Implications of Non-uniqueness of Solutions in Cancer Phylogenetics

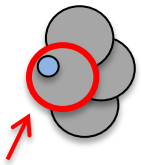
Mohammed El-Kebir

University of Illinois at Urbana Champaign,
Department of Computer Science

October, 2019

Tumorigenesis: Cell Mutation

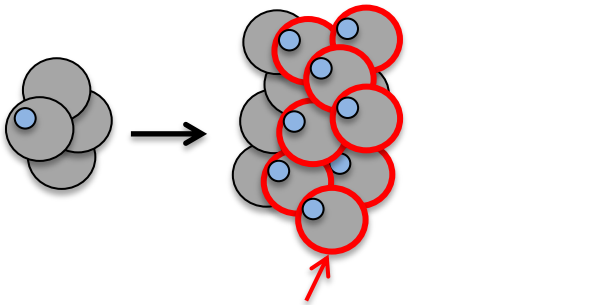
Clonal Evolution Theory of Cancer
[Nowell, 1976]



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

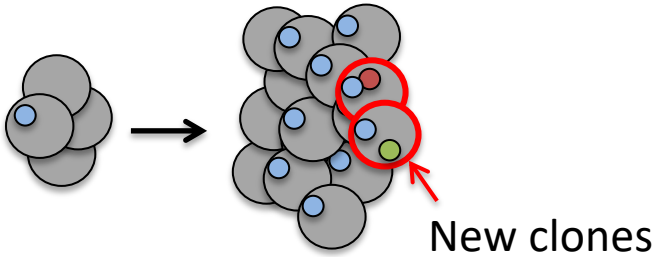
Tumorigenesis: Cell Mutation

Clonal Evolution Theory of Cancer
[Nowell, 1976]



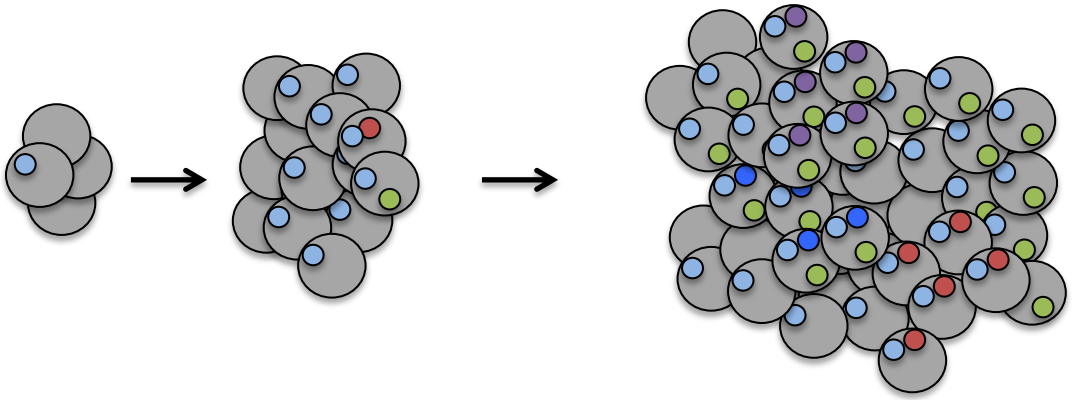
Tumorigenesis: Cell Mutation

Clonal Evolution Theory of Cancer
[Nowell, 1976]



Tumorigenesis: Cell Mutation & Division

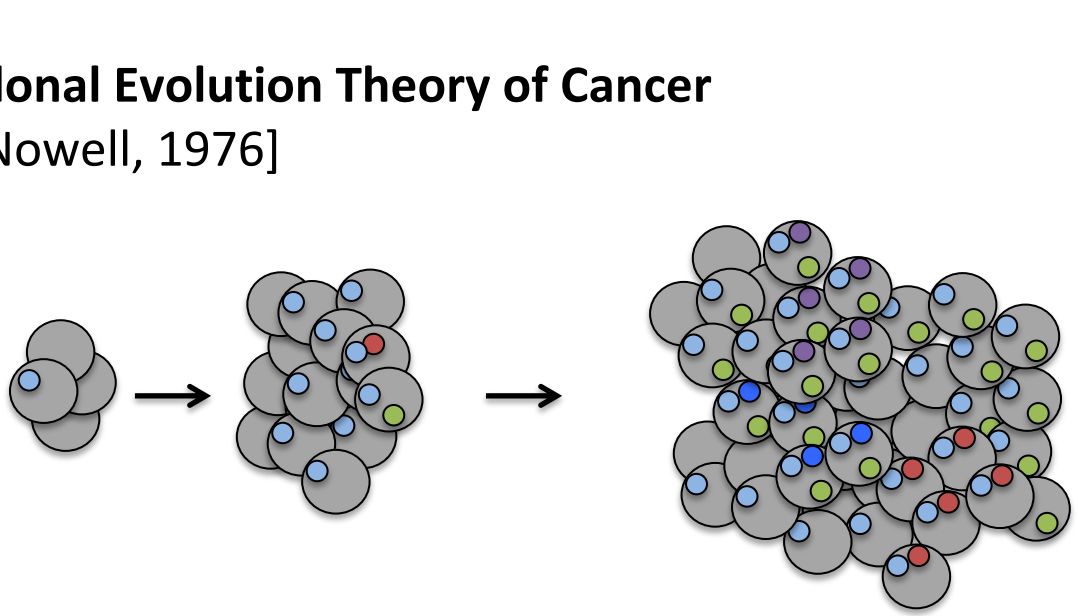
Clonal Evolution Theory of Cancer
[Nowell, 1976]



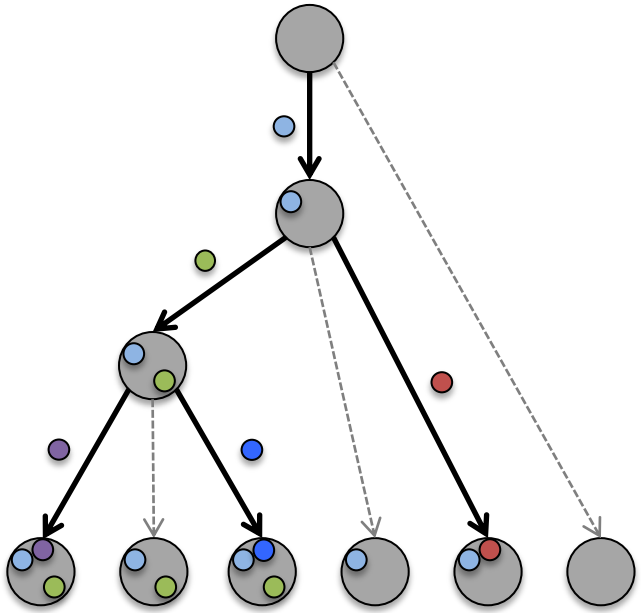
Intra-Tumor
Heterogeneity

Tumorigenesis: Cell Mutation & Division

Clonal Evolution Theory of Cancer
[Nowell, 1976]



Intra-Tumor Heterogeneity

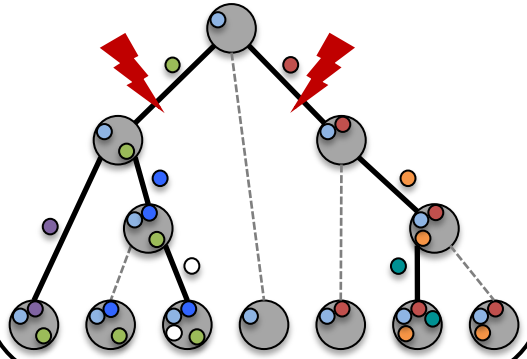


Phylogenetic Tree ***T***

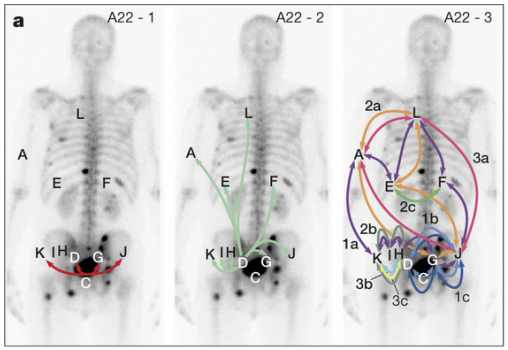
Question: Why are tumor phylogenies important?

Phylogenies are Key to Understanding Cancer

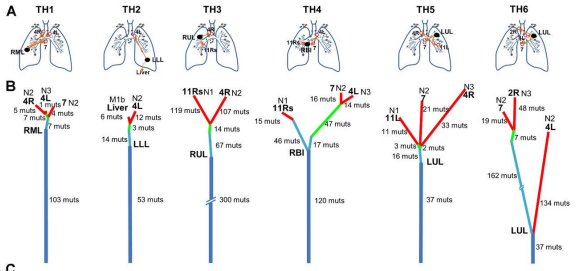
Identify targets for treatment



Understand metastatic development

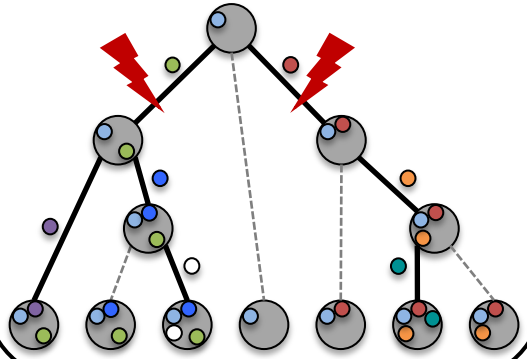


Recognize common patterns of tumor evolution across patients

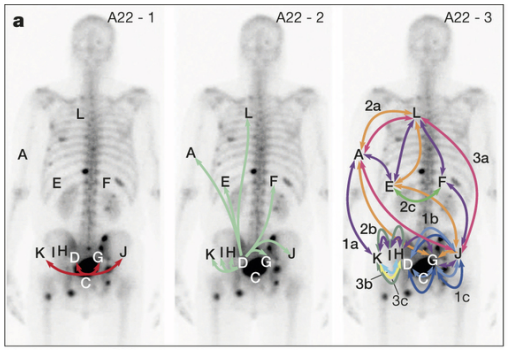


Phylogenies are Key to Understanding Cancer

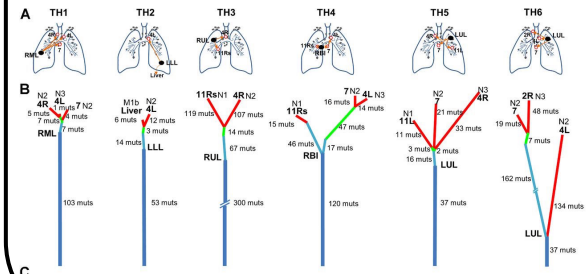
Identify targets for treatment



Understand metastatic development



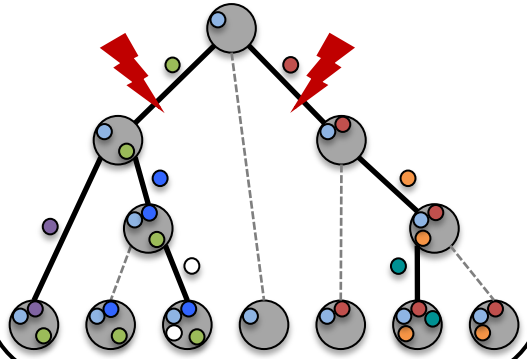
Recognize common patterns of tumor evolution across patients



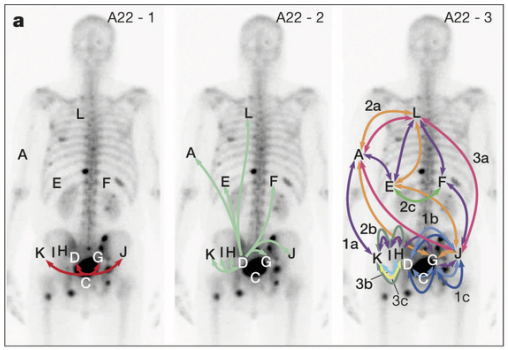
These downstream analyses **critically rely** on accurate tumor phylogeny inference

Phylogenies are Key to Understanding Cancer

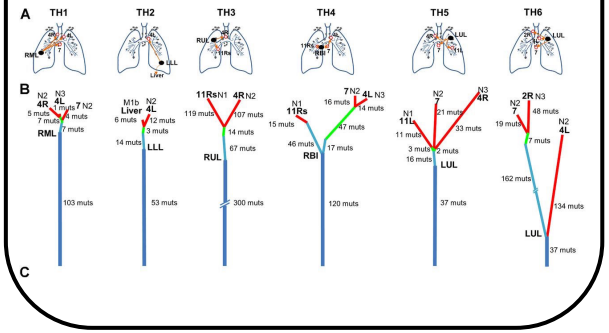
Identify targets for treatment



Understand metastatic development



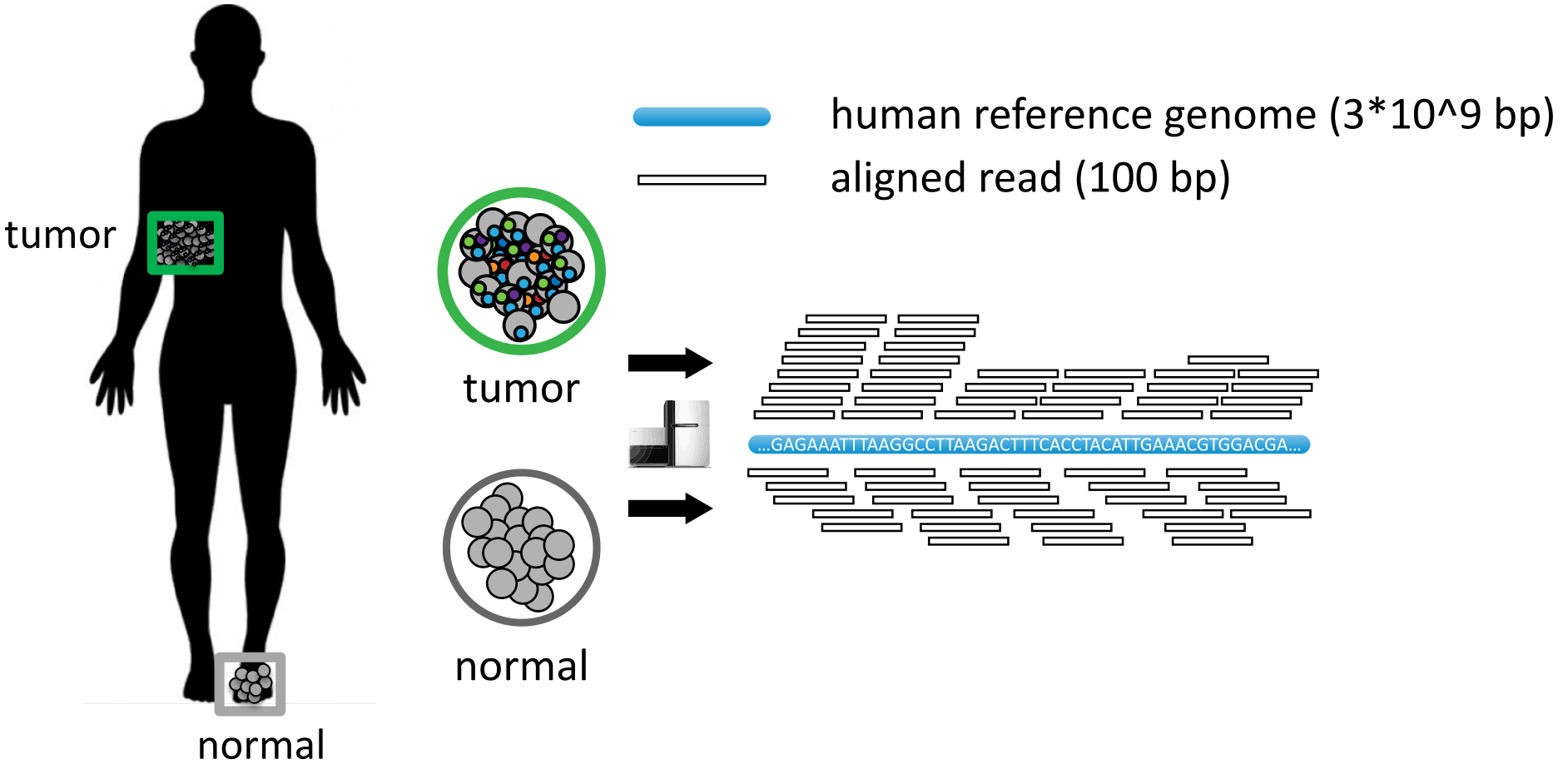
Recognize common patterns of tumor evolution across patients



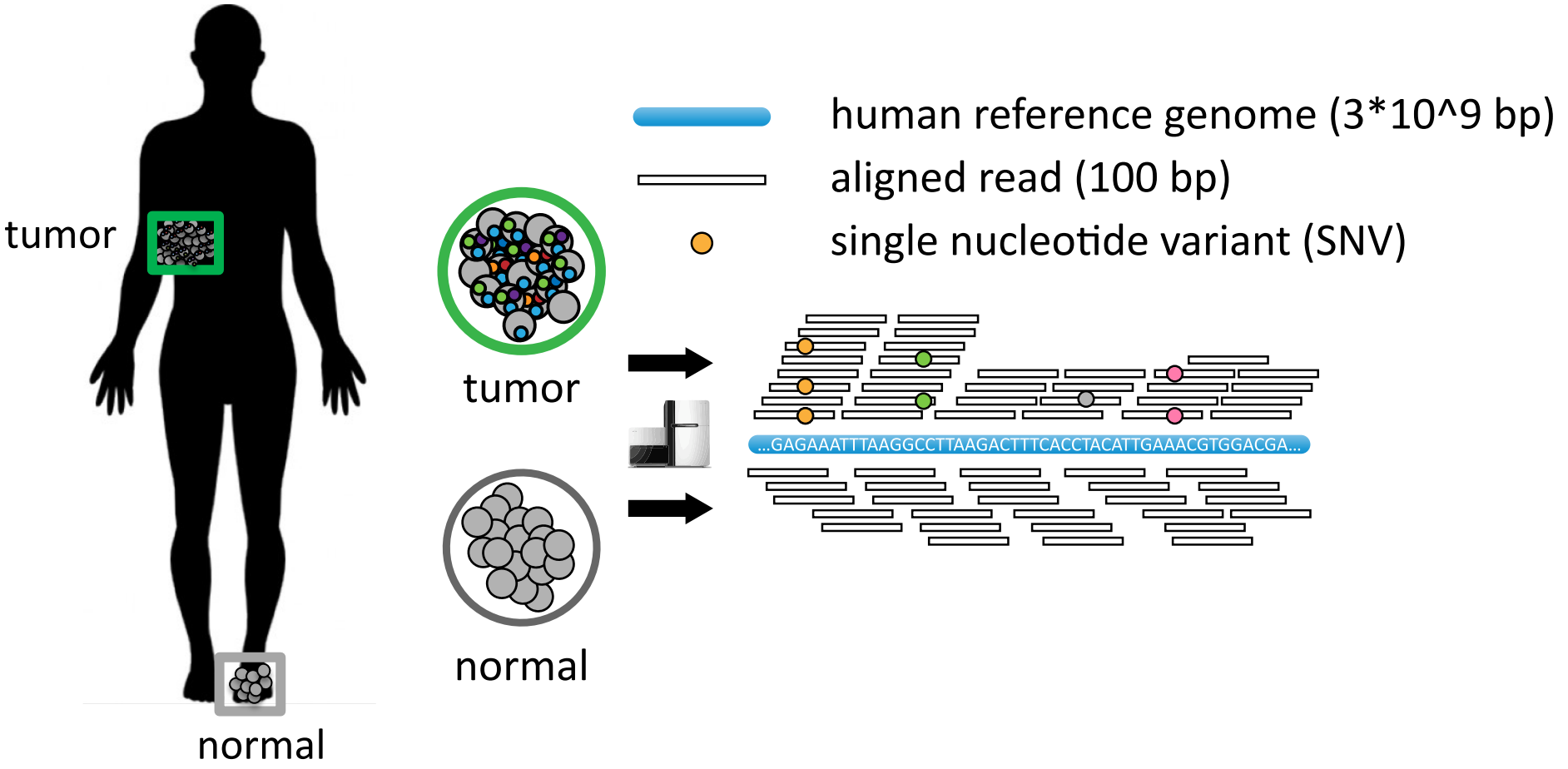
These downstream analyses **critically rely** on accurate tumor phylogeny inference

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

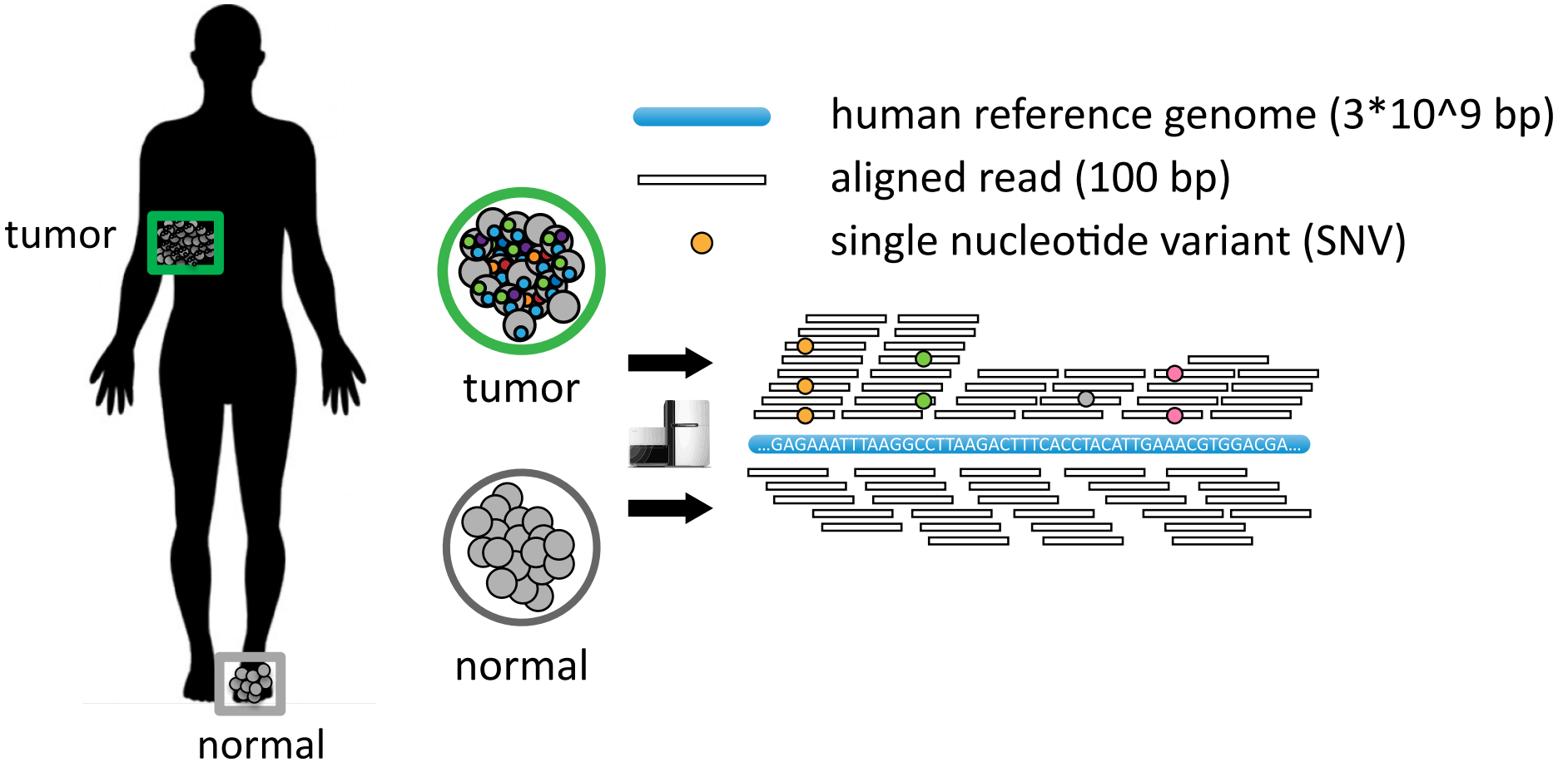
Additional Challenge in Cancer Phylogenetics



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, AMOB 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?
- How does non-uniqueness affect current methods?

