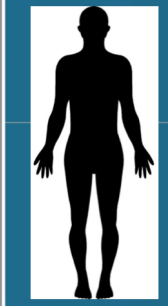
Combinatorial Algorithms for Tumor Phylogenetics

Mohammed El-Kebir

Cost per Genome

$100M
$10M
$1M
$100K
$10K
$1K

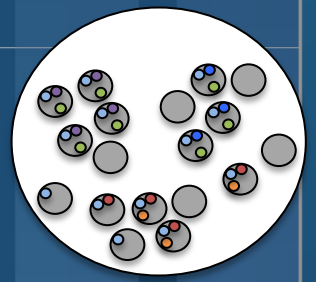
Moore's Law



2001: Human Genome

2007:
Next-gen sequencing

2013: The Cancer Genome Atlas
Inter and **Intra-Tumor Heterogeneity**



NIH National Human Genome Research Institute

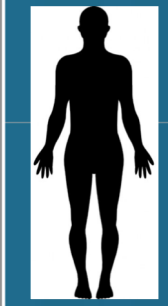
genome.gov/sequencingcosts

2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015

Cost per Genome

$100M
$10M
$1M
$100K
$10K
$1K

Moore's Law

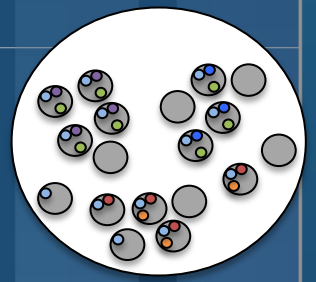


2001: Human Genome

2007:
Next-gen sequencing

2013: The Cancer Genome Atlas
Inter and **Intra-Tumor Heterogeneity**

Computational Approaches



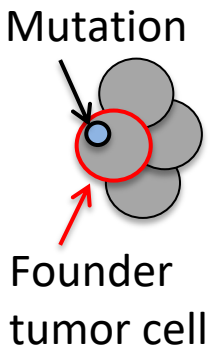
NIH National Human Genome Research Institute

genome.gov/sequencingcosts

2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015

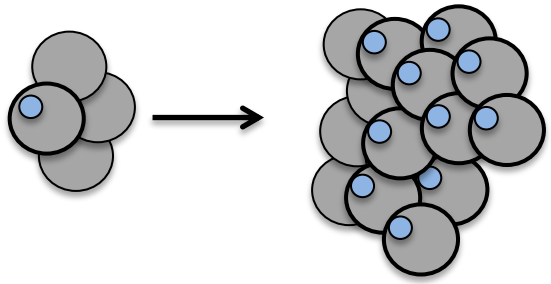
Cancer Evolution and Intra-Tumor Heterogeneity

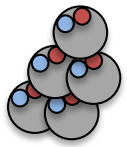
Clonal Theory of Cancer
[Nowell, 1976]



Cancer Evolution and Intra-Tumor Heterogeneity

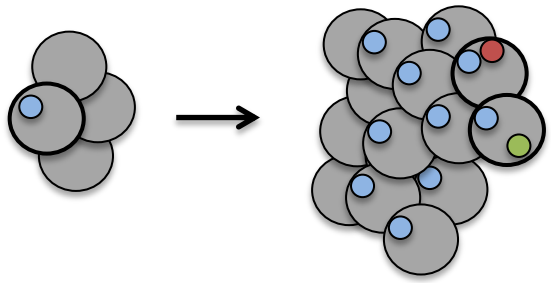
Clonal Theory of Cancer
[Nowell, 1976]

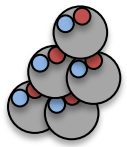


Clone  is a group  of cells with the same mutations {●, ●}

Cancer Evolution and Intra-Tumor Heterogeneity

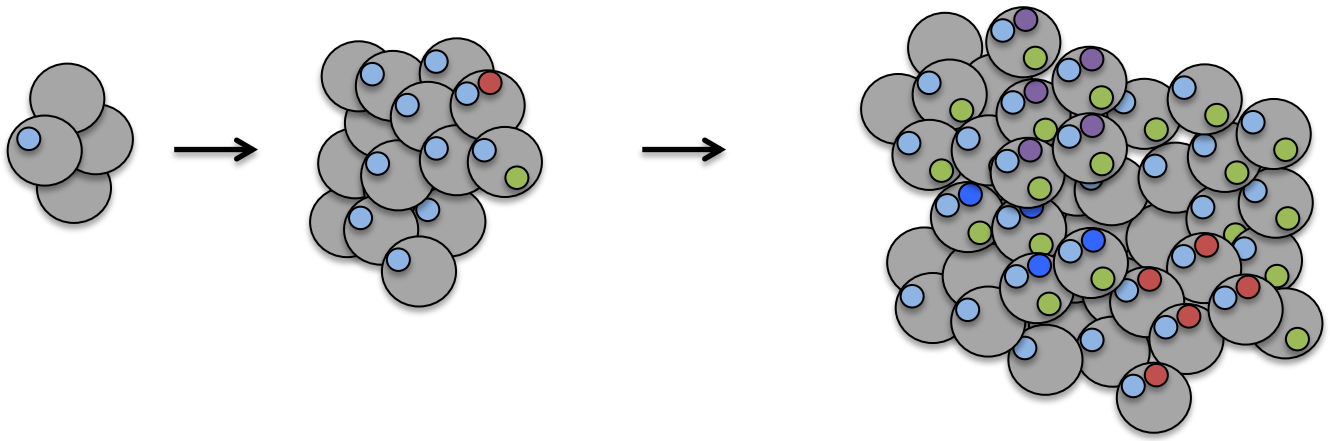
Clonal Theory of Cancer
[Nowell, 1976]



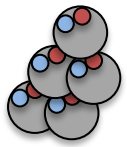
Clone  is a group  of cells with the same mutations {●, ●}

Cancer Evolution and Intra-Tumor Heterogeneity

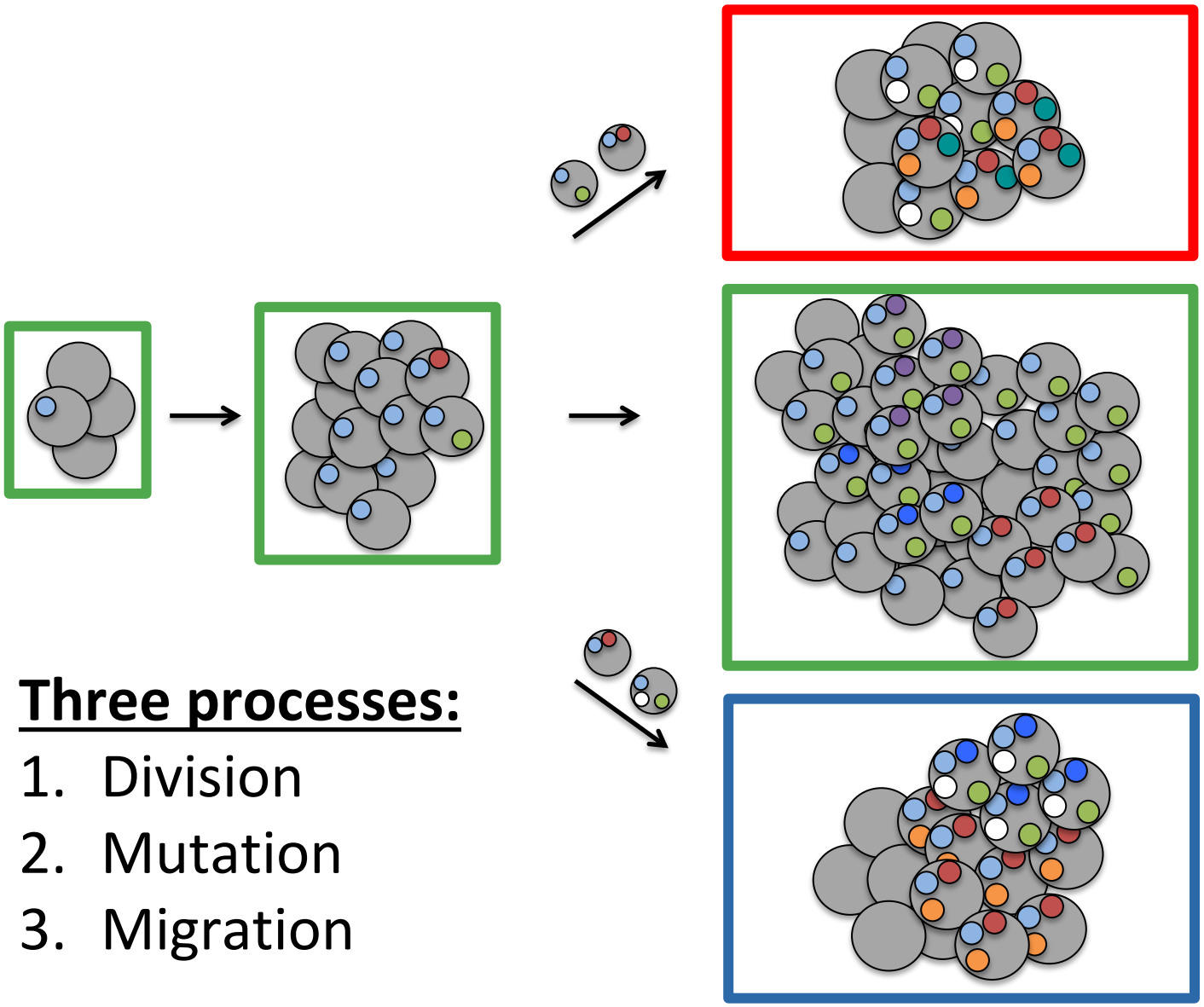
Clonal Theory of Cancer
[Nowell, 1976]



Heterogeneous Tumor

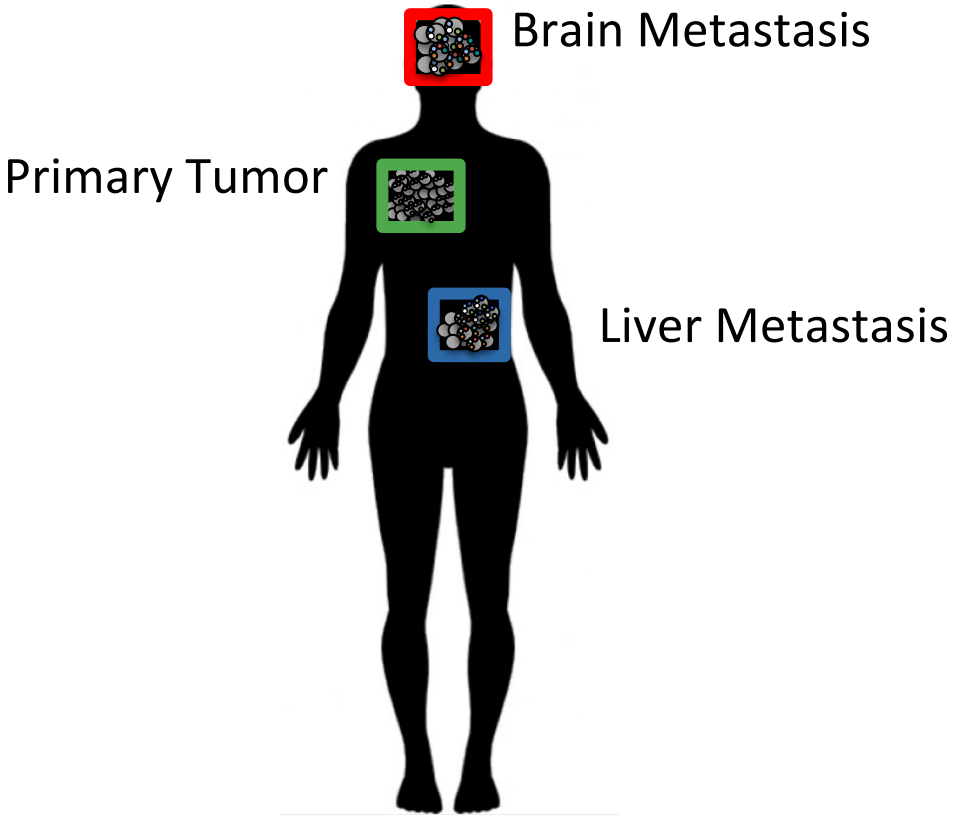
Clone  is a group  of cells with the same mutations {●, ●}

Cancer Evolution: Cell Division, Mutation & Migration



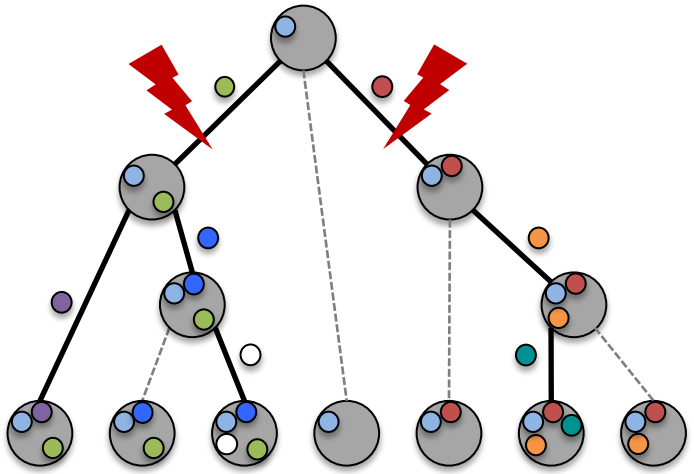
Three processes:

1. Division
2. Mutation
3. Migration

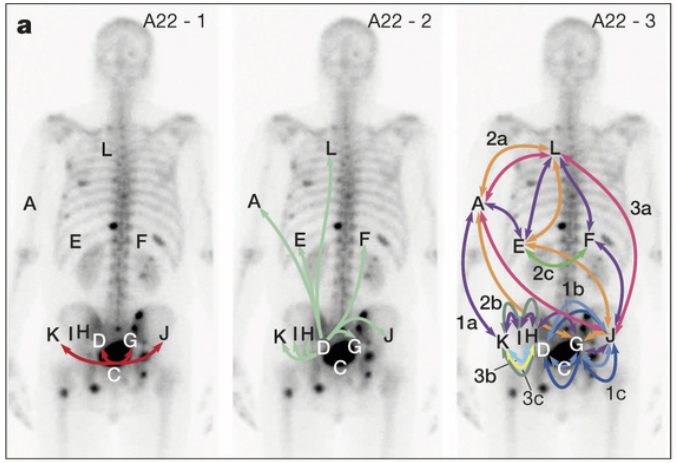


Understanding Tumor Life History Has Clinical Applications

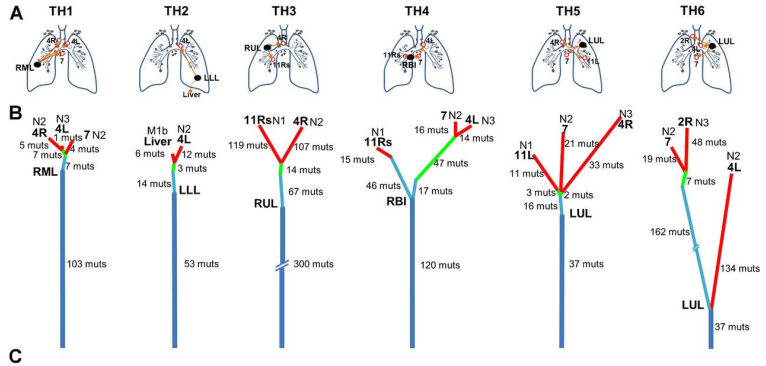
1. Identify targets for treatment



2. Understand metastatic development

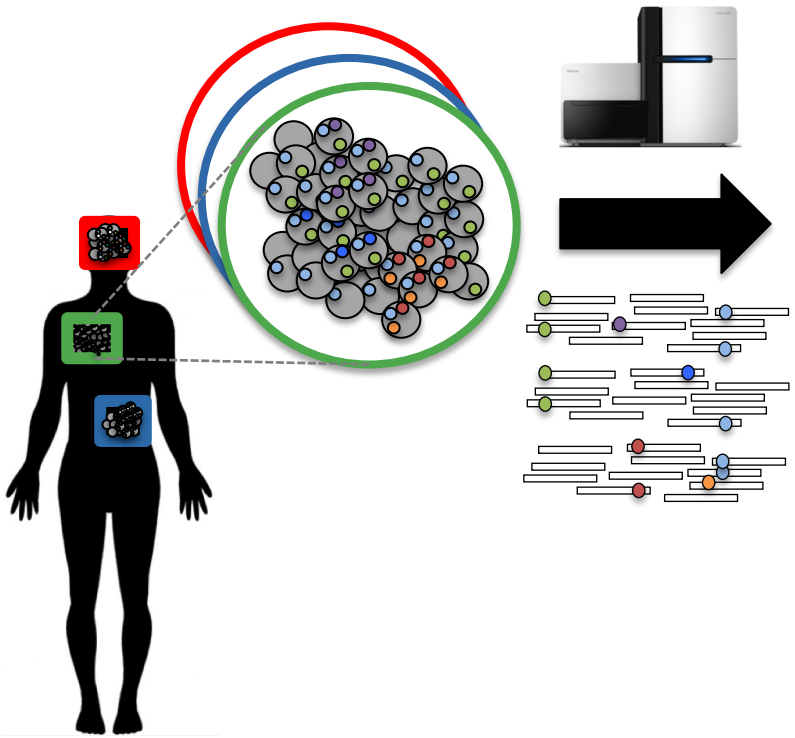


3. Recognize common patterns of tumor evolution across patients

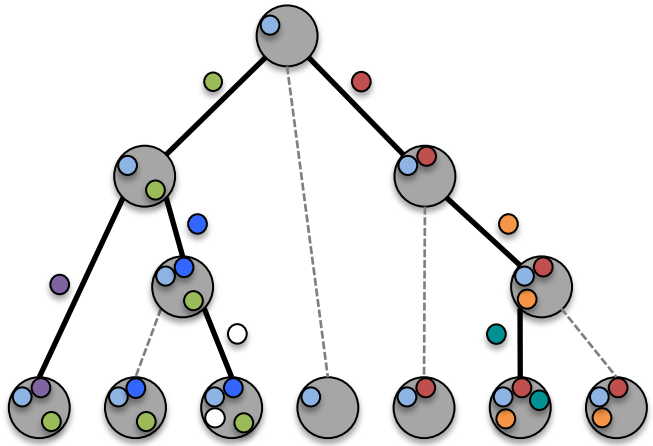


Figures adapted from [Um et al., 2016] and [Gundem et al., 2015]

Outline – Reconstructing Cancer Evolution



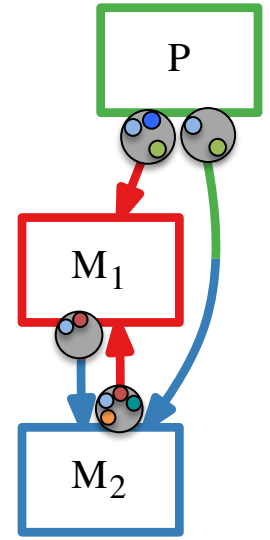
Cell division and mutation



April 10 & 12, 2018

[**El-Kebir***, Oesper* et al., *ISMB* 2015/*Bioinformatics*]
[**El-Kebir** et al., *RECOMB* 2016]; [**El-Kebir***, Satas* et al., *Cell Systems* 2016]

Cell migration

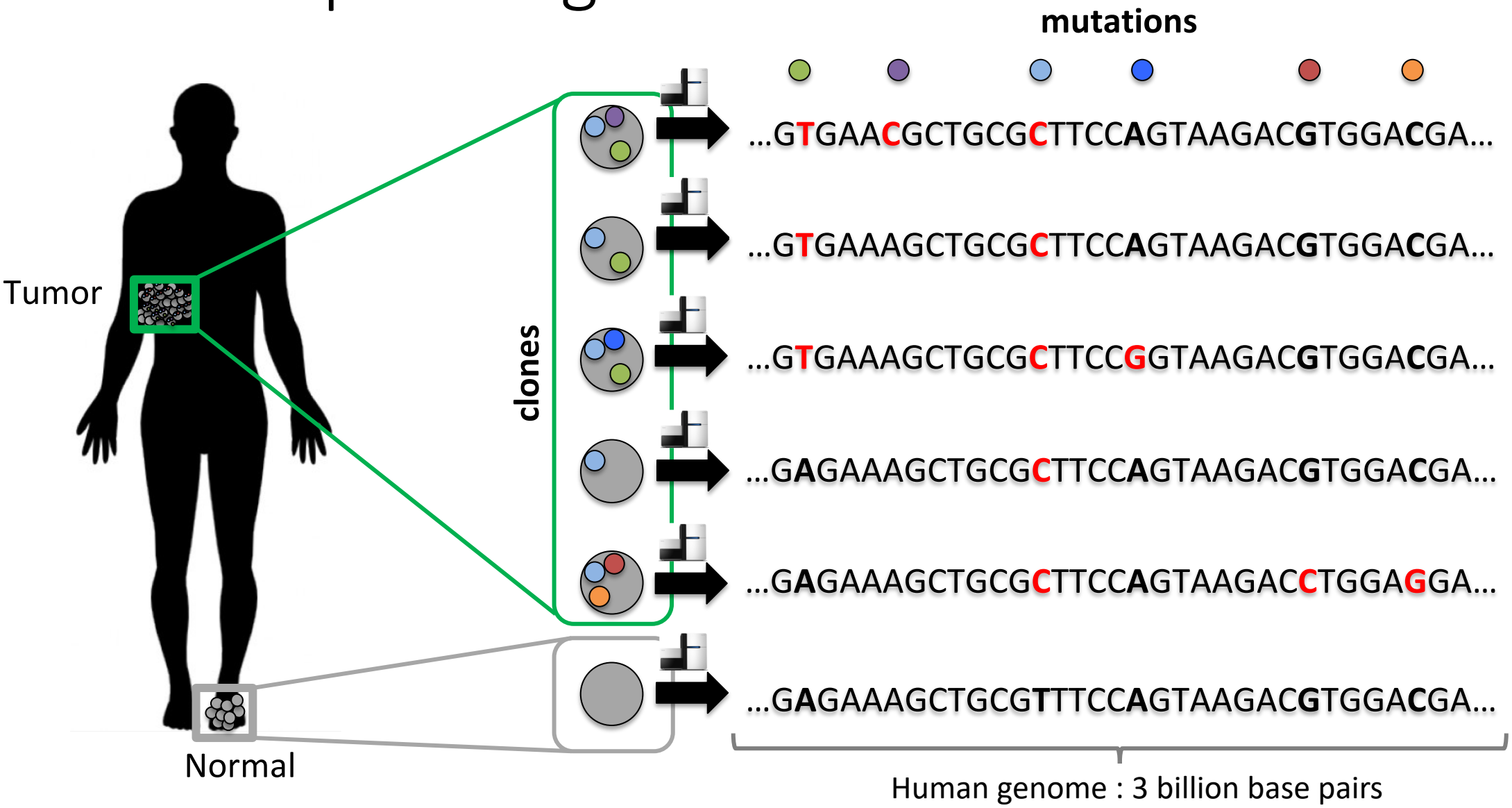


April 16, 2018

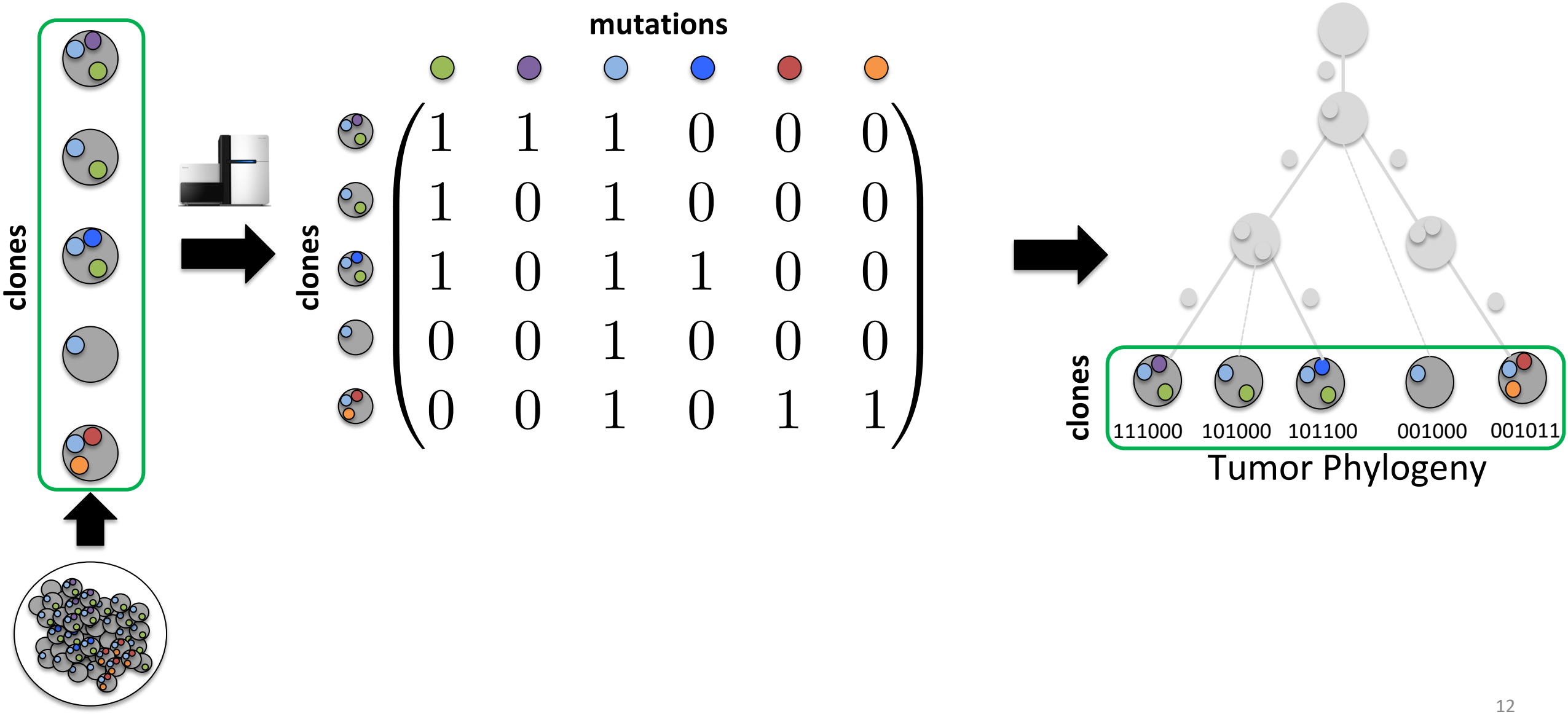
[**El-Kebir** et al., *Nature Genetics* 2018]

Precise mathematical models are needed to describe the evolutionary process in cancer

