CS 598MEB
Introduction to Bioinformatics
Lecture 7

Mohammed El-Kebir

February 11, 2020

Outline

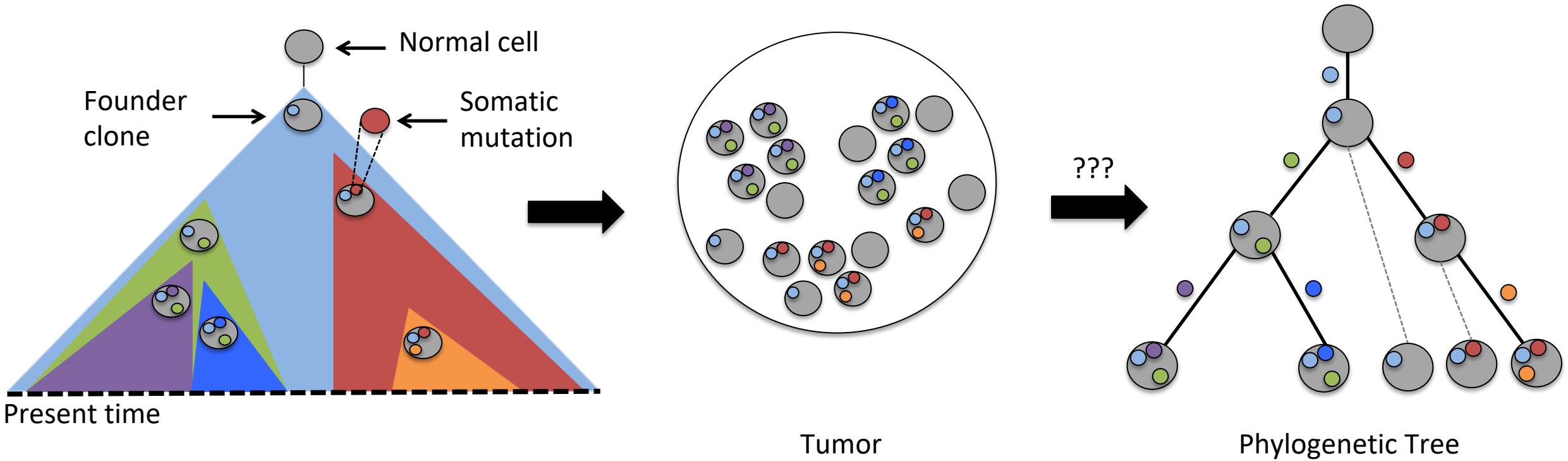
- Recap
- Tumor Phylogeny Inference from Bulk DNA-seq with Copy-Number Aberrations

Reading:

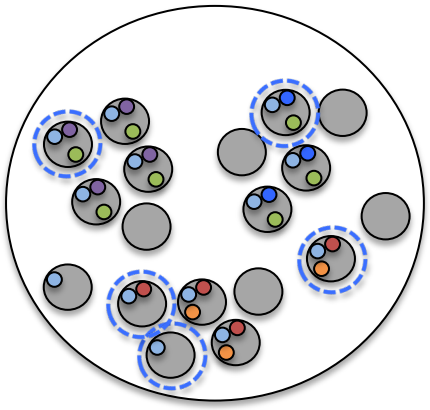
- **M. El-Kebir**, G. Satas , L. Oesper and B.J. Raphael. Inferring the Mutational History of a Tumor using Multi-State Perfect Phylogeny Mixtures. Cell Systems, 3(1):43-53, 2016.

Tumor Evolution as a Phylogenetic Tree

Clonal Theory [Nowell, 1976]



Observations are Leaves of a Perfect Phylogeny *T*



Tumor Snapshot

Single-cell sequencing

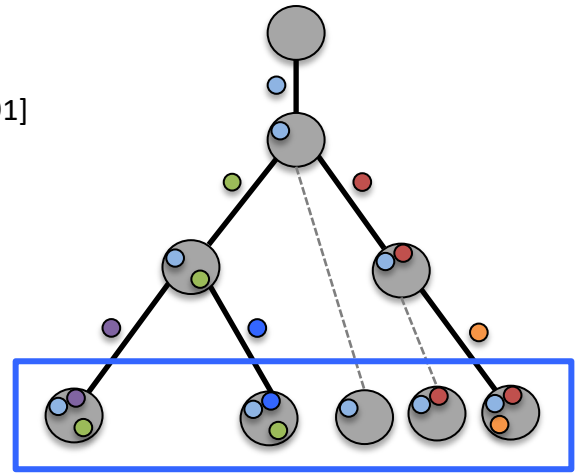
M =

SNVs						
						leaves of *T*
1	1	1	0	0	0	
1	1	0	1	0	0	
1	0	0	0	0	0	
1	0	0	0	1	0	
1	0	0	0	1	1	

Binary Matrix ***B***

O(mn)

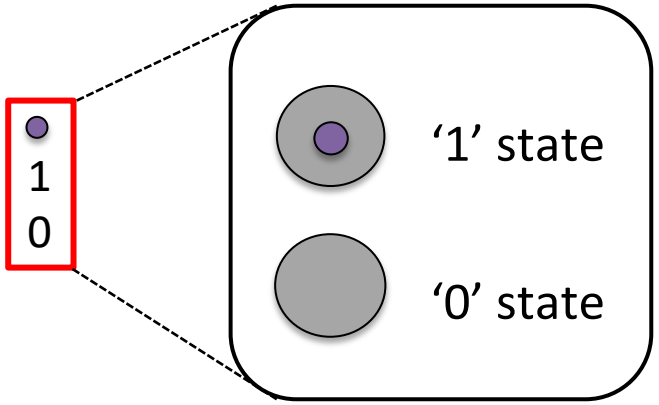
[Gusfield, 1991]



Two-State Perfect Phylogeny Tree ***T***

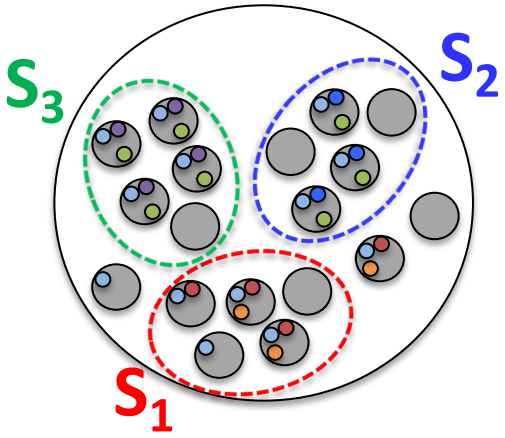
Assumptions:

- 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*



Tumor Snapshot

Bulk sequencing

F =

SNVs

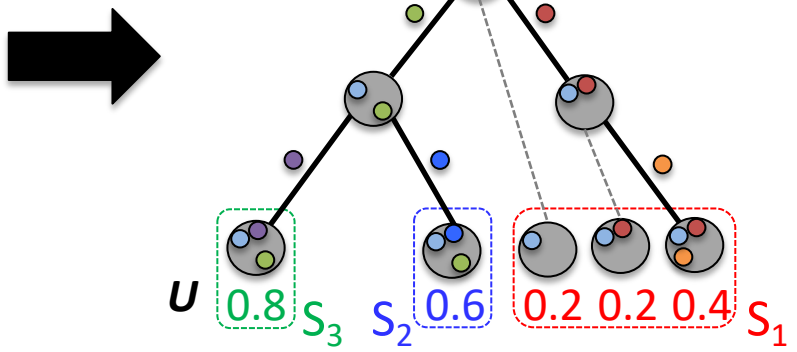
0.4	0.0	0.0	0.0	0.3	0.2	S_1
0.3	0.3	0.0	0.3	0.0	0.0	S_2
0.4	0.4	0.4	0.0	0.0	0.0	S_3

samples

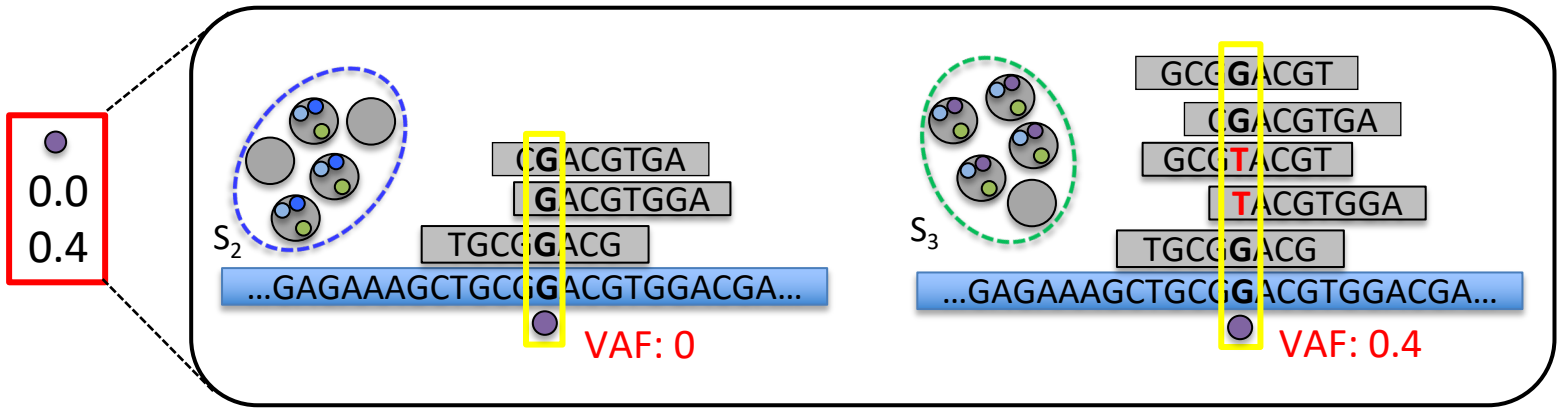
Variant Allele Frequencies Matrix ***F***

NP-complete

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



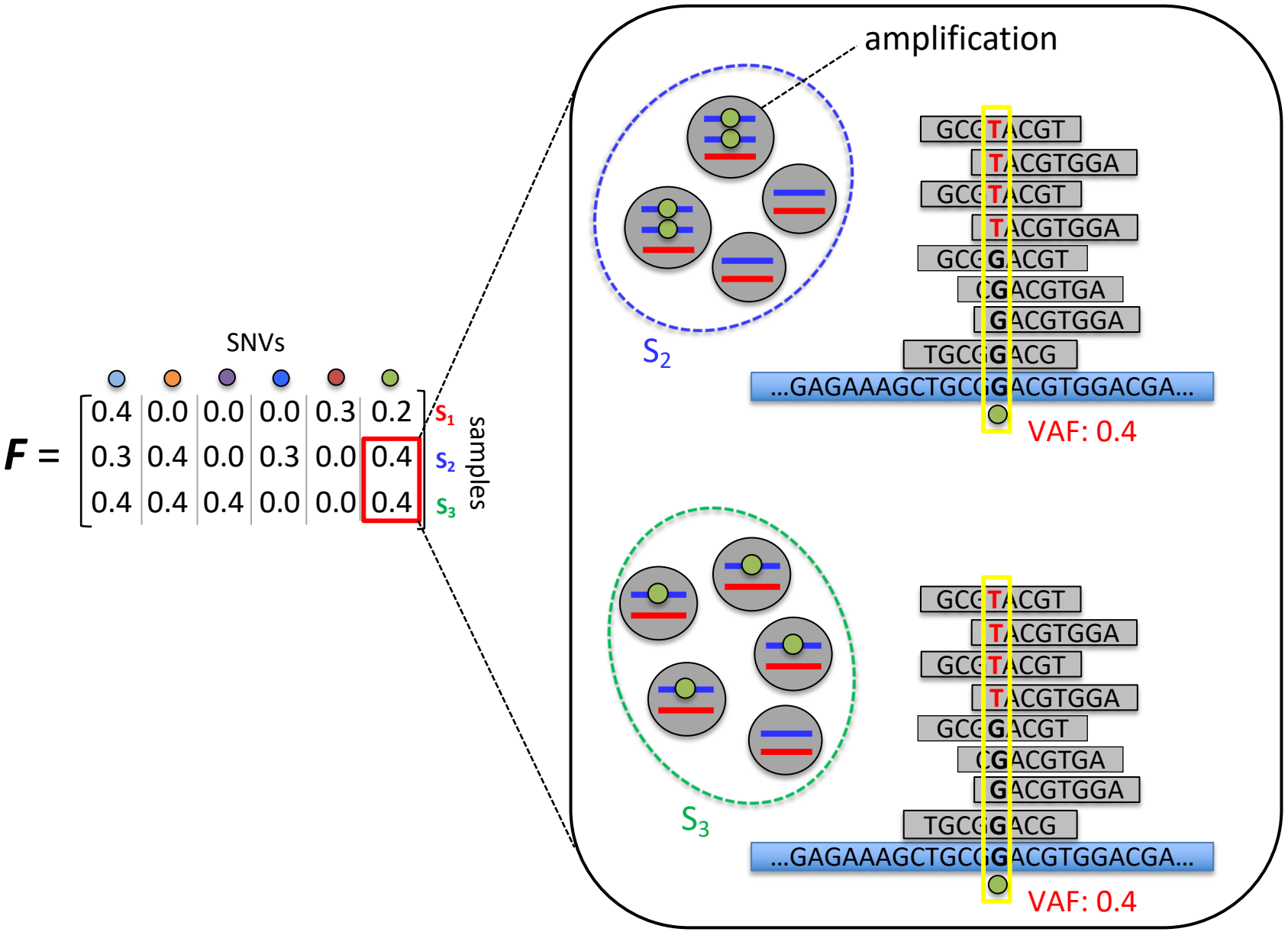
Two-State Perfect Phylogeny Tree ***T***
Mixing Proportions ***U***



