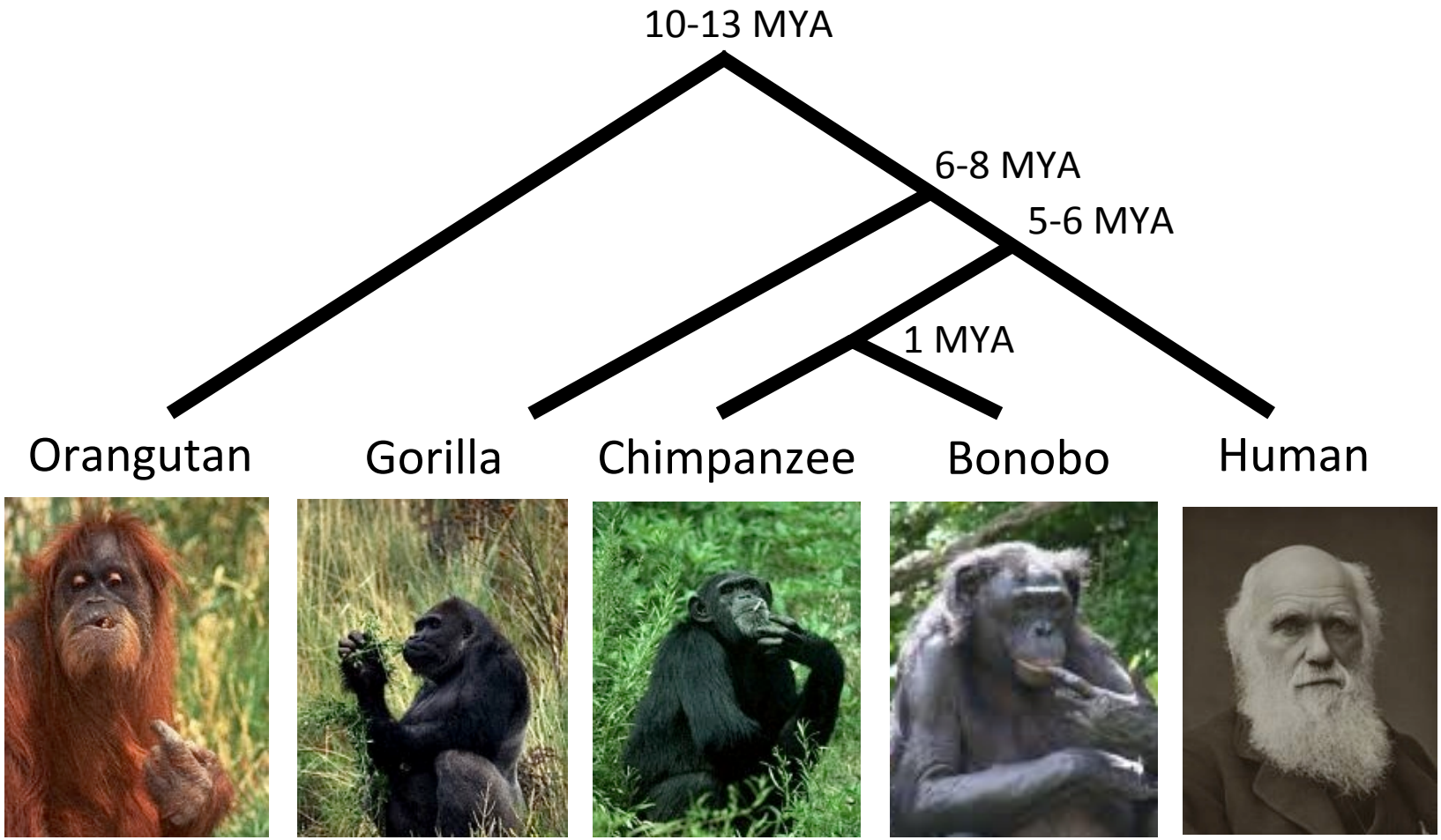
Estimating species trees from multiple gene trees in the presence of ILS

Tandy Warnow

Joint work with Siavash Mirarab,

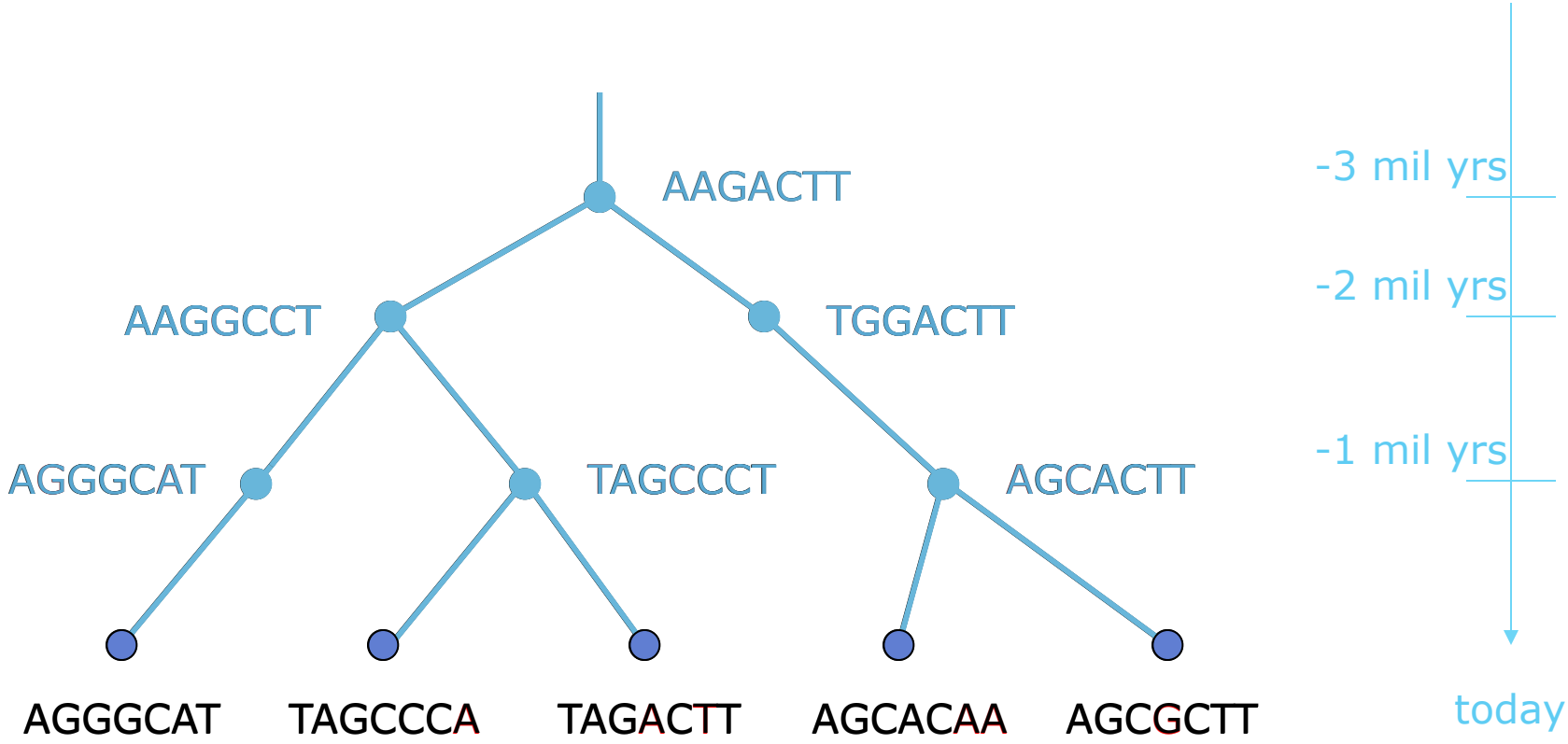
Md. S. Bayzid, and others

Species Tree



From the Tree of the Life Website, University of Arizona
Dates from Lock et al. Nature, 2011

DNA Sequence Evolution



Markov Model of Site Evolution

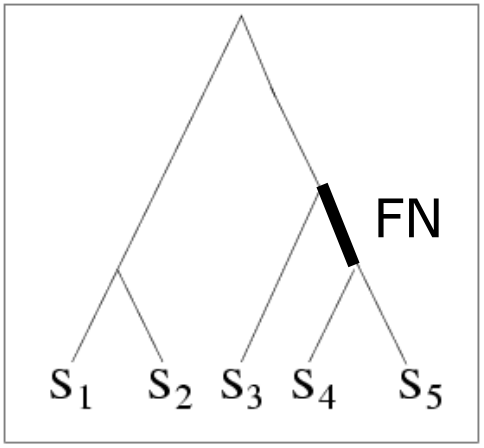
Simplest (Jukes-Cantor):

- The model tree T is binary and has substitution probabilities p(e) on each edge e.
- 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.

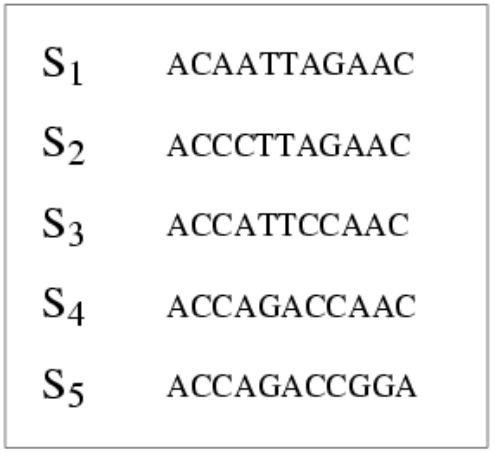
More complex models (such as the General Markov model) are also considered, often with little change to the theory.

Maximum Likelihood is a statistically consistent method under the JC model.

