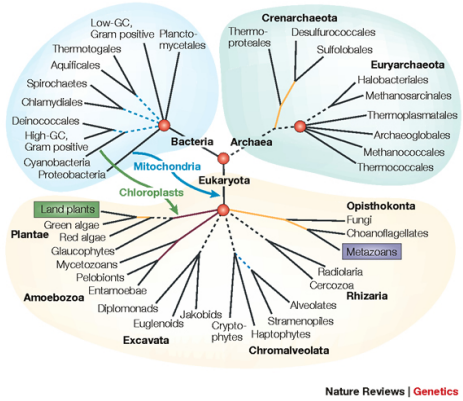


From Gene Trees to Species Trees

Tandy Warnow

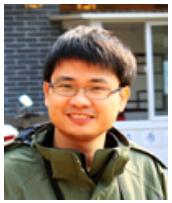
The University of Texas at Austin

Avian Phylogenomics Project

Erich Jarvis, HHMI

MTP Gilbert, Copenhagen

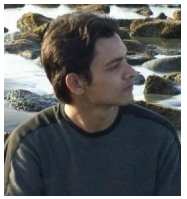
G Zhang, BGI



T. Warnow UT-Austin

S. Mirarab UT-Austin

Md. S. Bayzid, UT-Austin



Plus many many other people…

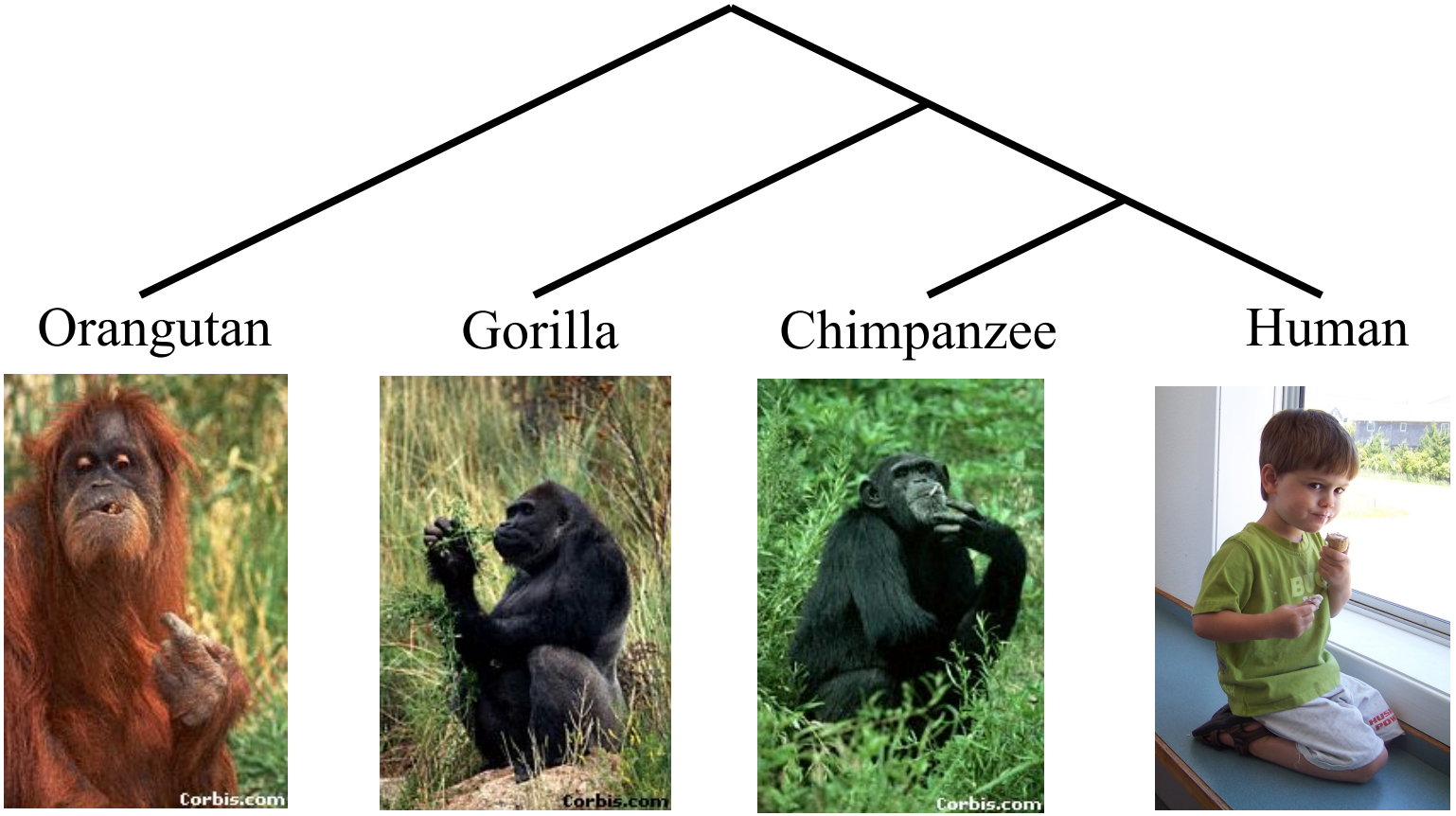
- Approx. 50 species, whole genomes
- 8000+ genes, UCEs
- Gene sequence alignments and trees computed using SATé (Liu et al., Science 2009 and Systematic Biology 2012)

Challenges:

Maximum likelihood on multi-million-site sequence alignments

Massive gene tree incongruence

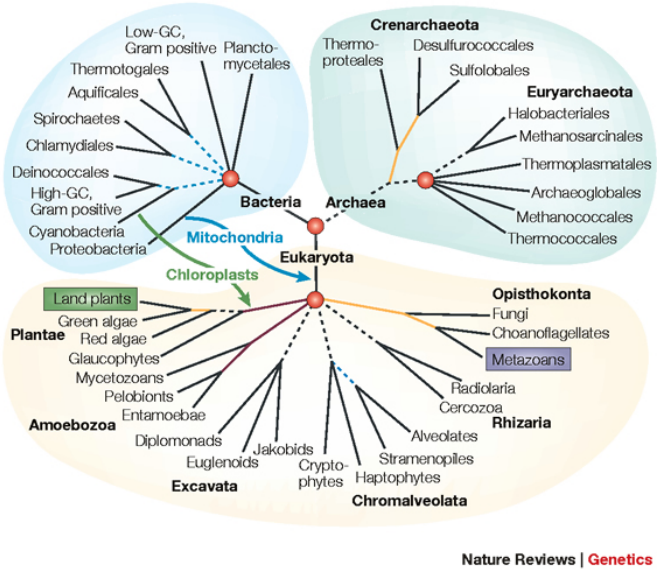
Phylogeny
(evolutionary tree)



From the Tree of the Life Website,
University of Arizona

Phylogenomics

(Phylogenetic estimation from whole genomes)



Nature Reviews | Genetics

Using multiple genes

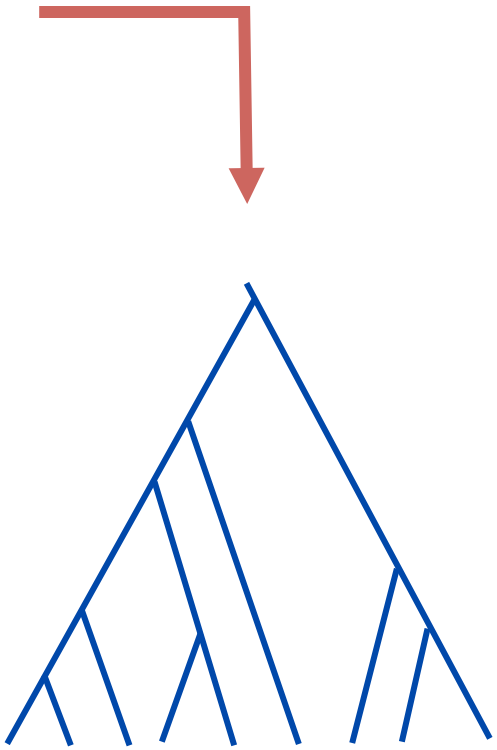
	gene 1
S_1	TCTAATGGAA
S_2	GCTAAGGGAA
S_3	TCTAAGGGAA
S_4	TCTAACGGAA
S_7	TCTAATGGAC
S_8	TATAACGGAA

	gene 2
S_4	GGTAACCCTC
S_5	GCTAAACCTC
S_6	GGTGACCATC
S_7	GCTAAACCTC

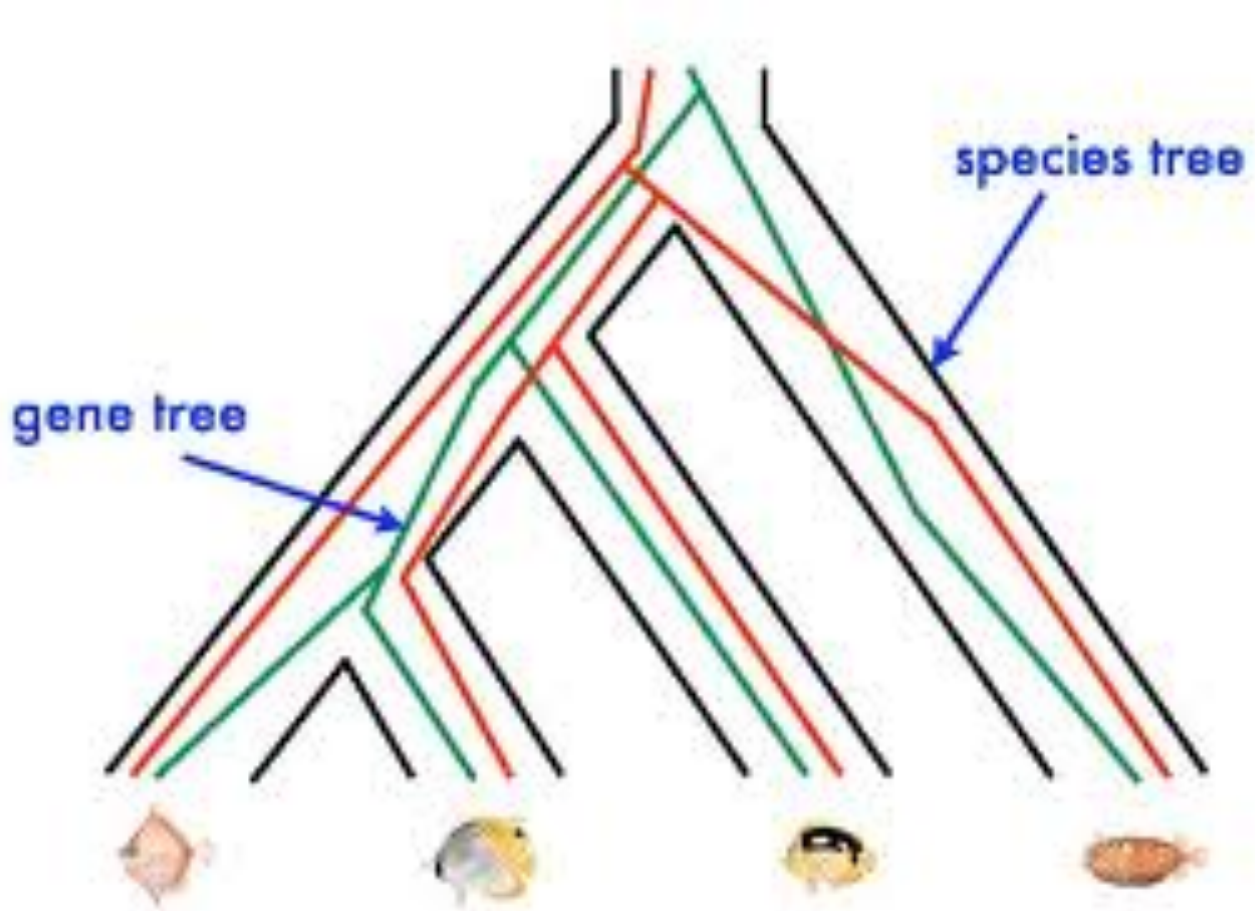
	gene 3
S_1	TATTGATACA
S_3	TCTTGATACC
S_4	TAGTGATGCA
S_7	TAGTGATGCA
S_8	CATTCATACC

