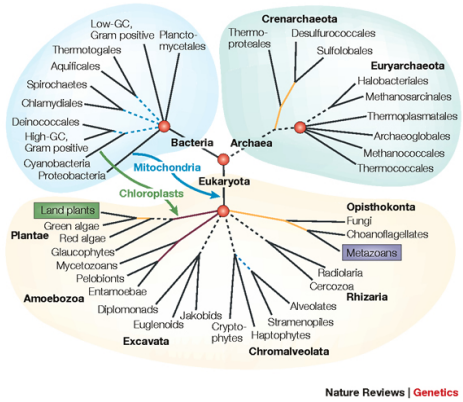


Nature Reviews | Genetics

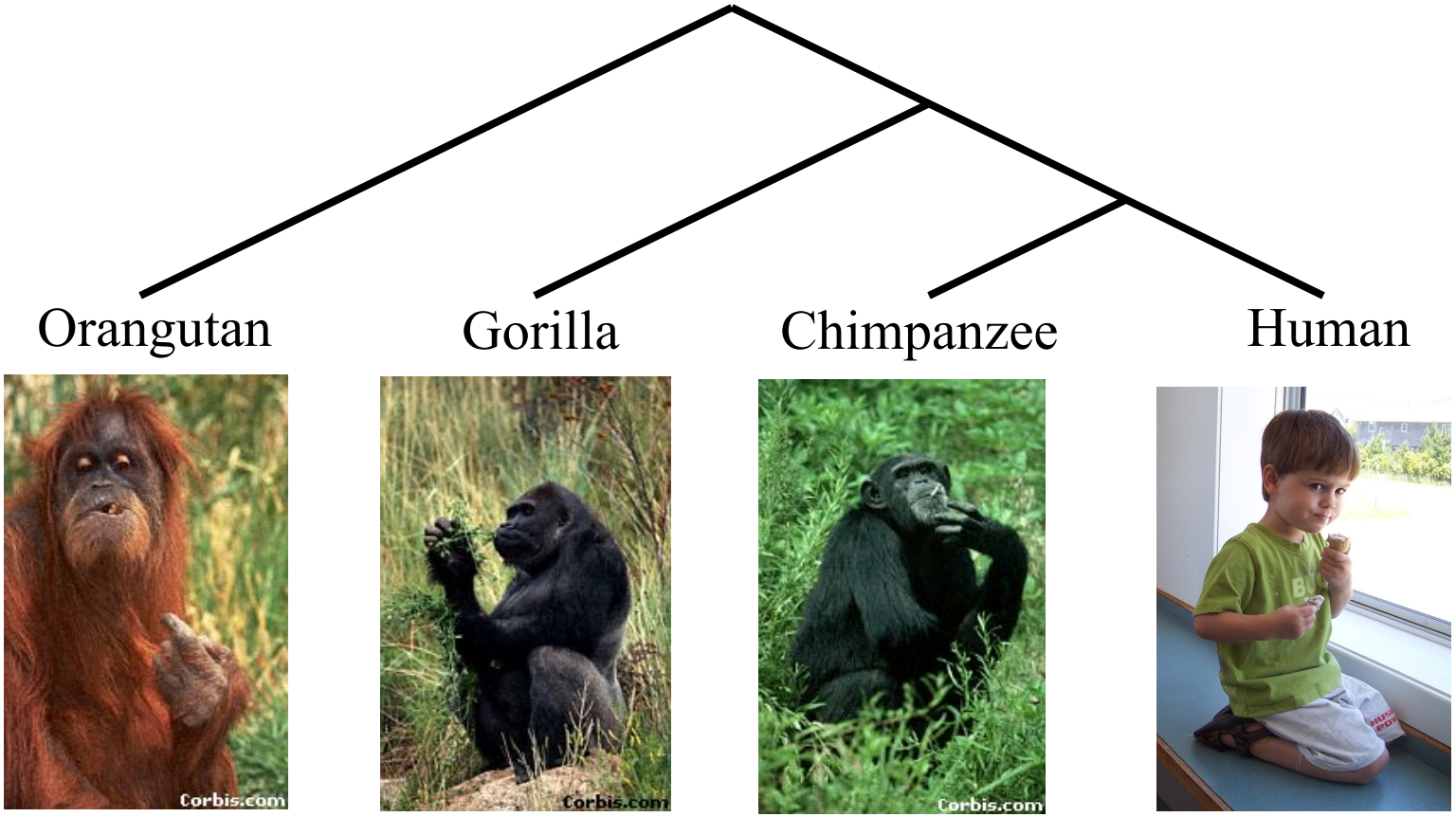
BBCA: Improving the scalability of *BEAST using random binning

Tandy Warnow

The University of Illinois at Urbana-Champaign

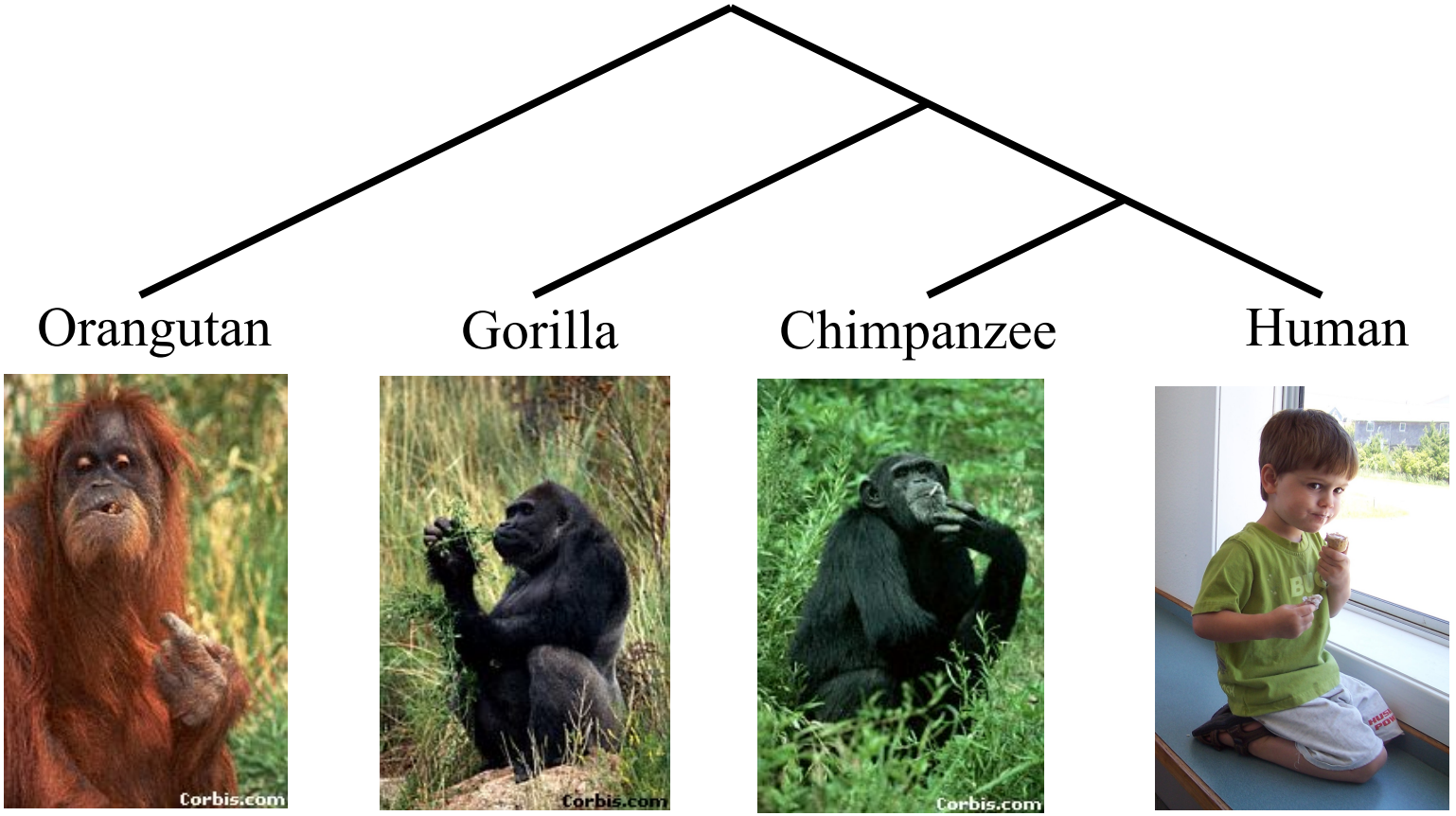
Co-authors: Theo Zimmermann (France) and Siavash Mirarab (Texas)

