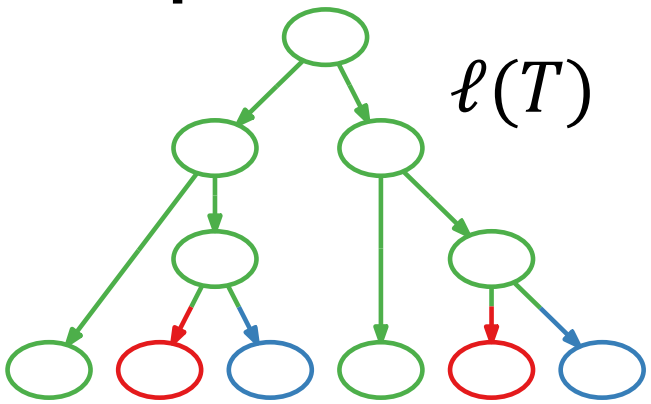
A labeling $\ell: V(T) \to S$ of vertices of rooted tree T induces a **directed multigraph** $G_\ell = (S, E)$ where $E = \{(\ell(u), \ell(v)) : (u, v) \in T\}$

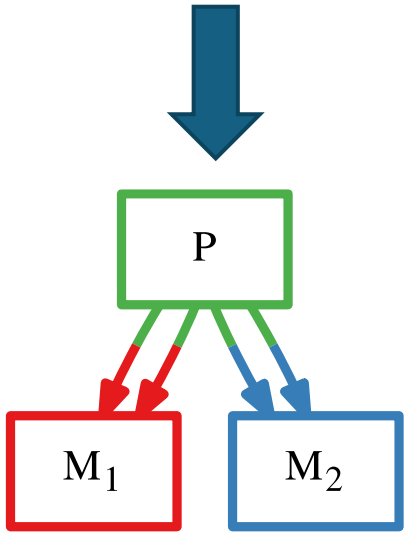
μ = 8 migrations

Migration graph G_ℓ

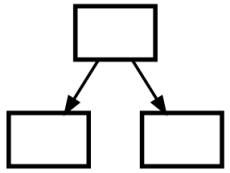
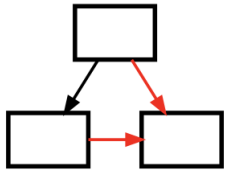
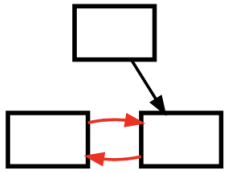
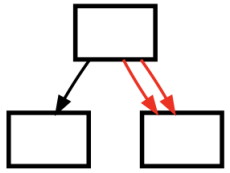
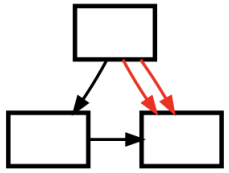
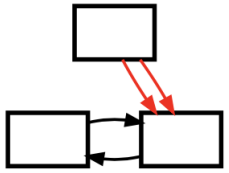


Migration graphs provide taxonomy of migration patterns





Migration Graph

	single-source seeding (S)	multi-source seeding (M)	reseeding (R)
monoclonal (m)	 tree **mS**	 directed acyclic graph **mM**	 directed graph **mR**
polyclonal (p)	 multi-tree **pS**	 directed acyclic multi-graph **pM**	 directed multi-graph **pR**

nature genetics

Analysis | Published: 26 April 2018

Inferring parsimonious migration histories for metastatic cancers

Mohammed El-Kebir, Gryte Satas & Benjamin J. Raphael

Additional constraints on migration patterns?

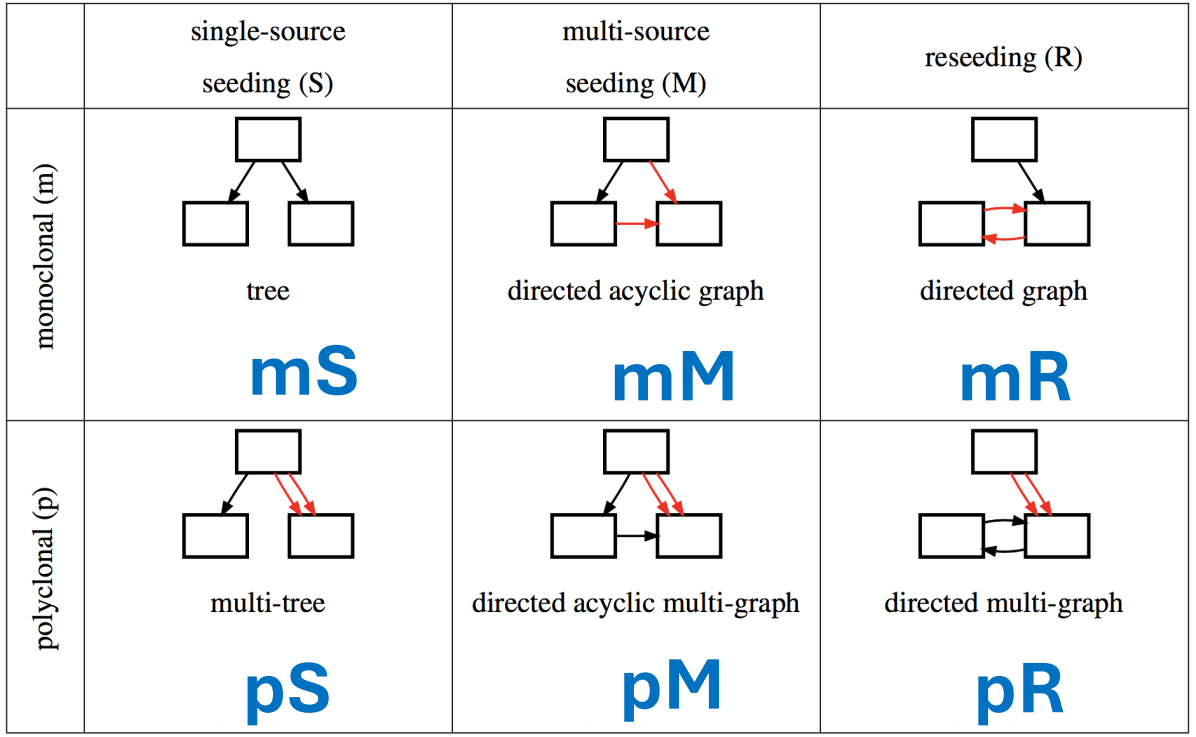
Clusters of cancer cells migrate simultaneously through the blood stream [Cheung et al., 2016]

→ Parallel edges more likely? ("comigrations" [El-Kebir et al. 2018])

Cell returning to site of origin may be

→ More likely?: "*seed and soil*" hypothesis

→ Less likely?: Needs to grow a subpopulation to be sampled



nature genetics

Analysis | Published: 26 April 2018

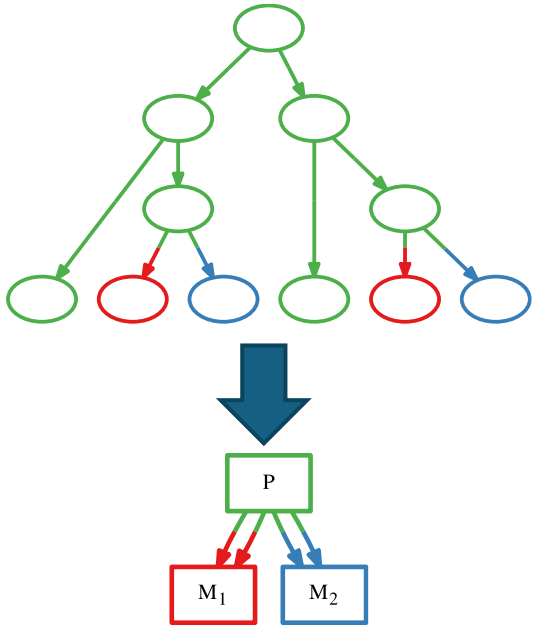
Inferring parsimonious migration histories for metastatic cancers

Mohammed El-Kebir, Gryte Satas & Benjamin J. Raphael

Parsimonious **Constrained** Migration History Problem

G_ℓ is migration graph induces by labeling ℓ

Let $\mathcal{G}(S)$ be set of *directed* graphs with vertex set S



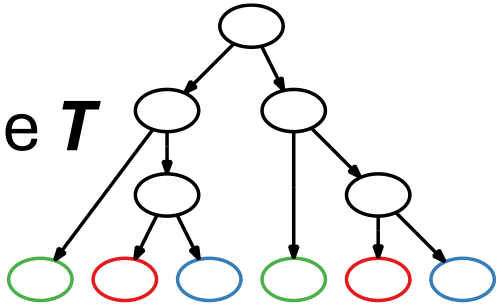
Given a rooted tree ***T*** with leaves labeled by S, and a subset $\mathcal{G} \subseteq \mathcal{G}(S)$, find a labeling ℓ of ancestral vertices that ***minimizes*** $\min_{\ell: G_\ell \in \mathcal{G}} \mu(\boldsymbol{l})$

$\mathcal{G} = \mathcal{G}(S)$ is unconstrained maximum parsimony → solved efficiently by Sankoff-Rousseau recurrence

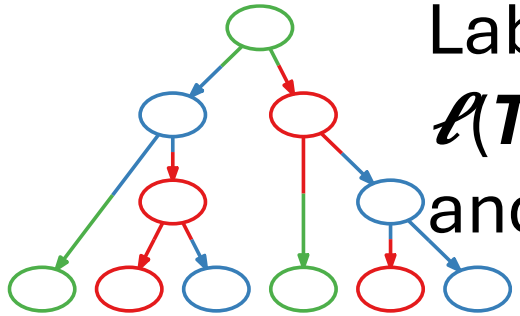
How to solve if $\mathcal{G}$ is set of trees on S, or $\mathcal{G}$ is set of DAGs on S?

MACHINA: Parsimonious constrained migration

Clone Tree $\boldsymbol{T}$



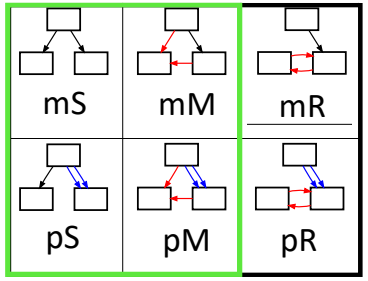
Integer linear program



Labeling $\ell(T)$ of ancestors

Minimum migrations μ^*

Subset $\mathcal{G}$ of migration patterns



Alternatively, a frequency matrix F

Mutations

Samples

$$\begin{bmatrix} 1 & 1 & .5 & .5 & .2 & .2 & 0 & 0 & 0 & .3 \\ 1 & 1 & .7 & .7 & .3 & .3 & .4 & .4 & .4 & 0 \end{bmatrix}$$

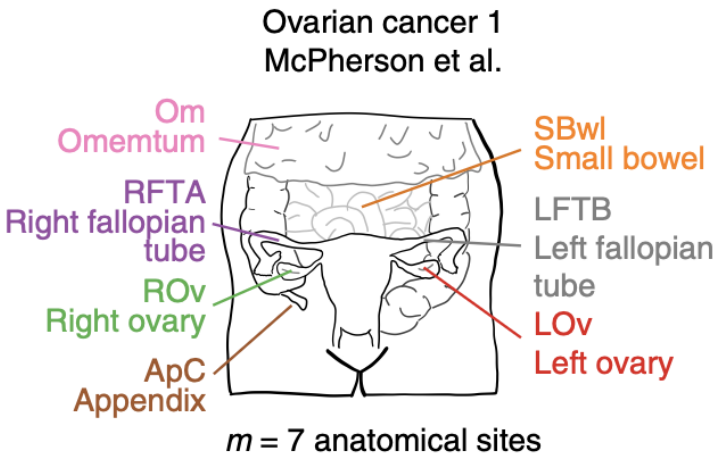
MACHINA [El-Kebir, Satas, R. *Nature Genetics* 2018]

