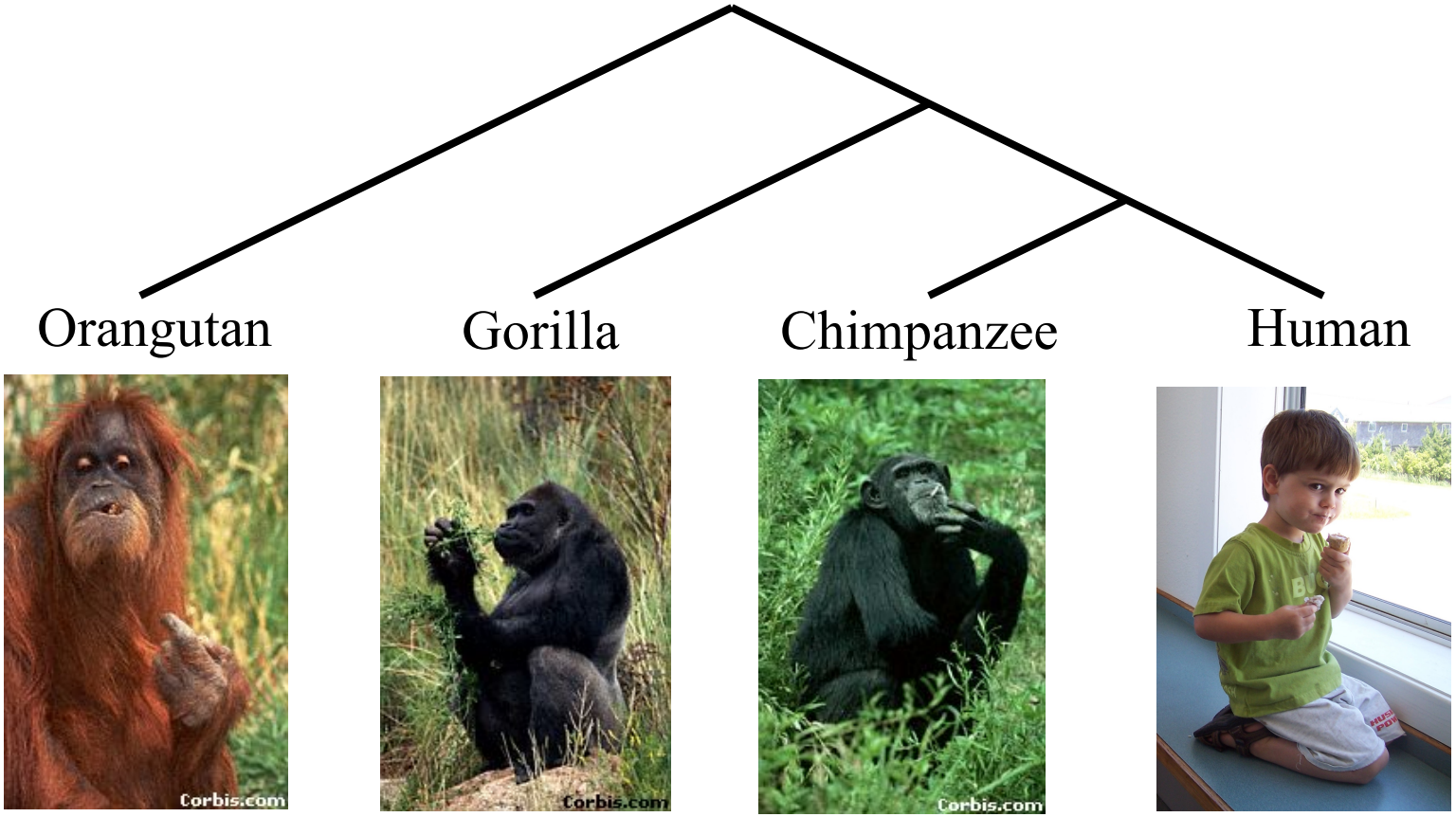
The Mathematics of Estimating the Tree of Life

Tandy Warnow
The University of Illinois

Phylogeny
(evolutionary tree)

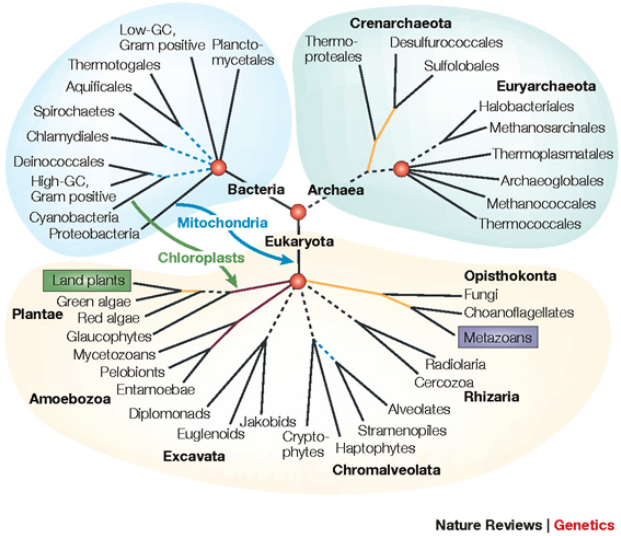


From the Tree of the Life Website,
University of Arizona

- "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

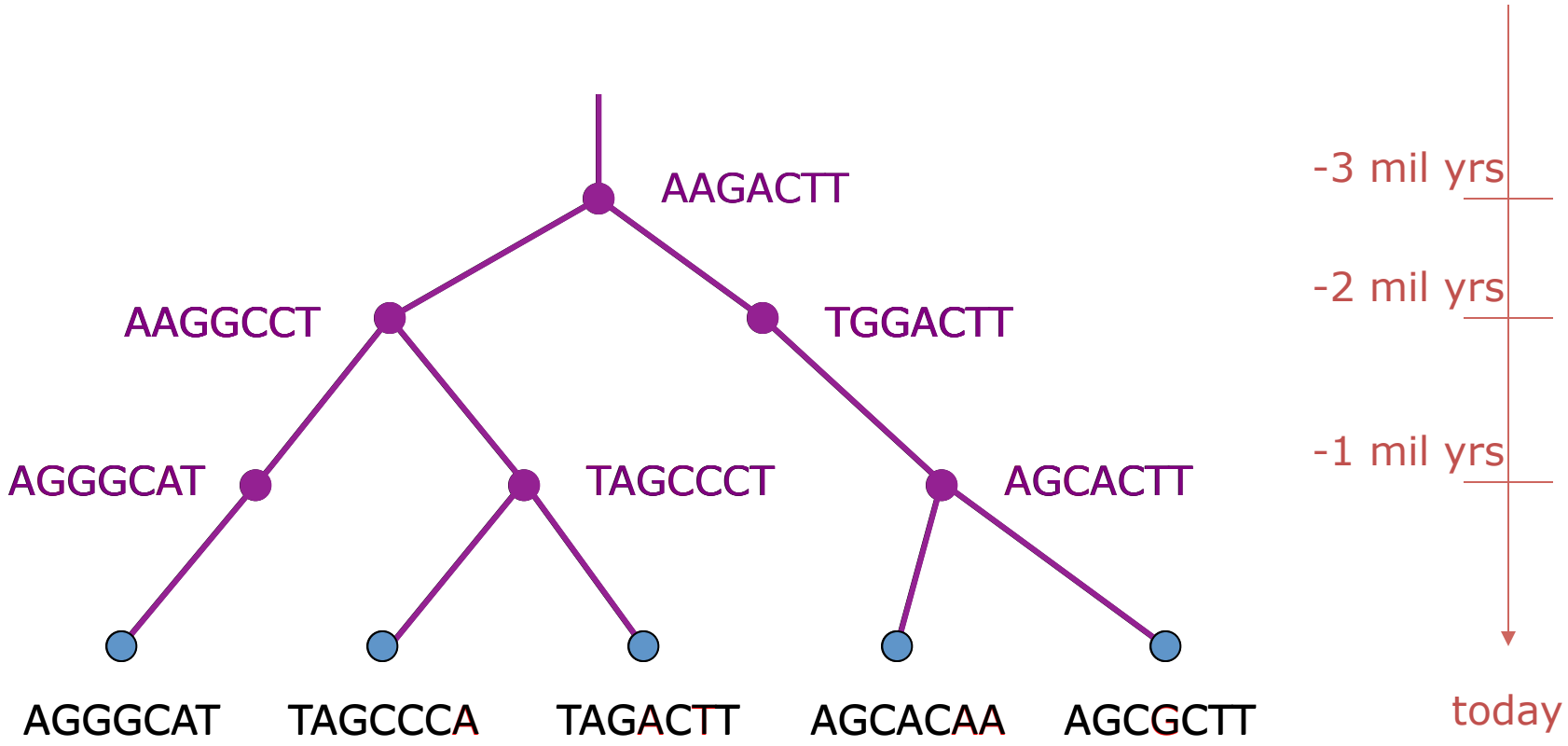
- *"…… nothing in evolution makes sense except in the light of phylogeny ..."*
  - *Society of Systematic Biologists, http://systbio.org/teachevolution.html*

Computing the Tree of Life



- 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

DNA Sequence Evolution (Idealized)