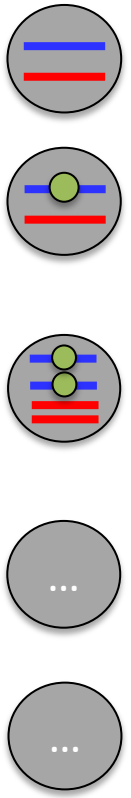
Seq. method	Inferring *T*	Complexity
single-cell	**unmixed** two-state perfect phylogeny	*O(mn)*
bulk	**mixed** two-state perfect phylogeny	NP-complete

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

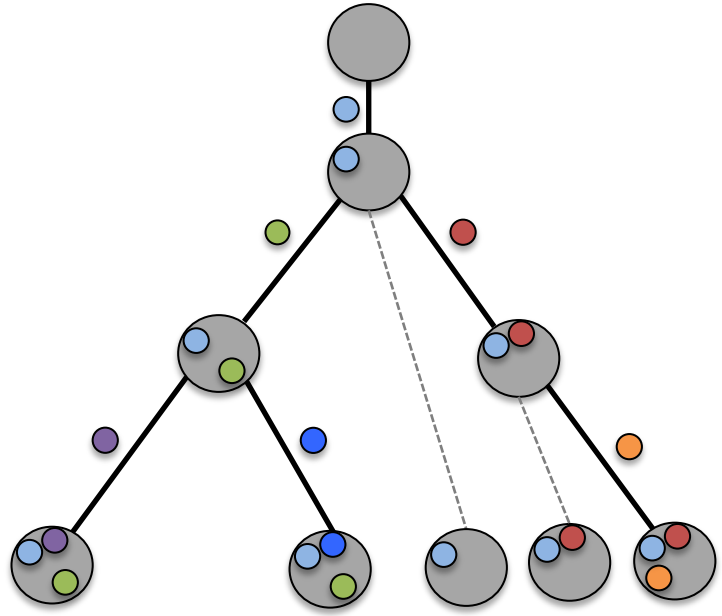


Need > 2 states:



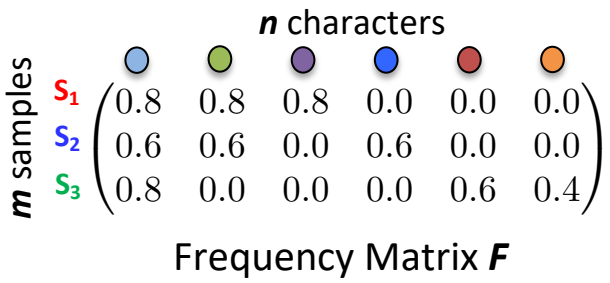
Outline

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

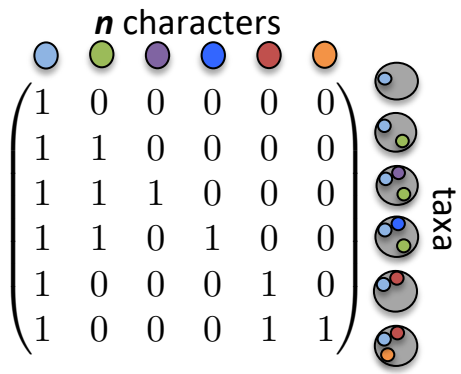


Infinite Sites Generalizes to Infinite Alleles

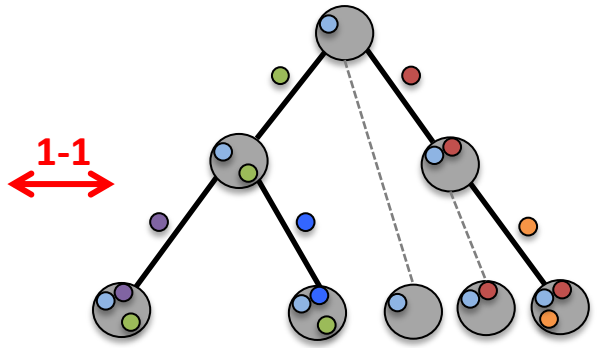
Two-State Perfect Phylogeny – Infinite sites assumption: a character changes state once



Frequency Matrix ***F***



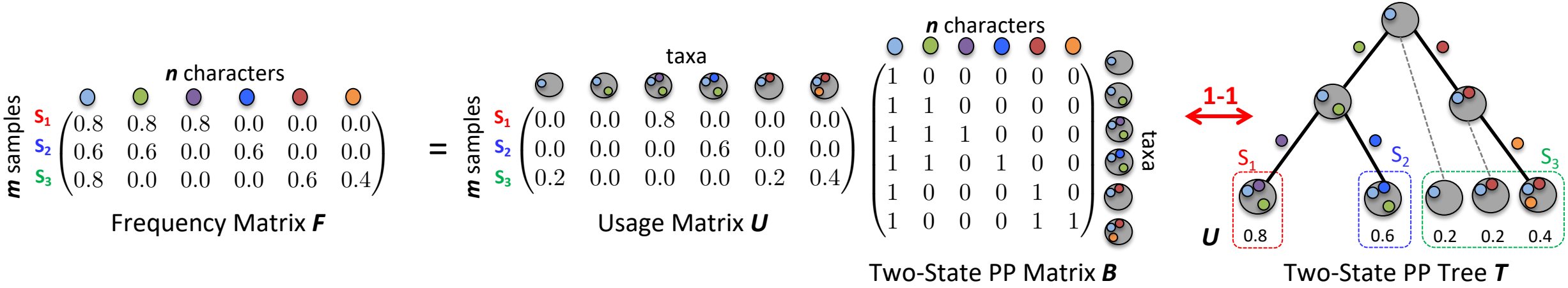
Two-State PP Matrix ***B***



Two-State PP Tree ***T***

Infinite Sites Generalizes to Infinite Alleles

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

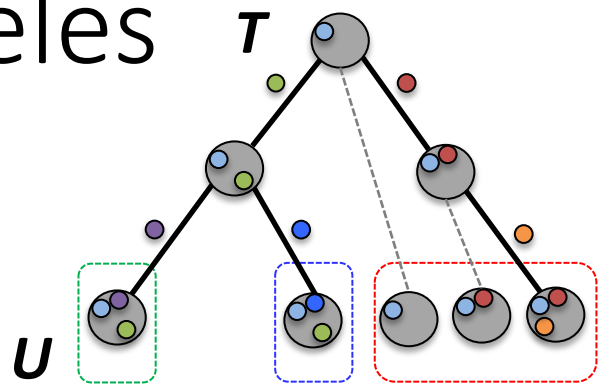


VAF Factorization Problem (VAFFP): Given ***F***, find ***U*** and ***B*** such that ***F = U B*** [El-Kebir, Oesper et al., 2015]

Infinite Sites Generalizes to Infinite Alleles

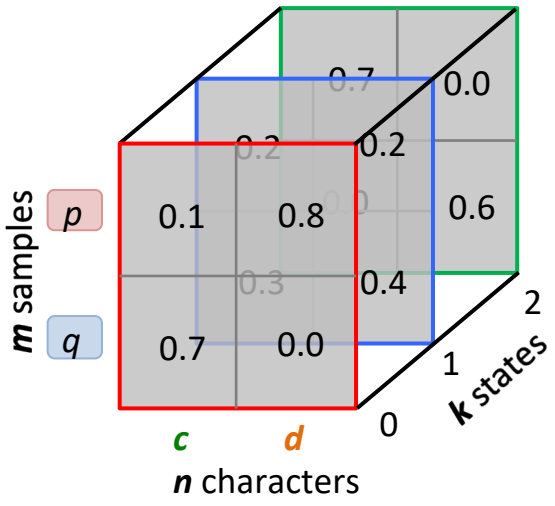
Two-State Perfect Phylogeny – Infinite sites assumption: a character changes state once

1-1



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

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



Frequency Tensor $\mathcal{F}$

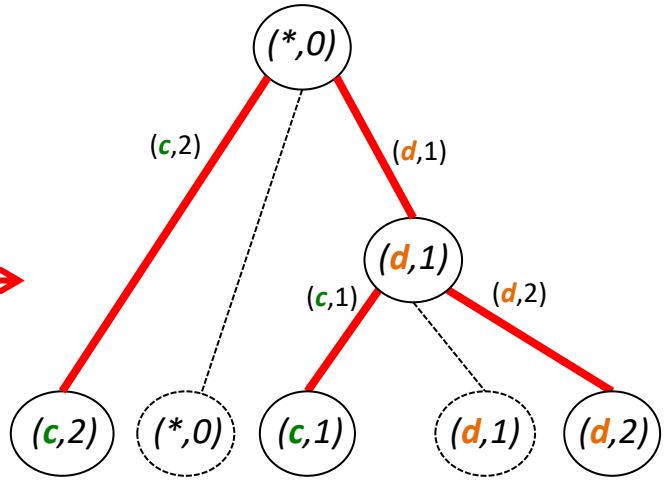
$\begin{pmatrix} 0.7 & 0.0 \\ 0.0 & 0.6 \end{pmatrix}$ F_2

$\begin{pmatrix} 0.2 & 0.2 \\ 0.3 & 0.4 \end{pmatrix}$ F_1

$\begin{pmatrix} 0.1 & 0.8 \\ 0.7 & 0.0 \end{pmatrix}$ F_0

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

Multi-State PP Matrix ***A***

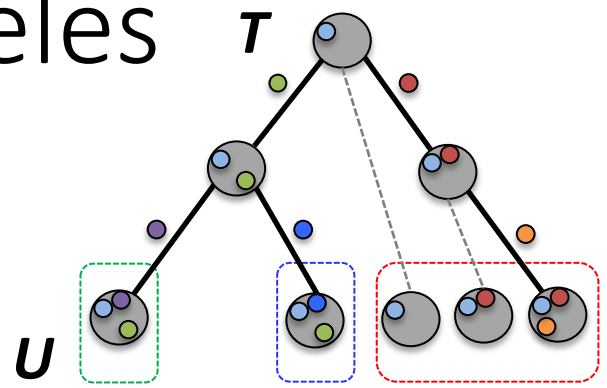


Multi-State Perfect Phylogeny Tree ***T***

Infinite Sites Generalizes to Infinite Alleles

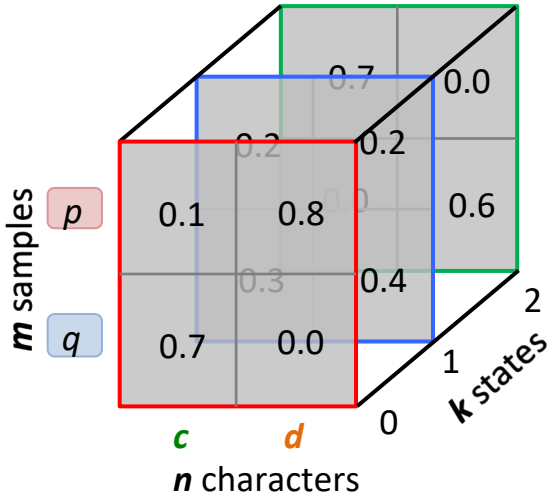
Two-State Perfect Phylogeny – Infinite sites assumption: a character changes state once

$$\underset{m \text{ samples}}{\underset{n \text{ characters}}{F}} = U \, B \quad \Leftrightarrow \text{1-1} \quad U$$



