

# CS 466

# Introduction to Bioinformatics

## Lecture 16

Mohammed El-Kebir

October 21, 2020



# Outline

- Character-based phylogeny (small)
- Application of small phylogeny maximum parsimony problem to cancer

## **Reading:**

- Chapters 10.2, 10.5-10.8, 10.9 in Jones and Pevzner

# Character-Based Tree Reconstruction

- Characters may be morphological features
  - Shape of beak {generalist, insect catching, ...}
  - Number of legs {2,3,4, ..}
  - Hibernation {yes, no}
- Character may be nucleotides/amino acids
  - {A, T, C, G}
  - 20 amino acids
- Values of a character are called states
  - We assume discrete states



Generalist



Insect catching



Grain eating



Coniferous-seed eating



Nectar feeding



Fruit eating



Chiseling



Dip netting



Surface skimming



Scything



Probing



Filter feeding



Aerial fishing



Pursuit fishing



Scavenging



Raptorial

# Character-Based Phylogeny Reconstruction

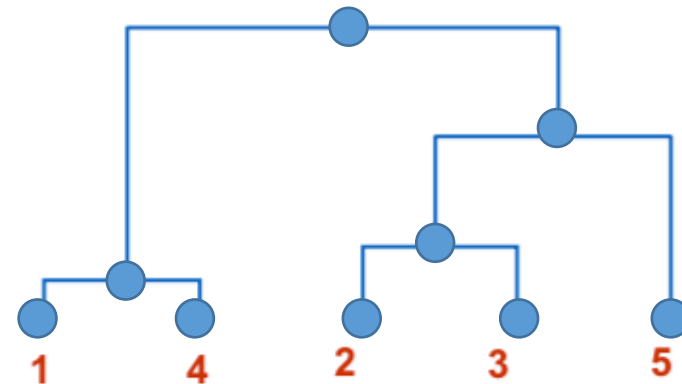
Input  
characters



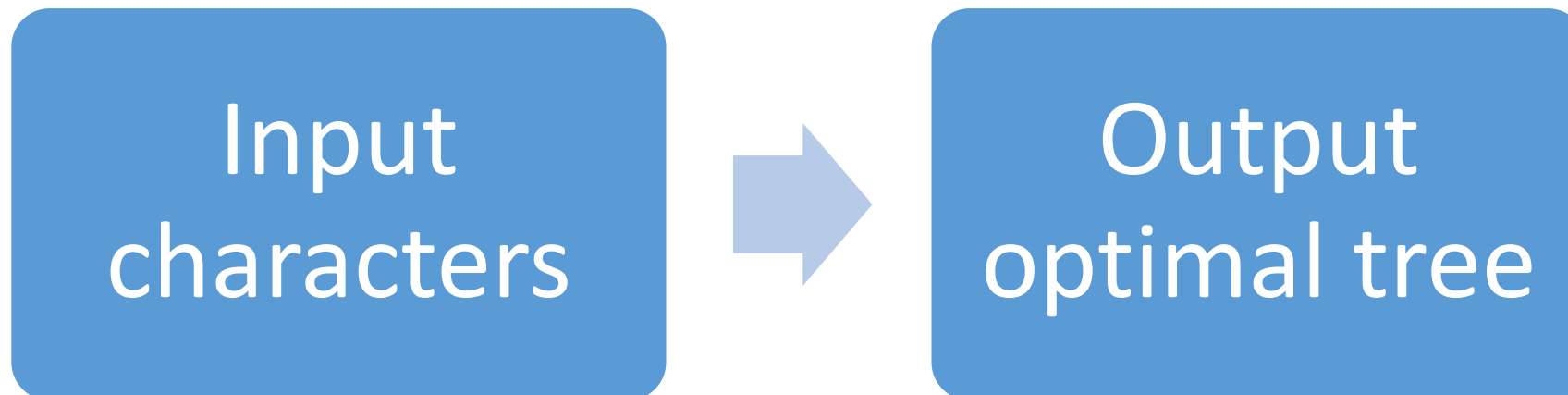
Output  
optimal tree

**Question:** What is optimal?

**Want:** Optimization criterion



# Character-Based Phylogeny Reconstruction

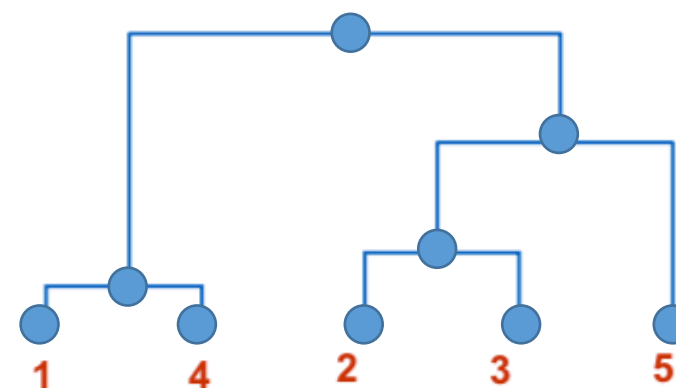


**Question:** What is optimal?

**Want:** Optimization criterion

**Question:** How to optimize this criterion?

**Want:** Algorithm

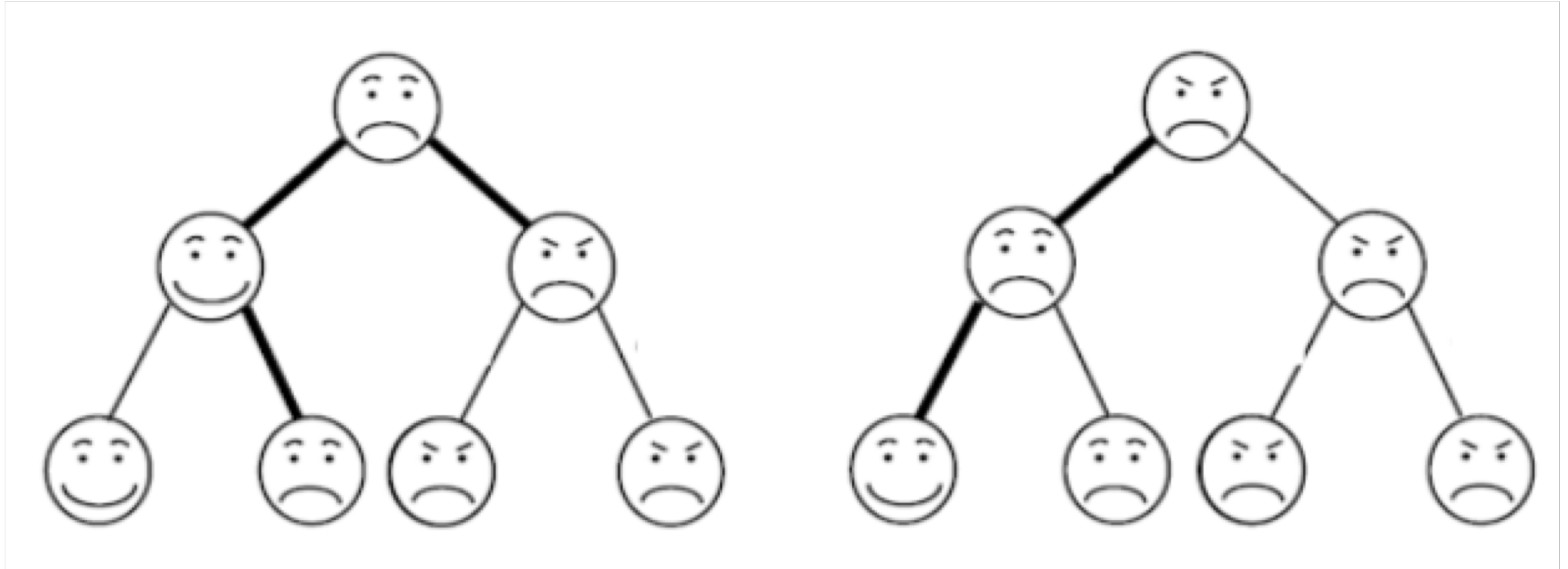


# Character-Based Phylogeny Reconstruction: Input

Characters / states	State 1	State 2
Mouth	Smile	Frown
Eyebrows	Normal	Pointed

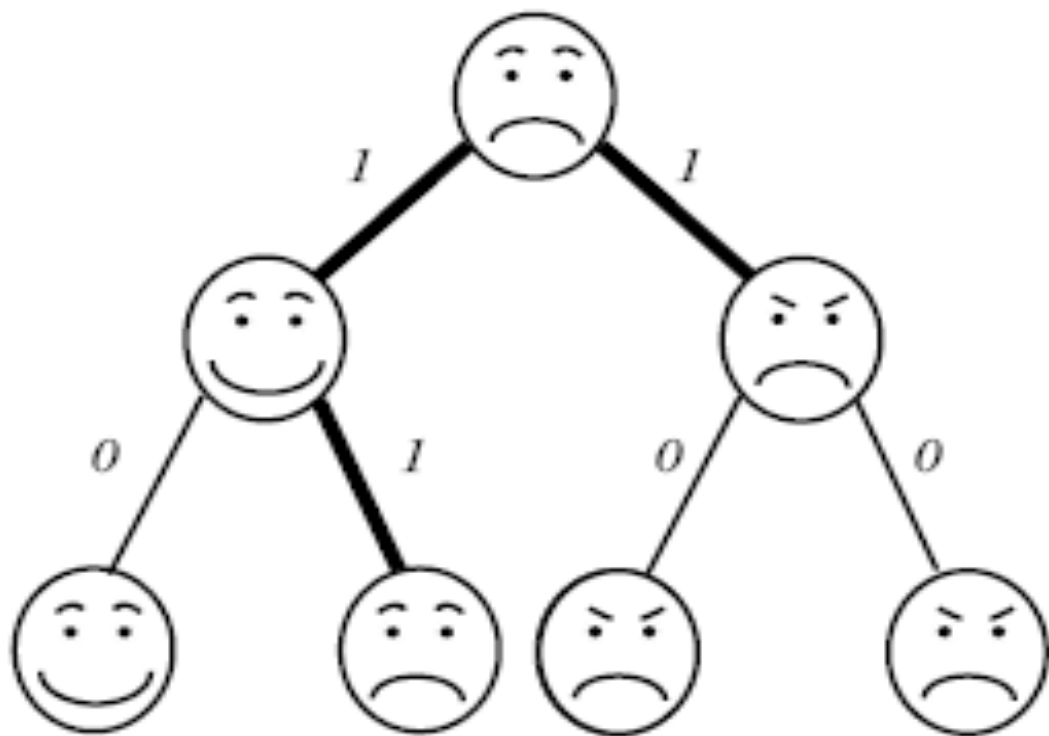


# Character-Based Phylogeny Reconstruction: Criterion

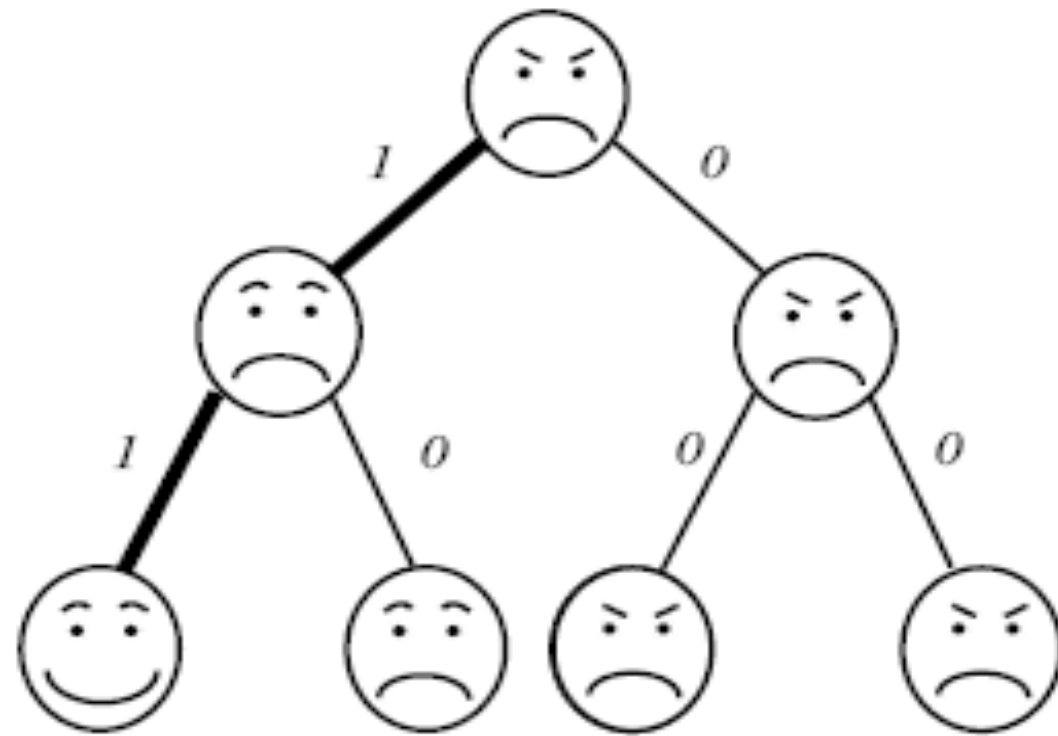


**Question:** Which tree is better?

# Character-Based Phylogeny Reconstruction: Criterion



(a) *Parsimony Score=3*



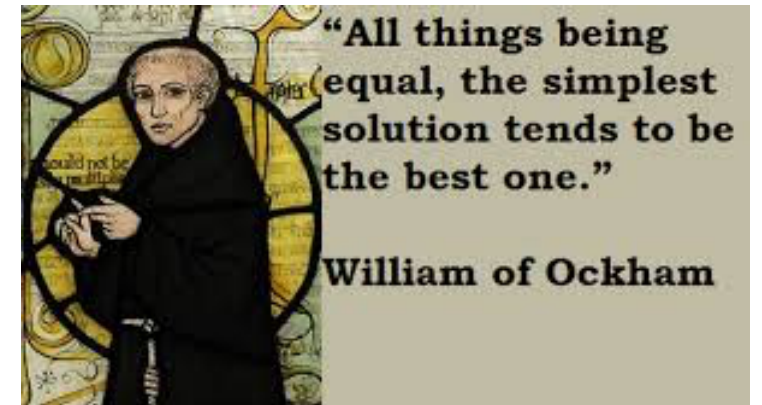
(b) *Parsimony Score=2*

**Parsimony:** minimize number of changes on edges of tree



# Why Parsimony?

- Ockham's razor: "simplest" explanation for data
- Assumes that observed character differences resulted from the fewest possible mutations
- Seeks tree with the lowest **parsimony score**, i.e. the sum of all (costs of) mutations in the tree.



