Anchoring quartet-based phylogenetic distances and applications to species tree reconstruction

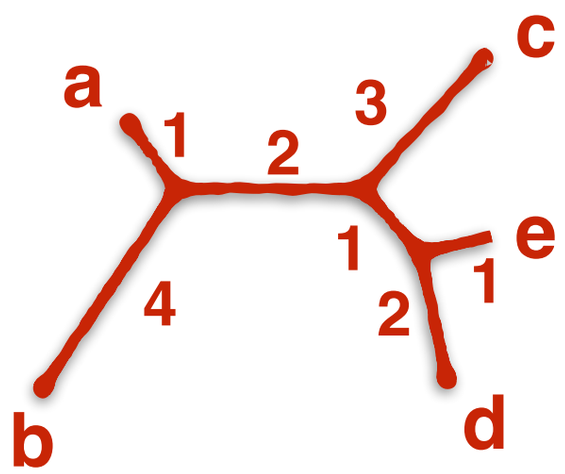
Erfan Sayyari and Siavash Mirarab
Department of Electrical and Computer Engineering
University of California at San Diego

Contributions

- Theoretical:
  A surprisingly “correct” way of computing distances between two leaves on a tree

- Practical:
  Using the new distance definition to reconstruct species trees from gene trees (DISTIQUE)

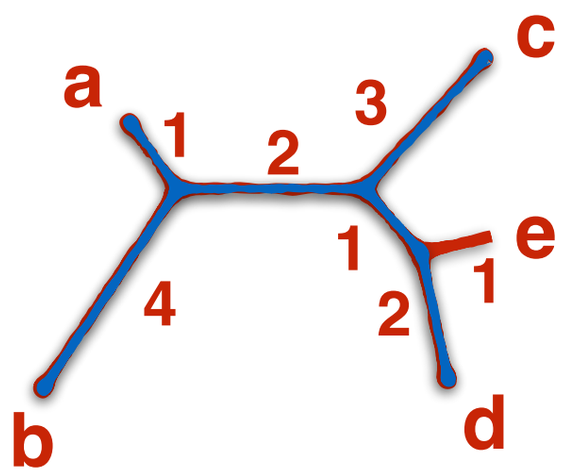
Phylogenetics 101: Distances



	a	b	c	d	e
a	0				
b	5	0			
c	6	9	0		
d	6	9	6	0	
e	4	8	5	3	0

- Additive matrices correspond to trees

Phylogenetics 101: Distances



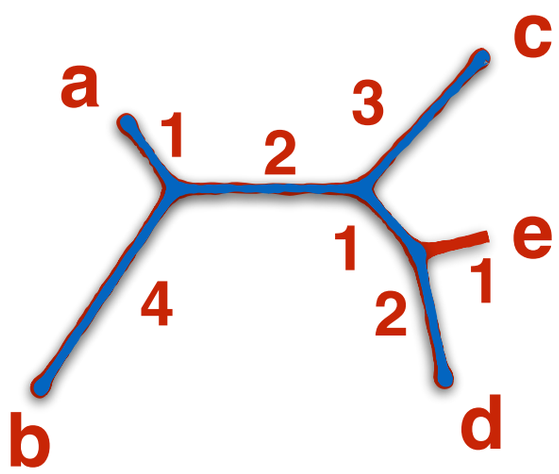
	a	b	c	d	e
a	0				
b	5	0			
c	6	9	0		
d	6	9	6	0	
e	4	8	5	3	0

- Additive matrices correspond to trees

- Test additivity using four point condition for all quartets of tips:

$$D(a,b) + D(c,d) \quad < \quad D(a,c) + D(b,d)$$
$$= \quad D(a,d) + D(b,c)$$

Phylogenetics 101: Distances



	a	b	c	d	e
a	0				
b	5	0			
c	6	9	0		
d	6	9	6	0	
e	4	8	5	3	0

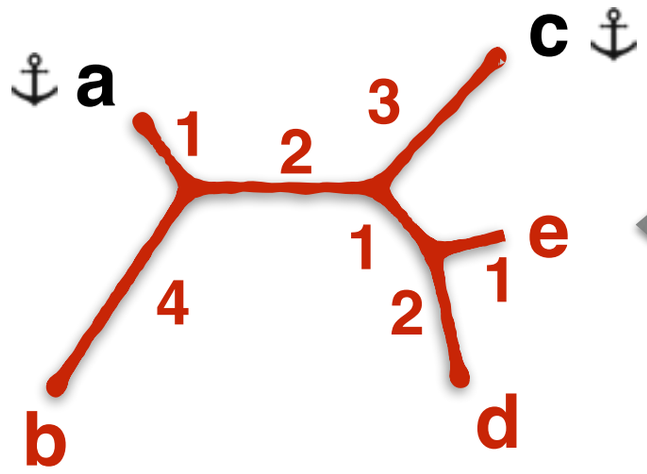
- Additive matrices correspond to trees

- Test additivity using four point condition for all quartets of tips:

$$D(a,b) + D(c,d) \quad < \quad D(a,c) + D(b,d)$$
$$= \quad D(a,d) + D(b,c)$$

- Distance methods (e.g., minimum evolution, Neighbor joining,…) can reconstruct trees correctly from (nearly) additive matrices

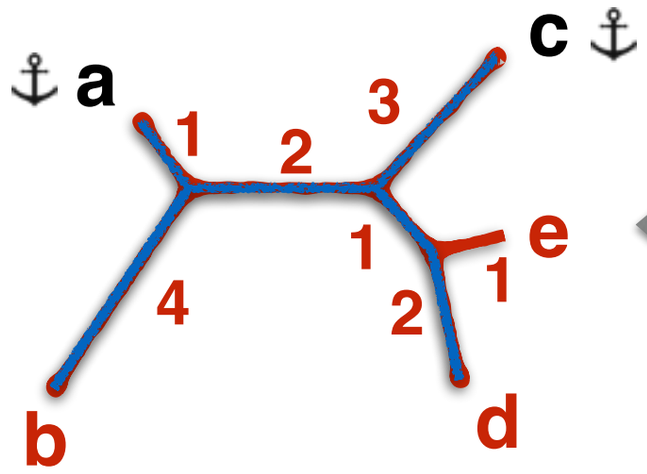
Contribution 1: Anchored distances



	a	b	c	d	e
a	-	-	-	-	-
b	-	0			
c	-	-	-	-	-
d	-	?	-	0	-
e	-	?	-	?	0

- Fix two arbitrary anchors

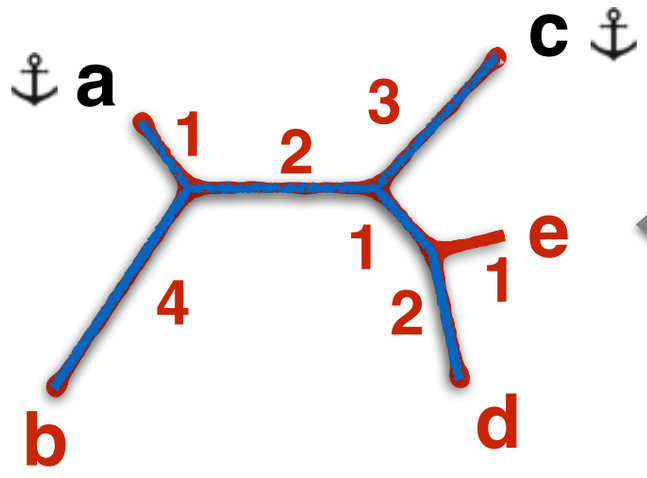
Contribution 1: Anchored distances



	a	b	c	d	e
a	-	-	-	-	-
b	-	0			
c	-	-	-	-	-
d	-	?	-	0	-
e	-	?	-	?	0

- Fix two arbitrary anchors
- Set the distances between *other* pairs of leaves based on the quartet they define together with anchors

Contribution 1: Anchored distances



	a	b	c	d	e
a	-	-	-	-	-
b	-	0			
c	-	-	-	-	-
d	-	?	-	0	-
e	-	?	-	?	0

- Fix two arbitrary anchors
- Set the distances between *other* pairs of leaves based on the quartet they define together with anchors

- For two tips u and v, we define:

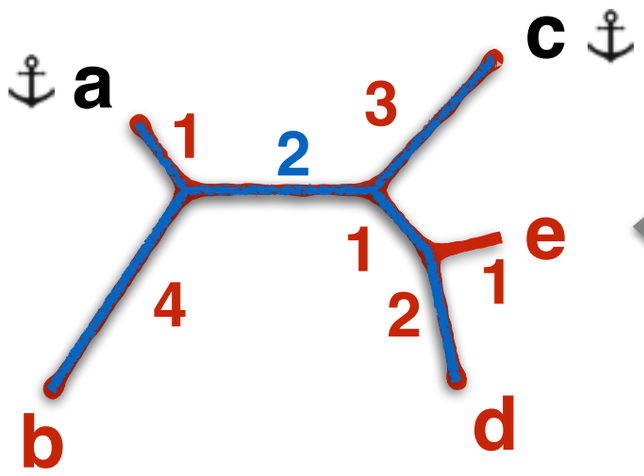
any >0 constant

— $D(u,v) = ß + \tau(ac.uv)$ *if anchors are not together*

— $D(u,v) = ß - f(\tau(ac.uv))$ *if anchors are together*

Any positive monotonically increasing function bounded above by $ß$

τ: Length of the internal branch in ab.uv quartet



	a	b	c	d	e
a	-	-	-	-	-
b	-	0			
c	-	-	-	-	-
d	-	4	-	0	-
e	-	?	-	?	0

- — Fix two arbitrary anchors —> **a** and **c**
  — Distance between **b** and **d**

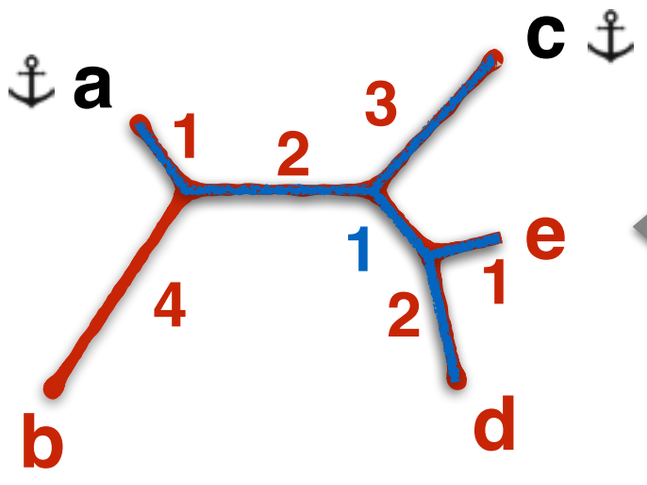
- $ß=2$

— $D(u,v) = ß + \tau(ac.bd)$ *if anchors are not together*

— $D(u,v) = ß - f(\tau(ac.bd))$ *if anchors are together*

$f(x) = 2 - 2^{-x}$

$\tau(ac.bd) = 2$



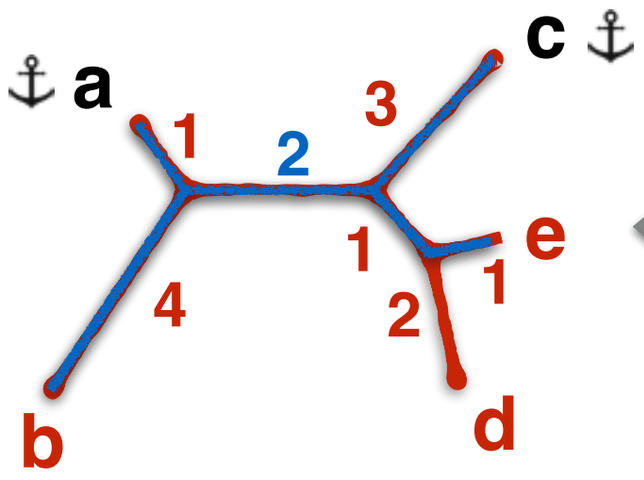
	a	b	c	d	e
a	-	-	-	-	-
b	-	0			
c	-	-	-	-	-
d	-	4	-	0	-
e	-	?	-	0.5	0

- — Fix two arbitrary anchors —> **a** and **c**
  — Distance between **d** and **e**

- $ß=2$
  - $D(u,v) = ß + \tau(ac.de)$ *if anchors are not together*
  - $D(u,v) = ß - f(\tau(ac.de))$ *if anchors are together*

$f(x) = 2 - 2^{-x}$

$\tau(ac.de) = 1$



	a	b	c	d	e
a	-	-	-	-	-
b	-	0			
c	-	-	-	-	-
d	-	4	-	0	-
e	-	4	-	0.5	0

- — Fix two arbitrary anchors —> **a** and **c**
  — Distance between **b** and **e**

- $ß=2$

  — $D(u,v) = ß + \tau(ac.be)$ *if anchors are not together*

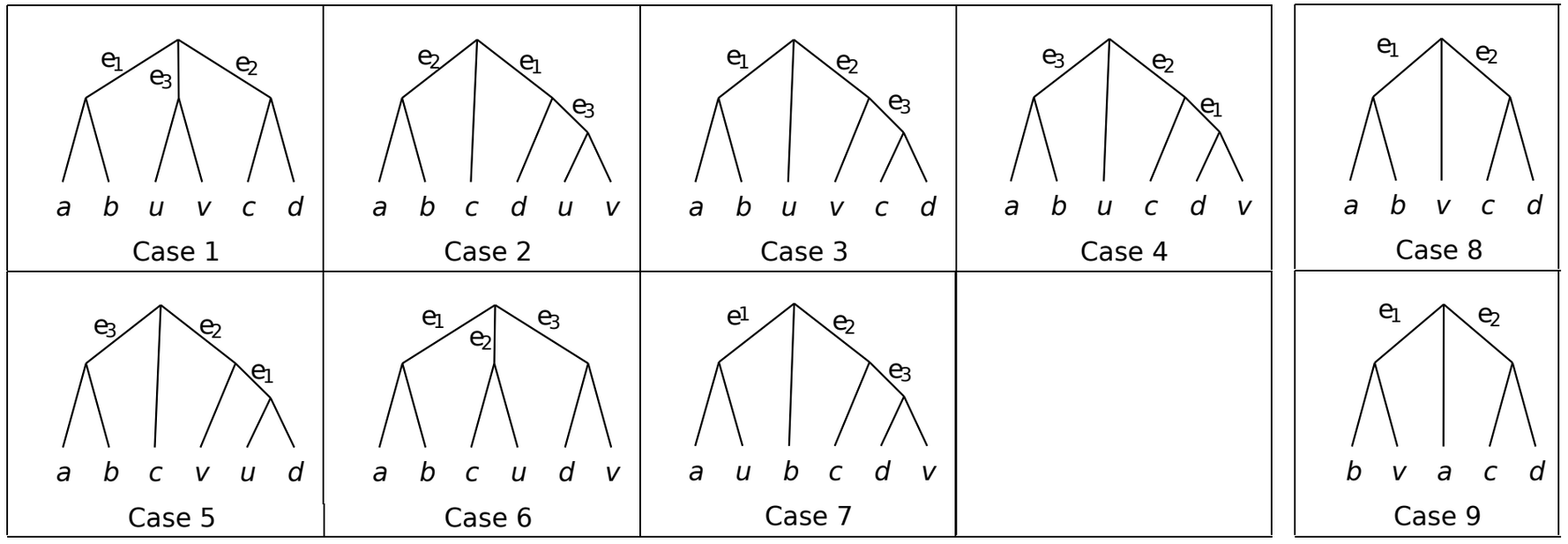
  — $D(u,v) = ß - f(\tau(ac.be))$ *if anchors are together*

