Counting and Sampling Problems in Computational Biology

Mohammed El-Kebir, Jackie Oh, Yuanyuan Qi and Palash Sashittal

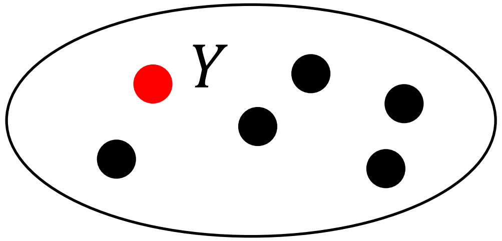
Department of Computer Science, University of Illinois, Urbana Champaign

MCW 2020, July 9th, 2020

Combinatorial Optimization in Computational Biology

- How similar are genome sequences? → Edit Distance
- What is the evolutionary history of all species? → Steiner Tree

Problem Π: Given input X find output Y such that Z.

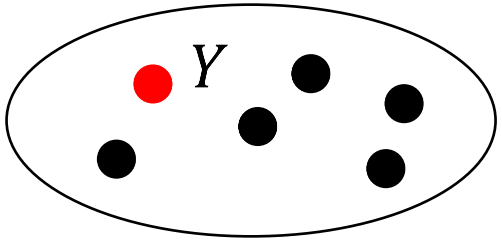


space of feasible solutions $\Pi(X)$

Combinatorial Optimization in Computational Biology

- How similar are genome sequences? → Edit Distance
- What is the evolutionary history of all species? → Steiner Tree

Problem Π: Given input X find output Y such that Z.



space of feasible solutions $\Pi(X)$

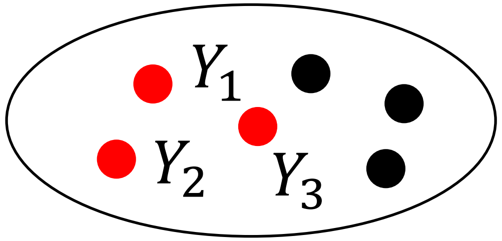
Challenge 1: Optimization problems inspired by biology often NP-hard

Integer linear programming

Combinatorial Optimization in Computational Biology

- How similar are genome sequences? → Edit Distance
- What is the evolutionary history of all species? → Steiner Tree

Problem Π: Given input X find output Y such that Z.



space of feasible solutions $\Pi(X)$

Challenge 1: Optimization problems inspired by biology often NP-hard

Integer linear programming

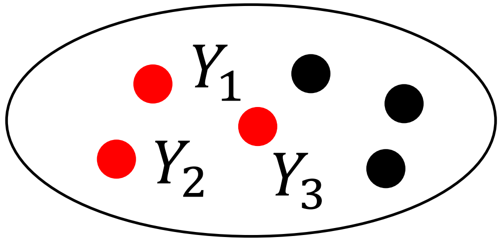
Challenge 2: Multiple solutions due to

- Problem itself (integer objective function)
- Interest in near-optimal solutions

Combinatorial Optimization in Computational Biology

- How similar are genome sequences? → Edit Distance
- What is the evolutionary history of all species? → Steiner Tree

Problem Π: Given input X
find output Y such that Z.



space of feasible solutions $\Pi(X)$

Challenge 1: Optimization problems inspired by biology often NP-hard

Integer linear programming

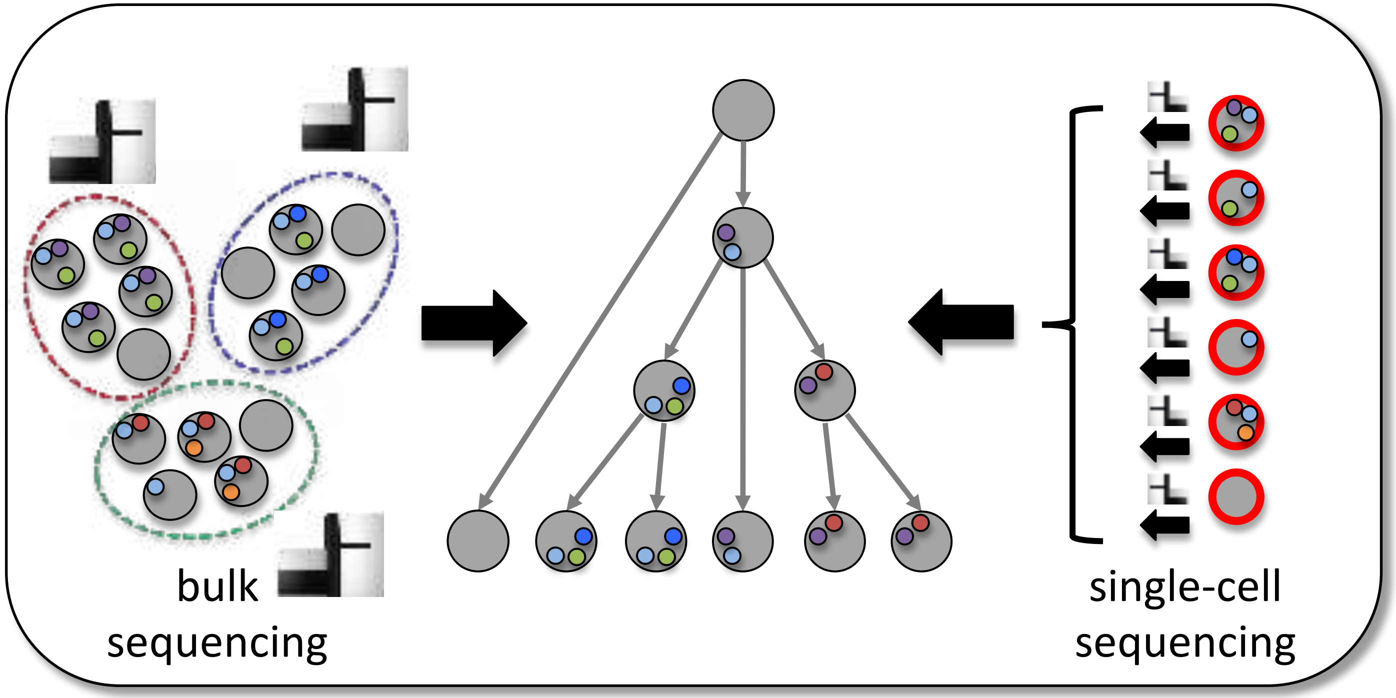
Challenge 2: Multiple solutions due to

- Problem itself
(integer objective function)
- Interest in near-optimal solutions

Satisfiability

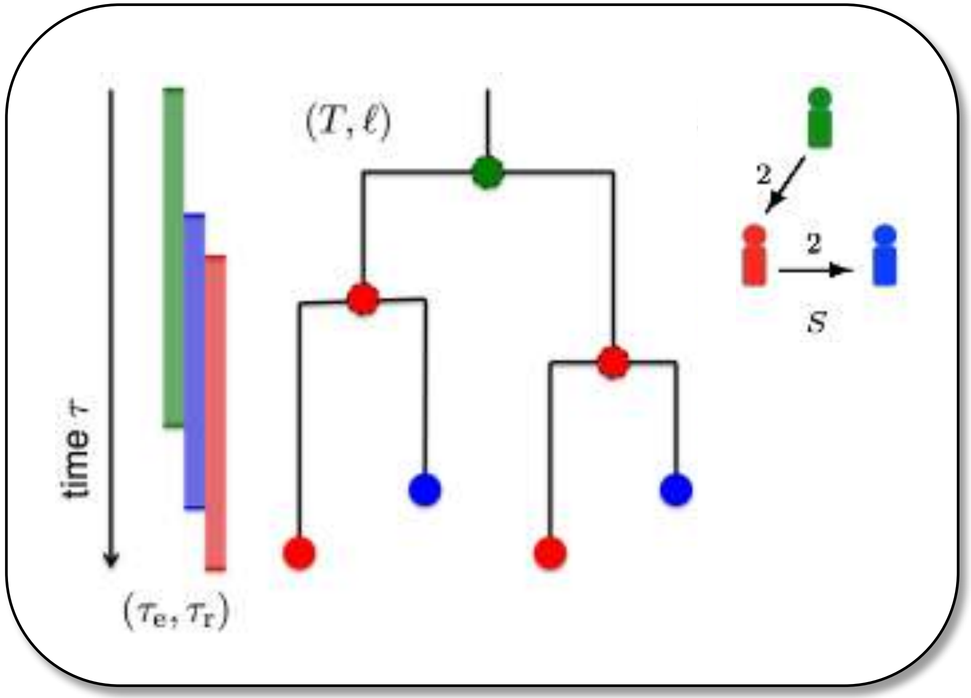
Outline

Solving problems in computational biology
via approximate model counting



bulk sequencing

single-cell sequencing



Reconstructing a tumor’s evolution
from sequencing data

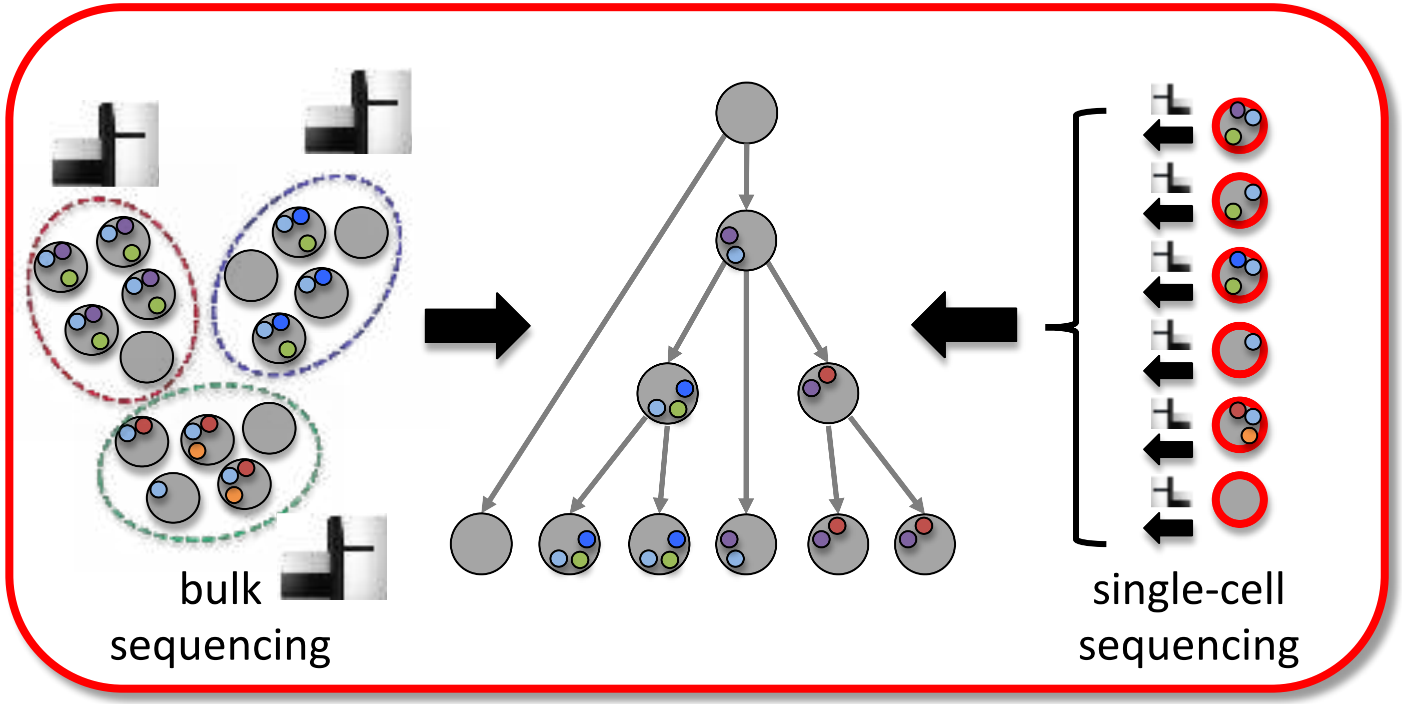
Reconstructing transmissions
during outbreaks

Outline

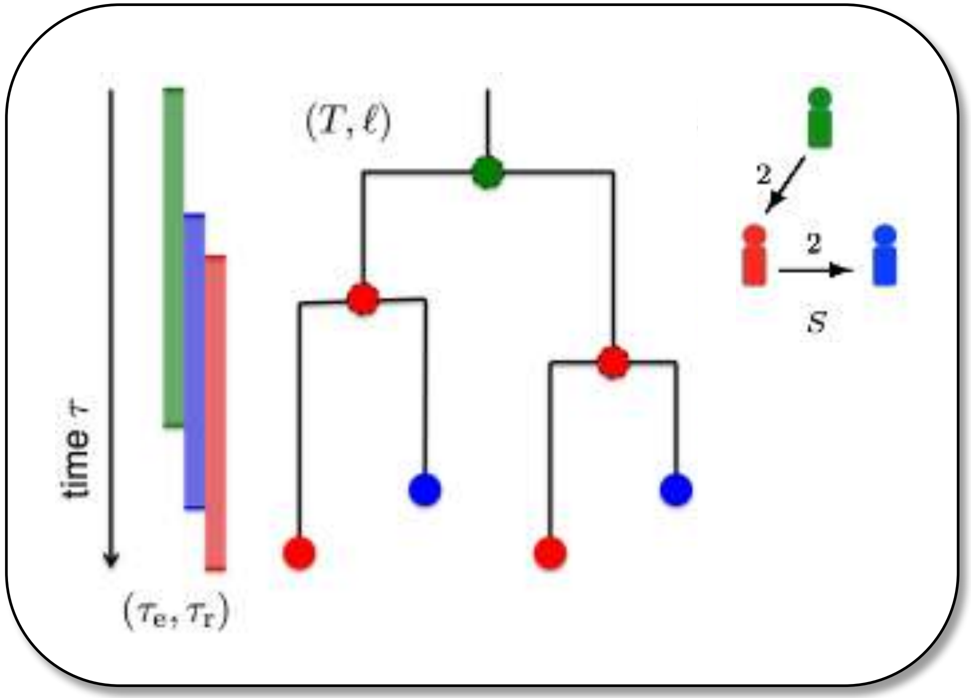
Solving problems in computational biology via approximate model counting

Yuanyuan Qi

Jackie Oh



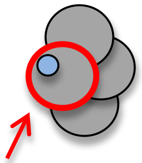
Reconstructing a tumor's evolution from sequencing data



Reconstructing transmissions during outbreaks

Cancer is an Evolutionary Process

Clonal Evolution Theory of Cancer
[Nowell, 1976]

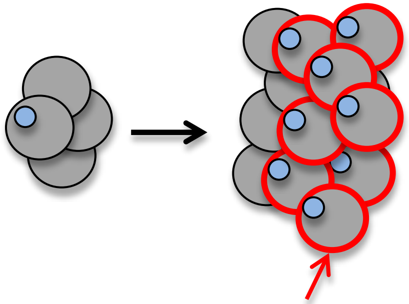


Founder
tumor cell
with somatic mutation: ●
(e.g. BRAF V600E)

Cancer is an Evolutionary Process

Clonal Evolution Theory of Cancer

[Nowell, 1976]

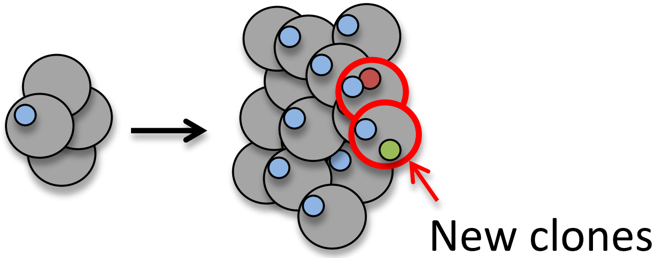


Clonal expansion

Cancer is an Evolutionary Process

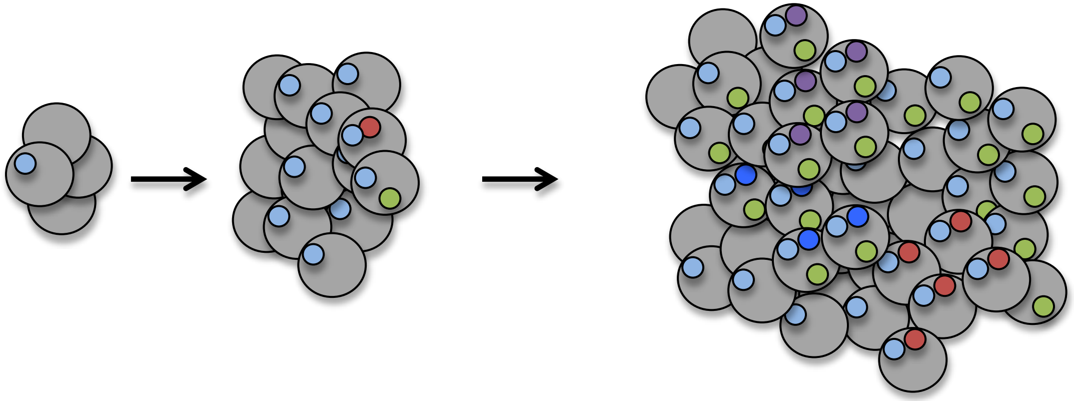
Clonal Evolution Theory of Cancer

[Nowell, 1976]

