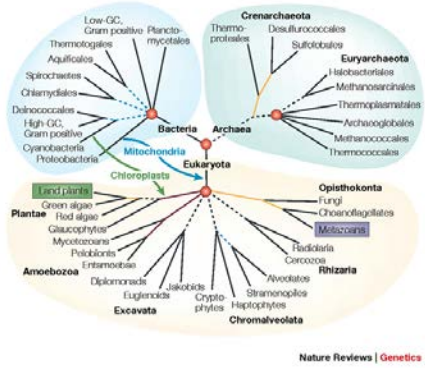


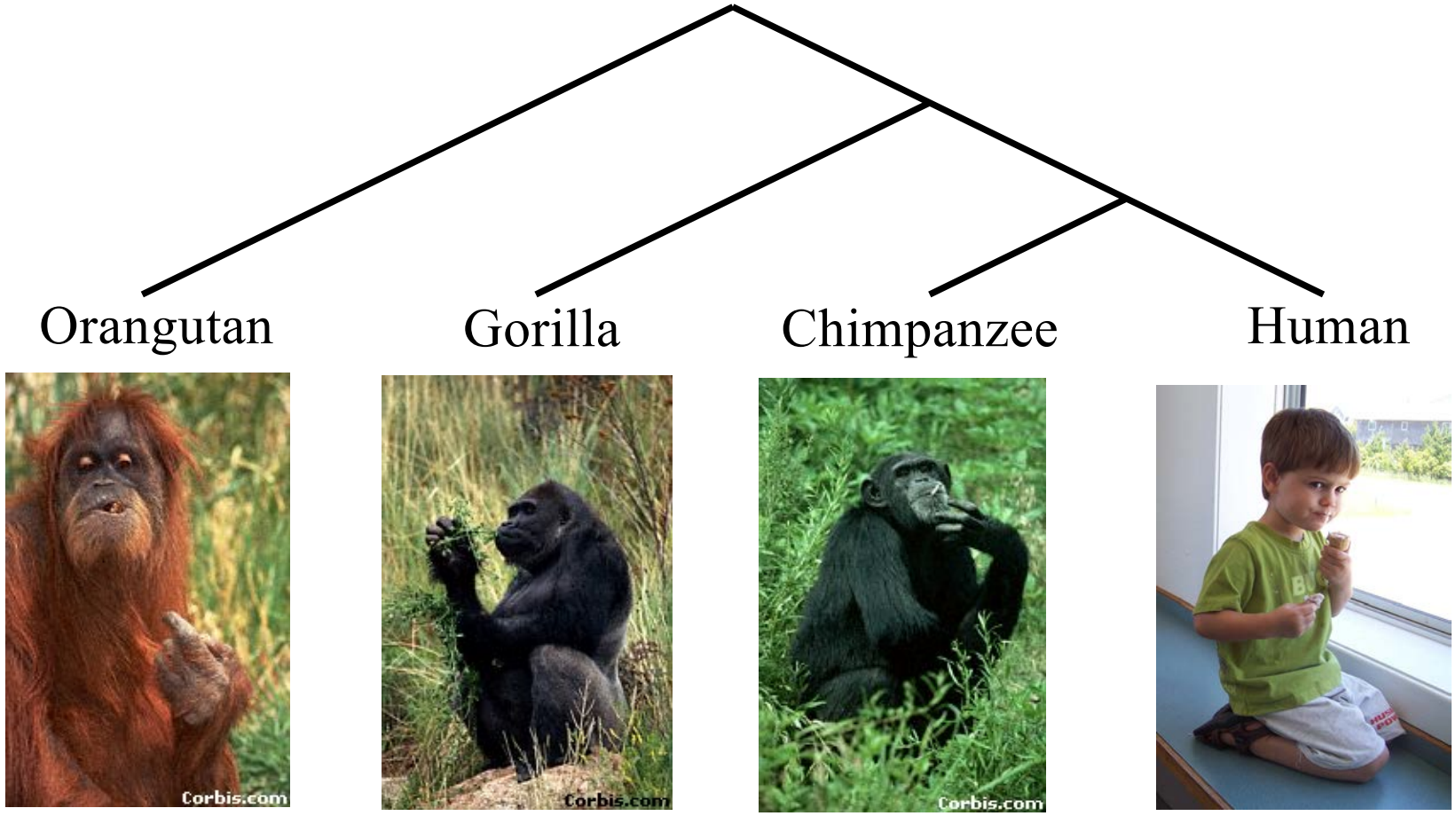
Scaling methods for phylogeny estimation to large datasets using divide-and-conquer

Tandy Warnow
University of Illinois at Urbana-Champaign

Joint work with Erin Molloy

.

Phylogeny
(evolutionary tree)

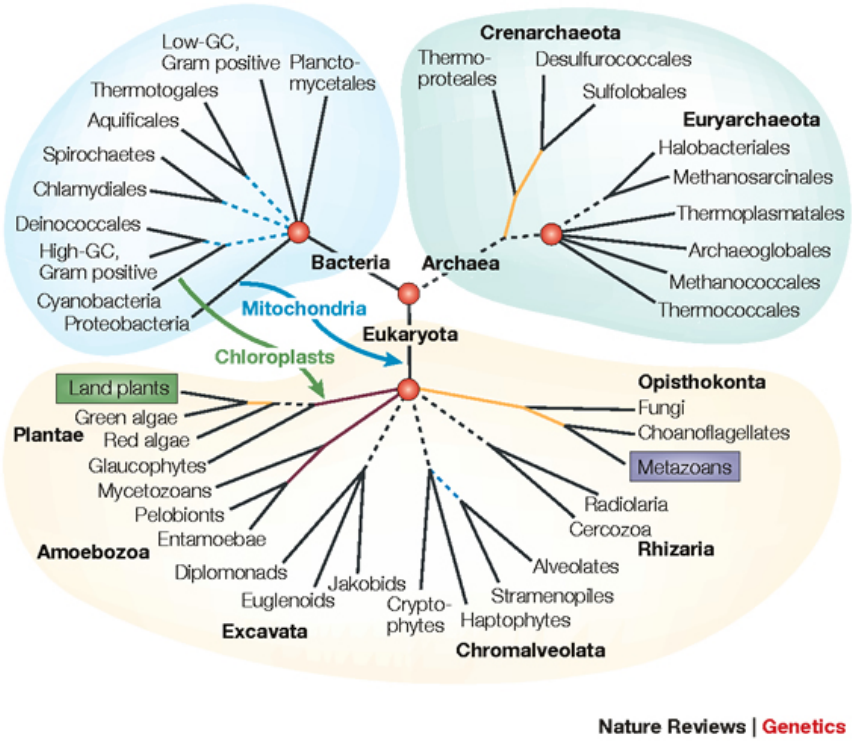


From the Tree of the Life Website,
University of Arizona

- "Nothing in biology makes sense except in the light of evolution"
  - Theodosius Dobzhansky, 1973 essay in the American Biology Teacher, vol. 35, pp 125-129

- *"…… nothing in evolution makes sense except in the light of phylogeny …"*
  - *Society of Systematic Biologists, http://systbio.org/teachevolution.html*

Phylogenomics



Nature Reviews | Genetics

Phylogeny + genomics = genome-scale phylogeny estimation

.

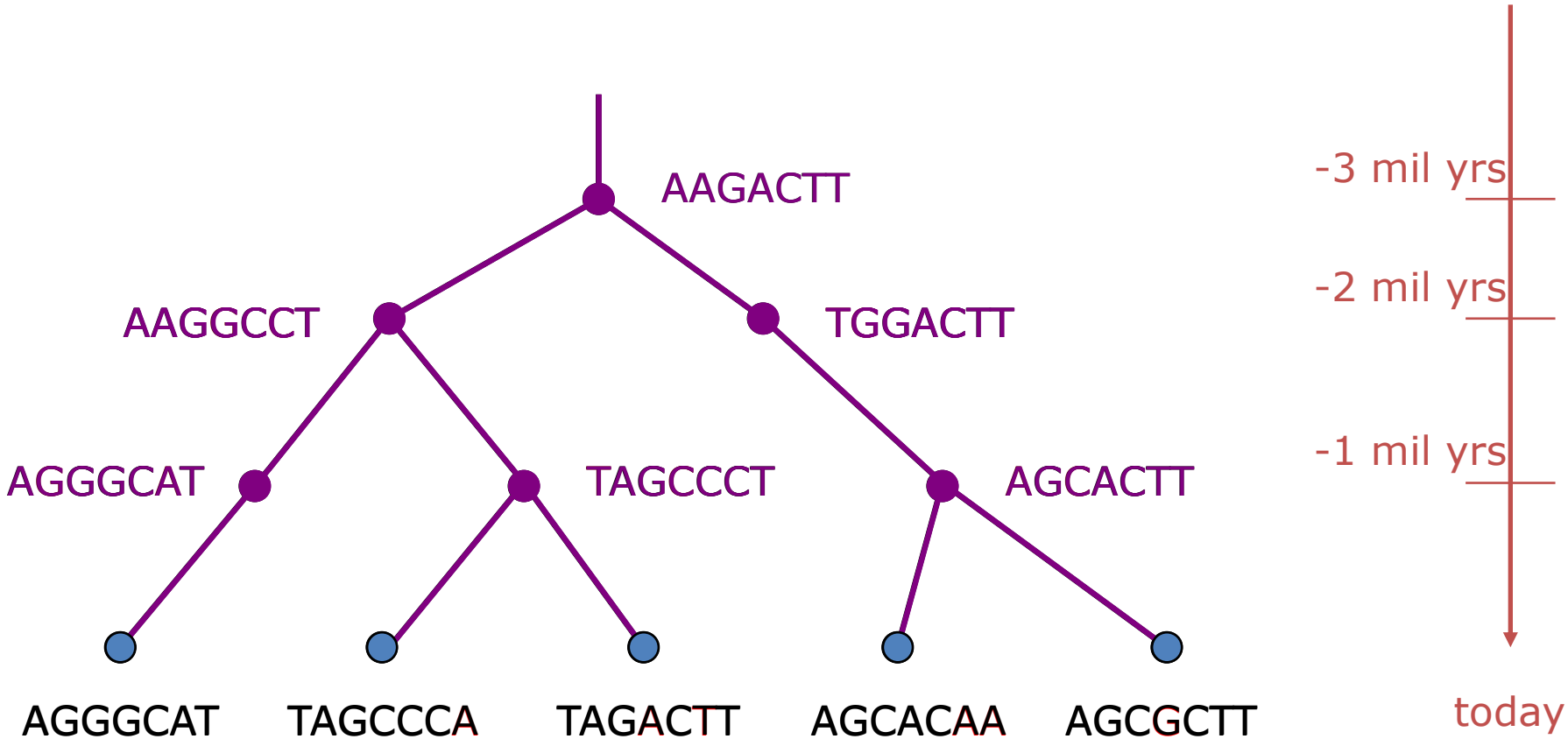
I use Blue Waters to:

- Design and test algorithms for core problems in phylogenomics and its applications

