

Advances in Phylogenomic Estimation

Tandy Warnow

The University of Illinois

Phylogenomic 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.)

Phylogenomic 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.)

Phylogenomic 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.)

Brief Comments on MSA Estimation

- Many good methods for aligning moderate sized datasets, but choice of alignment method depends on type of data (e.g., nucleotide vs. protein).
- MAFFT is very good for both types of data, and methods like PROMALS, T-Coffee, and ContrAlign are very good for proteins.
- Statistical co-estimation of alignments and trees (e.g., using BAli-Phy) also appealing, but expensive.



PASTA and UPP

scaling MSA to 1,000,000 sequences

- PASTA (Mirarab et al., J. Computational Biology 2014) – an improvement over SATe (Liu et al., Science 2009)
 - PASTA available at <https://github.com/smirarab/>
 - PASTA+BAlI-Phy at <http://github.com/MGNute/pasta>
- UPP (Nguyen et al., Genome Biology 2015) – uses PASTA for backbone, inserts sequences into the backbone using ensembles of HMMs, handles fragmentary sequences well
 - <https://github.com/smirarab/sepp>

Phylogenomic 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.)

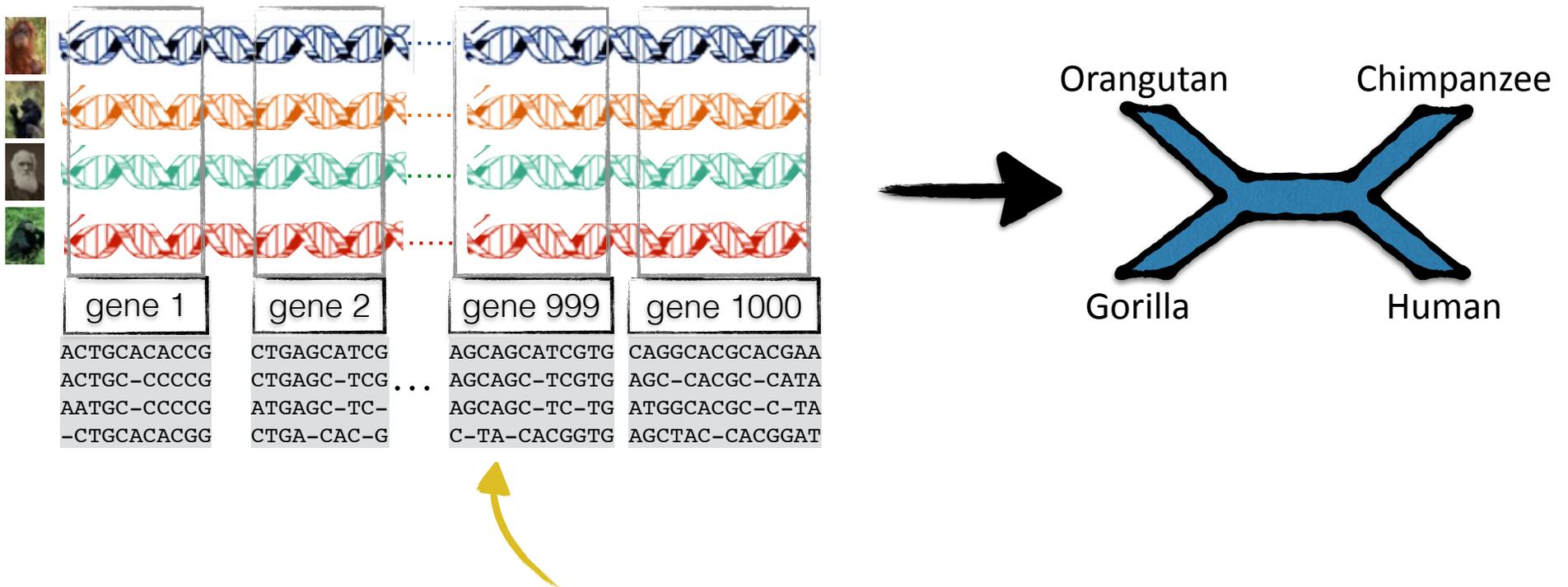
Brief Comments on Gene Tree Estimation

- Many methods: maximum parsimony, maximum likelihood, Bayesian tree estimation, distance methods, etc.
- Most good approaches are computationally intensive (e.g., heuristics for NP-hard optimization problems or Bayesian MCMC)
- Some of these are statistically consistent under standard Markov models of evolution
- Simulations under standard Markov models of evolution suggest maximum likelihood is a good approach to gene tree estimation

Phylogenomic Pipeline

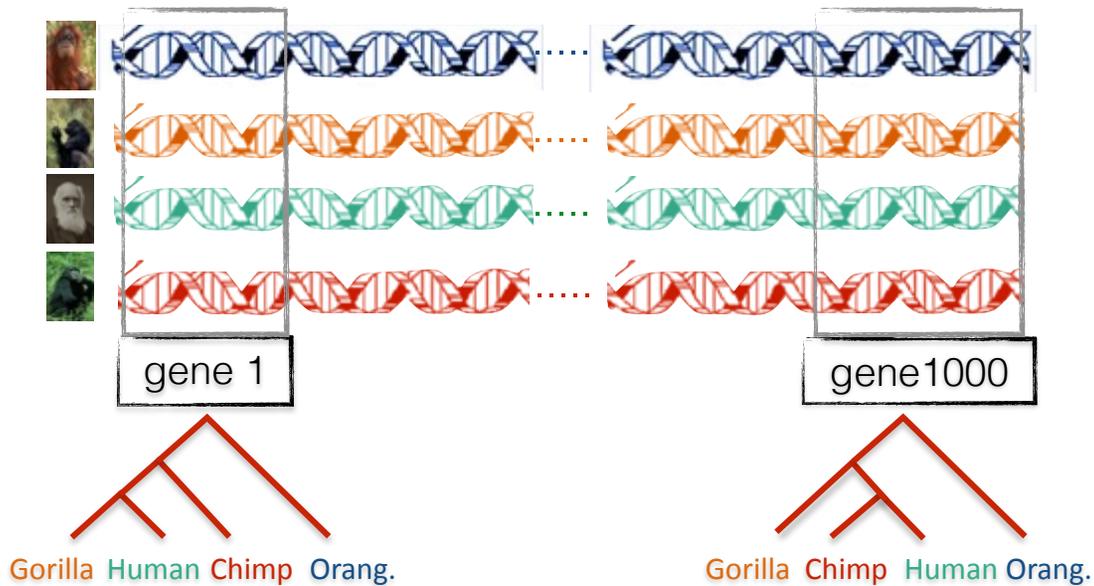
- Assemble and annotate genomes (e.g., determine orthologs)
- Compute multiple sequence alignments of individual loci
- Construct gene trees
- **Construct species tree**
- Perform post-tree analyses (e.g., estimate dates, infer selection, etc.)

phylogenomics



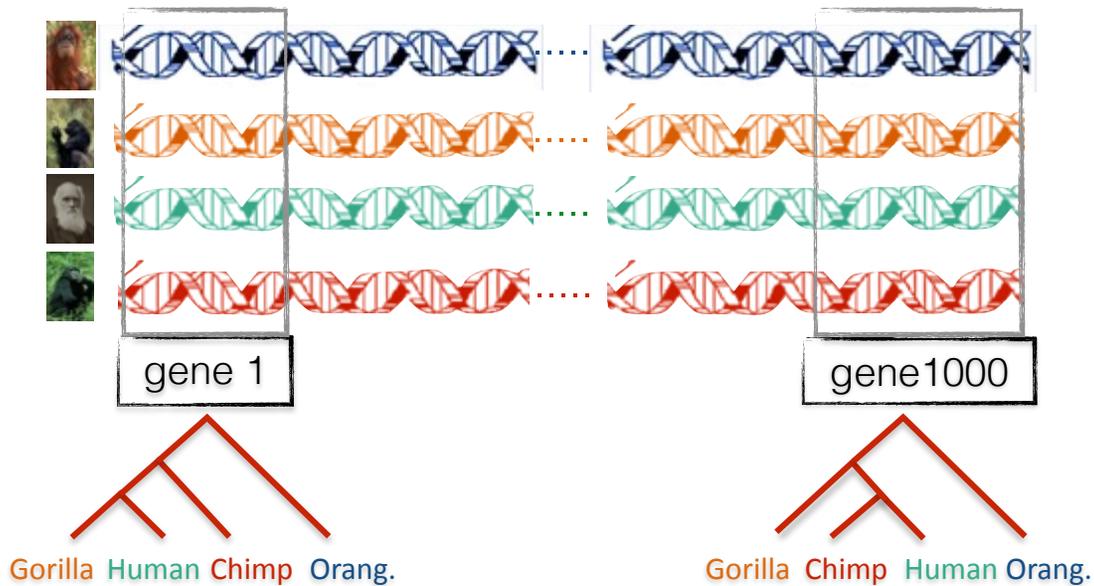
I'll use the term "gene" to refer to "c-genes":
recombination-free orthologous stretches of the genome

