Theoretical and Empirical Advances in Species Tree Estimation

Tandy Warnow

The University of Illinois

This talk

- Originally focused on species tree estimation, taking gene tree heterogeneity into account
- Expanded to address heterogeneity across the tree (e.g., heterotachy), which leads to model misspecification

Phyogenomic Pipeline

- Assemble and annotate genomes (e.g., determine orthologs)
- Compute multiple sequence alignments of individual loci
- Construct gene trees
- Construct species tree
- Perform post-tree analyses (e.g., estimate dates, infer selection, etc.)

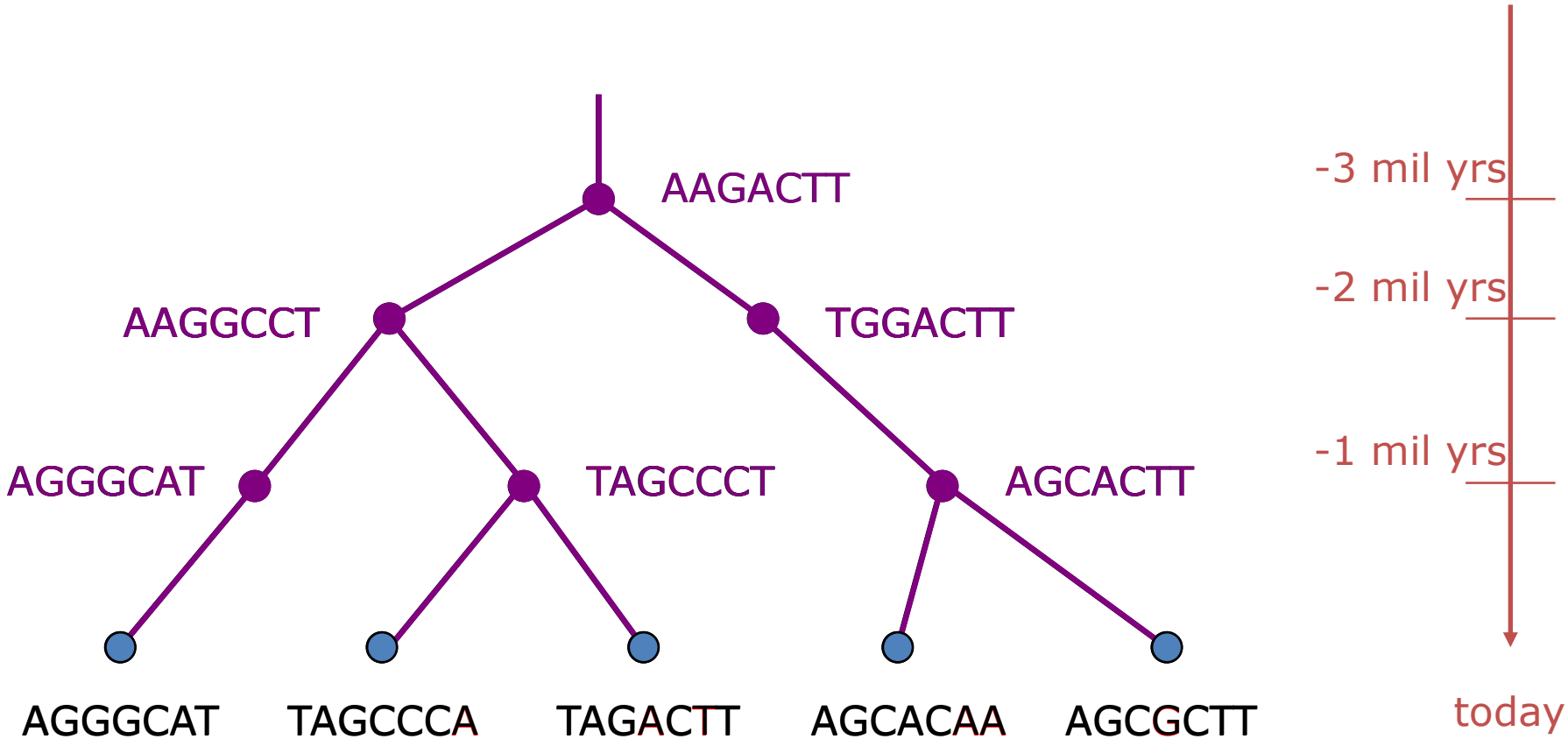
Phyogenomic Pipeline

- Assemble and annotate genomes (e.g., determine orthologs)
- Compute multiple sequence alignments of individual loci
- Construct gene trees
- Construct species tree
- Perform post-tree analyses (e.g., estimate dates, infer selection, etc.)

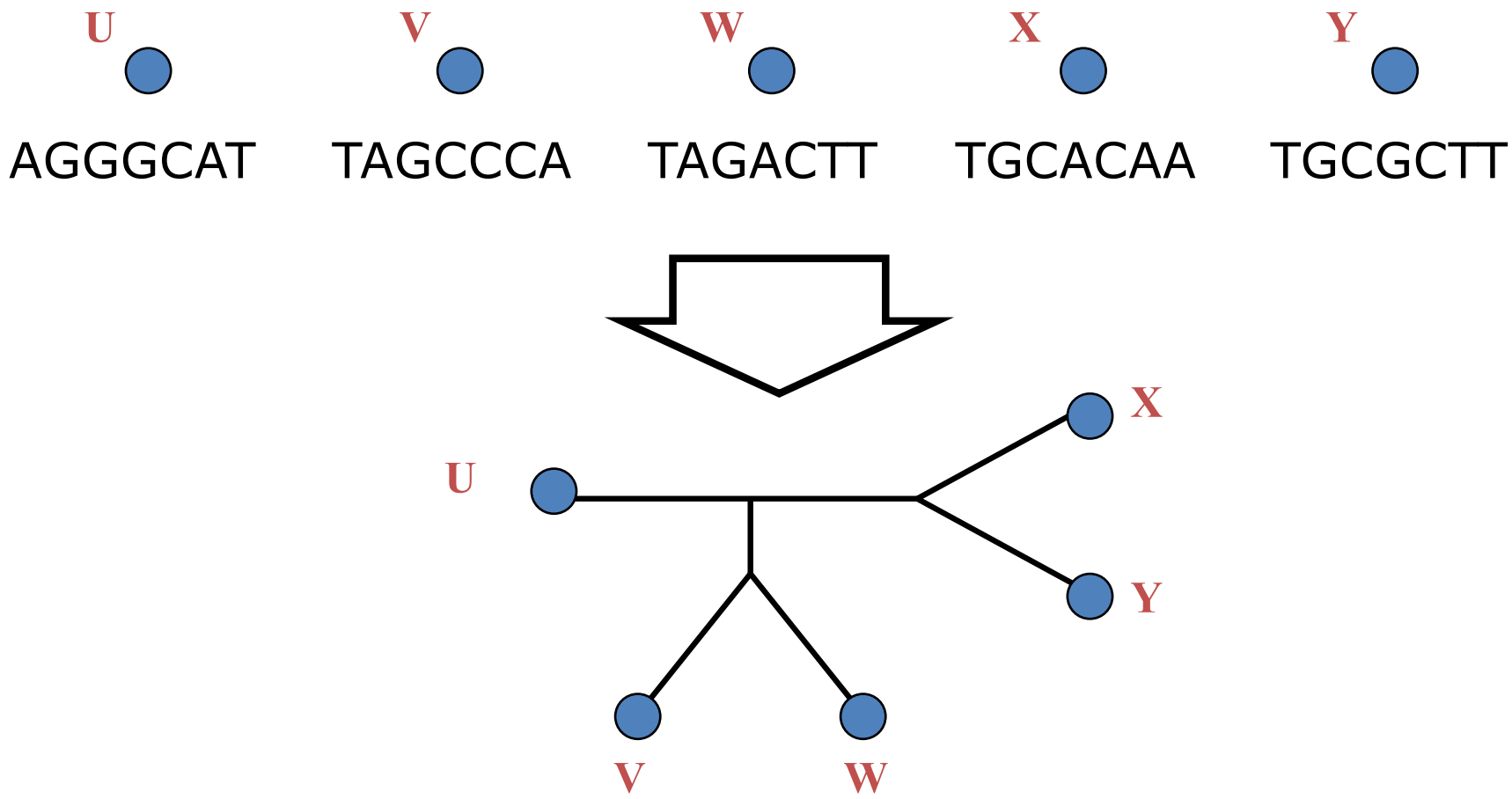
Phyogenomic Pipeline

- Assemble and annotate genomes (e.g., determine orthologs)
- Compute multiple sequence alignments of individual loci
- Construct gene trees
- Construct species tree
- Perform post-tree analyses (e.g., estimate dates, infer selection, etc.)

DNA Sequence Evolution (Idealized)



Phylogeny Problem



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):

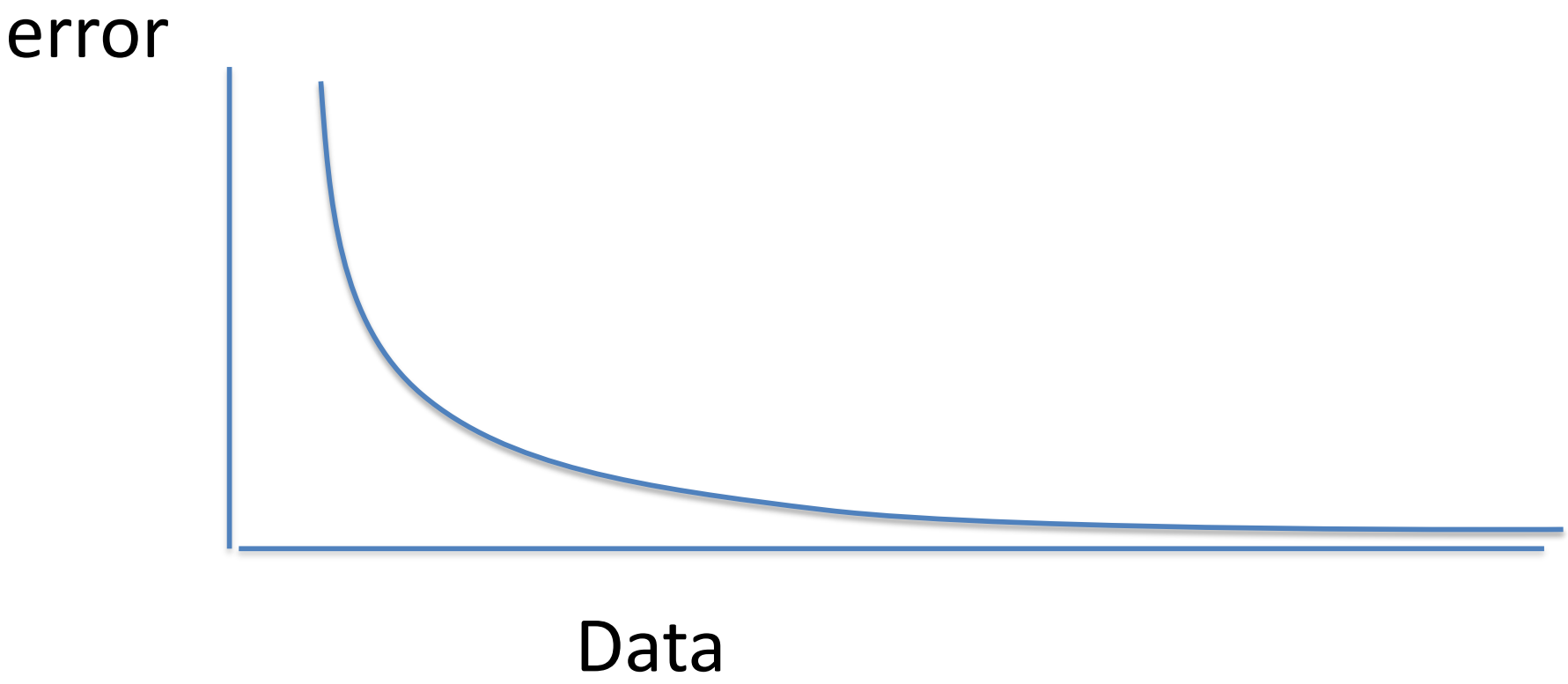
- The model tree T is binary and has substitution probabilities p(e) on each edge e, with $0<p(e)<3/4$.
- The state at the root is randomly drawn from {A,C,T,G} (nucleotides)
- If a site (position) changes on an edge, it changes with equal probability to each of the remaining states.
- The evolutionary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

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

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)?
- What are the computational issues?

Statistical Consistency/Identifiability



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 sequence length requirements for standard methods.
- The best methods (typically maximum likelihood or Bayesian estimation) are very computationally intensive.