Phylogeny
(evolutionary tree)



From the Tree of the Life Website,
University of Arizona

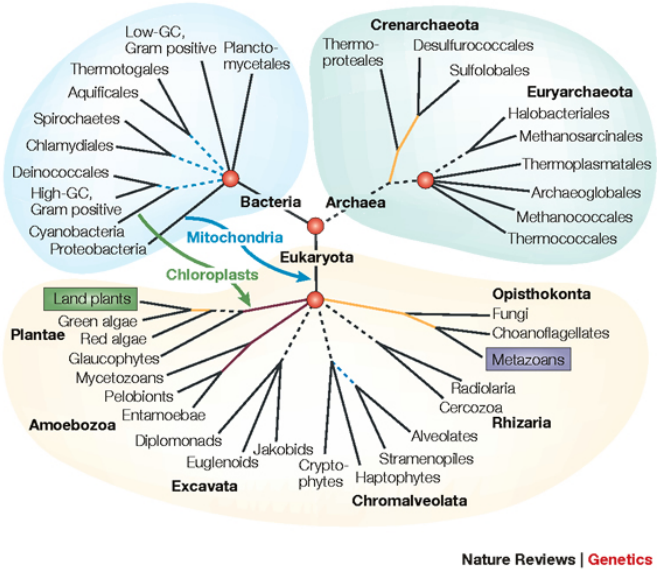
Sampling multiple genes from multiple species



From the Tree of the Life Website,
University of Arizona

Phylogenomics

(Phylogenetic estimation from whole genomes)



Nature Reviews | Genetics

Using multiple genes

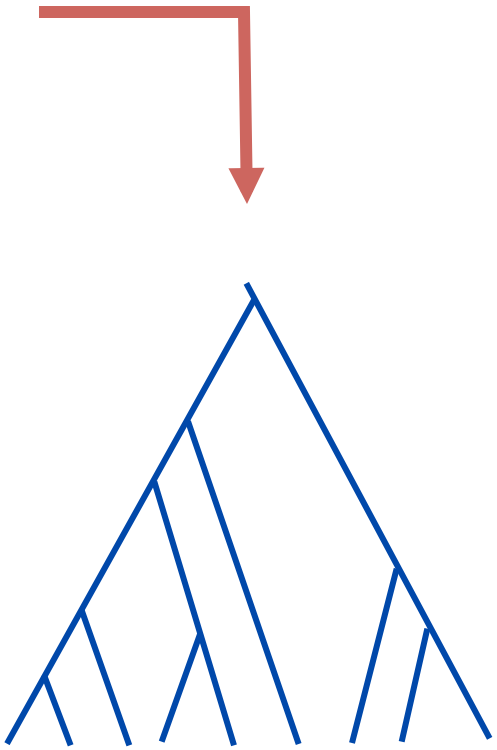
	gene 1
S_1	TCTAATGGAA
S_2	GCTAAGGGAA
S_3	TCTAAGGGAA
S_4	TCTAACGGAA
S_7	TCTAATGGAC
S_8	TATAACGGAA

	gene 2
S_4	GGTAACCCTC
S_5	GCTAAACCTC
S_6	GGTGACCATC
S_7	GCTAAACCTC

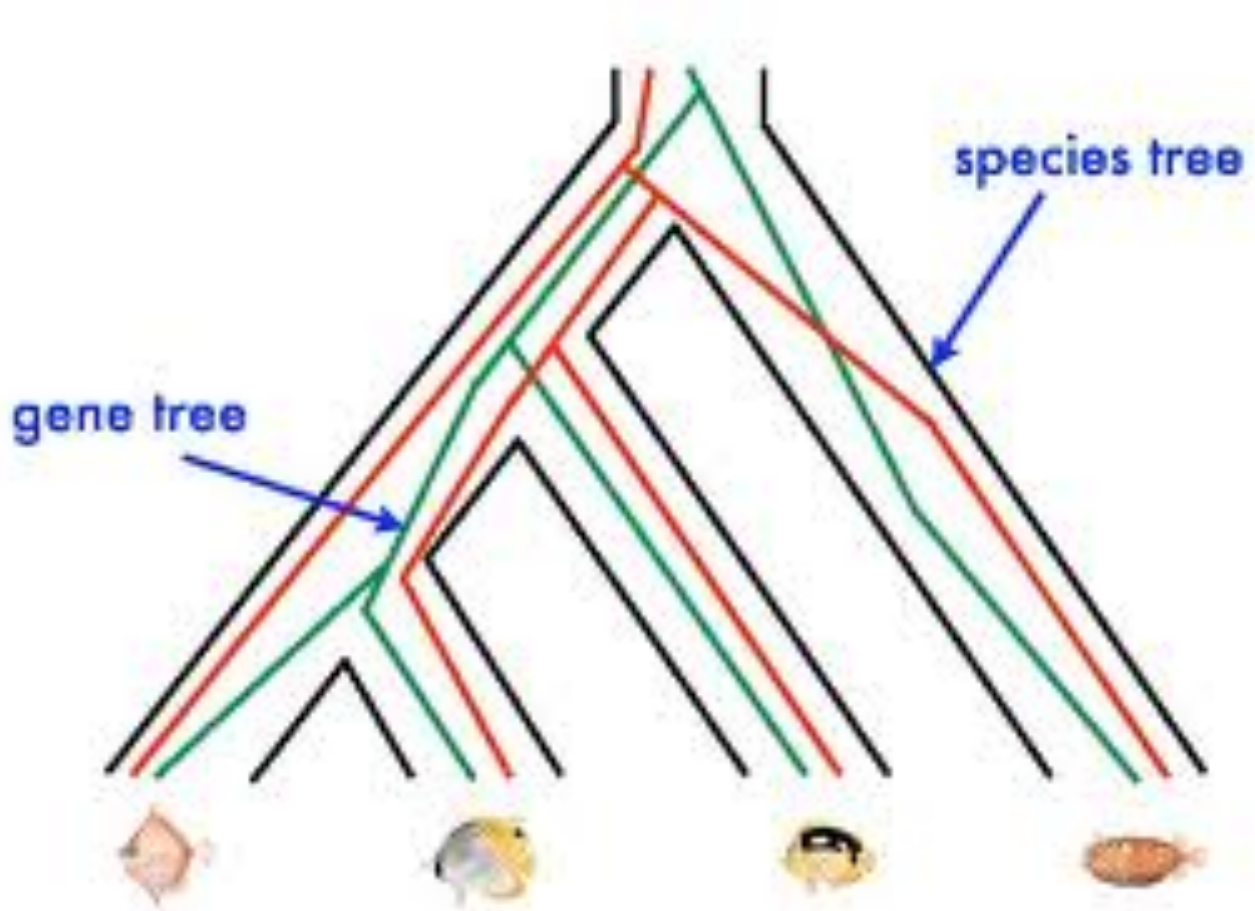
	gene 3
S_1	TATTGATACA
S_3	TCTTGATACC
S_4	TAGTGATGCA
S_7	TAGTGATGCA
S_8	CATTCATACC

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)

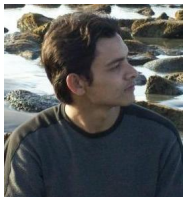


1KP: Thousand Transcriptome Project







G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

N. W[illegible]
Northwestern

N. Matasci
iPlant

T. Warnow,
UT-Austin

S. Mirarab,
UT-Austin

N. Nguyen,
UT-Austin

Md. S.Bayzid
UT-Austin

Gene Tree Incongruence

- 1200 plant transcriptomes
- More than 13,000 gene families (most not single copy)
- Multi-institutional project (10+ universities)
- iPLANT (NSF-funded cooperative)
- Gene sequence alignments and trees computed using SATe (Liu et al., Science 2009 and Systematic Biology 2012)

