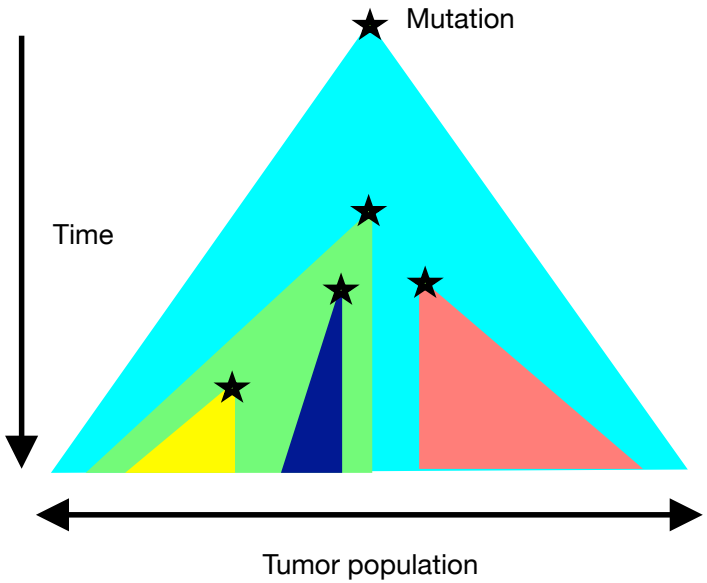
Examining Tumor Phylogeny Inference in Noisy Sequencing Data

Kiran Tomlinson and Layla Oesper

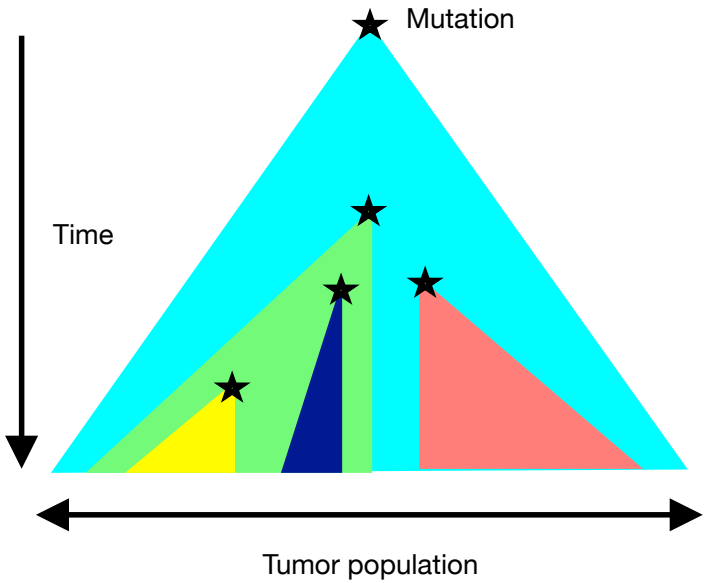
Department of Computer Science, Carleton College

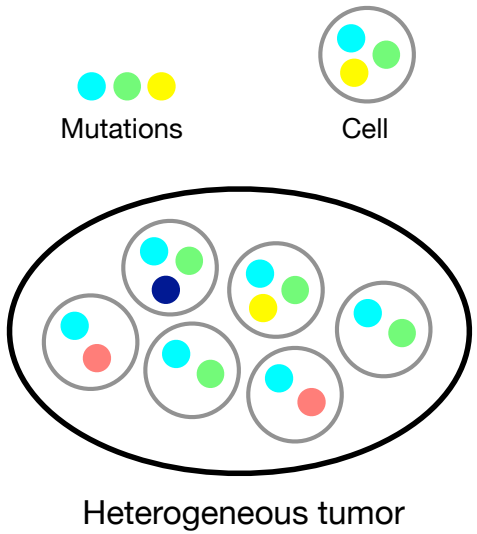
Dec. 4, 2018

Clonal theory (Nowell 1976)

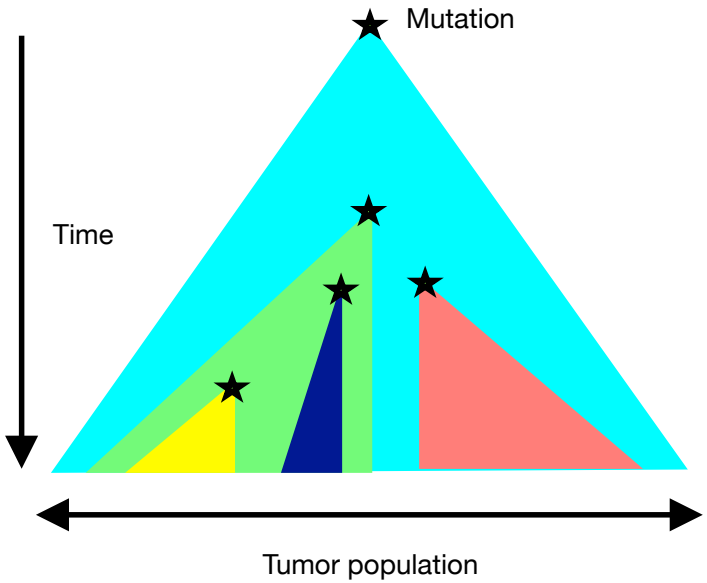


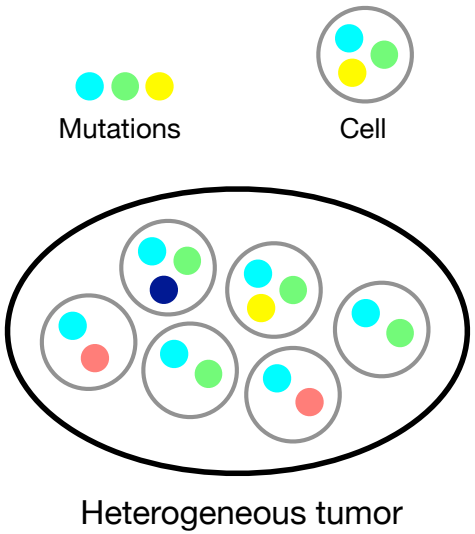
Clonal theory (Nowell 1976)

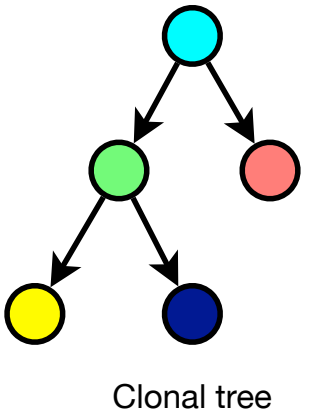




Clonal theory (Nowell 1976)

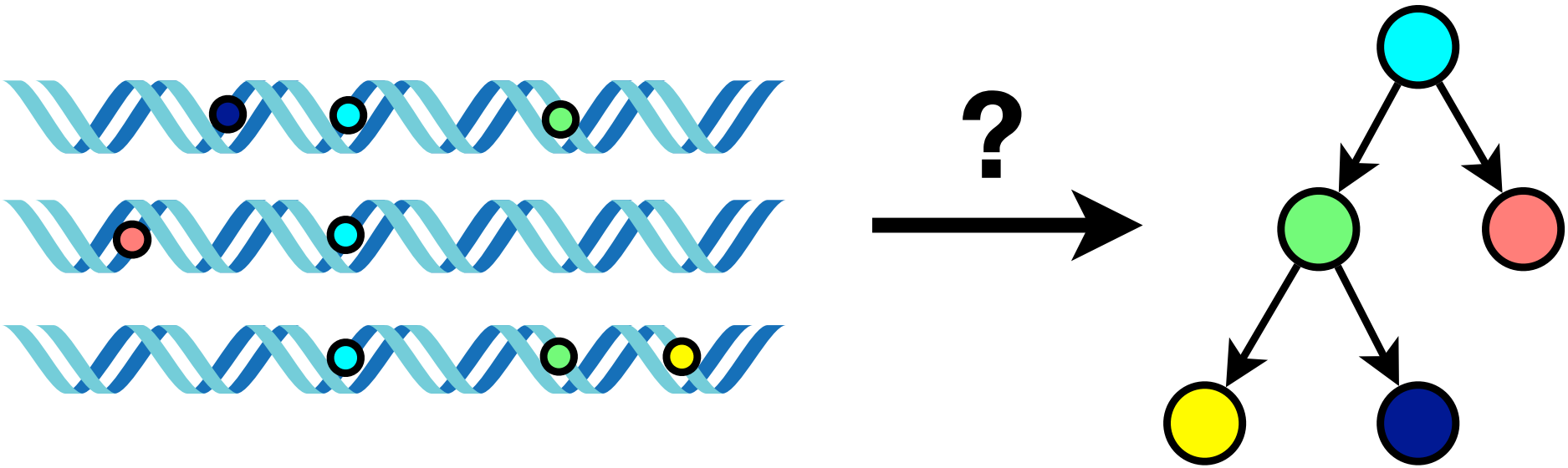






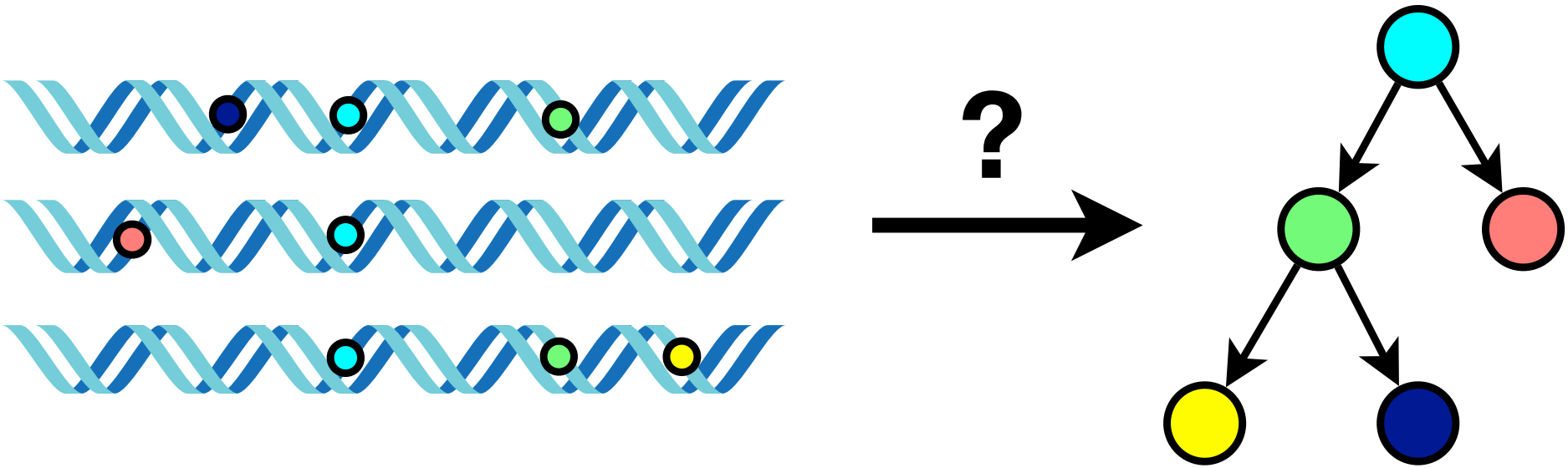
Inferring tumor phylogeny

How can we reconstruct a tumor's clonal tree from its genome?



