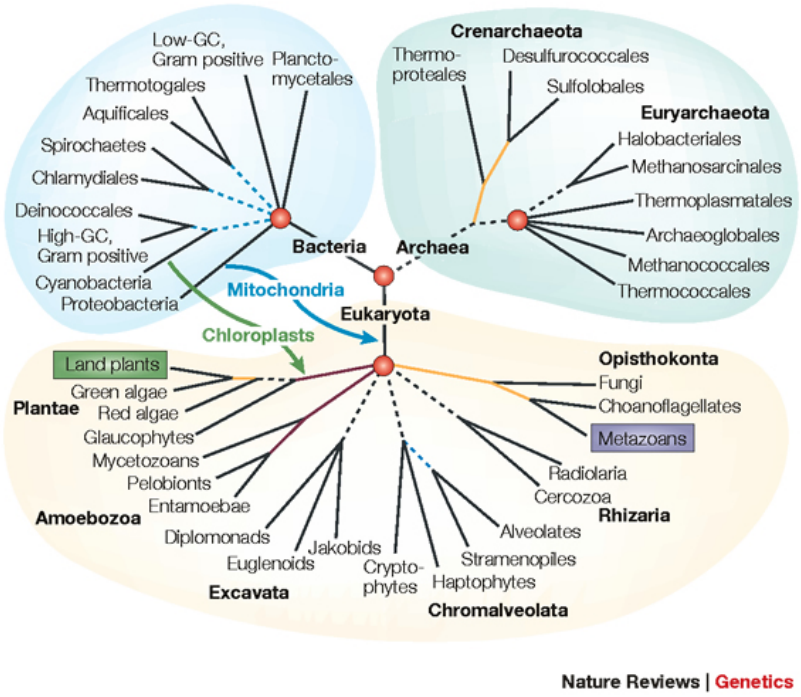
This young bird closely resembles its parents which are about a third larger. The white flecks on the face are mallophaga eggs. Most bird lice taxa found on Hoatzins are unique to this host.

Entering the genomic era of 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 <u>to understand biology</u>

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:

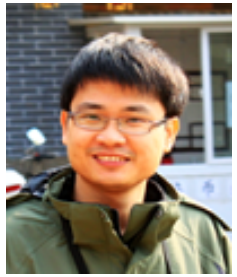
- Large alignments (and sequence length heterogeneity)
- Multi-copy genes omitted (9500 -> 400)
- Massive gene tree heterogeneity consistent with ILS

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 challenge:
Multi-copy genes omitted
Massive gene tree heterogeneity consistent with ILS

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

Phylogeny/MSA estimation: CS and Statistics

- Assume DNA sequences are generated on an unknown model tree, and try to infer the tree (and/or alignment) from the observed sequences seen at the leaves

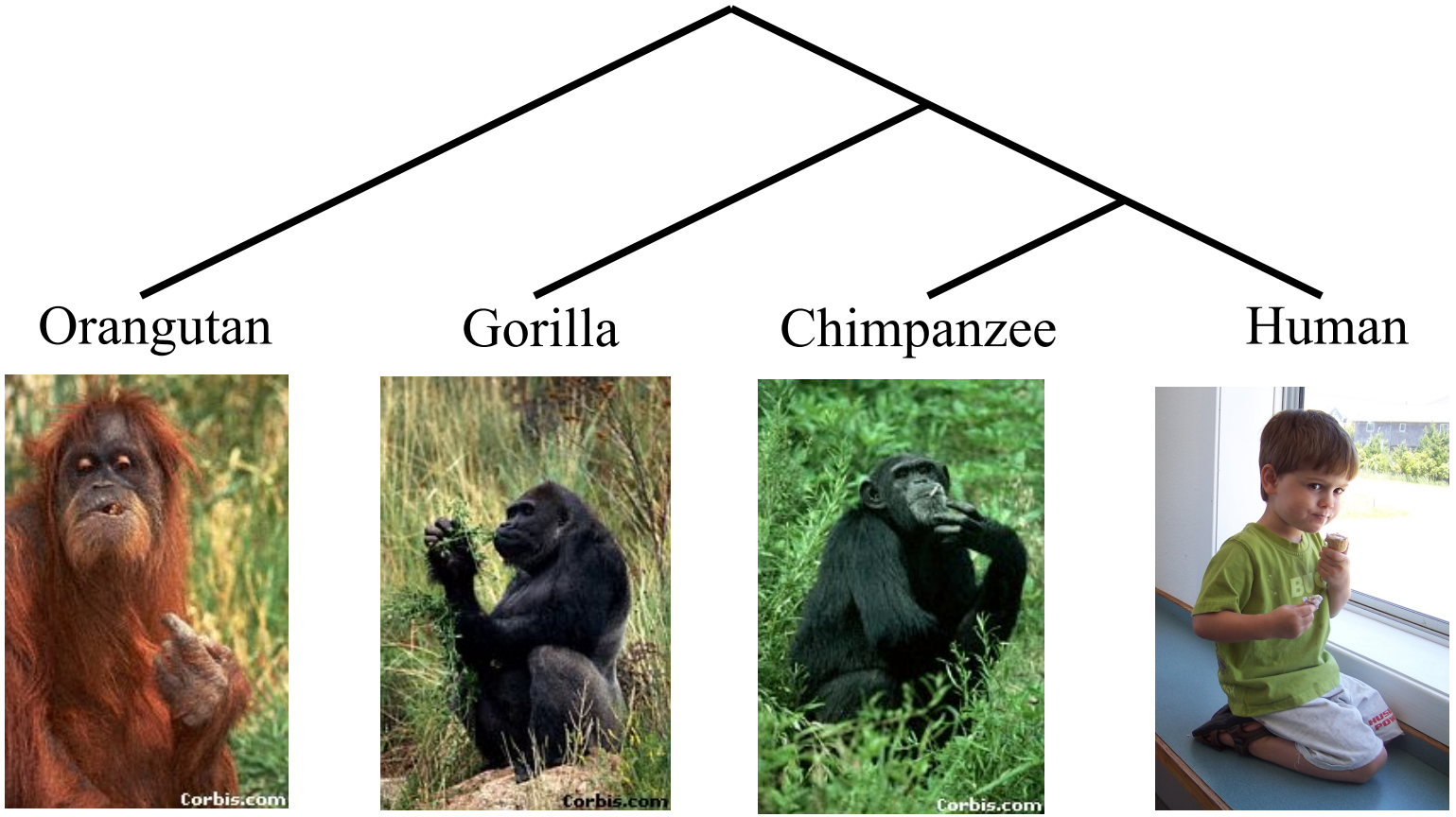
Large datasets are difficult

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

This talk

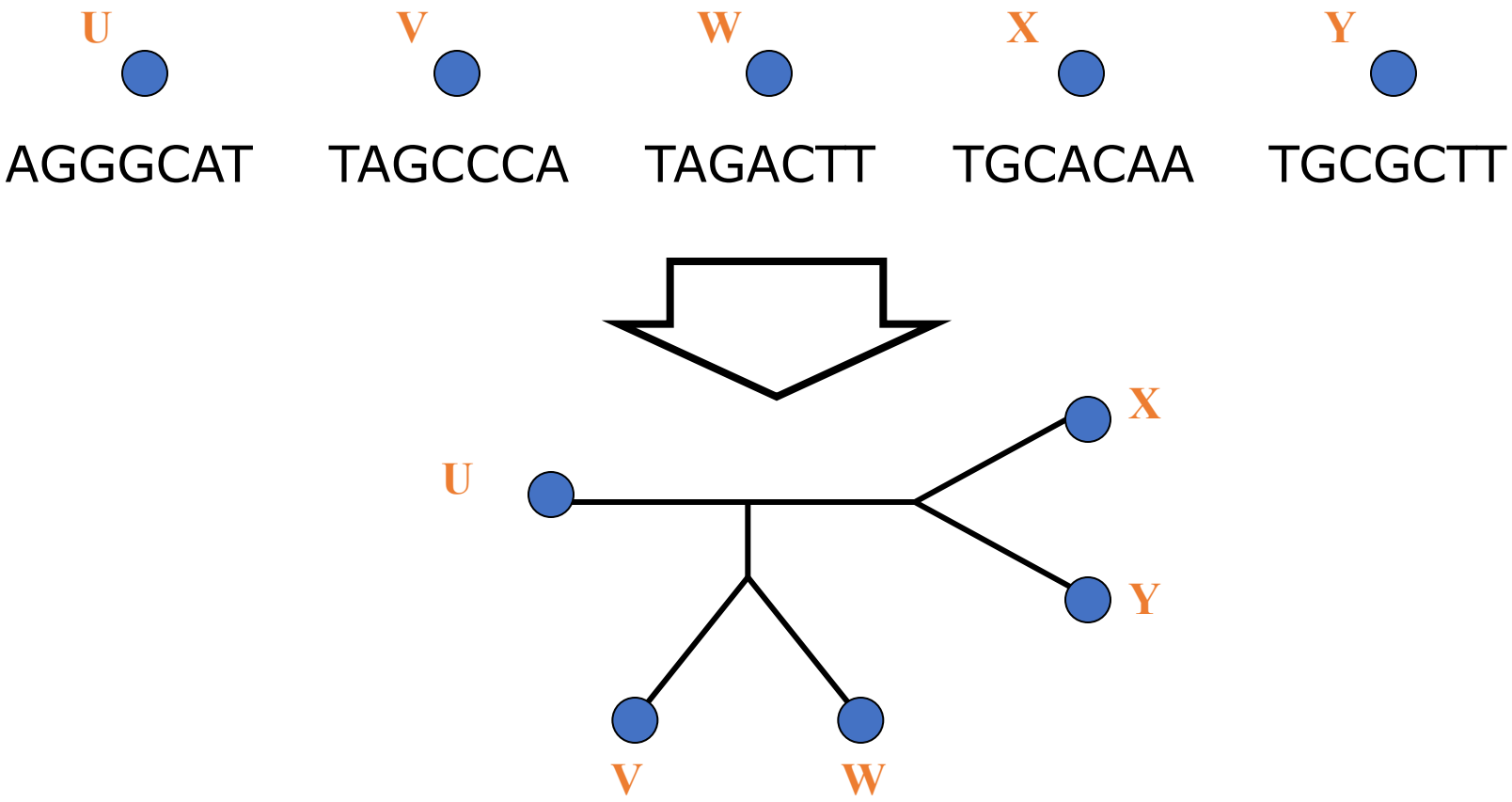
1. Estimating species trees from gene trees ("easy") – 15 minutes
2. Maximum Likelihood for estimating large gene trees (very hard) - 5 minutes
3. Multiple sequence alignment (harder, plus a conundrum) – 5 minutes