Quantifying Error



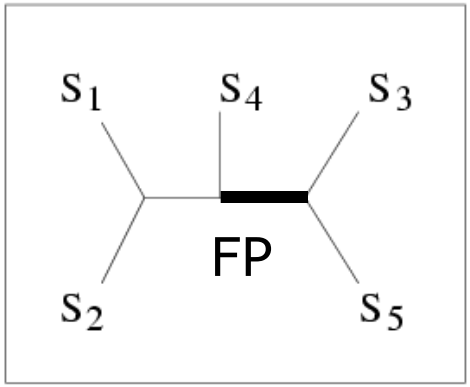
TRUE TREE



DNA SEQUENCES

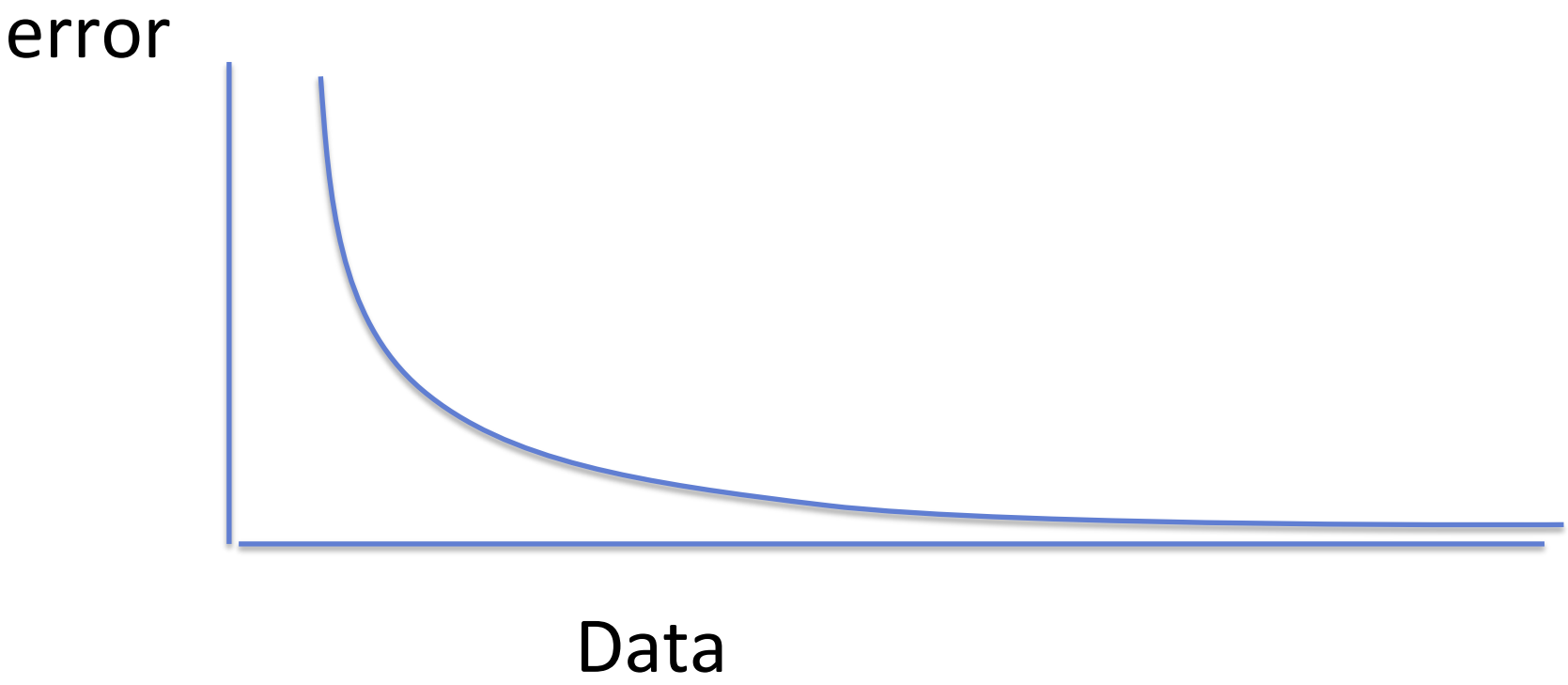
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate



INFERRED TREE

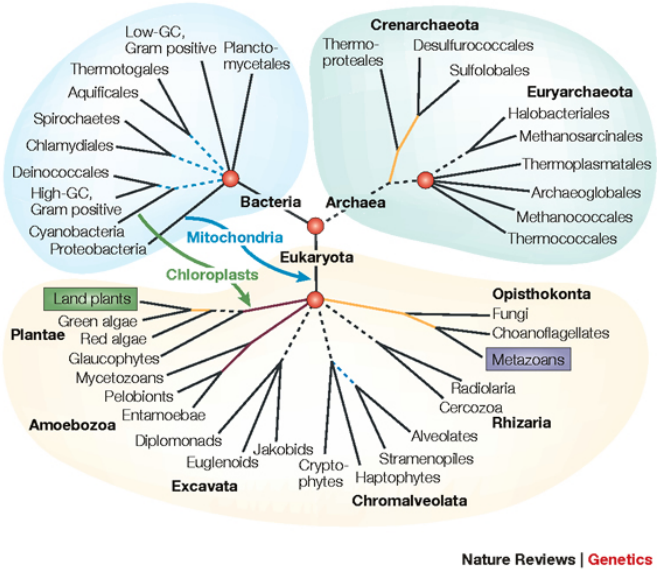
Statistical Consistency



Data are sites in an alignment

Phylogenomics

(Phylogenetic estimation from whole genomes)



Nature Reviews | Genetics

Not all genes present in all species

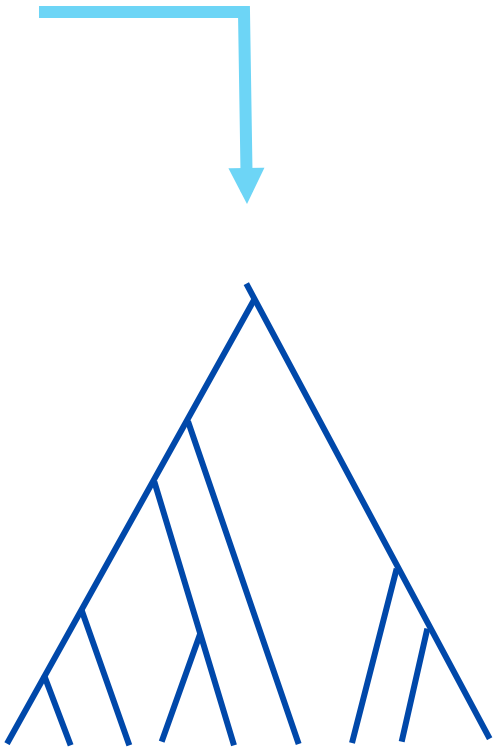
	gene 1
S_1	TCTAATGGAA
S_2	GCTAAGGGAA
S_3	TCTAAGGGAA
S_4	TCTAACGGAA
S_7	TCTAATGGAC
S_8	TATAACGGAA

	gene 2
S_4	GGTAACCCTC
S_5	GCTAAACCTC
S_6	GGTGACCATC
S_7	GCTAAACCTC

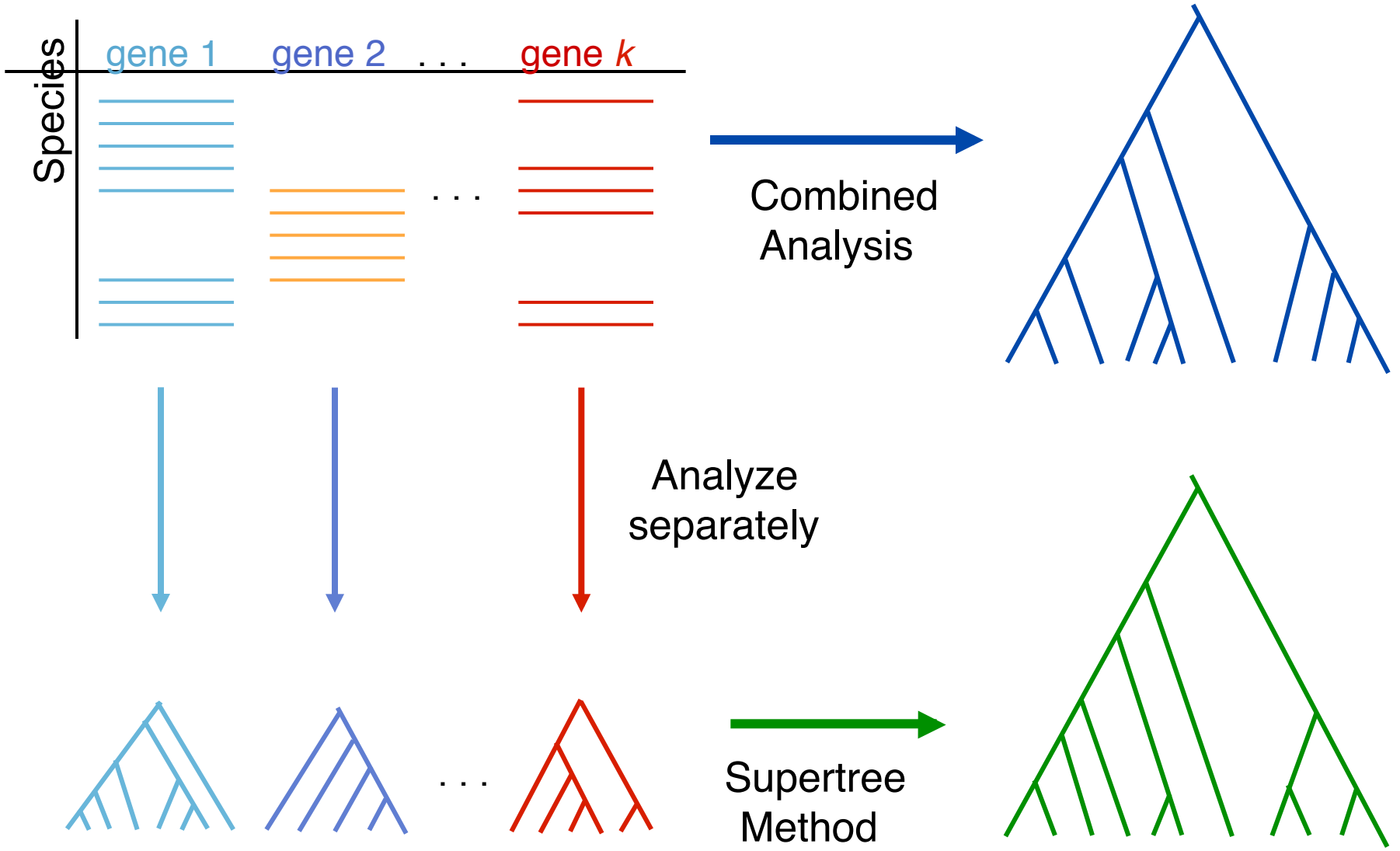
	gene 3
S_1	TATTGATACA
S_3	TCTTGATACC
S_4	TAGTGATGCA
S_7	TAGTGATGCA
S_8	CATTCATACC

