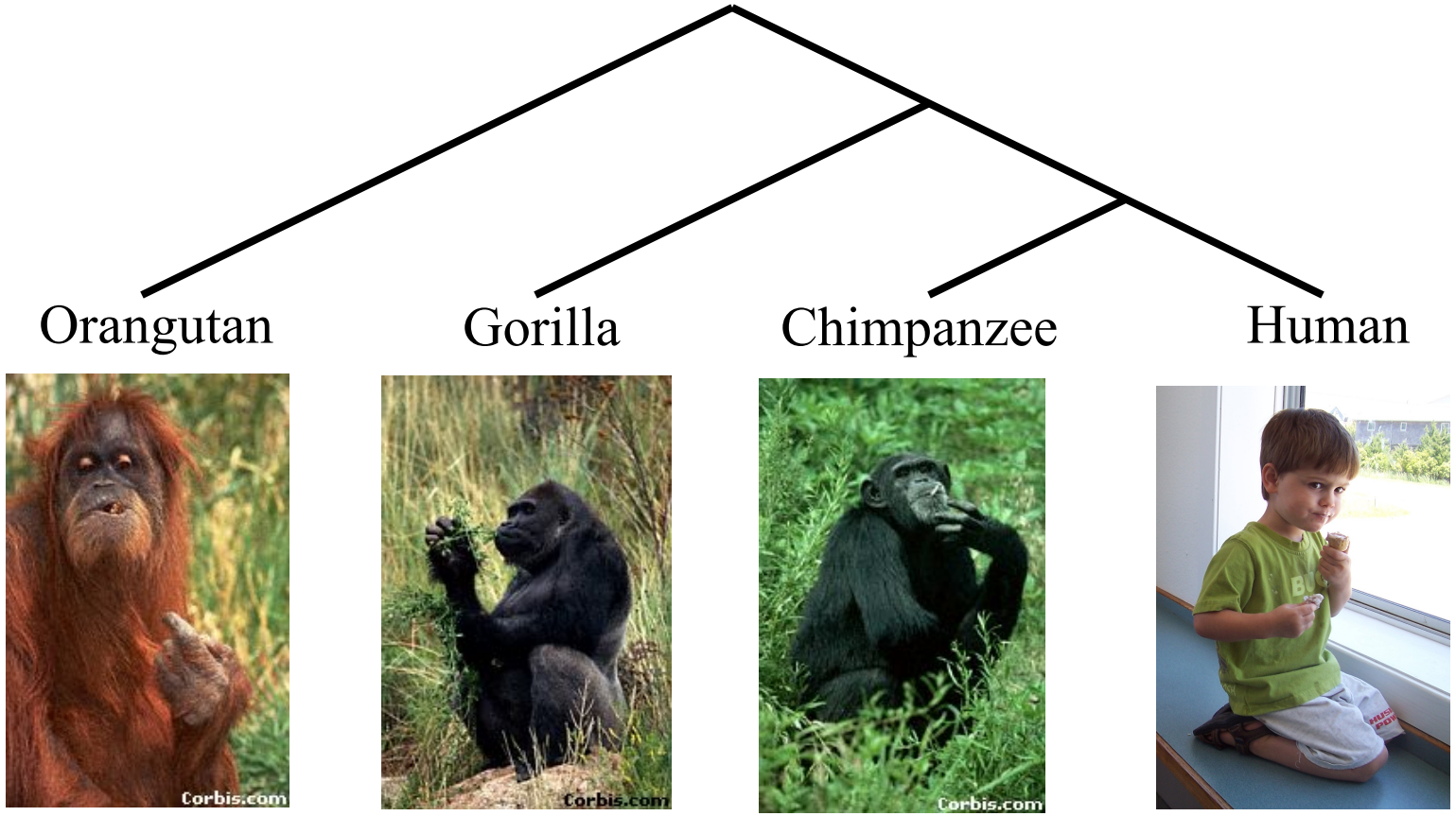
This young bird closely resembles its parents which are about a third larger. The white flecks on the face are mallophaga eggs. Most bird lice taxa found on Hoatzins are unique to this host.

Mathematical and Computational Challenges in Reconstructing Evolution

Tandy Warnow

The University of Illinois

Phylogeny (evolutionary tree)



From the Tree of the Life Website, University of Arizona

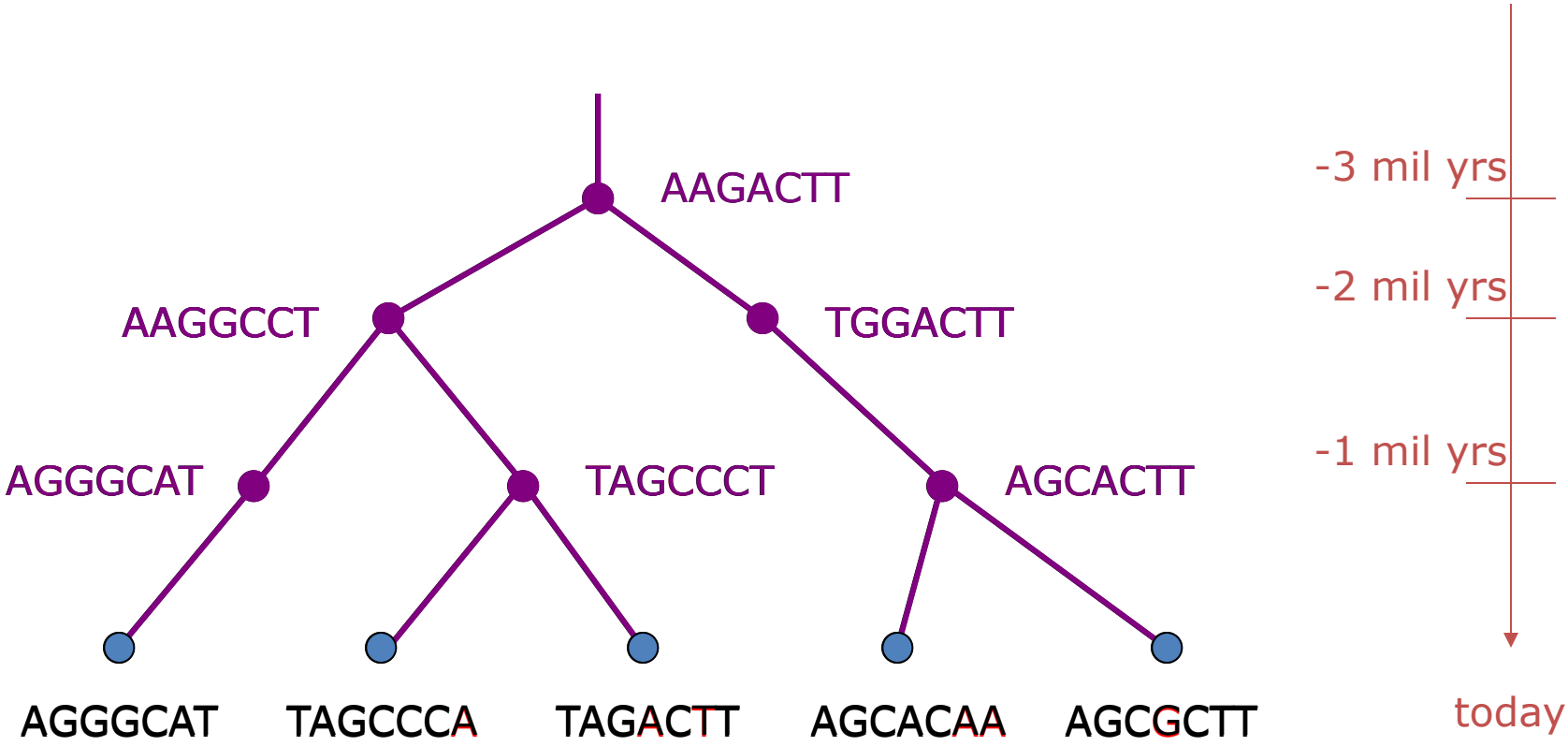
- “Nothing in biology makes sense except in the light of evolution”
  - Theodosius Dobzhansky, 1973 essay in the American Biology Teacher, vol. 35, pp 125-129

- *“…… nothing in evolution makes sense except in the light of phylogeny ...”*
  - *Society of Systematic Biologists, http://systbio.org/teachevolution.html*

This Talk

- Models of evolution, identifiability, statistical consistency
- Genome-scale phylogeny
- Open questions
- Future directions

DNA Sequence Evolution (Idealized)



Markov Models of Sequence Evolution

The different sites are assumed to evolve ***i.i.d***. down the model tree (with rates that are drawn from a gamma distribution).

Markov Models of Sequence Evolution

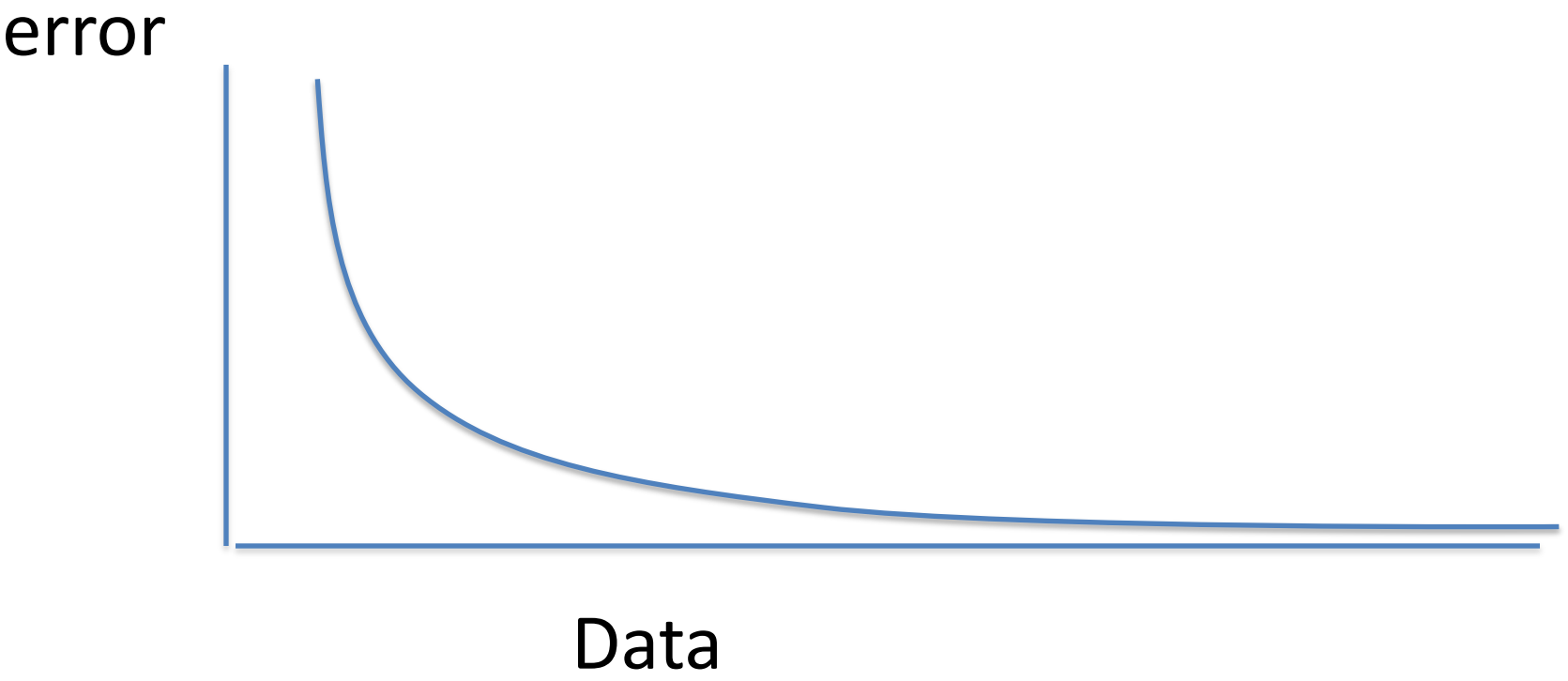
The different sites are assumed to evolve ***i.i.d***. down the model tree (with rates that are drawn from a gamma distribution).