Computational issues

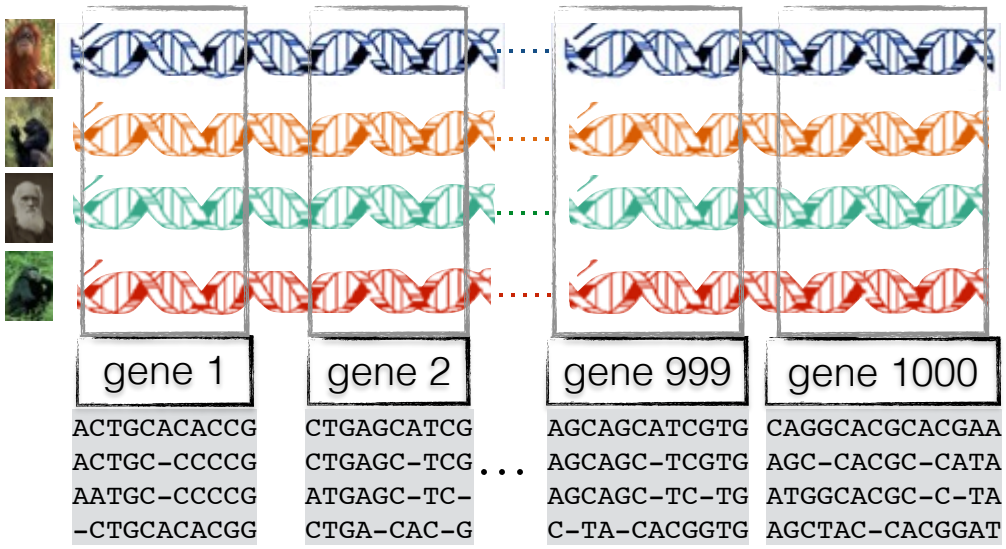
- Maximum likelihood, maximum parsimony, and even some distance-based methods: NP-hard, and tree-space grows exponentially with the number of leaves
- Bayesian estimation: need to run to convergence (may fail)
- Parallelism helps but is not enough

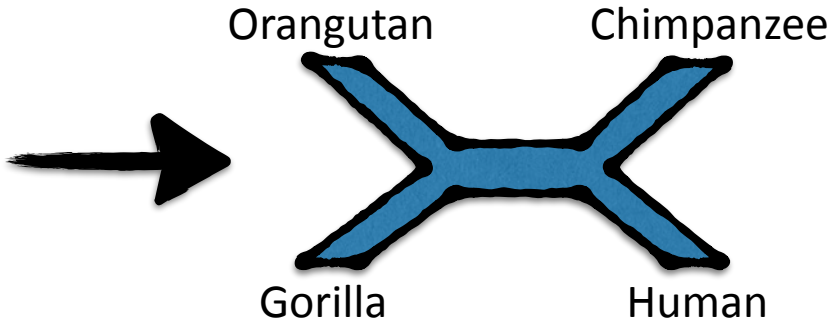
Take home message: large datasets are beyond the capability of current methods (even with supercomputers)

Phyogenomic Pipeline

- Assemble and annotate genomes (e.g., determine orthologs)
- Compute multiple sequence alignments of individual loci
- Construct gene trees
- Construct species tree
- Perform post-tree analyses (e.g., estimate dates, infer selection, etc.)

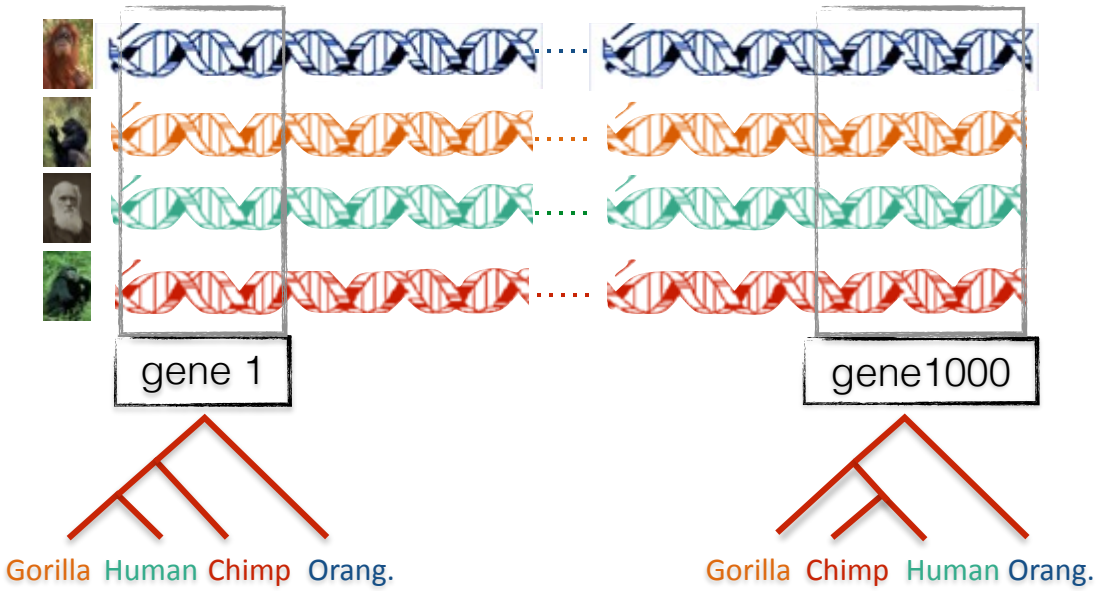
phylogenomics





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

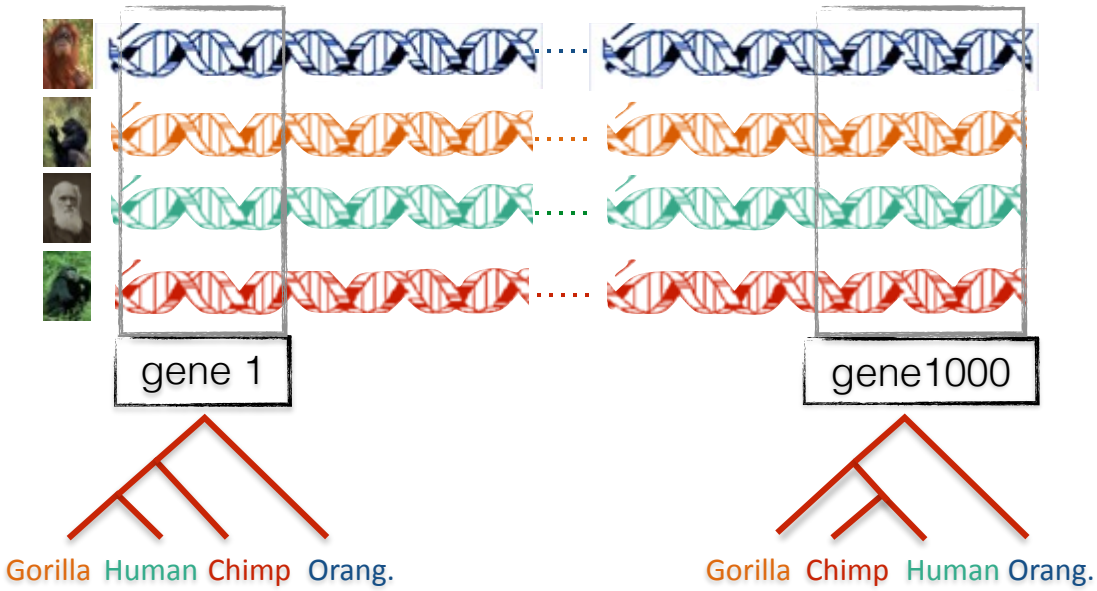
Gene tree discordance



Multiple causes for discord, including

- Incomplete Lineage Sorting (ILS),
- Gene Duplication and Loss (GDL), and
- Horizontal Gene Transfer (HGT)

Gene tree discordance

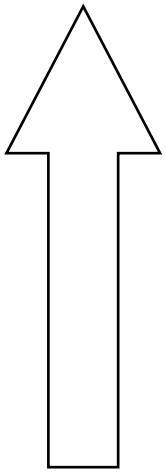


Multiple causes for discord, including

- Incomplete Lineage Sorting (ILS),
- Gene Duplication and Loss (GDL), and
- Horizontal Gene Transfer (HGT)

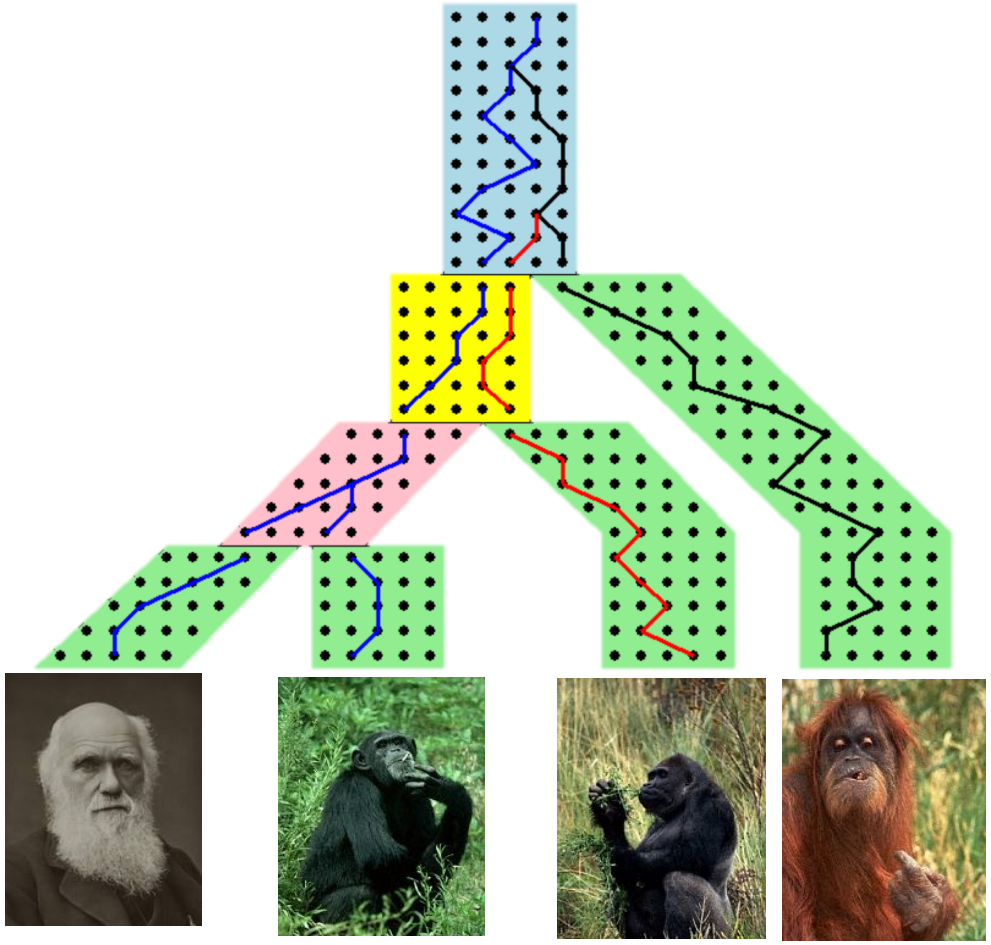
Gene trees inside the species tree (Coalescent Process)

Past

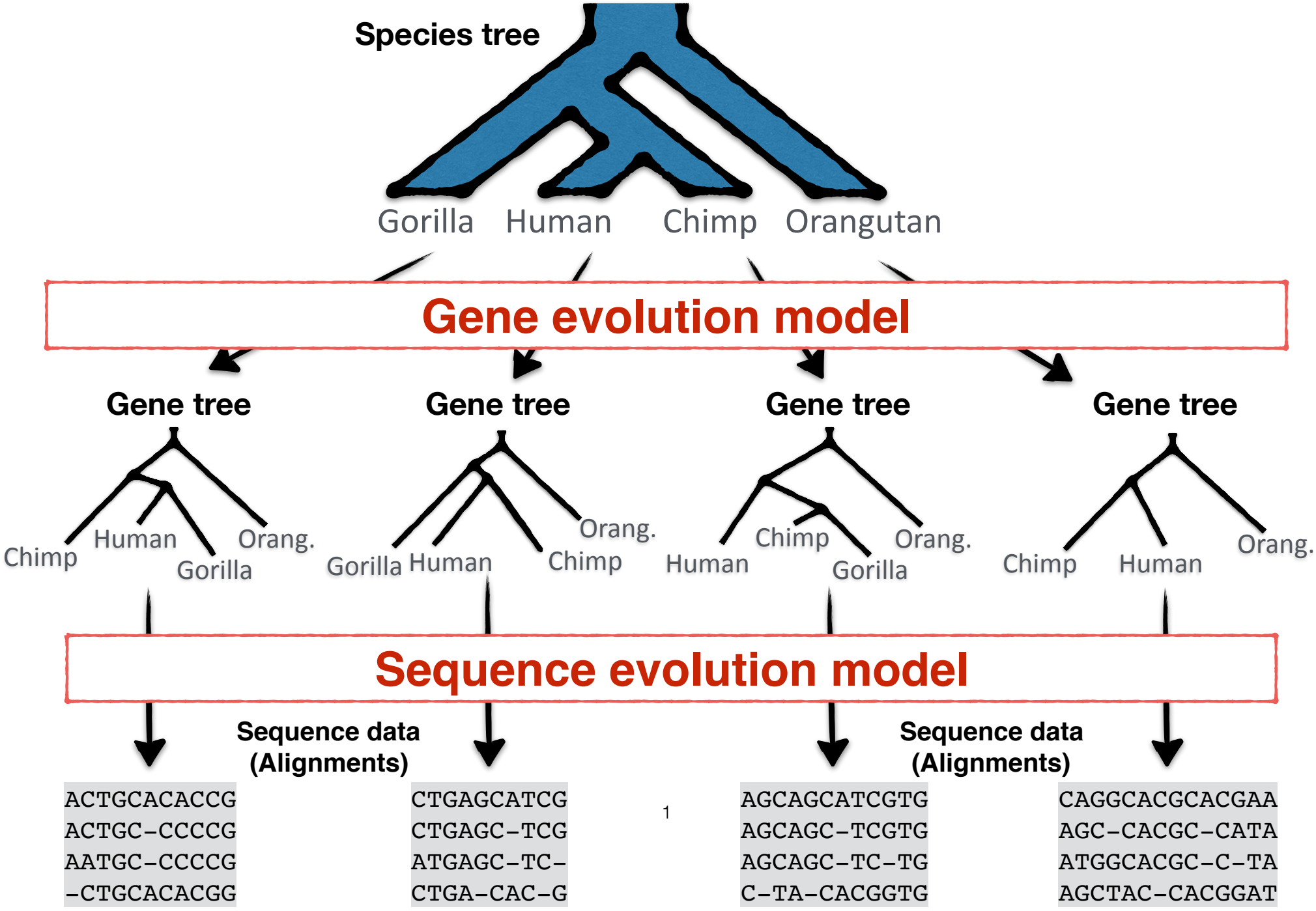


Present

Courtesy James Degnan



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


Species tree
Gorilla
Human
Chimp
Orangutan
Gene evolution model
Gene tree
Gene tree
Gene tree
Gene tree
Chimp
Human
Orang.
Gorilla
Gorilla
Human
Orang.
Chimp
Human
Chimp
Orang.
Gorilla
Chimp
Human
Orang.
Sequence evolution model
Sequence data
(Alignments)
Sequence data
(Alignments)
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT

Big picture challenge

- Multi-locus data, generated by a hierarchical model
  - Species tree generates gene trees
  - Gene trees generate sequences
- How can we estimate the species tree from the sequence data?