Combined analysis

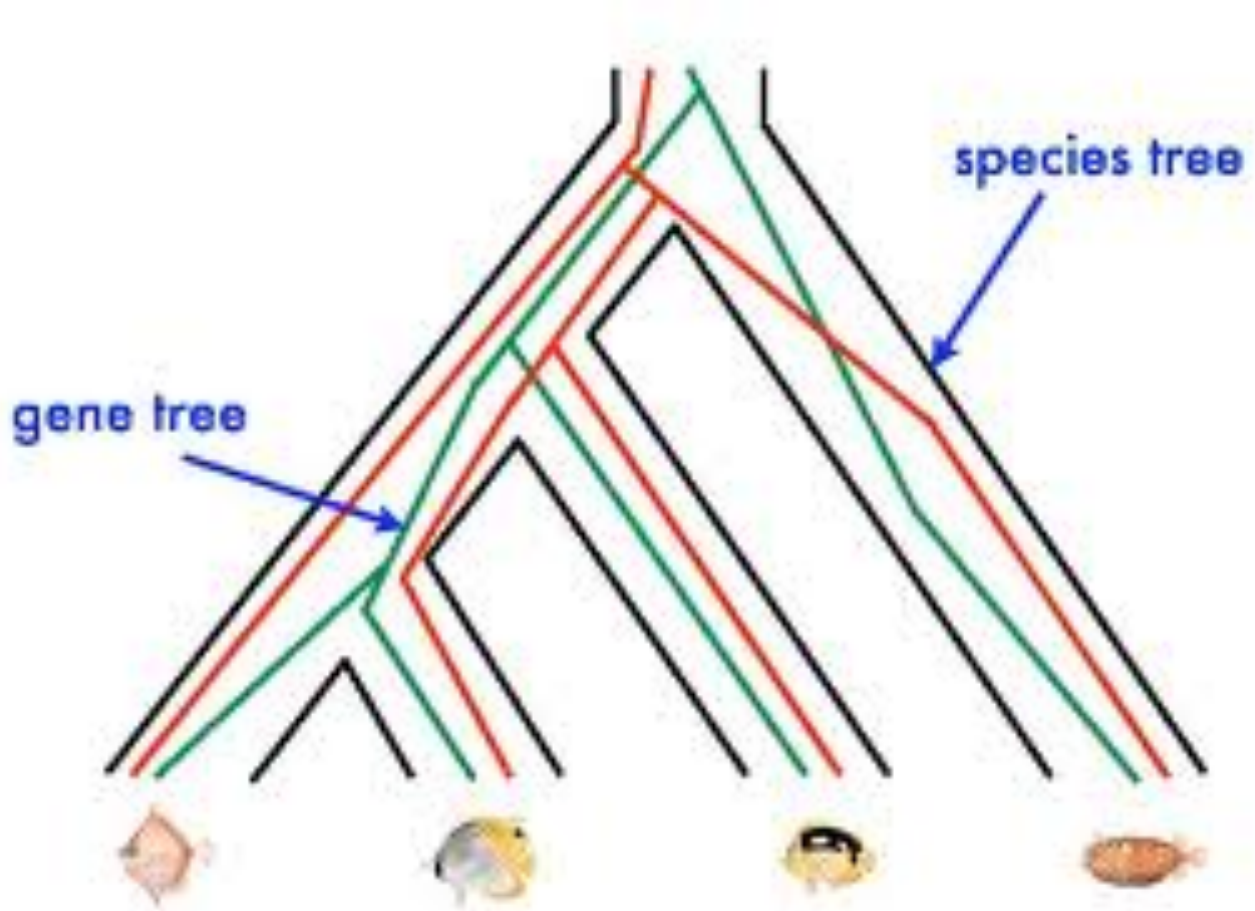
	gene 1	gene 2	gene 3
S_1	TCTAATGGAA	? ? ? ? ? ? ? ? ? ?	TATTGATACA
S_2	GCTAAGGGAA	? ? ? ? ? ? ? ? ? ?	? ? ? ? ? ? ? ? ? ?
S_3	TCTAAGGGAA	? ? ? ? ? ? ? ? ? ?	TCTTGATACC
S_4	TCTAACGGAA	GGTAACCCTC	TAGTGATGCA
S_5	? ? ? ? ? ? ? ? ? ?	GCTAAACCTC	? ? ? ? ? ? ? ? ? ?
S_6	? ? ? ? ? ? ? ? ? ?	GGTGACCATC	? ? ? ? ? ? ? ? ? ?
S_7	TCTAATGGAC	GCTAAACCTC	TAGTGATGCA
S_8	TATAACGGAA	? ? ? ? ? ? ? ? ? ?	CATTCATACC



Two competing approaches



Red gene tree ≠ species tree
(green gene tree okay)

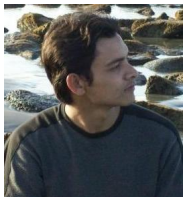


1KP: Thousand Transcriptome Project







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

Nguyen,
UT-Austin

Md. S.Bayzid
UT-Austin

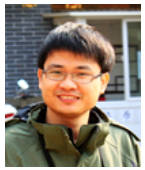
- 1200 plant transcriptomes
- More than 13,000 gene families (most not single copy)
- iPLANT (NSF-funded cooperative)
- Gene sequence alignments and trees computed using SATé

Gene Tree Incongruence

Avian Phylogenomics Project

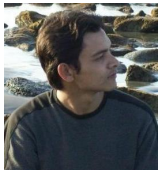

E Jarvis, HHMI


MTP Gilbert, Copenhagen


G Zhang, BGI


T. Warnow UT-Austin


S. Mirarab UT-Austin


Md. S. Bayzid, UT-Austin

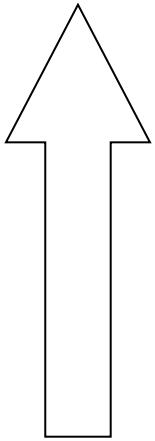
Plus many many other people…

- Approx. 50 species, whole genomes
- 8000+ genes, UCEs
- Gene sequence alignments computed using SATé

Gene Tree Incongruence

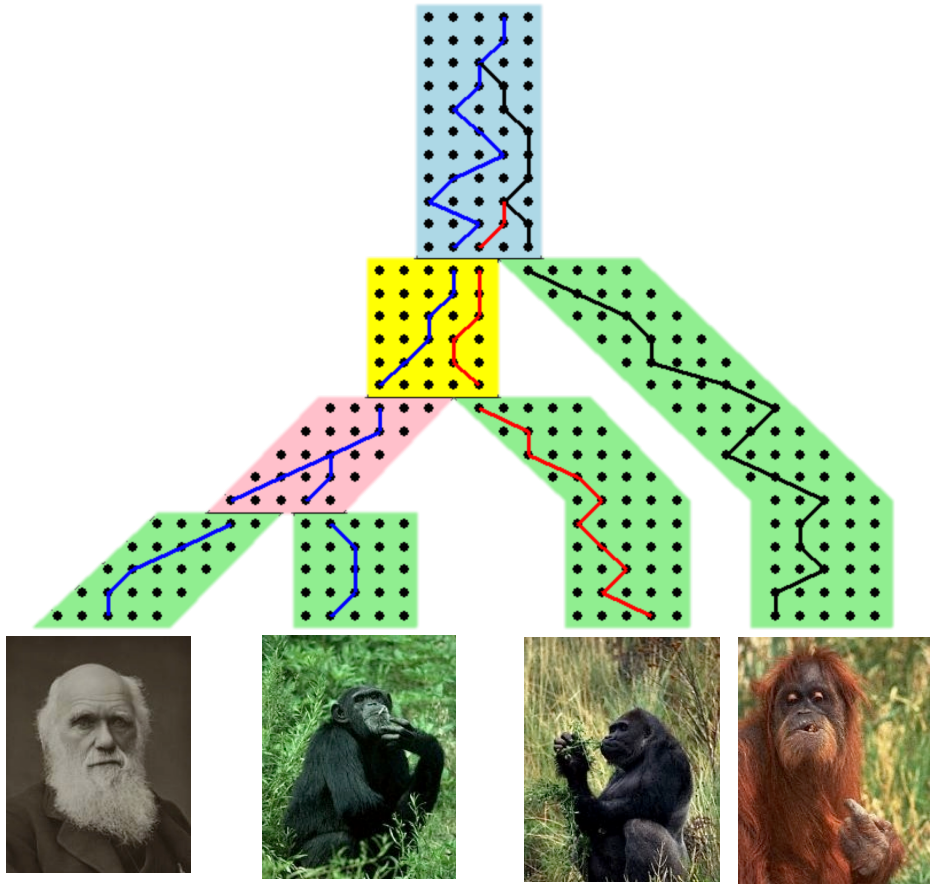
Gene trees inside the species tree (Coalescent Process)

