On the Non-uniqueness of Solutions to the Perfect Phylogeny Mixture Problem

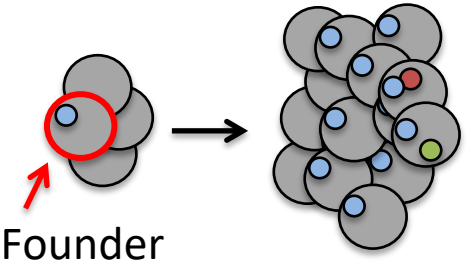
Dikshant Pradhan and Mohammed El-Kebir

University of Illinois at Urbana Champaign,
Department of Computer Science

RECOMB-CG 2018

Tumorigenesis: Cell Mutation

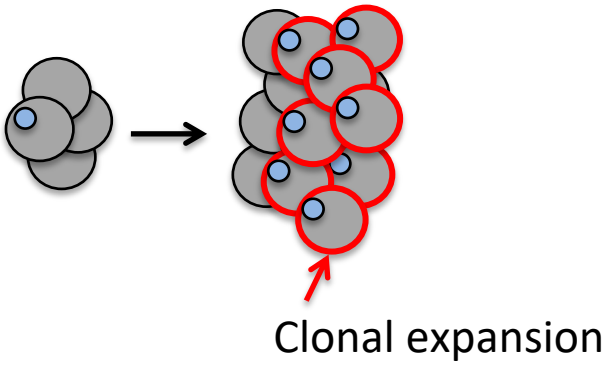
Clonal Evolution Theory of Cancer
[Nowell, 1976]



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

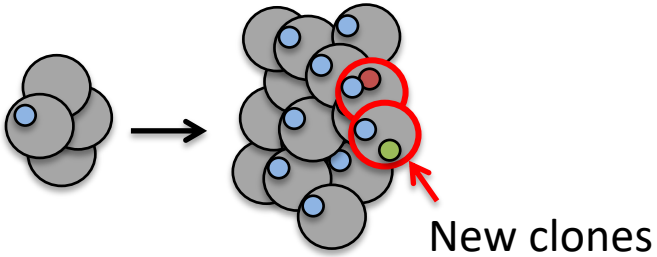
Tumorigenesis: Cell Mutation

Clonal Evolution Theory of Cancer
[Nowell, 1976]



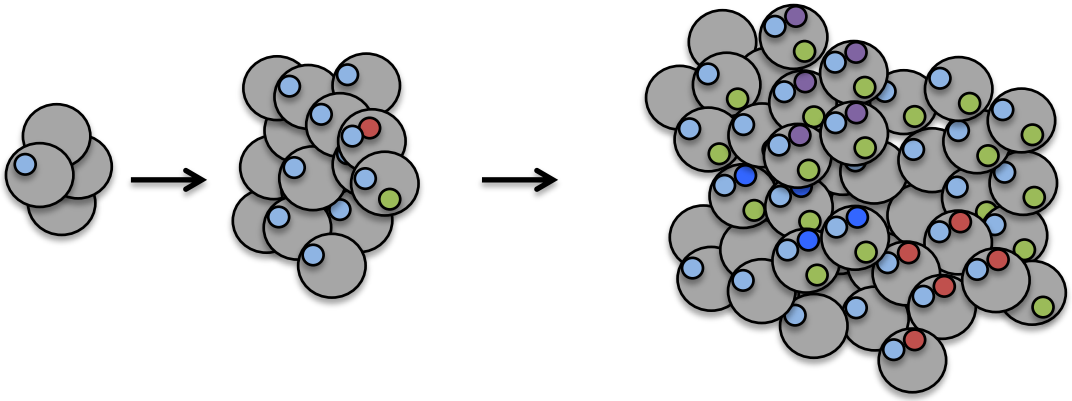
Tumorigenesis: Cell Mutation

Clonal Evolution Theory of Cancer
[Nowell, 1976]



Tumorigenesis: Cell Mutation & Division

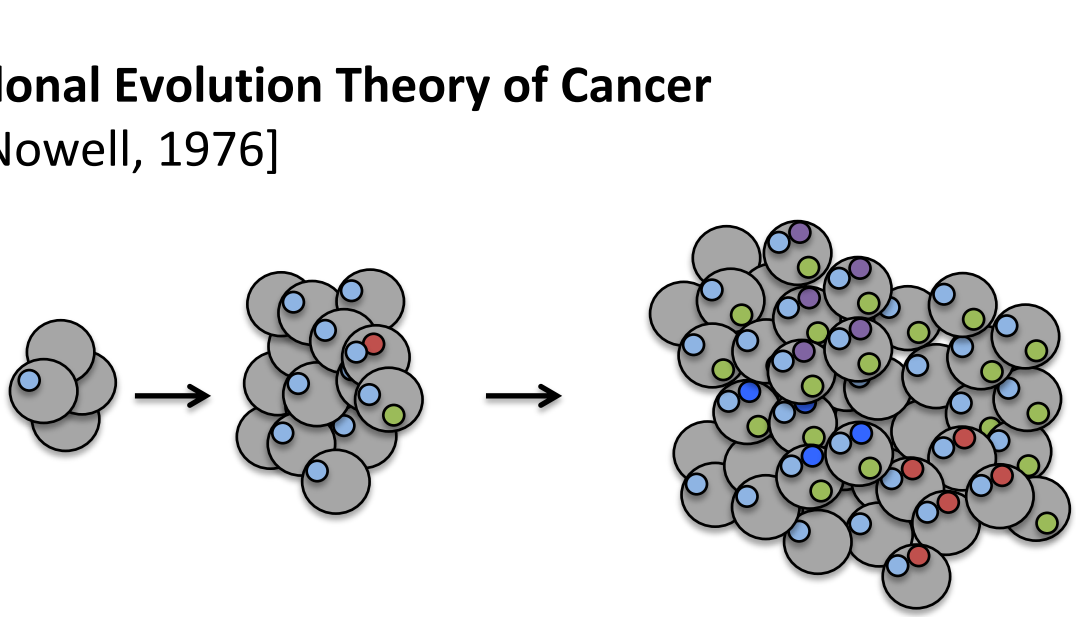
Clonal Evolution Theory of Cancer
[Nowell, 1976]



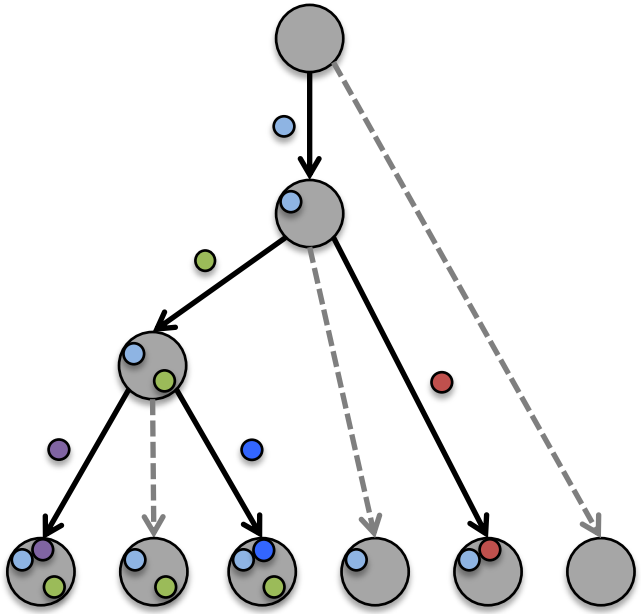
Intra-Tumor
Heterogeneity

Tumorigenesis: Cell Mutation & Division

Clonal Evolution Theory of Cancer
[Nowell, 1976]



Intra-Tumor Heterogeneity

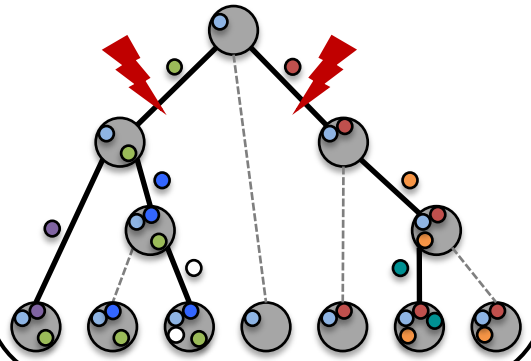


Phylogenetic Tree ***T***

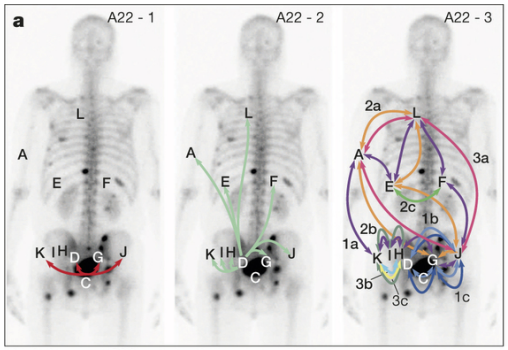
Question: Why are tumor phylogenies important?

Phylogenies are Key to Understanding Cancer

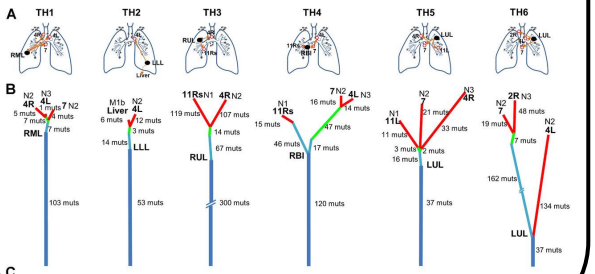
Identify targets for treatment



Understand metastatic development

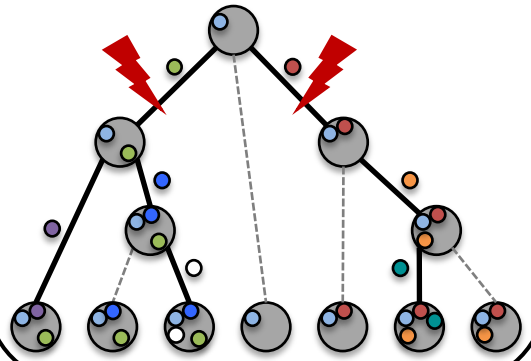


Recognize common patterns of tumor evolution across patients

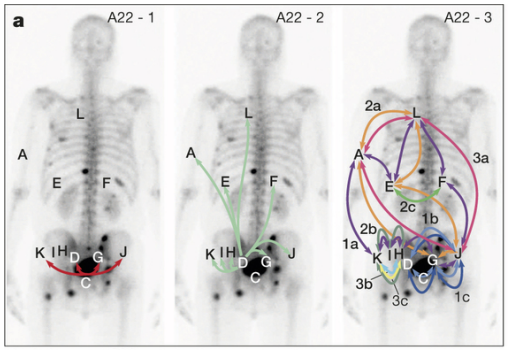


Phylogenies are Key to Understanding Cancer

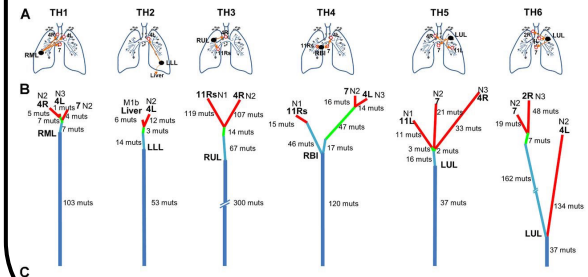
Identify targets for treatment



Understand metastatic development



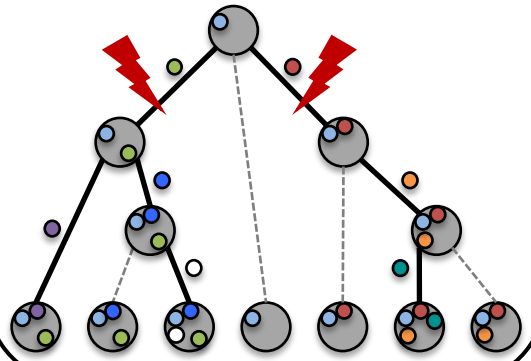
Recognize common patterns of tumor evolution across patients