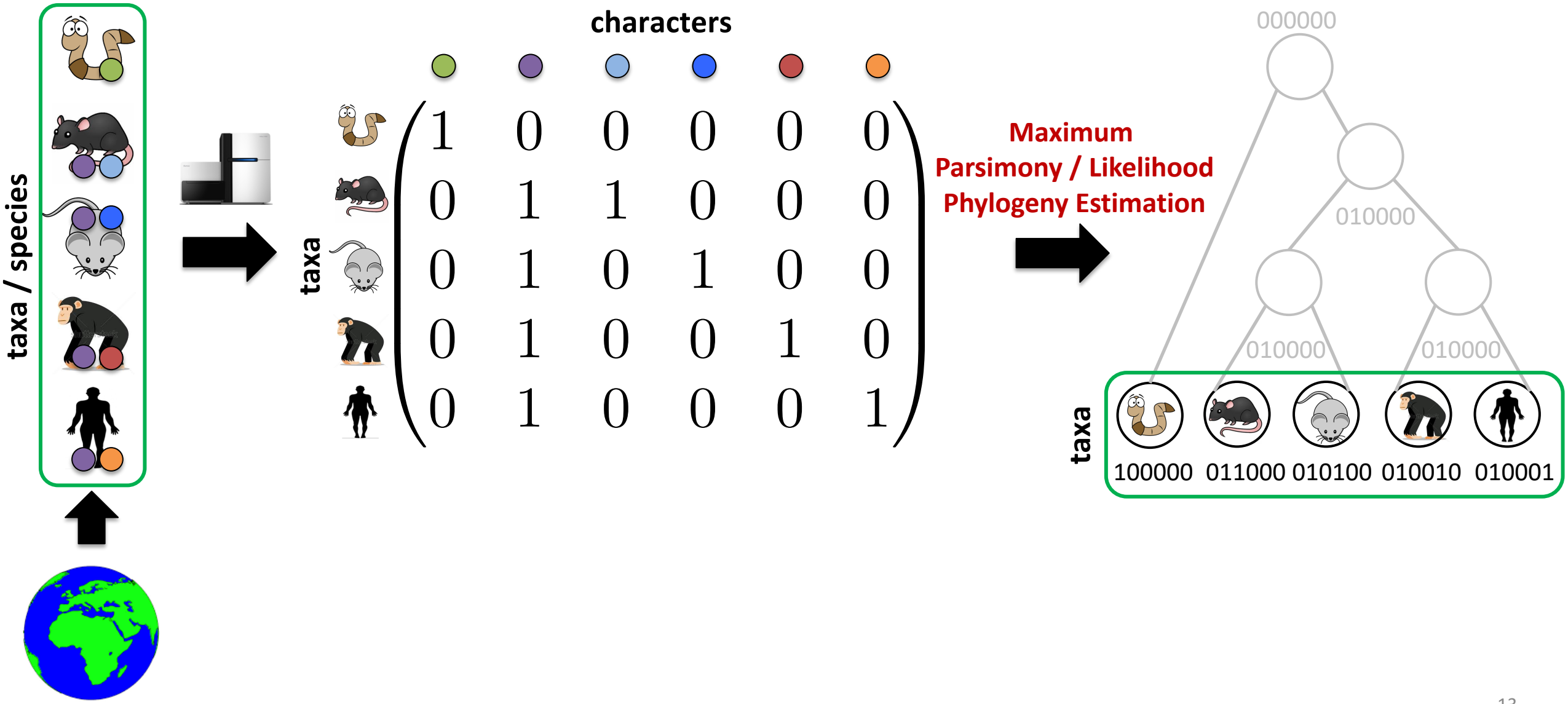
Tumor Sequencing



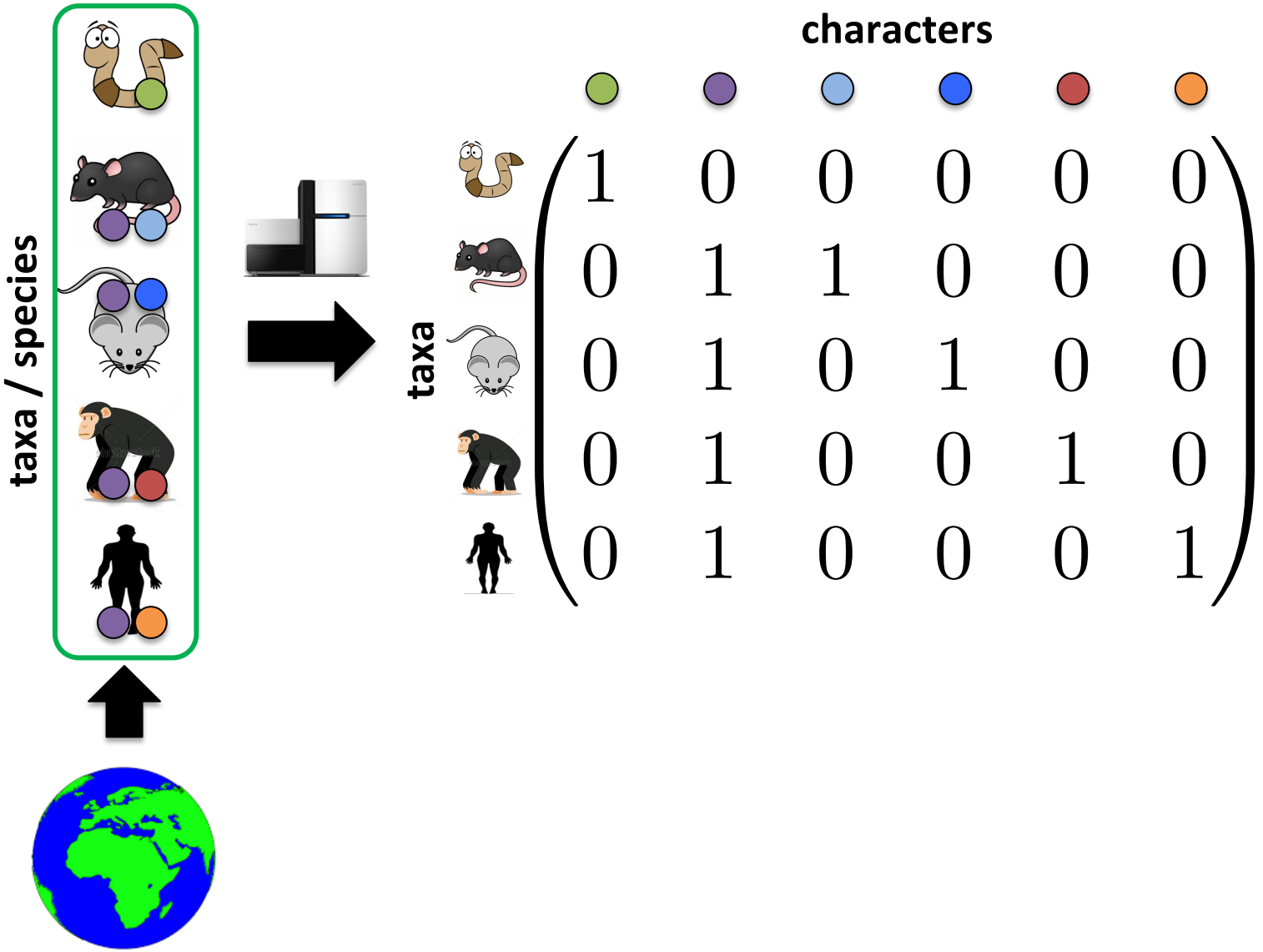
Sequencing and Tumor Phylogeny Estimation



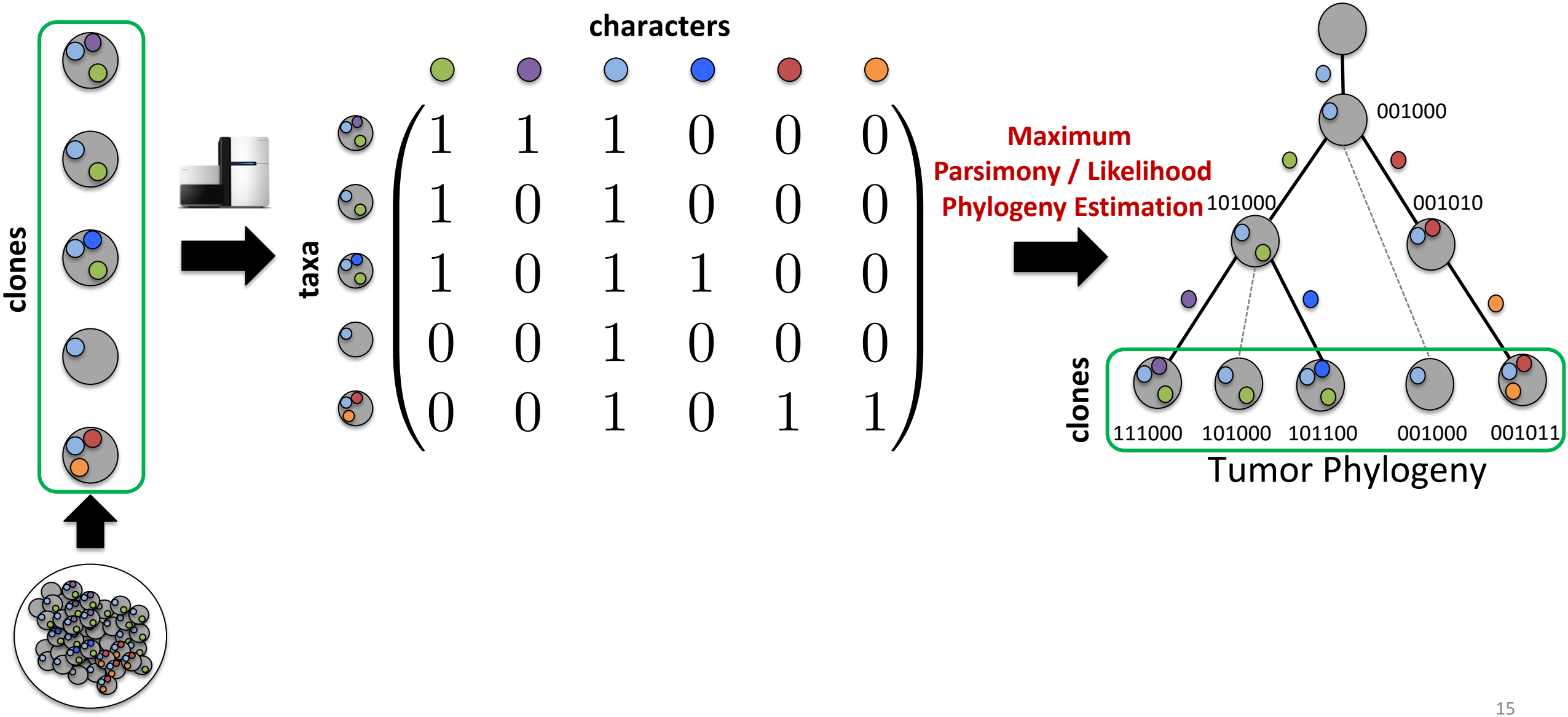
Sequencing and Tumor Phylogeny Estimation



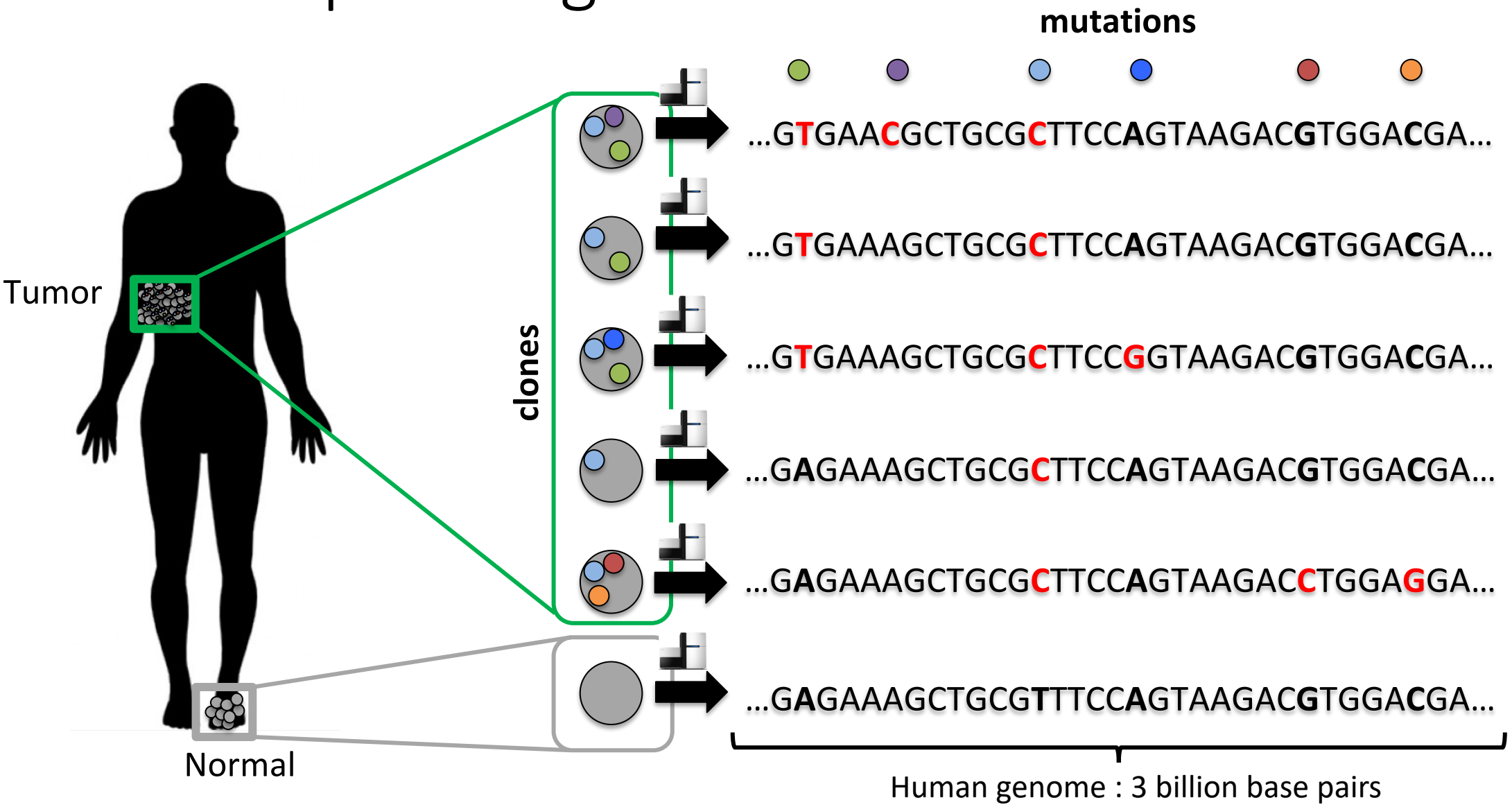
Sequencing and Tumor Phylogeny Estimation



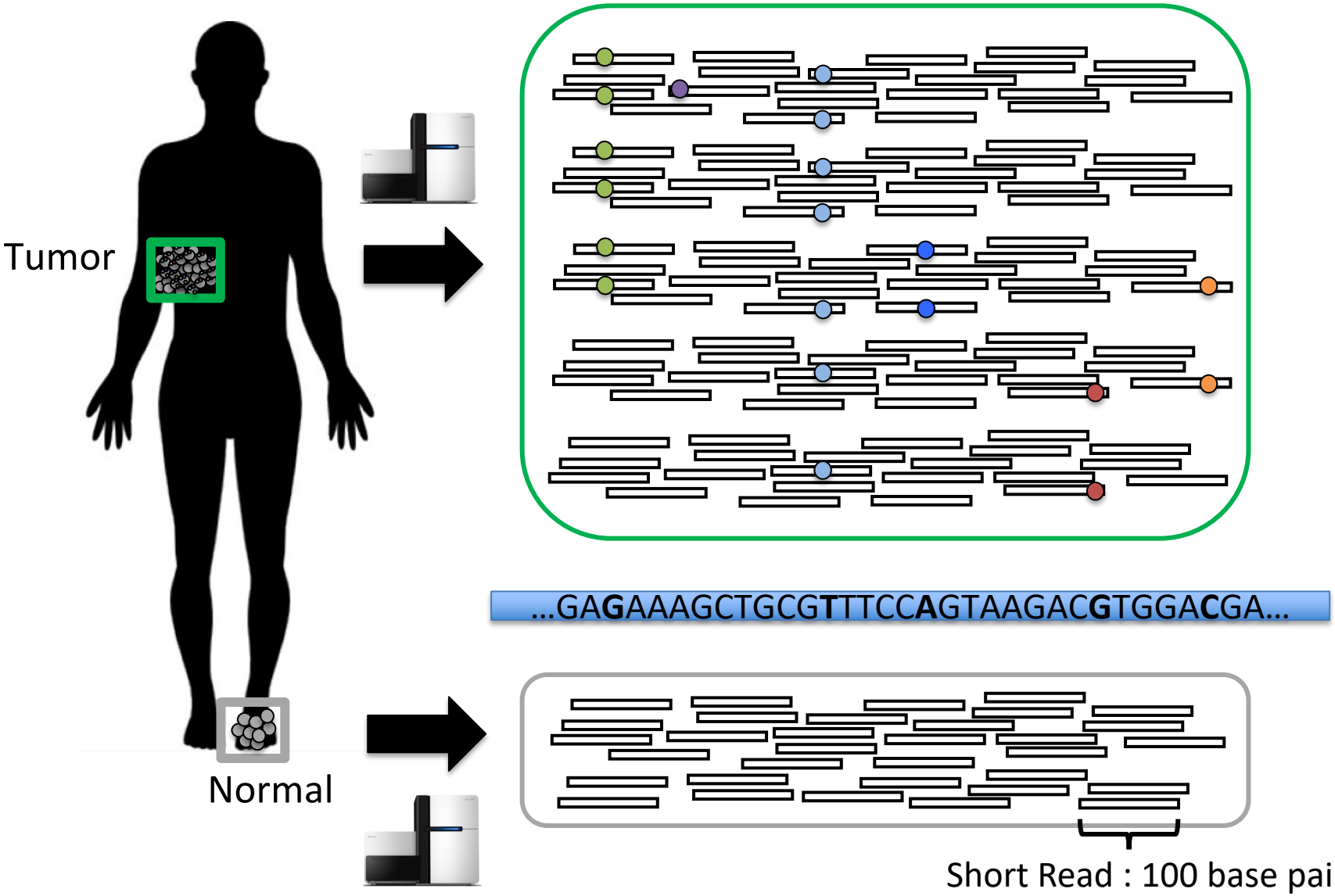
Sequencing and Tumor Phylogeny Estimation



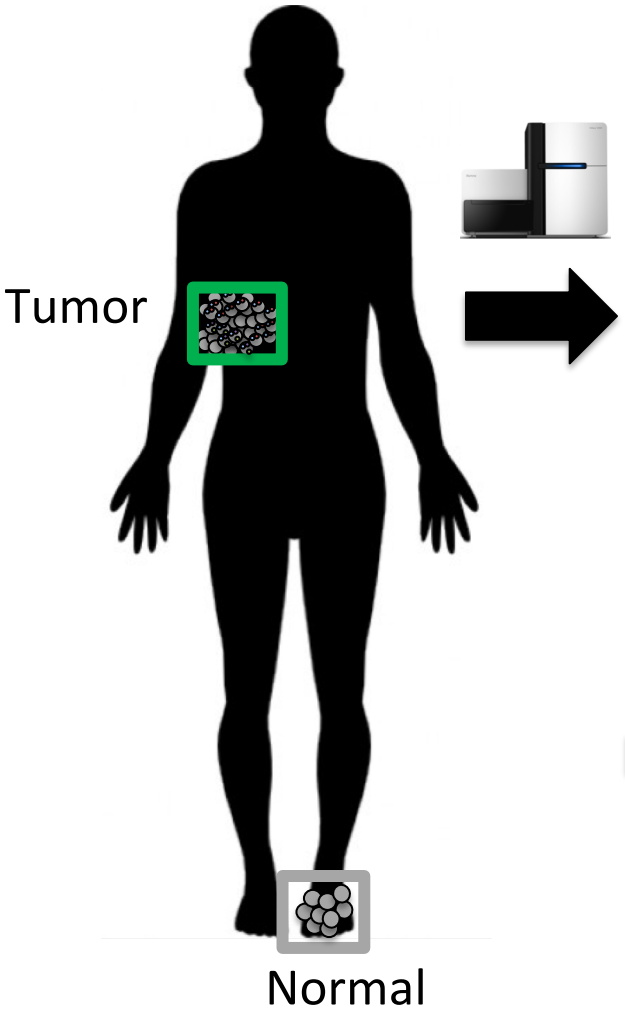
Tumor Sequencing

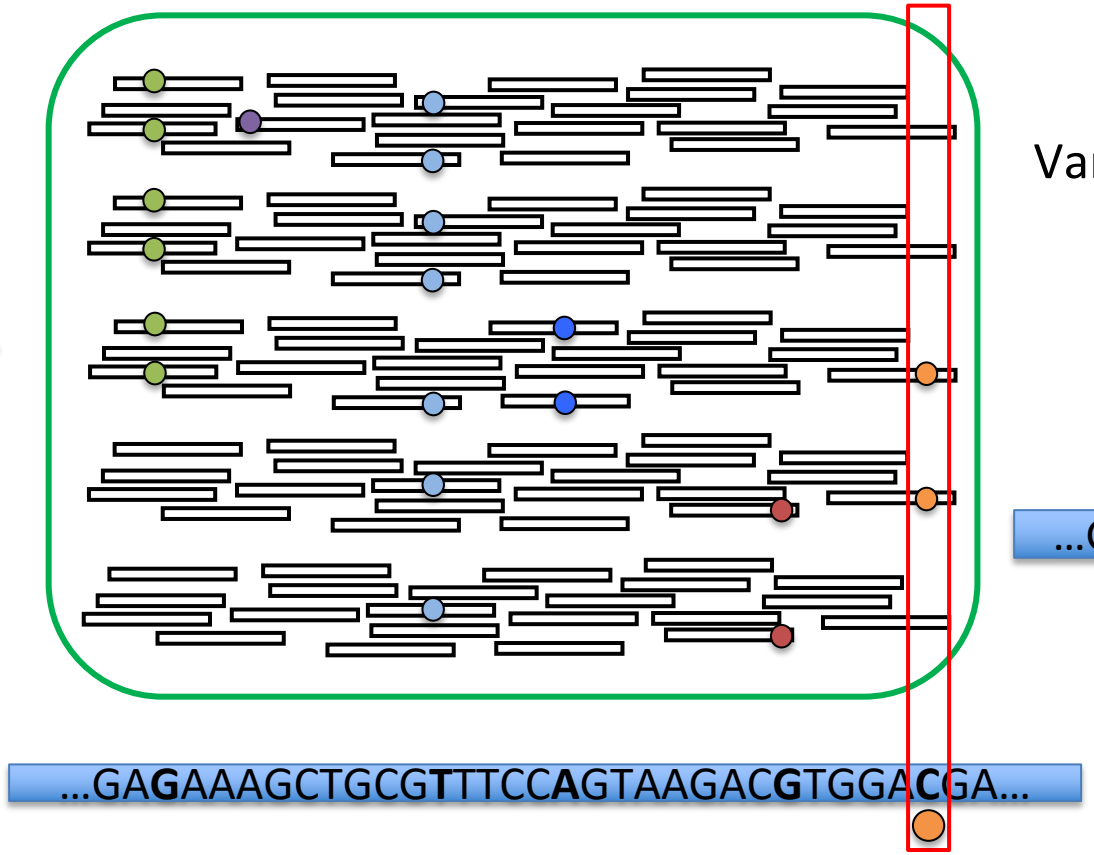


Sequencing and Tumor Phylogeny Estimation

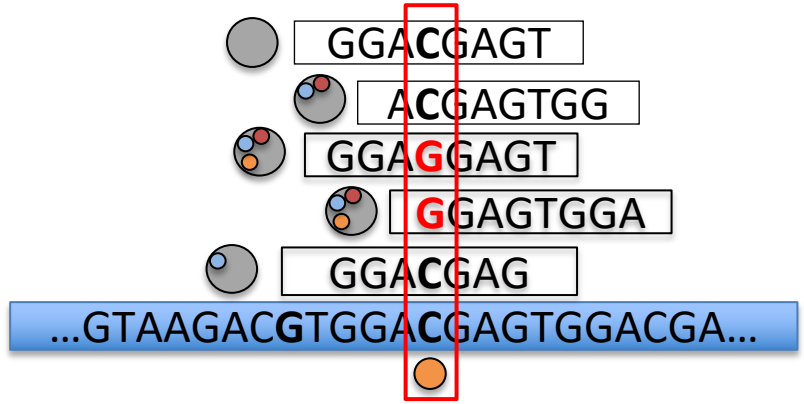


Sequencing and Tumor Phylogeny Estimation

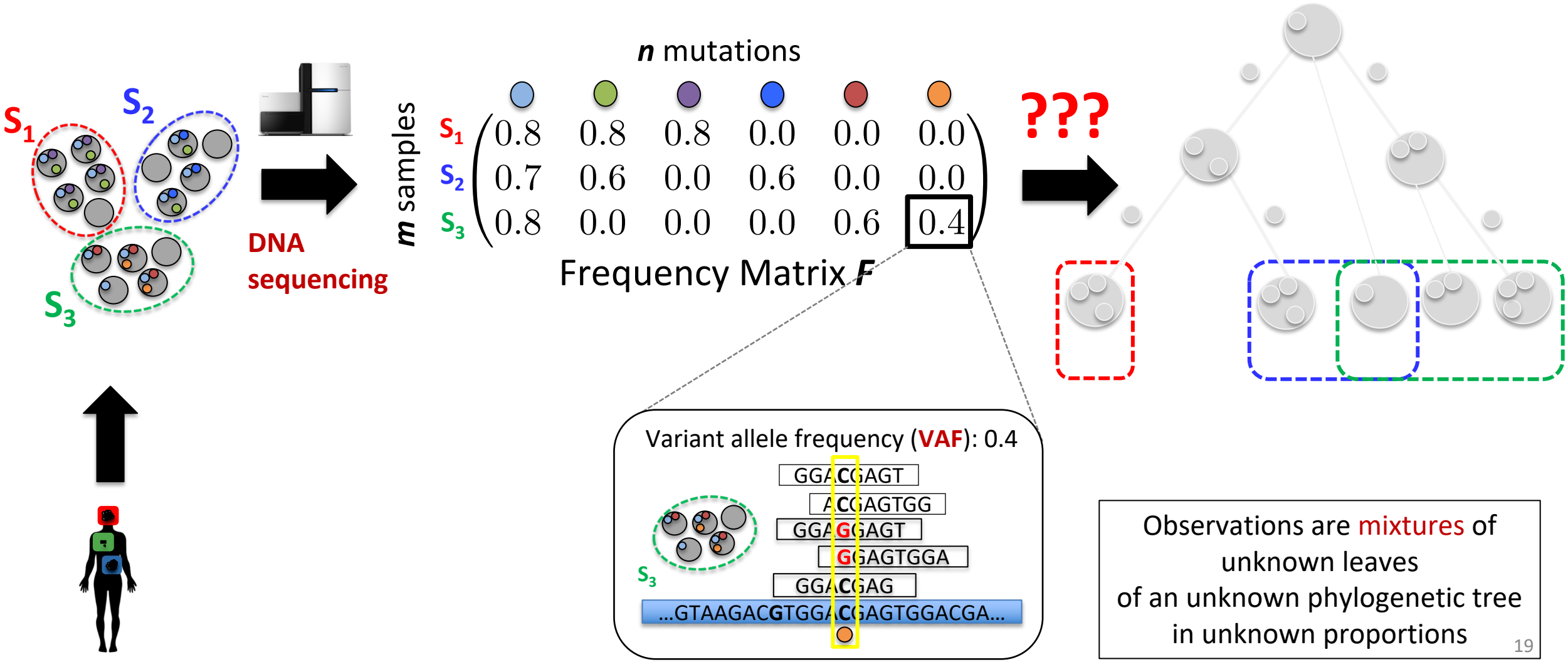




Variant allele frequency (**VAF**): 0.4

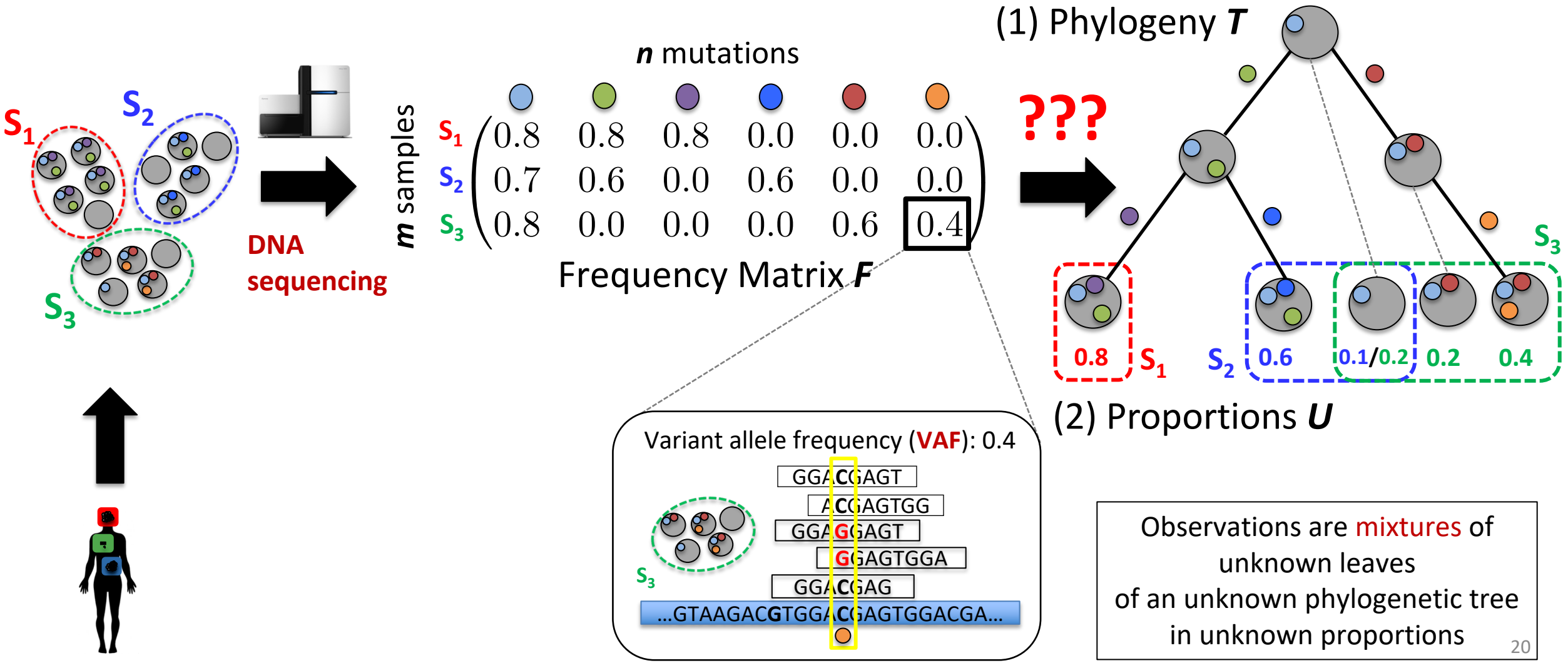


Sequencing and Tumor Phylogeny Estimation



Sequencing and Tumor Phylogeny Estimation

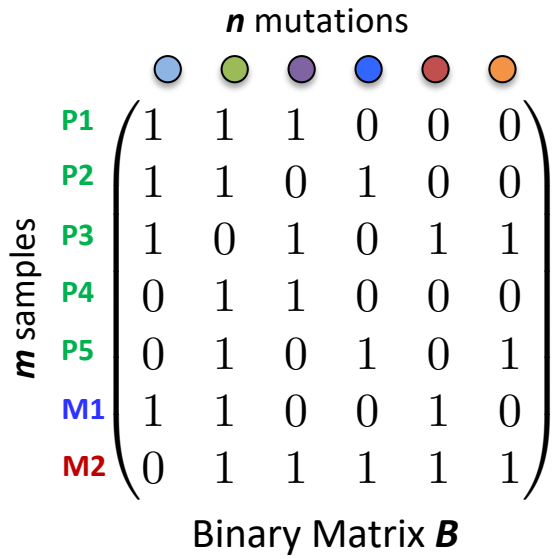
Tumor Phylogeny Estimation: Given frequencies ***F***, find (1) phylogeny ***T*** and (2) proportions ***U***