Past



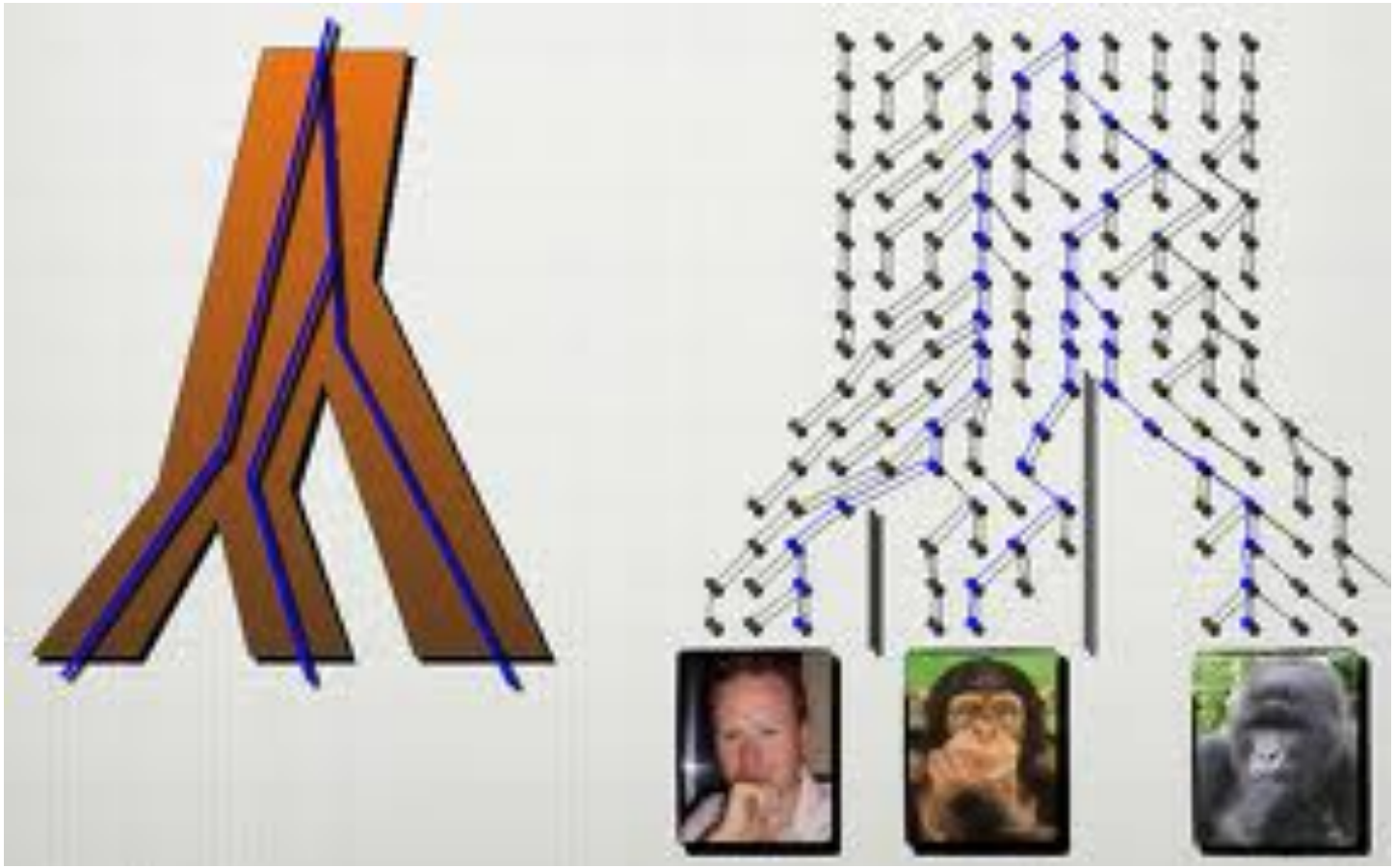
Present

Courtesy James Degnan



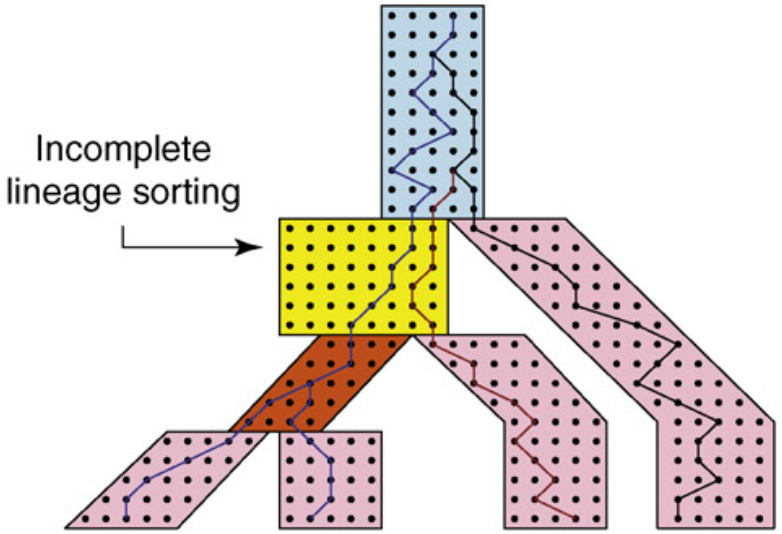
Gorilla and Orangutan are not siblings in the species tree, but they are in the gene tree.

Gene Tree inside a Species Tree



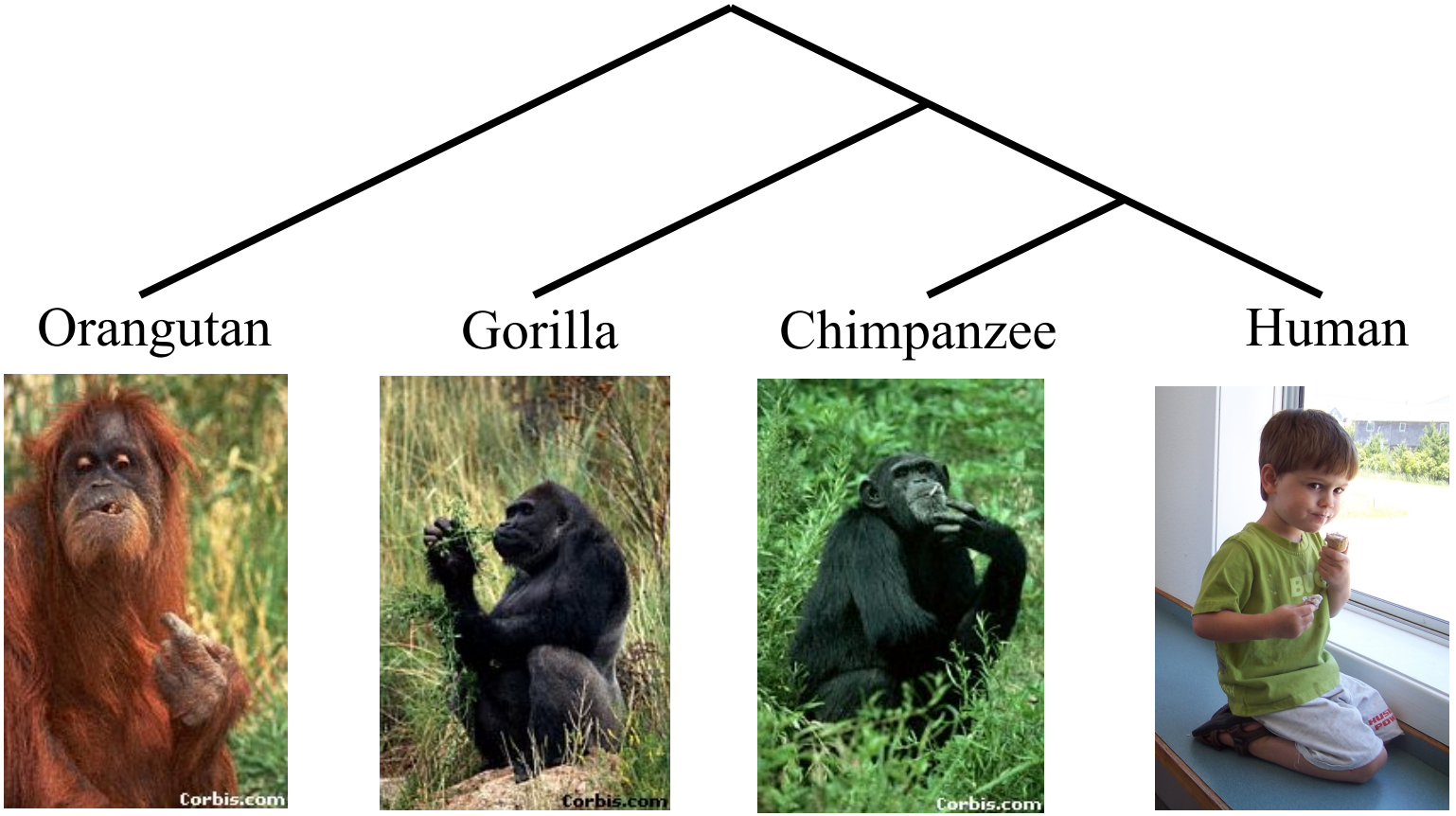
Incomplete Lineage Sorting (ILS)

- Two (or more) lineages fail to coalesce in their first common ancestral population
- Probability of ILS increases for short branches or large population size (wider branches)
- ~960 papers in 2013 include phrase “incomplete lineage sorting”



JH Degnan, NA Rosenberg –
Trends in ecology & evolution, 2009

Species tree estimation: difficult, even for small datasets



From the Tree of the Life Website, University of Arizona

How to compute a species tree?

Techniques:

- Most frequent gene tree?
- Consensus of gene trees?
- Other?

Anomaly Zone

- Under the multi-species coalescent model, the most probable gene tree may not be the true species tree (the “anomaly zone”) – Degnan & Rosenberg 2006, 2009.

Anomaly Zone

- Under the multi-species coalescent model, the most probable gene tree may not be the true species tree (the “anomaly zone”) – Degnan & Rosenberg 2006, 2009.
- Hence, selecting the most frequent gene tree is not a statistically consistent technique.

Anomaly Zone

- Under the multi-species coalescent model, the most probable gene tree may not be the true species tree (the “anomaly zone”) – Degnan & Rosenberg 2006, 2009.
- Hence, selecting the most frequent gene tree is not a statistically consistent technique.
- However, there are no anomalous rooted 3-taxon trees or unrooted 4-taxon trees – Allman et al. 2011, Degnan 2014.

Anomaly Zone

- Hence, for every 3 species, the most frequent rooted gene tree will be the true rooted species tree with high probability.
- (The same thing is true for unrooted 4-leaf gene trees.)

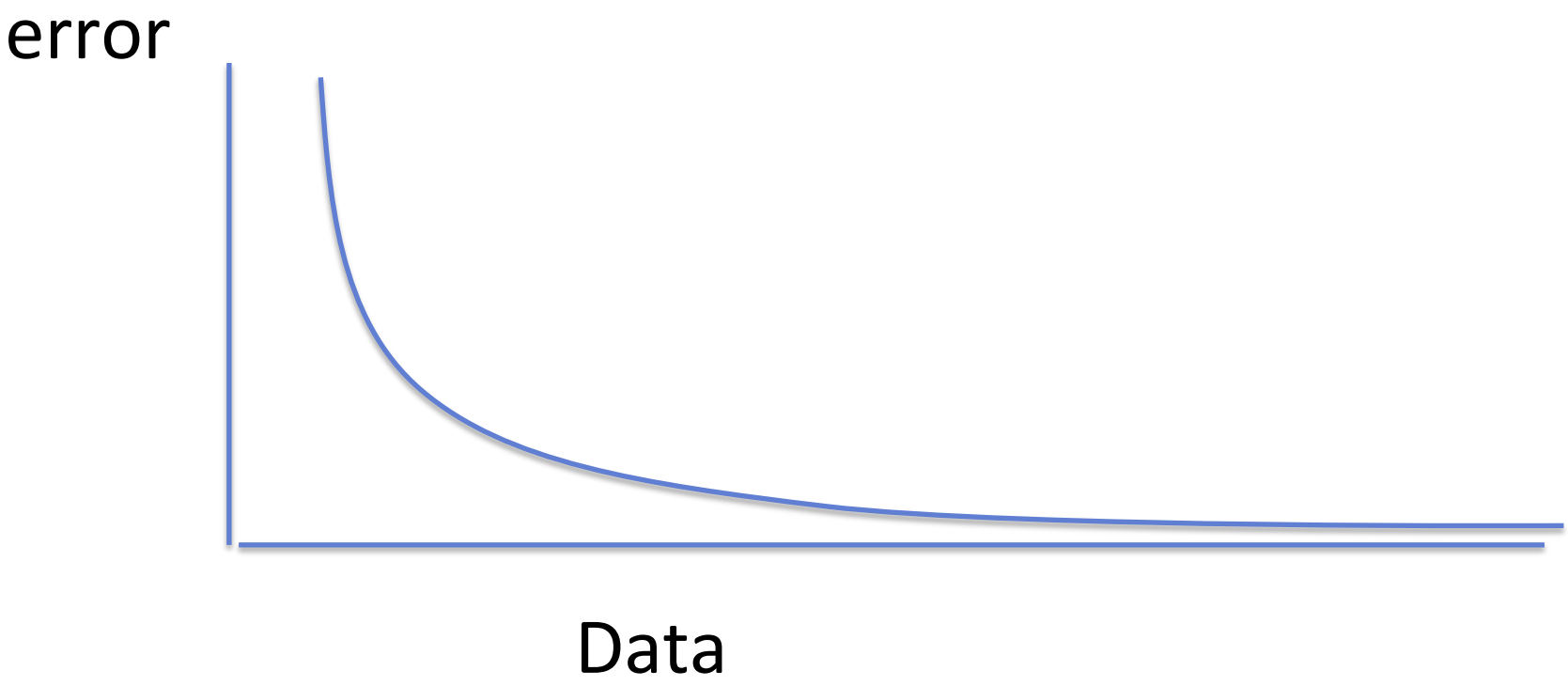
Statistically consistent method

- Given set of rooted gene trees, for every three species:
  - Compute the induced triplet trees in each gene tree
  - Find dominant triplet tree.
- If the triplet trees are compatible, it is easy to compute the tree they all agree with.
- Otherwise, apply a heuristic to find a tree that satisfies the largest number of dominant triplet trees (NP-hard).