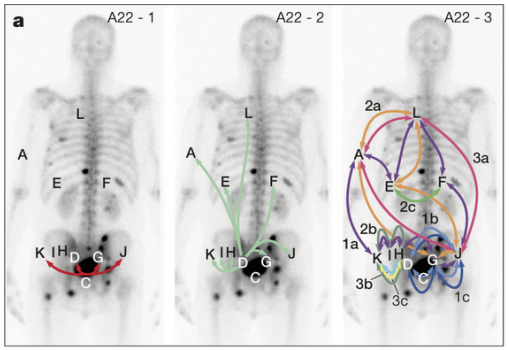
These downstream analyses **critically rely** on accurate tumor phylogeny inference

Phylogenies are Key to Understanding Cancer

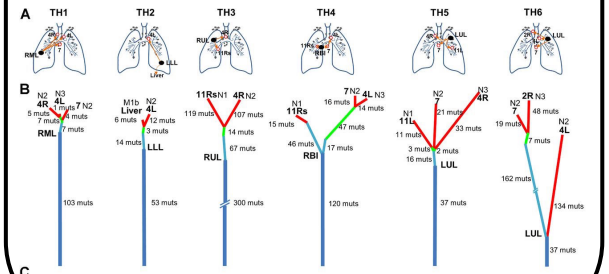
Identify targets for treatment



Understand metastatic development



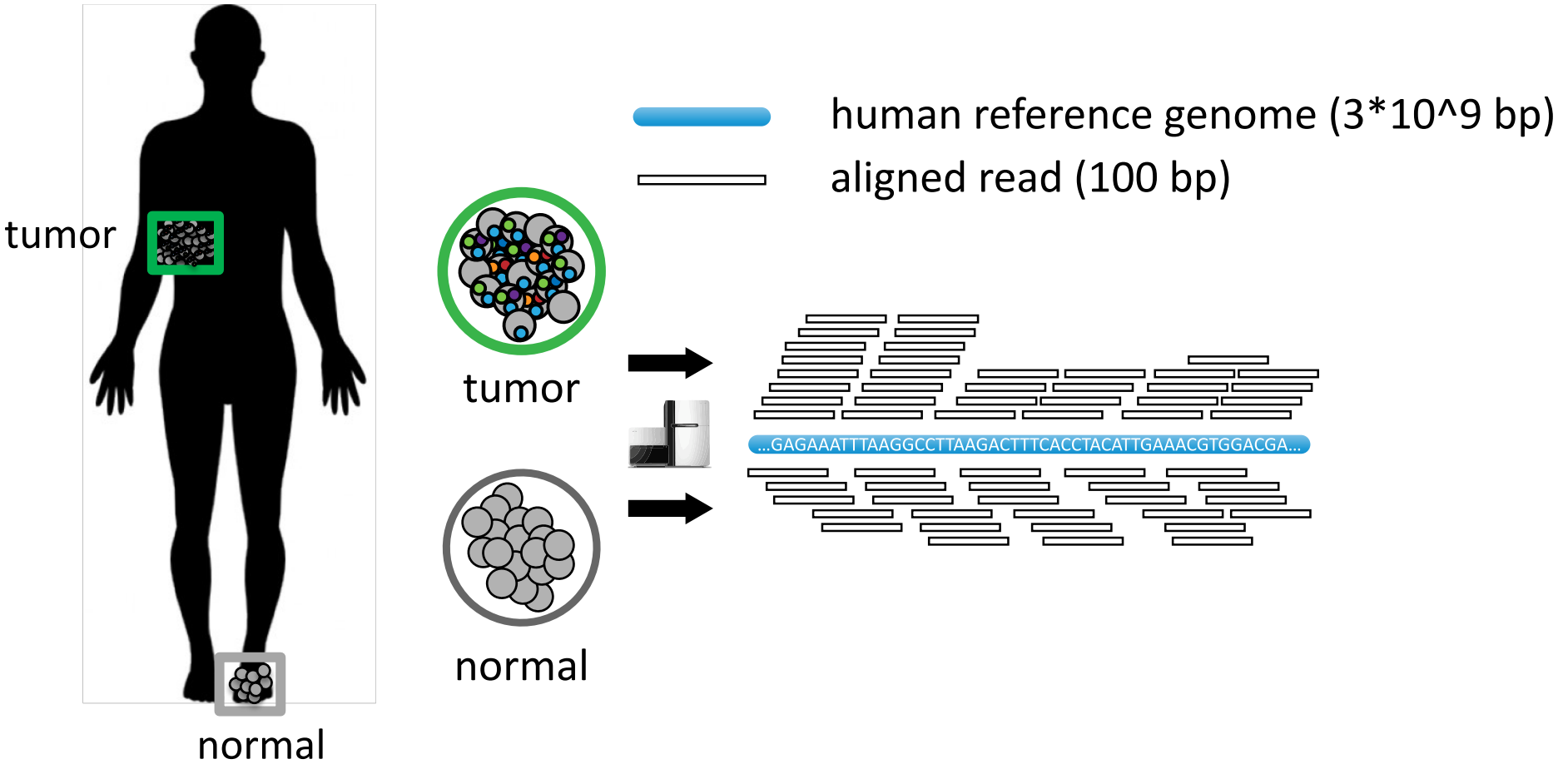
Recognize common patterns of tumor evolution across patients



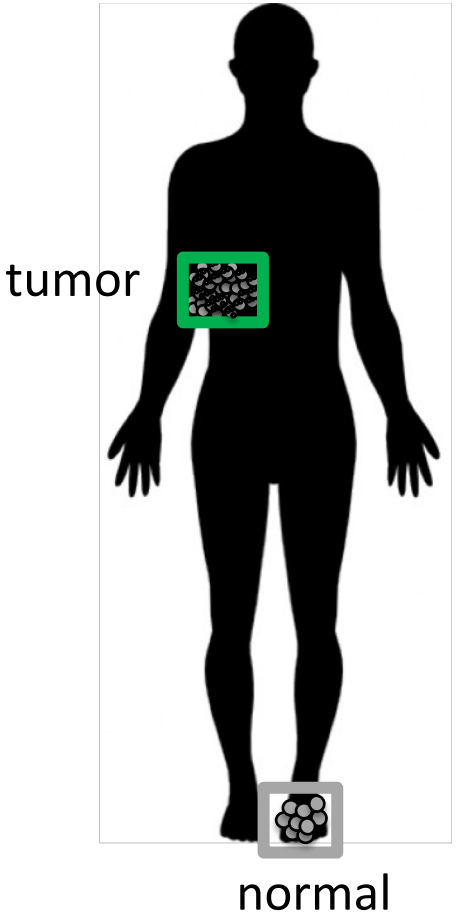
These downstream analyses **critically rely** on accurate tumor phylogeny inference

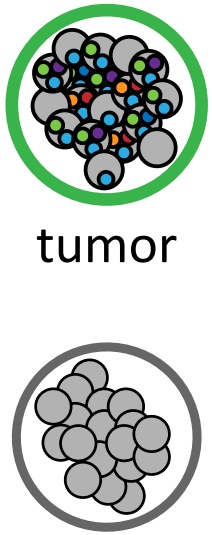
Key challenge in phylogenetics:
Accurate phylogeny inference from data at present time

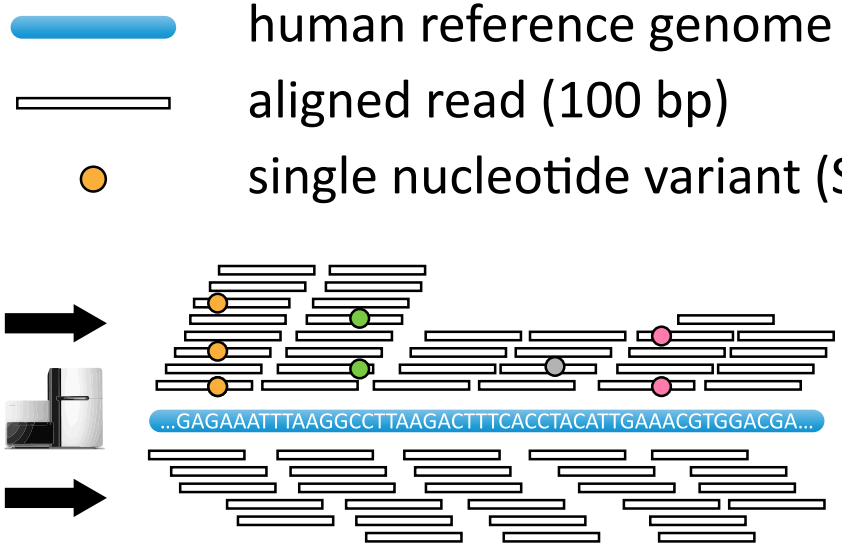
Additional Challenge in Cancer Phylogenetics



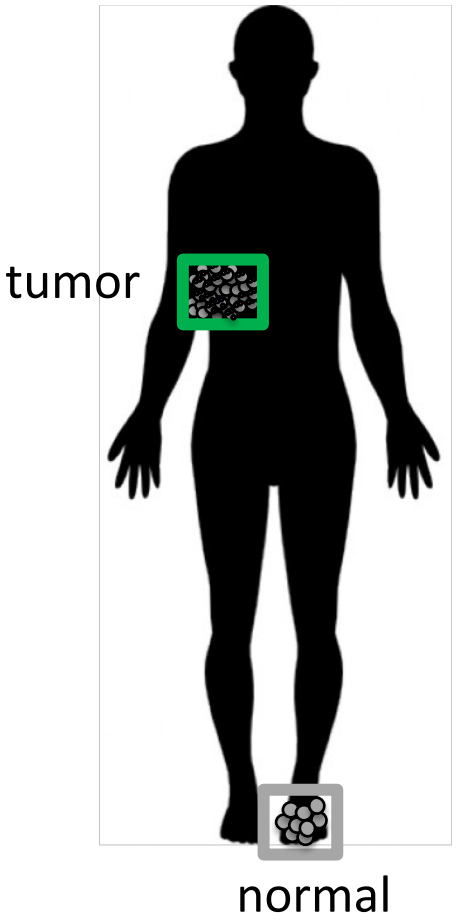
Additional Challenge in Cancer Phylogenetics

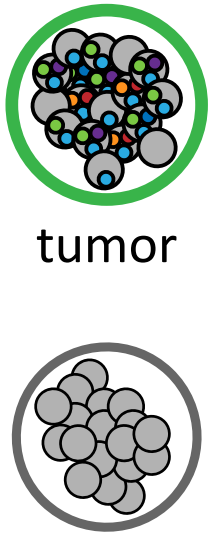






Additional Challenge in Cancer Phylogenetics





human reference genome (3*10^9 bp)

aligned read (100 bp)

single nucleotide variant (SNV)

Additional challenge in cancer phylogenetics:

