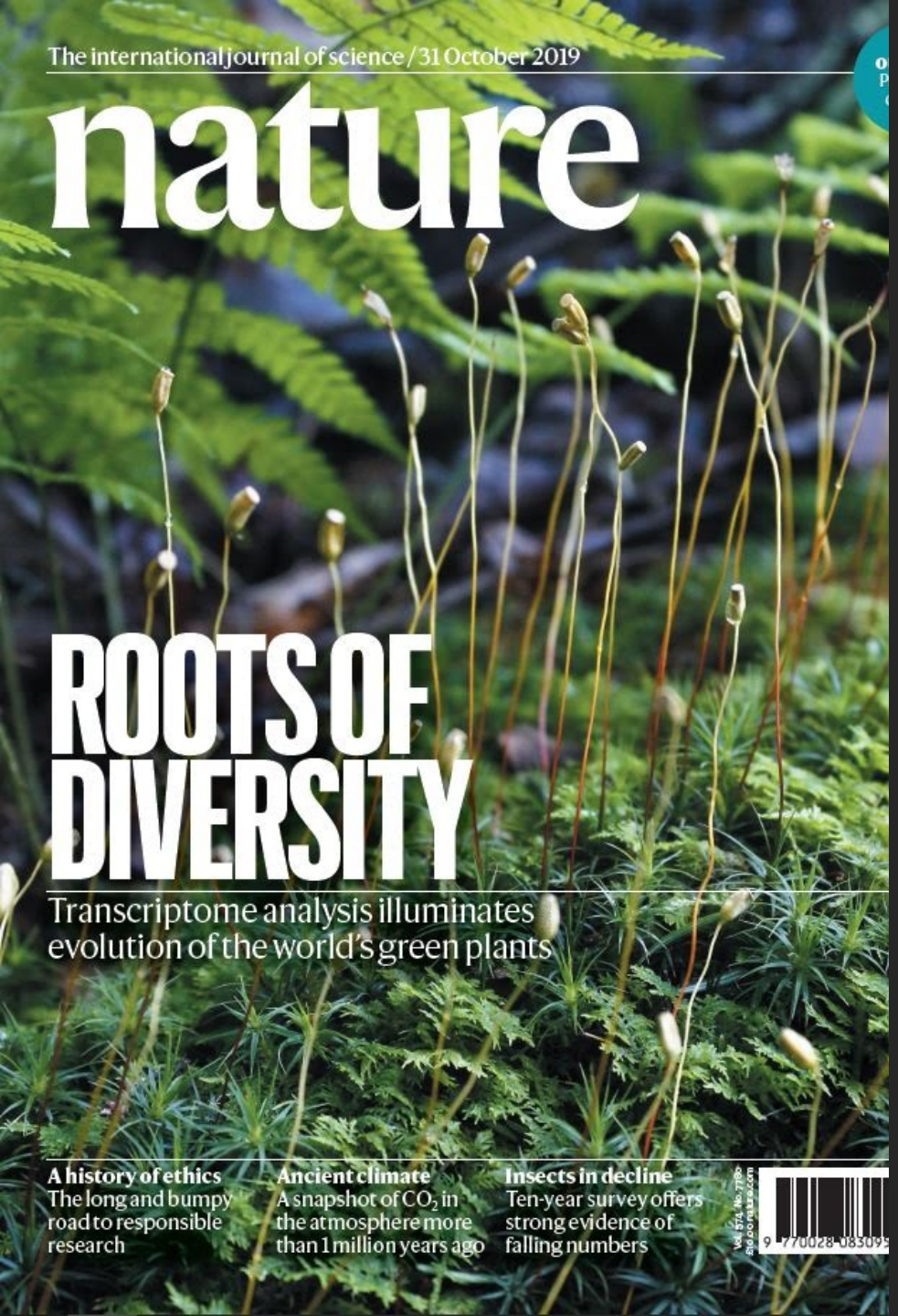
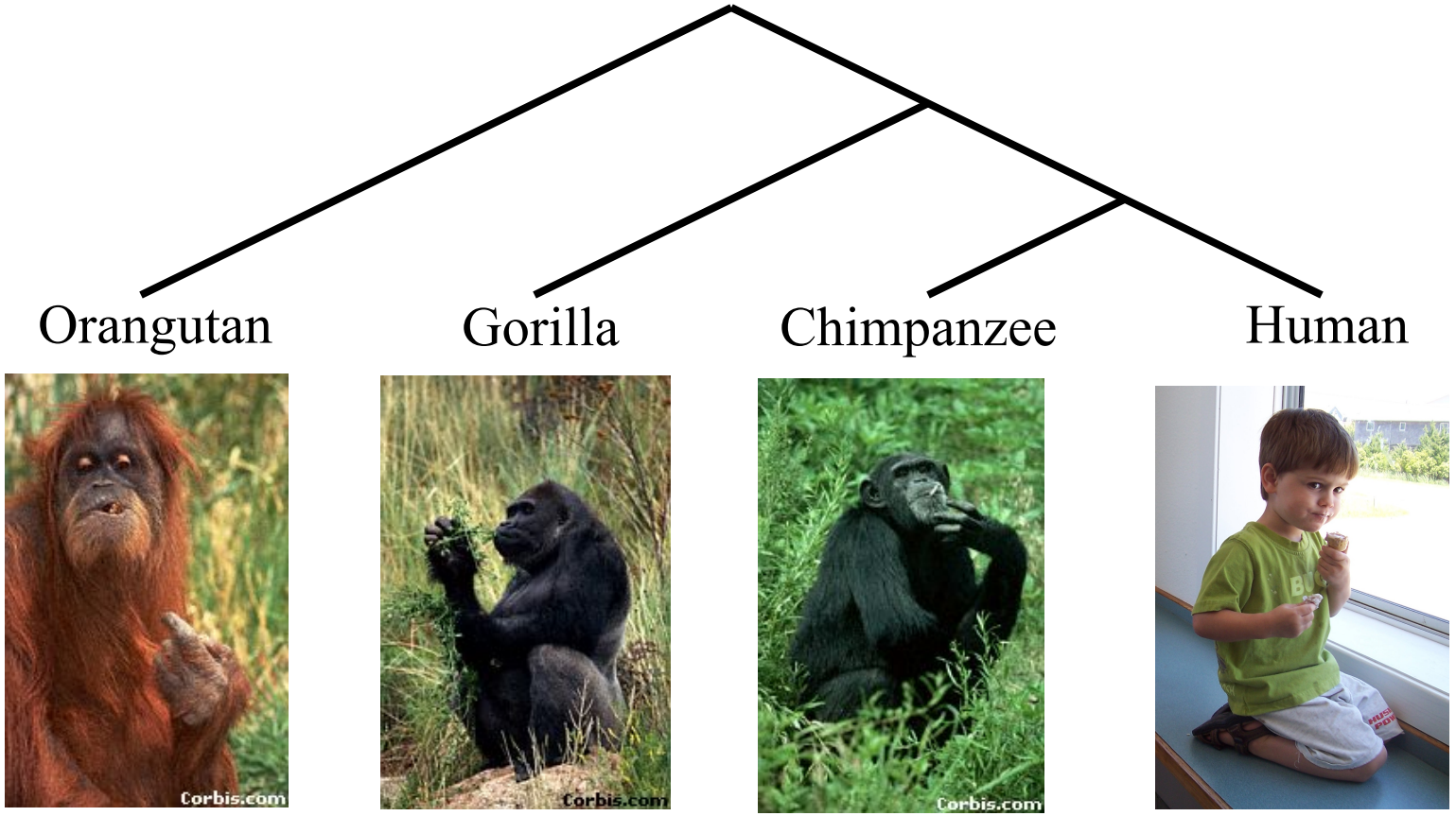


Machine Learning and Discrete Algorithms for Reconstructing the Tree of Life

Tandy Warnow

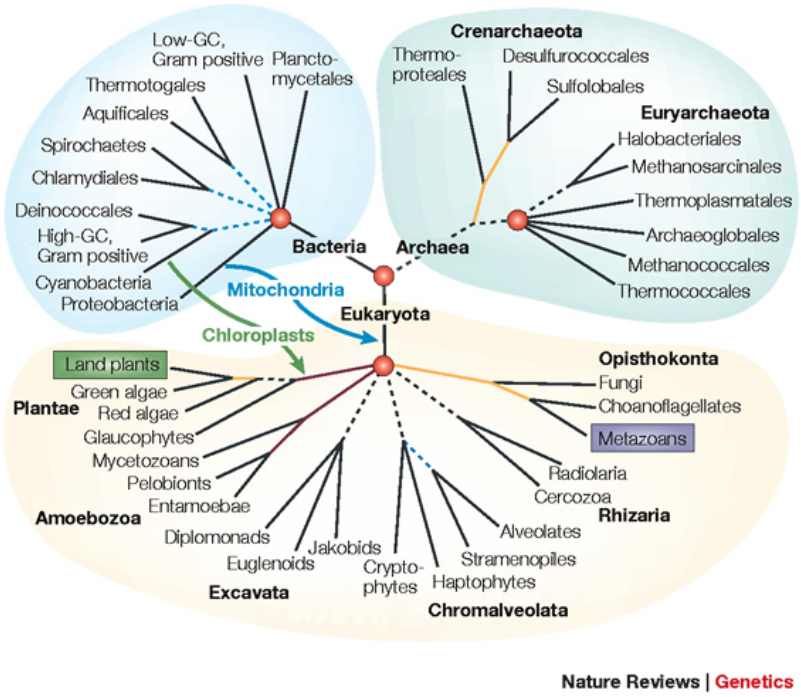
The University of Illinois

Phylogeny (evolutionary tree)



From the Tree of the Life Website,
University of Arizona

Phylogenomics



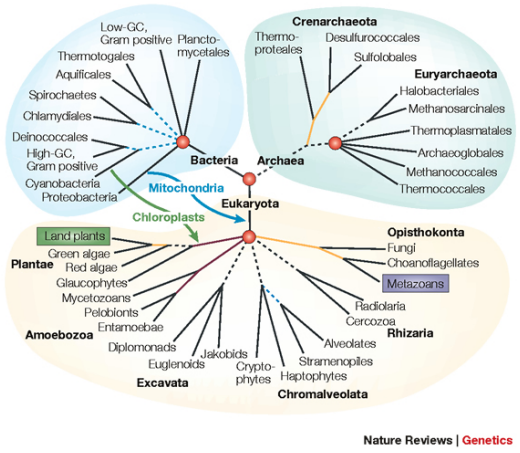
Phylogeny + genomics = genome-scale phylogeny estimation

.