Concatenation

	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



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



Gene Tree Incongruence

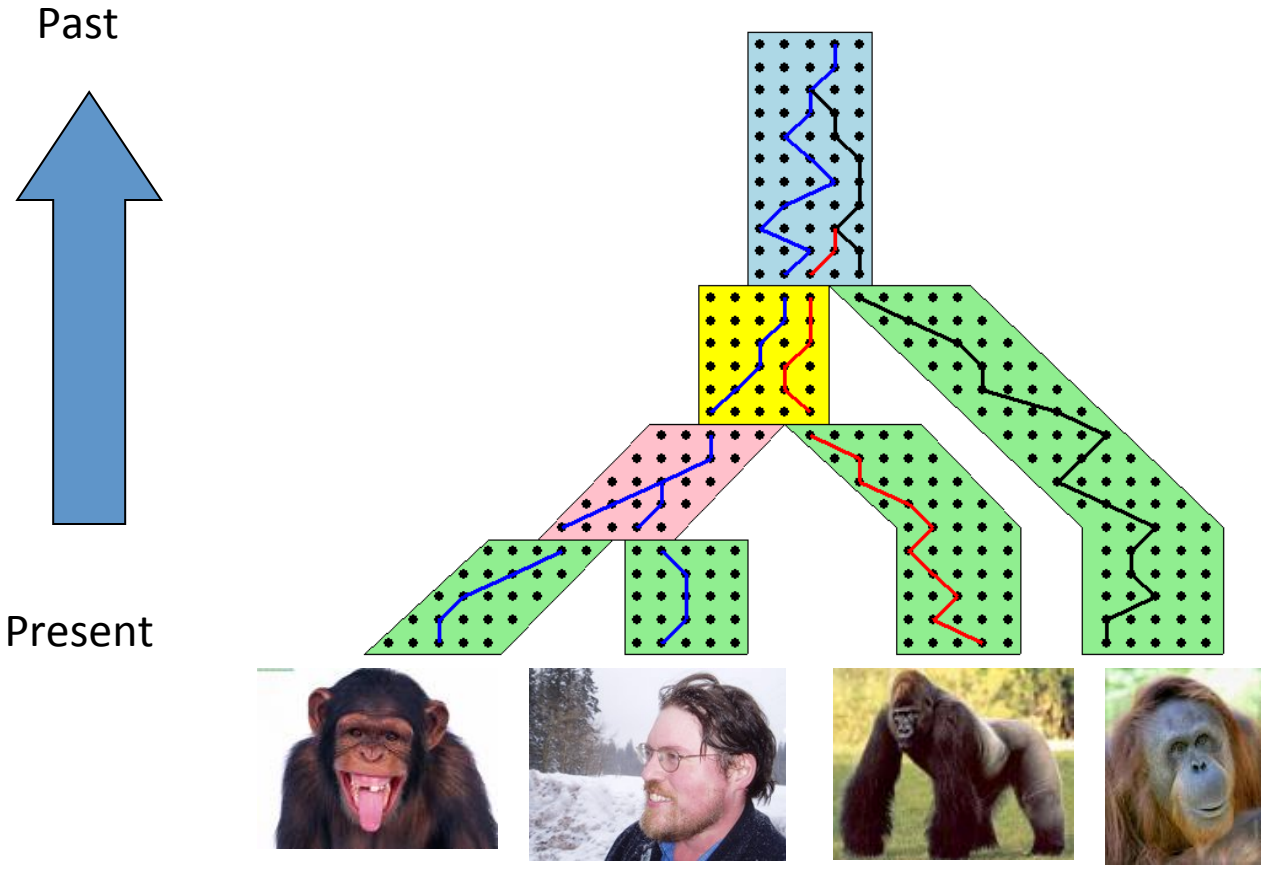
- Gene trees can differ from the species tree due to:
  - Duplication and loss
  - Horizontal gene transfer
  - Incomplete lineage sorting (ILS)

Lineage Sorting

- Population-level process, also called the "Multi-species coalescent" (Kingman, 1982)
- Gene trees can differ from species trees due to short times between speciation events or large population size; this is called "Incomplete Lineage Sorting" or "Deep Coalescence".

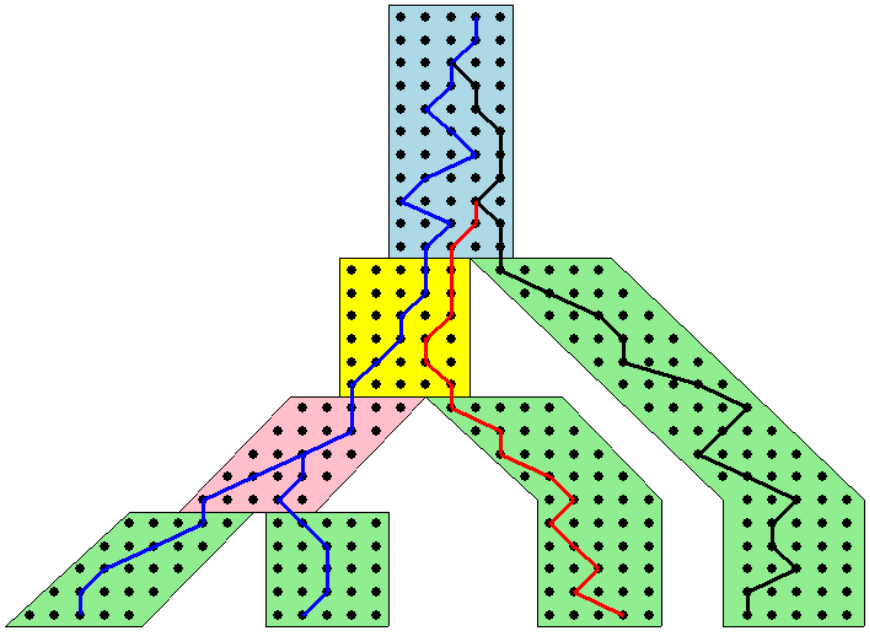
The Coalescent

Courtesy James Degnan

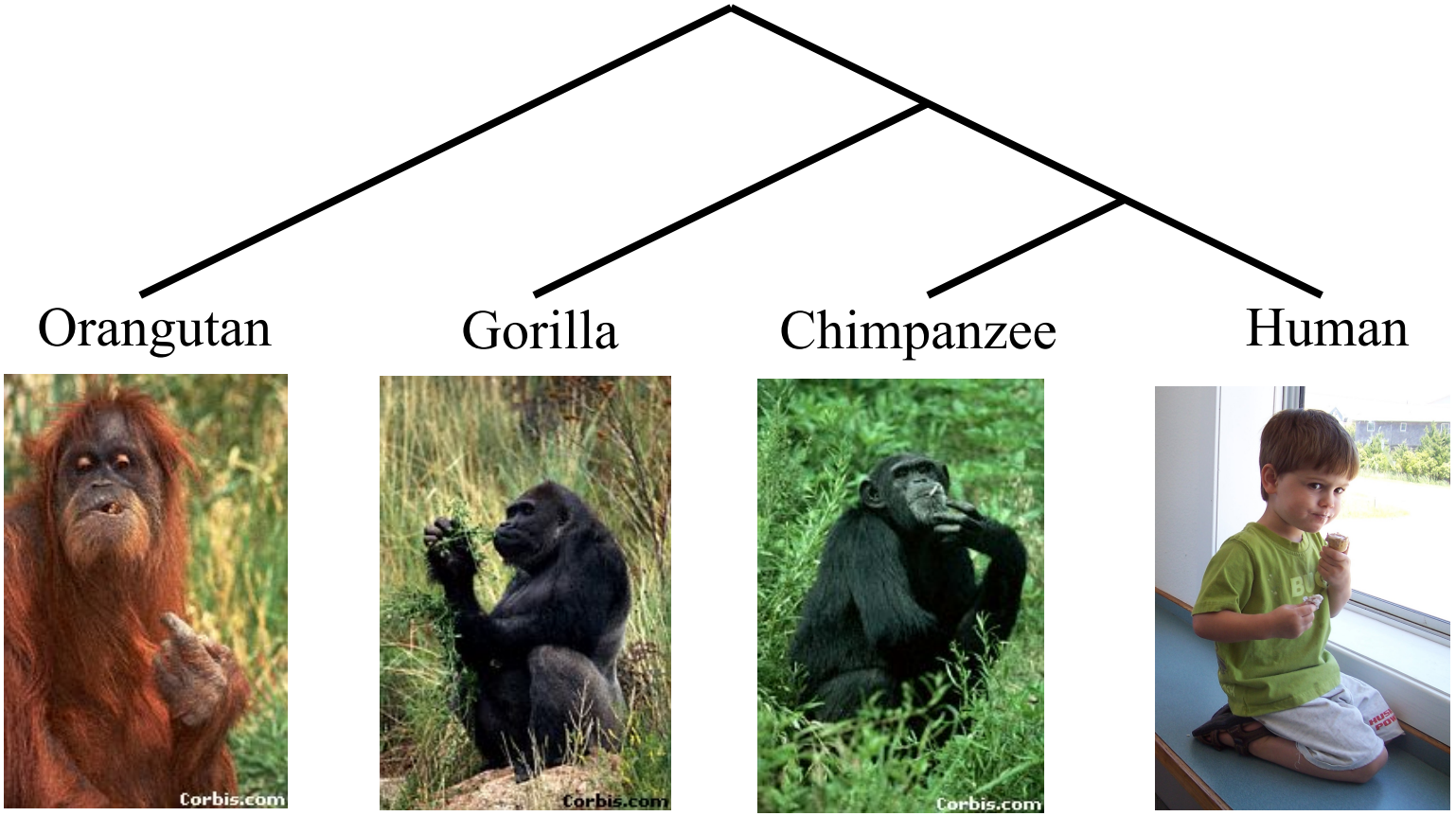


Gene tree in a species tree

Courtesy James Degnan



Species tree estimation: difficult, even for small datasets!

