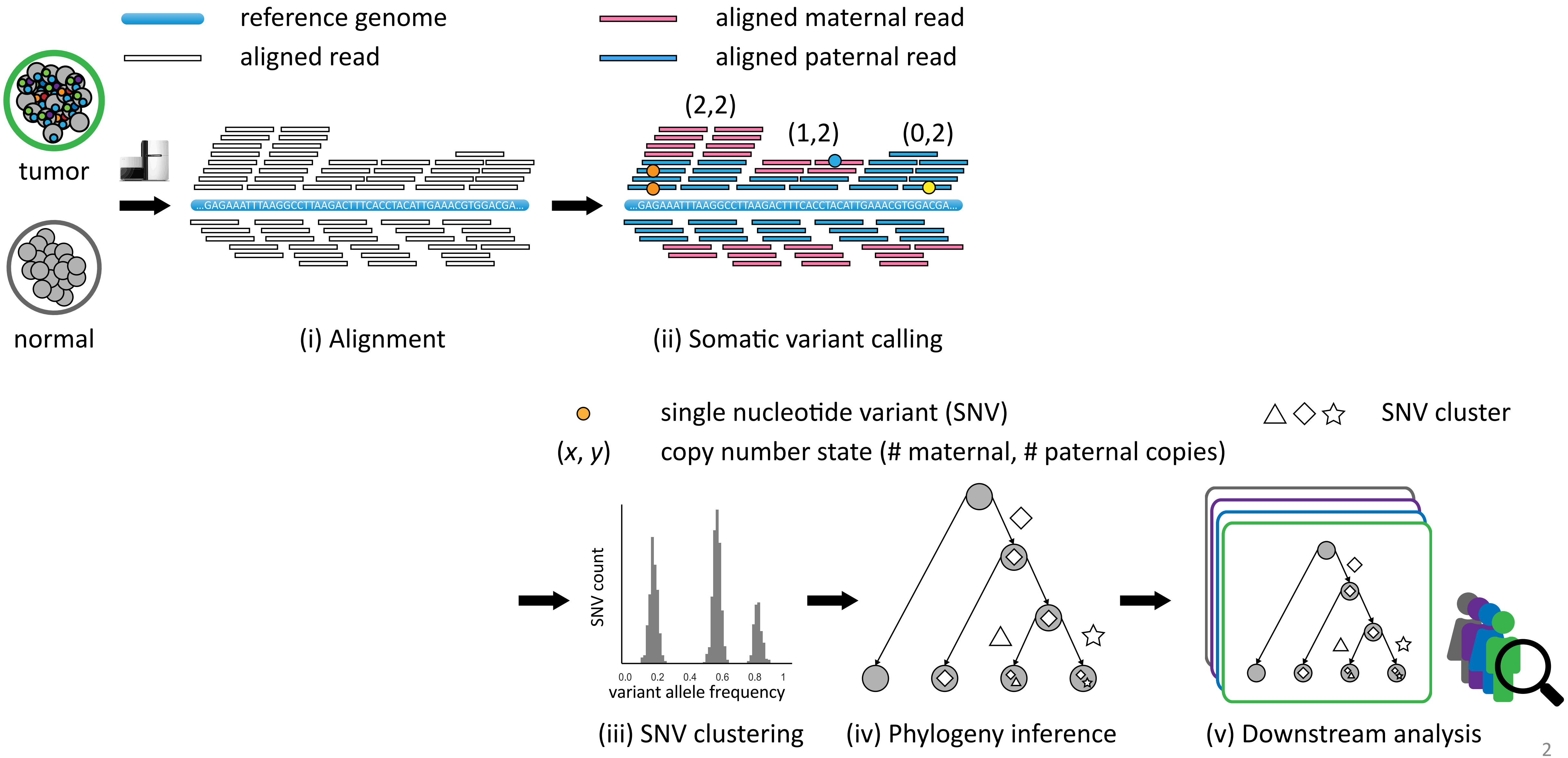
CS 598MEB
Computational Cancer Biology
Lecture 15

Mohammed El-Kebir

March 18, 2021

Cancer Phylogenetics Pipeline



reference genome

aligned read

aligned maternal read

aligned paternal read

(i) Alignment

(ii) Somatic variant calling

● single nucleotide variant (SNV)

(*x*, *y*) copy number state (# maternal, # paternal copies)

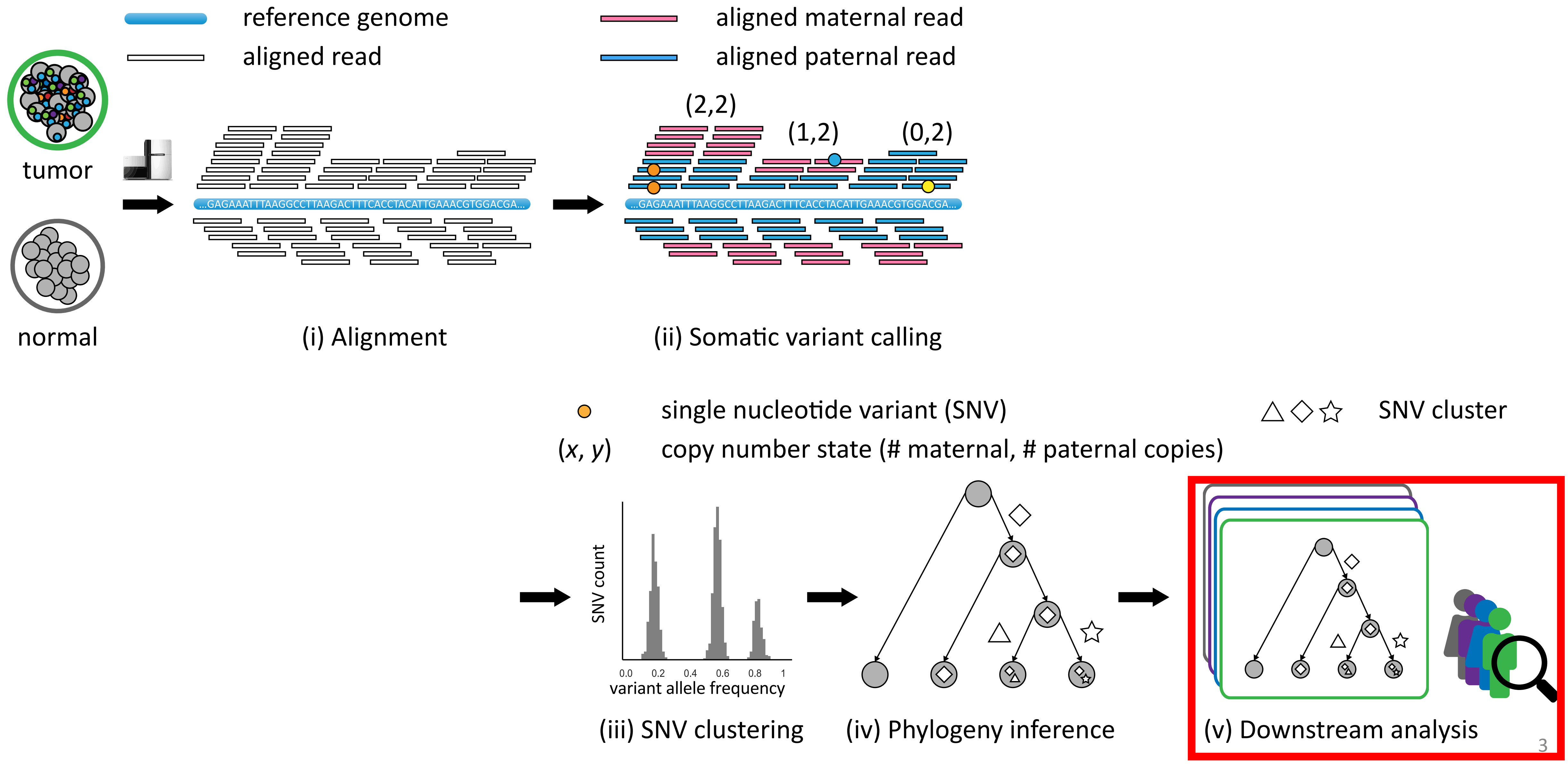
△ ◇ ☆ SNV cluster

(iii) SNV clustering

(iv) Phylogeny inference

(v) Downstream analysis

Cancer Phylogenetics Pipeline



reference genome

aligned read

aligned maternal read

aligned paternal read

(i) Alignment

(ii) Somatic variant calling

● single nucleotide variant (SNV)

(*x*, *y*) copy number state (# maternal, # paternal copies)

△ ◇ ☆ SNV cluster

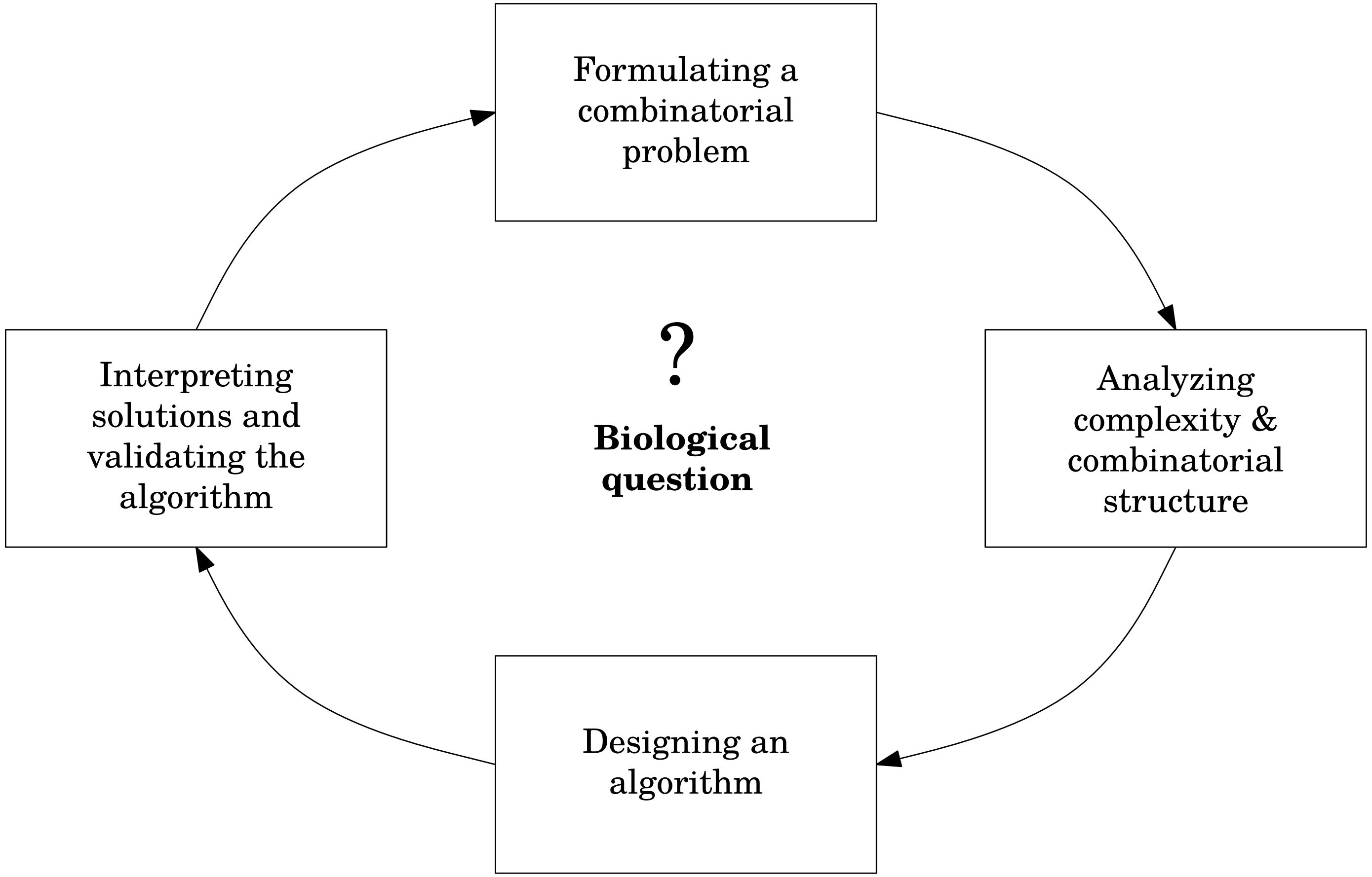
(iii) SNV clustering

(iv) Phylogeny inference

(v) Downstream analysis

Outline

- Metastasis
- Maximum parsimony
- Problem statement
- Complexity
- Algorithm & results
- Problem variants



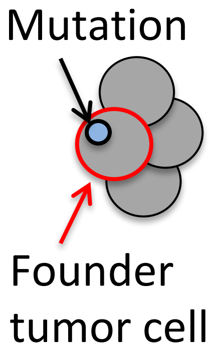
Reading:

- **M. El-Kebir**, G. Satas and B.J. Raphael. Inferring parsimonious migration histories for metastatic cancers. [Nature Genetics, 50:718-726, 2018]().
- **M. El-Kebir**†. [Parsimonious Migration History Problem: Complexity and Algorithms](). WABI 2018, Helsinki, Finland, August 20-22, 2018.

Tumorigenesis: (i) Cell Mutation

Clonal Theory of Cancer

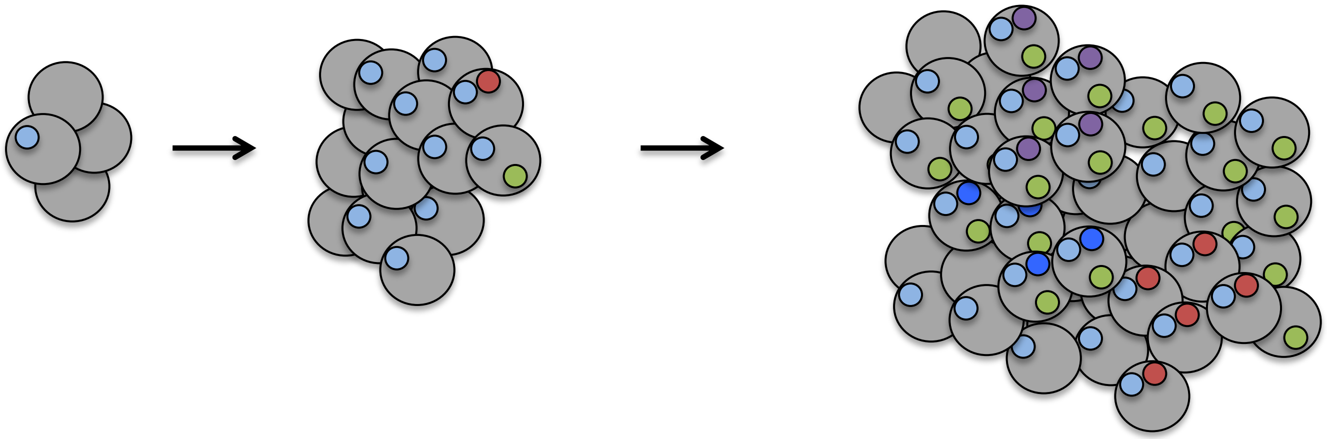
[Nowell, 1976]



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

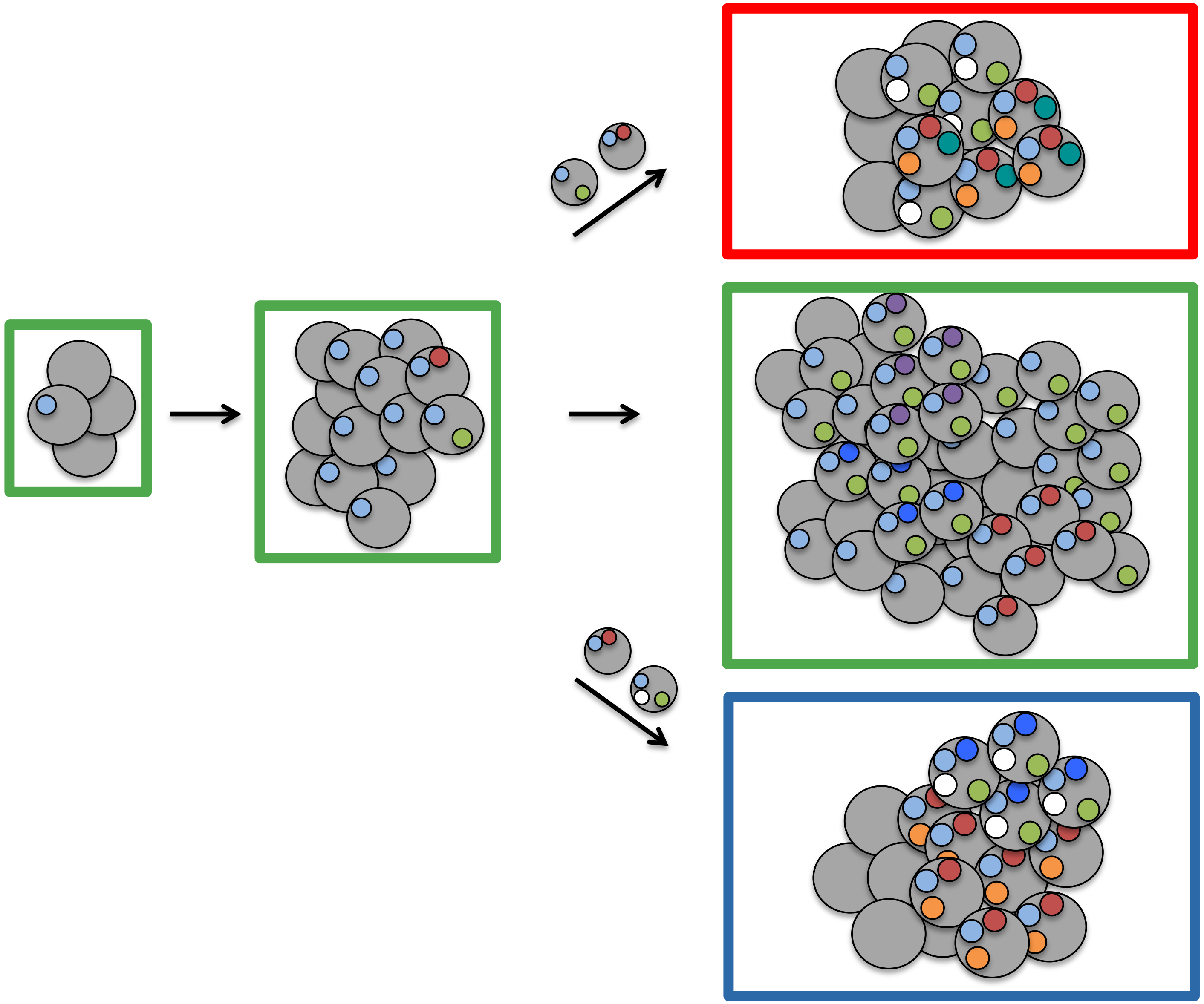
Clonal Theory of Cancer

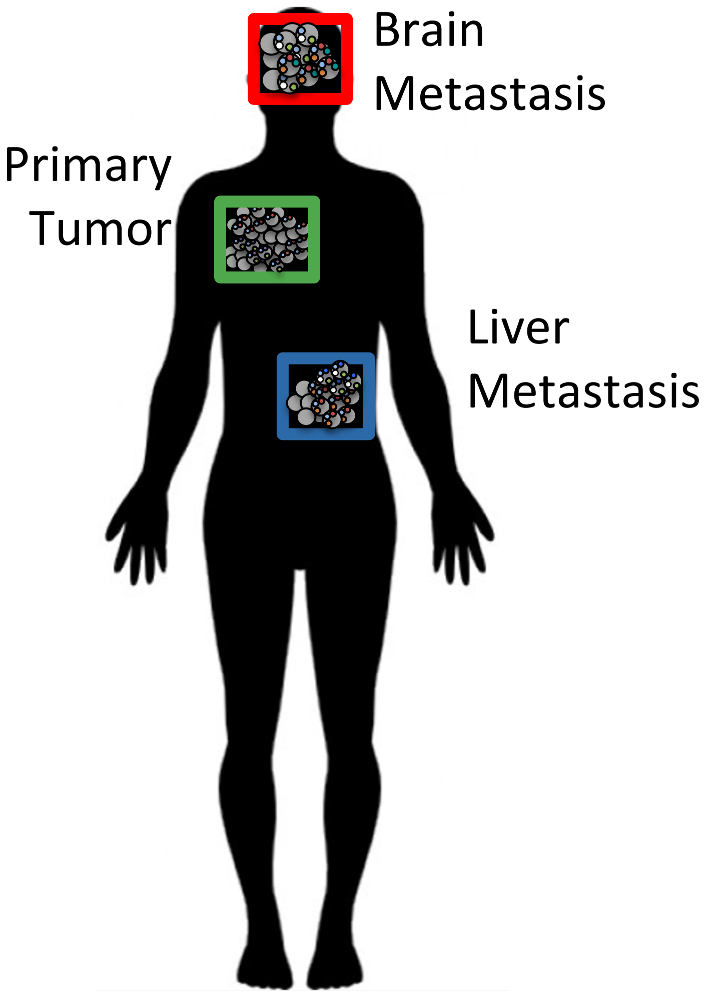
[Nowell, 1976]