Markov Model of Site Evolution

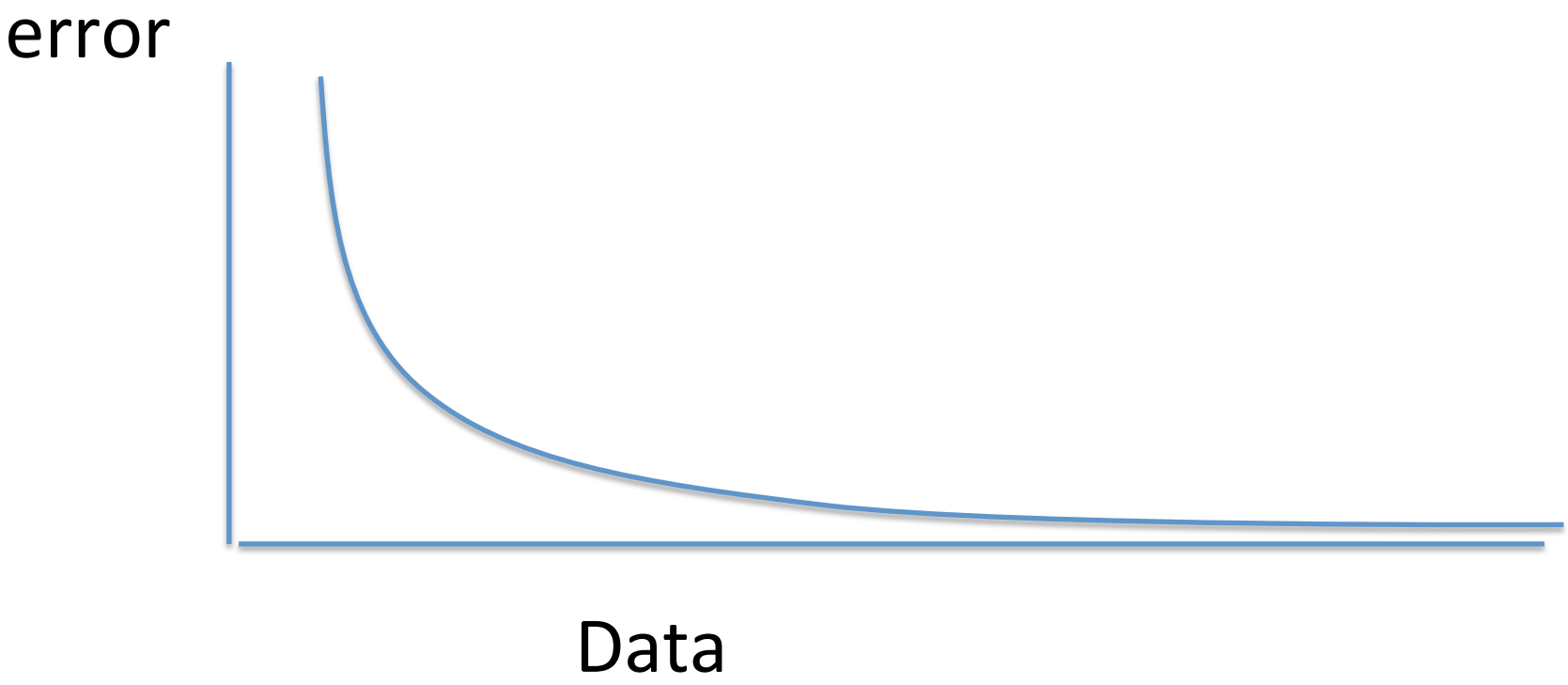
Simplest (Jukes-Cantor, 1969):

- 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.

The different sites are assumed to evolve independently and identically down the tree (with rates that are drawn from a gamma distribution).

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

Statistical Consistency



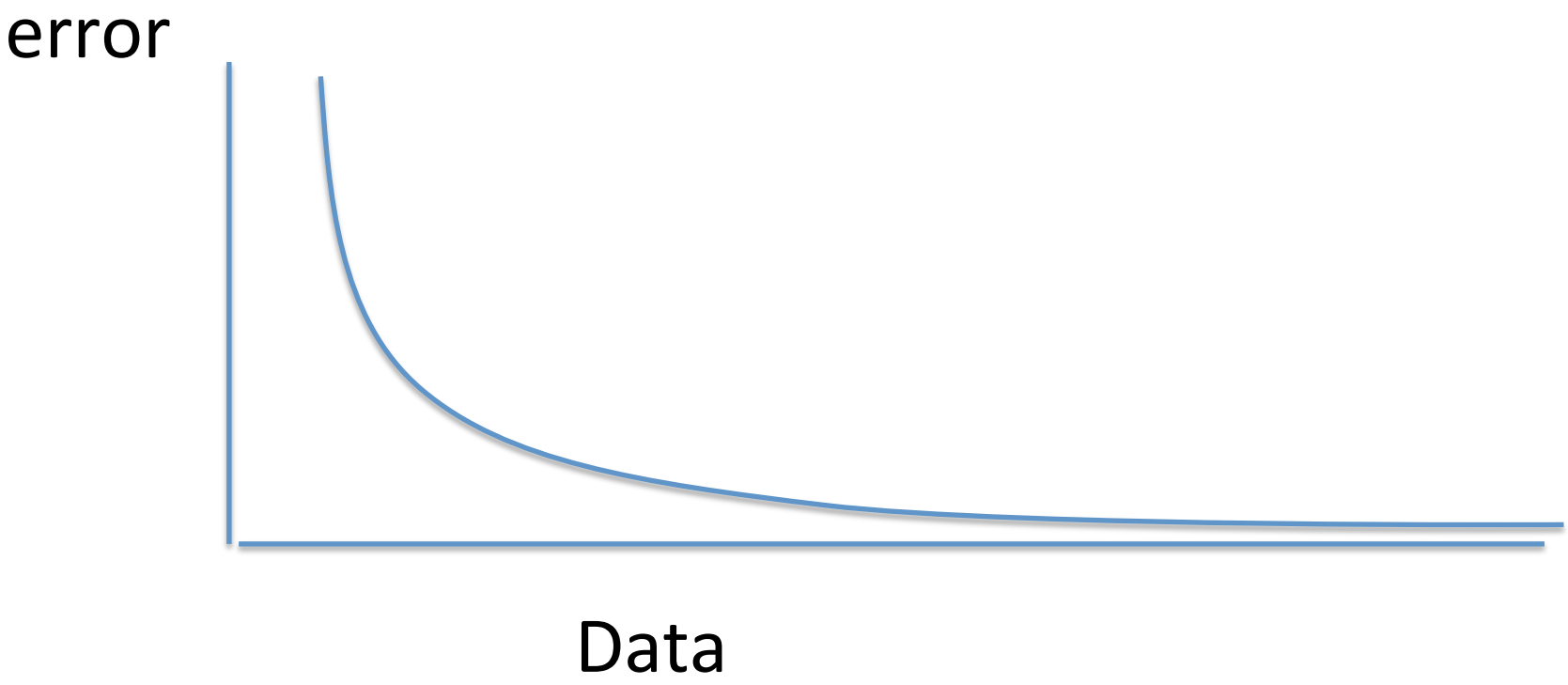
Questions

- 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 is the computational complexity of an estimation problem?

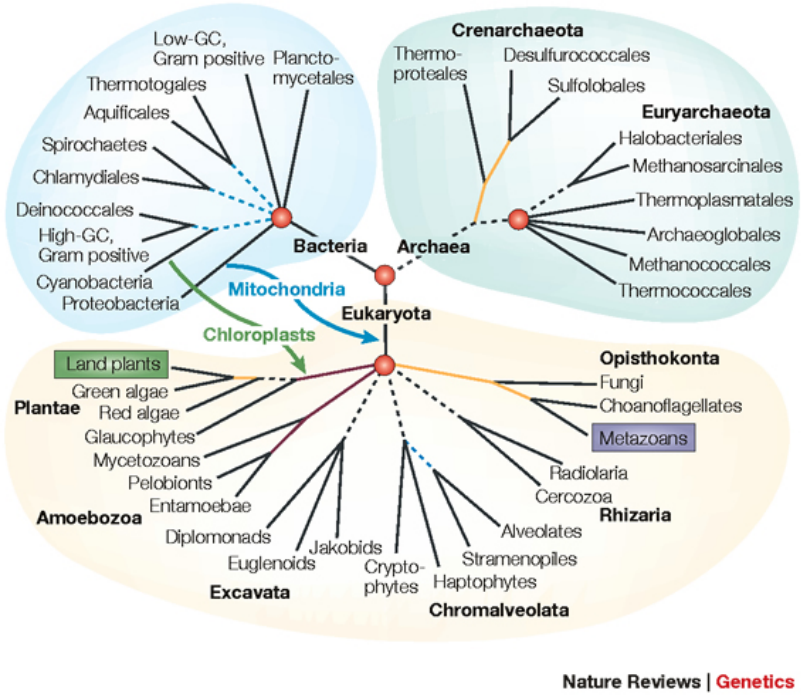
Answers?

- 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.
- Just about everything is NP-hard, and the datasets are big.

Statistical Consistency



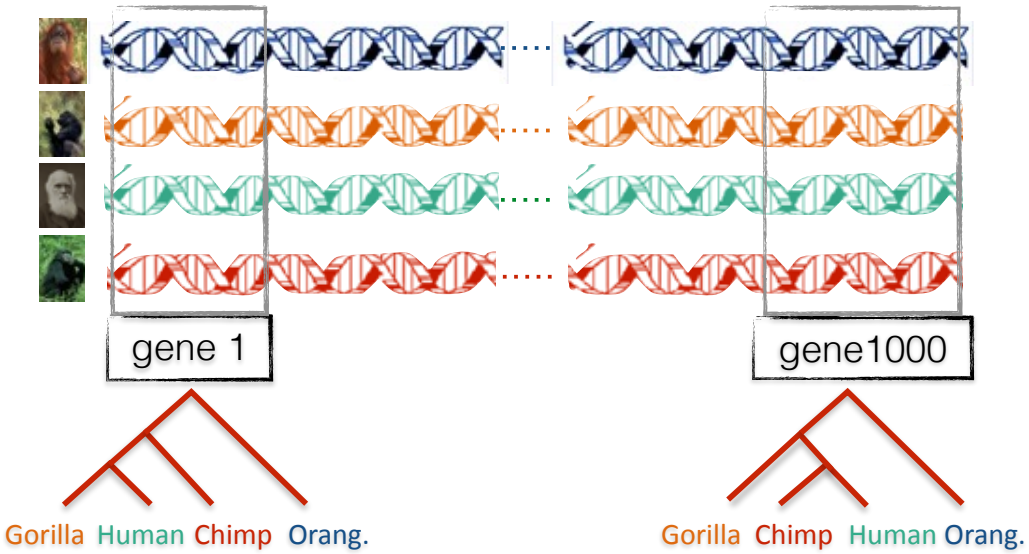
Phylogenomics



Phylogeny + genomics = genome-scale phylogeny estimation

.

Gene tree discordance



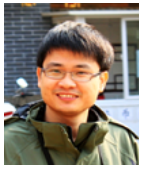
Incomplete Lineage Sorting (ILS) is a dominant cause of gene tree heterogeneity

Avian Phylogenomics Project

E Jarvis, HHMI

MTP Gilbert, Copenhagen

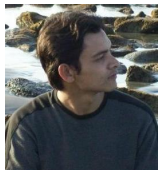
G Zhang, BGI



T. Warnow UT-Austin

S. Mirarab UT-Austin

Md. S. Bayzid, UT-Austin



Plus many many other people…

- Approx. 50 species, whole genomes, 14,000 loci
- Jarvis, Mirarab, et al., *Science 2*014

Major challenge:

- Massive gene tree heterogeneity consistent with incomplete lineage sorting

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

Challenge:
- Massive gene tree heterogeneity consistent with ILS

Incomplete Lineage Sorting (ILS)

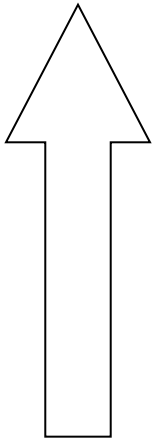
- Confounds phylogenetic analysis for many groups: Hominids, Birds, Yeast, Animals, Toads, Fish, Fungi, etc.
- There is substantial debate about how to analyze phylogenomic datasets in the presence of ILS, focused around statistical consistency guarantees (theory) and performance on data.