Big picture challenge

- Multi-locus data, generated by a hierarchical model
  - Species tree generates gene trees
  - Gene trees generate sequences
- How can we estimate the species tree from the sequence data?
- Suppose the number of genes and the sequence data per gene both go to infinity?

Four Basic Approaches

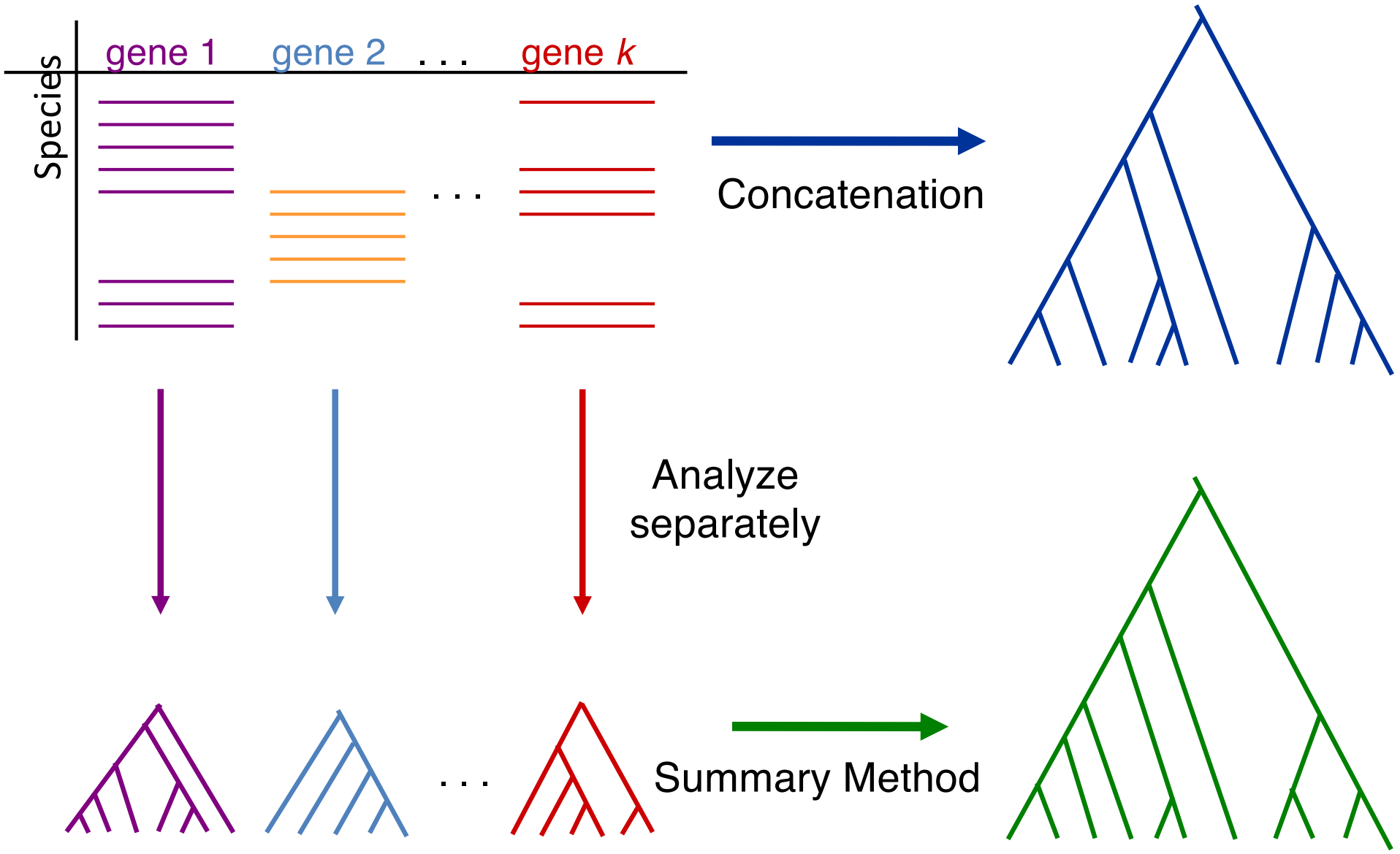
Statistically consistent methods:

- Co-estimate species tree and gene trees: e.g., *BEAST (Heled and Drummond)
- Site-based methods: e.g., SVDquartets (Chifman and Kubatko, implemented in PAUP*)
- Methods that combine gene trees (summary methods): e.g., NJst, MP-EST, ASTRAL, ASTRID, STEM, etc.

And of course

- Concatenation, but this isn't statistically consistent in the presence of ILS

Main competing approaches



Our Methods (all open source)

- BBCA: Improving scalability of *BEAST (Zimmermann, Mirarab, and Warnow 2015) to large numbers of loci
- SVDquest: Improving accuracy and scalability for SVDquartets (Vachaspati and Warnow, 2018)
- ASTRID (Vachaspati and Warnow 2015) and ASTRAL (Mirarab et al. 2014, 2015, etc.): summary methods that can analyze datasets with thousands of species and loci with high accuracy
- NJMerge: Improving scalability of species tree estimation methods to large numbers of species (Molloy and Warnow, 2018)
- TreeMerge: improvement on NJMerge (Molloy and Warnow, 2019)

All are statistically consistent under the MSC+GTR model

ASTRAL

- Mirarab and Warnow, Bioinformatics 2014
- https://github.com/smirarab/ASTRAL

Algorithmic approach:

- Given set of gene trees, find the species tree that agrees with the maximum number of quartet trees within a constrained search space.
- Polynomial time and statistically consistent in the presence of ILS.

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
- Prum et al, **198** avian species, 259 genes

X

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

eLIFE elifesciences.org

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

ELSEVIER

Molecular Phylogenetics and Evolution

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

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. Wend

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]

LETTER

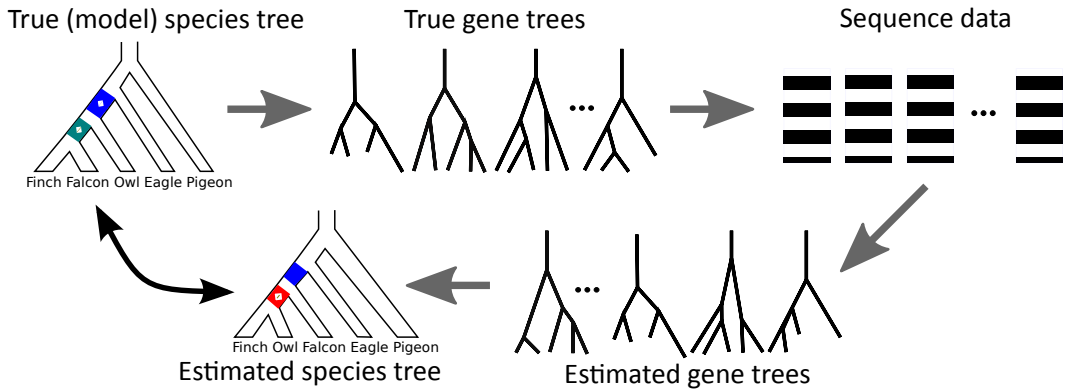
doi:10.1038/nature15697

A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing

Richard O. Prum[1,2]*, Jacob S. Berv[2]*, Alex Dornburg[1,2,4], Daniel J. Field[2,5], Jeffrey P. Townsend[3,6], Emily Moriarty Lemmon[7] & Alan R. Lemmon[8]

Simulation study

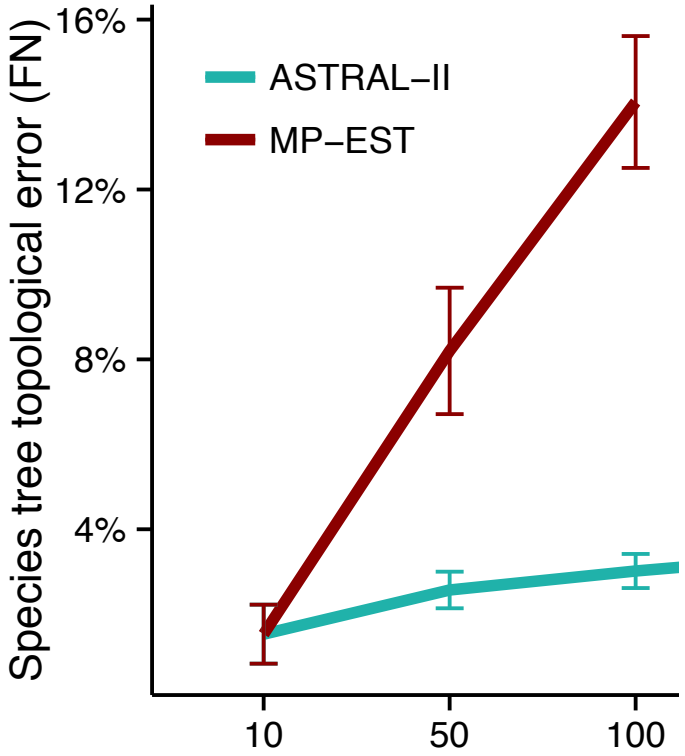
- Variable parameters:
  - Number of species: 10 – 1000
  - Number of genes: 50 – 1000
  - Amount of ILS: low, medium, high
  - Deep versus recent speciation



