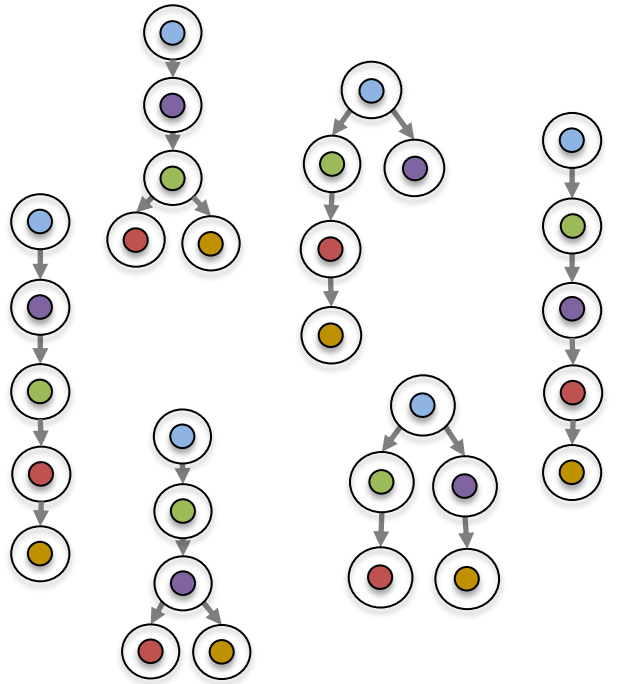
Summarizing the Solution Space in Tumor Phylogeny Inference by Multiple Consensus Trees

Nuraini Aguse*, Yuanyuan Qi* and Mohammed El-Kebir
University of Illinois at Urbana Champaign, Department of Computer Science

*Joint first authorship

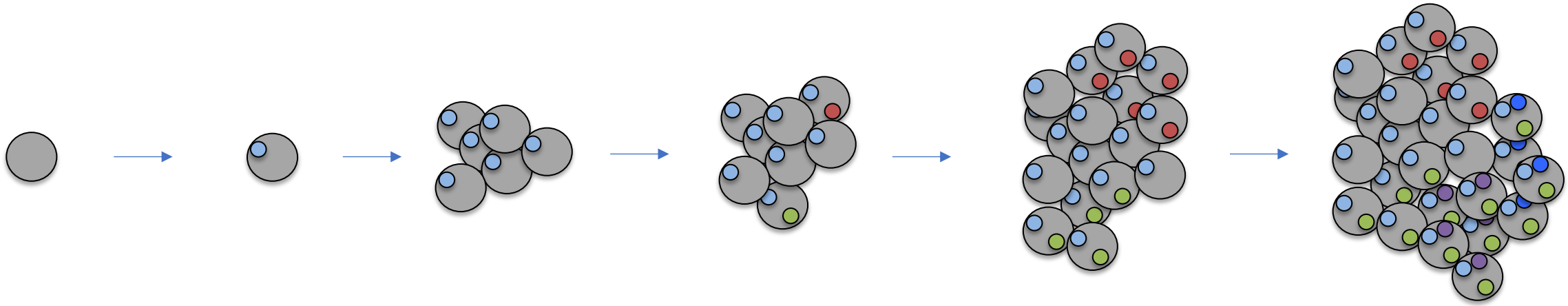
Outline

- Introduction to tumor phylogenies
- Previous work
- Multiple Consensus Trees problem
- Methods
  - MILP
  - Heuristic
  - Bayesian Information Criterion



Introduction to tumor phylogenies

Cancer results from an evolutionary process



Normal cell

Cell acquires cancerous mutation

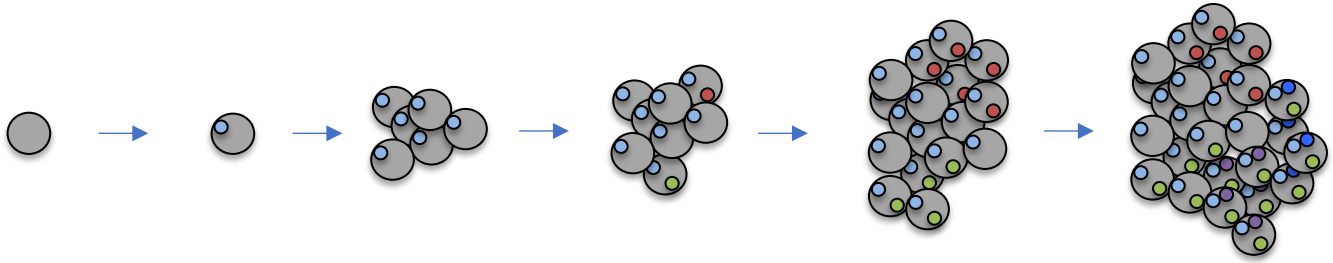
Cell with mutation replicate

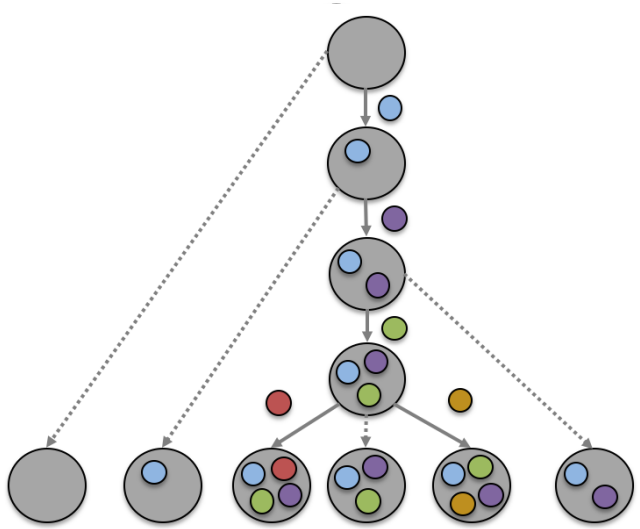
Some cells acquire more mutations

Each cell divide and duplicates

Heterogeneous tumor

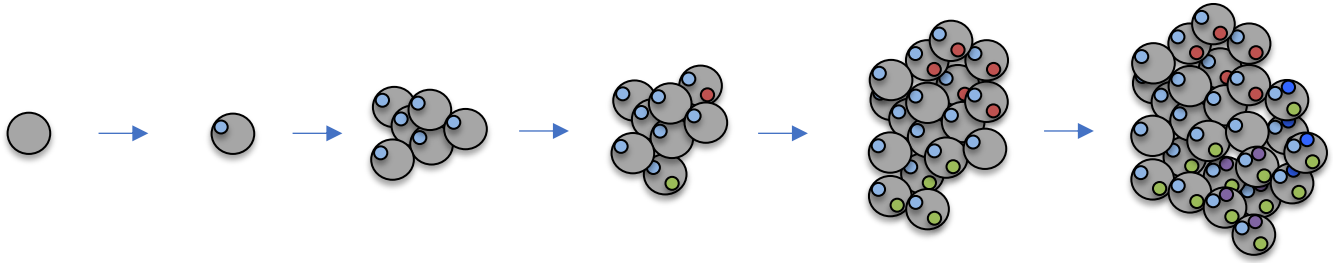
Tumor progression can be represented as a phylogenetic tree



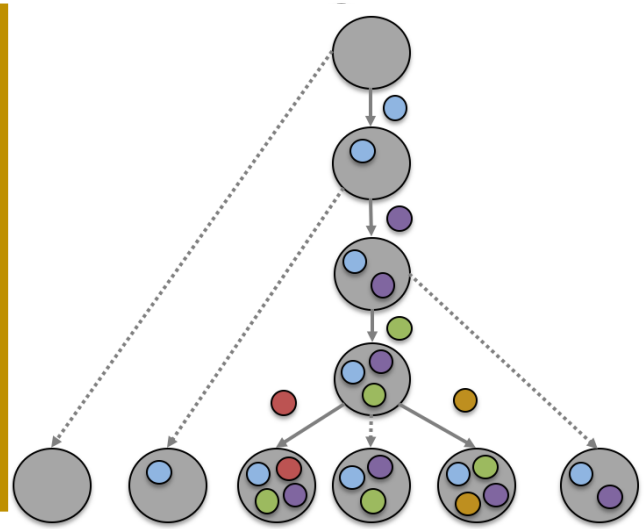


Phylogenetic tree

Tumor progression can be represented as a phylogenetic tree

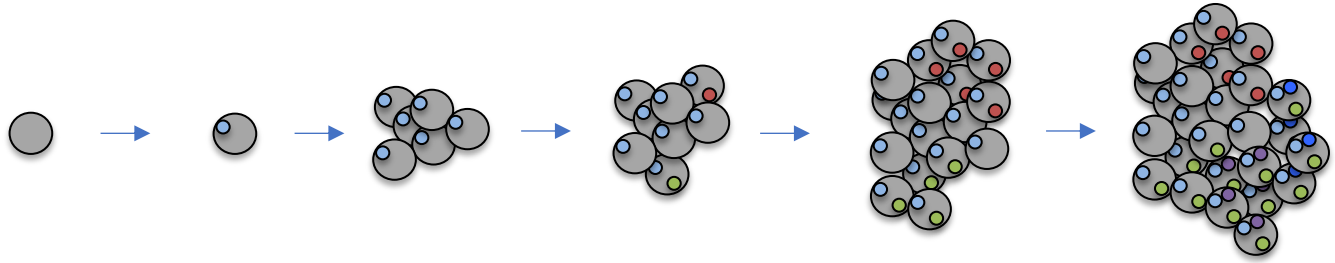


Infinite Sites Assumption (ISA) states that a mutation is gained only once and never subsequently lost

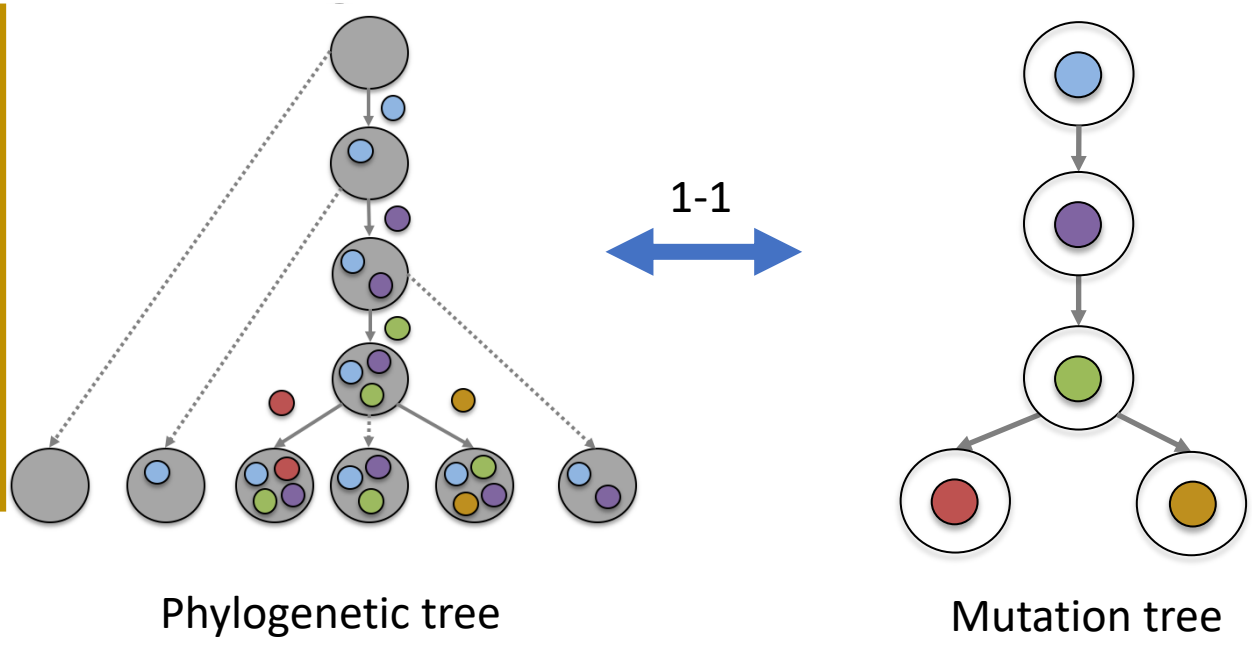


Phylogenetic tree

Tumor progression can be represented as a phylogenetic tree (or mutation tree)



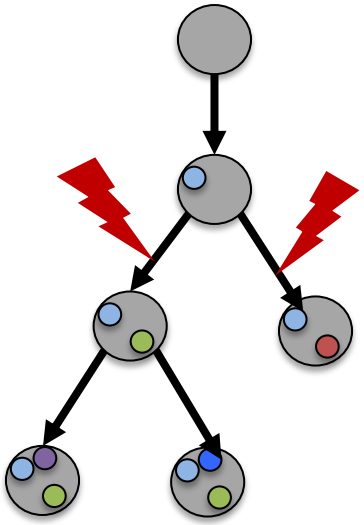
Infinite Sites Assumption (ISA) states that a mutation is gained only once and never subsequently lost



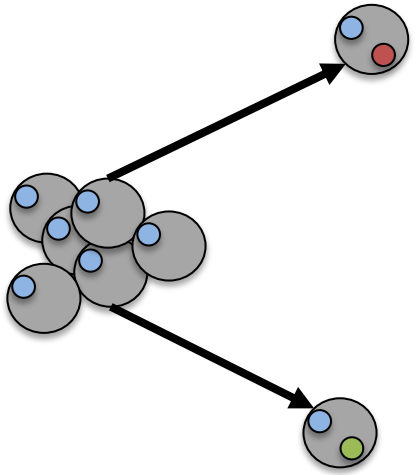
Under ISA, a phylogenetic tree can be represented as a mutation tree

Phylogenies enable us to understand and treat cancer

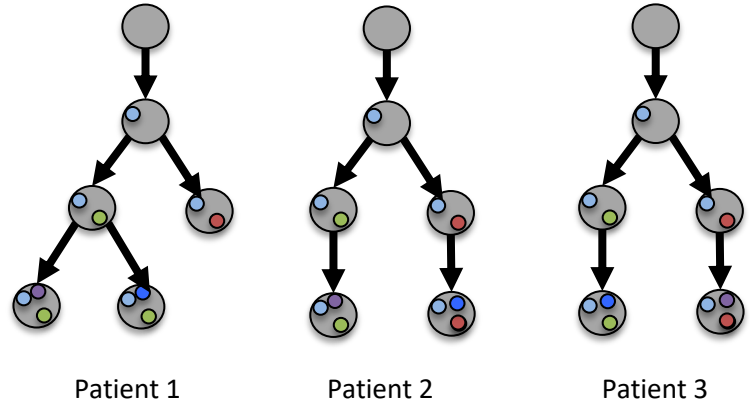
Identify targets for treatment



Understand metastatic development



Recognize common patterns of tumor evolution across patients

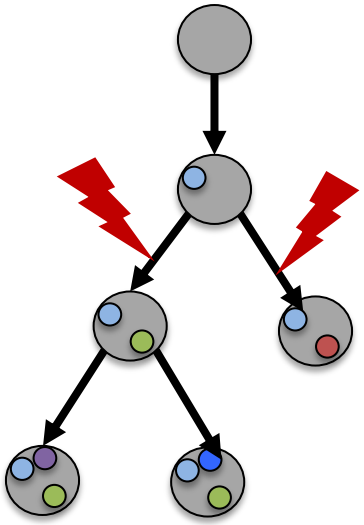


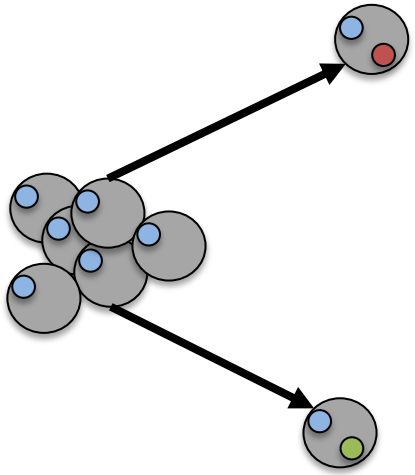
Phylogenies enable us to understand and treat cancer

