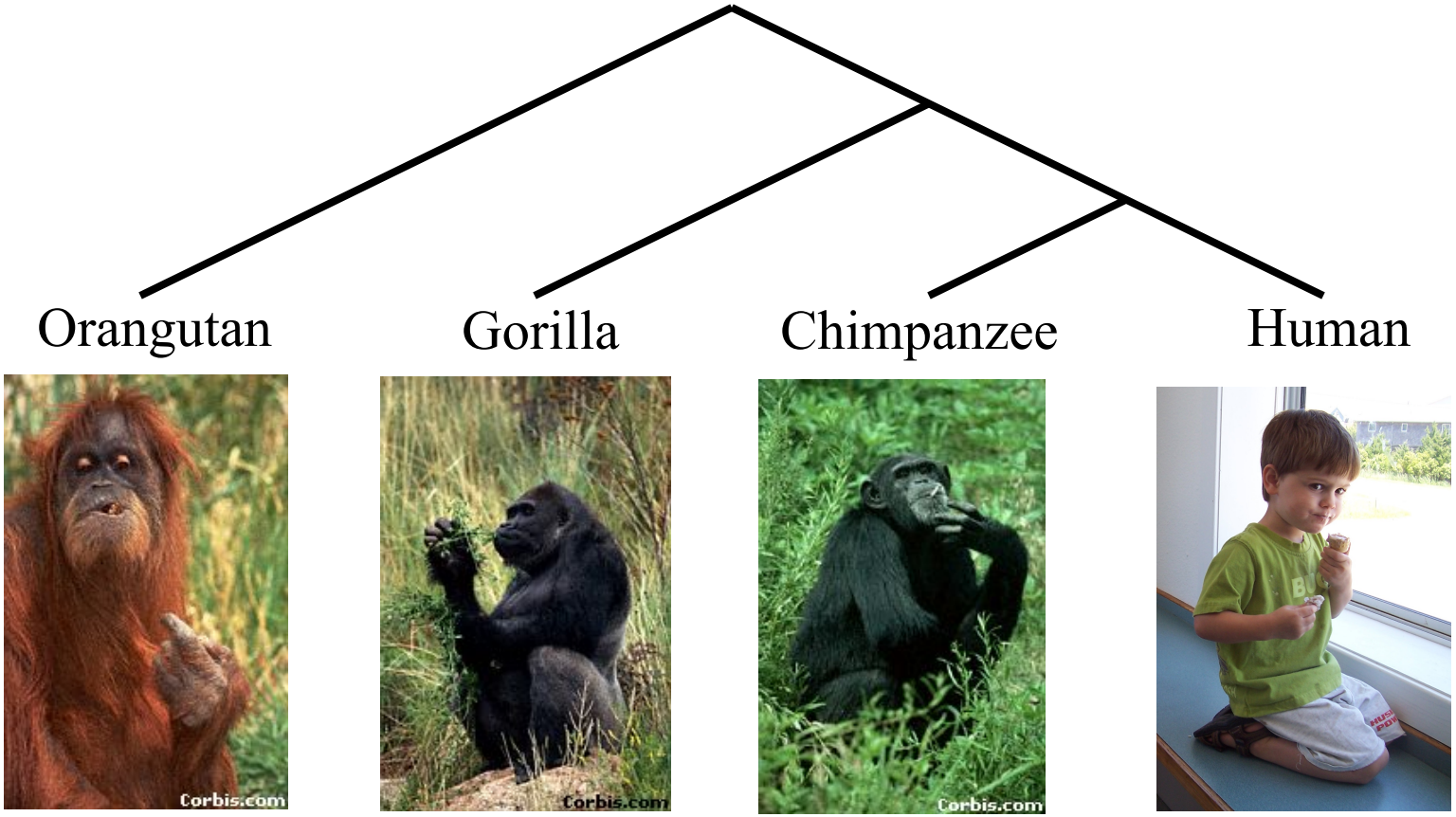
New methods for estimating species trees from genome-scale data