  $f(x) = 2 - 2^{-x}$

  $\tau(ac.be) = 2$

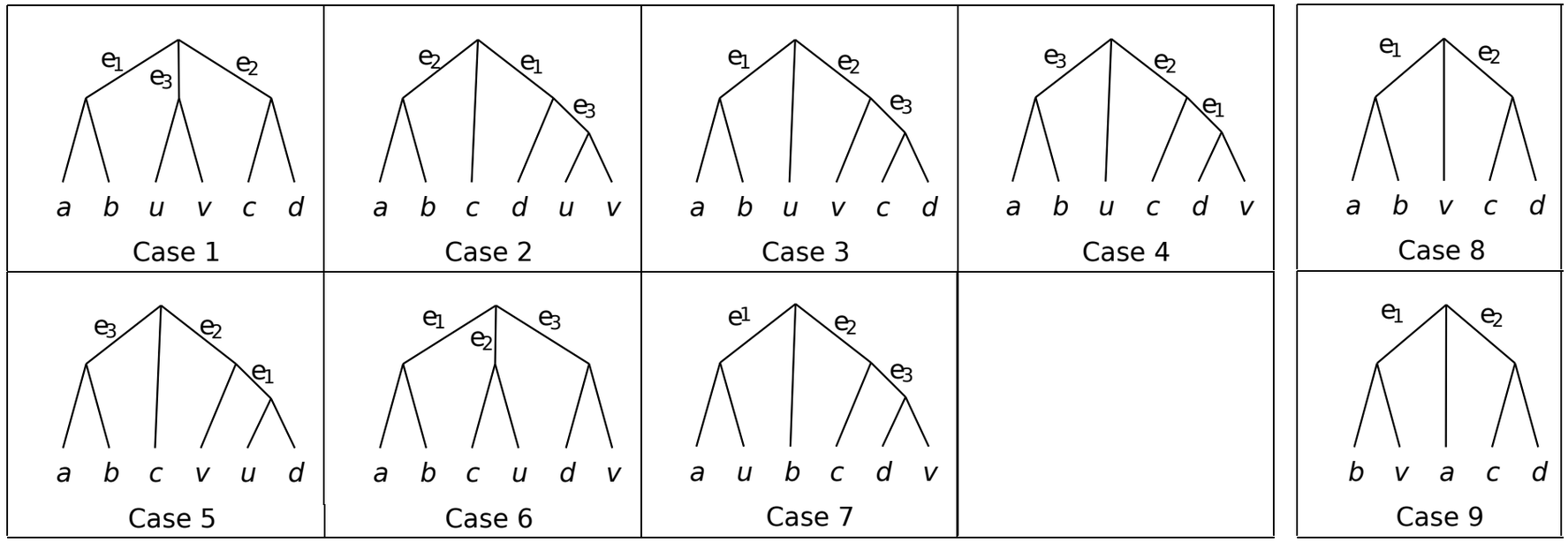
Main theoretical results

Anchored distances are additive and correspond to a tree that is identical to the original tree with respect to the topology but not branch lengths



Main theoretical results

Anchored distances are additive and correspond to a tree that is identical to the original tree with respect to the topology but not branch lengths



Proof

Using anchored distances

- To compute the distance between two leaves from data:
  - Pick two anchors
  - Compute all $O(n^2)$ quartet trees that include anchors (topology and internal branch length) from data
  - Computed anchored distances from quartet trees

Using anchored distances

- To compute the distance between two leaves from data:
  - Pick two anchors
  - Compute all $O(n^2)$ quartet trees that include anchors (topology and internal branch length) from data
  - Computed anchored distances from quartet trees

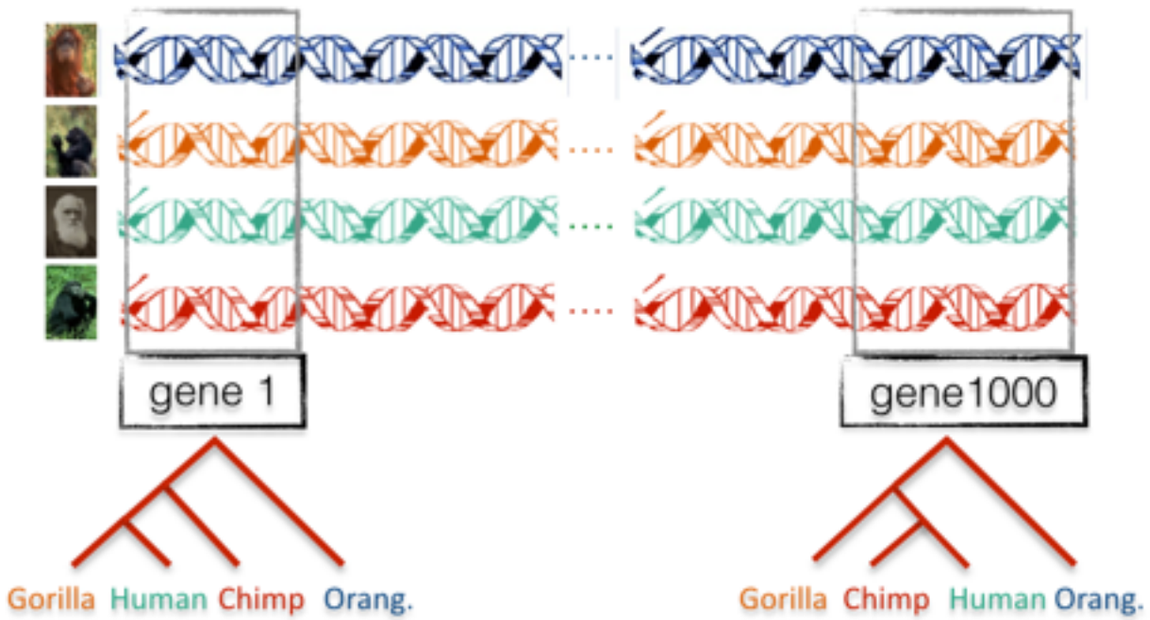
- With a statistically consistent quartet tree estimator, running a distance-based method (with a safety radius) on anchored distances gives an statistically consistent estimate of the tree

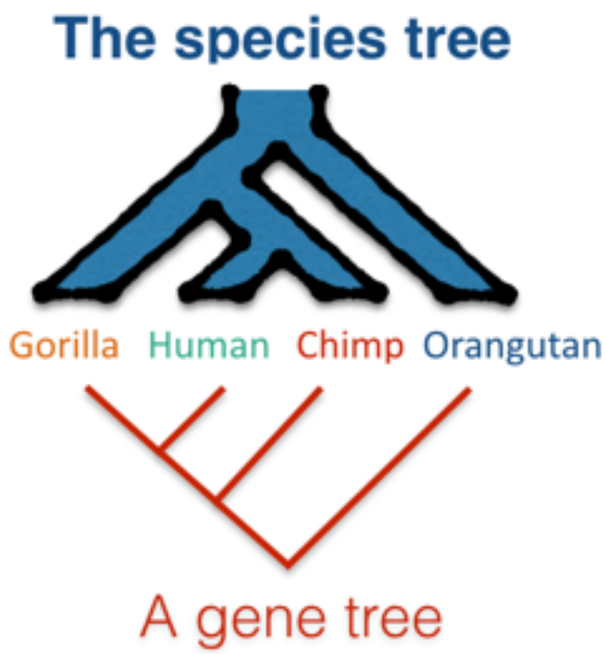
So what?

- Why not just compute pairwise distances directly?
  - Sometimes computing quartet trees is more straight-forward

So what?

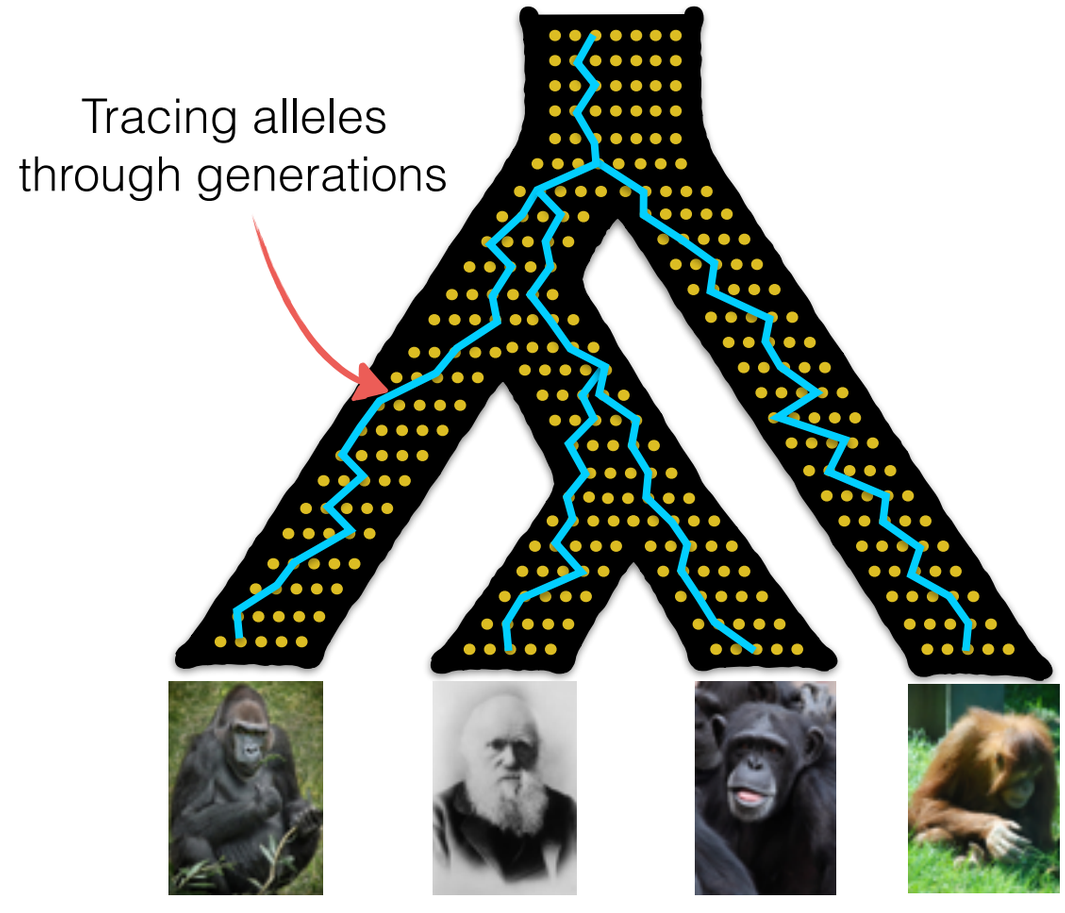
- Why not just compute pairwise distances directly?
  - Sometimes computing quartet trees is more straight-forward
- Example: genome-scale species tree estimation from gene trees according to the multi-species coalescent model (of ILS)





