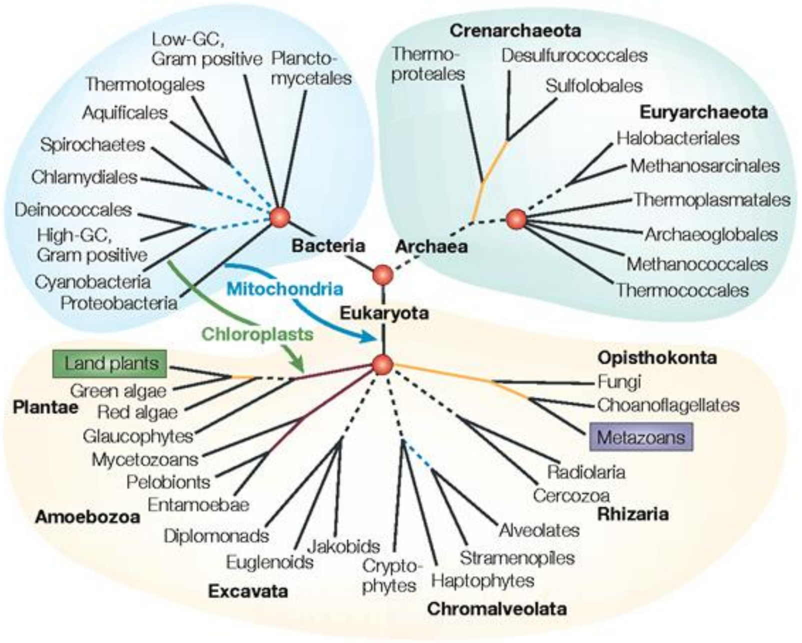
# New methods for very-large scale tree estimation

Tandy Warnow
School of Computing and Data Science
Grainger College of Engineering
The University of Illinois Urbana-Champaign

The international journal of science / 31 October 2019
outline Prostate cancer
nature
ROOTS OF DIVERSITY
Transcriptome analysis illuminates evolution of the world's green plants
A history of ethics
The long and bumpy road to responsible research
Ancient climate
A snapshot of CO₂ in the atmosphere more than 1 million years ago
Insects in decline
Ten-year survey offers strong evidence of falling numbers

# Phylogenomics



Nature Reviews | Genetics

Phylogeny + genomics = genome-scale phylogeny 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:

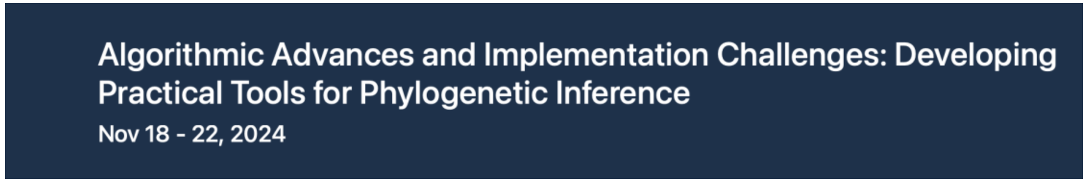
- Multi-copy genes omitted
- Massive gene tree heterogeneity consistent with ILS
- Concatenation analysis took 250 CPU years

# 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

- So many talks about large-scale phylogenetic tree estimation!

- Example topics
- NP-hard problems,
- species tree estimation,
- likelihood-based statistical estimation,
- model complexity,
- assessing branch support
- estimating dates
- distance-based estimation
- visualization of large trees
-
- And then the many talks about phylogenetic networks!



Algorithmic Advances and Implementation Challenges: Developing Practical Tools for Phylogenetic Inference

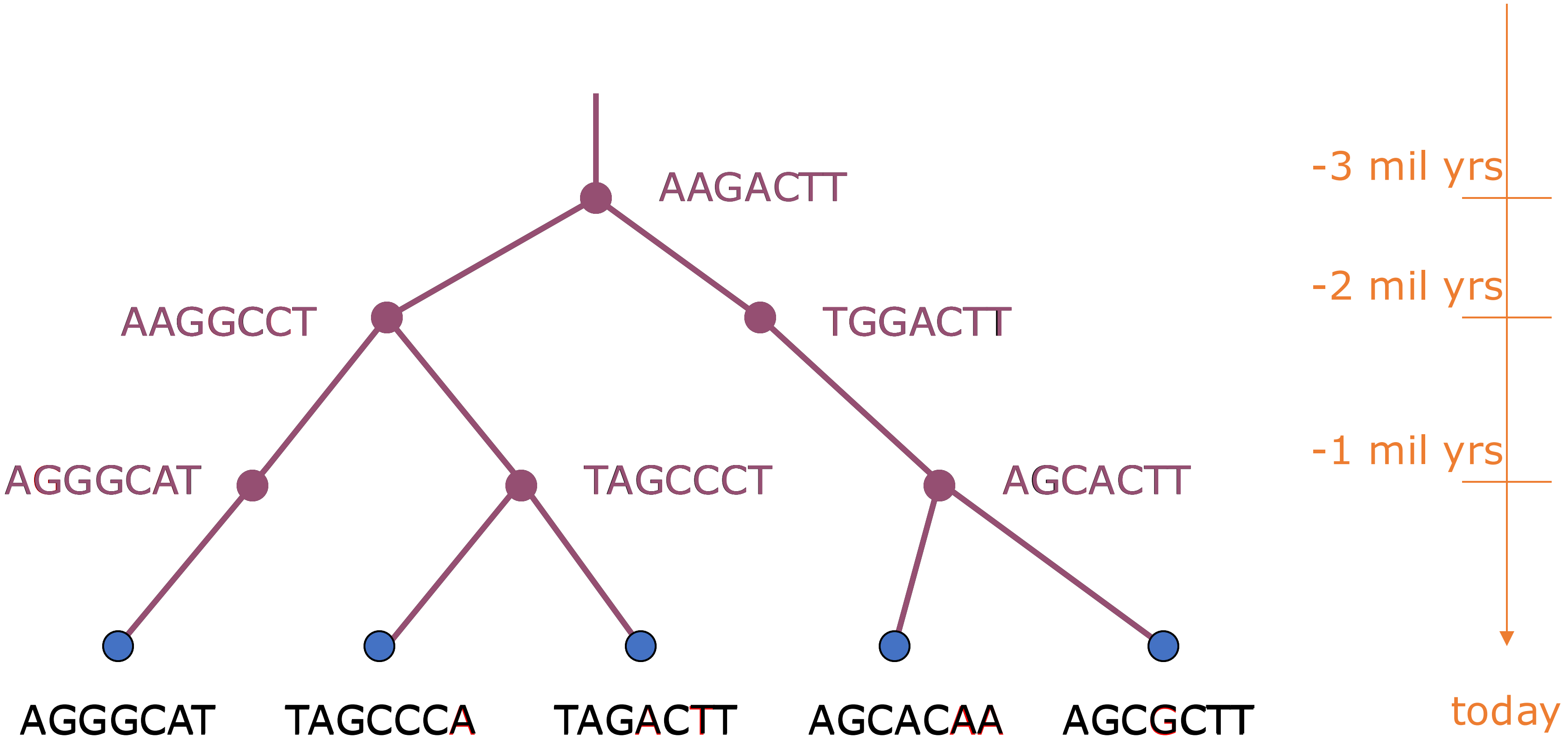
Nov 18 - 22, 2024

# This talk: Scaling methods to large trees

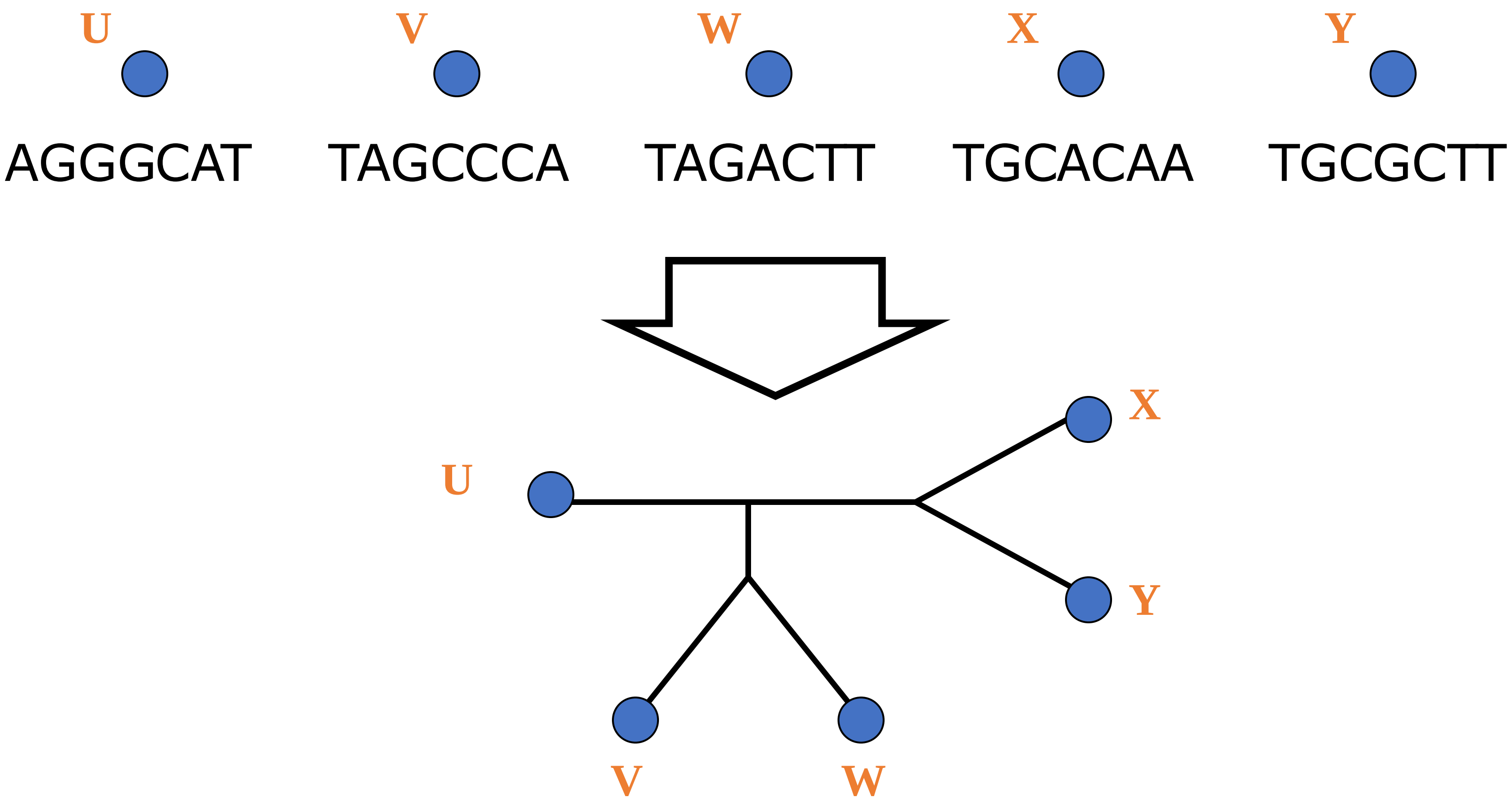
- Part I:  Divide-and-conquer using supertrees
- Part II: Divide-and-conquer using Disjoint Tree Mergers
- Part III: Discussion and open problems

# Part I: Divide-and-Conquer using Supertrees

# DNA Sequence Evolution (Idealized)



# Phylogeny Problem



AGGGCAT TAGCCCA TAGACTT TGCACAA TGCGCTT

# 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.

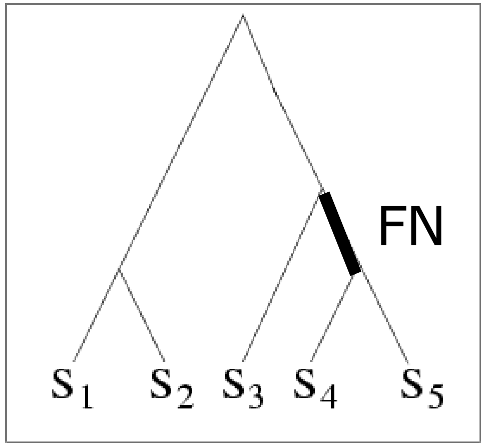
# Phylogeny estimation: statistical problem

- Assume DNA sequences are generated on an unknown model tree, infer the tree from the observed sequences seen at the leaves
- Many methods:
  - Maximum likelihood: Find the model tree that maximizes the probability of generating the observed sequences
  - Bayesian estimation
  - Distance-based methods (e.g., neighbor joining)
  - Maximum parsimony

NP-hard optimization problems, heuristics

# Phylogeny estimation method evaluation

- Statistical properties
  - consistency
  - sample complexity
- Computational performance
  - Most problems are NP-hard, so many methods are heuristics
- Accuracy
  - on simulated datasets
  - on biological datasets



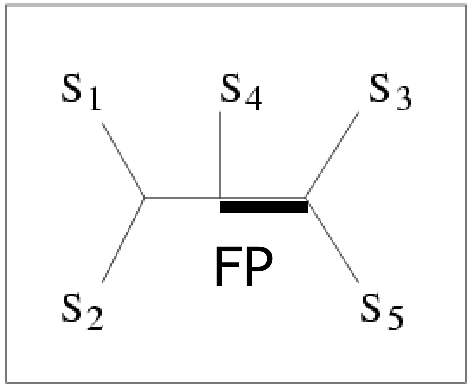
TRUE TREE

| | |
|---|---|
| $S_1$ | ACAATTAGAAC |
| $S_2$ | ACCCTTAGAAC |
| $S_3$ | ACCATTCCAAC |
| $S_4$ | ACCAGACCAAC |
| $S_5$ | ACCAGACCGGA |