"Resolving the Tree of Life is unquestionably among the most complex scientific problems facing biology and presents challenges much greater than sequencing the human genome."

From "Assembling the Tree of Life: Harnessing Life's History to Benefit Science and Society," National Science Foundation (2002), available at http://ucjeps.berkeley.edu/tol.pdf

Phylogenetic Inference



Nature Reviews | Genetics

“Big Data”:

- Heterogeneous
- Large
- Noisy
- Error-ridden
- Streaming
- Model-misspecification

Approaches:

- NP-hard optimization problems and large datasets
- Statistical estimation under stochastic models of evolution
- Probabilistic analysis of algorithms
- Graph-theoretic divide-and-conquer
- Chordal graph theory
- Combinatorial optimization

This talk

- Fast introduction to phylogenetic estimation, in a statistical framework
- ASTRAL – fast and accurate (and statistically consistent) species tree estimation addressing Incomplete Lineage Sorting (ILS)
- TreeMerge: enabling ASTRAL to run on large datasets
- FastMulRFS: fast and accurate (and statistically consistent) species tree estimation addressing Gene Duplication and Loss
- Discussion

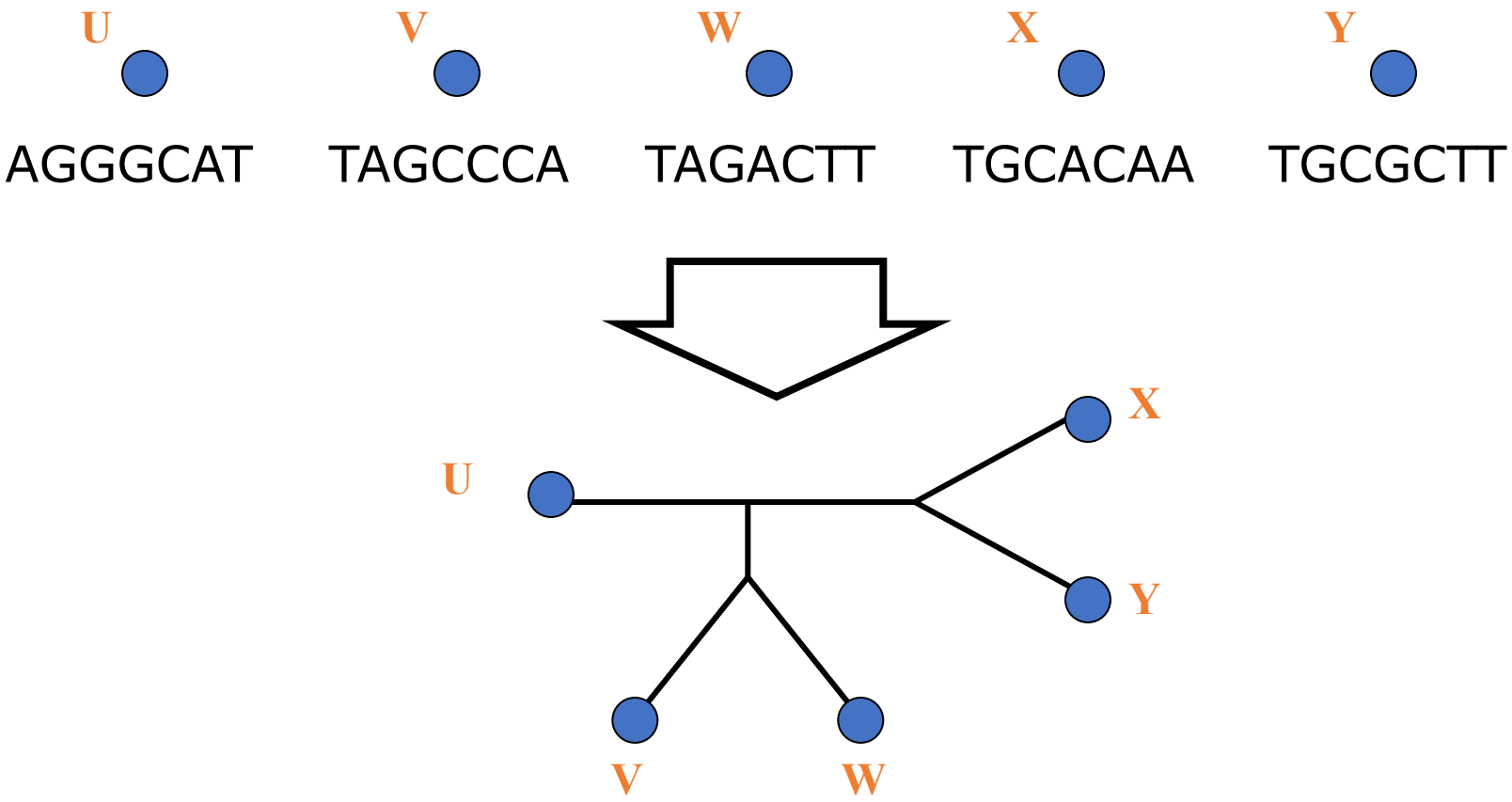
Future directions and themes

- Standard approaches take hundreds of CPU years (even for small numbers of species) for genome-scale data
- Even single genes can take weeks or months of CPU time
- Distributed computing and parallel computing inherent in phylogenomics
- Similar challenges for the problem of computing multiple sequence alignments

Divide-and-conquer approaches improve accuracy and running time

- Initial decomposition can be based on a tree, but for very large datasets novel clustering methods are needed
- Graph theory is used for statistical consistency guarantees
- Many open problems for phylogeny estimation and multiple sequence alignment

Phylogeny Problem



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

NP-hard optimization problems

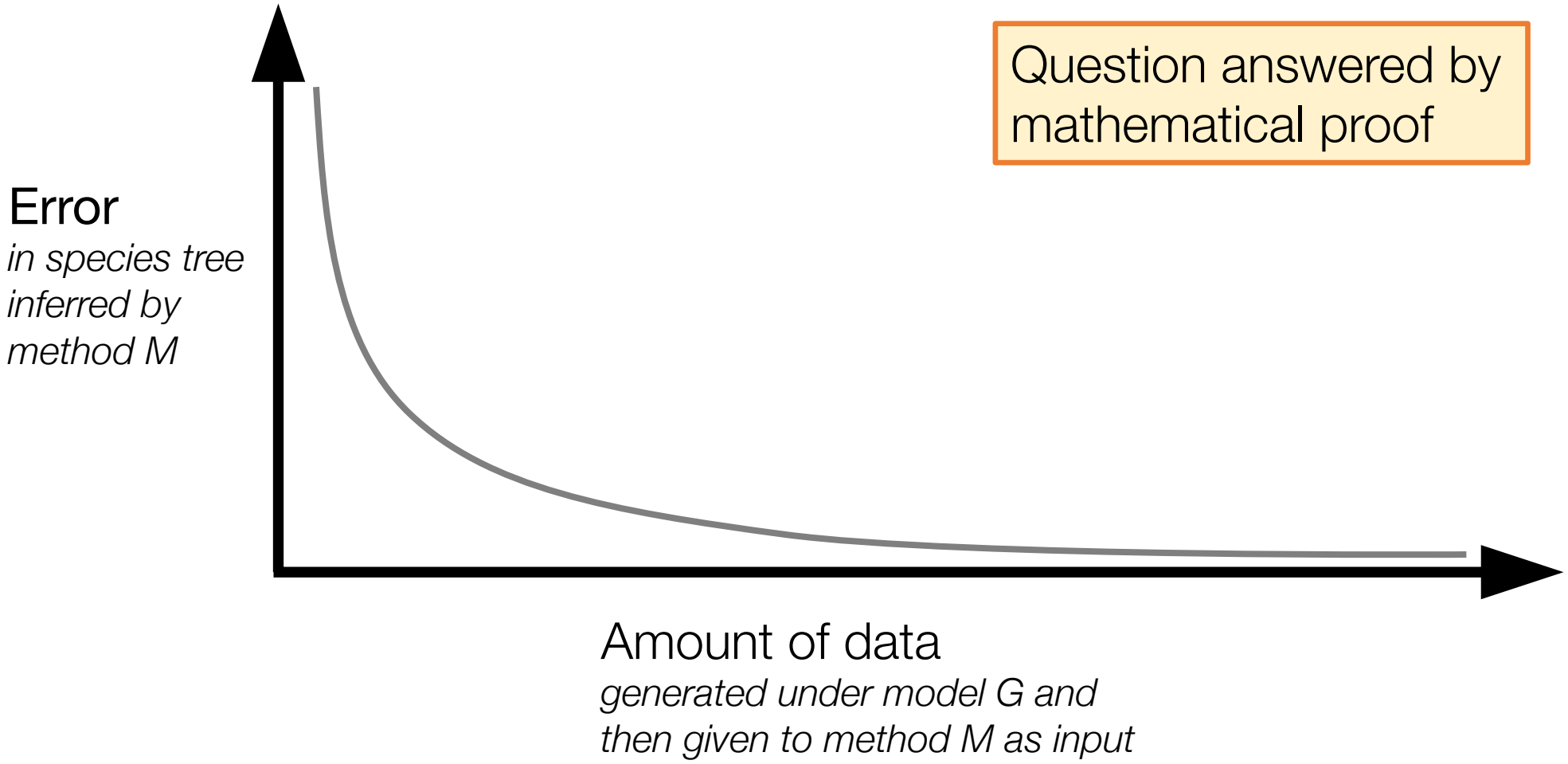
Large datasets

Years of CPU time for standard methods

This research combines many types of computer science:

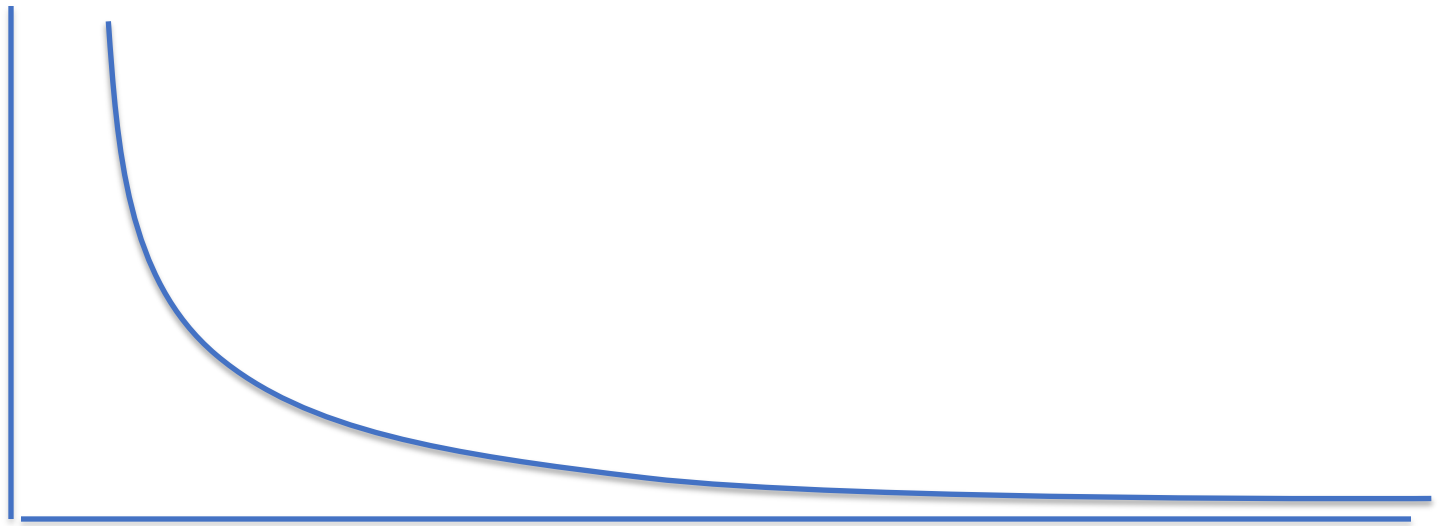
Algorithm design, **proofs**, implementation, simulations and testing

Is method M statistically consistent under model G?