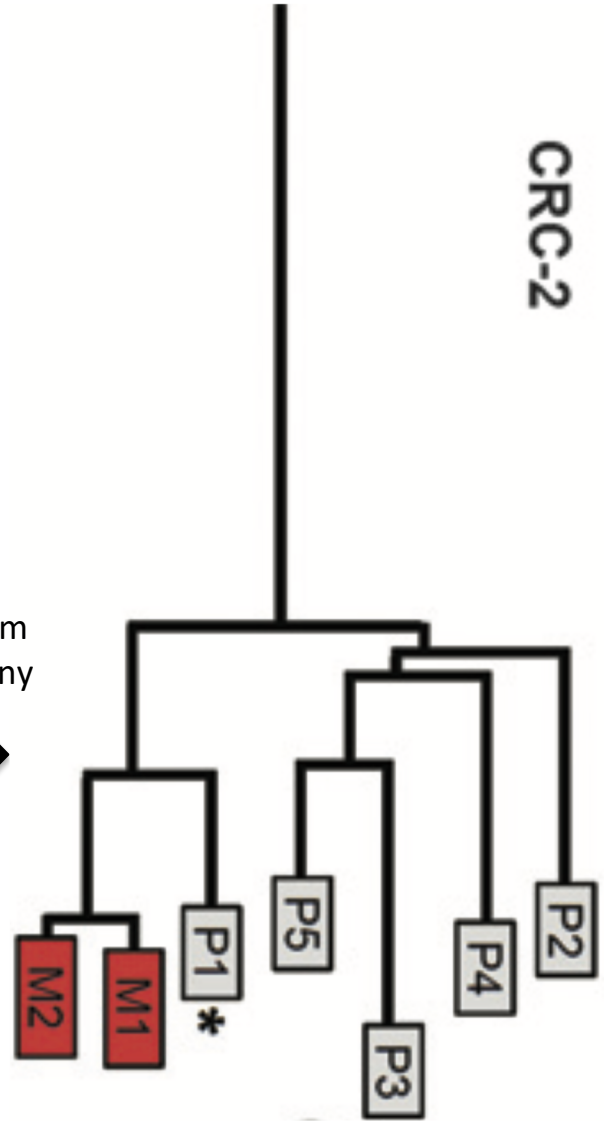
Tumor Phylogeny Estimation

Metastatic Colorectal Cancer (Patient CRC2)
[Kim et al., *Clin Cancer Res* 21(19), 2015]:

- 5 primary samples (P1-P5)
- 2 metastases (M1-M2)
- 412 single-nucleotide variants (SNVs)



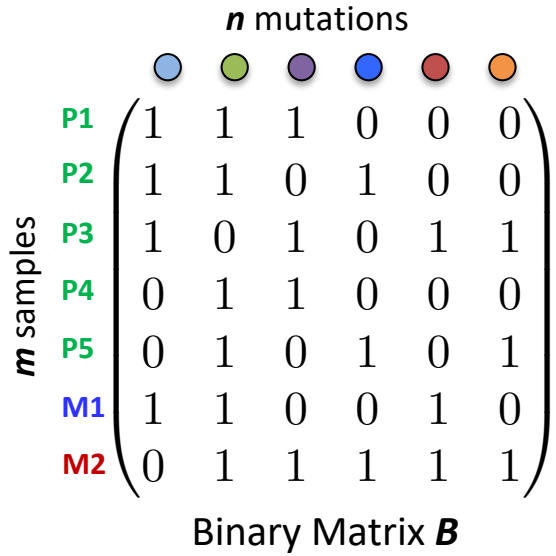
Maximum Parsimony



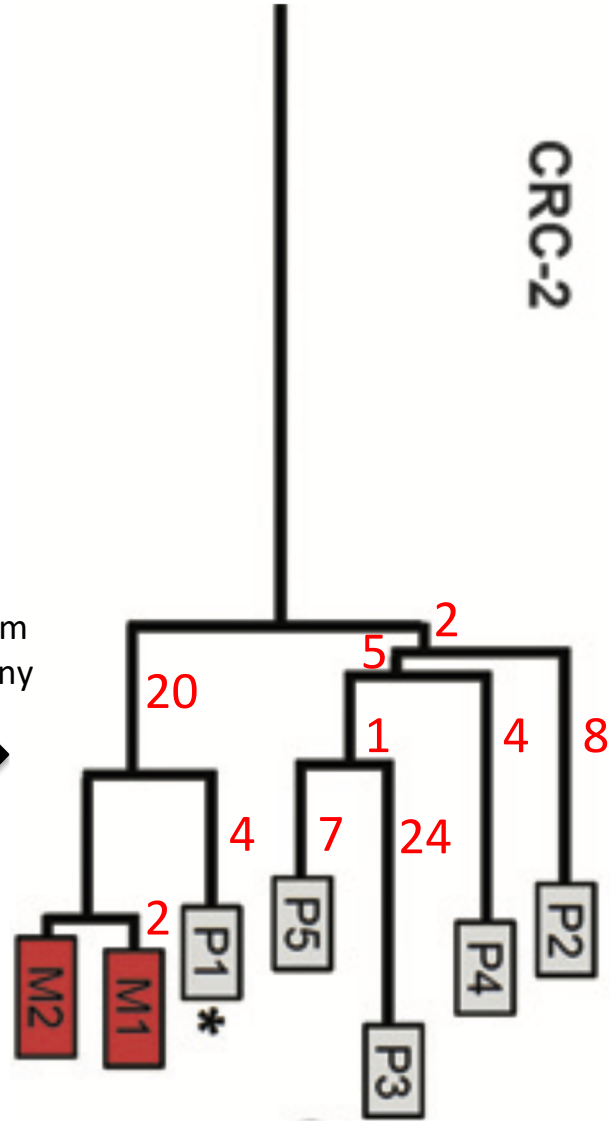
Tumor Phylogeny Estimation

Metastatic Colorectal Cancer (Patient CRC2)
[Kim et al., *Clin Cancer Res* 21(19), 2015]:

- 5 primary samples (P1-P5)
- 2 metastases (M1-M2)
- 412 single-nucleotide variants (SNVs)
- 41 mutate more than once (**homoplasy**)



Maximum Parsimony



Heuristic for Tumor Phylogeny Estimation

Metastatic Colorectal Cancer (Patient CRC2)
[Kim et al., *Clin Cancer Res* 21(19), 2015]:

- 5 primary samples (P1-P5)
- 2 metastases (M1-M2)
- 412 single-nucleotide variants (SNVs)
- 41 mutate more than once (**homoplasy**)

n mutations

m samples

P1	0.80	0.62	0.53	0.03	0.01	0.02
P2	0.45	0.34	0.04	0.43	0.02	0.02
P3	0.66	0.04	0.89	0.02	0.15	0.30
P4	0.04	0.72	0.54	0.01	0.01	0.00
P5	0.02	0.67	0.03	0.06	0.03	0.11
M1	0.74	0.43	0.01	0.02	0.14	0.00
M2	0.01	0.54	0.15	0.25	0.32	0.38

Frequency Matrix ***F***

Discretize

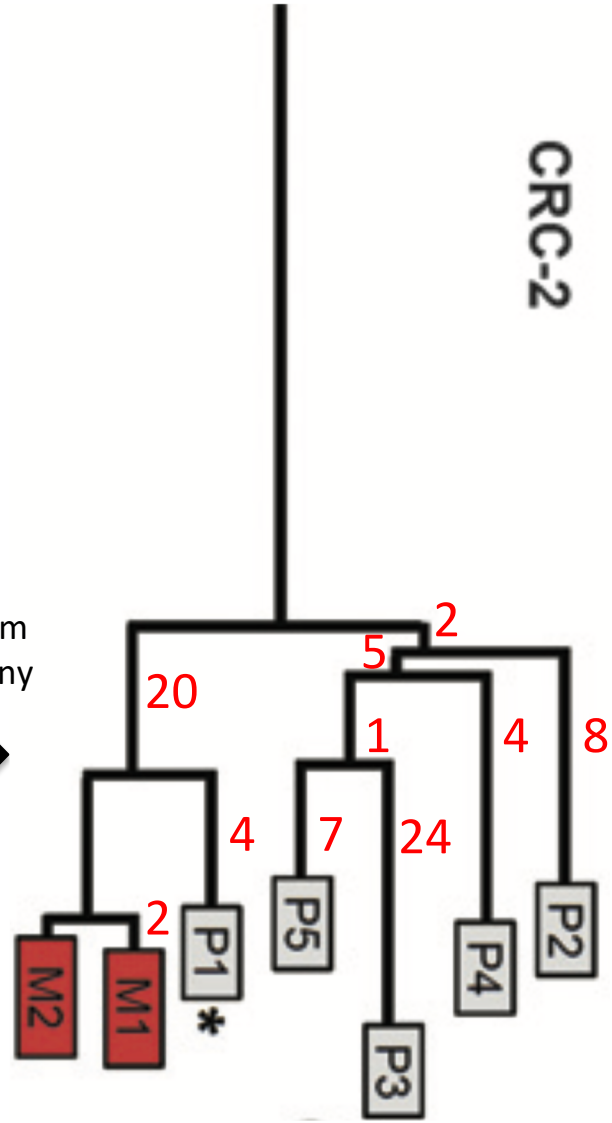
n mutations

m samples

P1	1	1	1	0	0	0
P2	1	1	0	1	0	0
P3	1	0	1	0	1	1
P4	0	1	1	0	0	0
P5	0	1	0	1	0	1
M1	1	1	0	0	1	0
M2	0	1	1	1	1	1

Binary Matrix ***B***

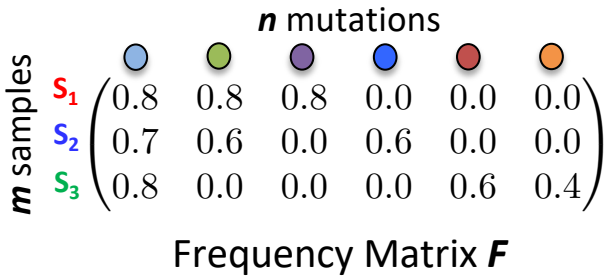
Maximum Parsimony



Resulting sample tree is **not** representative of the division/mutation history or the migration history

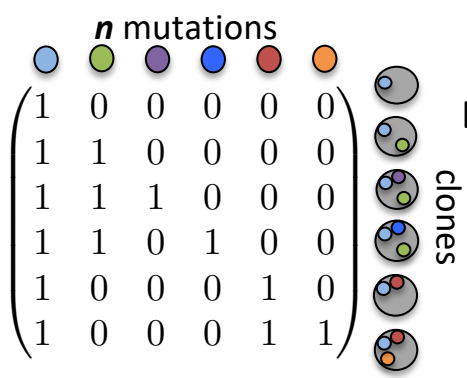
Two-State Perfect Phylogeny (PP) Mixtures

Assumptions:
- Infinite sites assumption: a character changes state once
- Error-free data

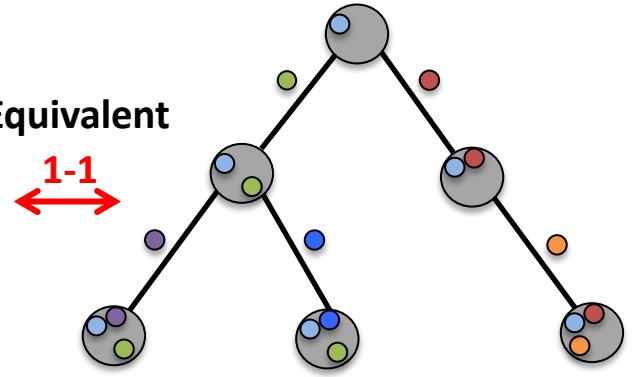


Frequency Matrix ***F***

Tumor Phylogeny Estimation: Given frequencies ***F***, find (1) phylogeny ***T*** and (2) proportions ***U***



Restricted PP Matrix ***B***



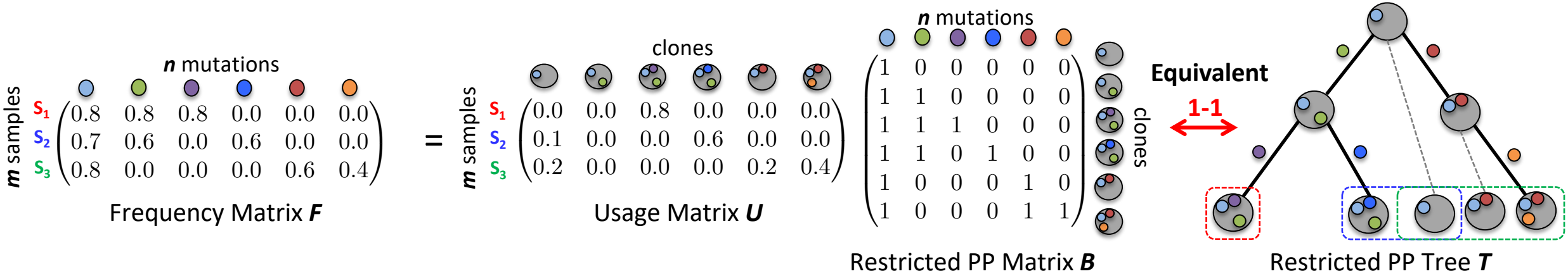
Restricted PP Tree ***T***

Perfect Phylogeny Theorem [Estabrook, 1971] [Gusfield, 1991]

Two-State Perfect Phylogeny (PP) Mixtures

Assumptions:
- Infinite sites assumption: a character changes state once
- Error-free data

Tumor Phylogeny Estimation: Given frequencies ***F***, find (1) phylogeny ***T*** and (2) proportions ***U***



Rows of ***U*** are mixture proportions

$$u_{pj} \geq 0 \text{ and } \sum_j u_{pj} \leq 1$$

Perfect Phylogeny Theorem [Estabrook, 1971] [Gusfield, 1991]

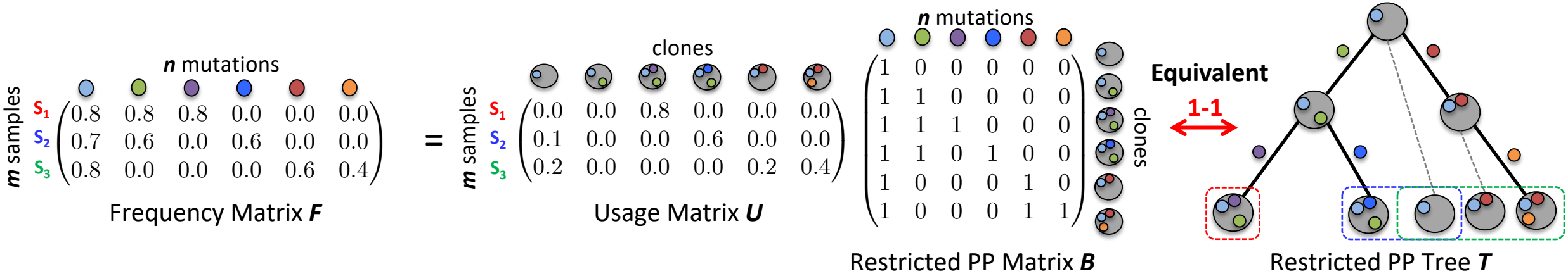
VAF Factorization Problem (VAFFP) [El-Kebir*, Oesper* et al., 2015]

Given ***F***, find ***U*** and ***B*** such that ***F = U B***

Two-State Perfect Phylogeny (PP) Mixtures

Assumptions:
- Infinite sites assumption: a character changes state once
- **Error-free data**

Variant of VAFFP:
TrAp [Strino *et al.*, 2013], PhyloSub [Jiao *et al.*, 2014]
CITUP [Malikic *et al.*, 2015], BitPhylogeny [Yuan *et al.*, 2015]
LICHeE [Popic e*t al.*, 2015], …



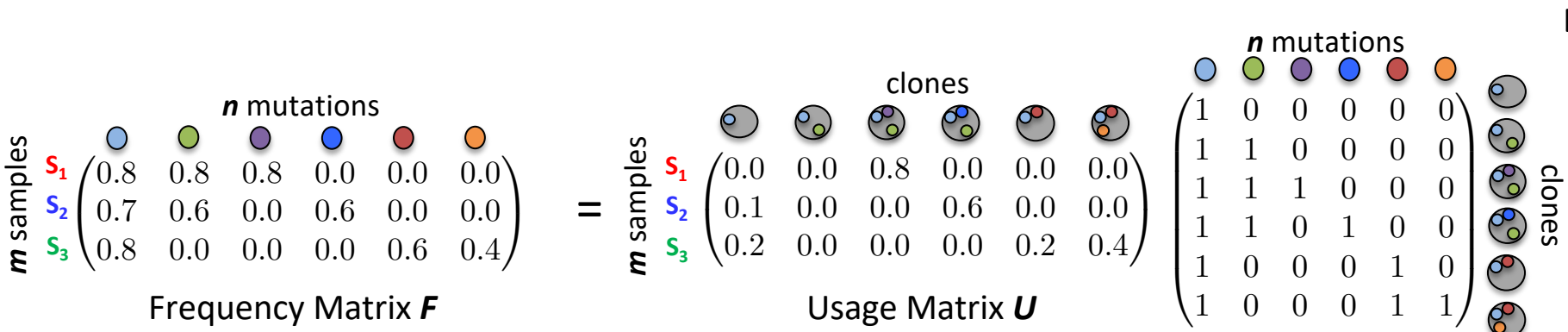
VAF Factorization Problem (VAFFP) [El-Kebir*, Oesper* et al., 2015]

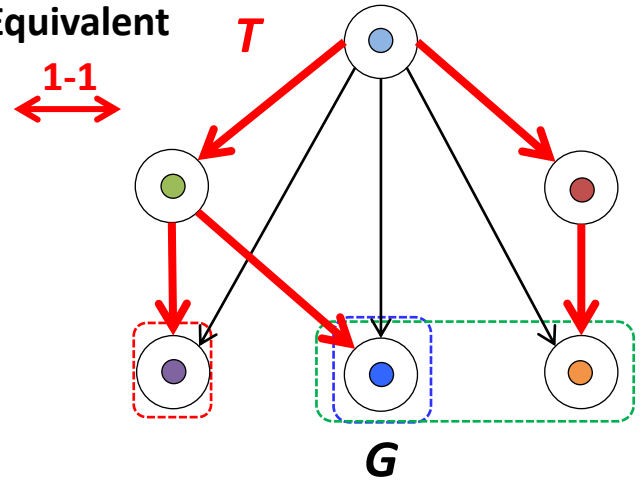
Given ***F***, find ***U*** and ***B*** such that ***F = U B***

Two-State Perfect Phylogeny (PP) Mixtures

Assumptions:
- Infinite sites assumption: a character changes state once
- **Error-free data**

Sum Condition:
Given ***F*** and ***T***, $f_{pj} \geq \sum_{k \text{ child of } j} f_{pk}$





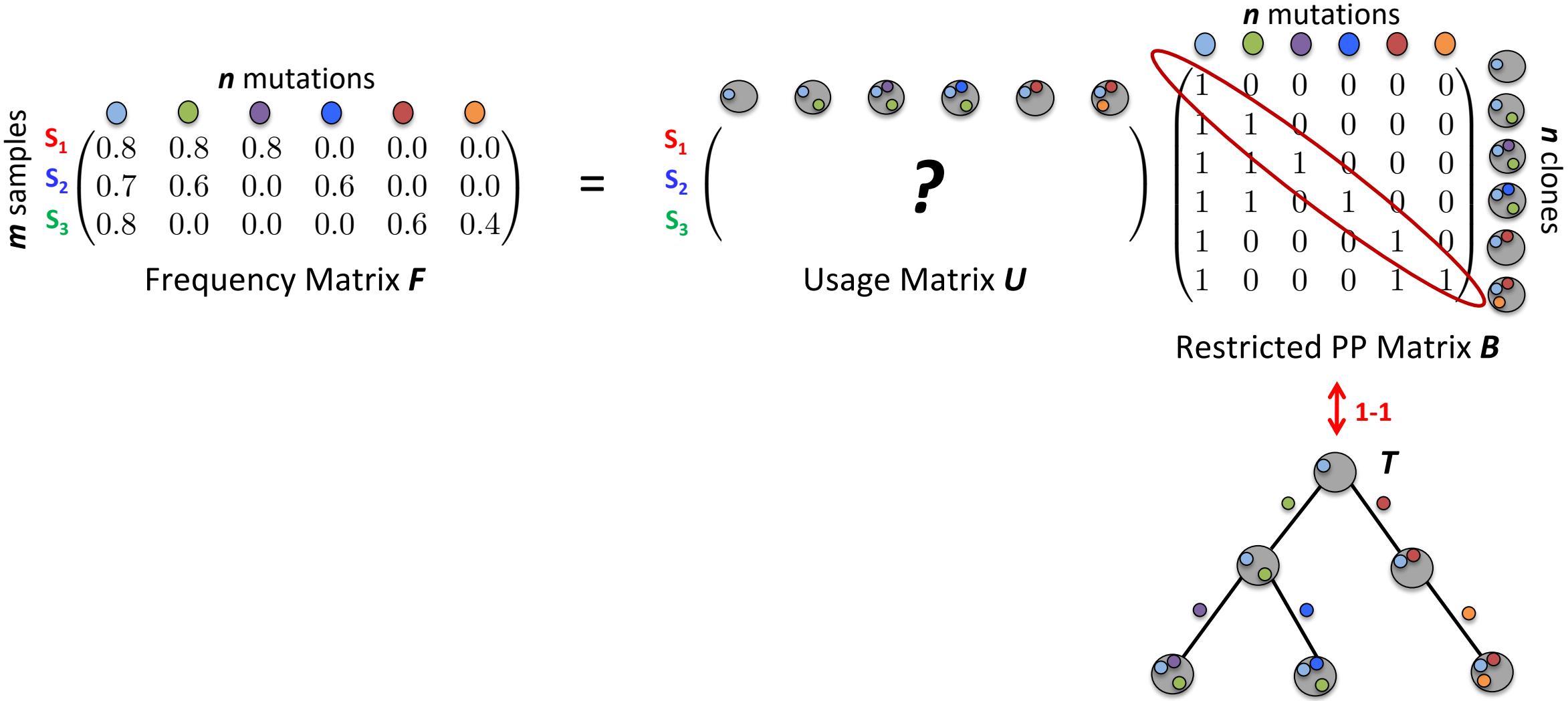
Theorem 1:
T is a solution to the VAFFP if and only if ***T*** is a spanning tree of ***G*** satisfying the sum condition

Theorem 2:
VAFFP is NP-complete

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

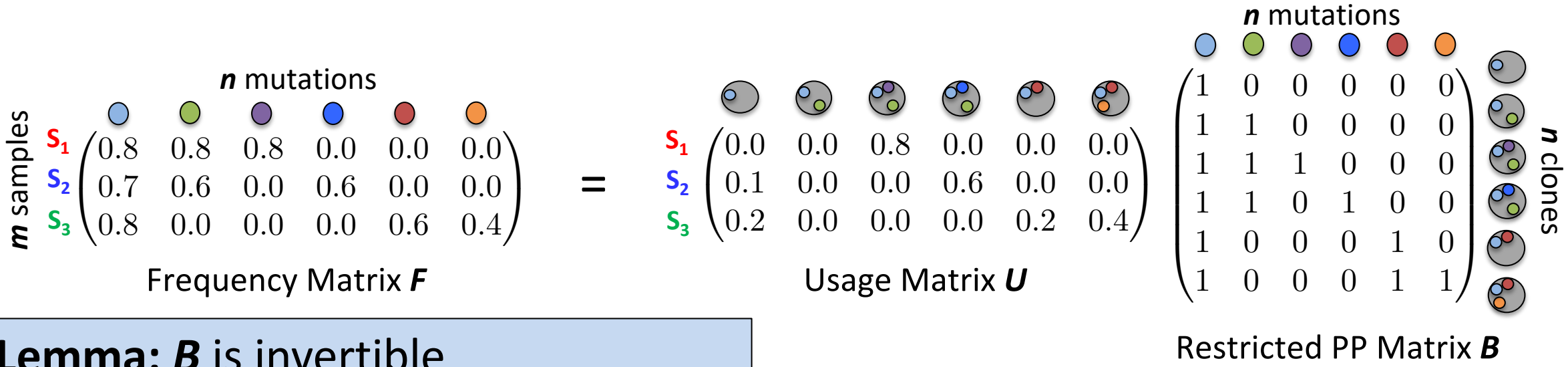
Given *F* and *T* (or *B*), is there a usage matrix *U*?

VAFFP: Given ***F***, find ***U*** and ***B*** such that ***F*** = ***U B***



Given *F* and *T* (or *B*), is there a usage matrix *U*?

VAFFP: Given ***F***, find ***U*** and ***B*** such that ***F*** = ***U B***



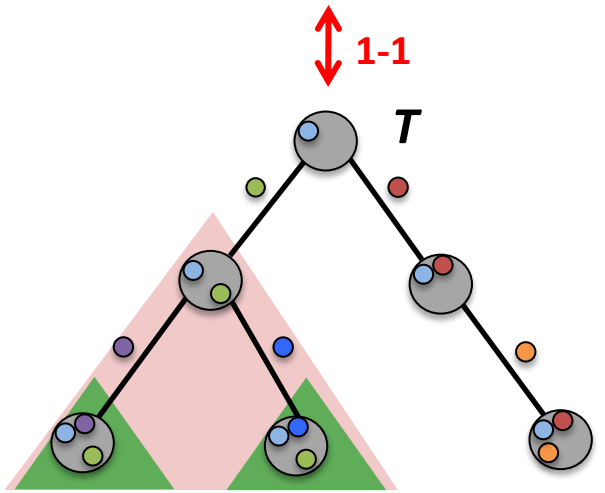
Frequency Matrix ***F*** Usage Matrix ***U*** Restricted PP Matrix ***B***

Lemma: *B* is invertible

➡ Given ***F*** and ***B***, ***U*** is **unique**: $U = F B^{-1}$

Lemma:

$$u_{pj} = f_{pj} - \sum_{k \text{ child of } j} f_{pk}$$



Given *F* and *T* (or *B*), is there a usage matrix *U*?

