

Combinatorial Algorithms for Tumor Phylogenetics

Mohammed El-Kebir

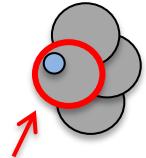
University of Illinois at Urbana Champaign,
Department of Computer Science

October 2022



Tumorigenesis: Cell Mutation

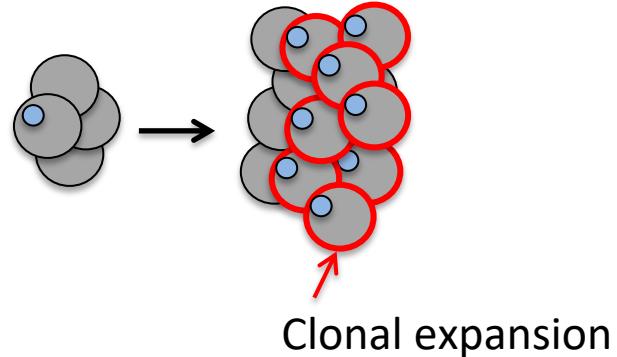
Clonal Evolution Theory of Cancer
[Nowell, 1976]



Founder
tumor cell
with somatic mutation: ●
(e.g. BRAF V600E)

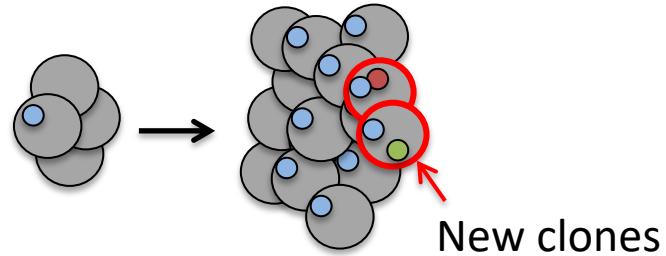
Tumorigenesis: Cell Mutation

Clonal Evolution Theory of Cancer
[Nowell, 1976]



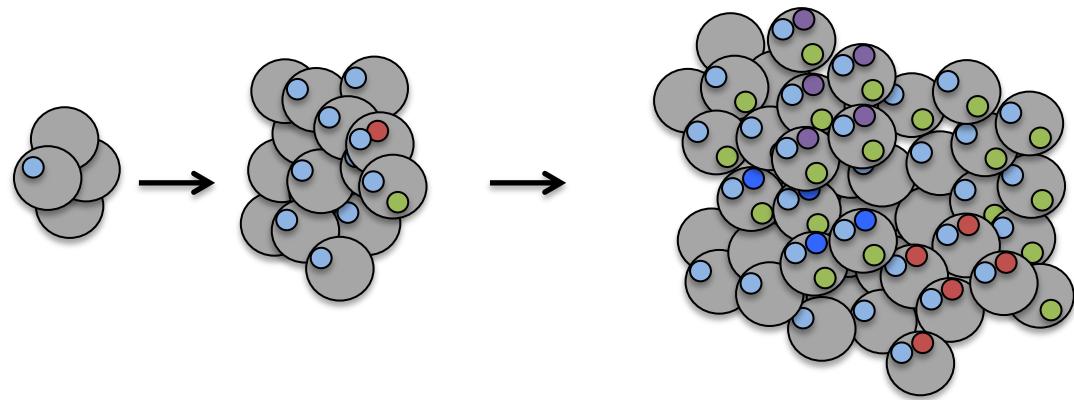
Tumorigenesis: Cell Mutation

Clonal Evolution Theory of Cancer
[Nowell, 1976]



Tumorigenesis: Cell Mutation & Division

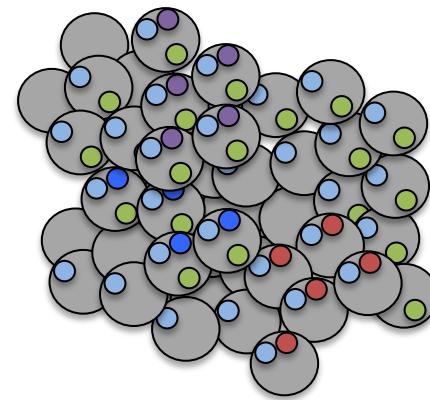
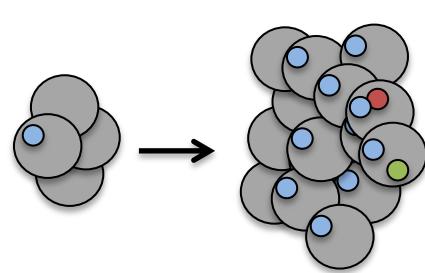
Clonal Evolution Theory of Cancer
[Nowell, 1976]



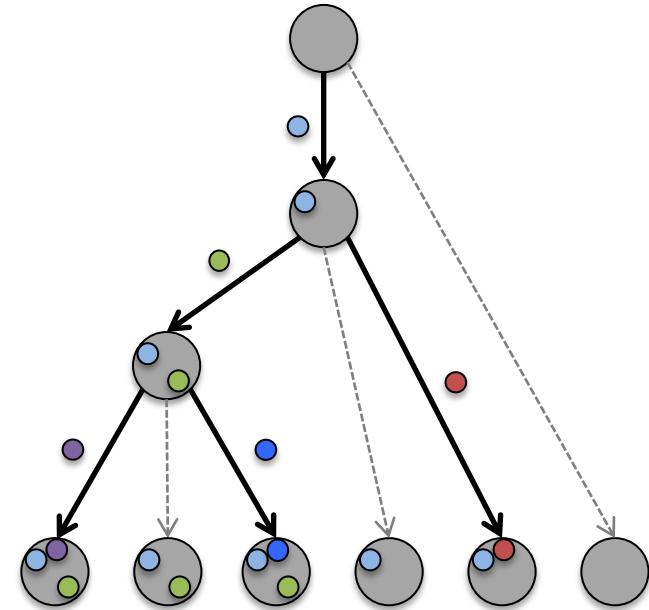
Intra-Tumor
Heterogeneity

Tumorigenesis: Cell Mutation & Division

Clonal Evolution Theory of Cancer
[Nowell, 1976]



Intra-Tumor
Heterogeneity

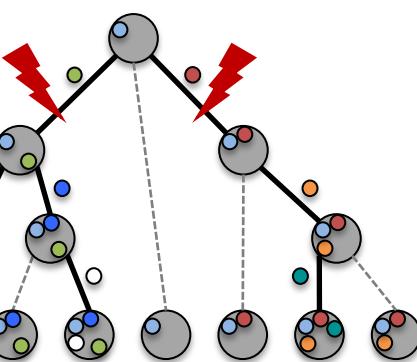


Phylogenetic
Tree T

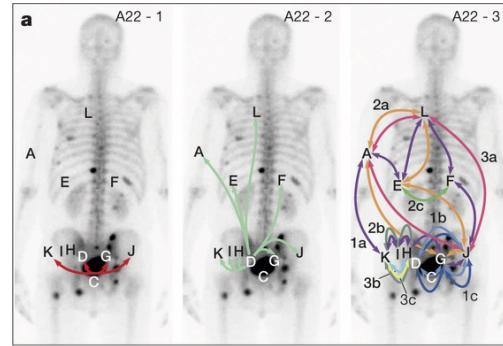
Question: Why are tumor phylogenies important?

Phylogenies are Key to Understanding Cancer

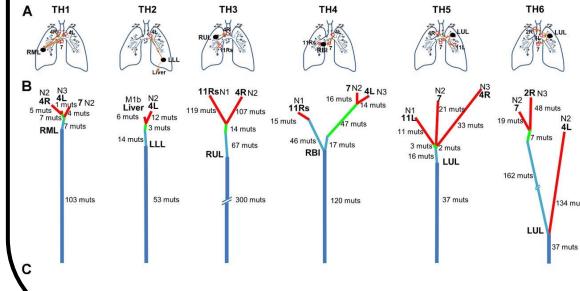
Identify targets for treatment



Understand metastatic development

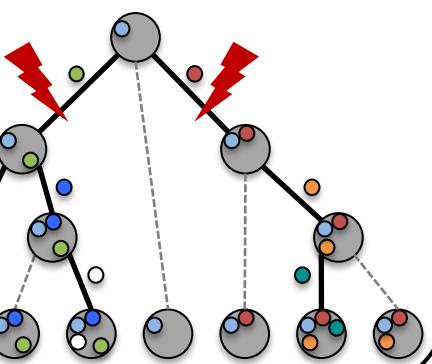


Recognize common patterns of tumor evolution across patients

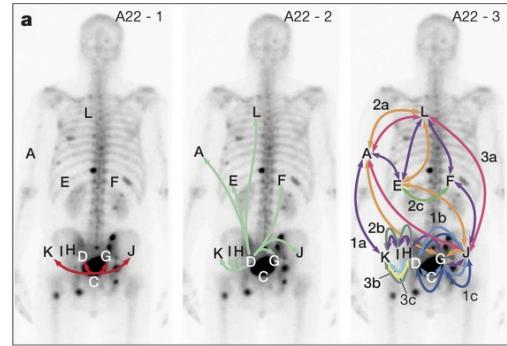


Phylogenies are Key to Understanding Cancer

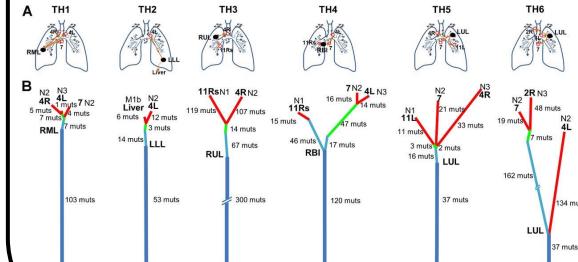
Identify targets for treatment



Understand metastatic development



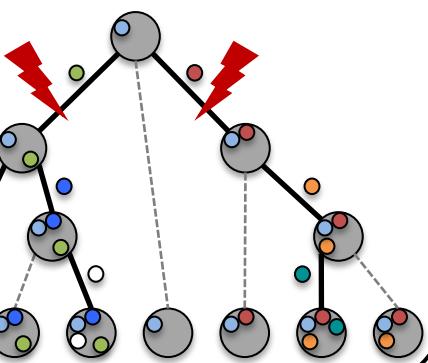
Recognize common patterns of tumor evolution across patients



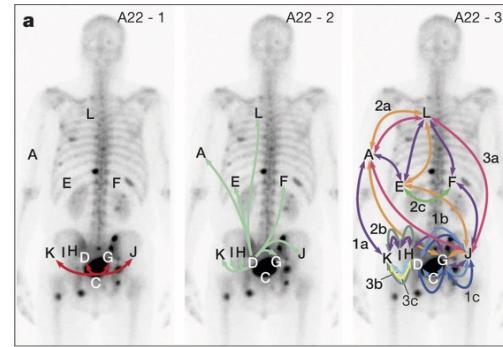
These downstream analyses **critically rely** on accurate tumor phylogeny inference

Phylogenies are Key to Understanding Cancer

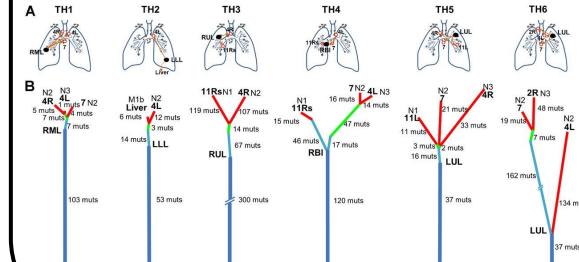
Identify targets for treatment



Understand metastatic development



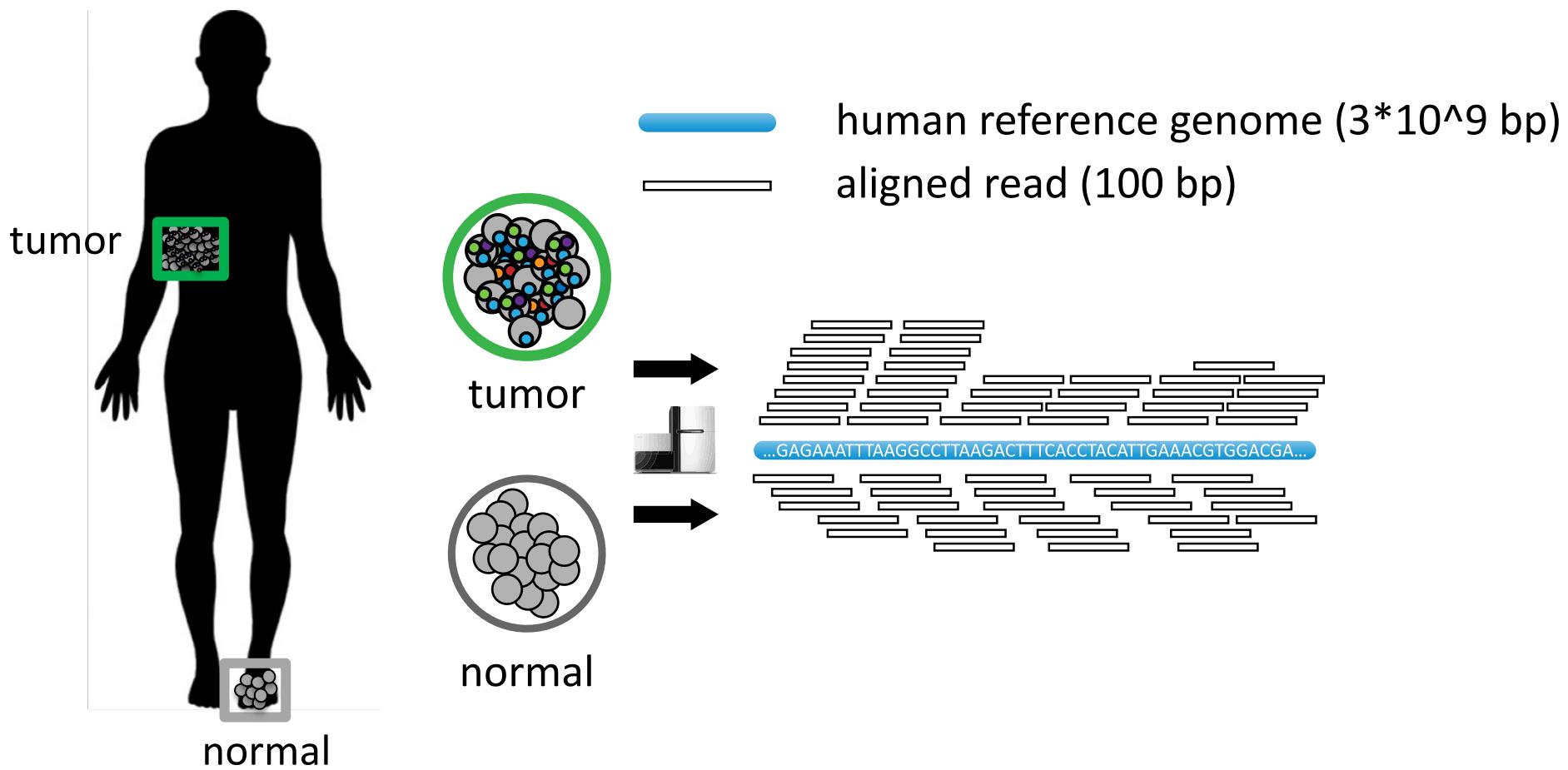
Recognize common patterns of tumor evolution across patients



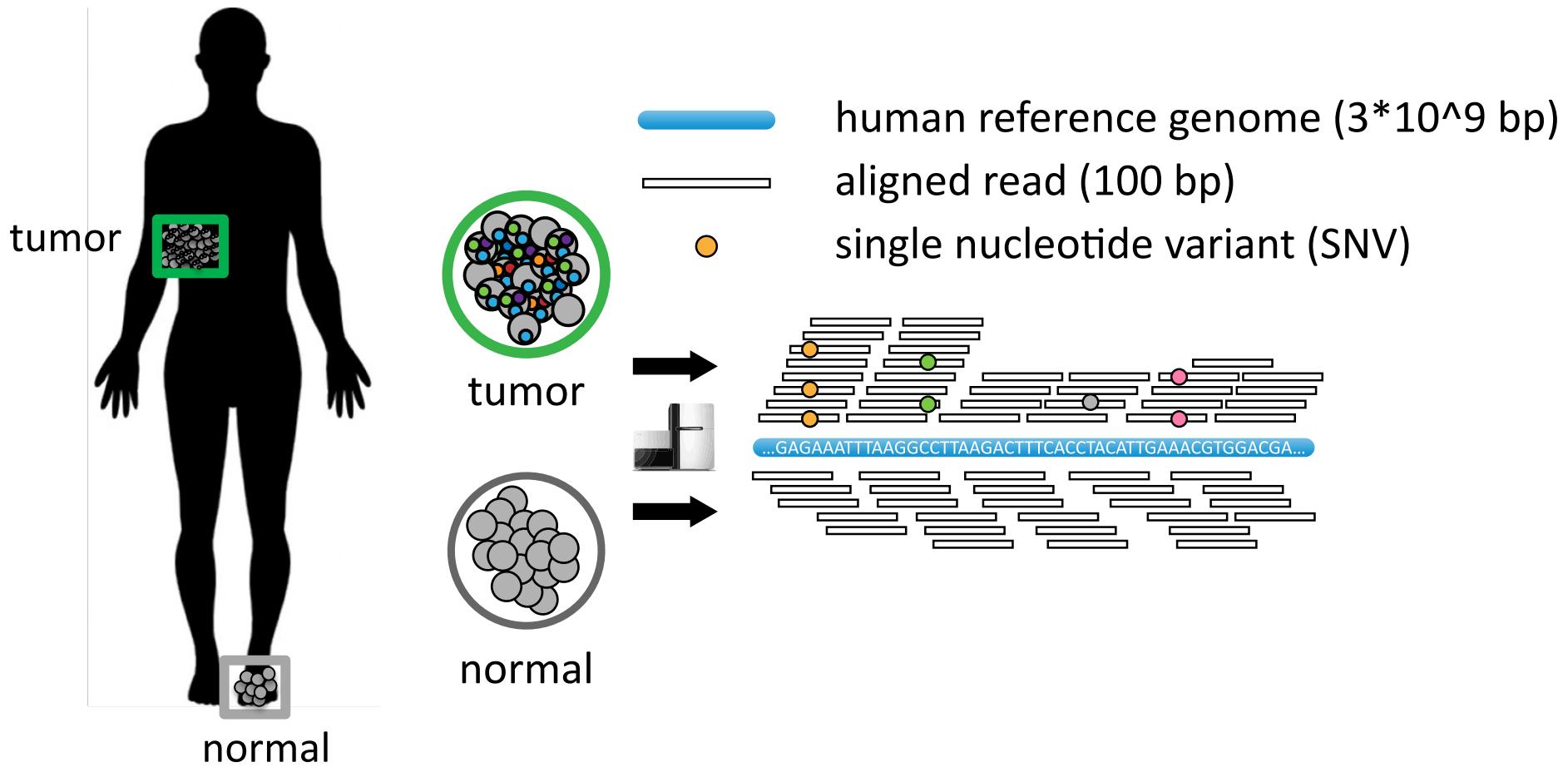
These downstream analyses **critically rely** on accurate tumor phylogeny inference

Key challenge in phylogenetics:
Accurate phylogeny inference from data at present time

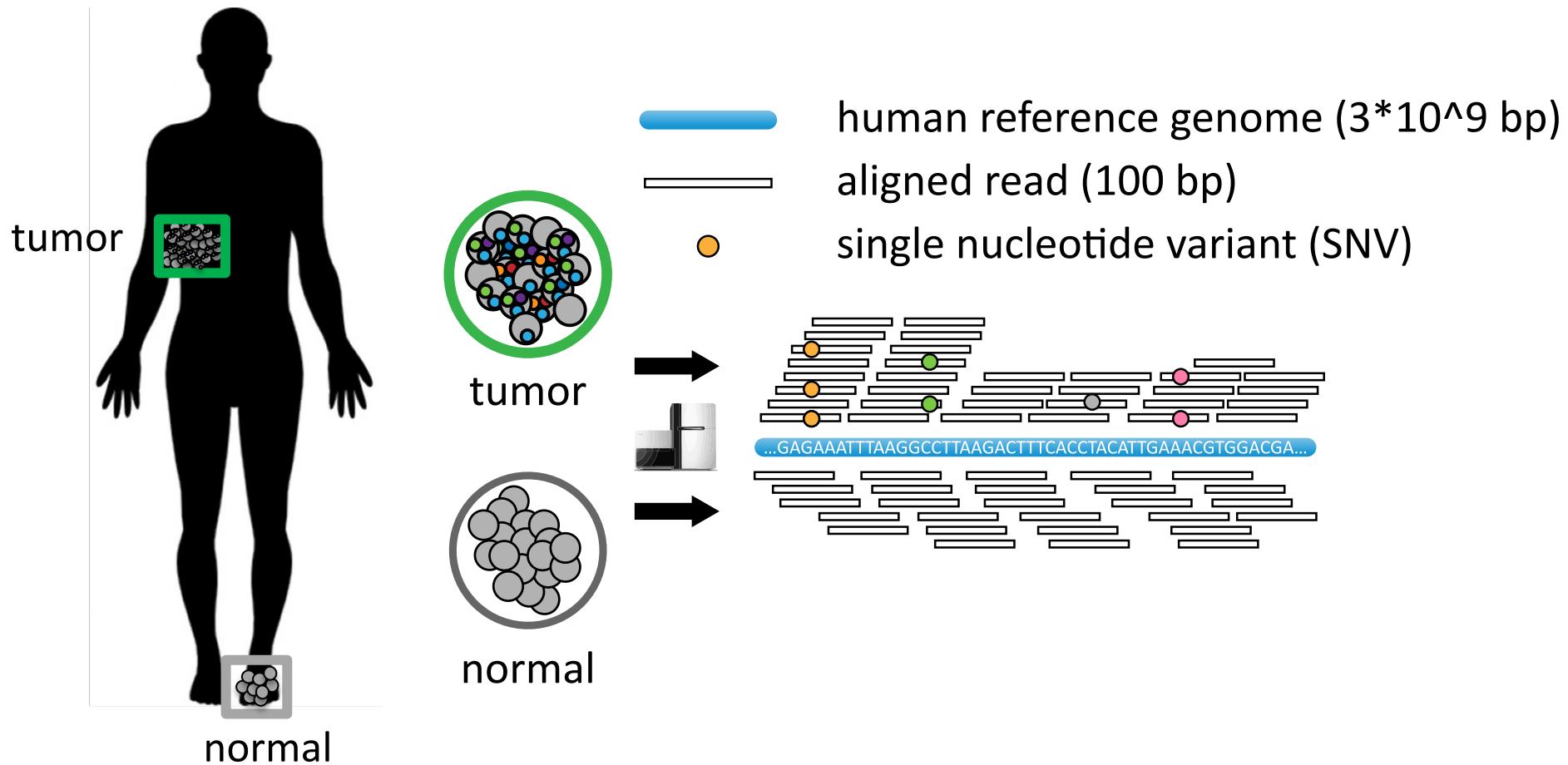
Additional Challenge in Cancer Phylogenetics



Additional Challenge in Cancer Phylogenetics



Additional Challenge in Cancer Phylogenetics



Additional challenge in cancer phylogenetics:
Phylogeny inference from **mixed bulk samples** at present time

Outline

1. Background and theory: [RECOMB-CG 2018, AMB 2019]

- Perfect Phylogeny Mixture (PPM) problem
- #PPM: exact counting and uniform sampling

2. Simulation results: [RECOMB-CG 2018, AMOB 2019]

- What contributes to non-uniqueness?
- How to reduce non-uniqueness?

3. Summarizing solution space:

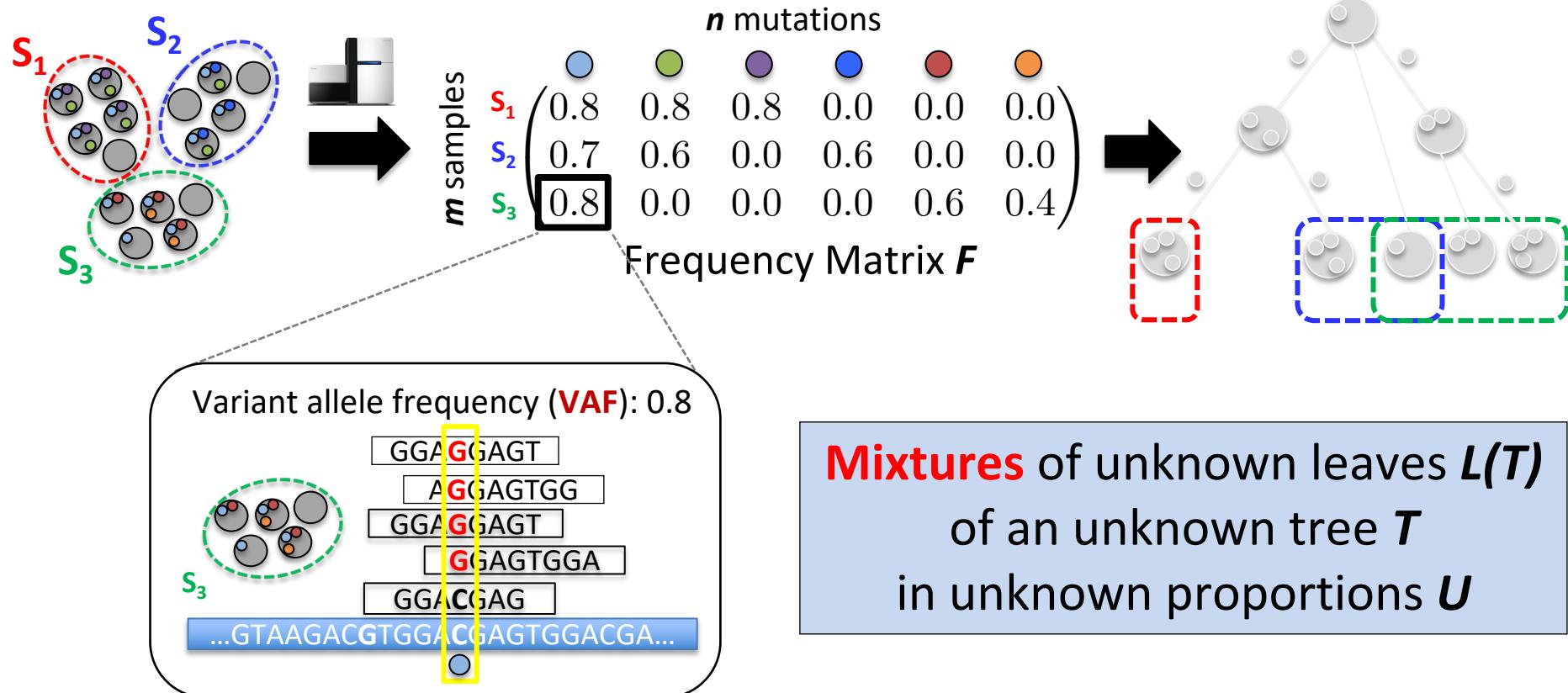
- Multiple consensus tree problem [ISMB/ECCB 2019]

4. Phylogenetic Integration: [WABI 2021, AMB 2022]

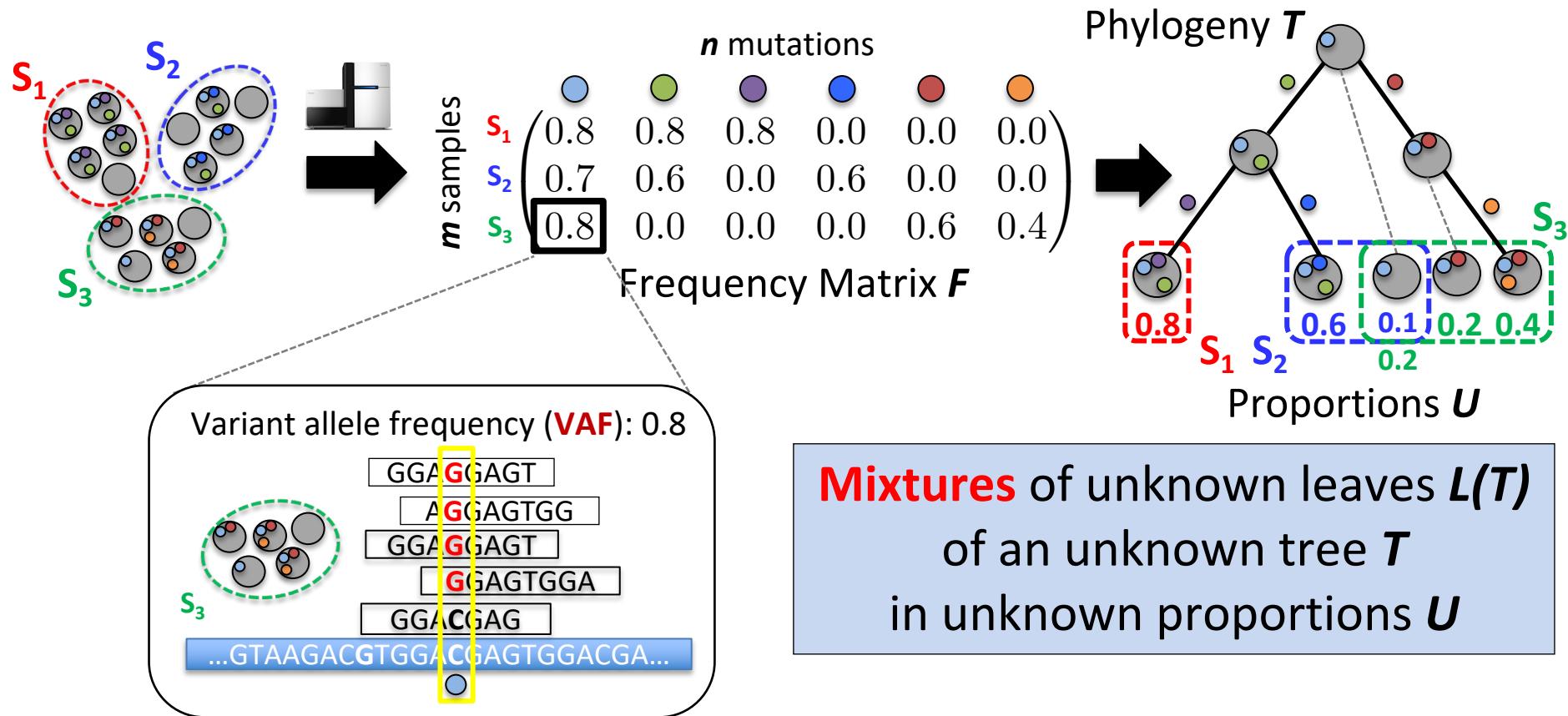
5. Applications

- Mutational signature dynamics [PSB 2020]
- Designing follow-up single-cell sequencing experiments [PLOS CB 2020]
- Visualizing spatial composition of a tumor [ISMB 2020]
- Identifying repeated evolutionary trajectories [ECCB 2020]