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

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



Frequency Tensor $\mathcal{F}$

$$\begin{pmatrix} 0.7 & 0.0 \\ 0.0 & 0.6 \end{pmatrix} = \begin{pmatrix} 0.1 & 0.2 & 0.7 & 0.0 & 0.0 \\ 0.0 & 0.3 & 0.0 & 0.1 & 0.6 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} A_2$$

$F_2 \qquad U$

$$\begin{pmatrix} 0.2 & 0.2 \\ 0.3 & 0.4 \end{pmatrix} = \begin{pmatrix} 0.1 & 0.2 & 0.7 & 0.0 & 0.0 \\ 0.0 & 0.3 & 0.0 & 0.1 & 0.6 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} A_1$$

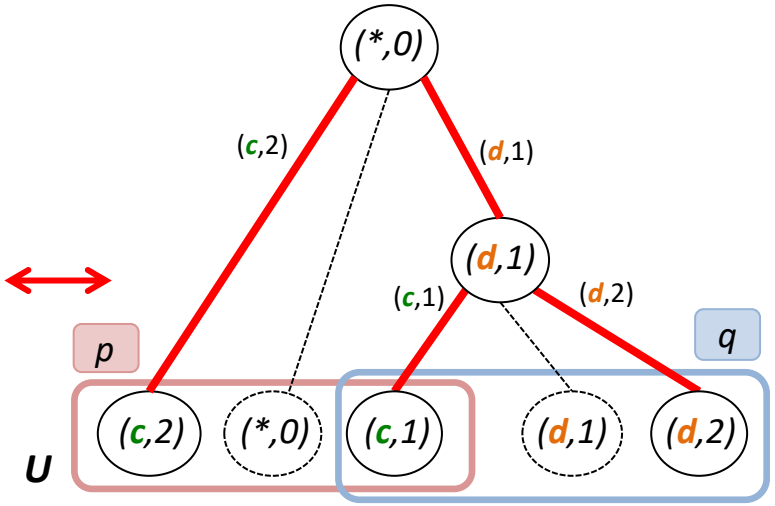
$F_1 \qquad U$

$$\begin{pmatrix} 0.1 & 0.8 \\ 0.7 & 0.0 \end{pmatrix} = \begin{pmatrix} 0.1 & 0.2 & 0.7 & 0.0 & 0.0 \\ 0.0 & 0.3 & 0.0 & 0.1 & 0.6 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} A_0$$

$F_0 \qquad U$

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

Multi-State PP Matrix ***A***

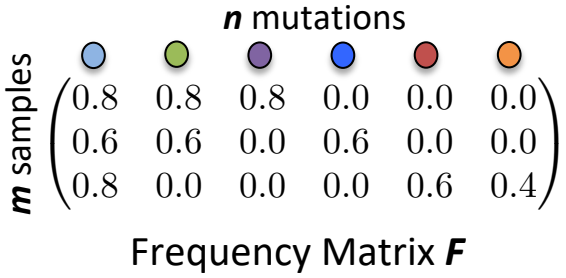


Multi-State Perfect Phylogeny Tree ***T***

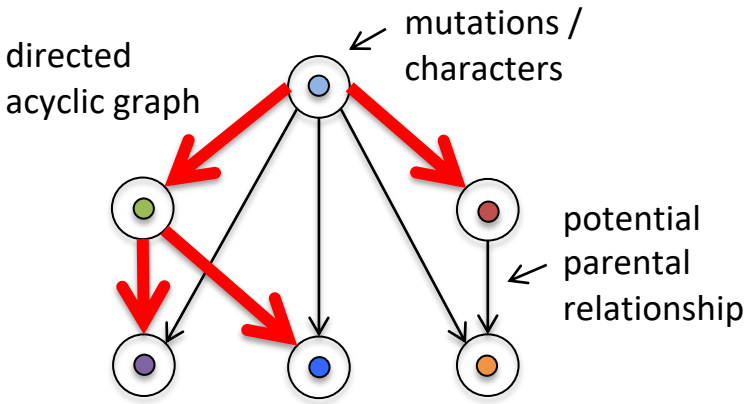
Multi-state Perfect Phylogeny Mixture Problem: Given ***F***, find ***U*** and ***A*** such that $F_i = U A_i$ for all states ***i***

Combinatorial Characterization of Solutions

Two-State Perfect Phylogeny



Frequency Matrix ***F***



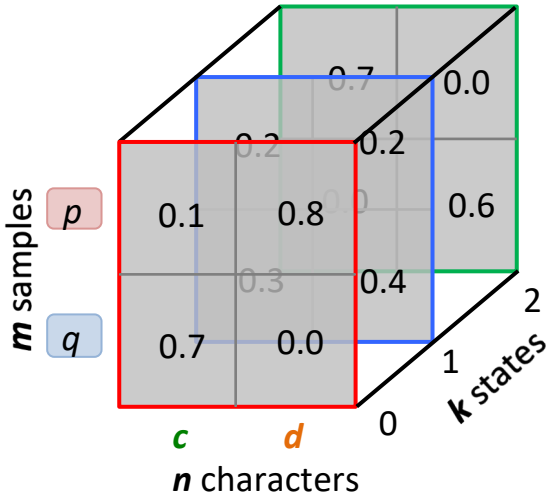
Ancestry Graph ***G***

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

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

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

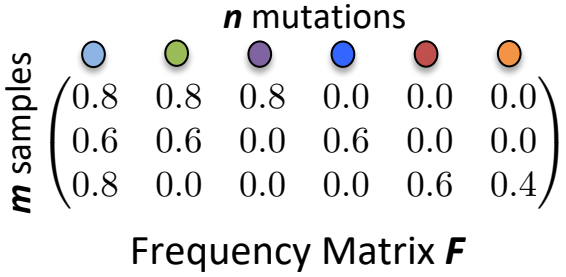
Multi-State Perfect Phylogeny



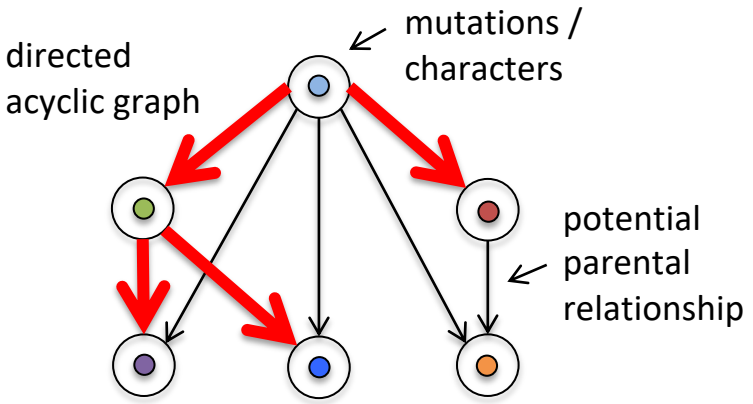
Frequency Tensor $\mathcal{F}$

Combinatorial Characterization of Solutions

Two-State Perfect Phylogeny



Frequency Matrix ***F***



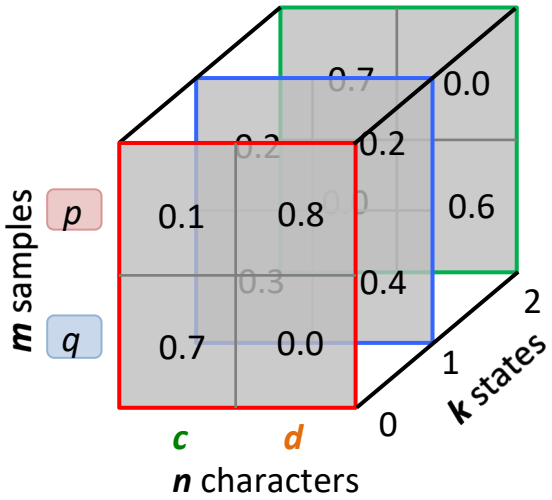
Ancestry Graph ***G***

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

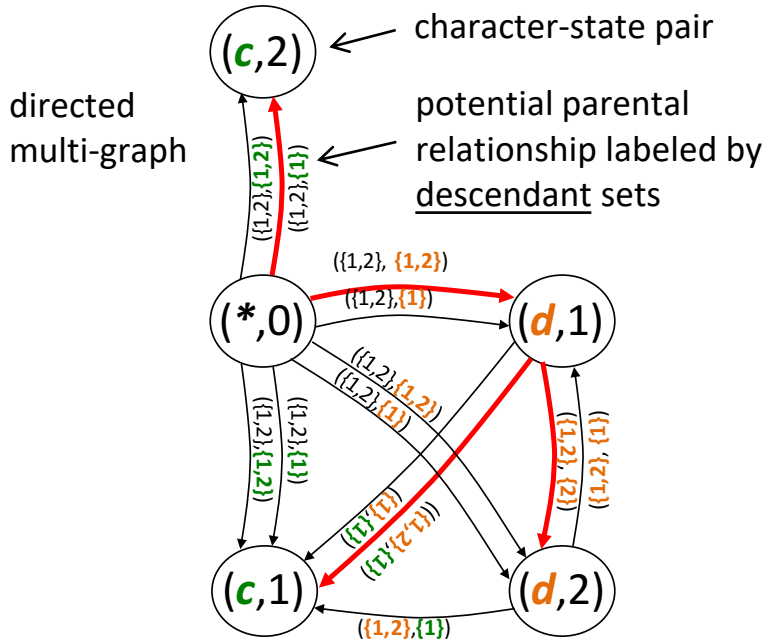
$$f_{p,(c,1)} \geq \sum_{(d,1)\in\delta(c,1)} f_{p,(d,1)} \qquad \text{(SC)}$$

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

Multi-State Perfect Phylogeny



Frequency Tensor $\mathcal{F}$



Multi-State Ancestry Graph ***G***

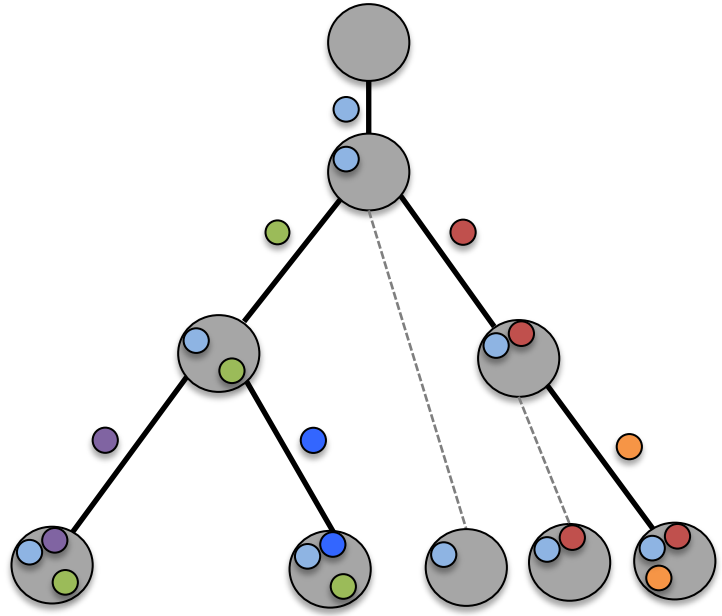
Theorem [El-Kebir et al., 2016]
Solutions are **threaded spanning trees** satisfying

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

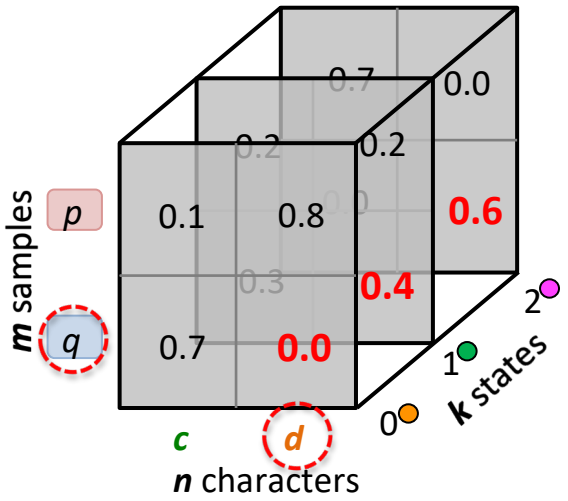
Theorem [El-Kebir et al., 2016]
PPMDP is NP-complete even for $m = 2$ and $k = 2$