This talk

- Gene tree heterogeneity due to incomplete lineage sorting, modelled by the multi-species coalescent (MSC)
- Statistically consistent estimation of species trees under the MSC, and the impact of gene tree estimation error
- ASTRAL (Bioinformatics 2014, 2015): coalescent-based species tree estimation method that has high accuracy on large datasets (1000 species and genes)
- Statistical binning (Science 2015 and PLOS One 2015)
- Open questions

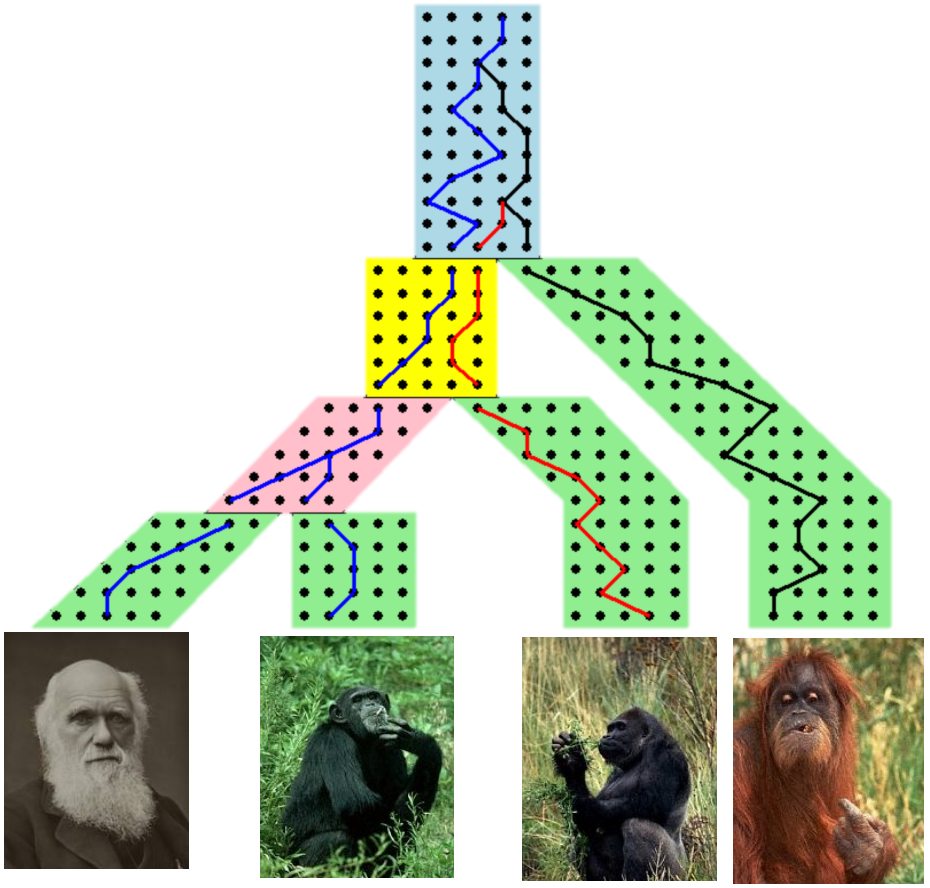
Gene trees inside the species tree (Coalescent Process)

Past



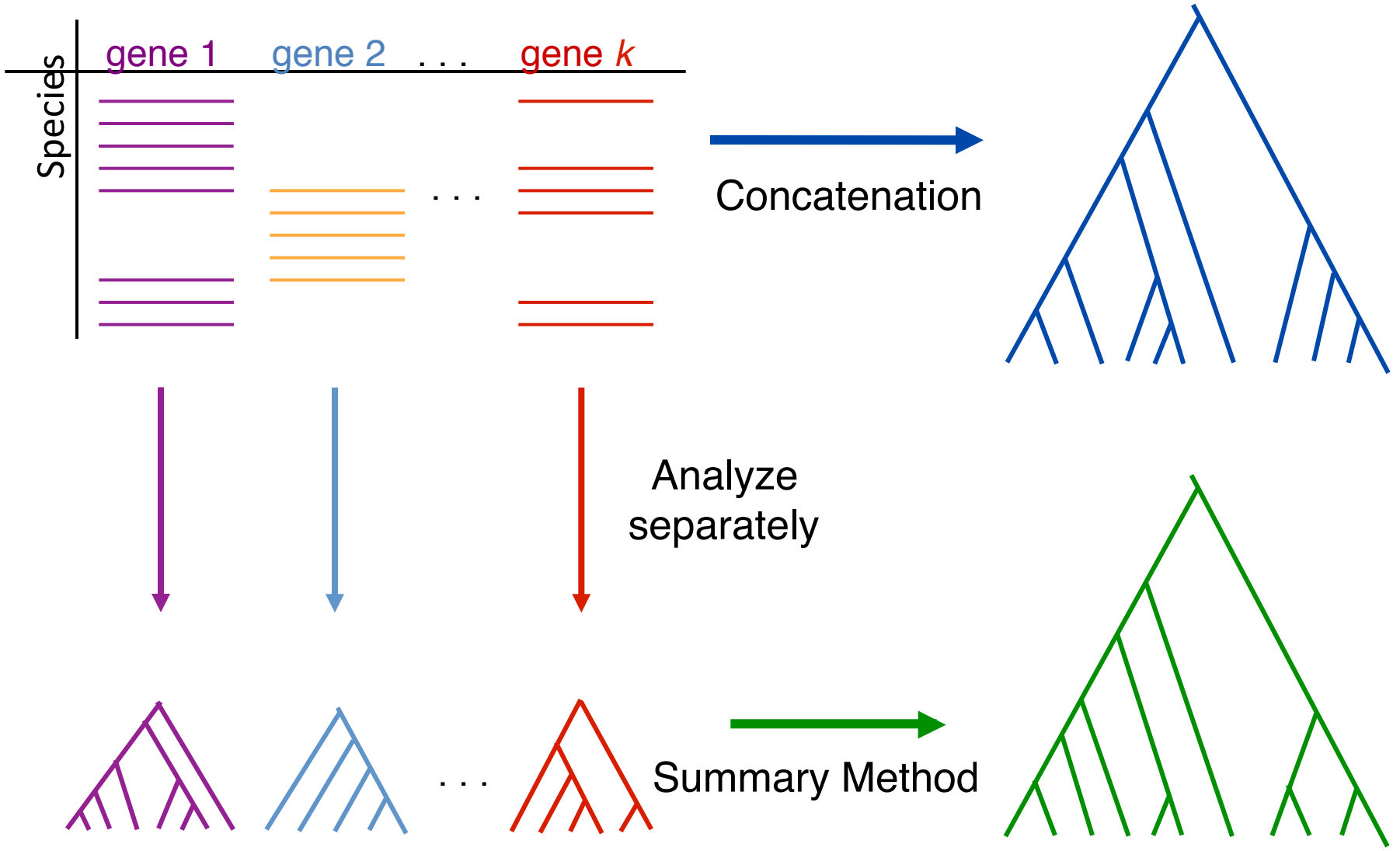
Present

Courtesy James Degnan

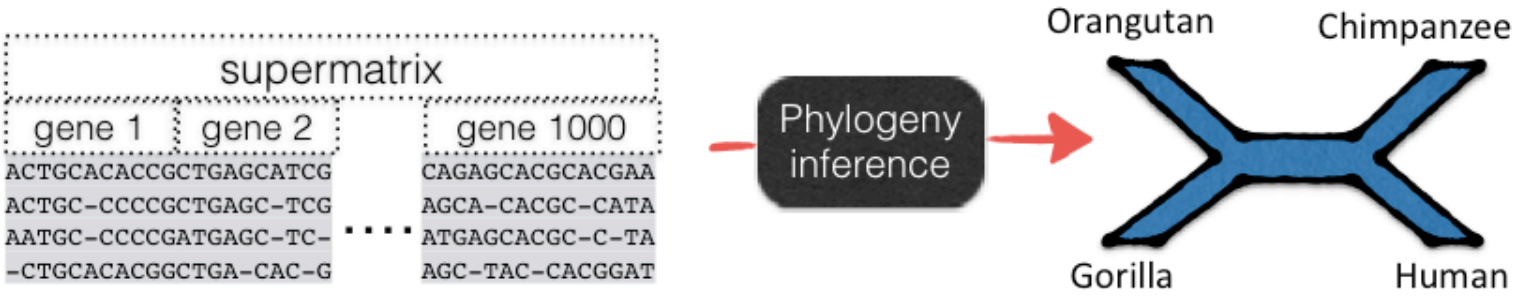


Gorilla and Orangutan are not siblings in the species tree, but they are in the gene tree.

Main competing approaches

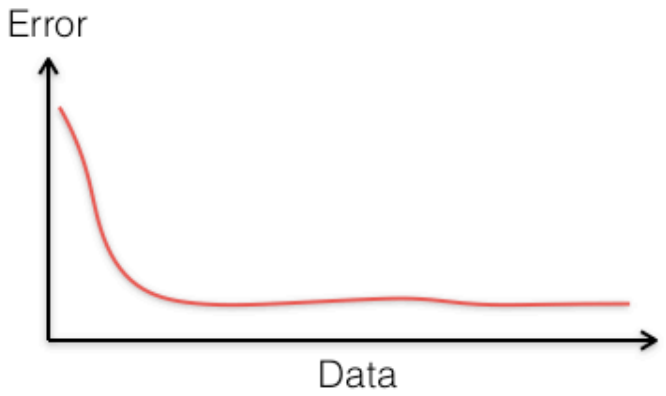


Traditional approach: concatenation



- 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]



What about summary methods?

What about summary methods?

