CS 598MEB
Introduction to Bioinformatics
Lecture 6

Mohammed El-Kebir

January 31, 2019

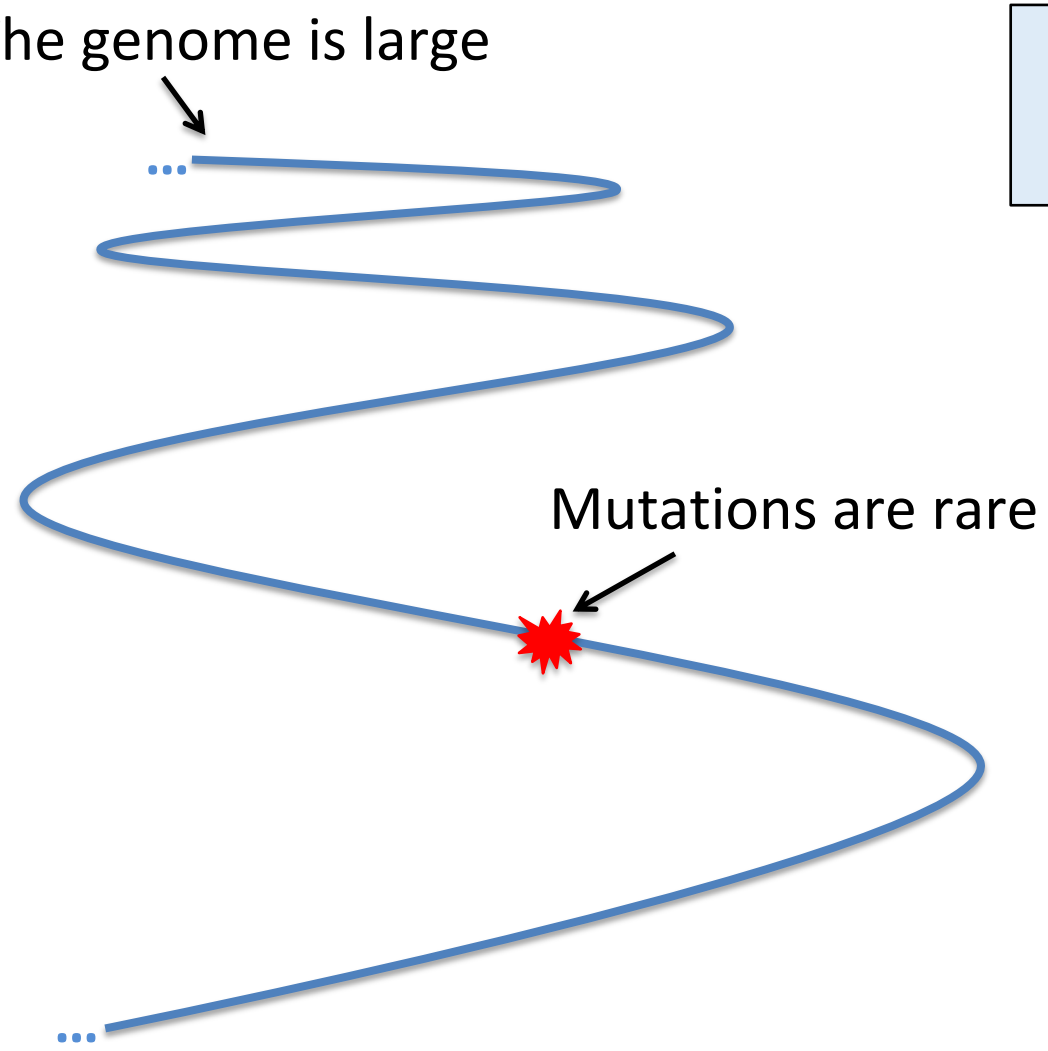
Outline

- Multi-State Perfect Phylogeny
- Tumor Phylogeny Inference from Single-cell DNA-seq

Reading:

- Lecture notes
- M. El-Kebir. SPhyR: Tumor Phylogeny Estimation from Single-Cell Sequencing Data under Loss and Error. Bioinformatics (ECCB 2018), 34(17):i671-679, 2018.

Infinite Sites Model = Two-state Perfect Phylogeny



[Kimura, 1969]

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

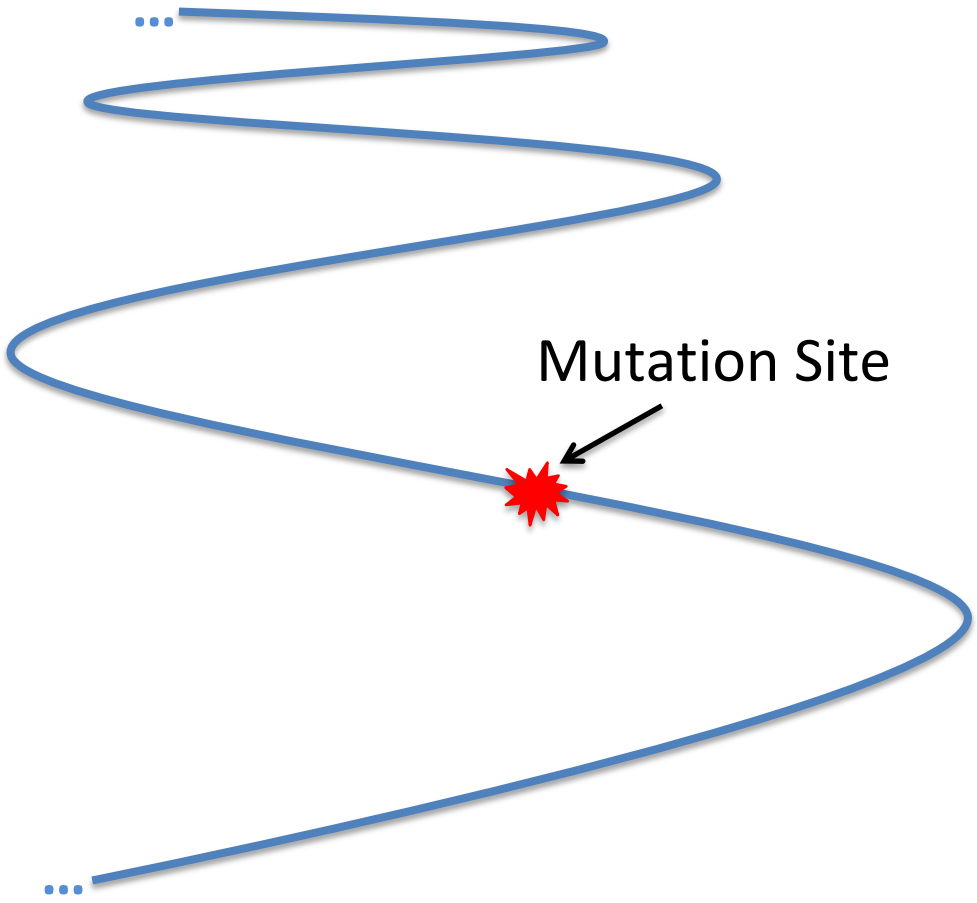
Mutated Loci

Species (cancer cells)						
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

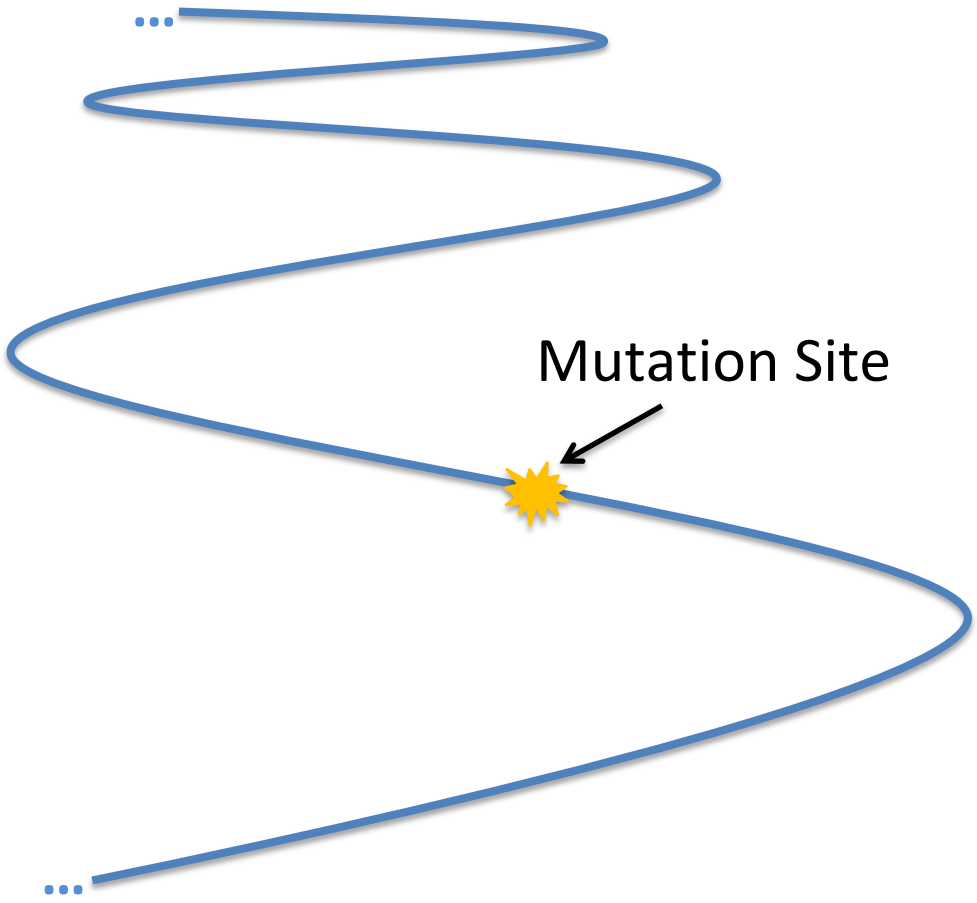


Infinite alleles model:

- 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.

