

CS 598MEB

Introduction to Bioinformatics

Lecture 7

Mohammed El-Kebir

February 5, 2019



Outline

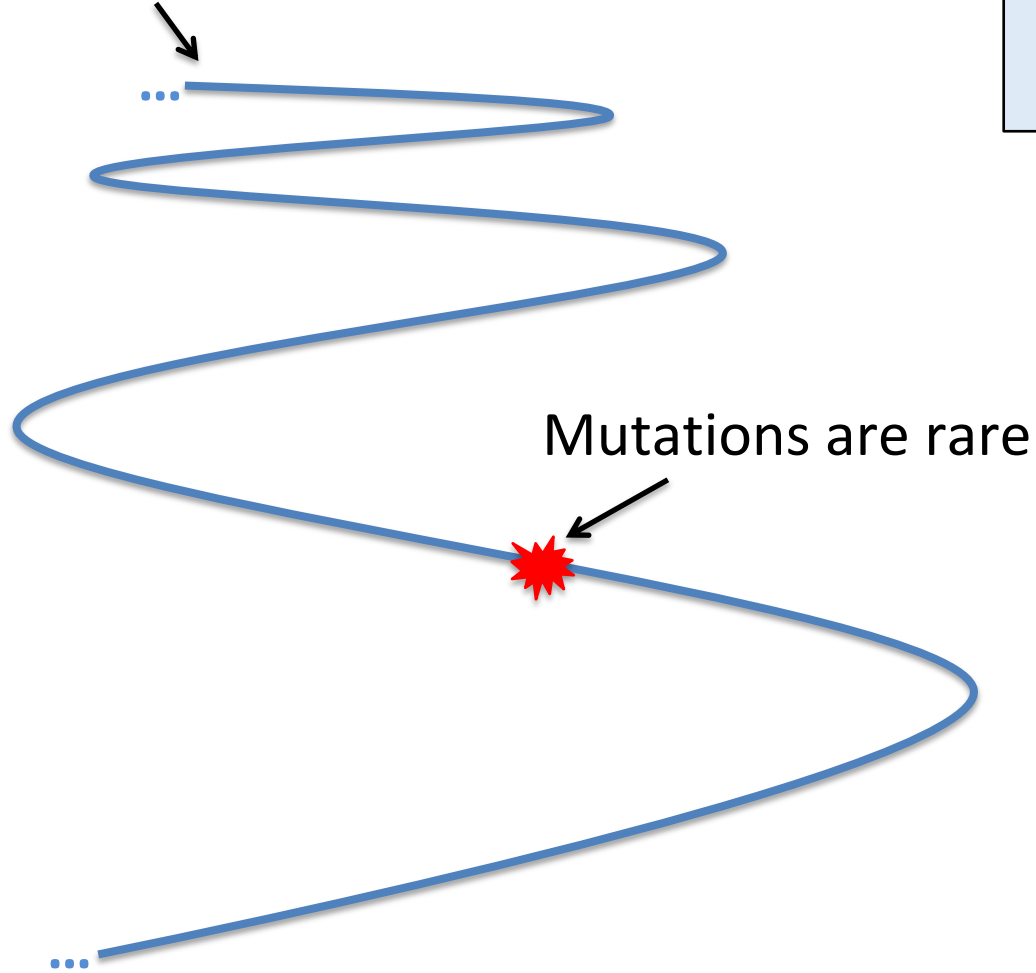
- Recap
- Tumor Phylogeny Inference from Bulk DNA-seq with Copy-Number Aberrations

Reading:

- **M. El-Kebir**, G. Satas , L. Oesper and B.J. Raphael. Inferring the Mutational History of a Tumor using Multi-State Perfect Phylogeny Mixtures. [Cell Systems, 3\(1\):43-53, 2016.](#)

Infinite Sites Model = Two-state Perfect Phylogeny

The genome is large



[Kimura, 1969]

Infinite sites model: multiple mutations never occur at the same position

Mutated Loci

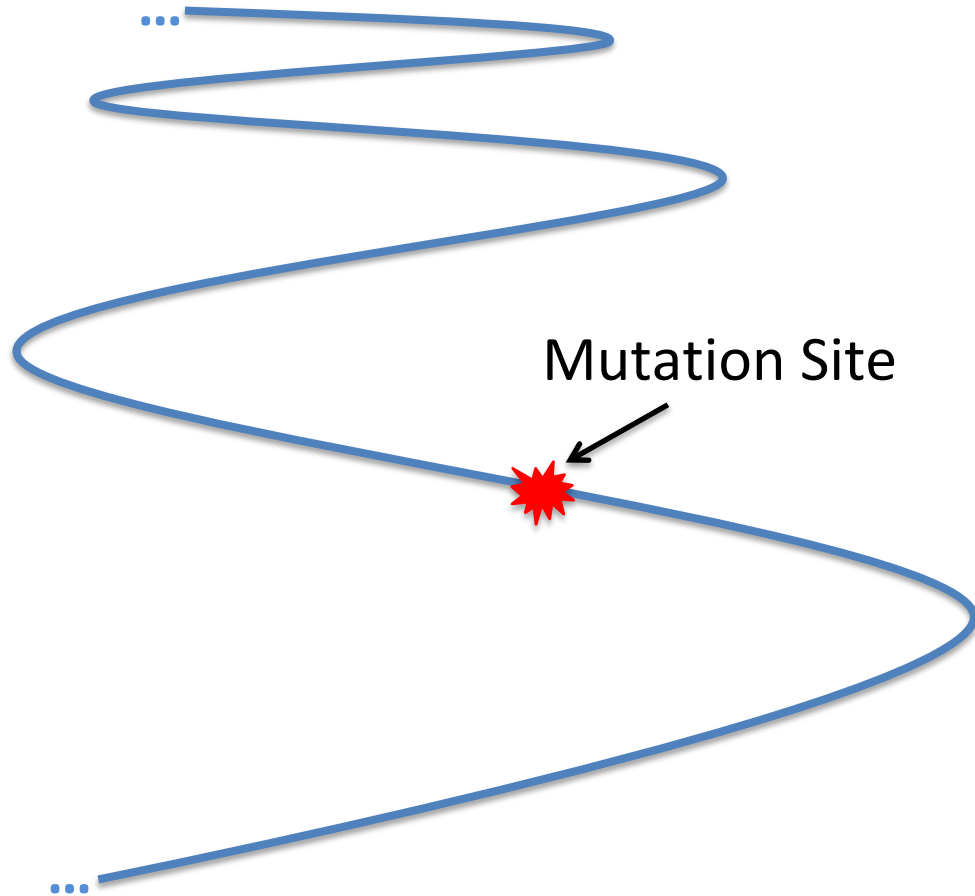
Species (cancer cells)	Red	Blue	Green	Purple	Orange	Yellow
A	0	0	0	0	1	1
B	0	0	0	1	1	1
C	0	0	1	0	1	0
D	1	0	0	0	0	0
E	1	1	0	0	0	0

1: mutated

0: not

All sites are bi-allelic: mutated or not.

Infinite Alleles Model = Multi-state Perfect Phylogeny



Infinite alleles model:

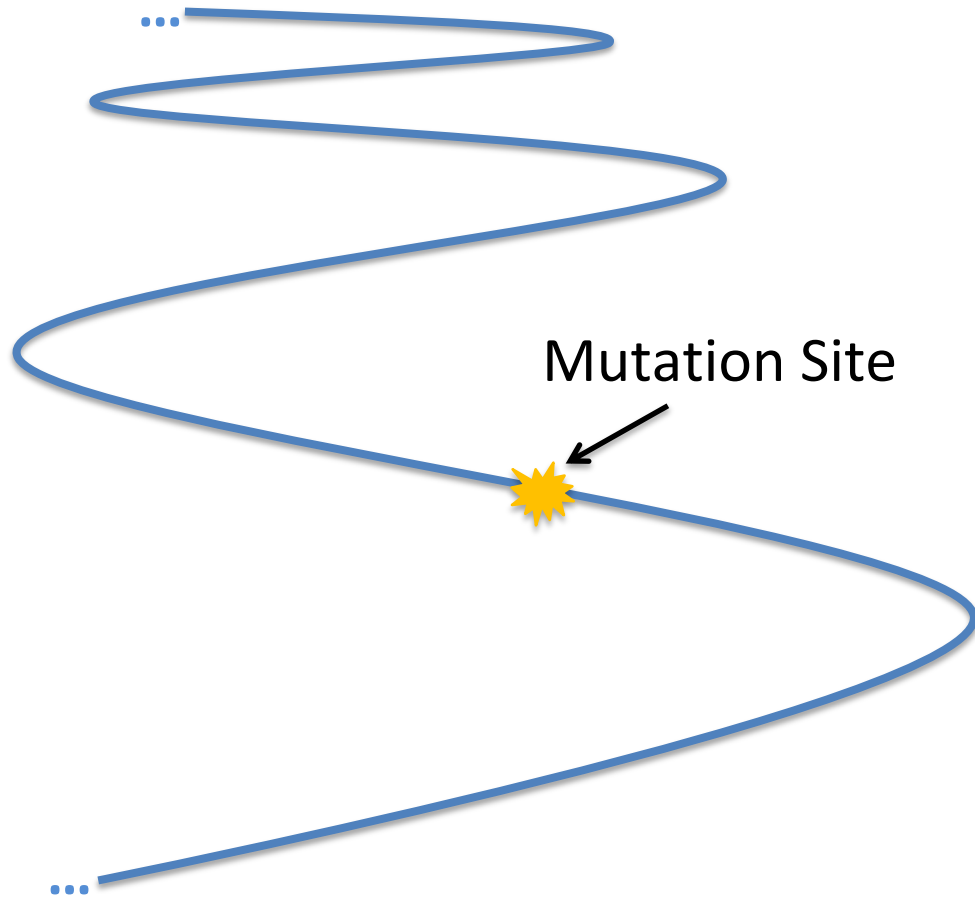
- For any mutation, there are an infinite number of possibilities of what mutation looks like (states).
- So, the same position can be mutated multiple times, but it never mutates to the same “allele” or state.

Site History:



Characters have integer states

Infinite Alleles Model = Multi-state Perfect Phylogeny



Infinite alleles model:

- For any mutation, there are an infinite number of possibilities of what mutation looks like (states).
- So, the same position can be mutated multiple times, but it never mutates to the same “allele” or state.

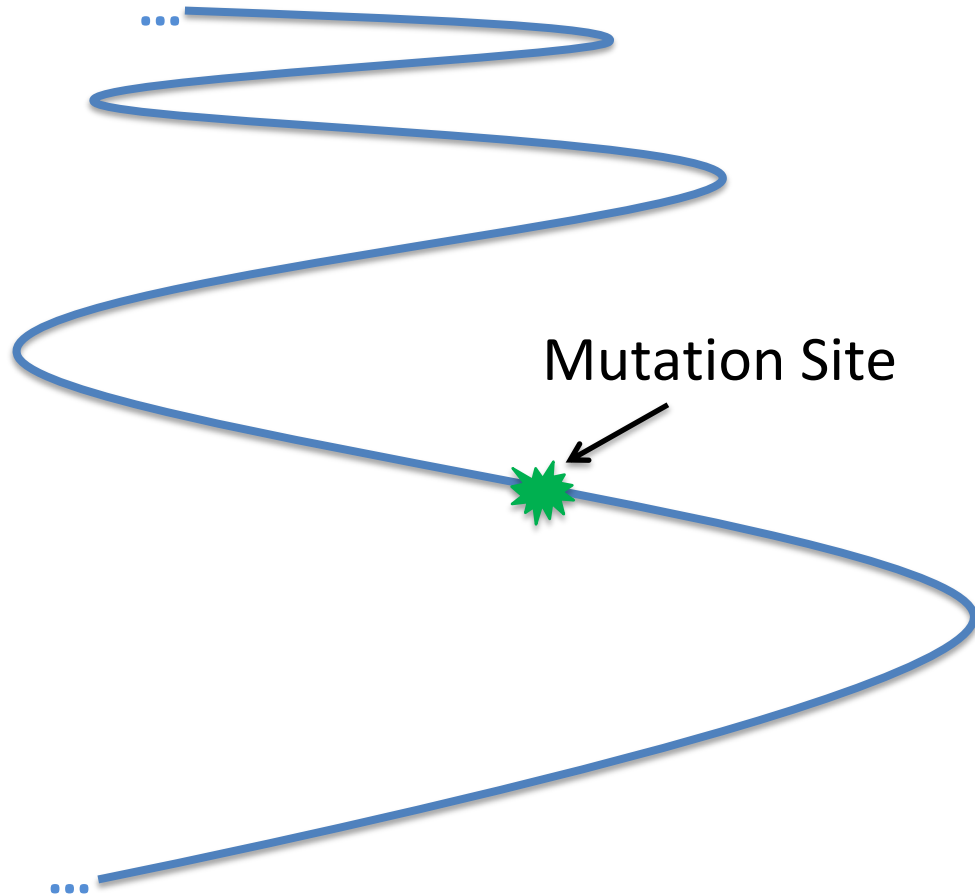
Site History:



Time

Characters have integer states

Infinite Alleles Model = Multi-state Perfect Phylogeny



Infinite alleles model:

- For any mutation, there are an infinite number of possibilities of what mutation looks like (states).
- So, the same position can be mutated multiple times, but it never mutates to the same “allele” or state.