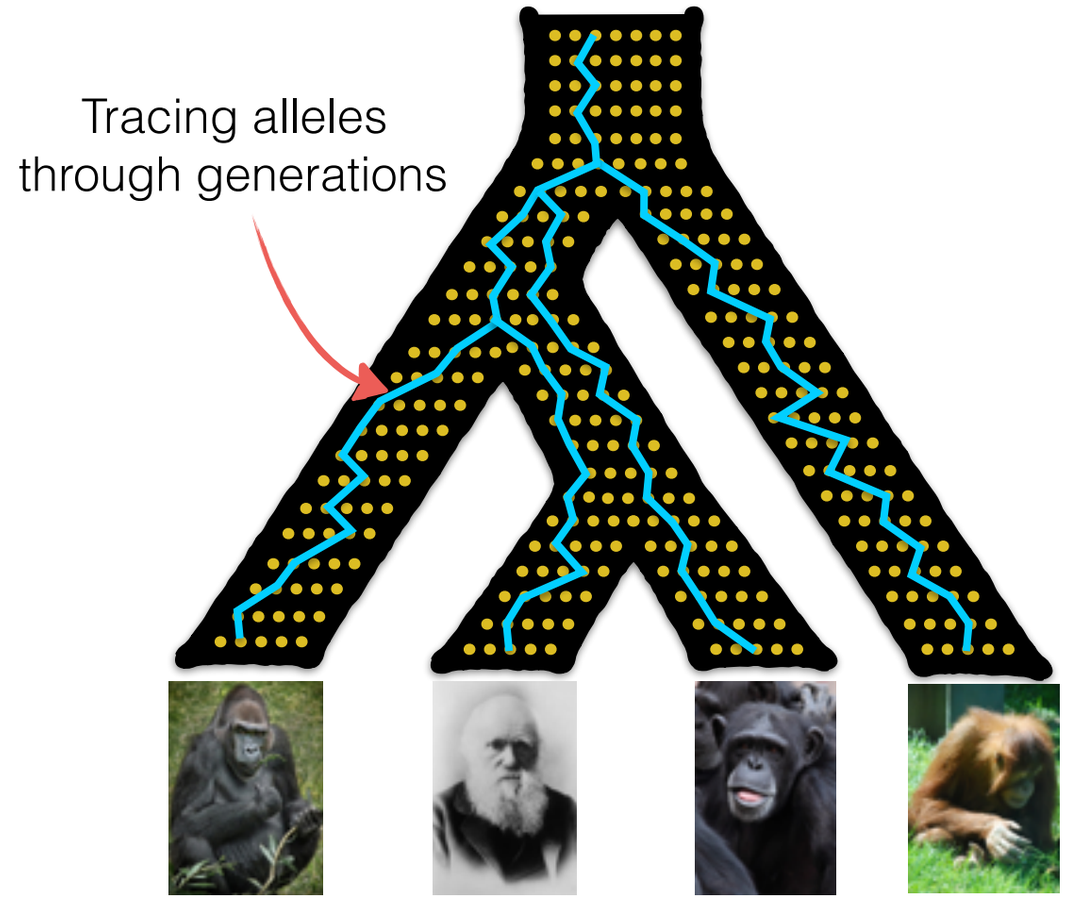
Incomplete Lineage Sorting (ILS)

- A random process related to the coalescence of alleles across various populations



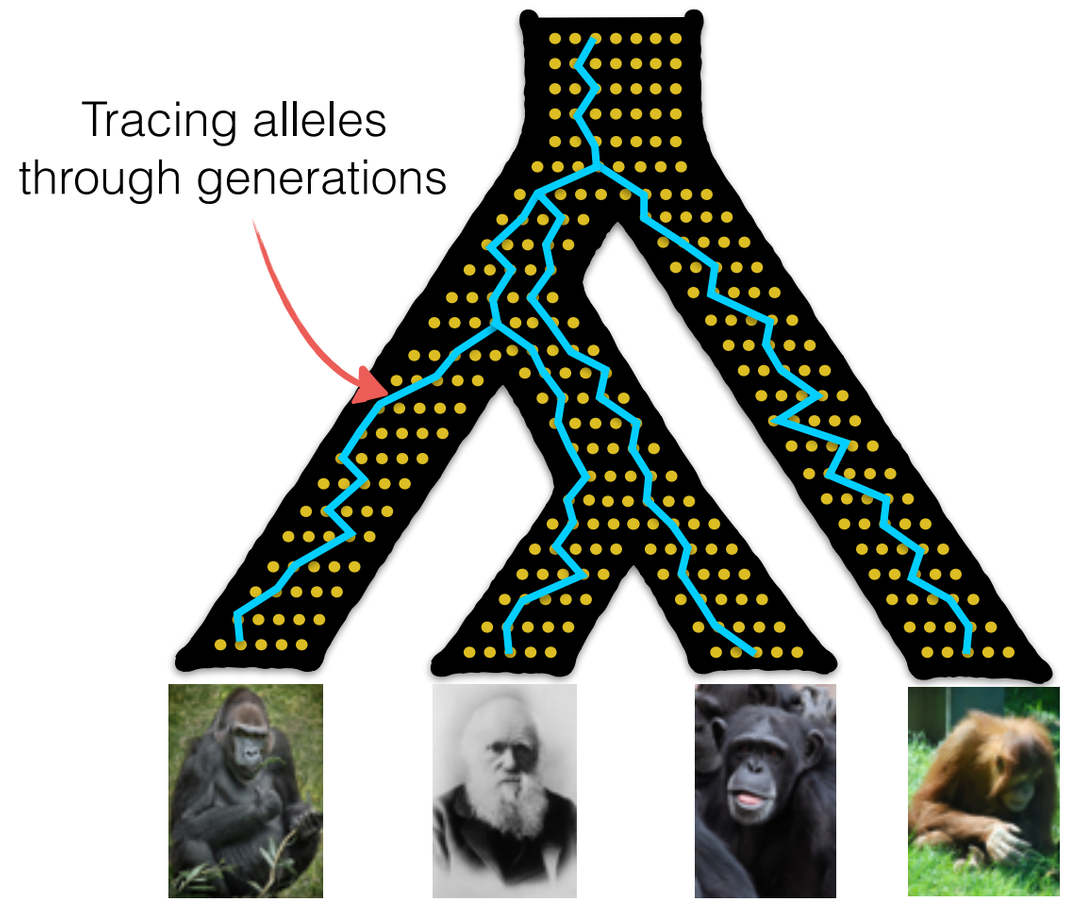
Incomplete Lineage Sorting (ILS)

- A random process related to the coalescence of alleles across various populations