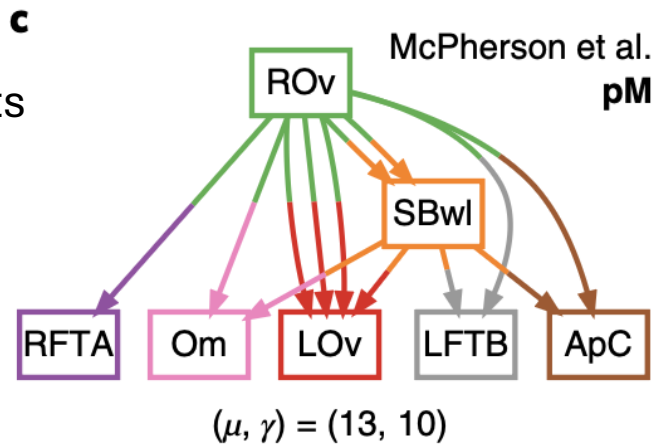
Mohammed El-Kebir

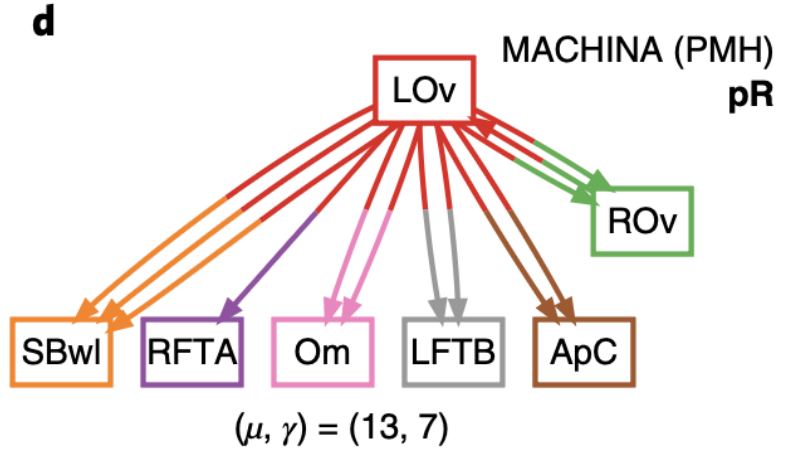
Gryte Satas

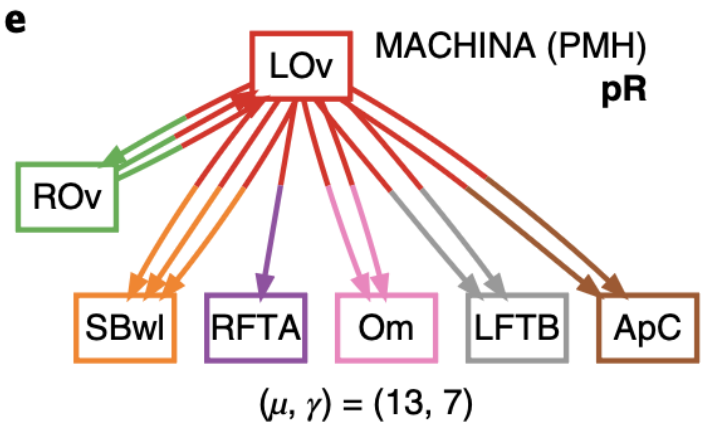
Ovarian Cancer Metastasis

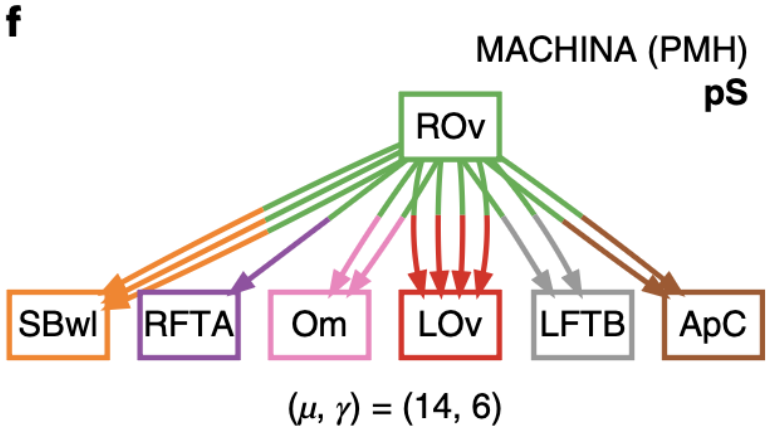
DNA sequencing of ovarian cancer mets
[McPhearson et al. Nature Genetics 2016]









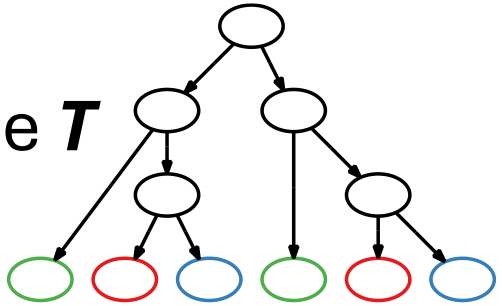


MACHINA [El-Kebir, Satas, R. *Nature Genetics* 2018]

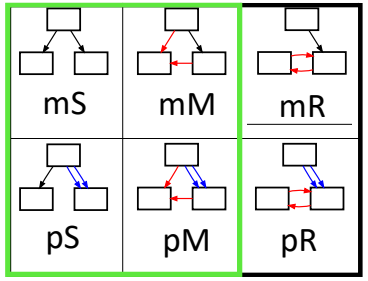
μ = number migrations γ = number comigrations

MACHINA: Parsimonious constrained migration

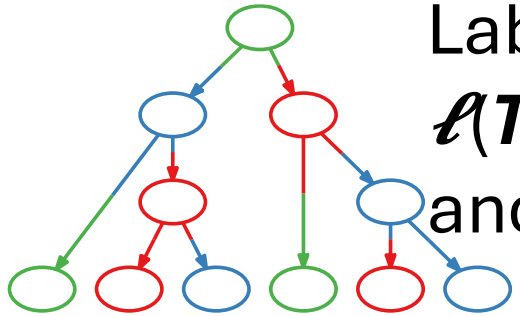
Clone Tree T



Subset $\mathcal{G}$ of migration patterns



Integer linear program



Labeling $\ell(T)$ of ancestors

Minimum migrations μ^*

MACHINA [El-Kebir, Satas, R. *Nature Genetics* 2018]

Mohammed El-Kebir

Gryte Satas

Alternatively, a frequency matrix F

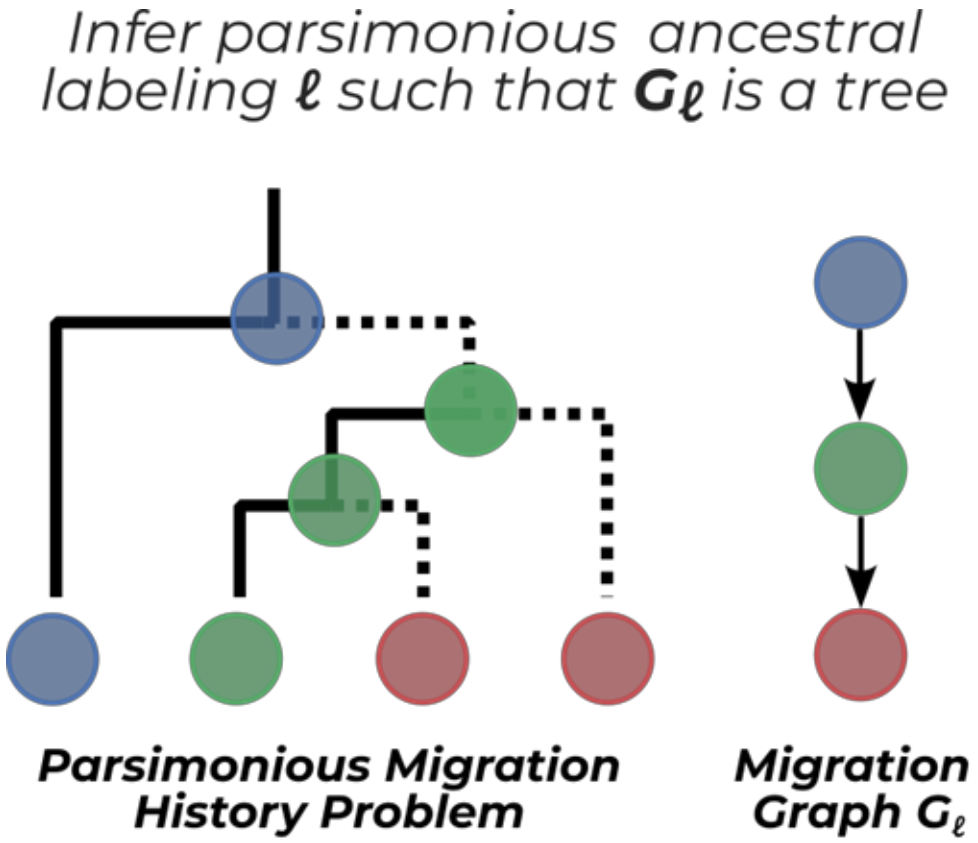
Mutations

Samples

$$\begin{bmatrix} 1 & 1 & .5 & .5 & .2 & .2 & 0 & 0 & 0 & .3 \\ 1 & 1 & .7 & .7 & .3 & .3 & .4 & .4 & .4 & 0 \end{bmatrix}$$

Metient [Koyyalagunta , Ganesh, Morris, *Nature Methods* 2025]
fastMACH [Schmidt and R. RECOMB 2025, *Cell Systems*, in press]
MACH2 M.S. Roddur, … El-Kebir. RECOMB 2025]
SMiTH [Kuzmin et al. *Nature Comm.* 2025)

Constrained ancestral reconstruction problems often cannot be solved with dynamic programming algorithms

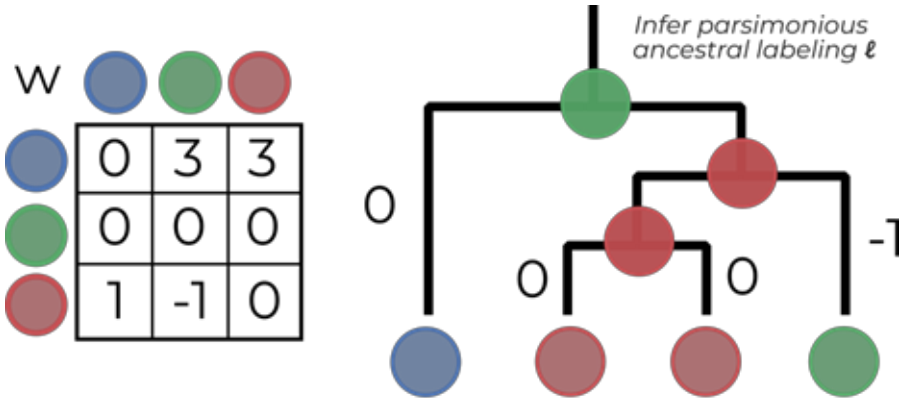


Global constraints on the ancestral labeling ***break*** the local optimality required for Sankoff-Rousseau recurrence

$$M_v(a) = \sum_{w \in C(v)} \min_b (M_w(b) + s(b \mid a))$$

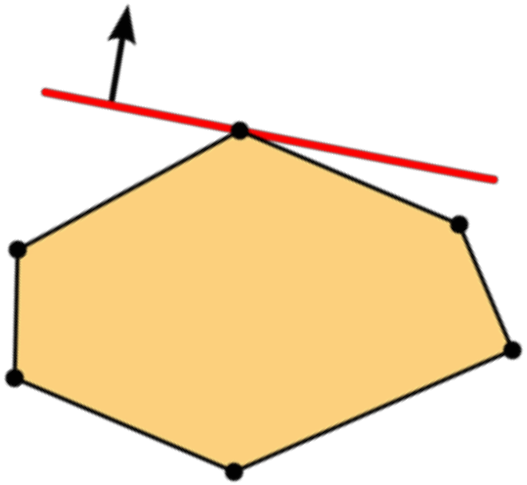
where $C(v)$ = children of v

A linear program for ancestral reconstruction problem



Small Parsimony Problem

Theorem. *Polynomial sized linear programming formulation* →

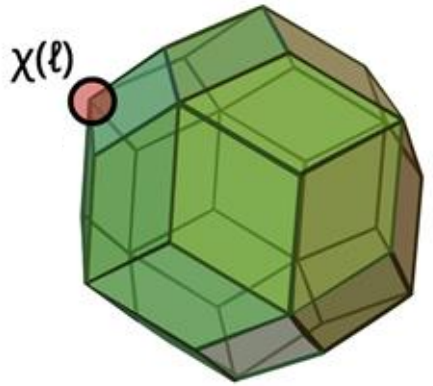


Henri Schmidt

Theorem (Informal). *Most parsimonious/likely ancestral labeling of T is solved by a linear program containing* $O(nm^2)$ *variables and constraints* [n = # *leaves,* m = # anatomical locations].

Key Technical Ideas:

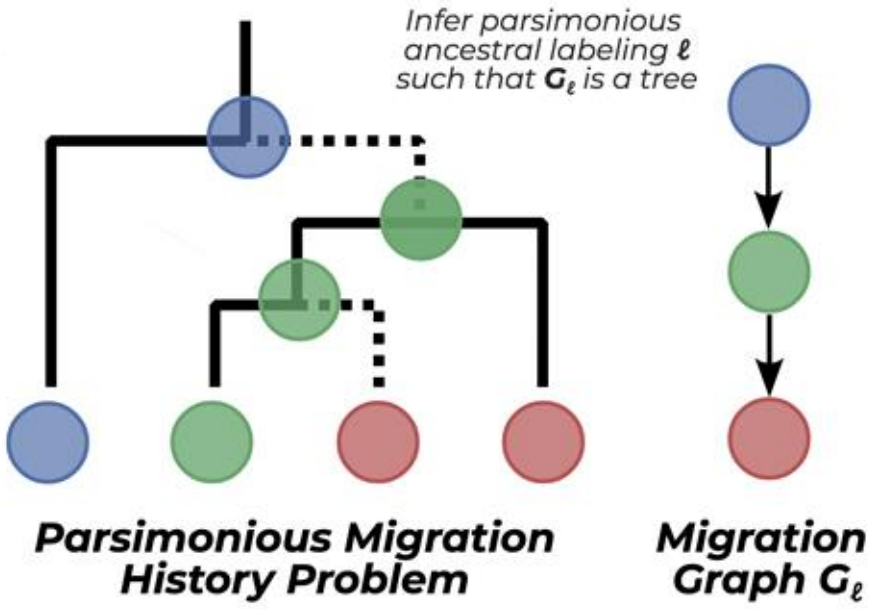
- Introduce the *tree labeling polytope P:*
  - *v*ertices ***χ(ℓ)*** of P ←→ labelings ***ℓ*** *of vertices of T.*
- Use a primal-dual packing argument to show that the tree labeling polytope P has a polynomial-sized description



Tree Labeling Polytope

Schmidt and R. RECOMB 2025 and *Cell Systems* (in revision)

Extend to mixed integer linear program (MILP) for constrained migration history problem (fastMach)



MILP

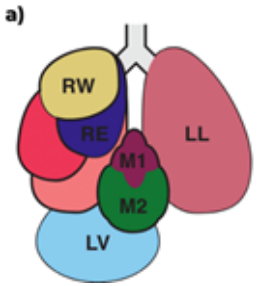
$$x_{e,i,j} = 1$$
$$y_{i,j} = 1$$
$$\min_{x_{e,i,j} \geq 0, y_i \in \{0,1\}} \sum_{\substack{e,i,j \\ e \neq (r',r)}} w(e,i,j) \cdot x_{e,i,j}$$
$$\text{s.t.} \quad x_{e,i,j} = 0 \quad \text{if} \quad w(e,i,j) = \infty,$$
$$Ax = b,$$
$$x_{e,i,j} \leq y_{i,j} \quad \text{for all } e \in E(\mathcal{T}), i, j \in \Sigma$$
$$\sum_{i,j} y_{i,j} = m - 1.$$

- ***Fixed parameter tractable*** with only m^2 binary variables (m = # anatomical sites)
- ***Strong linear relaxations***, quantified in terms of the Lagrangian relaxation

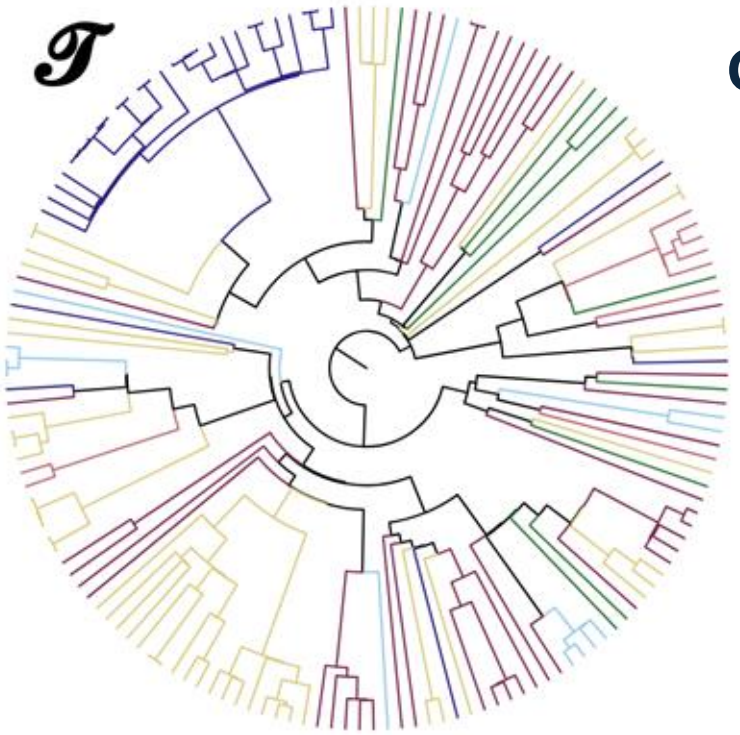
Contrasts state-of-the-art algorithm MACHINA (El-Kebir et al. 2018), which has $O(nm)$ binary variables and worse linear relaxations.