Part I: Species Tree Estimation

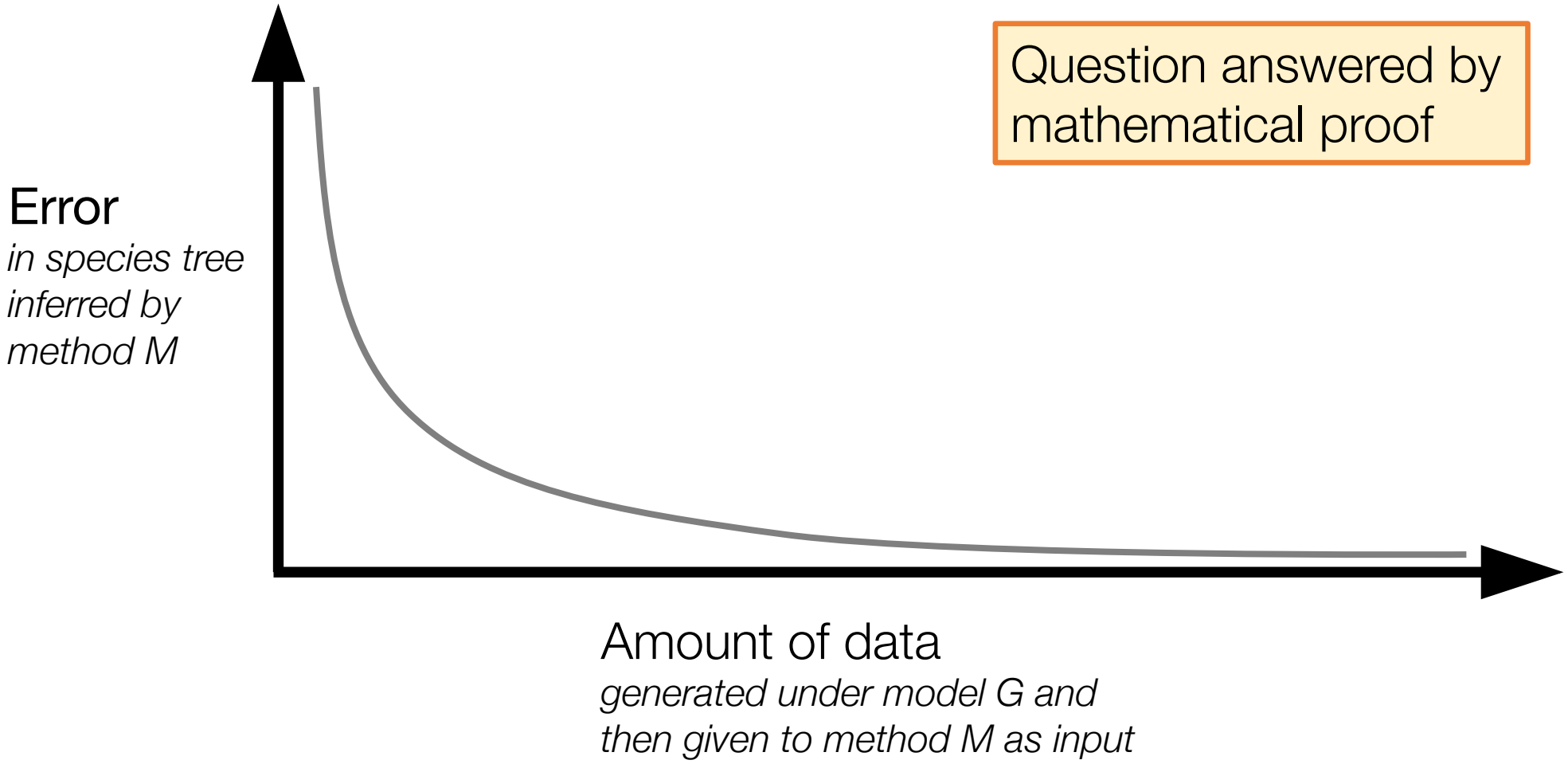


From the Tree of the Life Website,
University of Arizona

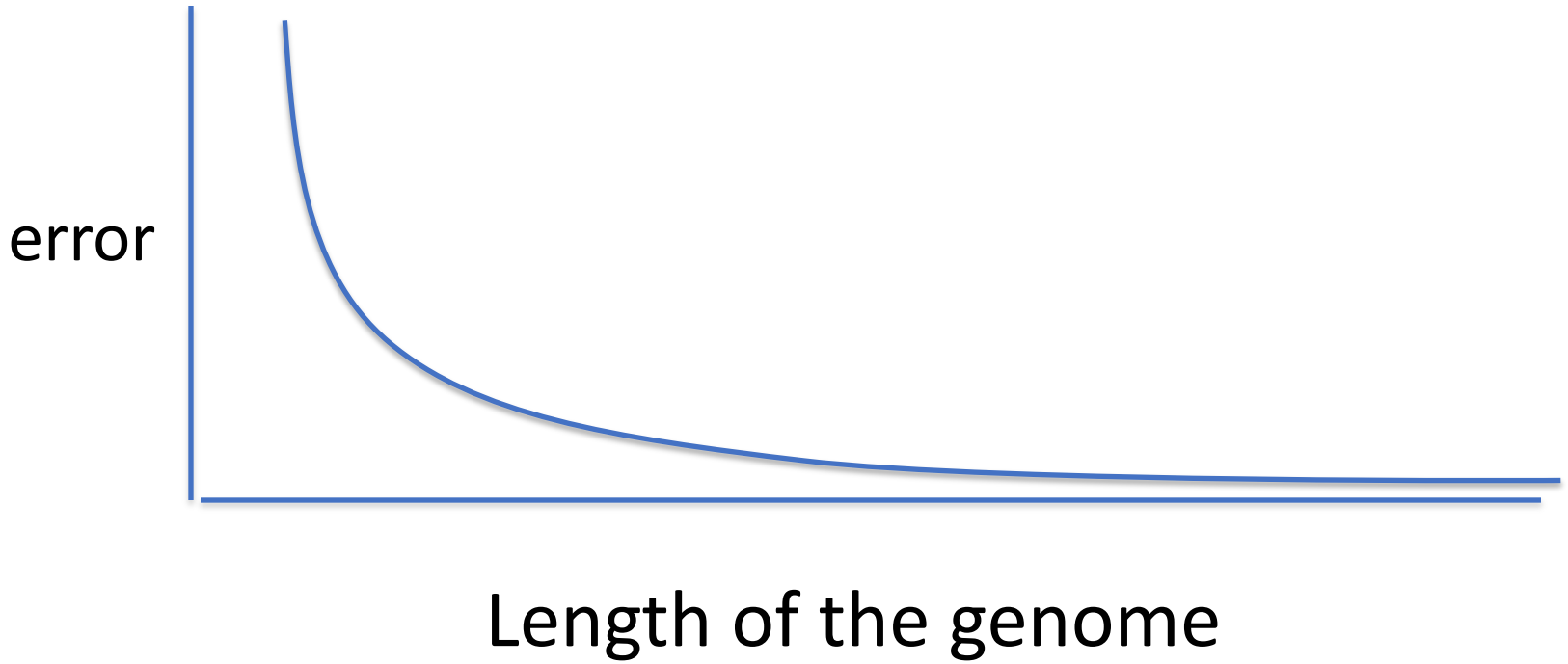
Phylogeny Estimation



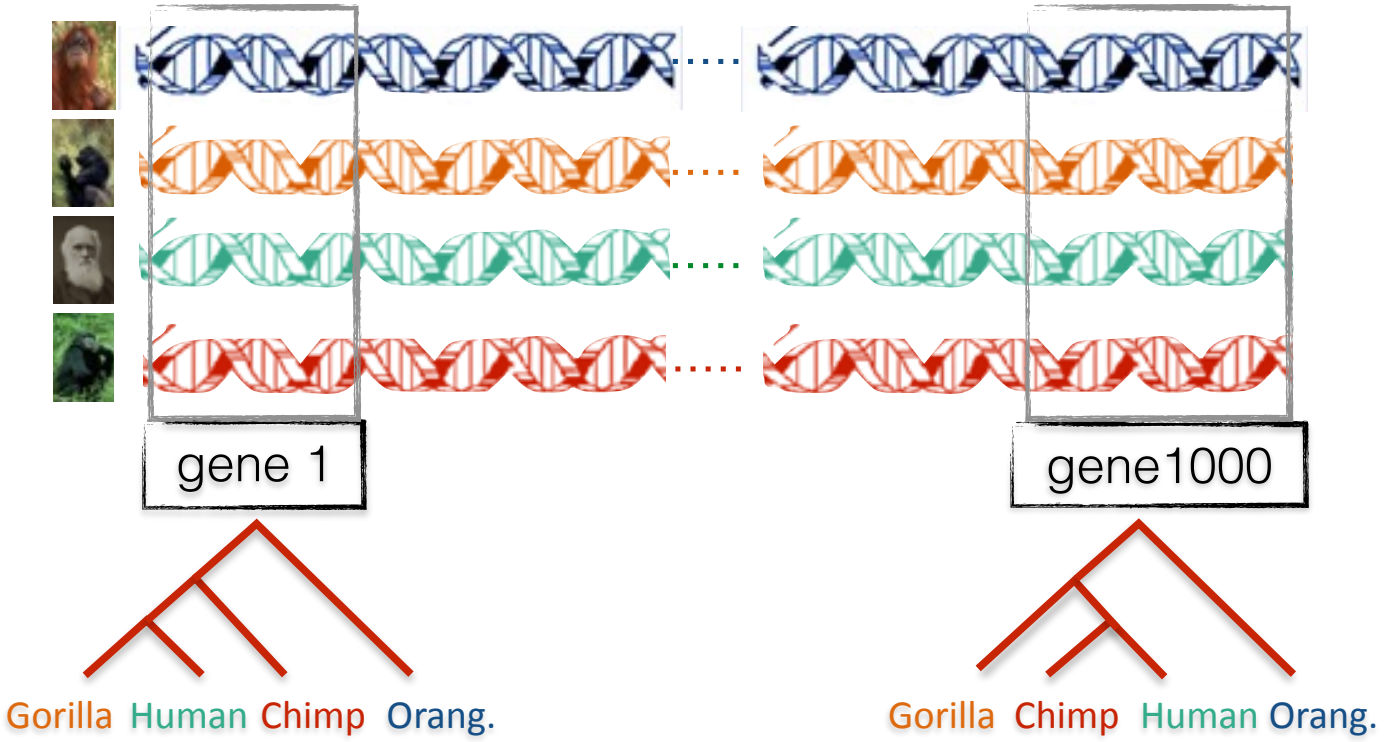
Is method M statistically consistent under model G?



Genome-scale data?



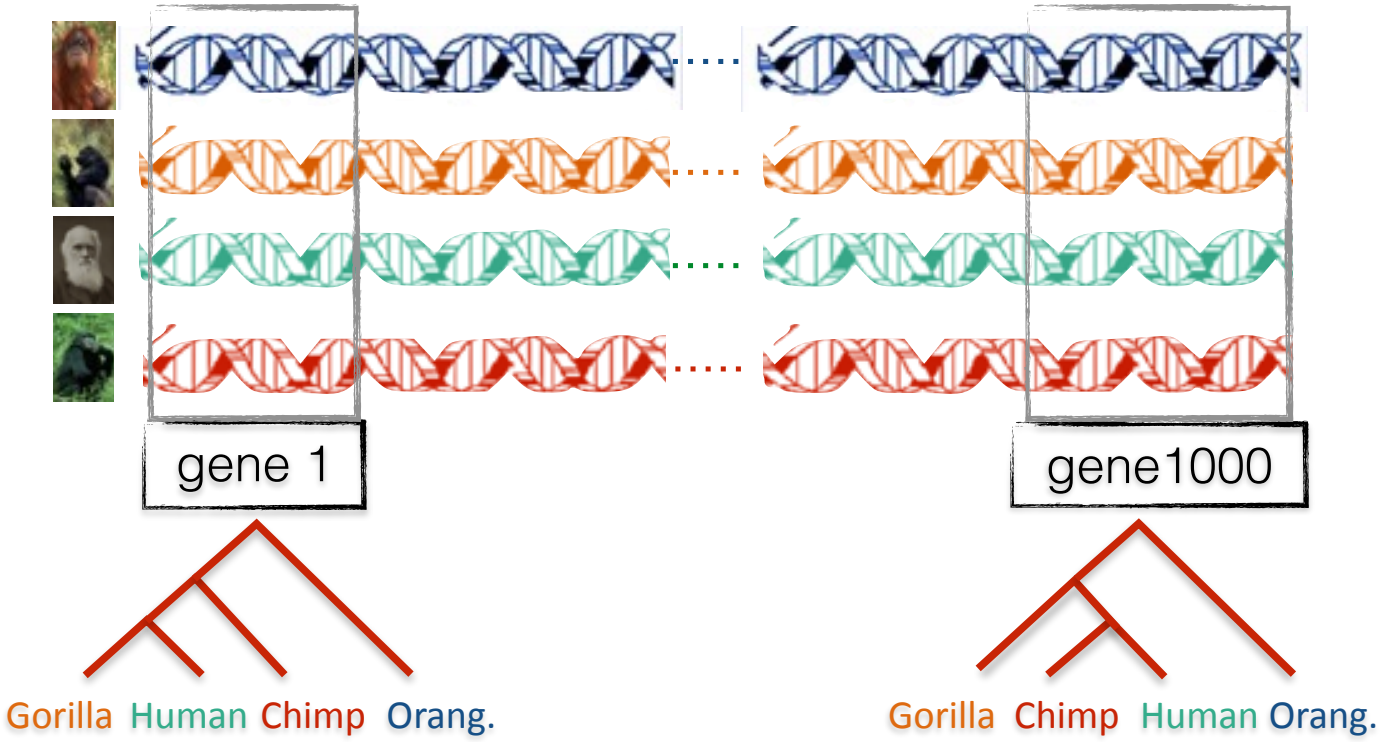
Gene tree discordance



Multiple causes for discord, including

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

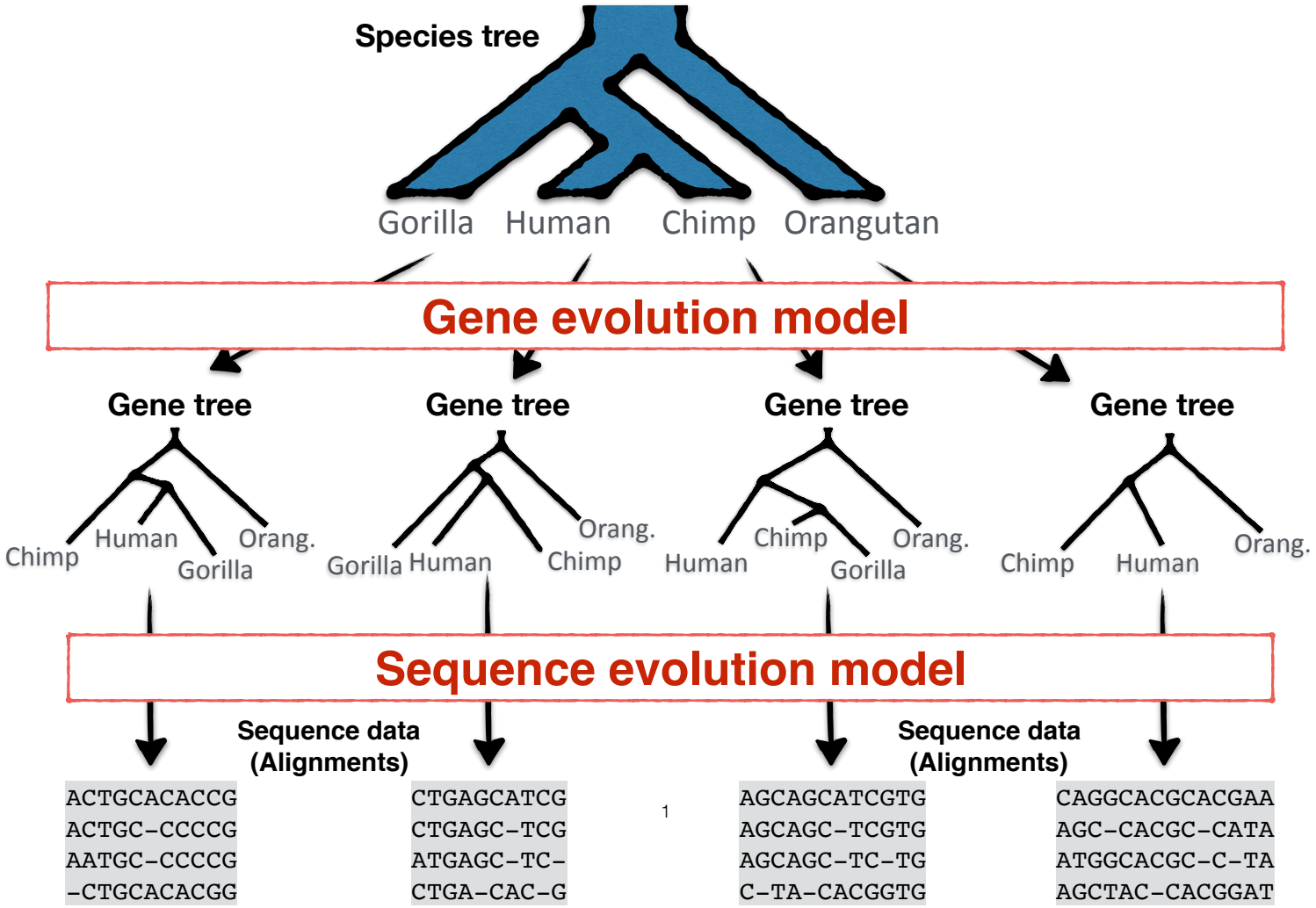
Gene tree discordance



Multiple causes for discord, including

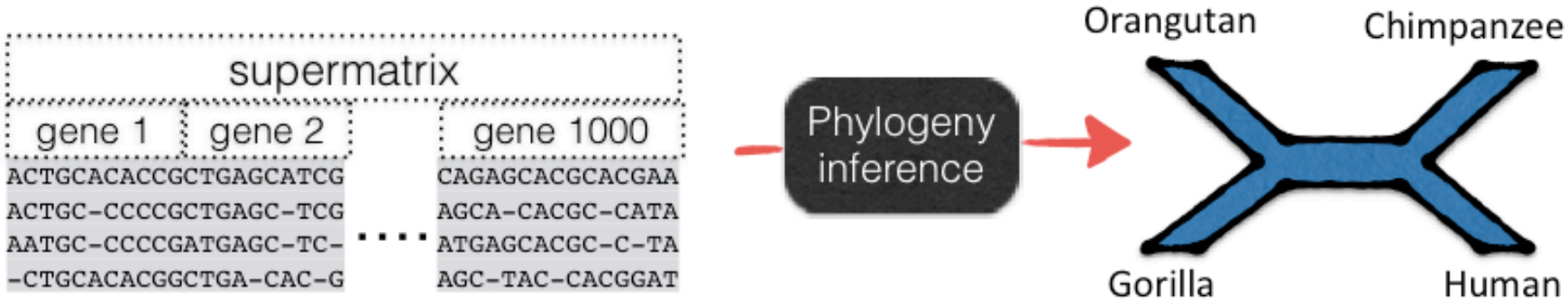
- Incomplete Lineage Sorting (ILS),
- Gene Duplication and Loss (GDL), and
- Horizontal Gene Transfer (HGT)

MSC+GTR Hierarchical Model



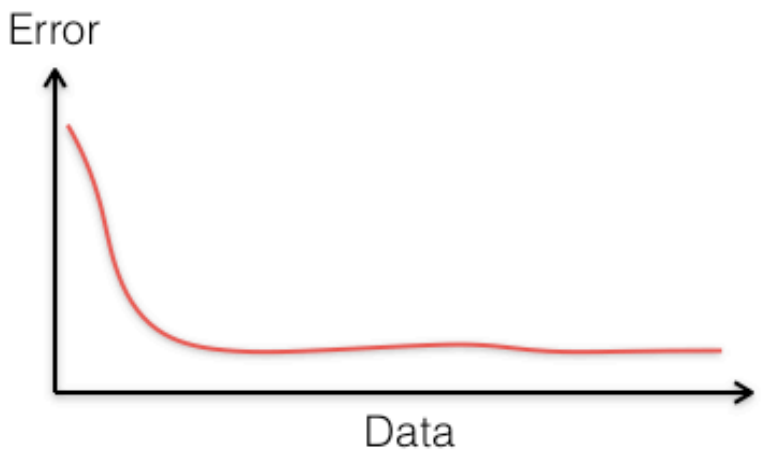
1. Gene trees evolve within the species tree (under the Multi-Species Coalescent model)
2. Sequences evolve down the gene trees (under GTR model)