Gene tree discordance



- Multiple causes for discord, including
- Incomplete Lineage Sorting (ILS),
 - Gene Duplication and Loss (GDL),
 - and
 - Horizontal Gene Transfer (HGT)

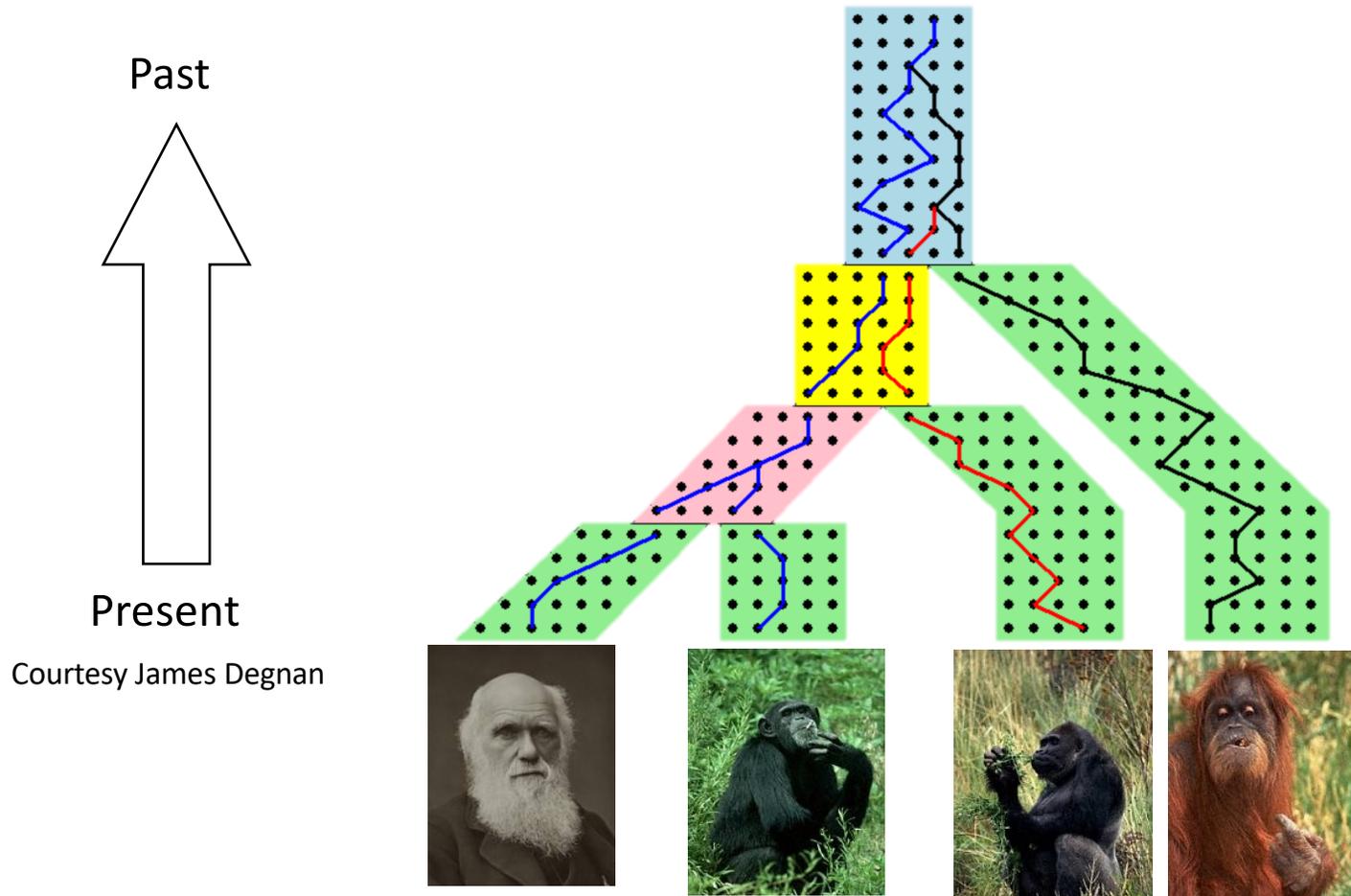
Gene tree discordance



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)



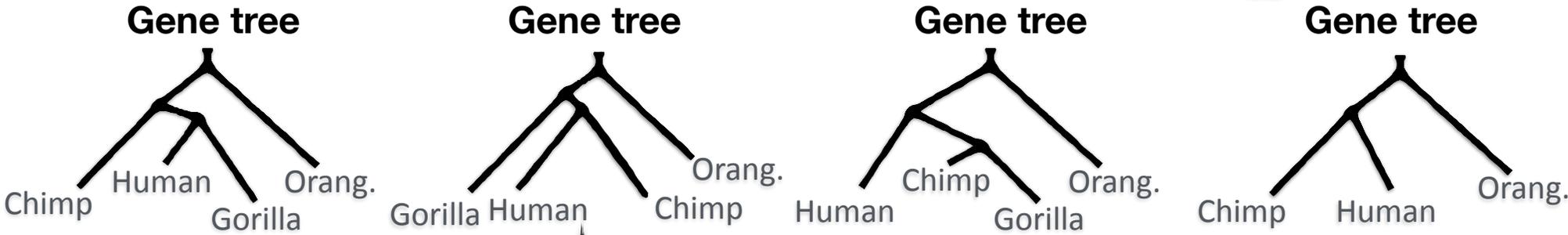
Gorilla and Orangutan are not siblings in the species tree, but they are in the gene tree.

Species tree



Gorilla Human Chimp Orangutan

Gene evolution model



Sequence evolution model

Sequence data (Alignments)

```
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
```

```
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
```