DNA SEQUENCES

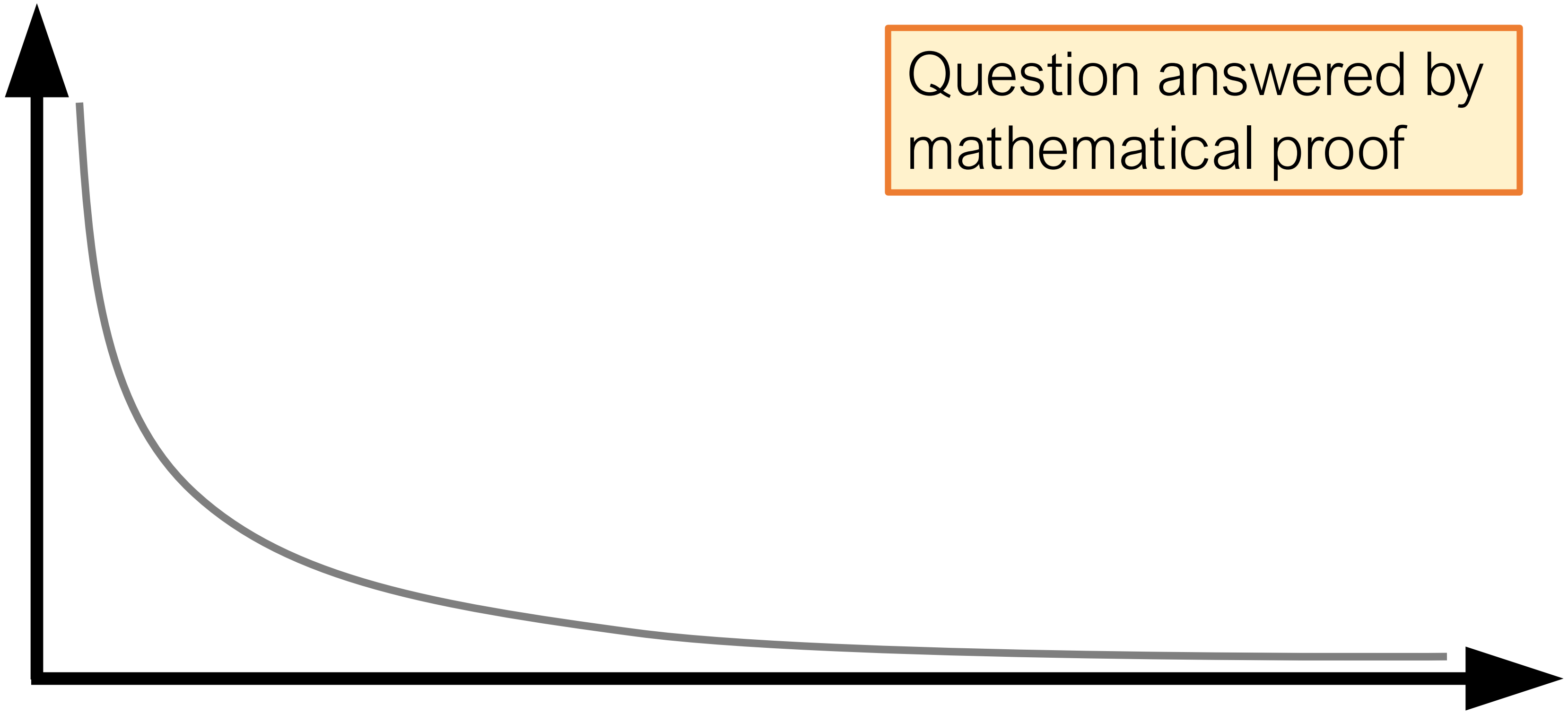
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

**50% error rate**



INFERRED TREE

# Statistical Consistency under model G?



Question answered by mathematical proof

# Sample Complexity

The sequence length (number of sites) that suffices for a phylogeny reconstruction method M to reconstruct the true tree with probability at least 1-ε depends on

- M (the method)
- ε
- f = min w(e),
- g = max w(e), and
- n, the number of leaves

We fix everything but n.

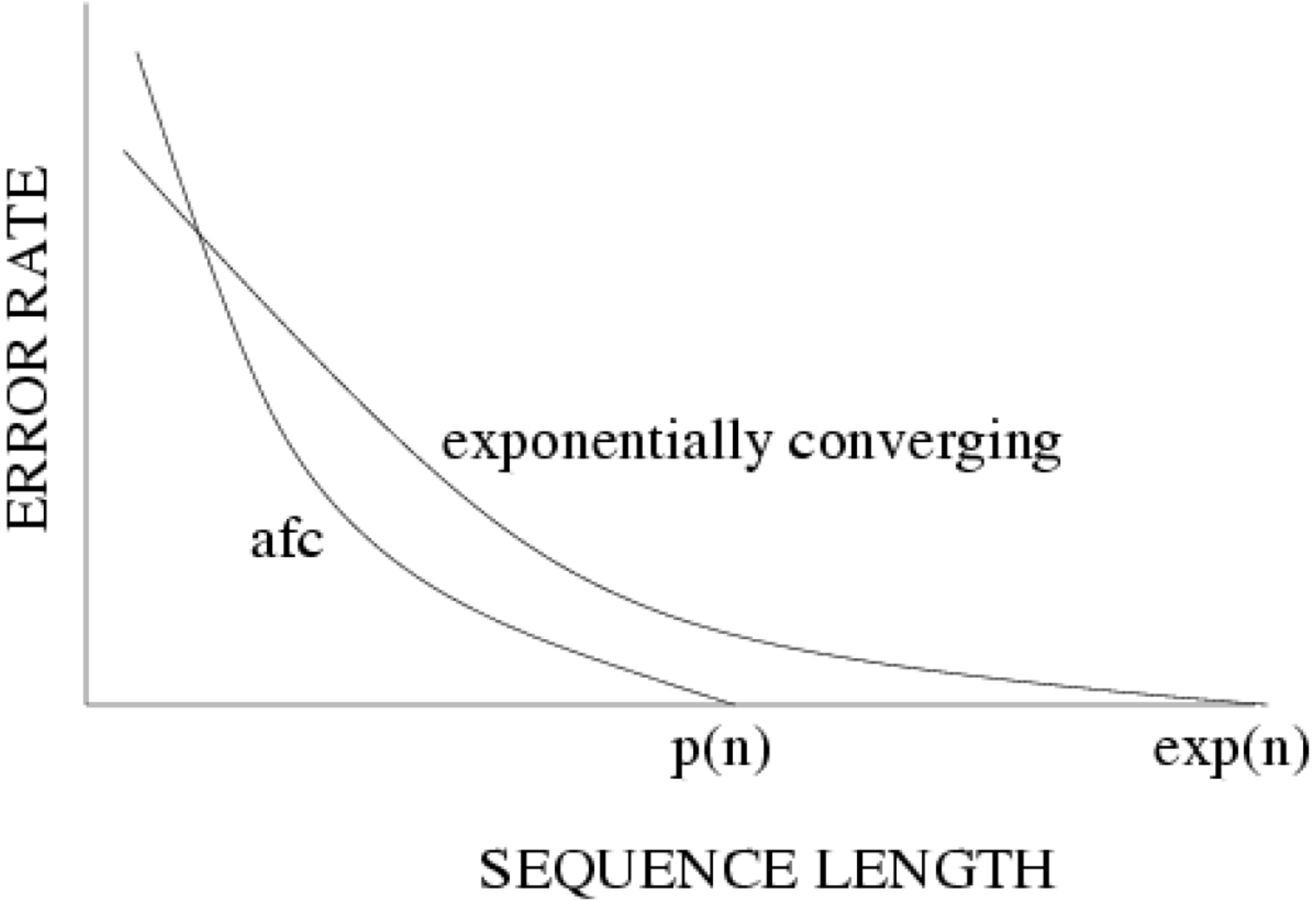
# Absolute Fast Converging (AFC) methods

A method M is “absolute fast converging”, or afc, if for all positive f, g, and $\varepsilon$, there is a polynomial p(n) s.t. Pr(M(S)=T) > 1- $\varepsilon$, when S is a set of sequences generated on T of length at least p(n).

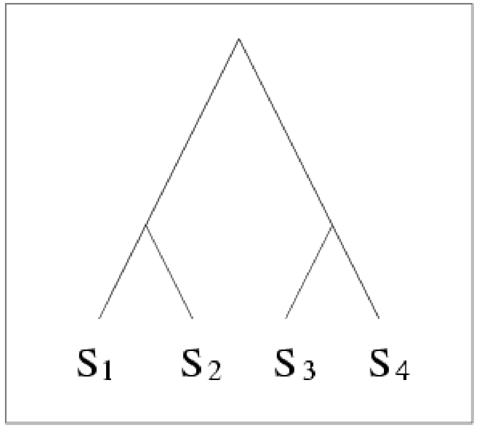
Notes:

1. The polynomial p(n) will depend upon M, f, g, and $\varepsilon$.
2. The method M is not “told” the values of f and g.

# Sample Complexity



# Distance-based estimation



TRUE TREE

→

| | |
|---|---|
| $S_1$ | ACAATTAGAAC |
| $S_2$ | ACCCTTAGAAC |
| $S_3$ | ACCATTCCAAC |
| $S_4$ | ACCAGACCAAC |

DNA SEQUENCES

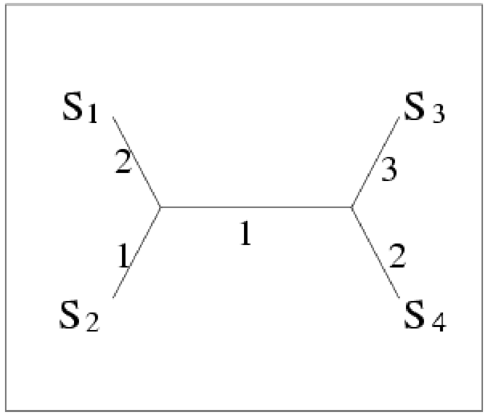
STATISTICAL ESTIMATION OF PAIRWISE DISTANCES

| | $S_1$ | $S_2$ | $S_3$ | $S_4$ |
|---|---|---|---|---|
| $S_1$ | 0 | 3 | 6 | 5 |
| $S_2$ | | 0 | 5 | 4 |
| $S_3$ | | | 0 | 5 |
| $S_4$ | | | | 0 |

DISTANCE MATRIX

METHODS SUCH AS NEIGHBOR JOINING

←



INFERRED TREE

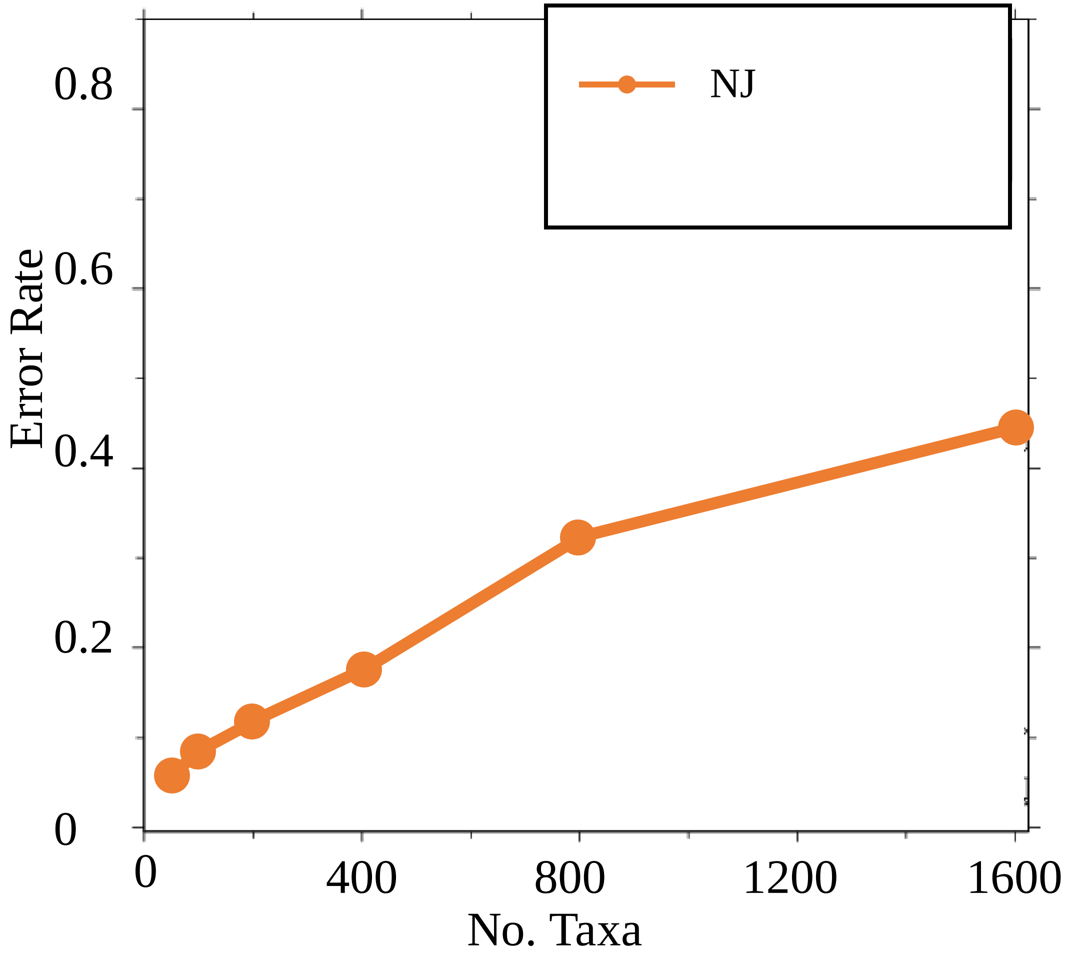
Theorem (Erdos et al., Atteson):

Neighbor joining (and some other methods) will return the true tree w.h.p. provided sequence lengths are exponential in the evolutionary diameter of the tree.

Sketch of proof:

- NJ (and other distance methods) guaranteed correct if *all* entries in the estimated distance matrix have sufficiently low error.
- Estimations of large distances require long sequences to have low error w.h.p.

# NJ has high error on large diameter trees



Simulation study based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.

Error rates reflect proportion of incorrect edges in inferred trees.

*[Nakhleh et al. ISMB 2001]*

# AFC methods (and related work)

- 1997: Erdos, Steel, Szekely, and Warnow (ICALP).
- 1999: Erdos, Steel, Szekely, and Warnow (RSA, TCS); Huson, Nettles and Warnow (J. Comp Bio.)
- 2001: Warnow, St. John, and Moret (SODA); Cryan, Goldberg, and Goldberg (SICOMP); Csuros and Kao (SODA); Nakhleh, St. John, Roshan, Sun, and Warnow (ISMB)
- 2002: Csuros (J. Comp. Bio.)
- 2006: Daskalakis, Mossel, Roch (STOC), Daskalakis, Hill, Jaffe, Mihaescu, Mossel, and Rao (RECOMB)
- 2007: Mossel (IEEE TCBB)
- 2008: Gronau, Moran and Snir (SODA)
- 2010: Roch (Science)
- 2017: Roch and Sly (Prob. Theory and Related Fields)

and others

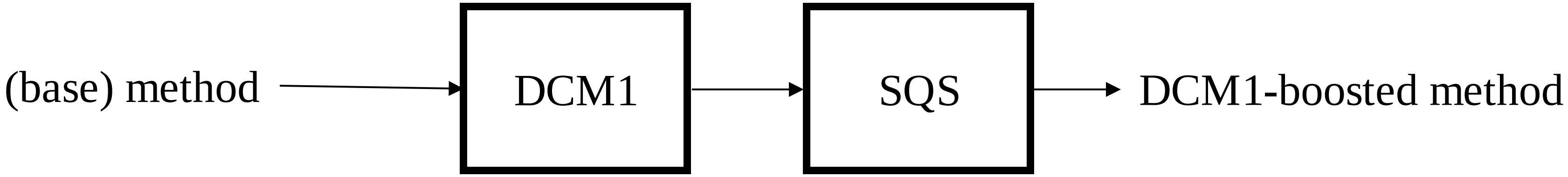
# DCM1: Divide-and-conquer AFC method

- DCM: disk-covering method

- Idea is to use divide-and-conquer to decompose a dataset into subsets, apply your favored method to construct trees on the subsets, and then combine these trees into a tree on the full dataset using a supertree method.