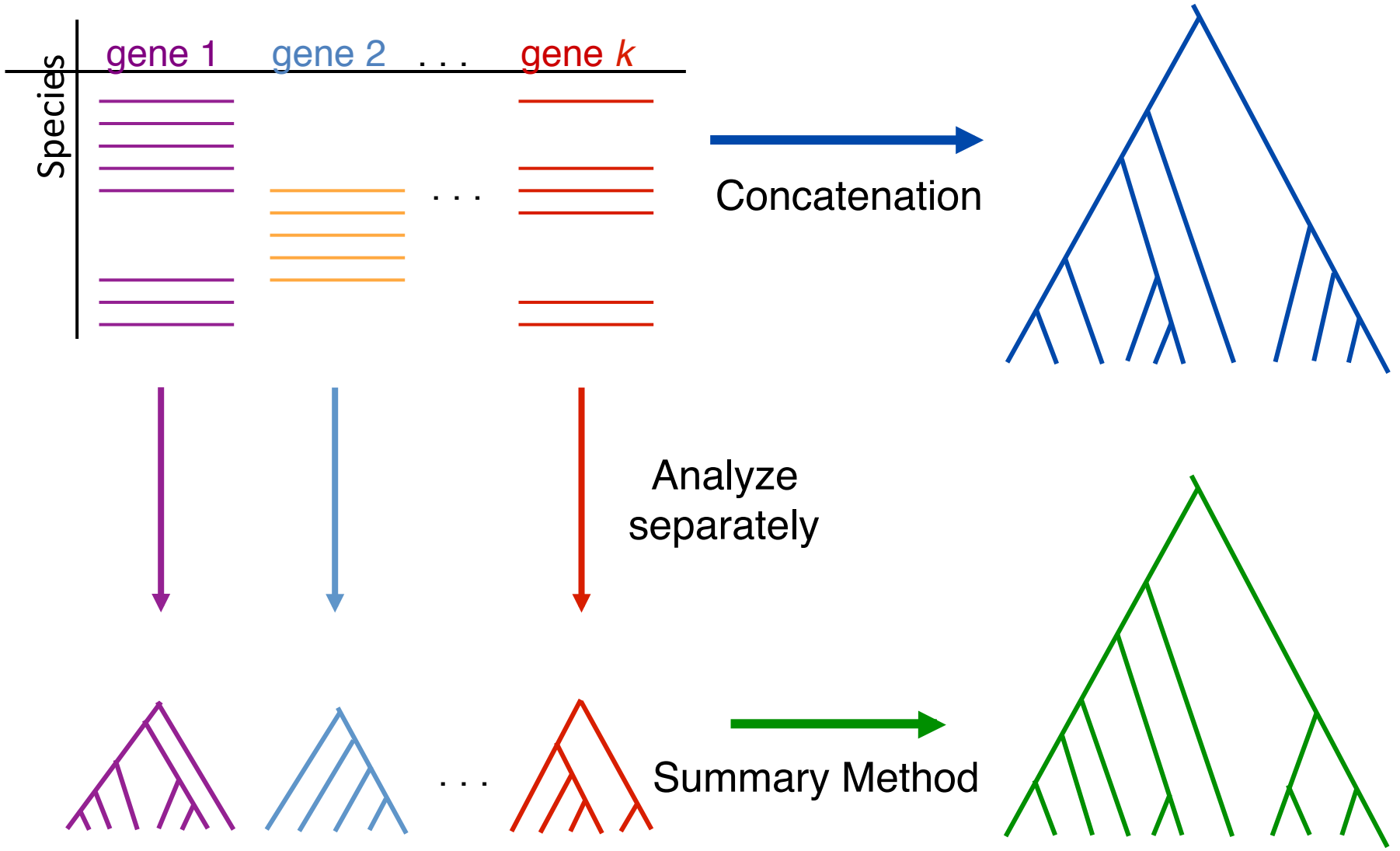


From the Tree of the Life Website, University of Arizona

Incomplete Lineage Sorting (ILS)

- 2000+ papers in 2013 alone
- Confounds phylogenetic analysis for many groups:
  - Hominids
  - Birds
  - Yeast
  - Animals
  - Toads
  - Fish
  - Fungi
- There is substantial debate about how to analyze phylogenomic datasets in the presence of ILS.

Two competing approaches



How to compute a species tree?

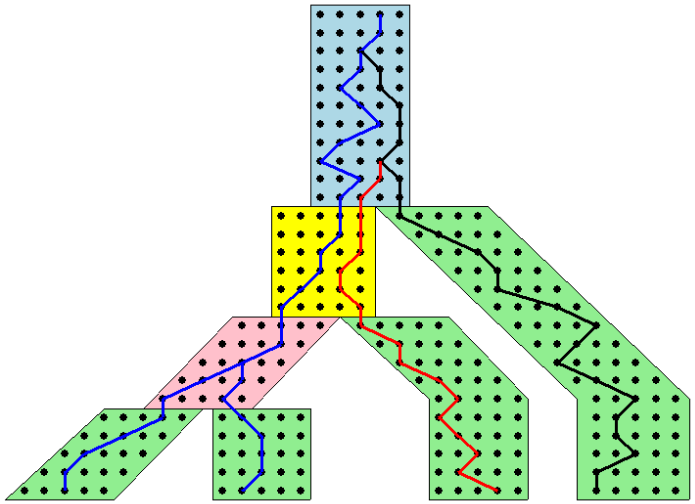
How to compute a species tree?

Techniques:

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

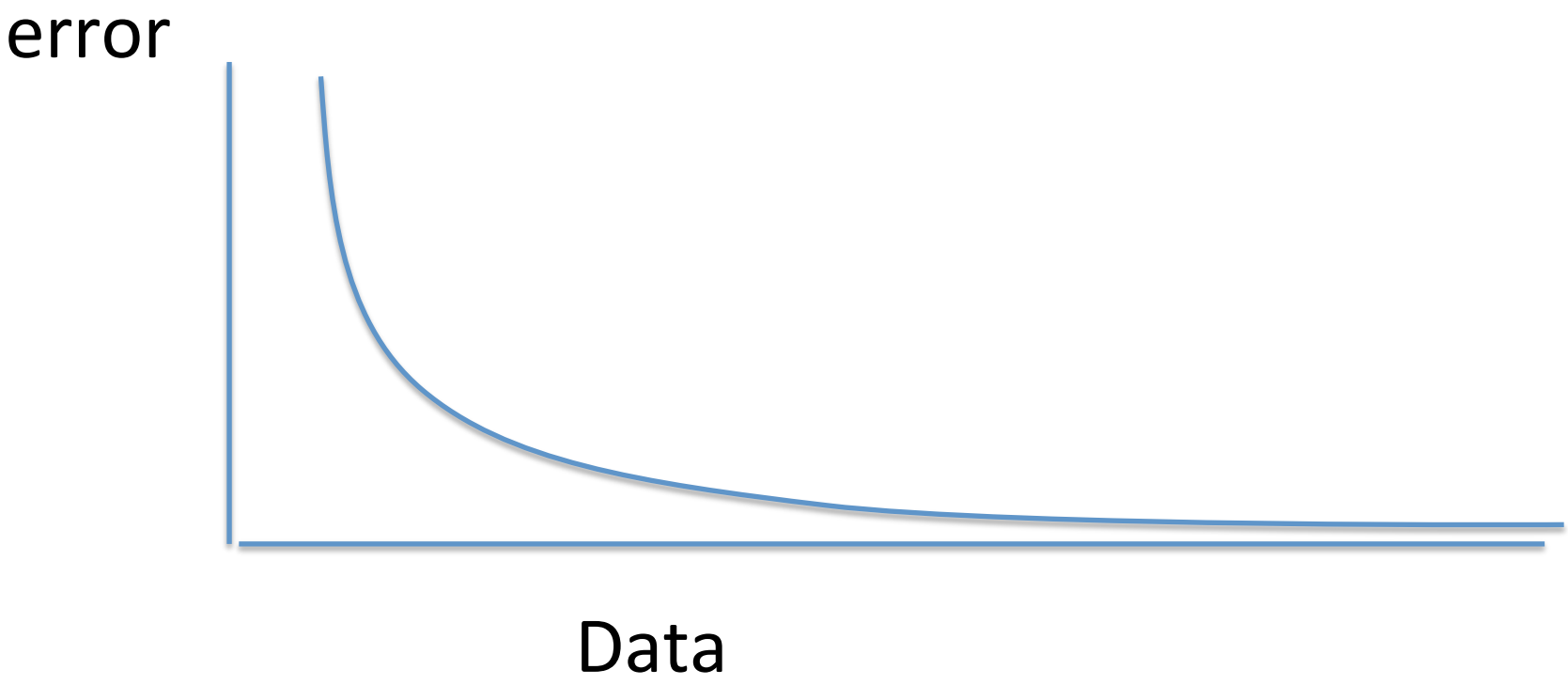
Key observation:

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

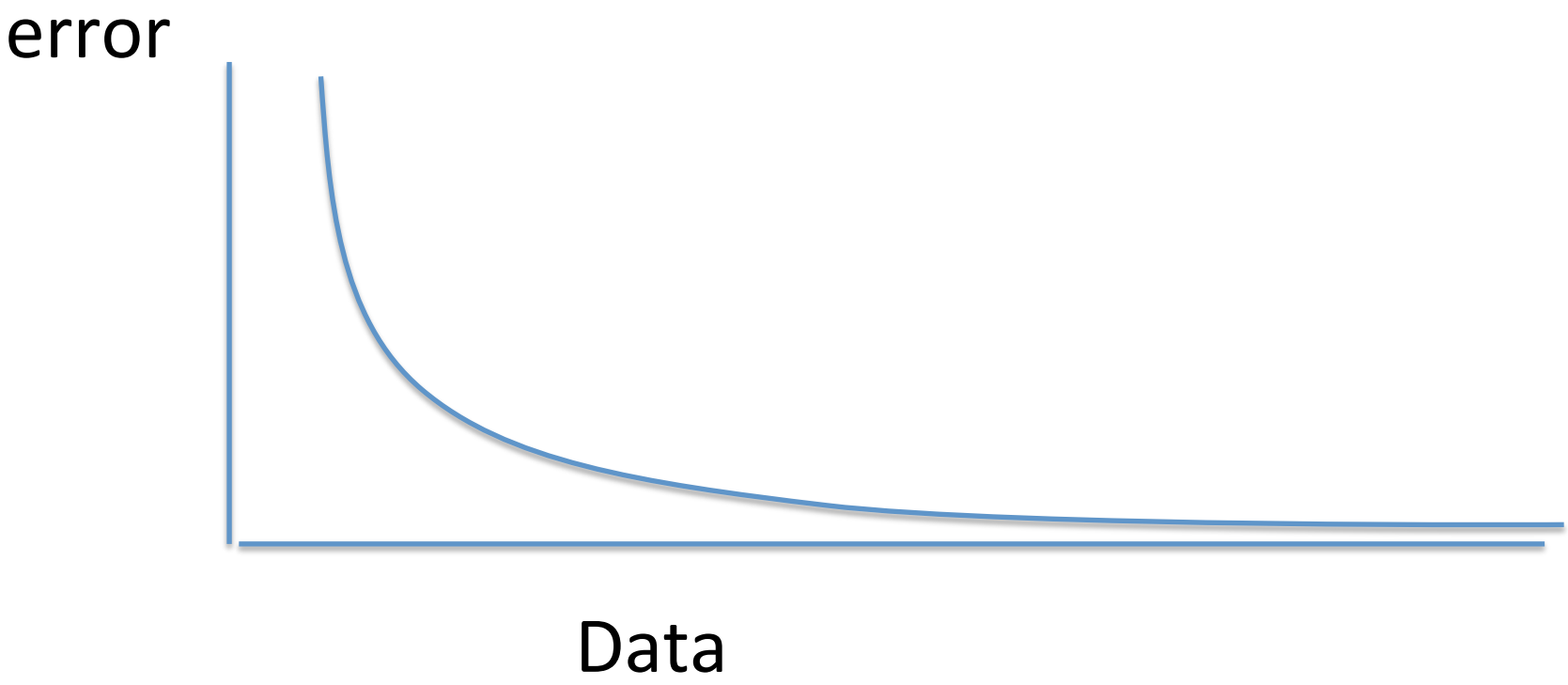


Courtesy James Degnan

Statistical Consistency



Statistical Consistency

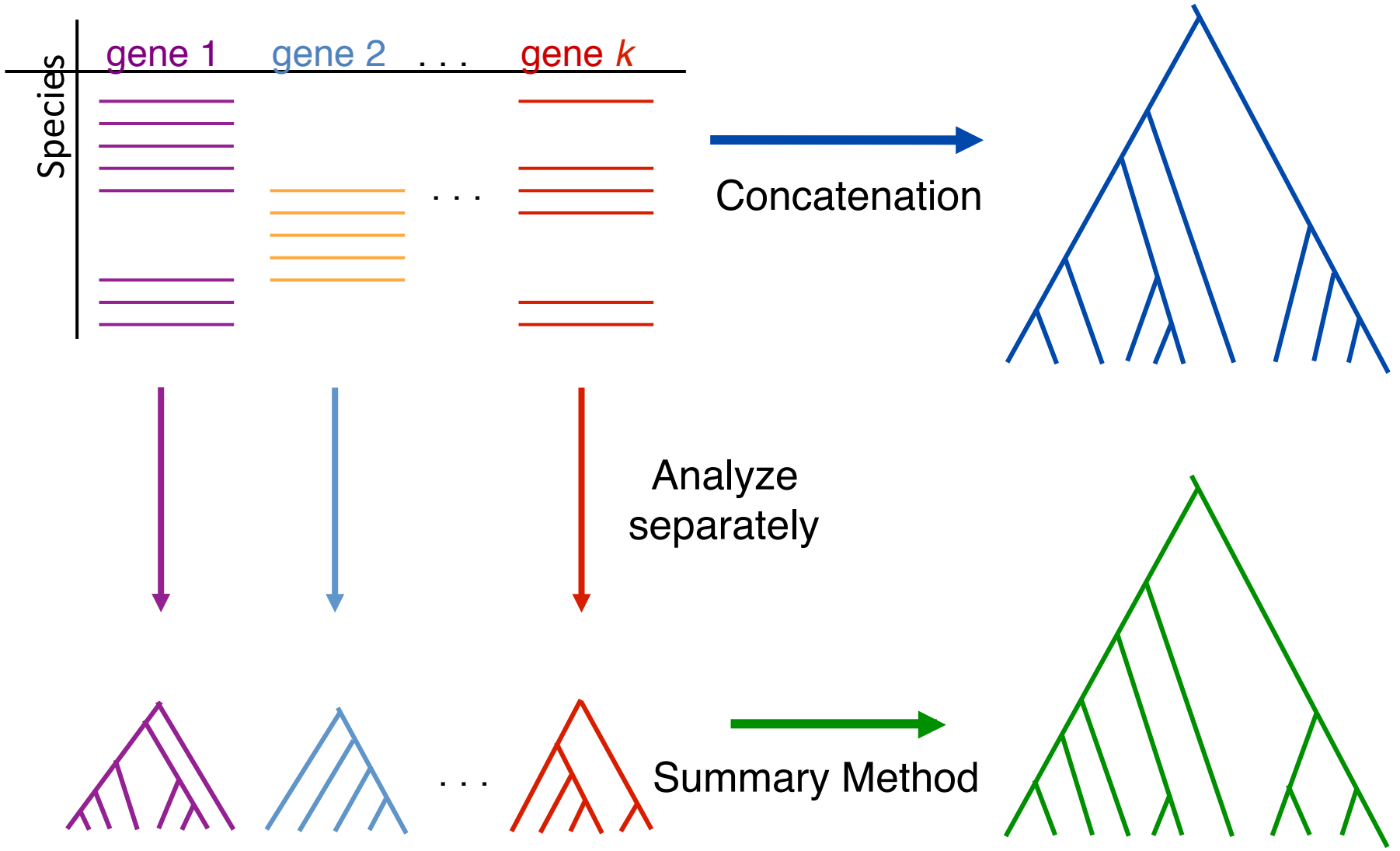


Data are gene trees, presumed to be randomly sampled <u>true gene trees.</u>

Statistically consistent under ILS?

- MP-EST (Liu et al. 2010): maximum likelihood estimation of rooted species tree – YES
- BUCKy-pop (Ané and Larget 2010): quartet-based Bayesian species tree estimation –YES
- MDC – NO
- Greedy – NO
- Concatenation under maximum likelihood – open
- MRP (supertree method) – open

Two competing approaches



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

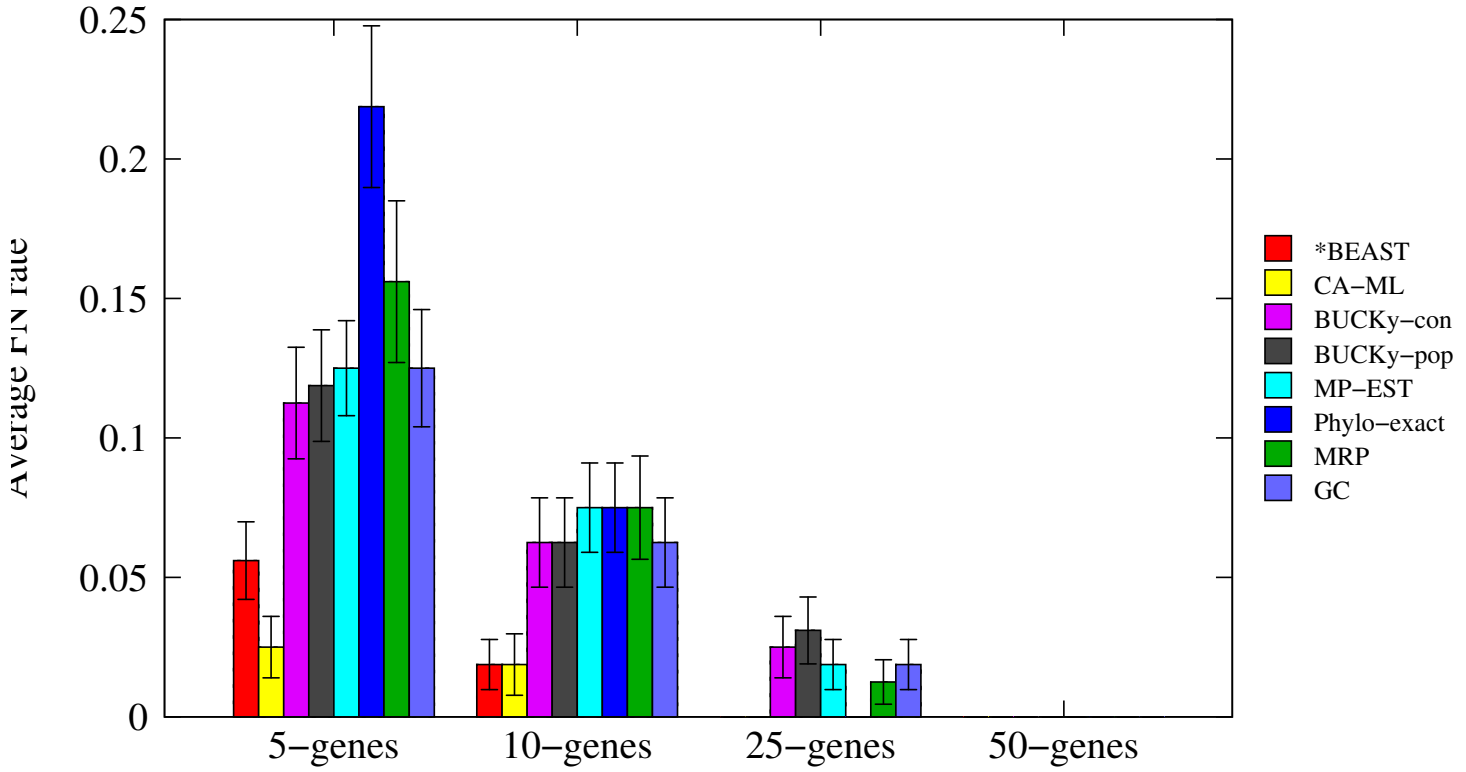
- In favor of concatenation
  - Reasonable results on data
  - High bootstrap support
  - Summary methods (that combine gene trees) can have poor support or miss well-established clades entirely
  - Some methods (such as *BEAST) are computationally too intensive to use

Is Concatenation Evil?