1

Sequence data (Alignments)

```
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?
- Suppose the number of genes and the sequence data per gene both go to infinity?

Four Basic Approaches

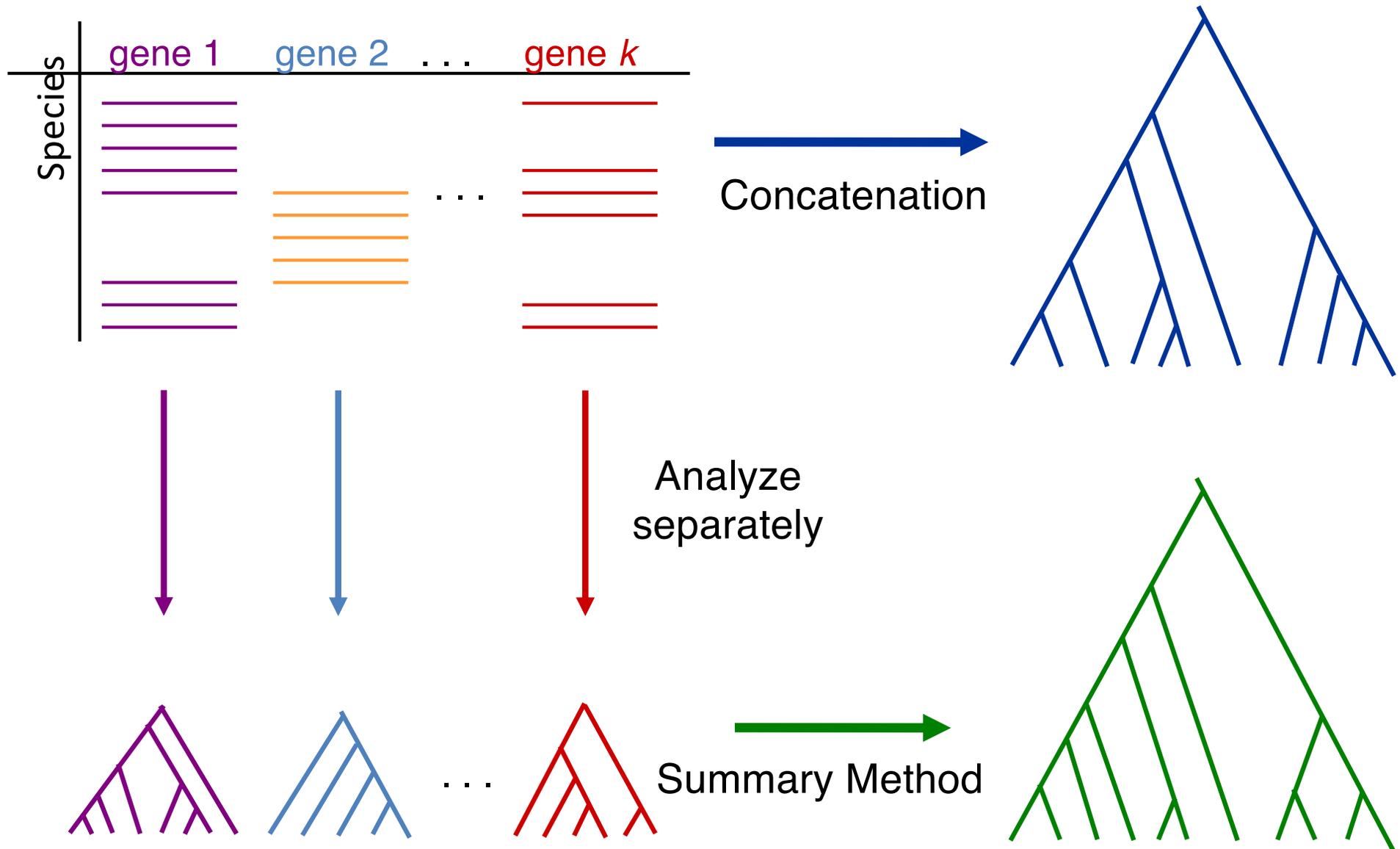
Statistically consistent methods:

- 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

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

Syst. Biol. 000:1–14, 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



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 Hejnol², Gonzalo Giribet¹



Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

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

Corrinne E. Grover^{1,2*}, Joseph P. Gallagher³, Josef J. Jareczek⁴, Justin T. Page⁵, Joshua A. Udall⁶, Michael A. Gore⁶, Jonathan F. Wendt⁷ *Journal of Biogeography* (2015)



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:10.1016/j.nature15697

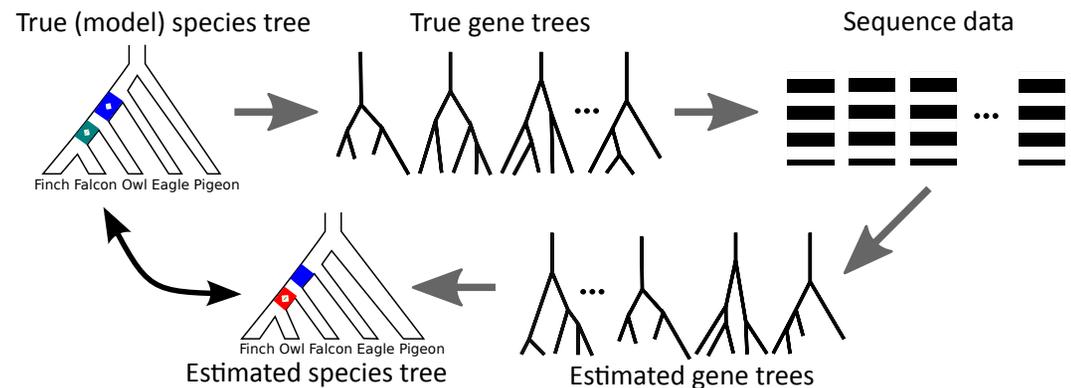
A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing

Richard O. Prum^{1,2*}, Jacob S. Berv^{3,4}, Alex Dornburg^{1,2,4}, Daniel J. Field^{1,5}, Jeffrey P. Townsend^{1,6}, Emily Moriarty Lemmon⁷ & Alan R. Lemmon⁸