Simplest site evolution model (Jukes-Cantor, 1969):

- The model tree T is binary and has substitution probabilities p(e) on each edge e, with $0<p(e)<3/4$.
- The state at the root is randomly drawn from {A,C,T,G} (nucleotides)
- If a site (position) changes on an edge, it changes with equal probability to each of the remaining states.
- The evolutionary process is Markovian.

More complex models (such as the General Markov model) are also considered, often with little change to the theory.

Statistical Consistency



Questions

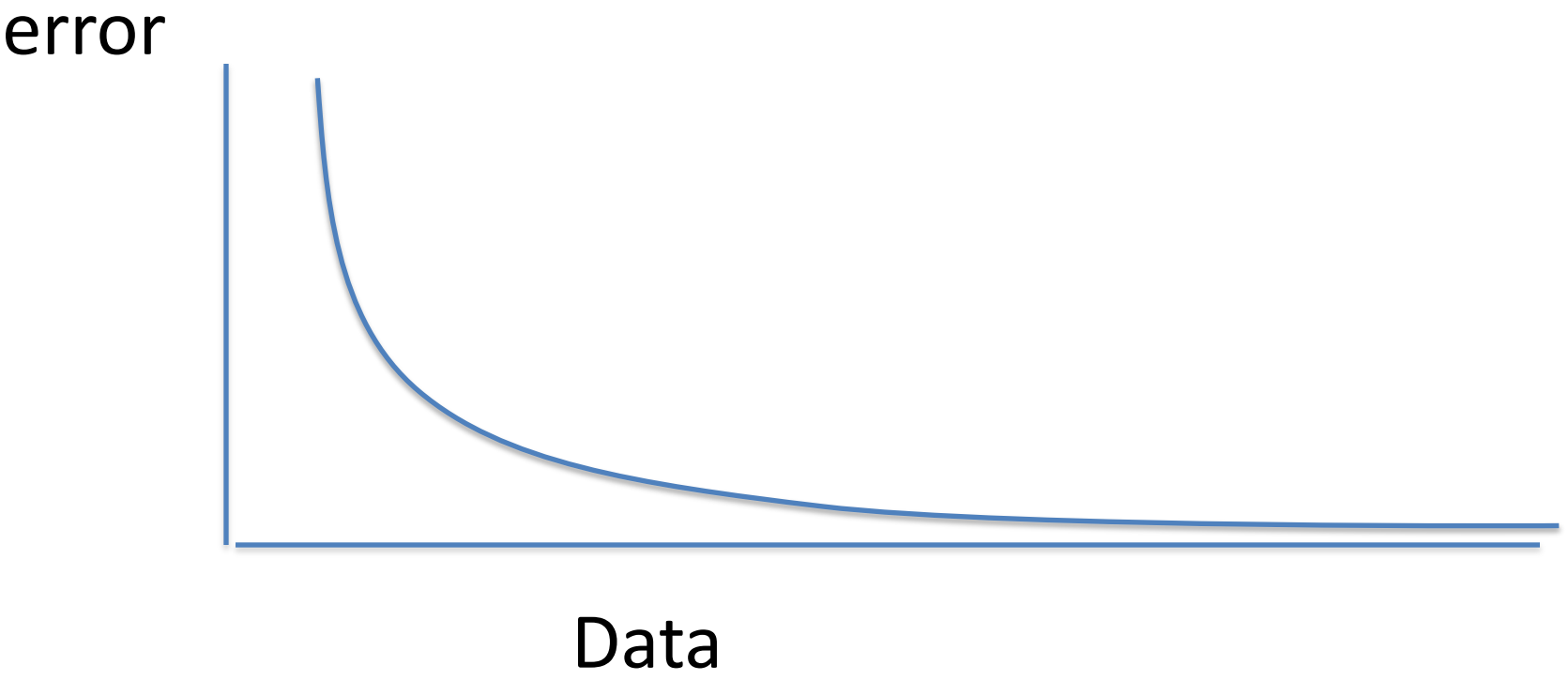
- Is the model tree identifiable?
- Which estimation methods are statistically consistent under this model?
- How much data does the method need to estimate the model tree correctly with high probability (i.e., what is the method's sample complexity)?

Answers?

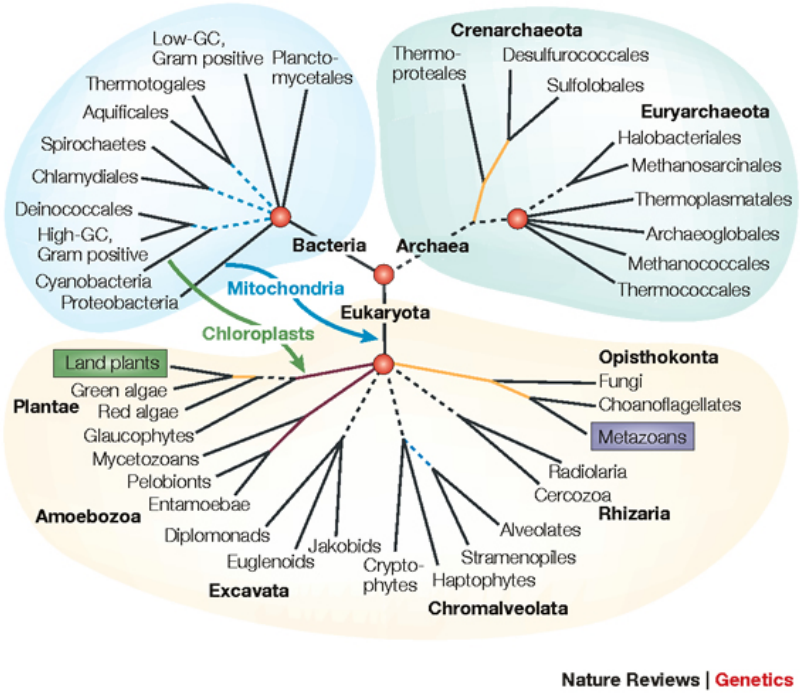
- We know a lot about which site evolution models are identifiable, and which methods are statistically consistent.
- We know a little bit about the sample complexity for standard methods.

Take home message: need to limit (or not allow) heterogeneity to get identifiability!

Genome-scale data?



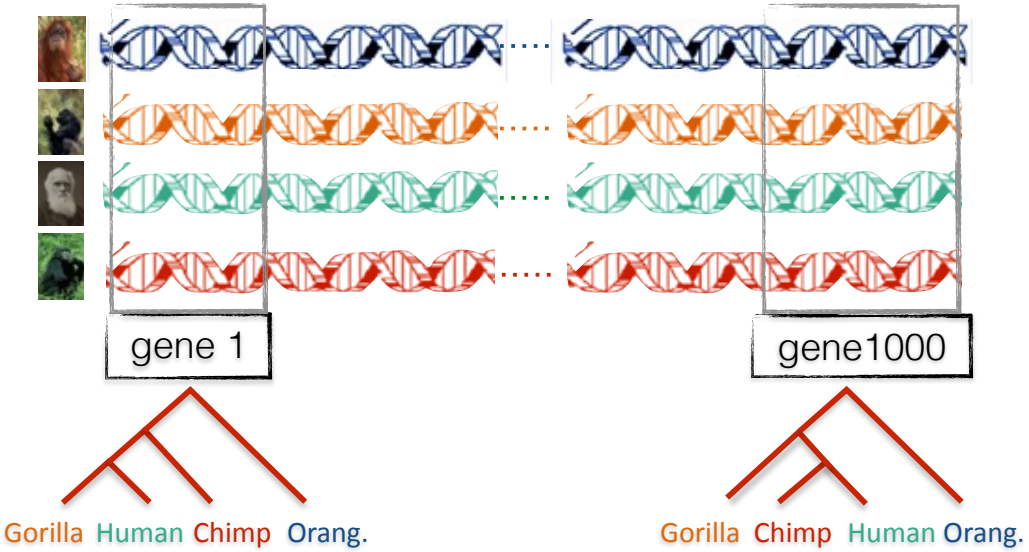
Phylogenomics



Phylogeny + genomics = genome-scale phylogeny estimation

.

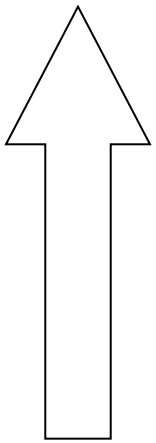
Gene tree discordance



Incomplete Lineage Sorting (ILS) is a dominant cause of gene tree heterogeneity

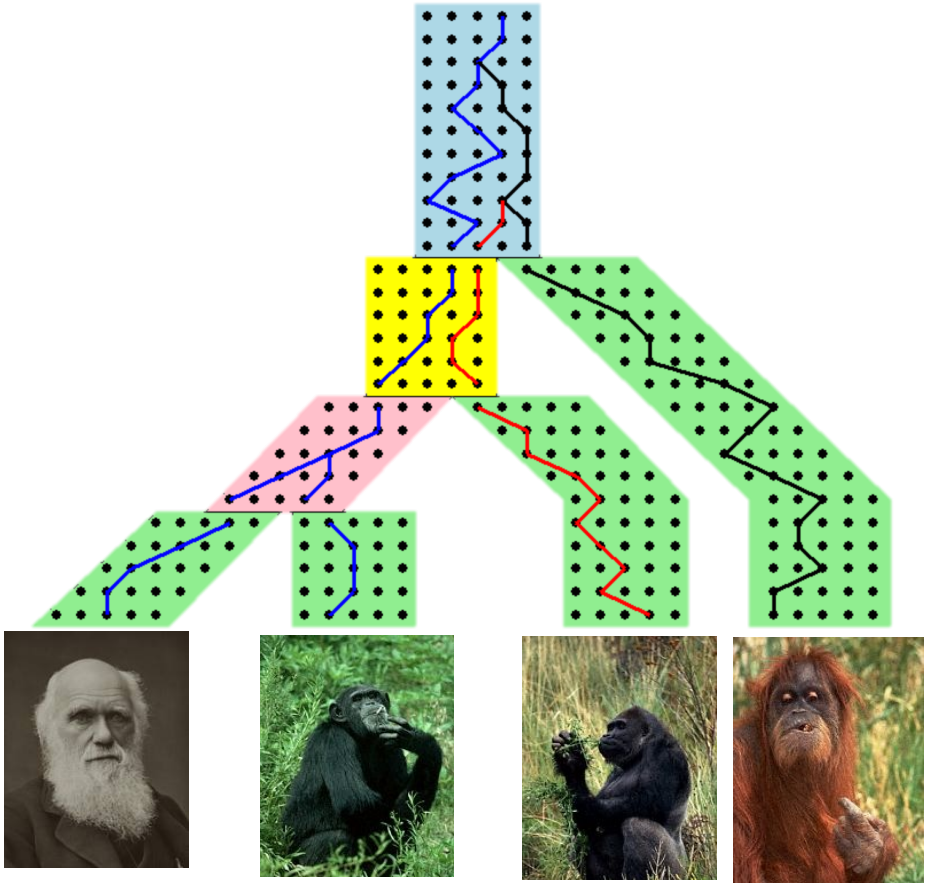
Gene trees inside the species tree (Coalescent Process)

