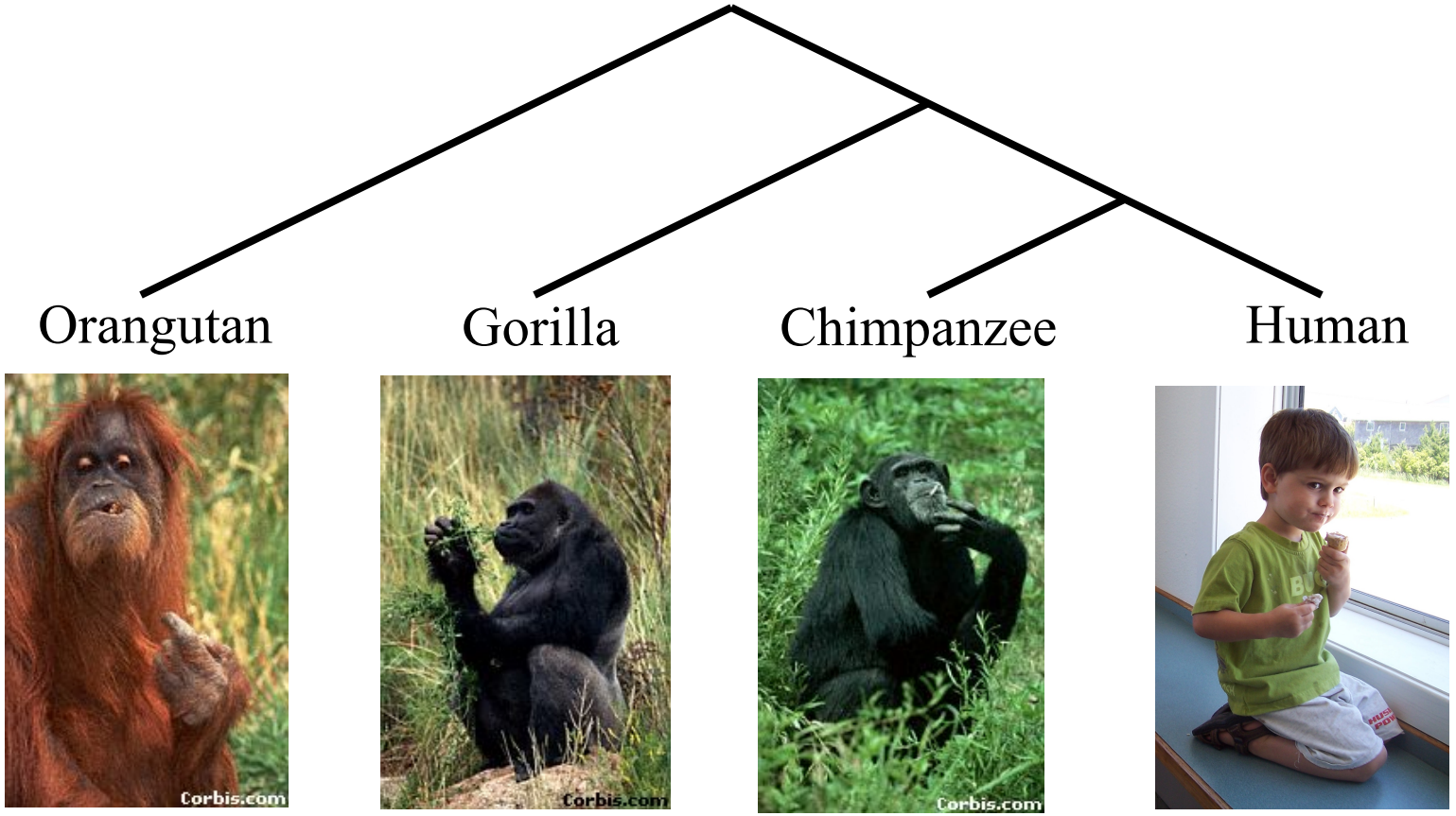
PPoSS Talk, Nov 2, 2023

Tandy Warnow

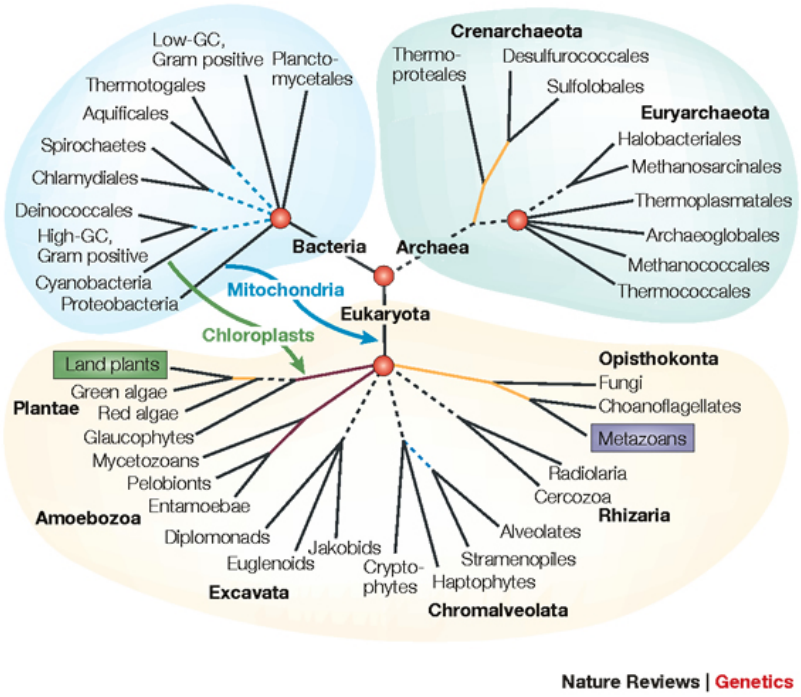
University of Illinois at Urbana-Champaign

Phylogeny (evolutionary tree)



From the Tree of the Life Website,
University of Arizona

Phylogenomics



Phylogeny + genomics = genome-scale phylogeny estimation

.

Estimating the Tree of Life

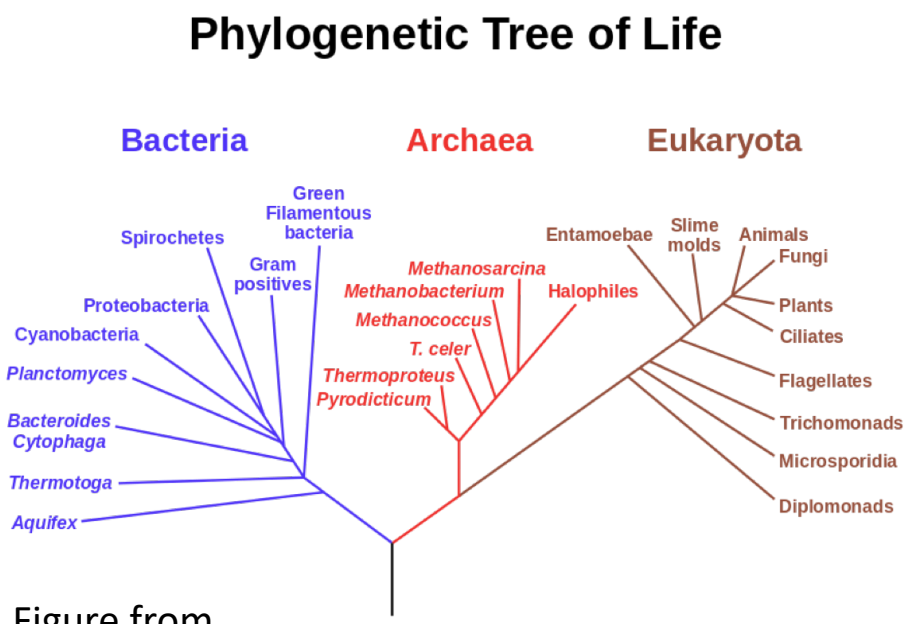


Figure from
https://en.wikipedia.org/wiki/Common_descent

Basic Biology:

How did life evolve?

Applications of phylogenies to:

protein structure and function
population genetics
human migrations
metagenomics

Estimating the Tree of Life

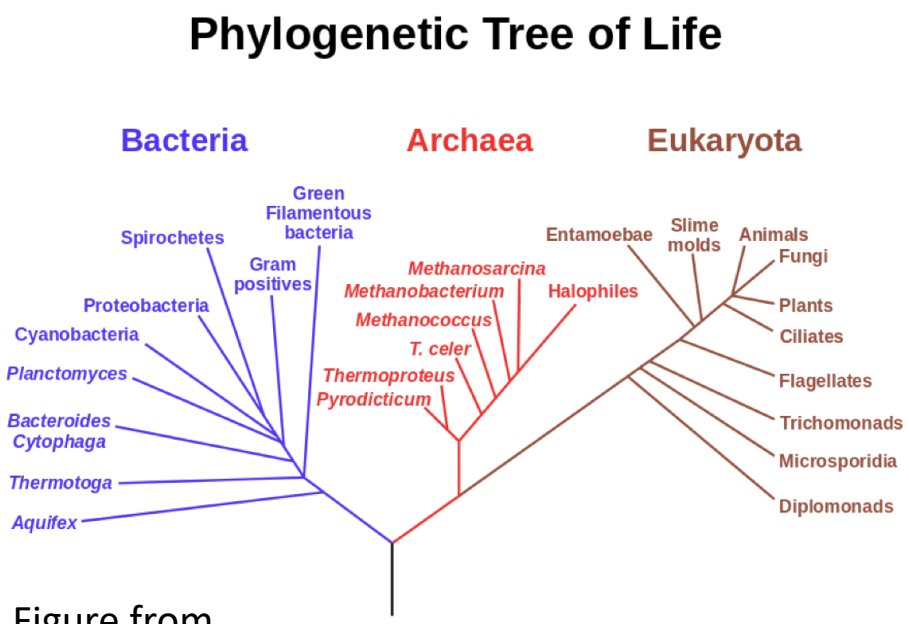


Figure from
https://en.wikipedia.org/wiki/Common_descent

Large datasets!
- Millions of species
- thousands of genes

NP-hard optimization problems
- Exact solutions infeasible
- Approximation algorithms
- Heuristics
- Multiple optima

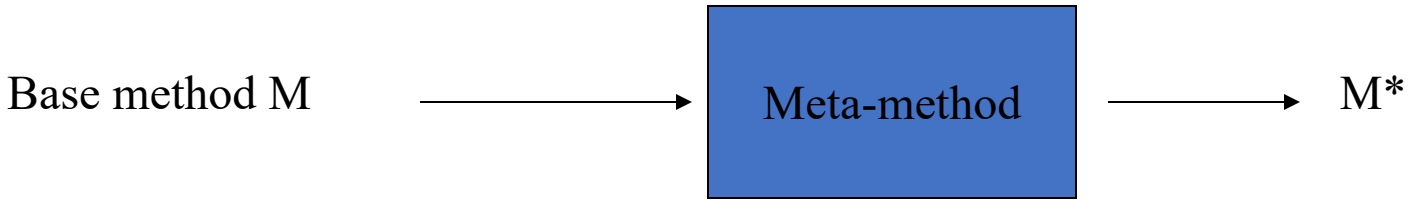
High Performance Computing:
- necessary
- but not sufficient

Goal of this PPoSS project (wrt Phylogeny)

- Phylogeny estimation can be seen as a statistical estimation problem.
- We want fast and accurate methods that are scalable to large datasets (thousands to hundreds of thousands of species, and genome-scale data).
- We also want these methods to have statistical guarantees (provably statistically consistent).
- The basic technique we will use is divide-and-conquer: using the best methods on subsets.