Simulation study

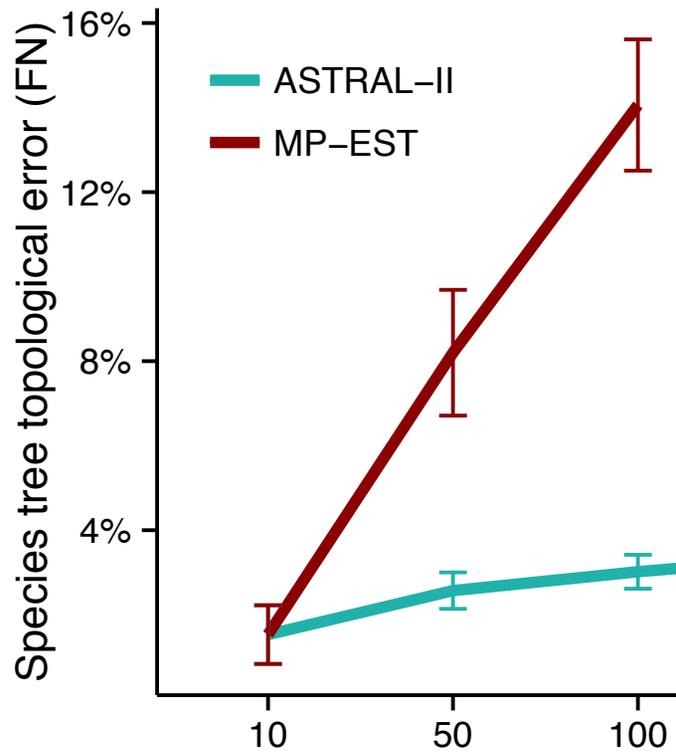
- Variable parameters:

- 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



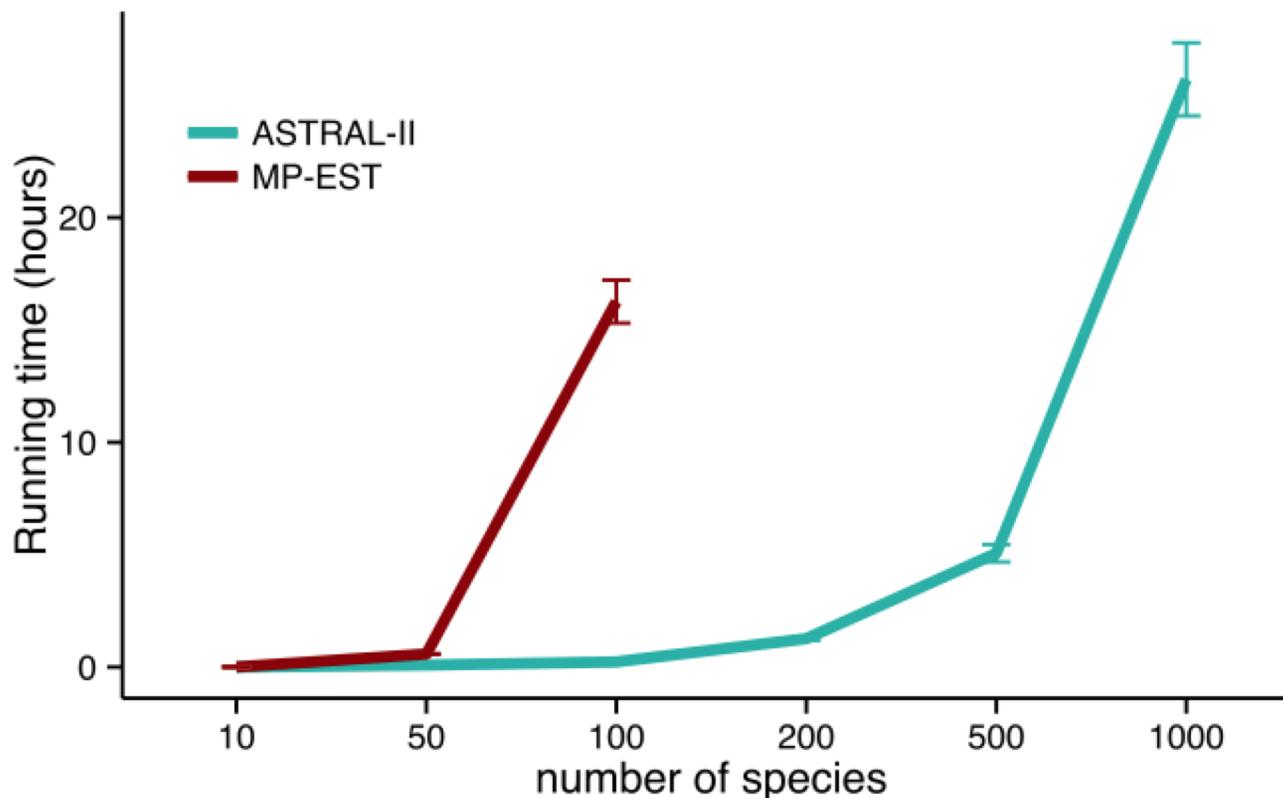
Used SimPhy, Mallo and Posada, 2015

Tree accuracy when varying the number of species



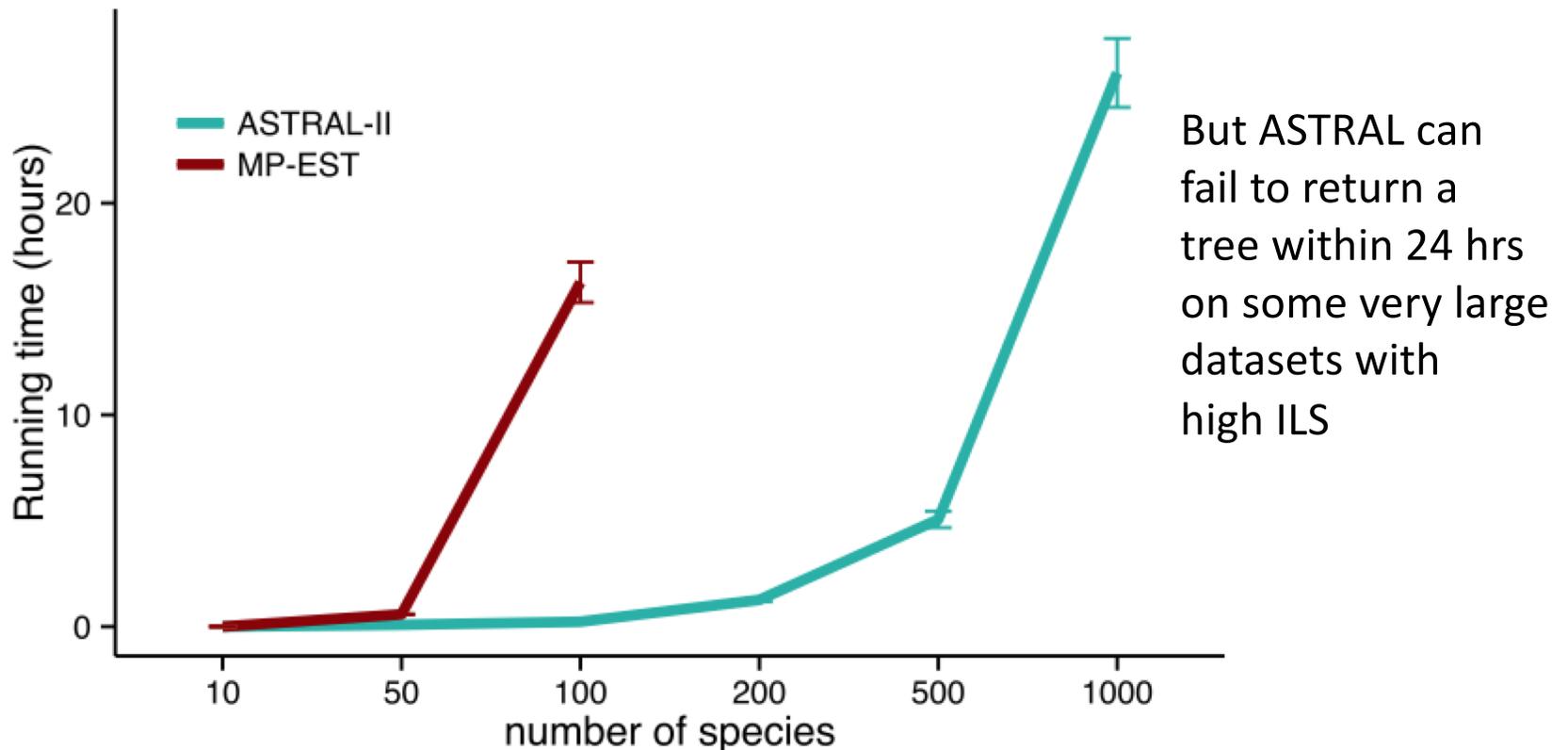
1000 genes, “medium” levels of recent ILS

Running time as function of # species



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

Running time as function of # species



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

ASTRID



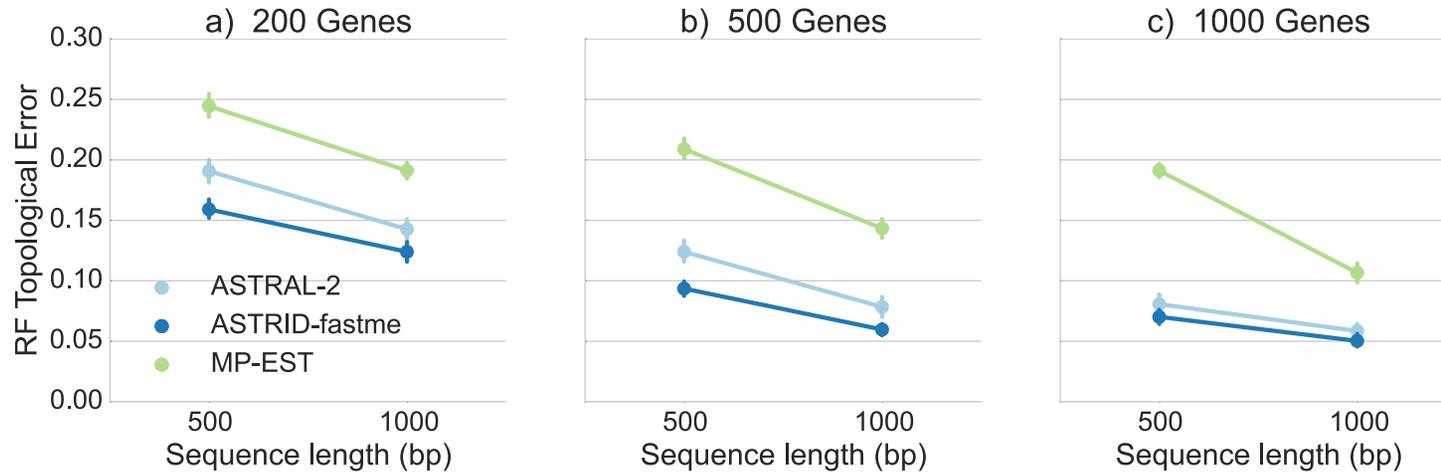
- ASTRID: Accurate species trees using internode distances, Vachaspati and Warnow, RECOMB-CG 2015 and BMC Genomics 2015
- Github site: <https://github.com/pranjalv123/ASTRID>

Algorithmic design:

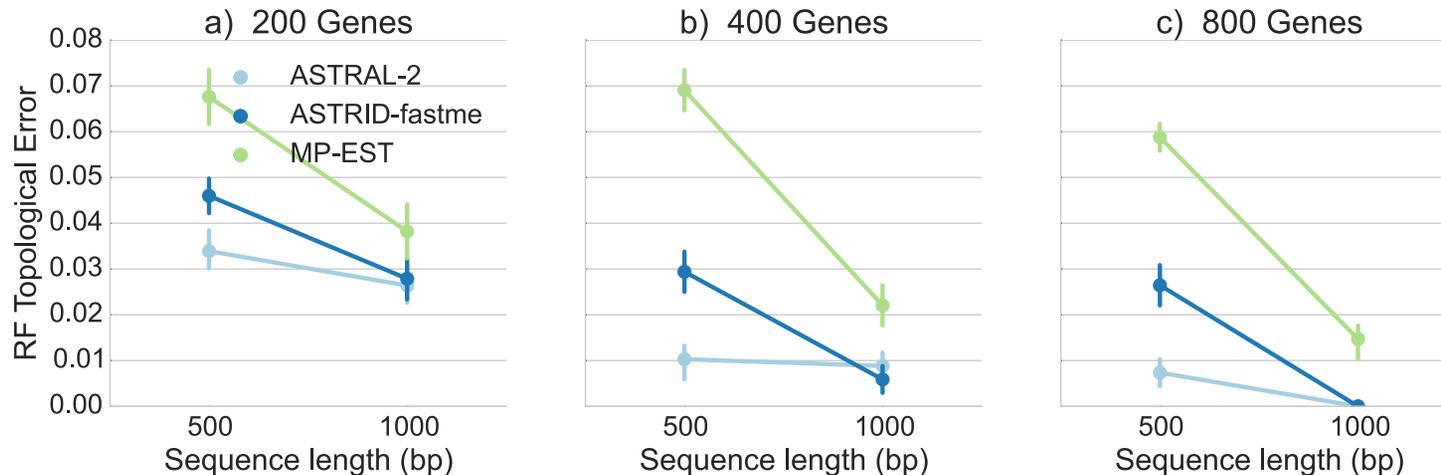
- Nearly the same as NJst (Liu and Yu, 2010)- computes a matrix of average leaf-to-leaf topological distances, and then computes a tree using FastME (more accurate than neighbor Joining and faster, too).
