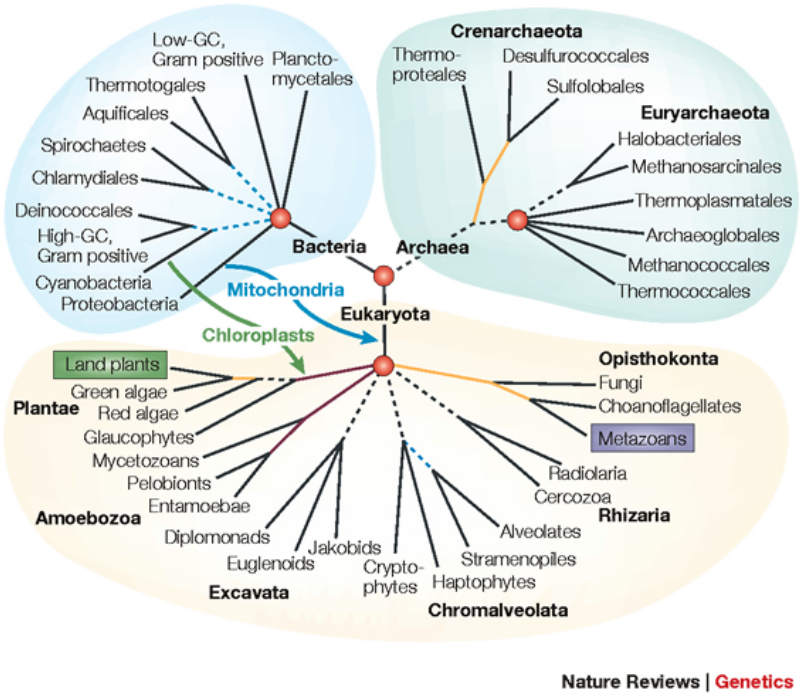
Using Disjoint Tree Mergers for Large Tree Estimation

Tandy Warnow

Siebel School of Computing and Data Science

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

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

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**

What I hope to convince you of:

- “Disjoint tree mergers” (DTMs) are generic methods, that can be used with any phylogeny estimation method (for any kind of data), and enable scalability to large datasets.
- The Guide Tree Merger (GTM) is the current leading DTM technique, based on empirical performance.
  - GTM improves maximum likelihood gene tree estimation and also species tree estimation.
  - However, GTM does NOT allow blending, and so should be able to be improved.

This talk

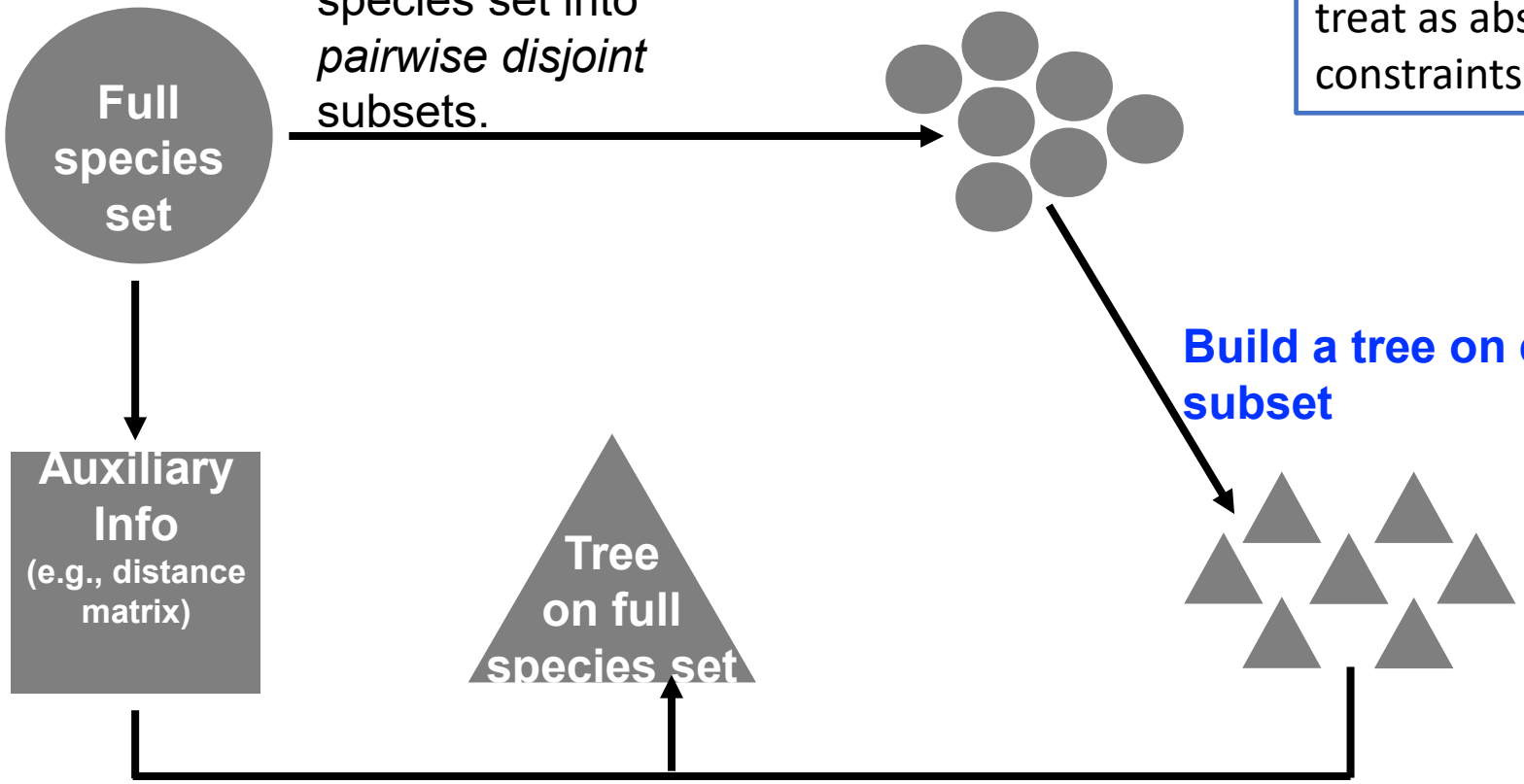
- Part I: Divide-and-conquer methods and Disjoint Tree Mergers
- Part II: Application to species tree estimation (e.g., ASTRAL and concatenation)
- Part III: Application to large-scale maximum likelihood tree estimation
- Part IV: Discussion

Part II: Disjoint Tree Mergers

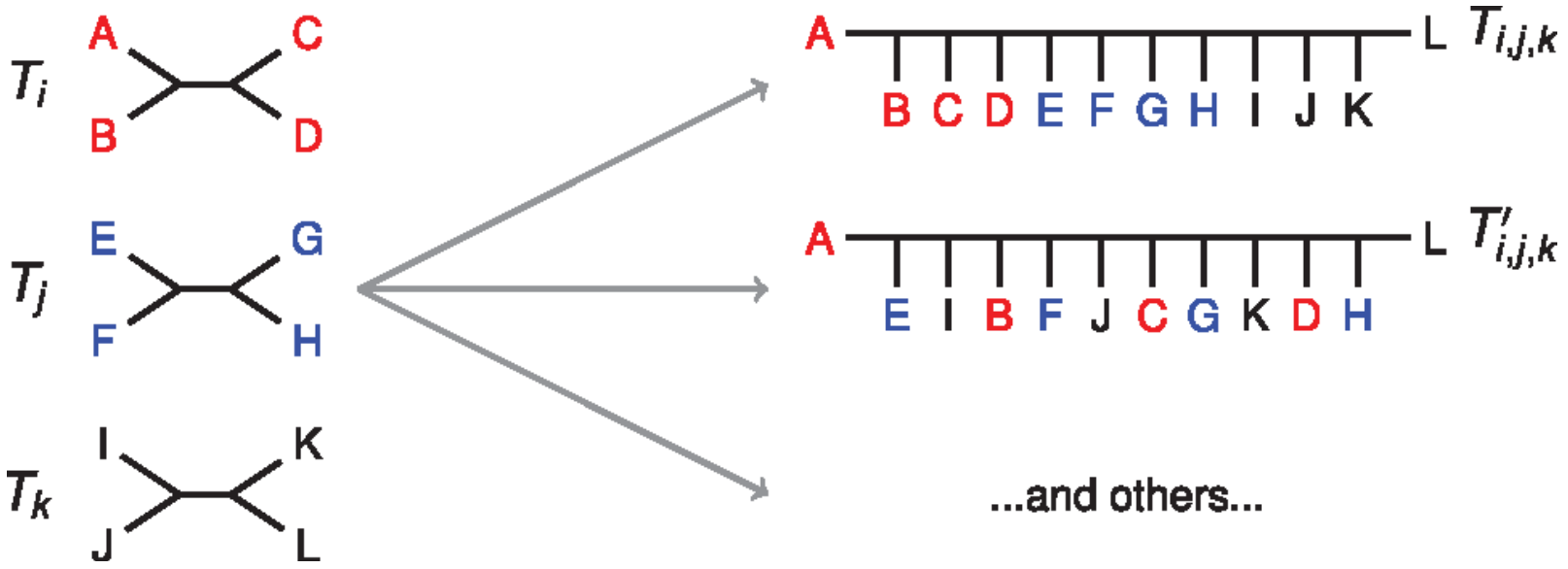
Divide-and-Conquer using Disjoint Tree Mergers

Erin Molloy, Introduced this approach

Note: use most accurate method on subsets, and treat as absolute constraints



DTMs Merge Subset Trees



Notes:

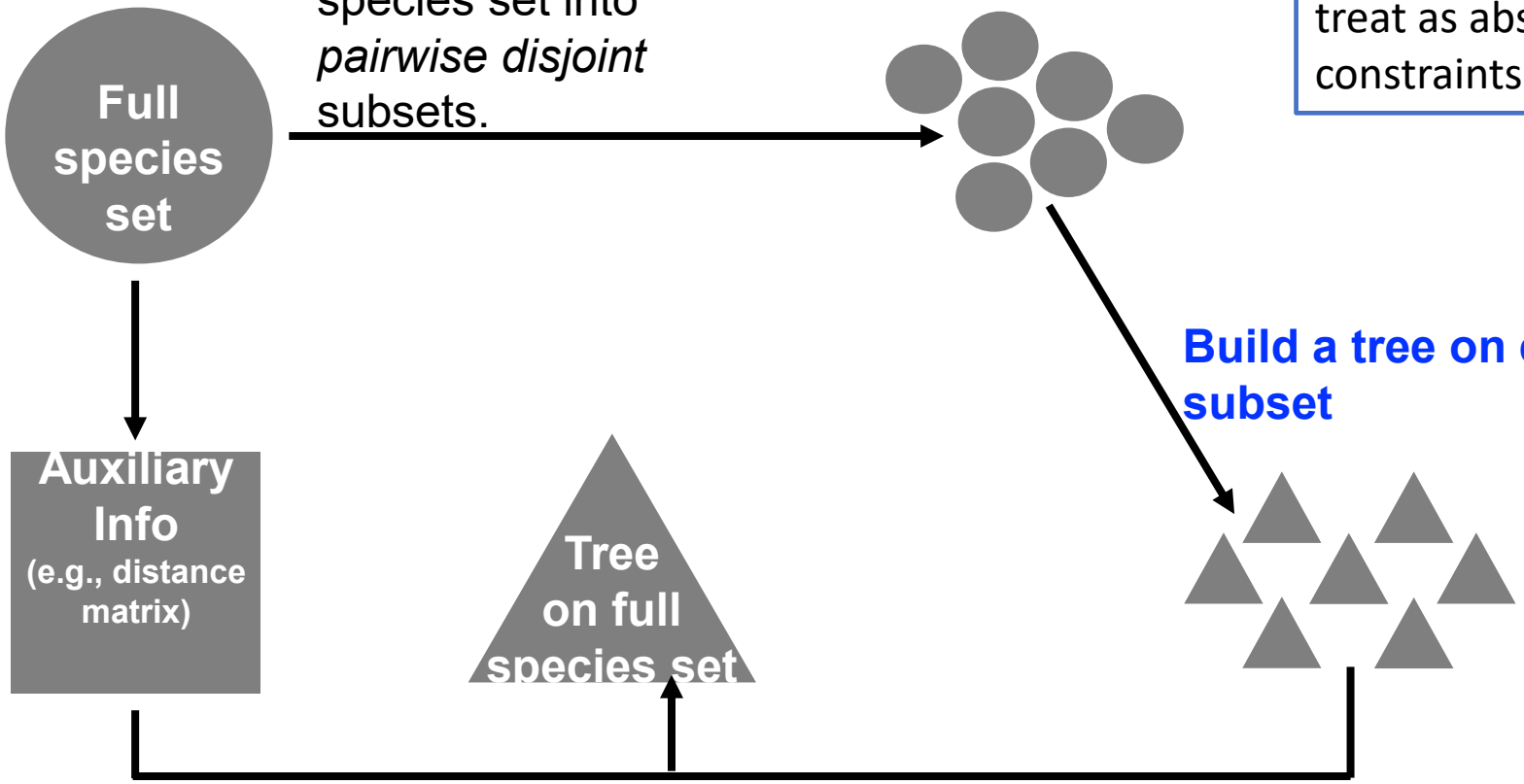
- Subset trees are requirements (constraint trees)
- Blending is permitted!