“Boosters”, or “Meta-Methods”

- Meta-methods use divide-and-conquer and iteration (or other techniques) to “boost” the performance of base methods (phylogeny reconstruction, alignment estimation, etc)



Today's ***Fast*** Intro to Phylogenetics Research

- Models of evolution, identifiability, statistical consistency
- Trees, additive matrices, and chordal graphs
- Divide-and-conquer phylogeny estimation: overlapping vs disjoint subsets
- Genome-scale phylogeny:
  - Incomplete lineage sorting and species tree estimation under the Multi-Species Coalescent model (MSC)
  - ASTRAL: non-parametric accurate and statistically consistent species tree estimation under the MSC
  - TreeMerge/GTM: scaling species tree methods to large datasets

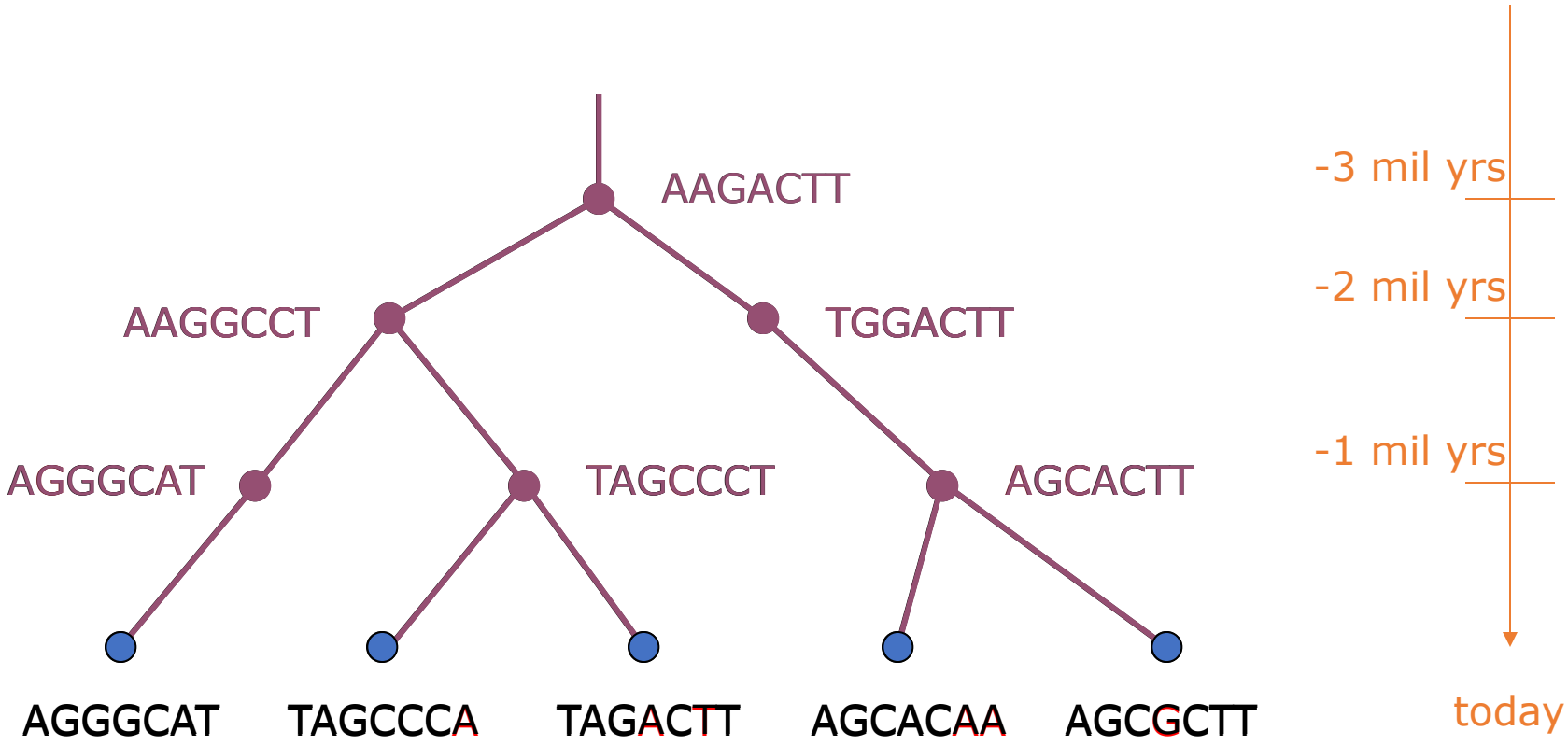
Phyogenomic Pipeline

- Assemble and annotate genomes (e.g., determine orthologs)
- Compute multiple sequence alignments of individual loci
- Construct gene trees
- Construct species tree
- Perform post-tree analyses (e.g., estimate dates, infer selection, etc.)

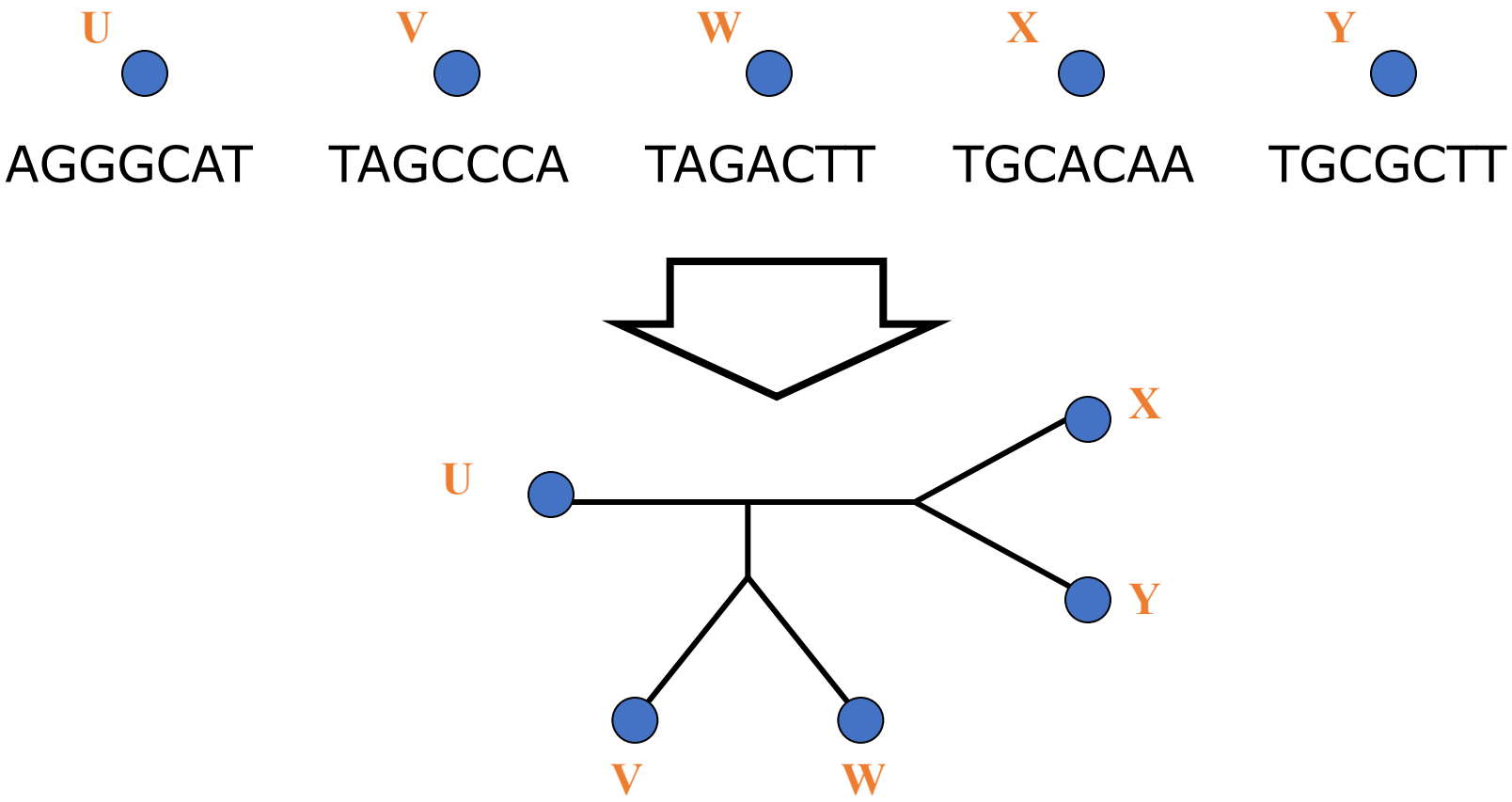
Phyogenomic Pipeline

- Assemble and annotate genomes (e.g., determine orthologs)
- Compute multiple sequence alignments of individual loci
- Construct gene trees
- Construct species tree
- Perform post-tree analyses (e.g., estimate dates, infer selection, etc.)

DNA Sequence Evolution (Idealized)



Phylogeny Problem



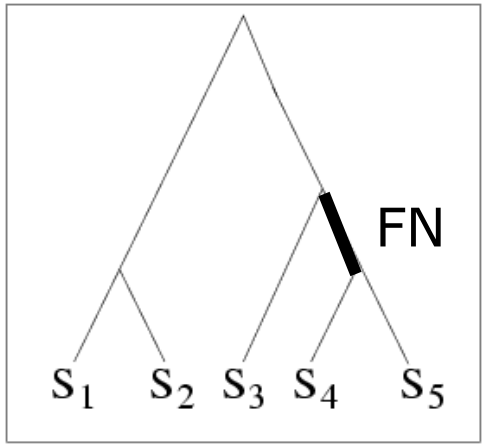
Markov Models of Sequence Evolution

The different sites are assumed to evolve ***i.i.d***. down the model tree

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



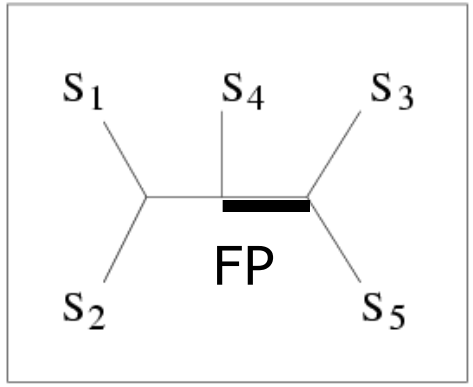
TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

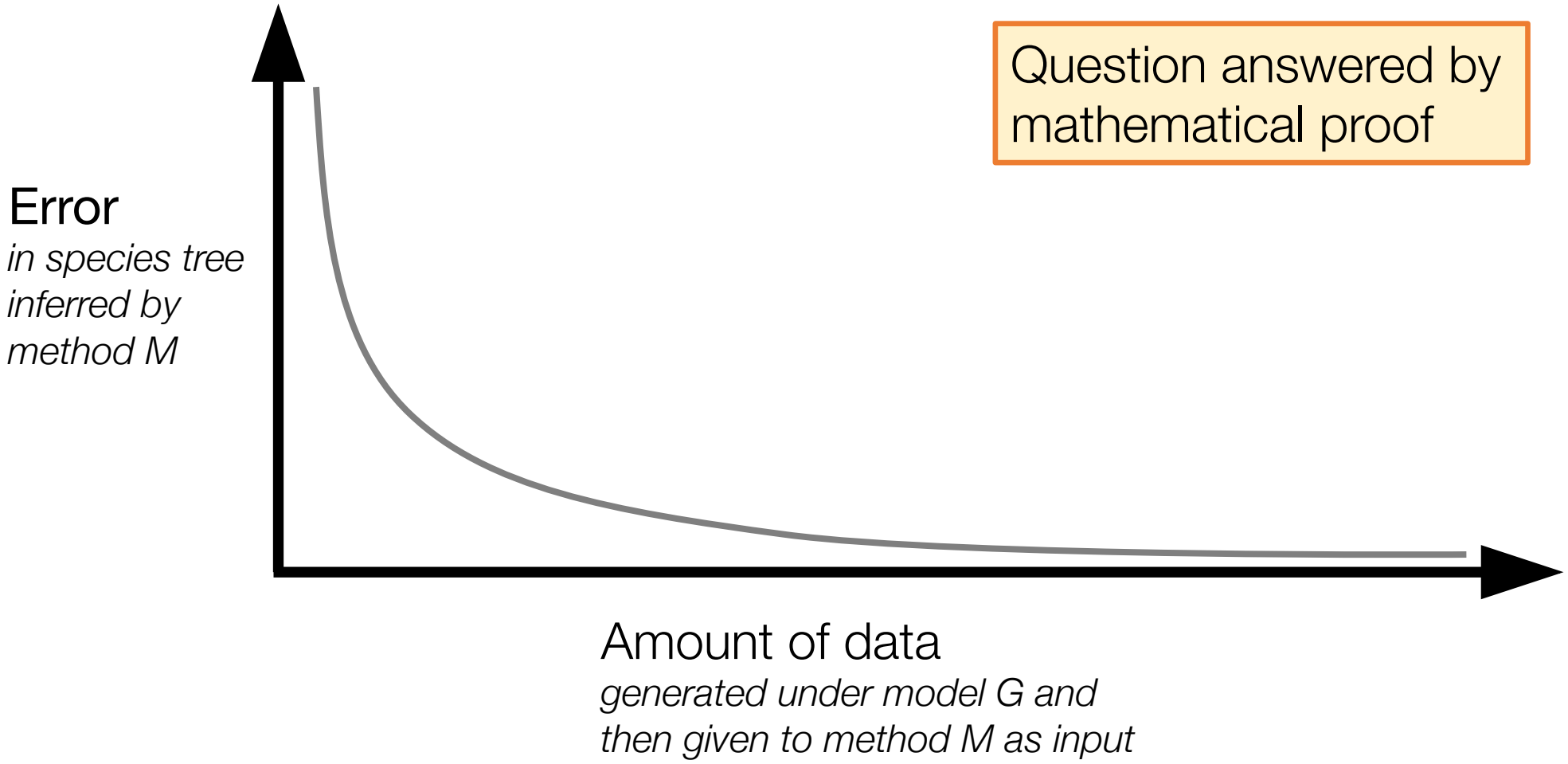
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate



INFERRED TREE

Is method M statistically consistent under model G?



