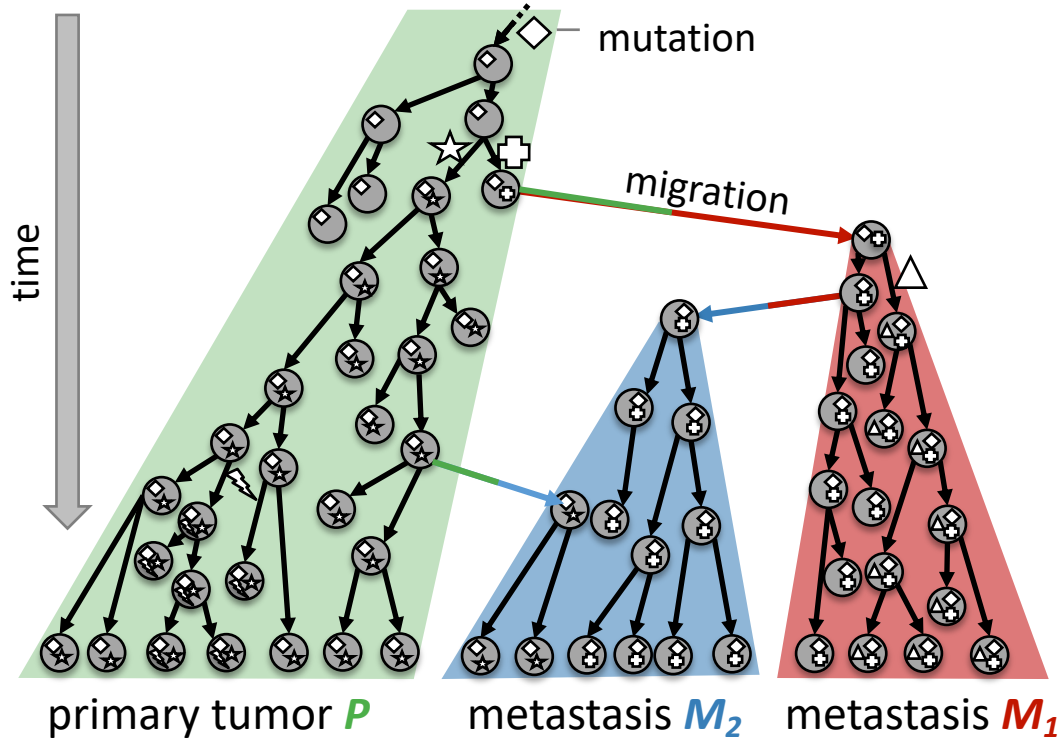
Parsimonious Migration History Problem: Complexity and Algorithms

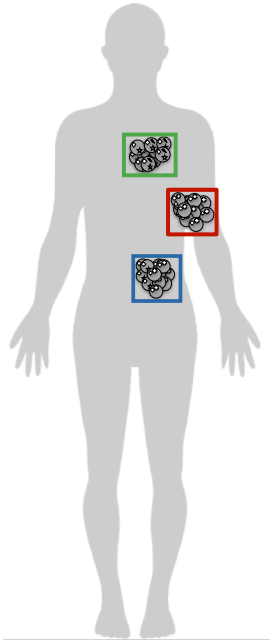
Mohammed El-Kebir, University of Illinois at Urbana-Champaign

WABI 2018

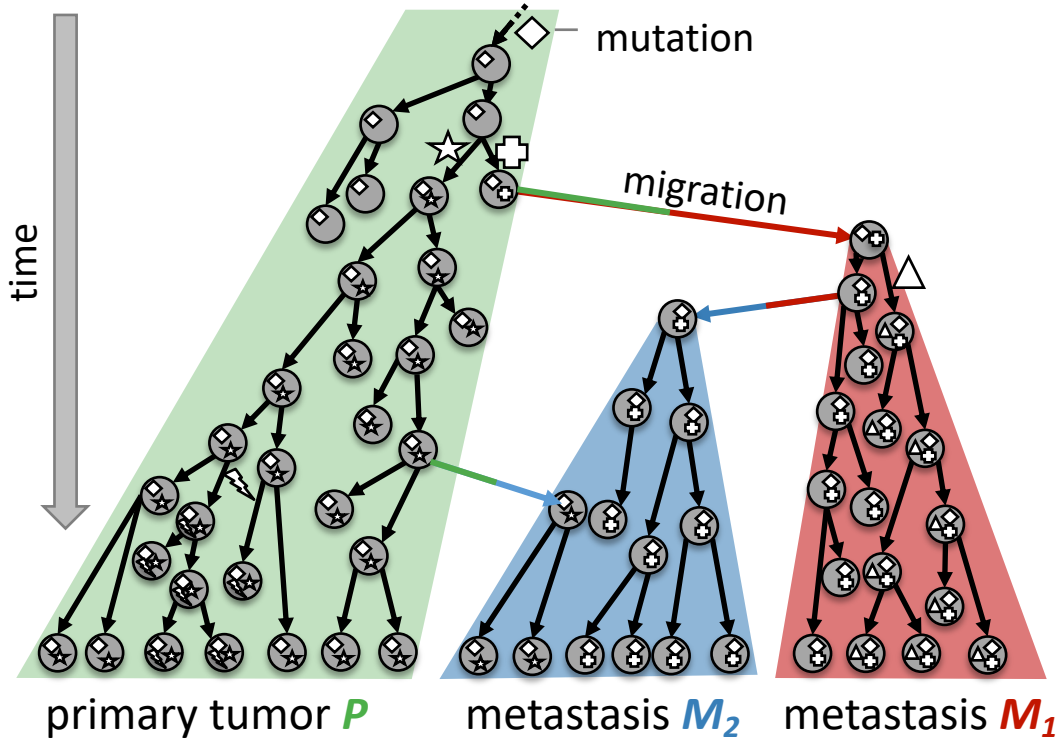
Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration



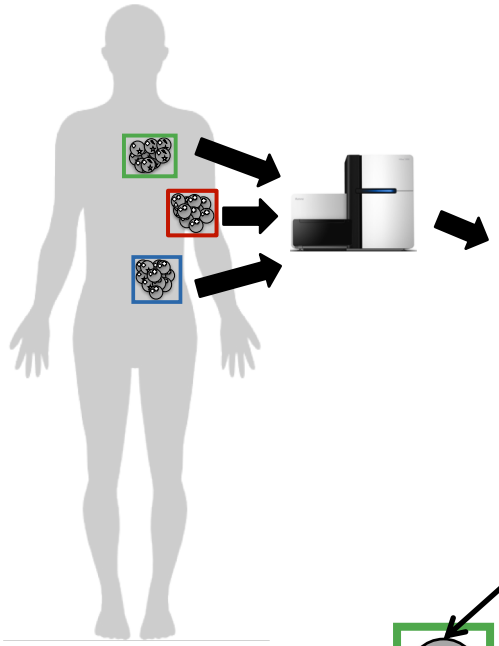
Cell Tree

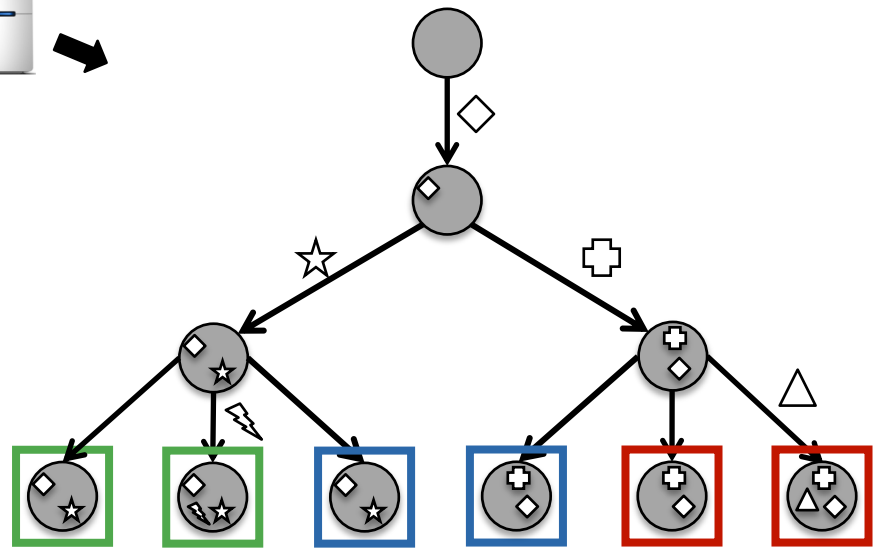


Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration



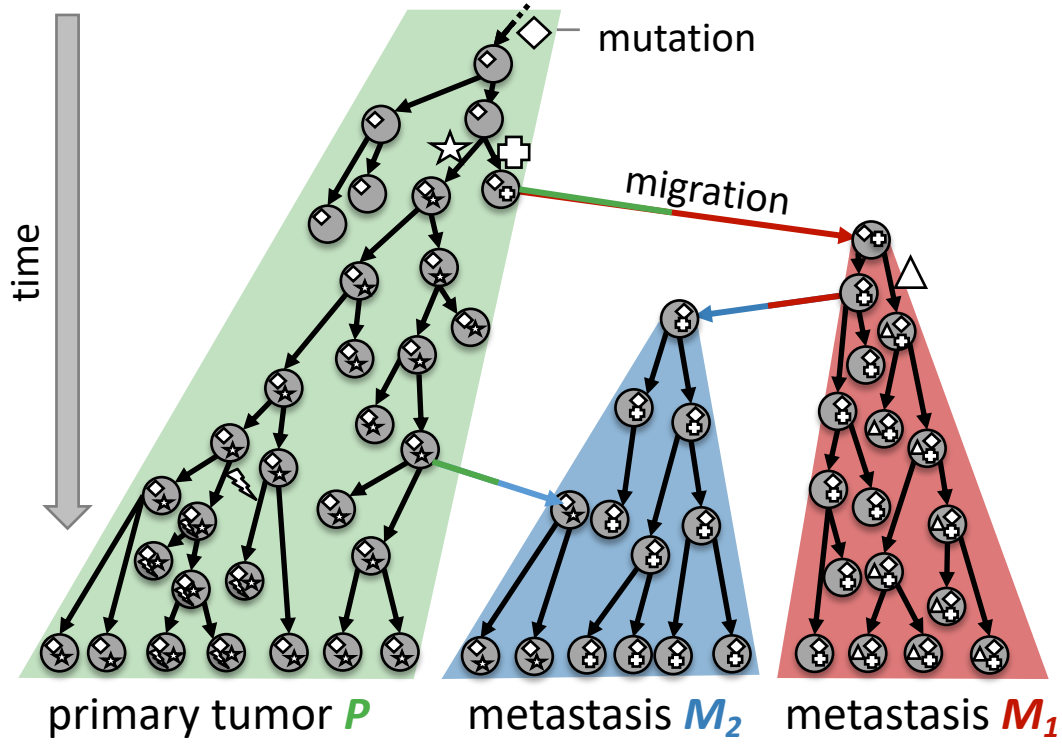
Cell Tree



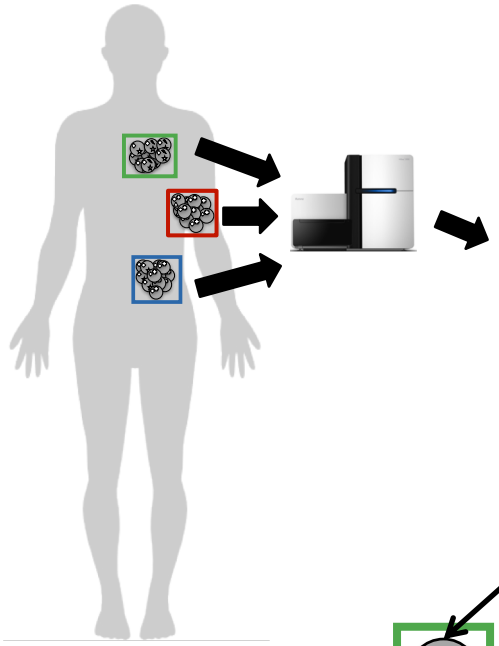


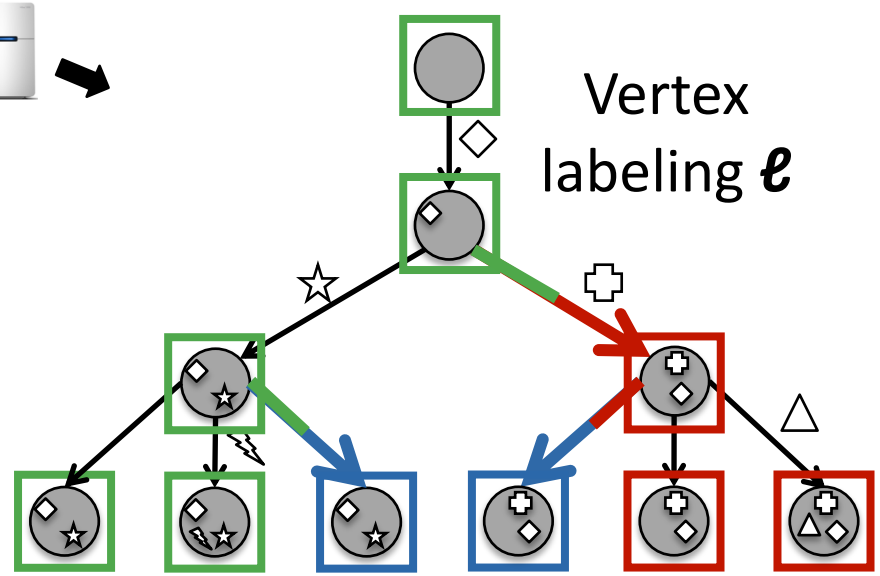
Phylogenetic Tree *T*

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



Cell Tree





Phylogenetic Tree *T*

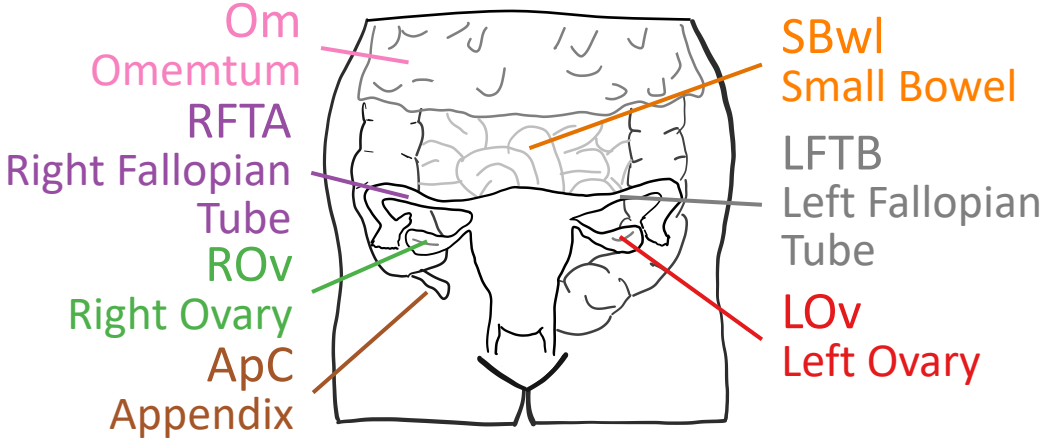
Goal: Given phylogenetic tree ***T***, find ***parsimonious*** vertex labeling ℓ with fewest migrations

Slatkin, M. and Maddison, W. P. (1989). A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics*, *123*(3), 603–613.

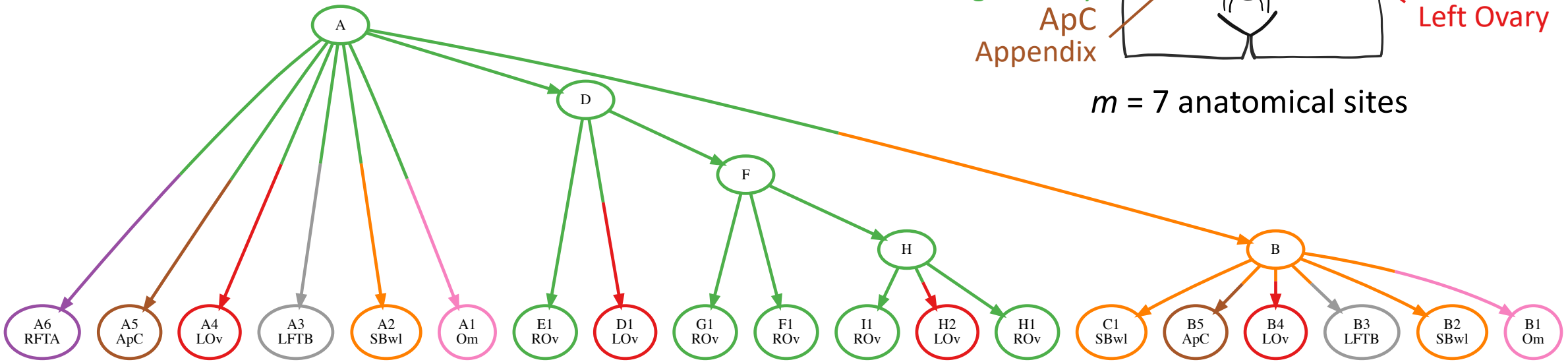
Minimum Migration Analysis in Ovarian Cancer

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]



m = 7 anatomical sites

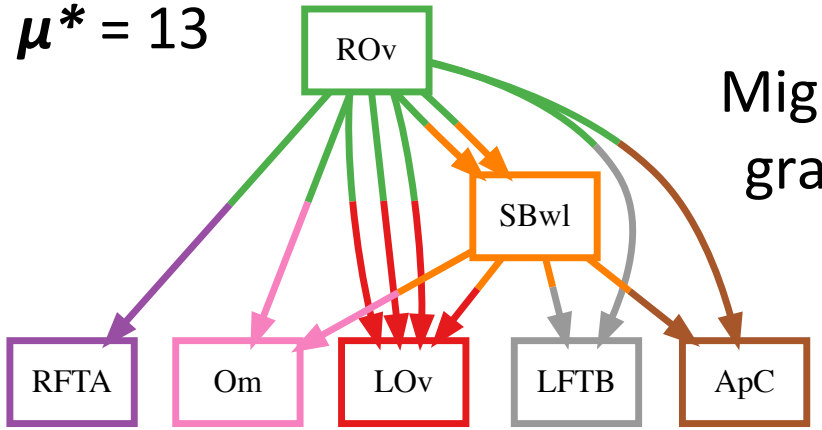


Minimum Migration Analysis in Ovarian Cancer

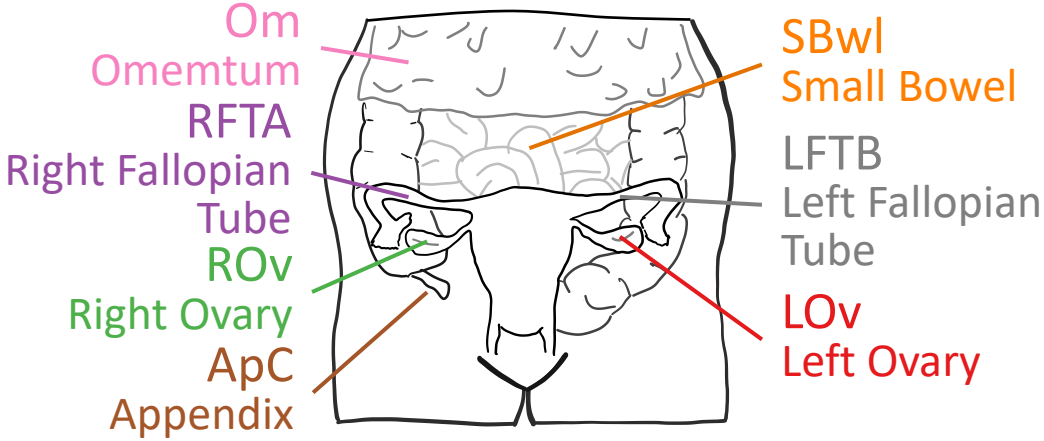
McPherson et al. (2016). Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics*.