Bioinformatics, Volume 35, Issue 14, July 2019, Pages i417–i426, https://doi.org/10.1093/bioinformatics/btz344

Divide-and-Conquer using Disjoint Tree Mergers

Erin Molloy, Introduced this approach

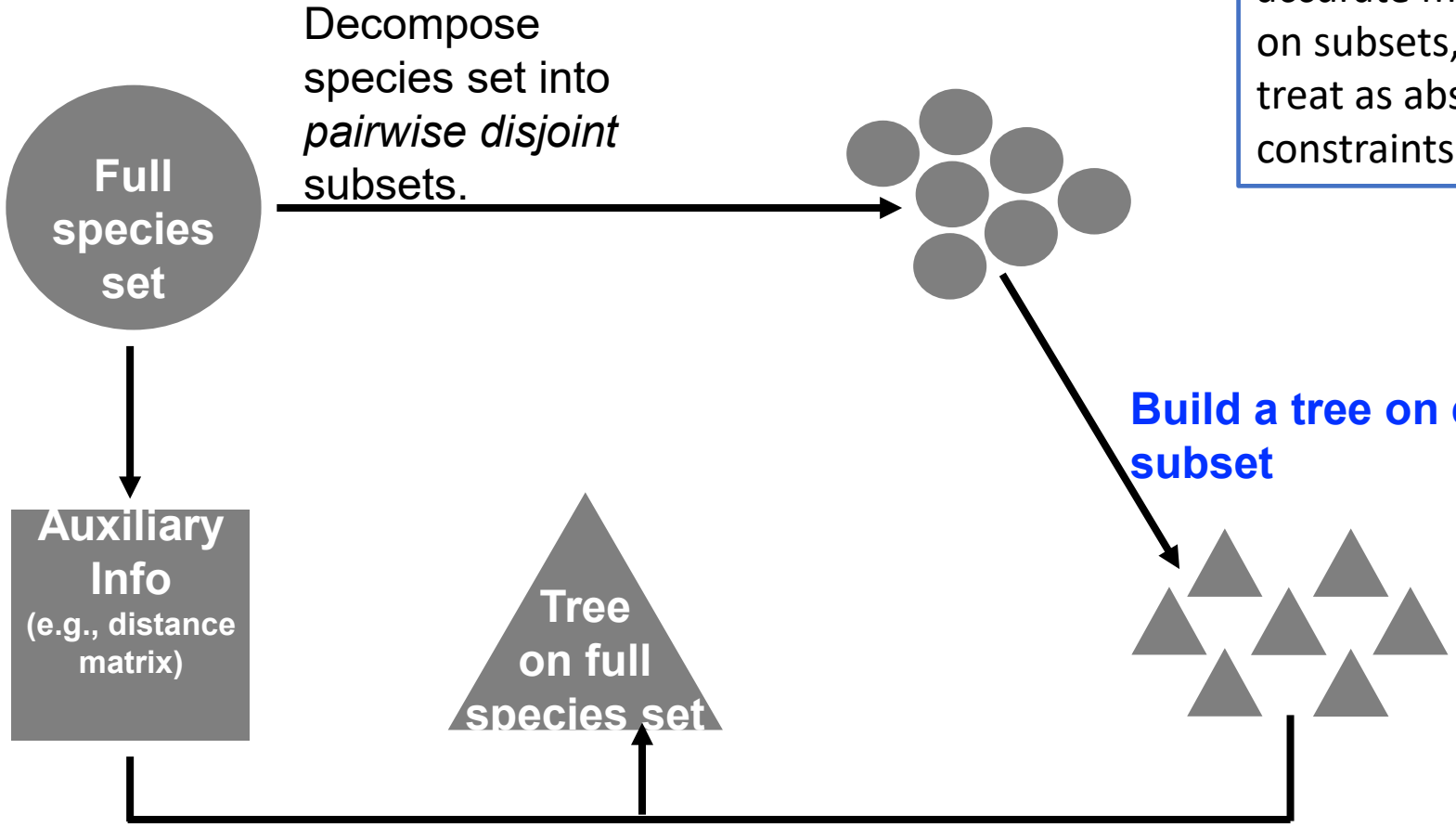
Note: use most accurate method on subsets, and treat as absolute constraints



Divide-and-Conquer using Disjoint Tree Mergers

Erin Molloy, Introduced this approach

Note: use most accurate method on subsets, and treat as absolute constraints



NJMerge

- Uses distance matrix for auxiliary info.
- Computes constraint trees on subsets
- Builds tree using agglomerative technique from NJ, as long as constraint trees not violated
- Statistically consistent

Disjoint Tree Mergers (DTMs)

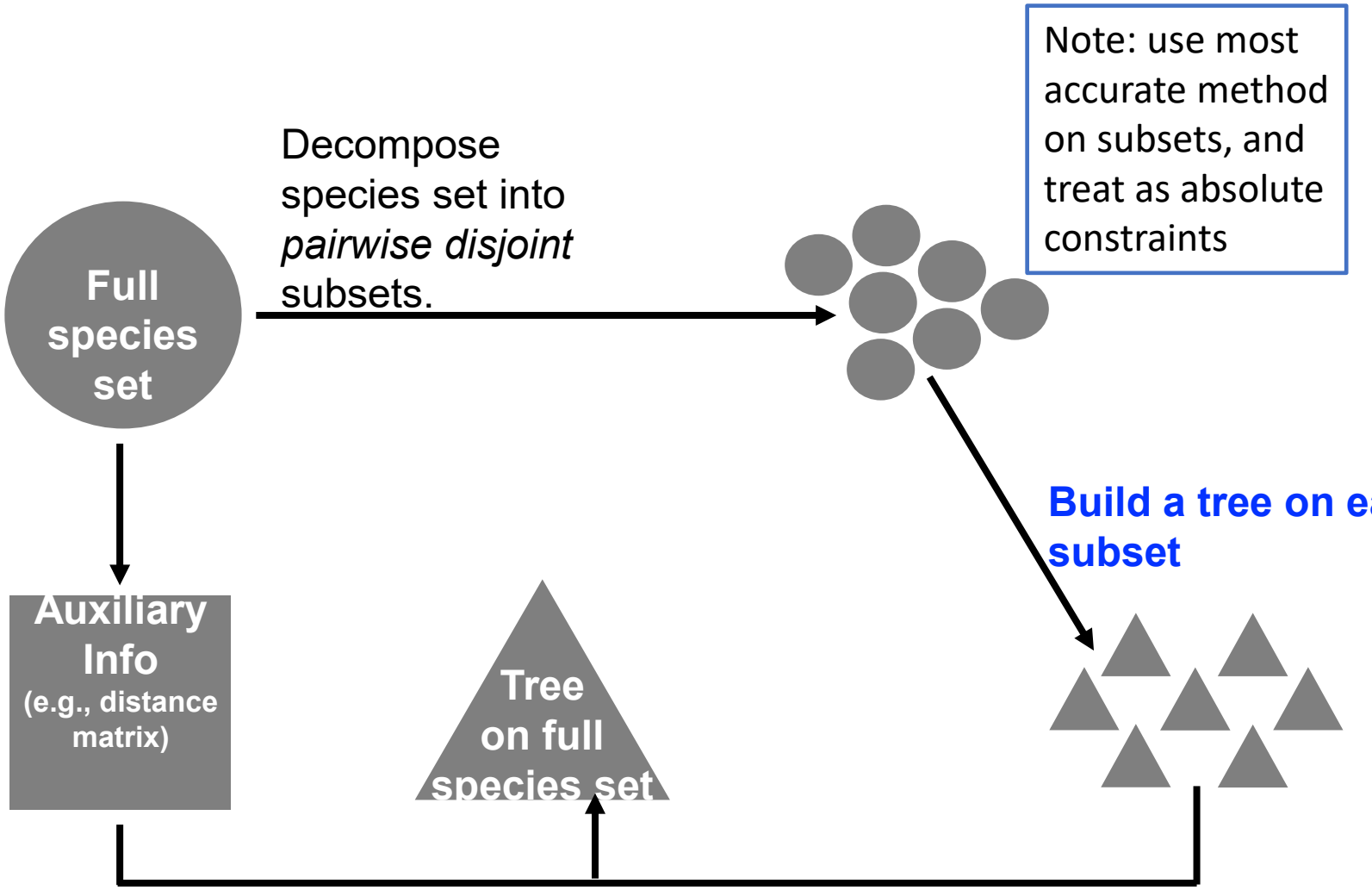
- NJMerge (Molloy and Warnow, Alg Mol Biol 2019)
- TreeMerge (Molloy and Warnow, Bioinf 2019)
- Constrained-INC (Zhang, Rao, and Warnow, Alg Mol Biol 2019)
- Guide Tree Merger (Smirnov and Warnow, BMC Genomics 2020)

Guide Tree Merger

- Input:
  - set *T* of trees T_i on leafset S_i (disjoint sets)
  - "guide tree" T on union of S_i
- Output: Tree T* that induces each T_i and minimizes the bipartition distance to T

- NP-hard
- If we constrain T* to be formed by adding edges between the trees T_i (i.e., no blending allowed), then solvable in polynomial time.
- Smirnov and Warnow, BMC Genomics 2020

Statistically consistent pipelines are easy to design!

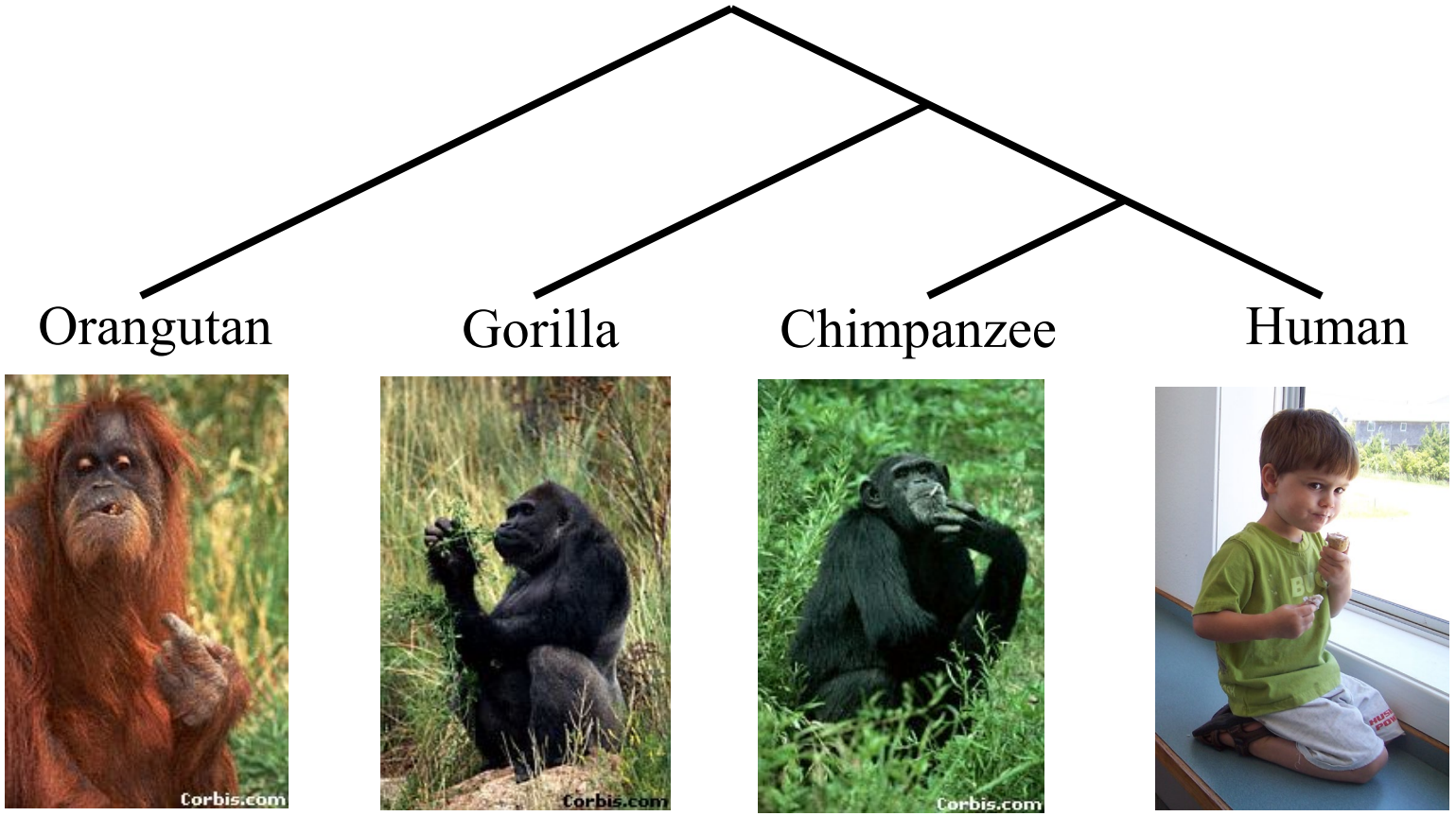


Statistically consistent pipelines are easy with any of the four DTMs, provided:

- **Statistically consistent auxiliary information**
- **Statistically consistent subtree calculation**