Schmidt and R. RECOMB 2025 and *Cell Systems* (in revision)

Application: Migration history inference on CRISPR lineage tracing of mouse lung xenograft

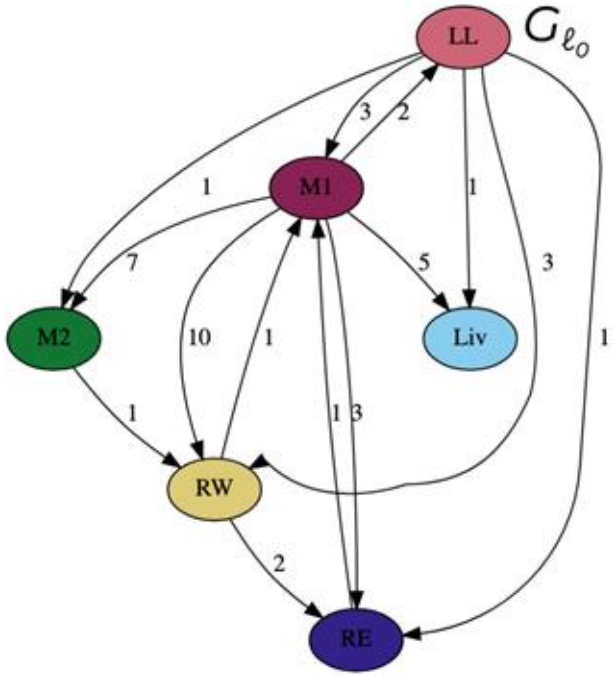
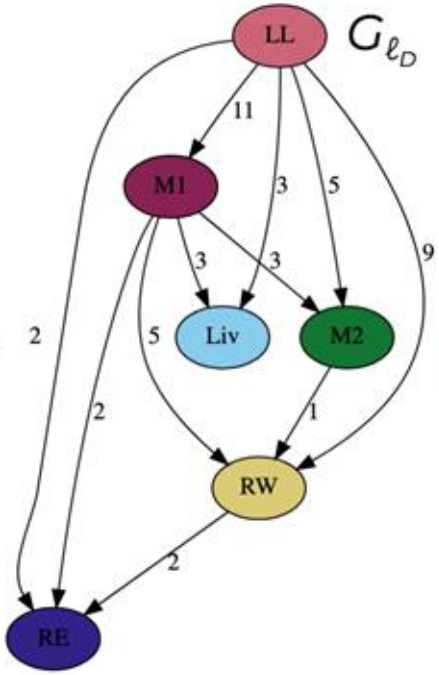
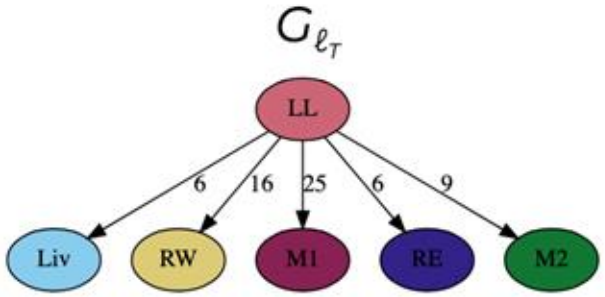


Inferred CP28 Migration Graphs



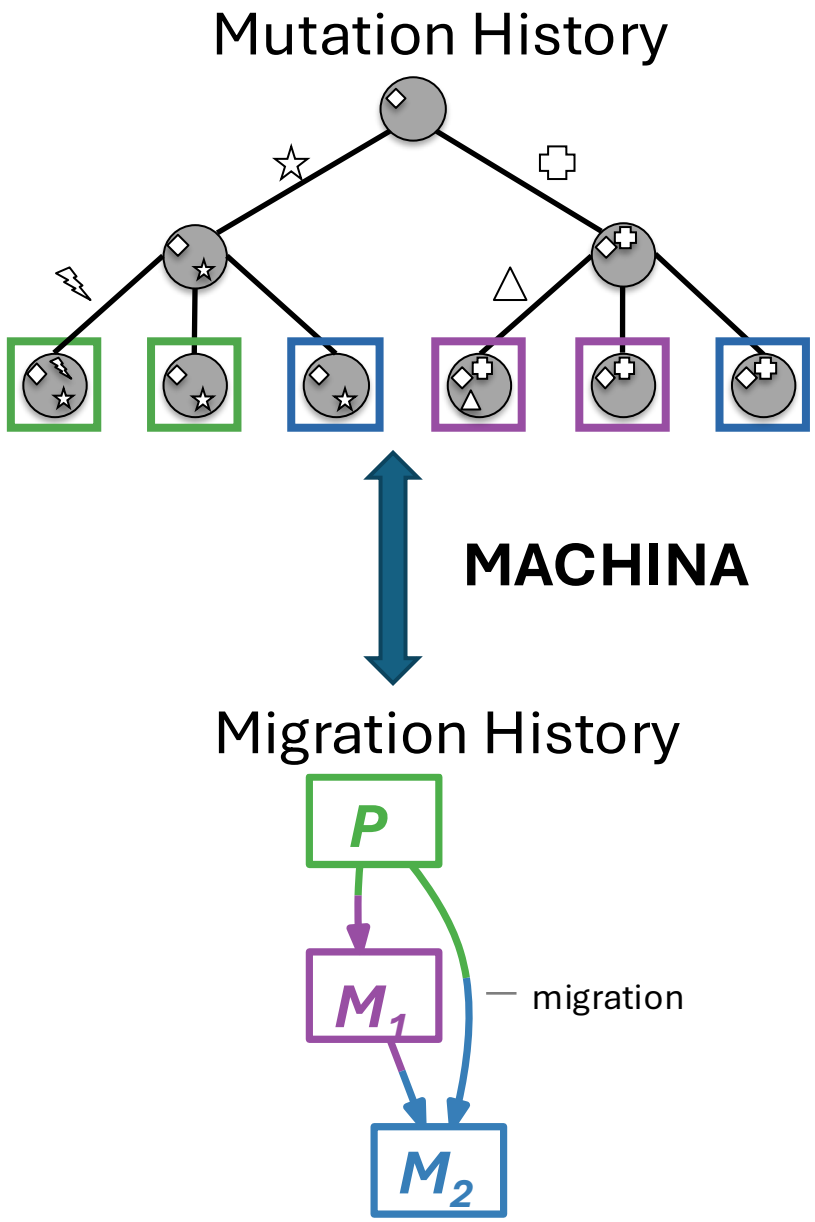
CP28 Lineage Tree (180 cells)

Quinn et al. *Science* 2018

Structural Constraints:	*None*	*DAG*	*Tree*
			
Number of Migrations:	*38*	*41*	*61*

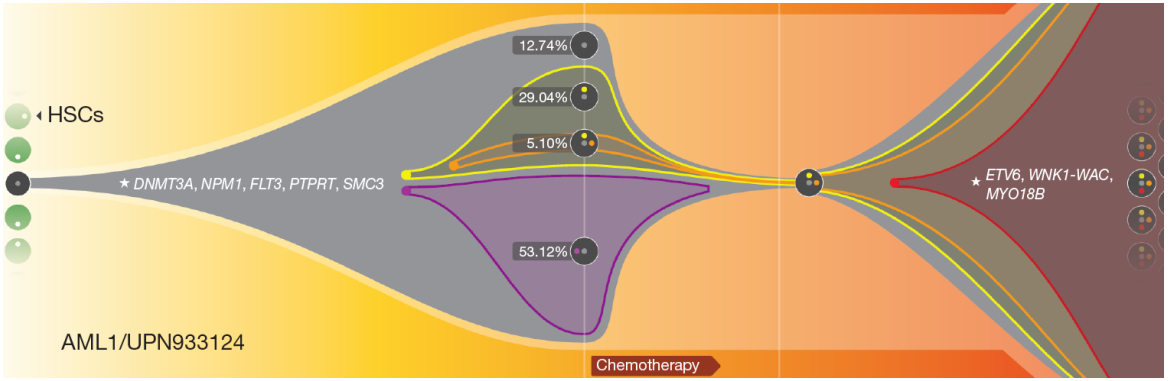
Summary

- Infer migration history of metastatic tumors with **maximum likelihood/parsimony**
- MACHINA [El Kebir, Satas, R. 2018]: Maximum parsimony with (global) constraints on observed migrations
- Derived a **polynomial-sized** linear programming algorithm for the small parsimony problem based on **tree labeling polytope** [Schmidt and R. 2025]
- fastMACHINA: MILP for parsimonious migration history problem studied in MACHINA [Schmidt and R. 2025]



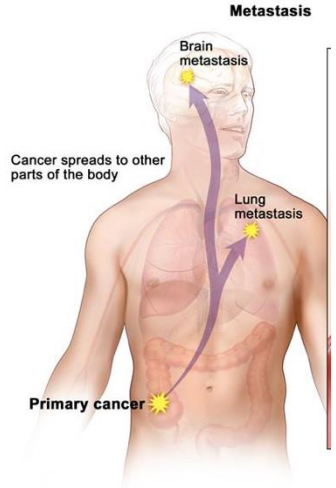
Outline

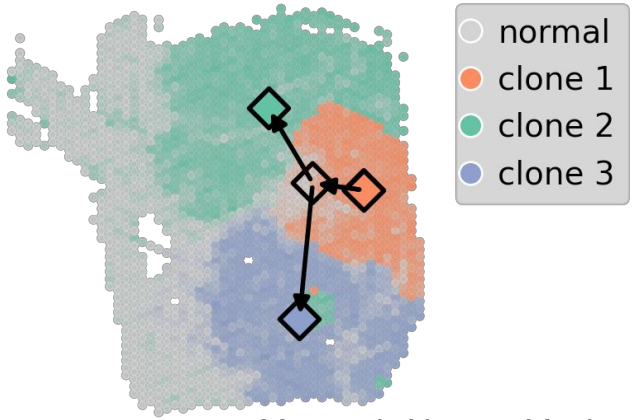
1. Constructing phylogenies from bulk and single-cell cancer sequencing



Ding et al. *Nature* 2012

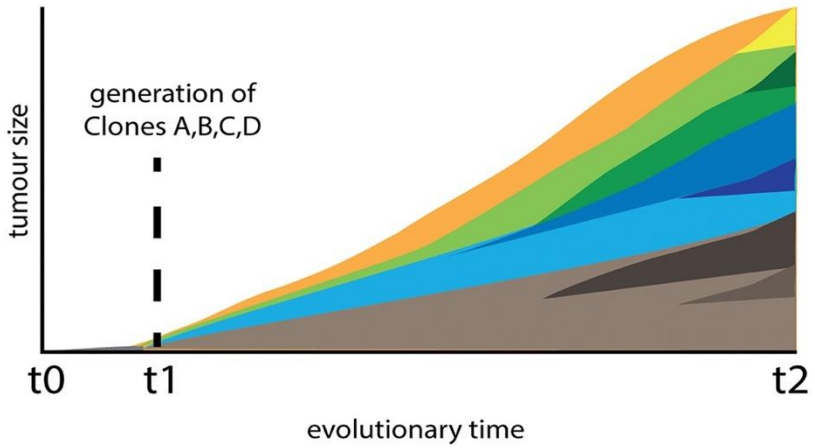
2. Migration: Metastasis and Spatial tumor evolution

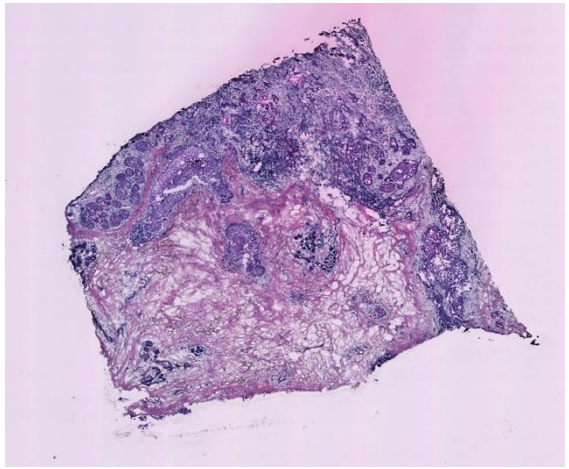




Ma et al. *Nature Methods* 2024

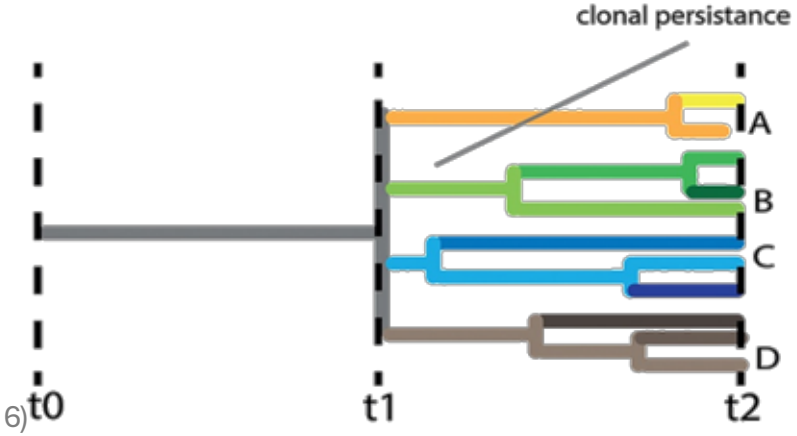
Spatial tumor evolution?