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

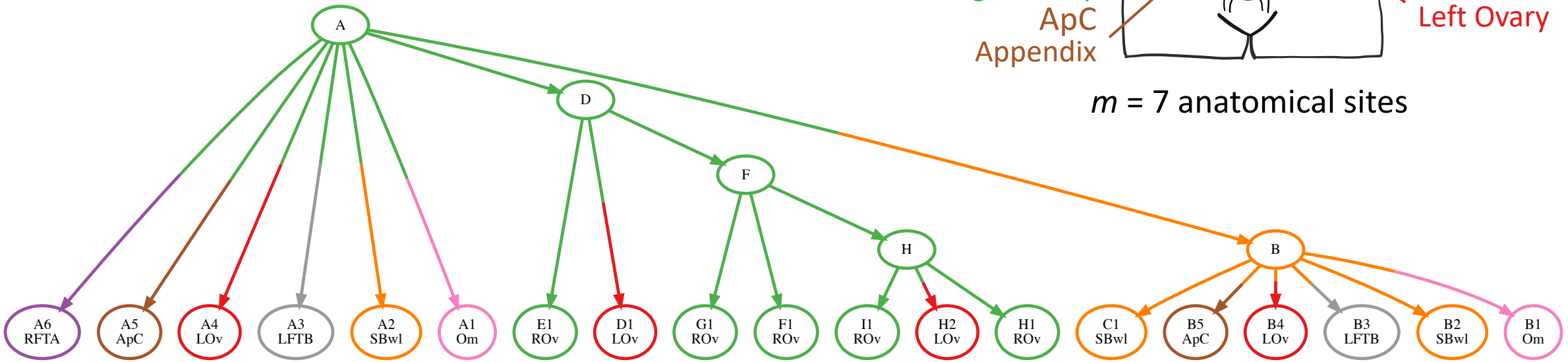
$\mu^* = 13$



Migration graph **G**



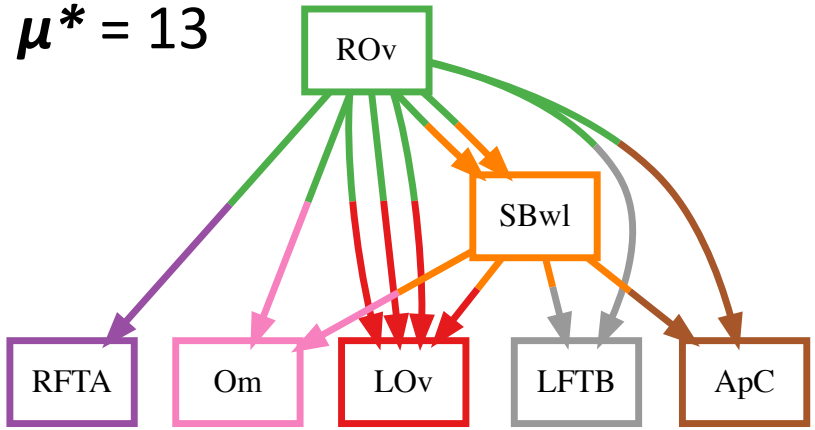
m = 7 anatomical sites



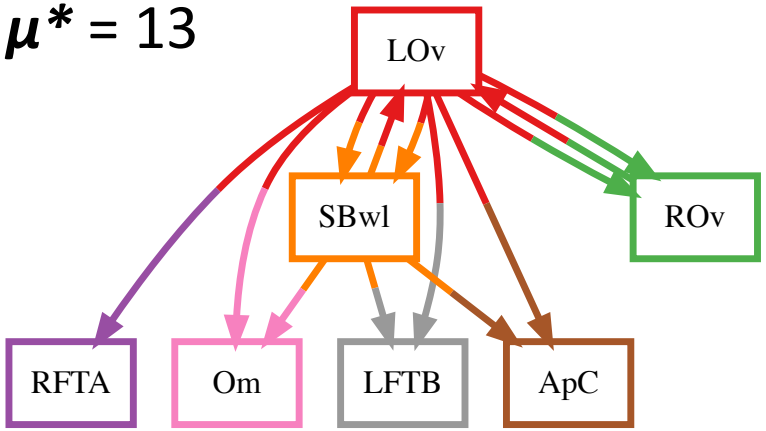
Minimum Migration History is ***Not*** Unique

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

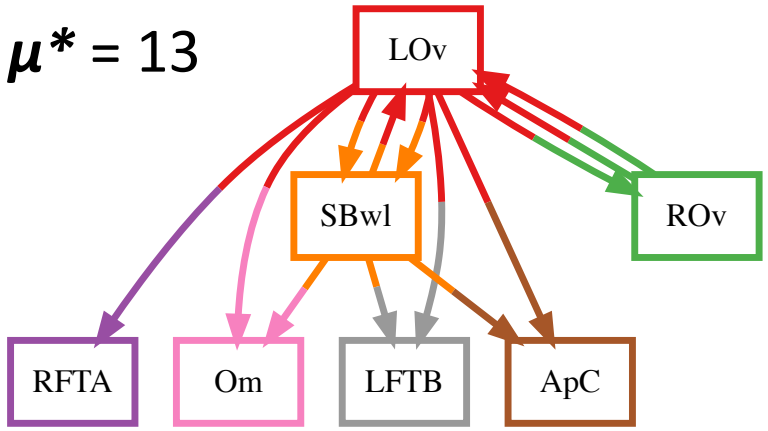
$\mu^* = 13$



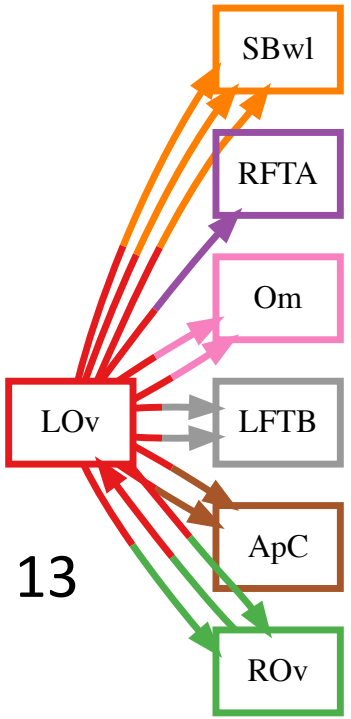
$\mu^* = 13$



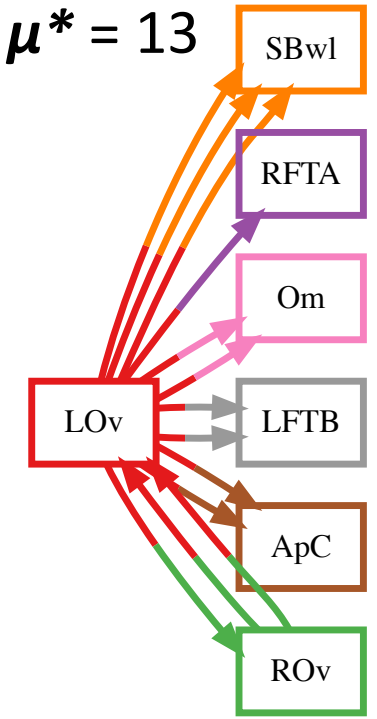
$\mu^* = 13$



$\mu^* = 13$



$\mu^* = 13$



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

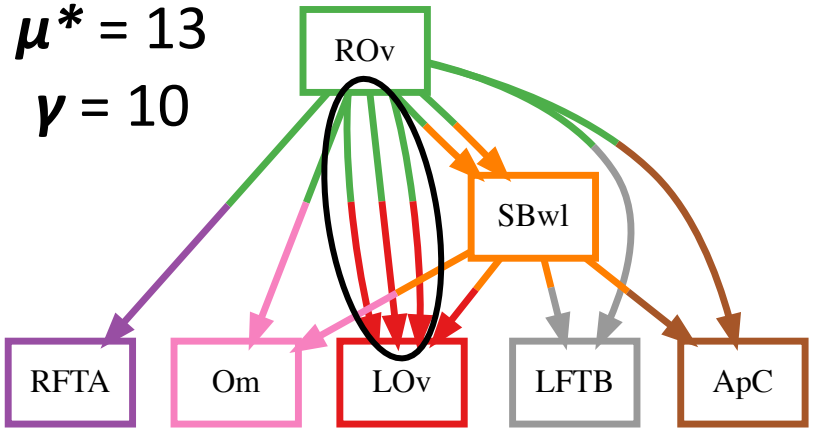
Comigrations: Simultaneous Migrations of Multiple Clones

- Multiple tumor cells migrate simultaneously through the blood stream [Cheung et al., 2016]
- Second objective: number γ of comigrations is the number of multi-edges in migration graph G †

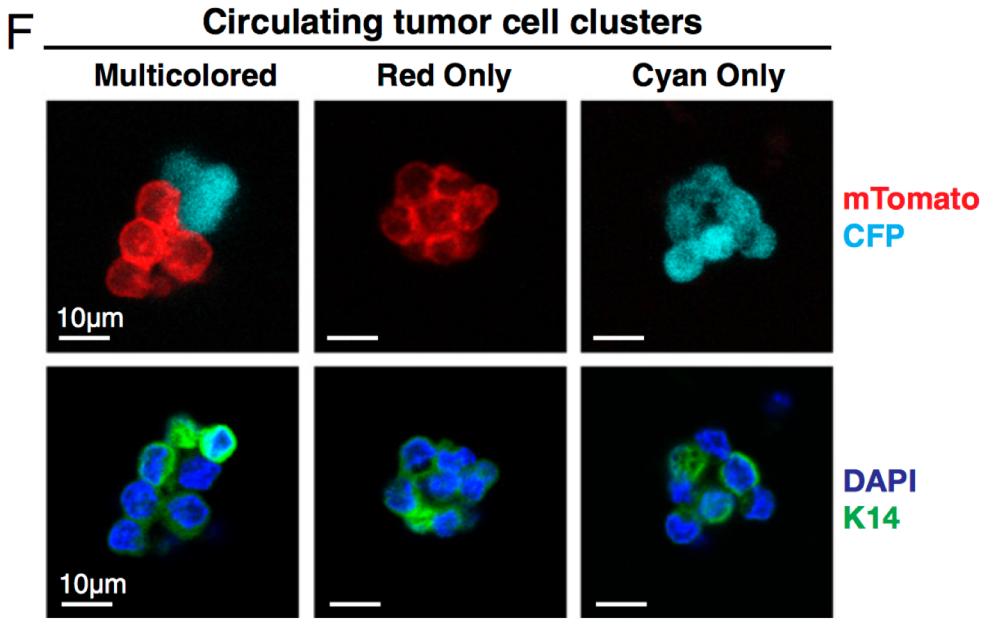
† Not necessarily true in the case of directed cycles