- Joseph Heled:
  - YES
- John Gatesy
  - No

- Data needed to held understand existing methods and their limitations
- Better methods are needed

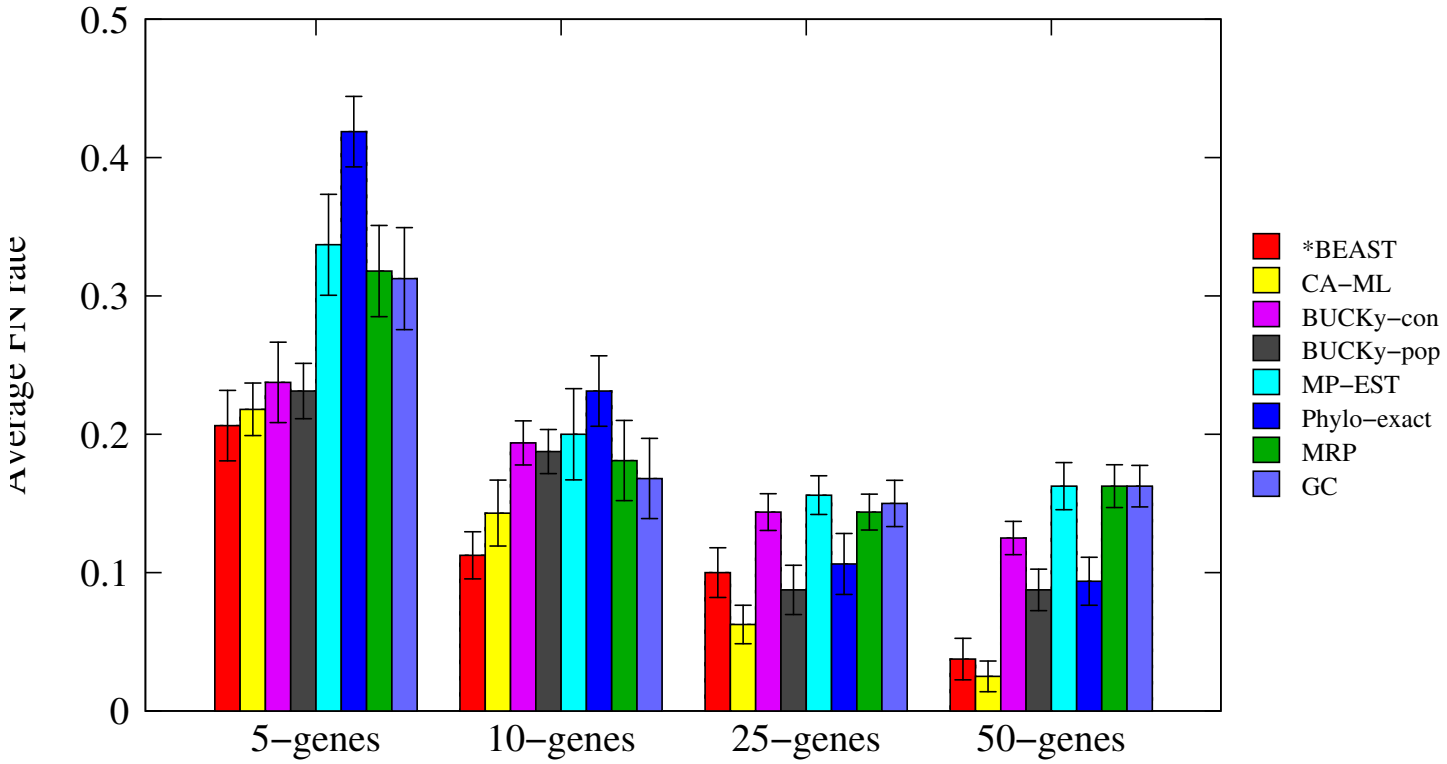
Results on 11-taxon datasets with weak ILS



***BEAST** more accurate than summary methods (MP-EST, BUCKy, etc)
CA-ML: (concatenated analysis) most accurate

Datasets from Chung and Ané, 2011
Bayzid & Warnow, Bioinformatics 2013

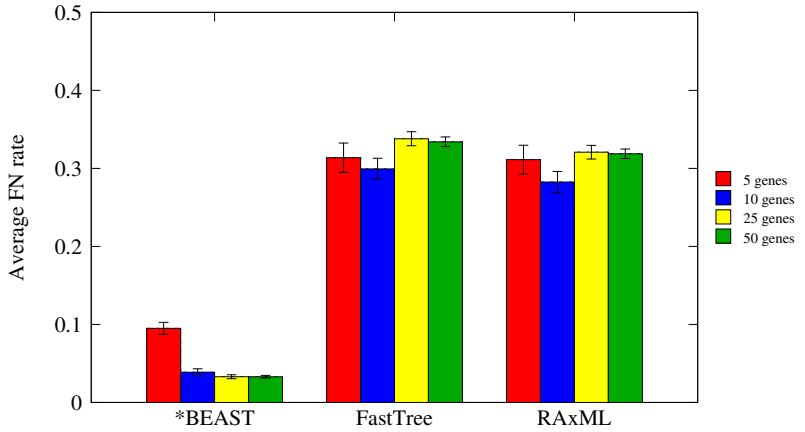
Results on 11-taxon datasets with strongILS



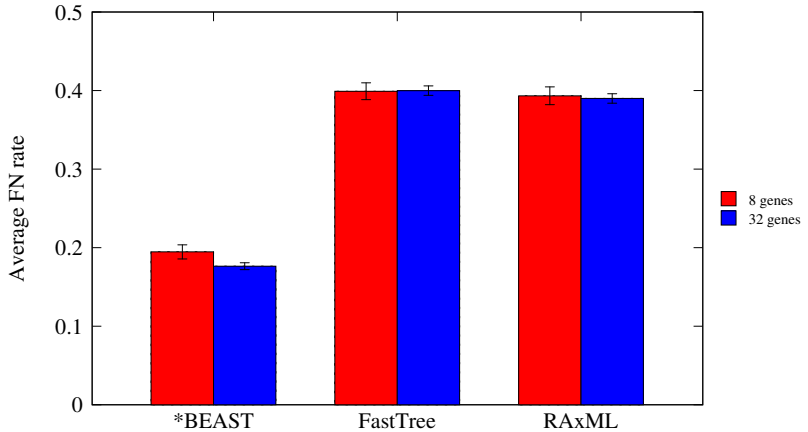
***BEAST** more accurate than summary methods (MP-EST, BUCKy, etc)
CA-ML: (concatenated analysis) also very accurate

Datasets from Chung and Ané, 2011
Bayzid & Warnow, Bioinformatics 2013

Gene Tree Estimation:
*BEAST vs. Maximum Likelihood



11-taxon weakILS datasets

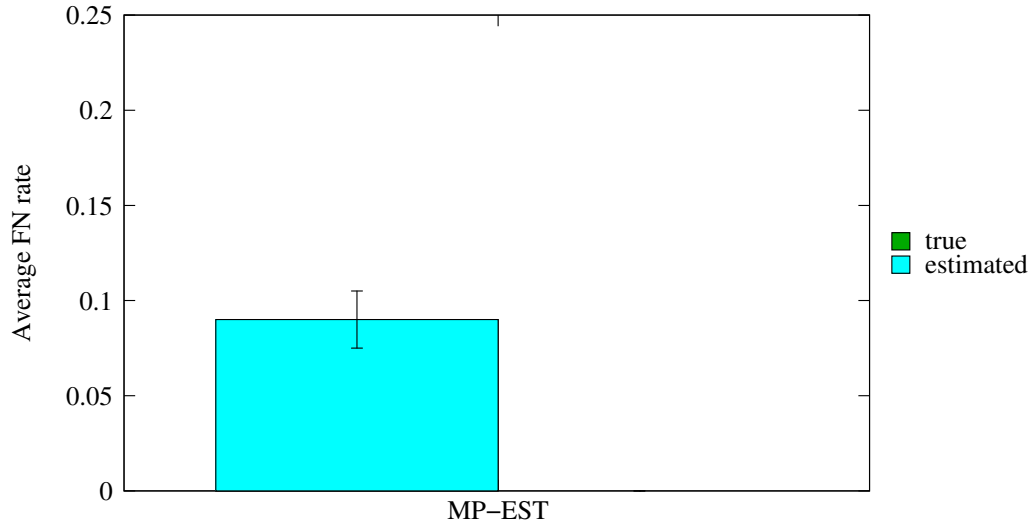


17-taxon (very high ILS) datasets

*BEAST produces more accurate gene trees than ML on gene sequence alignments

11-taxon datasets from Chung and Ané, Syst Biol 2012
17-taxon datasets from Yu, Warnow, and Nakhleh, JCB 2011

Impact of Gene Tree Estimation Error on MP-EST



MP-EST has no error on true gene trees, but
MP-EST has 9% error on estimated gene trees

Datasets: 11-taxon strongILS conditions with 50 genes

Similar results for other summary methods (MDC, Greedy, etc.).

Problem: poor gene trees

- Summary methods combine estimated gene trees, not true gene trees.

Problem: poor gene trees

- Summary methods combine estimated gene trees, not true gene trees.
- The individual gene sequence alignments in the 11-taxon datasets have poor phylogenetic signal, and result in poorly estimated gene trees.

Problem: poor gene trees

- Summary methods combine estimated gene trees, not true gene trees.
- The individual gene sequence alignments in the 11-taxon datasets have poor phylogenetic signal, and result in poorly estimated gene trees.
- Species trees obtained by combining poorly estimated gene trees have poor accuracy.

TYPICAL PHYLOGENOMICS PROBLEM: many poor gene trees

- Summary methods combine estimated gene trees, not true gene trees.
- The individual gene sequence alignments in the 11-taxon datasets have poor phylogenetic signal, and result in poorly estimated gene trees.
- Species trees obtained by combining poorly estimated gene trees have poor accuracy.

Addressing gene tree estimation error

- Get better estimates of the gene trees
- Restrict to subset of estimated gene trees
- Model error in the estimated gene trees
- Modify gene trees to reduce error
- “Bin-and-conquer”

Bin-and-Conquer?

1. Assign genes to "bins", creating "supergene alignments"
2. Estimate trees on each supergene alignment using maximum likelihood
3. Combine the supergene trees together using a summary method

Bin-and-Conquer?

1. Assign genes to "bins", creating "supergene alignments"

2. Estimate trees on each supergene alignment using maximum likelihood

3. Combine the supergene trees together using a summary method

Variants:

- Naïve binning (Bayzid and Warnow, Bioinformatics 2013)
- Statistical binning (Mirarab, Bayzid, Boussau, and Warnow, submitted)

Avian Phylogenomics Project

- Approx. 50 species, whole genomes
- 8000+ genes, UCEs
- Gene sequence alignments and trees computed using SATé (Liu et al., Science 2009 and Systematic Biology 2012)

- Approximately 14,000 "gene trees", all with very low support (exons average bootstrap support about 25%, introns about 47%)

- To concatenate or not to concatenate?

Avian Phylogeny

- GTRGAMMA Maximum likelihood analysis (RAxML) of 37 million basepair alignment (exons, introns, UCEs) – highly resolved tree with near 100% bootstrap support.