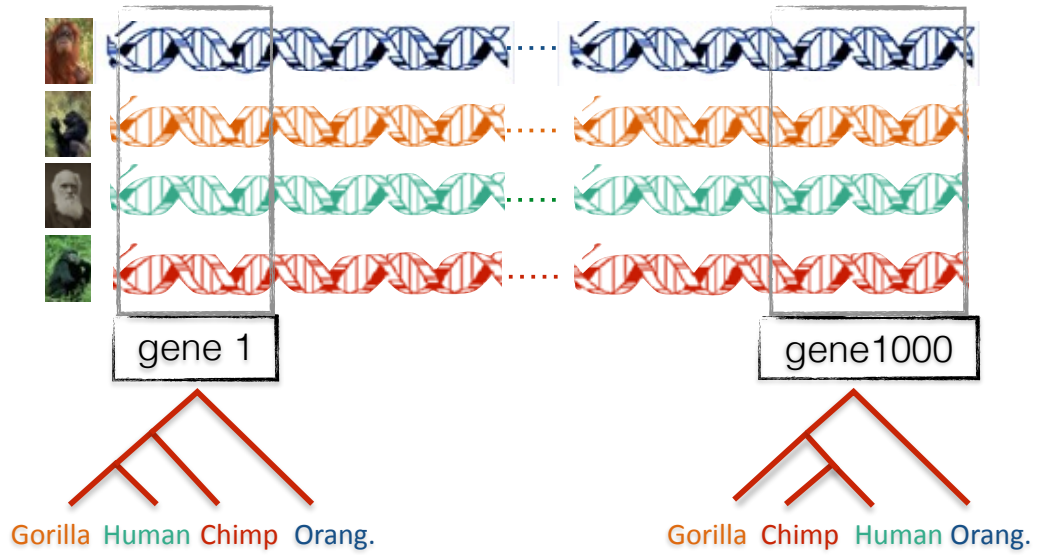
Genome-scale data?

error



Length of the genome

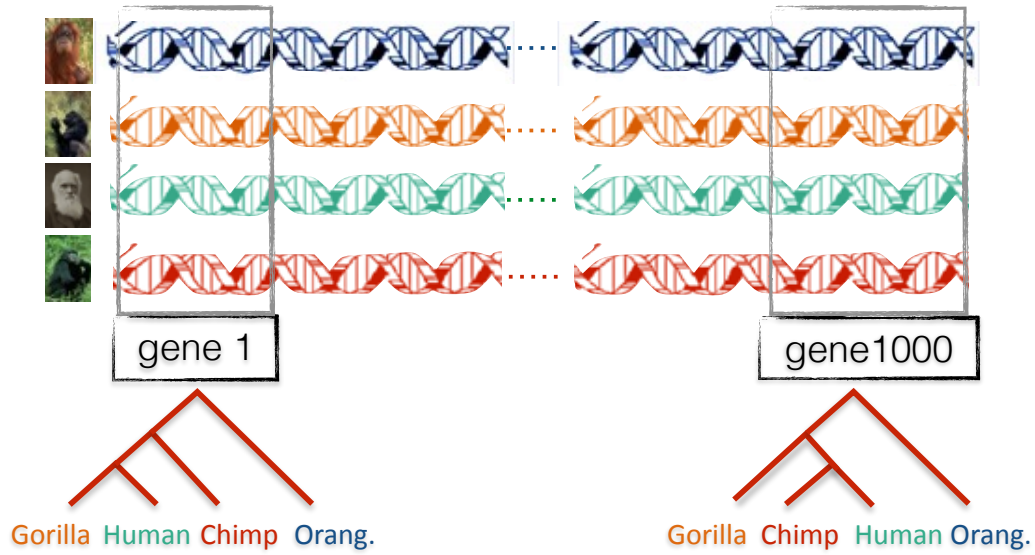
Gene tree discordance



Multiple causes for discord, including

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

Gene tree discordance



Multiple causes for discord, including

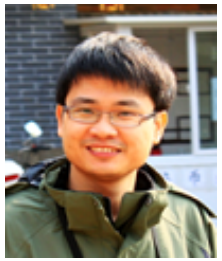
- Incomplete Lineage Sorting (ILS),
- Gene Duplication and Loss (GDL), and
- Horizontal Gene Transfer (HGT)

Avian Phylogenomics Project

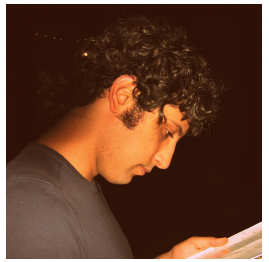
Erich Jarvis, HHMI

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
- 250 CPU years to estimate tree with heuristic maximum likelihood method