$\mu^* = 13$

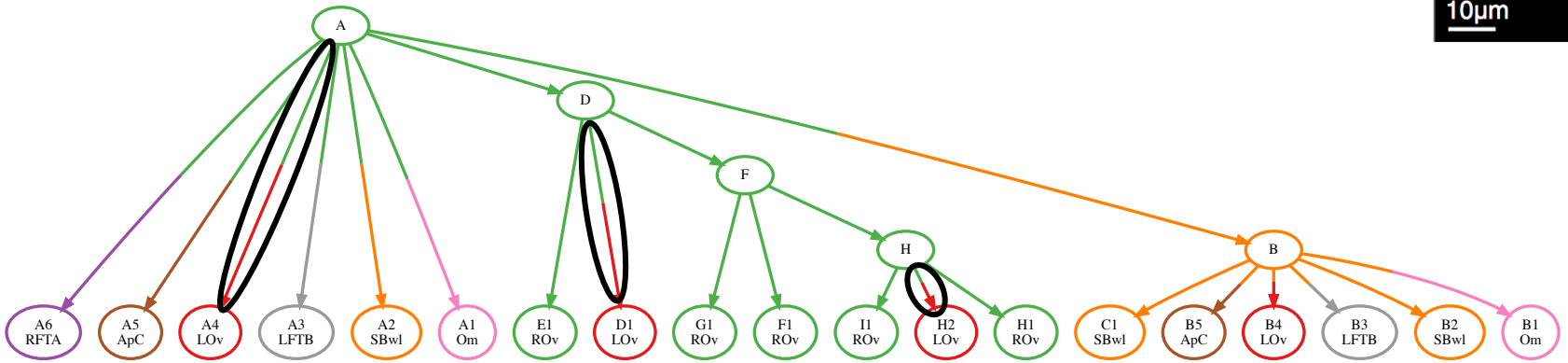
$\gamma = 10$



Migration Graph G



Clone Tree T



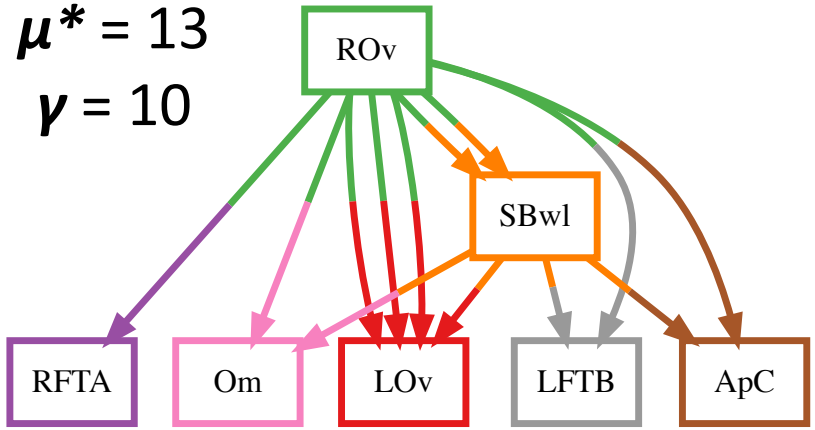
ApC	Appendix
LFTB	Left Fallopian Tube
LOv	Left Ovary
RFTA	Right Fallopian Tube
ROv	Right Ovary
SBwl	Small Bowel
Om	Omentum

Comigrations: Simultaneous Migrations of Multiple Clones

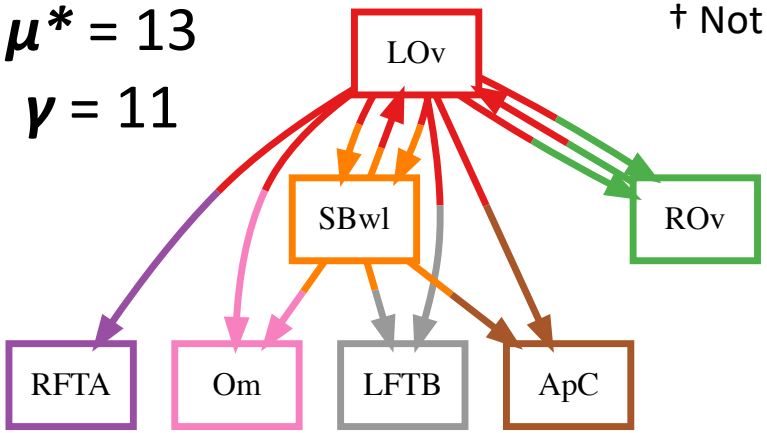
- Multiple tumor cells migrate simultaneously through the blood stream [Cheung et al., 2016]
- Second objective: number γ of comigrations is the number of multi-edges in migration graph G †

† Not necessarily true in the case of directed cycles

$\mu^* = 13$
$\gamma = 10$

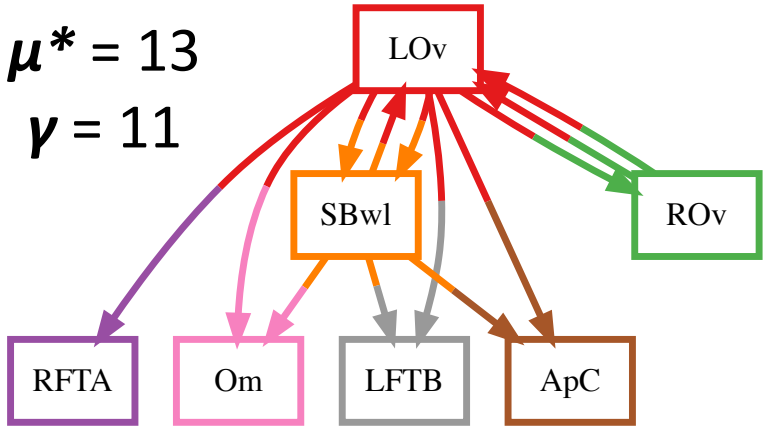


$\mu^* = 13$
$\gamma = 11$

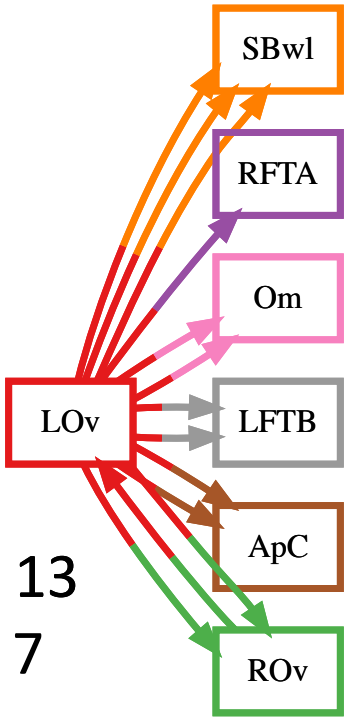


ApC	Appendix
LFTB	Left Fallopian Tube
LOv	Left Ovary
RFTA	Right Fallopian Tube
ROv	Right Ovary
SBwl	Small Bowel
Om	Omentum