Site History:



Characters have integer states

Outline

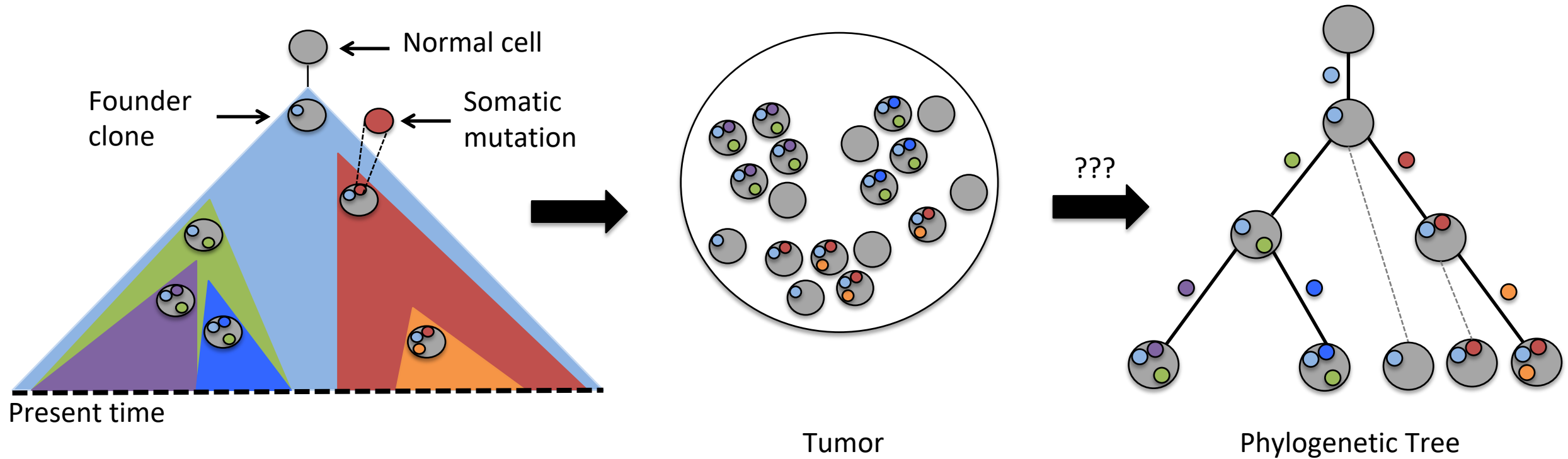
- Recap
- Tumor Phylogeny Inference from Bulk DNA-seq with Copy-Number Aberrations

Reading:

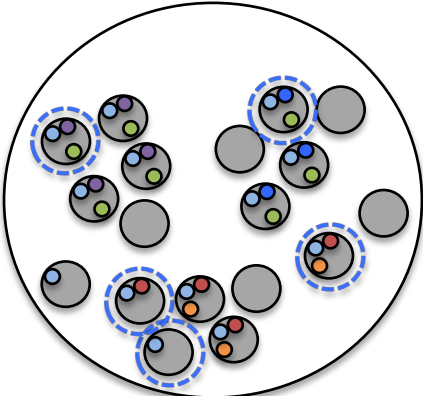
- **M. El-Kebir**, G. Satas , L. Oesper and B.J. Raphael. Inferring the Mutational History of a Tumor using Multi-State Perfect Phylogeny Mixtures. [Cell Systems, 3\(1\):43-53, 2016.](#)

Tumor Evolution as a Phylogenetic Tree

Clonal Theory [Nowell, 1976]

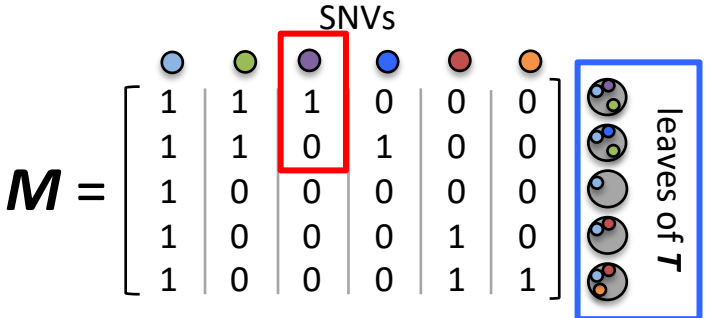


Observations are Leaves of a Perfect Phylogeny T



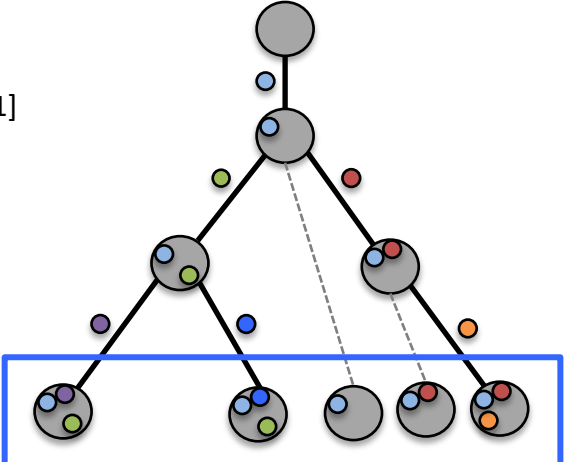
Tumor Snapshot

Single-cell sequencing



Binary Matrix B

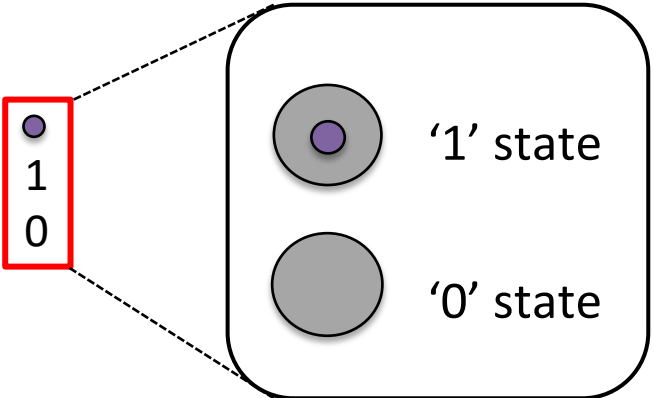
$O(mn)$
[Gusfield, 1991]



Two-State Perfect Phylogeny Tree T

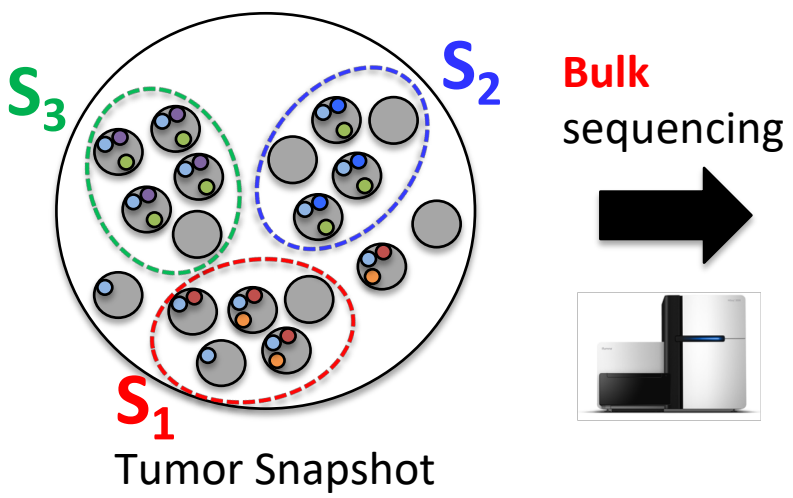
Assumptions:

- Mutations are single nucleotide variants (SNVs)
- Infinite sites assumption



Seq. method	Inferring T	Complexity
single-cell	unmixed two-state perfect phylogeny	$O(mn)$

Observations are Mixtures of the Leaves of T



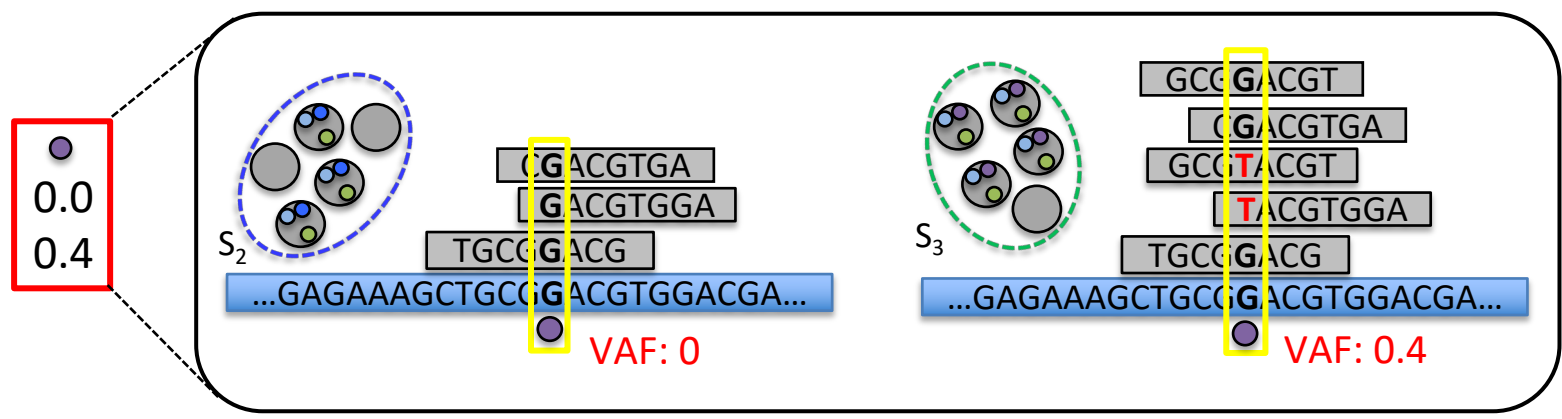
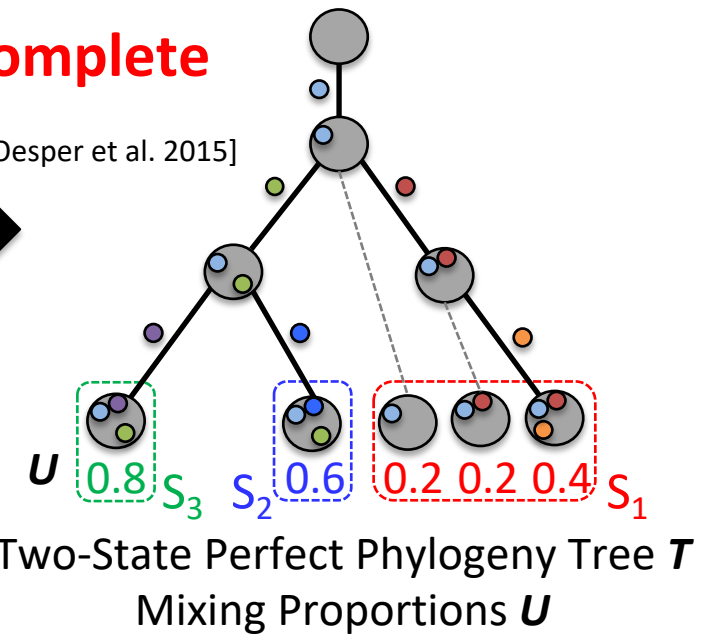
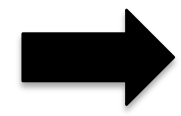
SNVs

$$F = \begin{bmatrix} 0.4 & 0.0 & 0.0 & 0.0 & 0.3 & 0.2 \\ 0.3 & 0.3 & 0.0 & 0.3 & 0.0 & 0.0 \\ 0.4 & 0.4 & 0.4 & 0.0 & 0.0 & 0.0 \end{bmatrix} \begin{matrix} S_1 \\ S_2 \\ S_3 \end{matrix} \text{ samples}$$

Variant Allele Frequencies Matrix F

NP-complete

AncesTree [El-Kebir, Oesper et al. 2015]



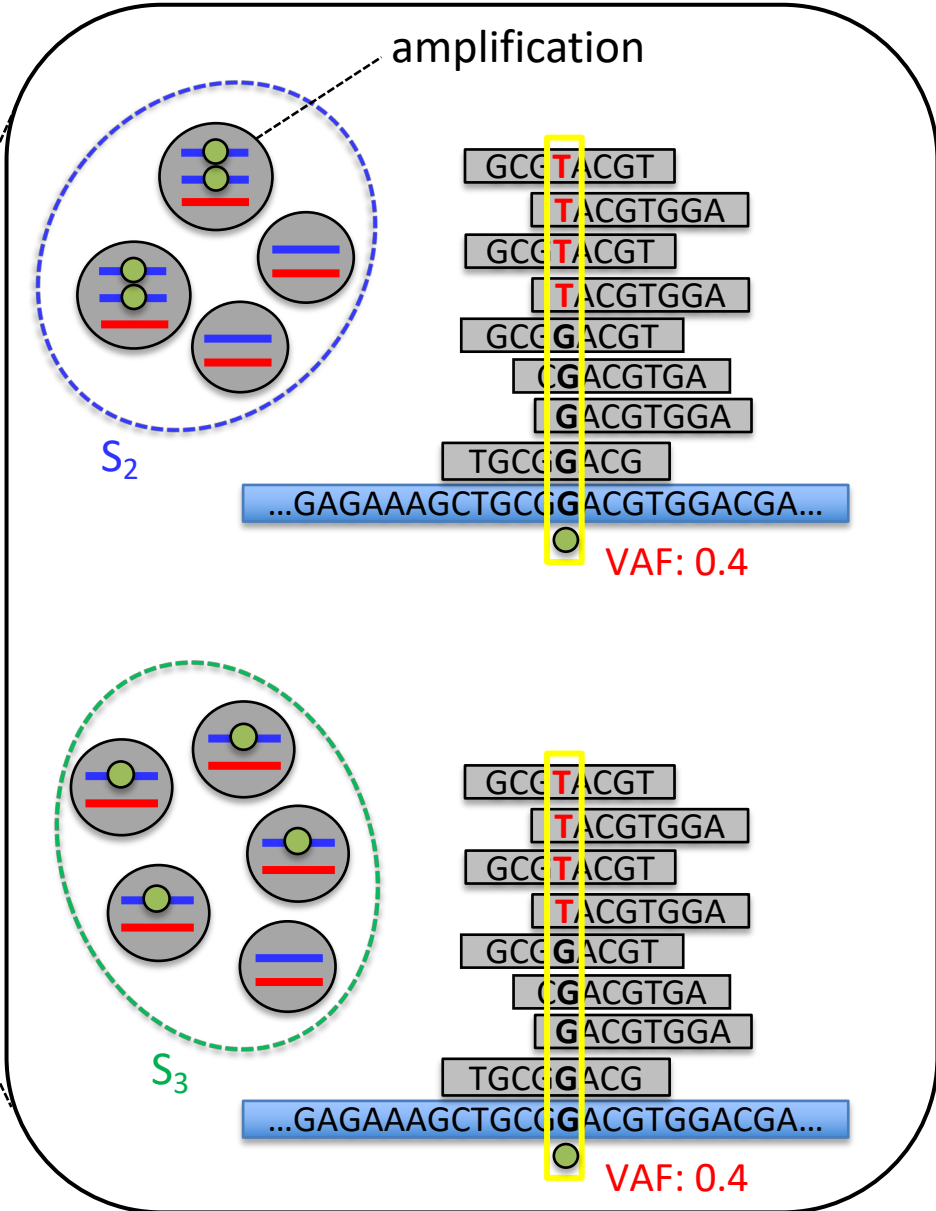
Seq. method	Inferring T	Complexity
single-cell	unmixed two-state perfect phylogeny	$O(mn)$
bulk	mixed two-state perfect phylogeny	NP-complete

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], ...