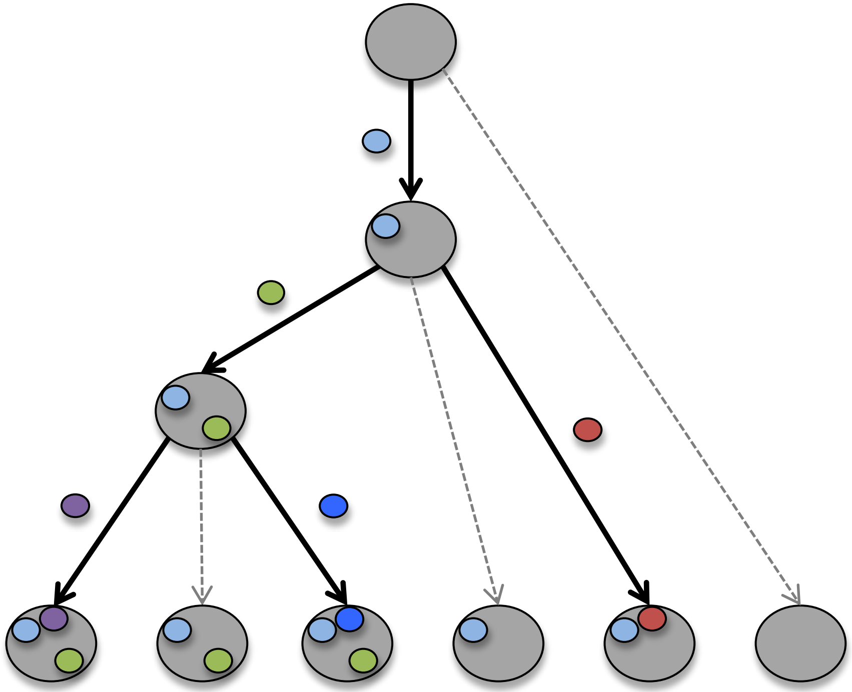


Cancer is an Evolutionary Process

Clonal Evolution Theory of Cancer
[Nowell, 1976]



Intra-Tumor Heterogeneity



Phylogenetic Tree ***T***

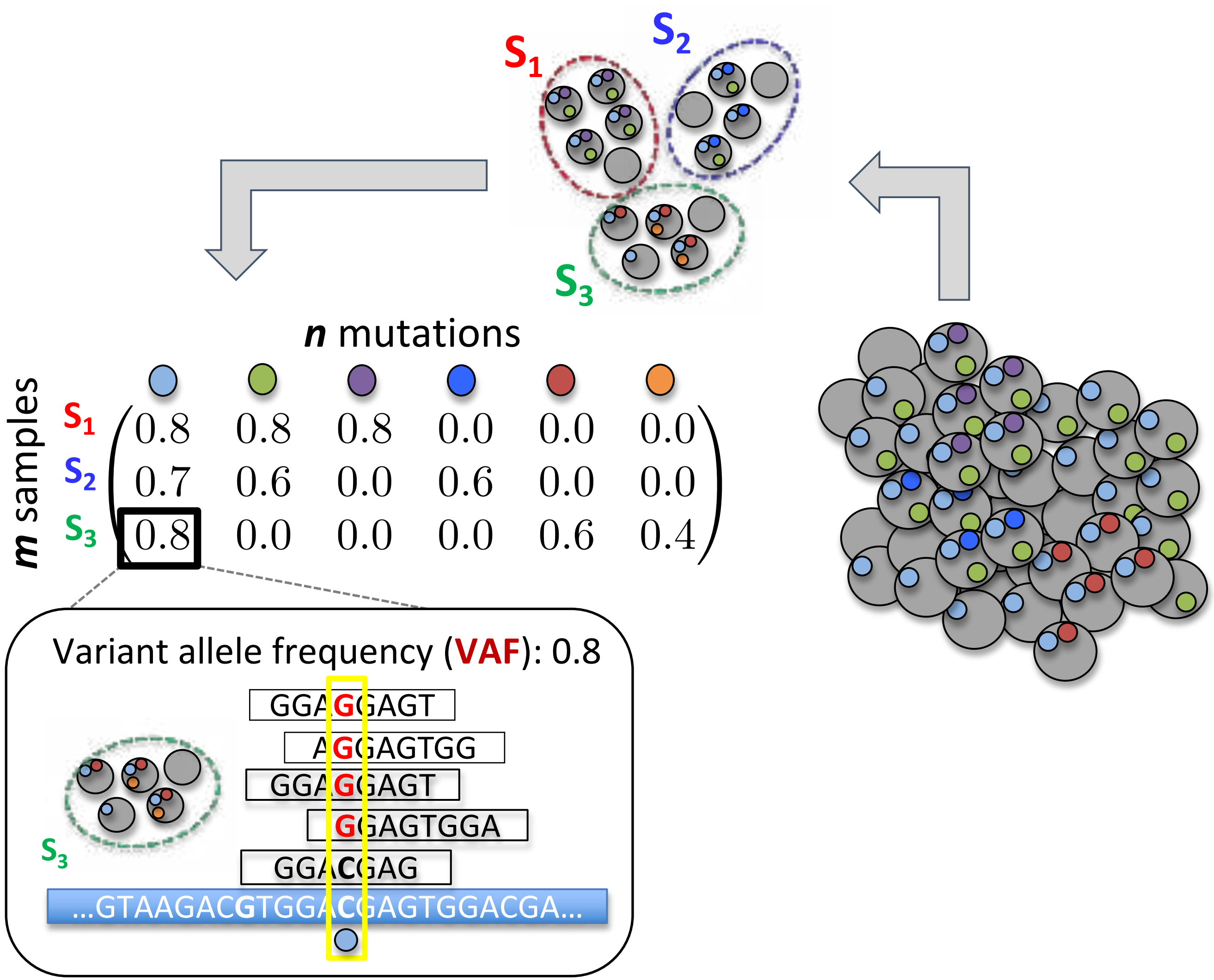
Identify treatment targets

Understand metastatic development

Compare evolutionary patterns across patients

DNA Sequencing of Tumors

Bulk DNA Sequencing



n **mutations**

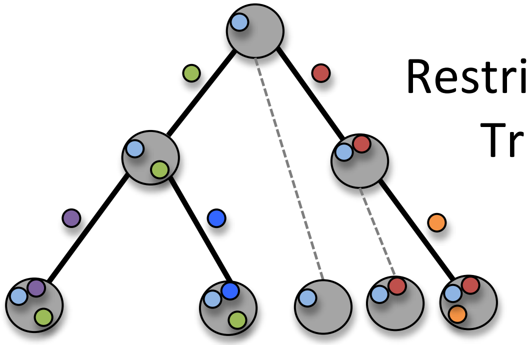
m **samples**

$$\begin{array}{c|cccccc} & ● & ● & ● & ● & ● & ● \\ \hline S_1 & 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ S_2 & 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ S_3 & \boxed{0.8} & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{array}$$

Variant allele frequency (**VAF**): 0.8

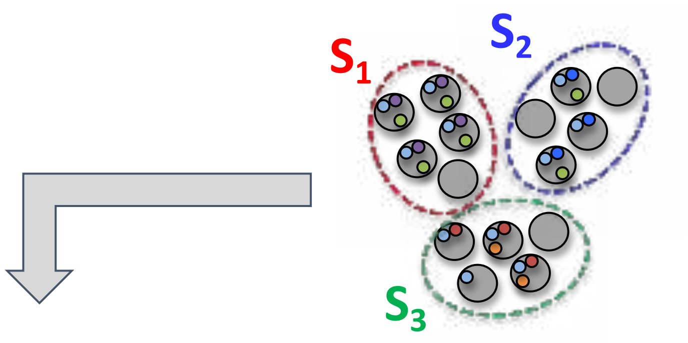
[El-Kebir et al., Bioinformatics/ISMB 2015]

Perfect Phylogeny Mixture (PPM)



Restricted PP Tree ***T***

1-1 ↕ **Equivalent**



n mutations

$$\begin{array}{c|cccccc} & ● & ● & ● & ● & ● & ● \\ \hline S_1 & 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ S_2 & 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ S_3 & 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{array}$$

m samples

Frequency Matrix ***F***

=

clones

$$\begin{array}{c|cccccc} & ● & ● & ● & ● & ● & ● \\ \hline S_1 & 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\ S_2 & 0.1 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 \\ S_3 & 0.2 & 0.0 & 0.0 & 0.0 & 0.2 & 0.4 \end{array}$$

m samples

Mixture Matrix ***U***

n mutations

$$\begin{array}{cccccc|c} ● & ● & ● & ● & ● & ● & \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & ● \\ 1 & 1 & 0 & 0 & 0 & 0 & ● \\ 1 & 1 & 1 & 0 & 0 & 0 & ● \\ 1 & 1 & 0 & 1 & 0 & 0 & ● \\ 1 & 0 & 0 & 0 & 1 & 0 & ● \\ 1 & 0 & 0 & 0 & 1 & 1 & ● \end{array}$$

clones

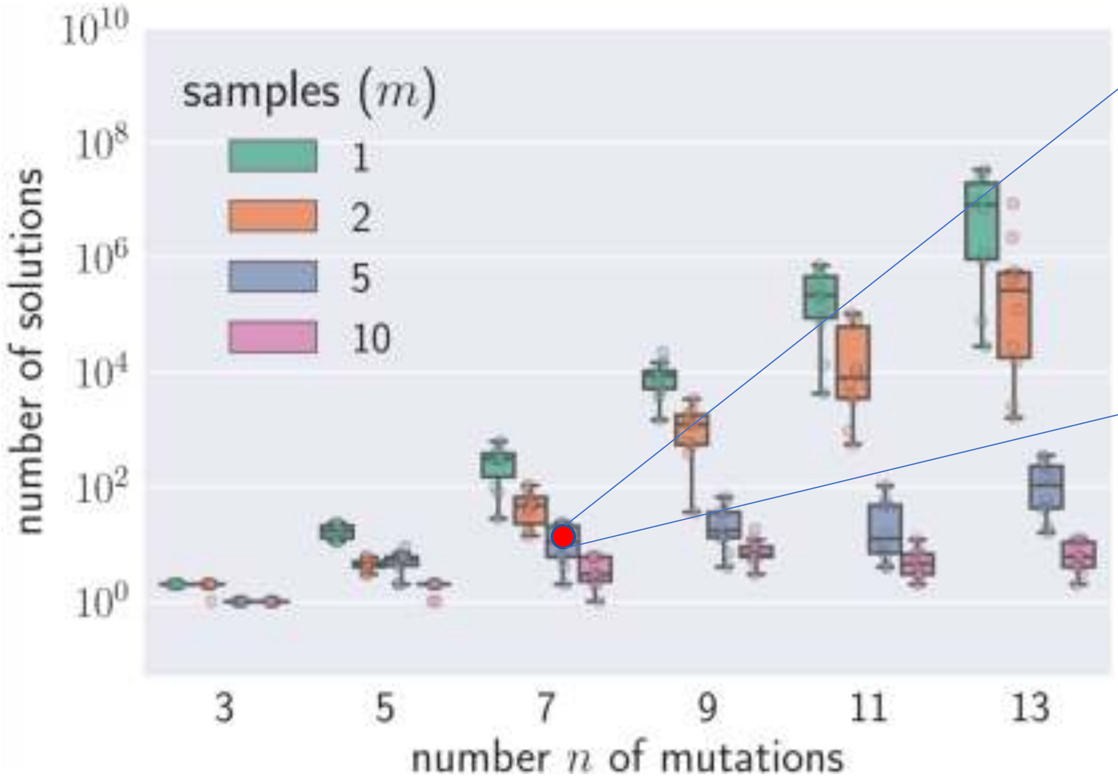
Restricted PP Matrix ***B***

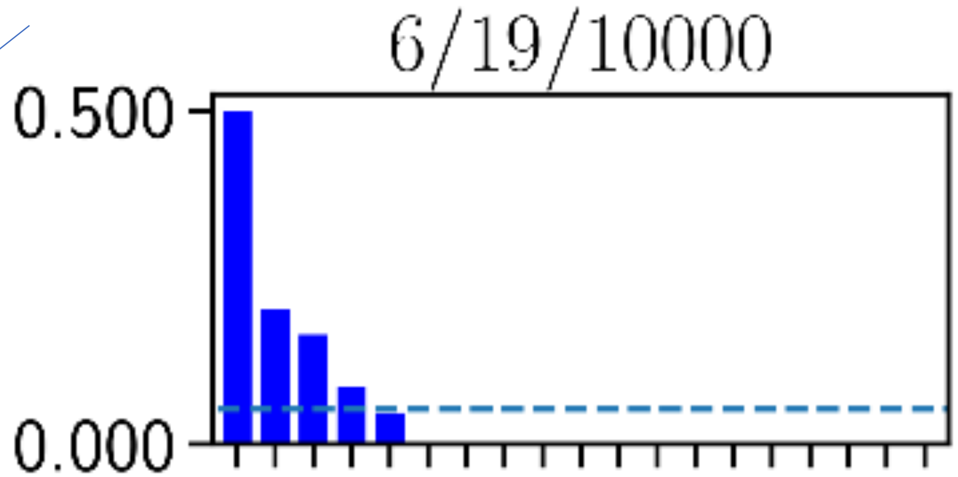
Perfect Phylogeny Mixture:

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

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

Sampling PPM Solutions





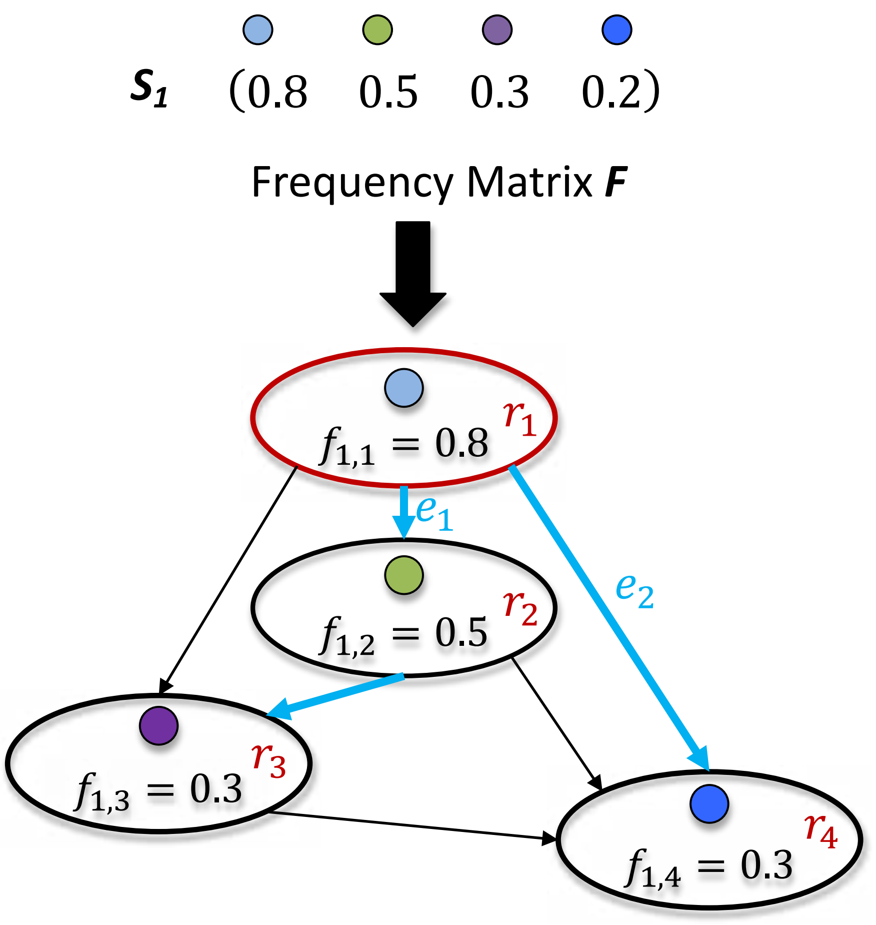
Sampling results by PhyloWGS
[Deshwar et al., 2015]

- PPM is NP-Complete (El-Kebir et al., 2015)
- #PPM is #P-Complete (Qi et al., 2019)

[Qi et al., Algorithms in Molecular Biology, 2019]

SAT Formulation

Sum condition: frequency of parent >= sum of frequencies of children



Frequency Matrix ***F***

Ancestry Graph ***G***

$(r_1 \vee r_2 \vee r_3 \vee r_4)$
$(\neg r_1 \vee \neg r_2)$
$(\neg r_1 \vee \neg r_3)$
$(\neg r_1 \vee \neg r_4)$
$(\neg r_2 \vee \neg r_3)$
$(\neg r_2 \vee \neg r_4)$
$(\neg r_3 \vee \neg r_4)$

$(r_1 \vee e_1 \vee e_2)$
$(\neg r_1 \vee \neg e_1)$
$(\neg r_1 \vee \neg e_2)$
$(\neg e_1 \vee \neg e_2)$