Inferring tumor phylogeny

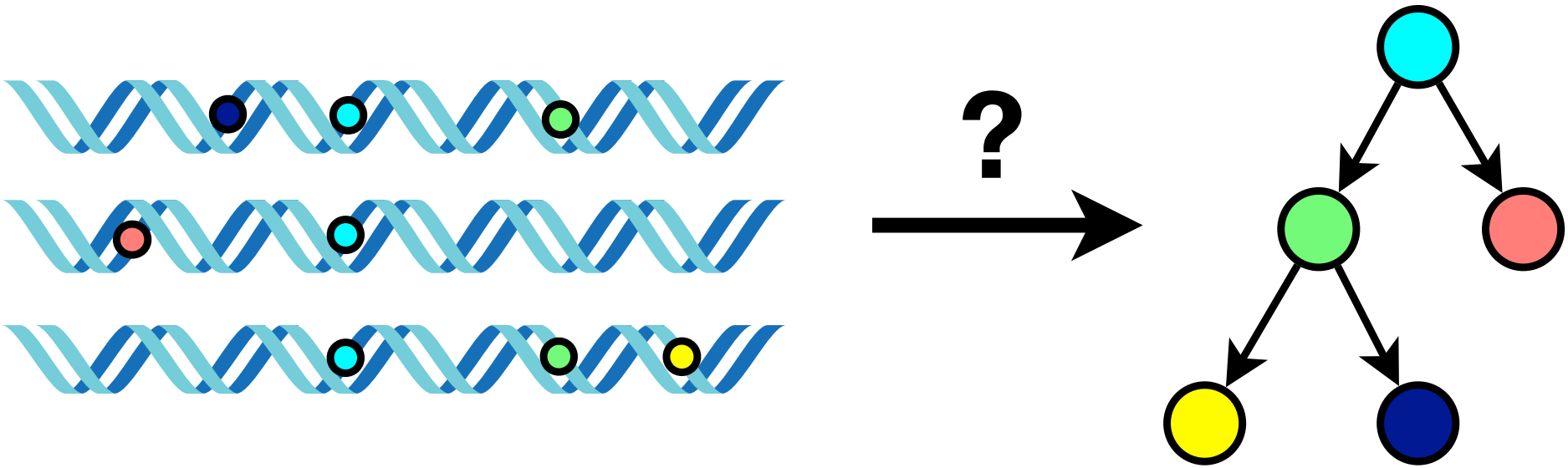
How can we reconstruct a tumor's clonal tree from its genome?



Why is this important?

Inferring tumor phylogeny

How can we reconstruct a tumor's clonal tree from its genome?

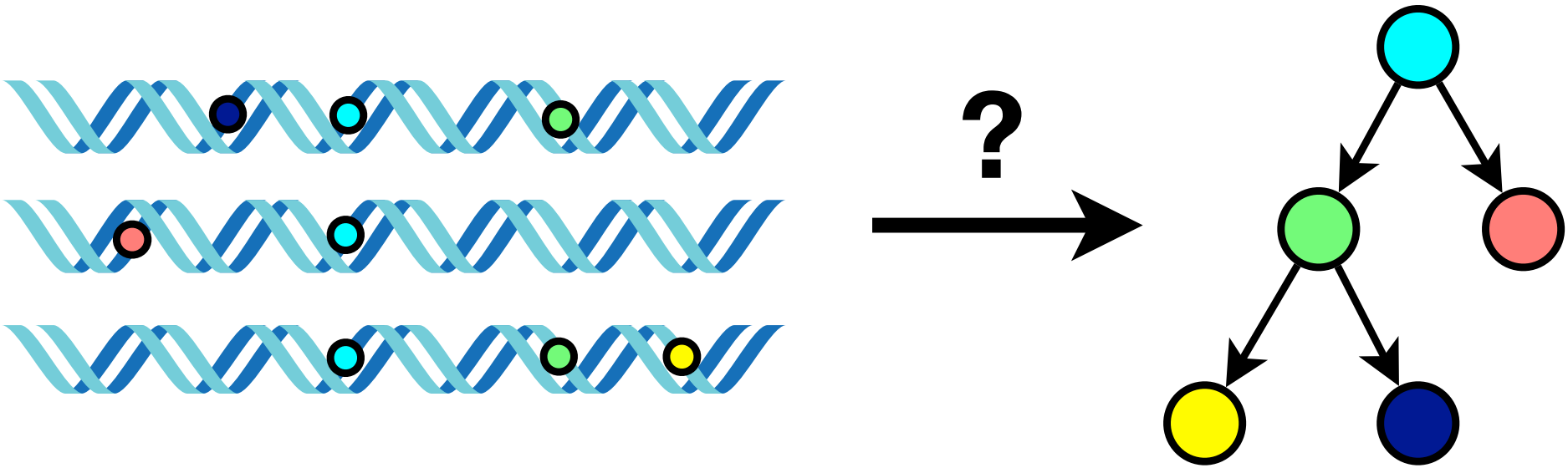


Why is this important?

1. Personalized medicine (Greaves 2015), (McGranahan and Swanton 2017)

Inferring tumor phylogeny

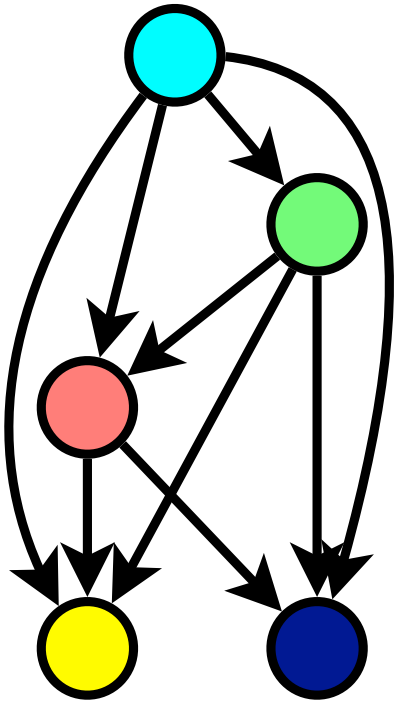
How can we reconstruct a tumor's clonal tree from its genome?



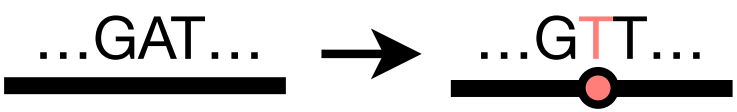
Why is this important?

1. Personalized medicine (Greaves 2015), (McGranahan and Swanton 2017)
2. Improved understanding of cancer development

Outline



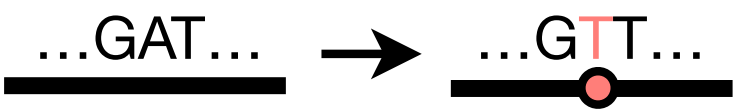
Previous work



Single nucleotide variants (SNV) only:

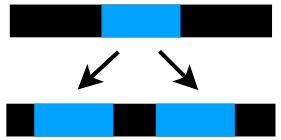
- PhyloSub (Jiao et al. 2014)
- Rec-BTP (Hajirasouliha et al. 2014)
- AncesTree (El-Kebir et al. 2015)
- CITUP (Malikic et al. 2015)
- LICHeE (Popic et al. 2015)
- BitPhylogeny (Yuan et al. 2015)

Previous work



Single nucleotide variants (SNV) only:

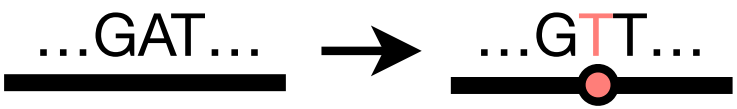
- PhyloSub (Jiao et al. 2014)
- Rec-BTP (Hajirasouliha et al. 2014)
- AncesTree (El-Kebir et al. 2015)
- CITUP (Malikic et al. 2015)
- LICHeE (Popic et al. 2015)
- BitPhylogeny (Yuan et al. 2015)



SNVs and CNAs/structural variants:

- SubcloneSeeker (Qiao et al. 2014)
- PhyloWGS (Deshwar et al. 2015)
- SPRUCE (El-Kebir et al. 2016)
- Canopy (Jiang et al. 2016)
- PASTRI (Satas and Raphael 2017)

Previous work

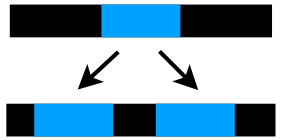


Single nucleotide variants (SNV) only:

- PhyloSub (Jiao et al. 2014)
- Rec-BTP (Hajirasouliha et al. 2014)
- AncesTree (El-Kebir et al. 2015)
- CITUP (Malikic et al. 2015)
- LICHeE (Popic et al. 2015)
- BitPhylogeny (Yuan et al. 2015)

Single-cell sequencing data:

- OncoNEM (Ross et al. 2016)
- SCITE (Jahn et al. 2016)
- SiFit (Zafar et al. 2017)



SNVs and CNAs/structural variants:

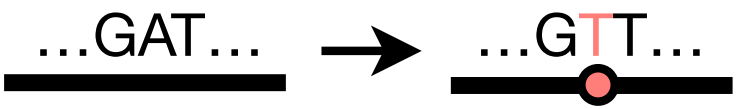
- SubcloneSeeker (Qiao et al. 2014)
- PhyloWGS (Deshwar et al. 2015)
- SPRUCE (El-Kebir et al. 2016)
- Canopy (Jiang et al. 2016)
- PASTRI (Satas and Raphael 2017)

Single-cell and bulk data:

- ddClone (Salehi et al. 2017)
- B-SCITE (Malikic et al. 2018)

and many more....

Previous work

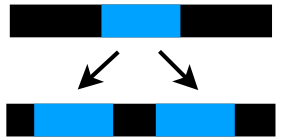


Single nucleotide variants (SNV) only:

- PhyloSub (Jiao et al. 2014)
- Rec-BTP (Hajirasouliha et al. 2014)
- **AncesTree** (El-Kebir et al. 2015)
- CITUP (Malikic et al. 2015)
- LICHeE (Popic et al. 2015)
- BitPhylogeny (Yuan et al. 2015)

Single-cell sequencing data:

- OncoNEM (Ross et al. 2016)
- SCITE (Jahn et al. 2016)
- SiFit (Zafar et al. 2017)



SNVs and CNAs/structural variants:

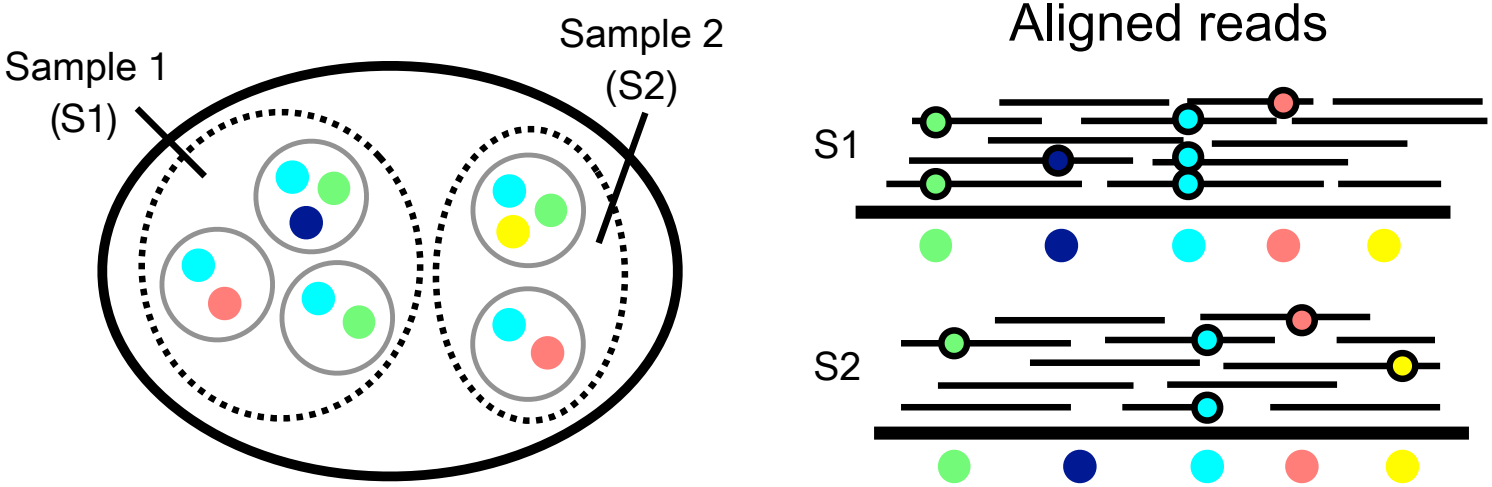
- SubcloneSeeker (Qiao et al. 2014)
- PhyloWGS (Deshwar et al. 2015)
- SPRUCE (El-Kebir et al. 2016)
- Canopy (Jiang et al. 2016)
- PASTRI (Satas and Raphael 2017)

