Statistically Consistent Rooting of Species Trees under the Multi-Species Coalescent Model

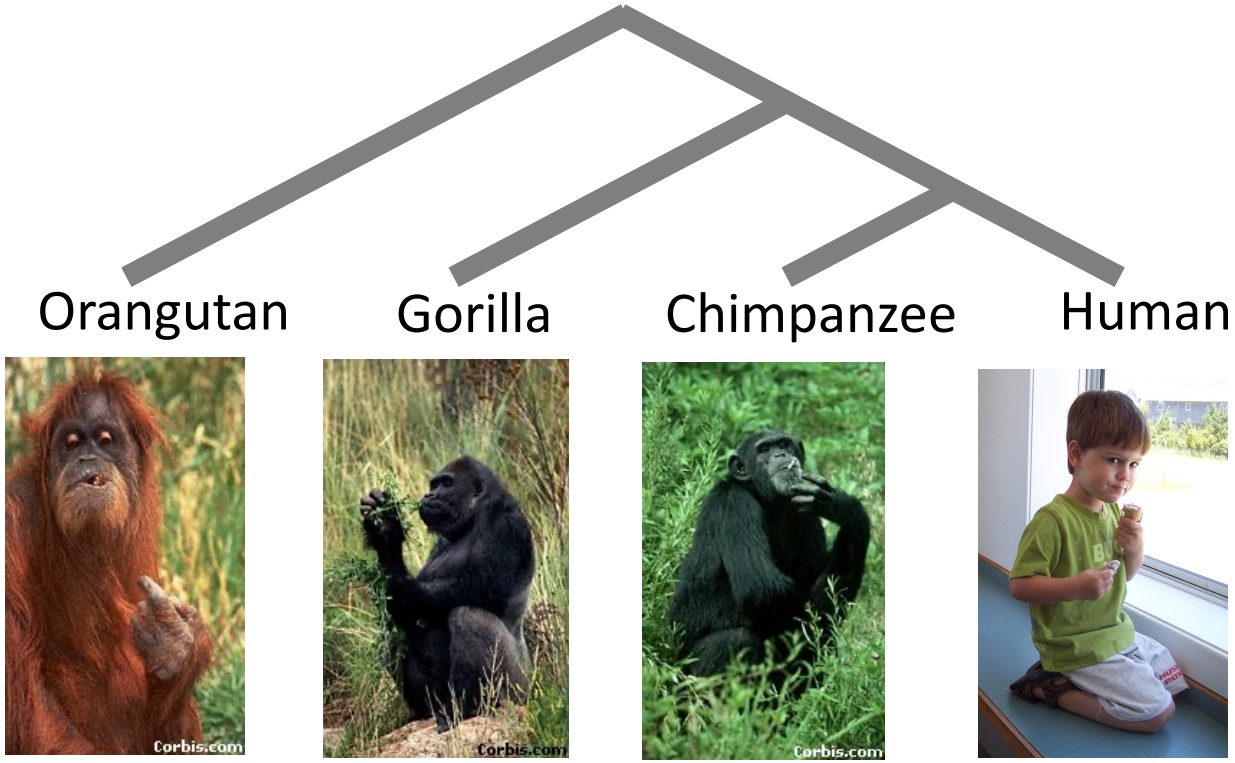
Yasamin Tabatabaee[1], Sebastien Roch[2], and Tandy Warnow[1]

[1] University of Illinois at Urbana-Champaign, [2] University of Wisconsin–Madison

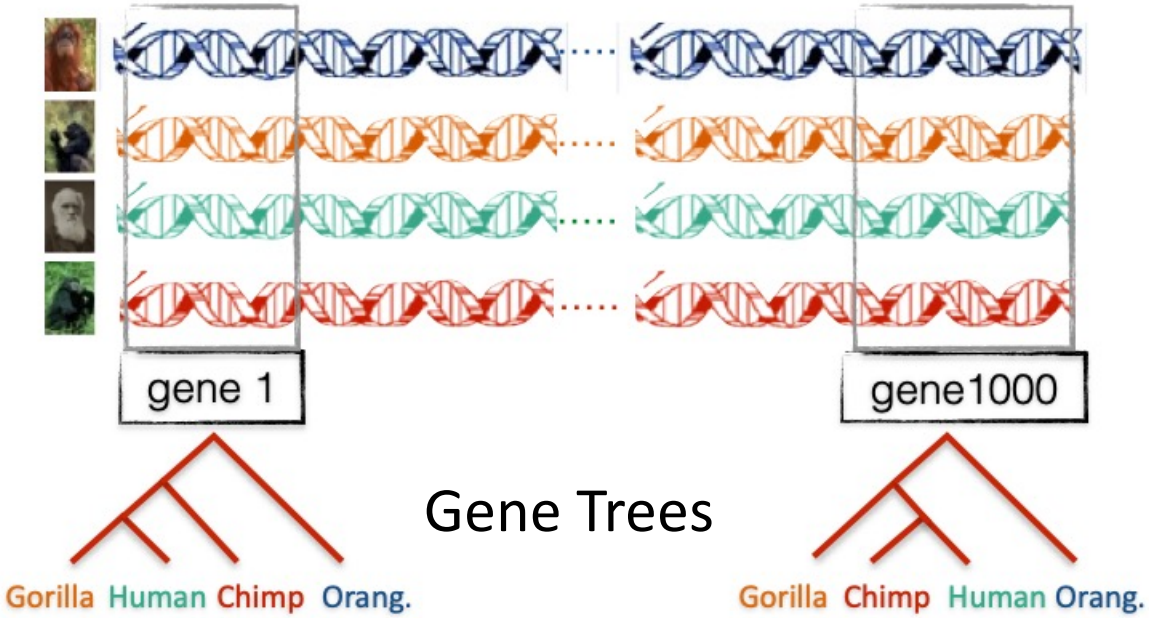
Research in Computational Molecular Biology (RECOMB)

April 17, 2023

Phylogenomics and Gene Tree Discordance



Species Tree



Gene Trees

Image Credit: the *Tree of Life Website, University of Arizona*

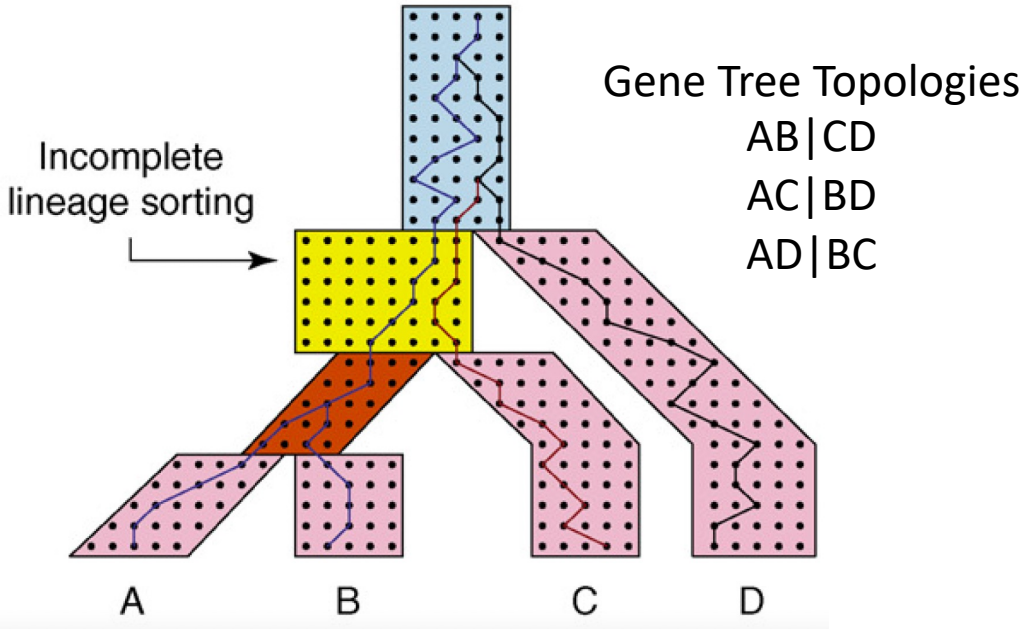
Image Credit: Siavash Mirarab

Gene Tree Discordance and ILS

Causes of gene tree discordance:

- Incomplete lineage sorting (ILS)
- Gene duplication and loss (GDL)
- Horizontal gene transfer (HGT)
- …

Modeled by the Multi-Species Coalescent (MSC) model

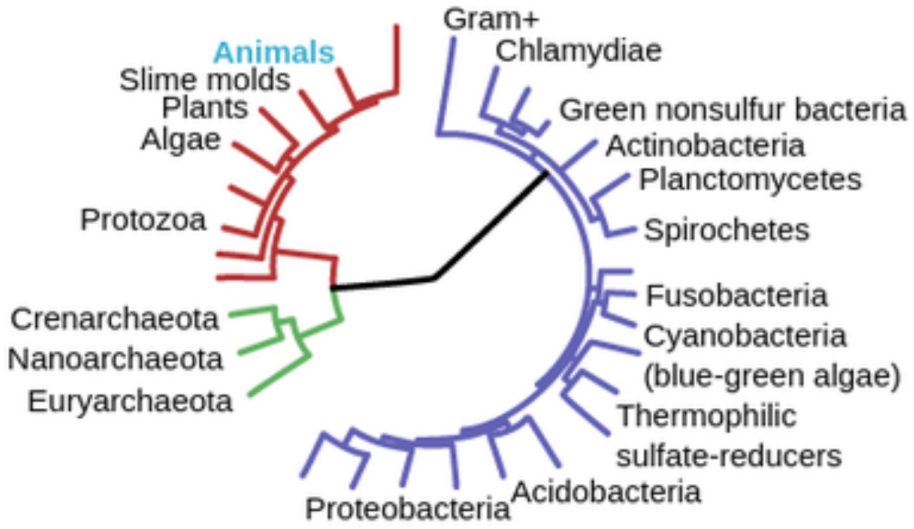


The model species tree defines a probability distribution on the gene tree topologies

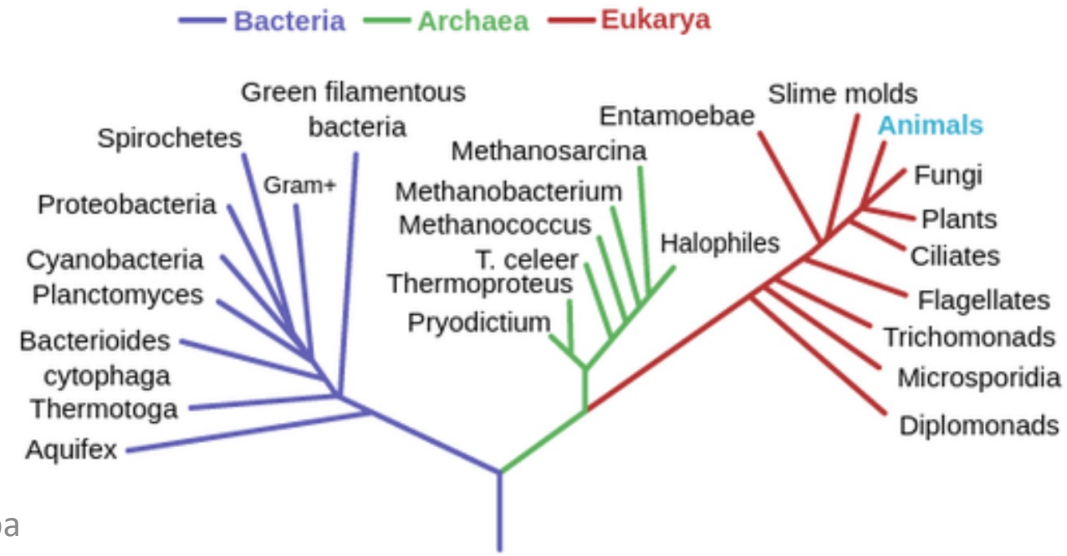
Image Credit: Degnan and Rosenberg, 2009, Trends in Ecology and Evolution

Why Rooting Species Trees?

- Multiple applications throughout biology
- Understanding
  - Adaptation
  - Biodiversity
  - Comparative genomics
  - Dating
- Most species tree estimation methods produce ***unrooted*** trees



Unrooted species tree



Rooted species tree

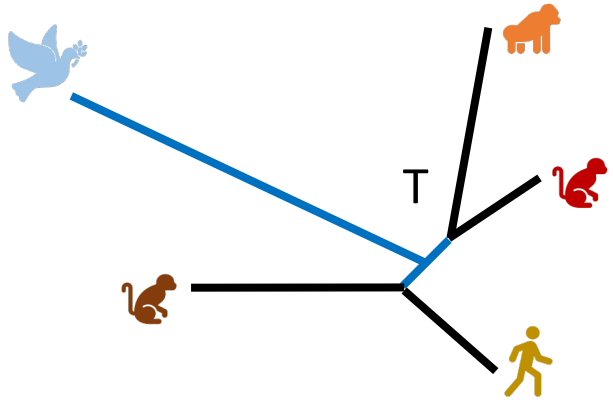
Image Credit: https://theory.labster.com/phylogenetic-tree/, modification of work by Eric Gaba

Current Approaches for Rooting Species Trees

Problem: Find the root position in an unrooted species tree T.

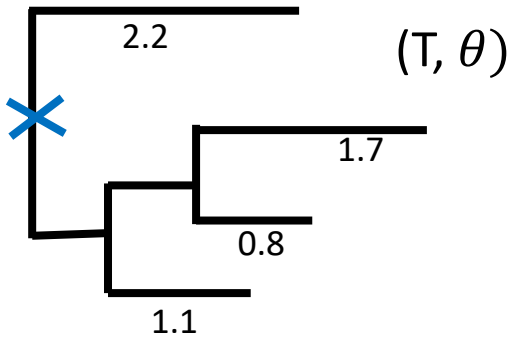
Outgroup Rooting

- Needs prior information about taxa
- Selecting a proper outgroup can be challenging



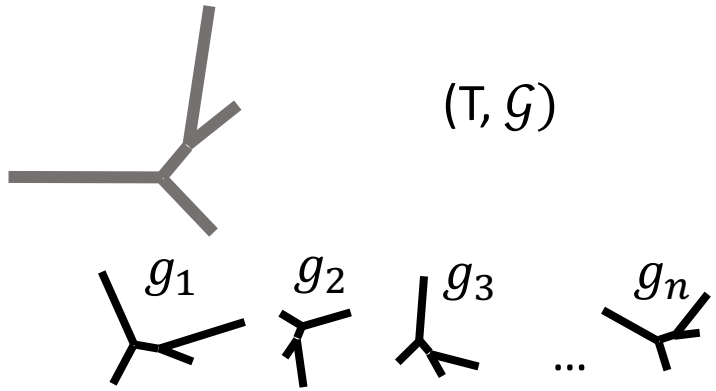
Distance-Based

- Species tree with branch lengths (e.g. Midpoint, MAD, MinVar, …)
- Most are sensitive to deviations from the molecular clock



Gene-Based

- QR (2022): ILS-based
- STRIDE (2017): GDL-based
- Tian & Kubatko (2017): site-based method, clock assumption

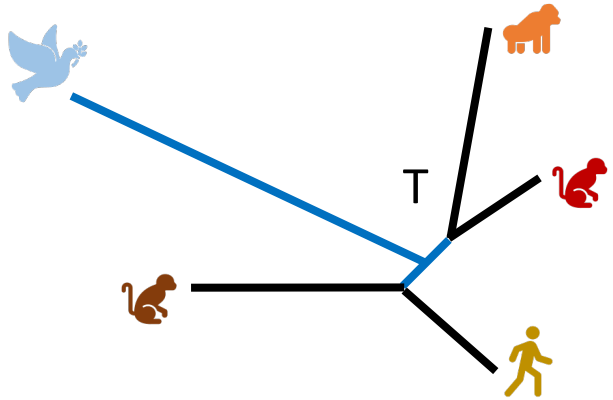


Current Approaches for Rooting Species Trees

Problem: Find the root position in an unrooted species tree T.