⋮

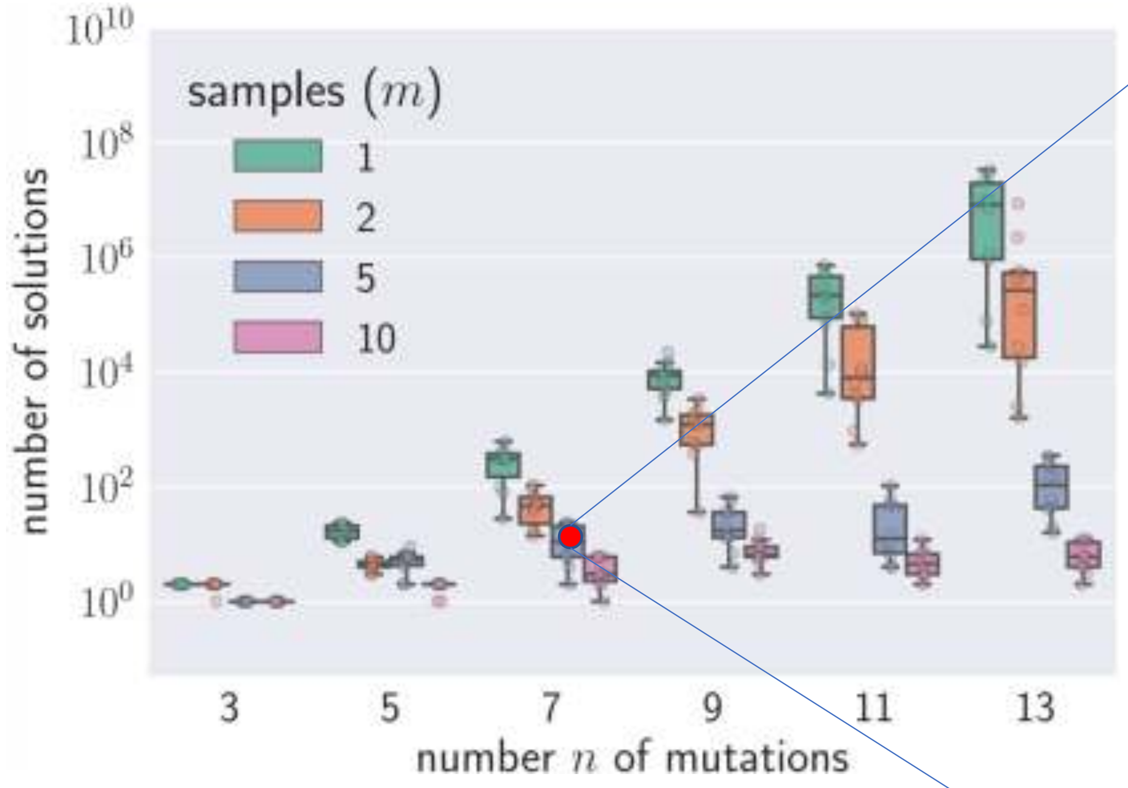
⋮

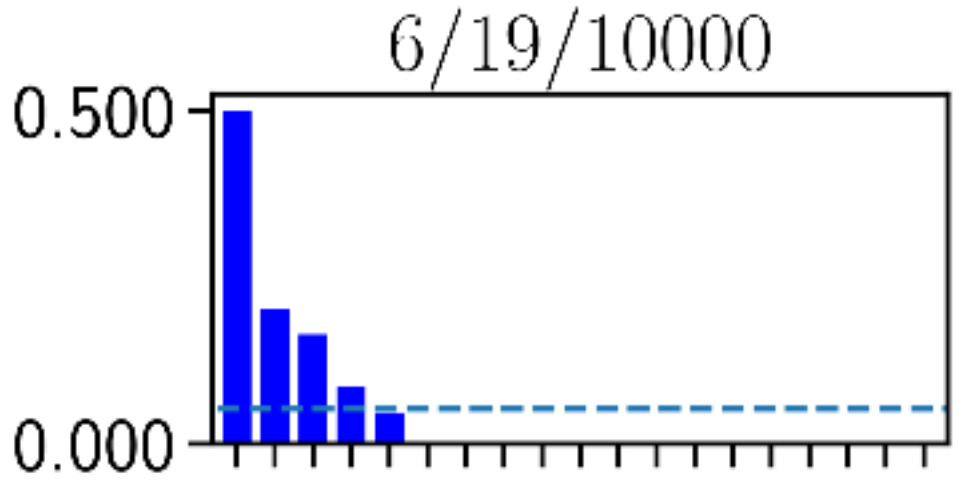
- Constraints:
  - Unique root
  - Unique parents
  - Cycle prevention
  - Sum condition

- Complexity:
  - $O(n|E| + Nm|E|)$ variables
  - $O(|E|^2 + Nm|E|)$ clauses

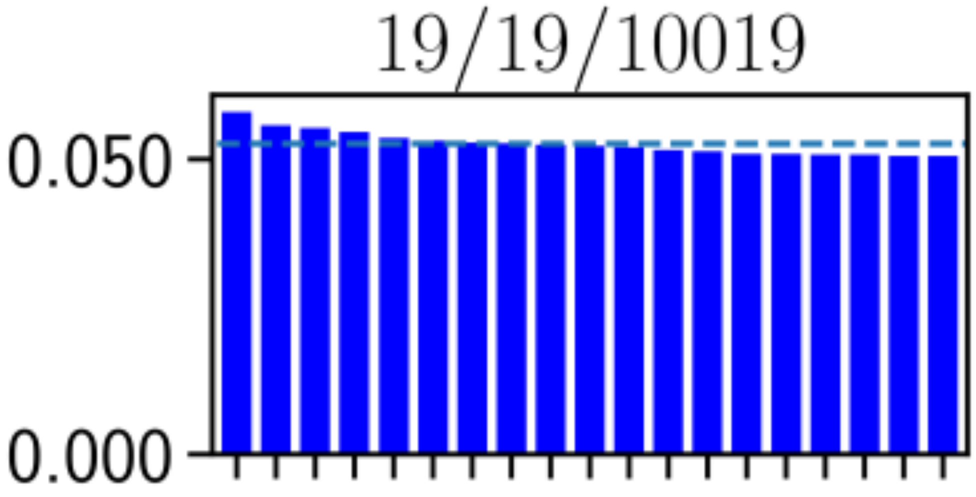
[Qi and El-Kebir, In preparation]

Sampling using UniGen v2





Sampling results by PhyloWGS
[Deshwar et al., 2015]

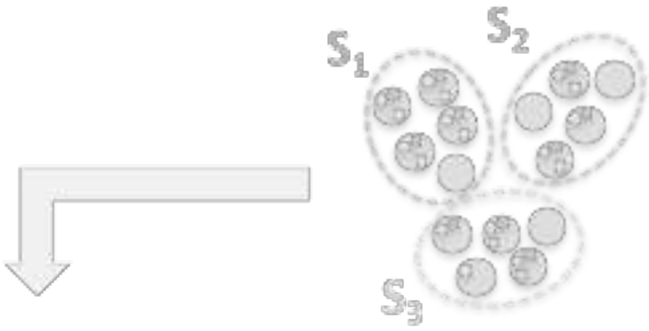


Sampling results using SAT formulation

[Qi and El-Kebir, In preparation]

DNA Sequencing of Tumors (2/2)

Bulk DNA Sequencing ($)

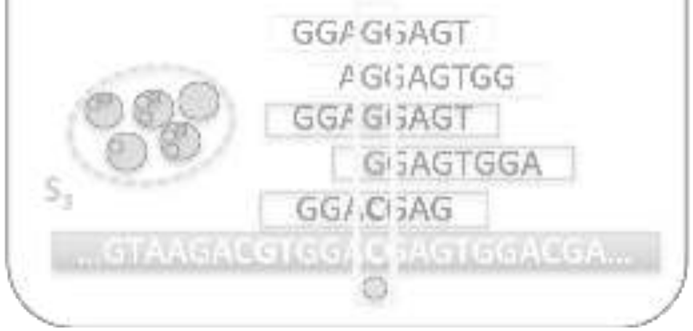


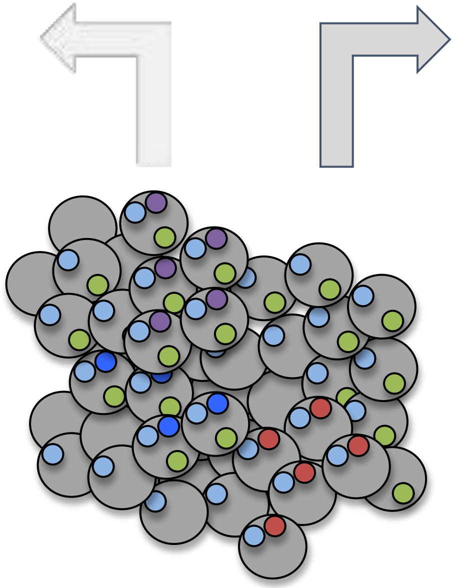
n mutations

m samples

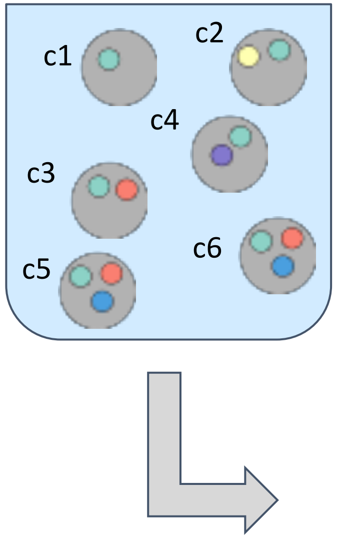
$$\begin{array}{c|cccccc} & ● & ● & ● & ● & ● & ● \\ \hline s_1 & 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ s_2 & 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ s_3 & \boxed{0.8} & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{array}$$

Variant allele frequency (VAF): 0.8





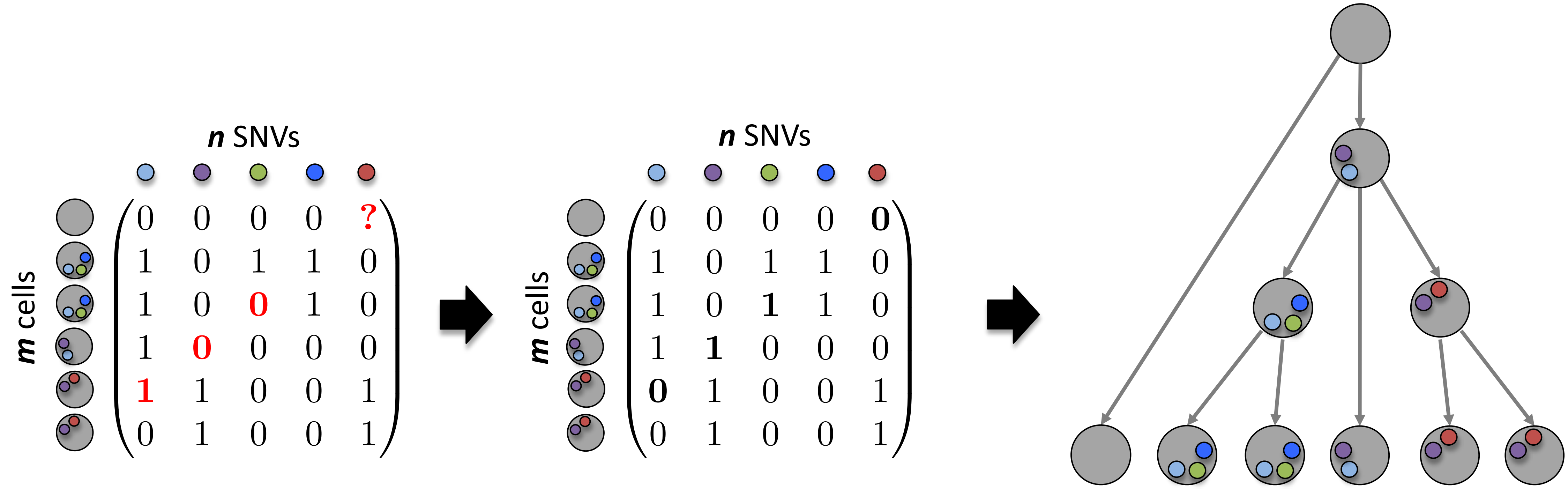
Single-cell DNA Sequencing ($$$)



	●	●	●	●	●
c1	1	0	0	0	0
c2	1	1	🟥 1	0	0
c3	🟥 0	0	0	1	0
c4	1	0	🟥 0	1	0
c5	1	🟥 ?	0	1	1
c6	1	0	0	1	🟥 0

🟥 0	False Negative
🟥 1	False Positive
🟥 ?	Missing Data

Phylogeny Inference from Single-cell Data



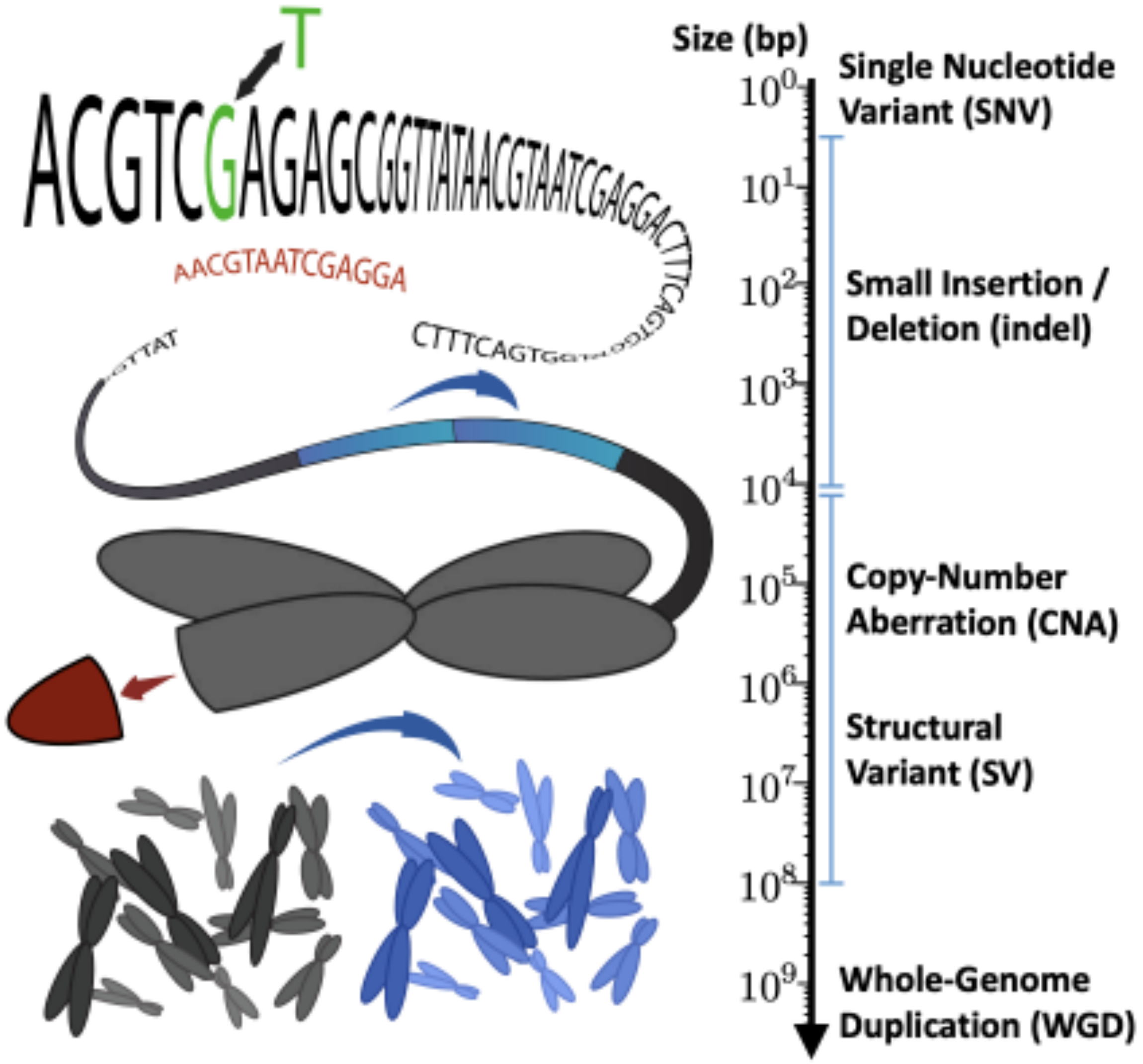
$$\begin{pmatrix} 0 & 0 & 0 & 0 & ? \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \Rightarrow \begin{pmatrix} 0 & 0 & 0 & 0 & \mathbf{0} \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & \mathbf{1} & 1 & 0 \\ 1 & \mathbf{1} & 0 & 0 & 0 \\ \mathbf{0} & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \Rightarrow$$

Input Matrix ***D*** Binary Matrix ***B*** Phylogenetic Tree ***T***

Goal: Given single-cell sequencing data, sample possible phylogenetic trees

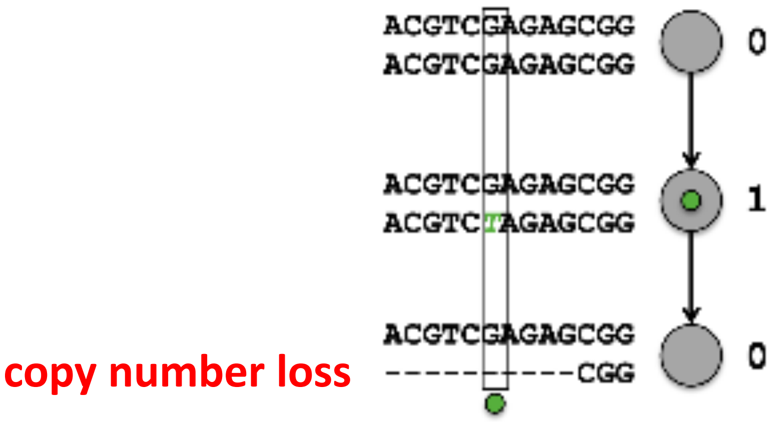
Requirement: Evolutionary model for somatic mutations

Infinite Sites Assumption vs *k*-Dollo Model


T
ACGTCGAGAGCGGTTATAACGTAATCGAGGACTTTCAGTG
AACGTAATCGAGGA
CTTTCAGTG
Size (bp)
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^7
10^8
10^9
Single Nucleotide Variant (SNV)
Small Insertion / Deletion (indel)
Copy-Number Aberration (CNA)
Structural Variant (SV)
Whole-Genome Duplication (WGD)

[El-Kebir, Bioinformatics/ECCB 2018]

SNVs can be **lost** due to CNAs



Infinite Sites Assumption:

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

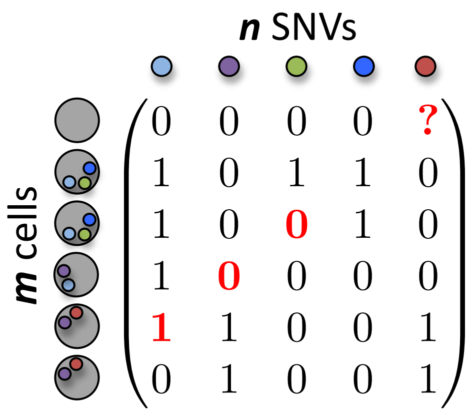
k-Dollo Parsimony Model:

- No parallel evolution of SNVs
- SNV can be lost up to k times

We will use the 1-Dollo model, where k=1

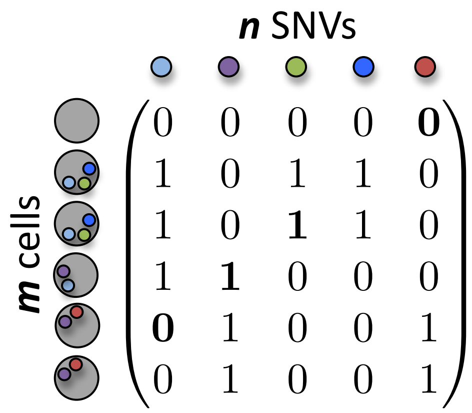
k-Dollo Phylogeny Flip and Cluster (k-DPFC) problem. Given matrix $D \in \{0,1,?\}^{m \times n}$, error rates $\alpha, \beta \in [0,1]$, integers $k, s, t \in \mathbb{N}$, find matrix $B \in \{0,1\}^{m \times n}$ and tree T such that: (1) B has at most s unique rows and at most t unique columns; (2)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

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



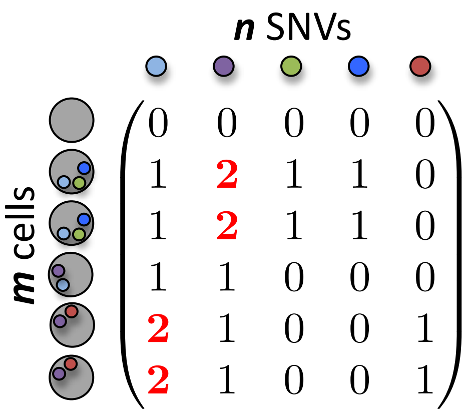
$$\begin{pmatrix} 0 & 0 & 0 & 0 & ? \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

Input Matrix ***D***



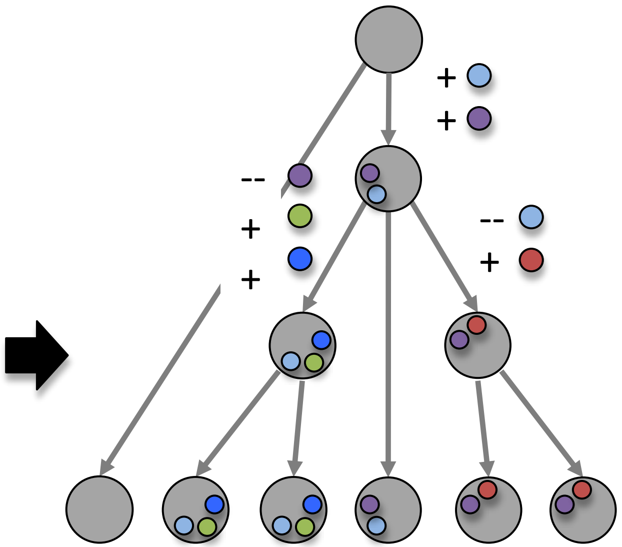
$$\begin{pmatrix} 0 & 0 & 0 & 0 & \mathbf{0} \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & \mathbf{1} & 1 & 0 \\ 1 & \mathbf{1} & 0 & 0 & 0 \\ \mathbf{0} & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

Binary Matrix ***B***



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

k*-Dollo Completion *A



k*-Dollo Phylogeny *T

[El-Kebir, Bioinformatics/ECCB 2018]

SAT Formulation

Variables

False positive and false negatives

$$\alpha_{i,j}, i \in [m], j \in [n]$$
$$\beta_{i,j}, i \in [m], j \in [n]$$

Losses

$$d_{i,j}, i \in [m], j \in [n]$$

Clustering (determine duplicate rows/columns)

$$c_j, j \in [m] \quad x_{i,k,l}, i \in [m], k,l \in [n], k < l$$
$$r_l, l \in [n] \quad y_{i,j,k}, i,j \in [m], l \in [n], i < j$$
$$p_{i,j}, i,j \in [m], i < j$$
$$q_{k,l}, k,l \in [n], k < l$$

Number of variables: $O(m^2n + mn^2)$

Clauses

Enforce absence of forbidden submatrices

- Enforce that any submatrix of A cannot equal any of the 25 submatrices
- Allow this constraint to be violated if a row or column of the submatrix is a duplicate

$$\neg \begin{bmatrix} \neg\alpha_{1,1} & \neg\beta_{1,2} \wedge \neg d_{1,2} \\ \neg\beta_{2,1} \wedge \neg d_{2,1} & d_{2,2} \\ \beta_{3,1} & \beta_{3,2} \end{bmatrix} \vee \boxed{c_1 \vee c_2 \vee r_1 \vee r_2 \vee r_3}$$

Determine whether two rows or columns are equal

Bound the number of false positives and false negatives

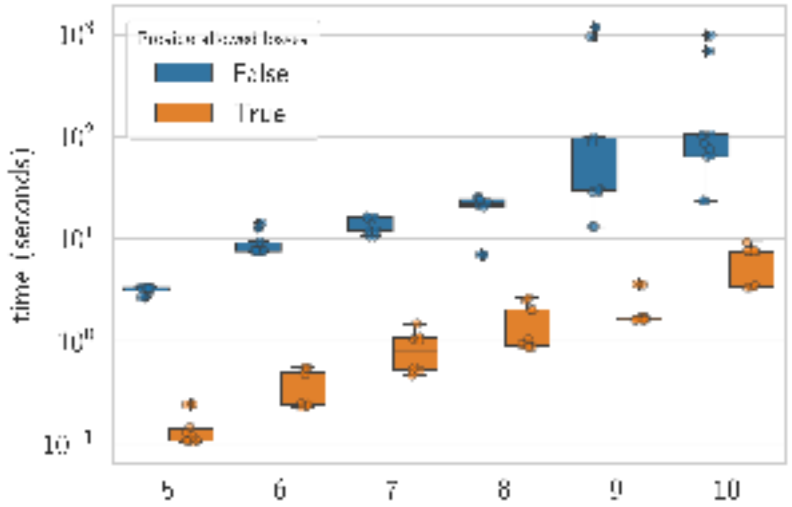
Enforce the number of cell and mutation clusters

- Encode sum of binary variables as a binary vector using a half/full adder

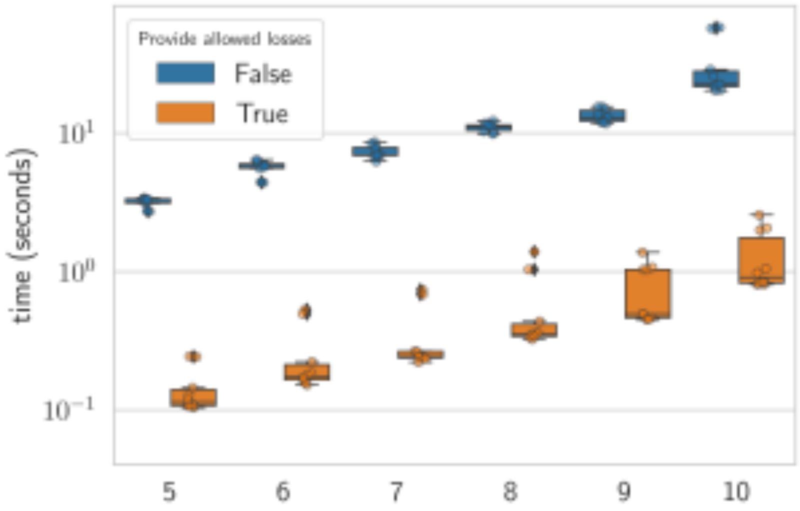
Number of clauses: $O(m^3n^2 + n^3)$

[Oh and El-Kebir, In preparation]

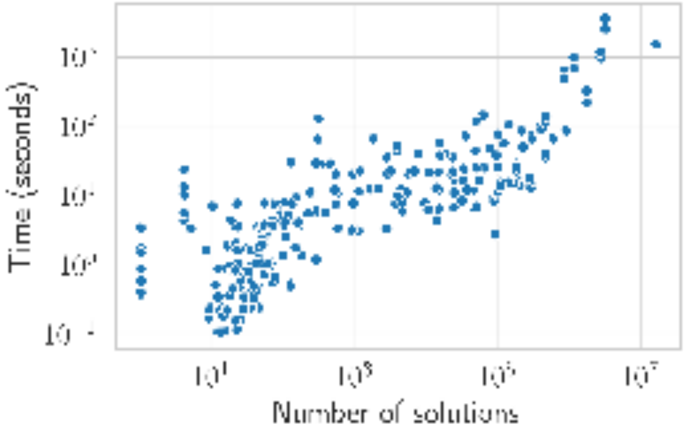
Results



number m of cells ($n = 5$ mutations)

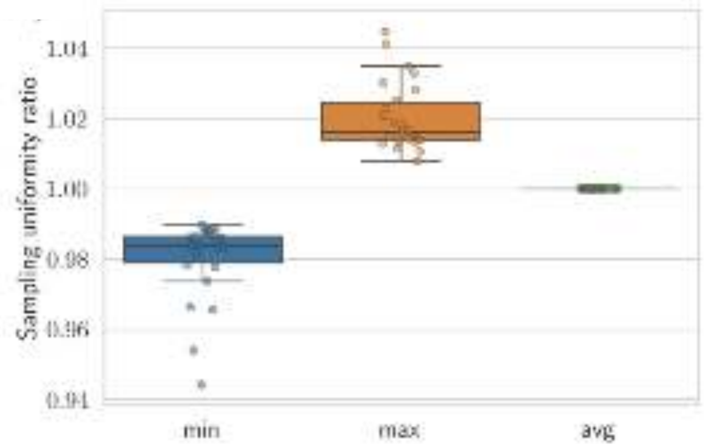


number n of mutations ($m = 5$ cells)



Simulations show:

- Runtime is reduced by providing the set of known allowed losses
  - Supplementing SCS data with copy number data could help improve runtime
- Runtime is roughly proportional to the number of solutions to a given formula
- DolloSAT is not yet feasible for real datasets (m > 100 cells)
  - Currently working on a cutting planes approach to reduce runtime

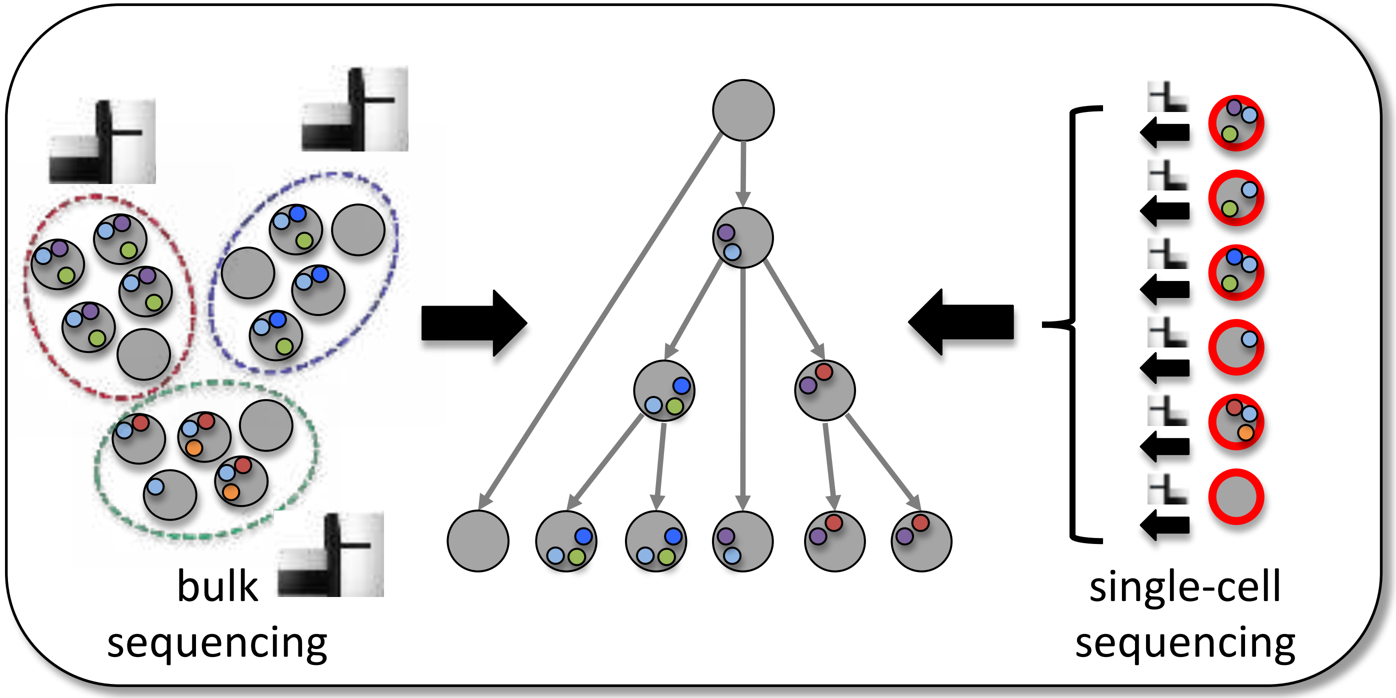


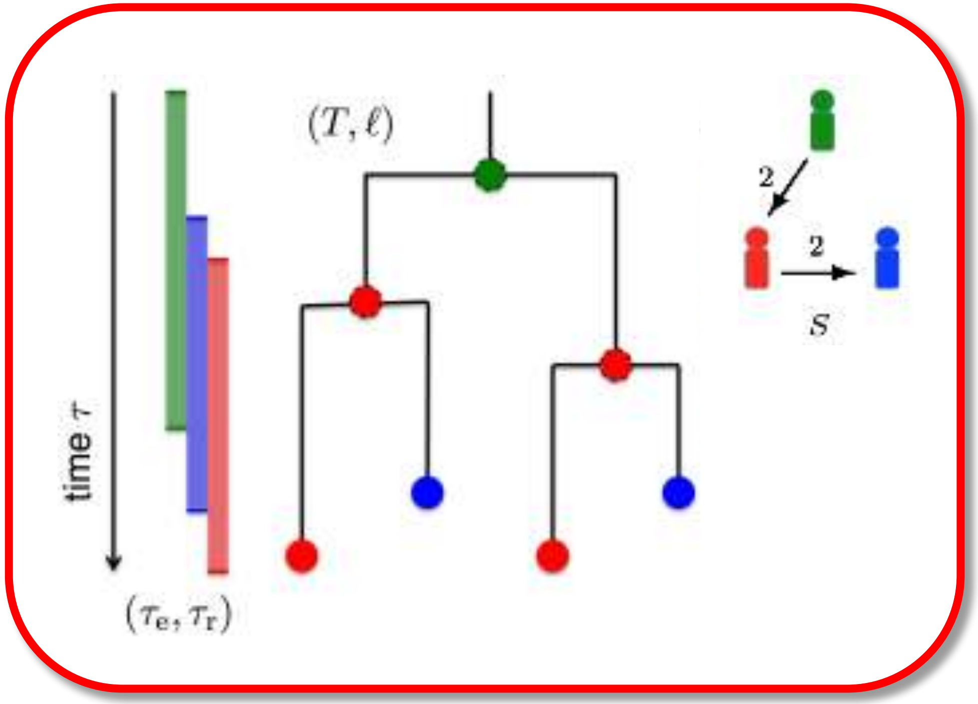
[Oh and El-Kebir, In preparation]