3. Almost uniform sampling: [To be submitted]

- Reducing PPM to SATISFIABILITY

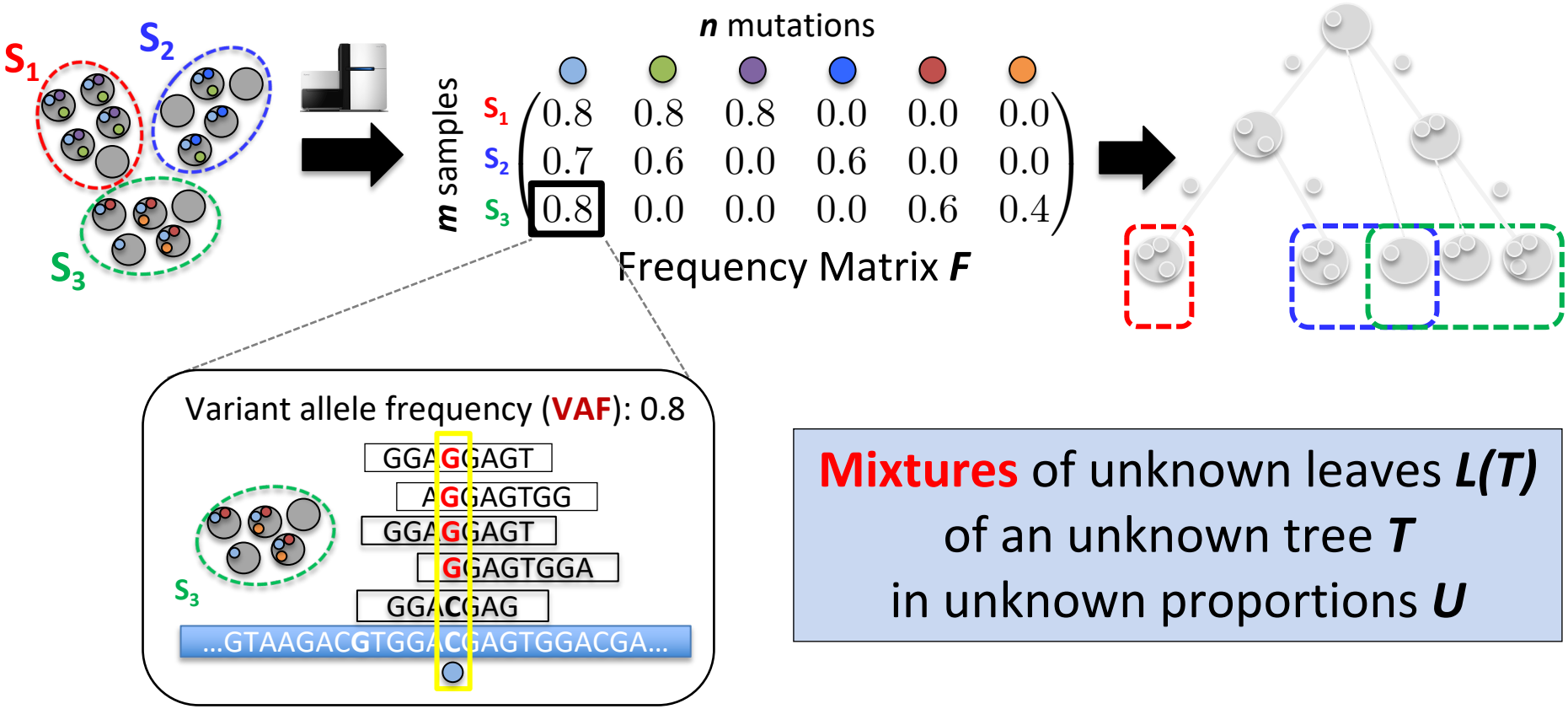
4. Summarizing solution space: [ISMB/ECCB 2019]

- Multiple consensus tree problem

5. Applications

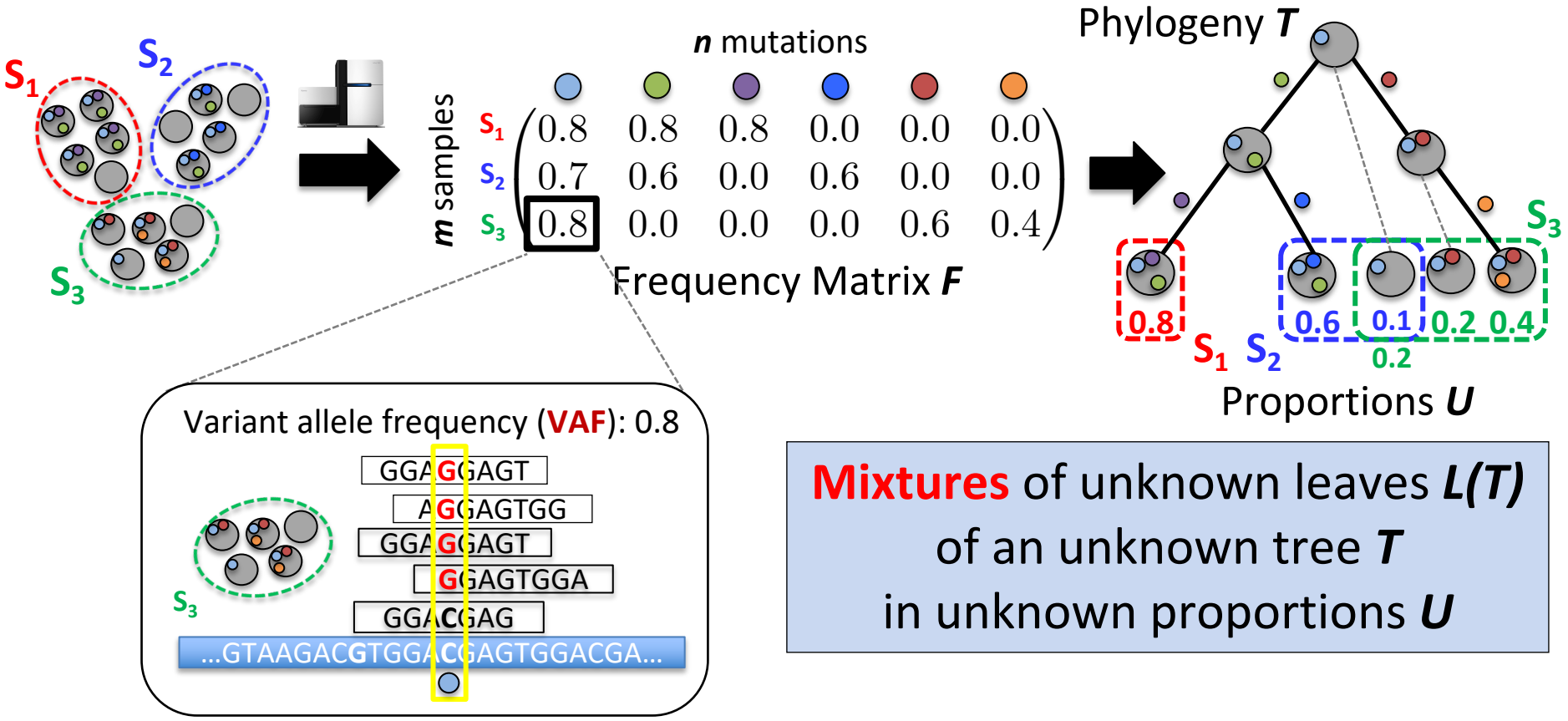
- Mutational signature dynamics [PSB 2020]

Sequencing and Tumor Phylogeny Inference



Mixtures of unknown leaves ***L(T)***
of an unknown tree ***T***
in unknown proportions ***U***

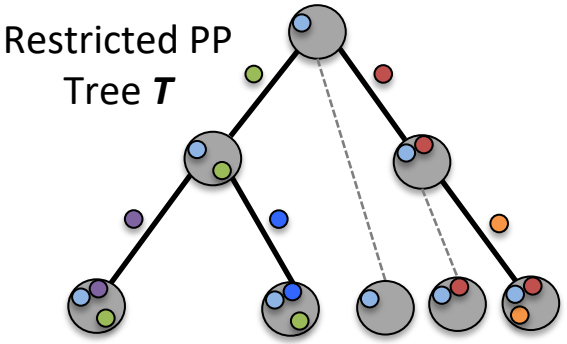
Sequencing and Tumor Phylogeny Inference



Mixtures of unknown leaves ***L(T)***
of an unknown tree ***T***
in unknown proportions ***U***

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

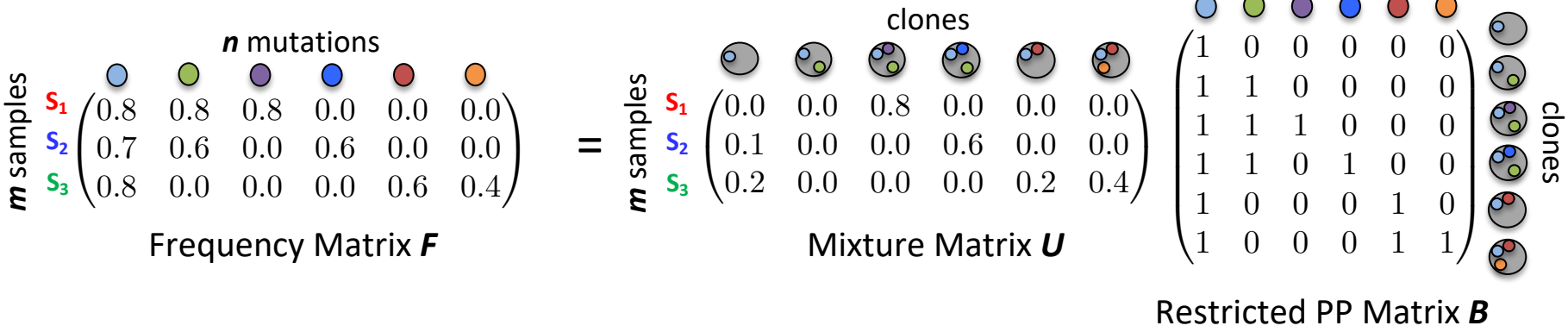
Perfect Phylogeny Mixture



Assumptions:

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

1-1 Equivalent



Rows of U are proportions:

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

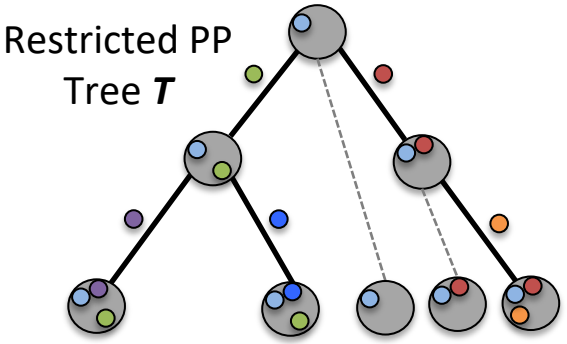
Perfect Phylogeny Theorem
[Estabrook, 1971]
[Gusfield, 1991]

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

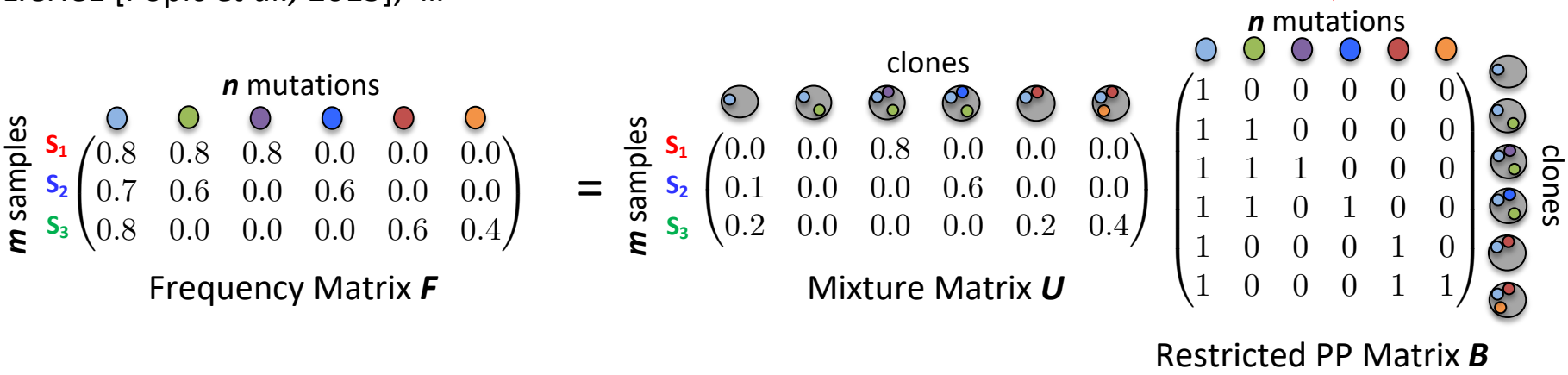
Previous Work

Variant of PPM:

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



1-1 Equivalent



Rows of ***U*** are proportions:

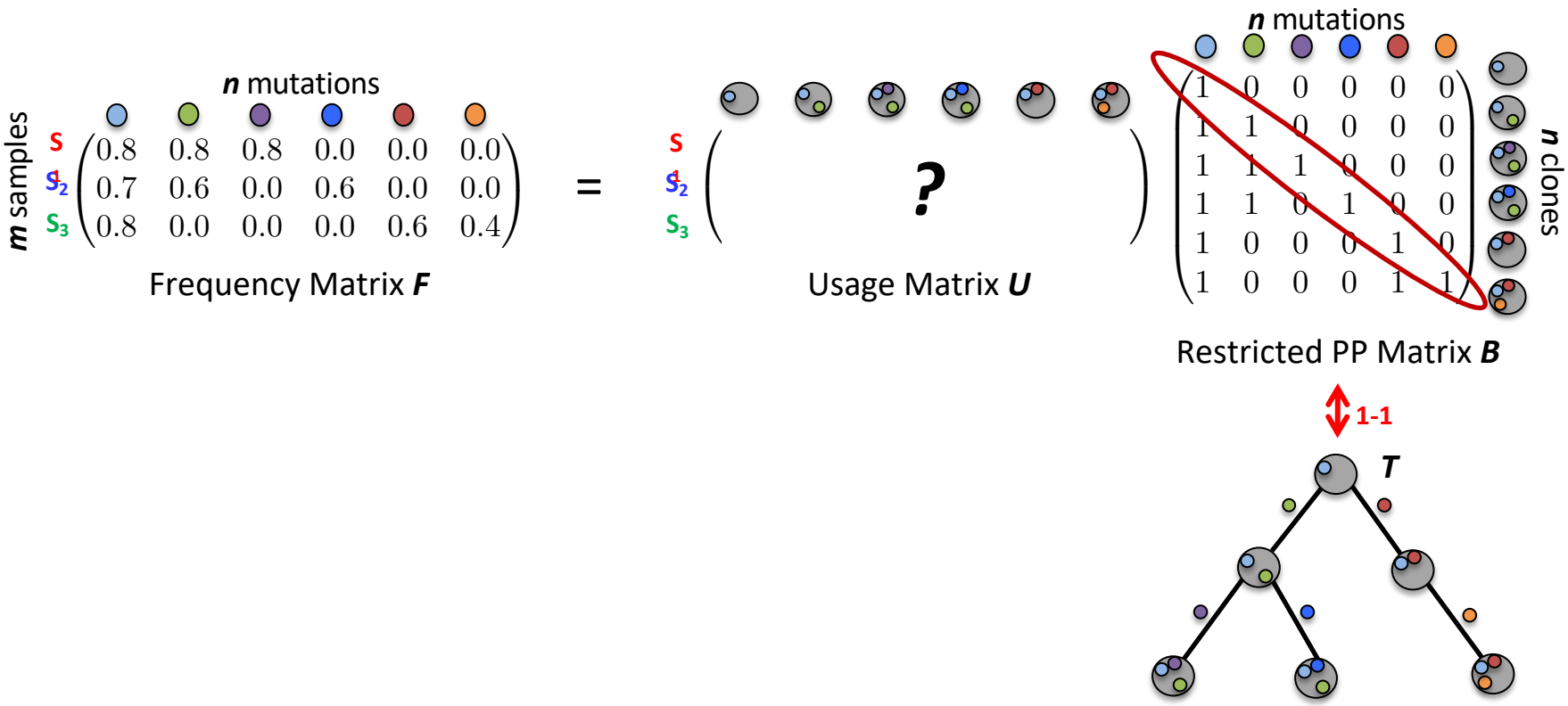
$$u_{pj} \geq 0 \text{ and } \sum_j u_{pj} \leq 1$$

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

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

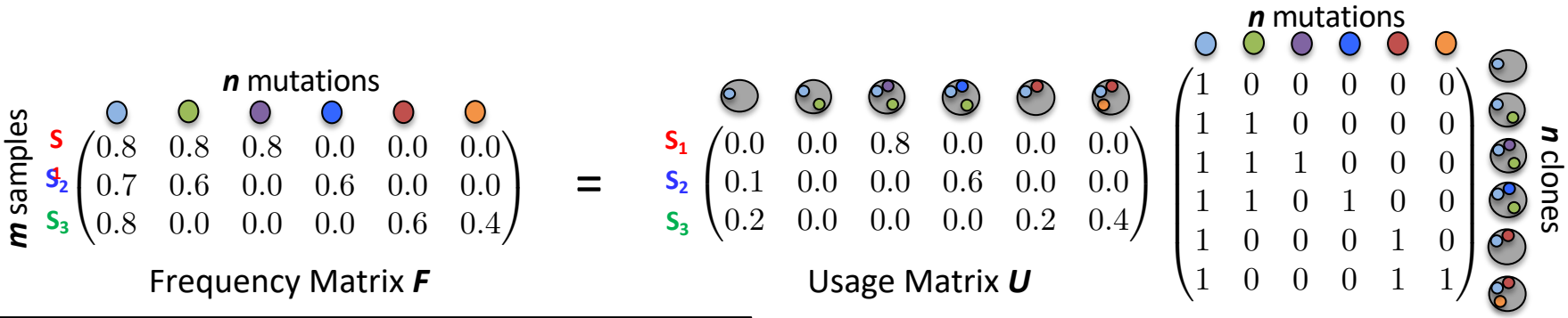
Given *F* and *T* (or *B*), is there a usage matrix *U*?

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



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

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



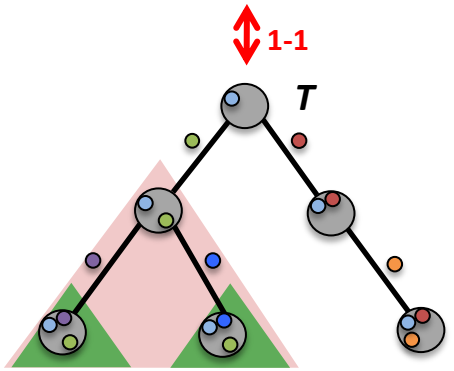
Frequency Matrix ***F*** | Usage Matrix ***U*** | Restricted PP Matrix ***B***

Lemma: *B* is invertible

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

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*?

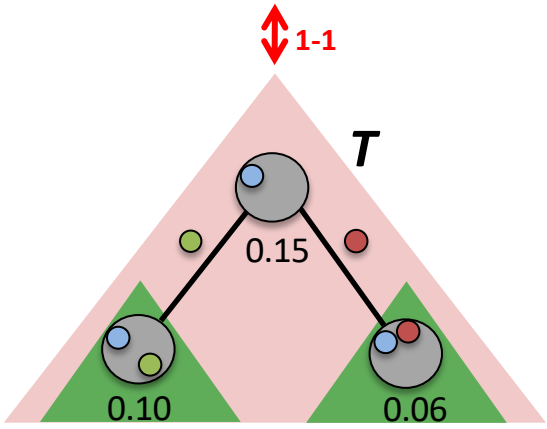
PPM: Given *F*, find *U* and *B* such that *F = U B*

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

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 = U B***

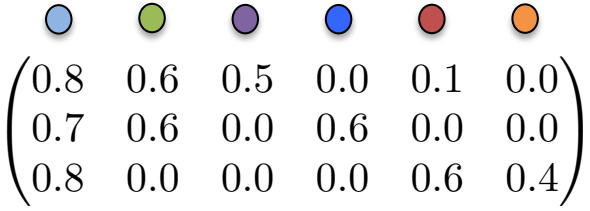
Lemma:

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

Lemma (Sum Condition):

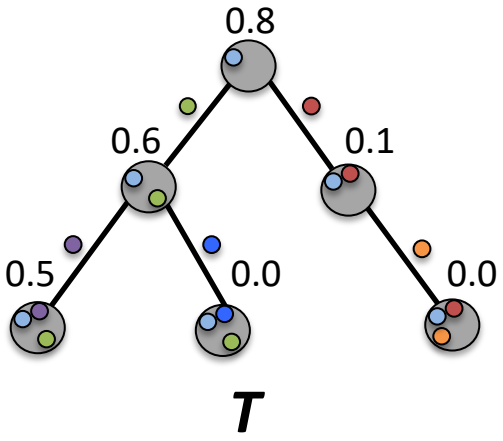
Given ***F*** and ***T***, for all samples *p* and mutations *j*, $f_{pj} \geq \sum_{k \text{ child of } j} f_{pk}$

necessary sufficient



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

F



T

Combinatorial Characterization of Solutions

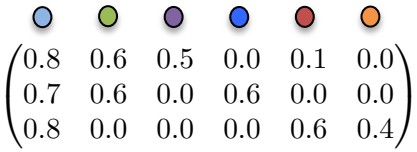
PPM: Given ***F***, find ***U*** and ***B*** such that ***F = U B***

Lemma (Sum Condition):
Given ***F*** and ***T***, for all samples *p* and mutations *j*, $f_{pj} \geq \sum_{k \text{ child of } j} f_{pk}$

necessary
sufficient

Lemma (Ancestry Condition):
Given ***F*** and ***T***, for all samples *p* and mutations *k* child of *j*, $f_{pj} \geq f_{pk}$

necessary



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

F

Combinatorial Characterization of Solutions

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

Lemma (Sum Condition):
Given ***F*** and ***T***, for all samples *p* and mutations *j*, $f_{pj} \geq \sum_{k \text{ child of } j} f_{pk}$

necessary
sufficient

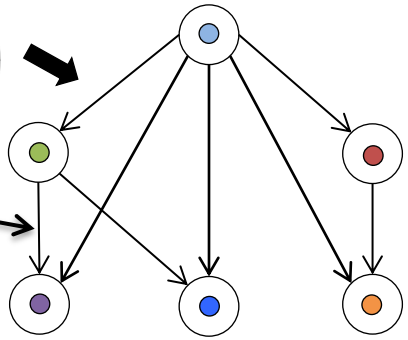
Lemma (Ancestry Condition):
Given ***F*** and ***T***, for all samples *p* and mutations *k* child of *j*, $f_{pj} \geq f_{pk}$

necessary

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

F

potential parental relationship



Ancestry Graph ***G* = (*V*, *A*)**

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 = U B***

Lemma (Sum Condition):
Given ***F*** and ***T***, for all samples *p* and mutations *j*, $f_{pj} \geq \sum_{k \text{ child of } j} f_{pk}$

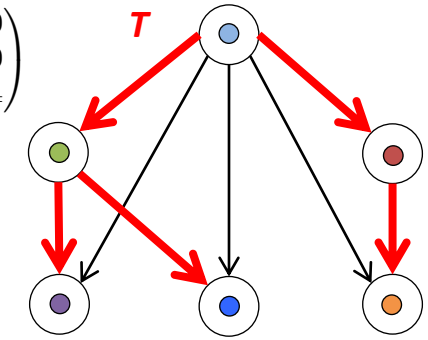
necessary
sufficient

Lemma (Ancestry Condition):
Given ***F*** and ***T***, for all samples *p* and mutations *k* child of *j*, $f_{pj} \geq f_{pk}$

necessary

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

F



Ancestry Graph ***G* = (*V*, *A*)**

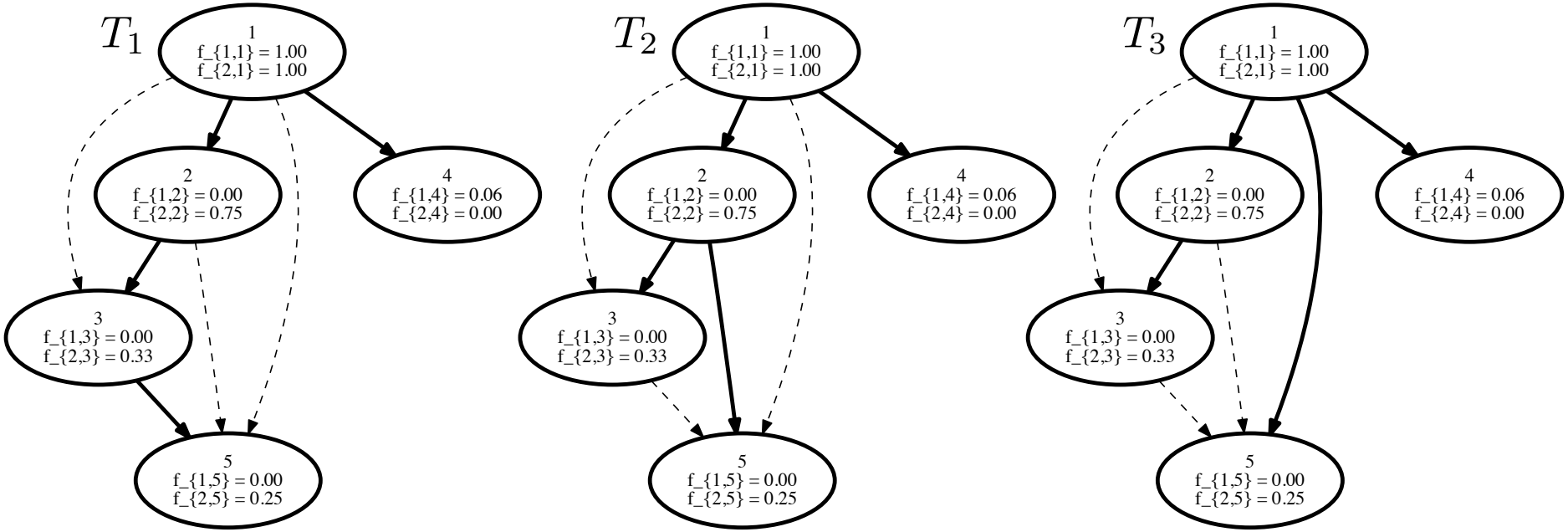
Ancestry graph *G* = (*V*, *A*); given ***F***

- 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

Theorem 2:
PPM is NP-complete even if *m* = 2

Non-uniqueness of Solutions to PPM



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

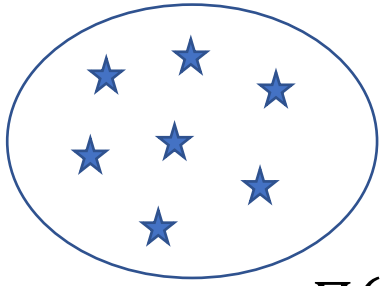
Question 1: Can we determine the number of solutions?

Question 2: Can we sample solutions uniformly at random?

Intermezzo: Problems/Algorithms in CS

Problem Π with instance/input X and feasible solution set $\Pi(X)$:

- Decision problem:
  - Is $\Pi(X) = \emptyset$?
- Optimization problem:
  - Find $y^* \in \Pi(X)$ s.t. $f(y^*)$ is optimum.
- Counting problem:
  - Compute $|\Pi(X)|$.
- Sampling problem:
  - Sample uniformly from $\Pi(X)$.
- Enumeration problem:
  - Enumerate all solutions in $\Pi(X)$.

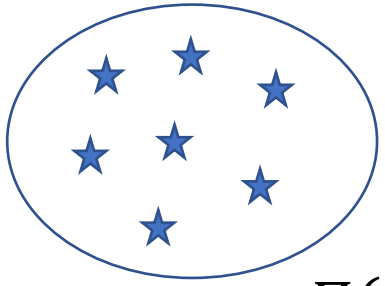


$\Pi(X)$

Intermezzo: Problems/Algorithms in CS

Problem Π with instance/input X and feasible solution set $\Pi(X)$:

- Decision problem:
  - Is $\Pi(X) = \emptyset$?
- Optimization problem:
  - Find $y^* \in \Pi(X)$ s.t. $f(y^*)$ is optimum.
- Counting problem:
  - Compute $|\Pi(X)|$.
- Sampling problem:
  - Sample uniformly from $\Pi(X)$.
- Enumeration problem:
  - Enumerate all solutions in $\Pi(X)$.



$\Pi(X)$

Algorithm:

Set of instructions for solving problem.

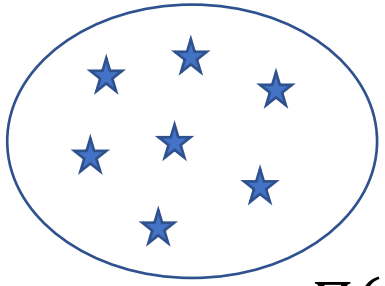
- Exact
- Heuristic

Running time: How does the number of steps scale as a function of $|X|$?

Intermezzo: Problems/Algorithms in CS

Problem Π with instance/input X and feasible solution set $\Pi(X)$:

- Decision problem:
  - Is $\Pi(X) = \emptyset$?
- Optimization problem:
  - Find $y^* \in \Pi(X)$ s.t. $f(y^*)$ is optimum.
- Counting problem:
  - Compute $|\Pi(X)|$.
- Sampling problem:
  - Sample uniformly from $\Pi(X)$.
- Enumeration problem:
  - Enumerate all solutions in $\Pi(X)$.



$\Pi(X)$

Algorithm:

Set of instructions for solving problem.