Sequencing and Tumor Phylogeny Inference



Sequencing and Tumor Phylogeny Inference



Tumor Phylogeny Inference: Given frequencies F , find phylogeny T and proportions U

Perfect Phylogeny Mixture

Assumptions:

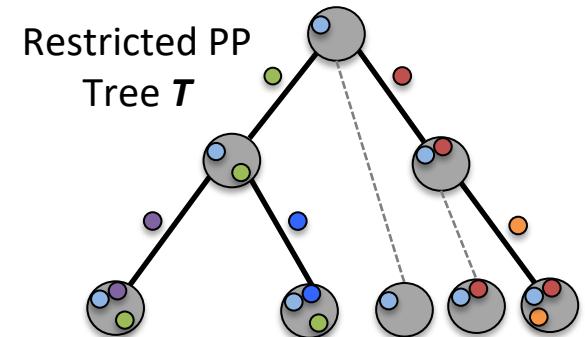
- Infinite sites assumption:
a character changes state once
- Error-free data

$$\begin{array}{c}
 \text{Frequency Matrix } \mathbf{F} \\
 \begin{array}{c}
 \text{m samples} \\
 \text{n mutations} \\
 \begin{matrix} S_1 & S_2 & S_3 \end{matrix} \\
 \begin{pmatrix} 0.8 & 0.8 & 0.8 \\ 0.7 & 0.6 & 0.0 \\ 0.8 & 0.0 & 0.0 \end{pmatrix} \quad \begin{pmatrix} 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.1 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.2 & 0.0 & 0.0 & 0.0 & 0.2 & 0.4 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \\
 \text{clones} \\
 \text{m samples} \\
 \text{n mutations}
 \end{array}
 \end{array}$$

= Mixture Matrix \mathbf{U}

1-1 Equivalent

Restricted PP Matrix \mathbf{B}



Rows of \mathbf{U} are proportions:

$$u_{pj} \geq 0 \text{ and } \sum_j u_{pj} \leq 1$$

Perfect Phylogeny Theorem
 [Estabrook, 1971]
 [Gusfield, 1991]

Perfect Phylogeny Mixture: [El-Kebir*, Oesper* et al., 2015]
 Given \mathbf{F} , find \mathbf{U} and \mathbf{B} such that $\mathbf{F} = \mathbf{U} \mathbf{B}$

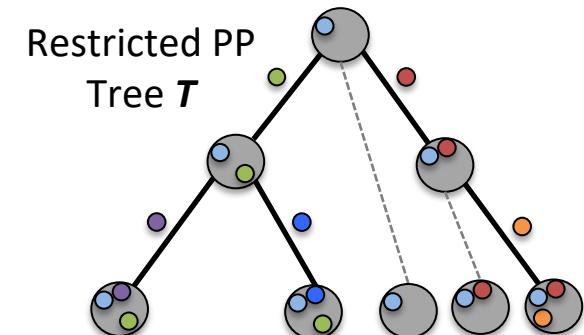
Previous Work

Variant of PPM:

TrAp [Strino *et al.*, 2013], PhyloSub [Jiao *et al.*, 2014]
 CITUP [Malikic *et al.*, 2015], BitPhylogeny [Yuan *et al.*, 2015]
 LICHHeE [Popic *et al.*, 2015], ...

$$\begin{matrix} m \text{ samples} \\ \text{Frequency Matrix } \mathbf{F} \end{matrix} = \begin{matrix} n \text{ mutations} \\ \mathbf{S}_1 \\ \mathbf{S}_2 \\ \mathbf{S}_3 \end{matrix} \left(\begin{matrix} 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{matrix} \right)$$

$$\begin{matrix} m \text{ samples} \\ \text{clones} \\ \text{Mixture Matrix } \mathbf{U} \end{matrix} = \begin{matrix} n \text{ mutations} \\ \mathbf{S}_1 \\ \mathbf{S}_2 \\ \mathbf{S}_3 \end{matrix} \left(\begin{matrix} 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.1 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.2 & 0.0 & 0.0 & 0.0 & 0.2 & 0.4 \end{matrix} \right) \quad \begin{matrix} n \text{ mutations} \\ \text{clones} \\ \text{Restricted PP Matrix } \mathbf{B} \end{matrix}$$



1-1 Equivalent

Rows of \mathbf{U} are proportions:

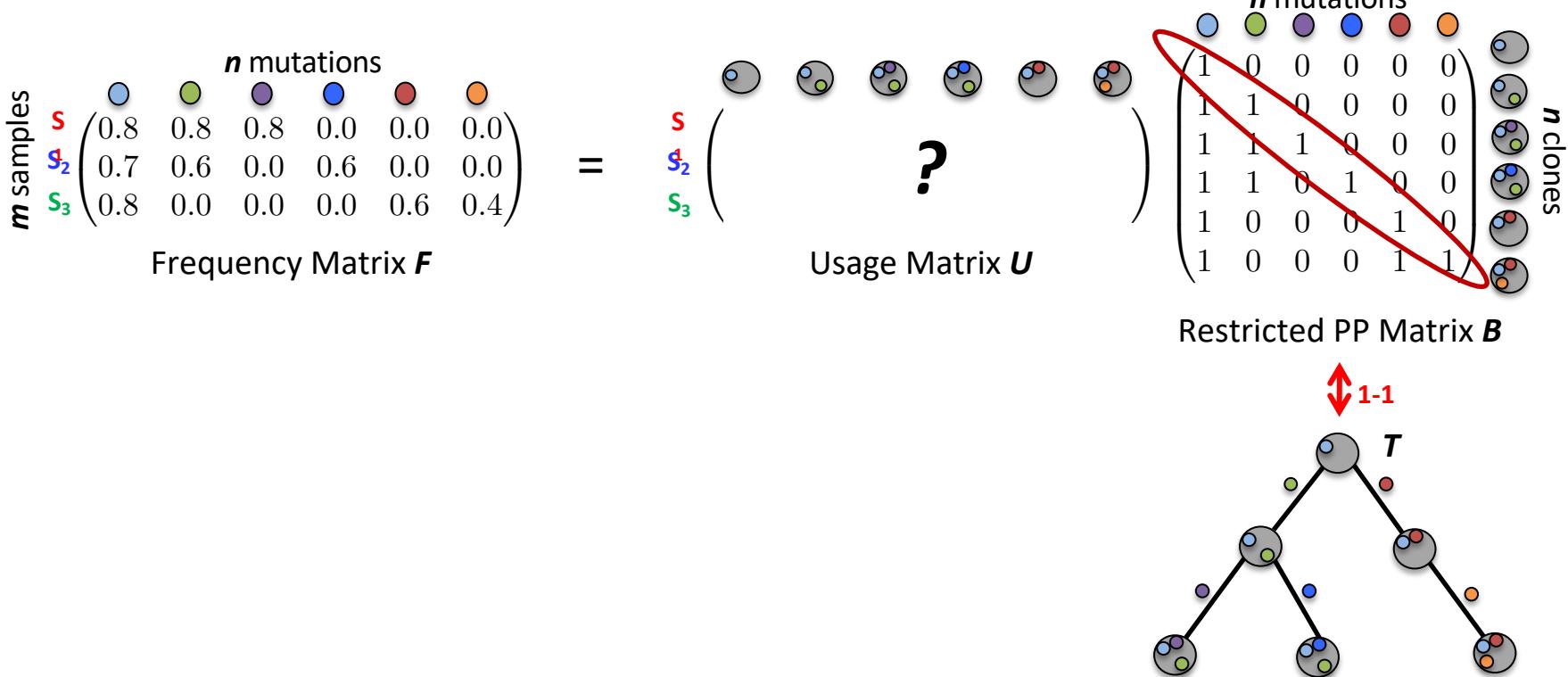
$$u_{pj} \geq 0 \text{ and } \sum_j u_{pj} \leq 1$$

Perfect Phylogeny Theorem
 [Estabrook, 1971]
 [Gusfield, 1991]

Perfect Phylogeny Mixture: [El-Kebir*, Oesper* et al., 2015]
 Given \mathbf{F} , find \mathbf{U} and \mathbf{B} such that $\mathbf{F} = \mathbf{U} \mathbf{B}$

Given F and T (or B), is there a usage matrix U ?

PPM: Given F , find U and B such that $F = UB$



Given F and T (or B), is there a usage matrix U ?

PPM: Given F , find U and B such that $F = UB$

$$\begin{matrix} m \text{ samples} \\ \text{Frequency Matrix } F \end{matrix} = \begin{matrix} n \text{ mutations} \\ S_1 \\ S_2 \\ S_3 \end{matrix} \begin{pmatrix} 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

$$\begin{matrix} n \text{ mutations} \\ U \\ \text{Usage Matrix } U \end{matrix} = \begin{matrix} n \text{ mutations} \\ S_1 \\ S_2 \\ S_3 \end{matrix} \begin{pmatrix} 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.1 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.2 & 0.0 & 0.0 & 0.0 & 0.2 & 0.4 \end{pmatrix}$$

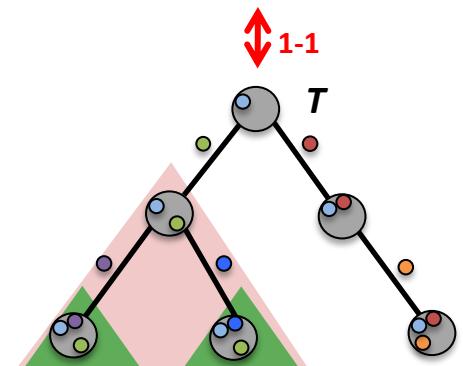
$$\begin{matrix} n \text{ mutations} \\ B \\ \text{Restricted PP Matrix } B \end{matrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Lemma: B is invertible

→ Given F and B , U is **unique**: $U = FB^{-1}$

Lemma:

$$u_{pj} = f_{pj} - \sum_{k \text{ child of } j} f_{pk}$$



Given F and T (or B), is there a usage matrix U ?

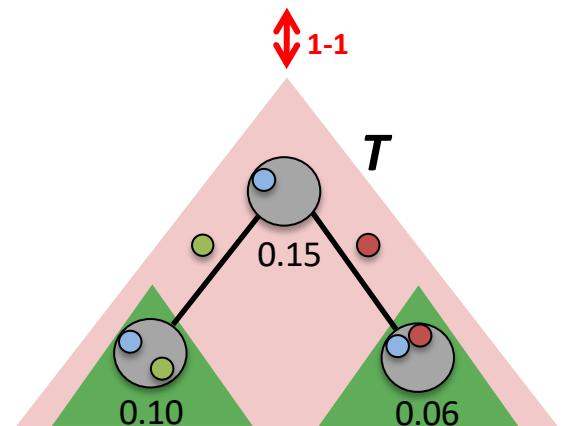
PPM: Given F , find U and B such that $F = UB$

$$\begin{pmatrix} 0.15 & 0.1 & 0.06 \\ 0.2 & 0.05 & 0.04 \end{pmatrix} \quad = \quad \begin{pmatrix} -0.01 & 0.1 & 0.06 \\ 0.11 & 0.05 & 0.04 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

Frequency Matrix F Usage Matrix U Restricted PP Matrix B

Lemma:

$$u_{pj} = f_{pj} - \sum_{k \text{ child of } j} f_{pk}$$



Combinatorial Characterization of Solutions

PPM: Given F , find U and B such that $F = UB$

Lemma:

$$u_{pj} = f_{pj} - \sum_{k \text{ child of } j} f_{pk}$$

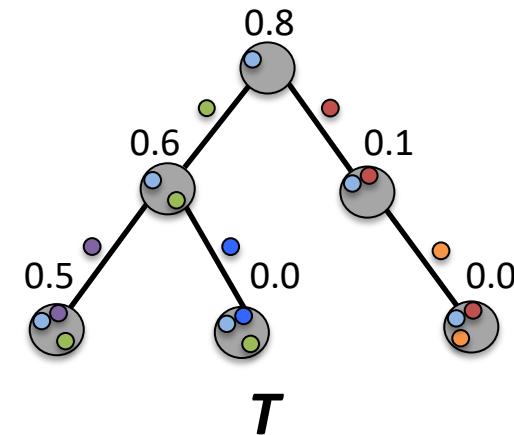
Lemma (Sum Condition):

Given F and T , for all samples p and mutations j ,

$$f_{pj} \geq \sum_{k \text{ child of } j} f_{pk}$$


$$\begin{array}{cccccc} \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ (0.8 & 0.6 & 0.5 & 0.0 & 0.1 & 0.0) \\ (0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0) \\ (0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4) \end{array}$$

F



Combinatorial Characterization of Solutions

PPM: Given F , find U and B such that $F = UB$

Lemma (Sum Condition):

Given F and T , for all samples p and

mutations j , $f_{pj} \geq \sum_{k \text{ child of } j} f_{pk}$



$\begin{matrix} \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ (0.8 & 0.6 & 0.5 & 0.0 & 0.1 & 0.0) \\ (0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0) \\ (0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4) \end{matrix}$

F

Lemma (Ancestry Condition):

Given F and T , for all samples p and

mutations k child of j , $f_{pj} \geq f_{pk}$



Combinatorial Characterization of Solutions

PPM: Given F , find U and B such that $F = UB$

Lemma (Sum Condition):

Given F and T , for all samples p and

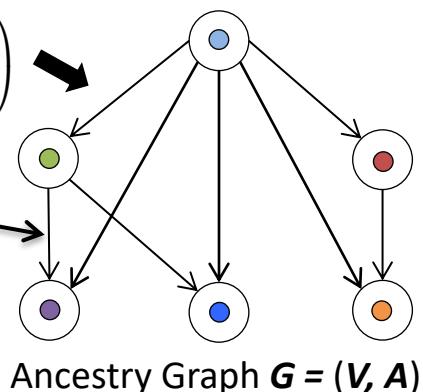
mutations j , $f_{pj} \geq \sum_{k \text{ child of } j} f_{pk}$



$$\begin{pmatrix} 0.8 & 0.6 & 0.5 & 0.0 & 0.1 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

F

potential parental
relationship



Ancestry Graph $G = (V, A)$

Lemma (Ancestry Condition):

Given F and T , for all samples p and

mutations k child of j , $f_{pj} \geq f_{pk}$



Ancestry graph $G = (V, A)$; given F

- Vertex for every mutation
- Edge $(j, k) \in A$ iff $f_{pj} \geq f_{pk}$
for all samples p

Combinatorial Characterization of Solutions

PPM: Given F , find U and B such that $F = UB$

Lemma (Sum Condition):

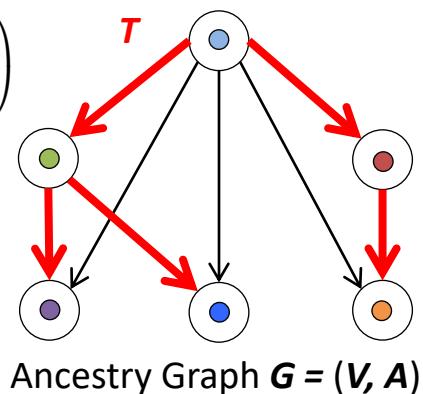
Given F and T , for all samples p and

mutations j , $f_{pj} \geq \sum_{k \text{ child of } j} f_{pk}$



$$\begin{matrix} F \\ \begin{pmatrix} 0.8 & 0.6 & 0.5 & 0.0 & 0.1 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix} \end{matrix}$$

F



Theorem 2:

PPM is NP-complete even if $m = 2$

Lemma (Ancestry Condition):

Given F and T , for all samples p and

mutations k child of j , $f_{pj} \geq f_{pk}$



Ancestry graph $G = (V, A)$; given F

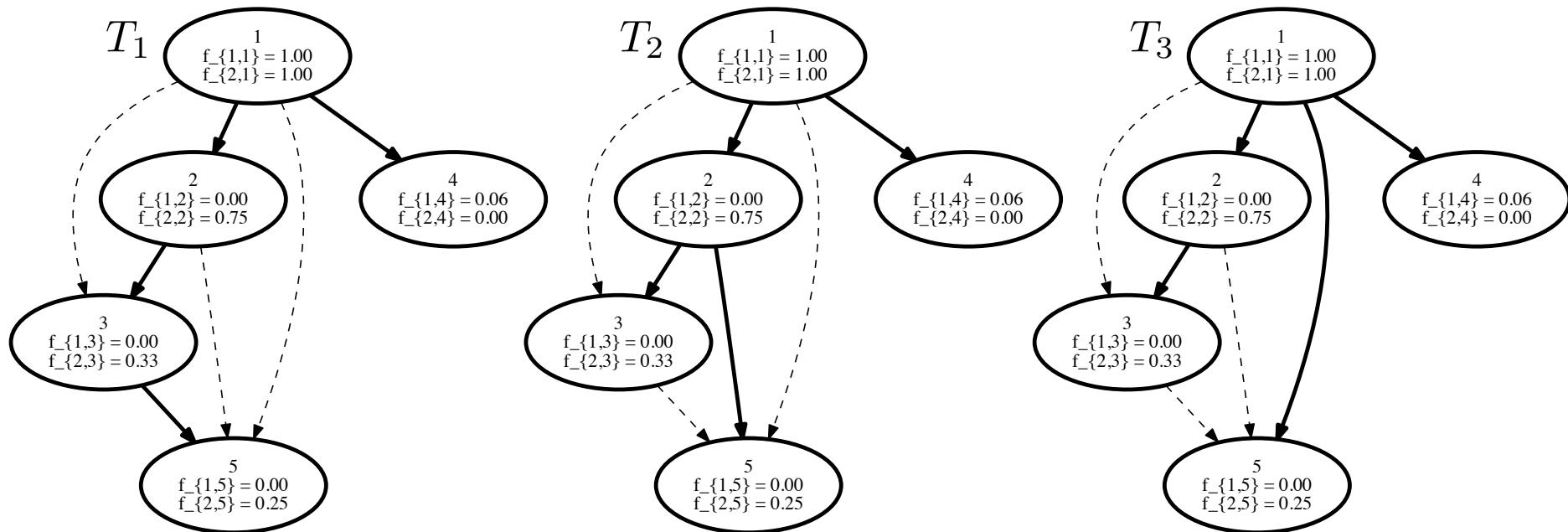
- Vertex for every mutation
- Edge $(j, k) \in A$ iff $f_{pj} \geq f_{pk}$ for all samples p

Theorem 1:

T is a solution to the PPM if and only if

T is a spanning tree of G satisfying the Sum Condition

Non-uniqueness of Solutions to PPM



$$F = \begin{pmatrix} 1 & 0 & 0 & 0.06 & 0 \\ 1 & 0.75 & 0.33 & 0 & 0.25 \end{pmatrix}$$

Question 1: Can we determine the number of solutions?

Question 2: Can we sample solutions uniformly at random?

On the Complexity of #PPM

Question 1: Can we determine the number of solutions?

Question 2: Can we sample solutions uniformly at random?

#PPM: Given F , count the number of pairs (U, B) composed of mixture matrix U and perfect phylogeny matrix B such that $F = UB$

On the Complexity of #PPM

Question 1: Can we determine the number of solutions?

Question 2: Can we sample solutions uniformly at random?

#PPM: Given F , count the number of pairs (U, B) composed of mixture matrix U and perfect phylogeny matrix B such that $F = UB$

#P is the complexity class of counting problems whose decision problems are in NP

Every problem in #P can be reduced in polynomial time to any problem in #P-complete, preserving cardinalities

On the Complexity of #PPM

Question 1: Can we determine the number of solutions?

Question 2: Can we sample solutions uniformly at random?

#PPM: Given F , count the number of pairs (U, B) composed of mixture matrix U and perfect phylogeny matrix B such that $F = UB$

#P is the complexity class of counting problems whose decision problems are in NP

Every problem in #P can be reduced in polynomial time to any problem in #P-complete, preserving cardinalities

Theorem: #PPM is #P-complete

Theorem: There is no FPRAS for #PPM

Theorem: There is no FPAUS for PPM



Yuanyuan Qi

Outline

1. Background and theory: [RECOMB-CG 2018, AMB 2019]

- Perfect Phylogeny Mixture (PPM) problem
- #PPM: exact counting and uniform sampling

2. Simulation results: [RECOMB-CG 2018]

- What contributes to non-uniqueness?
- How to reduce non-uniqueness?



3. Summarizing solution space: [ISMB/ECCB 2019]

- Multiple consensus tree problem

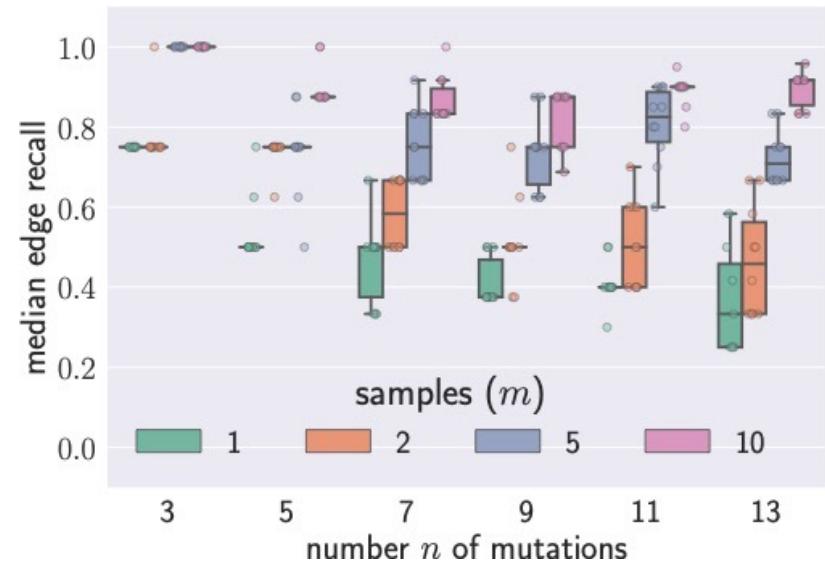
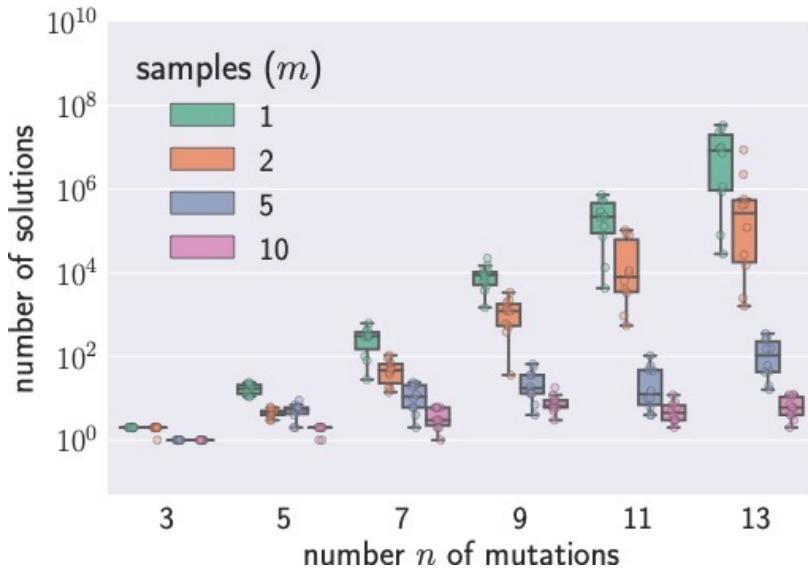
Dikshant Pradhan

4. Phylogenetic integration: [WABI 2021/AMB 2022]

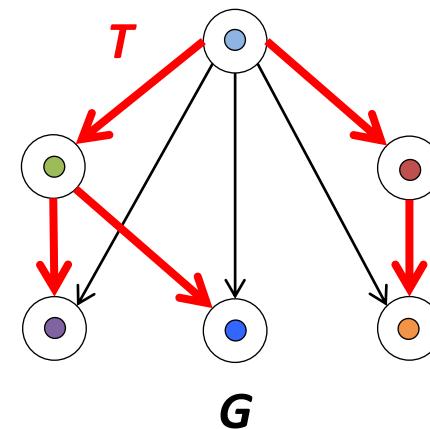
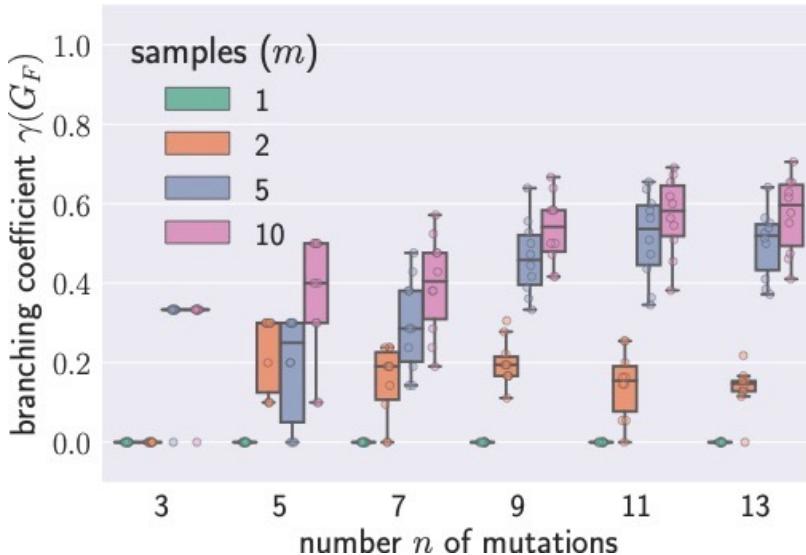
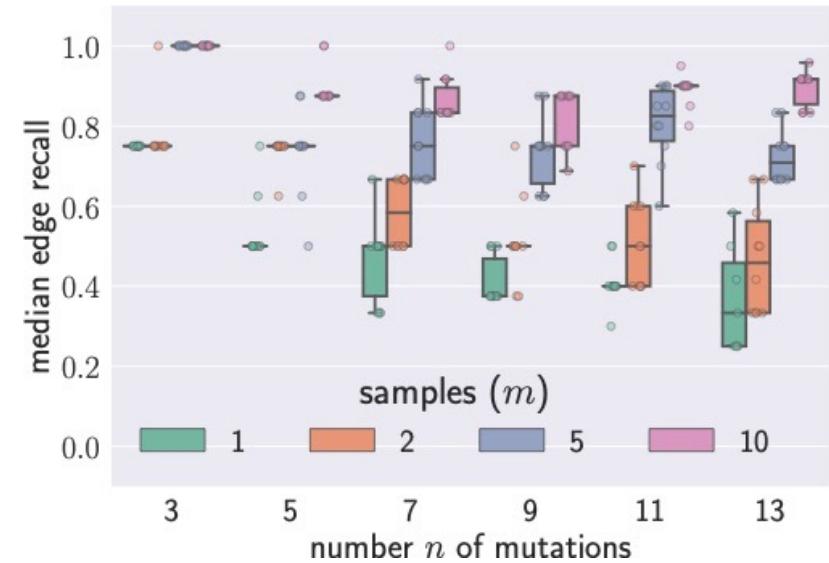
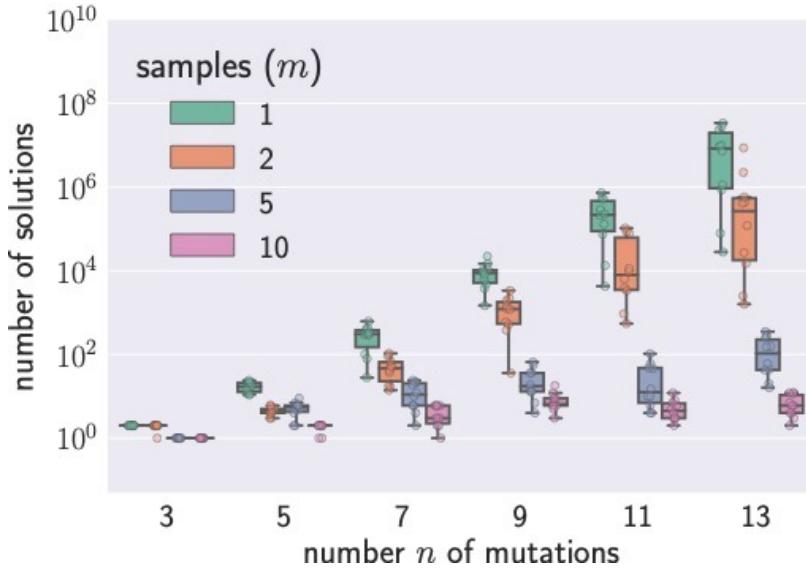
5. Applications

- Mutational signature dynamics [PSB 2020]
- Designing follow-up single-cell sequencing experiments [PLOS CB 2020]
- Visualizing spatial composition of a tumor [ISMB 2020]
- Identifying repeated evolutionary trajectories [ECCB 2020]

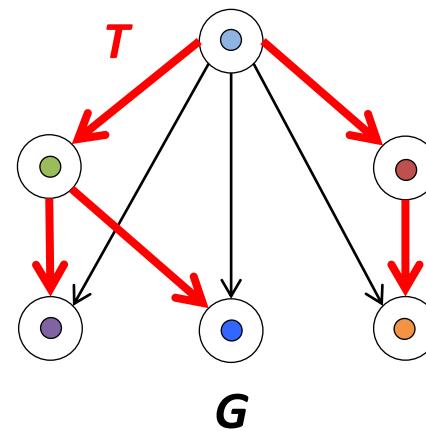
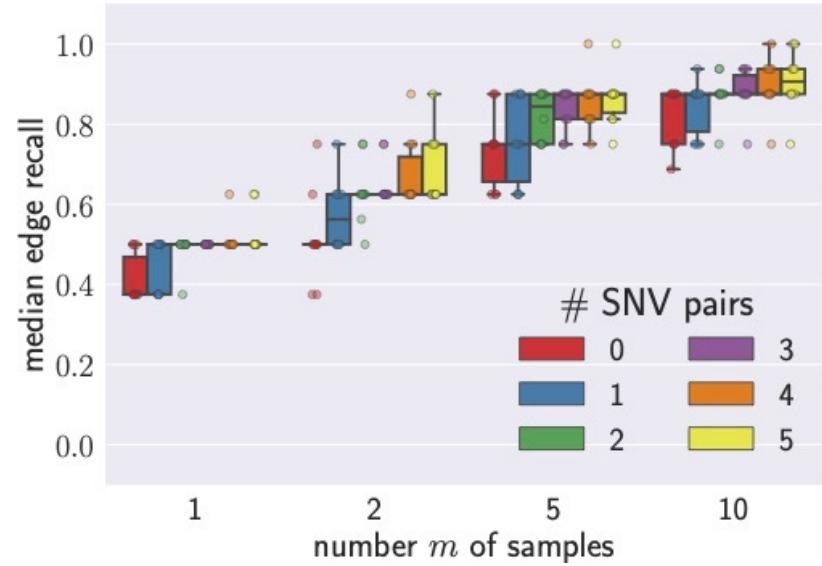
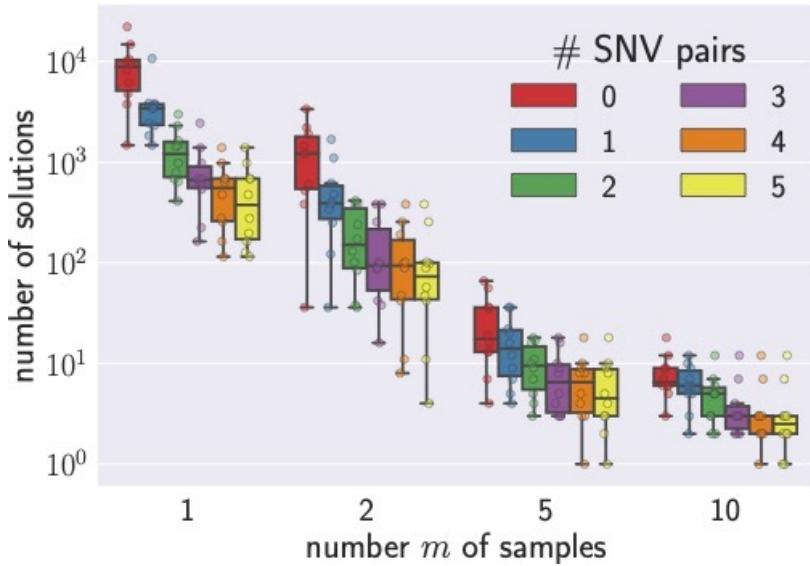
What Contributors to Non-uniqueness?