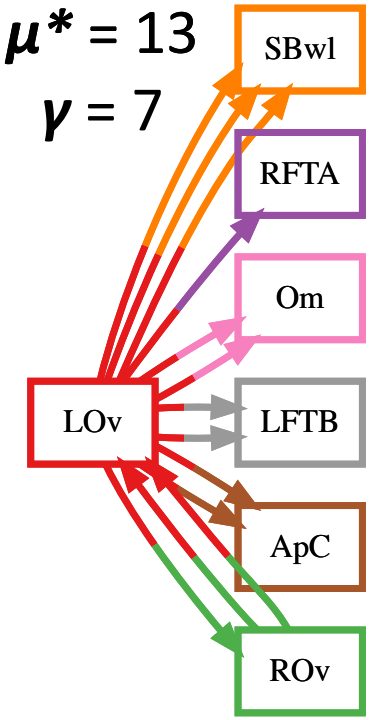
$\mu^* = 13$
$\gamma = 11$



$\mu^* = 13$
$\gamma = 7$

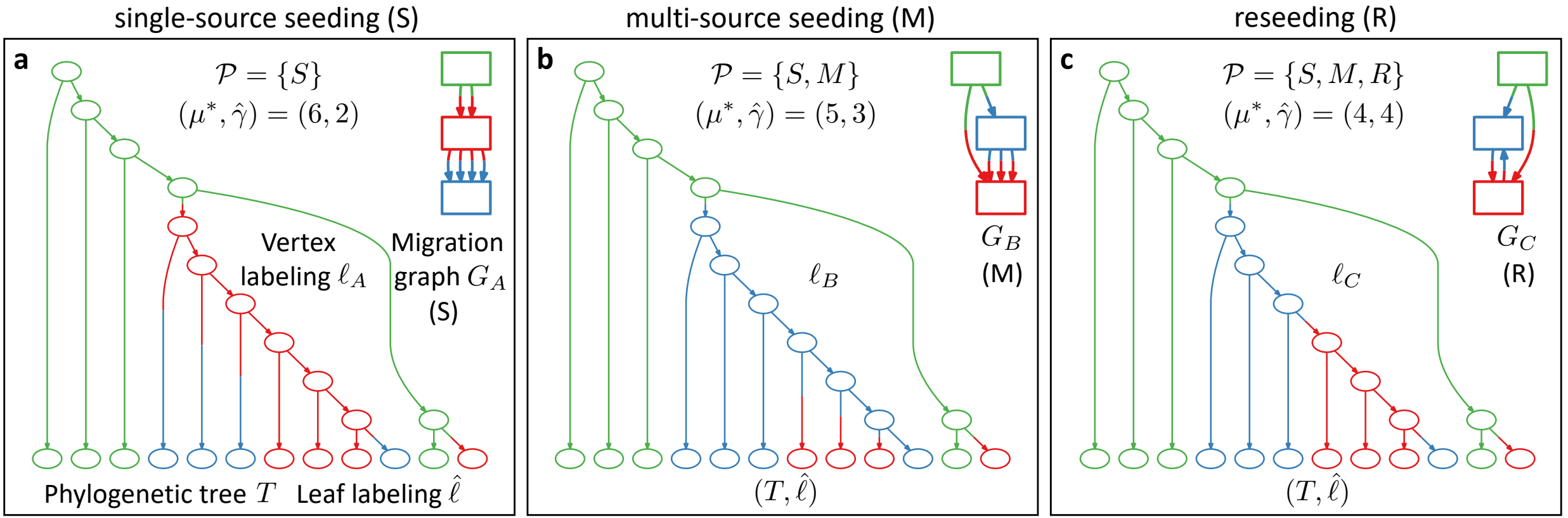


$\mu^* = 13$
$\gamma = 7$



Constrained Multi-objective Optimization Problem

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

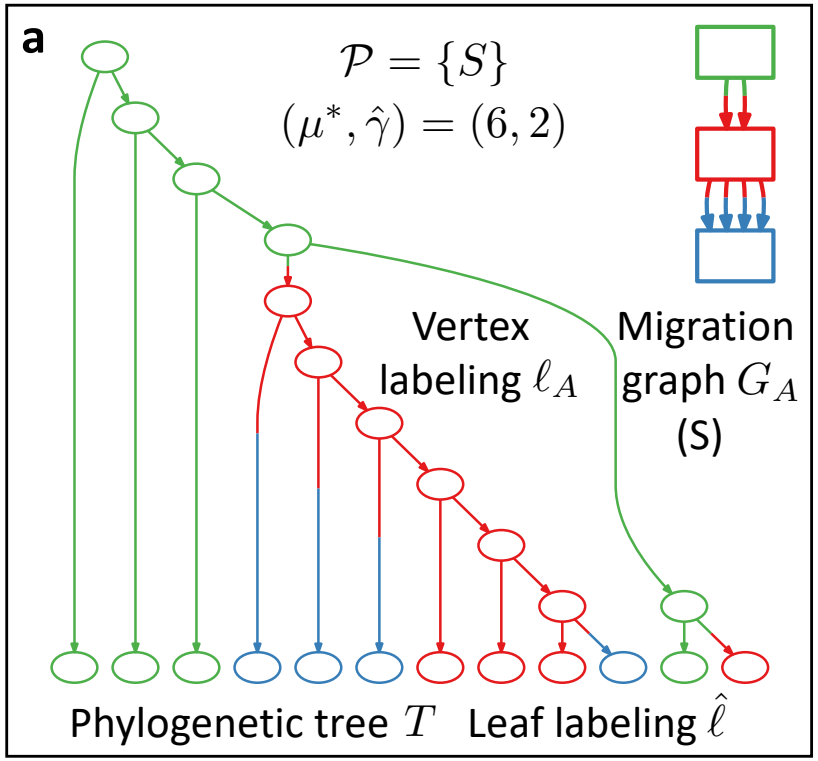


El-Kebir, M., Satas, G., & Raphael, B. J. (2018). Inferring parsimonious migration histories for metastatic cancers. *Nature Genetics*, *50*(5), 718–726.

Results

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

single-source seeding (S)

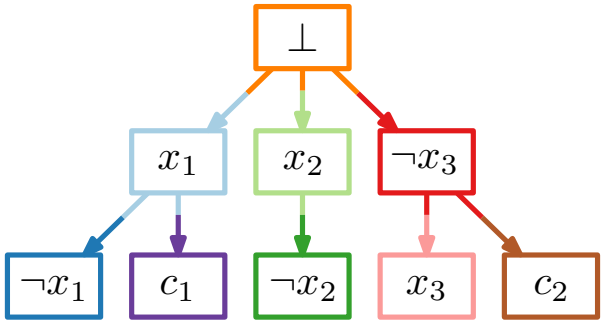


Theorem 1: PMH is NP-hard when $\mathcal{P} = \{\mathrm{S}\}$

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

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

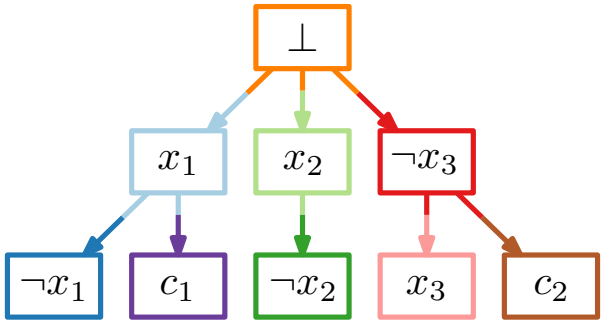
3-SAT: Given $\varphi = \Lambda_{i=1}^{k}(y_{i,1} \vee y_{i,2} \vee y_{i,3})$ with variables $\{x_1, \dots, x_n\}$ and k clauses, find $\phi : [n] \to \{0,1\}$ satisfying φ



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

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

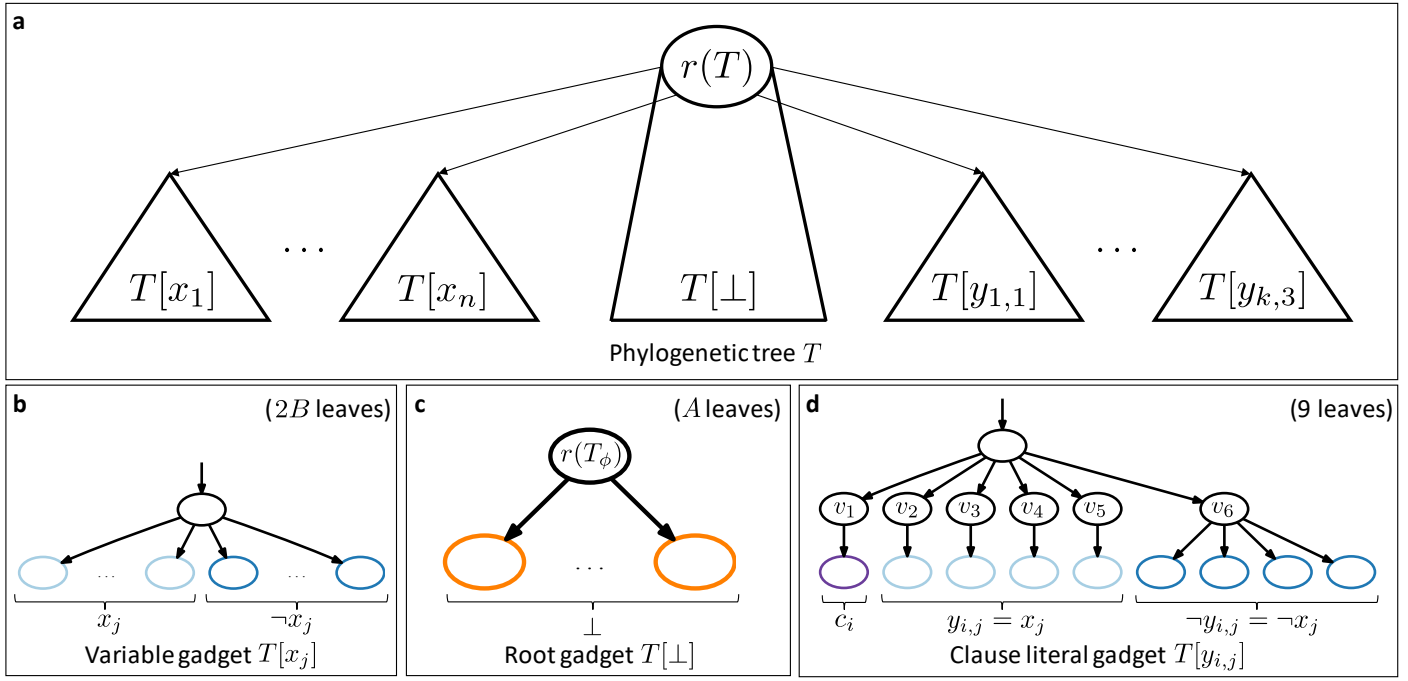
3-SAT: Given $\varphi = \Lambda_{i=1}^{k}(y_{i,1} \vee y_{i,2} \vee y_{i,3})$ with variables $\{x_1, \ldots, x_n\}$ and k clauses, find $\phi : [n] \to \{0,1\}$ satisfying φ



$$\Sigma = \{x_1, \ldots, x_n, \neg x_1, \ldots, \neg x_n, c_1, \ldots c_k, \bot\}$$

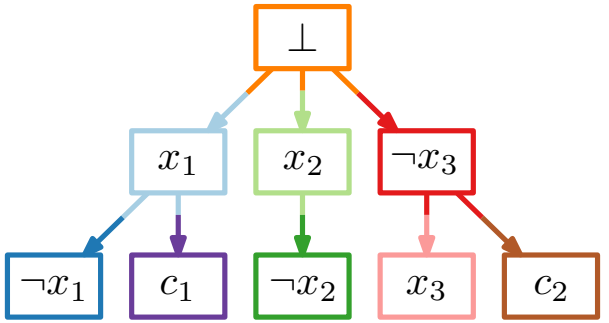
Three ideas:

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



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

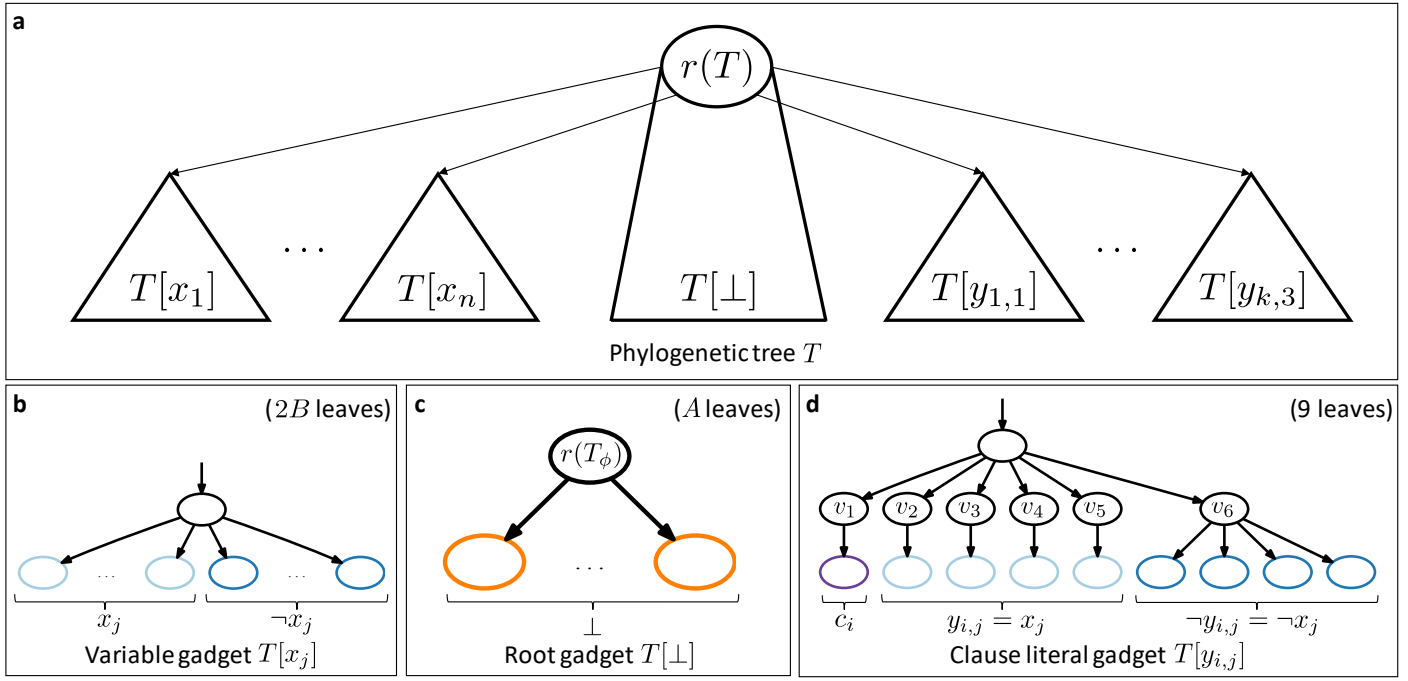
3-SAT: Given $\varphi = \Lambda_{i=1}^{k}(y_{i,1} \vee y_{i,2} \vee y_{i,3})$ with variables $\{x_1, \dots, x_n\}$ and k clauses, find $\phi : [n] \to \{0,1\}$ satisfying φ