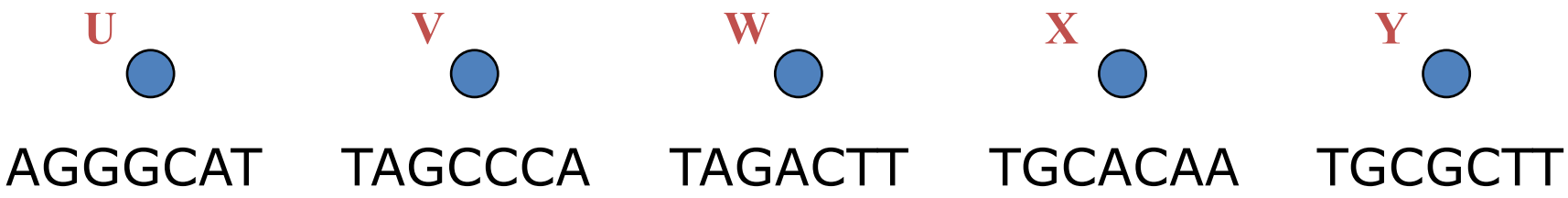
This Talk

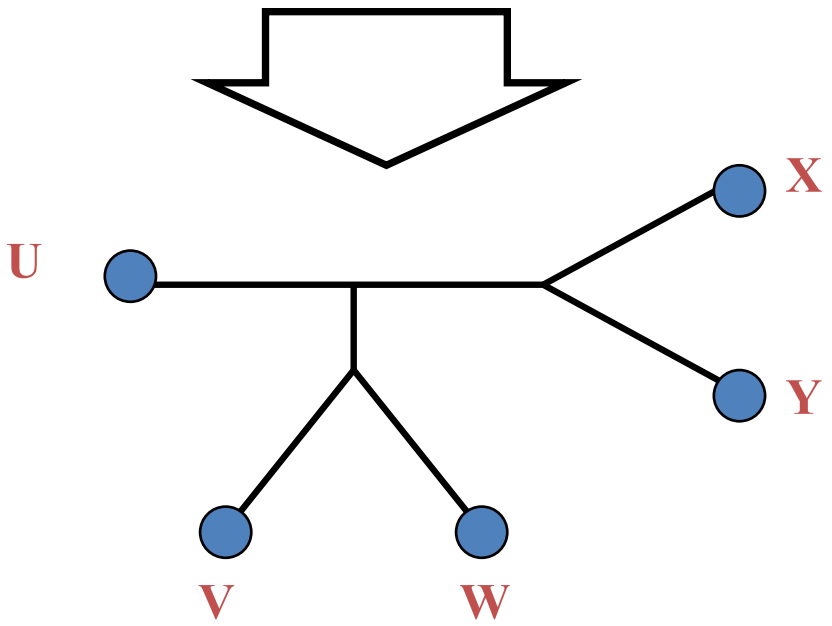
- Genome-scale species tree estimation
  - The pipeline: Statistical estimation and NP-hard optimization problems
  - Incomplete lineage sorting and species tree estimation under the Multi-Species Coalescent model (MSC)
  - Statistically consistent methods (ASTRAL and ASTRID)
  - NJMerge and TreeMerge: scaling species tree methods to large datasets
- Discussion and Future directions

DNA Sequence Evolution (Idealized)



Phylogeny Problem





Markov Models of Sequence Evolution

The different sites are assumed to evolve ***i.i.d***. down the model tree (with rates that are drawn from a gamma distribution).

Markov Models of Sequence Evolution

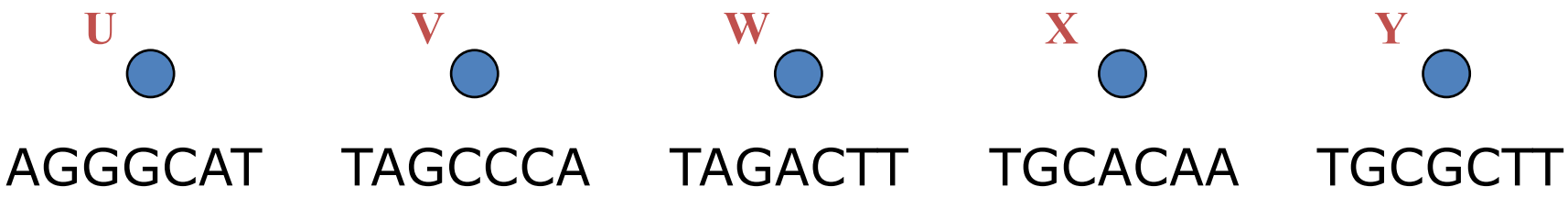
The different sites are assumed to evolve ***i.i.d***. down the model tree (with rates that are drawn from a gamma distribution).

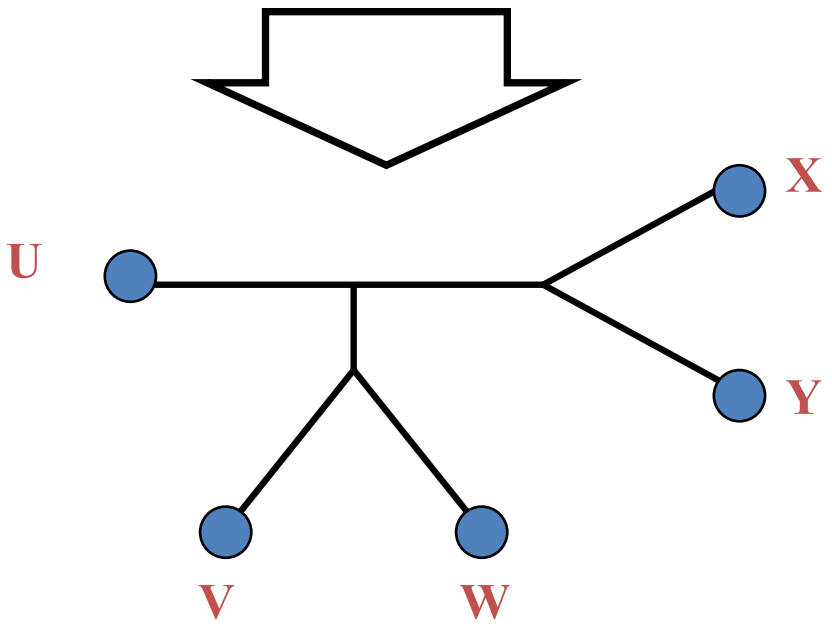
Simplest site evolution model (Jukes-Cantor, 1969):

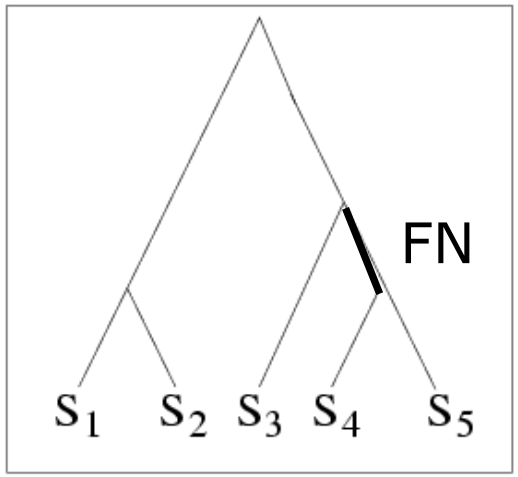
- The model tree T is binary and has substitution probabilities p(e) on each edge e, with $0<p(e)<3/4$.
- The state at the root is randomly drawn from {A,C,T,G} (nucleotides)
- If a site (position) changes on an edge, it changes with equal probability to each of the remaining states.
- The evolutionary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

More complex models (such as the General Markov model) are also considered, often with little change to the theory.

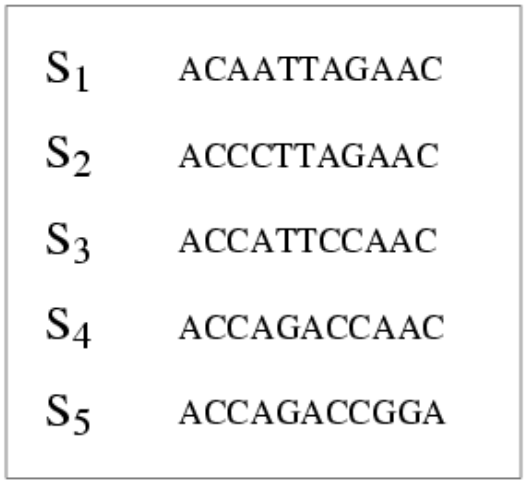
Phylogeny Problem





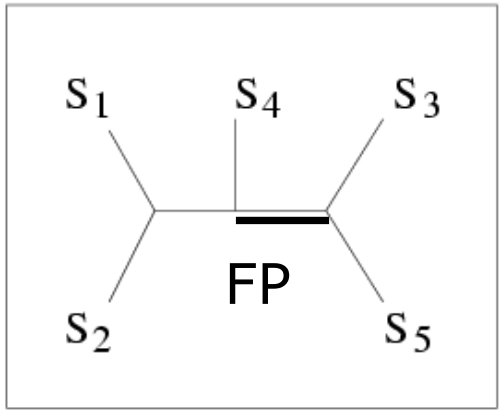


TRUE TREE



DNA SEQUENCES

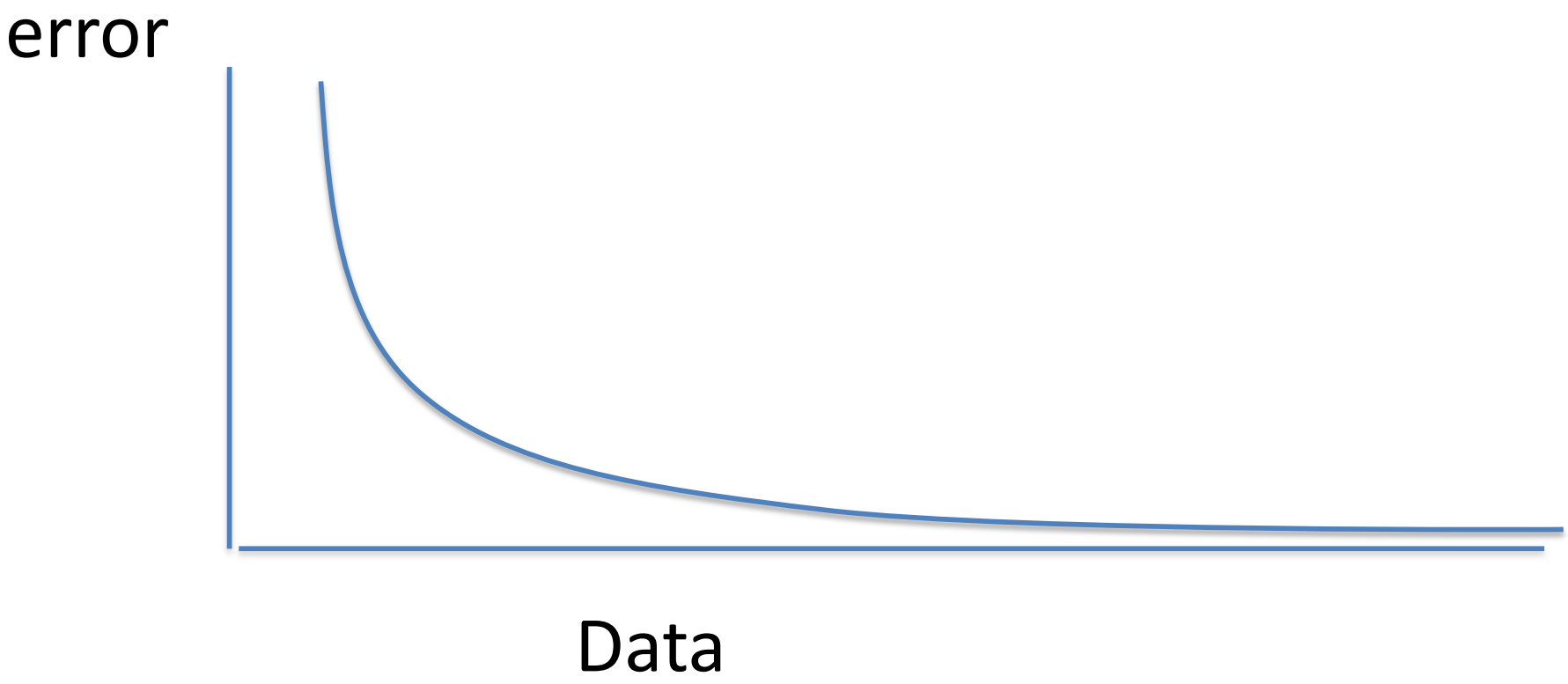
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)



INFERRED TREE

50% error rate

Statistical Consistency/Identifiability



Questions

- Is the model tree identifiable?
- Which estimation methods are statistically consistent under this model?
- How much data does the method need to estimate the model tree correctly (with high probability)?
- What are the computational issues?