Heterogeneous Tumor

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



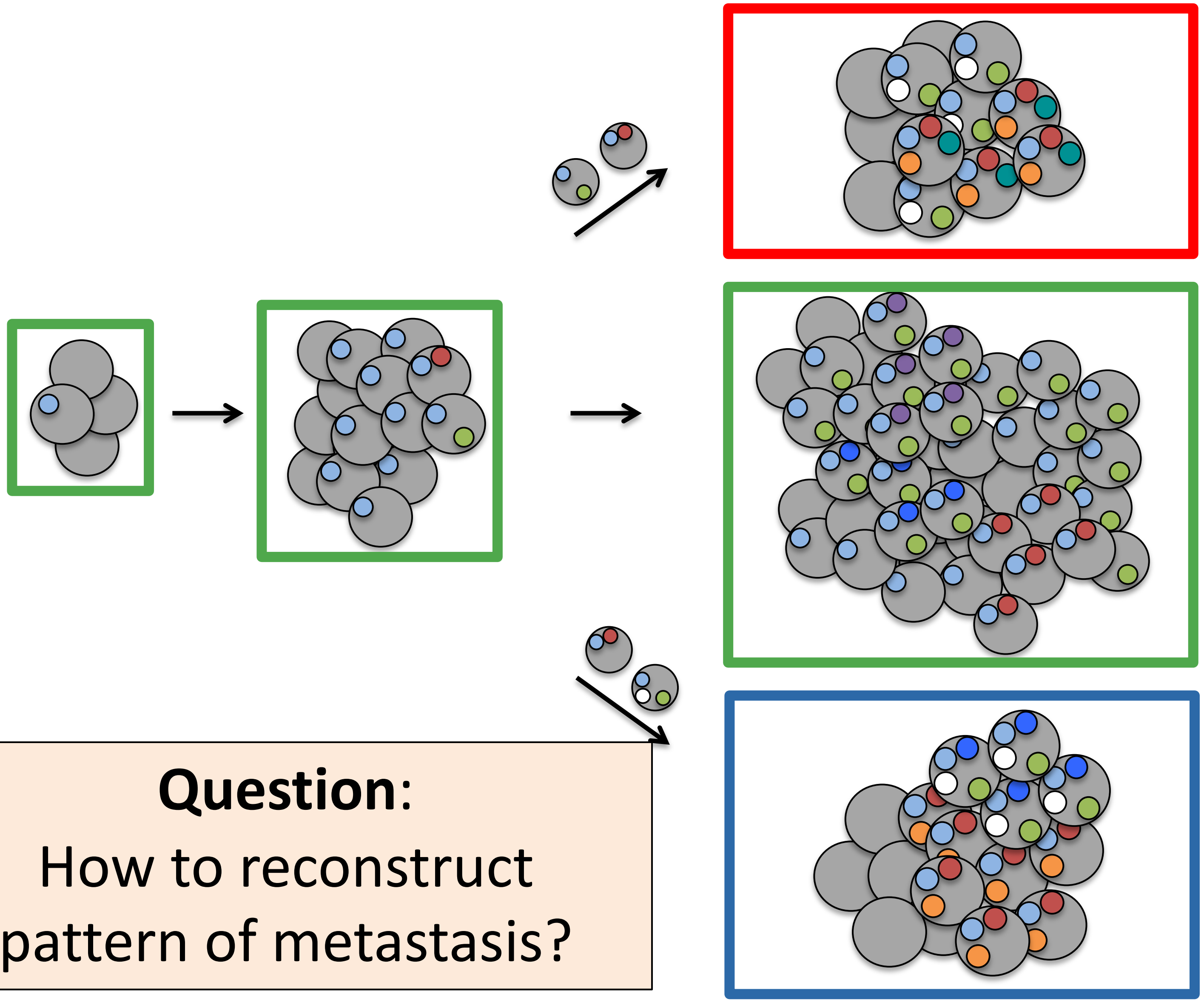


Brain Metastasis

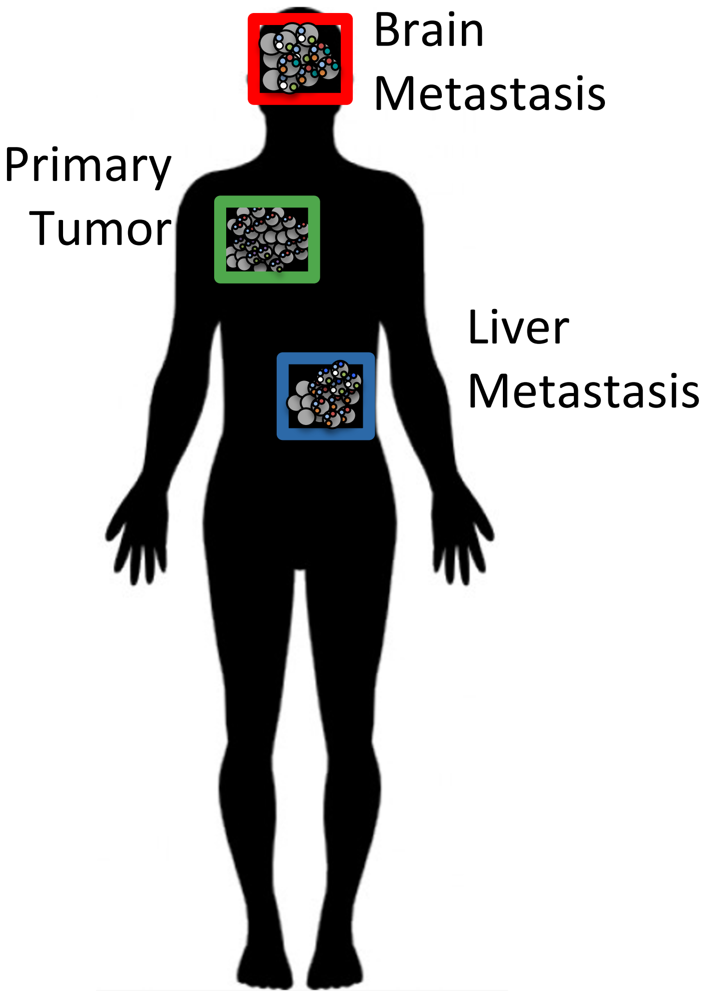
Primary Tumor

Liver Metastasis

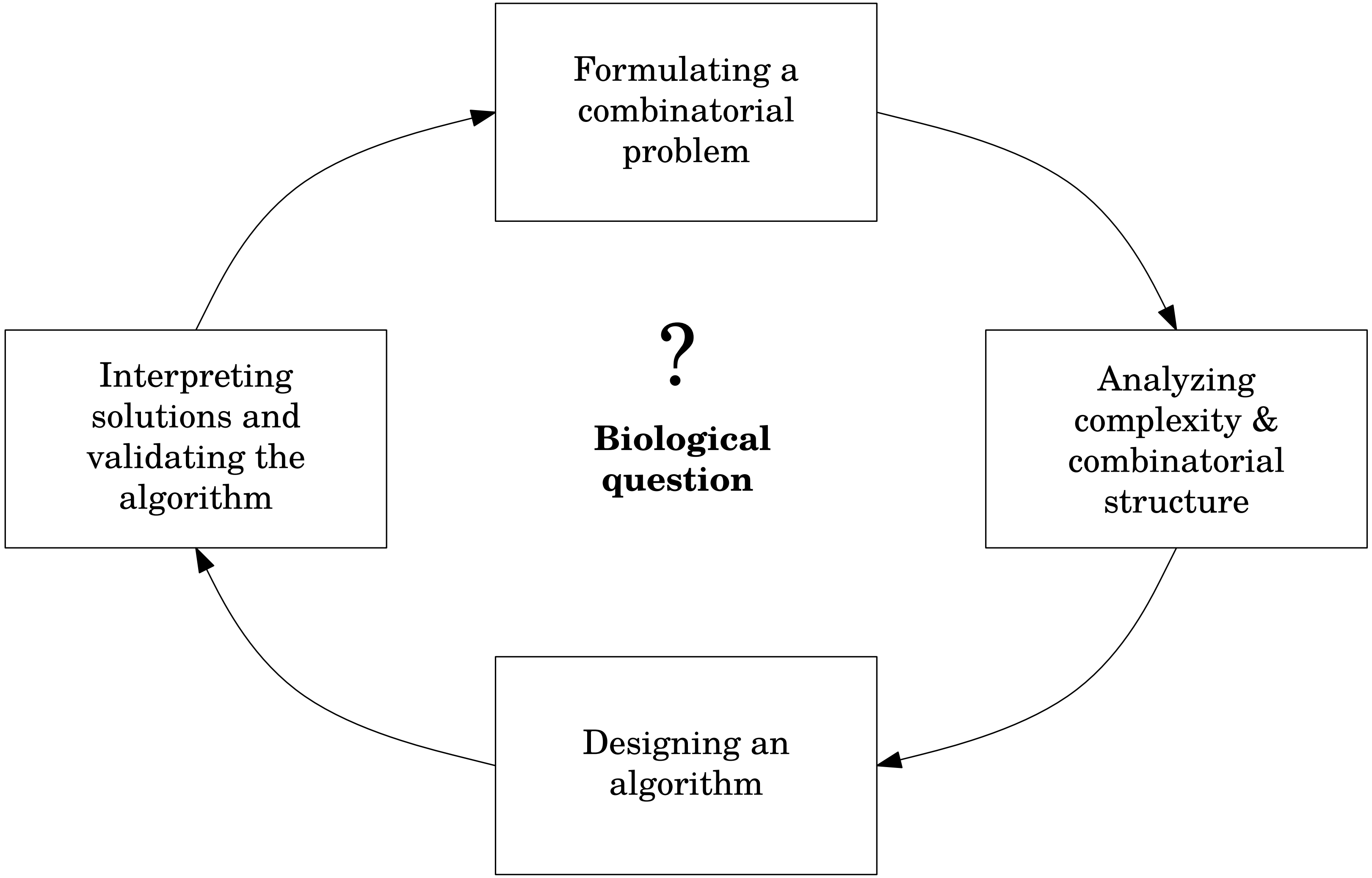
Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration



Question:
How to reconstruct
pattern of metastasis?



Key Challenge in Computational Biology



Formulating a combinatorial problem

Analyzing complexity & combinatorial structure

Designing an algorithm

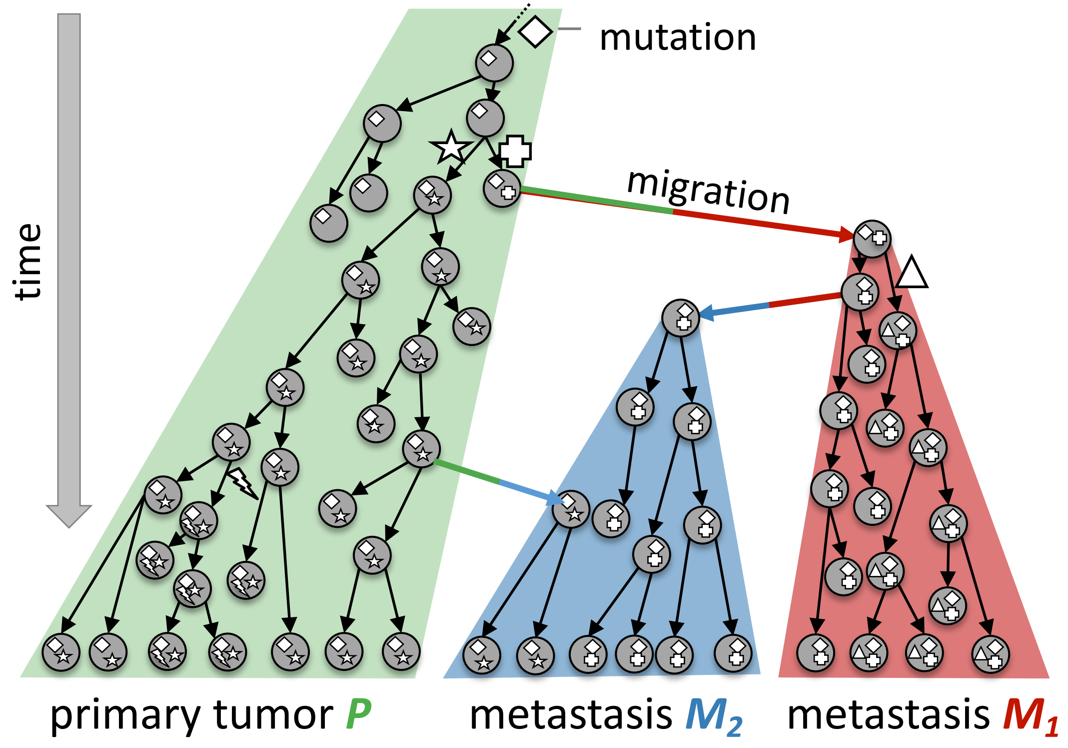
Interpreting solutions and validating the algorithm

?

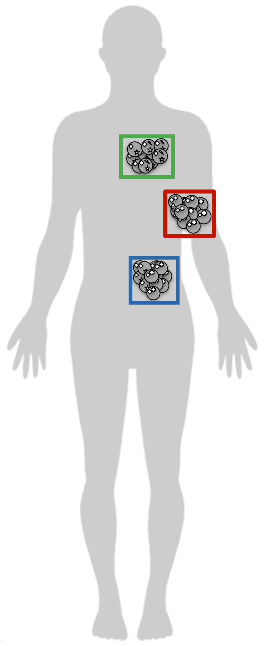
Biological question

Translating a biological problem into computer science

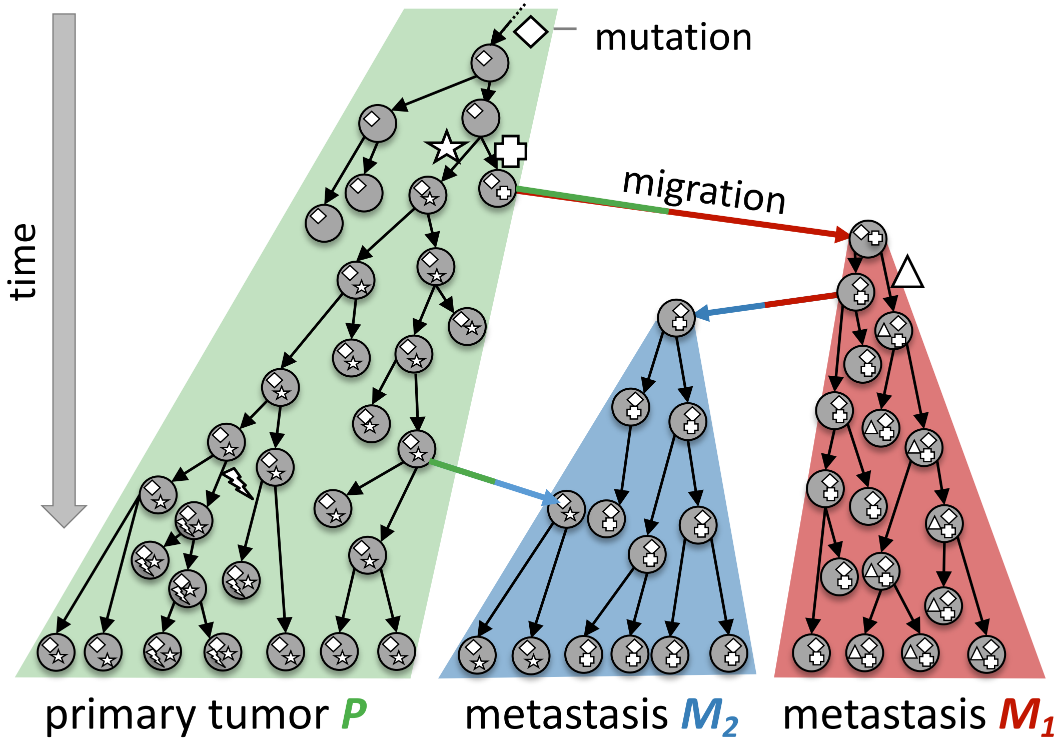
Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration



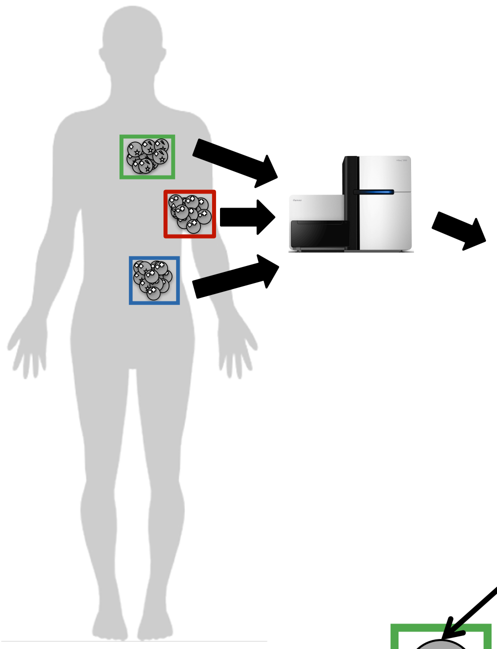
Cell Tree

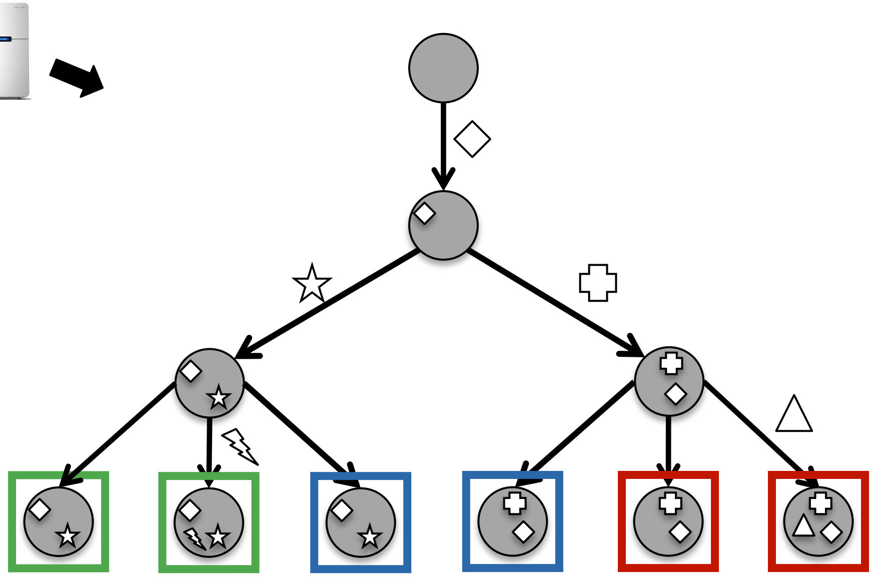


Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration



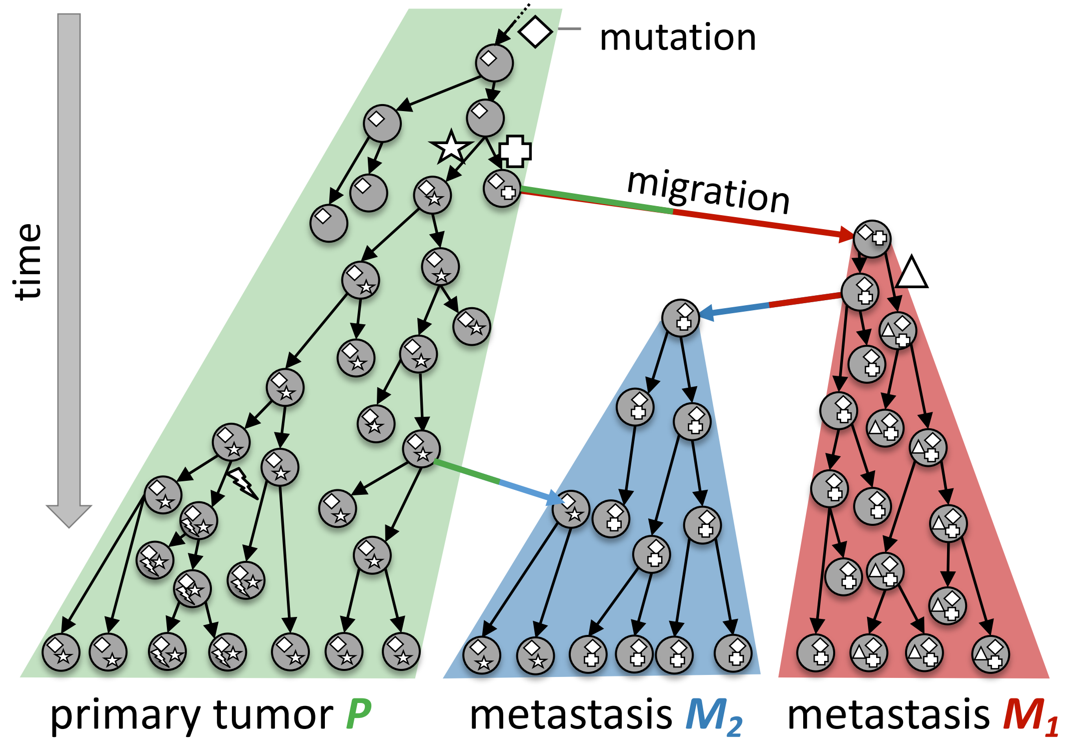
Cell Tree



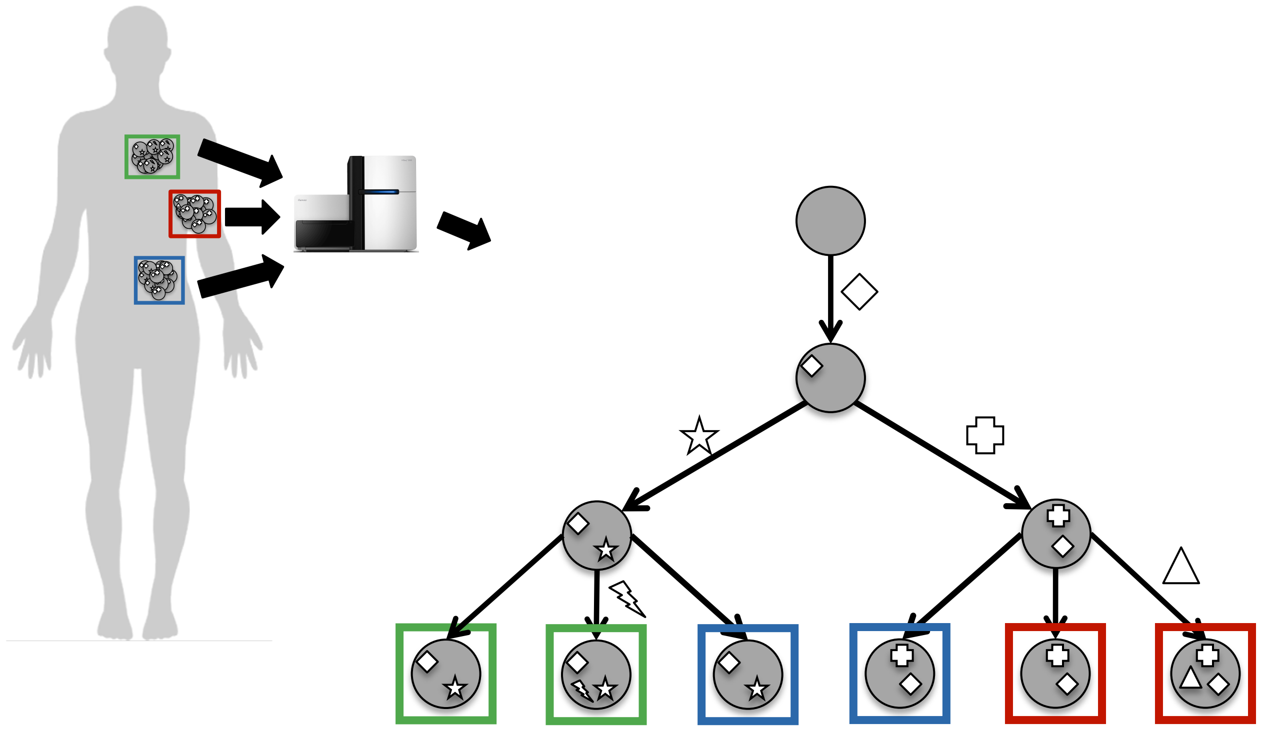


Phylogenetic Tree *T*

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



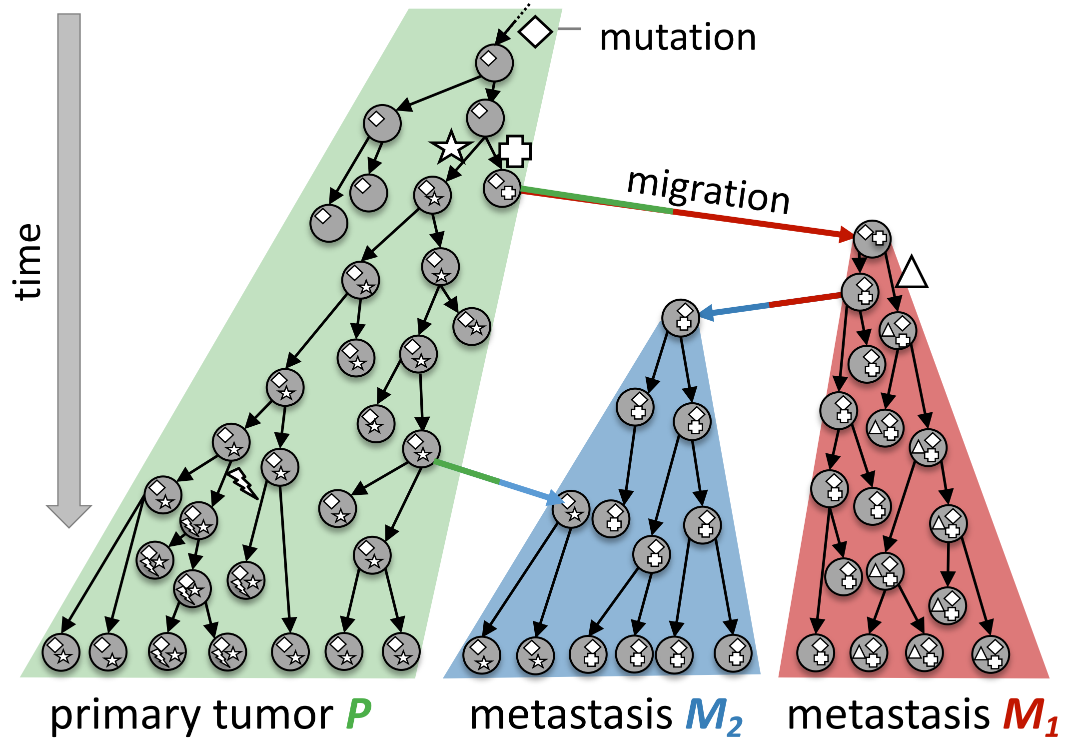
Cell Tree



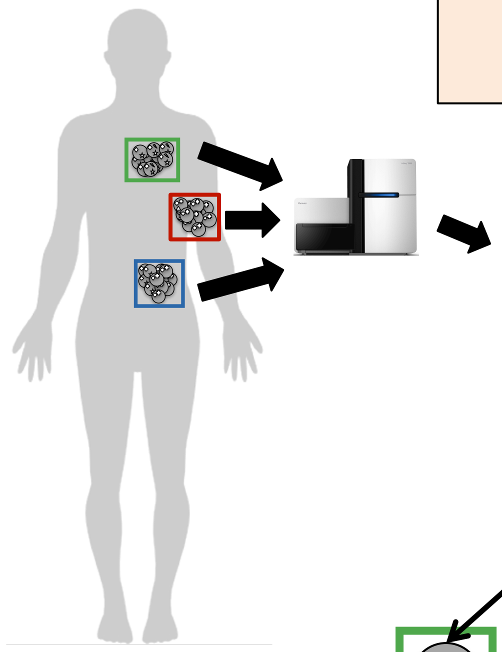
Phylogenetic Tree ***T***

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

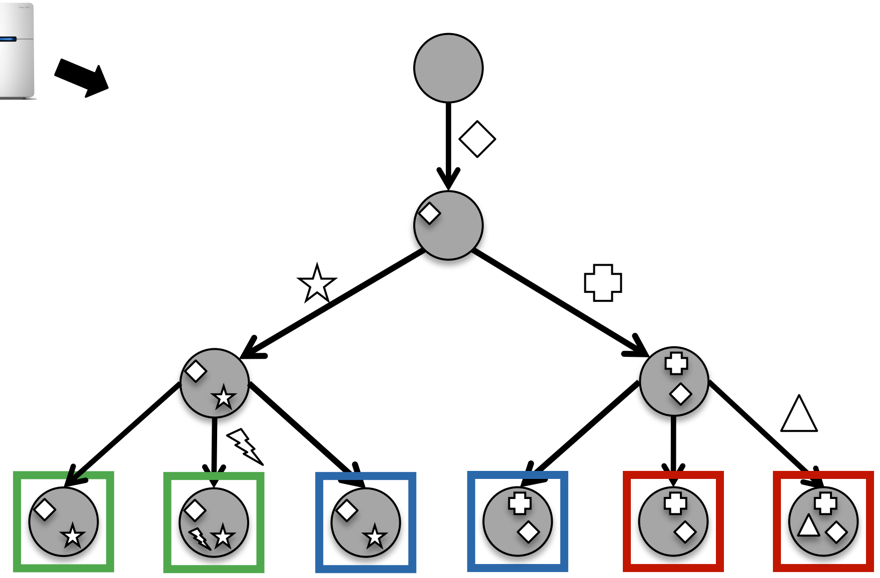
Tumorigenesis: (i) Cell Division, (ii) Mutation & (iii) Migration



Cell Tree



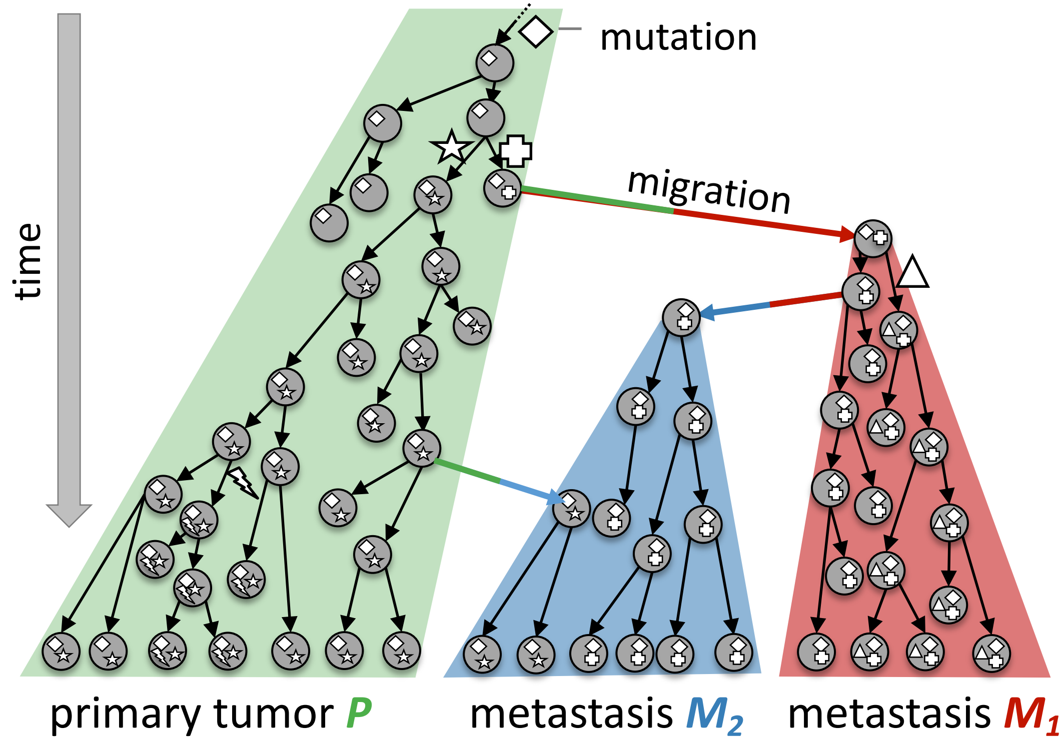
Question: Find parsimonious vertex labeling



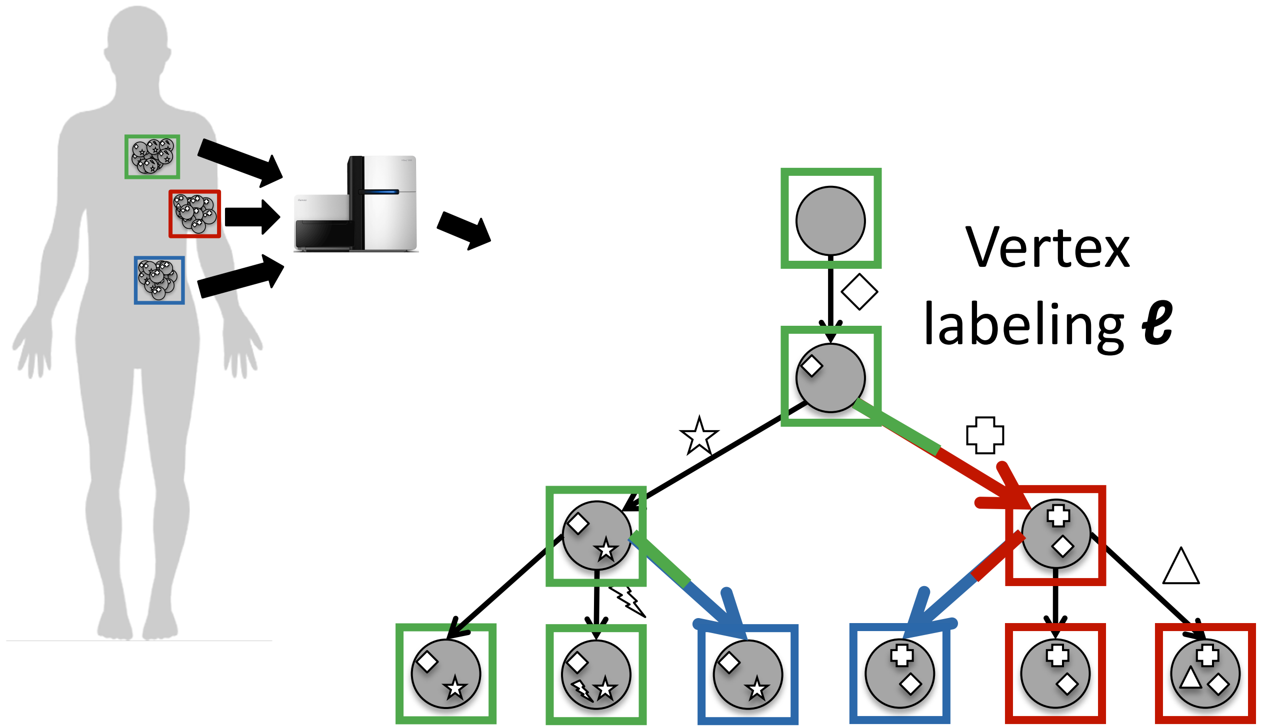
Phylogenetic Tree ***T***

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

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



Cell Tree

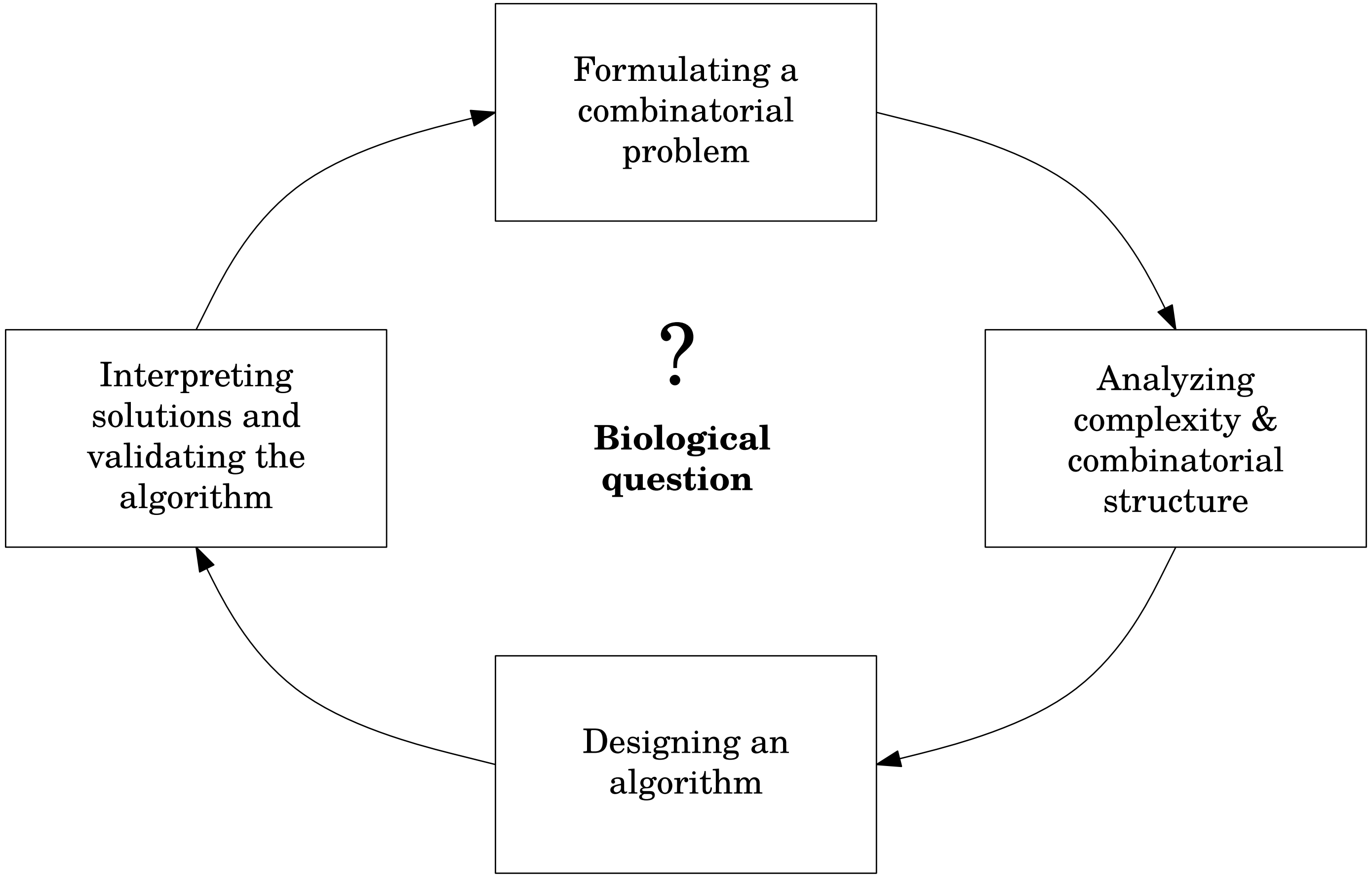


Phylogenetic Tree ***T***

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

Outline

- Metastasis
- Maximum parsimony
- Problem statement
- Complexity
- Algorithm & results
- Problem variants


Formulating a combinatorial problem

Analyzing complexity & combinatorial structure

Designing an algorithm

Interpreting solutions and validating the algorithm

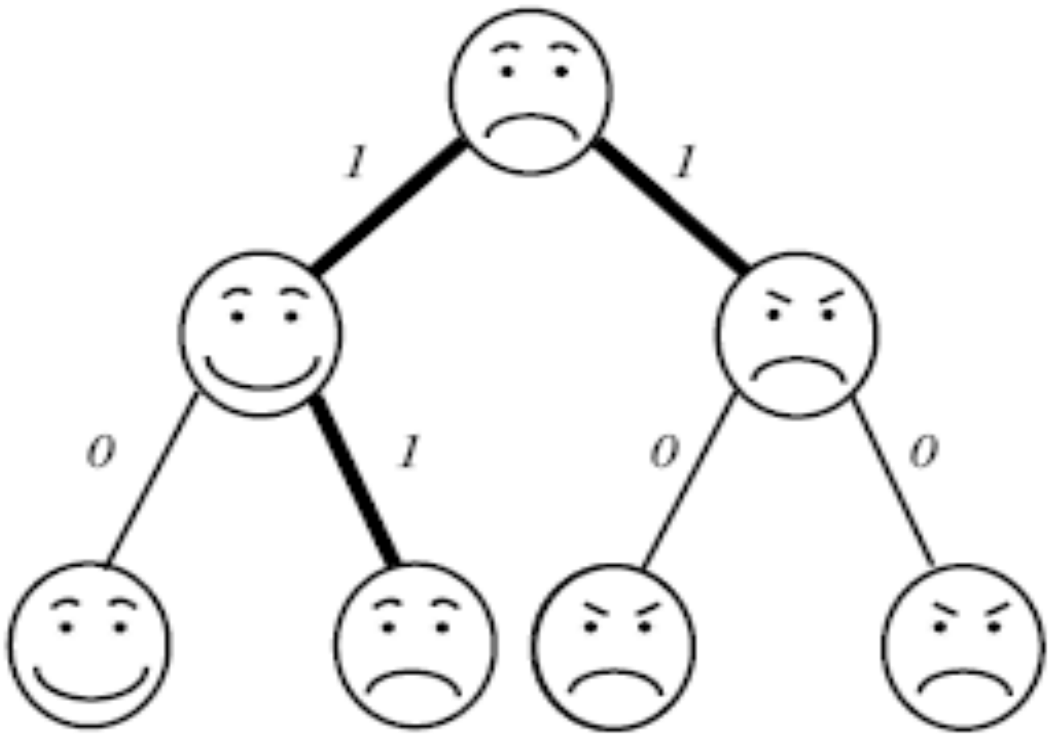
?