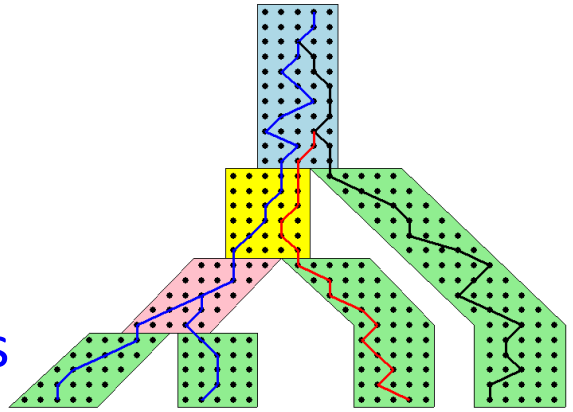
Techniques:

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

Under the multi-species coalescent model, the species tree defines a probability distribution on the gene trees

Courtesy James Degnan

Theorem (Degnan et al., 2006, 2009): Under the multi-species coalescent model, for any three taxa A, B, and C, the most probable rooted gene tree on {A,B,C} is identical to the rooted species tree induced on {A,B,C}.



How to compute a species tree?

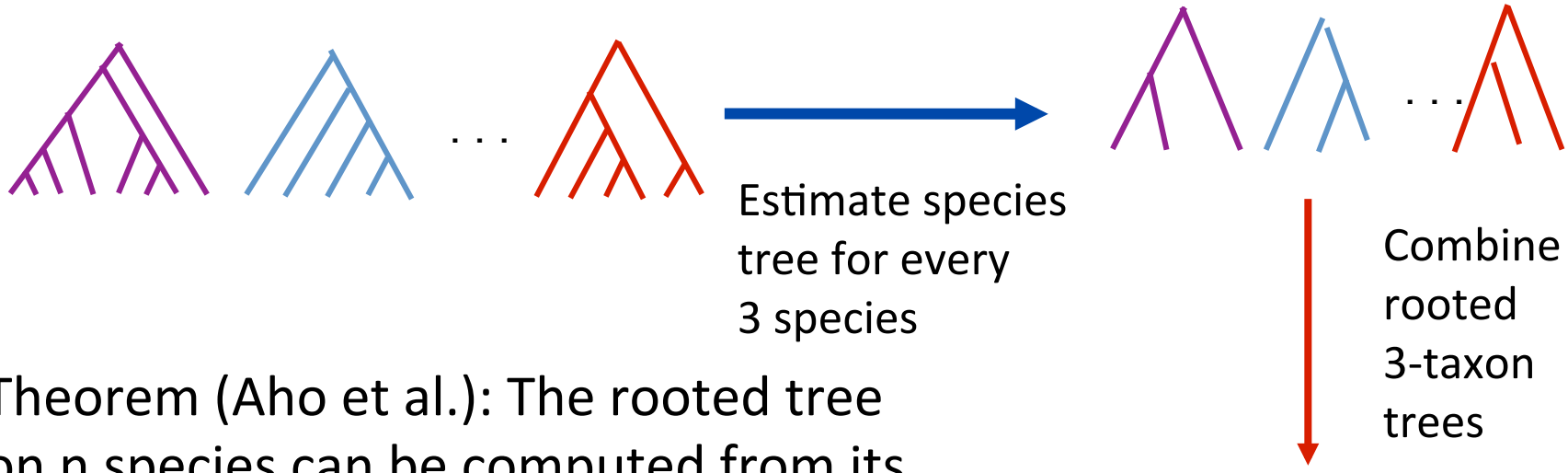
Theorem (Degnan et al., 2006, 2009): Under the multi-species coalescent model, for any three taxa A, B, and C, the most probable rooted gene tree on {A,B,C} is identical to the rooted species tree induced on {A,B,C}.

How to compute a species tree?

Estimate species tree for every 3 species

Theorem (Aho et al.): The rooted tree on n species can be computed from its set of 3-taxon rooted subtrees in polynomial time.

How to compute a species tree?



Theorem (Aho et al.): The rooted tree on n species can be computed from its set of 3-taxon rooted subtrees in polynomial time.

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

Challenges:

- Massive gene tree heterogeneity consistent with ILS
- At the time, MP-EST was the leading method.
- However, we could not use MP-EST due to missing data (many gene trees could not be rooted) and large number of species.

Species tree estimation from unrooted gene trees

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}.

Hence: Statistically consistent methods for estimating unrooted species trees from unrooted gene trees:

BUCKy-pop (Larget et al., 2010) and

ASTRAL (Mirarab et al., 2014 and Mirarab and Warnow 2015)

Minimum Quartet Distance

Input: Set of gene trees, t_1, t_2, …, t_k, each on leafset S

Output: Tree T on leafset S that minimizes

$$\Sigma_t \, d(t,T)$$

where d(t,T) is the quartet distance between the two trees, t and T.

Notes:

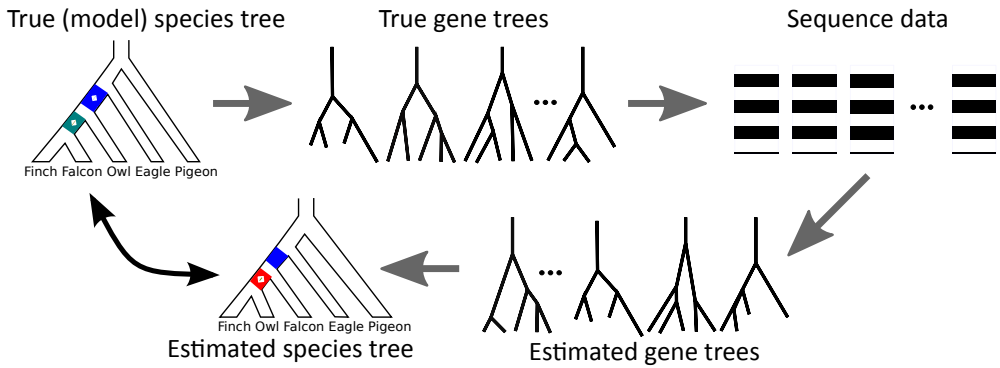
- The computational complexity of this problem is open.
- An exact solution to this problem is a statistically consistent method for species tree estimation under the multi-species coalescent model.

ASTRAL and ASTRAL-2

- Exactly solves the minimum quartet distance problem subject to a constraint on the solution space, given by input set X of bipartitions (find tree T taking its bipartitions from a set X, that minimizes the quartet distance).
- Theorem: ASTRAL is statistically consistent under the MSC, even when solved in constrained mode (drawing bipartitions from the input gene trees)
- The constrained version of ASTRAL runs in polynomial time
- Open source software at https://github.com/smirarab
- Published in Bioinformatics 2014 and 2015
- Used in Wickett, Mirarab et al. (PNAS 2014) and Prum, Berv et al. (Nature 2015)

Simulation study

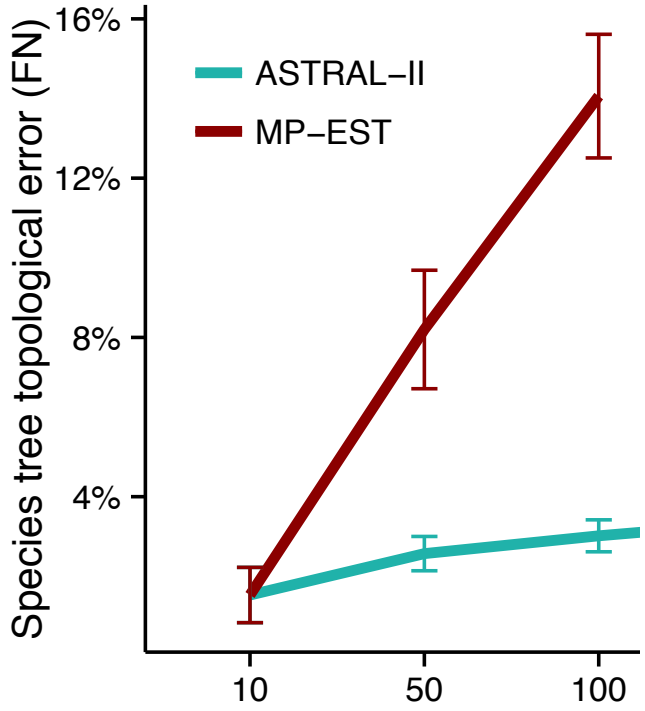
- Variable parameters:
  - Number of species: 10 – 1000
  - Number of genes: 50 – 1000
  - Amount of ILS: low, medium, high
  - Deep versus recent speciation



- 11 model conditions (50 replicas each) with heterogenous gene tree error
- Compare to NJst, MP-EST, concatenation (CA-ML)
- Evaluate accuracy using FN rate: the percentage of branches in the true tree that are missing from the estimated tree