But, the details matter (see Stendhal)

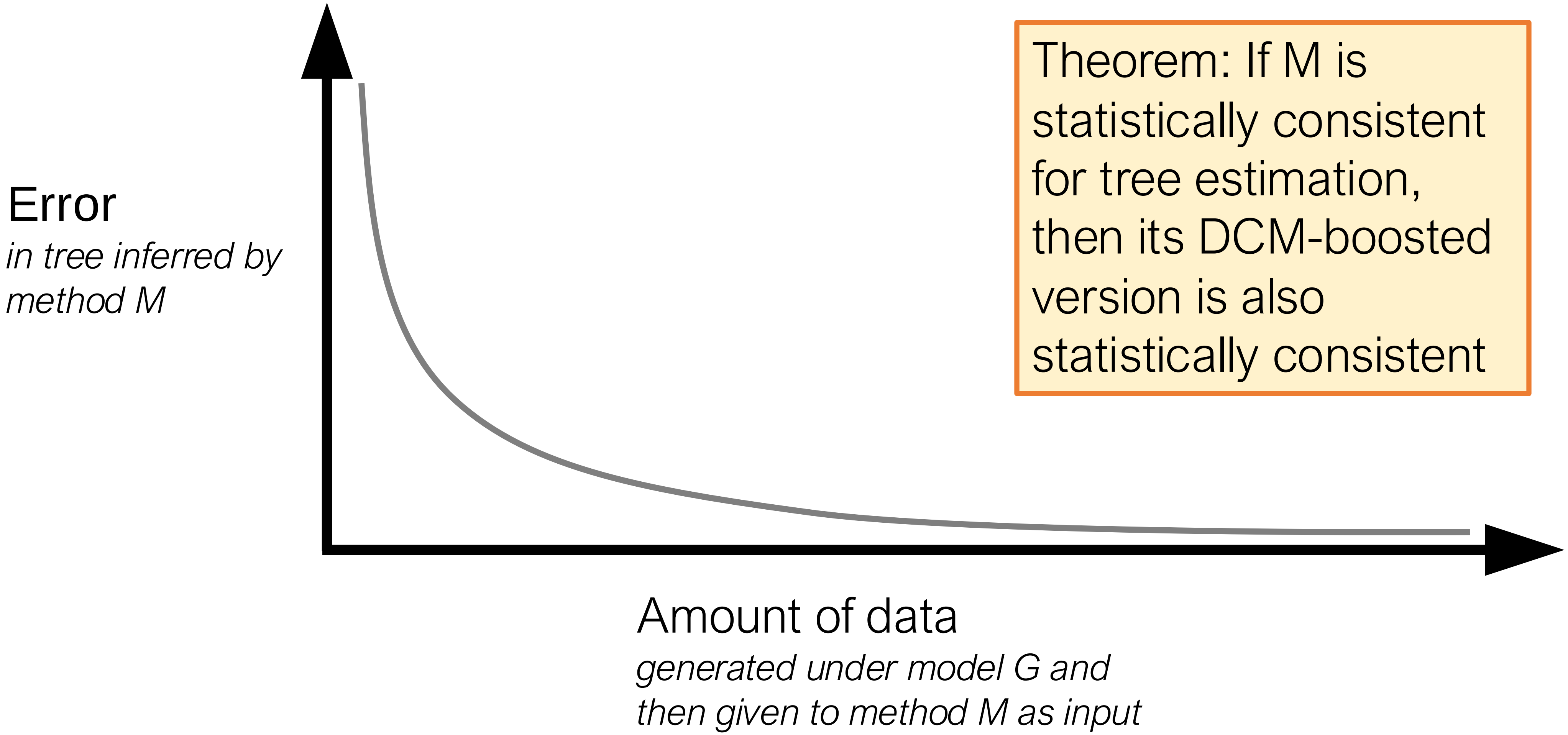
# DCM1-boosting



- The DCM1 phase produces a collection of trees (one for each threshold), and the SQS phase picks the “best” tree.
- For a given threshold, the base method is used to construct trees on small subsets (defined by the threshold) of the taxa. These small trees are then combined into a tree on the full set of taxa.

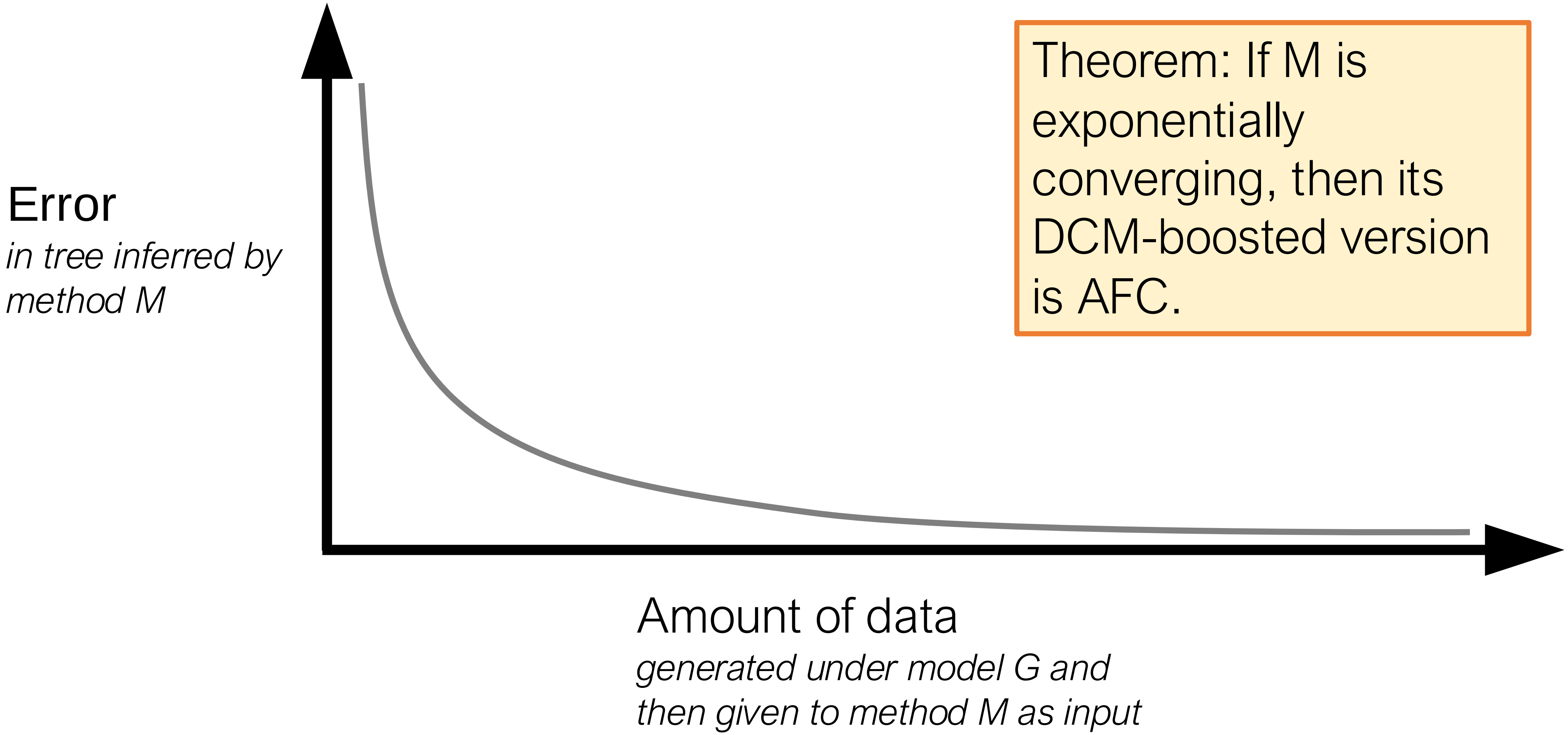
*Warnow, St. John, and Moret, SODA 2001*

# DCM-boosting maintains statistical consistency



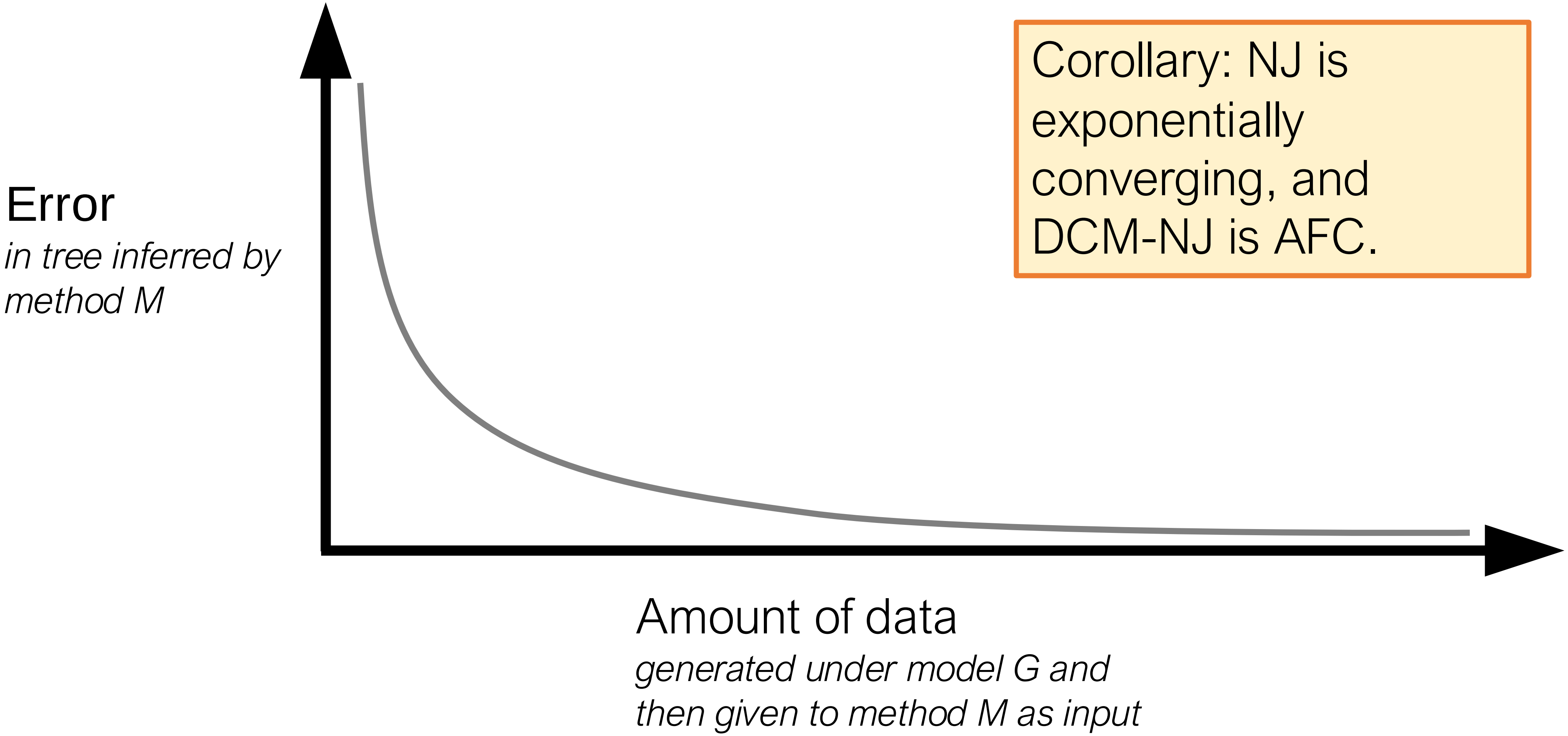
Theorem: If M is statistically consistent for tree estimation, then its DCM-boosted version is also statistically consistent

# DCM-boosting improves sample complexity



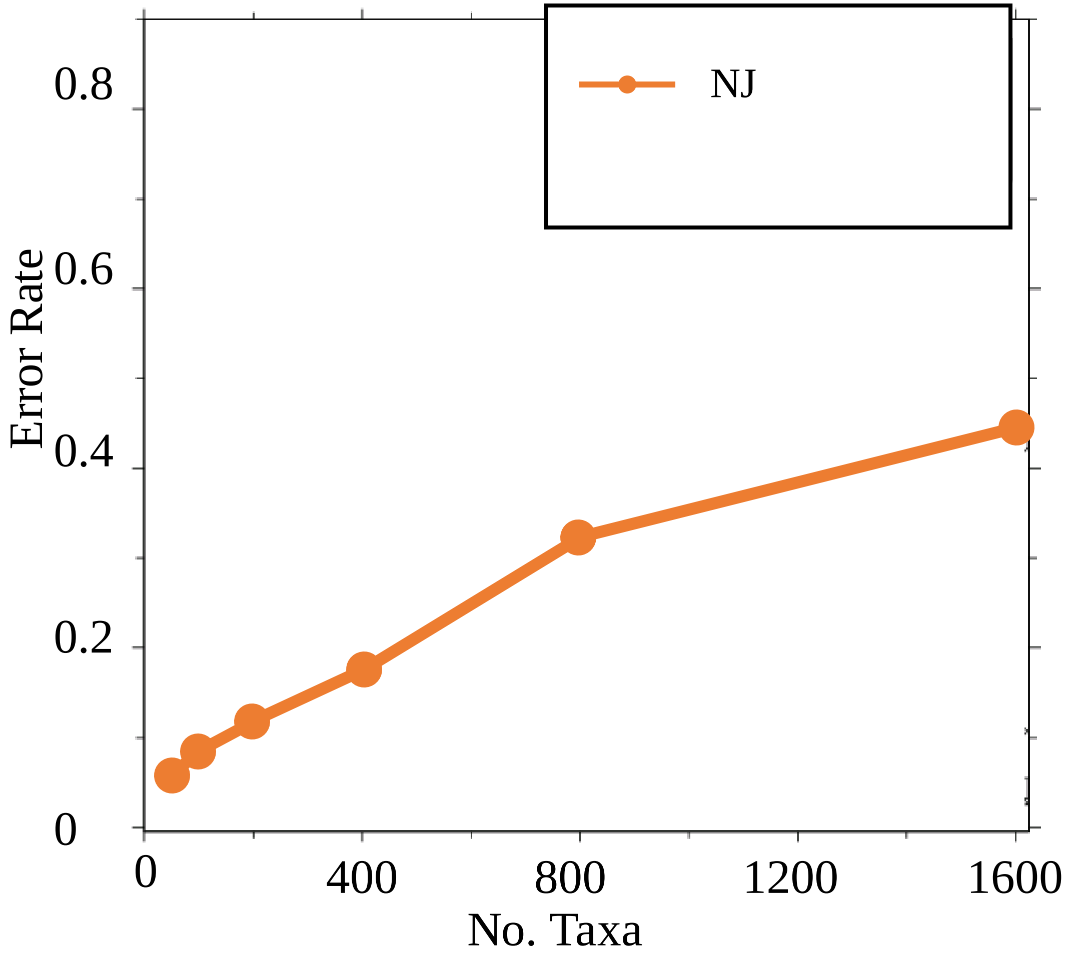
Theorem: If M is exponentially converging, then its DCM-boosted version is AFC.

# NJ is exp. convg., DCM-NJ is AFC



Corollary: NJ is exponentially converging, and DCM-NJ is AFC.