Answers?

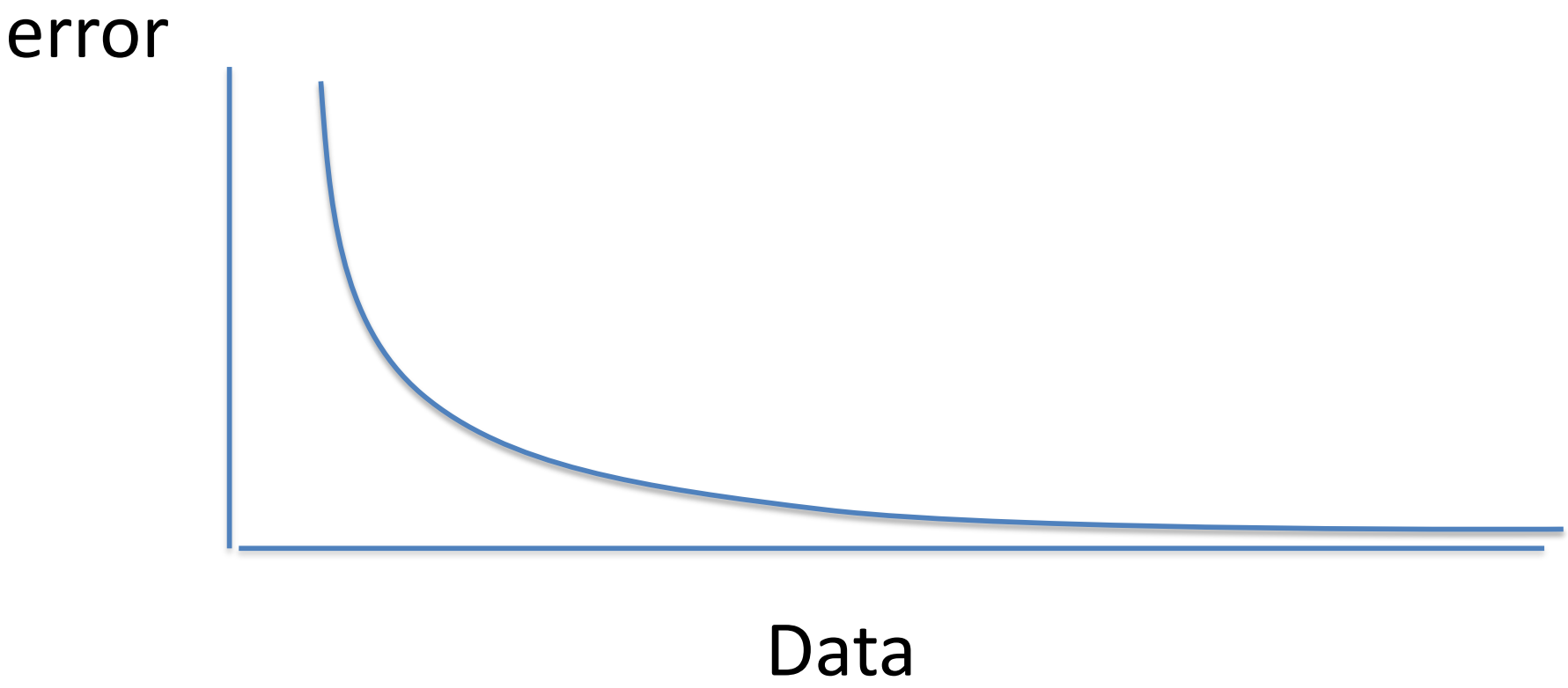
- We know a lot about which site evolution models are identifiable, and which methods are statistically consistent.
- We know a little bit about the sequence length requirements for standard methods.
- The best methods (typically maximum likelihood or Bayesian estimation) are very computationally intensive.

Computational issues

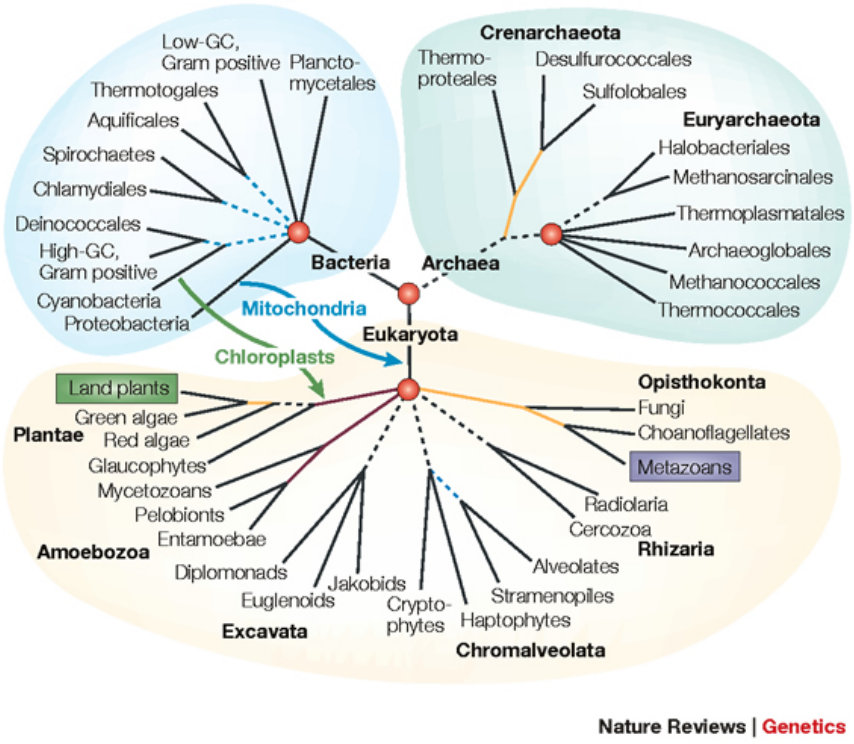
- Maximum likelihood: NP-hard, and tree-space grows exponentially with the number of leaves
- Bayesian estimation: need to run to convergence (may fail)
- Parallelism helps but is not enough

Take home message: large datasets are beyond the capability of current methods (perhaps even with Blue Waters)

Genome-scale data?



Phylogenomics

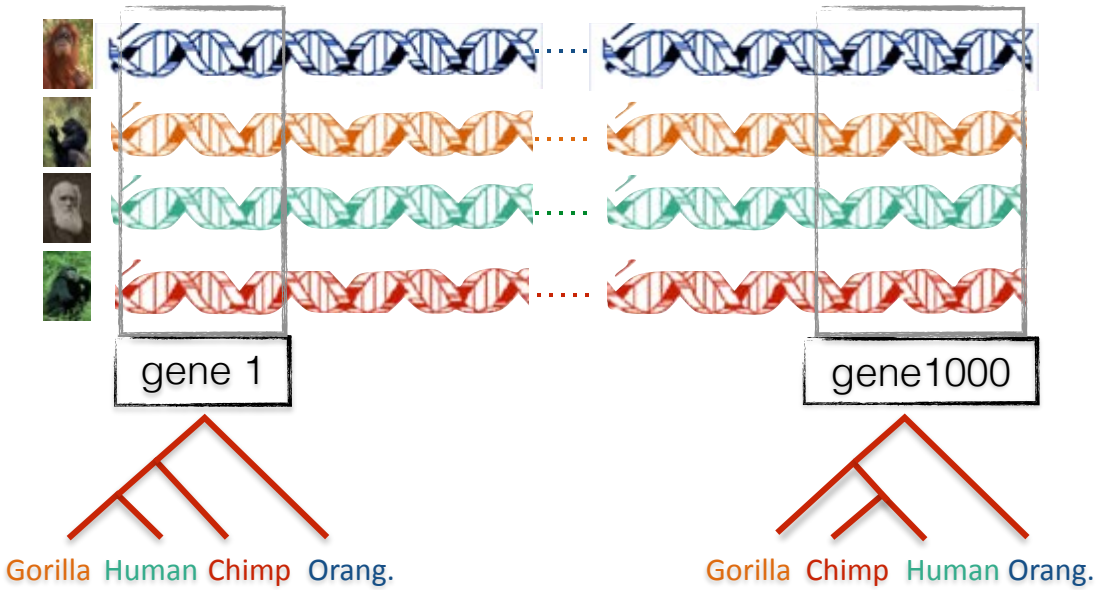


Nature Reviews | Genetics

Phylogeny + genomics = genome-scale phylogeny estimation

.

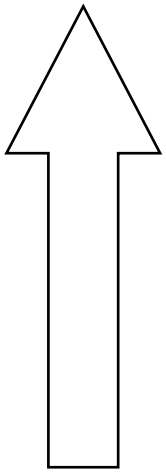
Gene tree discordance



Incomplete Lineage Sorting (ILS) is a dominant cause of gene tree heterogeneity

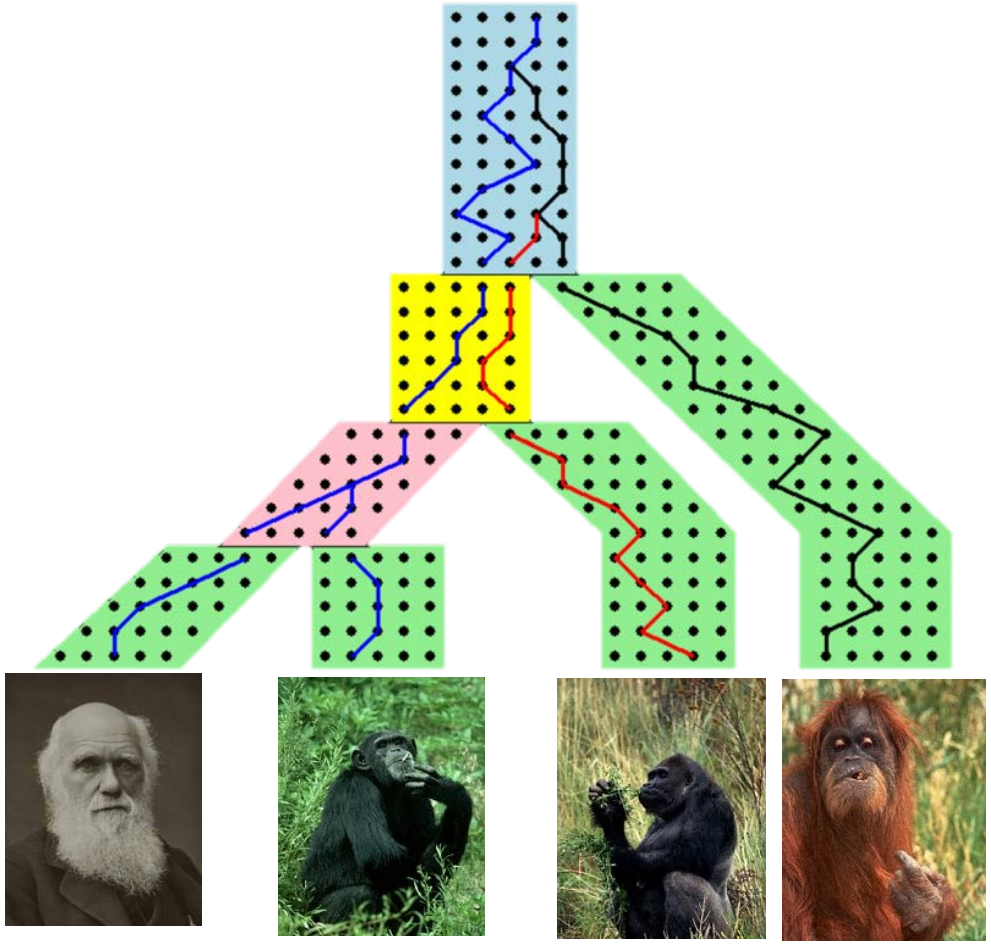
Gene trees inside the species tree (Coalescent Process)