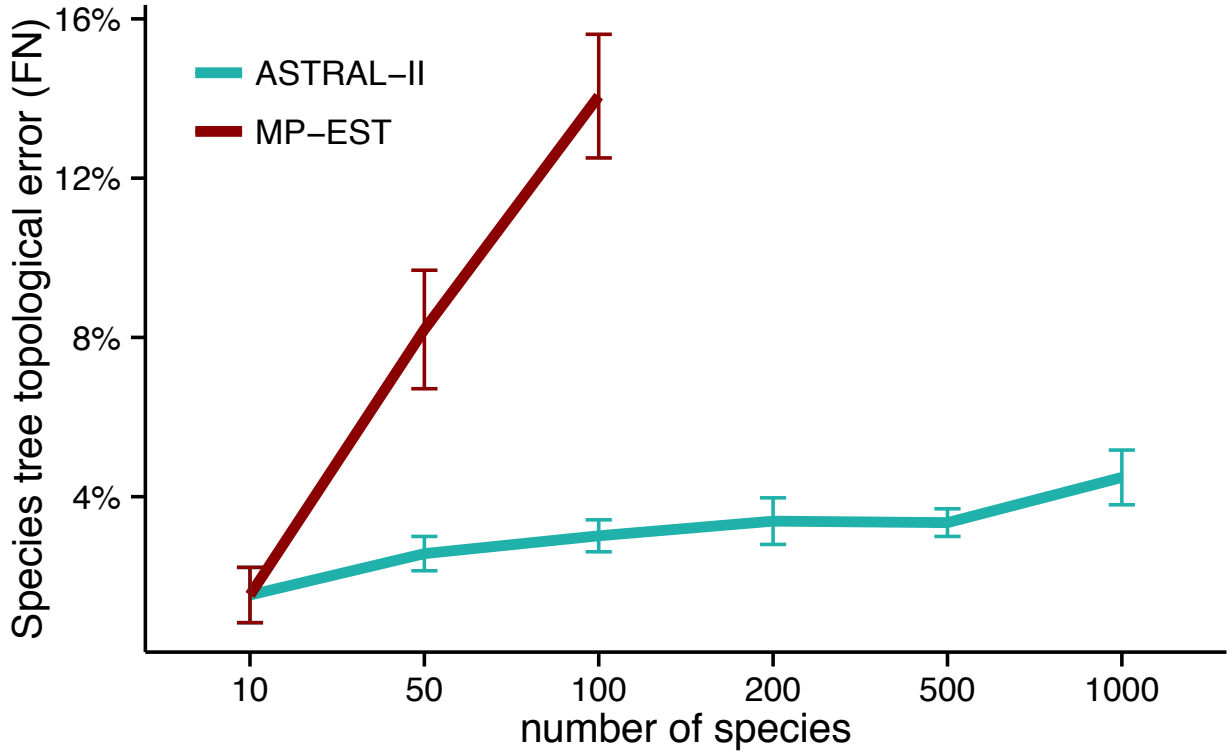
Used SimPhy, Mallo and Posada, 2015

Tree accuracy when varying the number of species



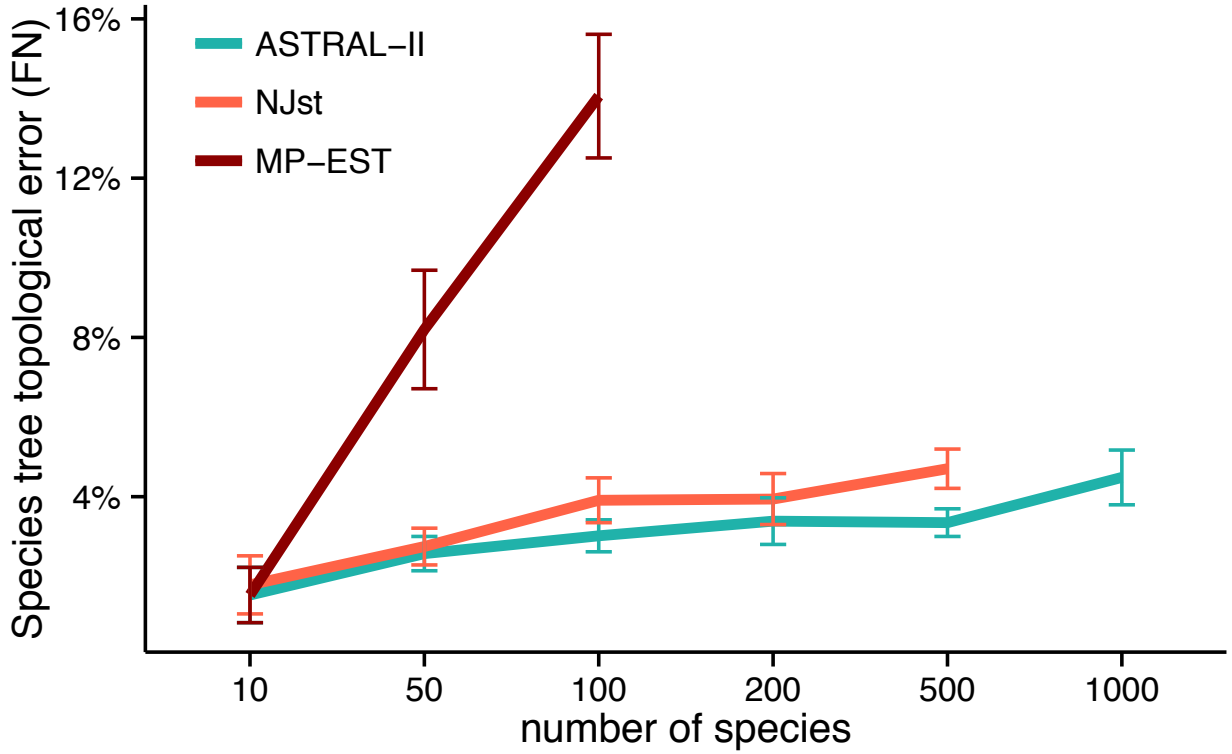
1000 genes, “medium” levels of recent ILS

Tree accuracy when varying the number of species



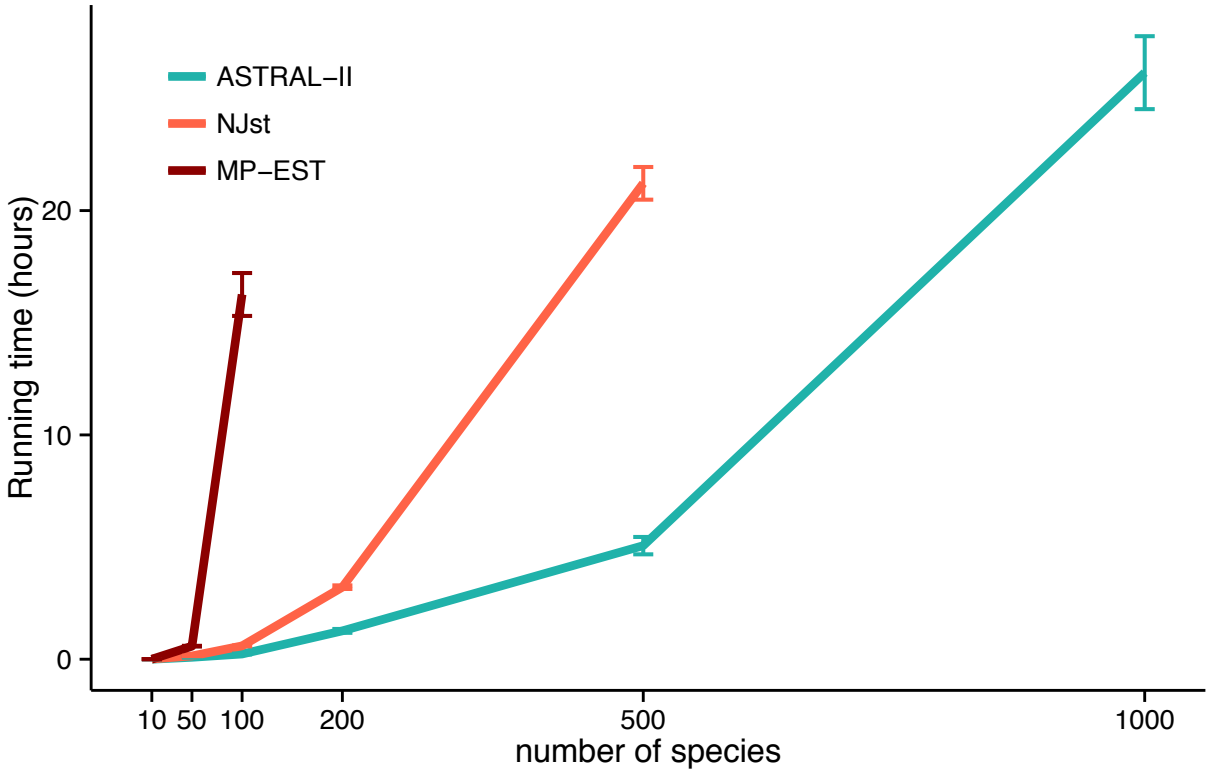
1000 genes, “medium” levels of recent ILS

Tree accuracy when varying the number of species



1000 genes, “medium” levels of recent ILS

Running time when varying the number of species



1000 genes, “medium” levels of recent ILS

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

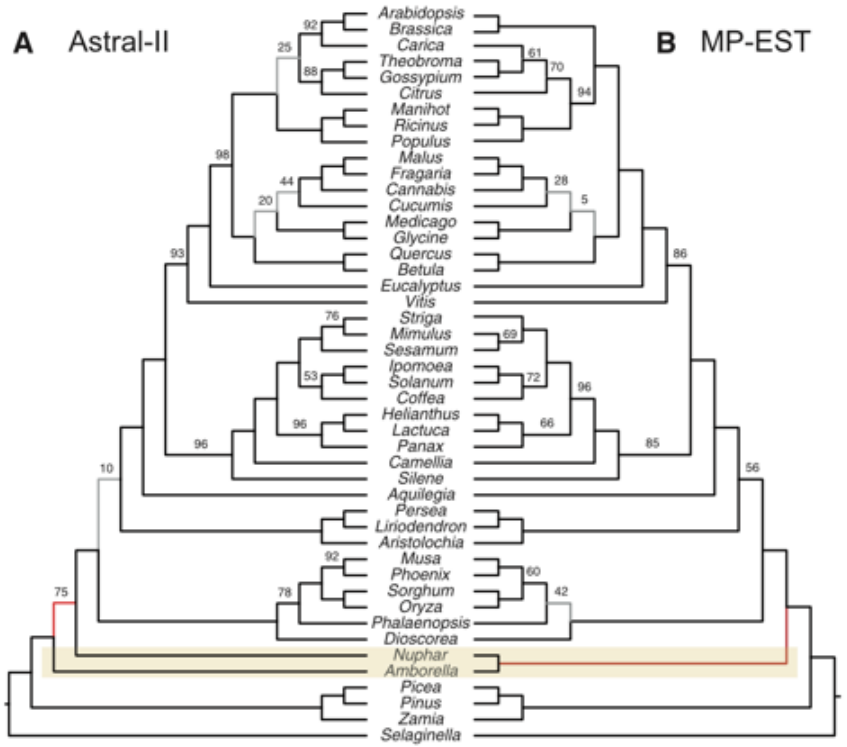
- 103 plant transcriptomes, 400-800 single copy "genes"
- Next phase will be much bigger
- Wickett, Mirarab et al., *PNAS* 2014

Two trees presented: one based on concatenation, and the other based on ASTRAL.

The two trees were highly consistent.