Past



Present

Courtesy James Degnan



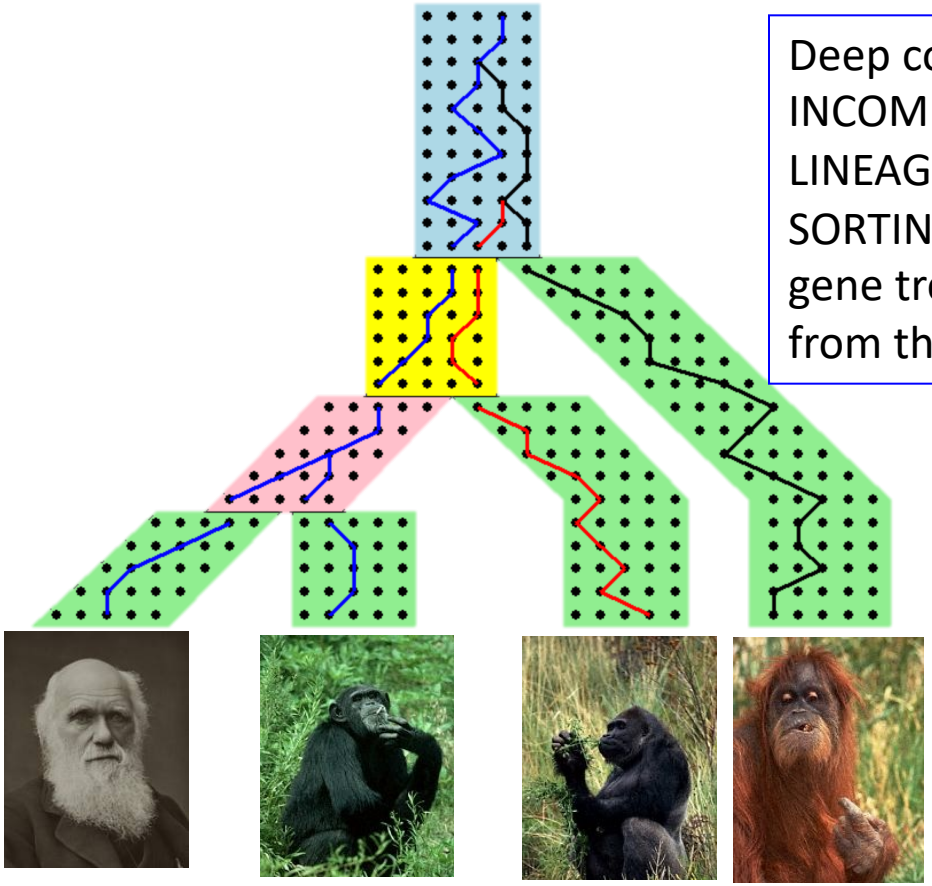
Gorilla and Orangutan are not siblings in the species tree, but they are in the gene tree.

Gene trees inside the species tree (Coalescent Process)

Past

Present

Courtesy James Degnan



Deep coalescence = INCOMPLETE LINEAGE SORTING (ILS): gene tree can be different from the species tree

Gorilla and Orangutan are not siblings in the species tree, but they are in the gene tree.

1KP: Thousand Transcriptome Project







G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

N. Wickett
Northwestern

N. Matasci
iPlant

T. Warnow,
UT-Austin

S. Mirarab,
UT-Austin

N. Nguyen
UT-Austin

- 103 plant transcriptomes, 400-800 single copy "genes"
- Next phase will be much bigger
- Wickett, Mirarab et al., *PNAS* 2014

Major Challenge:

- Massive gene tree heterogeneity consistent with ILS

Avian Phylogenomics Project

Erich Jarvis, HHMI

MTP Gilbert, Copenhagen

Guojie Zhang, BGI

Siavash Mirarab, Texas

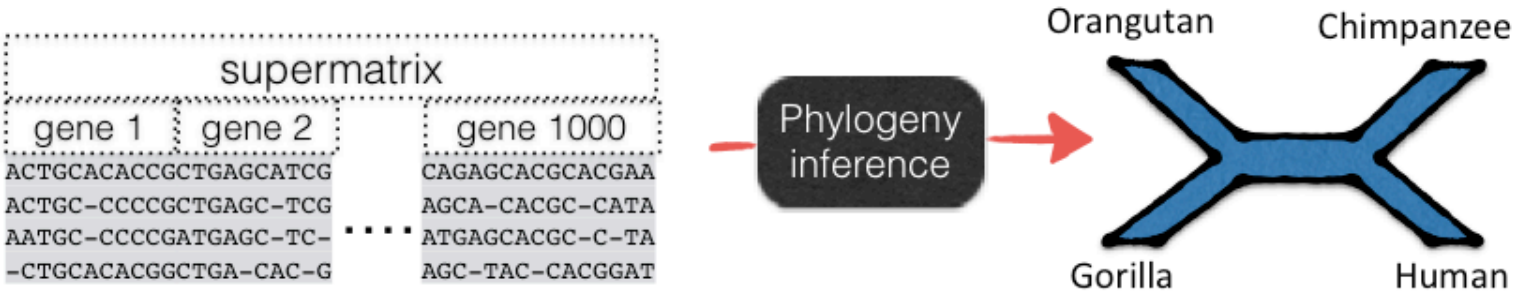
Tandy Warnow, Texas and UIUC

- Approx. 50 species, whole genomes
- 14,000 loci
- Multi-national team (100+ investigators)
- 8 papers published in special issue of Science 2014

Major challenge:

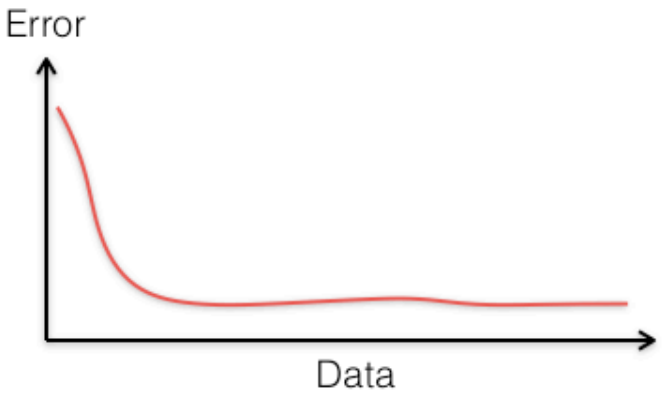
- Massive gene tree heterogeneity consistent with ILS.

Traditional approach: concatenation



- Statistically <u>inconsistent</u> and can even be positively misleading (proved for unpartitioned maximum likelihood)
  [Roch and Steel, Theo. Pop. Gen., 2014]

- Mixed accuracy in simulations
  [Kubatko and Degnan, Systematic Biology, 2007]
  [<u>Mirarab</u>, et al., Systematic Biology, 2014]



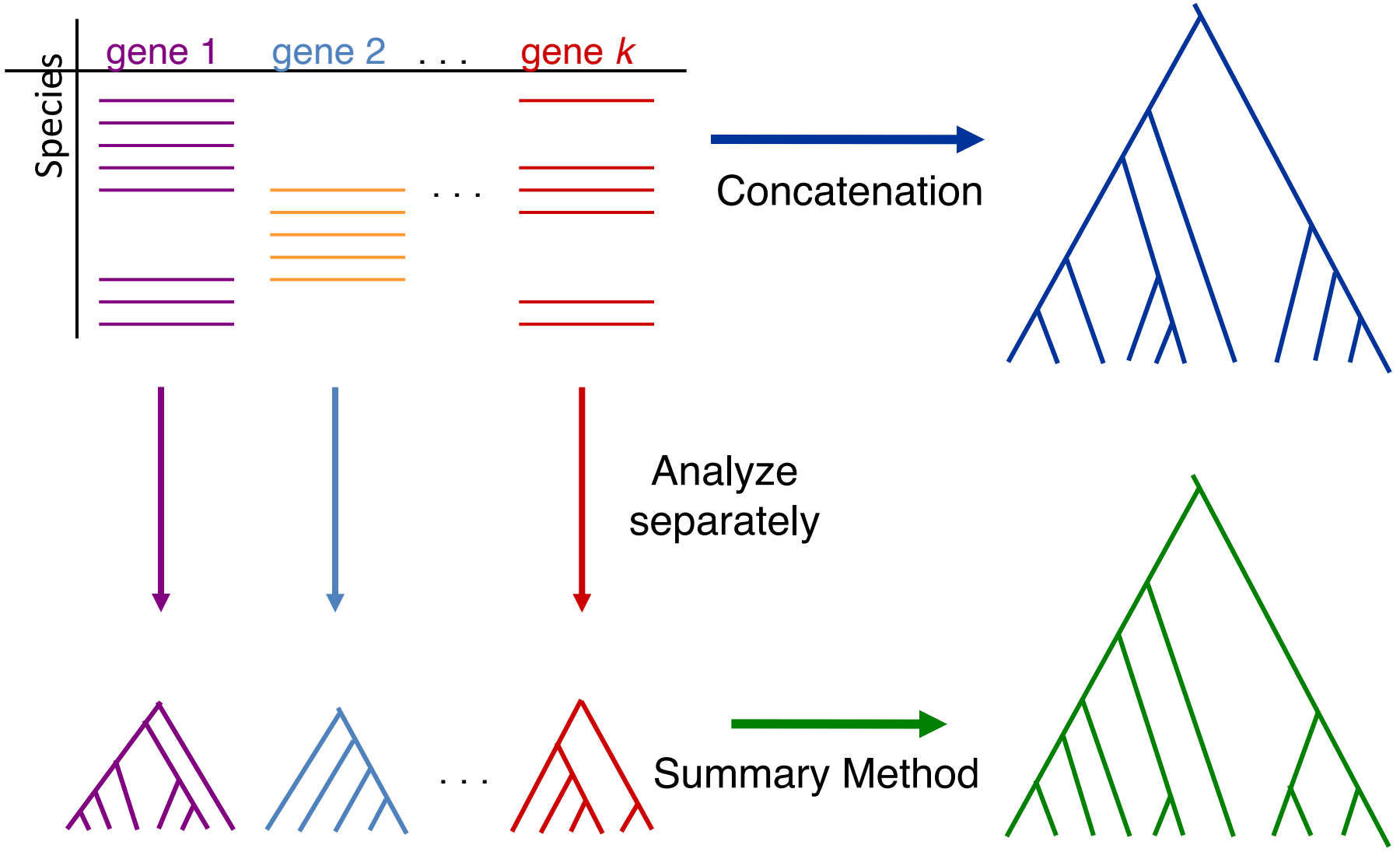
Statistically consistent methods

- **Coalescent-based summary methods:** Estimate gene trees, and then combine together (ASTRAL, ASTRID, MP-EST, NJst, and others)

- **Co-estimation methods**: Co-estimate gene trees and species trees (TOO EXPENSIVE)

- **Site-based methods:** estimate the species tree from the concatenated alignment, and do not estimate gene trees (NOT WELL STUDIED)

Main competing approaches



What about summary methods?