Copy-Number Aberrations Confound VAFs

SNVs

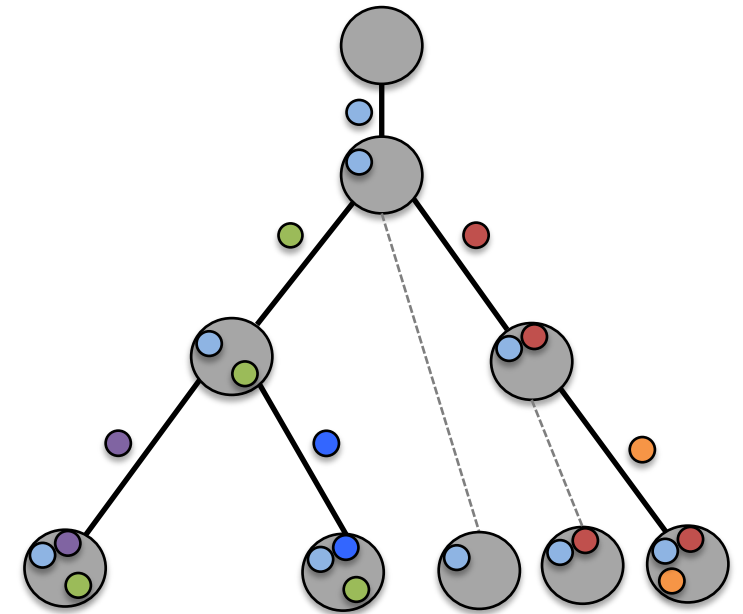
$F =$	0.4	0.0	0.0	0.0	0.3	0.2	s_1
	0.3	0.4	0.0	0.3	0.0	0.4	s_2
	0.4	0.4	0.4	0.0	0.0	0.4	s_3
							samples



- Need > 2 states:**
- 0 : non-mutated
 - 1 : mutated
 - 2 : single-copy amplification
 - 3 : ...
 - ...

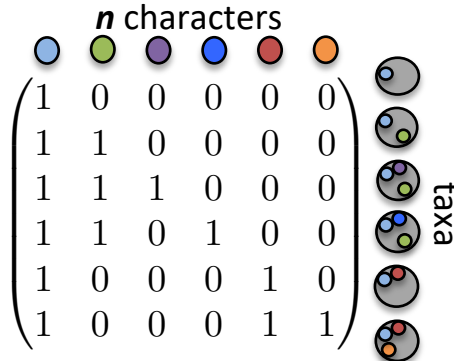
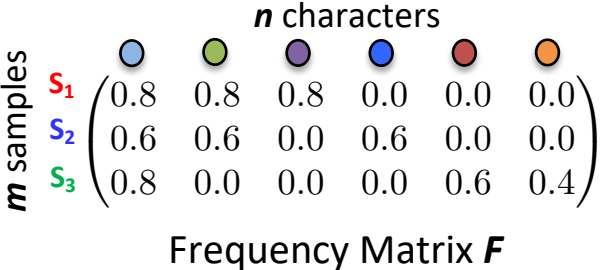
Outline

- Multi-State Perfect Phylogeny Mixture Problem
- Combinatorial Characterization of Solutions
- Application to Cancer Bulk-Sequencing Data
- Results

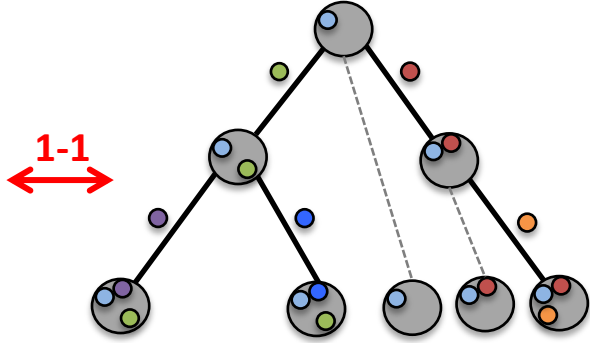


Infinite Sites Generalizes to Infinite Alleles

Two-State Perfect Phylogeny – Infinite sites assumption: a character changes state once



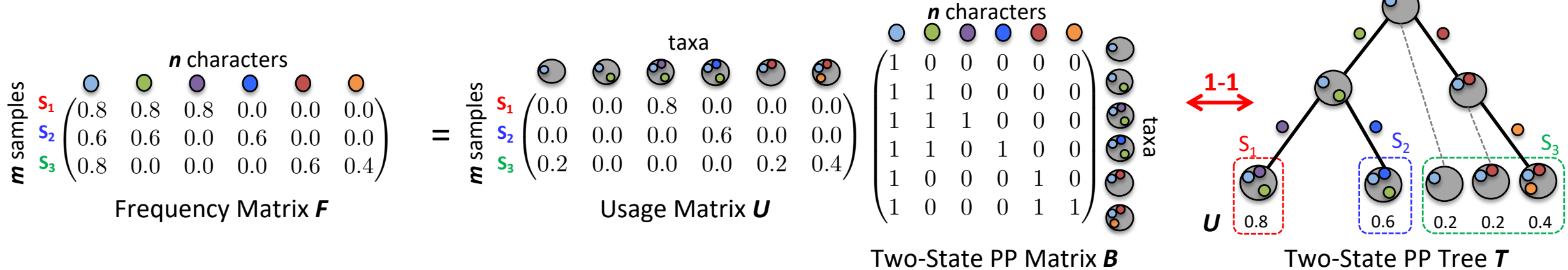
Two-State PP Matrix **B**



Two-State PP Tree **T**

Infinite Sites Generalizes to Infinite Alleles

Two-State Perfect Phylogeny – **Infinite sites assumption**: a character changes state once

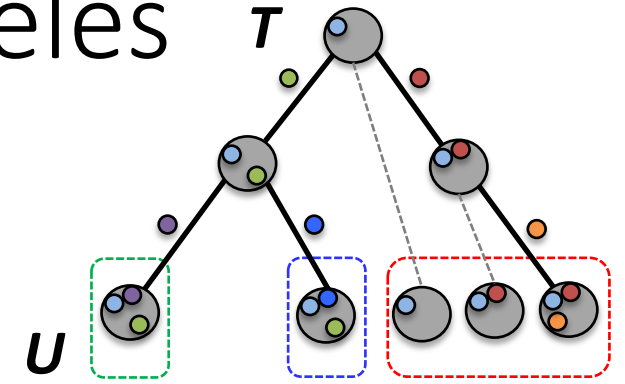


VAF Factorization Problem (VAFFP): Given F , find U and B such that $F = UB$ [El-Kebir, Oesper et al., 2015]

Infinite Sites Generalizes to Infinite Alleles

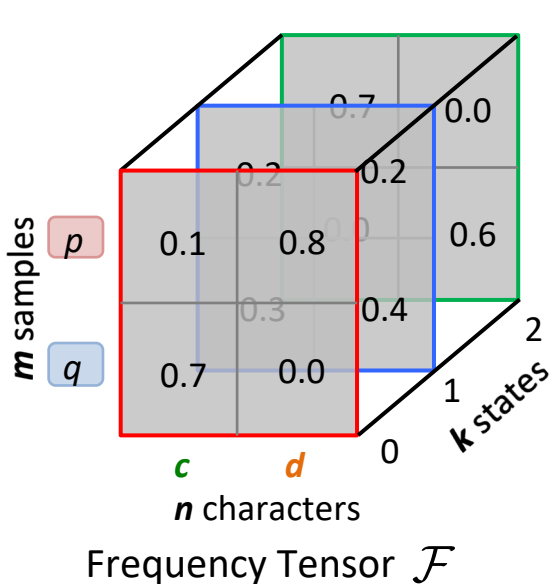
Two-State Perfect Phylogeny – **Infinite sites assumption**: a character changes state once

$$m \text{ samples} \begin{pmatrix} n \text{ characters} \\ \mathbf{F} \end{pmatrix} = \begin{pmatrix} \mathbf{U} \end{pmatrix} \begin{pmatrix} \mathbf{B} \end{pmatrix} \quad \longleftrightarrow \quad \mathbf{1-1}$$



VAF Factorization Problem (VAFFP): Given \mathbf{F} , find \mathbf{U} and \mathbf{B} such that $\mathbf{F} = \mathbf{U}\mathbf{B}$ [El-Kebir, Oesper et al., 2015]

Multi-State Perfect Phylogeny – **Infinite alleles assumption**: a character changes to a state once



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

\mathbf{F}_2

$$\begin{pmatrix} 0.2 & 0.2 \\ 0.3 & 0.4 \end{pmatrix}$$

\mathbf{F}_1

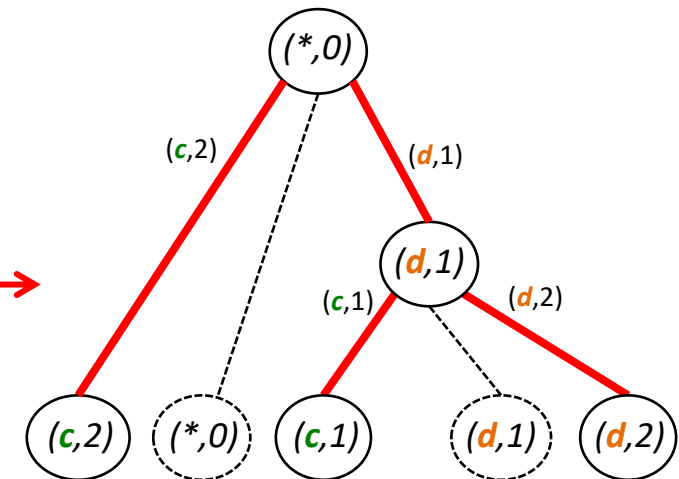
$$\begin{pmatrix} 0.1 & 0.8 \\ 0.7 & 0.0 \end{pmatrix}$$

\mathbf{F}_0

Frequency Tensor \mathcal{F}

$$\begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 0 \\ 0 & 1 \\ 0 & 2 \end{pmatrix}$$

Multi-State
PP Matrix \mathbf{A}



Multi-State Perfect Phylogeny Tree T

Infinite Sites Generalizes to Infinite Alleles

Two-State Perfect Phylogeny – **Infinite sites assumption**: a character changes state once

$$\begin{matrix} m \text{ samples} \\ \left(\begin{matrix} & n \text{ characters} \\ & \mathbf{F} \end{matrix} \right) = \left(\begin{matrix} & \mathbf{U} \end{matrix} \right) \left(\begin{matrix} & \mathbf{B} \end{matrix} \right) \xleftrightarrow{1-1} \mathbf{U}
 \end{matrix}$$

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

Multi-State Perfect Phylogeny – **Infinite alleles assumption**: a character changes to a state once

Frequency Tensor \mathcal{F}

$$\begin{matrix} \mathbf{F}_2 \\ \mathbf{F}_1 \\ \mathbf{F}_0 \end{matrix} = \begin{matrix} \mathbf{U} \\ \mathbf{U} \\ \mathbf{U} \end{matrix} \begin{matrix} \mathbf{A}_2 \\ \mathbf{A}_1 \\ \mathbf{A}_0 \end{matrix}$$

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} \mathbf{A}_2 \leftrightarrow \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 0 \\ 0 & 1 \\ 0 & 2 \end{pmatrix} \mathbf{A}$$

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 1 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} \mathbf{A}_1 \leftrightarrow \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 0 \\ 0 & 1 \\ 0 & 2 \end{pmatrix} \mathbf{A}$$

$$\begin{pmatrix} 1 & 1 \\ 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \mathbf{A}_0 \leftrightarrow \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 0 \\ 0 & 1 \\ 0 & 2 \end{pmatrix} \mathbf{A}$$