Biological question

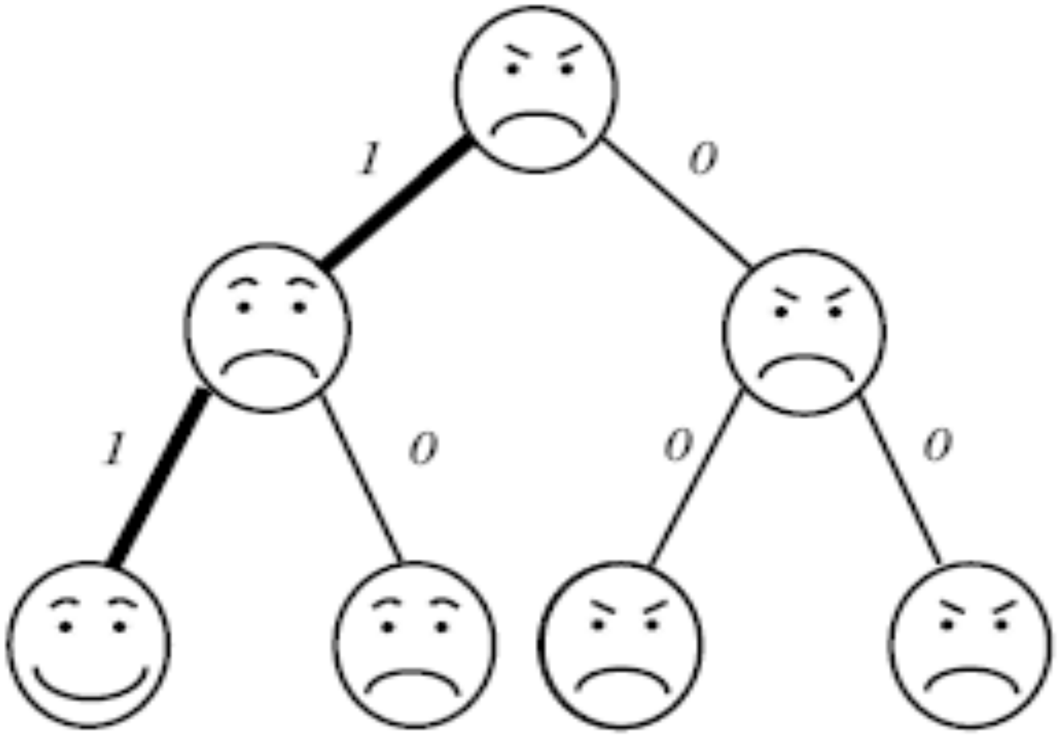
Reading:

- **M. El-Kebir**, G. Satas and B.J. Raphael. Inferring parsimonious migration histories for metastatic cancers. Nature Genetics, 50:718-726, 2018.
- **M. El-Kebir**[†]. Parsimonious Migration History Problem: Complexity and Algorithms. WABI 2018, Helsinki, Finland, August 20-22, 2018.

Character-Based Phylogeny Reconstruction: Criterion



(a) *Parsimony Score=3*



(b) *Parsimony Score=2*

Parsimony: minimize number of changes on edges of tree

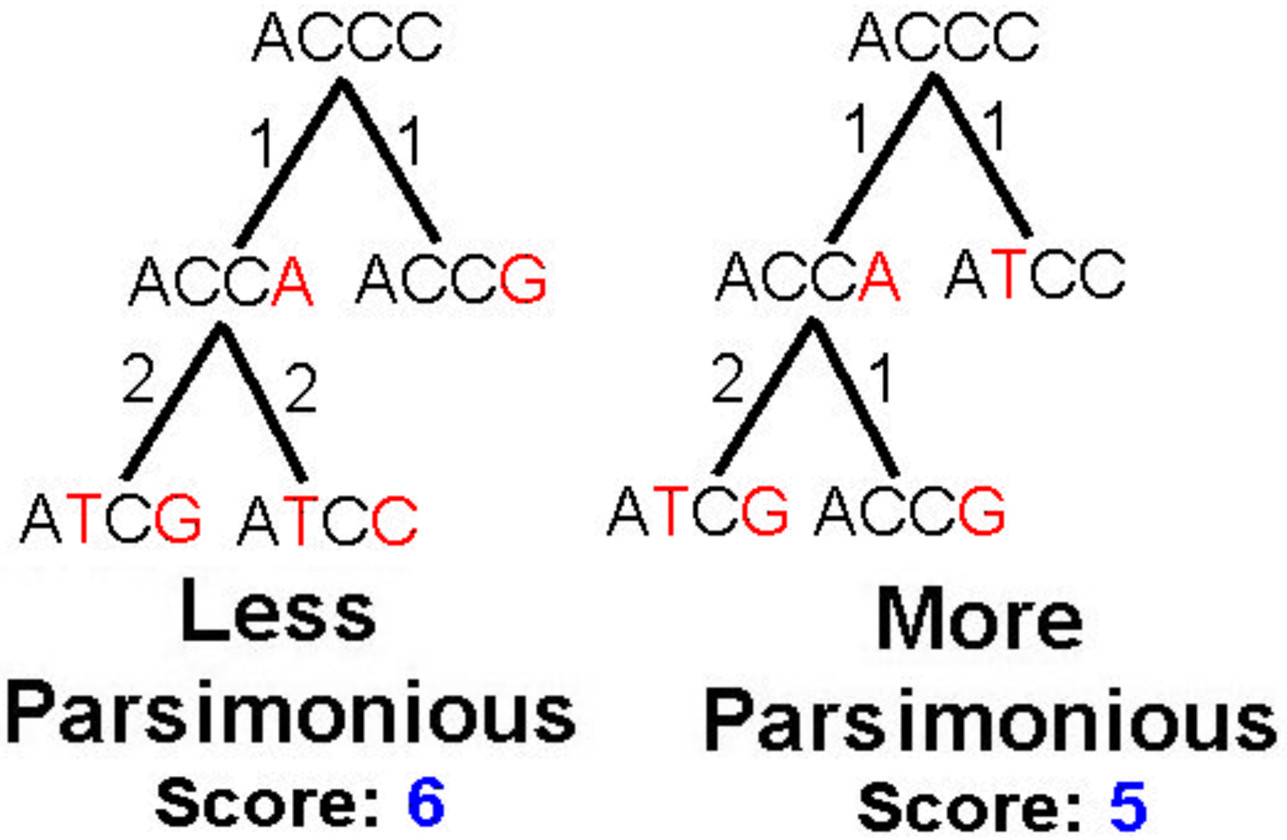
A Small and a Large Problem

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



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

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.

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.

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

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.

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

Filling out DP Table and Traceback

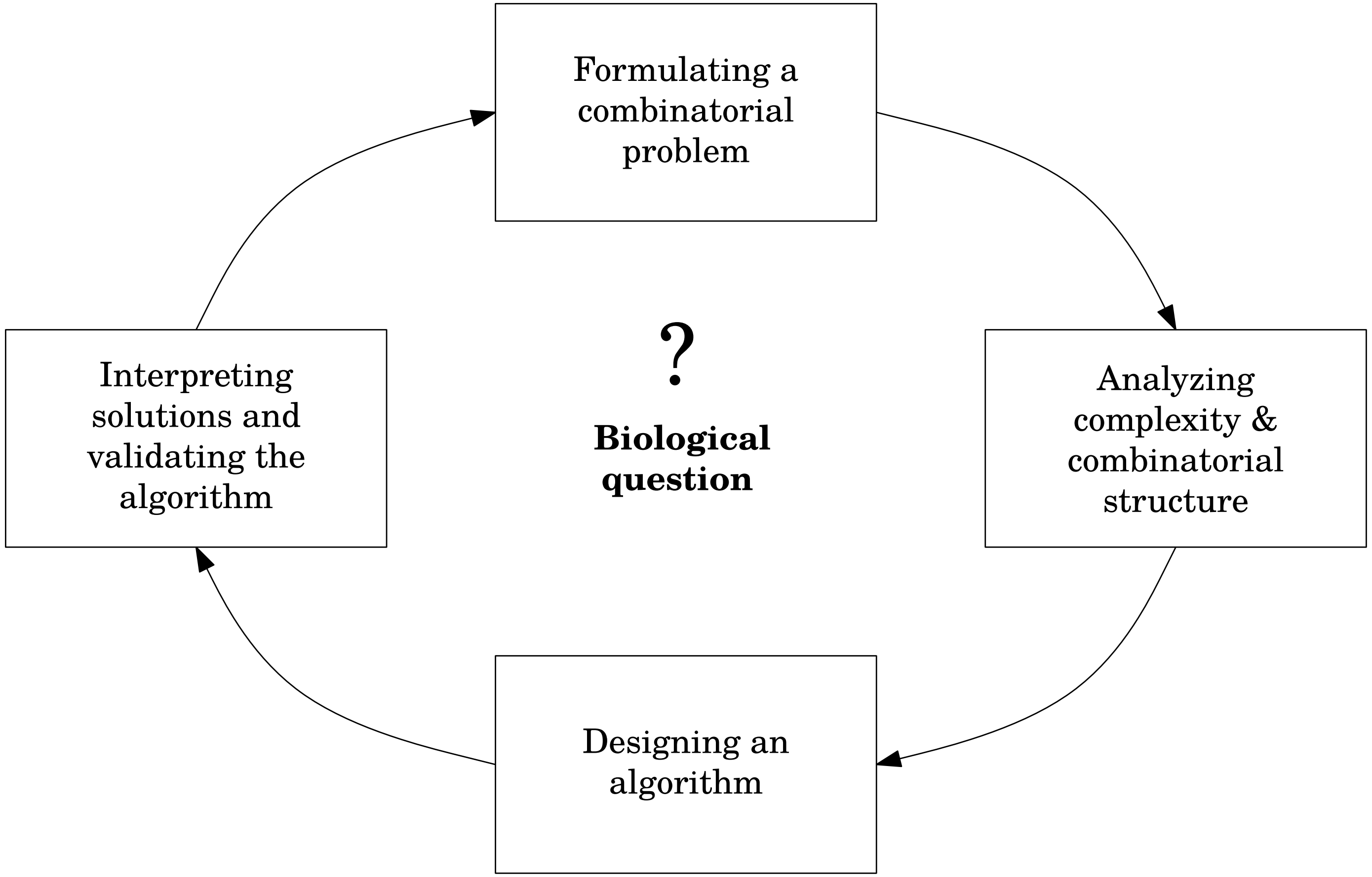
```
Filling out $\mu$        $Fill(T, r(T), \sigma, \Sigma)$

$Fill(T, v, \sigma, \Sigma)$        $O(m|\Sigma|^2)$
if $v \in L(T)$ then
    For $s \in \Sigma$
        if $s = \sigma(v)$ then
            $\mu(v,s) = 0$
        else
            $\mu(v,s) = \infty$
else
    For $w \in \delta(v)$        // children
        $Fill(T, w, \sigma, \Sigma)$
    $\mu(v,s) = 0$
    For $w \in \delta(v)$
        $\mu(v,s) \mathrel{+}= \min_{t \in \Sigma} \{c(s,t) + \mu(w,t)\}$
```

```
Backtrace$(T, v, \mu)$
if $v = r(T)$
    $\sigma(r(T)) = \arg\min_{s \in \Sigma} \{\mu(r(T), s)\}$
else Let $u$ be the parent of $v$ and let $s = \sigma(u)$ be the state of $u$
    $\sigma(v) = \arg\min_{t \in \Sigma} \{c(s,t) + \mu(v,t)\}$
For $w \in \delta(v)$
    Backtrace$(T, w, \mu)$
```

Let $r(T)$ be the root vertex

Outline

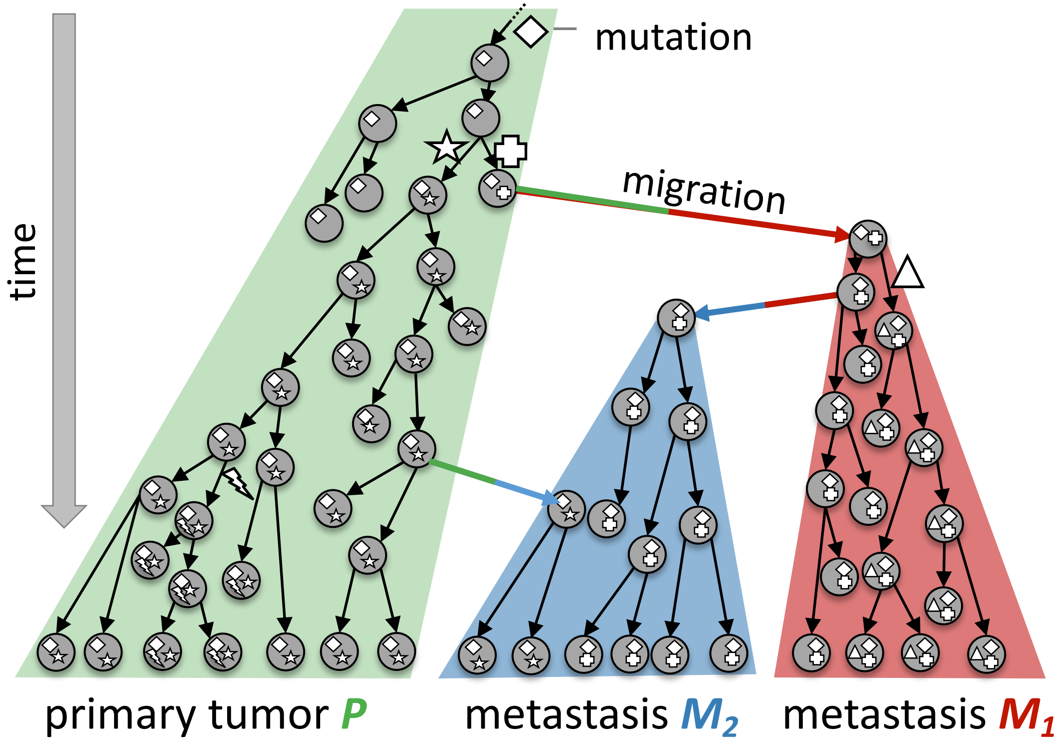
- Metastasis
- Maximum parsimony
- Problem statement
- Complexity
- Algorithm & results
- Problem variants



Reading:

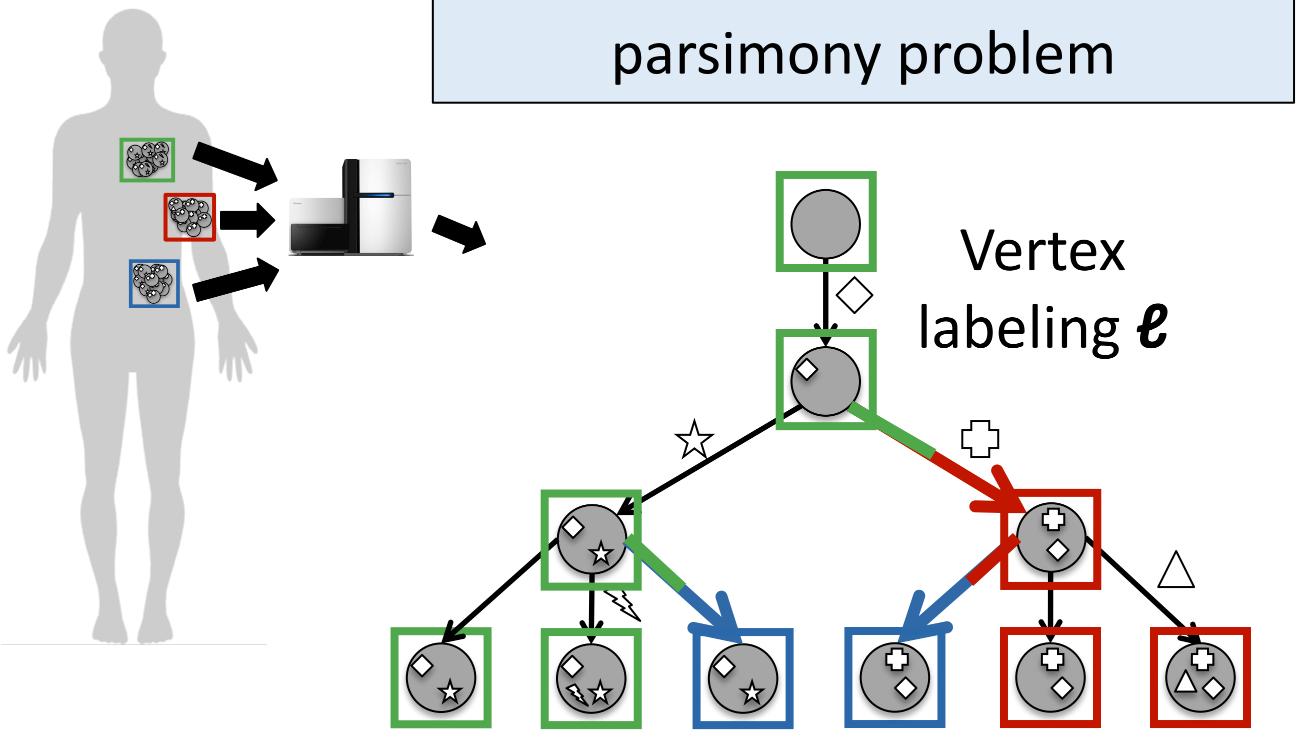
- **M. El-Kebir**, G. Satas and B.J. Raphael. Inferring parsimonious migration histories for metastatic cancers. Nature Genetics, 50:718-726, 2018.
- **M. El-Kebir**†. Parsimonious Migration History Problem: Complexity and Algorithms. WABI 2018, Helsinki, Finland, August 20-22, 2018.

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



Cell Tree

Instance of small maximum parsimony problem



Phylogenetic Tree ***T***

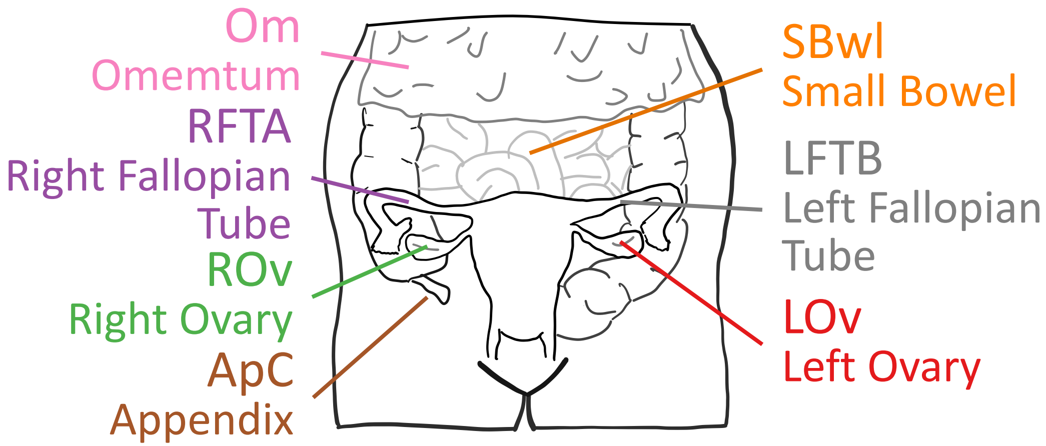
Goal: Given phylogenetic tree ***T***, find ***parsimonious*** vertex labeling ***ℓ*** with fewest migrations

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

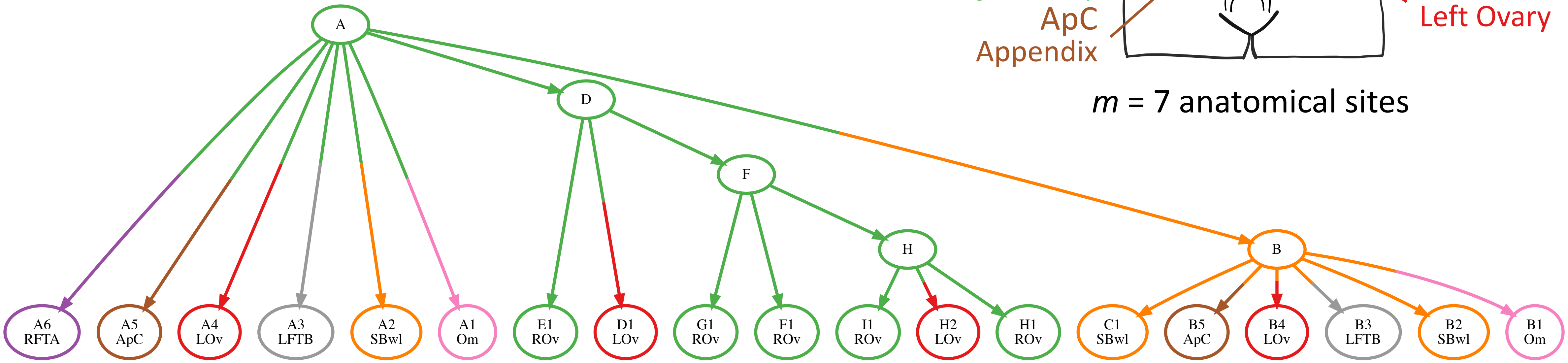
Minimum Migration Analysis in Ovarian Cancer

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

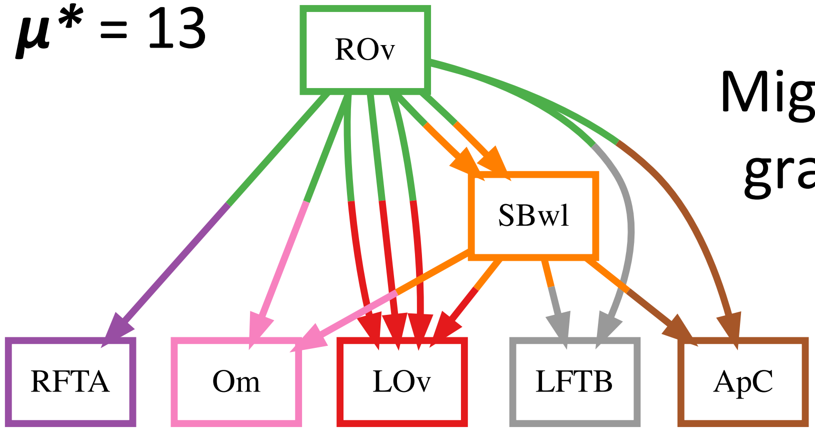


Minimum Migration Analysis in Ovarian Cancer

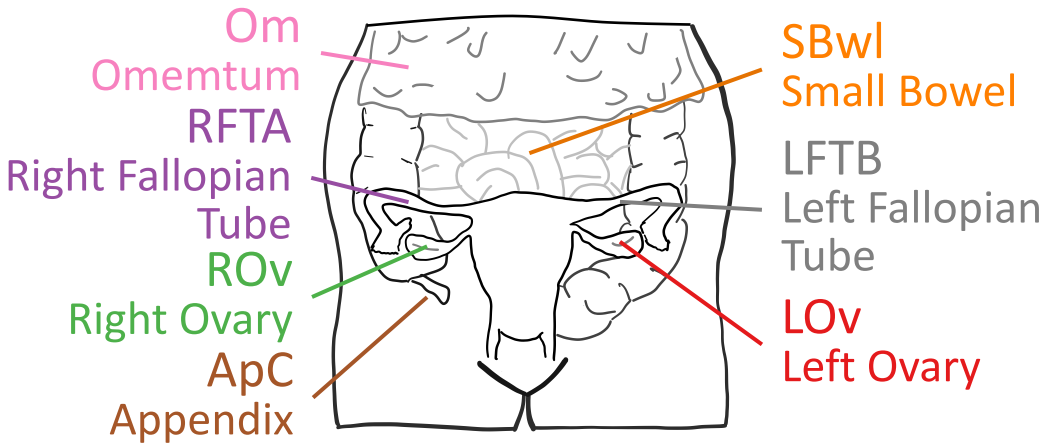
McPherson et al. (2016). Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics*.