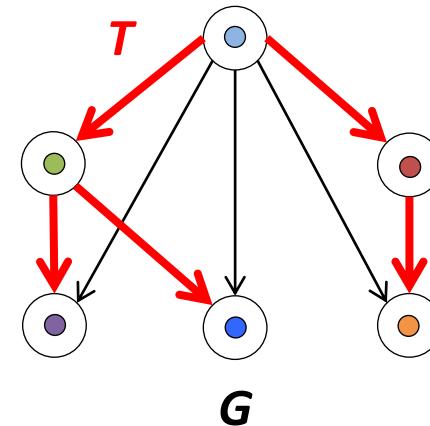
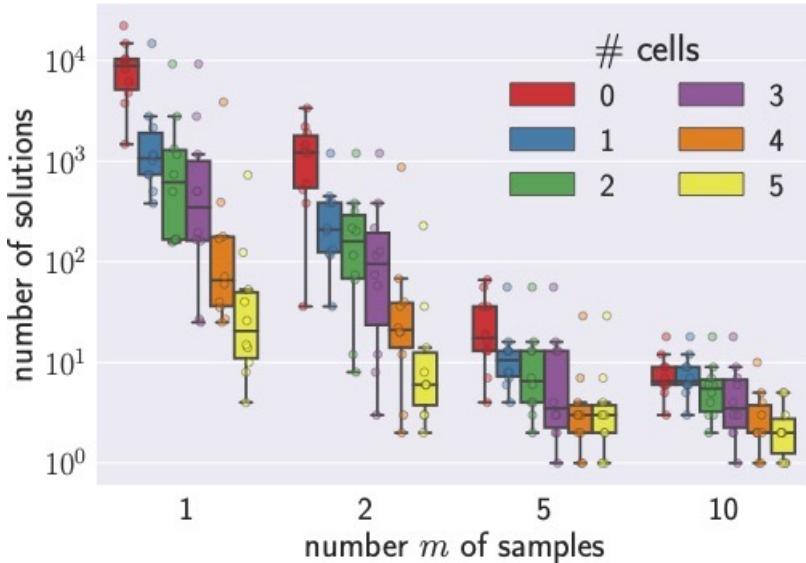
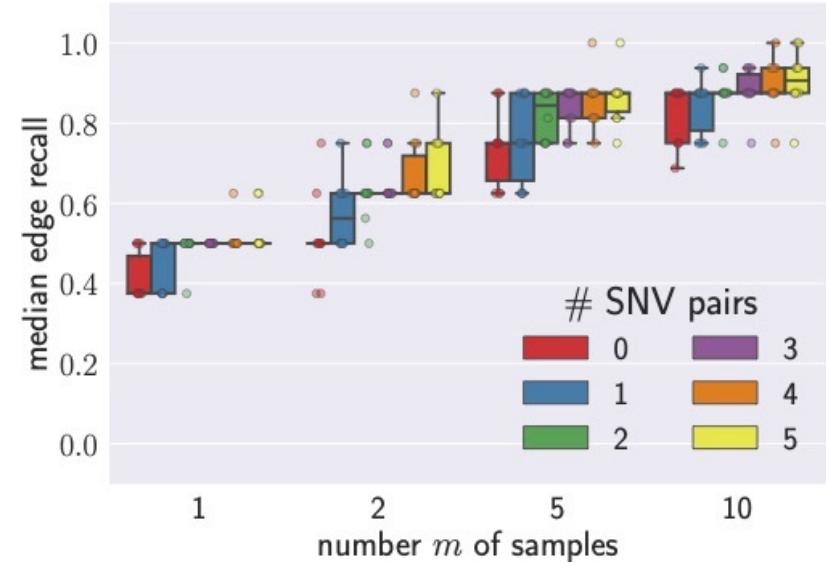
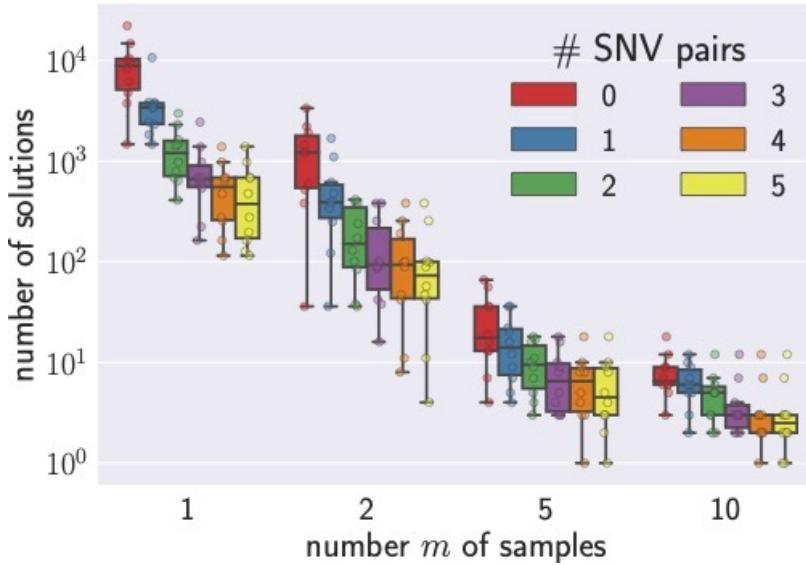
What Contributors to Non-uniqueness?



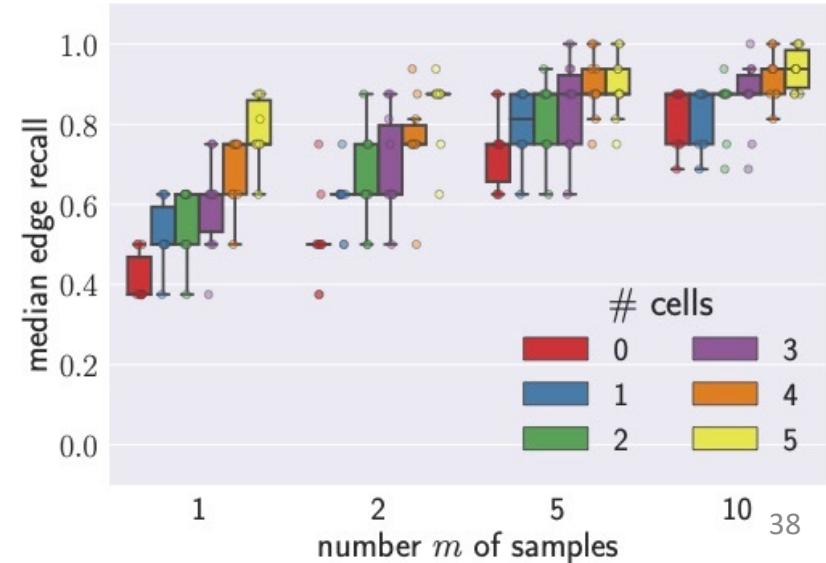
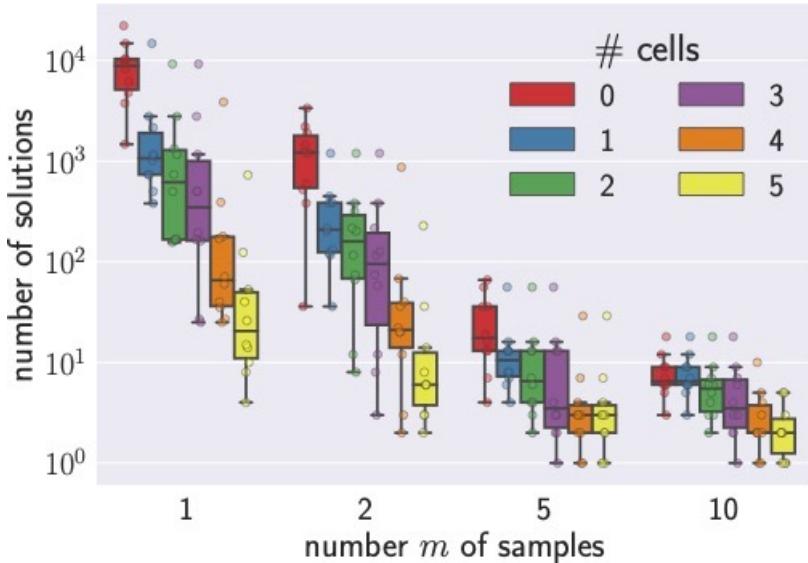
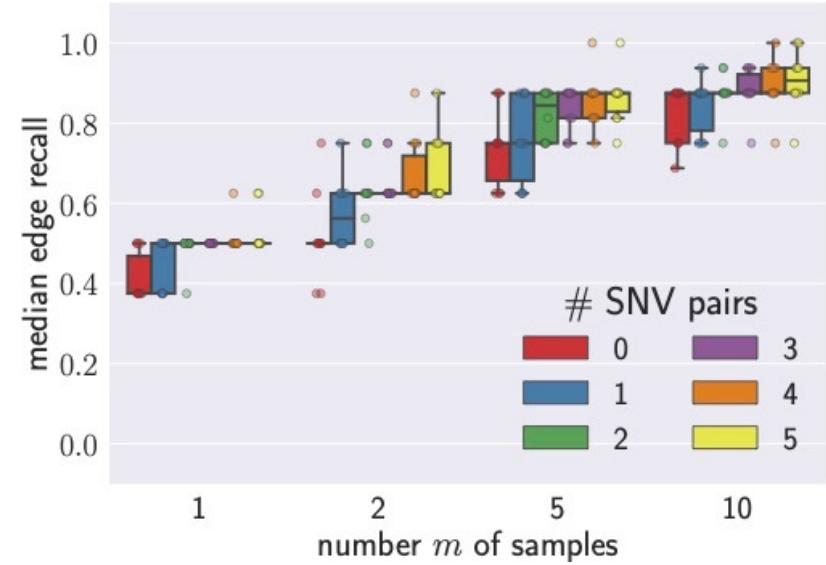
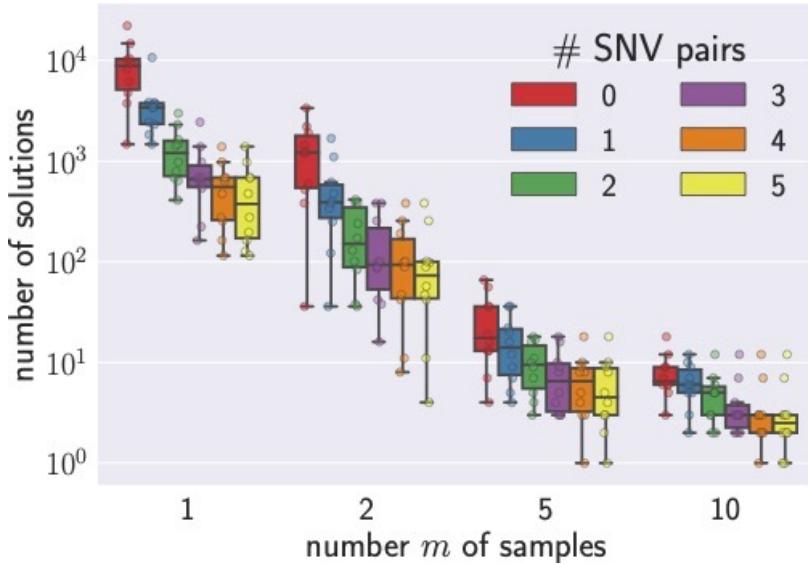
How to Reduce Non-Uniqueness?



How to Reduce Non-Uniqueness?



How to Reduce Non-Uniqueness?



Outline

1. Background and theory: [RECOMB-CG 2018 / AMB 2019]

- Perfect Phylogeny Mixture (PPM) problem
- #PPM: exact counting and uniform sampling

2. Simulation results: [RECOMB-CG 2018]

- What contributes to non-uniqueness?
- How to reduce non-uniqueness?

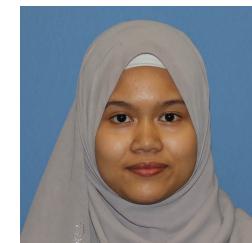
3. Summarizing solution space: [ISMB/ECCB 2019]

- Multiple consensus tree problem

4. Phylogenetic integration: [WABI 2021 / AMB 2022]

5. Applications

- Mutational signature dynamics [PSB 2020]
- Designing follow-up single-cell sequencing experiments [PLOS CB 2020]
- Visualizing spatial composition of a tumor [ISMB 2020]
- Identifying repeated evolutionary trajectories [ECCB 2020]



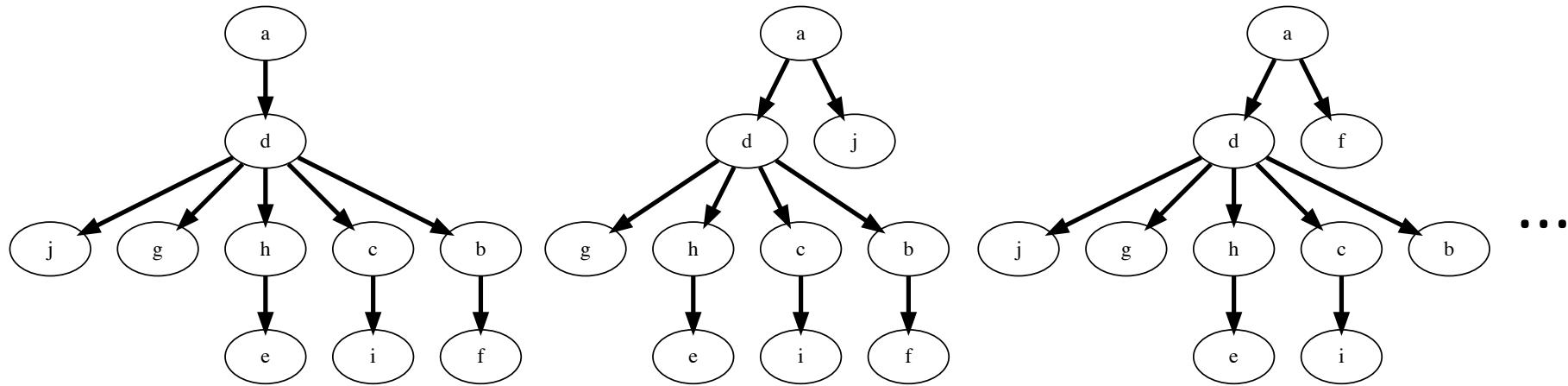
Nuraini Aguse



Yuanyuan Qi

Lung Cancer Patient: CRUK0037

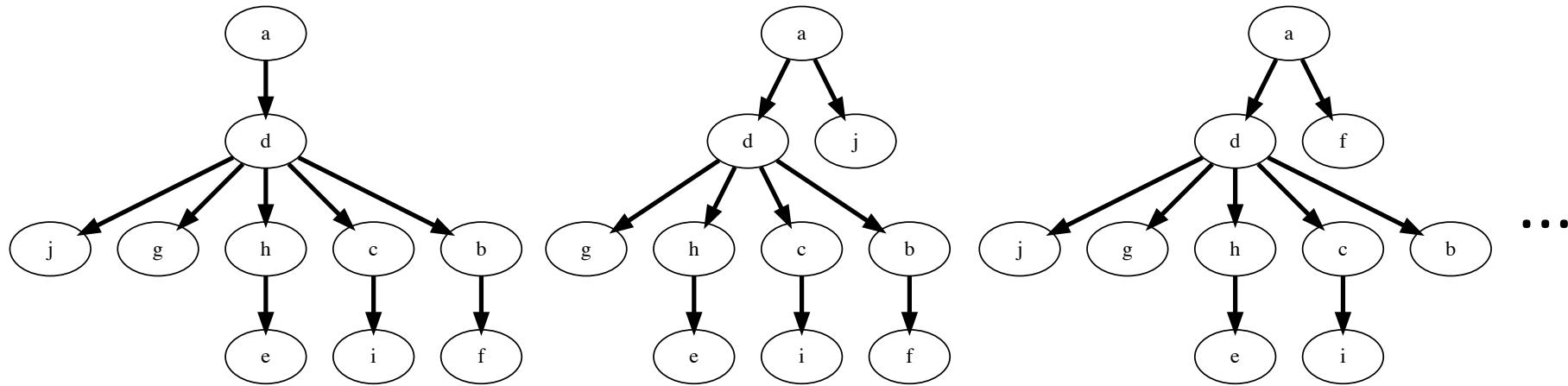
Jamal-Hanjani et al. (2017). *New England Journal of Medicine*, 376(22), 2109–2121.



Authors inferred 17 trees

Lung Cancer Patient: CRUK0037

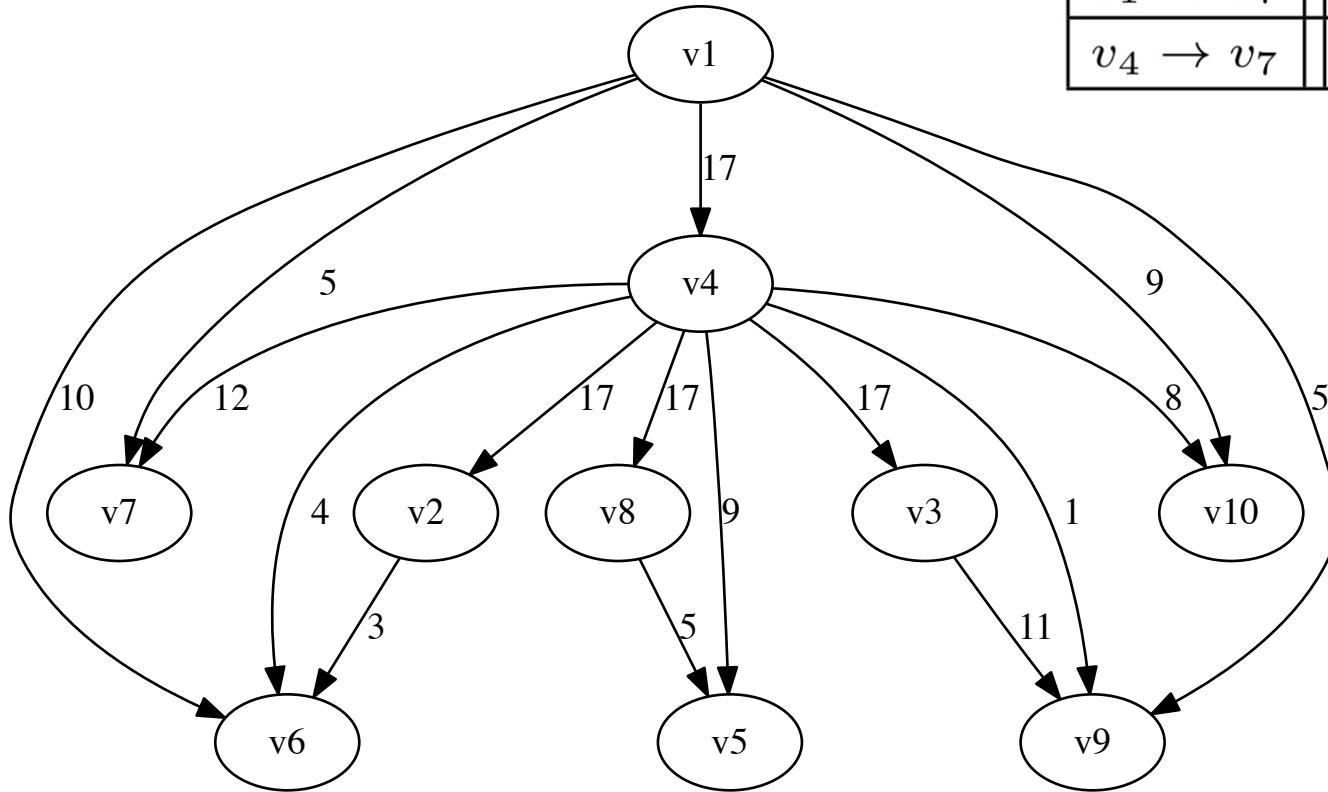
Jamal-Hanjani et al. (2017). *New England Journal of Medicine*, 376(22), 2109–2121.



Authors inferred 17 trees

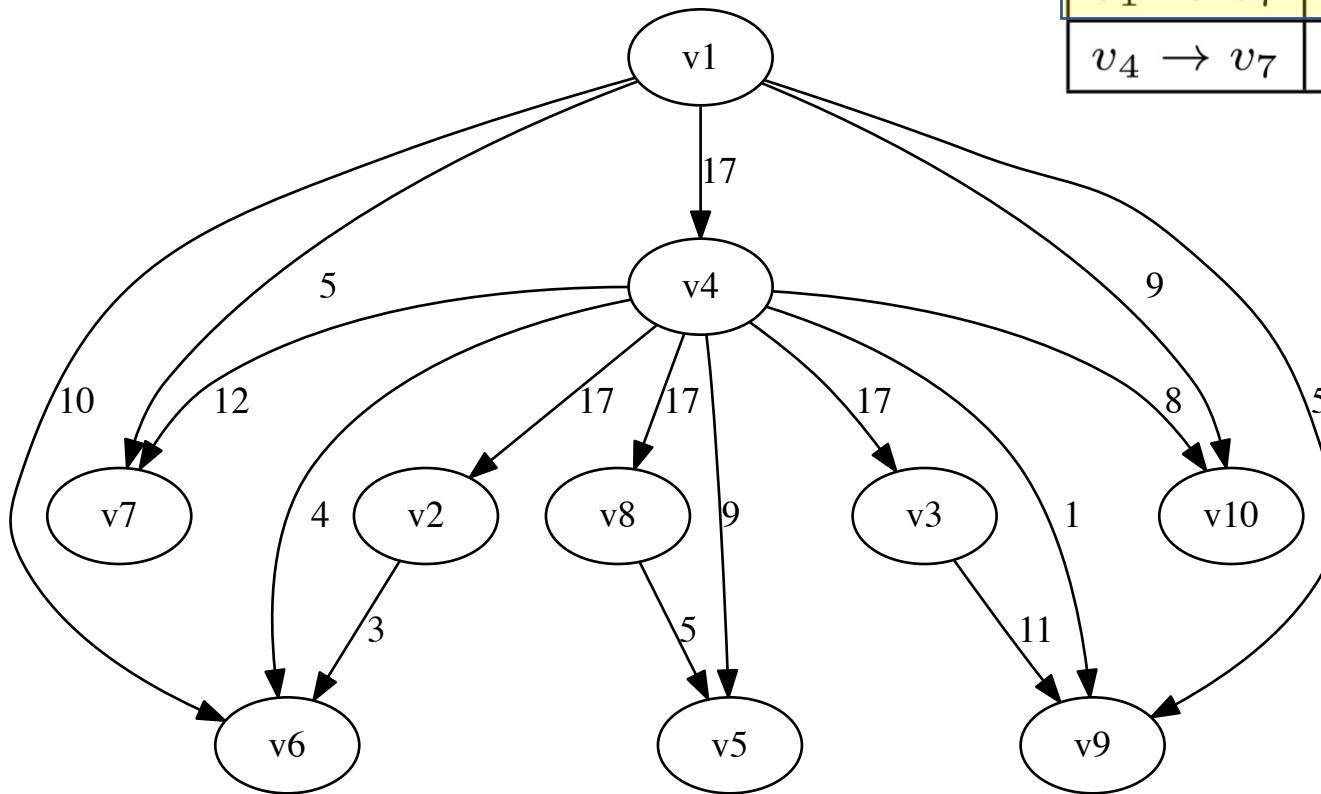
Question: How to summarize solution space in order to remove inference errors and identify dependencies among mutations?