# Again, a Small and a Large Problem

## **Small Maximum Parsimony Phylogeny Problem:**

Given  $m \times n$  matrix  $A = [a_{i,j}]$  and tree  $T$  with  $m$  leaves, find assignment of character states to each internal vertex of  $T$  with minimum parsimony score.

## **Large Maximum Parsimony Phylogeny Problem:**

Given  $m \times n$  matrix  $A = [a_{i,j}]$ , find a tree  $T$  with  $m$  leaves labeled according to  $A$  and an assignment of character states to each internal vertex of  $T$  with minimum parsimony score.

**Question:** Are both problems easy (i.e. in P)?

# Again, a Small and a Large Problem

## **Small Maximum Parsimony Phylogeny Problem:**

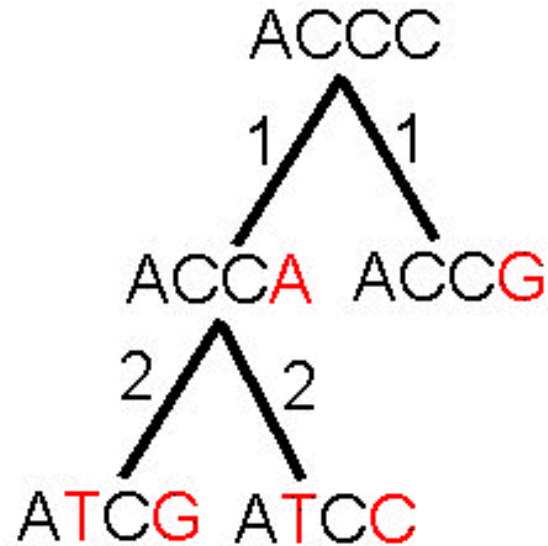
Given  $m \times n$  matrix  $A = [a_{i,j}]$  and tree  $T$  with  $m$  leaves, find assignment of character states to each internal vertex of  $T$  with minimum parsimony score.

## **Large Maximum Parsimony Phylogeny Problem:**

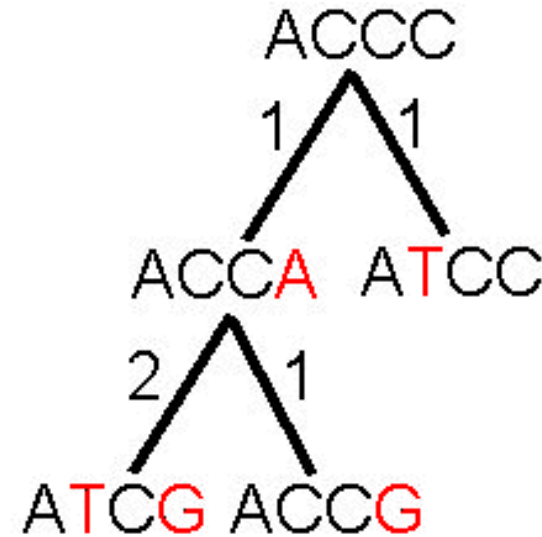
Given  $m \times n$  matrix  $A = [a_{i,j}]$ , find a tree  $T$  with  $m$  leaves labeled according to  $A$  and an assignment of character states to each internal vertex of  $T$  with minimum parsimony score.

**Question:** Are both problems easy (i.e. in P)?

# Small Maximum Parsimony Phylogeny Problem



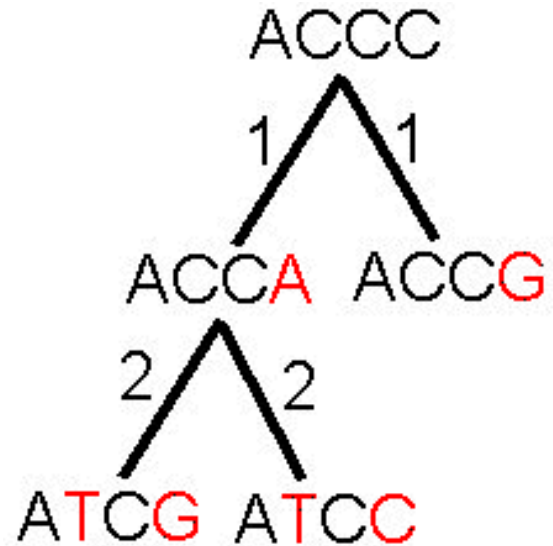
**Less**  
**Parsimonious**  
**Score: 6**



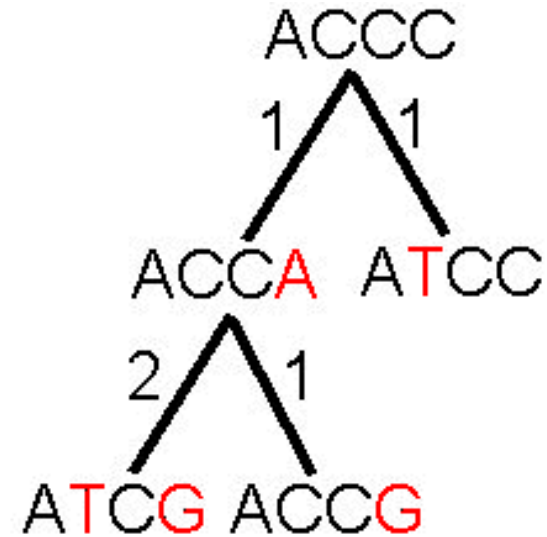
**More**  
**Parsimonious**  
**Score: 5**

**Question:** There are  $n = 4$  characters in the  $m = 2$  taxa (leaves).  
Can we solve each character separately?

# Small Maximum Parsimony Phylogeny Problem



**Less**  
**Parsimonious**  
**Score: 6**



**More**  
**Parsimonious**  
**Score: 5**

**Key observations:** (1) Characters can be solved independently.  
(2) Optimal substructure in subtrees.

# Recurrence

# Recurrence for Small Maximum Parsimony Problem

## Small Maximum Parsimony Phylogeny Problem:

Given rooted tree  $T$  whose leaves are labeled by  $\sigma : L(T) \rightarrow \Sigma$ , find assignment of states to each internal vertex of  $T$  with minimum parsimony score.

Let  $\mu(v, s)$  be the minimum number of mutations in the subtree rooted at  $v$  when assigning state  $s$  to  $v$ .

$$c(s, t) = \begin{cases} 0, & \text{if } s = t \\ 1, & \text{if } s \neq t, \end{cases}$$

Let  $\delta(v)$  be the set of children of  $v$ .

$$\mu(v, s) = \min \begin{cases} \infty, & \text{if } v \in L(T) \text{ and } s \neq \sigma(v), \\ 0, & \text{if } v \in L(T) \text{ and } s = \sigma(v), \\ \sum_{w \in \delta(v)} \min_{t \in \Sigma} \{c(s, t) + \mu(w, t)\}, & \text{if } v \notin L(T). \end{cases}$$

# Example

$$c(s, t) = \begin{cases} 0, & \text{if } s = t \\ 1, & \text{if } s \neq t, \end{cases}$$

$$\mu(v, s) = \min \begin{cases} \infty, & \text{if } v \in L(T) \text{ and } s \neq \sigma(v), \\ 0, & \text{if } v \in L(T) \text{ and } s = \sigma(v), \\ \sum_{w \in \delta(v)} \min_{t \in \Sigma} \{c(s, t) + \mu(w, t)\}, & \text{if } v \notin L(T). \end{cases}$$

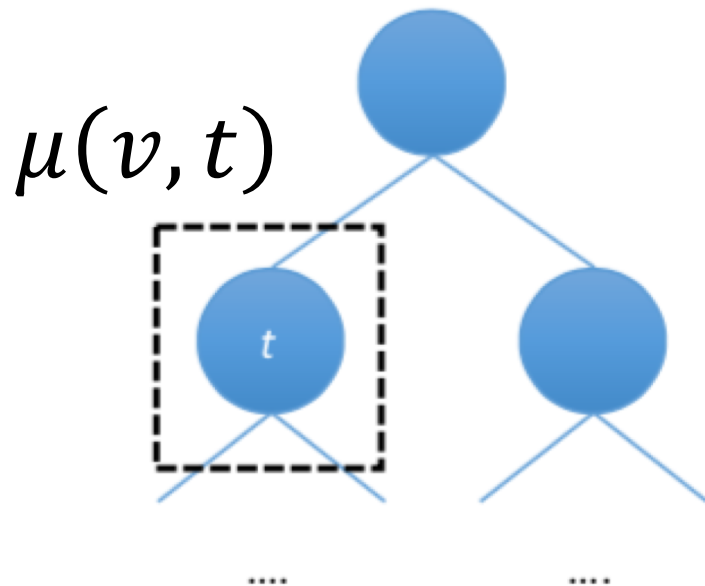


# Pseudocode for Filling and Traceback

# Sankoff Algorithm (Sankoff 1975)

## Small Maximum Parsimony Phylogeny Problem:

Given  $m \times n$  matrix  $A = [a_{i,j}]$  and tree  $T$  with  $m$  leaves, find assignment of character states to each internal vertex of  $T$  with minimum parsimony score.



One example:



# Outline

- Recap character-based phylogeny
- Application of small phylogeny maximum parsimony problem to cancer

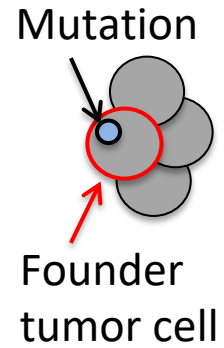
## **Reading:**

- Chapters 10.2, 10.5-10.8, 10.9 in Jones and Pevzner

# Tumorigenesis: (i) Cell Mutation

## Clonal Theory of Cancer

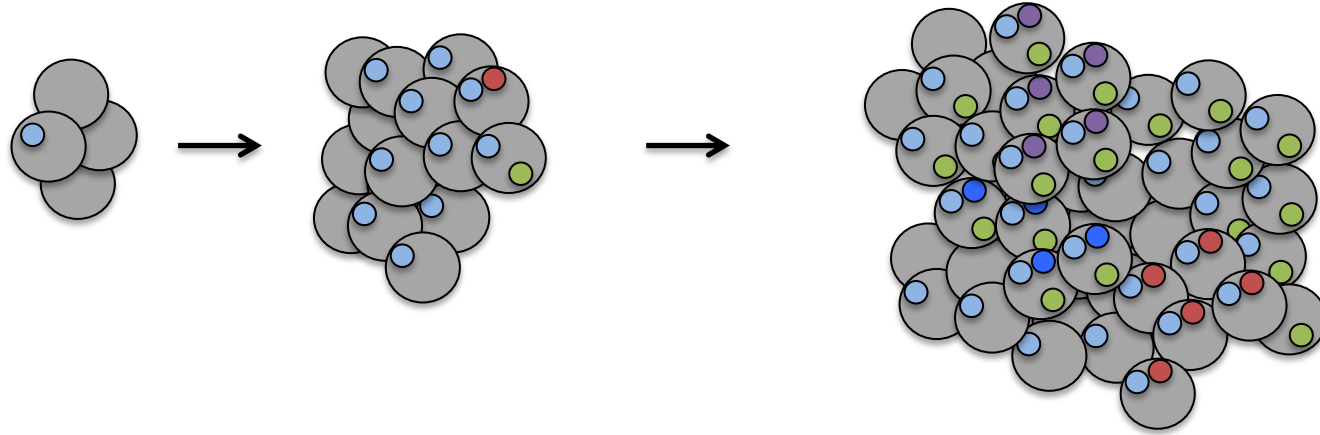
[Nowell, 1976]