Questions

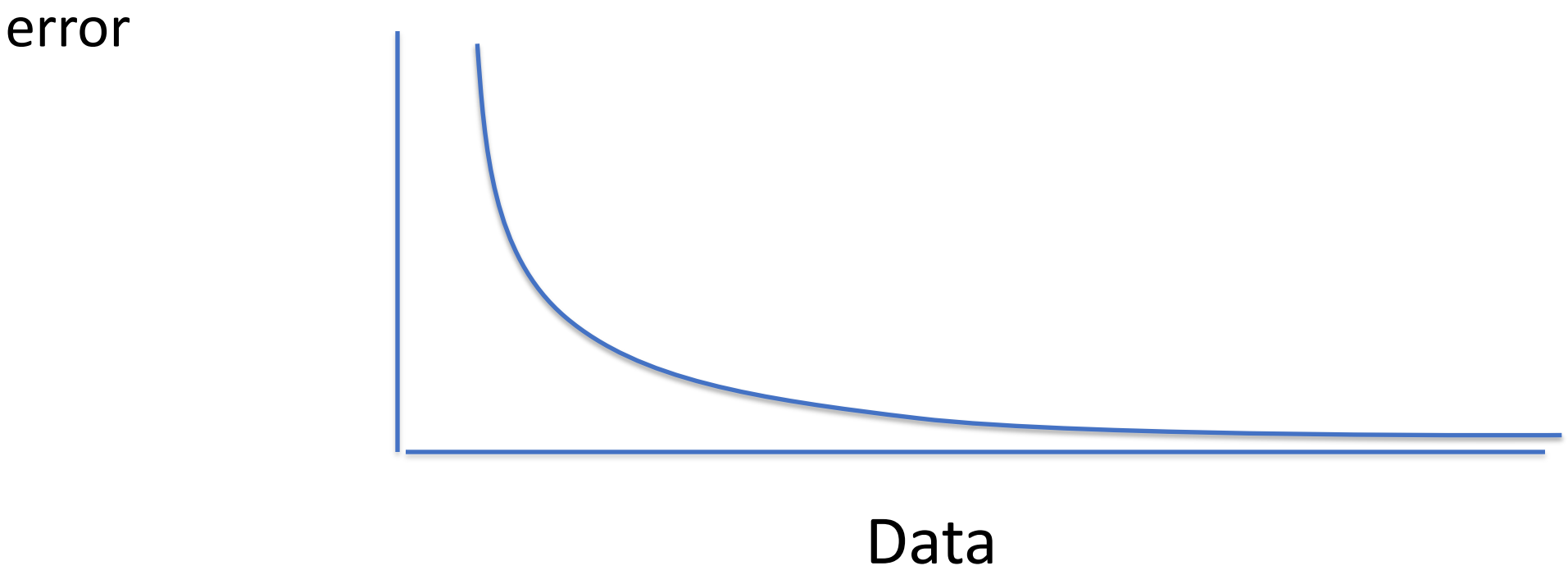
- Is the model tree identifiable?
- Which estimation methods are statistically consistent under this model?
- How much data does the method need to estimate the model tree correctly (with high probability)?
- What are the computational issues?

Answers for Gene Tree Evolution?

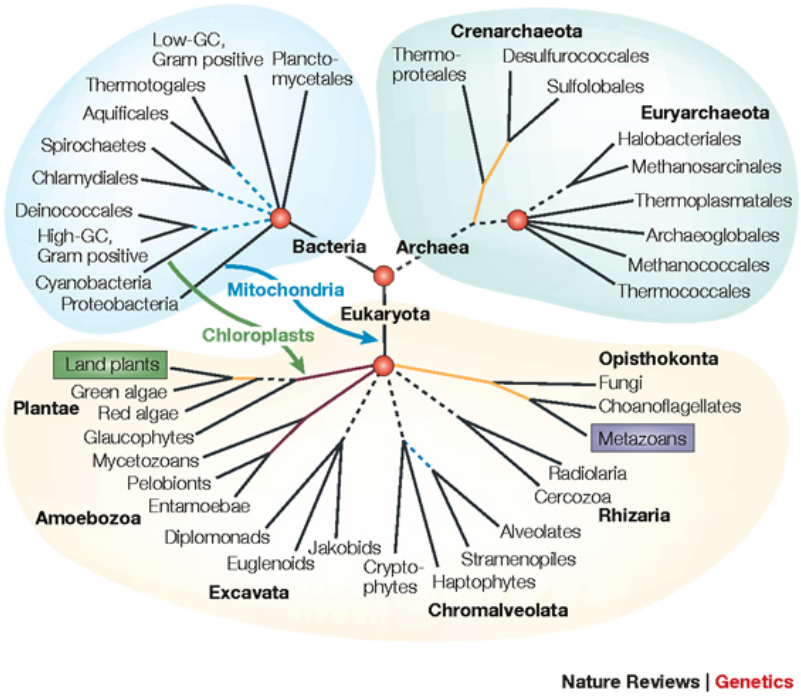
- We know a lot about which site evolution models are identifiable, and which methods are statistically consistent.
  - Maximum Likelihood statistically consistent, but NP-hard (good heuristics)
  - Distance-based methods also statistically consistent and typically polynomial time, but generally less accurate than maximum likelihood
- We know a little bit about the sample complexity (i.e. sequence length requirements) for standard methods.
  - Maximum likelihood has optimal sample complexity, standard distance-based methods do not

Take home message: maximum likelihood preferred, even though hard to find good solutions

Genome-scale data?



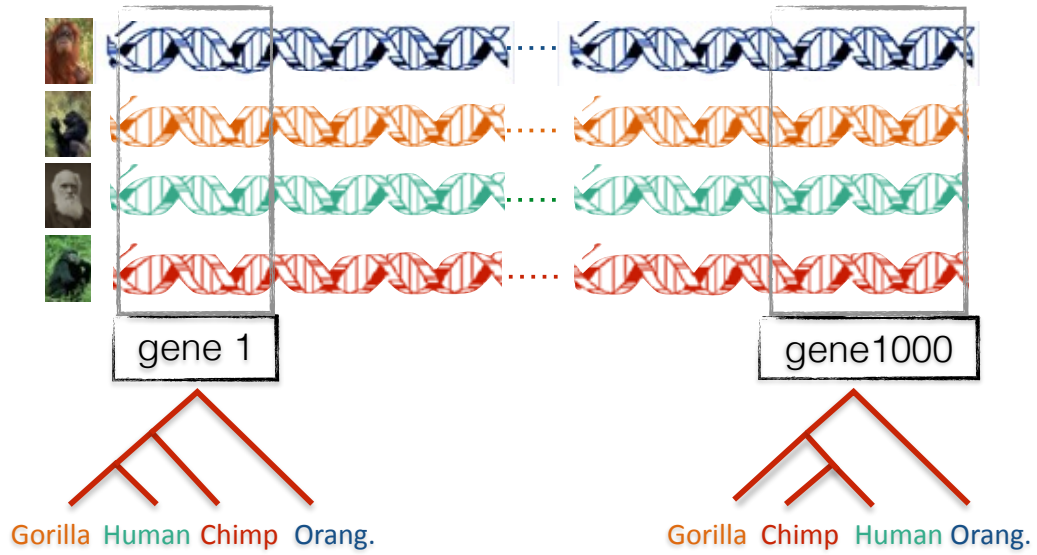
Phylogenomics



Phylogeny + genomics = genome-scale phylogeny estimation

.

Gene tree discordance



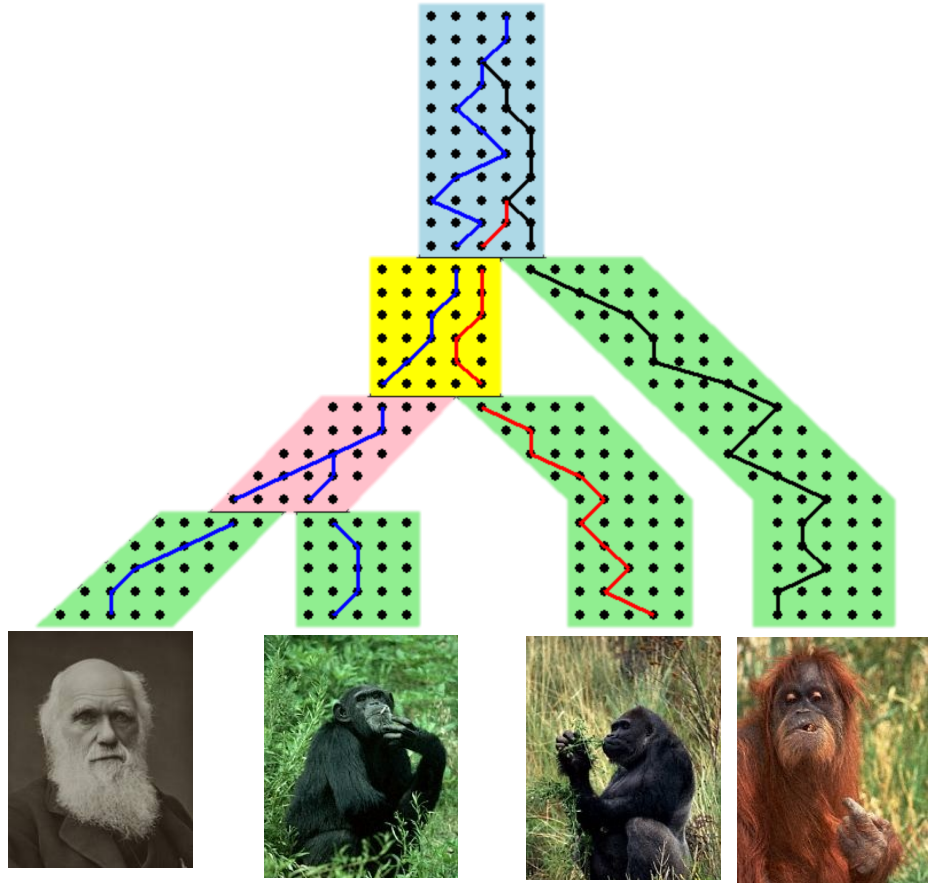
Incomplete Lineage Sorting (ILS) is a dominant cause of gene tree heterogeneity

Gene Trees inside the Species Tree (Multi-Species Coalescent)