Past



Present

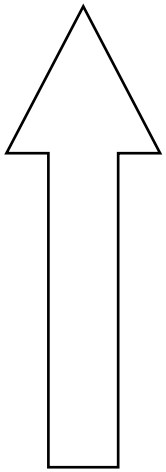
Courtesy James Degnan



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

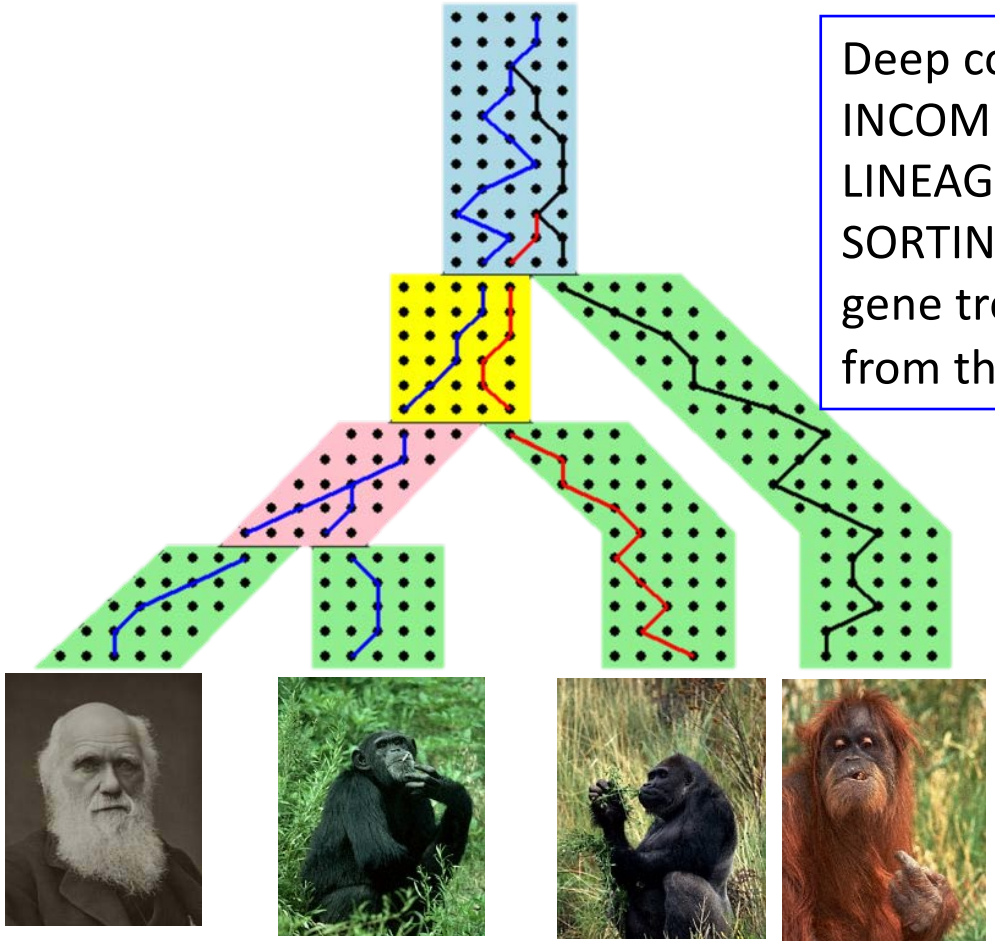
Gene trees inside the species tree (Coalescent Process)

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.

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


S. Mirarab,
UT-Austin


N. Nguyen
UT-Austin

- 103 plant transcriptomes, 400-800 single copy “genes”
- Next phase will be much bigger
- Wickett, Mirarab et al., *PNAS* 2014

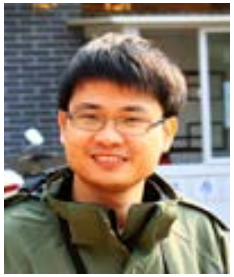
Major Challenge:
- Massive gene tree heterogeneity consistent with ILS

Avian Phylogenomics Project

Erich Jarvis, HHMI

MTP Gilbert, Copenhagen

Guojie Zhang, BGI



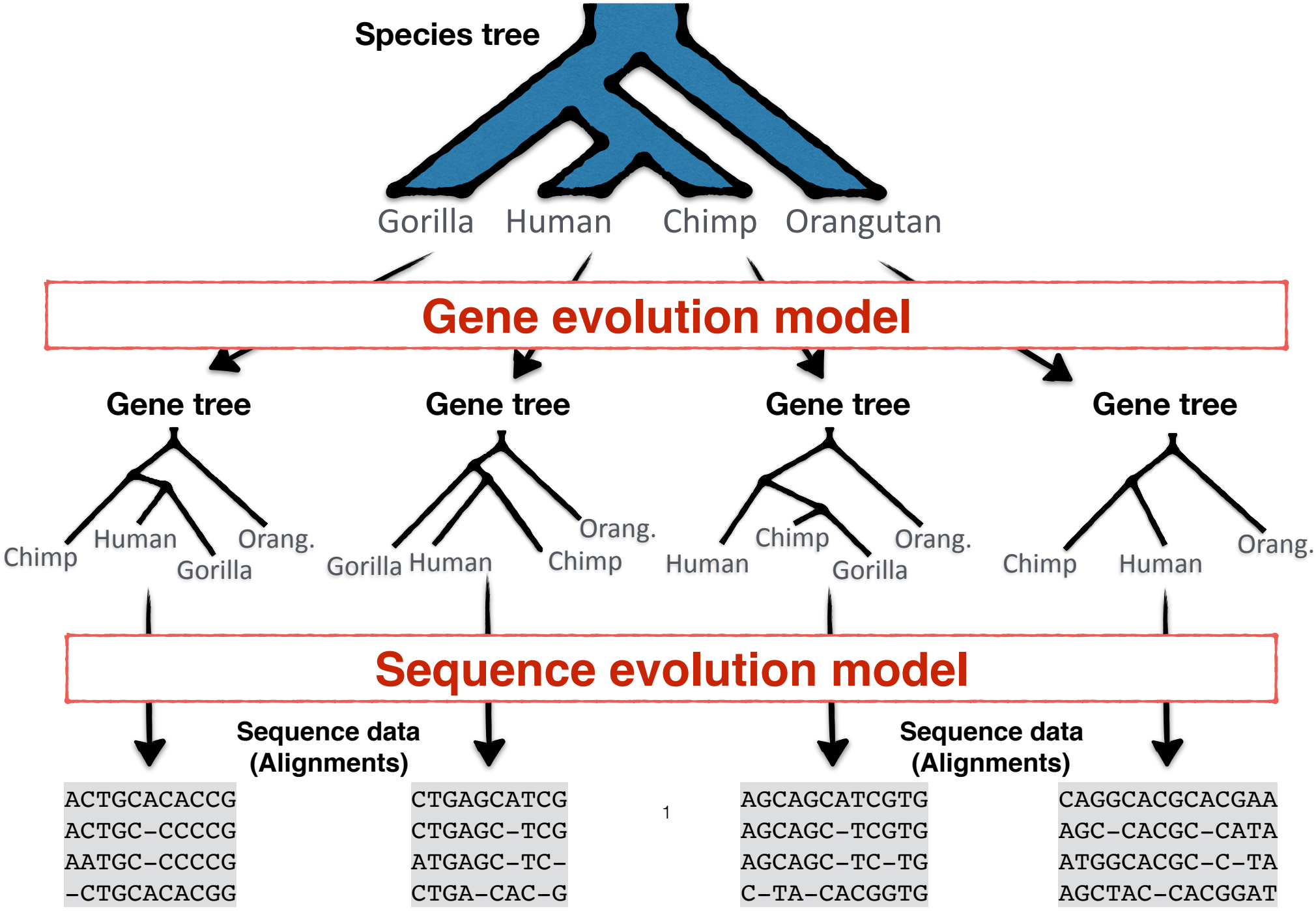
Siavash Mirarab, Texas

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:

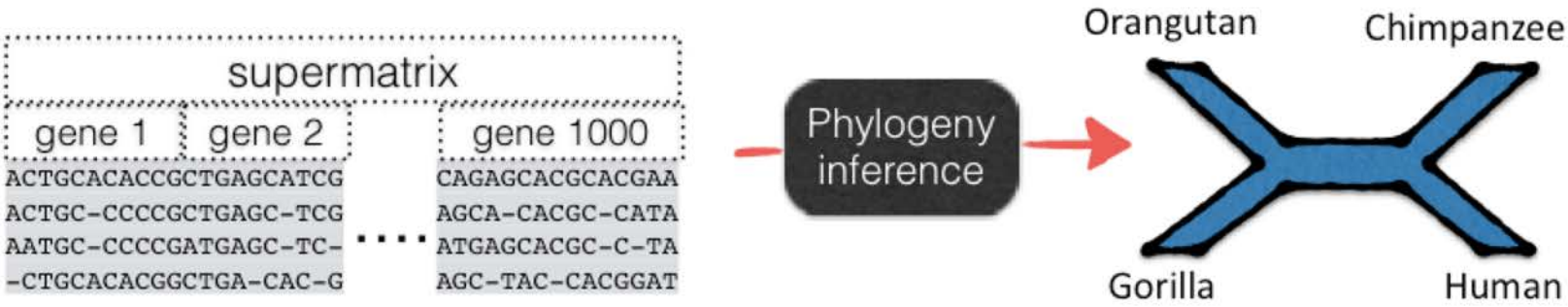
- Massive gene tree heterogeneity consistent with ILS.


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

Big picture challenge

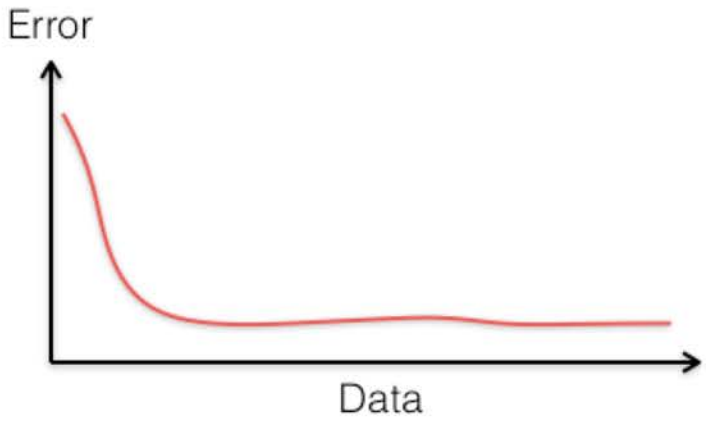
- Multi-locus data, generated by a hierarchical model
  - Species tree generates gene trees
  - Gene trees generate sequences
- How can we estimate the species tree from the sequence data?