# Tumorigenesis: (i) Cell Mutation, (ii) Cell Division

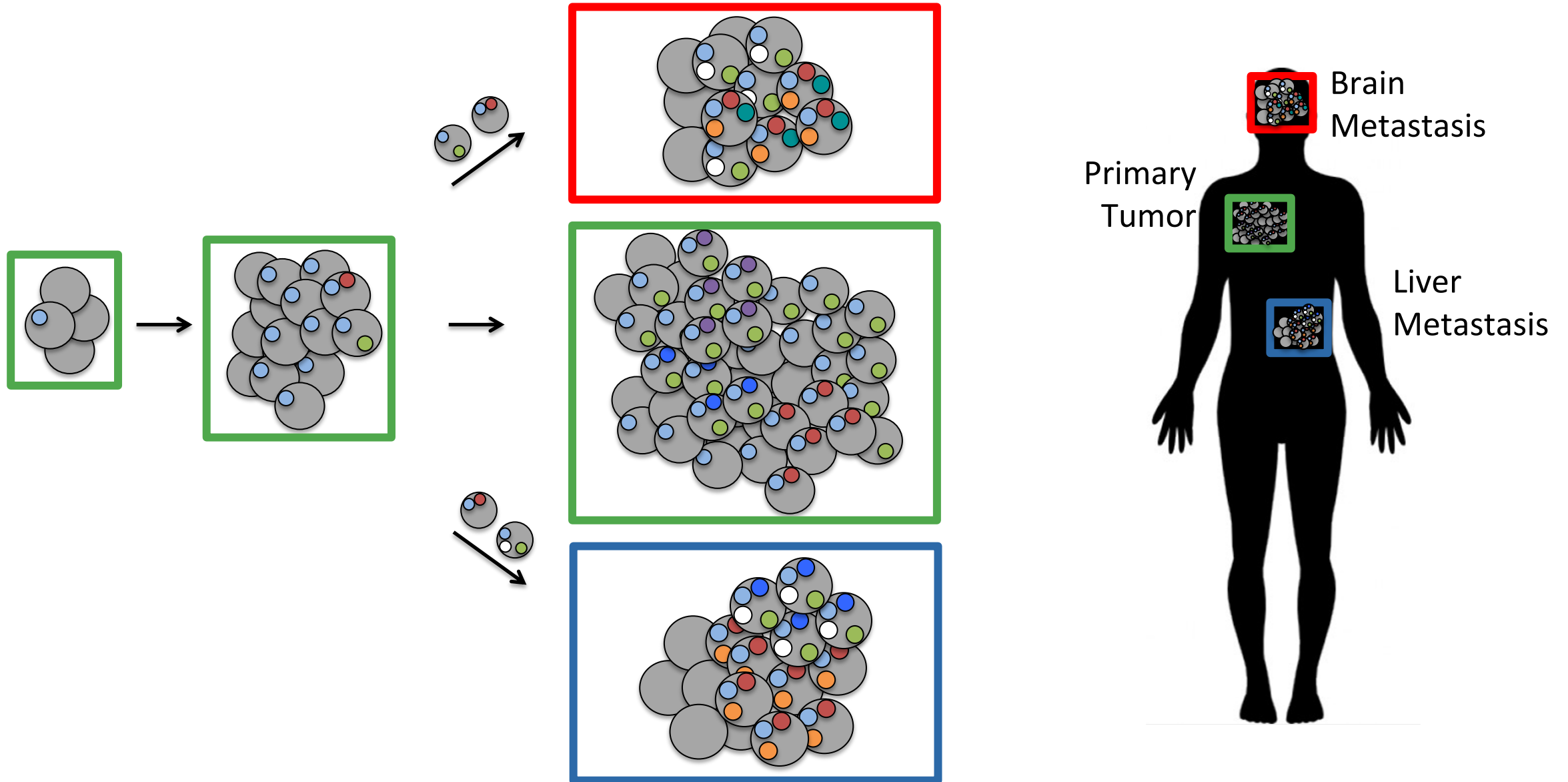
## Clonal Theory of Cancer

[Nowell, 1976]

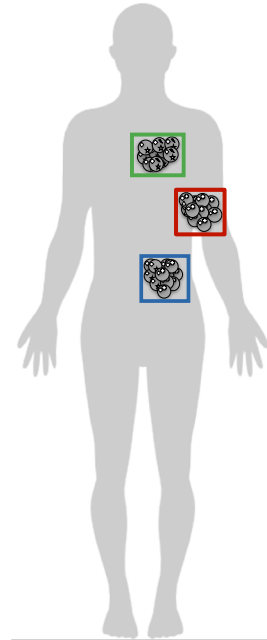
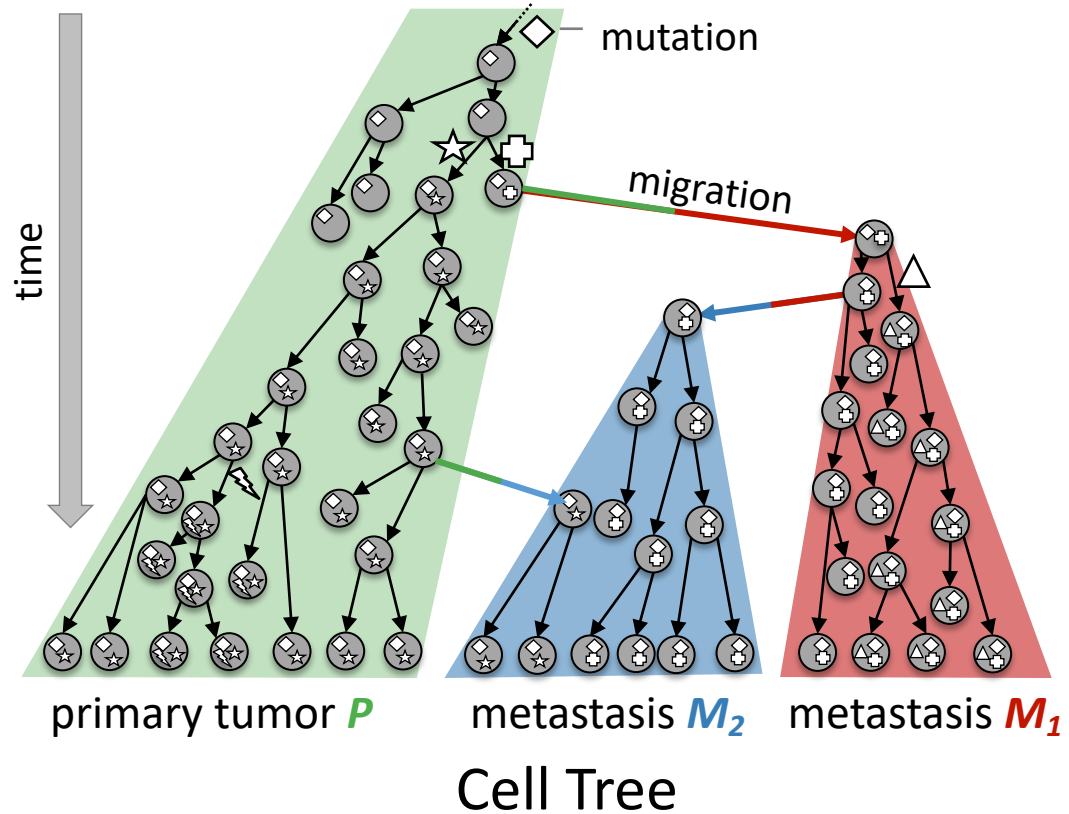


Heterogeneous Tumor

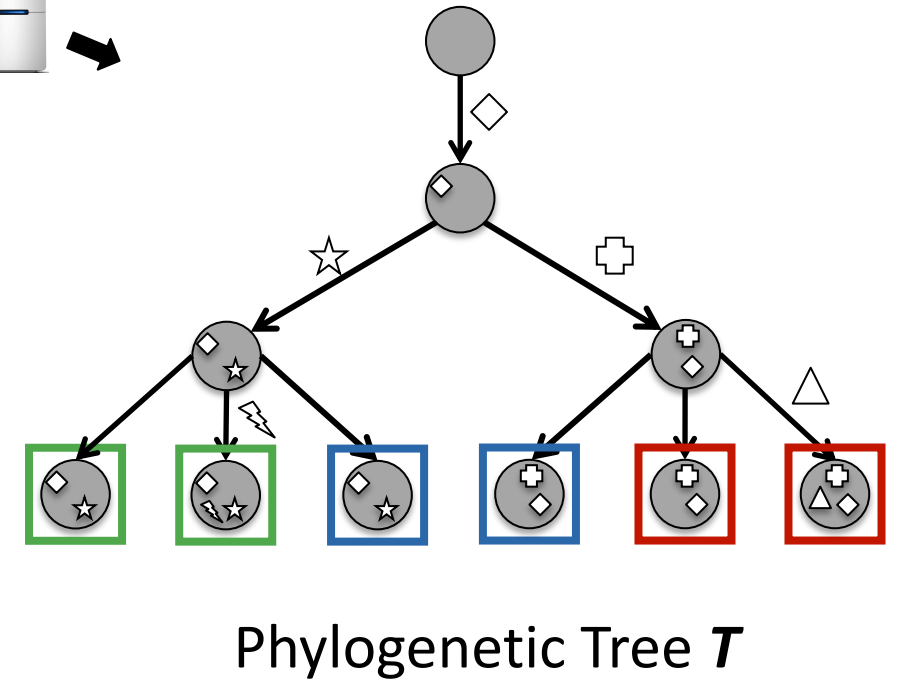
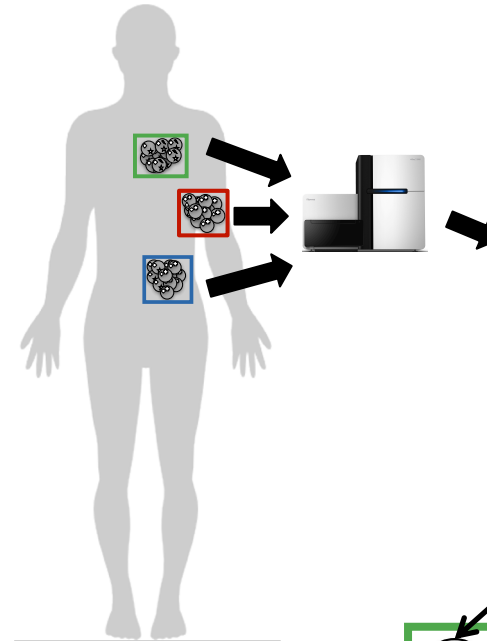
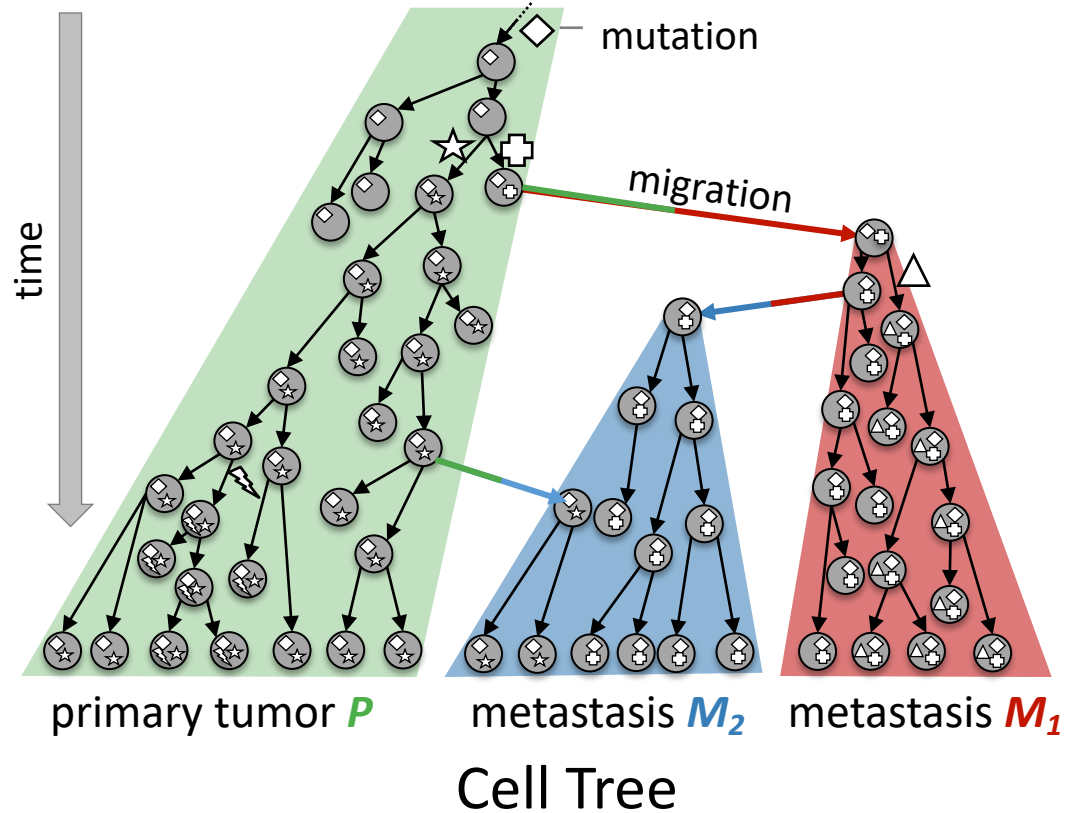
# Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration



# Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration

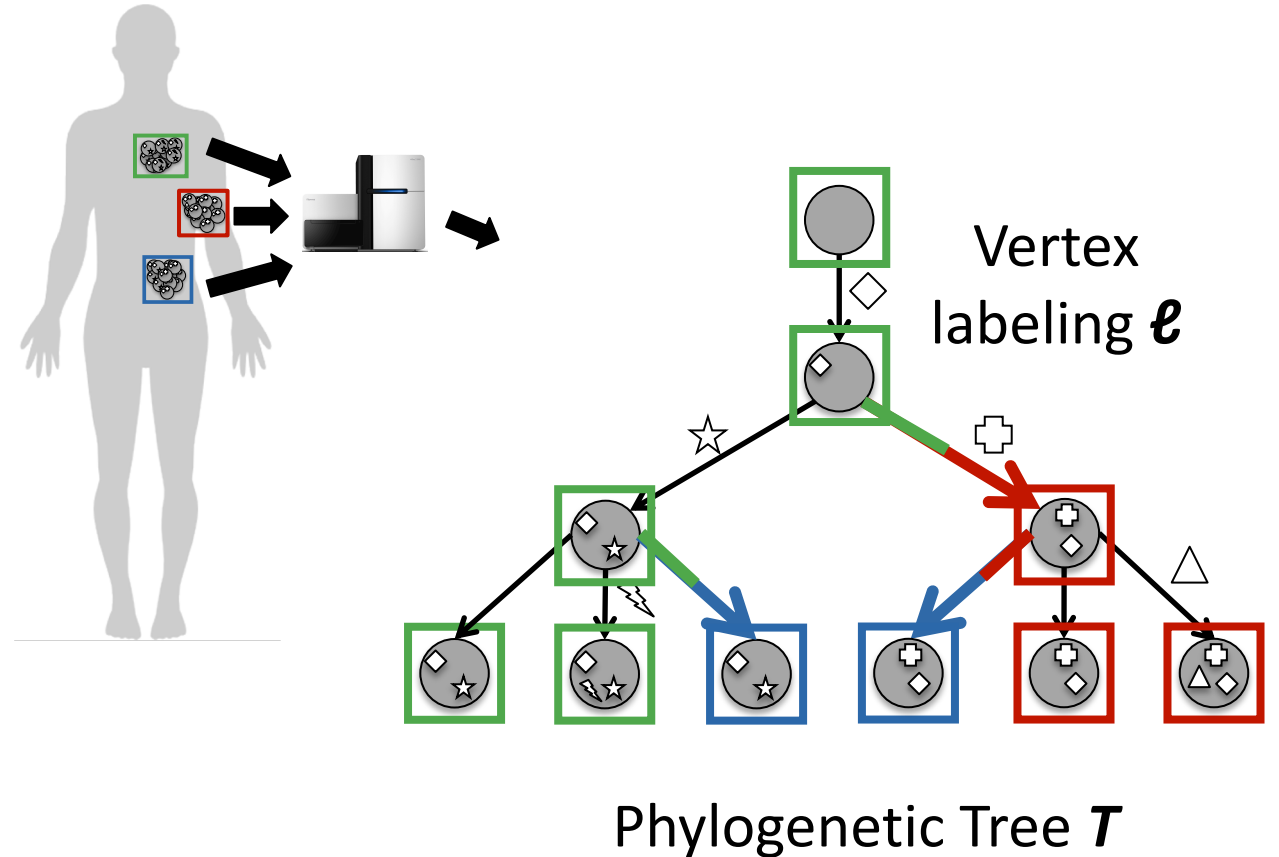
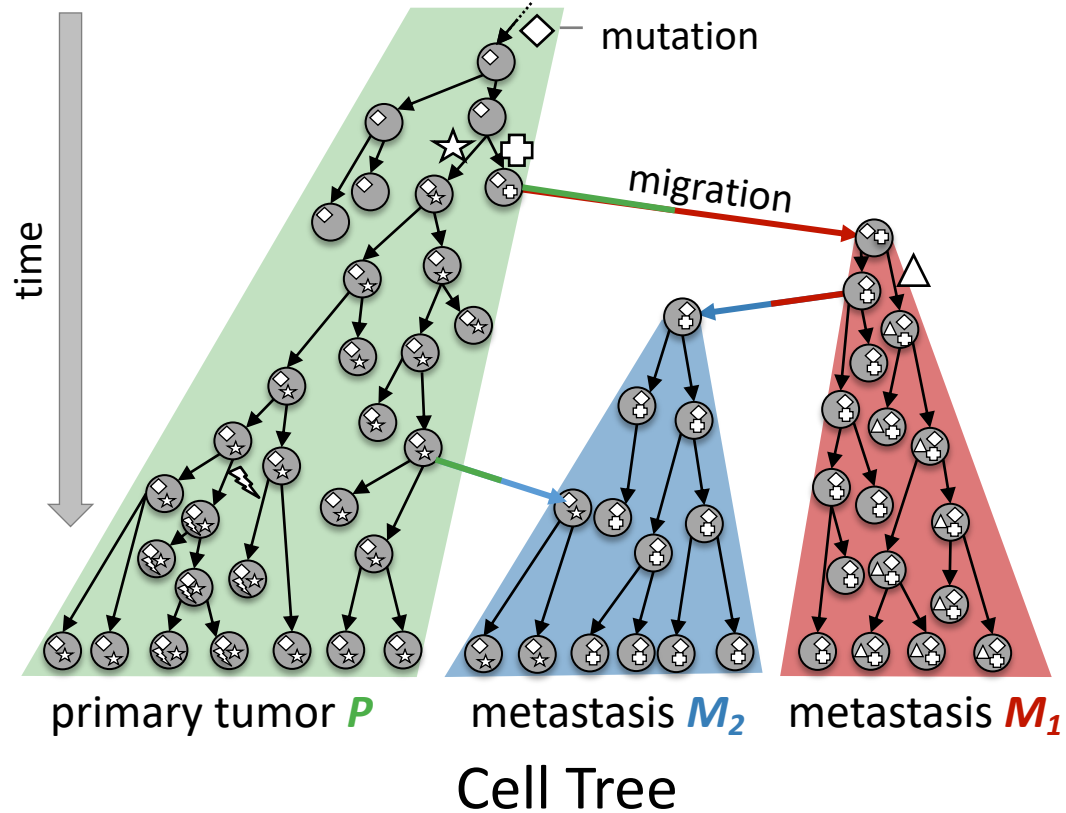


# Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration





# Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration

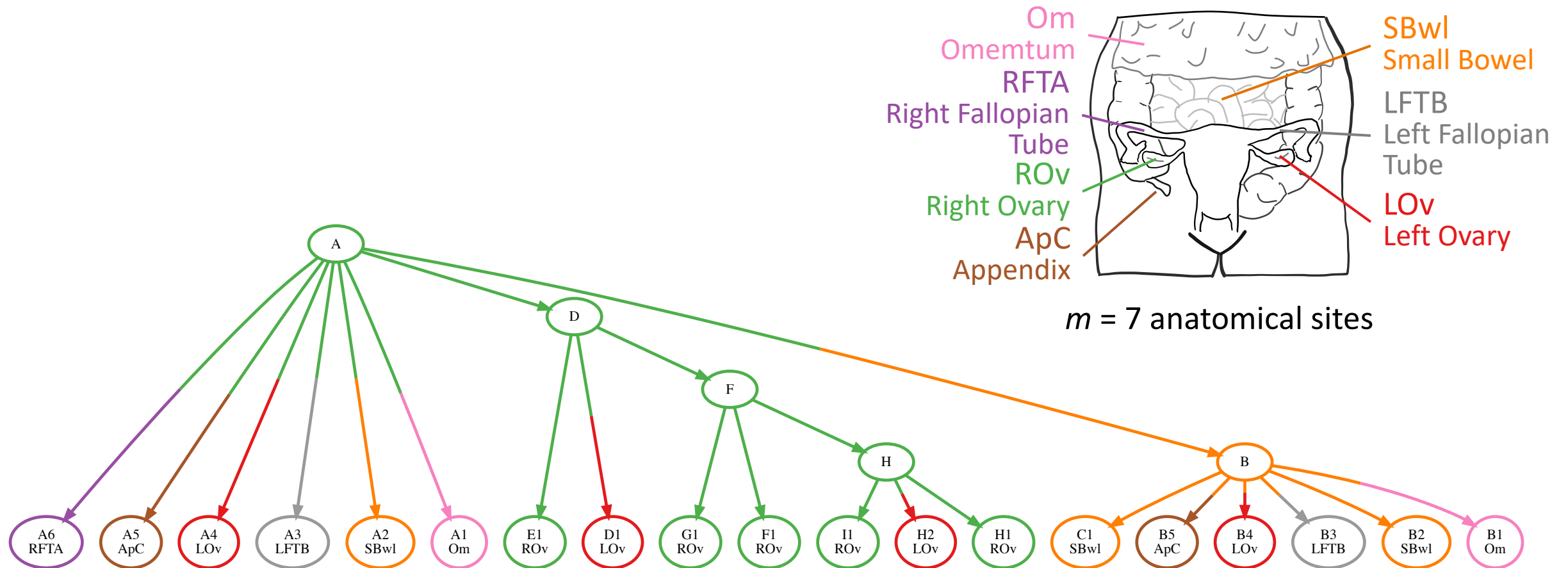


**Goal:** Given phylogenetic tree  $T$ , find *parsimonious* vertex labeling  $\ell$  with fewest migrations

# Minimum Migration Analysis in Ovarian Cancer

McPherson et al. (2016). Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics*.

- Instance of the maximum parsimony small phylogeny problem [Fitch, 1971; Sankoff, 1975]

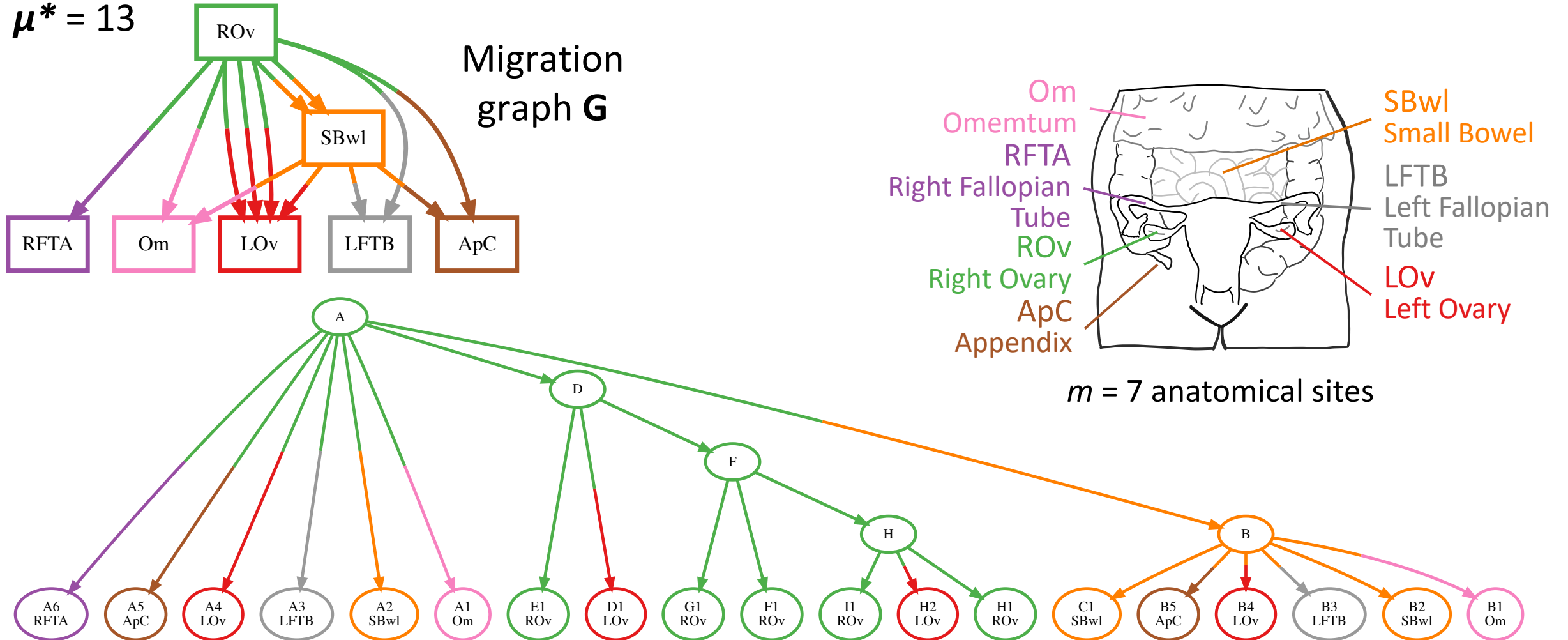


# Minimum Migration Analysis in Ovarian Cancer

McPherson et al. (2016). Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics*.