Single-cell and bulk data:

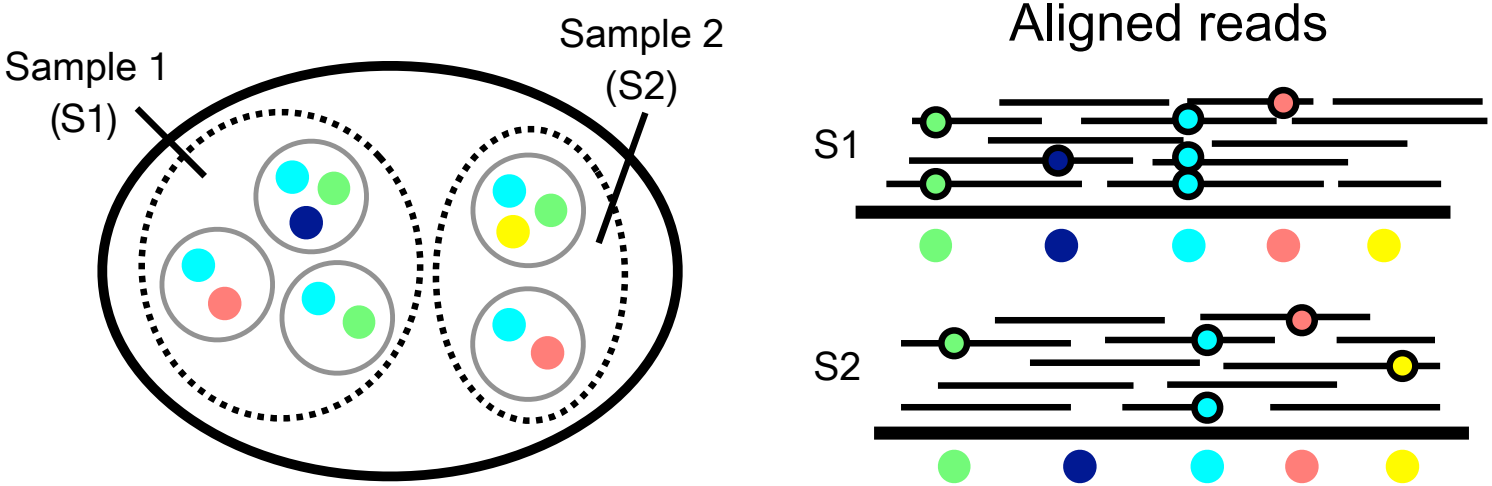
- ddClone (Salehi et al. 2017)
- B-SCITE (Malikic et al. 2018)

and many more....

Bulk sequencing data



Bulk sequencing data

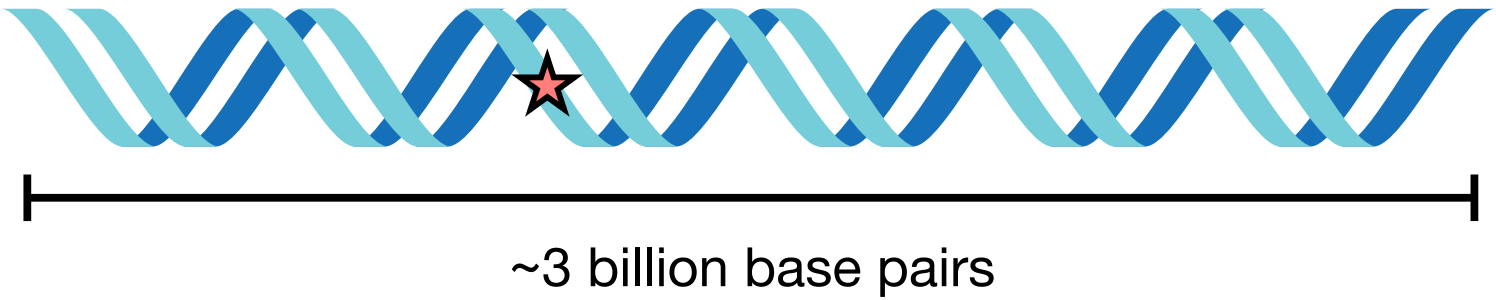


$$\begin{array}{c} \\ \text{S1} \\ \text{S2} \end{array} \begin{array}{c} \begin{array}{ccccc} \color{cyan}{\bullet} & \color{blue}{\bullet} & \color{green}{\bullet} & \color{red}{\bullet} & \color{yellow}{\bullet} \end{array} \\ \begin{bmatrix} 0.5 & 0.17 & 0.33 & 0.17 & 0 \\ 0.5 & 0 & 0.25 & 0.25 & 0.25 \end{bmatrix} \end{array}$$

VAF matrix F

(# variant reads / # total reads)

ISA

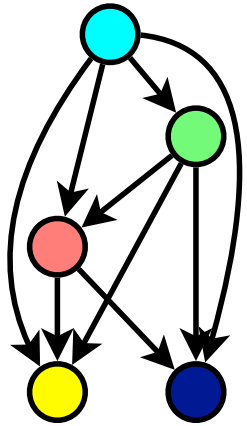


Infinite Sites Assumption (Kimura 1969)

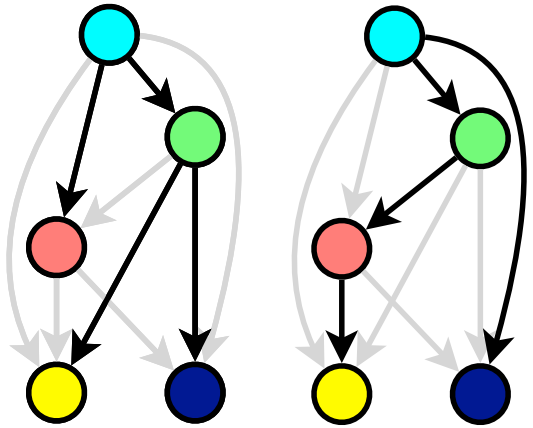
No position in the genome mutates more than once.

AncesTree (El-Kebir et al. 2015)

VAF matrix F



Ancestry graph (AG)

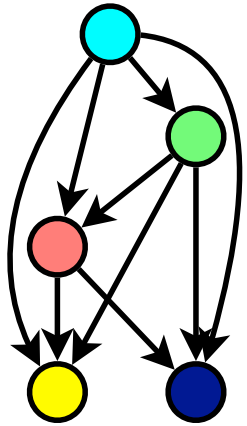


Clonal trees

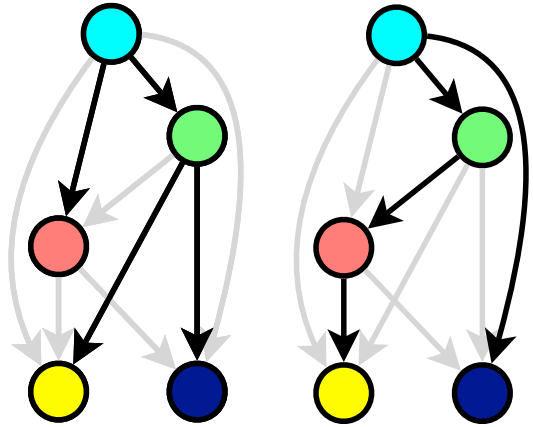
AncesTree (El-Kebir et al. 2015)

$$\begin{array}{c} \\ s1 \\ s2 \end{array}\begin{bmatrix} \color{cyan}{\bullet} & \color{blue}{\bullet} & \color{green}{\bullet} & \color{red}{\bullet} & \color{yellow}{\bullet} \\ 0.5 & 0.17 & 0.33 & 0.17 & 0 \\ 0.5 & 0 & 0.25 & 0.25 & 0.25 \end{bmatrix}$$

VAF matrix F



Ancestry graph (AG)



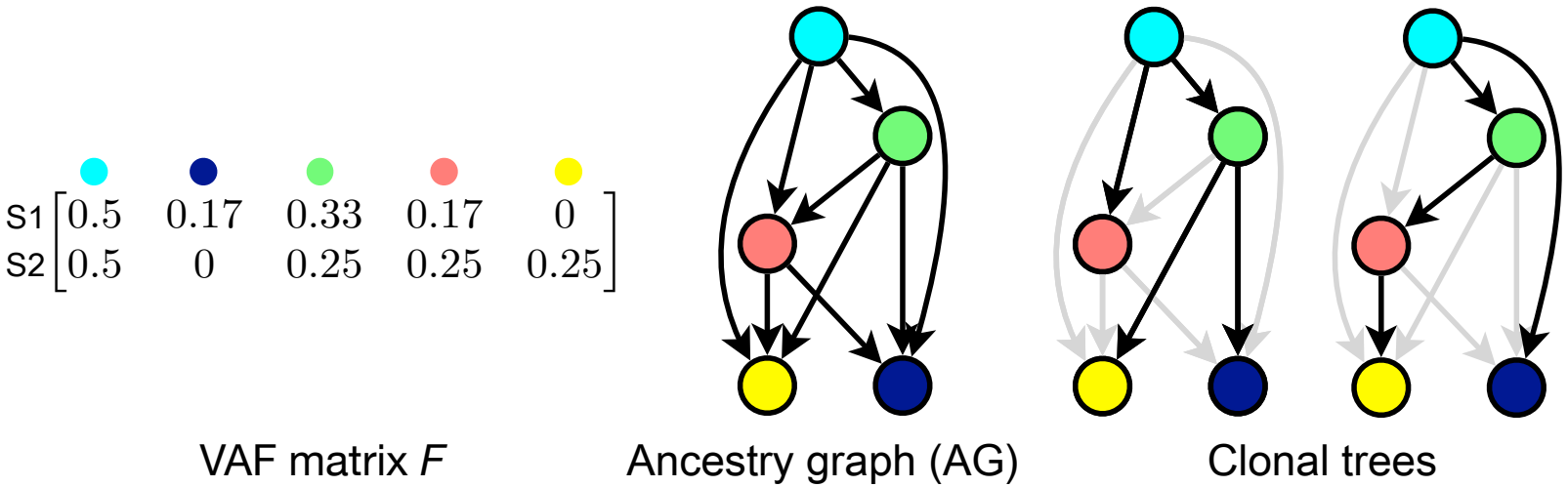
Clonal trees

Observation

Possible clonal trees $\equiv$ AG spanning trees satisfying the *sum condition*:

$$F_{ij} \geq \sum_{k \text{ child of } j} F_{ik} \qquad \forall i \in \{1, \ldots, s\}.$$

AncesTree (El-Kebir et al. 2015)



VAF matrix F | Ancestry graph (AG) | Clonal trees

Variant Allele Frequency Factorization Problem (VAFFP)

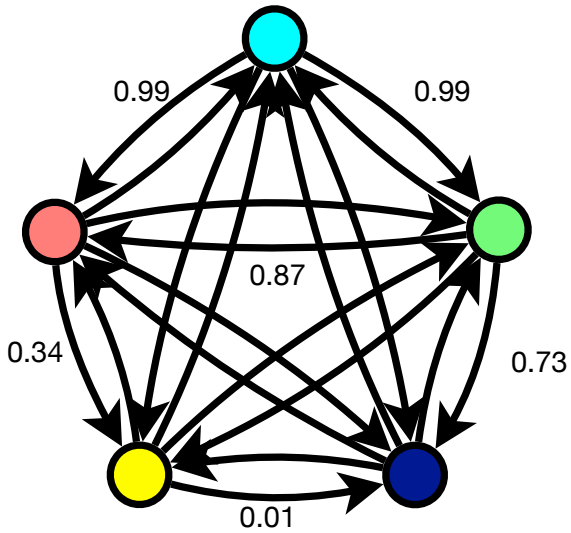
Given: VAF matrix F.