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

Three ideas:

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



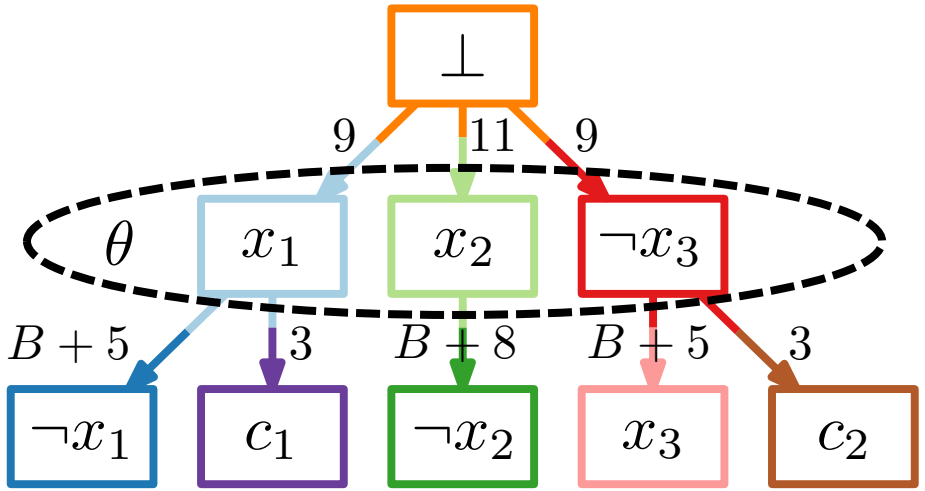
Lemma: Let $B > 10k + 1$ and $A > 2Bn + 27k$.
Then, φ is satisfiable if and only if $\mu^*(T) = (B + 1)n + 25k$

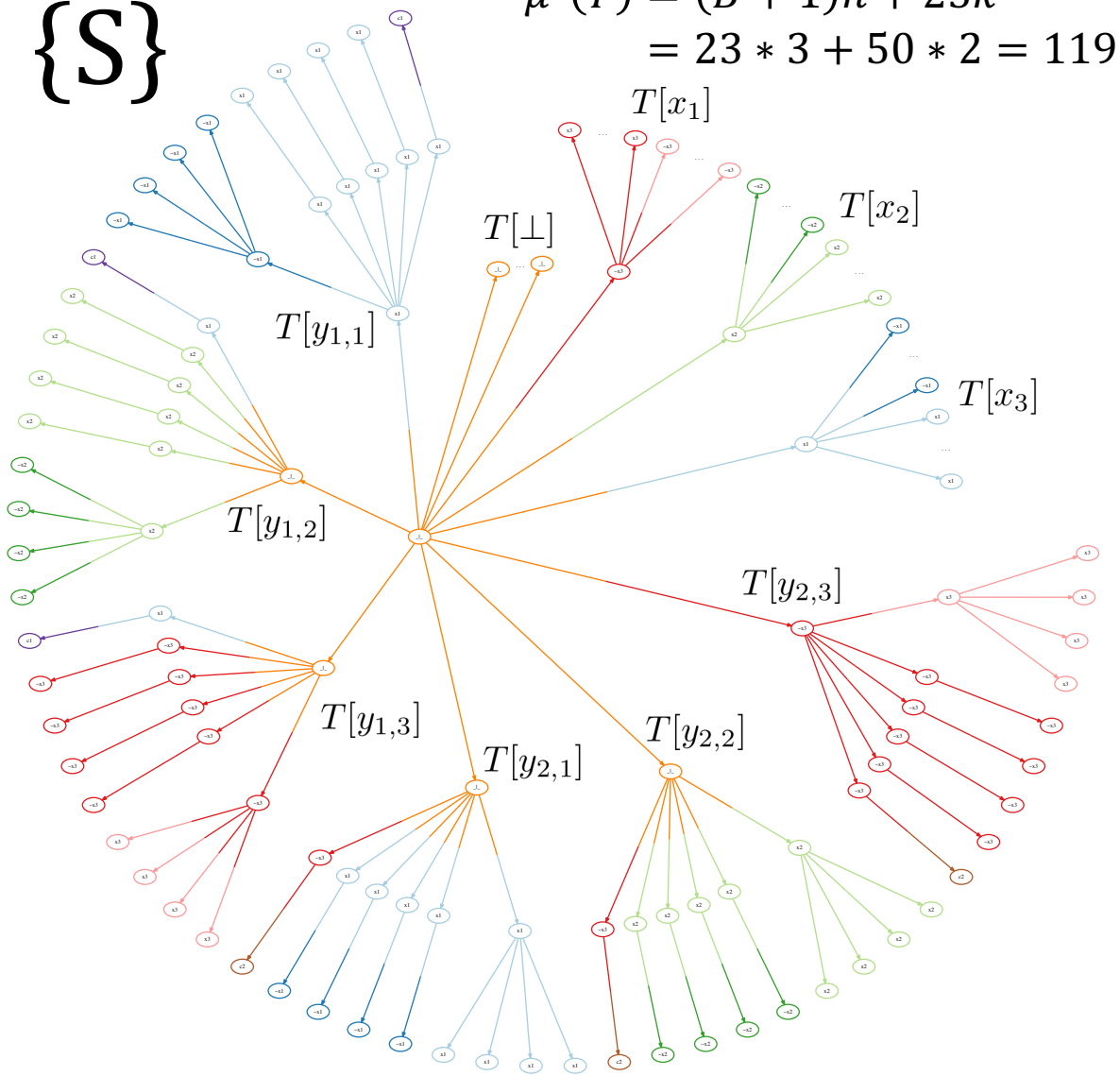
PMH is NP-hard when $\mathcal{P} = \{S\}$

$$\mu^*(T) = (B+1)n + 25k$$
$$= 23 * 3 + 50 * 2 = 119$$

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

$$k = 2, n = 3$$
$$B = 10k + 2 = 22$$
$$A = 2Bn + 27\mathrm{k} + 1 = 187$$

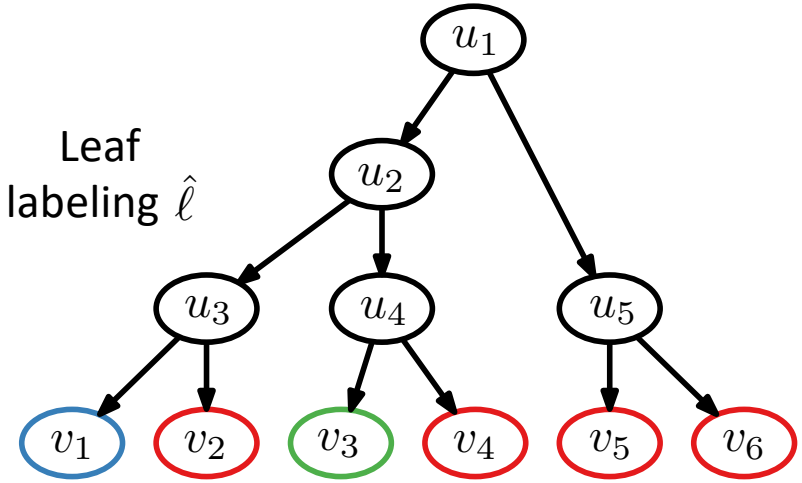




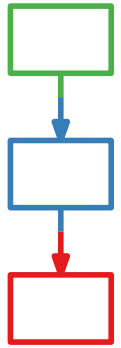
$$\Sigma = \{x_1, x_2, x_3, \neg x_1, \neg x_2, \neg x_3, c_1, c_2, \bot\}$$

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

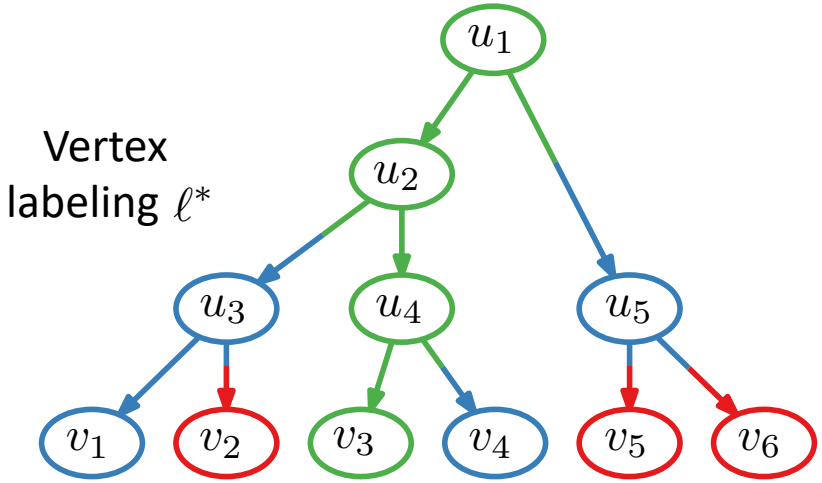
PMH is FPT in number m of locations when $\mathcal{P} = \{S\}$



Phylogenetic tree T



Migration tree $\hat{G}$



Phylogenetic tree T

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

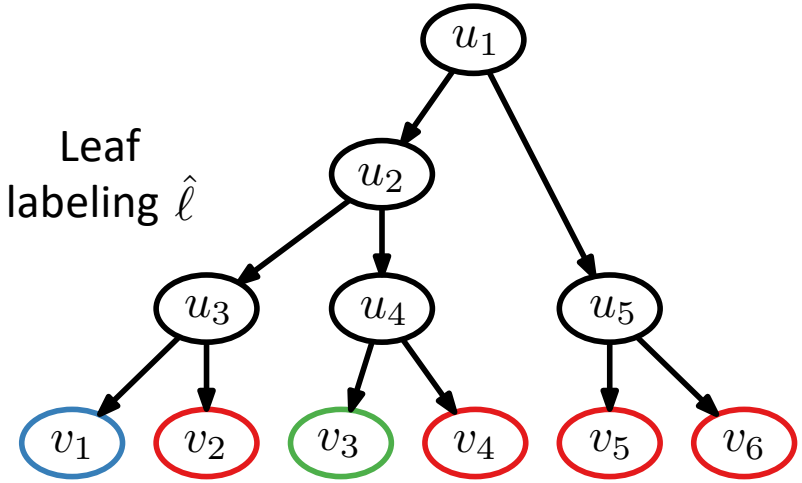
$$d_T(u,v) \geq d_{\hat{G}}(\mathrm{lca}_{\hat{G}}(u), \hat{\ell}(v)) \qquad \forall u, v \in V(T) \text{ such that } u \preceq_T v. \quad (1)$$

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

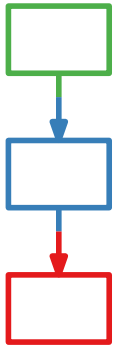
where $\sigma(s,t) = s$ if $s = t$ and otherwise $\sigma(s,t)$ is the unique child of s that lies on the path from s to t in $\hat{G}$.