Multi-State PP Matrix \mathbf{A}

Multi-State Perfect Phylogeny Tree \mathbf{T}

Multi-state Perfect Phylogeny Mixture Problem: Given \mathbf{F} , find \mathbf{U} and \mathbf{A} such that $\mathbf{F}_i = \mathbf{U}\mathbf{A}_i$ for all states i

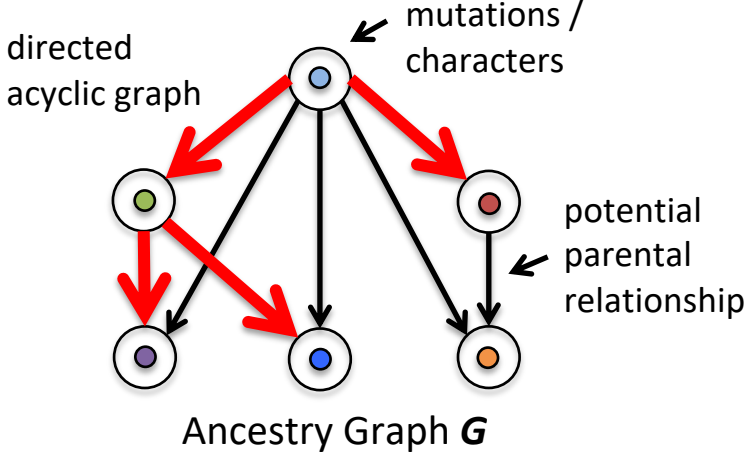
Combinatorial Characterization of Solutions

Two-State Perfect Phylogeny

n mutations

m samples						
	0.8	0.8	0.8	0.0	0.0	0.0
	0.6	0.6	0.0	0.6	0.0	0.0
	0.8	0.0	0.0	0.0	0.6	0.4

Frequency Matrix F



Theorem [El-Kebir, Oesper et al., 2015; Popic et al., 2015]
 Solutions are **spanning trees** that satisfy

$$f_{p,(c,1)} \geq \sum_{(d,1) \in \delta(c,1)} f_{p,(d,1)} \quad (SC)$$

Theorem [El-Kebir, Oesper et al., 2015]
 VAFFP is NP-complete for $m = O(n)$

Multi-State Perfect Phylogeny

m samples

n characters

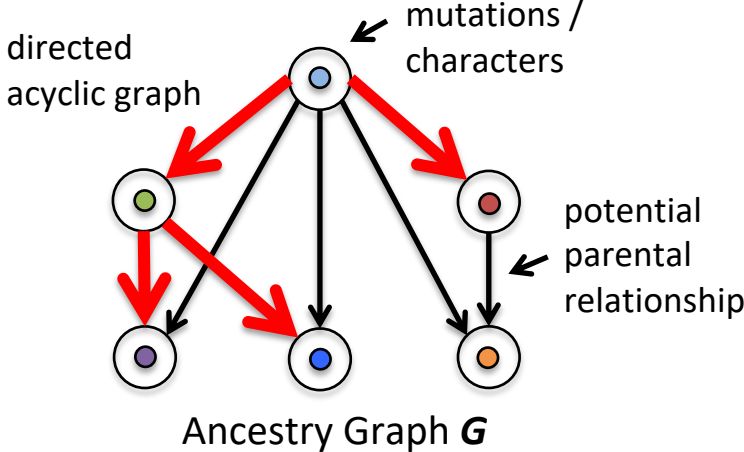
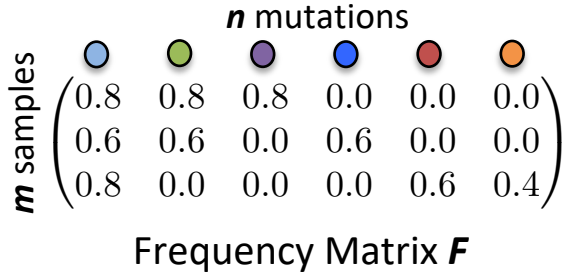
k states

p	0.1	0.8	0.0	0.6
q	0.7	0.0	0.4	0.2
	0.2	0.2	0.7	0.0
	0.3	0.4	0.0	0.2

Frequency Tensor \mathcal{F}

Combinatorial Characterization of Solutions

Two-State Perfect Phylogeny

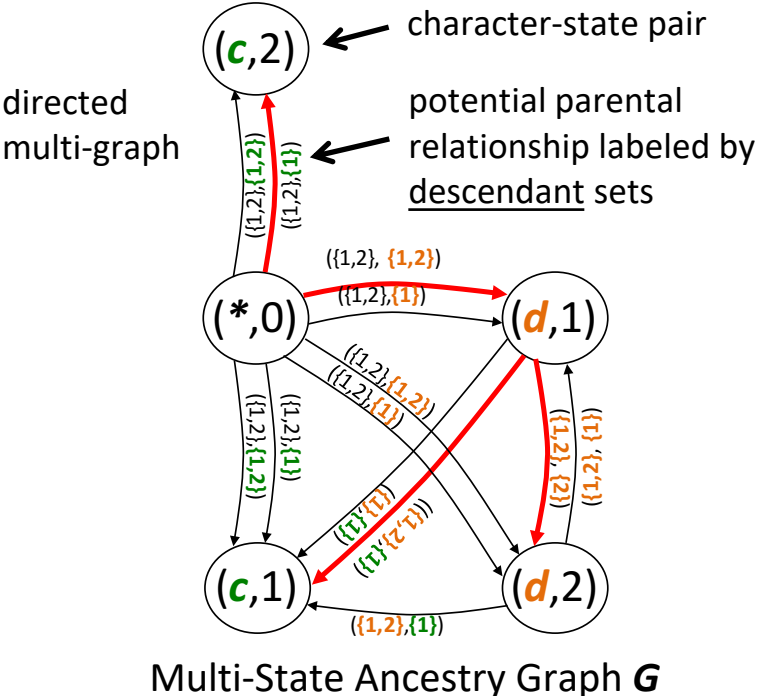
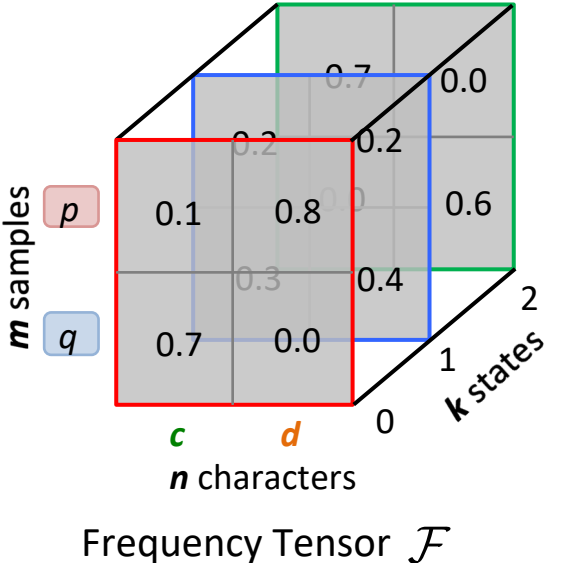


Theorem [El-Kebir, Oesper et al., 2015; Popic et al., 2015]
Solutions are **spanning trees** that satisfy

$$f_{p,(c,1)} \geq \sum_{(d,1) \in \delta(c,1)} f_{p,(d,1)} \quad (\text{SC})$$

Theorem [El-Kebir, Oesper et al., 2015]
VAFFP is NP-complete for $m = O(n)$

Multi-State Perfect Phylogeny



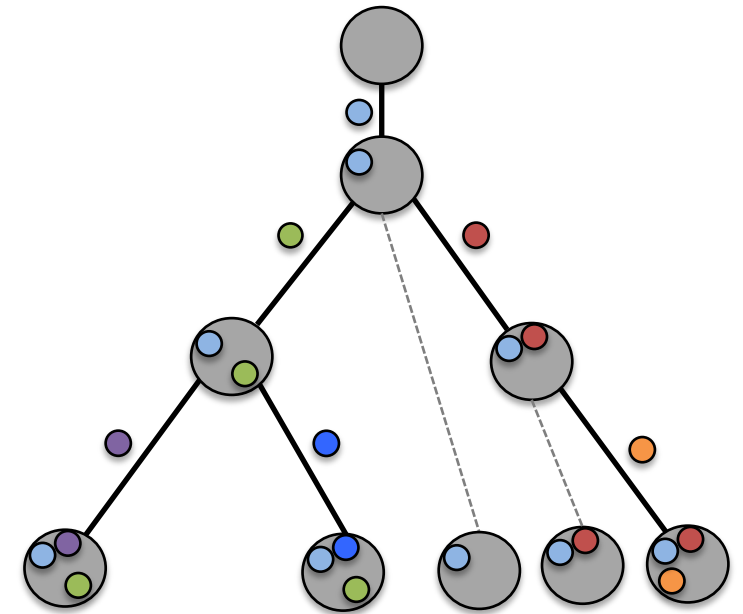
Theorem [El-Kebir et al., 2016]
Solutions are **threaded spanning trees** satisfying

$$f_p^+(D_{(c,i)}) \geq \sum_{(d,j) \in \delta(c,i)} f_p^+(D_{(d,j)}) \quad (\text{MSSC})$$

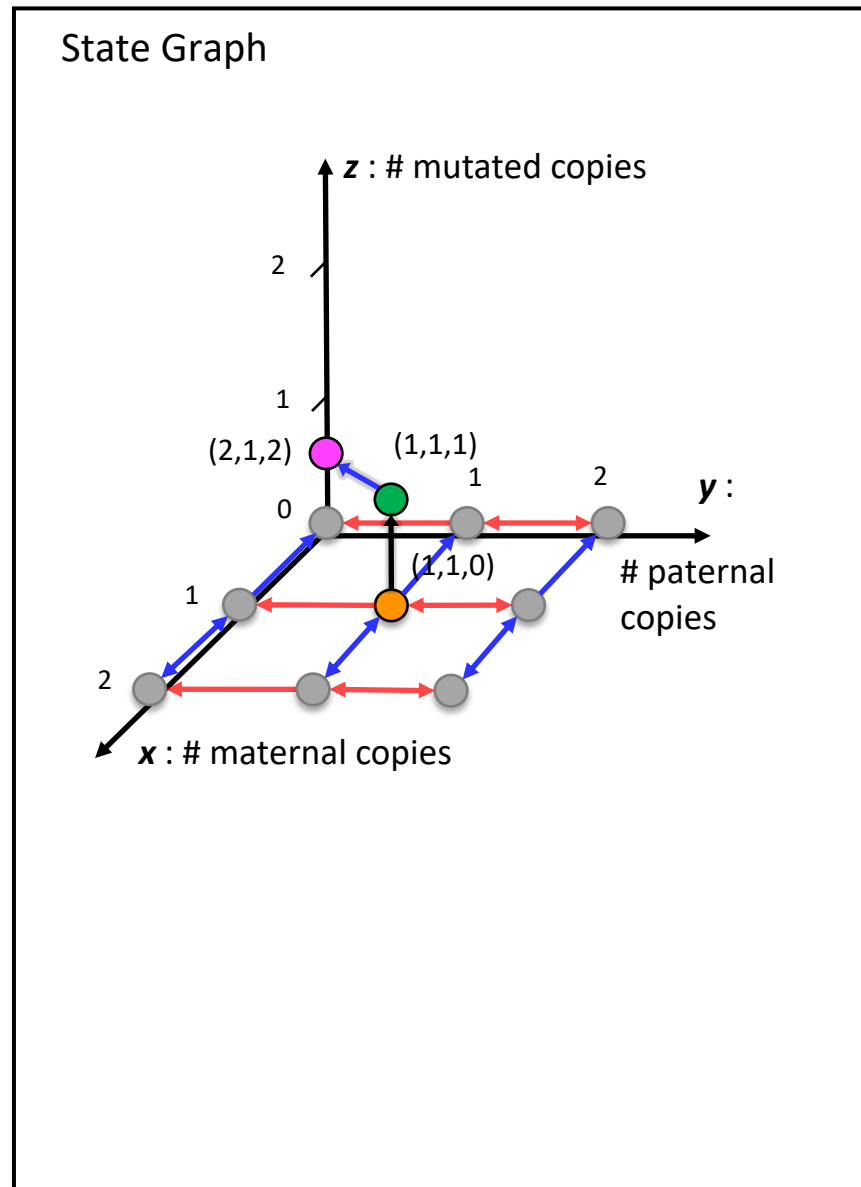
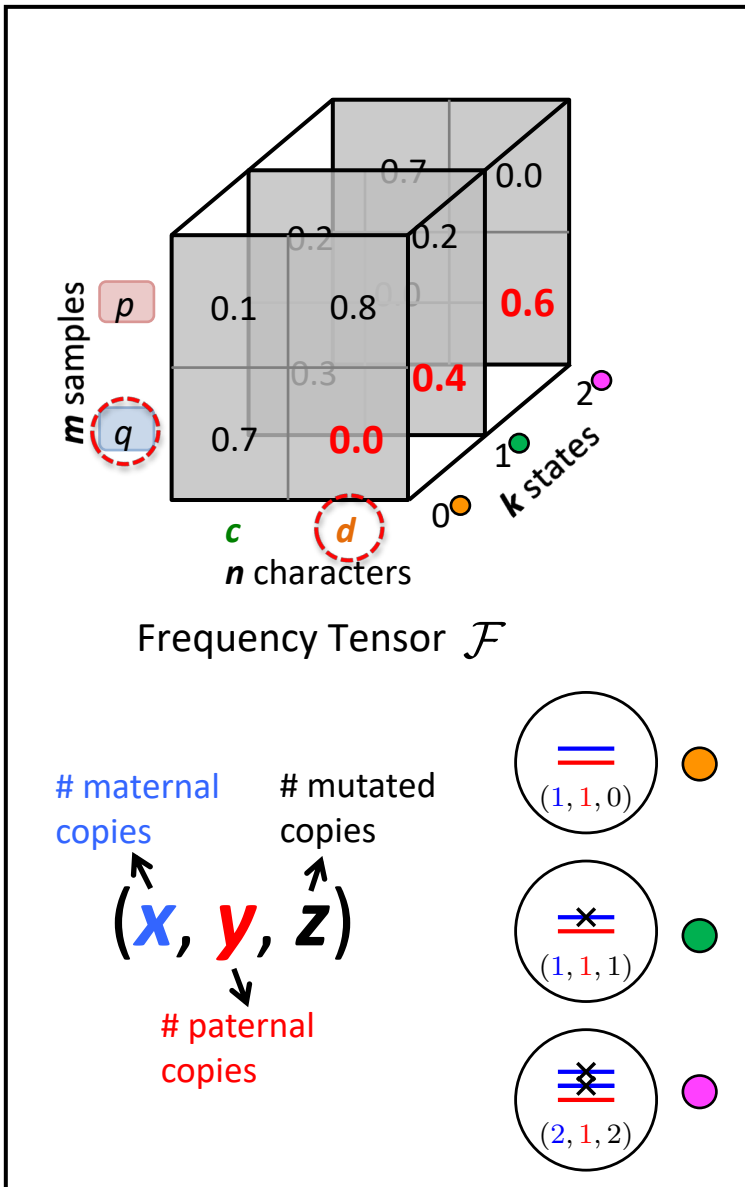
Theorem [El-Kebir et al., 2016]
PPMDP is NP-complete even for $m = 2$ and $k = 2$