Tandy Warnow

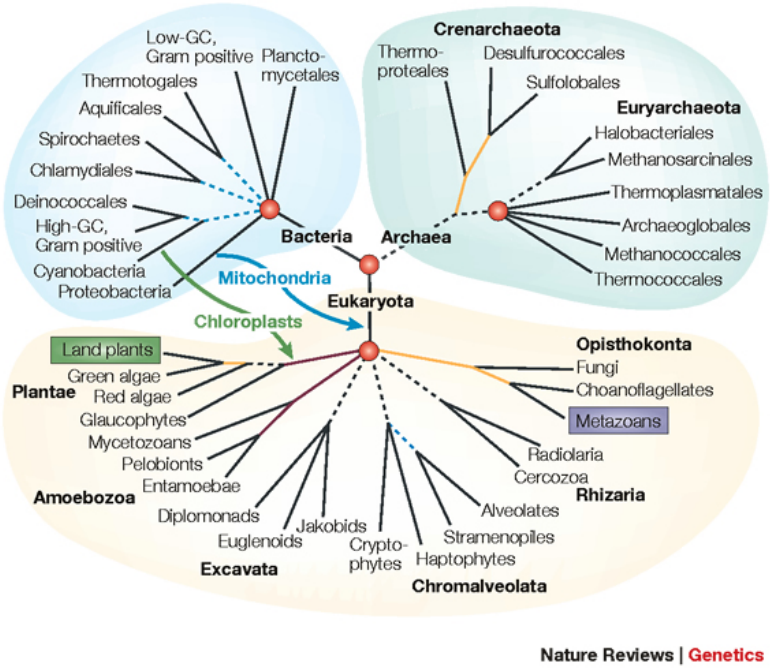
The University of Illinois

Species Tree Estimation



From the Tree of the Life Website,
University of Arizona

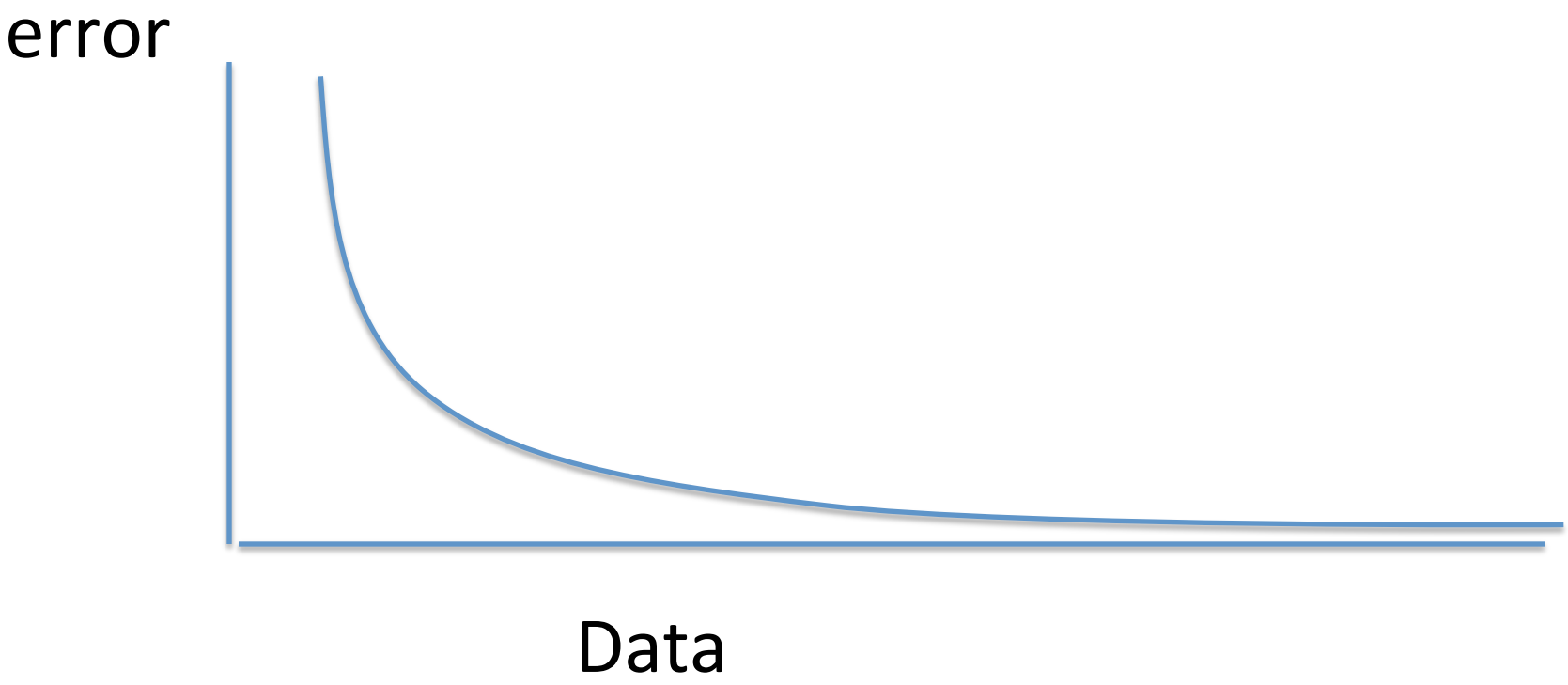
Phylogenomics = Species trees from whole genomes



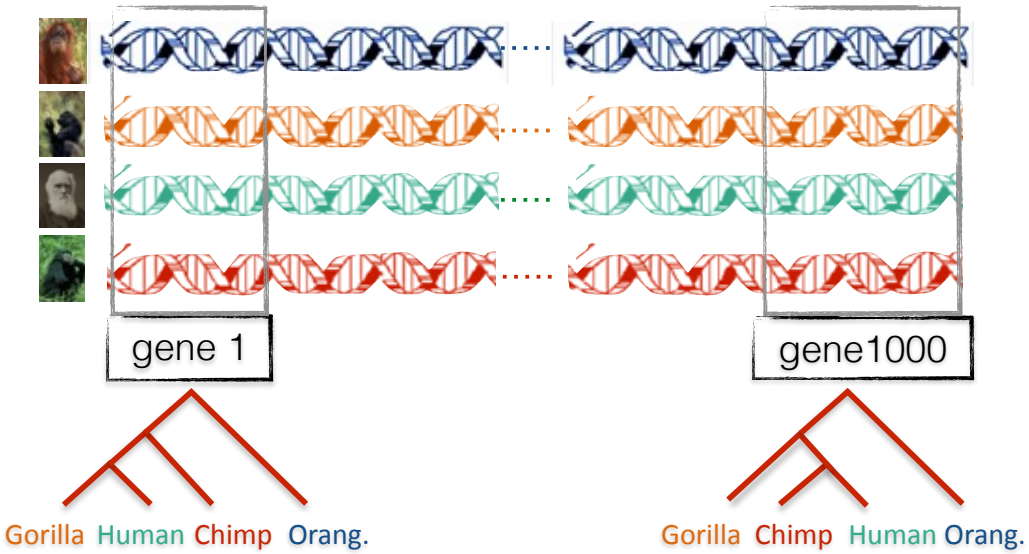
Multiple applications, including

- Metagenomics,
- Protein Structure and Function,
- Conservation Biology
- Adaptation