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

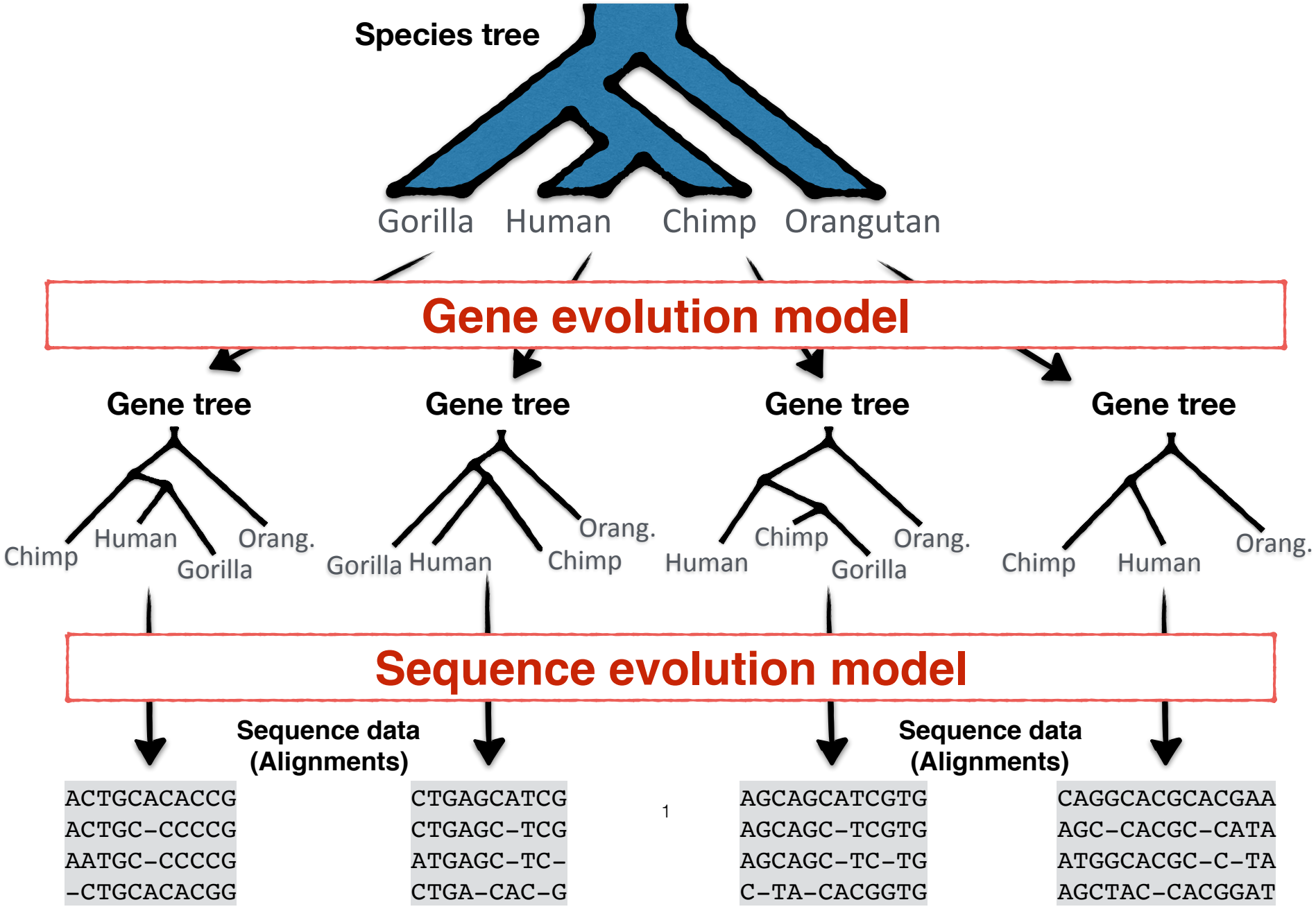
S. Mirarab,
UT-Austin /UCSD

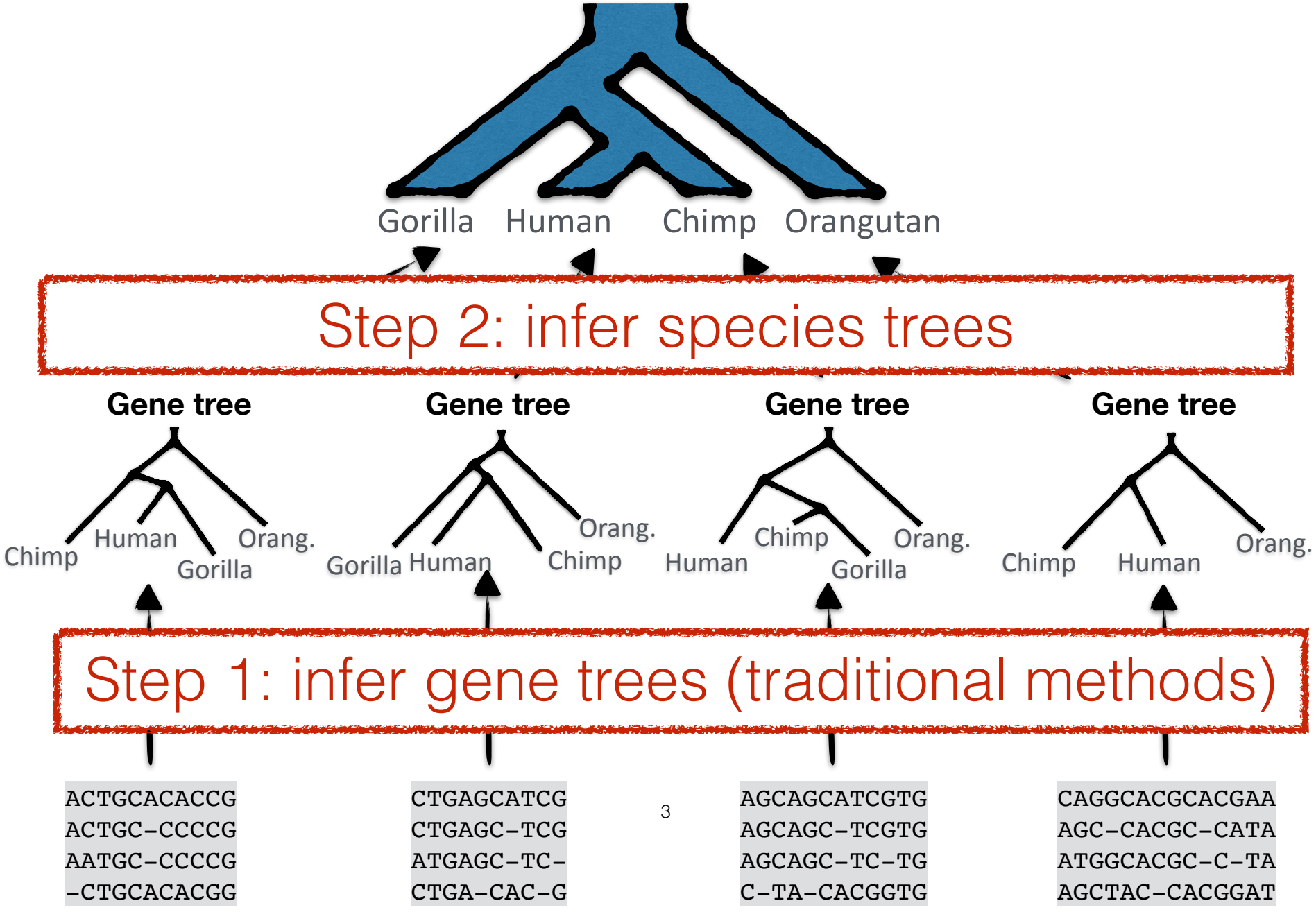
N. Nguyen
UT-Austin/UCSD

- 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


Species tree
Gorilla
Human
Chimp
Orangutan
Gene evolution model
Gene tree
Gene tree
Gene tree
Gene tree
Chimp
Human
Orang.
Gorilla
Gorilla
Human
Orang.
Chimp
Human
Chimp
Orang.
Gorilla
Chimp
Human
Orang.
Sequence evolution model
Sequence data
(Alignments)
Sequence data
(Alignments)
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT


Gorilla Human Chimp Orangutan
Step 2: infer species trees
Gene tree
Gene tree
Gene tree
Gene tree
Chimp Human Orang. Gorilla
Gorilla Human Orang. Chimp
Human Chimp Orang. Gorilla
Chimp Human Orang.
Step 1: infer gene trees (traditional methods)
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT

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

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**: 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

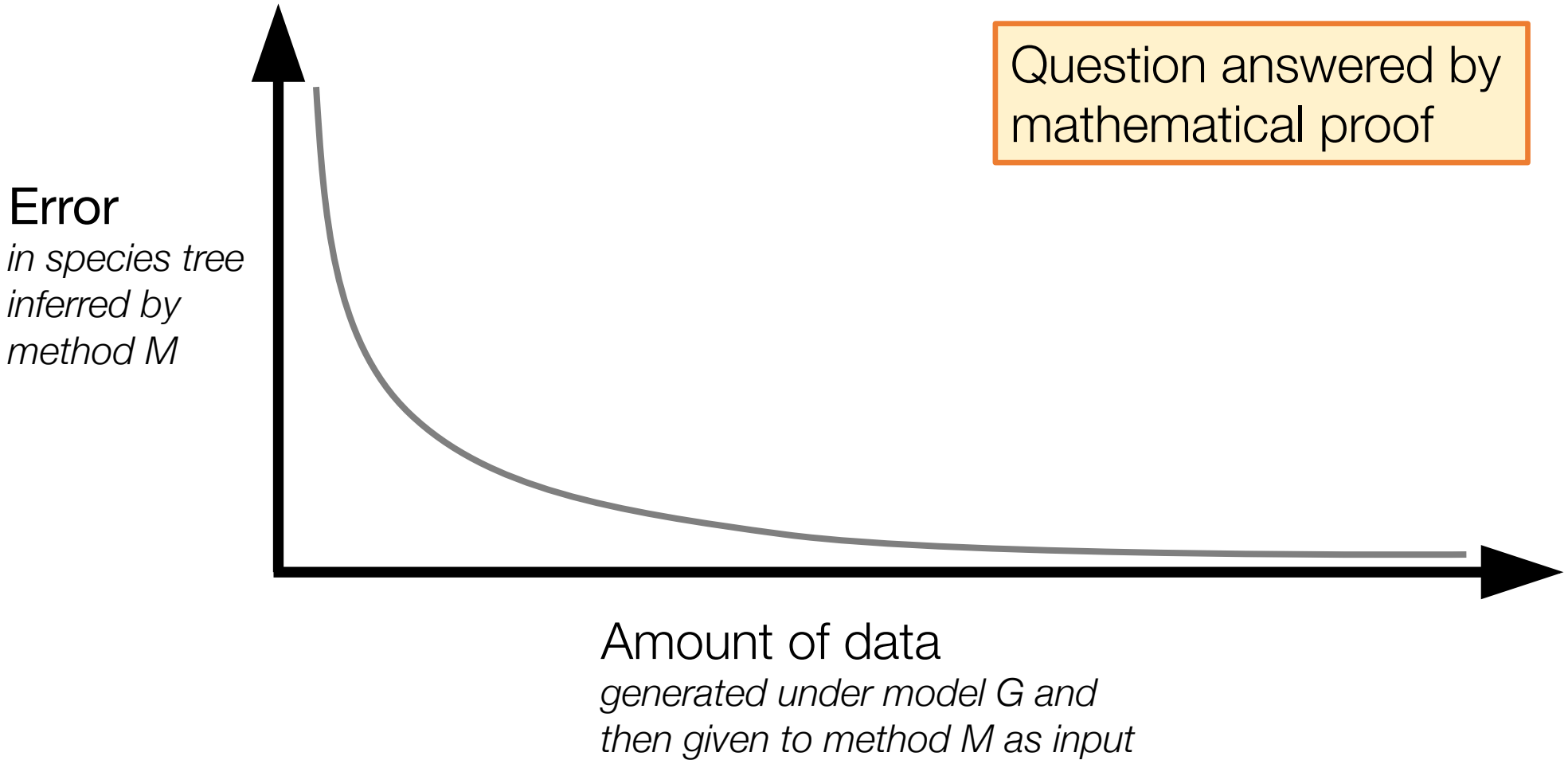
a gene tree

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

Is method M statistically consistent under model G?



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

DOI:10.1093/sysbio/syv029

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,*], Joseph P. Gallagher[a], Josef J. Jareczek[a], Justin T. Page[b], Joshua A. Udall
Michael A. Gore[c], Jonathan F. Wend[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[1,2,3,4], Daniel J. Field[1,2,5], Jeffrey P. Townsend[1,6],
Emily Moriarty Lemmon[7] & Alan R. Lemmon[8]

X

ASTRAL – pros and cons

- The good: ASTRAL is
  - Most popular statistically consistent method for species tree estimation among biologists
  - Very fast for many datasets (much faster than concatenation)
- The mixed:
  - Concatenation can be more accurate under some conditions
- The bad:
  - ASTRAL can fail to complete on large enough datasets within reasonable time frames (days of computation)