Find: Usage matrix U and clonal matrix B such that

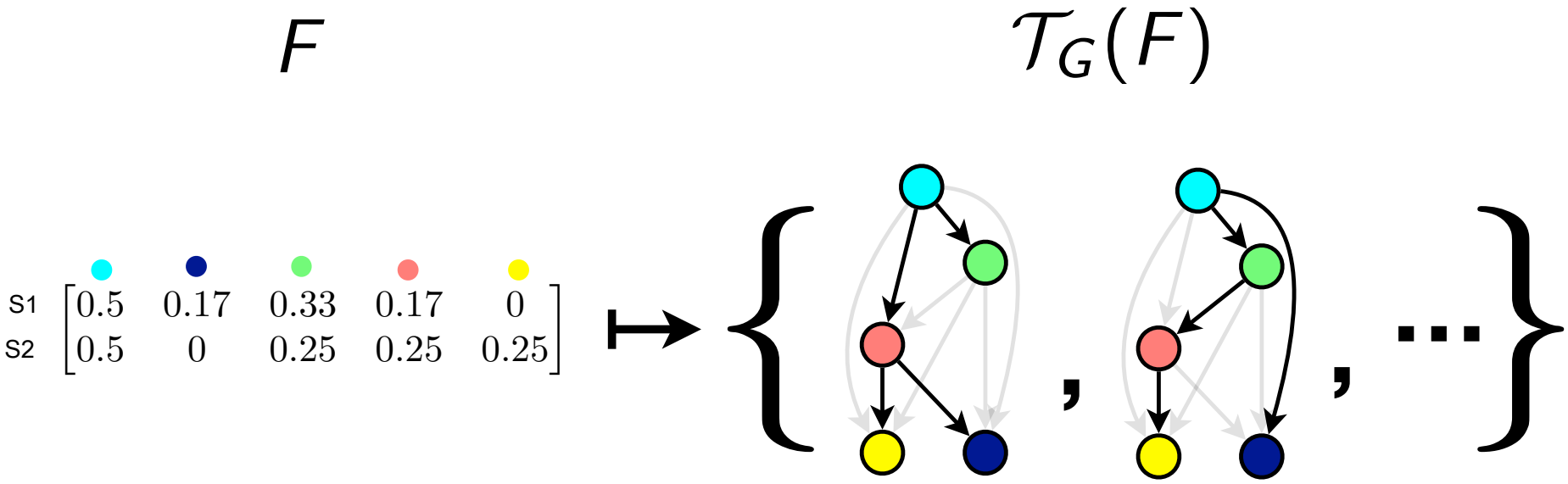
$$F = \frac{1}{2}UB.$$

Outline

1. Background
2. Methods
   - Enumeration VAFFP
   - Noise in sequencing data
   - Handling noise
   - Shrinking the search space
3. Results



E-VAFFP



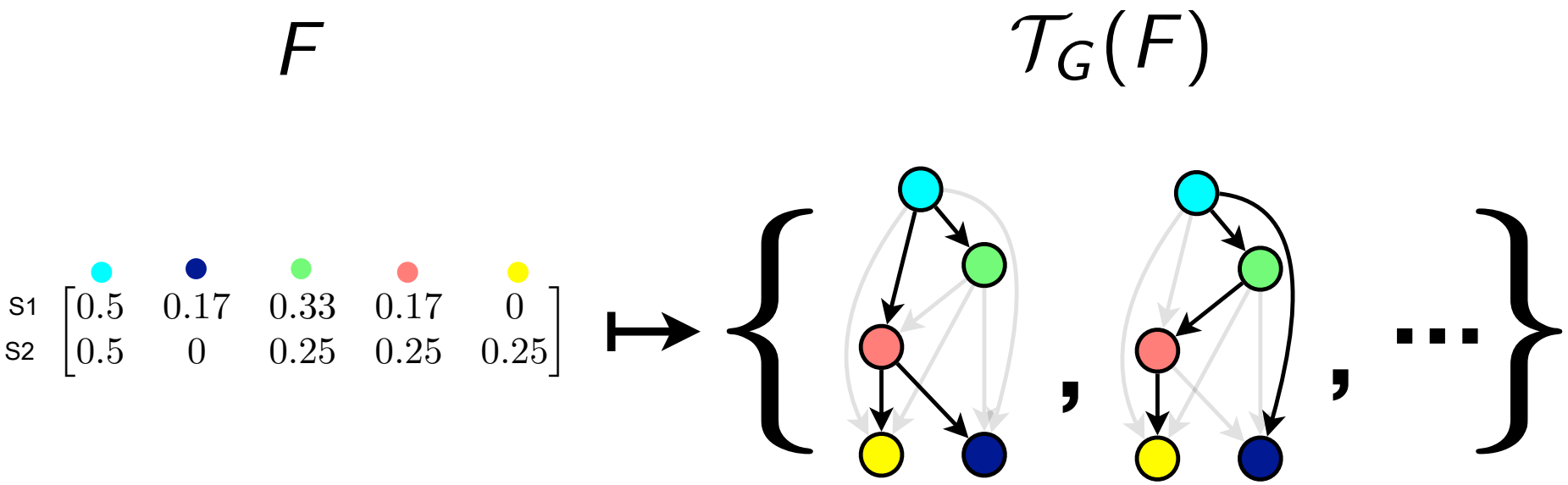
Enumeration VAFFP

Given: VAF matrix F.

Find: The set $\mathcal{T}(G_F)$ of *all* ancestry graph spanning trees that satisfy the sum condition.

How: Modified version of (Gabow and Myers 1978)

E-VAFFP



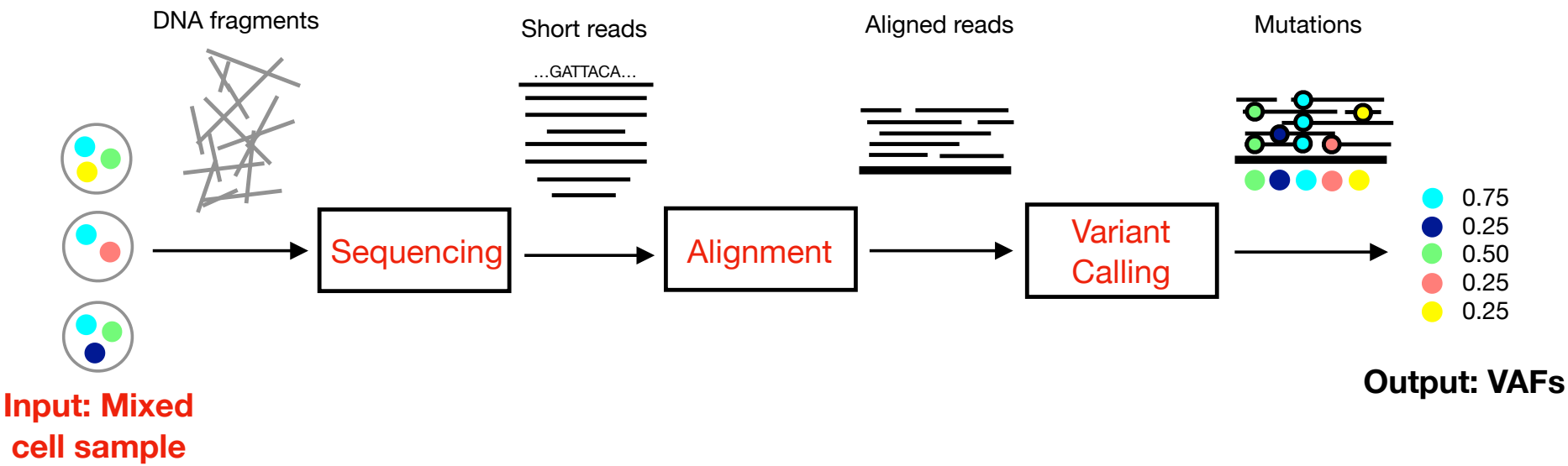
Enumeration VAFFP (strict)

Given: VAF matrix F.

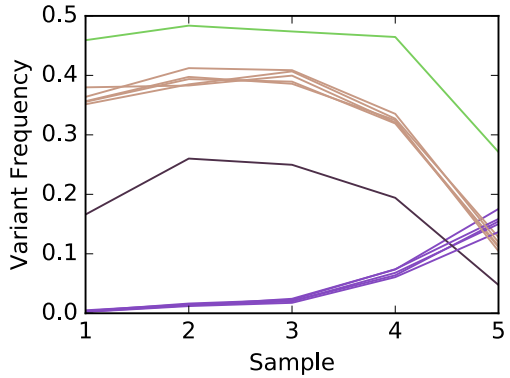
Find: The set $\mathcal{T}(G_F)$ of *all* ancestry graph spanning trees that satisfy the sum condition.

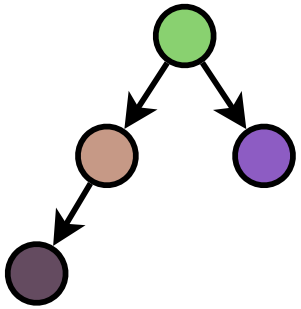
How: Modified version of (Gabow and Myers 1978)