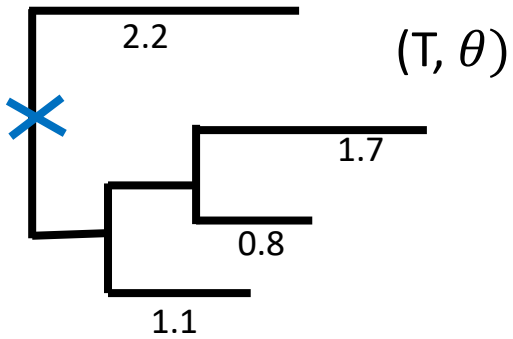
Outgroup Rooting

- Needs prior information about taxa
- Selecting a proper outgroup can be challenging



Distance-Based

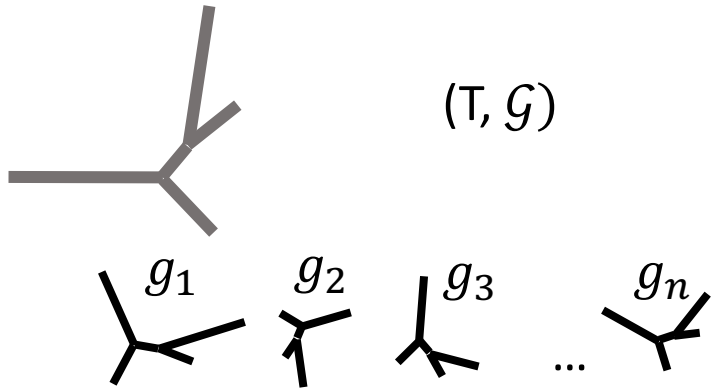
- Species tree with branch lengths (e.g. Midpoint, MAD, MinVar, …)
- Most are sensitive to deviations from the molecular clock



Do not consider sources of gene tree discordance

Gene-Based

- QR (2022): ILS-based
- STRIDE (2017): GDL-based
- Tian & Kubatko (2017): site-based method, clock assumption

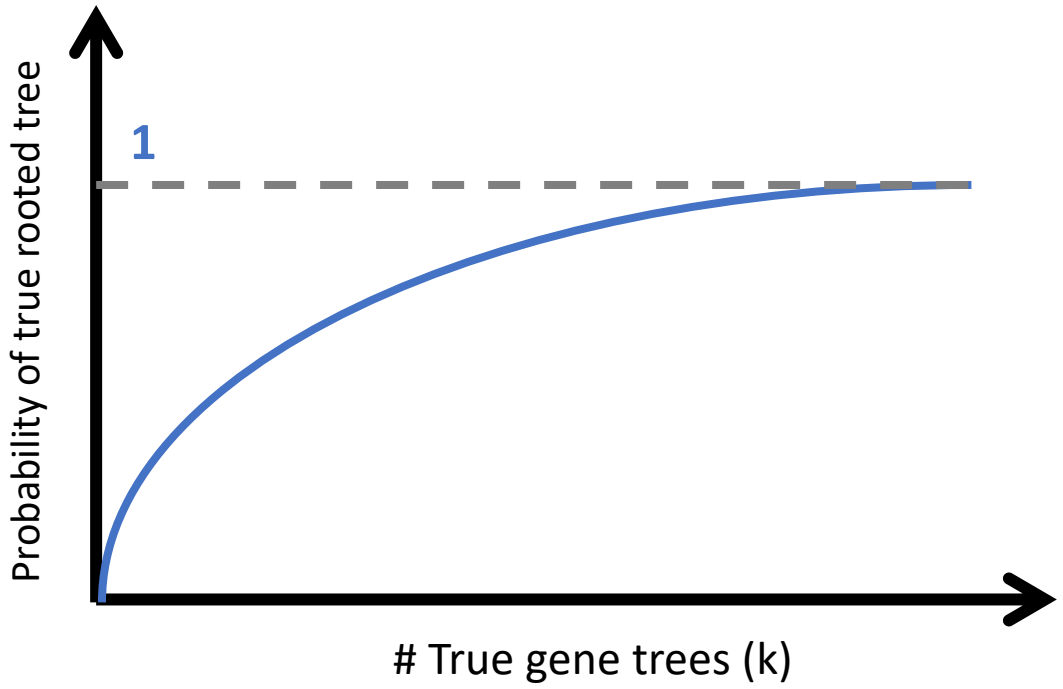


Focus of this talk

Statistical Consistency

An estimation method is statistically consistent under a model, if its output converges to the true parameter as the number of input samples increase.
(based on proof)

- Several methods proven statistically consistent estimators of ***unrooted*** species tree topology under MSC (ASTRAL, SVDQuartets, BUCKy)
- No consistency result for ***rooting*** methods

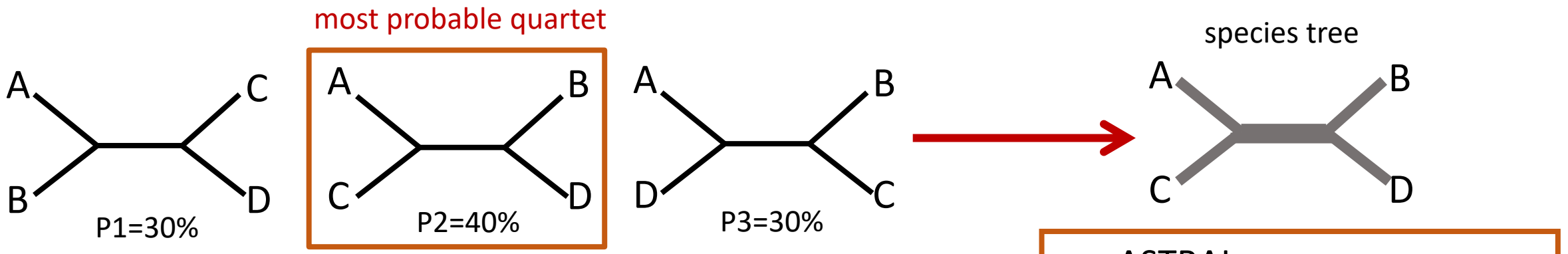


ADR: Identifiability of Unrooted Topology under MSC

***Theorem*: For 4 or more species, the unrooted topology of the species tree is identifiable from the probability distribution of the unrooted gene trees.** [Allman, Degnan and Rhodes (ADR), J. Math. Biol, 2011]

Key property: For 4 species, the most probable unrooted gene tree has the same topology as the unrooted species tree

- Does not hold for more than 4 species



Statistically consistent **quartet-based** species tree estimation methods

ASTRAL [Mirarab et al, 2014]
BUCKy-pop [Larget et al, 2010]
wQFM [Mahbub et al, 2021]
...

ADR: Identifiability of Rooted Topology under MSC

***Theorem*: For 5 or more species, the rooted topology of the species tree is identifiable from the probability distribution of the unrooted gene trees.** [Allman, Degnan and Rhodes (ADR), J. Math. Biol, 2011]

- ADR derive linear invariants and inequalities on the probability distribution of unrooted gene trees.
- They prove that these inequalities and invariants suffice to identify the rooted species tree topology
- Recently used in developing **Quintet Rooting (QR)** (Tabatabaee et al, ISMB'22)

QR: Rooting Species Trees under MSC

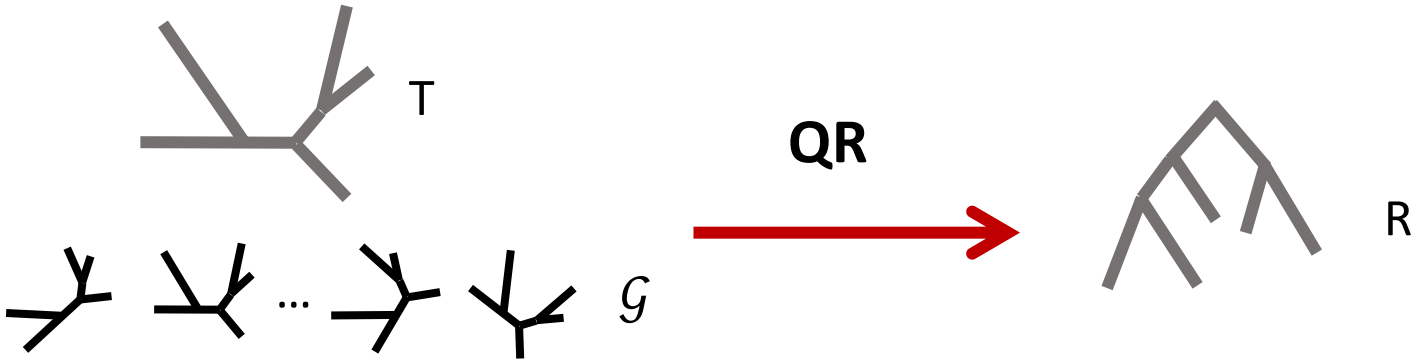
Input

- An unrooted species tree T.
- A set of k unrooted single-copy gene trees $\mathcal{G}$ on $\mathcal{L}(T)$.
- A cost function $Cost(R, \vec{u})$.

Output

- A rooted version of T that minimizes

$$Score(R,T) = \sum_{q \in Q^*} Cost(q, \overrightarrow{u_q})$$

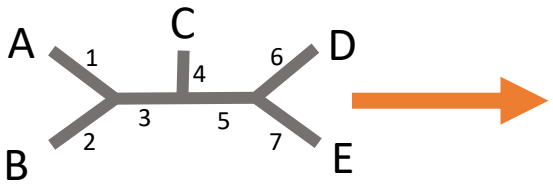


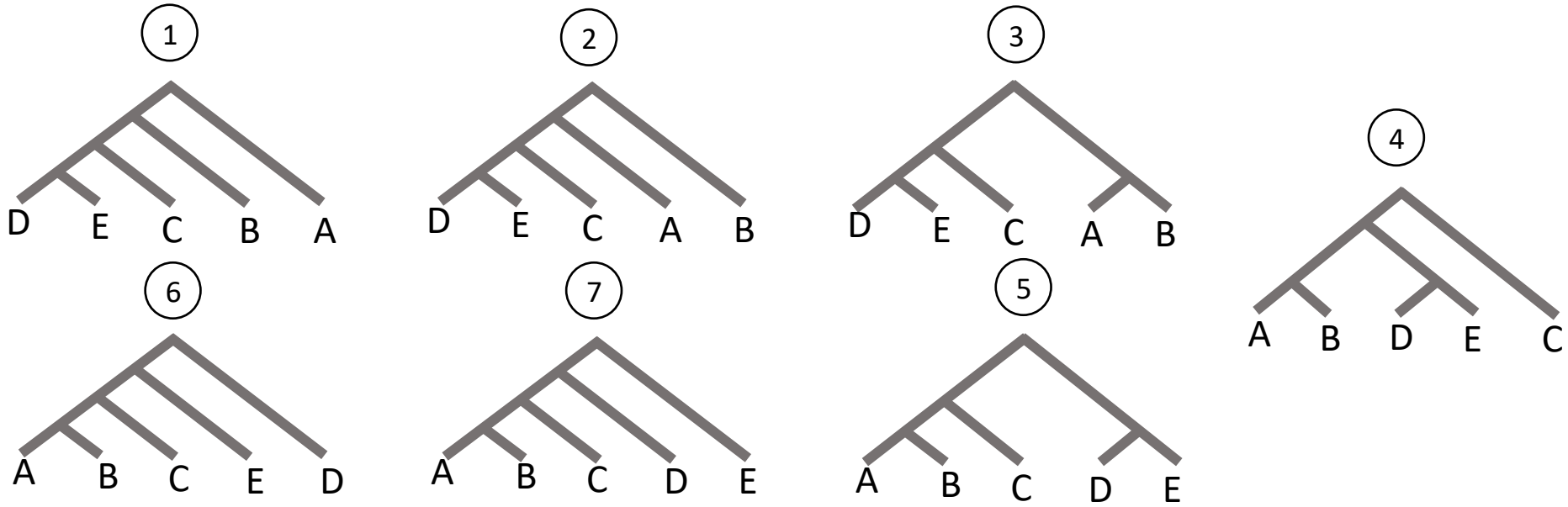
Overview of Results

- QR is not statistically consistent under MSC
  - **Proof outline**: there are two rooted quintet trees where QR cannot distinguish them given finite data (despite identifiability proof)
- New method: QR-STAR
  - **Basic approach**: Modify the QR cost function to include a penalty for the rooted shape + additional step for determining the shape + different weighting
- QR-STAR is statistically consistent under MSC
- QR-STAR has improved accuracy over QR in most model conditions

Properties of Quintet Trees

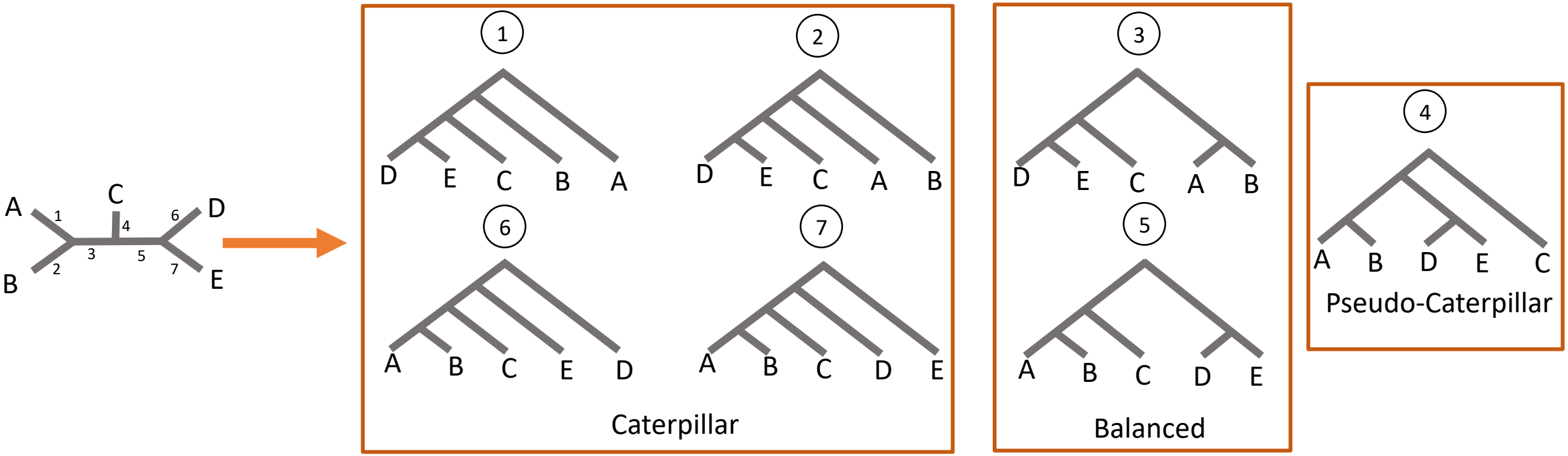
- There are **105** rooted binary trees and **15** unrooted binary trees on 5 taxa
- Each unrooted 5-taxon tree can be rooted on any of its **7** edges





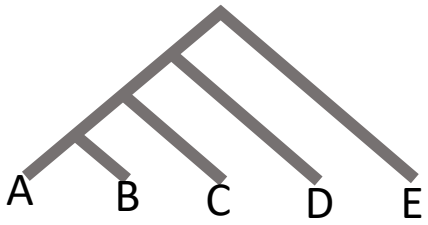
Properties of Quintet Trees

- There are **105** rooted binary trees and **15** unrooted binary trees on 5 taxa
- Each unrooted 5-taxon tree can be rooted on any of its **7** edges
- Rooted 5-taxon trees fall into **three** different shapes: caterpillar, balanced and pseudo-caterpillar [Rosenberg, 2007]

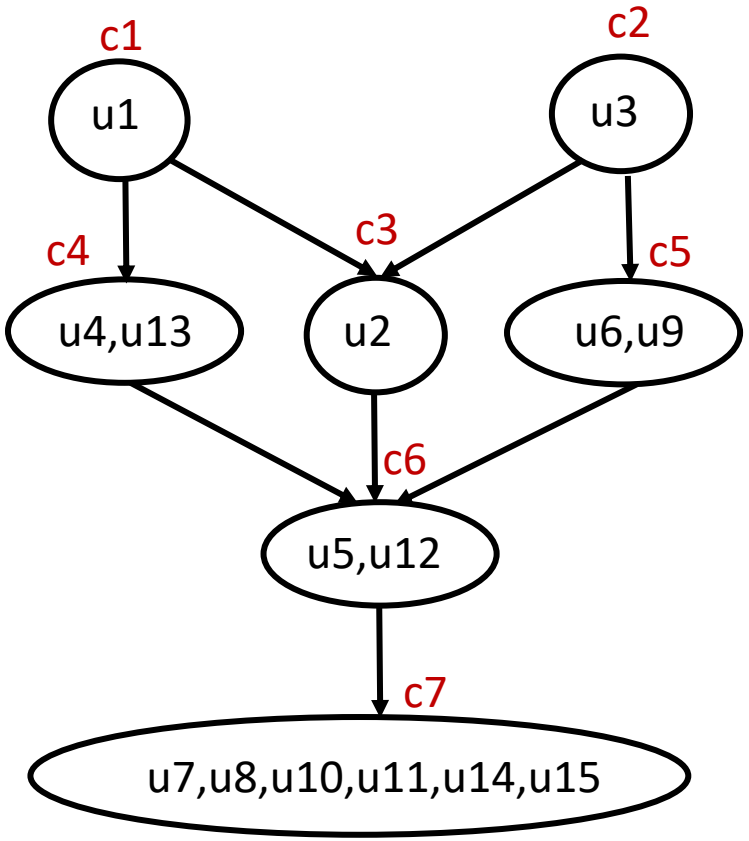


ADR Invariants & Inequalities

- ADR invariants and inequalities define a partial order on the distribution of unrooted gene trees $\vec{u}$
- The partial order for each tree shape can be shown with a Hasse diagram