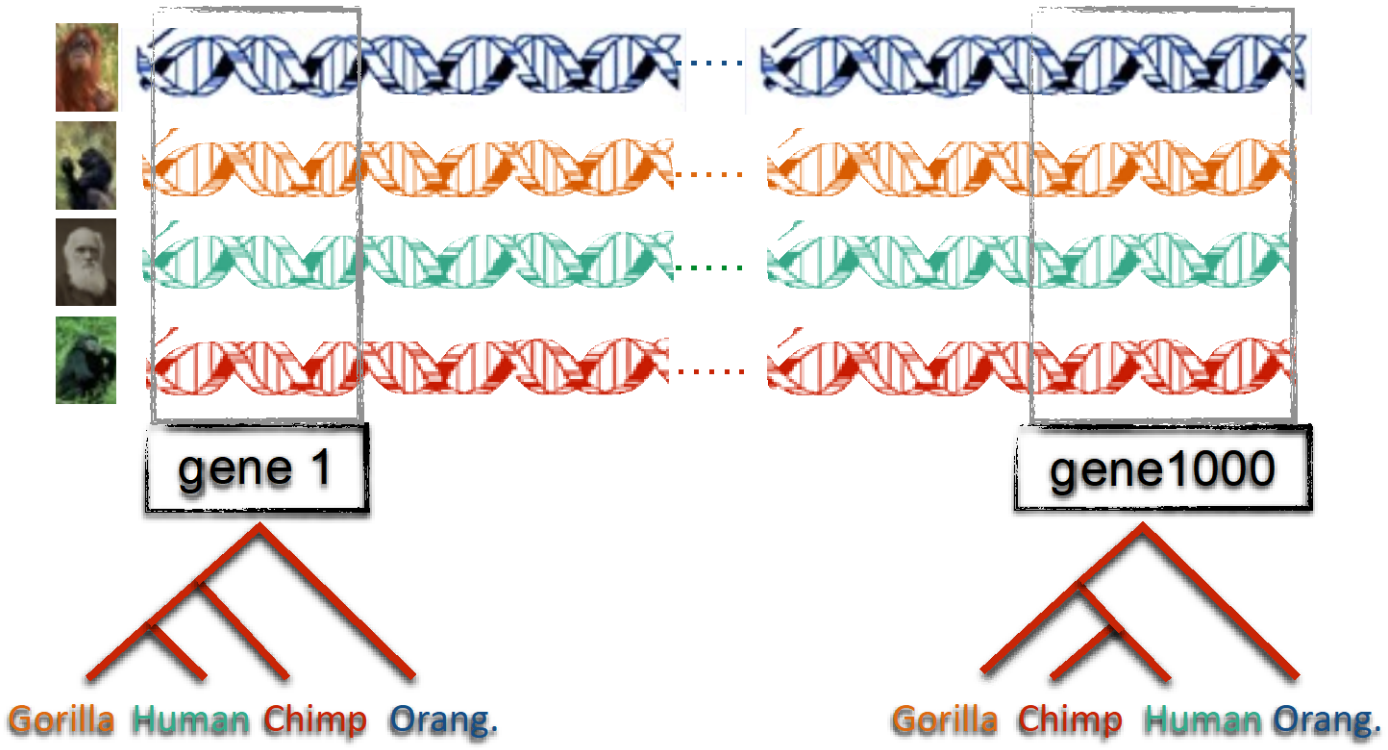
Part II: DTMs and Species Tree Estimation

Species Tree Estimation



From the Tree of the Life Website,
University of Arizona

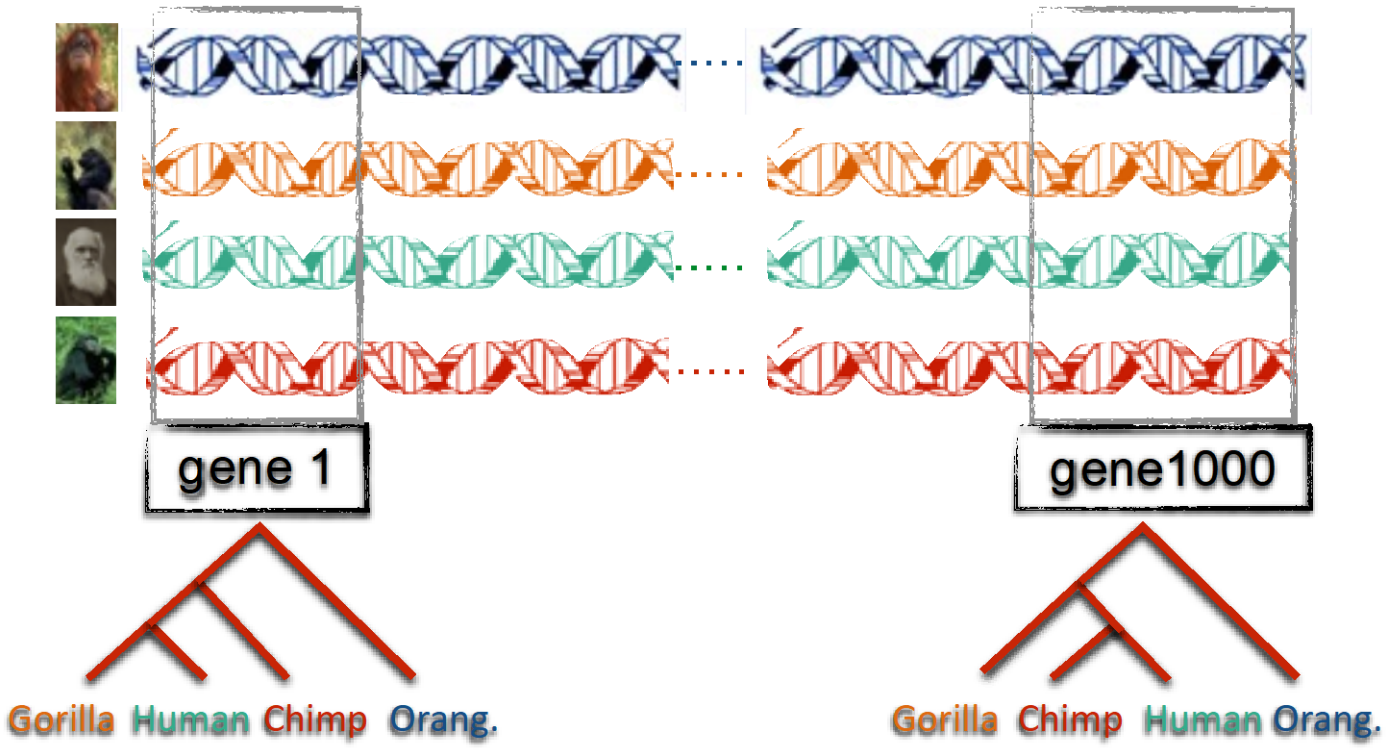
Gene tree discordance



Multiple causes for discord, including

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

Gene tree discordance

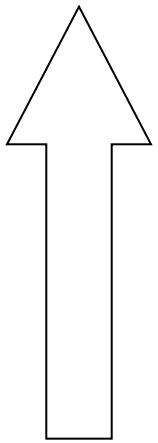


Multiple causes for discord, including

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

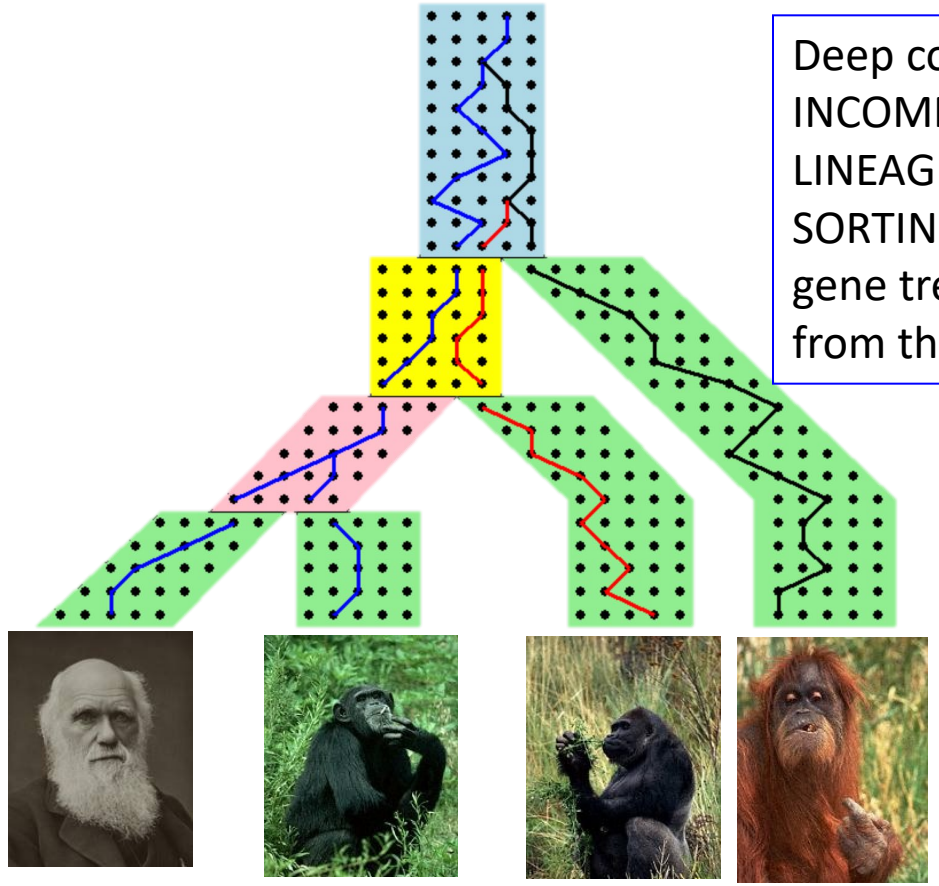
Gene trees inside the species tree (Coalescent Process)