Characters have integer states

Infinite Alleles Model = Multi-state Perfect Phylogeny

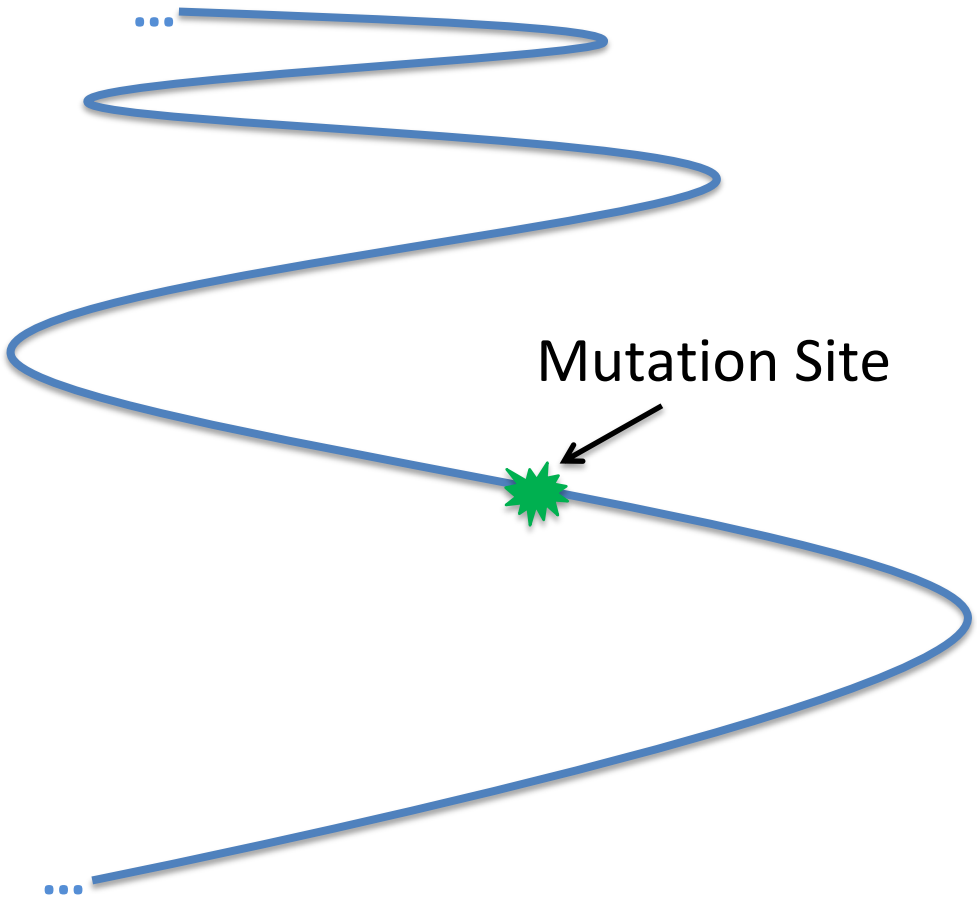


Infinite alleles model:

- 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.

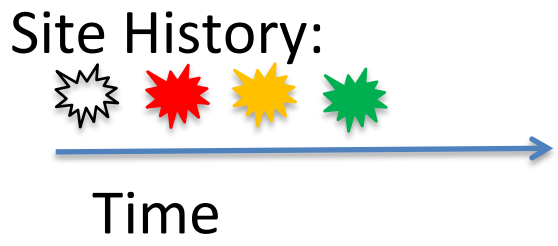
Characters have integer states

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.



Characters have integer states

Multi-state Perfect Phylogeny

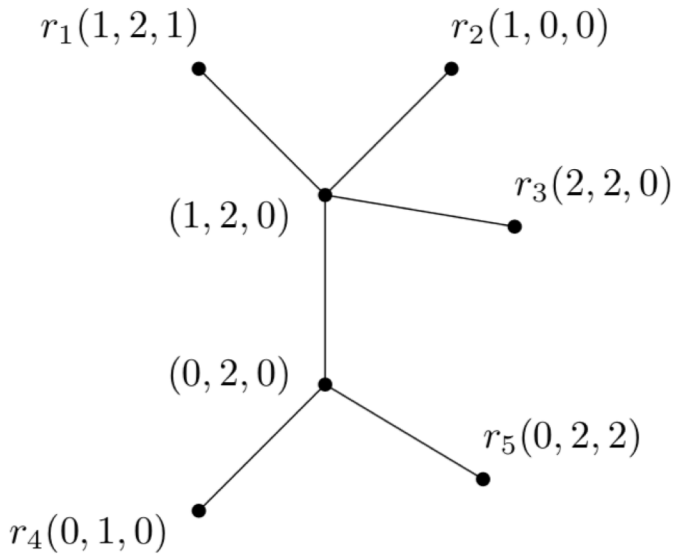
Matrix $M \in \{0, \ldots, k-1\}^{n \times m}$ has
n taxa and m characters

	c_1	c_2	c_3
r_1	1	2	1
r_2	1	0	0
r_3	2	2	0
r_4	0	1	0
r_5	0	2	2

Definition

A multi-state perfect phylogeny for M is a tree T with n leaves such that:

1. Each taxon labels exactly one leaf
2. Each node is labeled by $\{0, \ldots, k-1\}^m$
3. Nodes labeled with state i for character c form a connected subtree $T_c(i)$



Theorem (Bodlaender et al., 1992) [Bodlaender, Fellows and Warnow]

For general k, the multi-state perfect phylogeny problem is NP-complete

Cladistic vs. Qualitative Characters

Definition

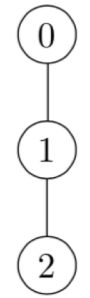
A multi-state perfect phylogeny for M is a tree T with n leaves such that:

1. Each taxon labels exactly one leaf
2. Each node is labeled by $\{0, \ldots, k-1\}^m$
3. Nodes with state i for character c form a connected subtree $T_c(i)$

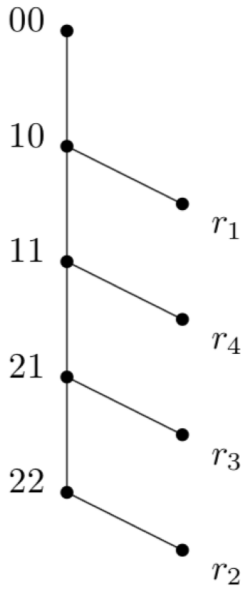
A cladistic character c has a state tree t_c on its states

A phylogeny T is consistent if the reduced tree $\sigma(T, c)$ is identical with t_c for all c

	a	b
r_1	1	0
r_2	2	2
r_3	2	1
r_4	1	1



t_a, t_b



Cladistic vs. Qualitative Characters

Definition

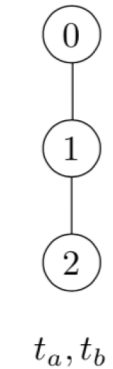
A multi-state perfect phylogeny for M is a tree T with n leaves such that:

1. Each taxon labels exactly one leaf
2. Each node is labeled by $\{0, \ldots, k-1\}^m$
3. Nodes with state i for character c form a connected subtree $T_c(i)$

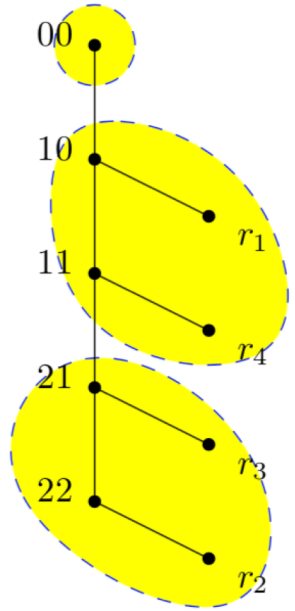
A cladistic character c has a state tree t_c on its states

A phylogeny T is consistent if the reduced tree $\sigma(T, c)$ is identical with t_c for all c

	a	b
r_1	1	0
r_2	2	2
r_3	2	1
r_4	1	1



t_a, t_b



Cladistic vs. Qualitative Characters

Definition

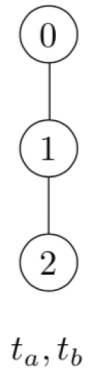
A multi-state perfect phylogeny for M is a tree T with n leaves such that:

1. Each taxon labels exactly one leaf
2. Each node is labeled by $\{0, \ldots, k-1\}^m$
3. Nodes with state i for character c form a connected subtree $T_c(i)$

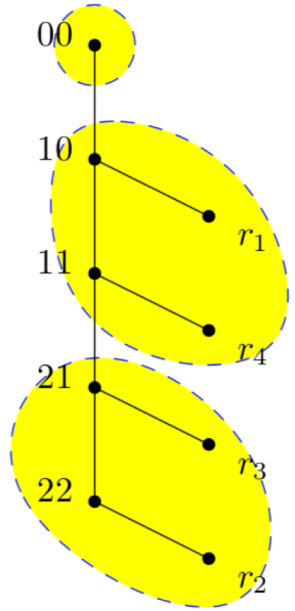
A cladistic character c has a state tree t_c on its states

A phylogeny T is consistent if the reduced tree $\sigma(T, c)$ is identical with t_c for all c

	a	b
r_1	1	0
r_2	2	2
r_3	2	1
r_4	1	1



t_a, t_b



Outline

- Multi-State Perfect Phylogeny
- Tumor Phylogeny Inference from Single-cell DNA-seq

Reading:

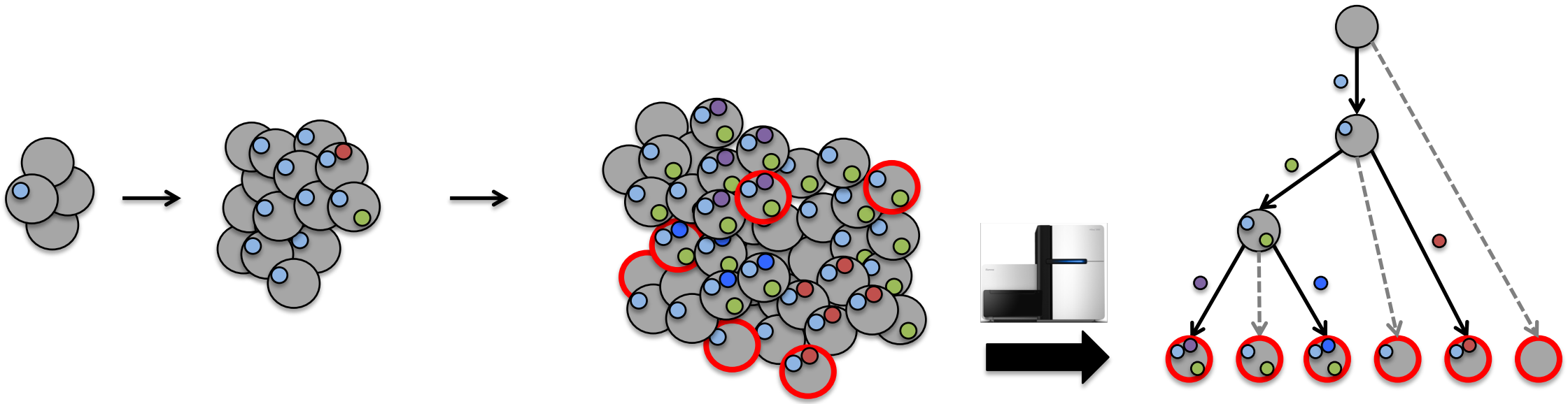
- Lecture notes
- M. El-Kebir. SPhyR: Tumor Phylogeny Estimation from Single-Cell Sequencing Data under Loss and Error. Bioinformatics (ECCB 2018), 34(17):i671-679, 2018.