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

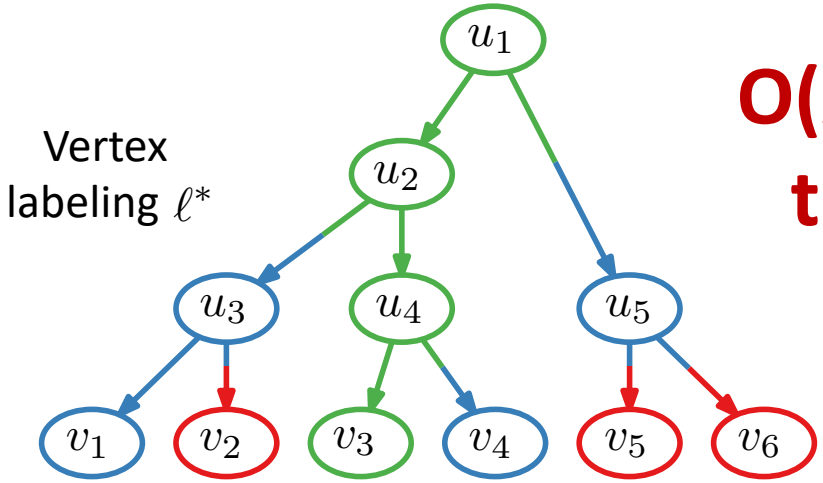
PMH is FPT in number m of locations when $\mathcal{P} = \{S\}$



Phylogenetic tree T



Migration tree $\hat{G}$



Phylogenetic tree T

O(nm^m) time

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

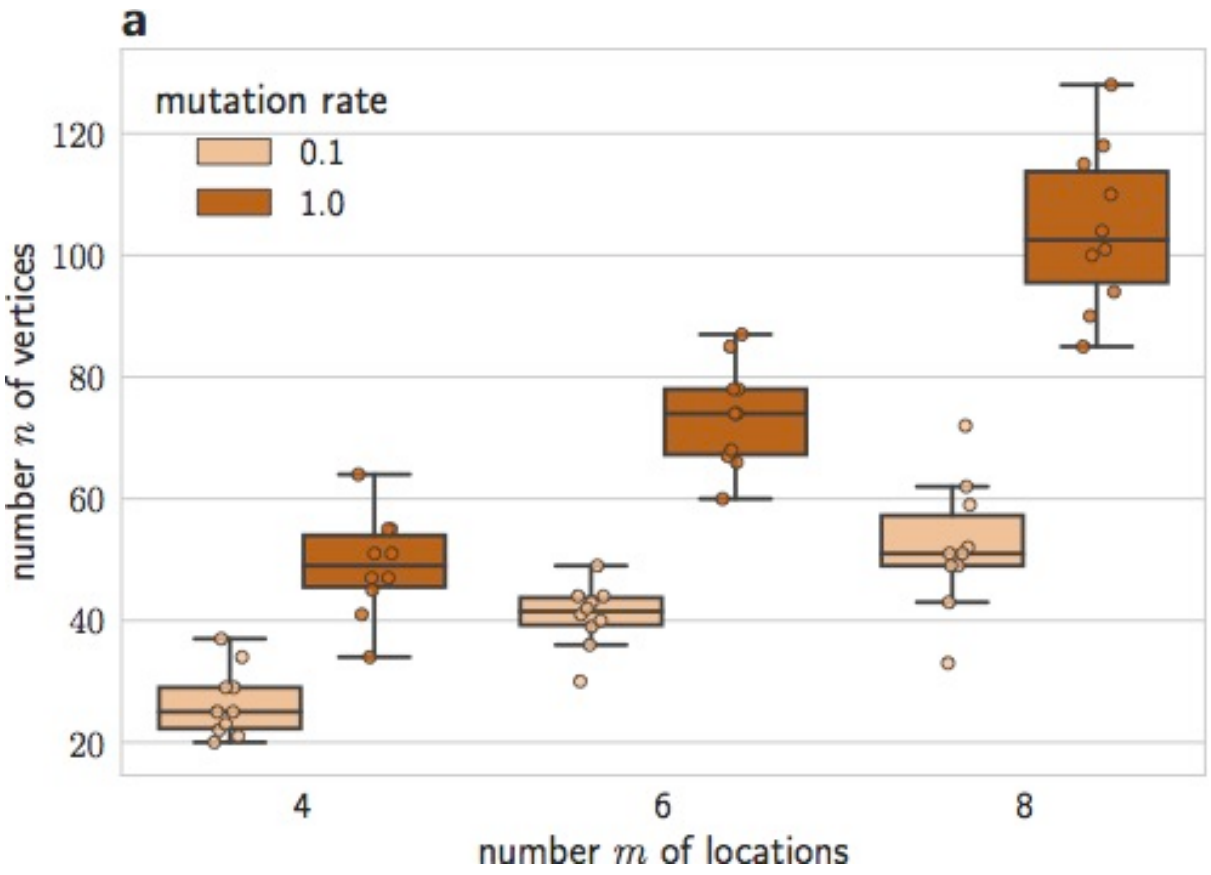
$$d_T(u,v) \geq d_{\hat{G}}(\mathrm{lca}_{\hat{G}}(u), \hat{\ell}(v)) \qquad \forall u, v \in V(T) \text{ such that } u \preceq_T v. \quad (1)$$

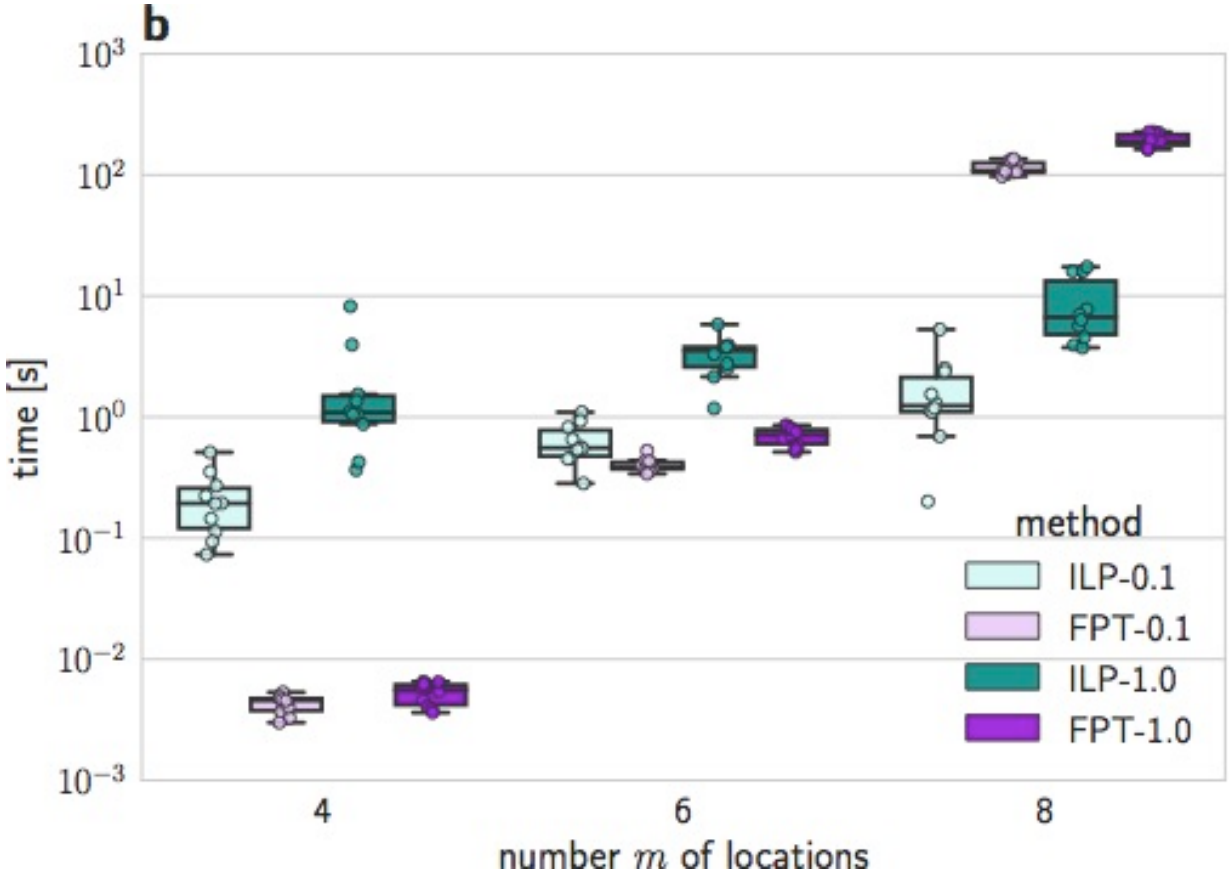
$$\ell^*(v) = \begin{cases} \mathrm{LCA}_{\hat{G}}(r(T)), & \text{if } v = r(T), \\ \sigma(\ell^*(\pi(v)), \mathrm{LCA}_{\hat{G}}(v)), & \text{if } v \neq r(T), \end{cases}$$

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 ℓ^* is a minimum migration labeling consistent with $\hat{G}$.