Past



Present

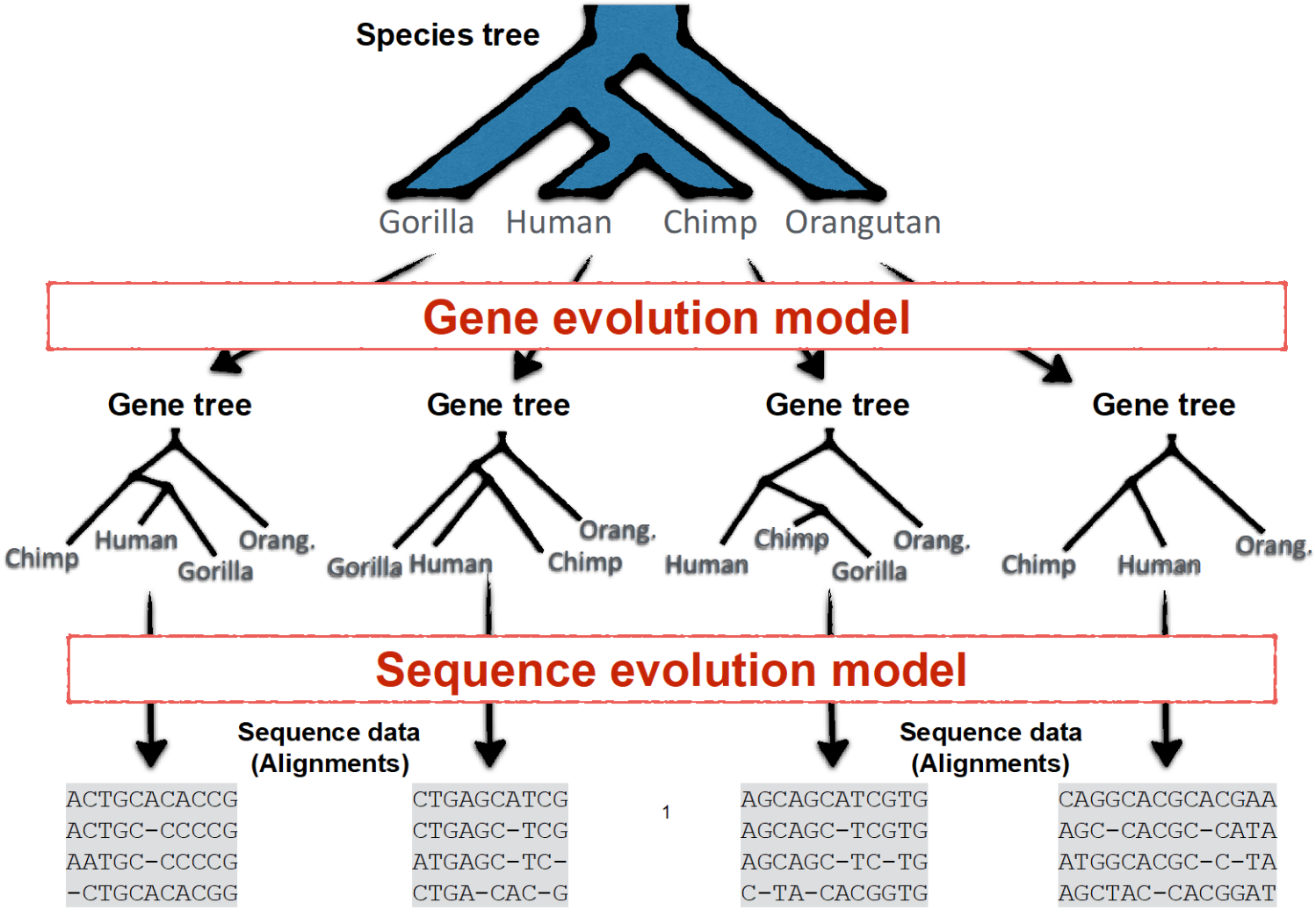
Courtesy James Degnan



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

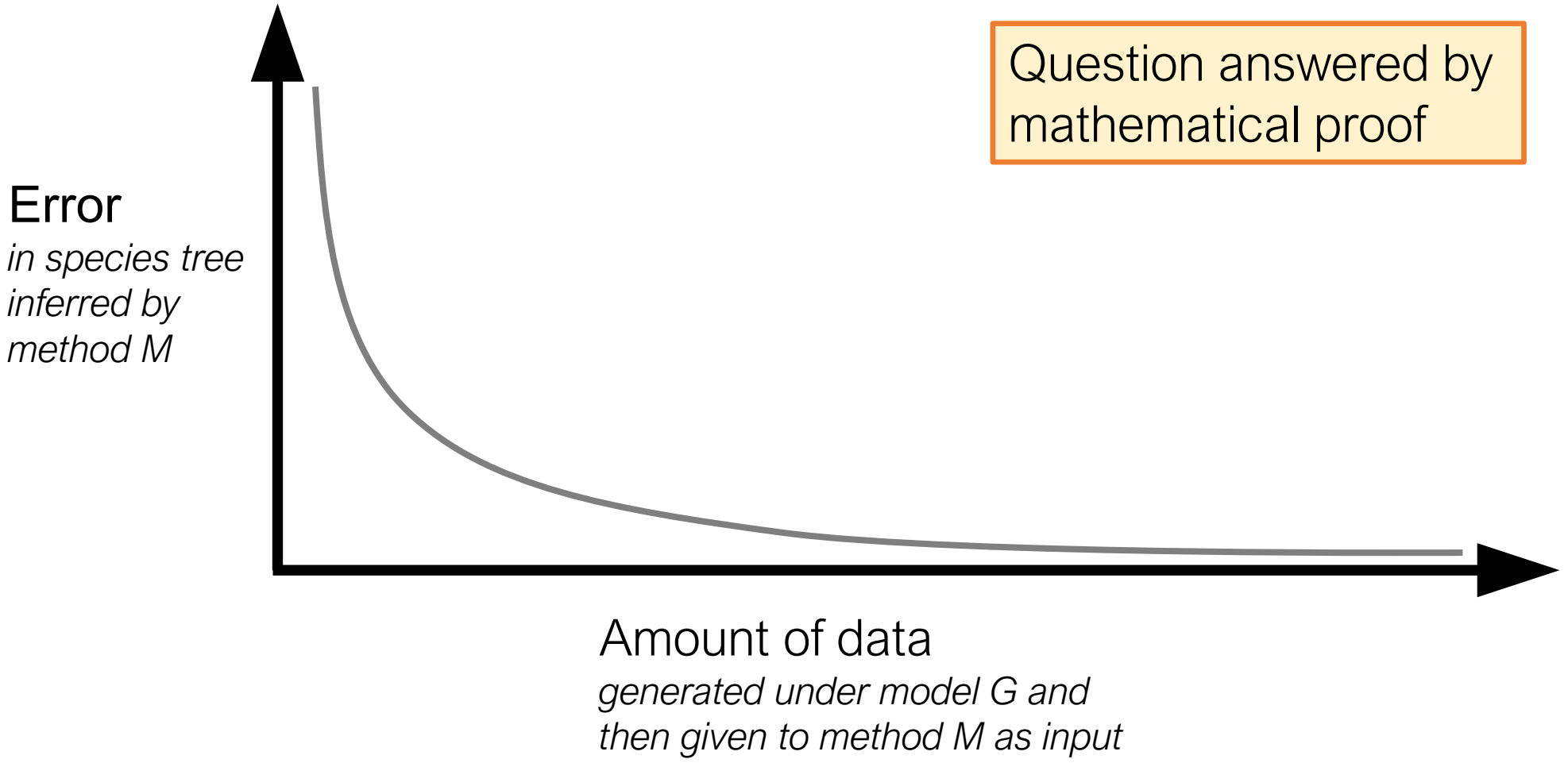
Gorilla and Orangutan are not siblings in the species tree, but they are in the gene tree.

MSC+GTR Hierarchical Model

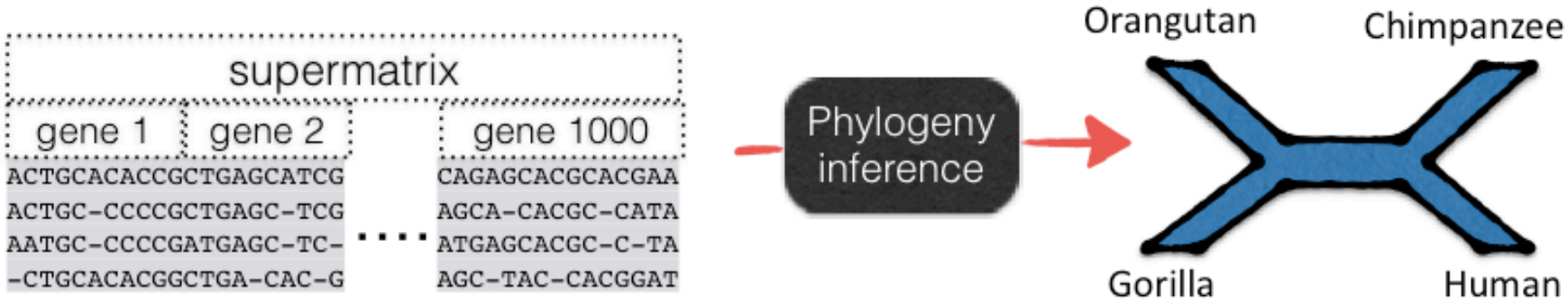


1. Gene trees evolve within the species tree (under the Multi-Species Coalescent model)
2. Sequences evolve down the gene trees (under GTR model)

Is method M statistically consistent under model G?

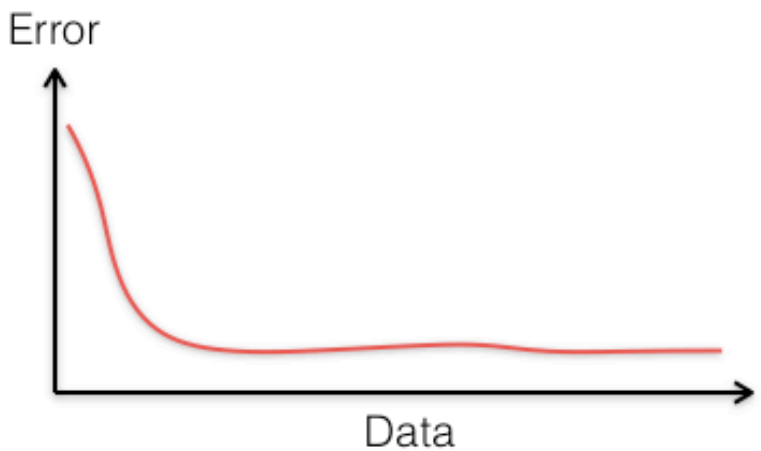


Traditional approach: concatenation

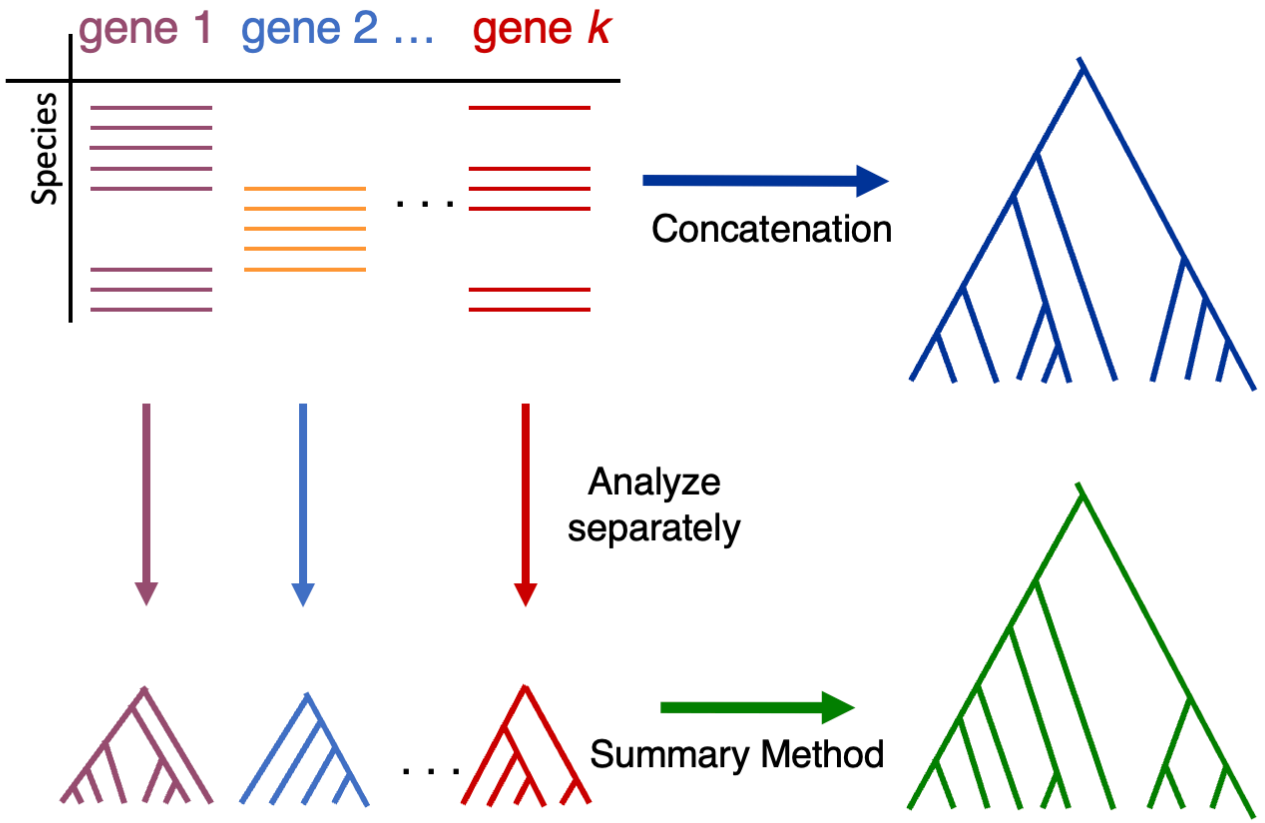


- Statistically <u>inconsistent</u> and can even be positively misleading (proved for unpartitioned maximum likelihood)
  [Roch and Steel, Theo. Pop. Gen., 2014]