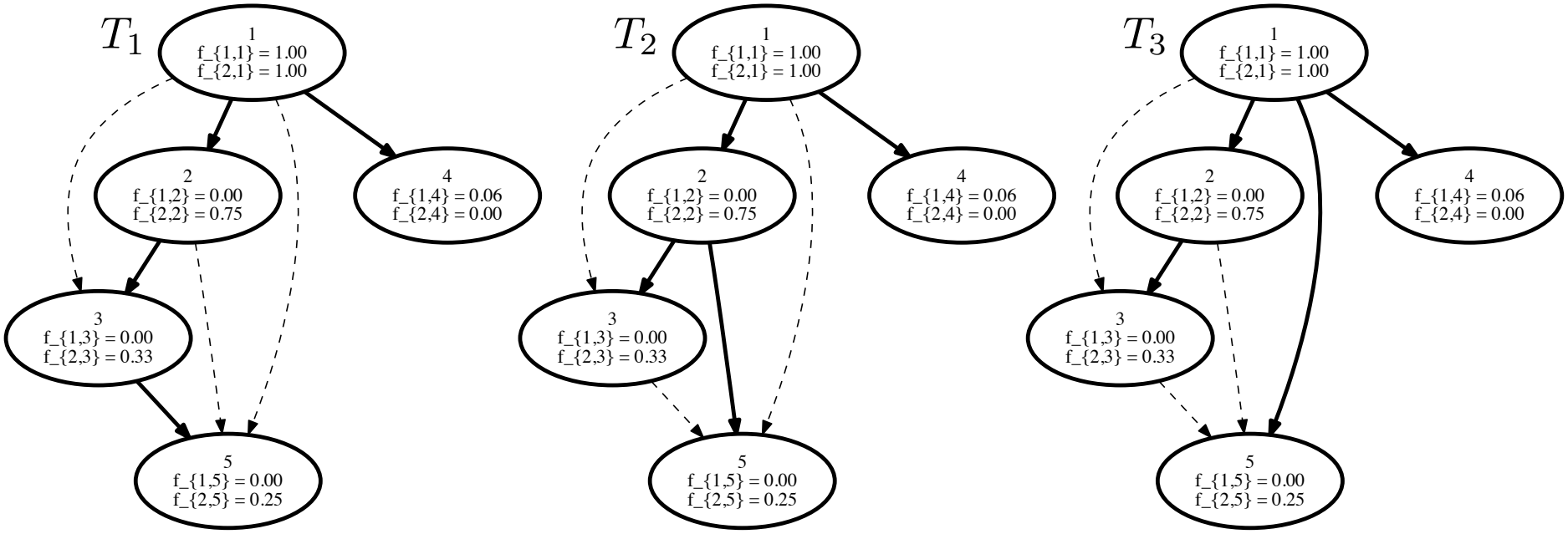
- Exact
- Heuristic

Running time: How does the number of steps scale as a function of $|X|$?

Problem != algorithm

Some problems do not admit efficient algorithms

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?

Counting problem

Sampling problem

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 = U B***

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 = U B***

#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 = U B***

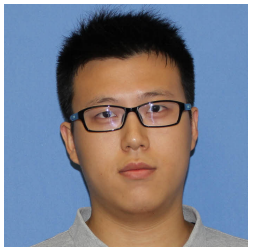
#P is the complexity class of counting problems whose decision problems are in NP

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]
- 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?

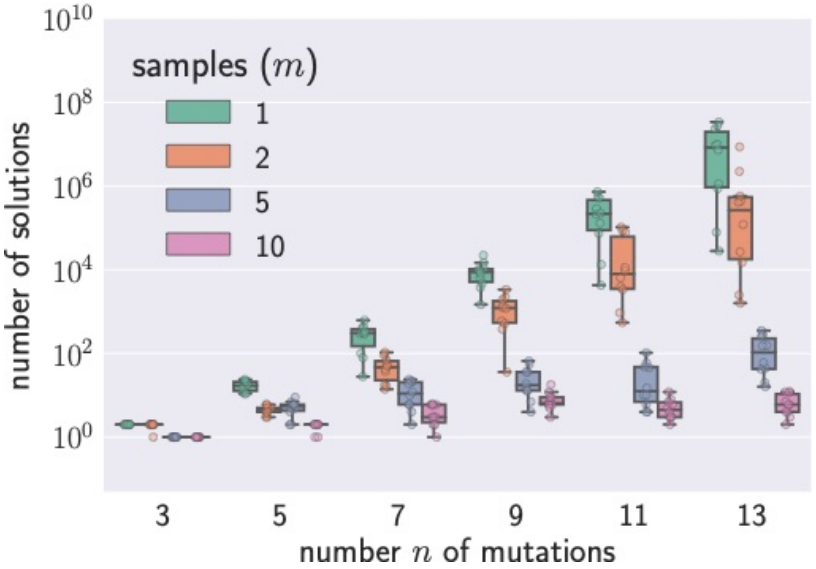
Dikshant Pradhan

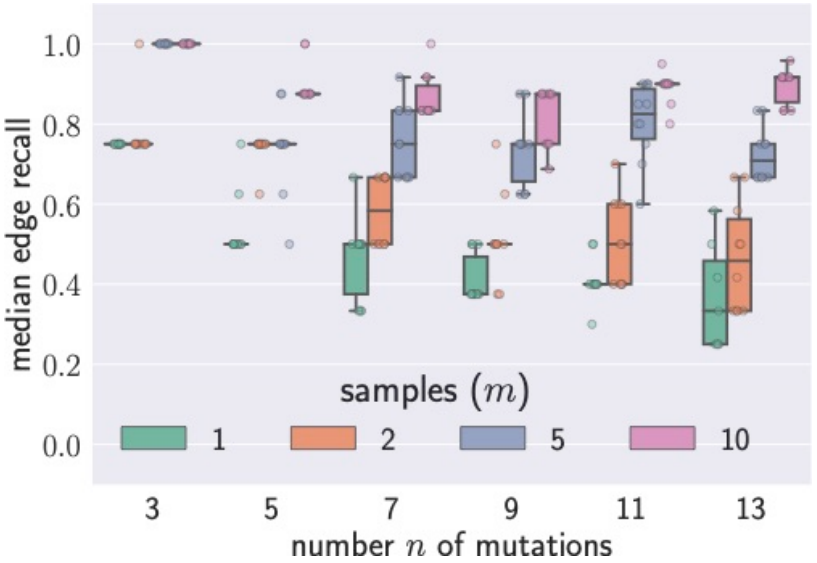
3. Almost uniform sampling: [To be submitted]
- Reducing PPM to SATISFIABILITY

4. Summarizing solution space: [ISMB/ECCB 2019]
- Multiple consensus tree problem

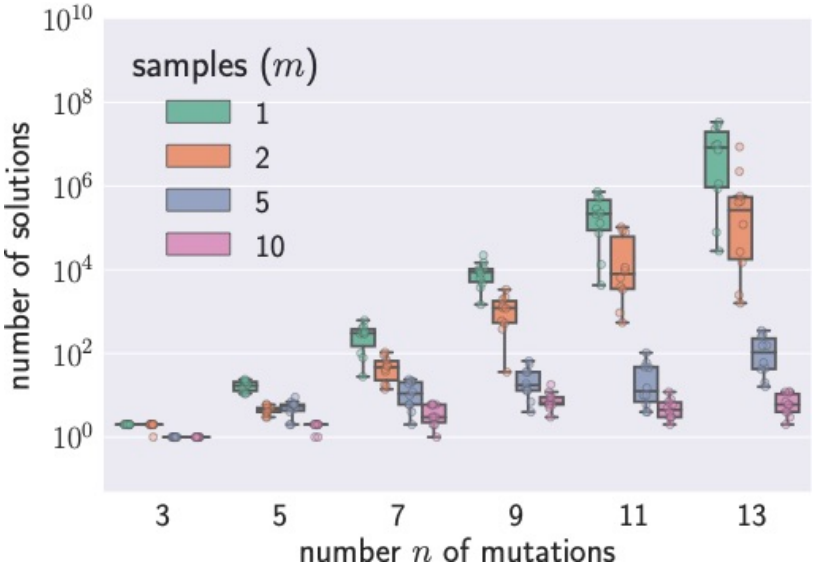
5. Applications
- Mutational signature dynamics [PSB 2020]

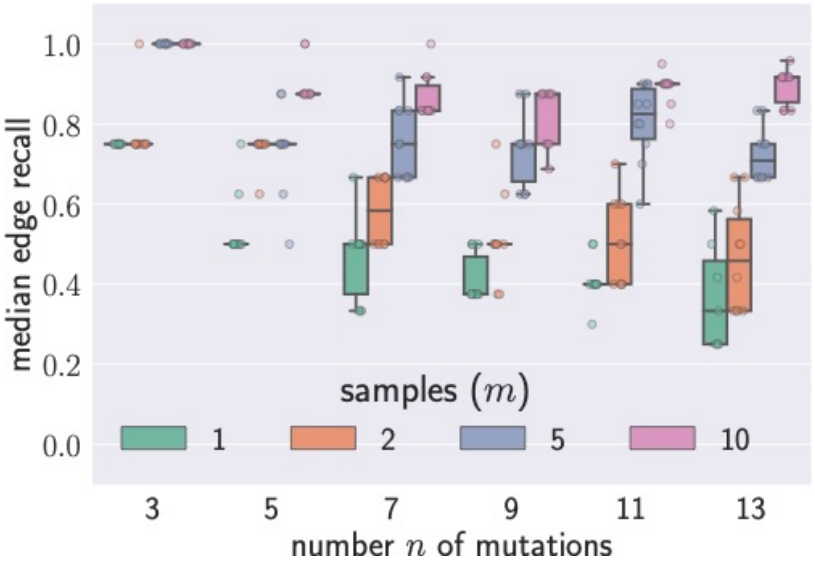
What Contributes to Non-uniqueness?

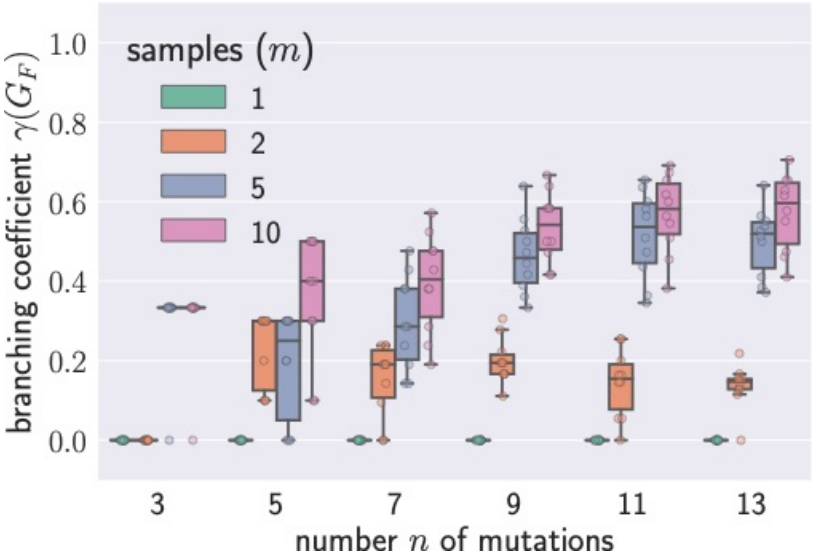


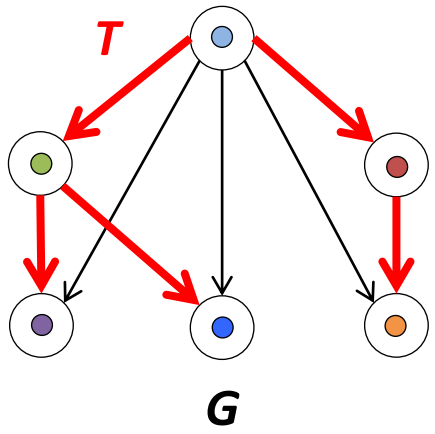


What Contributes to Non-uniqueness?

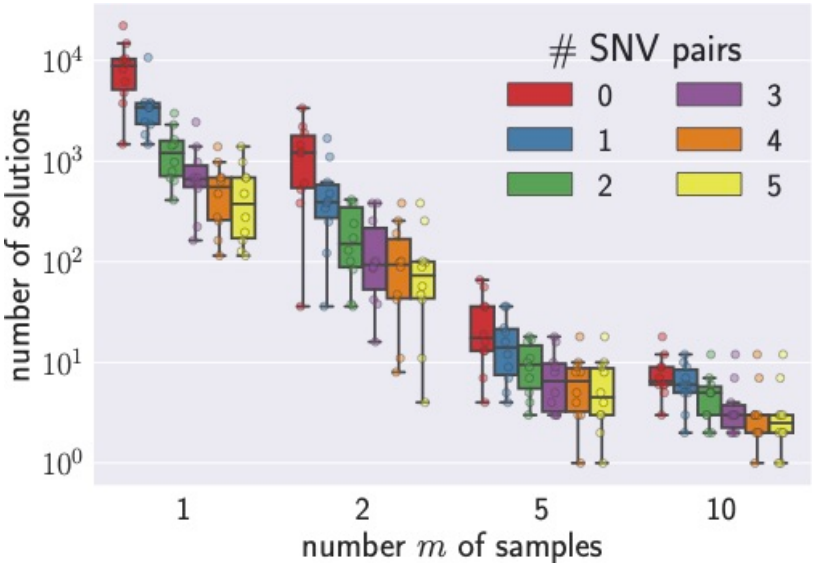


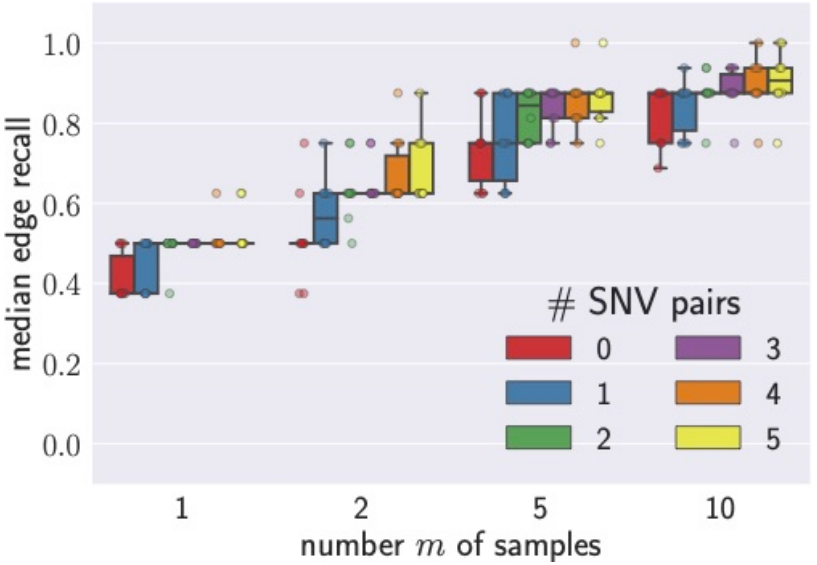


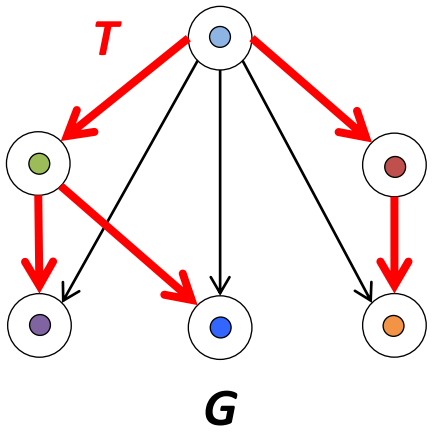




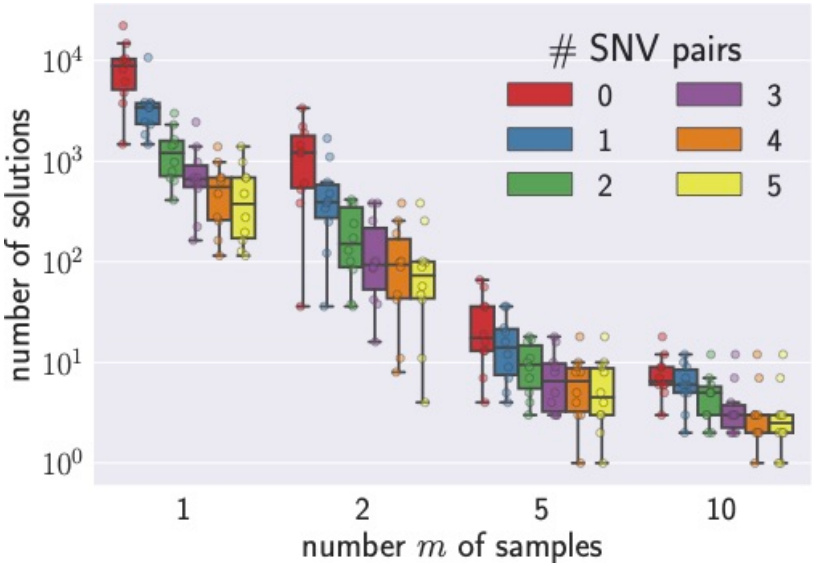
How to Reduce Non-Uniqueness?

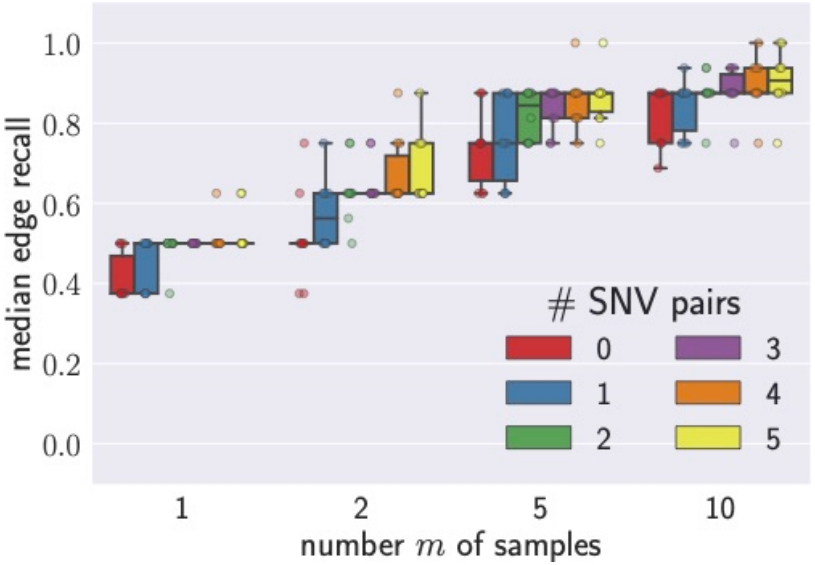


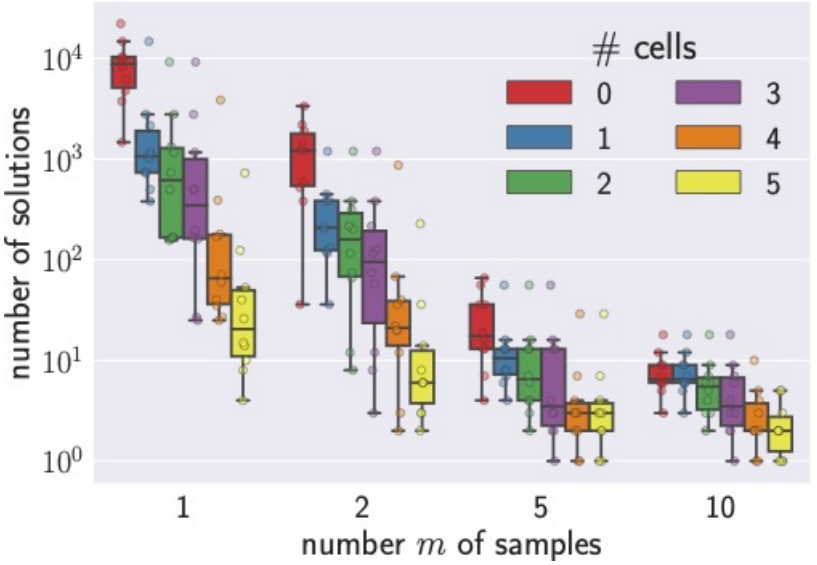


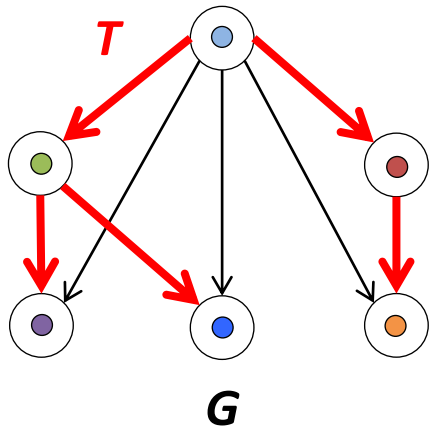


How to Reduce Non-Uniqueness?

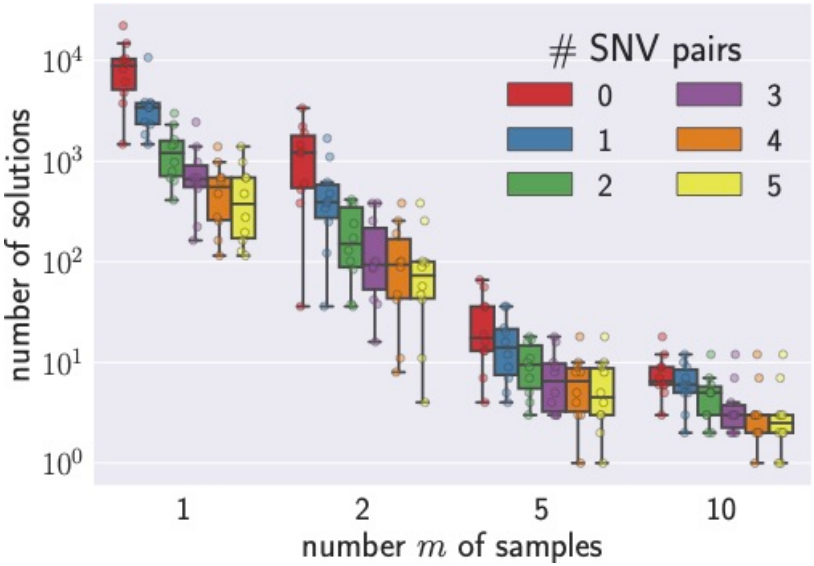


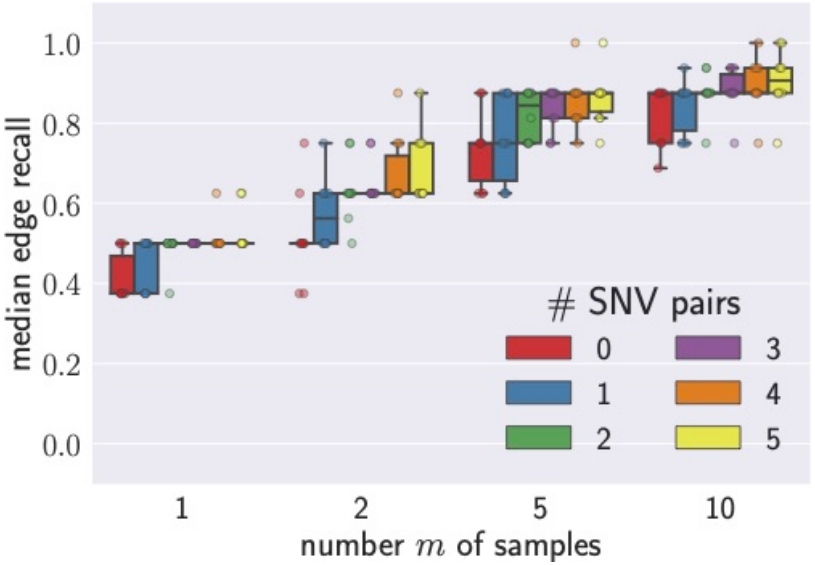


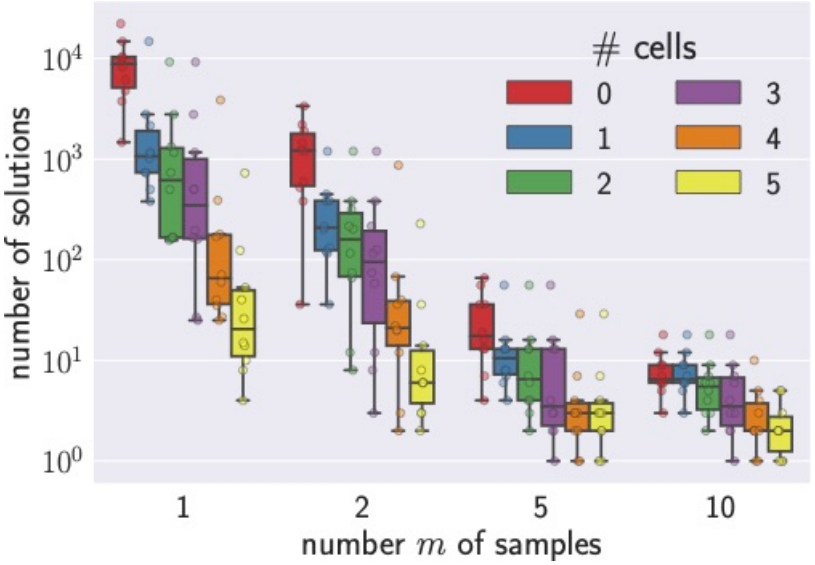


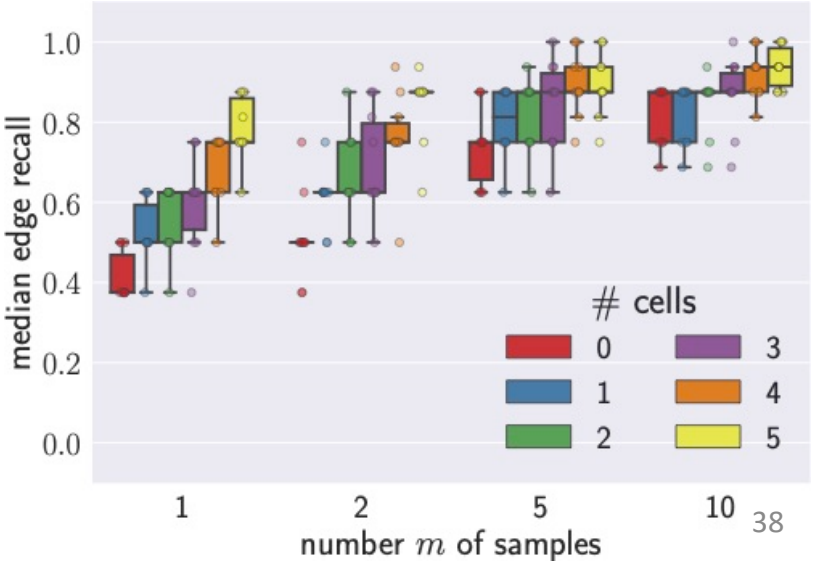


How to Reduce Non-Uniqueness?