What about summary methods?

Techniques:

- Most frequent gene tree?
- Consensus of gene trees?
- Other?

Species tree estimation from unrooted gene trees

Theorem (Allman et al.): Under the multi-species coalescent model, for any four taxa A, B, C, and D, the most probable unrooted gene tree on {A,B,C,D} is identical to the unrooted species tree induced on {A,B,C,D}.

Species tree estimation from unrooted gene trees

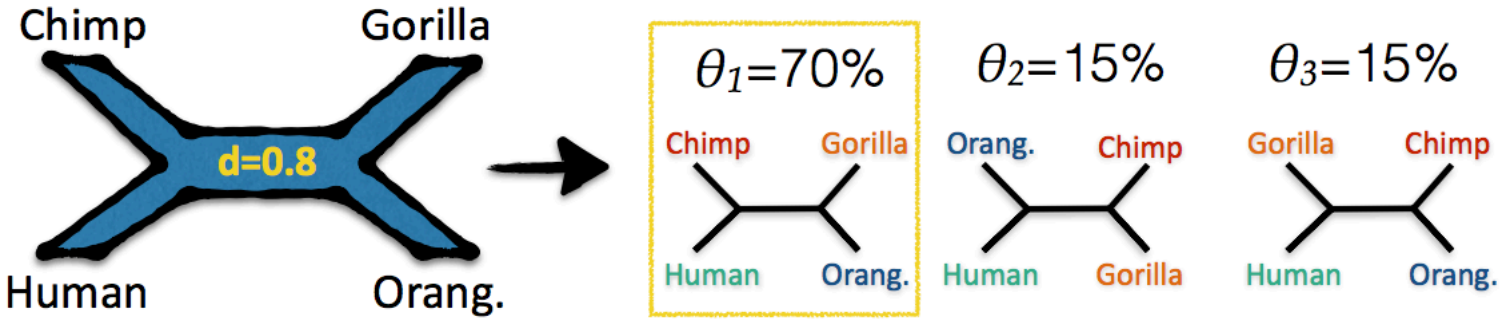
Theorem (Allman et al.): Under the multi-species coalescent model, for any four taxa A, B, C, and D, the most probable unrooted gene tree on {A,B,C,D} is identical to the unrooted species tree induced on {A,B,C,D}.

Proof: For every four species, select most frequently observed tree as the species tree. Then combine quartet trees!

Species tree estimation from unrooted gene trees

Theorem (Allman et al.): Under the multi-species coalescent model, for any four taxa A, B, C, and D, the most probable unrooted gene tree on {A,B,C,D} is identical to the unrooted species tree induced on {A,B,C,D}.

Proof: For every four species, select most frequently observed tree as the species tree. Then combine quartet trees!



ASTRAL

[Mirarab, et al., ECCB/Bioinformatics, 2014]

- Optimization Problem (NP-Hard):

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

Set of quartet trees induced by T

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

a gene tree

all input gene trees

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

Constrained Maximum Quartet Support Tree

- Input: Set $\mathcal{T} = \{t_1, t_2, \ldots, t_k\}$ of unrooted gene trees, with each tree on set S with n species, and set X of allowed bipartitions
- Output: Unrooted tree T on leafset S, maximizing the total quartet tree similarity to $\mathcal{T}$, subject to T drawing its bipartitions from X.

Constrained Maximum Quartet Support Tree

- Input: Set $\mathcal{T} = \{t_1, t_2, \ldots, t_k\}$ of unrooted gene trees, with each tree on set S with n species, and set X of allowed bipartitions
- Output: Unrooted tree T on leafset S, maximizing the total quartet tree similarity to $\mathcal{T}$, subject to T drawing its bipartitions from X.

Theorems (Mirarab et al., 2014):

- If X contains the bipartitions from the input gene trees (and perhaps others), then an exact solution to this problem is statistically consistent under the MSC.

Constrained Maximum Quartet Support Tree

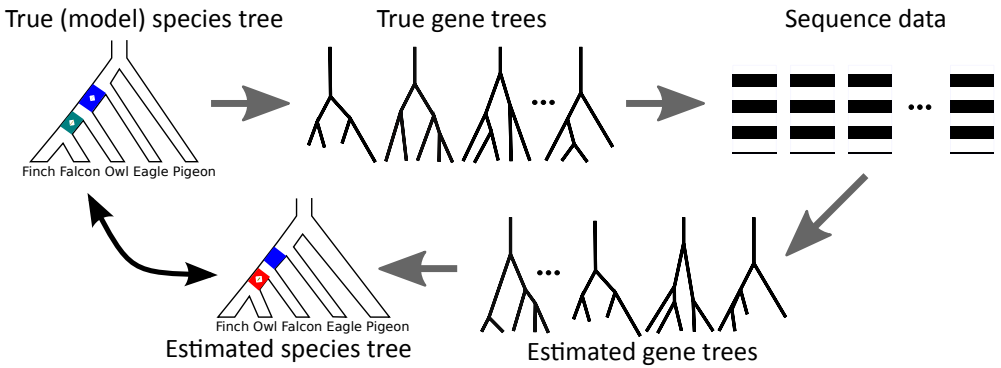
- Input: Set $\mathcal{T} = \{t_1, t_2, \ldots, t_k\}$ of unrooted gene trees, with each tree on set S with n species, and set X of allowed bipartitions
- Output: Unrooted tree T on leafset S, maximizing the total quartet tree similarity to $\mathcal{T}$, subject to T drawing its bipartitions from X.

Theorems (Mirarab et al., 2014):

- If X contains the bipartitions from the input gene trees (and perhaps others), then an exact solution to this problem is statistically consistent under the MSC.
- The constrained MQST problem can be solved in $O(|X|^2 nk)$ time. (We use dynamic programming, and build the unrooted tree from the bottom-up, based on "allowed clades" – halves of the allowed bipartitions.)

Simulation study

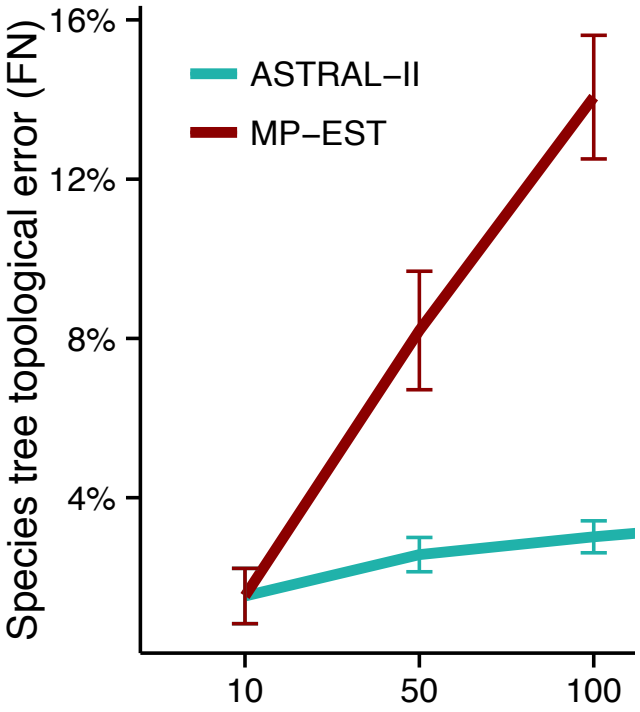
- Variable parameters:
  - Number of species: 10 – 1000
  - Number of genes: 50 – 1000
  - Amount of ILS: low, medium, high
  - Deep versus recent speciation