Parent-child Graph: Union of all Edges



$v_1 \rightarrow v_{10}$	$v_4 \rightarrow v_{10}$
$v_1 \rightarrow v_7$	2
$v_4 \rightarrow v_7$	2

Parent-child Graph: Union of all Edges



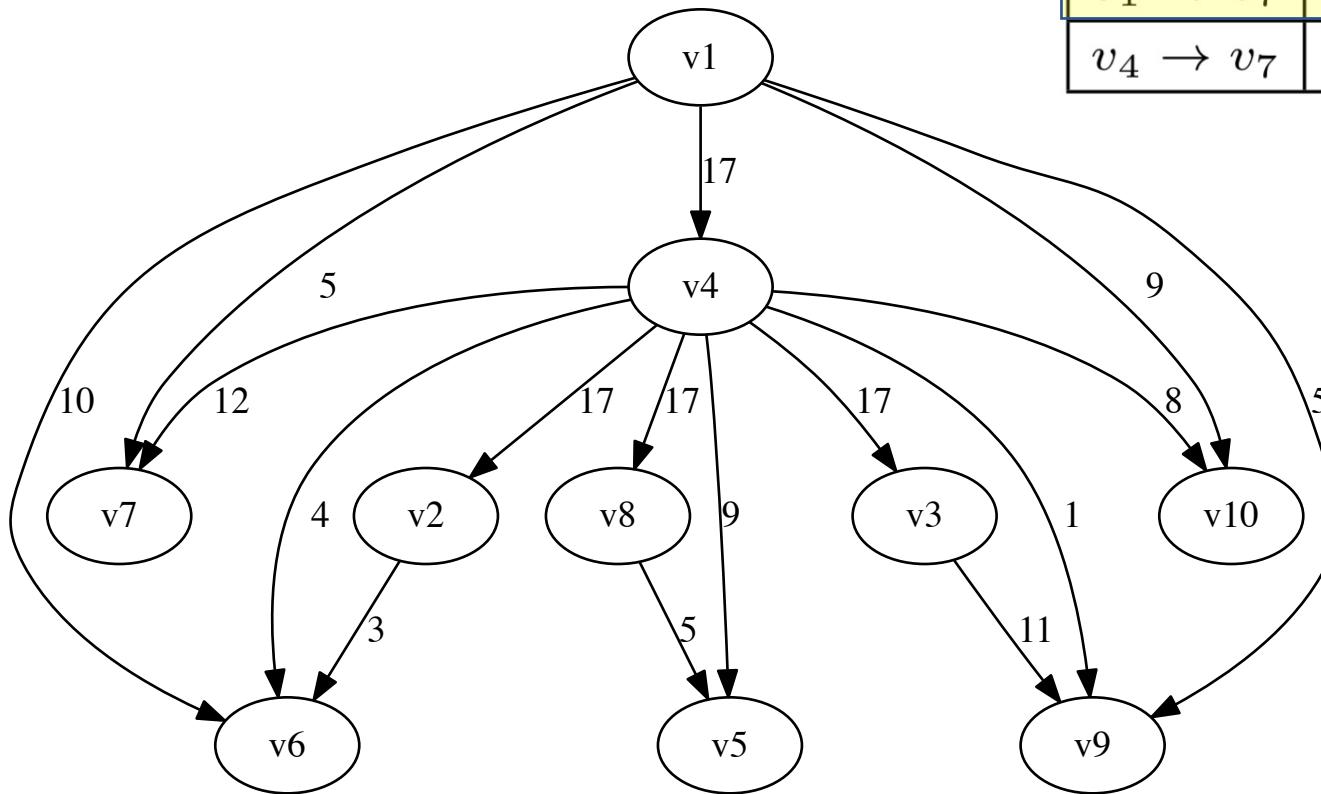
$v_1 \rightarrow v_{10}$	$v_4 \rightarrow v_{10}$
$v_1 \rightarrow v_7$	2
$v_4 \rightarrow v_7$	2

0

5

The parent-child graph does capture patterns of mutual exclusivity

Parent-child Graph: Union of all Edges



$v_1 \rightarrow v_{10}$	$v_4 \rightarrow v_{10}$
$v_1 \rightarrow v_7$	2
$v_4 \rightarrow v_7$	2

0

5

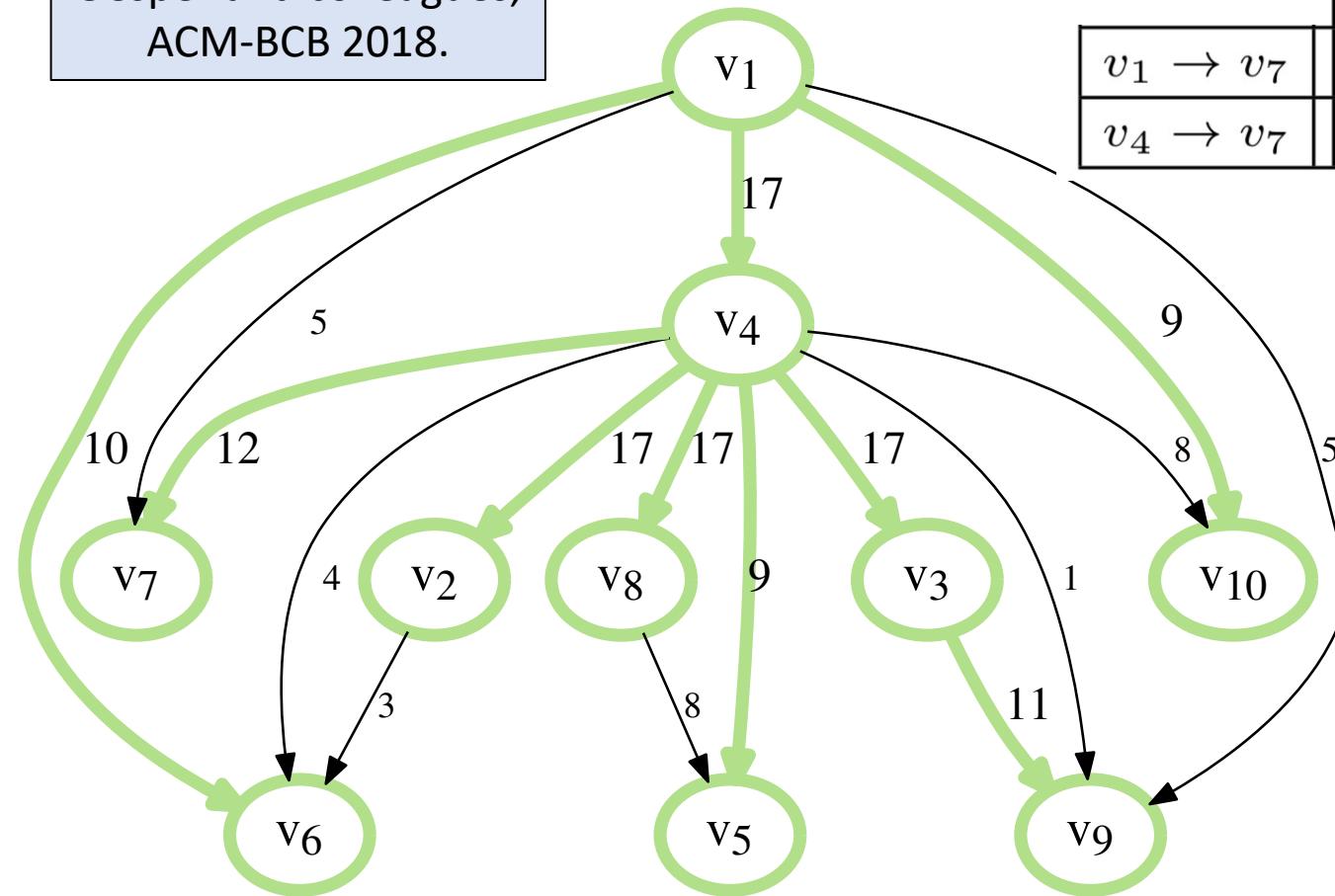
The parent-child graph does capture patterns of mutual exclusivity

Question: Can we infer a single consensus tree?

Single Consensus Tree: Max Weight Spanning Tree

Oesper and colleagues,
ACM-BCB 2018.

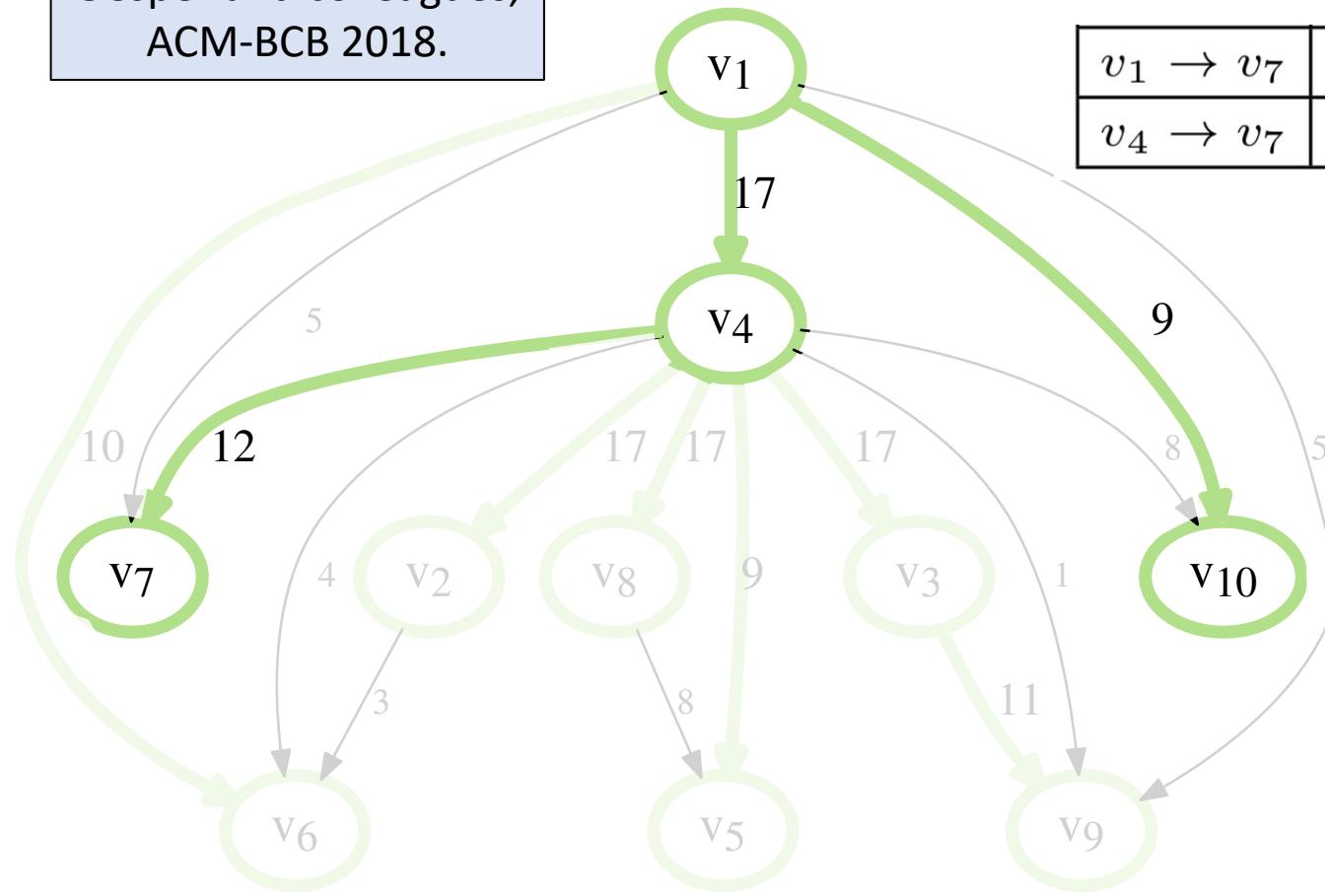
$v_1 \rightarrow v_{10}$	$v_4 \rightarrow v_{10}$
2	0
2	5



Single Consensus Tree: Max Weight Spanning Tree

Oesper and colleagues,
ACM-BCB 2018.

	$v_1 \rightarrow v_{10}$	$v_4 \rightarrow v_{10}$
$v_1 \rightarrow v_7$	2	0
$v_4 \rightarrow v_7$	2	5

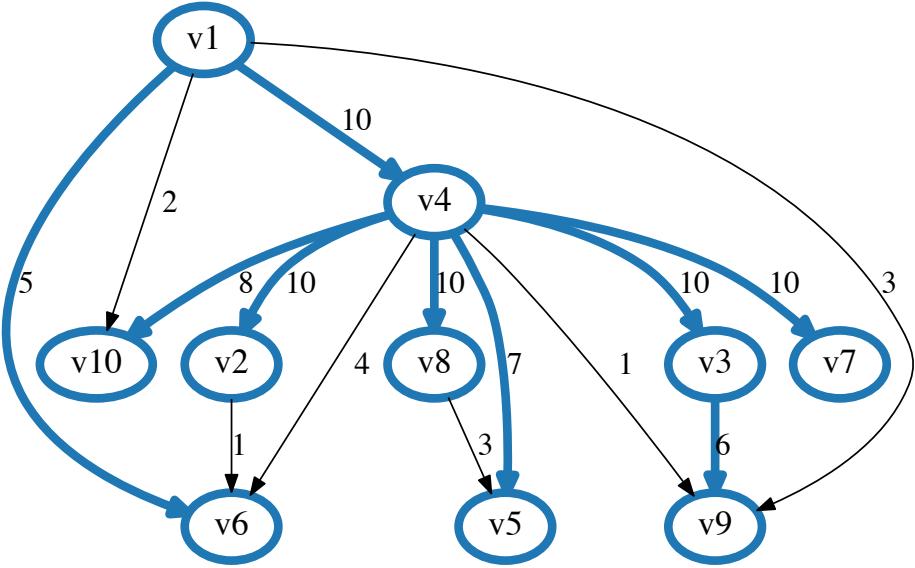
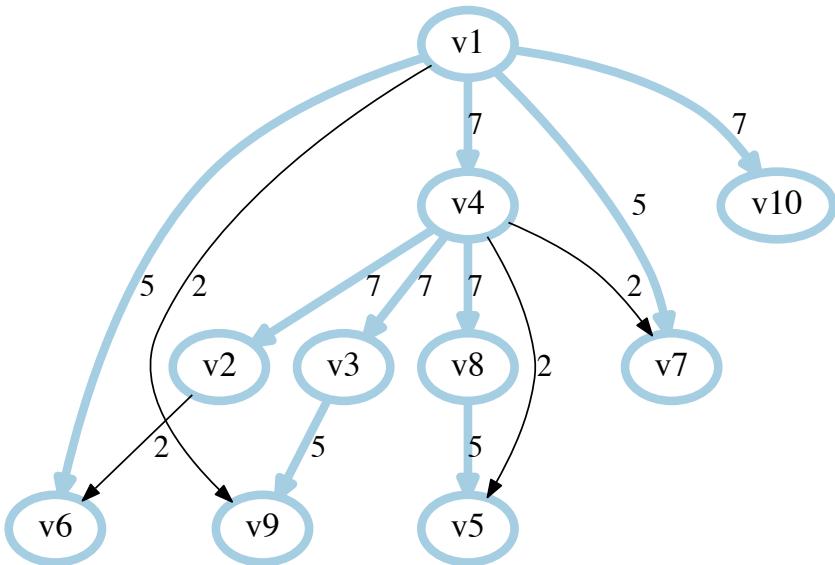


Inaccurate summary for diverse solution spaces

Question: How about inferring multiple consensus trees?

Multiple Consensus Trees

Simultaneous clustering and consensus tree inference

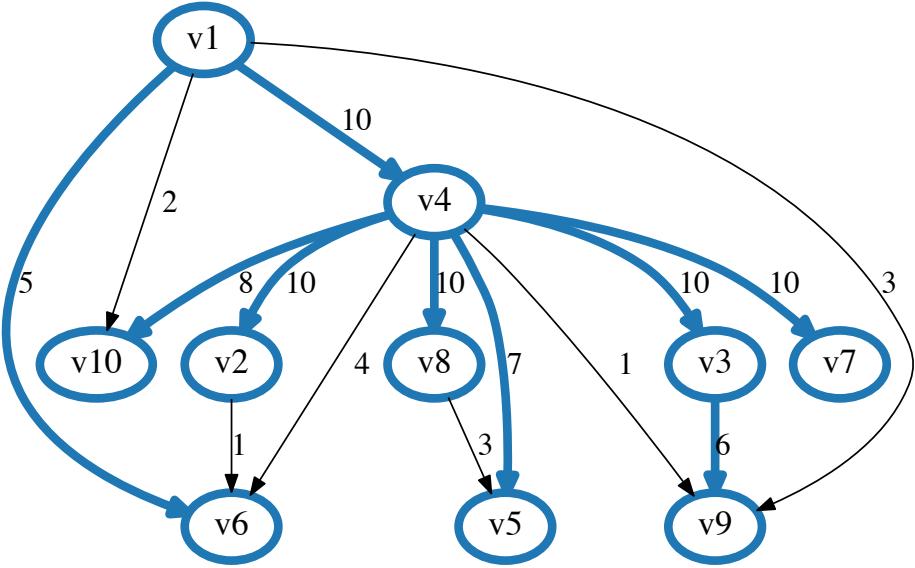
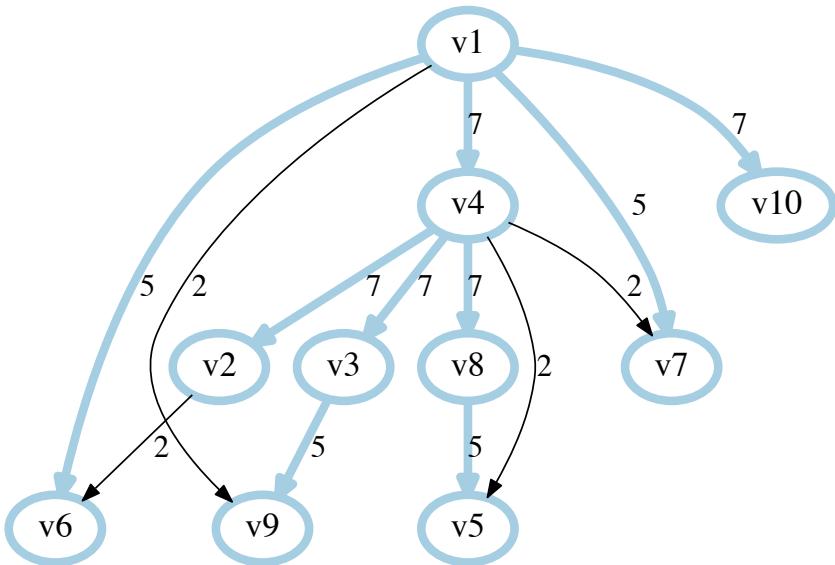


Multiple Consensus Trees (MCT): [ISMB 2019]

Given trees $\mathcal{T} = \{T_1, \dots, T_n\}$, find surjective clustering $\sigma : [n] \rightarrow [k]$ and consensus trees $\mathcal{R} = \{R_1, \dots, R_k\}$ such that $\sum_{i=1}^n d(T_i, R_{\sigma(i)})$ is minimum

Multiple Consensus Trees

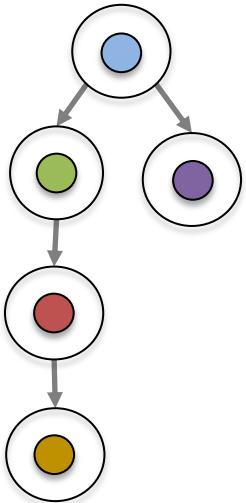
Simultaneous clustering and consensus tree inference



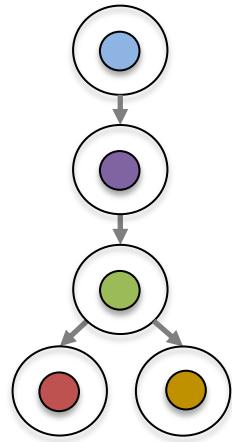
Multiple Consensus Trees (MCT): [ISMB 2019]

Given trees $\mathcal{T} = \{T_1, \dots, T_n\}$, find surjective clustering $\sigma : [n] \rightarrow [k]$ and consensus trees $\mathcal{R} = \{R_1, \dots, R_k\}$ such that $\sum_{i=1}^n d(T_i, R_{\sigma(i)})$ is minimum

Parent-child Distance Function

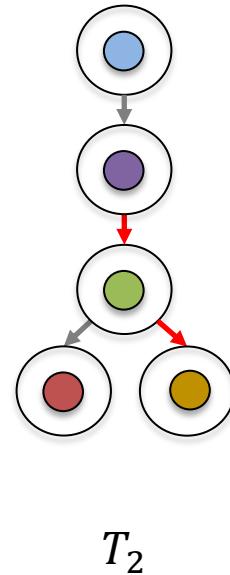
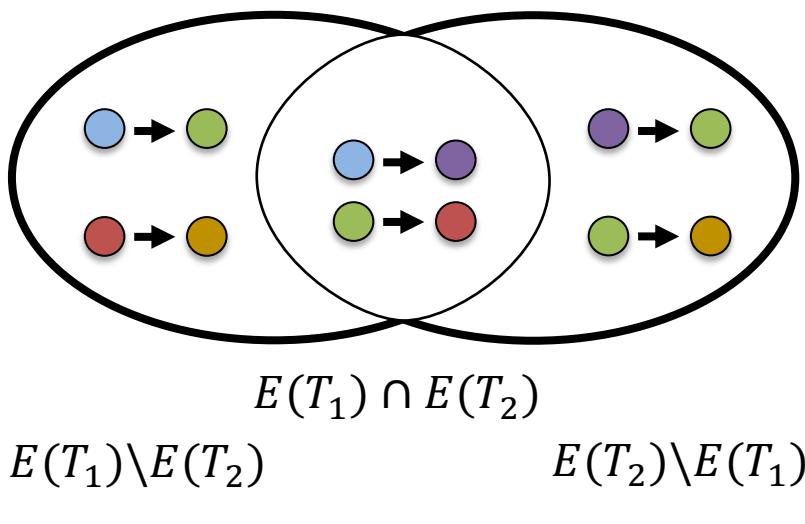
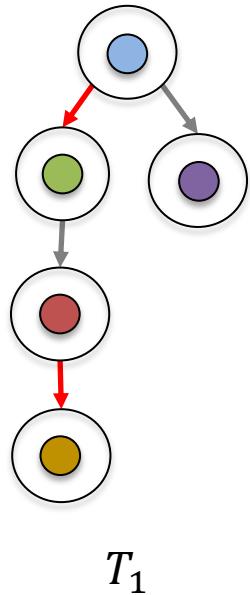


T_1

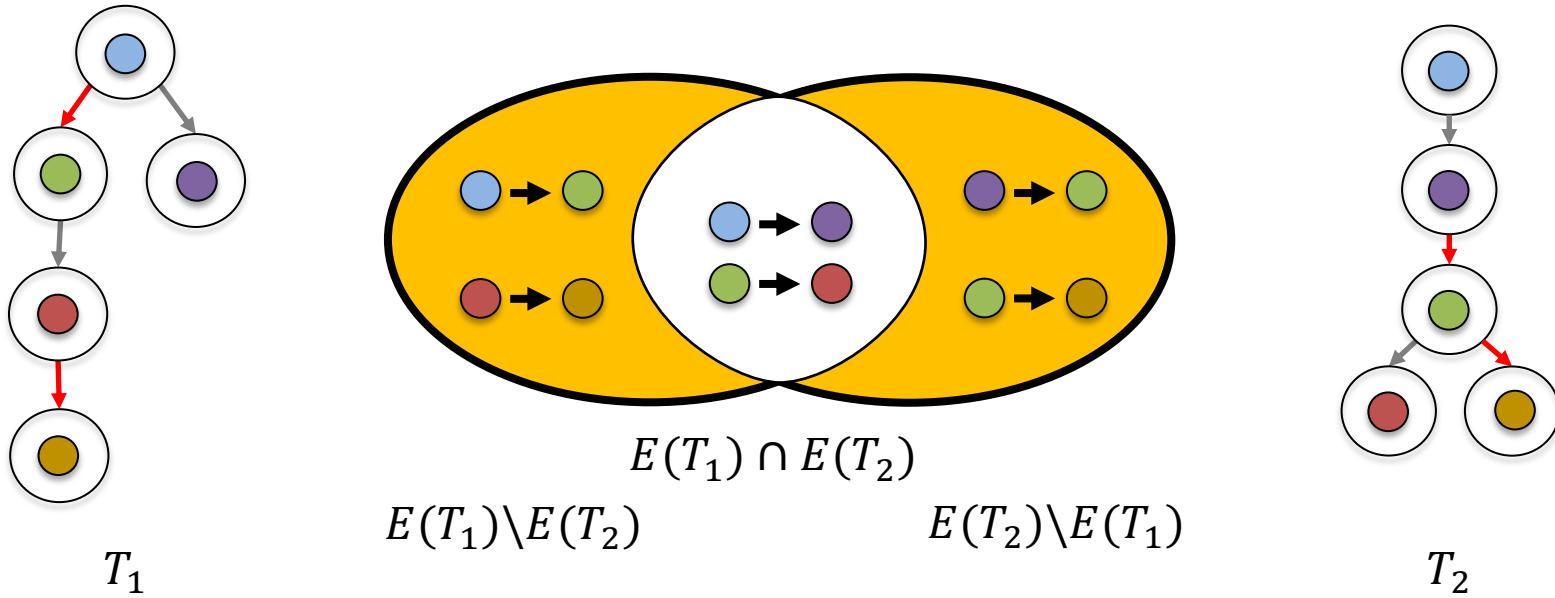


T_2

Parent-child Distance Function



Parent-child Distance Function



Parent-child distance $d(T_1, T_2)$ is the size of the **symmetric difference** of the edge sets

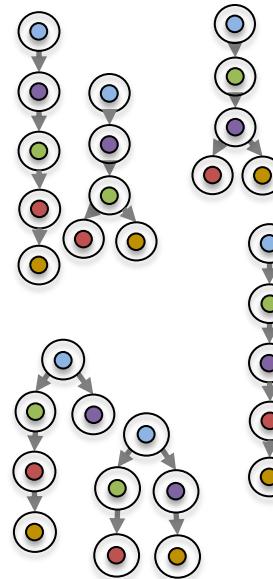
Here, $d(T_1, T_2) = |E(T_1) \setminus E(T_2)| + |E(T_2) \setminus E(T_1)| = 4$.

Combinatorial Characterization of Solutions

Single Consensus Trees (SCT):

[Govek et al., ACM-BCB 2018]

Given $\mathcal{T} = \{T_1, \dots, T_n\}$, find consensus tree R
s.t. $\sum_{i=1}^n d(T_i, R)$ is minimum



Solution Space \mathcal{T}

Combinatorial Characterization of Solutions

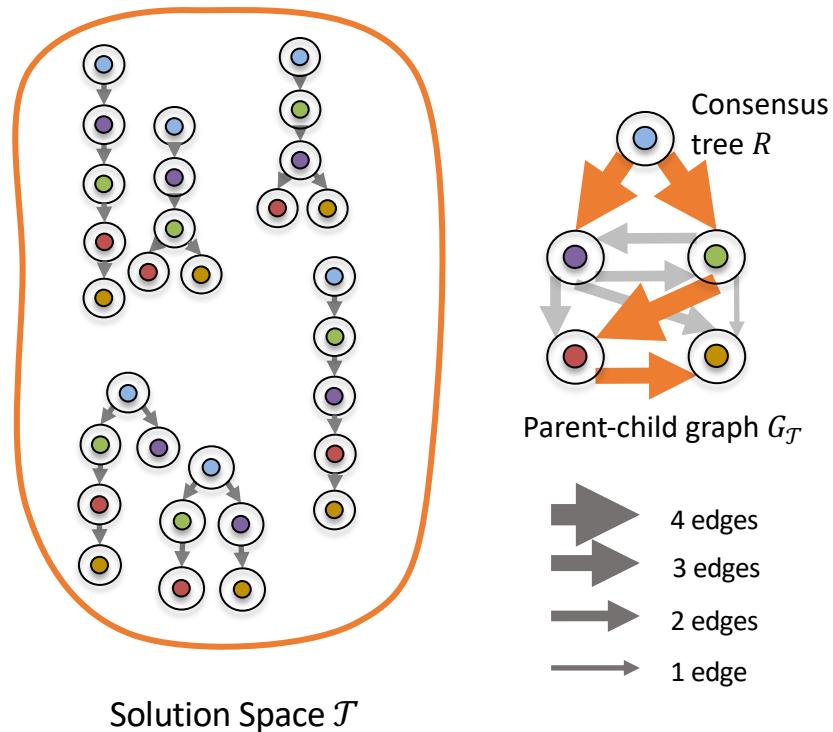
Single Consensus Trees (SCT):

[Govek et al., ACM-BCB 2018]

Given $\mathcal{T} = \{T_1, \dots, T_n\}$, find consensus tree R
s.t. $\sum_{i=1}^n d(T_i, R)$ is minimum

Theorem: [Govek et al., ACM-BCB 2018]

Max weight spanning arborescences
of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT



Combinatorial Characterization of Solutions

Single Consensus Trees (SCT):

[Govek et al., ACM-BCB 2018]

Given $\mathcal{T} = \{T_1, \dots, T_n\}$, find consensus tree R
s.t. $\sum_{i=1}^n d(T_i, R)$ is minimum

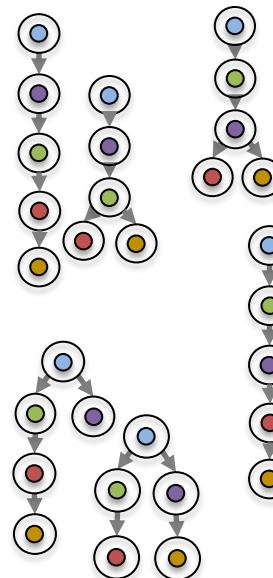
Theorem: [Govek et al., ACM-BCB 2018]

Max weight spanning arborescences
of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT

Multiple Consensus Trees (MCT):

[Aguse et al., ISMB 2019]

Given $\mathcal{T} = \{T_1, \dots, T_n\}$ and $k > 0$, find surjective
clustering $\sigma : [n] \rightarrow [k]$ and consensus trees
 $\mathcal{R} = \{R_1, \dots, R_k\}$ s.t. $\sum_{i=1}^n d(T_i, R_{\sigma(i)})$ is
minimum



Solution Space \mathcal{T}

Combinatorial Characterization of Solutions

Single Consensus Trees (SCT):

[Govek et al., ACM-BCB 2018]

Given $\mathcal{T} = \{T_1, \dots, T_n\}$, find consensus tree R
s.t. $\sum_{i=1}^n d(T_i, R)$ is minimum

Theorem: [Govek et al., ACM-BCB 2018]

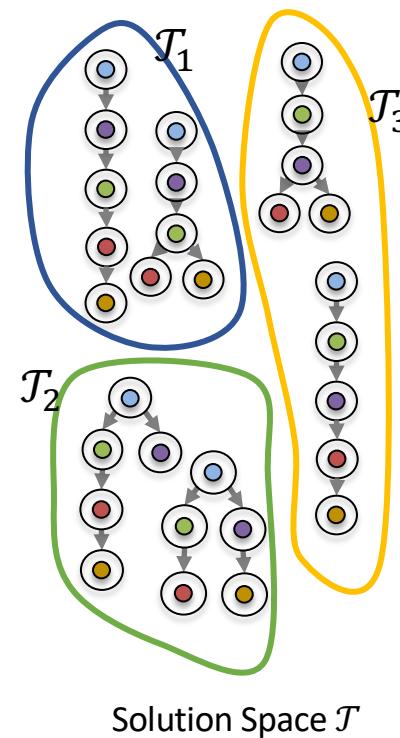
Max weight spanning arborescences
of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT

Multiple Consensus Trees (MCT):

[Aguse et al., ISMB 2019]

Given $\mathcal{T} = \{T_1, \dots, T_n\}$ and $k > 0$, find surjective
clustering $\sigma : [n] \rightarrow [k]$ and consensus trees

$\mathcal{R} = \{R_1, \dots, R_k\}$ s.t. $\sum_{i=1}^n d(T_i, R_{\sigma(i)})$ is
minimum



Proposition: [Aguse et al., ISMB 2019]

Given fixed clustering $\sigma : [n] \rightarrow [k]$, MCT
decomposes into k independent SCT instances

Combinatorial Characterization of Solutions

Single Consensus Trees (SCT):

[Govek et al., ACM-BCB 2018]

Given $\mathcal{T} = \{T_1, \dots, T_n\}$, find consensus tree R
s.t. $\sum_{i=1}^n d(T_i, R)$ is minimum

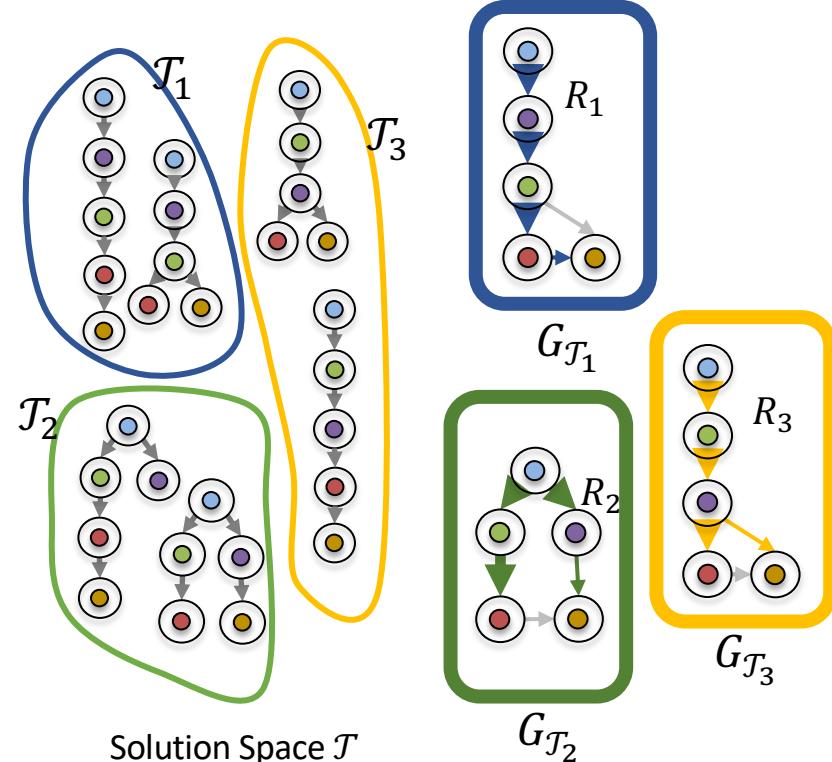
Theorem: [Govek et al., ACM-BCB 2018]