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

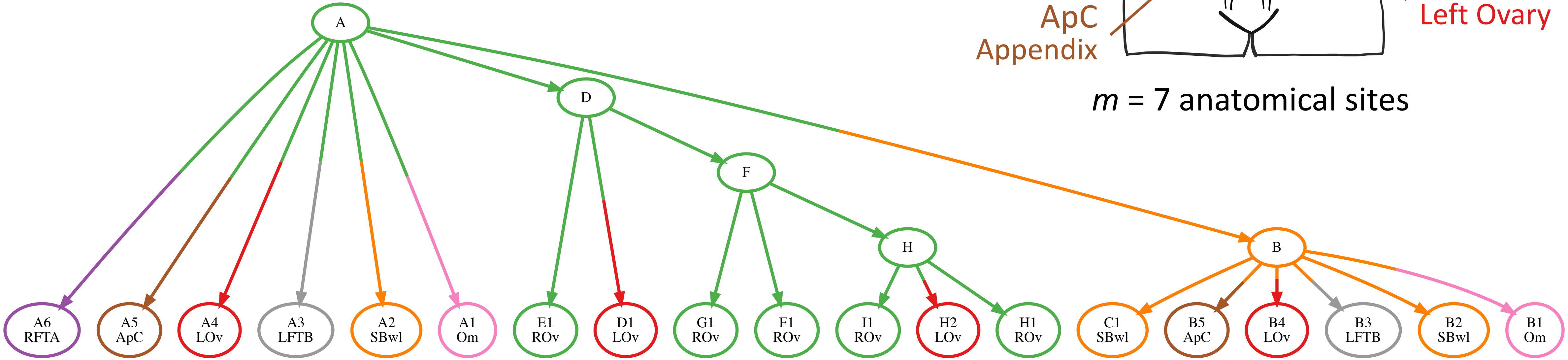
$\mu^* = 13$



Migration graph **G**



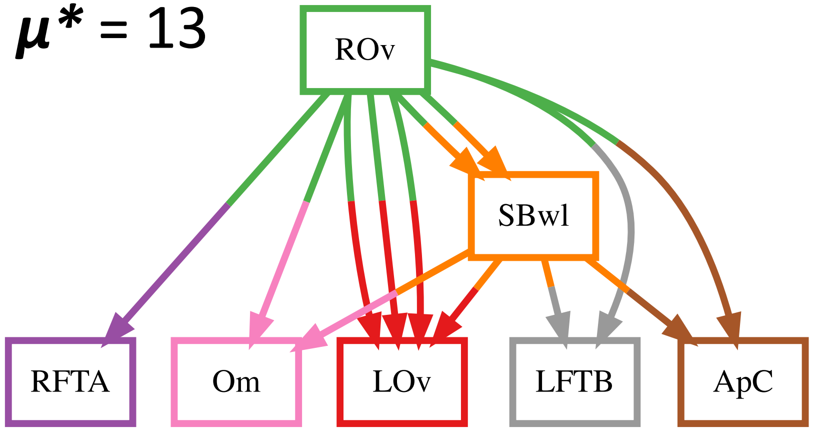
m = 7 anatomical sites



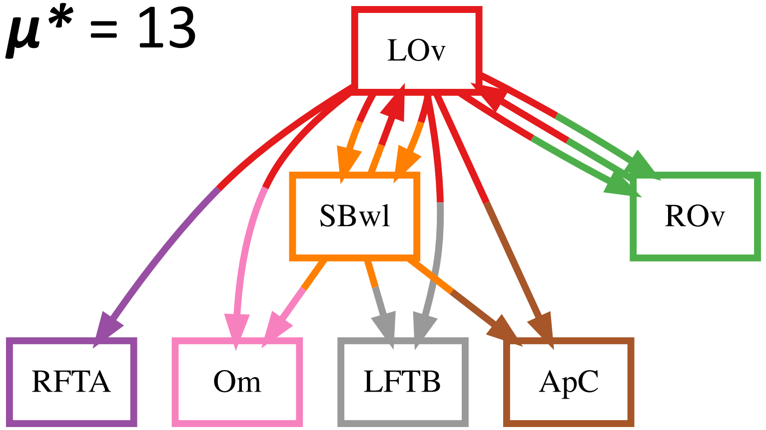
Minimum Migration History is *Not* Unique

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

***μ^** = 13**

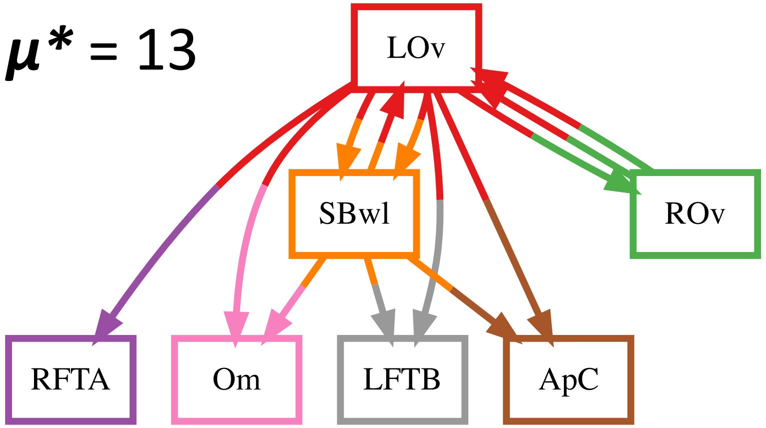


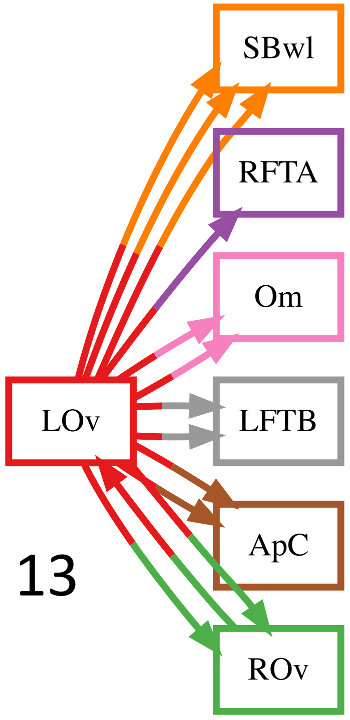
***μ^** = 13**



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

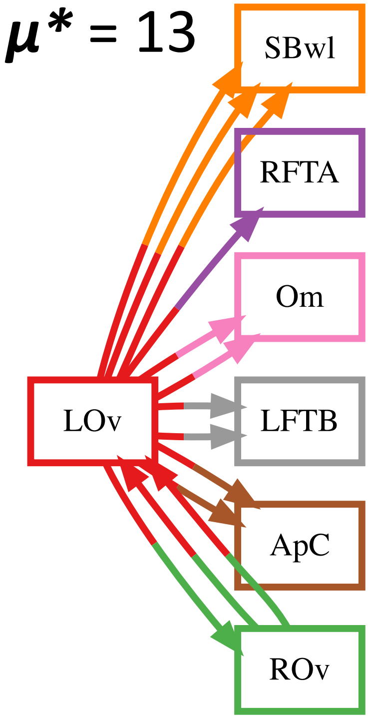
***μ^** = 13**





***μ^** = 13**

***μ^** = 13**



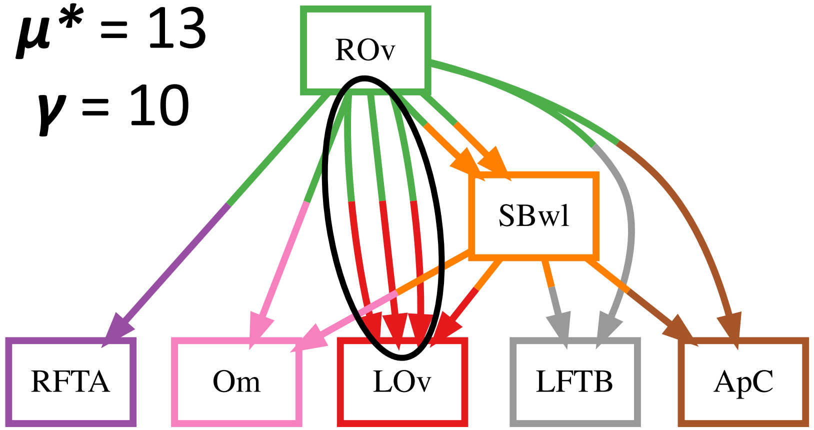
Comigrations: Simultaneous Migrations of Multiple Clones

- 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

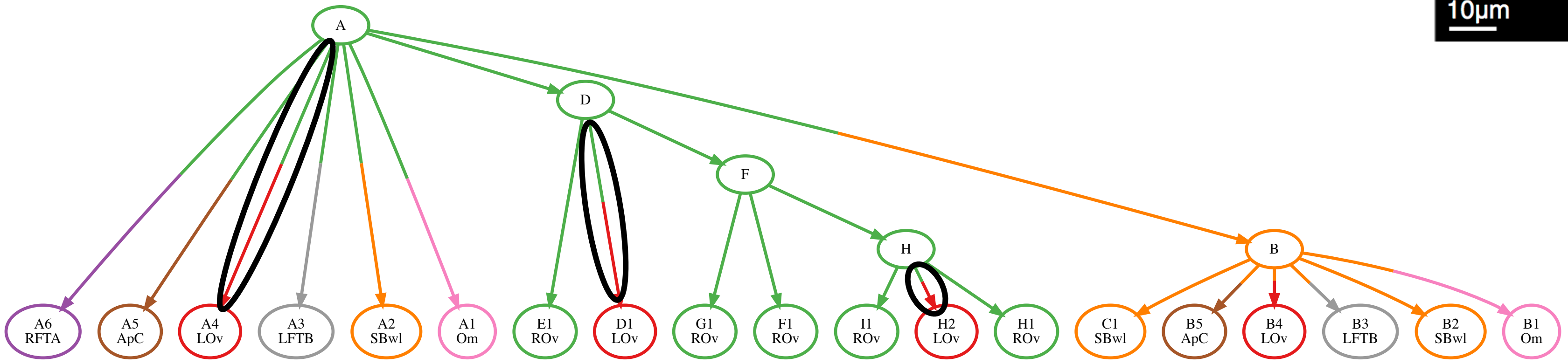
$\mu^* = 13$

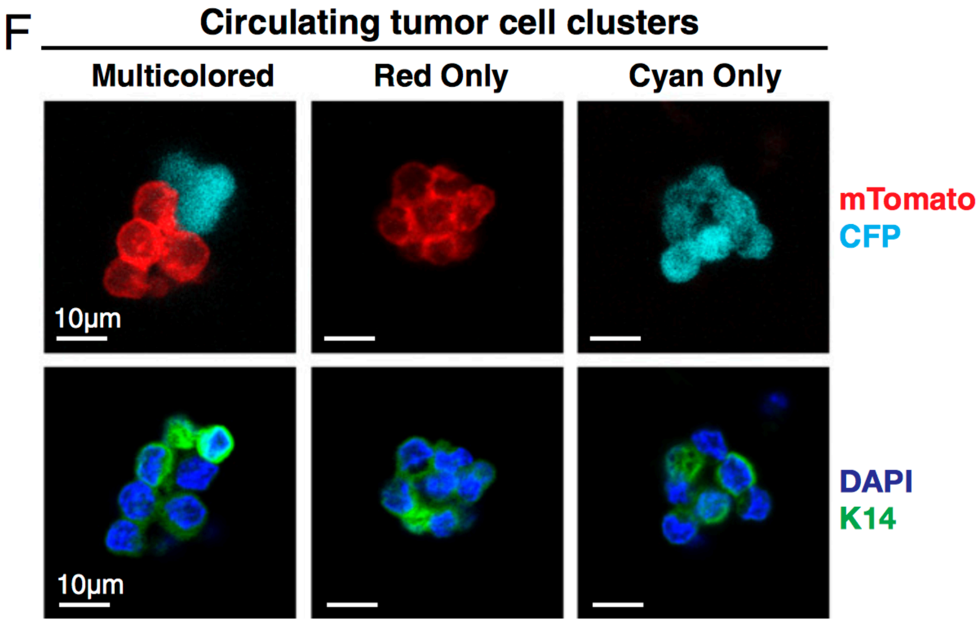
$\gamma = 10$



Migration Graph G

Clone Tree T



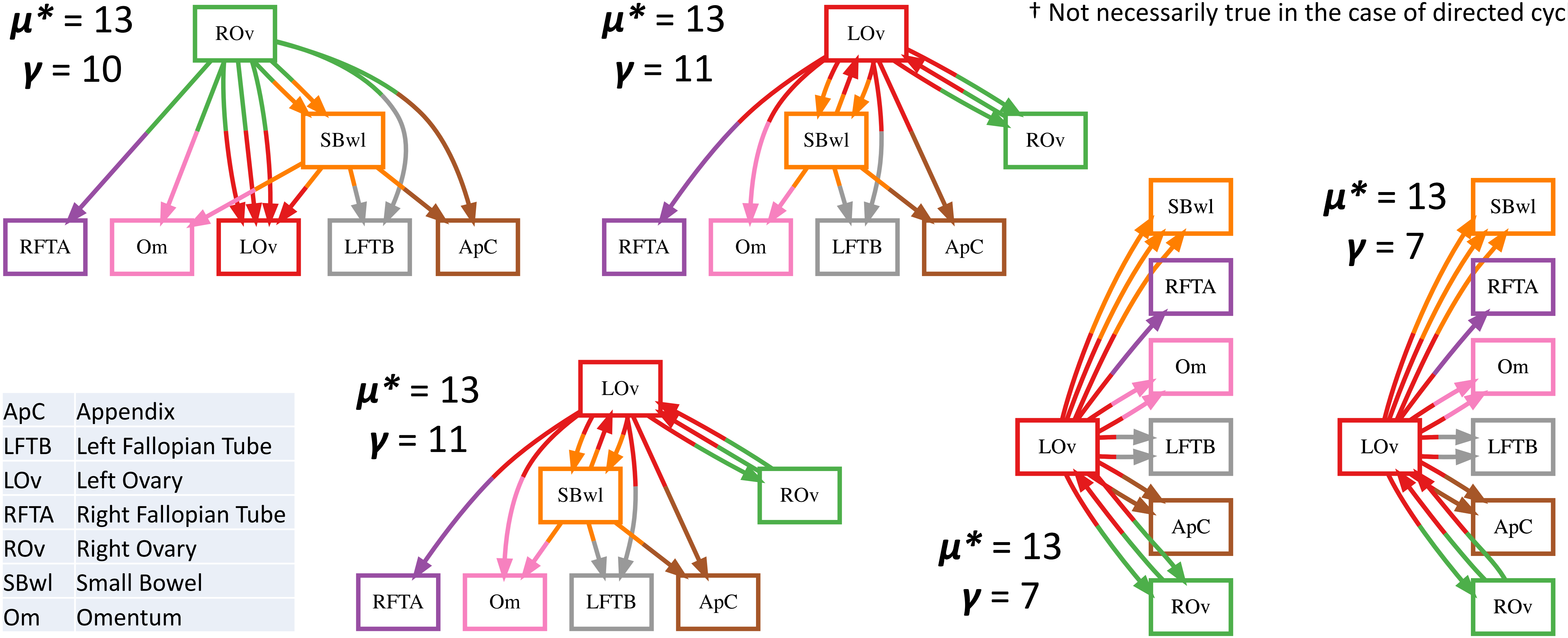


ApC	Appendix
LFTB	Left Fallopian Tube
LOv	Left Ovary
RFTA	Right Fallopian Tube
ROv	Right Ovary
SBwl	Small Bowel
Om	Omentum

Comigrations: Simultaneous Migrations of Multiple Clones

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

† Not necessarily true in the case of directed cycles



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

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

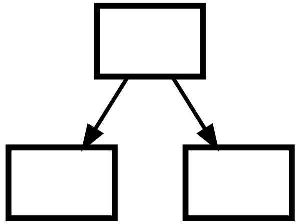
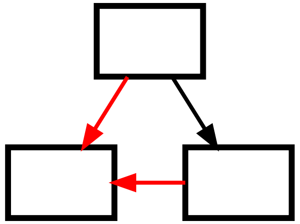
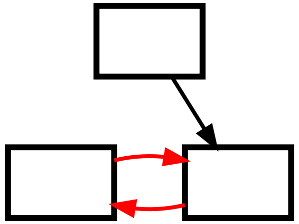
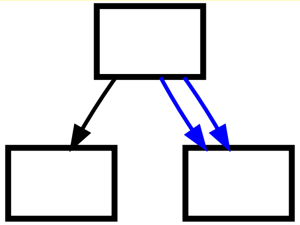
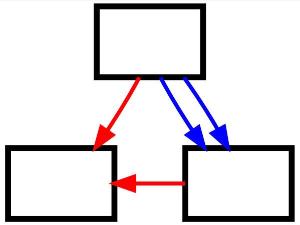
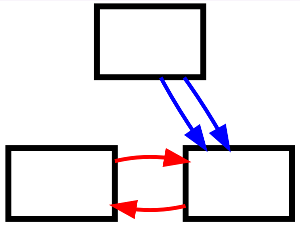
$\mu^* = 13$
$\gamma = 11$