Past

Present

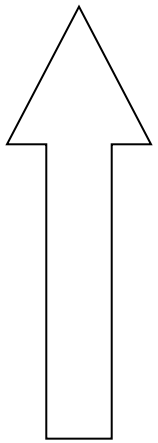
Courtesy James Degnan



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

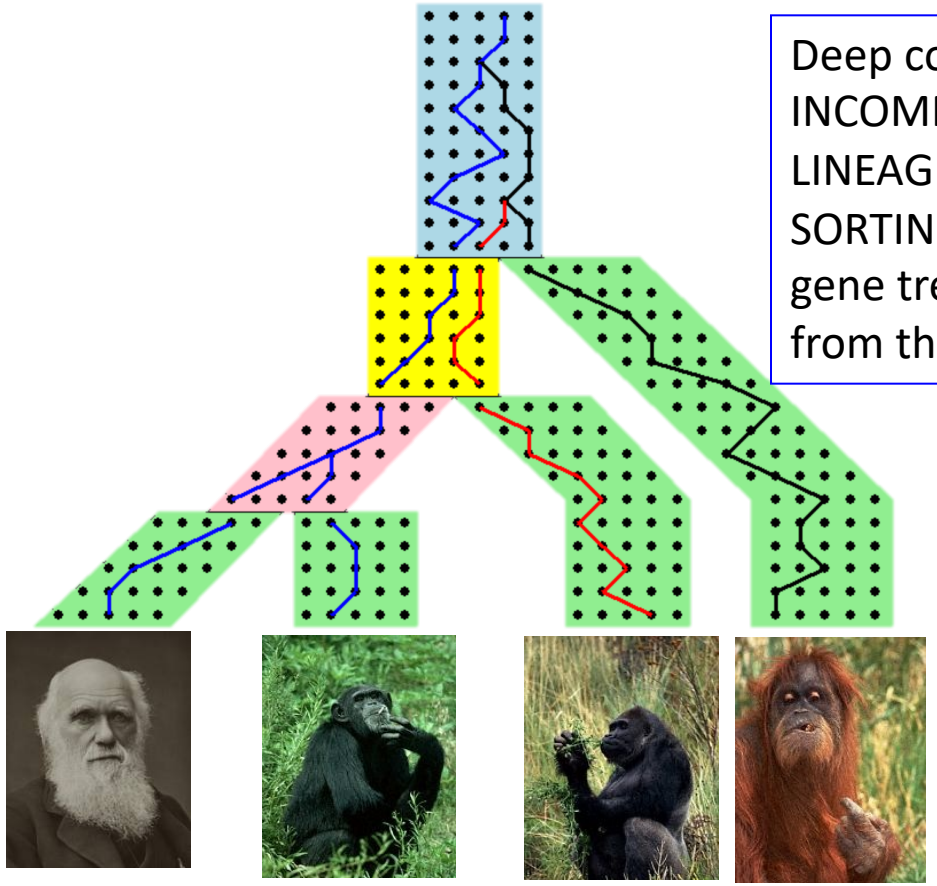
Gene Trees inside the Species Tree (Multi-Species Coalescent)

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

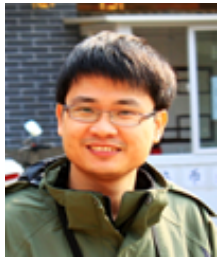
Major Challenge:
- Massive gene tree heterogeneity consistent with ILS

Avian Phylogenomics Project

Erich Jarvis, HHMI

MTP Gilbert, Copenhagen

Guojie Zhang, BGI



Siavash Mirarab, Texas

Tandy Warnow, Texas and UIUC

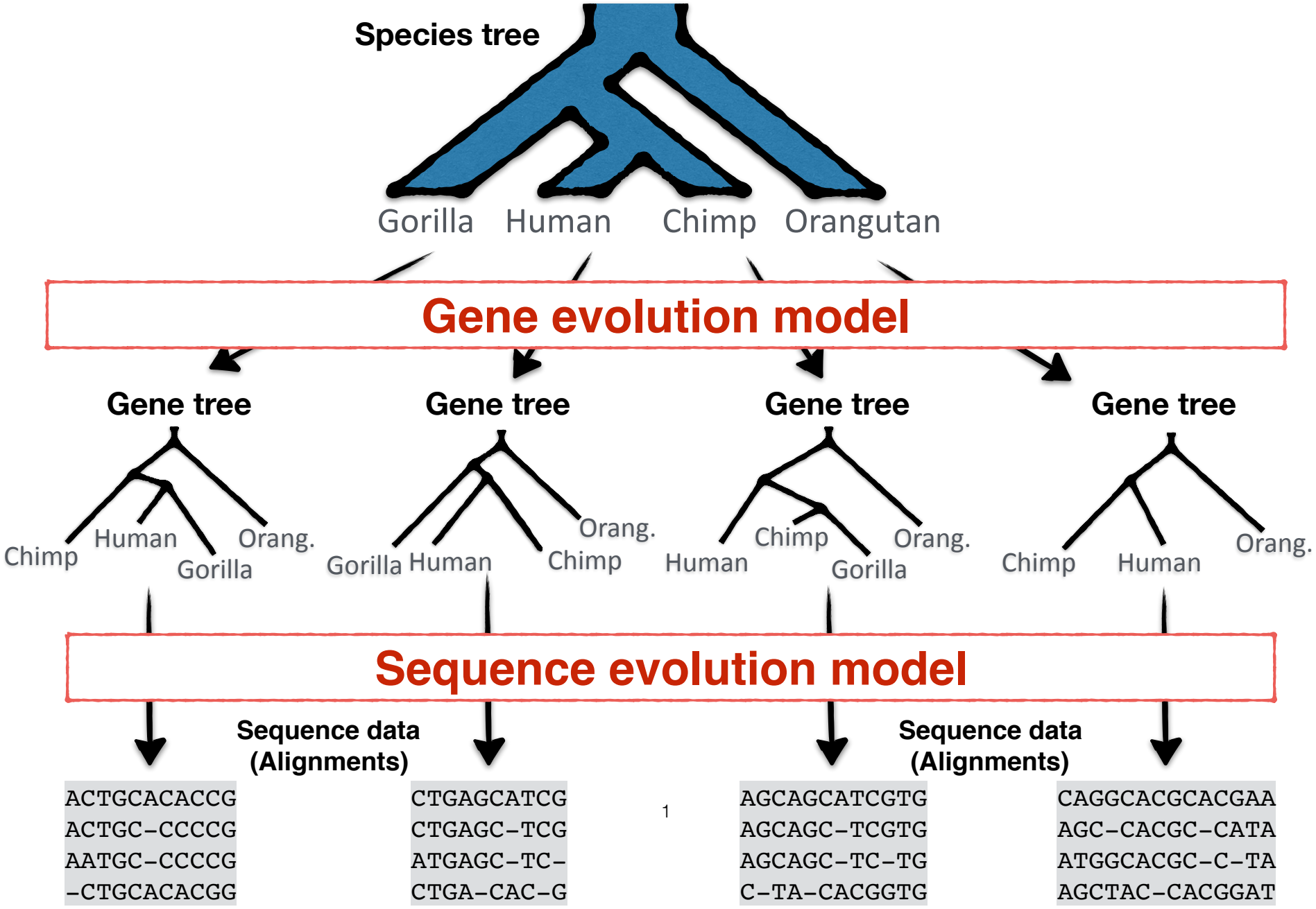
- Approx. 50 species, whole genomes
- 14,000 loci
- Multi-national team (100+ investigators)
- 8 papers published in special issue of Science 2014

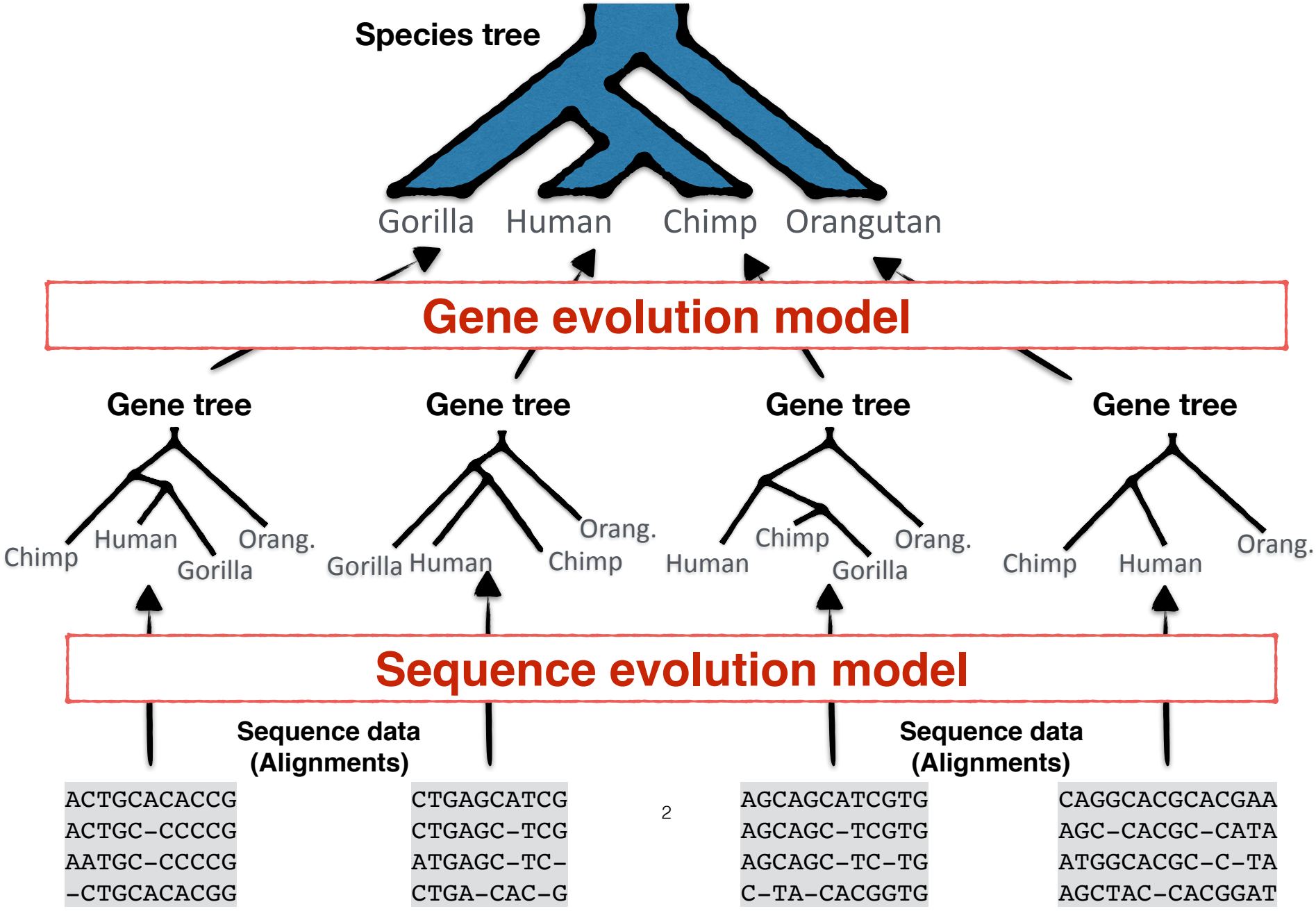
Major challenge:

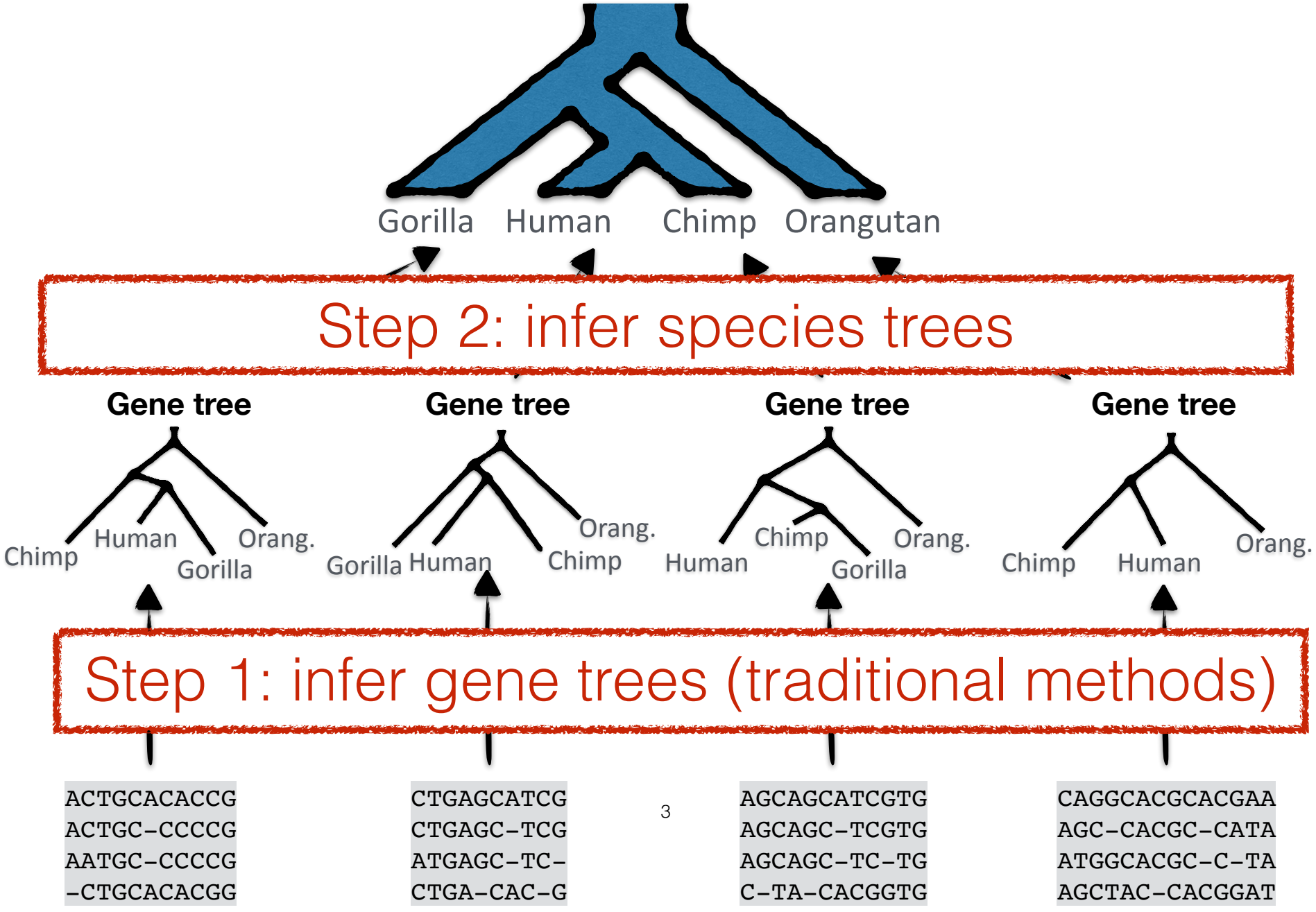
- Massive gene tree heterogeneity consistent with ILS.

Hierarchical Model: MSC+GTR

- Multi-locus data, generated by a hierarchical model
  - Species tree generates gene trees under Multi-Species Coalescent (MSC)
  - Gene trees generate sequences under the Generalized Time Reversible (GTR) model

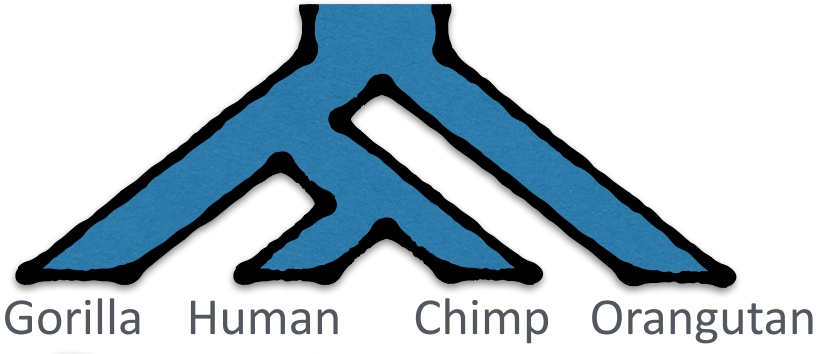

Species tree
Gorilla
Human
Chimp
Orangutan
Gene evolution model
Gene tree
Gene tree
Gene tree
Gene tree
Chimp
Human
Orang.
Gorilla
Gorilla
Human
Orang.
Chimp
Human
Chimp
Orang.
Gorilla
Chimp
Human
Orang.
Sequence evolution model
Sequence data
(Alignments)
Sequence data
(Alignments)
ACTGCACACCG
ACTGC-CCCCG
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

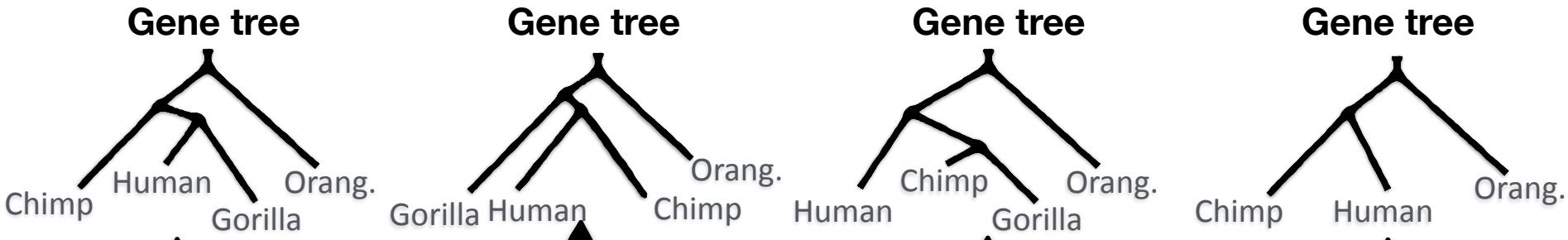

Gorilla Human Chimp Orangutan
Step 2: infer species trees
Gene tree
Gene tree
Gene tree
Gene tree
Chimp Human Orang. Gorilla
Gorilla Human Orang. Chimp
Human Chimp Orang. Gorilla
Chimp Human Orang.
Step 1: infer gene trees (traditional methods)
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
CTGAGCATCG
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

Suppose we can estimate all the gene trees correctly.

Can we estimate the species trees from lots of true gene trees?