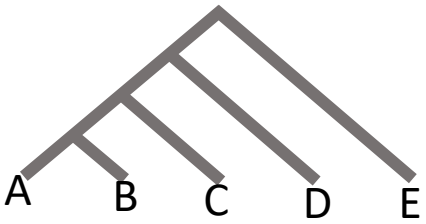
Caterpillar



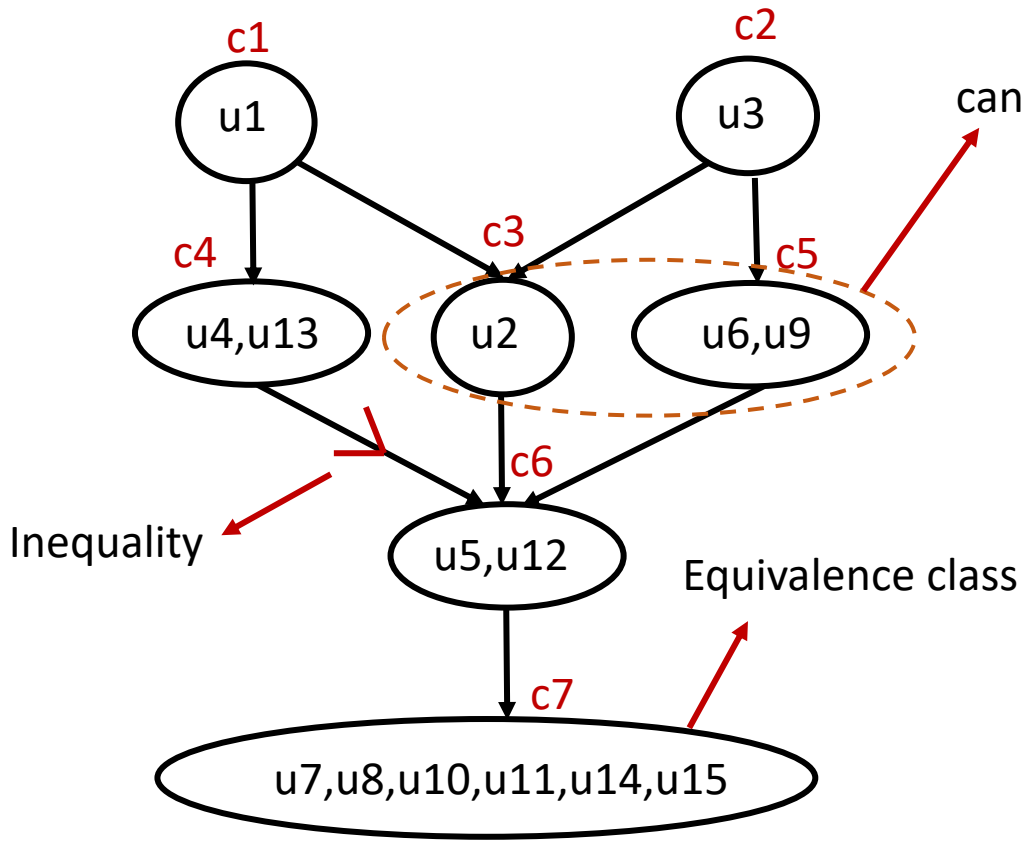
- 15 5-taxon unrooted topologies $T_1, \ldots, T_{15}$
- $u_i = \mathbb{P}(T_i)$

ADR Invariants & Inequalities

- ADR invariants and inequalities define a partial order on the distribution of unrooted gene trees $\vec{u}$
- The partial order for each tree shape can be shown with a Hasse diagram



Caterpillar



- 15 5-taxon unrooted topologies $T_1, \ldots, T_{15}$
- $u_i = \mathbb{P}(T_i)$

Example of invariants:

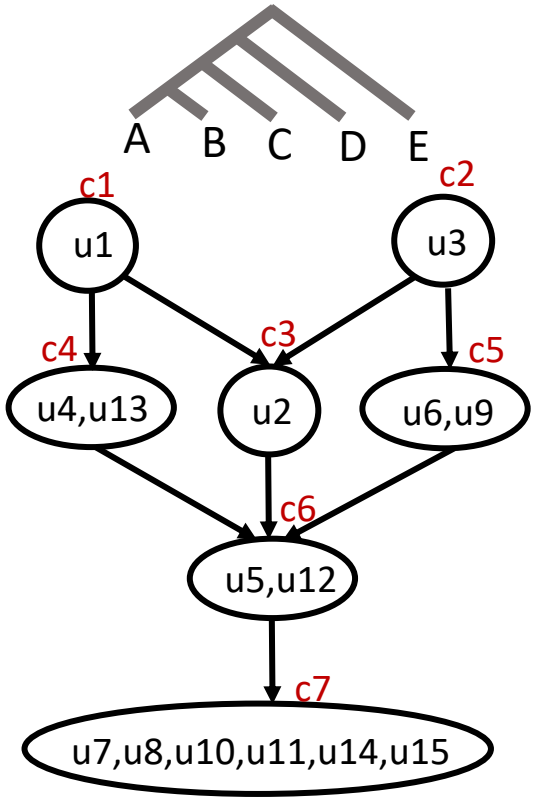
- $u_4 = u_{13}$

Example of inequalities:

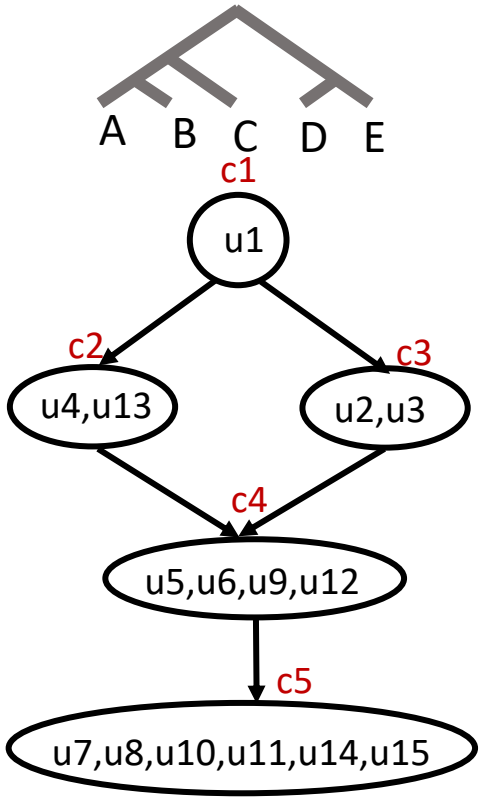
- $u_4 > u_5$

- Equivalence classes that are not related by inequalities can merge for some values of branch lengths

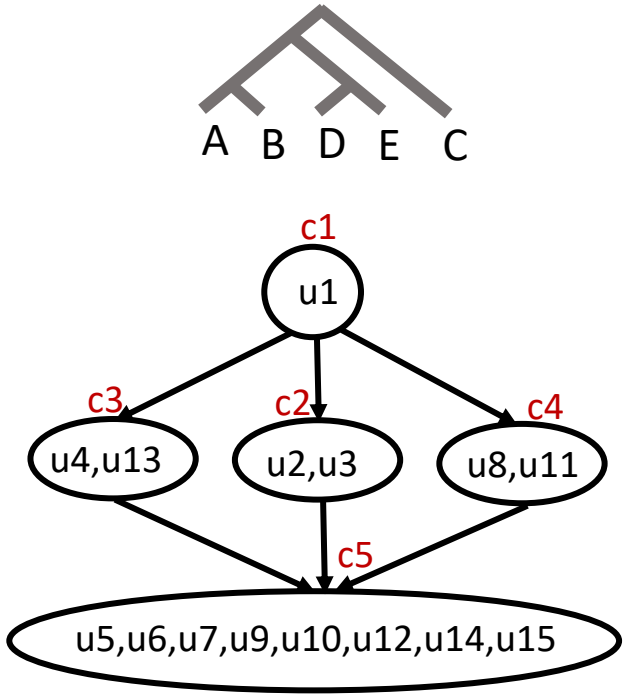
ADR Invariants & Inequalities



Caterpillar



Balanced

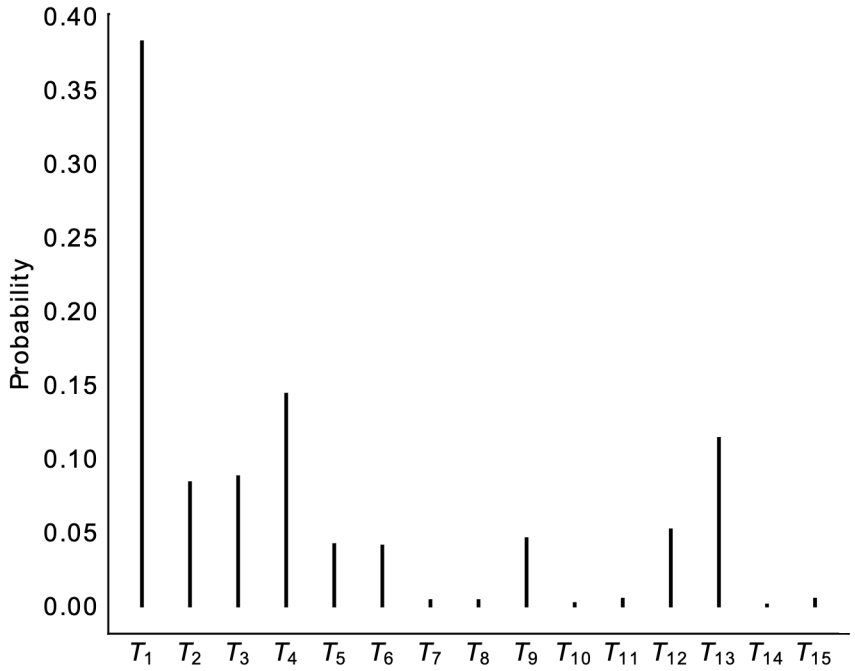


Pseudo-Caterpillar

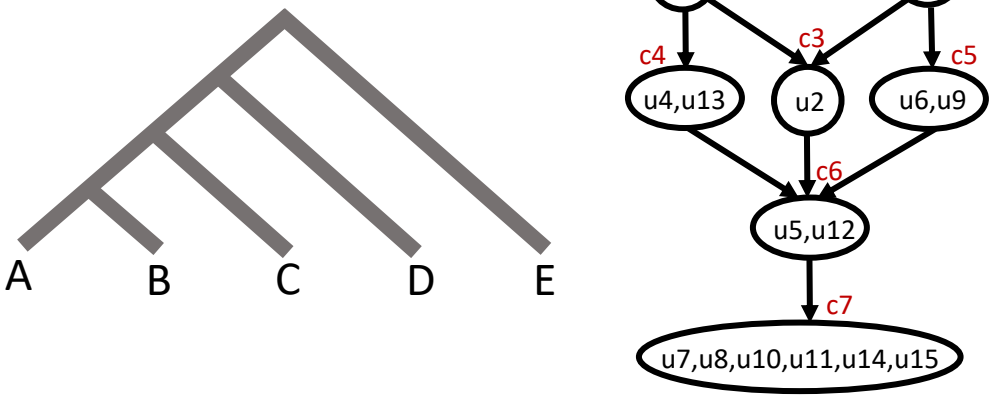
- According to ADR theory, each 105 rooted binary tree corresponds to a unique Hasse diagram
- The shape of this diagram only depends on the topological shape of the tree

Cost : Fitness between a Tree and a Distribution

- Measures the fitness between a distribution and a tree (i.e. its partial order)
- Linear combination of invariant and inequality penalty terms



Estimated gene tree distribution $\vec{u}$

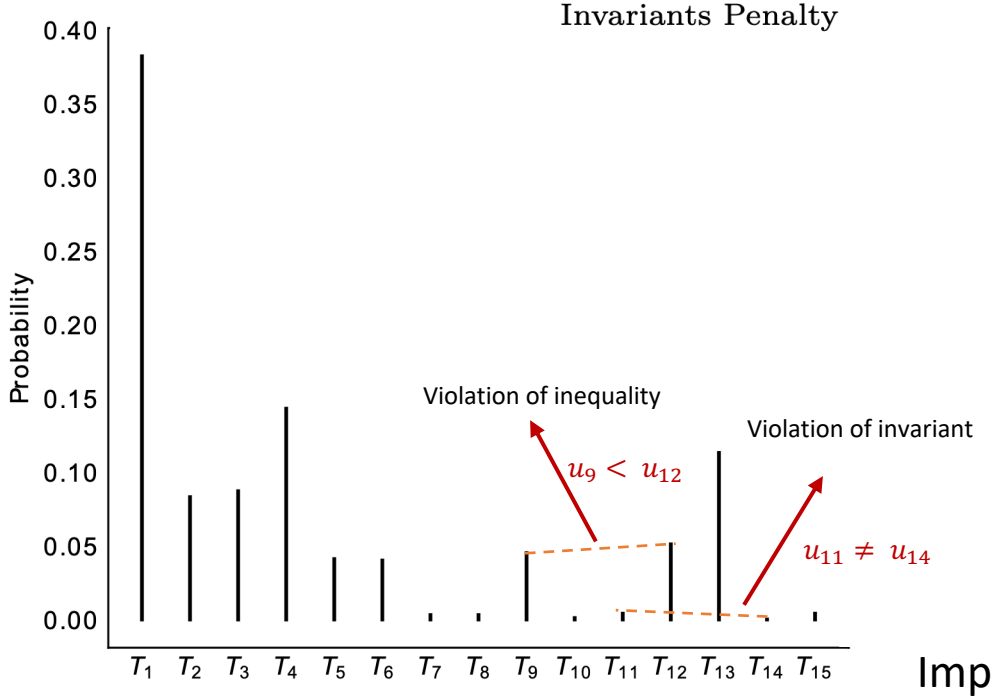


A model tree R and its partial order

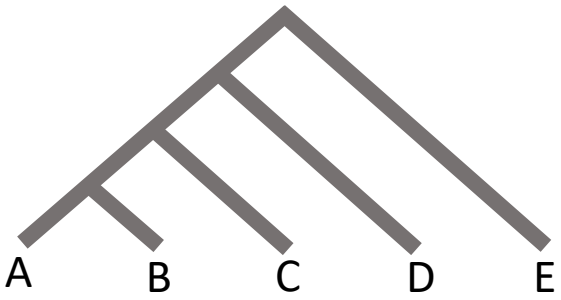
Cost : Fitness between a Tree and a Distribution

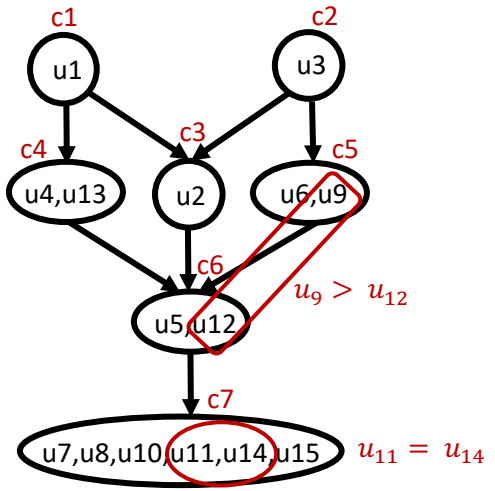
- Measures the fitness between a distribution and a tree (i.e. its partial order)
- Linear combination of invariant and inequality penalty terms

$$\mathrm{Cost}(r,\vec{\hat{u}}) = \underbrace{\sum_{c\in C_r}\frac{1}{|c|}\sum_{u_a,u_b\in c}|\hat{u}_a-\hat{u}_b|}_{\text{Invariants Penalty}} + \underbrace{\sum_{c>c'\in C_r}\frac{1}{|c'|}\sum_{u_a\in c,u_b\in c'}\max(0,\hat{u}_b-\hat{u}_a)}_{\text{Inequalities Penalty}}.$$



Estimated gene tree distribution $\vec{u}$





A model tree R and its partial order

Implied by the distribution:

- $u_{11} \neq u_{14}$
- $u_9 < u_{12}$

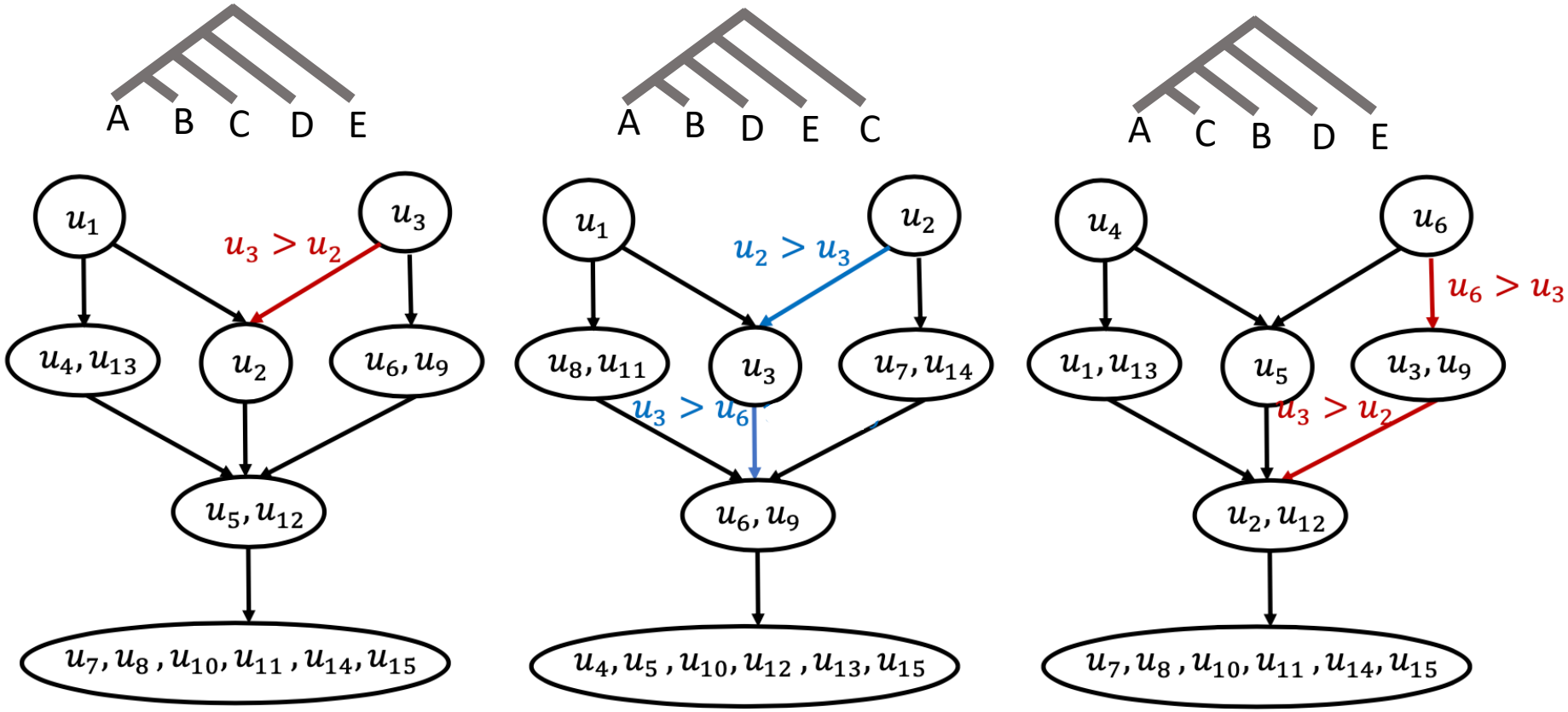
violations

Implied by the partial order:

- $u_{11} = u_{14}$
- $u_9 > u_{12}$

Conflicting inequalities between rooted quintets

$\boldsymbol{V(R, R')}$:
Set of violated inequalities between two rooted quintet trees

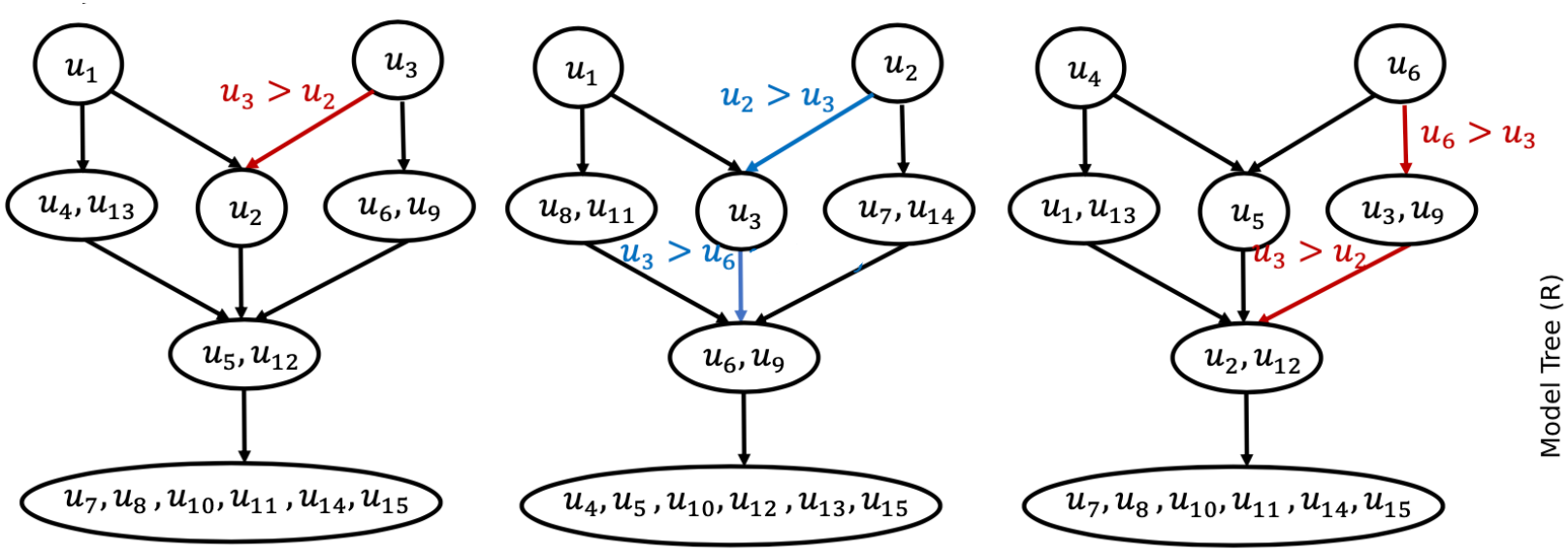


$R_1 = (((a, b), c), d), e)$ $R_4 = (((a, b), d), e), c)$ $R_7 = (((a, c), b), d), e)$

$$V(R_1, R_4) = \{\{2, 3\}\} \rightarrow |V(R_1, R_4)| = 1$$

$$V(R_7, R_4) = \{\{2, 3\}, \{3, 6\}\} \rightarrow |V(R_7, R_4)| = 2$$

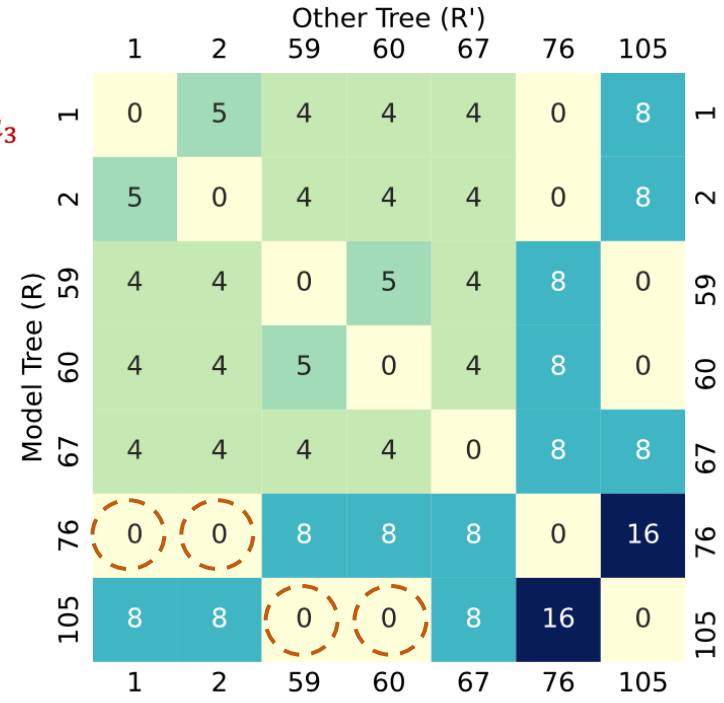
Violation of Invariants and Inequalities

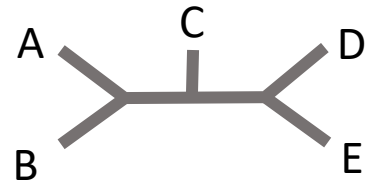


$R_1 = (((a, b), c), d), e)$ $R_4 = (((a, b), d), e), c)$ $R_7 = (((a, c), b), d), e)$

$$V(R_1, R_4) = \{\{2, 3\}\} \rightarrow |V(R_1, R_4)| = 1$$
$$V(R_7, R_4) = \{\{2, 3\}, \{3, 6\}\} \rightarrow |V(R_7, R_4)| = 2$$

Heatmap showing |V(R,R')|





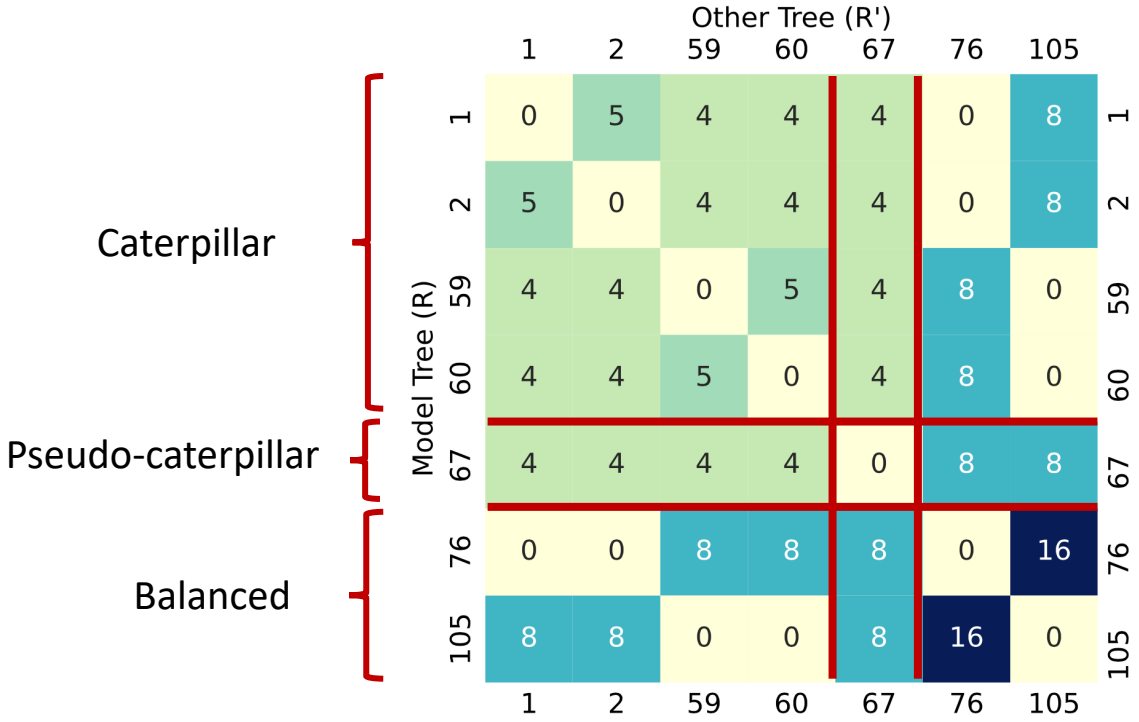
- There are pairs of trees whose partial orders have no conflicts → QR is not consistent.

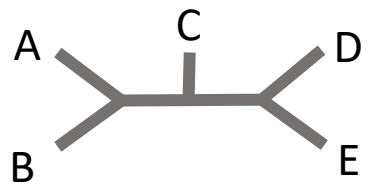
Key Idea behind QR-STAR

- Pairs of trees with the same rooted topological shape (caterpillar, balanced, pseudo-caterpillar) always have conflicting distributions

- Idea:
  - Determine the topological shape of each quintet
  - Incorporate the topological shape in the cost function

Heatmap of the number of conflicts between pairs of trees with the same unrooted topology





New Cost Function

$$\text{Cost}(r,\vec{\hat{u}}) = \underbrace{\sum_{c\in C_r}\frac{1}{|c|}\sum_{u_a,u_b\in c}|\hat{u}_a-\hat{u}_b|}_{\text{Invariants Penalty}} + \underbrace{\sum_{c>c'\in C_r}\frac{1}{|c'|}\sum_{u_a\in c,u_b\in c'}\max(0,\hat{u}_b-\hat{u}_a)}_{\text{Inequalities Penalty}}$$

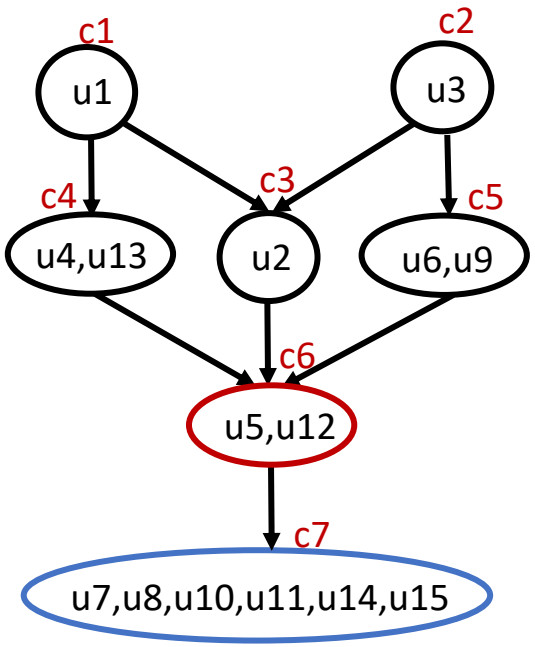
→ **QR**

$$\text{Cost}^*(r,\vec{\hat{u}}) = \underbrace{\sum_{c\in C_r}\sum_{u_a,u_b\in c}\alpha_{a,b}|\hat{u}_a-\hat{u}_b|}_{\text{Invariants Penalty}} + \underbrace{\sum_{c>c'\in C_r}\sum_{u_a\in c,u_b\in c'}\beta_{a,b}\max(0,\hat{u}_b-\hat{u}_a)}_{\text{Inequalities Penalty}} + \underbrace{C\mathbb{1}|S(r)\neq\hat{S}(\hat{u})|}_{\text{Shape Penalty}}$$

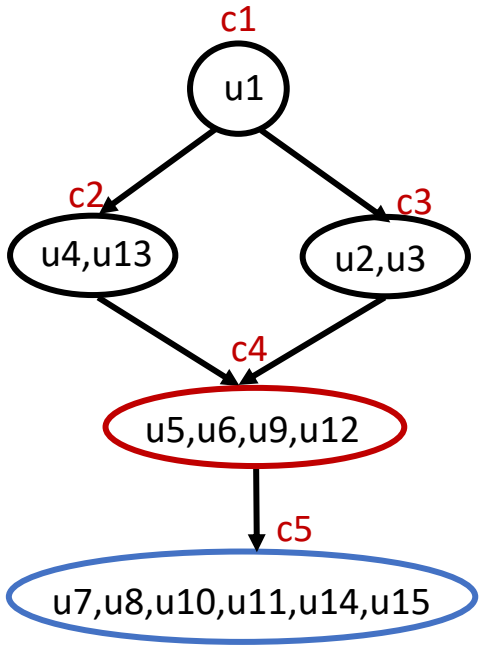
→ **QR-STAR**

Constraints: $\alpha_{a,b} \geq 0, \beta_{a,b}, C > 0$

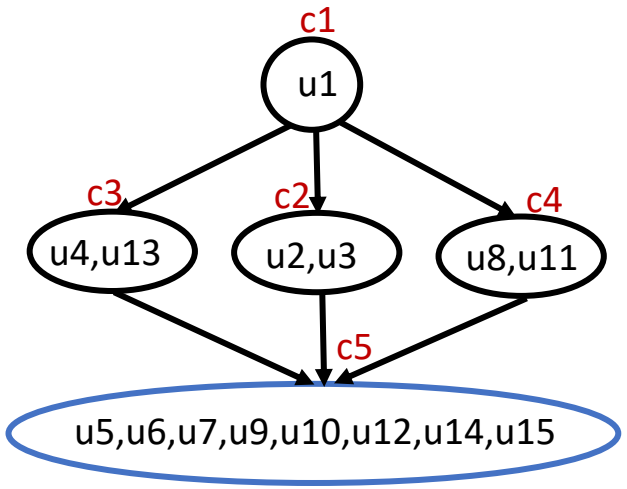
How to differentiate between different rooted shapes?



Caterpillar



Balanced



Pseudo-Caterpillar

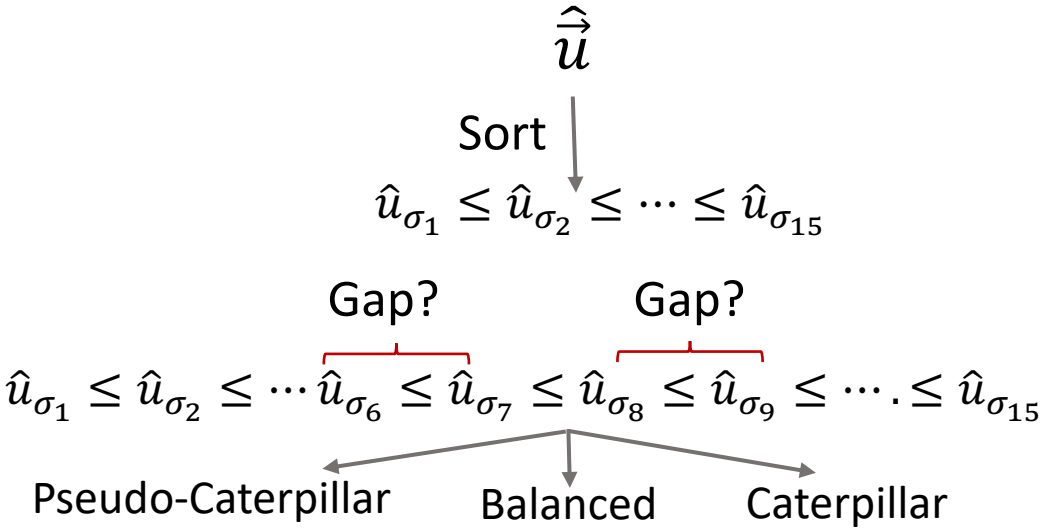
Size of the class with the

- smallest probability: **8** for pseudo-caterpillar, **6** for other tree shapes
- second smallest probability: **4** for balanced, **2** for caterpillar

How to differentiate between different rooted shapes given *finite data*?

- It is likely that none of the invariants derived from the ADR theory exactly hold
- Class sizes can not be directly determined given finite data

- **Idea**: Look for significant gaps between quintet gene tree probabilities.