Simulations





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

Conclusions

- PMH is NP-hard when restricted to a multi-tree
- PMH is FPT in *m* when restricted to a multi-tree
- FPT algorithm is practical for modest number *m* of locations

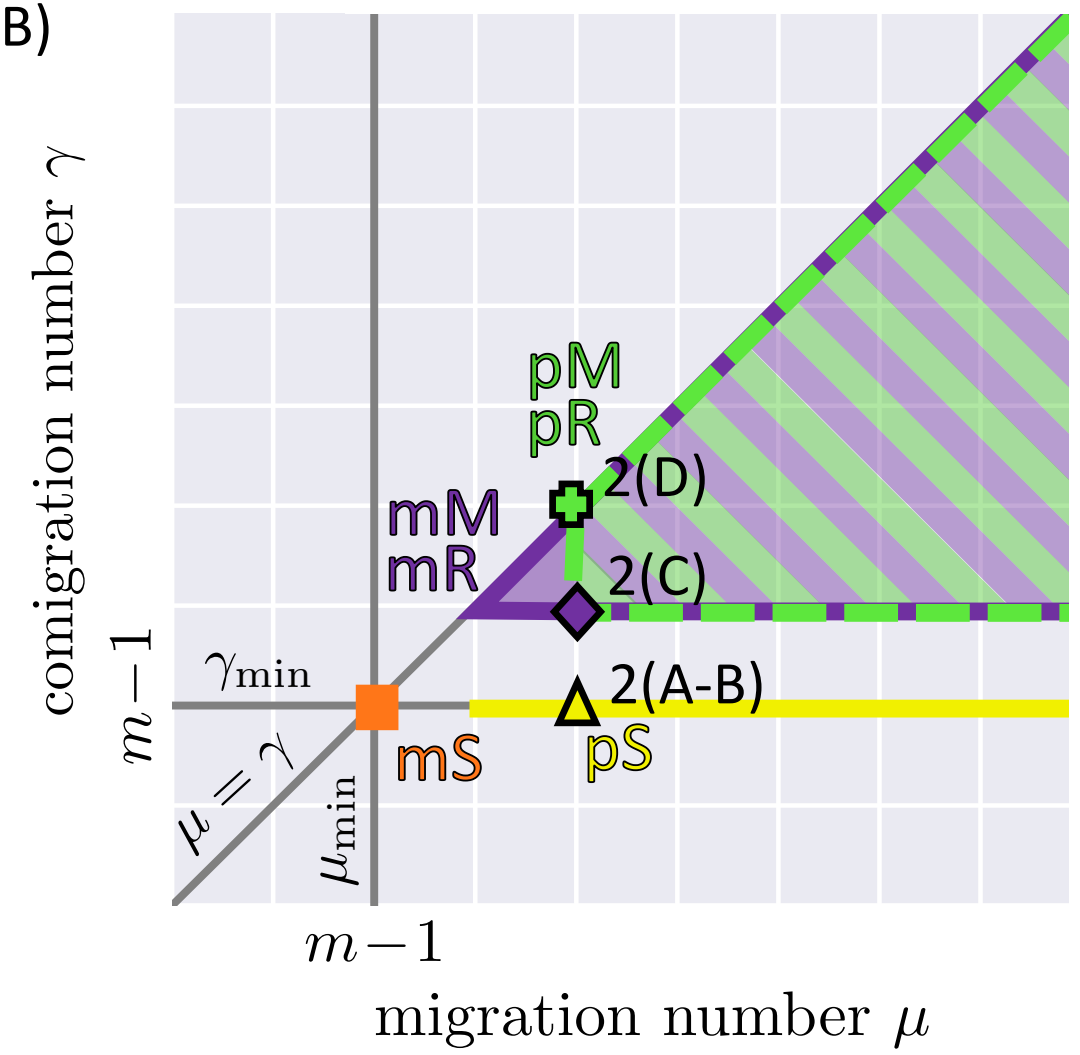
Discussion:

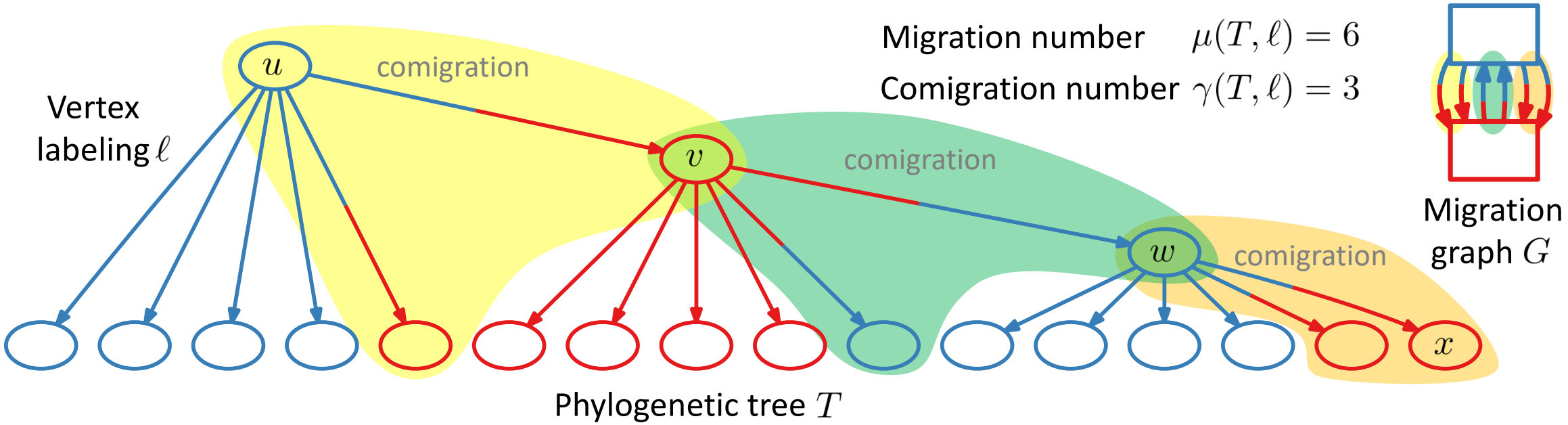
- Applicable to other domains, i.e. comigration of bacterial/viral strains through a single transmission event
- Polytomy resolution version is hard as well
- Open question: Hardness when ***G*** is unrestricted or restricted to a DAG?

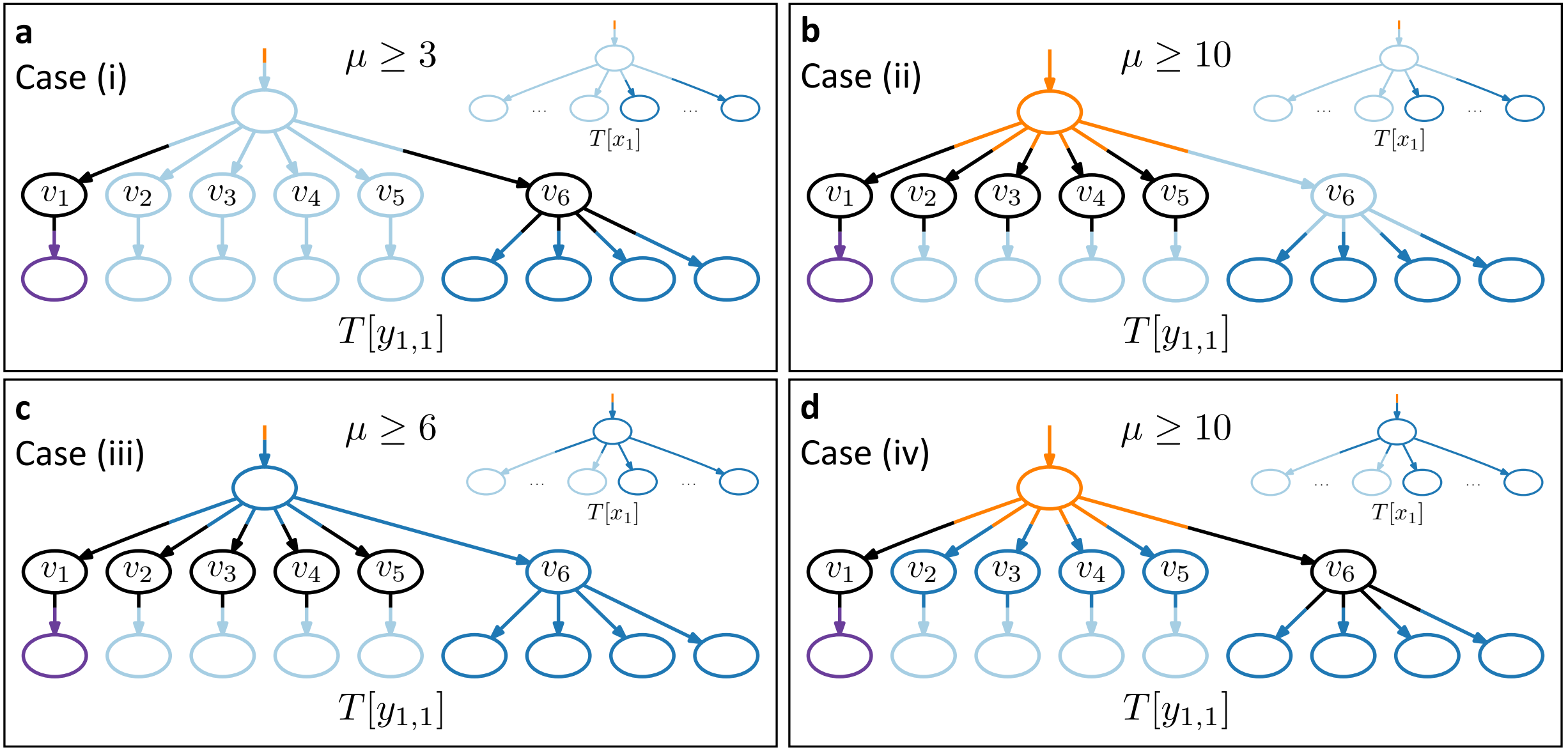
(A)

	single-source seeding (S)	multi-source seeding (M)	reseeding (R)
monoclonal (m)	tree	directed acyclic graph	directed graph
polyclonal (p)	multi-tree	directed acyclic multi-graph	directed multi-graph

(B)




Migration number $\mu(T, \ell) = 6$
Comigration number $\gamma(T, \ell) = 3$
Vertex labeling ℓ
u
comigration
v
comigration
w
comigration
x
Migration graph G
Phylogenetic tree T


a
Case (i)
$\mu \geq 3$
$T[x_1]$
v_1 v_2 v_3 v_4 v_5 v_6
$T[y_{1,1}]$
b
Case (ii)
$\mu \geq 10$
$T[x_1]$
v_1 v_2 v_3 v_4 v_5 v_6
$T[y_{1,1}]$
c
Case (iii)
$\mu \geq 6$
$T[x_1]$
v_1 v_2 v_3 v_4 v_5 v_6
$T[y_{1,1}]$
d
Case (iv)
$\mu \geq 10$
$T[x_1]$
v_1 v_2 v_3 v_4 v_5 v_6
$T[y_{1,1}]$