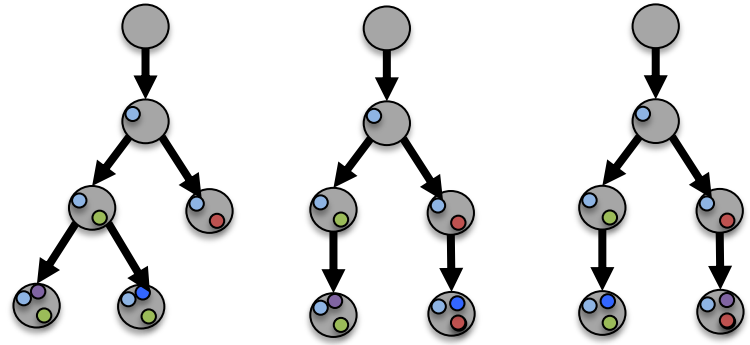
Identify targets for treatment

Understand metastatic development

Recognize common patterns of tumor evolution across patients

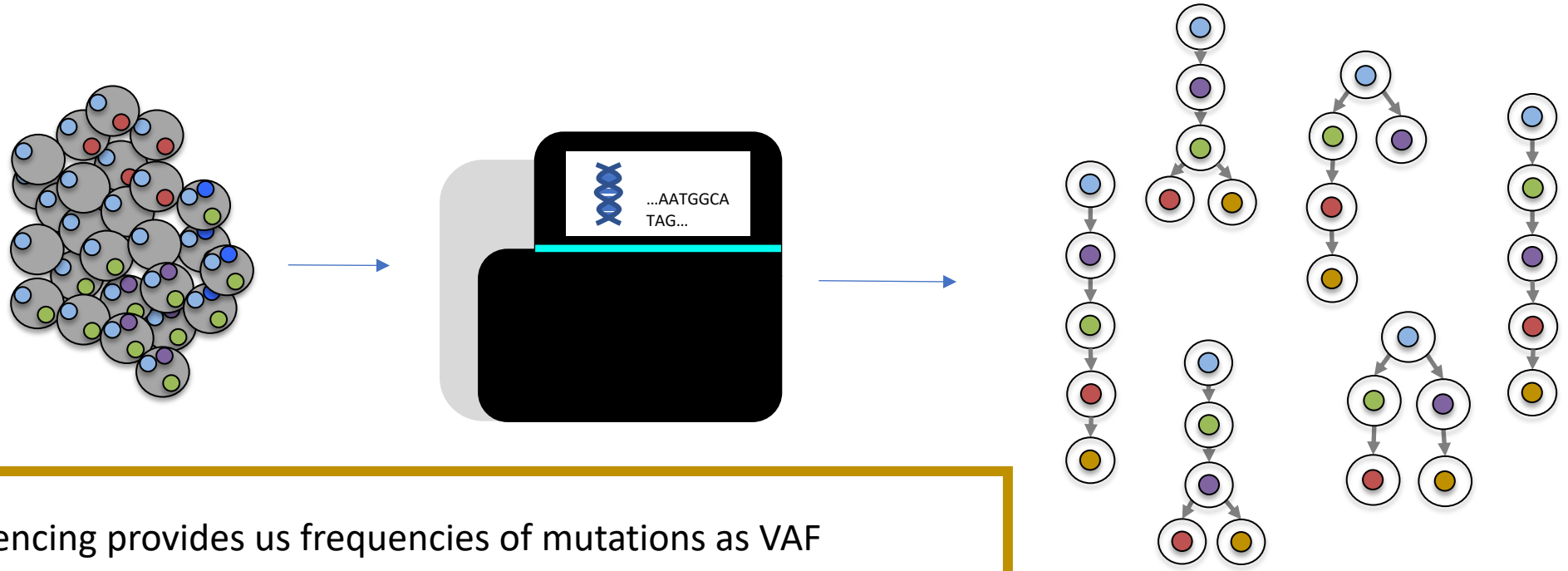






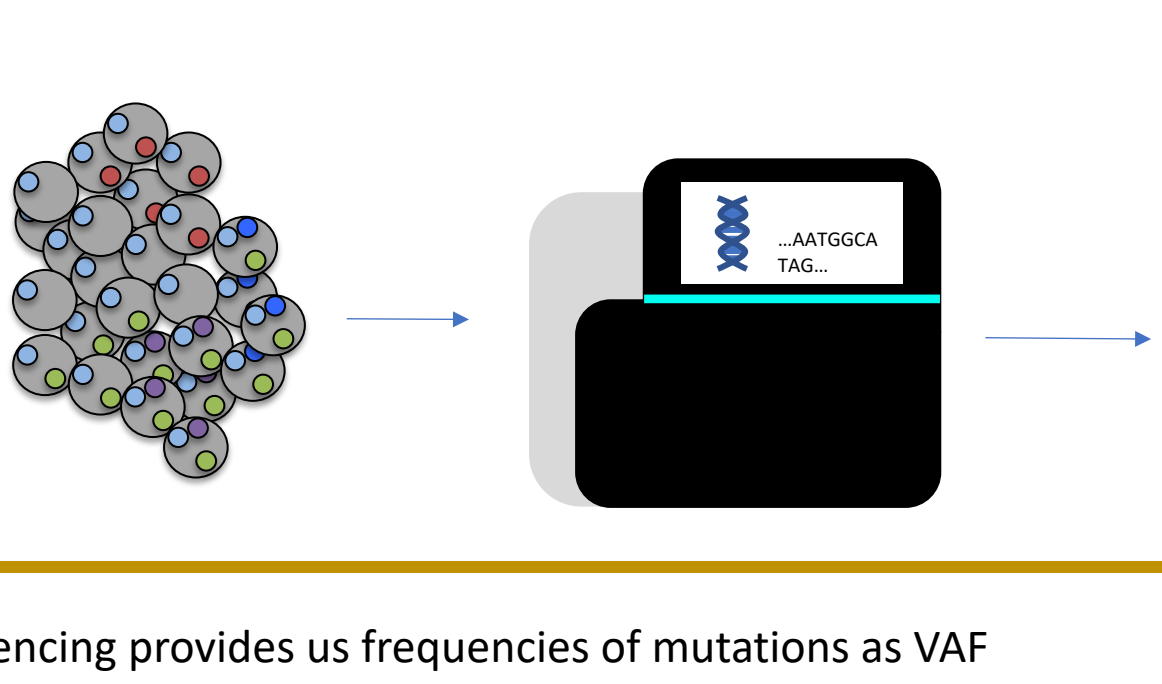
How do we obtain the phylogenetic tree of a tumor?

Current tumor phylogeny inference results are not unique



- Bulk sequencing provides us frequencies of mutations as VAF
- Single-cell sequencing has a high false positive and false negative rate
- Both contribute to non-uniqueness of tumor phylogeny inference

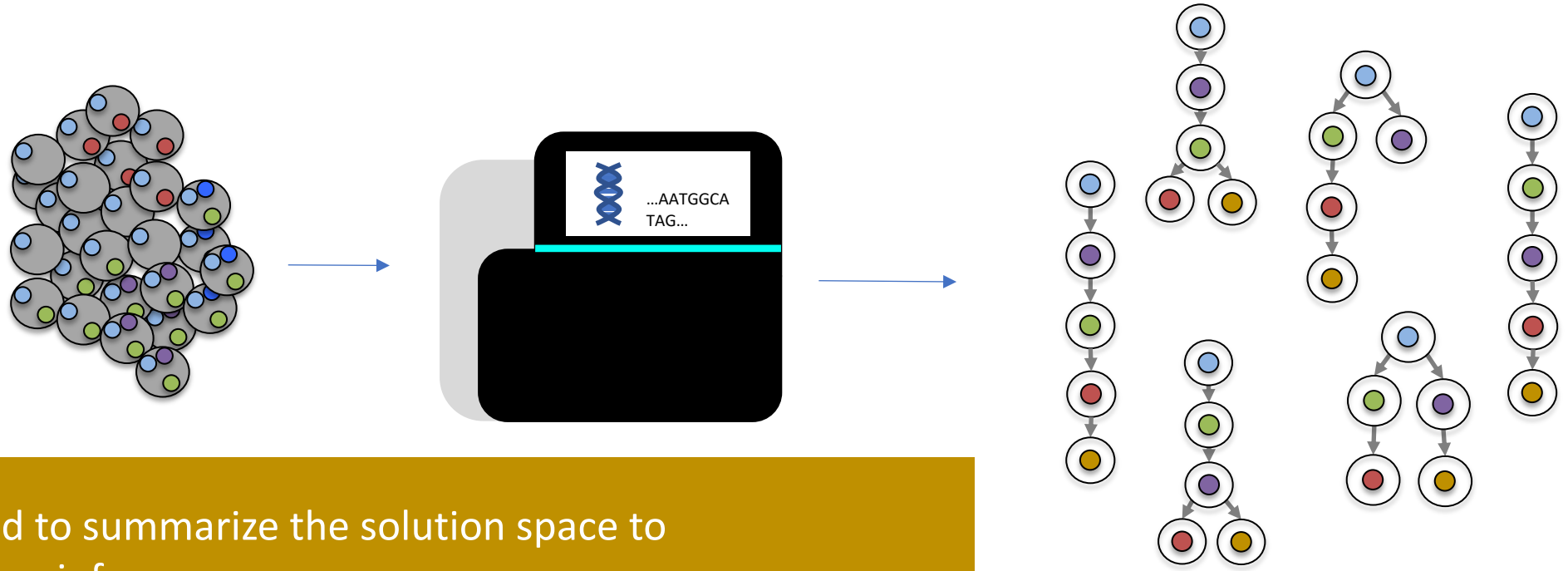
Current tumor phylogeny inference results are not unique



- Bulk sequencing provides us frequencies of mutations as VAF
- Single-cell sequencing has a high false positive and false negative rate
- Both contribute to non-uniqueness of tumor phylogeny inference

Downstream analyses rely on accurate phylogeny inference

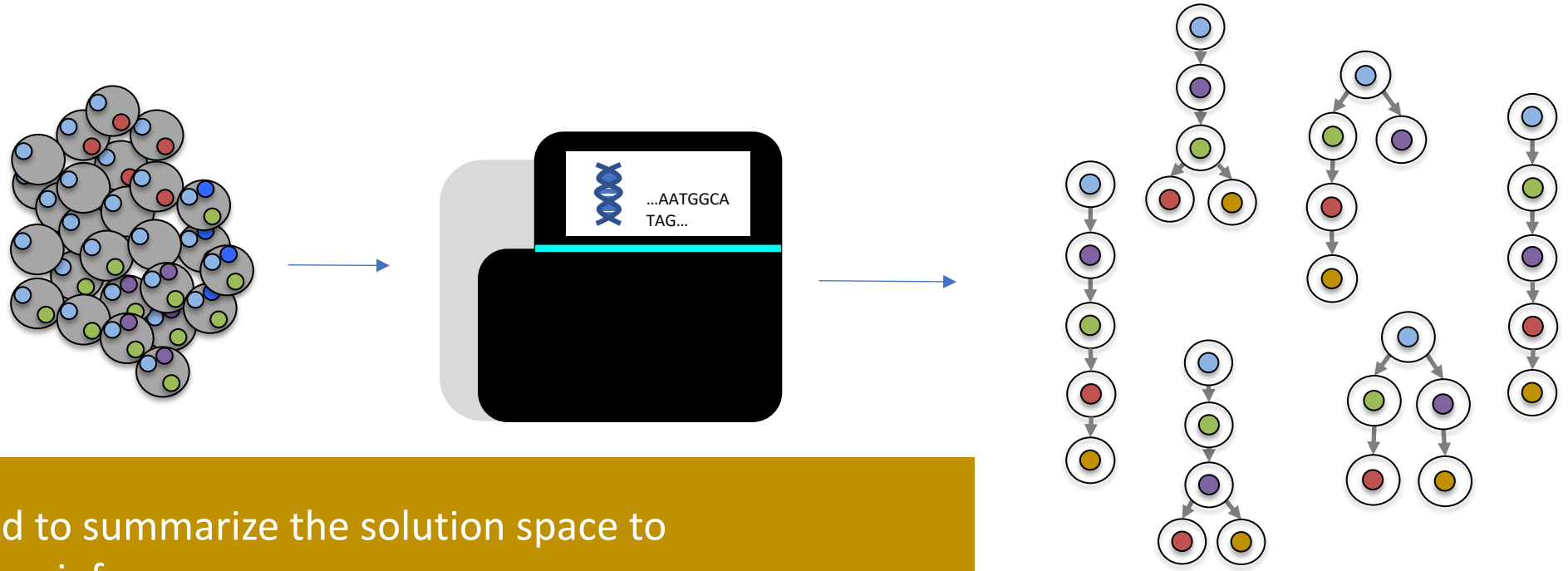
Current tumor phylogeny inference results are not unique



We need to summarize the solution space to

- Reduce inference errors
- Identify dependencies among mutations

Current tumor phylogeny inference results are not unique



We need to summarize the solution space to

- Reduce inference errors
- Identify dependencies among mutations

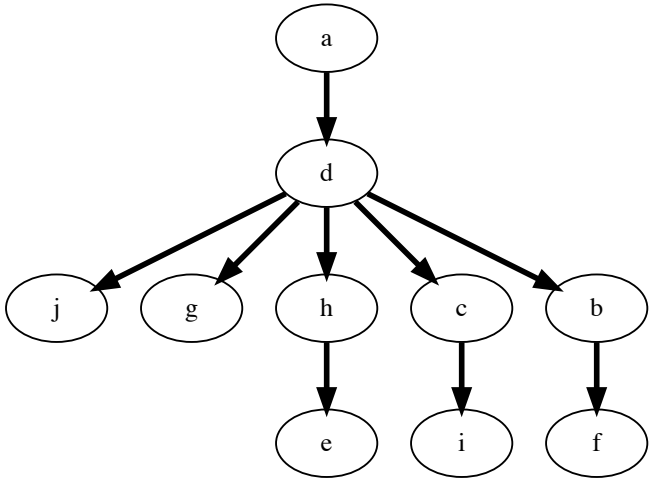
How do we best summarize the solution space?