Simple algorithm to construct species tree from unrooted gene trees

- Given set of gene trees, for every four species:
  - Compute the induced quartet trees in each gene tree
  - Find dominant quartet tree
- If the quartet trees are compatible, it is easy to compute the tree they all agree with.
- Otherwise, apply a heuristic to find a tree that satisfies most of the dominant quartet trees.

Statistical Consistency



Data are gene trees, presumed to be randomly sampled <u>true gene trees.</u>

Some statistically consistent methods

- Rooted gene trees:
  - Simple triplet-based methods for rooted gene trees
  - MP-EST (Liu et al. 2010): maximum likelihood estimation of rooted species tree
- Unrooted gene trees:
  - Simple quartet-based methods for unrooted gene trees
  - BUCKy-pop (Ané and Larget 2010): quartet-based Bayesian species tree estimation– ?
- Sequence alignments
  - *BEAST (Heled and Drummond): co-estimates gene trees and species tree

Some statistically consistent methods

- Rooted gene trees:
  - Simple triplet-based methods for rooted gene trees
  - MP-EST (Liu et al. 2010): maximum likelihood estimation of rooted species tree
- Unrooted gene trees:
  - Simple quartet-based methods for unrooted gene trees
  - BUCKy-pop (Ané and Larget 2010): quartet-based Bayesian species tree estimation– ?
- Sequence alignments
  - *BEAST (Heled and Drummond): co-estimates gene trees and species tree

Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model

Sen Song[a,1], Liang Liu[b,1], Scott V. Edwards[c], and Shaoyuan Wu[b,d,2]

[a]Department of Biomedical Engineering, School of Medicine, Tsinghua University, Beijing 100084, China; [d]Institute of Paleontology, Shenyang Normal University, Shenyang 110034, China; [b]Department of Statistics and Institute of Bioinformatics, University of Georgia, Athens, GA 30606; and [c]Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138

Edited by David M. Hillis, University of Texas at Austin, Austin, TX, and approved August 2, 2012 (received for review July 11, 2012)

The reconstruction of the Tree of Life has relied almost entirely on concatenation methods, which do not accommodate gene tree heterogeneity, a property that simulations and theory have identified as a likely cause of incongruent phylogenies. However, this incongruence has not yet been demonstrated in empirical studies. Several key relationships among eutherian mammals remain controversial and conflicting among previous studies, including the root of eutherian tree and the relationships within Euarchontoglires and Laurasiatheria. Both Bayesian and maximum-likelihood analysis of genome-wide data of 447 nuclear genes from 37 species show that concatenation methods indeed yield strong incongruence in the phylogeny of eutherian mammals, as revealed by subsampling analyses of loci and taxa, which produced strongly conflicting topologies. In contrast, the coalescent methods, which accommodate gene tree heterogeneity, yield a phylogeny that is robust to variable gene and taxon sampling and is congruent with geographic data. The data also demonstrate that incomplete lineage sorting, a major source of gene tree heterogeneity, is relevant to deep-level phylogenies, such as those among eutherian mammals. Our results firmly place the eutherian root between Atlantogenata and Boreoeutheria and support ungulate polyphyly and a sister-group relationship between Scandentia and Primates. This study demonstrates that the incongruence introduced by concatenation methods is a major cause of long-standing uncertainty in the phylogeny of eutherian mammals, and the same may apply to other clades. Our analyses suggest that such incongruence can be resolved using phylogenomic data and coalescent methods that deal explicitly with gene tree heterogeneity.

gene tree heterogeneity | incomplete lineage sorting | multispecies coalescent model | phylogenetic incongruence

To date, phylogenetic studies using DNA sequence data have been based almost entirely on concatenation methods. Concatenation methods infer phylogenies from multilocus sequences that are combined to form a single supermatrix (1), based on the assumption that all genes have the same or similar phylogenies (1, 2). However, empirical studies have shown widespread presence of gene tree heterogeneity within mammals and other clades (3, 4). When a high level of gene tree heterogeneity occurs in multilocus sequence data, theory and simulations have predicted that concatenation methods can yield misleading results (5, 6). By contrast, more recently developed coalescence-based methods estimate a species phylogeny from a collection of gene trees, an approach that allows different genes to have different topologies (4, 7–10). Simulations and theory have shown that coalescent methods can produce accurate phylogenies from multilocus sequence data that are subject to incomplete lineage sorting (ILS), a major cause of gene tree heterogeneity (4, 7–10). However, the superior performance of coalescent methods relative to concatenation methods in the face of substantial gene tree heterogeneity remains to be demonstrated in empirical studies.

Resolving the phylogeny of eutherian mammals has been challenging due to conflicting results from previous studies (11–20). In the past decade, the division of eutherian mammals into four superorders—Euarchontoglires, Laurasiatheria, Afrotheria, and Xenarthra—has been well supported (11–20). However, some key elements of eutherian mammal relationships, including the root of the eutherian tree and the interordinal relationships within Euarchontoglires and Laurasiatheria, remain unresolved or unstable (20). Resolving these incongruences is crucial not only for understanding the evolutionary history and dynamics of Eutheria, but also for revealing the source of contradictions on eutherian phylogeny in previous studies. Using a phylogenetic, DNA-based analysis of eutherian mammal relationships as a case study, we empirically demonstrate that concatenation methods can lead to phylogenetic results that are inherently incongruent, in that different subsamples of the same data set tend to produce strongly divergent topologies. Analyzing and subsampling the same data using coalescent methods yield more consistent results, and the resulting phylogeny suggests possible resolutions to persistent controversies regarding the position of the root of Eutheria and key relationships within Laurasiatheria and Euarchontoglires.

Results

Conflict Between Concatenation and Coalescent Phylogenetic Analyses. We analyzed sequence data from 447 nuclear genes from 33 eutherian species representing 16 of 18 eutherian orders and four outgroups including two marsupials, one monotreme, and chicken. The 447 orthologous genes in the data are distributed across all 22 autosomes and the X chromosome in the human genome, allowing us to access the phylogenetic utility of different parts across the genome.

Our analyses used two recently developed coalescent methods: the Maximum Pseudolikelihood Estimation of the Species Tree (MP-EST) method (8) and the Species Tree Estimation using Average Ranks of coalescence (STAR) method, used here with the neighbor-joining algorithm (9). MP-EST uses the frequencies of gene trees of triplets of taxa to estimate the topology and branch lengths (in coalescent units) of the overall species tree (8), whereas STAR computes the topological distances among pairs of taxa as the average of the ranks (number of nodes toward the root node) of those taxon pairs across nodes in the collected gene trees (9). MP-EST and STAR are partially parametric methods that reconstruct species phylogenies using

Author contributions: S.W. designed research; S.S., L.L., S.V.E., and S.W. performed research; S.S., L.L., S.V.E., and S.W. analyzed data; and S.V.E. and S.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]S.S. and L.L. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: shaoyuan5@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1211733109/-/DCSupplemental.

Song et al., PNAS 2012

Introduced statistically consistent method, MP-EST

Used MP-EST to analyze a mammalian dataset with 37 species and 447 genes

Song et al. PNAS 2012

- "This study demonstrates that the incongruence introduced by concatenation methods is a major cause of longstanding uncertainty in the phylogeny of eutherian mammals, and the same may apply to other clades. Our analyses suggest that such incongruence can be resolved using phylogenomic data and coalescent methods that deal explicitly with gene tree heterogeneity."

Springer and Gatesy (TPS 2014)

- "The poor performance of coalescence methods [5–8] presumably reflects their incorrect assumption that all conflict among gene trees is attributable to deep coalescence, whereas a multitude of other problems (long branches, mutational saturation, weak phylogenetic signal, model misspecification, poor taxon sampling) negatively impact reconstruction of accurate gene trees and provide more cogent explanations for incongruence [6,7]."

- "Shortcut coalescence methods are not a reliable remedy for persistent phylogenetic problems that extend back to the Precambrian. "

The Debate: Concatenation vs. Coalescent Estimation

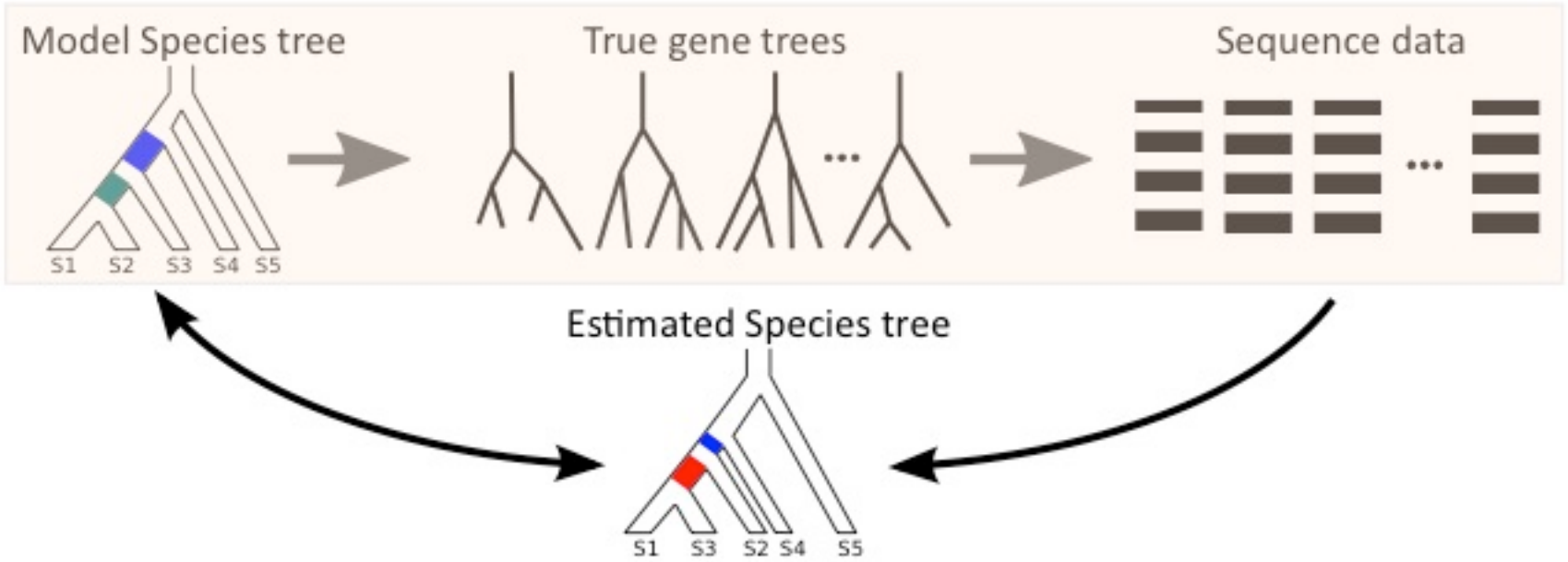
- In favor of coalescent-based estimation
  - Statistical consistency guarantees
  - Addresses gene tree incongruence resulting from ILS
  - Some evidence that concatenation can be positively misleading

- In favor of concatenation
  - Reasonable results on data
  - High bootstrap support
  - Summary methods (that combine gene trees) can have poor support or miss well-established clades entirely
  - Some methods (such as *BEAST) are computationally too intensive to use

Is Concatenation Evil?