Step 2: infer species trees



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
```

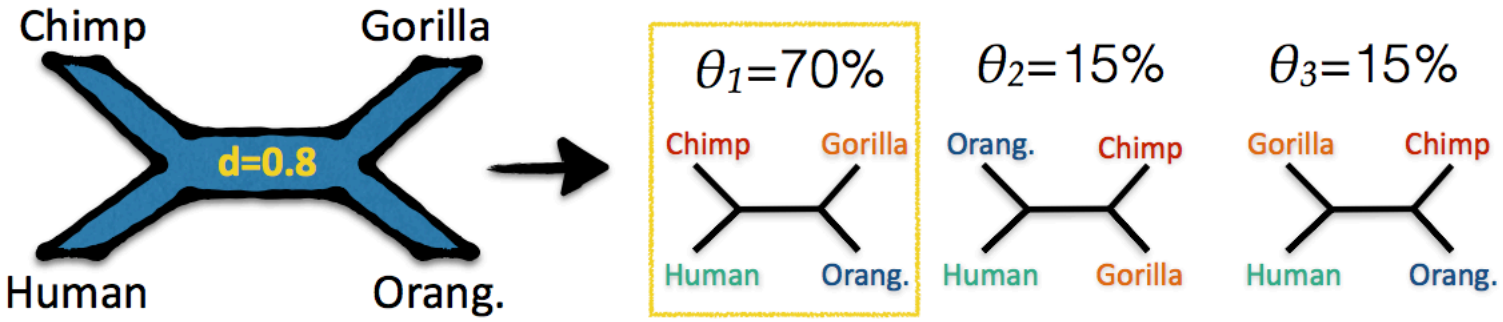
```
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT
```

How to estimate a 4-leaf species tree

Theorem (Allman et al.): Under the multi-species coalescent model, for any four taxa A, B, C, and D, the most probable unrooted gene tree on {A,B,C,D} is identical to the unrooted species tree induced on {A,B,C,D}.

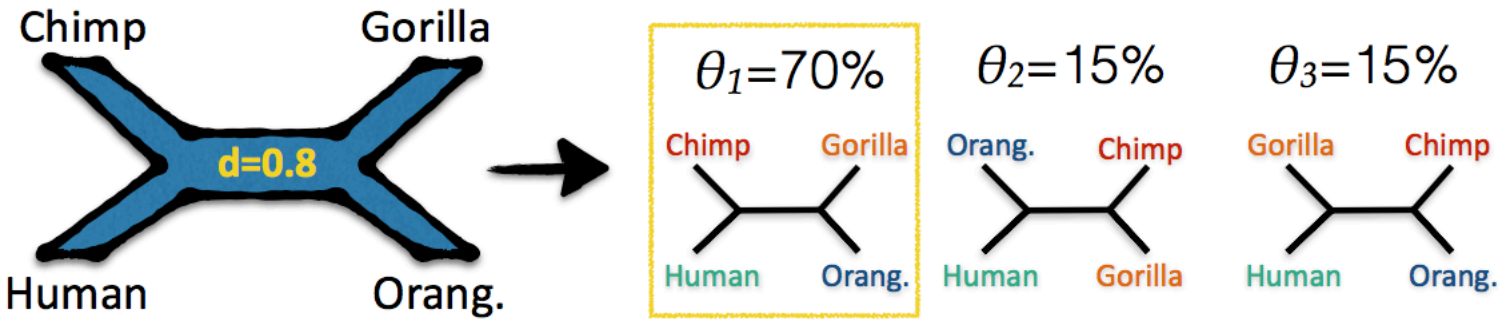
How to estimate a 4-leaf species tree

Theorem (Allman et al.): Under the multi-species coalescent model, for any four taxa A, B, C, and D, the most probable unrooted gene tree on {A,B,C,D} is identical to the unrooted species tree induced on {A,B,C,D}.



Species tree estimation from unrooted gene trees

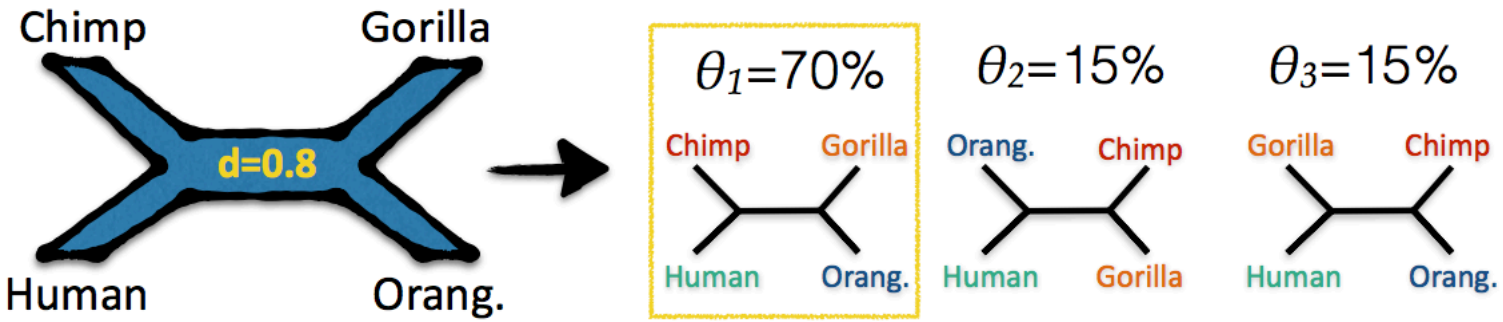
Corollary: Under the multi-species coalescent model, the species tree is identifiable from the gene tree distribution



Species tree estimation from unrooted gene trees

Corollary: Under the multi-species coalescent model, the species tree is identifiable from the gene tree distribution

Proof: For every four species, select most frequently observed tree as the species tree. Then combine quartet trees!



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

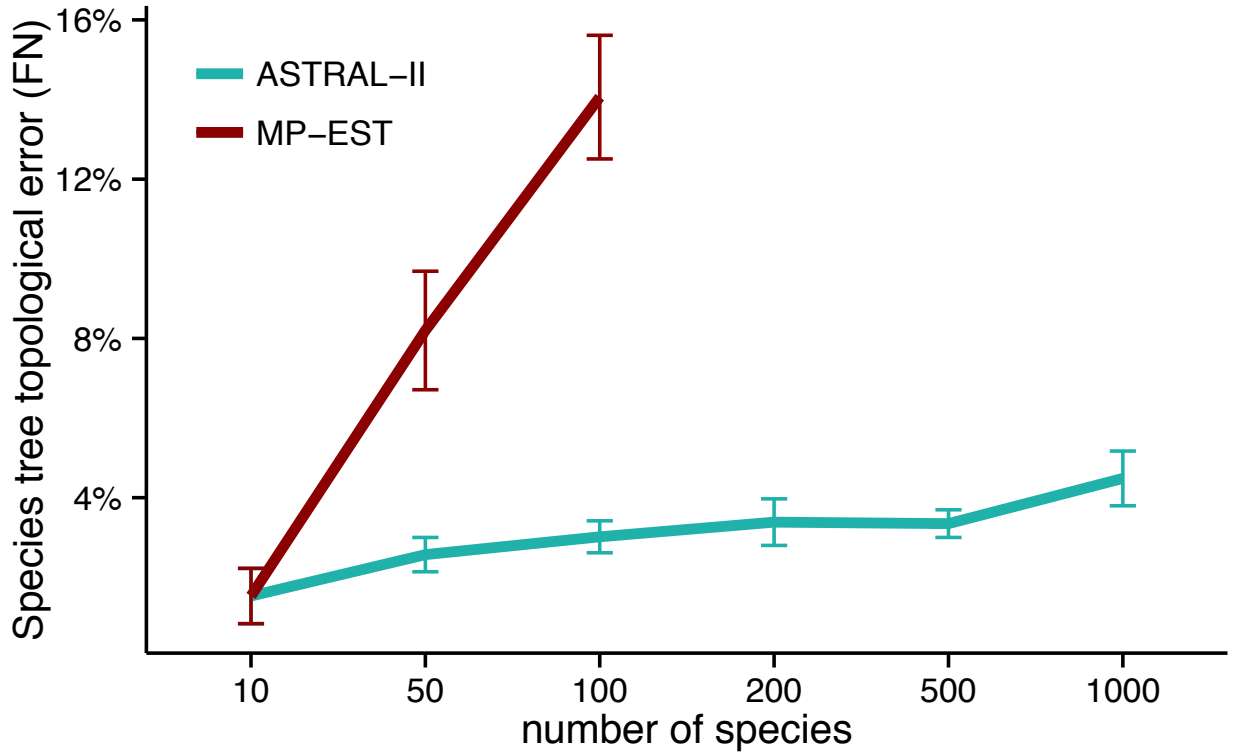
all input gene trees

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

ASTRAL

- 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

Tree accuracy when varying the number of species



1000 genes, “medium” levels of recent ILS

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 – great, but…

- The good: ASTRAL is
  - increasingly used in practice
  - statistically consistent given true gene trees
  - sometimes more accurate than concatenation, but impacted by gene tree estimation error
  - very fast for many datasets (faster than concatenation)
- 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

Goal of this PPoSS project (wrt Phylogeny)

- Phylogeny estimation can be seen as a statistical estimation problem.
- We want fast and accurate methods that are scalable to large datasets (thousands to hundreds of thousands of species, and genome-scale data).
- We also want these methods to have statistical guarantees (provably statistically consistent).
- The basic technique we will use is divide-and-conquer: using the best methods on subsets.

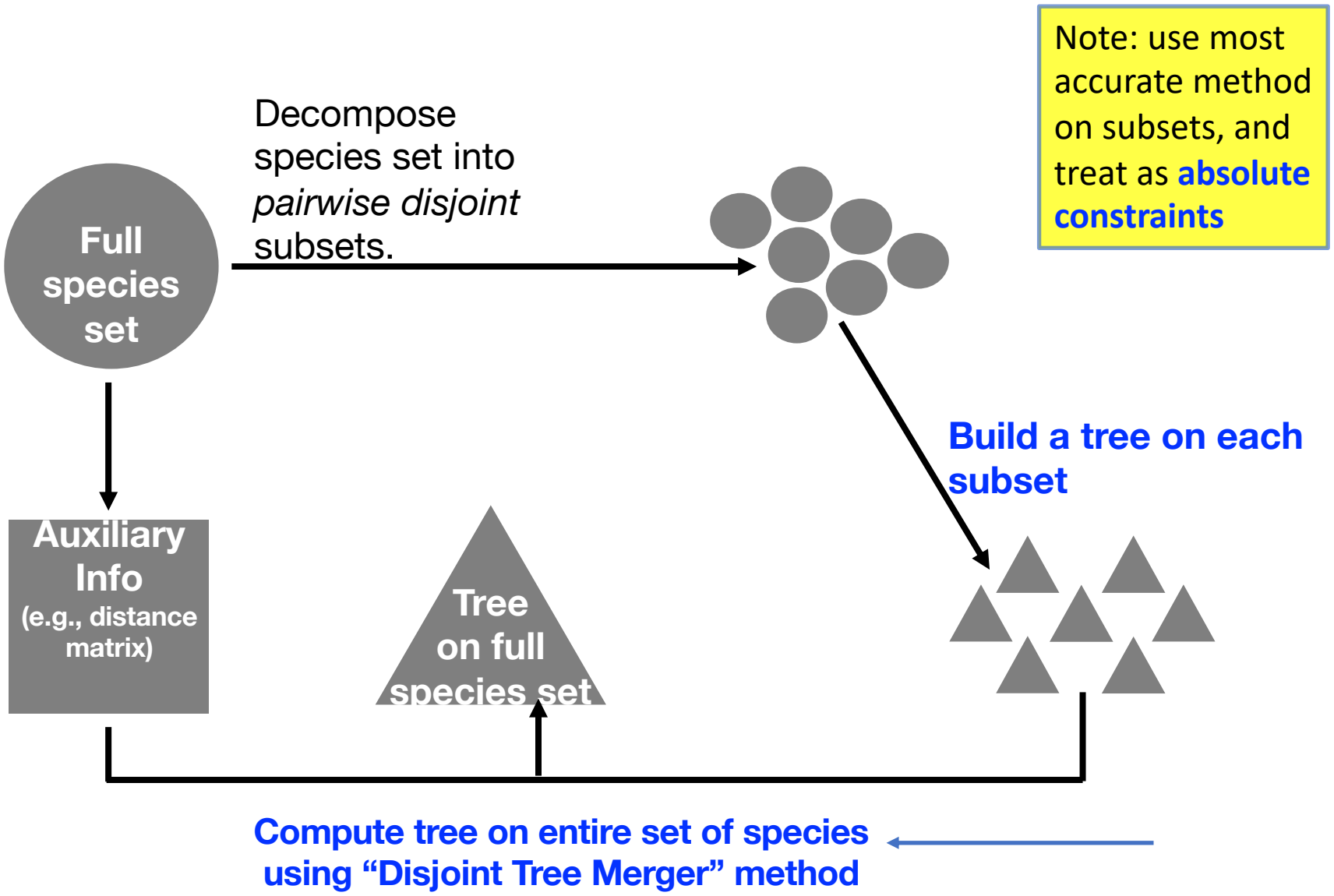
Divide-and-conquer using Chordal Graphs

- A matrix is additive if it equals path lengths in an edge-weighted tree
- Distances calculated in phylogenetics (from sequence data) converge to additive matrices, as the sequence length increases
- If we threshold an additive matrix, we obtain a chordal graph: one that has no simple cycles of size four or larger
- Chordal graphs have lovely properties
  - Can list all maximal cliques in polynomial time
  - Minimum vertex separators are maximal cliques