Additional Step: Determining the Rooted Tree shape

- **Idea**: Look for significant gaps between quintet gene tree probabilities.
- Let $A(k) = \sqrt{\frac{2}{k}\ln(30|Q|k)}$

k = number of gene trees
Q = set of sampled quintets

$$\text{Pseudo-caterpillar:}\quad \hat{u}_{\sigma_1} \le \hat{u}_{\sigma_2} \le \cdots \le \underbrace{\hat{u}_{\sigma_6} \le \hat{u}_{\sigma_7}}_{\hat{u}_{\sigma_7}-\hat{u}_{\sigma_6}<A(k)} \le \underbrace{\hat{u}_{\sigma_8} < \hat{u}_{\sigma_9}}_{\hat{u}_{\sigma_9}-\hat{u}_{\sigma_8}>A(k)} \le \hat{u}_{\sigma_{10}} \le \cdots \le \hat{u}_{\sigma_{15}}$$

$$\text{Balanced:}\quad \hat{u}_{\sigma_1} \le \hat{u}_{\sigma_2} \le \cdots \le \underbrace{\hat{u}_{\sigma_6} < \hat{u}_{\sigma_7}}_{\hat{u}_{\sigma_7}-\hat{u}_{\sigma_6}>A(k)} \le \underbrace{\hat{u}_{\sigma_8} \le \hat{u}_{\sigma_9}}_{\hat{u}_{\sigma_9}-\hat{u}_{\sigma_8}<A(k)} \le \hat{u}_{\sigma_{10}} \le \cdots \le \hat{u}_{\sigma_{15}}$$

$$\text{Caterpillar:}\quad \hat{u}_{\sigma_1} \le \hat{u}_{\sigma_2} \le \cdots \le \underbrace{\hat{u}_{\sigma_6} < \hat{u}_{\sigma_7}}_{\hat{u}_{\sigma_7}-\hat{u}_{\sigma_6}>A(k)} \le \underbrace{\hat{u}_{\sigma_8} < \hat{u}_{\sigma_9}}_{\hat{u}_{\sigma_9}-\hat{u}_{\sigma_8}>A(k)} \le \hat{u}_{\sigma_{10}} \le \cdots \le \hat{u}_{\sigma_{15}}$$

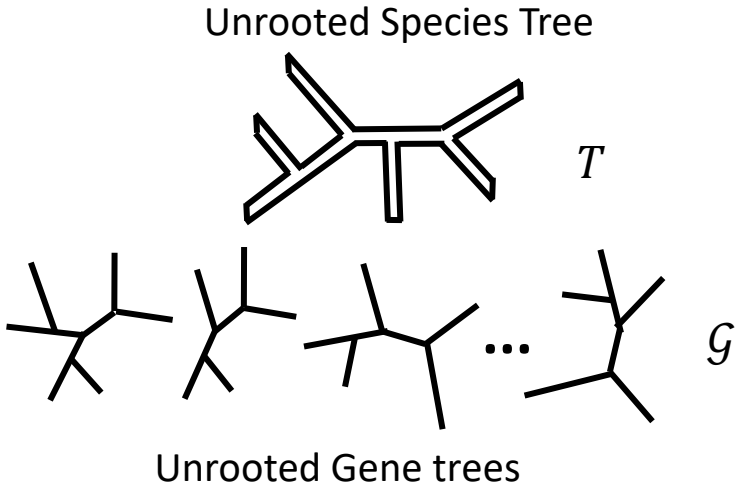
- Estimate the rooted shape $\hat{S}(\hat{u})$ as pseudo-caterpillar if $\hat{u}_{\sigma_7} - \hat{u}_{\sigma_6} < A(k)$
- Estimate the rooted shape $\hat{S}(\hat{u})$ as balanced if $\hat{u}_{\sigma_7} - \hat{u}_{\sigma_6} \ge A(k)$ and $\hat{u}_{\sigma_9} - \hat{u}_{\sigma_8} < A(k)$
- Estimate the rooted shape $\hat{S}(\hat{u})$ as caterpillar if $\hat{u}_{\sigma_7} - \hat{u}_{\sigma_6} \ge A(k)$ and $\hat{u}_{\sigma_9} - \hat{u}_{\sigma_8} \ge A(k)$

Statistical Consistency of QR-STAR (Proof Sketch)

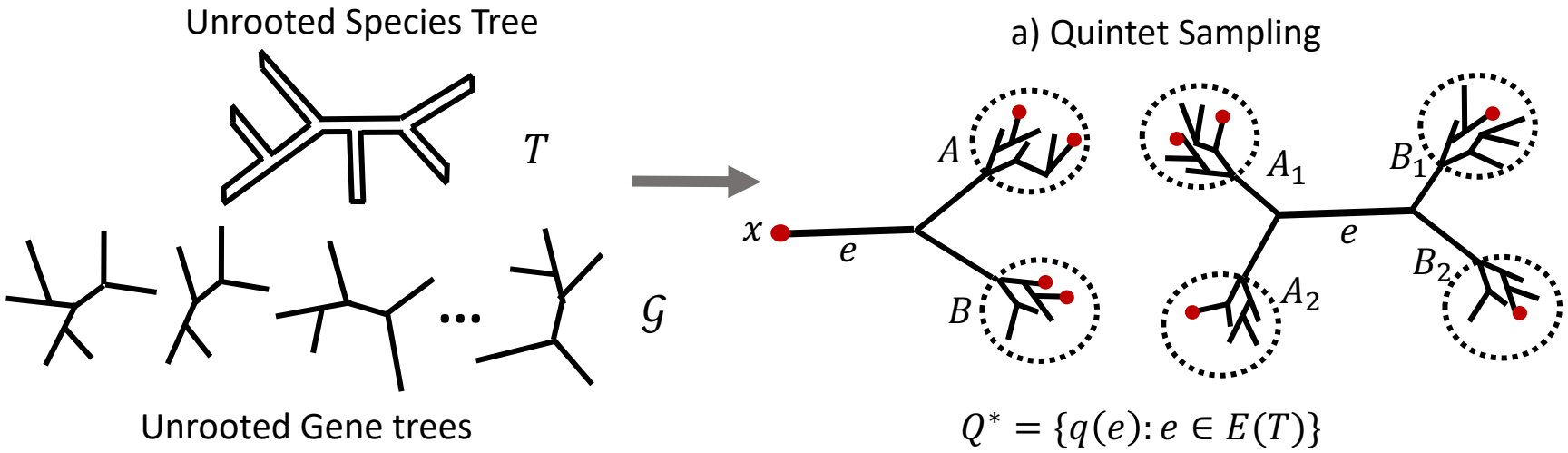
- As the number of input gene trees increase
  - Probability that the first step of QR-STAR correctly determines the rooted shape of each quintet converges to 1
  - The cost of true rooted quintet becomes arbitrarily close to zero
  - The cost of any other rooted quintet is bounded away from zero, where the bound depends on the *path length parameter* of the model tree $h(R)$
  - The set of quintets sampled in QR-STAR is selected so that each two different rooted trees define different set of quintets

- Therefore, the probability that QR-STAR correctly roots the given unrooted tree converges to 1

QR-STAR Algorithm

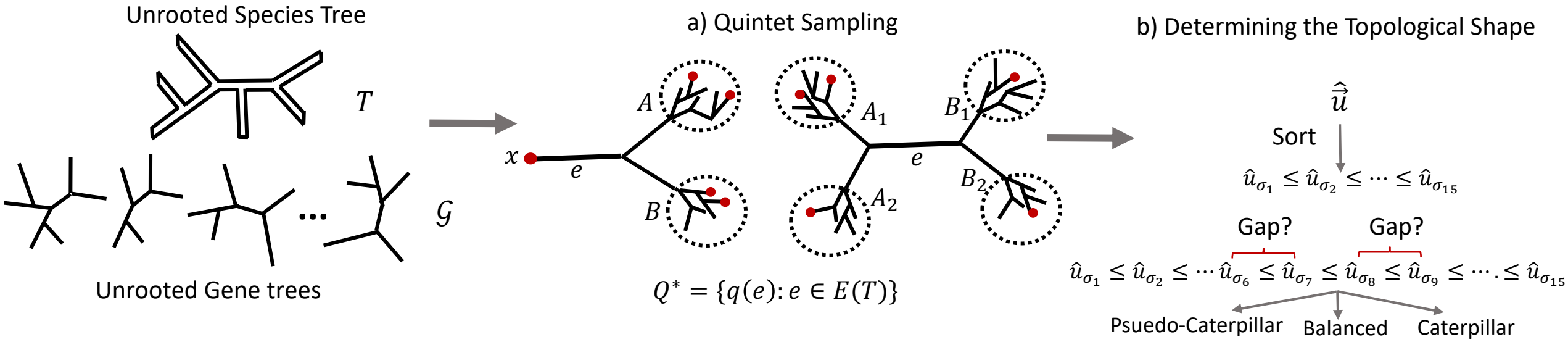


QR-STAR Algorithm

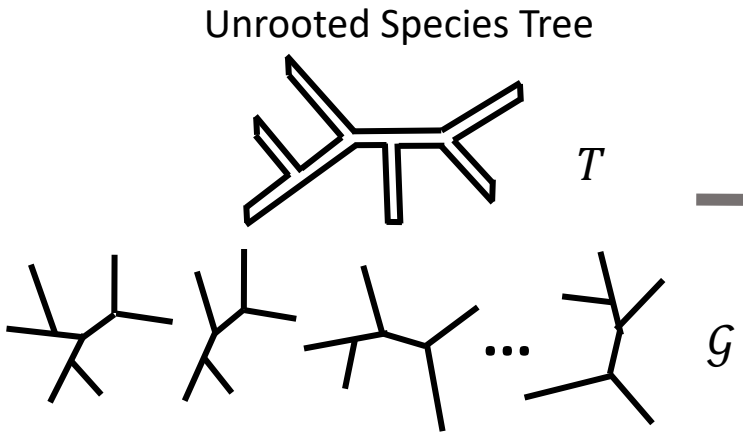


$Q^* = \{q(e): e \in E(T)\}$

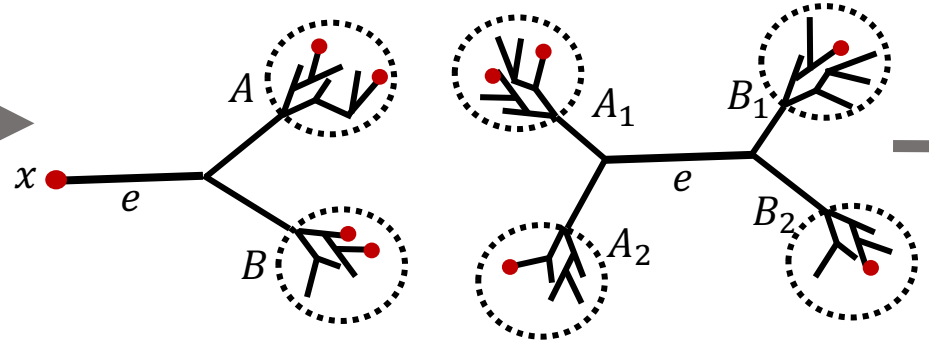
QR-STAR Algorithm



QR-STAR Algorithm



a) Quintet Sampling



$Q^* = \{q(e): e \in E(T)\}$

b) Determining the Topological Shape

$\hat{\vec{u}}$

Sort

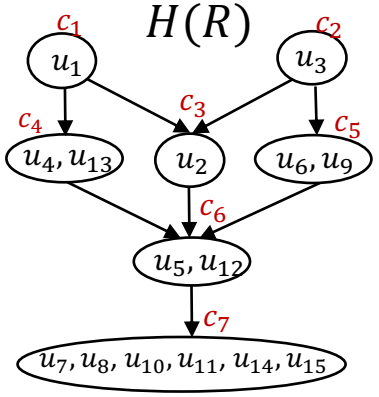
$\hat{u}_{\sigma_1} \leq \hat{u}_{\sigma_2} \leq \cdots \leq \hat{u}_{\sigma_{15}}$

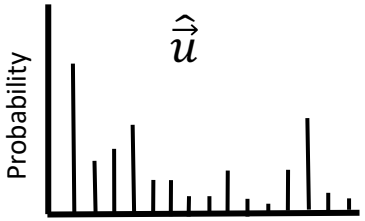
Gap? Gap?

$\hat{u}_{\sigma_1} \leq \hat{u}_{\sigma_2} \leq \cdots \hat{u}_{\sigma_6} \leq \hat{u}_{\sigma_7} \leq \hat{u}_{\sigma_8} \leq \hat{u}_{\sigma_9} \leq \cdots . \leq \hat{u}_{\sigma_{15}}$

Psuedo-Caterpillar | Balanced | Caterpillar

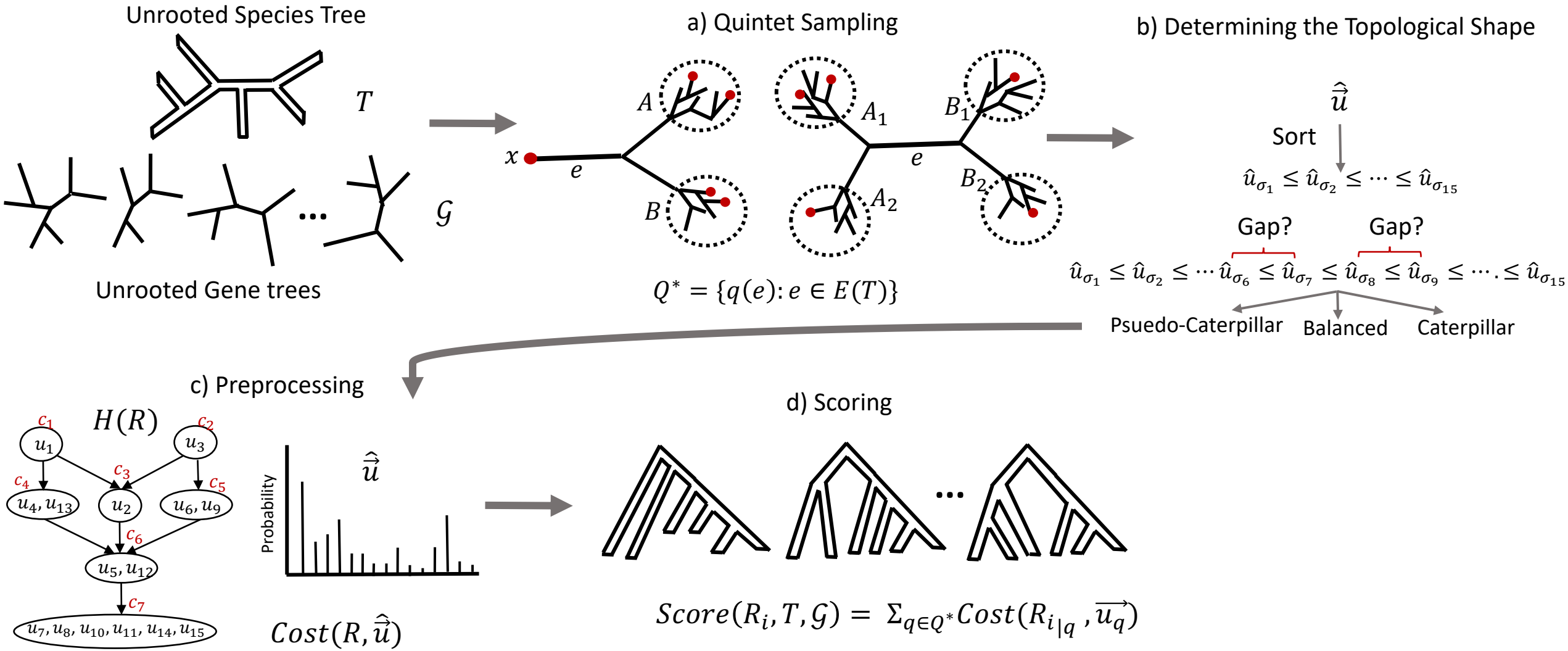
c) Preprocessing





$Cost(R, \hat{\vec{u}})$

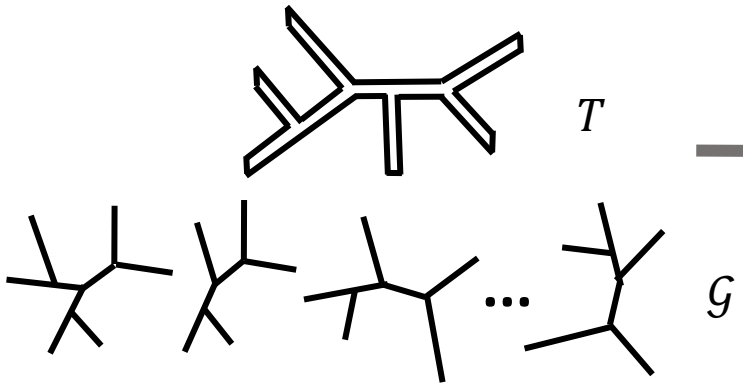
QR-STAR Algorithm



QR-STAR Algorithm

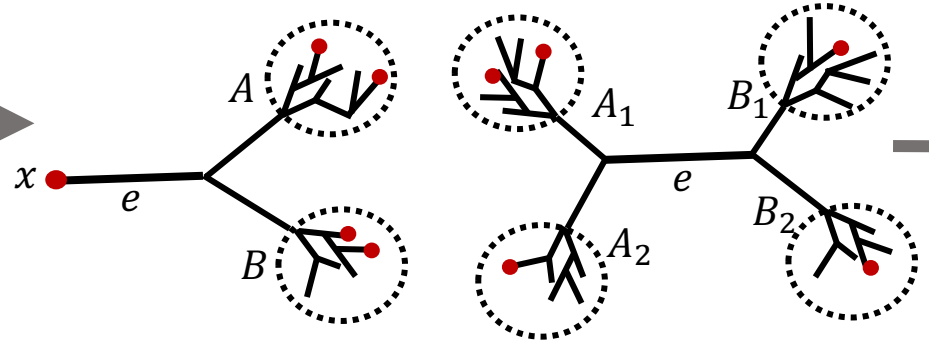
Runtime: O(nk)

Unrooted Species Tree



Unrooted Gene trees

a) Quintet Sampling



$Q^* = \{q(e): e \in E(T)\}$

b) Determining the Topological Shape

$\hat{\vec{u}}$

Sort

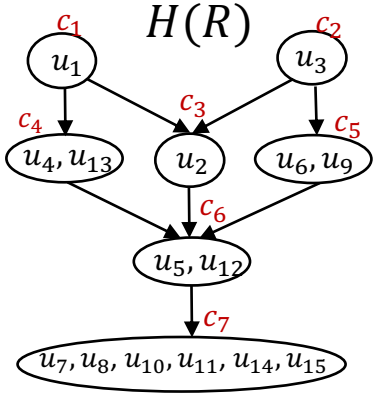
$\hat{u}_{\sigma_1} \leq \hat{u}_{\sigma_2} \leq \cdots \leq \hat{u}_{\sigma_{15}}$

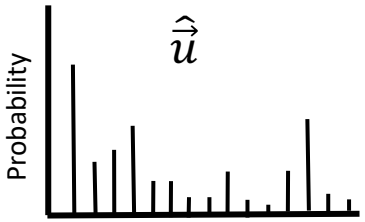
Gap? Gap?