Insights on biological data

- Main question: The placement of Amborella at the base of angiosperms
- Xi et al. (2014) used a collection of 310 genes sampled from 46 species.
- Conflicting results:
  - Concatenation puts Amborella at the base (H1)
  - MP-EST puts Amobrella+water lilies at the base (H2)
  - Xi et al. conclude ILS is the cause
- ASTRAL like many other recent studies (e.g., 1KP) recovers H1
  - ILS is not necessarily the cause

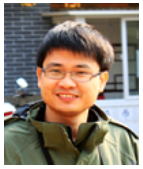


Avian Phylogenomics Project

E Jarvis, HHMI

MTP Gilbert, Copenhagen

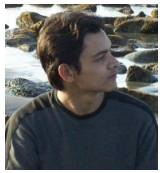
G Zhang, BGI



T. Warnow UT-Austin

S. Mirarab UT-Austin

Md. S. Bayzid, UT-Austin



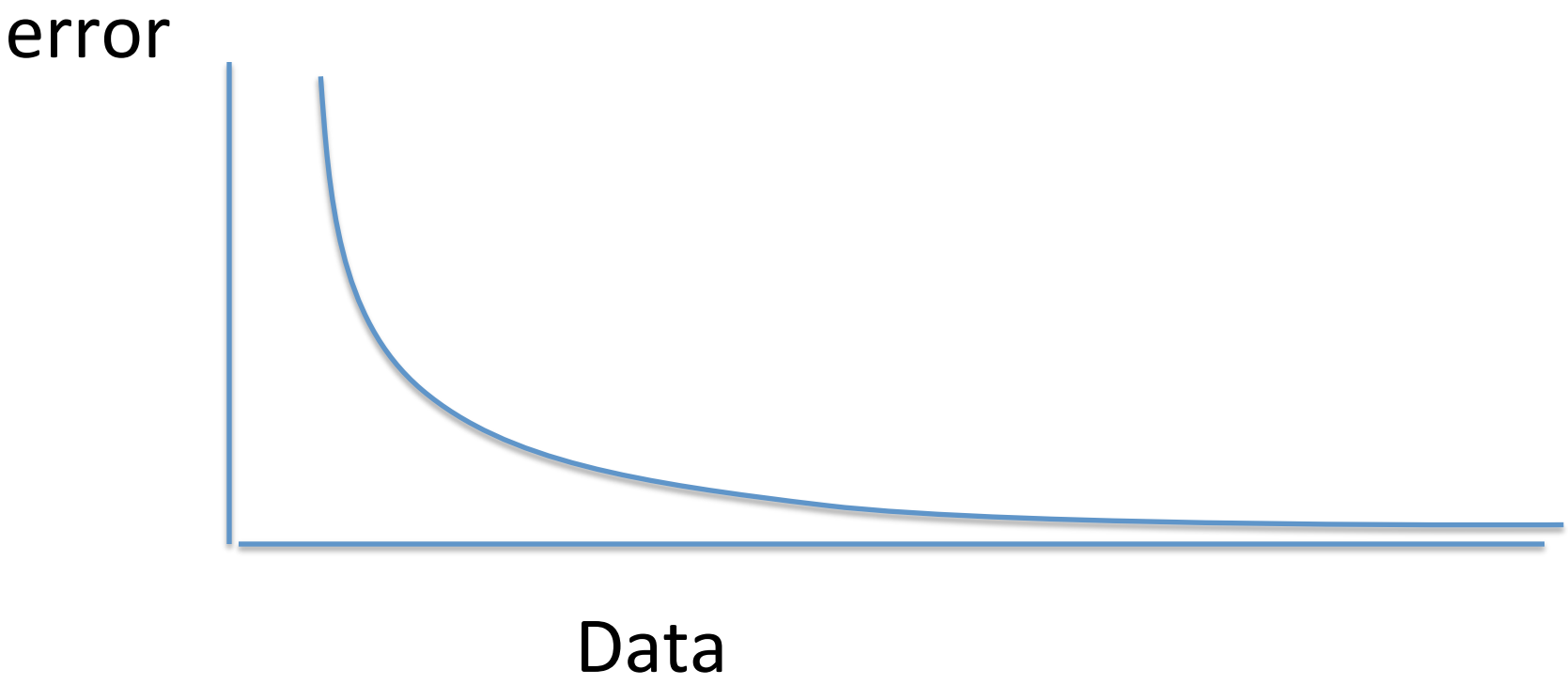
Plus many many other people…

- Approx. 50 species, whole genomes, 14,000 loci
- Jarvis, Mirarab, et al., Science 2014

Major challenge:

- Massive gene tree heterogeneity consistent with incomplete lineage sorting
- Very poor resolution in the 14,000 gene trees (average bootstrap support 25%)
- Standard coalescent-based species tree estimation methods contradicted concatenation analysis and prior studies

Statistical Consistency for summary methods



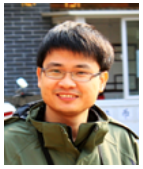
Data are gene trees, presumed to be randomly sampled <u>true gene trees.</u>

Avian Phylogenomics Project

E Jarvis, HHMI

MTP Gilbert, Copenhagen

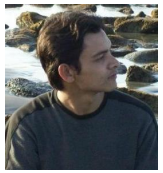
G Zhang, BGI



T. Warnow UT-Austin

S. Mirarab UT-Austin

Md. S. Bayzid, UT-Austin



Plus many many other people…

- Approx. 50 species, whole genomes, 14,000 loci
- Published Science 2014

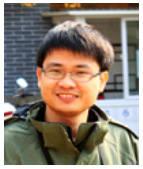
Most gene trees had very low bootstrap support, suggestive of gene tree estimation error

Avian Phylogenomics Project

E Jarvis, HHMI

MTP Gilbert, Copenhagen

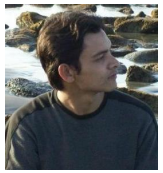
G Zhang, BGI



T. Warnow UT-Austin

S. Mirarab UT-Austin

Md. S. Bayzid, UT-Austin



Plus many many other people…

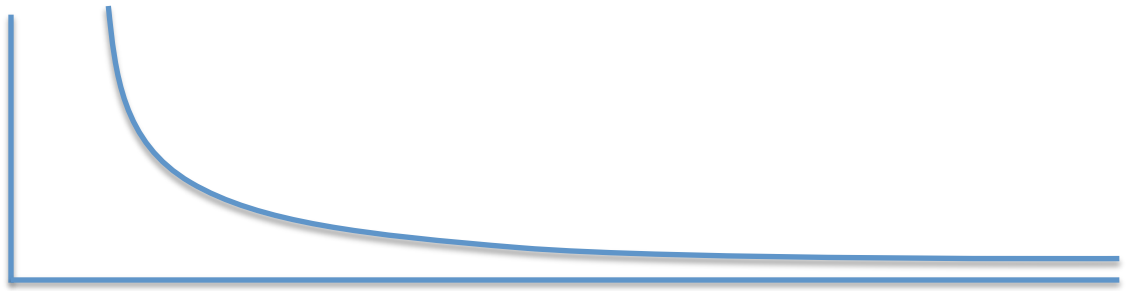
- Approx. 50 species, whole genomes, 14,000 loci

Solution: **Statistical Binning**

- Improves coalescent-based species tree estimation by improving gene trees (Mirarab, Bayzid, Boussau, and Warnow, *Science* 2014)
- Avian species tree estimated using **Statistical Binning with MP-EST** (Jarvis, Mirarab, et al., *Science* 2014)

