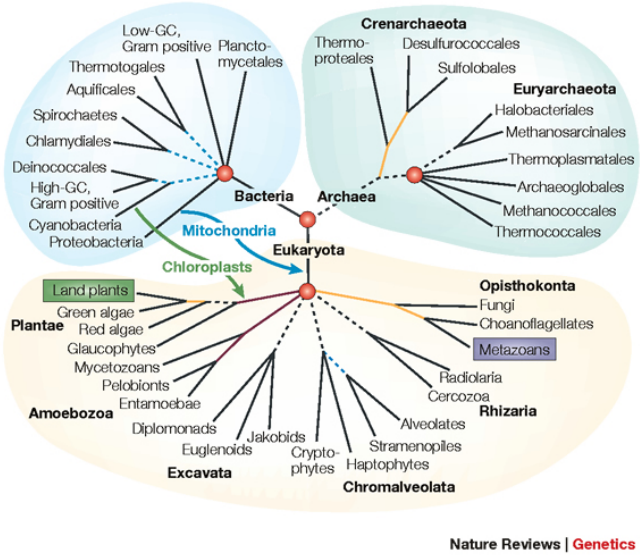
New Methods for Estimating the Tree of Life

Tandy Warnow

The University of Illinois

Phylogenomics



Phylogeny + genomics = genome-scale phylogeny estimation

.

Phylogenomic pipeline

- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- Compute multiple sequence alignments for each locus, and construct gene trees
- Compute species tree or network:
  - Combine the estimated gene trees, OR
  - Estimate a tree from a concatenation of the multiple sequence alignments
- Get statistical support on each branch (e.g., bootstrapping)
- Estimate dates on the nodes of the phylogeny
- Use species tree with branch support and dates to understand biology

Phylogenomic pipeline

- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- Compute multiple sequence alignments for each locus, and construct gene trees
- Compute species tree or network:
  - Combine the estimated gene trees, OR
  - Estimate a tree from a concatenation of the multiple sequence alignments
- Get statistical support on each branch (e.g., bootstrapping)
- Estimate dates on the nodes of the phylogeny
- Use species tree with branch support and dates to understand biology

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/UIUC


S. Mirarab,
UT-Austin /UCSD


N. Nguyen
UT-Austin/UCSD

- 2014 *PNAS* study: 103 plant transcriptomes, 400-800 single copy "genes"
- 2019 *Nature* study: much larger!

Major Challenges:

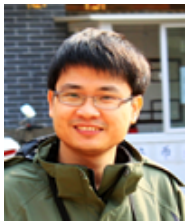
- Large alignments (and sequence length heterogeneity)
- Multi-copy genes omitted (9500 -> 400)
- Massive gene tree heterogeneity consistent with ILS

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 challenges:

- Multi-copy genes omitted
- Massive gene tree heterogeneity consistent with ILS
- Concatenation analysis took 250 CPU years

Large datasets are difficult

- Two dimensions:
  - Number of loci
  - Number of species (or individuals)
- Missing data
- Heterogeneity
- Many analytical pipelines involve Maximum likelihood and Bayesian estimation

This talk

- Part I: New methods for multiple sequence alignment
- Part II: New methods for maximum likelihood phylogenetic placement
- Part III: New methods for maximum likelihood tree estimation
- Part IV: New methods for species tree estimation

Some of this work is Not Yet Published (NYP), but all the codes described are available in open-source form on github

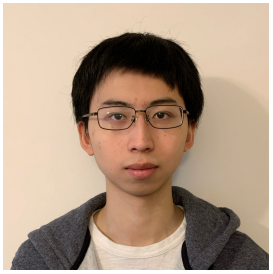
Please contact me if you wish to collaborate!

Part I: Multiple sequence alignment

- Aligning large datasets:
  - SATé (2009), PASTA (2014), MAGUS (2021) and **recursive MAGUS (2022**)
- Constructing alignments with sequence length heterogeneity:
  - UPP (2015), WITCH (2022), WITCH-ng (2023), UPP2 (2023), HMMerge (2023), and **EMMA (2023)**
  - These methods can also be used to add sequences into an existing alignment

Smirnov
MAGUS

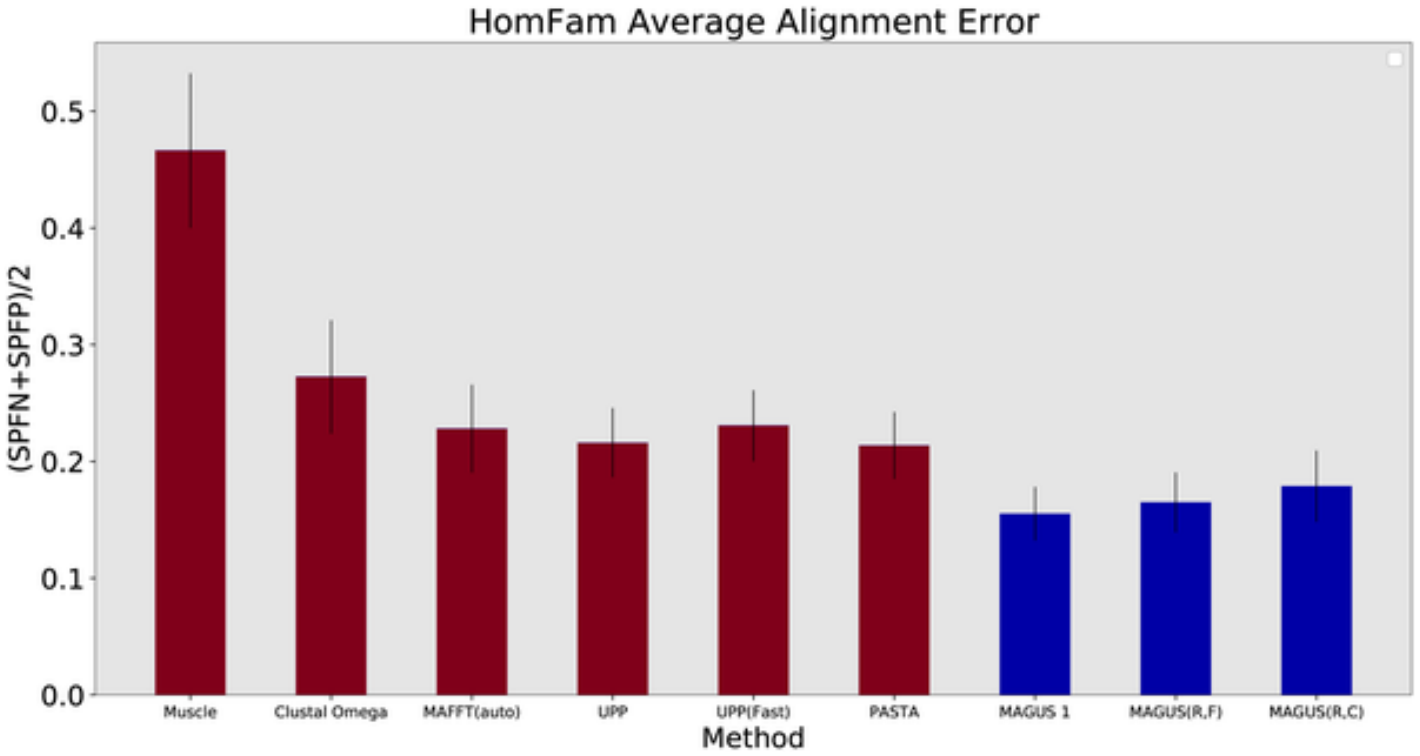
Shen
WITCH, EMMA



Liu
WITCH-ng

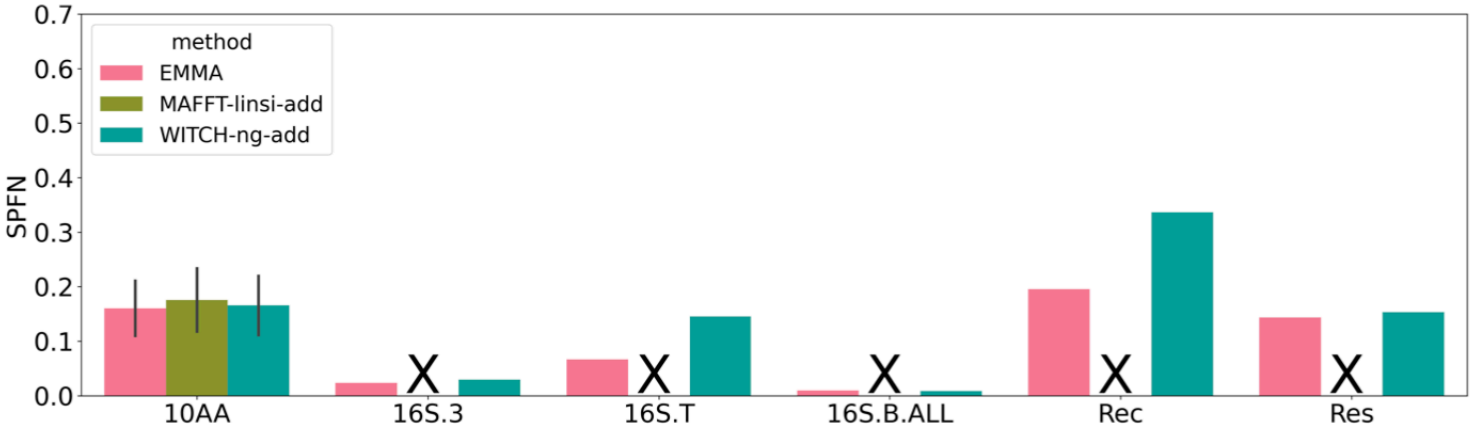
Park
HMMerge, UPP2

MAGUS – Highly Accurate Multiple Sequence Alignment for large datasets



Smirnov V (2021) Recursive MAGUS: Scalable and accurate multiple sequence alignment. PLOS Computational Biology 17(10): e1008950. https://doi.org/10.1371/journal.pcbi.1008950
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008950

EMMA: Extending Multiple alignments using MAFFT--add)



Biological datasets ranging from ~300 to ~170K sequences

Recombinase and Resolvase are datasets studied by K.P. Williams (SNL-Livermore)

Authors: C. Shen, B. Liu, K.P. Williams, and T. Warnow

To appear: Workshop on Algorithms for Bioinformatics 2023

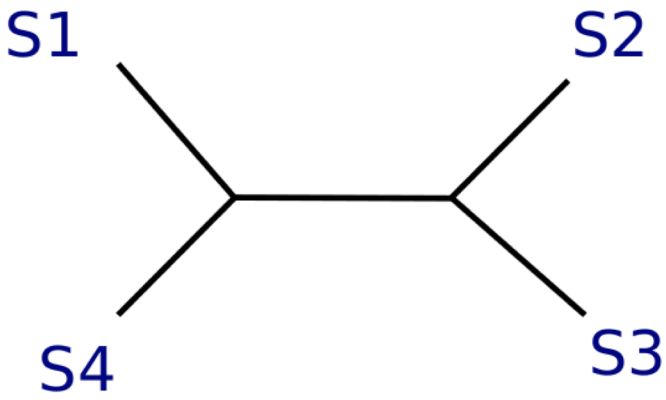
Part II: Phylogenetic placement

- Adding aligned sequences into a tree
- Applications:
  - Taxonomic identification of reads in metagenomics and microbiome analysis
  - Updating large trees