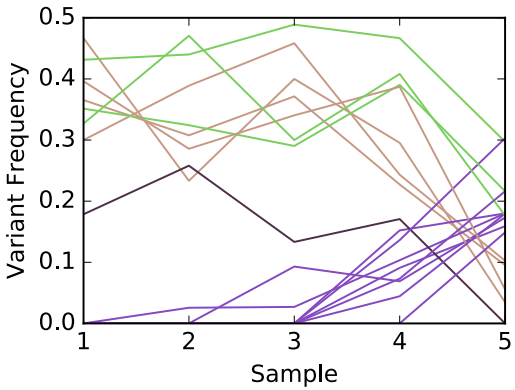
Sources of noise

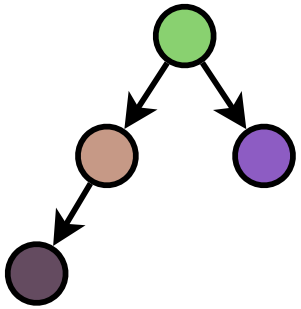


Relaxed sum condition



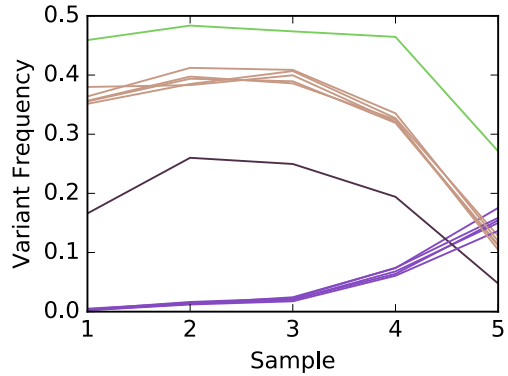


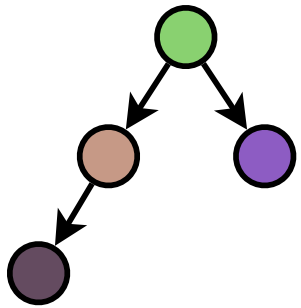


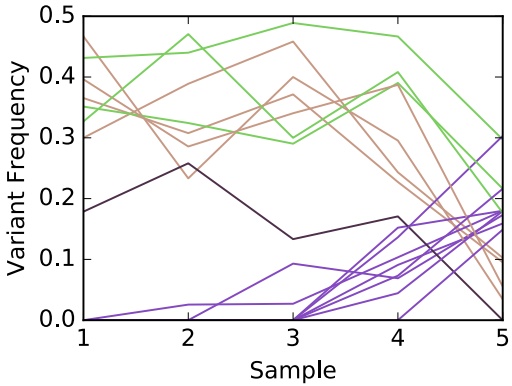


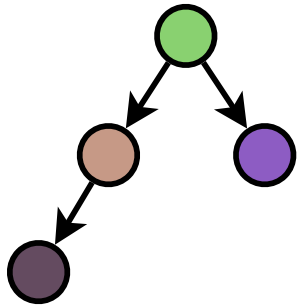
$$F_{ij} \geq \sum_{k \text{ child of } j} F_{ik} \qquad \forall i \in \{1, \ldots, s\}$$

Relaxed sum condition



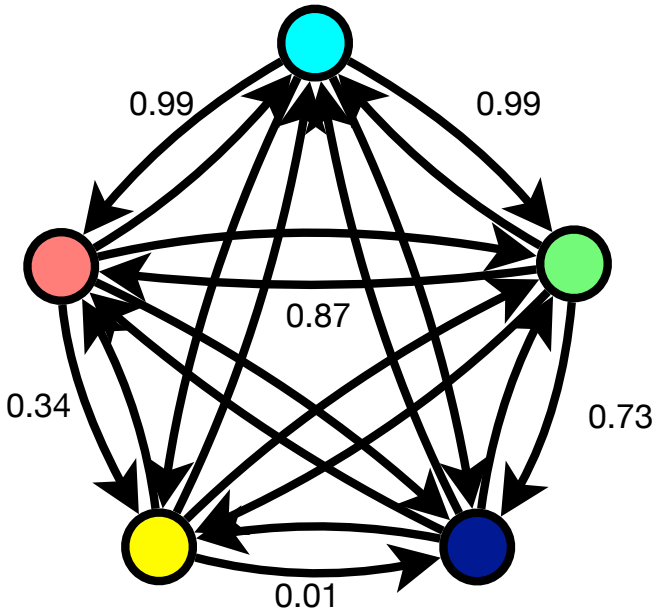






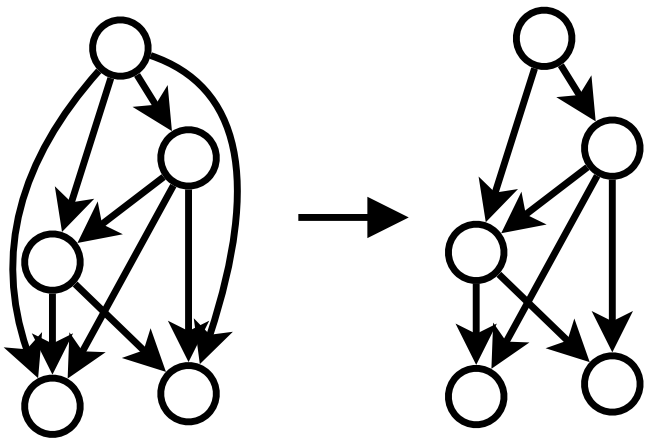
$$F_{ij} \boxed{+\varepsilon} \geq \sum_{k \text{ child of } j} F_{ik} \qquad \forall i \in \{1, \ldots, s\}$$

Approximate ancestry graph



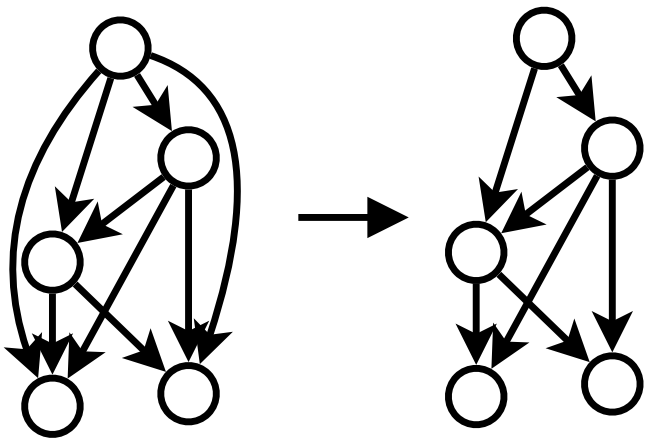
1. Complete weighted digraph
2. Posterior robability of ancestry: beta-binomial model (El-Kebir et al. 2015)
3. Enumerate spanning trees in weight order (Camerini et al. 1980)

Partial transitive reduction

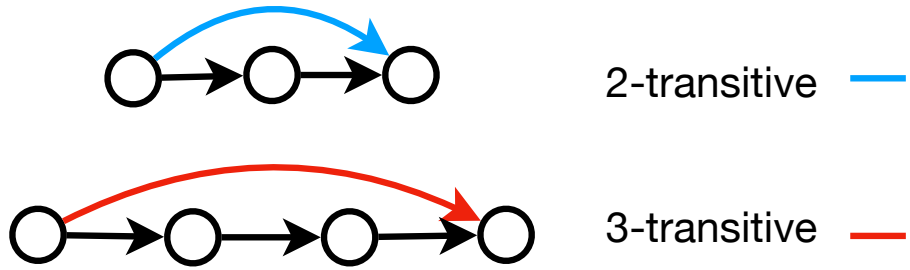


Goal: simplify ancestry graph

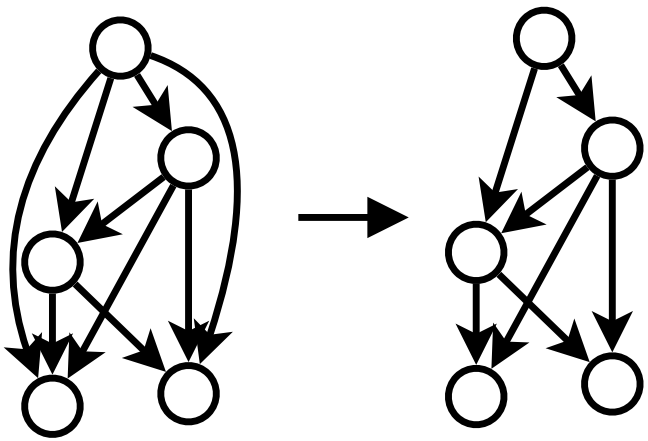
Partial transitive reduction



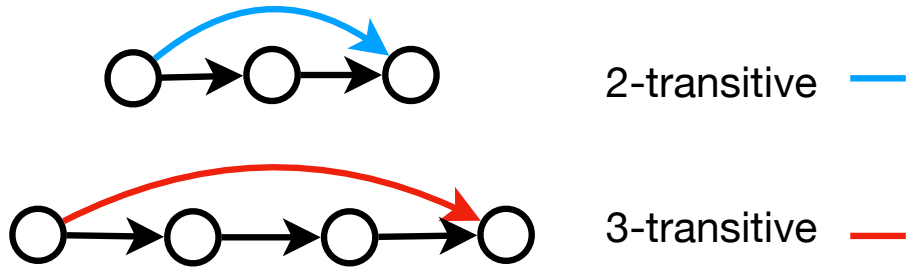
Goal: simplify ancestry graph



Partial transitive reduction



Goal: simplify ancestry graph

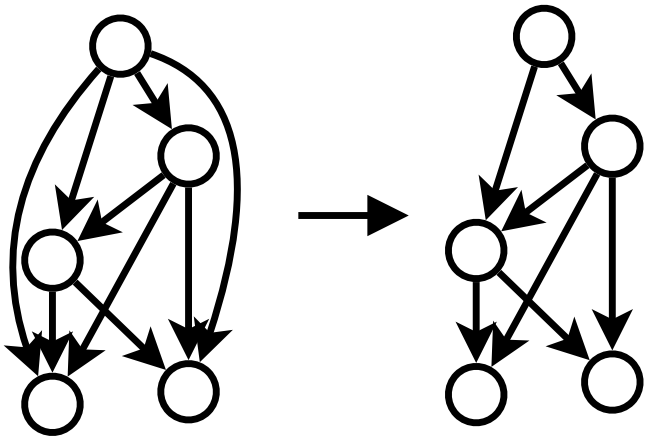


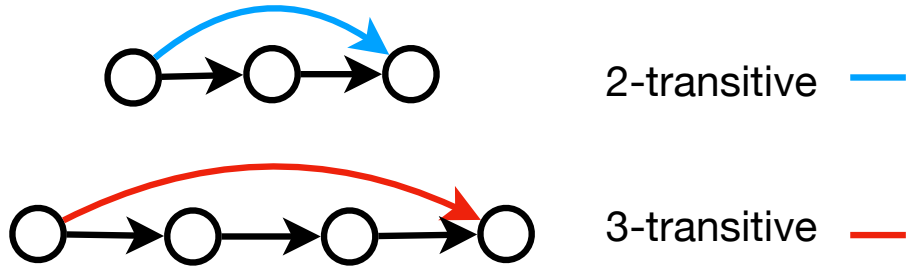
k-PTR

Subgraph resulting from removing all $\geq k$-transitive edges.

Partial transitive reduction

3-PTR



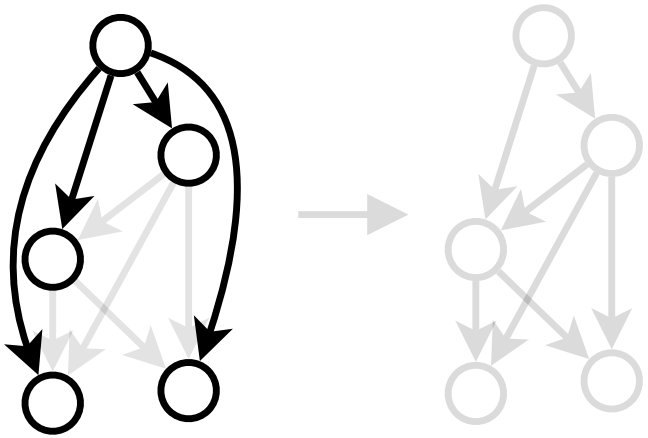


Goal: simplify ancestry graph

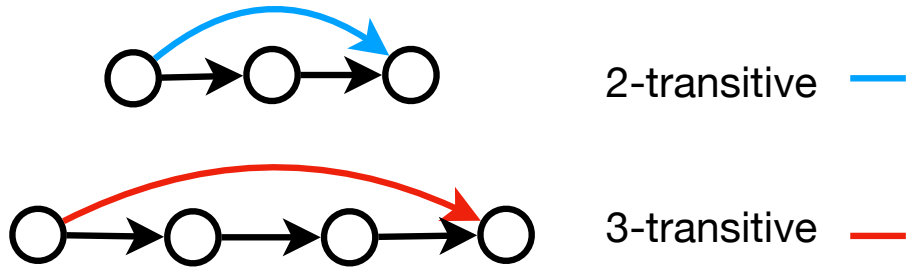
k-PTR

Subgraph resulting from removing all $\geq k$-transitive edges.