The alternatives are worse

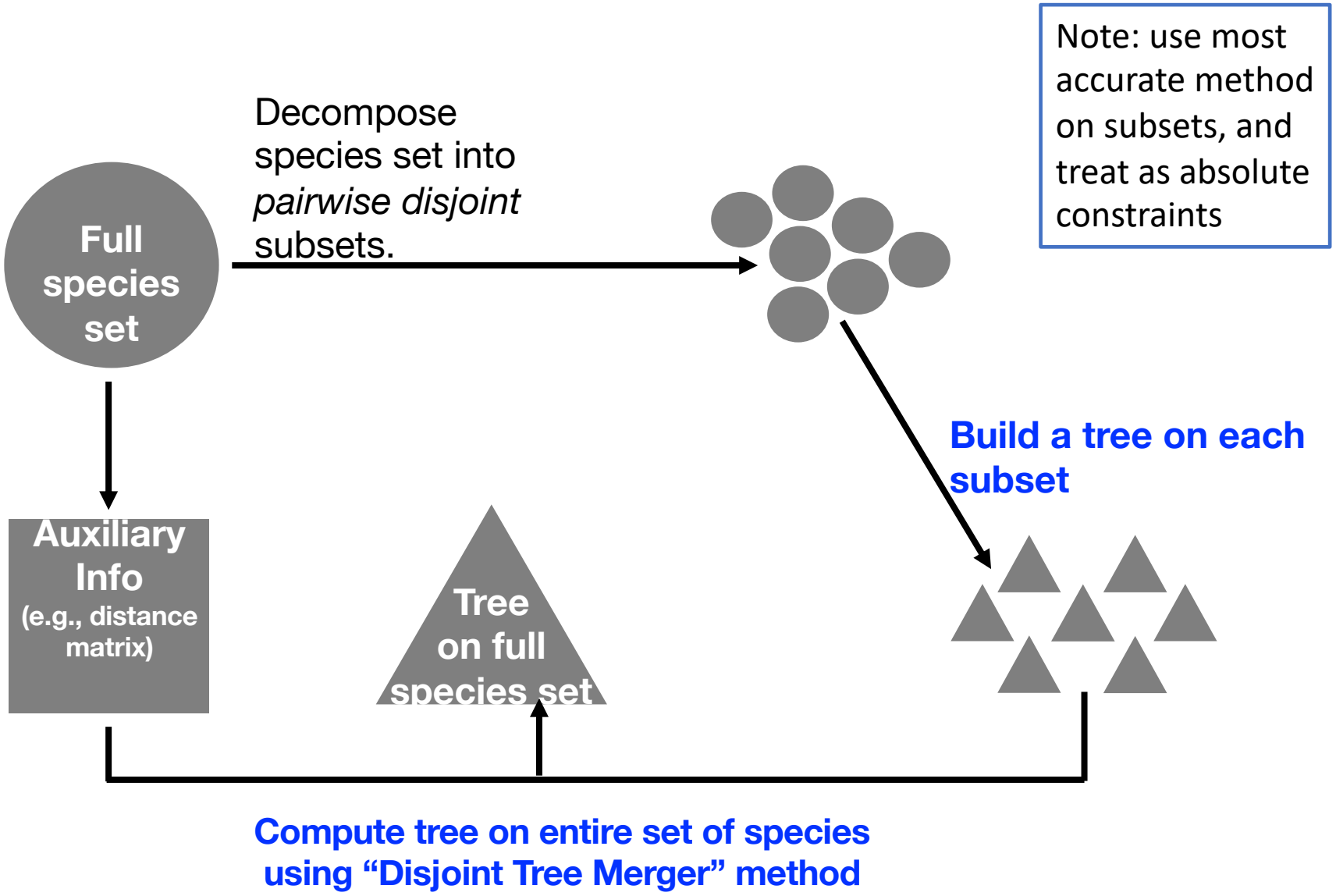
- Concatenation Analyses (e.g., using RAxML):
  - most commonly used method, not statistically consistent, sometimes more accurate than summary methods
  - computationally intensive (e.g., 250 CPU years for the Avian Phylogenomics project with only 48 species) and do not scale to large numbers of species
- Co-estimation of gene trees and species trees: too expensive
- Other statistically consistent methods: not as accurate as ASTRAL

So we need to make ASTRAL truly scalable to large datasets!

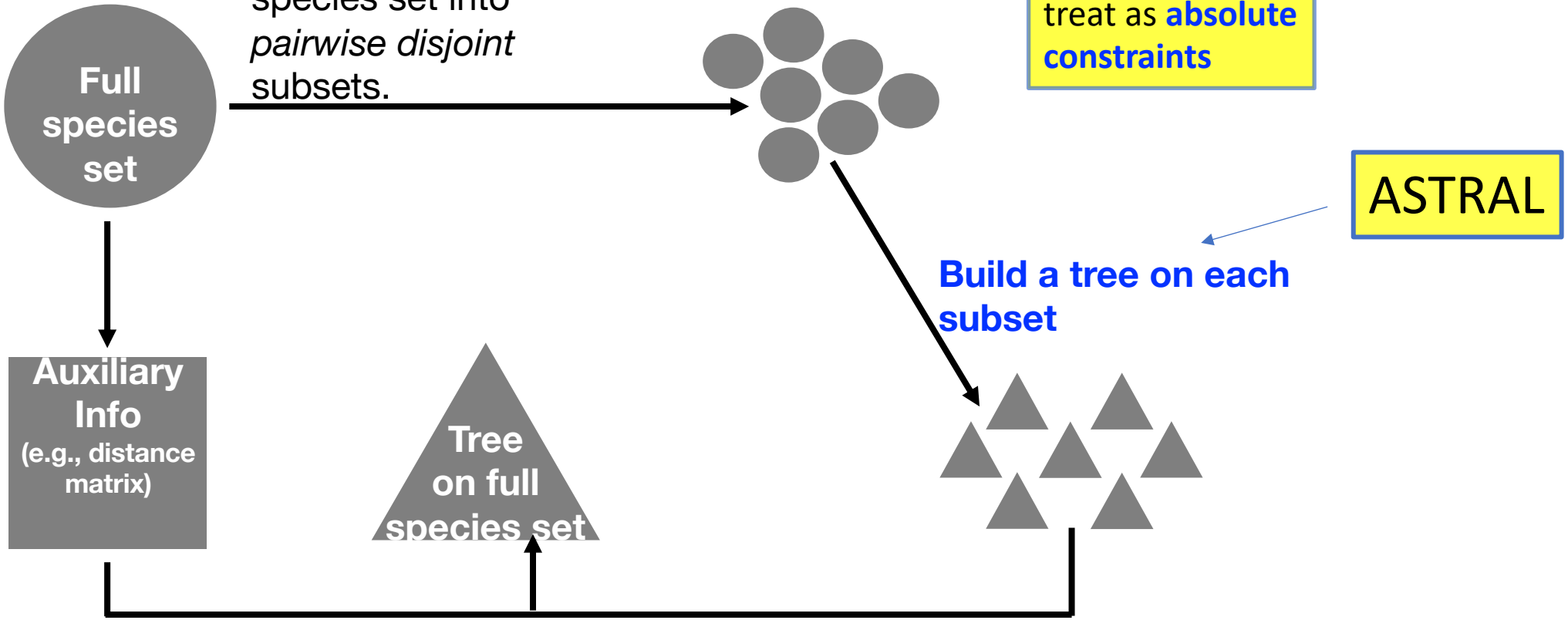
Disjoint Tree Mergers

- Molloy and Warnow, introduced in RECOMB-CG 2018
- Divide-and-conquer:
  - divides species set into disjoint subsets,
  - computes species trees on the subsets using selected species tree method (e.g., ASTRAL, RAxML, SVDquartets),
  - then merges subset trees using other information (computed from input)

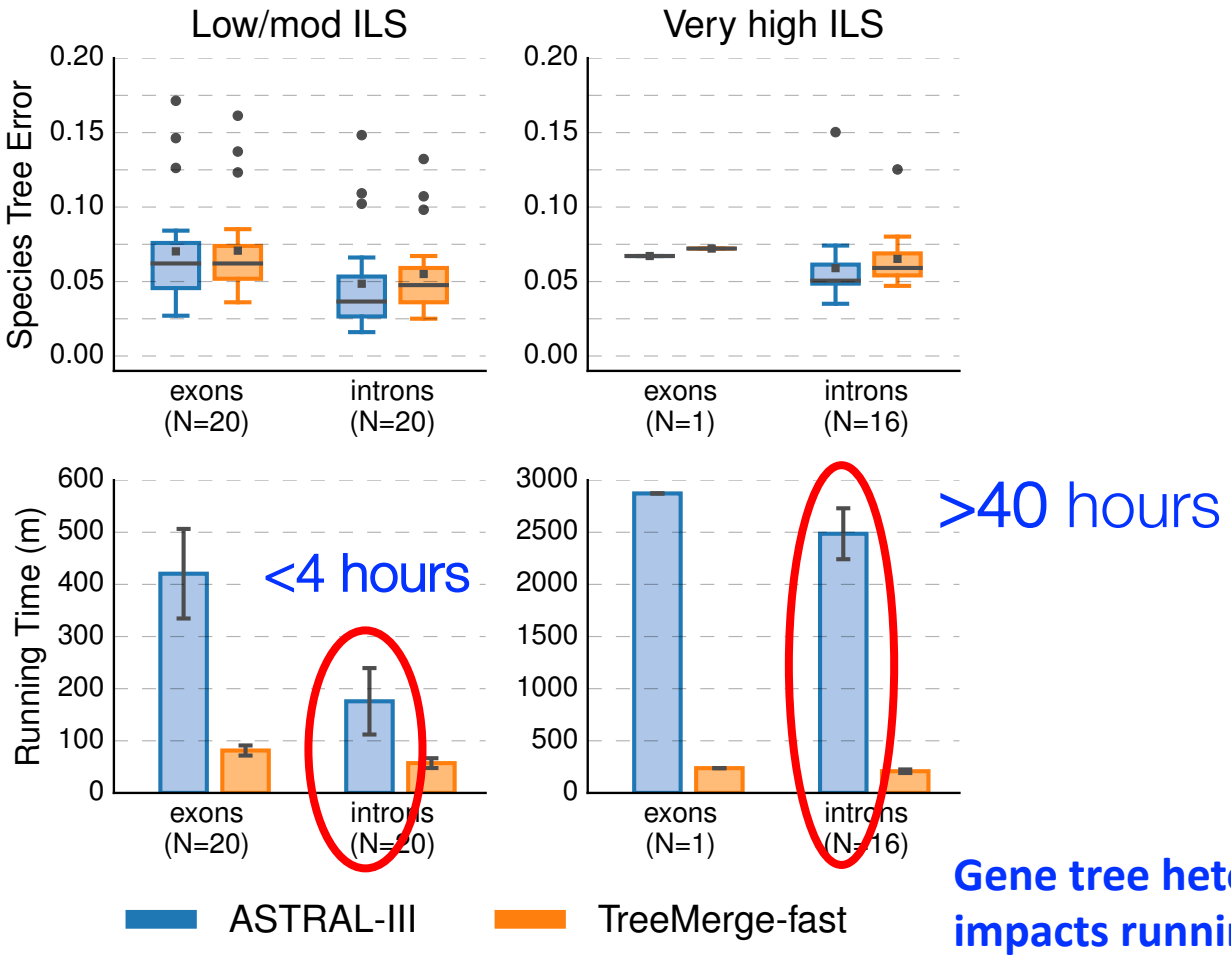
Divide-and-Conquer using Disjoint Tree Mergers



Divide-and-Conquer using Disjoint Tree Mergers



TreeMerge with ASTRAL



Theorem: TreeMerge enables polynomial time and statistically consistent species tree estimation pipelines.

Empirical: TreeMerge improves running time without sacrificing accuracy.

Gene tree heterogeneity impacts running time of ASTRAL.

Molloy and Warnow, ISMB 2019.