Statistical Consistency



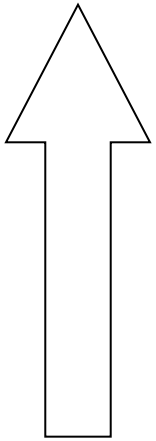
Gene tree discordance



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

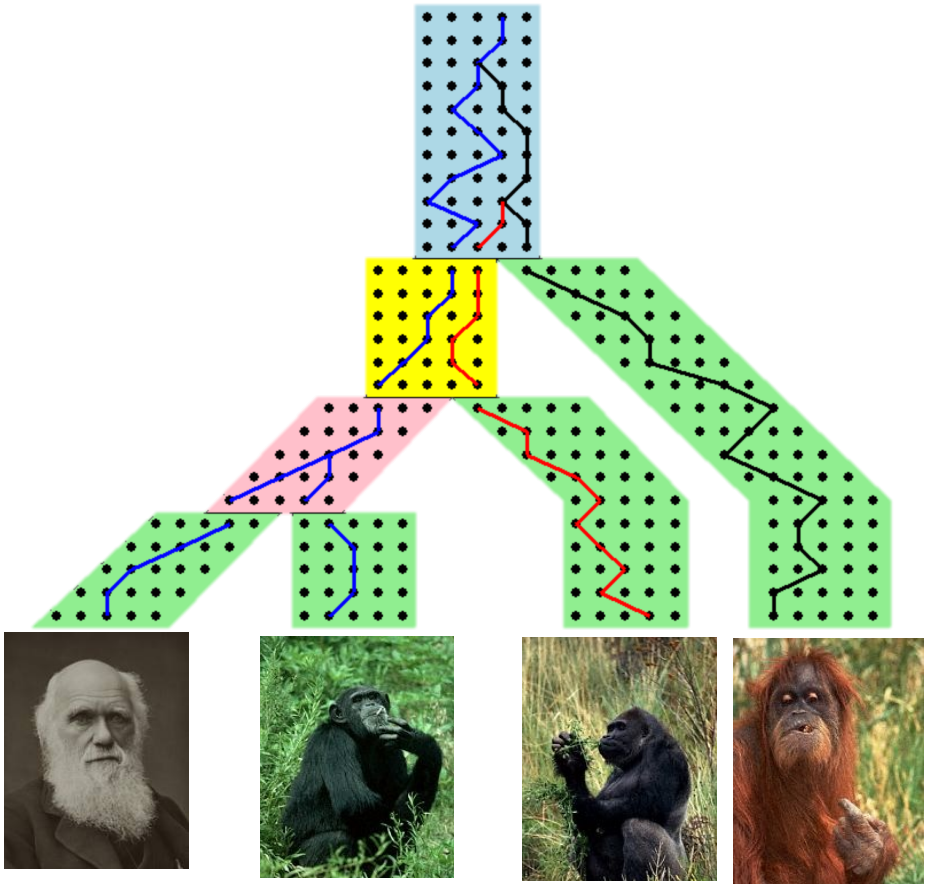
Gene trees inside the species tree (Coalescent Process)

