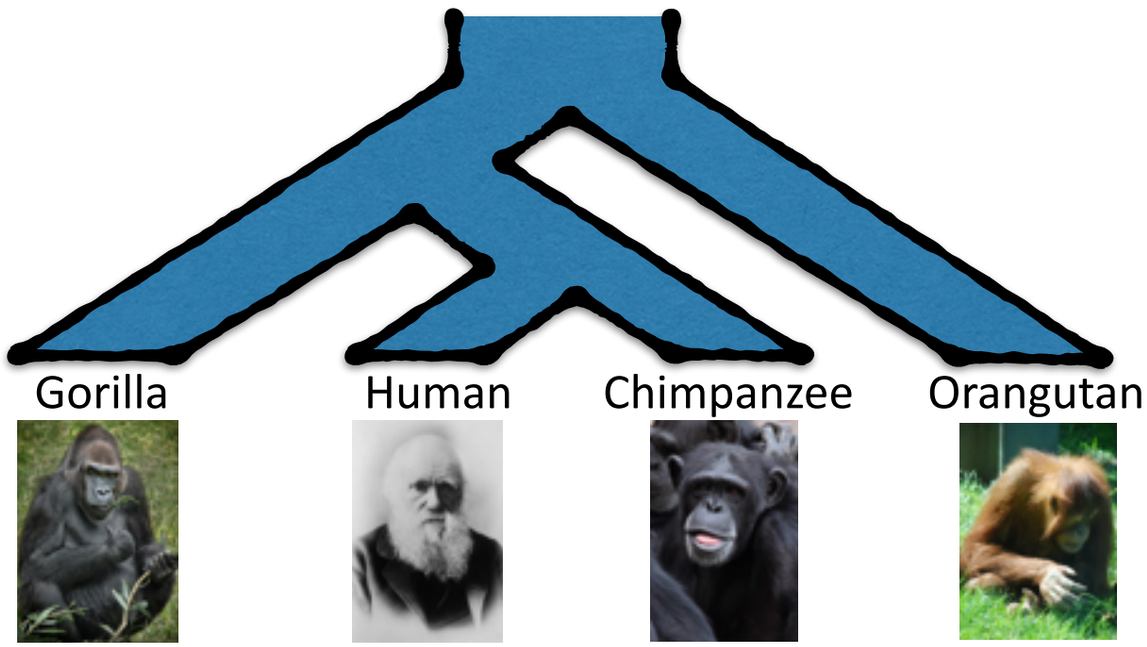
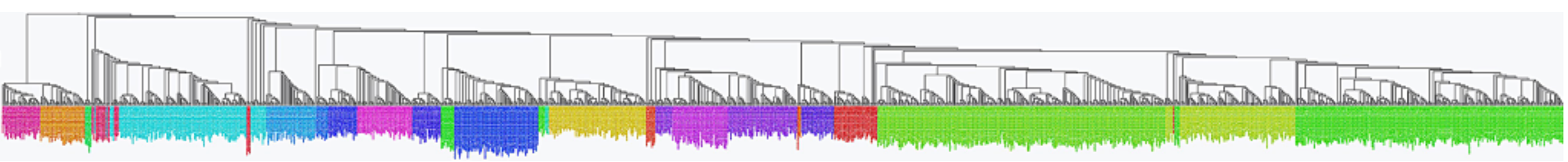
ASTRAL: Fast coalescent-based computation of the species tree topology, branch lengths, and local branch support

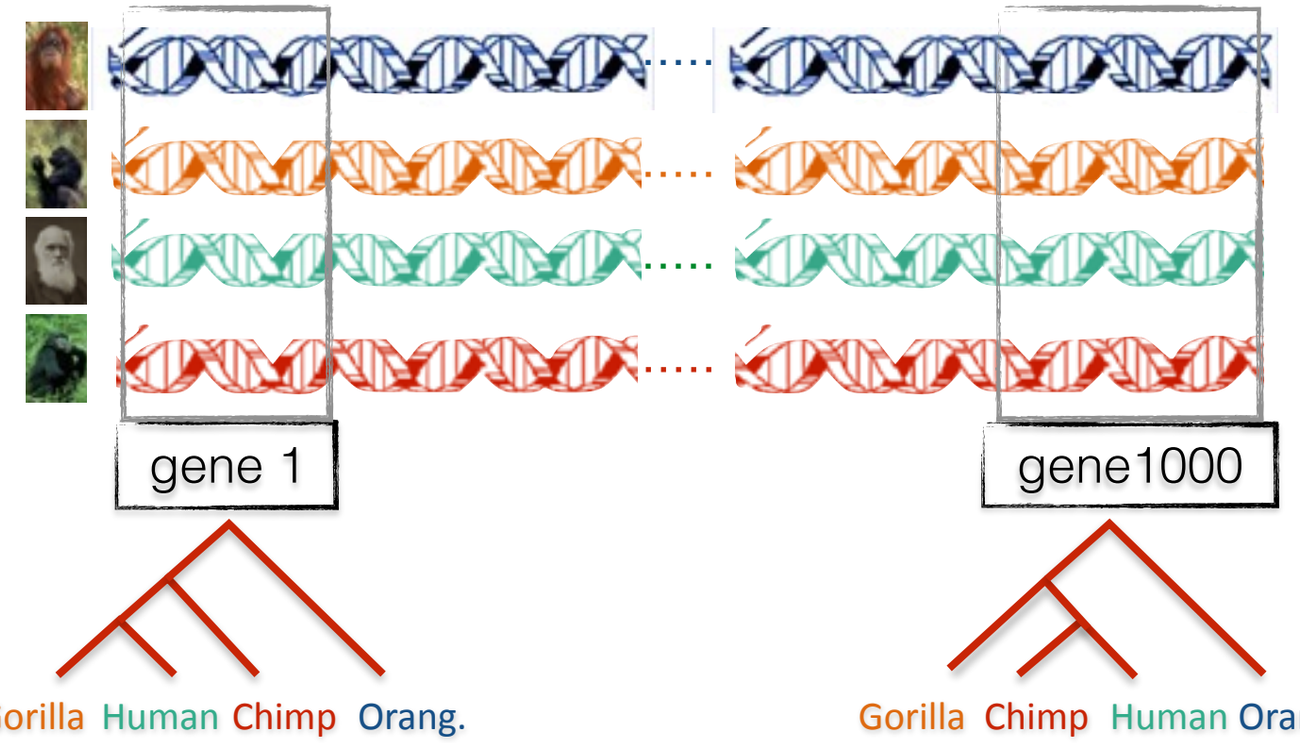
Siavash Mirarab
University of California, San Diego

Joint work with
Tandy Warnow
Erfan Sayyari

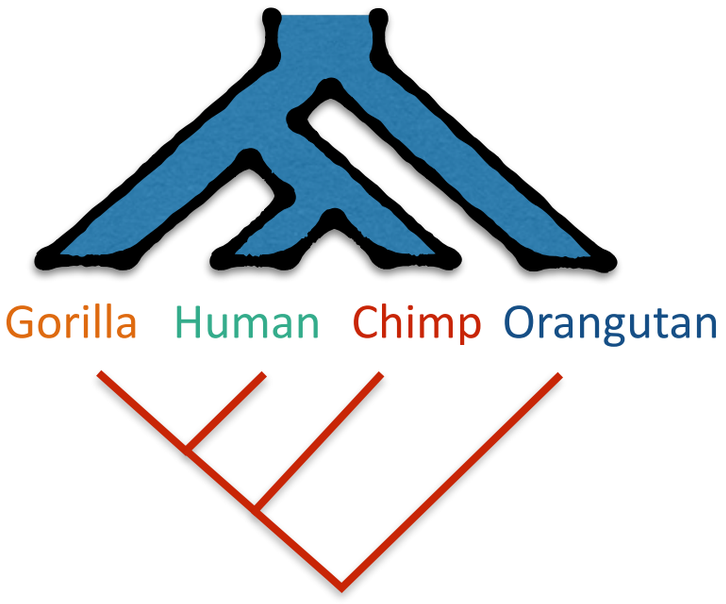




Gene tree discordance



The species tree



A gene tree

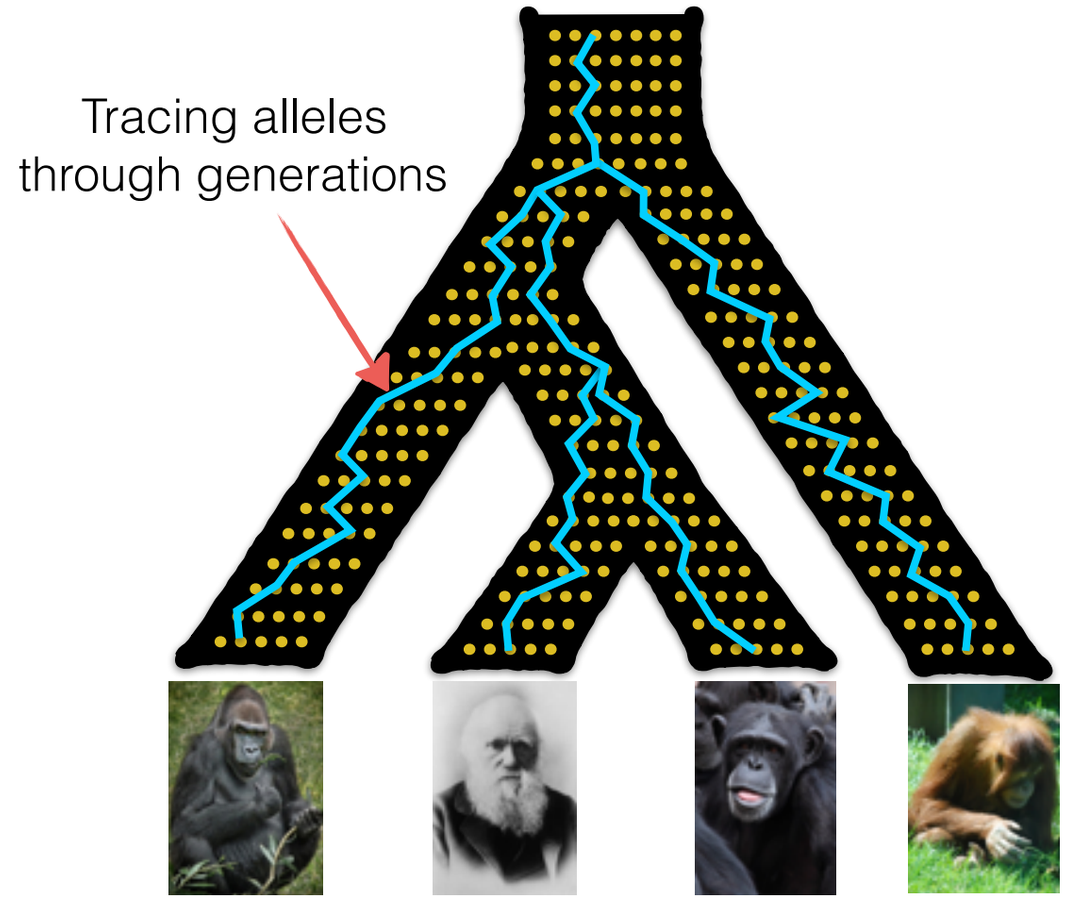
Causes of gene tree discordance include:

- Incomplete Lineage Sorting (ILS)
- Duplication and loss
- Horizontal Gene Transfer (HGT)

"gene": recombination-free orthologous stretches of the genome

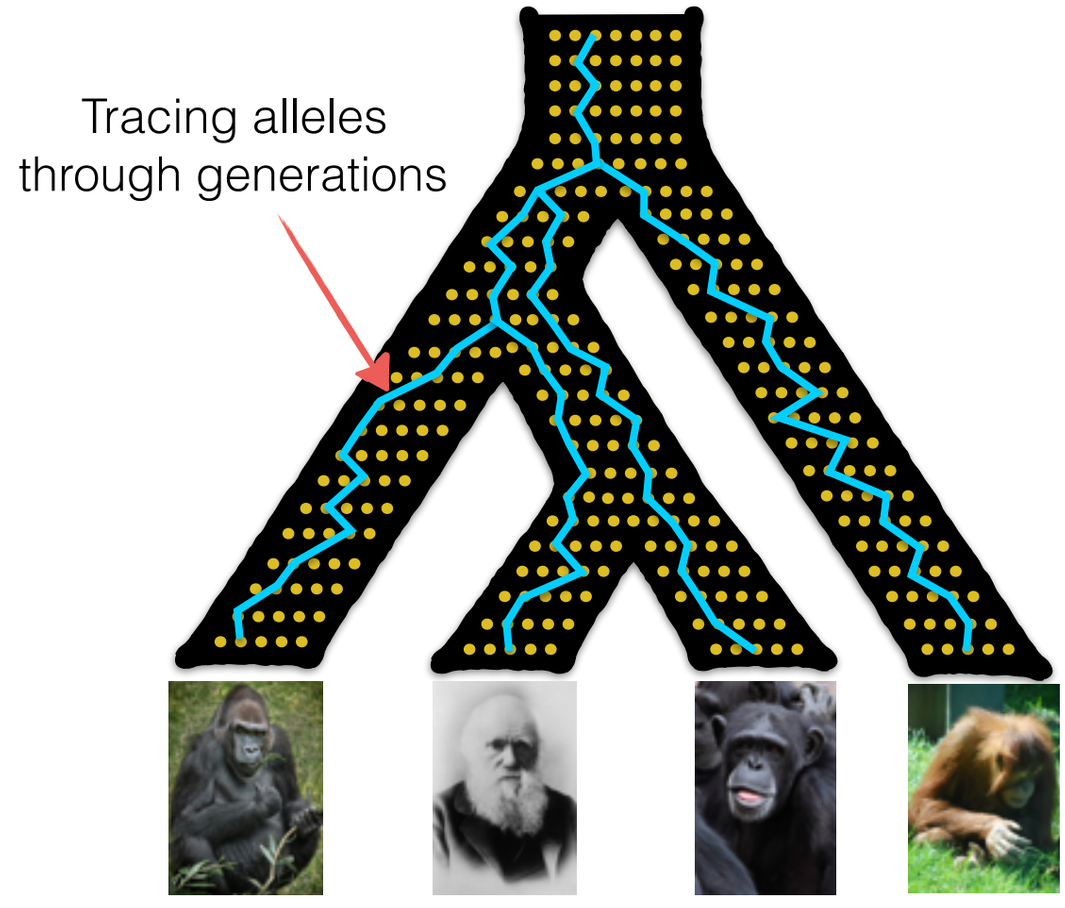
Incomplete Lineage Sorting (ILS)

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



Incomplete Lineage Sorting (ILS)

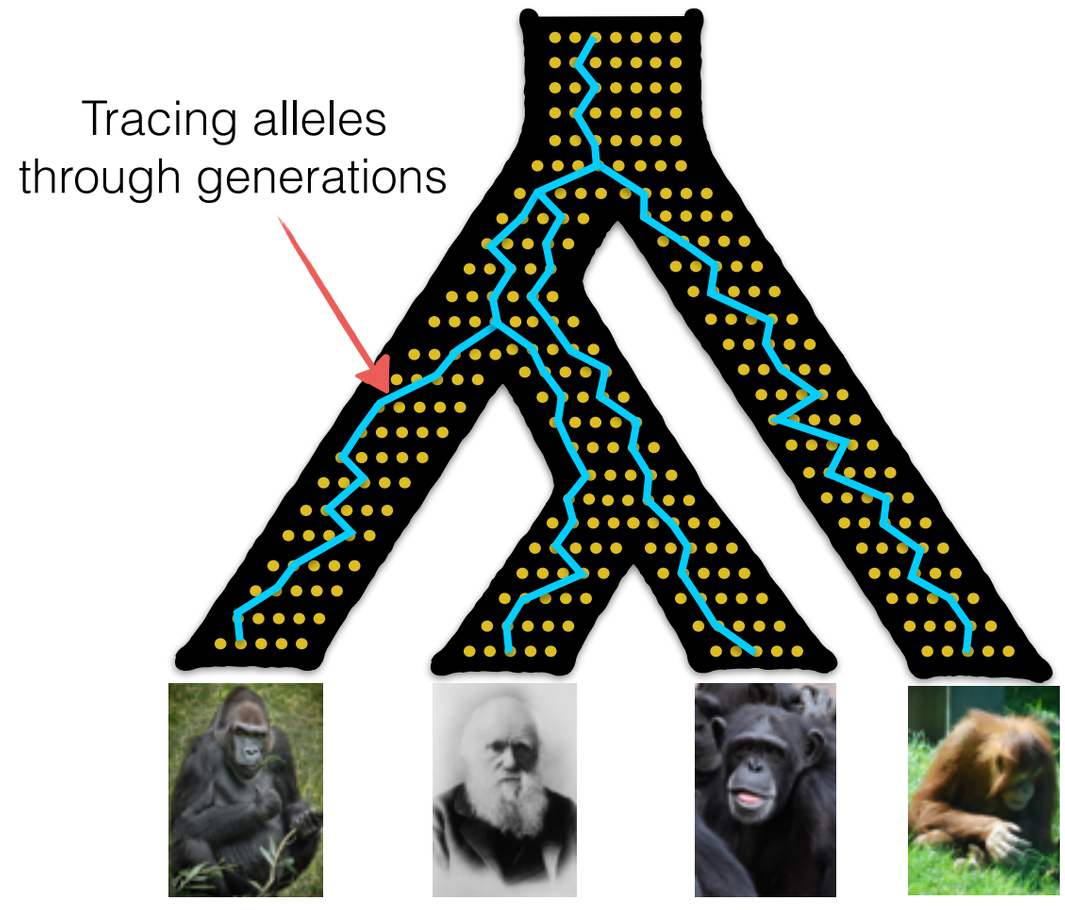
- A random process related to the coalescence of alleles across various populations



Incomplete Lineage Sorting (ILS)