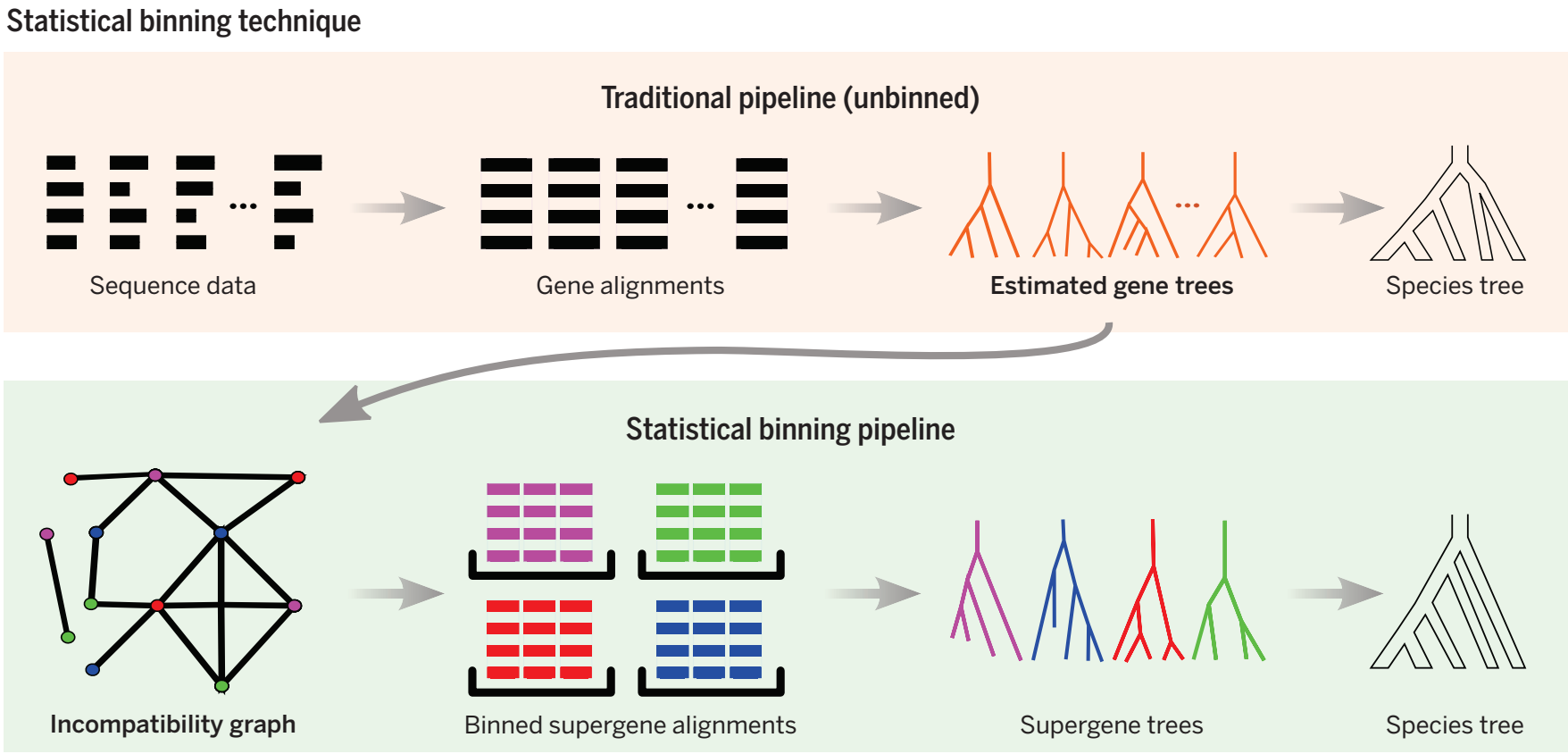
Ideas behind statistical binning

- "Gene tree" error tends to decrease with the number of sites in the alignment



Number of sites in an alignment

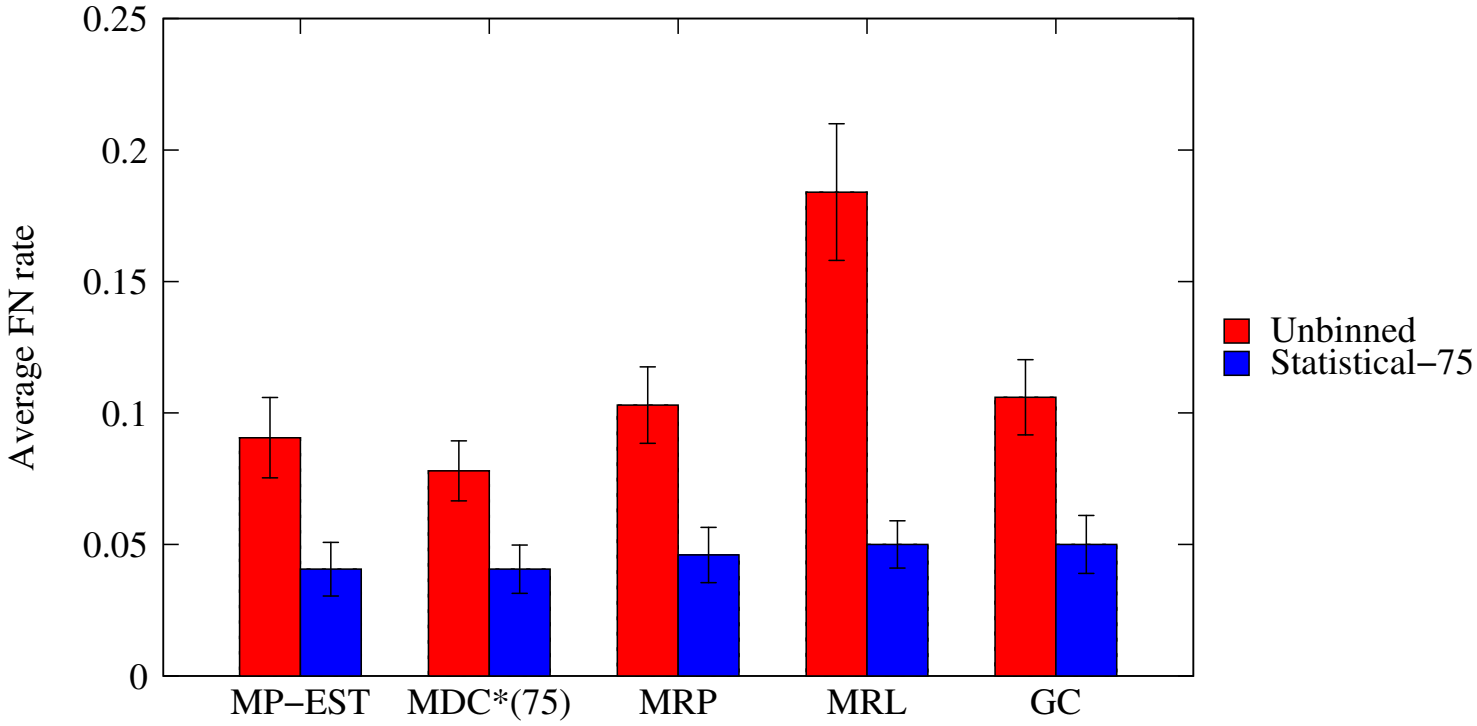
- Concatenation (even if not statistically consistent) tends to be reasonably accurate when there is not too much gene tree heterogeneity



The statistical binning pipeline for estimating species trees from gene trees. Loci are grouped into bins based on a statistical test for combinabilty, before estimating gene trees.

Note: Supergene trees computed using fully partitioned maximum likelihood
Vertex-coloring graph with balanced color classes is NP-hard; we used heuristic.

Statistical binning vs. unbinned



Datasets: 11-taxon strongILS datasets with 50 genes from Chung and Ané, Systematic Biology

Binning produces bins with approximate 5 to 7 genes each

Theorem 3 (PLOS One, Bayzid et al. 2015): Unweighted statistical binning pipelines are not statistically consistent under GTR+MSC

As the number of sites per locus increase:

- All estimated gene trees converge to the true gene tree and have bootstrap support that converges to 1 (Steel 2014)
- For each bin, with probability converging to 1, the genes in the bin have the same tree topology (but can have different numeric parameters), and there is only one bin for any given tree topology
- For each bin, a fully partitioned maximum likelihood (ML) analysis of its supergene alignment converges to a tree with the common gene tree topology.

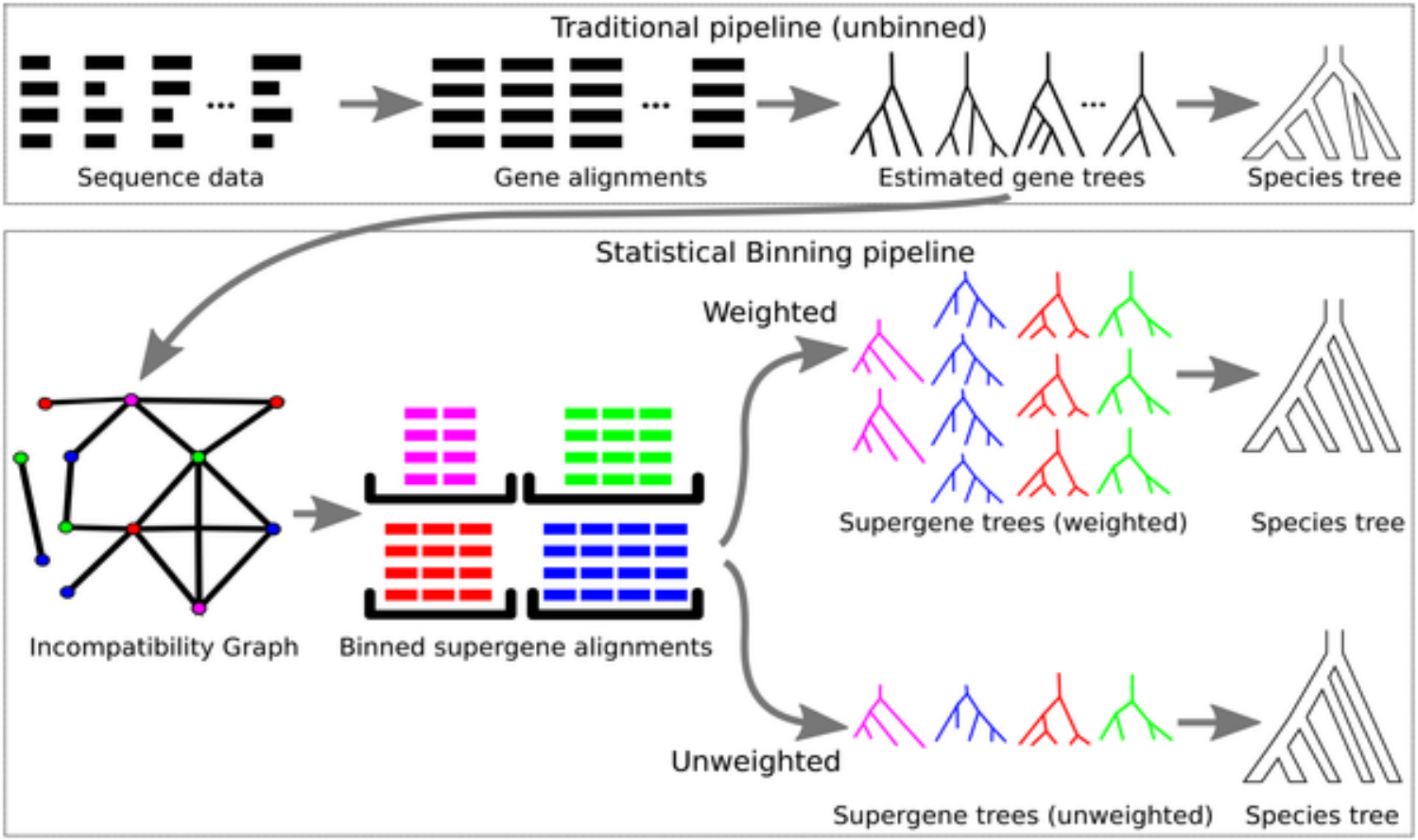
As the number of loci increase:

- every gene tree topology appears with probability converging to 1.

Hence as both the number of loci and number of sites per locus increase, with probability converging to 1, every gene tree topology appears exactly once in the set of supergene trees.

It is impossible to infer the species tree from the flat distribution of gene trees!

Fig 1. Pipeline for unbinned analyses, unweighted statistical binning, and weighted statistical binning.



Bayzid MS, Mirarab S, Boussau B, Warnow T (2015) Weighted Statistical Binning: Enabling Statistically Consistent Genome-Scale Phylogenetic Analyses. PLoS ONE 10(6): e0129183. doi:10.1371/journal.pone.0129183
http://127.0.0.1:8081/plosone/article?id=info:doi/10.1371/journal.pone.0129183

Theorem 2 (PLOS One, Bayzid et al. 2015): WSB pipelines are statistically consistent under GTR+MSC

Easy proof:

As the number of sites per locus increase

- All estimated gene trees converge to the true gene tree and have bootstrap support that converges to 1 (Steel 2014)
- For every bin, with probability converging to 1, the genes in the bin have the same tree topology
- Fully partitioned GTR ML analysis of each bin converges to a tree with the common topology of the genes in the bin

Hence as the number of sites per locus and number of loci both increase, WSB followed by a statistically consistent summary method will converge in probability to the true species tree. Q.E.D.

Table 1. Model trees used in the Weighted Statistical Binning study. We show number of taxa, species tree branch length (relative to base model), and average topological discordance between true gene trees and true species tree.