Avian Phylogenomics Project

E Jarvis, HHMI

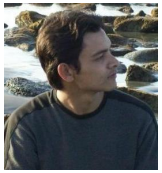
MTP Gilbert, Copenhagen

G Zhang, BGI

T. Warnow UT-Austin

S. Mirarab UT-Austin

Md. S. Bayzid, UT-Austin



Gene Tree Incongruence

Plus many many other people…

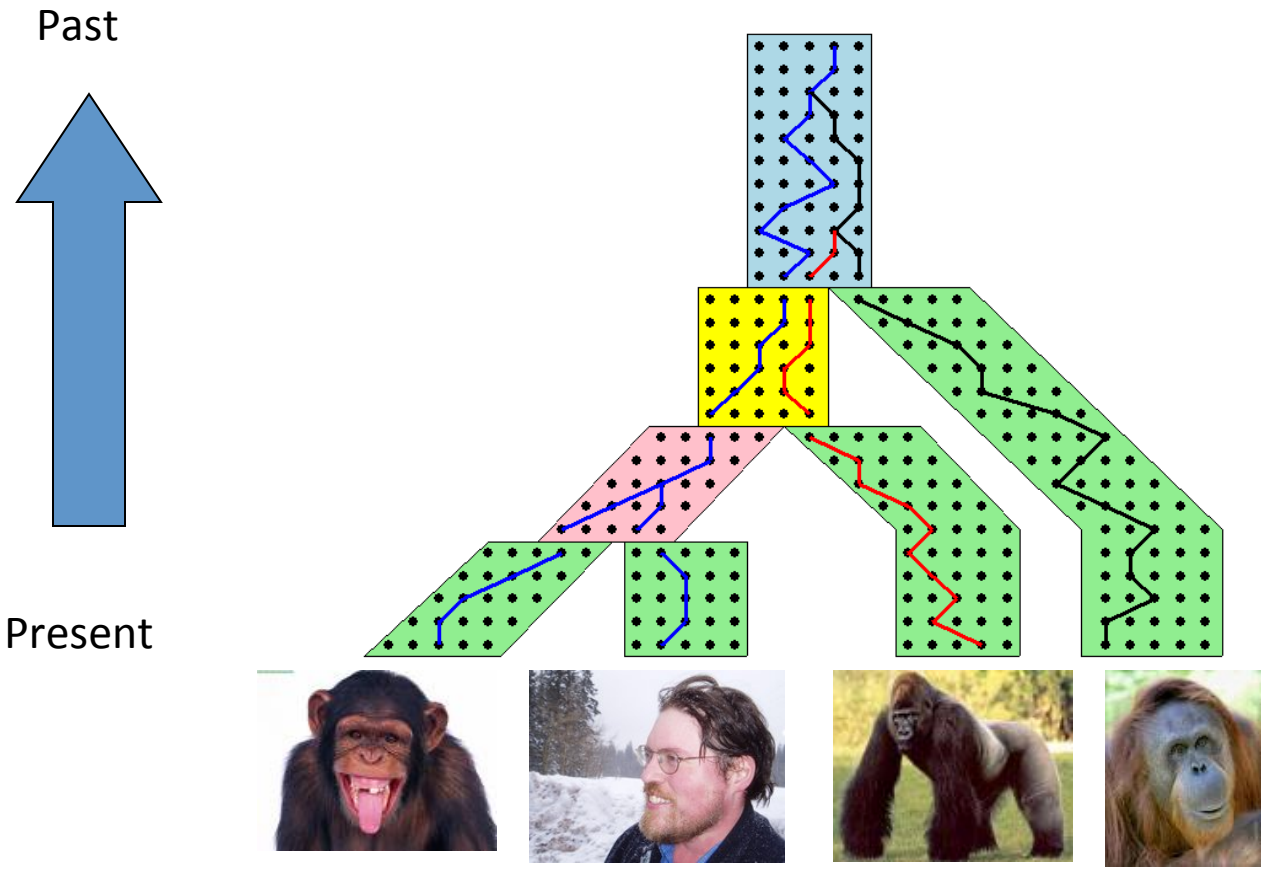
- Approx. 50 species, whole genomes
- 8000+ genes, UCEs
- Gene sequence alignments computed using SATé (Liu et al., Science 2009 and Systematic Biology 2012)

Gene Tree Incongruence

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

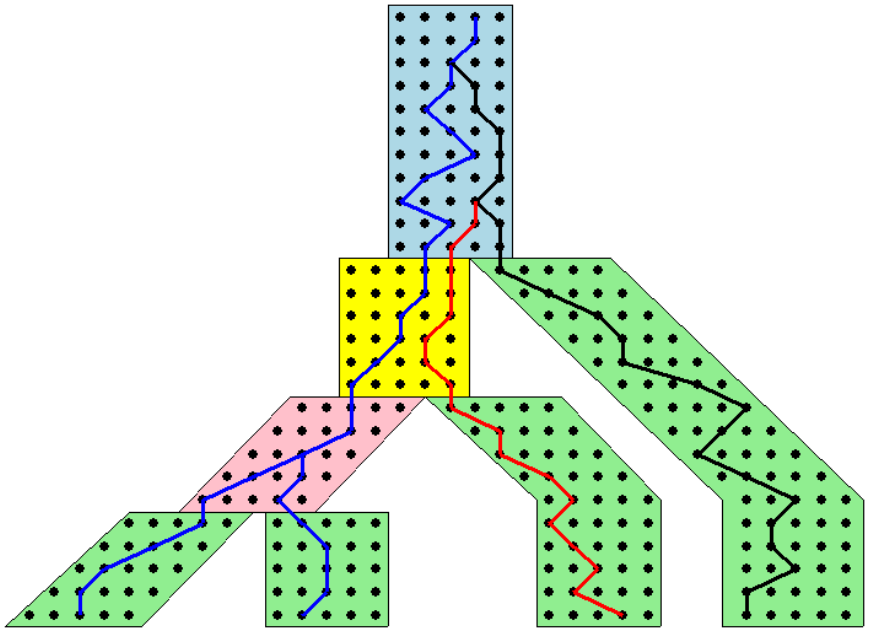
The Coalescent

Courtesy James Degnan



Gene tree in a species tree

Courtesy James Degnan



Lineage Sorting

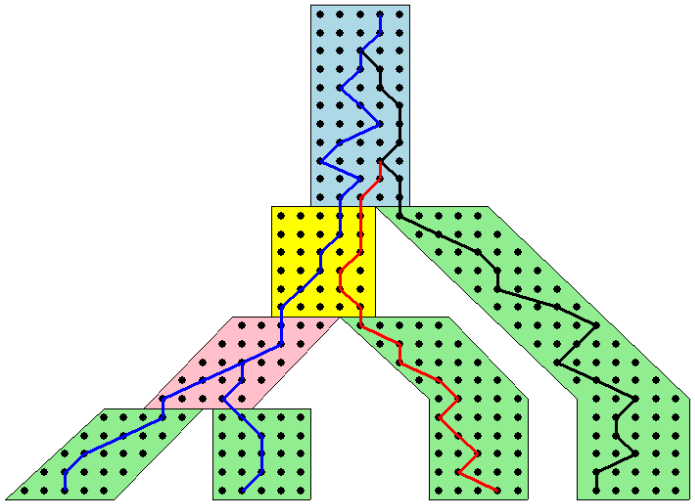
- Population-level process, also called the "Multi-species coalescent" (Kingman, 1982)
- Gene trees can differ from species trees due to short times between speciation events or large population size; this is called "Incomplete Lineage Sorting" or "Deep Coalescence".

Incomplete Lineage Sorting (ILS)

- 1000+ 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.