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]



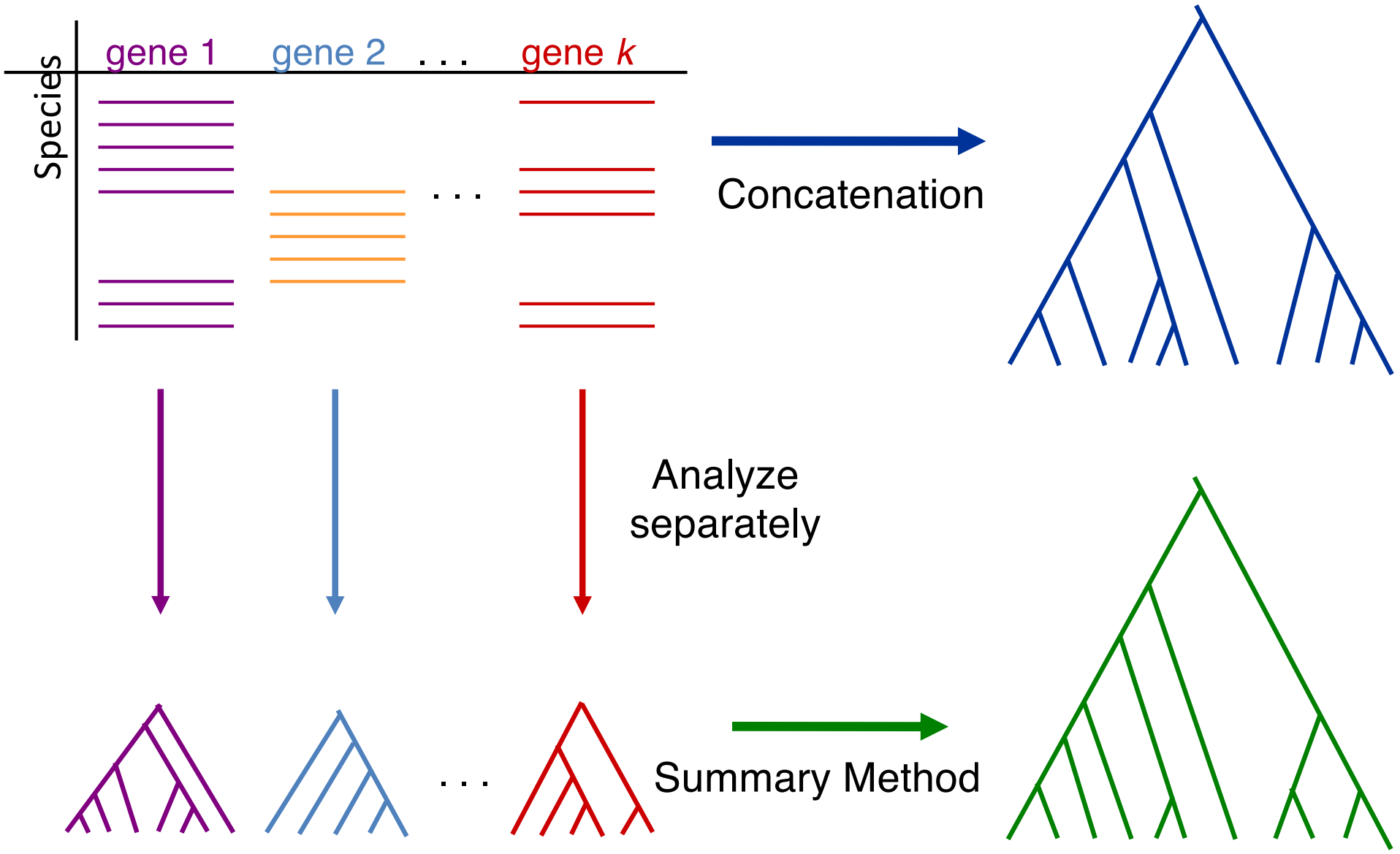
Statistically consistent methods

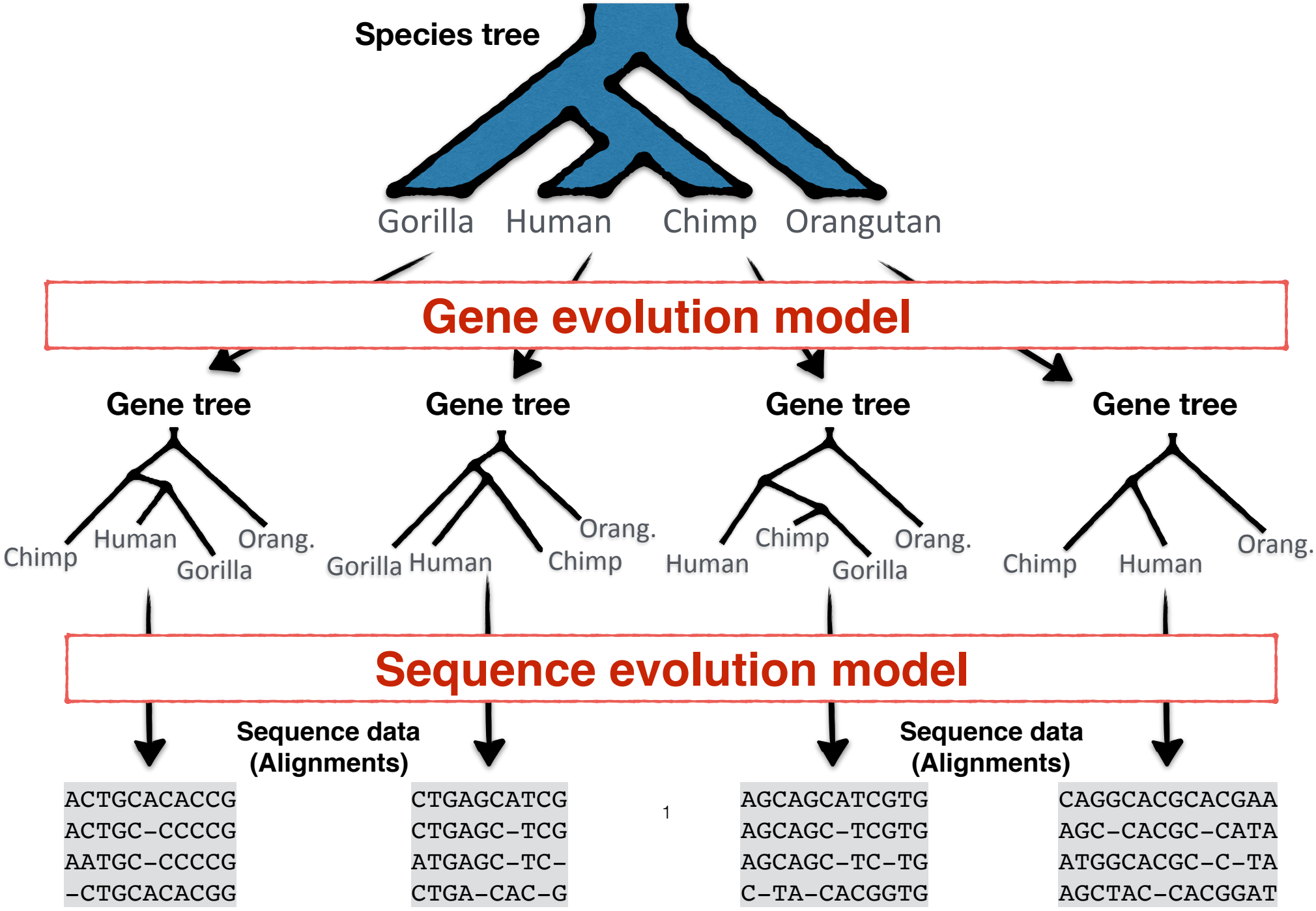
- **Coalescent-based summary methods:** Estimate gene trees, and then combine together (ASTRAL, ASTRID, MP-EST, NJst, and others)

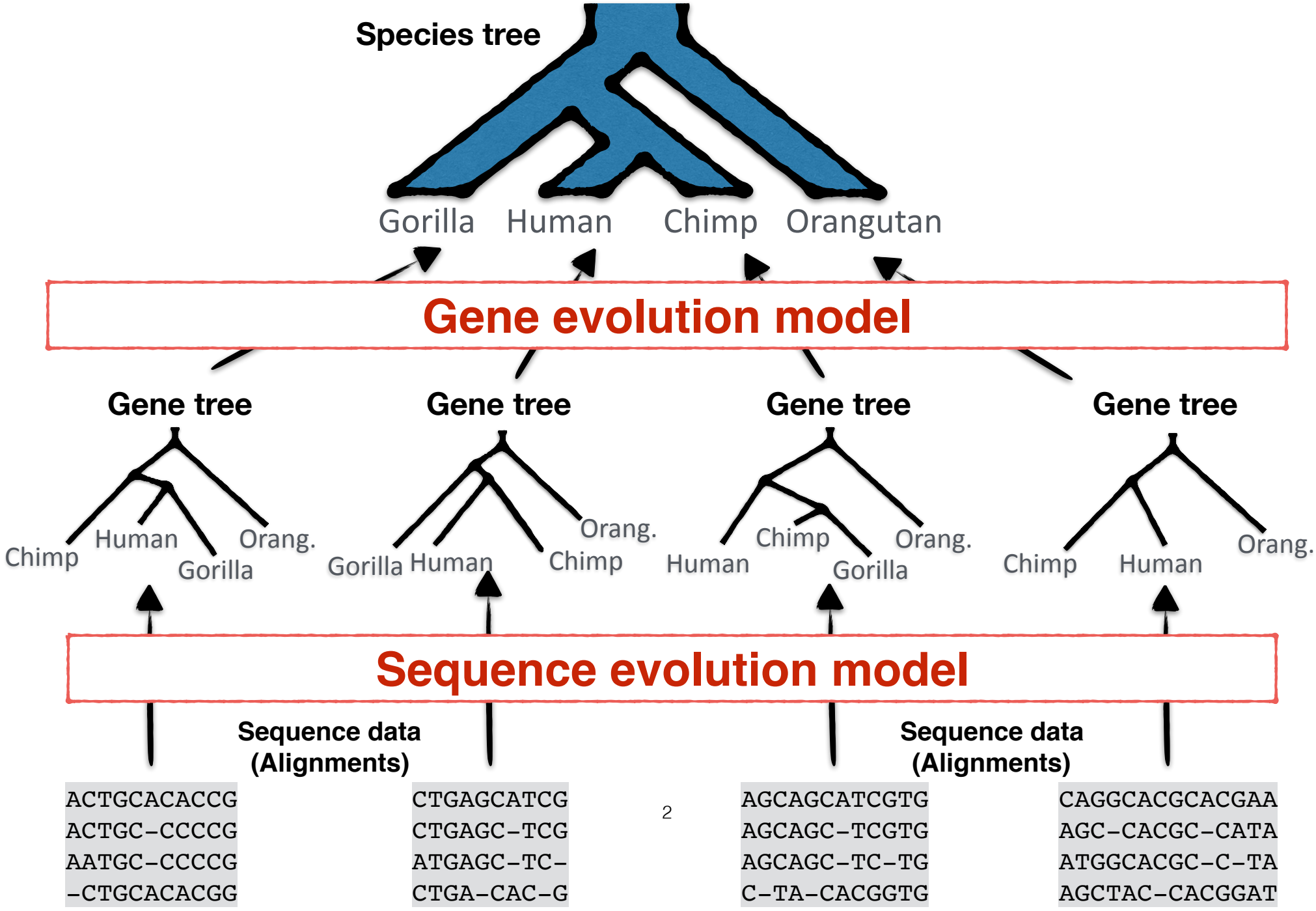
- **Co-estimation methods**: Co-estimate gene trees and species trees (TOO EXPENSIVE)