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



Main Approaches for Species Tree Estimation



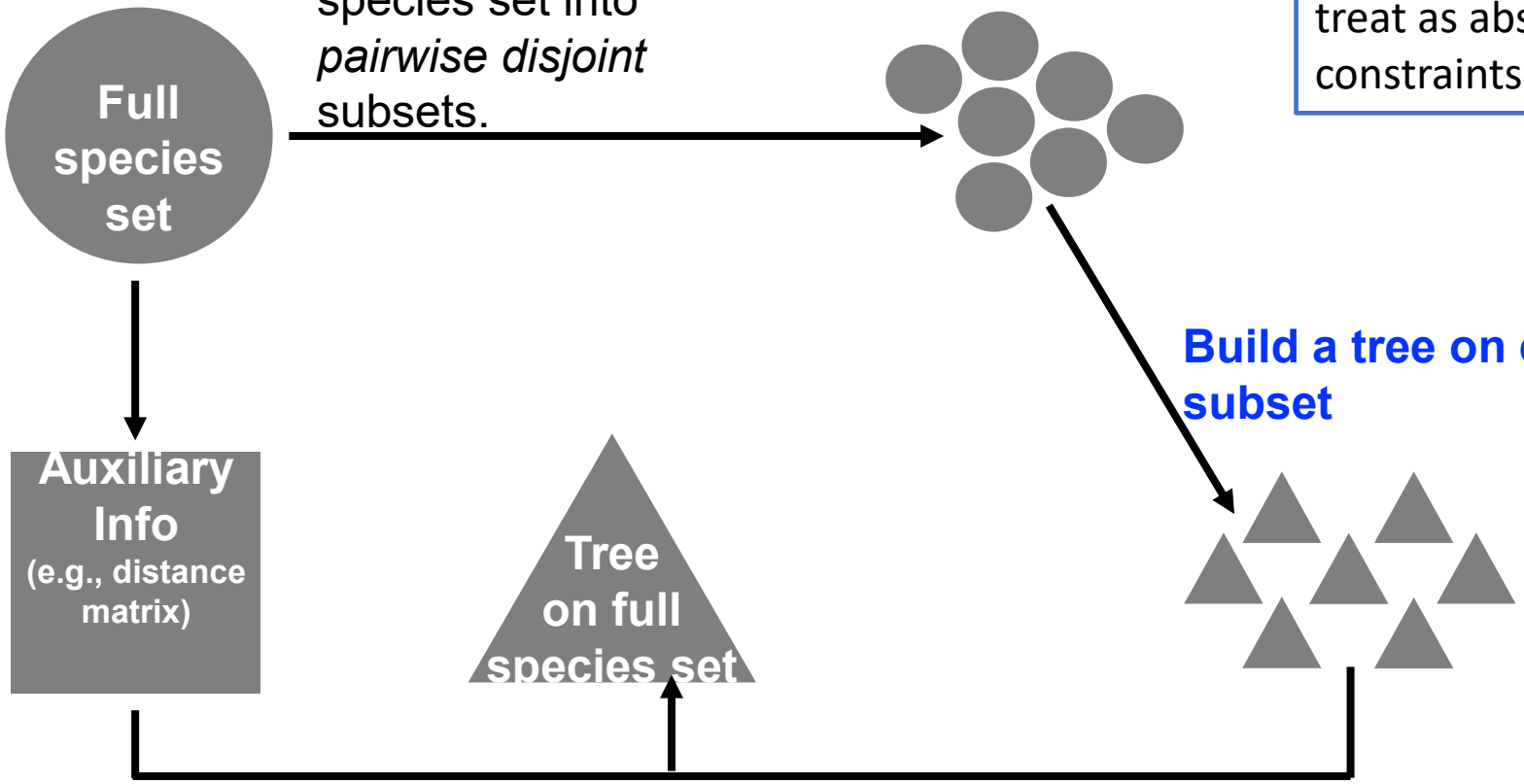
e.g., RAxML

e.g., ASTRAL

DTMs for Species Tree Estimation

Erin Molloy, Introduced this approach

Note: use most accurate method on subsets, and treat as absolute constraints



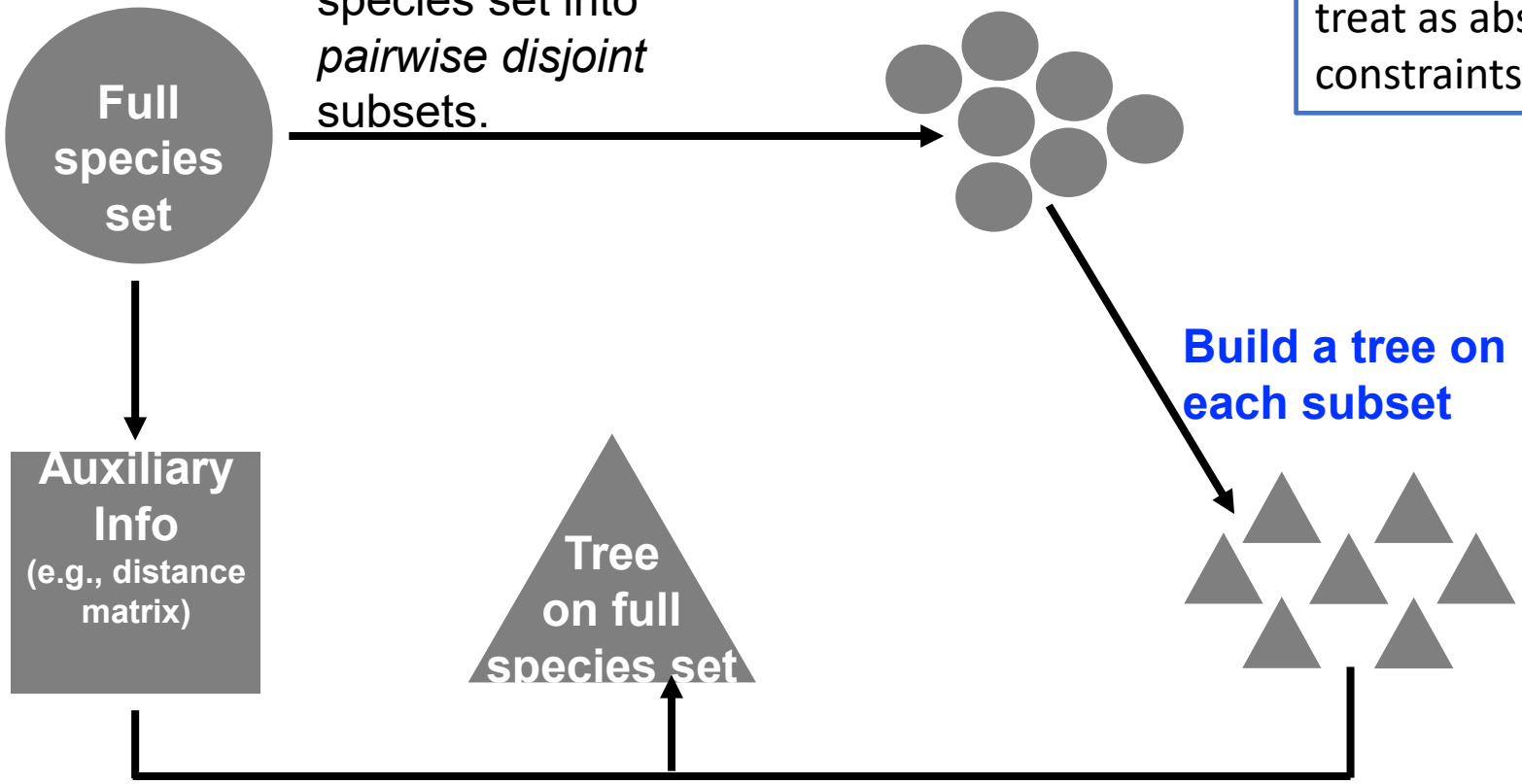
Use summary method or concatenation for subtree construction!

Combine with DTM method.

DTMs for Species Tree Estimation are Consistent

Erin Molloy, Introduced this approach

Note: use most accurate method on subsets, and treat as absolute constraints

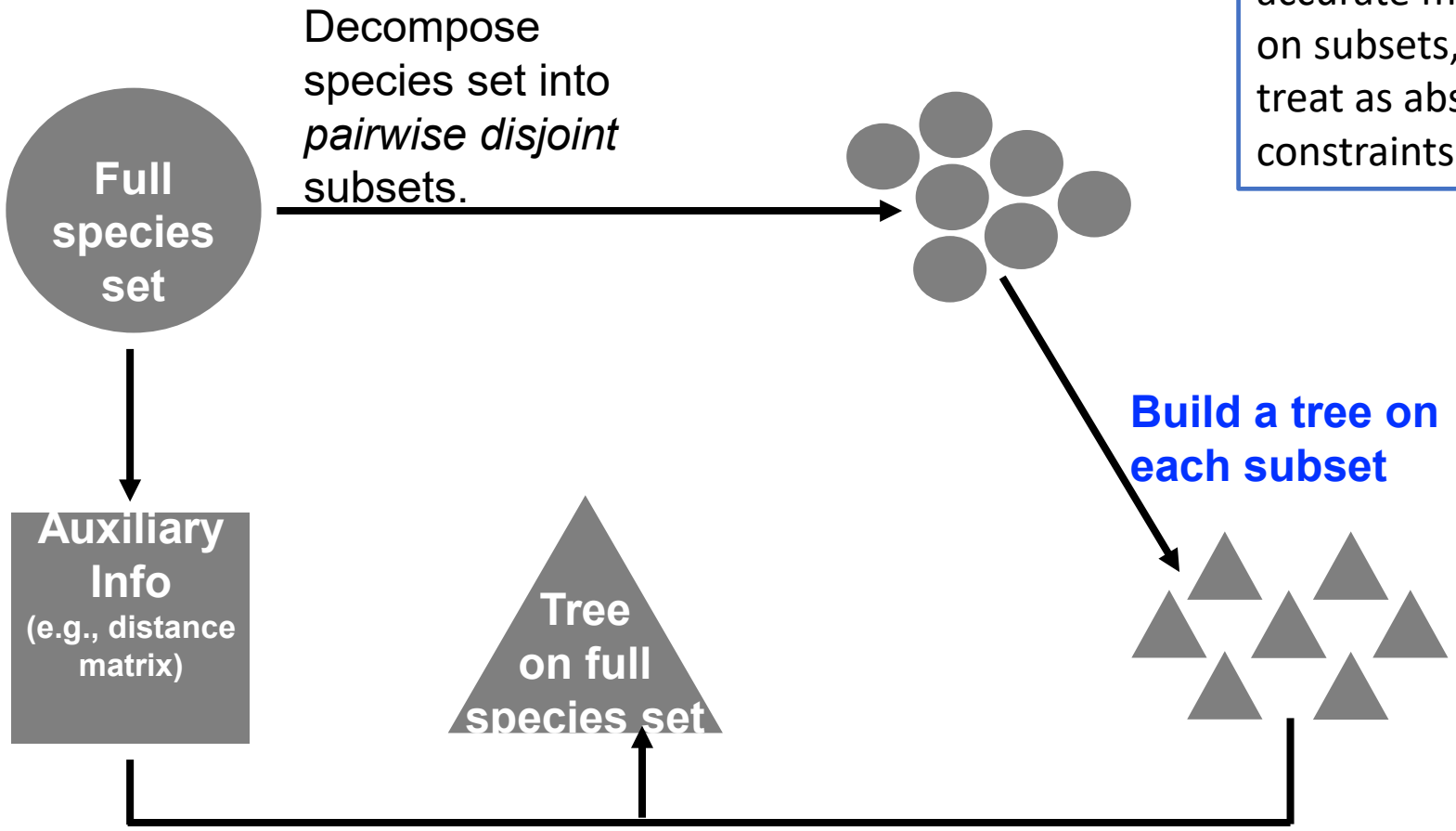


Theorem: Pipelines based on internode distance for auxiliary info, subtree calculation using summary methods, and then NJMerge or TreeMerge, are statistically consistent.

DTMs for Species Tree Estimation are Consistent

Vladimir Smirnov

Note: use most accurate method on subsets, and treat as absolute constraints



Theorem: Pipelines based on NJst/ASTRID for auxiliary info (guide tree), subtree calculation using summary methods, and then **Guide Tree Merger (GTM)** are statistically consistent.

GTM pipelines for Species Tree Estimation

- ASTRAL, NJst, and ASTRID are statistically consistent
- Concatenation using maximum likelihood (CA-ML) is not consistent
- GTM Pipelines we studied:
  - Guide tree is NJst or ASTRID
  - Subtrees computed using ASTRAL or CA-ML
  - Combined using GTM
- We evaluate accuracy and runtime under conditions with varying ILS levels

GTM+ASTRAL: faster and more accurate than ASTRAL

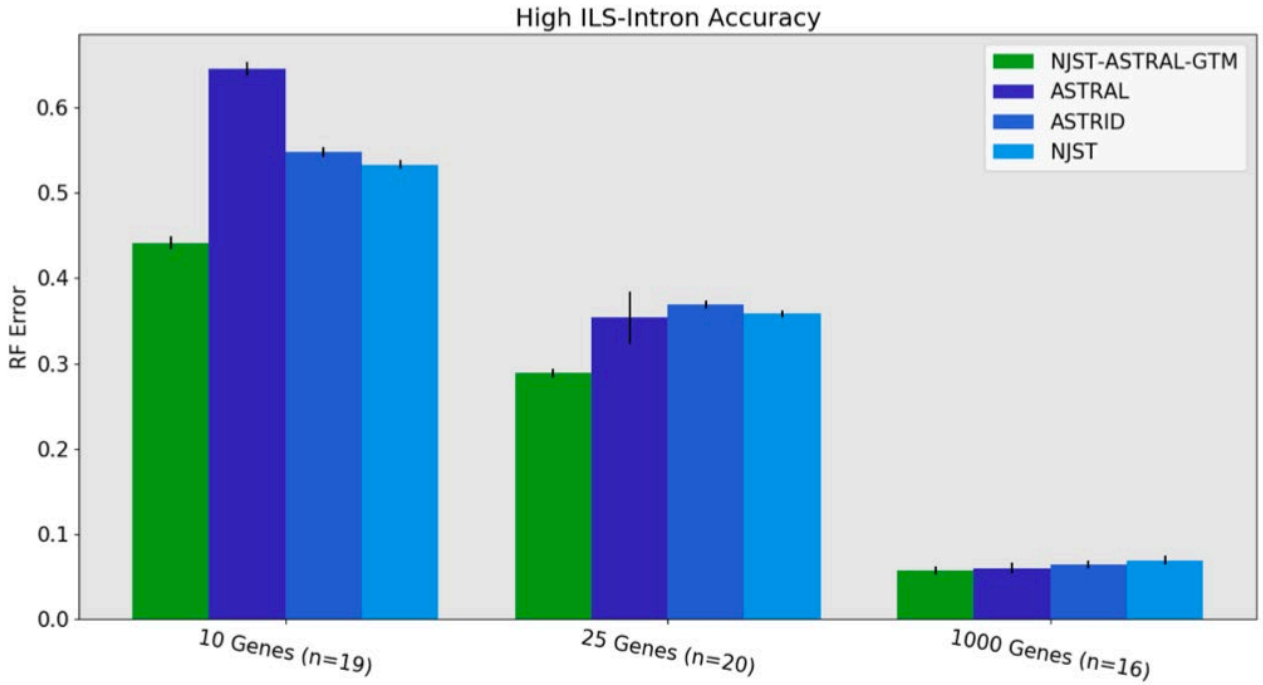


Table 3 Comparison of average runtime (seconds) of GTM+ASTRAL vs ASTRAL for high ILS conditions with introns on 1000 species. The value for n is the number of replicates being compared (i.e., where ASTRAL trees are available). Pre-GTM covers computing gene trees using FastTree, the NJst starting tree, and ASTRAL subset trees; the gap between "total" and "ASTRAL" for the right hand column reflects the time to compute gene trees using FastTree, which is 3.9 seconds per gene. Results for the 1000-gene ASTRAL trees are taken from the NJMerge study [2].