- Joseph Heled:
  - YES
- John Gatesy
  - No

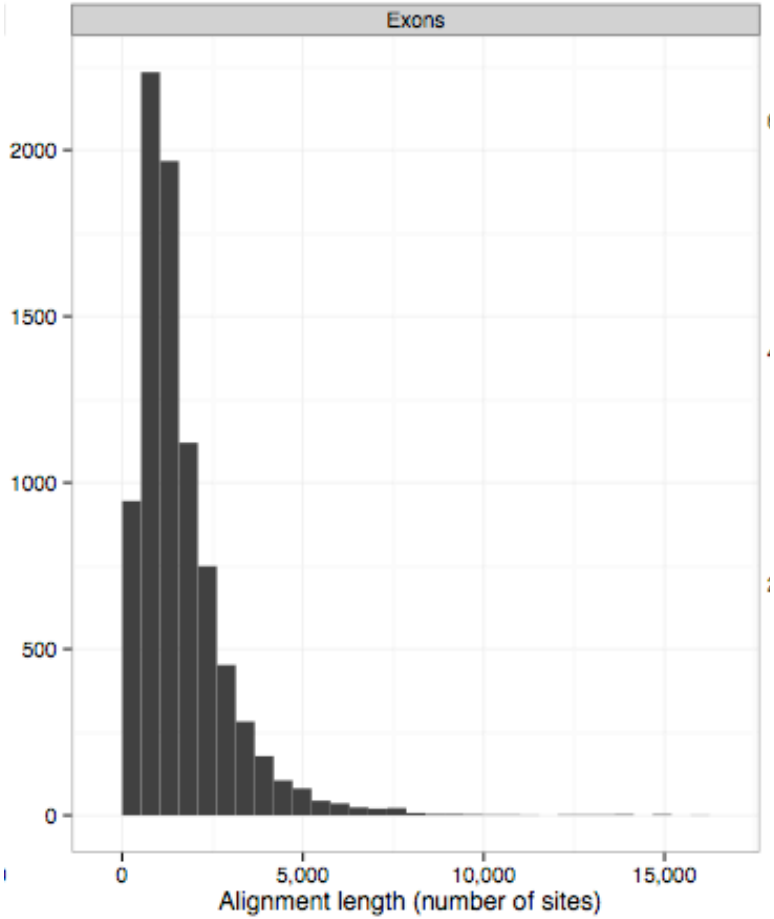
Evaluating methods using simulation



- Summary method
  - MP-EST (statistically consistent, increasingly popular)
- Concatenation
  - RAxML (among many good tools for ML)

Data quality: the flip side of phylogenomics

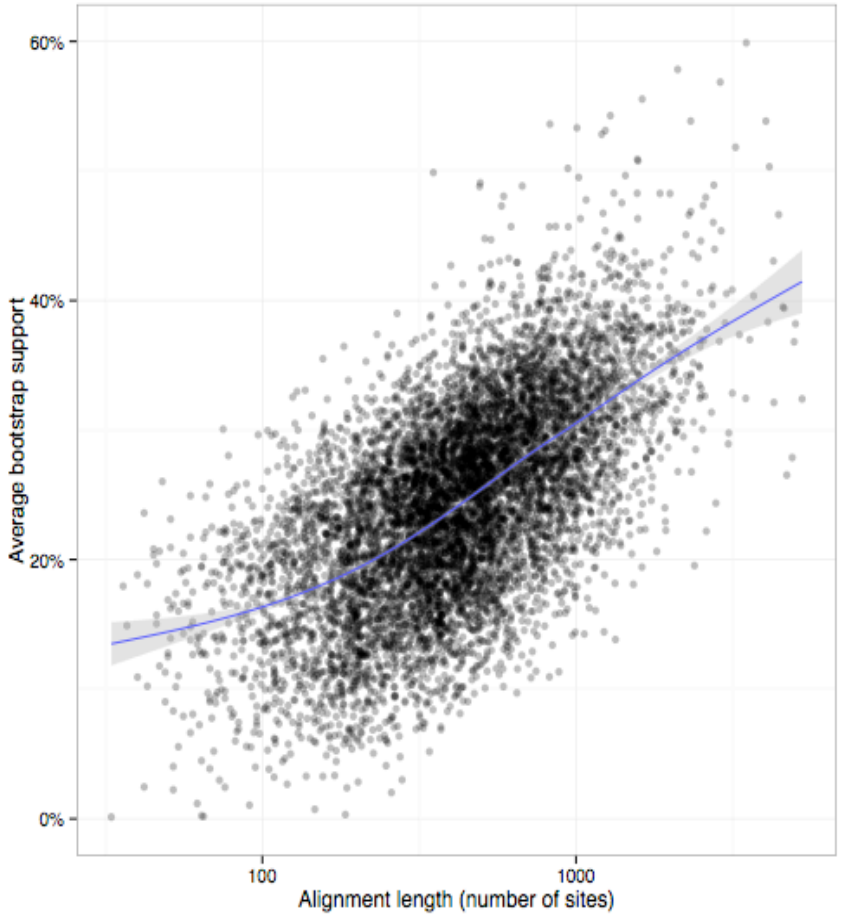
- As more genes are sampled, many of them have low quality
  - Short sequences
  - Uninformative sites



8,500 exons from the Avian project

Poorly resolved gene trees

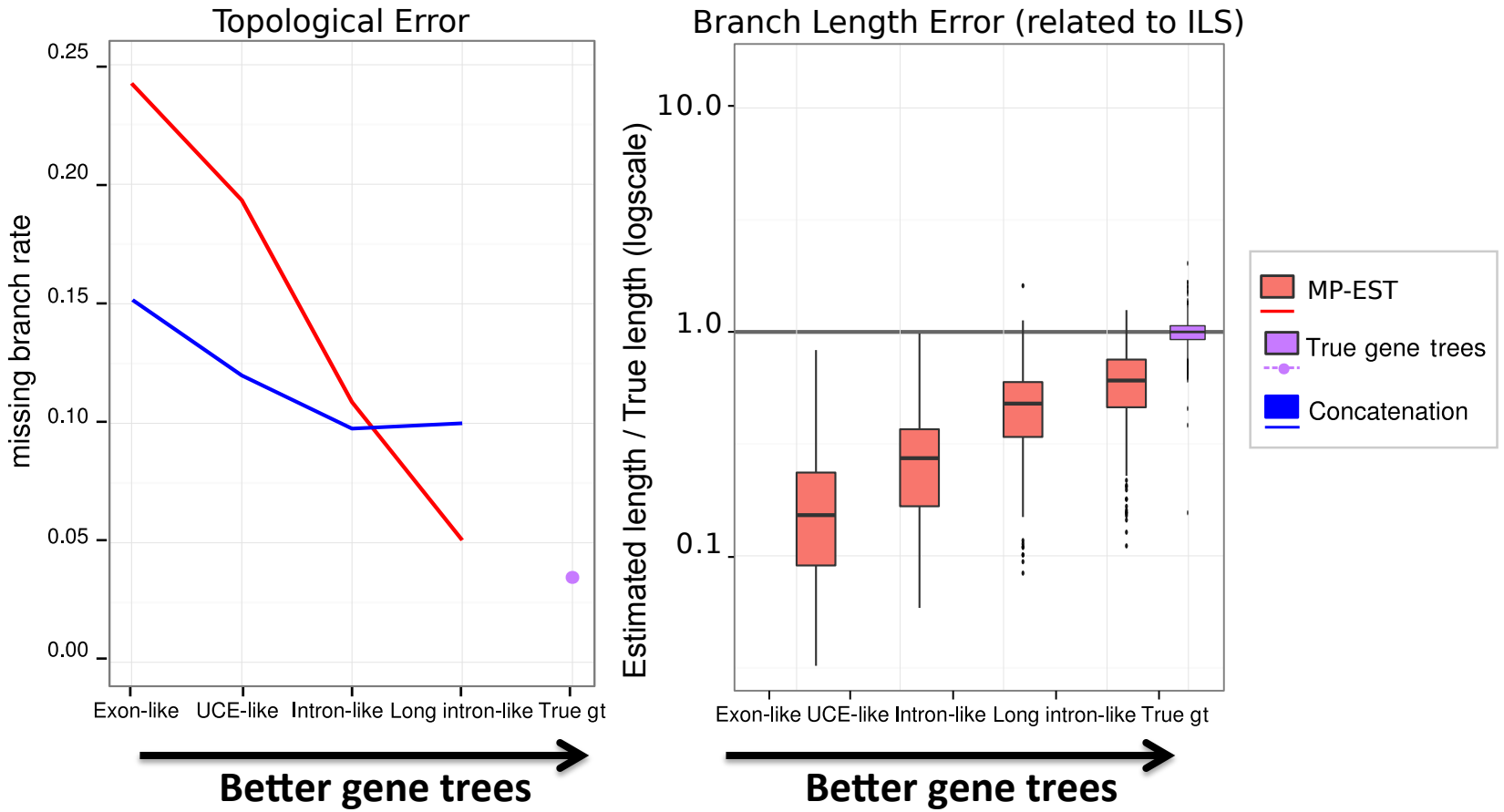
- As more genes are sampled, many of them have low quality
- which leads to gene trees with low support (and hence high error)



8,500 exons from the Avian project

Avian-like simulation results

- Avian-like simulation; 1000 genes, 48 taxa, high levels of ILS



Is Concatenation Evil?