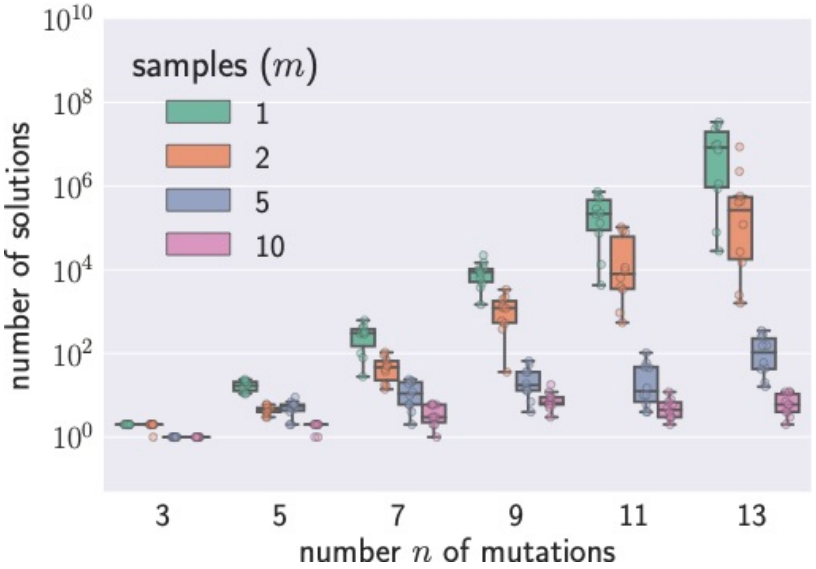


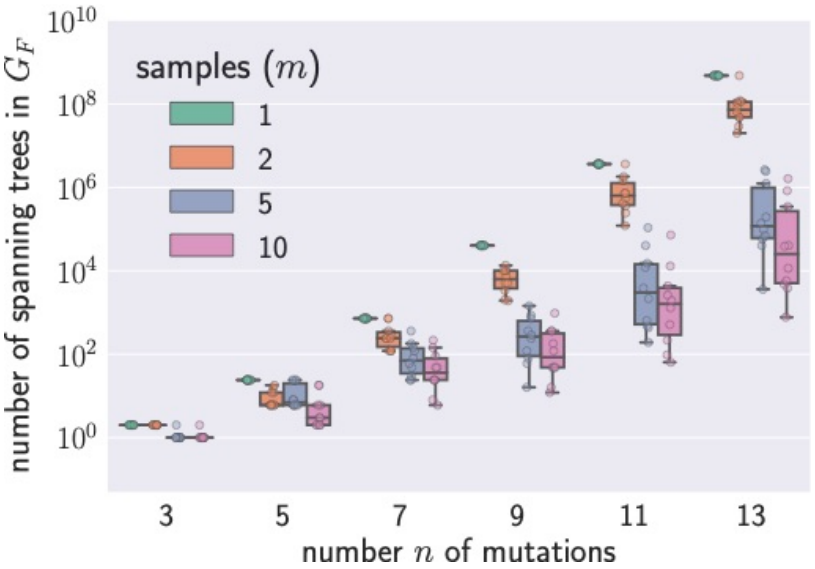


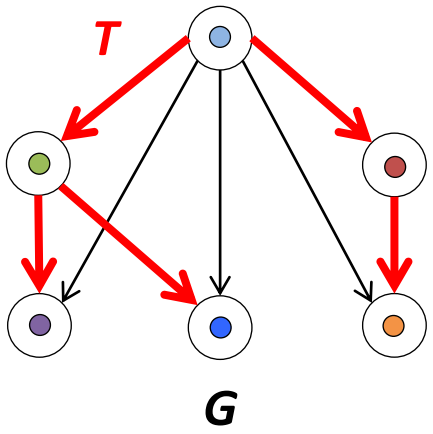




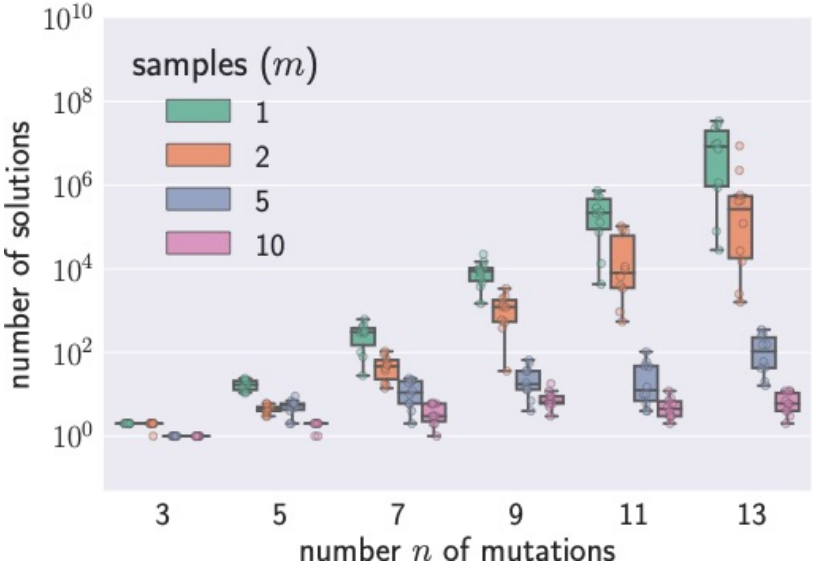
An Upper Bound for Number of Solutions

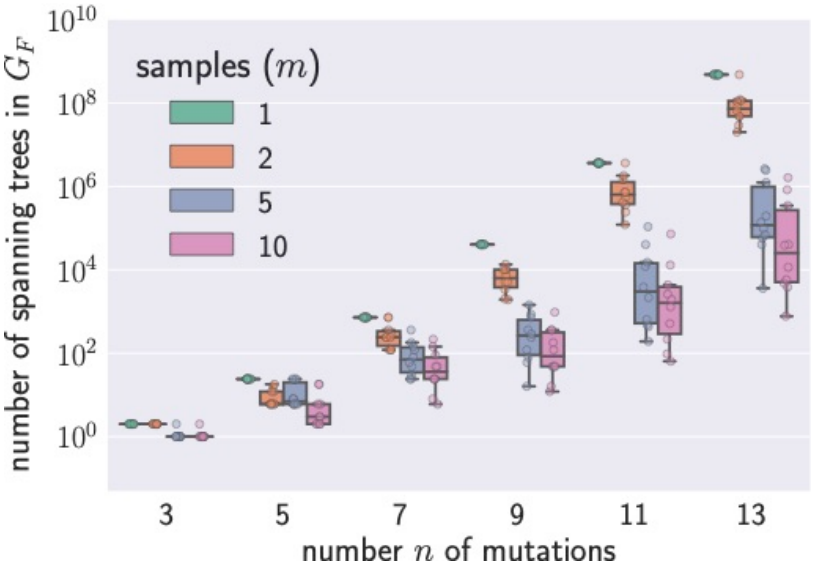


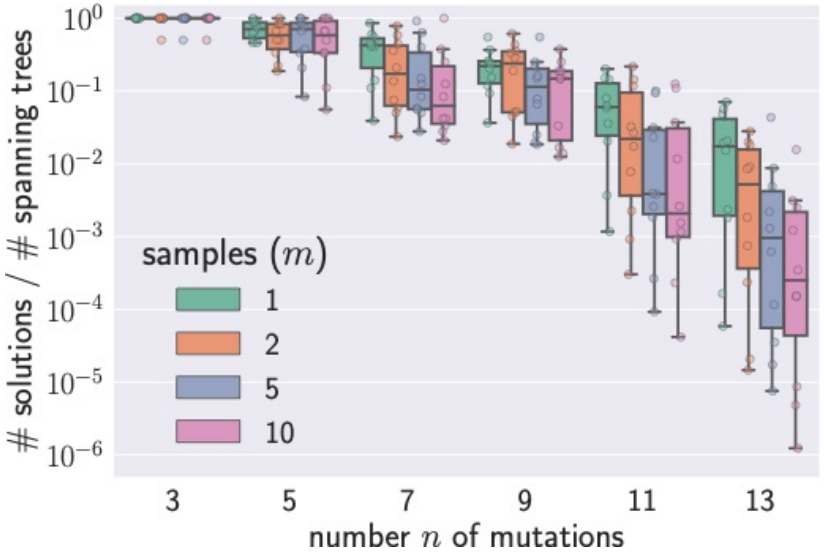


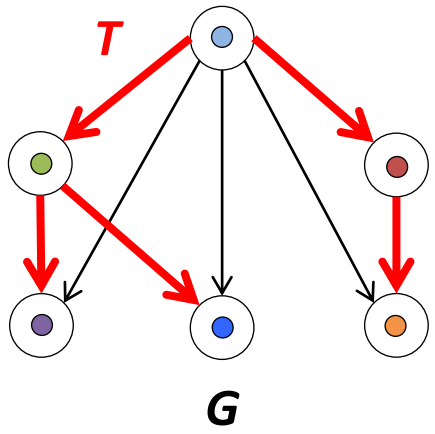


An Upper Bound for Number of Solutions





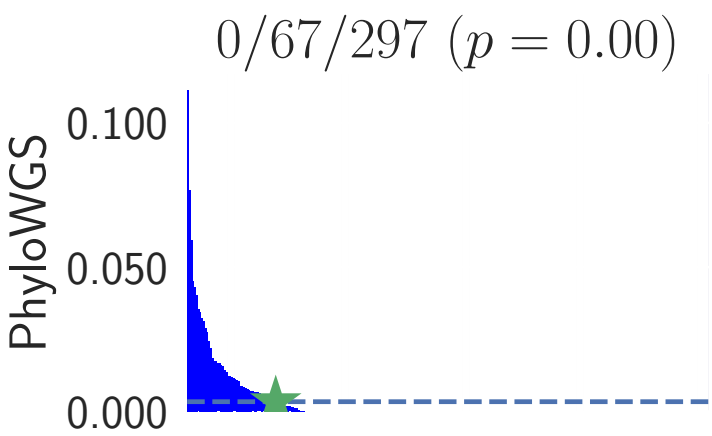


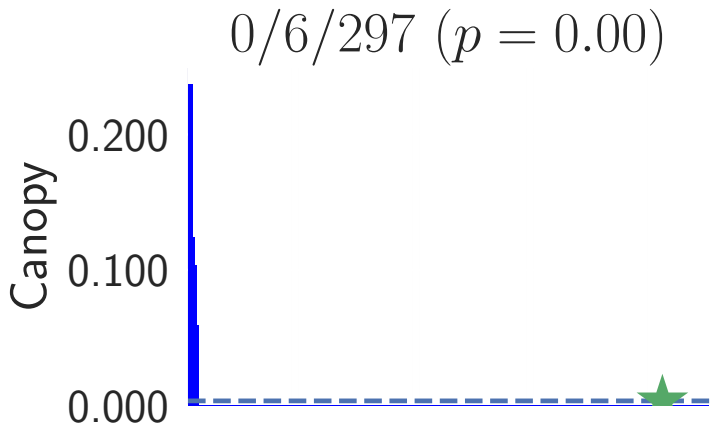


How Does Non-uniqueness affect Methods?

Two current MCMC methods using default parameters:

- PhyloWGS, Deshwar et al., Genom. Biol., 2015 [10,000 samples]
- Canopy, Jiang et al., PNAS, 2016 [~300 samples]

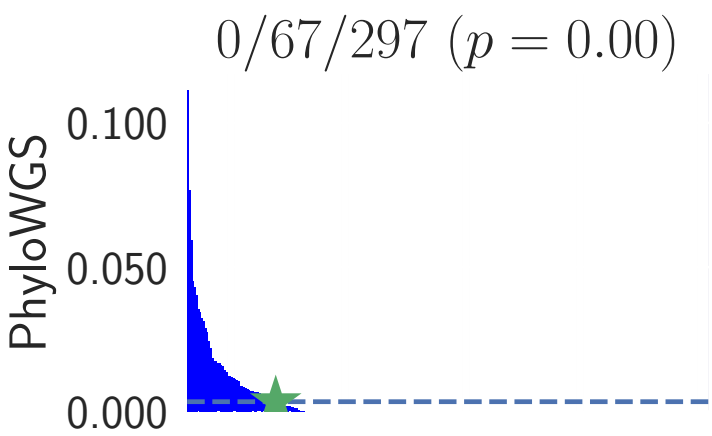


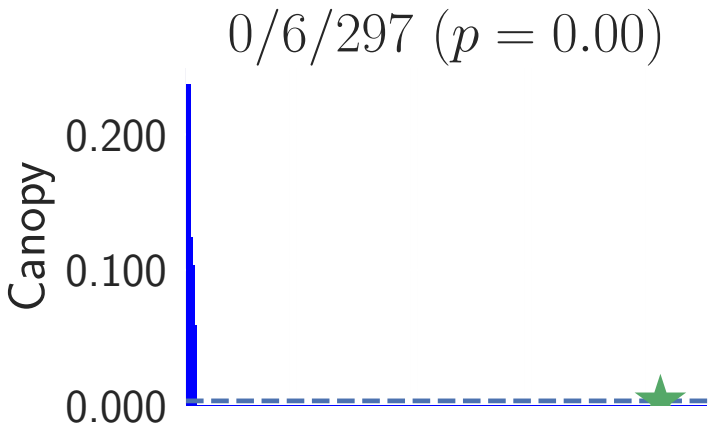


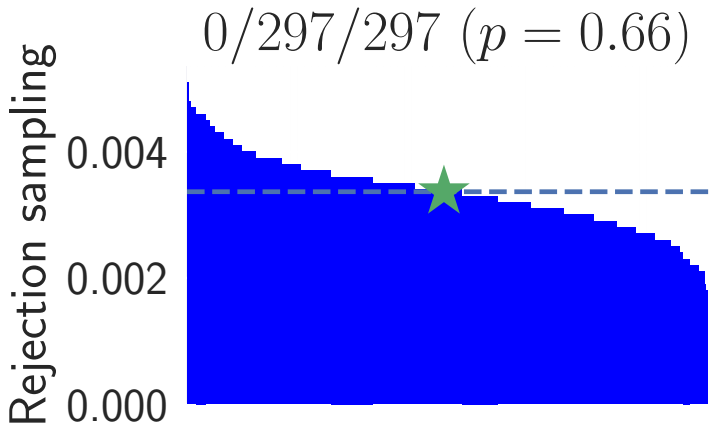
How Does Non-uniqueness affect Methods?

Two current MCMC methods using default parameters:

- PhyloWGS, Deshwar et al., Genom. Biol., 2015 [10,000 samples]
- Canopy, Jiang et al., PNAS, 2016 [~300 samples]



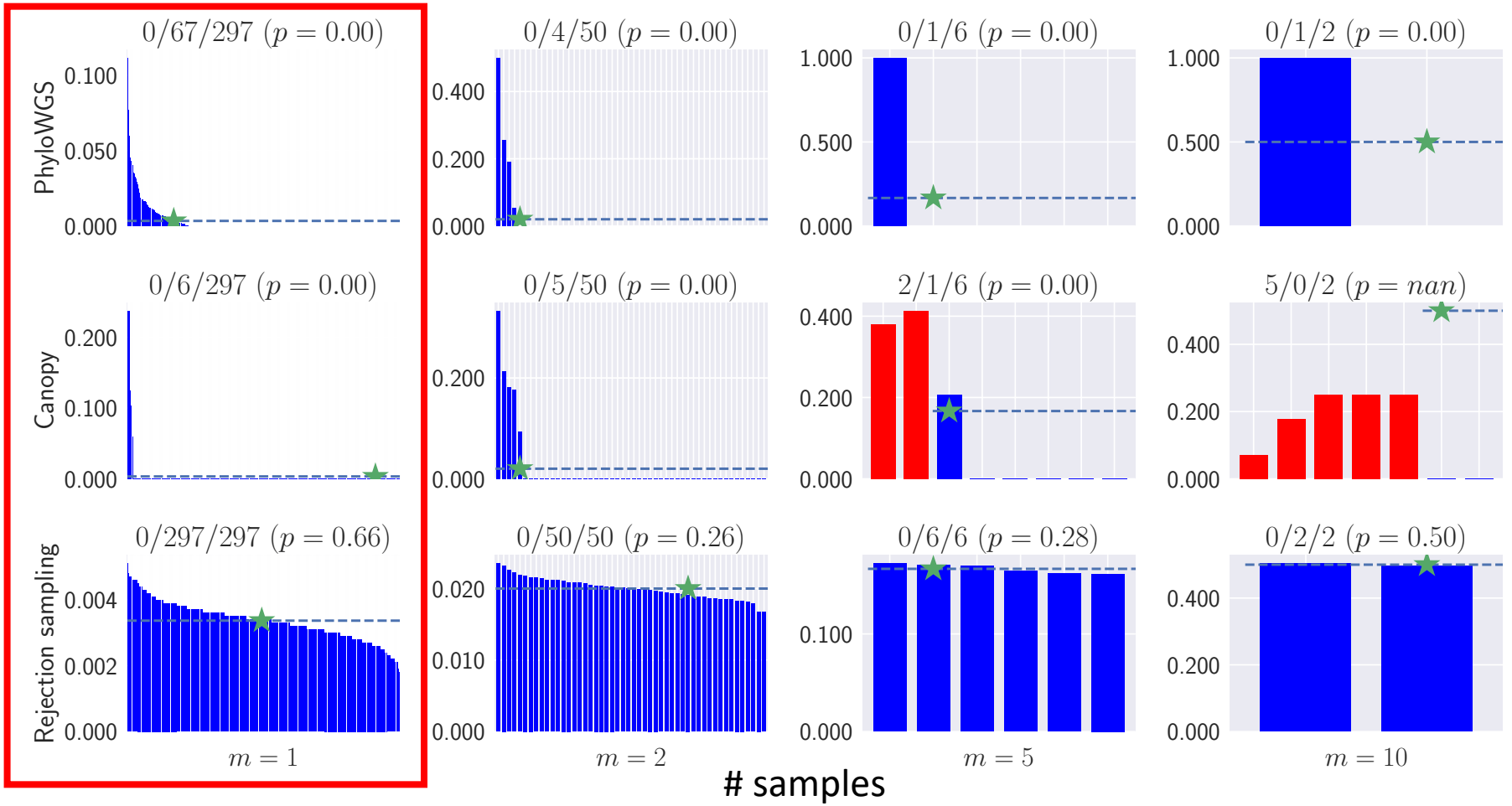




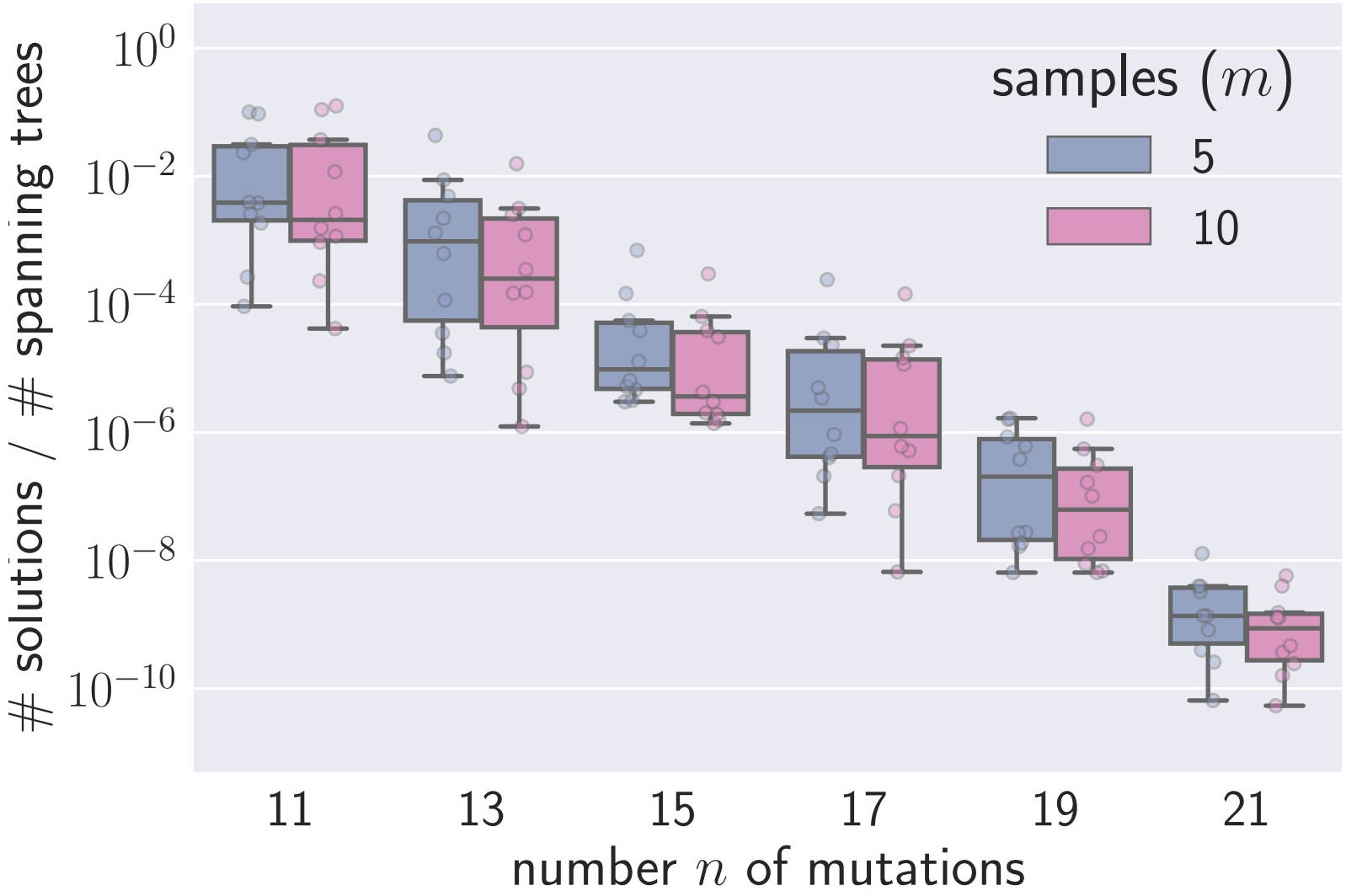
How Does Non-uniqueness affect Methods?

Two current MCMC methods using default parameters:

- PhyloWGS, Deshwar et al., Genom. Biol., 2015 [10,000 samples]
- Canopy, Jiang et al., PNAS, 2016 [~300 samples]



Rejection Sampling Does Not Scale



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. Almost uniform sampling: [To be submitted]

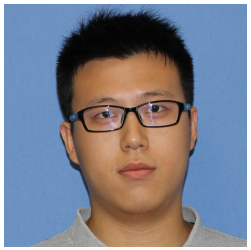
- Reducing PPM to SATISFIABILITY

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

- Multiple consensus tree problem

5. Applications

- Mutational signature dynamics [PSB 2020]

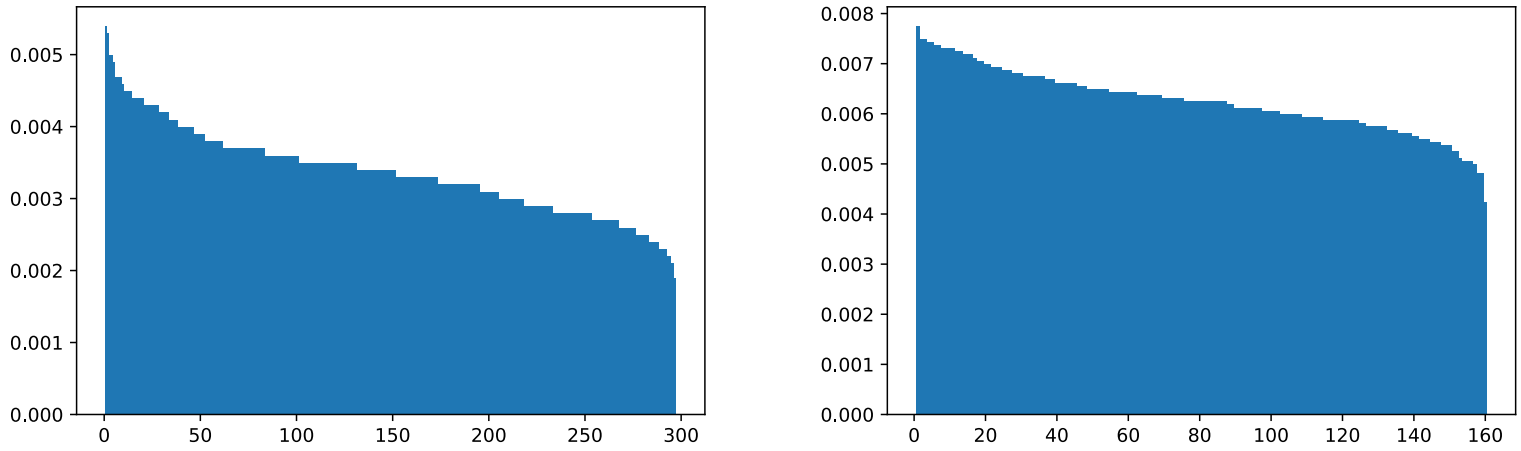


Yuanyuan Qi

Result

Generating 10,000 trees from the solution space almost uniformly at random

3 seconds **4 minutes**



(a) (b)

Figure 4: Out method samples uniformly:
(a) the same simulated instance (m=1, n=7, 297 solution trees);
(b) the ranged version of PPM of CRUK0062 (90% confidence interval, m=7, n=15, 160 solutions)

Method

Unigen

An almost uniform sampler for SAT (satisfiability) problem[4].

Reduce PPM to SAT.

[4] Chakraborty, S., Fremont, D. J., Meel, K. S., Seshia, S. A., & Vardi, M. Y. (2015, April). On parallel scalable uniform SAT witness generation.

Boolean Satisfiability

- Goal: find a model that satisfies a propositional formula.

$a \wedge (\neg b \vee c) \longrightarrow a \mapsto T, b \mapsto F, c \mapsto F$

$a \wedge b \wedge (\neg b \vee \neg a) \longrightarrow$ unsatisfiable

- The original NP-complete problem.
  - "As hard as any other problem in NP."
- S. Cook, *The complexity of theorem proving procedures*, STOC 1971.

Method

Spanning tree of ancestry graph

Consider simple case:

- F is precise
- No repeated columns (G_F is a DAG)

Variables:

- r_p: $r_p = 1$ iff p is the root.
- $e_{p,q}$: $e_{p,q} = 1$ iff p is the parent of q.

Restrictions:

- only one of r_p for all $p \in [n]$ is true.
- for all $p \in [n]$, only one of $e_{q,p} \forall q \in [n]$ and r_p is true.

Method

Clauses for restriction "only one of $v_1, \ldots, v_n$ is true":

- $\bigvee_{p \in [n]} v_p$
- $\neg v_p \vee \neg v_q$, for all $p \neq q$

Method

Sum Condition

- Discretize the frequencies (multiply by $(2^N - 1)$).
- Represent the integers with a vector of binary vriables.

We need to represent addition and comparison in CNF form (clauses).

Method

Addition

half_adder(a, b, r, c):

- $r = a \oplus b$:
  - $r \vee \neg a \vee b$
  - $r \vee a \vee \neg b$
  - $\neg r \vee a \vee b$
  - $\neg r \vee \neg a \vee \neg b$
- $c = a \vee b$:
  - $\neg c \vee a \vee b$
  - $c \vee \neg a$
  - $c \vee \neg b$

full_adder($a, b, last\_c, r, c$)

Comparison

leq(a,b): check the carry of $\sim a + b + 1$.

Method

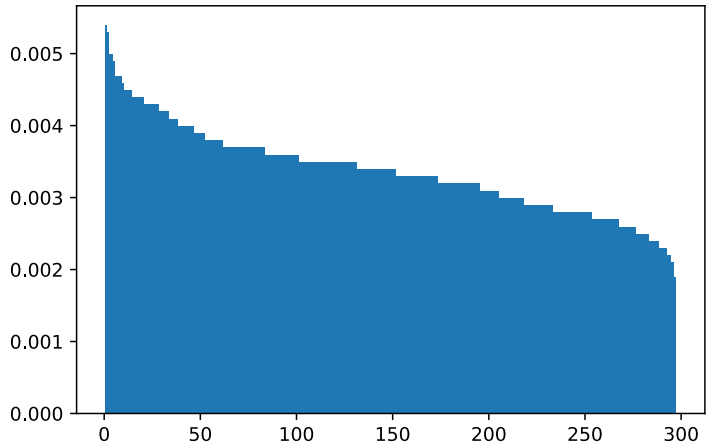
More complicated case: given F^- and F^+ instead of a single F.

Approach

- Cycle prevention: $depth(p) = depth(q) + 1$ if $e_{q,p} = 1$.
- F is not determined: use the minimum frequency possible.