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

- Omnipresent; most likely for short branches or large population sizes



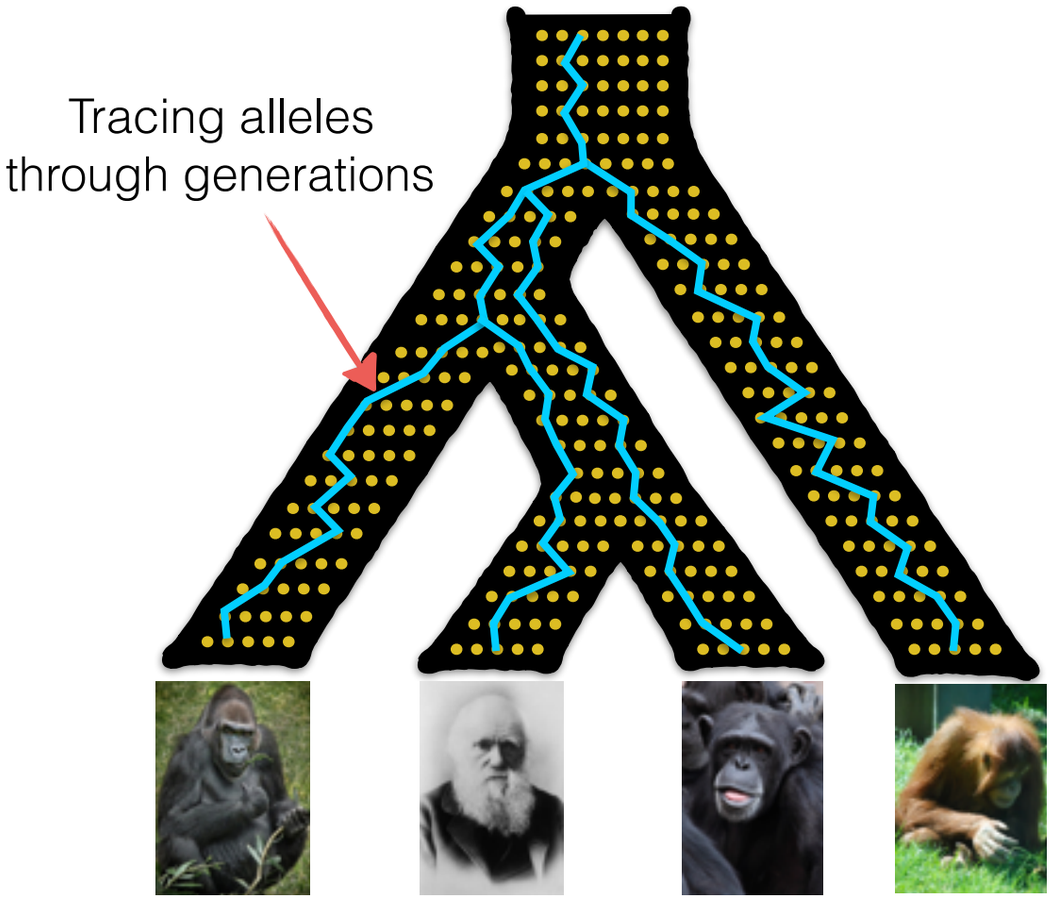
Incomplete Lineage Sorting (ILS)

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

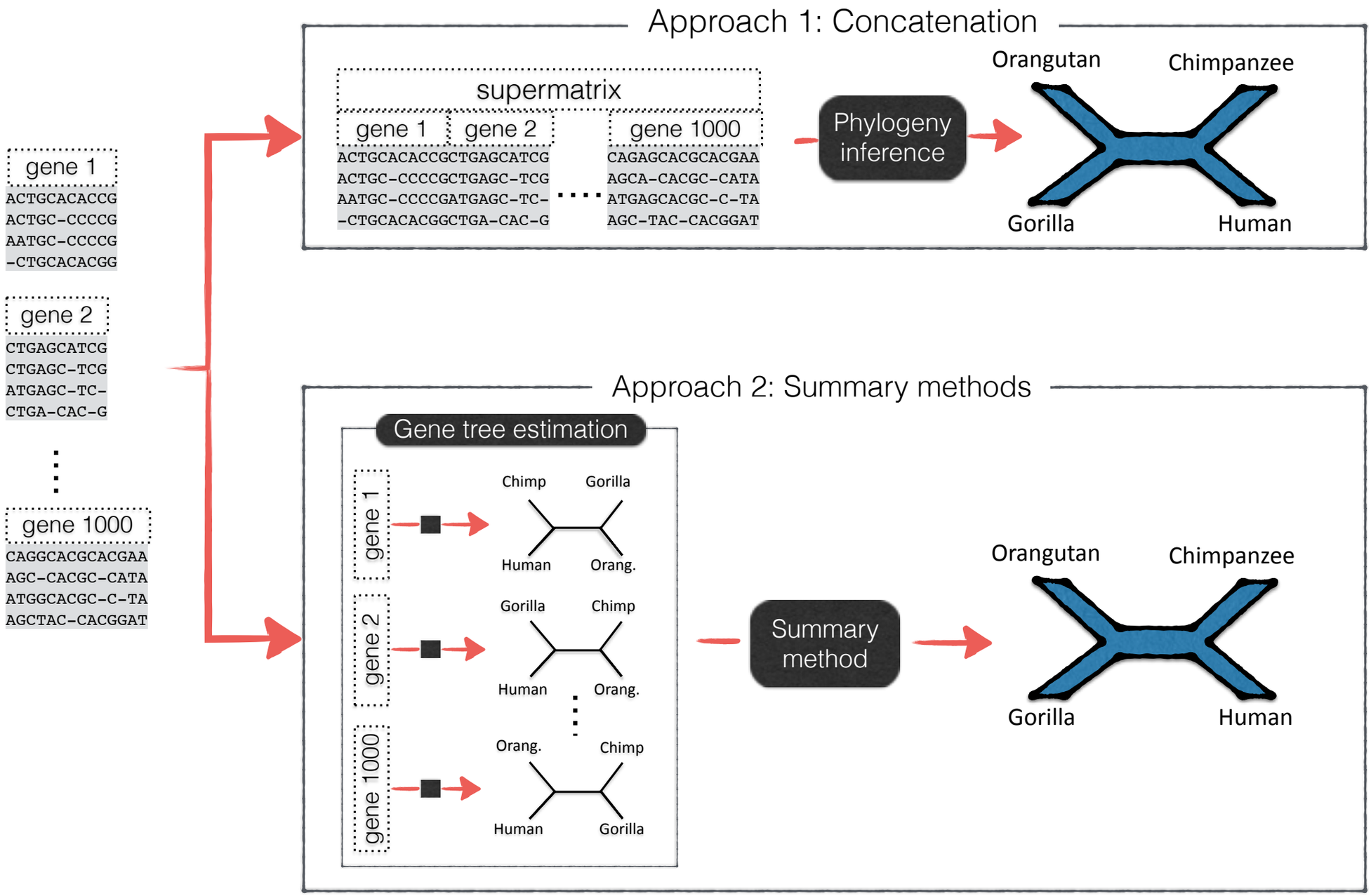
- Omnipresent; most likely for short branches or large population sizes

- Multi-species coalescent: models ILS

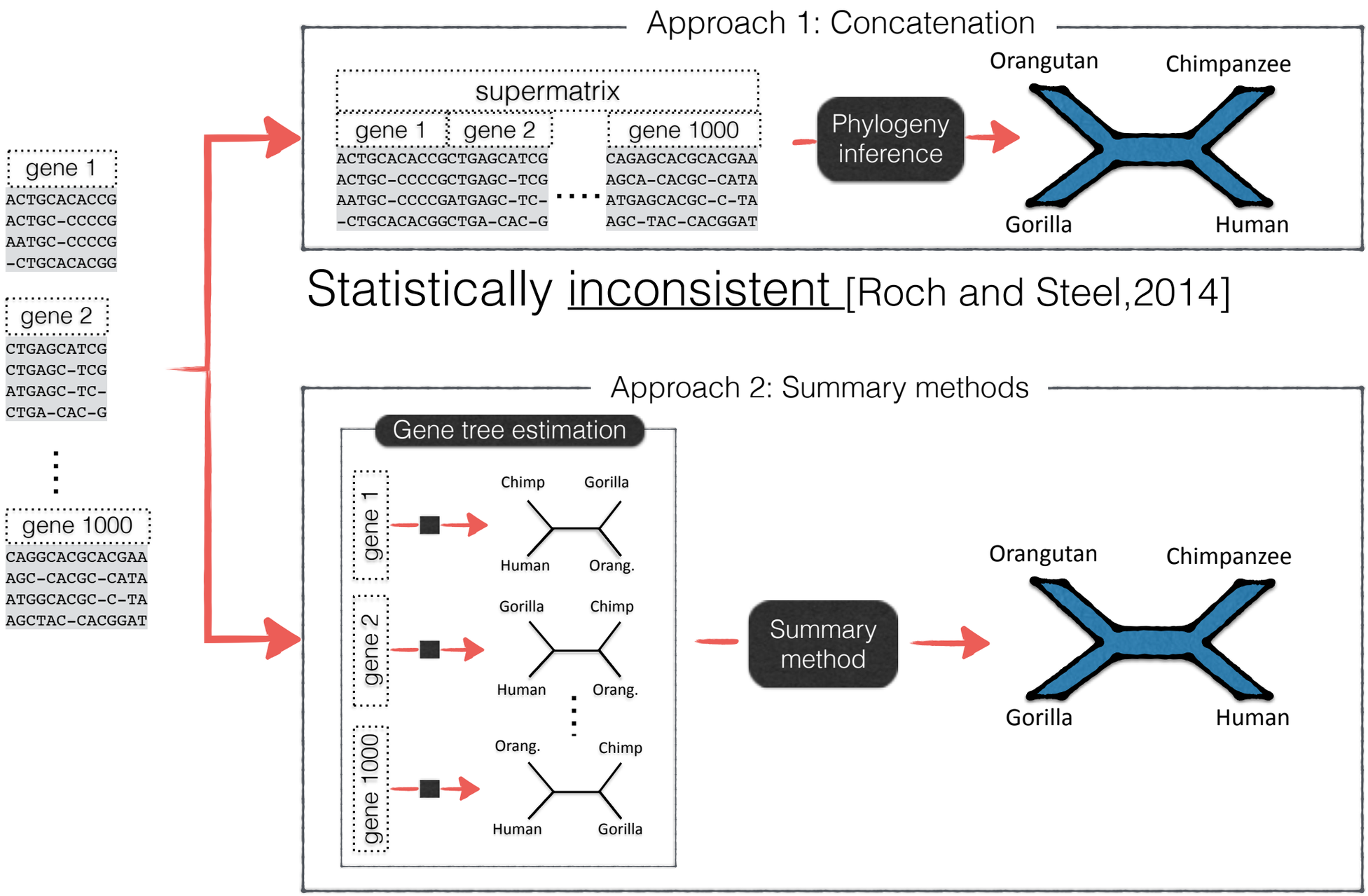
  - The species tree defines the probability distribution on gene trees, and is identifiable from the distribution on gene tree topologies
    [Degnan and Salter, Int. J. Org. Evolution, 2005]



Multi-gene species tree estimation



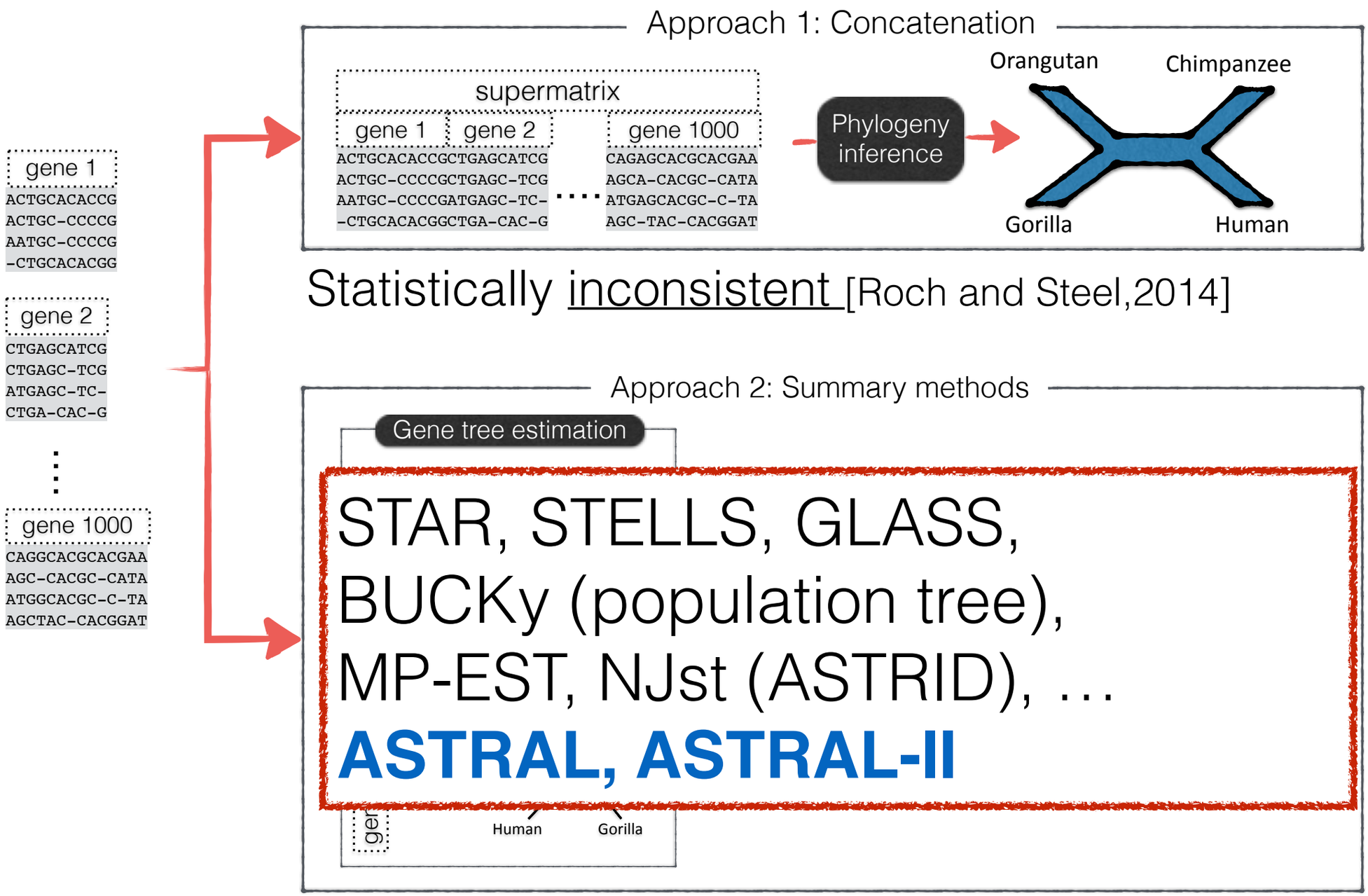
Multi-gene species tree estimation



Statistically <u>inconsistent</u> [Roch and Steel,2014]

Can be statistically consistent given <u>true gene trees</u>

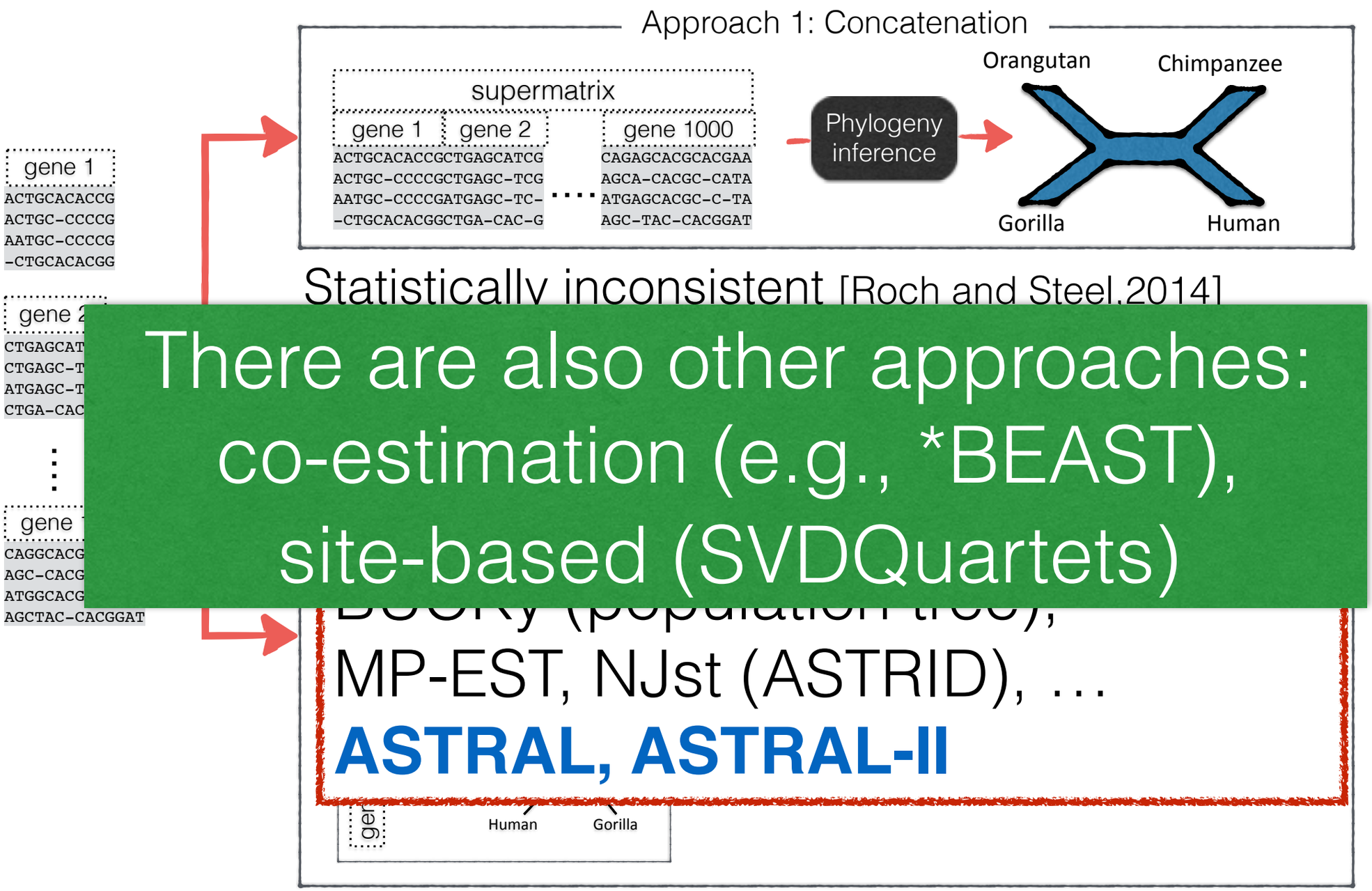
Multi-gene species tree estimation



Statistically <u>inconsistent</u> [Roch and Steel,2014]