Max weight spanning arborescences
of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT

Multiple Consensus Trees (MCT):

[Aguse et al., ISMB 2019]

Given $\mathcal{T} = \{T_1, \dots, T_n\}$ and $k > 0$, find surjective clustering $\sigma : [n] \rightarrow [k]$ and consensus trees
 $\mathcal{R} = \{R_1, \dots, R_k\}$ s.t. $\sum_{i=1}^n d(T_i, R_{\sigma(i)})$ is
minimum where $R_{\sigma(i)}$ is max weight spanning
arborescence of $G_{\mathcal{T}_{\sigma(i)}}$



Proposition: [Aguse et al., ISMB 2019]

Given fixed clustering $\sigma : [n] \rightarrow [k]$, MCT
decomposes into k independent SCT instances

Combinatorial Characterization of Solutions

Single Consensus Trees (SCT):

[Govek et al., ACM-BCB 2018]

Given $\mathcal{T} = \{T_1, \dots, T_n\}$, find consensus tree R
s.t. $\sum_{i=1}^n d(T_i, R)$ is minimum

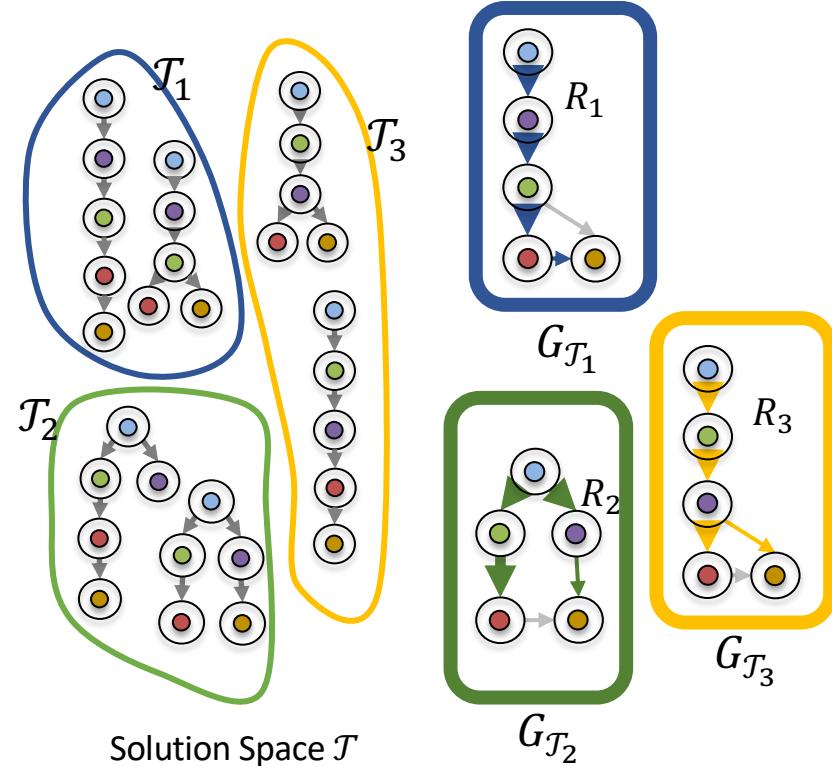
Theorem: [Govek et al., ACM-BCB 2018]

Max weight spanning arborescences
of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT

Multiple Consensus Trees (MCT):

[Aguse et al., ISMB 2019]

Given $\mathcal{T} = \{T_1, \dots, T_n\}$ and $k > 0$, find surjective clustering $\sigma : [n] \rightarrow [k]$ and consensus trees
 $\mathcal{R} = \{R_1, \dots, R_k\}$ s.t. $\sum_{i=1}^n d(T_i, R_{\sigma(i)})$ is
minimum where $R_{\sigma(i)}$ is max weight spanning
arborescence of $G_{\mathcal{T}_{\sigma(i)}}$



Proposition: [Aguse et al., ISMB 2019]

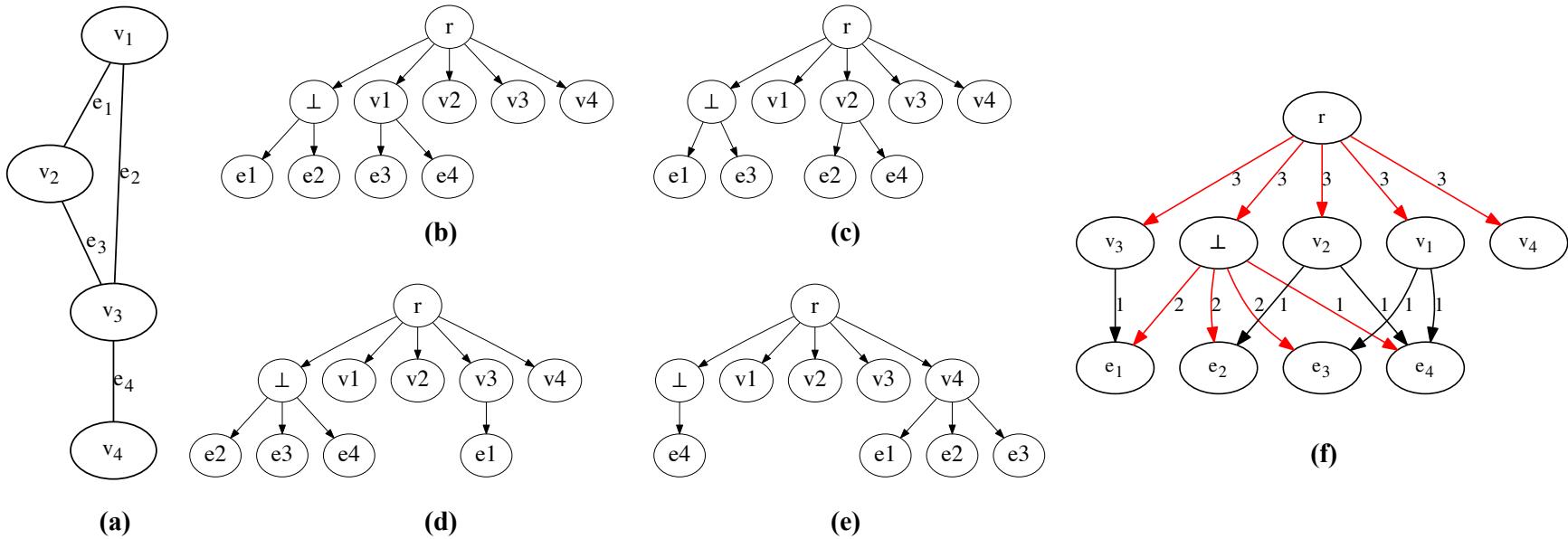
Given fixed clustering $\sigma : [n] \rightarrow [k]$, MCT
decomposes into k independent SCT instances

Question: How to find σ^* ?

Complexity

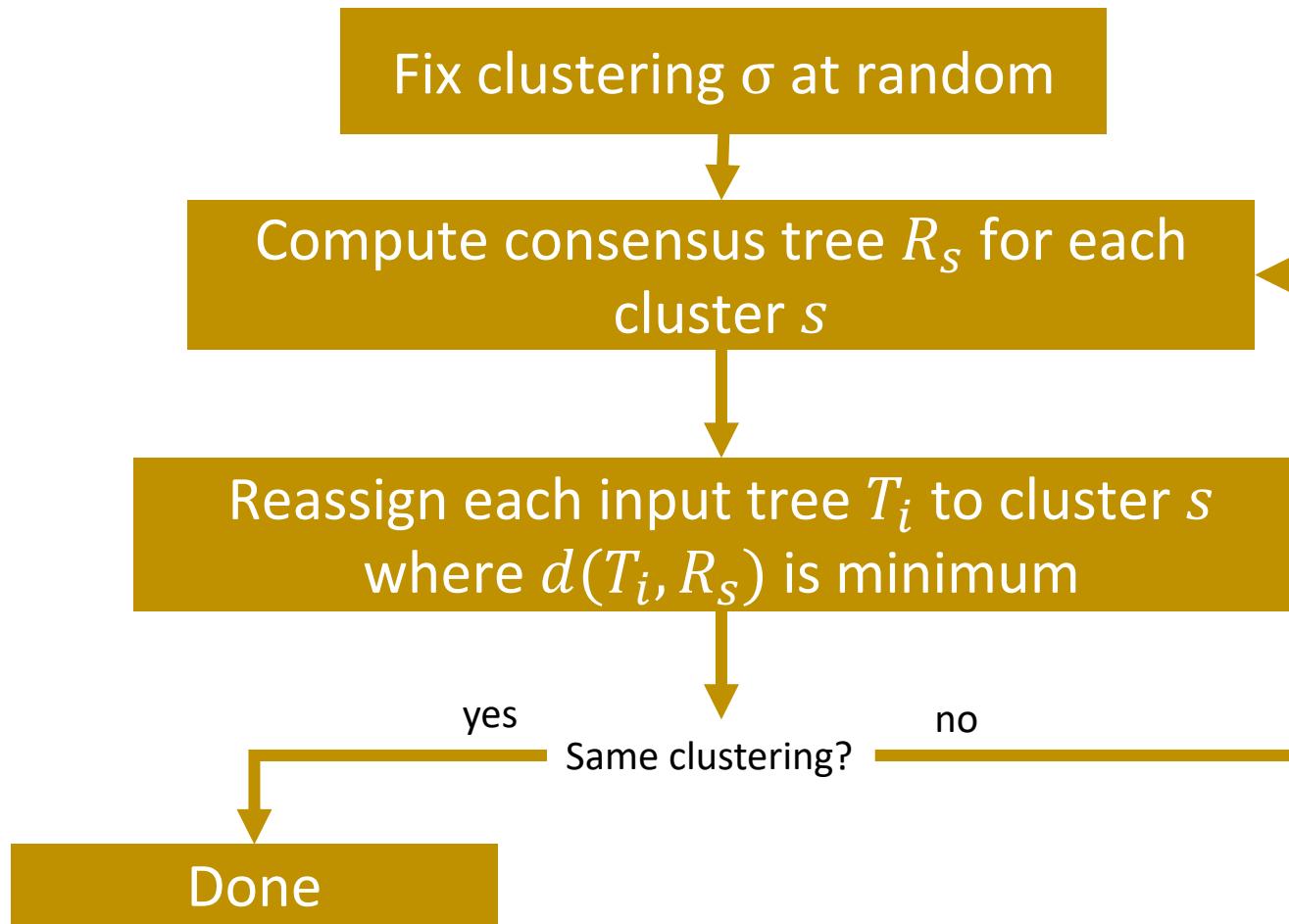
Multiple Consensus Trees (MCT):

Given $\mathcal{T} = \{T_1, \dots, T_n\}$ and $k > 0$, find surjective clustering $\sigma : [n] \rightarrow [k]$ s.t. $\sum_{i=1}^n d(T_i, R_{\sigma(i)})$ is minimum where $R_{\sigma(i)}$ is max weight spanning arborescence of $G_{\mathcal{T}_{\sigma(i)}}$



Theorem: MCT is NP-hard for general k (by reduction from CLIQUE).

Alternating optimization heuristic



Heuristic finds optimal solutions efficiently

Small, medium, and large simulated instances



	#clusters k	MILP (1 h)	CA (100 r.)
small (16)	2	16	16
	3	16	16
	4	16	16
	5	16	16
	2	15	15
medium (15)	3	13	13
	4	12	12
	5	10	10
	2	3	3
	3	0	0
large (14)	4	0	0
	5	0	0



Number of instances solved by MILP
to provable optimality

Heuristic finds optimal solutions efficiently

Small, medium, and large simulated instances



	#clusters k	MILP (1 h)	CA (100 r.)
small (16)	2	16	16
	3	16	16
	4	16	16
	5	16	16
	2	15	15
medium (15)	3	13	13
	4	12	12
	5	10	10
	2	3	3
	3	0	0
large (14)	4	0	0
	5	0	0



Number of instances where heuristic returned MILP's optimal solution

Heuristic finds optimal solutions efficiently

Small, medium, and large simulated instances



	#clusters k	MILP (1 h)	CA (100 r.)
small (16)	2	16	16
	3	16	16
	4	16	16
	5	16	16
medium (15)	2	15	15
	3	13	13
	4	12	12
	5	10	10
large (14)	2	3	3
	3	0	0
	4	0	0
	5	0	0



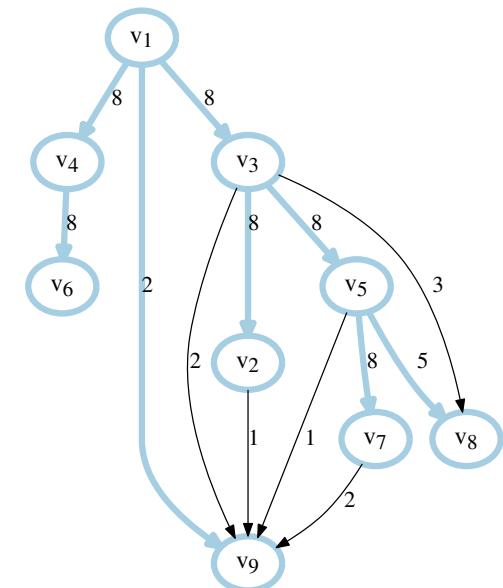
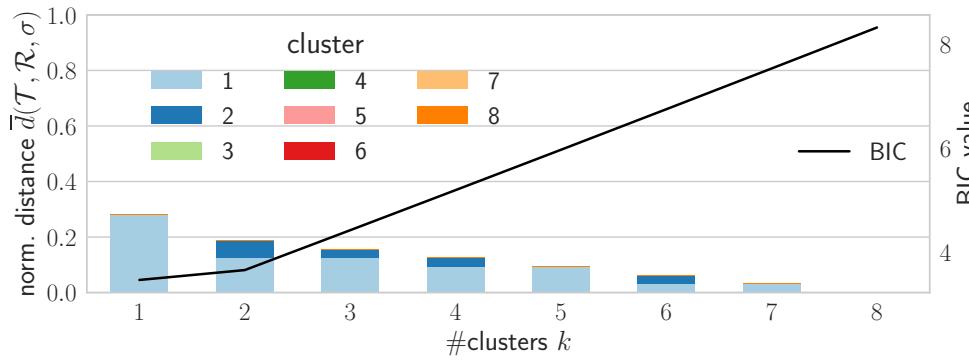
Number of instances where heuristic returned MILP's optimal solution

Question: How to determine k ?

Bayesian Information Criterion determines the number of clusters for each solution space

Jamal-Hanjani et al. (2017). NEJM.

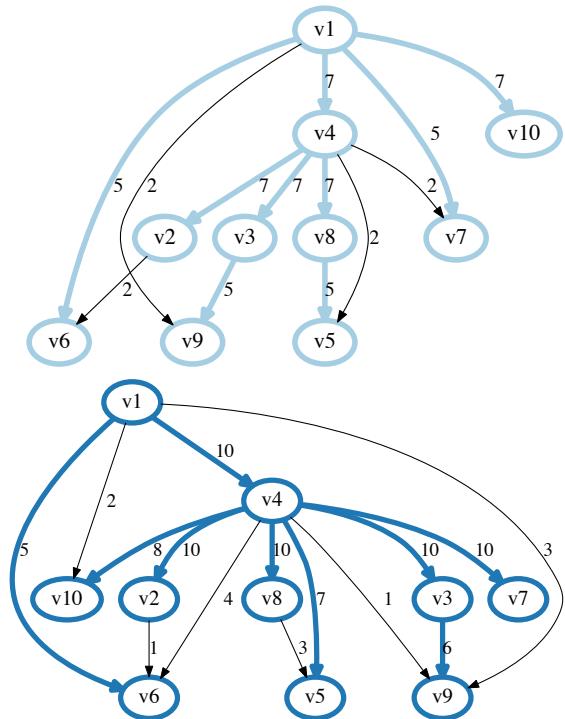
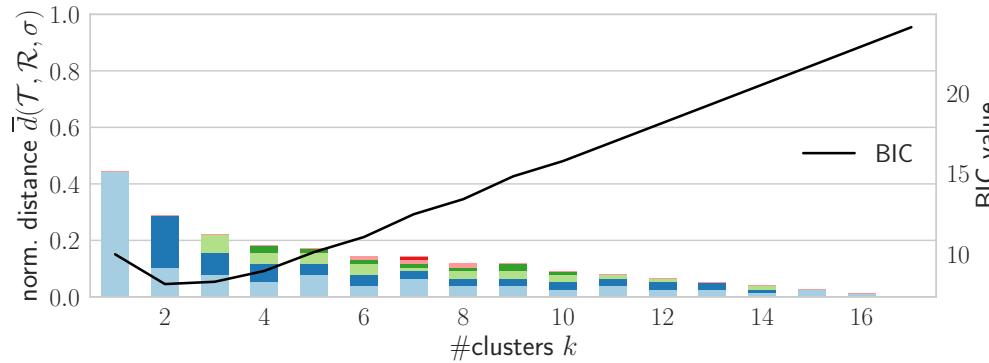
Jamal-Hanjani et al. inferred 8 trees for patient CRUK0013



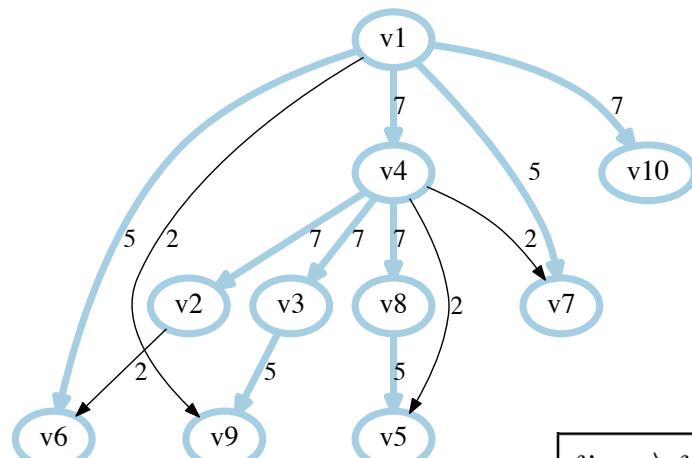
Bayesian Information Criterion determines the number of clusters for each solution space

Jamal-Hanjani et al. (2017). NEJM.

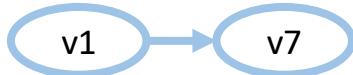
Jamal-Hanjani et al. inferred 17 trees for patient CRUK0037



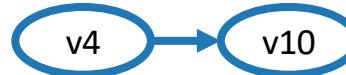
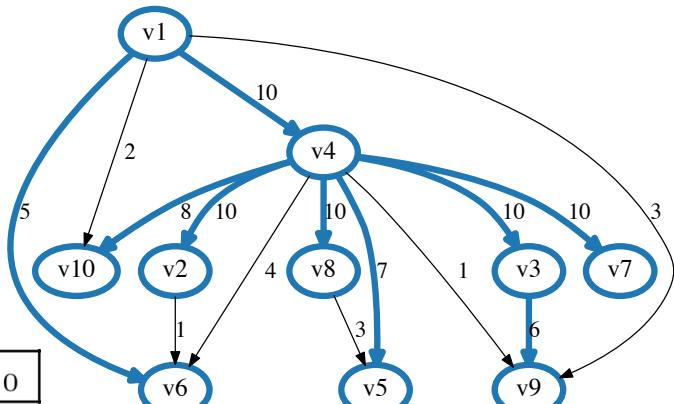
Multiple Consensus Trees capture patterns of mutual exclusivity and co-occurrence



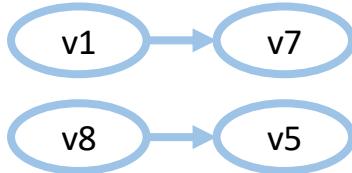
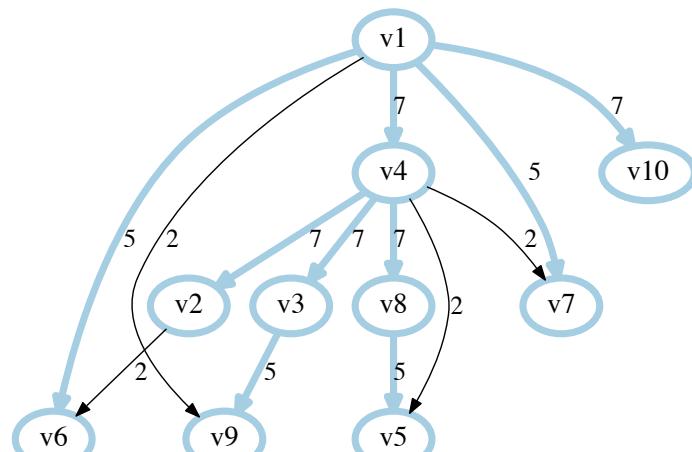
	$v_1 \rightarrow v_{10}$	$v_4 \rightarrow v_{10}$
$v_1 \rightarrow v_7$	2	0
$v_4 \rightarrow v_7$	2	5



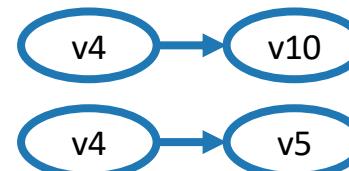
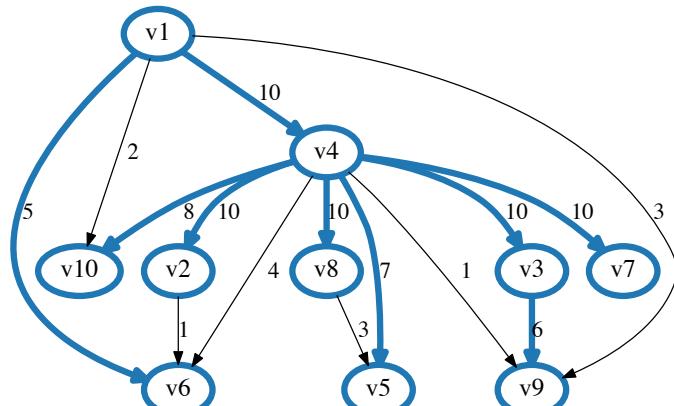
These edges are
mutually exclusive



Multiple Consensus Trees capture patterns of mutual exclusivity and co-occurrence

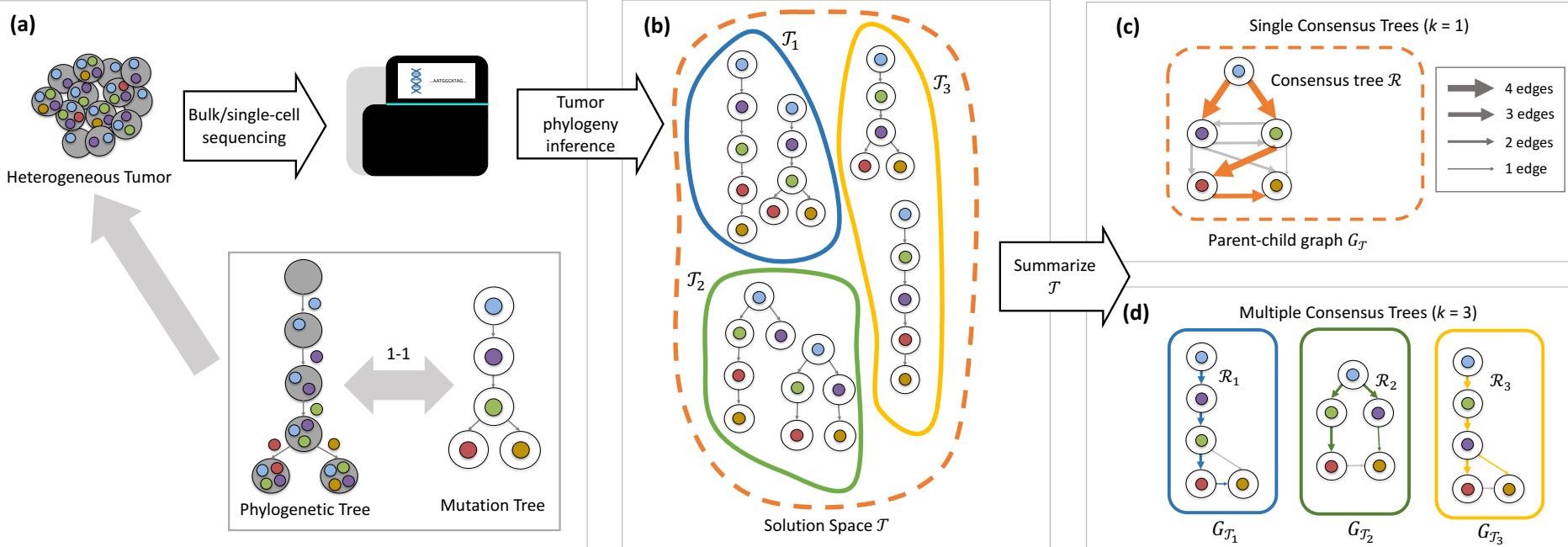


These edges tend to co-occur in the trees in the solution space



Multiple Consensus Trees (MCT): [ISMB 2019]

Given trees $\mathcal{T} = \{T_1, \dots, T_n\}$, find surjective clustering $\sigma : [n] \rightarrow [k]$ and consensus trees $\mathcal{R} = \{R_1, \dots, R_k\}$ such that $\sum_{i=1}^n d(T_i, R_{\sigma(i)})$ is minimum



- Characterize combinatorial structure of optimal solutions
- Show that MCT is NP-hard for general k
- Introduce a heuristic that returns optimal solution in most cases
- Model selection for k

Outline

1. Background and theory: [RECOMB-CG 2018]

- Perfect Phylogeny Mixture (PPM) problem
- #PPM: exact counting and uniform sampling

2. Simulation results: [RECOMB-CG 2018]

- What contributes to non-uniqueness?
- How to reduce non-uniqueness?