DTM Methods

- **NJMerge** (Molloy, Warnow 2019, Algorithms for Molecular Biology)
  - Can fail given more than two trees
- **TreeMerge** (Molloy, Warnow 2019, Bioinformatics)
  - Uses NJMerge to combine pairs of constraint trees, then merges via shared backbones
  - Cannot fail, faster than NJMerge, but slightly less accurate
- **Constrained INC** (Zhang, Rao, Warnow 2019, Algorithms for Molecular Biology)
  - Incrementally assembles a tree through a quartet voting process, cannot fail
- **Guide Tree Merger** (Smirnov and Warnow, 2020, BMC Genomics)
  - Uses computed guide tree to merge constraint tree, does not allow blending
  - Faster than the other methods, and as accurate

All these methods have strong statistical properties (e.g., maintaining statistical consistency) and are polynomial time

Runtime Results

- **DTM Methods**
  - TreeMerge: **10 min**
  - NJMerge: **10-30 min**
  - GTM: **0.5 sec**
- **ASTRAL-GTM vs ASTRAL**
  - 10 genes: **98 sec vs 2.4 hours**
  - 1000 genes: **2.2 hours vs 42.5 hours**
- **RAxML-GTM vs RAxML**
  - 10 genes: **15 min vs. 2 hours**
  - 1000 genes: **20 hours (completed) vs 48 hours (capped)**

GTM-ASTRAL vs ASTRAL accuracy

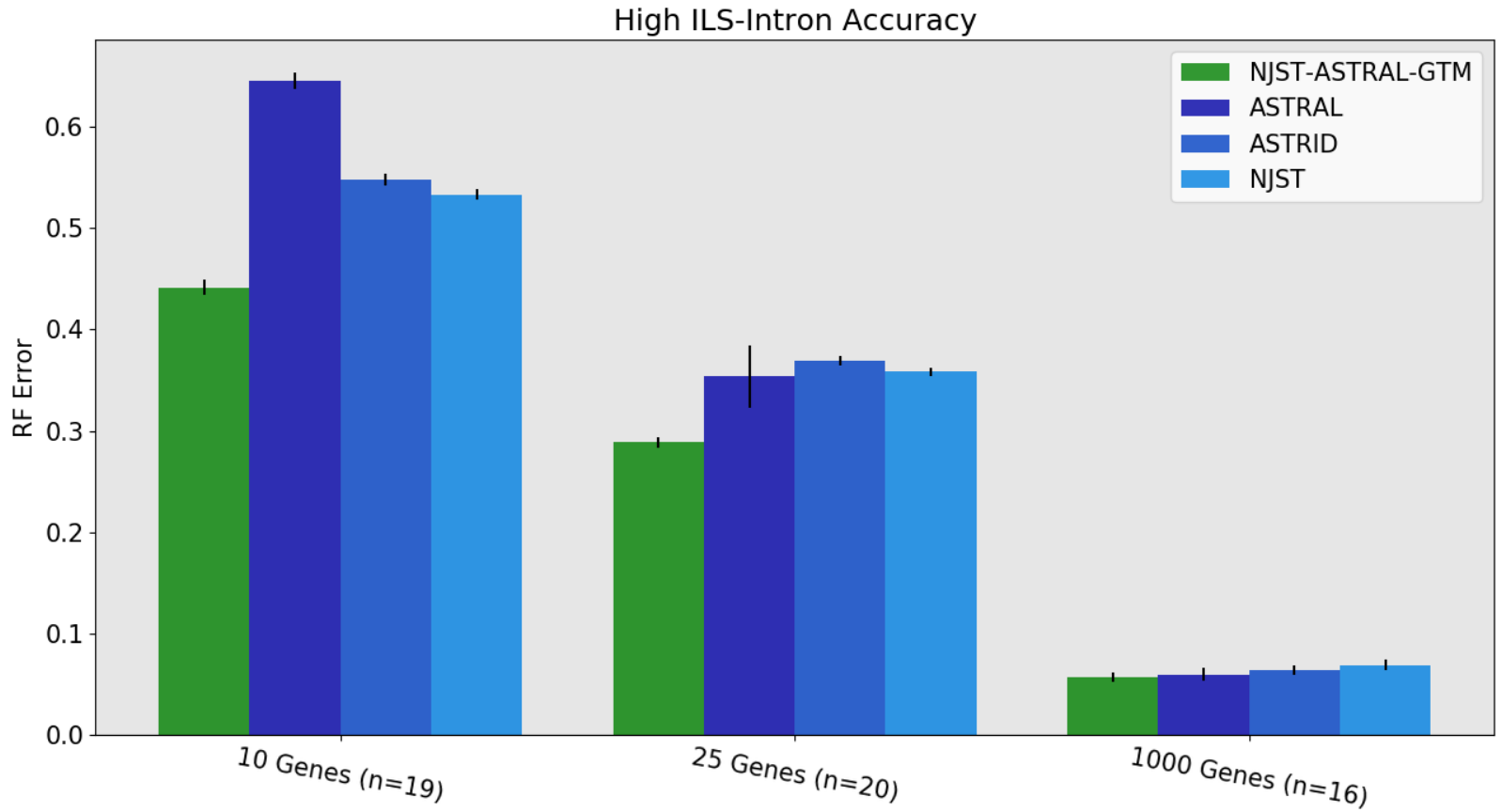
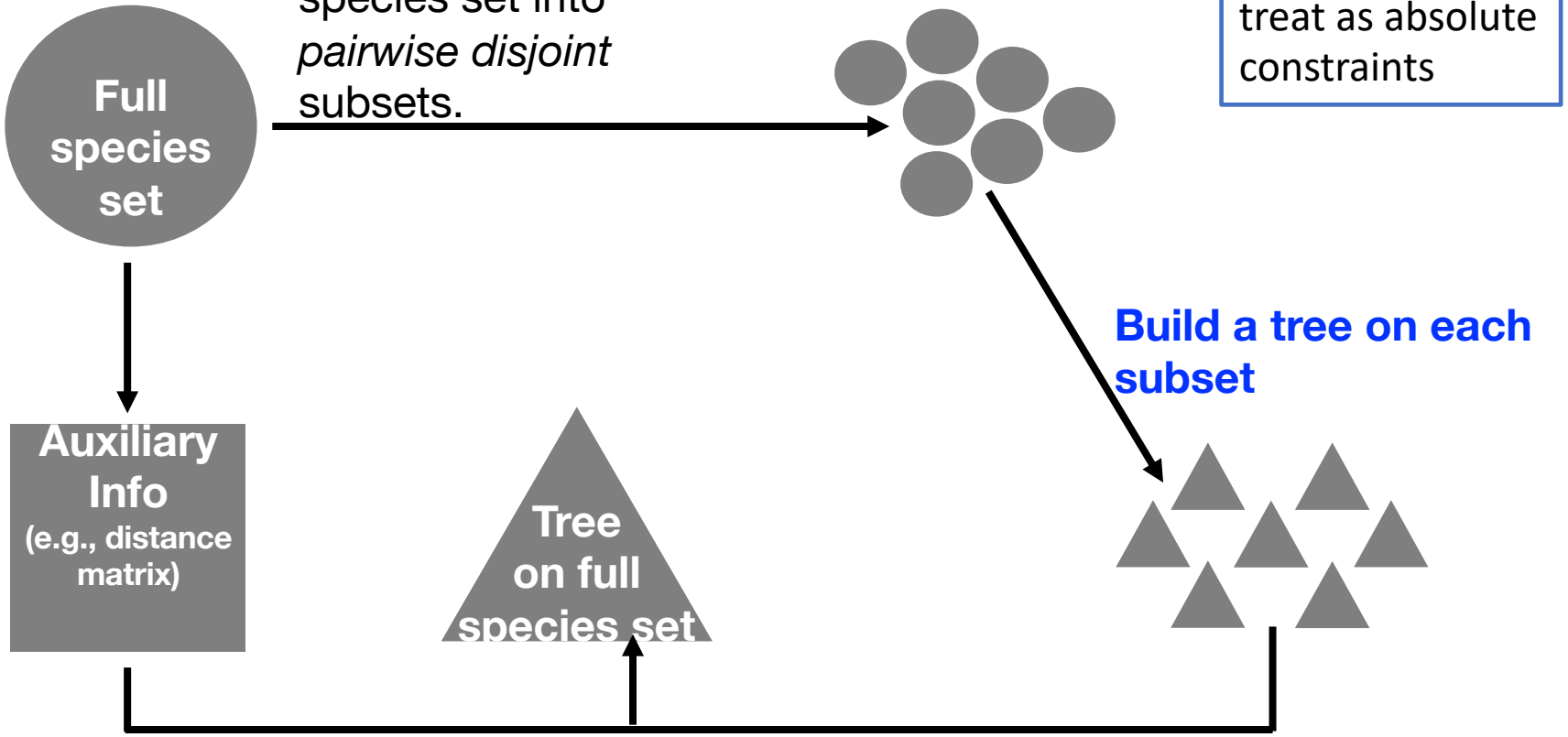


Table 4 Comparison of average runtime (seconds) of NJst-ASTRAL-GTM and ASTRAL for high ILS conditions with introns on 1000 species

	NJst-ASTRAL-GTM	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

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 [3]

DTMs can be used for any tree estimation problem



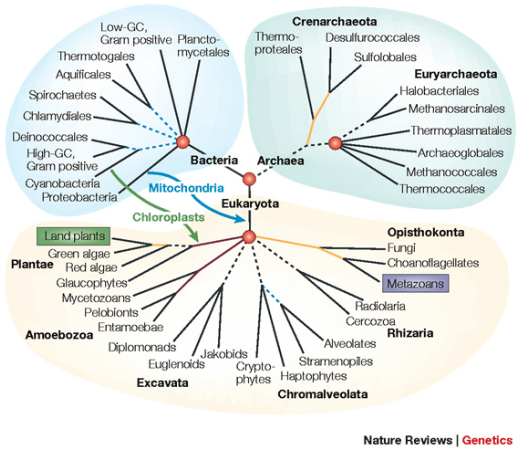
Future directions and themes

- Standard approaches take hundreds of CPU years (even for small numbers of species) for genome-scale data
- Even single genes can take weeks or months of CPU time
- Distributed computing and parallel computing inherent in phylogenomics
- Similar challenges for the problem of computing multiple sequence alignments (PASTAspark)

Divide-and-conquer approaches improve accuracy and running time

- Initial decomposition can be based on a tree, but for very large datasets novel clustering methods are needed
- Graph theory is used for statistical consistency guarantees
- Many open problems for phylogeny estimation and multiple sequence alignment