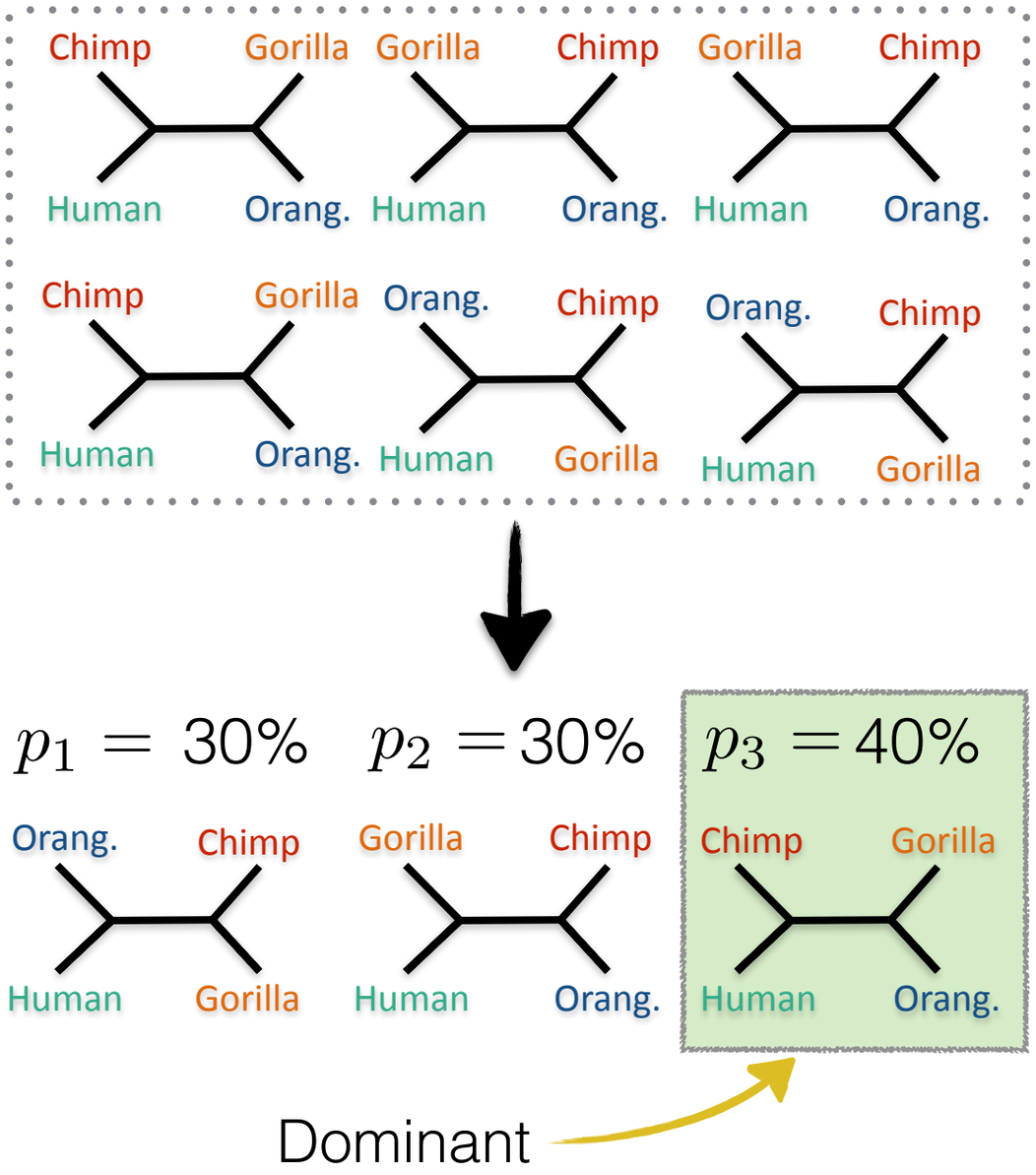
Can be statistically consistent given <u>true gene trees</u>

Multi-gene species tree estimation



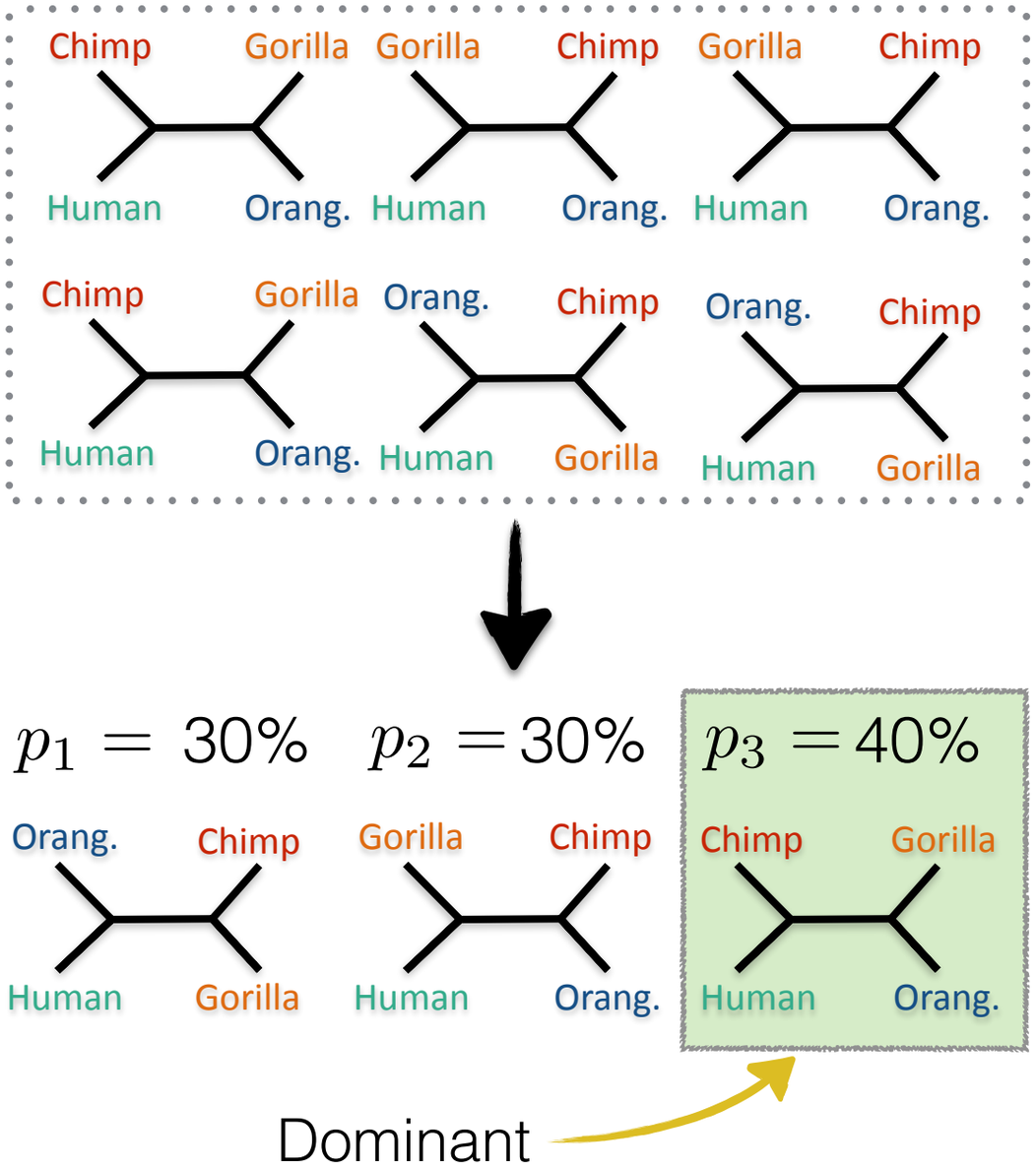
Can be statistically consistent given <u>true gene trees</u>

Properties of quartet trees in presence of ILS



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

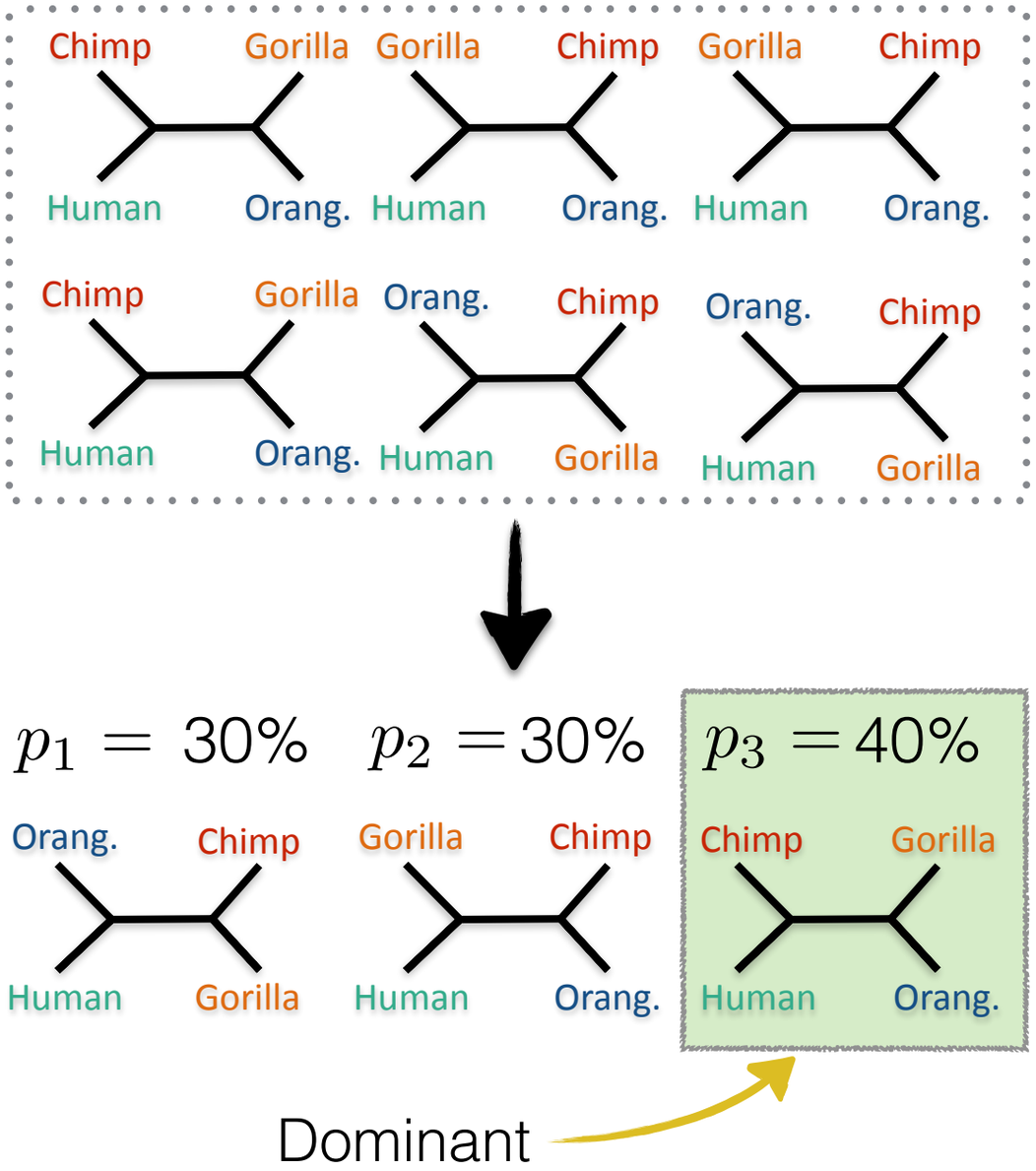
Properties of quartet trees in presence of ILS



- For 4 species, the dominant quartet topology is the species tree [Allman, et al. 2010]
- For >4 species, the dominant topology may be different from the species tree [Degnan and Rosenberg, 2006]

1. Breakup input each gene tree into $\binom{n}{4}$ trees on 4 taxa (quartet trees)
2. Find all $\binom{n}{4}$ dominant quartet topologies
3. Combine dominant quartet trees

Properties of quartet trees in presence of ILS



- For 4 species, the dominant quartet topology is the species tree [Allman, et al. 2010]
- For >4 species, the dominant topology may be different from the species tree [Degnan and Rosenberg, 2006]
  1. Breakup input each gene tree into $\binom{n}{4}$ trees on 4 taxa (quartet trees)
  2. Find all $\binom{n}{4}$ dominant quartet topologies
  3. Combine dominant quartet trees
- Alternative: weight 3 $\binom{n}{4}$ quartet topology by their frequency and find the optimal tree

Maximum Quartet Support Species Tree

[Mirarab, et al., ECCB, 2014]

- Optimization Problem (suspected NP-Hard):

Find the species tree with the maximum number of induced quartet trees shared with the collection of input gene trees

$$Score(T) = \sum_{t \in \mathcal{T}} |Q(T) \cap Q(t)|$$

Set of quartet trees induced by T

a gene tree

all input gene trees

- **Theorem**: Statistically consistent under the multi-species coalescent model when solved exactly

ASTRAL-I

[Mirarab, et al., Bioinformatics, 2014]

- ASTRAL solves the problem exactly using dynamic programming:
  - Exponential running time (feasible for <18 species)

ASTRAL-I

[Mirarab, et al., Bioinformatics, 2014]

- ASTRAL solves the problem exactly using dynamic programming:
  - Exponential running time (feasible for <18 species)

- Introduced a constrained version of the problem
  - Draws the set of branches in the species tree from a given set $\mathcal{X}$ = {all bipartitions in all gene trees}
  - Given many genes, each species tree branch likely appears in at least one of the gene trees
  - Theorem: the constrained version remains statistically consistent
  - Running time: $O(n^2 k|\mathcal{X}|^2)$ for *n* species and *k* species

ASTRAL-II

[Mirarab and Warnow, Bioinformatics, 2015]

1. Faster calculation of the score function inside DP

   - $O(nk|\mathcal{X}|^2)$ instead of $O(n^2k|\mathcal{X}|^2)$ for *n* species and *k* genes

ASTRAL-II

[Mirarab and Warnow, Bioinformatics, 2015]

1. Faster calculation of the score function inside DP

   - $O(nk|\mathcal{X}|^2)$ instead of $O(n^2k|\mathcal{X}|^2)$ for *n* species and *k* genes

2. Add extra bipartitions to the set $\mathcal{X}$ using heuristic approaches

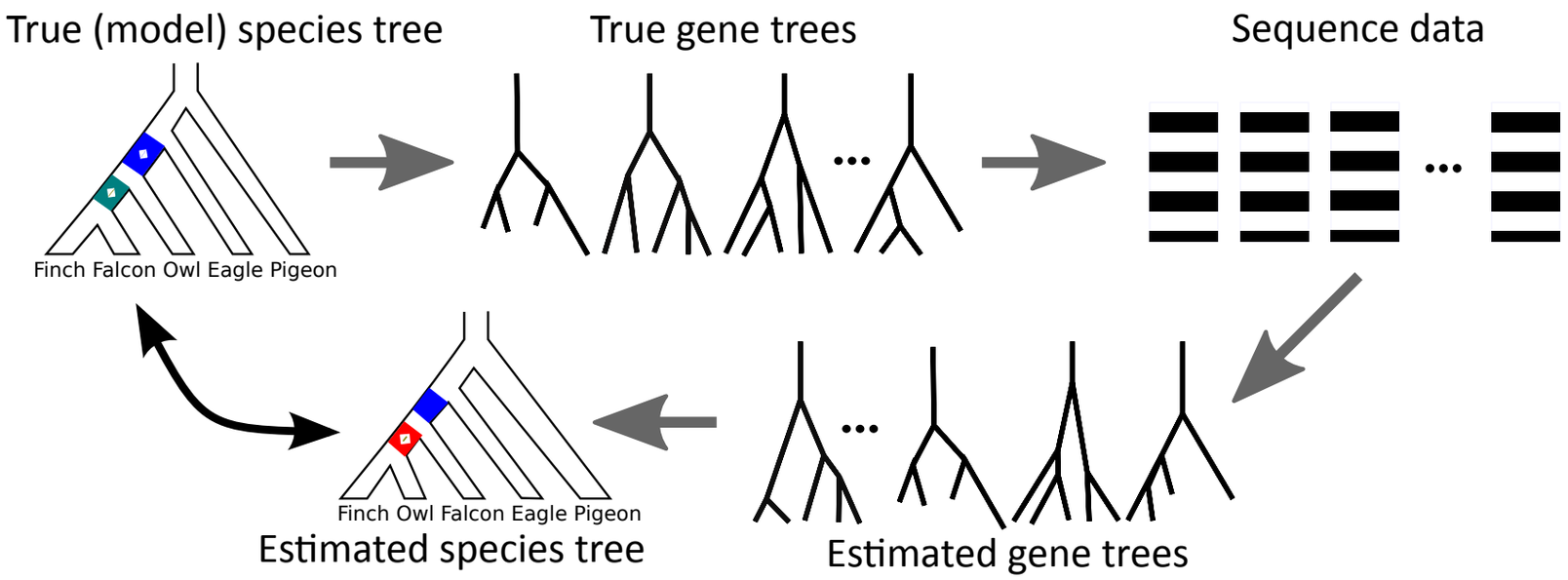
   - Resolving consensus trees by subsampling taxa
   - Using quartet-based distances to find likely branches
   - Complete incomplete gene trees

ASTRAL-II

[Mirarab and Warnow, Bioinformatics, 2015]

1. Faster calculation of the score function inside DP
   - $O(nk|\mathcal{X}|^2)$ instead of $O(n^2k|\mathcal{X}|^2)$ for *n* species and *k* genes
2. Add extra bipartitions to the set $\mathcal{X}$ using heuristic approaches
   - Resolving consensus trees by subsampling taxa
   - Using quartet-based distances to find likely branches
   - Complete incomplete gene trees
3. Ability to take as input gene trees with polytomies