SPhyR: Tumor Phylogeny Estimation from Single-Cell Sequencing Data under Loss and Error

Mohammed El-Kebir – University of Illinois at Urbana Champaign, Department of Computer Science

ECCB 2018

Single-cell Tumor Phylogeny Inference



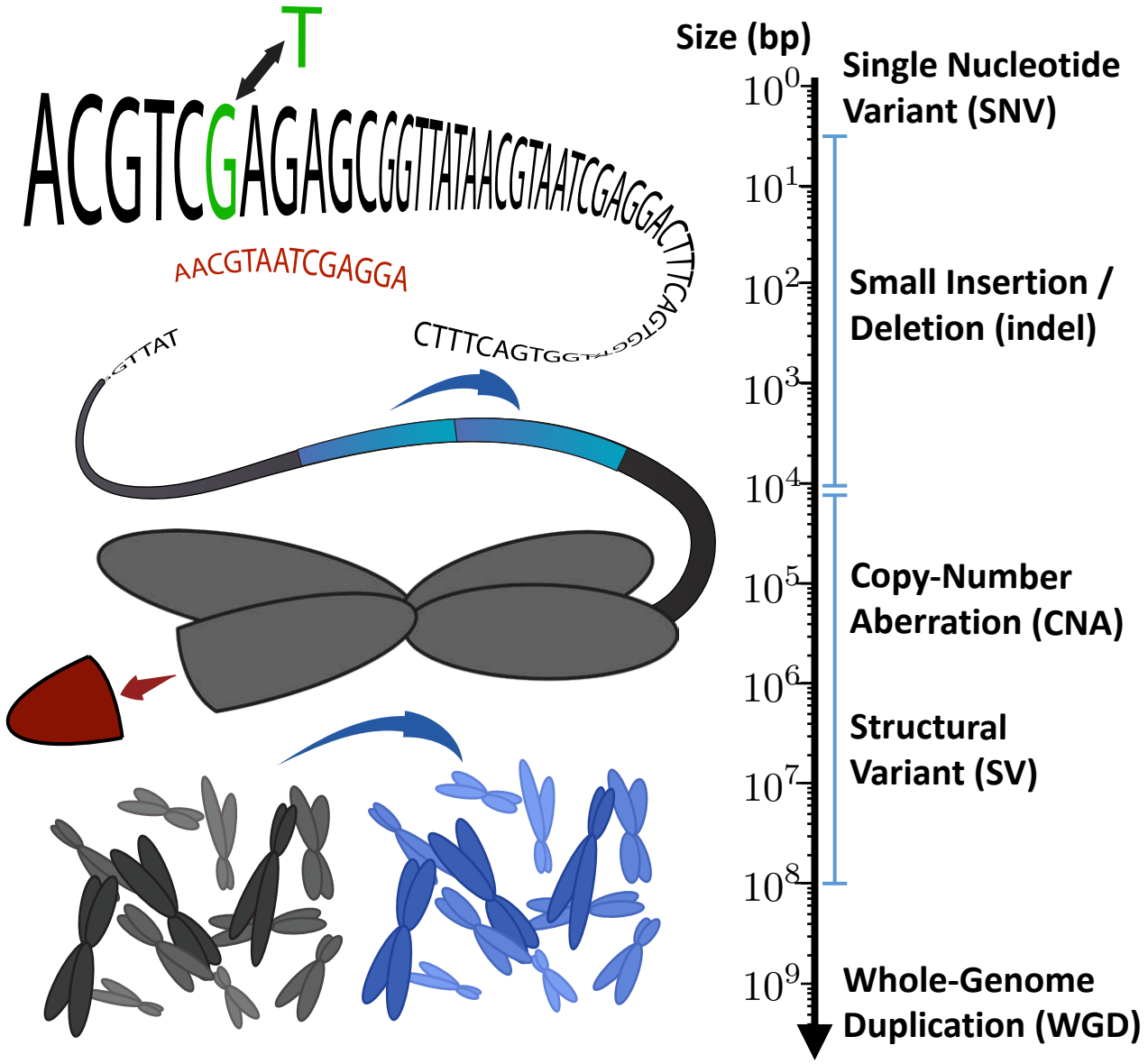
Heterogeneous Tumor

Phylogenetic Tree ***T***

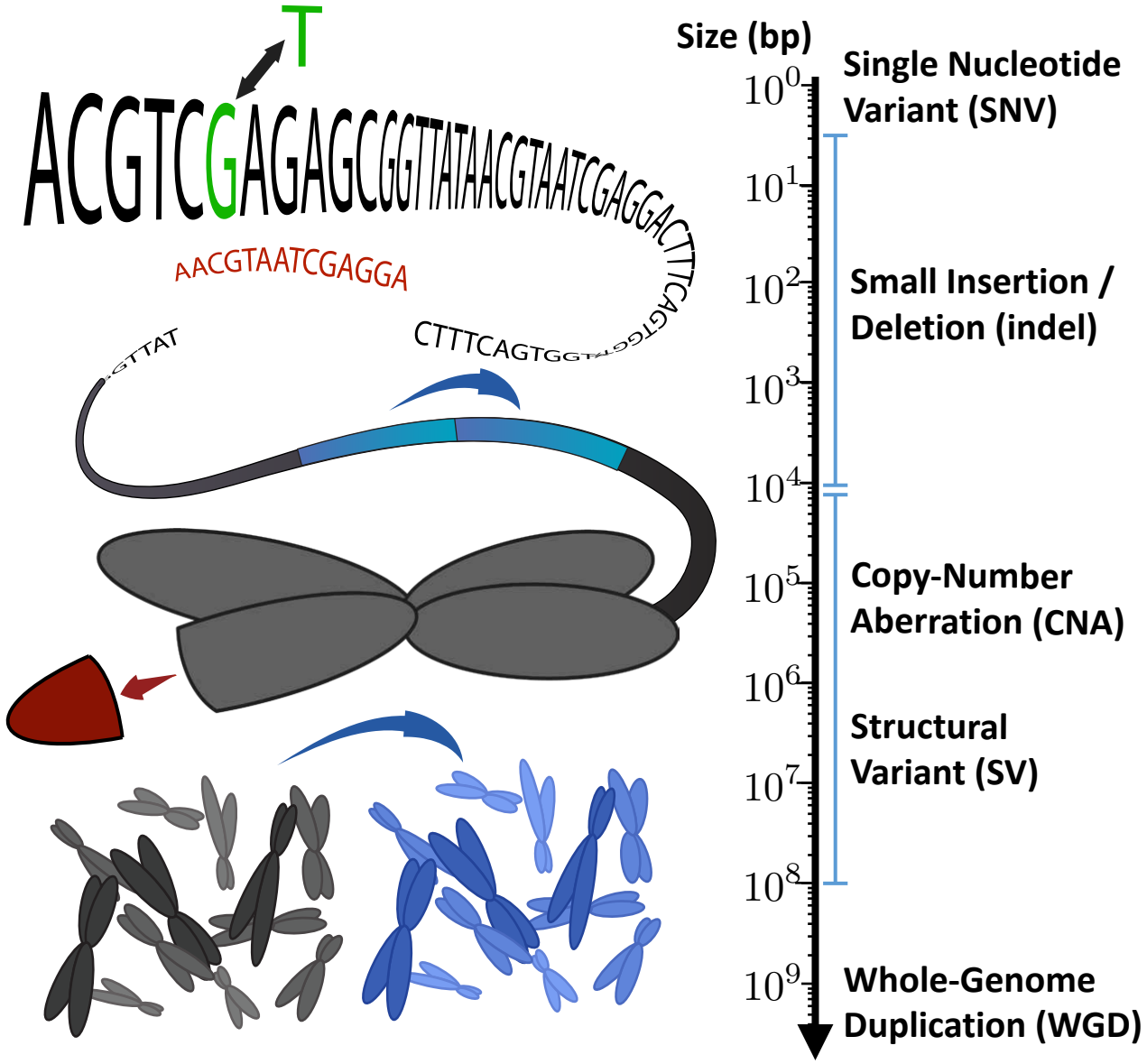
Goal: Given single-cell DNA sequencing data, find phylogenetic tree ***T***

Requirement: Evolutionary model for somatic mutations

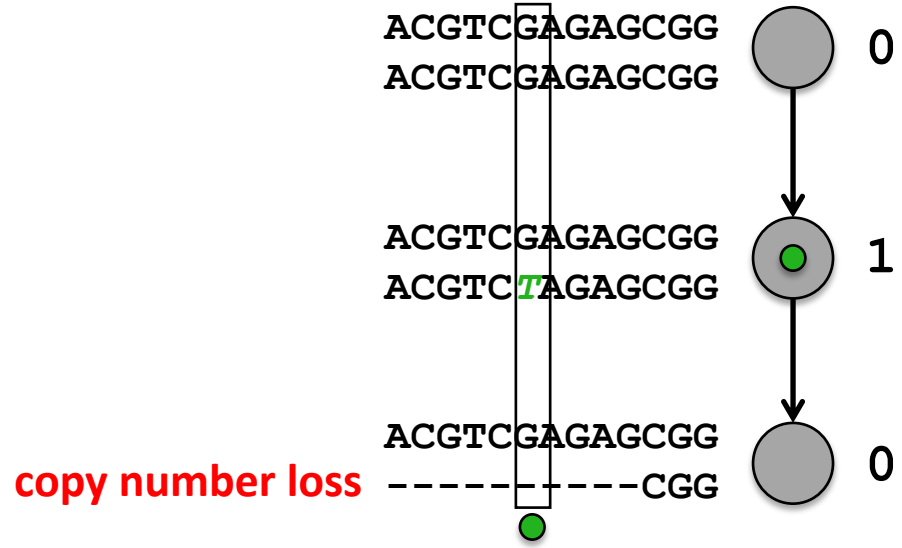
Somatic Mutations Occur at Different Genomic Scales