Phylogeny inference from **mixed bulk samples** at present time

Outline

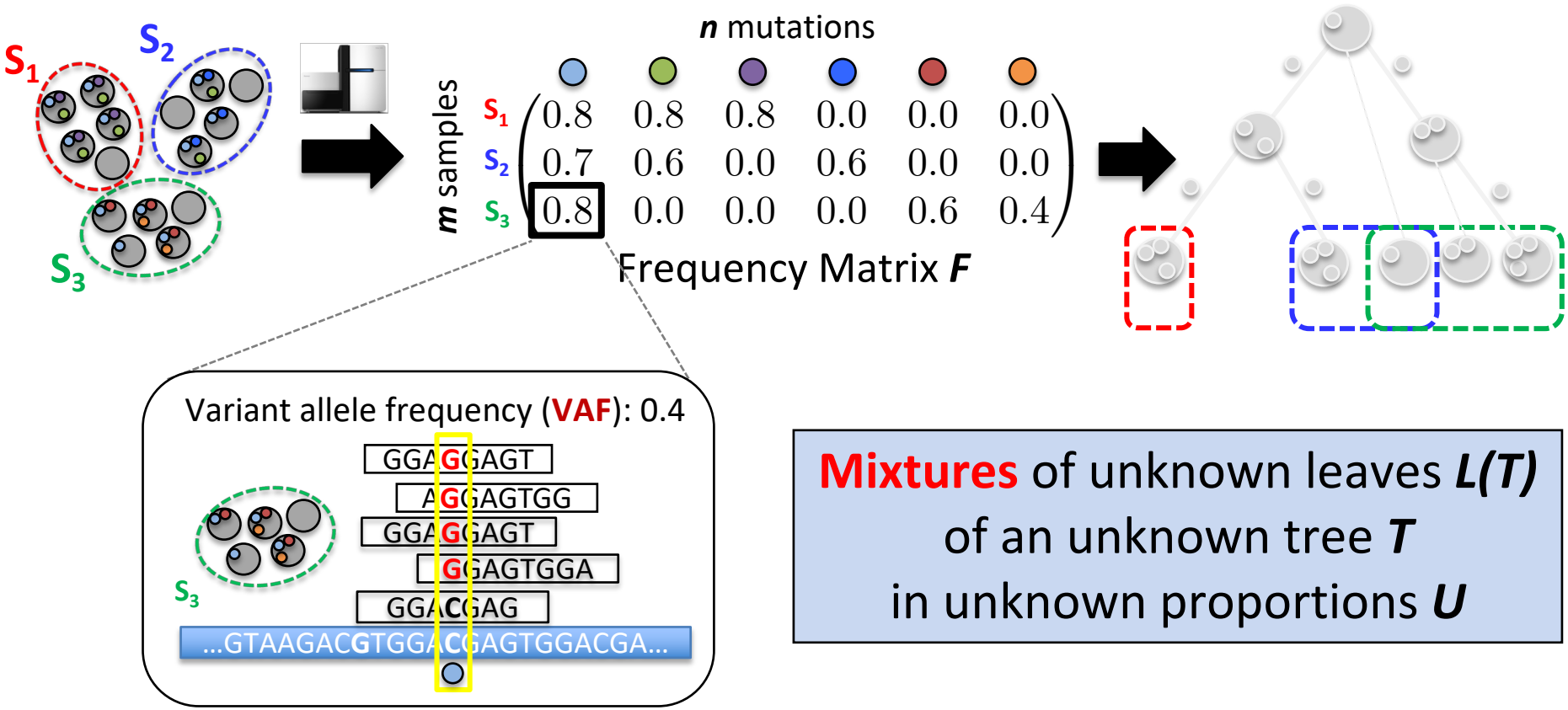
Background and theory:

- Perfect Phylogeny Mixture (PPM) problem
- Combinatorial characterization of solutions
- #PPM: exact counting and uniform sampling

Simulation results:

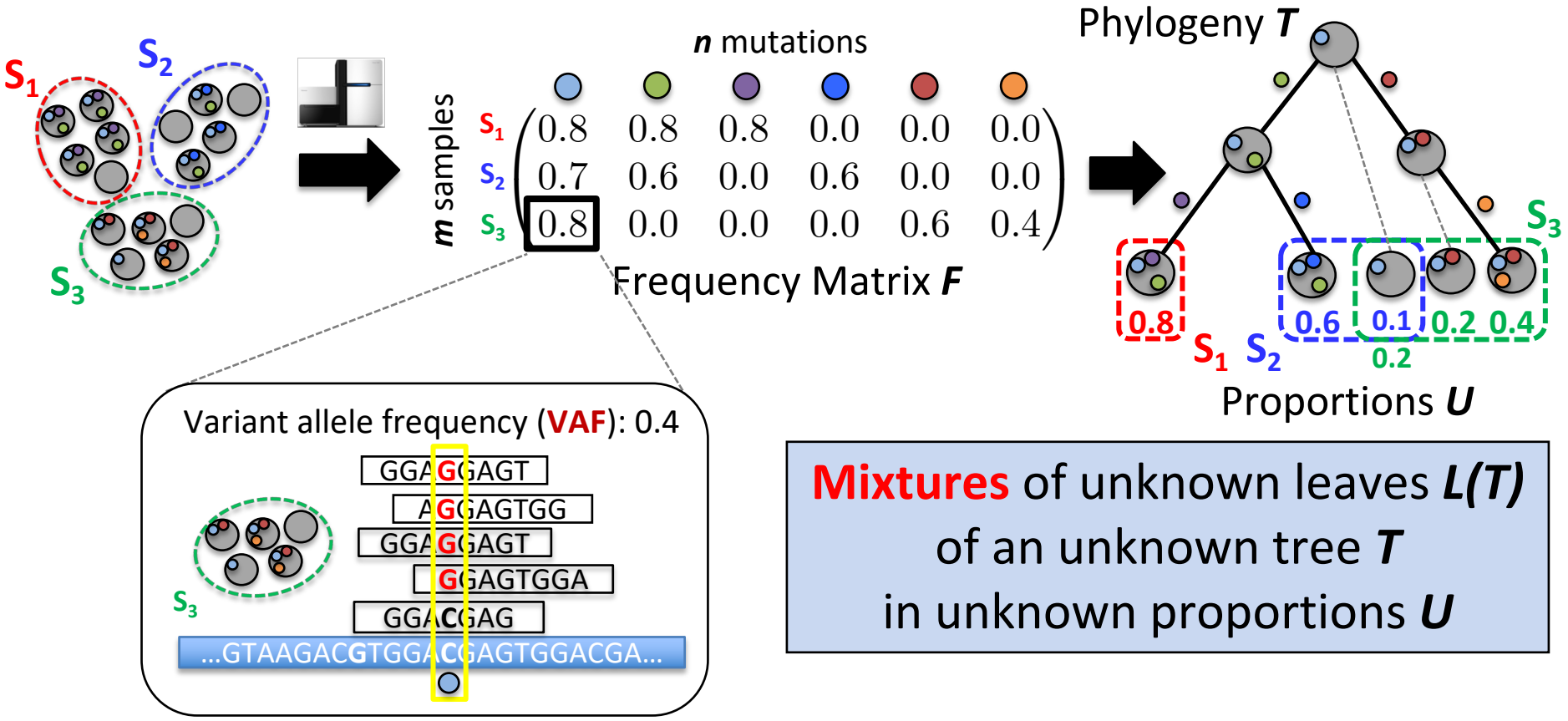
- What contributes to non-uniqueness?
- How to reduce non-uniqueness?
- How does non-uniqueness affect current methods?

Sequencing and Tumor Phylogeny Inference



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

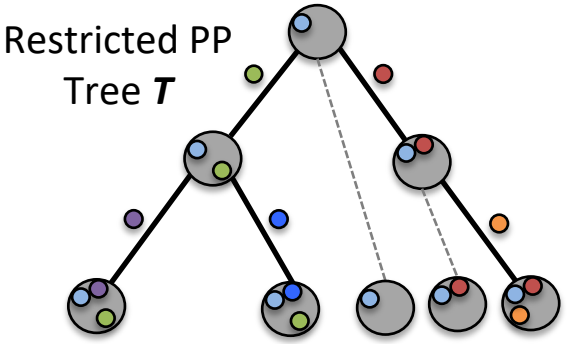
Sequencing and Tumor Phylogeny Inference



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***

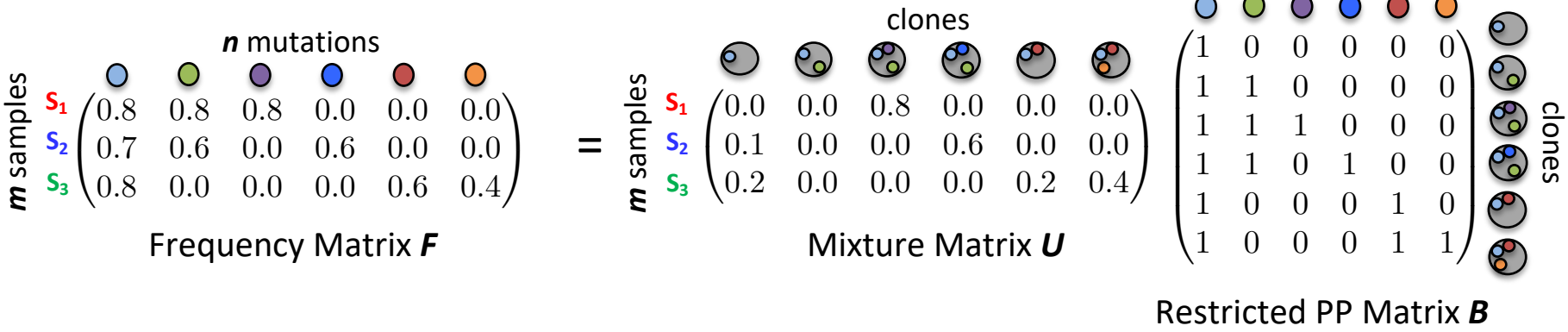
Perfect Phylogeny Mixture



Assumptions:

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

1-1 Equivalent



Rows of U are proportions:

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

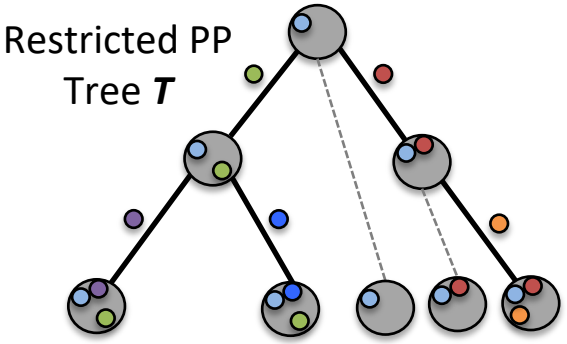
Perfect Phylogeny Theorem
[Estabrook, 1971]
[Gusfield, 1991]

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

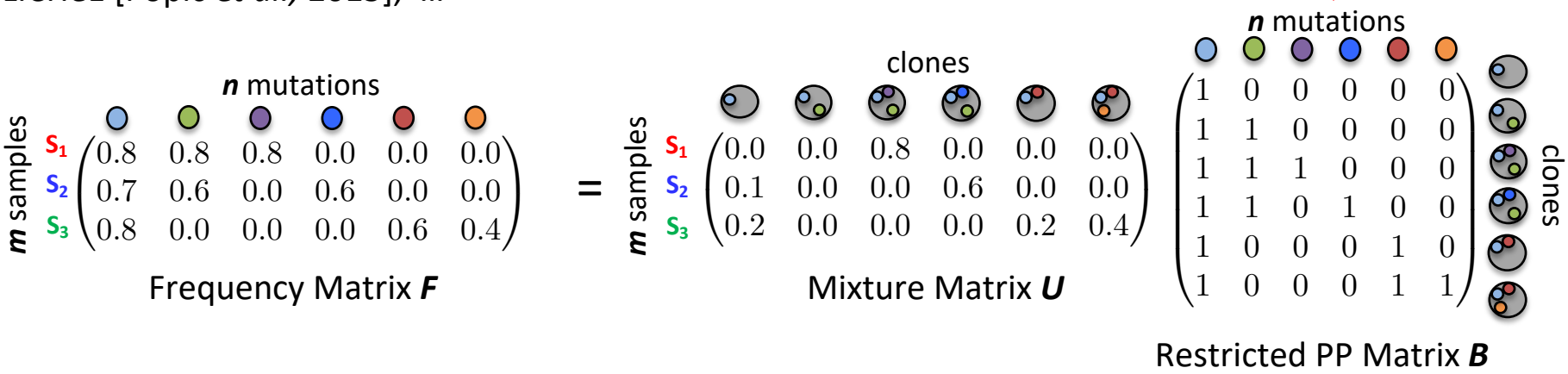
Previous Work

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**



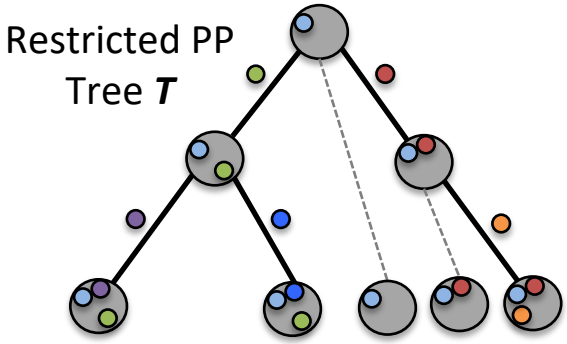
Rows of ***U*** are proportions:

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

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

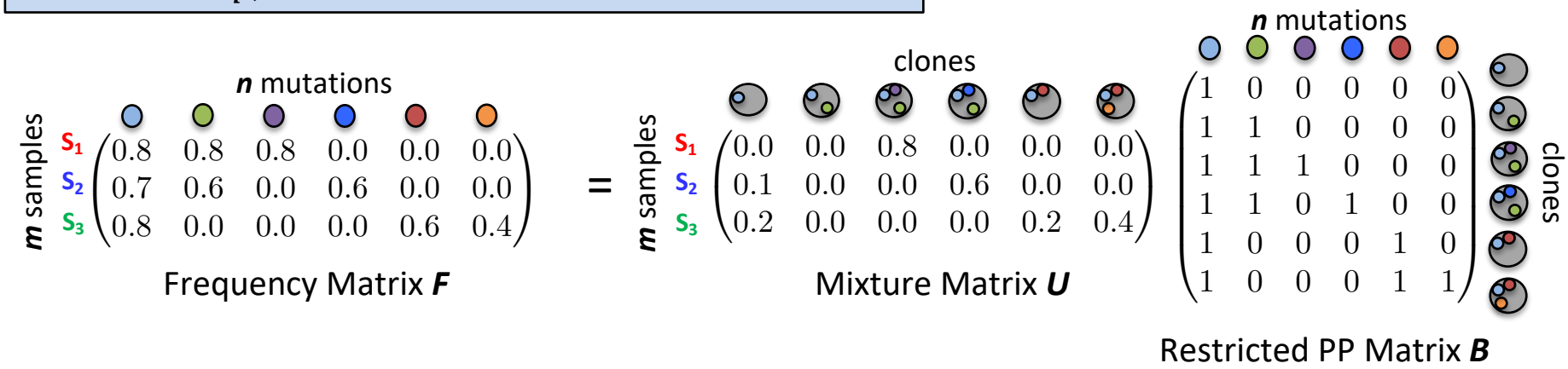
Perfect Phylogeny Mixture: [El-Kebir*, Oesper* et al., 2015]
Given ***F***, find ***U*** and ***B*** such that ***F = U B***

Combinatorial Characterization



- Frequency $f_{p,i}$ is mass of subtree rooted at node that introduced i
- Usage $u_{p,i}$ is mass of node that introduced i

1-1 **Equivalent**



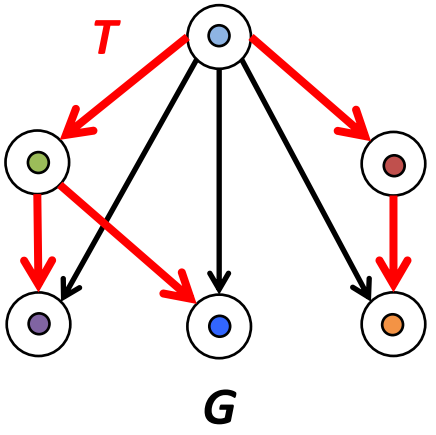
Rows of ***U*** are proportions:

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

Perfect Phylogeny Mixture: [El-Kebir*, Oesper* et al., 2015]

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

Combinatorial Characterization



- Frequency $f_{p,i}$ is mass of subtree rooted at node that introduced i
- Usage $u_{p,i}$ is mass of node that introduced i

n mutations

m samples

$$\begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{pmatrix} 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

Frequency Matrix ***F***

=

clones

m samples

$$\begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{pmatrix} 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.1 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.2 & 0.0 & 0.0 & 0.0 & 0.2 & 0.4 \end{pmatrix}$$

Mixture Matrix ***U***

n mutations

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

clones

Restricted PP Matrix ***B***

Theorem 1:

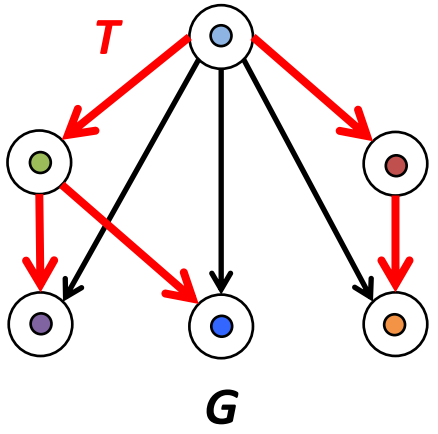
T is a solution to the PPM if and only if ***T*** is a spanning tree of ***G*** satisfying the sum condition

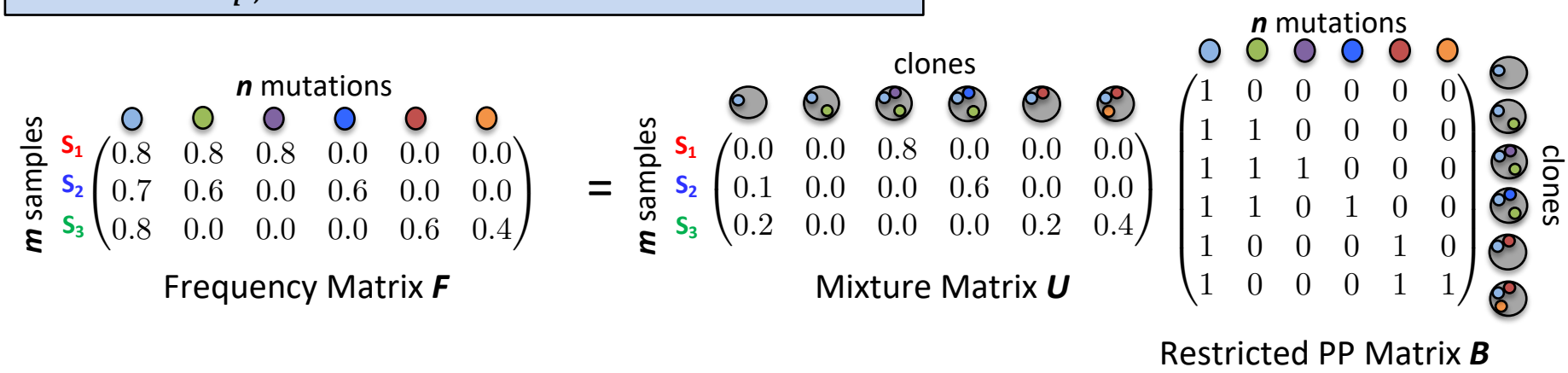
Perfect Phylogeny Mixture: [El-Kebir*, Oesper* et al., 2015]

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

Combinatorial Characterization

- Frequency $f_{p,i}$ is mass of subtree rooted at node that introduced i
- Usage $u_{p,i}$ is mass of node that introduced i



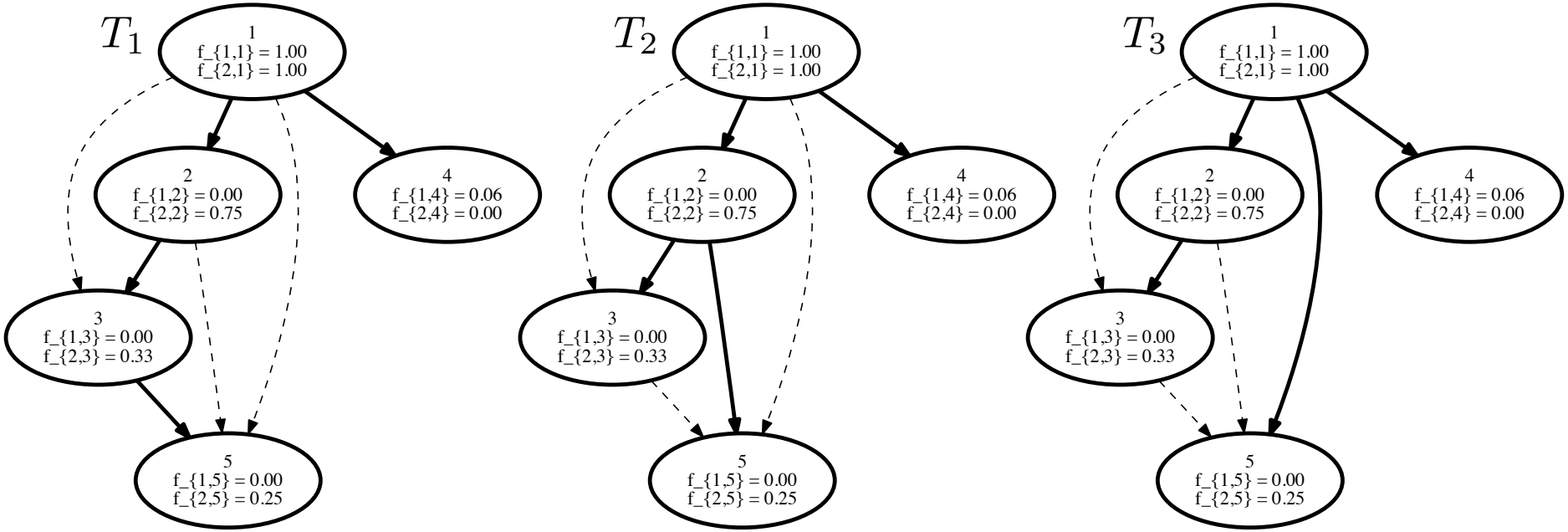


Theorem 1:
T is a solution to the PPM if and only if ***T*** is a spanning tree of ***G*** satisfying the sum condition

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

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

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 sample solutions uniformly at random?

On the Complexity of #PPM (new results)

Question 1: Can we determine the number of solutions?

Question 2: Can 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 (new results)

Question 1: Can we determine the number of solutions?

Question 2: Can 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 (new results)

Question 1: Can we determine the number of solutions?

Question 2: Can 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***

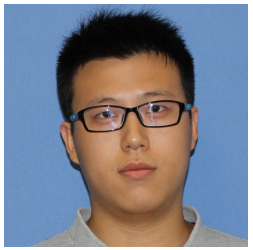
#P is the complexity class of counting problems whose decision problems are in NP

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

Background and theory:

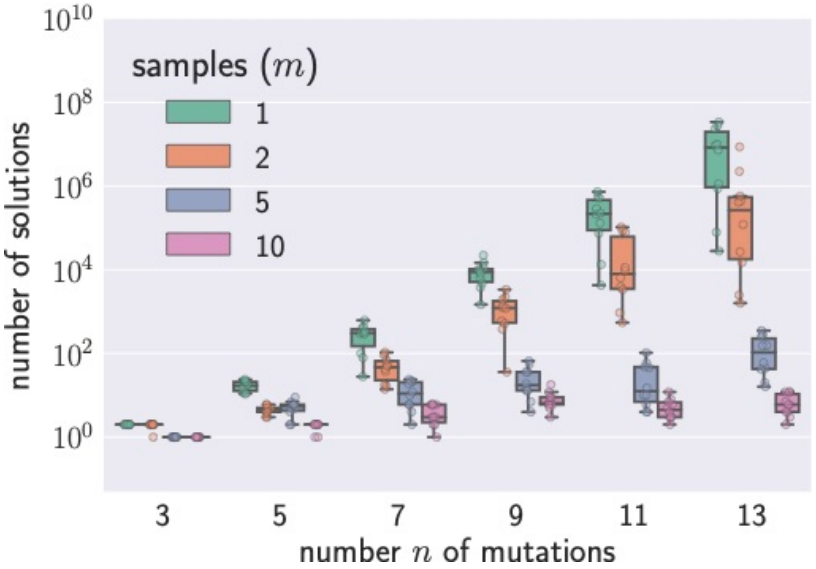
- Perfect Phylogeny Mixture (PPM) problem
- Combinatorial characterization of solutions
- #PPM: exact counting and uniform sampling

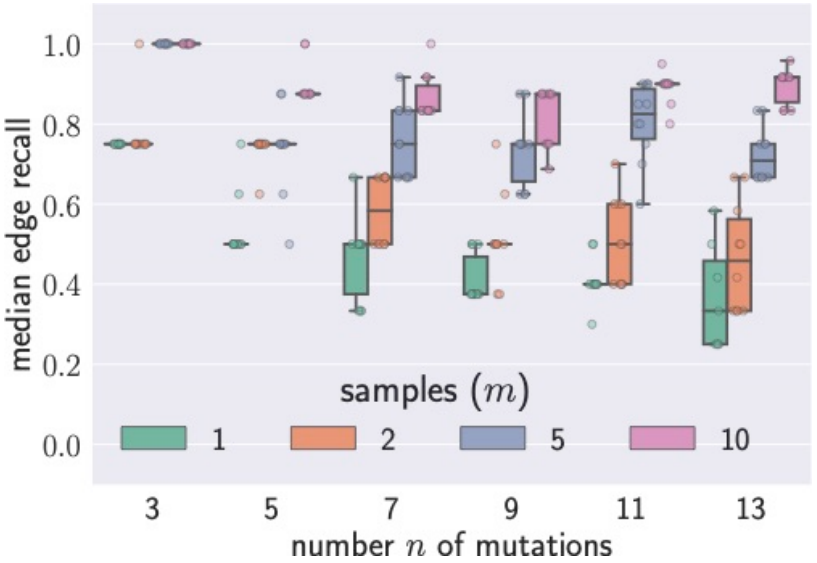
Simulation results:

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

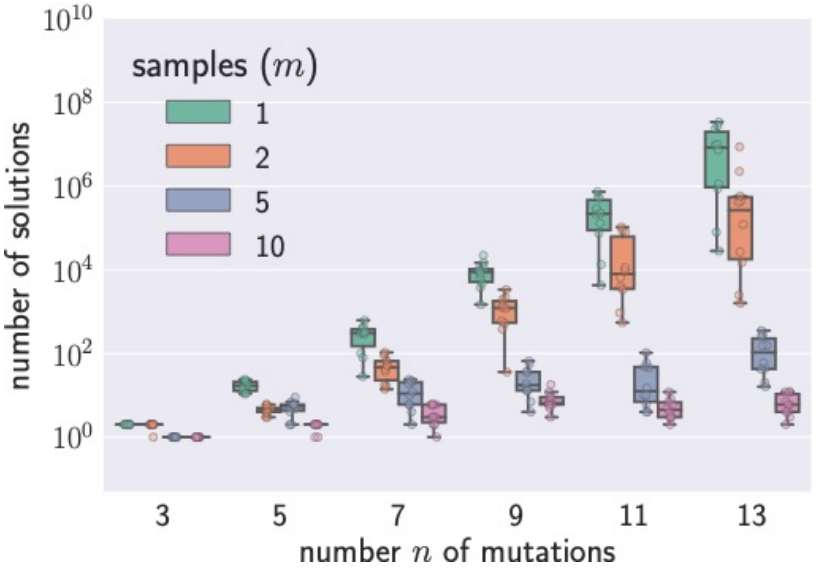
Dikshant Pradhan

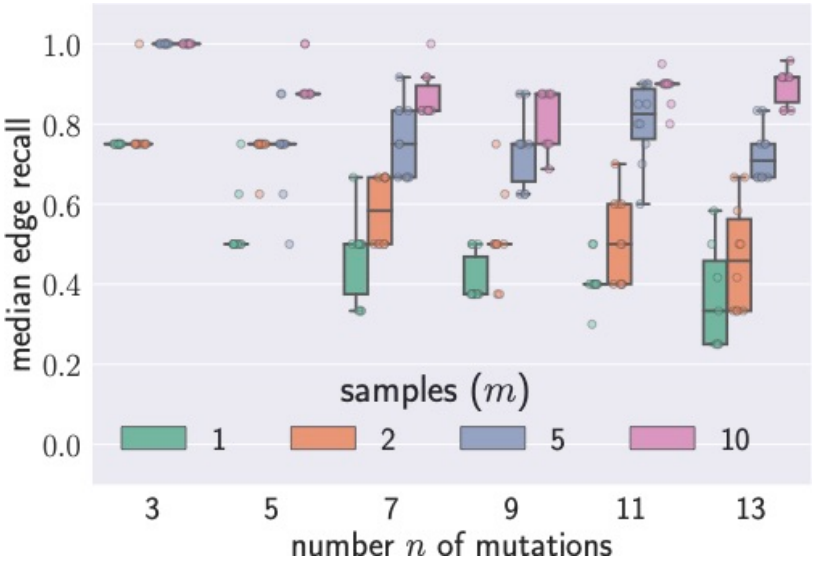
What Contributes to Non-uniqueness?

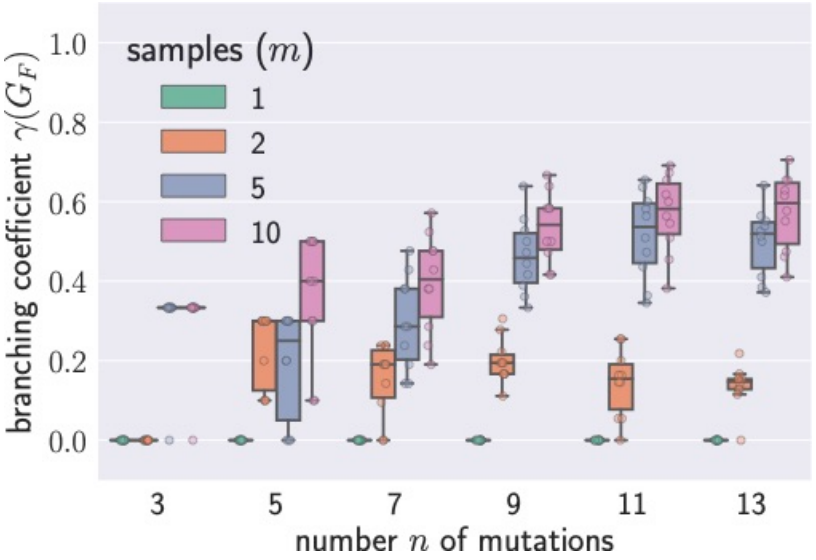


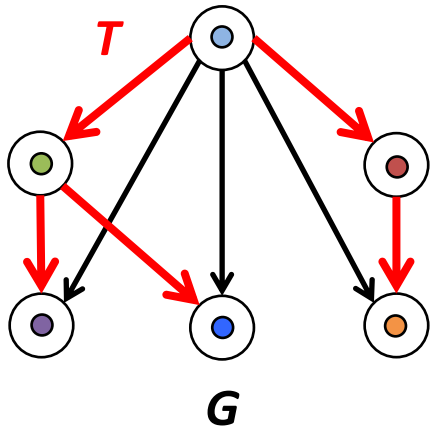


What Contributes to Non-uniqueness?