	GTM+ASTRAL	ASTRAL
10 Genes (n=18)		
-Pre-GTM	97.4	n.a.
-ASTRAL	n.a.	8,617.0
-GTM	0.4	n.a.
-Total	97.8	8,656.0
25 Genes (n=20)		
-Pre-GTM	174.7	n.a.
-ASTRAL	n.a.	5,441.4
-GTM	0.4	n.a.
-Total	175.1	5,539.4
1000 Genes (n=16)		
-Pre-GTM	7,948.9	n.a.
-ASTRAL	n.a.	149,145.9
-GTM	0.4	n.a.
-Total	7,949.3	153,045.9

GTM pipelines for CA-ML

- CA-ML: Concatenation using maximum likelihood
  - Not guaranteed statistically consistent
  - Can be highly accurate
- GTM Pipelines:
  - Guide tree is FastTree
  - Subtrees computed using CA-ML (using RAxML)
  - Combined using GTM
- We evaluate accuracy and runtime under conditions with varying ILS levels

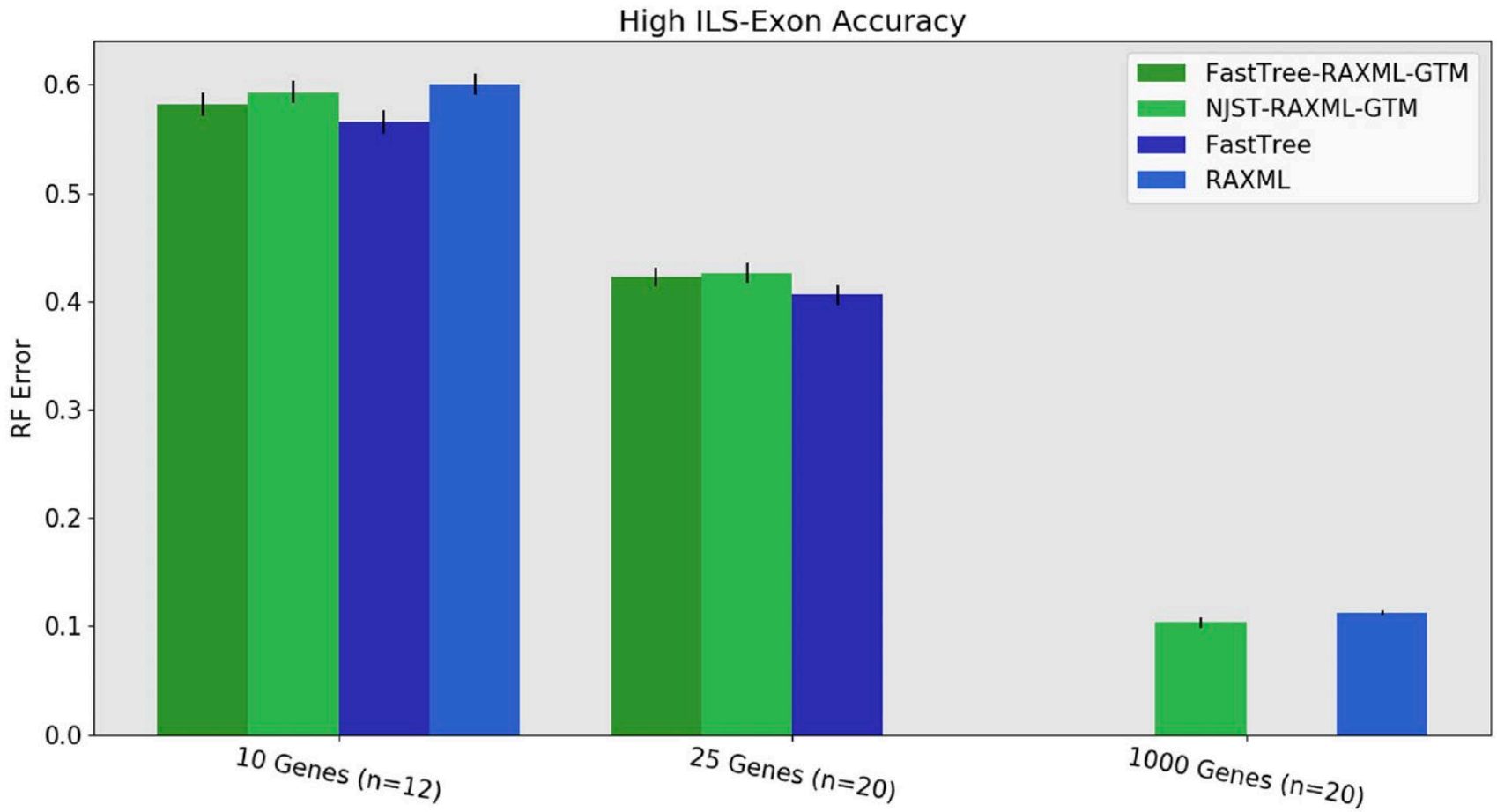


Fig. 9 Experiment 3: Comparison of FastTree-RAxML-GTM and NJst-RAxML-GTM to RAxML and FastTree on 1000-species datasets with high ILS exons. The value for *n* is the number of replicates on which RAxML completed; missing replicates indicate RAxML exceeding runtime limits on 10 and 25 genes (the 1000-gene RAxML trees are taken from [3]). FastTree was not used for 1000 genes. Error bars show standard error of the replicate average

Table 6 Average runtime (seconds) of FastTree-RAxML-GTM (GTM(RAxML)) and RAxML on 1000-species exon datasets

	GTM(RAxML)	RAxML
Low ILS 10 Genes (*n*=19)		
-FastTree	279.6	n.a.
-RAxML subtrees	831.3	n.a.
-GTM	0.4	n.a.
-Total	1,111.3	7,313.7
Low ILS 25 Genes (*n*=10)		
-FastTree	686.3	n.a.
-RAxML subtrees	1,460.6	n.a.
-GTM	0.4	n.a.
-Total	2,147.3	10,539.4
High ILS 10 Genes (*n*=12)		
-FastTree	283.7	n.a.
-RAxML subtrees	637.5	n.a.
-GTM	0.4	n.a.
-Total	921.6	10,135.6
High ILS 25 Genes (*n*=20)		
-FastTree	731.5	n.a.
-RAxML subtrees	1363.1	n.a.
-GTM	0.4	n.a.
-Total	2,095	n.a.

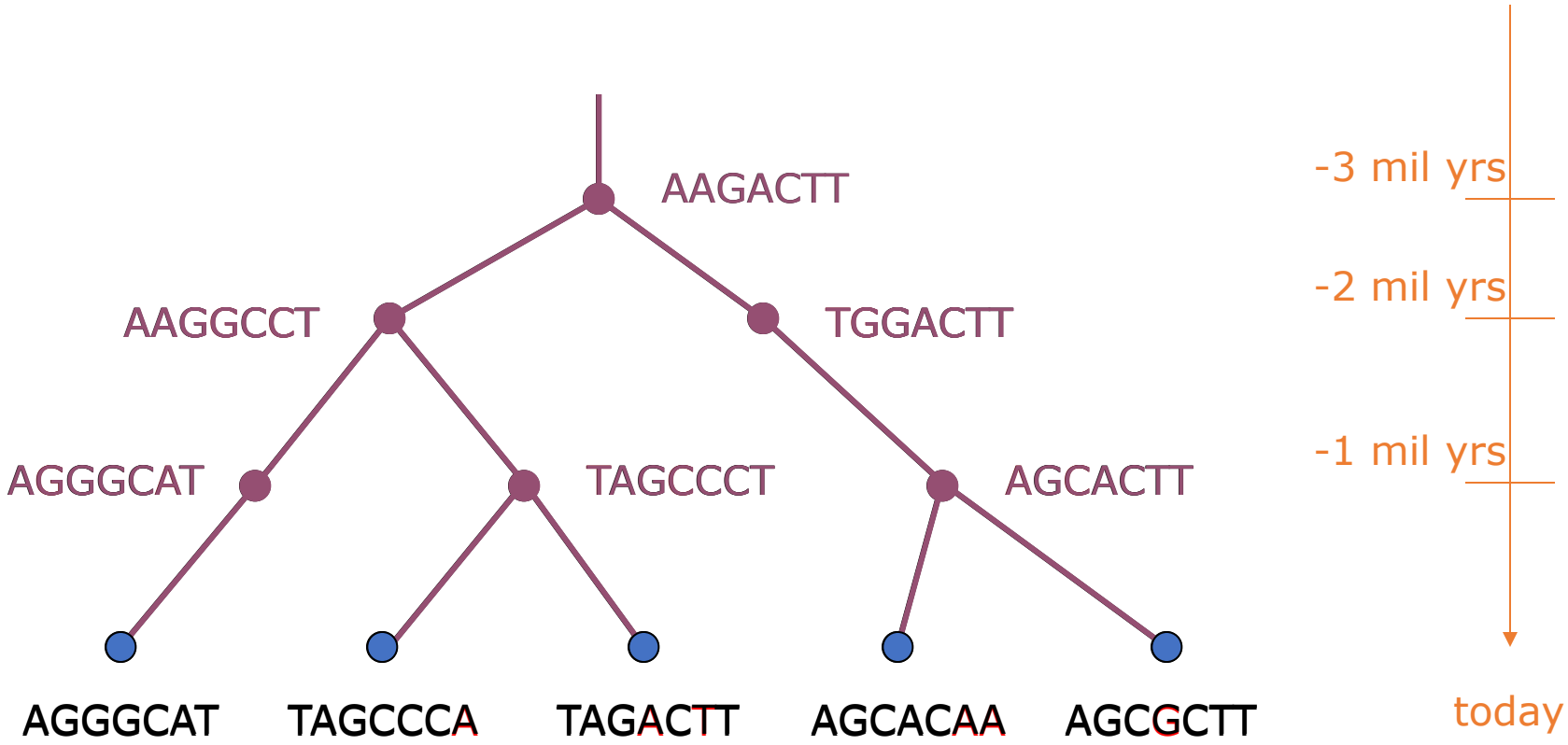
The value for *n* is the number of replicates being compared, i.e., where a RAxML tree is available

GTM pipelines improve running time for CA-ML

(Could make large-scale CA-ML feasible)

Part III: DTMs and Maximum Likelihood Tree Estimation

DNA Sequence Evolution (Idealized)



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 (e.g., Generalized Time Reversible) are also considered, often with little change to the theory.

Maximum likelihood tree estimation

- **Input**: multiple sequence alignment and “model” (e.g., GTR, Jukes-Cantor)
- **Output**: Model tree (rooted binary tree with numeric parameters) that maximizes the probability of producing the alignment