Outline

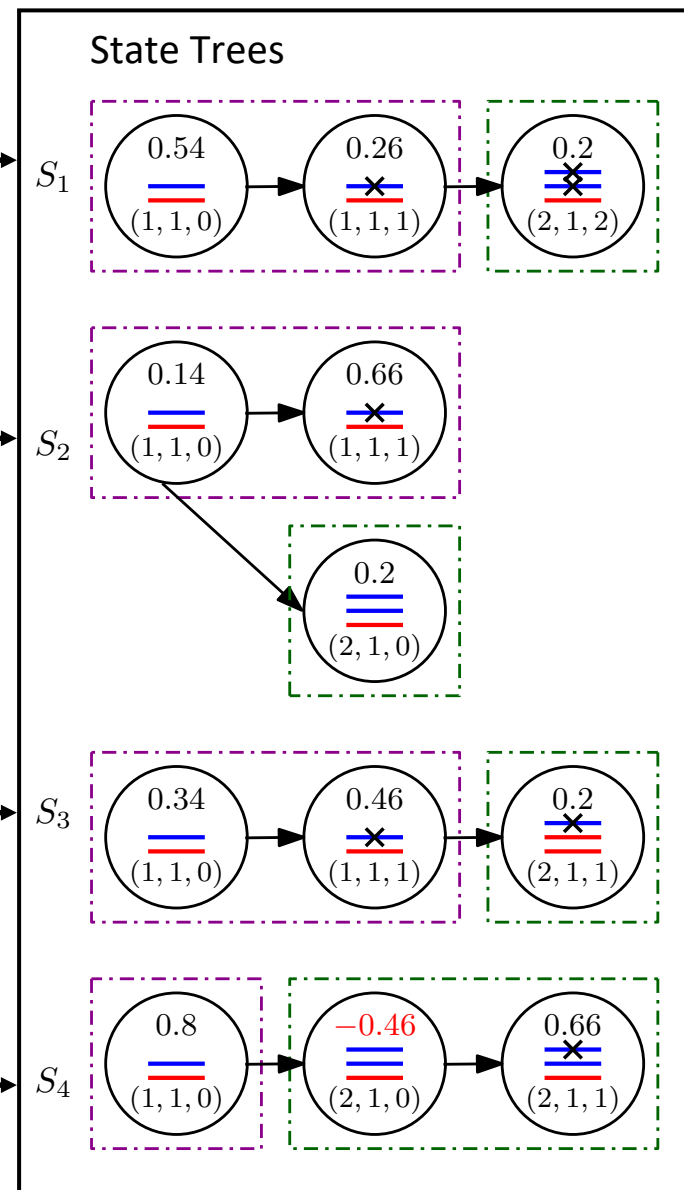
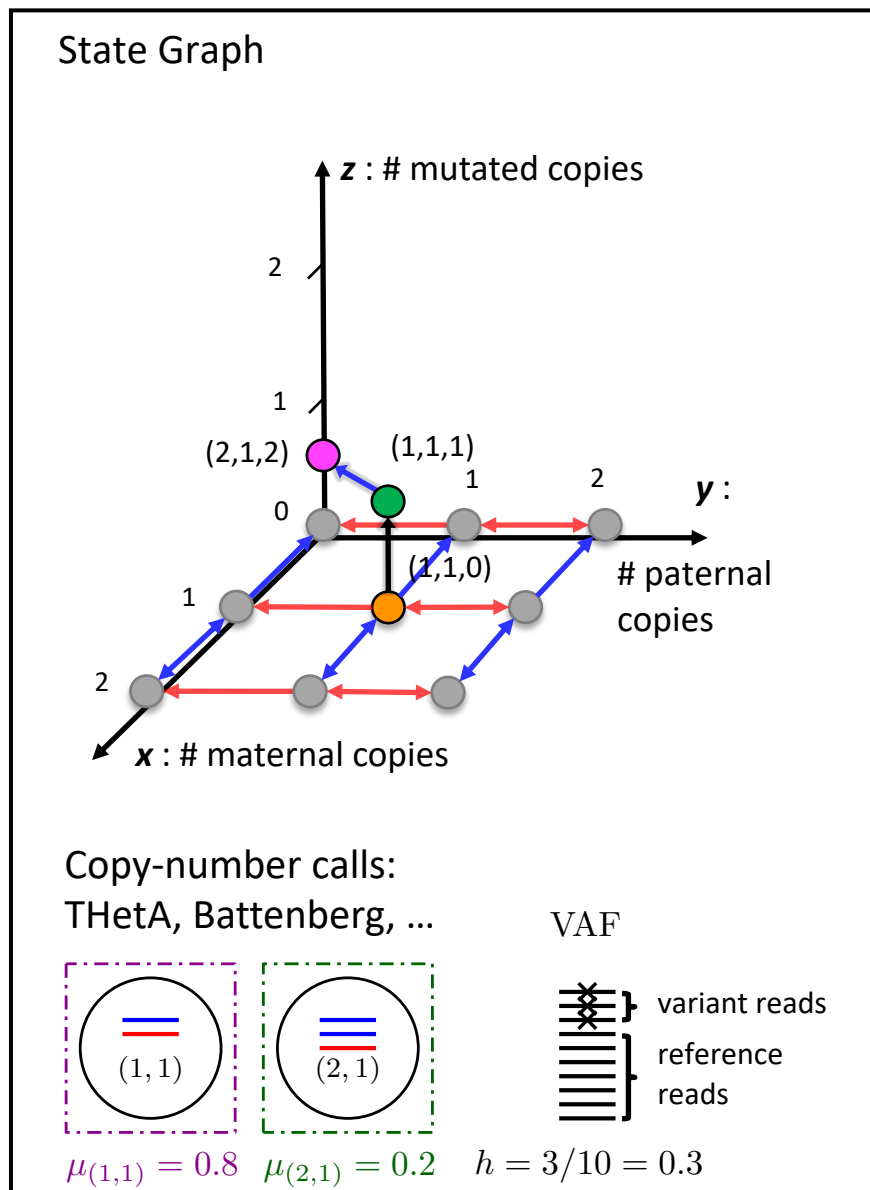
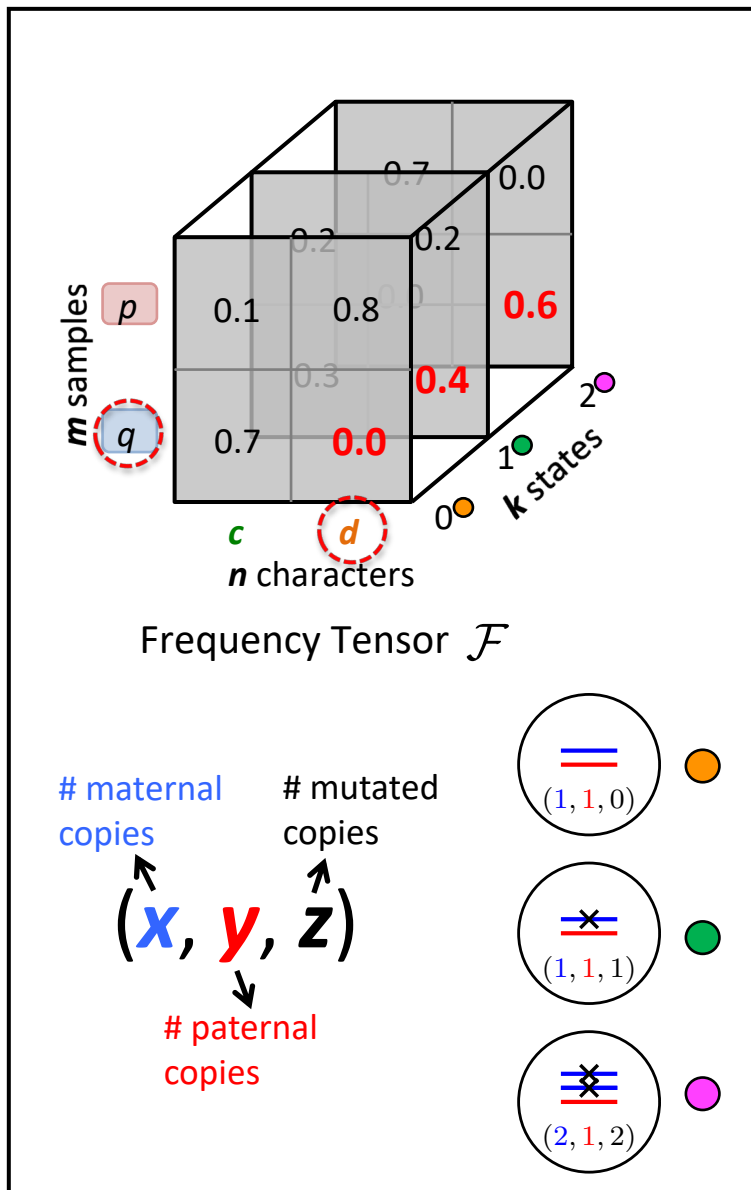
- Multi-State Perfect Phylogeny Mixture Deconvolution Problem
- Combinatorial Characterization of Solutions
- Application to Cancer Bulk-Sequencing Data
- Results



Application to Cancer Bulk Sequencing



Application to Cancer Bulk Sequencing

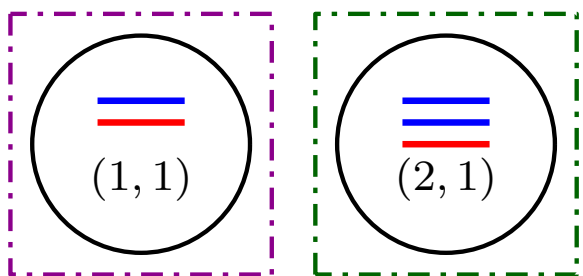


SPRUCE Enumerates Phylogenies

Somatic Phylogeny Reconstruction Using Combinatorial Enumeration

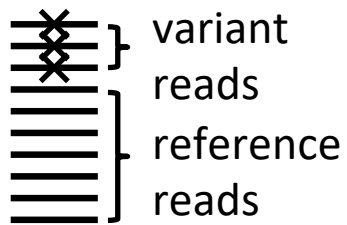
Available at: <http://compbio.cs.brown.edu/projects/spruce/>

Copy-number calls
and mixing proportions



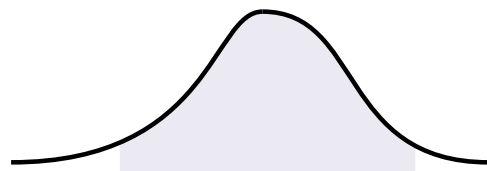
$\mu_{(1,1)} = 0.8$ $\mu_{(2,1)} = 0.2$

VAF



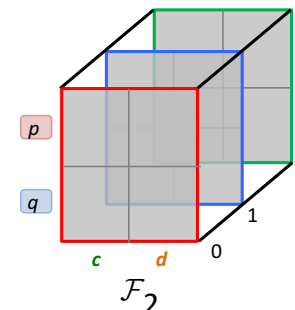
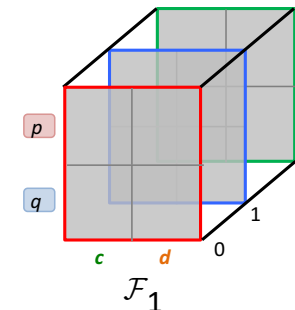
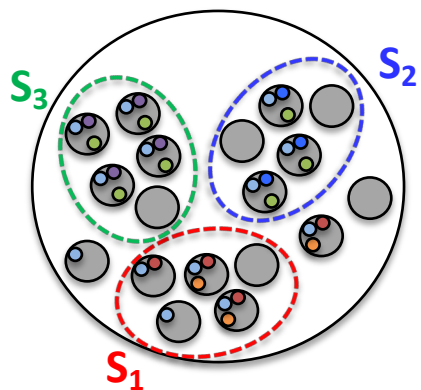
$h = 3/10 = 0.3$