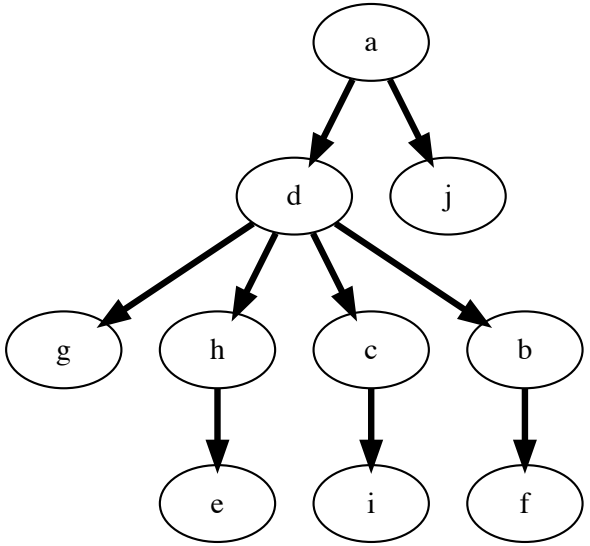
Previous work

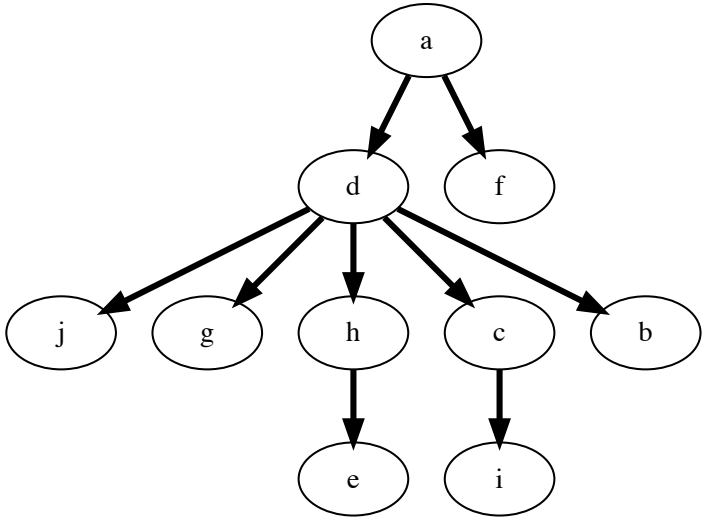
Solution space of lung cancer patient CRUK0037

Jamal-Hanjani et al. (2017). *New England Journal of Medicine*, *376*(22), 2109–2121.

Jamal-Hanjani et al. inferred 17 trees for patient CRUK0037





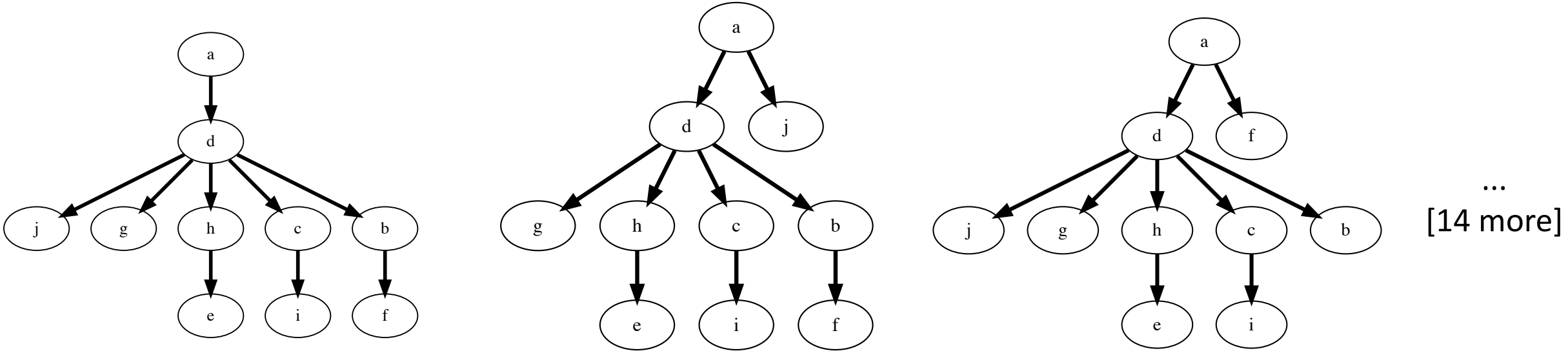


...
[14 more]

Solution space of lung cancer patient CRUK0037

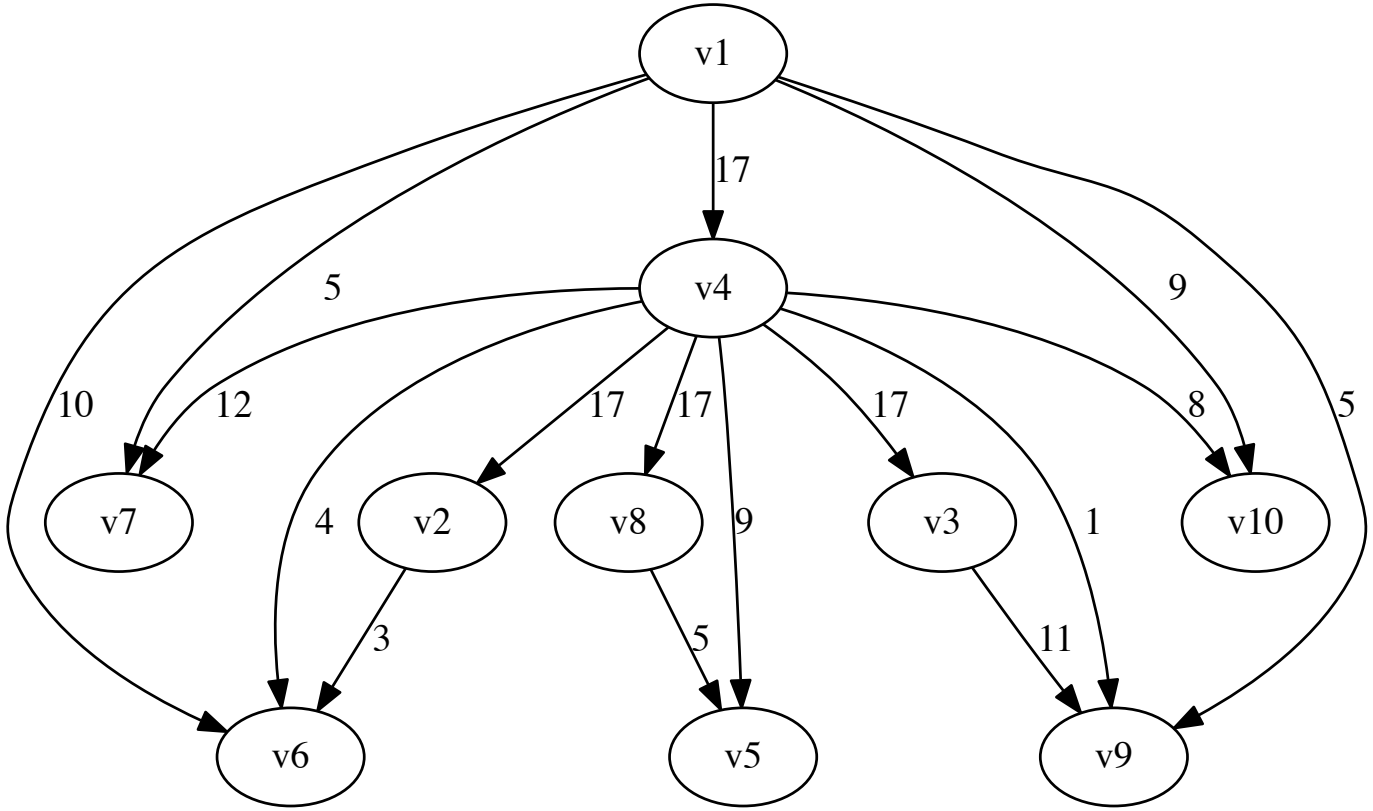
Jamal-Hanjani et al. (2017). *New England Journal of Medicine*, *376*(22), 2109–2121.