Infinite Sites Assumption is too Restrictive for SNVs



SNVs can be **lost** due to CNAs

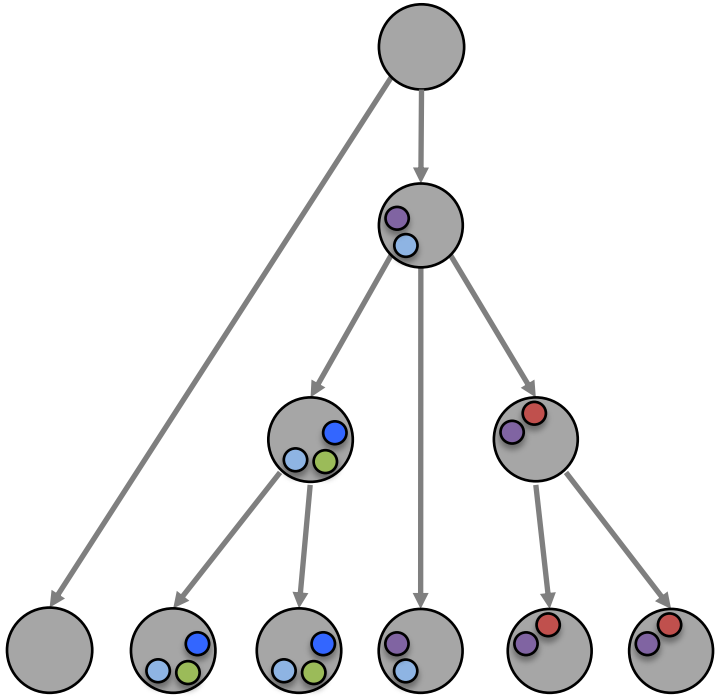


Infinite sites assumption:

- No parallel evolution of SNVs
- No loss of SNVs
- SCITE [Jahn et al. 2016]
- OncoNEM [Ross and Markowetz, 2016]

Outline

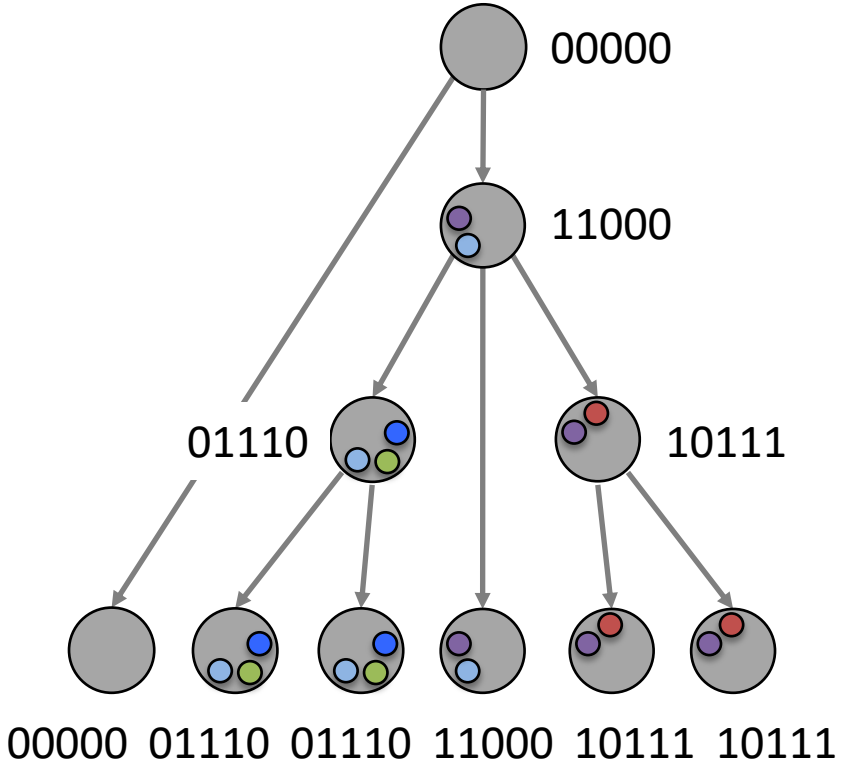
- Perfect data (error free)
  - Problem statement
  - Combinatorial characterization of solutions
  - Exact algorithm
  - Results
- Real data (with errors)
  - Problem statement
  - Heuristic algorithm
  - Results
- Conclusions



k-Dollo Phylogeny (*k*-DP) Problem

Definition 1. A *k-Dollo phylogeny* T is a rooted, node-labeled tree subject to the following conditions.

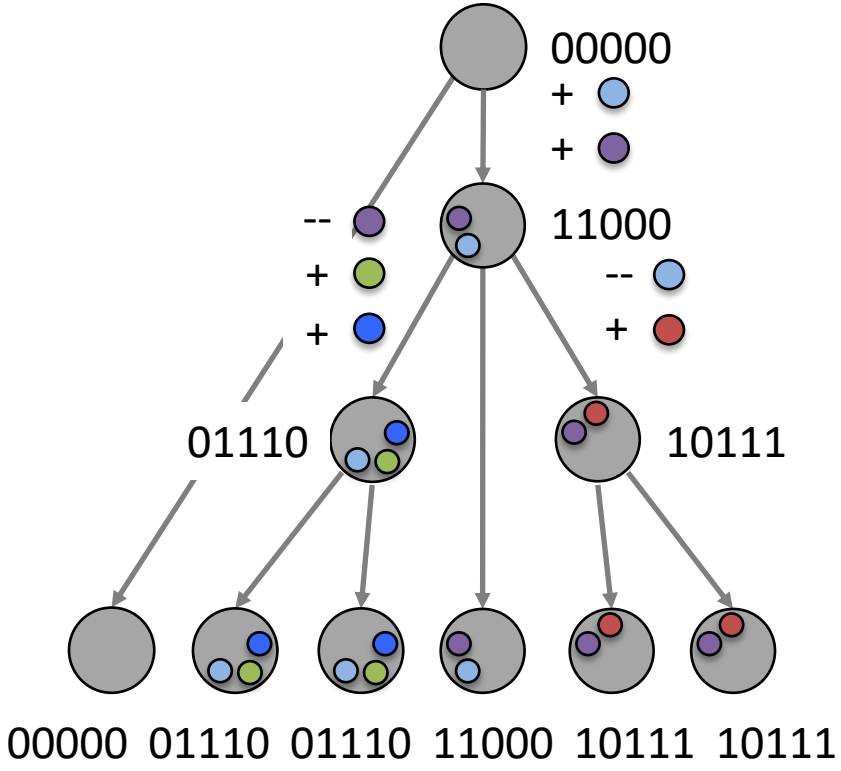
1. Each node v of T is labeled by a vector $\mathbf{b}_v \in \{0,1\}^n$.
2. The root r of T is labeled by vector $\mathbf{b}_r = [0,\ldots,0]^T$.



k-Dollo Phylogeny (*k*-DP) Problem

Definition 1. A *k-Dollo phylogeny* T is a rooted, node-labeled tree subject to the following conditions.

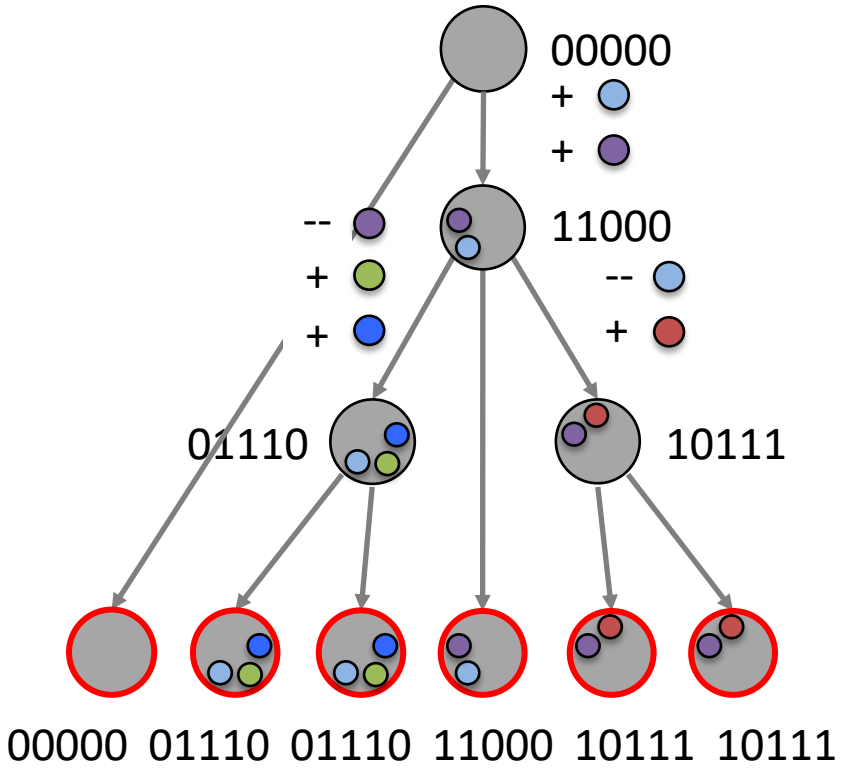
1. Each node v of T is labeled by a vector $\mathbf{b}_v \in \{0,1\}^n$.
2. The root r of T is labeled by vector $\mathbf{b}_r = [0,\ldots,0]^T$.
3. For each character $c \in [n]$, there is exactly one *gain edge* (v,w) in T such that $b_{v,c} = 0$ and $b_{w,c} = 1$.
4. For each character $c \in [n]$, there are at most k *loss edges* (v,w) in T such that $b_{v,c} = 1$ and $b_{w,c} = 0$.



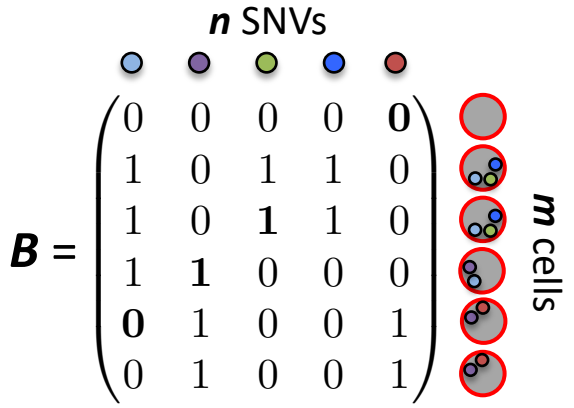
k-Dollo Phylogeny (*k*-DP) Problem

Definition 1. A *k-Dollo phylogeny* T is a rooted, node-labeled tree subject to the following conditions.