Partial transitive reduction



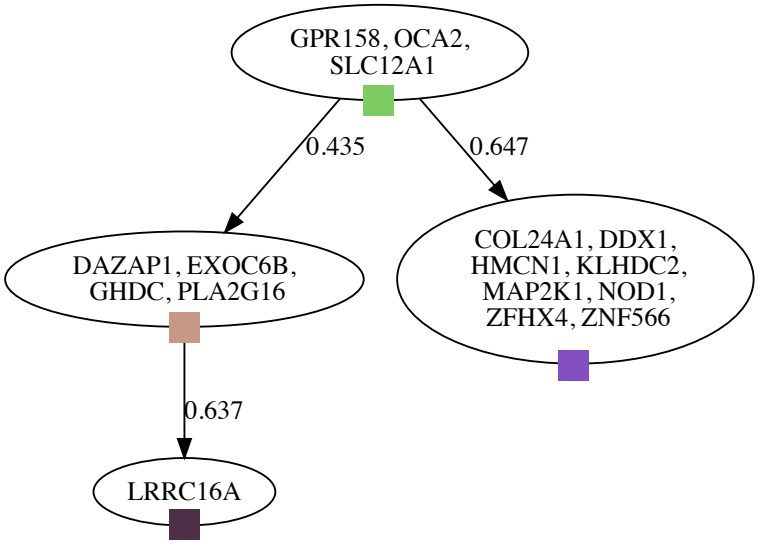
Goal: simplify ancestry graph



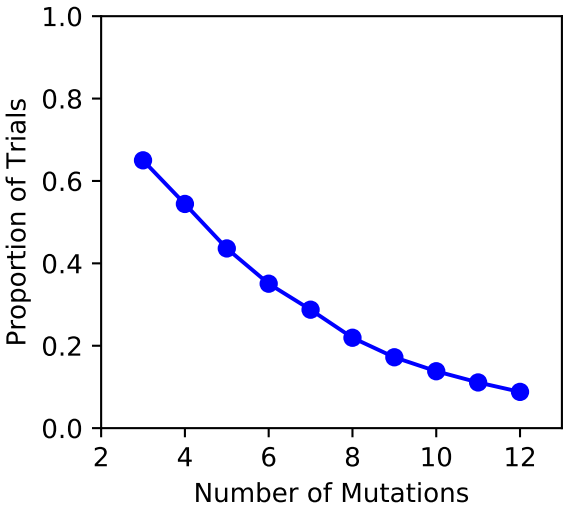
k-PTR

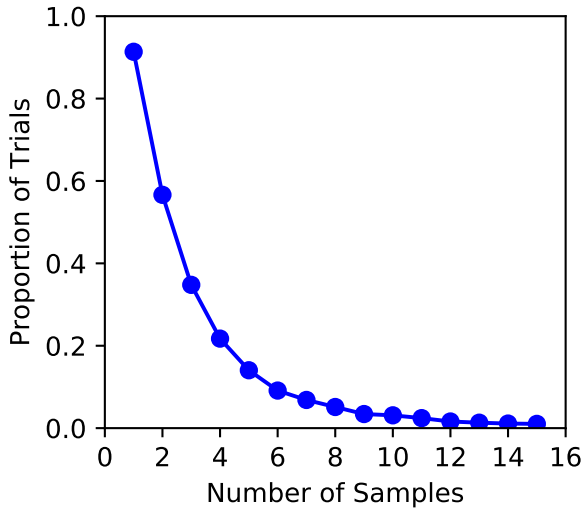
Subgraph resulting from removing all $\geq k$-transitive edges.

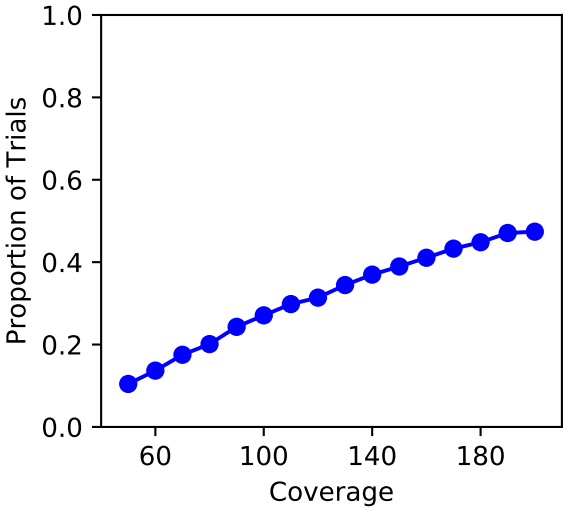
Outline

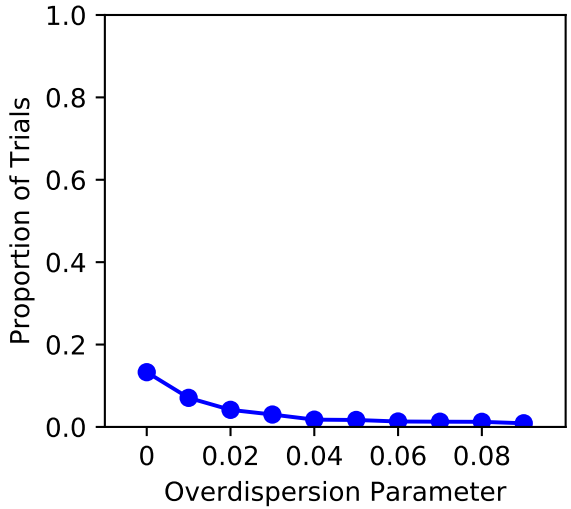


Simulated data: solution existence



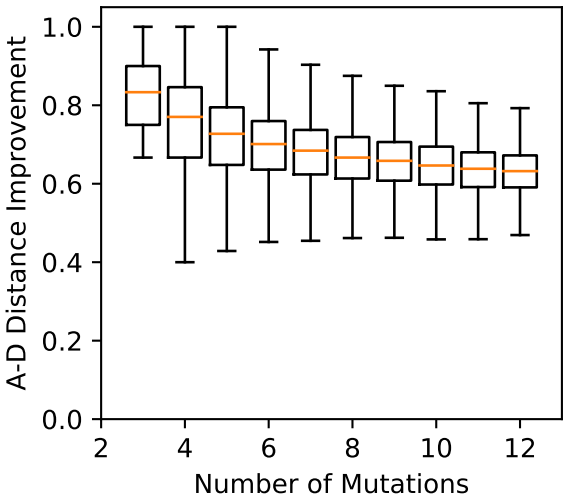


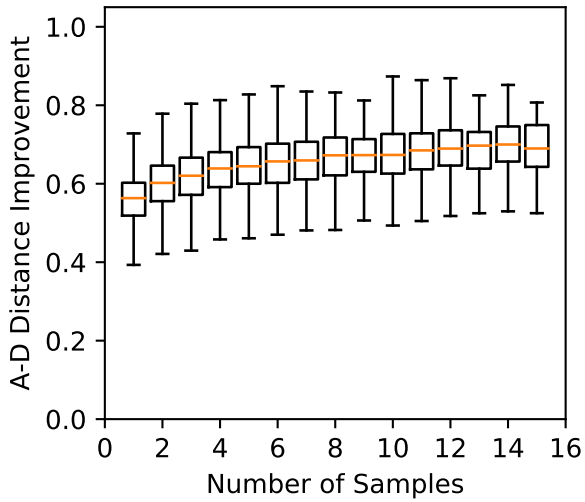


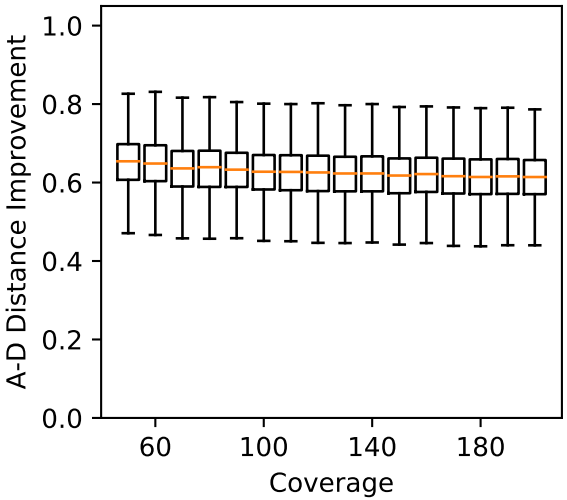


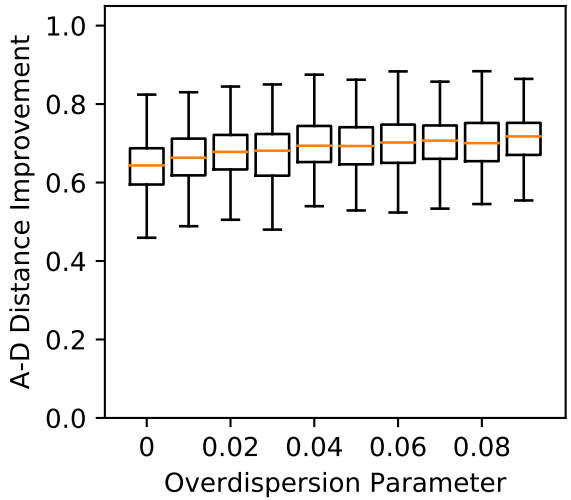
Defaults:
10 mutation clusters
5 samples
60× coverage
No overdispersion

Simulated data: solution quality





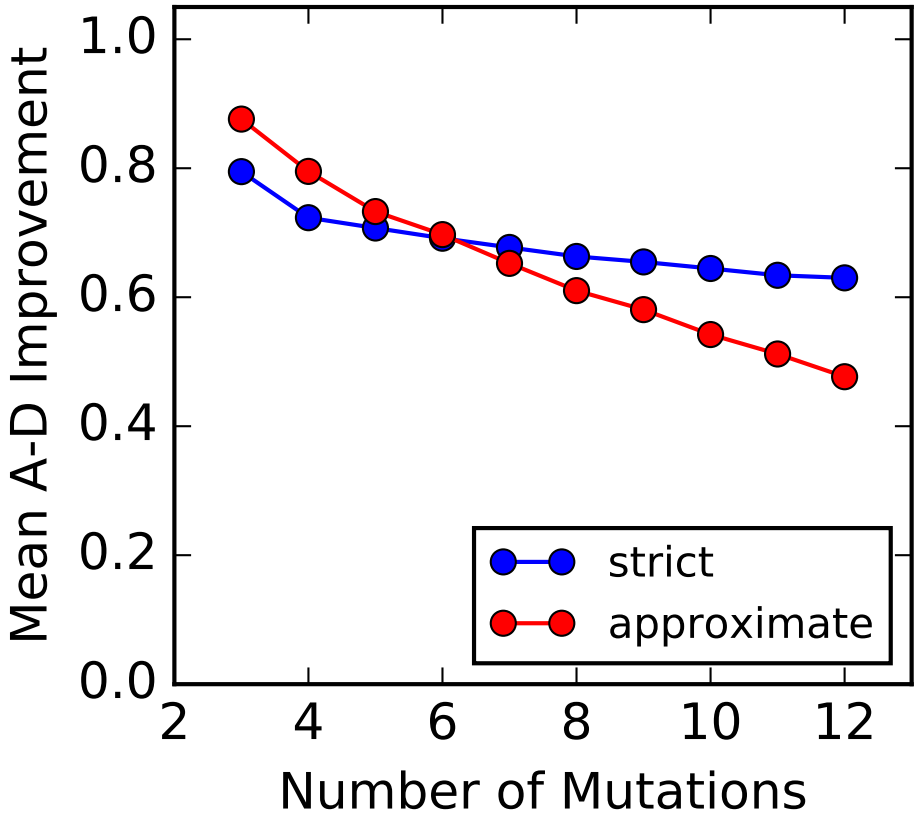


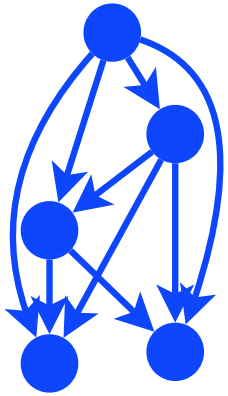


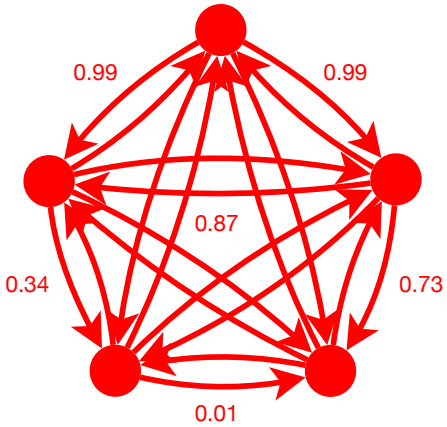
Defaults:
10 mutation clusters
5 samples
60× coverage
No overdispersion