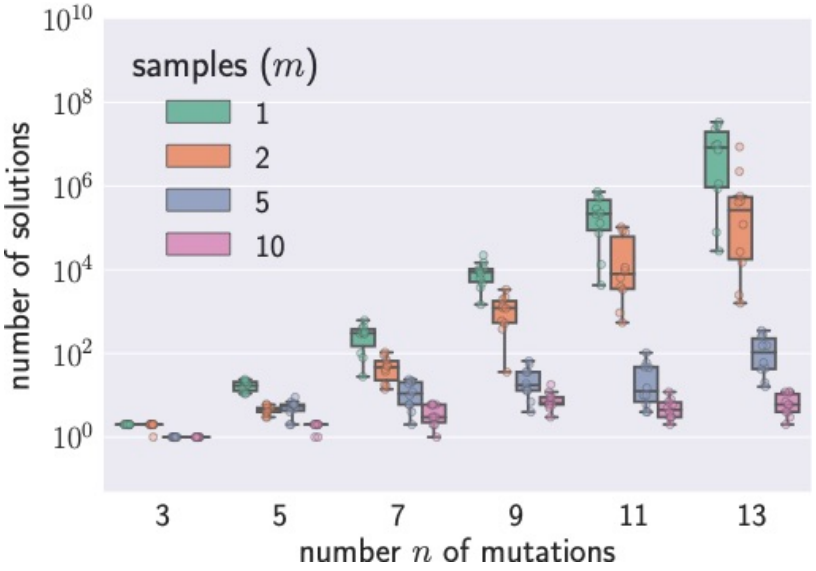


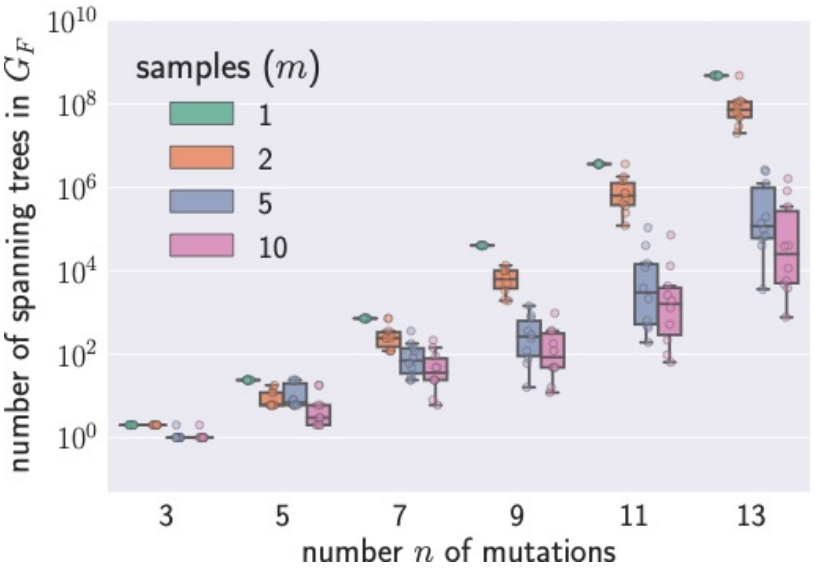


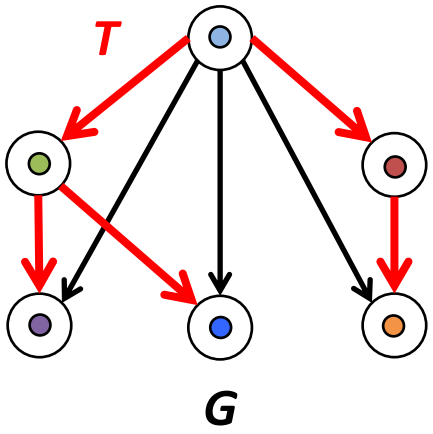




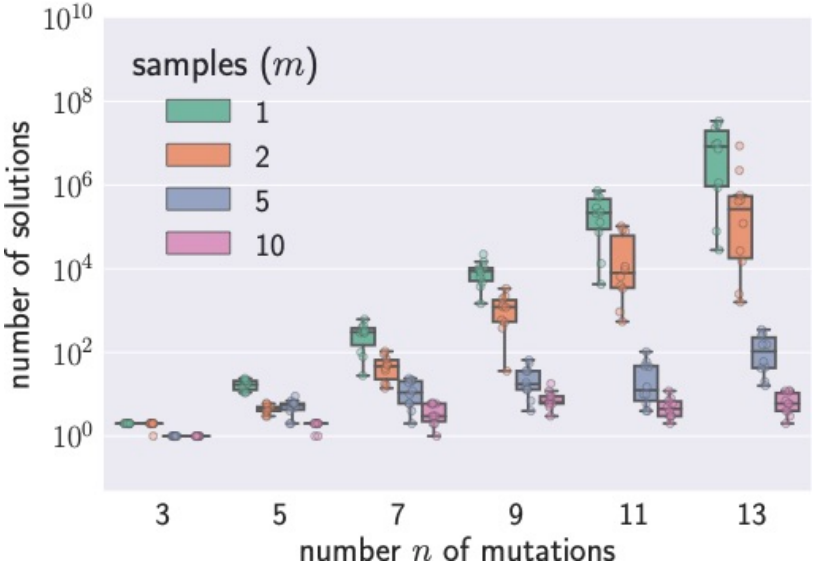
An Upper Bound for Number of Solutions

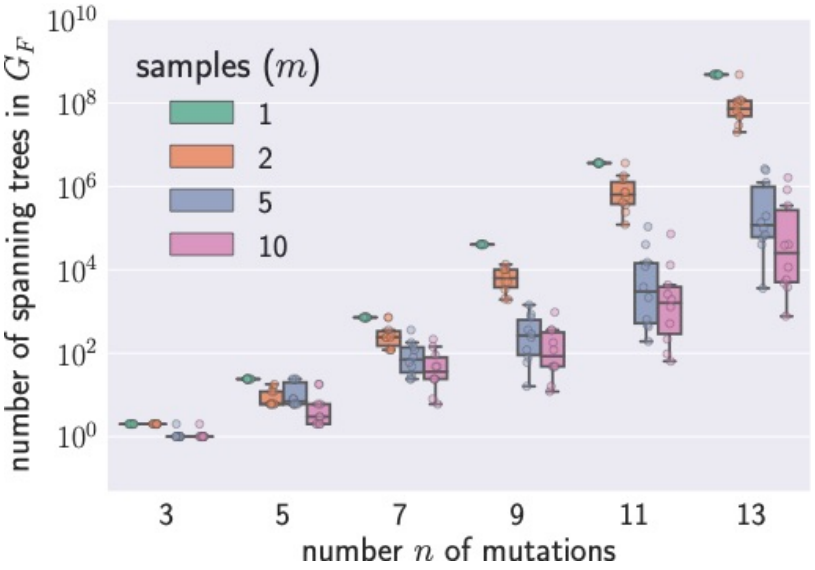


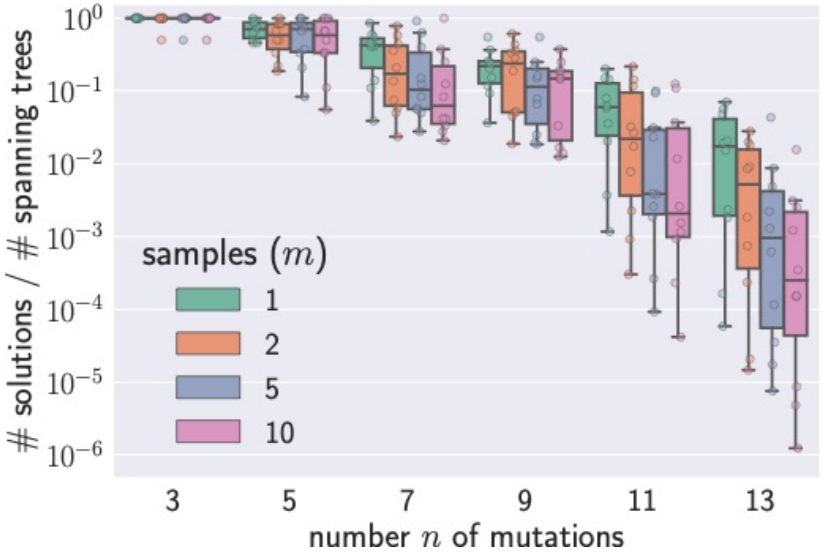


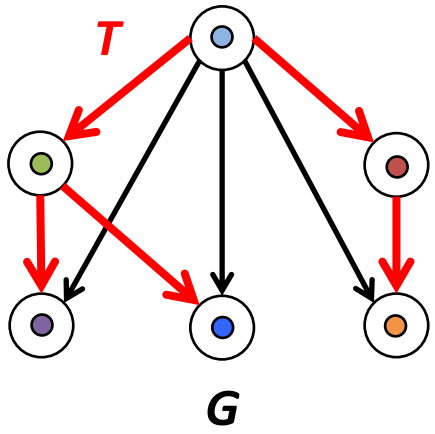


An Upper Bound for Number of Solutions

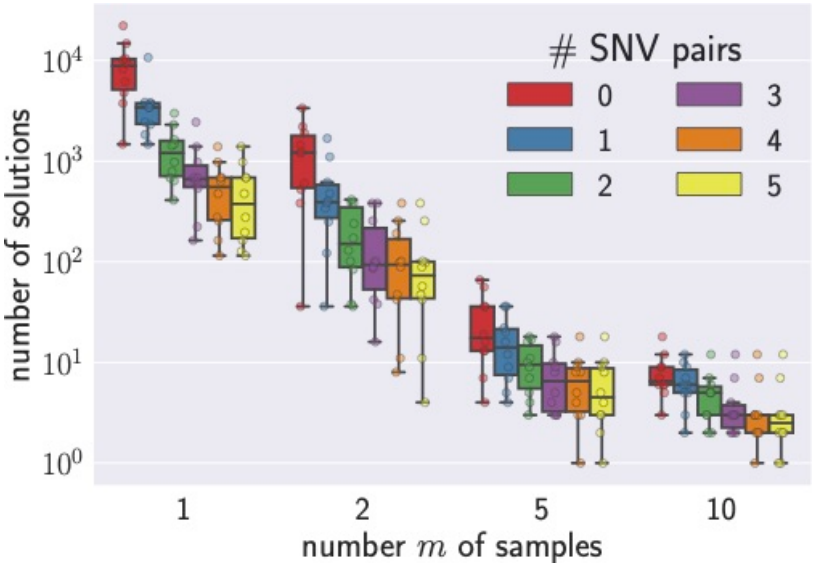


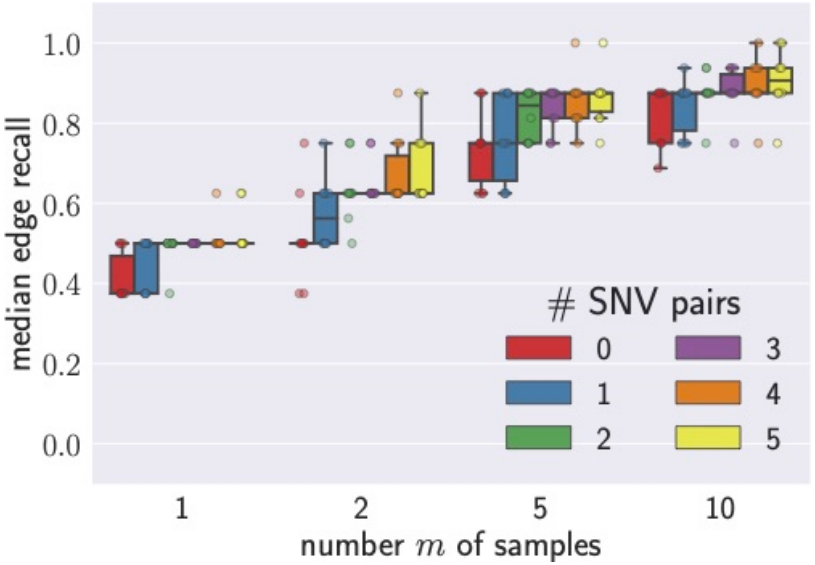


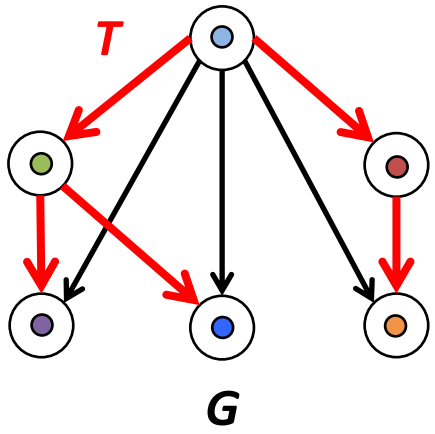




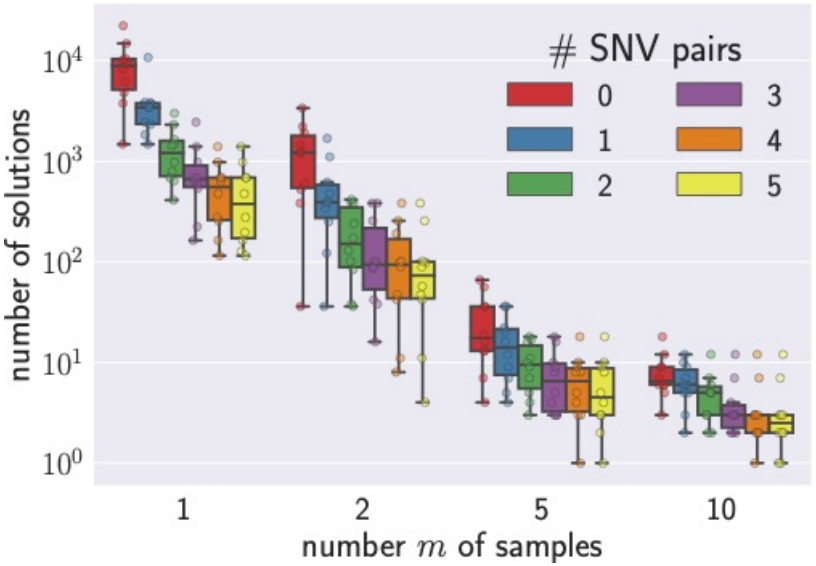
How to Reduce Non-Uniqueness?

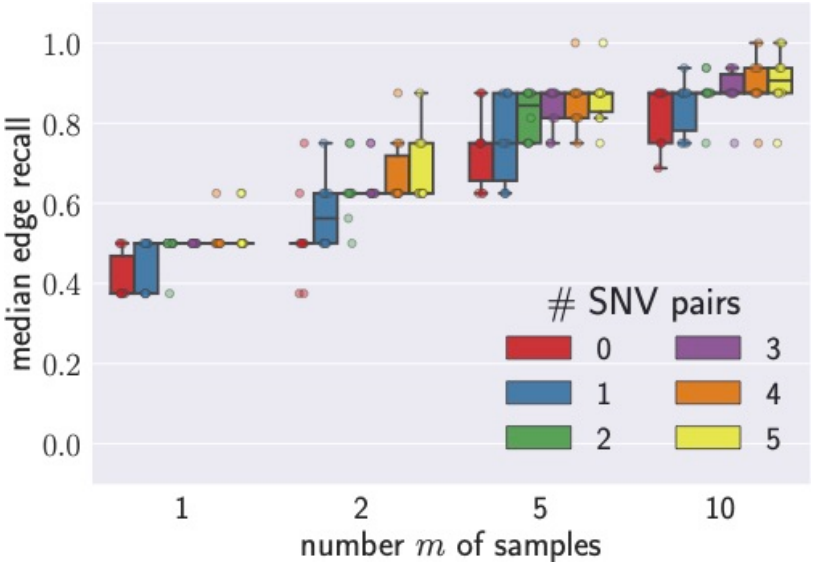


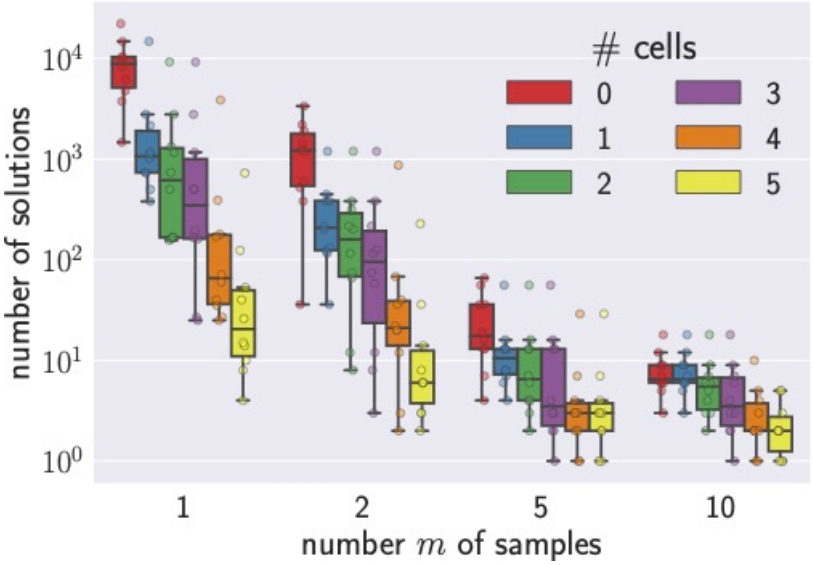


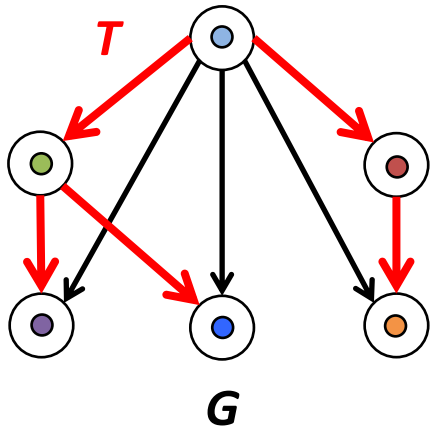


How to Reduce Non-Uniqueness?