Key observation:

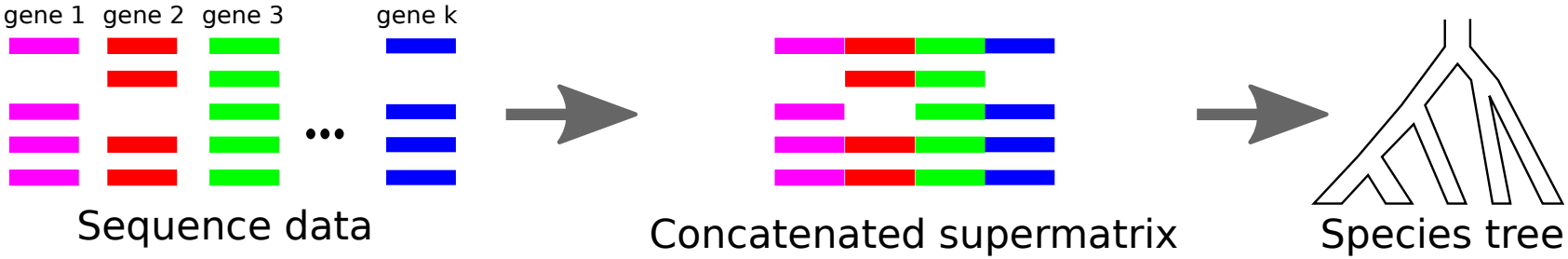
Under the multi-species coalescent model, the species tree defines a *probability distribution on the gene trees*



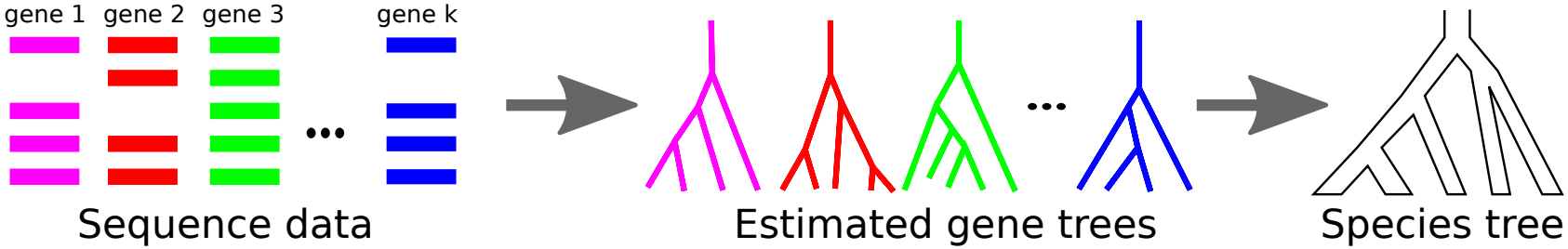
Courtesy James Degnan

Species tree estimation

1- Concatenation: statistically inconsistent (Roch & Steel 2014)

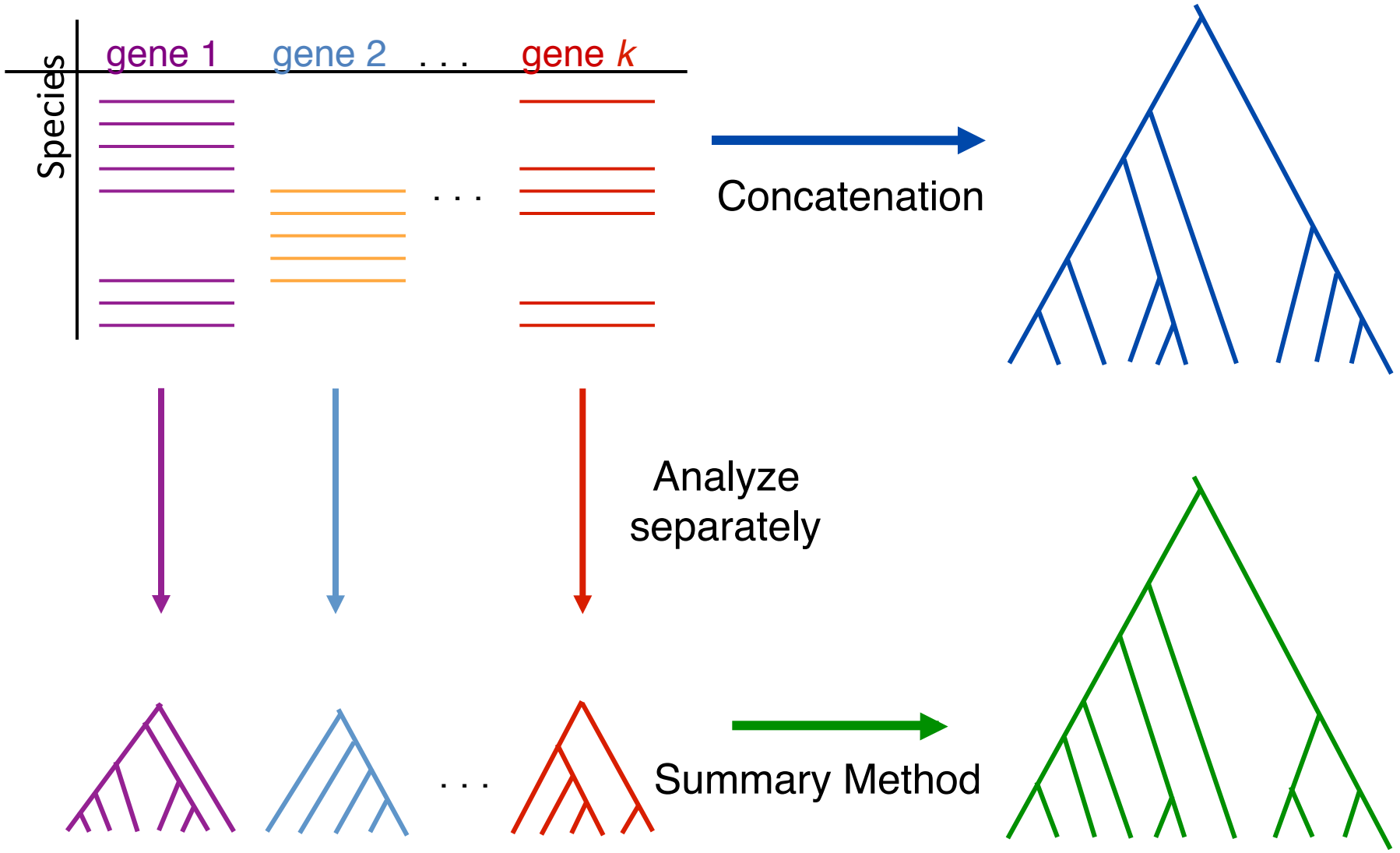


2- Summary methods: can be statistically consistent



3- Co-estimation methods: too slow for large datasets

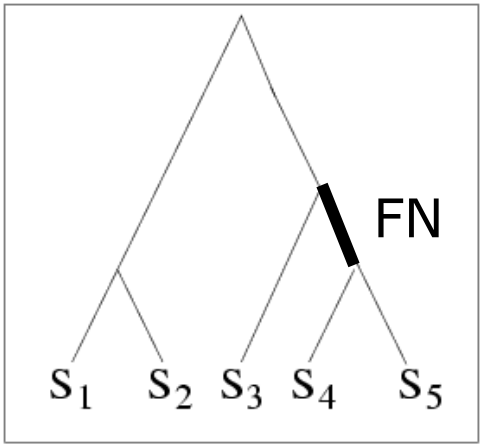
Two competing approaches



Statistically consistent under ILS?

- *BEAST (Heled and Drummond 2010): Bayesian co-estimation of gene trees and species trees -- YES
- 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
- ASTRAL (Mirarab et al. 2014): quartet-based method - YES
- MDC – NO
- Greedy – NO
- Concatenation under maximum likelihood – NO (Roch & Steel, submitted)
- MRP (supertree method) – open

Quantifying Error



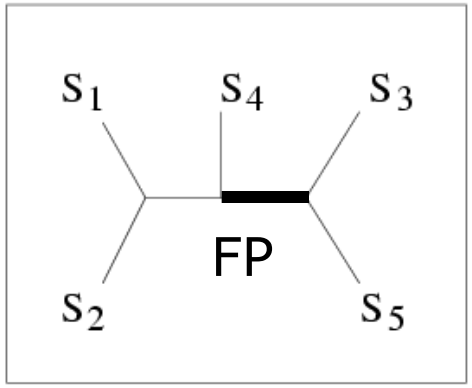
TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