# NJ has high error on large diameter trees

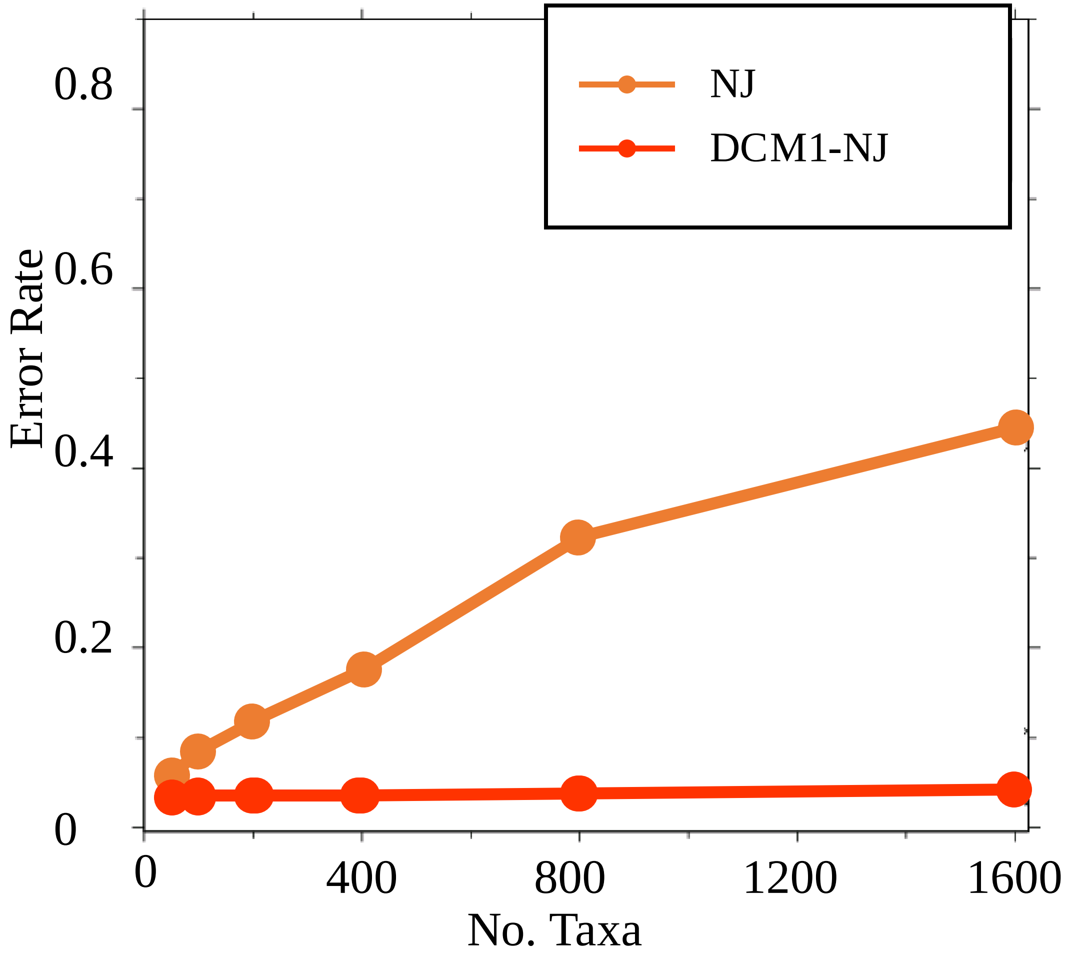


- Simulation study based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.
- Error rates reflect proportion of incorrect edges in inferred trees.

*[Nakhleh et al. ISMB 2001]*

# DCM1-boosting distance-based methods

*[Nakhleh et al. ISMB 2001]*



Theorem (Warnow et al., SODA 2001): DCM1-NJ converges to the true tree from polynomial length sequences

# Are we done? Unfortunately, no.

# 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
- Empirical (based on heuristics) – using **RAxML** (leading ML heuristic)
  - Outstanding accuracy on simulated data (e.g., better than DCM-NJ)
  - Challenging on large datasets (best methods can take CPU years or fail to run on large datasets)

# DCM-NJ vs. Maximum Likelihood

- DCM-NJ is polynomial time and scales to large datasets
- Maximum likelihood is an NP-hard optimization problem and its heuristics can be slow
- In simulation, *Maximum Likelihood is usually more accurate than DCM-NJ*

Question: Are there other Divide-and-Conquer approaches that improve maximum likelihood scalability and speed?

# Divide-and-conquer using supertree methods

- Given input dataset
  - Divide into overlapping subsets
  - Construct trees on subsets
  - Combine the overlapping subset trees using a supertree method
- Studied most in comparison to maximum parsimony and maximum likelihood on sequence alignments

# Divide-and-conquer using supertree methods

- Examples of standard supertree methods:
  - Robinson-Foulds Supertrees (minimize total RF distance to source trees)
  - Matrix Representation using Parsimony (MRP): represent the input source trees as a matrix with 0,1,?, and then solve for maximum parsimony
  - Matrix Representation using Likelihood (MRL): construct same matrix, but then run solve for maximum likelihood
- All NP-hard problems, so heuristics are used
- Excellent accuracy but slow and not scalable

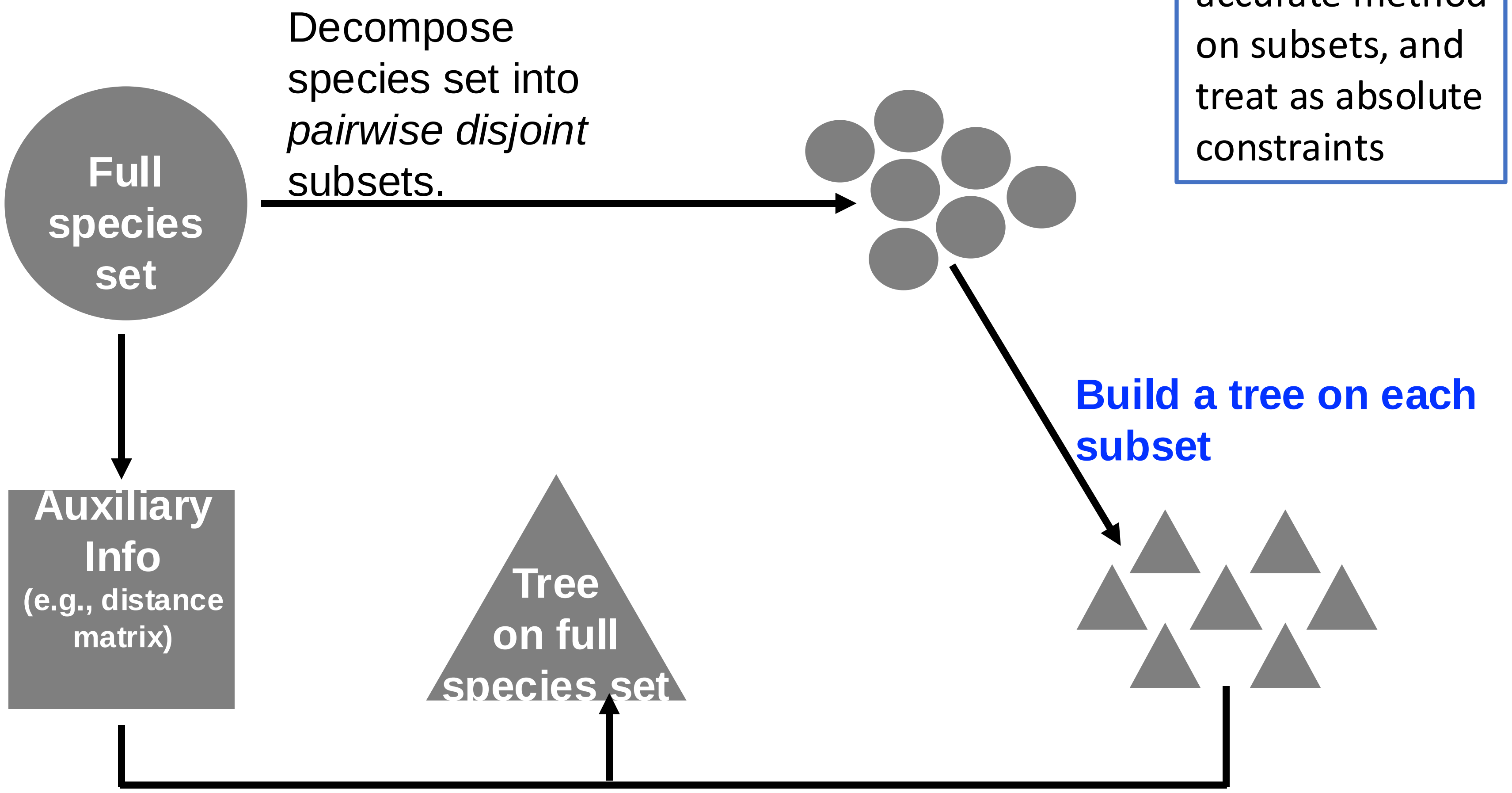
Summary: insufficient scalability/accuracy for large-scale phylogeny

# Part II: Divide-and-conquer using DTMs

# Divide-and-Conquer using Disjoint Tree Mergers

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

Erin Molloy, Introduced this approach



Full species set

Decompose species set into *pairwise disjoint* subsets.

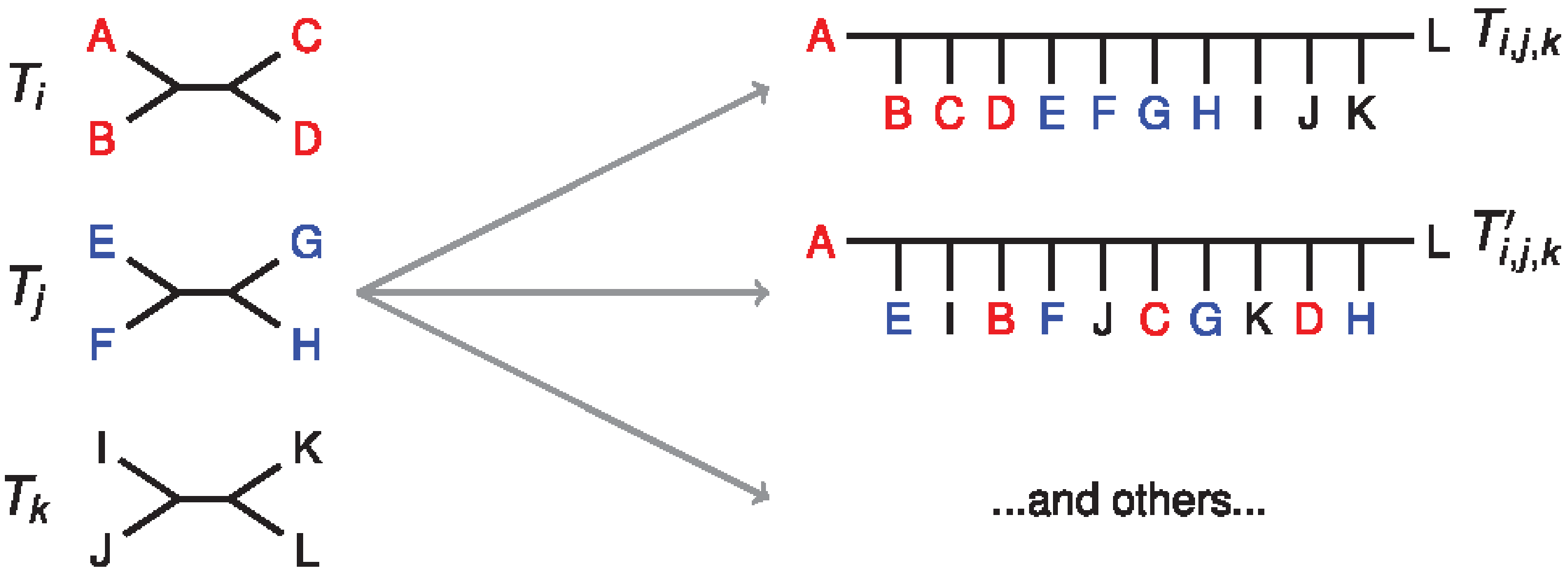
**Build a tree on each subset**

Auxiliary Info (e.g., distance matrix)

Tree on full species set

**Compute tree on entire set of species using "Disjoint Tree Merger" method**

# DTMs Merge Subset Trees



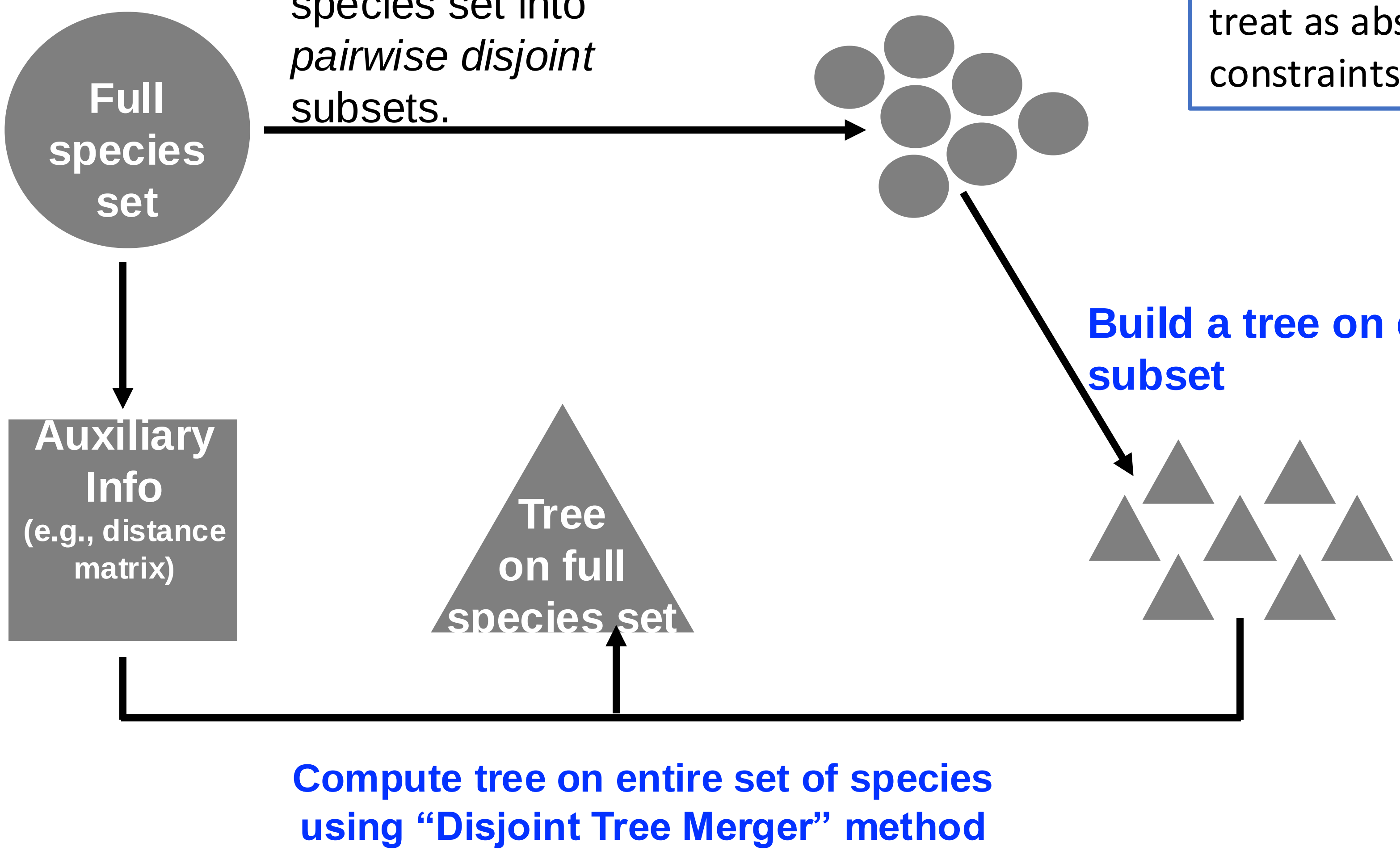
Notes:

- 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

OXFORD
UNIVERSITY PRESS

# Divide-and-Conquer using Disjoint Tree Mergers



Decompose species set into *pairwise disjoint* subsets.