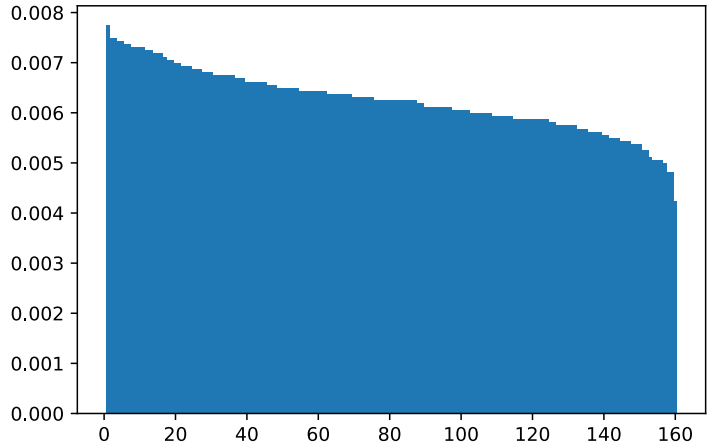
Result

$5m \cdot |E| \cdot N$ variables
$17m \cdot |E| \cdot N$ clauses **3 seconds**



(a)

$6m \cdot |E| \cdot N$ variables
$20m \cdot |E| \cdot N$ clauses **4 minutes**



(b)

Figure 4: Out method samples uniformly:
(a) the same simulated instance (m=1, n=7, 297 solution trees);
(b) the ranged version of PPM of CRUK0062 (90% confidence interval, m=7, n=15, 160 solutions)

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]

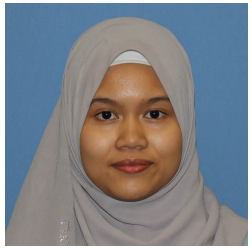
- What contributes to non-uniqueness?
- How to reduce non-uniqueness?
- How does non-uniqueness affect current methods?

3. Almost uniform sampling: [To be submitted]

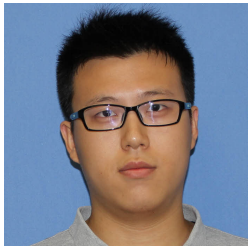
- Reducing PPM to SATISFIABILITY

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

- Multiple consensus tree problem



Nuraini Aguse



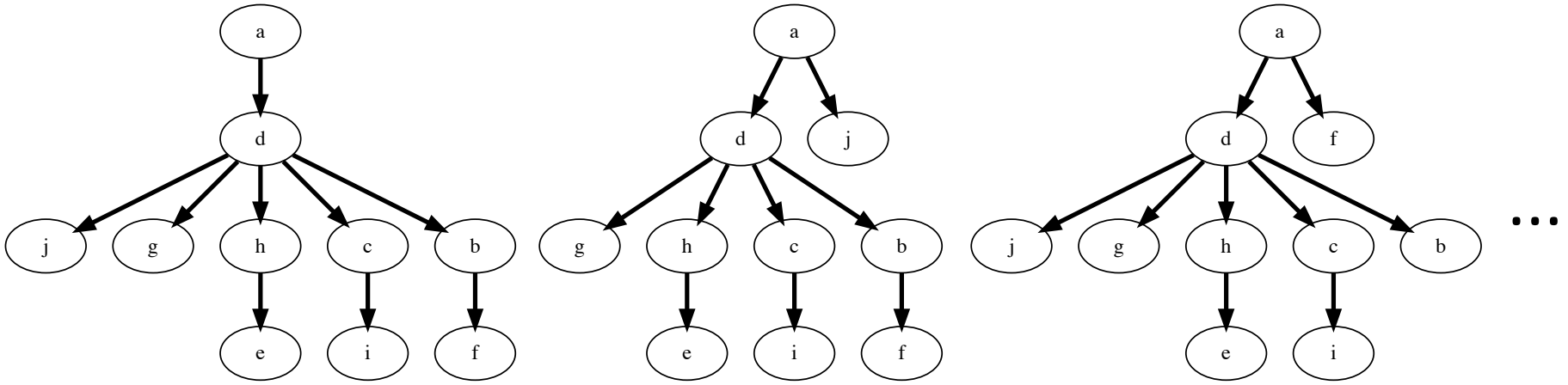
Yuanyuan Qi

5. Applications

- Visualizing spatial clonal architecture of tumors: [To be submitted]
- Mutational signature dynamics [PSB 2020]

Lung Cancer Patient: CRUK0037

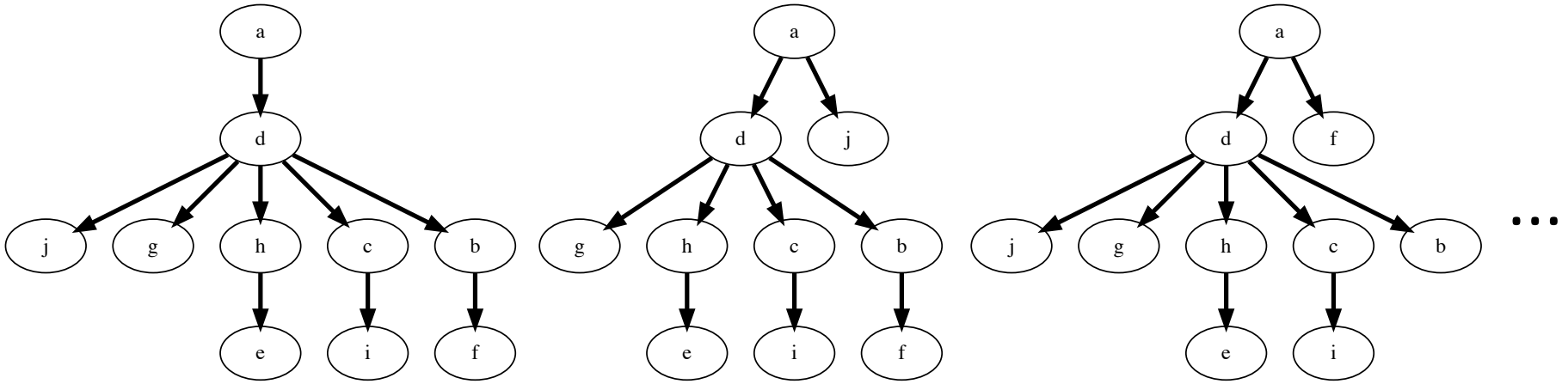
Jamal-Hanjani et al. (2017). *New England Journal of Medicine*, *376*(22), 2109–2121.



Authors inferred 17 trees

Lung Cancer Patient: CRUK0037

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

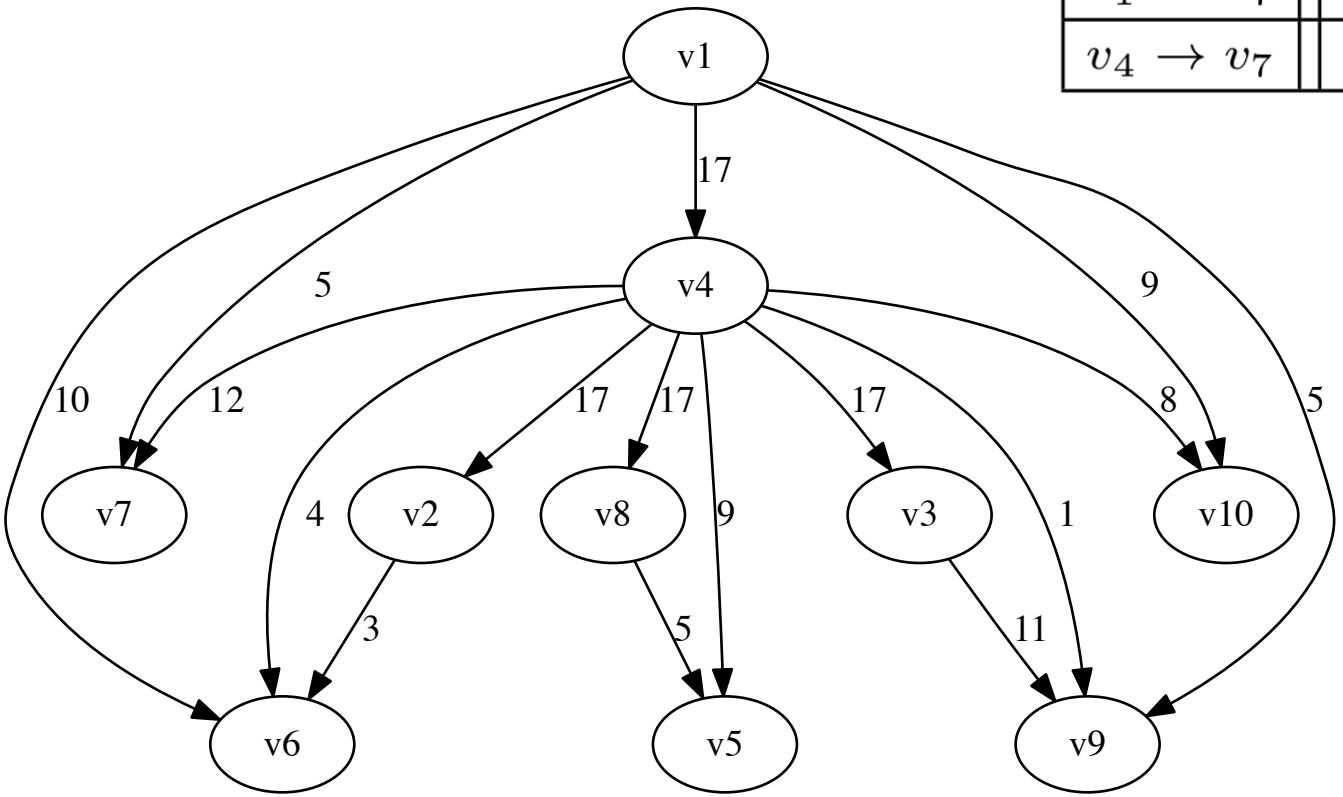


Authors inferred 17 trees

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

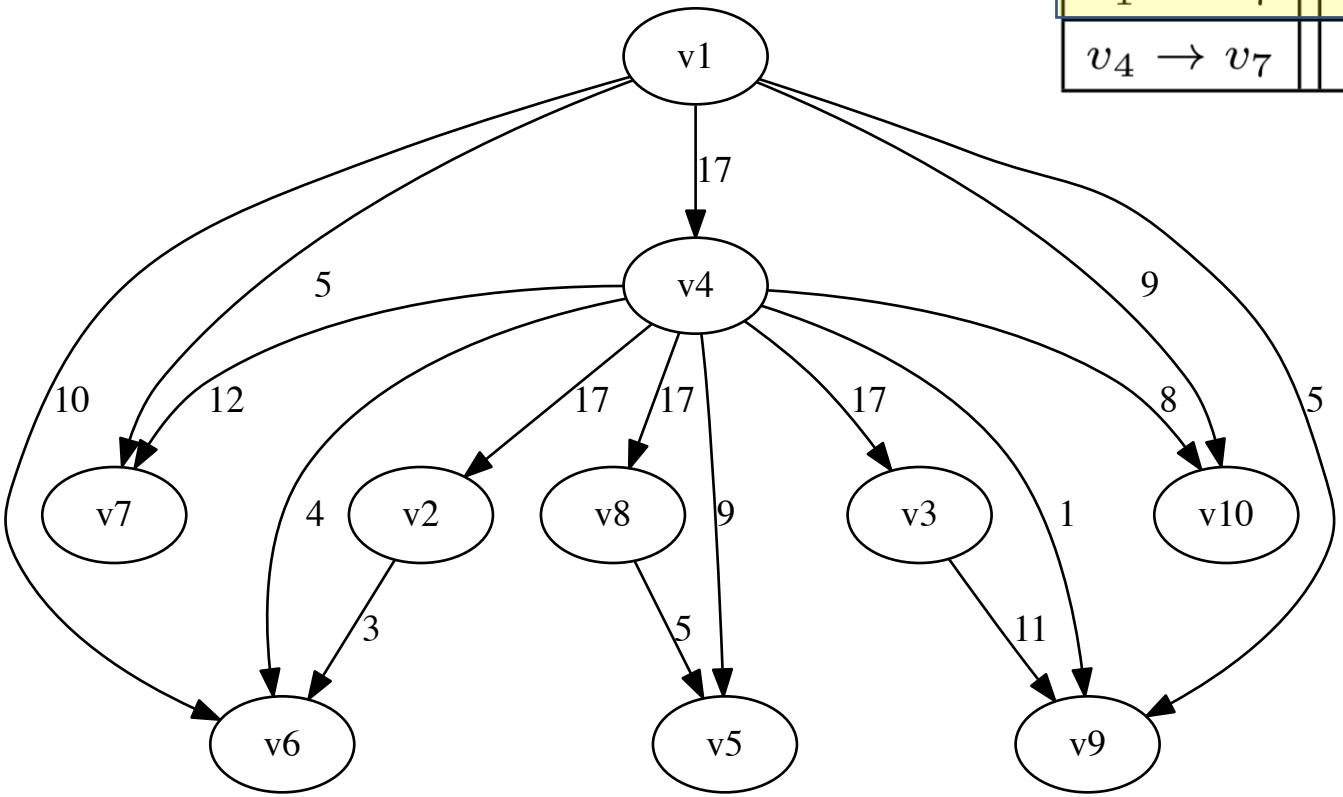
Parent-child Graph: Union of all Edges

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



Parent-child Graph: Union of all Edges

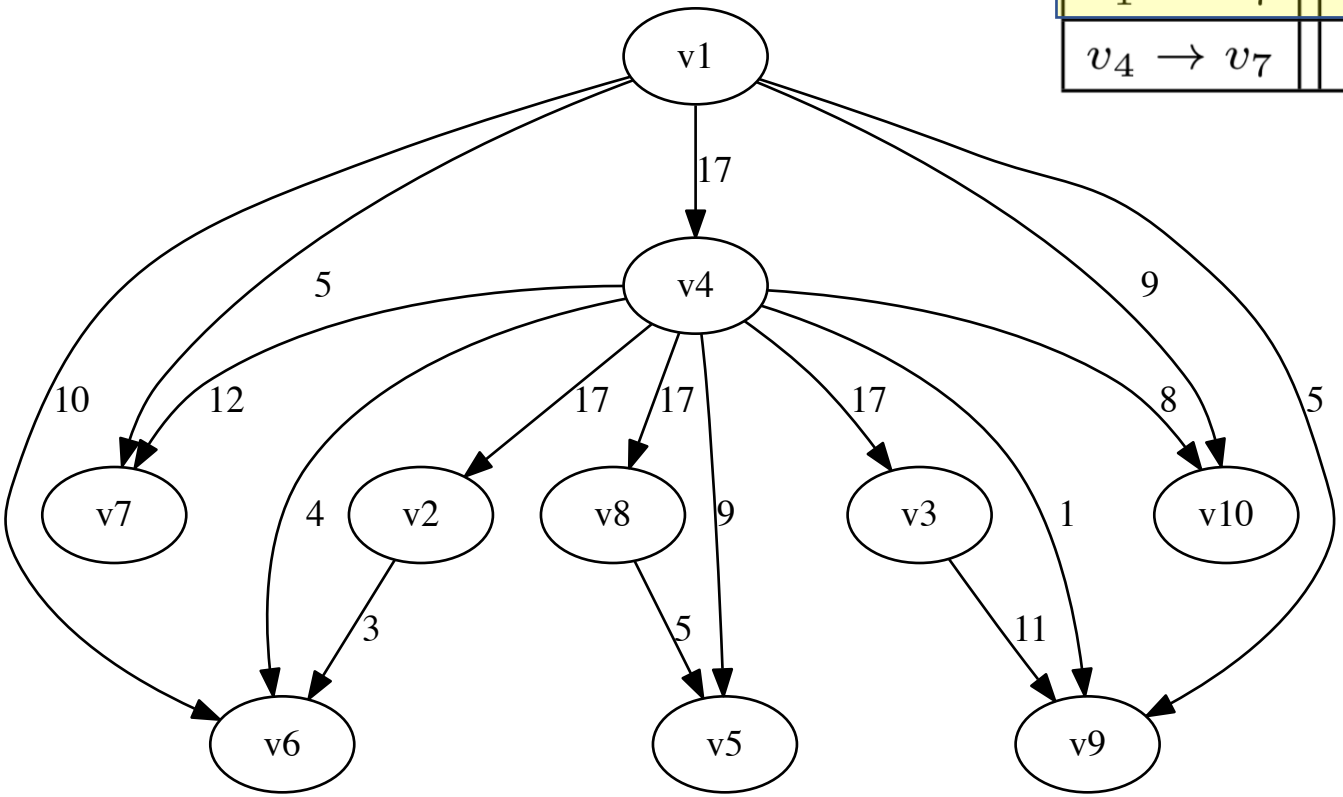
	$v_1 \rightarrow v_{10}$	$v_4 \rightarrow v_{10}$
$v_1 \rightarrow v_7$	2	0
$v_4 \rightarrow v_7$	2	5



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	0
$v_4 \rightarrow v_7$	2	5



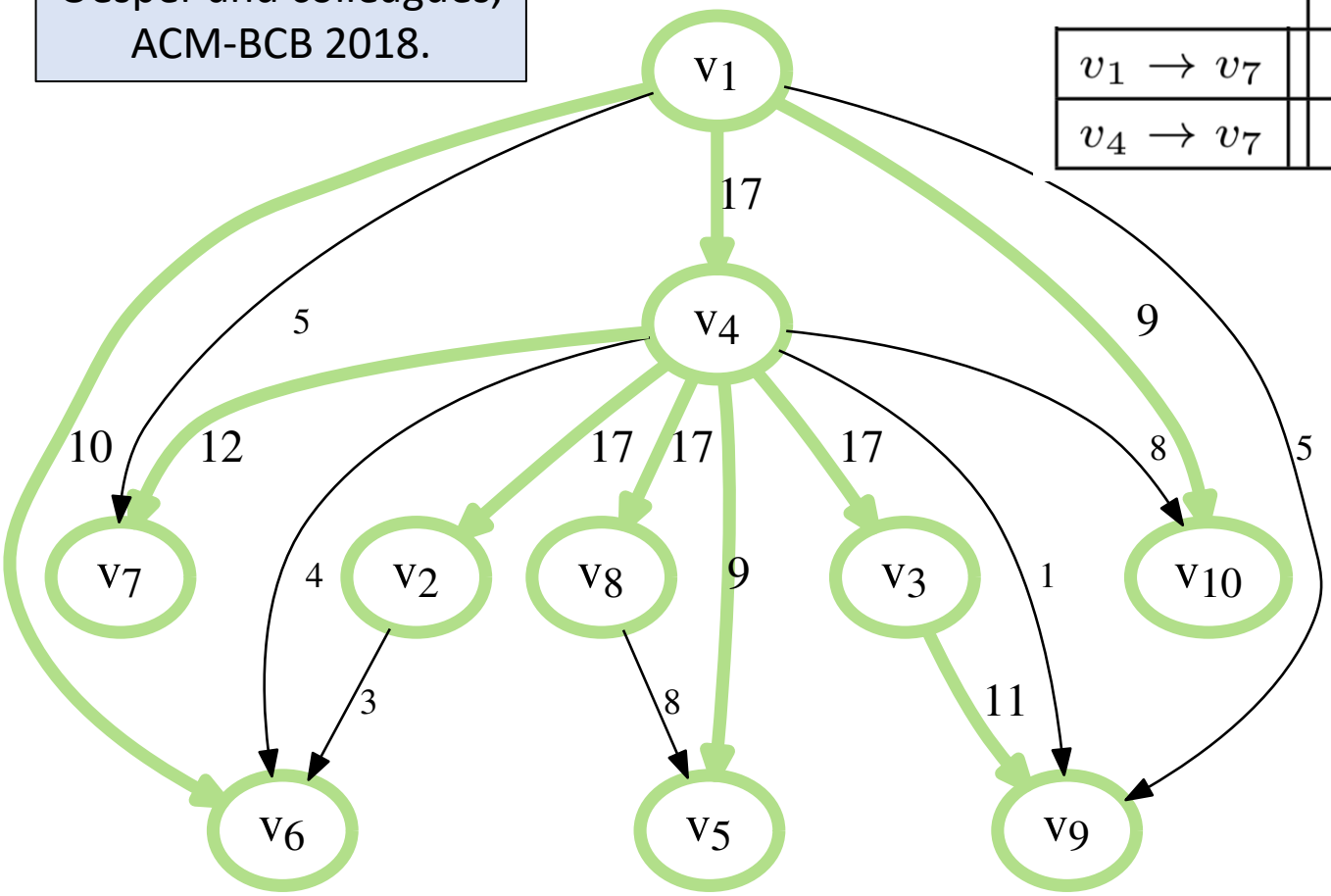
The parent-child graph does capture patterns of mutual exclusivity

Question: Can we infer a single consensus tree?

Single Consensus Tree: Max Weight Spanning Tree

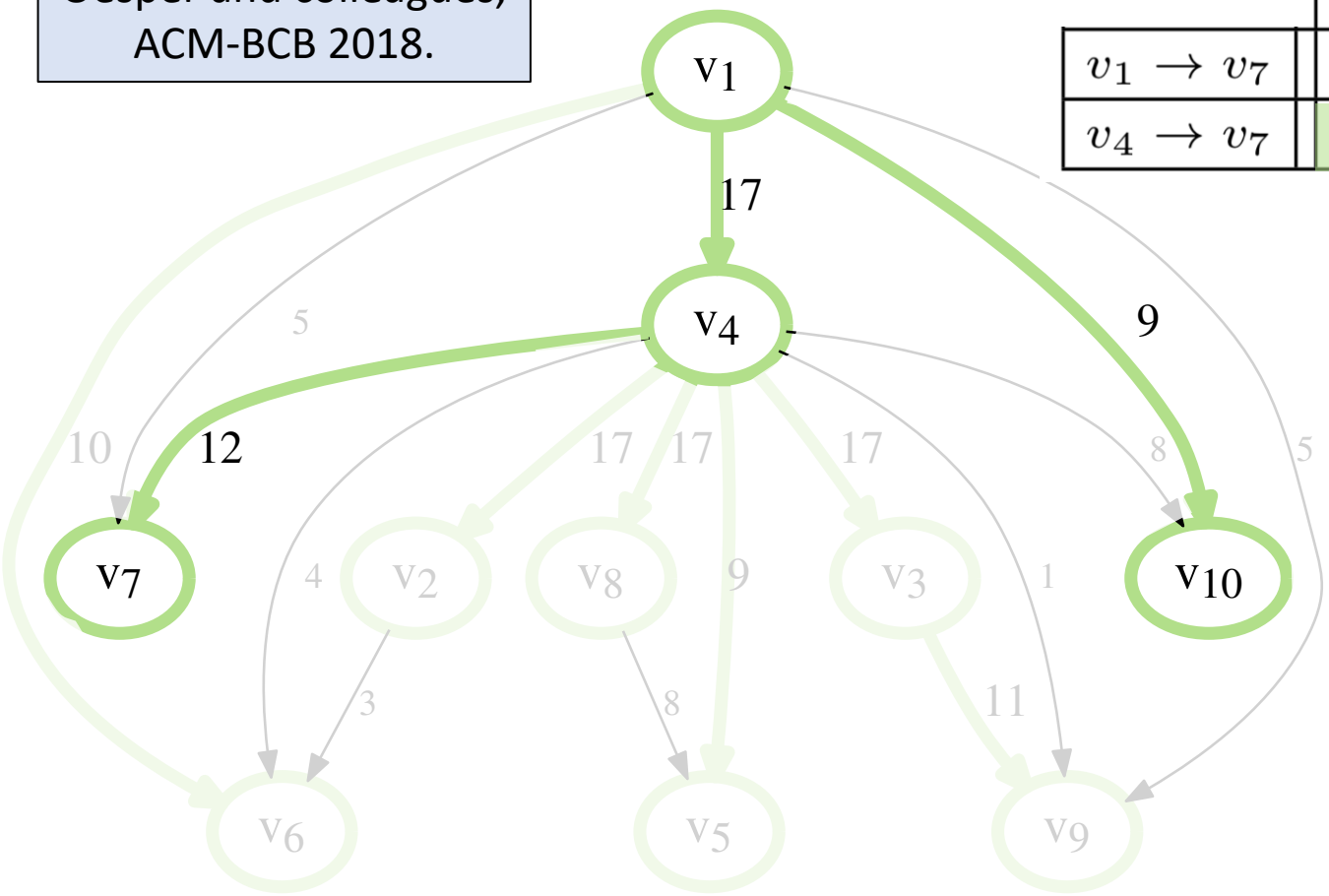
Oesper and colleagues, ACM-BCB 2018.

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



Single Consensus Tree: Max Weight Spanning Tree

Oesper and colleagues, ACM-BCB 2018.



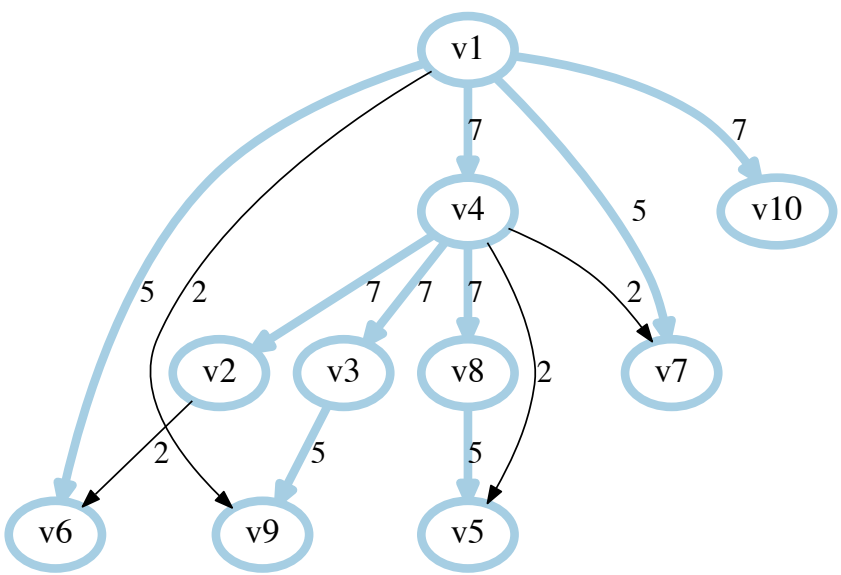
	$v_1 \rightarrow v_{10}$	$v_4 \rightarrow v_{10}$
$v_1 \rightarrow v_7$	2	0
$v_4 \rightarrow v_7$	2	5

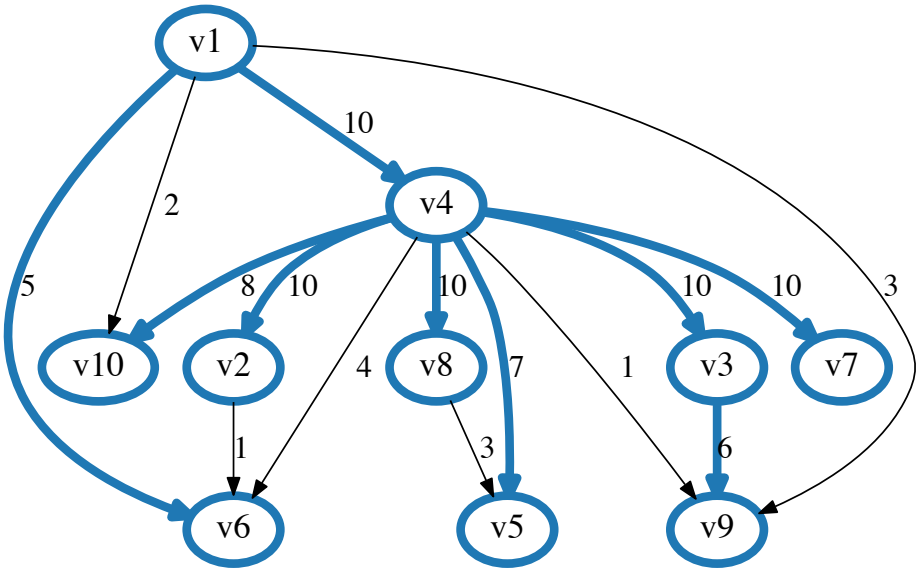
Inaccurate summary for diverse solution spaces

Question: How about inferring multiple consensus trees?