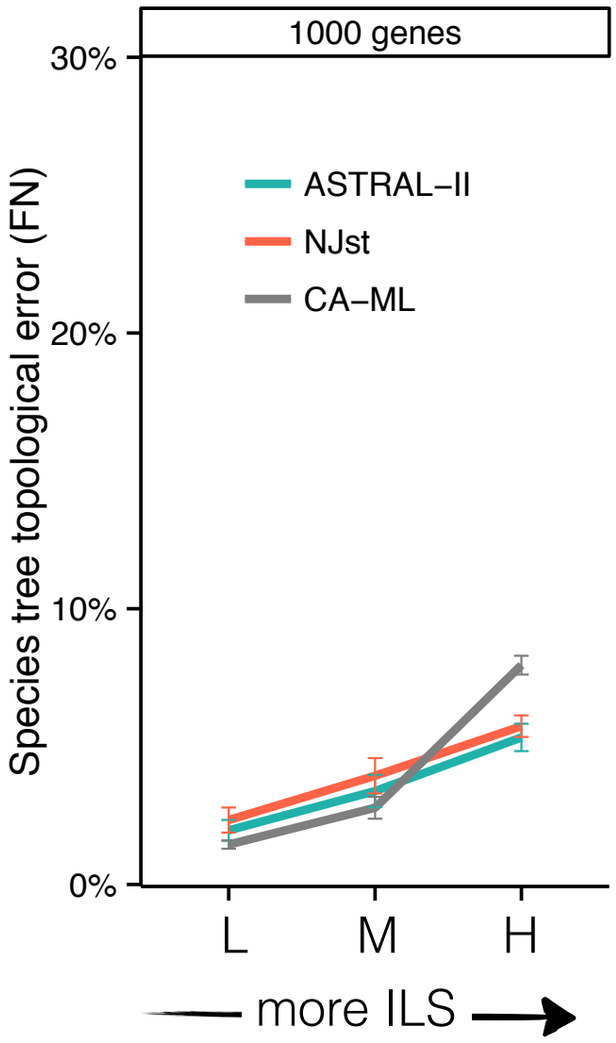
Simulation study



- Using SimPhy. Vary many parameters:
  - Number of species: 10 – 1000
  - Number of genes: 50 – 1000
  - Amount of ILS: low, medium, high
  - Deep versus recent speciation

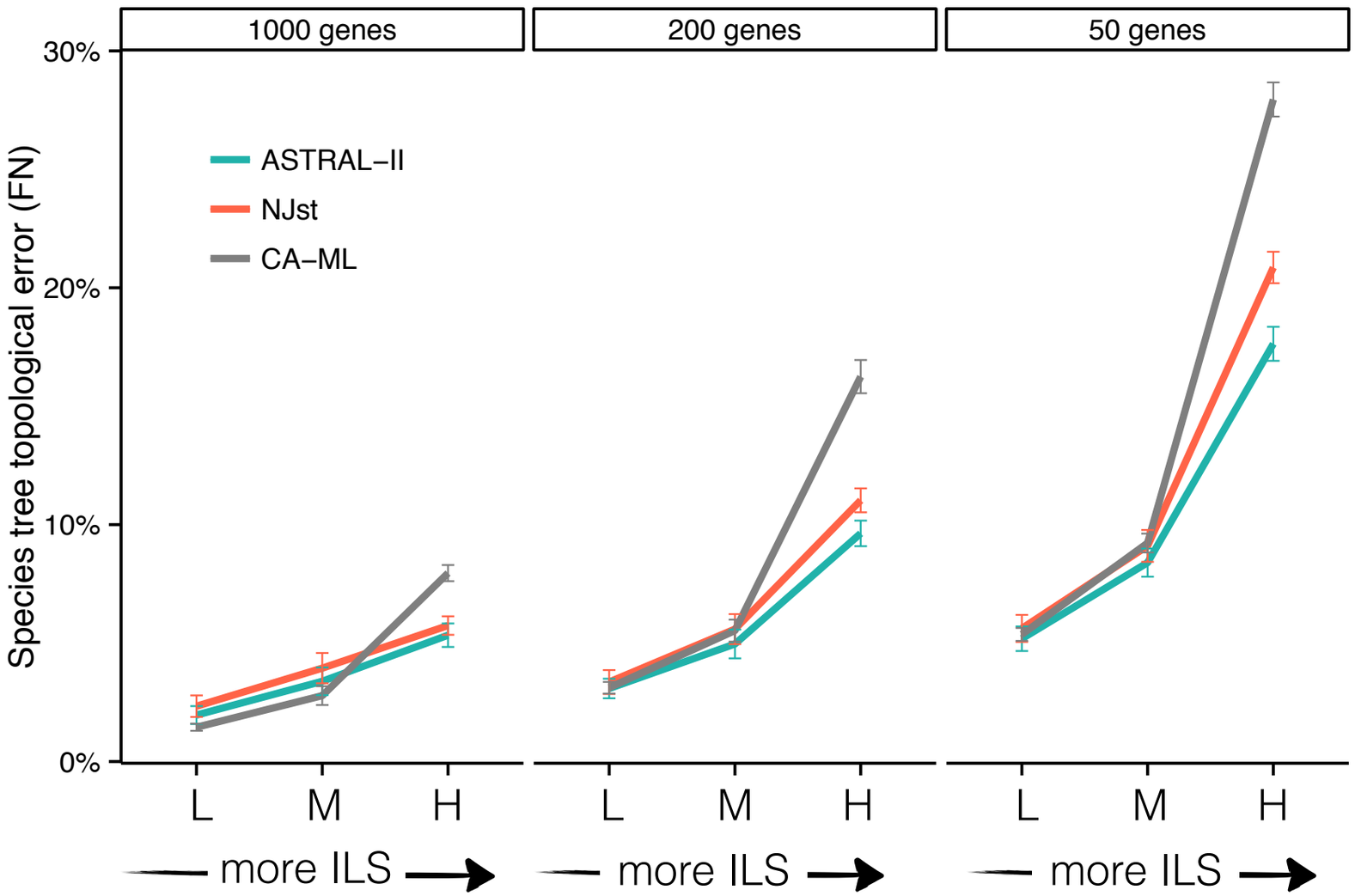
- 11 model conditions (50 replicas each) with heterogenous gene tree error
- Compare to NJst, MP-EST, concatenation (CA-ML) using FastTree-II
- Evaluate accuracy using FN rate: the percentage of branches in the true tree that are missing from the estimated tree

Tree error, varying level of ILS



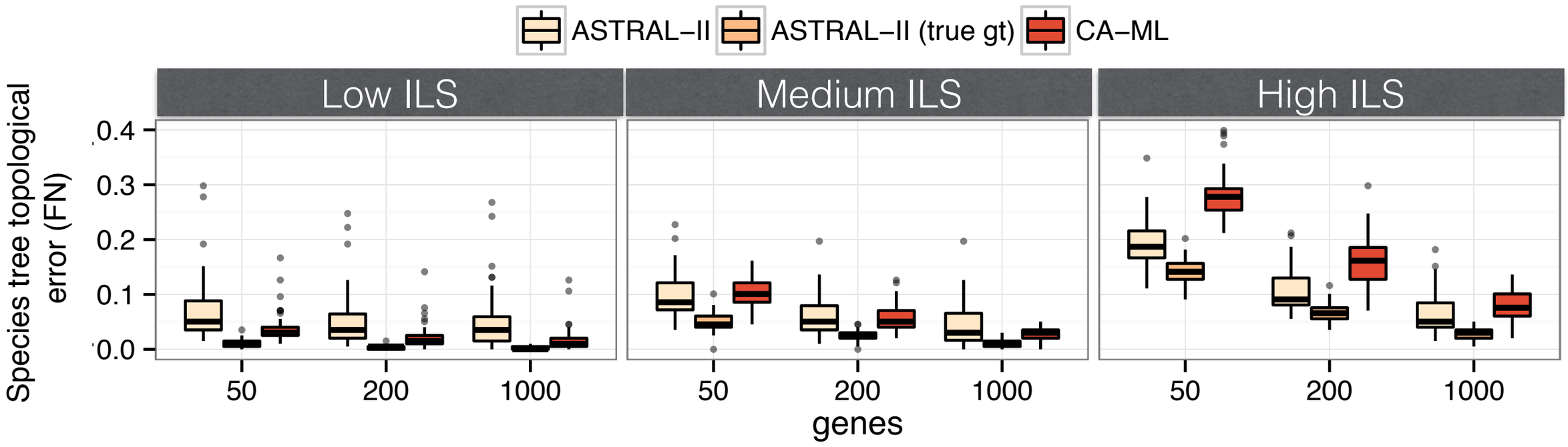
200 species, recent ILS

Tree error, varying level of ILS

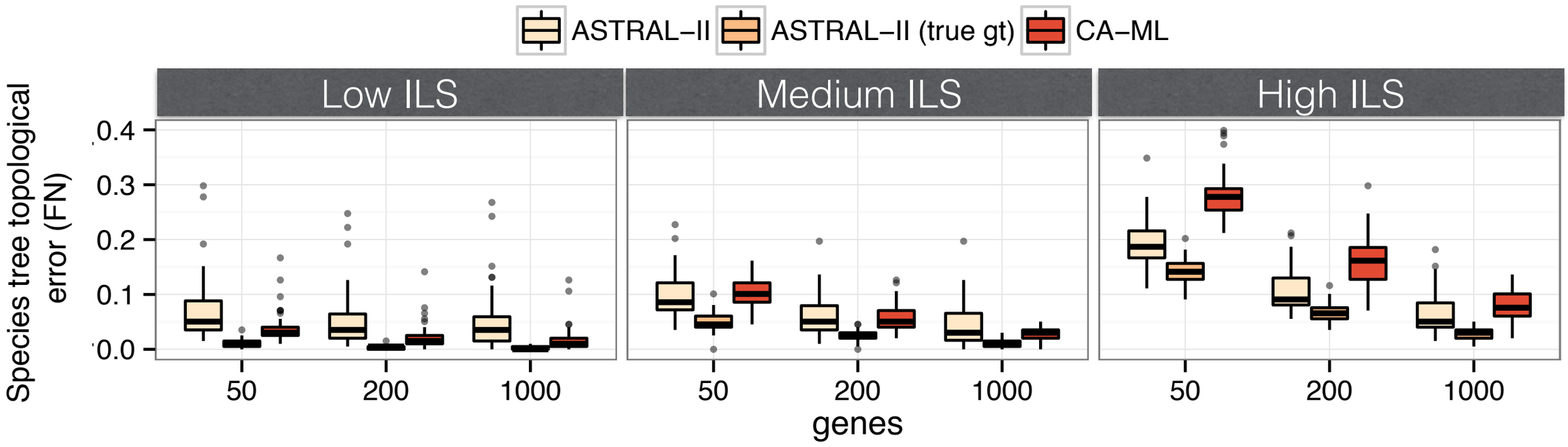


200 species, recent ILS

Impact of gene tree error (using true gene trees)



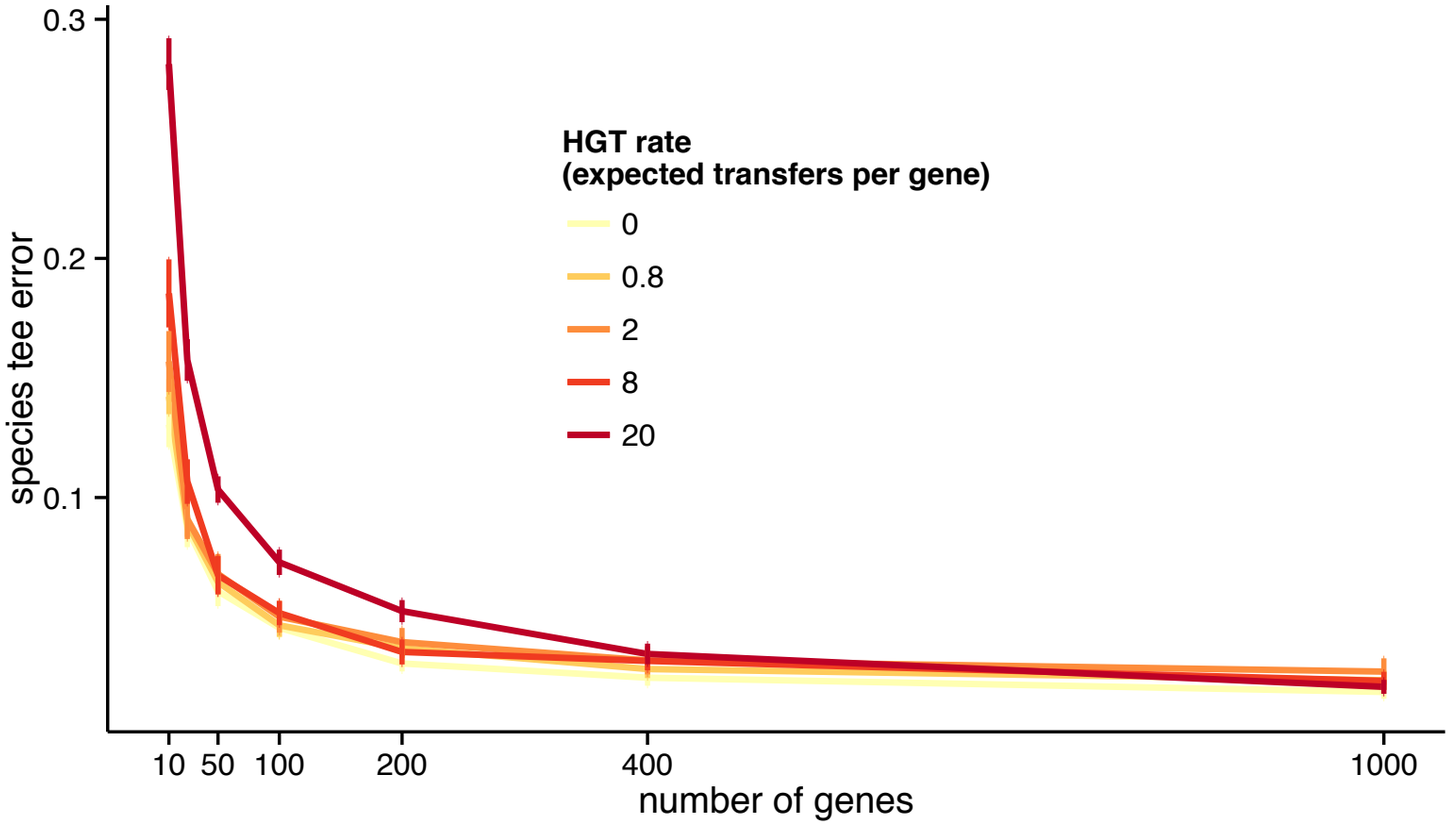
Impact of gene tree error (using true gene trees)



- When we divide our 50 replicates into low, medium, or high gene tree estimation error, ASTRAL tends to be better with low error

Impact of HGT

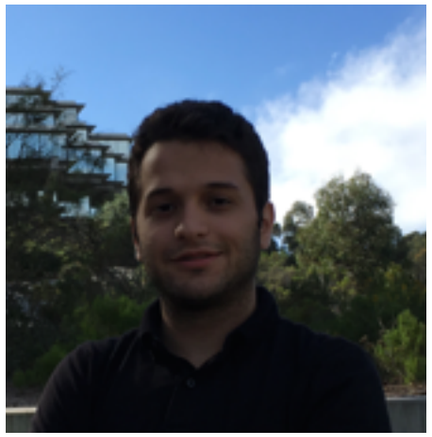
[Davidson, et al., BMC-genomics, 2015]



- Random HGT, with 50 species, moderate levels of ILS
- ASTRAL does well on ILS+HGT

Going beyond the topology

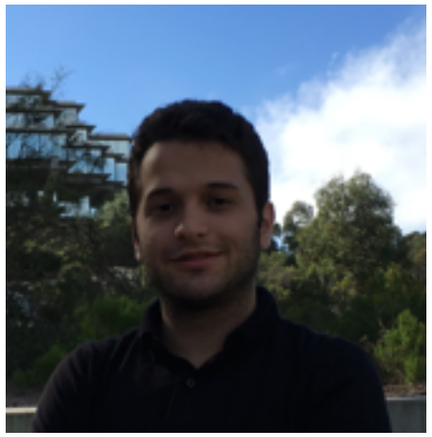
[Sayyari and Mirarab, MBE, 2016]



- Branch length (BL):
  - ASTRAL can compute BL in coalescent units for internal branches
  - Simply a function of the amount of discordance

Going beyond the topology

[Sayyari and Mirarab, MBE, 2016]

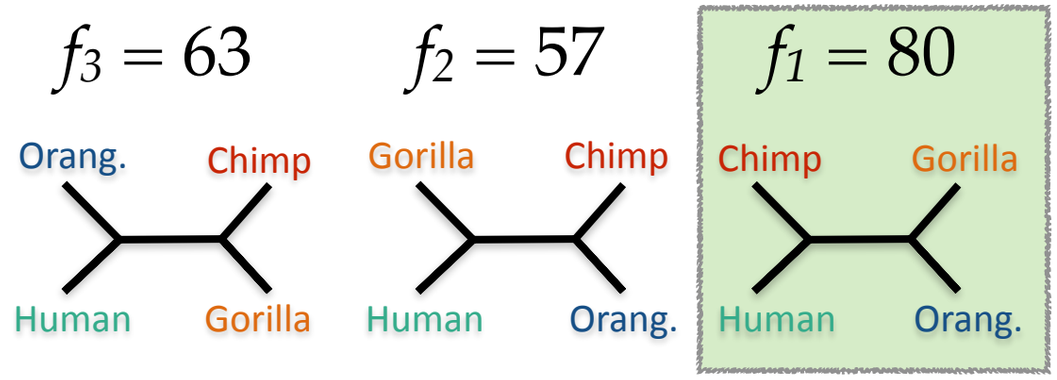


- Branch length (BL):
  - ASTRAL can compute BL in coalescent units for internal branches
  - Simply a function of the amount of discordance