- Polynomial time and statistically consistent in the presence of ILS.

Both ASTRAL and ASTRID substantially outperform MP-EST

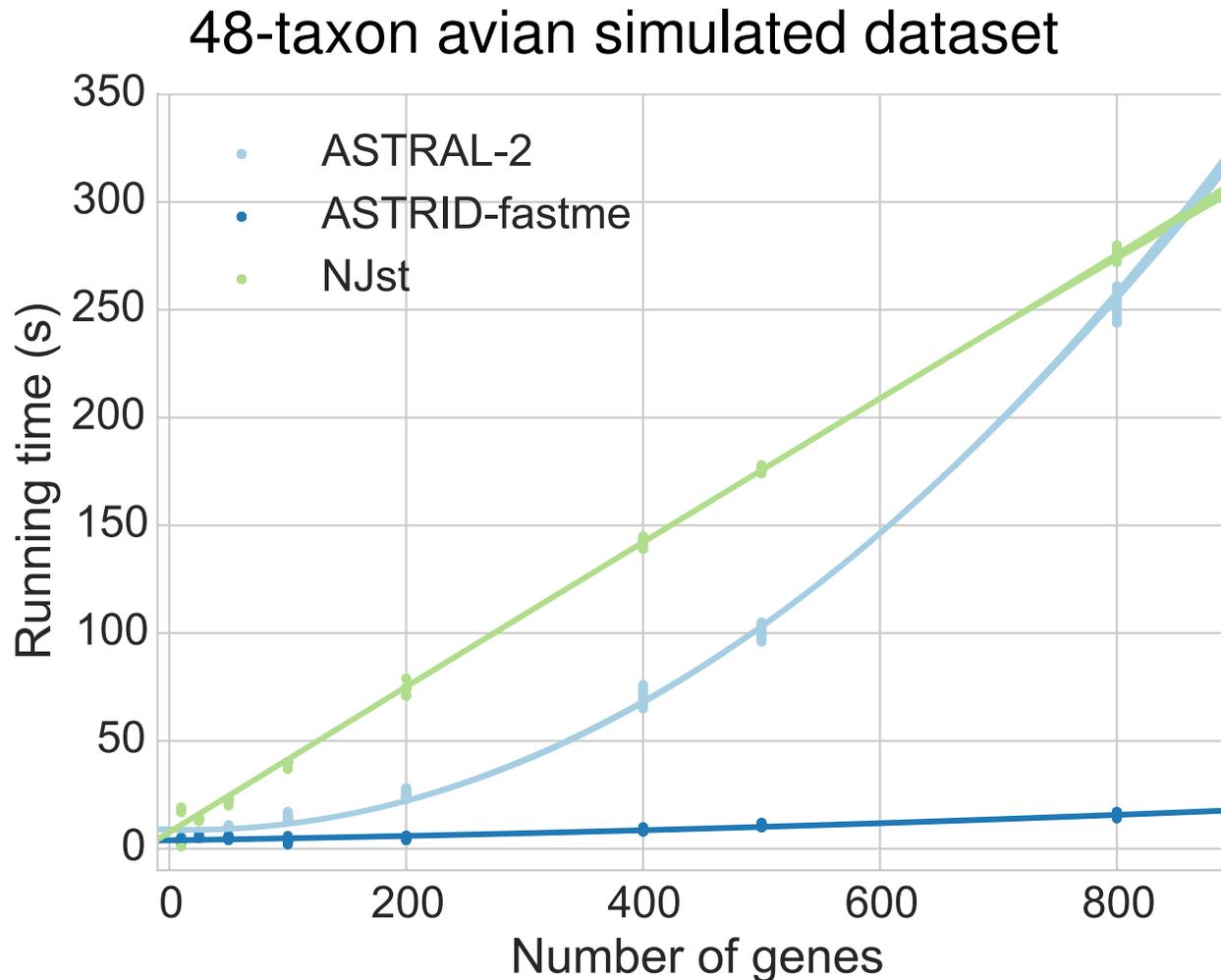
Avian simulated dataset



Mammalian simulated dataset



ASTRID is very fast



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

Should you filter?



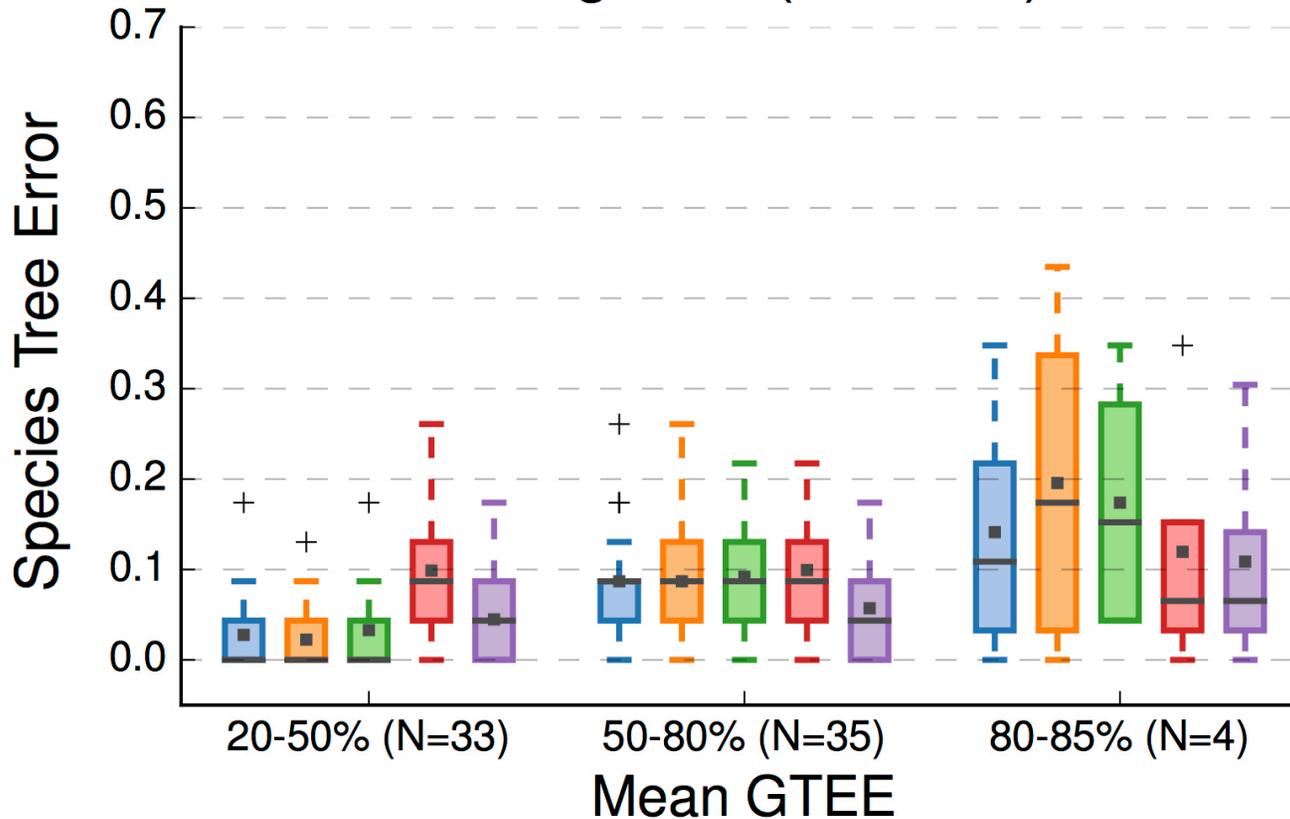
- Filtering genes based on missing data?
 - Generally not beneficial (see Molloy and Warnow, *Systematic Biology* 2018)
- Filtering genes based on gene tree estimation error?
 - Depends on conditions (see Molloy and Warnow, *Systematic Biology* 2018)
- Filtering fragmentary sequences from genes while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

Impact of Gene Tree Estimation Error

(from Molloy and Warnow 2017)



High ILS (41% AD)



Error is fraction of bipartitions that are not recovered

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

Legend: ASTRAL (blue), ASTRID (orange), MP-EST (green) are Summary Methods; SVDquartets (red), CA-ML (purple) are Site-based Method.

Which method to use?