Ancestor-descendant distance (Govek et al. 2018)

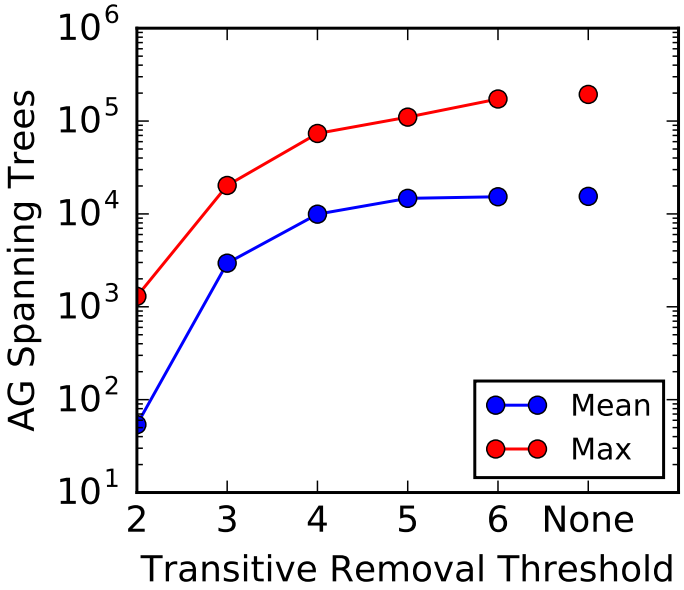
Simulated data: approximate vs strict

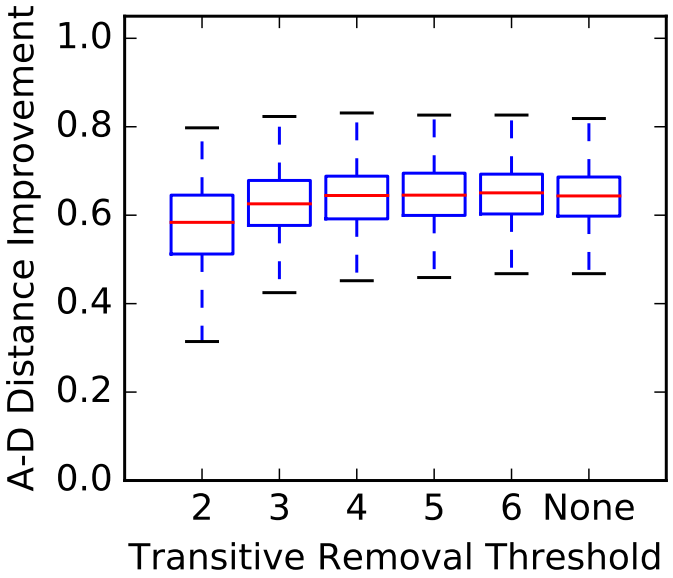


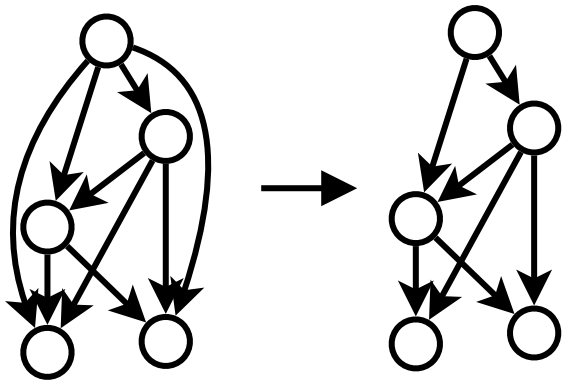




Simulated data: PTR







Real data

Chronic lymphocytic leukemia (Schuh et al. 2012)

- 3 patients (CLL003, CLL006, CLL077)
- 5 samples each, spaced over time
- WGS ($40\times$ coverage) and deep sequencing ($100000\times$ coverage)

Clear cell renal carcinoma (Gerlinger et al. 2014)

- 8 patients (EV003, EV005, EV006, EV007, RK26, RMH002, RMH004, RMH008)
- 5-11 samples from different regions
- Amplicon sequencing ($> 400\times$ coverage)

Real data: strict solution rarity

Patient	**Samples**	**Mutations**[1]	**# Clusters**	$\lvert\mathcal{T}(G_F)\rvert$
CLL003 (deep)	5	15/20	4	0
CLL003 (WGS)	5	13/30	4	0
CLL006 (deep)	5	5/10	5	2
CLL006 (WGS)	5	6/16	5	0
CLL077 (deep)	5	12/16	4	1
CLL077 (WGS)	5	16/20	4	0
EV003	8	12/16	4, 5, 6	0
EV005	7	61/64	5, 6	0
EV006	9	52/57	5	0
EV007	8	54/56	4, 5	0
RK26	11	62/62	4, 5, 6	0
RMH002	5	48/48	5, 6	0
RMH004	6	126/126	5, 6	0
RMH008	8	69/71	5, 6	0

[1] After/before filtering out mutations with VAF above 0.5.

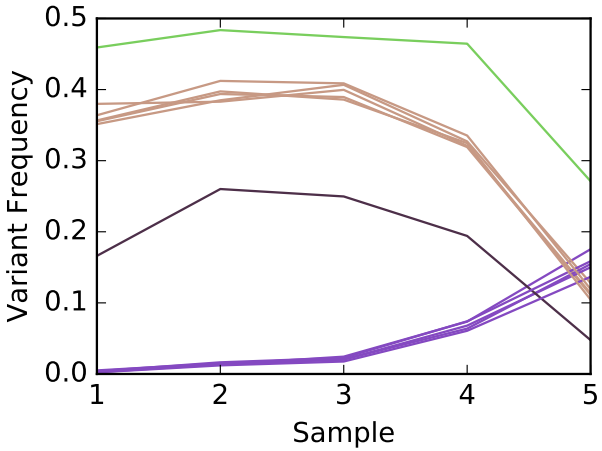
Real data: strict solution rarity

Patient	Samples	Mutations[1]	# Clusters	$\lvert\mathcal{T}(G_F)\rvert$
CLL003 (deep)	5	15/20	4	0
CLL003 (WGS)	5	13/30	4	0
CLL006 (deep)	5	5/10	5	2
CLL006 (WGS)	5	6/16	5	0
CLL077 (deep)	5	12/16	4	1
CLL077 (WGS)	5	16/20	4	0
EV003	8	12/16	4, 5, 6	0
EV005	7	61/64	5, 6	0
EV006	9	52/57	5	0
EV007	8	54/56	4, 5	0
RK26	11	62/62	4, 5, 6	0
RMH002	5	48/48	5, 6	0
RMH004	6	126/126	5, 6	0
RMH008	8	69/71	5, 6	0

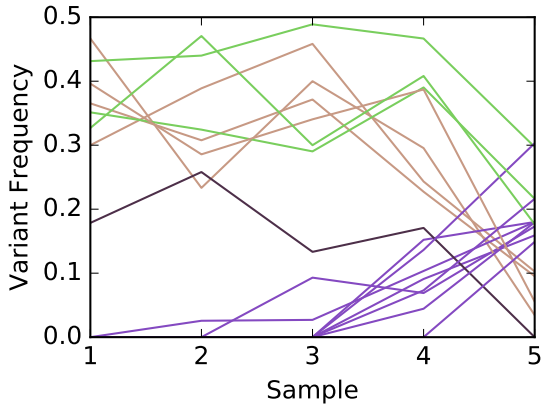
[1] After/before filtering out mutations with VAF above 0.5.