- More than 17 years of compute time, and used 256 GB. Run at HPC centers.

Avian Phylogenomics Project, in preparation

Avian Phylogeny

- GTRGAMMA Maximum likelihood analysis (RAxML) of 37 million basepair alignment (exons, introns, UCEs) – highly resolved tree with near 100% bootstrap support.

- More than 17 years of compute time, and used 256 GB. Run at HPC centers.

- Unbinned MP-EST on 14000+ genes: highly incongruent with the concatenated maximum likelihood analysis, poor bootstrap support.

Avian Phylogenomics Project, in preparation

Avian Phylogeny

- GTRGAMMA Maximum likelihood analysis (RAxML) of 37 million basepair alignment (exons, introns, UCEs) – highly resolved tree with near 100% bootstrap support.
- More than 17 years of compute time, and used 256 GB. Run at HPC centers.
- Unbinned MP-EST on 14000+ genes: highly incongruent with the concatenated maximum likelihood analysis, poor bootstrap support.
- Statistical binning version of MP-EST on 14000+ gene trees – highly resolved tree, largely congruent with the concatenated analysis, good bootstrap support

Avian Phylogenomics Project, in preparation

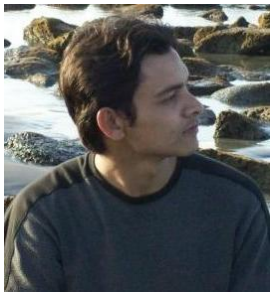
To consider

- Binning *reduces the amount* of data (number of gene trees) but can improve the accuracy of individual "supergene trees". The response to binning differs between methods. Thus, there is a trade-off between data *quantity* and *quality, and not all methods respond the same to the trade-off*.

- We know very little about the impact of data *error* on methods. We do not even have proofs of statistical consistency in the presence of data error.

Warnow Laboratory





PhD students: Siavash Mirarab*, Nam Nguyen, and Md. S. Bayzid**
Undergrad: Keerthana Kumar
Lab Website: http://www.cs.utexas.edu/users/phylo

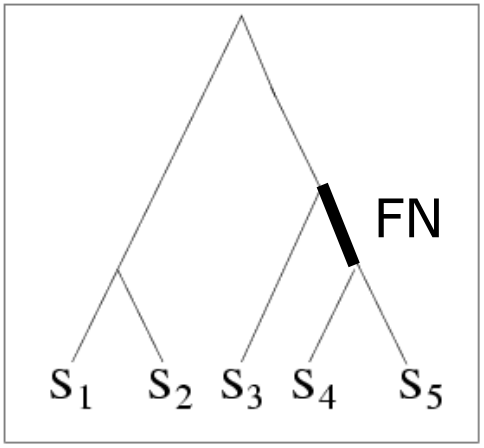
Funding: Guggenheim Foundation, Packard, NSF, Microsoft Research New England, David Bruton Jr. Centennial Professorship, and TACC (Texas Advanced Computing Center)

TACC and UTCS computational resources

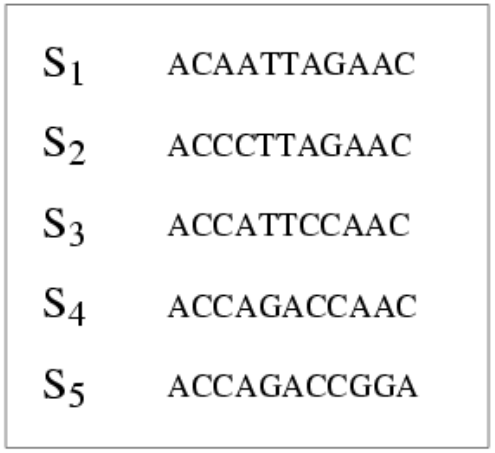
* Supported by HHMI Predoctoral Fellowship

** Supported by Fulbright Foundation Predoctoral Fellowship

Quantifying Error



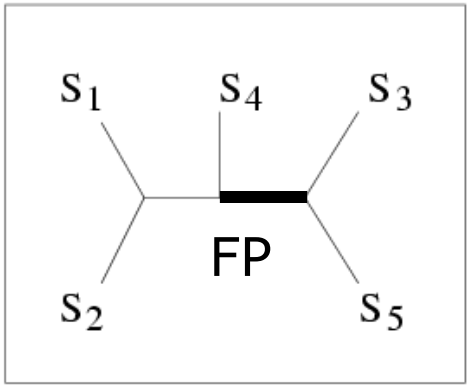
TRUE TREE



DNA SEQUENCES

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

50% error rate



INFERRED TREE

Avian Simulation – 14,000 genes

- **MP-EST:**
  - Unbinned ~ 11.1% error
  -
- **Greedy:**
  - Unbinned ~ 26.6% error
  -

- 8250 exon-like genes (27% avg. bootstrap support)
- 3600 UCE-like genes (37% avg. bootstrap support)
- 2500 intron-like genes (51% avg. bootstrap support)

Avian Simulation – 14,000 genes

- **MP-EST:**
  - Unbinned ~ 11.1% error
  - Binned ~ 6.6% error
- **Greedy:**
  - Unbinned ~ 26.6% error
  - Binned ~ 13.3% error

- 8250 exon-like genes (27% avg. bootstrap support)
- 3600 UCE-like genes (37% avg. bootstrap support)
- 2500 intron-like genes (51% avg. bootstrap support)

Avian Phylogeny

- GTRGAMMA Maximum likelihood analysis (RAxML) of 37 million basepair alignment (exons, introns, UCEs) – highly resolved tree with near 100% bootstrap support.

- More than 17 years of compute time, and used 256 GB. Run at HPC centers.

- Unbinned MP-EST on 14000+ genes: highly incongruent with the concatenated maximum likelihood analysis, poor bootstrap support.

Avian Phylogenomics Project, in preparation

Basic 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?

Additional Statistical Questions

- Trade-off between data quality and quantity
- Impact of data selection
- Impact of data error
- Performance guarantees on finite data (e.g., prediction of error rates as a function of the input data and method)

We need a solid mathematical framework for these problems.

Summary

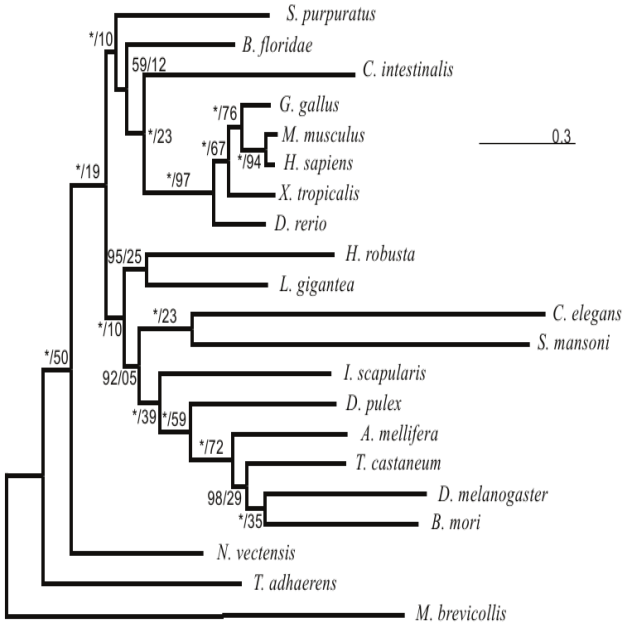
- DCM1-NJ: an absolute fast converging (afc) method, uses chordal graph theory and probabilistic analysis of algorithms to prove performance guarantees

- Binning: species tree estimation from multiple genes, can improve coalescent-based species tree estimation methods.

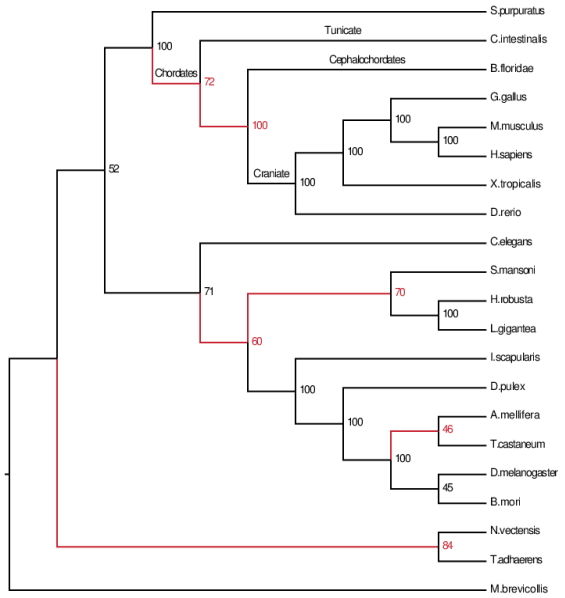
- New questions in phylogenetic estimation about impact of error in input data.

Metazoa Dataset from Salichos & Rokas - Nature 2013 225 genes and 21 species

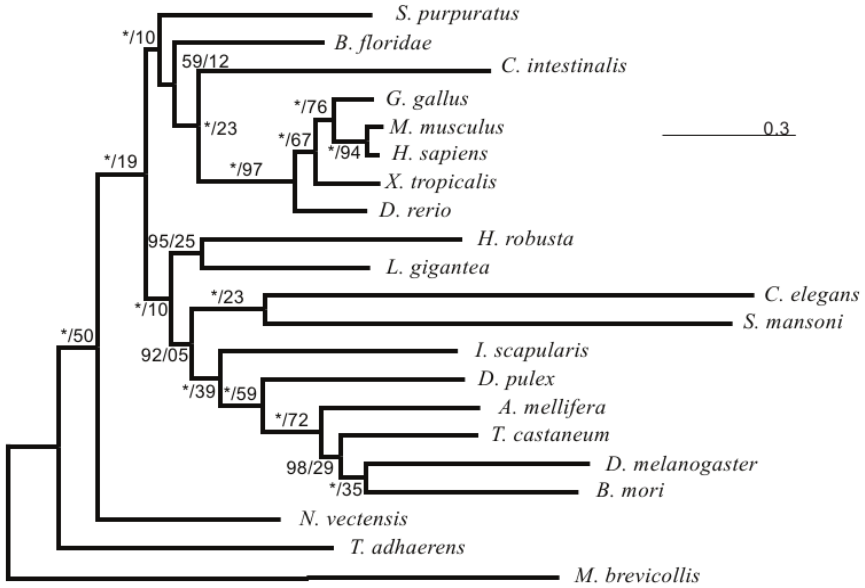
- UnBinned MP-EST compared to Concatenation using RAxML
  - Poor bootstrap support
  - Substantial conflict with concatenation (red is conflict - green/black is congruence)
  - Strongly rejects (Tunicate,Craniate), a subgroup that is strongly supported in the literature [Bourlat, Sarah J., et al *Nature* 444.7115 (2006); Delsuc, Frédéric, et al. *Genesis* 46.11 (2008); Singh, Tiratha R., et al. *BMC genomics* 10.1 (2009): 534.]