Traditional approach: concatenation

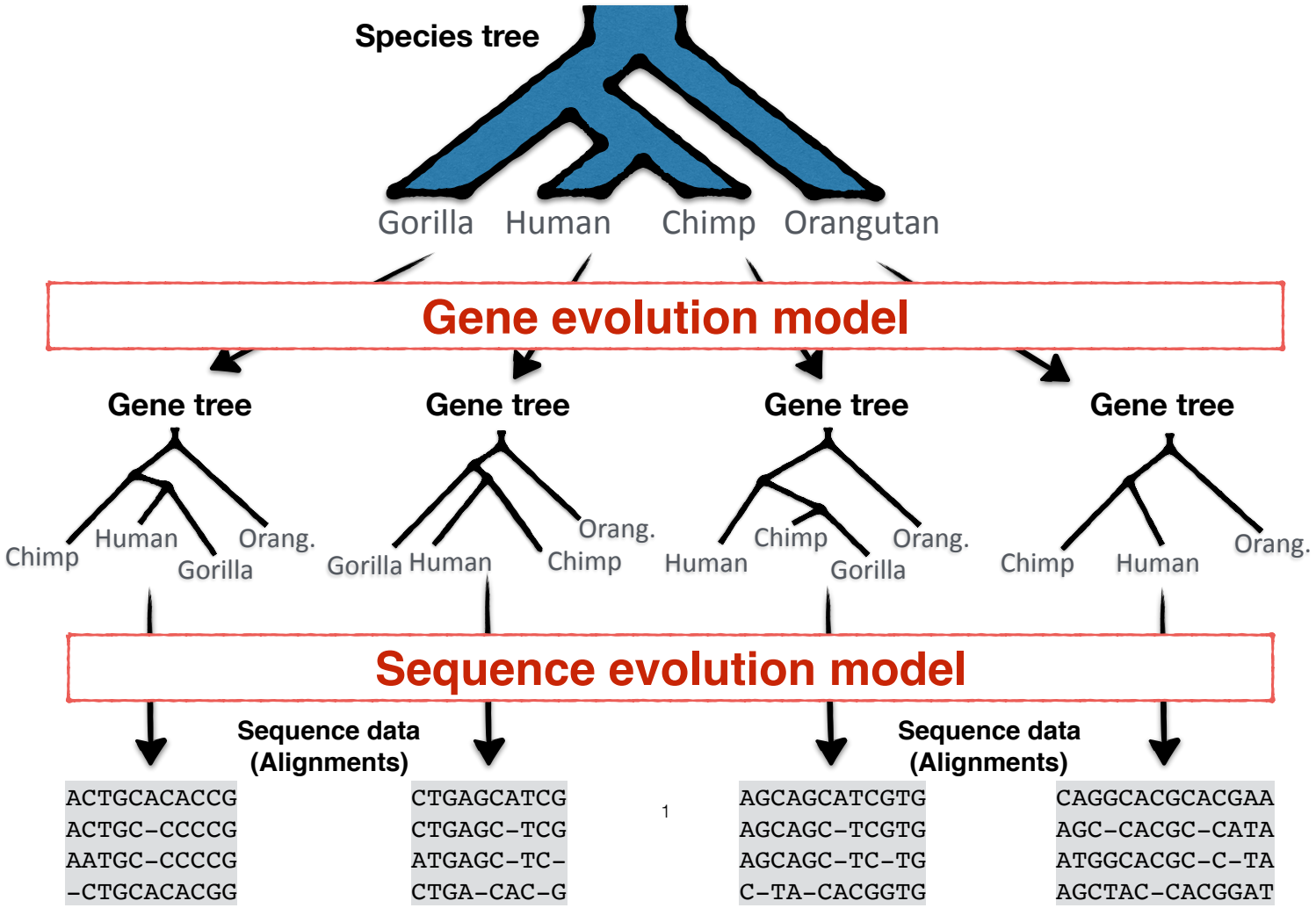


- Statistically <u>inconsistent</u> and can even be positively misleading (proved for unpartitioned maximum likelihood)
  [Roch and Steel, Theo. Pop. Gen., 2014]

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

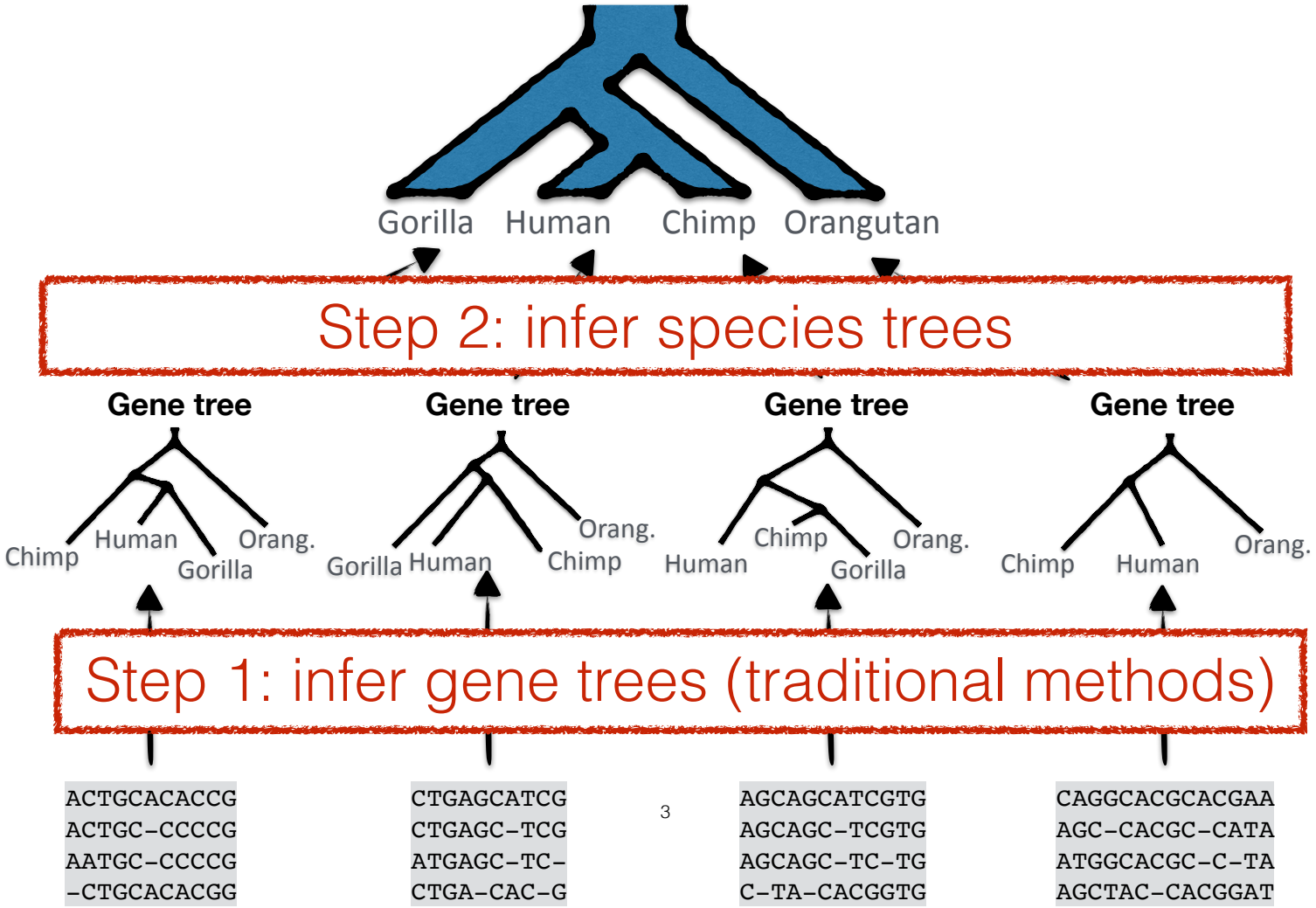


MSC+GTR Hierarchical Model



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

Summary Method Protocol



1. Given gene sequence alignments, compute gene trees
2. Given gene trees, combine into species tree

Faster than concatenation, and can be parallelized

Quartet trees and the MSC

- Allman, Degnan, and Rhodes (J. Mathematical Biology 2011) proved:
  - For every four species, the most probable unrooted gene tree is topologically identical to the true species tree
- This is not true for five species (anomaly zone)

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

[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

Theorem: Under MSC, the most probable quartet tree is the true species tree

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

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

[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

ASTRAL uses dynamic programming to solve a **constrained version** of this problem, and is provably **statistically consistent**

- **Theorem**: Statistically consistent under the multi-species coalescent model when solved exactly

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

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

- **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

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

eLIFE

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]

ELSEVIER

Molecular Phylogenetics and Evolution

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

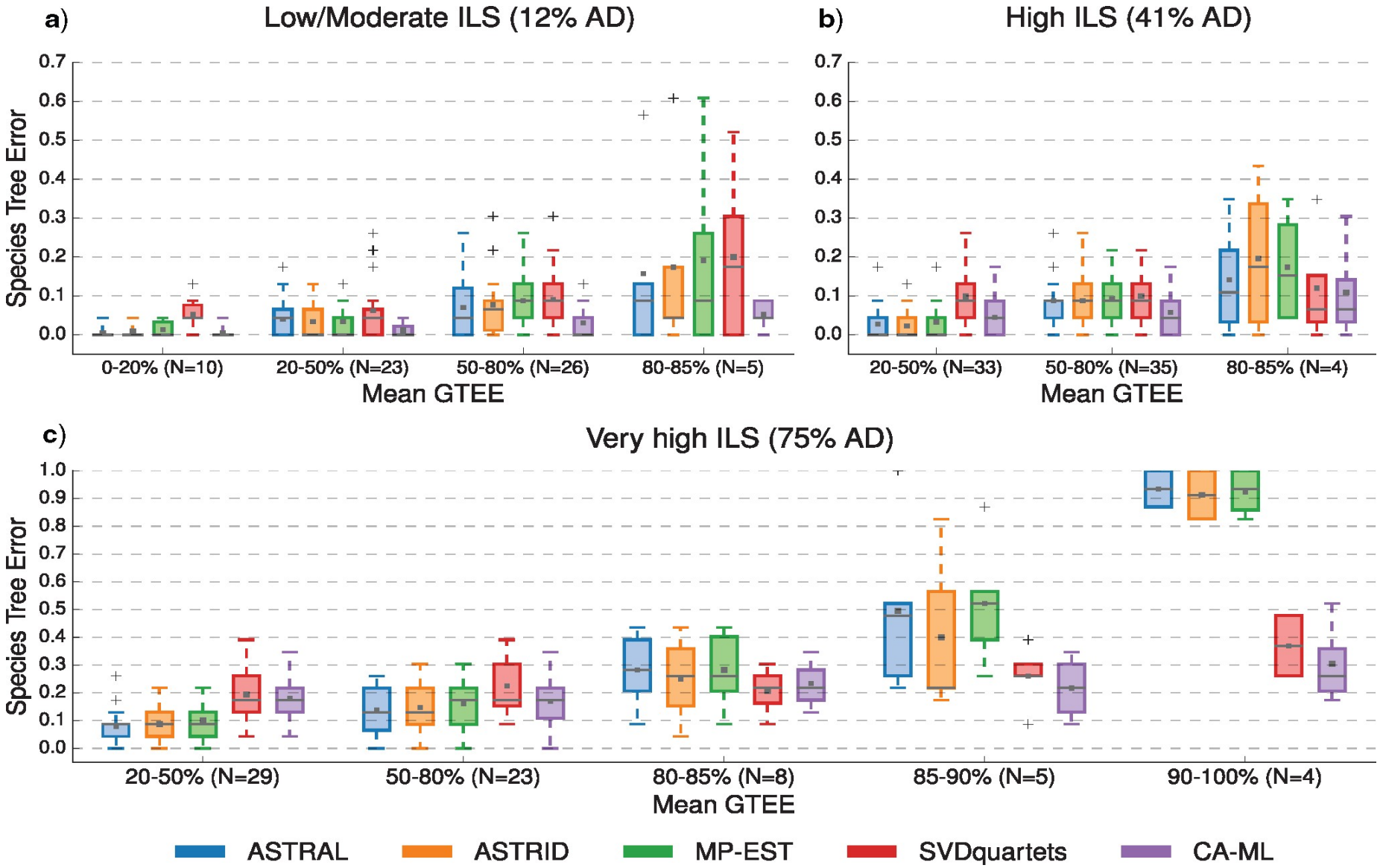
ORIGINAL ARTICLE

Land connectivity changes and global cooling shaped the colonization history and diversification of New World quail (Aves: Galliformes: Odontophoridae)

LETTER

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

X



Concatenation using Maximum Likelihood (RAxML) vs. ASTRAL: depends on levels of ILS and GTEE

Figure 1. The impact of gene tree estimation error (GTEE) and incomplete lineage sorting (ILS) on species tree error, all datasets with 26 species and 1000 genes.

Syst Biol, Volume 67, Issue 2, March 2018, Pages 285–303, https://doi.org/10.1093/sysbio/syx077

Summary and two key ideas for ILS-based species tree estimation

- Theorem: For every four species, the most probable (unrooted) gene tree is topologically identical to the true species tree
- ASTRAL can find the "Maximum Quartet Support Species Tree" (MQSST) in low degree polynomial time, in a constrained search space
  - Maintains statistical consistency if the constraint space is defined from the gene trees
  - Fast in practice
  - Highly accurate if there are enough gene trees (even when there is gene tree estimation error and missing data)
- Concatenation using ML (e.g., RAxML) sometimes more accurate, despite not having guarantee of statistical consistency

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

Gene Family Trees

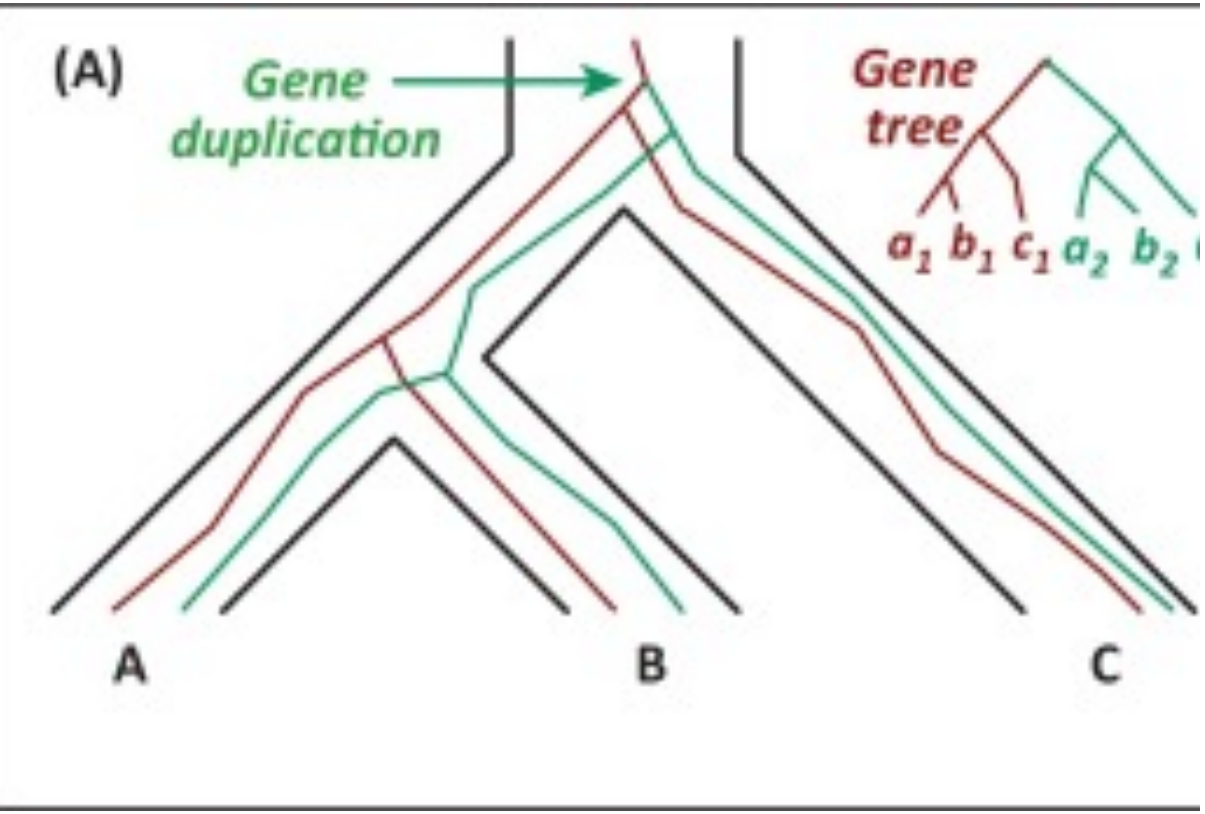


Figure by Luay Nakhleh, TREE 2013

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

Species tree estimation under GDL

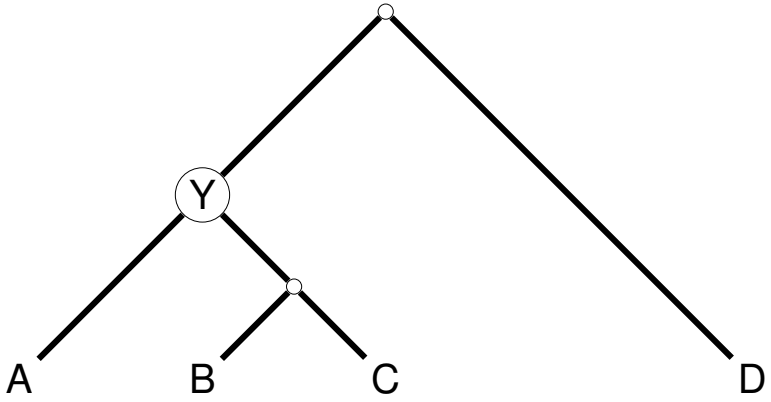
Options:

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

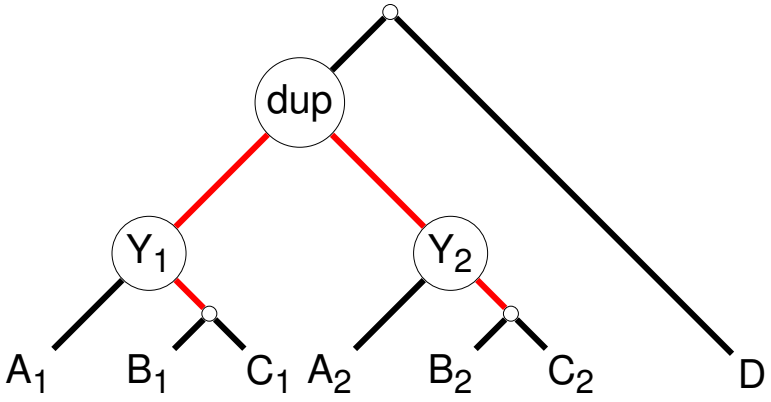
Note:

Nothing proven to be statistically consistent under GDL… until 2019

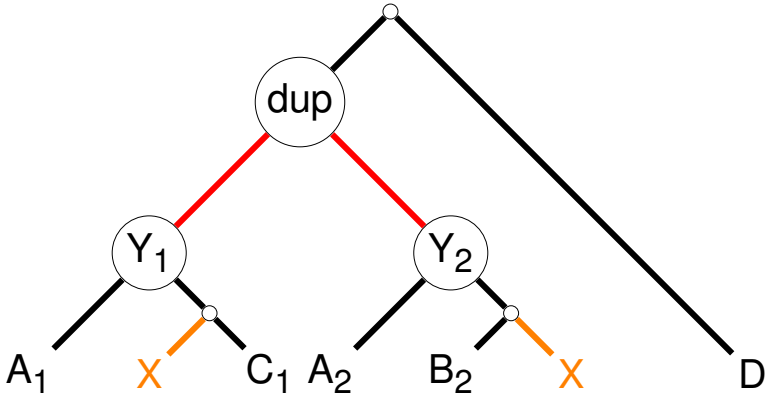
Problem: Given set of MUL-trees, infer the species tree



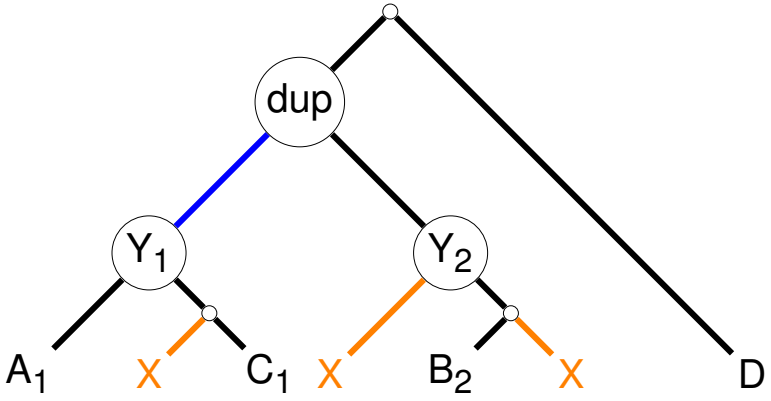
(a) Species tree T^*



(b) Gene tree M_1 with one duplication.



(c) Gene tree M_2 with one duplication and two losses.



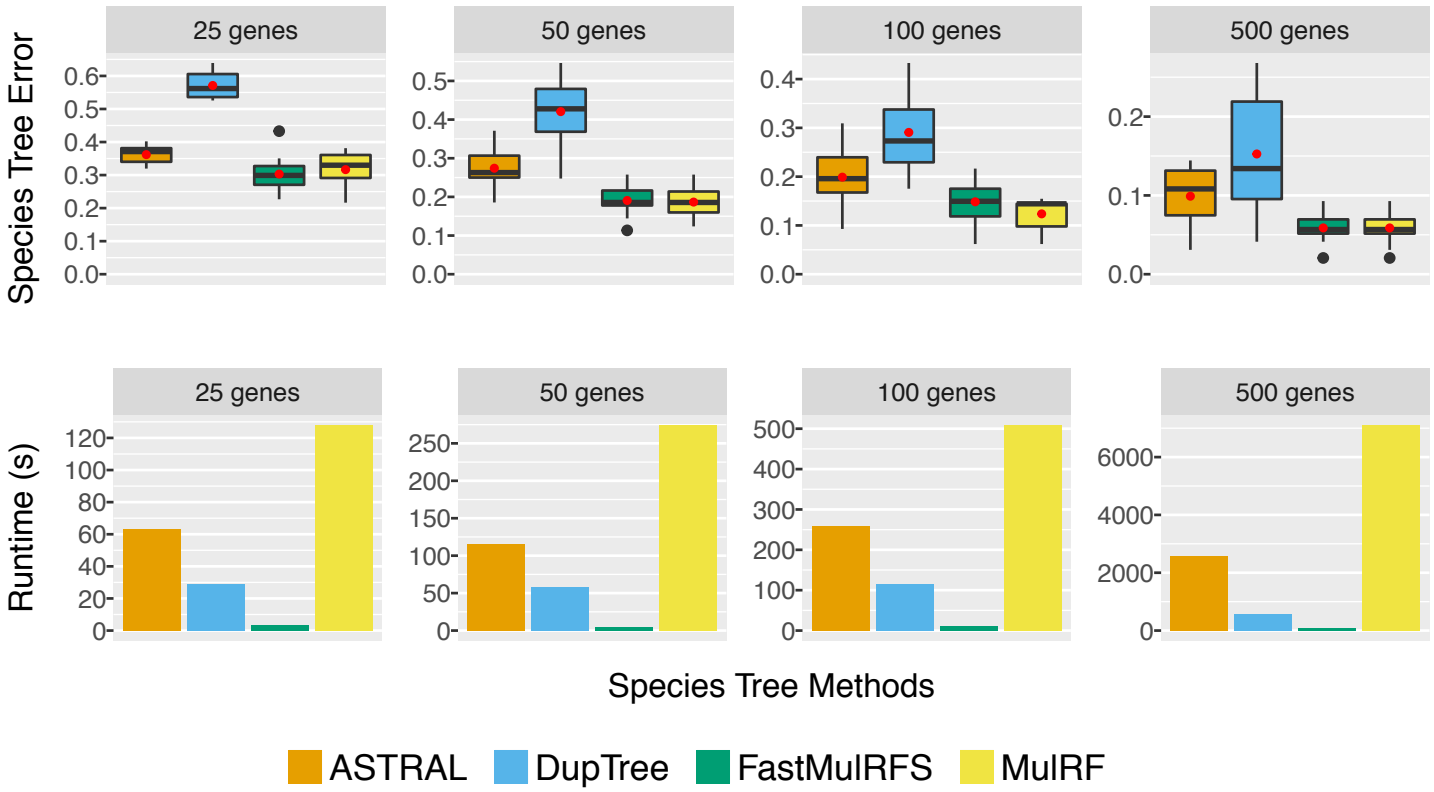
(d) Gene tree with one duplication and three losses.

Note: no orthology detection

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

But…ASTRAL-multi is not as accurate as other methods!



Results on 100-species datasets with moderate GDL, moderately high ILS, and high GTEE

MOLECULAR BIOLOGY AND EVOLUTION

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

ASTRAL-pro

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

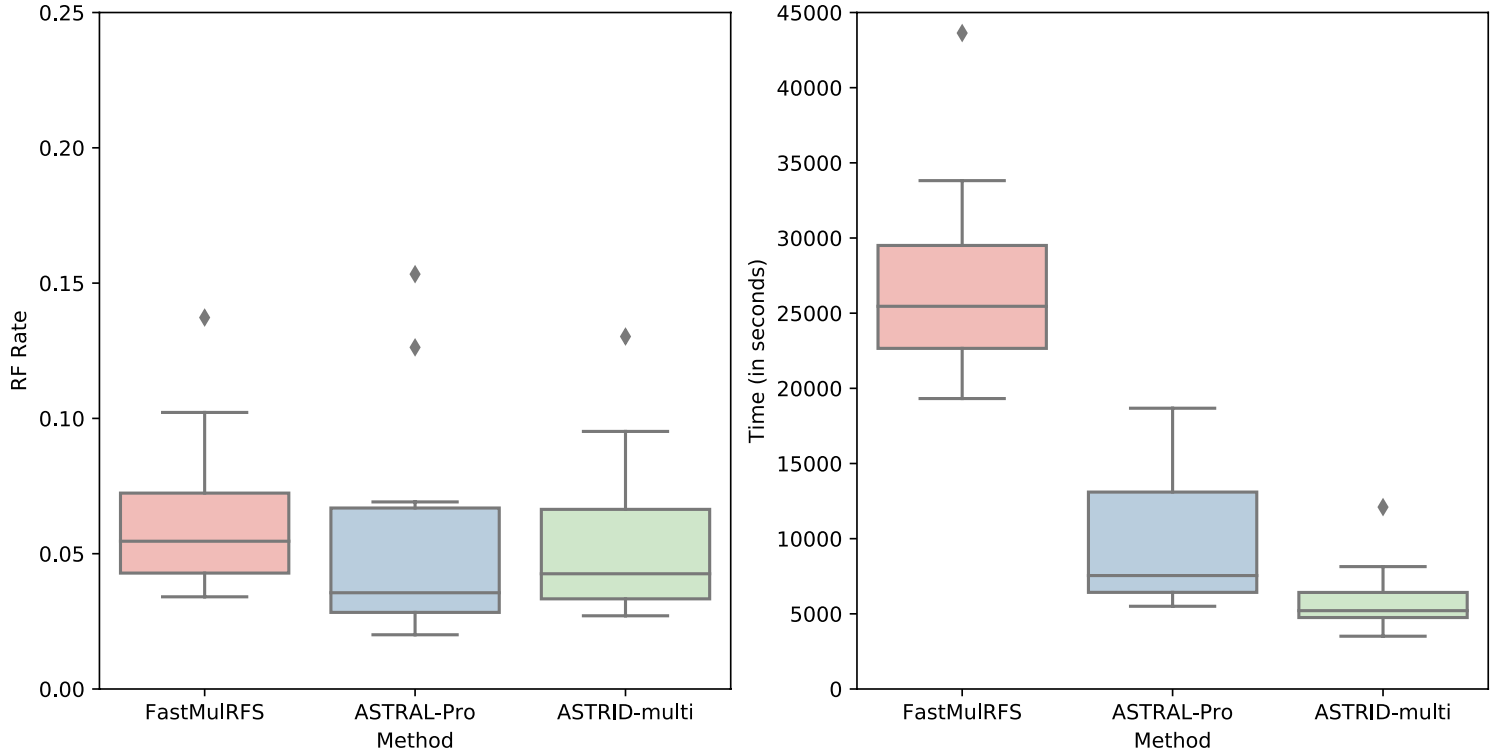
- Step 1: ”root and tag” every mul-tree
- Step 2: Use the rooting to define “speciation quartets”
- Step 3: Run ASTRAL’s DP algorithm with modified weights, reflecting speciation quartets

Theorem: ASTRAL-pro is statistically consistent if it correctly roots-and-tags every mul-tree

Results on 1000-taxon Species Trees

Accuracy: no important accuracy differences on these large datasets.

Speed: ASTRID-multi fastest, but ASTRAl-Pro nearly same. FastMulRFS slowest (max about 7 hours).



1000 species and 1000 gene trees estimated from 100bp alignments (approx. 44% mean gene tree error), AD=20%, duplication rate of 5.0 × 10−10, and loss/dup = 1.
James Willson et al. AlCoB 2021. Lecture Notes in Computer Science, vol 12715.

Summary and two key ideas for species tree estimation under ILS

- Theorem: For every four species, the most probable (unrooted) gene tree is topologically identical to the true species tree
- ASTRAL can find the "Maximum Quartet Support Species Tree" (MQSST) in low degree polynomial time, in a constrained search space
  - Maintains statistical consistency if the constraint space is defined from the gene trees
  - Fast in practice
  - Highly accurate if there are enough gene trees (even when there is gene tree estimation error and missing data)
- Concatenation using ML (not consistent) sometimes more accurate, despite not having guarantee of statistical consistency

Summary and three key ideas for species tree estimation under ILS+GDL

- Theorem: For every four species, the most probable (unrooted) gene tree is topologically identical to the true species tree
- ASTRAL-multi and ASTRAL-one: can find the "Maximum Quartet Support Species Tree" (MQSST) in low degree polynomial time, in a constrained search space – and are statistically consistent
- ASTRAL-Pro (not guaranteed statistically consistent) but:
  - Roots and tags gene trees
  - Statistically consistent if rooting and tagging is correct
  - Highly accurate if there are enough gene trees (even when there is gene tree estimation error and missing data)

What about HGT?

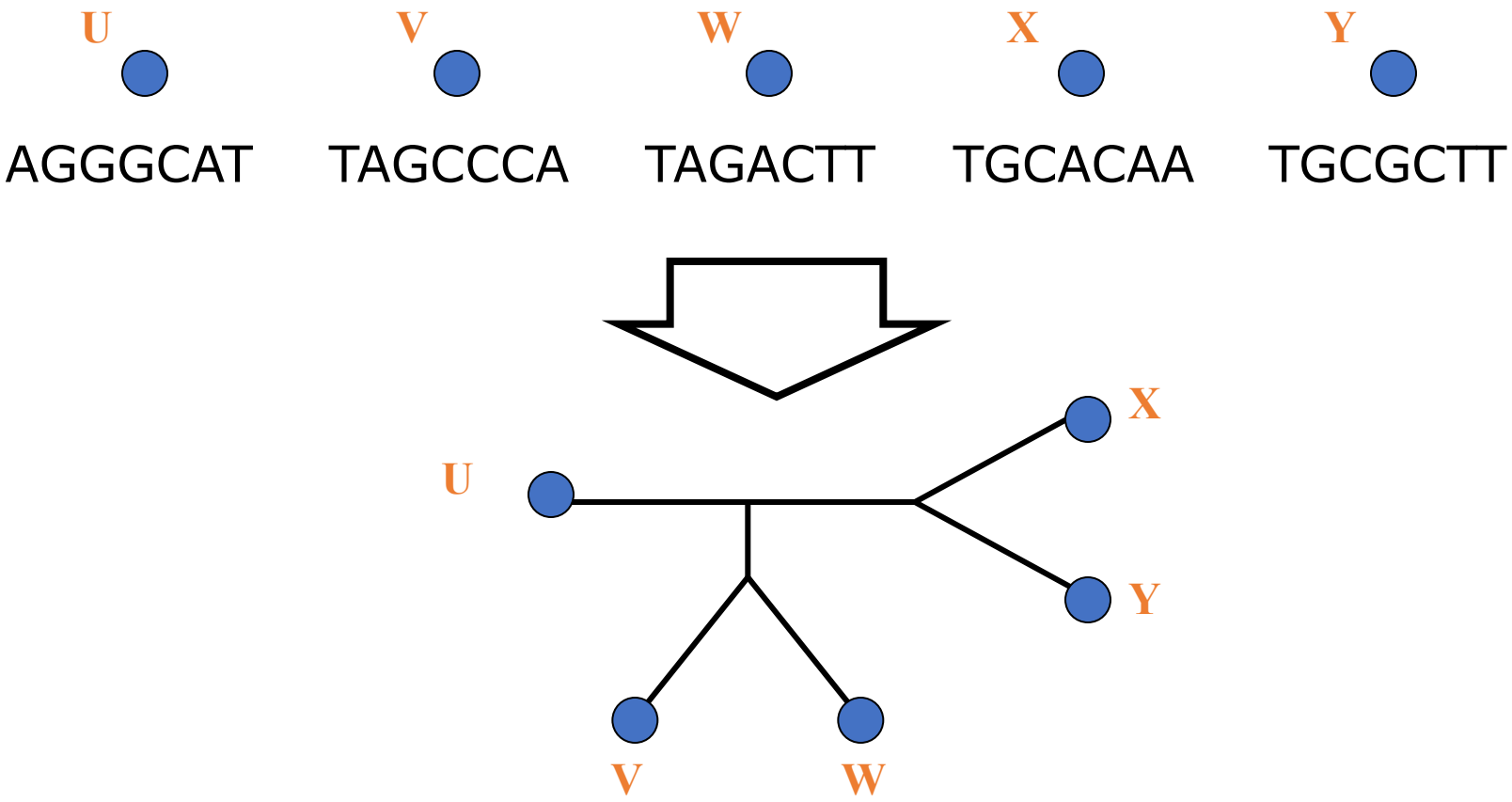
- HGT also makes heterogeneous gene trees
- Under some assumptions of random HGT operating, it may be possible to define the “underlying” species tree.
- Statistical consistency of quartet-based methods for computing the underlying species tree established by:
    - Roch & Snir, JCB 2013
    - Daskalakis & Roch, arXiv 2015
- Simulation study (Davidson et al. 2015) shows ASTRAL and wQMC (Snir et al.) more accurate than concatenation and NJst when ILS+HGT is present