1. Each node v of T is labeled by a vector $\mathbf{b}_v \in \{0,1\}^n$.
2. The root r of T is labeled by vector $\mathbf{b}_r = [0,\ldots,0]^T$.
3. For each character $c \in [n]$, there is exactly one *gain edge* (v,w) in T such that $b_{v,c} = 0$ and $b_{w,c} = 1$.
4. For each character $c \in [n]$, there are at most k *loss edges* (v,w) in T such that $b_{v,c} = 1$ and $b_{w,c} = 0$.



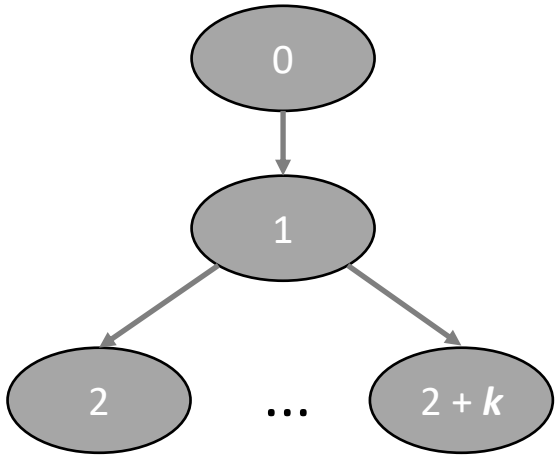
***k*-Dollo Phylogeny problem** (*k*-DP). Given a binary matrix $B \in \{0,1\}^{m\times n}$ and parameter $k \in \mathbb{N}$, determine whether there exists a *k*-Dollo phylogeny for B, and if so construct one.



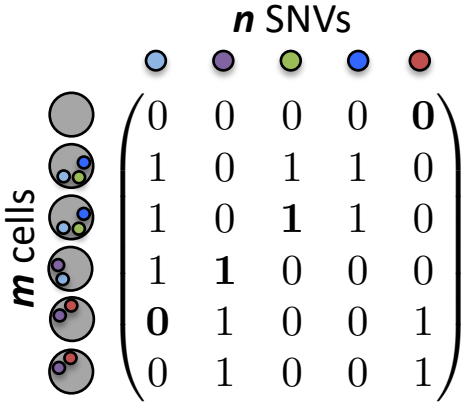
Combinatorial Characterization of *k*-DP

Theorem 3. Let $B \in \{0,1\}^{m \times n}$. The following statements are equivalent.

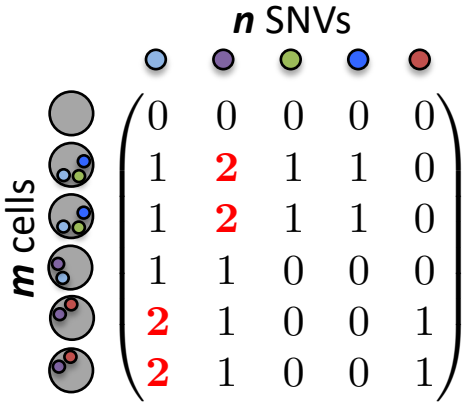
1. There exists a k-Dollo phylogeny T for B.
2. There exists a k-Dollo completion A of B.
3. There exists a k-completion A of B such that the binary factor matrix B' of $(A, \mathcal{S}[k])$ is a perfect phylogeny matrix.
4. There exists a k-completion A of B, and perfect phylogeny T for A whose characters are consistent with $S[k]$.



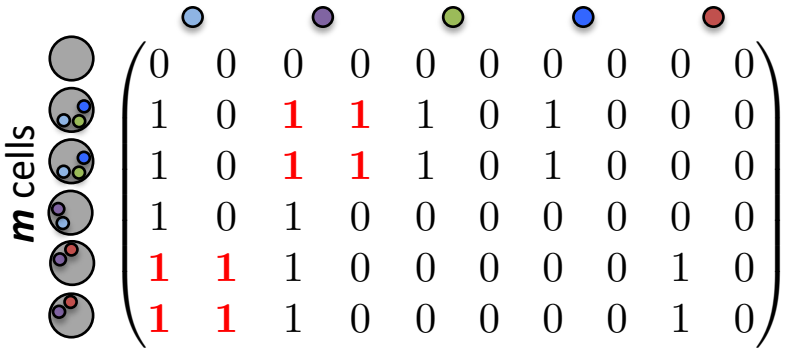
***k*-Dollo State Tree *S*[*k*]**



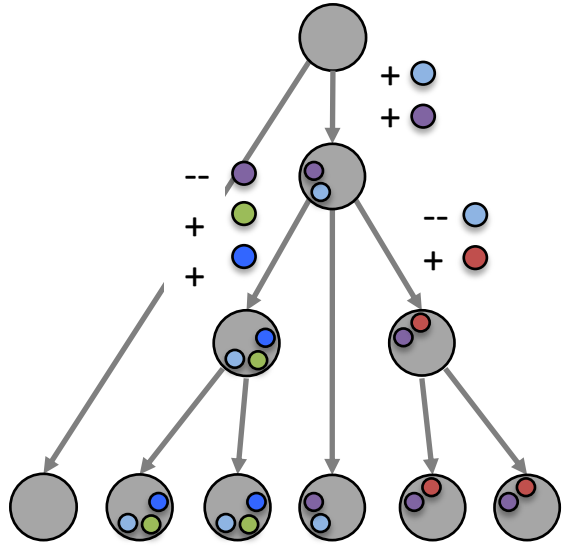
Input Matrix ***B***



k*-Dollo Completion *A



Binary Factor Matrix ***B'***



k*-Dollo Phylogeny *T

Forbidden Submatrices in Solutions A to k-DP

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

$k = 0$

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

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

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

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

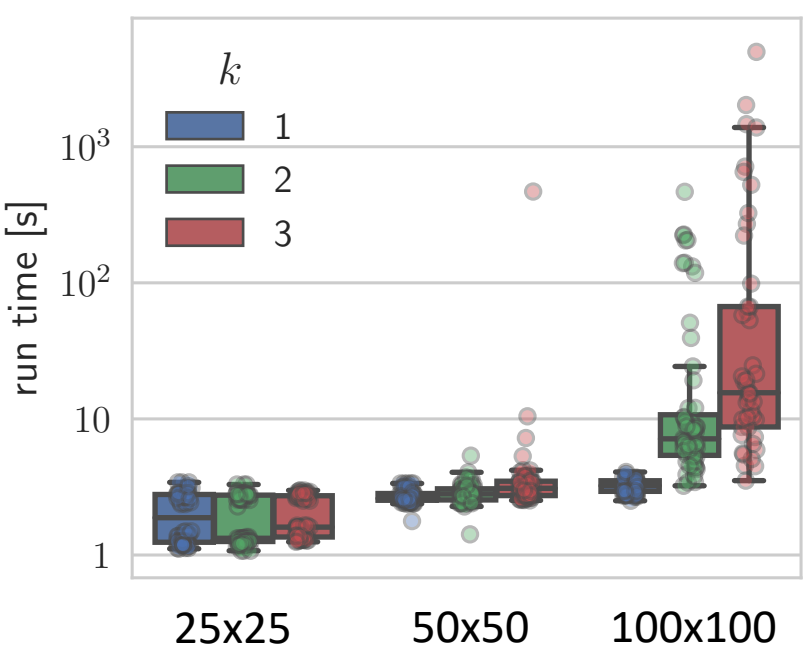
$k = 1$

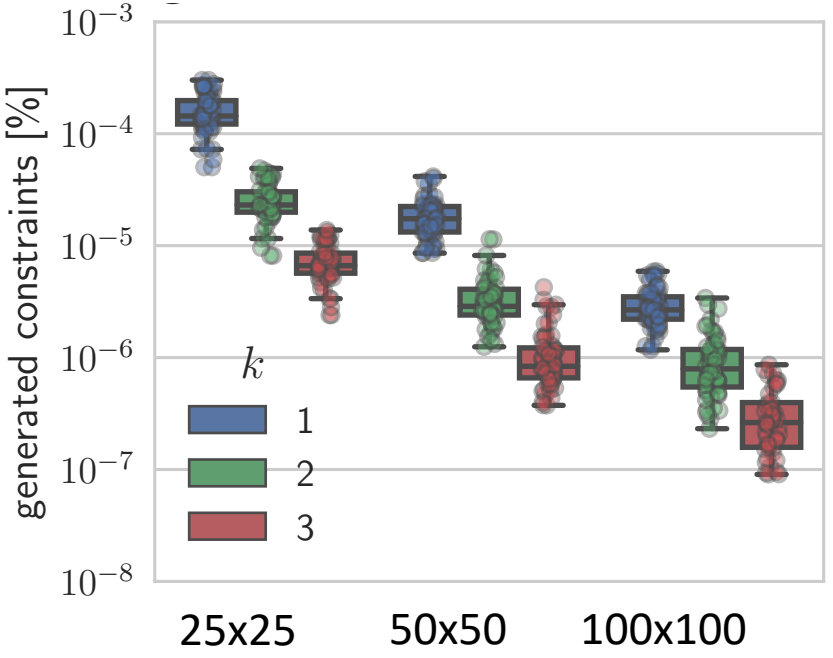
Number of forbidden submatrices is $4k^4 + 8k^3 + 8k^2 + 4k + 1$

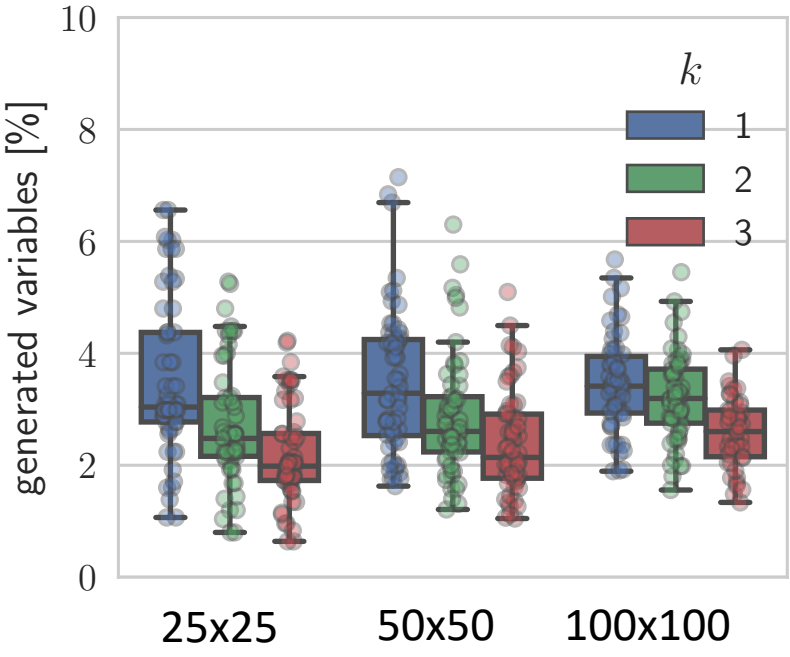
Open question: Hardness of deciding whether B admits a k-Dollo completion A