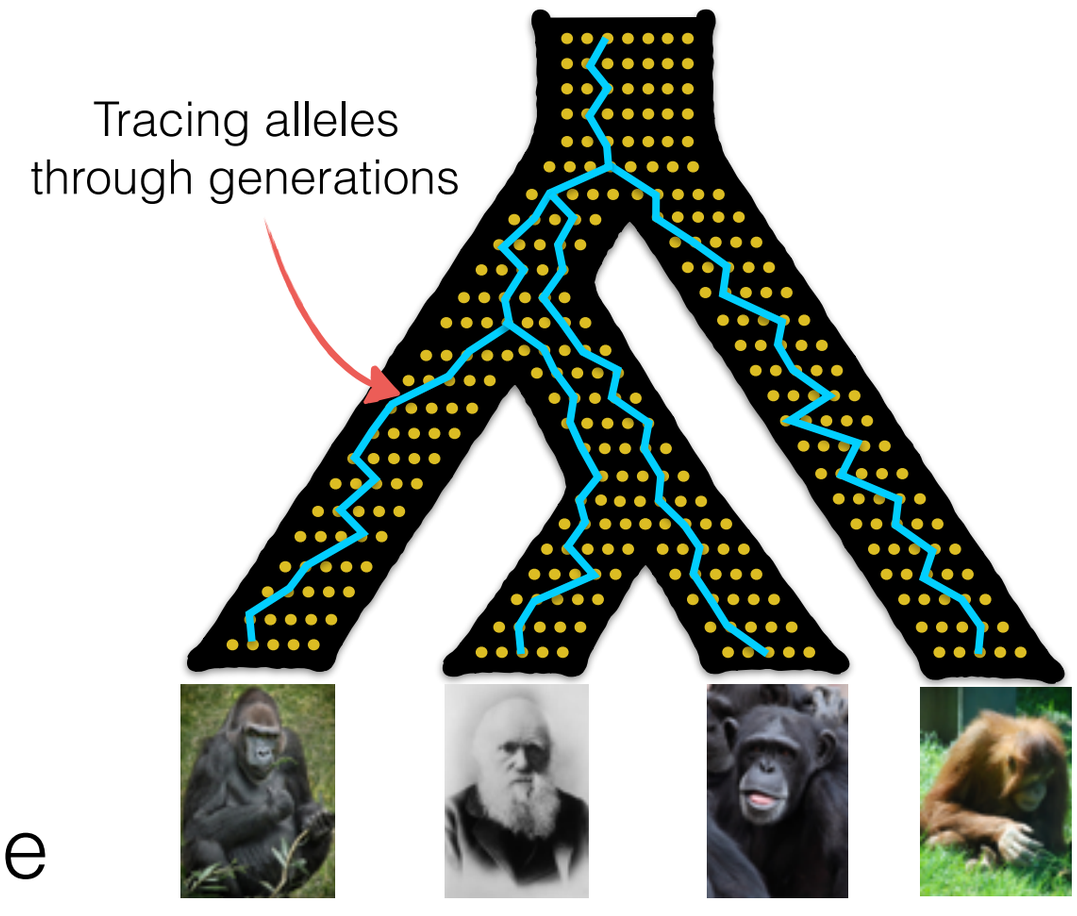
Incomplete Lineage Sorting (ILS)

- A random process related to the coalescence of alleles across various populations
- Modeled by the multi-species coalescent (MSC) model



Incomplete Lineage Sorting (ILS)

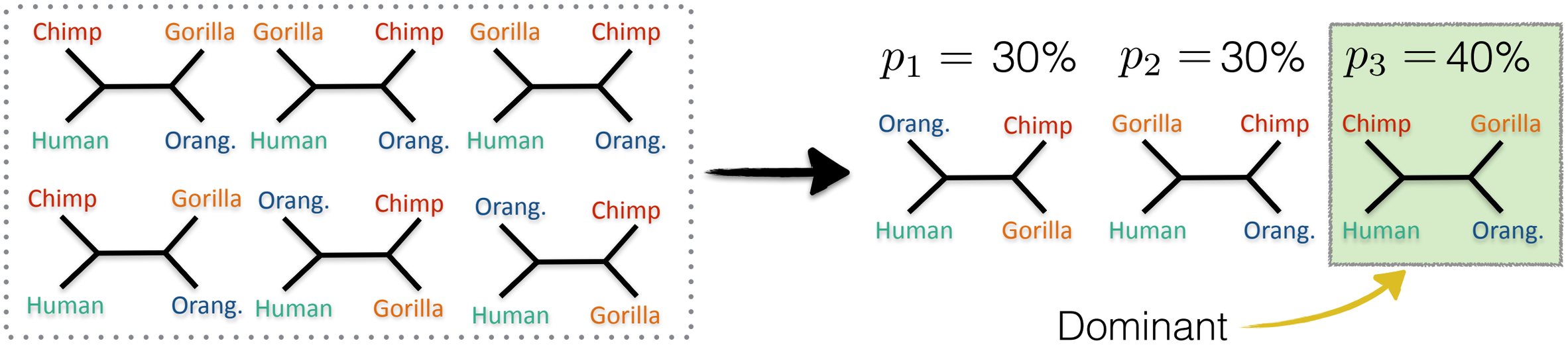
- A random process related to the coalescence of alleles across various populations



- Modeled by the multi-species coalescent (MSC) model

- The species tree defines the probability distribution on gene trees, and is identifiable from the distribution on gene trees
[Degnan and Salter, 2005]

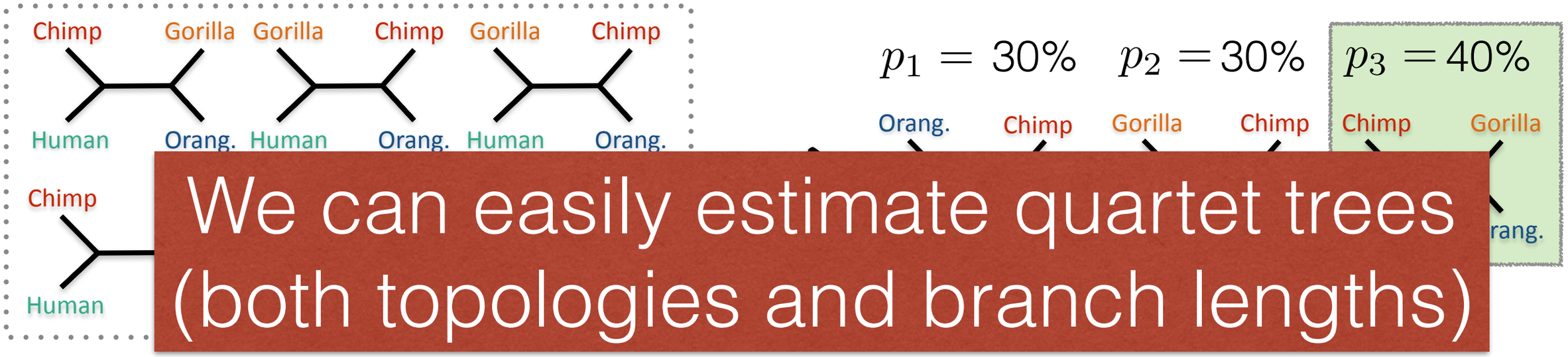
Properties of quartet trees in presence of ILS



- For 4 species, the dominant quartet topology is the species tree
  [Allman, et al. 2010]

- $P(\text{gene tree} = \text{species tree}) = 1 - 2/3\ e^{-d}$
  $P(\text{gene tree} \neq \text{species tree}) = 1/3\ e^{-d}$

Properties of quartet trees in presence of ILS



We can easily estimate quartet trees (both topologies and branch lengths)

- For 4 species, the dominant quartet topology is the species tree [Allman, et al. 2010]
- $P(\text{gene tree} = \text{species tree}) = 1 - 2/3\ e^{-d}$
  $P(\text{gene tree} \neq \text{species tree}) = 1/3\ e^{-d}$

Anchored distances for the MSC model

- Input: a set of gene trees
  Quartet estimator: expected value under the MSC model

Anchored distances for the MSC model

- Input: a set of gene trees
  Quartet estimator: expected value under the MSC model

$$D_{u,v}[a,b] = -ln(p(ab.uv))$$

The frequency of a quartet tree in input gene trees
(we have efficient algorithms to compute this)

$\beta = ln(3)$
$f(x) = \ln(3 - 2e^{-x})$

Anchored distances for the MSC model

- Input: a set of gene trees
  Quartet estimator: expected value under the MSC model

$$D_{u,v}[a,b] = -ln(p(ab.uv))$$

The frequency of a quartet tree in input gene trees (we have efficient algorithms to compute this)

- Statistically consistent for the MSC model

$$\beta = ln(3)$$
$$f(x) = \ln(3 - 2e^{-x})$$

DISTIQUE

- **Input**: A set of gene trees
- **Output**: A species tree
- **Approach**:
  Use the MSC-based distance metric together with a distance-based method (by default we use FastME)

Algorithmic details

- How about the two missing anchors?
  Compute distances/trees using multiple anchors and combine