- Joseph Heled:
  - YES
- John Gatesy
  - No

Objective

- Fast, and able to analyze genome-scale data (thousands of loci) quickly
- Highly accurate
- Statistically consistent
- Convince Gatesy that coalescent-based estimation is okay

ASTRAL

- ASTRAL = Accurate Species TRee Algorithm
- Authors: S. Mirarab, R. Reaz, Md. S. Bayzid, T. Zimmerman, S. Swenson, and T. Warnow
- To appear, Bioinformatics and ECCB 2014
- Tutorial on using ASTRAL at Evolution 2014
- Open source and freely available

Simple algorithm

- Given set of gene trees, for every four species:
  - Compute the induced quartet trees in each gene tree
  - Find which quartet tree is dominant
- If the quartet trees are compatible, it is easy to compute the tree they all agree with.
- Otherwise, apply a heuristic to find a tree that satisfies most of the dominant quartet trees.

Simple algorithm

- Given set of gene trees, for every four species:
  - Compute the induced quartet trees in each gene tree
  - Find which quartet tree is dominant
- If the quartet trees are compatible, it is easy to compute the tree they all agree with.
- Otherwise, apply a heuristic to find a tree that satisfies most of the dominant quartet trees.

Problem: loss of information about confidence/support in the quartet tree

Median Tree

- Define the cost of a species tree T on set S with respect to a set of unrooted gene trees $\{t_1, t_2, \ldots t_k\}$ on set S by:

  – $\text{Cost}(T,S) = d(T,t_1) + d(T,t_2) + \ldots + d(T,t_k)$

  where $d(T,t_i)$ is the number of quartets of taxa that T and t_i have different topologies.

- The optimization problem is to find a tree T of minimum cost with respect to the input set of unrooted gene trees.

Statistical Consistency

- Theorem: Let $\{t_1, t_2, \ldots t_k\}$ be a set of unrooted gene trees on set S. Then the median tree is a statistically consistent estimator of the unrooted species tree, under the multi-species coalescent.

Statistical Consistency

- Theorem: Let $\{t_1, t_2, \ldots t_k\}$ be a set of unrooted gene trees on set S. Then the median tree is a statistically consistent estimator of the unrooted species tree, under the multi-species coalescent.
- Proof: Given a large enough number of genes, then with high probability the most frequent gene tree on any four species is the true species tree. When this holds, then the true species tree has the minimum cost, because it agrees with the largest number of quartet trees.

Computing the median tree

- This is likely to be an NP-hard problem, so we don't try to solve it exactly. Instead, we solve a constrained version:
  - Input: set of unrooted gene trees $\{t_1, t_2, \ldots t_k\}$ on set S, and set X of bipartitions on S.
  - Output: tree T that has minimum cost, subject to T drawing its bipartitions from X.

Default ASTRAL

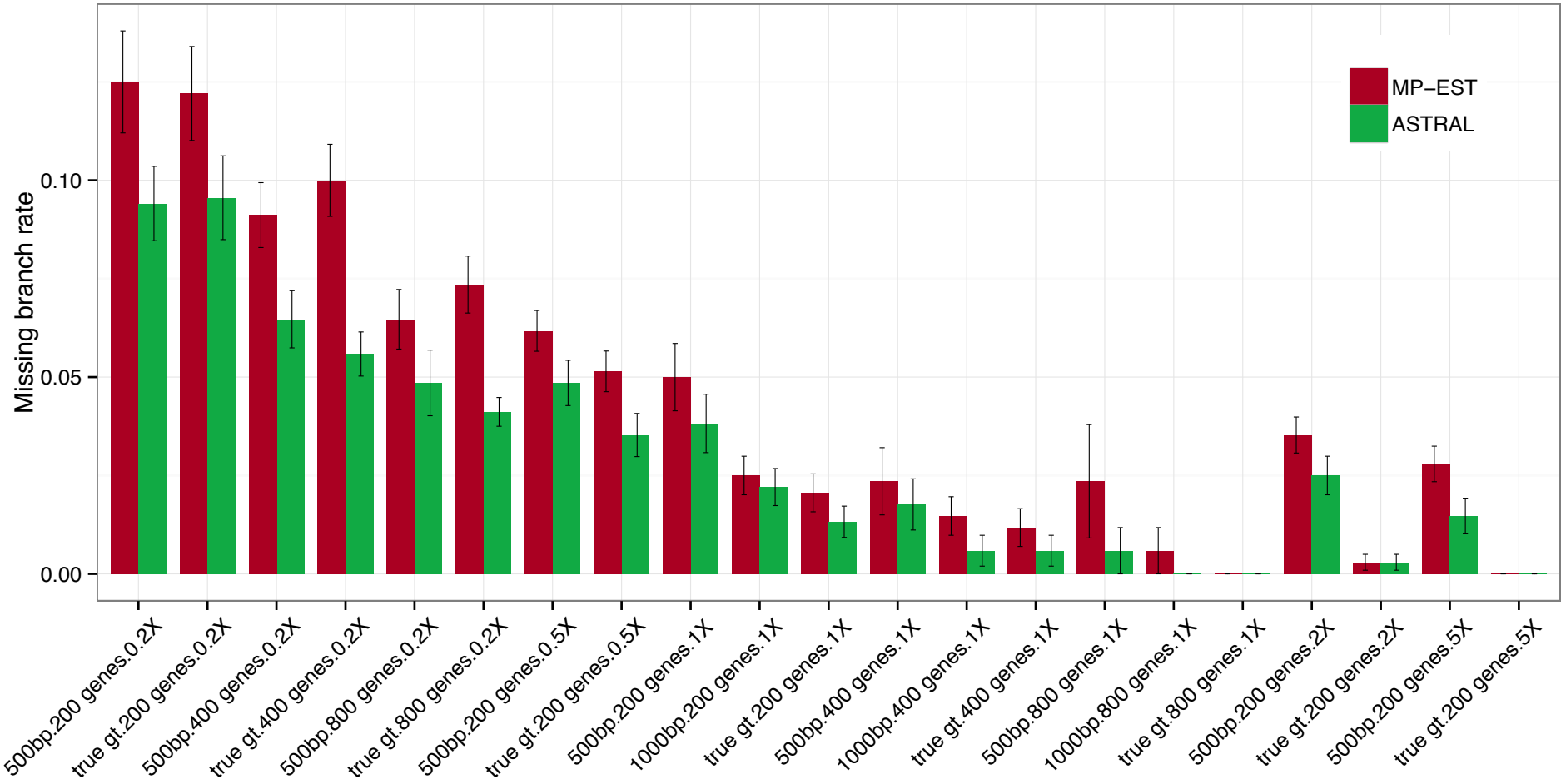
- The default setting for ASTRAL sets X to be the bipartitions in the input set of gene trees.
- Theorem: Default ASTRAL is statistically consistent under the multi-species coalescent model.
- Proof: given a large enough number of gene trees, some gene tree will have the same topology as the species tree with high probability.

ASTRAL running time

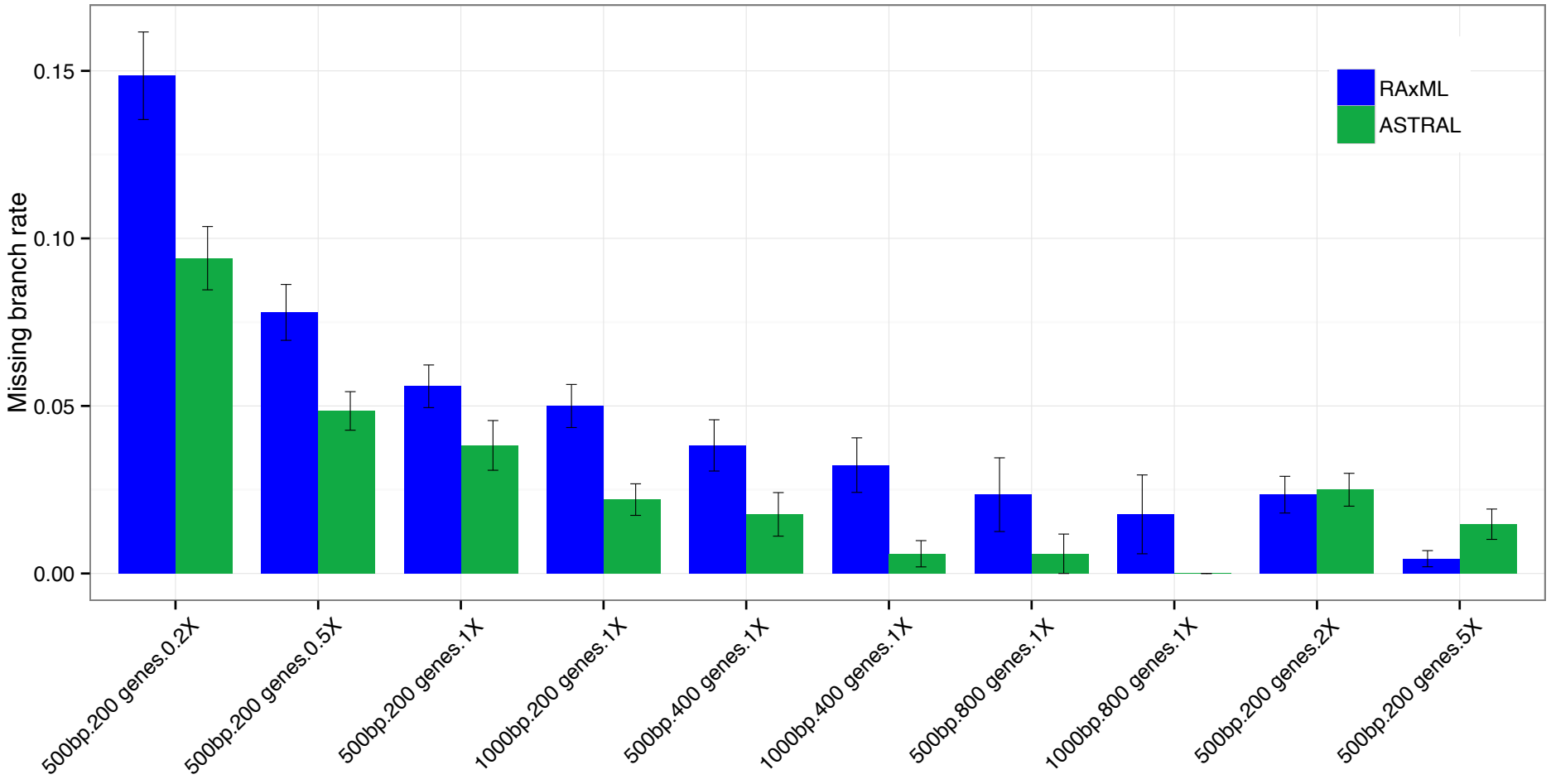
- ASTRAL has $O(nk|X|^2)$ running time for k genes of n taxa and bipartitions from set X
  - $O(n^3k^3)$ if X is the set of bipartitions from gene trees