VAF confidence interval

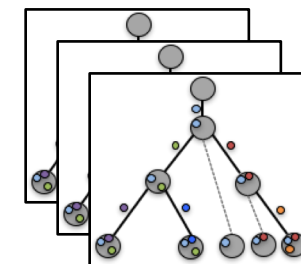
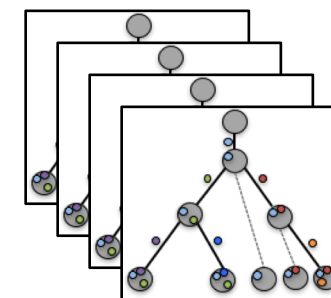
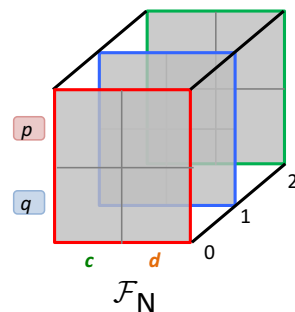


$[h, \bar{h}] = [0.18, 0.32]$

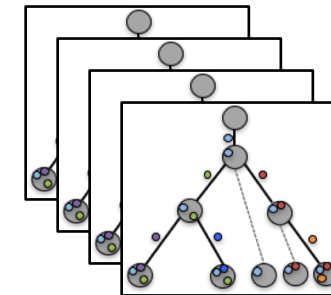
Error-Model for Variant Allele Frequencies



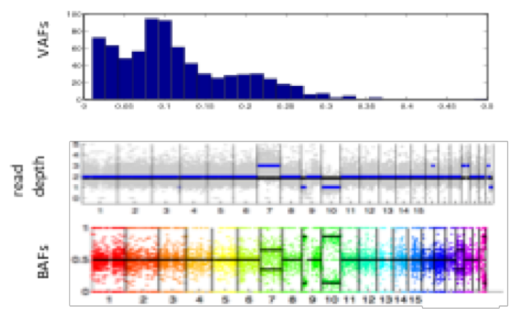
⋮



⋮



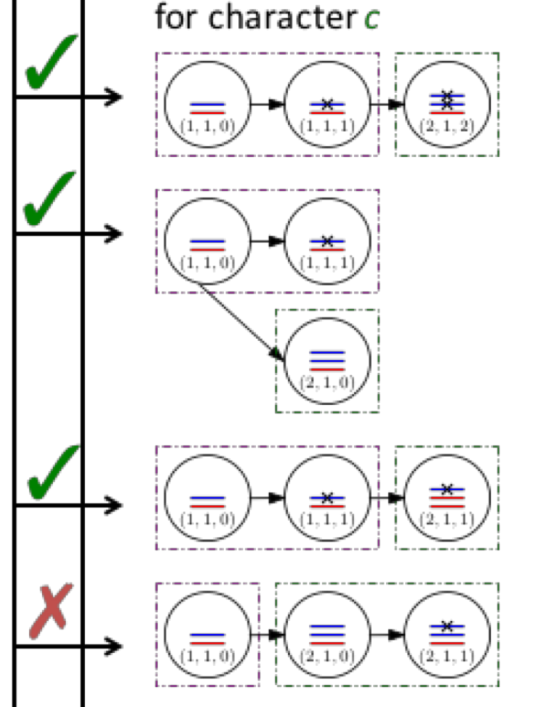
(A) Multi-Sample Sequencing Data



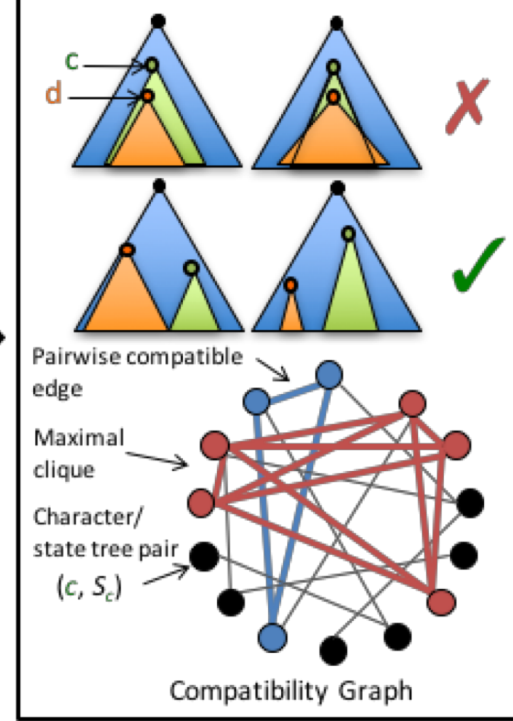
Input to SPRUCE

Char.	Sample	VAF	$\mu(1,1)$	$\mu(2,1)$
c	p	0.3	0.2	0.8
c	q	0.1	0.6	0.4
d	p	0.6	0.3	0.7
\vdots	\vdots	\vdots	\vdots	\vdots

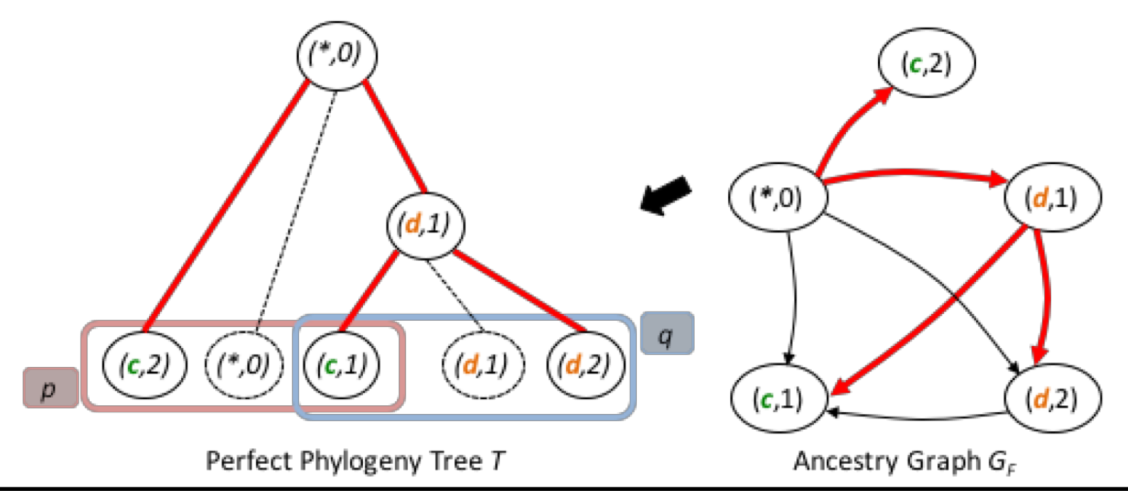
(B) Compatible State Trees for character c



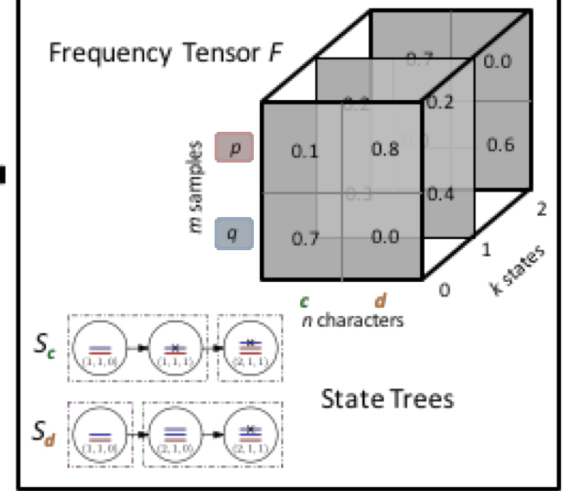
(C) Pairwise Compatibility



(E) Multi-State Perfect Phylogeny Enumeration



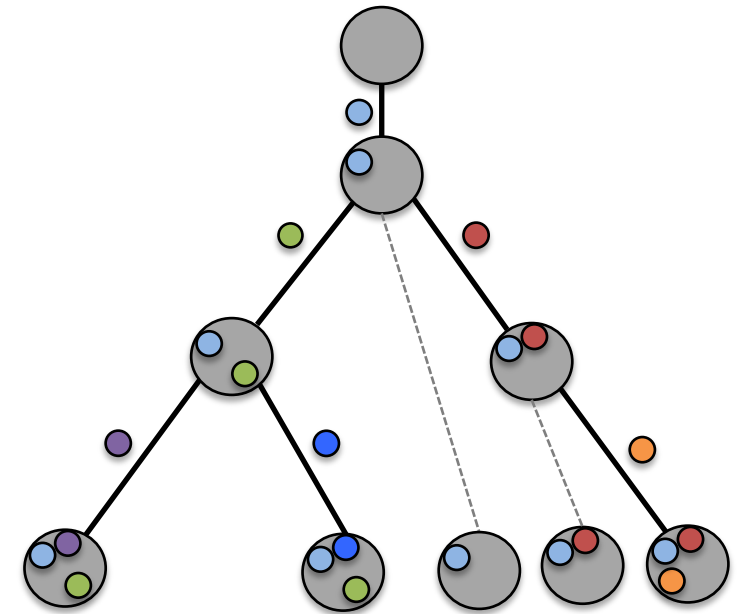
(D) Instances of Cladistic-PPMDP



SPRUCE:
Somatic Phylogeny
Reconstruction
Using Combinatorial
Enumeration

Outline

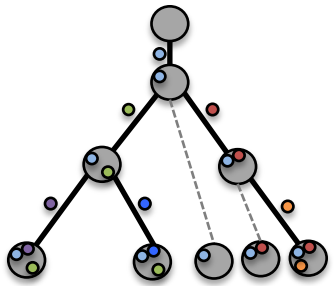
- Multi-State Perfect Phylogeny Mixture Deconvolution Problem
- Combinatorial Characterization of Solutions
- Application to Cancer Sequencing
- Results



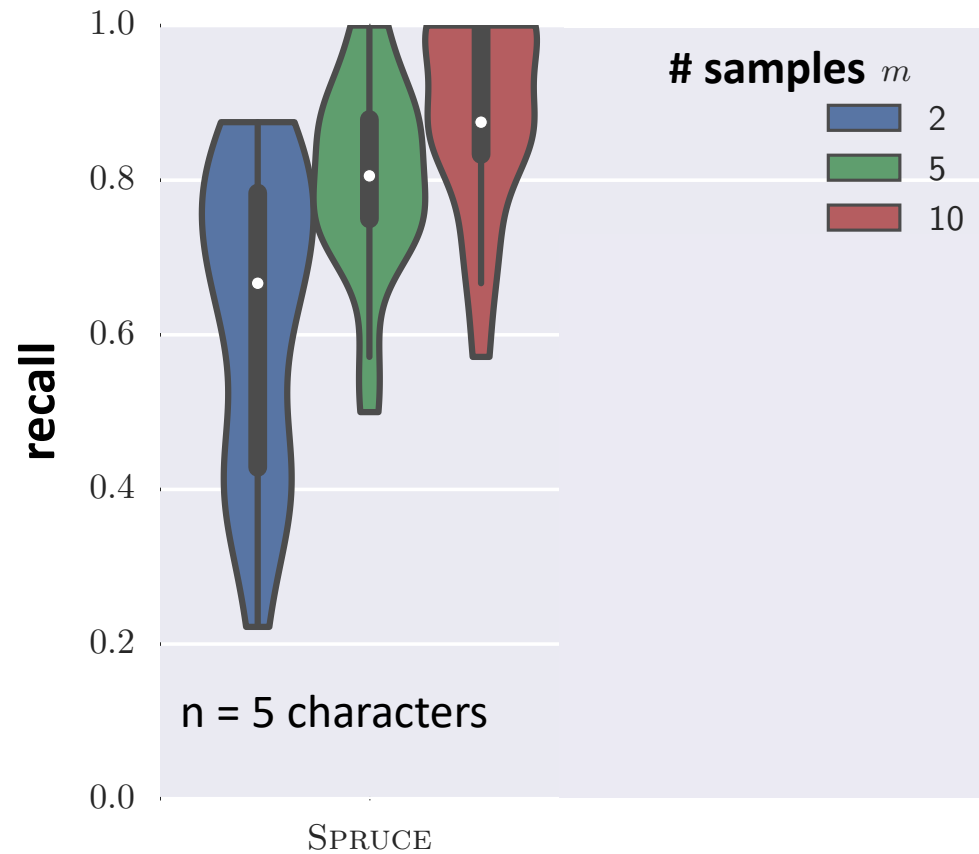
SPRUCE Accurately Recovers Simulated Trees

Parameters:

- # characters n : 5, 15



- # samples m : 2, 5, 10



Methods:

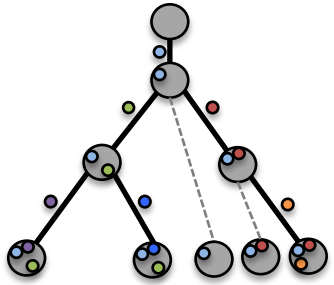
- SPRUCE
- PhyloWGS [Deshwar et al., 2015]