Results for *k*-DP

- Naive ILP does not scale and has O(*mnk*) variables and O($m^3n^2k^4$) constraints
- Column and cutting plane generation
  - Introduce variables and constraints only when needed
- Simulations with 60 instances for each *m*, *n* and *k*





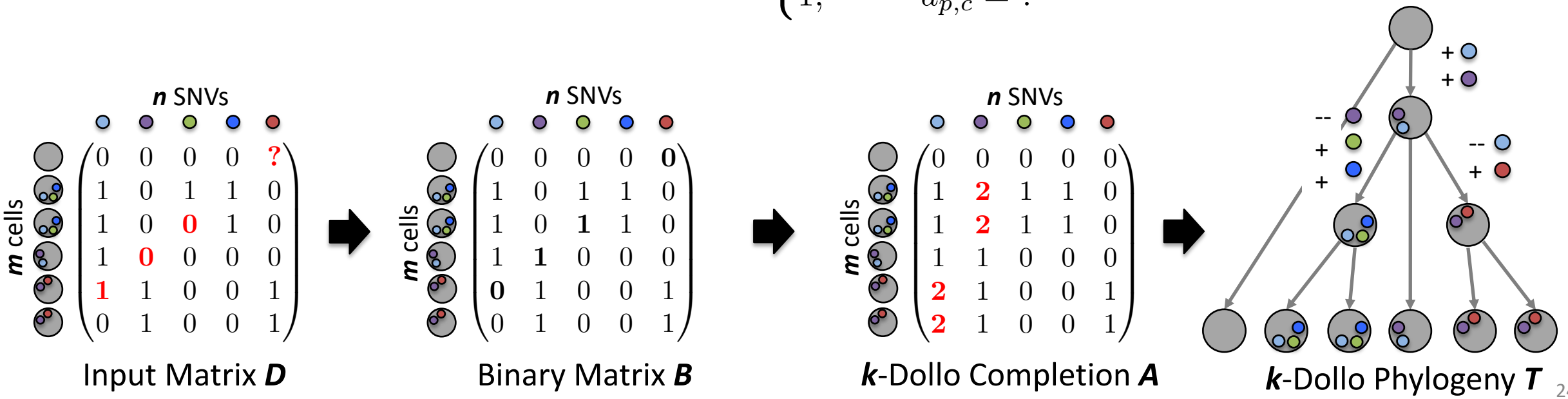


Outline

- Perfect data (error free)
  - Problem statement
  - Combinatorial characterization of solutions
  - Exact algorithm
  - Results
- Real data (with errors)
  - Problem statement
  - Heuristic algorithm
  - Results
- Conclusions

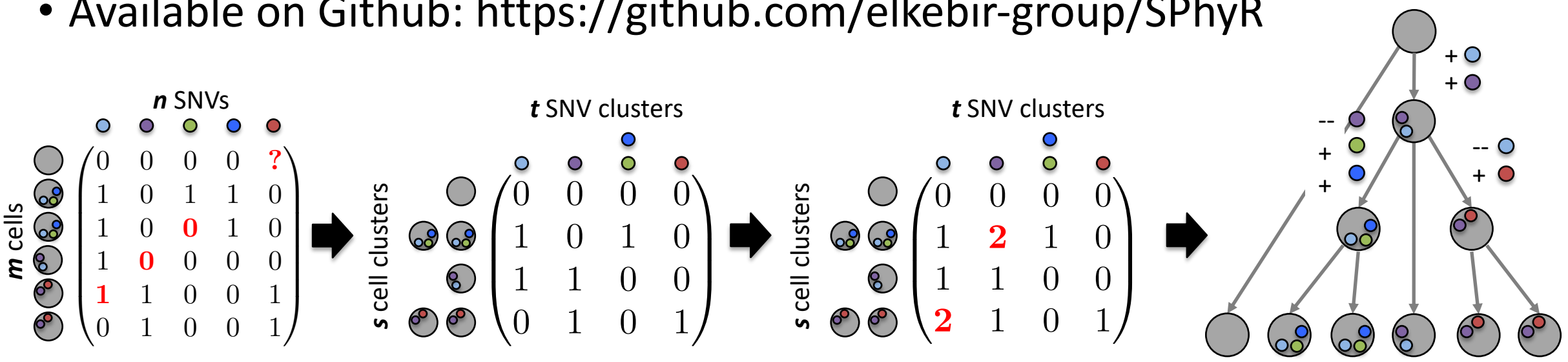
k-Dollo Phylogeny Flip and Cluster (k-DPFC) problem. Given matrix $D \in \{0,1,?\}^{m \times n}$, error rates $\alpha, \beta \in [0,1]$, integers $k, s, t \in \mathbb{N}$, find matrix $B \in \{0,1\}^{m \times n}$ and tree T such that: (1) B has at most s unique rows and at most t unique columns; (2) $\Pr(D \mid B, \alpha, \beta)$ is maximum; and (3) T is a k-Dollo phylogeny for B.

$$\Pr(D \mid B, \alpha, \beta) = \prod_{p=1}^{m} \prod_{c=1}^{n} \begin{cases} \alpha, & d_{p,c} = 1 \text{ and } b_{p,c} = 0 \\ 1-\alpha, & d_{p,c} = 1 \text{ and } b_{p,c} = 1, \\ \beta, & d_{p,c} = 0 \text{ and } b_{p,c} = 1, \\ 1-\beta, & d_{p,c} = 0 \text{ and } b_{p,c} = 0, \\ 1, & d_{p,c} = ? \end{cases}$$

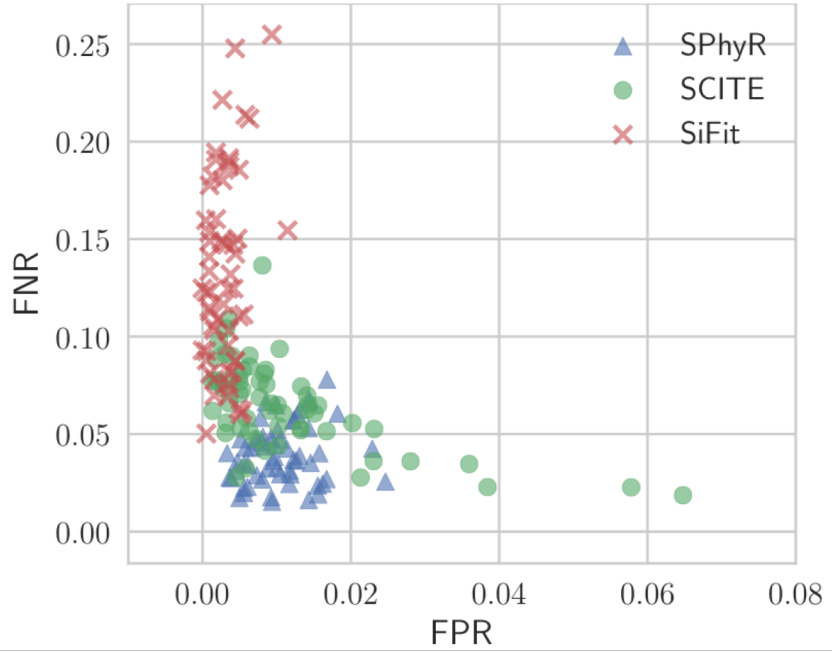


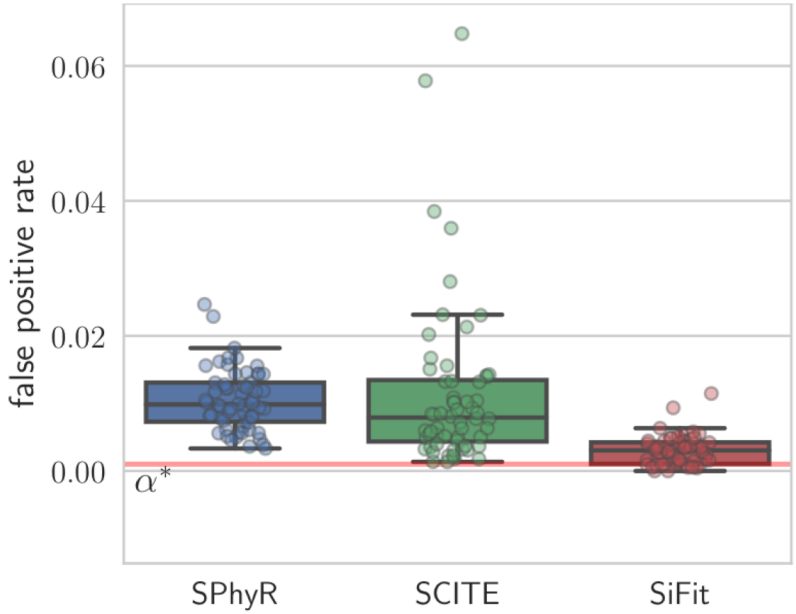
SPhyR: Single-cell Phylogeny Reconstruction

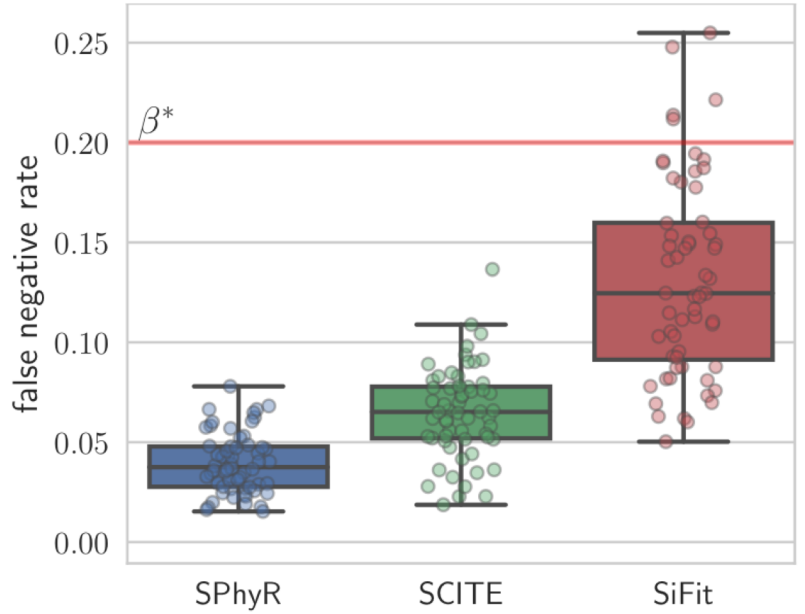
- Coordinate ascent:
  1. k-Means with random seed to obtain cell clustering $\boldsymbol{\pi}$ and SNV clustering $\boldsymbol{\psi}$
  2. ILP to obtain maximum likelihood $\boldsymbol{k}$-Dollo completion $\boldsymbol{A}$ given $\boldsymbol{D}$, $\boldsymbol{\pi}$ and $\boldsymbol{\psi}$
  3. Identify maximum likelihood $\boldsymbol{\pi}$ given $\boldsymbol{A}$ and $\boldsymbol{\psi}$
  4. Identify maximum likelihood $\boldsymbol{\psi}$ given $\boldsymbol{A}$ and $\boldsymbol{\pi}$
  5. Repeat until convergence
- Available on Github: https://github.com/elkebir-group/SPhyR