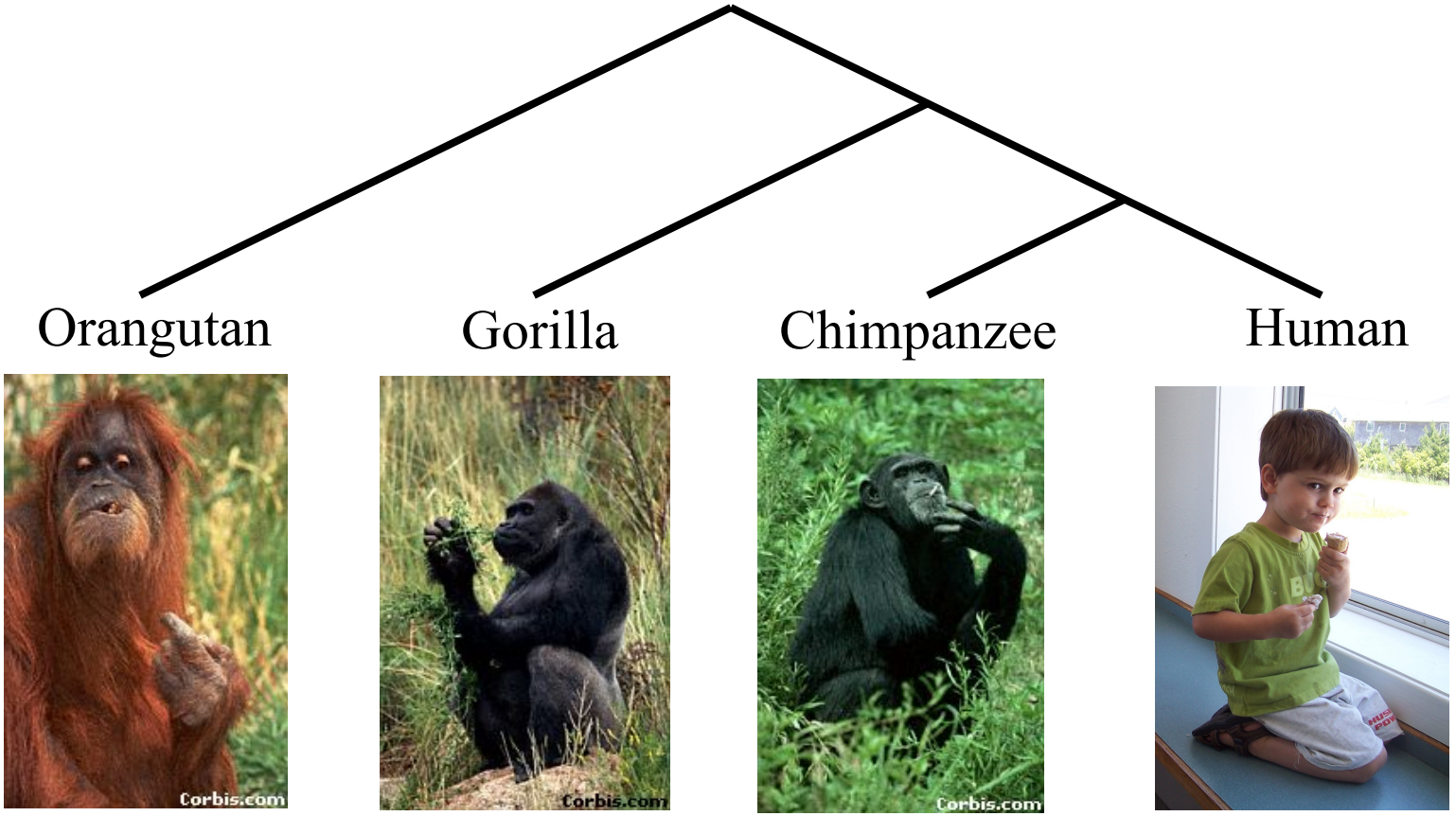
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate



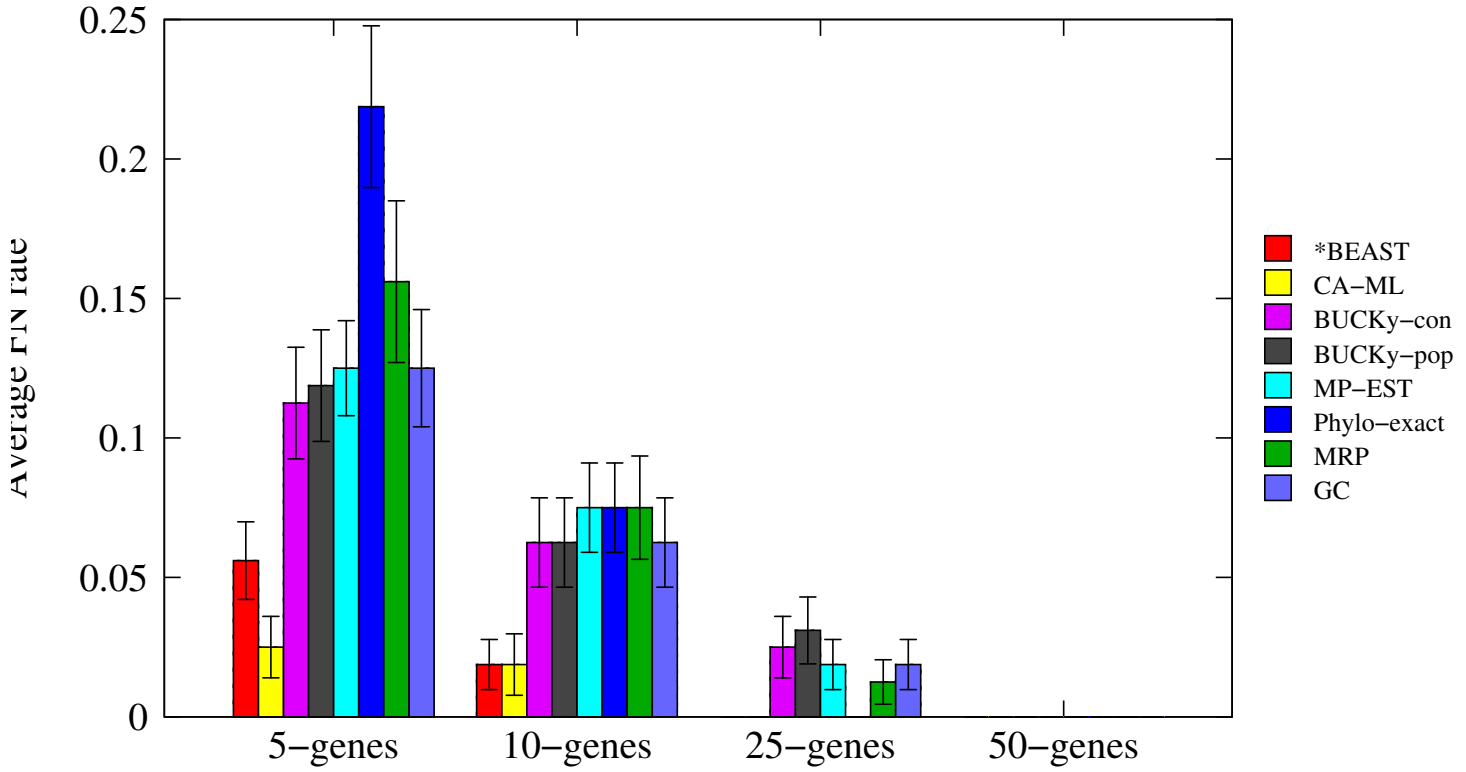
INFERRED TREE

Species tree estimation: difficult, even for small datasets!