3. Summarizing solution space: [ISMB/ECCB 2019]

- Multiple consensus tree problem

4. Phylogenetic integration: [WABI 2021 / AMB 2022]

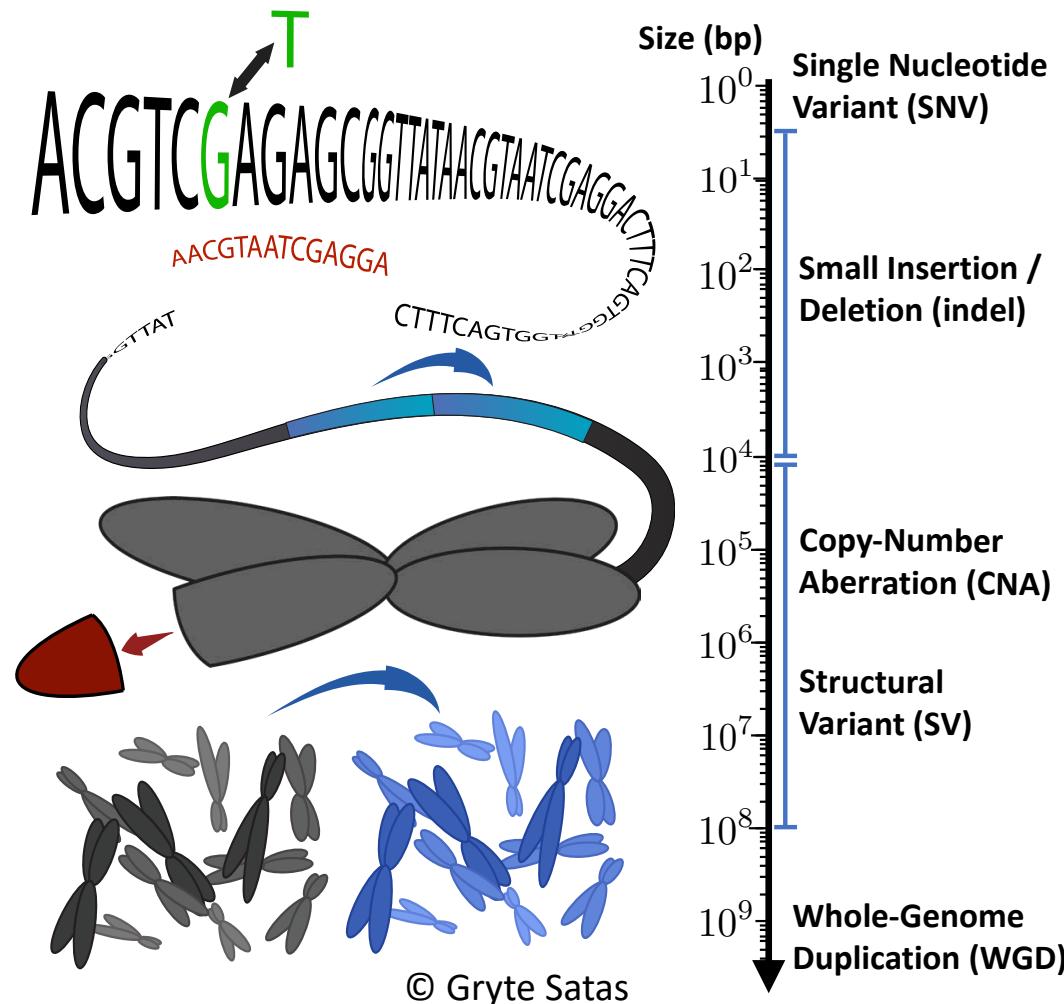
5. Applications

- Mutational signature dynamics [PSB 2020]
- Designing follow-up single-cell sequencing experiments [PLOS CB 2020]
- Visualizing spatial composition of a tumor [ISMB 2020]
- Identifying repeated evolutionary trajectories [ECCB 2020]

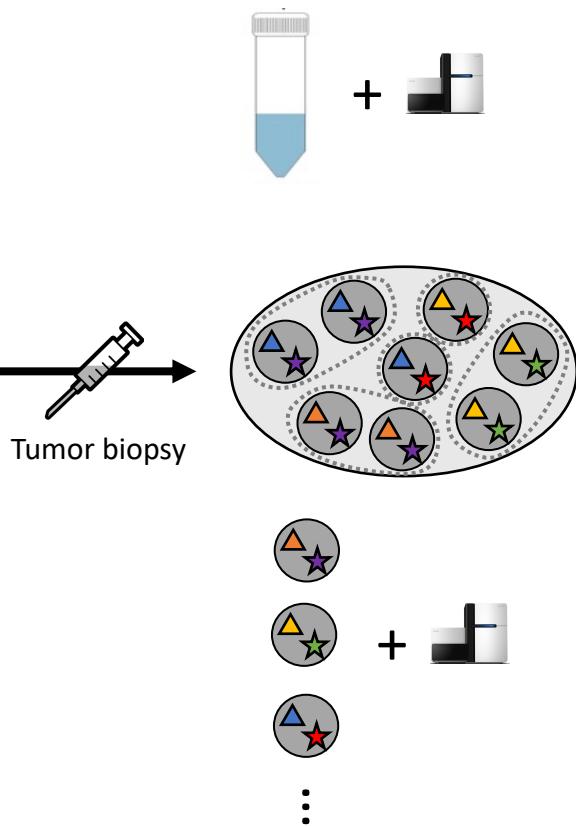


Palash Sashittal

Somatic Mutations Occur at Different Genomic Scales



Challenges in cancer tumor phylogeny estimation



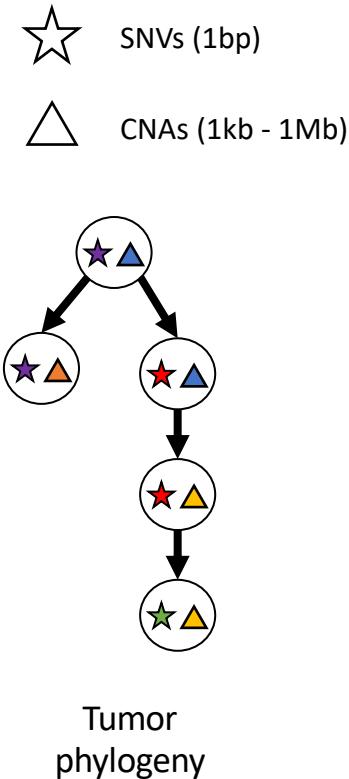
Bulk sequencing technologies

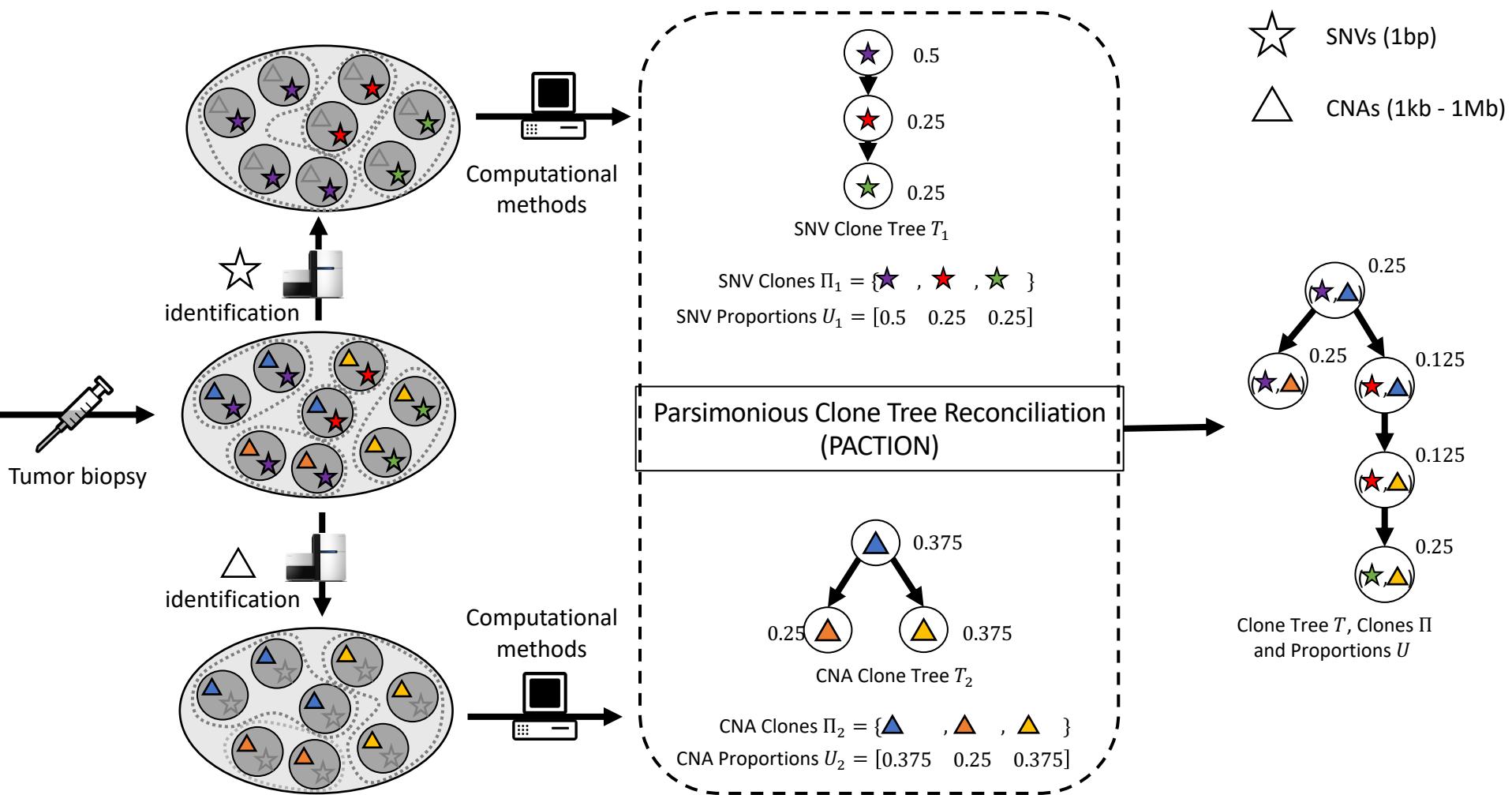
- Requires deconvolution of observed read counts
- Majority of methods only focus on either SNVs or CNAs, but not both
- Methods that account for both SNVs and CNAs do not scale to size of current cancer datasets

HARD

Single cell sequencing technologies

- Current technologies allow reliable detection of either SNVs or CNAs, but not both
- Targeted MDA, Tapestry suitable for identification of SNVs
- DOP-PCR, DLP, ACT, 10X Genomics CNV suitable for identification of CNAs



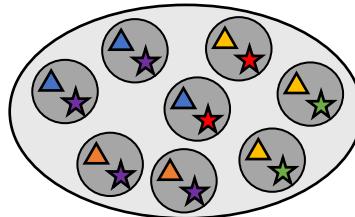


Parsimonious Clone Tree Integration

Clones $\Pi_1 = \{\triangle, \square, \triangledown\}$

Proportions $U_1 = [0.375 \quad 0.25 \quad 0.375]$

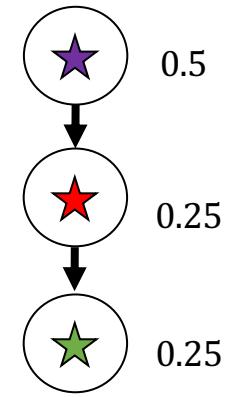
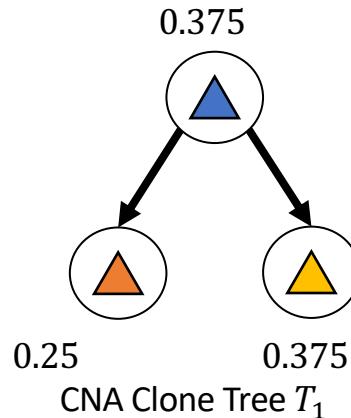
- 1.1. Input clones and proportions
- 2.1. Input clone trees



Clones $\Pi_2 = \{\star, \text{red star}, \text{green star}\}$

Proportions $U_2 = [0.5 \quad 0.25 \quad 0.25]$

- 1.2. Input clones and proportions
- 2.2. Input clone trees

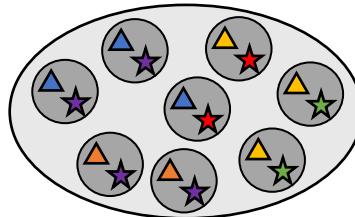


Parsimonious Clone Tree Integration

Clones $\Pi_1 = \{\Delta, \square, \triangleright\}$

Proportions $U_1 = [0.375 \quad 0.25 \quad 0.375]$

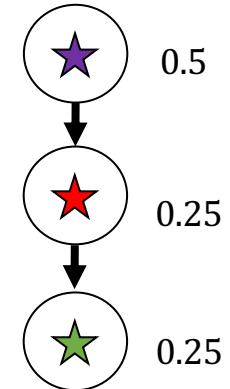
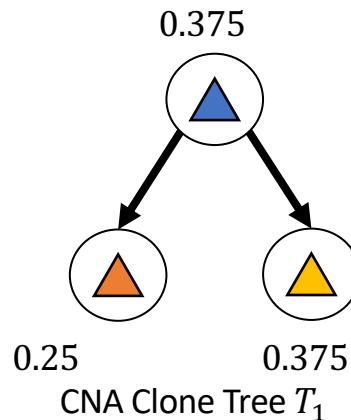
- 1.1. Input clones and proportions
- 2.1. Input clone trees



Clones $\Pi_2 = \{\star, \star, \star\}$

Proportions $U_2 = [0.5 \quad 0.25 \quad 0.25]$

- 1.2. Input clones and proportions
- 2.2. Input clone trees

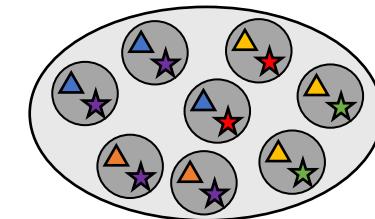
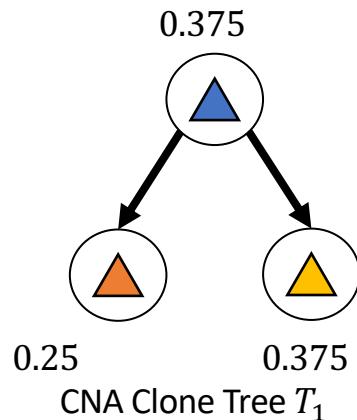


Parsimonious Clone Tree Integration

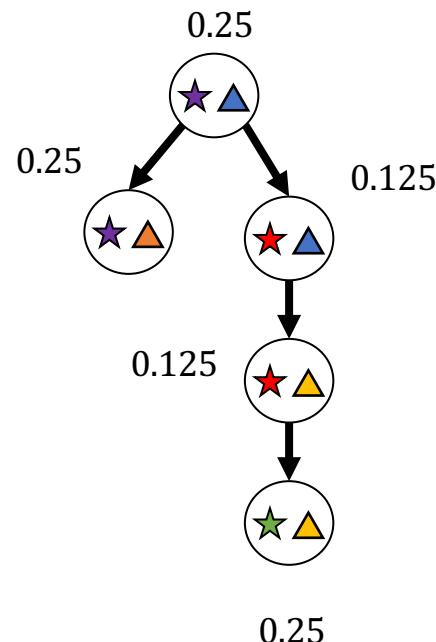
Clones $\Pi_1 = \{\Delta, \square, \triangleright\}$

Proportions $U_1 = [0.375 \quad 0.25 \quad 0.375]$

- 1.1. Input clones and proportions
- 2.1. Input clone trees



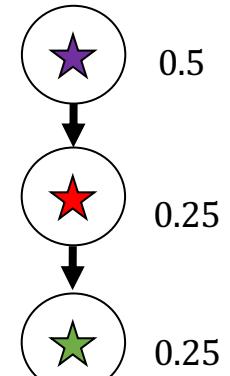
3. Output clone tree and proportions



Clones $\Pi_2 = \{\star, \star, \star\}$

Proportions $U_2 = [0.5 \quad 0.25 \quad 0.25]$

- 1.2. Input clones and proportions
- 2.2. Input clone trees

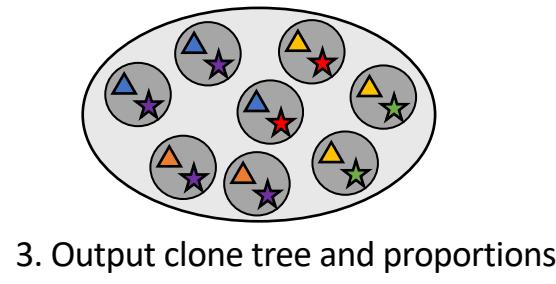
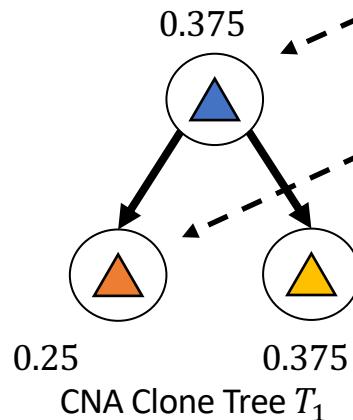


Parsimonious Clone Tree Integration

Clones $\Pi_1 = \{\Delta, \Delta, \Delta\}$

Proportions $U_1 = [0.375 \quad 0.25 \quad 0.375]$

- 1.1. Input clones and proportions
- 2.1. Input clone trees

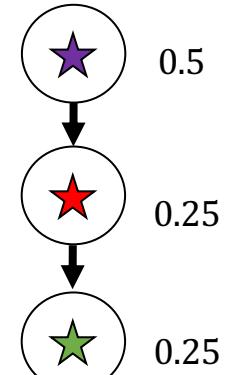


3. Output clone tree and proportions

Clones $\Pi_2 = \{\star, \star, \star\}$

Proportions $U_2 = [0.5 \quad 0.25 \quad 0.25]$

- 1.2. Input clones and proportions
- 2.2. Input clone trees



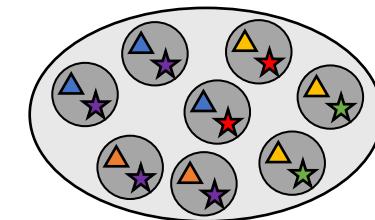
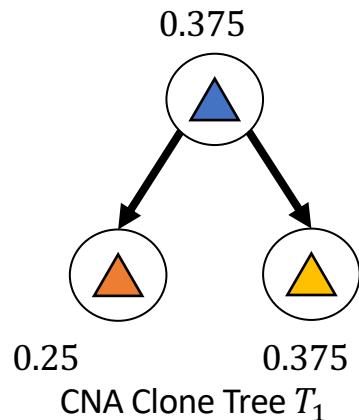
SNV Clone Tree T_2

Parsimonious Clone Tree Integration

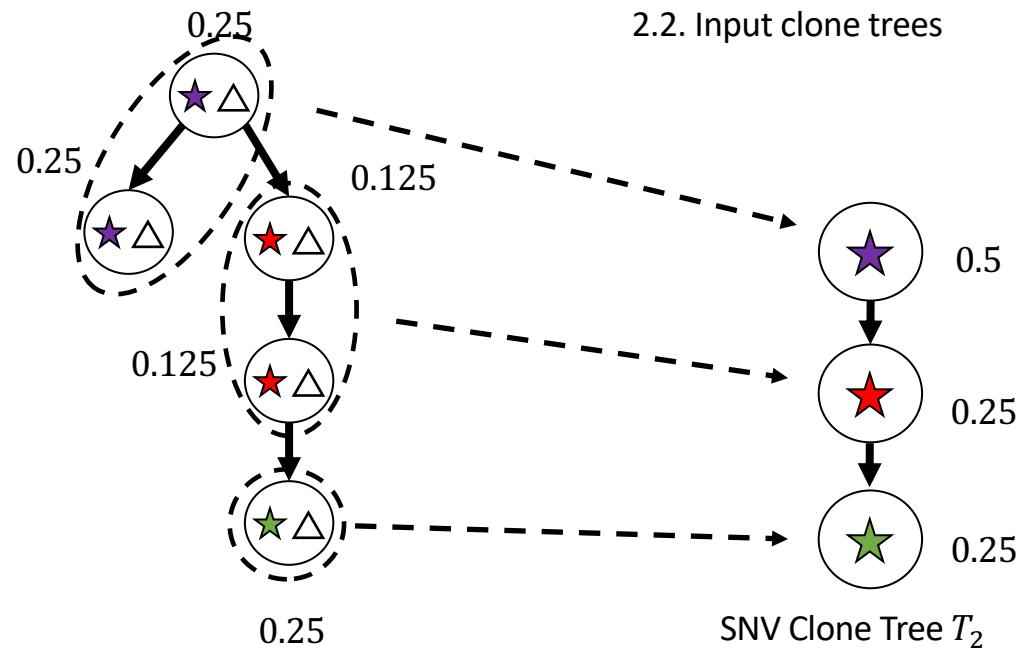
Clones $\Pi_1 = \{\Delta, \Delta, \Delta\}$

Proportions $U_1 = [0.375 \quad 0.25 \quad 0.375]$

- 1.1. Input clones and proportions
- 2.1. Input clone trees



3. Output clone tree and proportions



Clones $\Pi_2 = \{\star, \star, \star\}$

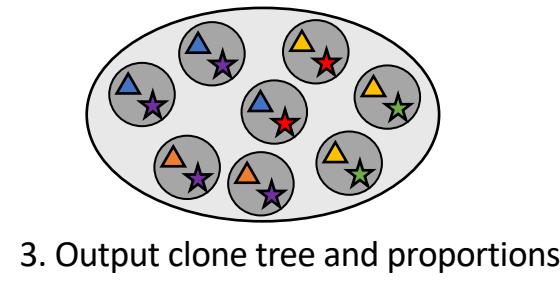
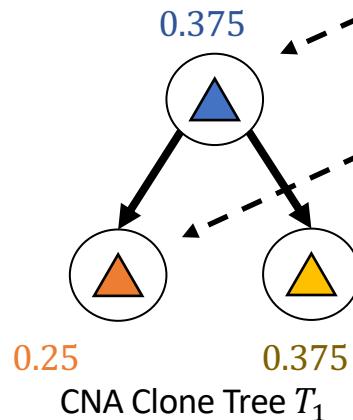
Proportions $U_2 = [0.5 \quad 0.25 \quad 0.25]$

Parsimonious Clone Tree Integration

Clones $\Pi_1 = \{\Delta, \square, \triangleright\}$

Proportions $U_1 = [0.375 \quad 0.25 \quad 0.375]$

- 1.1. Input clones and proportions
- 2.1. Input clone trees

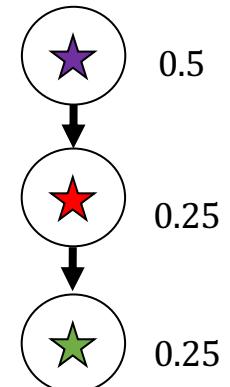


3. Output clone tree and proportions

Clones $\Pi_2 = \{\star, \star, \star\}$

Proportions $U_2 = [0.5 \quad 0.25 \quad 0.25]$

- 1.2. Input clones and proportions
- 2.2. Input clone trees



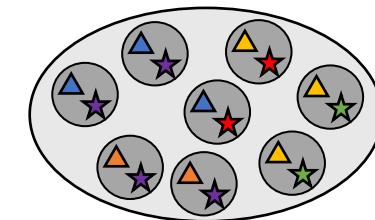
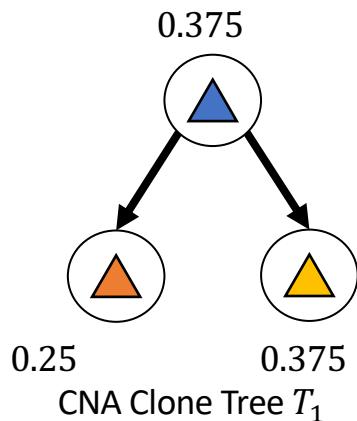
SNV Clone Tree T_2

Parsimonious Clone Tree Integration

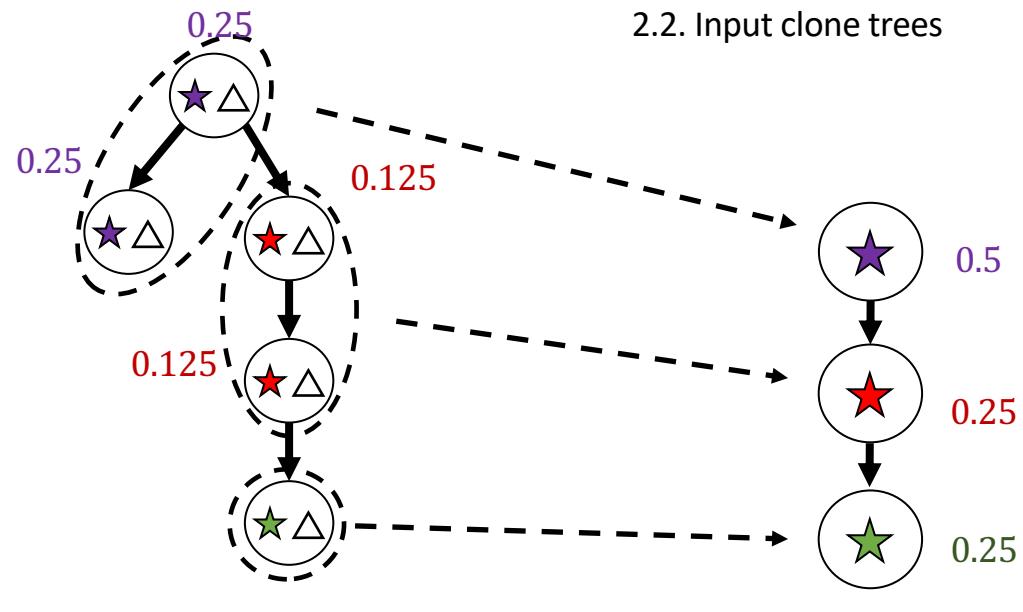
Clones $\Pi_1 = \{\Delta, \Delta, \Delta\}$

Proportions $U_1 = [0.375 \quad 0.25 \quad 0.375]$

- 1.1. Input clones and proportions
- 2.1. Input clone trees



3. Output clone tree and proportions

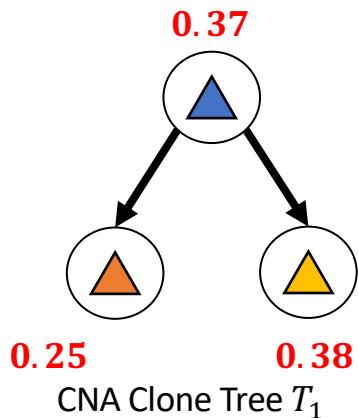


Parsimonious Clone Tree Integration

Clones $\Pi_1 = \{ \triangle, \square, \triangledown \}$

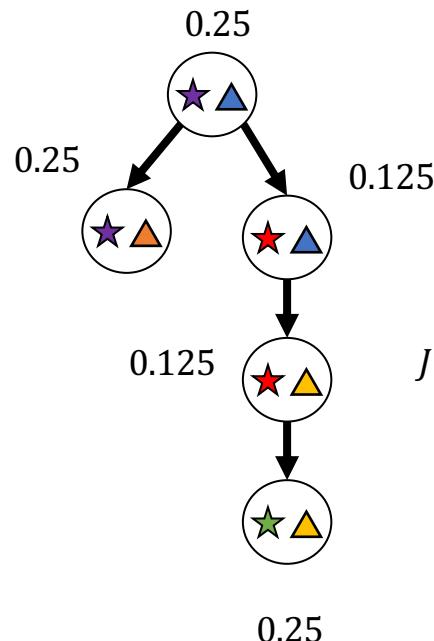
Proportions $U_1 = [0.375 \quad 0.25 \quad 0.375]$

- 1.1. Input clones and proportions
- 2.1. Input clone trees



* In practice we minimize the L_1 norm of difference between input and output clone proportions

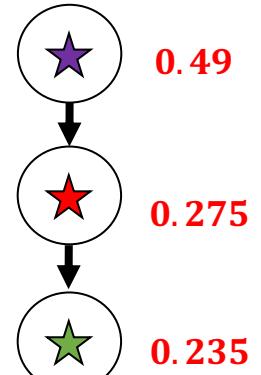
- 3. Output clone tree and proportions



Clones $\Pi_2 = \{ \star, \textcolor{red}{\star}, \textcolor{green}{\star} \}$

Proportions $U_2 = [0.5 \quad 0.25 \quad 0.25]$

- 1.2. Input clones and proportions
- 2.2. Input clone trees

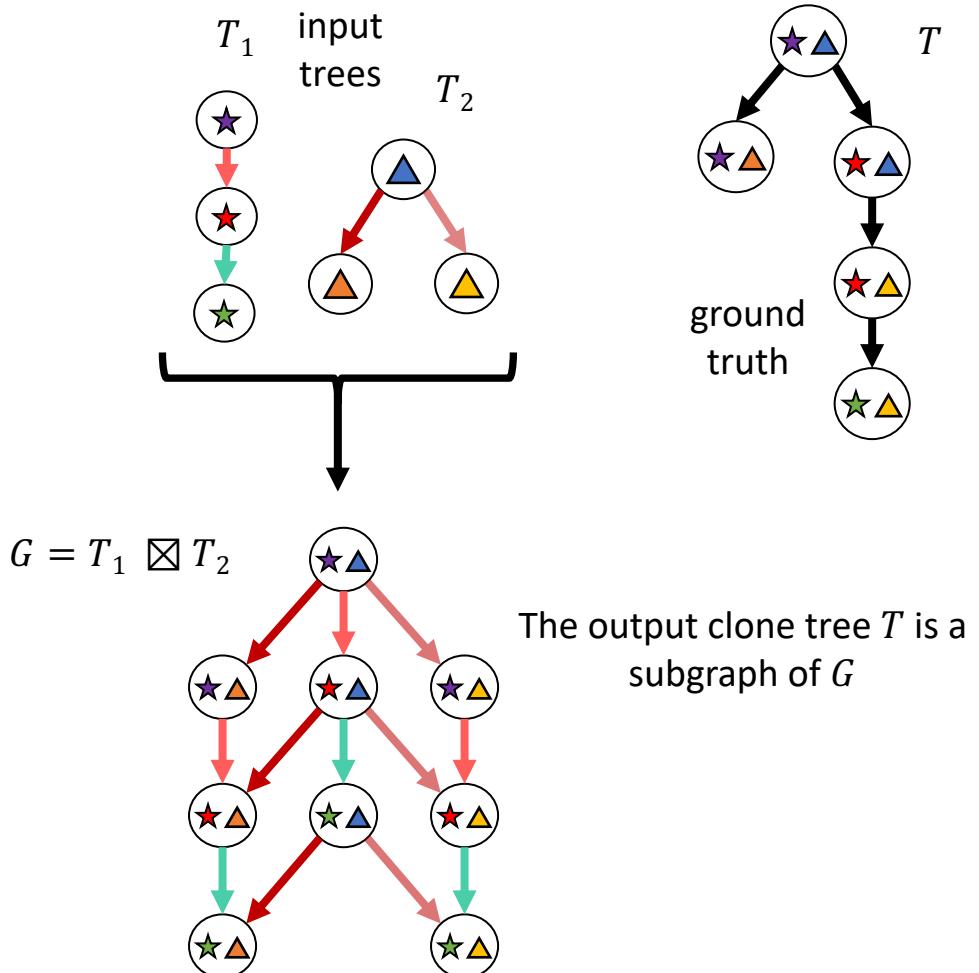


PACTION: Mixed-Integer Linear Programming

Vertex set

$$V(G) = V(T_1) \times V(T_2)$$

This gives the number of binary variables



PACTION: Mixed-Integer Linear Programming

Vertex set

$$V(G) = V(T_1) \times V(T_2)$$

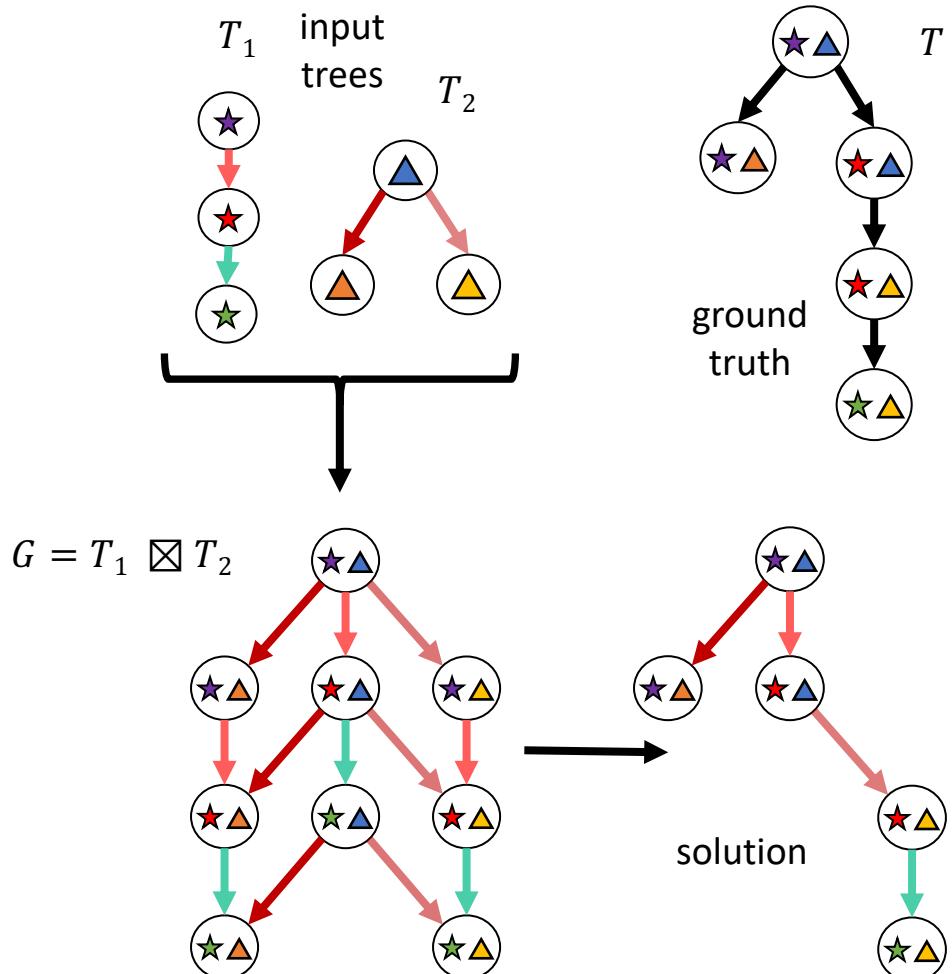
This gives the number of binary variables