Outline

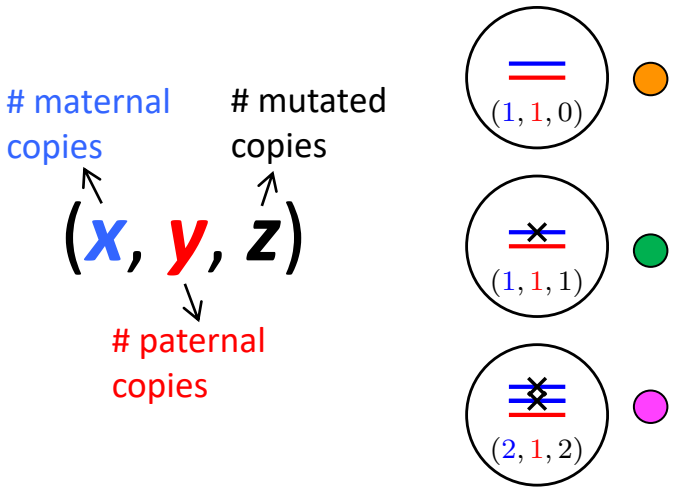
- Multi-State Perfect Phylogeny Mixture Deconvolution Problem
- Combinatorial Characterization of Solutions
- Application to Cancer Bulk-Sequencing Data
- Results