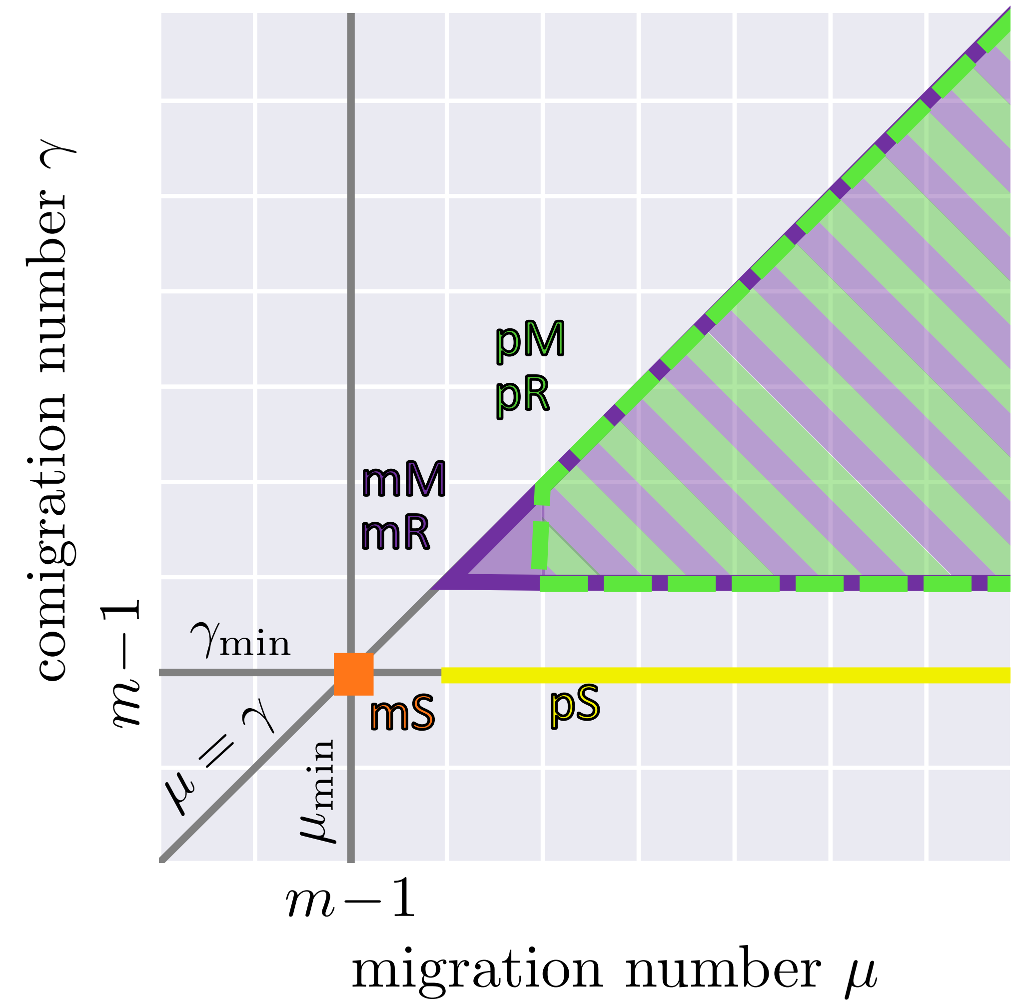
$\mu^* = 13$
$\gamma = 7$

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

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

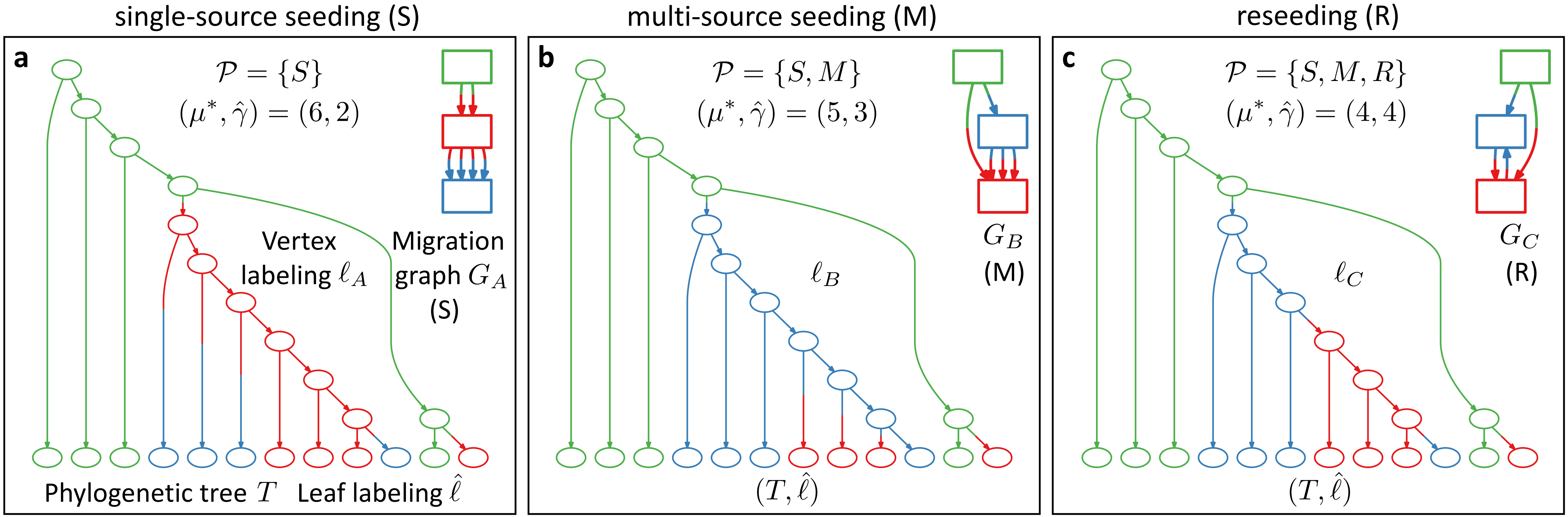
Tradeoffs between Migrations, Comigrations and Migration Pattern

	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



Constrained Multi-objective Optimization Problem

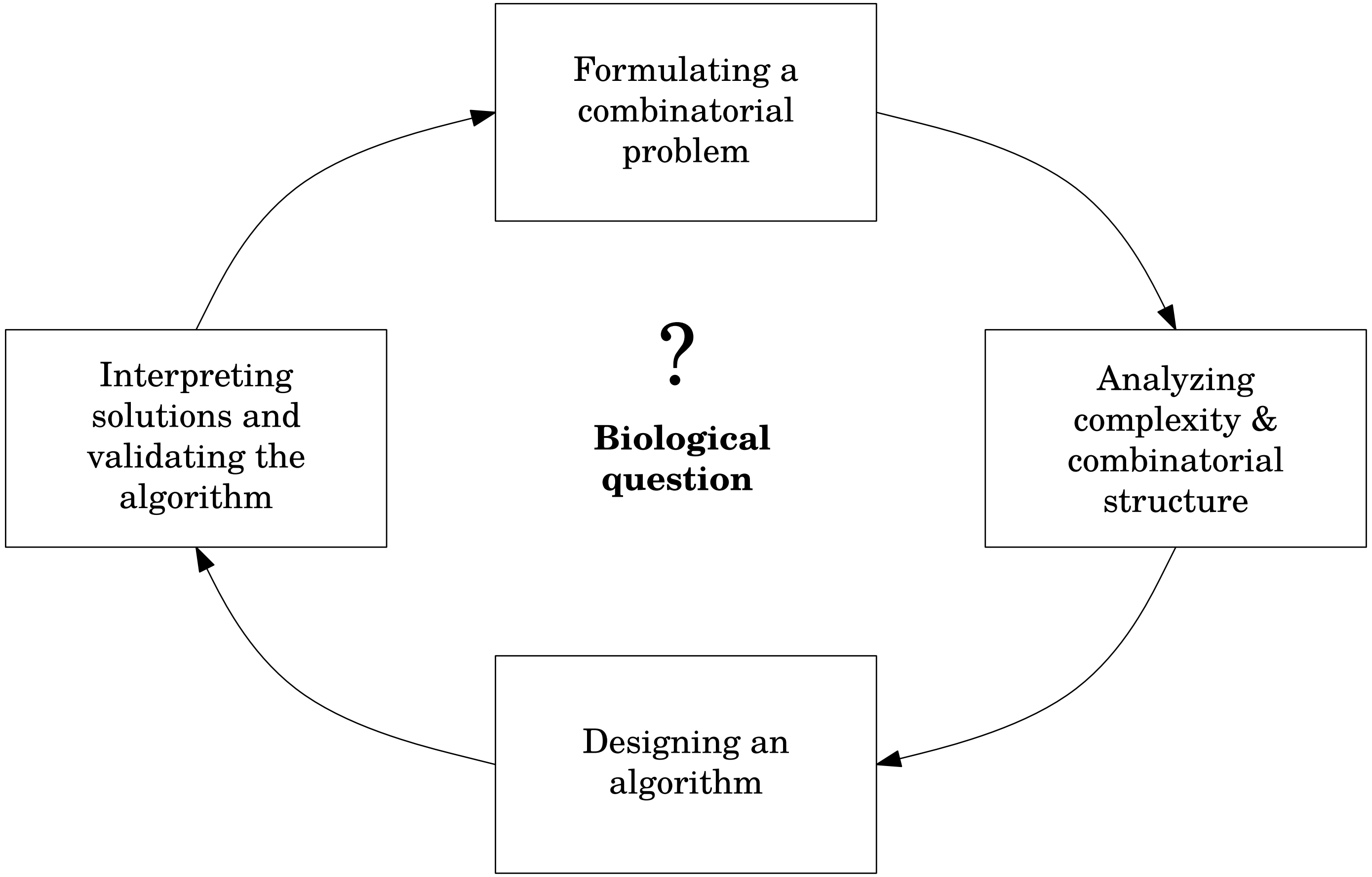
Parsimonious Migration History (PMH): Given a phylogenetic tree T and a set $\mathcal{P} \subseteq \{\mathrm{S}, \mathrm{M}, \mathrm{R}\}$ of allowed migration patterns, find vertex labeling ℓ with minimum migration number $\mu^*(T)$ and smallest comigration number $\hat{\gamma}(T)$.



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

Outline

- Metastasis
- Maximum parsimony
- Problem statement
- Complexity
- Algorithm & results
- Problem variants



Formulating a combinatorial problem

Analyzing complexity & combinatorial structure

Designing an algorithm

Interpreting solutions and validating the algorithm

?

Biological question

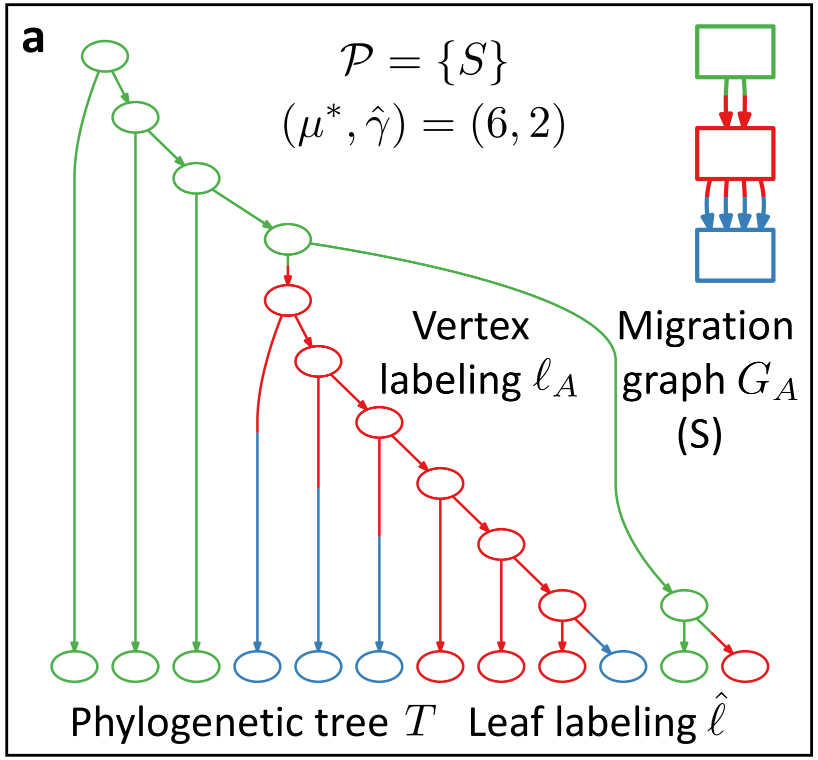
Reading:

- **M. El-Kebir**, G. Satas and B.J. Raphael. Inferring parsimonious migration histories for metastatic cancers. Nature Genetics, 50:718-726, 2018.
- **M. El-Kebir**†. Parsimonious Migration History Problem: Complexity and Algorithms. WABI 2018, Helsinki, Finland, August 20-22, 2018.

Results [El-Kebir, WABI 2018]

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)

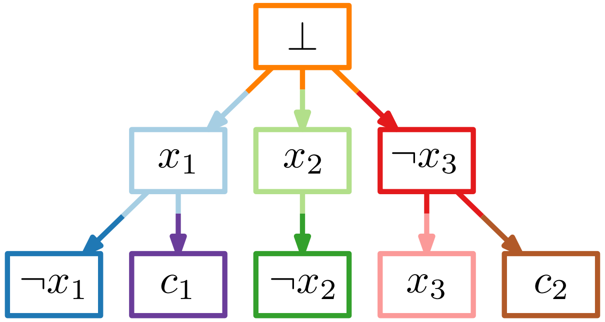


Theorem 1: PMH is NP-hard when $\mathcal{P} = \{S\}$

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

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

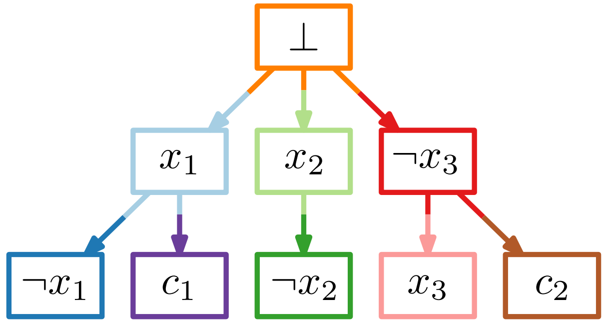
3-SAT: Given $\varphi = \bigwedge_{i=1}^{k} (y_{i,1} \vee y_{i,2} \vee y_{i,3})$ with variables $\{x_1, \dots, x_n\}$ and k clauses, find $\phi : [n] \to \{0,1\}$ satisfying φ



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

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

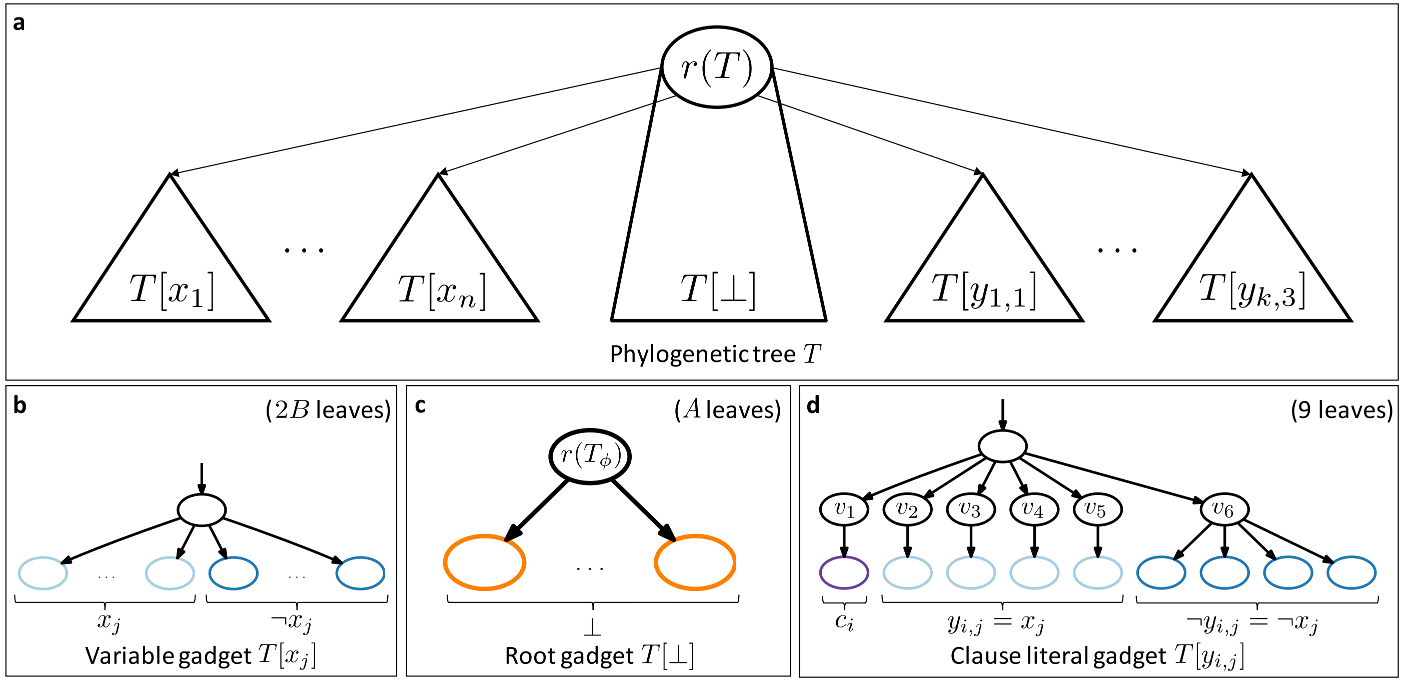
3-SAT: Given $\varphi = \wedge_{i=1}^{k}(y_{i,1} \vee y_{i,2} \vee y_{i,3})$ with variables $\{x_1, \dots, x_n\}$ and k clauses, find $\phi : [n] \rightarrow \{0,1\}$ satisfying φ



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

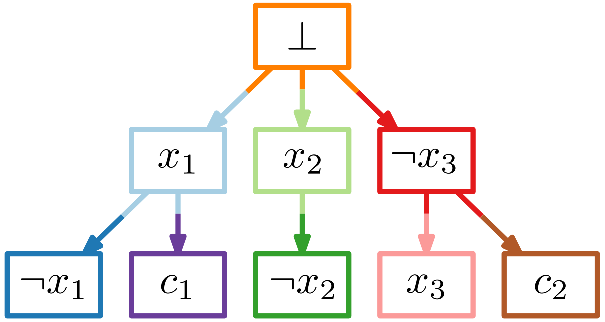
Three ideas:

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



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

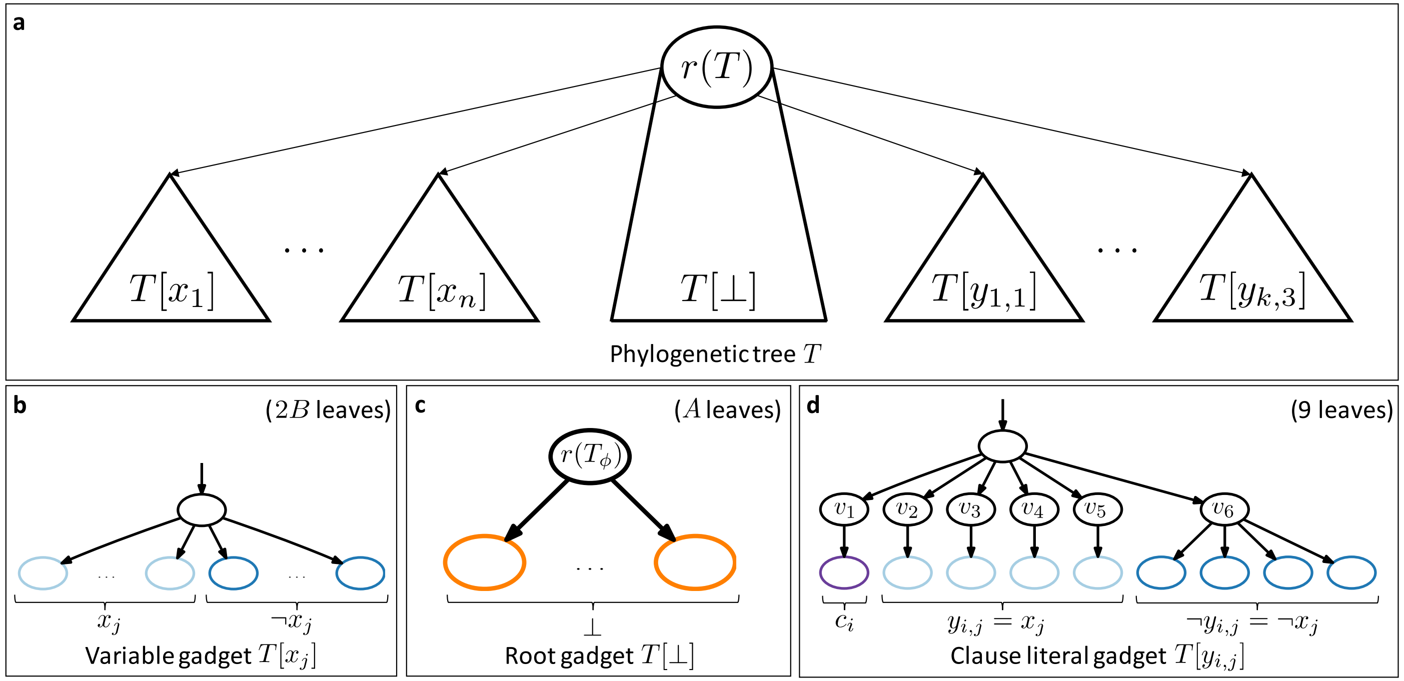
3-SAT: Given $\varphi = \wedge_{i=1}^{k}(y_{i,1} \vee y_{i,2} \vee y_{i,3})$ with variables $\{x_1, \ldots, x_n\}$ and k clauses, find $\phi : [n] \to \{0,1\}$ satisfying φ