Gene trees -> Species trees, using Quartet Trees

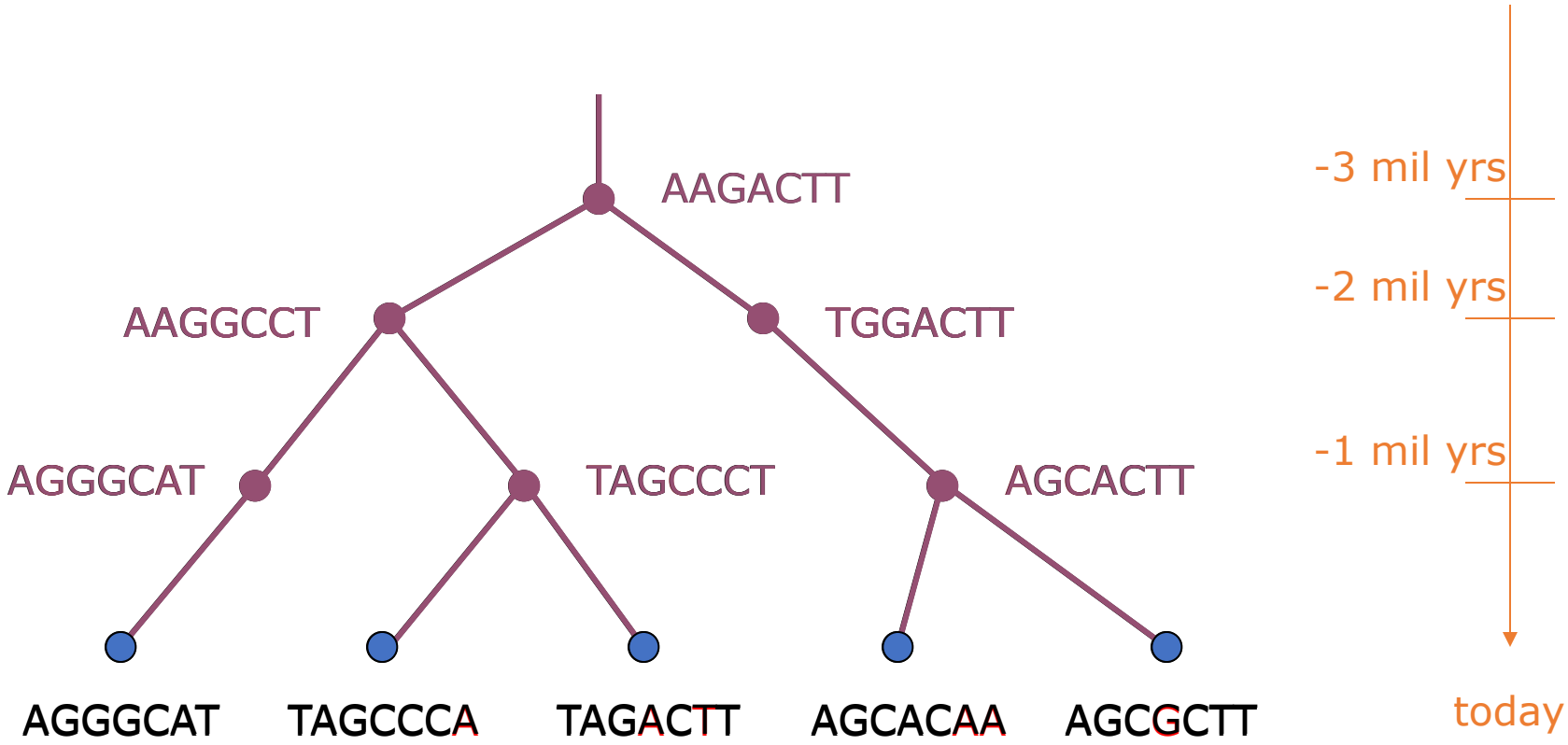
- Recurring theme: the most probable quartet tree is the species tree for:
  - ILS (Multi-species coalescent)
  - ILS+GDL (DLCOAL)
  - Random (but bounded) HGT
- Hence, quartet-based species tree estimation (from gene trees) is often statistically consistent, if performed properly.
- Finding MQSST (within a constrained search space) a powerful approach that maintains consistency and is computationally efficient
- Interestingly, MQSST is also fairly robust

Part II: Large-scale gene tree estimation

Phylogeny Problem



DNA Sequence Evolution (Idealized)



Markov Models of Sequence Evolution (Gene Tree)

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 as a statistical problem

- Assume DNA sequences are generated on an unknown model tree, and try to 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

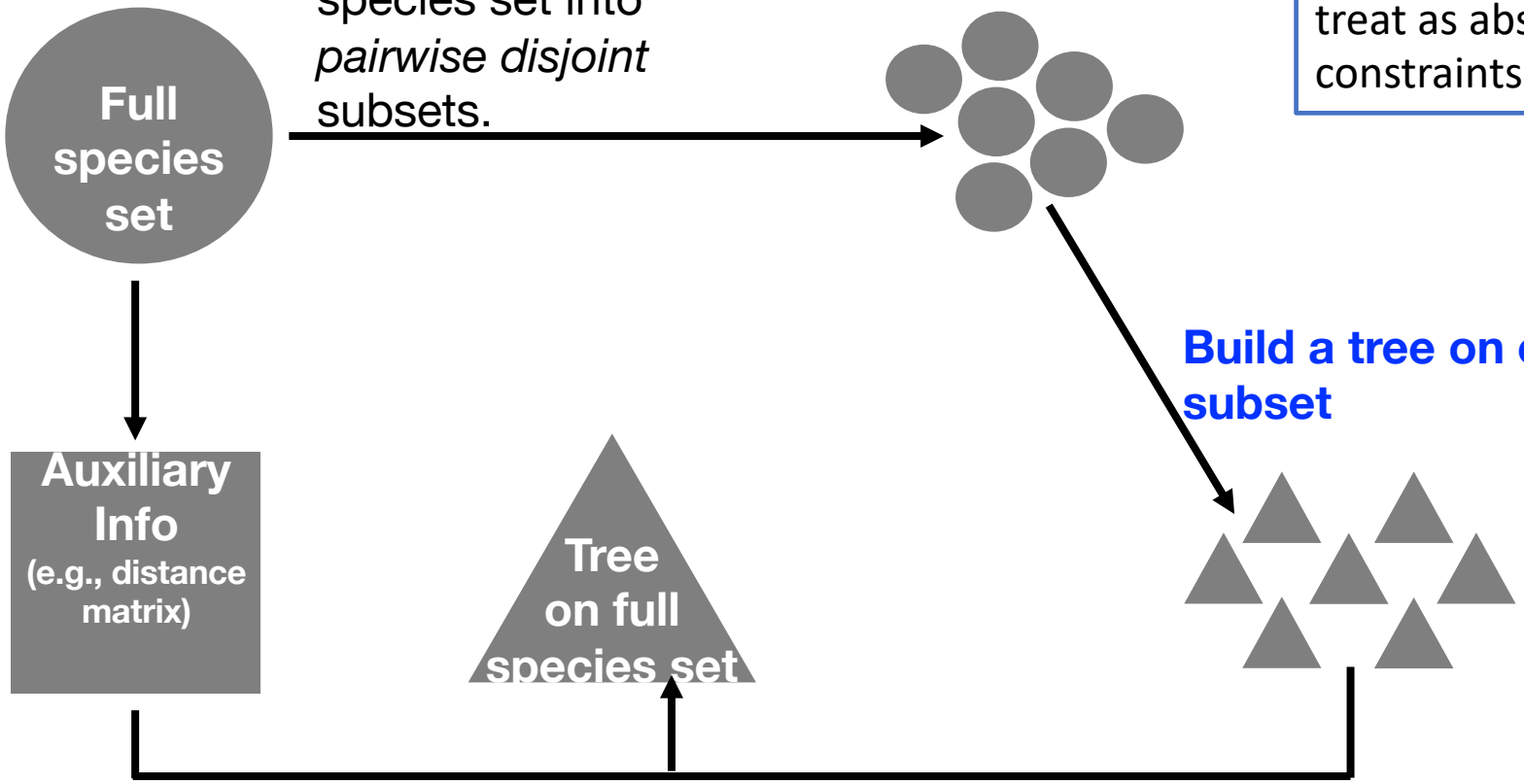
Maximum likelihood for gene tree estimation

- 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)

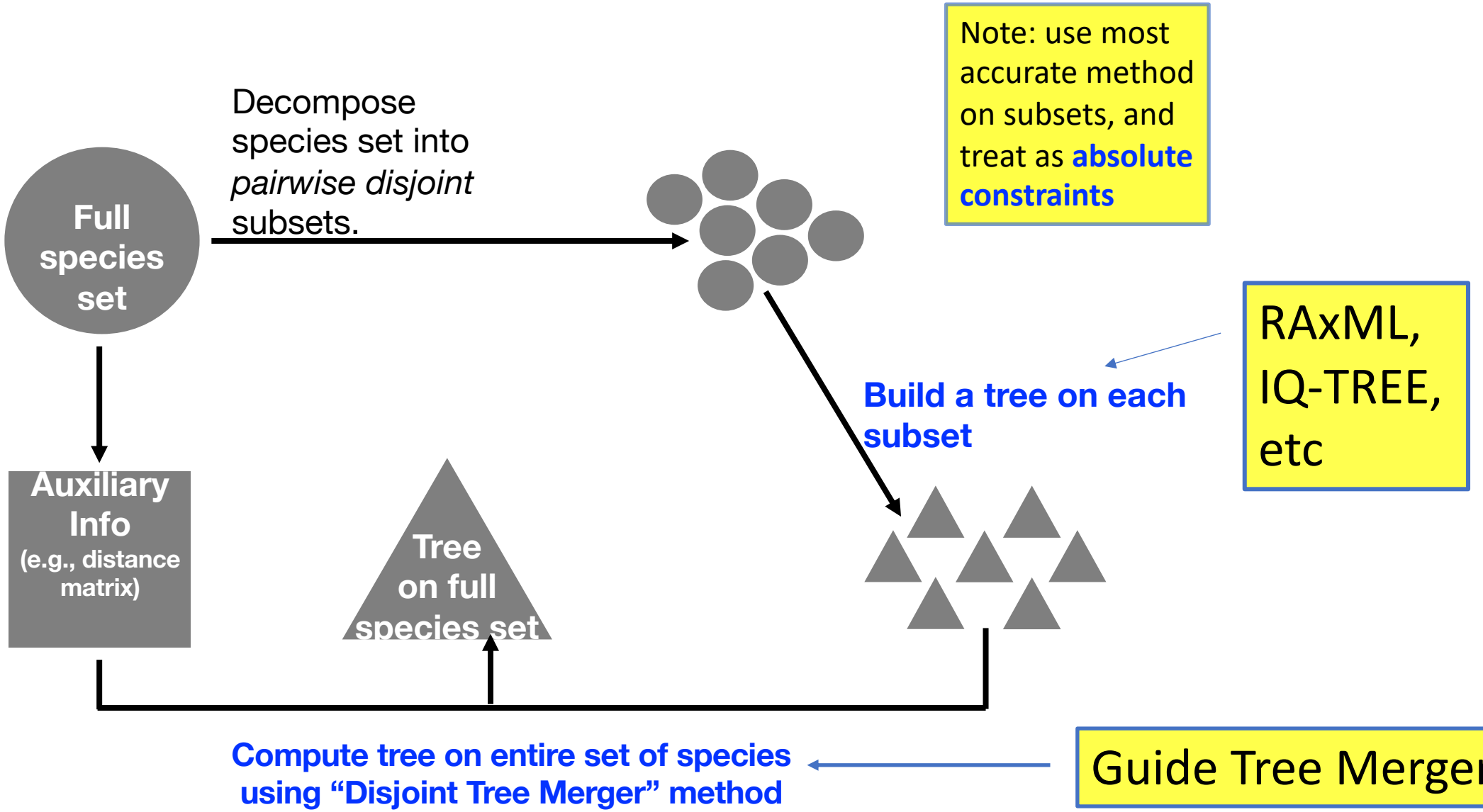
Divide-and-Conquer using Disjoint Tree Mergers