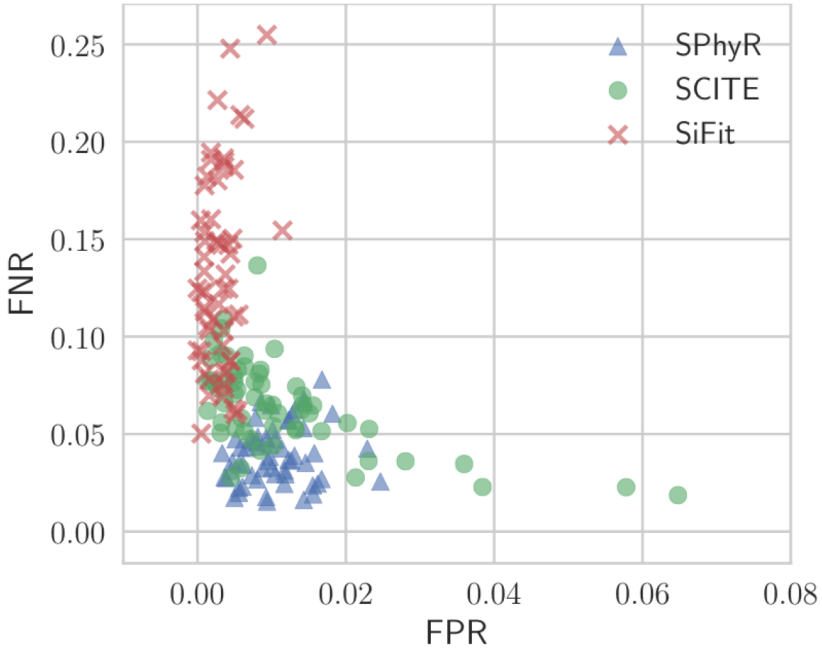
Simulation Results (m = 50, n = 50, k = 1)

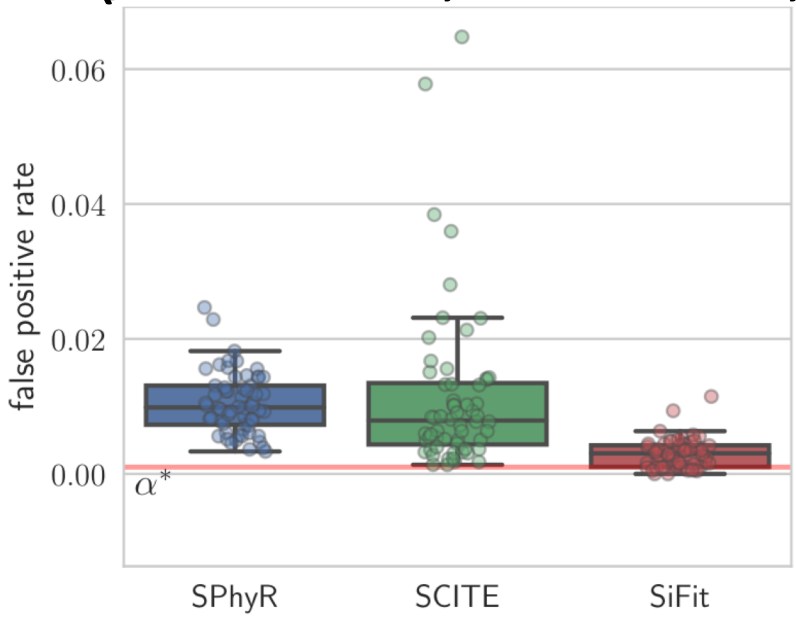


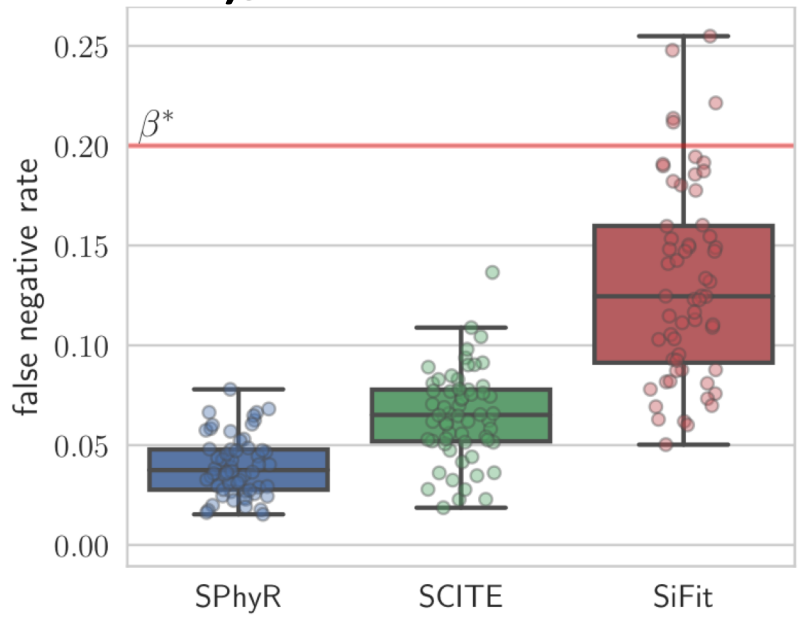


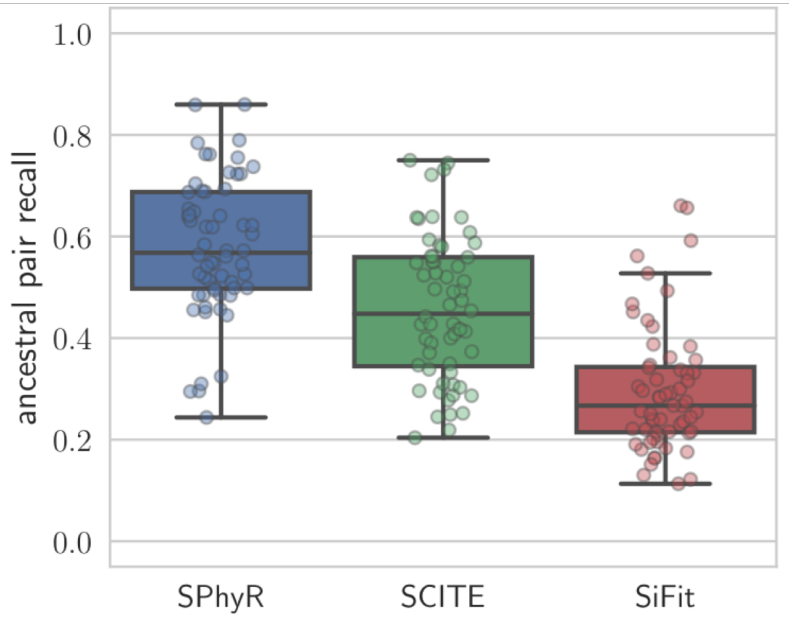


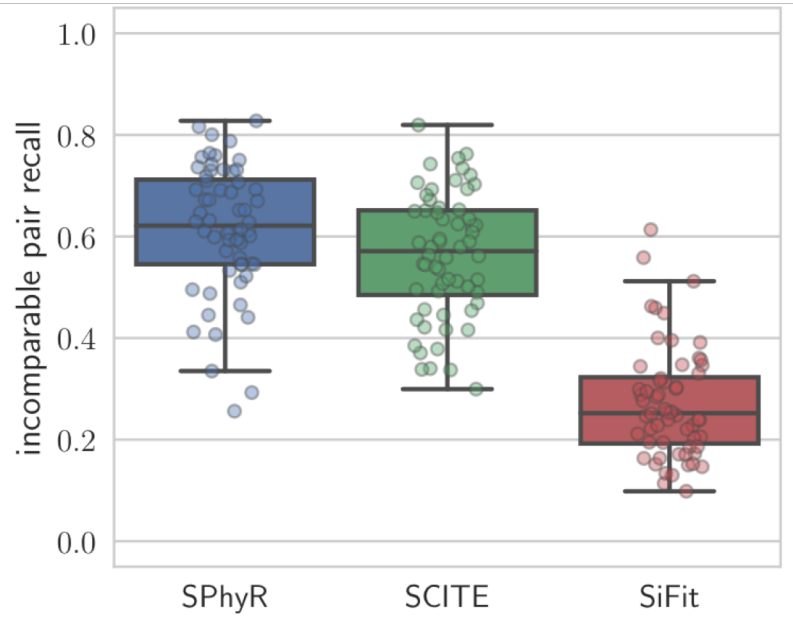
Simulation Results (m = 50, n = 50, k = 1)