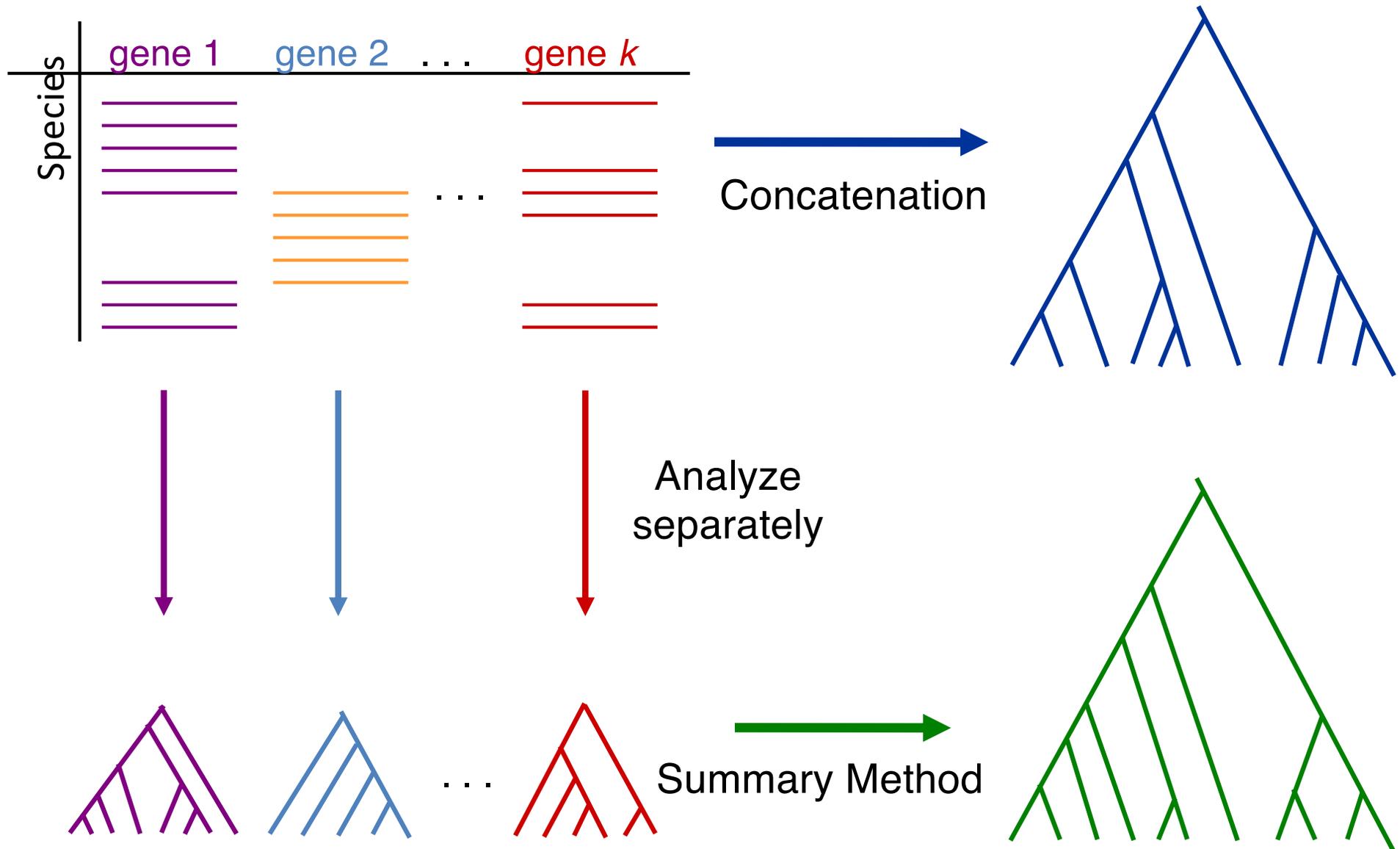
- Concatenation can be more accurate than alternative methods under some conditions (notably when gene tree estimation error is high, ILS is high, and number of loci is low)
- Site-based methods (such as SVDquartets and SVDquest) are expected to provide improved accuracy (compared to summary methods and concatenation) under some conditions
- Summary methods (ASTRAL, ASTRID, etc.) can be very accurate given highly accurate gene trees, and are fast
- Choice of method depends on the data (degree and cause of gene tree heterogeneity, gene tree estimation error, and number of species).

But -- species tree estimation for a large numbers of species is challenging

Scaling to large numbers of species

- Concatenation analyses: very expensive (e.g., 250 CPU years for the Avian Phylogenomics project with only 48 species)
- Site-based methods: don't seem to scale (SVDquartets and SVDquest require all quartet trees)
- 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.

Main competing approaches



NJMerge



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