- 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

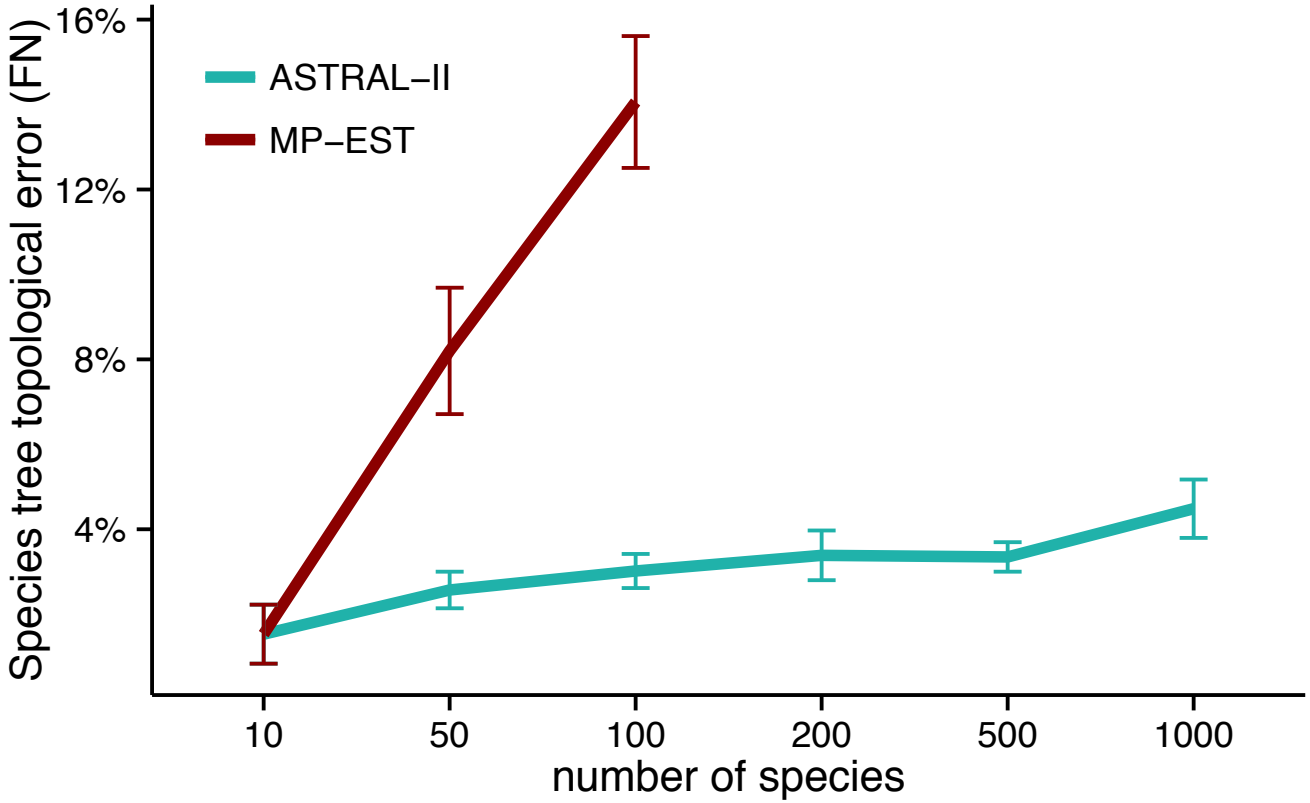
Used SimPhy, Mallo and Posada, 2015

Tree accuracy when varying the number of species



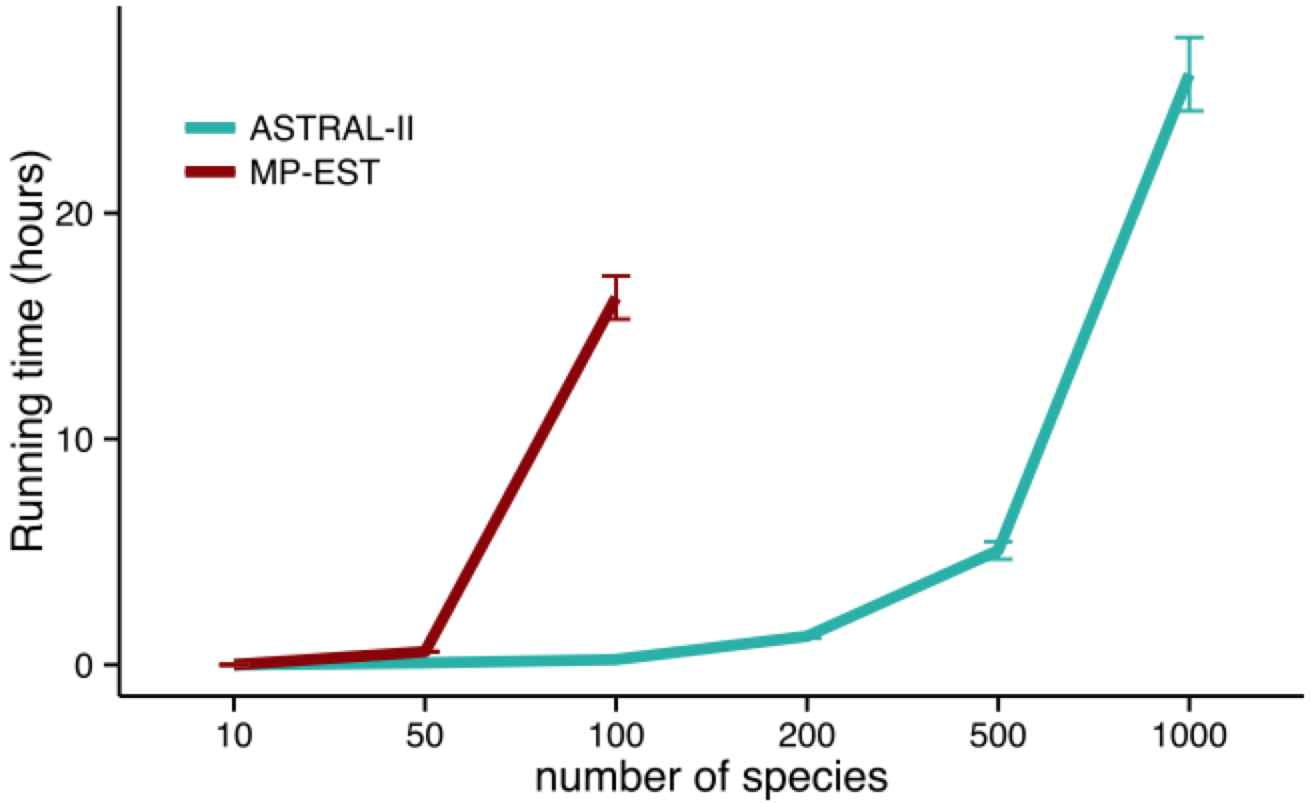
1000 genes, “medium” levels of recent ILS

Tree accuracy when varying the number of species



1000 genes, “medium” levels of recent ILS

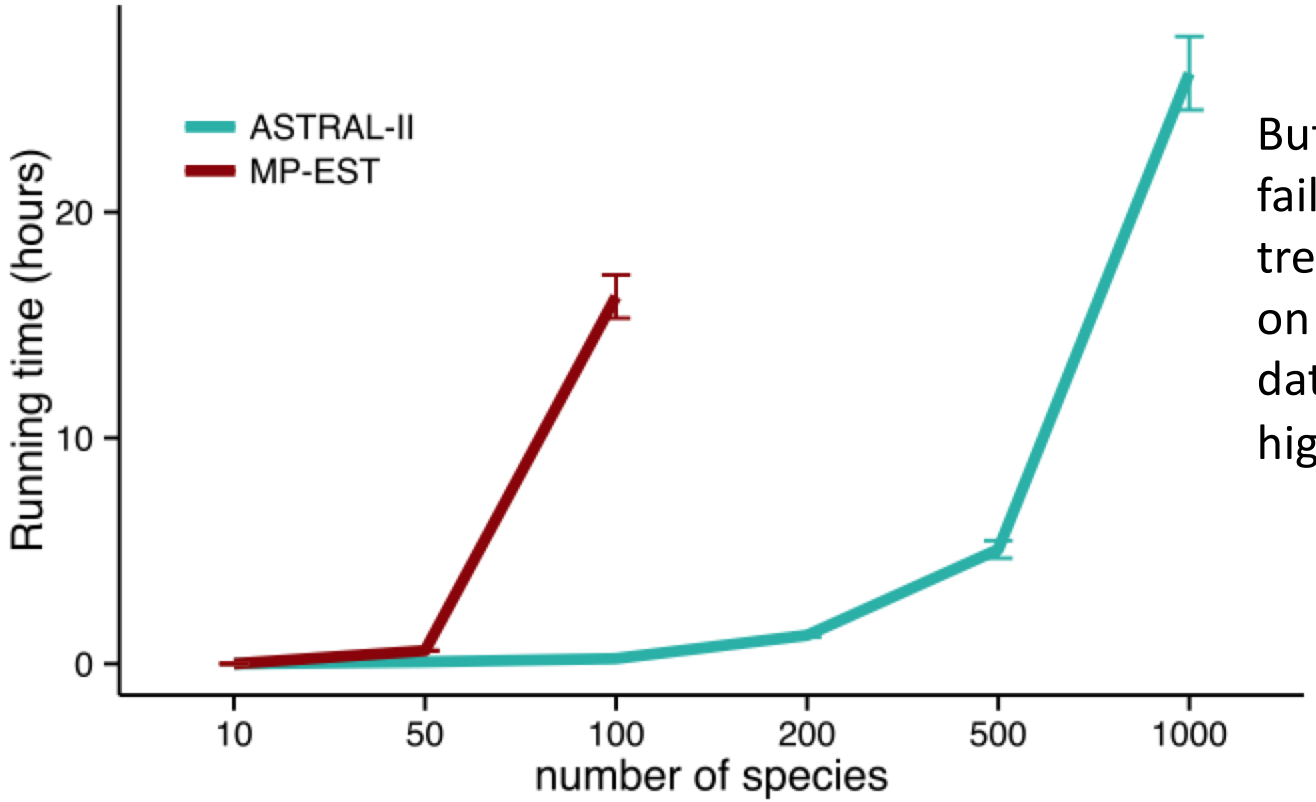
Running time as function of # species



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

X

Running time as function of # species



But ASTRAL can fail to return a tree within 24 hrs on some very large datasets with high ILS

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

X

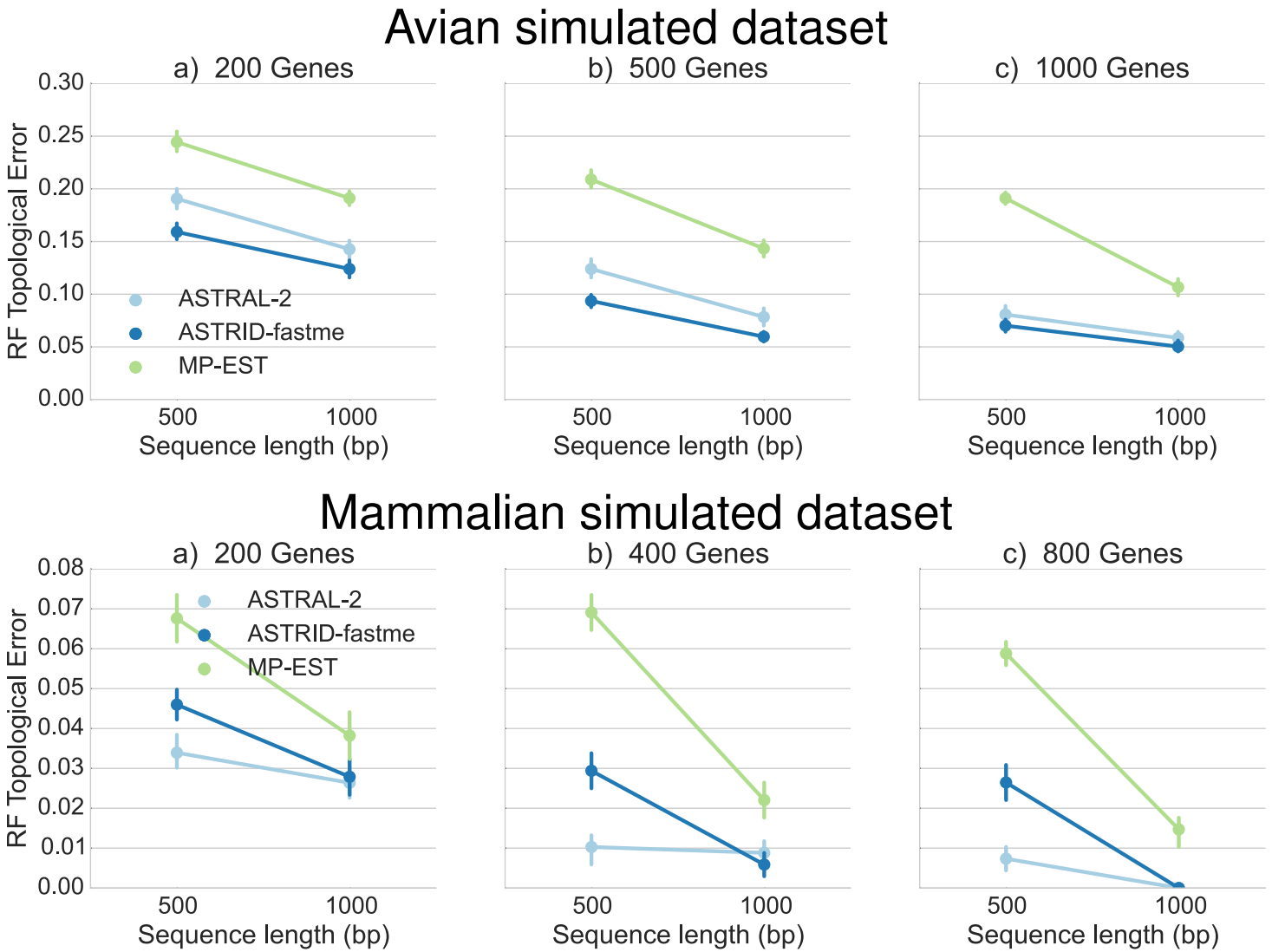
ASTRID

- ASTRID: Accurate species trees using internode distances, Vachaspati and Warnow, RECOMB-CG 2015 and BMC Genomics 2015
- Github site: https://github.com/pranjalv123/ASTRID