Dataset	Species tree branch length scaling	Average Discordance (%)
Avian (48)	2X	35
Avian (48)	1X	47
Avian (48)	0.5X	59
Mammalian (37)	2X	18
Mammalian (37)	1X	32
Mammalian (37)	0.5X	54
10-taxon	“Lower ILS"	40
10-taxon	“Higher ILS"	84
15-taxon	“High ILS"	82

doi:10.1371/journal.pone.
0129183.t001

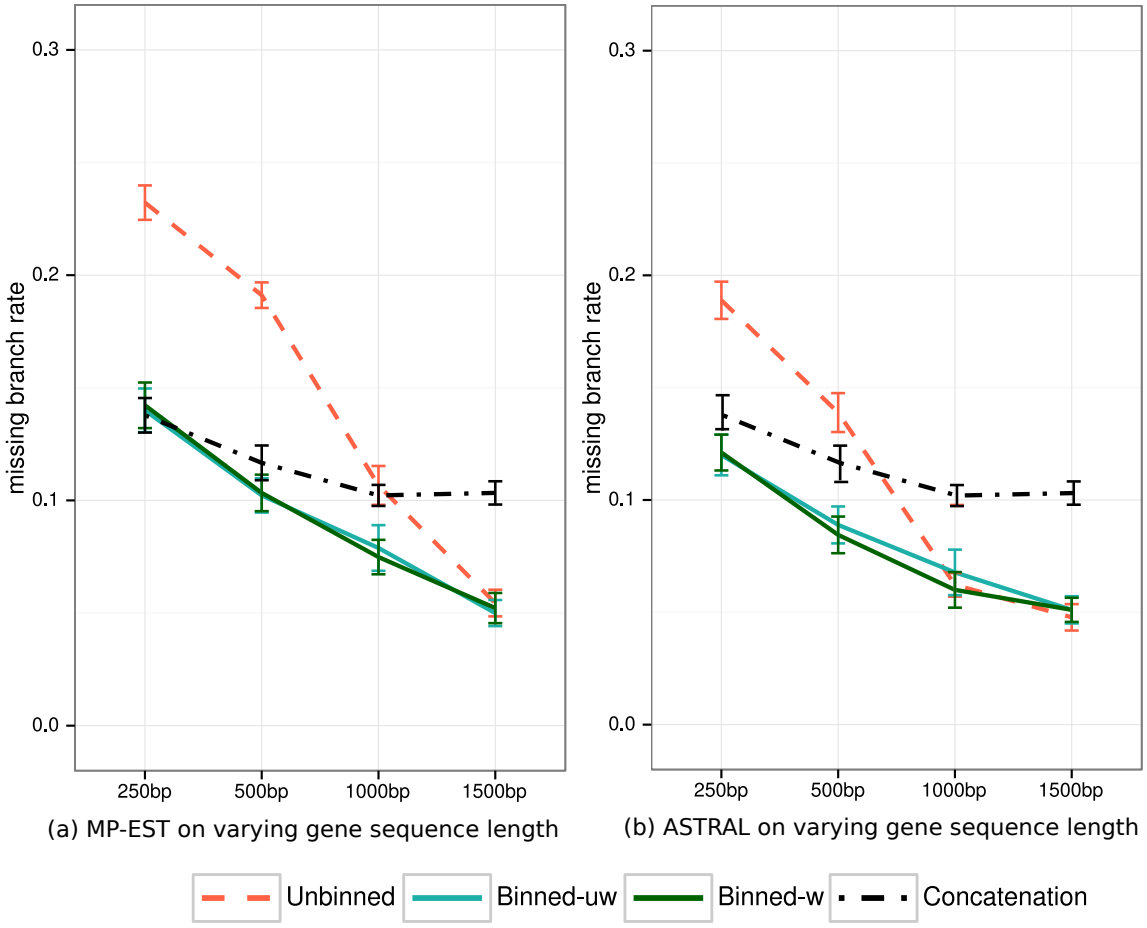
Weighted Statistical Binning: empirical

WSB generally benign to highly beneficial for moderate to large datasets:

- Improves gene tree estimation
- Improves species tree topology
- Improves species tree branch length
- Reduces incidence of highly supported false positive branches

- WSB can be hurtful on very small datasets with very high ILS levels.

Binning can improve species tree topology estimation

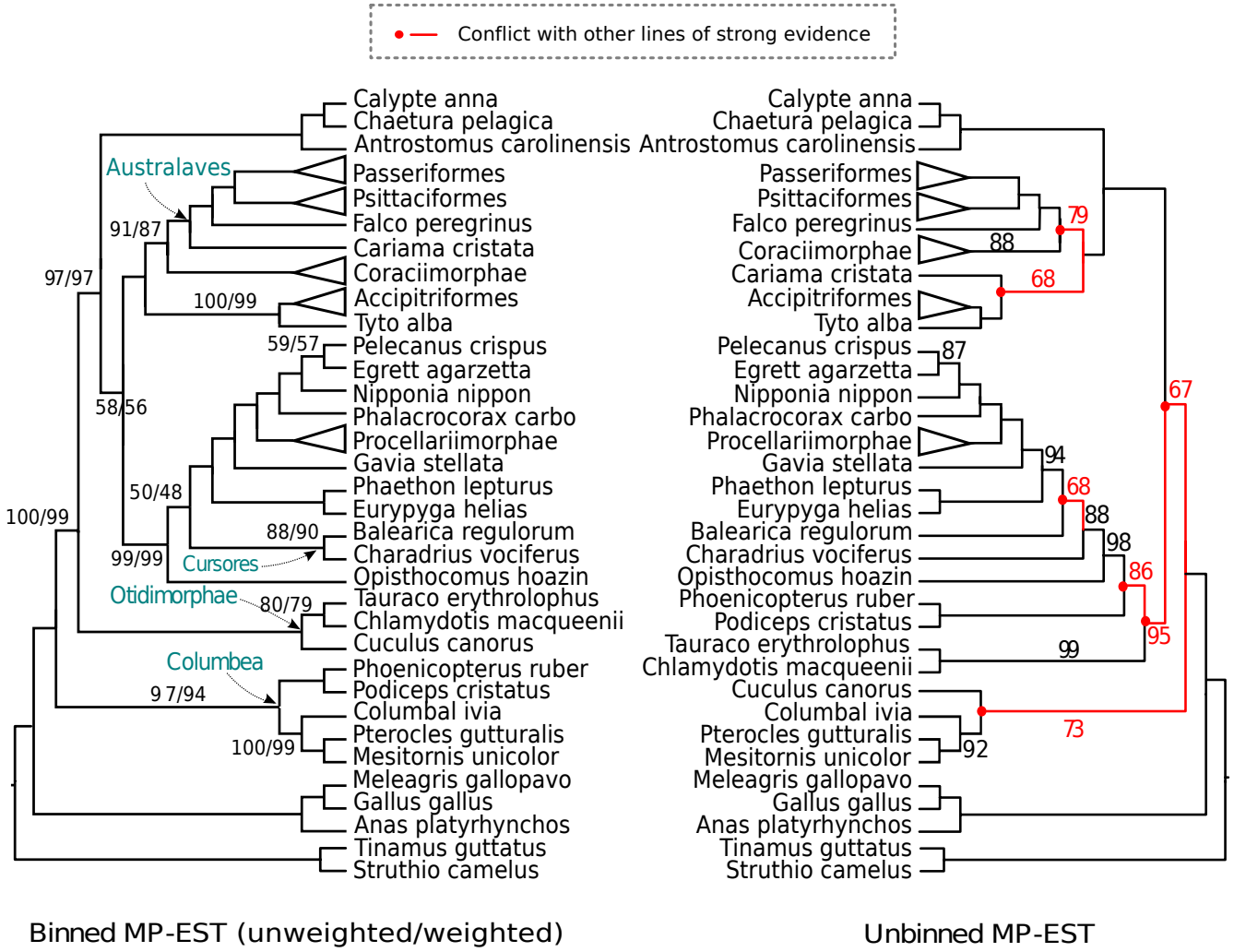


Species tree estimation error for MP-EST and ASTRAL, and also concatenation using ML, on avian simulated datasets: 48 taxa, moderately high ILS (AD=47%), 1000 genes, and varying gene sequence length.

Bayzid et al., (2015). PLoS ONE 10(6): e0129183

PLOS | ONE

Comparing Binned and Un-binned MP-EST on the Avian Dataset



Binned MP-EST (unweighted/weighted)

Unbinned MP-EST

Binned MP-EST is largely consistent with the ML concatenation analysis.

The trees presented in Science 2014 were the ML concatenation and Binned MP-EST

Bayzid et al., (2015). PLoS ONE 10(6): e0129183

PLOS ONE

What about performance on bounded number of sites?

- Question: Do any summary methods converge to the species tree as the number of loci increase, but where each locus has only a constant number of sites?
- Answers: Roch & Warnow, Syst Biol, March 2015:
  - Strict molecular clock: Yes for some new methods, even for a single site per locus
  - No clock: Unknown for all methods, including MP-EST, ASTRAL, etc.

S. Roch and T. Warnow. "On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods", Systematic Biology, 64(4):663-676, 2015, (PDF)

Future Directions

- Better coalescent-based summary methods (that are more robust to gene tree estimation error)

Future Directions

- Better coalescent-based summary methods (that are more robust to gene tree estimation error)
- Better techniques for estimating gene trees given multi-locus data, or for co-estimating gene trees and species trees

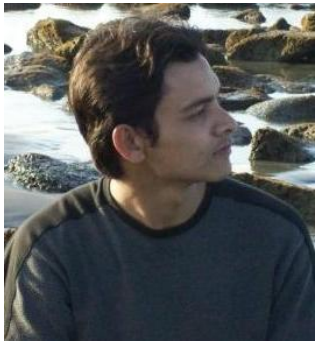
Future Directions

- Better coalescent-based summary methods (that are more robust to gene tree estimation error)
- Better techniques for estimating gene trees given multi-locus data, or for co-estimating gene trees and species trees
- Better theory about robustness to gene tree estimation error (or lack thereof) for coalescent-based summary methods

Future Directions

- Better coalescent-based summary methods (that are more robust to gene tree estimation error)
- Better techniques for estimating gene trees given multi-locus data, or for co-estimating gene trees and species trees
- Better theory about robustness to gene tree estimation error (or lack thereof) for coalescent-based summary methods
- Better “single site” methods

Acknowledgments

Mirarab et al., Science 2014 (Statistical Binning)

Roch and Warnow, Systematic Biology 2014 (Points of View)

Bayzid et al., Science 2015 (Response to Liu and Edwards Comment)

Mirarab and Warnow, Bioinformatics 2015 (ASTRAL-2)

Warnow PLOS Currents: Tree of Life 2014 (concatenation analysis)

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

ASTRAL and statistical binning software at https://github.com/smirarab

Funding: NSF, David Bruton Jr. Centennial Professorship, TACC (Texas Advanced Computing Center), Grainger Foundation, and HHMI (to SM).

ASTRAL-II on biological datasets (ongoing collaborations)

- 1200 plants with ~ 400 genes (1KP consortium)
- 250 avian species with 2000 genes (with LSU, UF, and Smithsonian)
- 200 avian species with whole genomes (with Genome 10K, international)
- 250 suboscine species (birds) with ~2000 genes (with LSU and Tulane)
- 140 Insects with 1400 genes (with U. Illinois at Urbana-Champaign)
- 50 Hummingbird species with 2000 genes (with U. Copenhagen and Smithsonian)
- 40 raptor species (birds) with 10,000 genes (with U. Copenhagen and Berkeley)
- 38 mammalian species with 10,000 genes (with U. of Bristol, Cambridge, and Nat. Univ. of Ireland)