Outline

Solving problems in computational biology via approximate model counting

Palash Sashittal

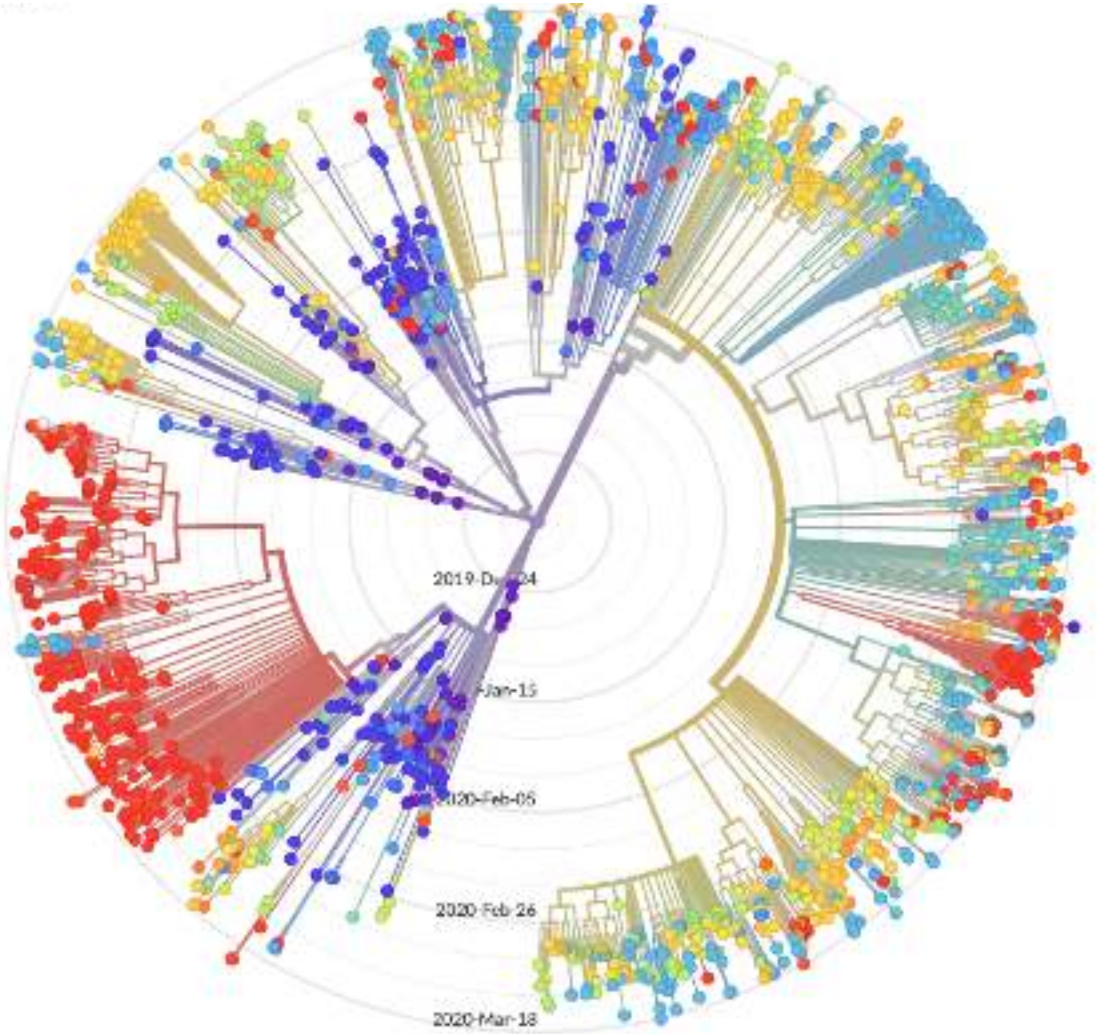




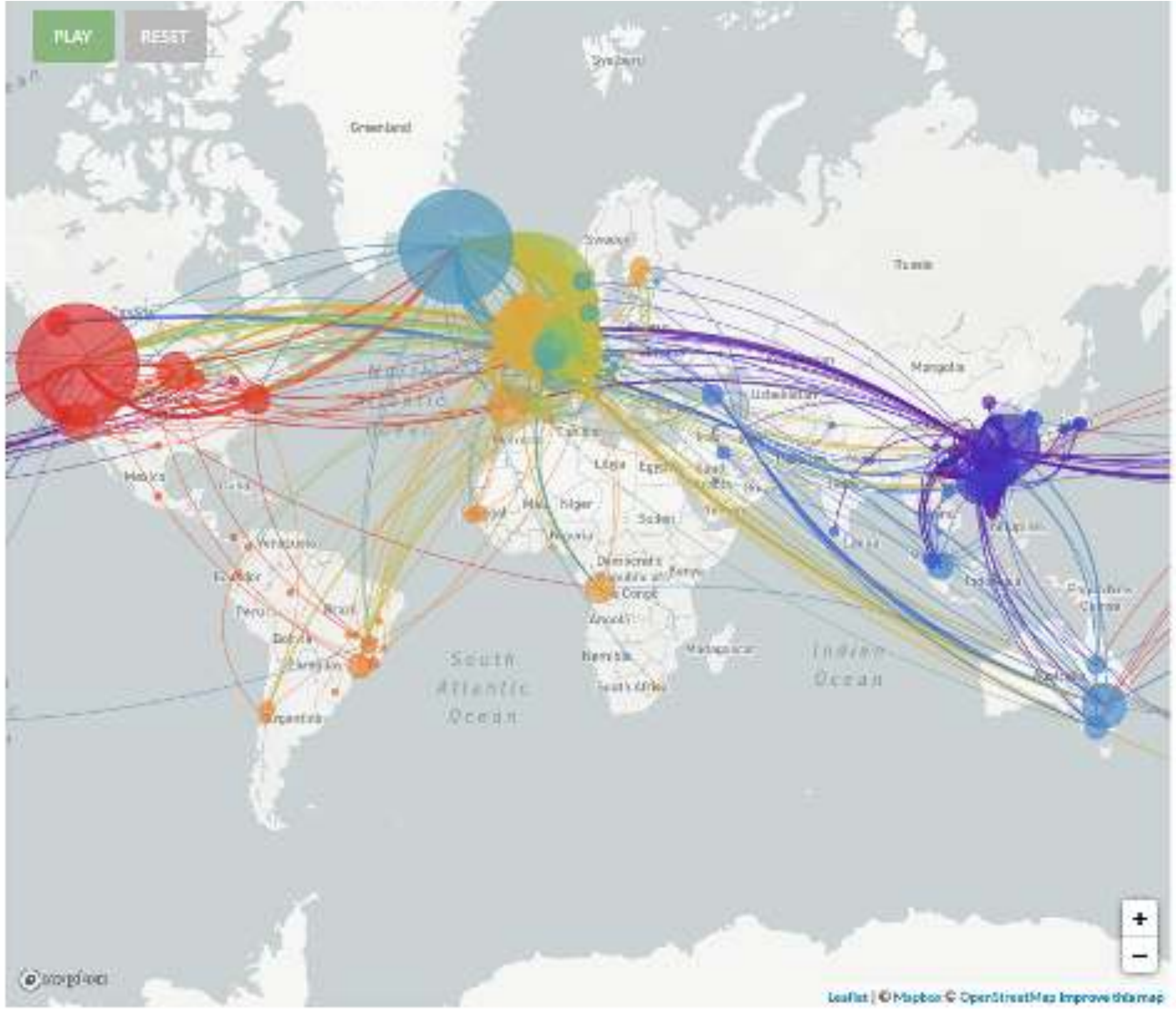
Reconstructing a tumor’s evolution from sequencing data

Reconstructing transmissions during outbreaks

Evolution & Transmission during an Outbreak



https://nextstrain.org/ncov?l=radial

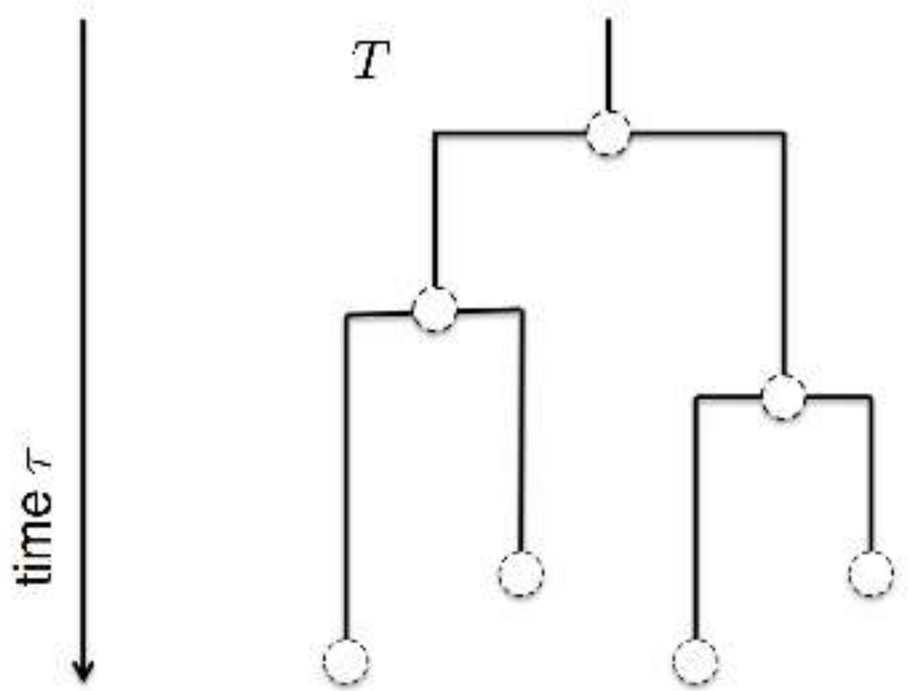


Evolutionary history: Phylogeny

Transmission history: Transmission graph

Directed Transmission Inference (DTI): Input

Timed Phylogeny:
A rooted tree T whose vertices are labeled by time-stamps $\tau : V(T) \longrightarrow \mathbb{R}^{\geq 0}$ s.t. $\tau(u) < \tau(v)$ for all pairs (u, v) where u is an ancestor of v.

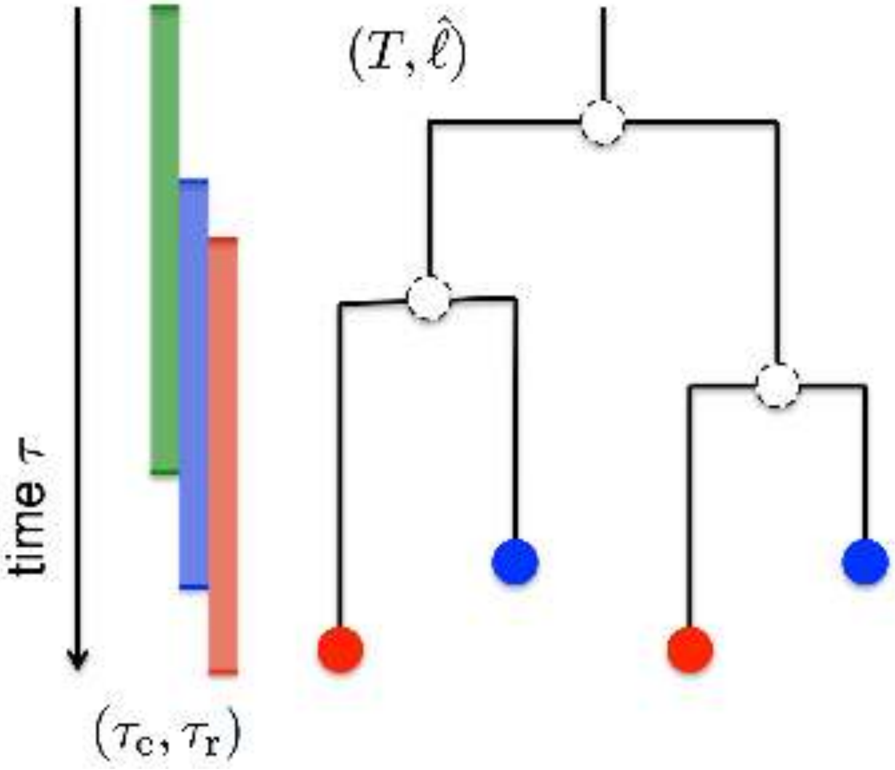


[Sashittal and El-Kebir, Bioinformatics/ISMB 2020]

Directed Transmission Inference (DTI): Input

Timed Phylogeny:
A rooted tree T whose vertices are labeled by time-stamps $\tau : V(T) \longrightarrow \mathbb{R}^{\geq 0}$ s.t. $\tau(u) < \tau(v)$ for all pairs (u, v) where u is an ancestor of v.

Epidemiological Data:
For each host $s \in \Sigma$, we have an entrance time $\tau_e(s)$ and removal time $\tau_r(s)$ and leaves are labeled by hosts.



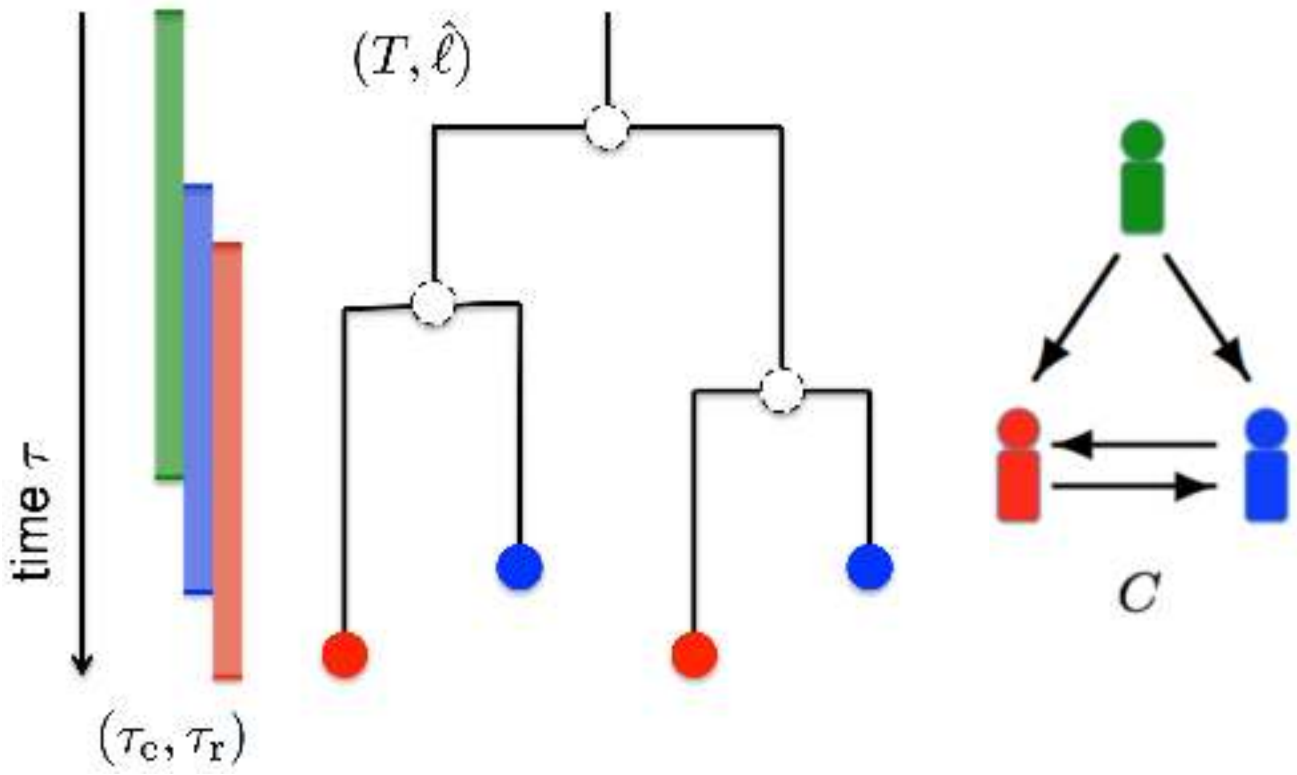
[Sashittal and El-Kebir, Bioinformatics/ISMB 2020]

Directed Transmission Inference (DTI): Input

Timed Phylogeny:
A rooted tree T whose vertices are labeled by time-stamps $\tau : V(T) \longrightarrow \mathbb{R}^{\geq 0}$ s.t. $\tau(u) < \tau(v)$ for all pairs (u, v) where u is an ancestor of v.

Epidemiological Data:
For each host $s \in \Sigma$, we have an entrance time $\tau_e(s)$ and removal time $\tau_r(s)$ and leaves are labeled by hosts.

Contact Map:
A directed graph with vertex set given by the set of hosts Σ indicating putative transmission pairs.



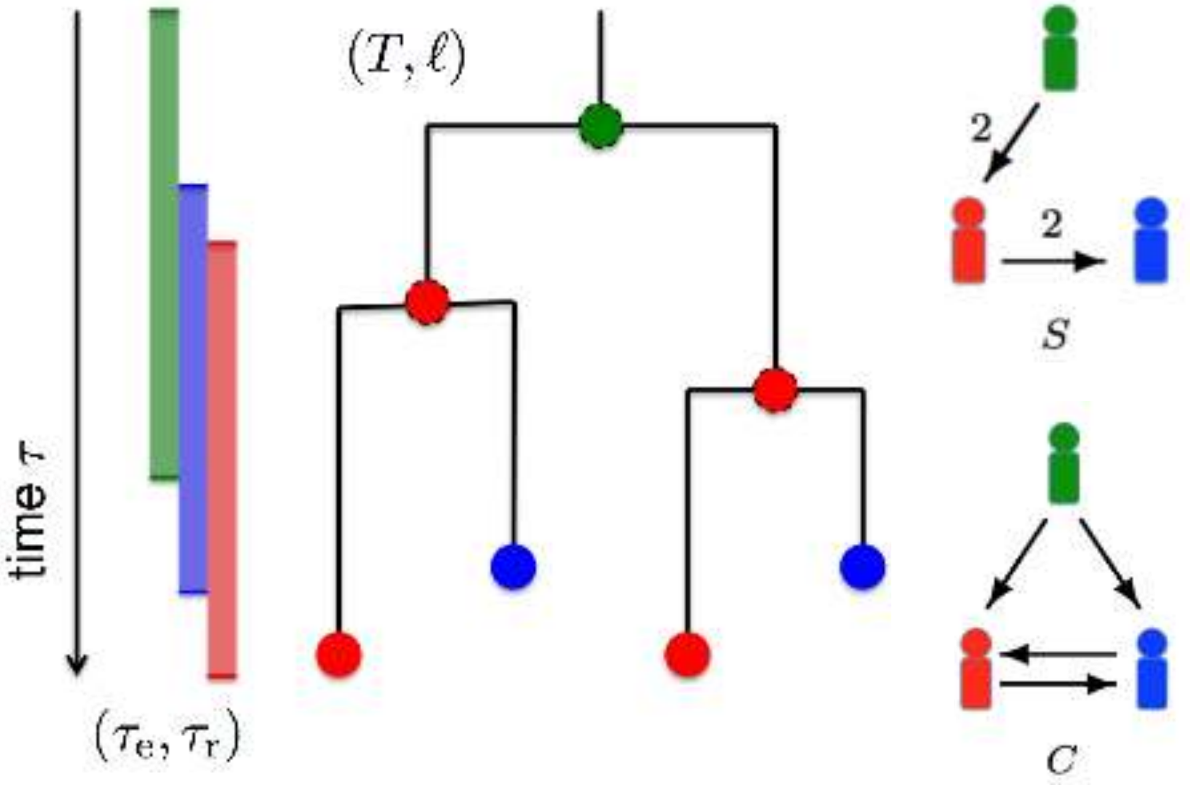
[Sashittal and El-Kebir, Bioinformatics/ISMB 2020]

Directed Transmission Inference (DTI): Output

Timed Phylogeny:
A rooted tree T whose vertices are labeled by time-stamps $\tau : V(T) \longrightarrow \mathbb{R}^{\geq 0}$ s.t. $\tau(u) < \tau(v)$ for all pairs (u, v) where u is an ancestor of v.