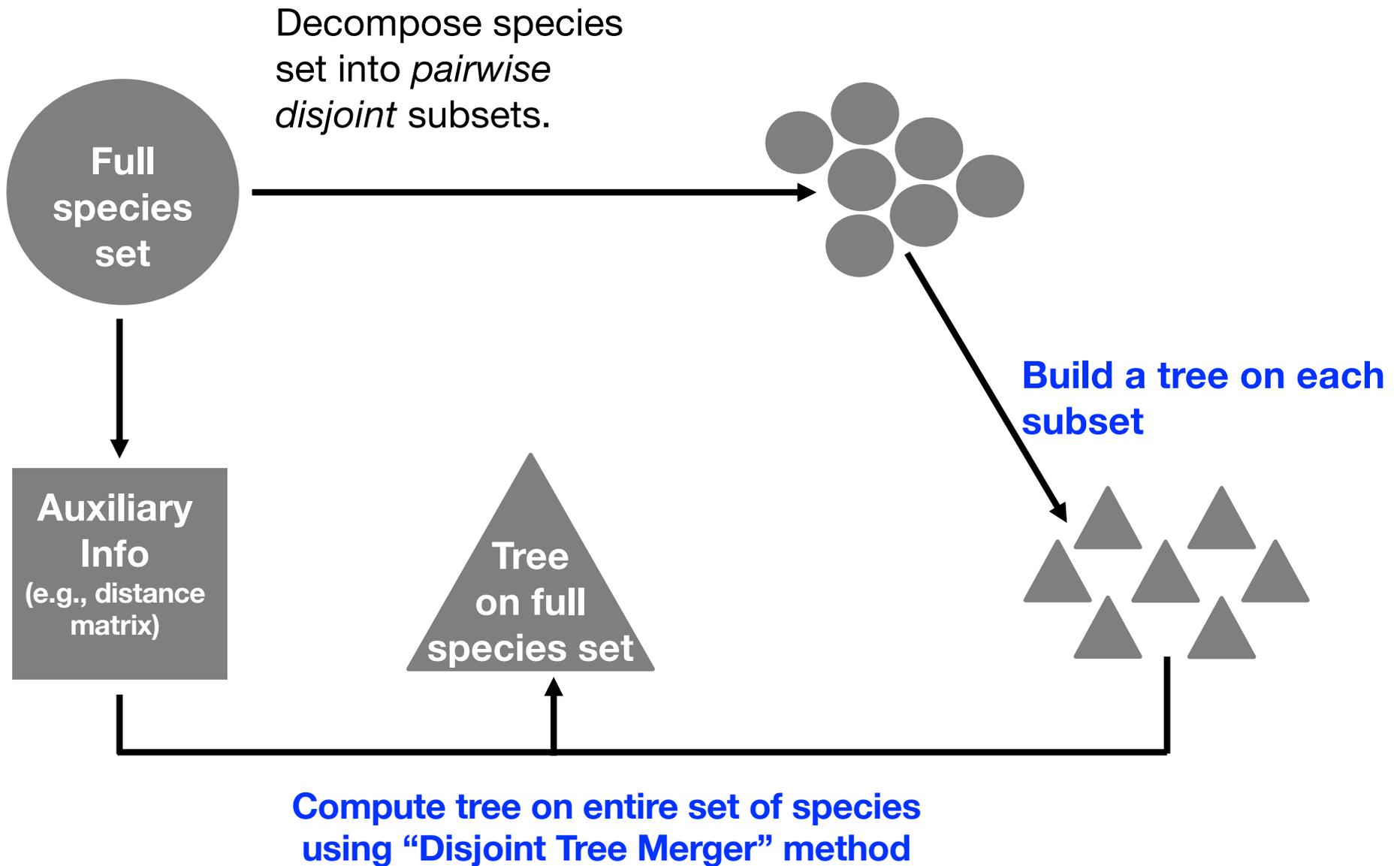
Algorithmic strategy:

- 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.

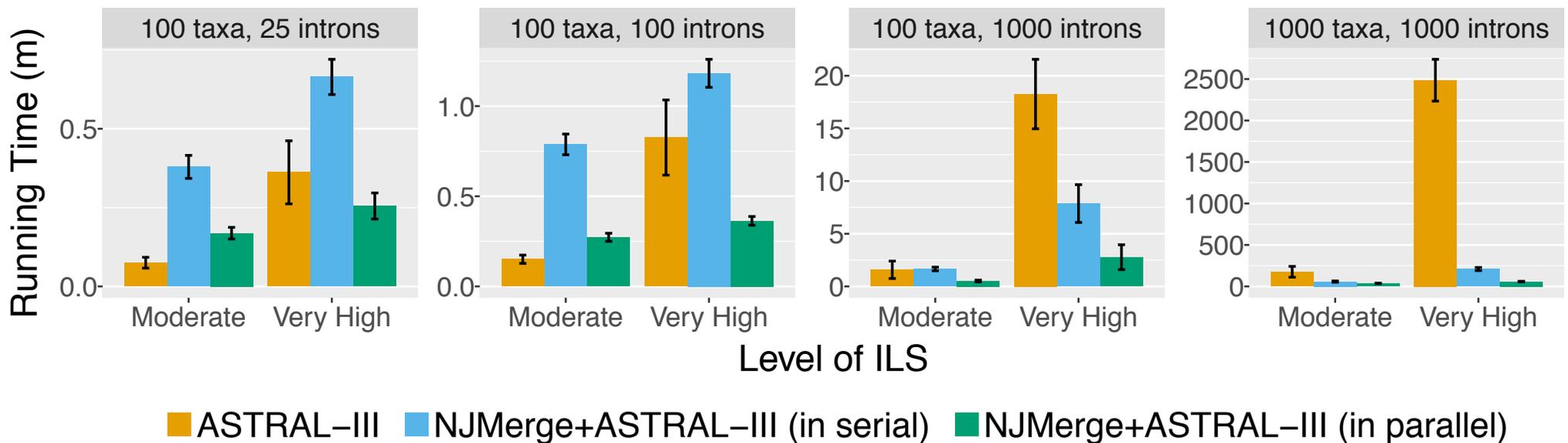
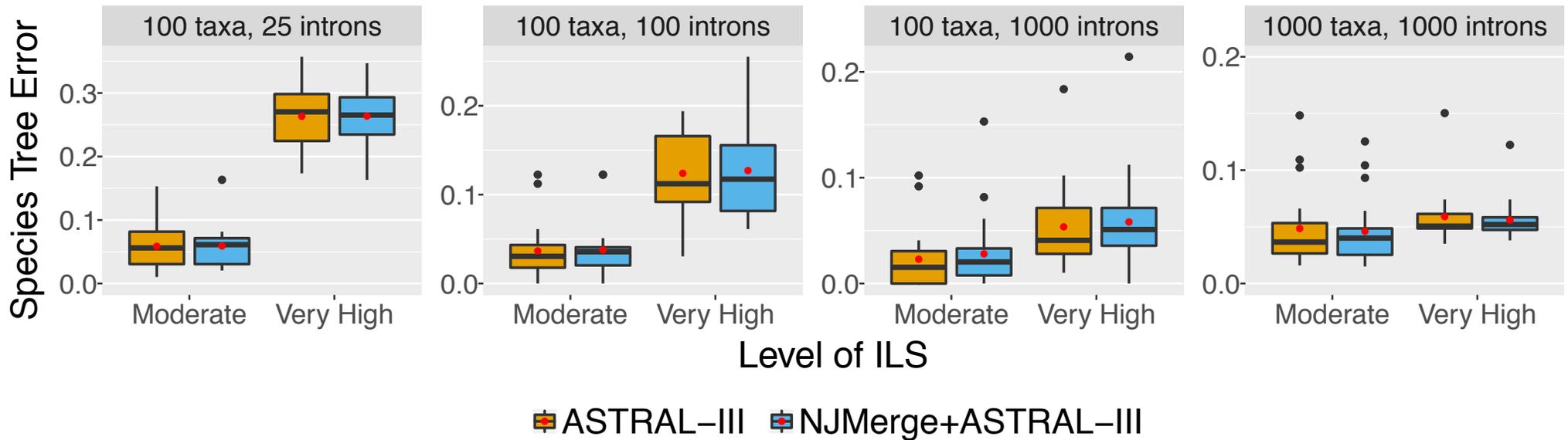
TreeMerge

- 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

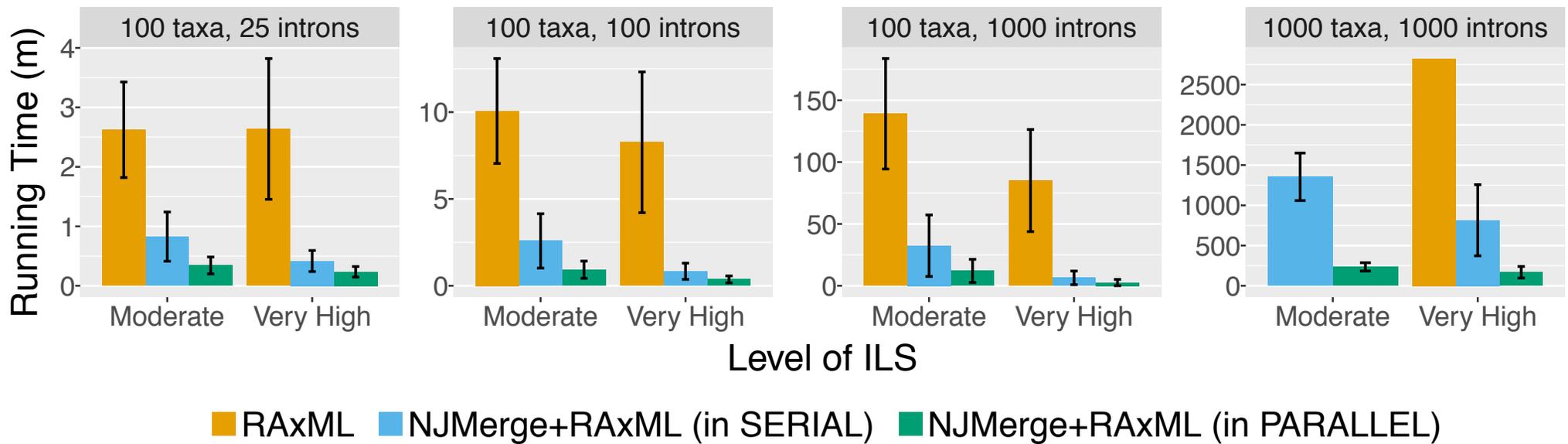
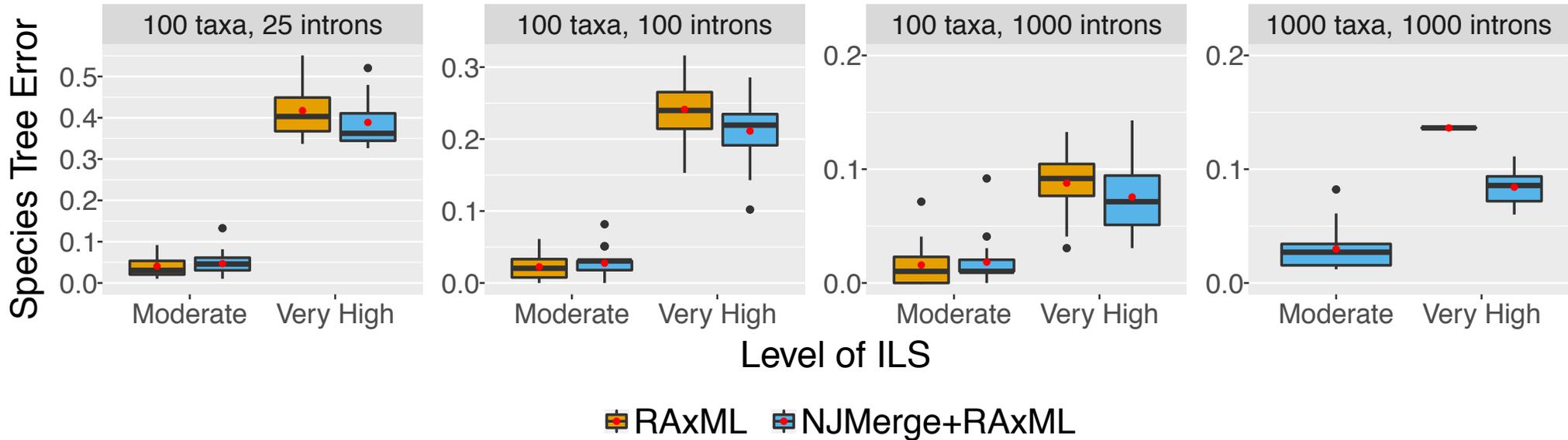
Divide-and-Conquer Pipeline



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, 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)



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

Phylogenomic 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.)

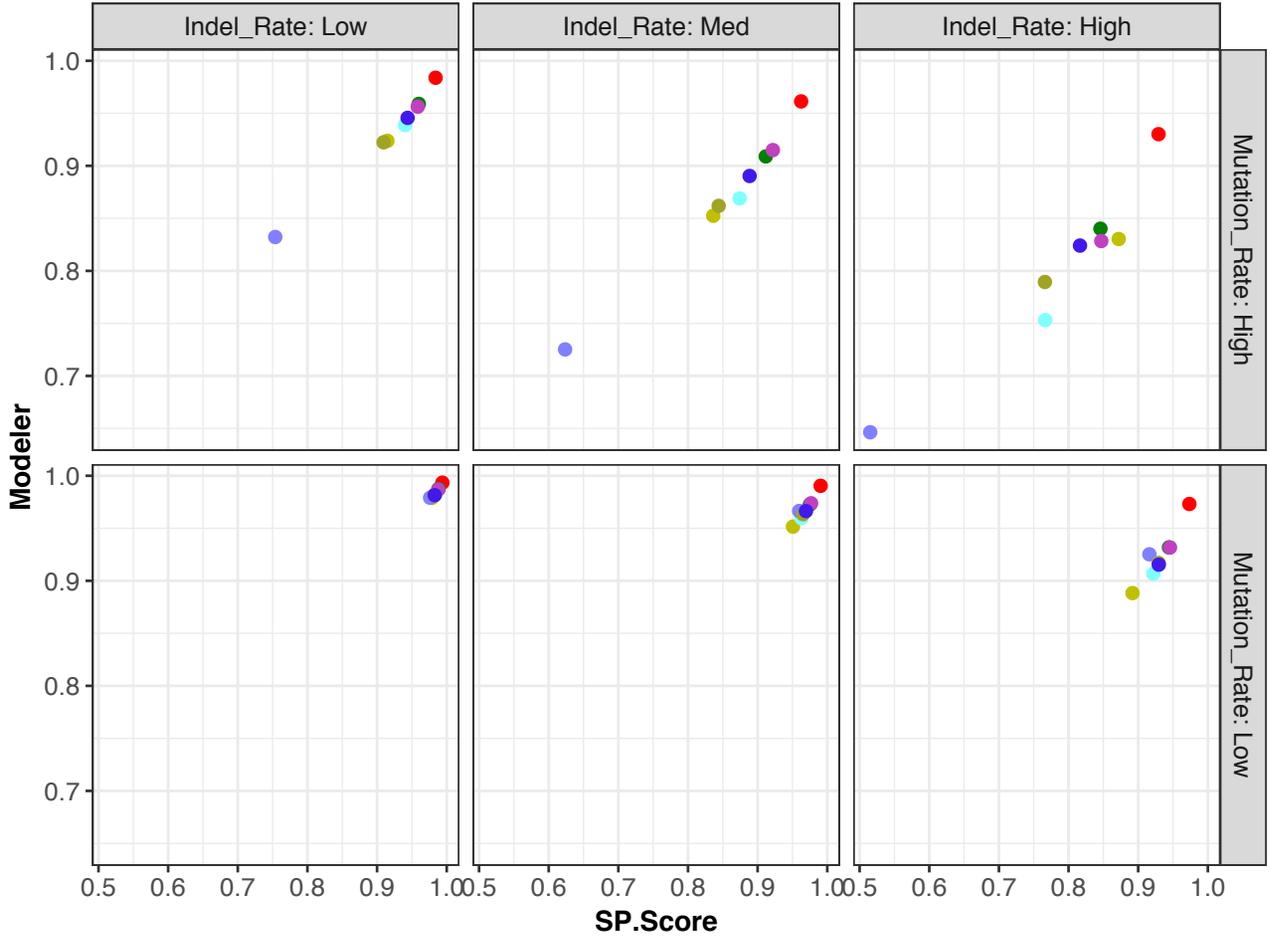
Benchmarking Statistical Multiple Sequence Alignment

Nute, Saleh, and Warnow 2018

Systematic Biology syy068, 2018, doi:10.1093/sysbio/syy068.



Alignment Accuracy on Simulated Datasets (120 datasets)

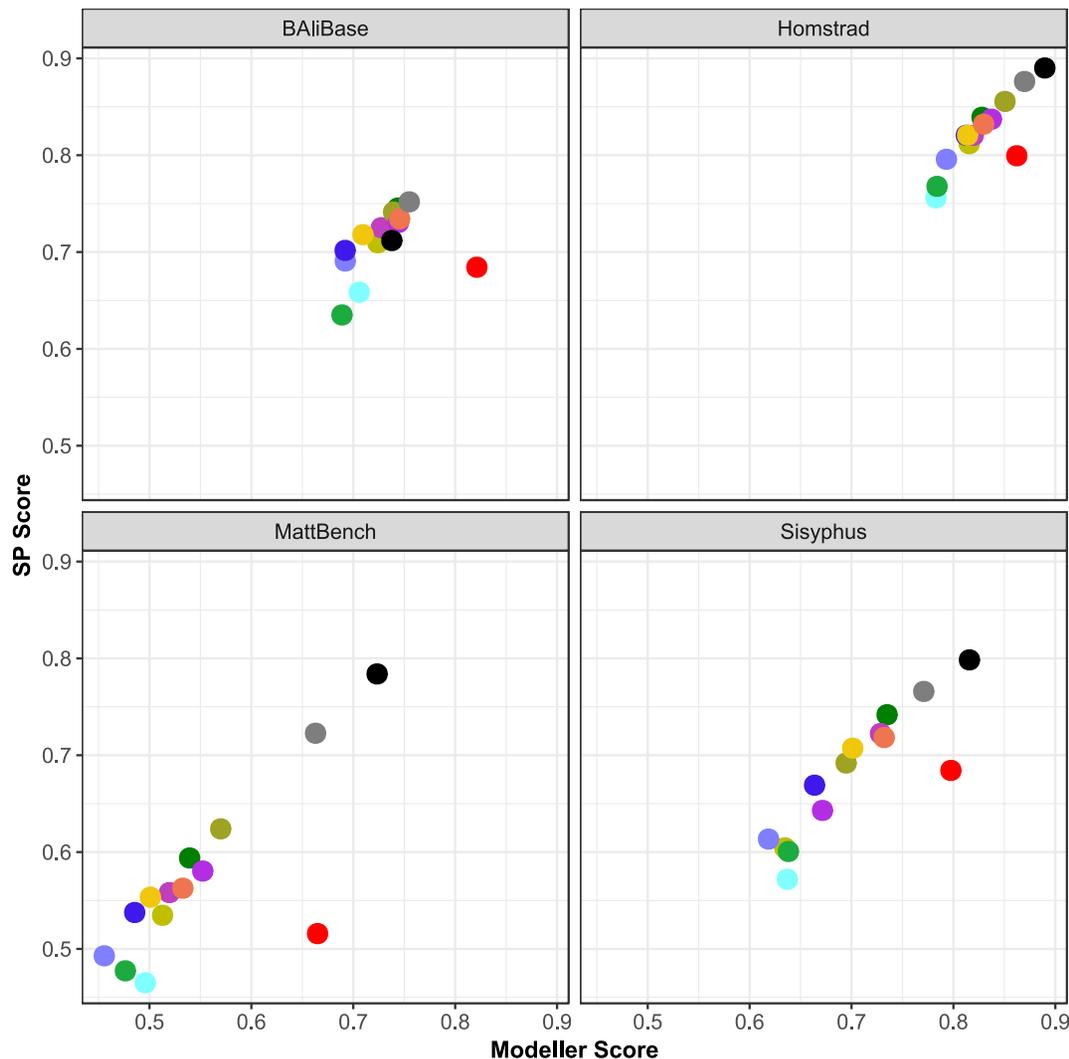


BAli-Phy is best!

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

- BAli-Phy
- MAFFT G-INS-i
- PRANK
- Probalign
- Clustal
- MUSCLE
- PRIME
- PROBCONS