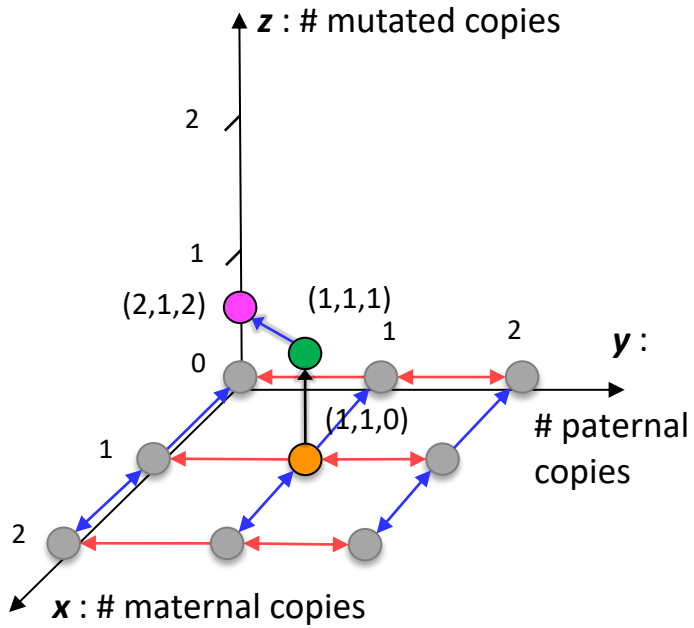
Application to Cancer Bulk Sequencing



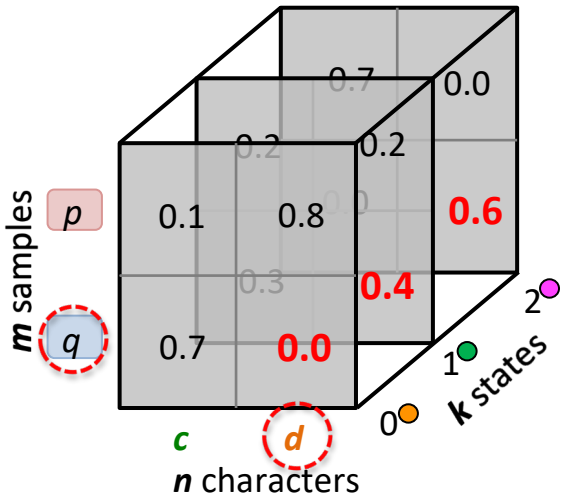
Frequency Tensor $\mathcal{F}$



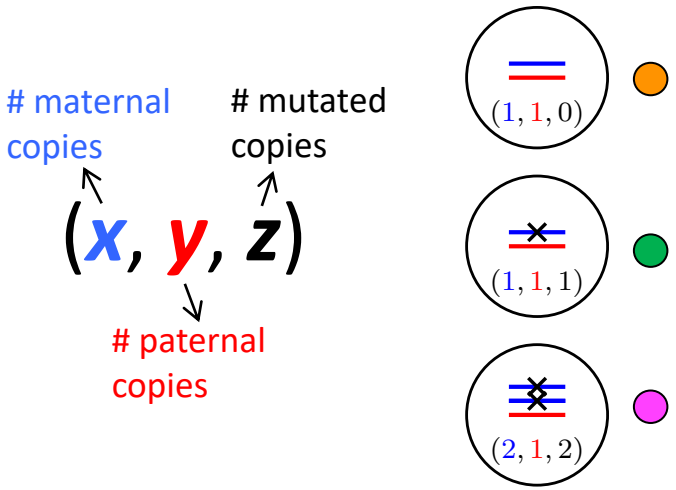
State Graph



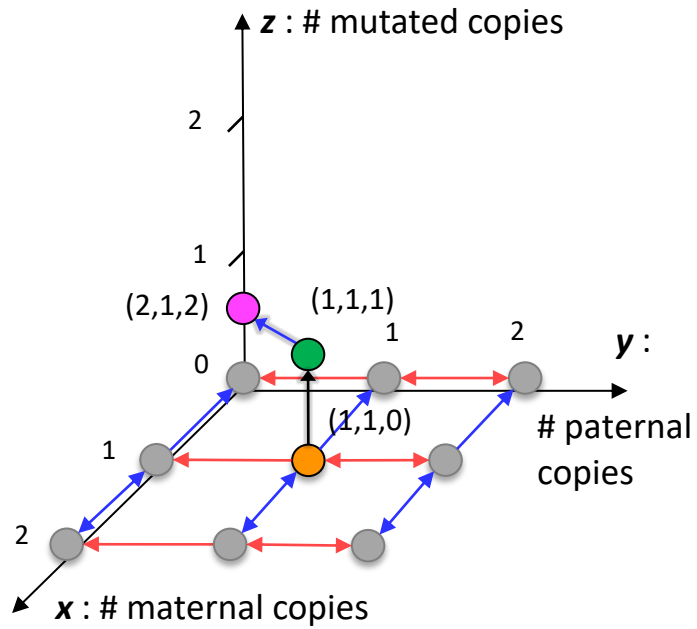
Application to Cancer Bulk Sequencing



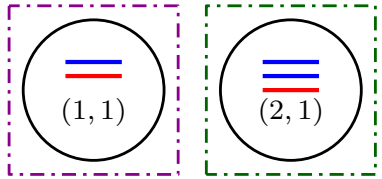
Frequency Tensor $\mathcal{F}$



State Graph

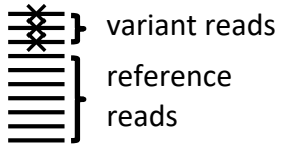


Copy-number calls:
THetA, Battenberg, …



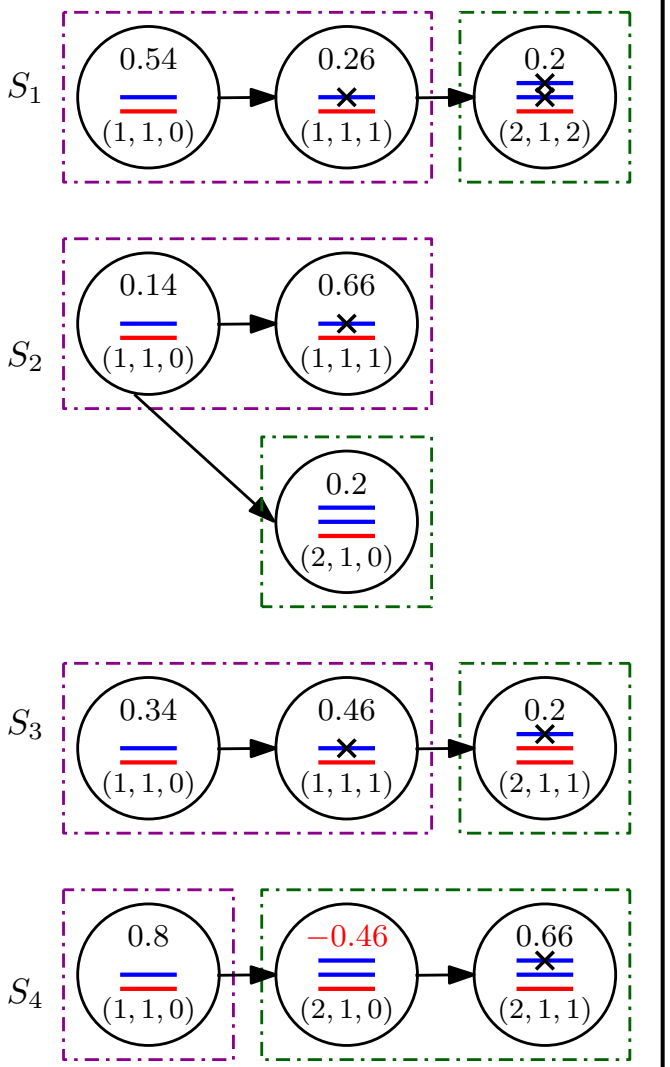
$\mu_{(1,1)} = 0.8$ $\mu_{(2,1)} = 0.2$

VAF



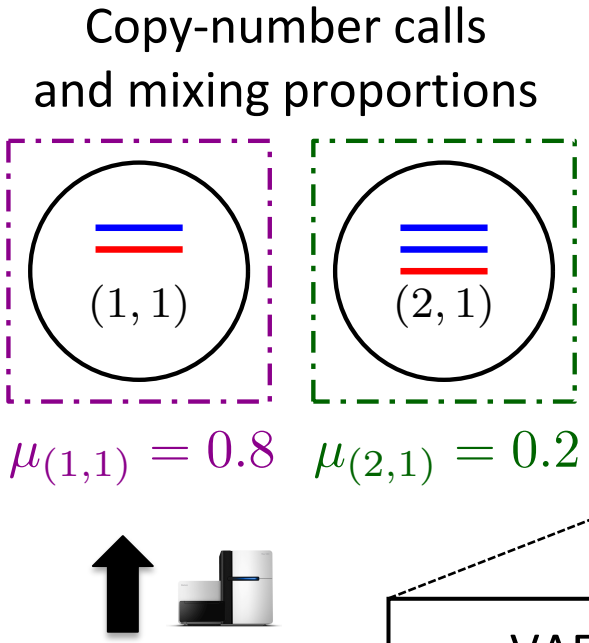
$h = 3/10 = 0.3$

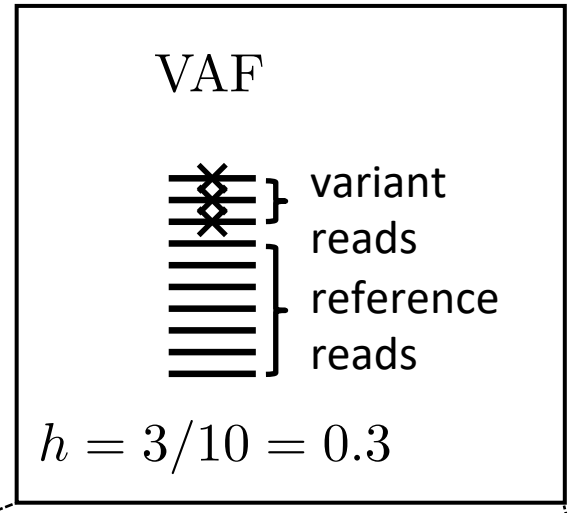
State Trees

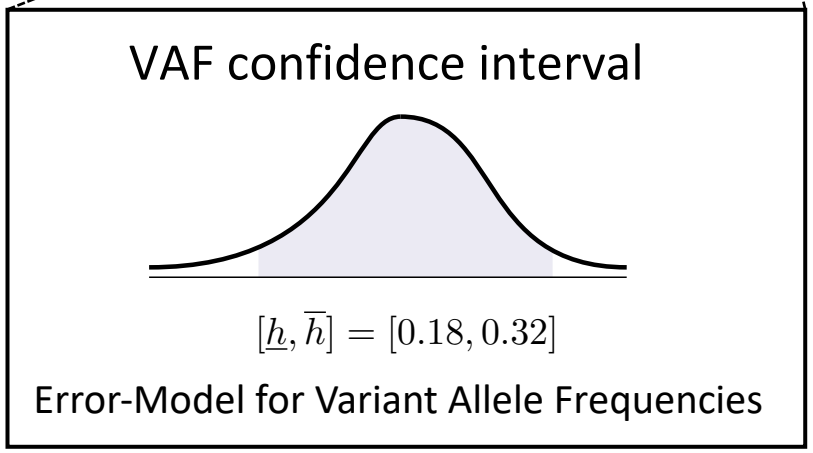


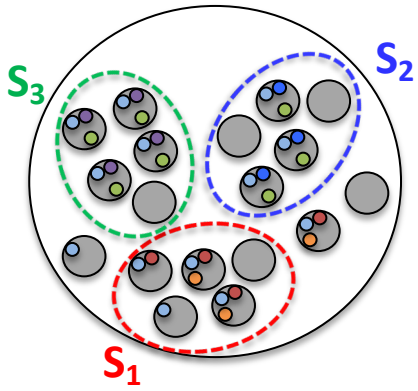
SPRUCE Enumerates Phylogenies

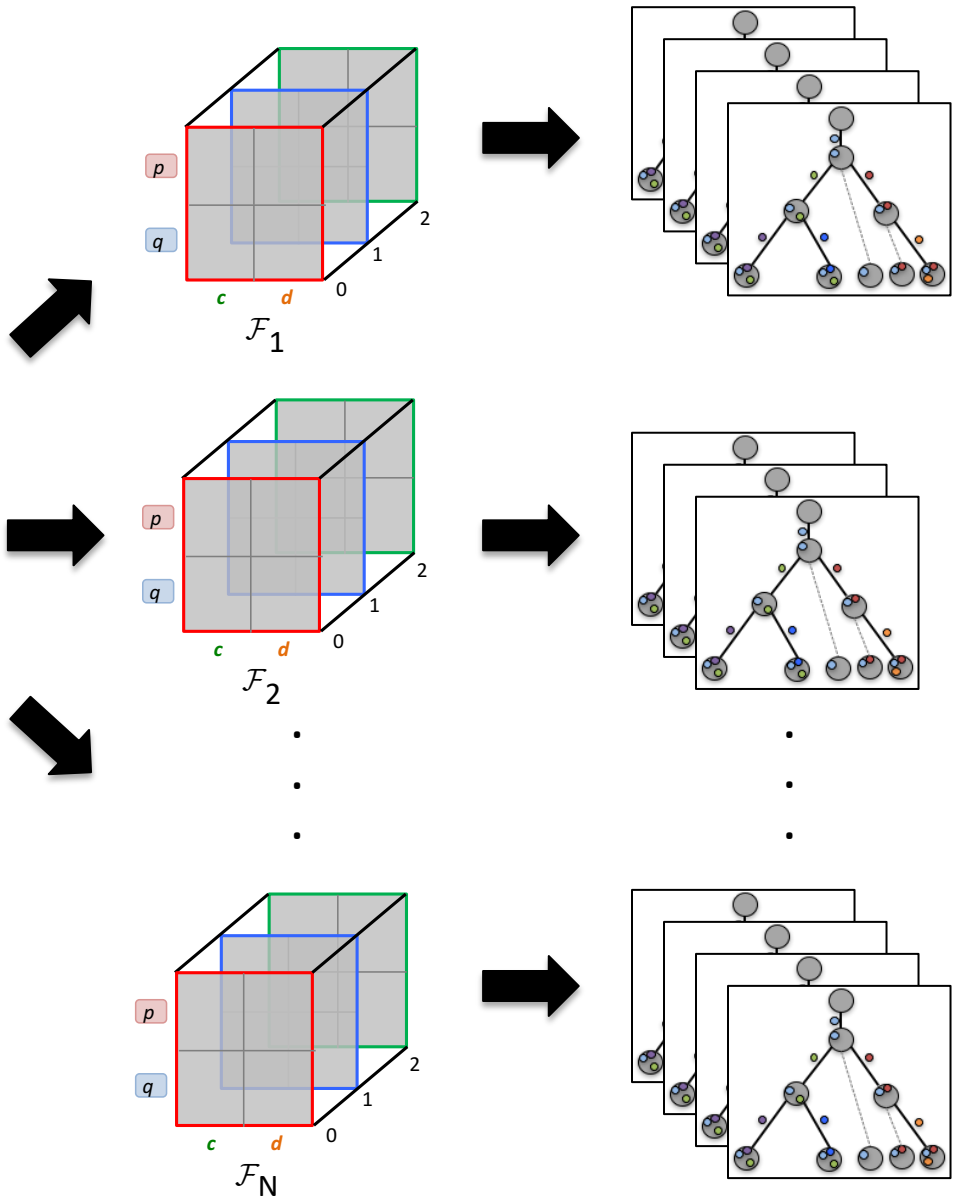
Somatic **P**hylogeny **R**econstruction **U**sing **C**ombinatorial **E**numeration
Available at: http://compbio.cs.brown.edu/projects/spruce/