Phylogenetic Placement

Phylogenetic placement problem: *Given a query sequence and multiple sequence alignment, determine the placement into an existing reference tree.*

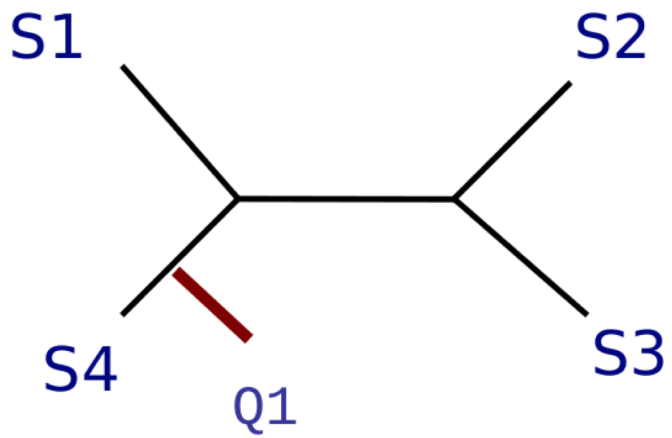
```
S1  = -AGGCTATCACCTGACCTCCA-AA
S2  = TAG-CTATCAC--GACCGC--GCA
S3  = TAG-CT-------GACCGC--GCT
S4  = TAC----TCAC--GACCGACAGCT
Q1  = -------T-A--AAAC--------
```



Phylogenetic Placement

Phylogenetic placement problem: *Given a query sequence and multiple sequence alignment, determine the placement into an existing reference tree.*

```
S1  = -AGGCTATCACCTGACCTCCA-AA
S2  = TAG-CTATCAC--GACCGC--GCA
S3  = TAG-CT-------GACCGC--GCT
S4  = TAC----TCAC--GACCGACAGCT
Q1  = -------T-A--AAAC--------
```



Existing Methods for Phylogenetic Placement

Maximum likelihood methods (expensive to run):

- **pplacer** (Matsen et al., 2010) is currently the most accurate method, but fails on large trees (e.g., some with 4000 leaves)
- **EPA-ng** (Barbera et al., 2019), designed for speed with large numbers of query sequnces, but can fail on trees with 10,000 or more leaves

Distance-based methods:

- **APPLES-2** (Balaban et al., 2021), one of the only methods that can place onto large backbone trees (200K sequences)

Other methods haven't been as scalable as APPLES-2 or as accurate or as accurate as pplacer/EPA-ng

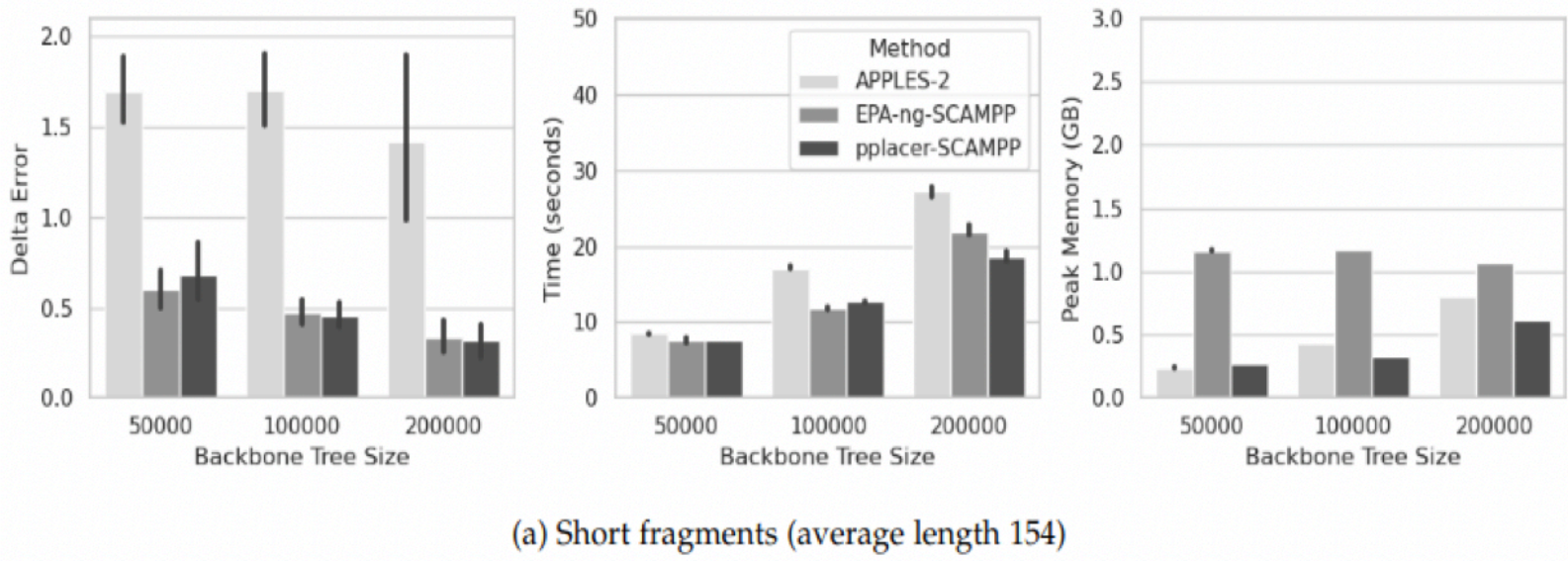
SCAMPP Framework (Wedell et al., TCBB 2022)

Used with selected phylogenetic placement method (e.g., pplacer or EPA-ng)

Input: Backbone tree with branch lengths, alignment and aligned query sequences, and a subtree size.

- ***Stage 1***- Extract placement subtree of 2000 leaves from backbone tree
- ***Stage 2*** - Use pplacer to find edge in placement subtree and location and distal length along placement edge.
- ***Stage 3*** - Find edge in backbone tree using branch lengths.

Placing short sequences: SCAMPP accuracy, scalabilty, and speed



(a) Short fragments (average length 154)

APPLES-2 has high error when placing fragmentary sequences

SCAMPP enables maximum likelihood methods to place into very large trees (200K sequences)

Runtime (per sequence!) and memory usage increases with backbone tree size

Delta-error decreases with the backbone tree size: *beneficial impact of increased taxon sampling!*

Batch-SCAMPP (WABI 2023) : Placing many short sequences

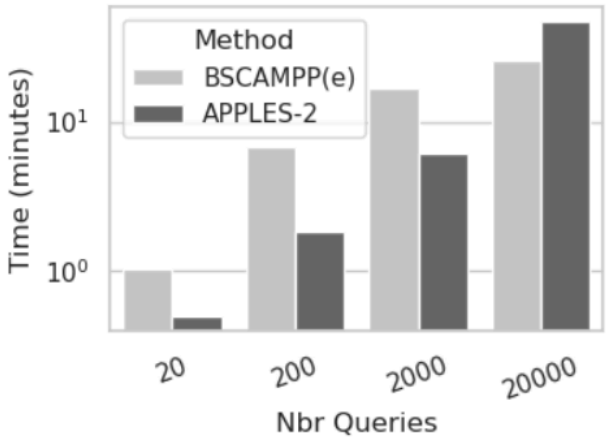
Table 3 Testing Data Results for Method Comparison on RNASim (50,000 sequences in the backbone tree with 10,000 fragmentary query sequences).

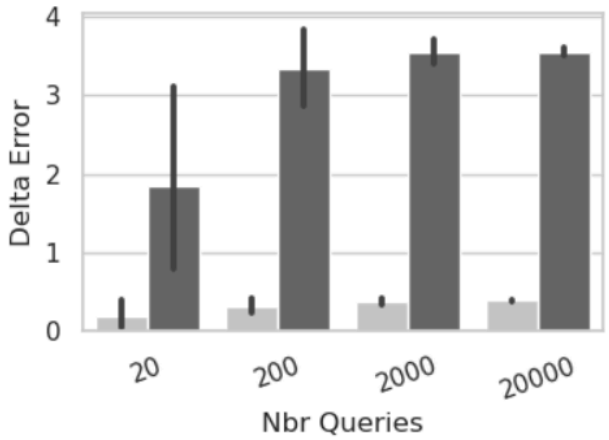
	RNASim		
Method	Delta Error	Runtime (minutes)	Memory (GB)
BSCAMPP(e)	0.50	7.2	3.0
SCAMPP(e)	0.51	466.0	1.2
SCAMPP(p)	0.46	1421.3	0.2
APPLES-2	1.52	4.8	1.1
EPA-ng	X	X	X

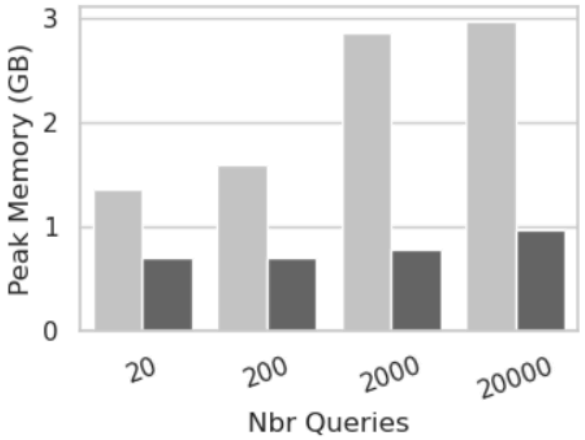
BATCH-SCAMPP (2023) is a modified version of SCAMPP that is designed for use with EPA-ng, which scales sublinearly with number of query sequences, but cannot place into large trees

Note:
- APPLES-2 is very fast (uses parallelism well) and has low memory requirement, but has much higher placement error than Batch-SCAMPP(EPA-ng)
- EPA-ng fails to run on this backbone tree

Batch-SCAMPP: Scalability with number of query sequences







APPLES-2 is very fast and has low memory requirement, but has much higher placement error than Batch-SCAMPP(EPA-ng)

BSCAMPP(EPA-ng) runtime is sublinear with number of query sequences

Part III: Large-scale maximum likelihood trees

Markov Models of Sequence Evolution

The different sites are assumed to evolve ***i.i.d***. down the model tree, so it suffices to model a single site

Jukes-Cantor, 1969 (simplest DNA site evolution model):