- Branch support:
  - Multi-locus bootstrapping:
    — Slow (requires bootstrapping each gene)
    — Inaccurate [Mirarab et al., Sys bio, 2014; Bayzid et al., PLoS One, 2015]
  - Local posterior probability:
    ASTRAL's own support values that don't require bootstrapping

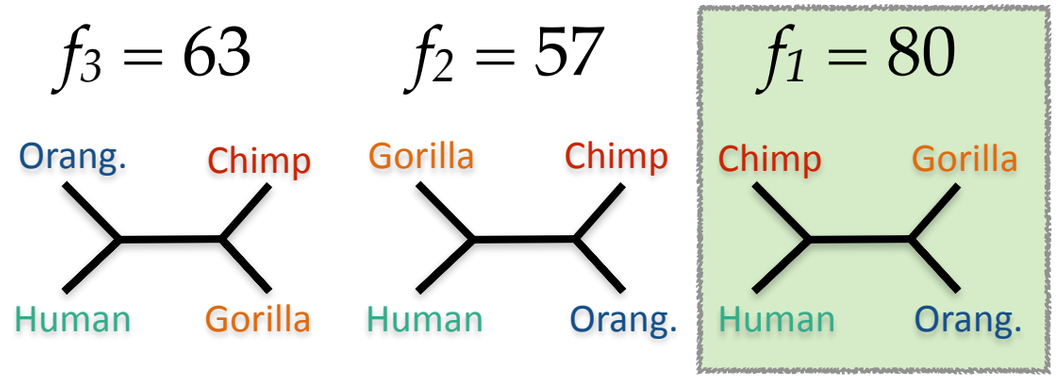
Branch support: idea

- Observing quartet topologies in gene trees is just like tossing a three-sided unbalanced die



Branch support: idea

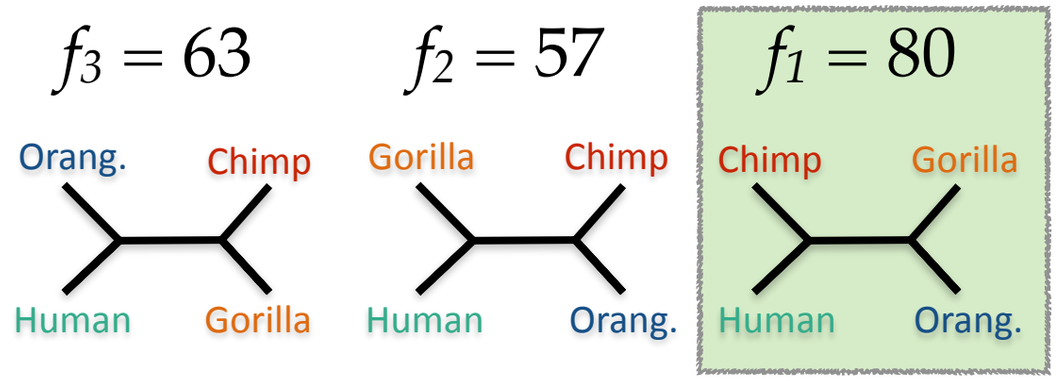
- Observing quartet topologies in gene trees is just like tossing a three-sided unbalanced die



- P (a quartet tree seen f_1 times in k gene trees is in the species tree) =
  P (a three-sided die tossed k times is biased towards a side that shows up f_1 times)

Branch support: idea

- Observing quartet topologies in gene trees is just like tossing a three-sided unbalanced die

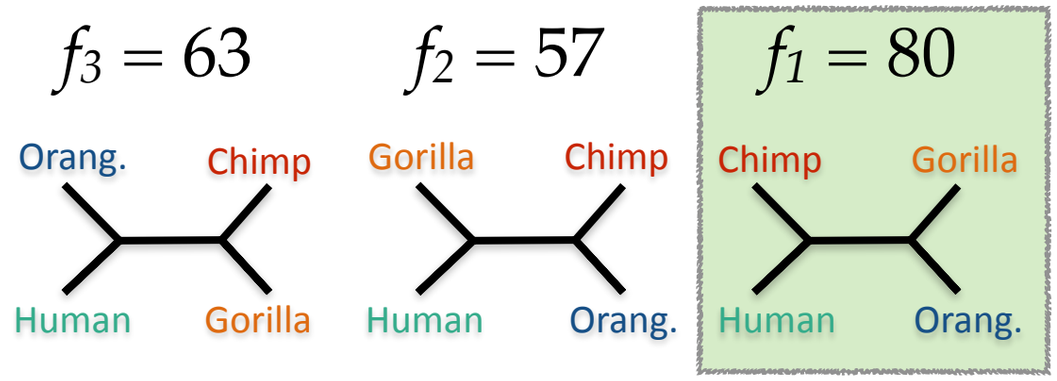


- P (a quartet tree seen f_1 times in k gene trees is in the species tree) =
  P (a three-sided die tossed k times is biased towards a side that shows up f_1 times)

- Can be analytically solved (Bayesian with a prior)

Branch support: idea

- Observing quartet topologies in gene trees is just like tossing a three-sided unbalanced die



- P (a quartet tree seen f_1 times in k gene trees is in the species tree) =
  P (a three-sided die tossed k times is biased towards a side that shows up f_1 times)

  - Can be analytically solved (Bayesian with a prior)

- ML or MAP BL can be computed as a function of f_1

Prior

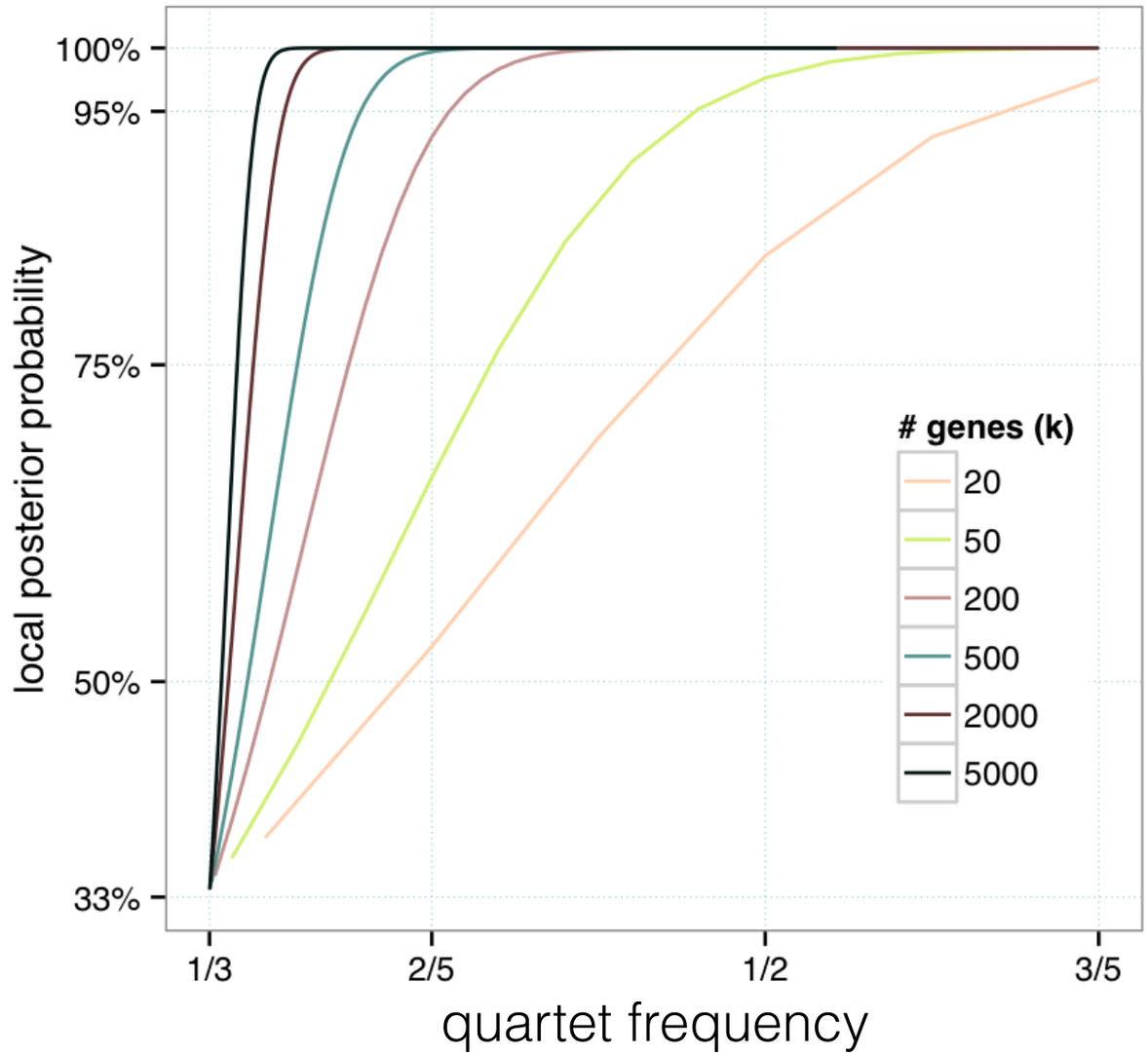
- We assume a-priori all three topologies are equally likely

$$Pr(\theta_1 > \frac{1}{3}) = Pr(\theta_2 > \frac{1}{3}) = Pr(\theta_3 > \frac{1}{3}) = \frac{1}{3}$$

- We assume branch lengths are generated through a Yule process with rate λ

  - Default: λ =0.5 —> branch lengths become uniformly distributed

Quartet support versus posterior probability



ASTRAL

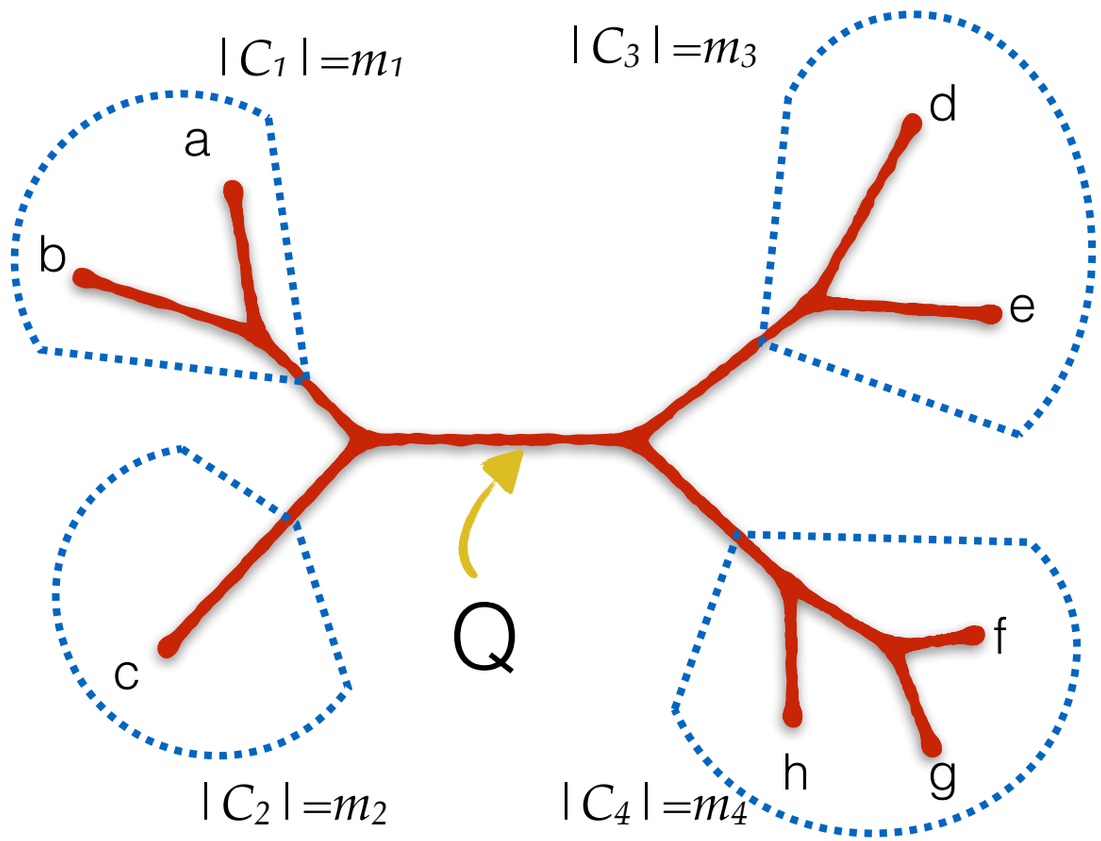
- **Input:** a set of *unrooted* gene trees

- **Output:** an *unrooted* species tree

- **Optimization score:** The number of quartet trees shared between the species tree and the input gene trees

Multiple quartets

- **Locality Assumption**: All four clusters around a branch Q are correct
    - Compute support for each branch independently from others



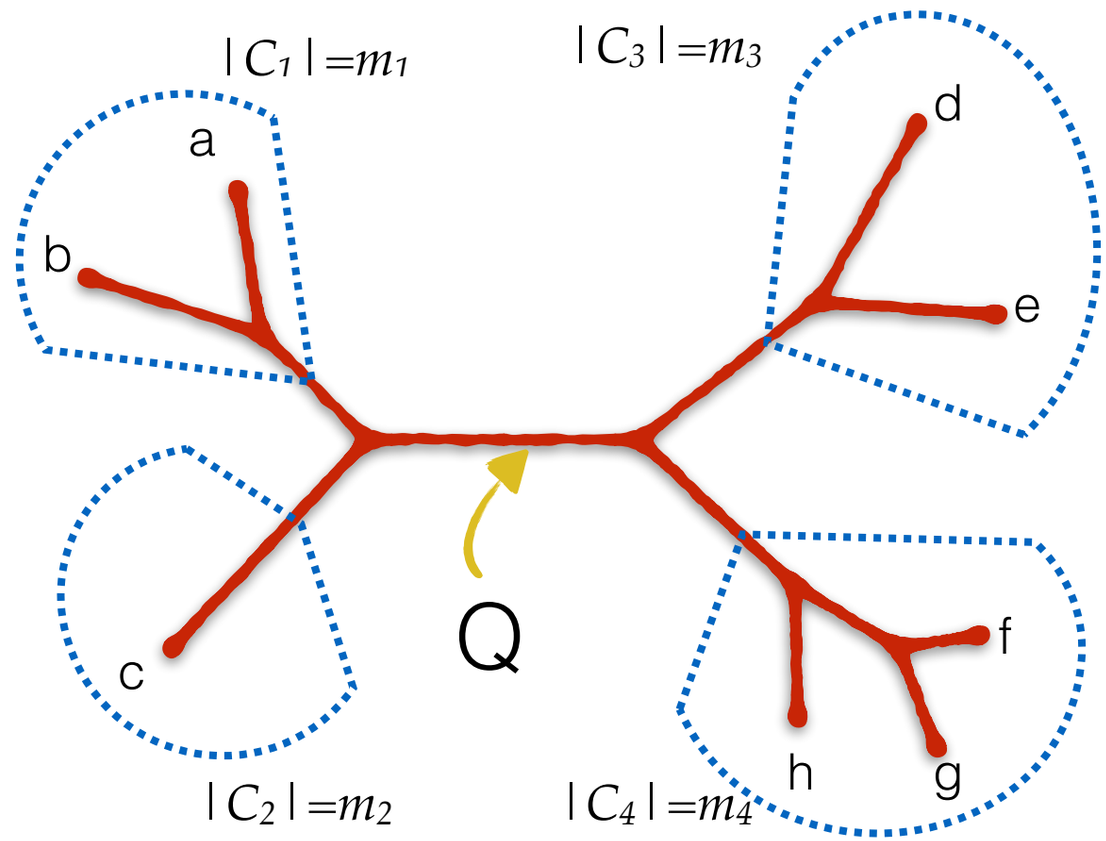
$$SQ = \{ac|dh, ac|df, bc|ef, \ldots\}$$

$$m = \#SQ = m_1 * m_2 * m_3 * m_4$$

$$F_i = (f_{i1}, f_{i2}, f_{i3}) \text{ for } 1 \leq i \leq m$$

Multiple quartets

- **Locality Assumption**: All four clusters around a branch Q are correct
    - Compute support for each branch independently from others
- Assuming independence between quartets would be too liberal (m x k tosses)



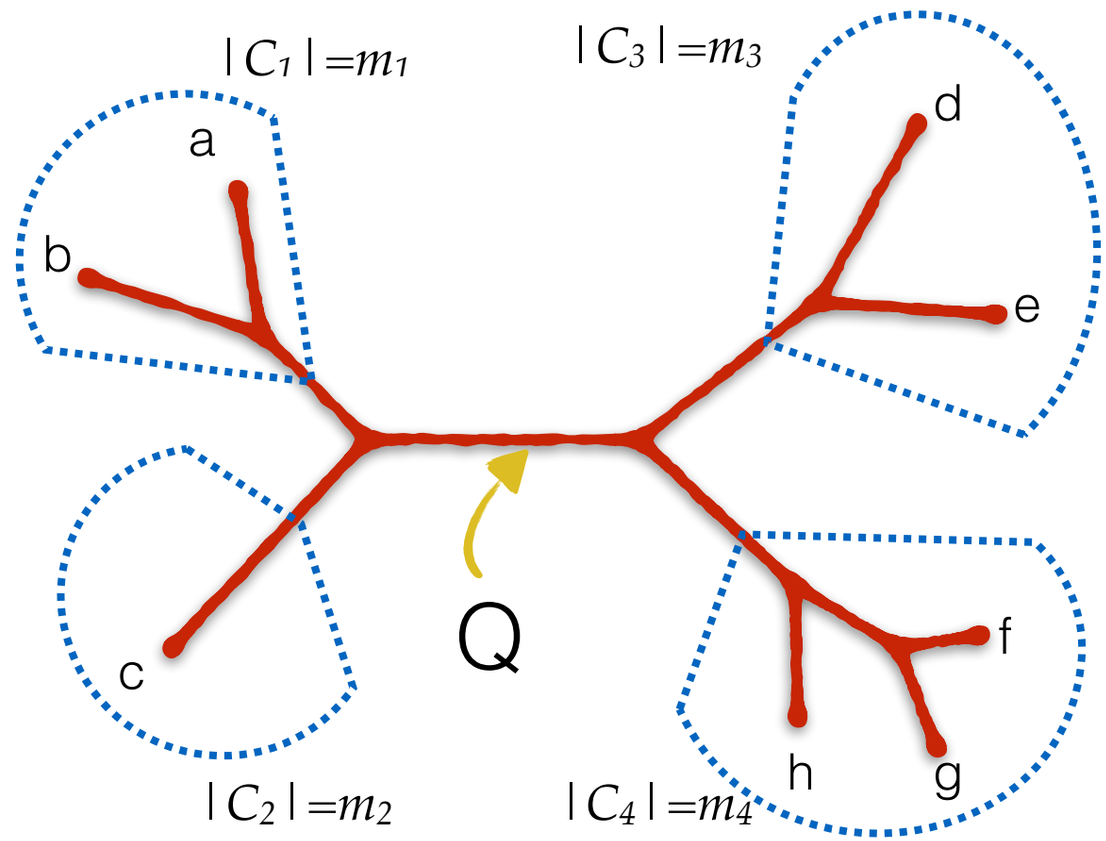
$$SQ = \{ac|dh, ac|df, bc|ef, \ldots\}$$

$$m = \#SQ = m_1 * m_2 * m_3 * m_4$$

$$F_i = (f_{i1}, f_{i2}, f_{i3}) \text{ for } 1 \leq i \leq m$$

Multiple quartets

- **Locality Assumption**: All four clusters around a branch Q are correct
  - Compute support for each branch independently from others
- Assuming independence between quartets would be too liberal (m x k tosses)
- Conservative approach: All quartet frequencies are noisy estimates of a single hidden true frequency
  - Simply average quartet frequencies
  - k tosses, each time reading the results m times with some noise



$$SQ = \{ac|dh, ac|df, bc|ef, \ldots\}$$

$$m = \#SQ = m_1 * m_2 * m_3 * m_4$$

$$F_i = (f_{i1}, f_{i2}, f_{i3}) \text{ for } 1 \leq i \leq m$$

Speed

- Can analytically compute local posterior probabilities and branch lengths (using results from Allman, et. al, 2012 for BL).