Increasing number of samples decreases ambiguity

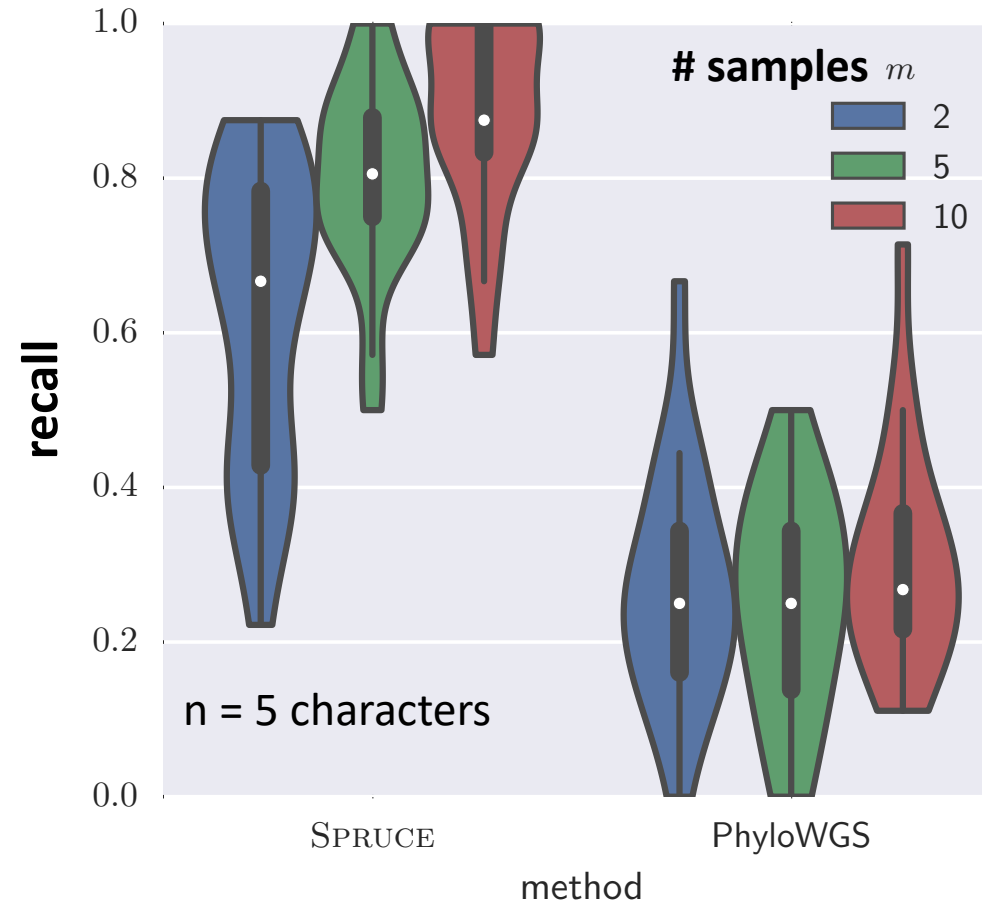
SPRUCE Accurately Recovers Simulated Trees

Parameters:

- # characters n : 5, 15



- # samples m : 2, 5, 10



Methods:

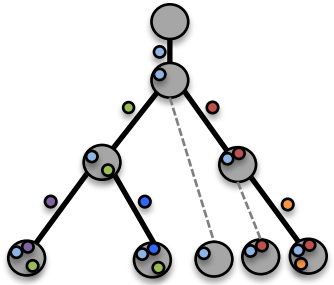
- SPRUCE
- PhyloWGS [Deshwar et al., 2015]

Increasing number of samples decreases ambiguity

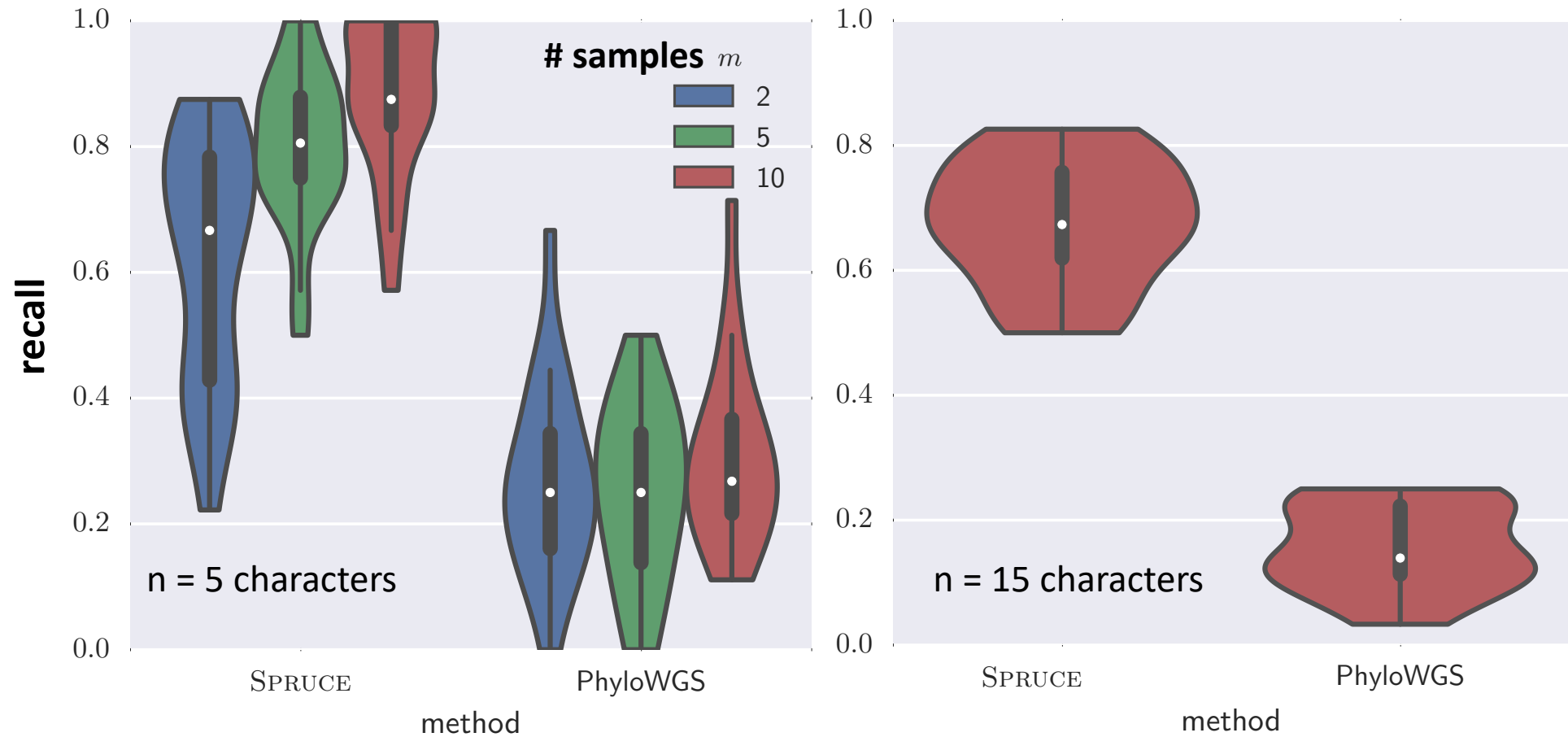
SPRUCE Accurately Recovers Simulated Trees

Parameters:

- # characters n : 5, 15



- # samples m : 2, 5, 10

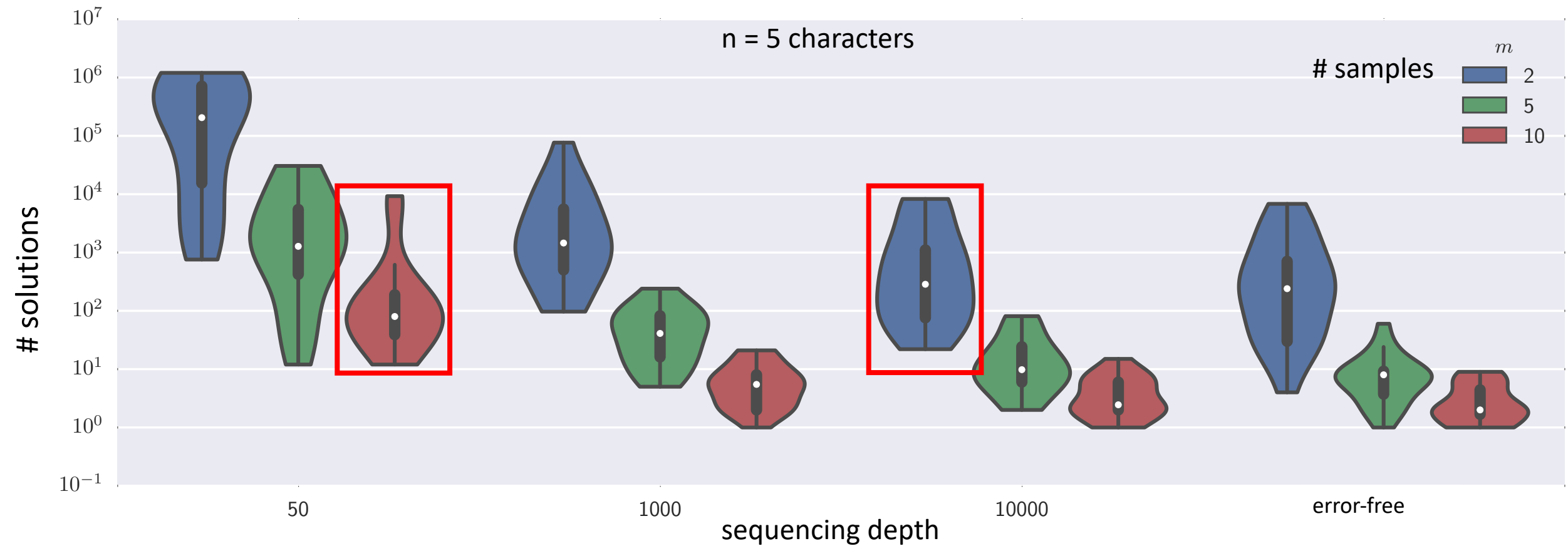


Methods:

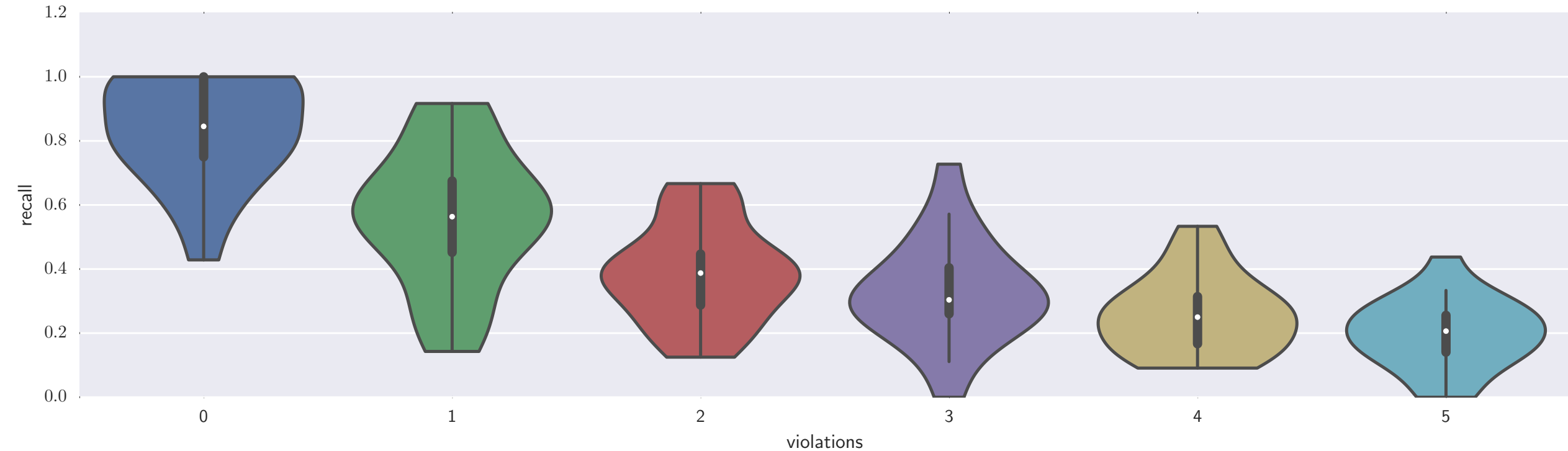
- SPRUCE
- PhyloWGS [Deshwar et al., 2015]