Past



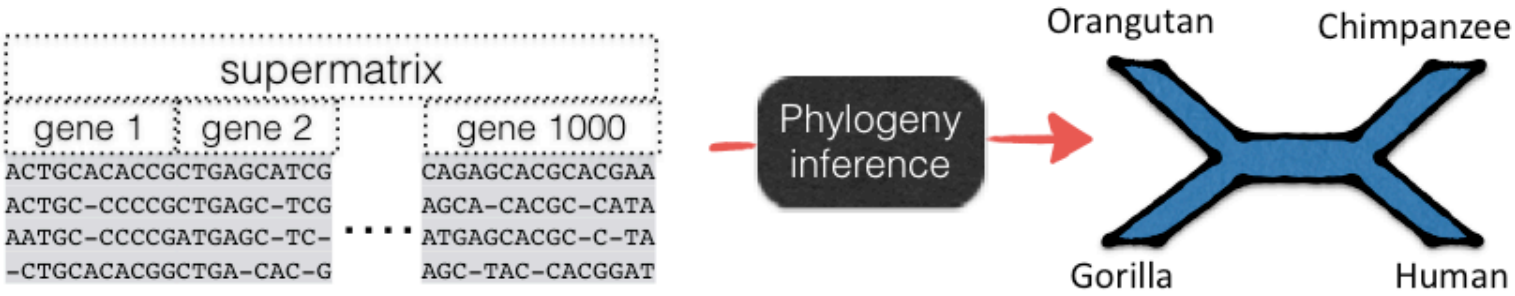
Present

Courtesy James Degnan



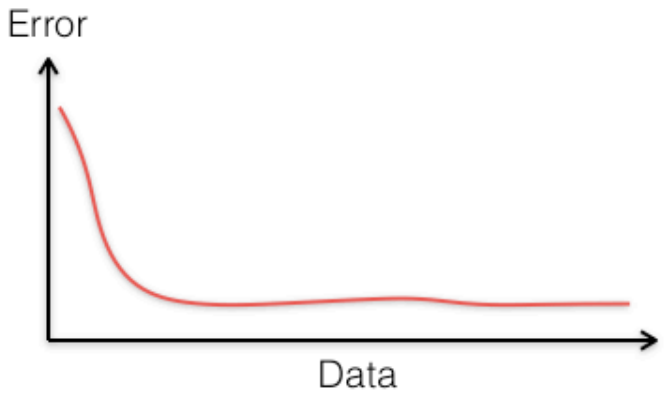
Gorilla and Orangutan are not siblings in the species tree, but they are in the gene tree.

Traditional approach: concatenation

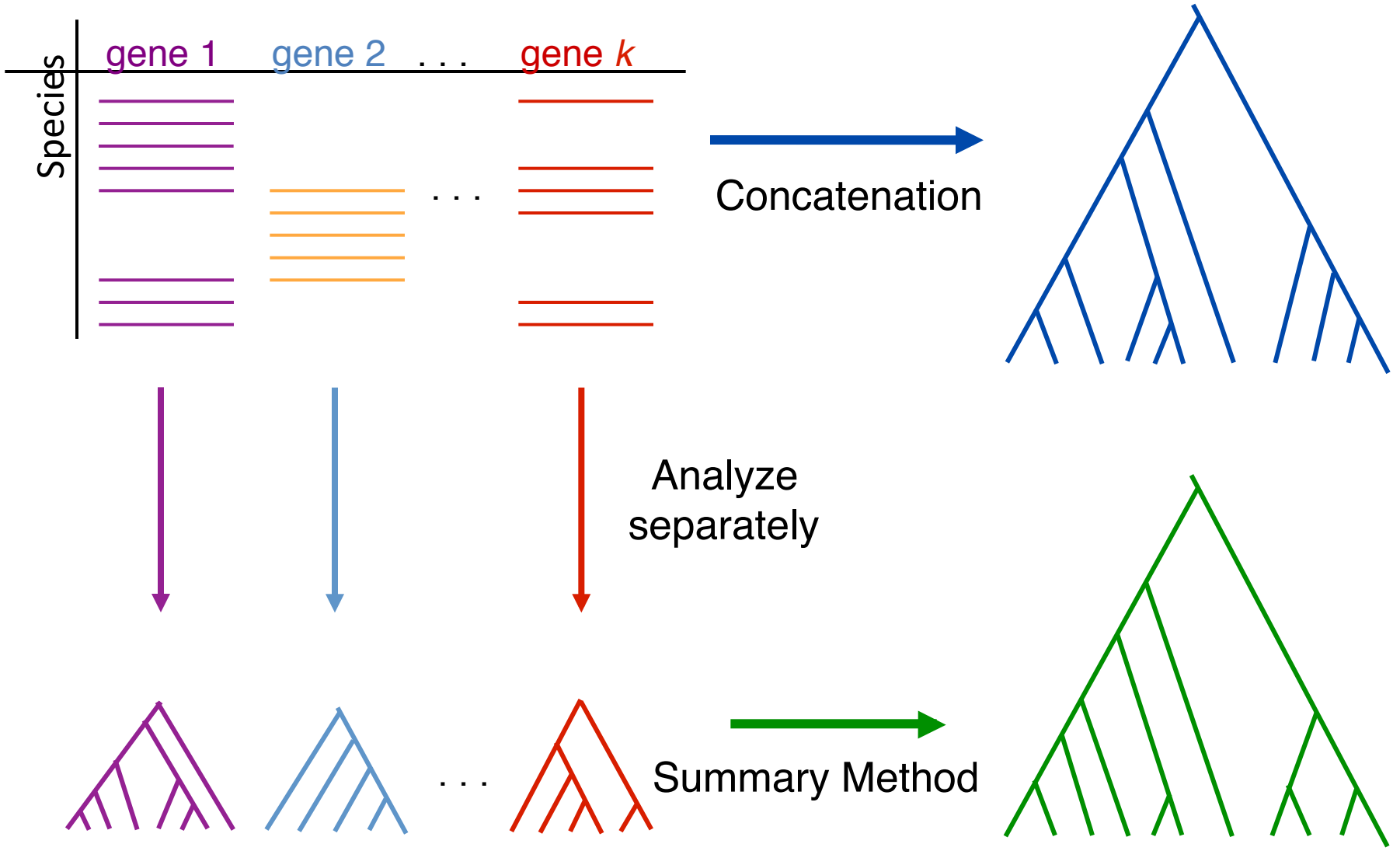


- Statistically <u>inconsistent</u> and can even be positively misleading (proved for unpartitioned maximum likelihood)
  [Roch and Steel, Theo. Pop. Gen., 2014]

- Mixed accuracy in simulations
  [Kubatko and Degnan, Systematic Biology, 2007]
  [<u>Mirarab</u>, et al., Systematic Biology, 2014]