  - Can obtain decompositions into overlapping subsets, and employ in divide-and-conquer strategies

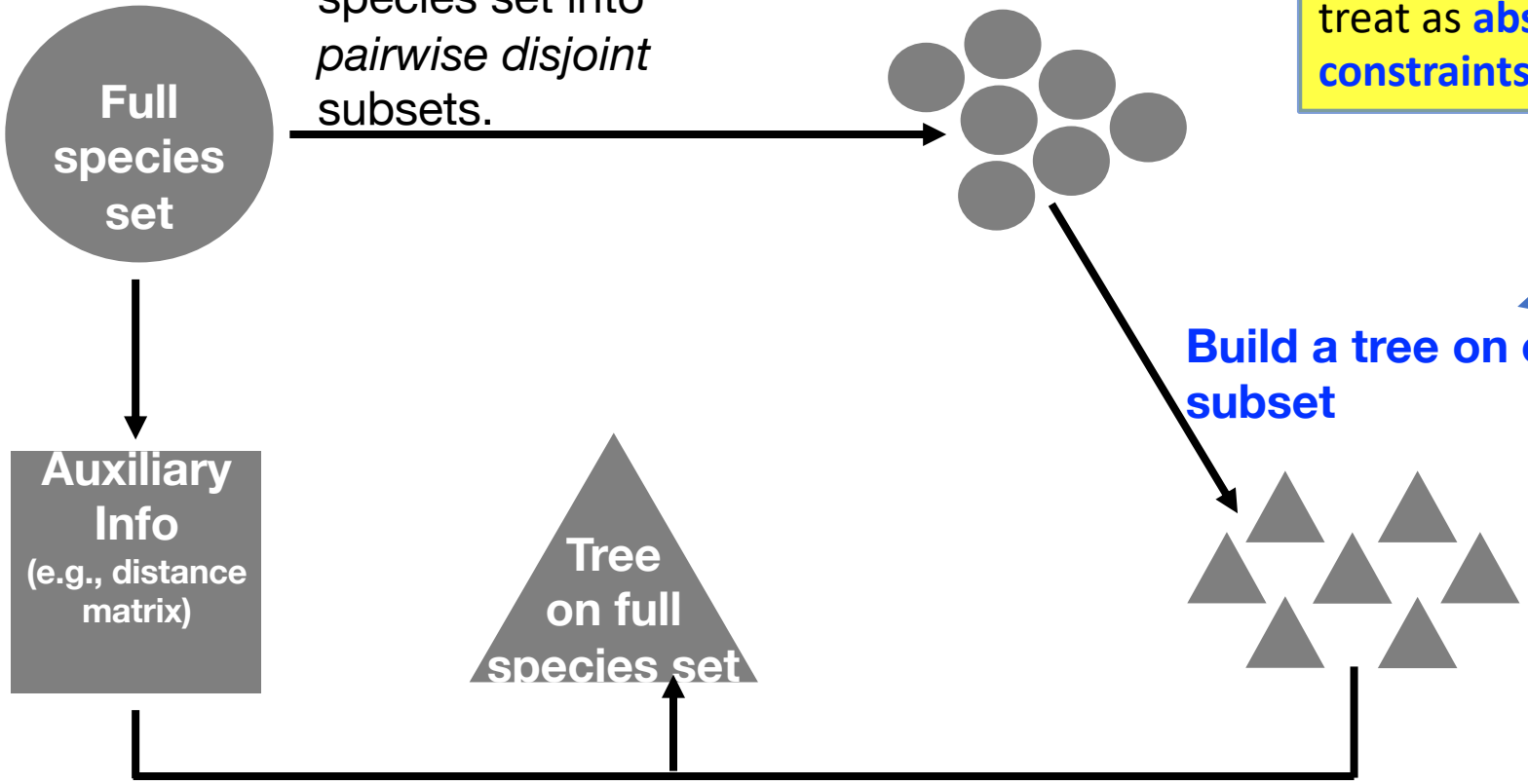
Divide-and-conquer using Chordal Graphs

- Chordal graphs have lovely properties
  - Can list all maximal cliques in polynomial time
  - Minimum vertex separators are maximal cliques
  - Can obtain decompositions into overlapping subsets, and employ in divide-and-conquer strategies
- If we do this, we need methods that combine overlapping subset trees, i.e., “supertree” methods
- These approaches have not been as scalable as needed.

Divide-and-Conquer using Disjoint Tree Mergers



Divide-and-Conquer using Disjoint Tree Mergers



Divide-and-Conquer using Disjoint Tree Mergers

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

Decompose species set into *pairwise disjoint* sub

Full species set

Auxiliary Info (e.g., distance matrix)

Theorem: TreeMerge+ASTRAL and GTM+ASTRAL are both **statistically consistent** and **polynomial time**

ASTRAL

tree on each

species set

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

TreeMerge or GTM

Guide Tree Merger

- Smirnov and Warnow, RECOMB-Comparative Genomics
- Guide Tree Merger (GTM): Another Disjoint Tree Merger method… unlike TreeMerge, it does not allow blending
- Github site: https://github.com/vlasmirnov/GTM

Algorithmic strategy:

- divide species set into disjoint subsets,
- compute species trees on the subsets using selected species tree method, and
- connect subset trees by adding edges (no blending!), so as to minimize distance to the given guide tree (polynomial time!)

ASTRAL+GTM: better than ASTRAL!

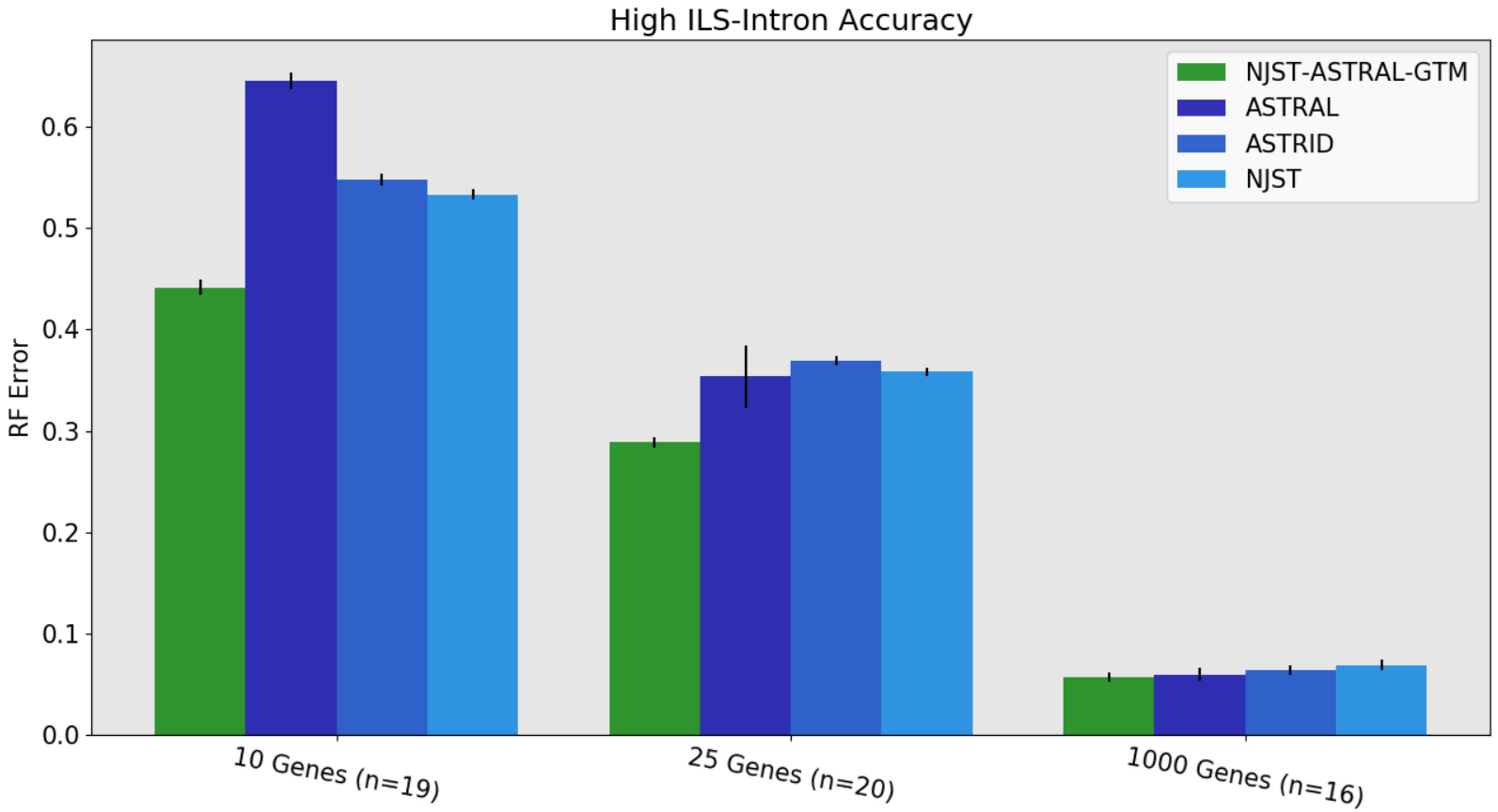
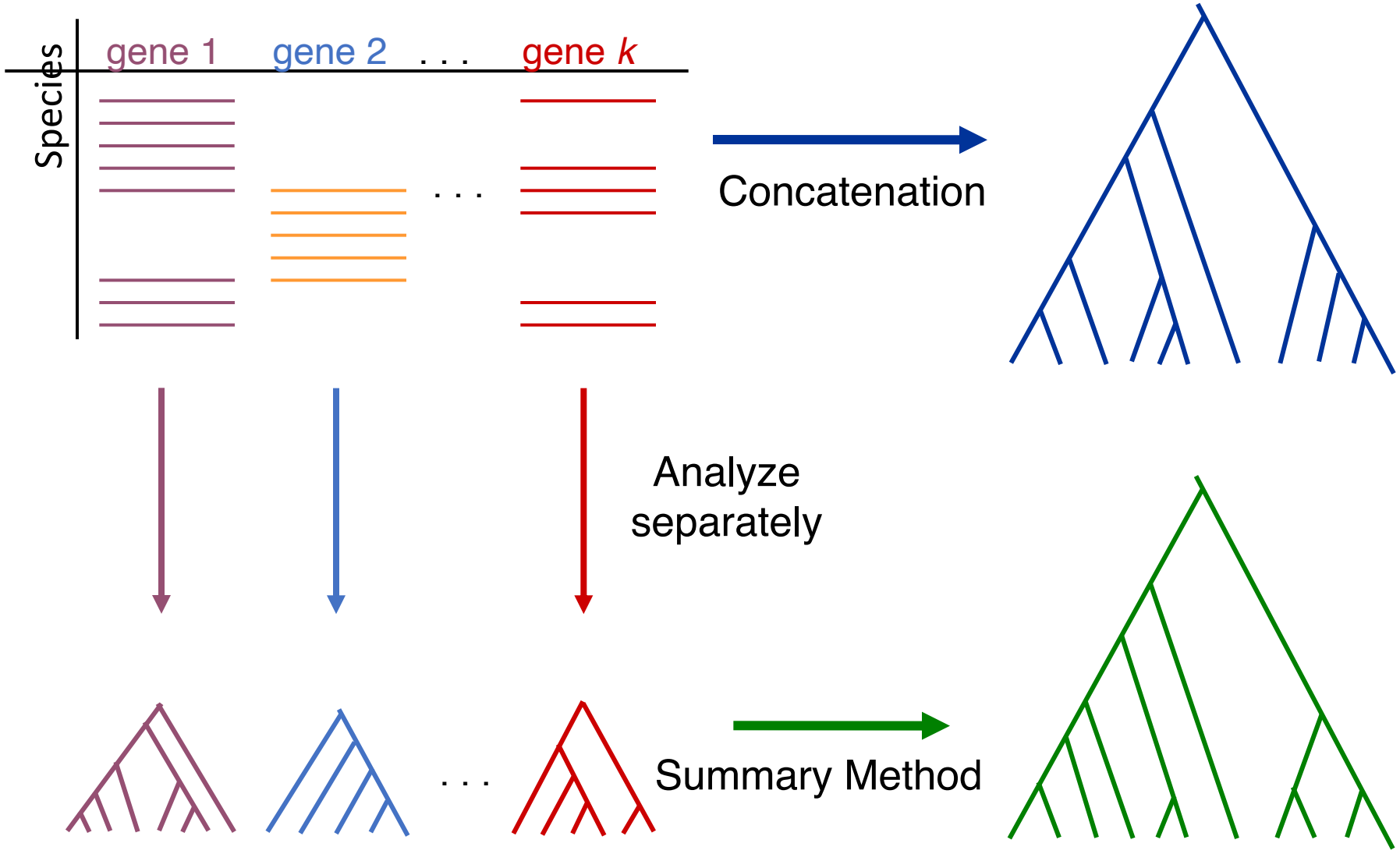


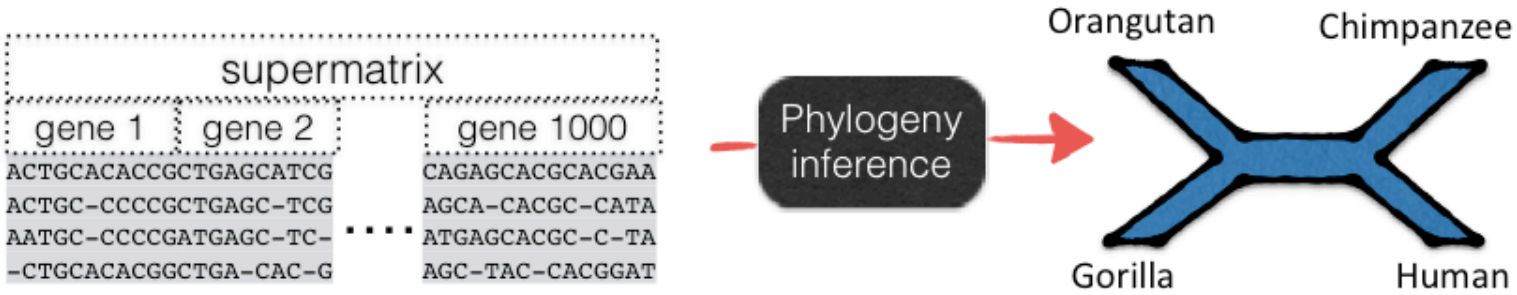
Table 4 Comparison of average runtime (seconds) of NJst-ASTRAL-GTM and ASTRAL for high ILS conditions with introns on 1000 species

	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

Main competing approaches



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]

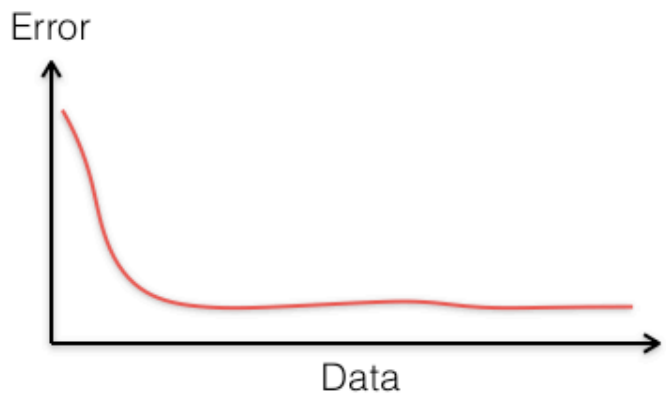


Table 6 Average runtime (seconds) of FastTree-RAxML-GTM (GTM(RAxML)) and RAxML on 1000-species exon datasets

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

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

The GTM pipeline used with RAxML on subsets matches accuracy with RAxML, but is much faster

Summary about phylogenetic tree estimation

- The best tree estimation methods are computationally intensive, and tree-space grows exponentially.
- The Divide-and-Conquer pipelines we are developing (especially GTM) maintain statistical consistency, maintain or improve accuracy and are much faster.
- In addition, they naturally enable parallel implementations.

What about Community Detection?

- I also work on community detection in large networks, largely in the context of Scientometrics