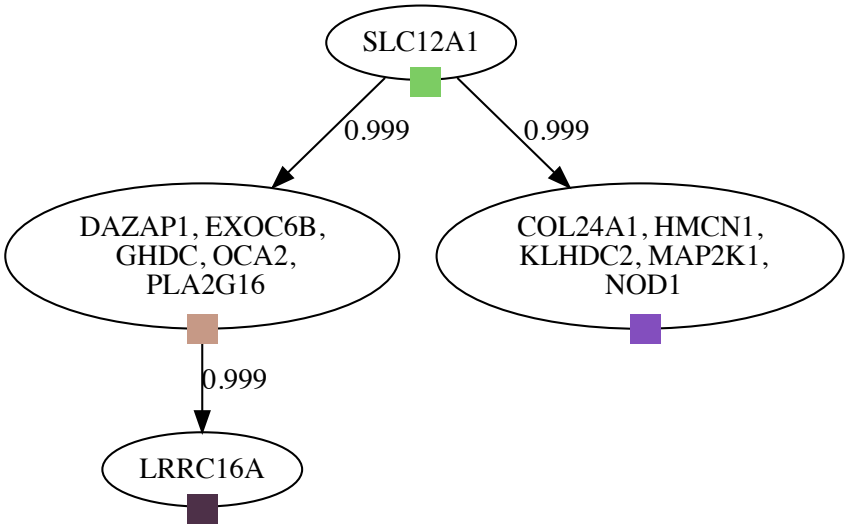
CLL077

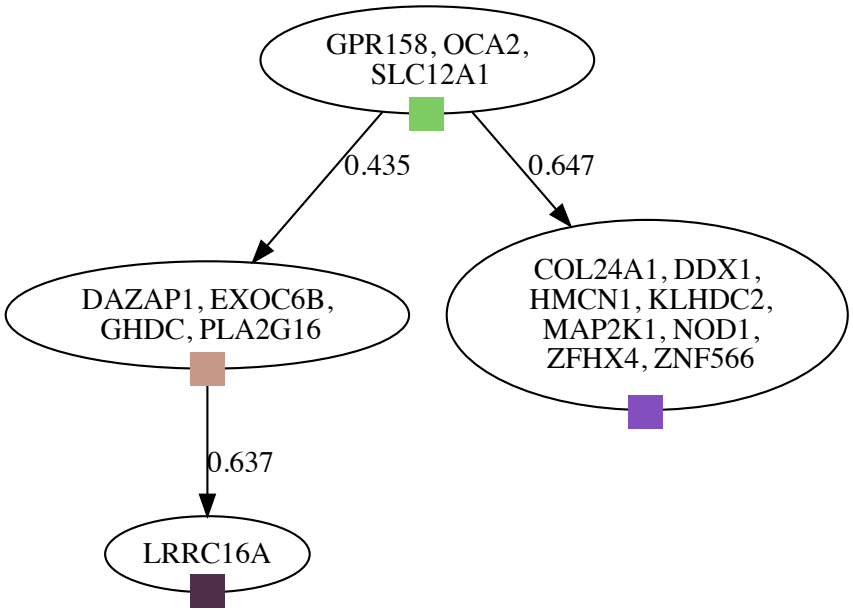
100000× coverage



40× coverage

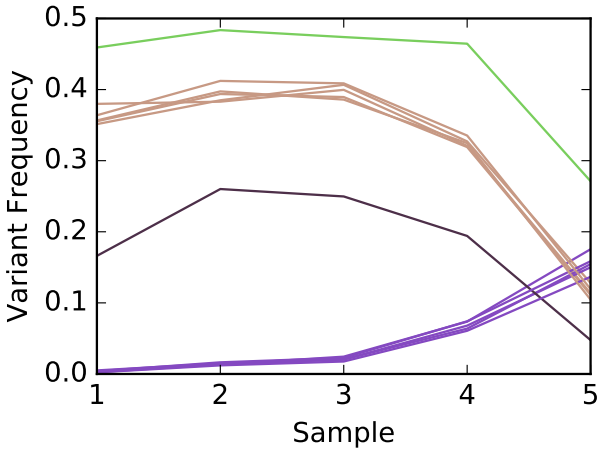




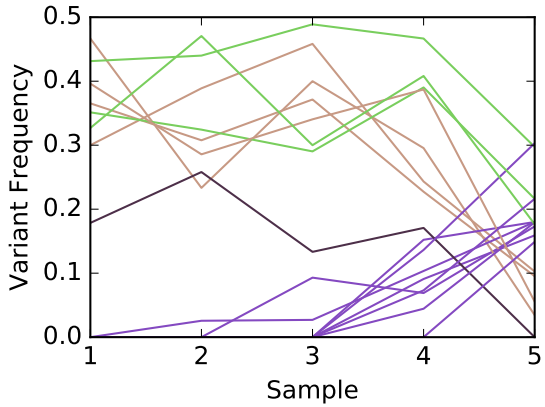


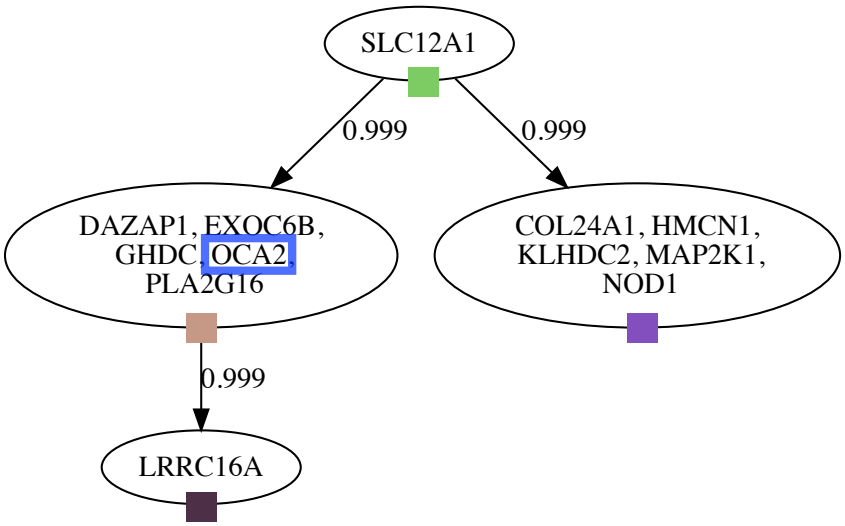
CLL077

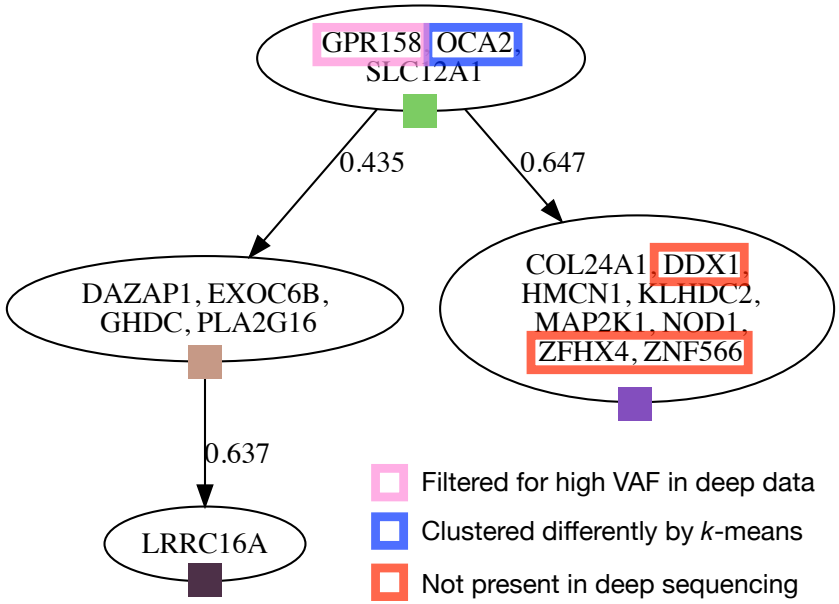
100000× coverage



40× coverage

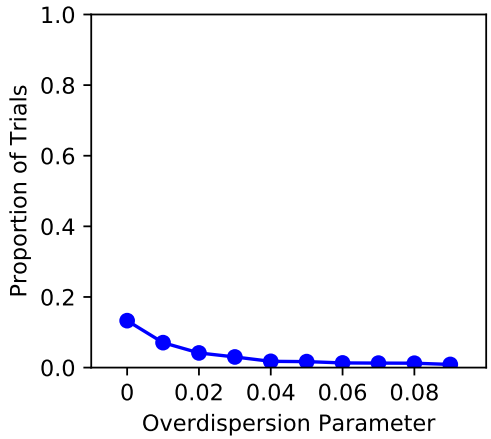






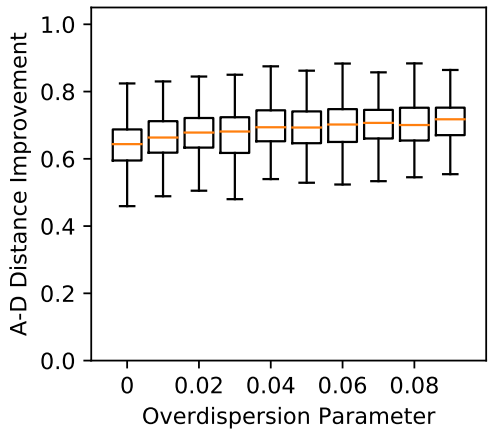
Conclusions

1. Strict ISA-based trees are rare in simulated and real data
2. Overdispersion makes solutions rarer, but not worse
3. Approximate AG and relaxed sum condition increase robustness
4. PTR simplifies AG with minor quality impact (skews topology)
5. Approximate AG outperforms strict for few mutations and vice versa



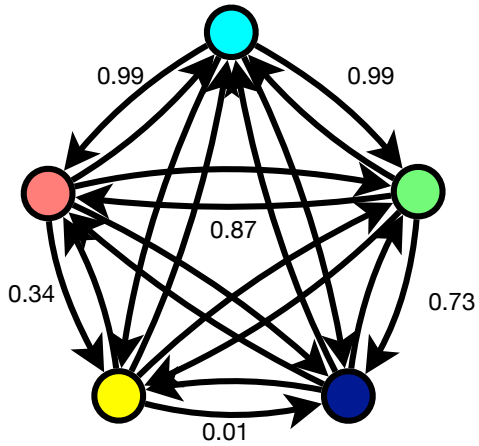
Conclusions

1. Strict ISA-based trees are rare in simulated and real data
2. Overdispersion makes solutions rarer, but not worse
3. Approximate AG and relaxed sum condition increase robustness
4. PTR simplifies AG with minor quality impact (skews topology)
5. Approximate AG outperforms strict for few mutations and vice versa



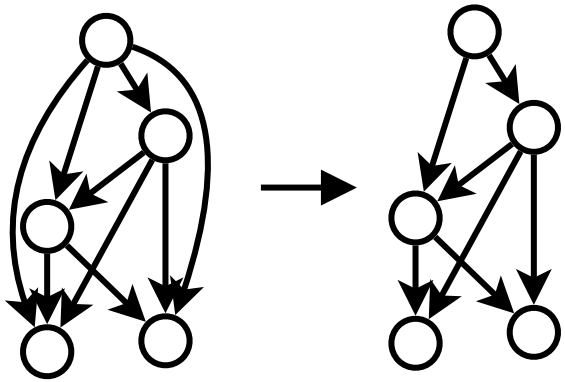
Conclusions

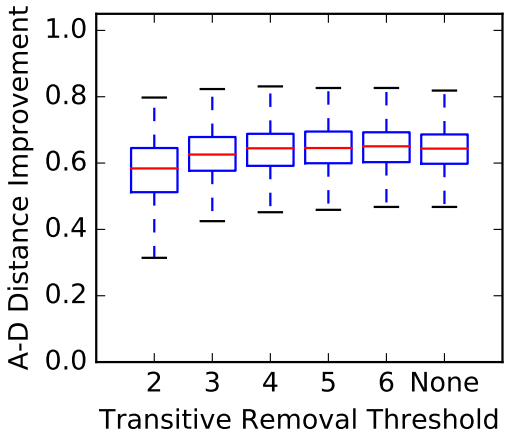
1. Strict ISA-based trees are rare in simulated and real data
2. Overdispersion makes solutions rarer, but not worse
3. Approximate AG and relaxed sum condition increase robustness
4. PTR simplifies AG with minor quality impact (skews topology)
5. Approximate AG outperforms strict for few mutations and vice versa



Conclusions

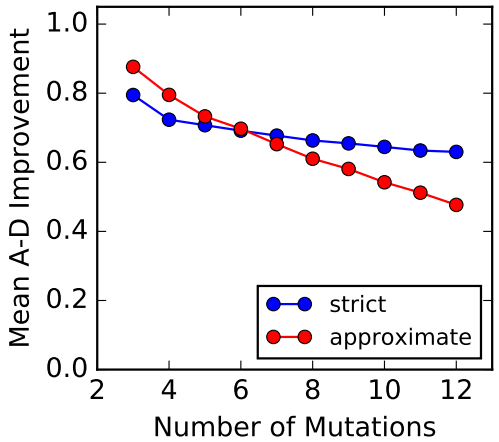
1. Strict ISA-based trees are rare in simulated and real data
2. Overdispersion makes solutions rarer, but not worse
3. Approximate AG and relaxed sum condition increase robustness
4. PTR simplifies AG with minor quality impact (skews topology)
5. Approximate AG outperforms strict for few mutations and vice versa





Conclusions

1. Strict ISA-based trees are rare in simulated and real data
2. Overdispersion makes solutions rarer, but not worse
3. Approximate AG and relaxed sum condition increase robustness
4. PTR simplifies AG with minor quality impact (skews topology)
5. Approximate AG outperforms strict for few mutations and vice versa



Acknowledgment

- This project is supported by NSF CRII award IIS-1657380 and by Elledge, Eugster, and Class of '49 Fellowships from Carleton College (to LO).
- Thanks to Zach DiNardo, Thais Del Rosario Hernandez, and Rosa Zhou for helpful conversations.
- Special thanks to Layla Oesper for her mentorship, support, and feedback.