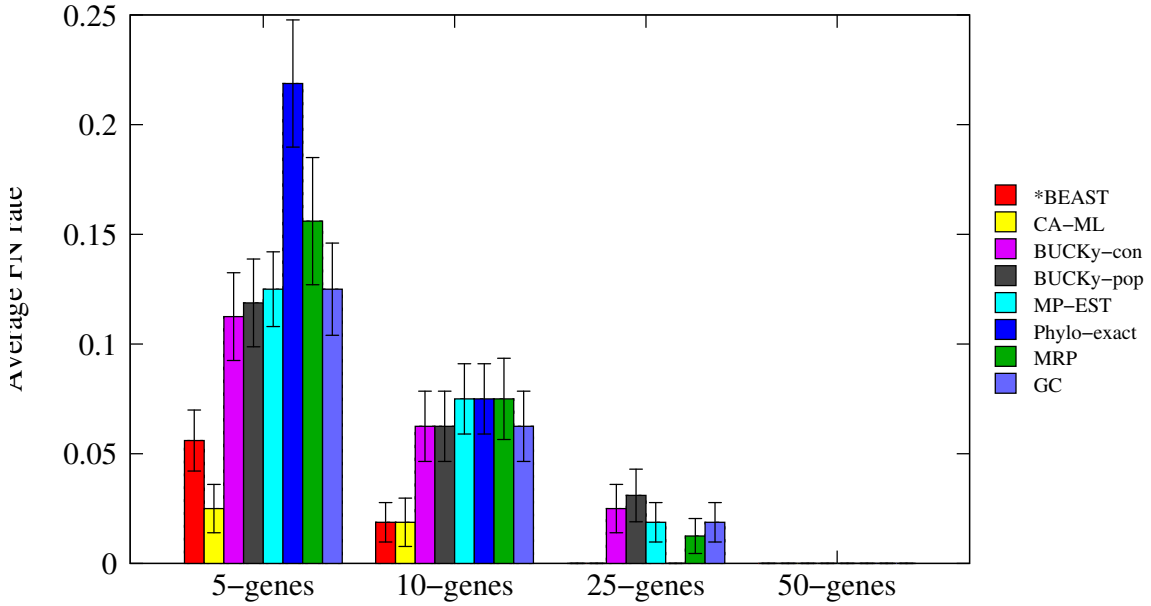
Main competing approaches



Species Tree Estimation Methods

- CA-ML (concatenation using maximum likelihood) – not consistent
- *BEAST (co-estimation of gene trees and species trees)– consistent, but very slow and limited to small datasets
- Summary methods: some are consistent, e.g.,
  - BUCKy, MP-EST, …
  - But not all

Results on 11-taxon datasets with weak ILS

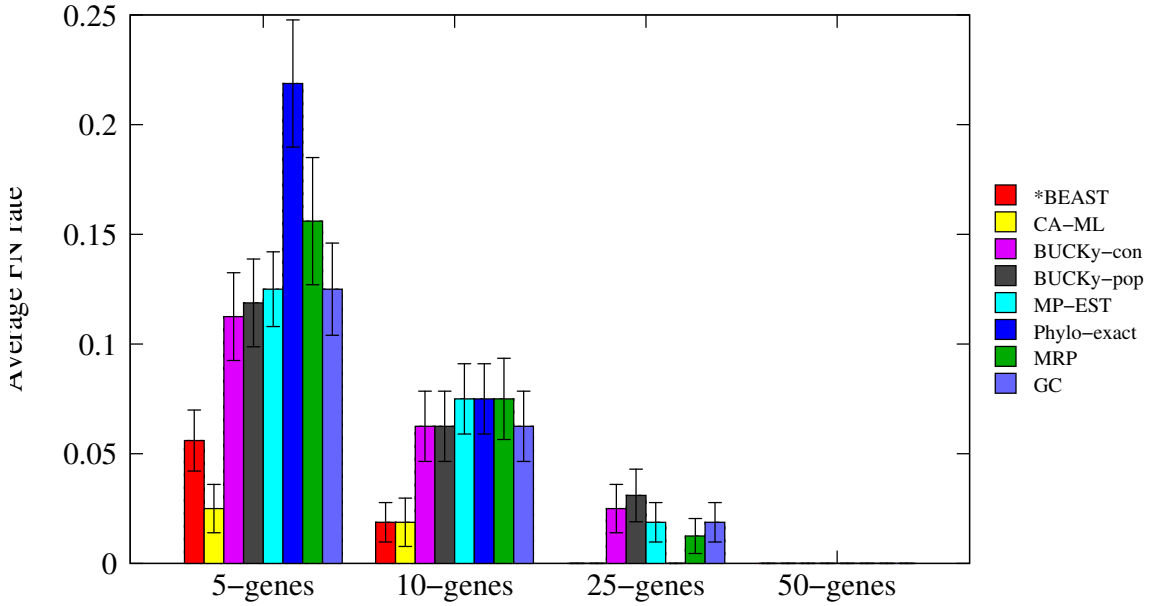


*BEAST MORE ACCURATE than summary methods, because *BEAST gets more accurate gene trees!