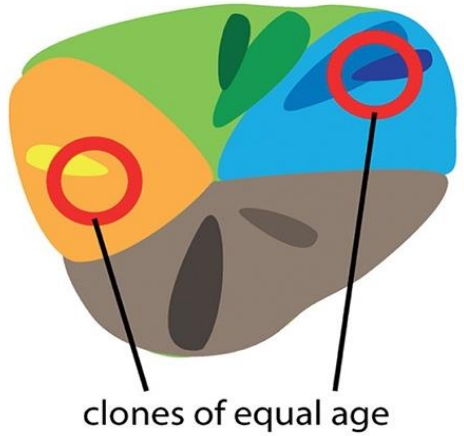
Tumor sample

DNA sequencing +
Phylogenetic inference



Spatial sequencing

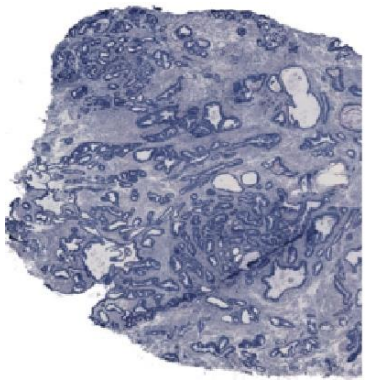
Spatial evolution



Images: Cross et al. J. Pathol. (2016)

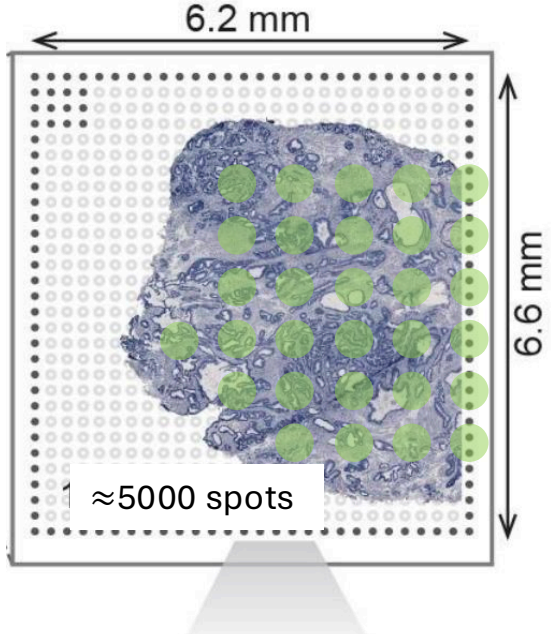
Spatial transcriptomics (*-omics)

Tissue sample



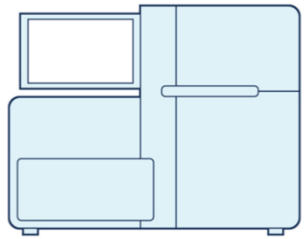
Sequencing

Barcoded Grid of Spots

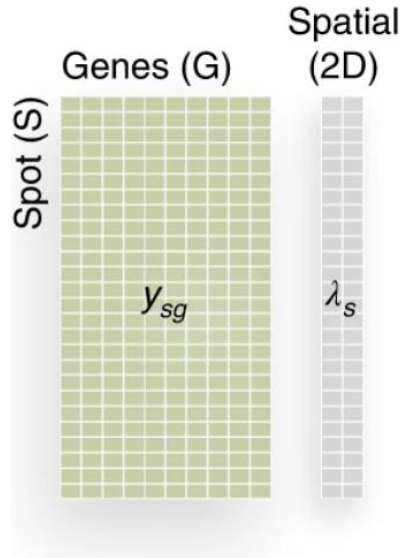


Berglund et al. Nat Com. 2018

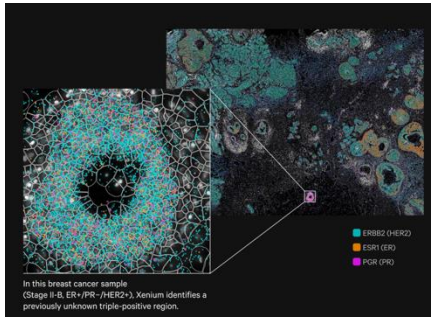
RNA (DNA) sequencing



Spot × gene **transcript count matrix** with **spatial coordinates**



Imaging

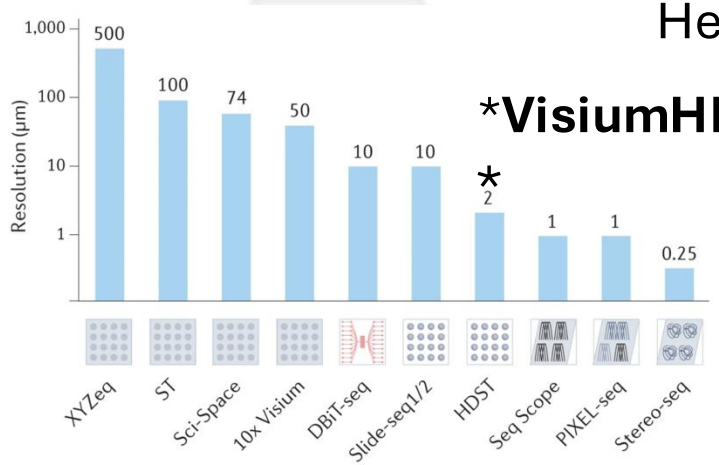


10x Genomics Xenium

10x Genomics Visium

55 µm spots

Hexagonal grid



*Moffit et al. *Nature Reviews Genetics* 2022

***VisiumHD** (2 – 16 µm)

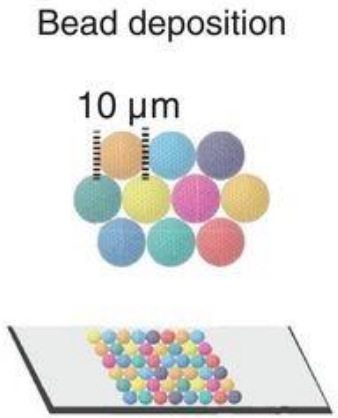
Slide-seqV2

10 µm beads

Variable spacing

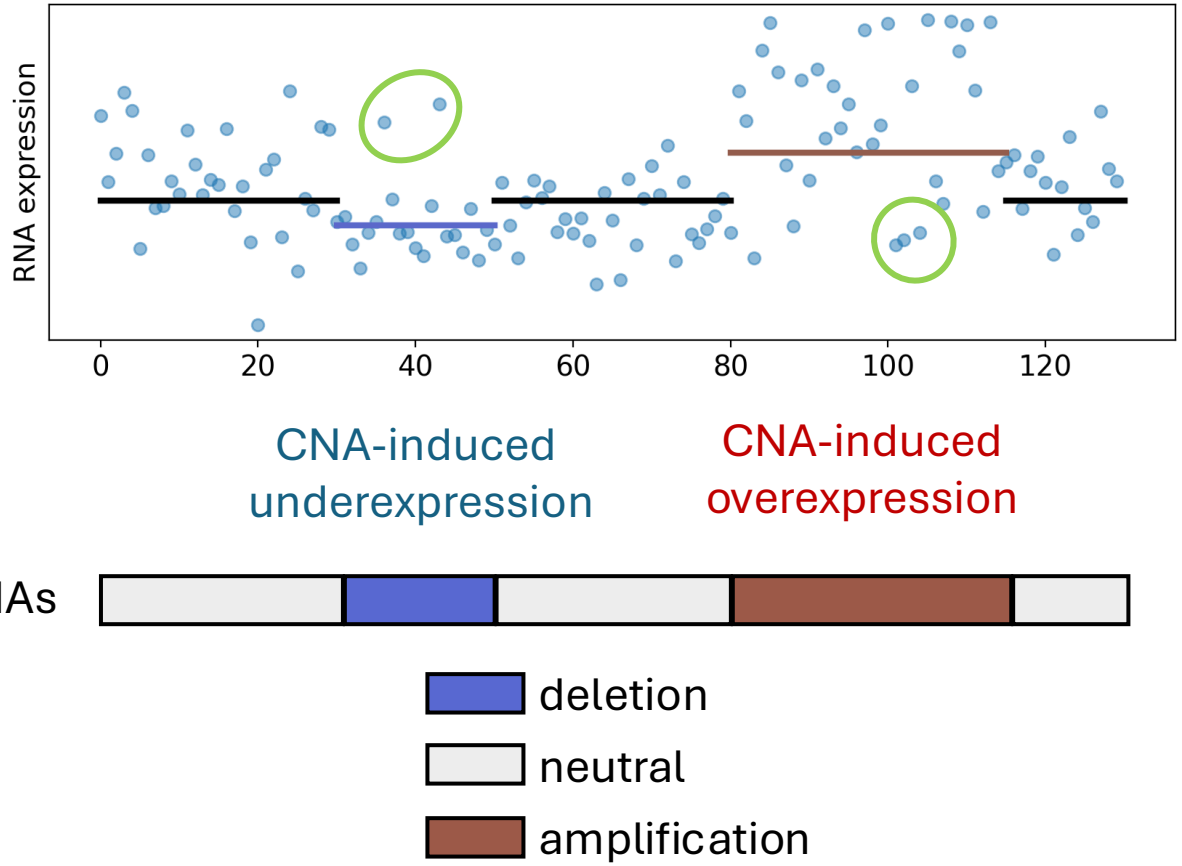
Rodriques et al. *Science* 2019

Stickels et a. *Nat. Biotech*. 2021



Footprints of copy number aberrations in (spatial) transcriptomics data

Large CNAs alter expression of multiple adjacent genes



CNA inference from single-cell RNA-seq:

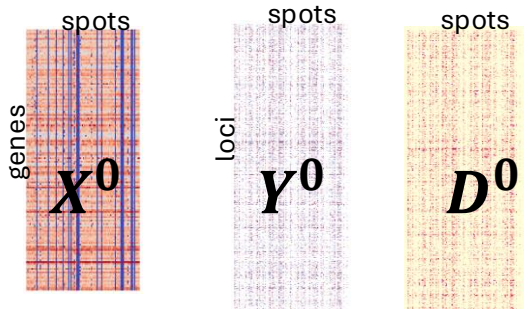
HoneyBADGER [Fan et al. 2018], InferCNV [Trinity-CTAT 2018], CopyKAT [Gao et al., 2021], Clonalscope [Wu et al., 2022], Numbat [Gao et al., 2022], Xclone [Huang et al., 2023]

Other expression variation

Distinguishing effects of CNAs from other expression variation is challenging.

CalicoST: **Allele-specific CNAs** and **spatial tumor evolution** from spatial transcriptomics

Multi-sample SRT

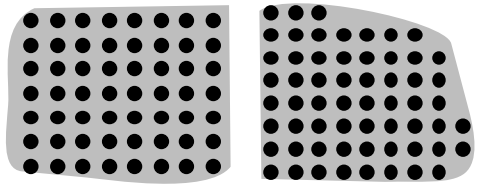


Transcript counts

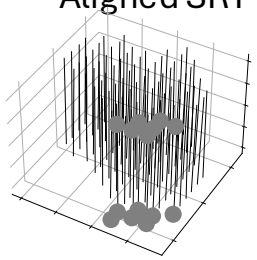
ALT allele counts

Total allele counts

Coordinates S

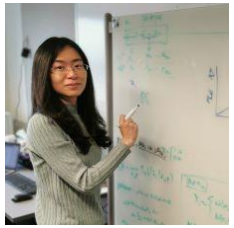


Aligned SRT



(Optional) Tumor proportion θ, alignment W across multiple slices

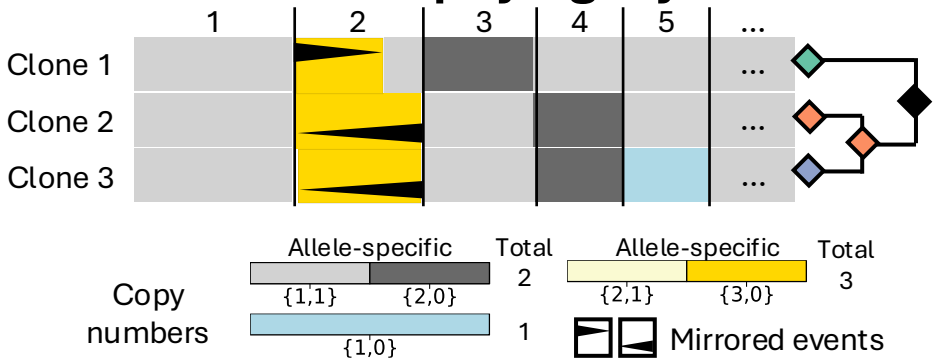
Calico-ST



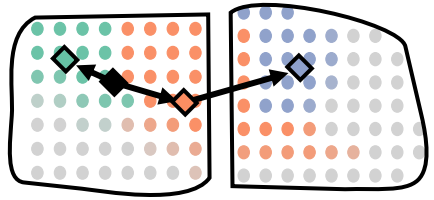
Cong Ma

Ma et al., *RECOMB* (2024) and *Nature Methods* (2024)