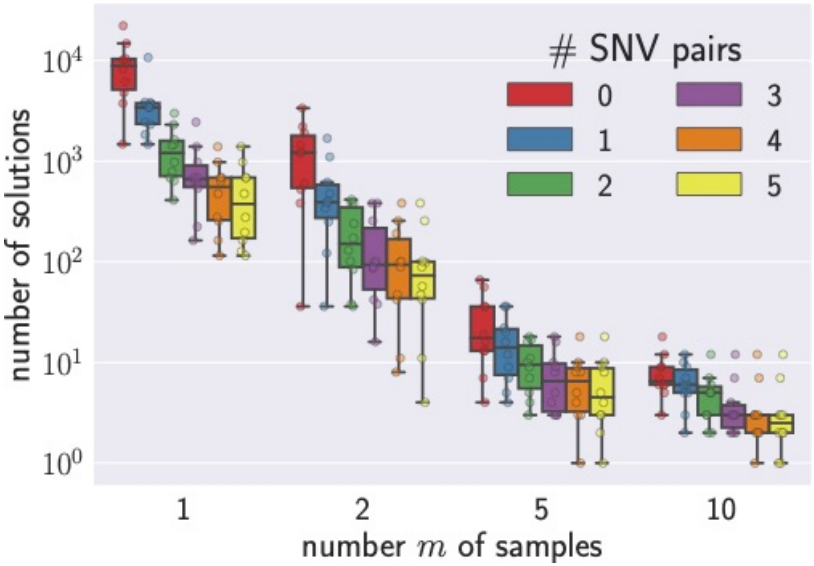


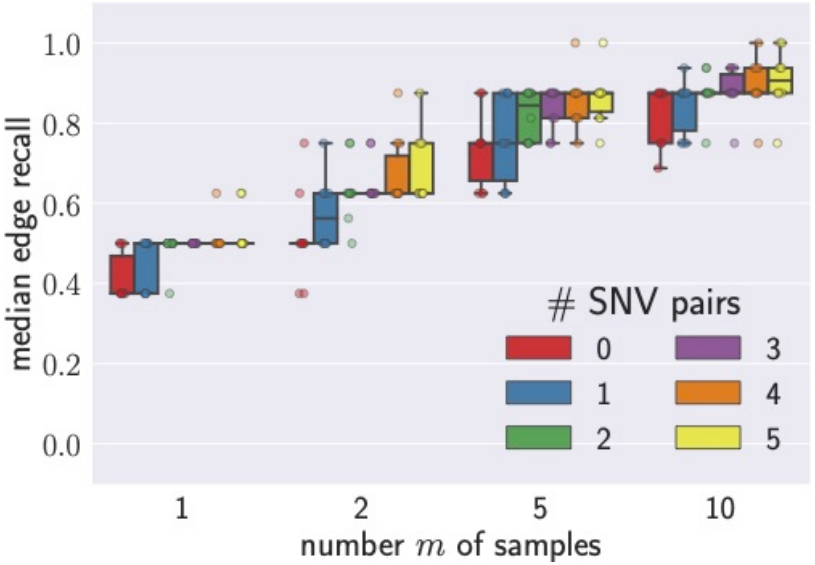


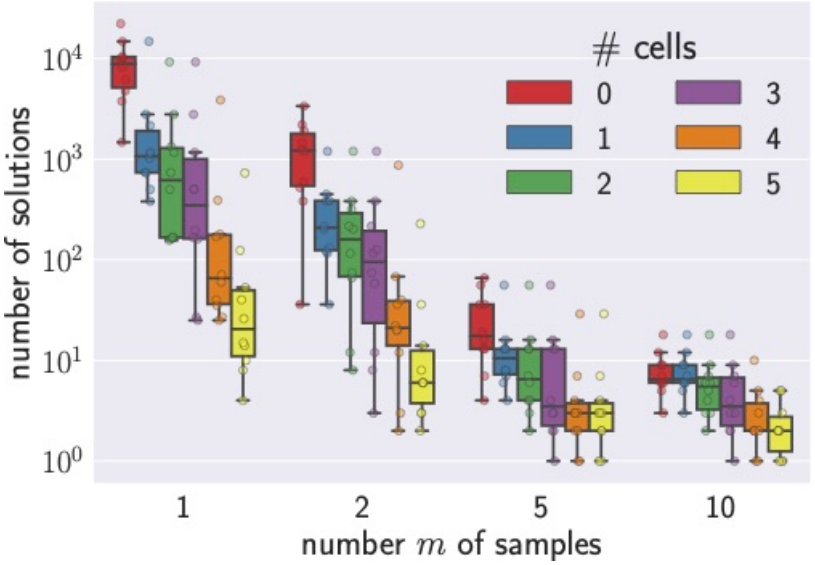


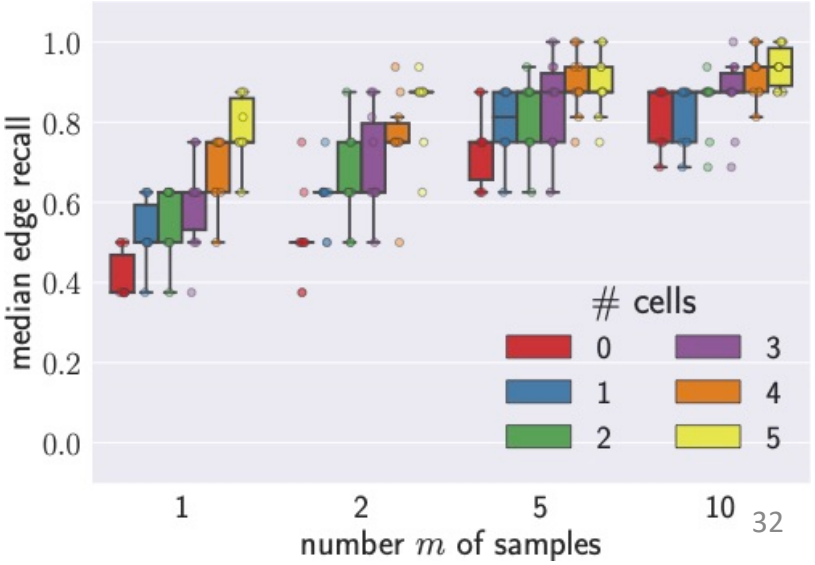


How to Reduce Non-Uniqueness?





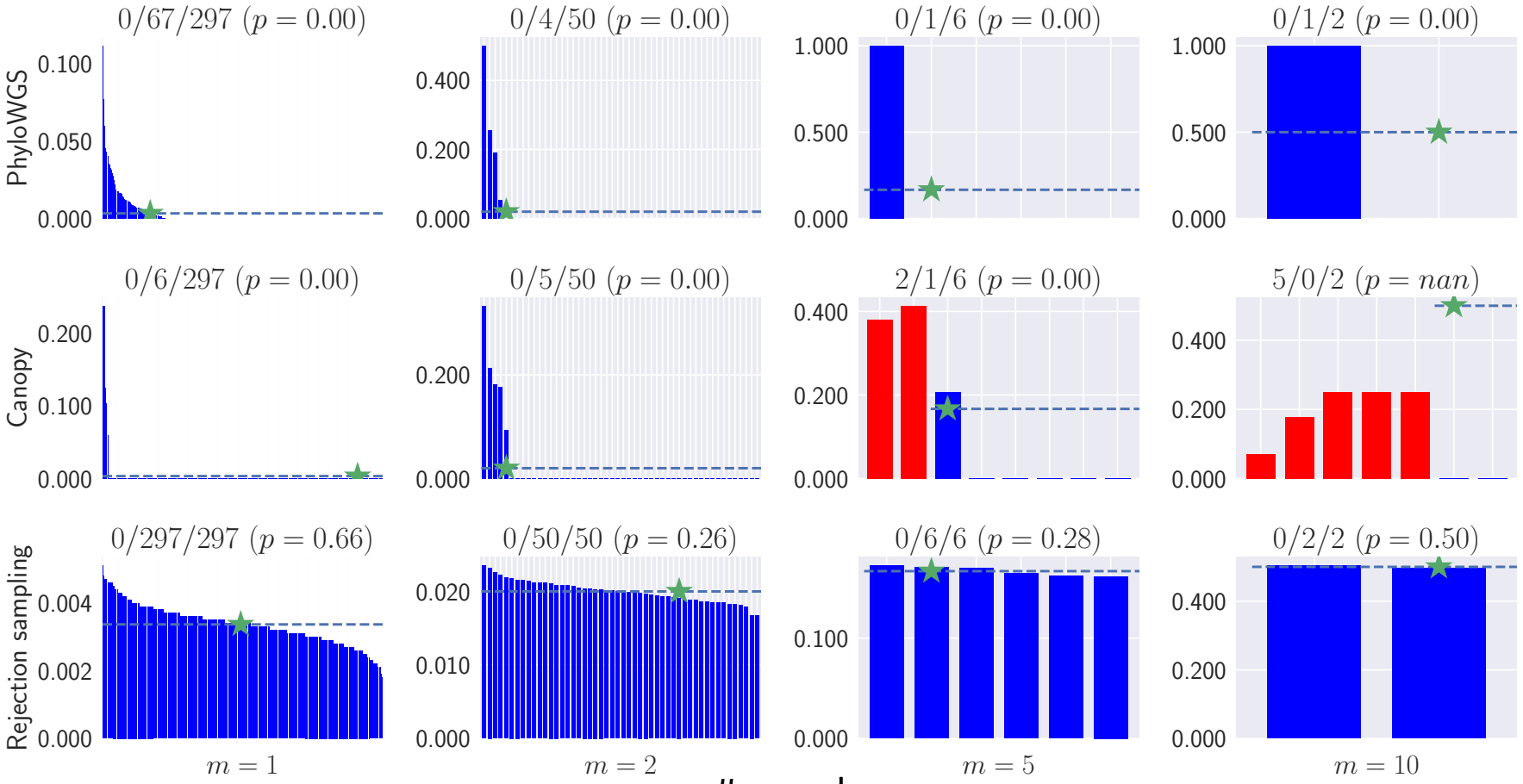


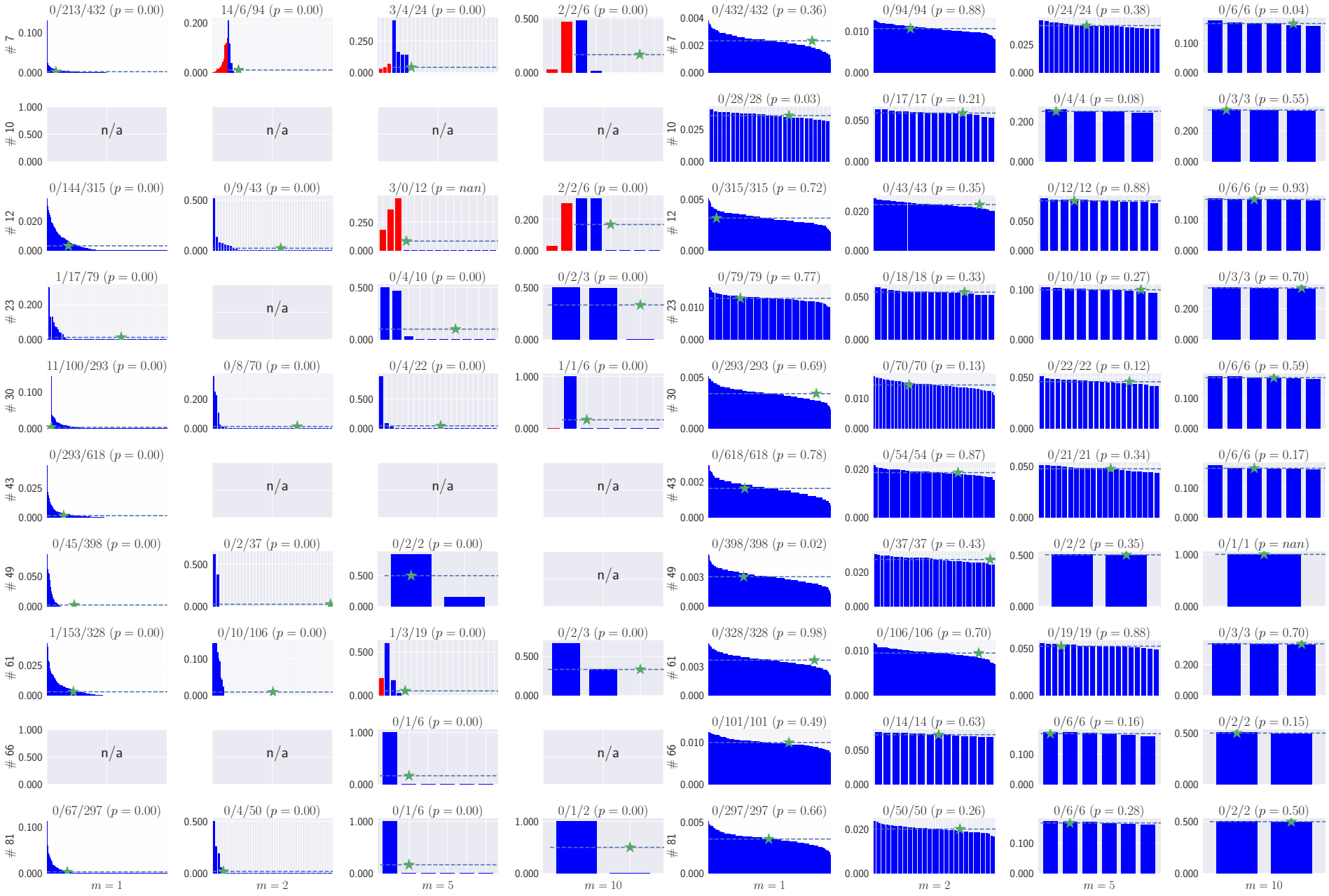


How Does Non-uniqueness affect Methods?