Epidemiological Data:
For each host $s \in \Sigma$, we have an entrance time $\tau_e(s)$ and removal time $\tau_r(s)$ and leaves are labeled by hosts.

Contact Map:
A directed graph with vertex set given by the set of hosts Σ indicating putative transmission pairs.



Internal Vertex Labeling and Transmission Tree:
A host labeling of a timed phylogeny T is a function $\ell : L(T) \longrightarrow \Sigma$, assigning a host $\ell(u)$ to each vertex u of T such that the resulting transmission network S is a spanning tree of the contact map C.

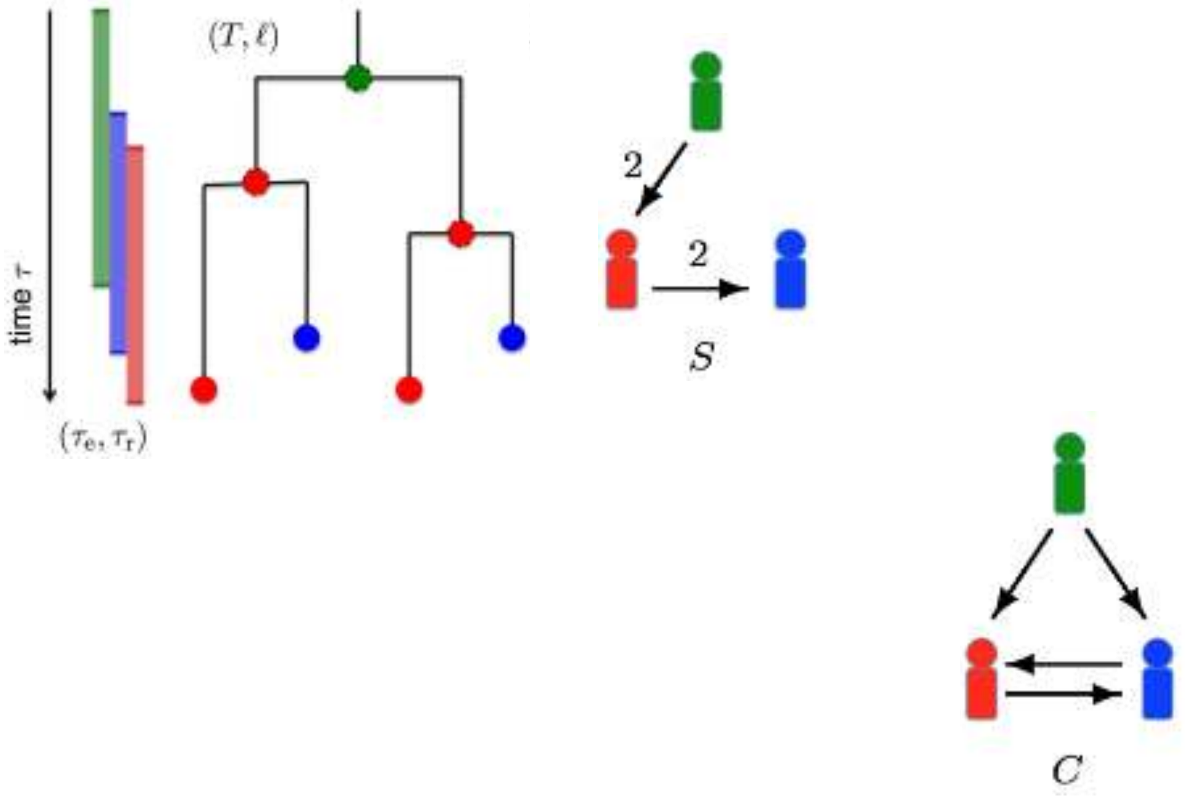
[Sashittal and El-Kebir, Bioinformatics/ISMB 2020]

Directed Transmission Inference (DTI): Output

Timed Phylogeny:
A rooted tree T whose vertices are labeled by time-stamps $\tau : V(T) \longrightarrow \mathbb{R}^{\geq 0}$ s.t. $\tau(u) < \tau(v)$ for all pairs (u, v) where u is an ancestor of v.

Epidemiological Data:
For each host $s \in \Sigma$, we have an entrance time $\tau_e(s)$ and removal time $\tau_r(s)$ and leaves are labeled by hosts.

Contact Map:
A directed graph with vertex set given by the set of hosts Σ indicating putative transmission pairs.



Internal Vertex Labeling and Transmission Tree:
A host labeling of a timed phylogeny T is a function $\ell : L(T) \longrightarrow \Sigma$, assigning a host $\ell(u)$ to each vertex u of T such that the resulting transmission network S is a spanning tree of the contact map C.

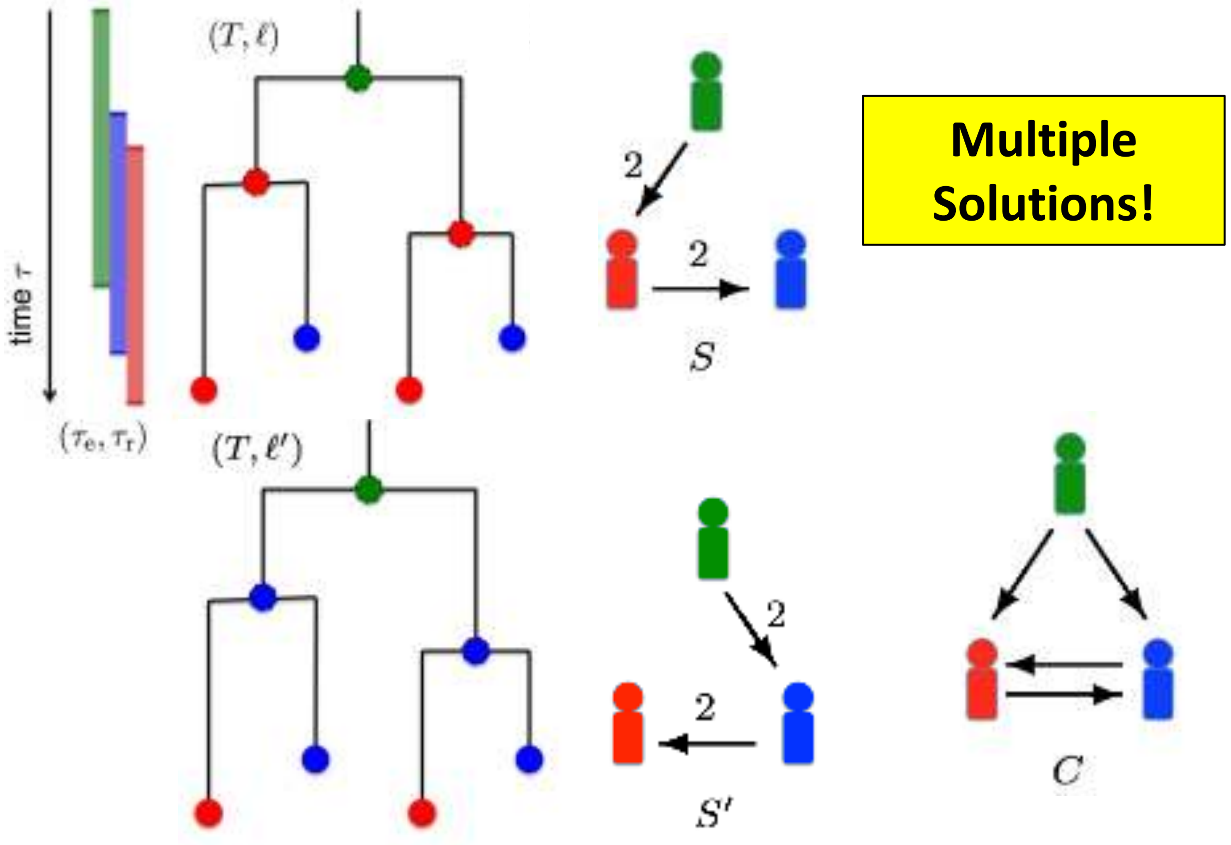
[Sashittal and El-Kebir, Bioinformatics/ISMB 2020]

Directed Transmission Inference (DTI): Output

Timed Phylogeny:
A rooted tree T whose vertices are labeled by time-stamps $\tau : V(T) \longrightarrow \mathbb{R}^{\geq 0}$ s.t. $\tau(u) < \tau(v)$ for all pairs (u, v) where u is an ancestor of v.

Epidemiological Data:
For each host $s \in \Sigma$, we have an entrance time $\tau_e(s)$ and removal time $\tau_r(s)$ and leaves are labeled by hosts.

Contact Map:
A directed graph with vertex set given by the set of hosts Σ indicating putative transmission pairs.



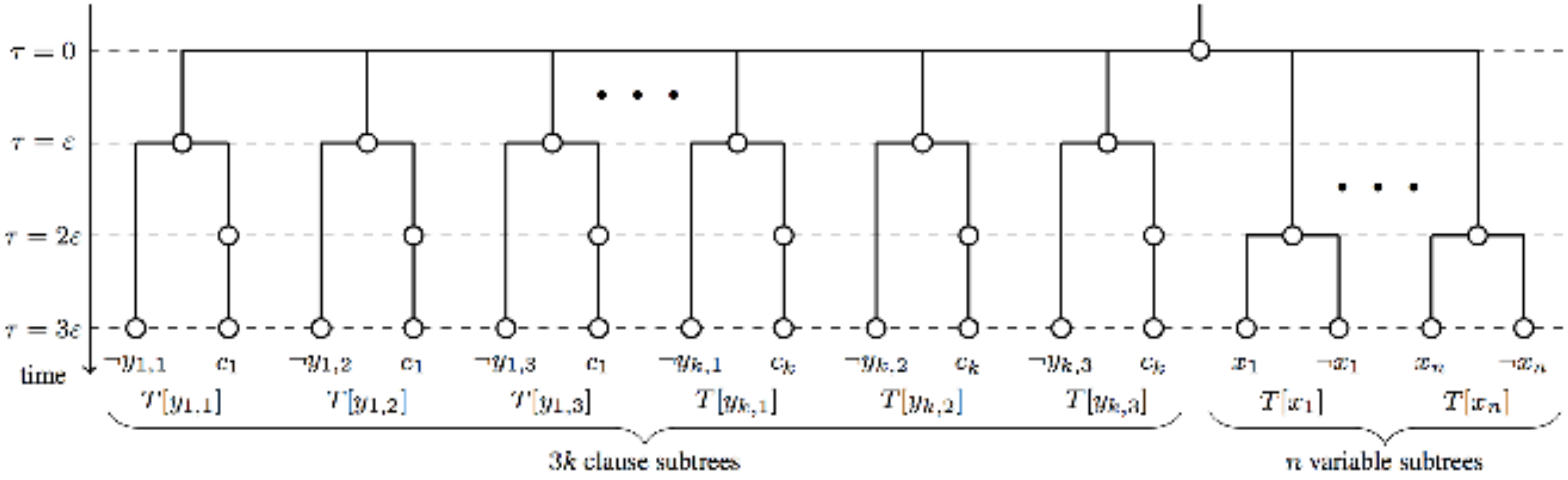
Multiple Solutions!

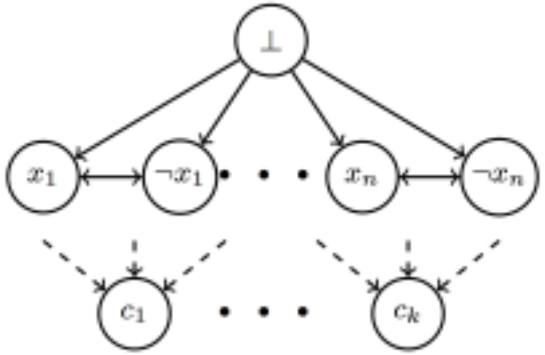
Internal Vertex Labeling and Transmission Tree:
A host labeling of a timed phylogeny T is a function $\ell : L(T) \longrightarrow \Sigma$, assigning a host $\ell(u)$ to each vertex u of T such that the resulting transmission network S is a spanning tree of the contact map C.

[Sashittal and El-Kebir, Bioinformatics/ISMB 2020]

Complexity

$$\phi = \bigwedge_{i=1}^{k}(y_{i,1} \vee y_{i,2} \vee y_{i,3})$$





Contact Map

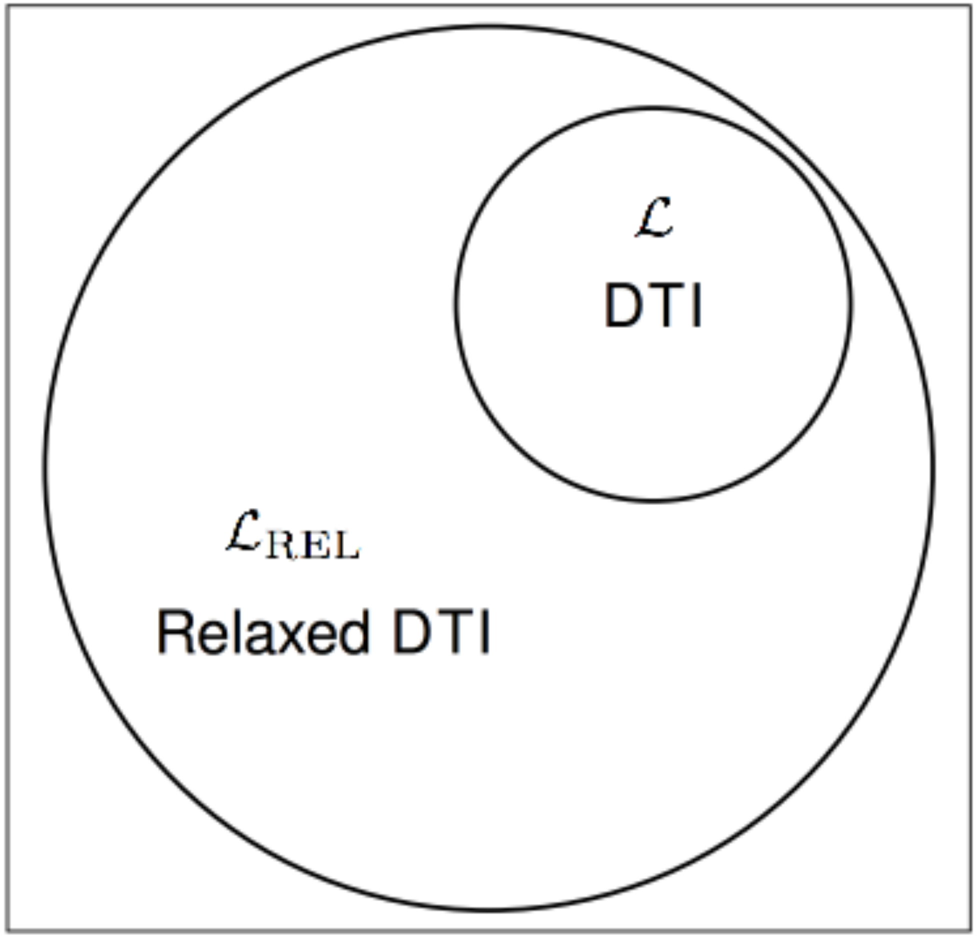
Timed Phylogeny and epidemiological data

We show that Transmission Tree Inference Problem is NP-complete and the corresponding counting problem is #P-complete by reduction from the 1-in-3SAT problem

[Sashittal and El-Kebir, Bioinformatics/ISMB 2020]

Sampling DTI Solutions

Naïve Rejection Sampling



SAT based Almost Uniform Sampling (UniGen)

Vertex Labeling

$$\mathrm{onehot}(\{x_{i,1}, \cdots, x_{i,m}\}), \quad \forall v_i \in V(T).$$

Transmission Edges

$$(x_{i,s} \wedge x_{j,t}) \implies c_{s,t}, \quad \forall (v_i, v_j) \in E(T) \text{ and } s, t \in \Sigma.$$

Root Host Constraint

$$x_{i,t} \implies \neg c_{s,t}, \quad \forall s, t \in \Sigma, s \neq t,$$

Unique Infector Constraint

$$\neg c_{s,t} \vee \neg c_{s,t'}, \quad t, t' \in \Sigma \text{ and } t \neq t'$$