Speed

- Can analytically compute local posterior probabilities and branch lengths (using results from Allman, et. al, 2012 for BL).

- We don’t need to list all n choose 4 quartet frequencies.

  - We can use extensions of algorithmic tricks in ASTRAL to compute average quartet frequencies for each branch in $O(nk)$

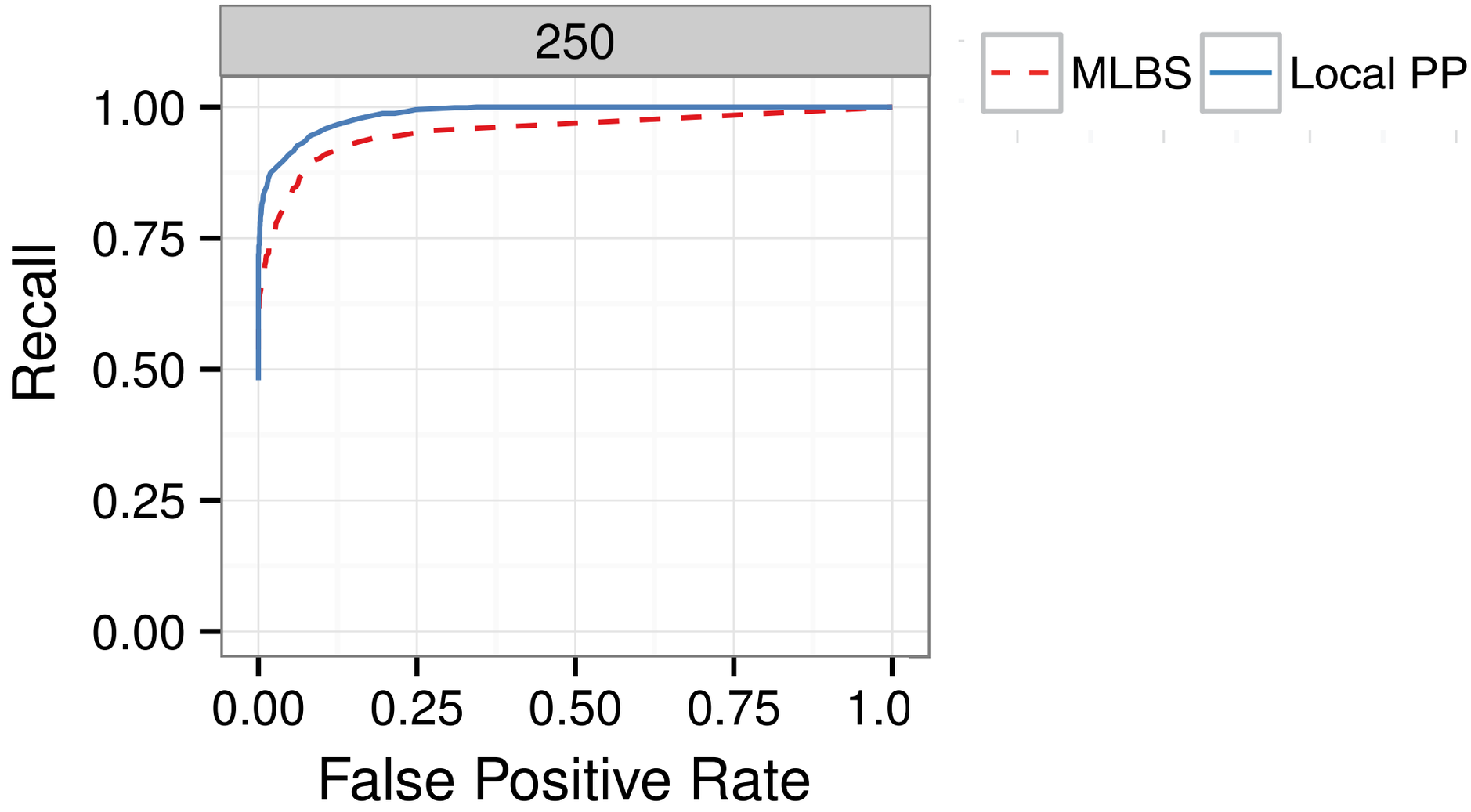
Simulation studies

- Violations of assumptions
  - Estimated gene trees instead of true gene trees
  - Estimated species trees, where the locality assumption is not always correct

- Datasets:
  — 201-taxon ASTRAL-II dataset
  — An avian simulated dataset

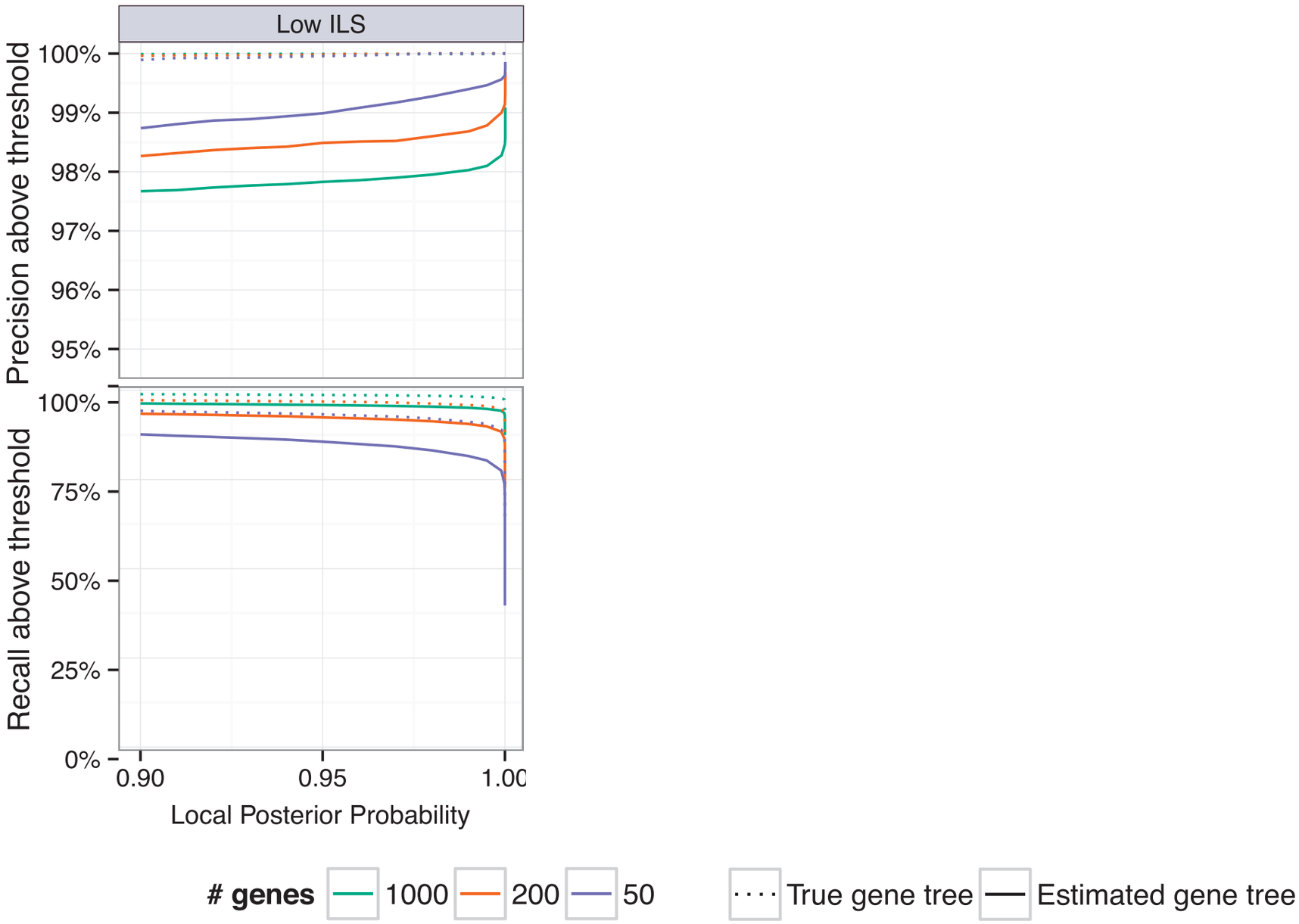
- Support accuracy:
  The number of false positive and false negatives above a certain threshold of support

Results (Avian, ROC)



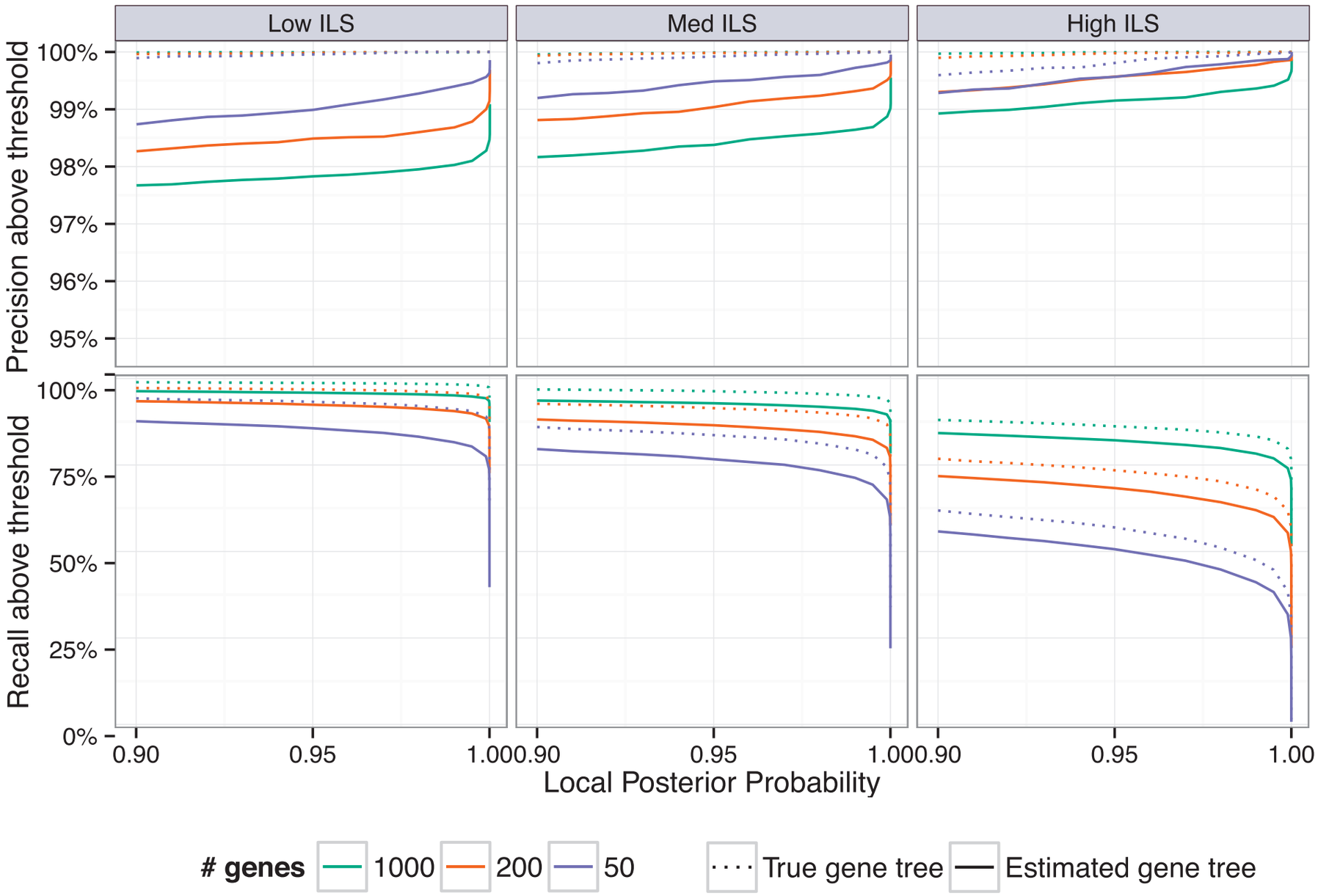
Avian simulated dataset (48 taxa, 1000 genes)

Precision and recall at high support



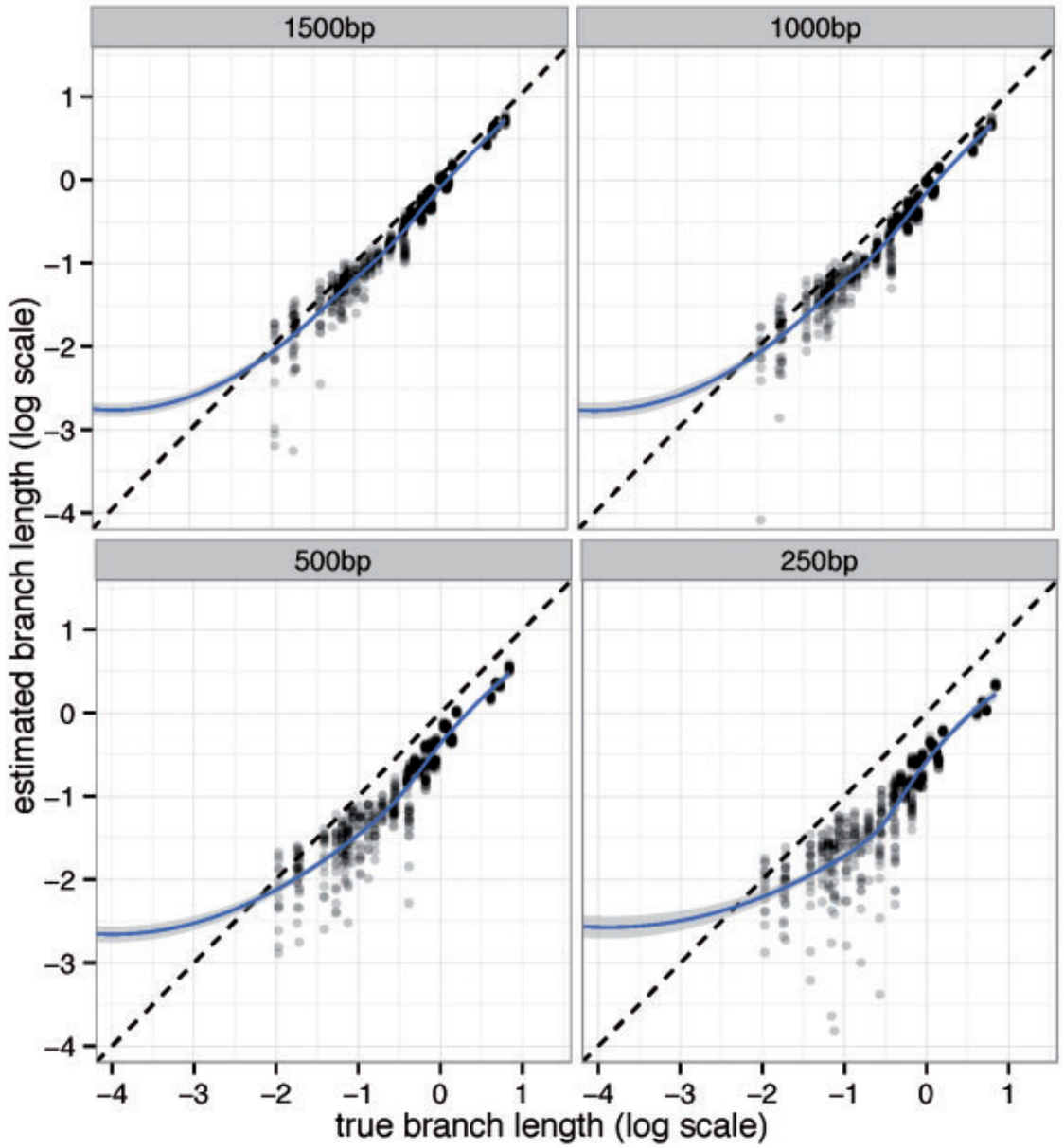
201-taxon datasets (simPhy)