$\hat{u}_{\sigma_1} \leq \hat{u}_{\sigma_2} \leq \cdots \hat{u}_{\sigma_6} \leq \hat{u}_{\sigma_7} \leq \hat{u}_{\sigma_8} \leq \hat{u}_{\sigma_9} \leq \cdots . \leq \hat{u}_{\sigma_{15}}$

Psuedo-Caterpillar | Balanced | Caterpillar

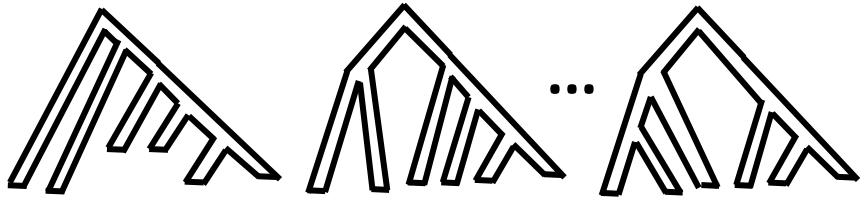
c) Preprocessing





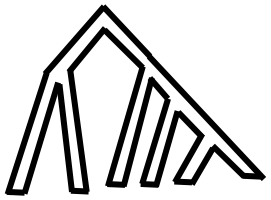
$Cost(R, \hat{\vec{u}})$

d) Scoring



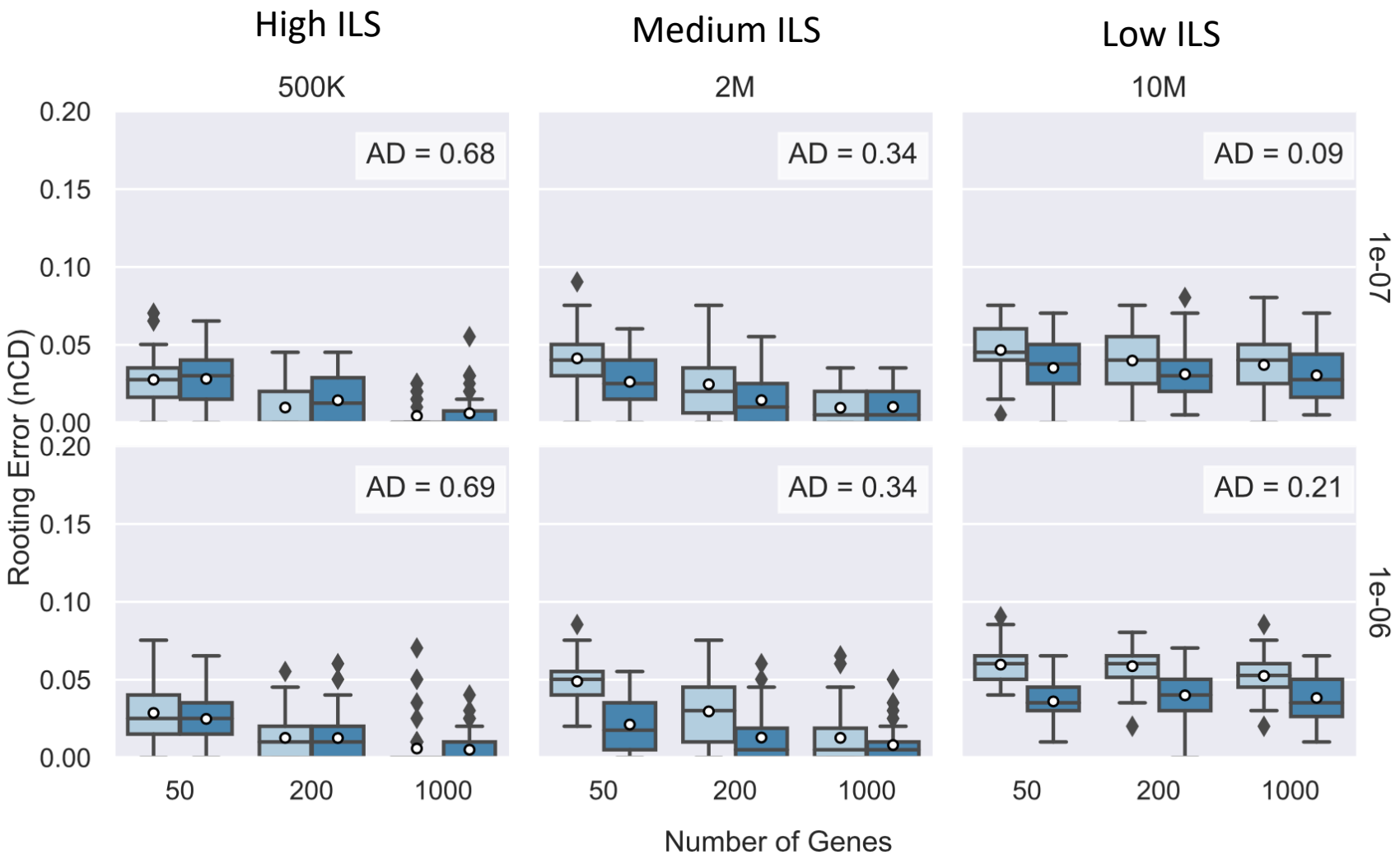
$Score(R_i, T, \mathcal{G}) = \Sigma_{q \in Q^*} Cost(R_{i|q}, \overrightarrow{u_q})$

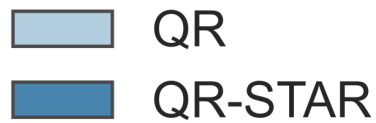
Rooted Species Tree



$argmin_R Score(R, T, \mathcal{G})$

Simulation Study on 201-taxon ILS dataset





Speciation rate

AD: true gene tree discordance, measures the level of ILS

GTEE: gene tree estimation error, varies between 0.22 and 0.49

- QR-STAR is run with parameters C=1e-02 and $\frac{\alpha}{\beta}$ = 0.

Summary & Future Directions

Summary

- QR-STAR is a polynomial-time statistically consistent method for rooting species trees in the presence of ILS
- It is based on the ADR theory of identifiability of rooted 5-leaf species trees from unrooted gene trees under MSC
- QR-STAR has improved accuracy over QR (ISMB’22) in most model conditions

Future Directions

- Consistency of rooting methods in the presence of other sources of gene tree discordance (e.g. GDL)
- Develop consistent methods for *inferring* the rooted tree topology *directly* from unrooted gene trees

Acknowledgements

Thank you!


Members of Warnow Lab

Full paper is available at:
https://www.biorxiv.org/content/biorxiv/early/2023/01/06/2022.10.26.513897.full.pdf

Software is available on Github:
https://github.com/ytabatabaee/Quintet-Rooting

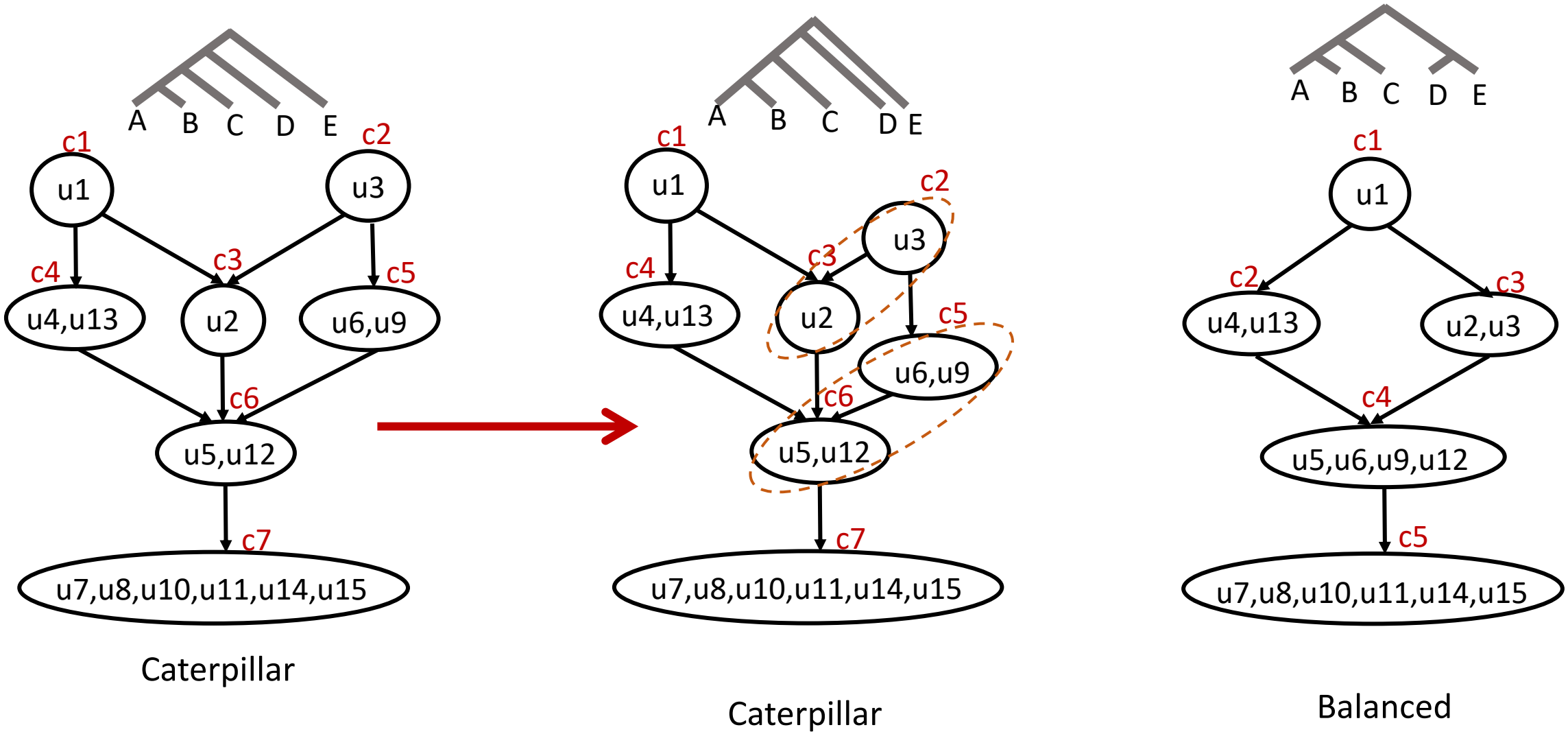
Funding:

- NSF grants DMS-1902892, DMS1916378 and DMS-2023239 (TRIPODS Phase II), Vilas Associates Award to SR
- Grainger Foundation to TW

Computing Resources:
UIUC Campus Cluster

Backup Slides

Example: Quintets with no violation



No violation between inequalities (the caterpillar partial order **degenerates** into the balanced partial order)

Observations

- Rooting under ILS is easier with more discordance
  - unlike species tree estimation
  - consider trade-off in biological analysis

- Rooting is difficult with very long and very short branches, although species tree estimation is mainly impacted by short branches
  - Sample complexity of QR-STAR depends on both shortest branch and longest path

- It is possible to consistently infer the rooted topology *directly* from unrooted gene trees for 5-taxon trees
  - Rather than the typical two-step approach

Simulation Study

Simulated Datasets:

- ILS-only datasets
- Training: 101-taxon dataset [Zhang et al, 2018]
- Test: 201-taxon dataset [Mirarab et al, 2015]
- Varying
  - Number of genes: 50-1000
  - Gene tree discordance (ILS): 0.09 to 0.69
  - Gene tree estimation error (GTEE): 0.0 to 0.49
  - Tree height
  - Speciation rate

Pipeline:

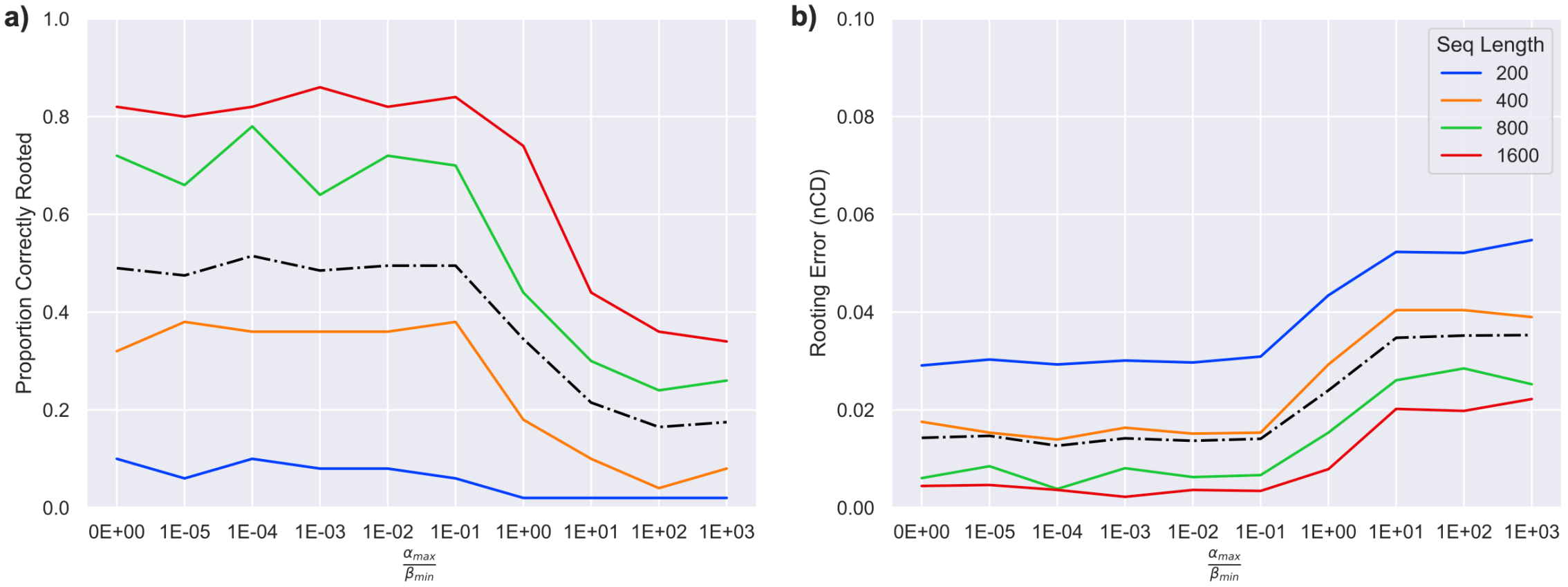
- Rooting true species tree with true and estimated gene trees
  - Only rooting error
- Rooting estimated (ASTRAL) species tree with true and estimated gene trees
  - Species tree estimation + rooting error

Evaluation Criteria:

Average normalized clade distance (nCD)

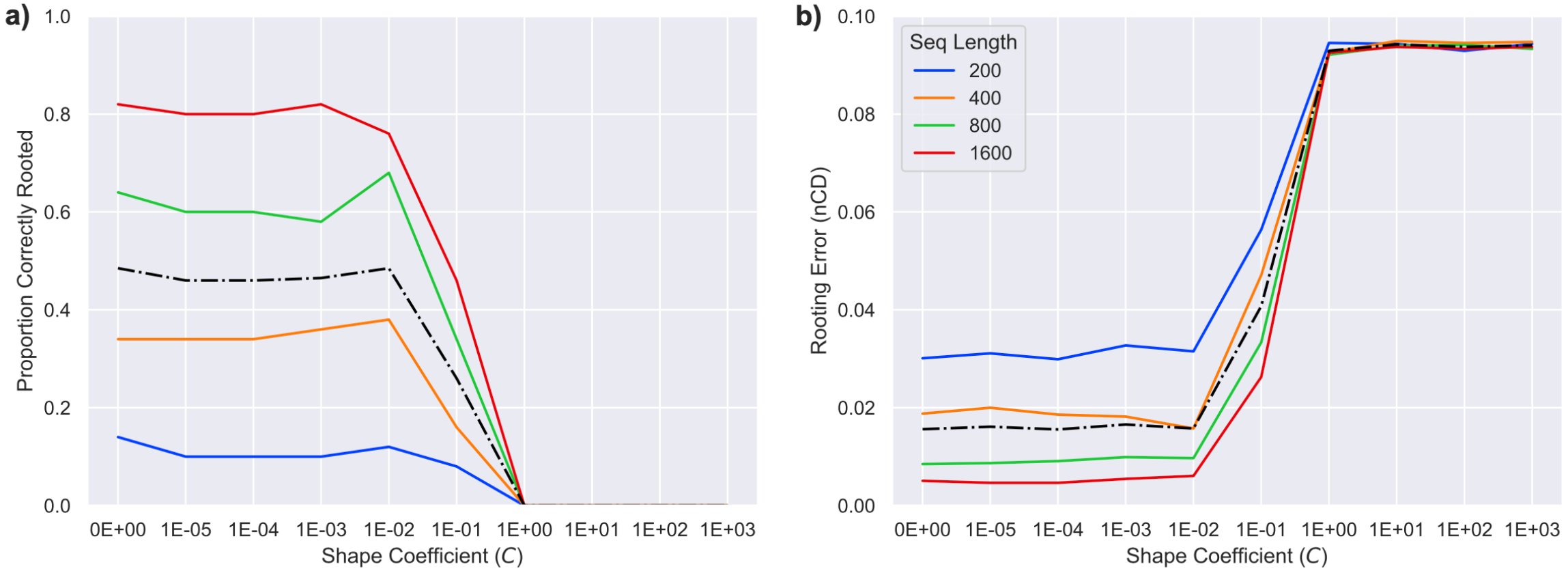
- Number of clades differing between the true and estimated trees

Impact of invariants and inequalities (α_{max}/β_{min} ratio) on QR-STAR cost function