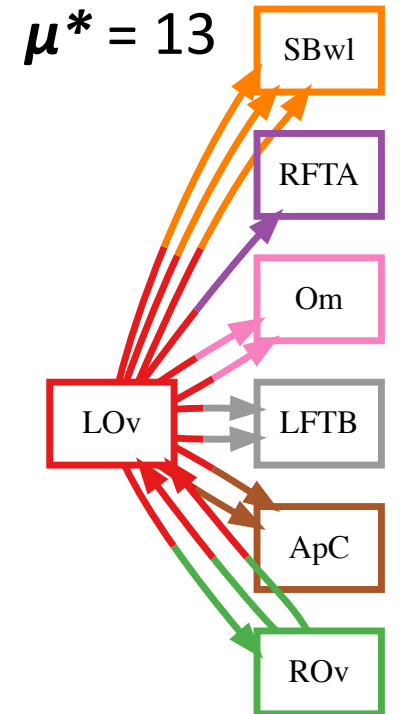
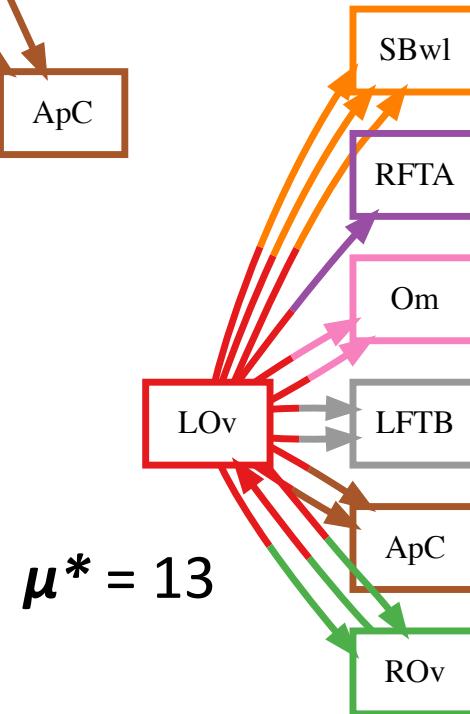
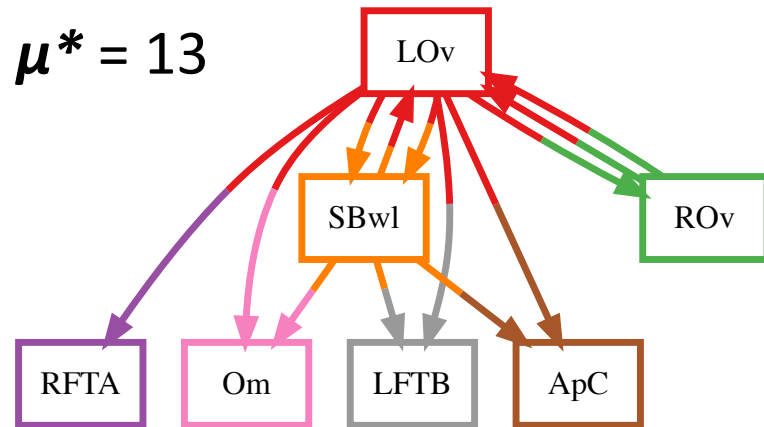
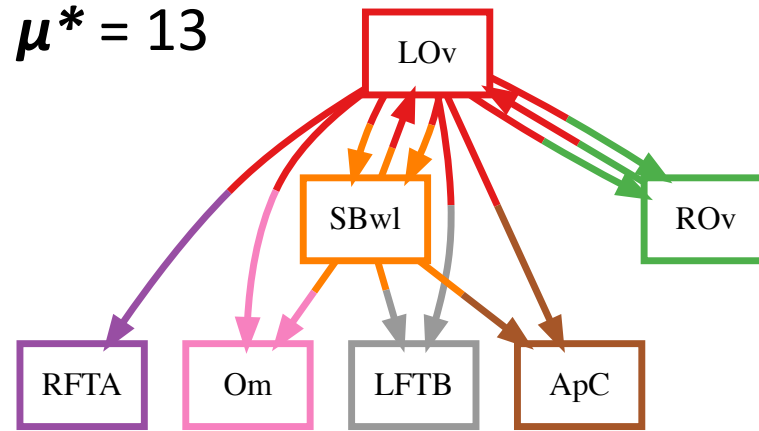
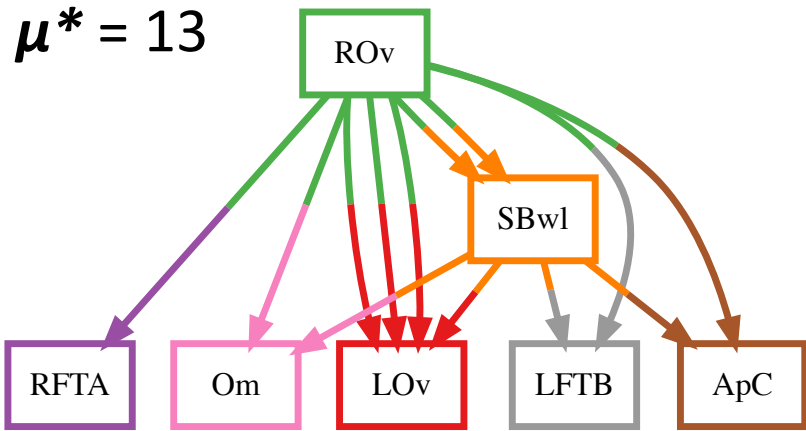
- Instance of the maximum parsimony small phylogeny problem [Fitch, 1971; Sankoff, 1975]

$\mu^* = 13$



# Minimum Migration History is *Not* Unique

- Enumerate all minimum-migration vertex labelings in the backtrace step

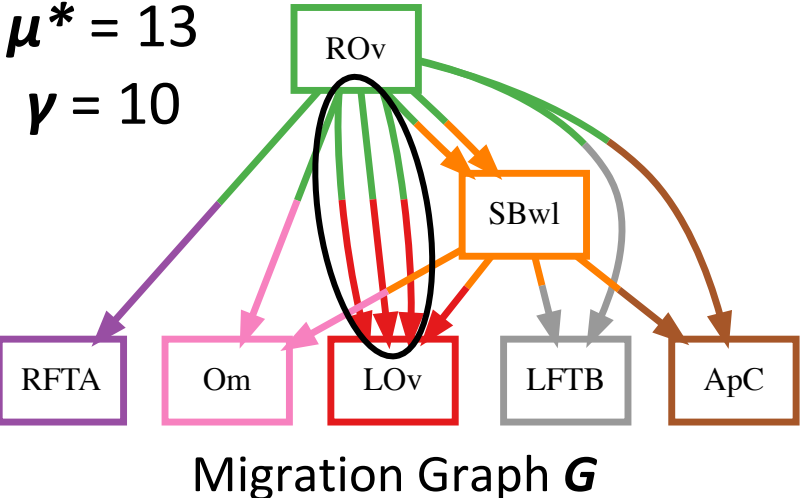


ApC	Appendix
LFTB	Left Fallopian Tube
LOv	Left Ovary
RFTA	Right Fallopian Tube
ROv	Right Ovary
SBwl	Small Bowel
Om	Omentum

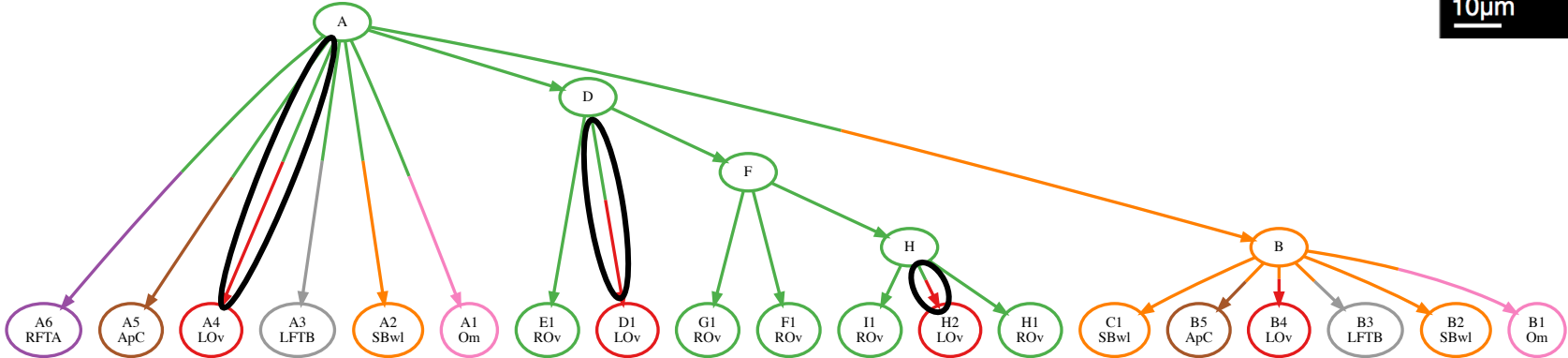
# Comigrations: Simultaneous Migrations of Multiple Clones

- Multiple tumor cells migrate simultaneously through the blood stream [Cheung et al., 2016]
- Second objective: number  $\gamma$  of **comigrations** is the number of multi-edges in migration graph  $G^\dagger$

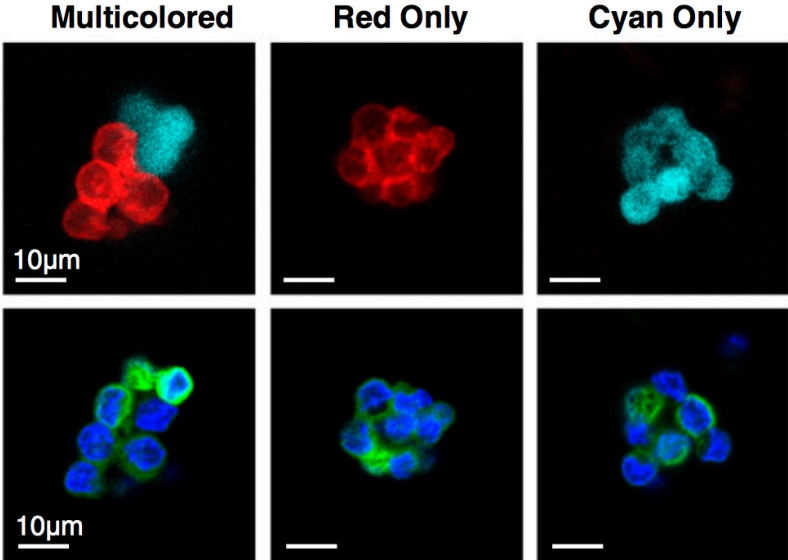
† Not necessarily true in the case of directed cycles



Clone Tree  $T$



F Circulating tumor cell clusters



mTomato  
CFP

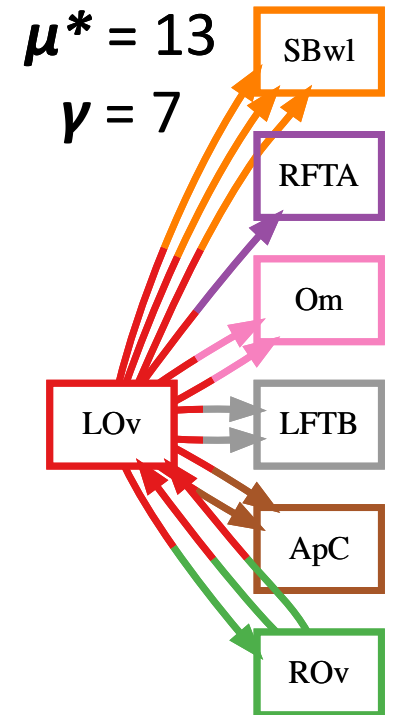
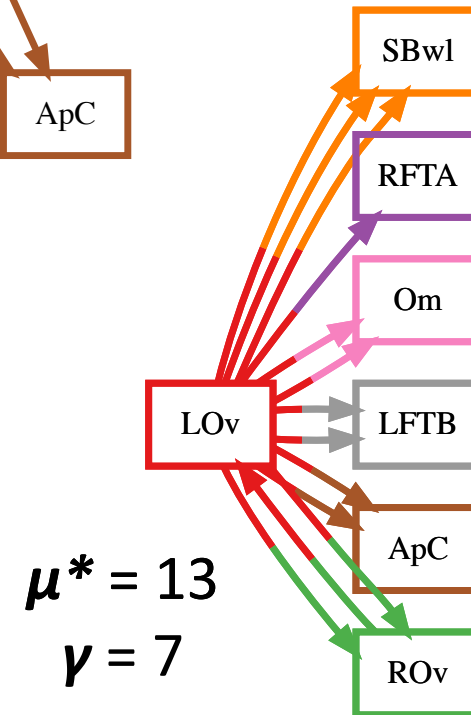
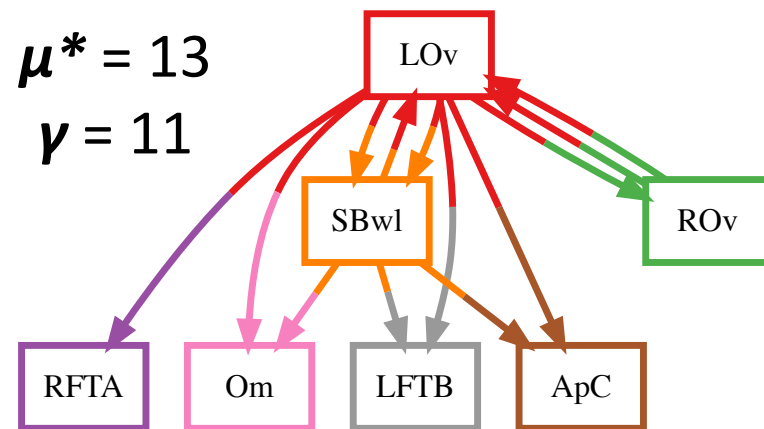
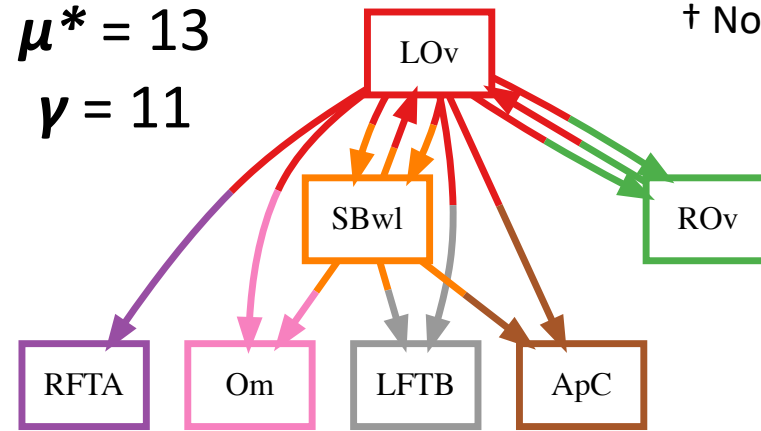
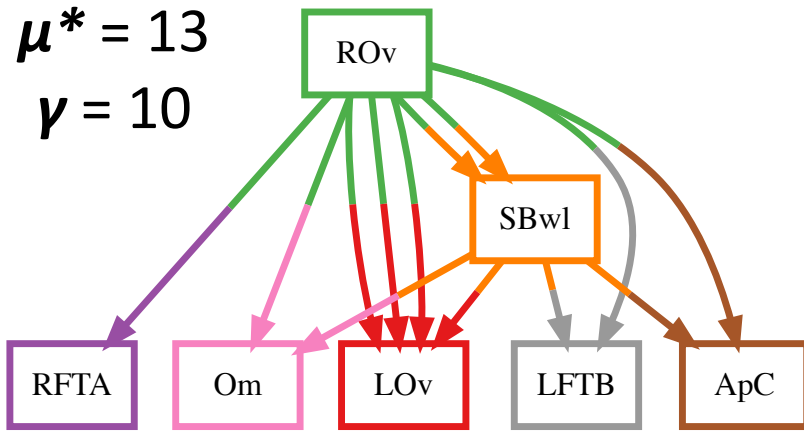
DAPI  
K14