- Runs in ~3 minutes for 800 genes and 103 taxa
  - In contrast, MP-EST takes around a day and does not converge (multiple searches result in widely different trees)

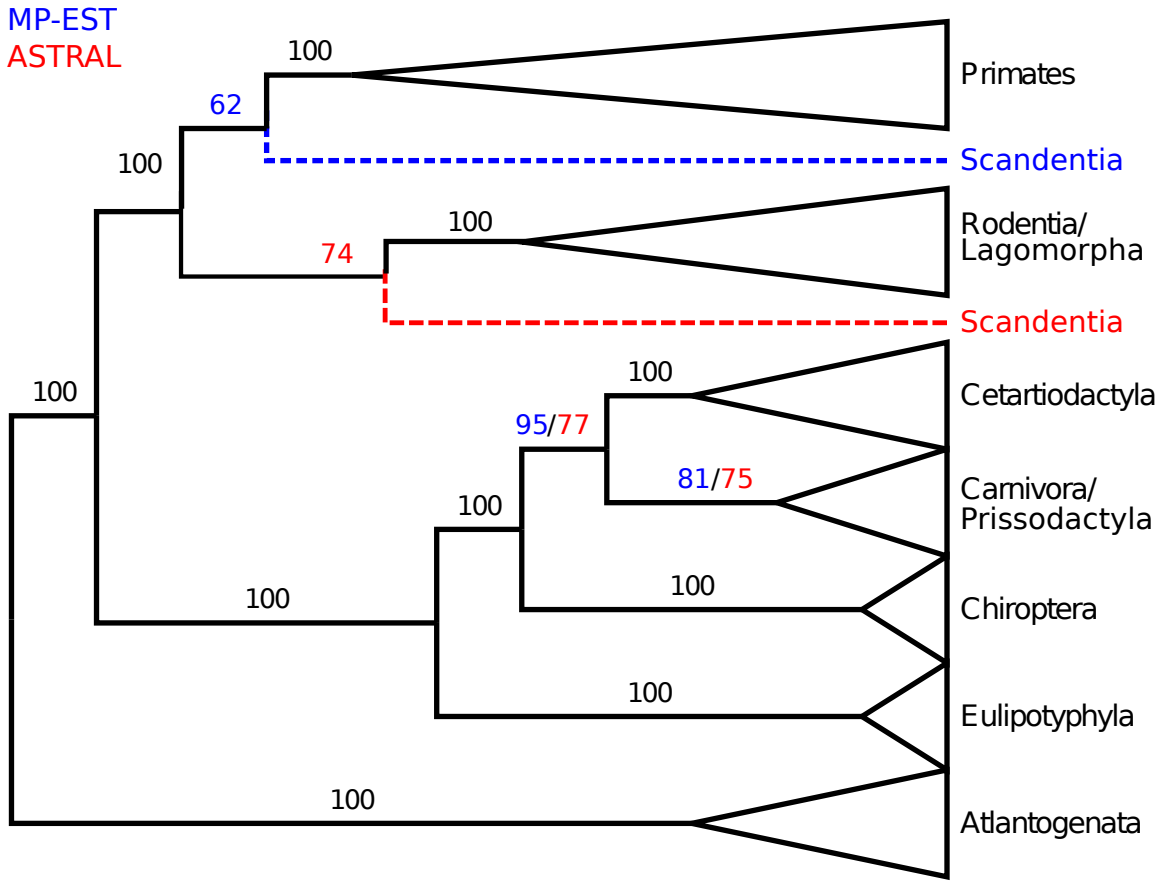
ASTRAL vs. MP-EST (mammalian simulation)



ASTRAL vs. Concatenation (mammalian simulation)



Analyses of the Song et al. Mammalian dataset



The placement of Scandentia (Tree Shrew) is controversial.
The ASTRAL analysis agrees with maximum likelihood concatenation analysis of this dataset.

Letter

CellPress

Land plant origins and coalescence confusion

Mark S. Springer and John Gatesy

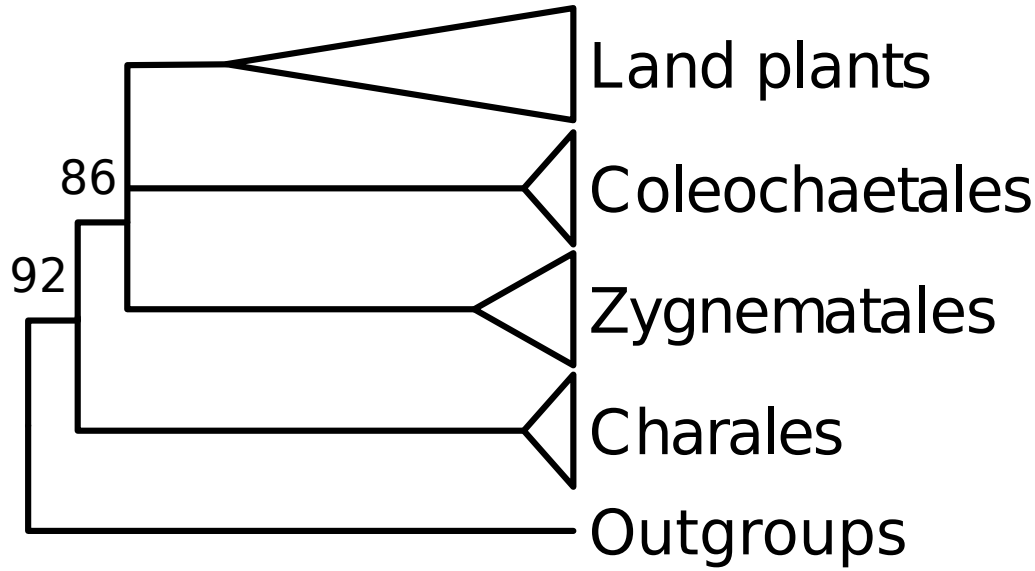
Department of Biology, University of California, Riverside, CA 92521, USA

Coalescence methods have emerged as an alternative to concatenation methods for reconstructing species trees [1,2]. Zhong *et al.* [3] advocated the coalescence approach for resolving early branching events in plant phylogeny. We show that different coalescence methods yield discordant results and call attention to fundamental problems with the application of coalescence to deep phylogenetic questions such as the origin of land plants.

conclusions regarding land plant origins and the utility of coalescence methods for deep phylogenetic problems.

MP-EST and STAR are both statistically consistent methods when their underlying assumptions are upheld and in these instances may yield more accurate species trees than concatenation [1,2]. However, theoretical guarantees are empty when assumptions are violated and should be trumped by empirical performance. Zhong

ASTRAL Analysis of Zhong et al. dataset



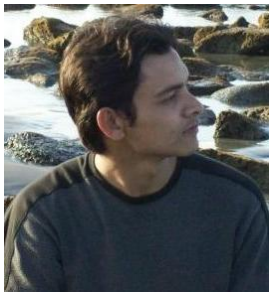
MP-EST analysis supported Zygnematales as the sister to Land Plants.
The ASTRAL analysis leaves the sister to Land Plants open: it produced one low support branch (18% BS); collapsing that branch rules out Charales as the possible sister to land plants.

Summary

- ASTRAL is statistically consistent under the multi-species coalescent model.
- On the datasets we studied, ASTRAL is more than other summary methods, and typically more accurate than concatenation (except under low levels of ILS). It is also very fast and can analyze very large datasets.
- ASTRAL analyses of biological datasets are often closer to concatenation analyses than MP-EST analyses of the same datasets.
- Lots of low hanging fruit.

Warnow Laboratory





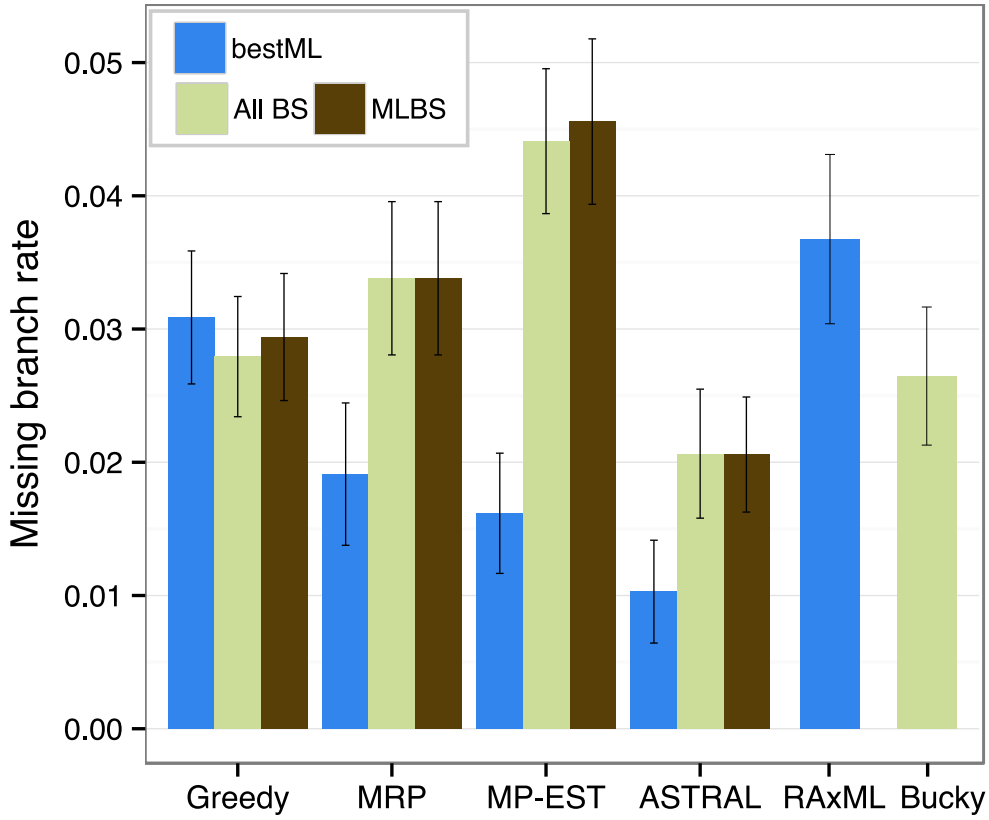
PhD students: Siavash Mirarab, Nam Nguyen, and Md. S. Bayzid

Undergrad: Keerthana Kumar

Lab Website: http://www.cs.utexas.edu/users/phylo

Funding: Guggenheim Foundation, Packard Foundation, NSF, Microsoft Research New England, David Bruton Jr. Centennial Professorship, and TACC (Texas Advanced Computing Center). HHMI graduate fellowship to Siavash Mirarab and Fulbright graduate fellowship to Md. S. Bayzid.

Impact of using bootstrap gene trees instead of best ML gene trees



Mammalian simulated dataset with 400 genes, 1X ILS level