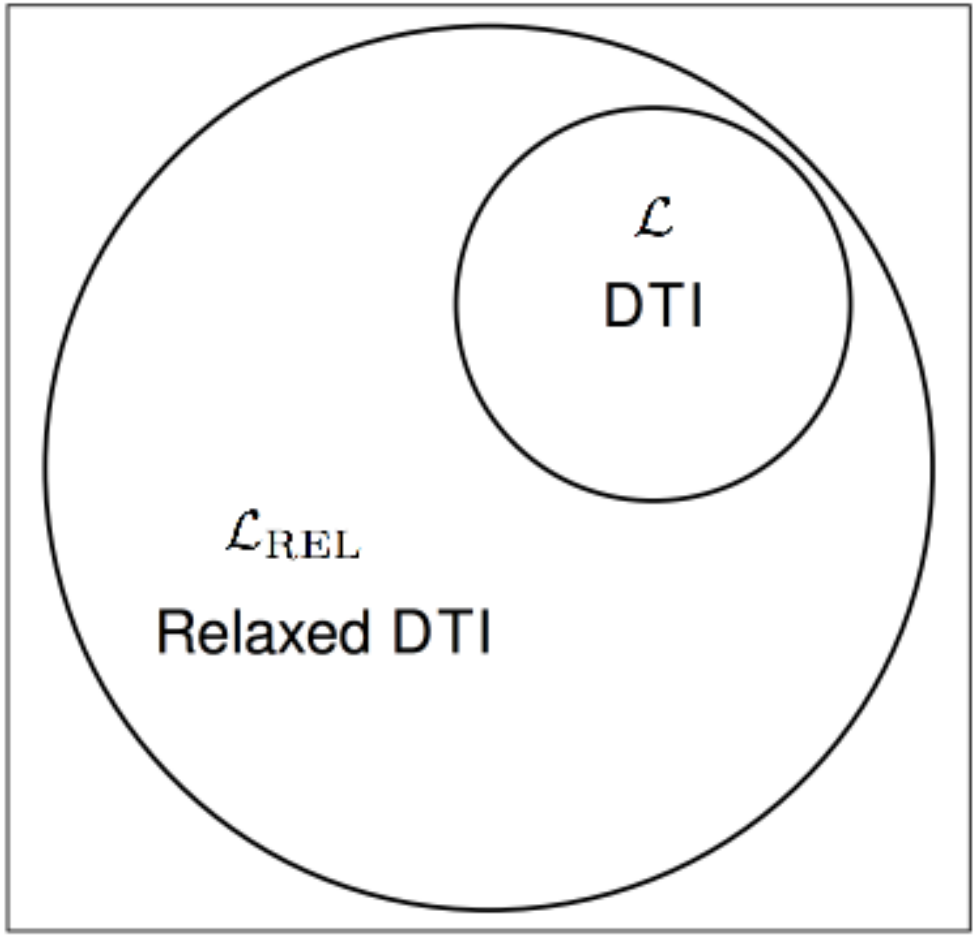
$$\neg x_{i,s} \vee \neg x_{j,t} \vee \neg x_{k,s} \vee \neg x_{l,t}, \quad \forall s, t \in \Sigma, s \neq t.$$

[Sashittal and El-Kebir, Bioinformatics/ISMB 2020]

Sampling DTI Solutions

Not Efficient

Naïve Rejection Sampling



[Sashittal and El-Kebir, Bioinformatics/ISMB 2020]

Efficient and Accurate

$O(nm + m^2)$ *variables and* $O(nm^2 + n^2m^2)$ *constraints*

Vertex Labeling

$$\text{onehot}(\{x_{i,1}, \cdots, x_{i,m}\}), \quad \forall v_i \in V(T).$$

Transmission Edges

$$(x_{i,s} \wedge x_{j,t}) \implies c_{s,t}, \quad \forall (v_i, v_j) \in E(T) \text{ and } s, t \in \Sigma.$$

Root Host Constraint

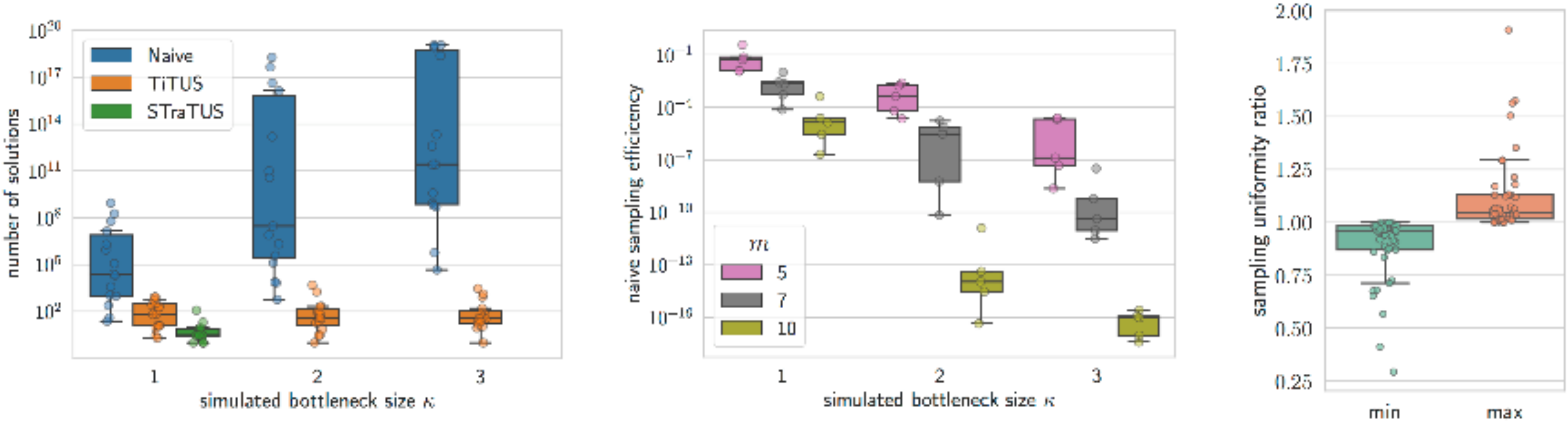
$$x_{i,t} \implies \neg c_{s,t}, \quad \forall s, t \in \Sigma, s \neq t,$$

Unique Infector Constraint

$$\neg c_{s,t} \vee \neg c_{s,t'}, \quad t, t' \in \Sigma \text{ and } t \neq t'$$

$$\neg x_{i,s} \vee \neg x_{j,t} \vee \neg x_{k,s} \vee \neg x_{l,t}, \quad \forall s, t \in \Sigma, s \neq t.$$

Simulation Results



Simulations (with complete sampling) show that:

(a) Weak Transmission Bottleneck needs to be considered for inferring and sampling the solutions.

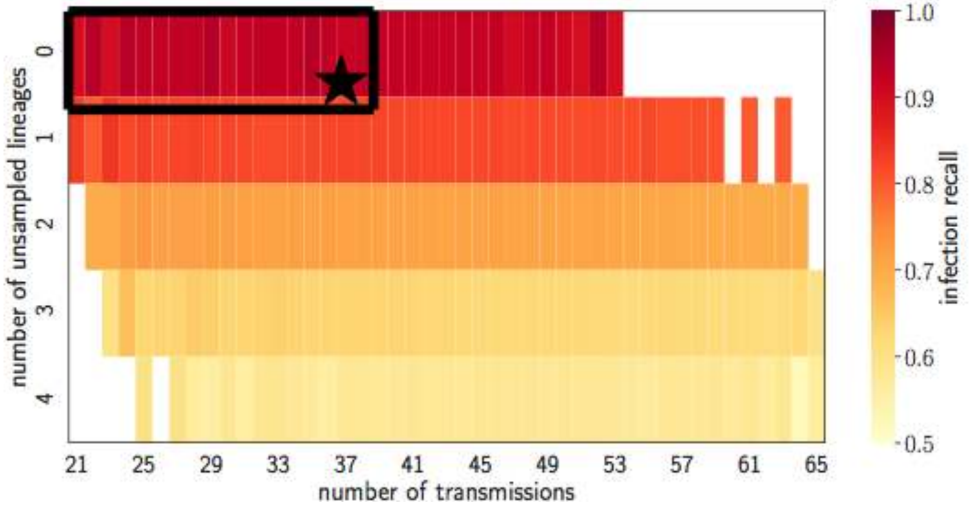
(b) Naïve sampling is infeasible for large outbreaks

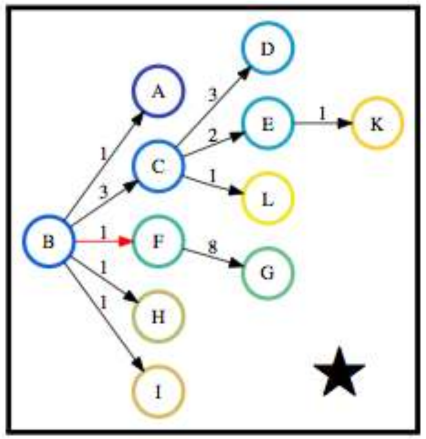
(c) TiTUS uniformly samples the solution space

[Sashittal and El-Kebir, Bioinformatics/ISMB 2020]

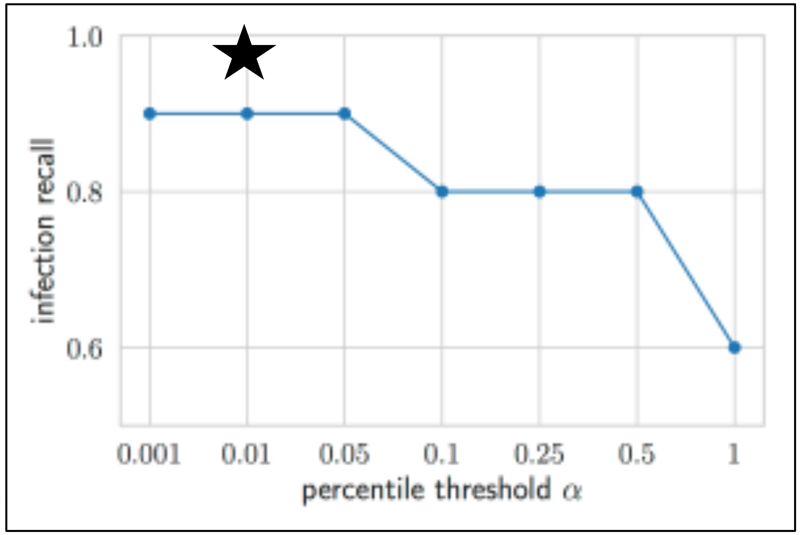
HIV Outbreak in 1988-2006 among 11 patients

α = 0.01





Consensus Transmission Tree



TiTUS reconstruct the transmission history of a HIV outbreak:

(a) We generate 100,000 samples from the solution space and build a consensus of the selected solutions

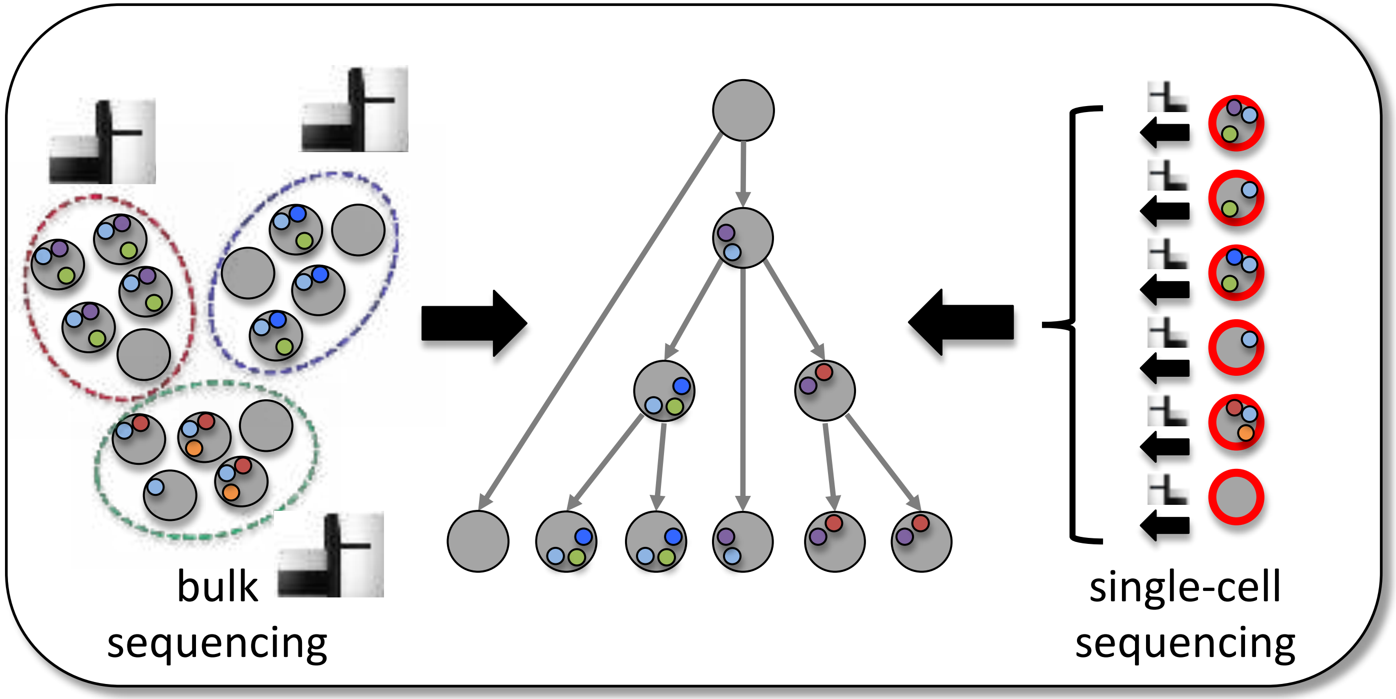
(b) Consensus transmission tree recovers 9/10 transmission pairs in the outbreak

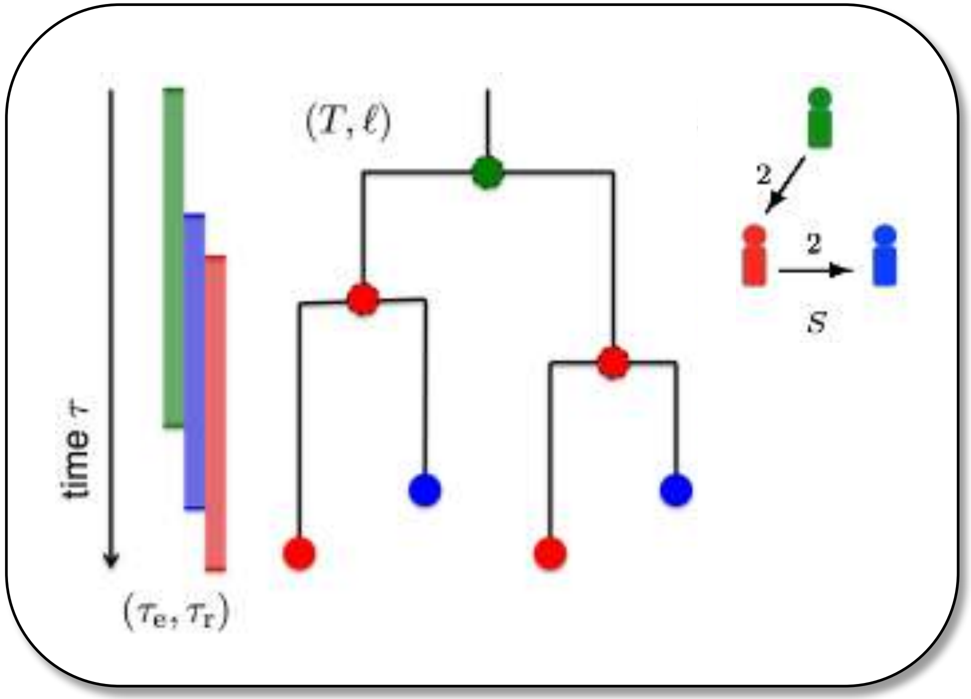
(c) Our method is robust for the choice of percentile threshold

[Sashittal and El-Kebir, Bioinformatics/ISMB 2020]

Conclusions and Future Directions

Solving problems in computational biology via approximate model counting





Cutting planes & column generation

Weighted model counting

Guidance/best practices on efficient SAT formulations

Acknowledgements

This work was funded by the National Science Foundation (CCF-1850502 and CCF-2027669)

- Kuldeep Meel
- Mate Soos

El-Kebir group

- **Jackie Oh**
- **Palash Sashittal**
- **Yuanyuan Qi**
- Chuanyi Zhang
- Jiaqi Wu
- Juho Kim
- Leah Weber
- Nuraini Aguse
- Sarah Christensen

BACKUP

Problem Statement

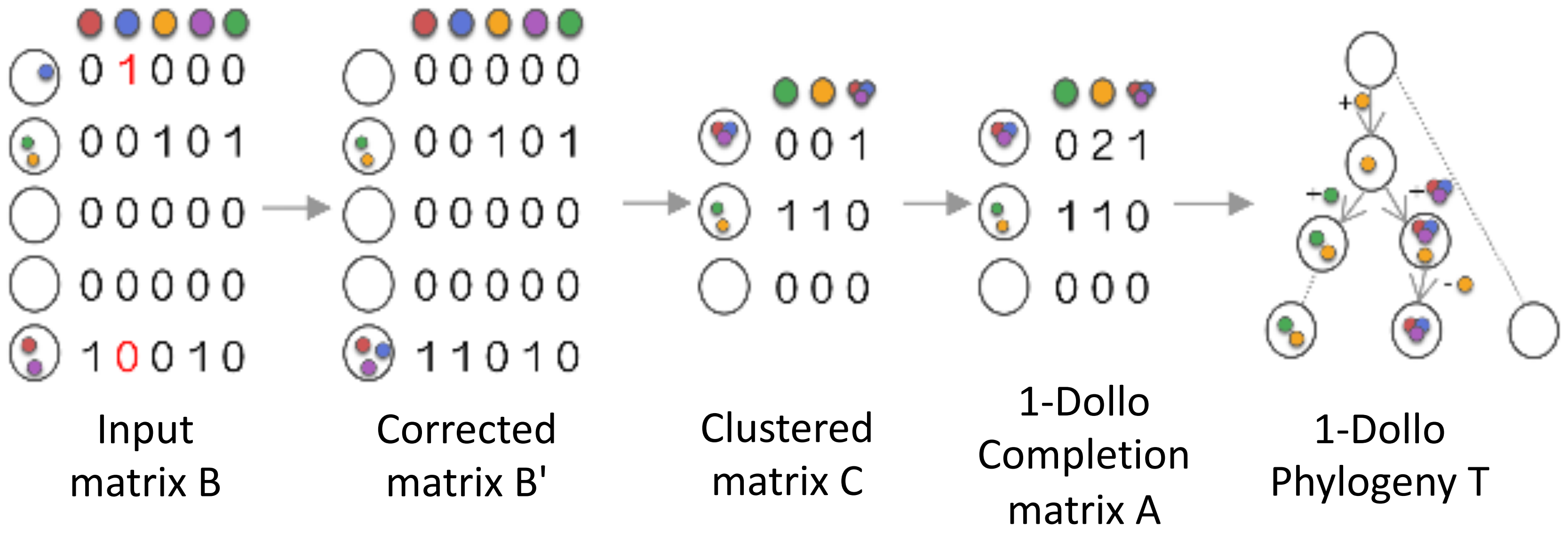
Inputs:

- Binary matrix $B \in \{0,1\}^{m \times n}$ where entry $b_{i,j} = 1$ if and only if cell i contains mutation j
- A set L of mutations that can be lost
- Number of mutation clusters s
- Number of cell clusters t
- False positive rate α, false negative rate β

Desired output:

A rooted tree T that meets the following conditions:

- Each vertex is labeled by a vector $v \in \{0,1\}^t$
- The root of T is labeled by the zero vector
- Each mutation in $[n]$ labels exactly one gain edge
- Each mutation in L labels at most one loss edge
- Each leaf of T is labeled by a row of matrix C $\in \{0,1\}^{s \times t}$
  - C is the result of correcting errors in B and clustering so that there are s distinct rows and t distinct columns



$$\begin{matrix}0&1&0&0&0\\0&0&1&0&1\\0&0&0&0&0\\0&0&0&0&0\\1&0&0&1&0\end{matrix} \rightarrow \begin{matrix}0&0&0&0&0\\0&0&1&0&1\\0&0&0&0&0\\0&0&0&0&0\\1&1&0&1&0\end{matrix} \rightarrow \begin{matrix}0&0&1\\1&1&0\\0&0&0\end{matrix} \rightarrow \begin{matrix}0&2&1\\1&1&0\\0&0&0\end{matrix} \rightarrow$$

A matrix is a 1-Dollo Completion **<u>if and only if</u>** it does not contain any forbidden submatrices

There are 25 forbidden submatrices [El Kebir et al.]