ApC	Appendix
LFTB	Left Fallopian Tube
LOv	Left Ovary
RFTA	Right Fallopian Tube
ROv	Right Ovary
SBwl	Small Bowel
Om	Omentum

# Comigrations: Simultaneous Migrations of Multiple Clones

- Multiple tumor cells migrate simultaneously through the blood stream [Cheung et al., 2016]
- Second objective: number  $\gamma$  of **comigrations** is the number of multi-edges in migration graph  $G^\dagger$

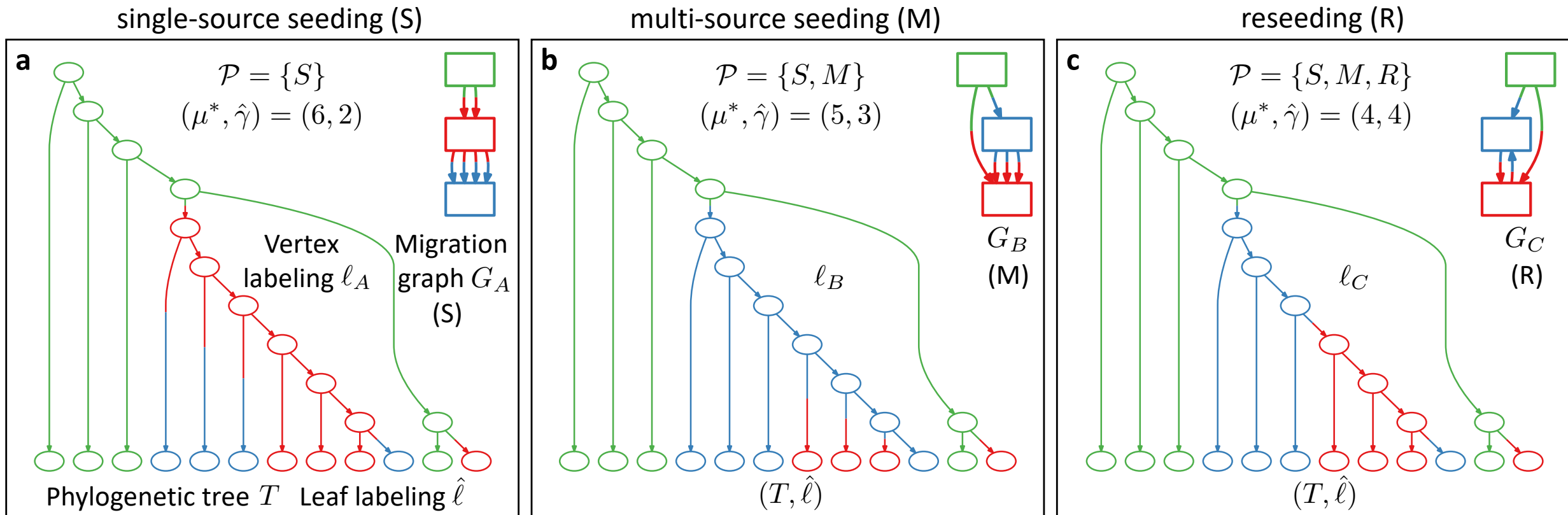
† Not necessarily true in the case of directed cycles



ApC	Appendix
LFTB	Left Fallopian Tube
LOv	Left Ovary
RFTA	Right Fallopian Tube
ROv	Right Ovary
SBwl	Small Bowel
Om	Omentum

# Constrained Multi-objective Optimization Problem

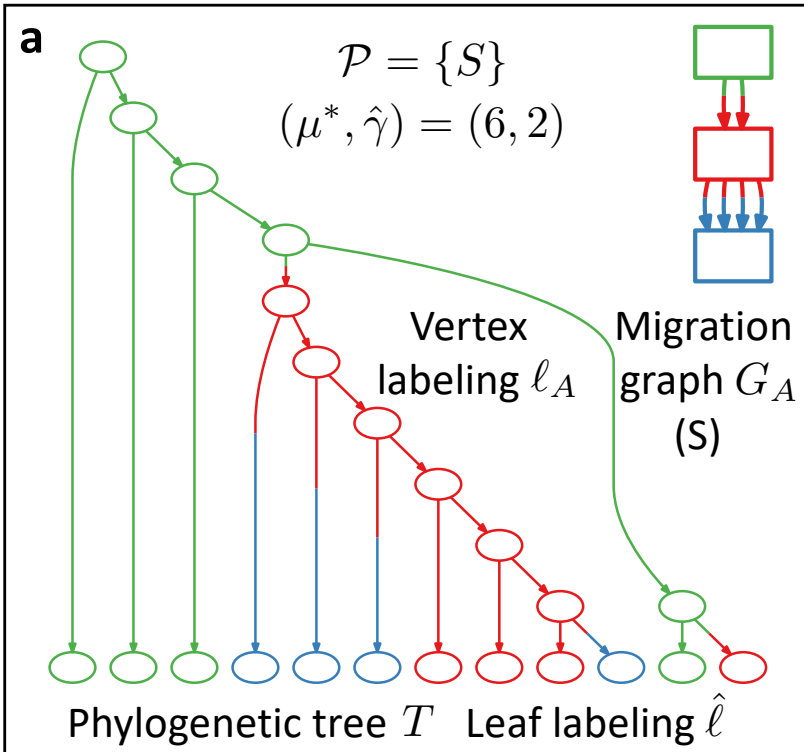
**Parsimonious Migration History (PMH):** Given a phylogenetic tree  $T$  and a set  $\mathcal{P} \subseteq \{S, M, R\}$  of allowed migration patterns, find vertex labeling  $\ell$  with minimum migration number  $\mu^*(T)$  and smallest comigration number  $\hat{\gamma}(T)$ .



# Results [El-Kebir, WABI 2018]

**Parsimonious Migration History (PMH):** Given a phylogenetic tree  $T$  and a set  $\mathcal{P} \subseteq \{S, M, R\}$  of allowed migration patterns, find vertex labeling  $\ell$  with minimum migration number  $\mu^*(T)$  and smallest comigration number  $\hat{\gamma}(T)$ .

single-source seeding (S)



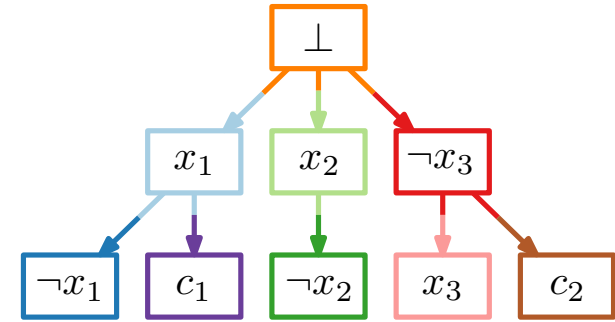
**Theorem 1:** PMH is NP-hard when  $\mathcal{P} = \{S\}$

**Theorem 2:** PMH is fixed parameter tractable in the number  $m$  of locations when  $\mathcal{P} = \{S\}$



# PMH is NP-hard when $\mathcal{P} = \{S\}$

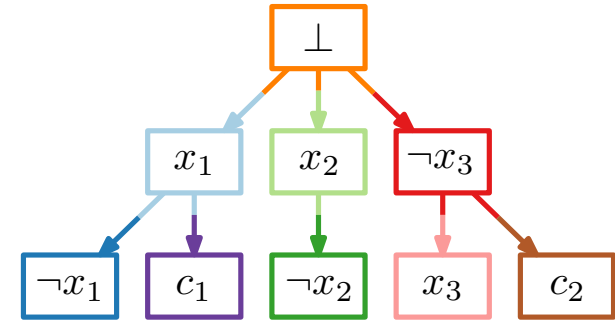
**3-SAT:** Given  $\varphi = \bigwedge_{i=1}^k (y_{i,1} \vee y_{i,2} \vee y_{i,3})$  with variables  $\{x_1, \dots, x_n\}$  and  $k$  clauses, find  $\phi : [n] \rightarrow \{0,1\}$  satisfying  $\varphi$



$$\Sigma = \{x_1, \dots, x_n, \neg x_1, \dots, \neg x_n, c_1, \dots, c_k, \perp\}$$

# PMH is NP-hard when $\mathcal{P} = \{S\}$

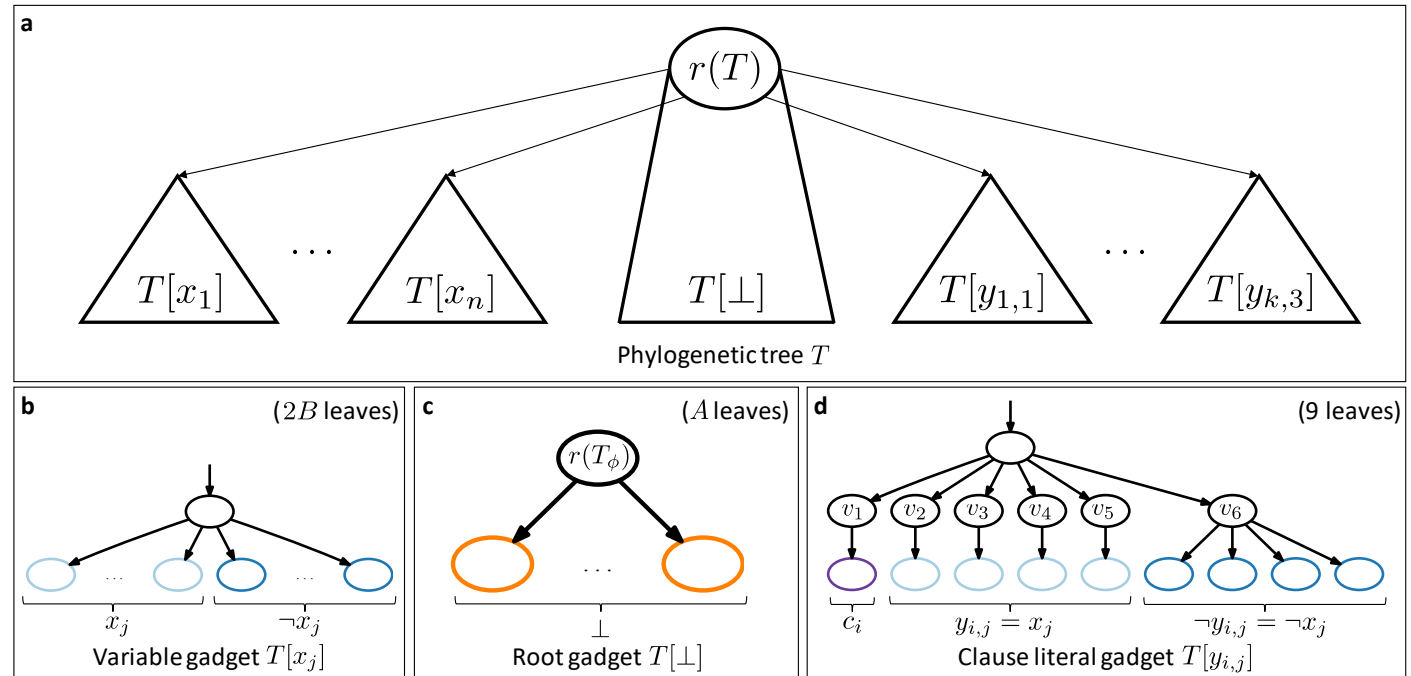
**3-SAT:** Given  $\varphi = \bigwedge_{i=1}^k (y_{i,1} \vee y_{i,2} \vee y_{i,3})$  with variables  $\{x_1, \dots, x_n\}$  and  $k$  clauses, find  $\phi : [n] \rightarrow \{0,1\}$  satisfying  $\varphi$



$$\Sigma = \{x_1, \dots, x_n, \neg x_1, \dots, \neg x_n, c_1, \dots, c_k, \perp\}$$

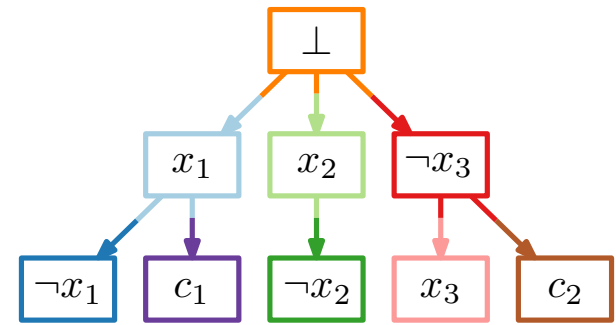
Three ideas:

1. Ensure that  $(x, \neg x) \in E(G)$  or  $(\neg x, x) \in E(G)$
2. Ensure that  $\ell^*(r(T)) = \perp$
3. Ensure that  $\varphi$  is satisfiable if and only if  $\ell^*$  encodes a satisfying truth assignment



# PMH is NP-hard when $\mathcal{P} = \{S\}$

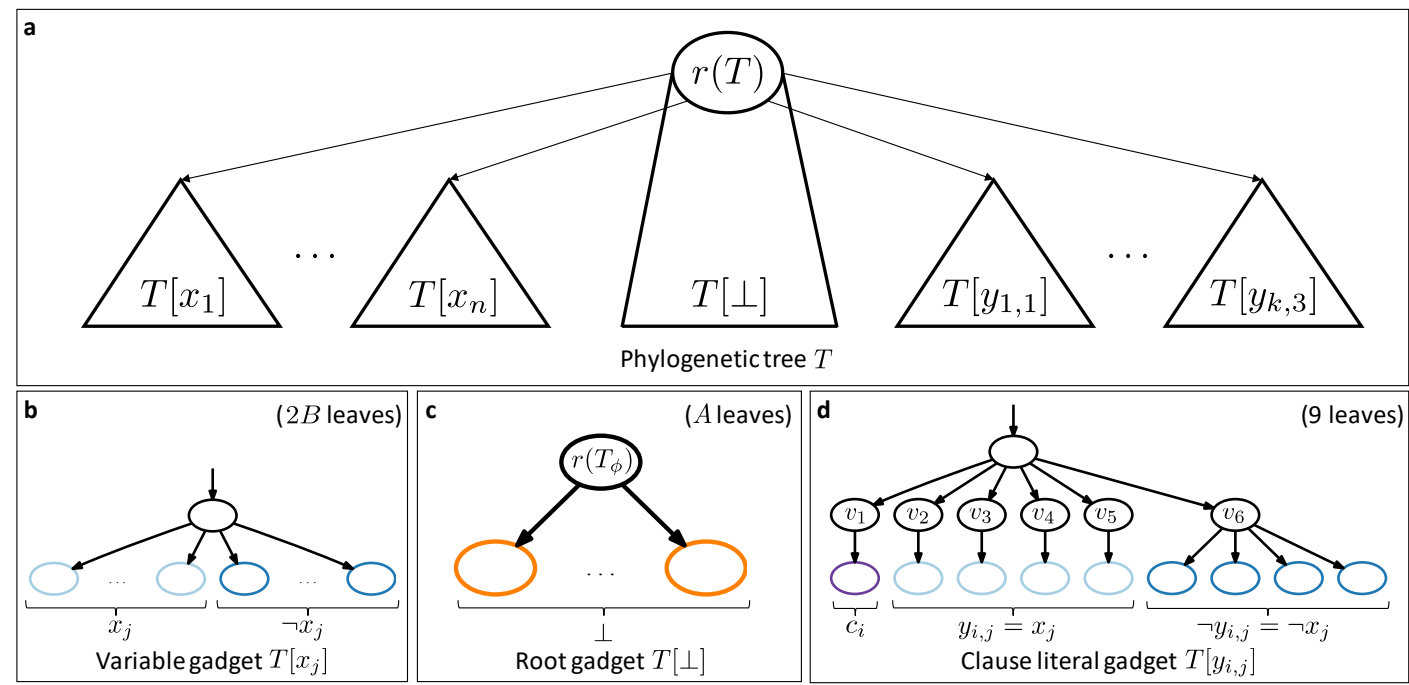
**3-SAT:** Given  $\varphi = \bigwedge_{i=1}^k (y_{i,1} \vee y_{i,2} \vee y_{i,3})$  with variables  $\{x_1, \dots, x_n\}$  and  $k$  clauses, find  $\phi : [n] \rightarrow \{0,1\}$  satisfying  $\varphi$



$$\Sigma = \{x_1, \dots, x_n, \neg x_1, \dots, \neg x_n, c_1, \dots, c_k, \perp\}$$

Three ideas:

1. Ensure that  $(x, \neg x) \in E(G)$  or  $(\neg x, x) \in E(G)$
2. Ensure that  $\ell^*(r(T)) = \perp$
3. Ensure that  $\varphi$  is satisfiable if and only if  $\ell^*$  encodes a satisfying truth assignment



**Lemma:** Let  $B > 10k + 1$  and  $A > 2Bn + 27k$ .  
 Then,  $\varphi$  is satisfiable if and only if  $\mu^*(T) = (B + 1)n + 25k$

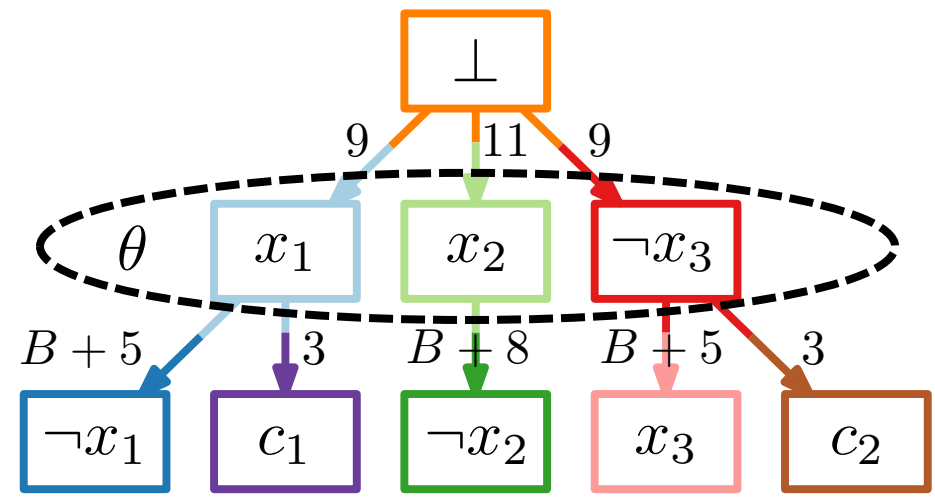
# PMH is NP-hard when $\mathcal{P} = \{S\}$

$$\varphi = (x_1 \vee x_2 \vee \neg x_3) \wedge (\neg x_1, \neg x_2, \neg x_3)$$

$$k = 2, n = 3$$

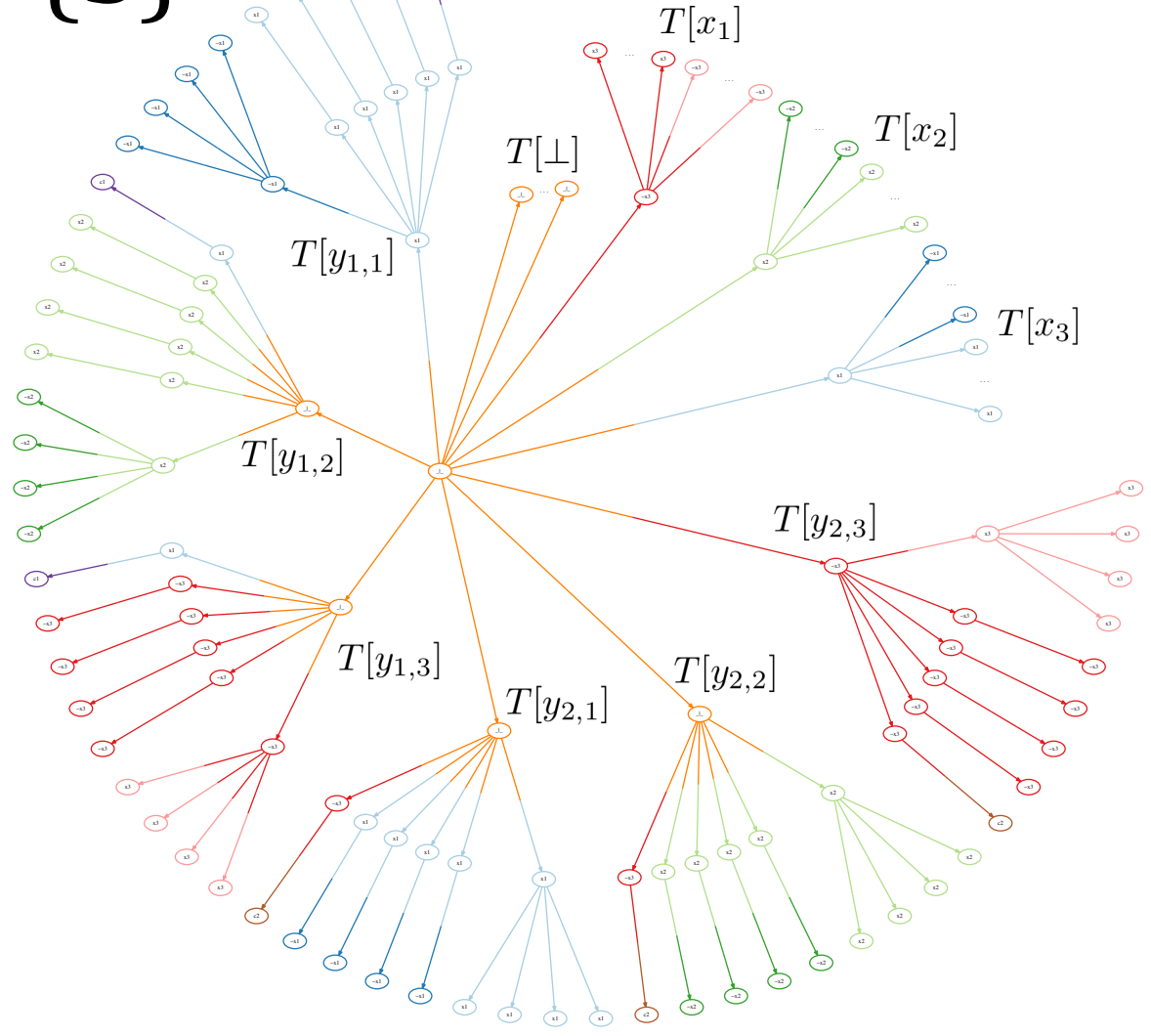
$$B = 10k + 2 = 22$$

$$A = 2Bn + 27k + 1 = 187$$



$$\Sigma = \{x_1, x_2, x_3, \neg x_1, \neg x_2, \neg x_3, c_1, c_2, \perp\}$$

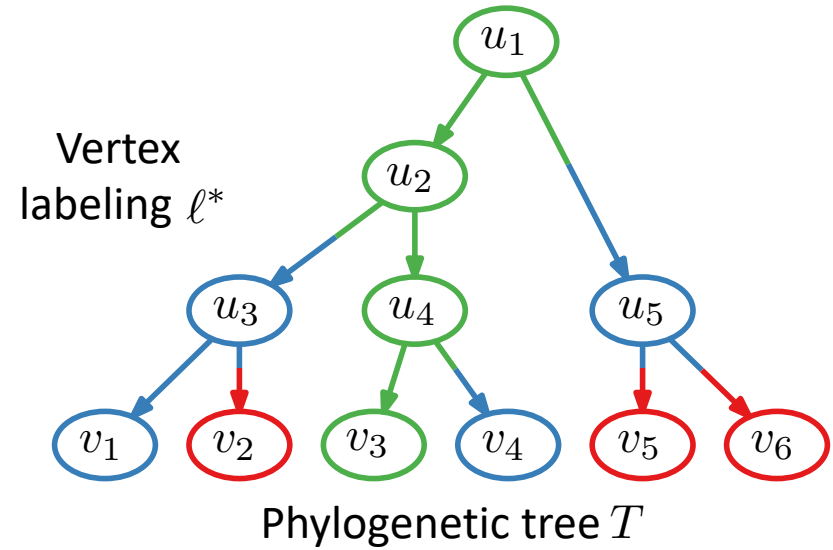
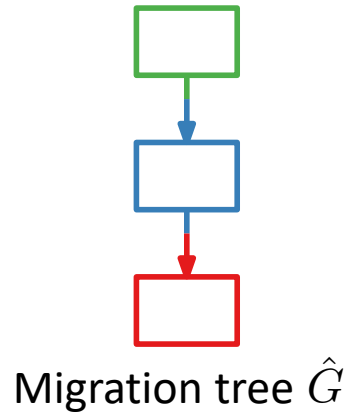
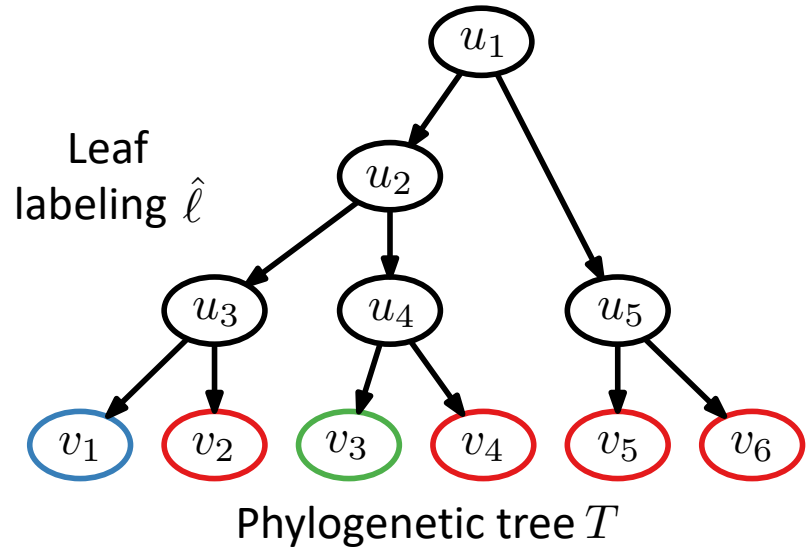
$$\begin{aligned} \mu^*(T) &= (B + 1)n + 25k \\ &= 23 * 3 + 50 * 2 = 119 \end{aligned}$$