VAFFP: Given ***F***, find ***U*** and ***B*** such that ***F*** = ***U B***

$$\begin{pmatrix} \boxed{0.15} & 0.1 & 0.06 \\ 0.2 & 0.05 & 0.04 \end{pmatrix} = \begin{pmatrix} \boxed{-0.01} & 0.1 & 0.06 \\ 0.11 & 0.05 & 0.04 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

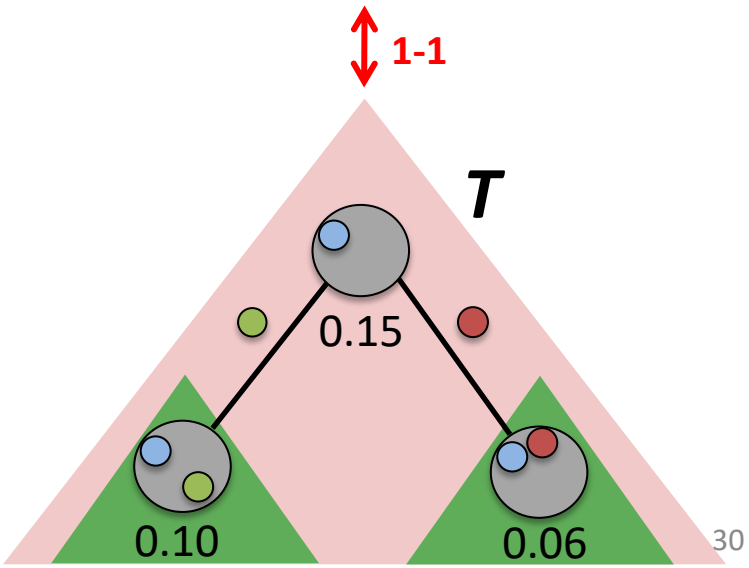
Frequency Matrix ***F***

Usage Matrix ***U***

Restricted PP Matrix ***B***

Lemma:

$$u_{pj} = \boxed{f_{pj}} - \boxed{\sum_{k \text{ child of } j} f_{pk}}$$



Combinatorial Characterization of Solutions

Lemma:

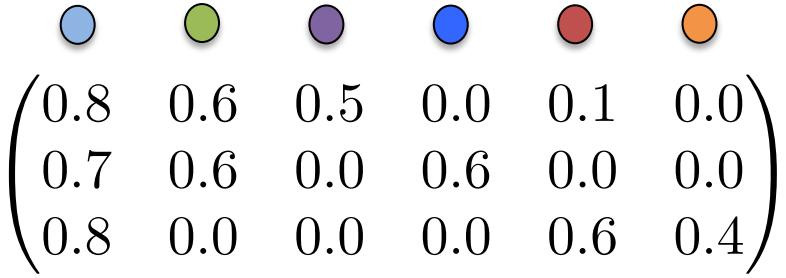
$$u_{pj} = f_{pj} - \sum_{k \text{ child of } j} f_{pk}$$

Lemma (Sum Condition):

Given ***F*** and ***T***, for all samples *p* and mutations *j*,

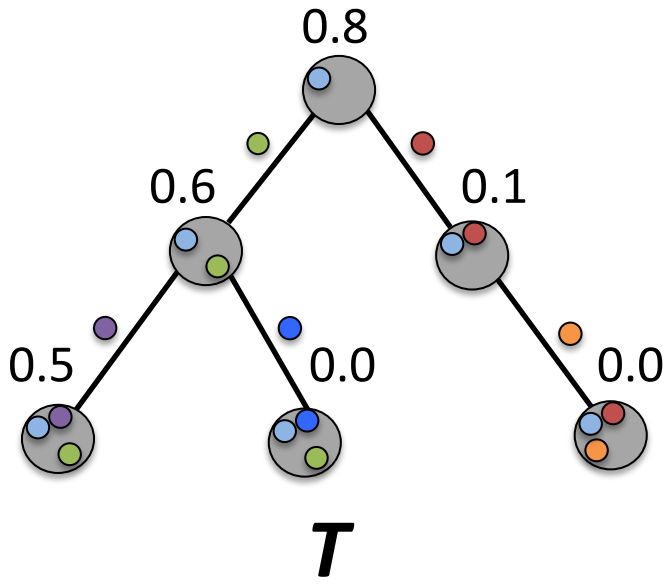
$$f_{pj} \geq \sum_{k \text{ child of } j} f_{pk}$$

necessary
sufficient



$$\begin{pmatrix} 0.8 & 0.6 & 0.5 & 0.0 & 0.1 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

F



T

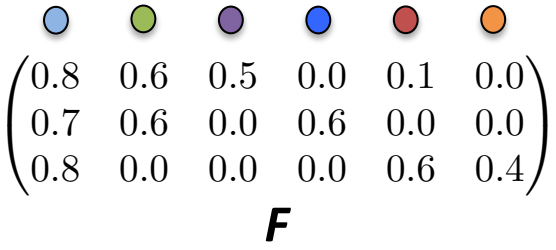
Combinatorial Characterization of Solutions

Lemma (Sum Condition):
Given ***F*** and ***T***, for all samples *p* and mutations *j*, $f_{pj} \geq \sum_{k \text{ child of } j} f_{pk}$

necessary
sufficient

Lemma (Ancestry Condition):
Given ***F*** and ***T***, for all samples *p* and mutations *k* child of *j*, $f_{pj} \geq f_{pk}$

necessary



$$\begin{pmatrix} 0.8 & 0.6 & 0.5 & 0.0 & 0.1 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

F

Combinatorial Characterization of Solutions

Lemma (Sum Condition):
Given ***F*** and ***T***, for all samples *p* and mutations *j*, $f_{pj} \geq \sum_{k \text{ child of } j} f_{pk}$

necessary
sufficient

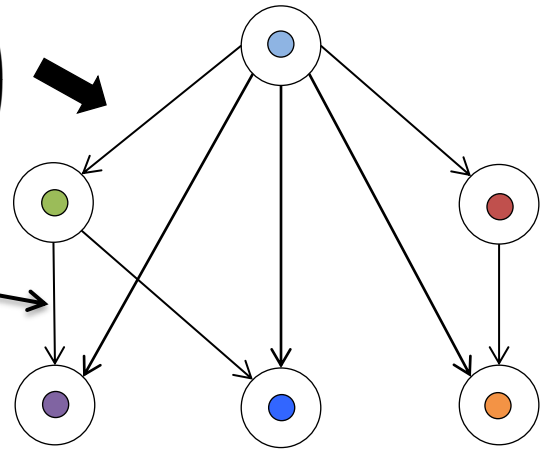
Lemma (Ancestry Condition):
Given ***F*** and ***T***, for all samples *p* and mutations *k* child of *j*, $f_{pj} \geq f_{pk}$

necessary

$$\begin{pmatrix} 0.8 & 0.6 & 0.5 & 0.0 & 0.1 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

F

potential parental relationship



Ancestry Graph ***G*** = (***V, A***)

Ancestry graph *G* = (*V*, *A*); given ***F***

- Vertex for every mutation
- Edge $(j, k) \in A$ iff $f_{pj} \geq f_{pk}$ for all samples *p*

Combinatorial Characterization of Solutions

Lemma (Sum Condition):
Given ***F*** and ***T***, for all samples *p* and mutations *j*, $f_{pj} \geq \sum_{k \text{ child of } j} f_{pk}$

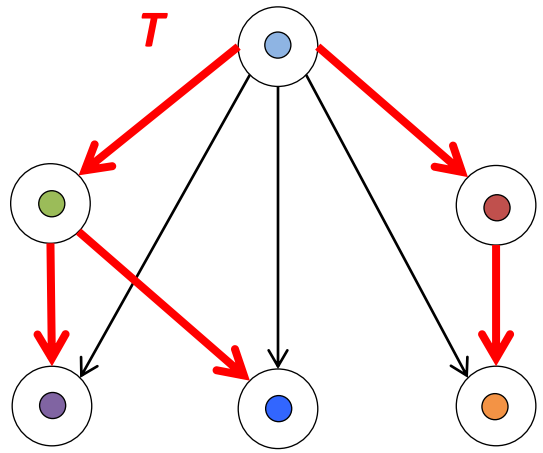
necessary
sufficient

Lemma (Ancestry Condition):
Given ***F*** and ***T***, for all samples *p* and mutations *k* child of *j*, $f_{pj} \geq f_{pk}$

necessary

$$\begin{pmatrix} 0.8 & 0.6 & 0.5 & 0.0 & 0.1 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

F



Ancestry Graph ***G*** = (***V, A***)

Ancestry graph *G* = (*V, A*); given ***F***

- Vertex for every mutation
- Edge $(j, k) \in A$ iff $f_{pj} \geq f_{pk}$ for all samples *p*

Theorem 1:
T is a solution to the VAFFP if and only if ***T*** is a spanning tree of ***G*** satisfying the Sum Condition

Theorem 2:
VAFFP is NP-complete

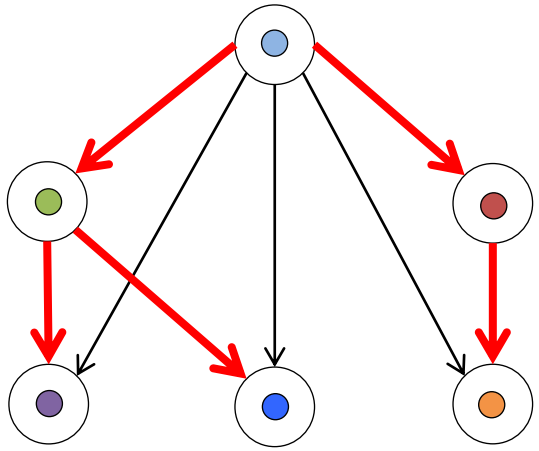
Solving the VAFFP: ILP formulation

$$\max \sum_{(v_j,v_k)\in A'} x_{jk}$$

Find the largest set of edges in *G*

$$\text{s.t.} \sum_{v_j\in\delta^+(v_r)} x_{rj} = 1$$

Exactly one root node



***G* = (*V*, *A*)**

$$x_{kl} \leq \sum_{v_j\in\delta^-(v_k)} x_{jk} \qquad \forall (v_k, v_l) \in A$$

Connectivity

$$\sum_{v_j\in\delta^-(v_k)} x_{jk} \leq 1 \qquad \forall v_k \in V$$

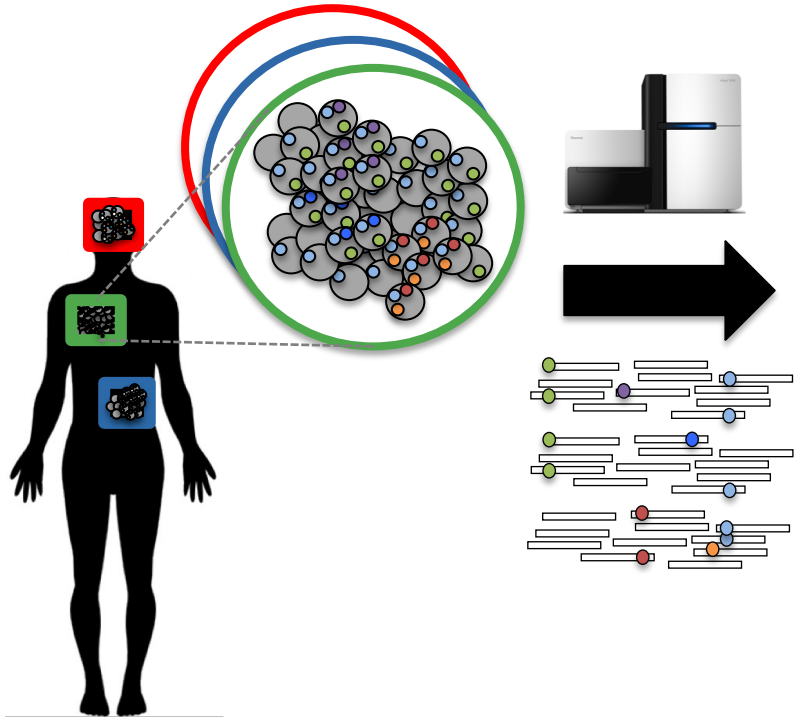
Tree

$$\sum_{v_j\in\delta^-(v_k)} f_{pk}x_{jk} \geq \sum_{v_l\in\delta^+(v_k)} f_{pl}x_{kl} \qquad \forall p \in [m],\, v_k \in V$$

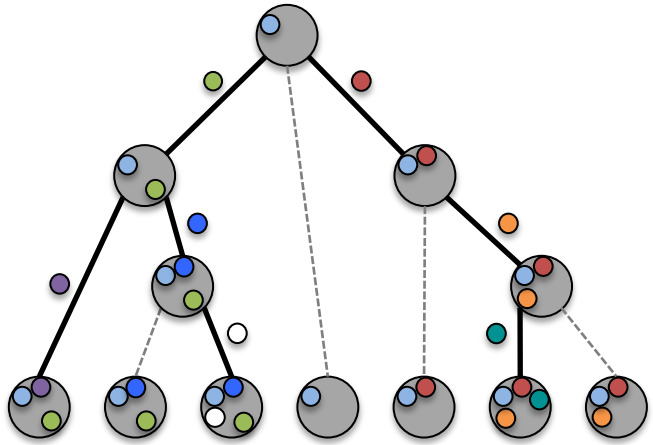
Sum condition

$$x_{jk} \in \{0,1\} \qquad \forall (v_j, v_k) \in A'$$

Outline



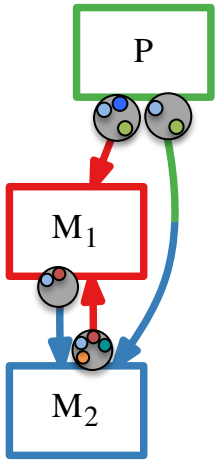
Cell division and mutation



Tumor Phylogeny Estimation

[**El-Kebir***, Oesper* et al., *ISMB* 2015/*Bioinformatics*]

Cell migration



Tumor Migration Analysis

Biological Problem

Mixed tumor samples require specialized phylogeny algorithms

Assumptions

- Infinite sites assumption
- ***Error-free data***

Computational Problem

VAF Factorization Problem (VAFFP)
Given ***F***, find ***U*** and ***B*** such that ***F = U B***

Approach

- Combinatorial characterization as constrained spanning trees in a directed acyclic graph
- Integer linear programming

Probabilistic Model for Noisy Measurements

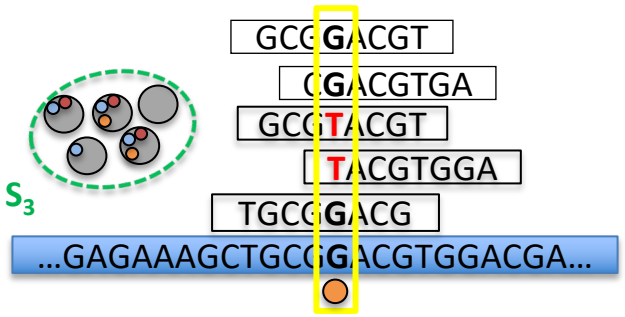
mutations

$$\boldsymbol{F} = [f_{pi}] = \begin{pmatrix} 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

S_1, S_2, S_3 — samples

VAF of mutation *i* in sample *p*

Variant allele frequency (**VAF**): 0.4



$$\text{VAF}(\bullet) = \frac{\text{\# reads with } \bullet}{\text{\# reads}}$$

mutations

$$[\boldsymbol{X}_{pi}] = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$$

samples

VAF posterior distribution (beta) given the reads of mutation *i* in sample *p*

Probabilistic Model for Noisy Measurements

mutations

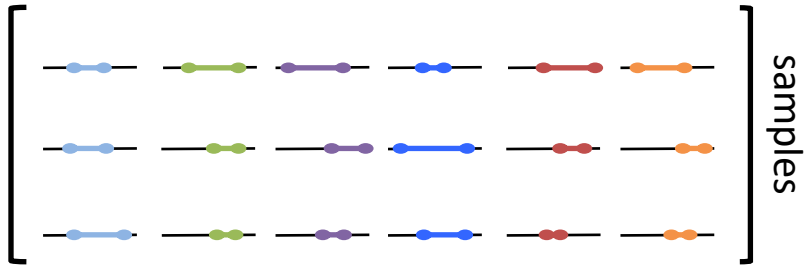
$$F^- = [f_{pi}^-] = \begin{pmatrix} 0.75 & 0.78 & 0.77 & 0.0 & 0.0 & 0.0 \\ 0.55 & 0.43 & 0.0 & 0.54 & 0.0 & 0.0 \\ 0.56 & 0.0 & 0.0 & 0.0 & 0.57 & 0.34 \end{pmatrix}$$ S_1 S_2 S_3 samples

Interval VAF Factorization Problem (I-VAFFP)

Given F^- and F^+, find F, U and B such that $F = U\,B$ and $f_{pi}^- \leq f_{pi} \leq f_{pi}^+$ for all samples p and mutations i

Consider (1 - α) confidence intervals:

mutations

$[X_{pi}] =$  samples

VAF posterior distribution (beta) given the reads of mutation *i* in sample *p*

mutations

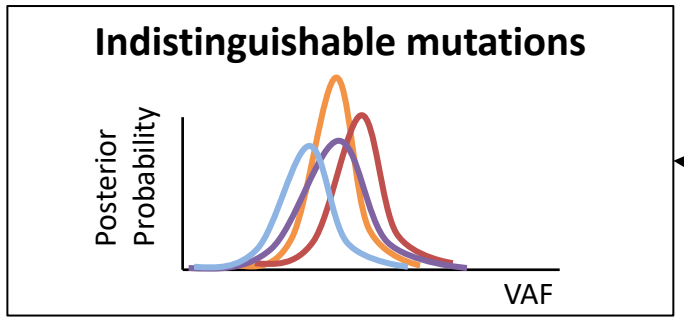
$$F^+ = [f_{pi}^+] = \begin{pmatrix} 0.9 & 0.85 & 0.87 & 0.05 & 0.0 & 0.0 \\ 0.75 & 0.65 & 0.05 & 0.68 & 0.0 & 0.0 \\ 0.83 & 0.0 & 0.04 & 0.0 & 0.67 & 0.48 \end{pmatrix}$$ S_1 S_2 S_3 samples

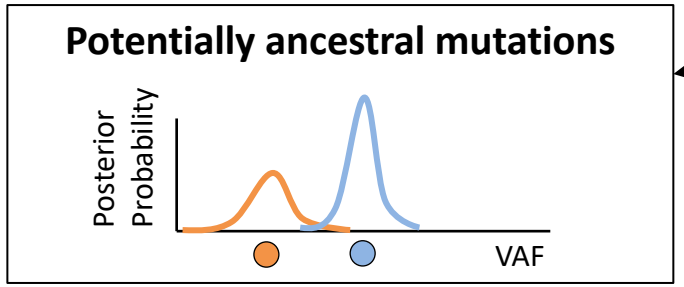
Approximate Ancestry Graph

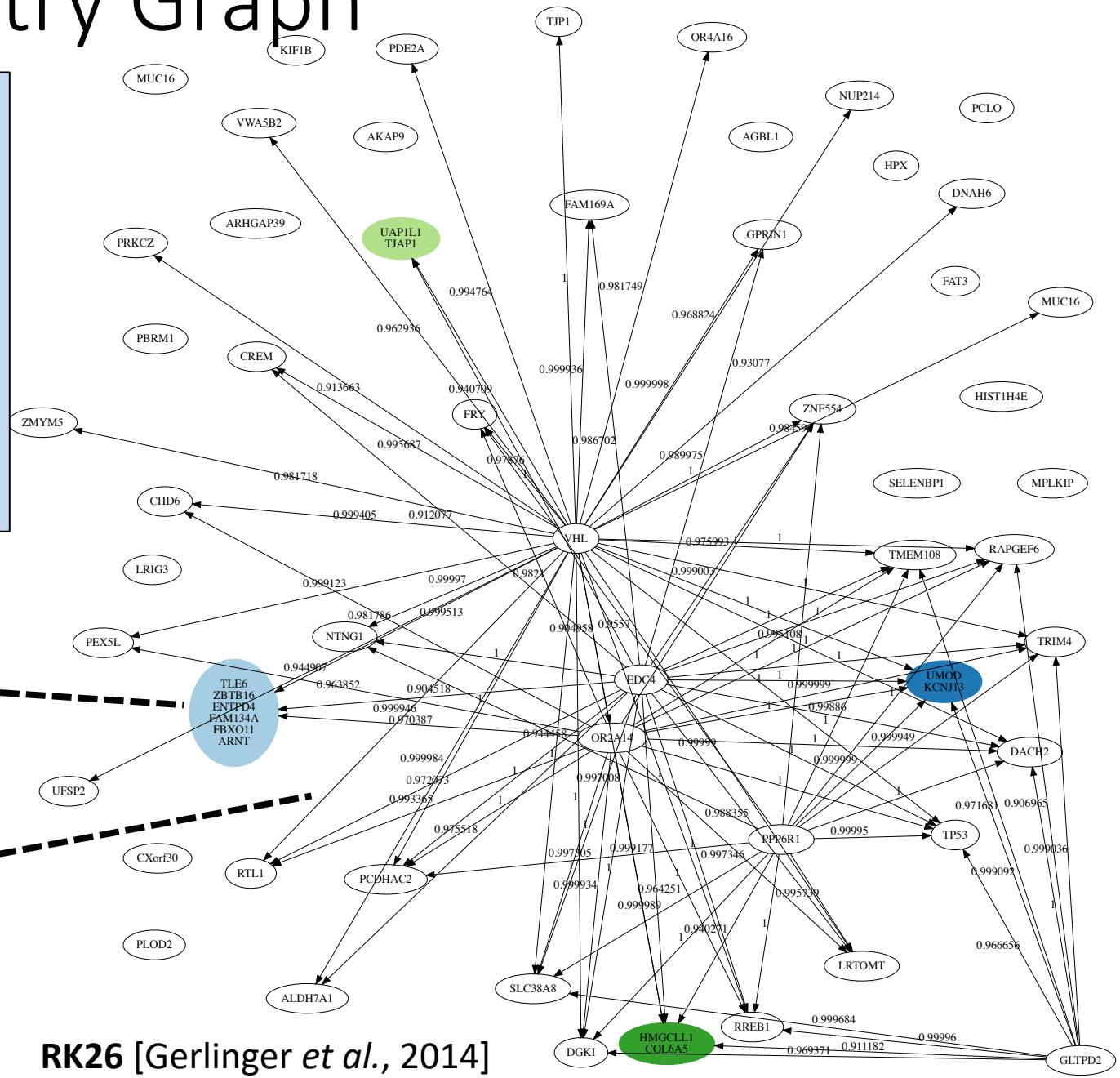
Approximate ancestry graph *G*:

V: indistinguishable mutations

E: connect vertices (*j*,*k*) with high posterior ancestral probability $\Pr[X_{pj} > X_{pk}]$





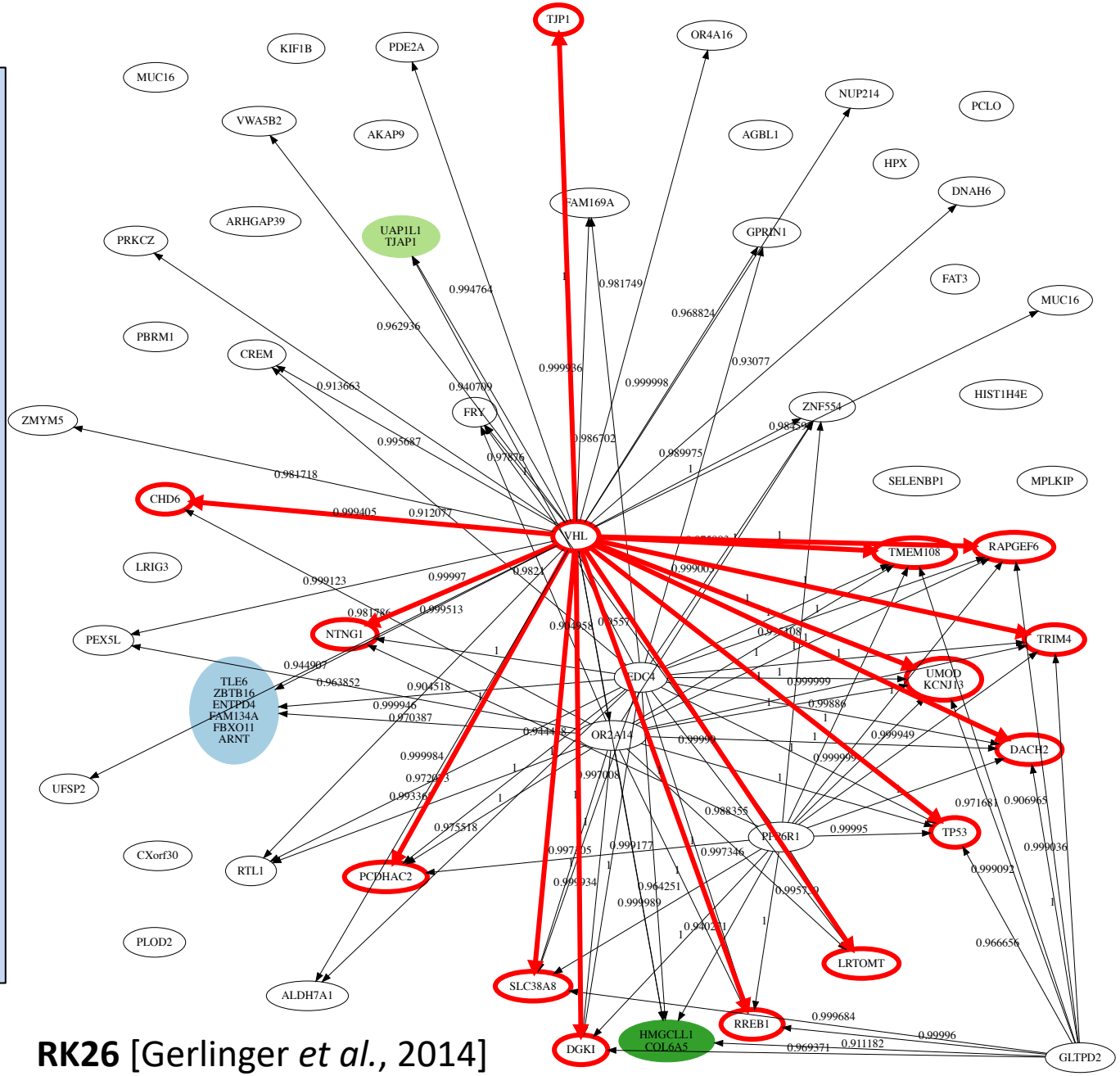


RK26 [Gerlinger *et al.*, 2014]

AncesTree

AncesTree:

1. Build approximate ancestry graph ***G***
2. Find the largest tree ***T*** in the approximate ancestry graph ***G*** and matrix ***F*** that satisfy the sum condition for ***T*** and ***F*** (mixed integer linear programming)
3. Compute the usage matrix ***U***



RK26 [Gerlinger *et al.*, 2014]

Solving the I-VAFFP: MILP formulation

$$\max \sum_{(v_j,v_k)\in A'} x_{jk} \quad -\frac{1}{mn}\sum_{p=1}^{m}\sum_{j=1}^{n}|\tilde{f}_{pj} - f_{pj}|$$

Find the largest set of edges with a tie-breaker

$$\text{s.t.} \sum_{v_j\in\delta^+(v_r)} x_{rj} = 1$$

Exactly one root node

$$x_{kl} \le \sum_{v_j\in\delta^-(v_k)} x_{jk} \qquad \forall (v_k, v_l) \in A$$

Connectivity

$$\sum_{v_j\in\delta^-(v_k)} x_{jk} \le 1 \qquad \forall v_k \in V$$

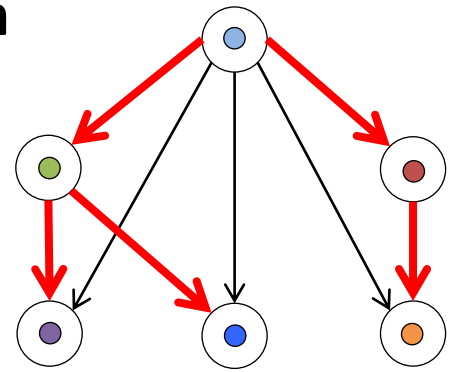
Tree

$$\sum_{v_j\in\delta^-(v_k)} f_{pk}x_{jk} \ge \sum_{v_l\in\delta^+(v_k)} f_{pl}x_{kl} \qquad \forall p \in [m],\ v_k \in V$$