***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 weak ILS

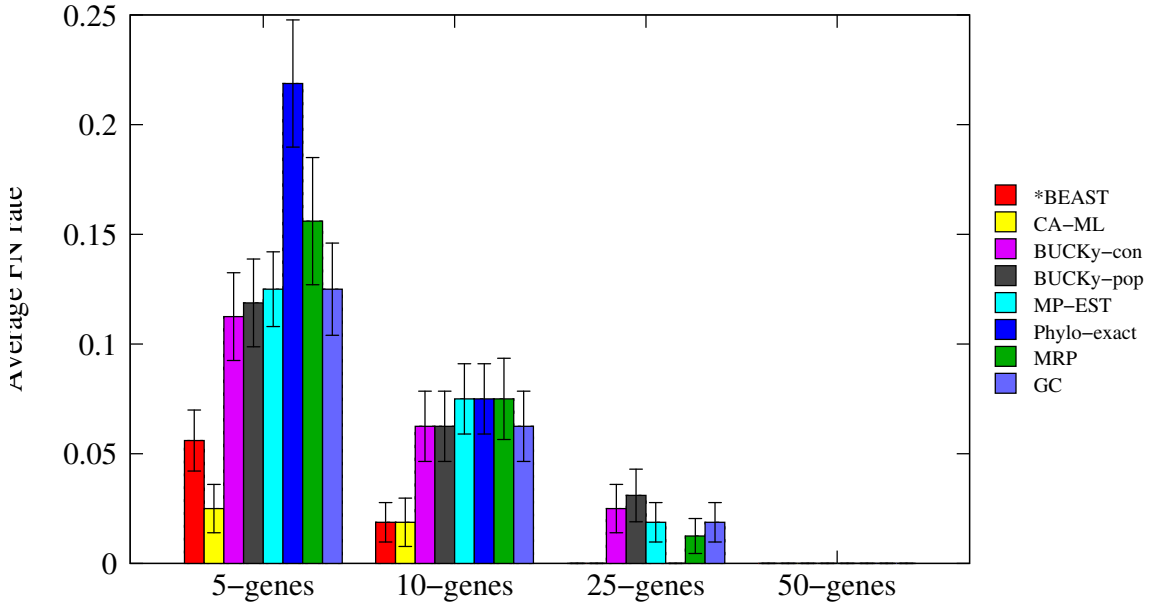


Summary methods (BUCKy-pop, MP-EST) are both statistically consistent under the MSC but are impacted by gene tree estimation error

***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 weak ILS



Concatenation (RAxML) best of all methods on these data!
(However, for high enough ILS, concatenation is not as accurate as the best summary methods.)

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

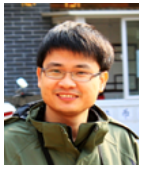
Datasets from Chung and Ané, 2011
Bayzid & Warnow, Bioinformatics 2013

Avian Phylogenomics Project

E Jarvis, HHMI

MTP Gilbert, Copenhagen

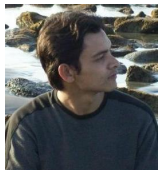
G Zhang, BGI



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 challenges:

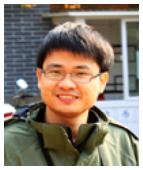
- Concatenation analysis took > 250 CPU years, and suggested a rapid radiation
- We observed massive gene tree heterogeneity consistent with incomplete lineage sorting
- Very poor resolution in the 14,000 gene trees (average bootstrap support 25%)
- Standard coalescent-based species tree estimation methods contradicted concatenation analysis and prior studies

Avian Phylogenomics Project

E Jarvis, HHMI

MTP Gilbert, Copenhagen

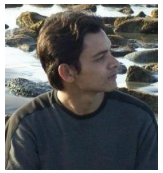
G Zhang, BGI



T. Warnow UT-Austin

S. Mirarab UT-Austin

Md. S. Bayzid, UT-Austin



Plus many many other people…

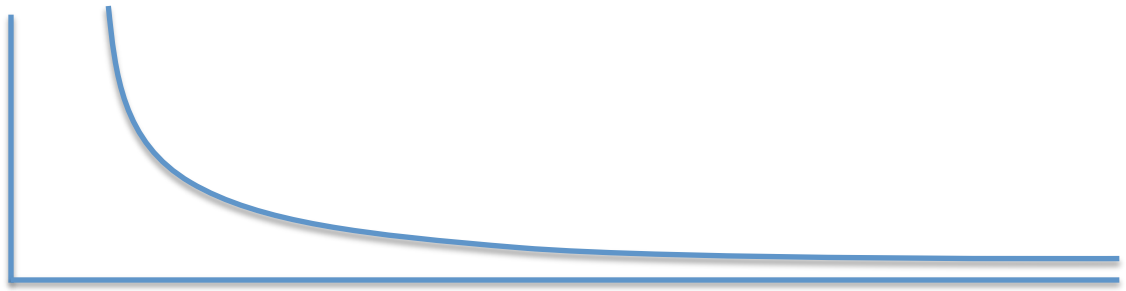
- Approx. 50 species, whole genomes, 14,000 loci

Solution: **Statistical Binning**

- Improves coalescent-based species tree estimation by improving gene trees (Mirarab, Bayzid, Boussau, and Warnow, *Science* 2014), see also weighted statistical binning (Bayzid et al., PLOS One 2015)
- Avian species tree estimated using **Statistical Binning with MP-EST** (Jarvis, Mirarab, et al., *Science* 2014)