- Our recent paper (Park et al., Complex Networks 2023) addresses failure of standard community detection methods (aka clustering methods) to produce well-connected clusters.
- See https://tandy.cs.illinois.edu/bibliometrics.html for papers

Networks we studied

network	nodes	edges	avg_deg	ref
Open Citations	75,025,194	1,363,605,603	36.35	(*17*)
CEN	13,989,436	92,051,051	13.16	(*35*)
cit_hepph	34,546	420,877	24.37	(*36*)
cit_patents	3,774,768	16,518,947	8.75	(*36*)
orkut	3,072,441	117,185,083	76.28	(*37*)
wiki_talk	2,394,385	4,659,565	3.89	(*38*)
wiki_topcats	1,791,489	25,444,207	28.41	(*39*)

Table 1: Summary statistics for networks used in this study. Average degree is the average of the node degrees across the network.

Many small edge cuts in Leiden clusters on real-world networks

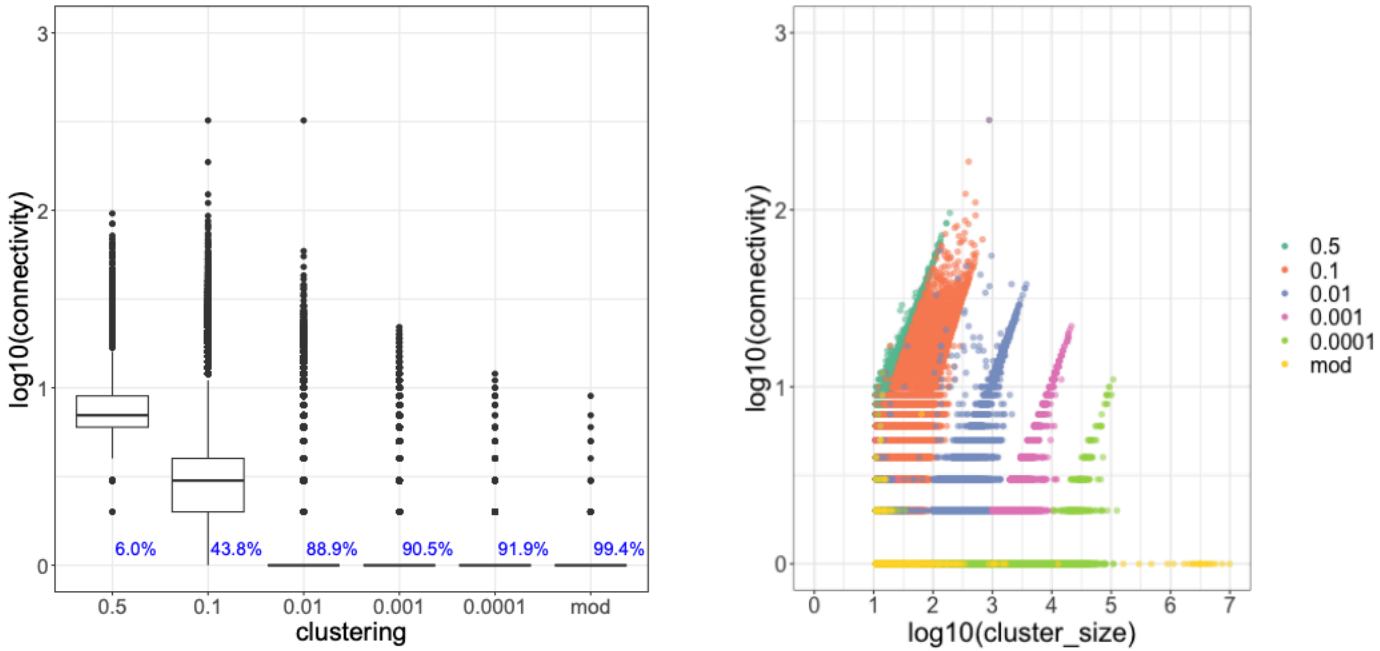
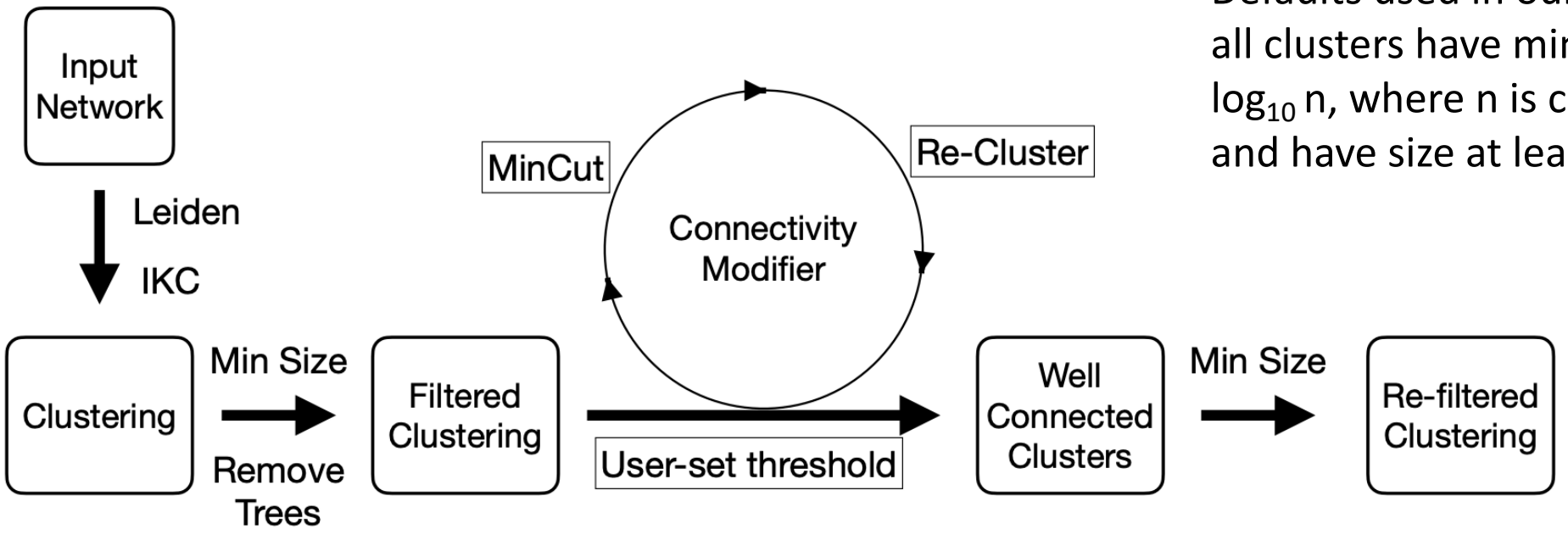


Figure 1: *Node coverage, connectivity, and size distribution of clusters generated by Leiden optimizing either CPM or modularity on the Open Citations network (75,025,194 nodes).* Con-

Leiden optimizing the Constant Potts Model (CPM) or modularity (mod)

Results using other clustering methods are similar

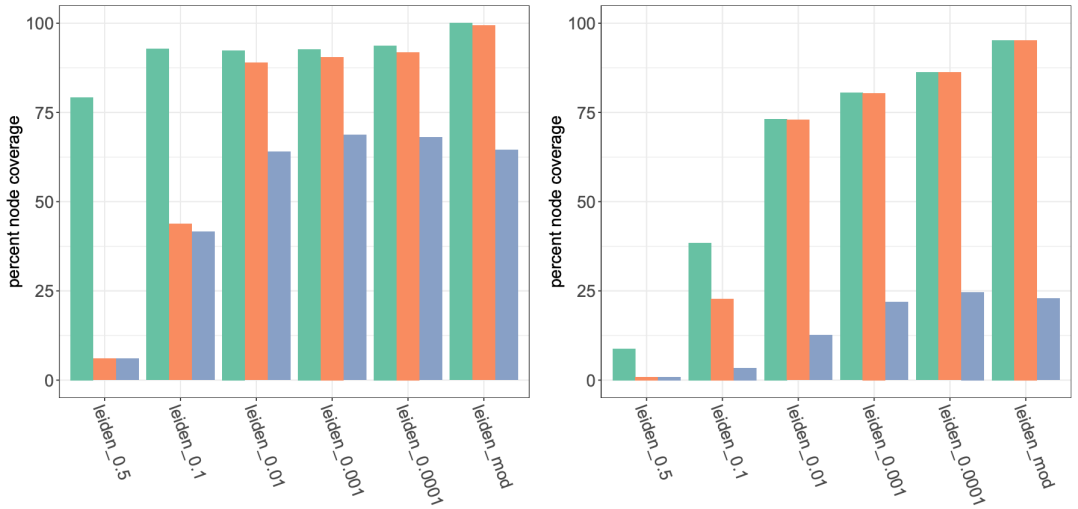
The Connectivity Modifier (CM) Pipeline



Defaults used in our study:
all clusters have min cuts above $\log_{10} n$, where n is cluster size, and have size at least 11

Figure 3: *Connectivity Modifier Pipeline Schematic.* The four-stage pipeline depends on user-

Impact of the Connectivity Modifier



(a) Open Citations

(b) CEN

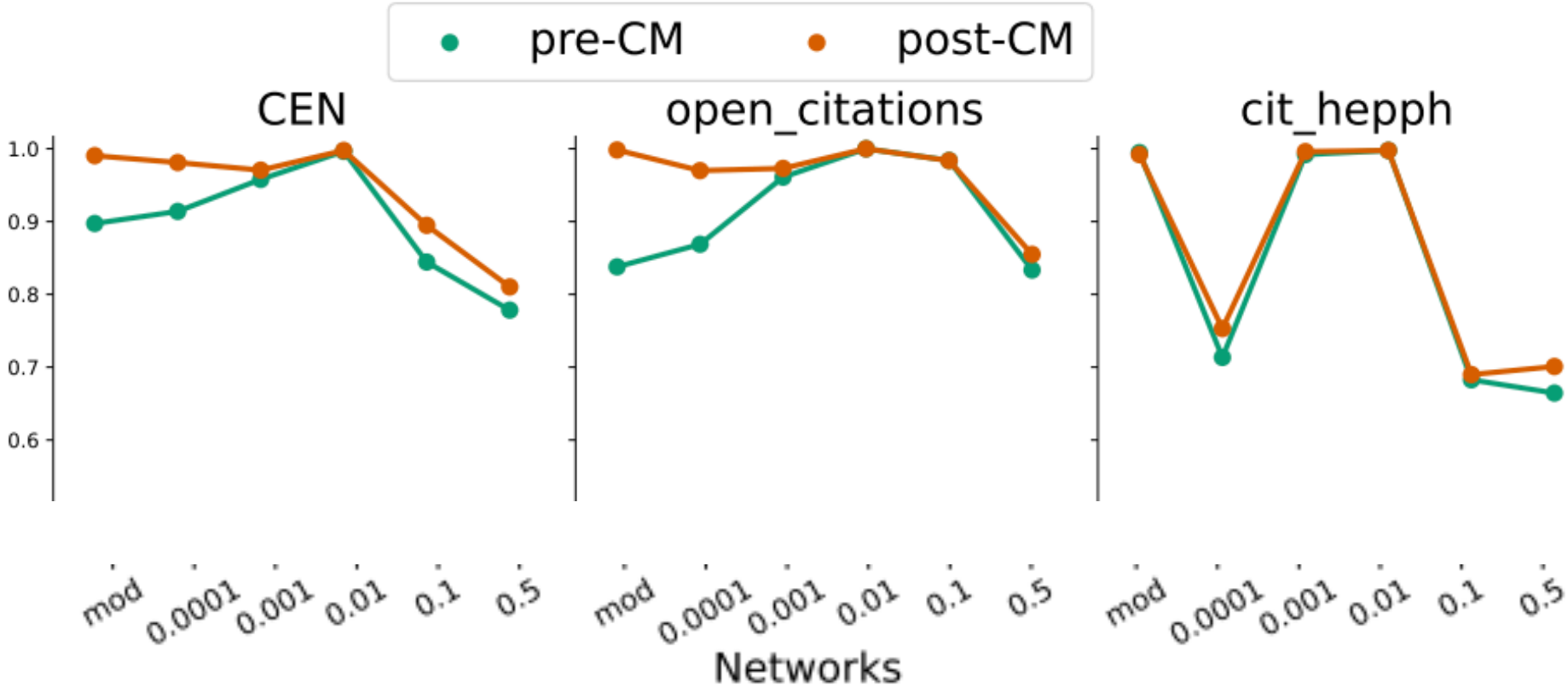
Figure 4: *Reduction in node coverage after CM treatment of Leiden clusters.* The Open Citations (left panel) and CEN (right panel) networks were clustered using the Leiden algorithm under CPM at five different resolution values or modularity. Node coverage (defined as the percentage of nodes in cluster of size at least 2) was computed for Leiden clusters • (lime green), Leiden clusters with trees and/or clusters of size 10 or less filtered out • (soft orange), and after CM treatment of filtered clusters • (desaturated blue).

CPM clustering is impacted by the resolution parameter: small values give high node coverage, but many of these clusters are poorly connected (even trees).

Modularity-optimization is similar to CPM with a small resolution parameter.

Using CM reduces node coverage

The CM pipeline improves accuracy



Results for NMI accuracy on LFR networks. Results for other criteria are similar.

Summary

- The tendency for standard clustering methods to have poorly connected clusters (or else have low node coverage) is striking.
- CM ensures that all returned clusters are well-connected, according to the user specified bound
- CM improves accuracy on LFR networks
- But after CM, there is a drop in node coverage that can be large.
- How do we explain the drop in node coverage?
  - Perhaps not the case that the entire network is covered by communities?