We select a tree in G while minimizing the difference in the clone proportions

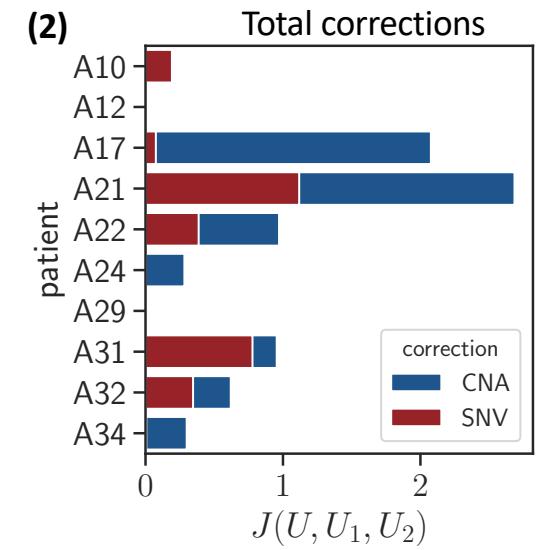
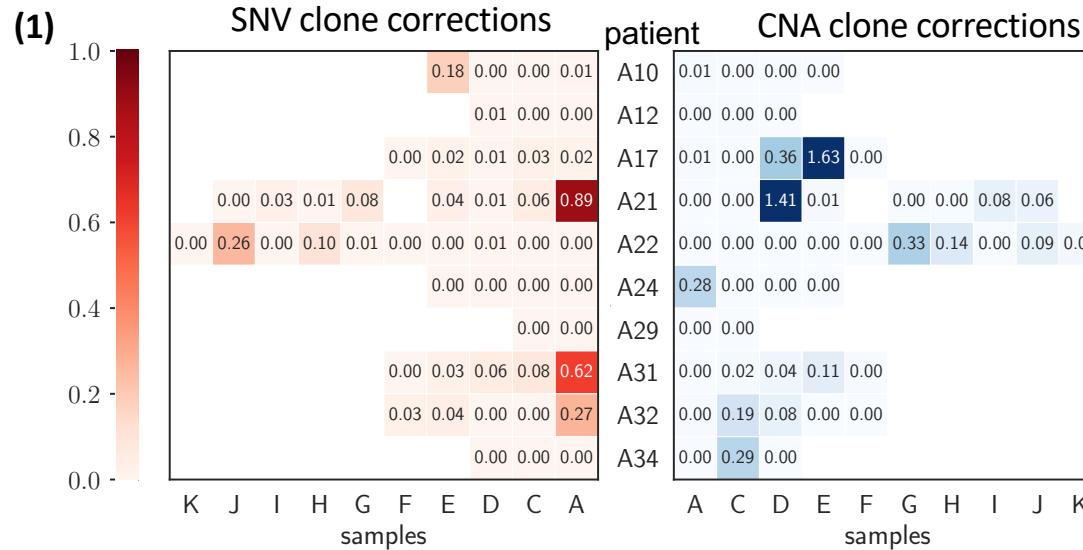
$$\min \underbrace{\sum_{p=1}^m \sum_{i=1}^{n_1} c_{p,i}^{(1)} + \sum_{p=1}^m \sum_{j=1}^{n_2} c_{p,j}^{(2)}}_{J(U, U_1, U_2)}$$

Difference in SNV clone proportions Difference in CNA clone proportions

such that the clone tree T is a refinement of T_1 and T_2



49 samples from 10 Metastatic prostate cancer patients [1]



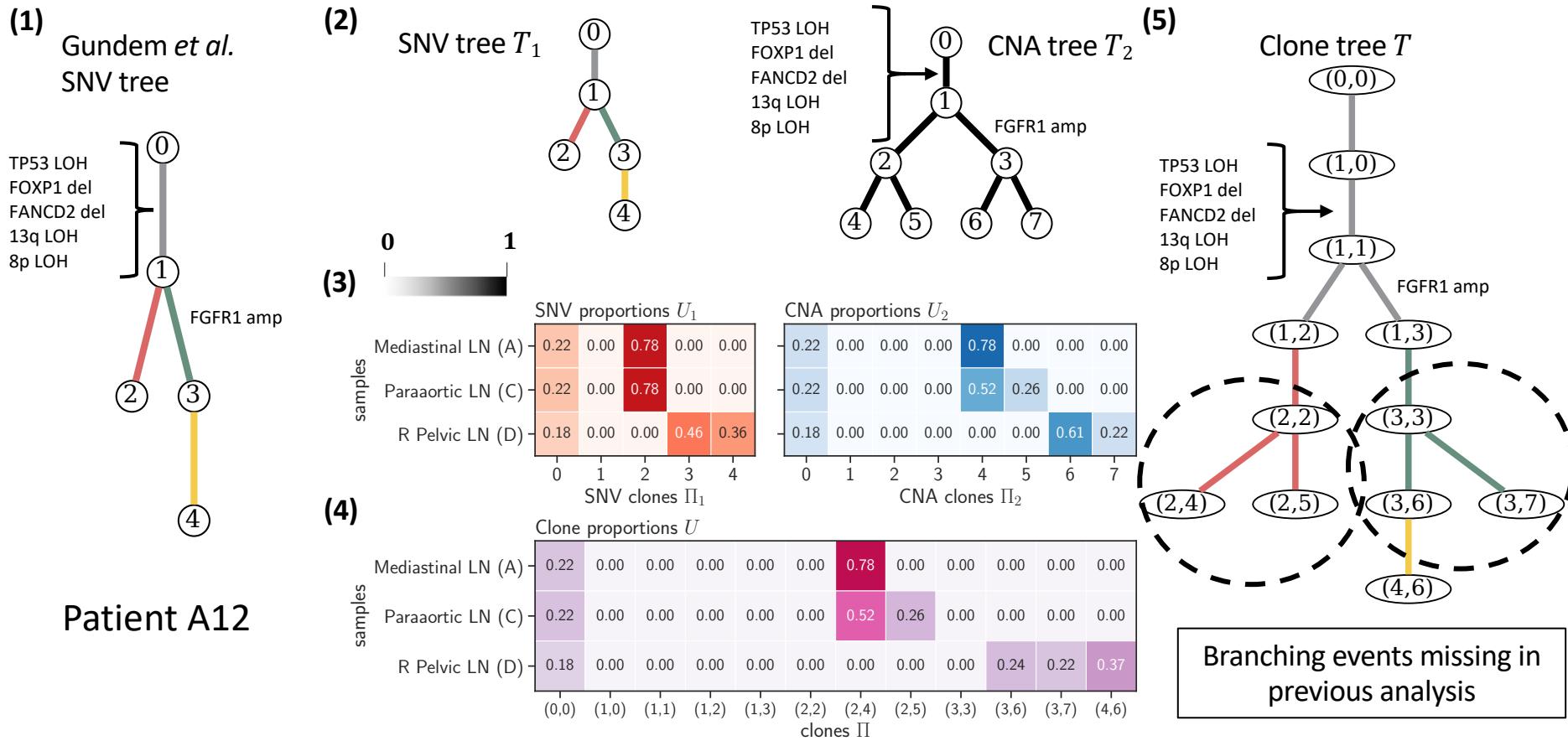
1. Corrections made to clone proportions for each patient at localized.
 - a. SNV clone trees and proportions from Gundem *et al.* [1]
 - b. CNA clones and proportions obtained from previous HATCHet analysis [2].
 - c. CNA clone trees selected via exhaustive enumeration.

2. Almost all patients required corrections for feasible reconciliation of SNV and CNA clone trees

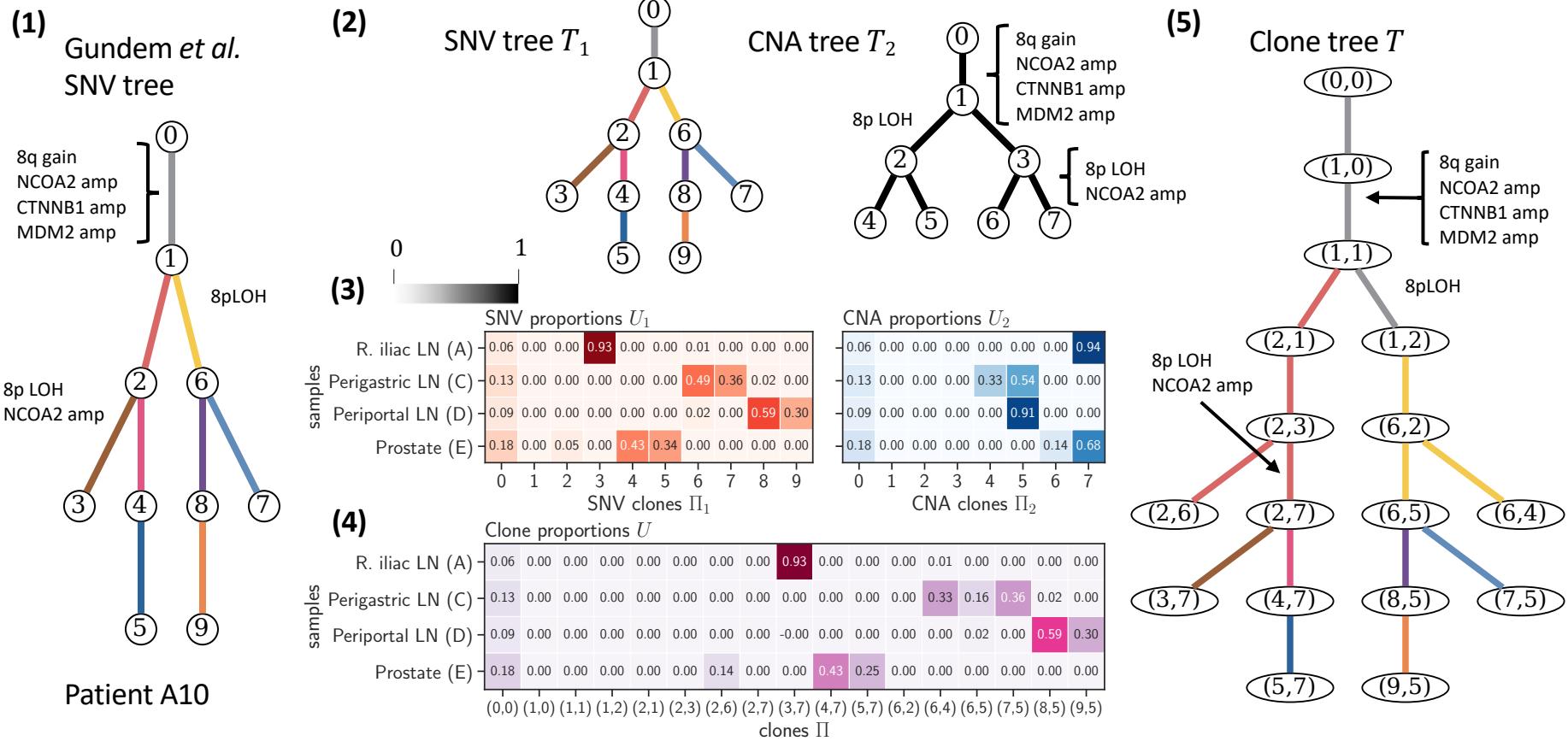
[1] Gundem *et al.* Nature, 520(7547):353–357, 2015.

[2] Simone Zaccaria and Benjamin J Raphael, Nature communications, 11(1):1–13, 2020.

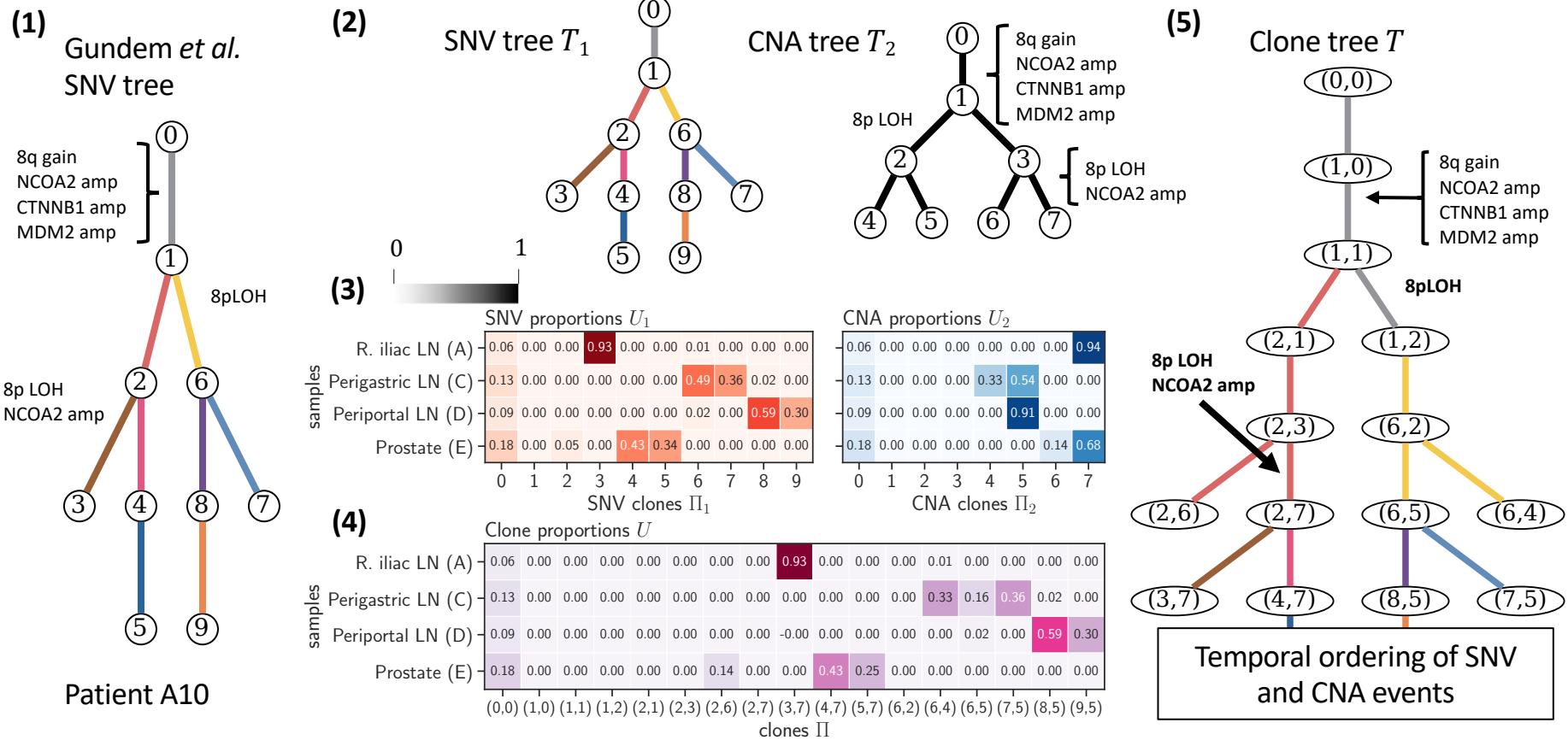
PACTION reveals previously missing branching events



PACTION reveals temporal ordering of SNVs and CNAs



PACTION reveals temporal ordering of SNVs and CNAs



Outline

1. Background and theory: [RECOMB-CG 2018]

- Perfect Phylogeny Mixture (PPM) problem
- #PPM: exact counting and uniform sampling

2. Simulation results: [RECOMB-CG 2018]

- What contributes to non-uniqueness?
- How to reduce non-uniqueness?
- How does non-uniqueness affect current methods?

3. Summarizing solution space: [ISMB/ECCB 2019]

- Multiple consensus tree problem

4. Phylogenetic integration: [WABI 2021 / AMB 2022]

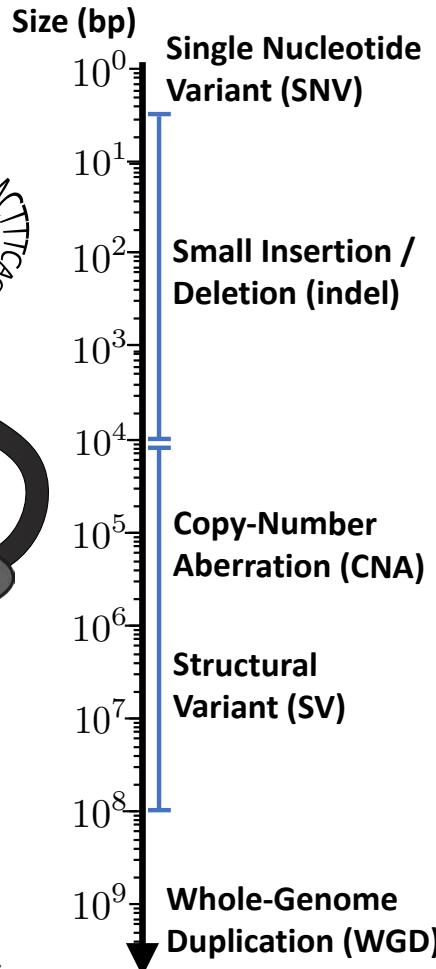
5. Applications

- Mutational signature dynamics [PSB 2020]
- Designing follow-up single-cell sequencing experiments [PLOS CB 2020]
- Visualizing spatial composition of a tumor [ISMB 2020]
- Identifying repeated evolutionary trajectories [ECCB 2020]



Sarah Christensen

Mutational Signatures



There are $4 \times 6 \times 4 = 96$ SNV categories:
AC>AA, ... GT>GG

Distinct mutational processes =>
distinct patterns of SNV categories

Mutational Signatures – NMF

P : Feature Matrix

$$\begin{matrix} \text{AC>AA} & \text{TC>GG} \\ \left(\begin{matrix} 0 & 0 & 1 \\ 2 & 1 & 1 \end{matrix} \right) & \approx \end{matrix}$$

$m \times n$

S : Signature Matrix

$$\begin{matrix} \text{AC>AA} & \text{TC>GG} \\ \left(\begin{matrix} \text{Sig. 1} & \text{Sig. 2} \\ 0 & .5 \\ 1 & .5 \end{matrix} \right) & \end{matrix}$$

$m \times r$

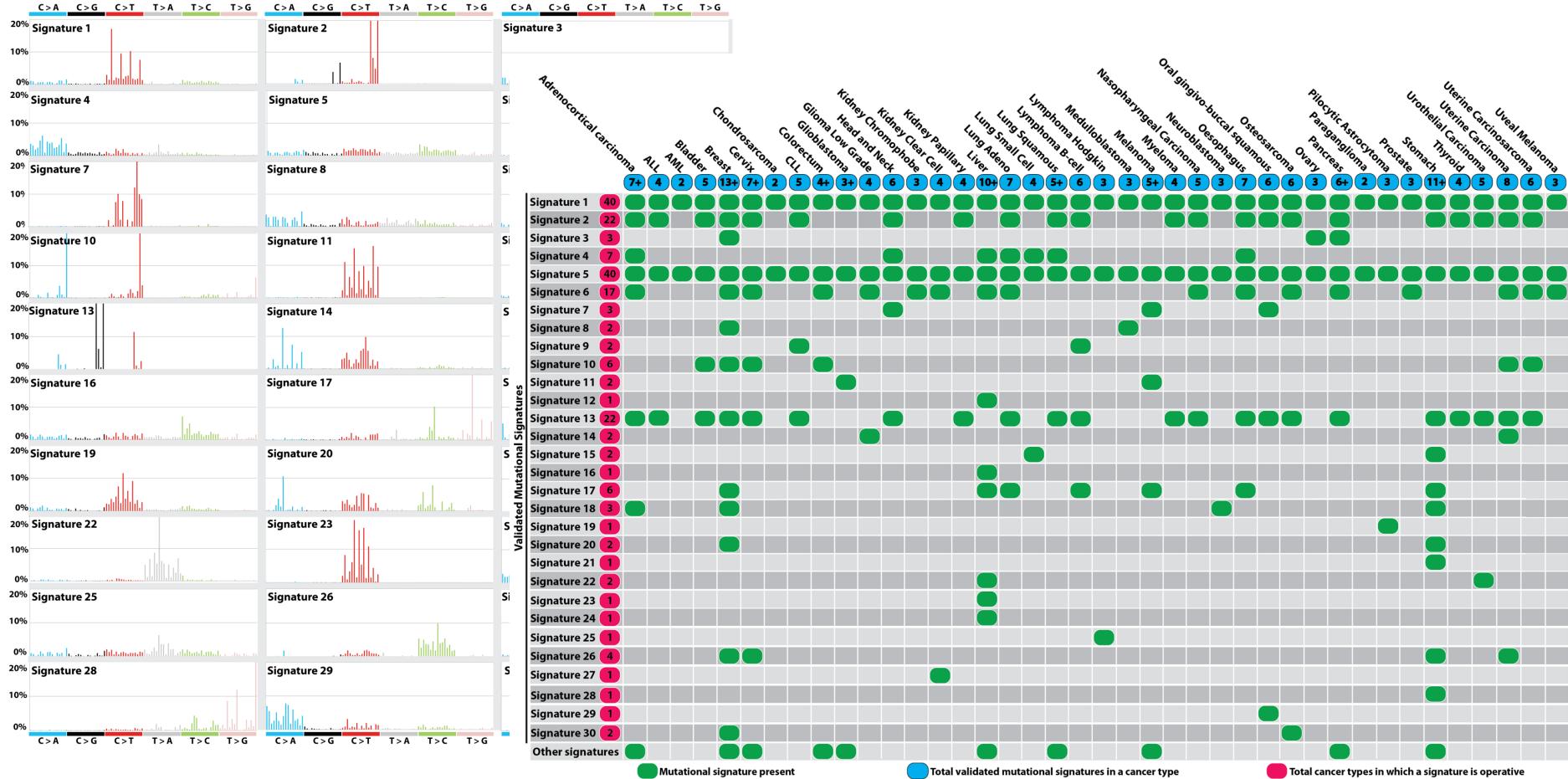
E : Exposure Matrix

$$\left(\begin{matrix} \text{Silhouette 1} & \text{Silhouette 2} & \text{Silhouette 3} \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{matrix} \right)$$

$r \times n$

Alexandrov et al. [Nature, 2013] performed nonnegative matrix factorization on a large patient cohort ($n = 10,000$) identifying $r = 30$ signatures and exposures

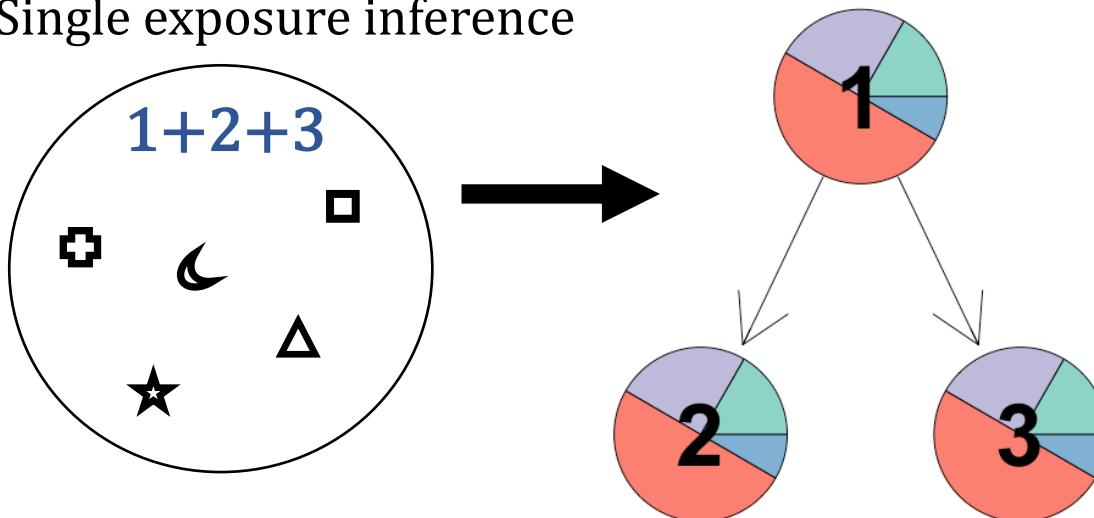
Mutational Signatures



Alexandrov et al. [Nature, 2013] performed nonnegative matrix factorization on a large patient cohort ($n = 10,000$) identifying $r = 30$ signatures and exposures

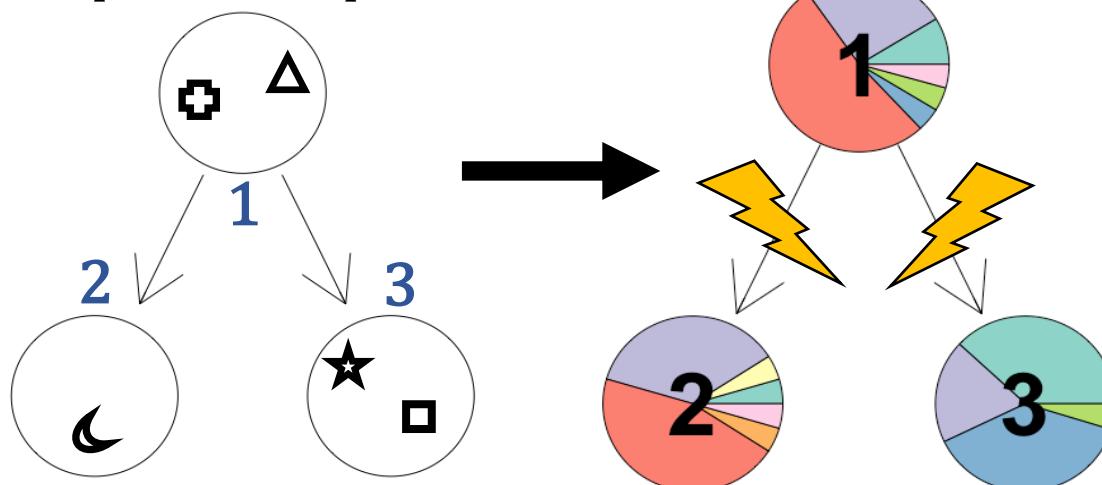
Clone-specific exposure inference problems

Single exposure inference



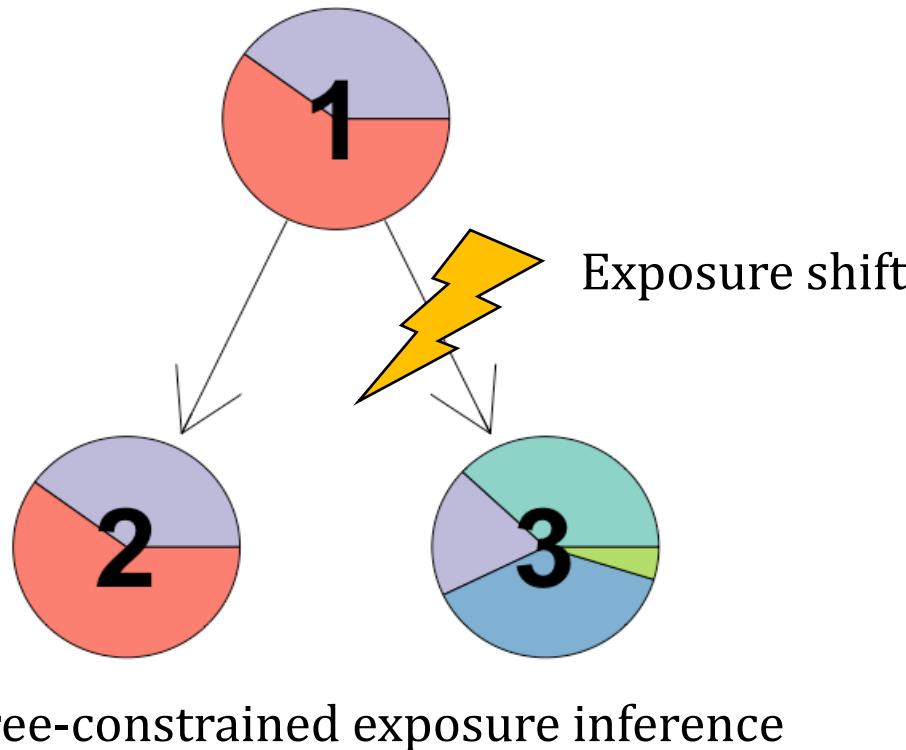
- Alexandrov et al., Cell reports 2013
- Rosenthal et al., Genome Biology 2016
- Huang et al., Bioinformatics, 2017
- ...

Independent exposure inference



- McPherson et al., Nature Genetics, 2016
- Jamal-Hanjani et al., NEJM 2017
- ...