Algorithmic details

- How about the two missing anchors?
  Compute distances/trees using multiple anchors and combine
- How about $p(ab.uv)=0$?
  Use a pseudo-count —> helps but not enough
  - $p=0$ for branches with no discordance —> they are “easy”.
  - Compute a majority consensus of gene trees and then resolve polytomies in the consensus (remains statistically consistent)

Algorithmic details

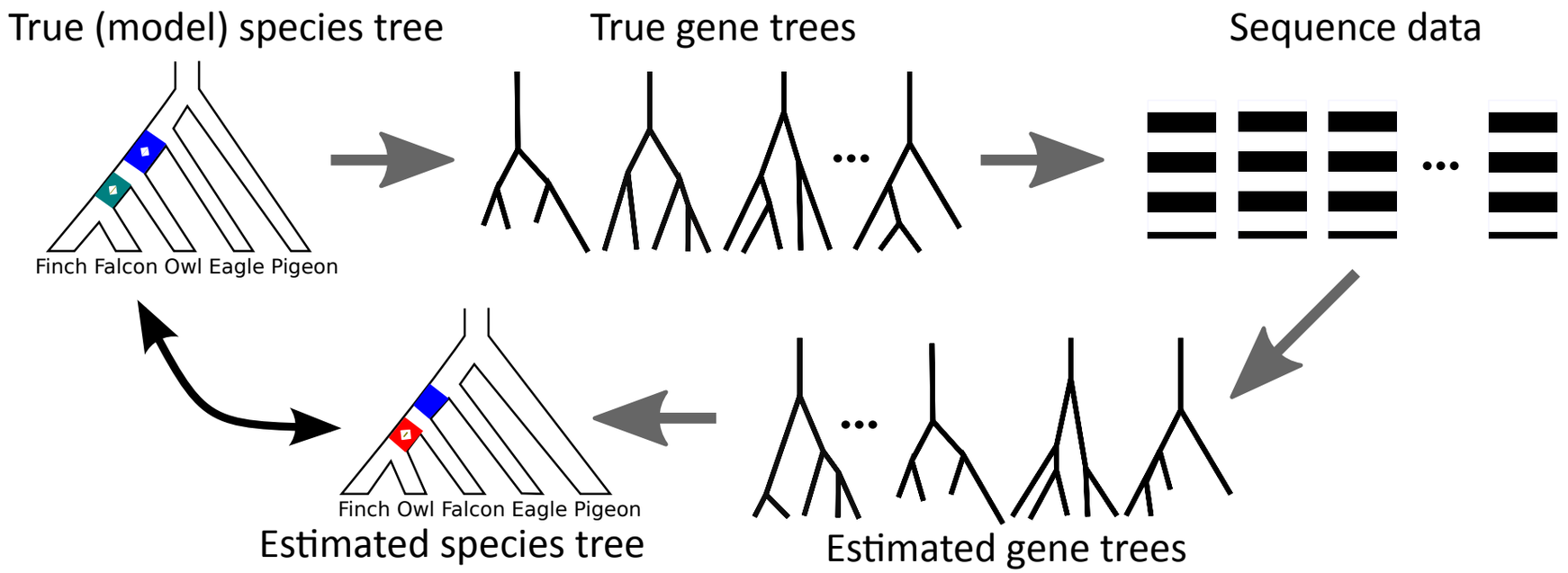
- How about the two missing anchors?
  Compute distances/trees using multiple anchors and combine
- How about $p(ab.uv)=0$?
  Use a pseudo-count —> helps but not enough
  - $p=0$ for branches with no discordance —> they are “easy”.
  - Compute a majority consensus of gene trees and then resolve polytomies in the consensus (remains statistically consistent)
- How many anchor pairs?
  You get to choose, but $O(n)$ seems to work well

Algorithmic details

- How about the two missing anchors?
  Compute distances/trees using multiple anchors and combine
- How about $p(ab.uv)=0$?
  Use a pseudo-count —> helps but not enough
  - $p=0$ for branches with no discordance —> they are “easy”.
  - Compute a majority consensus of gene trees and then resolve polytomies in the consensus (remains statistically consistent)
- How many anchor pairs?
  You get to choose, but $O(n)$ seems to work well
- How do you choose anchor pairs?
  Sample $O(d)$ around each polytomy of degree d.

How well does it work?

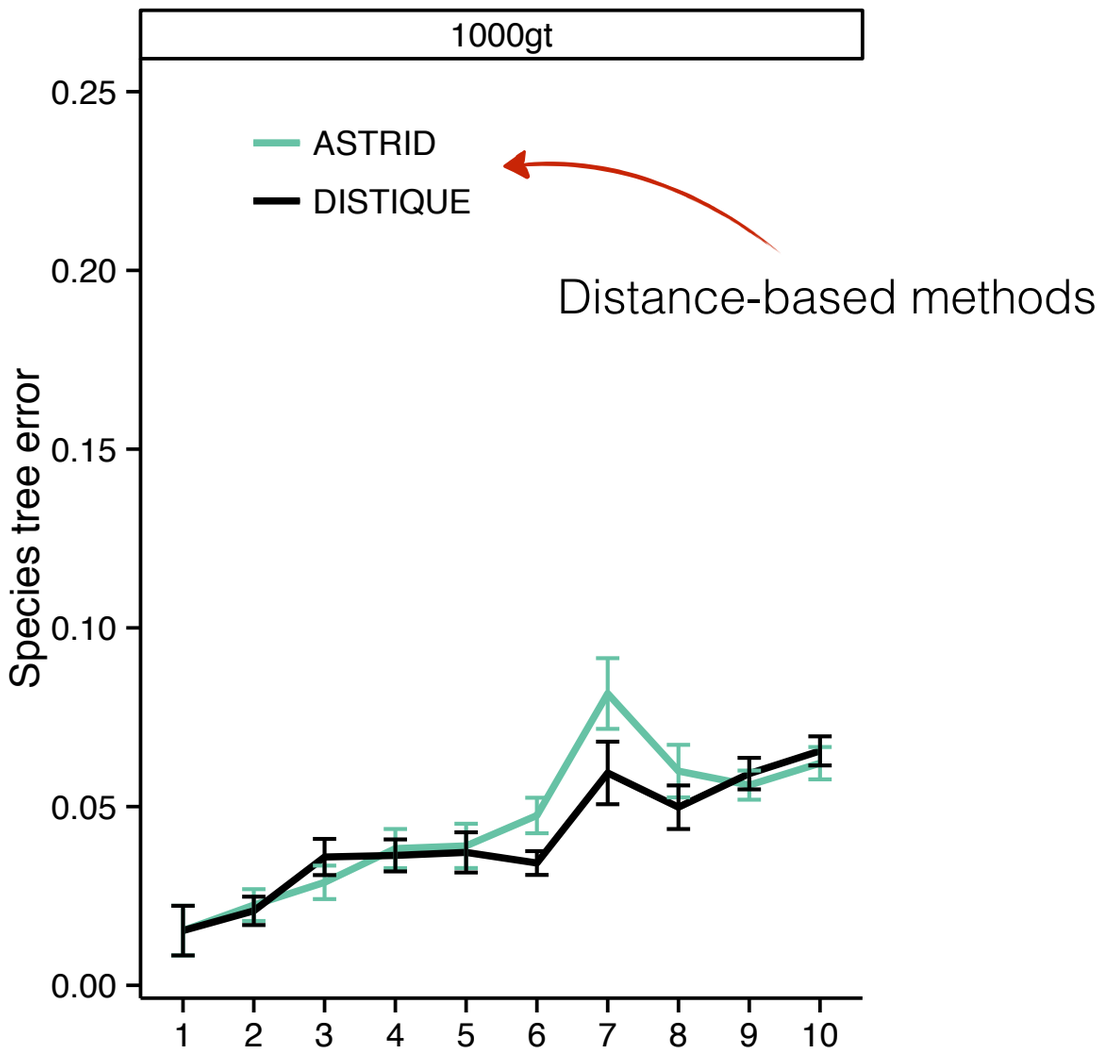
Simulation study



- Datasets:
  - Heterogenous: simphy
  - Homogenous:
    Avian and mammalian
- Estimated gene trees from data

- Compare to:
  NJst (distance-based), ASTRAL (quartet-based), concatenation
- Accuracy measure — FN rate:
  the percentage of branches in the true tree that are missing from the estimated tree