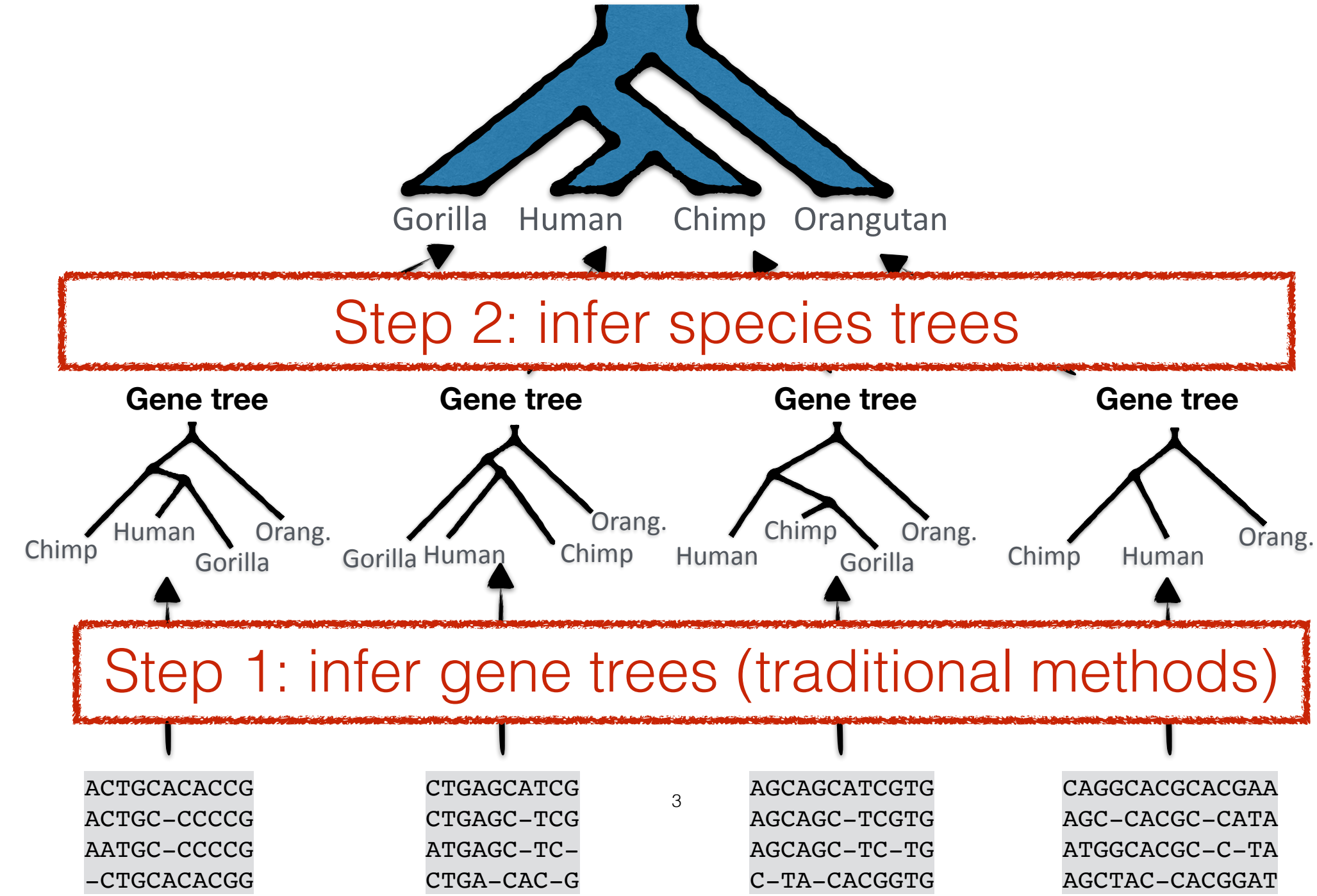
- **Site-based methods:** estimate the species tree from the concatenated alignment, and do not estimate gene trees (NOT WELL STUDIED)

Main competing approaches




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


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

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

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

- Statistically consistent under the MSC, and runs in polynomial time
- Solves constrained version of the NP-hard Maximum Quartet Support problem using dynamic programming
  - Input: Gene trees and set X of allowed bipartitions
  - Output: Species tree T that maximizes the quartet support criterion, subject to drawing its bipartitions from the set X

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

Syst. Biol. 0(0):1–14, 2015
© The Author(s) 2015. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
For Permissions, please email: journals.permissions@oup.com
DOI:10.1093/sysbio/syv029

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]

Contents lists available at ScienceDirect

ELSEVIER

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

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[b], Michael A. Gore[c], Jonathan F. Wendel[a]

Journal of Biogeography (J. Biogeogr.) (2015)

ORIGINAL ARTICLE

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

Peter A. Hosner[1]*, Edward L. Braun[1,2,3] and Rebecca T. Kimball[1,2,3]

LETTER

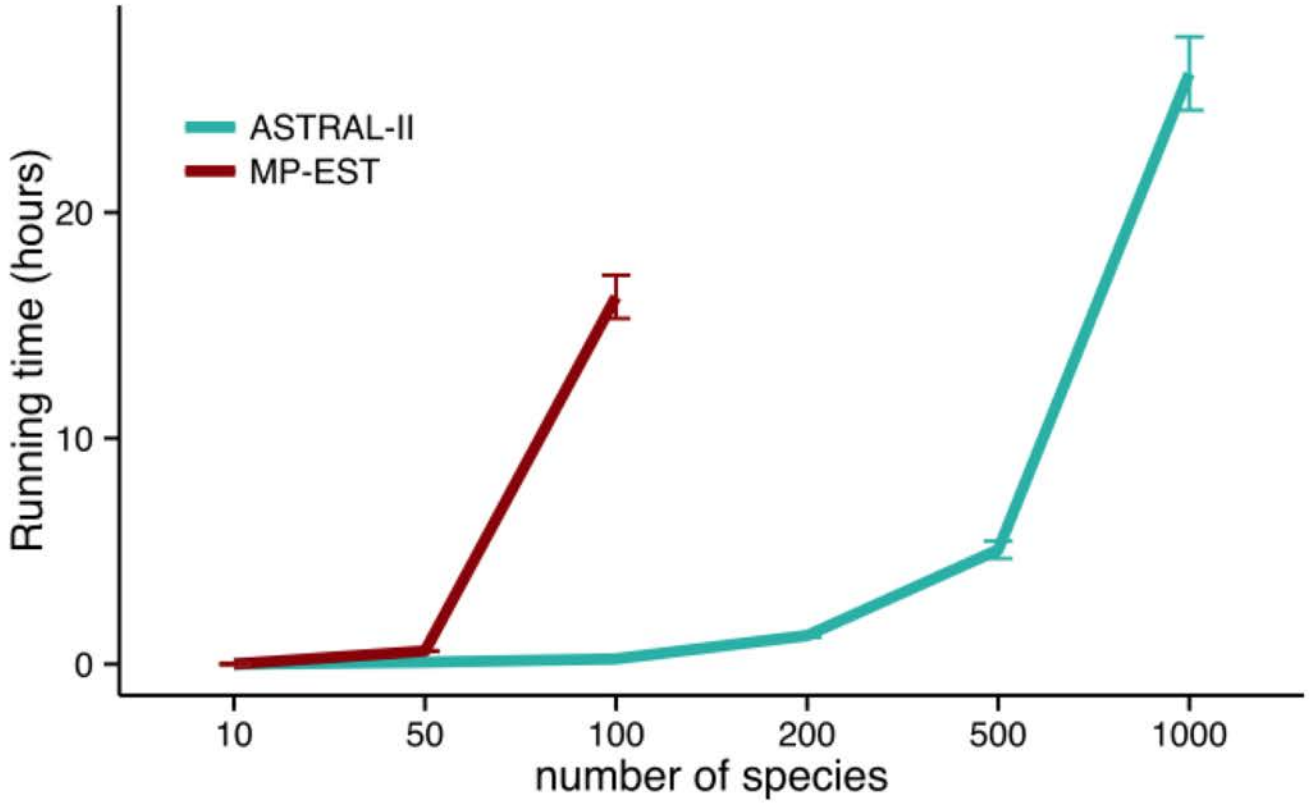
doi:10.1038/nature15697

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

Richard O. Prum[1,2]*, Jacob S. Berv[2]*, Alex Dornburg[1,2,3], Daniel J. Field[2,3], Jeffrey P. Townsend[1,3], Emily Moriarty Lemmon[4] & Alan R. Lemmon[5]

X

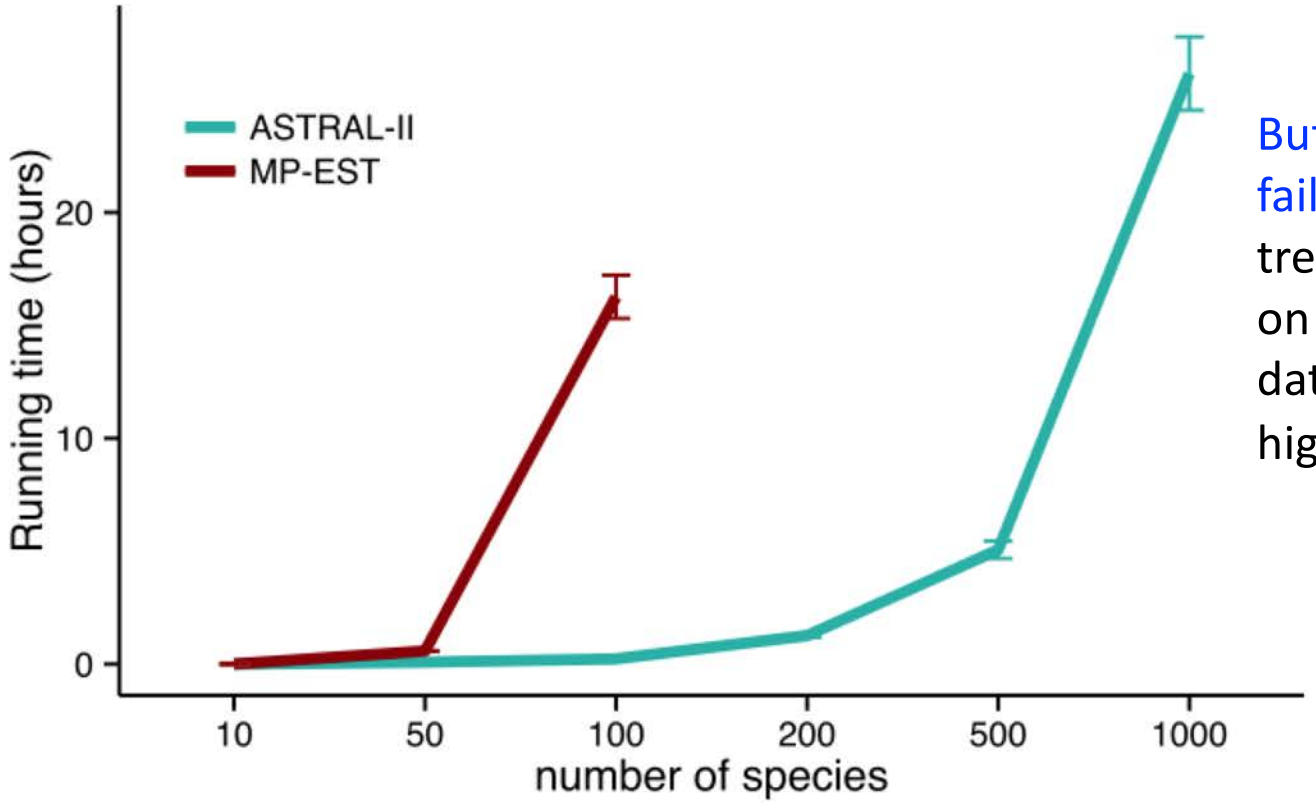
Running time as function of # species



1000 genes, "medium" levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]

X

Running time as function of # species



But ASTRAL can fail to return a tree within 24 hrs on some very large datasets with high ILS

1000 genes, "medium" levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]

X

Scalability to large datasets

- ASTRAL can fail on some datasets with many species and genes (constraint space too big)

- Concatenation using Maximum Likelihood (inconsistent, because it assumes all sites evolve down the same model tree): attempts to solve NP-hard optimization problem, and no current method scales to large numbers of species and genes