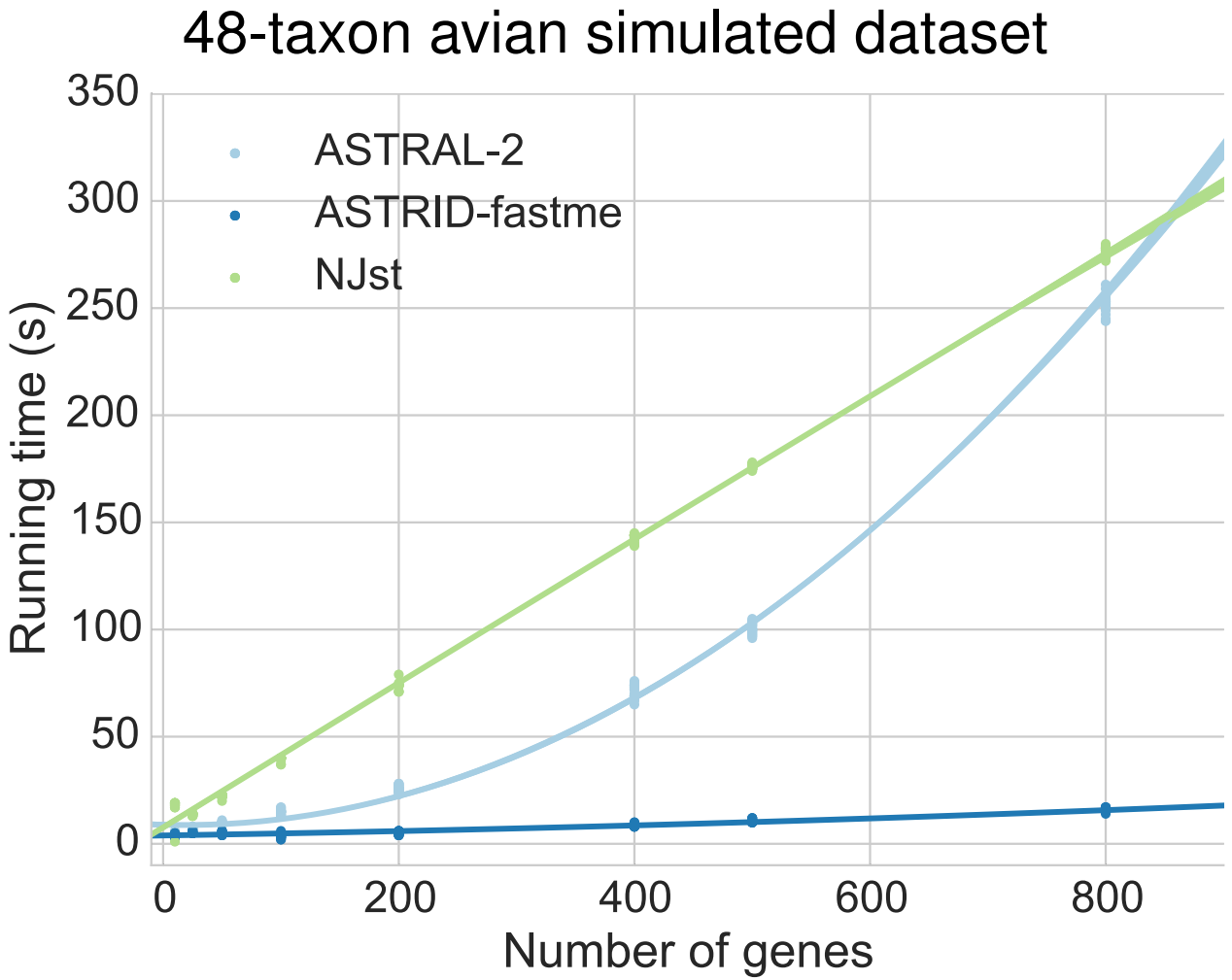
Algorithmic design:

- Nearly the same as NJst (Liu and Yu, 2010)- computes a matrix of average leaf-to-leaf topological distances, and then computes a tree using FastME (more accurate than neighbor Joining and faster, too).
- Polynomial time and statistically consistent in the presence of ILS.

Both ASTRAL and ASTRID substantially outperform MP-EST



ASTRID is very fast

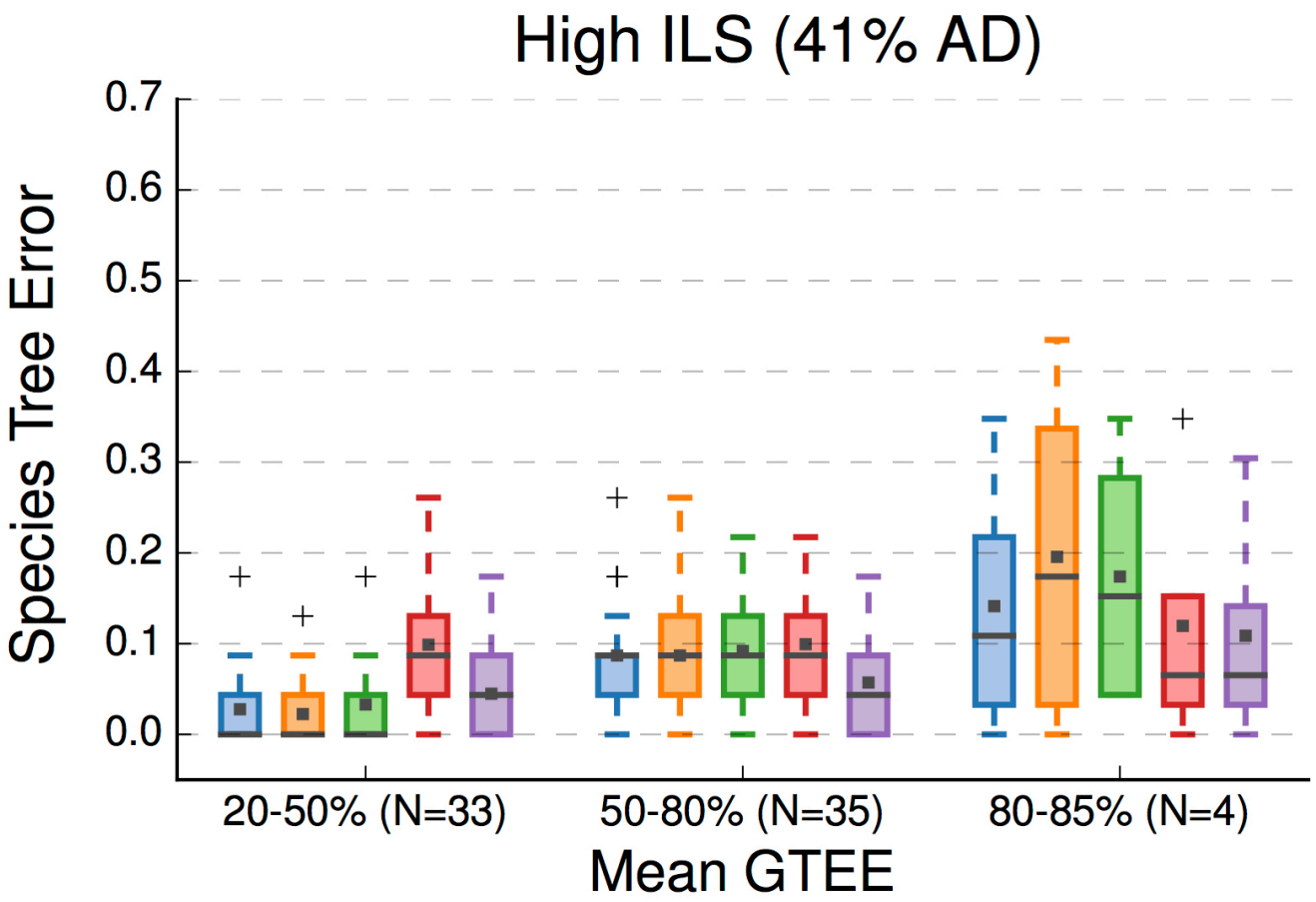


- On the ASTRAL-2 dataset with 1000 taxa, 1000 genes, ASTRID-FastME takes 33 minutes, ASTRAL takes 12 hours.

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!

ASTRAL ASTRID MP-EST SVDquartets CA-ML

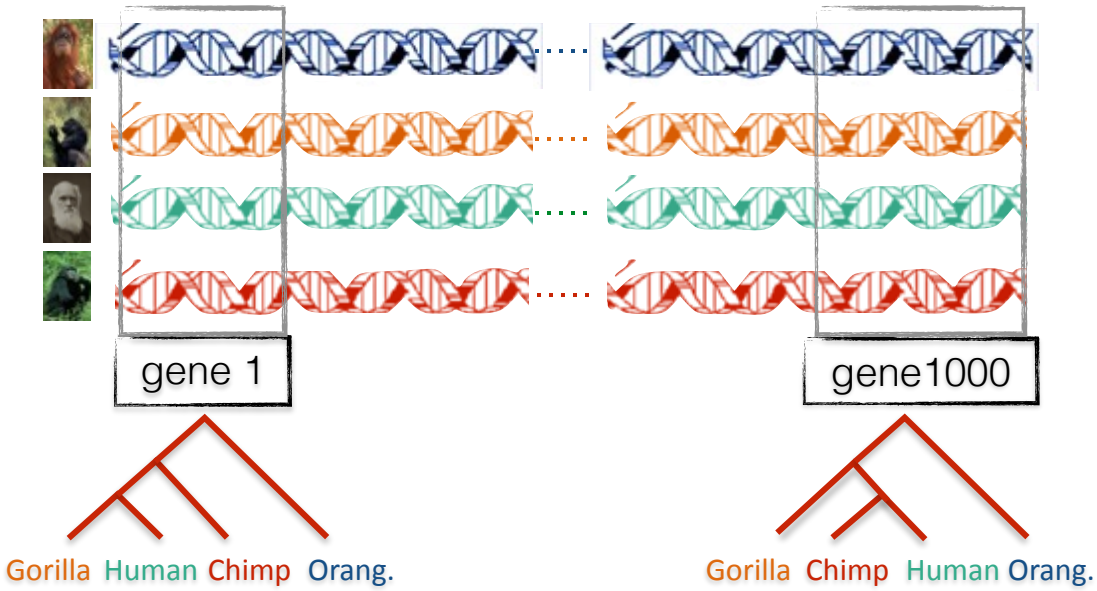
Summary Methods

Site-based Method

Should you filter?

- Filtering genes based on missing data?
  - Generally not beneficial (see Molloy and Warnow, Systematic Biology 2018)
- Filtering genes based on gene tree estimation error?
  - Depends on conditions (see Molloy and Warnow, Systematic Biology 2018)
- Filtering fragmentary sequences from genes while keeping the gene?
  - Often beneficial (see Sayyari, Whitfield, and Mirarab, MBE 2018)

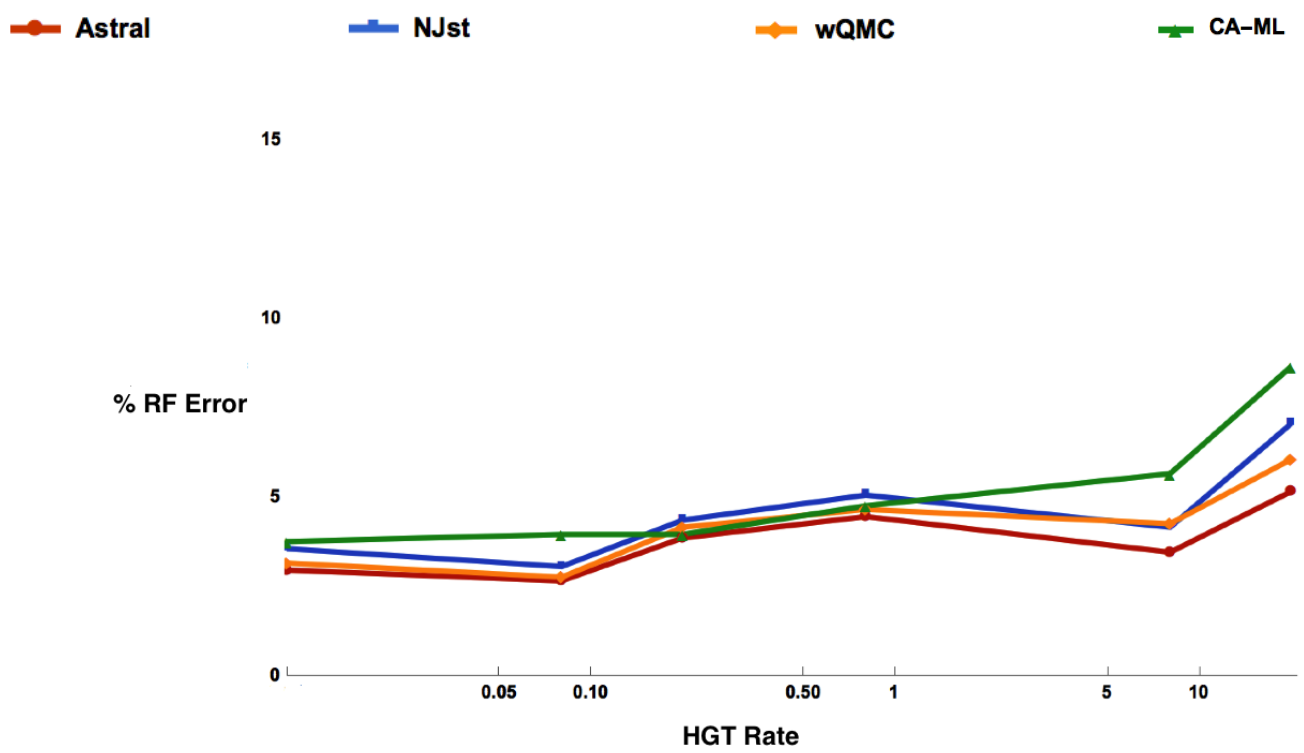
Gene tree discordance



Multiple causes for discord, including

- Incomplete Lineage Sorting (ILS),
- Gene Duplication and Loss (GDL), and
- Horizontal Gene Transfer (HGT)

Accuracy in the presence of HGT + ILS



ASTRAL and wQMC both good (a bit more robust to HGT than NJst and CA-ML)