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

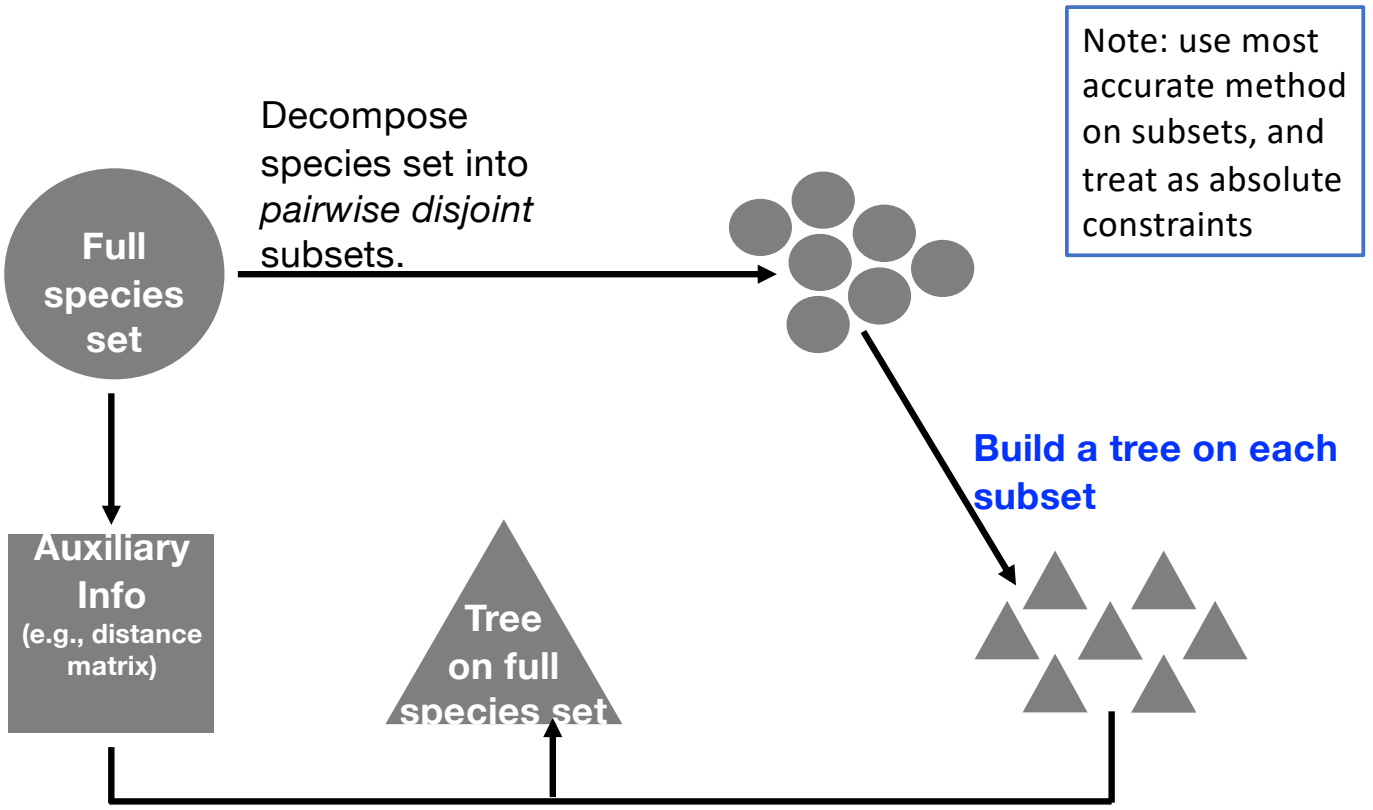
More complex models are also considered, often with little change to the theory.

Maximum likelihood for gene tree estimation

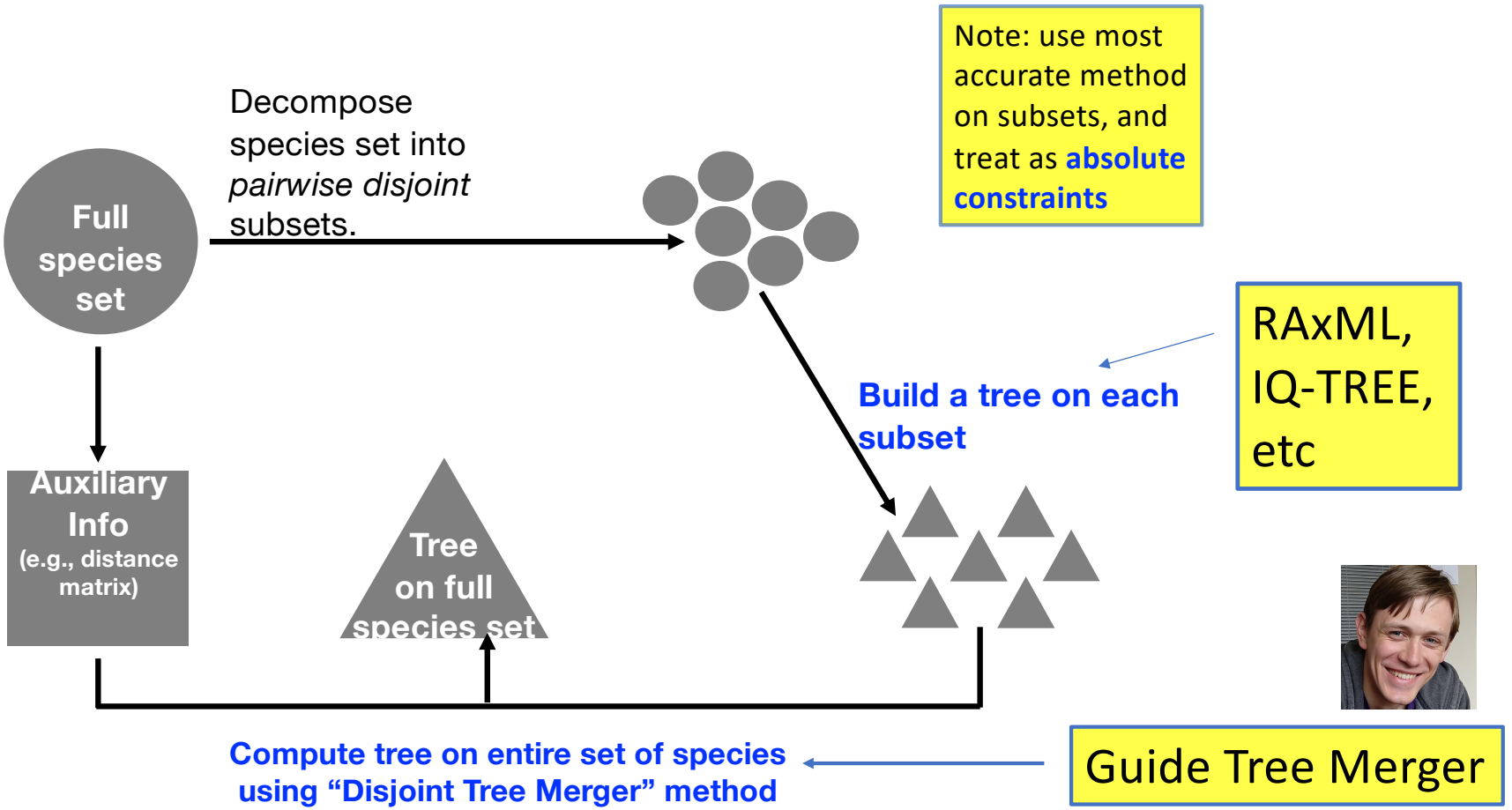
- Theory:
  - Statistically consistent
  - Low sample complexity (Roch & Sly, Prob. Theory and Related Fields, 2017): phase transition (logarithmic then polynomial)
  - NP-hard
- Empirical (based on heuristics) – using **RAxML** (leading ML heuristic)
  - Outstanding accuracy on simulated data
  - Challenging on large datasets (best methods can take CPU years or fail to run on large datasets)

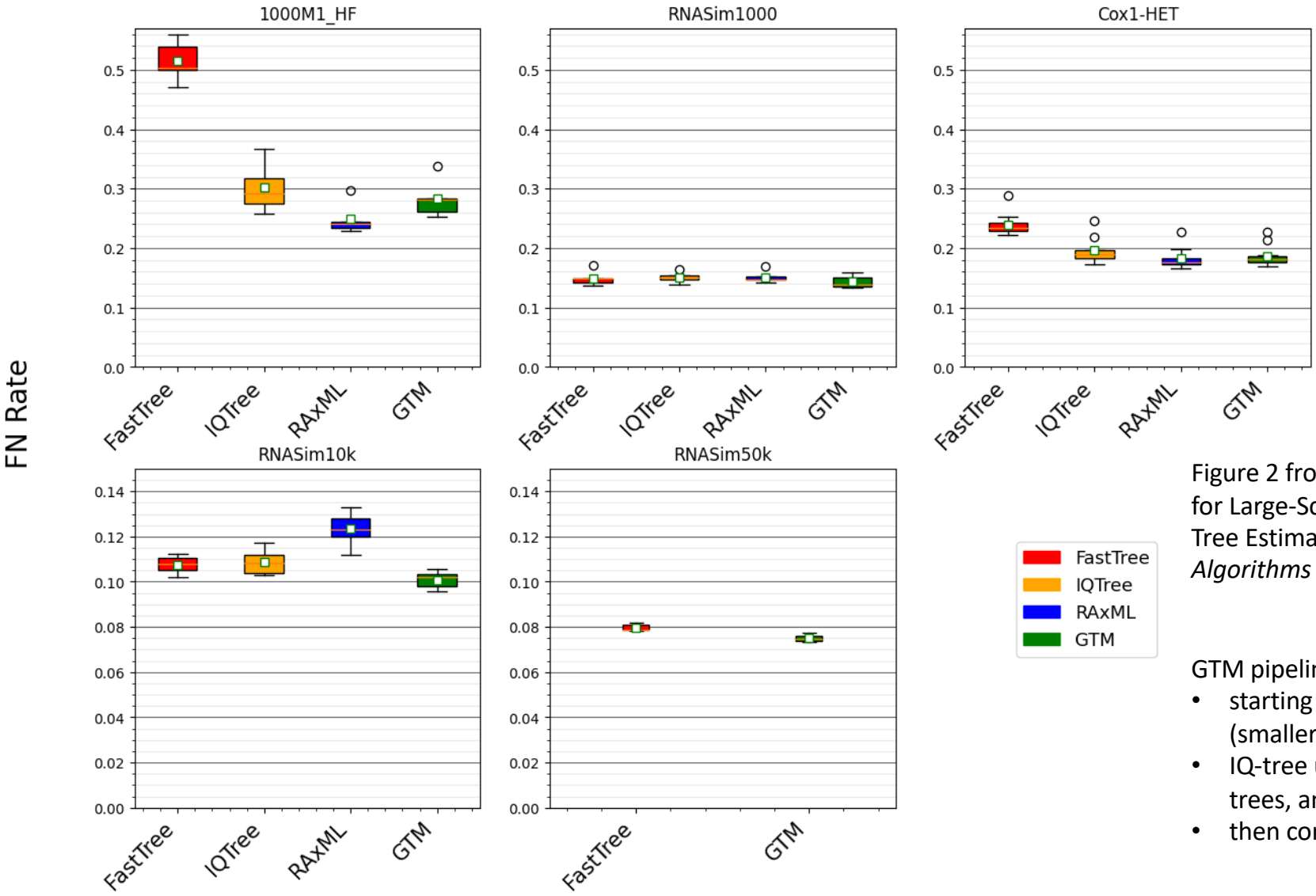
Divide-and-Conquer using Disjoint Tree Mergers