Jamal-Hanjani et al. inferred 17 trees for patient CRUK0037

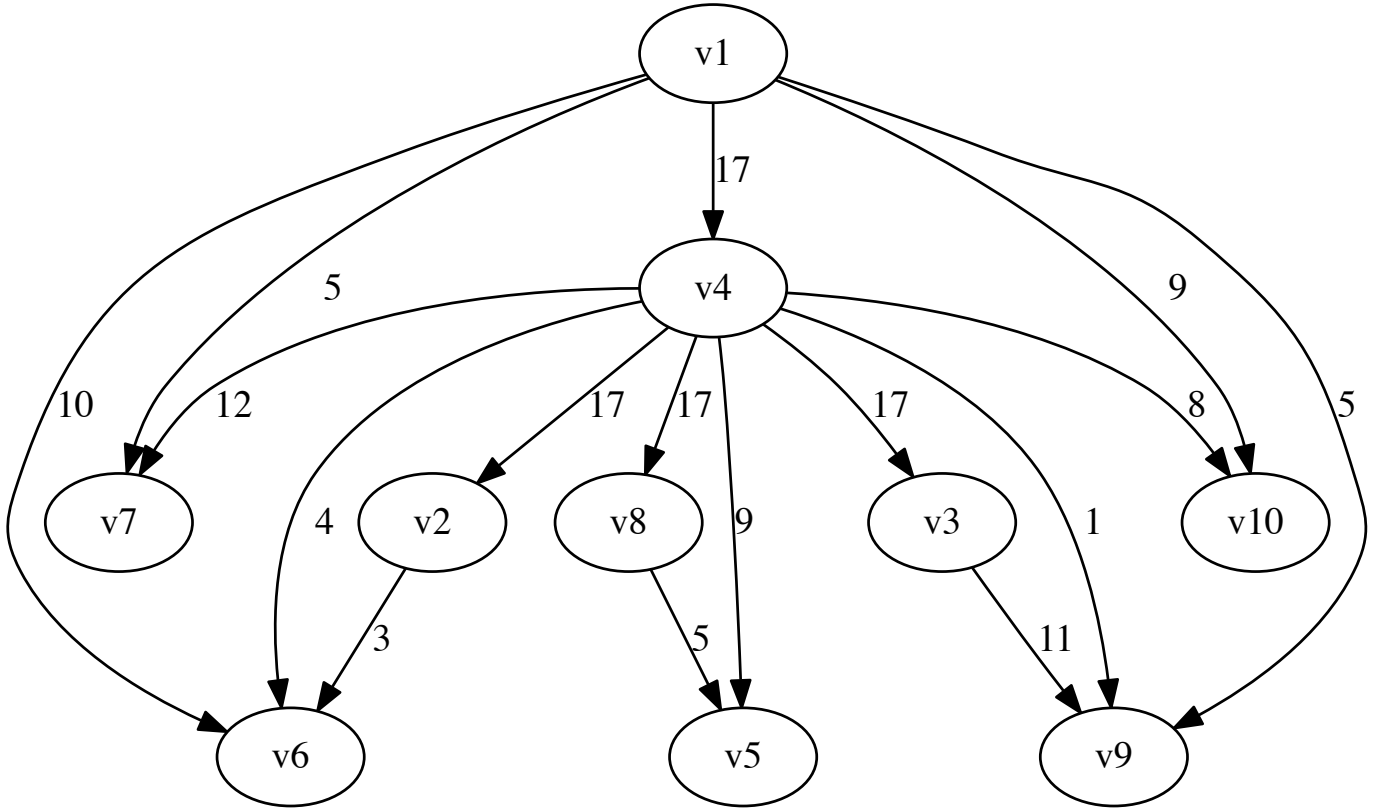


Two summary methods in previous work:
Parent-child graph & single consensus tree

Summarizing the solution space using parent-child graph

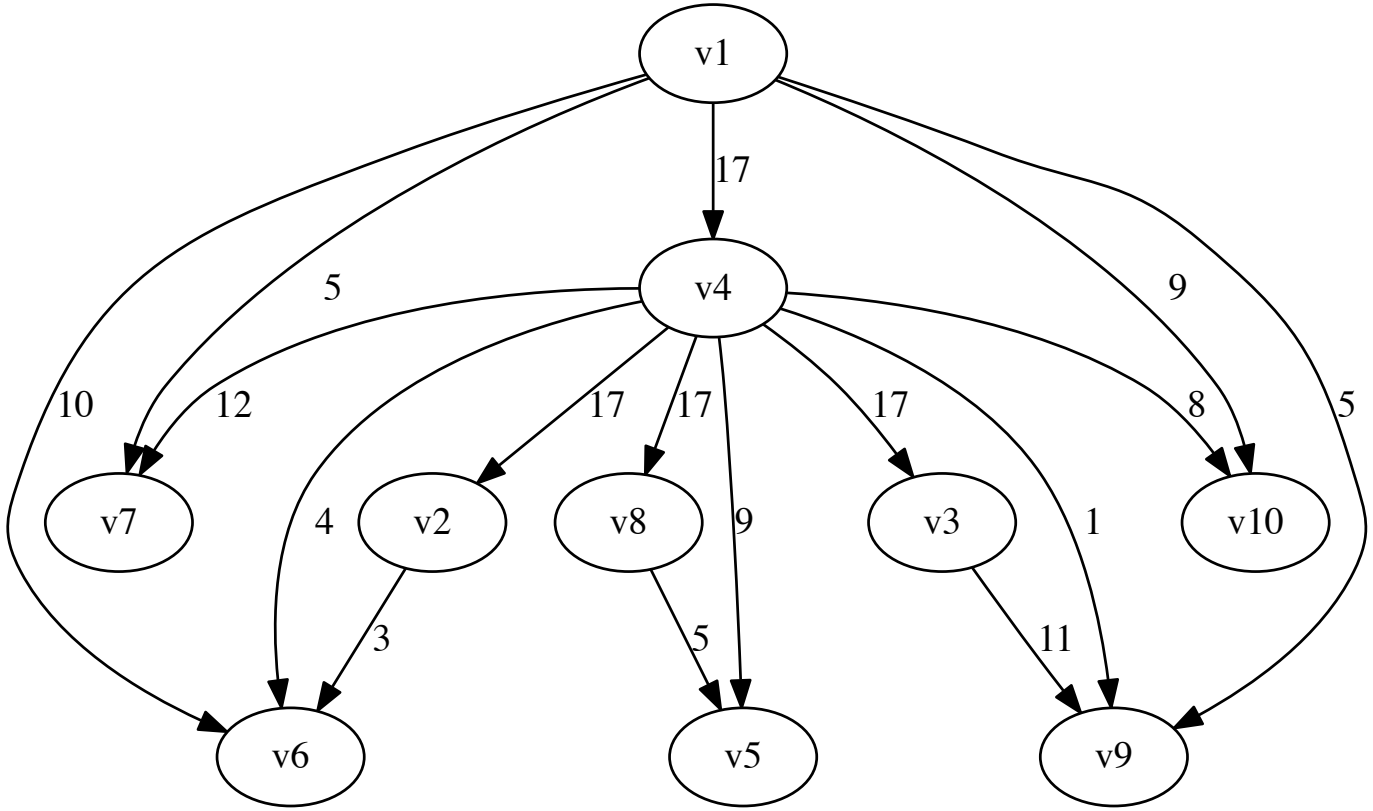


Summarizing the solution space using parent-child graph



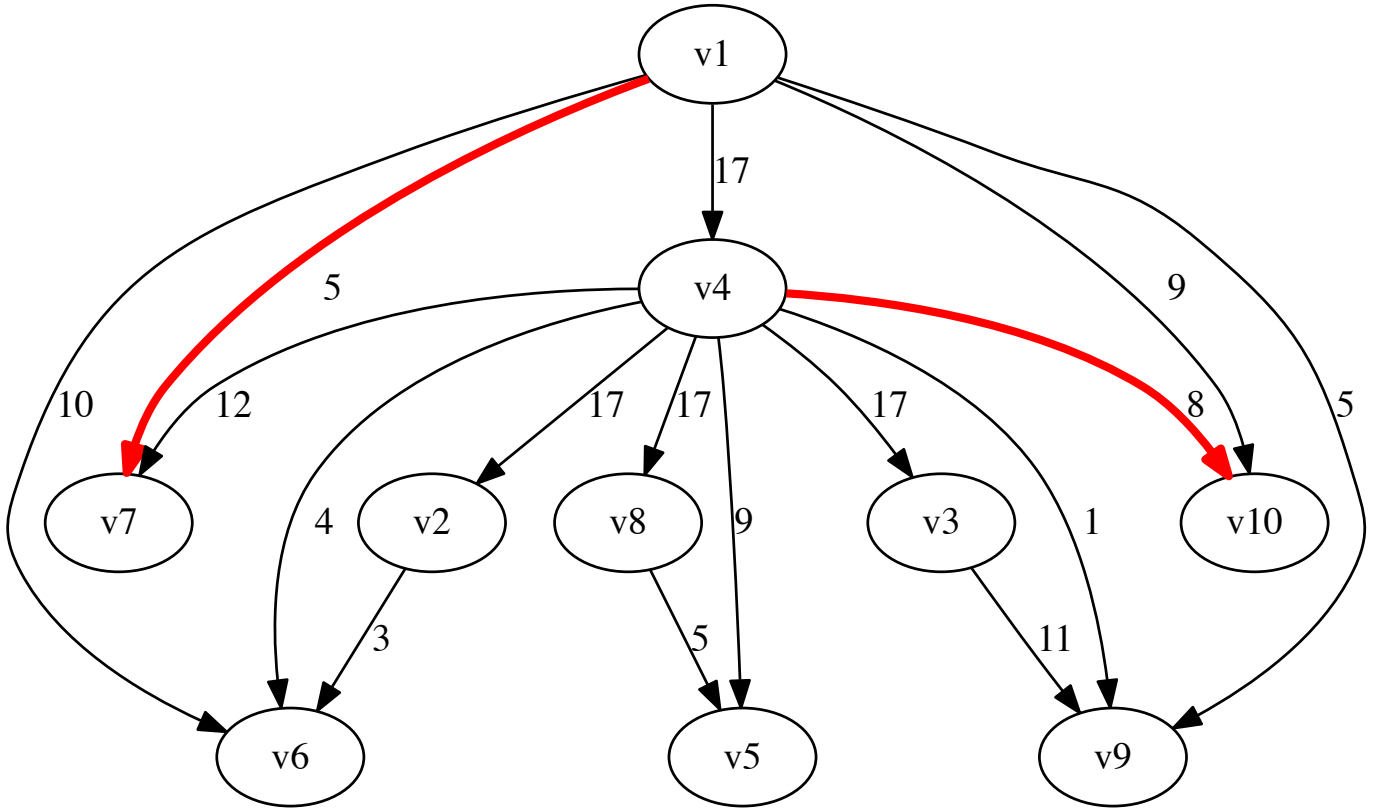
	$v_1 \rightarrow v_{10}$	$v_4 \rightarrow v_{10}$
$v_1 \rightarrow v_7$	2	0
$v_4 \rightarrow v_7$	2	5

Summarizing the solution space using parent-child graph



	$v_1 \rightarrow v_{10}$	$v_4 \rightarrow v_{10}$
$v_1 \rightarrow v_7$	2	0
$v_4 \rightarrow v_7$	2	5

Summarizing the solution space using parent-child graph

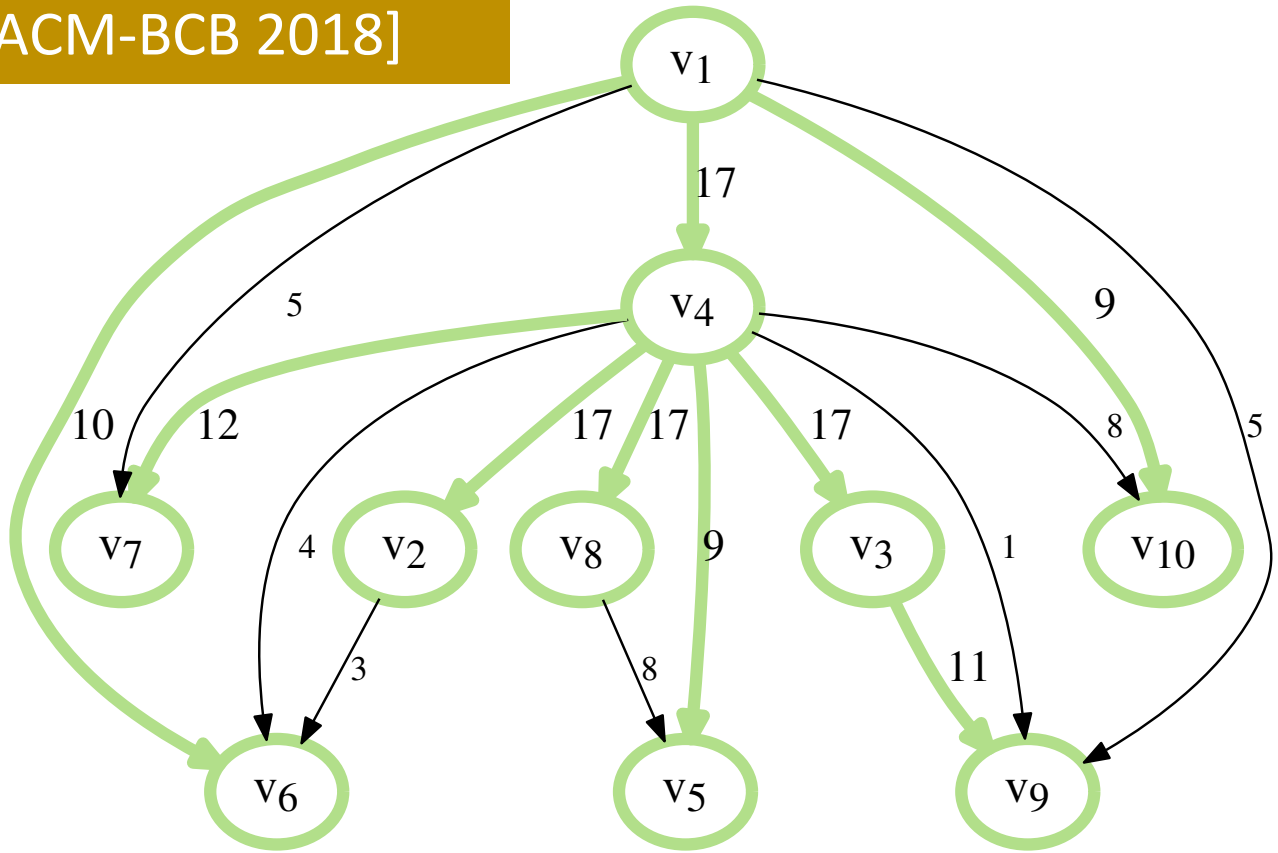


	$v_1 \rightarrow v_{10}$	$v_4 \rightarrow v_{10}$
$v_1 \rightarrow v_7$	2	0
$v_4 \rightarrow v_7$	2	5

Patterns of mutual exclusivity are not captured in parent-child graph

Summarizing the solution space using a single consensus tree

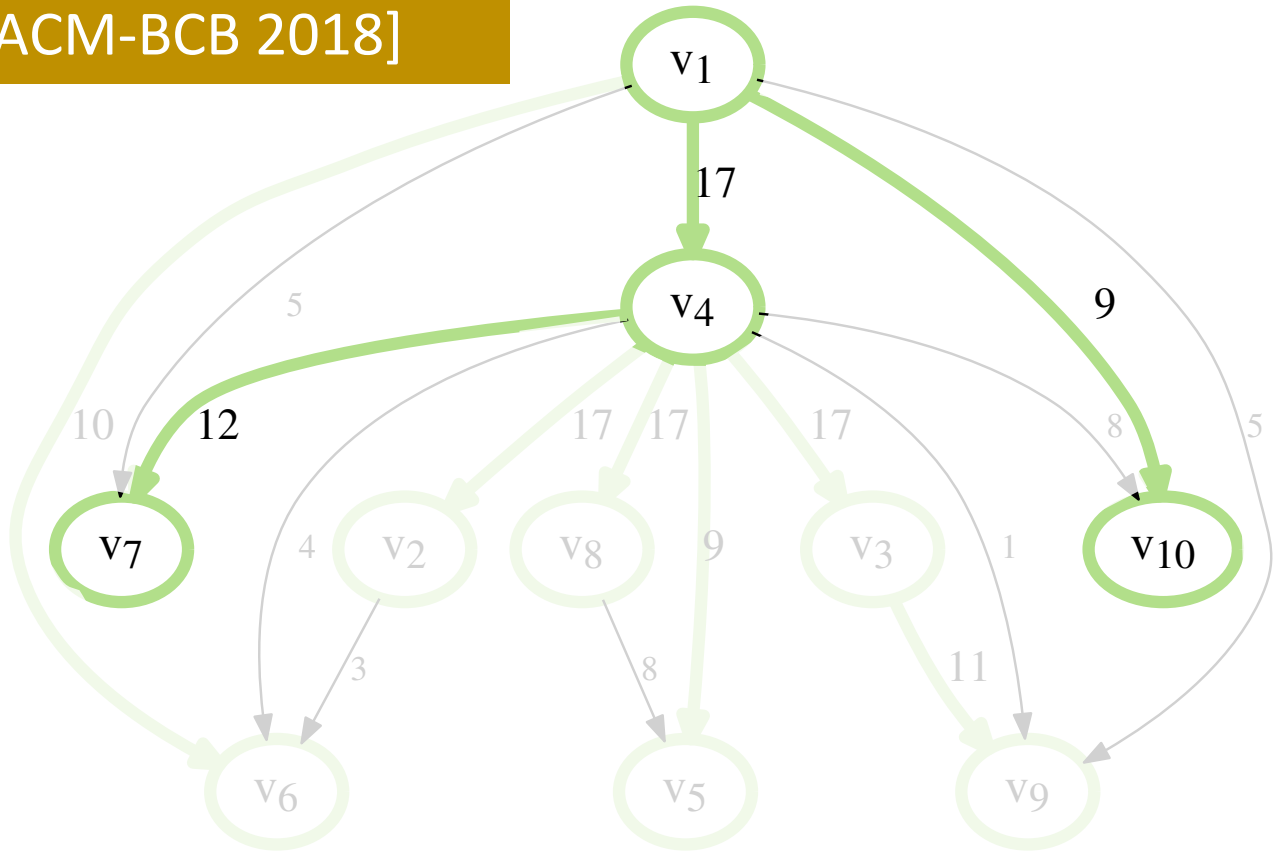
Oesper and colleagues. [ACM-BCB 2018]



	$v_1 \rightarrow v_{10}$	$v_4 \rightarrow v_{10}$
$v_1 \rightarrow v_7$	2	0
$v_4 \rightarrow v_7$	2	5

Summarizing the solution space using a single consensus tree

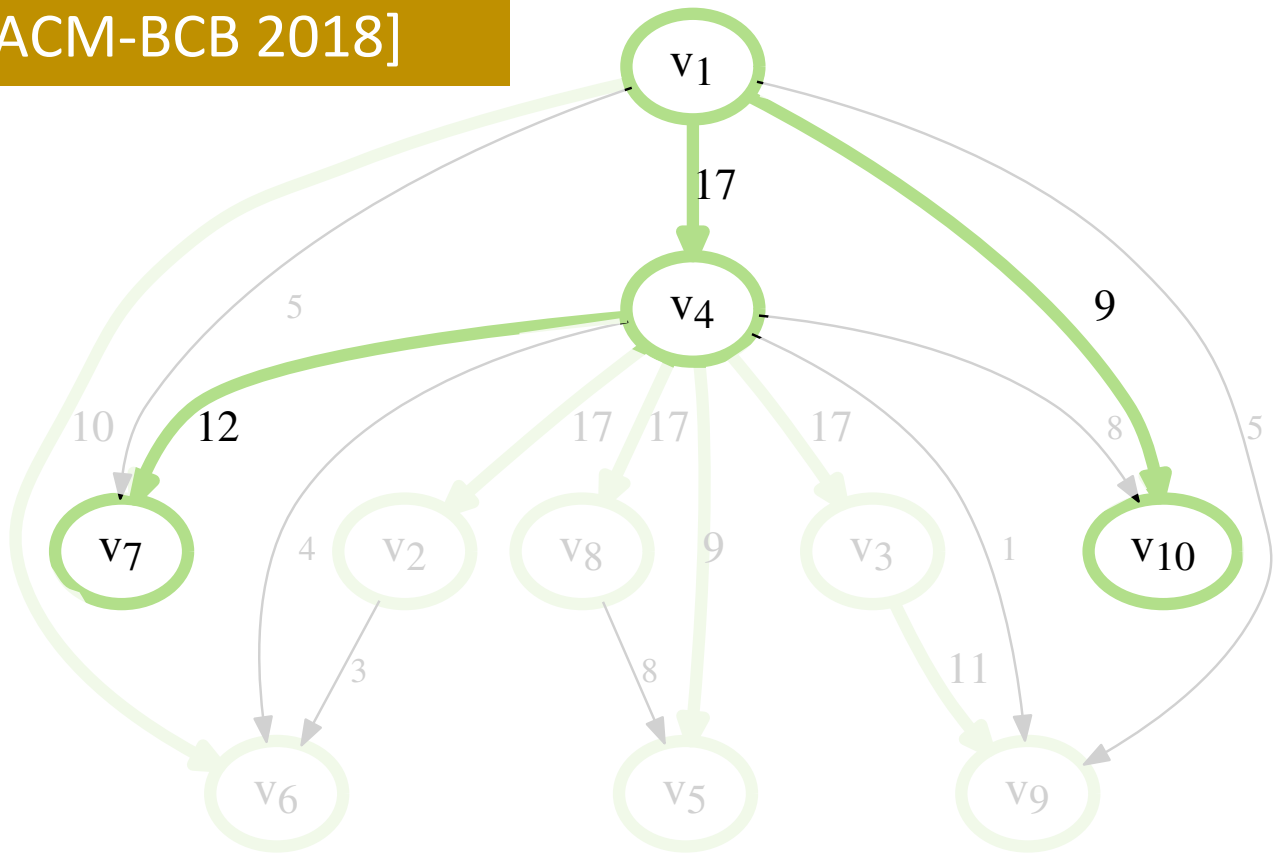
Oesper and colleagues. [ACM-BCB 2018]



	$v_1 \rightarrow v_{10}$	$v_4 \rightarrow v_{10}$
$v_1 \rightarrow v_7$	2	0
$v_4 \rightarrow v_7$	2	5

Summarizing the solution space using a single consensus tree

Oesper and colleagues. [ACM-BCB 2018]