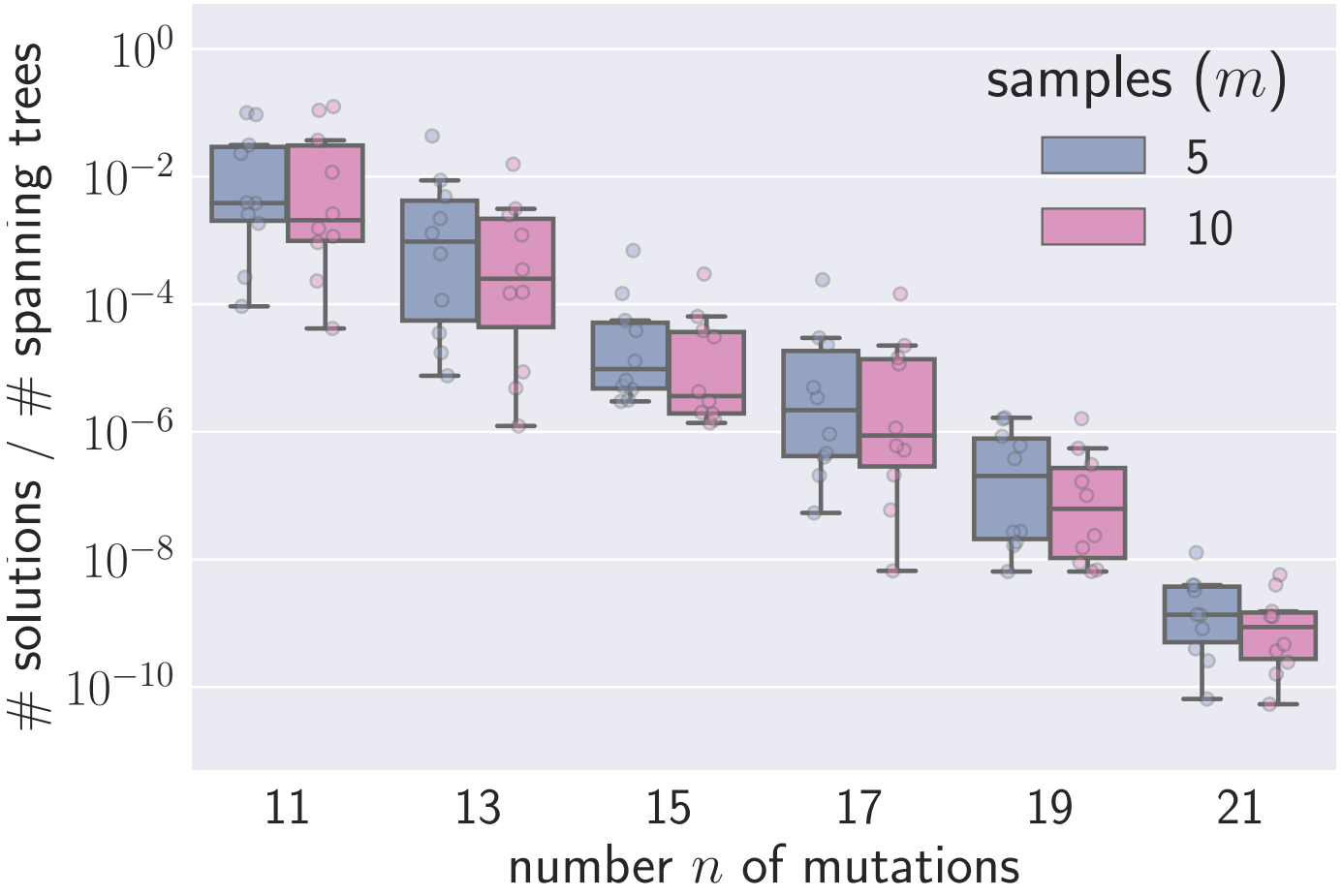
Two current MCMC methods using default parameters:

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



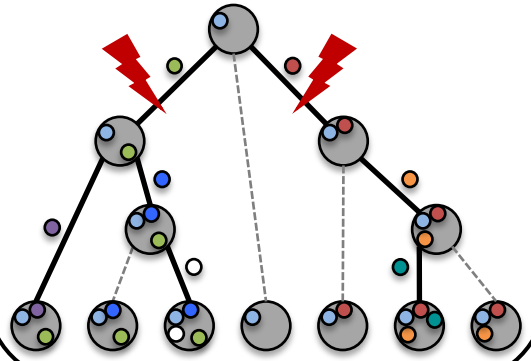

PhyloWGS
Rejection Sampling
0/213/432 (p = 0.00)
14/6/94 (p = 0.00)
3/4/24 (p = 0.00)
2/2/6 (p = 0.00)
0/432/432 (p = 0.36)
0/94/94 (p = 0.88)
0/24/24 (p = 0.38)
0/6/6 (p = 0.04)
7
n/a
0/28/28 (p = 0.03)
0/17/17 (p = 0.21)
0/4/4 (p = 0.08)
0/3/3 (p = 0.55)
10
0/144/315 (p = 0.00)
0/9/43 (p = 0.00)
3/0/12 (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
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)
23
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
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)
43
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
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)
61
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)
66
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

Rejection Sampling Does Not Scale

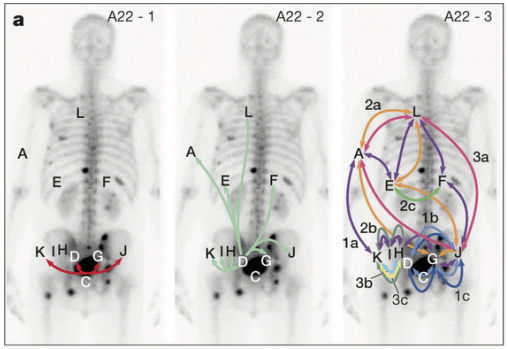


Challenge

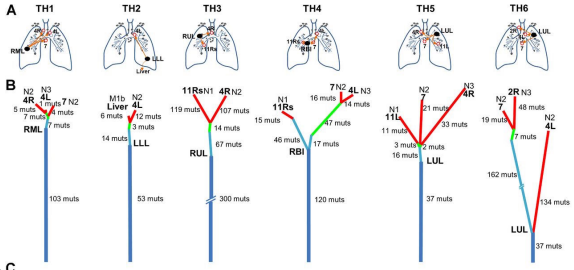
Identify targets for treatment



Understand metastatic development



Recognize common patterns of tumor evolution across patients



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

Key challenge:

Novel algorithms that sample uniformly at random from the space of PPM solutions

Conclusion

Background and theory:

- Perfect Phylogeny Mixture (PPM) problem
- Combinatorial characterization of solutions
- #PPM: exact counting and uniform sampling

Simulation results:

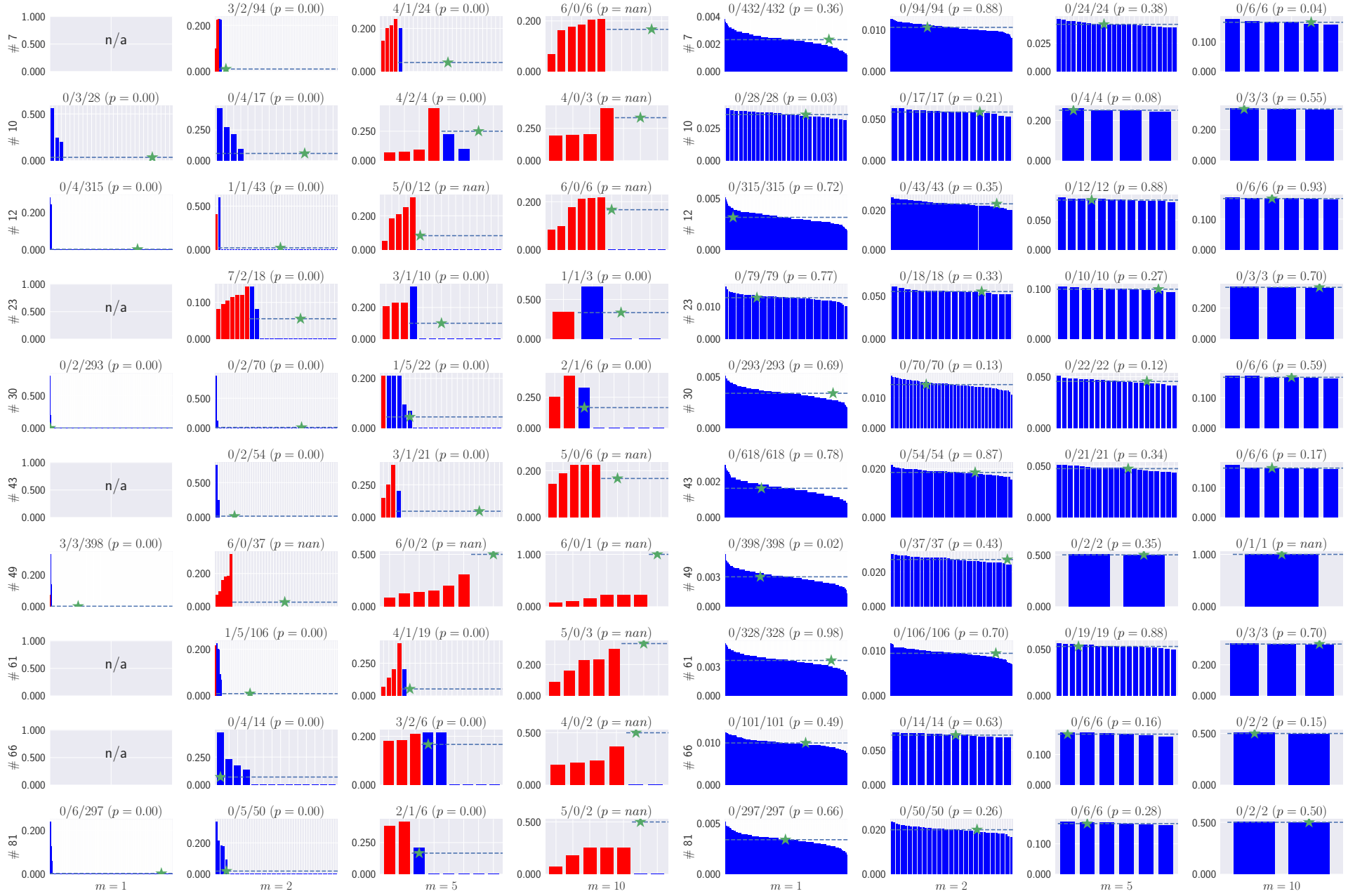
- What contributes to non-uniqueness?
- How to reduce non-uniqueness?
- How does non-uniqueness affect current methods?

Future Work

- Better sampling and counting algorithms
- Non-uniqueness in an infinite setting: characterize statistical consistency
- Constrained tumor phylogeny inference
  - Metastasis [El-Kebir, Satas & Raphael, Nature Genetics, 2018]
  - SCS + bulk, long-read sequencing
  - Other constraints…

Acknowledgments:

- Experiments were run on NCSA's Blue Waters supercomputer


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