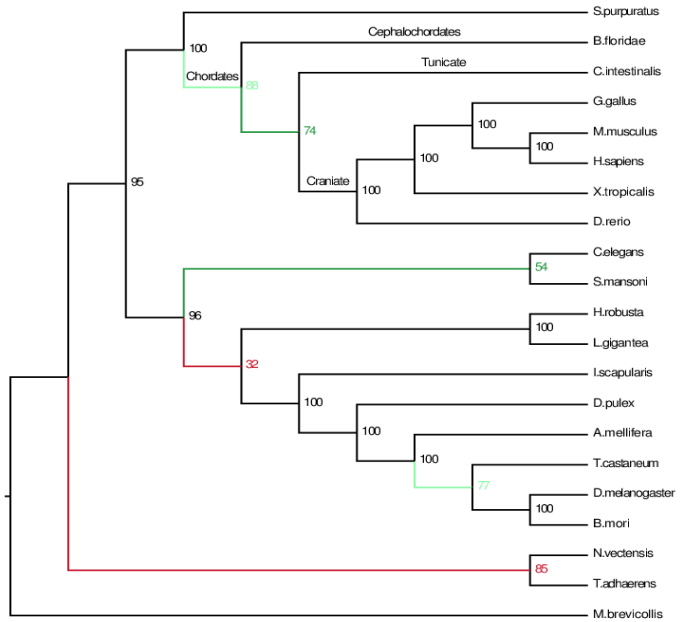
Concatenation



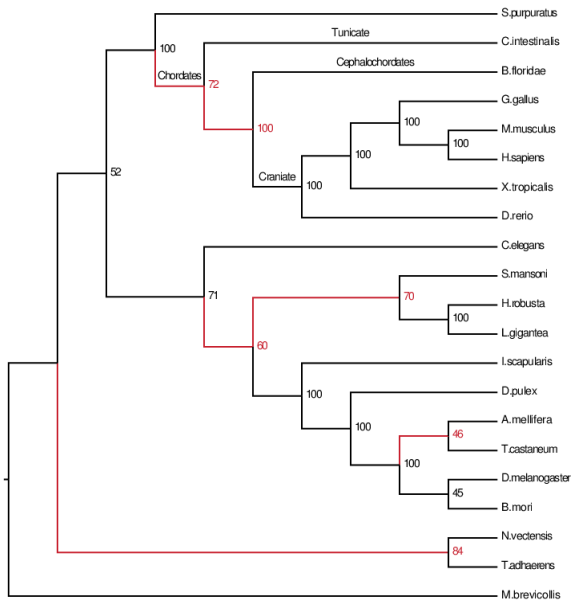
Unbinned MPEST



RAxML on combined datamatrix



Binned MP-EST

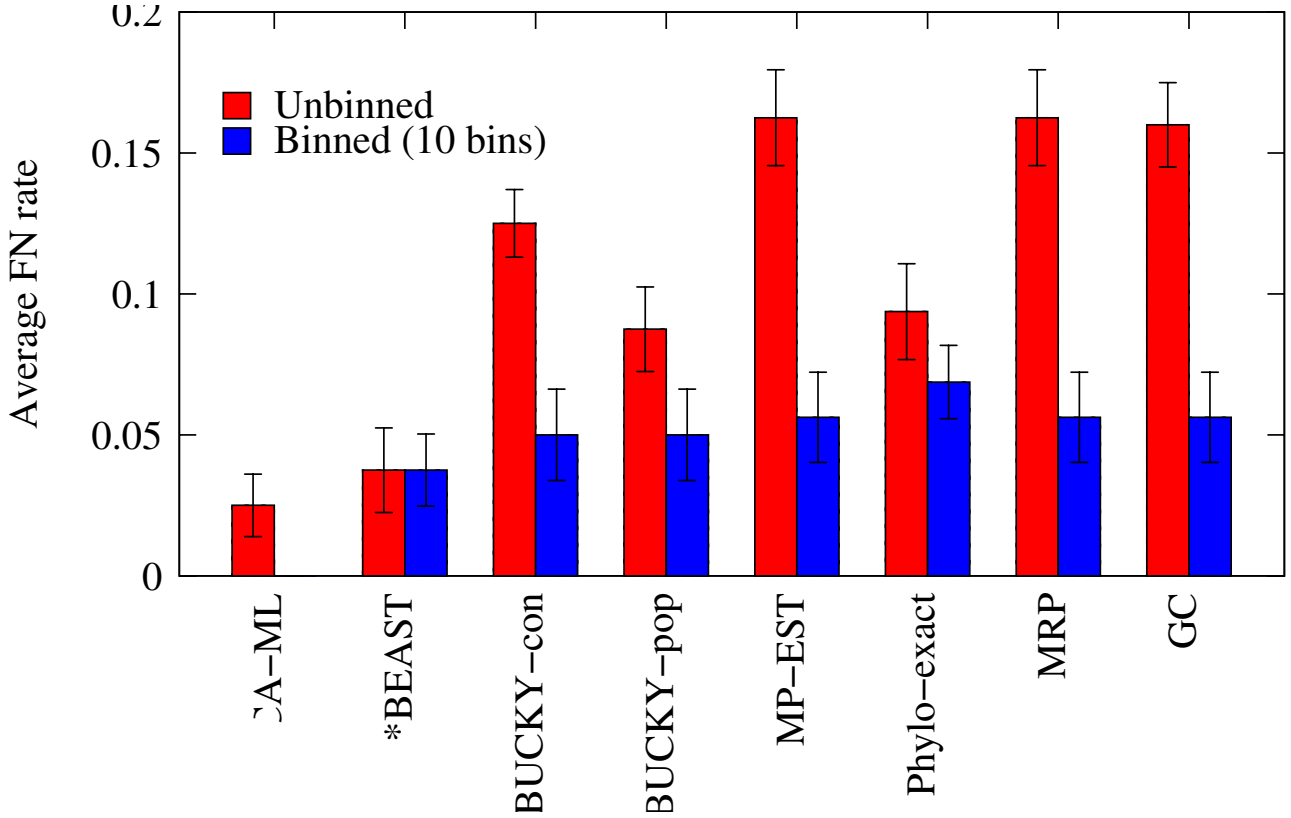


MP-EST unbinned

Binned vs. unbinned analyses

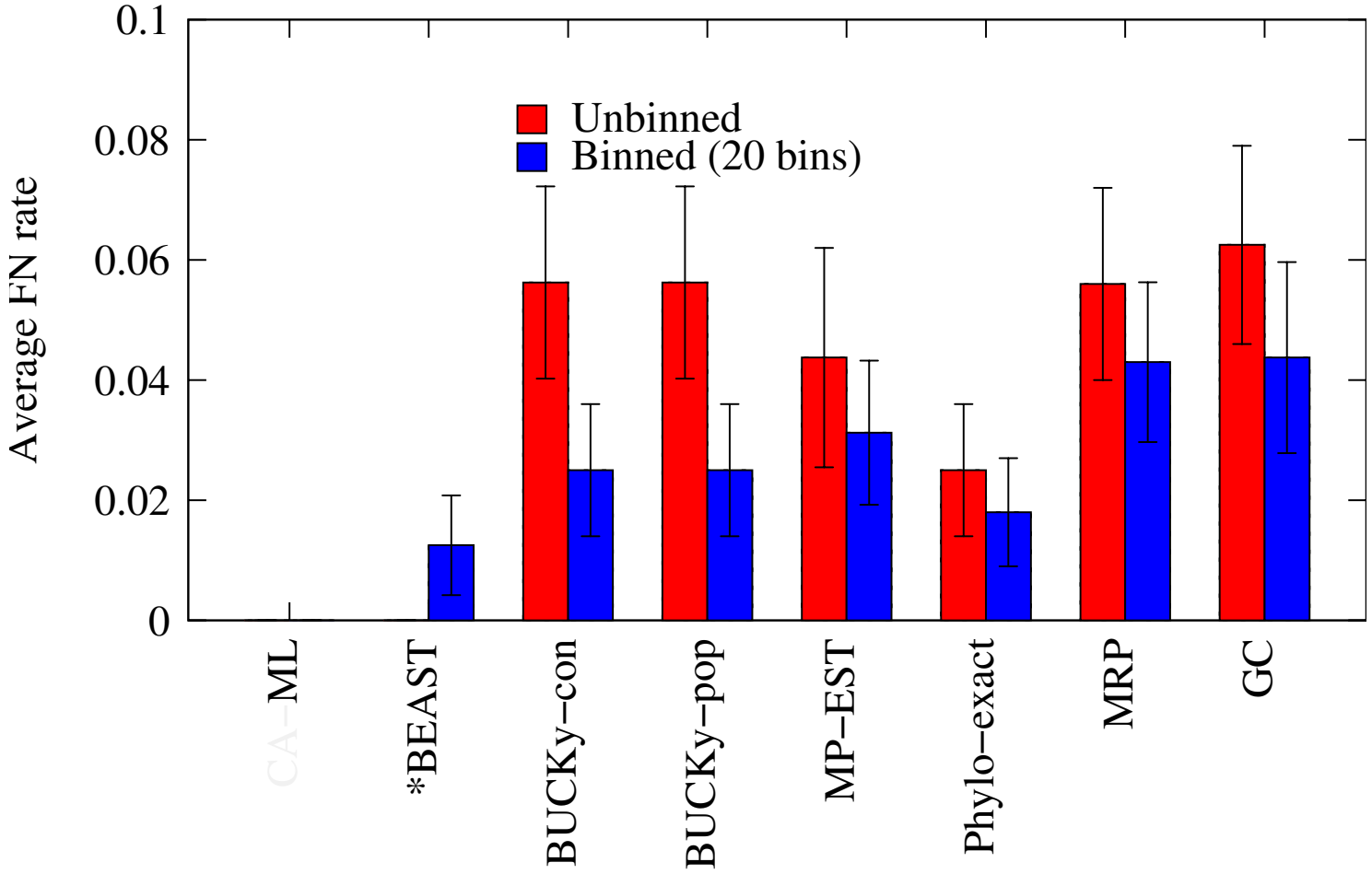
- 75%-threshold for binning
- Number of species: 21 for both
- Number of “genes”
  - Unbinned: 225 genes
  - Binned:17 supergenes
- Gene tree average bootstrap support
  - Unbinned: 47%
  - Binned: 78%
- Species tree bootstrap support
  - Unbinned: avg 83%, 11 above 75%, 10 above 90%
  - Binned: avg 89%, 15 above 75%, 12 above 90%