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

Three ideas:

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



Phylogenetic tree T

Variable gadget $T[x_j]$

Root gadget $T[\perp]$

Clause literal gadget $T[y_{i,j}]$

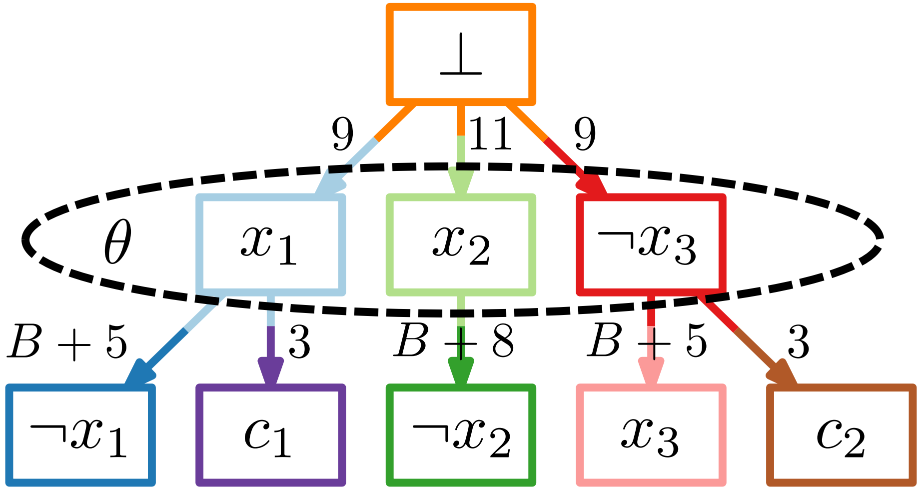
Lemma: Let $B > 10k + 1$ and $A > 2Bn + 27k$.
Then, φ is satisfiable if and only if $\mu^*(T) = (B+1)n + 25k$

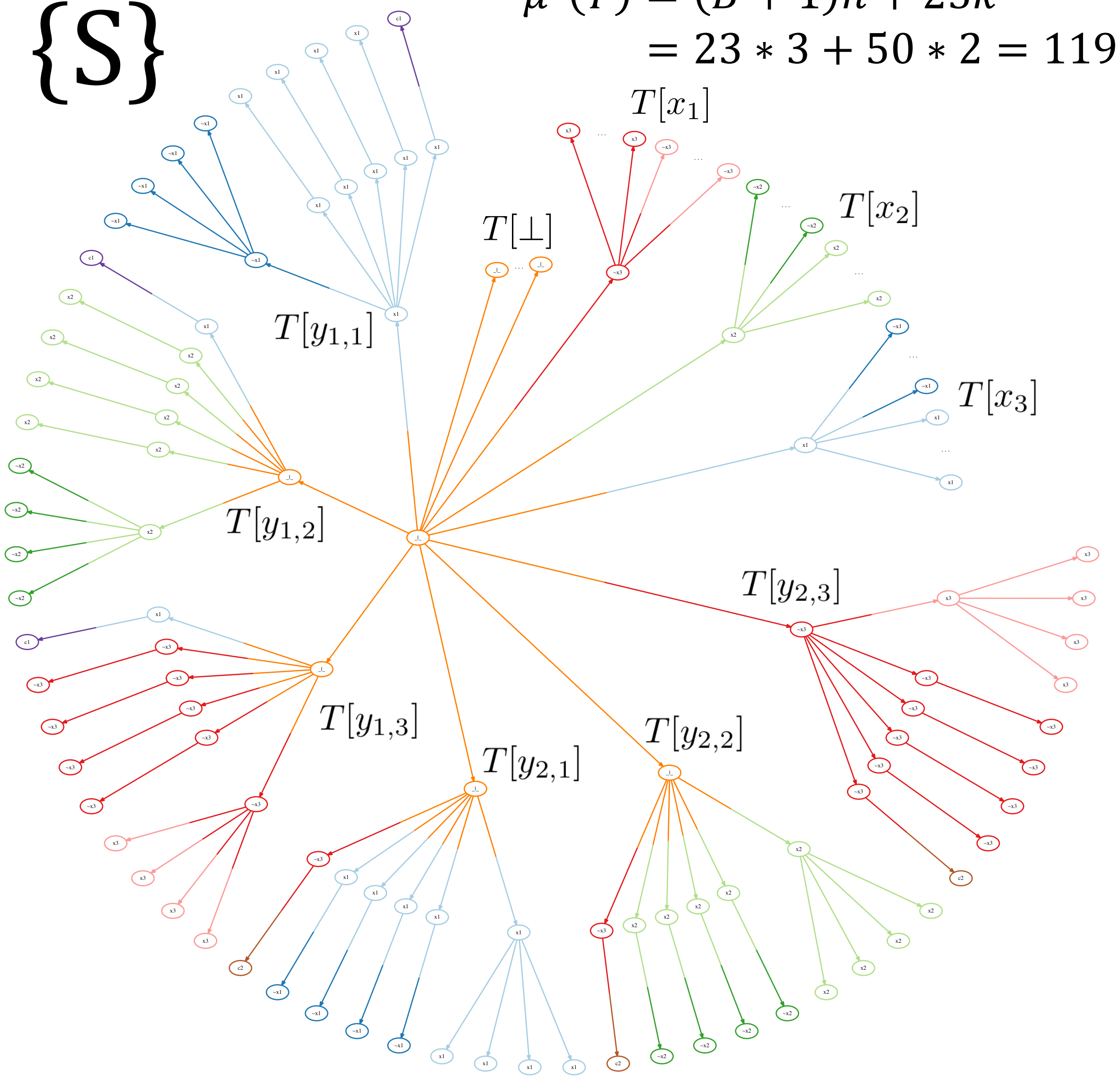
PMH is NP-hard when $\mathcal{P} = \{\mathrm{S}\}$

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

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

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



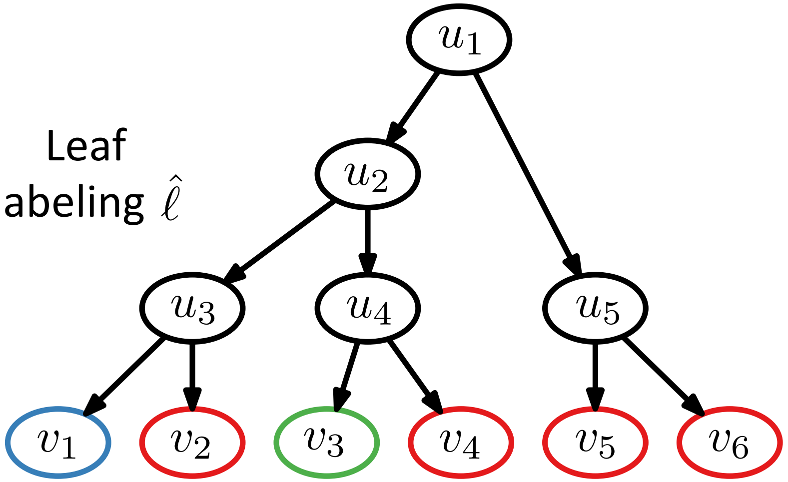


$$\Sigma = \{x_1, x_2, x_3, \neg x_1, \neg x_2, \neg x_3, c_1, c_2, \perp\}$$

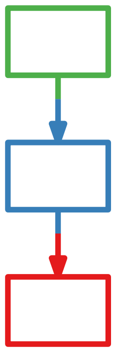
Lemma: Let $B > 10k + 1$ and $A > 2Bn + 27k$.
Then, φ is satisfiable if and only if $\mu^*(T) = (B+1)n + 25k$

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

Leaf labeling $\hat{\ell}$

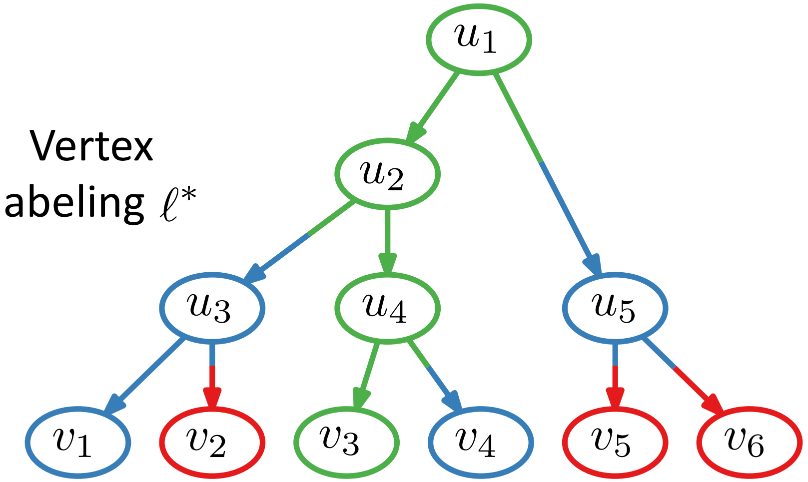


Phylogenetic tree T



Migration tree $\hat{G}$

Vertex labeling ℓ^*



Phylogenetic tree T

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

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

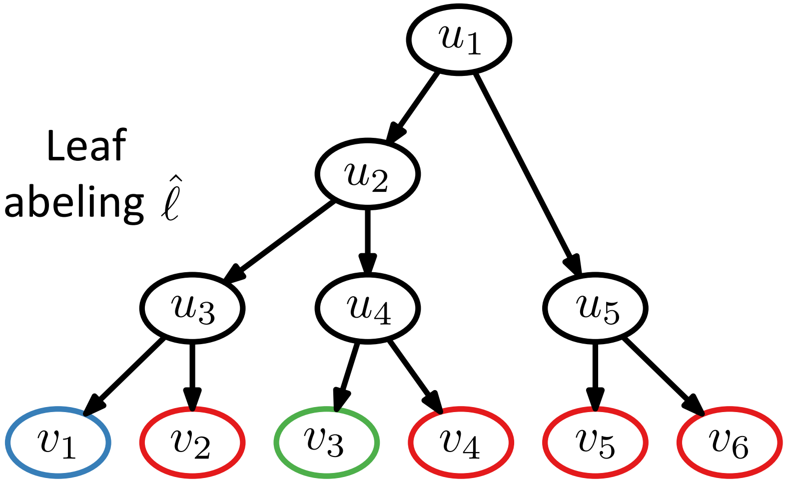
$$\ell^*(v) = \begin{cases} \mathrm{LCA}_{\hat{G}}(r(T)), & \text{if } v = r(T), \\ \sigma(\ell^*(\pi(v)), \mathrm{LCA}_{\hat{G}}(v)), & \text{if } v \neq r(T), \end{cases}$$

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 $\hat{G}$.

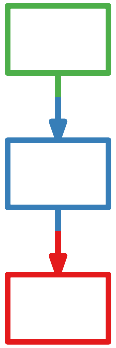
Lemma: If (1) holds then ℓ^* is a minimum migration labeling consistent with $\widehat{G}$.

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

Leaf labeling $\hat{\ell}$

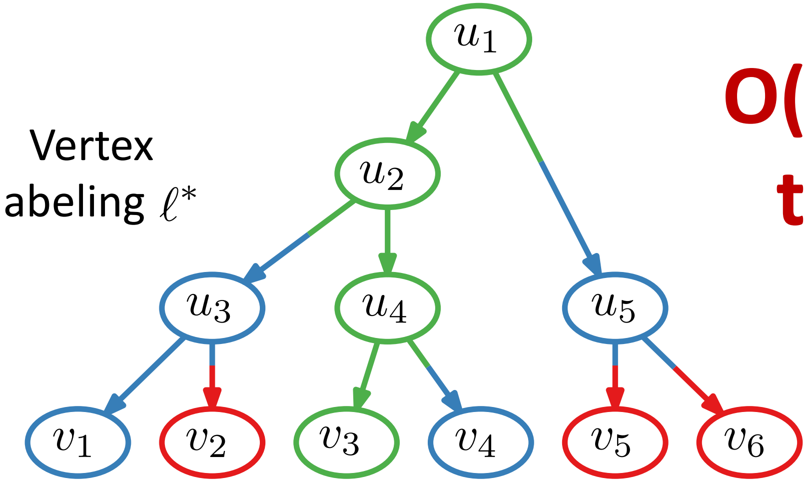


Phylogenetic tree T



Migration tree $\hat{G}$

Vertex labeling ℓ^*



Phylogenetic tree T

O(nm^m) time

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

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

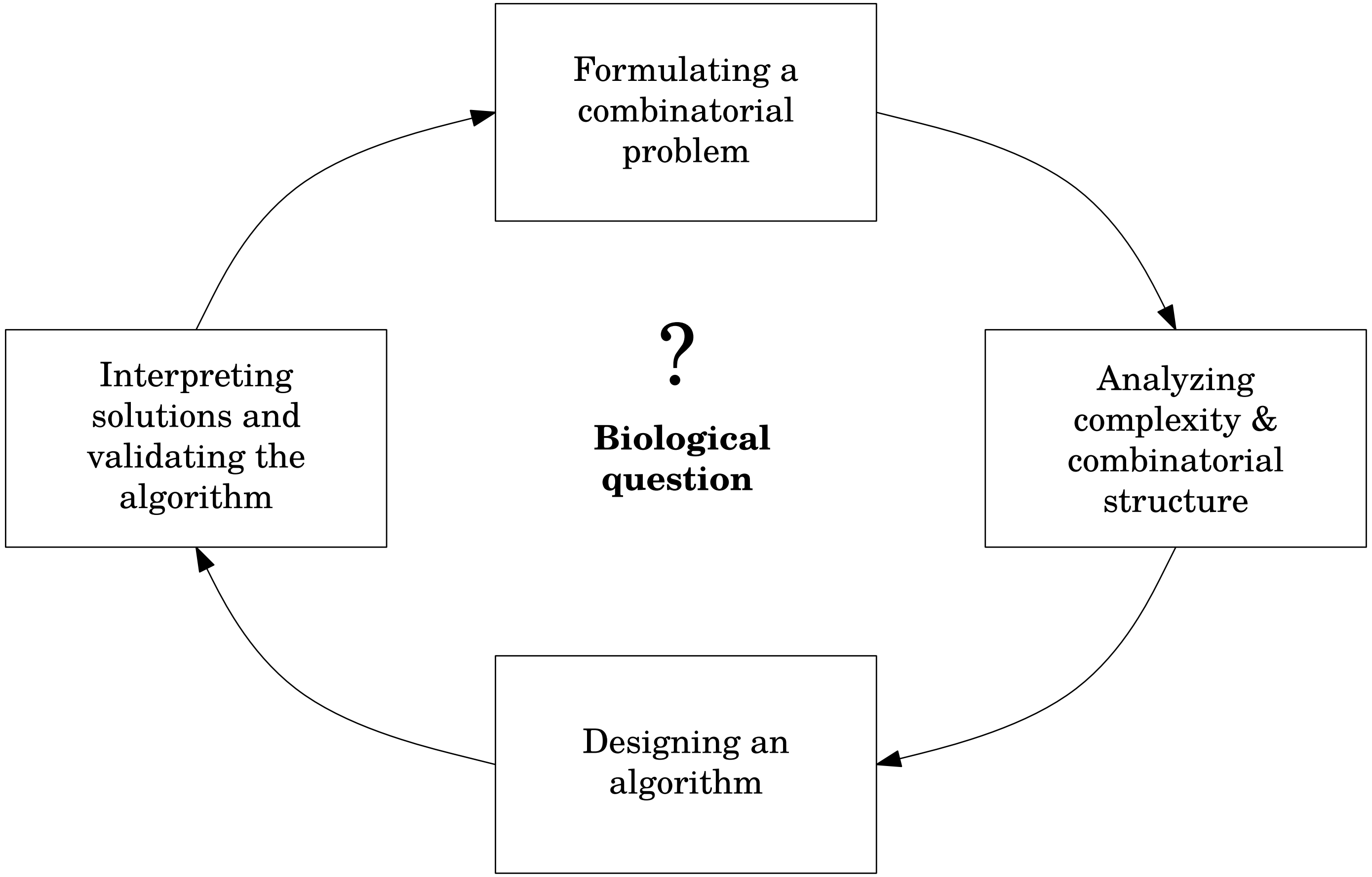
$$\ell^*(v) = \begin{cases} \mathrm{LCA}_{\hat{G}}(r(T)), & \text{if } v = r(T), \\ \sigma(\ell^*(\pi(v)), \mathrm{LCA}_{\hat{G}}(v)), & \text{if } v \neq r(T), \end{cases}$$

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 $\hat{G}$.

Outline

- Metastasis
- Maximum parsimony
- Problem statement
- Complexity
- Algorithm & results
- Problem variants



Formulating a combinatorial problem

Analyzing complexity & combinatorial structure

Designing an algorithm

Interpreting solutions and validating the algorithm

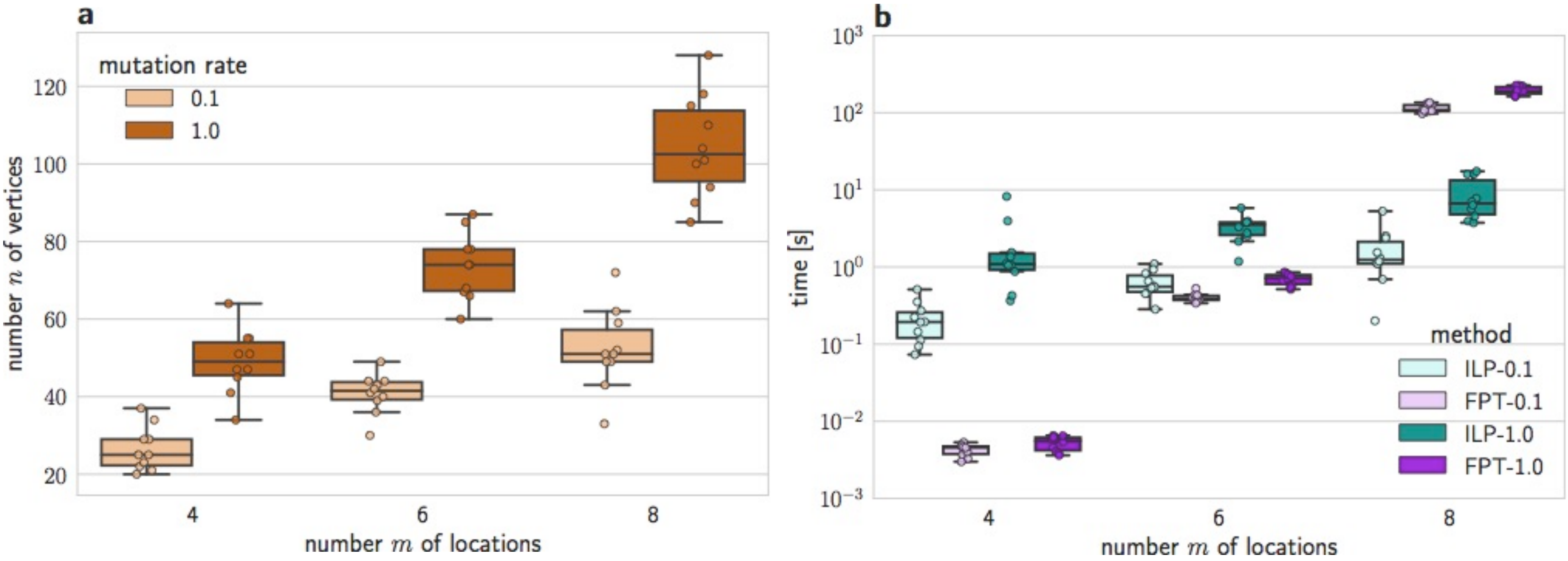
?

Biological question

Reading:

- **M. El-Kebir**, G. Satas and B.J. Raphael. Inferring parsimonious migration histories for metastatic cancers. Nature Genetics, 50:718-726, 2018.
- **M. El-Kebir**[†]. Parsimonious Migration History Problem: Complexity and Algorithms. WABI 2018, Helsinki, Finland, August 20-22, 2018.