Sum condition

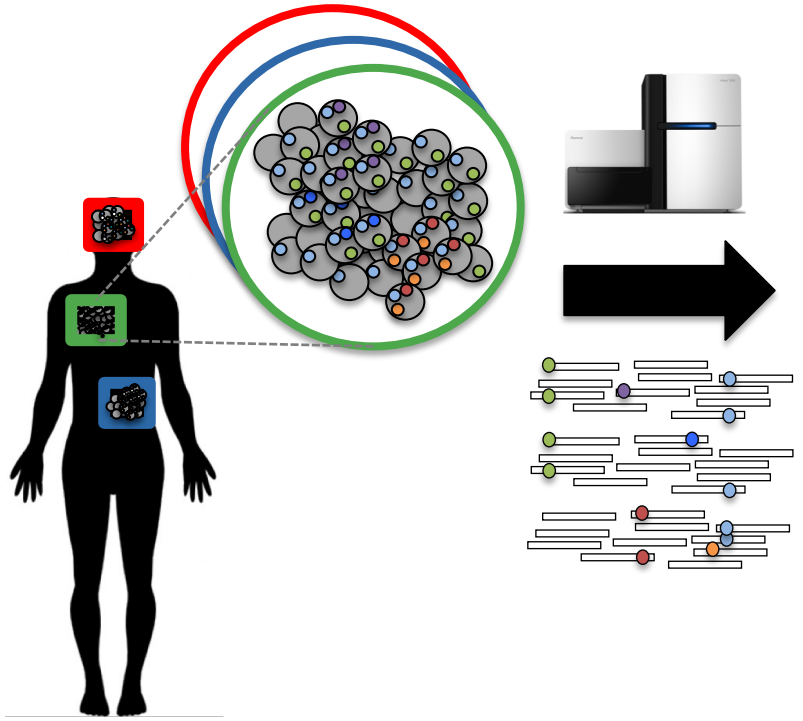
$$x_{jk} \in \{0,1\} \qquad \forall (v_j, v_k) \in A'$$

$$f_{pj} \in [f^-_{pj}, f^+_{pj}] \qquad \forall p \in [m], v_j \in V$$

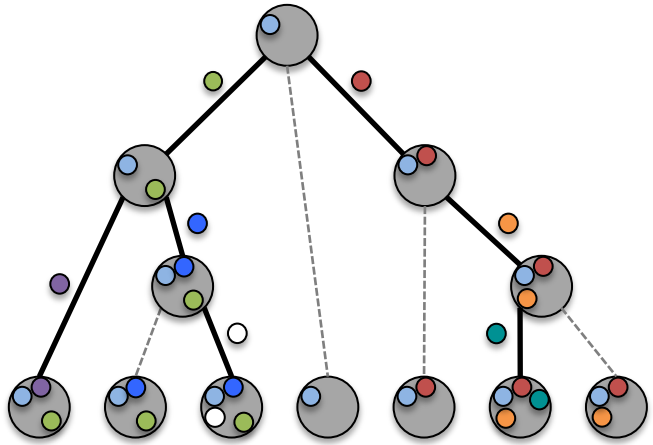


G = (V, A)

Outline

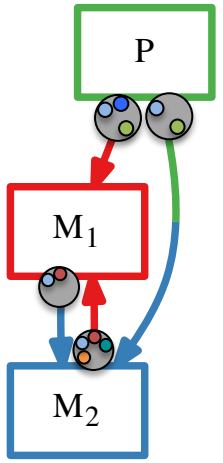


Cell division and mutation



Tumor Phylogeny Estimation

Cell migration



Tumor Migration Analysis

[**El-Kebir***, Oesper* et al., *ISMB* 2015/*Bioinformatics*]

Biological Problem

Mixed tumor samples require specialized phylogeny algorithms

Assumptions

- Infinite sites assumption

Computational Problem

Interval VAF Factorization Problem (I-VAFFP)
Given F^- and F^+, find F, U and B such that $F = U B$

Approach

- Combinatorial characterization as constrained spanning trees in a directed acyclic graph
- Probabilistic model for noisy variant allele frequencies
- **AncesTree** : mixed integer linear programming

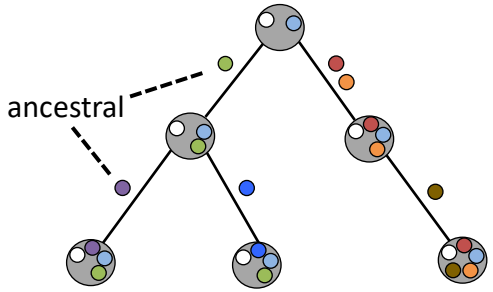
Results on Simulated Data

Generated:

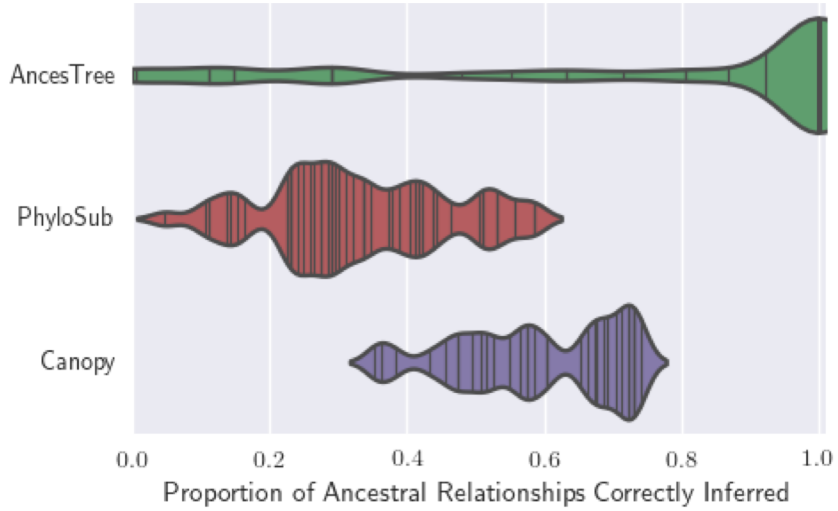
- 3,4,5 node trees
- 20 Mutations
- 5 Samples
- Average coverage: 200x

Ran:

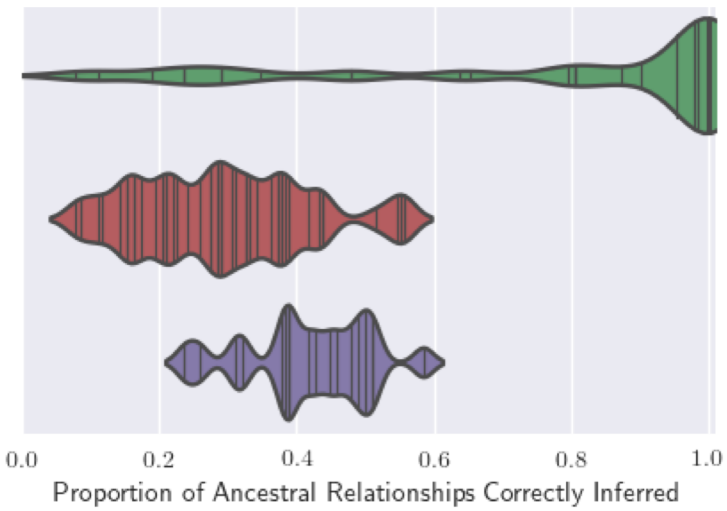
- AncesTree (Combinatorial)
- PhyloSub (Probabilistic) [Jiao *et al.,* 2014]
- Canopy (Probabilistic) [Jiang *et al.*, 2016]



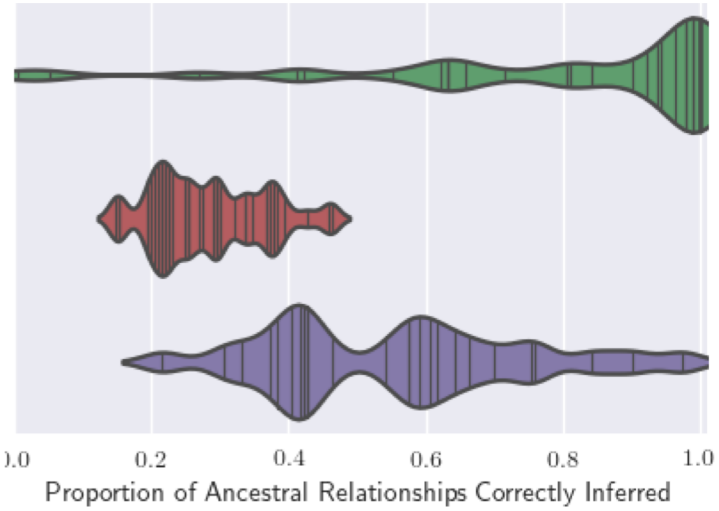
3 nodes



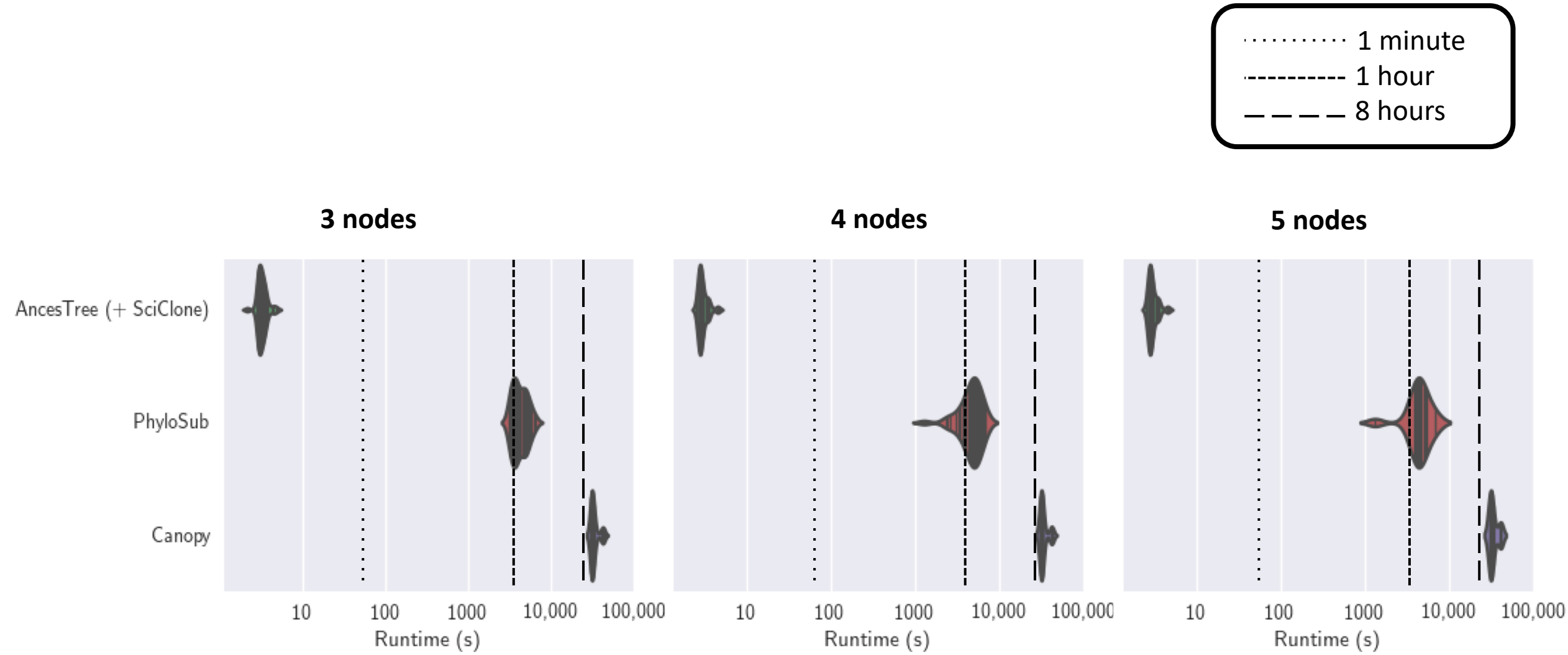
4 nodes



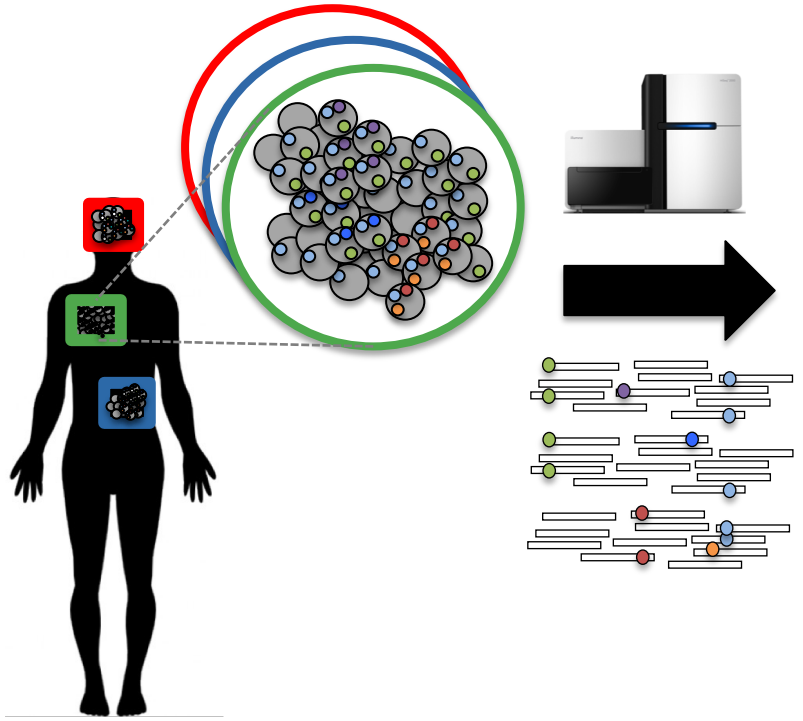
5 nodes



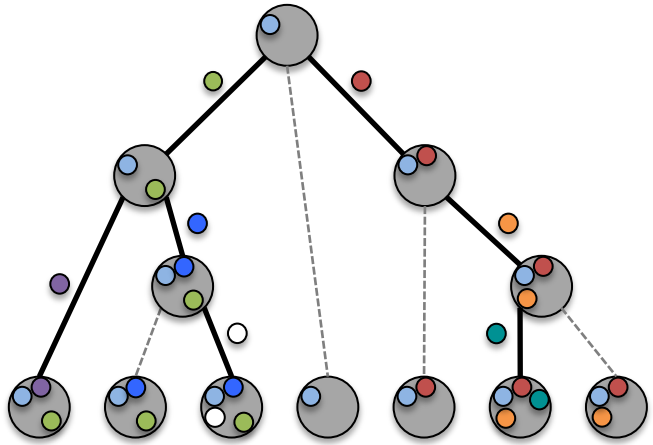
Runtime on Simulated Data



Outline



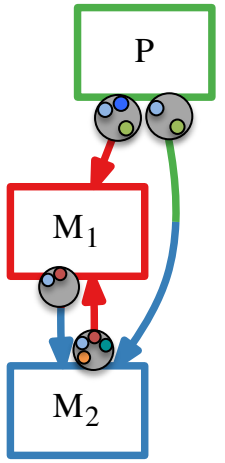
Cell division and mutation



Tumor Phylogeny Estimation

[**El-Kebir***, Oesper* et al., *ISMB* 2015/*Bioinformatics*]

Cell migration



Tumor Migration Analysis

[**El-Kebir** et al., *Nature Genetics* 2018]

Biological Problem

Mixed tumor samples require specialized phylogeny algorithms

Assumptions

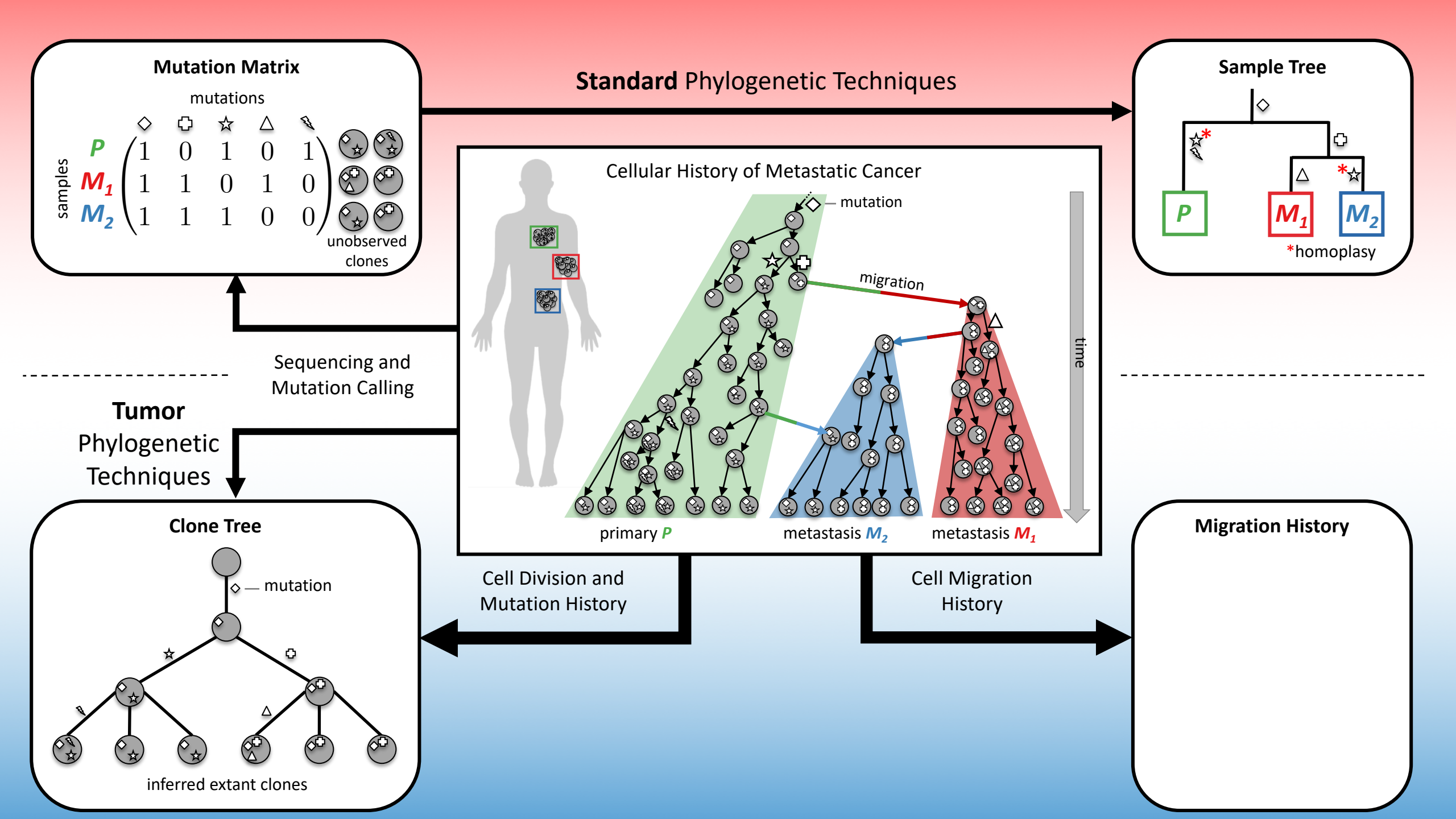
- ***Infinite sites assumption***

Computational Problem

Interval VAF Factorization Problem (I-VAFFP)
Given F^- and F^+, find F, U and B such that $F = U B$

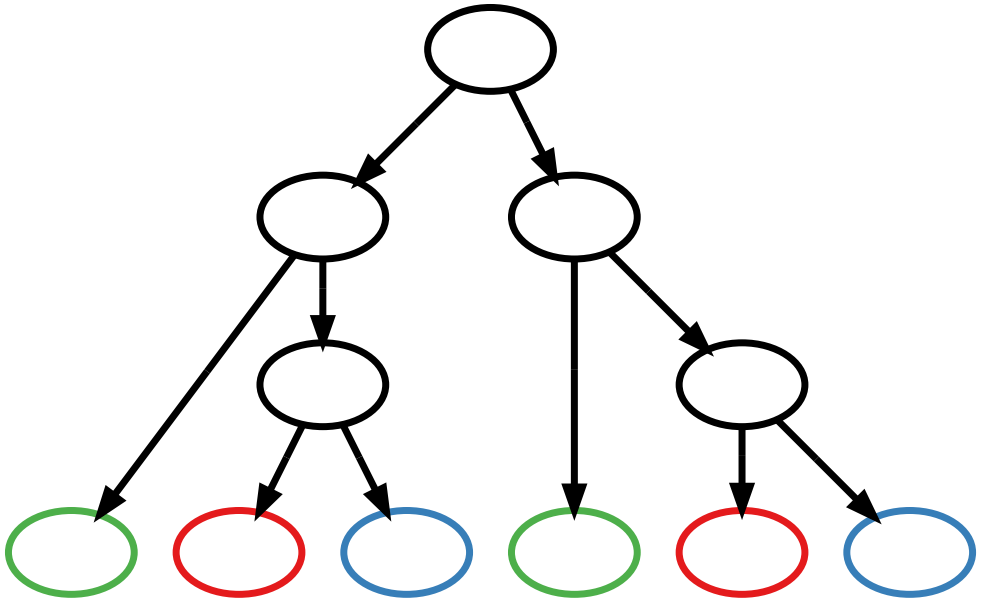
Approach

- Combinatorial characterization as constrained spanning trees in a directed acyclic graph
- Probabilistic model for noisy variant allele frequencies
- **AncesTree** : mixed integer linear programming


Mutation Matrix
mutations
samples
P
M1
M2
1 0 1 0 1
1 1 0 1 0
1 1 1 0 0
unobserved clones
Standard Phylogenetic Techniques
Sample Tree
P
M1
M2
*homoplasy
Cellular History of Metastatic Cancer
mutation
migration
time
primary P
metastasis M2
metastasis M1
Sequencing and Mutation Calling
Tumor Phylogenetic Techniques
Clone Tree
mutation
inferred extant clones
Cell Division and Mutation History
Cell Migration History
Migration History

Cell Division/Mutation and Migration are Separate Processes

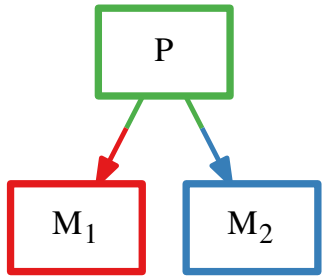
(1) Cell division and mutation



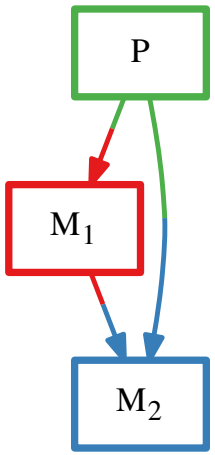
Leaf-labeled Clone Tree ***T***

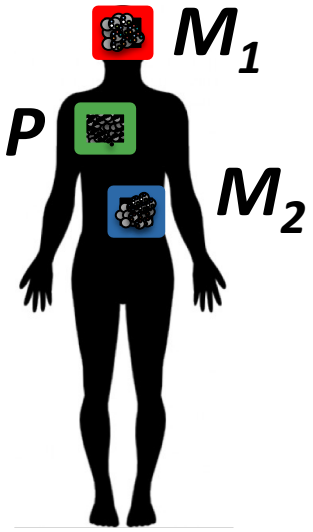
???