Tree-constrained exposure (TE) inference



TE problem: [Christensen, Leiserson and El-Kebir, PSB 2020]:
Given phylogenetic tree T and feature matrix P ,
find a small number of exposure shifts along edges of T

PhySigs solves the TE problem to optimality

P : Feature Matrix

$$\begin{matrix} \text{AC>AA} & \text{TC>GG} \\ \left(\begin{array}{ccc} \text{△+} & \text{∅} & \text{☆□} \\ 0 & 0 & 1 \\ 2 & 1 & 1 \end{array} \right) \end{matrix} \approx \begin{matrix} \text{AC>AA} & \text{TC>GG} \\ \left(\begin{array}{cc} \text{Sig. 1} & \text{Sig. 2} \\ 0 & .5 \\ 1 & .5 \end{array} \right) \end{matrix}$$

$m \times n$

S : Signature Matrix

$$\begin{matrix} \text{AC>AA} & \text{TC>GG} \\ \left(\begin{array}{cc} \text{Sig. 1} & \text{Sig. 2} \\ 0 & .5 \\ 1 & .5 \end{array} \right) \end{matrix}$$

$m \times r$

D : Relative Exposure Matrix

$$\begin{matrix} \text{AC>AA} & \text{TC>GG} \\ \left(\begin{array}{cc} \text{△+} & \text{∅} \\ 1 & 1 \\ 0 & 0 \end{array} \right) \end{matrix}$$

$r \times n$

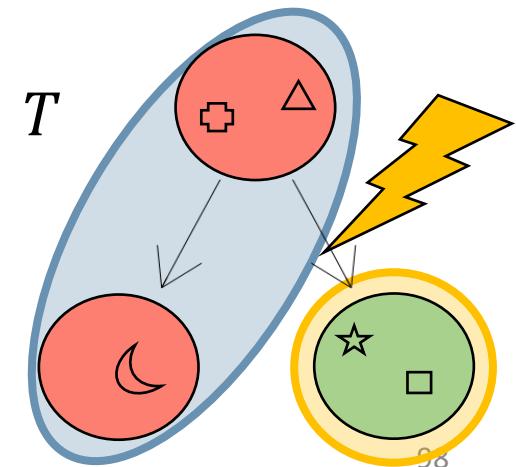
C : Count Matrix

$$\begin{matrix} \text{AC>AA} & \text{TC>GG} \\ \left(\begin{array}{ccc} \text{△+} & \text{∅} & \text{☆□} \\ 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{array} \right) \end{matrix}$$

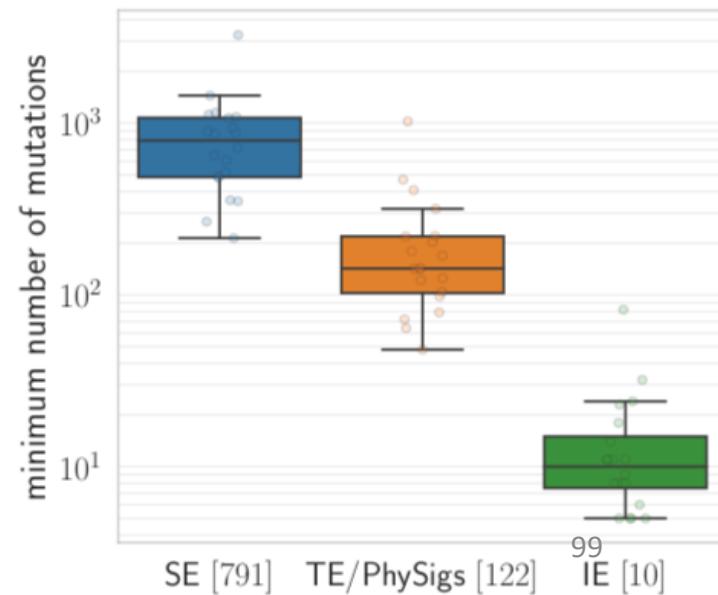
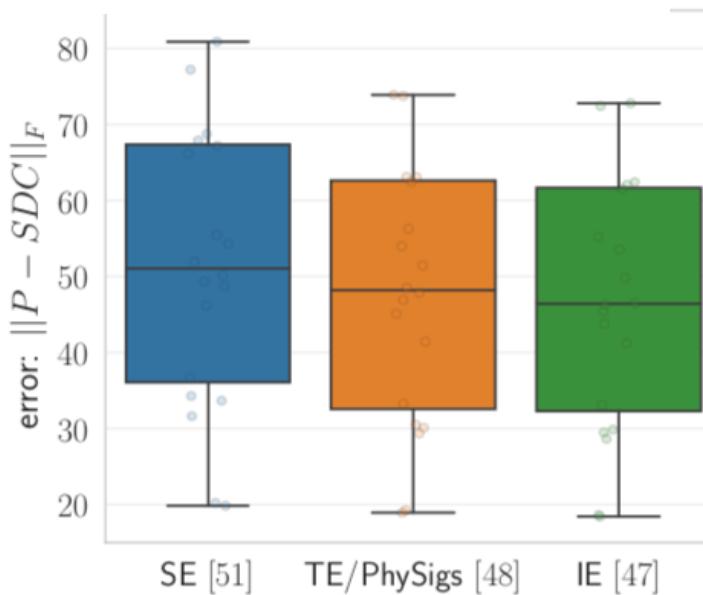
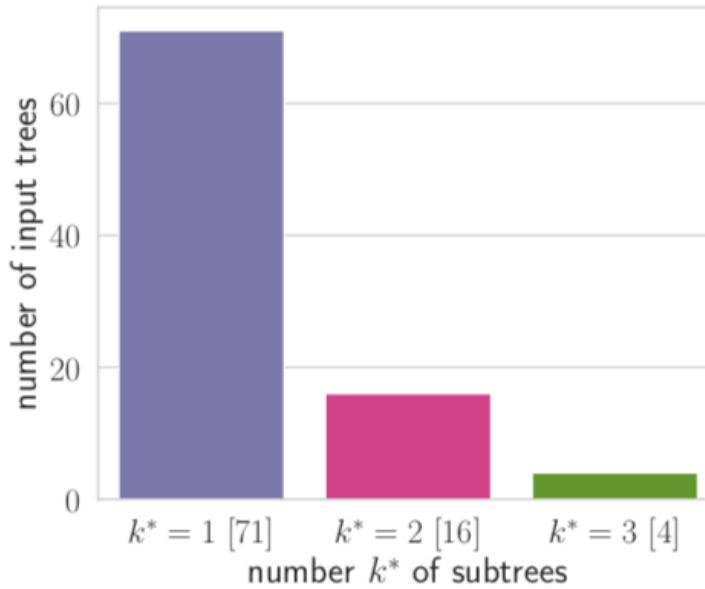
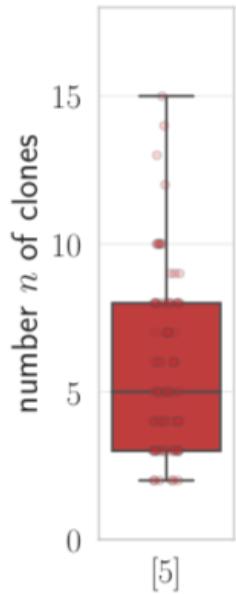
$n \times n$

TE problem:

Given feature matrix P , corresponding count matrix C , signature matrix S , phylogenetic tree T and integer $k \geq 1$, find relative exposure matrix D such that $\|P - SDC\|_F$ is minimum and D is composed of k sets of identical columns, each corresponding to a connected subtree of T .



PhySigs identifies accurate exposures without overfitting in a lung cancer cohort



PhySigs identifies an exposure shift supported by a subclonal driver

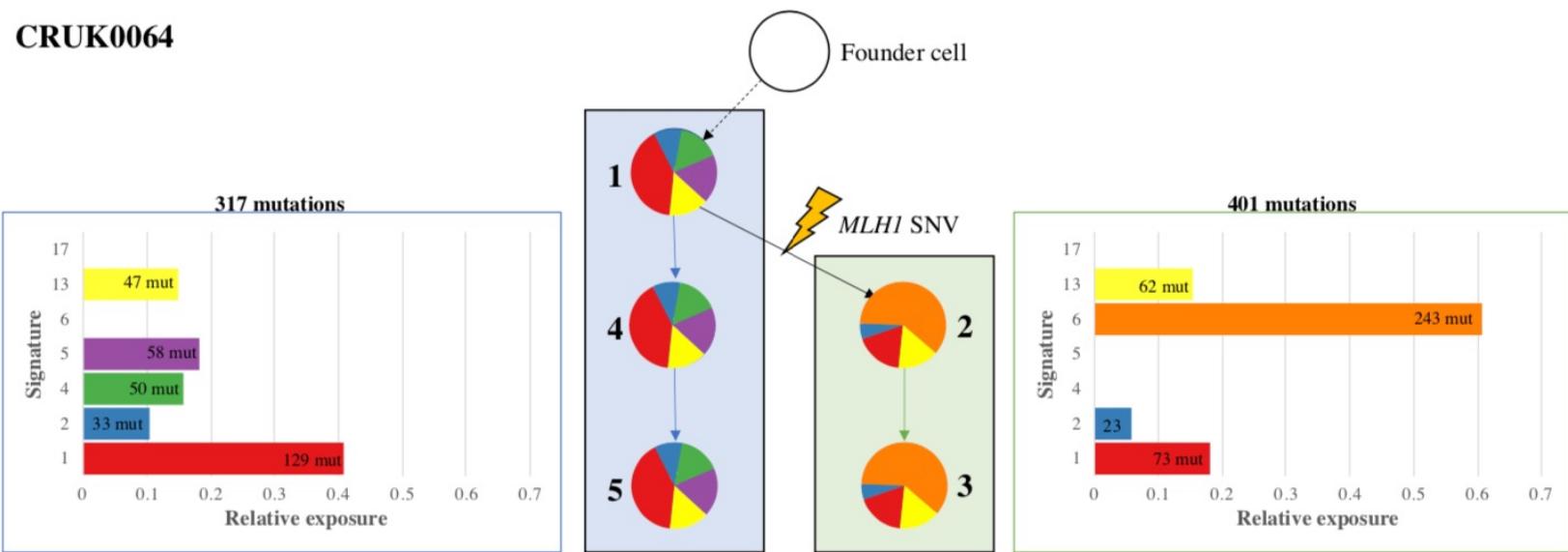
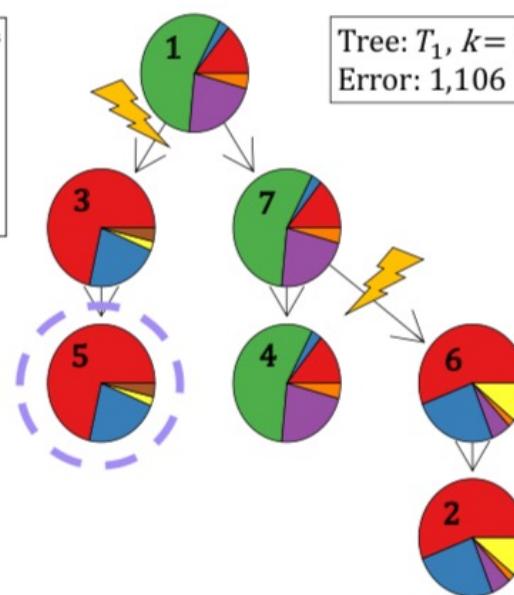


Fig. 5. PhySigs detects a large increase in DNA mismatch repair-associated Signature 6 (orange) along one branch (clusters 2 and 3; green) of the CRUK 0064 tree. In support of this finding, the branch includes a subclonal driver mutation to DNA mismatch repair gene *MLH1*.

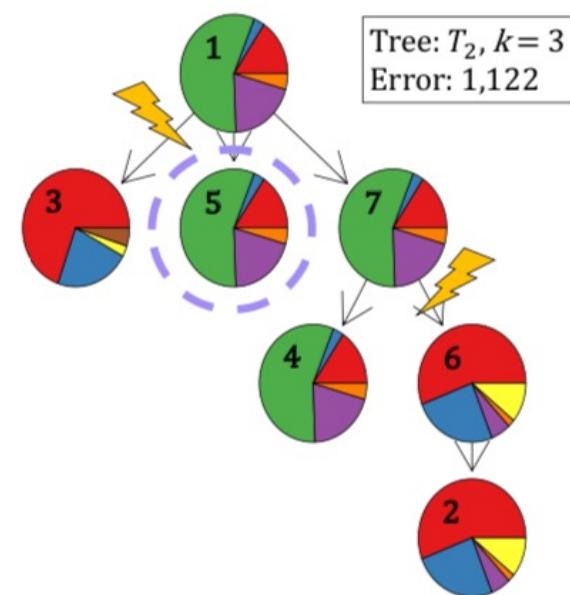
PhySigs enables prioritization of trees in solution space

CRUK0025		
k	Tree T_1	Tree T_2
	Error	
1	1,344	1,344
2	1,199	1,199
$k^* = 3$	1,106	1,122
4	1,096	1,105
5	1,095	1,095
6	1,095	1,095
7	1,094	1,094

(a)



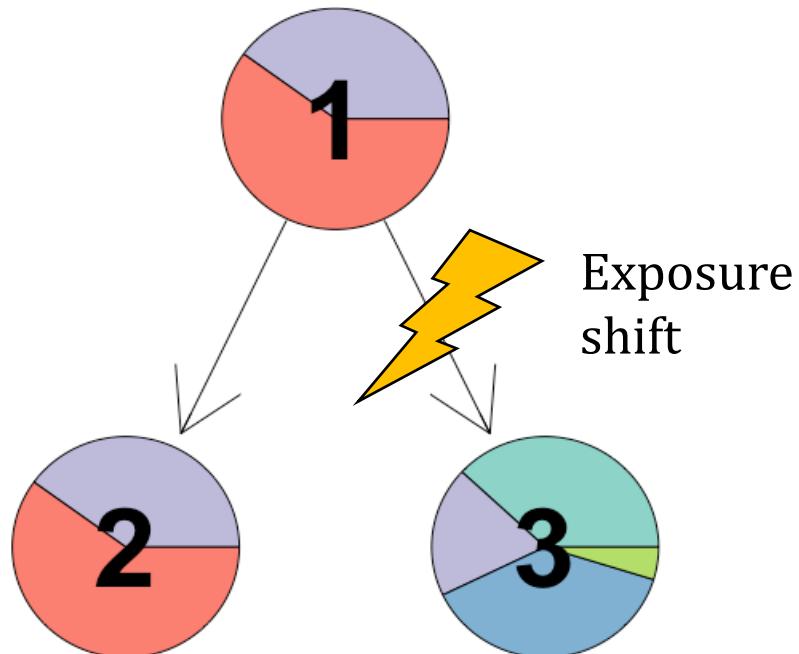
(b)



(c)

Tree-constrained exposure (TE) inference

TE problem: [Christensen, Leiserson and El-Kebir, PSB 2020]:
Given phylogenetic tree T and feature matrix P ,
find k exposure shifts along edges of T



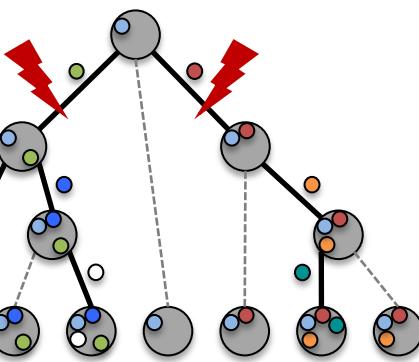
- **Key idea:** exposure may change along edges of phylogenetic tree
- TE interpolates between single exposure (SE) and independent exposure (IE) problems
- Model selection for k

Tree-constrained exposure inference

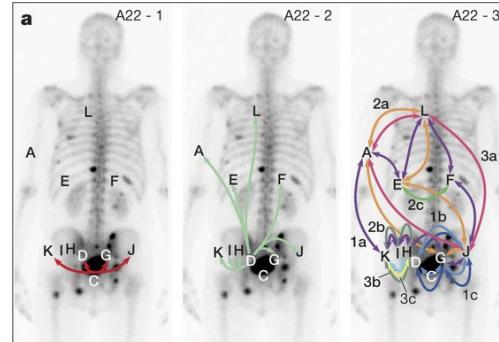
Conclusion

Downstream analyses in cancer genomics **critically rely** on accurate tumor phylogeny inference

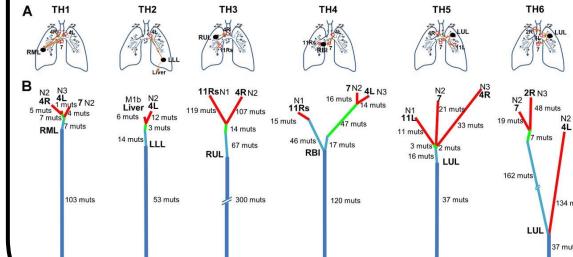
Identify targets for treatment



Understand metastatic development



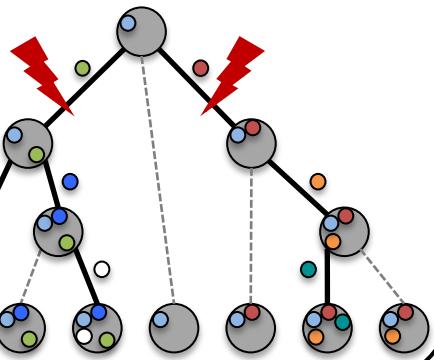
Recognize common patterns of tumor evolution across patients



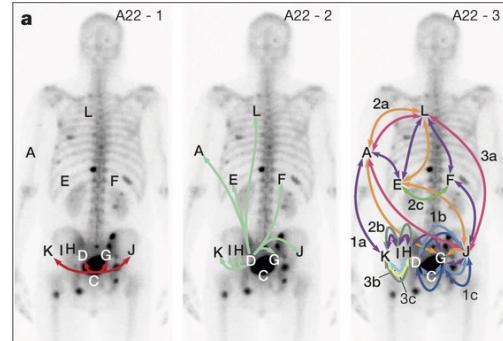
Conclusion

Downstream analyses in cancer genomics **critically rely** on accurate tumor phylogeny inference

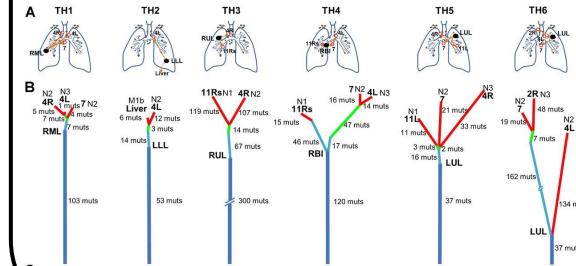
Identify targets for treatment



Understand metastatic development



Recognize common patterns of tumor evolution across patients



1. Theory and background of perfect phylogeny mixture (PPM) problem

2. Simulation study to assess factors contributing to and impact of non-uniqueness

3. Summarizing solution space using multiple consensus trees

4. Phylogenetic integration

5. Example of downstream application: Mutational signature dynamics

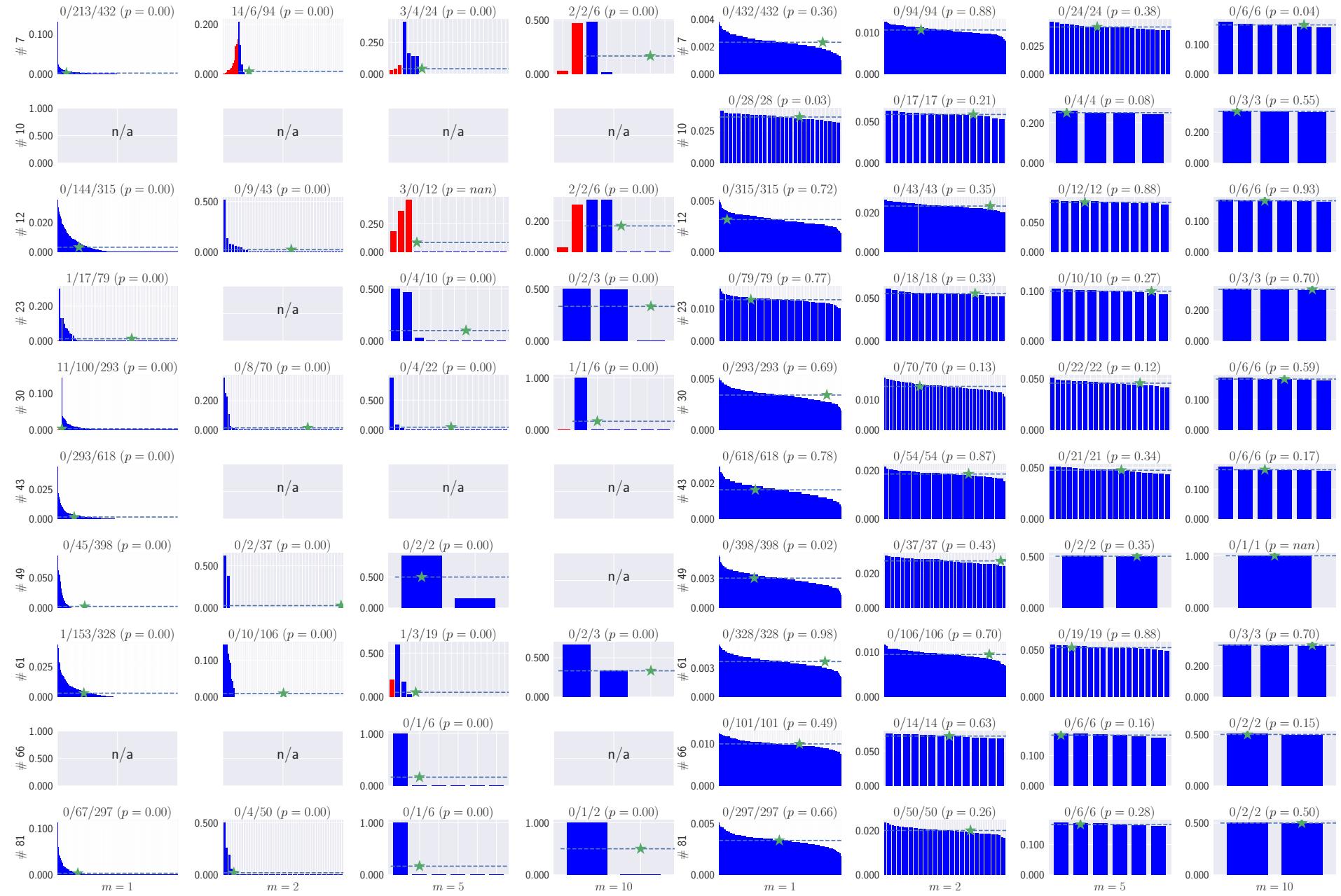
Acknowledgments

El-Kebir group

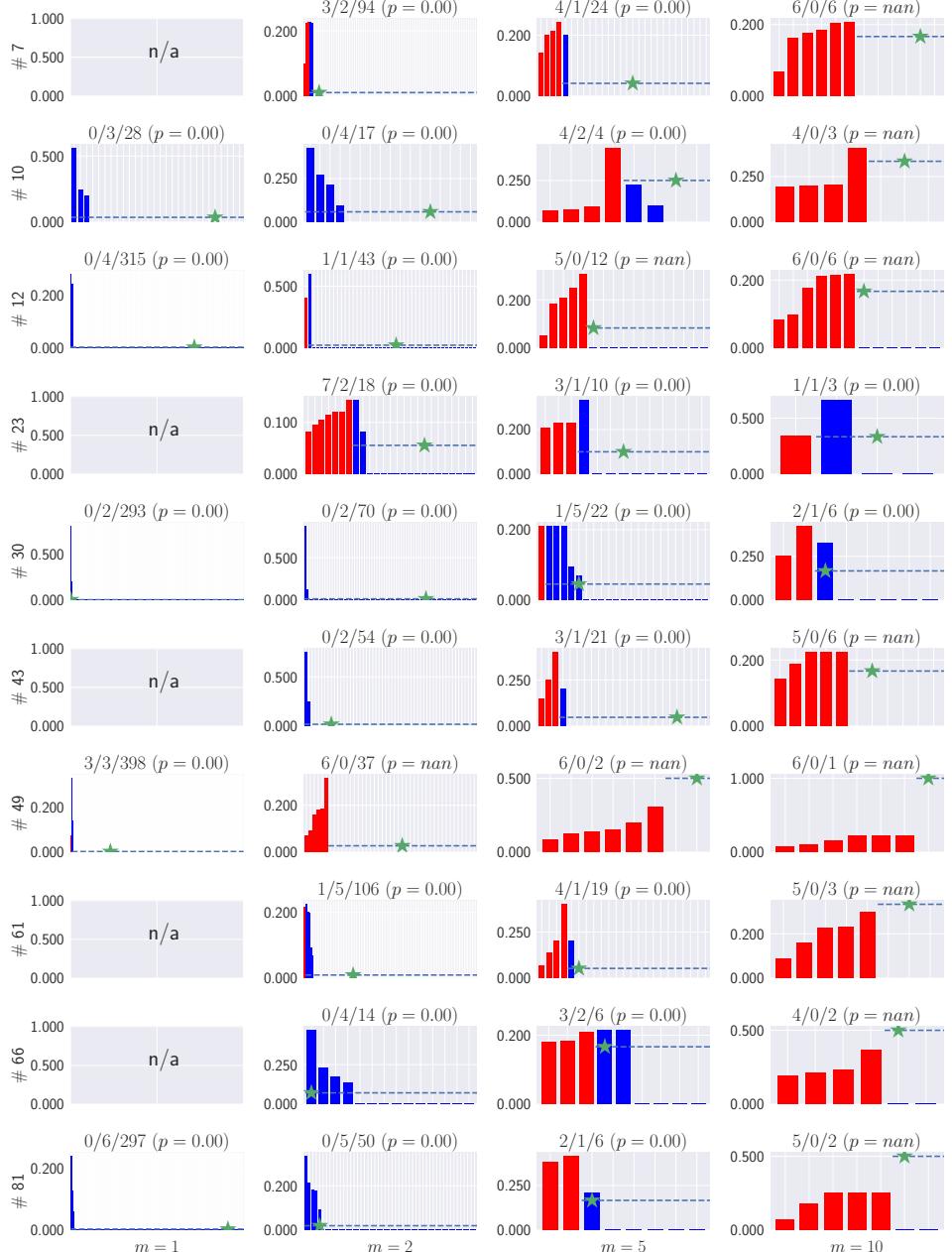
- **Nuraini Aguse** (graduated)
- **Sarah Christensen** (graduated)
- **Dikshant Pradhan** (graduated)
- **Yuanyuan Qi**
- Jiaqi Wu (graduated)
- Juho Kim (graduated)
- Palash Sashittal (graduated)
- Leah Weber
- Stefan Ivanovic
- Chuanyi Zhang
- Experiments were run on NCSA's Blue Waters supercomputer
- This work was supported by:
 - UIUC Center for Computational Biotechnology and Genomic Medicine (grant: CSN 1624790)
 - National Science Foundation (CCF 1850502, CCF 2046488)

PhyloWGS

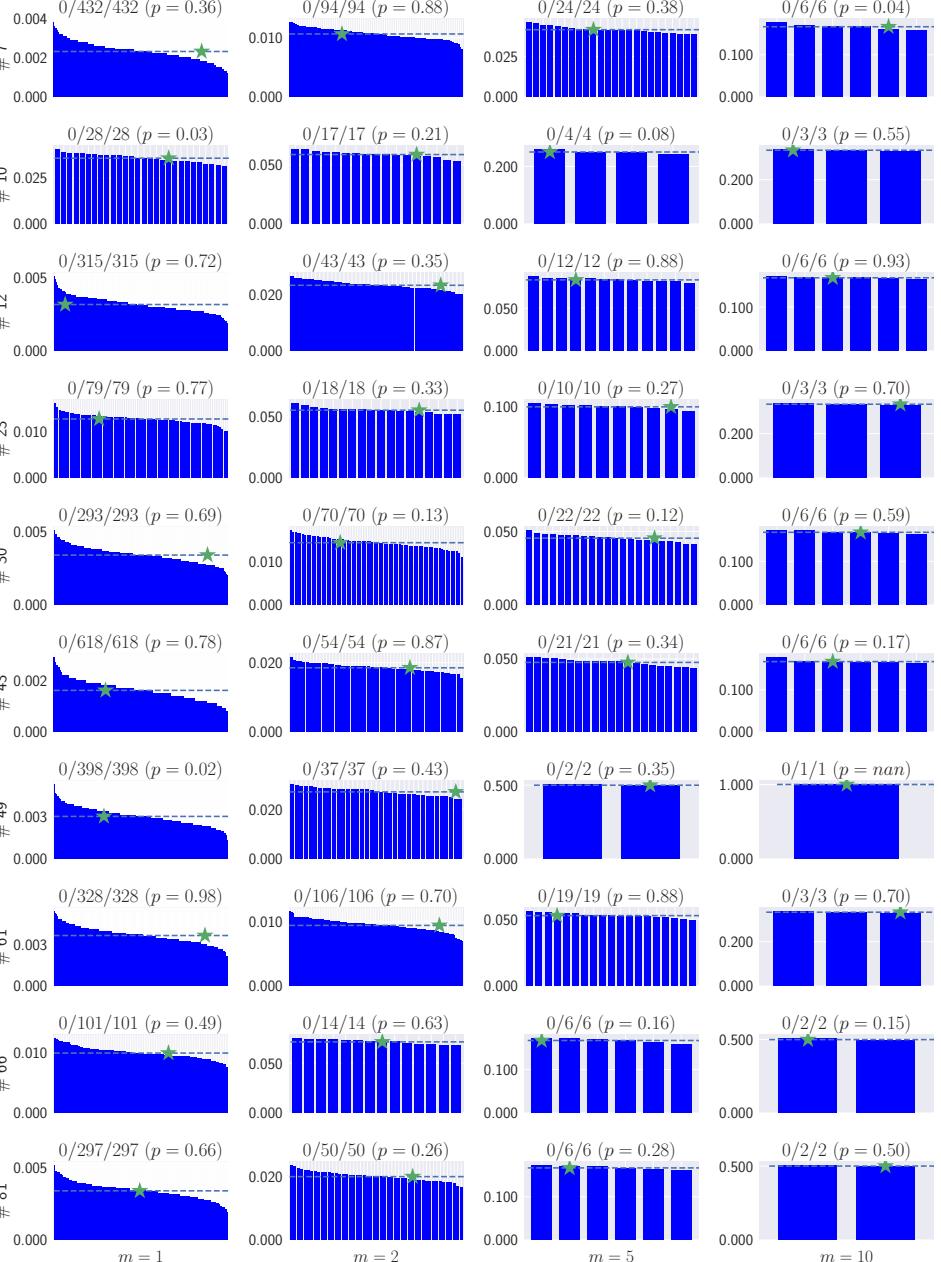
Rejection Sampling



Canopy



Rejection Sampling



Somatic Mutations Occur at Different Genomic Scales