Phylogenetic Inference



Nature Reviews | Genetics

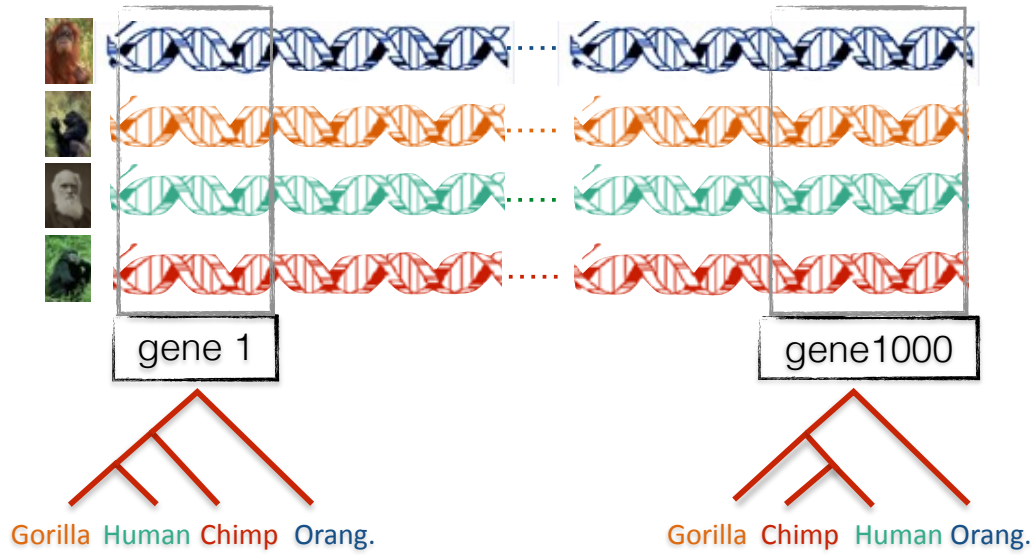
“Big Data”:

- Heterogeneous
- Large
- Noisy
- Error-ridden
- Streaming
- Model-misspecification

Approaches:

- NP-hard optimization problems and large datasets
- Statistical estimation under stochastic models of evolution
- Probabilistic analysis of algorithms
- Graph-theoretic divide-and-conquer
- Chordal graph theory
- Combinatorial optimization

Gene tree discordance



Multiple causes for discord, including

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

Gene Family Trees

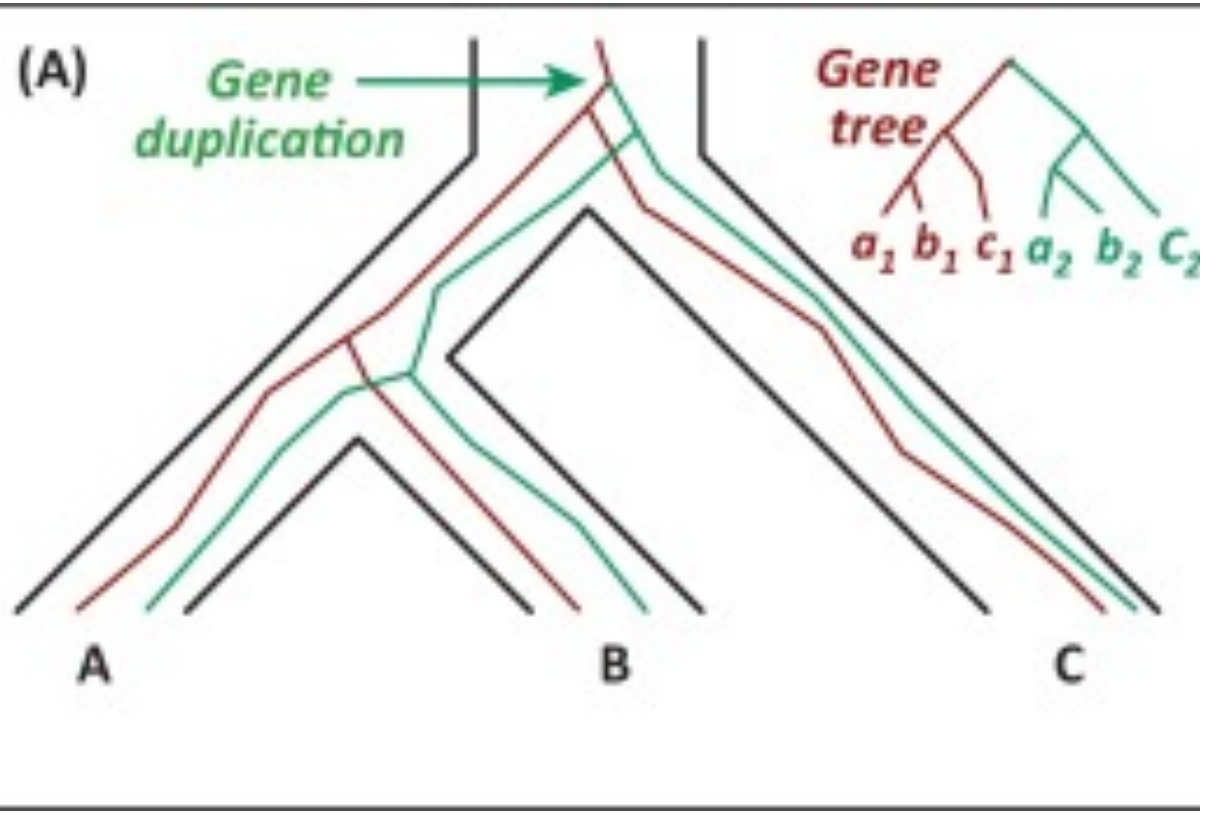
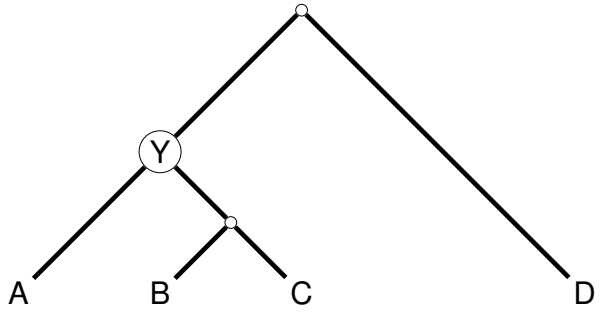


Figure by Luay Nakhleh, TREE 2013

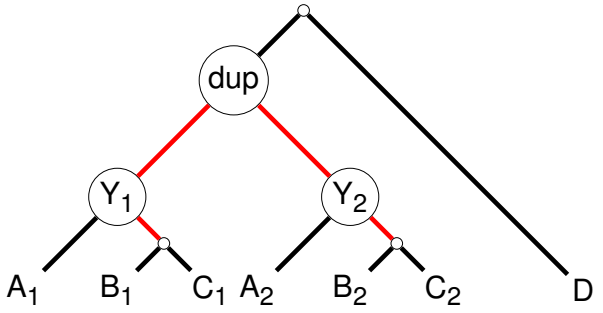
The species tree has one duplication (at the root), which produces a gene family tree that has two copies of the species tree!

Multi-copy trees: MUL-trees

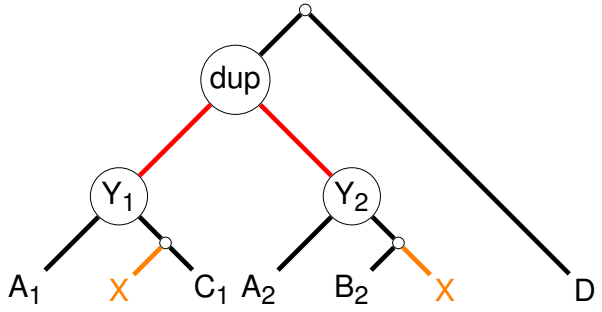
Problem: Given set of MUL-trees, infer the species tree



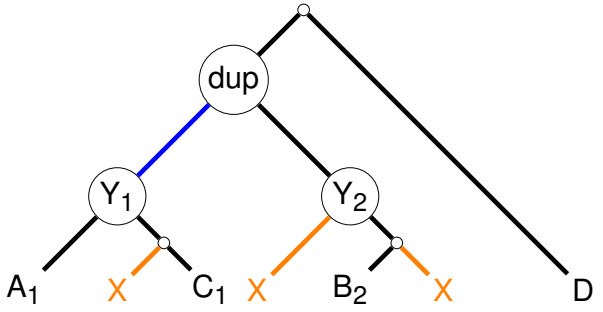
(a) Species tree T^*



(b) Gene tree M_1 with one duplication.



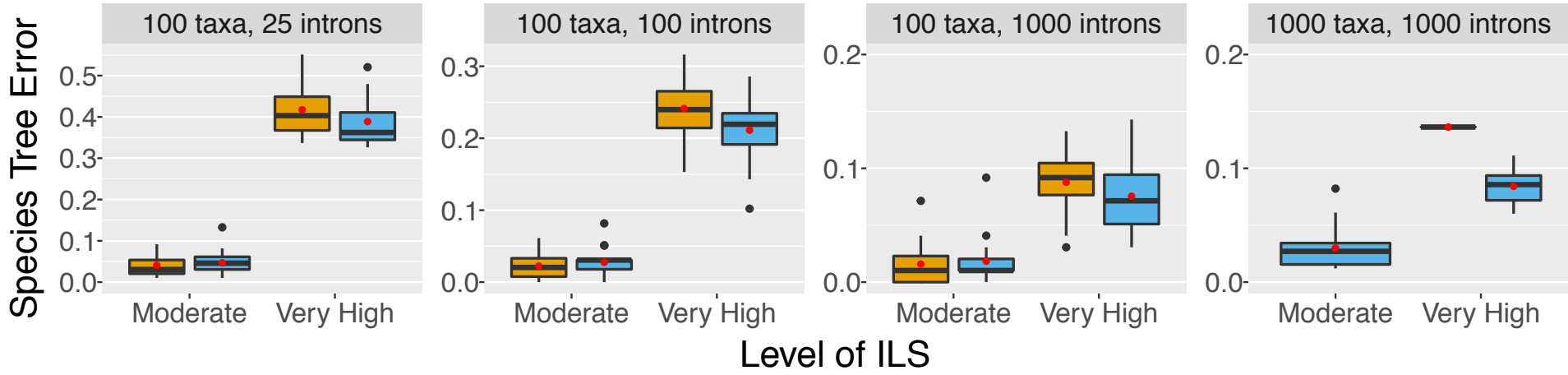
(c) Gene tree M_2 with one duplication and two losses.