(2) Cell migration

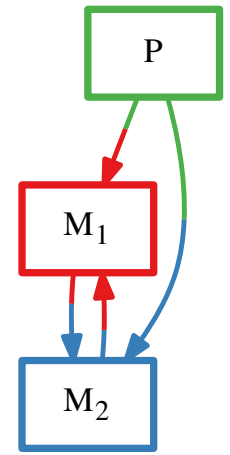


Parallel Seeding

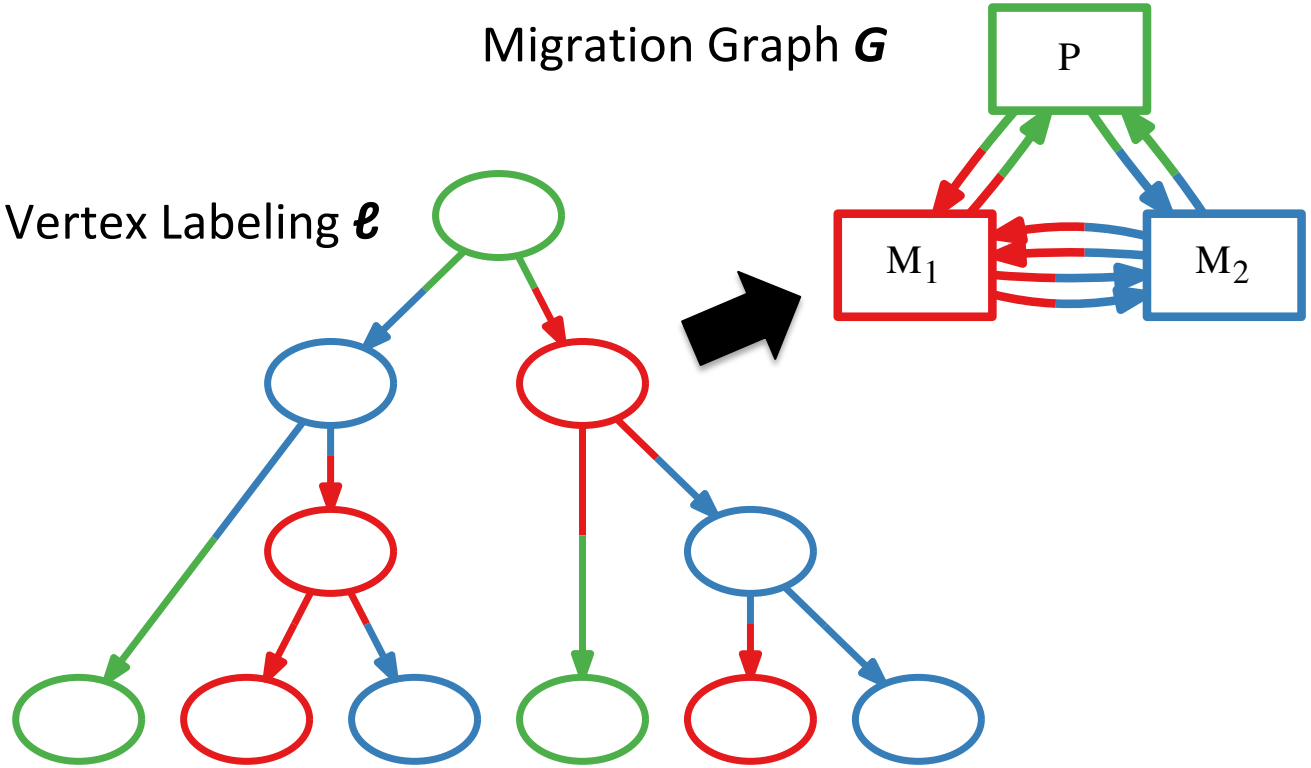




Cascading Seeding



Cell Division/Mutation and Migration are Separate Processes

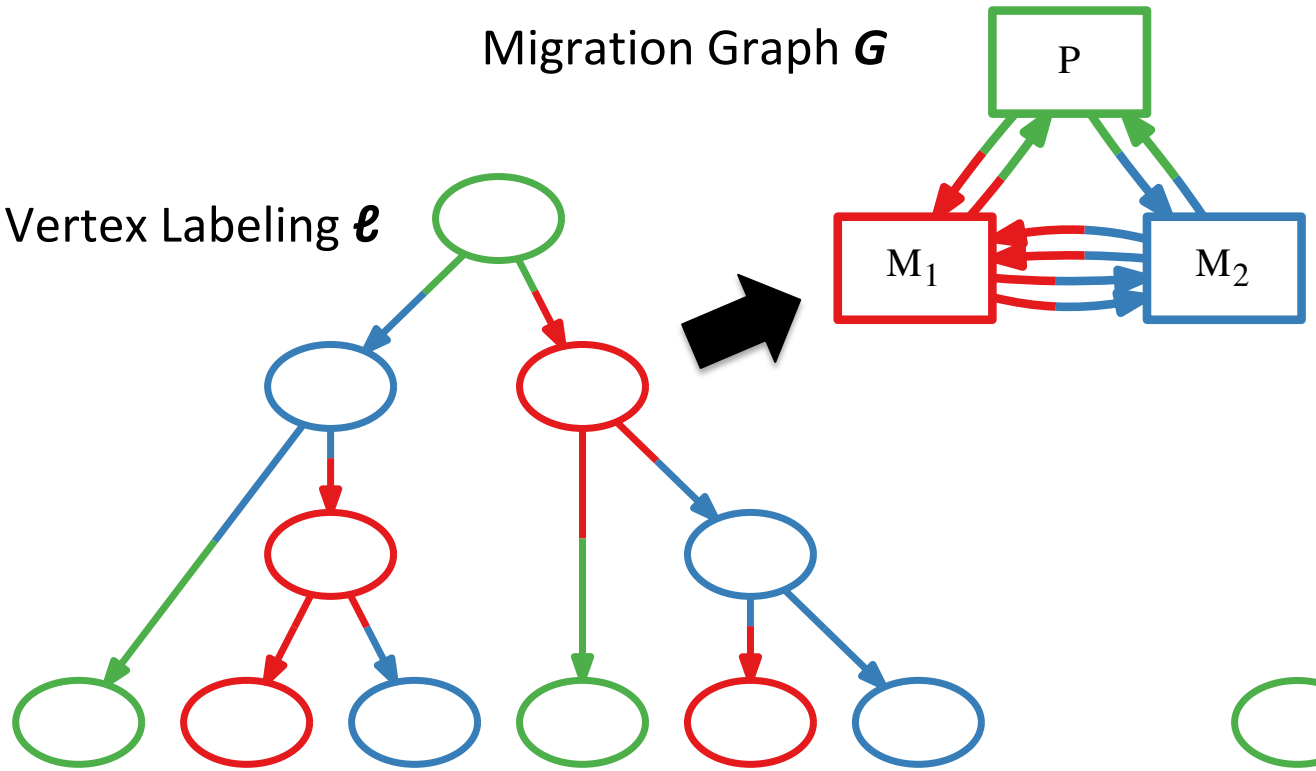


Leaf-labeled Clone Tree T

$\mu(T, \ell)$ = 8 migrations

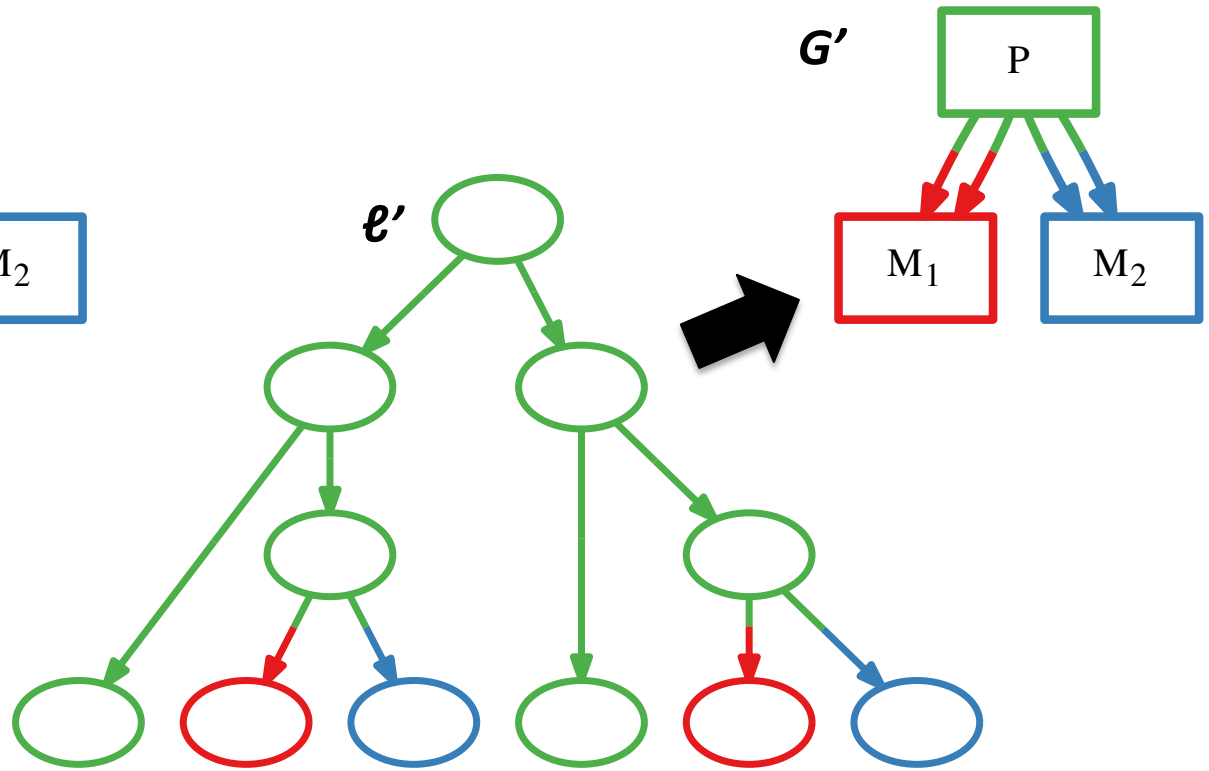
Given T and ℓ, migrations are bichromatic edges in T, or edges in G

Cell Division/Mutation and Migration are Separate Processes



Leaf-labeled Clone Tree T

$\mu(T, \ell) = 8$ migrations



T

$\mu(T, \ell') = 4$ migrations

Given T and ℓ, migrations are bichromatic edges in T, or edges in G

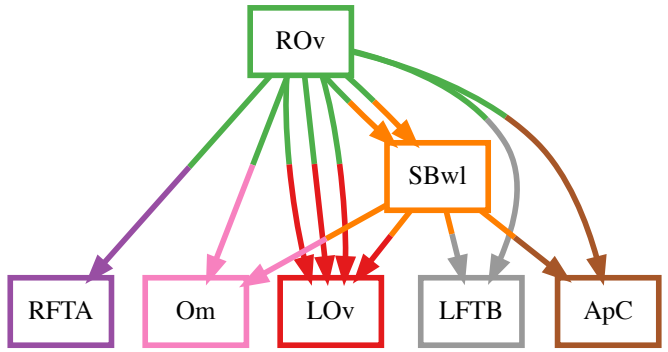
Goal: Given T, find vertex labeling ℓ with minimum number of migrations

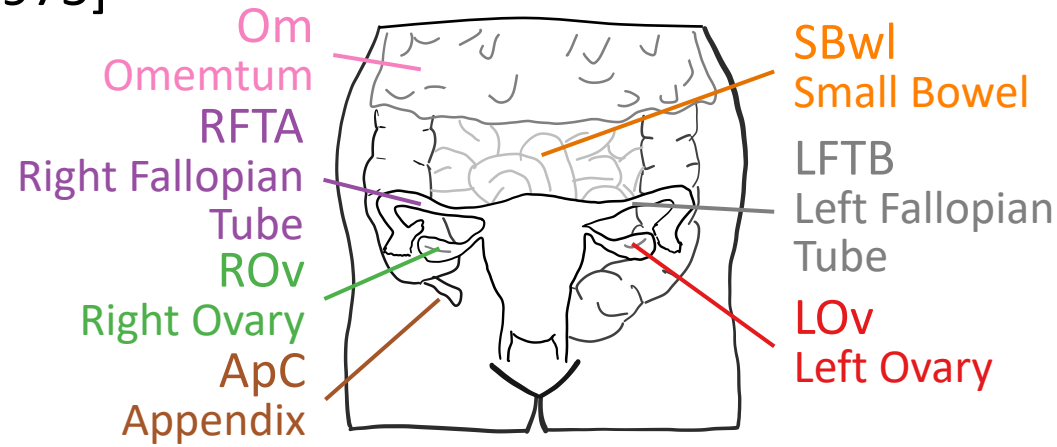
Minimum Migration Analysis in Ovarian Cancer

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

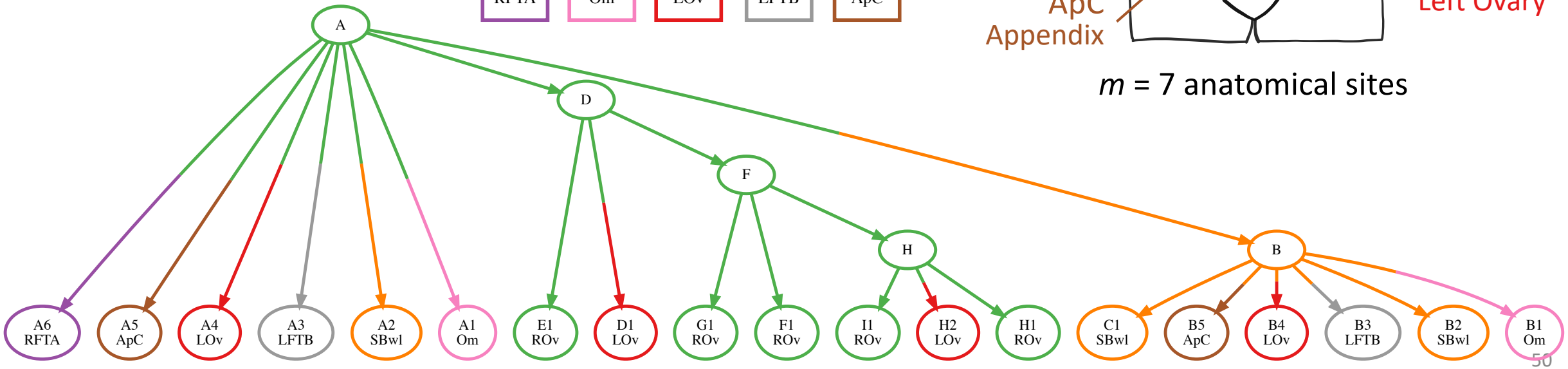
- Instance of the maximum parsimony small phylogeny problem
- Can be solved in polynomial time [Fitch, 1971; Sankoff, 1975]

$\mu^* = 13$ migrations

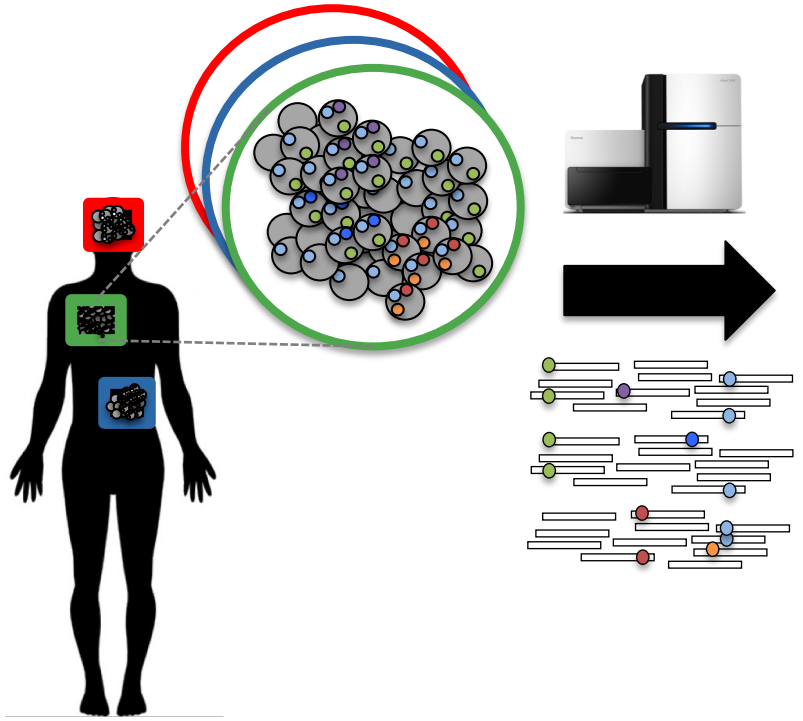




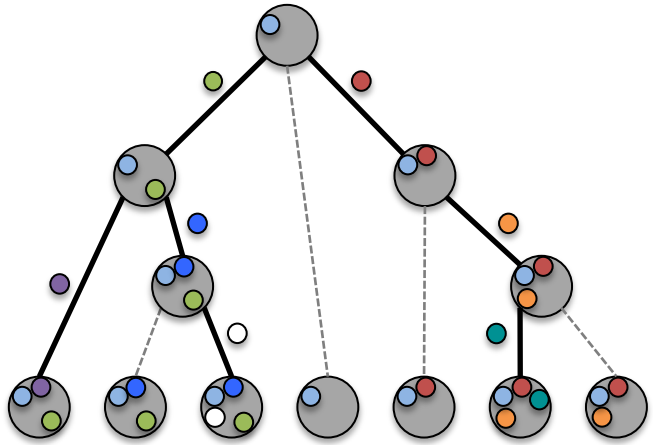
m = 7 anatomical sites



Outline



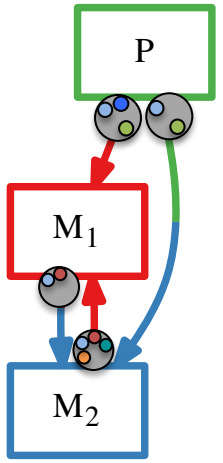
Cell division and mutation



Tumor Phylogeny Estimation

[**El-Kebir** et al., *RECOMB* 2016]; [**El-Kebir***, Satas* et al., *Cell Systems* 2016]

Cell migration



Tumor Migration Analysis

[**El-Kebir** et al., *Nature Genetics* 2018]

Biological Problem

Reconstructing migration history of a metastatic cancer

Assumptions

- Migrations are rare
- Migrations are independent

Computational Problem

Parsimonious Migration History: Given T, find vertex labeling ℓ with minimum number of migrations

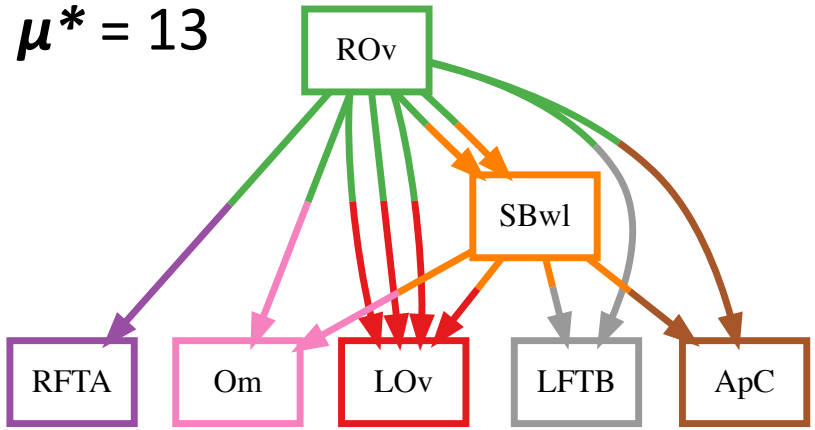
Approach

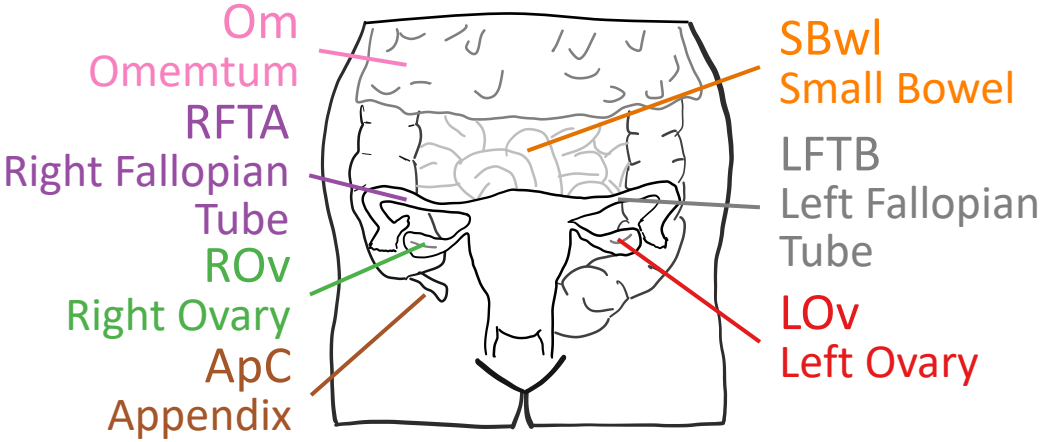
- Dynamic programming (Sankoff/Fitch algorithm)

Are there multiple vertex labelings with $\mu^* = 13$ migrations?

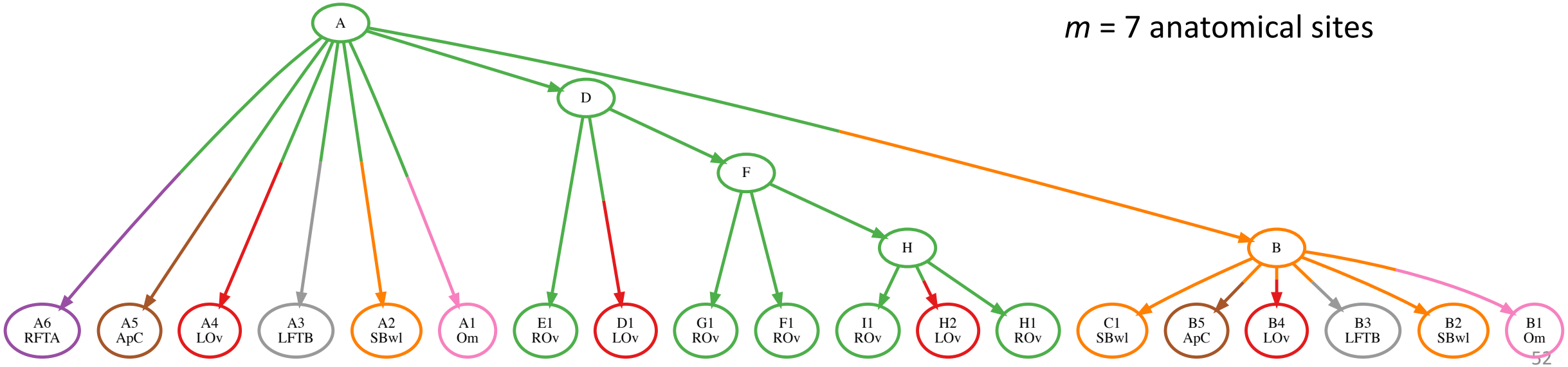
McPherson et al. (2016). Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics*.

$\mu^* = 13$





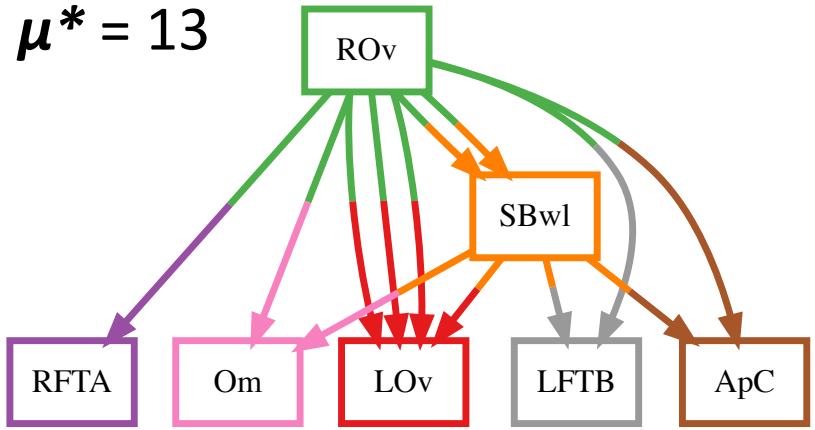
m = 7 anatomical sites



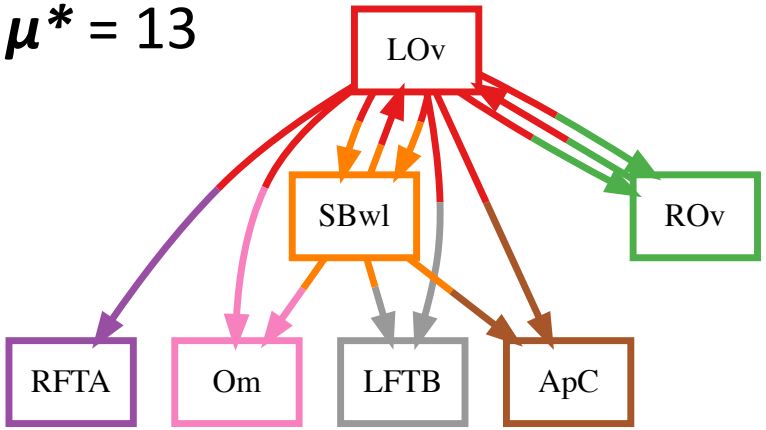
Minimum Migration History is ***Not*** Unique

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

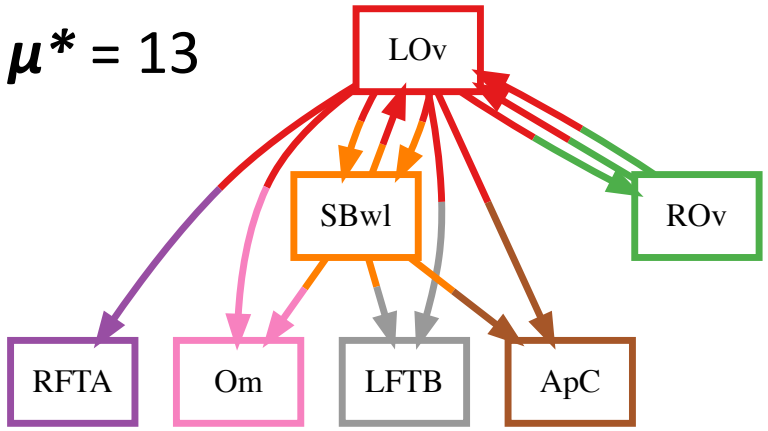
$\mu^* = 13$



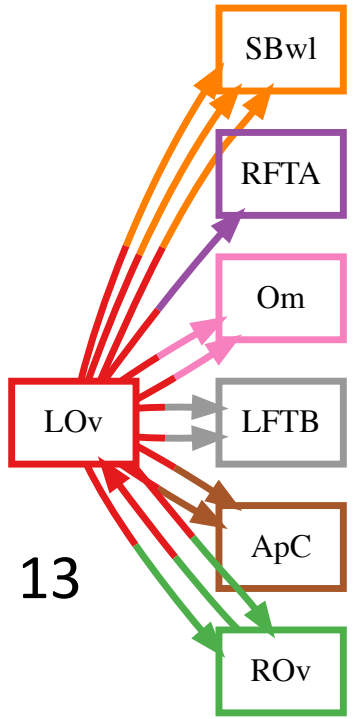
$\mu^* = 13$



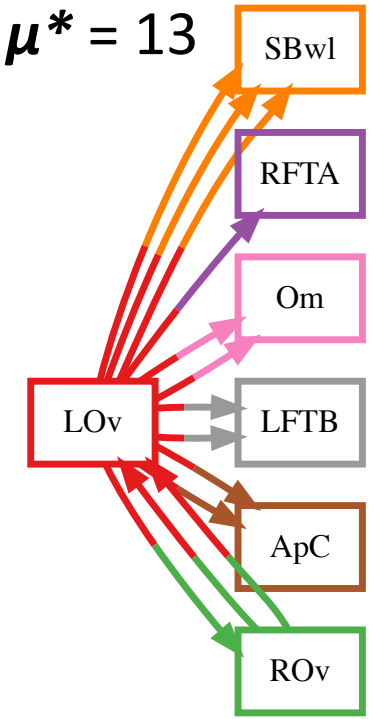
$\mu^* = 13$



$\mu^* = 13$



$\mu^* = 13$



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

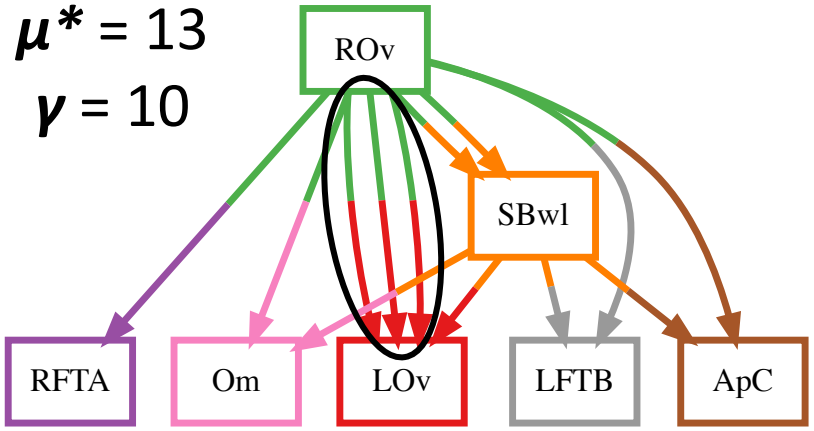
Comigrations: Simultaneous Migrations of Multiple Clones

- 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 †

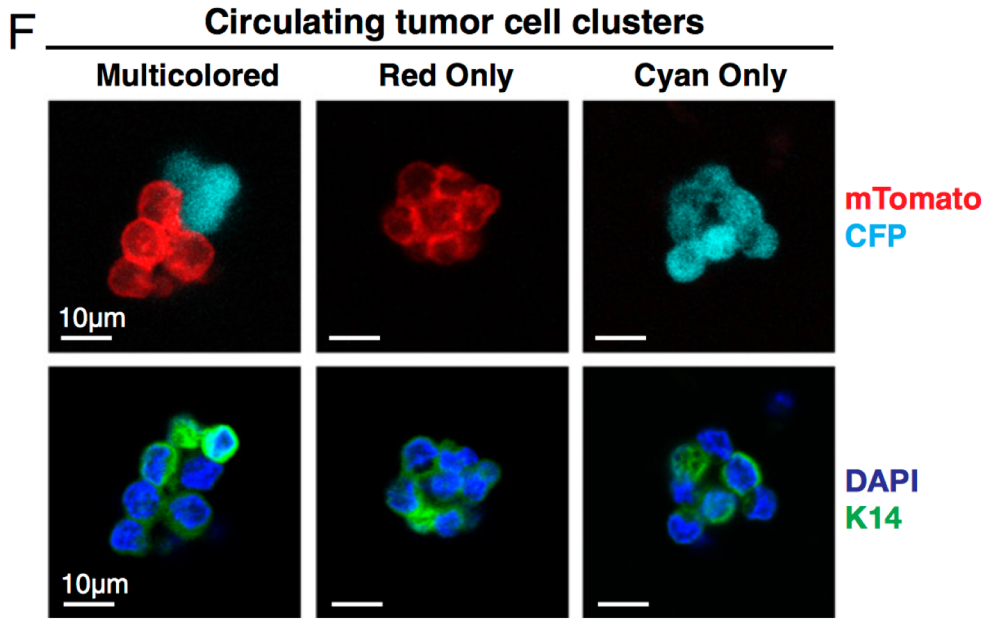
† Not necessarily true in the case of directed cycles

$\mu^* = 13$

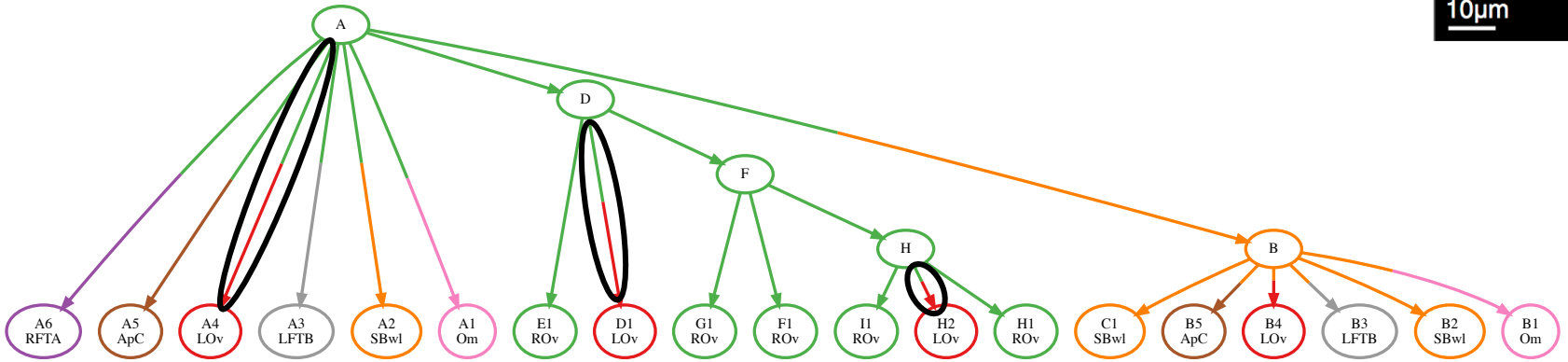
$\gamma = 10$



Migration Graph G



Clone Tree T



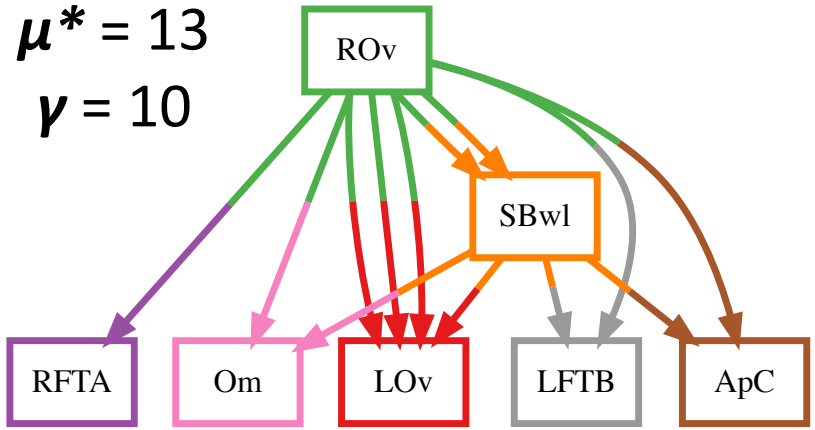
ApC	Appendix
LFTB	Left Fallopian Tube
LOv	Left Ovary
RFTA	Right Fallopian Tube
ROv	Right Ovary
SBwl	Small Bowel
Om	Omentum