Multiple Consensus Trees

Simultaneous clustering and consensus tree inference



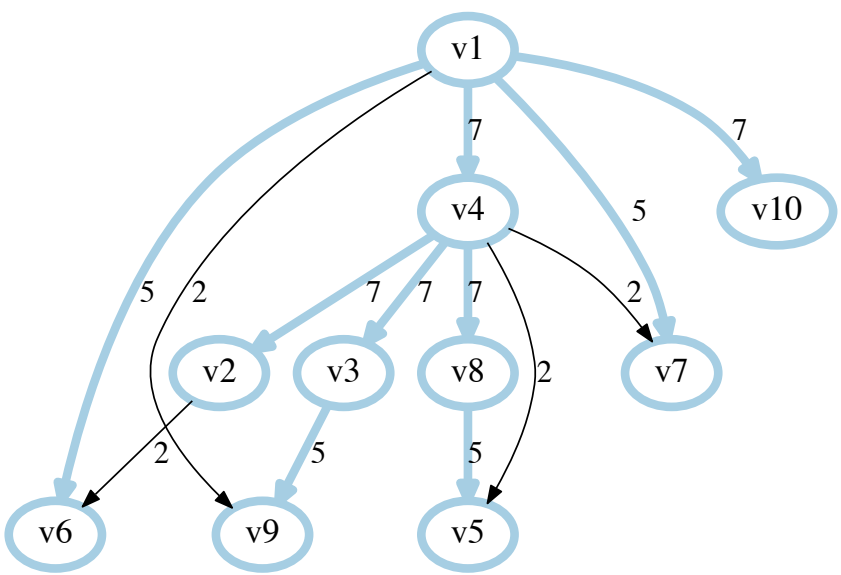


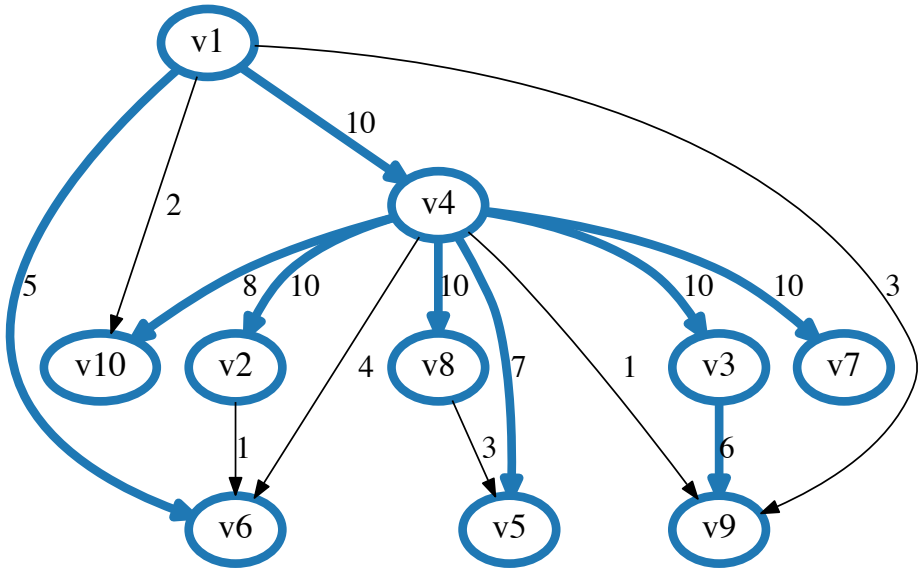
Multiple Consensus Trees (MCT): [ISMB 2019]

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

Multiple Consensus Trees

Simultaneous clustering and consensus tree inference

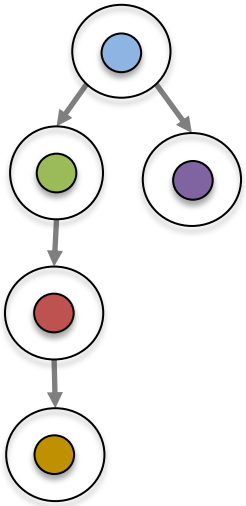




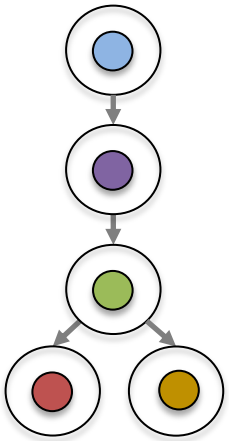
Multiple Consensus Trees (MCT): [ISMB 2019]

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

Parent-child Distance Function

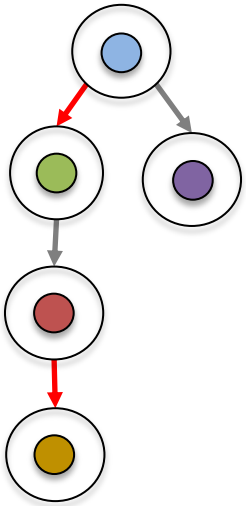


T_1

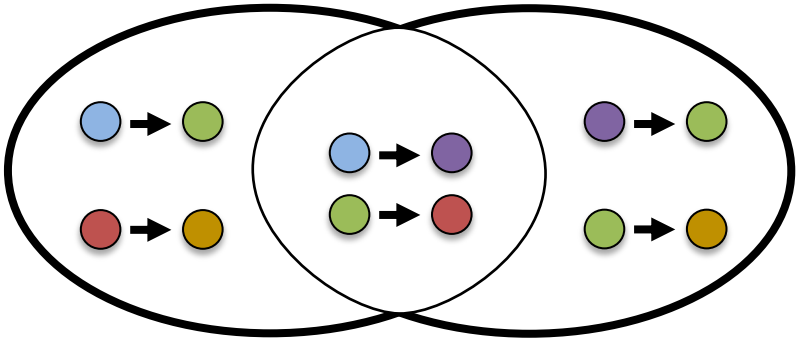


T_2

Parent-child Distance Function



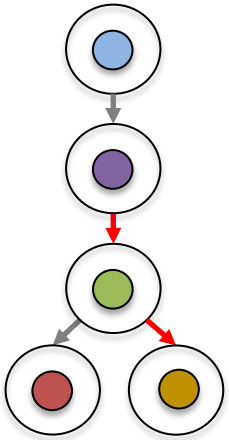
T_1



$E(T_1) \cap E(T_2)$

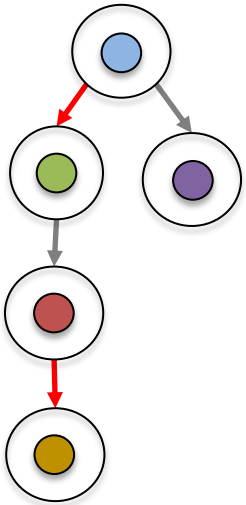
$E(T_1) \backslash E(T_2)$

$E(T_2) \backslash E(T_1)$

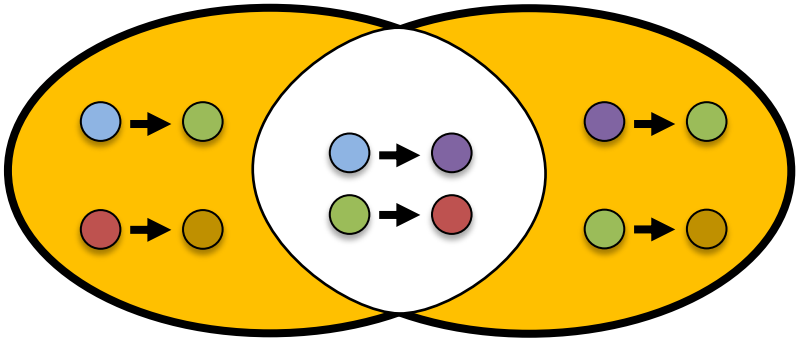


T_2

Parent-child Distance Function



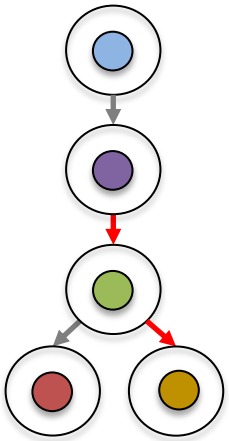
T_1



$E(T_1) \cap E(T_2)$

$E(T_1) \backslash E(T_2)$

$E(T_2) \backslash E(T_1)$



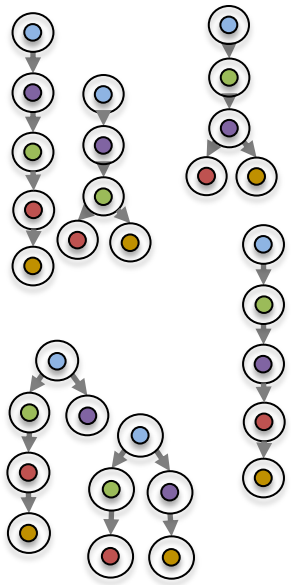
T_2

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

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

Combinatorial Characterization of Solutions

Single Consensus Trees (SCT):
[Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \ldots, T_n\}$, find consensus tree R
s.t. $\sum_{i=1}^{n} d(T_i, R)$ is minimum

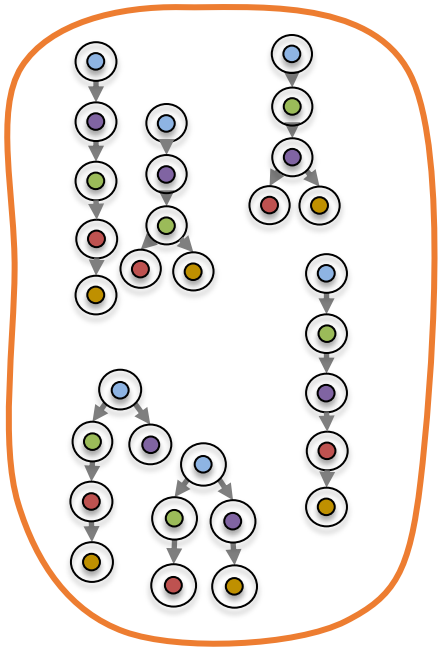


Solution Space $\mathcal{T}$

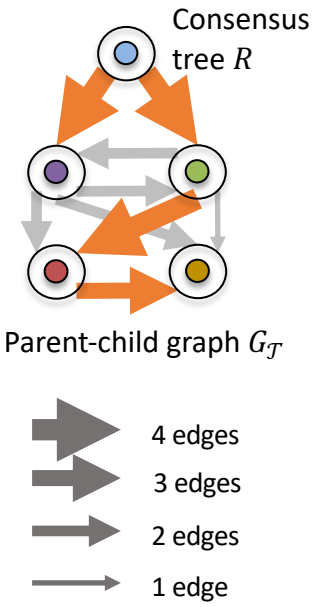
Combinatorial Characterization of Solutions

Single Consensus Trees (SCT):
[Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \ldots, T_n\}$, 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



Solution Space $\mathcal{T}$

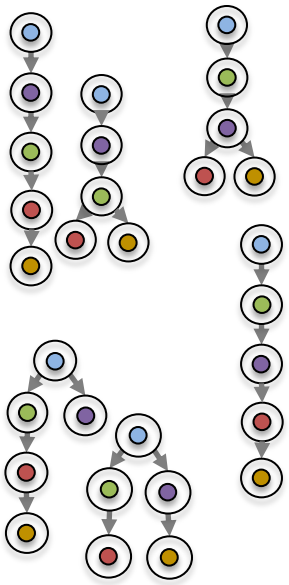


Combinatorial Characterization of Solutions

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 $\sigma : [n] \to [k]$ 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 $\mathcal{T}$

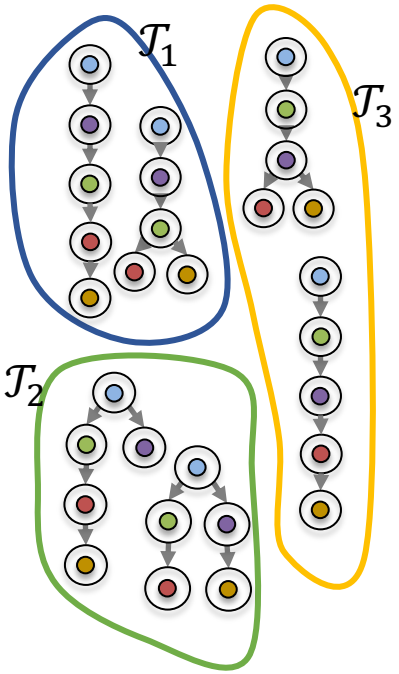
Combinatorial Characterization of Solutions

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 $\sigma : [n] \to [k]$ 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 $\sigma : [n] \to [k]$, MCT decomposes into k independent SCT instances



Solution Space $\mathcal{T}$

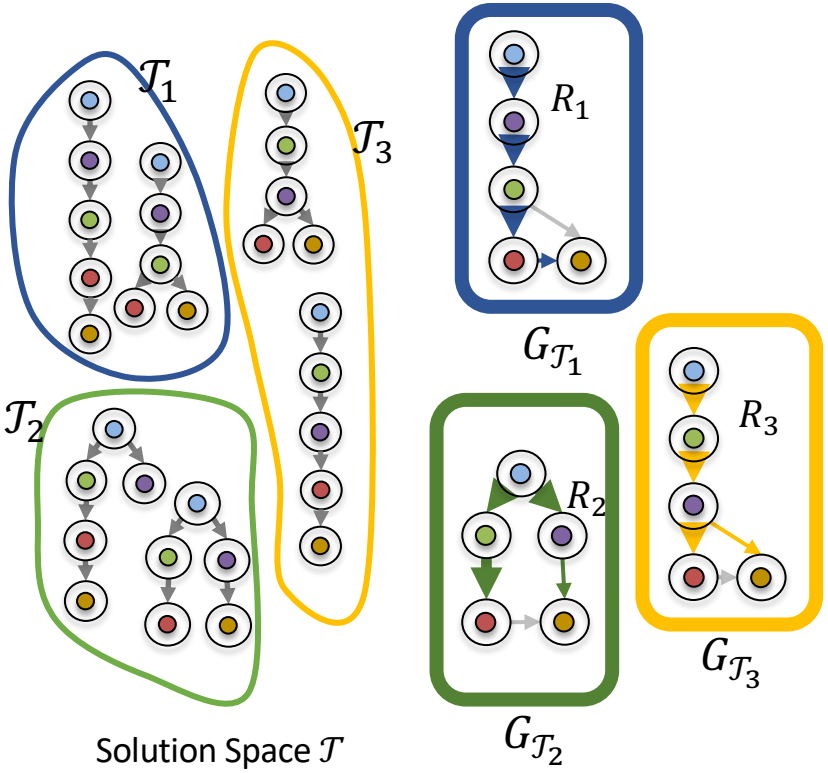
Combinatorial Characterization of Solutions

Single Consensus Trees (SCT):
[Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \ldots, T_n\}$, 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 $\mathcal{T} = \{T_1, \ldots, T_n\}$ and $k > 0$, find surjective clustering $\sigma : [n] \rightarrow [k]$ ~~and consensus trees $\mathcal{R} = \{R_1, \ldots, R_k\}$~~ 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



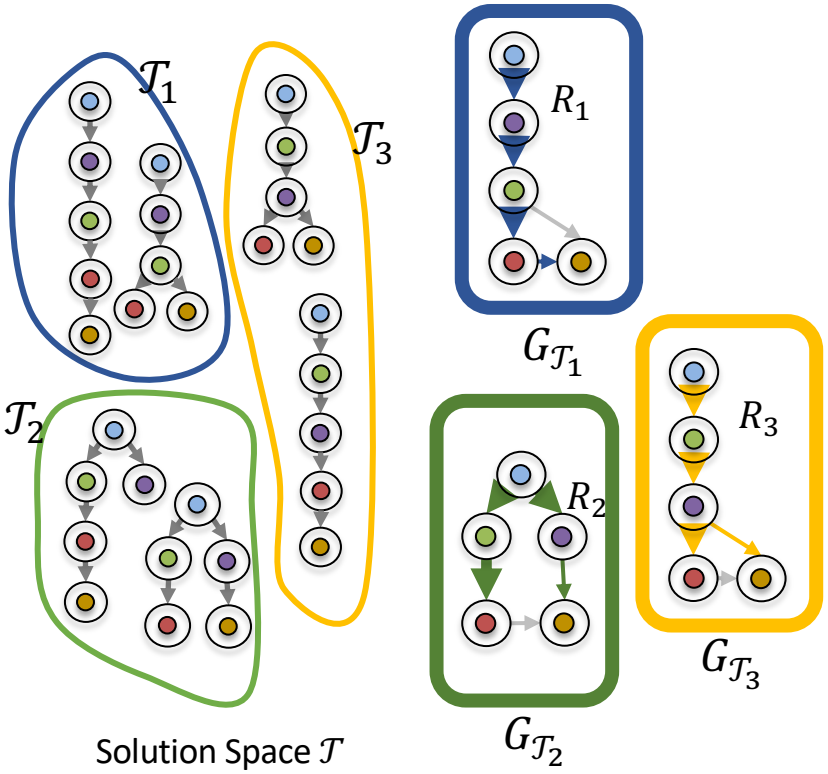
Combinatorial Characterization of Solutions

Single Consensus Trees (SCT):
[Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \ldots, T_n\}$, 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 $\mathcal{T} = \{T_1, \ldots, T_n\}$ and $k > 0$, find surjective clustering $\sigma : [n] \rightarrow [k]$ ~~and consensus trees $\mathcal{R} = \{R_1, \ldots, R_k\}$~~ 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

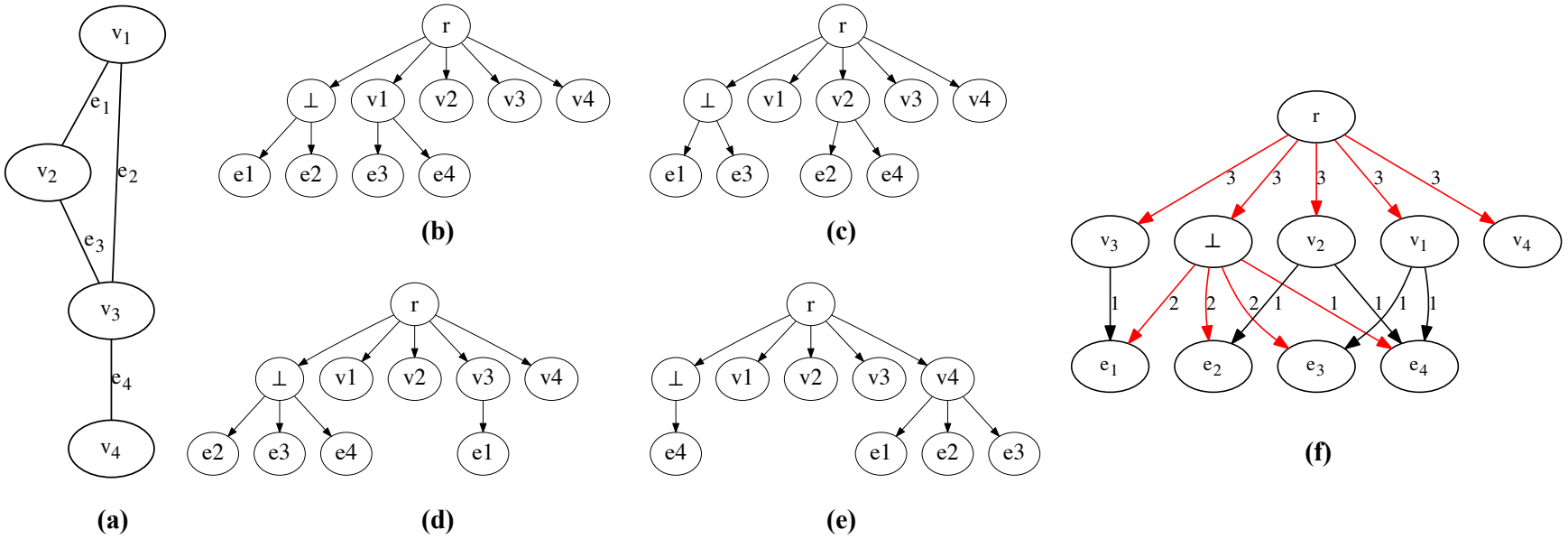


Question: How to find σ^*?

Complexity

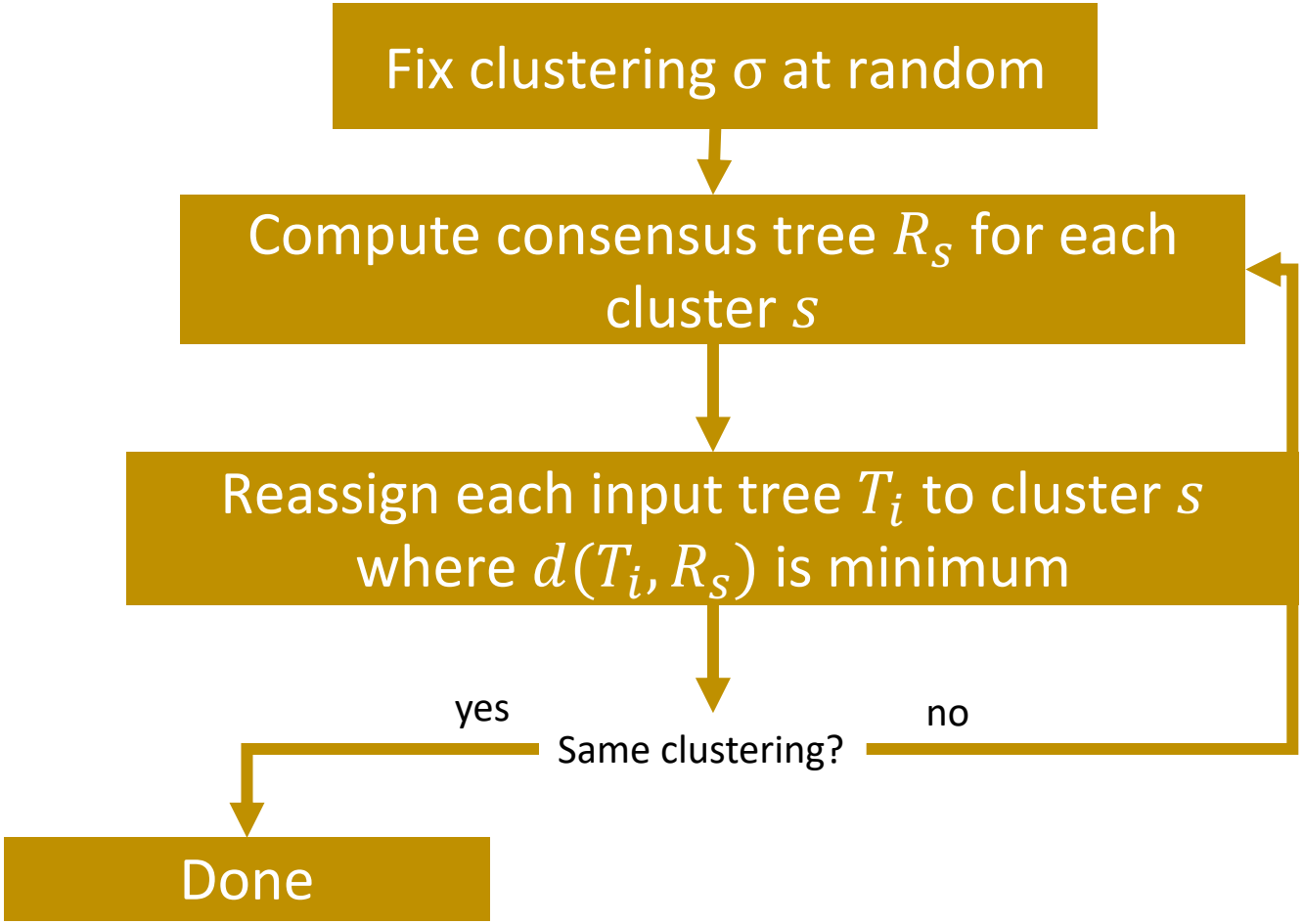
Multiple Consensus Trees (MCT):

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

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

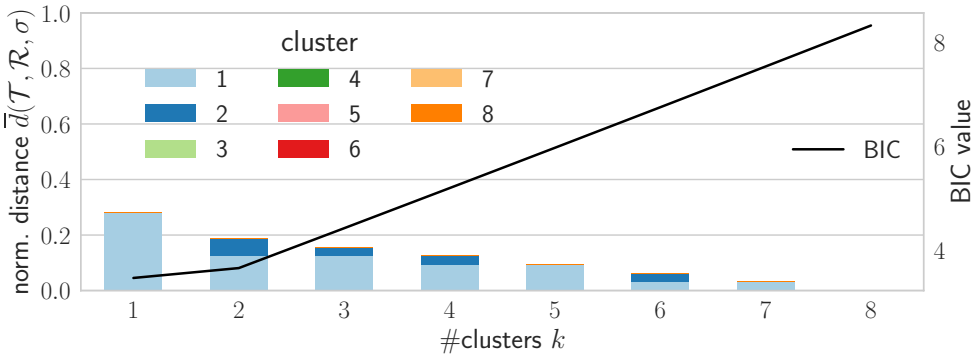
Number of instances where heuristic returned MILP's optimal solution

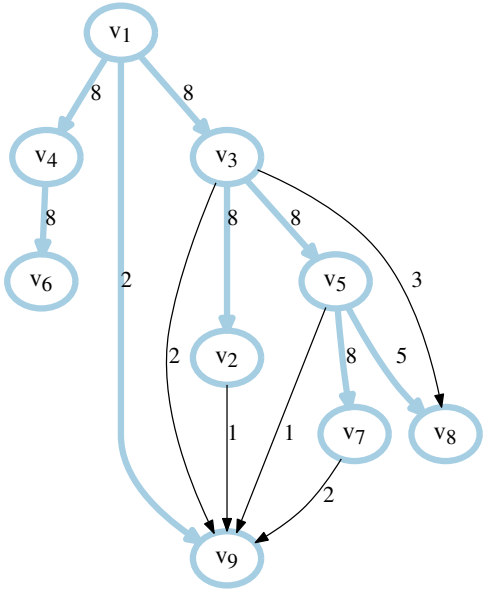
Question: How to determine k?