Erin Molloy, Introduced this approach



Divide-and-Conquer Gene Tree Estimation





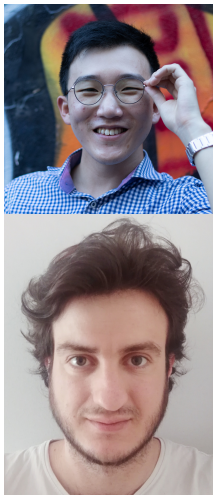
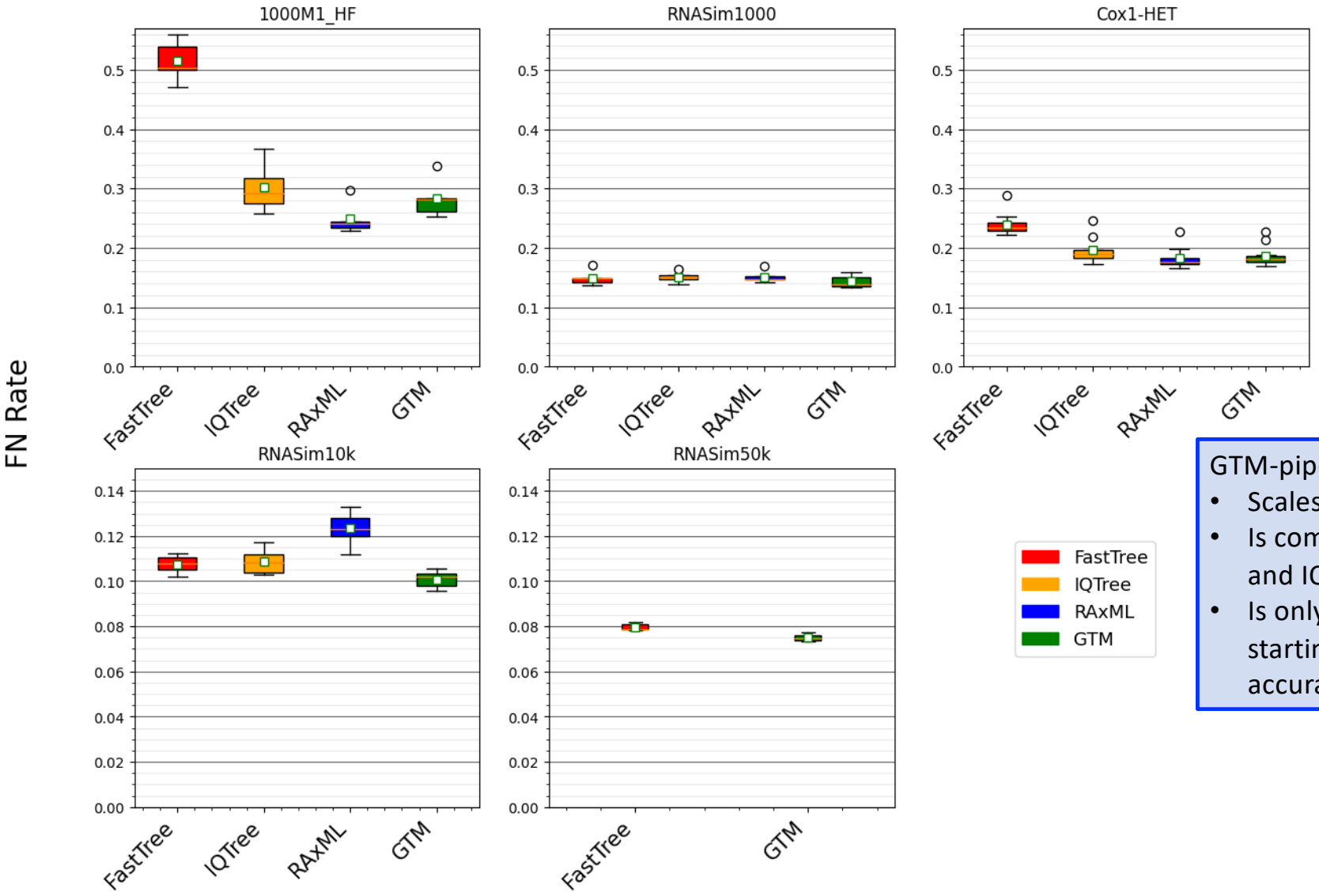
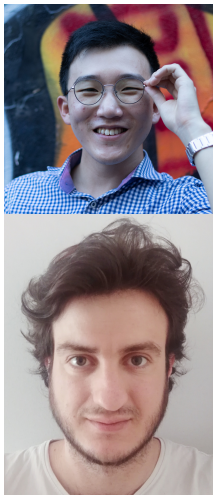


Figure 2 from "Disjoint Tree Mergers for Large-Scale Maximum Likelihood Tree Estimation", Park et al., *Algorithms 2021*

GTM pipeline:

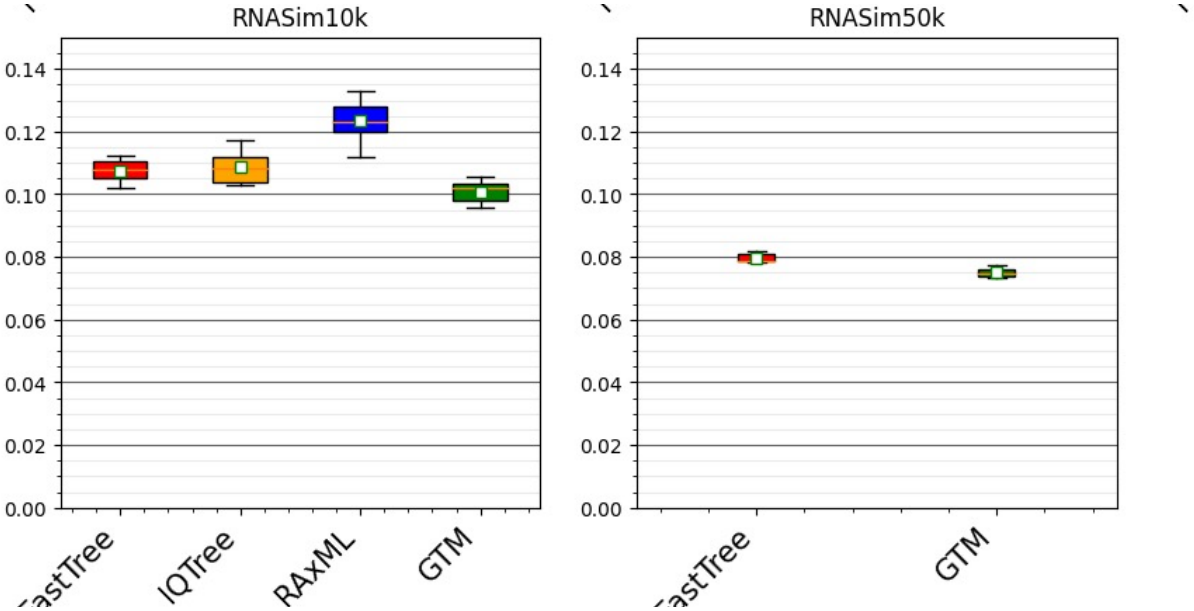
- starting tree is IQ-Tree or FastTree (smaller datasets),
- IQ-tree used to compute subset trees, and
- then combined using GTM





GTM-pipeline:

- Scales to large datasets
- Is competitive with RAxML and IQ-TREE for accuracy
- Is only slightly slower than starting tree (but more accurate)



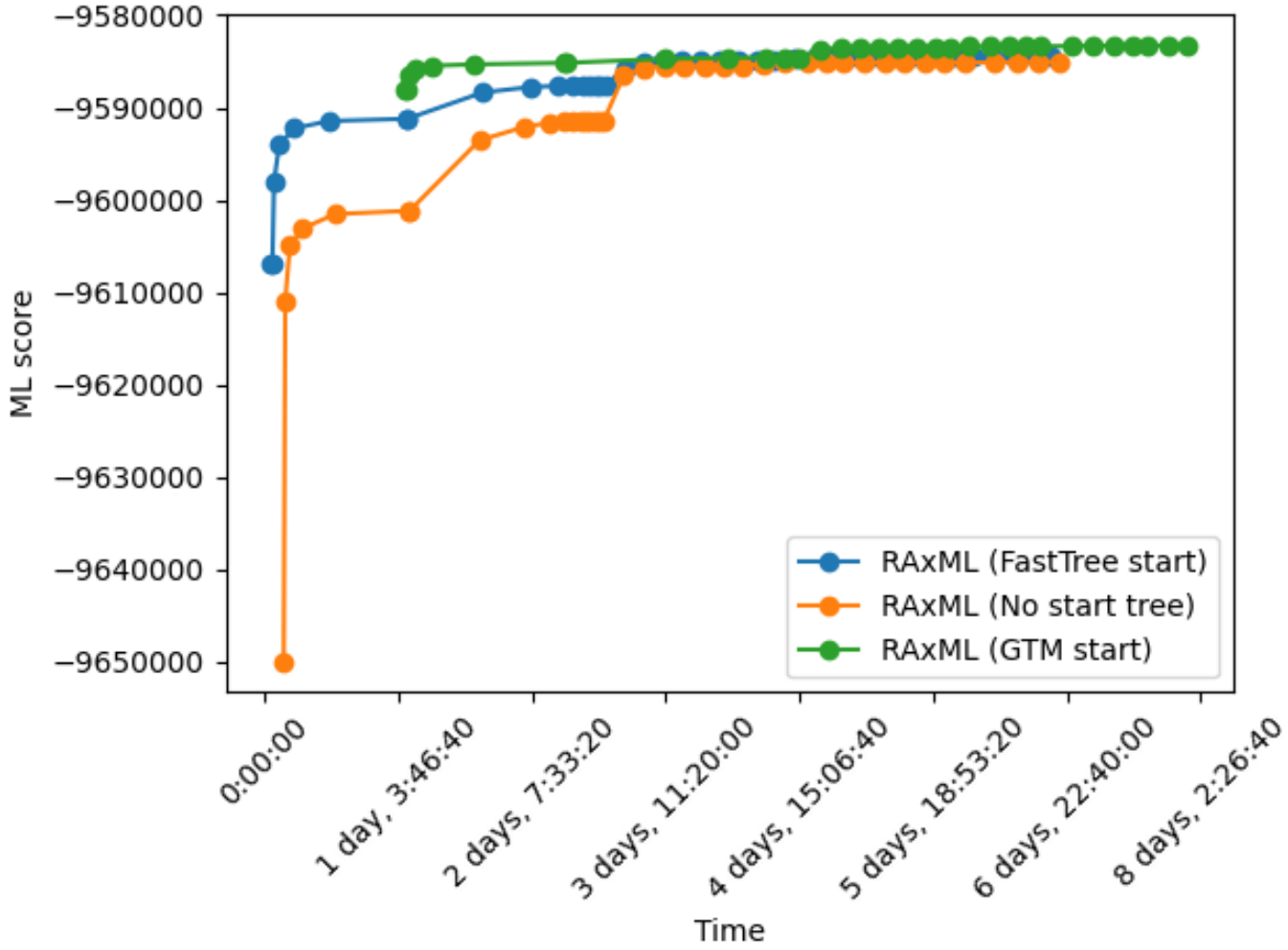
FastTree
IQTree
RAxML
GTM

Trends

- On RNASim10k: GTM most accurate topology
- On RNASim50K:
  - IQTree failed
  - RAxML had nearly 100% error
  - GTM most accurate

What about maximum likelihood score?