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

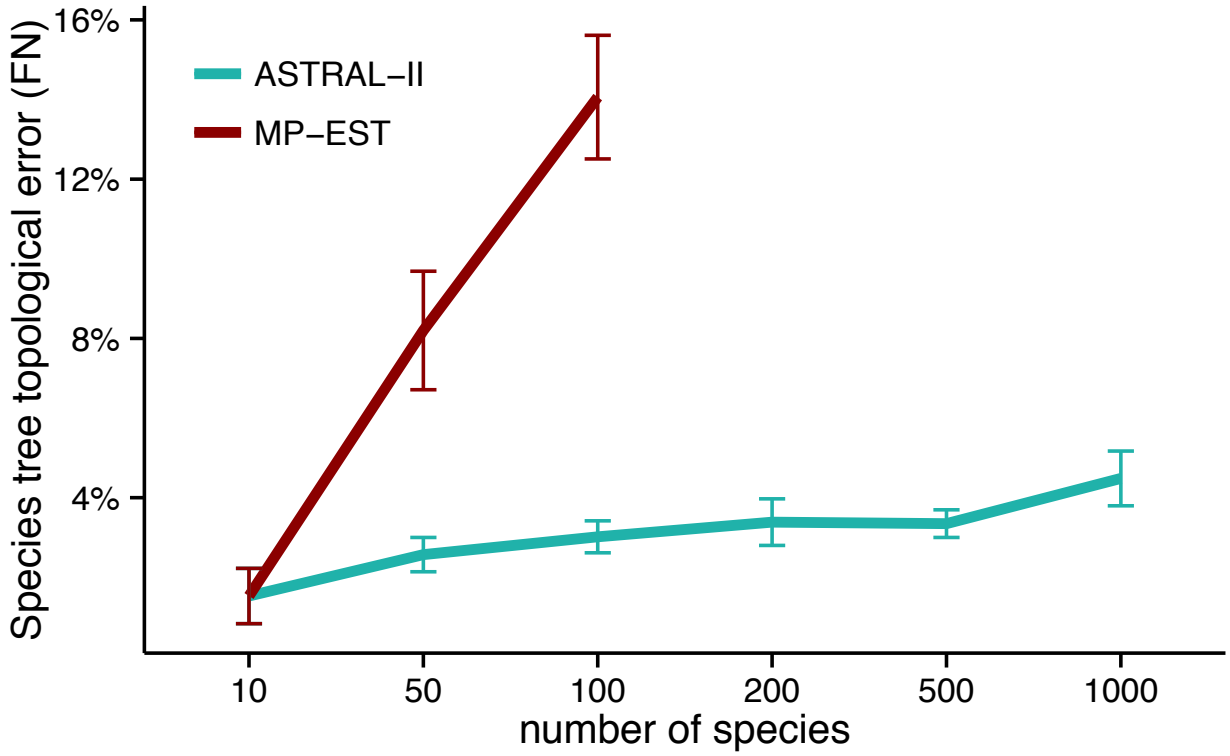
Used SimPhy, Mallo and Posada, 2015

Tree accuracy when varying the number of species



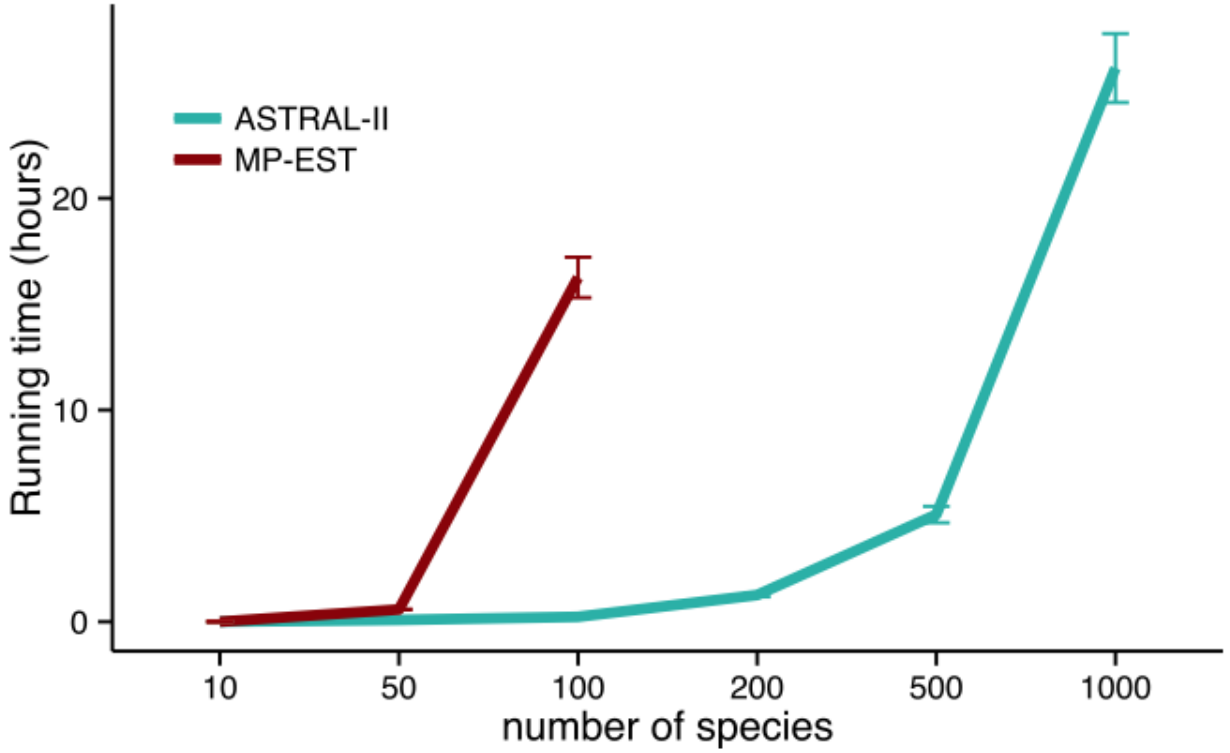
1000 genes, “medium” levels of recent ILS

Tree accuracy when varying the number of species



1000 genes, “medium” levels of recent ILS

Running time as function of # species



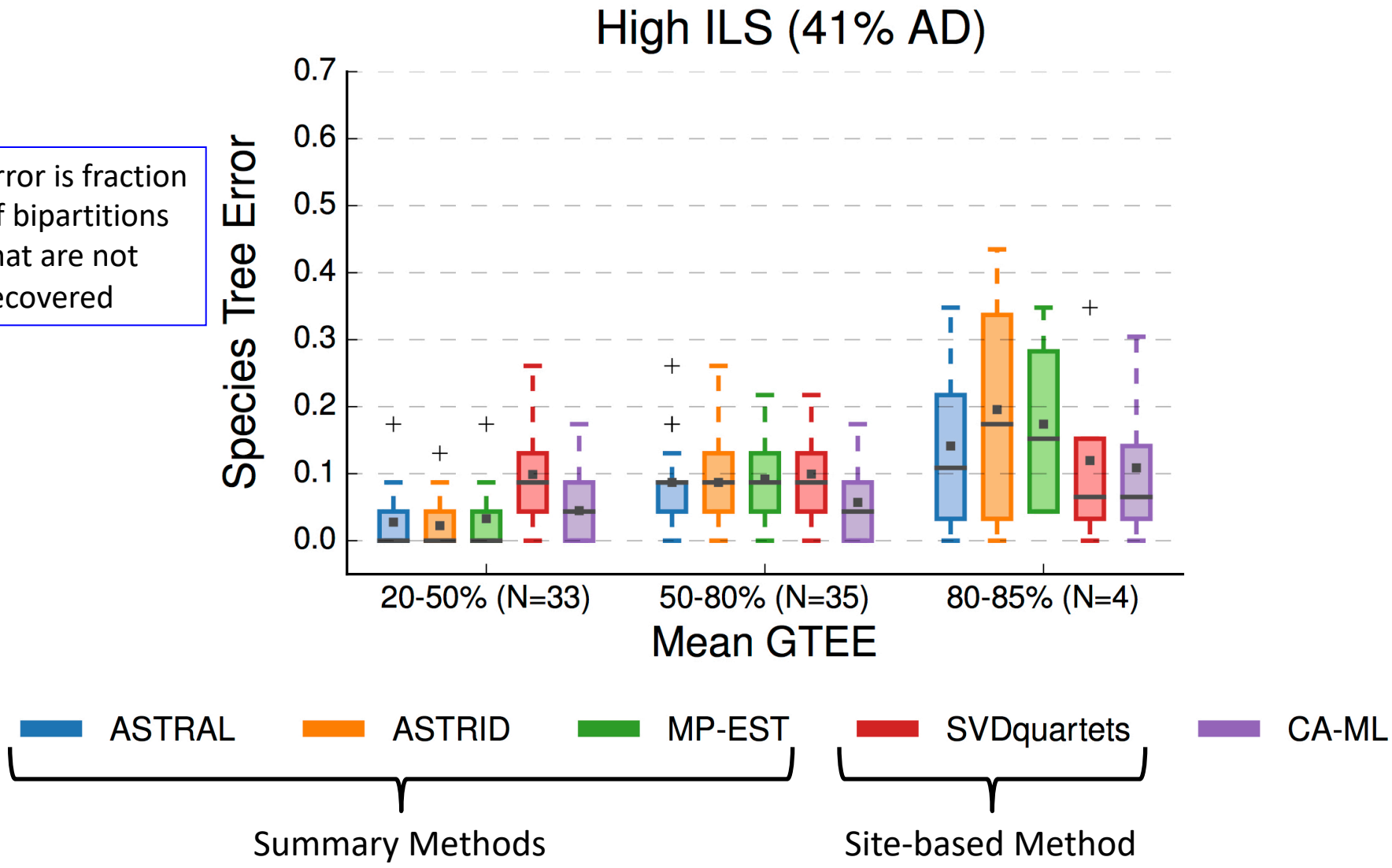
1000 genes, "medium" levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]

X

Impact of Gene Tree Estimation Error

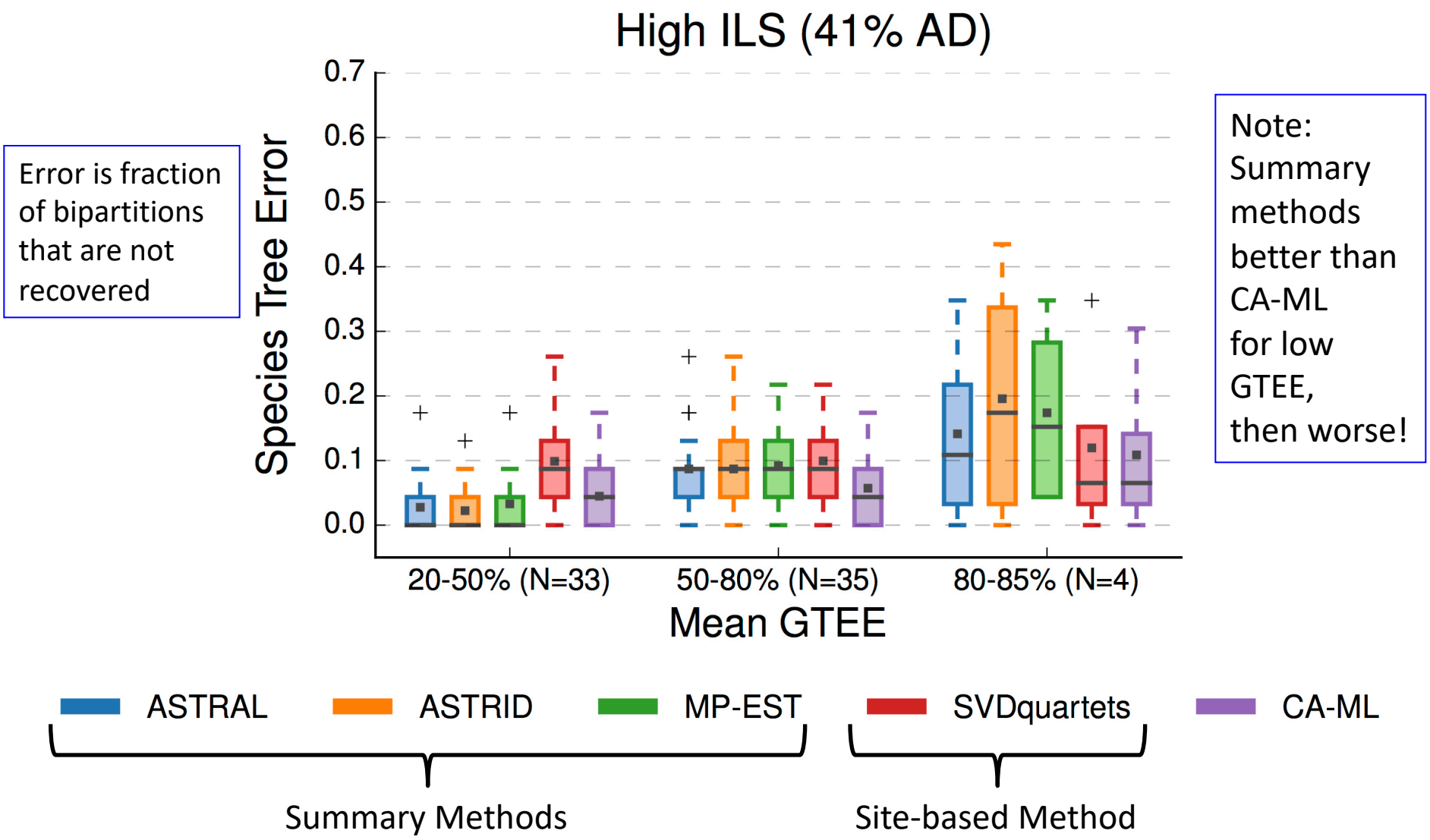
(from Molloy and Warnow 2017)