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

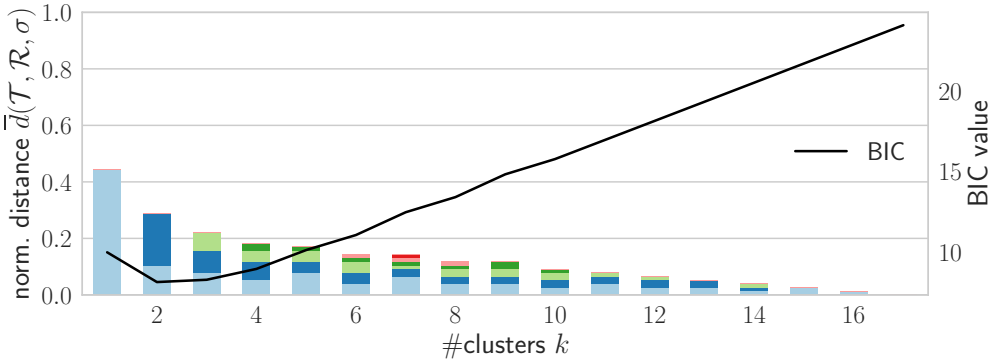


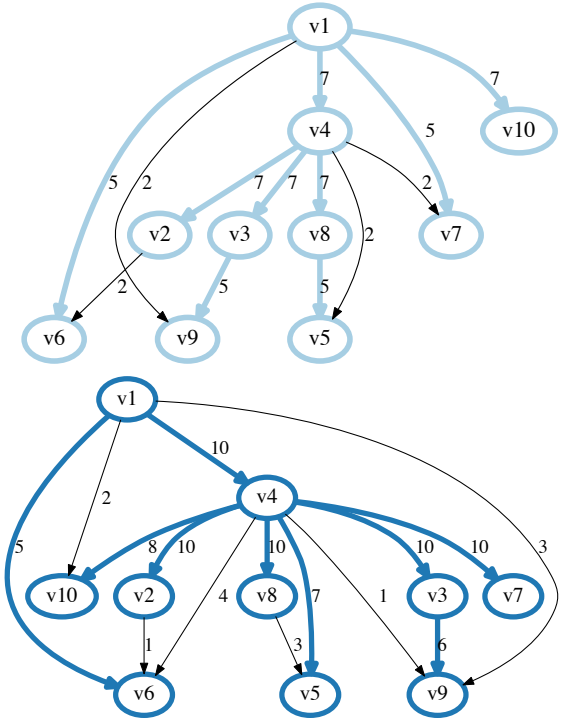


Bayesian Information Criterion determines the number of clusters for each solution space

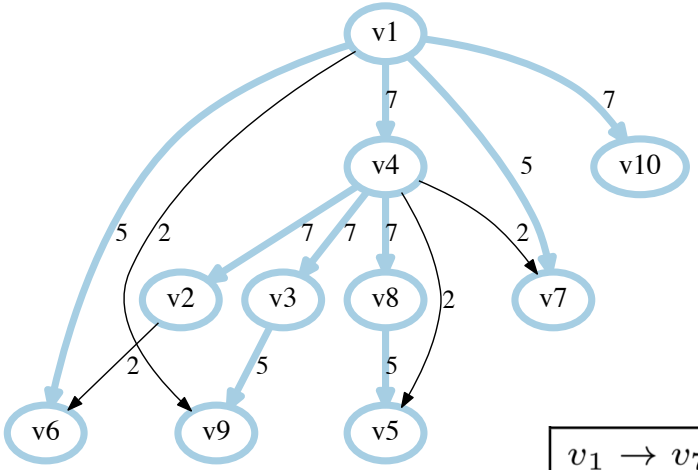
Jamal-Hanjani et al. (2017). *NEJM*.

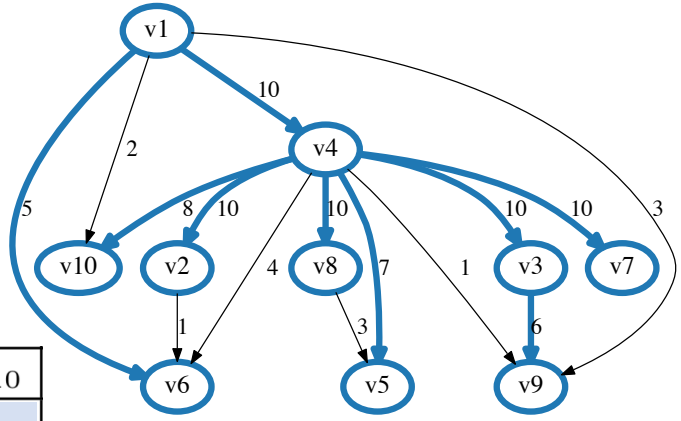
Jamal-Hanjani et al. inferred 17 trees for patient CRUK0037



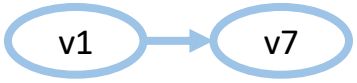


Multiple Consensus Trees capture patterns of mutual exclusivity and co-occurrence

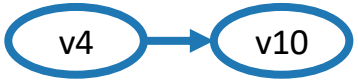




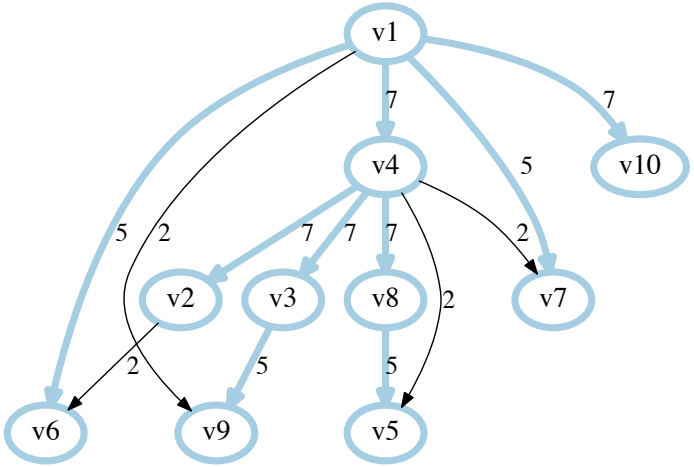
	$v_1 \rightarrow v_{10}$	$v_4 \rightarrow v_{10}$
$v_1 \rightarrow v_7$	2	0
$v_4 \rightarrow v_7$	2	5

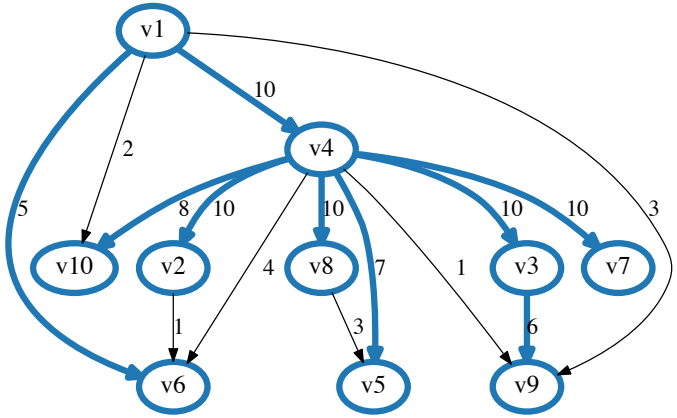


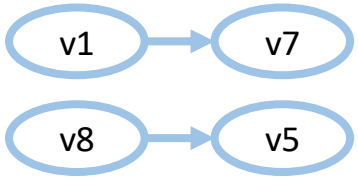
These edges are mutually exclusive



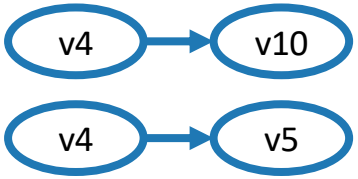
Multiple Consensus Trees capture patterns of mutual exclusivity and co-occurrence





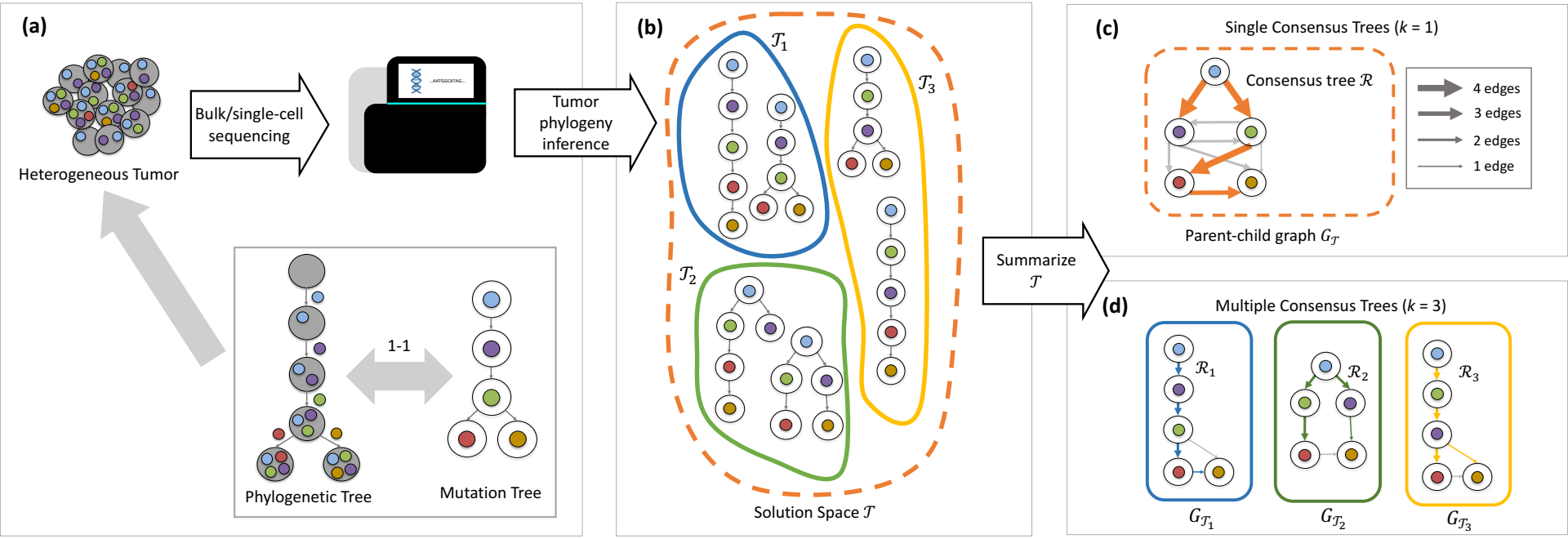


These edges tend to co-occur in the trees in the solution space



Multiple Consensus Trees (MCT): [ISMB 2019]

Given trees $\mathcal{T} = \{T_1, \ldots, T_n\}$, find surjective clustering $\sigma : [n] \to [k]$ and consensus trees $\mathcal{R} = \{R_1, \ldots, R_k\}$ 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?
- How does non-uniqueness affect current methods?

3. Almost uniform sampling: [To be submitted]

- Reducing PPM to SATISFIABILITY

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

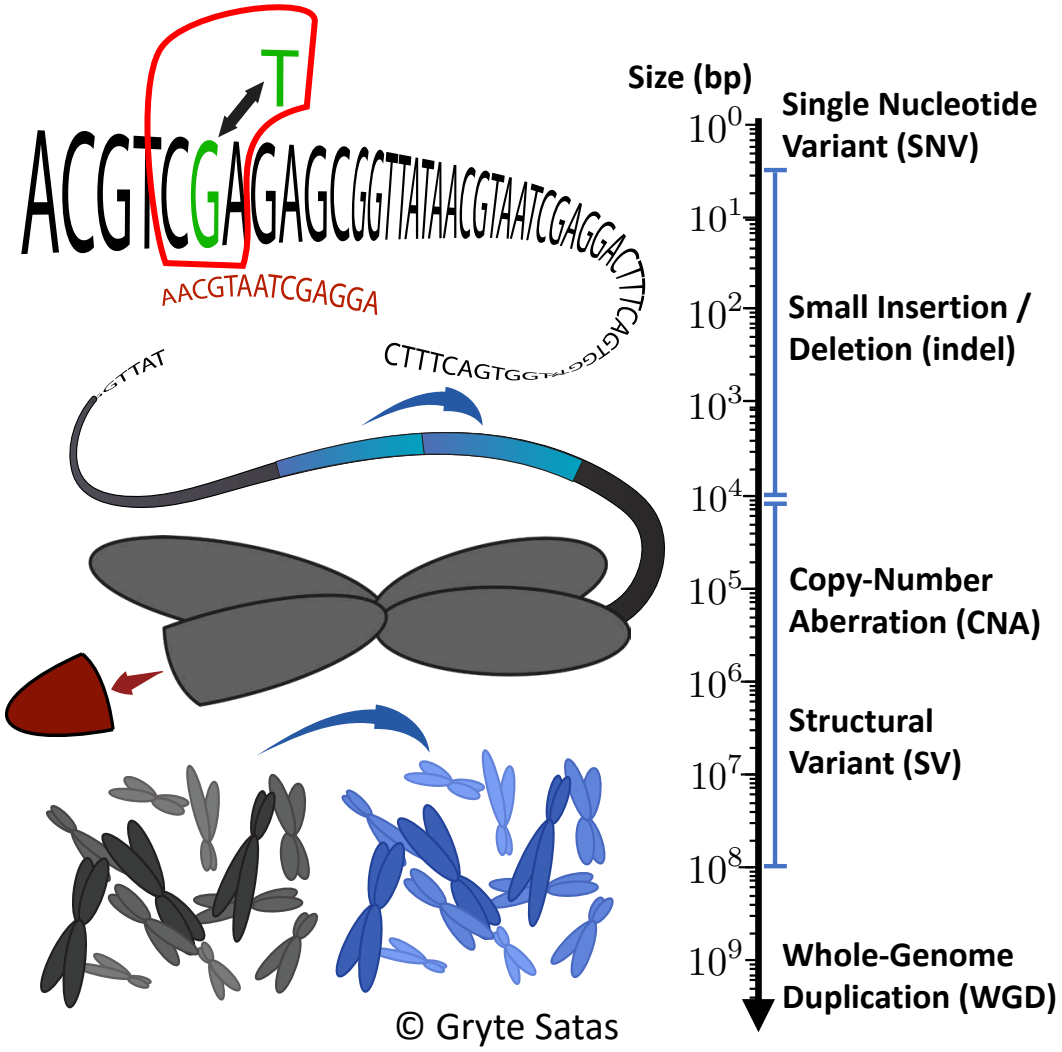
- Multiple consensus tree problem

5. Applications

- Mutational signature dynamics [PSB 2020]

Sarah Christensen

Mutational Signatures

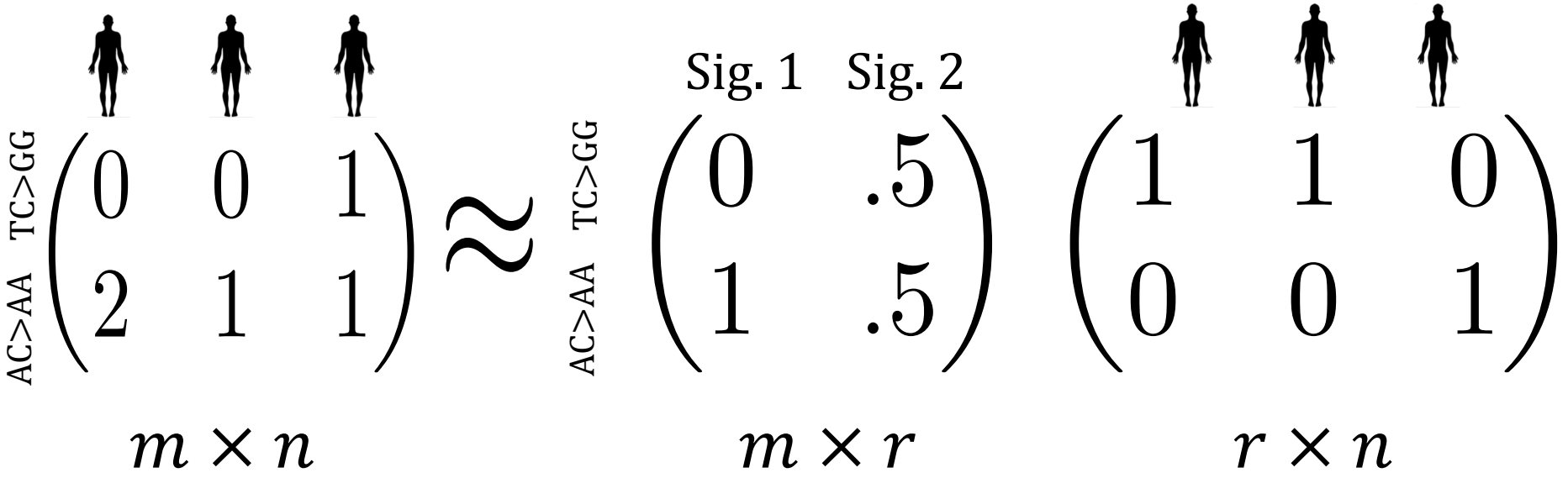


There are 4*6*4 = 96 SNV categories: AC>AA, ... GT>GG

Distinct mutational processes => distinct patterns of SNV categories

Mutational Signatures – NMF

P: Feature Matrix $\quad$ S: Signature Matrix $\quad$ E: Exposure Matrix



$$\begin{matrix} \text{AC>AA} \\ \text{TC>GG} \end{matrix} \begin{pmatrix} 0 & 0 & 1 \\ 2 & 1 & 1 \end{pmatrix} \approx \begin{matrix} \text{AC>AA} \\ \text{TC>GG} \end{matrix} \begin{pmatrix} 0 & .5 \\ 1 & .5 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$m \times n$ $\quad$ $m \times r$ $\quad$ $r \times n$

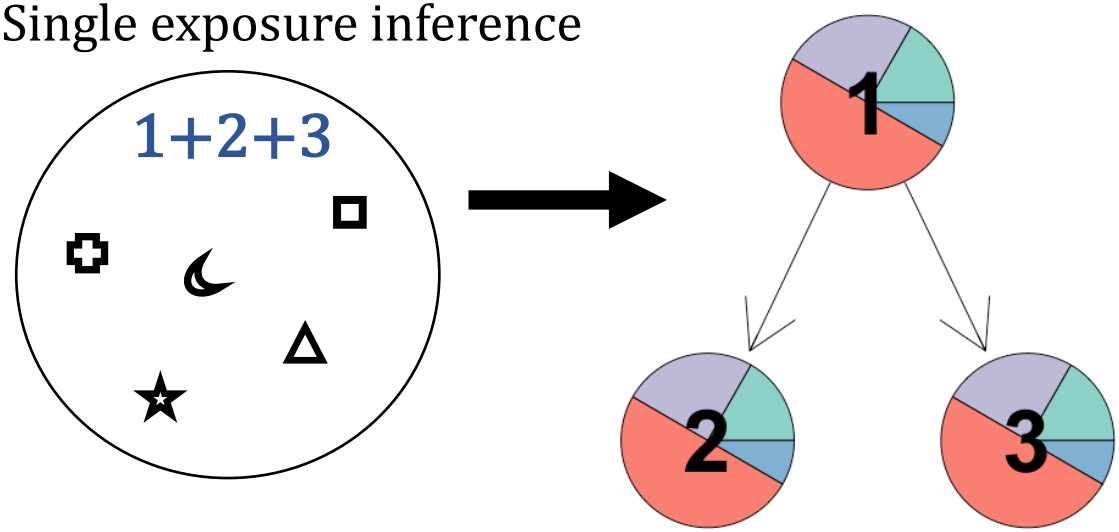
Alexandrov et al. [Nature, 2013] performed nonnegative matrix factorization on a large patient cohort (n = 10,000) identifying r = 30 signatures and exposures

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

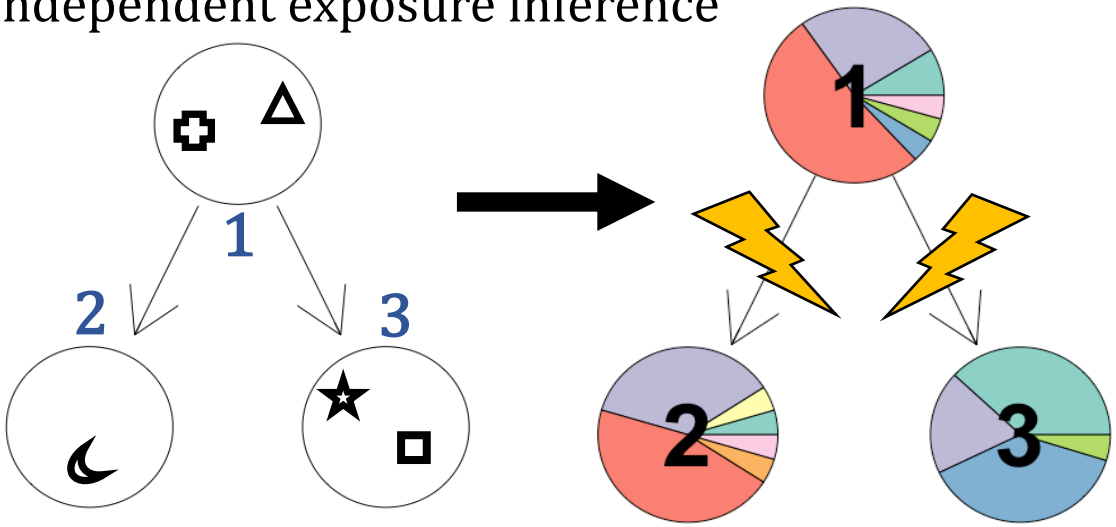
Clone-specific exposure inference problems

Single exposure inference



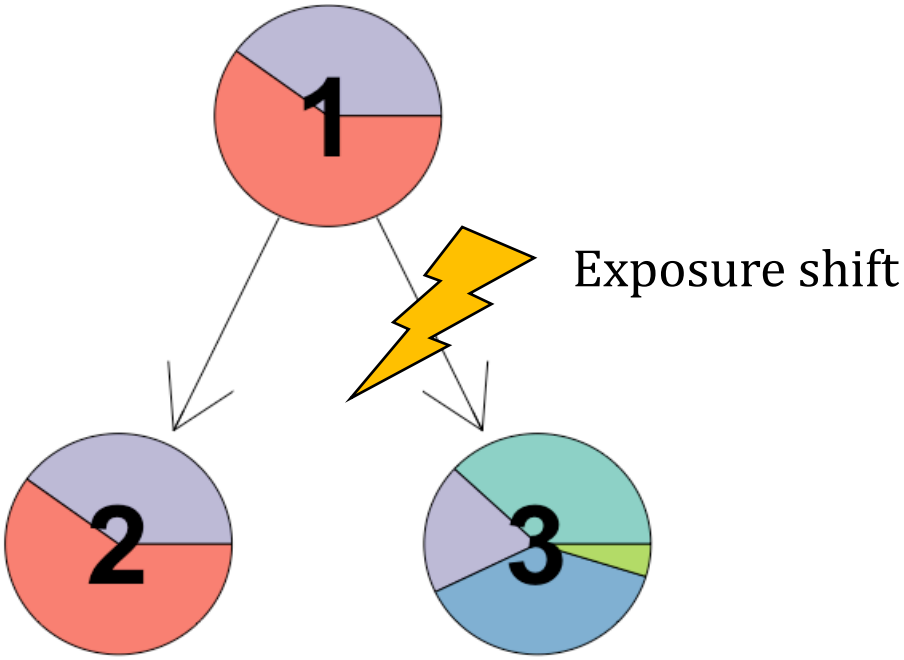
- Alexandrov et al., Cell reports 2013
- Rosenthal et al., Genome Biology 2016
- Huang et al., Bioinformatics, 2017
- …

Independent exposure inference



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

Tree-constrained exposure (TE) inference



Tree-constrained exposure 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*

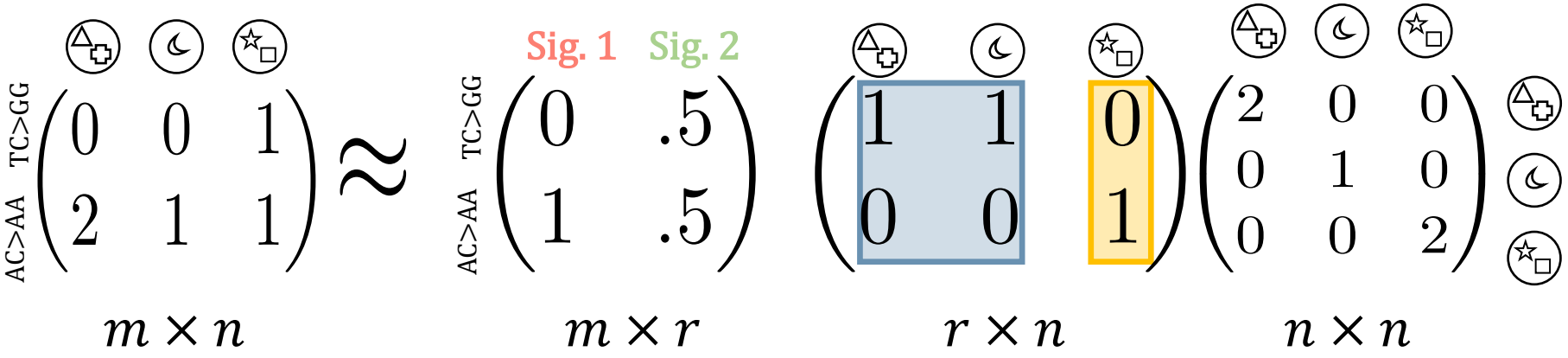
PhySigs solves the TE problem to optimality

P: Feature Matrix

S: Signature Matrix

D: Relative Exposure Matrix

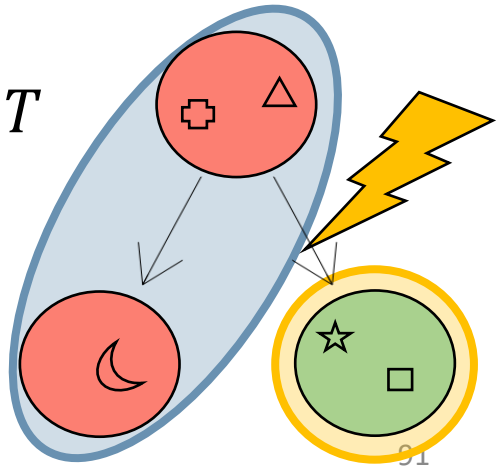
C: Count Matrix



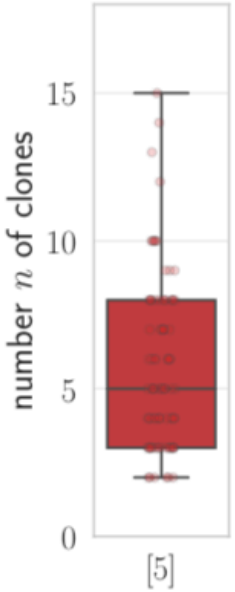
$m \times n \quad m \times r \quad r \times n \quad n \times n$

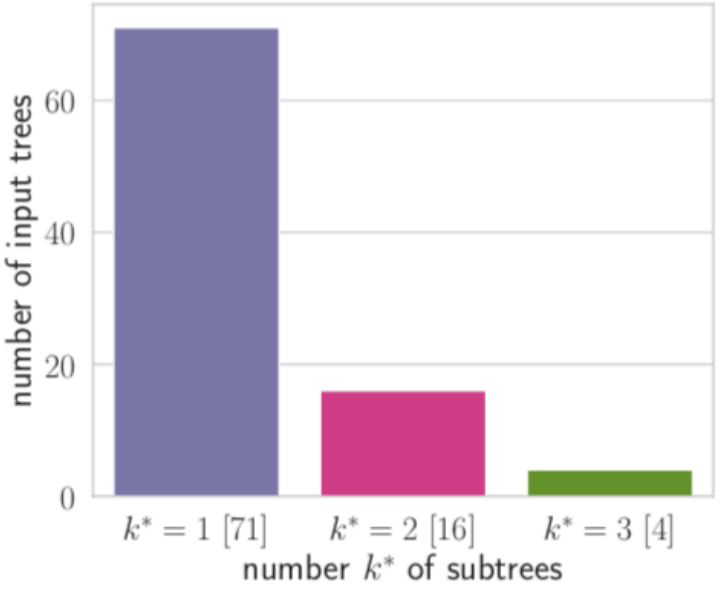
TE problem:

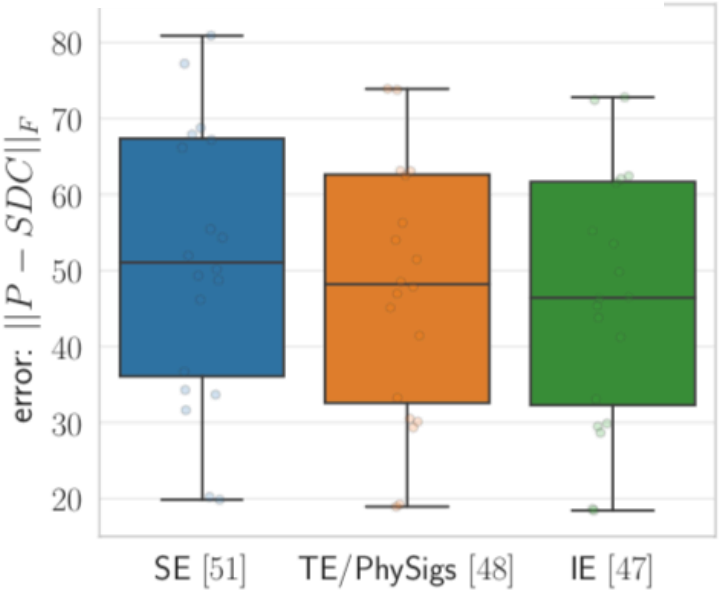
Given feature matrix *P*, corresponding count matrix *C*, signature matrix *S*, phylogenetic tree *T* and integer $k \geq 1$, find relative exposure matrix D such that $||P - SDC||_F$ is minimum and *D* is composed of *k* sets of identical columns, each corresponding to a connected subtree of *T*.