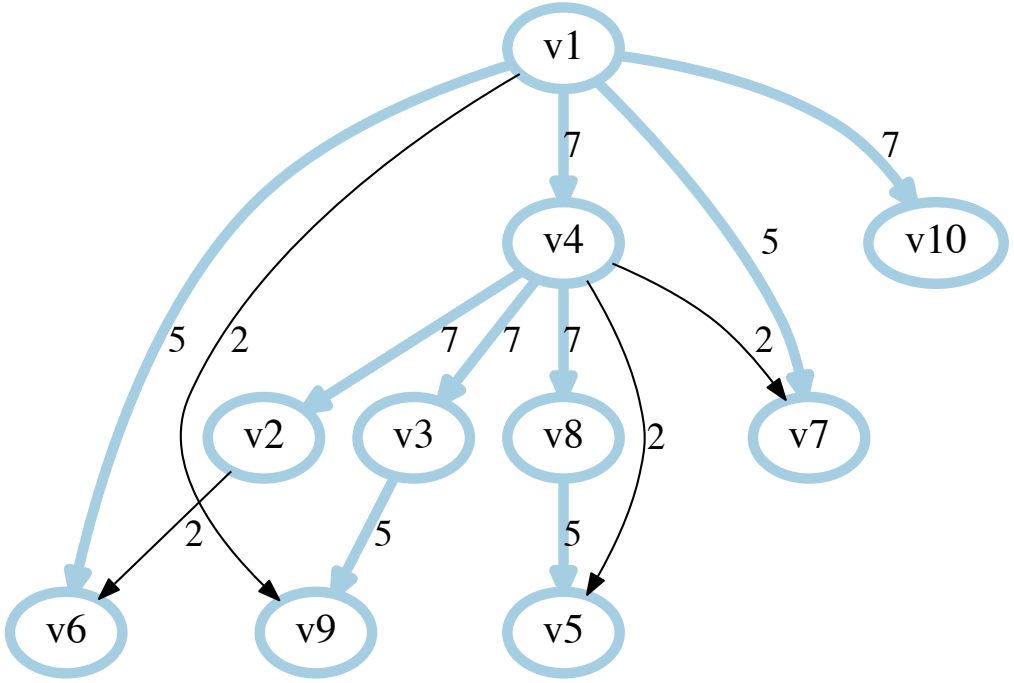
	$v_1 \rightarrow v_{10}$	$v_4 \rightarrow v_{10}$
$v_1 \rightarrow v_7$	2	0
$v_4 \rightarrow v_7$	2	5

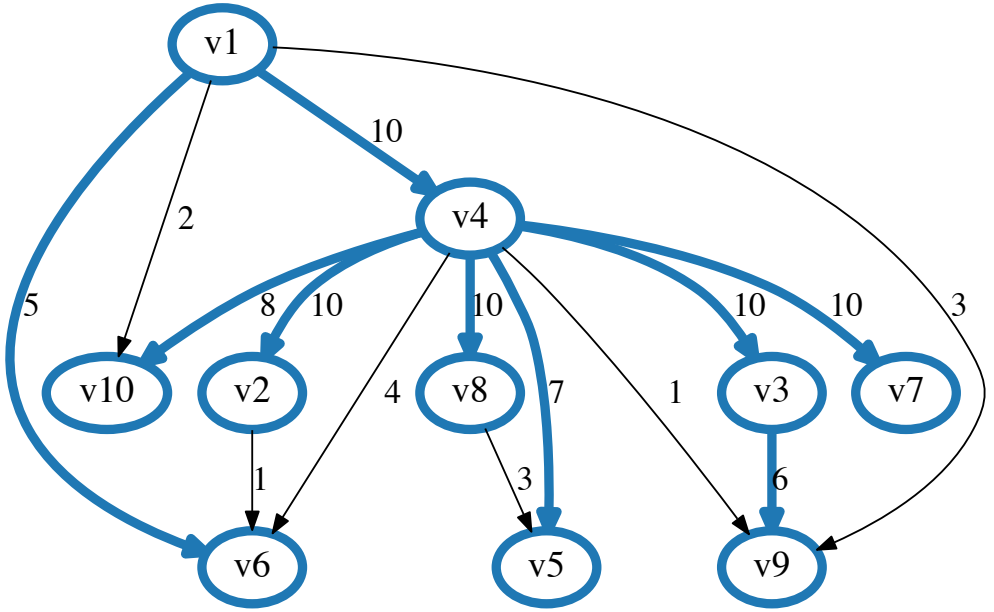
Single consensus tree results in inaccurate summary of diverse solution spaces

Multiple Consensus Trees problem

Multiple Consensus Trees problem

Simultaneous clustering and consensus tree inference



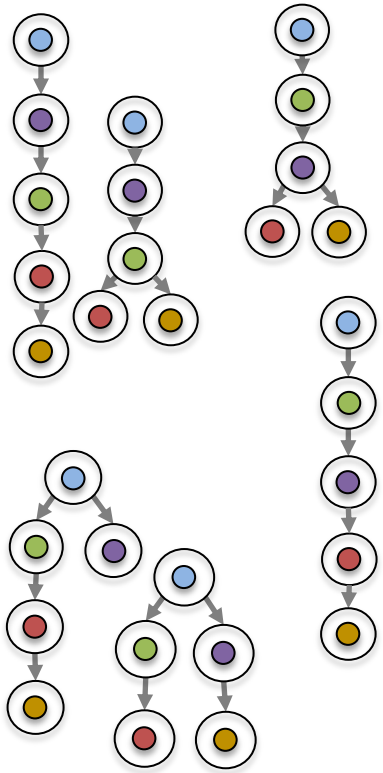


Multiple Consensus Trees (MCT): [ISMB/ECCB 2019]

Given trees $\mathcal{T} = \{T_1, \ldots, T_n\}$ and $k > 0$, find surjective clustering $\sigma : [n] \to [k]$ and consensus trees $\mathcal{R} = \{R_1, \ldots, R_k\}$ s.t. $\sum_{i=1}^{n} d(T_i, R_{\sigma(i)})$ is minimum

Combinatorial characterization of solutions to MCT

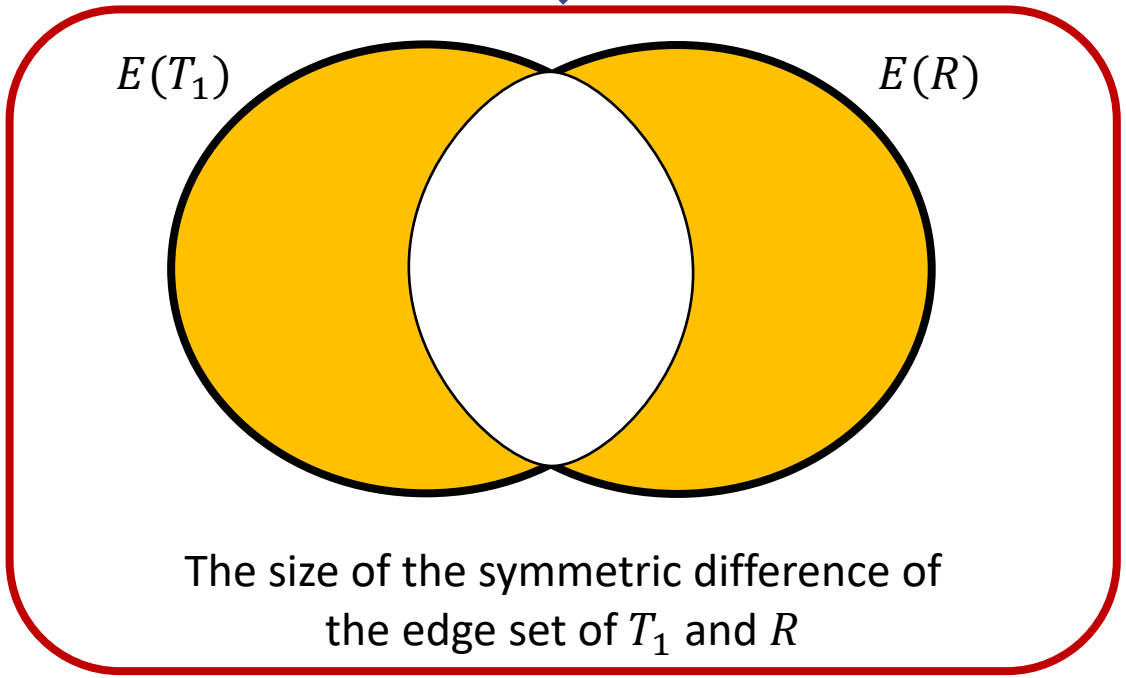
Single Consensus Trees (SCT): [Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \dots, T_n\}$, find consensus tree R s.t.
$\sum_{i=1}^{n} d(T_i, R)$ is minimum



Solution Space $\mathcal{T}$

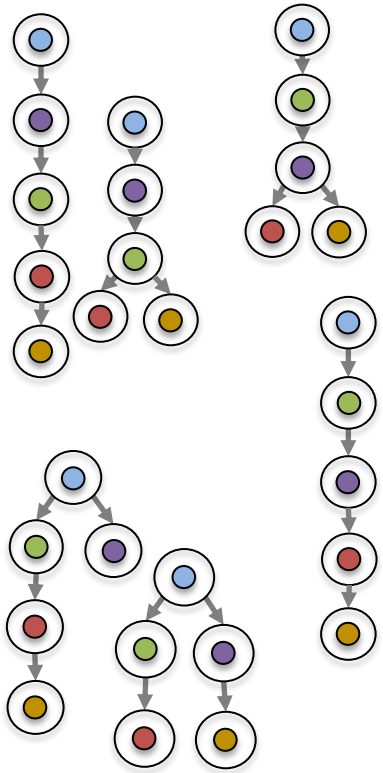
Combinatorial characterization of solutions to MCT

Single Consensus Trees (SCT): [Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \dots, T_n\}$, find consensus tree R s.t.
$\sum_{i=1}^{n} d(T_i, R)$ is minimum



The size of the symmetric difference of the edge set of T_1 and R

Parent-Child distance

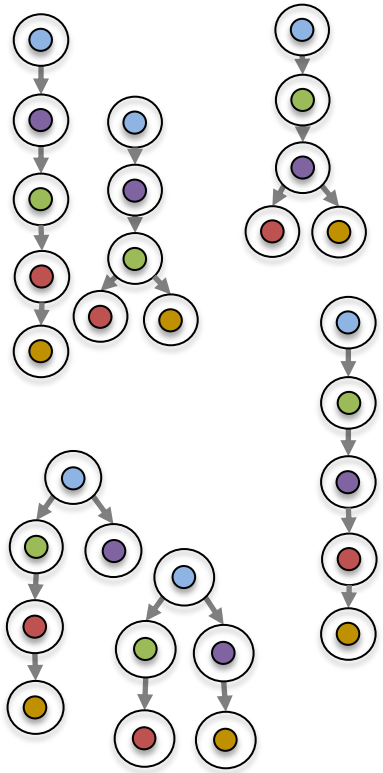


Solution Space $\mathcal{T}$

Combinatorial characterization of solutions to MCT

Single Consensus Trees (SCT): [Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \dots, T_n\}$, find consensus tree R s.t.
$\sum_{i=1}^{n} d(T_i, R)$ is minimum

Theorem: [Govek et al., ACM-BCB 2018]
Max weight spanning arborescences
of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT

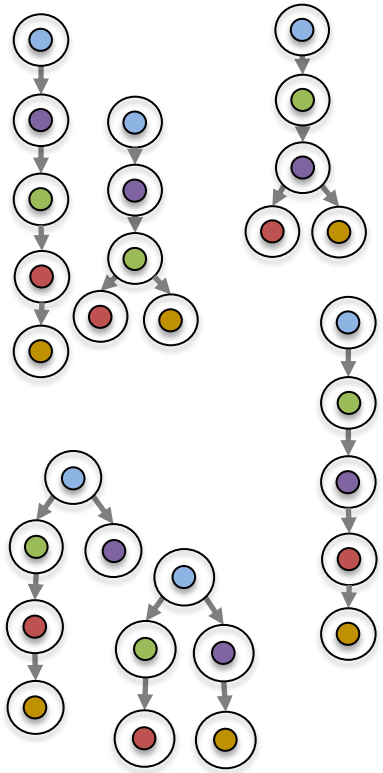


Solution Space $\mathcal{T}$

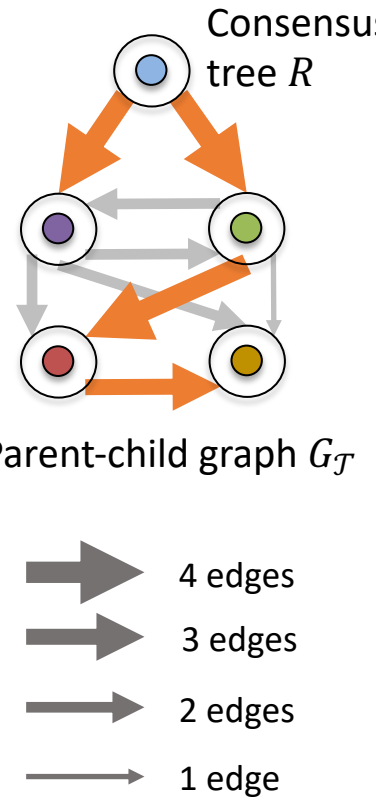
Combinatorial characterization of solutions to MCT

Single Consensus Trees (SCT): [Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \ldots, T_n\}$, find consensus tree R s.t.
$\sum_{i=1}^{n} d(T_i, R)$ is minimum

Theorem: [Govek et al., ACM-BCB 2018]
Max weight spanning arborescences
of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT



Solution Space $\mathcal{T}$



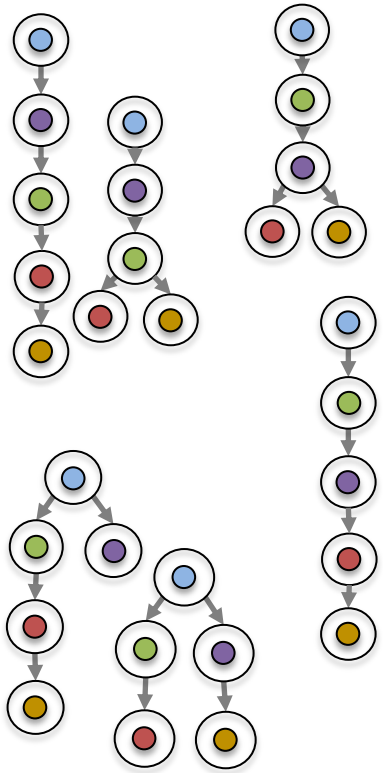
Parent-child graph $G_{\mathcal{T}}$

Combinatorial characterization of solutions to MCT

Single Consensus Trees (SCT): [Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \dots, T_n\}$, find consensus tree R s.t.
$\sum_{i=1}^{n} d(T_i, R)$ is minimum

Theorem: [Govek et al., ACM-BCB 2018]
Max weight spanning arborescences
of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT

Multiple Consensus Trees (MCT): [Aguse et al., ISMB 2019]
Given $\mathcal{T} = \{T_1, \dots, T_n\}$ and $k > 0$, find surjective clustering
$\sigma : [n] \to [k]$ and consensus trees $\mathcal{R} = \{R_1, \dots, R_k\}$
s.t. $\sum_{i=1}^{n} d(T_i, R_{\sigma(i)})$ is minimum