Precision and recall at high support

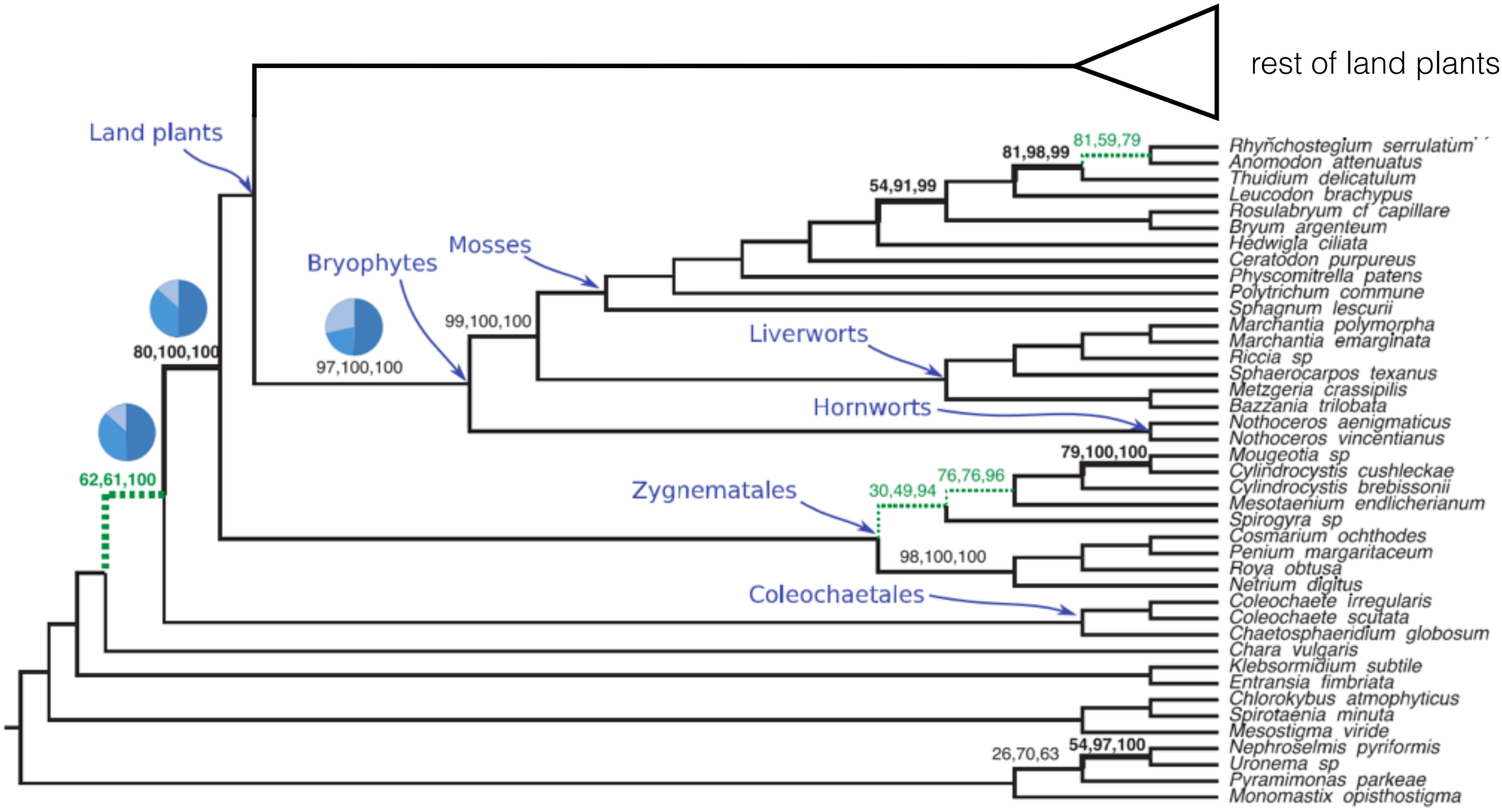


201-taxon datasets (simPhy)

Branch length accuracy



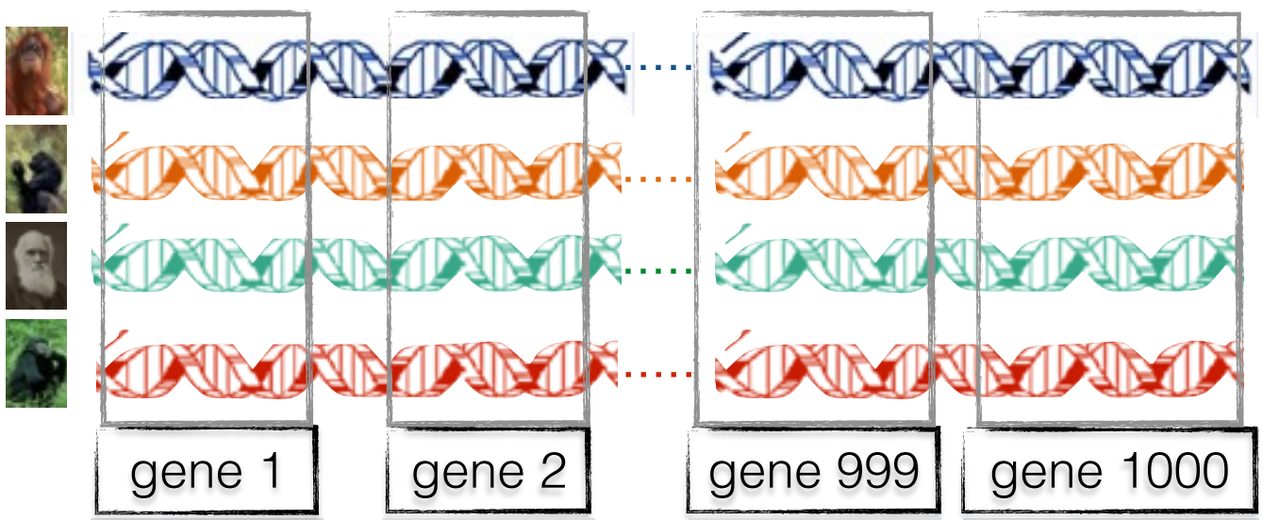
1KP dataset


rest of land plants
Land plants
Bryophytes
Mosses
Liverworts
Hornworts
Zygnematales
Coleochaetales
80,100,100
97,100,100
99,100,100
62,61,100
81,98,99
81,59,79
54,91,99
79,100,100
76,76,96
30,49,94
98,100,100
26,70,63
54,97,100
Rhynchostegium serrulatum
Anomodon attenuatus
Thuidium delicatulum
Leucodon brachypus
Rosulabryum cf capillare
Bryum argenteum
Hedwigia ciliata
Ceratodon purpureus
Physcomitrella patens
Polytrichum commune
Sphagnum lescurii
Marchantia polymorpha
Marchantia emarginata
Riccia sp
Sphaerocarpos texanus
Metzgeria crassipilis
Bazzania trilobata
Nothoceros aenigmaticus
Nothoceros vincentianus
Mougeotia sp
Cylindrocystis cushleckae
Cylindrocystis brebissonii
Mesotaenium endlicherianum
Spirogyra sp
Cosmarium ochthodes
Penium margaritaceum
Roya obtusa
Netrium digitus
Coleochaete irregularis
Coleochaete scutata
Chaetosphaeridium globosum
Chara vulgaris
Klebsormidium subtile
Entransia fimbriata
Chlorokybus atmophyticus
Spirotaenia minuta
Mesostigma viride
Nephroselmis pyriformis
Uronema sp
Pyramimonas parkeae
Monomastix opisthostigma

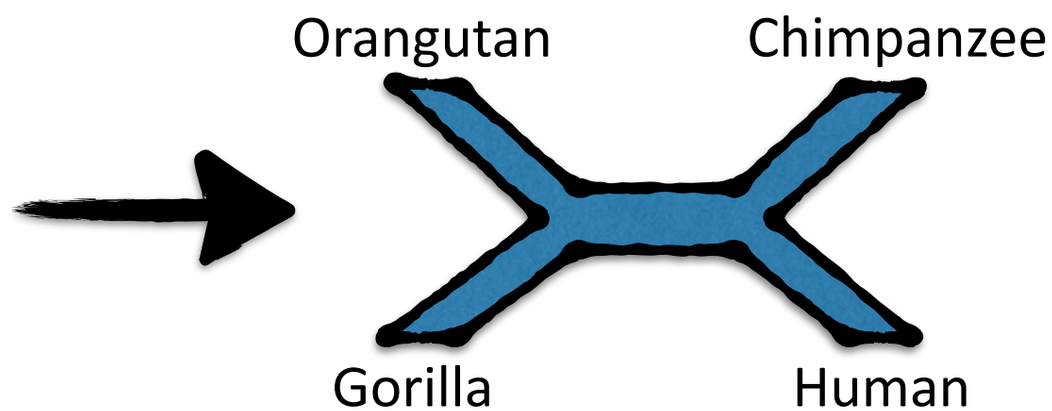
Summary

- ASTRAL-II can infer species tree from gene trees in datasets with 1000 genes and 1000 taxa in a day of single cpu running time.
- ASTRAL mostly dominates other summary methods. However, Concatenation is better when gene trees have high error.
- ASTRAL's way of computing support is more accurate than MLBS while being much faster

Phylogenomic species tree reconstruction




```
ACTGCACACCG   CTGAGCATCG     AGCAGCATCGTG  CAGGCACGCACGAA
ACTGC-CCCCG   CTGAGC-TCG ... AGCAGC-TCGTG  AGC-CACGC-CATA
AATGC-CCCCG   ATGAGC-TC-     AGCAGC-TC-TG  ATGGCACGC-C-TA
-CTGCACACGG   CTGA-CAC-G     C-TA-CACGGTG  AGCTAC-CACGGAT
```



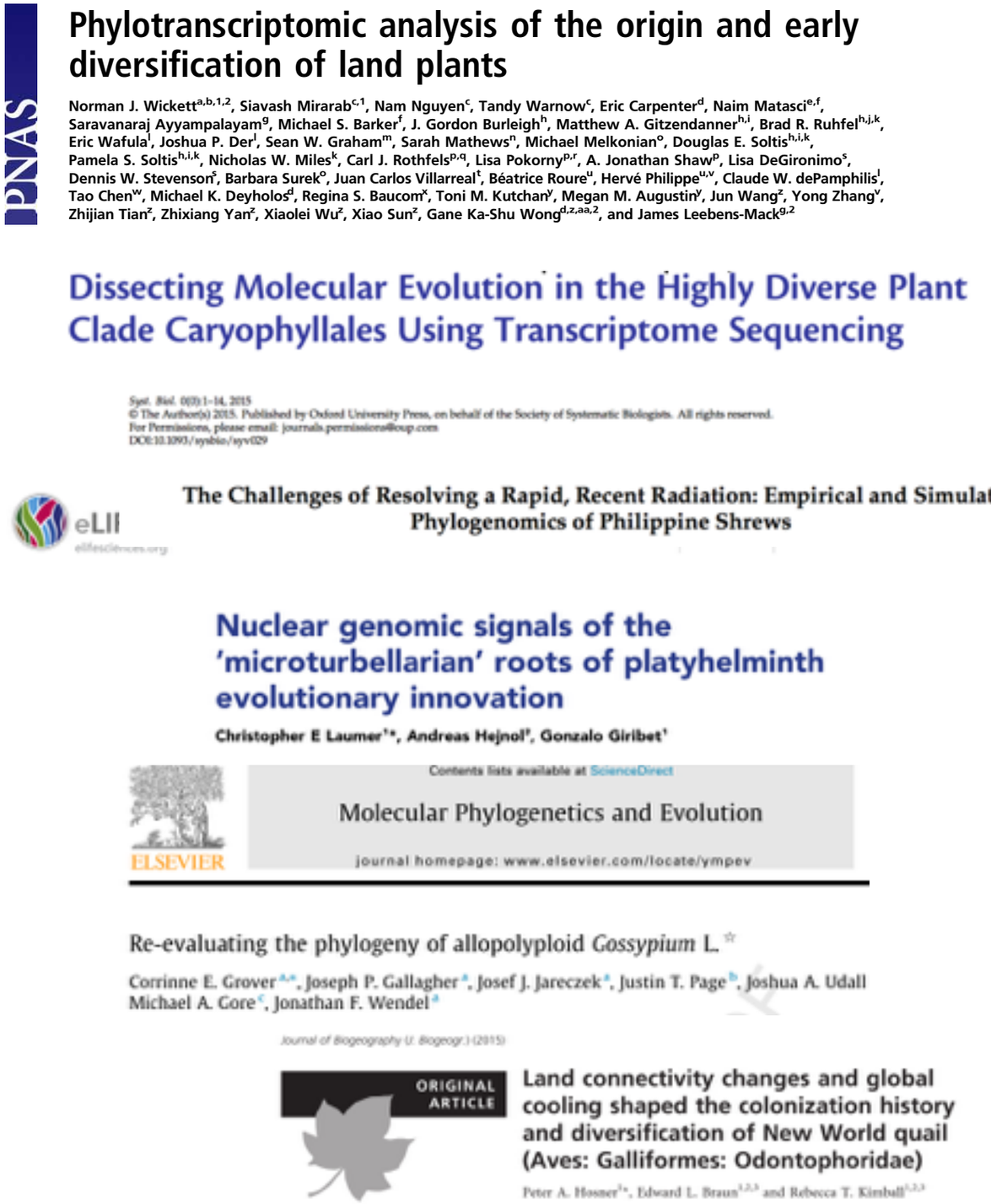
I'll use the term "gene" to refer to "c-genes":
recombination-free orthologous stretches of the genome

X

ASTRAL on biological datasets

- 1KP: **103** plant species, 400-800 genes
- Yang, et al. **96** Caryophyllales species, 1122 genes
- Dentinger, et al. **39** mushroom species, 208 genes
- Giarla and Esselstyn. **19** Philippine shrew species, 1112 genes
- Laumer, et al. **40** flatworm species, 516 genes
- Grover, et al. **8** cotton species, 52 genes
- Hosner, Braun, and Kimball. **28** quail species, 11 genes
- Simmons and Gatesy. **47** angiosperm species, 310 genes

X


PNAS

Phylotranscriptomic analysis of the origin and early diversification of land plants

Norman J. Wickett[a,b,1,2], Siavash Mirarab[c,1], Nam Nguyen[c], Tandy Warnow[c], Eric Carpenter[d], Naim Matasci[e,f], Saravanaraj Ayyampalayam[g], Michael S. Barker[f], J. Gordon Burleigh[h], Matthew A. Gitzendanner[h,i], Brad R. Ruhfel[h,j,k], Eric Wafula[l], Joshua P. Der[l], Sean W. Graham[m], Sarah Mathews[n], Michael Melkonian[o], Douglas E. Soltis[h,i,k], Pamela S. Soltis[h,i,k], Nicholas W. Miles[k], Carl J. Rothfels[p,q], Lisa Pokorny[p,r], A. Jonathan Shaw[p], Lisa DeGironimo[s], Dennis W. Stevenson[s], Barbara Surek[o], Juan Carlos Villarreal[t], Béatrice Roure[u], Hervé Philippe[u,v], Claude W. dePamphilis[l], Tao Chen[w], Michael K. Deyholos[d], Regina S. Baucom[x], Toni M. Kutchan[y], Megan M. Augustin[y], Jun Wang[z], Yong Zhang[y], Zhijian Tian[z], Zhixiang Yan[z], Xiaolei Wu[z], Xiao Sun[z], Gane Ka-Shu Wong[d,z,aa,2], and James Leebens-Mack[g,2]

Dissecting Molecular Evolution in the Highly Diverse Plant Clade Caryophyllales Using Transcriptome Sequencing

Syst. Biol. 0(0):1–14, 2015
© The Author(s) 2015. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
For Permissions, please email: journals.permissions@oup.com
DOI:10.1093/sysbio/syv029

eLH

The Challenges of Resolving a Rapid, Recent Radiation: Empirical and Simulated Phylogenomics of Philippine Shrews

Nuclear genomic signals of the 'microturbellarian' roots of platyhelminth evolutionary innovation

Christopher E Laumer[1*], Andreas Hejnol[2], Gonzalo Giribet[1]

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

ELSEVIER

Re-evaluating the phylogeny of allopolyploid *Gossypium* L.☆

Corrinne E. Grover[a,*], Joseph P. Gallagher[a], Josef J. Jareczek[a], Justin T. Page[b], Joshua A. Udall, Michael A. Gore[c], Jonathan F. Wendel[a]

Journal of Biogeography (*J. Biogeogr.*) (2015)

ORIGINAL ARTICLE

Land connectivity changes and global cooling shaped the colonization history and diversification of New World quail (Aves: Galliformes: Odontophoridae)

Peter A. Hosner[1*], Edward L. Braun[1,2,3] and Rebecca T. Kimball[1,2,3]

Future datasets

- **1200** plants with ~ 400 genes (1KP consortium)
- **250** avian species with 2000 genes (with LSU, UF, and Smithsonian)
- **200** avian species with whole genomes (with Genome 10K, international)
- **250** suboscine species (birds) with ~2000 genes (with LSU and Tulane)
- **140** Insects with 1400 genes (with U. Illinois at Urbana-Champaign)

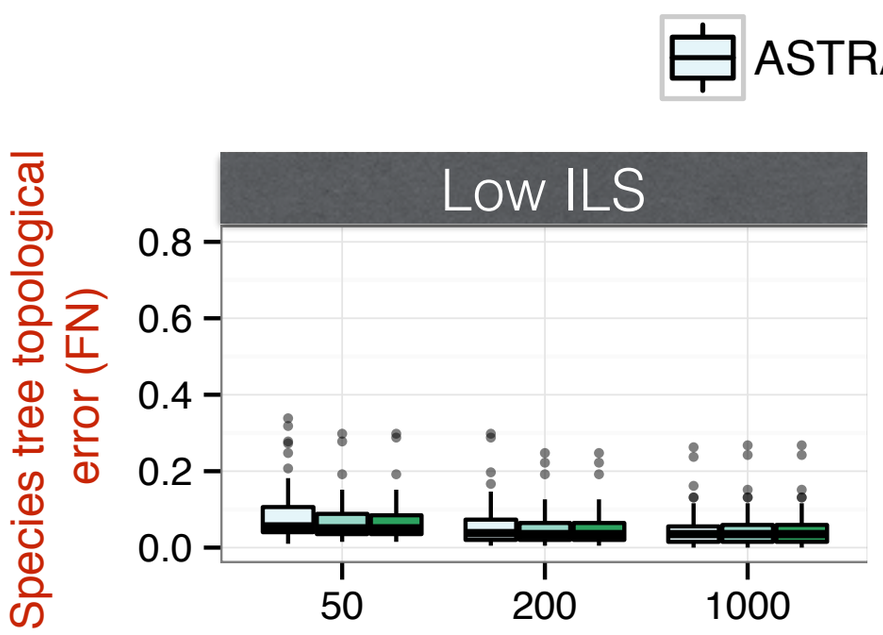
Shortcomings of ASTRAL-I

- Even the constrained version was too slow for more than about 200 species and hundreds of genes