Allele-specific copy number profiles and phylogeny

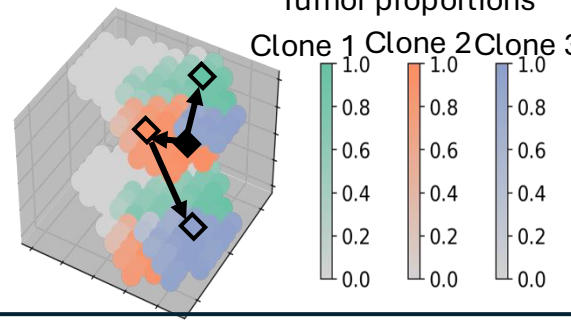


Cancer clones, tumor proportions, and phylogeography

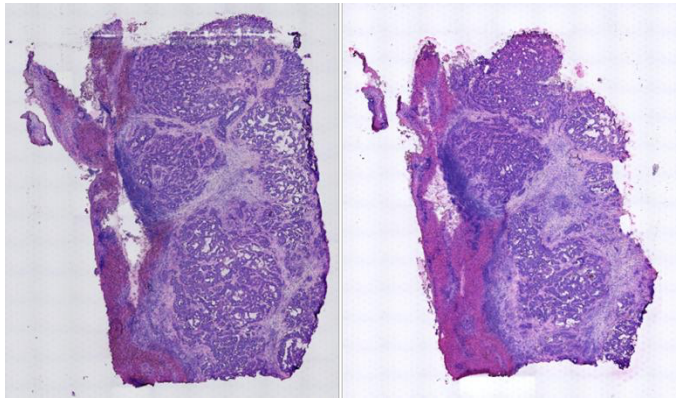


Phylogeography

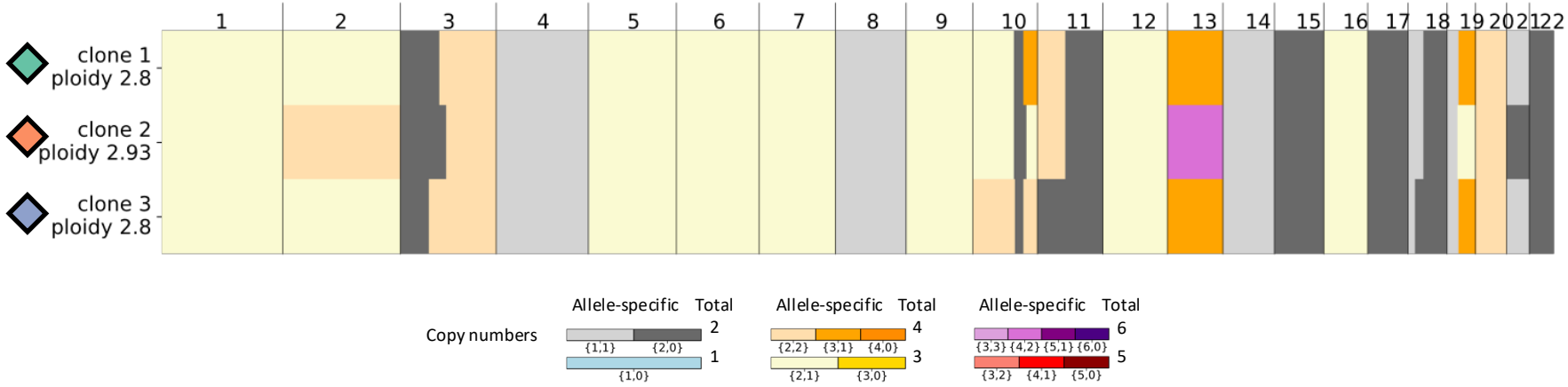
Tumor proportions



CalicoST reconstructs spatial evolution from loss of heterozygosity (LOH) events

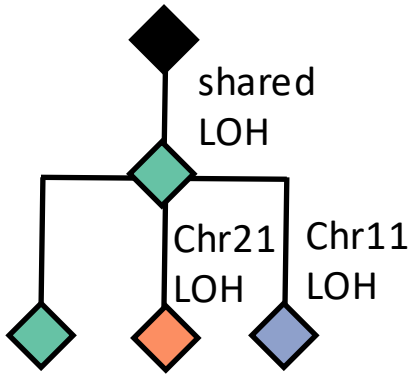


Replicate 1 Replicate 2



Colorectal liver-met (10x Visium)
(Li Ding, WUSTL, Human Tumor Atlas Network)

Infer tumor phylogeny

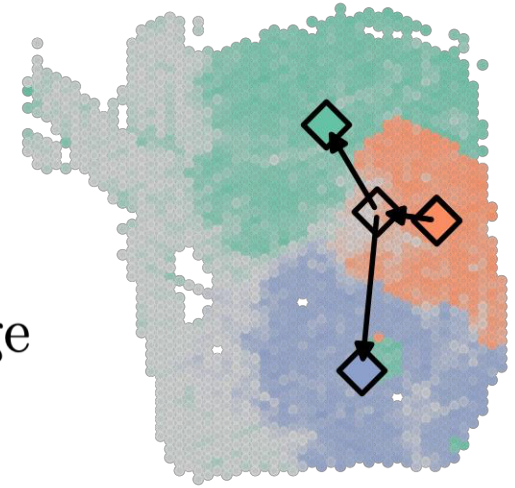


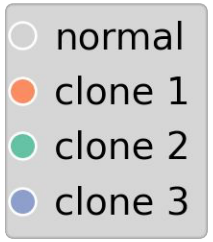
Infer ancestral spatial locations

Gaussian diffusion model

$$s_{\text{child}} \sim \mathcal{N}(s_{\text{parent}}, wI)$$

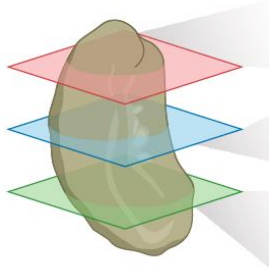
w : number of events on the edge

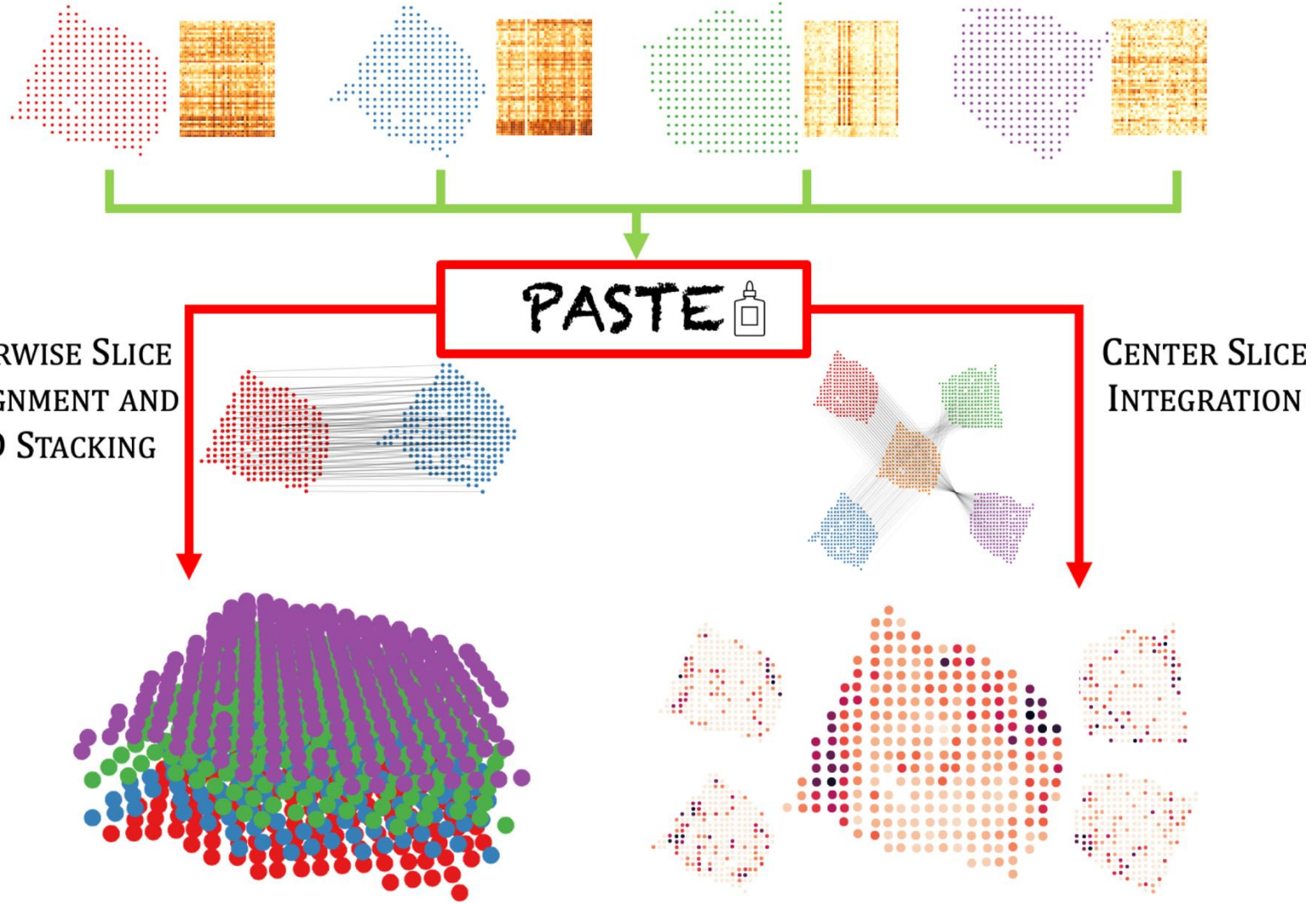




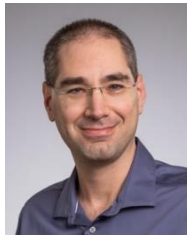
PASTE: Probabilistic Alignment of ST Experiments

Align and **integrate** spatial transcriptomics data from multiple slices, leveraging ***both*** spatial and gene expression information





Fused Gromov-Wasserstein Optimal Transport [Titouan et al, ICML 2019]



Ron Zeira

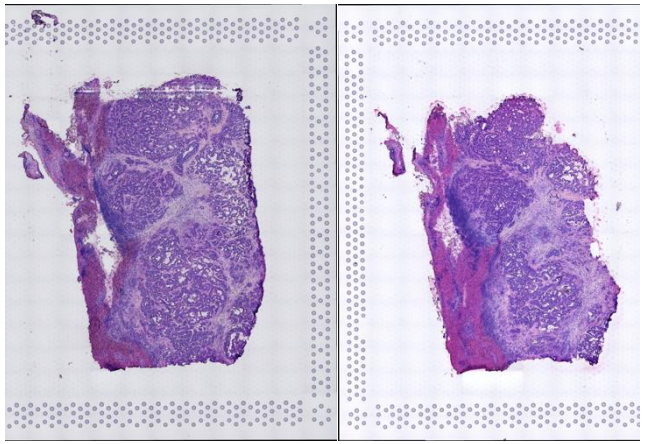
Max Land

Alexander Strzalkowski

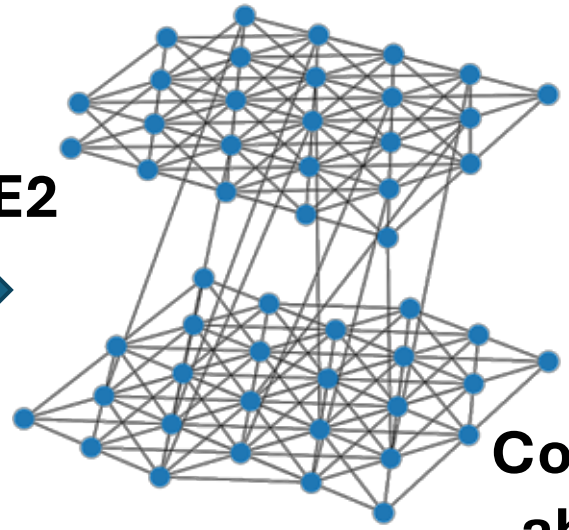
PASTE: Zeira, R. et al. *Nature Methods* **19,** 567–575 (2022) and *RECOMB* 2021]