Solution Space $\mathcal{T}$

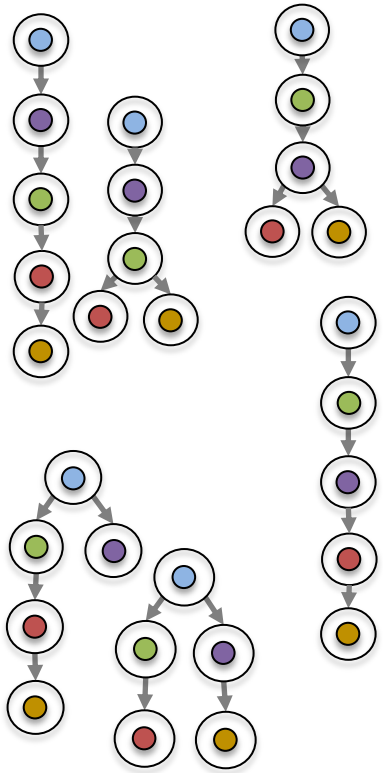
Combinatorial characterization of solutions to MCT

Single Consensus Trees (SCT): [Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \dots, T_n\}$, find consensus tree R s.t.
$\sum_{i=1}^{n} d(T_i, R)$ is minimum

Theorem: [Govek et al., ACM-BCB 2018]
Max weight spanning arborescences
of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT

Multiple Consensus Trees (MCT): [Aguse et al., ISMB 2019]
Given $\mathcal{T} = \{T_1, \dots, T_n\}$ and $k > 0$, find surjective clustering
$\sigma : [n] \to [k]$ and consensus trees $\mathcal{R} = \{R_1, \dots, R_k\}$
s.t. $\sum_{i=1}^{n} d(T_i, R_{\sigma(i)})$ is minimum

Proposition: [Aguse et al., ISMB 2019]
Given fixed clustering $\sigma : [n] \to [k]$, MCT decomposes into
k independent SCT instances



Solution Space $\mathcal{T}$

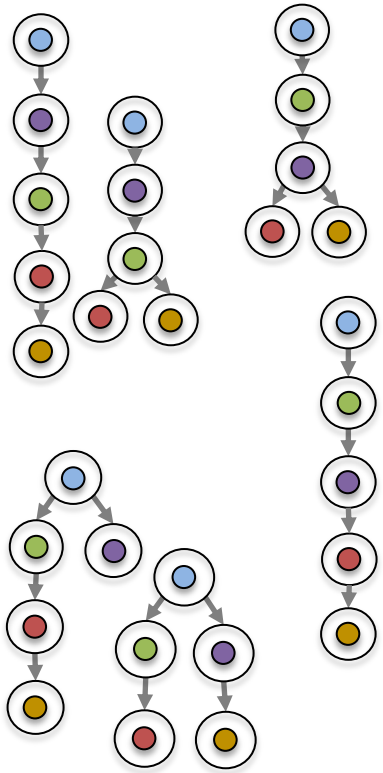
Combinatorial characterization of solutions to MCT

Single Consensus Trees (SCT): [Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \dots, T_n\}$, find consensus tree R s.t.
$\sum_{i=1}^{n} d(T_i, R)$ is minimum

Theorem: [Govek et al., ACM-BCB 2018]
Max weight spanning arborescences
of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT

Multiple Consensus Trees (MCT): [Aguse et al., ISMB 2019]
Given $\mathcal{T} = \{T_1, \dots, T_n\}$ and $k > 0$, find surjective clustering
$\sigma : [n] \rightarrow [k]$ ~~and consensus trees $\mathcal{R} = \{R_1, \dots, R_k\}$~~
s.t. $\sum_{i=1}^{n} d(T_i, R_{\sigma(i)})$ is minimum
where $R_{\sigma(i)}$ is max weight spanning arborescence of $G_{\mathcal{T}_{\sigma(i)}}$

Proposition: [Aguse et al., ISMB 2019]
Given fixed clustering $\sigma : [n] \rightarrow [k]$, MCT decomposes into
k independent SCT instances



Solution Space $\mathcal{T}$

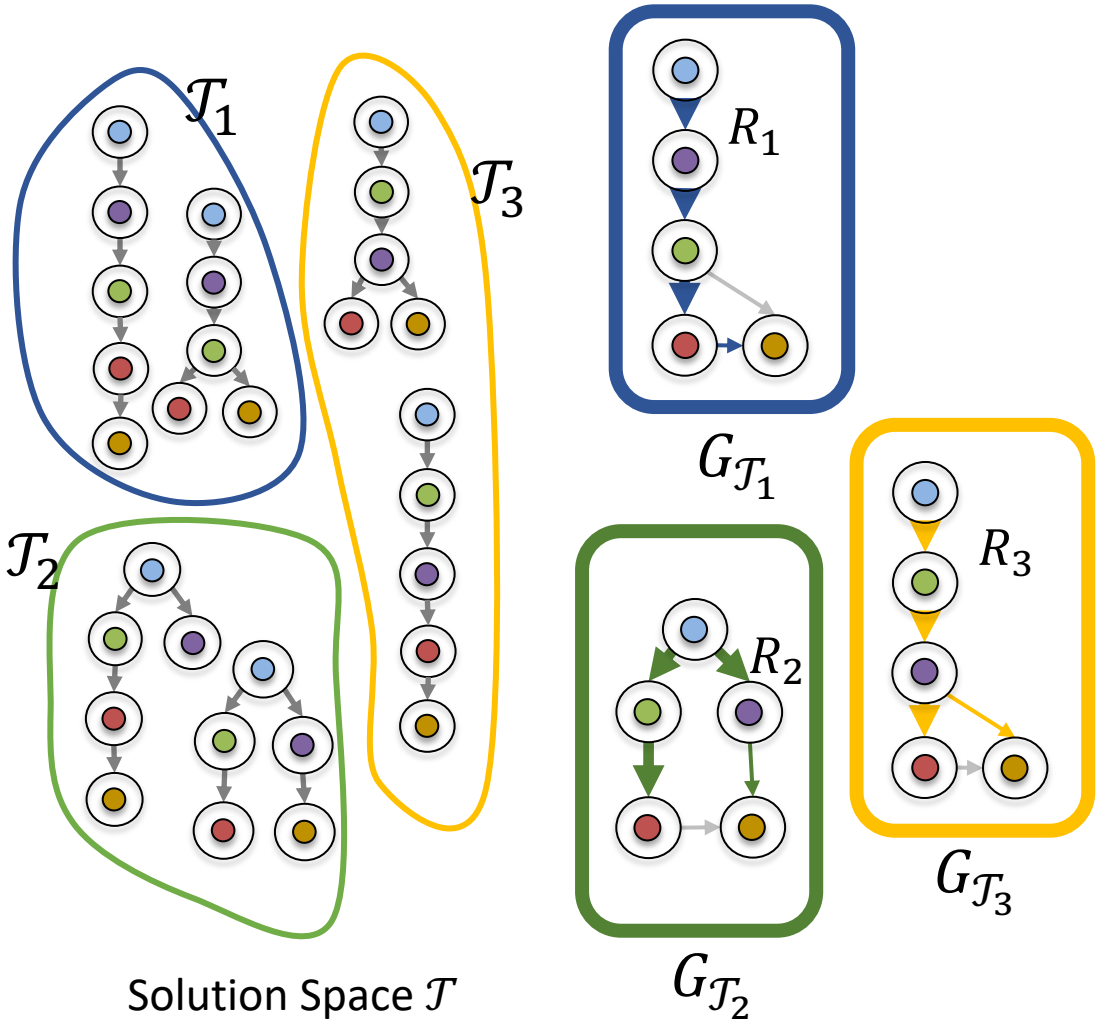
Combinatorial characterization of solutions to MCT

Single Consensus Trees (SCT): [Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \ldots, T_n\}$, find consensus tree R s.t.
$\sum_{i=1}^{n} d(T_i, R)$ is minimum

Theorem: [Govek et al., ACM-BCB 2018]
Max weight spanning arborescences
of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT

Multiple Consensus Trees (MCT): [Aguse et al., ISMB 2019]
Given $\mathcal{T} = \{T_1, \ldots, T_n\}$ and $k > 0$, find surjective clustering
$\sigma : [n] \rightarrow [k]$ ~~and consensus trees $\mathcal{R} = \{R_1, \ldots, R_k\}$~~
s.t. $\sum_{i=1}^{n} d(T_i, R_{\sigma(i)})$ is minimum
where $R_{\sigma(i)}$ is max weight spanning arborescence of $G_{\mathcal{T}_{\sigma(i)}}$

Proposition: [Aguse et al., ISMB 2019]
Given fixed clustering $\sigma : [n] \rightarrow [k]$, MCT decomposes into
k independent SCT instances



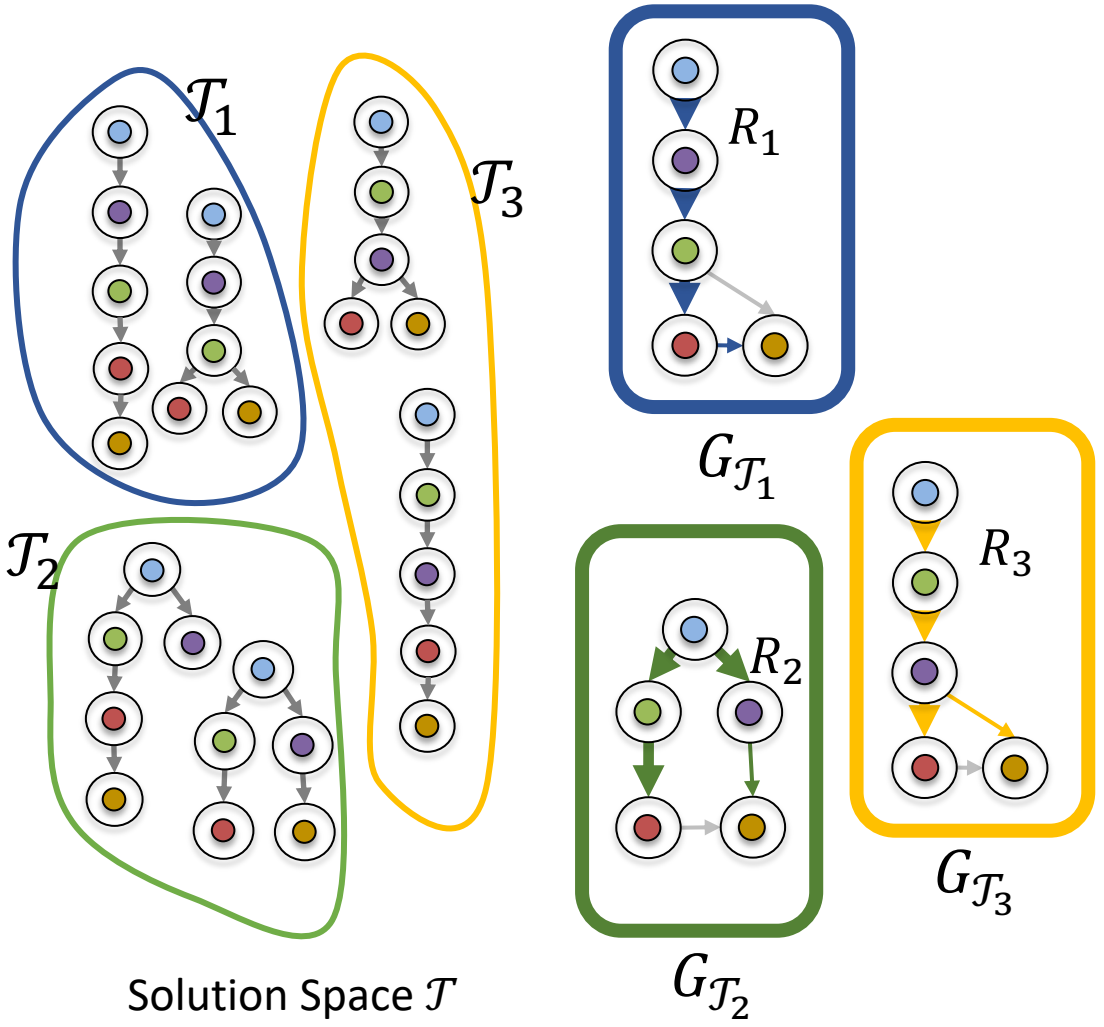
Combinatorial characterization of solutions to MCT

Single Consensus Trees (SCT): [Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \dots, T_n\}$, find consensus tree R s.t.
$\sum_{i=1}^{n} d(T_i, R)$ is minimum

Theorem: [Govek et al., ACM-BCB 2018]
Max weight spanning arborescences
of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT

Multiple Consensus Trees (MCT): [Aguse et al., ISMB 2019]
Given $\mathcal{T} = \{T_1, \dots, T_n\}$ and $k > 0$, find surjective clustering
$\sigma : [n] \rightarrow [k]$ ~~and consensus trees $\mathcal{R} = \{R_1, \dots, R_k\}$~~
s.t. $\sum_{i=1}^{n} d(T_i, R_{\sigma(i)})$ is minimum
where $R_{\sigma(i)}$ is max weight spanning arborescence of $G_{\mathcal{T}_{\sigma(i)}}$

Proposition: [Aguse et al., ISMB 2019]
Given fixed clustering $\sigma : [n] \rightarrow [k]$, MCT decomposes into
k independent SCT instances



How do we find σ^*?

Methods and results

Mixed Integer Linear Program

Theorem: MCT is NP-hard for general k
(by reduction from CLIQUE).

$$\min n(m-1) - \sum_{i=1}^{n}\sum_{s=1}^{k}\sum_{p=1}^{m}\sum_{q=1}^{m} w_{i,s,p,q}$$

$$\text{s.t.} \quad \sum_{s=1}^{k} x_{i,s} = 1 \qquad \forall i \in [n]$$

$$\sum_{i=1}^{n} x_{i,s} \geq 1 \qquad \forall s \in [k]$$

$$\sum_{p=1}^{m} z_{s,p} = 1 \qquad \forall s \in [k]$$

$$\sum_{q=1}^{m} y_{s,p,q} = 1 - z_{s,p} \qquad \forall s \in [k], p \in [m]$$

$$y_{s,p,q} \leq b_{p,q} \qquad \forall s \in [k], p, q \in [m]$$

$$\sum_{(p,q)\in\delta^-(U)} y_{s,p,q} + \sum_{p\in U} z_{s,p} \geq 1 \qquad \forall s \in [k], U \subseteq [m]$$

$$w_{i,s,p,q} \leq a_{i,p,q} \qquad \forall i \in [n], s \in [k], p, q \in [m]$$

$$w_{i,s,p,q} \leq x_{i,s} \qquad \forall i \in [n], s \in [k], p, q \in [m]$$

$$w_{i,s,p,q} \leq y_{s,p,q} \qquad \forall i \in [n], s \in [k], p, q \in [m]$$

$$w_{i,s,p,q} \geq 0 \qquad \forall i \in [n], s \in [k], p, q \in [m]$$

$$y_{s,p,q} \leq \sum_{i=1}^{n} a_{i,p,q} x_{i,s} \qquad \forall s \in [k], p, q \in [m]$$

$$y_{s,p,q} \geq \sum_{i=1}^{n} a_{i,p,q} x_{i,s} - \sum_{i=1}^{n} x_{i,s} + 1 \qquad \forall s \in [k], p, q \in [m]$$

$$\sum_{i=1}^{n} x_{i,s} \geq \sum_{i=1}^{n} x_{i,s+1} + 1 \qquad \forall s \in [k-1]$$

$$x_{i,s} \in \{0,1\} \qquad \forall i \in [n], s \in [k]$$

$$y_{s,p,q} \geq 0 \qquad \forall s \in [k], p, q \in [m]$$

$$z_{s,p} \geq 0 \qquad \forall s \in [k], p \in [m]$$

Mixed Integer Linear Program

Theorem: MCT is NP-hard for general k (by reduction from CLIQUE).