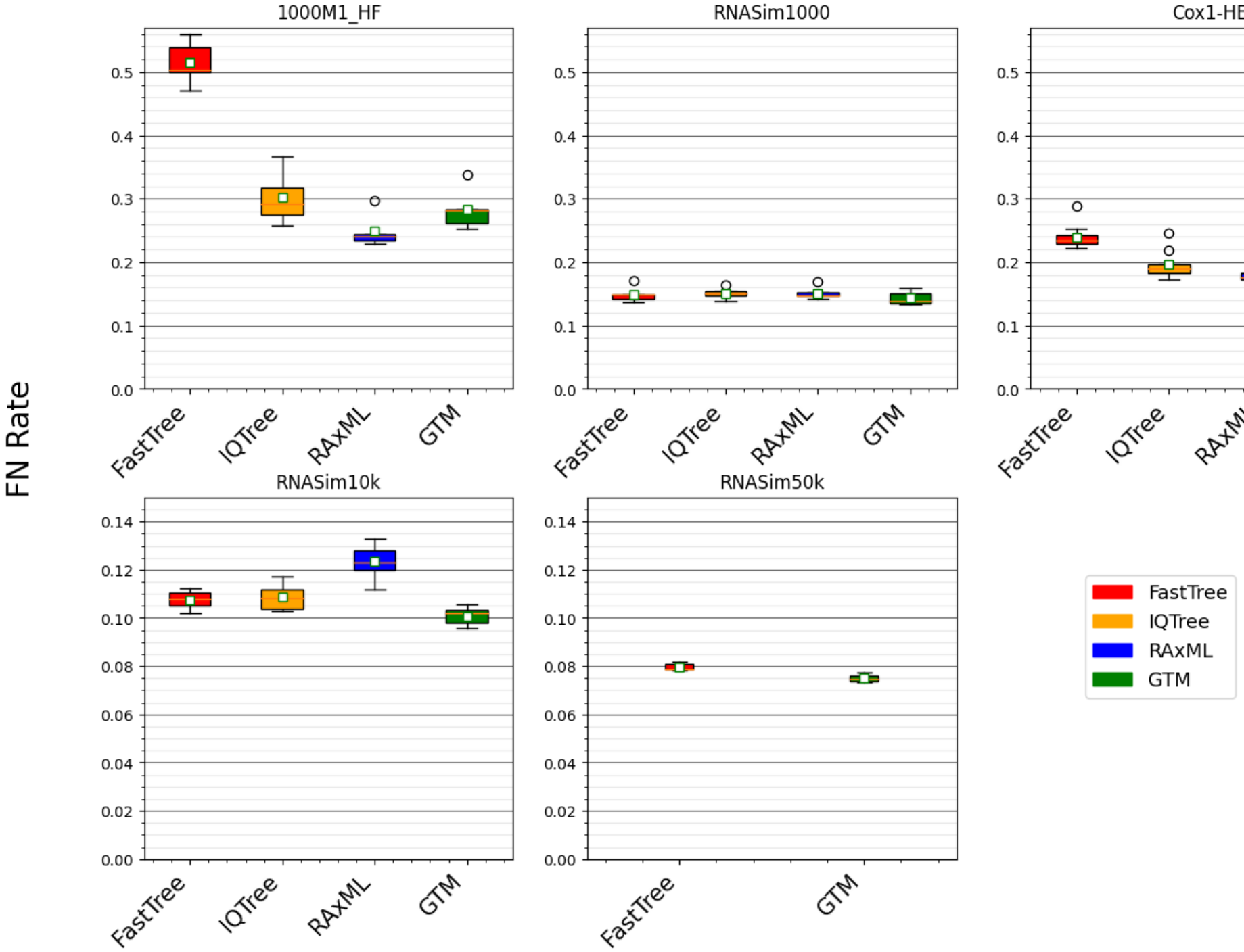
Erin Molloy, Introduced this approach

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



Divide-and-Conquer Gene Tree Estimation





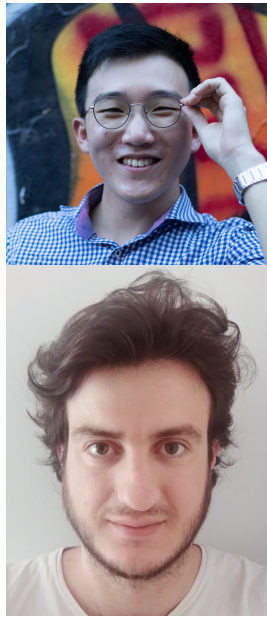
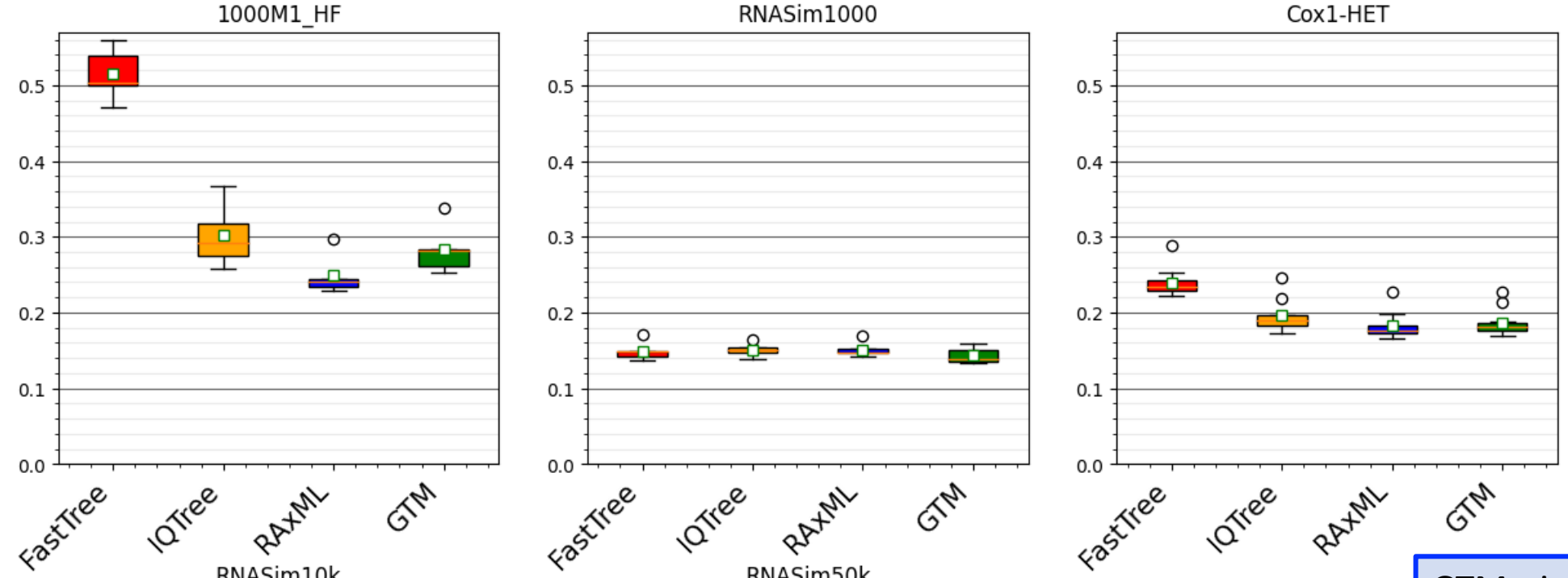


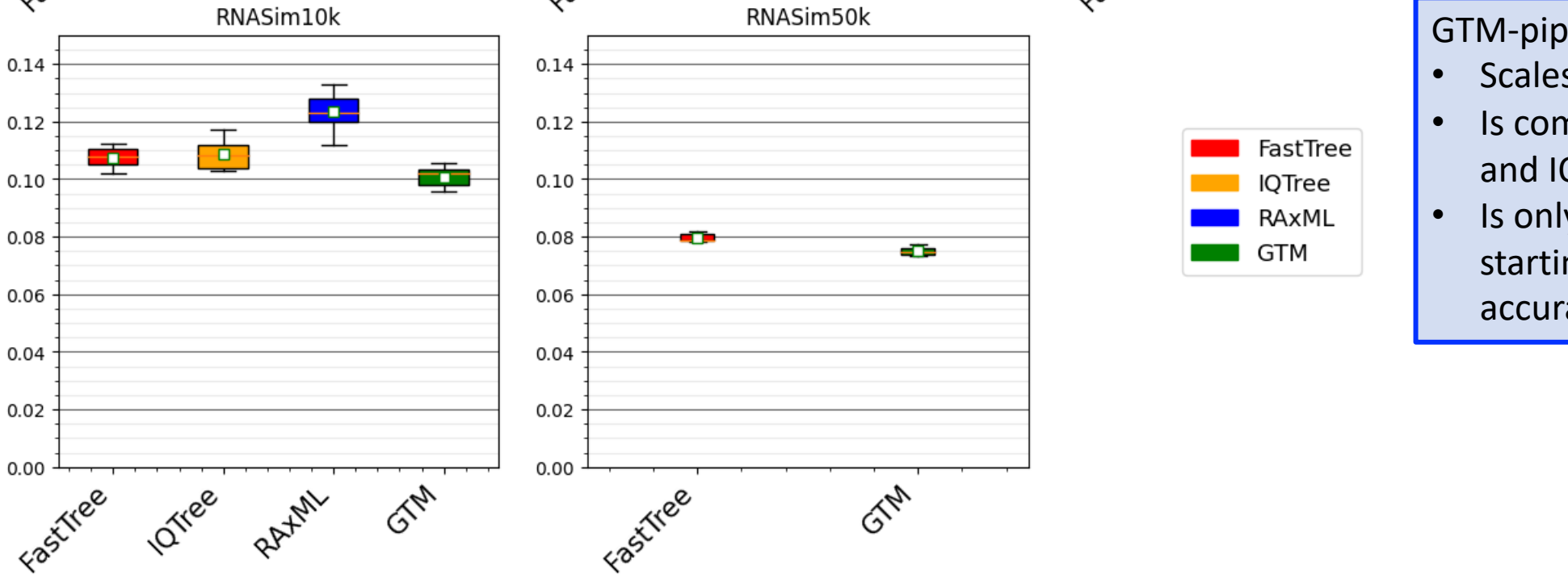
Figure 2 from "Disjoint Tree Mergers for Large-Scale Maximum Likelihood Tree Estimation", Park et al., *Algorithms 2021*

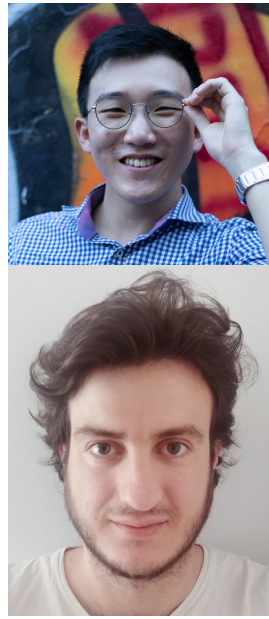
GTM pipeline:

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

FN Rate

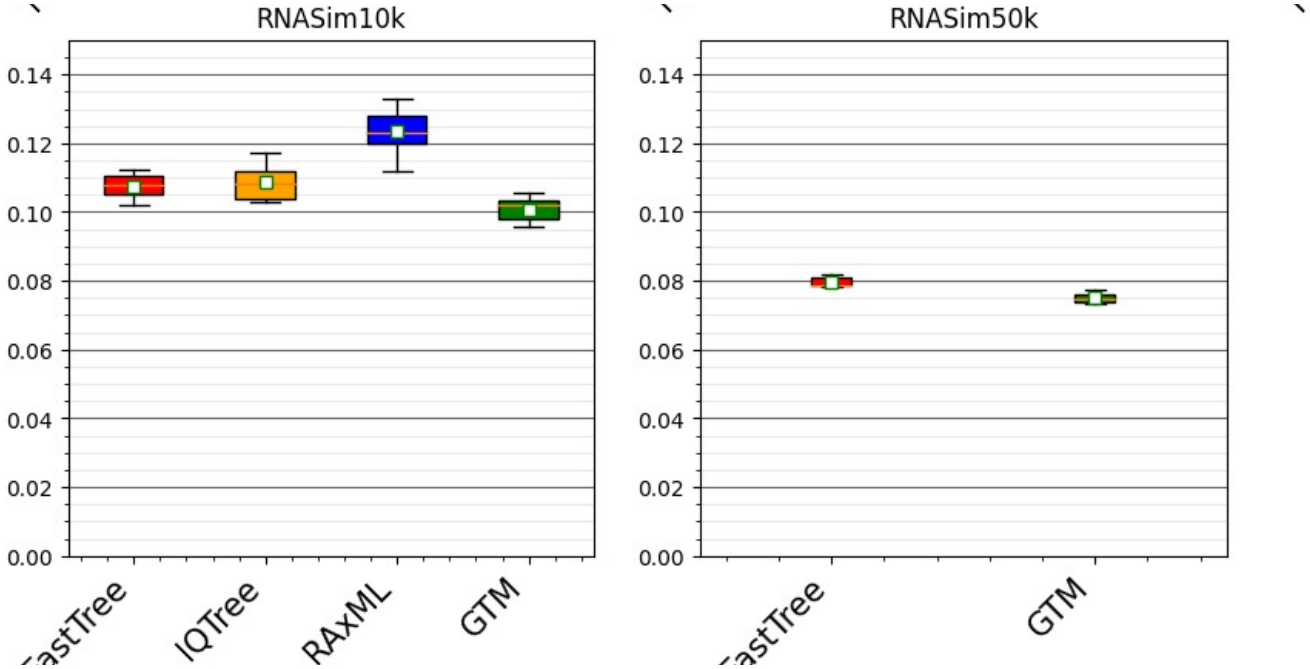


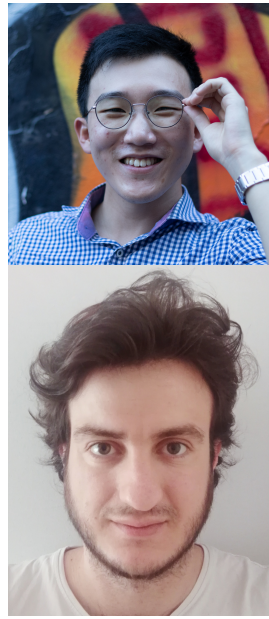




GTM-pipeline:

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



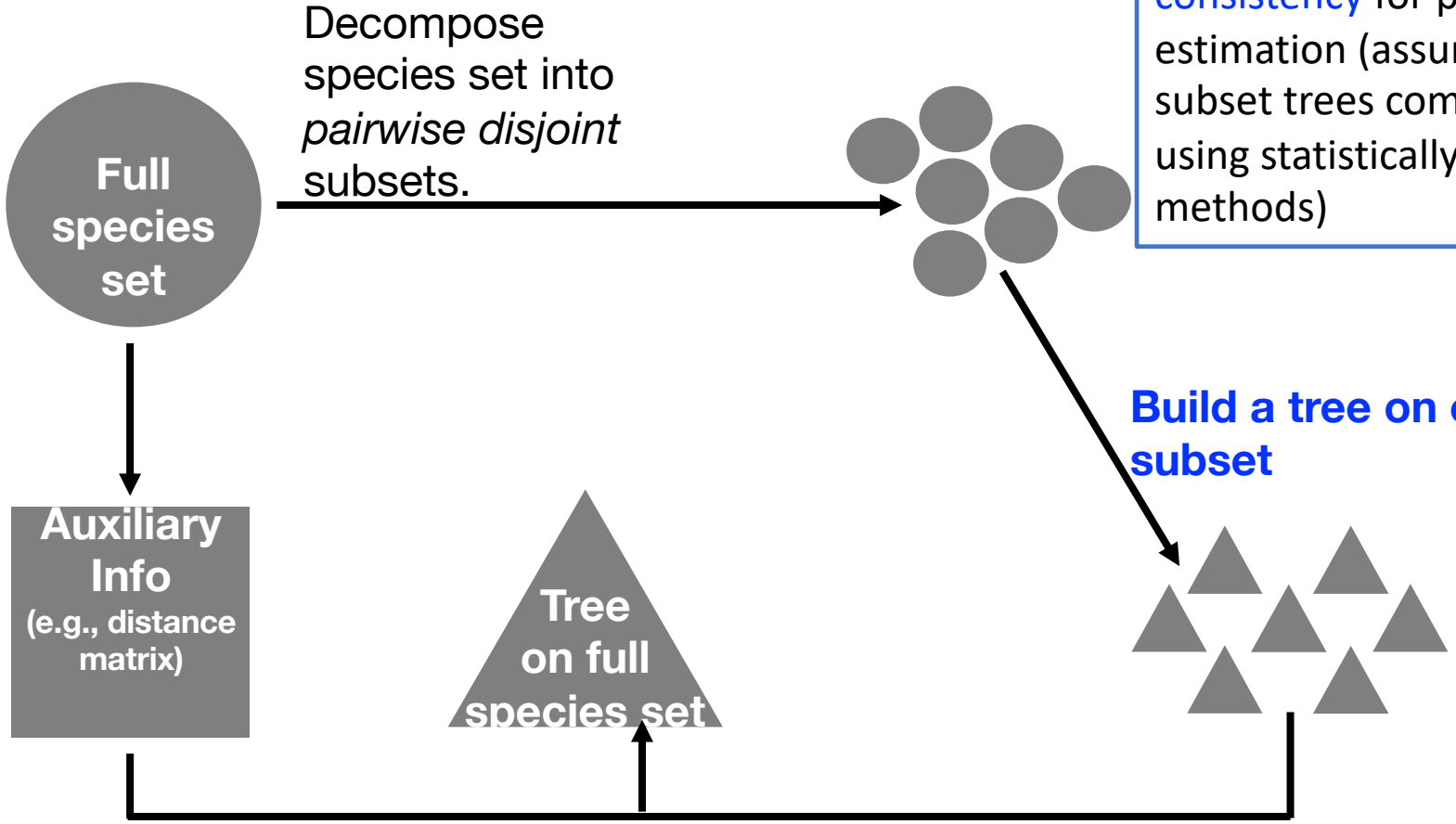


GTM-pipeline:

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

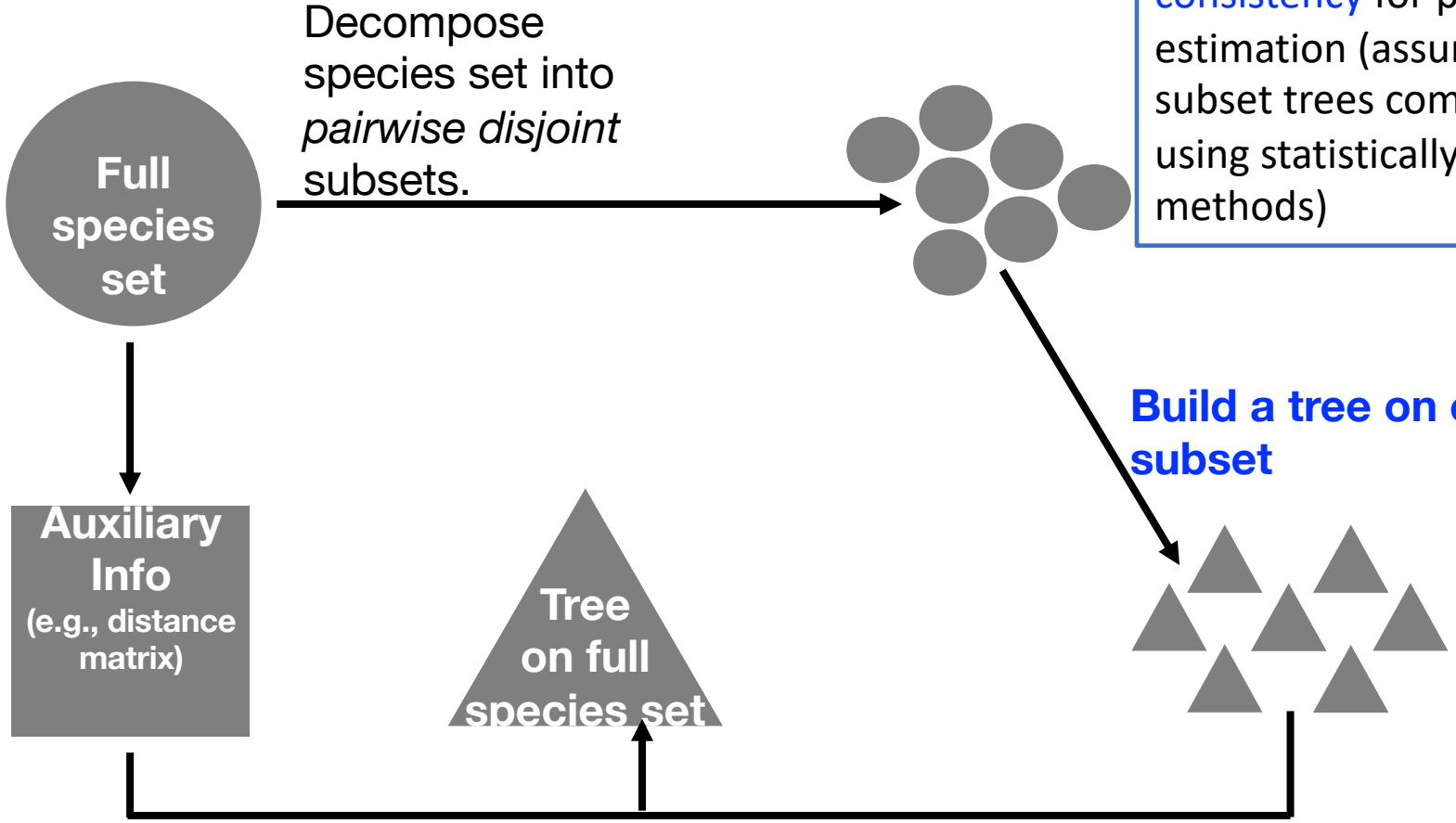
Divide-and-Conquer using Disjoint Tree Mergers

Theorem: DTM pipelines maintain statistical consistency for phylogeny estimation (assuming subset trees computed using statistically consistent methods)



Divide-and-Conquer using Disjoint Tree Mergers

Theorem: DTM pipelines maintain statistical consistency for phylogeny estimation (assuming subset trees computed using statistically consistent methods)



Also: has been used for species Tree estimation (e.g., ASTRAL and concatenation) and shows Improvements in runtime and sometimes accuracy

Part III: Multiple Sequence Alignment

Multiple Sequence Alignment (MSA): *a scientific grand challenge*[1]

```
S1 = AGGCTATCACCTGACCTCCA        S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC            S2 = TAG-CTATCAC--GACCGC--
S3 = TAGCTGACCGC                 S3 = TAG-CT-------GACCGC--
  ...                              ...
Sn = TCACGACCGACA          →     Sn = -------TCAC--GACCGACA
```

Novel techniques needed for scalability and accuracy

NP-hard problems and large datasets
Current methods do not provide good accuracy
Few methods can analyze even moderately large datasets

Many important applications besides phylogenetic estimation

[1] Frontiers in Massive Data Analysis, National Academies Press, 2013

The true pairwise alignment

- **Reflects historical substitution, insertion, and deletion events**
- **Letters (nucleotides or amino acids) in the same column are supposed to be homologs**

What makes MSA difficult?

- Large numbers of sequences
- High rates of substitutions and indels
- Sequence length heterogeneity
- Very long sequences (e.g., genome-scale)

Two-phase estimation

Alignment methods

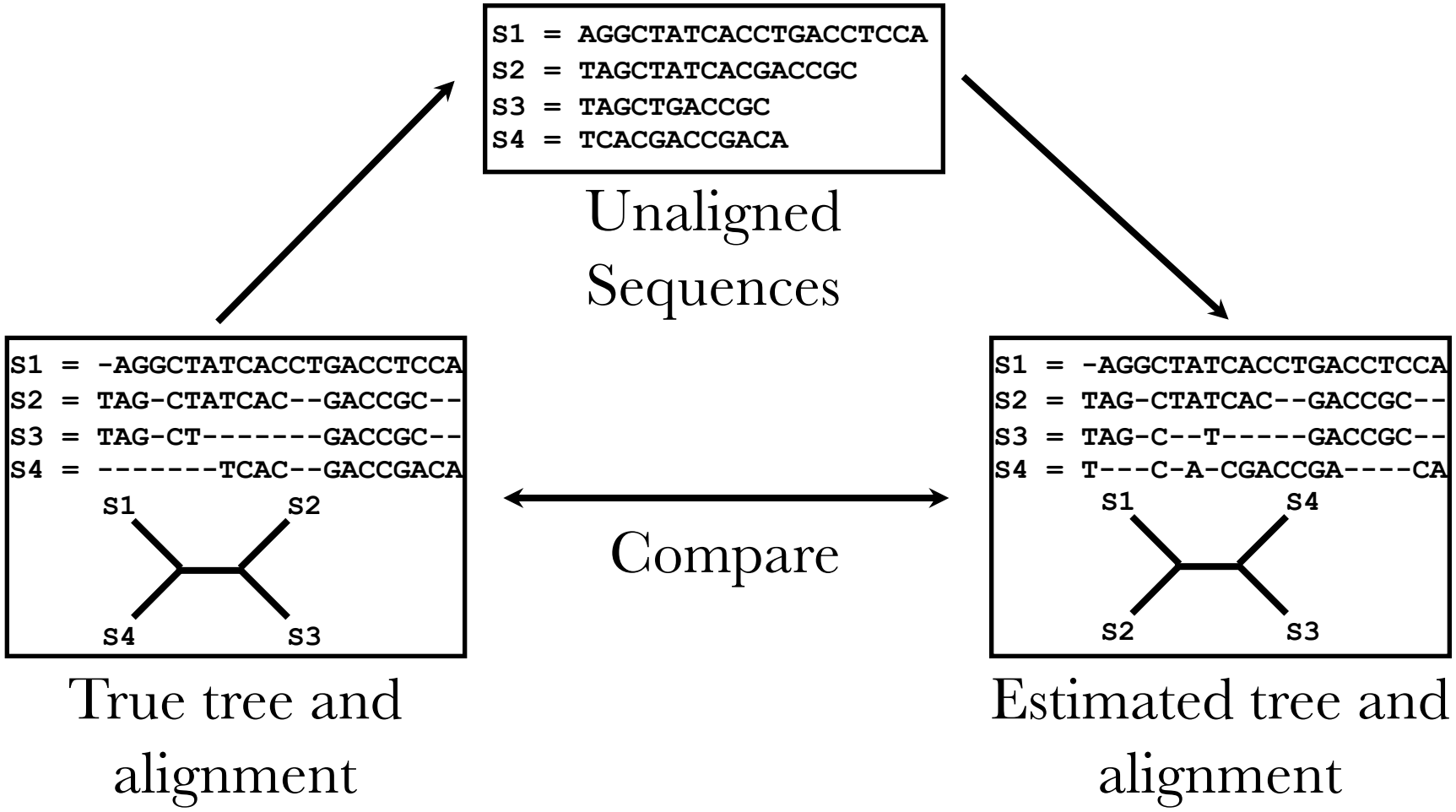
- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- Probalign
- MAFFT
- Muscle
- Di-align
- T-Coffee
- Prank (PNAS 2005, Science 2008)
- Opal (ISMB and Bioinf. 2007)
- *FSA (PLoS Comp. Bio. 2009)*
- *Infernal (Bioinf. 2009)*
- Etc.

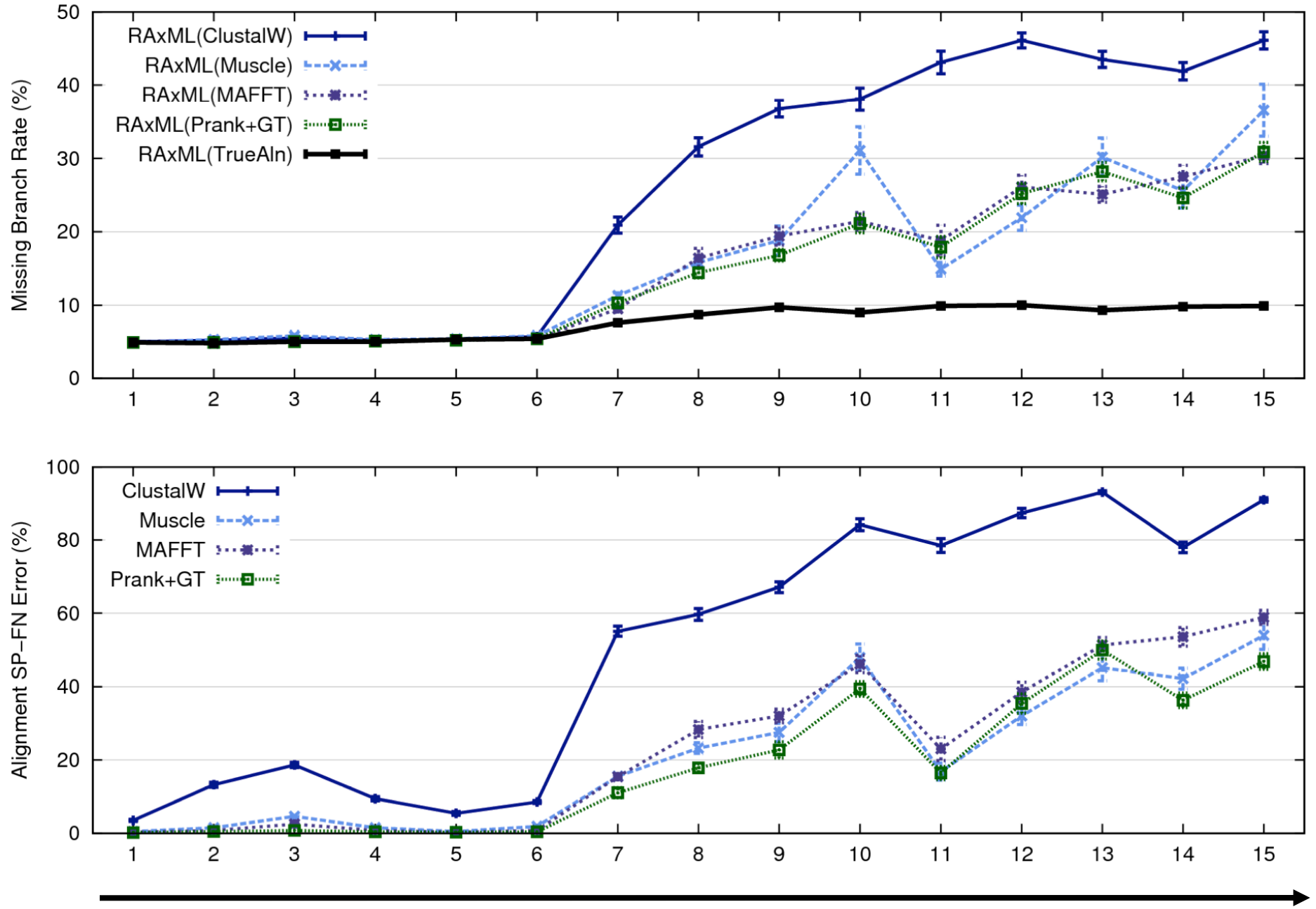
Phylogeny methods

- Bayesian MCMC
- Maximum parsimony
- Maximum likelihood
- Neighbor joining
- FastME
- UPGMA
- Quartet puzzling
- Etc.

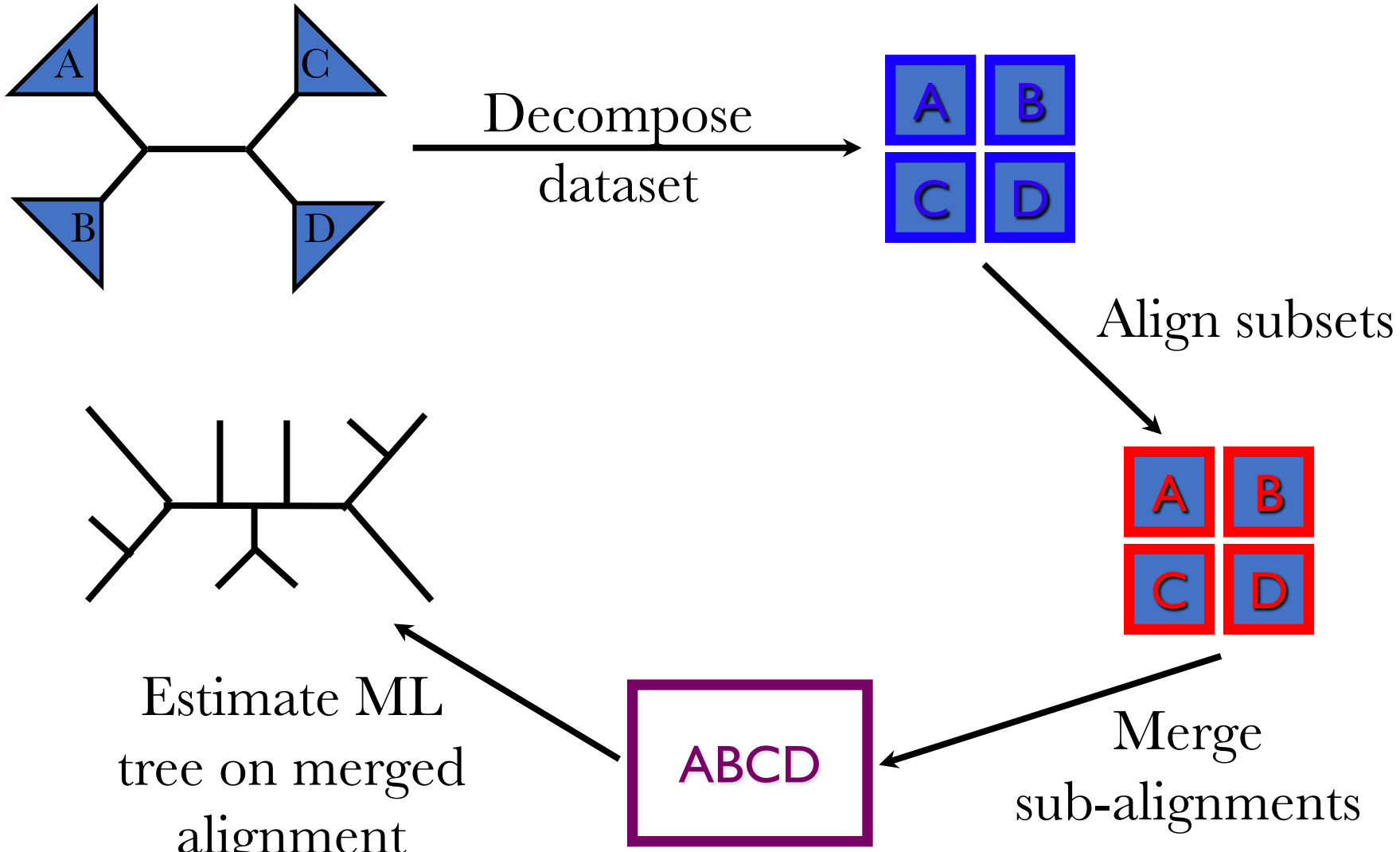
***RAxML**: heuristic for large-scale ML optimization*