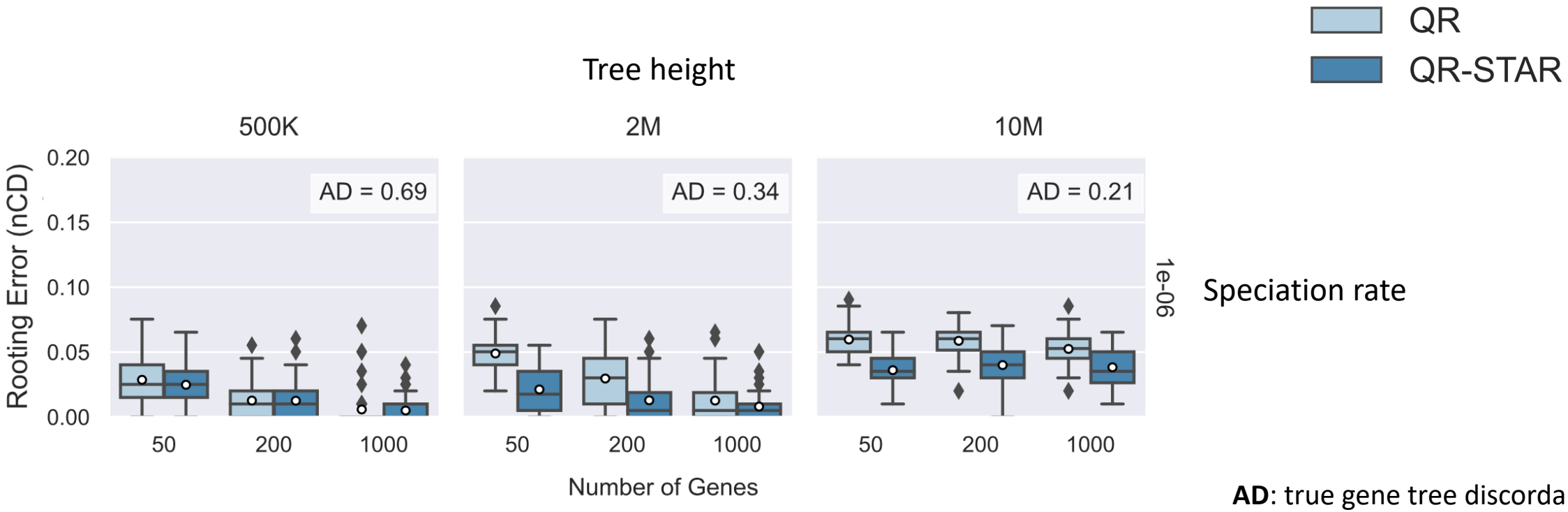
- 101-taxon trees, 0.46 AD ILS, GTEE varies between 0.23 to 0.55

Impact of shape coefficient (C) on QR-STAR cost function



- 101-taxon trees, 0.46 AD ILS, GTEE varies between 0.23 to 0.55

Simulations on 201-taxon ILS dataset

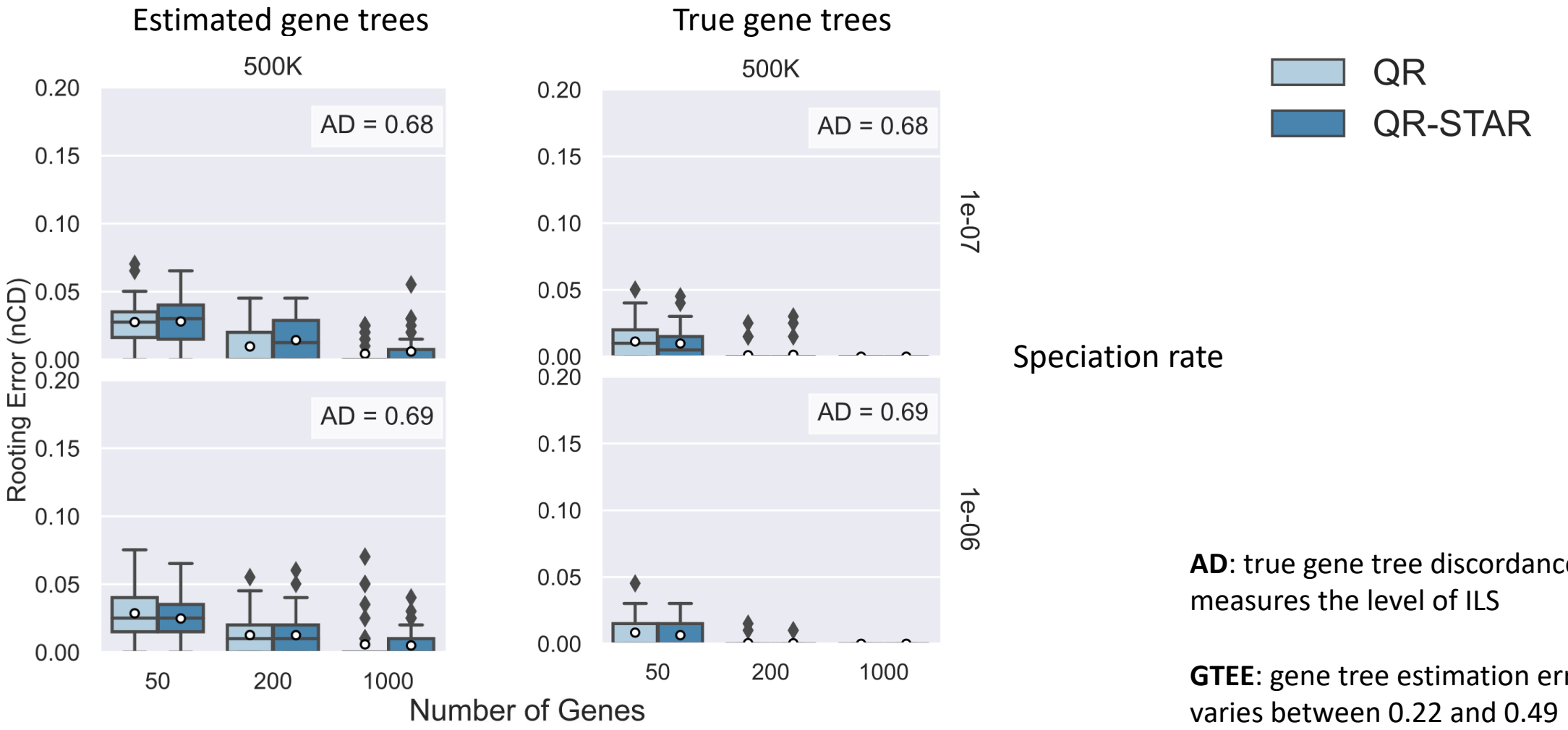


AD: true gene tree discordance, measures the level of ILS

GTEE: gene tree estimation error, varies between 0.22 and 0.49

- QR-STAR is run with parameters C=1e-02 and $\frac{\alpha}{\beta}$ = 0.

Rooting the true species tree topology with true/estimated gene trees

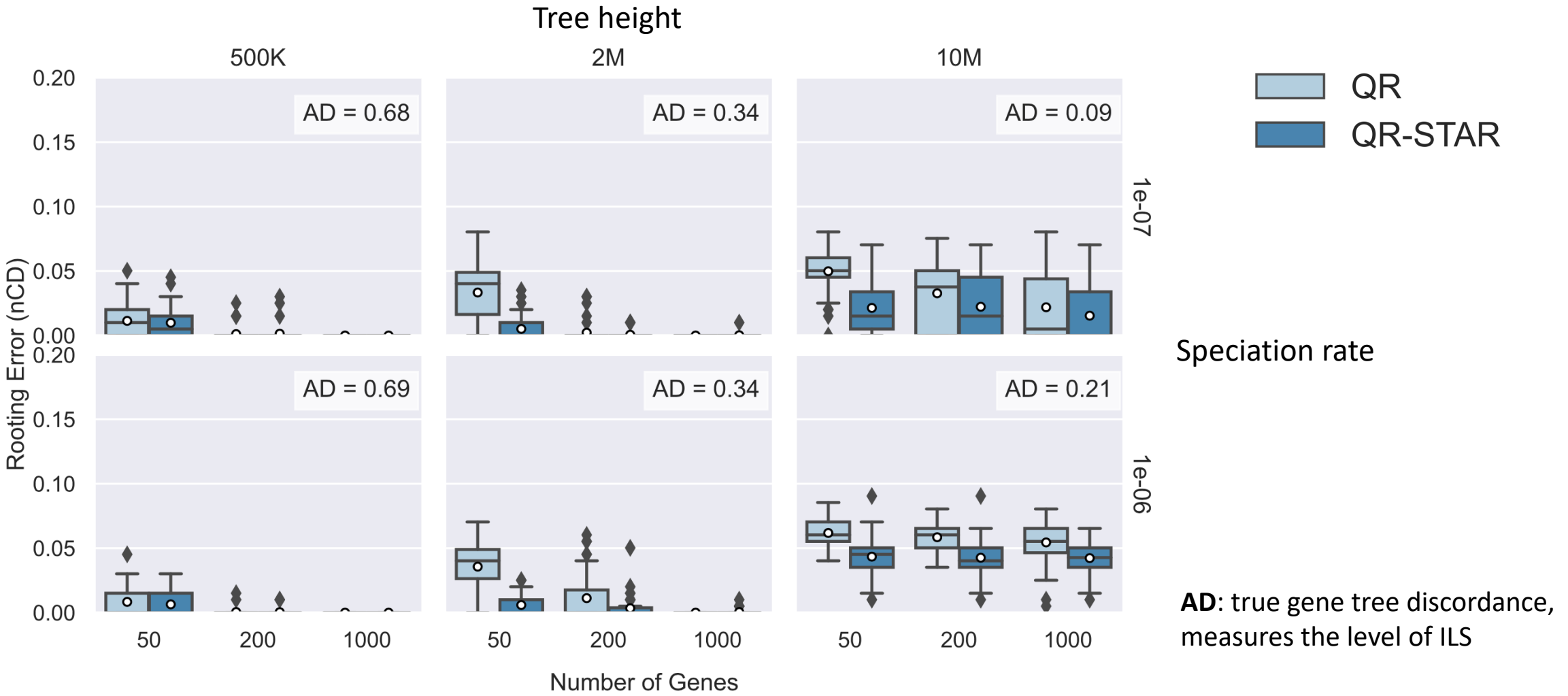


AD: true gene tree discordance, measures the level of ILS

GTEE: gene tree estimation error, varies between 0.22 and 0.49

- QR-STAR is run with parameters C=1e-02 and $\frac{\alpha}{\beta}$ = 0.

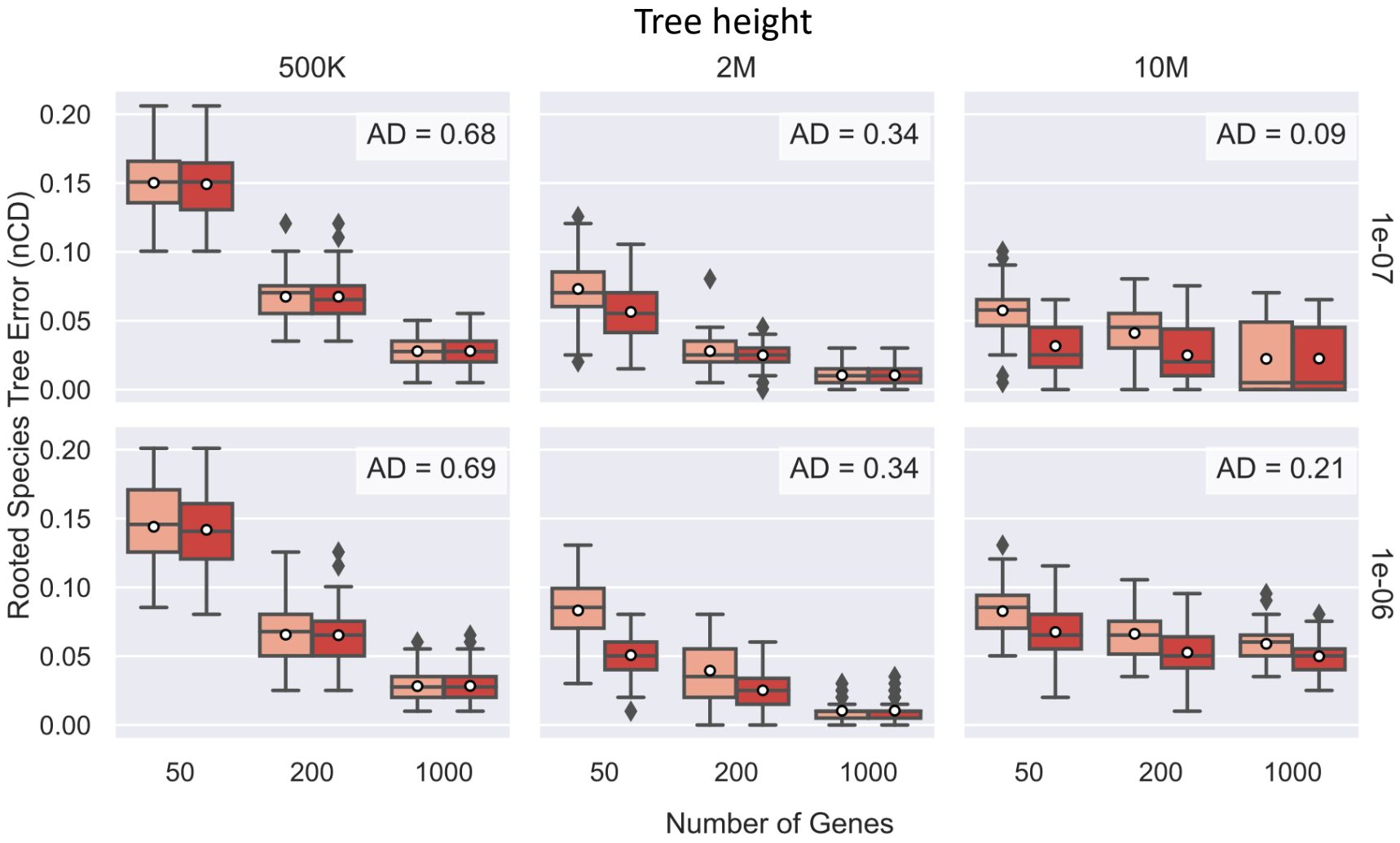
Rooting the true species tree topology with true gene trees

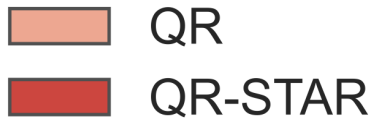


AD: true gene tree discordance, measures the level of ILS

- QR-STAR is run with parameters C=1e-02 and $\frac{\alpha}{\beta}$ = 0.

Rooting the estimated species tree topology with true gene trees



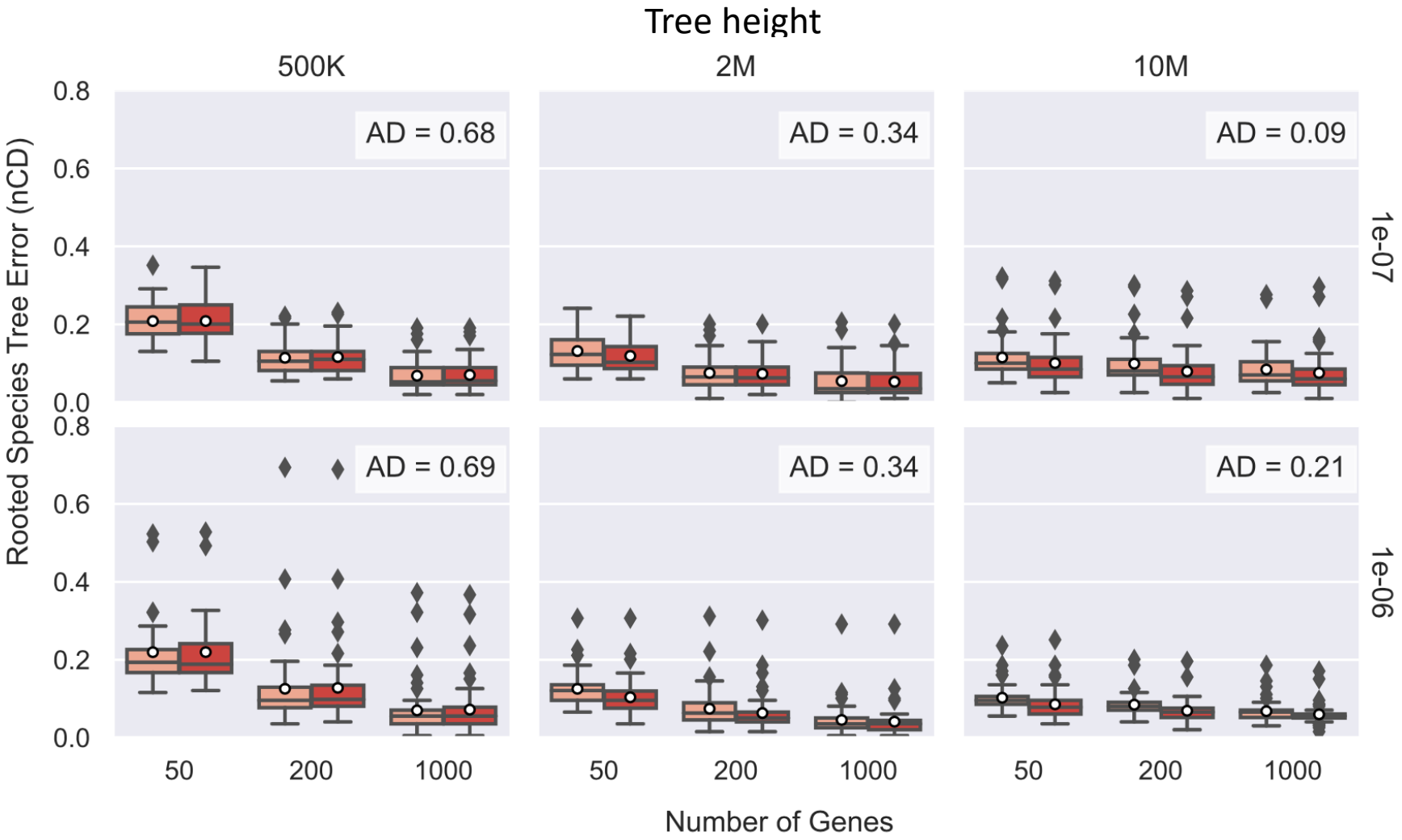


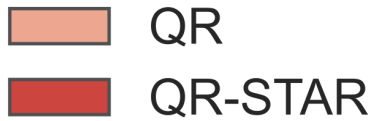
Speciation rate

AD: true gene tree discordance, measures the level of ILS

- QR-STAR is run with parameters C=1e-02 and $\frac{\alpha}{\beta}$ = 0.