**Lemma:** Let  $B > 10k + 1$  and  $A > 2Bn + 27k$ .

Then,  $\varphi$  is satisfiable if and only if  $\mu^*(T) = (B + 1)n + 25k$

PMH is FPT in number  $m$  of locations when  $\mathcal{P} = \{S\}$



**Lemma:** If there exists labeling  $\ell$  consistent with  $\hat{G}$  then

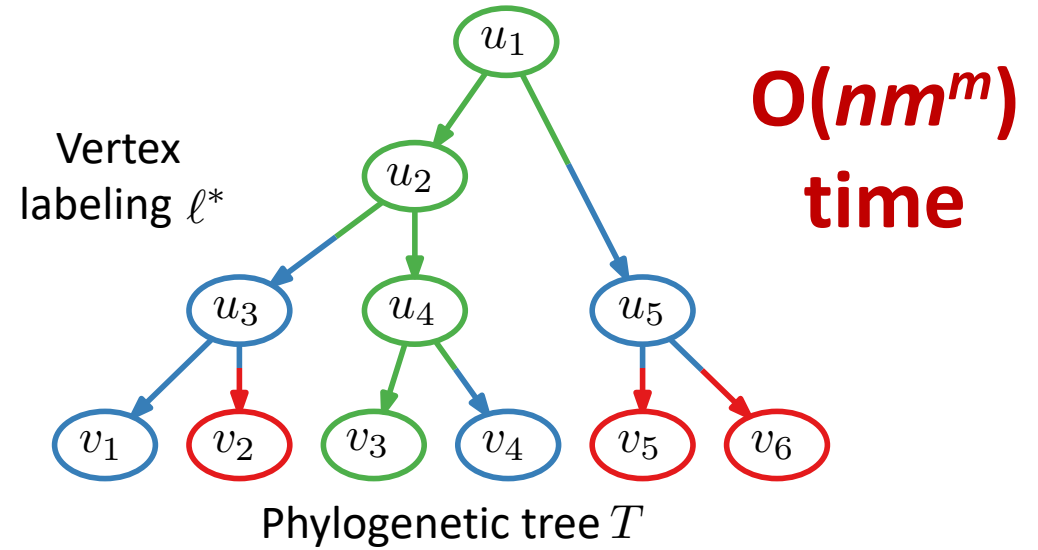
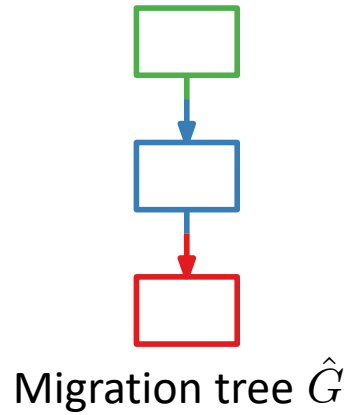
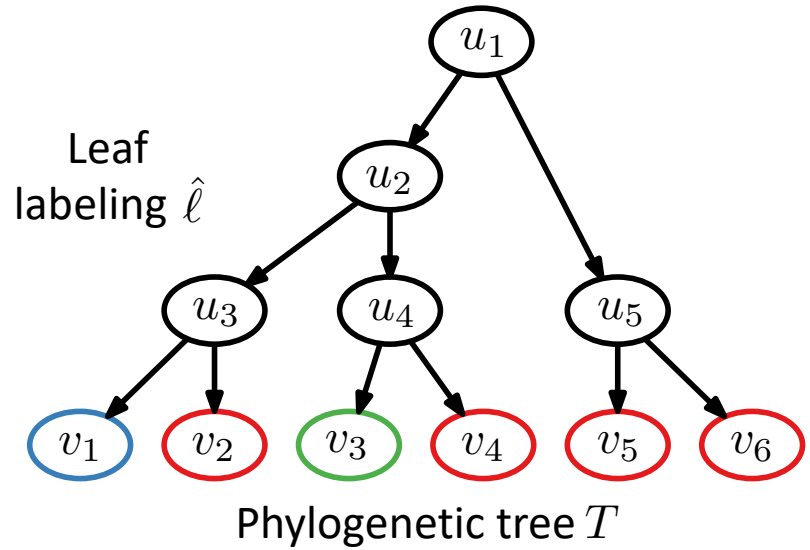
$$d_T(u, v) \geq d_{\hat{G}}(\text{lca}_{\hat{G}}(u), \hat{\ell}(v)) \quad \forall u, v \in V(T) \text{ such that } u \preceq_T v. \quad (1)$$

$$\ell^*(v) = \begin{cases} \text{LCA}_{\hat{G}}(r(T)), & \text{if } v = r(T), \\ \sigma(\ell^*(\pi(v)), \text{LCA}_{\hat{G}}(v)), & \text{if } v \neq r(T), \end{cases}$$

where  $\sigma(s, t) = s$  if  $s = t$  and otherwise  $\sigma(s, t)$  is the unique child of  $s$  that lies on the path from  $s$  to  $t$  in  $\hat{G}$ .

**Lemma:** If (1) holds then  $\ell^*$  is a minimum migration labeling consistent with  $\hat{G}$ .

PMH is FPT in number  $m$  of locations when  $\mathcal{P} = \{S\}$



**Lemma:** If there exists labeling  $\ell$  consistent with  $\hat{G}$  then

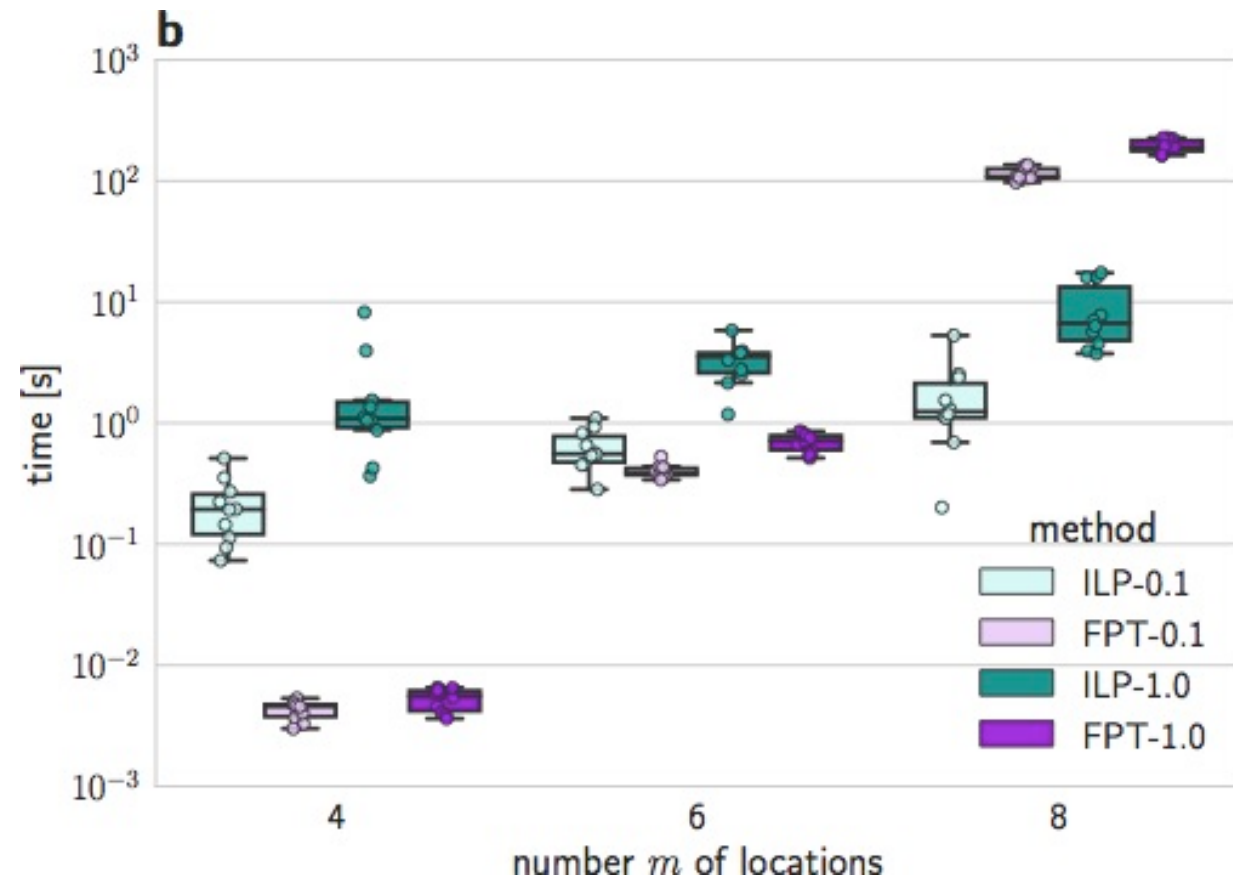
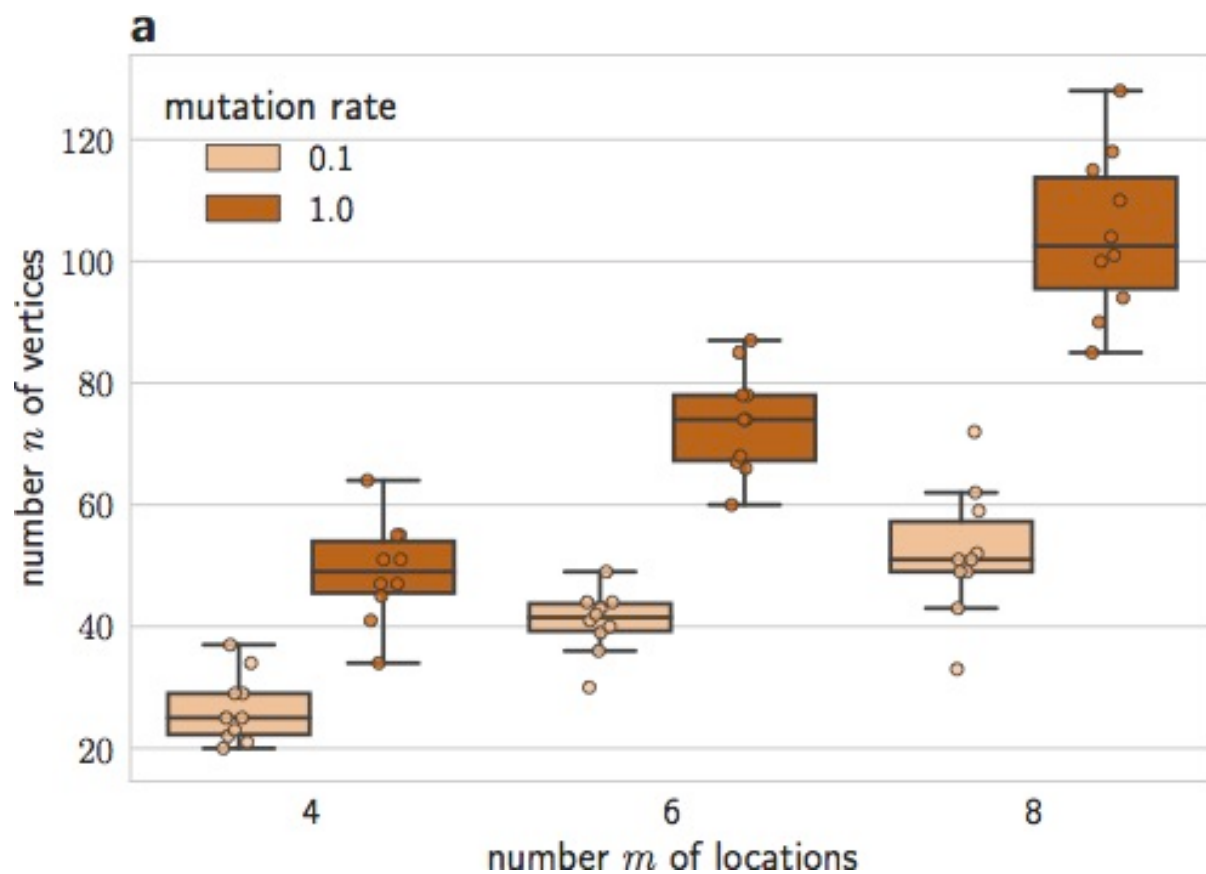
$$d_T(u, v) \geq d_{\hat{G}}(\text{lca}_{\hat{G}}(u), \hat{\ell}(v)) \quad \forall u, v \in V(T) \text{ such that } u \preceq_T v. \quad (1)$$

$$\ell^*(v) = \begin{cases} \text{LCA}_{\hat{G}}(r(T)), & \text{if } v = r(T), \\ \sigma(\ell^*(\pi(v)), \text{LCA}_{\hat{G}}(v)), & \text{if } v \neq r(T), \end{cases}$$

where  $\sigma(s, t) = s$  if  $s = t$  and otherwise  $\sigma(s, t)$  is the unique child of  $s$  that lies on the path from  $s$  to  $t$  in  $\hat{G}$ .

**Lemma:** If (1) holds then  $\ell^*$  is a minimum migration labeling consistent with  $\hat{G}$ .

# Simulations



Available on: <https://github.com/elkebir-group/PMH-S>