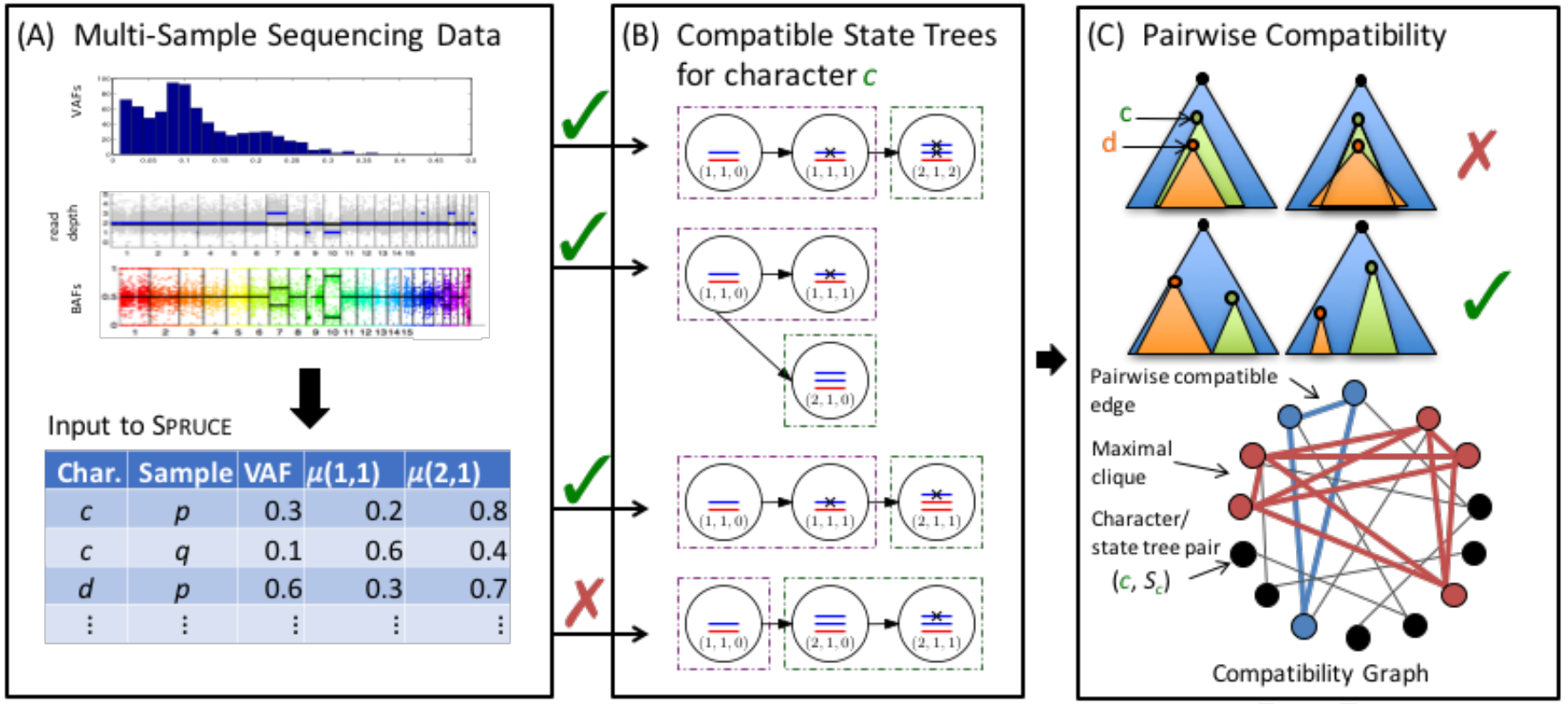


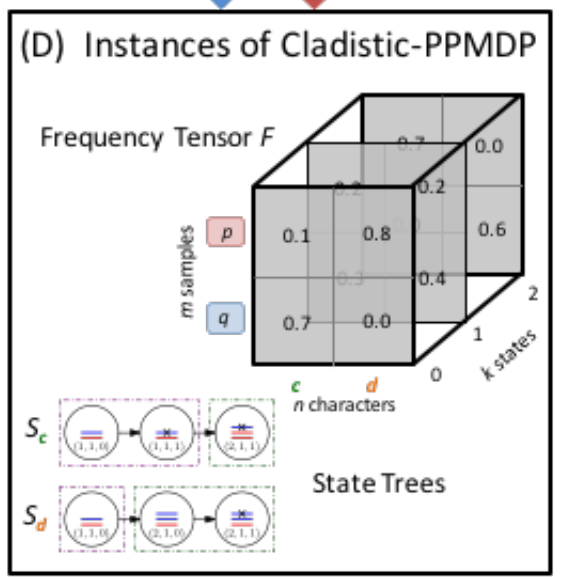


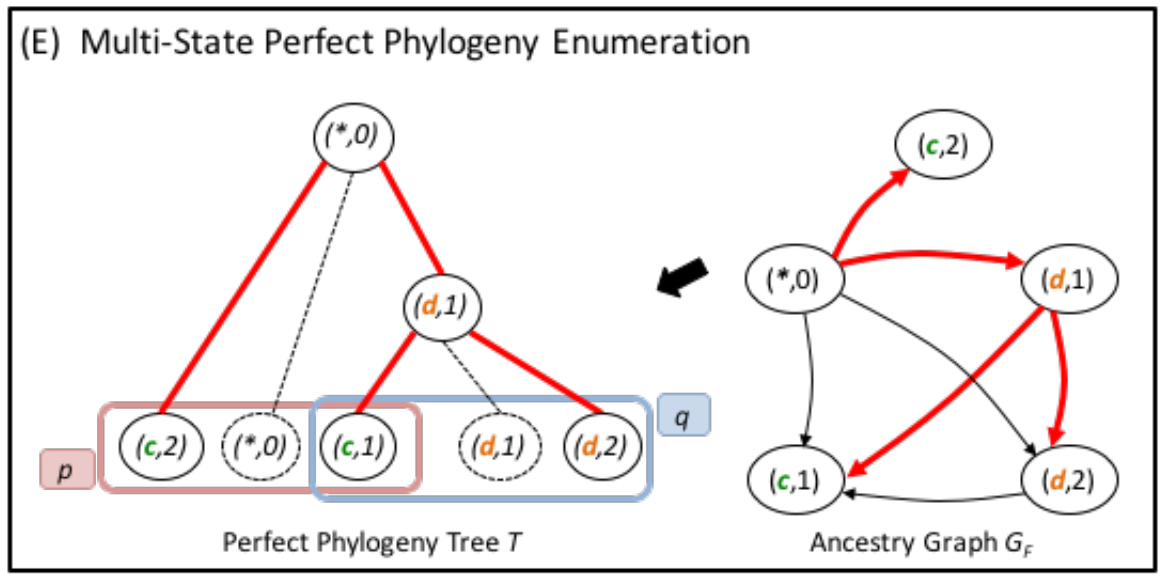








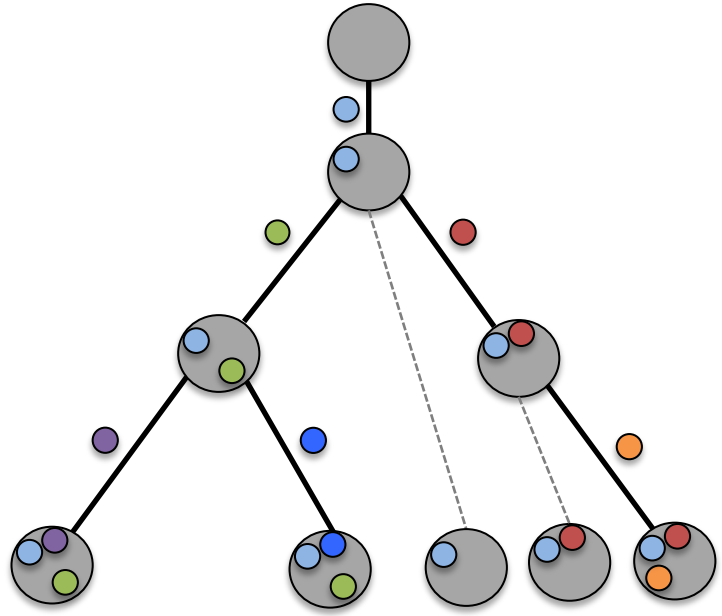




SPRUCE:
Somatic Phylogeny Reconstruction Using Combinatorial Enumeration

Outline

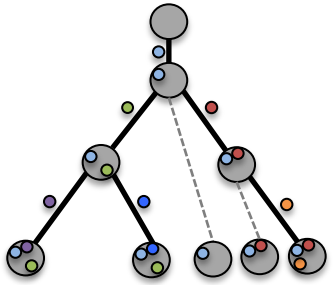
- Multi-State Perfect Phylogeny Mixture Deconvolution Problem
- Combinatorial Characterization of Solutions
- Application to Cancer Sequencing
- Results



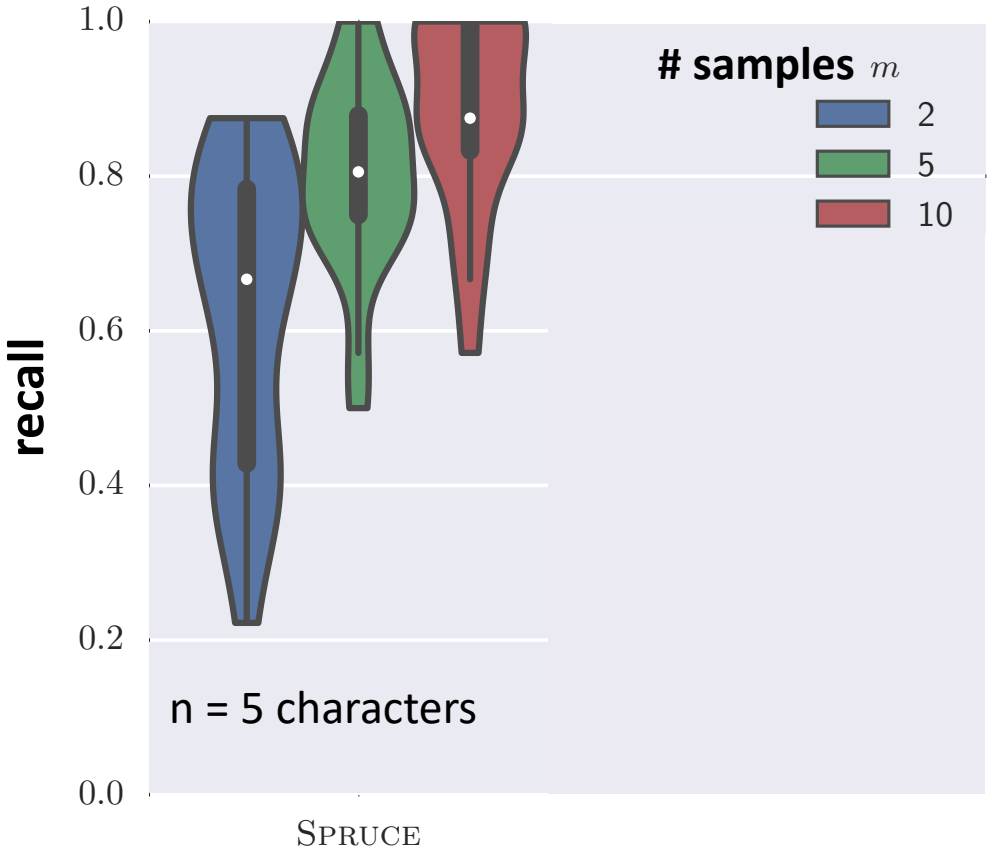
SPRUCE Accurately Recovers Simulated Trees

Parameters:

- # characters ***n***: 5, 15



- # samples ***m***: 2, 5, 10



Methods:

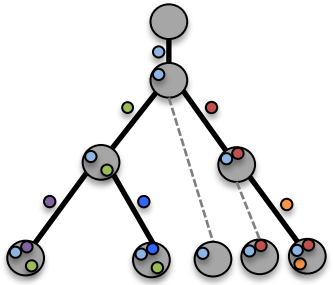
- SPRUCE
- PhyloWGS [Deshwar et al., 2015]

Increasing number of samples decreases ambiguity

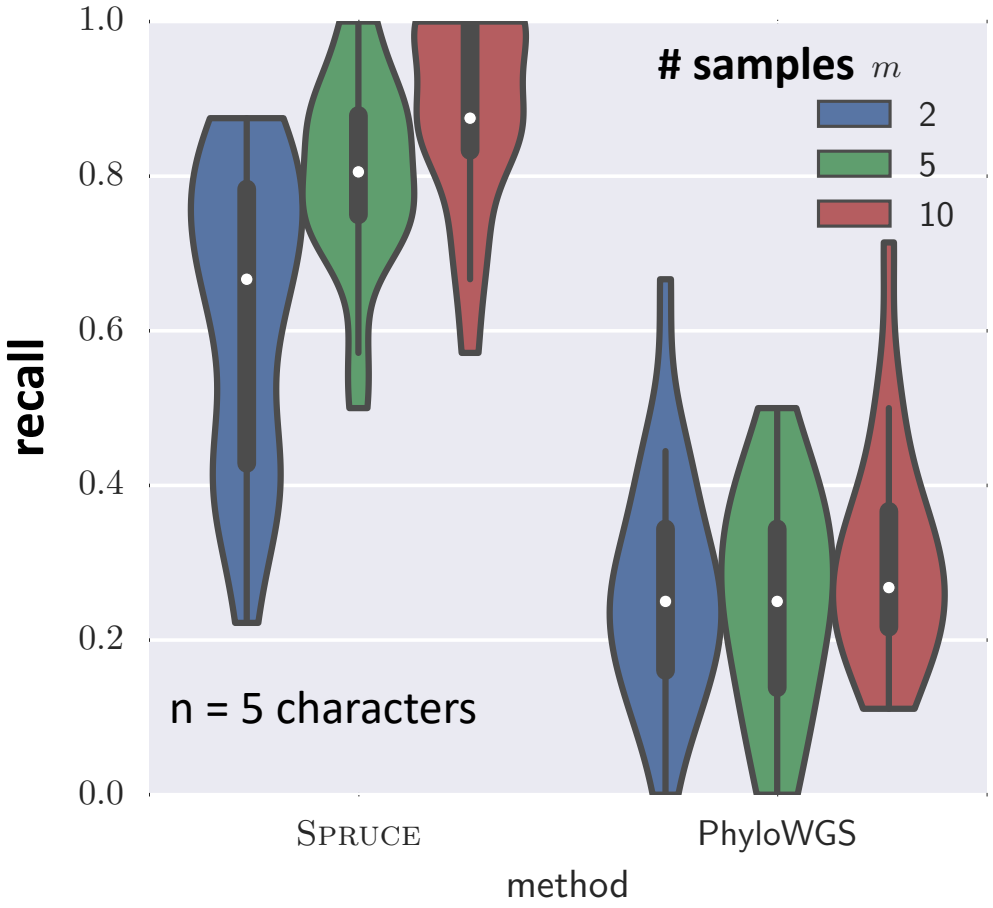
SPRUCE Accurately Recovers Simulated Trees

Parameters:

- # characters ***n***: 5, 15



- # samples ***m***: 2, 5, 10



Methods:

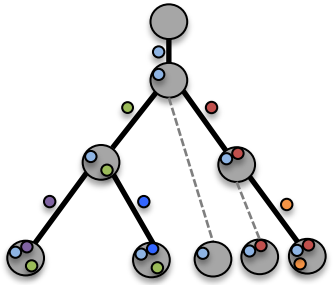
- SPRUCE
- PhyloWGS [Deshwar et al., 2015]

Increasing number of samples decreases ambiguity

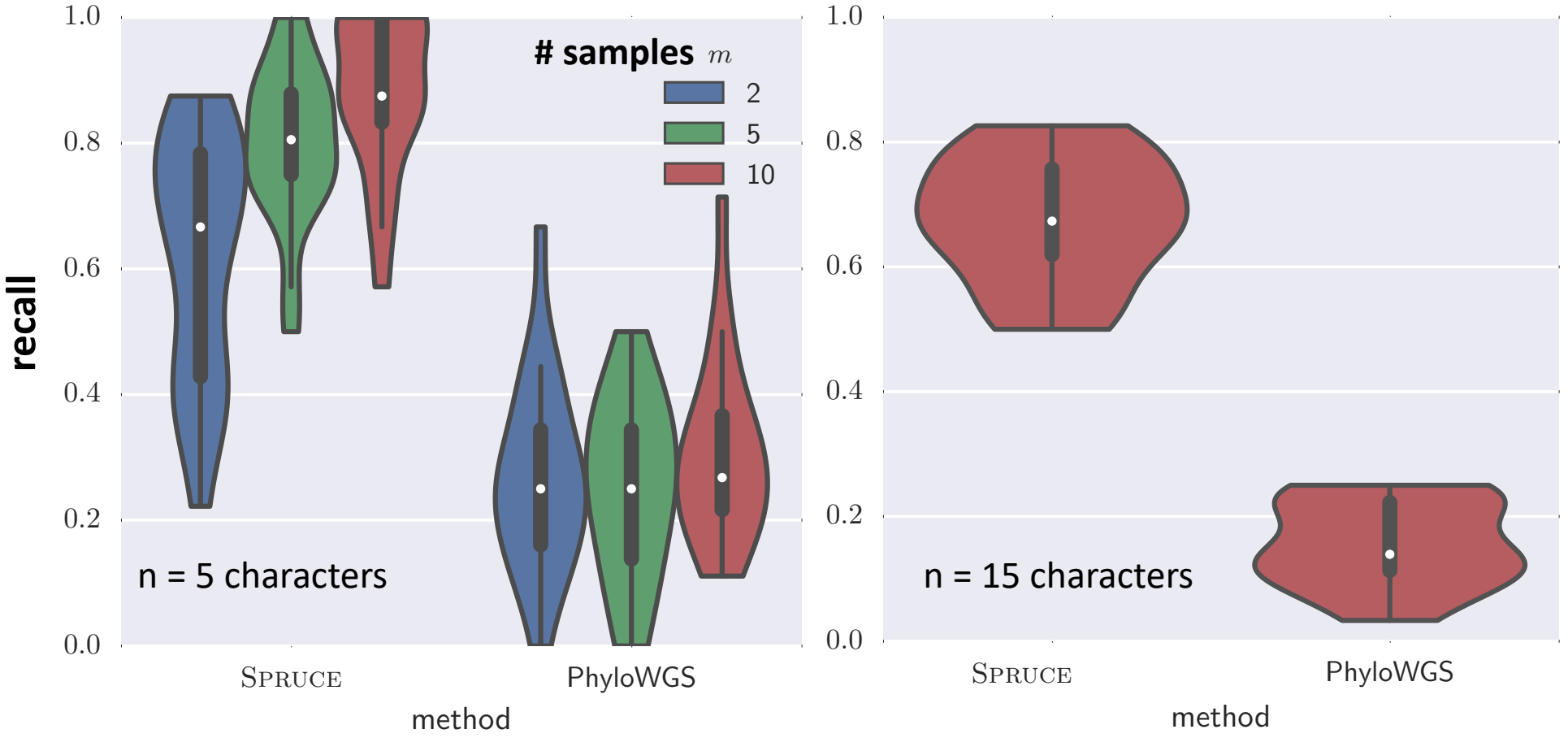
SPRUCE Accurately Recovers Simulated Trees