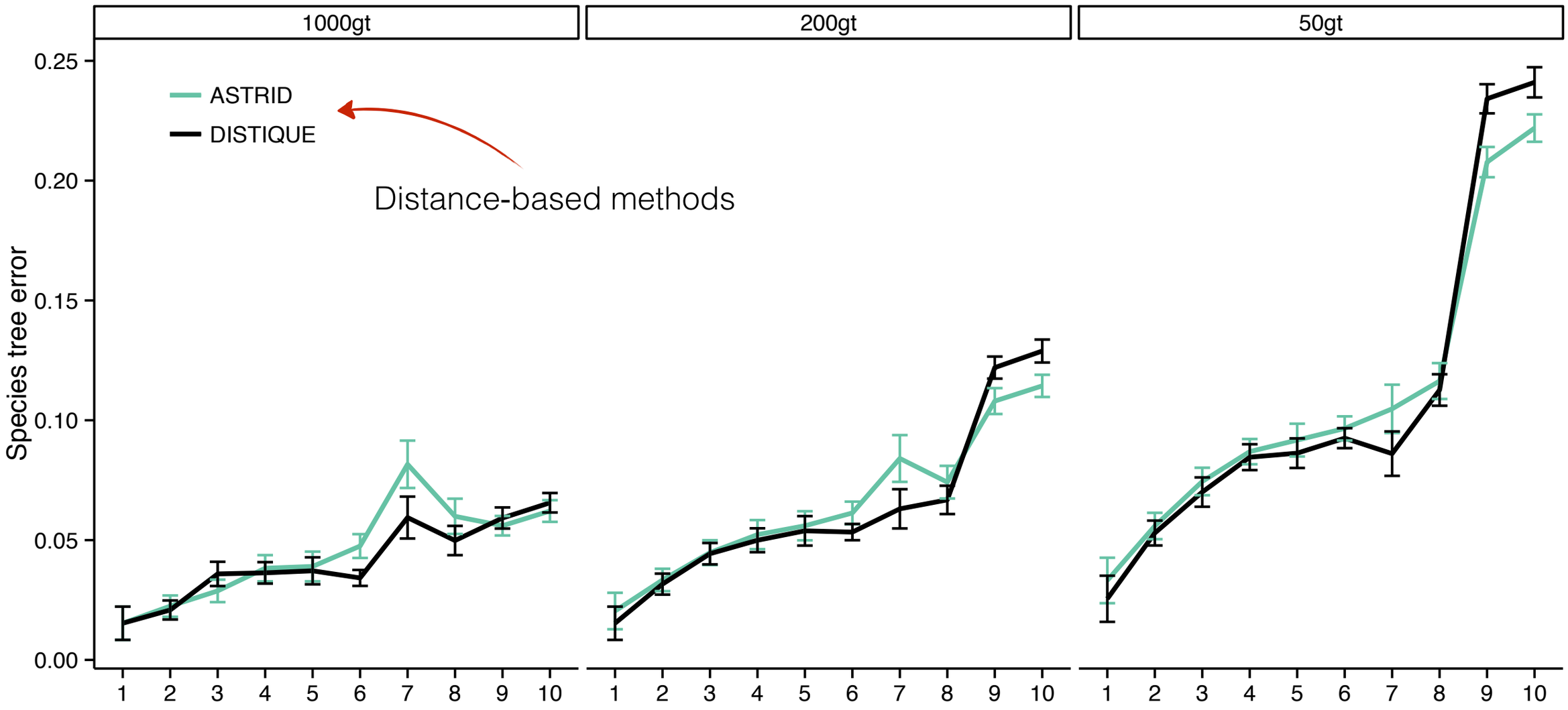
Heterogenous simulations



Distance-based methods

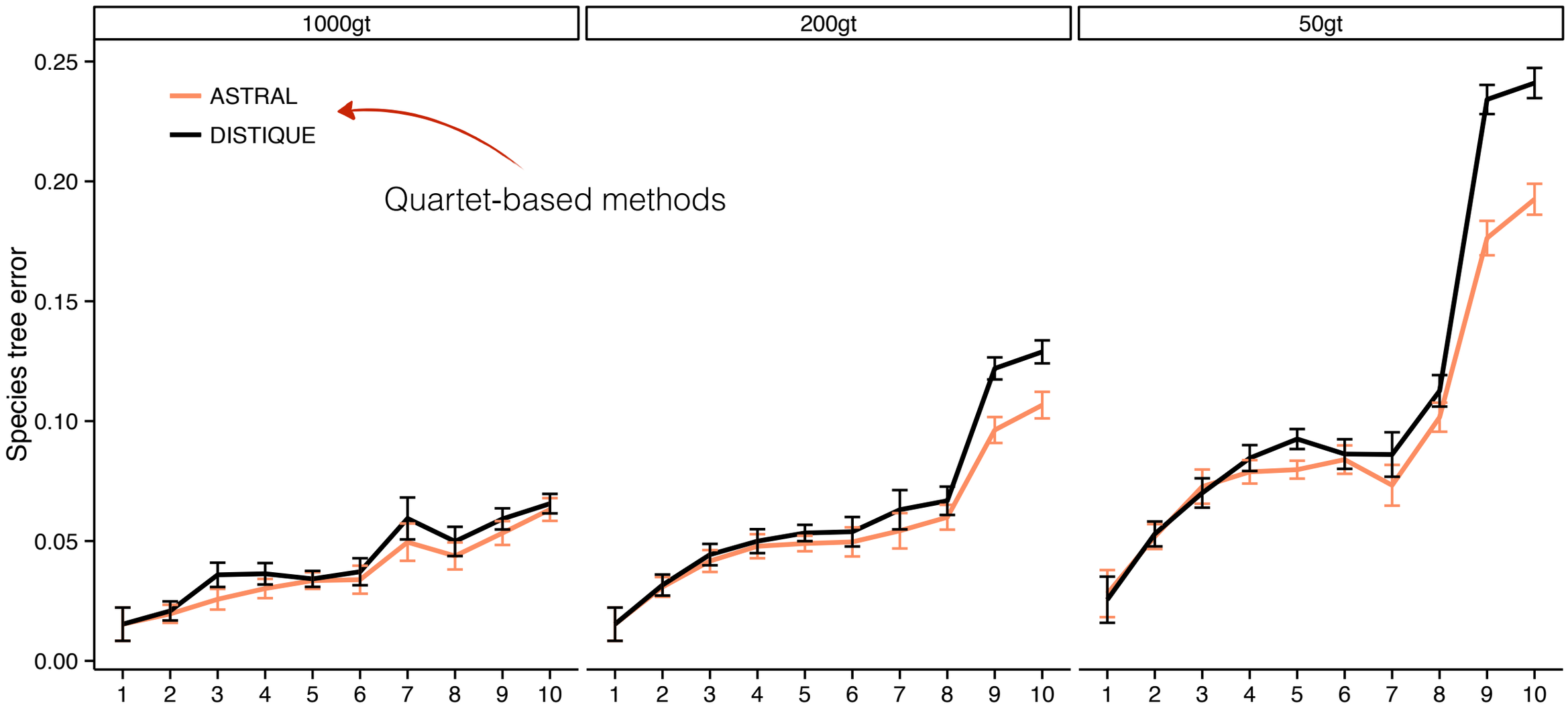
Model conditions: changing
— # taxa (10 to 1000)
— ILS level (low to high) and ILS depth (deep and shallow)