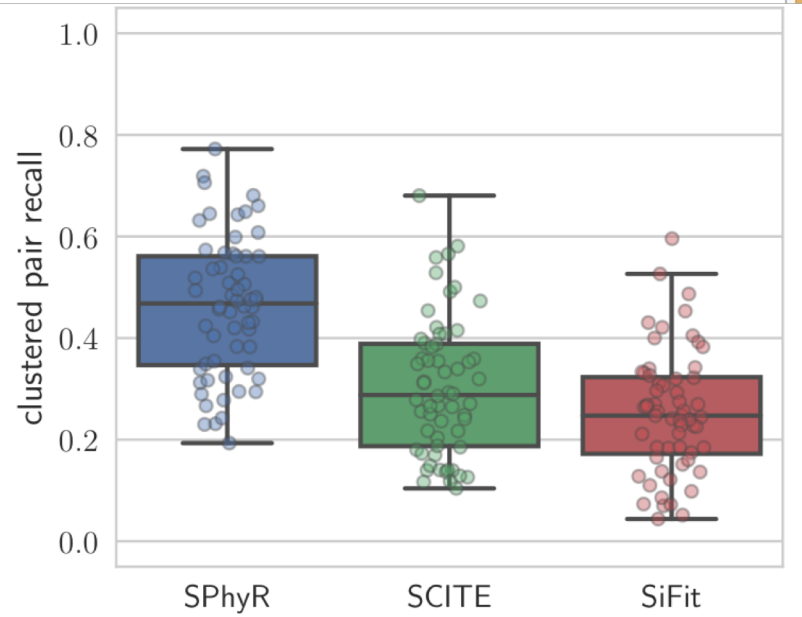












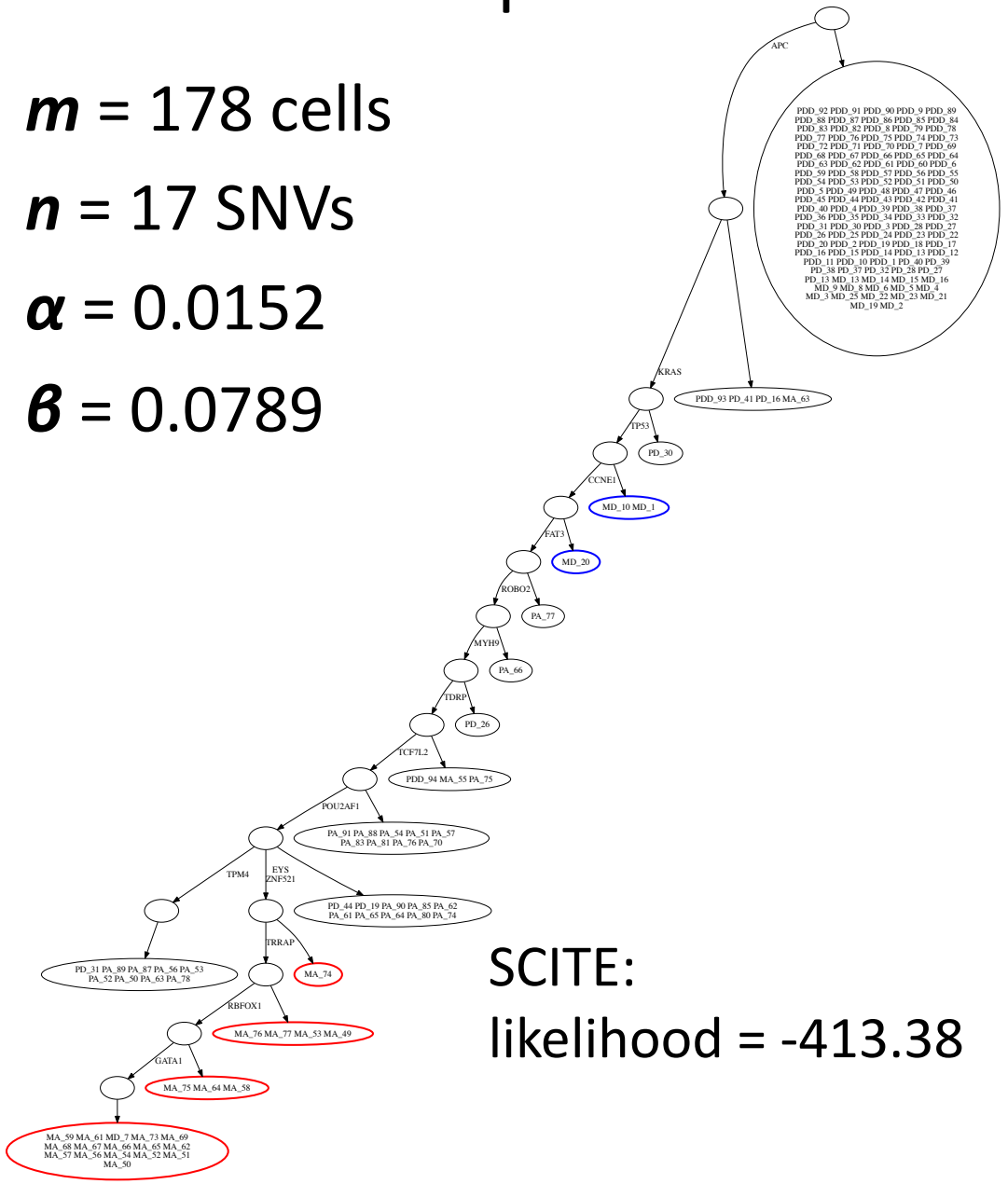
Colorectal patient CRC1 [Leung et al., 2017]

m = 178 cells

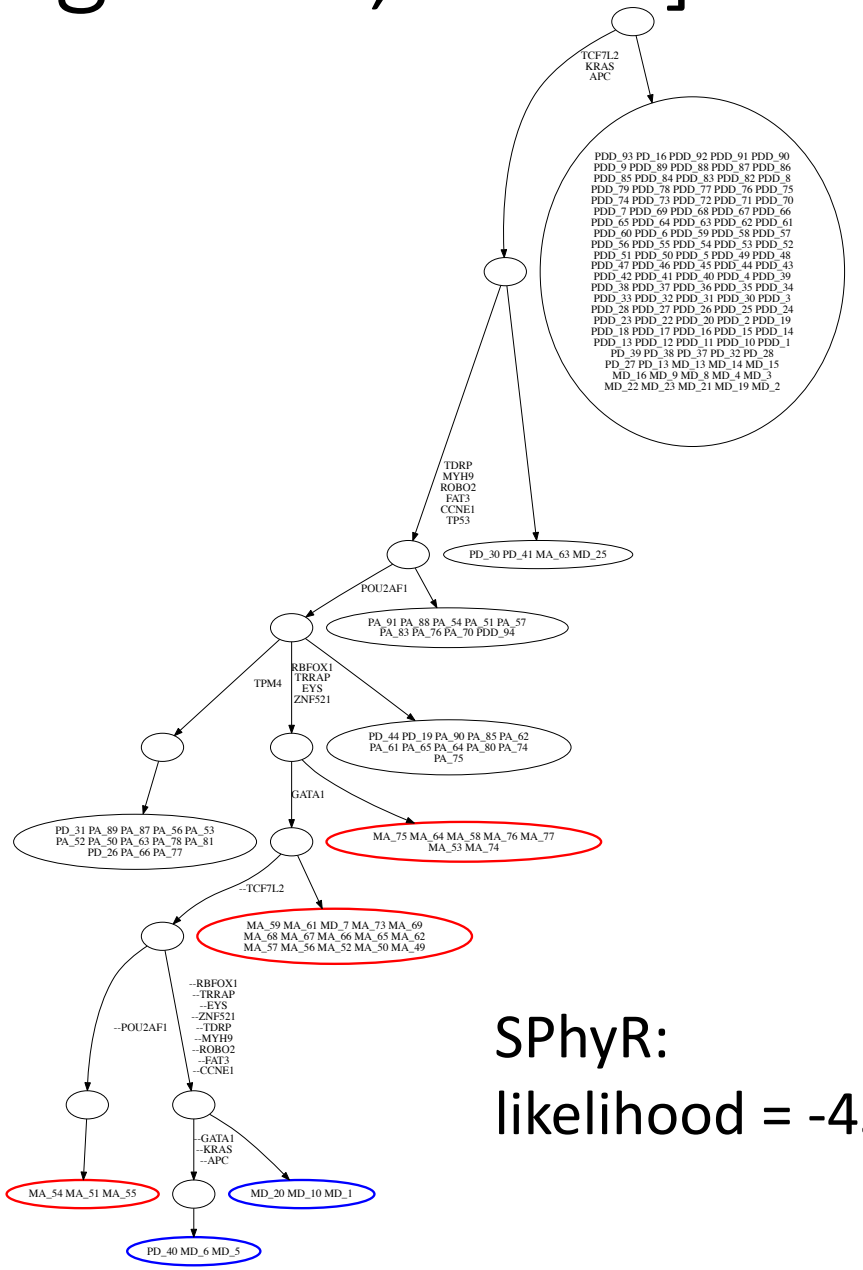
n = 17 SNVs

α = 0.0152

β = 0.0789



SCITE:
likelihood = -413.38



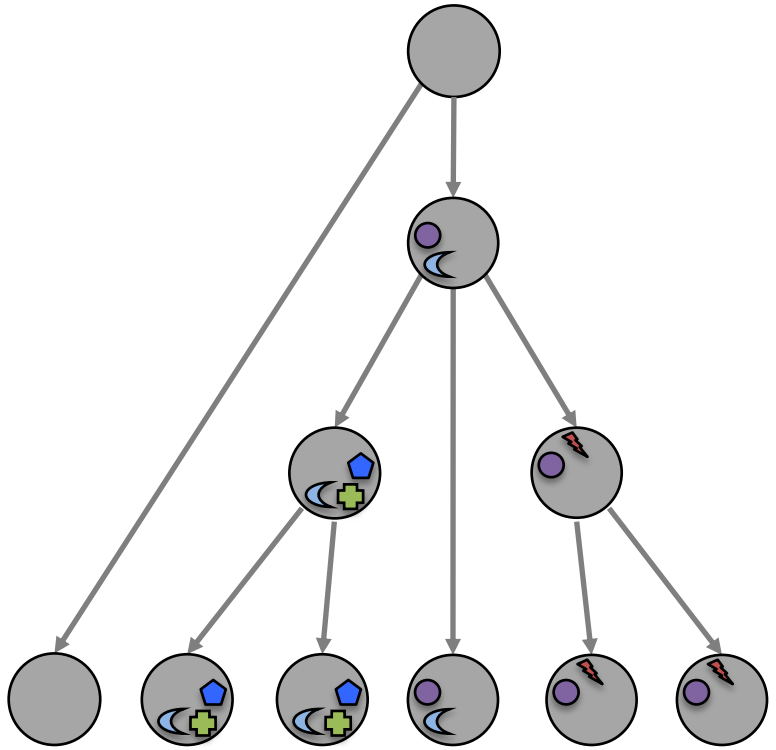
SPhyR:
likelihood = -450.70

Conclusions

- ***k***-Dollo parsimony model strikes a balance between realistic and yet sufficiently constrained
- Solutions are integer matrix completions
- SPhyR outperformed existing methods

Future work:

- Include $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ into optimization
- Model selection for ***s***, ***t*** and ***k***
- Hardness is open



Acknowledgments

- Experiments were run on NCSA Blue Waters