- The constraint set $\mathcal{X}$ did not include true species tree branches for some challenging datasets, resulting in low accuracy in some cases

- Input gene trees could not have polytomies

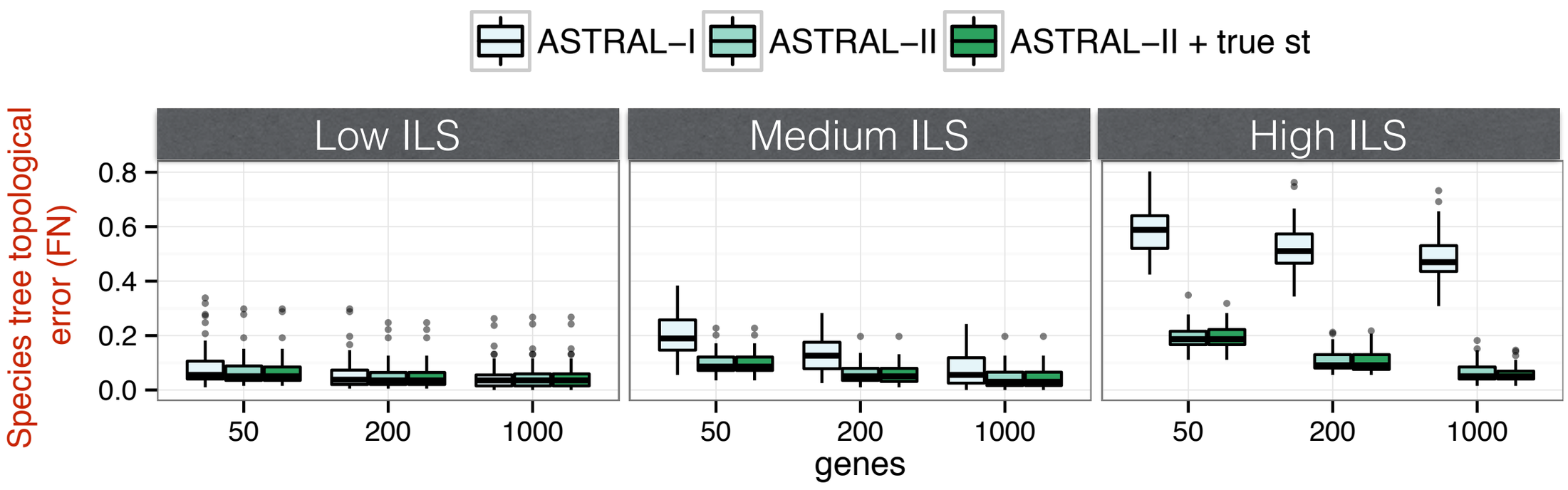
ASTRAL-I versus ASTRAL-II



200 species, deep ILS

X

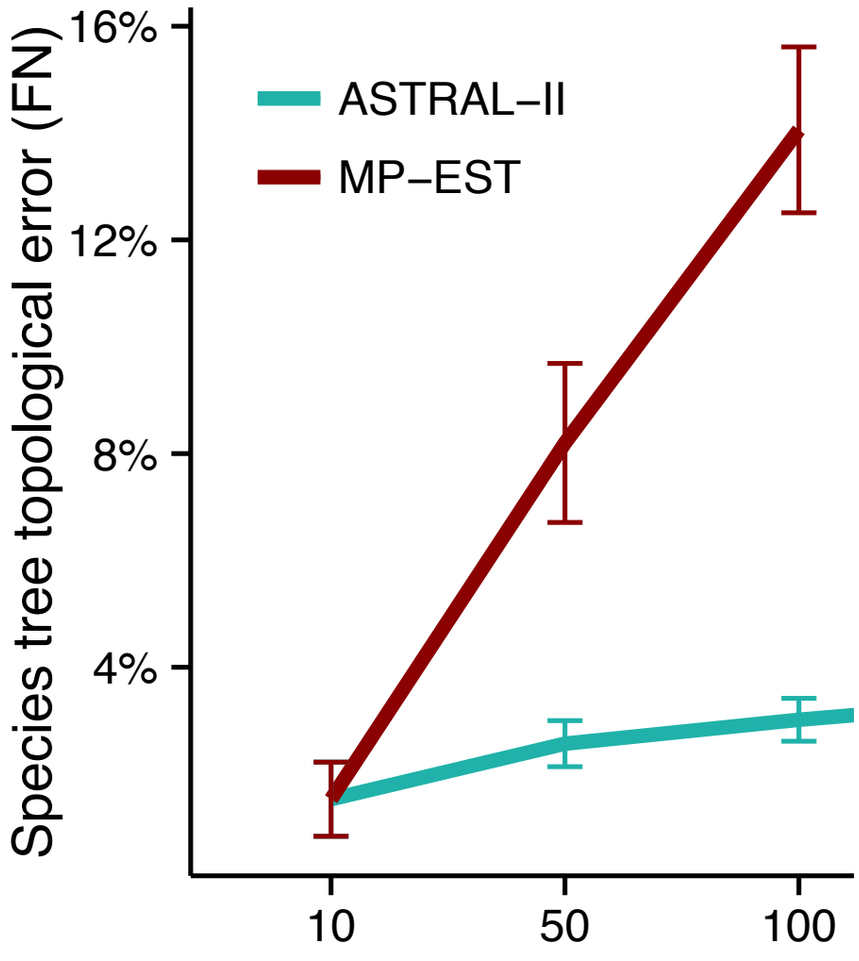
ASTRAL-I versus ASTRAL-II



200 species, deep ILS

X

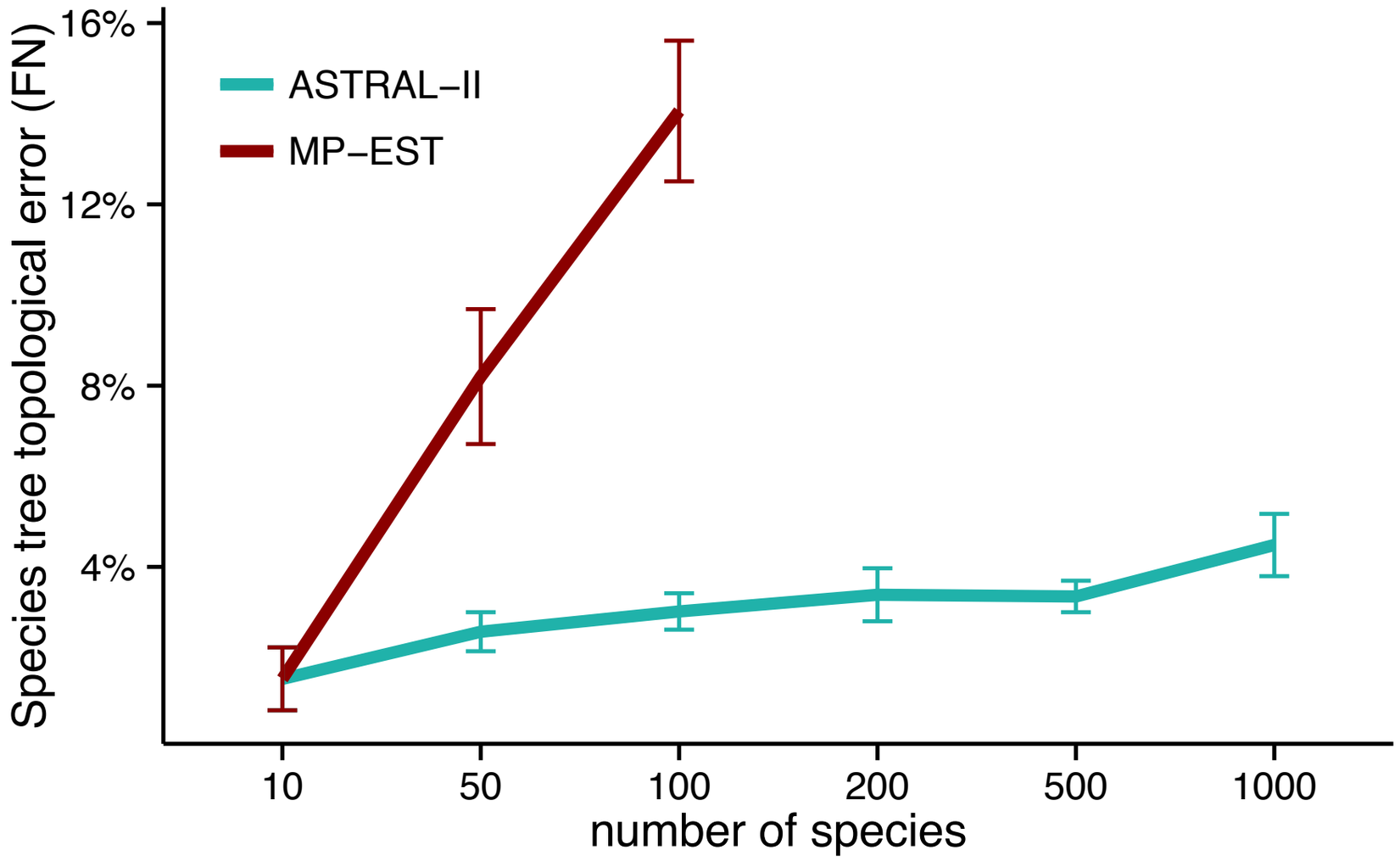
Tree error, varying # of species



1000 genes, “medium” levels of recent ILS

X

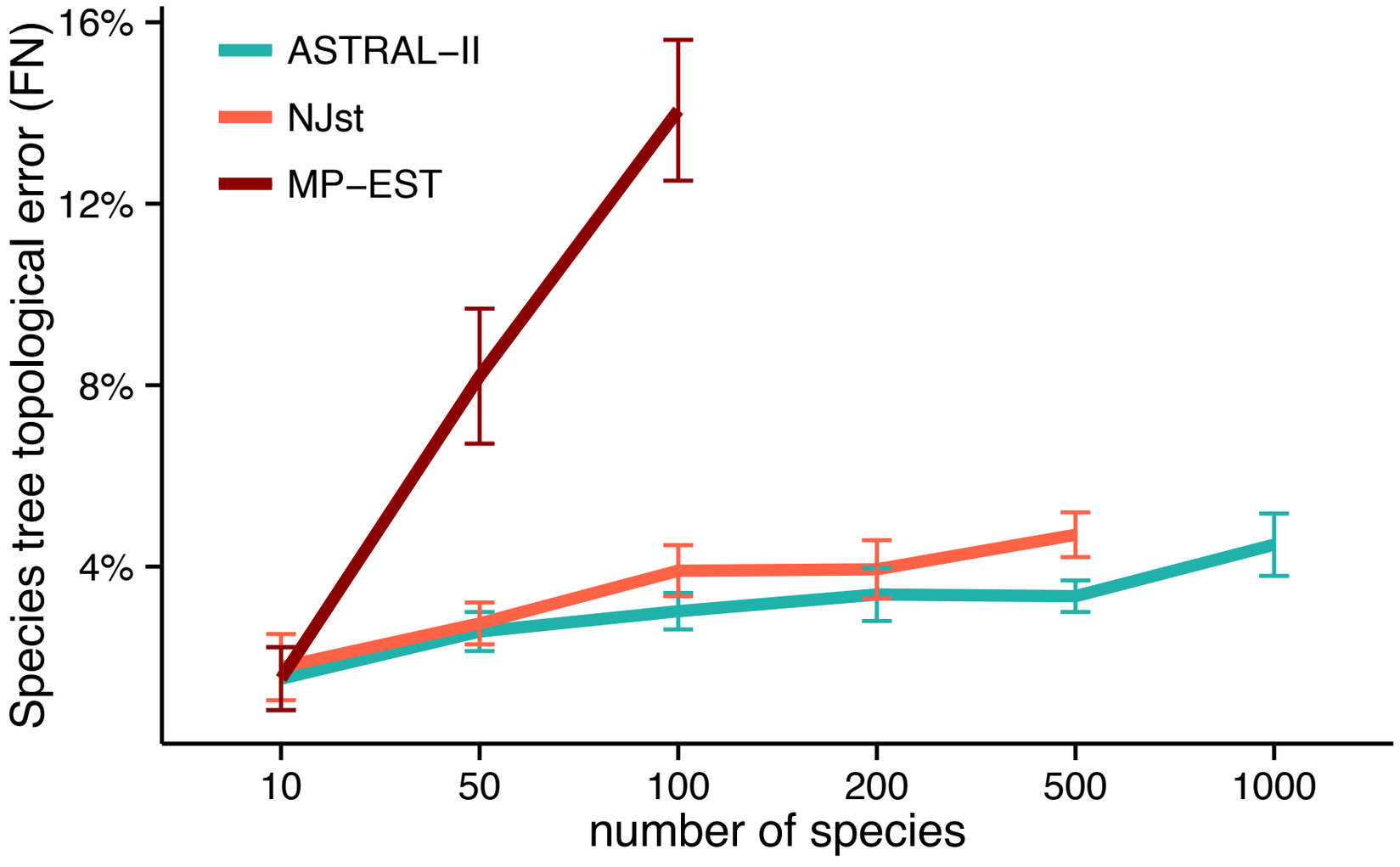
Tree error, varying # of species



1000 genes, “medium” levels of recent ILS

X

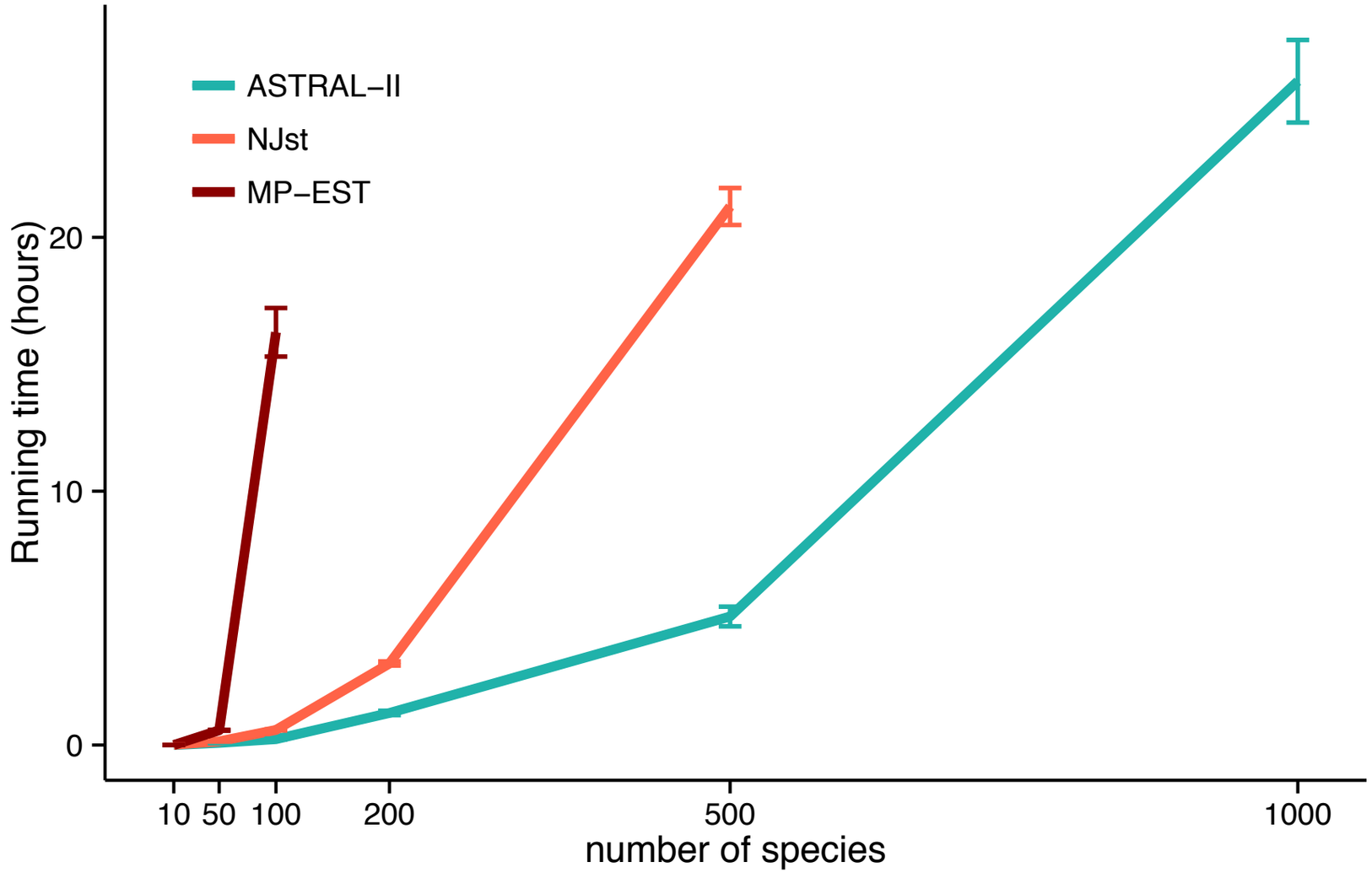
Tree error, varying # of species



1000 genes, “medium” levels of recent ILS

X

Running time, varying # of species



1000 genes, “medium” levels of recent ILS

X

Acknowledgments

Tandy Warnow

Keshav Pingali

Jim Leebens-mack
(UGA)

Norman Wickett
(U Chicago)

Gane Wong
(U of Alberta)

. . .

S.M. Bayzid

Théo
Zimmermann

Travel funding to ISMB/ECCB 2015
was generously provided by akamia.