- We used the same technique but evaluated maximum likelihood scores on an MAGUS+EMMA alignment of the Recombinase dataset (~70,000 protein sequences) from Kelly Williams, restricting the alignment to approximately 1000 sites.
- We let RAxML run under varying conditions: its default approach, using FastTree as a starting tree, and using our GTM tree as a starting tree.
- We compared these RAxML runs (different starting trees) to each other, using LG+Gamma(4) for the model

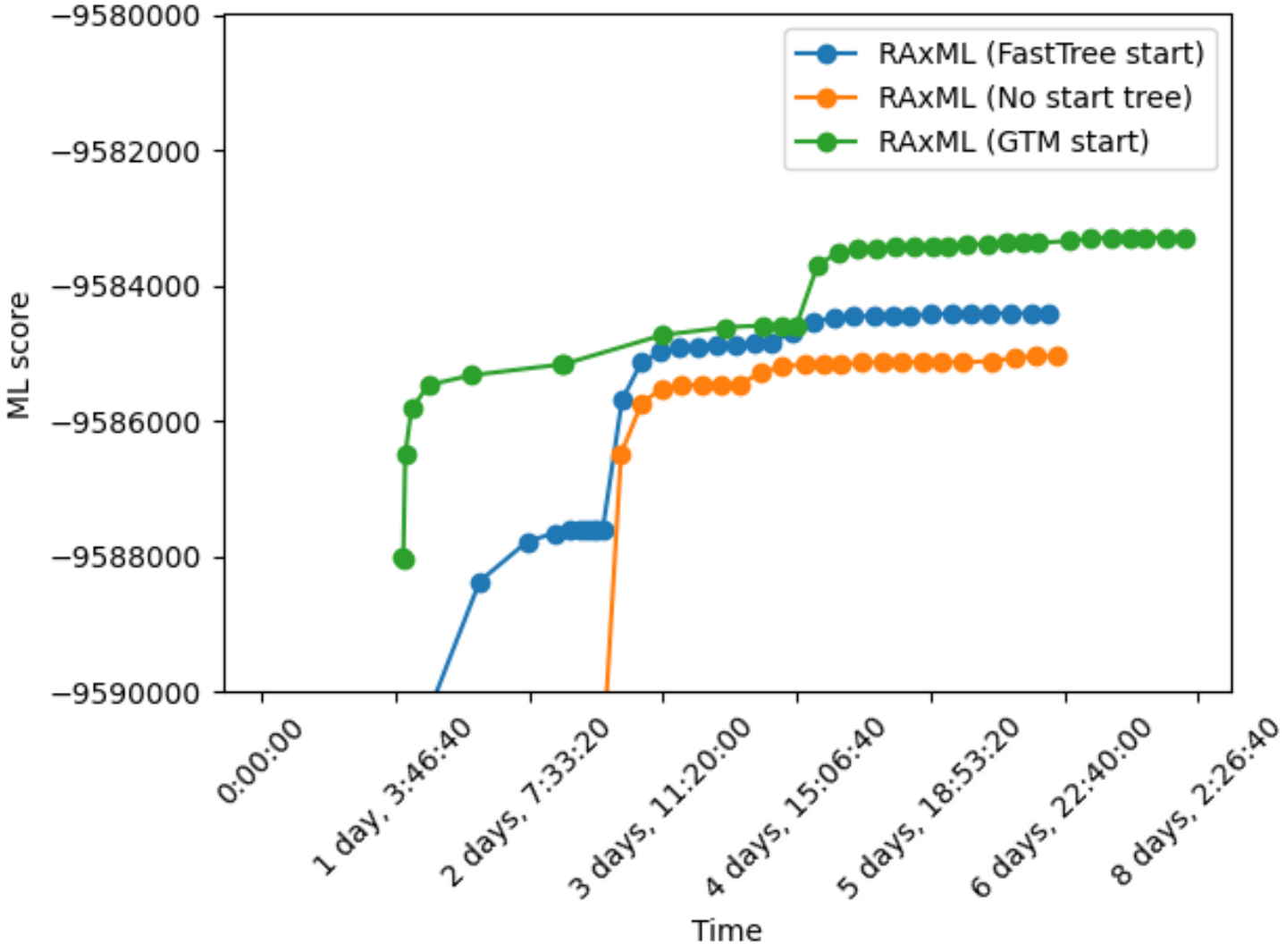


Analysis of Kelly Williams dataset (Minhyuk Park et al., NYP)

Choice of starting tree matters!

RAxML continues to improve its ML score during the entire 8 day period (but most gains are in the first 4 days)

GTM takes a bit more than 24 hours

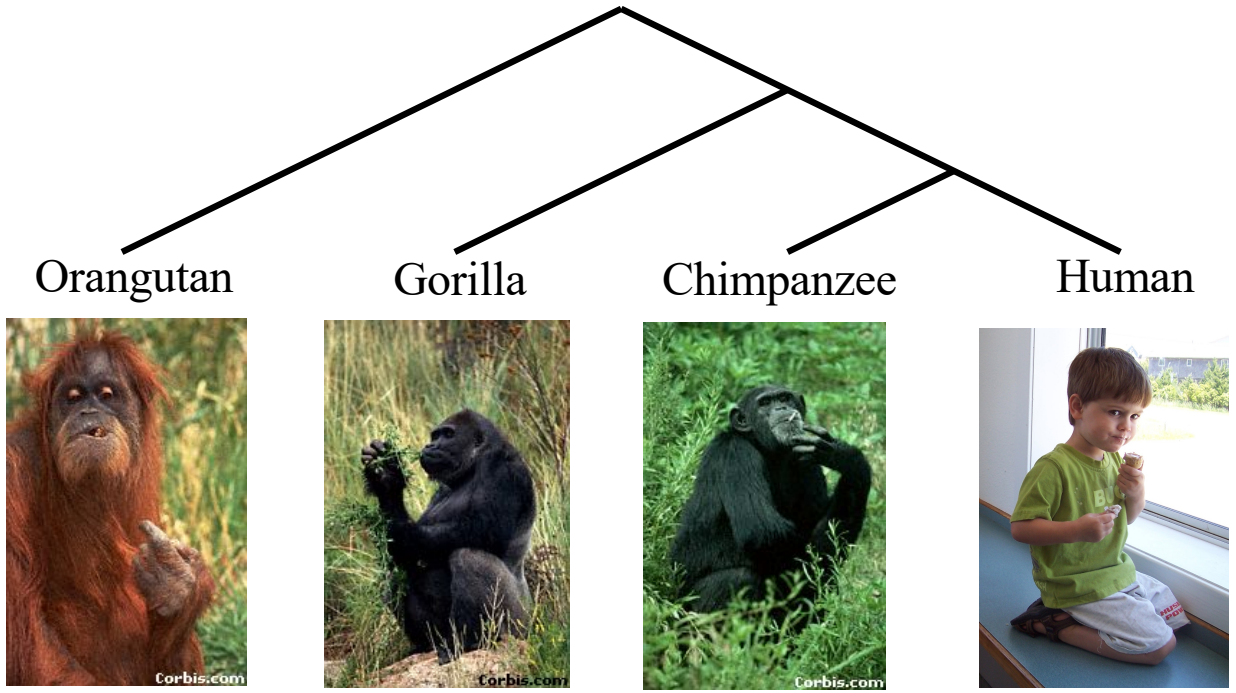


On this dataset,

- Default RAxML worst
- FastTree is a better starting tree
- GTM is much better

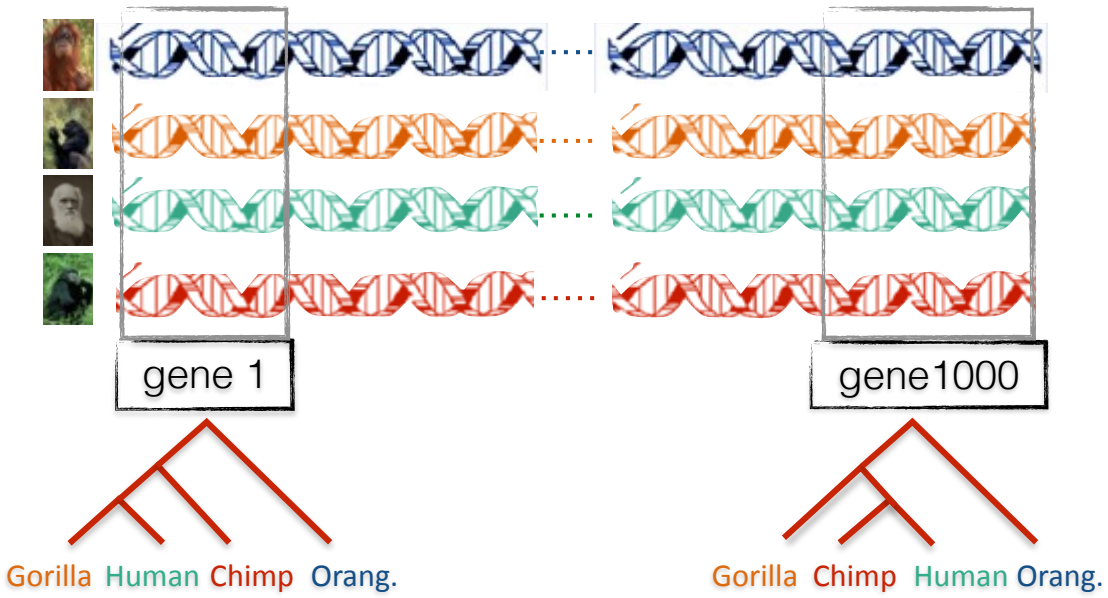
Large datasets need long running times and very good starting trees!