$x_{i,s} \in \{0,1\}$ Tree T_i is assigned to cluster s

$y_{s,p,q} \geq 0$ Edge (p,q) is present in consensus tree R_s

$z_{s,p} \geq 0$ Vertex p is root of consensus tree R_s

$$\min n(m-1) - \sum_{i=1}^{n}\sum_{s=1}^{k}\sum_{p=1}^{m}\sum_{q=1}^{m} w_{i,s,p,q}$$

$$\text{s.t.} \quad \sum_{s=1}^{k} x_{i,s} = 1 \qquad \forall i \in [n]$$

$$\sum_{i=1}^{n} x_{i,s} \geq 1 \qquad \forall s \in [k]$$

$$\sum_{p=1}^{m} z_{s,p} = 1 \qquad \forall s \in [k]$$

$$\sum_{q=1}^{m} y_{s,p,q} = 1 - z_{s,p} \qquad \forall s \in [k], p \in [m]$$

$$y_{s,p,q} \leq b_{p,q} \qquad \forall s \in [k], p,q \in [m]$$

$$\sum_{(p,q)\in\delta^-(U)} y_{s,p,q} + \sum_{p\in U} z_{s,p} \geq 1 \qquad \forall s \in [k], U \subseteq [m]$$

$$w_{i,s,p,q} \leq a_{i,p,q} \qquad \forall i \in [n], s \in [k], p,q \in [m]$$

$$w_{i,s,p,q} \leq x_{i,s} \qquad \forall i \in [n], s \in [k], p,q \in [m]$$

$$w_{i,s,p,q} \leq y_{s,p,q} \qquad \forall i \in [n], s \in [k], p,q \in [m]$$

$$w_{i,s,p,q} \geq 0 \qquad \forall i \in [n], s \in [k], p,q \in [m]$$

$$y_{s,p,q} \leq \sum_{i=1}^{n} a_{i,p,q} x_{i,s} \qquad \forall s \in [k], p,q \in [m]$$

$$y_{s,p,q} \geq \sum_{i=1}^{n} a_{i,p,q} x_{i,s} - \sum_{i=1}^{n} x_{i,s} + 1 \qquad \forall s \in [k], p,q \in [m]$$

$$\sum_{i=1}^{n} x_{i,s} \geq \sum_{i=1}^{n} x_{i,s+1} + 1 \qquad \forall s \in [k-1]$$

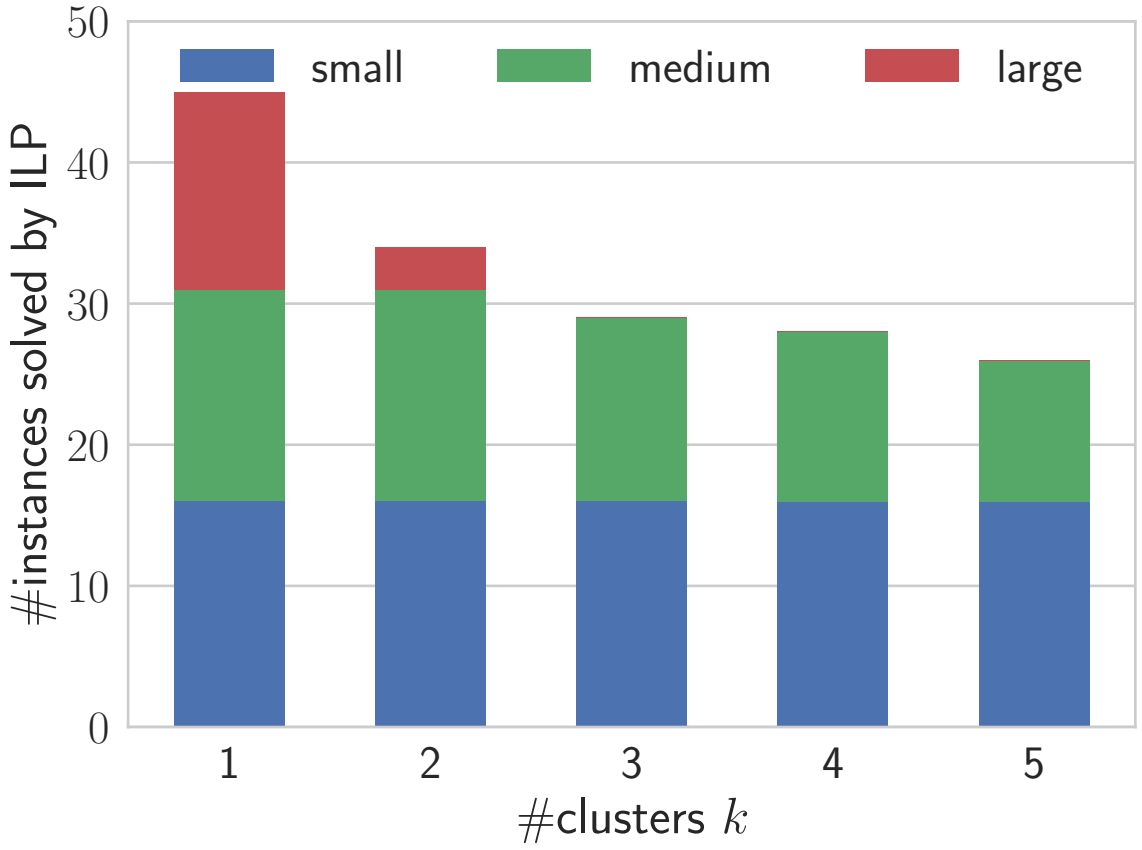
$$x_{i,s} \in \{0,1\} \qquad \forall i \in [n], s \in [k]$$

$$y_{s,p,q} \geq 0 \qquad \forall s \in [k], p,q \in [m]$$

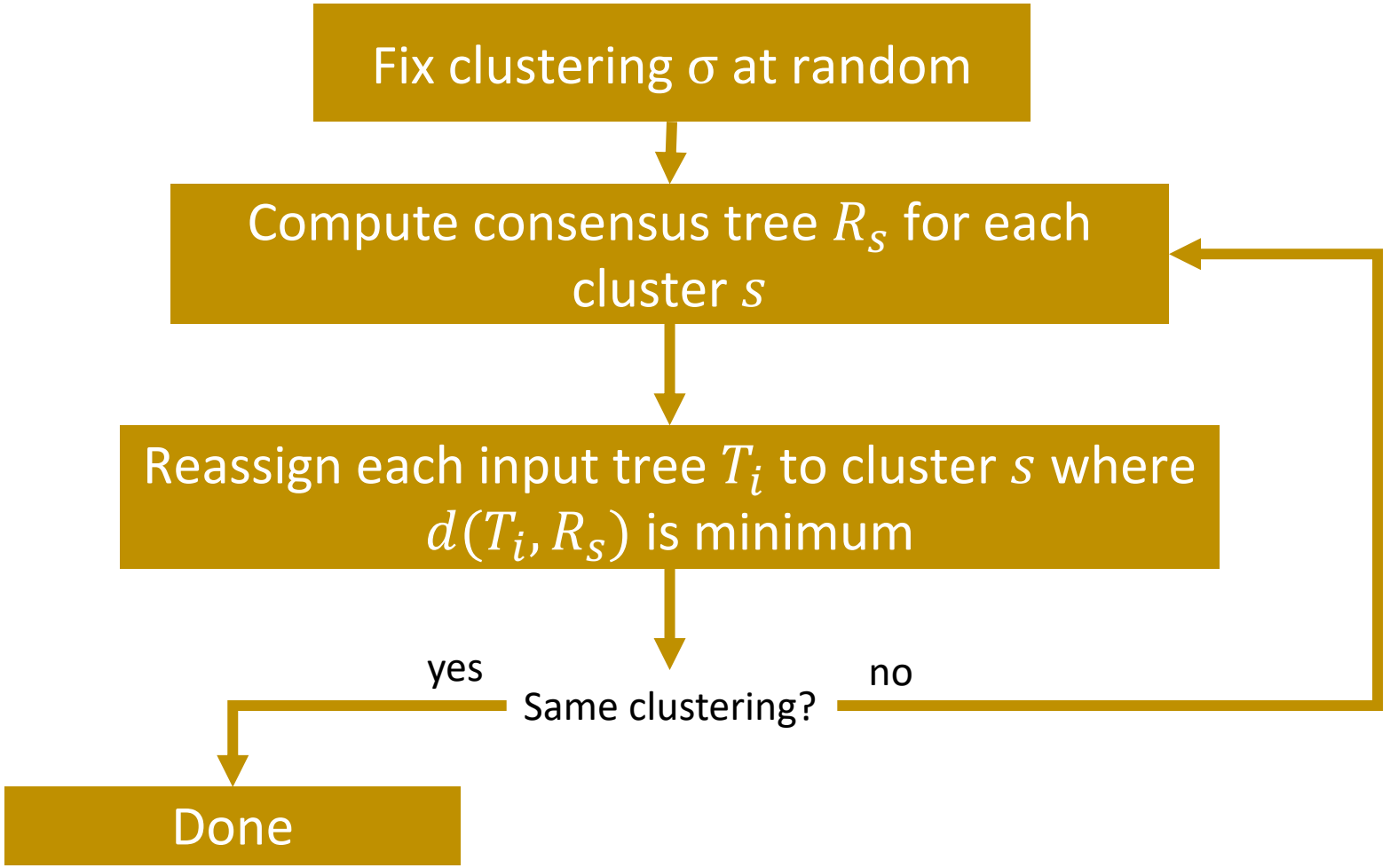
$$z_{s,p} \geq 0 \qquad \forall s \in [k], p \in [m]$$

MILP does not scale well with k and n

We simulated **small**, **medium** and **large** instances of phylogeny inference solution spaces



Coordinate Ascent heuristic



Coordinate Ascent heuristic finds optimal solutions efficiently

Small, medium, and large simulated instances

	#clusters k	MILP (1 h)	CA (100 r.)
small (16)	2	16	16
	3	16	16
	4	16	16
	5	16	16
medium (15)	2	15	15
	3	13	13
	4	12	12
	5	10	10
large (14)	2	3	3
	3	0	0
	4	0	0
	5	0	0

Number of instances solved by MILP to provable optimality

Coordinate Ascent heuristic finds optimal solutions efficiently

Small, medium, and large simulated instances

	#clusters k	MILP (1 h)	CA (100 r.)
small (16)	2	16	16
	3	16	16
	4	16	16
	5	16	16
medium (15)	2	15	15
	3	13	13
	4	12	12
	5	10	10
large (14)	2	3	3
	3	0	0
	4	0	0
	5	0	0

Number of instances where heuristic returned MILP's optimal solution

Coordinate Ascent heuristic finds optimal solutions efficiently

Small, medium, and large simulated instances

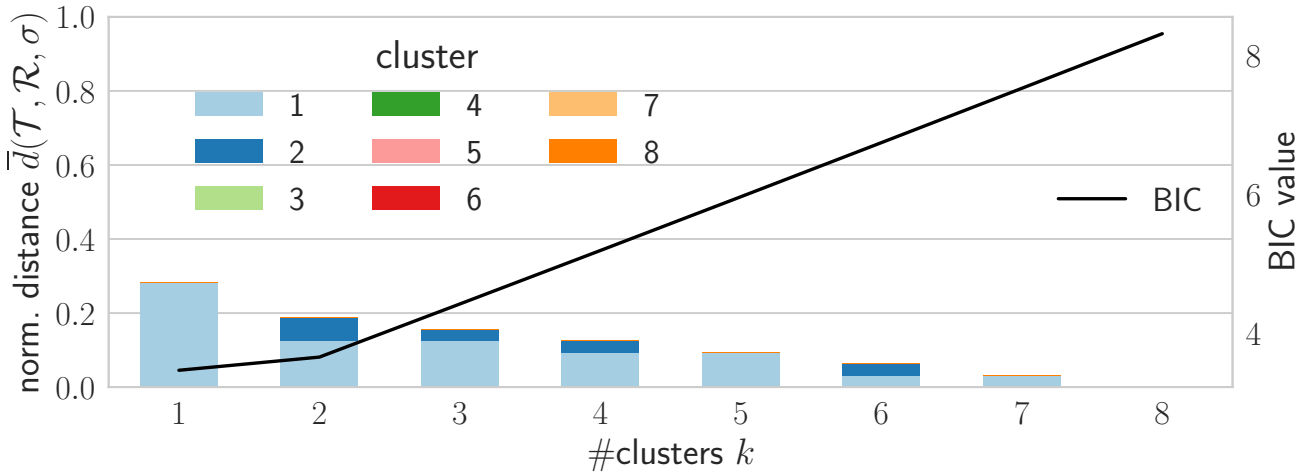
	#clusters k	MILP (1 h)	CA (100 r.)
small (16)	2	16	16
	3	16	16
	4	16	16
	5	16	16
medium (15)	2	15	15
	3	13	13
	4	12	12
	5	10	10
large (14)	2	3	3
	3	0	0
	4	0	0
	5	0	0

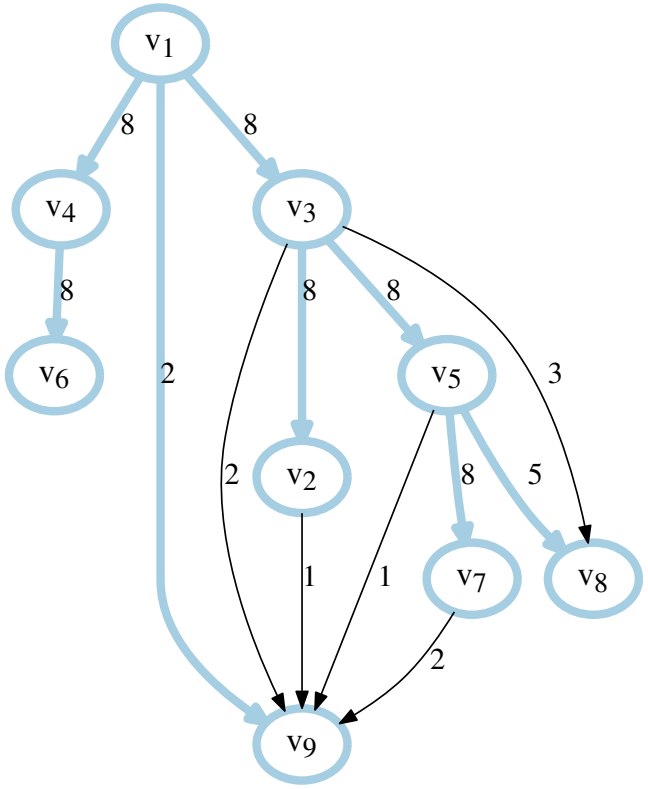
What about the number of clusters, k?

Bayesian Information Criterion determines the number of clusters for each solution space

Jamal-Hanjani et al. (2017). *NEJM*.