From the Tree of the Life Website, University of Arizona

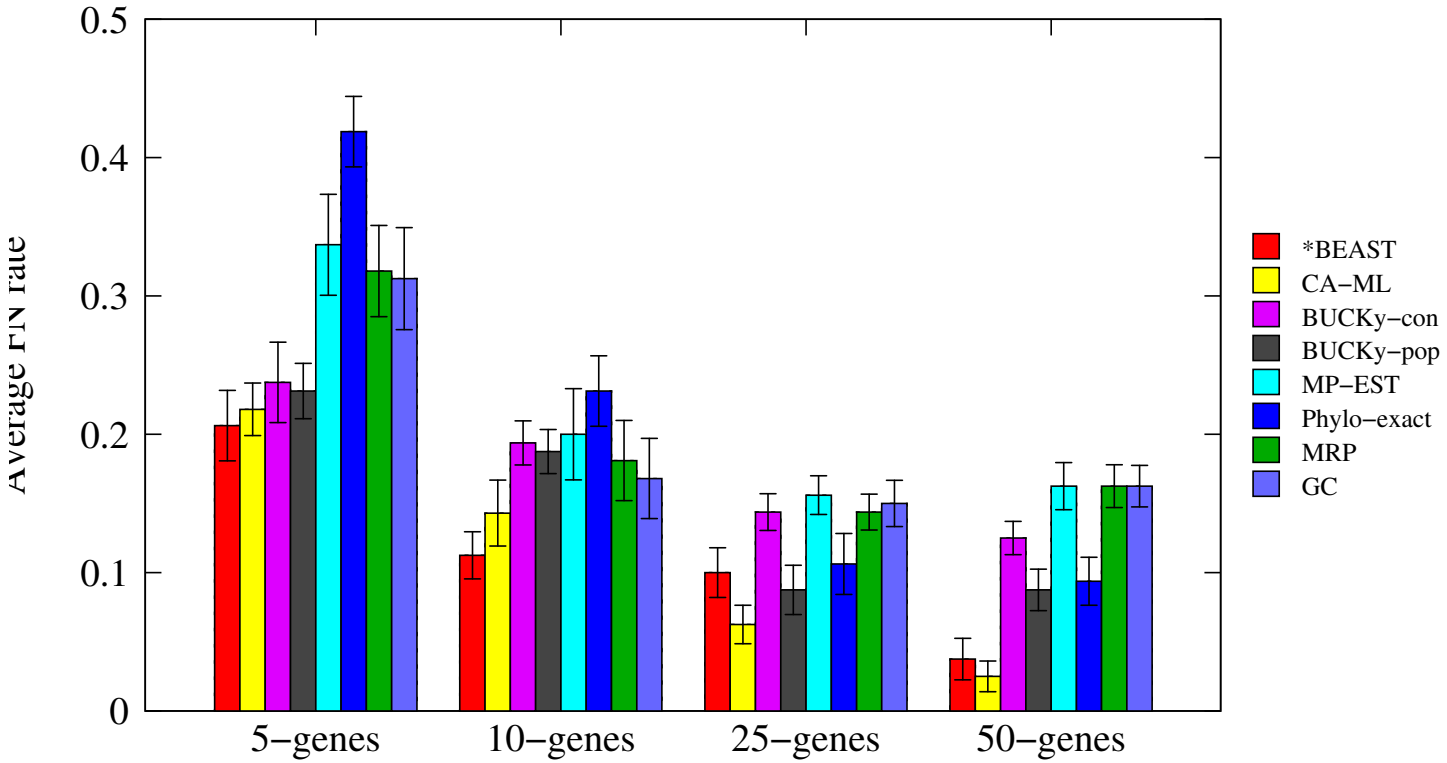
Results on 11-taxon datasets with weak ILS



***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

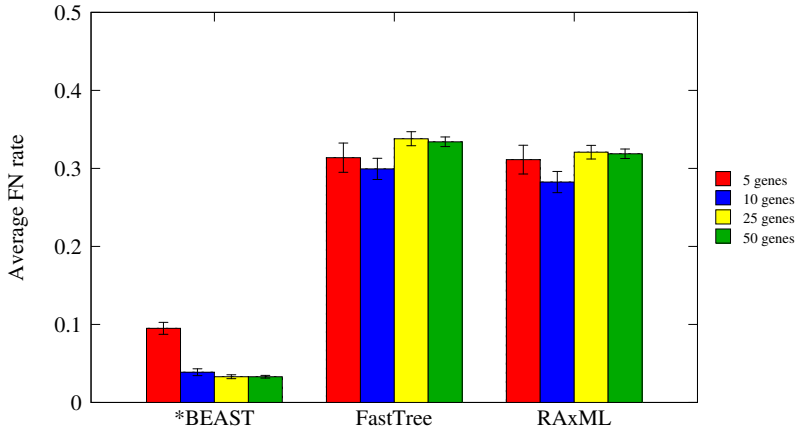
Results on 11-taxon datasets with strongILS



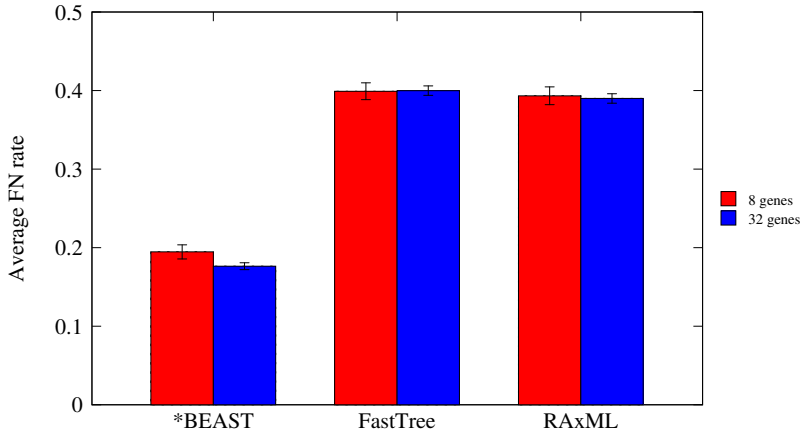
***BEAST** more accurate than summary methods (MP-EST, BUCKy, etc)
CA-ML: (concatenated analysis) also very accurate

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

Gene Tree Estimation:
*BEAST vs. Maximum Likelihood



11-taxon weakILS datasets

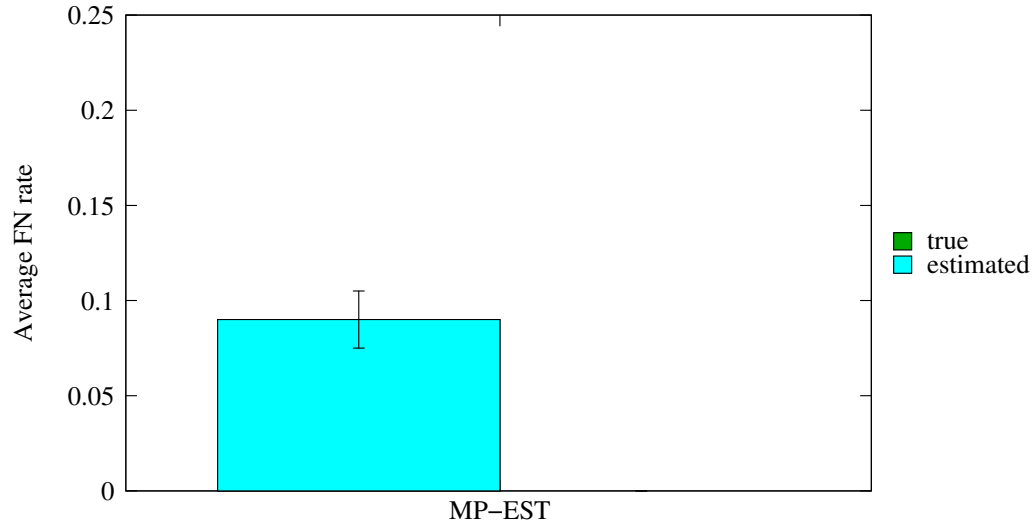


17-taxon (very high ILS) datasets

*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.

Species tree estimation

- Concatenation – not statistically consistent and can have poor accuracy
- Summary methods – can be statistically consistent, but typically have poor accuracy when gene trees have estimation error
- Co-estimation methods – can have outstanding accuracy but are too computationally intensive to use on datasets with more than 50 genes