Comigrations: Simultaneous Migrations of Multiple Clones

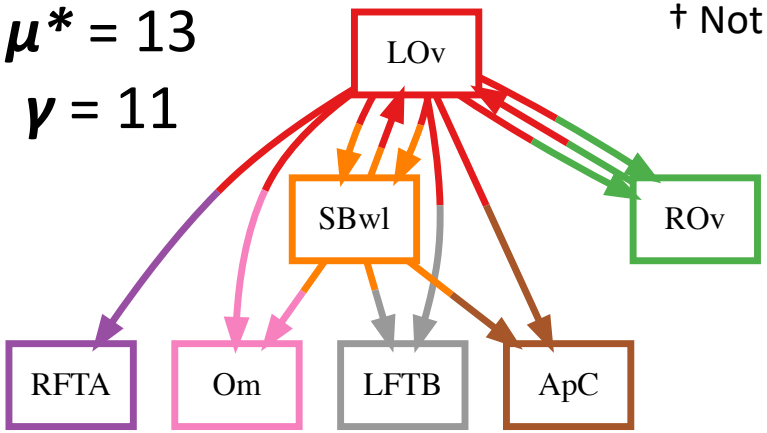
- Multiple tumor cells migrate simultaneously through the blood stream [Cheung et al., 2016]
- Second objective: number γ of comigrations is the number of multi-edges in migration graph $G^{\dagger}$

$\dagger$ Not necessarily true in the case of directed cycles

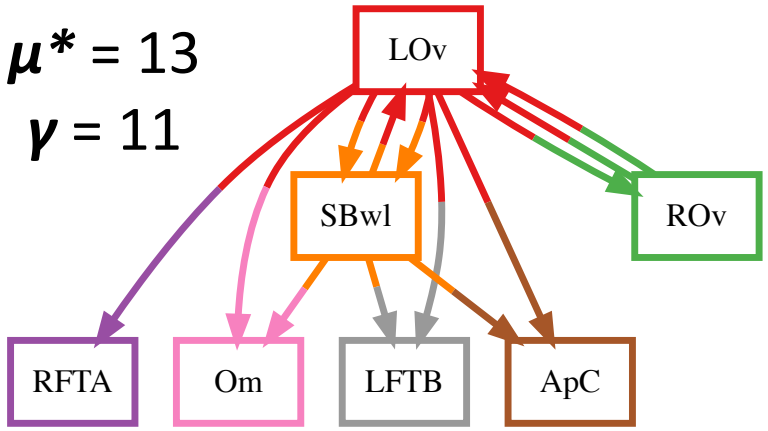
$\mu^* = 13$
$\gamma = 10$



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

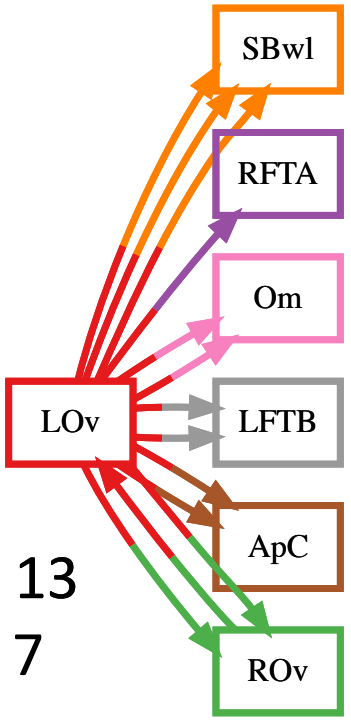


$\mu^* = 13$
$\gamma = 11$

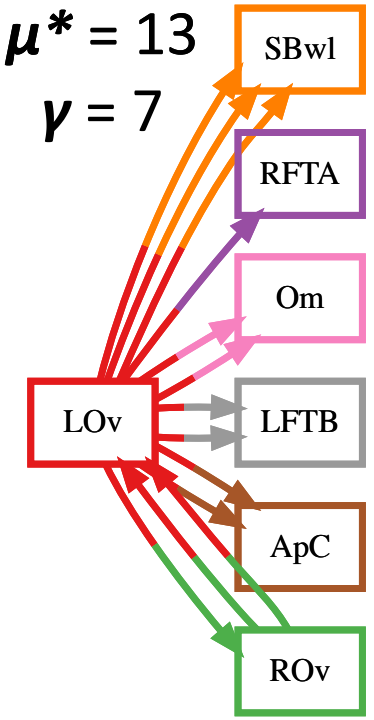


ApC	Appendix
LFTB	Left Fallopian Tube
LOv	Left Ovary
RFTA	Right Fallopian Tube
ROv	Right Ovary
SBwl	Small Bowel
Om	Omentum

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

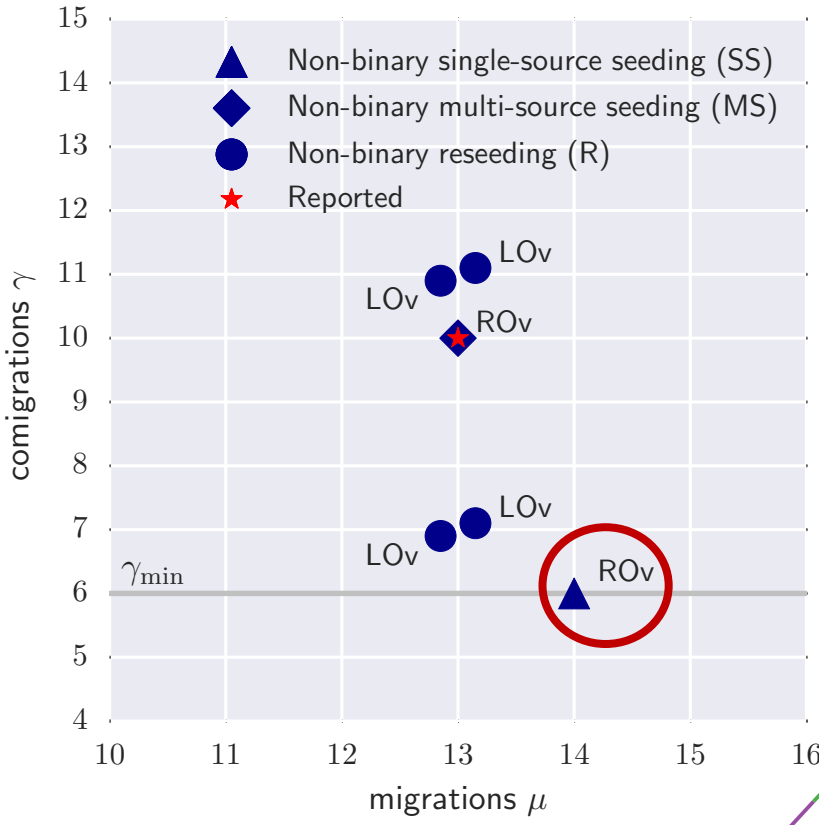


$\mu^* = 13$
$\gamma = 7$



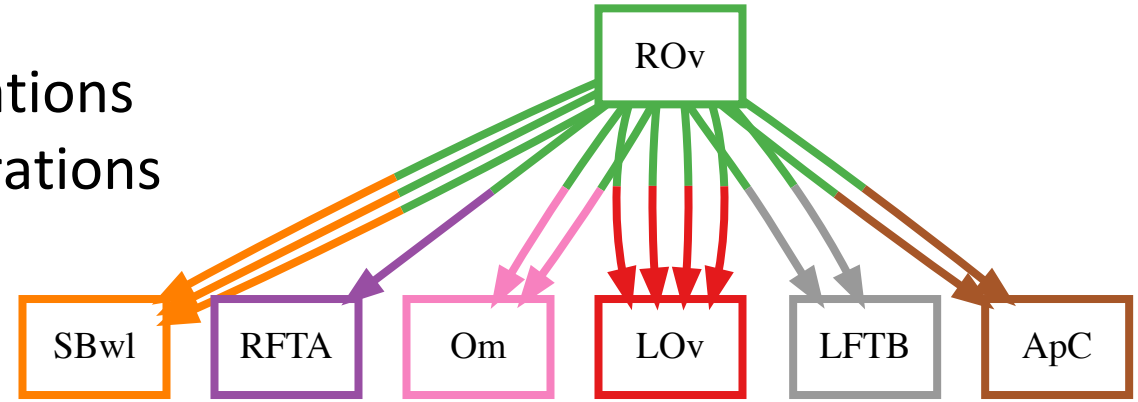
Tradeoff between Migrations and Comigrations

- Minimum number γ^* of comigrations is $m - 1$ (where m is #anatomical sites)

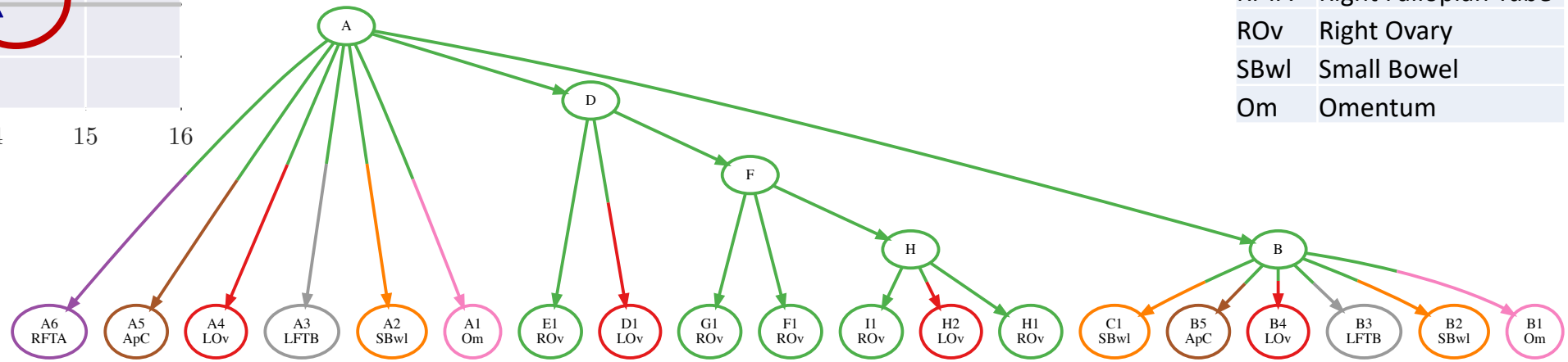


μ = 14 migrations

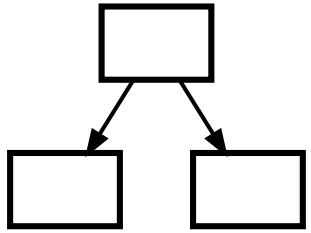
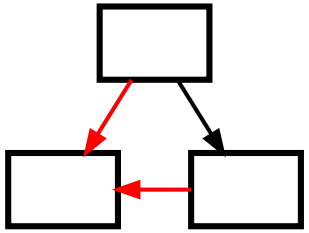
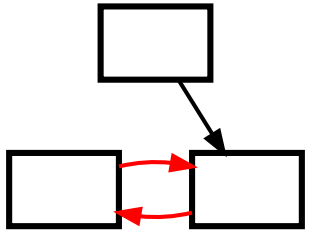
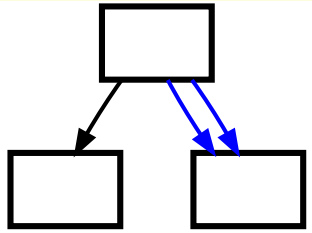
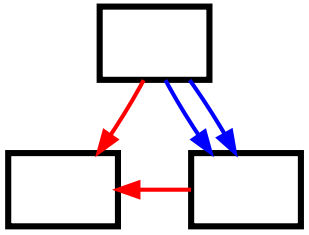
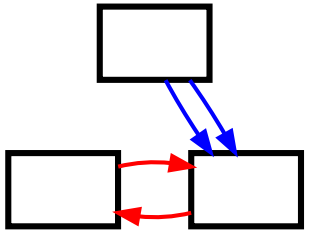
γ^* = 6 comigrations

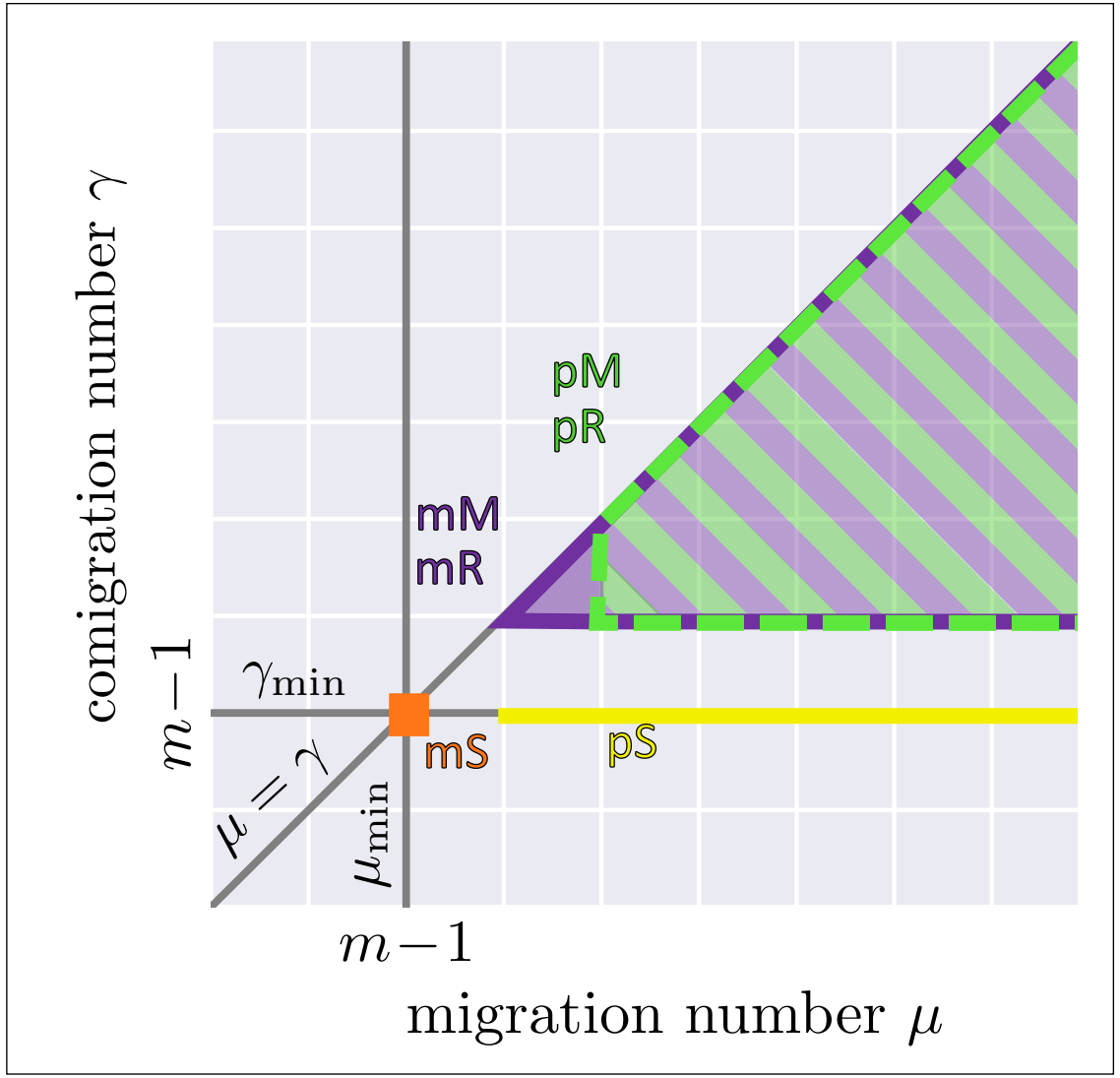


ApC	Appendix
LFTB	Left Fallopian Tube
LOv	Left Ovary
RFTA	Right Fallopian Tube
ROv	Right Ovary
SBwl	Small Bowel
Om	Omentum

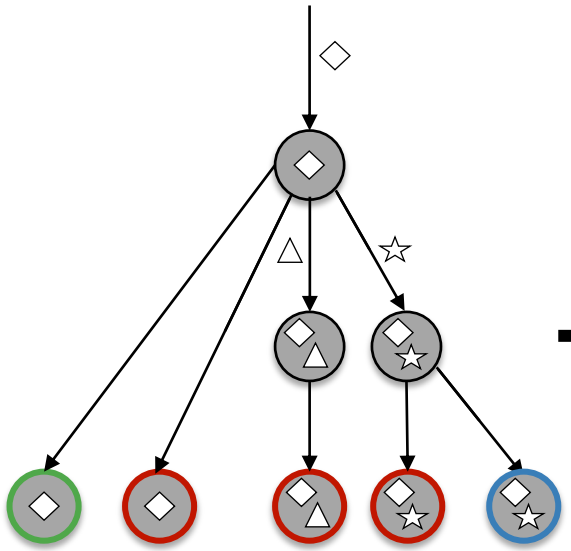


Tradeoff between Migrations, Comigrations and Migration Pattern

	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



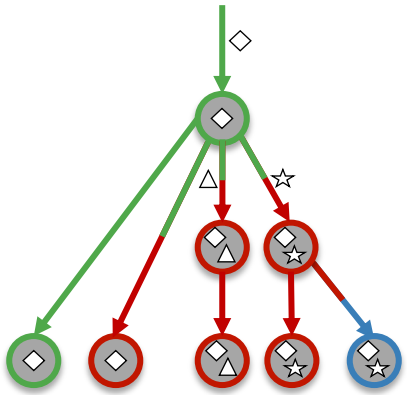
Parsimonious Migration History Problem



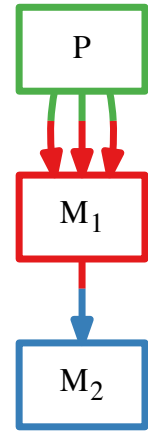
Clone Tree T

PMH Problem
Parsimonious
Migration History

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



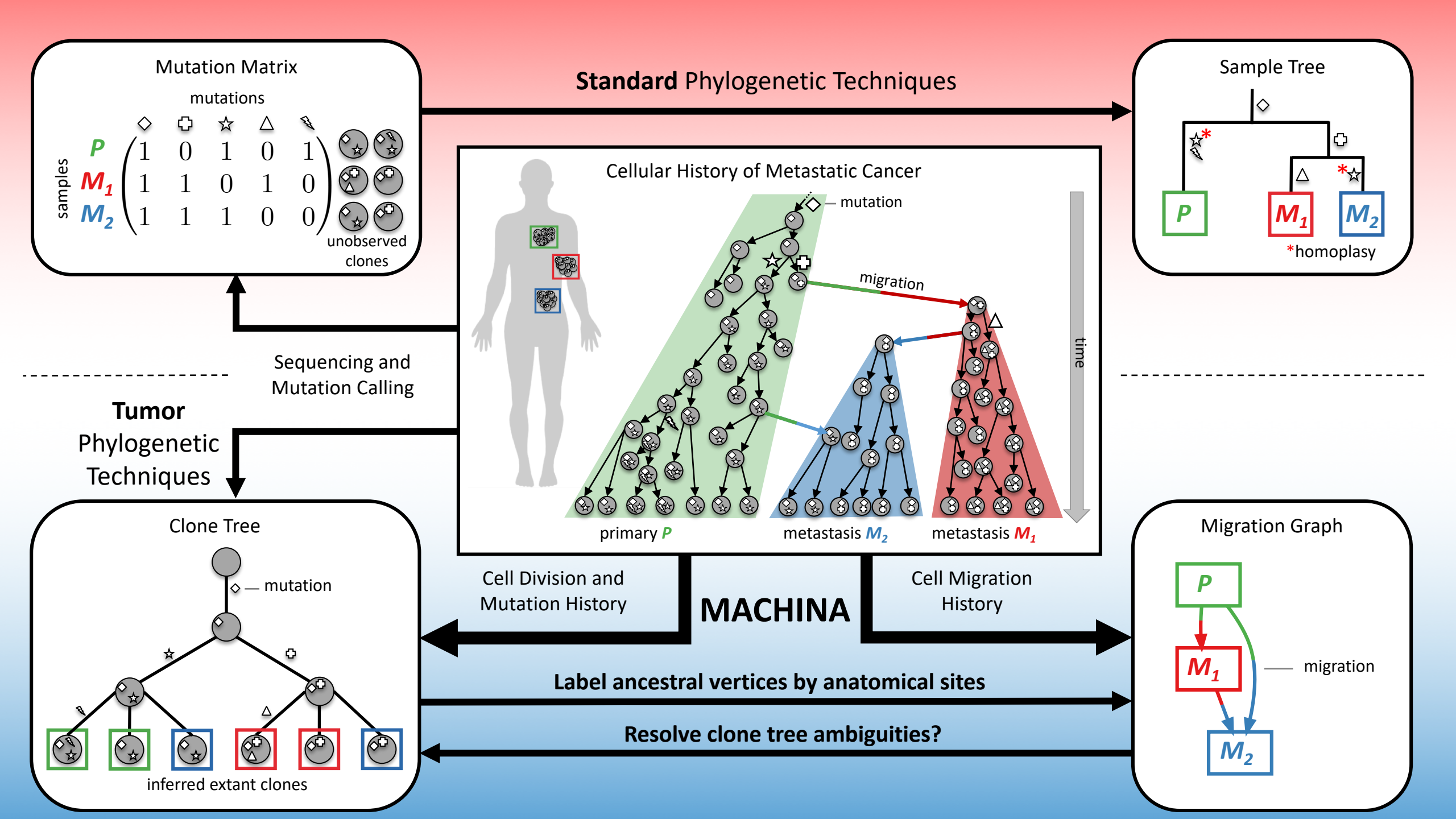
polyclonal single
source seeding (pS)



$\mu = 4$

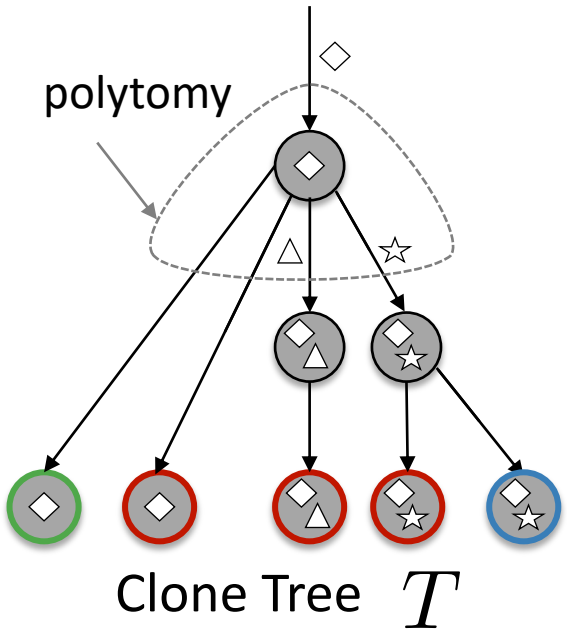
$\gamma = 2$

Parsimonious Migration History Problem: Given a clone tree T and a set $\mathcal{P}$ of allowed migration patterns, find a vertex labeling ℓ with the minimum migration number $\mu^*(T)$ and subsequently the smallest comigration number $\hat{\gamma}(T)$.


Mutation Matrix
mutations
samples
P
M1
M2
1 0 1 0 1
1 1 0 1 0
1 1 1 0 0
unobserved clones
Standard Phylogenetic Techniques
Sample Tree
P
M1
M2
*homoplasy
Cellular History of Metastatic Cancer
mutation
migration
time
primary P
metastasis M2
metastasis M1
Sequencing and Mutation Calling
Tumor Phylogenetic Techniques
Clone Tree
mutation
inferred extant clones
Cell Division and Mutation History
MACHINA
Cell Migration History
Label ancestral vertices by anatomical sites
Resolve clone tree ambiguities?
Migration Graph
P
M1
M2
migration

Resolving Clone Tree Ambiguities

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

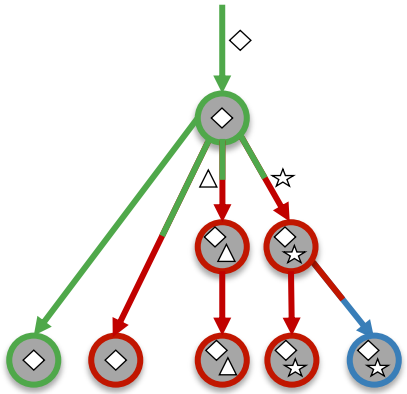


Clone Tree T

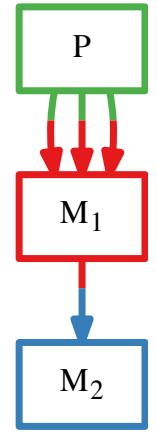
PMH

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

PMH-TR

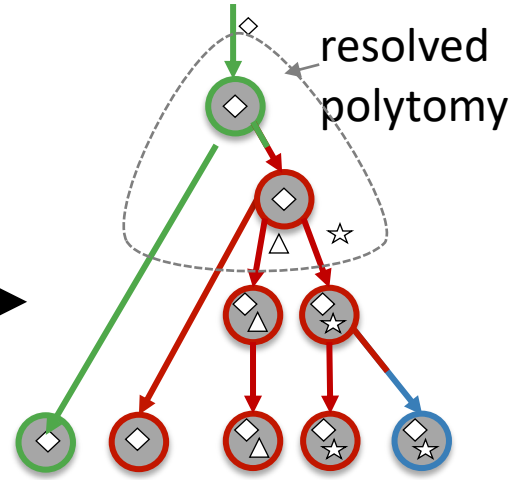


polyclonal single source seeding (pS)

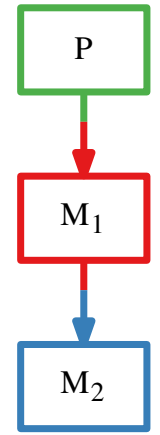


$\mu = 4$

$\gamma = 2$



monoclonal single source seeding (mS)



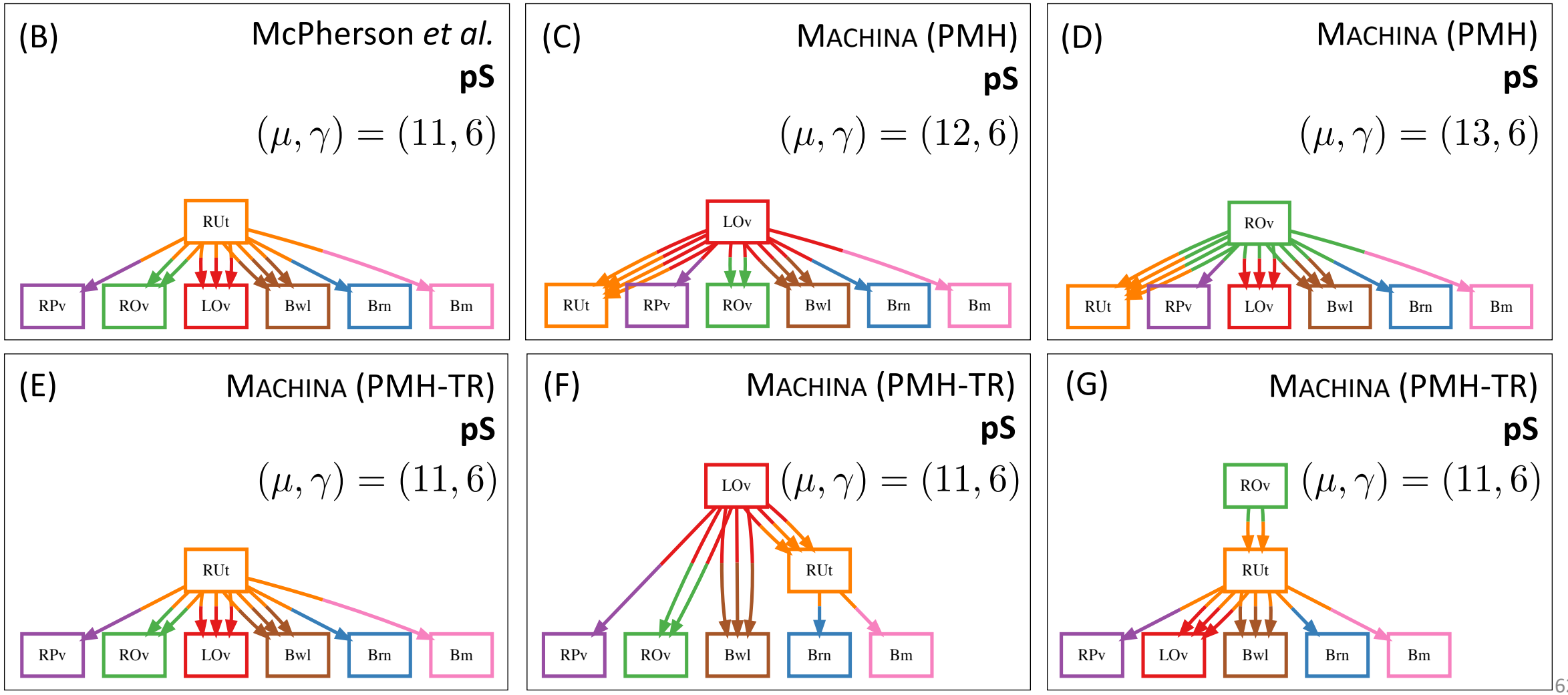
$\mu = 2$

$\gamma = 2$

Parsimonious Migration History with Tree Refinement (PMH-TR): Given a clone tree T and a set $\mathcal{P}$ of allowed migration patterns, find a refinement T' of T and vertex labeling ℓ of T' with the minimum migration number $\mu^*(T')$, and subsequently smallest comigration number $\hat{\gamma}(T')$.