NJMerge

- Molloy and Warnow, RECOMB-CG 2018
- Github site: https://github.com/ekmolloy/njmerge

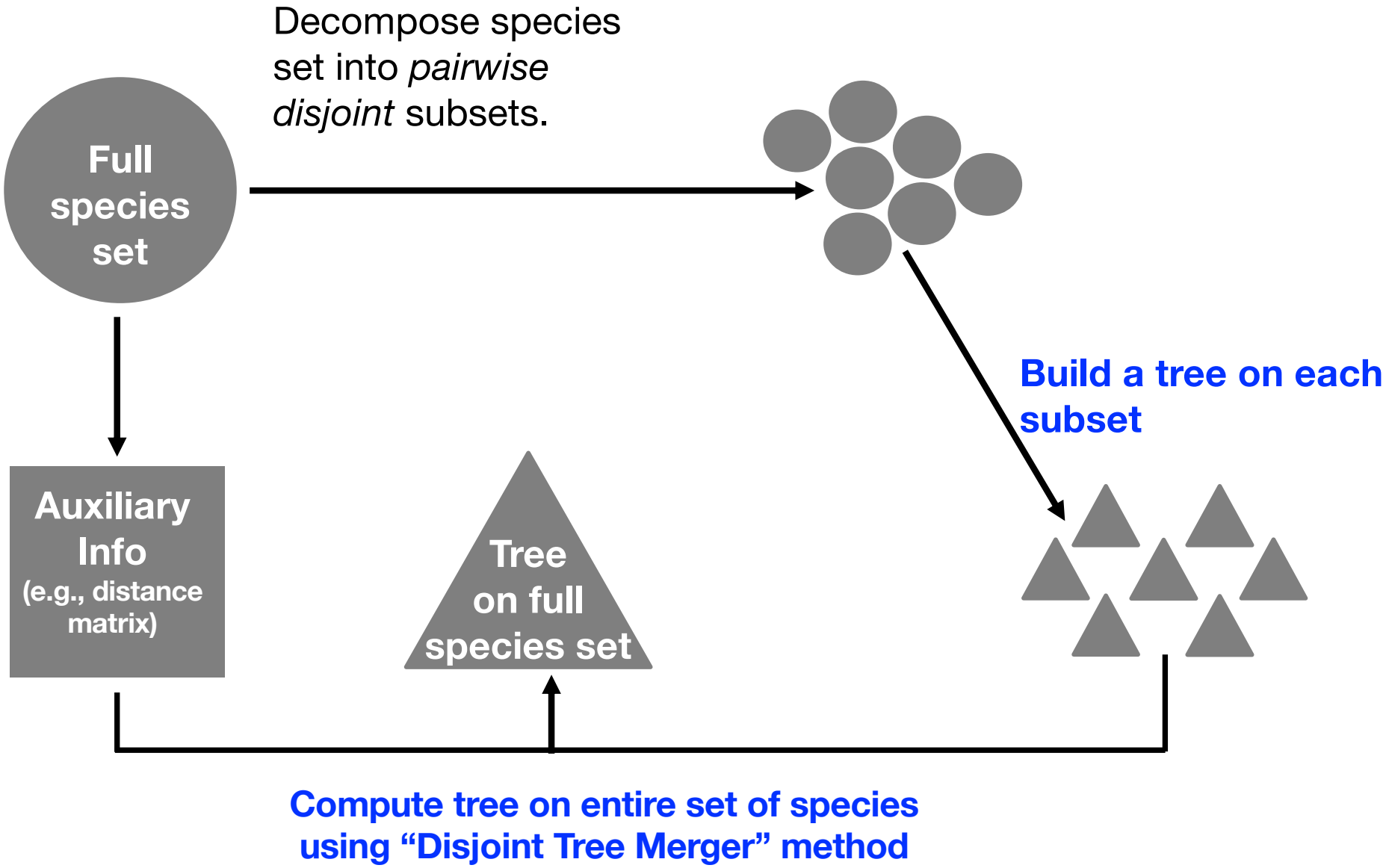
Algorithmic strategy:

- 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), and then merges subset trees using a distance-based method.

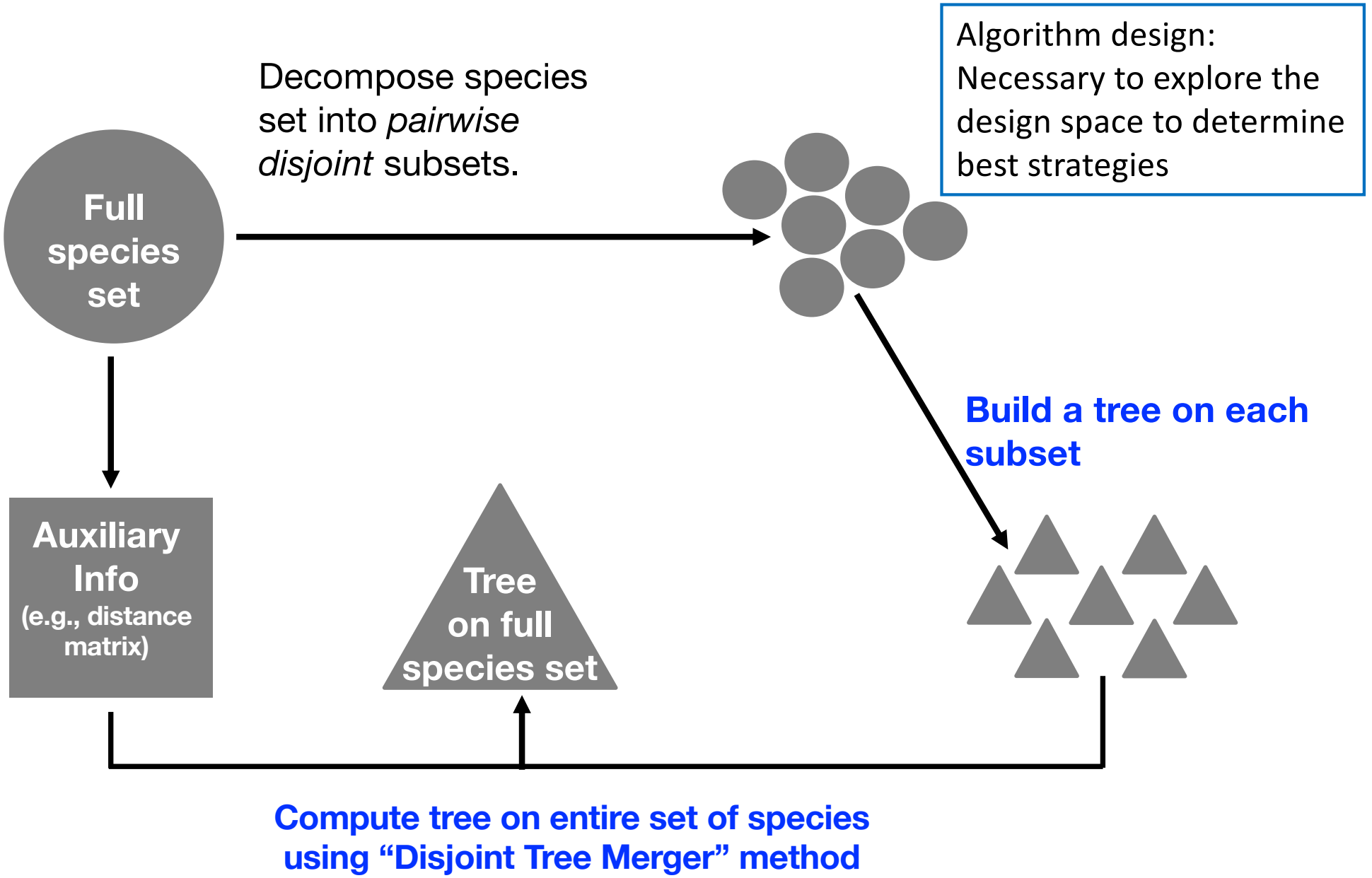
TreeMerge

- Molloy and Warnow, to appear, ISMB 2019
- Like NJMerge, it is statistically consistent under the MSC when used with ASTRAL or other statistically consistent methods
- Improves on NJMerge:
  - guaranteed to never fail
  - Asymptotically faster -- $O(n^2)$ in divide-and-conquer pipeline
- On github