PASTE2: Liu, Zeira, R [*Genome Research* 2023 and RECOMB 2023

Tumor evolution in 3D! PASTE + CalicoST

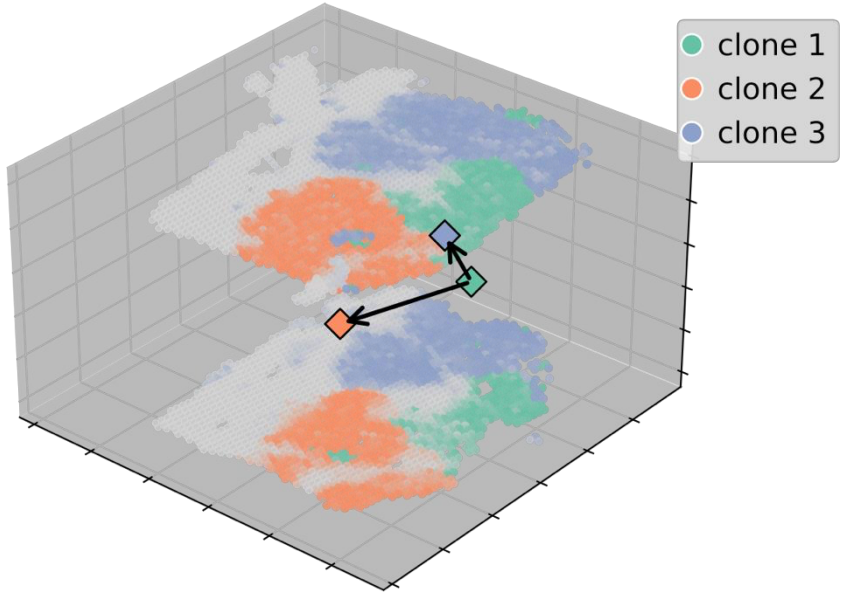


PASTE2



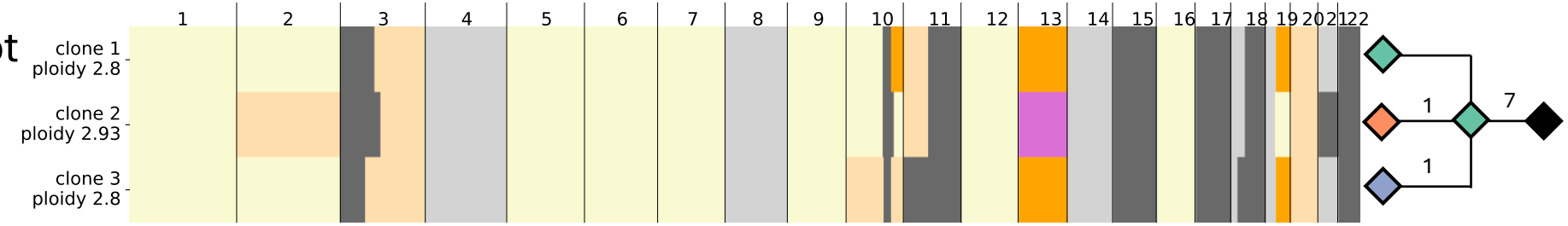
CalicoST

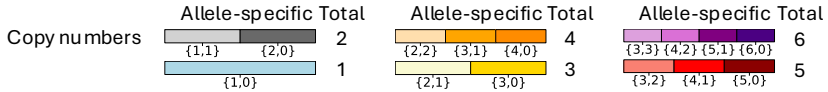
Copy number aberrations



2 slices of colorectal liver-met (10X Visium, **Li Ding, WUSTL**)

- 1975 and 1721 spots
- Median 28467 UMIs/spot





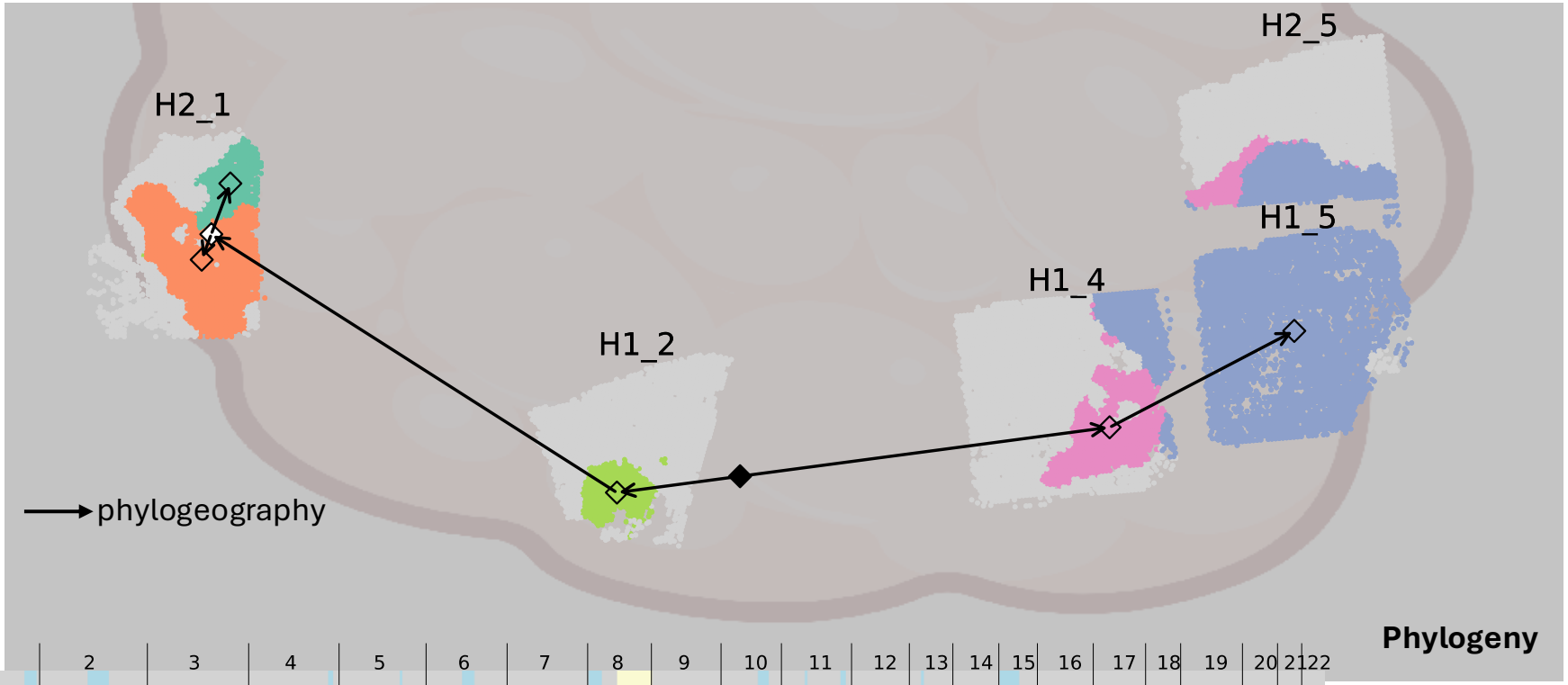
PASTE: Zeira, R. et al. *Nature Methods* **19,** 567–575 (2022) and RECOMB (2021)

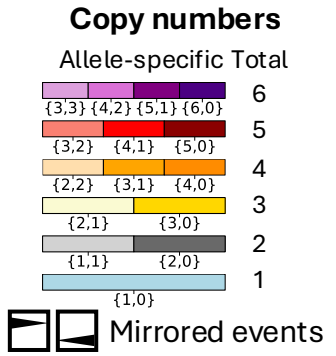
PASTE2: Liu, Zeira, R. *Genome Research* (2023) and RECOMB (2023)

CalicoST identifies mirrored events in prostate cancer

5 regions of a cancerous prostate organ
(10X Visium, [Erickson et al., Nature, 2022](#))

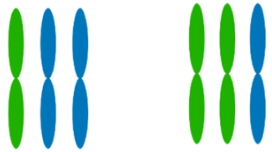
- 17372 spots in total
- Median 2683 UMIs/spot

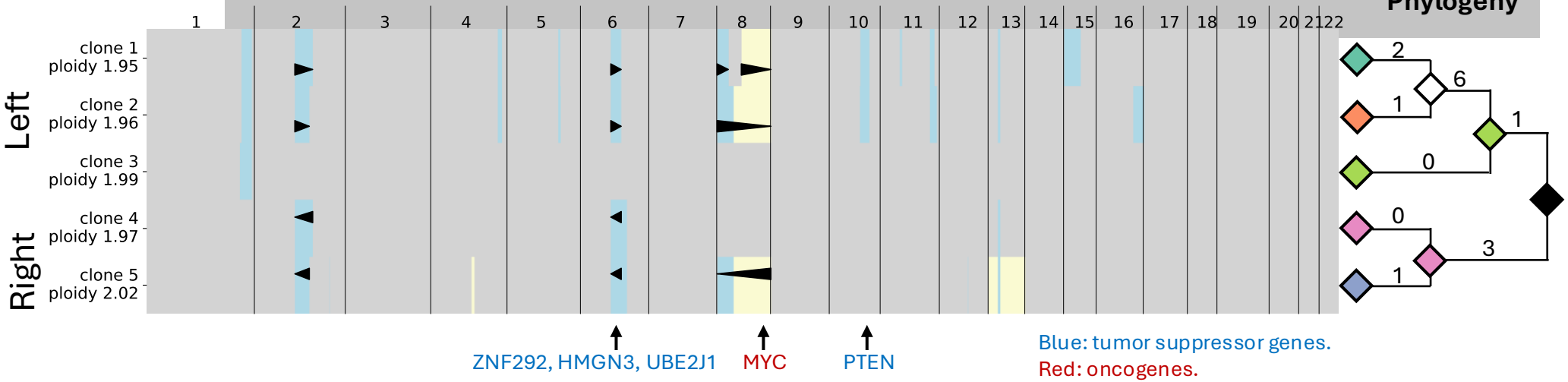




Chr8 mirrored amplification

Clones 1,2 Clone 5

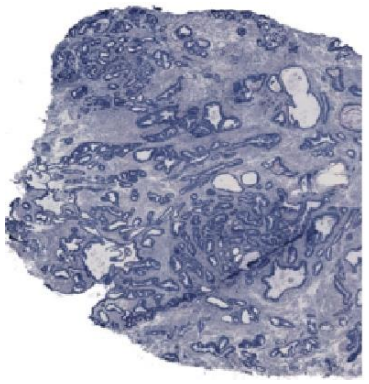




Blue: tumor suppressor genes.
Red: oncogenes.

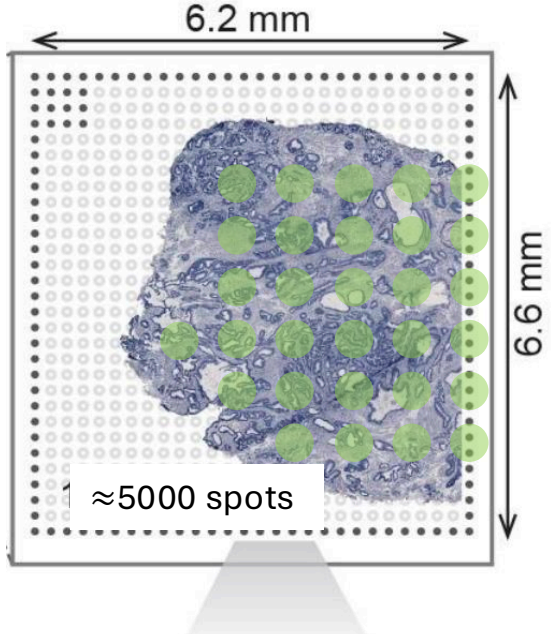
Spatial transcriptomics (*-omics)