Part IV: Species Tree Estimation



From the Tree of the Life Website,
University of Arizona

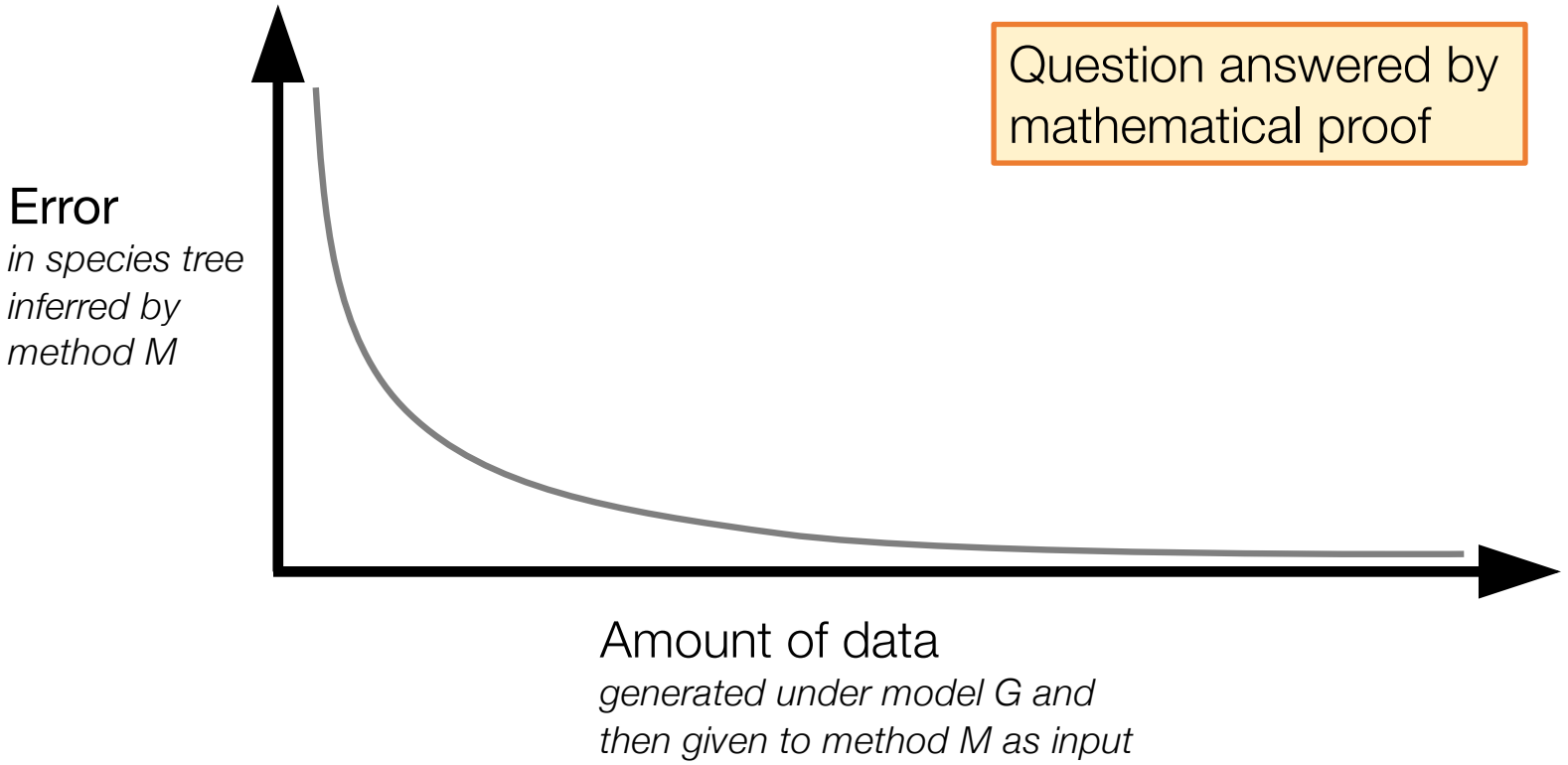
Gene tree discordance



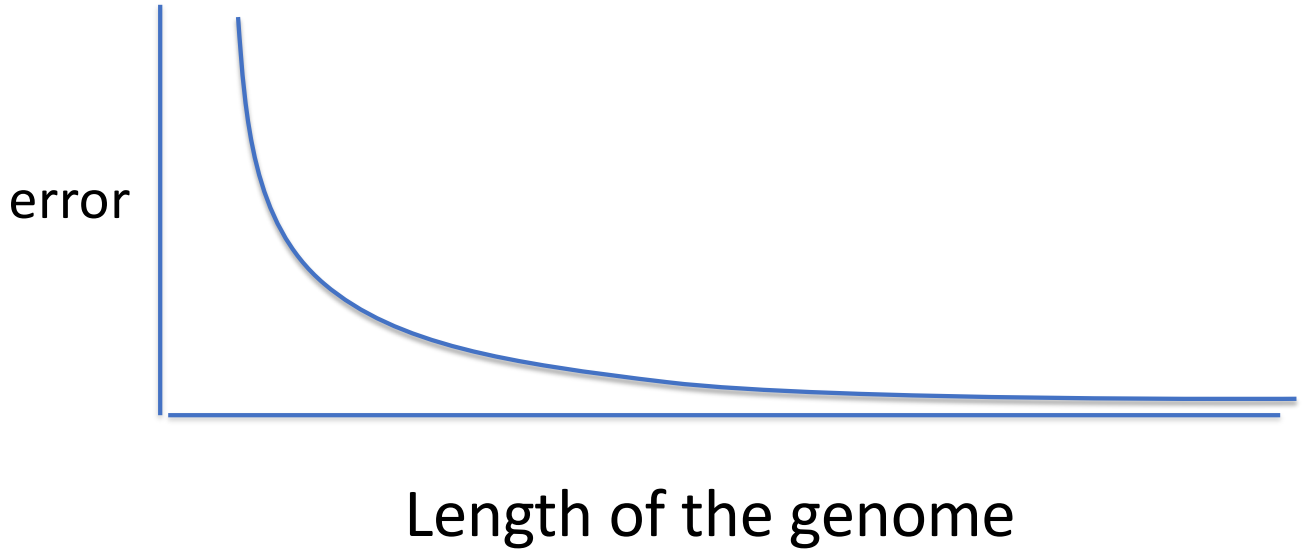
Multiple causes for discord, including

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

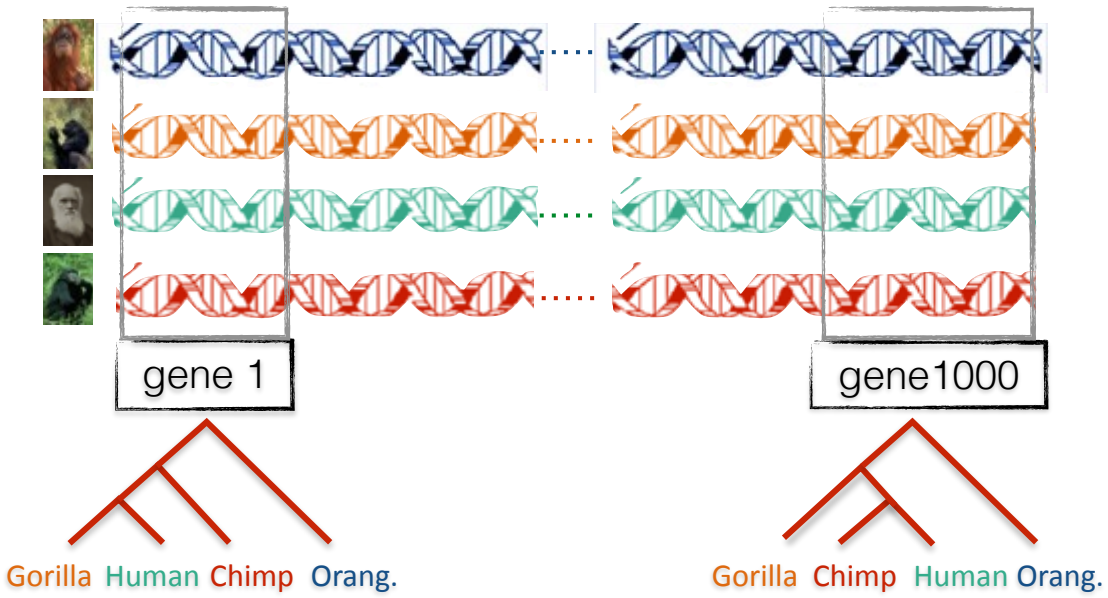
Is method M statistically consistent under model G?



Genome-scale data?



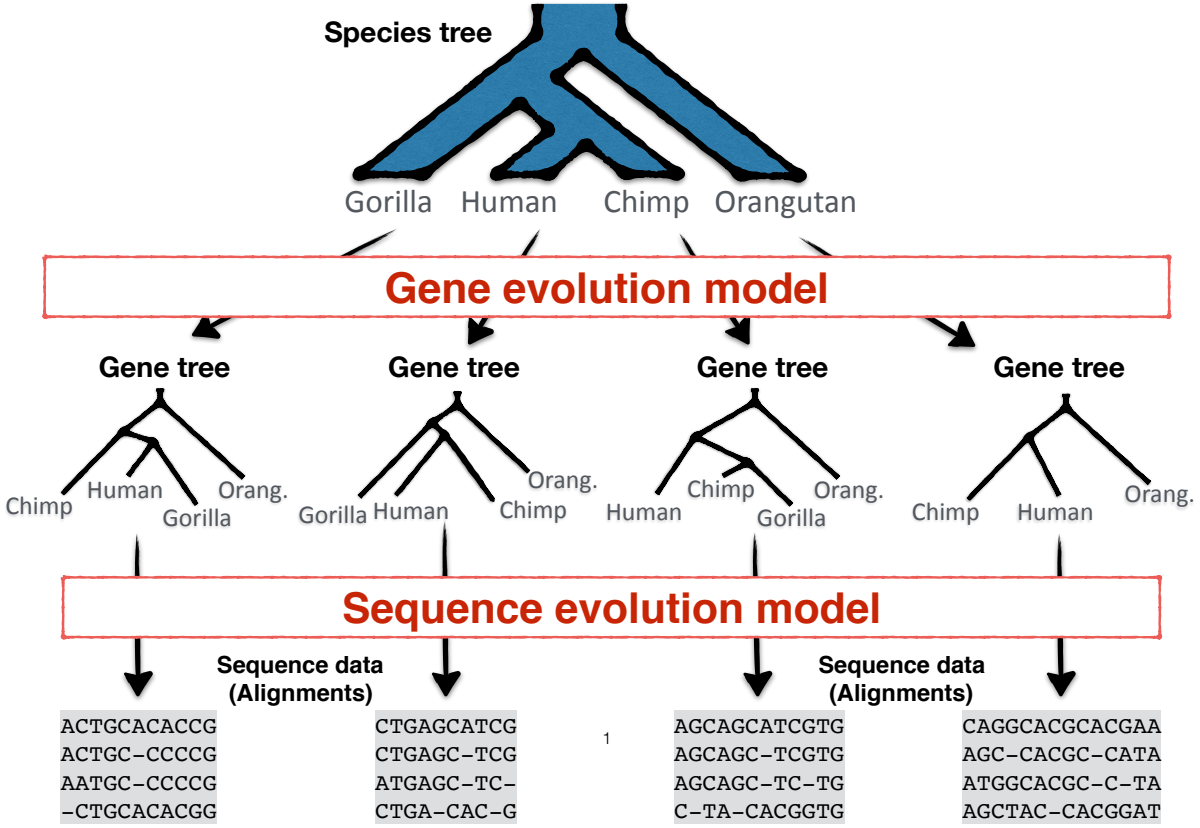
Gene tree discordance



Multiple causes for discord, including

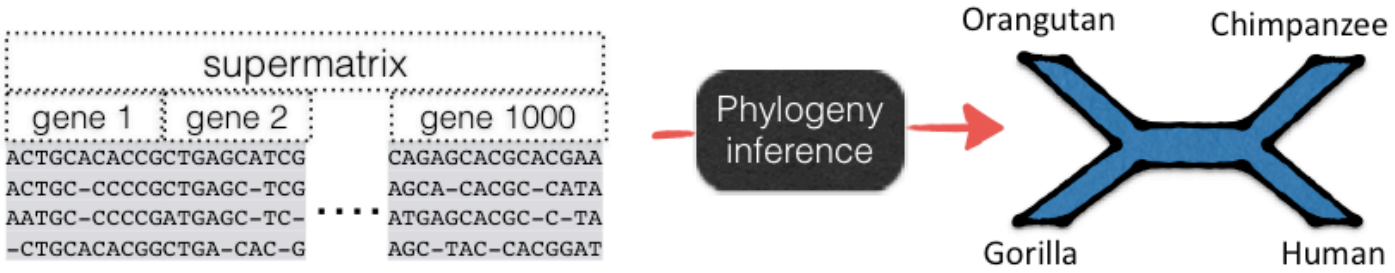
- Incomplete Lineage Sorting (ILS),
- Gene Duplication and Loss (GDL), and
- Horizontal Gene Transfer (HGT)