Rooting the estimated species tree topology with estimated gene trees





Speciation rate

AD: true gene tree discordance, measures the level of ILS

GTEE: gene tree estimation error, varies between 0.22 and 0.49

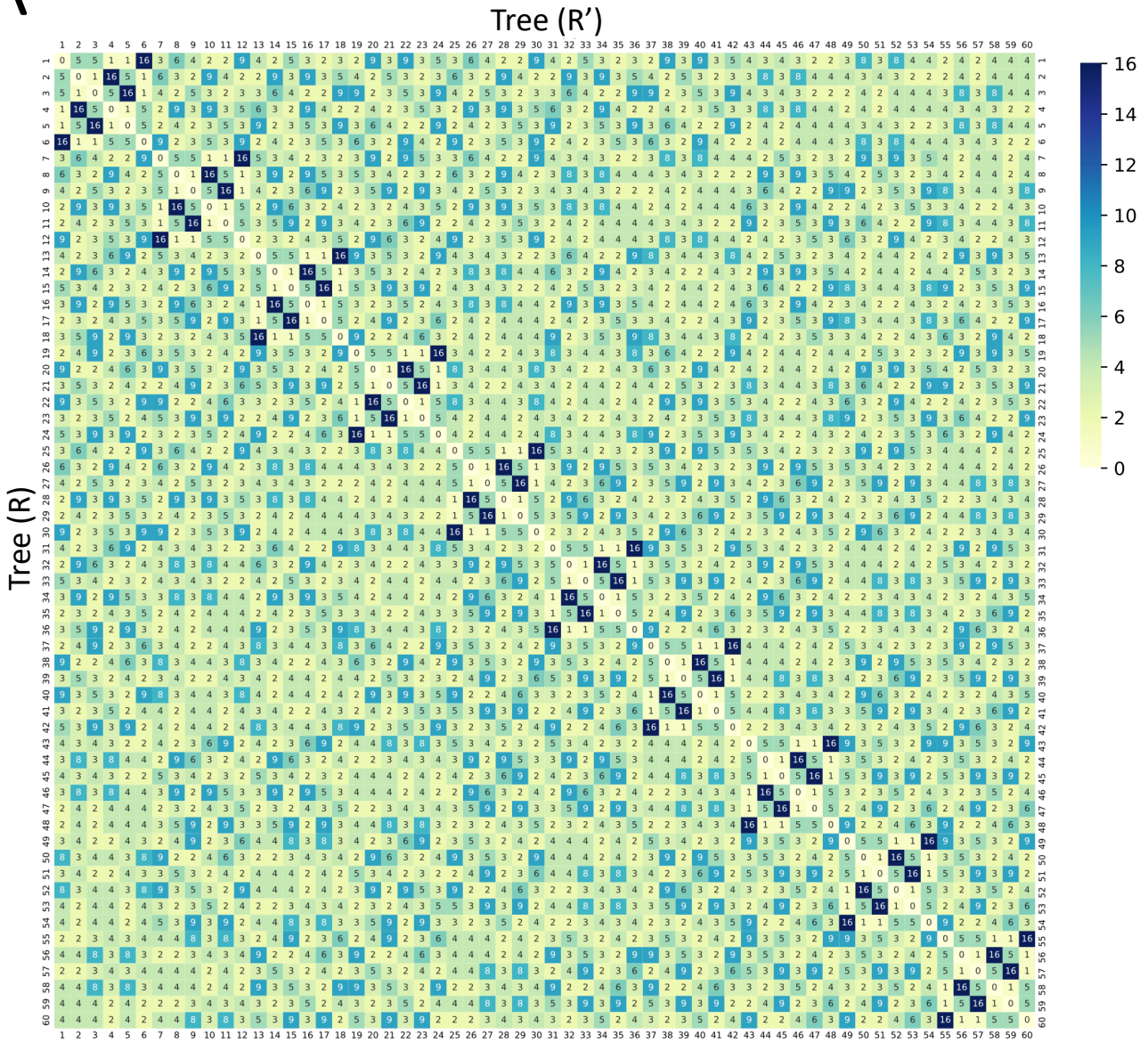
- QR-STAR is run with parameters C=1e-02 and $\frac{\alpha}{\beta} = 0$.

Key Idea behind QR-STAR

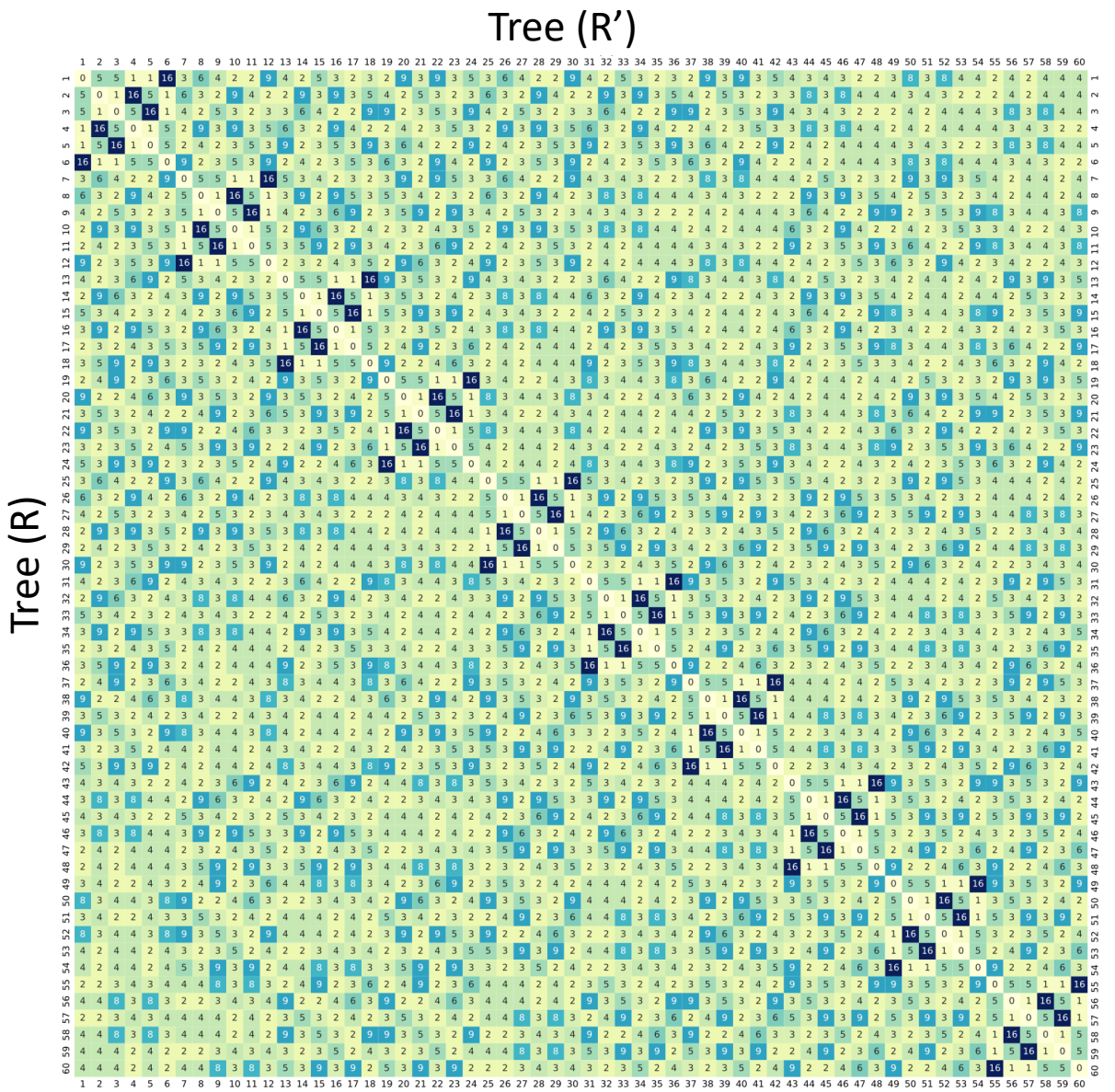
- No zeros except on the diagonal
- Pairs of trees with the same rooted topological shape (caterpillar, balanced, pseudo-caterpillar) always have conflicting distributions

- Idea:
  - Determine the topological shape of each quintet
  - Incorporate the topological shape in the cost function

Heatmap of the number of conflicts between all pairs of caterpillar trees



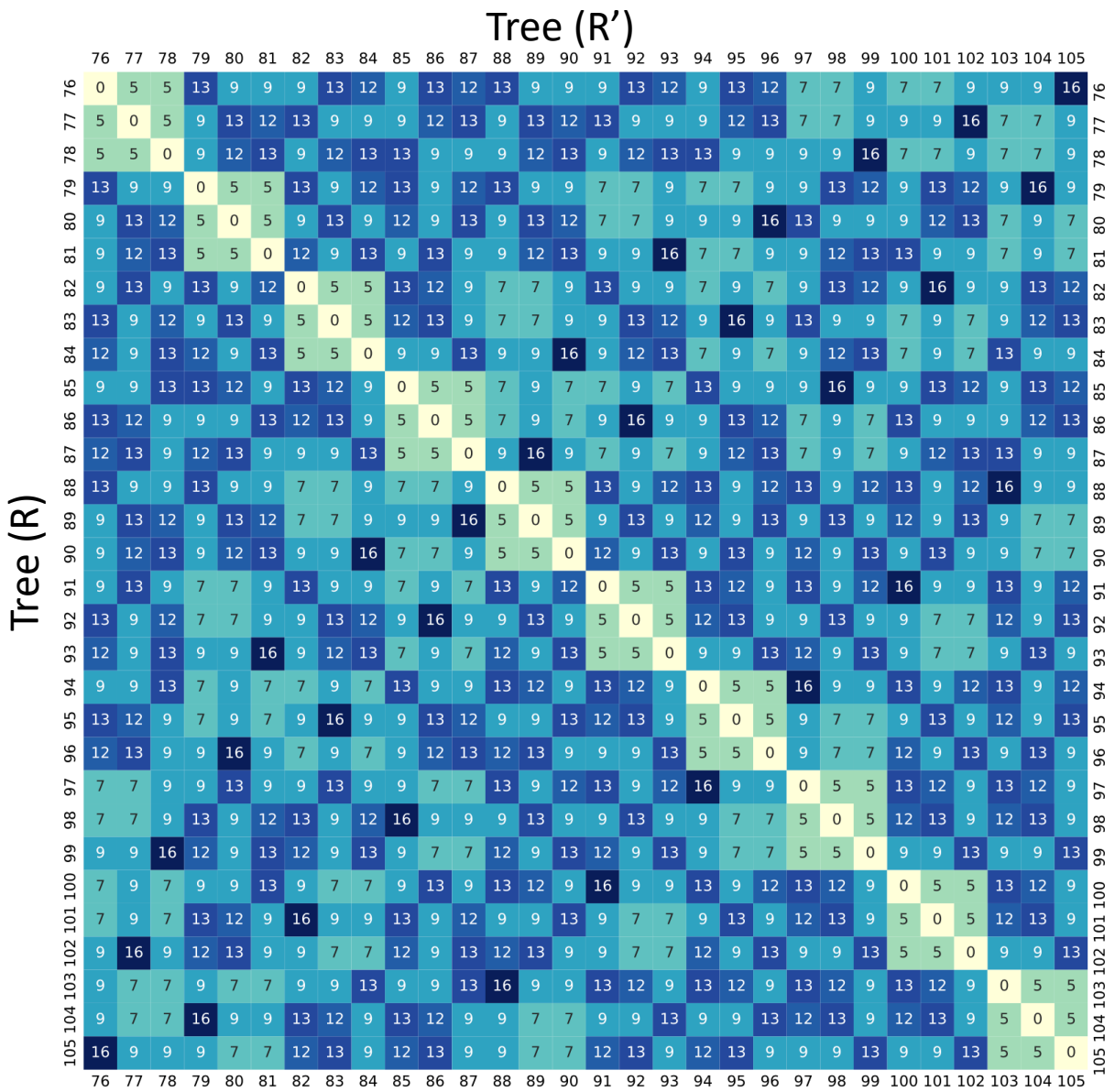
Conflicts between pairs of 5-taxon caterpillar trees



No zeros except on the diagonal

Heatmap showing the number of conflicting inequality penalty terms (the function $|V(R, R')|$) for pairs of caterpillar 5-taxon rooted trees.

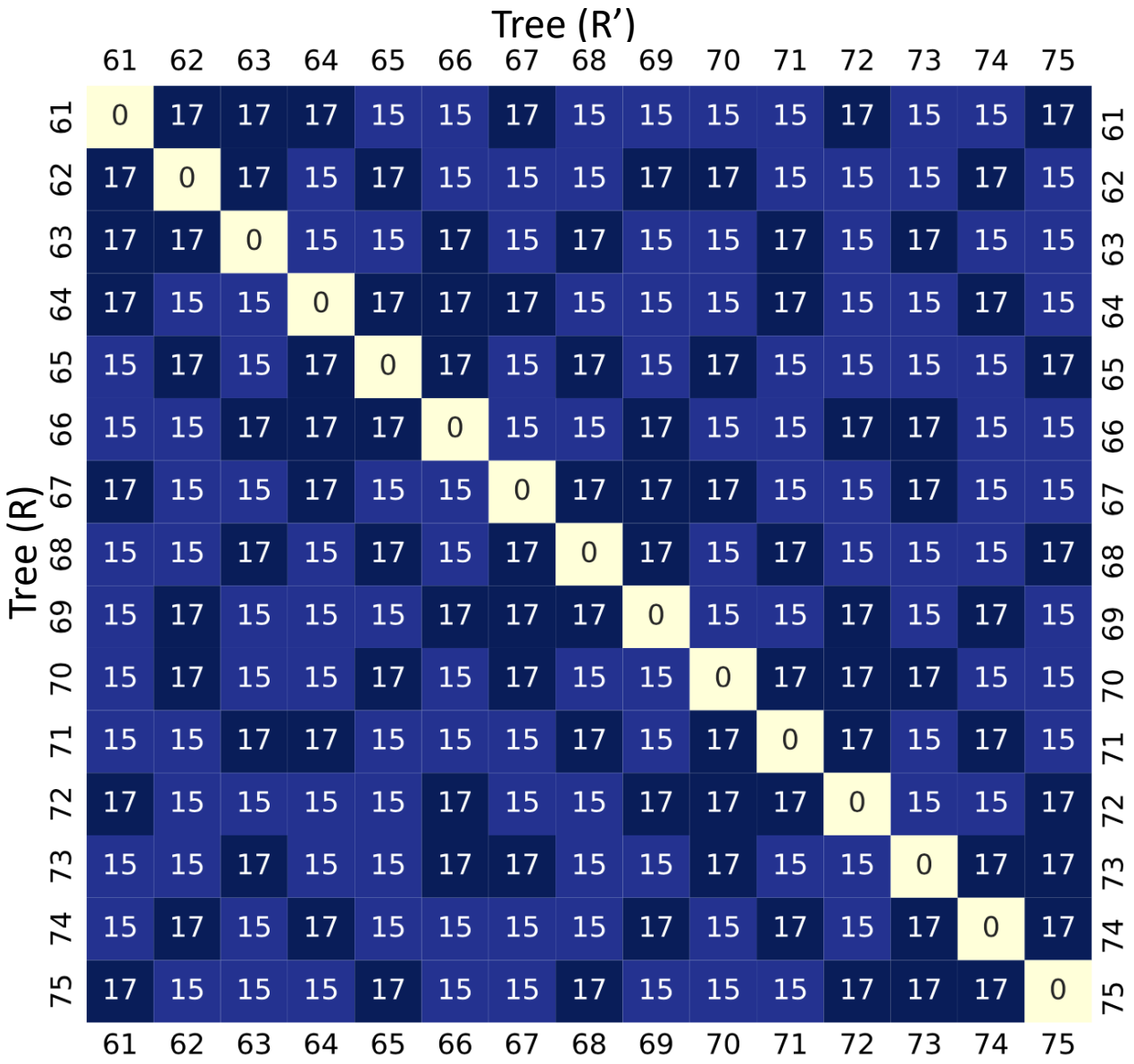
Conflicts between pairs of 5-taxon balanced trees



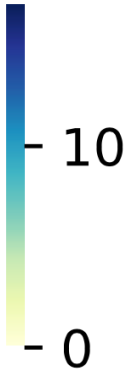
No zeros except on the diagonal

Heatmap showing the number of conflicting inequality penalty terms (the function $|V(R, R')|$) for pairs of balanced 5-taxon rooted trees.

Conflicts between pairs of 5-taxon pseudo-caterpillar trees



No zeros except on the diagonal



Heatmap showing the number of conflicting inequality penalty terms (the function $|V(R, R')|$) for pairs of pseudo-caterpillar 5-taxon rooted trees.