**Build a tree on each subset**

**Compute tree on entire set of species using "Disjoint Tree Merger" method**

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

Erin Molloy, Introduced this approach

Theorem:
If the subtree method is statistically consistent, then many DTM methods are **statistically consistent.**

# Disjoint Tree Mergers (DTMs)

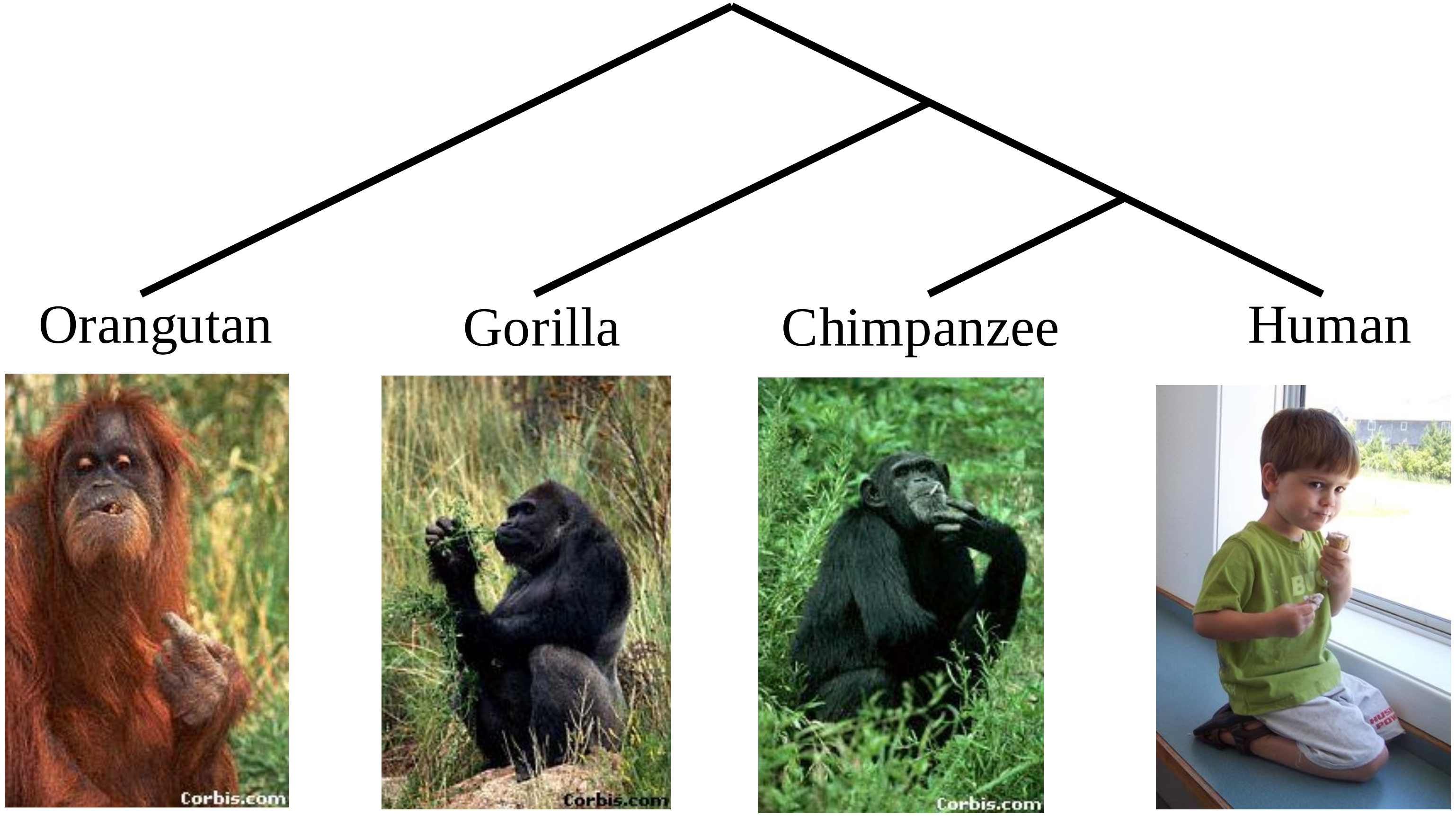
- NJMerge (Molloy and Warnow, Alg Mol Biol 2019)
- TreeMerge (Molloy and Warnow, Bioinf 2019)
- Constrained-INC (Zhang, Rao, and Warnow, Alg Mol Biol 2019)
  - The only one that allows full blending
- Guide Tree Merger (Smirnov and Warnow, 2020)
  - Does not allow blending

# 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

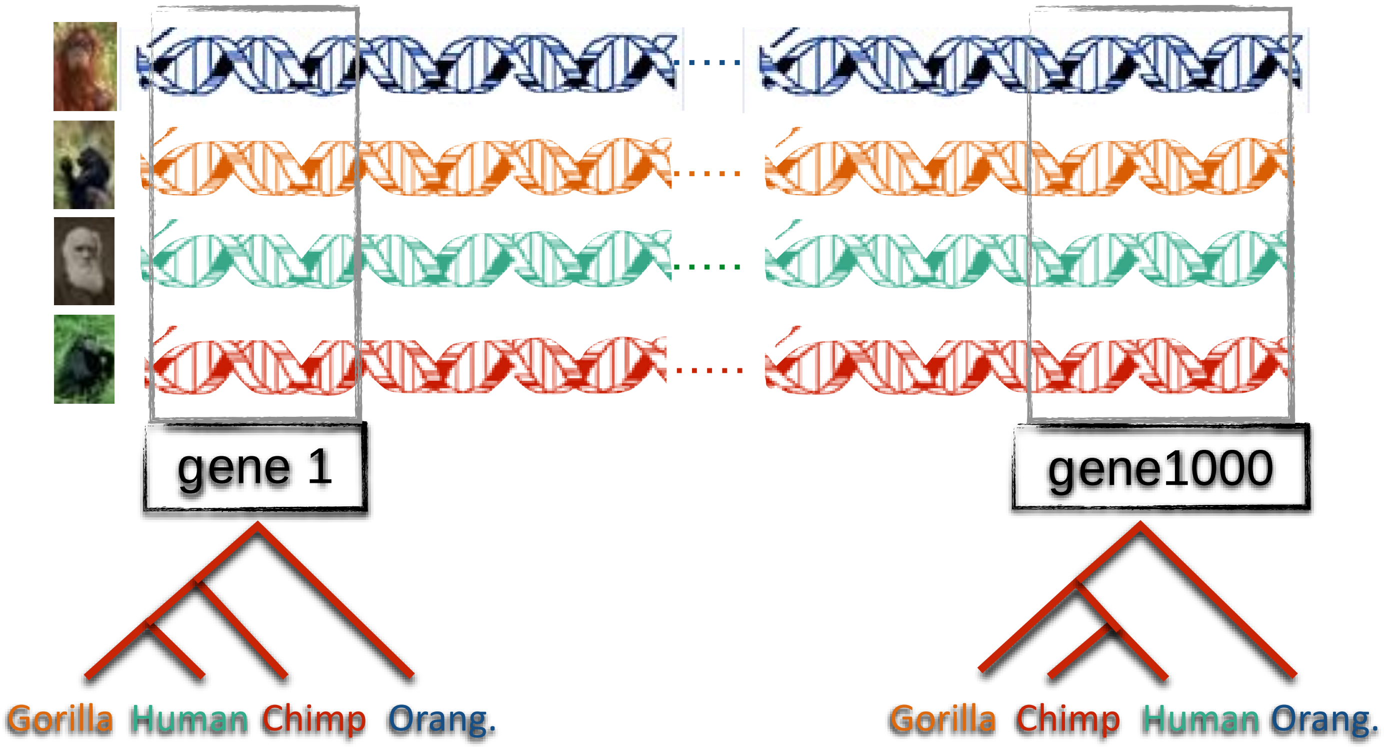
# Species Tree Estimation



Orangutan Gorilla Chimpanzee Human

*From the Tree of the Life Website,*
*University of Arizona*

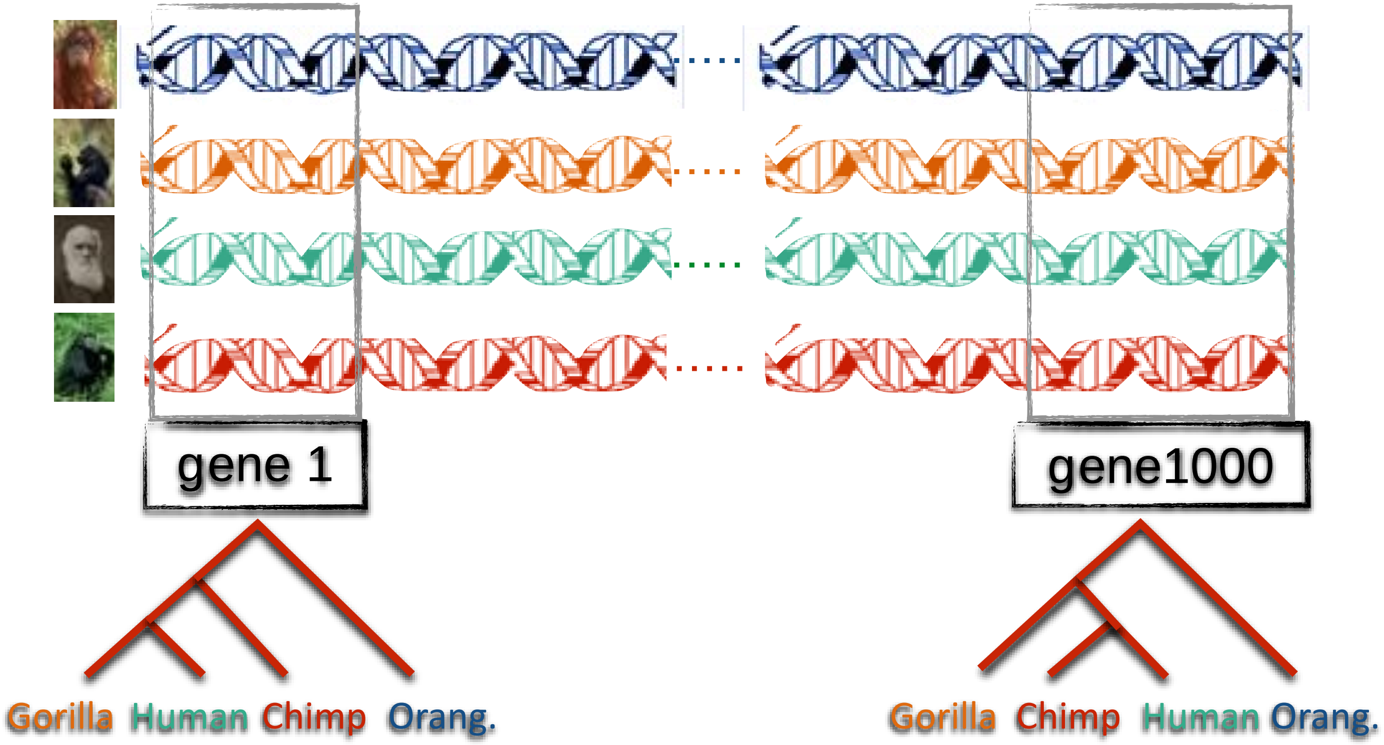
# Gene tree discordance



Multiple causes for discord, including

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

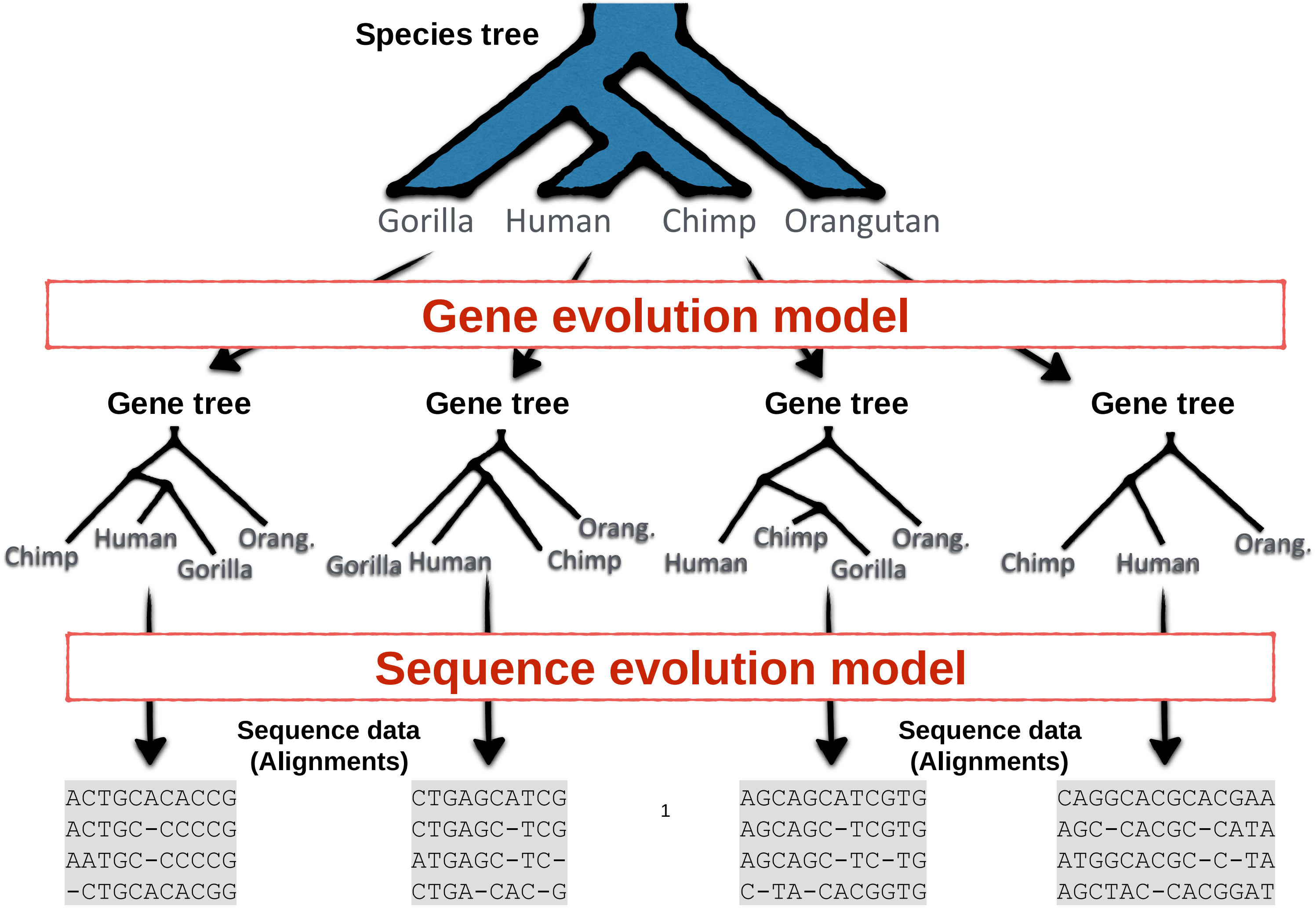
# Gene tree discordance



Multiple causes for discord, including

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

# MSC+GTR Hierarchical Model



Gene evolution model

Sequence evolution model

Sequence data (Alignments)