Polytomy Resolution in Ovarian Cancer 7

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



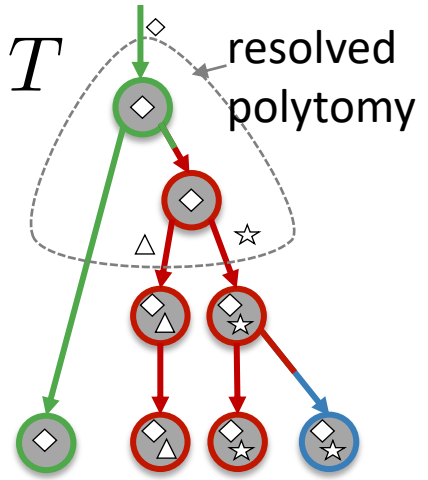
Resolving Clone Tree Ambiguities

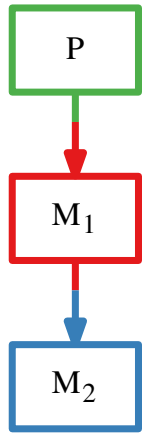
Parsimonious Migration History with Tree Refinement (PMH-TI):
Given a set $\mathcal{P}$ of allowed migration patterns and mutation frequency confidence intervals ($F^- = [f^-_{s,i}], F^+ = [f^+_{s,i}]$), find a frequency matrix $\hat{F} = [\hat{f}_{s,i}]$, a clone tree T, and a vertex labeling ℓ of T such that:

(1) $\hat{f}_{s,i} \in [f^-_{s,i}, f^+_{s,i}]$;

(2) $\hat{F}$ satisfies the sum condition for T; and

(3) vertex labeling ℓ of T has minimum migration number $\mu^*(T)$ and subsequently smallest comigration number $\hat{\gamma}(T)$.





$\mu = 2$

$\gamma = 2$

monoclonal single source seeding (mS)

$$\begin{array}{c} \\ P \\ M_1 \\ M_2 \end{array}\!\!\begin{array}{c} \diamond \quad \triangle \quad \star \\ \begin{bmatrix} 0.82 & 0 & 0 \\ 0.88 & 0.24 & 0.64 \\ 0.73 & 0 & 0.73 \end{bmatrix} \end{array} = \begin{array}{c} \\ P \\ M_1 \\ M_2 \end{array}\!\!\begin{array}{c} \diamond \quad \triangle \quad \star \\ \begin{bmatrix} 0.82 & 0 & 0 \\ 0 & 0.24 & 0.64 \\ 0 & 0 & 0.73 \end{bmatrix} \end{array} \begin{array}{c} \diamond \quad \triangle \quad \star \\ \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \end{array}$$

$\hat{F}$ $\qquad$ U $\qquad$ B

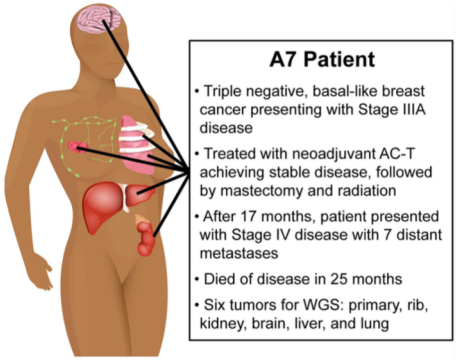
$$F^- = \begin{array}{c} \\ P \\ M_1 \\ M_2 \end{array}\!\!\begin{array}{c} \diamond \quad \triangle \quad \star \\ \begin{bmatrix} 0.75 & 0 & 0 \\ 0.82 & 0.15 & 0.56 \\ 0.66 & 0 & 0.68 \end{bmatrix} \end{array}$$

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

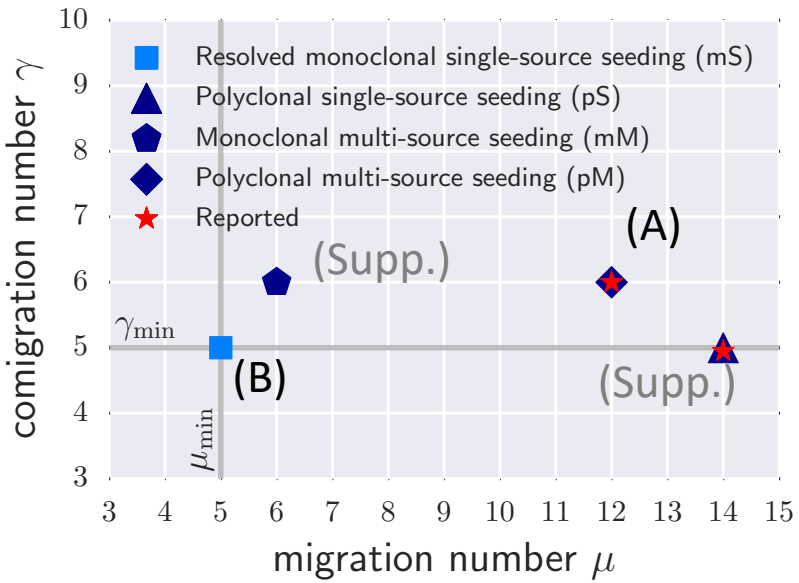
PMH-TI Problem
Parsimonious Migration History with Tree Inference

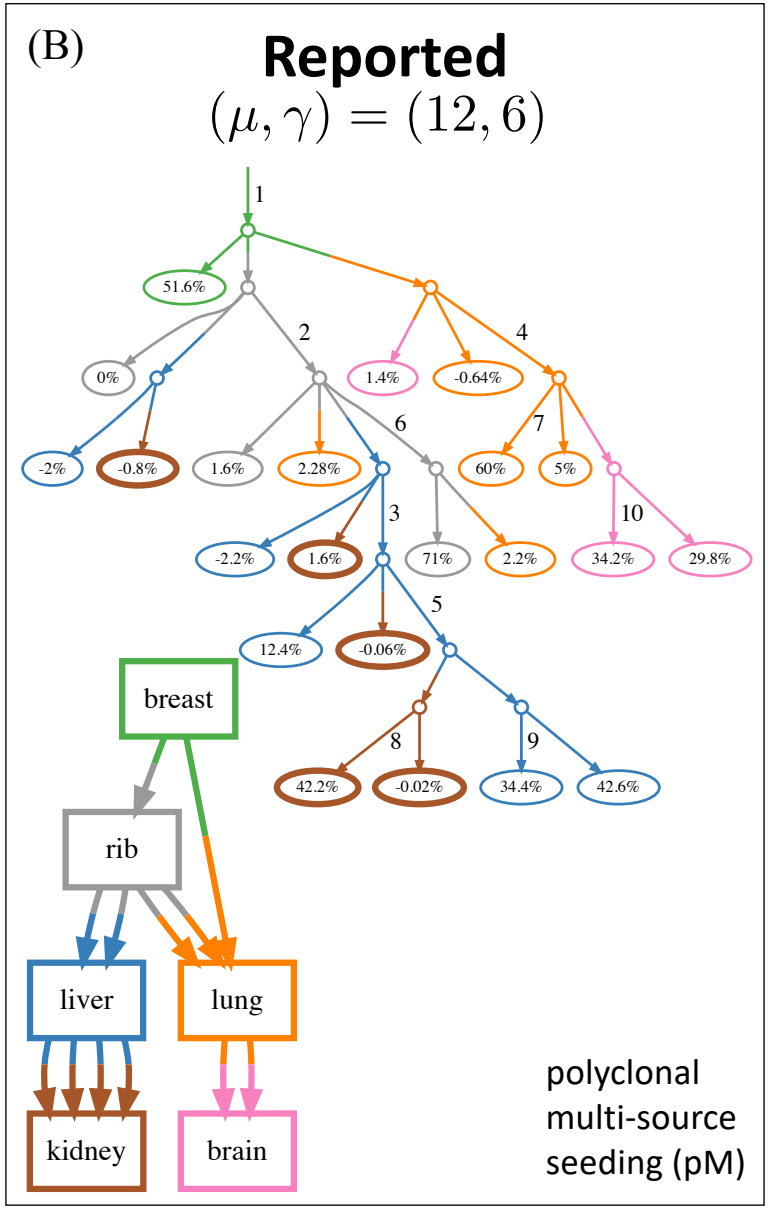
$\mathcal{P}$

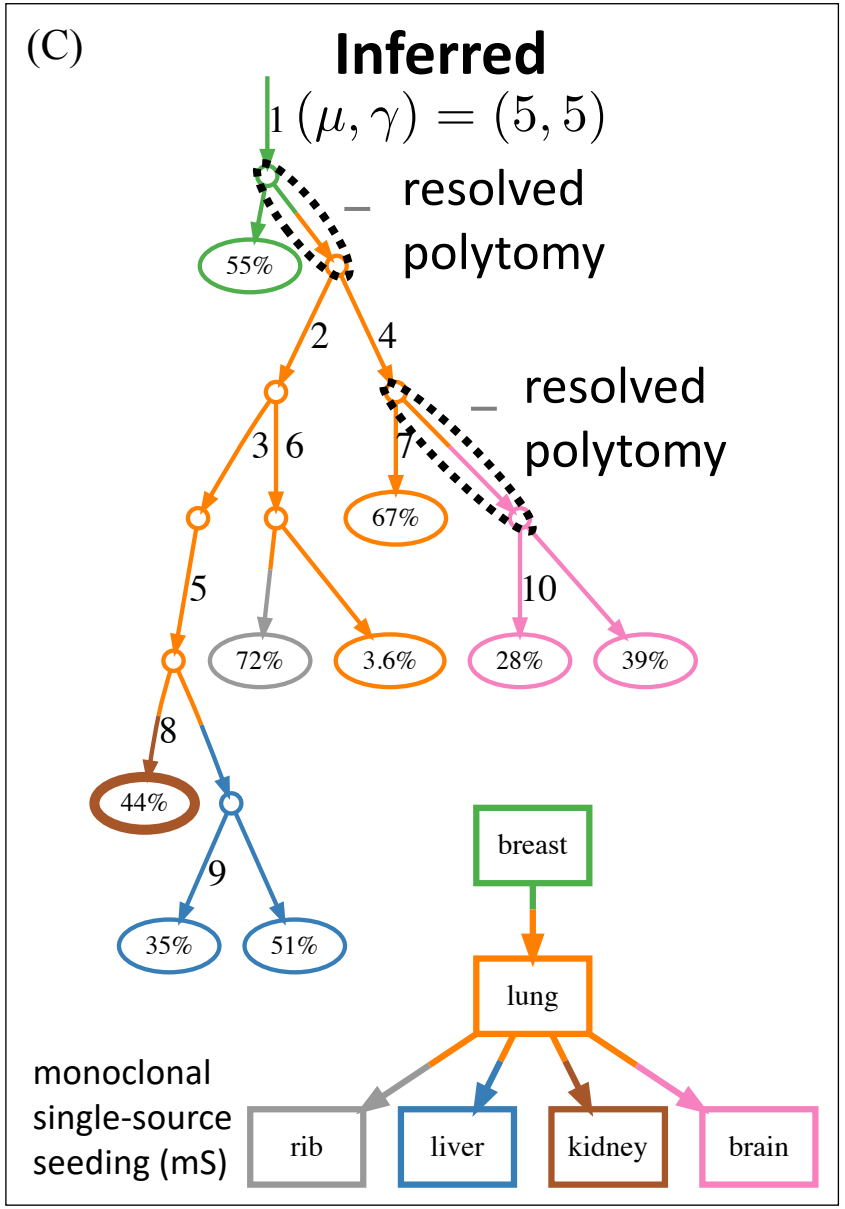
Applying MACHINA to Metastatic Breast Cancer

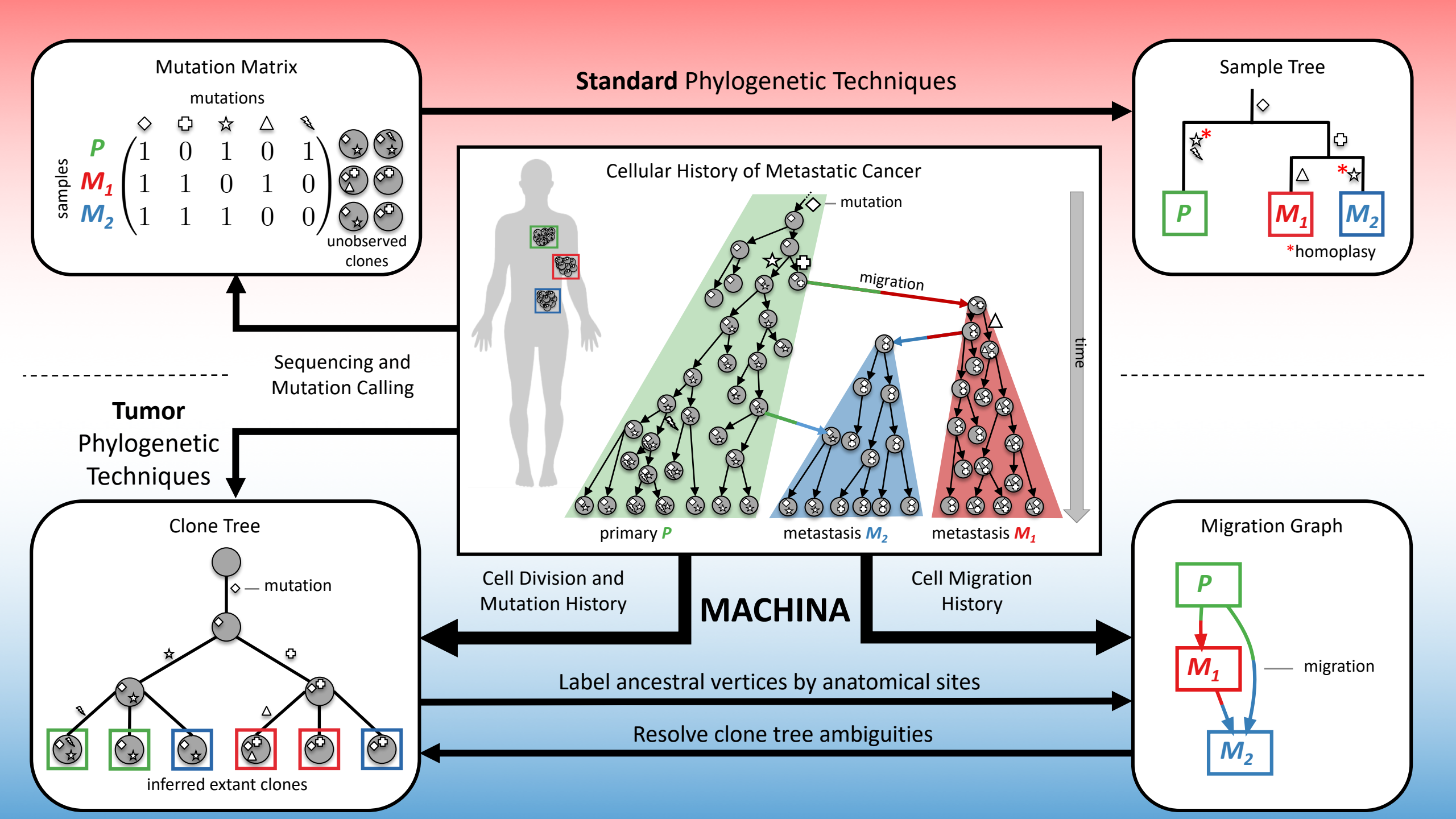


Hoadley et al. Tumor Evolution in Two Patients with Basal-like Breast Cancer: A Retrospective Genomics Study of Multiple Metastases. *PLOS Med*, *13*(12) 2016

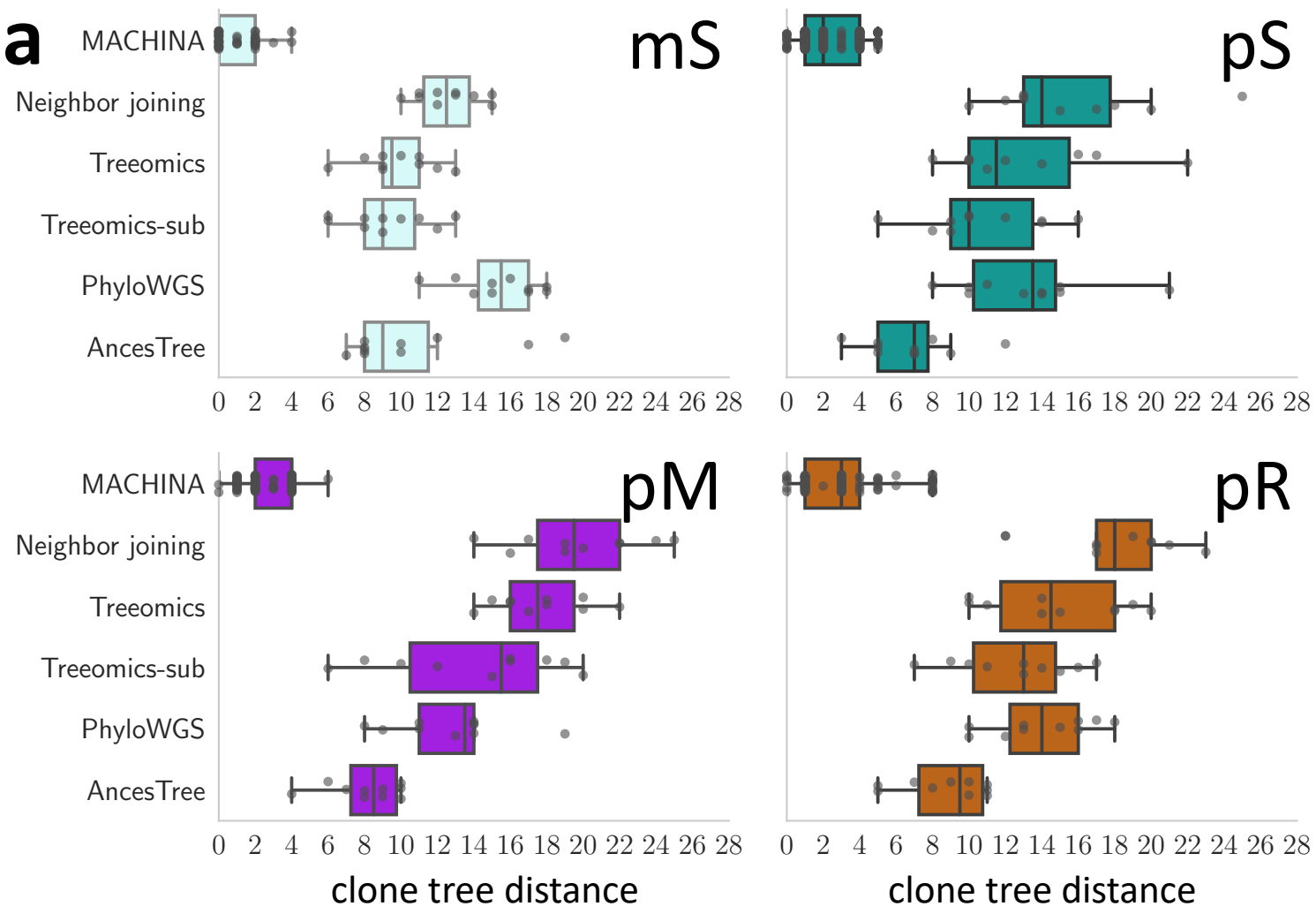


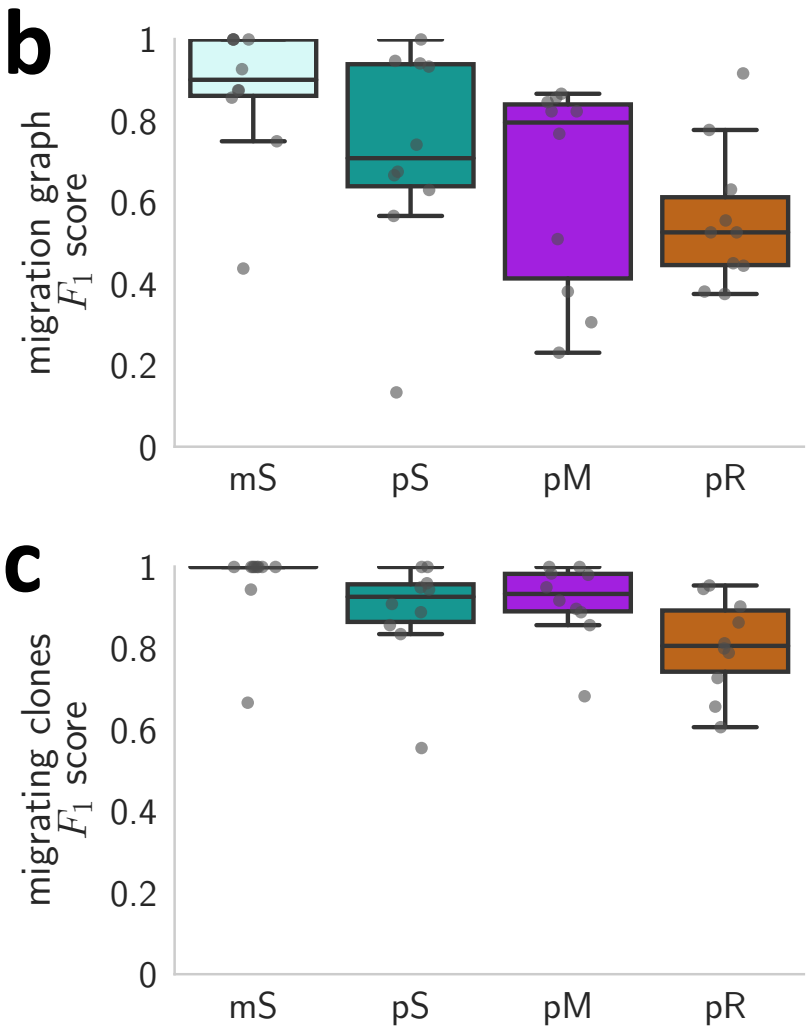




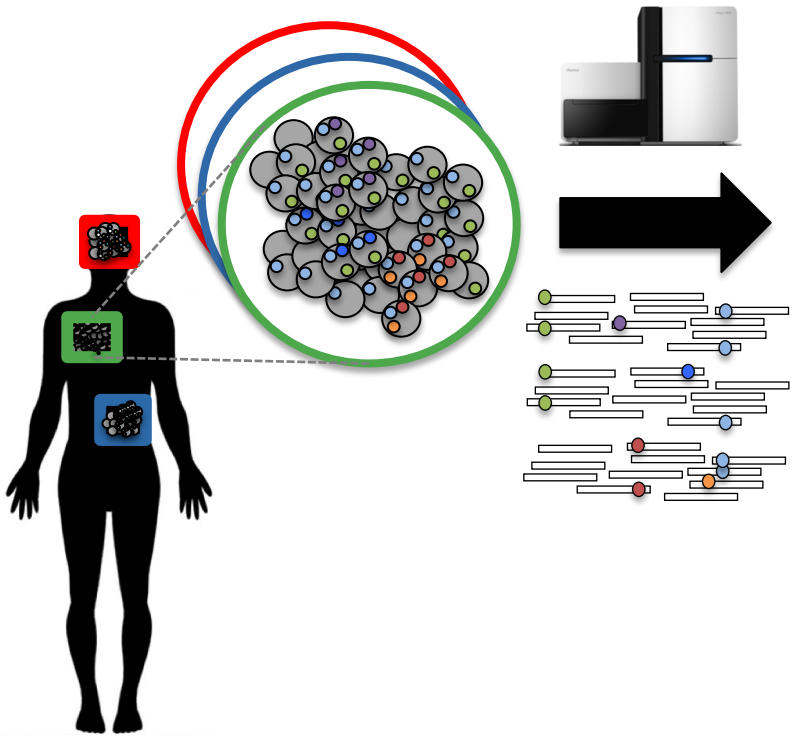

Mutation Matrix
mutations
samples
P
M1
M2
1 0 1 0 1
1 1 0 1 0
1 1 1 0 0
unobserved clones
Standard Phylogenetic Techniques
Sample Tree
P
M1
M2
*homoplasy
Cellular History of Metastatic Cancer
mutation
migration
time
primary P
metastasis M2
metastasis M1
Sequencing and Mutation Calling
Tumor Phylogenetic Techniques
Clone Tree
mutation
inferred extant clones
Cell Division and Mutation History
MACHINA
Cell Migration History
Label ancestral vertices by anatomical sites
Resolve clone tree ambiguities
Migration Graph
P
M1
M2
migration

MACHINA accurately infers clone trees and migration histories on simulated data

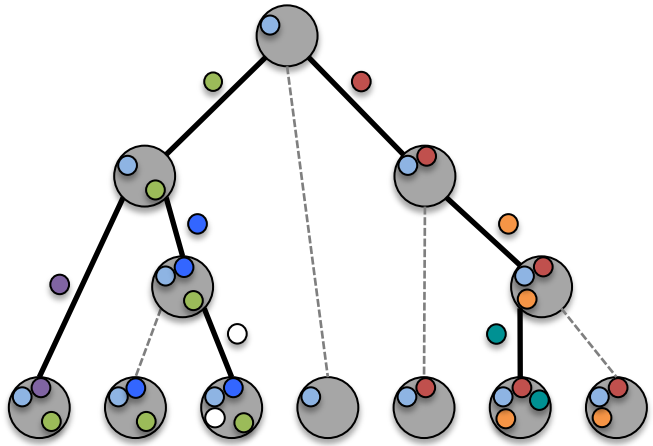




Conclusions



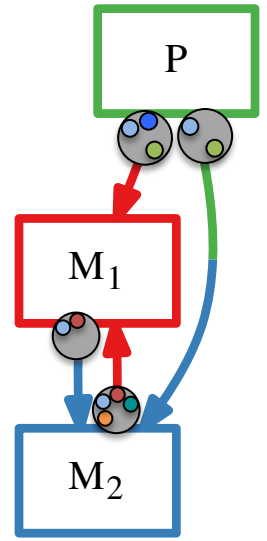
Cell division and mutation



Tumor Phylogeny Estimation

[**El-Kebir***, Oesper* et al., *ISMB* 2015/*Bioinformatics*]
[**El-Kebir** et al., *RECOMB* 2016]; [**El-Kebir***, Satas* et al., *Cell Systems* 2016]

Cell migration



Tumor Migration Analysis

[**El-Kebir** et al., *Nature Genetics* 2018]

Precise mathematical models are needed to describe the evolutionary process in cancer:

- Do not try to solve everything at once; it is OK to simplify and gradually add complexity
- Understanding combinatorial structure leads to a better understanding of the problem at hand
- This leads to better and efficient algorithms