Increasing number of samples decreases ambiguity

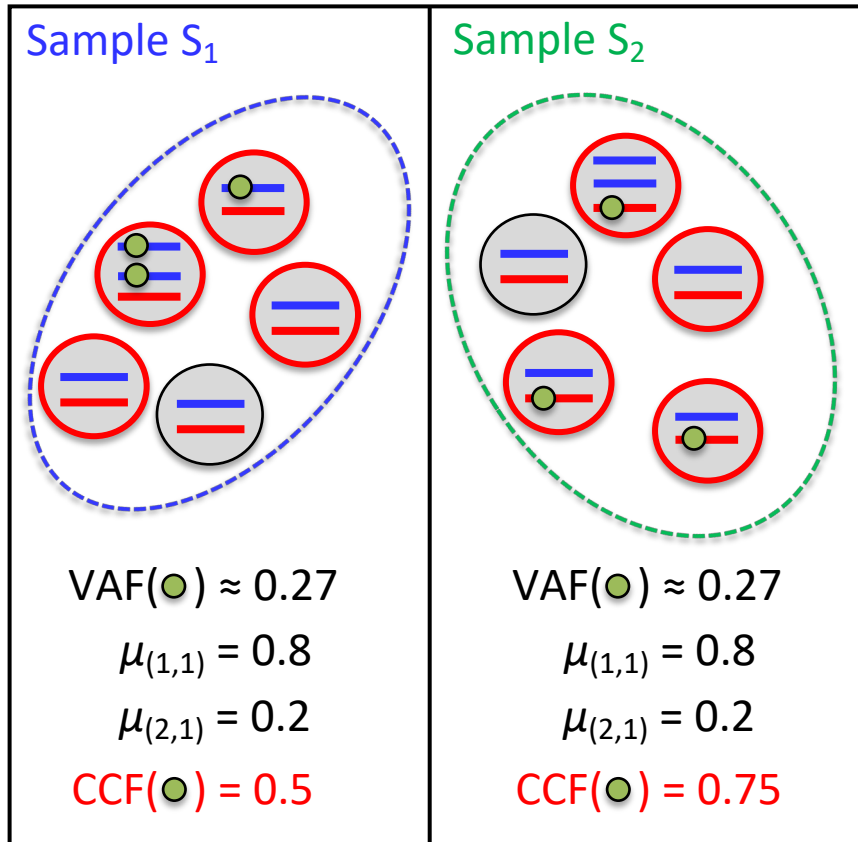
More Samples vs. More Coverage



Violations of Infinite Alleles Assumption



Cancer Cell Fractions (CCFs) Cannot Be Inferred A Priori

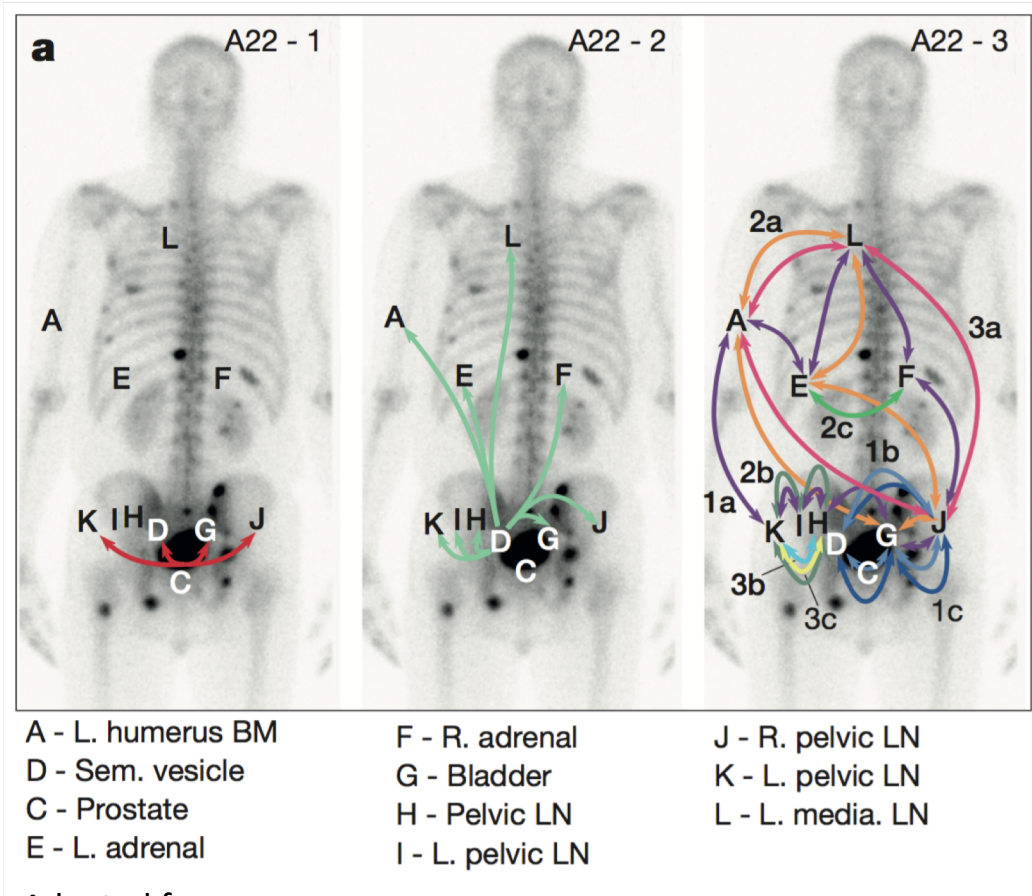


$$CCF(\bullet) = \frac{\# \text{ tumor cells with } \bullet}{\# \text{ tumor cells}}$$

CCFs are used extensively in studying intra-tumor heterogeneity and tumor evolution:

1. Timing of driver mutations
 - Andor *et al. Nature* (2015)
 - McGranahan *et al. Science Translational Medicine* (2015)
2. Tumor evolution and phylogeny reconstruction
 - Bolli *et al. Nature communications* (2014)
 - Nik-Zainal *et al. The life history of 21 breast cancers. Cell* (2012)
 - Sanborn *et al. PNAS* (2015)
 - Sottoriva *et al. Nature Genetics* (2015)
3. Developmental patterns of metastases
 - Brastianos *et al. Cancer Discovery*, (2015)
 - Gundem *et al. Nature* (2015)

Metastatic Evolution in Prostate Tumor



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

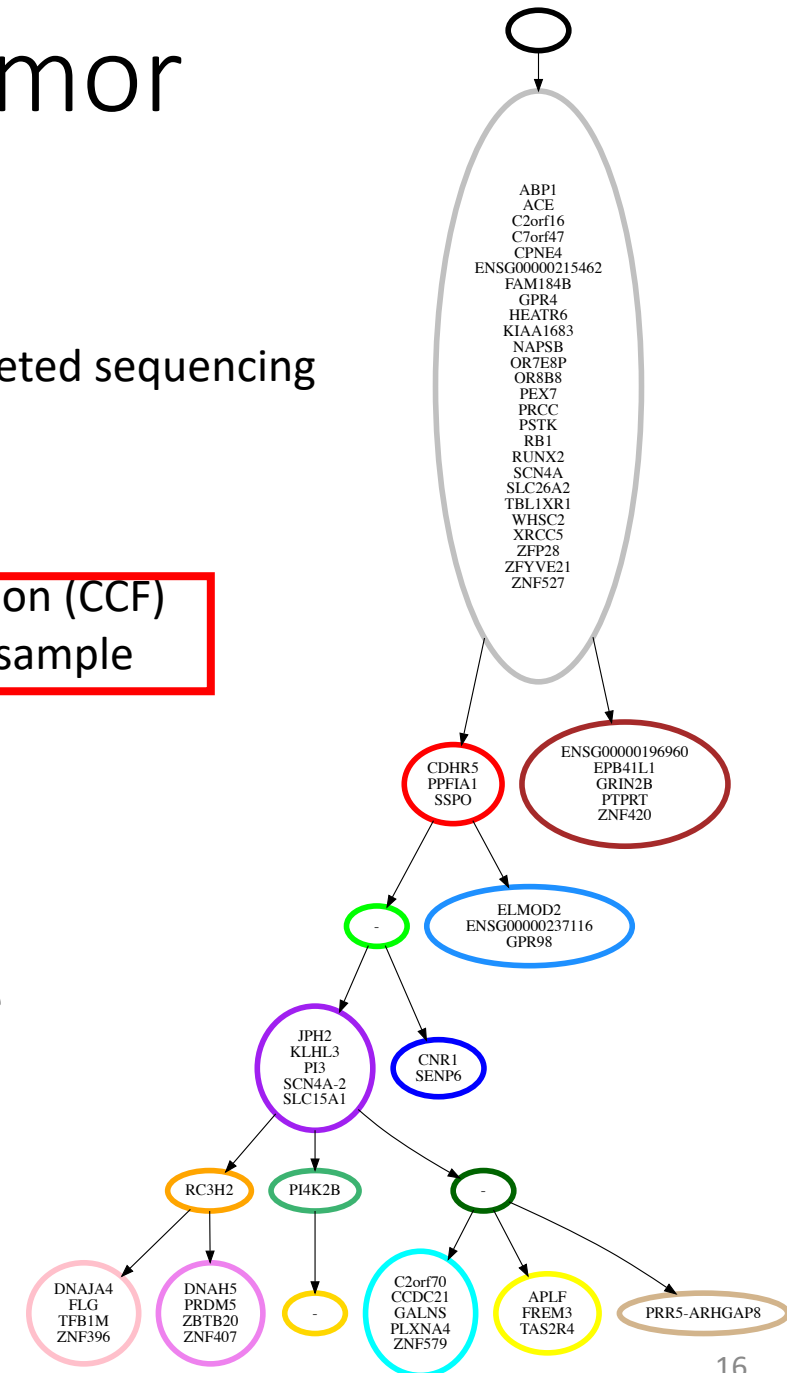
The evolutionary history of lethal metastatic prostate cancer

Input:

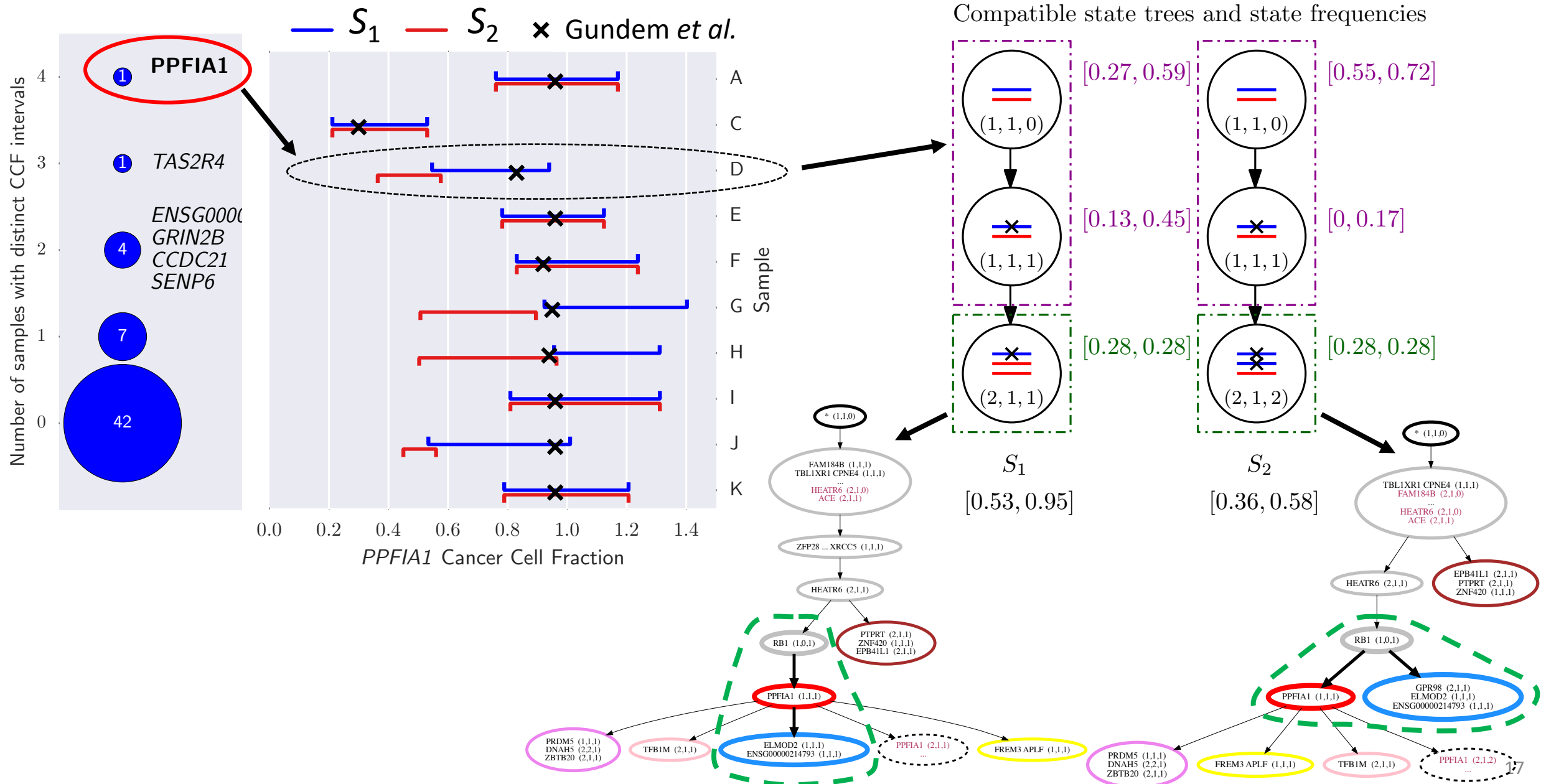
- 10 samples:
whole-genome & targeted sequencing
- ~110 SNVs

Tree building:

1. Infer cancer cell fraction (CCF) for each SNV in each sample
2. Cluster SNVs by CCFs across samples
3. Construct tree using Pigeon-Hole-Principle (Sum Condition)



Cancer Cell Fractions Cannot Be Inferred A Priori



Conclusions

- Copy-number aberrations confound variant allele frequencies
 - SNVs and CNAs must be considered jointly in phylogeny reconstruction
- Generalization of infinite sites for SNVs is infinite alleles for SNVs + CNAs
 - Multi-state Perfect Phylogeny Mixture Problem (PPM)
- Complete combinatorial characterization of the problem
 - Solutions are constrained spanning trees in a directed multi-graph
 - PPM is NP-complete for $k = 2$ and $m = 2$
- Using combinatorial structure, SPRUCE accurately recovers simulated trees
- Cancer cell fractions cannot be uniquely inferred *a priori* by considering SNVs in isolation
- Precise mathematical models are needed to describe evolutionary process in cancer