Sequence data (Alignments)

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

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

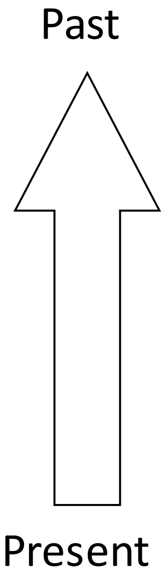
1

```
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
```

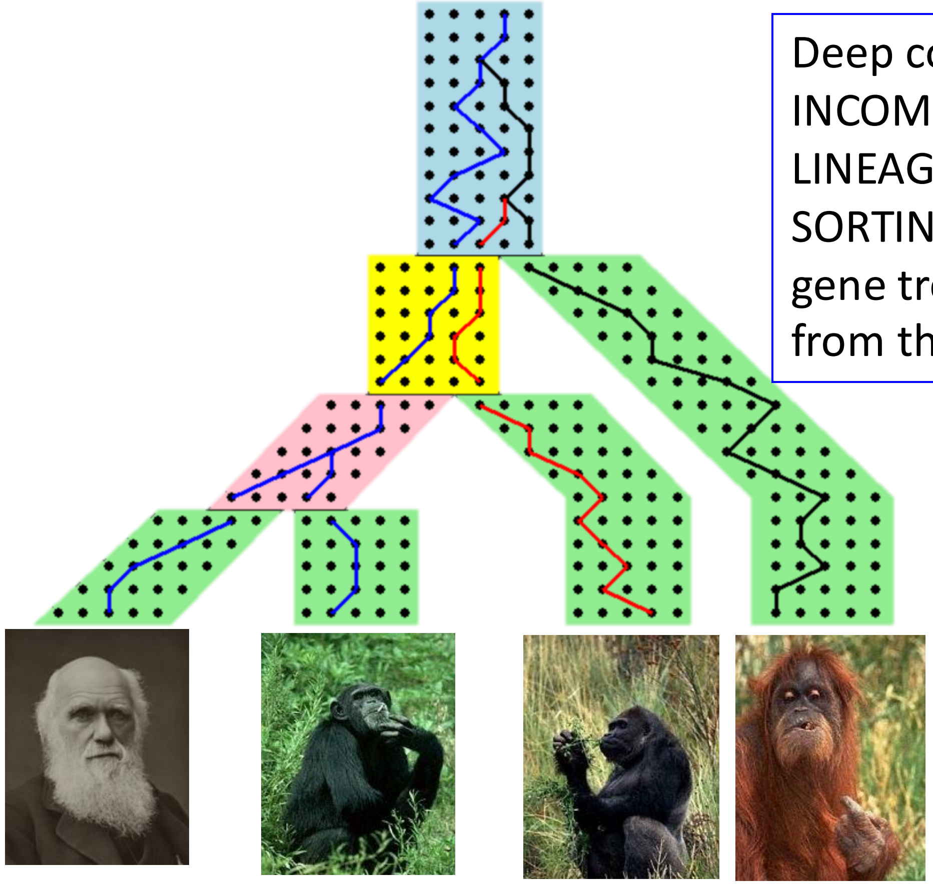
```
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT
```

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

# Gene trees inside the species tree (Coalescent Process)



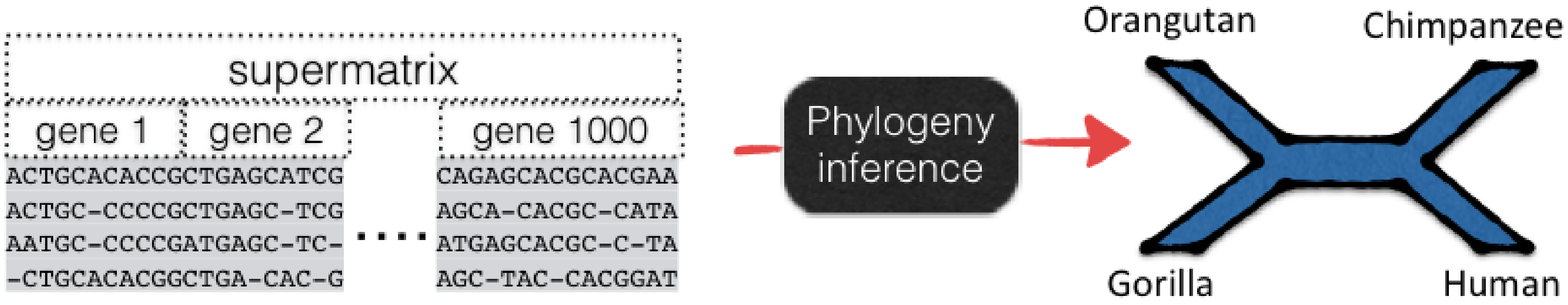
Courtesy James Degnan



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.

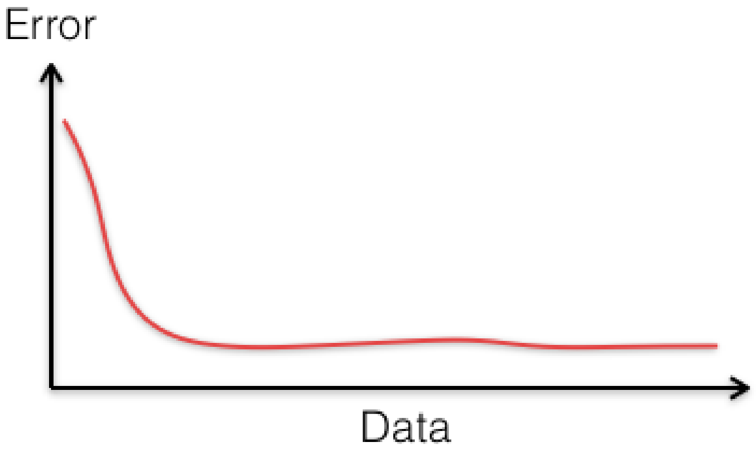
# Traditional approach: concatenation



```
ACTGCACACCGCTGAGCATCG      CAGAGCACGCACGAA
ACTGC-CCCCGCTGAGC-TCG      AGCA-CACGC-CATA
AATGC-CCCCGATGAGC-TC- .... ATGAGCACGC-C-TA
-CTGCACACGGCTGA-CAC-G      AGC-TAC-CACGGAT
```

- 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]



# ASTRAL

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

- Optimization Problem (NP-Hard):

Find the species tree with <u>the maximum number of induced quartet trees</u> shared with the collection of input gene trees

ASTRAL runs in $O(|X|^2kn)$ where there are n species and k genes, and X is the set of allowed bipartitions

Set of quartet trees induced by T

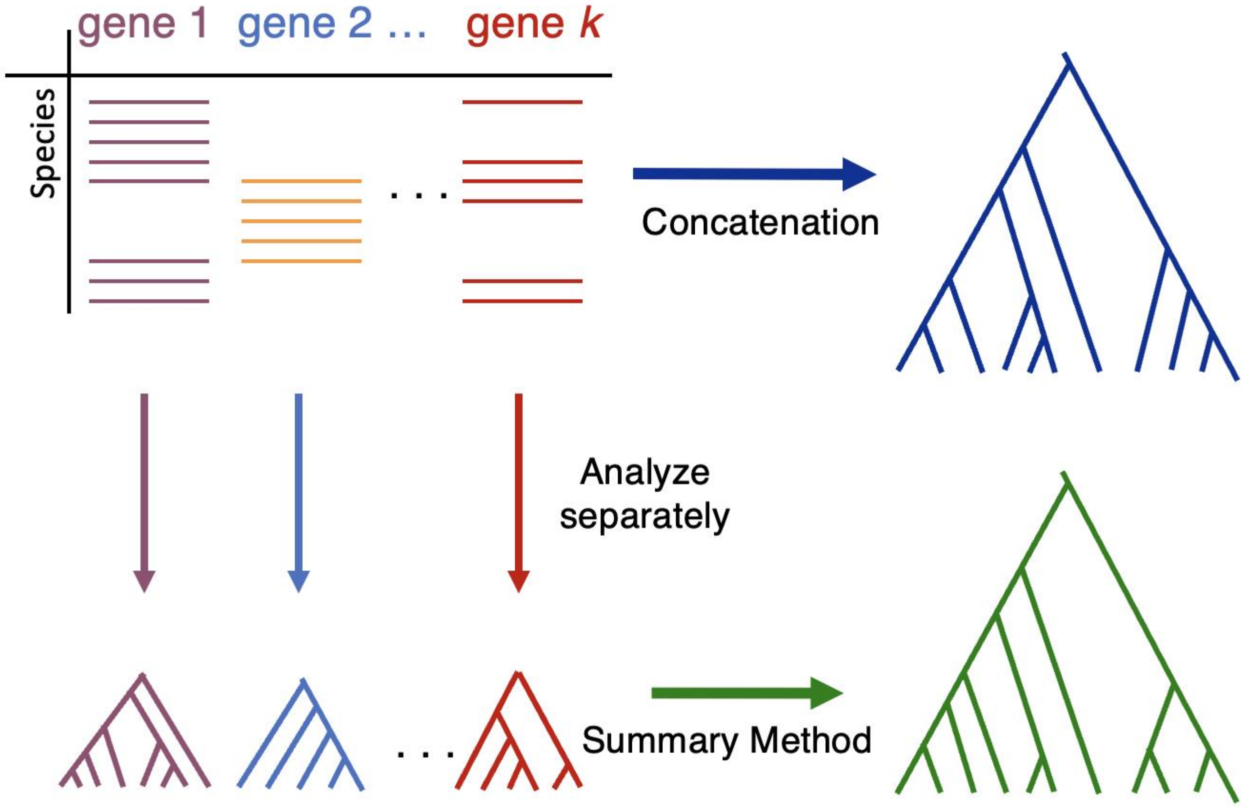
$$Score(T) = \sum_{t \in \mathcal{T}} |Q(T) \cap Q(t)|$$

a gene tree

all input gene trees

- **Theorem**: <u>Statistically consistent</u> under the multi-species coalescent model when solved exactly

# Main Approaches for Species Tree Estimation

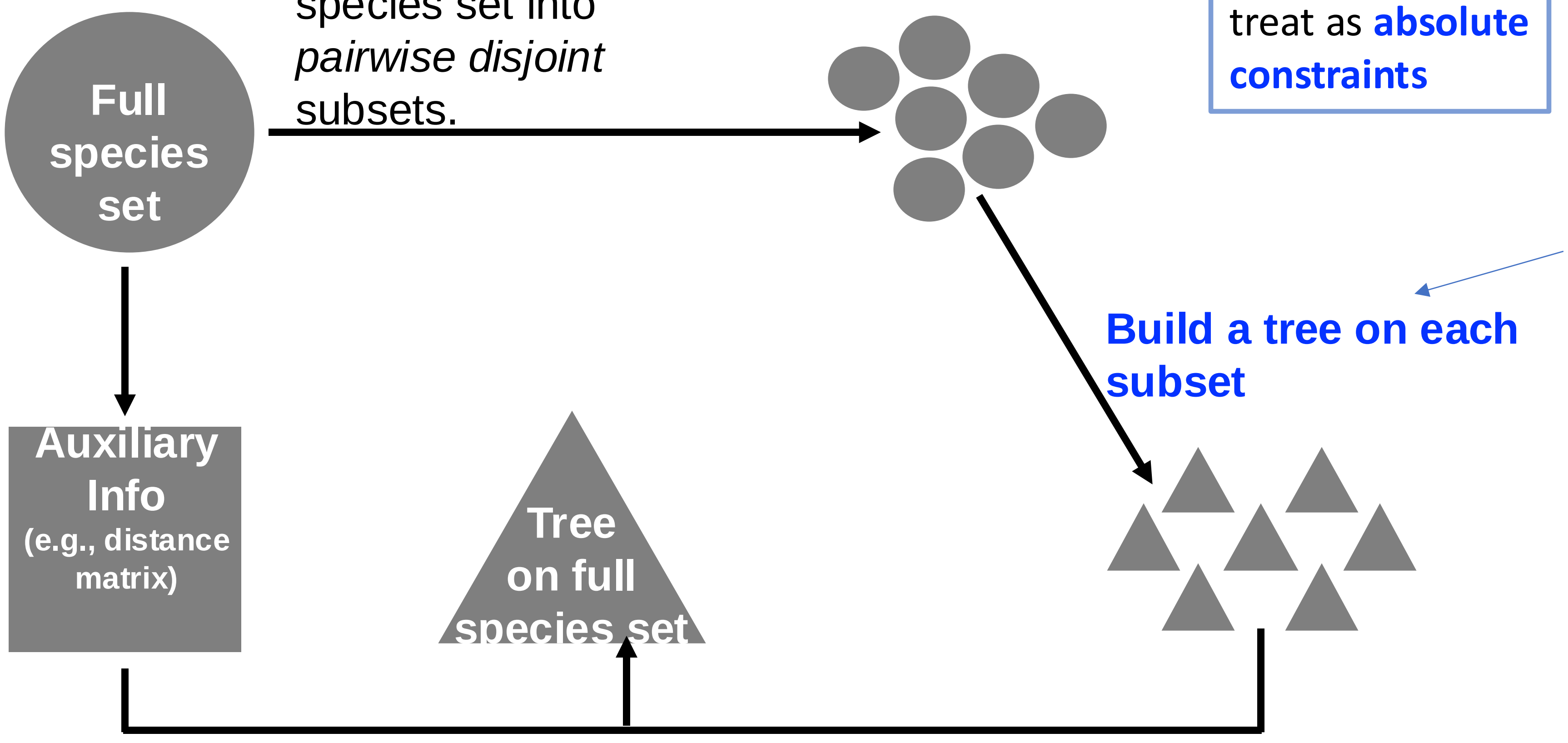


e.g., RAxML

e.g., ASTRAL

# Divide-and-Conquer Gene Tree Estimation

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


Full species set
Decompose species set into *pairwise disjoint* subsets.
Auxiliary Info (e.g., distance matrix)
Tree on full species set
Build a tree on each subset