MSC+GTR Hierarchical Model



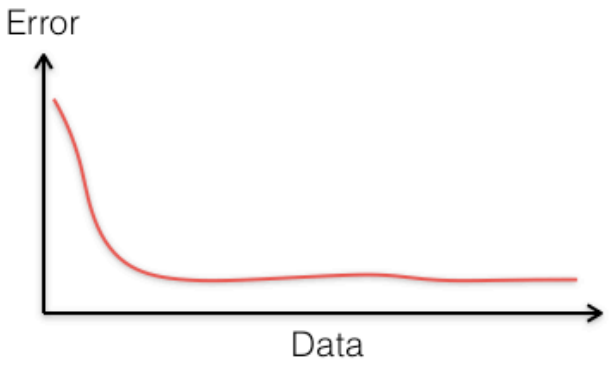
1. Gene trees evolve within the species tree (under the Multi-Species Coalescent model)
2. Sequences evolve down the gene trees (under GTR model)

Traditional approach: concatenation

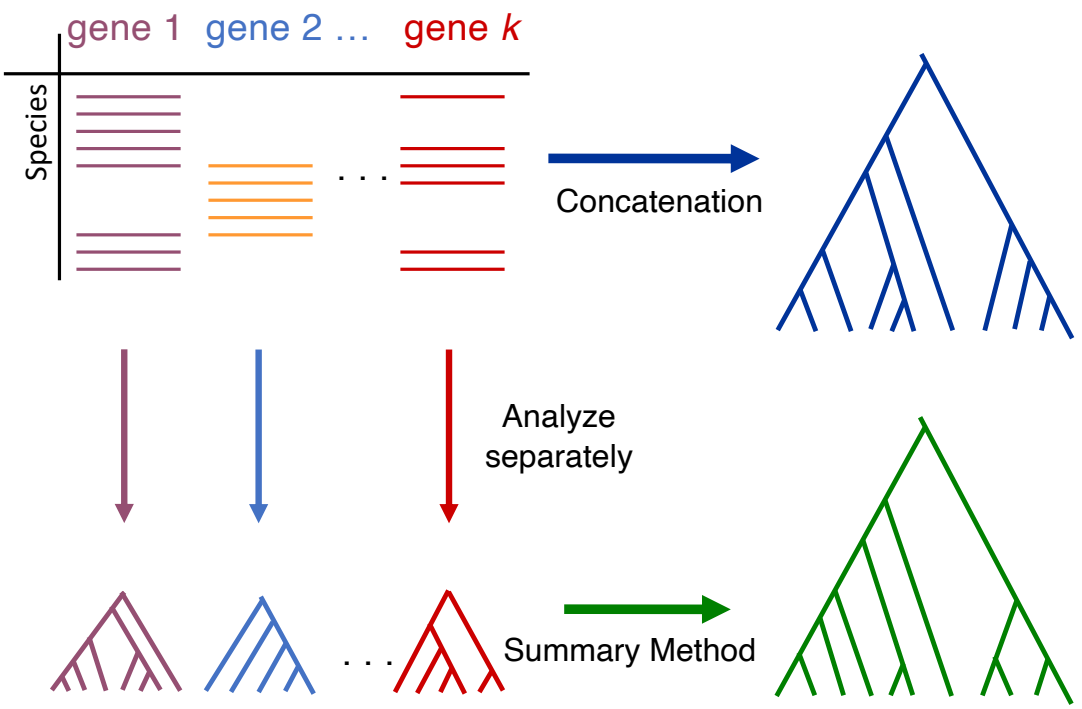


- Statistically <u>inconsistent</u> and can even be positively misleading (proved for unpartitioned maximum likelihood)
  [Roch and Steel, Theo. Pop. Gen., 2014]

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



Main Approaches for Species Tree Estimation under ILS



e.g., RAxML

e.g., ASTRAL

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

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

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 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

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
Molecular Phylogenetics and Evolution
journal homepage: www.elsevier.com/locate/ympev
ELSEVIER

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

Corrinne E. Grover[a,e], Joseph P. Gallagher[a], Josef J. Jareczek[a], Justin T. Page[b], Joshua A. Udall[c], 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]

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[2,3,4], Daniel J. Field[1,5], Jeffrey P. Townsend[1,6], Emily Moriarty Lemmon[7] & Alan R. Lemmon[8]

Gene Family Trees

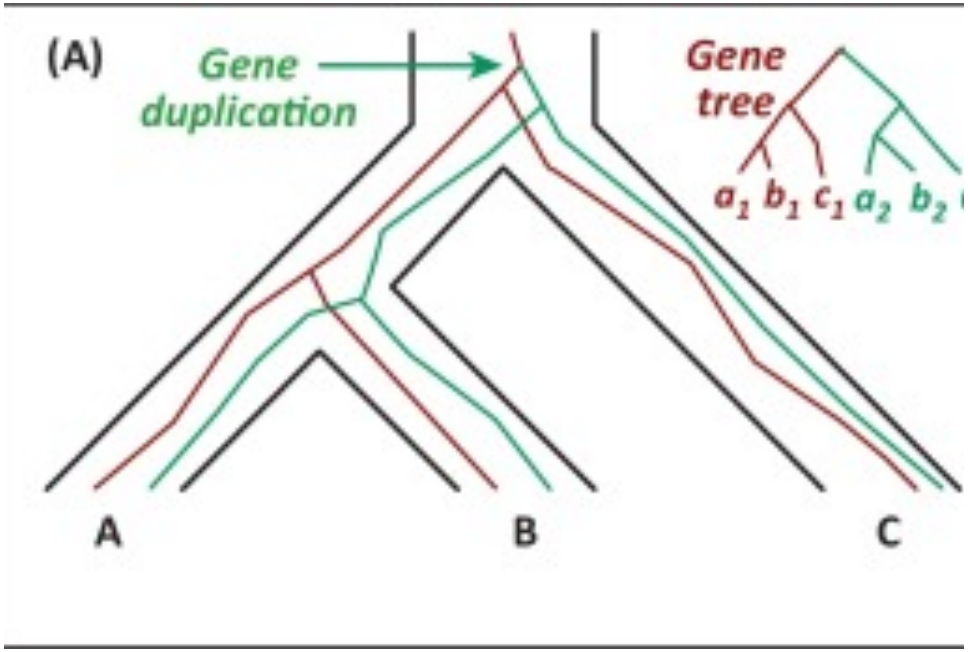


Figure by Luay Nakhleh, TREE 2013

The species tree has one duplication (at the root), which produces a gene family tree that has two copies of the species tree!

Multi-copy trees: MUL-trees

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/UIUC


S. Mirarab,
UT-Austin /UCSD

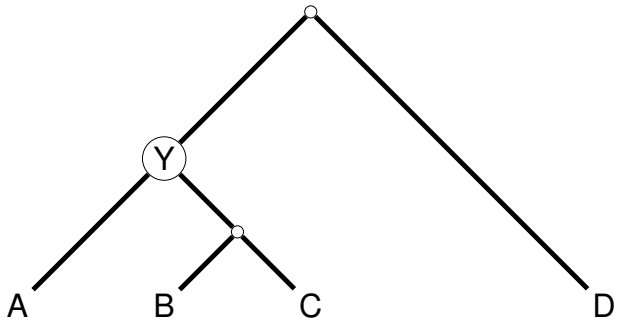

N. Nguyen
UT-Austin/UCSD

- 2014 *PNAS* study: 103 plant transcriptomes, 400-800 single copy "genes"
- 2019 *Nature* study: much larger!

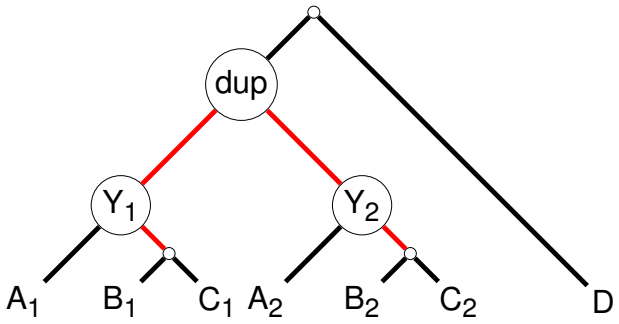
Major Challenges:

- **Multi-copy genes omitted (9500 -> 400)**
- Massive gene tree heterogeneity consistent with ILS

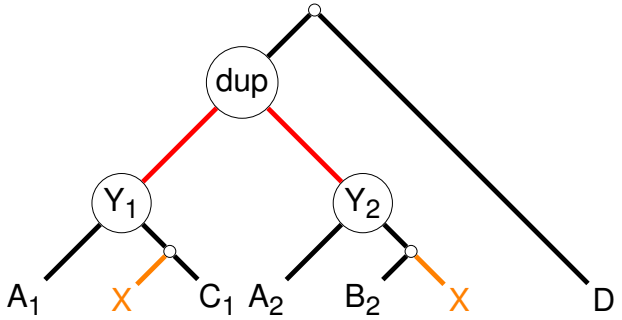
Problem: Given set of MUL-trees, infer the species tree



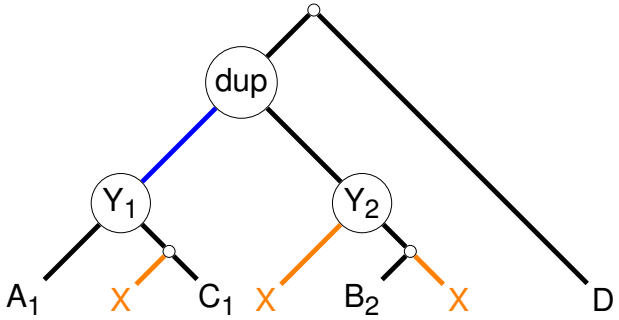
(a) Species tree T^*



(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

Species tree estimation under GDL

Options:

1. Throw out multi-copy genes
2. Figure out orthology
3. Run methods (like gene tree parsimony) that combine gene family trees into a species tree

Theorem (Legried, Molloy, Warnow, and Roch, 2019): **ASTRAL-multi** is statistically consistent under GDL and runs in polynomial time.