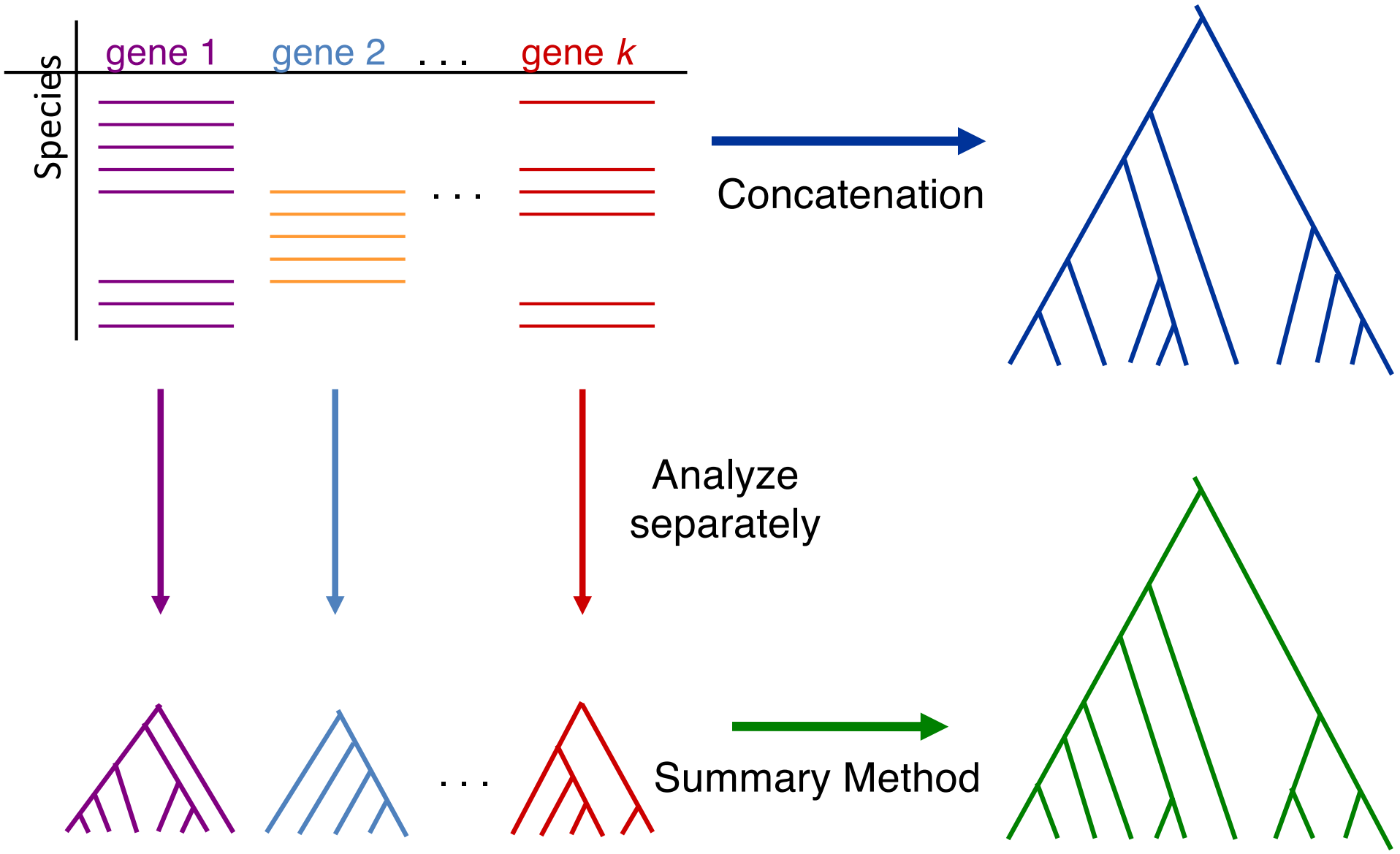
Data: Fixed, moderate ILS rate, 50 replicates per HGT rates (1)-(6), 1 model species tree per replicate on 51 taxa, 1000 true gene trees, simulated 1000 bp gene sequences using INDELible [8], 1000 gene trees estimated from GTR simulated sequences using FastTree-2[7]

[7] Price, Dehal, Arkin 2015

[8] Fletcher, Yang 2009

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

Main competing approaches



Scaling to large numbers of species

- Concatenation analyses: very expensive (e.g., 250 CPU years for the Avian Phylogenomics project with only 48 species)

- Summary methods: ASTRAL scales to large numbers of species but can fail on some large datasets, and running time increases with heterogeneity; ASTRID seems to be fine.

NJMerge

- Molloy and Warnow, RECOMB-CG 2018
- Github site: https://github.com/ekmolloy/njmerge

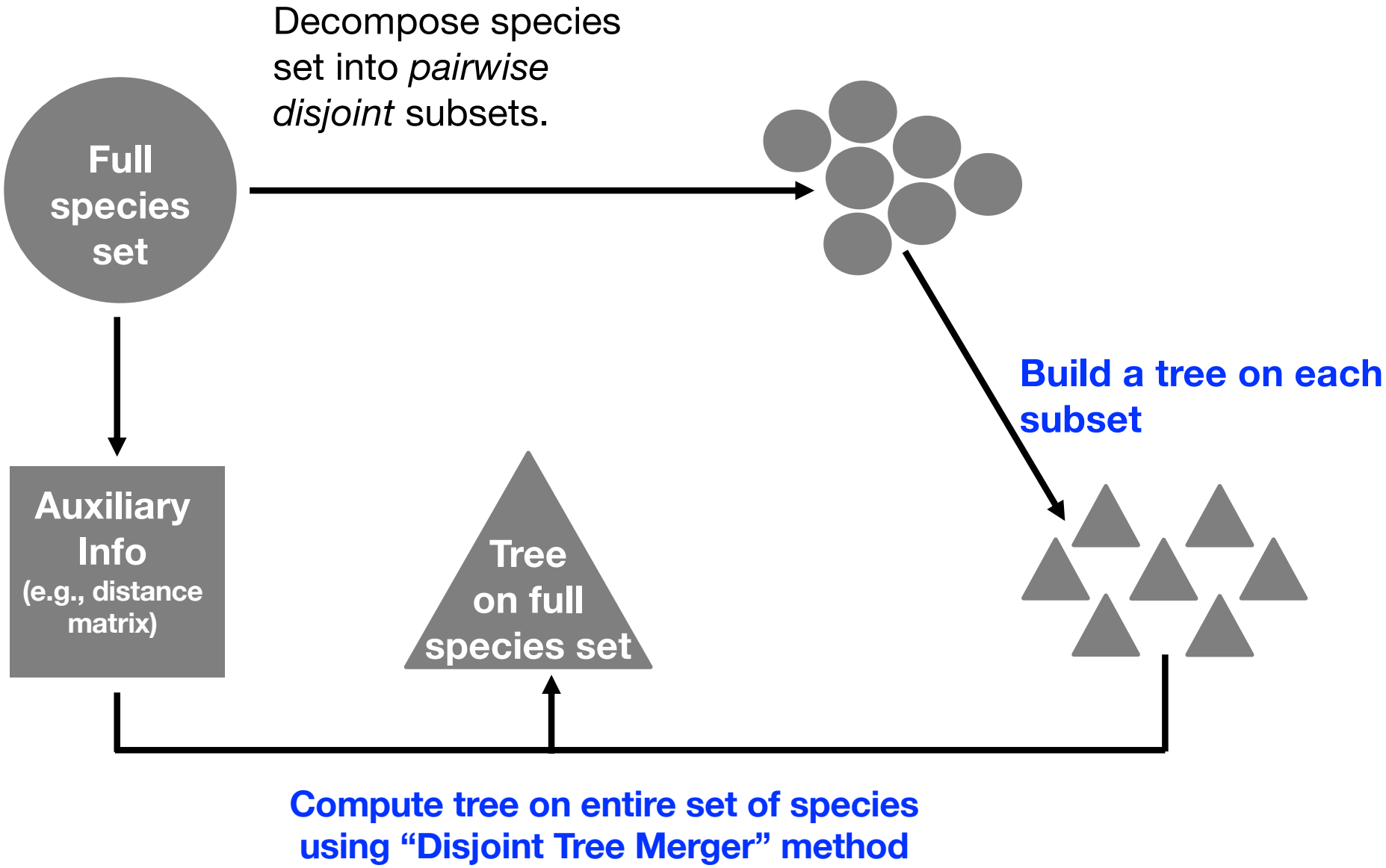
Algorithmic strategy:

- Divide-and-conquer: divides species set into disjoint subsets, computes species trees on the subsets using selected species tree method (e.g., ASTRAL, RAxML, SVDquartets), and then merges subset trees using a distance-based method.

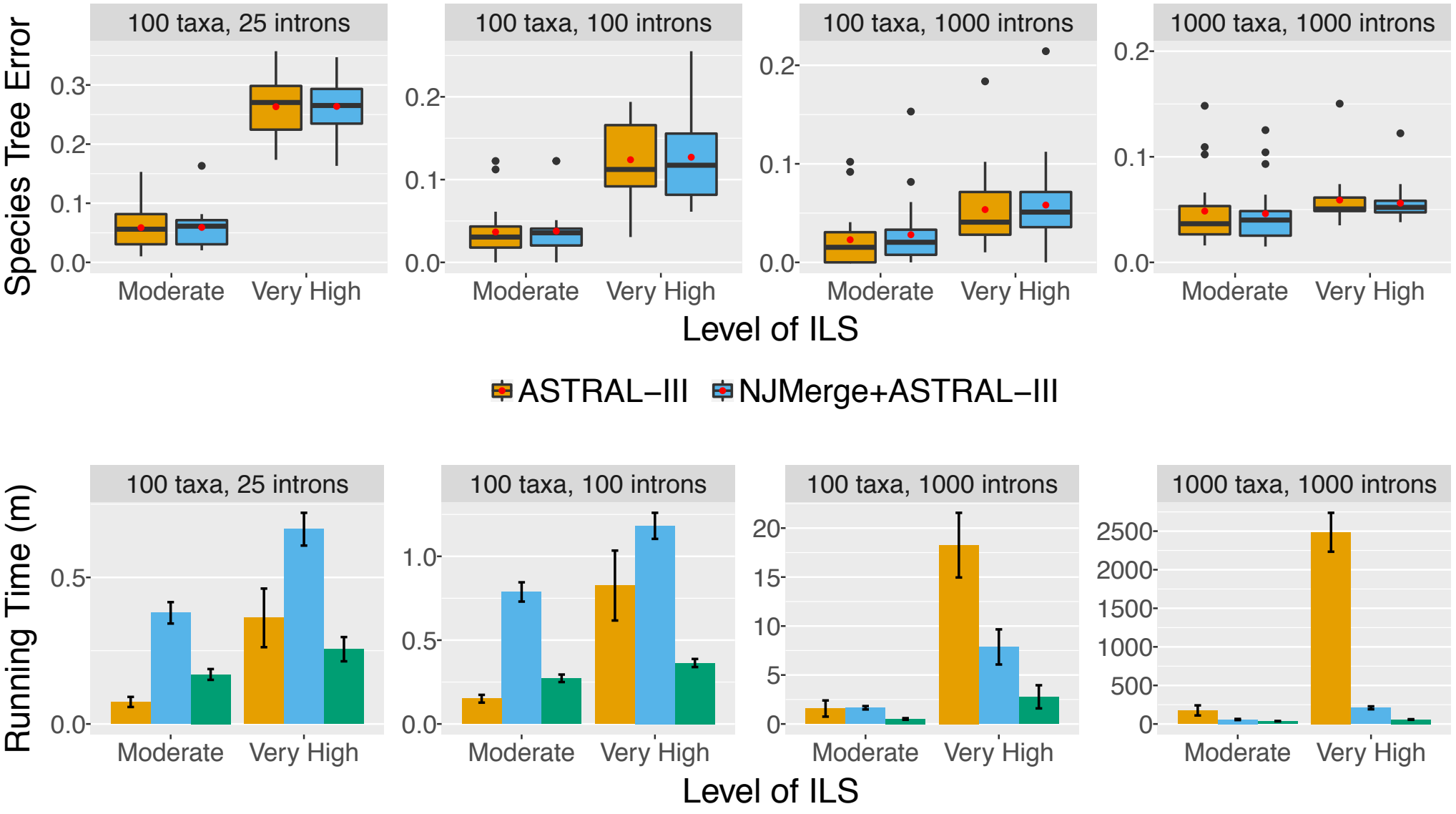
TreeMerge

- Molloy and Warnow, to appear, ISMB 2019
- Like NJMerge, it is statistically consistent under the MSC when used with ASTRAL or other statistically consistent methods
- Improves on NJMerge:
  - guaranteed to never fail
  - Asymptotically faster -- $O(n^2)$ in divide-and-conquer pipeline
- On github