Error is fraction of bipartitions that are not recovered



Impact of Gene Tree Estimation Error

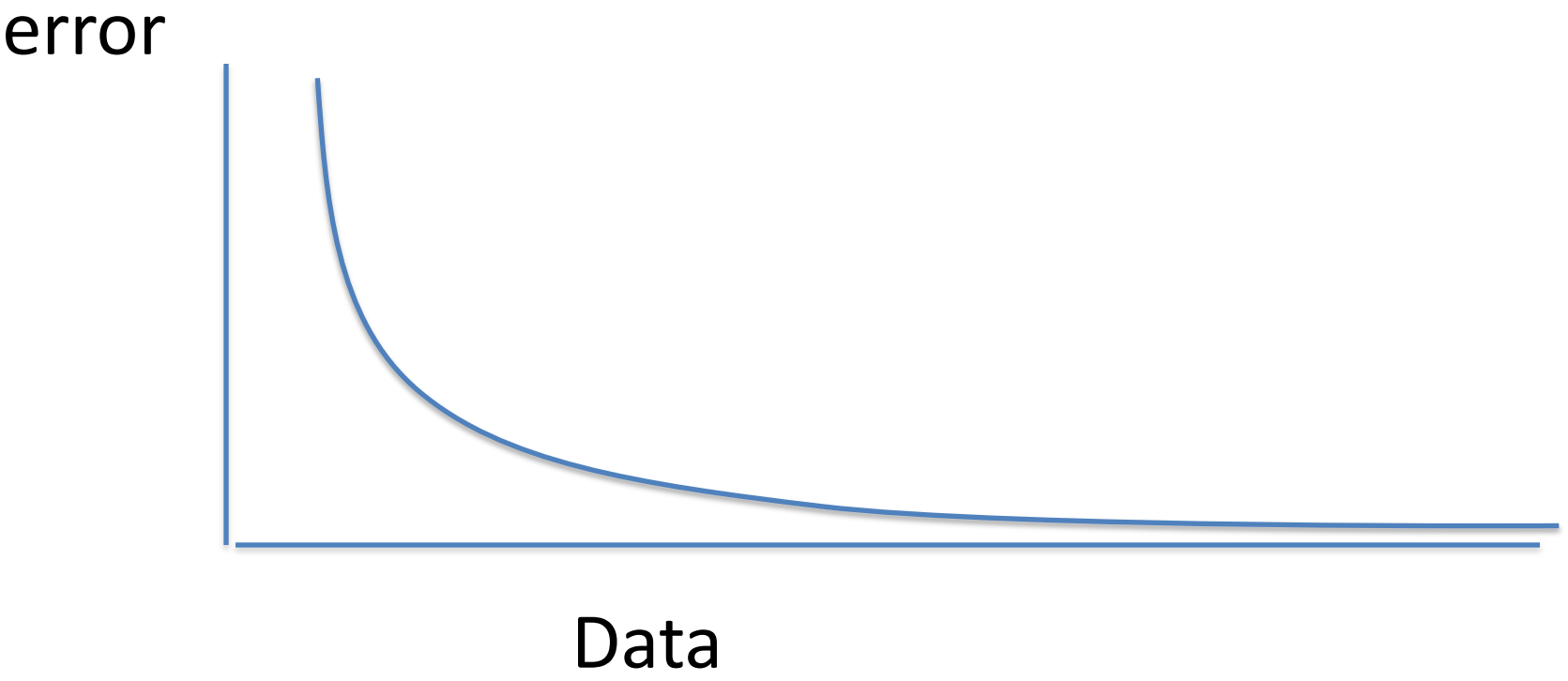
(from Molloy and Warnow 2017)



Error is fraction of bipartitions that are not recovered

Note: Summary methods better than CA-ML for low GTEE, then worse!

Statistical Consistency for summary methods



Data are gene trees, presumed to be randomly sampled true gene trees.

Gene tree estimation error: key issue in the debate

- Multiple studies show that *summary methods can be less accurate than concatenation* in the presence of high gene tree estimation error.
- Genome-scale data includes a range of markers, not all of which have substantial signal. Furthermore, removing sites due to model violations reduces signal.
- Some researchers also argue that "gene trees" should be based on very short alignments, to avoid intra-locus recombination.

What about performance on bounded number of sites?

- Question #1: Do any summary methods converge to the species tree as the number of loci increase, but where each locus has only a constant number of sites?
- Answer #1: Roch & Warnow, Syst Biol, March 2015:
  - Strict molecular clock: Yes for some new methods, even for a single site per locus
  - No clock: Unknown for all methods, including MP-EST, ASTRAL, etc.

S. Roch and T. Warnow. "On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods", Systematic Biology, 64(4):663-676, 2015

What about performance on bounded number of sites?

- Question #1: Do any summary methods converge to the species tree as the number of loci increase, but where each locus has only a constant number of sites?
- Answer #2: Roch, Nute, & Warnow, Syst Biol 2019.
  - No! Summary methods are not only not consistent, they can be positively misleading! (Felsenstein Zone)

What about performance on bounded number of sites?

- Question #2: What about concatenation using maximum likelihood?
- Answer: Roch, Nute, & Warnow, Syst Biol 2019
  - Not if fully partitioned! Concatenation using maximum likelihood, if fully partitioned is also not consistent and can be positively misleading (even if there is NO ILS)! (Felsenstein Zone)

Statistically consistent methods

- **Coalescent-based summary methods:** Estimate gene trees, and then combine together (ASTRAL, ASTRID, MP-EST, NJst, and others)

- **Co-estimation methods**: Co-estimate gene trees and species trees (TOO EXPENSIVE)

- **Site-based methods:** estimate the species tree from the concatenated alignment, and do not estimate gene trees (NOT WELL STUDIED)

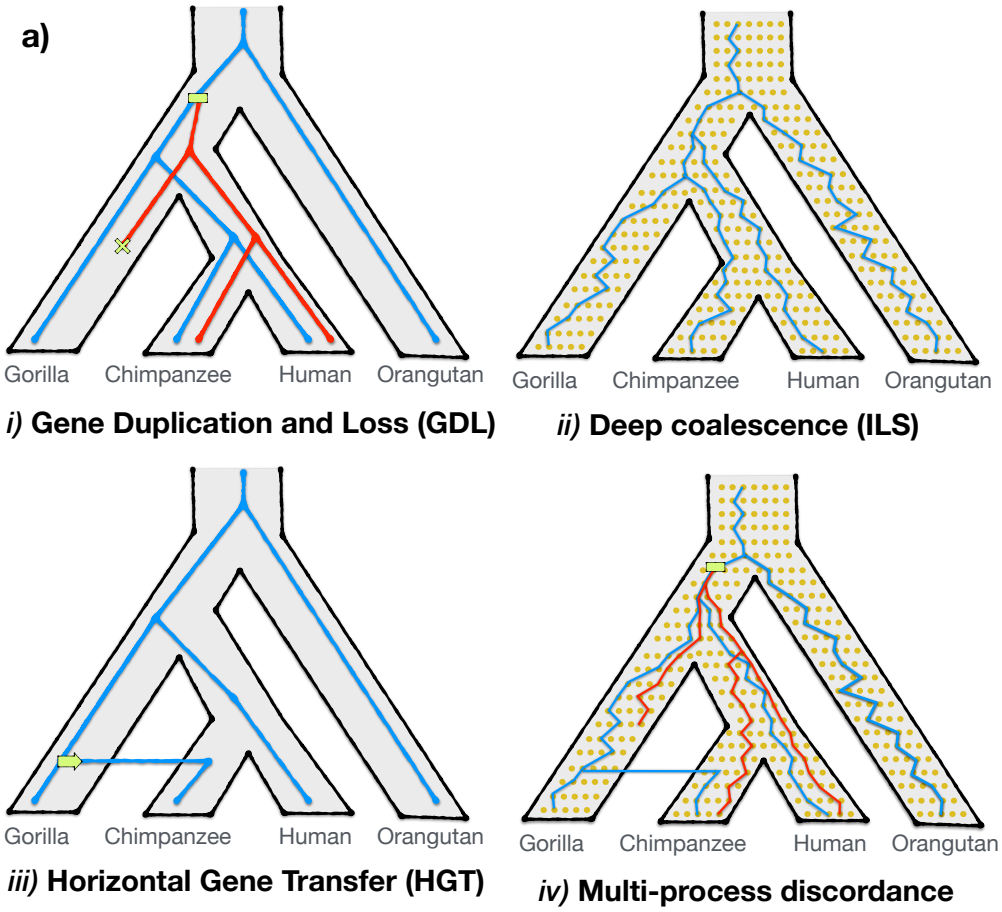
Statistically consistent methods (??)

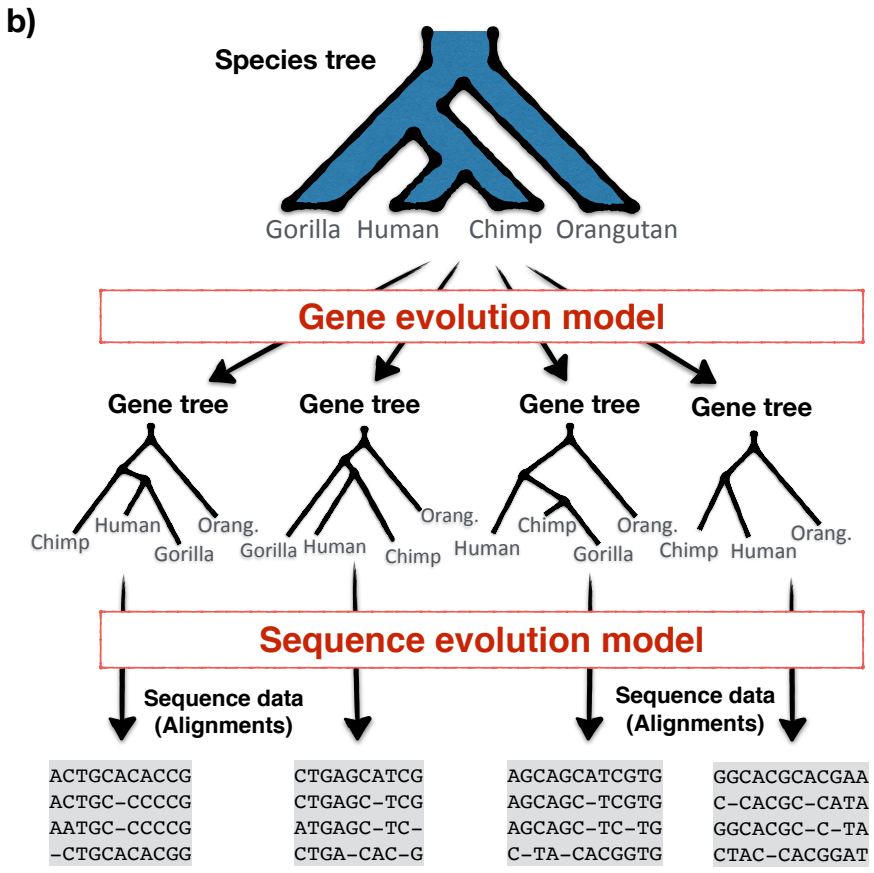
- **Coalescent-based summary methods:** Estimate gene trees, and then combine together (ASTRAL, ASTRID, MP-EST, NJst, and others)