Improving summary methods through binning

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

Improving summary methods through binning

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

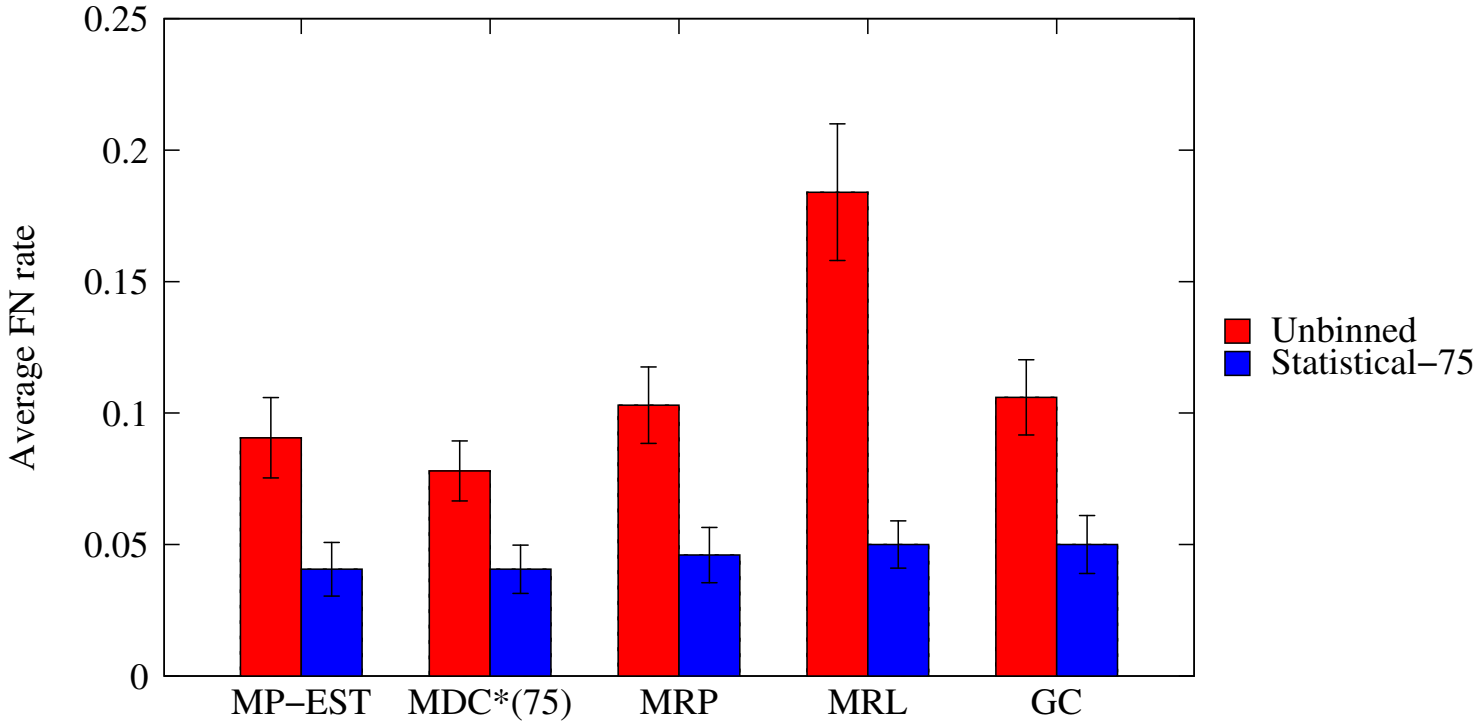
2. Estimate trees on each supergene alignment using maximum likelihood

3. Combine the supergene trees together using a summary method

Variants:

- Naïve binning (Bayzid and Warnow, Bioinformatics 2013)
- Statistical binning (Mirarab, Bayzid, and Warnow, in press)

Statistical binning vs. unbinned



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

*BEAST

- The input to *BEAST is a set of gene sequence alignments, and it uses them to co-estimate gene trees and species trees, using a Bayesian MCMC technique.
- Bayzid & Warnow (Bioinformatics 2013) showed:
  - *BEAST produces more accurate gene trees than maximum likelihood on gene sequence alignments.
  - the same accuracy species tree can be obtained by applying summary methods (e.g., MP-EST) to the gene trees produced by *BEAST.
  - *BEAST fails to converge on datasets with more than 25 species
  - **BEAST failed to converge on 100-gene datasets, even given a week of running time.*

BBCA

BBCA = Binned Boosted Coalescent Analysis

1. Assign genes to "bins" -- randomly

2. Apply *BEAST to each bin, co-estimating gene trees and species trees on each bin

3. Combine the gene trees together using a summary method (e.g., MP-EST)

Note: this is a statistically consistent way of using binning, if the size of the bins Is allowed to grow with the number of genes.

Evaluation

Datasets:

- 11-taxon simulated datasets, developed by Chung & Ané (Systematic Biology 2011) with 100 genes
- 12-taxon simulated datasets based on a coalescent analysis of a mammalian clade (Laurasiatheria), with 100 genes; sequence-lengths varied from 500 to 1500 nt.

Methods

- *BEAST (24-hour to 168-hour analyses)
- BBCA (24-hour and 48-hour analyses), using bins of 25 genes
- Concatenation using maximum likelihood (RAxML)

Criteria

- Species tree error (Robinson-Foulds distance to true tree)
- Failure to converge of MCMC run (% of ESS values below 100)

24-hour BBCA analysis has much lower error than 96-hour *BEAST analysis

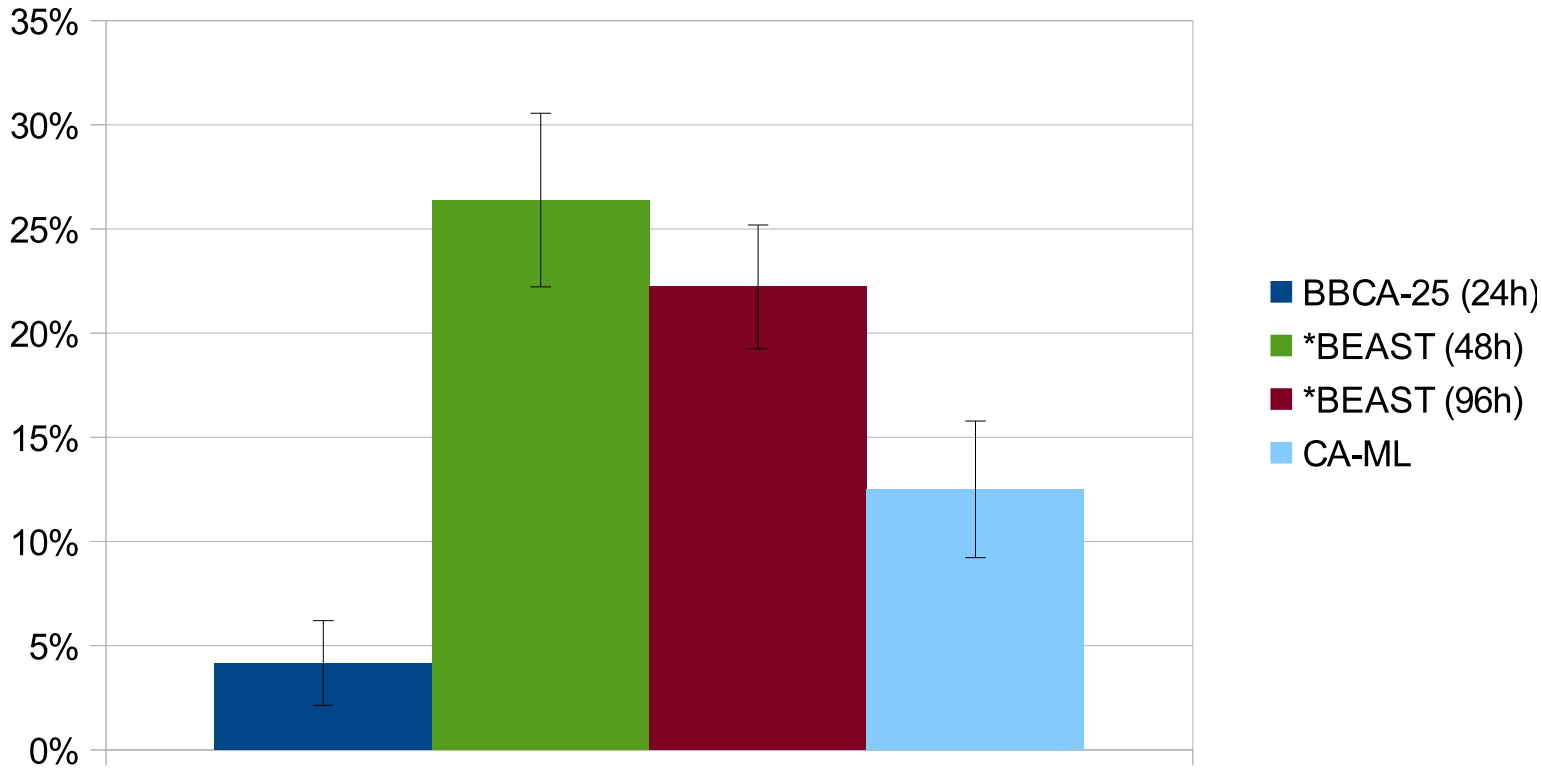


Figure 3 Average species tree estimation error (and standard error bars) on eight Laurasiatheria 1500bp simulated datasets using BBCA, *BEAST and concatenation; BBCA is run with a 24-hour time limit on each 25-gene bin, and *BEAST is run with a 48-hour or 96-hour time limit. Increasing the time per bin to 48 hours did not change the accuracy for BBCA.