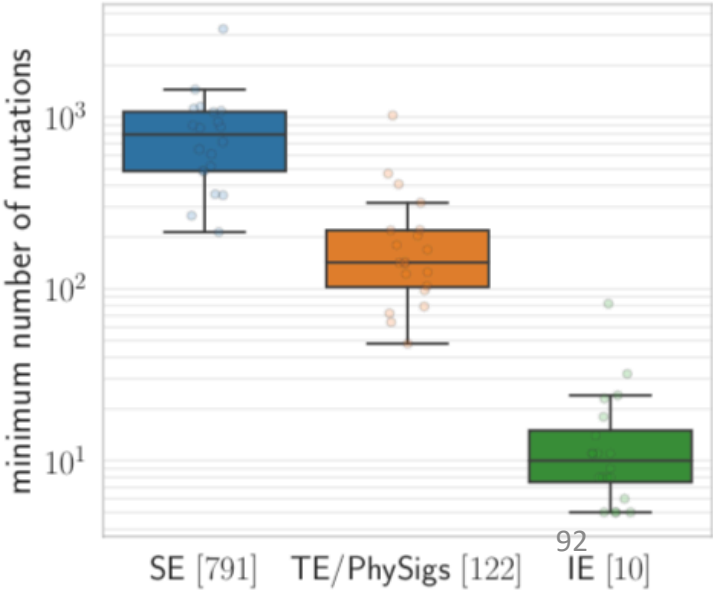


PhySigs identifies accurate exposures without overfitting in a lung cancer cohort









PhySigs identifies an exposure shift supported by a subclonal driver

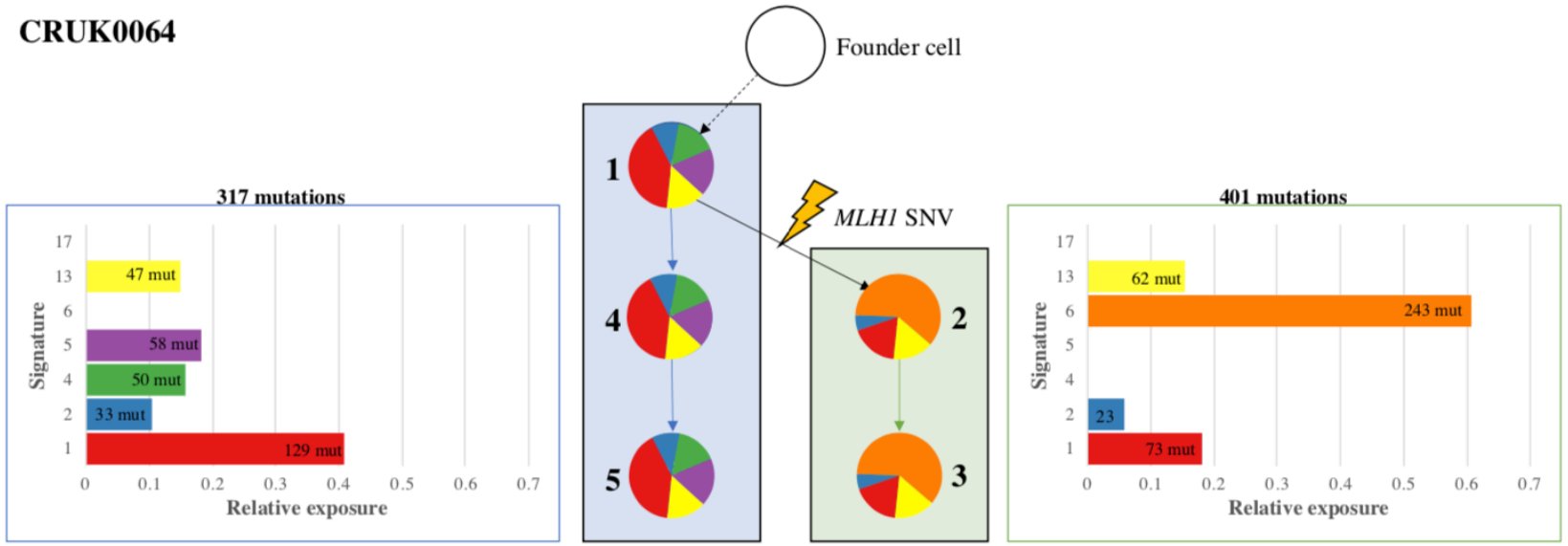
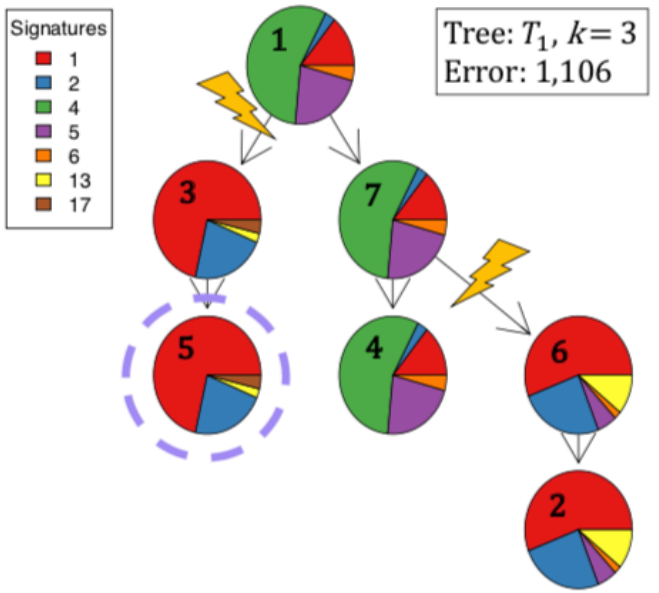


Fig. 5. **PhySigs detects a large increase in DNA mismatch repair-associated Signature 6 (orange) along one branch (clusters 2 and 3; green) of the CRUK 0064 tree.** In support of this finding, the branch includes a subclonal driver mutation to DNA mismatch repair gene *MLH1*.

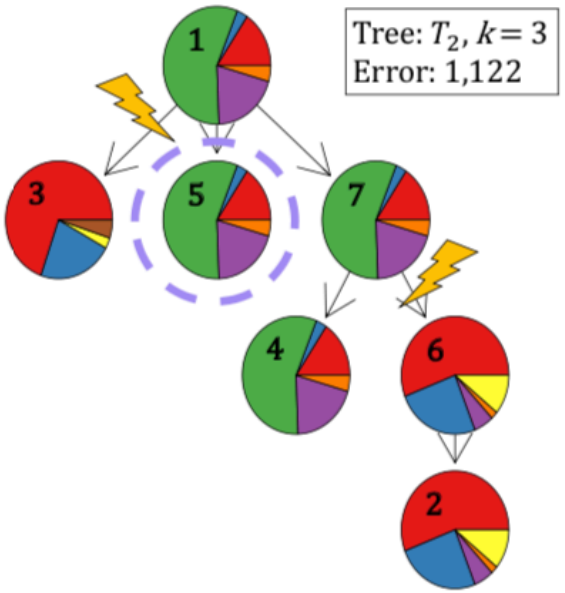
PhySigs enables prioritization of trees in solution space

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)



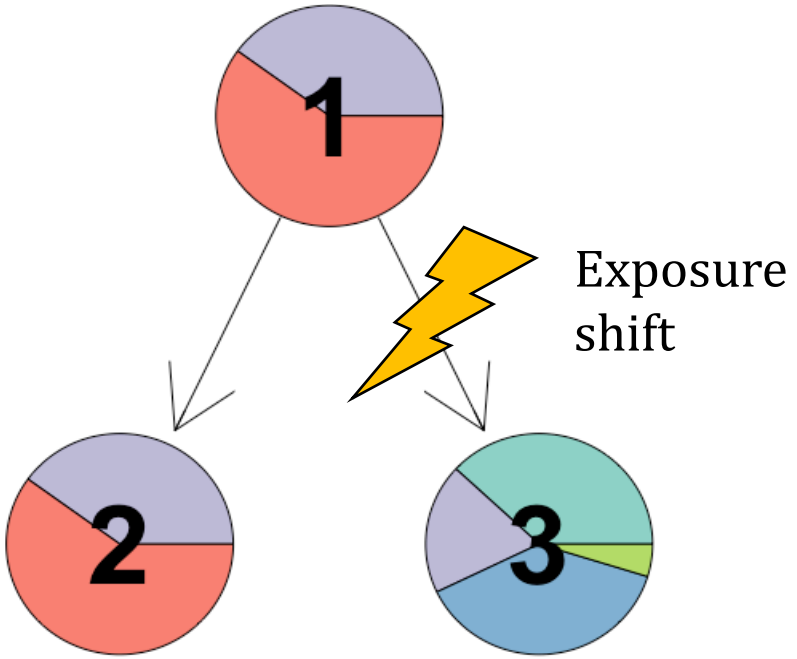
(b)



(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*



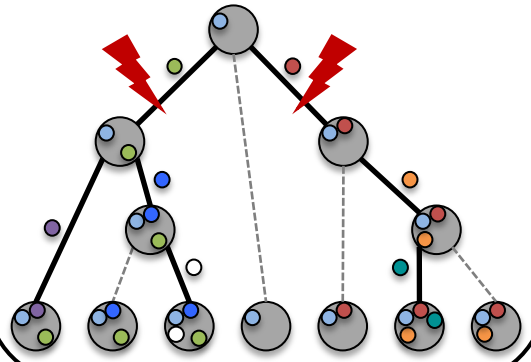
Tree-constrained exposure inference

- **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*

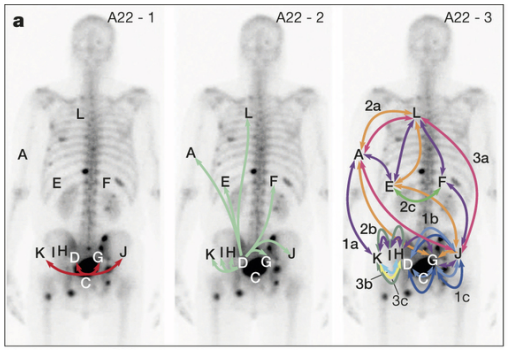
Conclusion

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

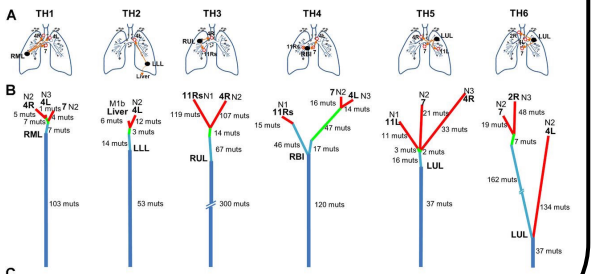
Identify targets for treatment



Understand metastatic development



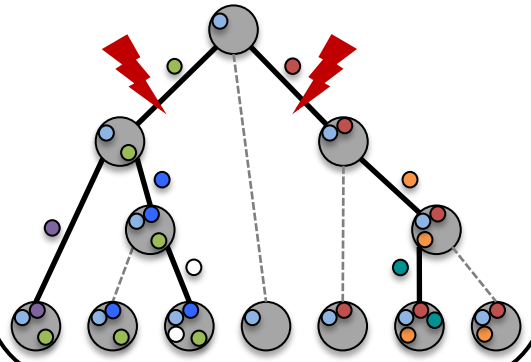
Recognize common patterns of tumor evolution across patients



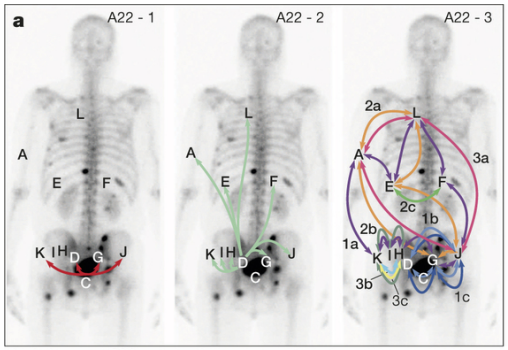
Conclusion

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

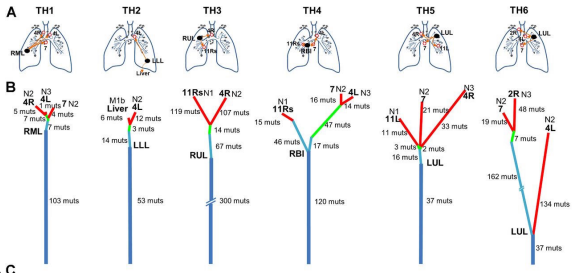
Identify targets for treatment



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. Almost uniform sampling of solutions to PPM

4. Summarizing solution space using multiple consensus trees

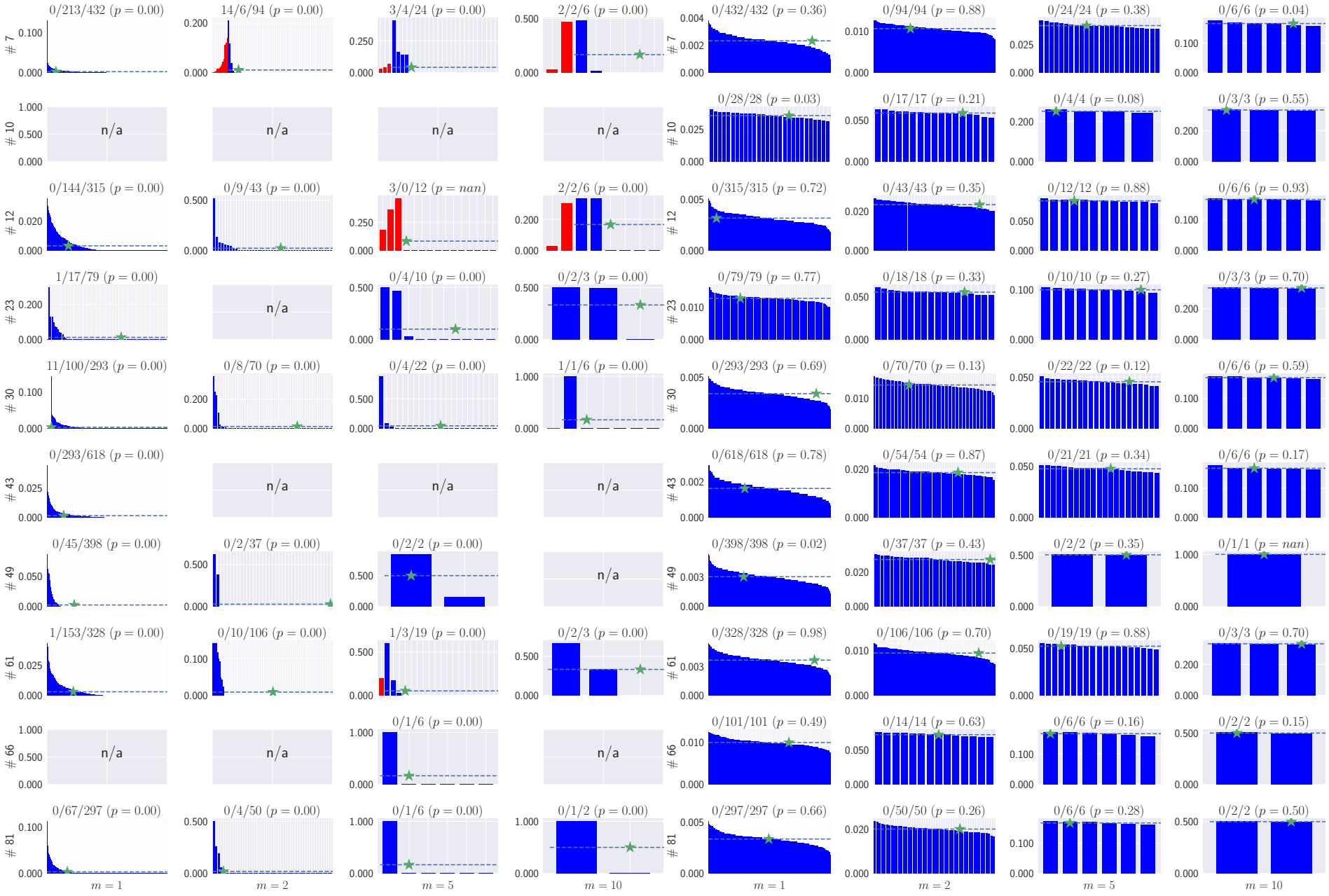
5. Example of downstream application: Mutational signature dynamics

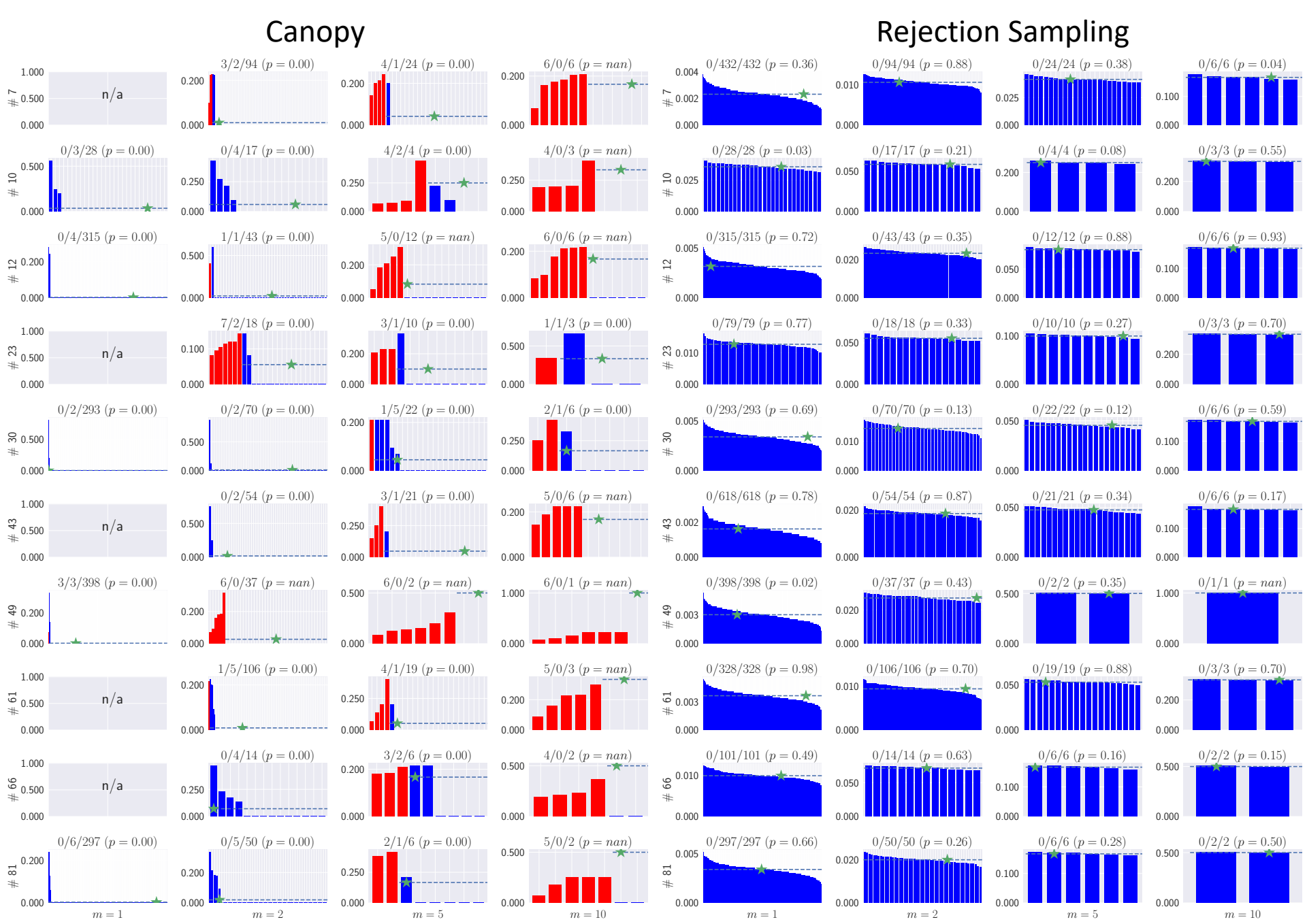
Acknowledgments

El-Kebir group

- **Yuanyuan Qi**
- **Nuraini Aguse**
- **Sarah Christensen**
- **Dikshant Pradhan**
- Jiaqi Wu
- Juho Kim
- Yerong Li
- 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 18-50502)


PhyloWGS
Rejection Sampling
0/213/432 (p = 0.00)
14/6/94 (p = 0.00)
3/4/24 (p = 0.00)
2/2/6 (p = 0.00)
0/432/432 (p = 0.36)
0/94/94 (p = 0.88)
0/24/24 (p = 0.38)
0/6/6 (p = 0.04)
7
10
n/a
0/28/28 (p = 0.03)
0/17/17 (p = 0.21)
0/4/4 (p = 0.08)
0/3/3 (p = 0.55)
0/144/315 (p = 0.00)
0/9/43 (p = 0.00)
3/0/12 (p = nan)
2/2/6 (p = 0.00)
0/315/315 (p = 0.72)
0/43/43 (p = 0.35)
0/12/12 (p = 0.88)
0/6/6 (p = 0.93)
12
23
1/17/79 (p = 0.00)
0/4/10 (p = 0.00)
0/2/3 (p = 0.00)
0/79/79 (p = 0.77)
0/18/18 (p = 0.33)
0/10/10 (p = 0.27)
0/3/3 (p = 0.70)
11/100/293 (p = 0.00)
0/8/70 (p = 0.00)
0/4/22 (p = 0.00)
1/1/6 (p = 0.00)
0/293/293 (p = 0.69)
0/70/70 (p = 0.13)
0/22/22 (p = 0.12)
0/6/6 (p = 0.59)
30
43
0/293/618 (p = 0.00)
0/618/618 (p = 0.78)
0/54/54 (p = 0.87)
0/21/21 (p = 0.34)
0/6/6 (p = 0.17)
0/45/398 (p = 0.00)
0/2/37 (p = 0.00)
0/2/2 (p = 0.00)
0/398/398 (p = 0.02)
0/37/37 (p = 0.43)
0/2/2 (p = 0.35)
0/1/1 (p = nan)
49
61
1/153/328 (p = 0.00)
0/10/106 (p = 0.00)
1/3/19 (p = 0.00)
0/2/3 (p = 0.00)
0/328/328 (p = 0.98)
0/106/106 (p = 0.70)
0/19/19 (p = 0.88)
0/3/3 (p = 0.70)
66
0/1/6 (p = 0.00)
0/101/101 (p = 0.49)
0/14/14 (p = 0.63)
0/6/6 (p = 0.16)
0/2/2 (p = 0.15)
0/67/297 (p = 0.00)
0/4/50 (p = 0.00)
0/1/6 (p = 0.00)
0/1/2 (p = 0.00)
0/297/297 (p = 0.66)
0/50/50 (p = 0.26)
0/6/6 (p = 0.28)
0/2/2 (p = 0.50)
# 81
m = 1
m = 2
m = 5
m = 10


Canopy
Rejection Sampling
7
10
12
23
30
43
49
61
66
81
n/a
3/2/94 (p = 0.00)
4/1/24 (p = 0.00)
6/0/6 (p = nan)
0/432/432 (p = 0.36)
0/94/94 (p = 0.88)
0/24/24 (p = 0.38)
0/6/6 (p = 0.04)
0/3/28 (p = 0.00)
0/4/17 (p = 0.00)
4/2/4 (p = 0.00)
4/0/3 (p = nan)
0/28/28 (p = 0.03)
0/17/17 (p = 0.21)
0/4/4 (p = 0.08)
0/3/3 (p = 0.55)
0/4/315 (p = 0.00)
1/1/43 (p = 0.00)
5/0/12 (p = nan)
6/0/6 (p = nan)
0/315/315 (p = 0.72)
0/43/43 (p = 0.35)
0/12/12 (p = 0.88)
0/6/6 (p = 0.93)
7/2/18 (p = 0.00)
3/1/10 (p = 0.00)
1/1/3 (p = 0.00)
0/79/79 (p = 0.77)
0/18/18 (p = 0.33)
0/10/10 (p = 0.27)
0/3/3 (p = 0.70)
0/2/293 (p = 0.00)
0/2/70 (p = 0.00)
1/5/22 (p = 0.00)
2/1/6 (p = 0.00)
0/293/293 (p = 0.69)
0/70/70 (p = 0.13)
0/22/22 (p = 0.12)
0/6/6 (p = 0.59)
0/2/54 (p = 0.00)
3/1/21 (p = 0.00)
5/0/6 (p = nan)
0/618/618 (p = 0.78)
0/54/54 (p = 0.87)
0/21/21 (p = 0.34)
0/6/6 (p = 0.17)
3/3/398 (p = 0.00)
6/0/37 (p = nan)
6/0/2 (p = nan)
6/0/1 (p = nan)
0/398/398 (p = 0.02)
0/37/37 (p = 0.43)
0/2/2 (p = 0.35)
0/1/1 (p = nan)
1/5/106 (p = 0.00)
4/1/19 (p = 0.00)
5/0/3 (p = nan)
0/328/328 (p = 0.98)
0/106/106 (p = 0.70)
0/19/19 (p = 0.88)
0/3/3 (p = 0.70)
0/4/14 (p = 0.00)
3/2/6 (p = 0.00)
4/0/2 (p = nan)
0/101/101 (p = 0.49)
0/14/14 (p = 0.63)
0/6/6 (p = 0.16)
0/2/2 (p = 0.15)
0/6/297 (p = 0.00)
0/5/50 (p = 0.00)
2/1/6 (p = 0.00)
5/0/2 (p = nan)
0/297/297 (p = 0.66)
0/50/50 (p = 0.26)
0/6/6 (p = 0.28)
0/2/2 (p = 0.50)
m = 1
m = 2
m = 5
m = 10

Somatic Mutations Occur at Different Genomic Scales