- **Co-estimation methods**: Co-estimate gene trees and species trees (TOO EXPENSIVE)

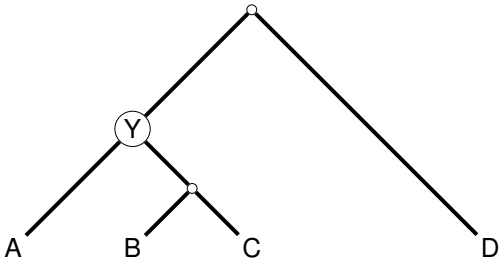
- **Site-based methods:** estimate the species tree from the concatenated alignment, and do not estimate gene trees (NOT WELL STUDIED)

Genome-scale discordance: very complex

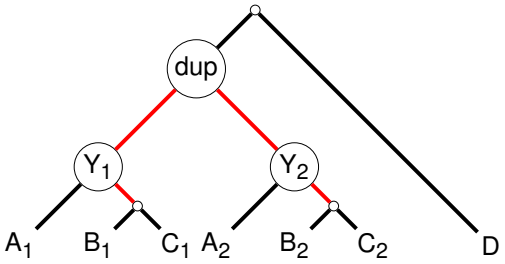




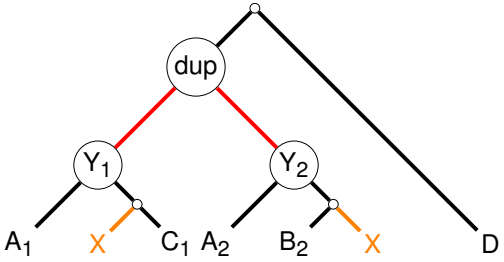
Problem: Given set of MUL-trees, infer the species tree



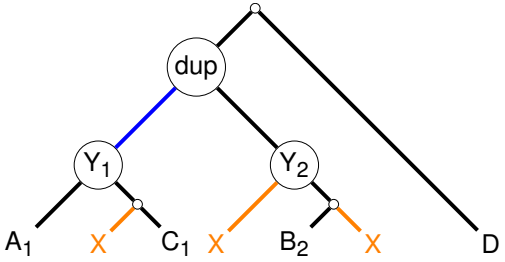
(a) Species tree T^*



(b) Gene tree M_1 with one duplication.



(c) Gene tree M_2 with one duplication and two losses.



(d) Gene tree with one duplication and three losses.

Note: no orthology detection

New Results (Legried et al.)

- Under the GDL model, the unrooted species tree topology is identifiable from the distribution of unrooted gene family trees (MUL-trees)
- ASTRAL-multi (i.e., ASTRAL designed for multi-individual gene trees) and ASTRAL-One are statistically consistent under GDL models

Legried, Molloy, Warnow, and Roch. RECOMB and J. Computational Biology 2020.

Empirical performance vs Theory

- Unknown if distance-based species tree estimation (e.g., ASTRID-multi) is statistically consistent under GDL models
- ASTRAL-multi is statistically consistent, but other methods that are not established to be statistically consistent have better accuracy

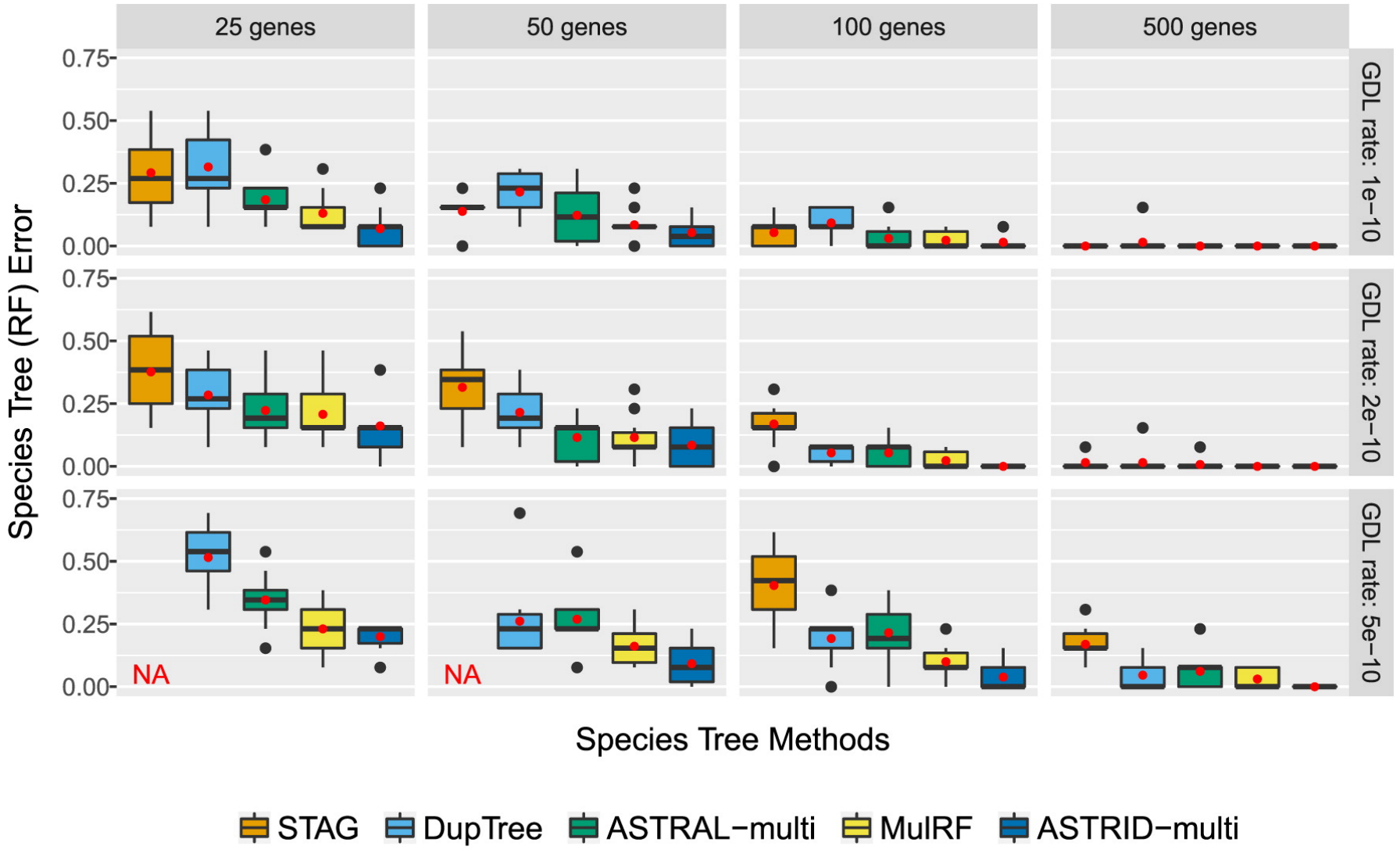
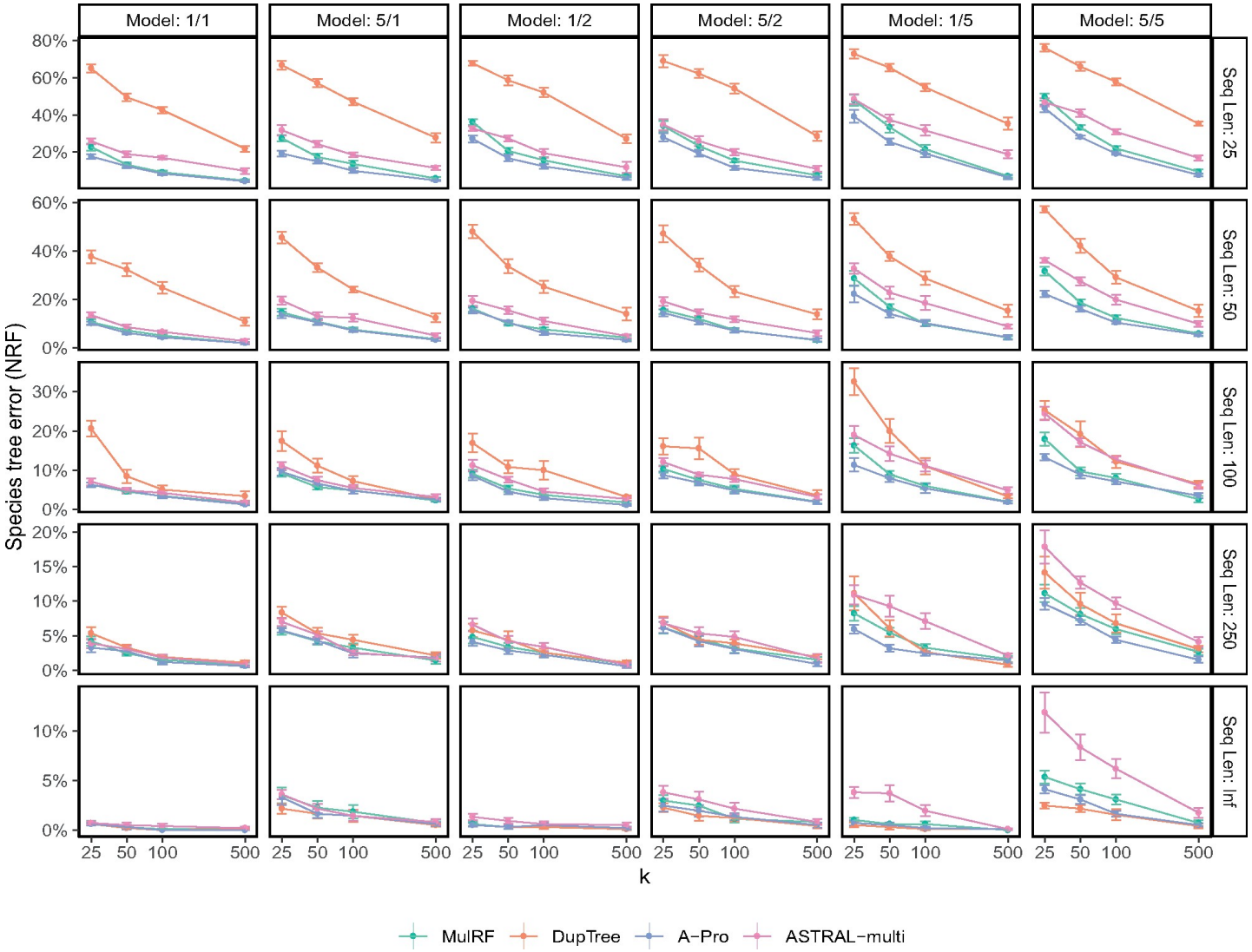


Figure 2 from Legried et al. Results are for trees estimated on estimated gene trees, 16-taxon fungal simulated datasets with varying numbers of genes.