Tissue sample



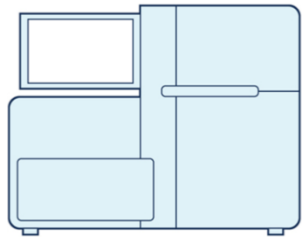
Sequencing

Barcoded Grid of Spots

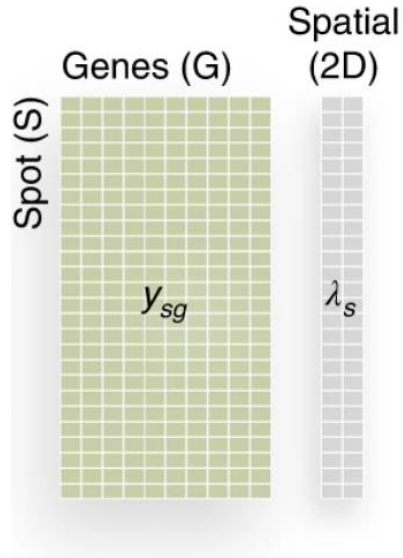


Berglund et al. Nat Com. 2018

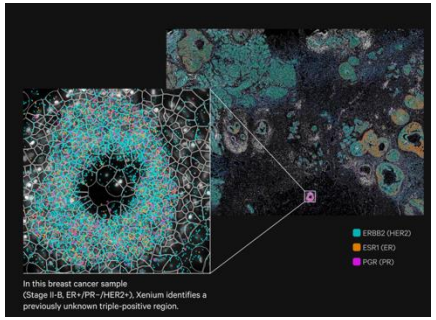
RNA (DNA) sequencing



Spot × gene **transcript count matrix** with **spatial coordinates**



Imaging

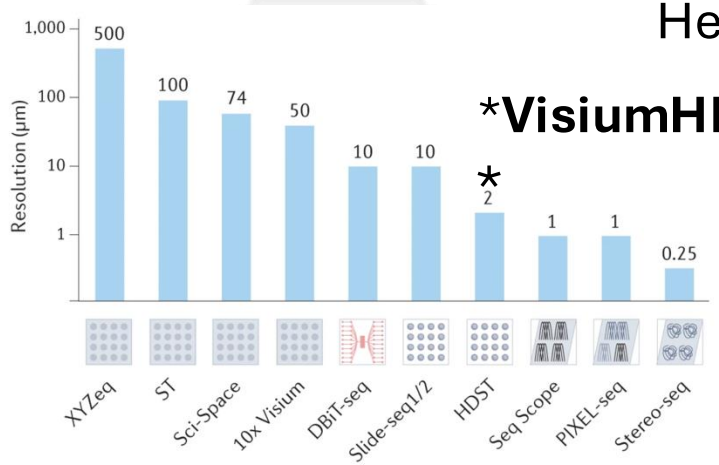


10x Genomics Xenium

10x Genomics Visium

55 µm spots

Hexagonal grid



*Moffit et al. *Nature Reviews Genetics* 2022

***VisiumHD** (2 – 16 µm)

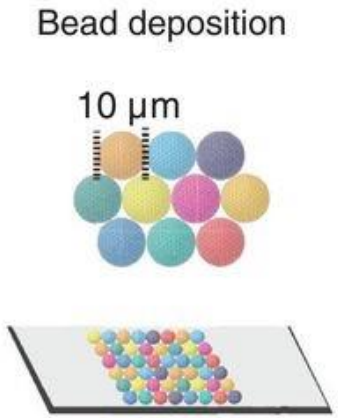
Slide-seqV2

10 µm beads

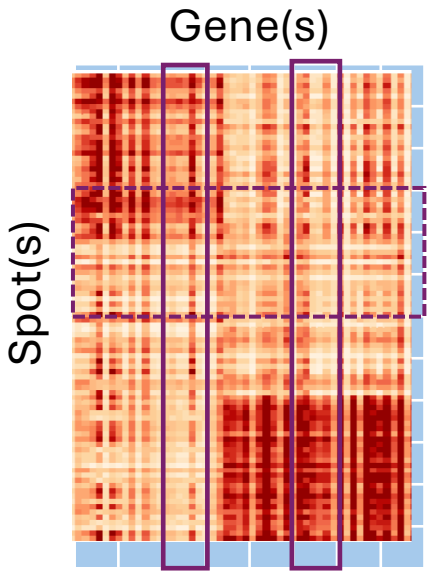
Variable spacing

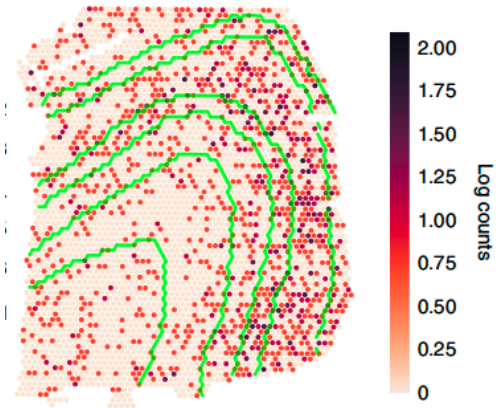
Rodriques et al. *Science* 2019

Stickels et a. *Nat. Biotech*. 2021



Spatial analyses complicated by low coverage





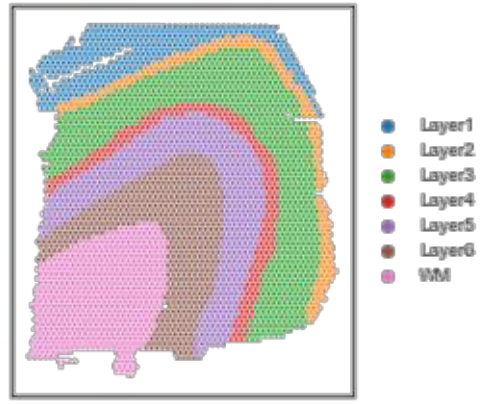
Spatial domains

Spatially varying genes

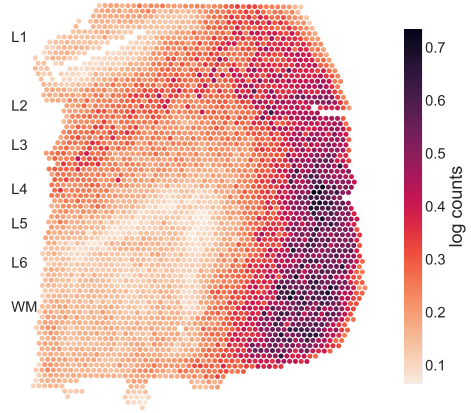
Cell-cell interactions

...

Human brain (prefrontal cortex)



Maynard *et al.*, Nat. Neuro. 2021



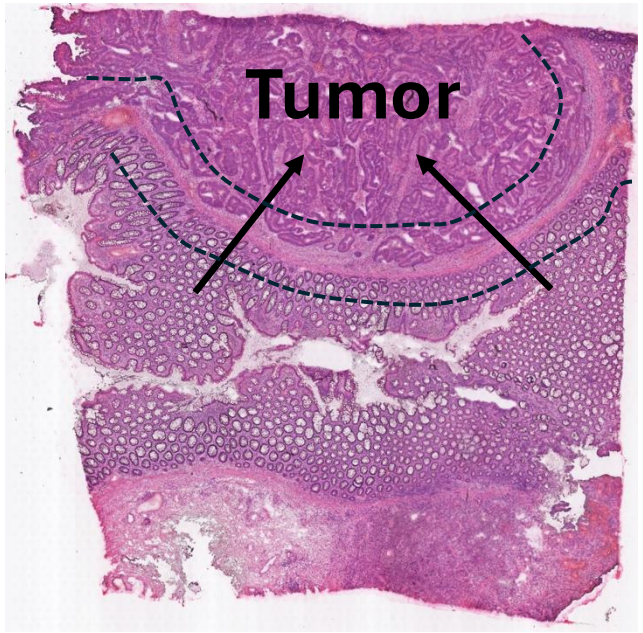
	Spot size	Num Spots	Total count / spot	Matrix sparsity
Visium	55 µm	4K	28000	0.62
Slide-seqV2	10 µm	33K	400	0.986
VisiumHD	2 µm	8.7M	31	0.998

Similar limitations for imaging-based technologies: MERFISH, 10x Genomics Xenium, STARmap, ...

Identifying spatial gradients of gene expression

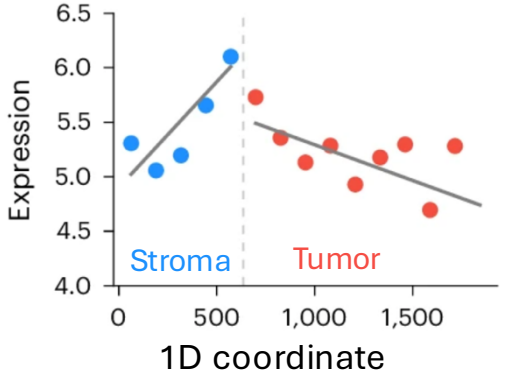
What genes change expression
In/near tumor boundary?
Within tumor microenvironment?

Need a coordinate system describing tissue geometry!

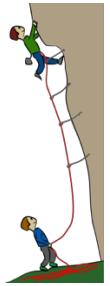


Colorectal tumor slice (stage IV)
(Wu et al, *Cancer Discovery* 2022)

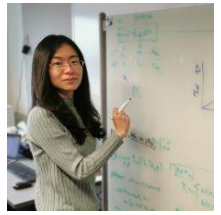
1D coordinate system?
Expression based on position relative to tumor boundary?



Gene expression

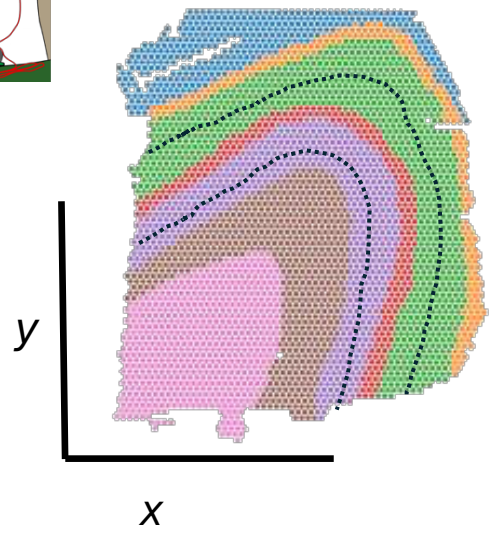


Belayer: Modeling layered tissues



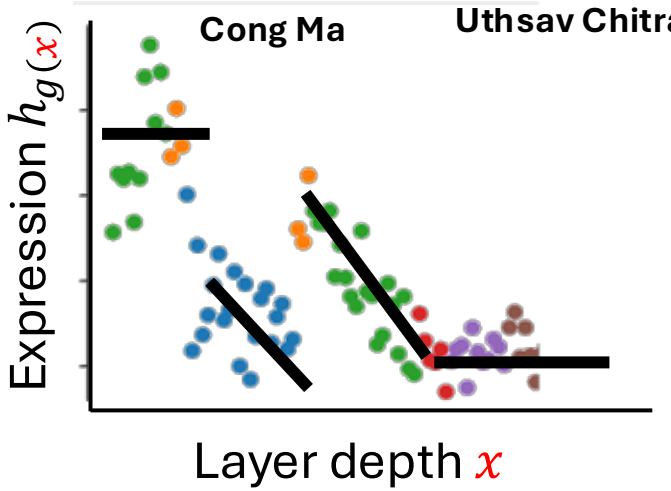
Cong Ma

Uthsav Chitra



DLPFC sample

(Marker) gene expression depends only on "***distance from layer boundary***"



Model (marker) gene expression:

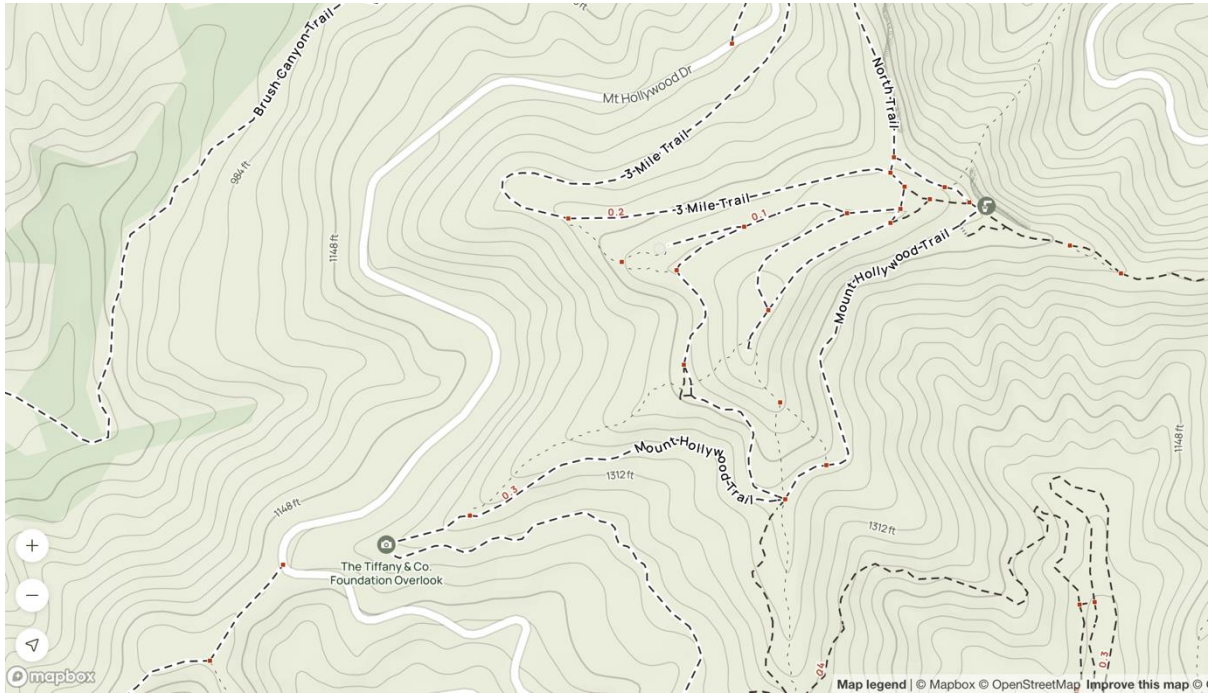
$f_g(x, y) = h_g(x)$ depends on x-coordinate

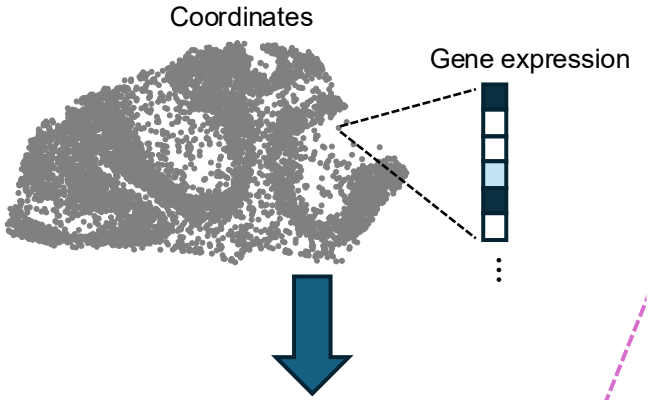
$f_g(x, y) = h_g(d(x, y))$ depends on value d

How to learn function $d(x, y): \mathbb{R}^2 \rightarrow \mathbb{R}$?

Ma, Chitra, et al. *Cell Systems* (2022) [Also: RECOMB 2022]
doi.org/10.1016/j.cels.2022.09.002

“Topography” of gene expression → **isodepth**

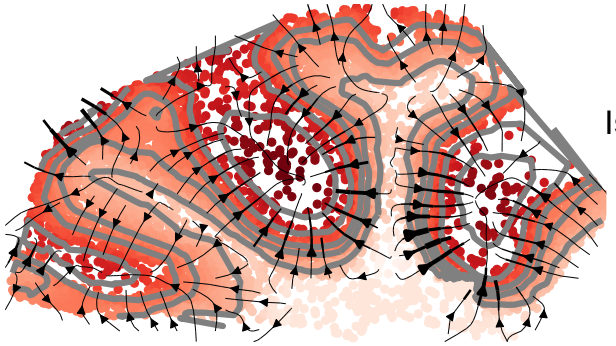


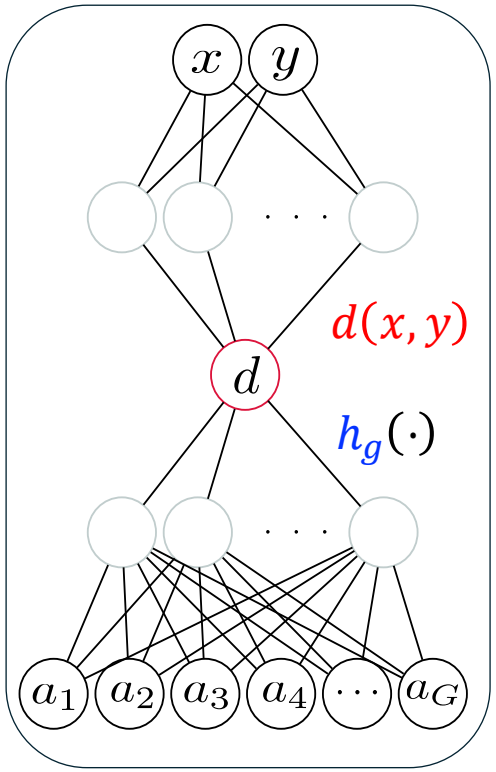


GASTON

Gradient Analysis of Spatial Transcriptomics Organization with Neural networks

Topographic map of tissue slice





Uthsav Chitra

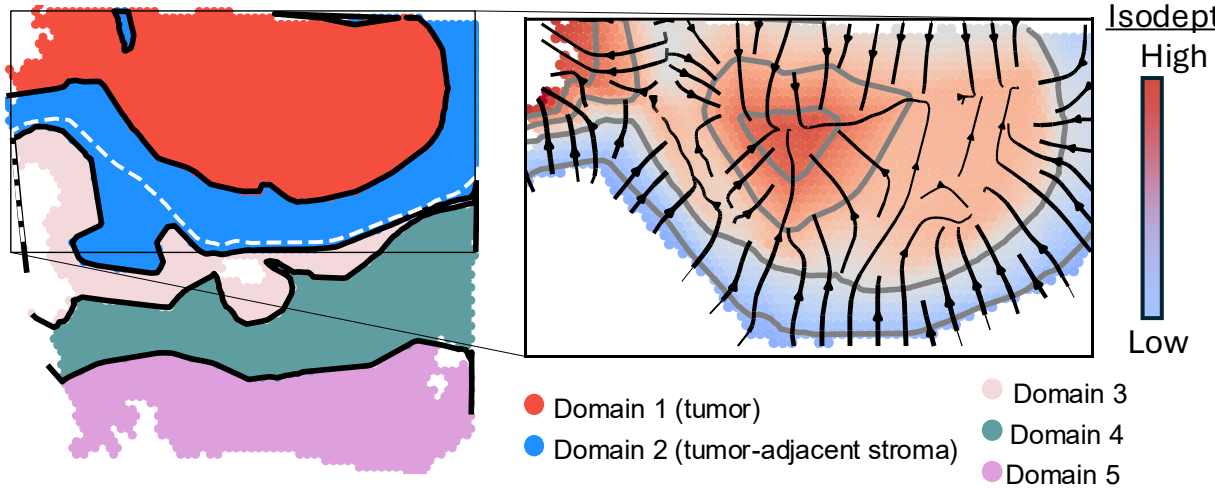
- **Isodepth** = contours of equal “potential”: $d(x,y) = c$
- Gene expression function: $h_g\big(d(x,y)\big)$
- See also: implicit neural representations