Jamal-Hanjani et al. inferred 8 trees for patient CRUK0013

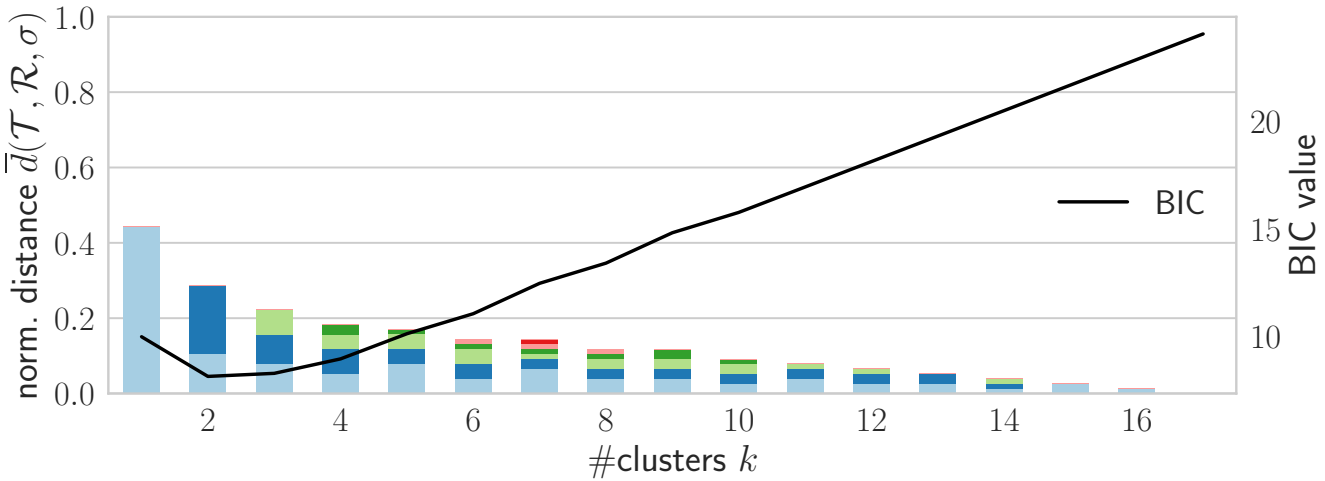


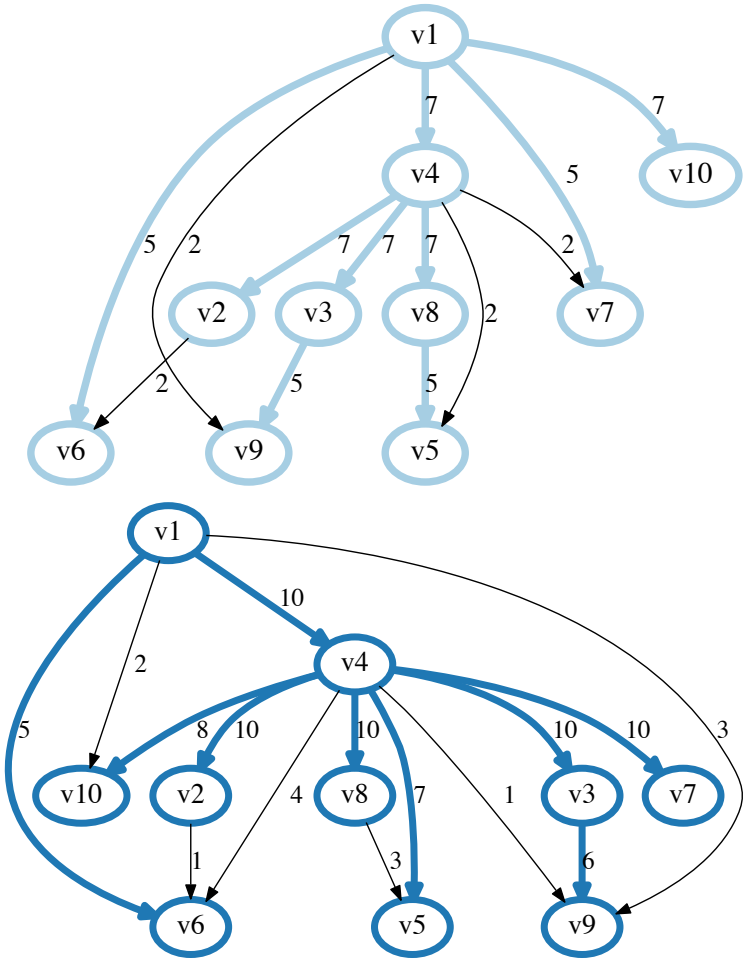


Bayesian Information Criterion determines the number of clusters for each solution space

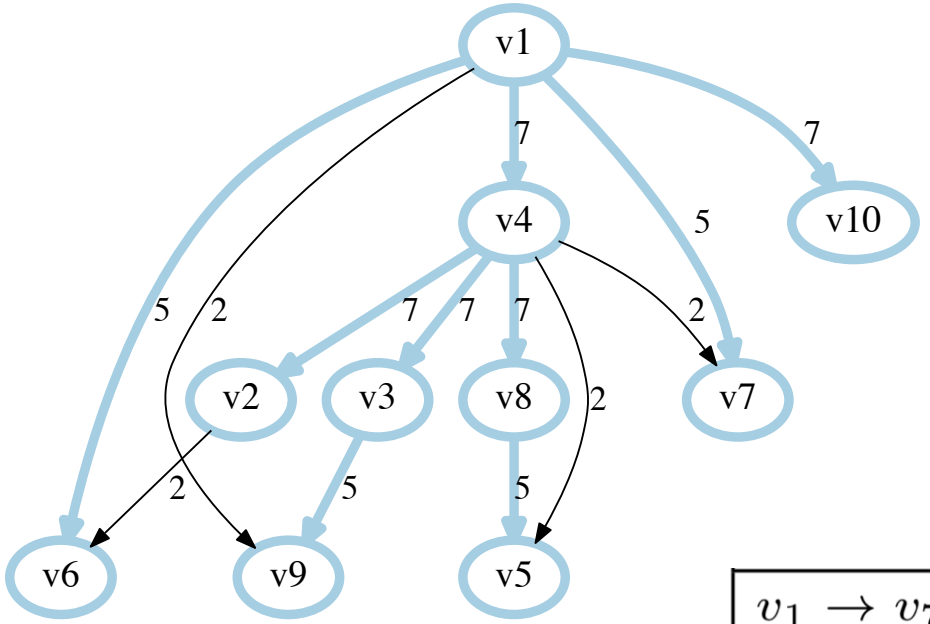
Jamal-Hanjani et al. (2017). *NEJM*.

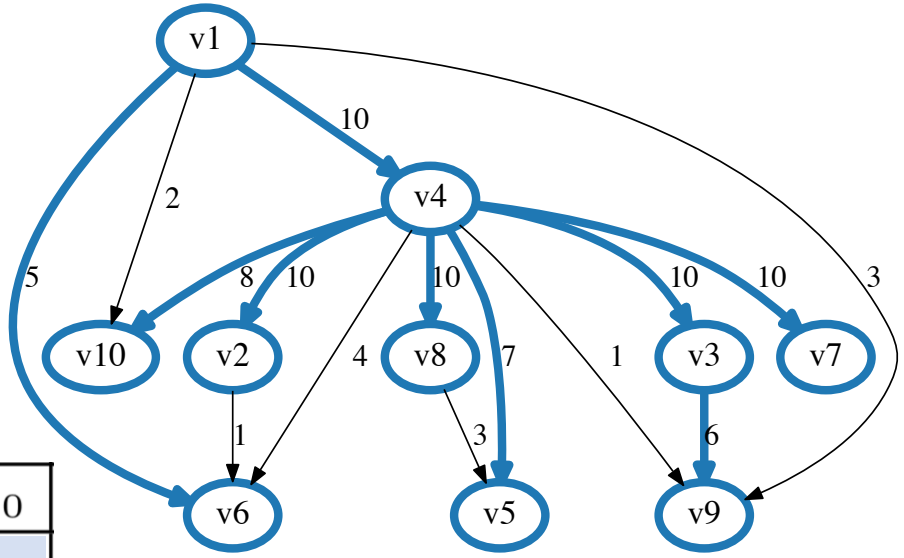
Jamal-Hanjani et al. inferred 17 trees for patient CRUK0037



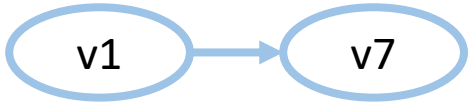


Multiple Consensus Trees capture patterns of mutual exclusivity and co-occurrence

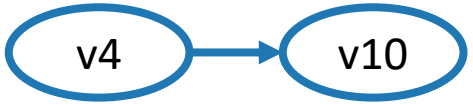




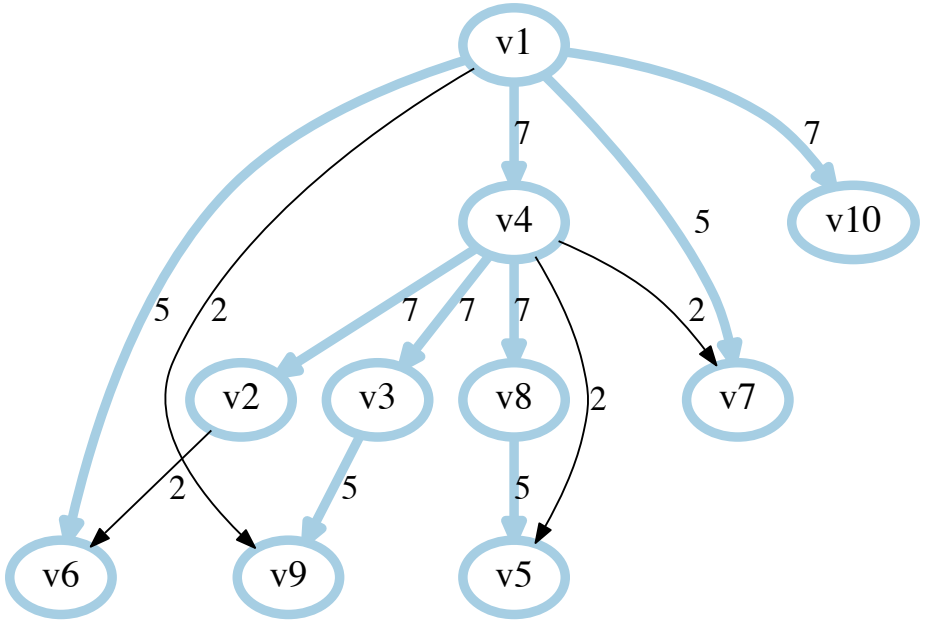
	$v_1 \rightarrow v_{10}$	$v_4 \rightarrow v_{10}$
$v_1 \rightarrow v_7$	2	0
$v_4 \rightarrow v_7$	2	5

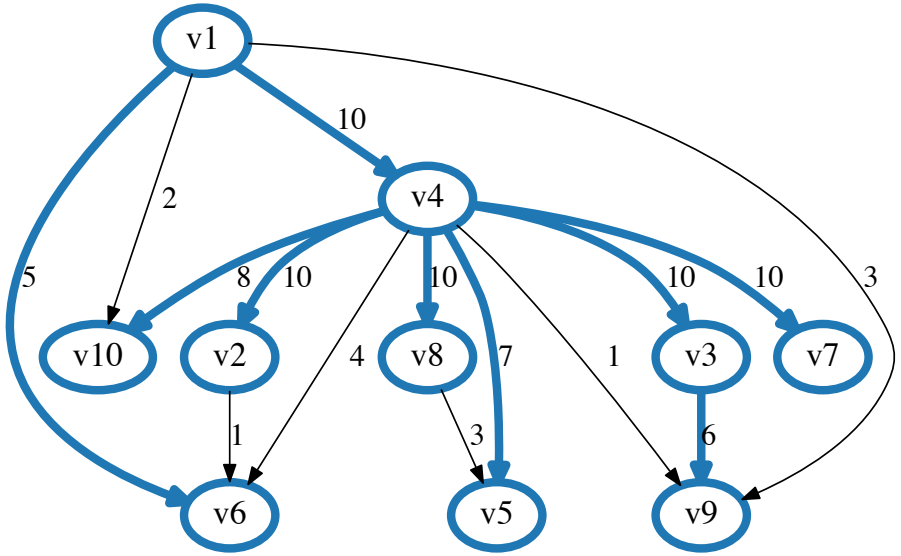


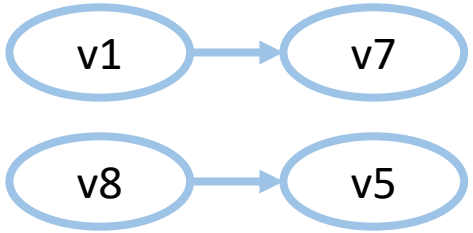
These edges are mutually exclusive



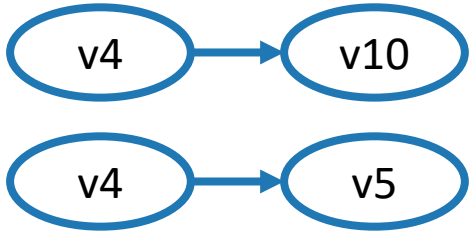
Multiple Consensus Trees capture patterns of mutual exclusivity and co-occurrence







These edges tend to co-occur in the trees in the solution space

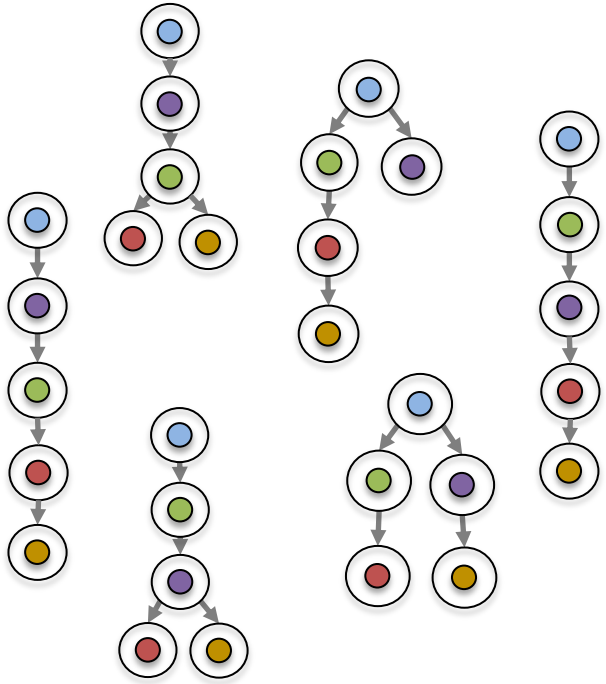


Conclusions

- Introduced the Multiple Consensus Trees (MCT) problem
- Showed hardness and presented a mixed integer linear program
- Presented an efficient heuristic and showed that it finds optimal solutions
- Model selection for the number of clusters

Future directions

- Relax infinite sites assumption
- Use alternative distance functions



Acknowledgements

- El-Kebir group
  - **Mohammed El-Kebir**
  - **Yuanyuan Qi**
  - Juho Kim
  - Jiaqi Wu

- ISCB for Travel Fellowship Award
- UIUC Center for Computational Biotechnology and Genomic Medicine (grant: CSN 1624790)
- National Science Foundation (CCF-1850502)