Heterogenous simulations



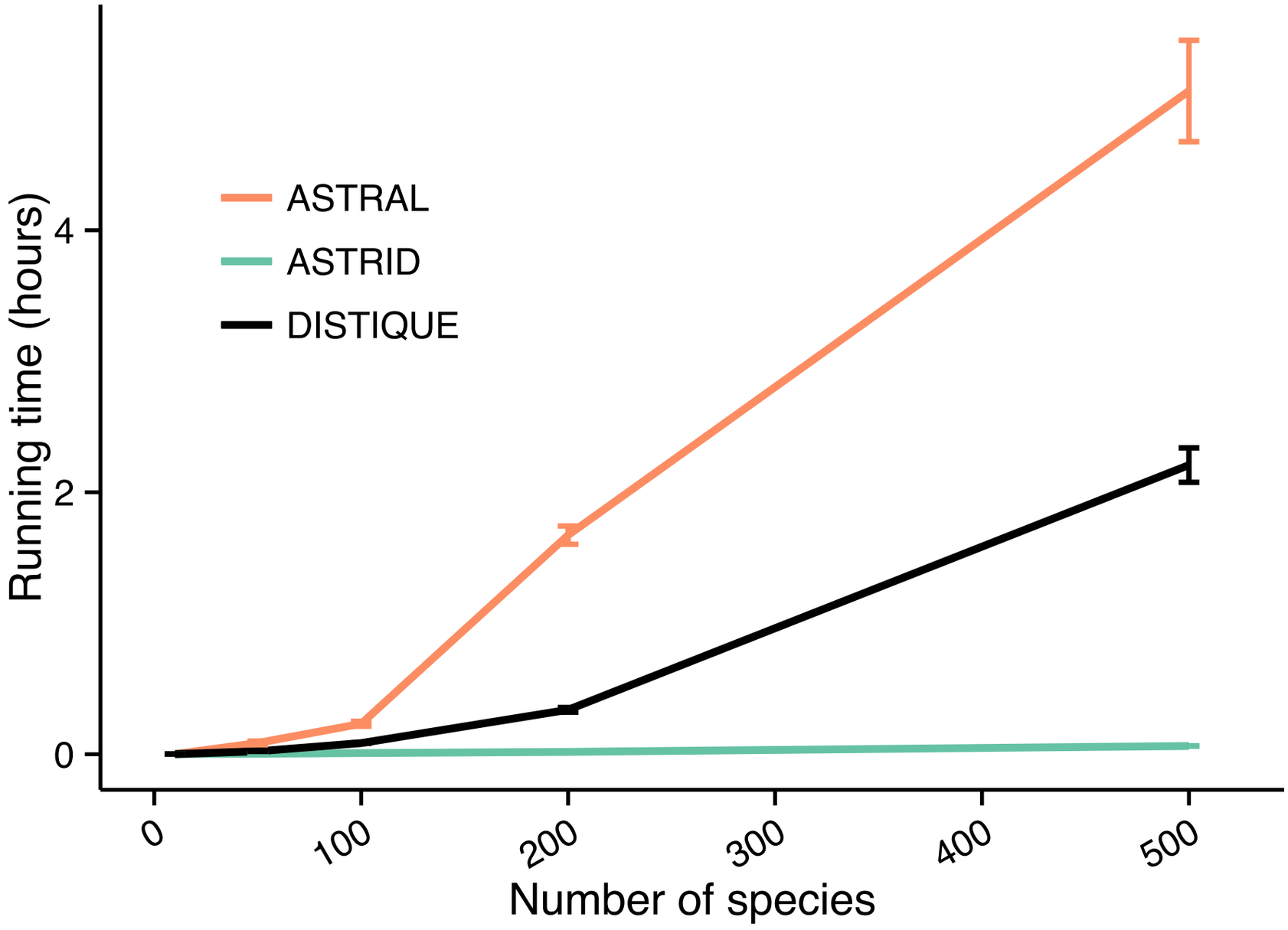
Model conditions: changing
— # taxa (10 to 1000)
— ILS level (low to high) and ILS depth (deep and shallow)

Heterogenous simulations

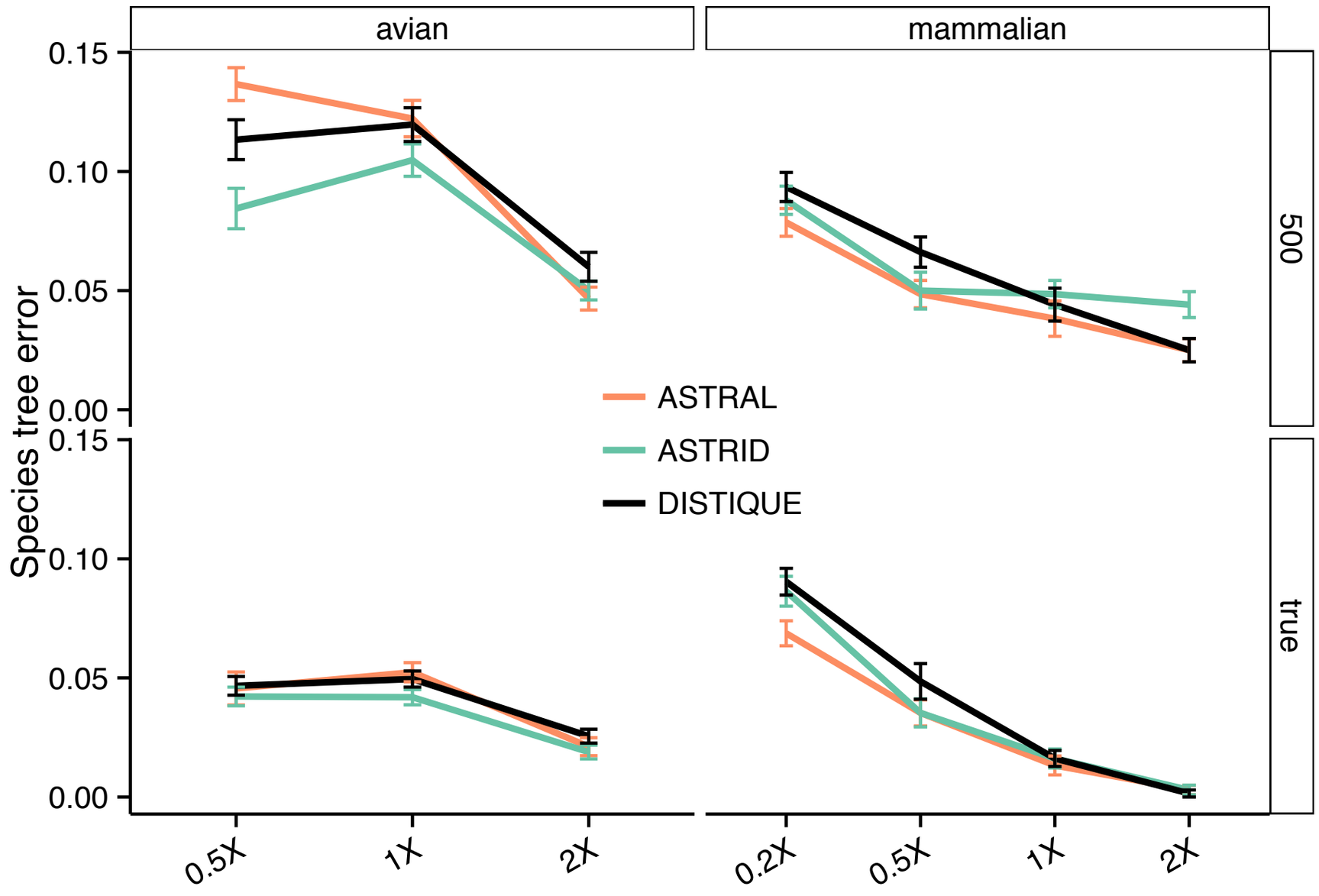


Model conditions: changing
— # taxa (10 to 1000)
— ILS level (low to high) and ILS depth (deep and shallow)

Running time



Homogenous (biological-like) simulations

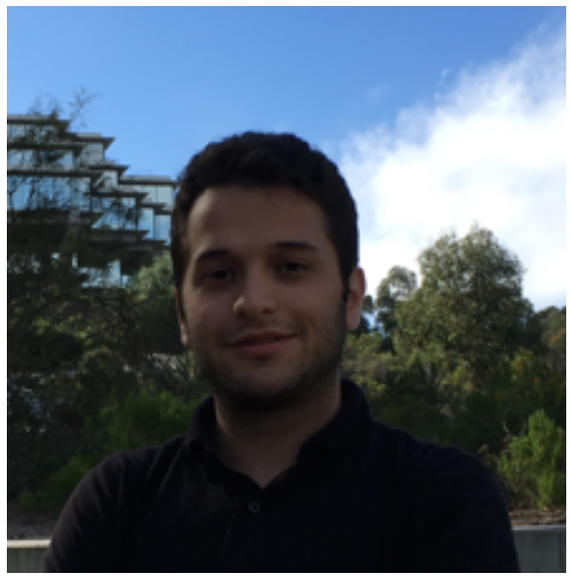


Summary

- Anchored distances:
  - Are computed by estimating quartet trees and transforming their branch length using certain transformations
  - Are (surprisingly) statistically consistent (additive for the correct topology but not branch length)

- Anchored distances can be used to build a ILS-based summary method (DISTIQUE) that is competitive with the best of the current methods
  — Code: https://github.com/esayyari/DISTIQUE
  — Data: http://esayyari.github.io/DISTIQUE.html

Acknowledgment



Erfan Sayyari

XSEDE
Extreme Science and Engineering
Discovery Environment

NSF