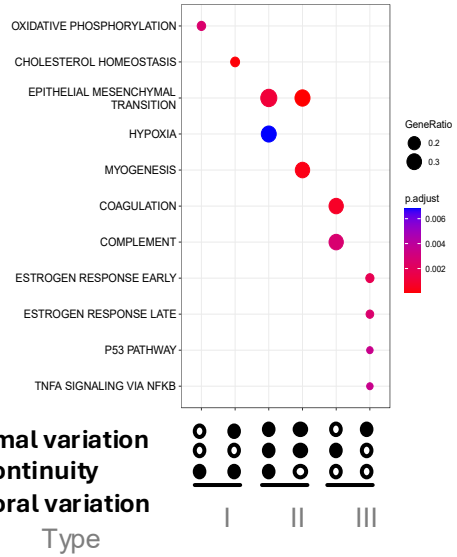
Chitra, et al. RECOMB (2024) and *Nature Methods* (2025)

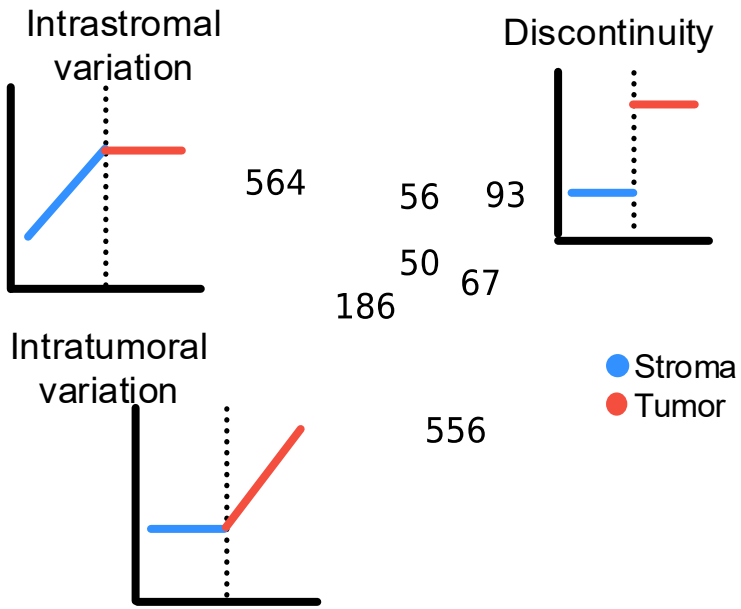
GASTON identifies gradients in tumor microenvironment

Colorectal tumor slice (stage IV)
(Wu et al, *Cancer Discovery* 2022)

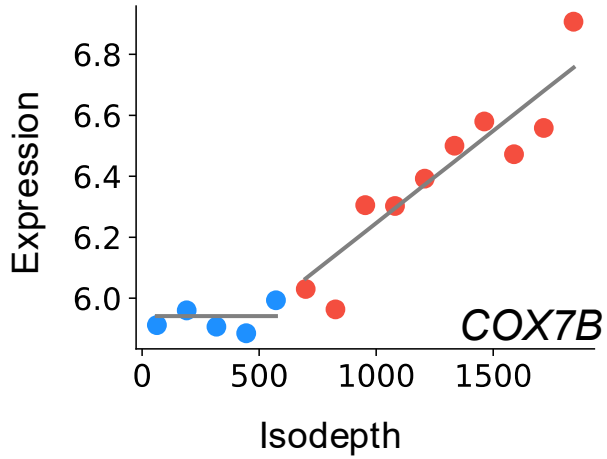
GASTON: spatial domains + isodepth



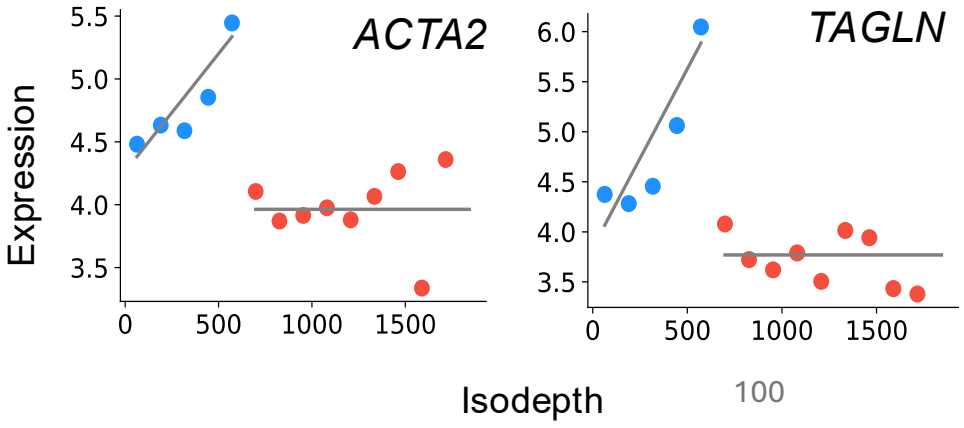




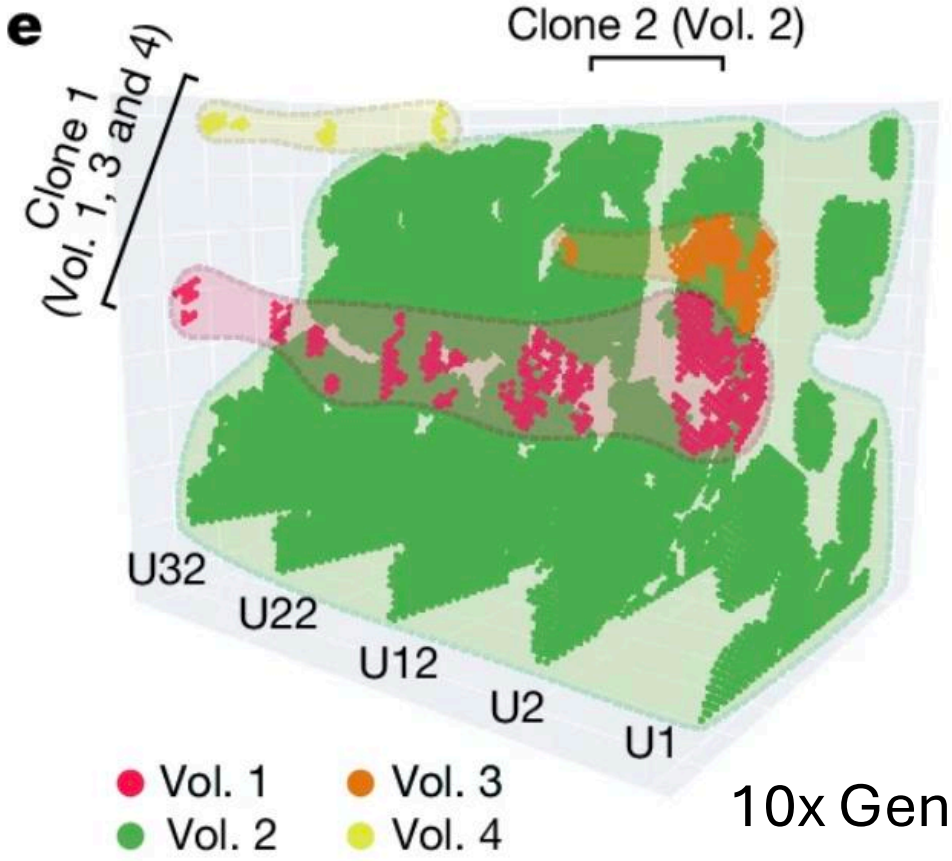
Intratumoral variation
aerobic metabolism



Discontinuity/Intrastromal variation-
epithelial-mesenchymal transition (EMT)



3D Tumor Atlas: Human Tumor Atlas Network (HTAN)



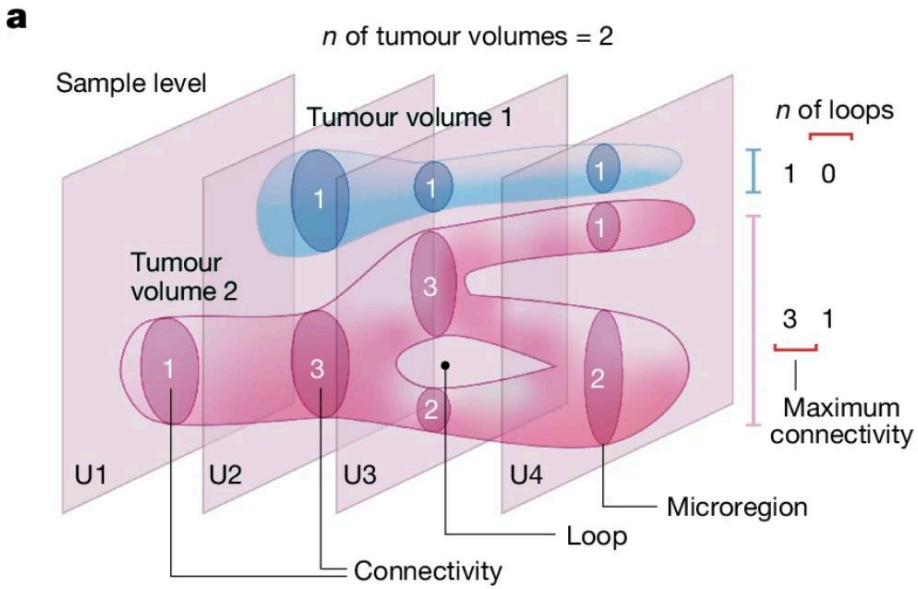
10x Genomics Visium slices from breast cancer liver metastasis

Article | Open access | Published: 30 October 2024

Tumour evolution and microenvironment interactions in 2D and 3D space

Chia-Kuei Mo, Jingxian Liu, Siqi Chen, Erik Storrs, Andre Luiz N. Targino da Costa, Andrew Houston, Michael C. Wendl, Reyka G. Jayasinghe, Michael D. Iglesia, Cong Ma, John M. Herndon, Austin N. Southard-Smith, Xinhao Liu, Jacqueline Mudd, Alla Karpova, Andrew Shinkle, S. Peter Goedegebuure, Abdurrahman Taha Mousa Ali Abdelzaher, Peng Bo, Lauren Fulghum, Samantha Livingston, Metin Balaban, Angela Hill, Joseph E. Ippolito, Vesteinn Thorsson, Jason M. Held, Ian S. Hagemann, Eric H. Kim, Peter O. Bayguinov, Albert H. Kim, Mary M. Mullen, Kooresh I. Shoghi, Tao Ju, Melissa A. Reimers, Cody Weimholt, Liang-I Kang, Sidharth V. Puram, Deborah J. Veis, Russell Pachynski, Katherine C. Fuh, Milan G. Chheda, William E. Gillanders, Ryan C. Fields, Benjamin J. Raphael, Feng Chen & Li Ding
— Show fewer authors

Nature **634**, 1178–1186 (2024) | Cite this article



Summary

Phylogeography: Metastasis, migration, and tumor growth

1. Infer history of metastatic migration between anatomical sites (MACHINA, fastMACH, etc.)
2. Cellular migration within a tumor (Calico-ST)
3. 3D tumor atlases (PASTE)

Future Directions

- What are constraints on metastatic seeding?
- Higher quality spatial and single-cell data
- Realistic (data informed) tumor growth models and interactions with microenvironment
- Multi-modal 3D tumor atlases (non-genetic evolution)

Acknowledgments

Raphael Group:

Viola Chen
Gillian Chu
Julian Gold
Peter Halmos
William Howard-Synder
Gary Hu
Akhil Jakatdar
Xinhao Liu
Sereno Lopez-Darwin
Runpeng Luo
Henri Schmidt
Yihang Shen
Hongyu Zheng
Clover Zheng

Washington University in St. Louis School of Medicine

Dr. Li Ding
Clara Liu
Siqi Chen
...and member of Ding group

NIH NATIONAL CANCER INSTITUTE

NIH NATIONAL CANCER INSTITUTE Informatics Technology for Cancer Research

DAMON RUNYON CANCER RESEARCH FOUNDATION

NSF

Alumni:

Metin Balaban
Mohammed El-Kebir
Cong Ma
Matthew Myers
Layla Oesper
Palash Sashittal
Gryte Satas
Ron Zeira
Simone Zaccaria

Software: github.com/raphael-group/

Contact: braphael@princeton.edu