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)

Observations

- Bali-Phy is much more accurate than all other methods on simulated datasets
- Bali-Phy is generally less accurate than the top half of these methods on biological datasets, especially with respect to SP-score (recall)

Why is there a difference in accuracy?

- Possible explanations:
 - Model misspecification (proteins don't evolve under the Bali-Phy model)
 - Structural alignments and evolutionary alignments are different
 - The structural alignments are not correct (perhaps over-aligned)
- All these explanations are likely true, but the relative contributions are unknown.

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

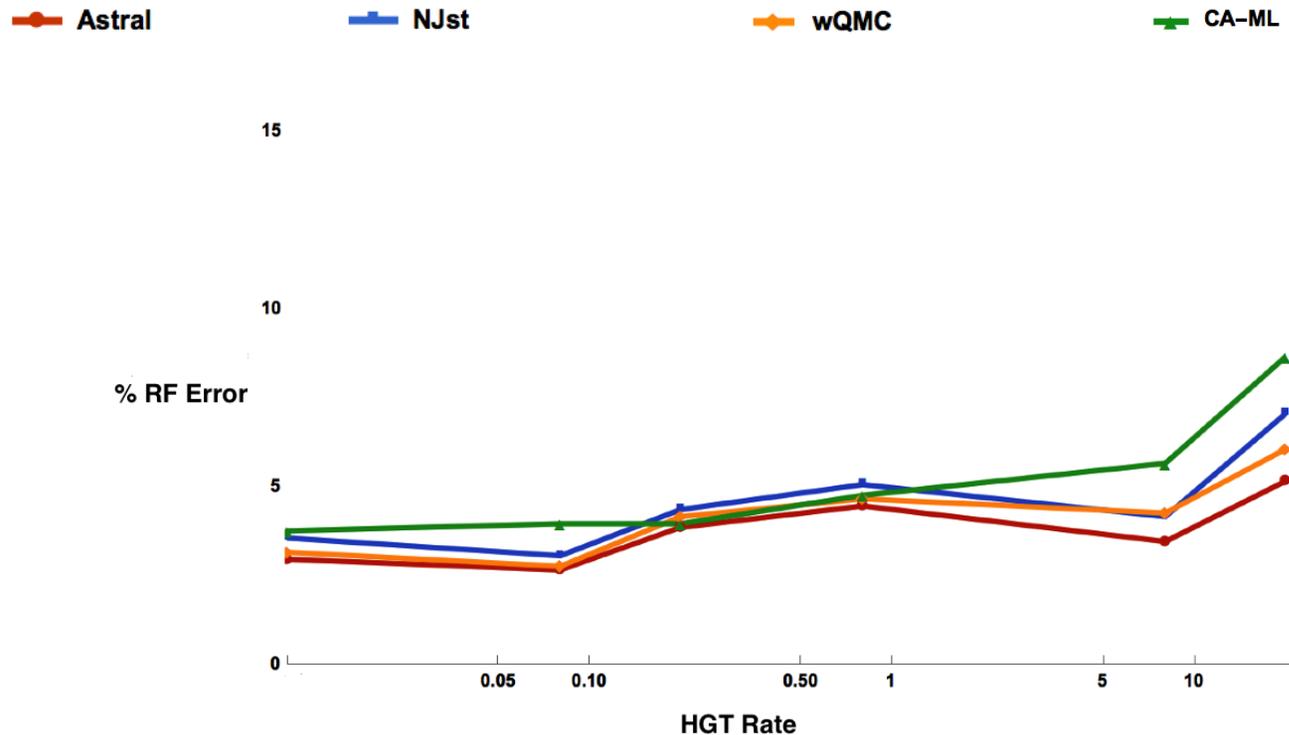
Computing done at Blue Waters (part of NCSA)

Papers available at <http://tandy.cs.illinois.edu/papers.html>

Presentations available at <http://tandy.cs.illinois.edu/talks.html>

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 INDELible⁸, 1000 gene trees estimated from GTR simulated sequences using FastTree-2⁷

⁷Price, Dehal, Arkin 2015

⁸Fletcher, Yang 2009