Naïve binning vs. unbinned: 50 genes



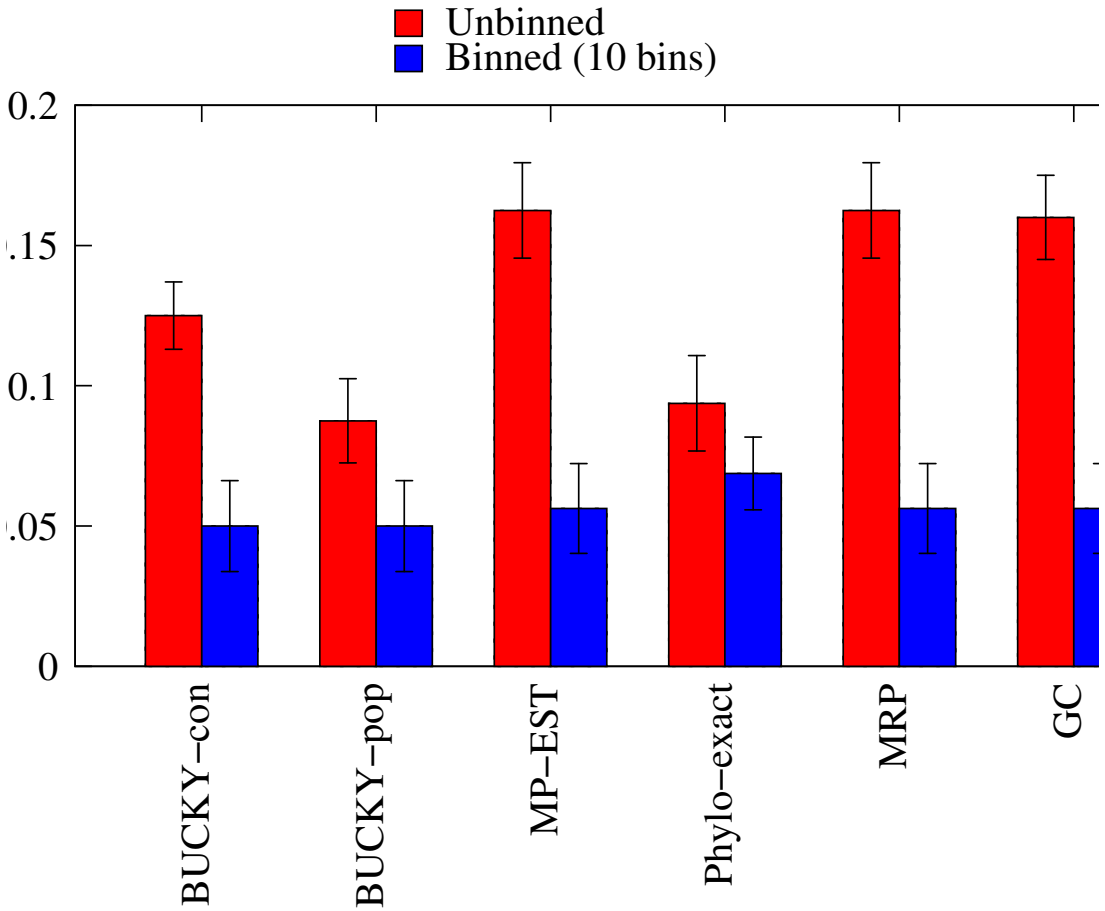
Bayzid and Warnow, Bioinformatics 2013
11-taxon strongILS datasets with 50 genes, 5 genes per bin

Naïve binning vs. unbinned, 100 genes



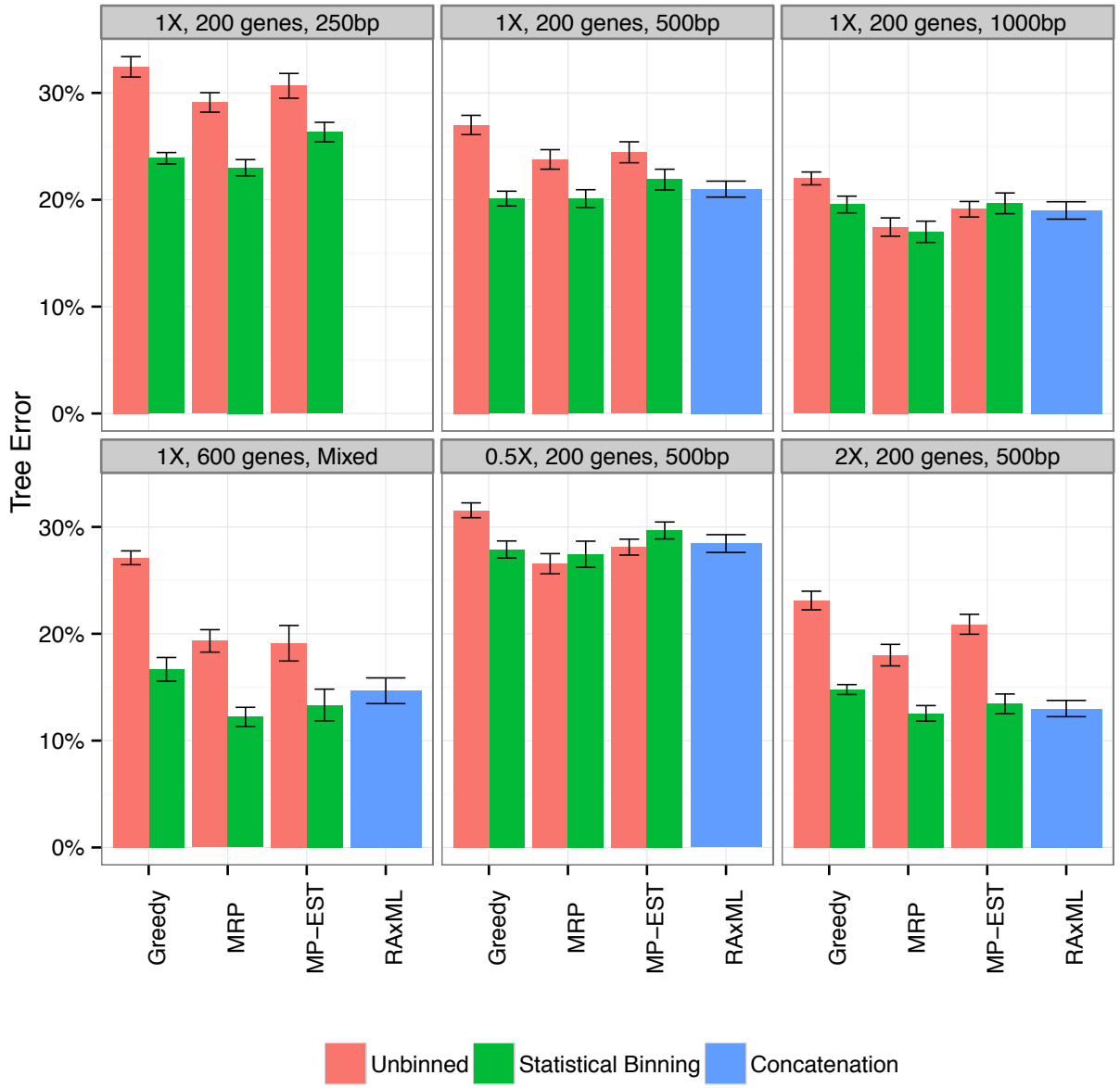
*BEAST did not converge on these datasets, even with 150 hours.
With binning, it converged in 10 hours.

Naïve binning vs. unbinned: 50 genes

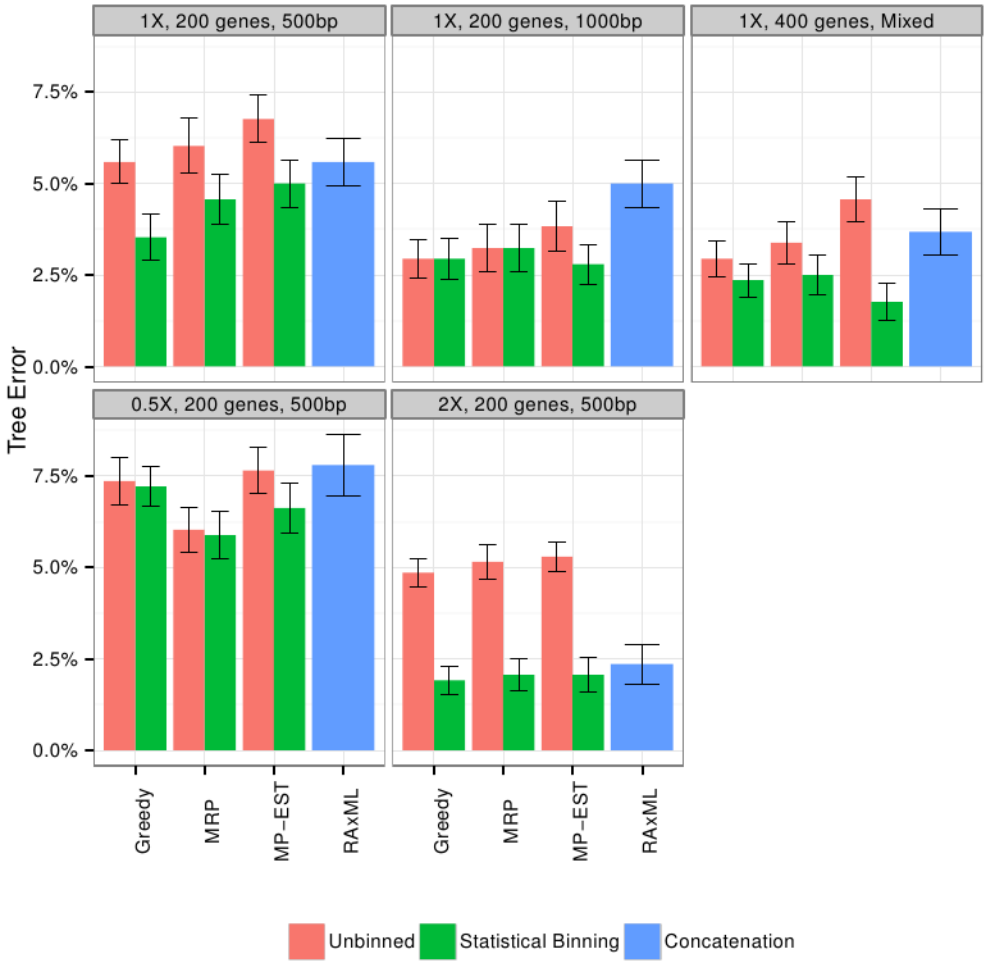


Bayzid and Warnow, Bioinformatics 2013
11-taxon strongILS datasets with 50 genes, 5 genes per bin

Avian Simulation study – binned vs. unbinned, and RAxML



Mammals Simulation



Avian Simulation