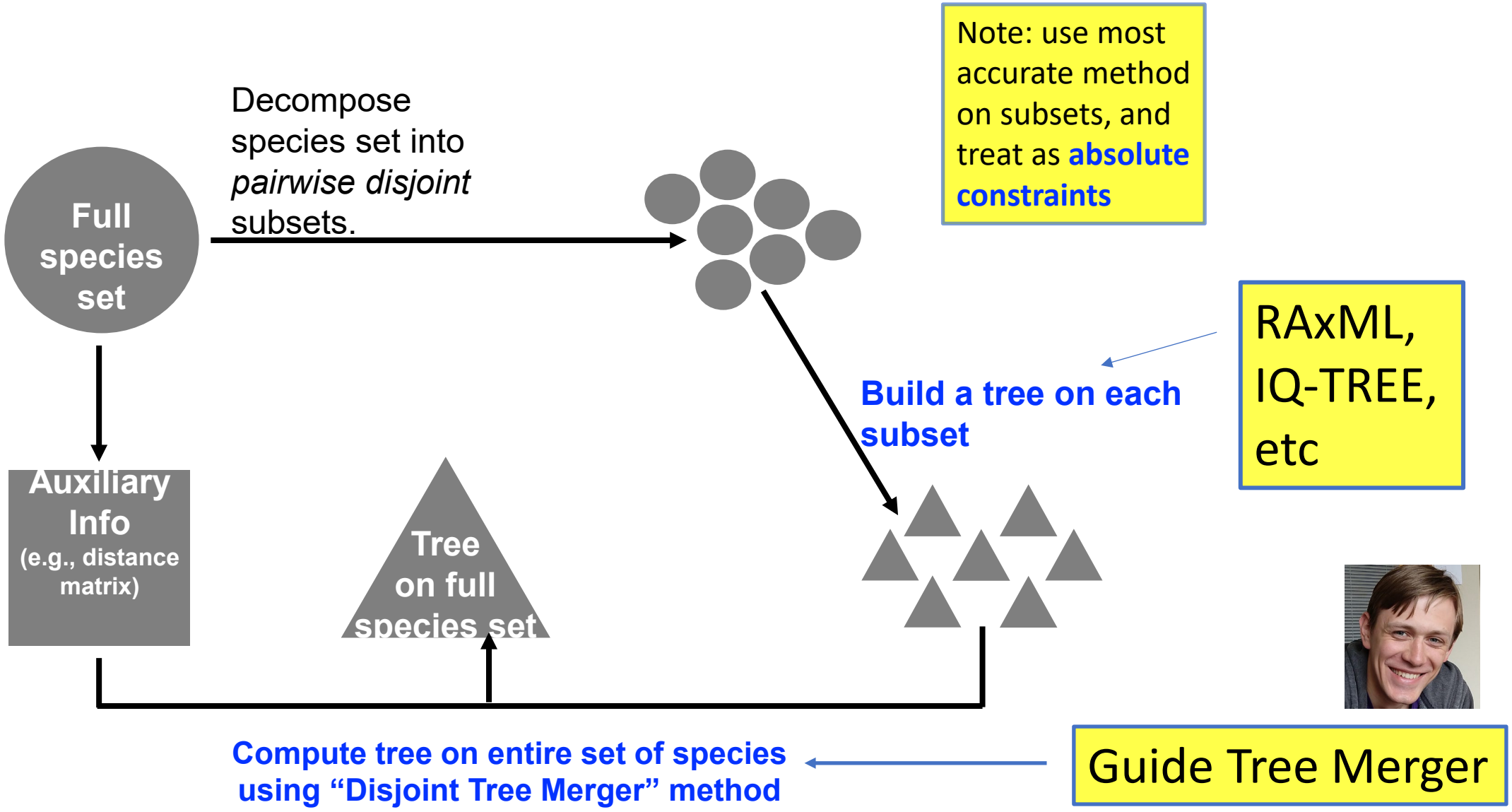
Maximum likelihood tree estimation

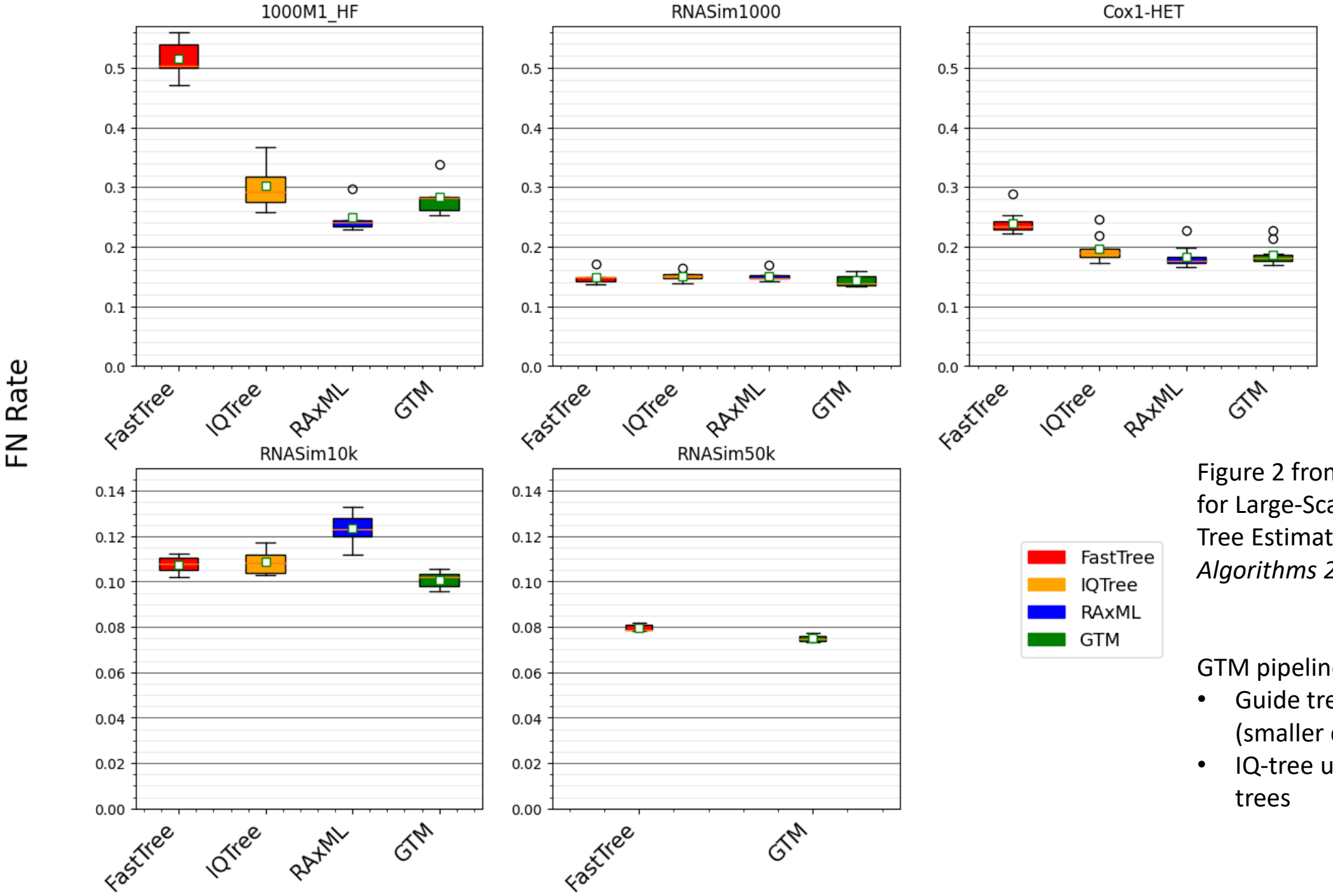
- Theory:
  - Statistically consistent under standard models
  - Excellent sample complexity (Roch & Sly, Prob. Theory and Related Fields, 2017): phase transition (logarithmic then polynomial)
  - NP-hard

Maximum Likelihood Software (heuristics)

- RAxML-ng (probably the best?)
- IQ-TREE2 (possibly competitive with RAxML-ng)
- FastTree 2 (extremely fast, not as accurate)
- And others, but none competitive with RAxML-ng

GTM for Maximum Likelihood Tree Estimation





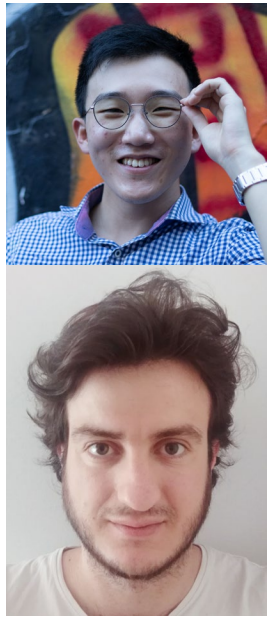
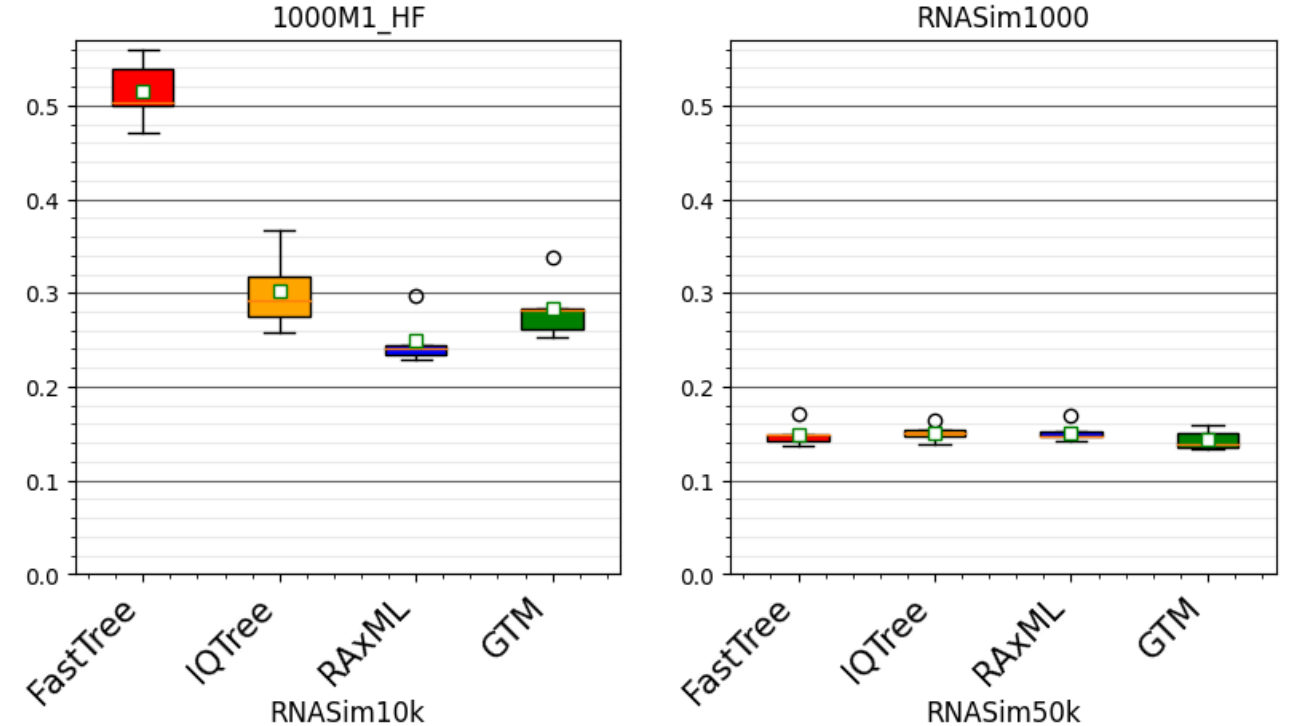


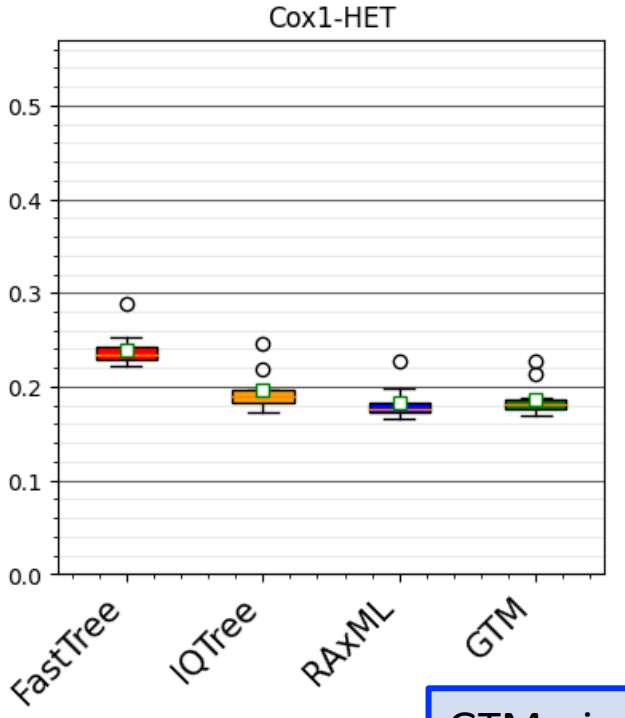
Figure 2 from "Disjoint Tree Mergers for Large-Scale Maximum Likelihood Tree Estimation", Park et al., *Algorithms 2021*

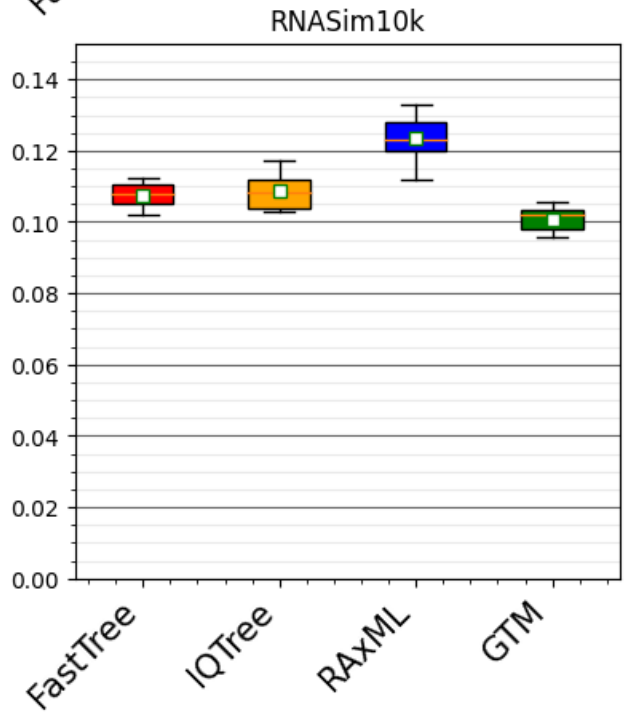
GTM pipeline:

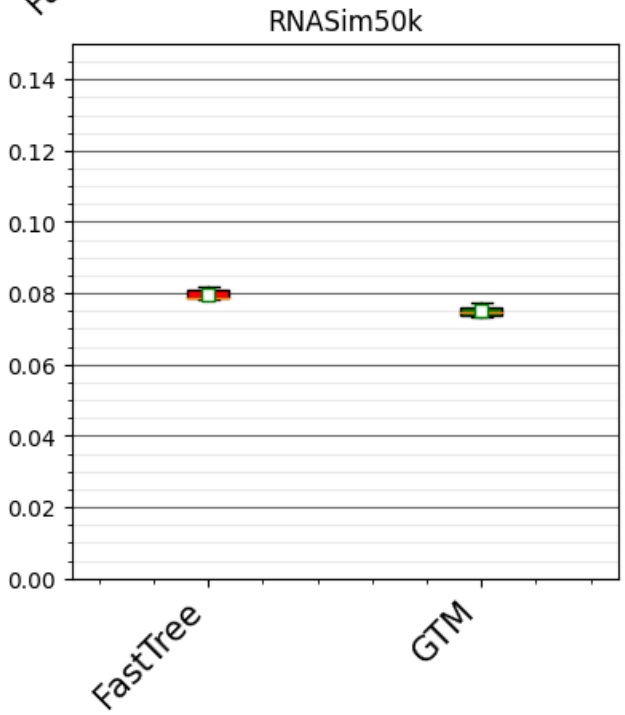
- Guide tree is IQ-Tree or FastTree (smaller datasets),
- IQ-tree used to compute subset trees

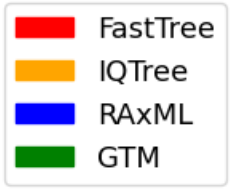
FN Rate





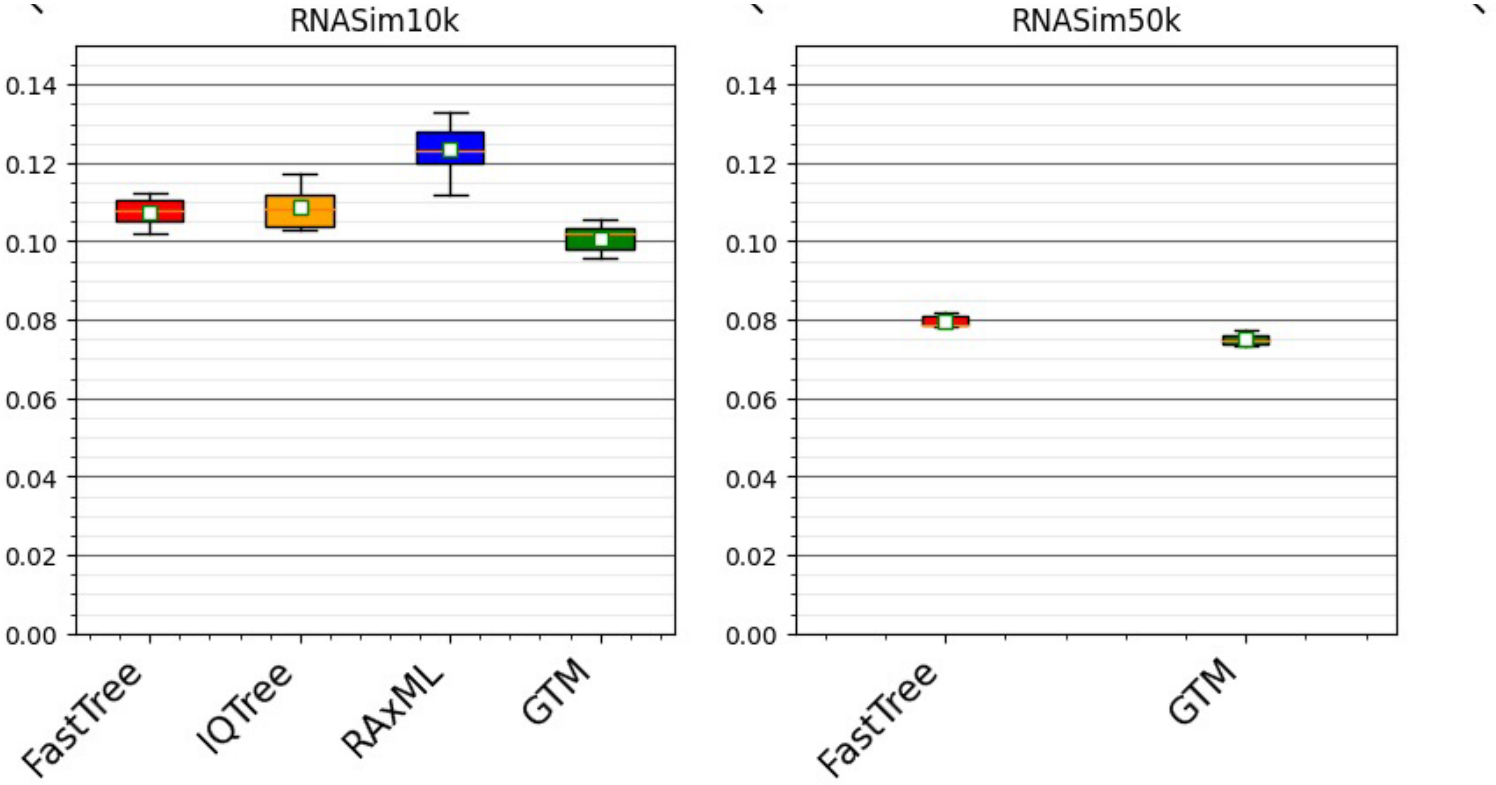


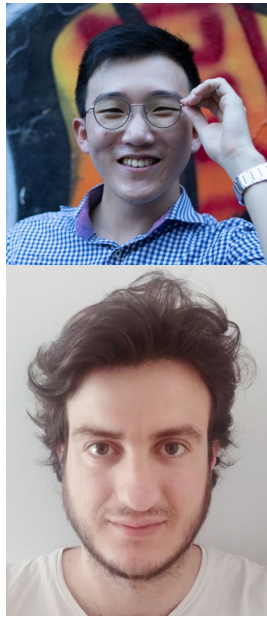




GTM-pipeline:

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

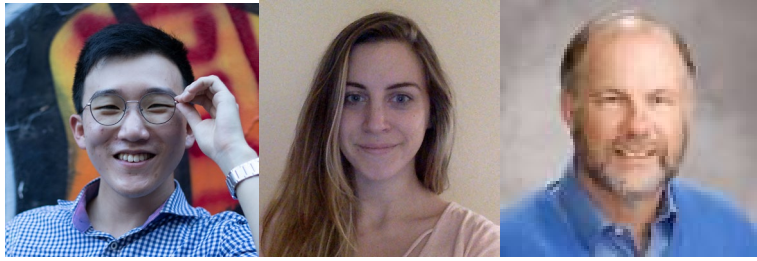




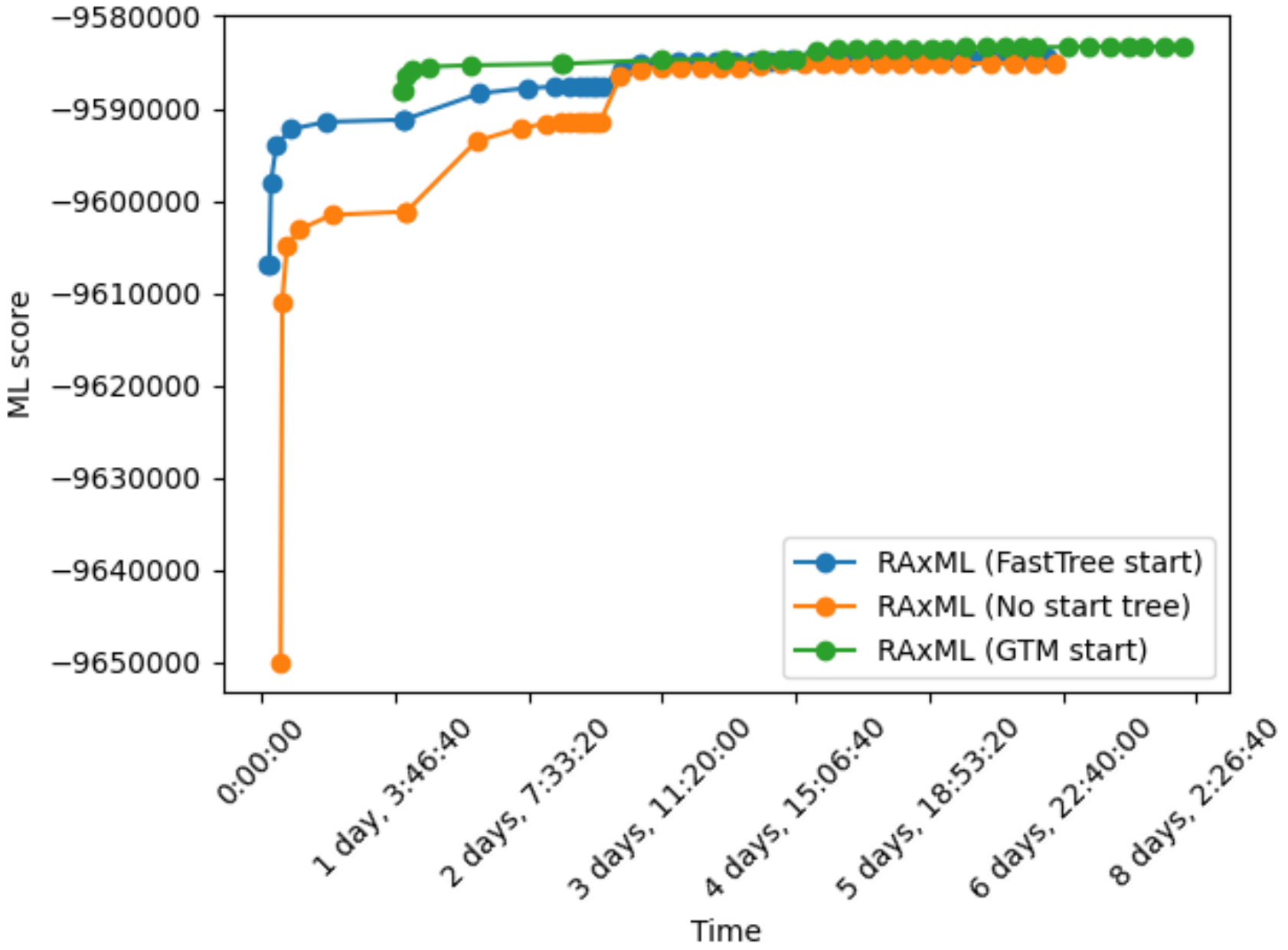
Trends on largest datasets:

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

What about biological data?



- We used the same technique but evaluated maximum likelihood scores on a MAGUS+EMMA alignment of the Recombinase dataset (~70,000 protein sequences) from Kelly Williams, restricting the alignment to approximately 1000 sites.
- Revised GTM pipeline: construct FastTree tree on full-length sequences, and add remaining sequences in using phylogenetic placement method BSCAMPP(EPA-ng) (tutorial on Thursday by Eleanor Wedell)
- We let RAxML run with different starting trees: 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.
- Unpublished analyses performed by Minhyuk Park.

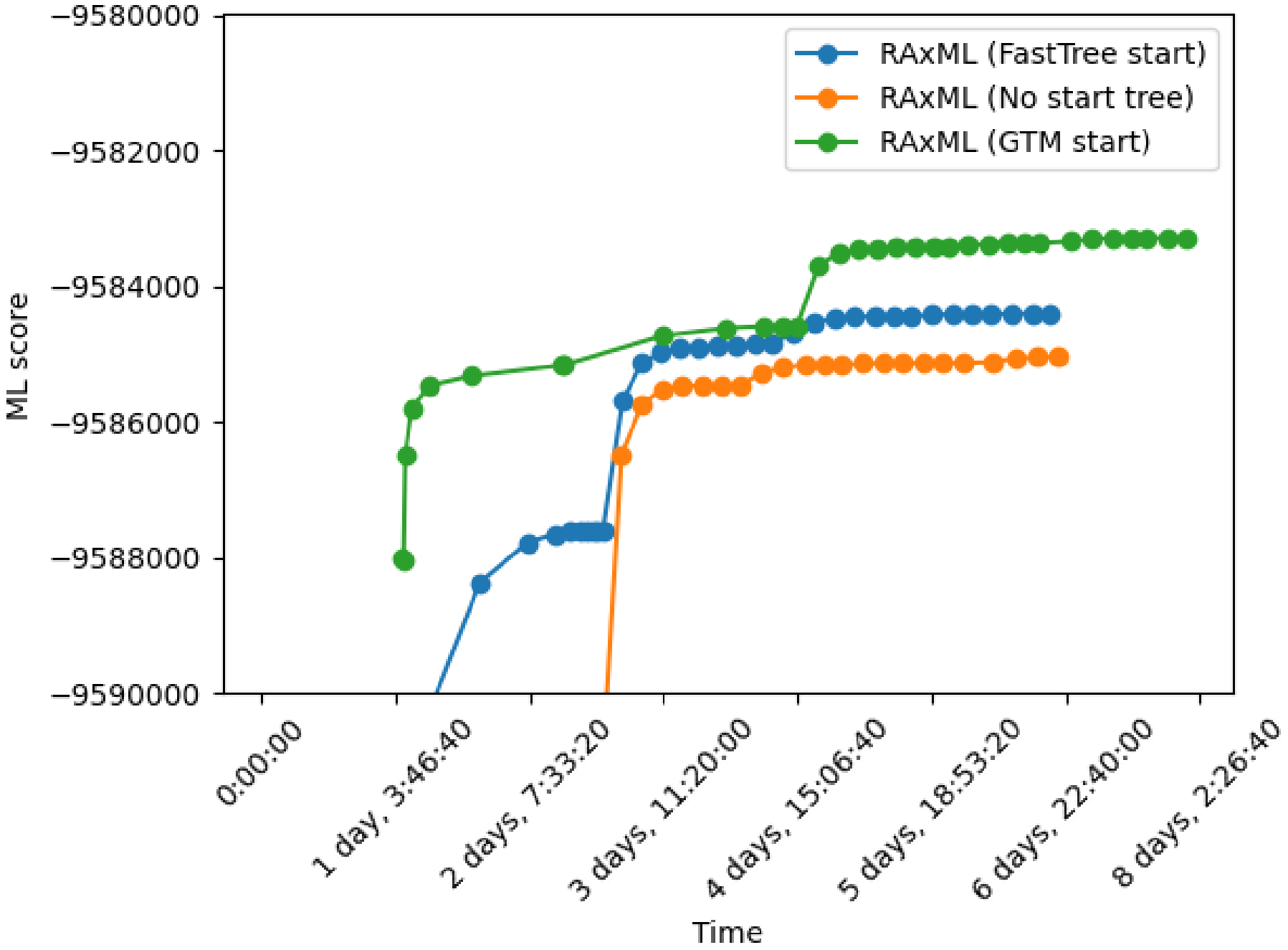


Analysis of Kelly Williams dataset (Minhyuk Park et al., NYP)

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!

Summary

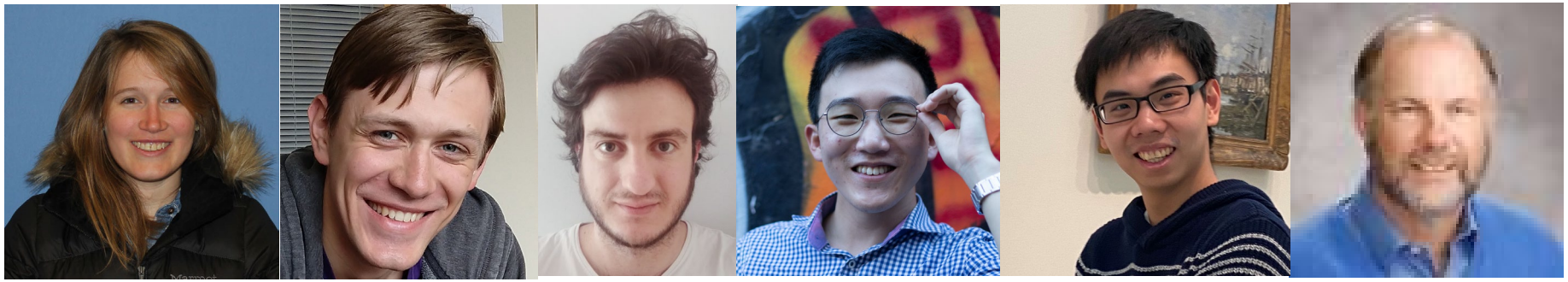
- “Disjoint tree mergers” (DTMs) are generic methods, that can be used with any phylogeny estimation method (for any kind of data).
- DTMs can be used in statistically consistent pipelines
- DTMs also provide empirical advantages:
  - DTMs enable scalability to large datasets.
  - DTMs improve gene tree and species tree estimation accuracy (based on simulation)
- GTM is the current leading DTM technique, based on empirical performance. However, because it does NOT allow blending, it is unlikely GTM is the best that can be done.

Open problems

- Open problems:
  - Develop a better DTM approach that allows blending.
  - Understand sample complexity
    - Impact of how division into subsets is done
    - Impact of subtree estimation method (e.g., maximum likelihood)
    - For GTM, evaluate impact of guide tree
  - Understand why GTM+ASTRAL is more accurate than ASTRAL
  - Examine use with Bayesian methods

- Not discussed here (and still needs work):
  - Phylogenetic networks
  - Genome rearrangement phylogeny

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, 2006069, Graduate Fellowship to Erin Molloy), the Grainger Foundation, the Ira and Debra Cohen Fellowship to Vlad Smirnov, and Sandia National Laboratories-Livermore (LDRD)

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

Write to me: warnow@Illinois.edu