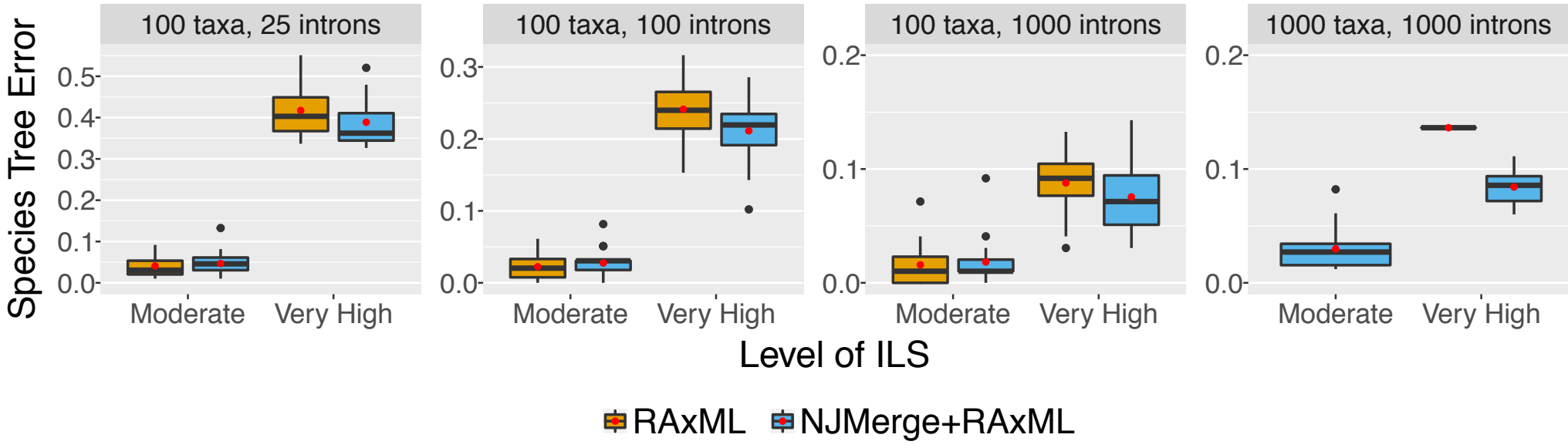
Divide-and-Conquer Pipeline

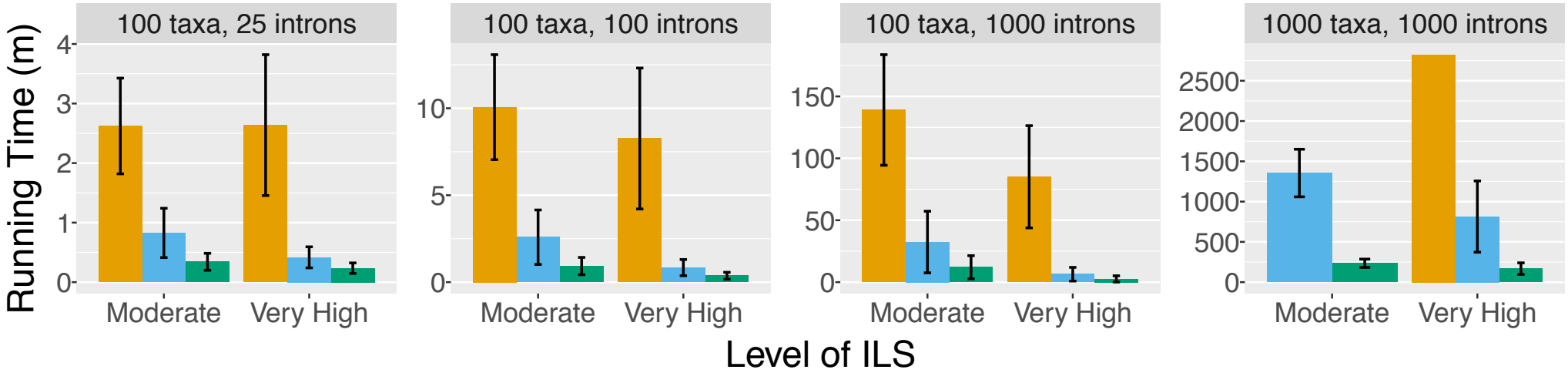


NJMerge + ASTRAL vs. ASTRAL:
Comparable accuracy and can analyze larger datasets



NJMerge + RAxML vs. RAxML: Better accuracy and faster!





NJMerge and TreeMerge

- Using NJMerge or TreeMerge with ASTRAL: generally as accurate and faster on large datasets than ASTRAL, and also statistically consistent under the Multi-Species Coalescent model
- Using NJMerge or TreeMerge with concatenation using maximum likelihood (CA-ML): more accurate and much faster, greater scalability than CA-ML
- Each has some advantages over the other, both available on Github

Statistical Consistency: assumptions

- Multi-locus data, generated by a hierarchical model
  - Species tree generates gene trees
  - Gene trees generate sequences
- Suppose the number of genes and the sequence data per gene both go to infinity?

Statistical consistency, given a 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: Roch, Nute, & Warnow, Syst. Biol. 2018
  - No! Summary methods are not only not consistent, they can be positively misleading! (Felsenstein Zone)

S. Roch, M. Nute, and T. Warnow. "Long-branch attraction in species tree estimation: inconsistency of partitioned likelihood and topology-based summary methods." Systematic Biology 2018

Statistical consistency, given a bounded number of sites?

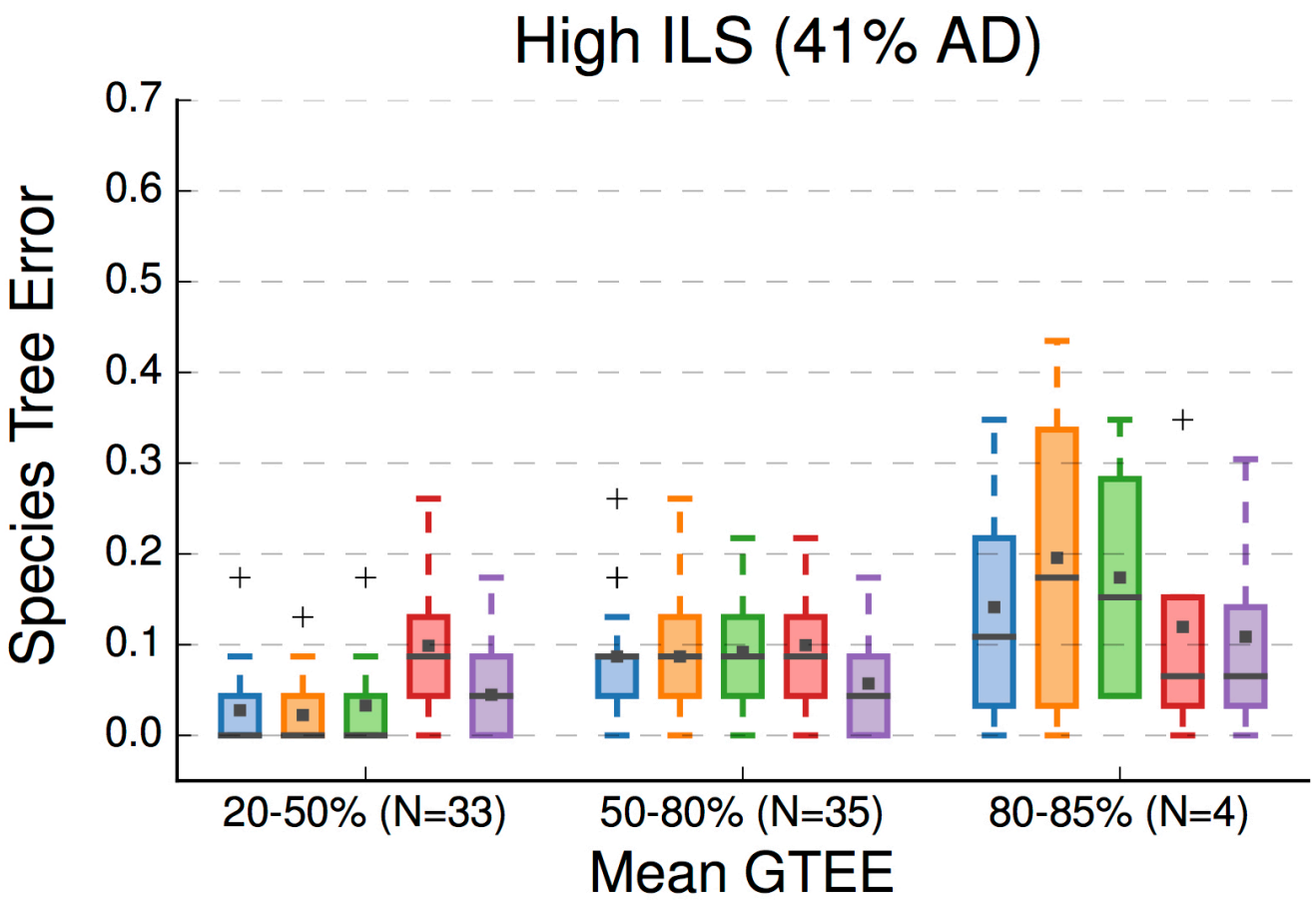
- Question #2: What about concatenation using maximum likelihood?
- Answer: Roch, Nute, & Warnow, Syst Biol. 2018
  - Not if fully partitioned! Concatenation using maximum likelihood, if fully partitioned is also not consistent and can be positively misleading (even if there is NO ILS)! (Felsenstein Zone)

S. Roch, M. Nute, and T. Warnow. "Long-branch attraction in species tree estimation: inconsistency of partitioned likelihood and topology-based summary methods." Systematic Biology 2018

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!

ASTRAL ASTRID MP-EST SVDquartets CA-ML

Summary Methods

Site-based Method

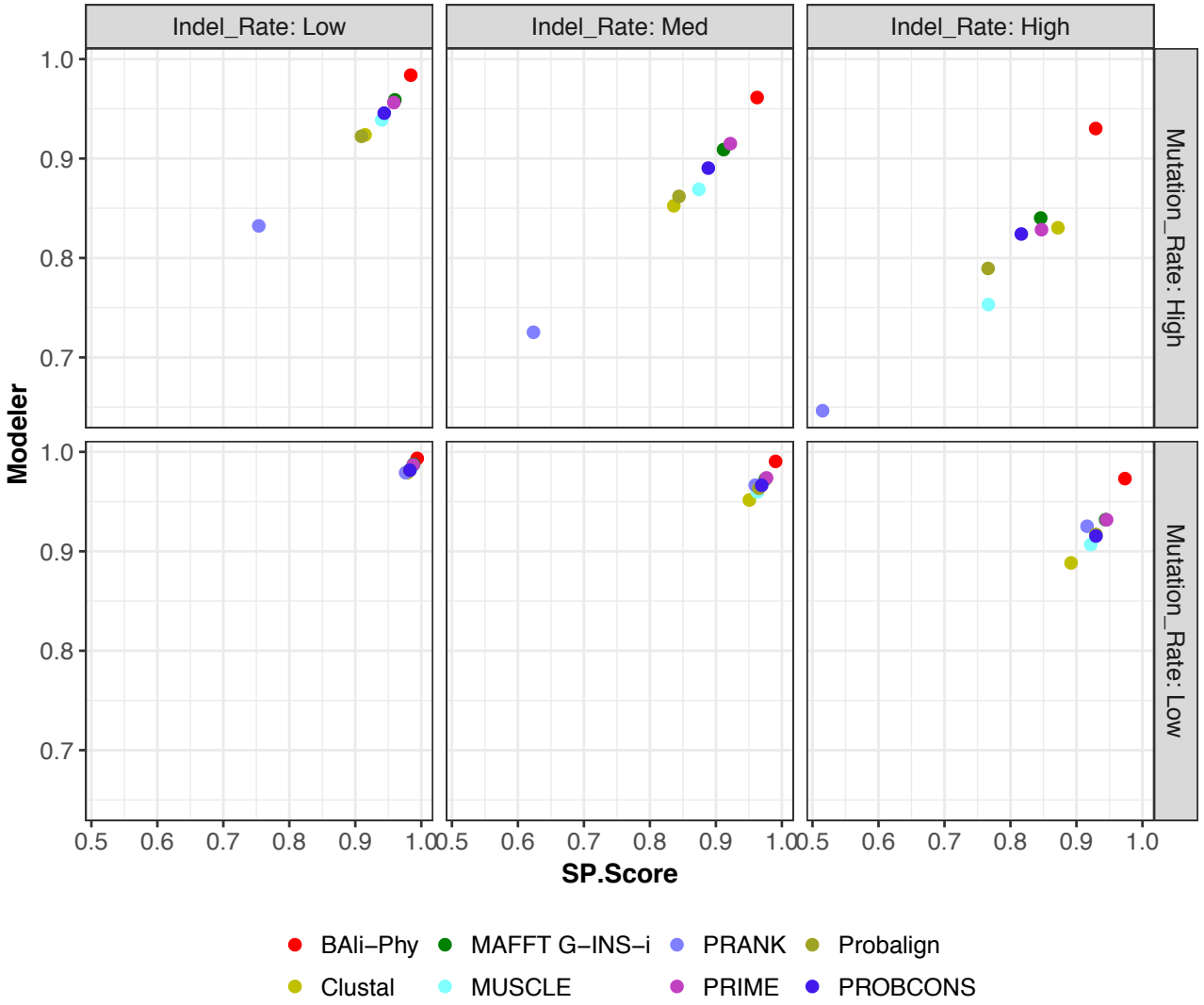
Error and Model Misspecification!!