Theorem: Under GDL, most probable quartet tree is the species tree

ASTRAL-Pro: Estimating species trees from gene family trees

MOLECULAR BIOLOGY AND EVOLUTION

Smbe Society for Molecular Biology & Evolution

ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy

Chao Zhang, Celine Scornavacca, Erin K Molloy, Siavash Mirarab

Molecular Biology and Evolution, Volume 37, Issue 11, November 2020, Pages 3292–3307, https://doi.org/10.1093/molbev/msaa139

Published: 04 September 2020

ASTRAL-pro

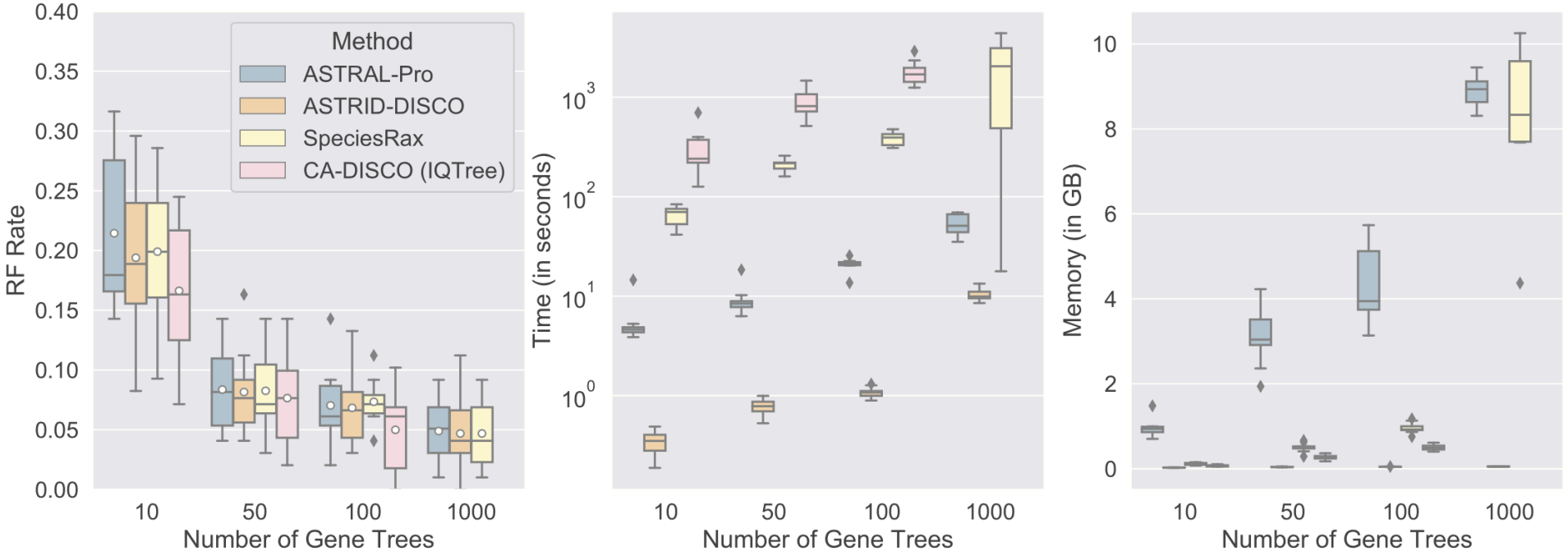
- Input: Set of unrooted multi-copy gene family trees (mul-trees)
- Output: Species tree

- Step 1: ”root and tag” every mul-tree
- Step 2: Use the rooting to define “speciation quartets”
- Step 3: Run ASTRAL’s DP algorithm with modified weights, reflecting speciation quartets

DISCO (Willson et al., Syst. Biol. 2022)

- Input: Set of gene family trees
- Output: Set of single copy gene trees (obtained by decomposing gene family trees)
- Technique:
  - Use ASTRAL-Pro to root and tag each gene family tree
  - Decompose from the “bottom-up”, aiming to keep at least one large subtree
  - Follow with method that requires single-copy genes (e.g., ASTRAL, ASTRID, Concatenation Analysis using maximum likelihood)
- Variants we examined: ASTRID-DISCO, ASTRAL-DISCO, CA-DISCO

Results on 101 species with GDL and ILS



Results on 1000 species and 1000 genes

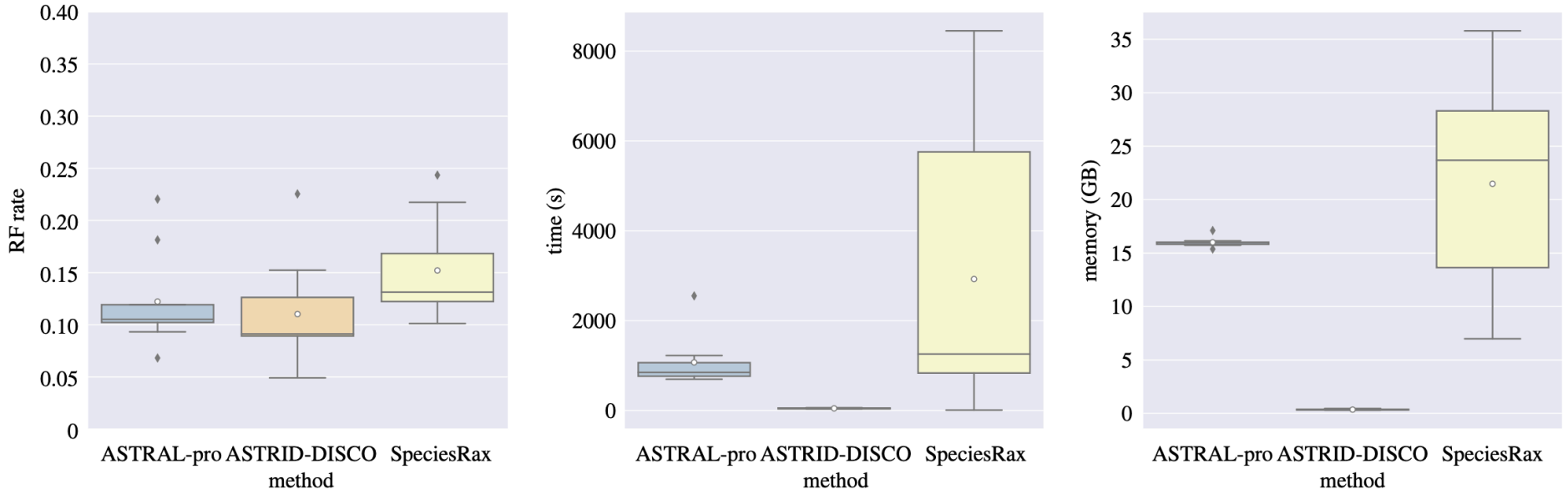
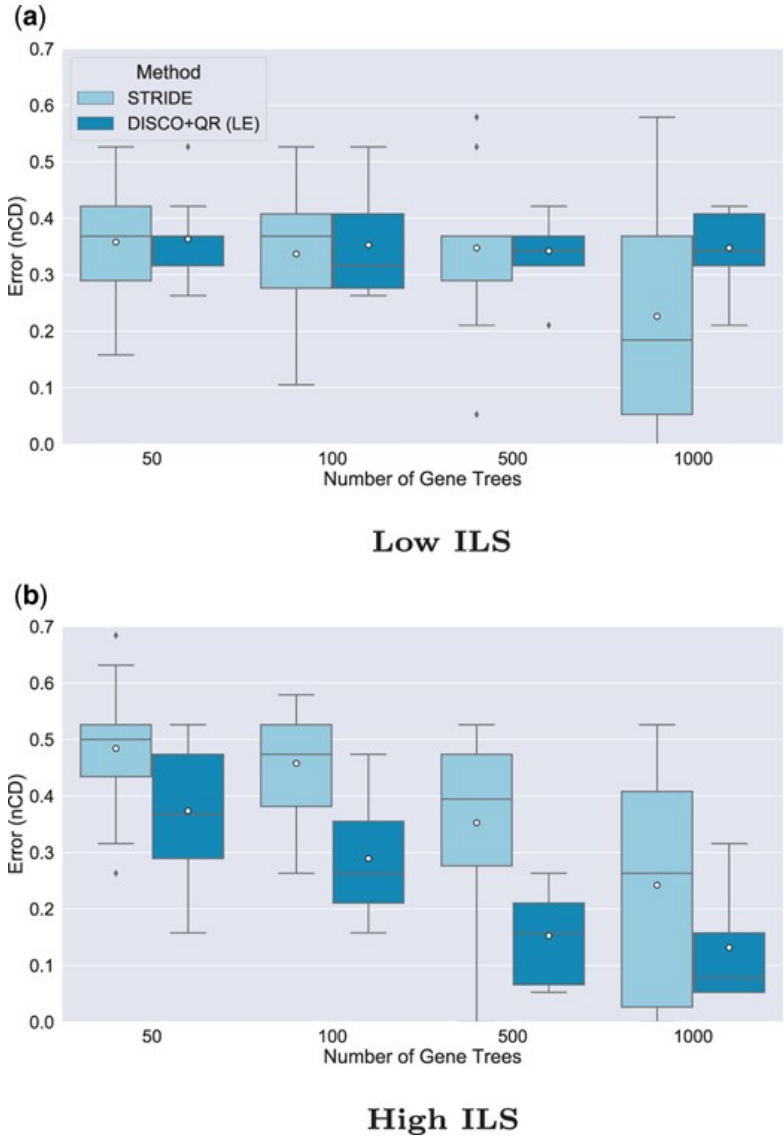


Figure 4. Species tree error (Robinson–Foulds (RF) error rates), wall clock running time (s) and peak memory usage of ASTRAL-Pro, ASTRID-DISCO and SpeciesRax

Rooting species trees

- QR-STAR (Tabatabaee et al.) is a statistically consistent method for rooting species trees when ILS is present, and uses the unrooted gene trees

- DISCO+QR-STAR (Willson et al.) combines DISCO and QR-STAR to root species trees when ILS and GDL are present, and uses the unrooted gene trees



Summary for species tree estimation

- If ILS but no GDL, then ASTRAL, ASTRID, and concatenation are all good (choice depends on data).
  - Not shown: FASTRAL and GTM can speed up ASTRAL
- If GDL as well, then ASTRID-DISCO or ASTRAL-Pro are good summary methods, and CA-DISCO (CA=RAxML on concatenated alignment) is excellent if runtime permits.
  - Note: no need to determine orthology – can use all your data!
- For rooting species trees:
  - If ILS but no GDL, then QR-STAR
  - If ILS and GDL, then STRIDE is good if ILS is low enough; otherwise, DISCO+QR is good

New software for phylogenomics

- New MSA methods: MAGUS, WITCH, WITCH-ng, HMMerge, EMMA
  - Some can be used to add sequences into alignments
  - MAGUS excellent if low sequence length heterogeneity
- New phylogenetic placement methods
  - SCAMPP and BATCH-SCAMPP
- New maximum likelihood gene tree estimation:
  - GTM pipeline (divide-and-conquer)
- New species tree estimation methods:
  - ASTRAL (for species trees under ILS)
  - ASTRAL-Pro (for species trees under GDL and ILS)
  - ASTRID-DISCO and CA-DISCO (for species trees under GDL and ILS)
- Species tree rooting methods: QR-STAR

Overall summary

- Large-scale phylogenetic tree estimation is becoming truly feasible!
  - Large numbers of sequences no longer a major impediment
  - Heterogeneity across the genome presents challenges, but methods are being developed that address biological heterogeneity
- Not discussed here (and still needs work):
  - Phylogenetic networks
  - Genome rearrangement phylogeny
  - Multiple whole genome alignment

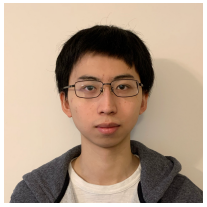
Acknowledgments










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

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

Software on github, links at http://tandy.cs.illinois.edu/software.html

Funding: NSF (CCF 1535977, ABI-1458652, 2006069, Graduate Fellowship to Erin Molloy), the Grainger Foundation, the Ira and Debra Cohen Fellowship to Vlad Smirnov and James Willson, and Sandia National Laboratories-Livermore (LDRD)

Supercomputers: Blue Waters and Campus Cluster, both supported by NCSA

Write to me: warnow@Illinois.edu