Simulation Studies



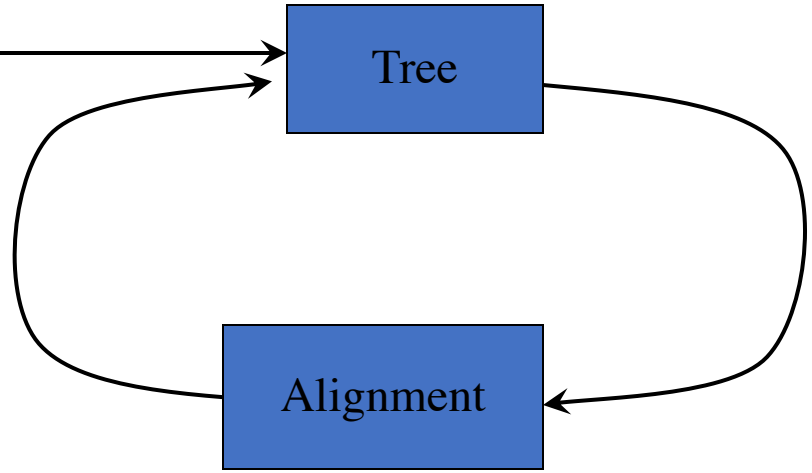

RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Missing Branch Rate (%)
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
ClustalW
Muscle
MAFFT
Prank+GT
Alignment SP–FN Error (%)
0
20
40
60
80
100
1000-taxon models, ordered by difficulty (Liu et al., 2009)

Re-aligning on a tree



SATé, PASTA, and MAGUS Algorithms

Obtain initial alignment and estimated ML tree

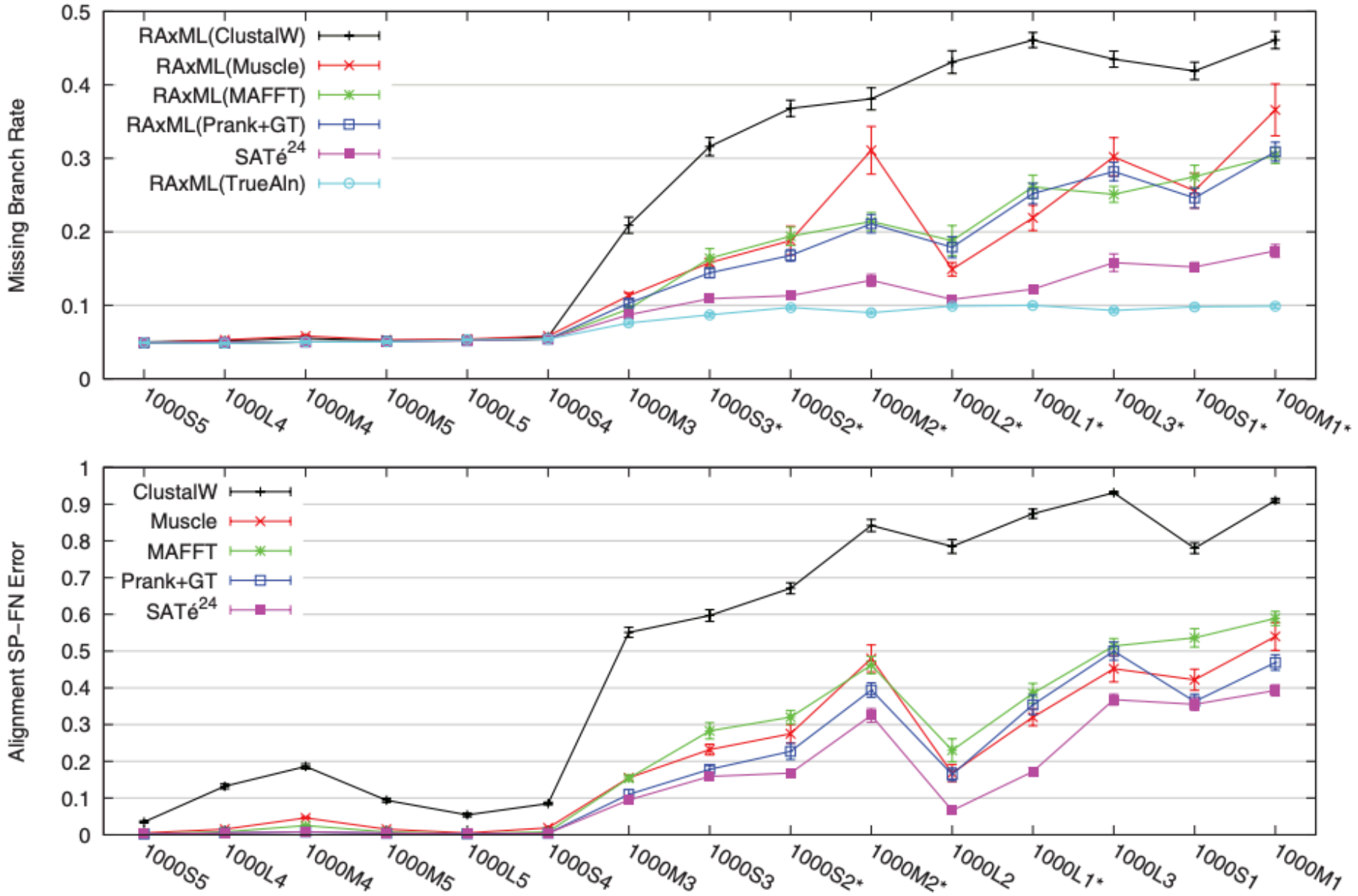


Estimate ML tree on new alignment

Use tree to compute new alignment

Repeat until termination condition, and

return the alignment/tree pair with the best ML score


Missing Branch Rate
0.5
0.4
0.3
0.2
0.1
0
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
SATé24
RAxML(TrueAln)
1000S5
1000L4
1000M4
1000M5
1000L5
1000S4
1000M3
1000S3*
1000S2*
1000M2*
1000L2*
1000L1*
1000L3*
1000S1*
1000M1*
Alignment SP−FN Error
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
ClustalW
Muscle
MAFFT
Prank+GT
SATé24
1000S5
1000L4
1000M4
1000M5
1000L5
1000S4
1000M3
1000S3
1000S2*
1000M2*
1000L2
1000L1*
1000L3
1000S1
1000M1

Improvement over time

- SATé-1 (Science 2009): up to about 8,000
- SATé-2 (Syst Biol 2012): up to 50,000
- PASTA (J Comp Biol 2014): up to 1,000,000
- MAGUS (Bioinformatics 2021): more accurate than PASTA (and one iteration suffices) – up to 1,000,000

Each method improved on the previous with respect to accuracy, speed, and scalability

Statistical Alignment

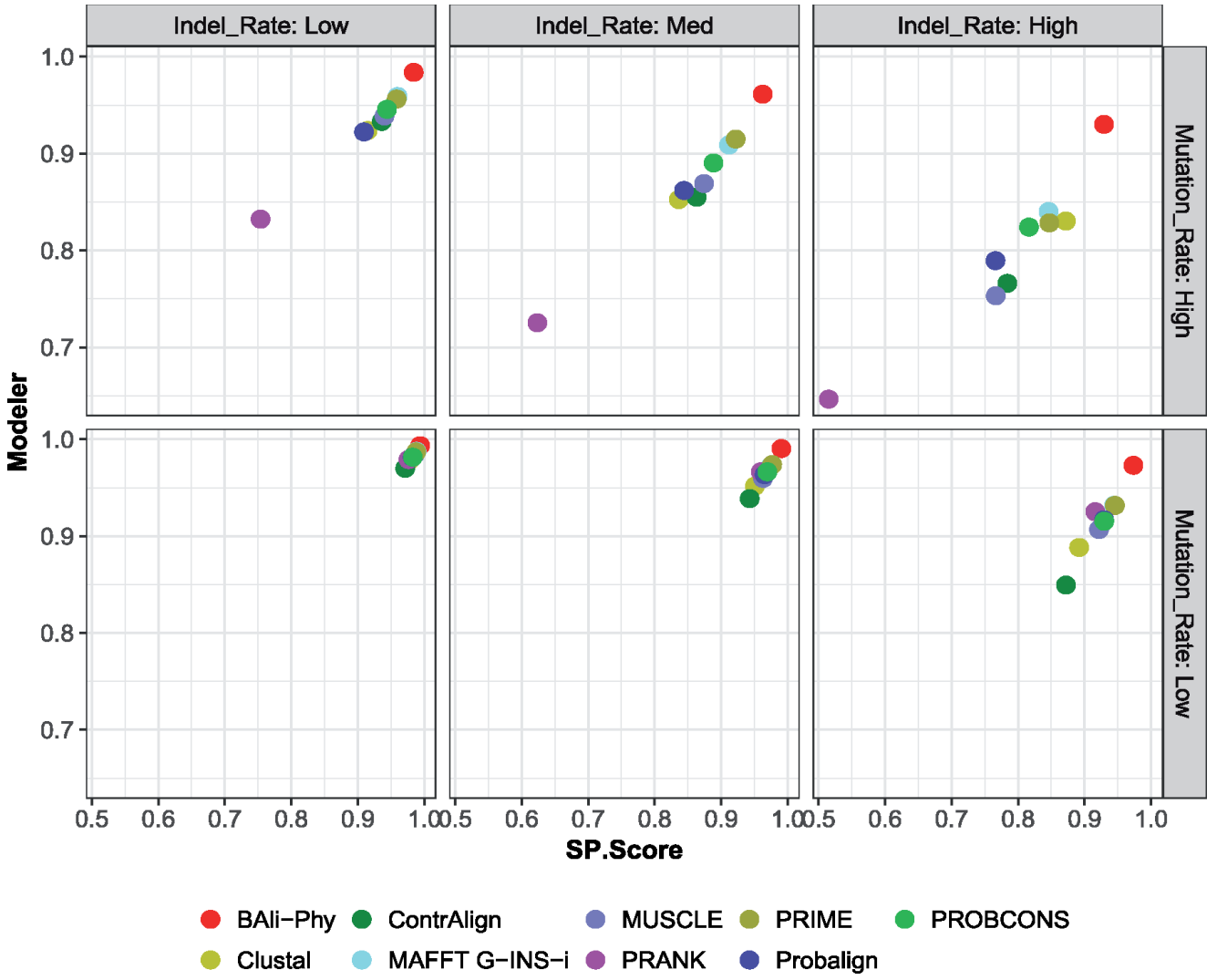
- Since MSA and tree estimation are both about evolution (recognition of homologies due to evolution), can we co-estimate them together, using a statistical model of evolution?
- BAli-Phy (Redelings and Suchard) is the main method for this problem.

But: BAli-Phy is limited to small datasets

From www.bali-phy.org/README.html, **5.2.1. Too many taxa?**

"BAli-Phy is quite CPU intensive, and so we recommend using 50 or fewer taxa in order to limit the time required to accumulate enough MCMC samples. (Despite this recommendation, data sets with more than 100 taxa have occasionally been known to converge.) We recommend initially pruning as many taxa as possible from your data set, then adding some back if the MCMC is not too slow."

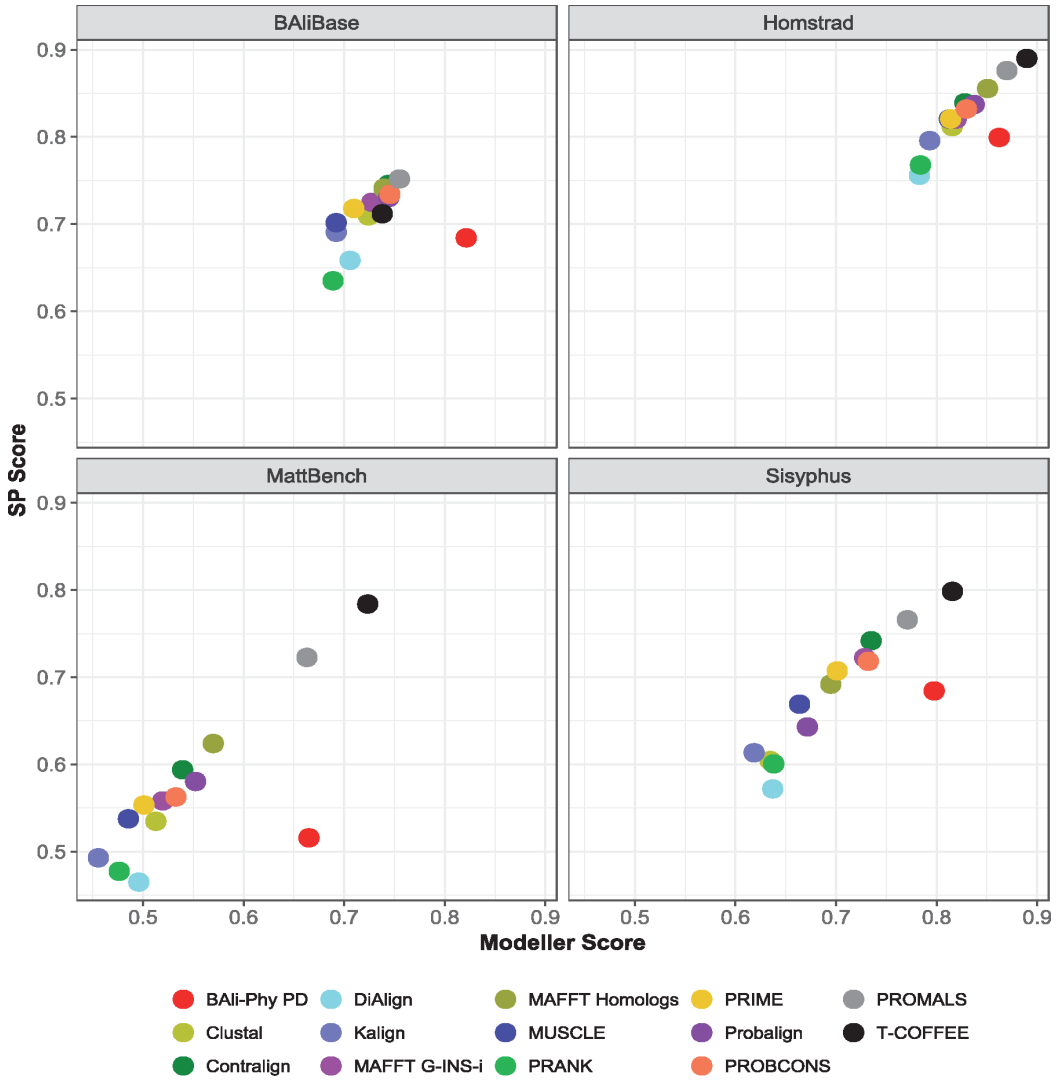
BAli-Phy is best on small simulated protein datasets!



BAli-Phy is best!

Systematic Biology, Volume 68, Issue 3, May 2019, Pages 396–411, https://doi.org/10.1093/sysbio/syy068

BAli-Phy not so great on on 1192 small biological protein datasets



T-Coffee and PROMALS are best!

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

Systematic Biology, Volume 68, Issue 3, May 2019, Pages 396–411, https://doi.org/10.1093/sysbio/syy068

OXFORD UNIVERSITY PRESS

Observations

- Simulated data: Bali-Phy is the best!
- Protein benchmarks: BAli-Phy in middle
  - Good for Modeler score (so low false positives)
  - Not good for SP-score (so high false negatives)
- BAli-Phy under-aligns on biological datasets, but not on simulated datasets

What is going on?

Most likely not an issue of failure of the MCMC analyses to converge (48 hours, 32 processors, < 30 sequences).

Possible explanations:

1. Model misspecification (i.e., BAli-Phy model not appropriate)
2. Structural alignments and evolutionary alignments different
3. The structural alignments are not correct

All these explanations are likely true, but the relative contributions are unknown.

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, and the Ira and Debra Cohen Fellowship to Vlad Smirnov

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

Write to me: warnow@Illinois.edu