- Assemble and annotate genomes (e.g., determine orthologs)
- Compute multiple sequence alignments of individual loci
- Construct gene trees
- Construct species tree
- Perform post-tree analyses (e.g., estimate dates, infer selection, etc.)

Error and Model Misspecification!!

- Assemble and annotate genomes (e.g., determine orthologs)
- Compute multiple sequence alignments of individual loci
- Construct gene trees
- Construct species tree
- Perform post-tree analyses (e.g., estimate dates, infer selection, etc.)

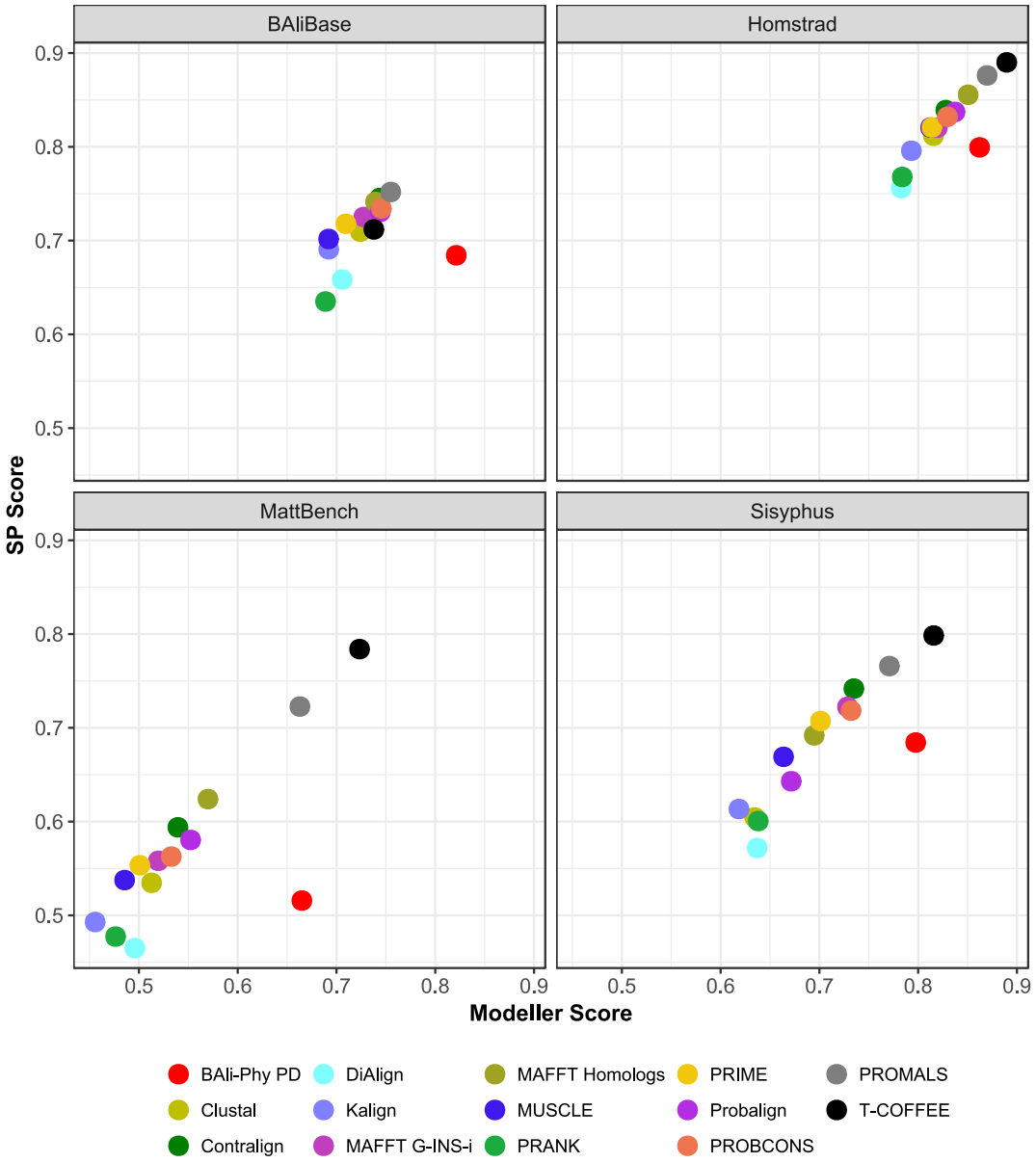
Alignment Accuracy on Simulated Datasets (120 datasets)



BAli-Phy is best!

Also: most methods produce short alignments, but not BAli-Phy or Prank (not shown here)

Alignment Accuracy on Protein Biological Benchmarks (1192 datasets)

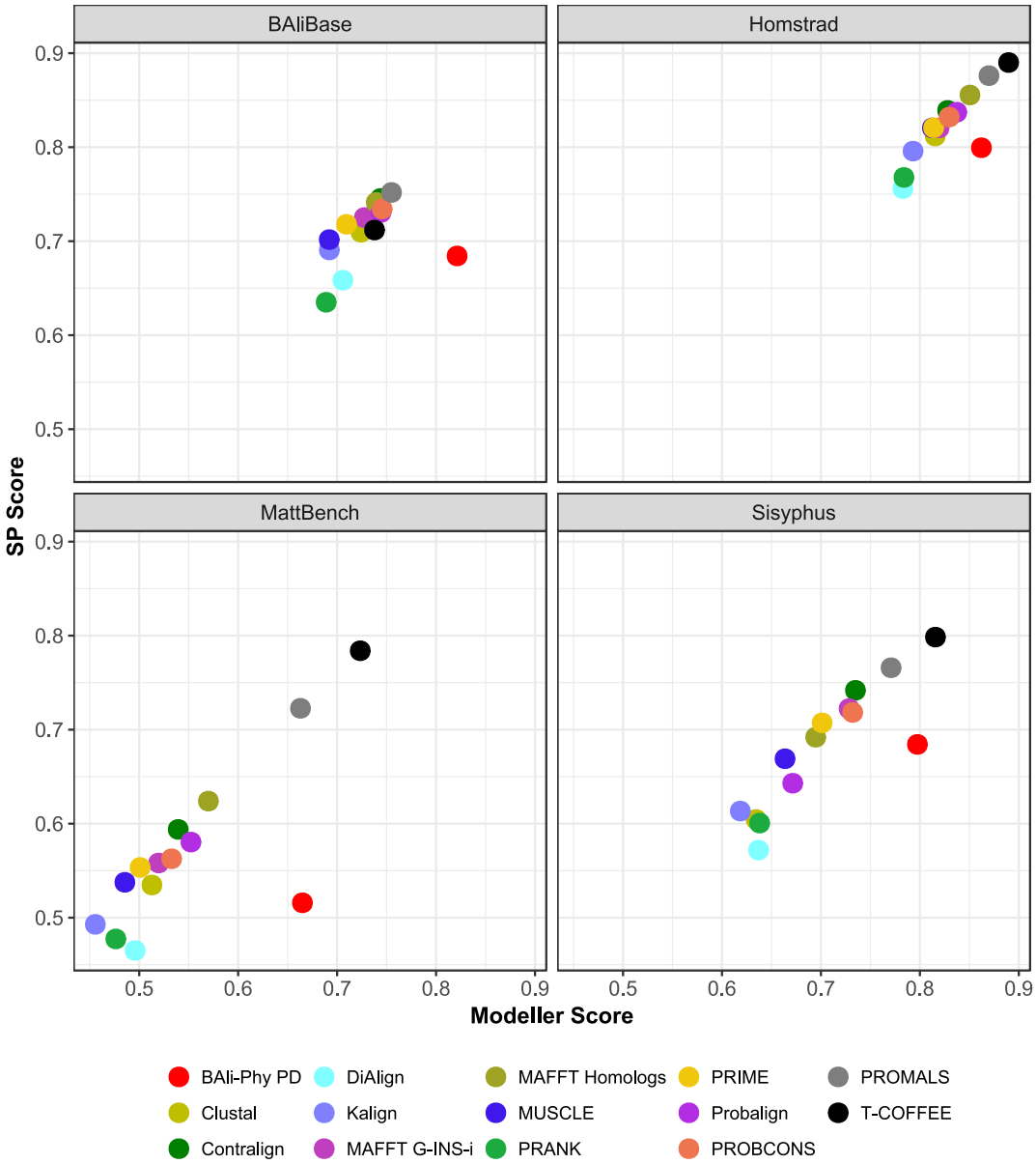


T-Coffee and PROMALS are best!

BAli-Phy good for Modeler score, but not so good for SP-Score (e.g., MAFFT better)

BAli-Phy produces longer alignments than the reference alignment (not shown here)

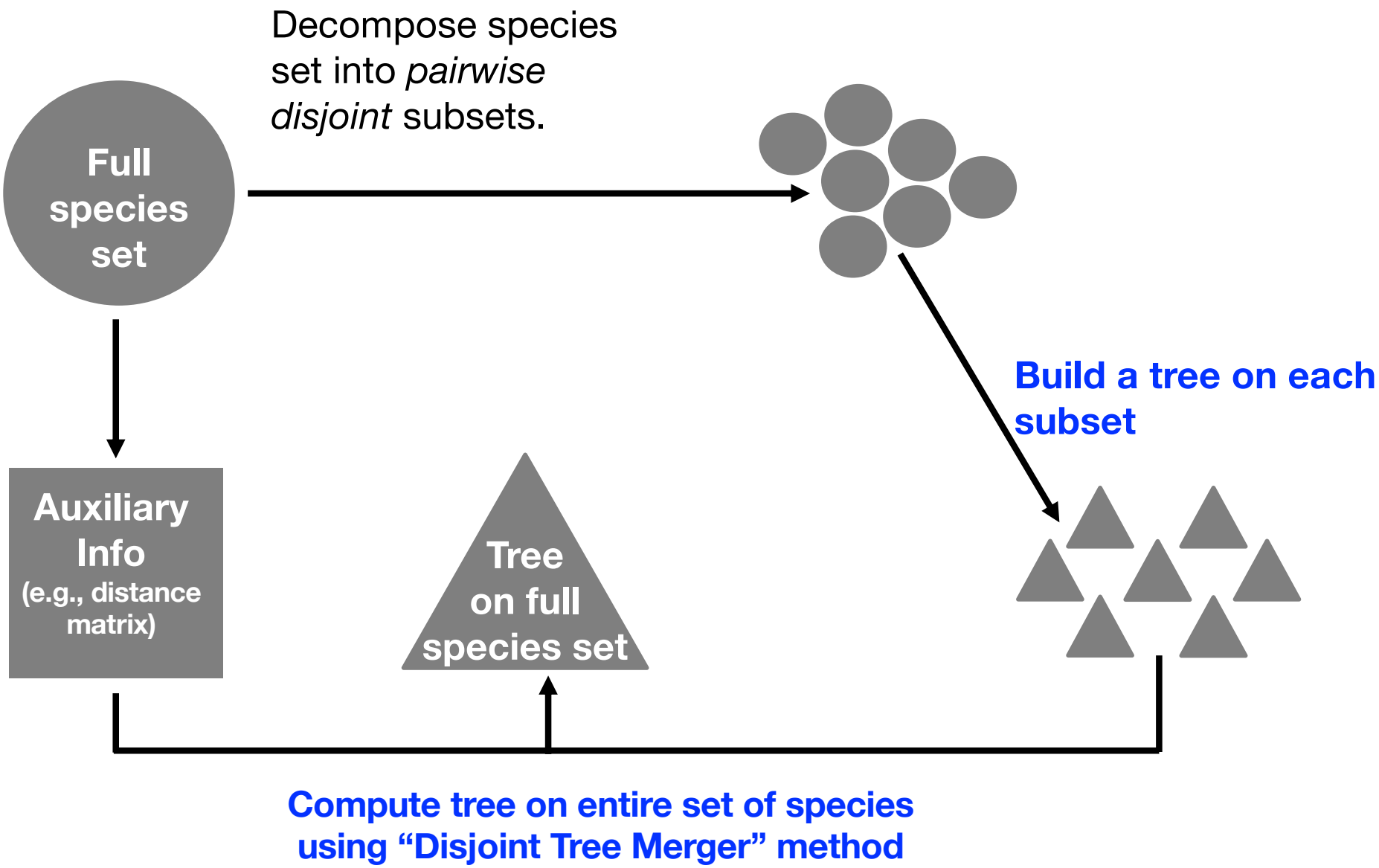
Alignment Accuracy on Protein Biological Benchmarks (1192 datasets)



Possible explanation:
Model misspecification.

But if this is the case, what does it mean for systematics in general?

Reducing Model Misspecification via Divide-and-Conquer



Closing Questions

- How should we evaluate accuracy?
- Do we need better models? Do they need to be identifiable?
- How should we select data?
- Taxon sampling helps – but increasing dataset size and evolutionary range increases heterogeneity and model misspecification: what adjustments are needed?
- Should we be confident, worried, or optimistic?

Acknowledgments






Software: all open source, see http://tandy.cs.illinois.edu/software.html

NSF grants DBI-1461364 and ABI-1458652
NSF graduate fellowships to Pranjal Vachaspati and Erin Molloy
HHMI graduate fellowship to Siavash Mirarab
Computing done at Blue Waters (part of NCSA)

Papers available at http://tandy.cs.illinois.edu/papers.html
Presentations available at http://tandy.cs.Illinois.edu/talks.html