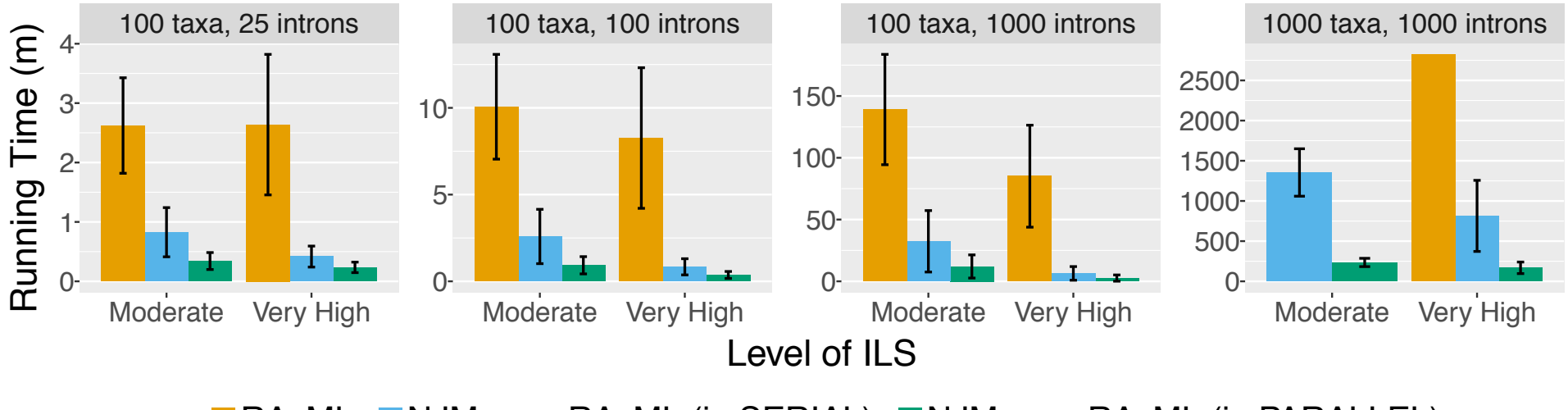


(d) Gene tree with one duplication and three losses.

Many methods, but until Fall 2019, none proven statistically consistent under GDL

NJMerge + RAxML vs. RAxML:
Better accuracy and faster!





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



Papers on bioRxiv (under review)

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

Theorem (Molloy and Warnow, 2019): **FastMulRFS** is statistically consistent under a generic duplication-only or loss-only model, and runs in polynomial time.

Papers on bioRxiv (under review)

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

Theorem (Molloy and Warnow, 2019): **FastMulRFS** is statistically consistent under a generic duplication-only or loss-only model, and runs in polynomial time.

Note: Both methods use dynamic programming to solve NP-hard **discrete optimization problems** within constrained search space in polynomial time.



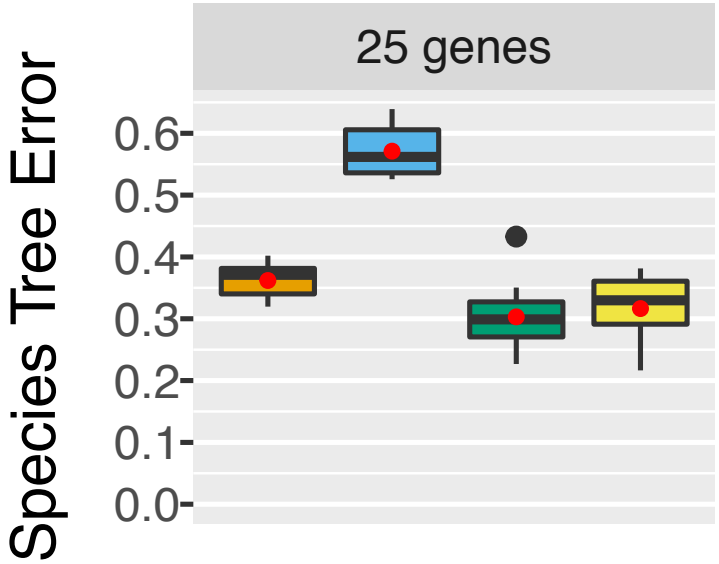
Papers on bioRxiv (under review)

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

Theorem (Molloy and Warnow, 2019): **FastMulRFS** is statistically consistent under a generic duplication-only or loss-only model, and runs in polynomial time.

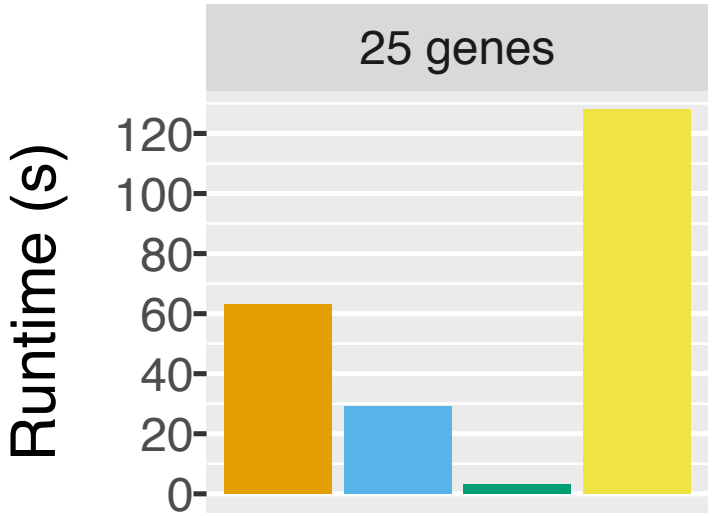
Note: Both methods use dynamic programming to solve NP-hard **discrete optimization problems** within constrained search space in polynomial time.

Papers on bioRxiv (under review)



Results on data (100 species):

- FastMulRFS and MulRF tied for best in terms of accuracy
- FastMulRFS is by far the fastest

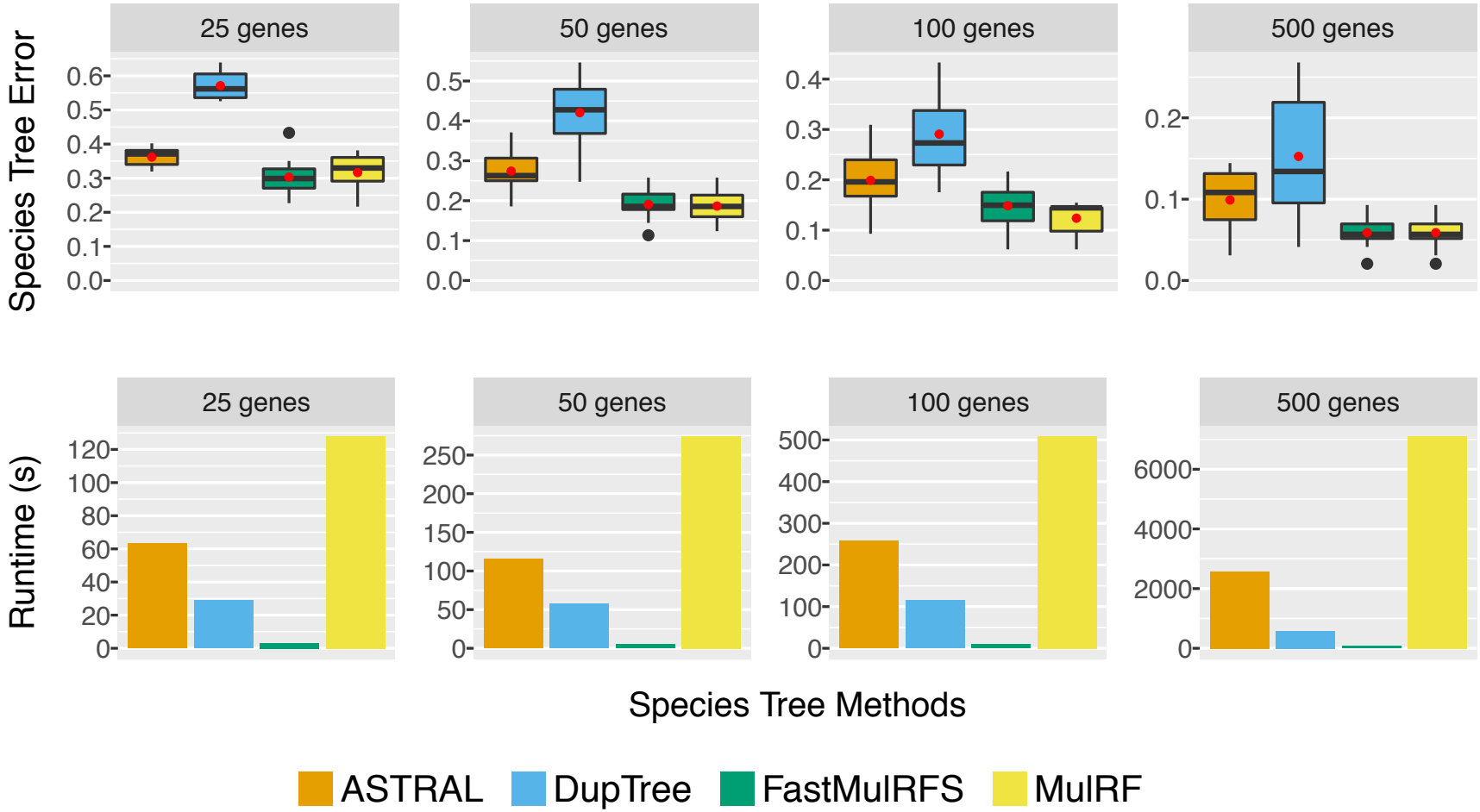


Data: 100 species, moderate GDL, moderately high ILS, high gene tree estimation error

ASTRAL DupTree FastMulRFS MulRF

Molloy and Warnow, bioRxiv 2019
doi: https://doi.org/10.1101/835553

FastMulRFS vs MulRF, ASTRAL-multi, and DupTree



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

This talk

- Fast introduction to phylogenetic estimation, in a statistical framework
- ASTRAL – fast and accurate (and statistically consistent) species tree estimation addressing Incomplete Lineage Sorting (ILS)
- TreeMerge: enabling ASTRAL to run on large datasets
- FastMulRFS: fast and accurate (and statistically consistent) species tree estimation addressing Gene Duplication and Loss
- Discussion

Acknowledgments

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

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

Funding: NSF (CCF 1535977 and also NSF Graduate Fellowship to Erin Molloy)

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

Write to me: warnow@Illinois.edu

Opportunities for PhD students:

- Large impact on biology through innovative algorithm design
- Interesting mathematical problems, including discrete algorithms and machine learning
- Not necessary to understand biology (seriously!)
- Most important skills: enjoying coding, testing, looking at data, and collaborating with other people.
- Many types of research: high performance computing, parallel algorithms, graph algorithms, combinatorial optimization, machine learning, etc.

My students go on to successful careers in academia (UCSD, Rice, etc.) and industry (Apple, Google, Amazon)