Ideas behind statistical binning

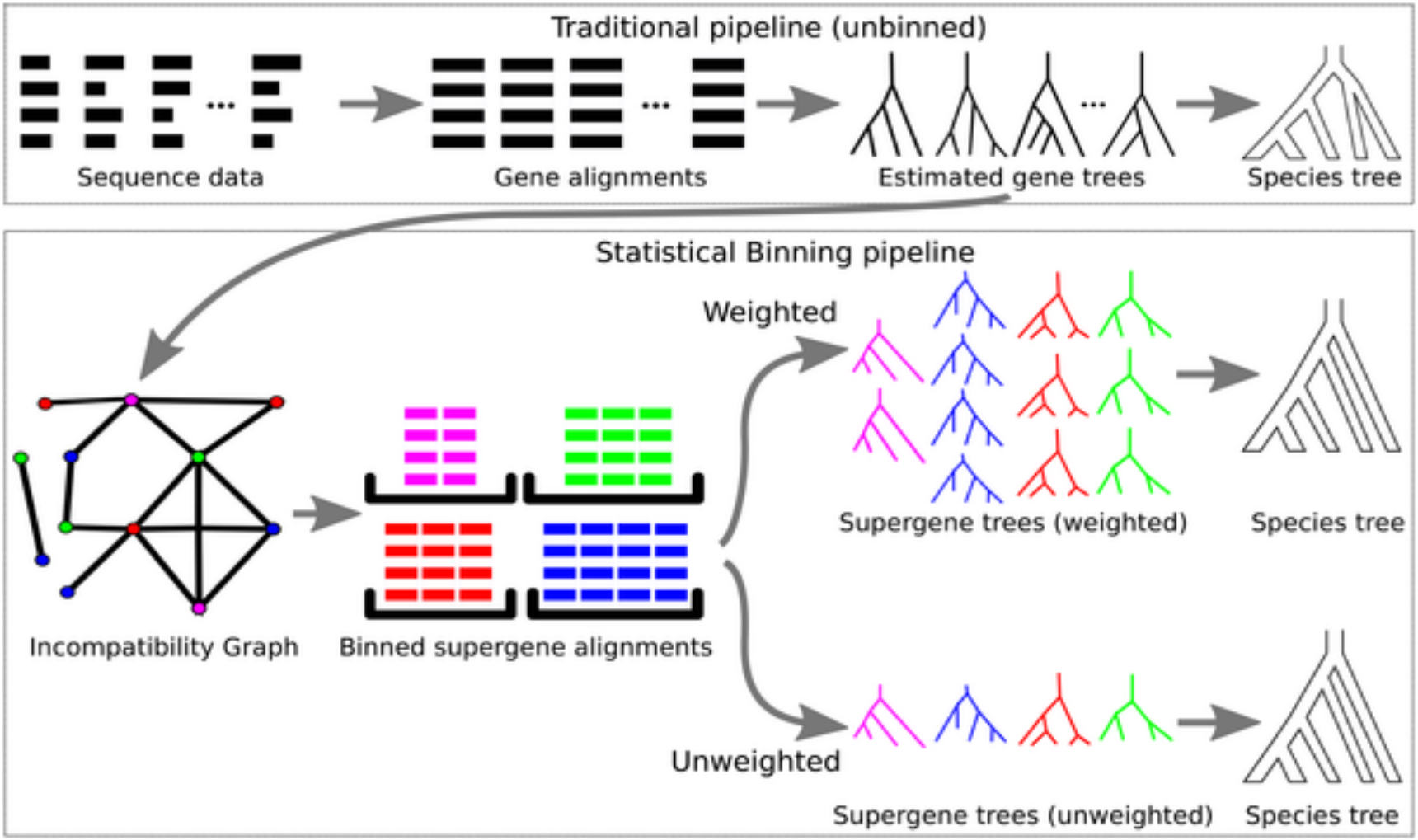
- "Gene tree" error tends to decrease with the number of sites in the alignment