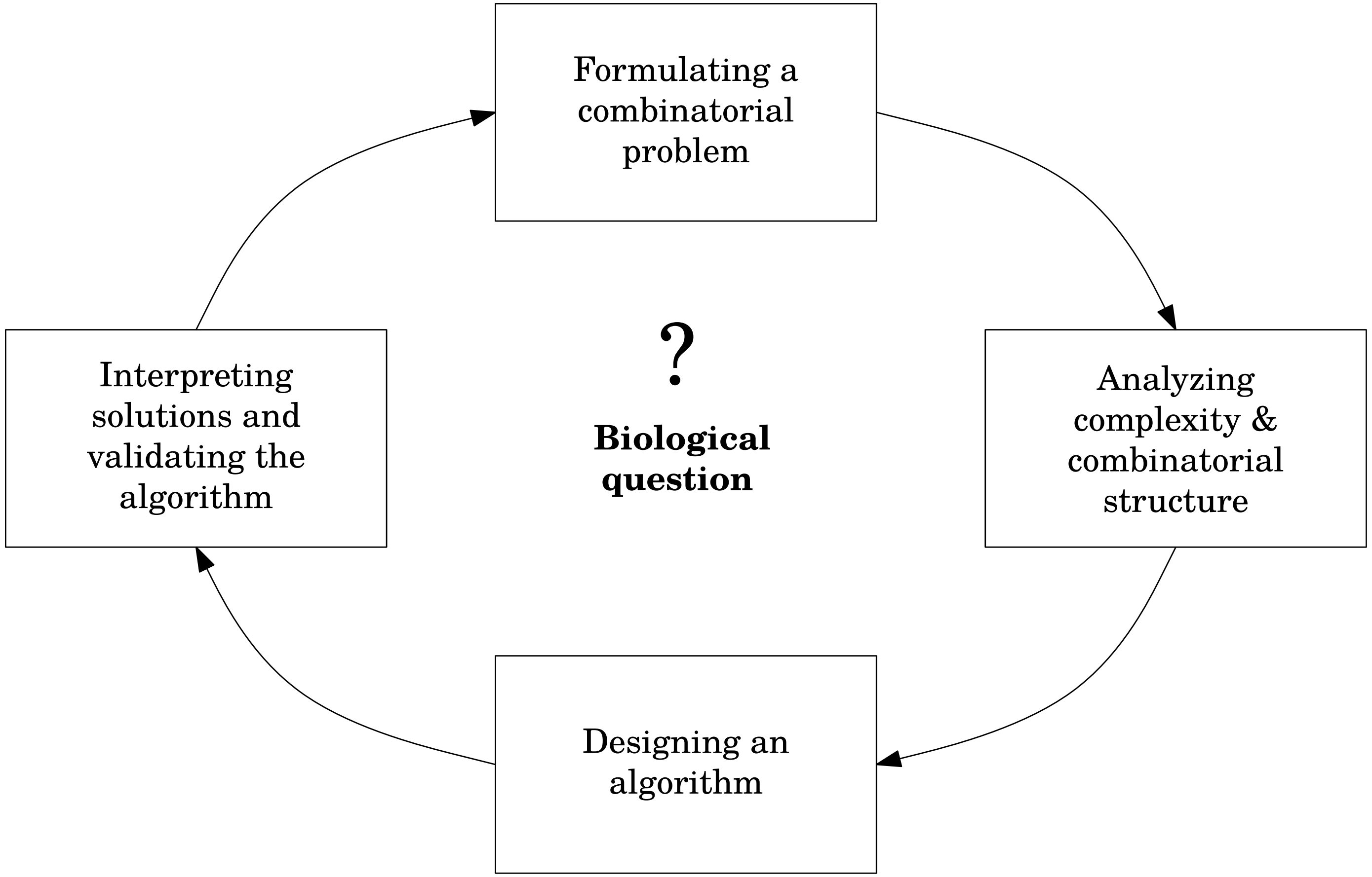
Simulations



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

Outline

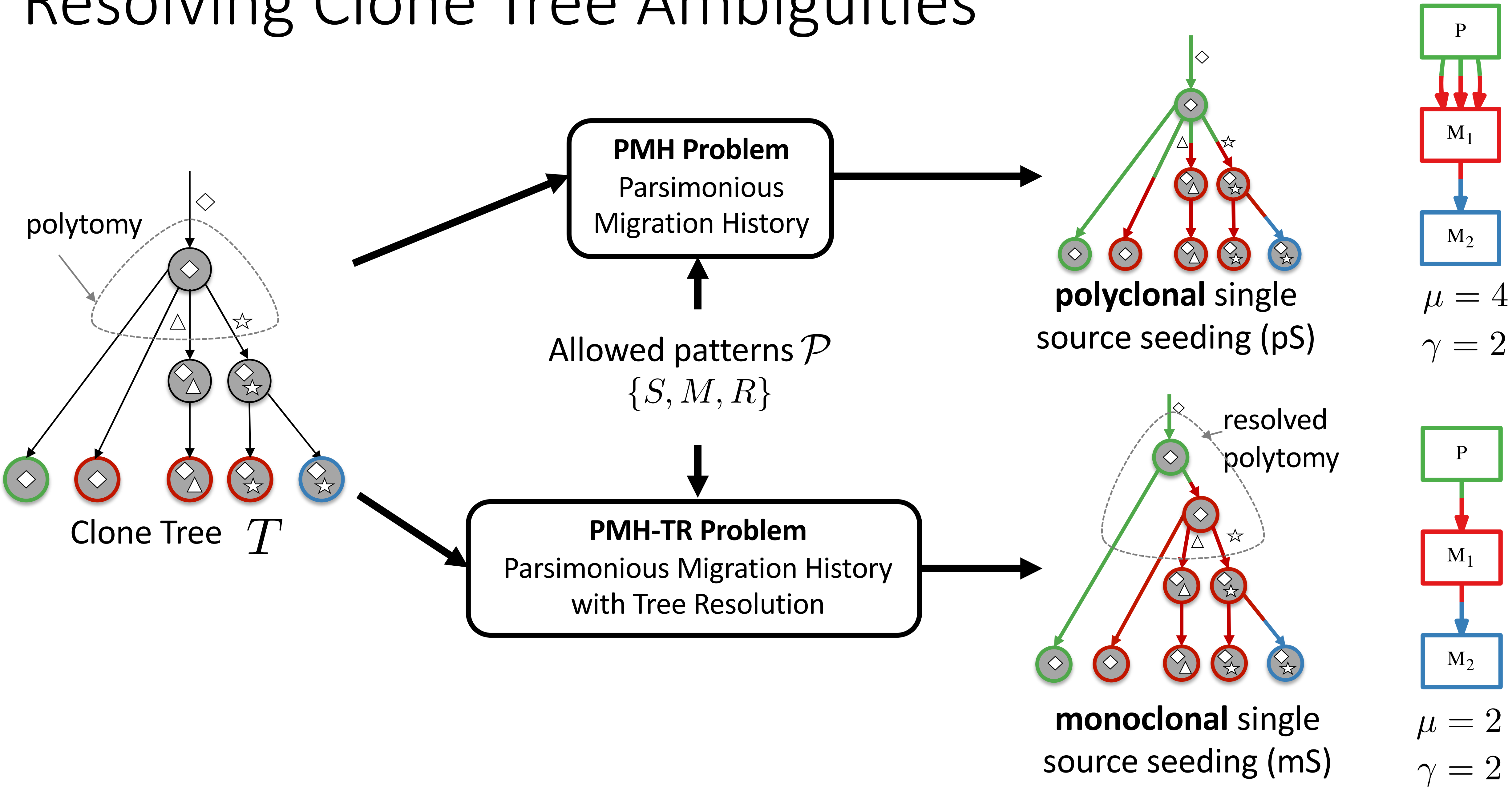
- Metastasis
- Maximum parsimony
- Problem statement
- Complexity
- Algorithm & results
- Problem variants



Reading:

- **M. El-Kebir**, G. Satas and B.J. Raphael. Inferring parsimonious migration histories for metastatic cancers. Nature Genetics, 50:718-726, 2018.
- **M. El-Kebir**†. Parsimonious Migration History Problem: Complexity and Algorithms. WABI 2018, Helsinki, Finland, August 20-22, 2018.

Resolving Clone Tree Ambiguities



Clone Tree T

PMH Problem
Parsimonious
Migration History

Allowed patterns $\mathcal{P}$
$\{S, M, R\}$

PMH-TR Problem
Parsimonious Migration History
with Tree Resolution

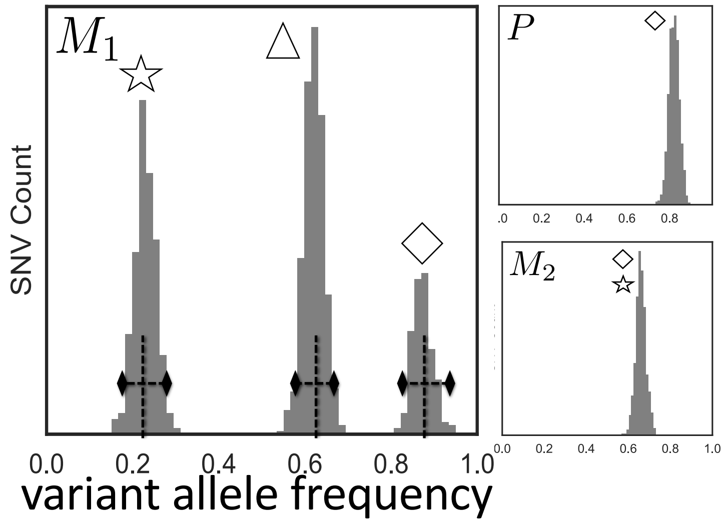
polyclonal single
source seeding (pS)

$\mu = 4$
$\gamma = 2$

monoclonal single
source seeding (mS)

$\mu = 2$
$\gamma = 2$

Resolving Clone Tree Ambiguities



confidence intervals

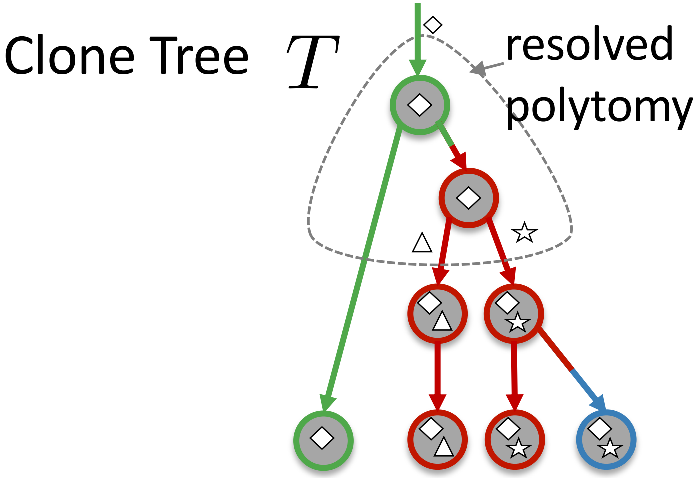
$$F^- = \begin{array}{c|ccc} & \diamond & \triangle & \star \\ P & 0.75 & 0 & 0 \\ M_1 & 0.82 & 0.15 & 0.56 \\ M_2 & 0.66 & 0 & 0.68 \end{array}$$

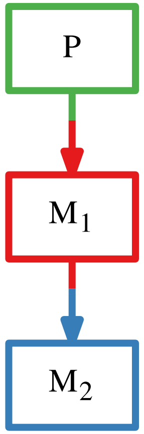
$$F^+ = \begin{array}{c|ccc} & \diamond & \triangle & \star \\ P & 0.86 & 0 & 0 \\ M_1 & 0.92 & 0.30 & 0.68 \\ M_2 & 0.78 & 0 & 0.77 \end{array}$$

$$\underbrace{\begin{array}{c|ccc} & \diamond & \triangle & \star \\ P & 0.82 & 0 & 0 \\ M_1 & 0.88 & 0.24 & 0.64 \\ M_2 & 0.73 & 0 & 0.73 \end{array}}_{\hat{F}} = \underbrace{\begin{array}{c|ccc} & \diamond & \diamond\triangle & \diamond\star \\ P & \boxed{0.82} & 0 & 0 \\ M_1 & 0 & \boxed{0.24} & \boxed{0.64} \\ M_2 & 0 & 0 & \boxed{0.73} \end{array}}_{U} \underbrace{\begin{array}{ccc} \diamond & \triangle & \star \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{array}}_{B}$$

PMH-TI Problem
Parsimonious Migration History
with Tree Inference

$\mathcal{P}$

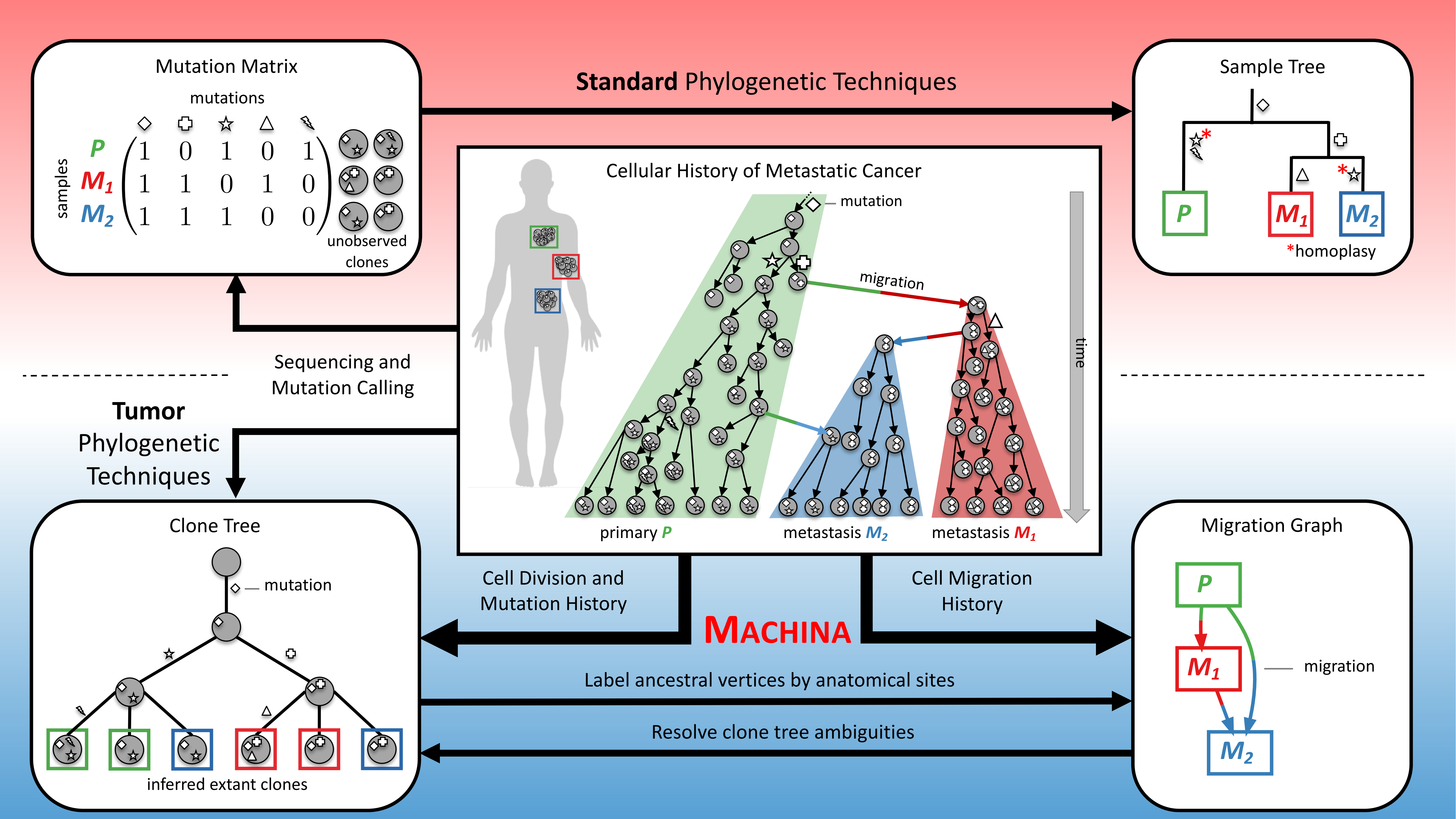




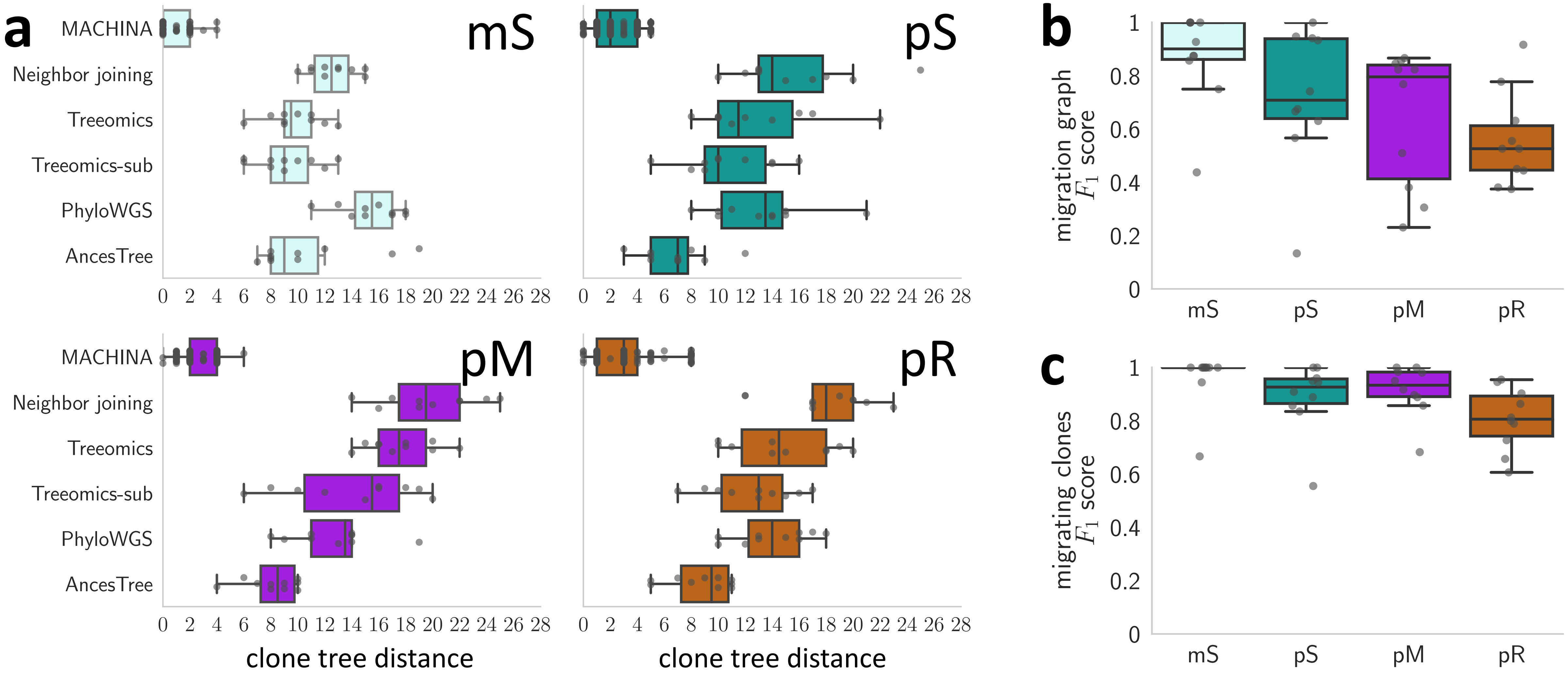
$\mu = 2$

$\gamma = 2$

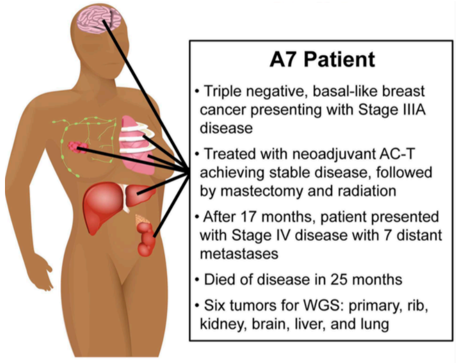
monoclonal single source seeding (mS)


Mutation Matrix
mutations
samples
P
M_1
M_2
1 0 1 0 1
1 1 0 1 0
1 1 1 0 0
unobserved clones
Standard Phylogenetic Techniques
Sample Tree
P
M_1
M_2
*homoplasy
Cellular History of Metastatic Cancer
mutation
migration
time
primary P
metastasis M_2
metastasis M_1
Sequencing and Mutation Calling
Tumor Phylogenetic Techniques
Clone Tree
mutation
inferred extant clones
Cell Division and Mutation History
Cell Migration History
MACHINA
Label ancestral vertices by anatomical sites
Resolve clone tree ambiguities
Migration Graph
P
M_1
M_2
migration

MACHINA accurately infers clone trees and migration histories on simulated data



Applying MACHINA to Metastatic Breast Cancer



A7 Patient

- Triple negative, basal-like breast cancer presenting with Stage IIIA disease
- Treated with neoadjuvant AC-T achieving stable disease, followed by mastectomy and radiation
- After 17 months, patient presented with Stage IV disease with 7 distant metastases
- Died of disease in 25 months
- Six tumors for WGS: primary, rib, kidney, brain, liver, and lung

Hoadley et al.
Tumor Evolution in Two Patients with Basal-like Breast Cancer: A Retrospective Genomics Study of Multiple Metastases.
PLOS Med, *13*(12) 2016