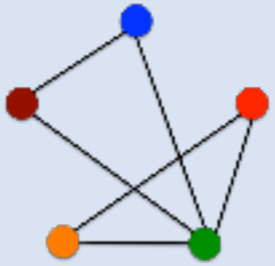
Background

Accurate inference of **transmission networks** if pivotal for

- real-time outbreak management,
- public health policies.

Traditional epidemiological approaches involve:

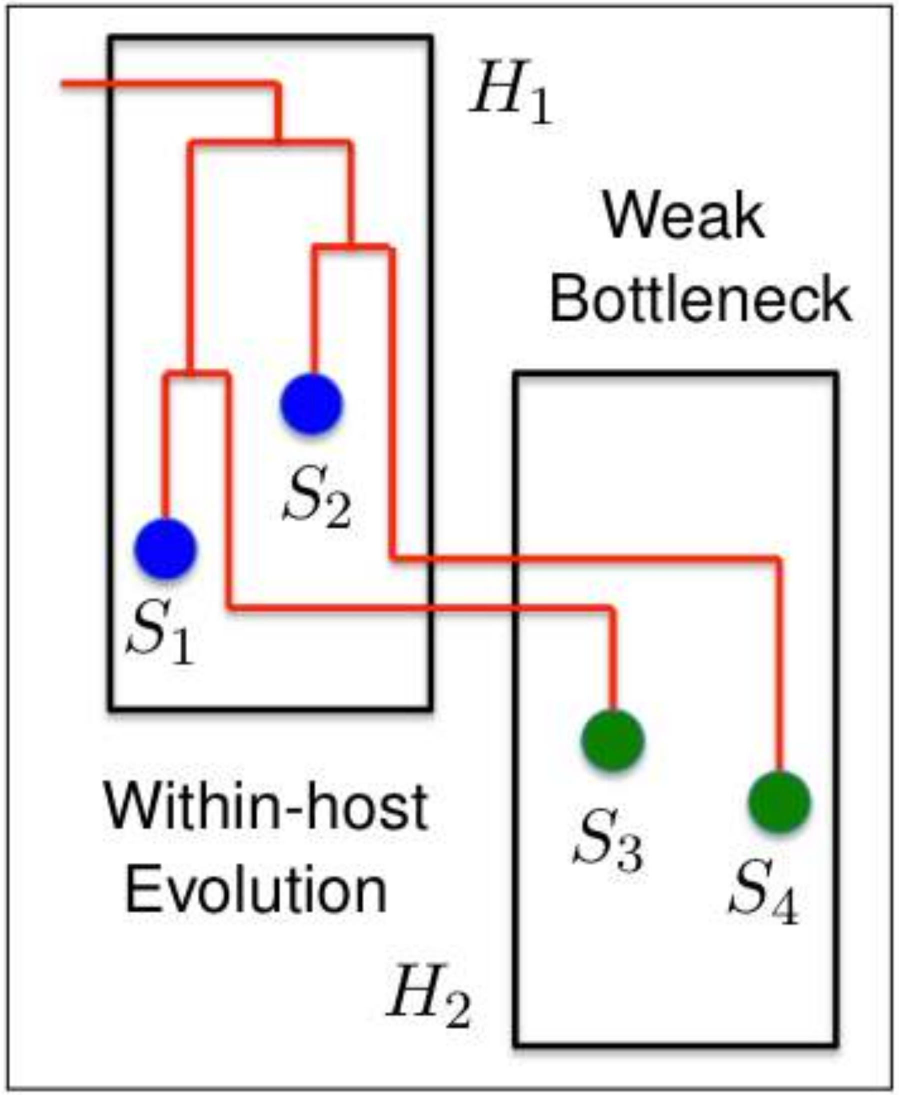
- fieldwork and interviews,
- contact tracing.



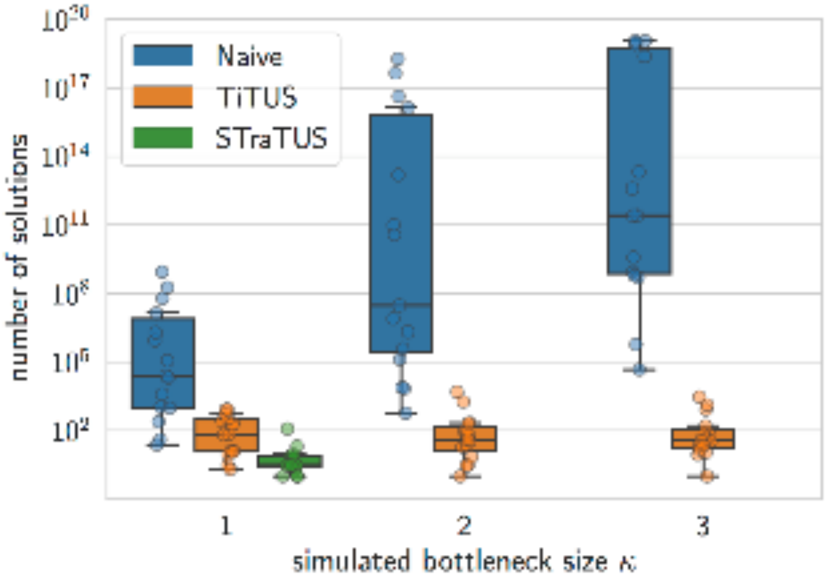
With decreasing costs of genomic sequencing, **molecular epidemiology** has become indispensable. (e.g. ~2500 SARS-CoV-2 sequences on GISAID.)

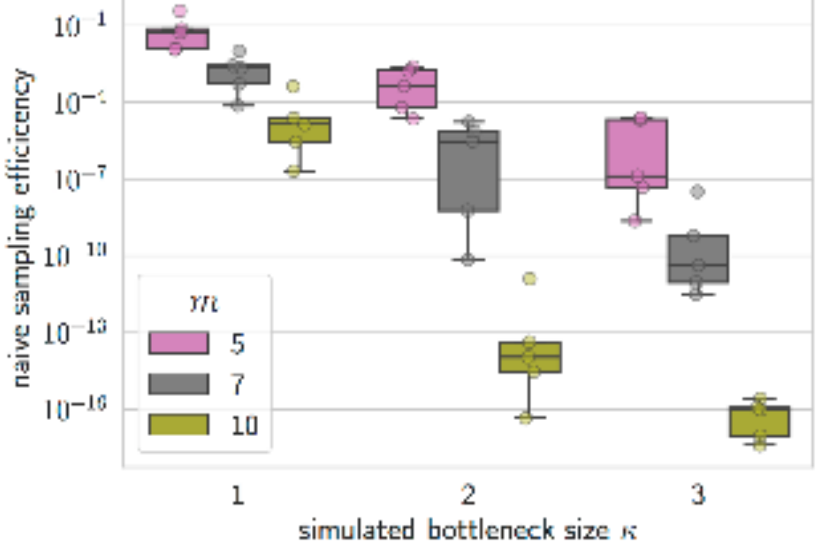
Challenges in Transmission Network Inference

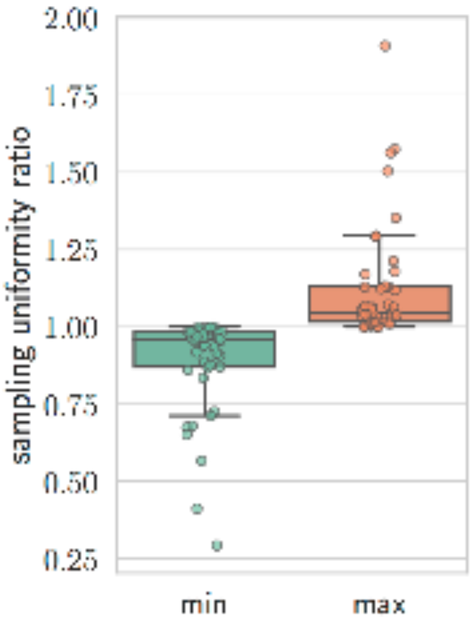
- **Incomplete lineage sorting**: pathogen evolutionary history does not match the transmission history of the outbreak.
- High mutation rates and/or long incubation times result in **within-host diversity**.
- Further complication arises due to multi-strain infection or **weak transmission bottleneck**.



Simulation Results







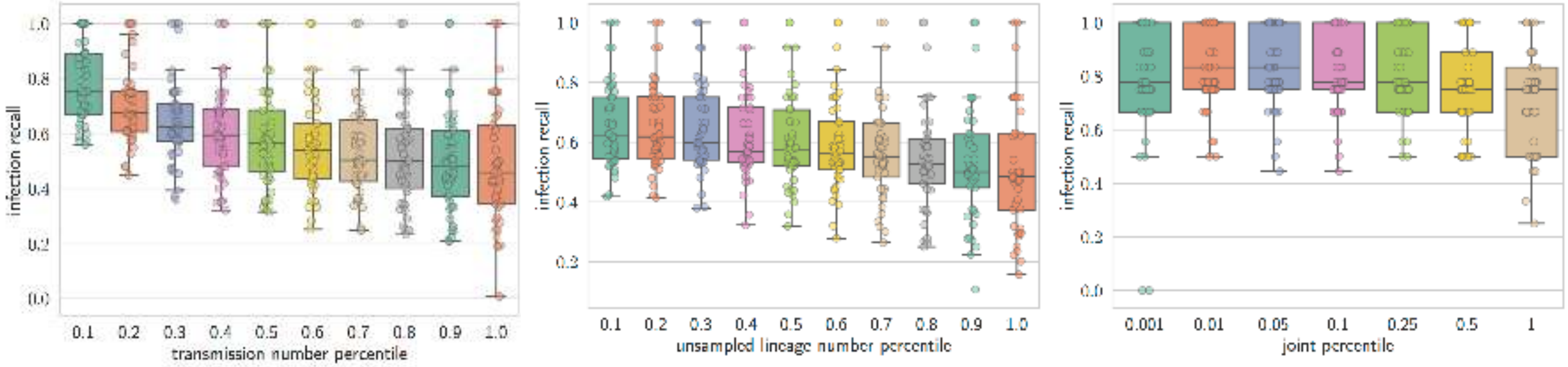
Simulations (with complete sampling) show that:

(a) Weak Transmission Bottleneck needs to be considered for inferring and sampling the solutions.

(b) Naïve sampling is infeasible for large outbreaks

(c) TiTUS uniformly samples the solution space

Selection Criteria

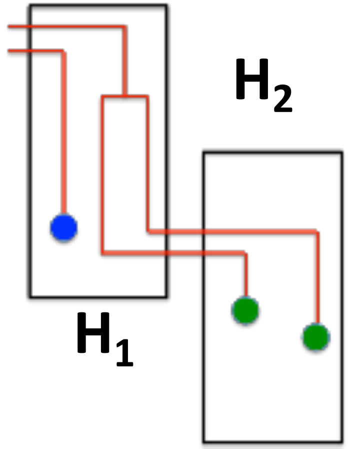


Following selection criteria are proposed (for a completely sampled outbreak):

(a) Number of transmitted strains in the outbreak

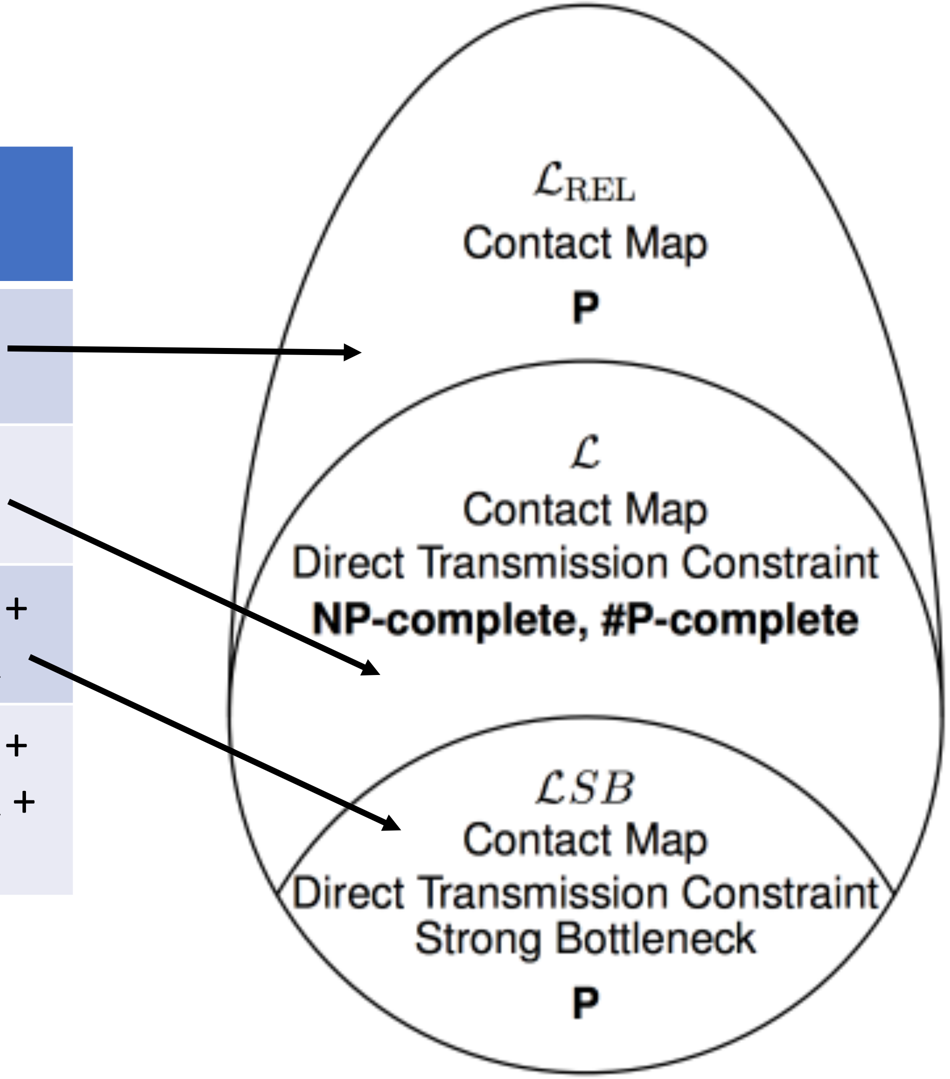
(b) Number of unsampled lineages in the outbreak

(c) We find that optimal performance is achieved at percentile threshold of 0.01



TiTUS vs. Previous Work

Method	Constraint
Simple Recursion	Contact Map
TiTUS	Contact Map + Unique Infector
STraTUS[2]	Contact Map + Unique Infector + Strong Transmission Bottleneck
Kenah[3]	Contact Map + Unique Infector + Strong Transmission Bottleneck + Order of Infection



$\mathcal{L}_{\text{REL}}$
Contact Map
P

$\mathcal{L}$
Contact Map
Direct Transmission Constraint
NP-complete, #P-complete

$\mathcal{L}SB$
Contact Map
Direct Transmission Constraint
Strong Bottleneck
P

[2] Matthew D Hall and Caroline Colijn. Molecular biology and Evolution (2019).
[3] Eben Kenah *et al.* PLoS Computational Biology (2016).