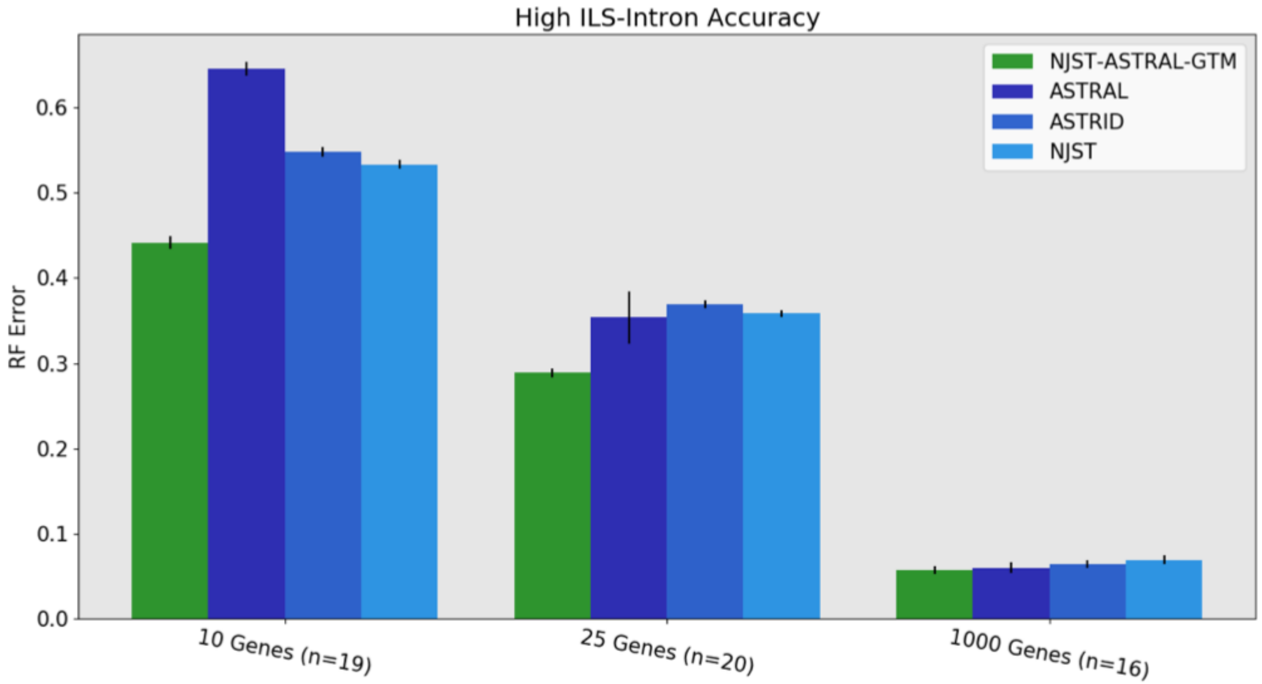
ASTRAL

NJst for guide tree

**Compute tree on entire set of species using "Disjoint Tree Merger" method**

Guide Tree Merger

# GTM+ASTRAL: faster and more accurate than ASTRAL



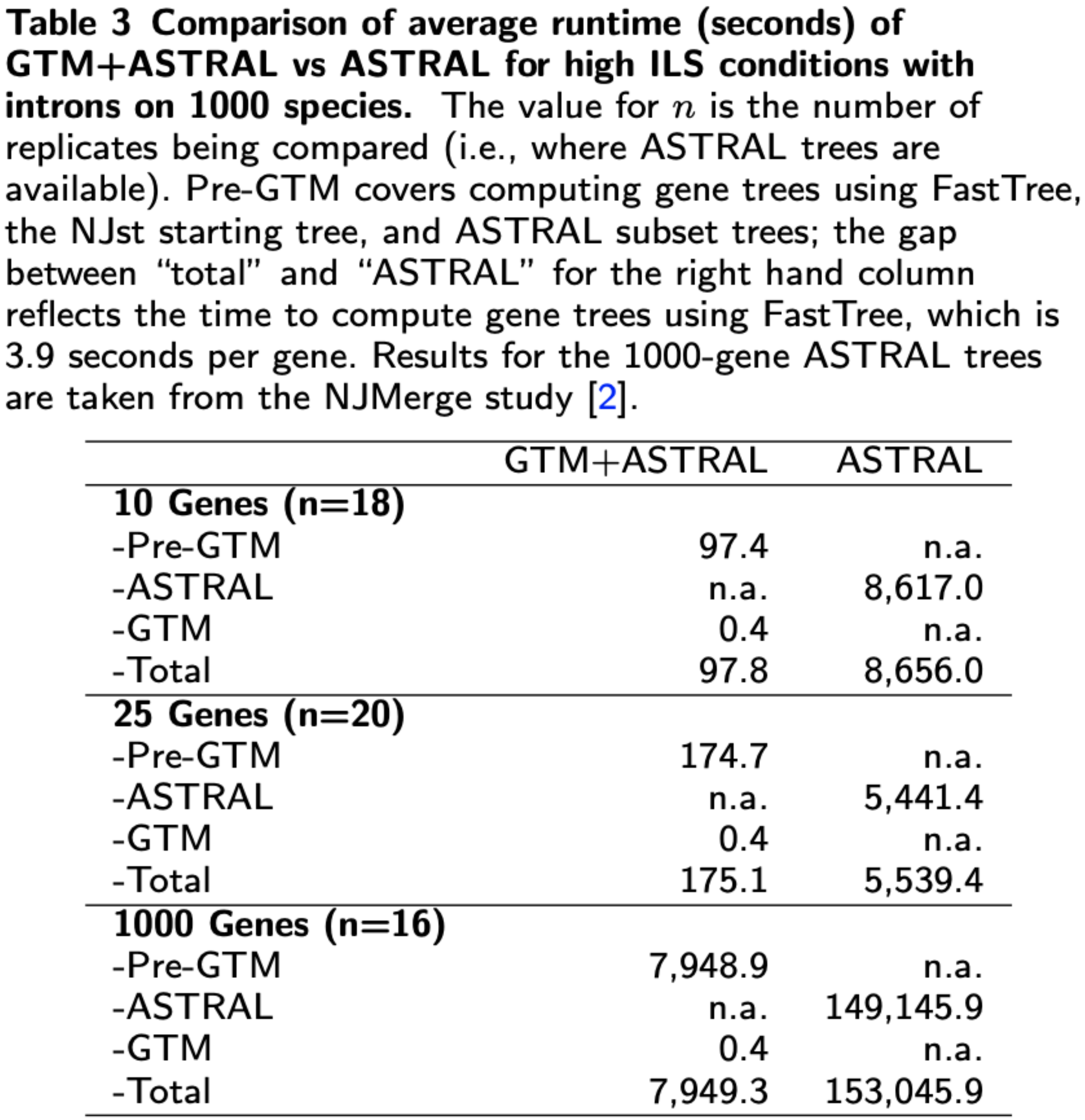


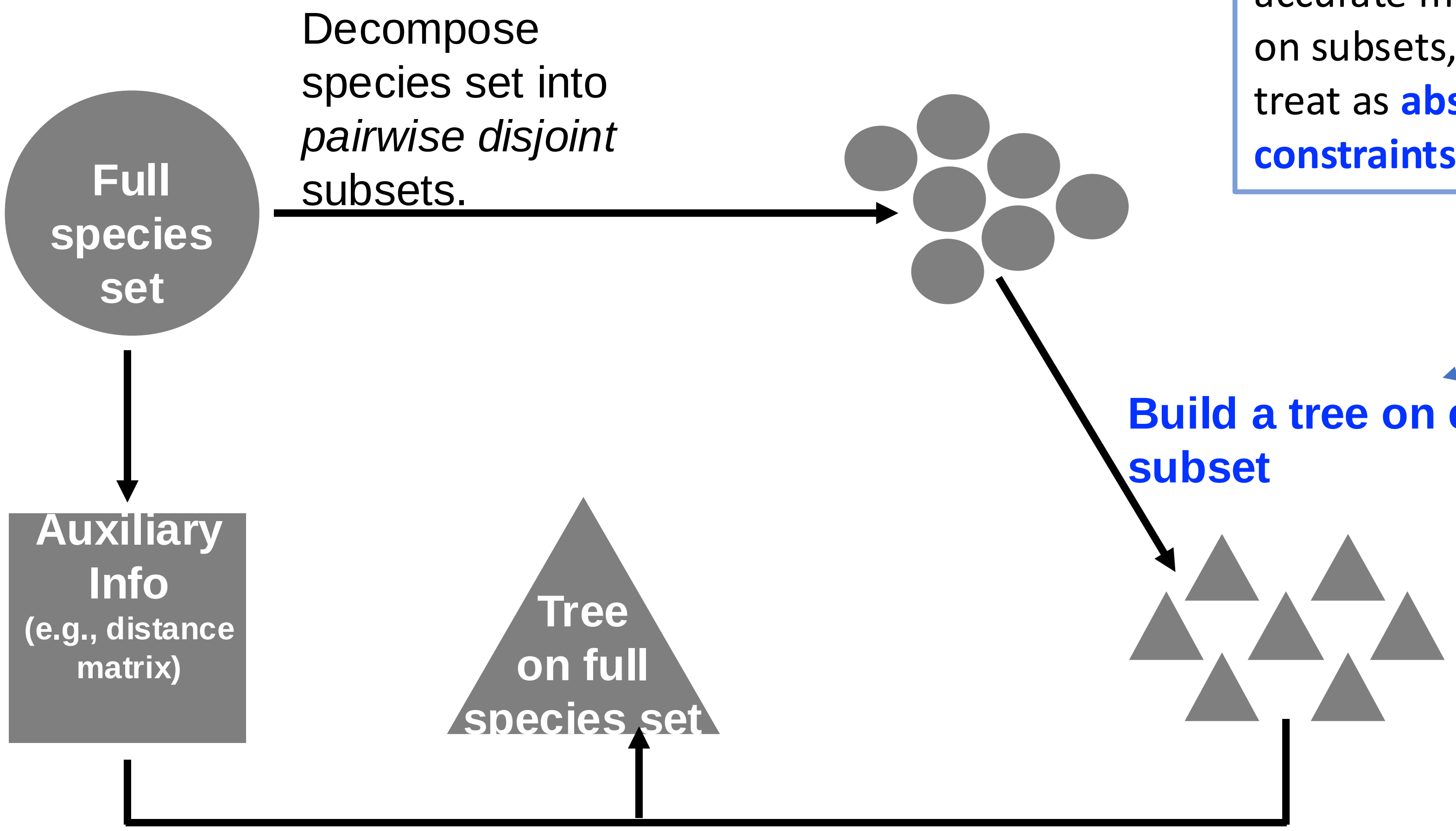
**Table 3 Comparison of average runtime (seconds) of GTM+ASTRAL vs ASTRAL for high ILS conditions with introns on 1000 species.** The value for $n$ is the number of replicates being compared (i.e., where ASTRAL trees are available). Pre-GTM covers computing gene trees using FastTree, the NJst starting 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 |

# What about scaling Maximum Likelihood?

# Divide-and-Conquer Gene Tree Estimation

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



Decompose species set into *pairwise disjoint* subsets.

Full species set

Auxiliary Info (e.g., distance matrix)

Tree on full species set

**Build a tree on each subset**

RAxML, IQ-TREE, etc

FastTree or IQ-Tree for guide tree

**Compute tree on entire set of species using "Disjoint Tree Merger" method**

Guide Tree Merger

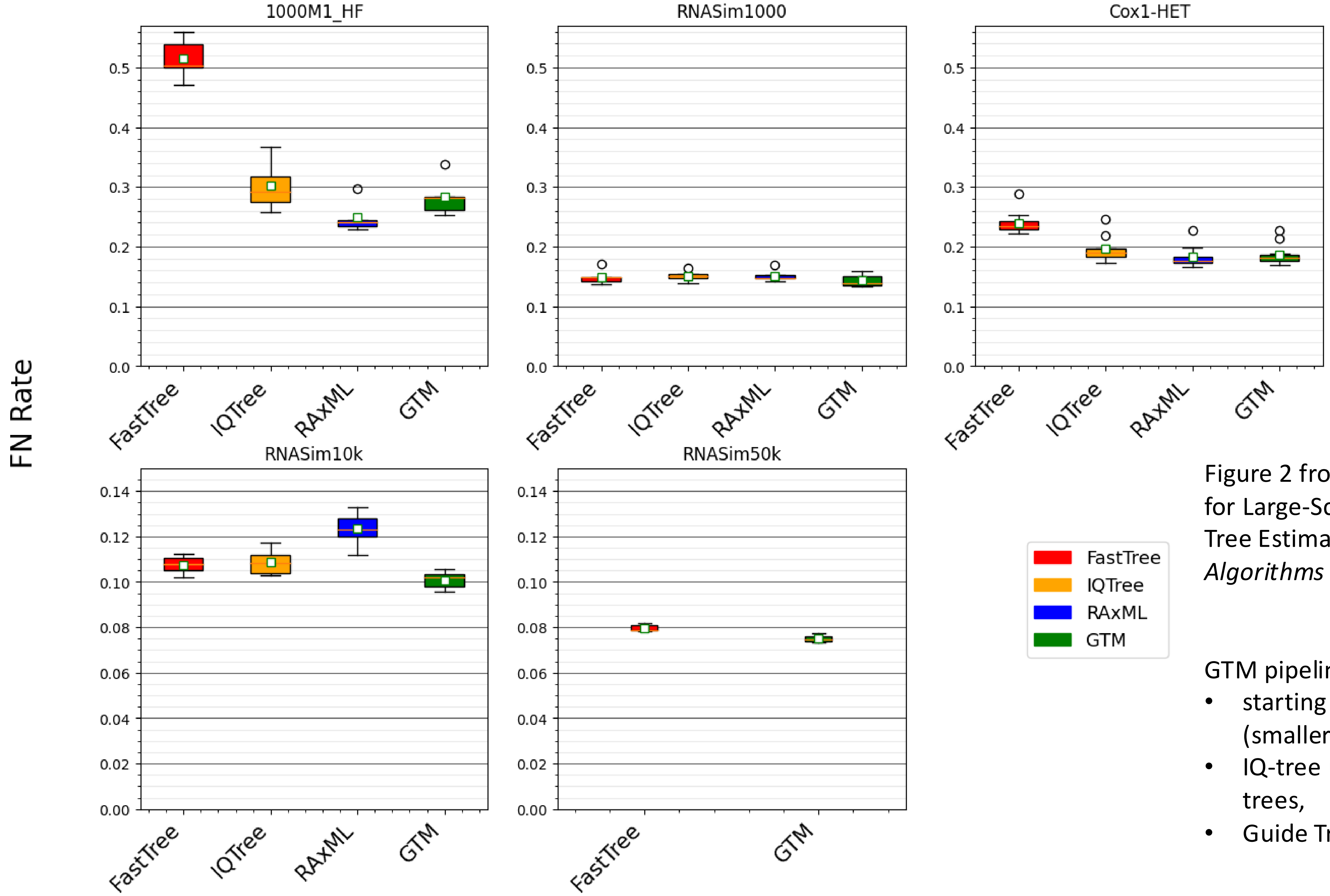
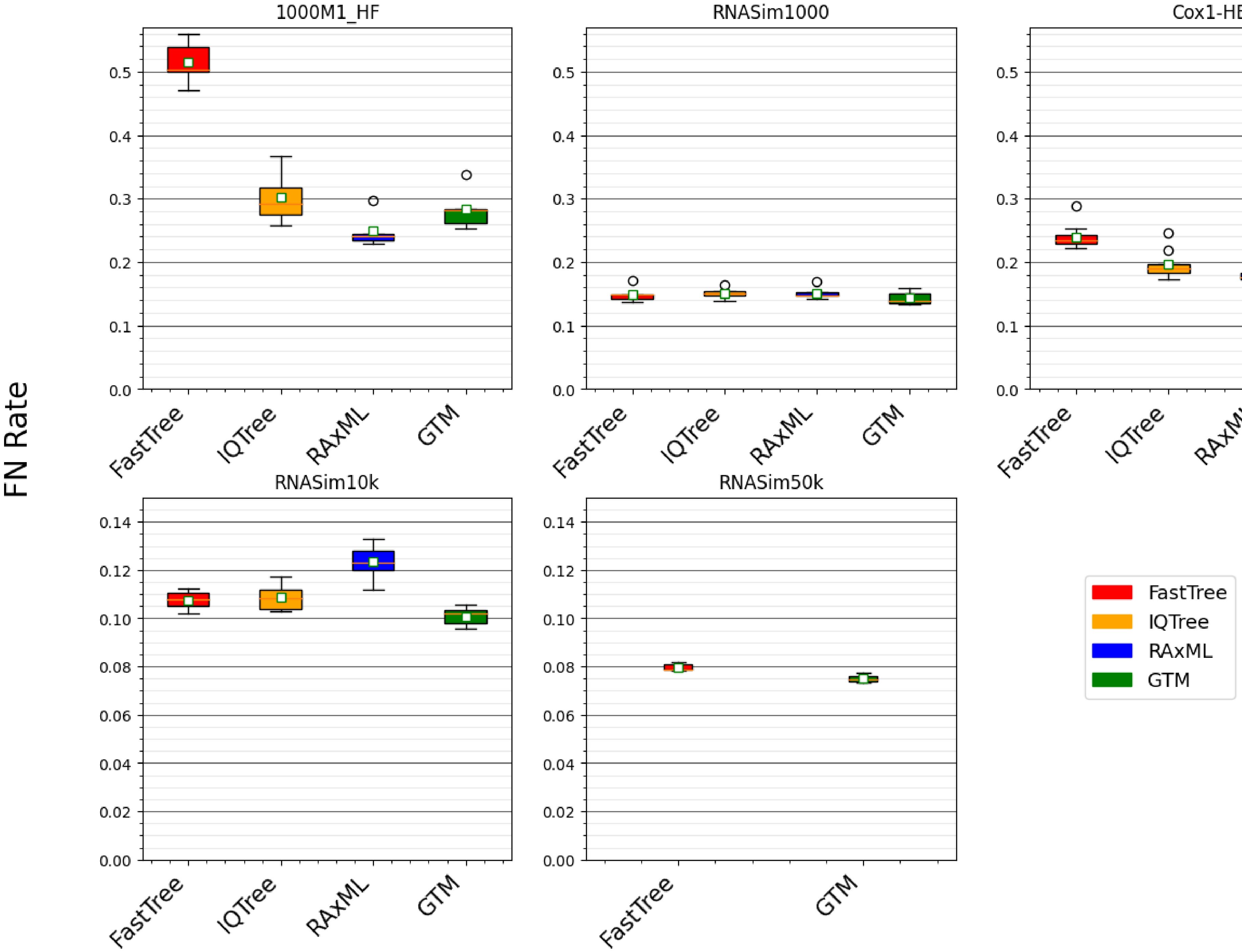