24-hour BBCA analysis more accurate than 168-hour *BEAST analysis

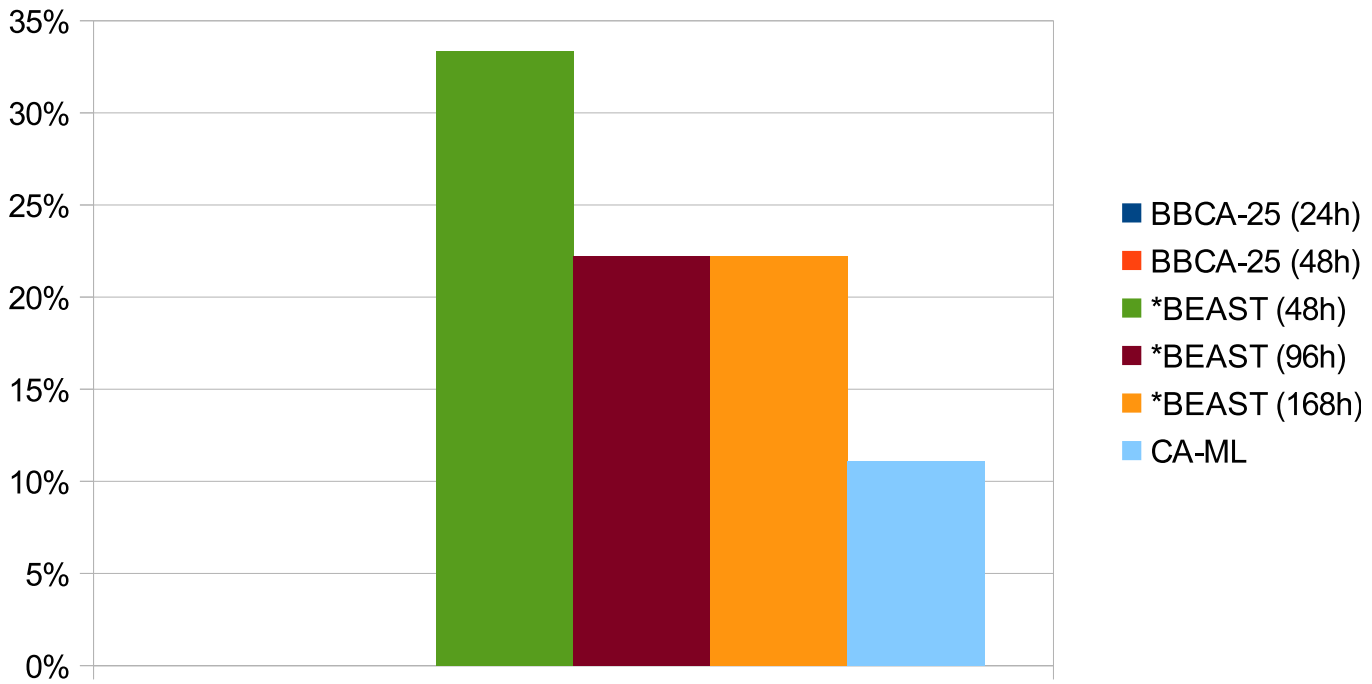


Figure 7 Average species tree estimation error on one Laurasiatheria 1000bp simulated dataset using BBCA, *BEAST and concatenation with maximum likelihood (CA-ML). BBCA is run with a 24-hour or 48-hour time limit on each 25-gene bin, and *BEAST is run with a 48-hour, 96-hour or 168-hour time limit. BBCA using either 24 hours or 48 hours per bin recovers the true species tree, but both *BEAST analyses and the CA-ML fail to recover the true species tree.

BBCA enables *BEAST to converge in less time

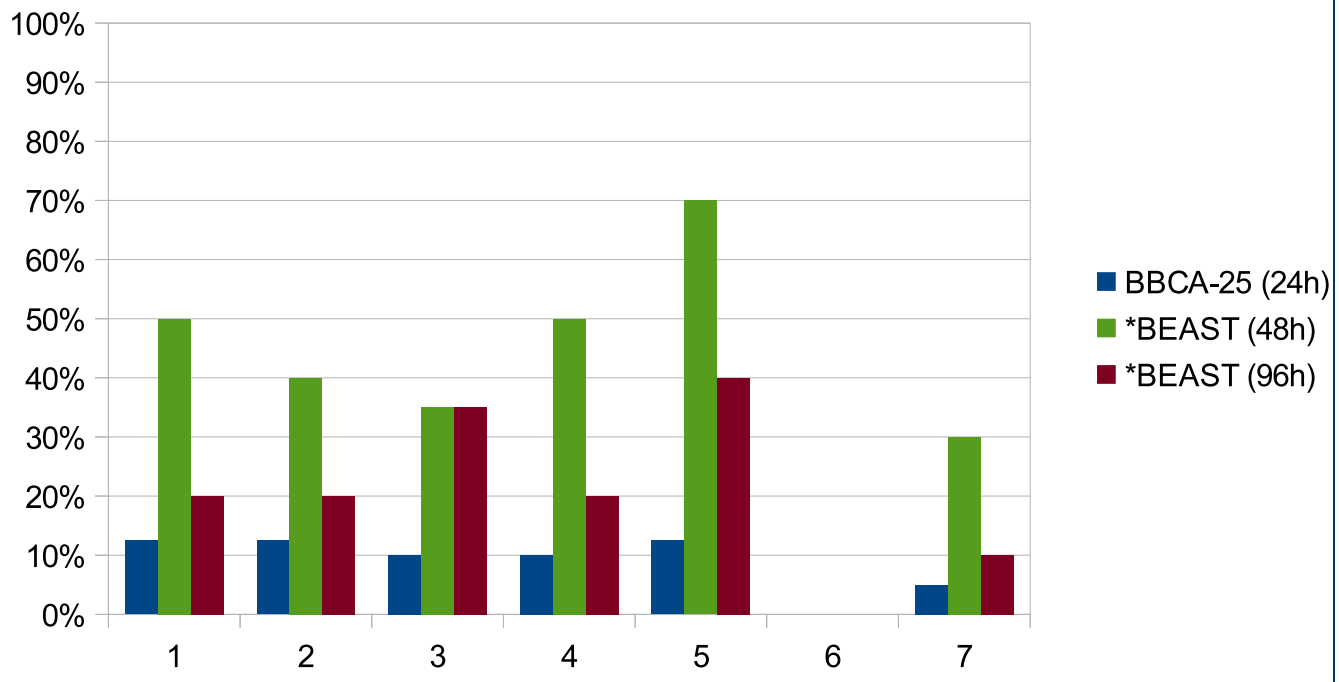


Figure 2 Proportion of ESS values below the minimum threshold (100) for convergence for the ten 11-taxon strongILS datasets. We show results when running BBCA (using 24 hours per 25-gene bin, in blue), *BEAST (using 48 hours on the sets of 100 genes, in green) or *BEAST (using 96 hours, in dark red). We report these proportions for seven different statistics: (1) posterior, (2) prior, (3) likelihood, (4) species.coalescent, (5) species.popSizesLikelihood, (6) speciation.likelihood, (7) species.popMean. Thus, BBCA has converged for 85-95% of the runs, using 24 hours per bin. In contrast, *BEAST has converged for only 60-90% of the runs after running for 96 hours.

Summary

- Coalescent-based species tree estimation: provides statistical guarantees in the presence of incomplete lineage sorting
- Summary methods (which combine gene trees) have low accuracy in the presence of gene tree estimation error
- *BEAST (which co-estimates gene trees and species trees) has higher accuracy, but fails to run on large datasets – convergence problems!
- BBCA enables *BEAST to converge on datasets with many genes, and improves topological accuracy
- New questions in phylogenetic estimation about impact of error in input data.

Acknowledgments

PhD students: Théo Zimmermann (Ecole Normale Superieure) and Siavash Mirarab (University of Texas at Austin)

Funding: NSF, David Bruton Jr. Centennial Professorship, TACC (Texas Advanced Computing Center), and HHMI Predoctoral Fellowship (to Siavash Mirarab)

TACC and UTCS computational resources