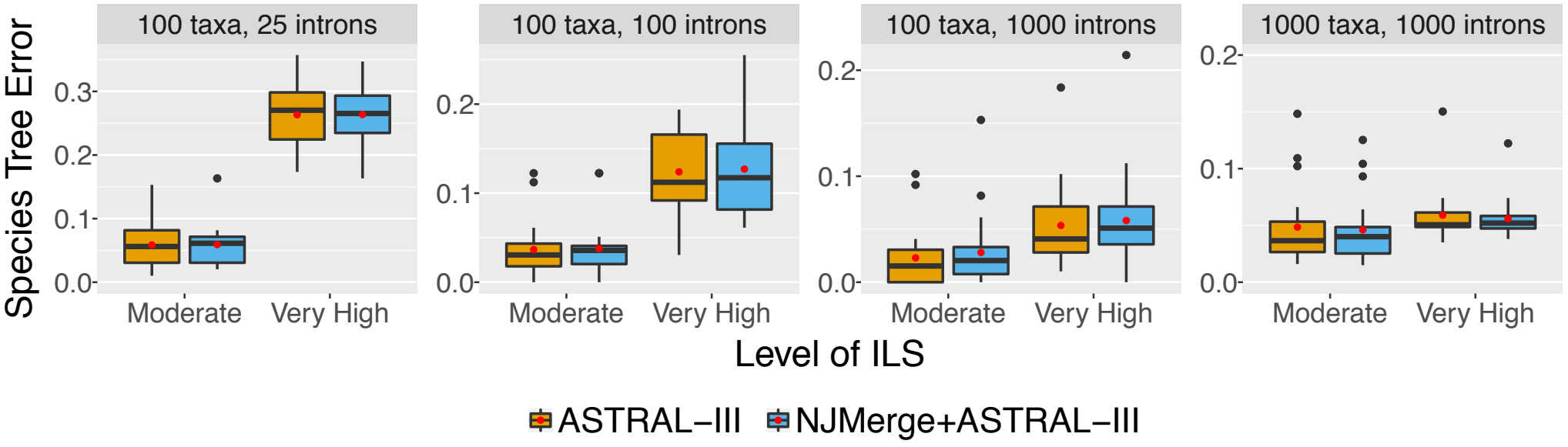
Divide-and-Conquer Pipeline

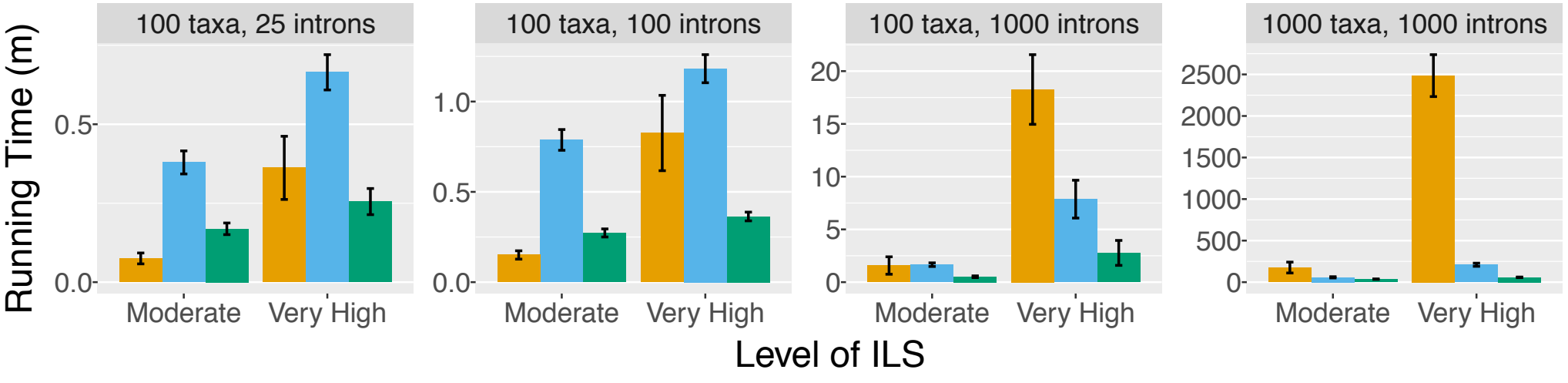


Divide-and-Conquer Pipeline

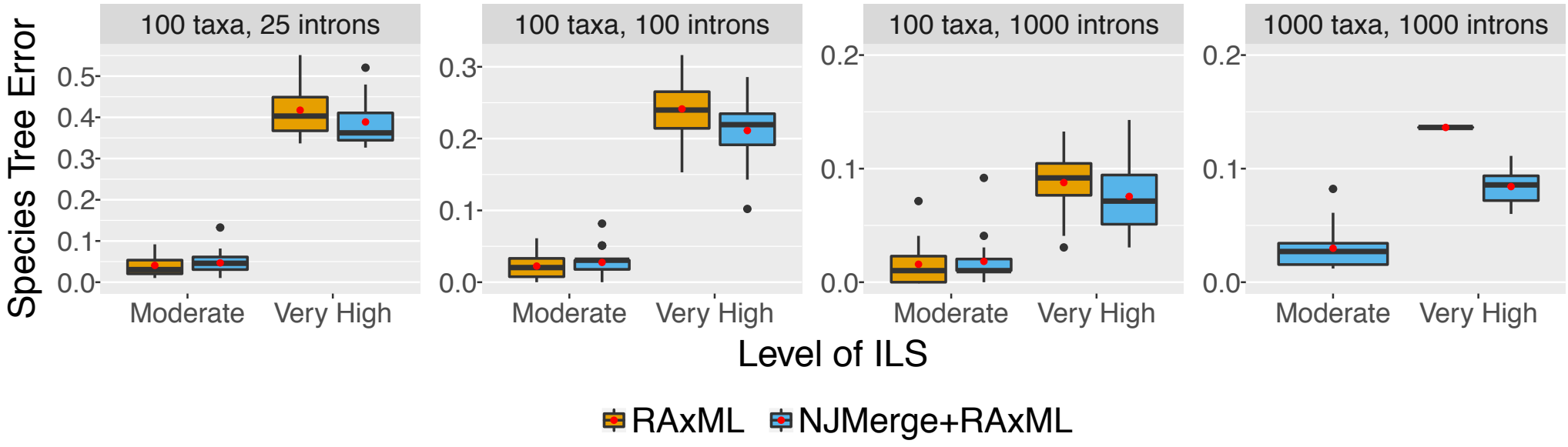


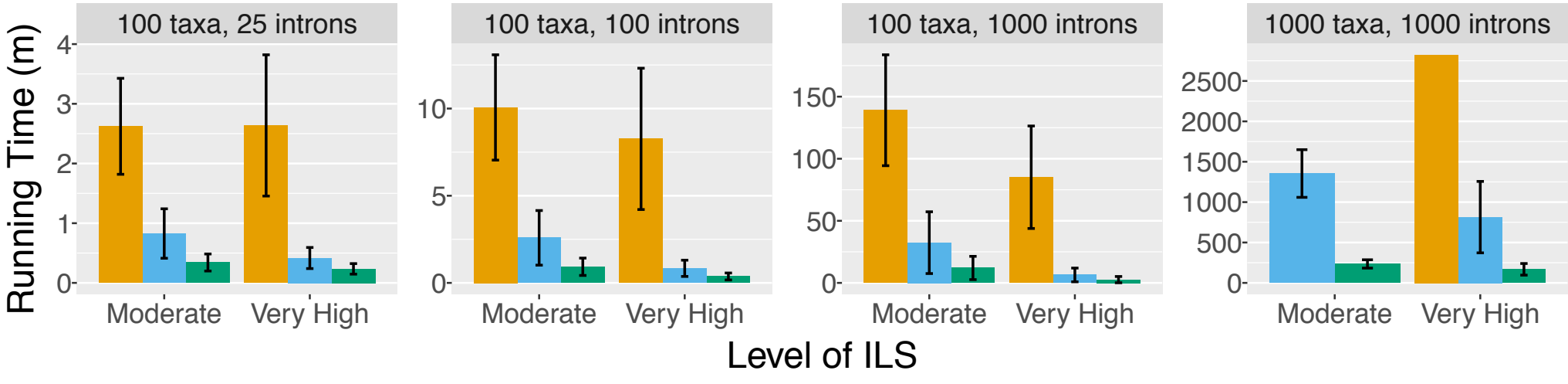
NJMerge + ASTRAL vs. ASTRAL:
Comparable accuracy and can analyze larger datasets





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





Summary

- Using NJMerge or TreeMerge with ASTRAL: generally as accurate and faster on large datasets than ASTRAL, and also statistically consistent under the Multi-Species Coalescent model
- Using NJMerge or TreeMerge with concatenation using maximum likelihood (CA-ML): more accurate and much faster, greater scalability than CA-ML

Summary

- The best tree estimation methods are computationally intensive, and tree-space grows exponentially

Summary

- The best tree estimation methods are computationally intensive, and tree-space grows exponentially
- Statistical consistency is important but not sufficient

Summary

- The best tree estimation methods are computationally intensive, and tree-space grows exponentially
- Statistical consistency is important but not sufficient
- Parallel implementations of expensive methods are helpful but not enough

Summary

- The best tree estimation methods are computationally intensive, and tree-space grows exponentially
- Statistical consistency is important but not sufficient
- Parallel implementations of expensive methods are helpful but not enough
- Divide-and-conquer improves scalability, maintains statistical consistency, and can maintain accuracy (or only lose a small amount)

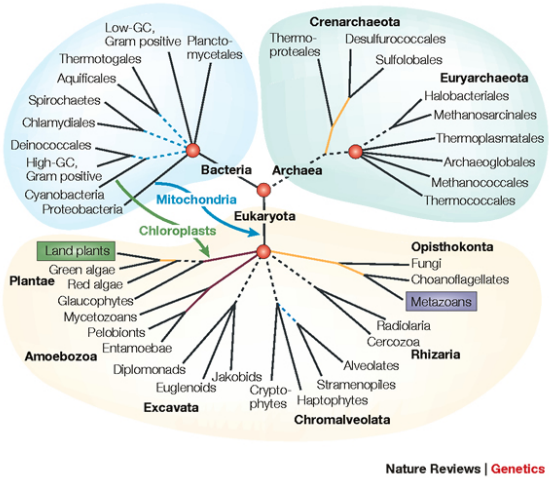
Summary

- The best tree estimation methods are computationally intensive, and tree-space grows exponentially
- Statistical consistency is important but not sufficient
- Parallel implementations of expensive methods are helpful but not enough
- Divide-and-conquer improves scalability, maintains statistical consistency, and can maintain accuracy (or only lose a small amount)
- Divide-and-conquer is highly parallelizable

What Blue Waters enabled

- Algorithm design is iterative, and requires evaluation using multiple variants on many datasets, each one taking potentially a very long time
- None of this would be feasible without Blue Waters
- Future phylogenomics projects will be able to use the methods developed using Blue Waters allocations.

Phylogenetic Inference



Nature Reviews | Genetics

Genomic data are:

- Heterogeneous
- Large
- Noisy
- Error-ridden
- Streaming

Approaches:

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

Acknowledgments

Mirarab and Warnow, Bioinformatics 2015 (ASTRAL-II)

Molloy and Warnow, Systematic Biology 2017

Molloy and Warnow, RECOMB-CG 2018 (and Algorithms for Molecular Biology)

Molloy and Warnow, ISMB 2019 (and Bioinformatics, to appear)

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

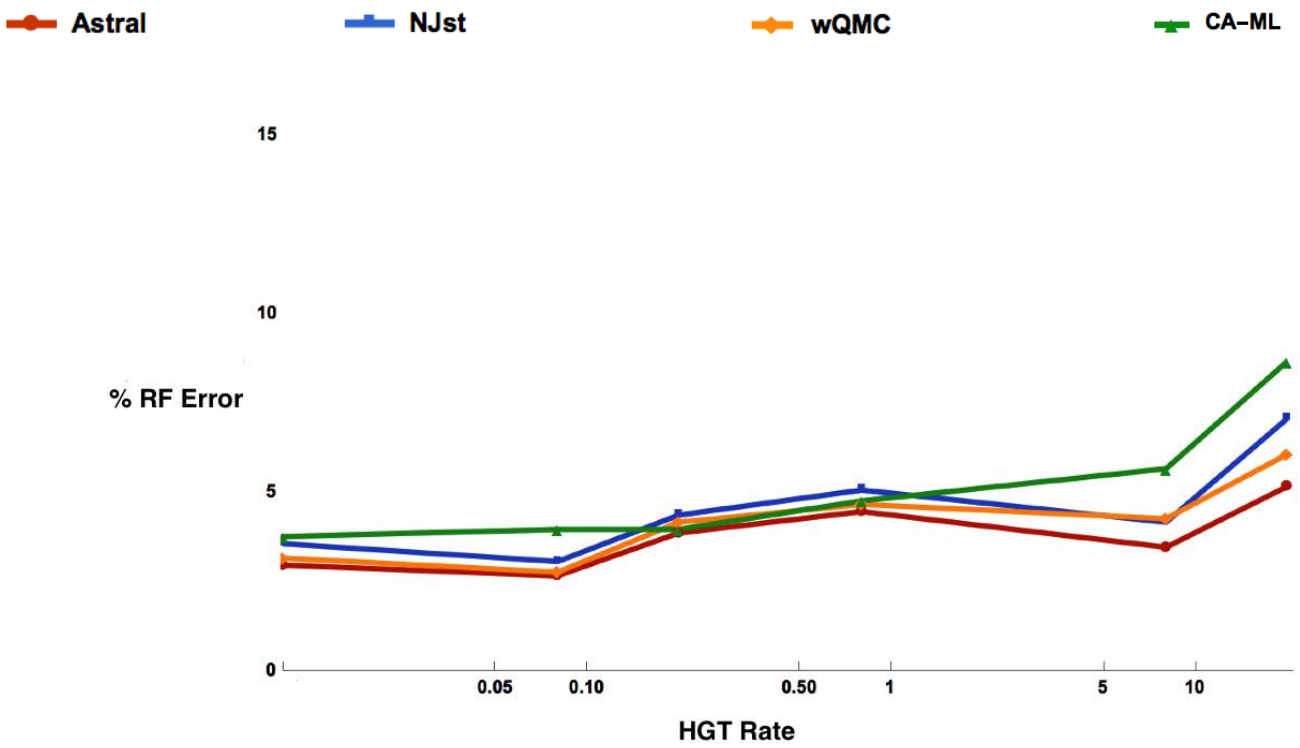
Presentations available at http://tandy.cs.illinois.edu/talks.html

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

Supercomputers: TACC (for ASTRAL) and BlueWaters (for NJMerge and TreeMerge)

Accuracy in the presence of HGT + ILS

200 Estimated Gene Trees



Data: Fixed, moderate ILS rate, 50 replicates per HGT rates (1)-(6), 1 model species tree per replicate on 51 taxa, 1000 true gene trees, simulated 1000 bp gene sequences using INDELible [8], 1000 gene trees estimated from GTR simulated sequences using FastTree-2[7]

[7] Price, Dehal, Arkin 2015

[8] Fletcher, Yang 2009

Davidson et al., RECOMB-CG, BMC Genomics 2015