ASTRAL-Pro

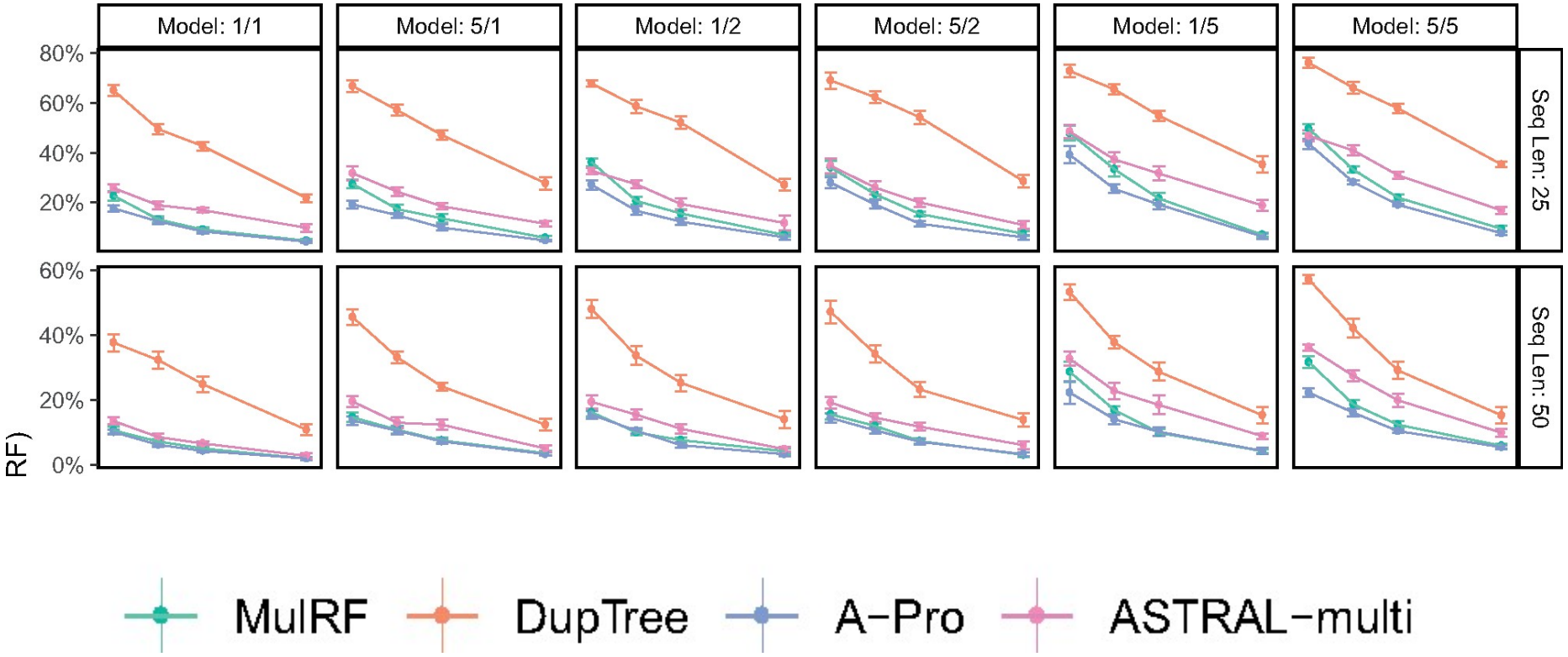
- New variant of ASTRAL for MUL-trees
- Algorithm:
  - ”Root and tag” the gene trees
  - Weight quartet trees under the assumption of correct rooting and tagging
  - Solve the same optimization problem (MQSST)

ASTRAL-Pro compared to other methods on 100 species



Mol Biol Evol, Volume 37, Issue 11, November 2020, Pages 3292–3307, https://doi.org/10.1093/molbev/msaa139

ASTRAL-Pro compared to other methods on 100 species (varying number genes)

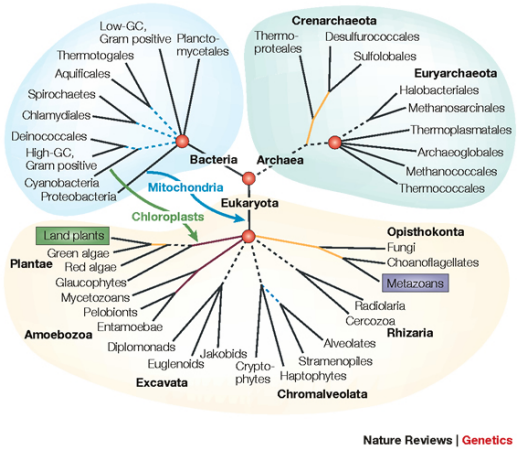


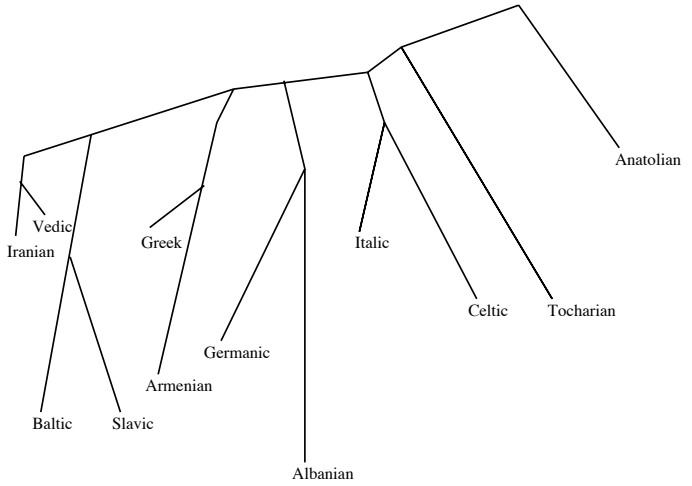
Mol Biol Evol, Volume 37, Issue 11, November 2020, Pages 3292–3307, https://doi.org/10.1093/molbev/msaa139

(Some) Open Questions

- Which species tree estimation methods are statistically consistent under GDL and DLCOAL models? (Gene tree parsimony? ASTRID-multi?)
- Is ASTRAL-Pro statistically consistent for GDL under some random model of error for rooting and tagging?
- What is the sample complexity for ASTRAL (in its variants and under different models)?
- What is the impact of gene tree estimation error, finite sequence length, and missing data on species tree estimation
- Why is concatenation using maximum likelihood so accurate, even under the ”anomaly zone”
- What are the pros and cons of marker selection practices

Phylogenetic Inference





- NP-hard optimization problems and large datasets
- Statistical estimation under stochastic models of evolution
- Probabilistic analysis of algorithms
- Graph-theoretic divide-and-conquer
- Chordal graph theory
- Combinatorial optimization

Acknowledgments



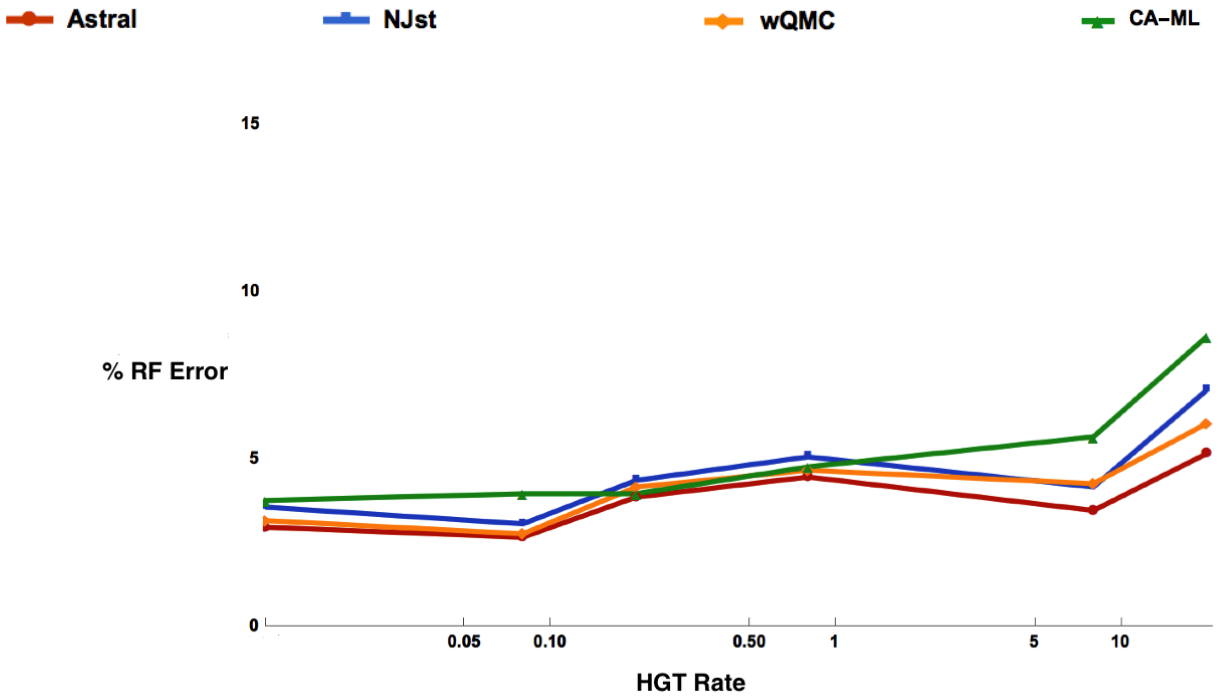


Roch and Warnow, Systematic Biology 2014
Mirarab and Warnow, Bioinformatics 2015
Molloy and Warnow, Systematic Biology 2017
Legried et al., J. Computational Biology 2020
Papers available at http://tandy.cs.illinois.edu/papers.html

Funding: NSF, Grainger Foundation, and HHMI (to SM).

Accuracy in the presence of HGT + ILS

200 Estimated Gene Trees



Data: Fixed, moderate ILS rate, 50 replicates per HGT rates (1)-(6), 1 model species tree per replicate on 51 taxa, 1000 true gene trees, simulated 1000 bp gene sequences using INDELible [8], 1000 gene trees estimated from GTR simulated sequences using FastTree-2[7]

[7] Price, Dehal, Arkin 2015

[8] Fletcher, Yang 2009

Davidson et al., RECOMB-CG, BMC Genomics 2015