Parameters:

- # characters ***n***: 5, 15



- # samples ***m***: 2, 5, 10

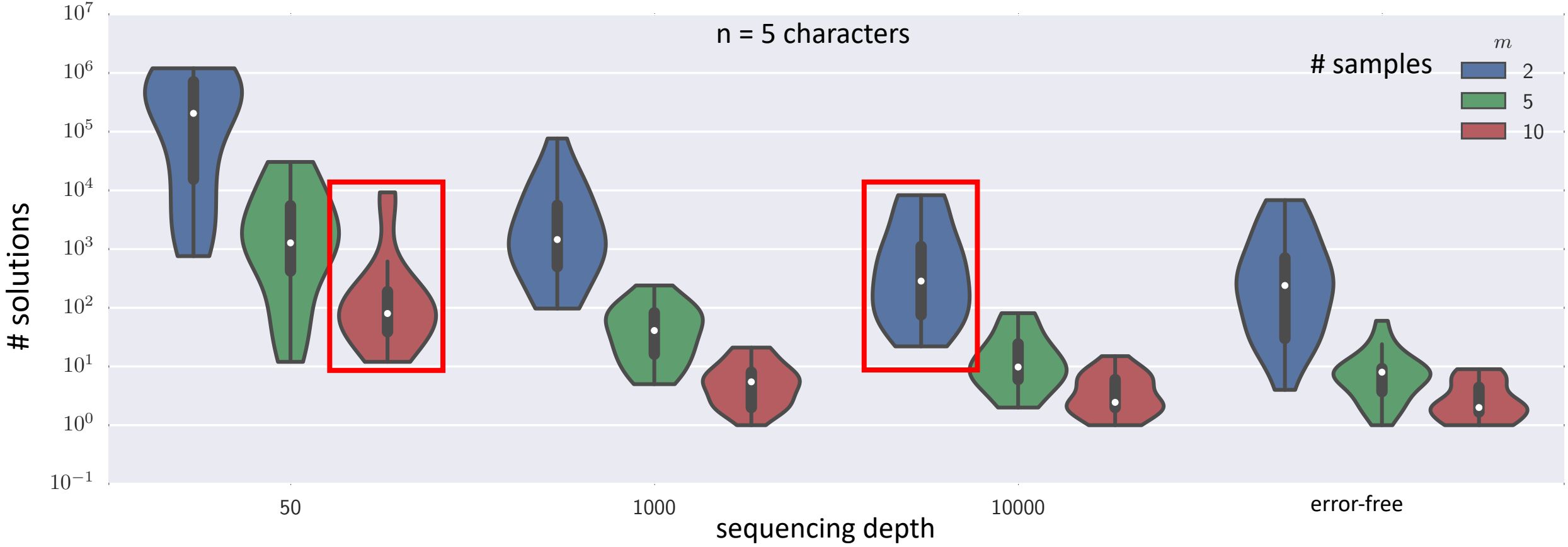


Methods:

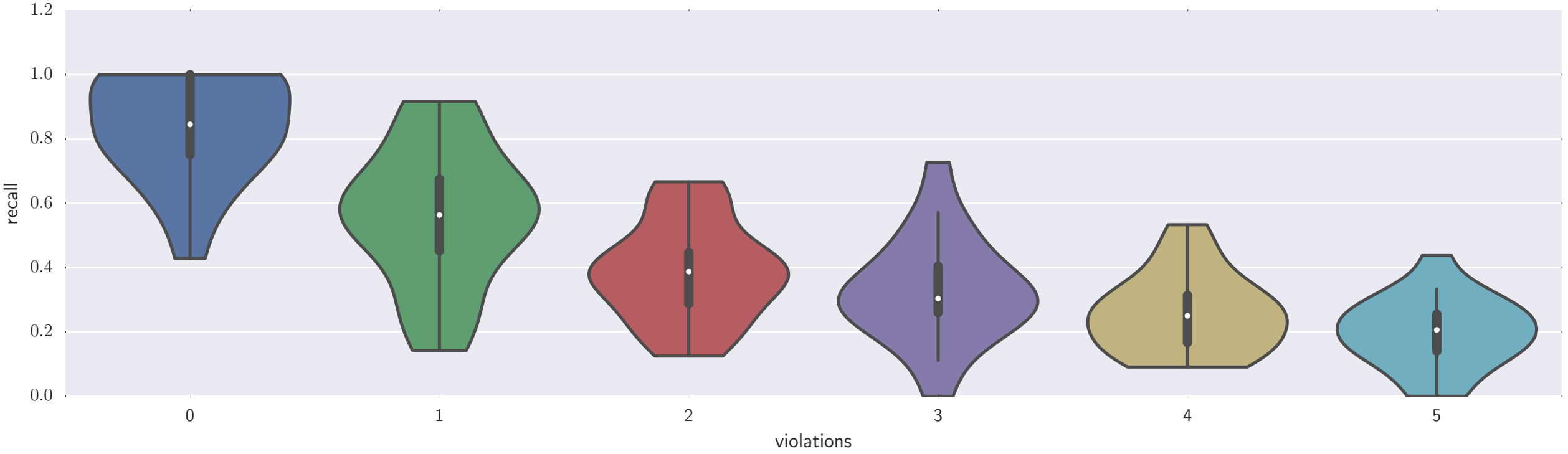
- SPRUCE
- PhyloWGS [Deshwar et al., 2015]

Increasing number of samples decreases ambiguity

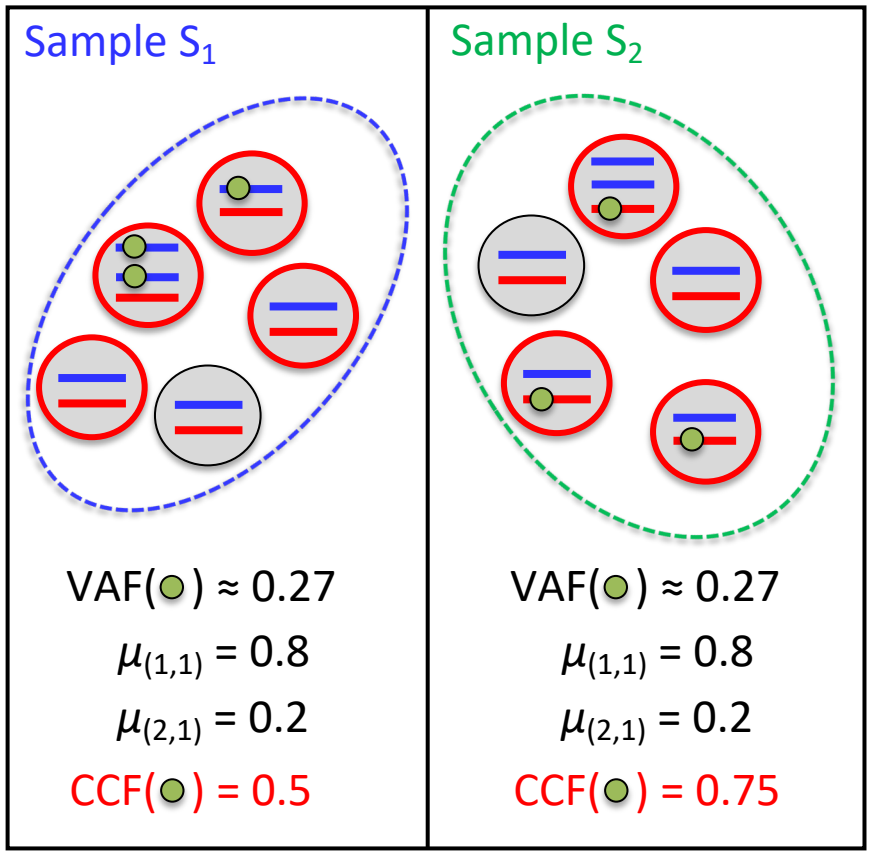
More Samples vs. More Coverage



Violations of Infinite Alleles Assumption



Cancer Cell Fractions (CCFs) Cannot Be Inferred A Priori

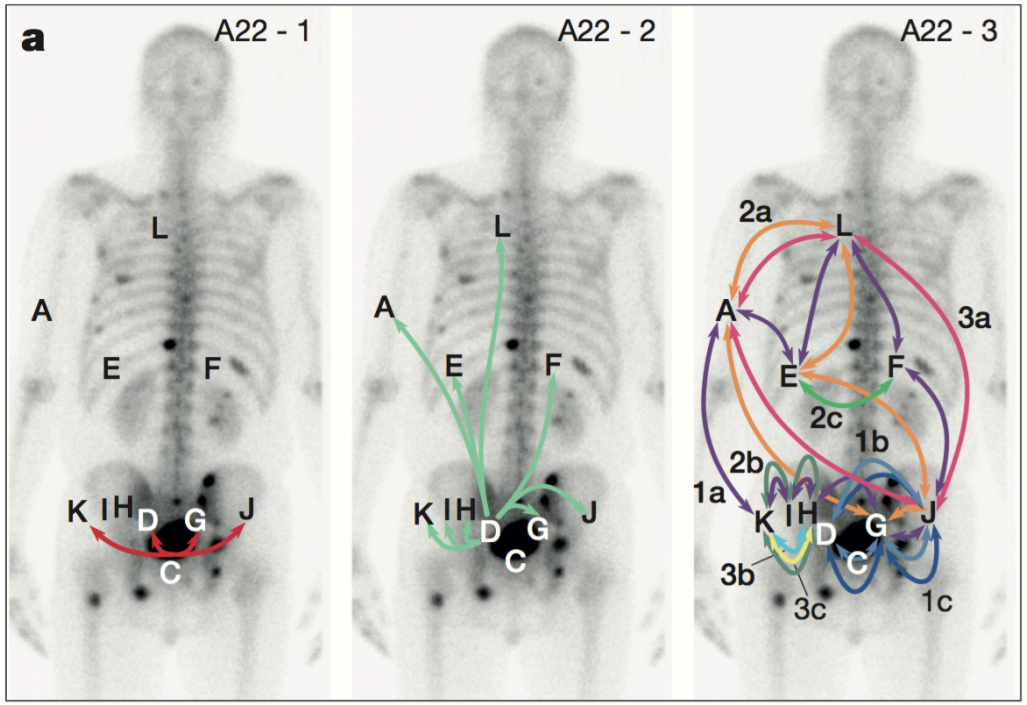


$$\text{CCF}(\bullet) = \frac{\text{\# tumor cells with } \bullet}{\text{\# tumor cells}}$$

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



A - L. humerus BM
D - Sem. vesicle
C - Prostate
E - L. adrenal

F - R. adrenal
G - Bladder
H - Pelvic LN
I - L. pelvic LN

J - R. pelvic LN
K - L. pelvic LN
L - L. media. LN

Adapted from:

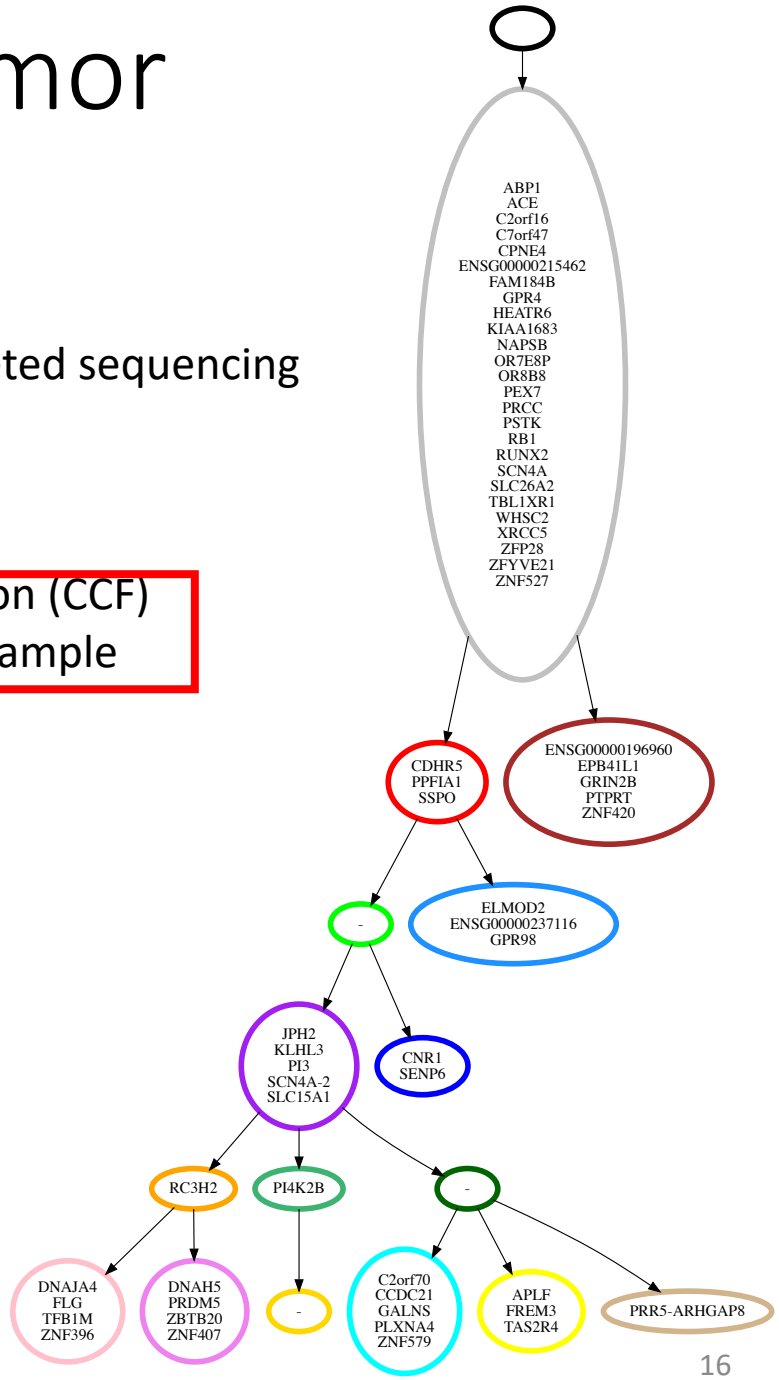
Gundem *et al.* (2015). *Nature*.
The evolutionary history of lethal metastatic prostate cancer

Input:

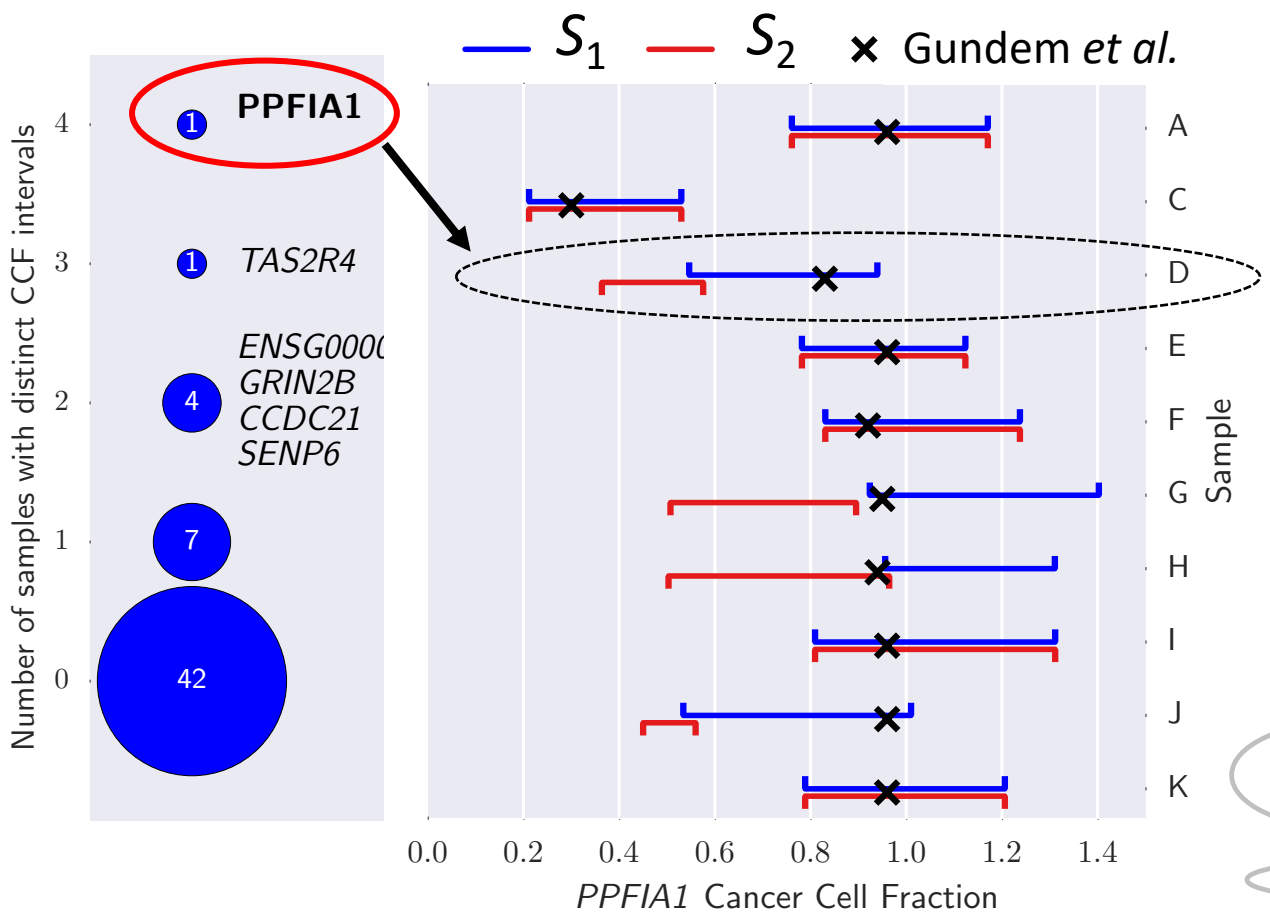
- 10 samples:
  whole-genome & targeted sequencing
- ~110 SNVs

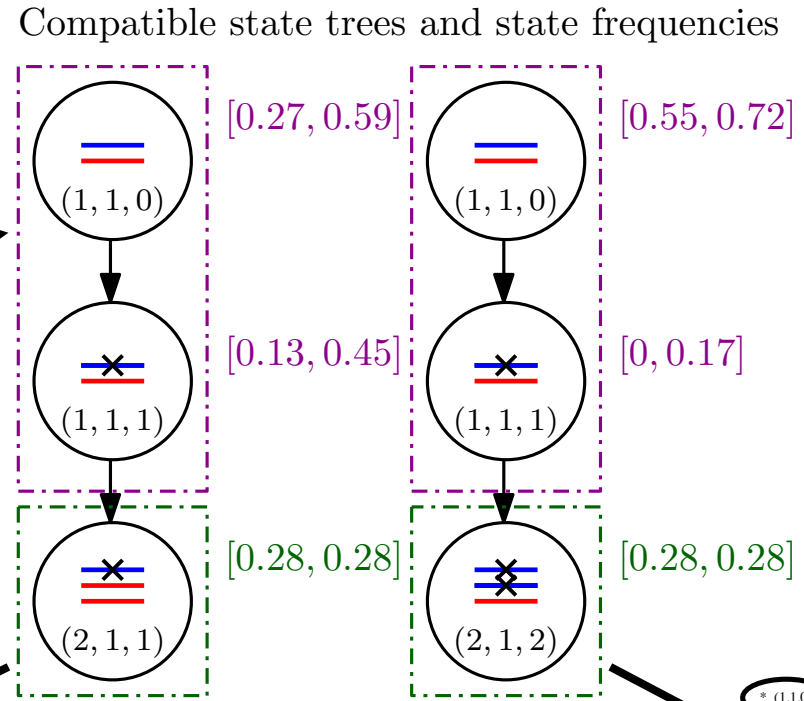
Tree building:

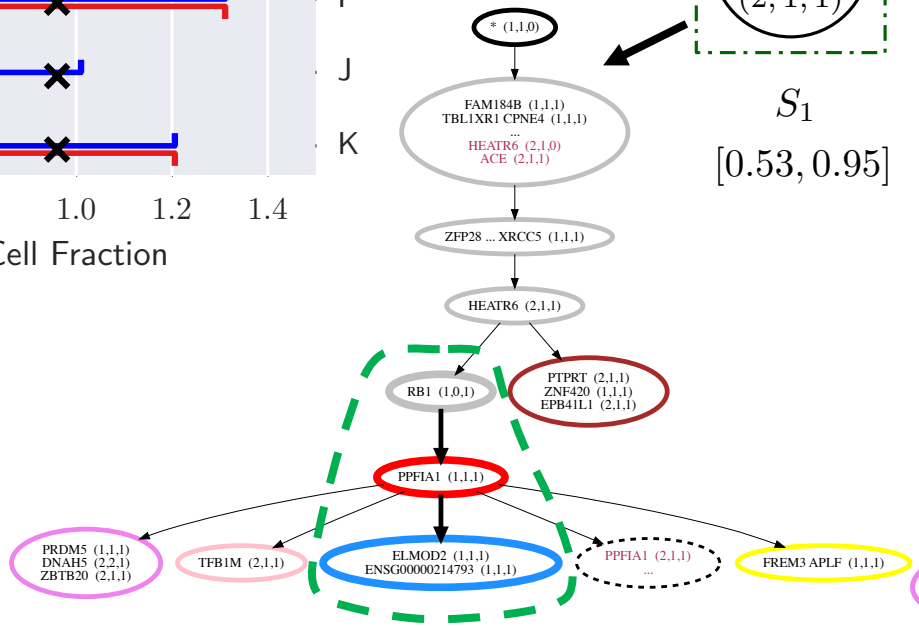
1. Infer cancer cell fraction (CCF) for each SNV in each sample
2. Cluster SNVs by CCFs across samples
3. Construct tree using Pigeon-Hole-Principle (Sum Condition)

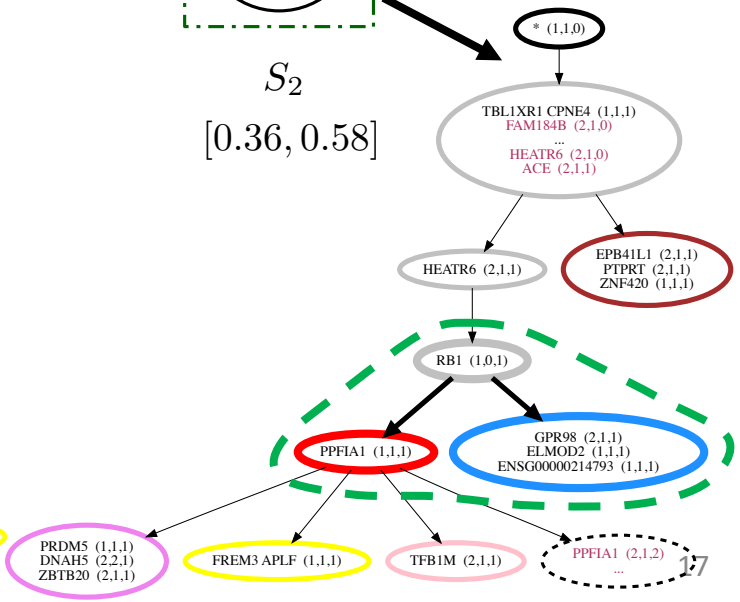


Cancer Cell Fractions Cannot Be Inferred A Priori

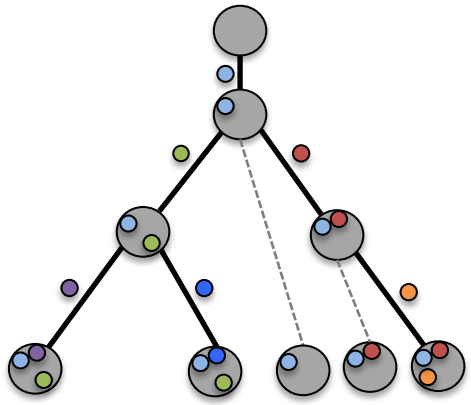








Conclusions



- 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