Figure 2 from "Disjoint Tree Mergers for Large-Scale Maximum Likelihood Tree Estimation", Park et al., *Algorithms 2021*

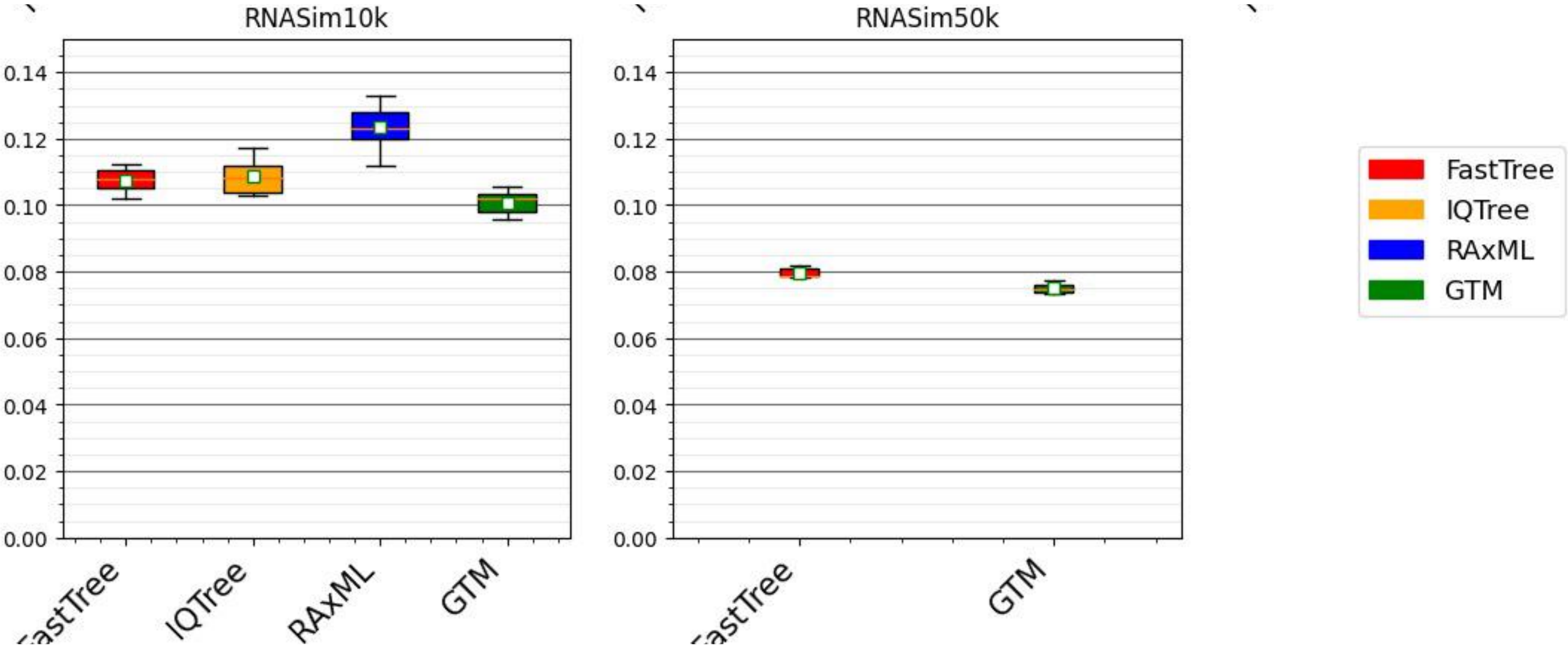
GTM pipeline:

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



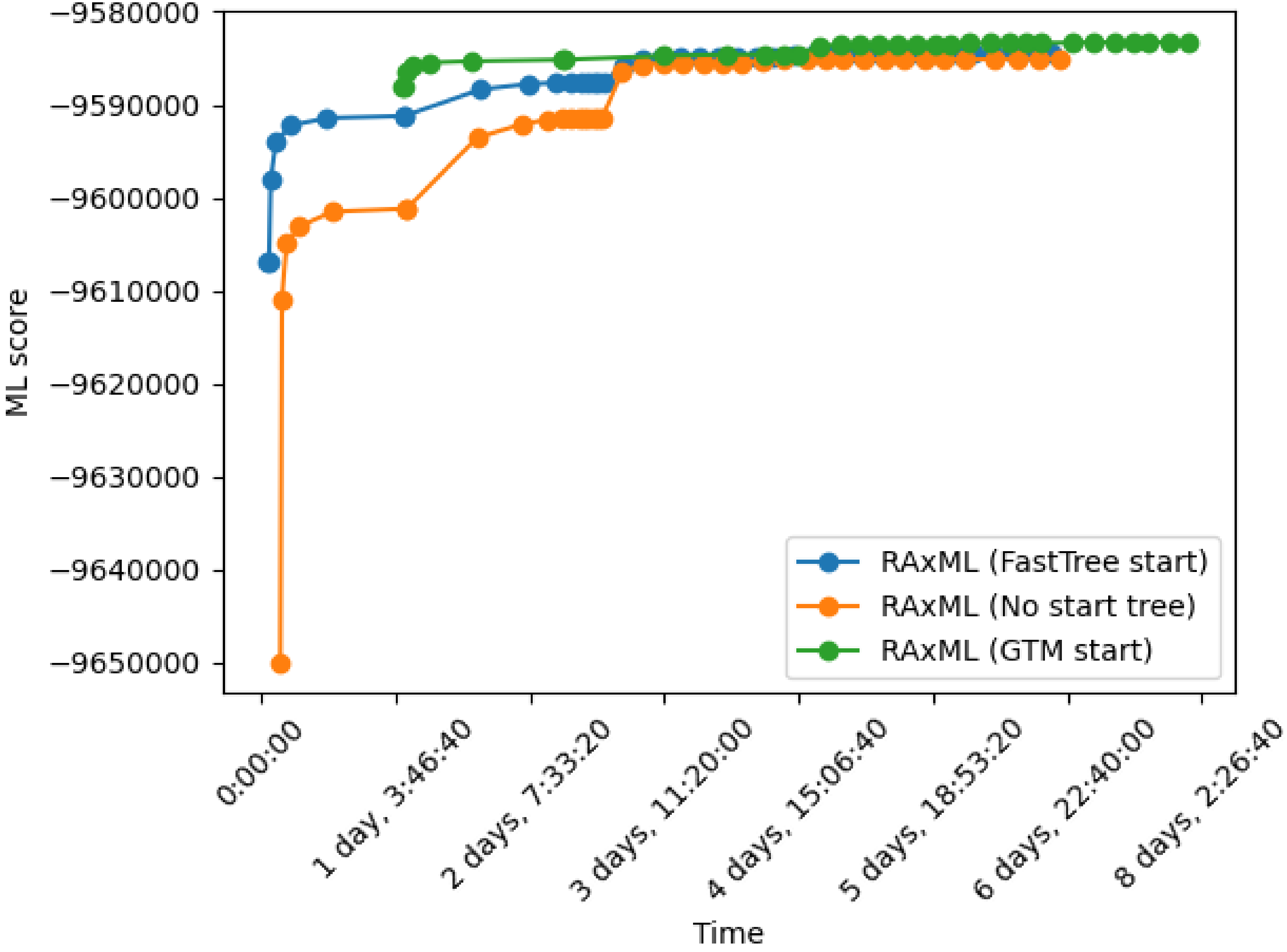
GTM-pipeline:

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



Trends

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

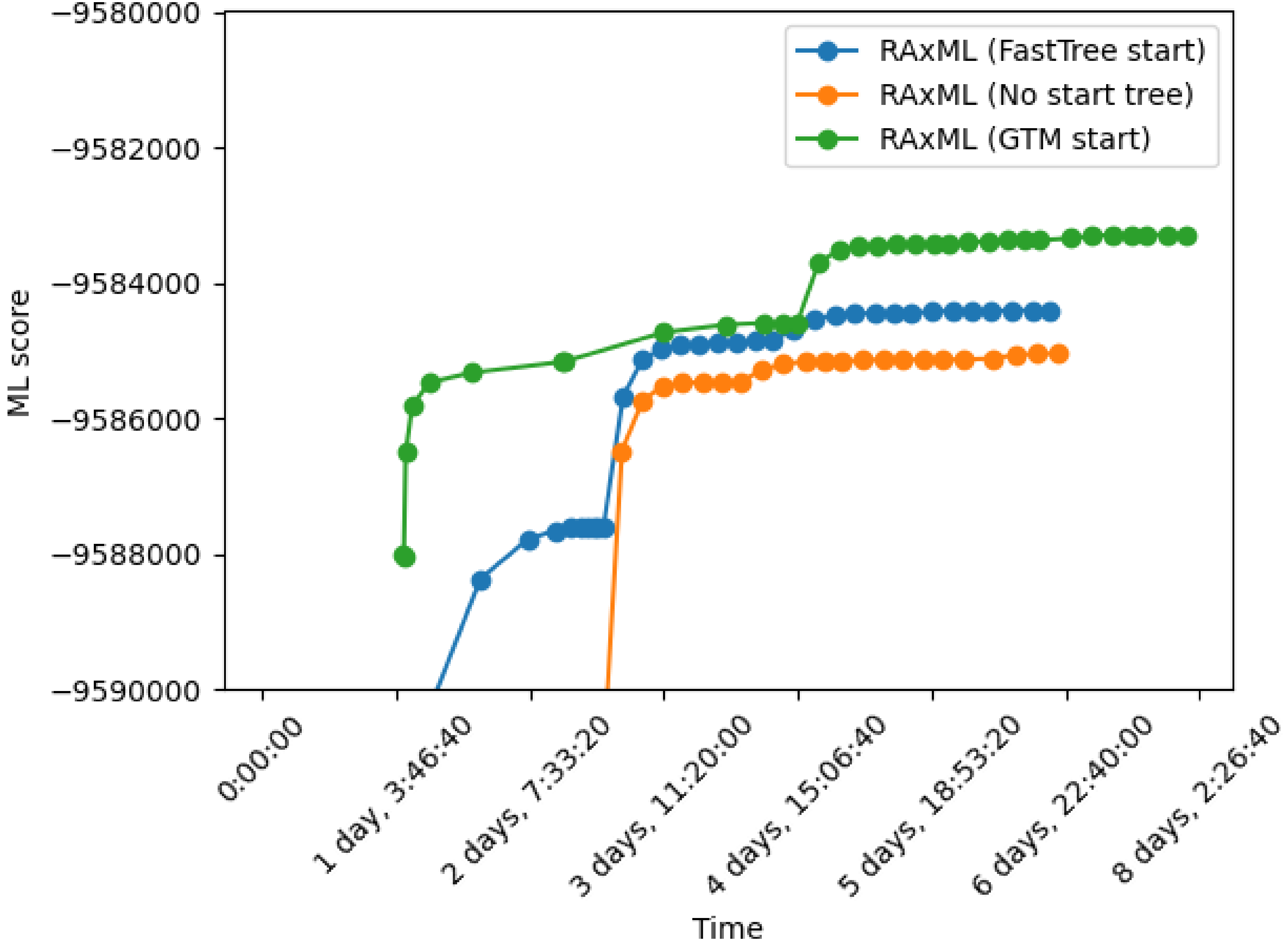


**Analysis of Kelly Williams dataset (Minhyuk Park et al., NYP)**

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!

# 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

# Disjoint Tree Mergers (summary)

- “Disjoint tree mergers” (DTMs) are generic methods, that can be used with any phylogeny estimation method (for any kind of data).
  - DTMs enable scalability to large datasets.
  - DTMs maintain statistical consistency
  - DTM-ASTRAL improves speed and accuracy compared to ASTRAL
  - Potential for improving maximum likelihood
  - 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

- Empirical:
  - Develop better Divide-and-Conquer strategies (e.g., improve on DTM)
  - Develop scalable and accurate supertree methods, and study them within divide-and-conquer pipelines.
  - Develop divide-and-conquer for phylogenetic network estimation
- Theoretical:
  - Are any divide-and-conquer pipelines AFC?
  - Can we bound error in Divide-and-Conquer pipelines analytically?
  - Develop theoretical framework for why GTM-boosting improves ASTRAL accuracy

# Resources

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

**Write to me: warnow@Illinois.edu**