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

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

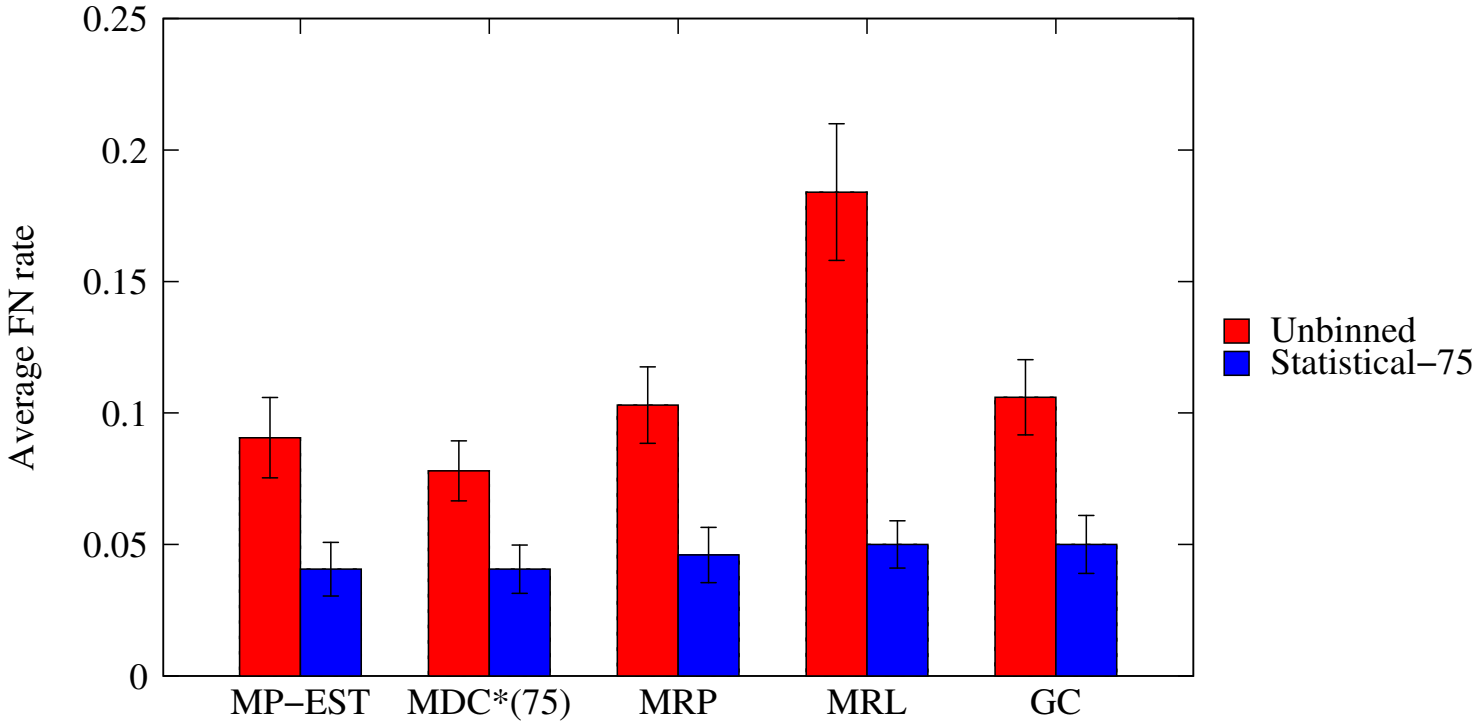
Easy proof:

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.

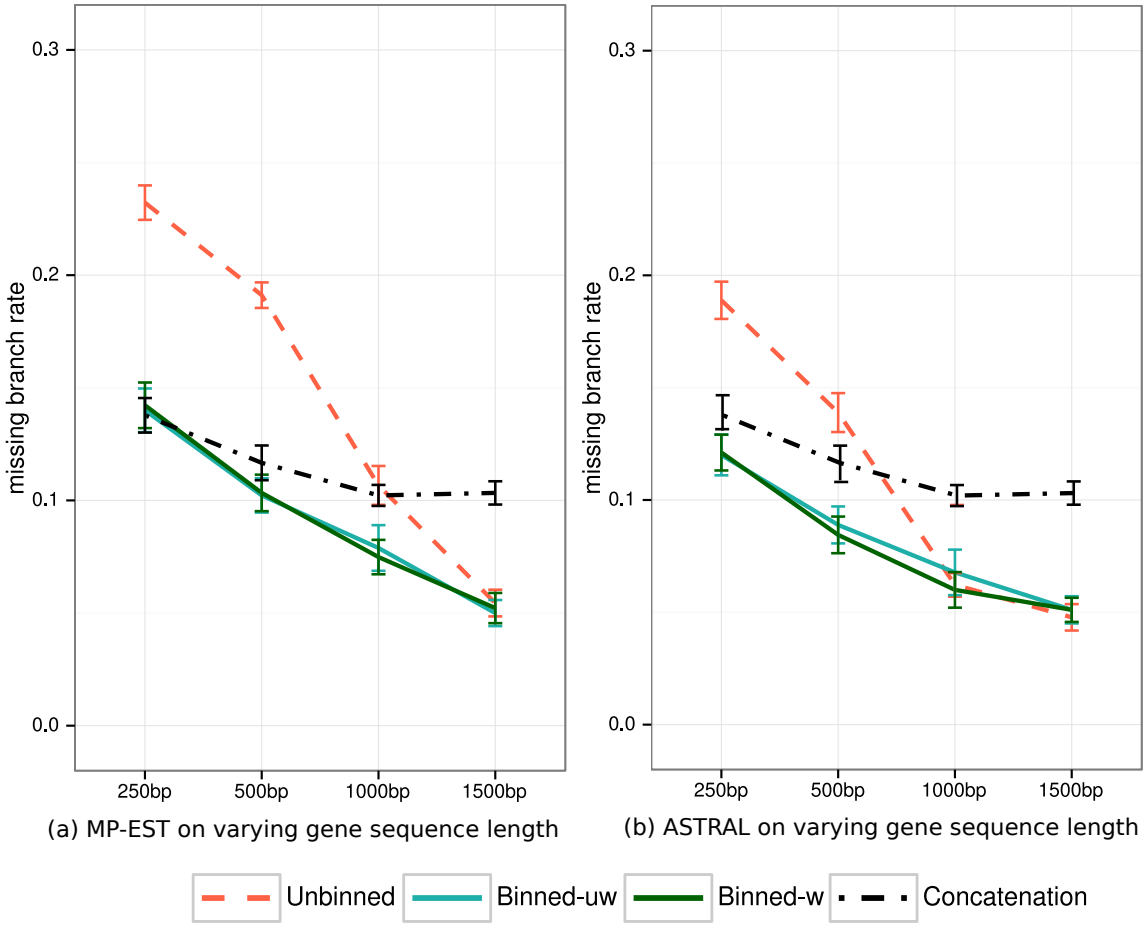
Statistical binning vs. unbinned



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

Binning produces bins with approximate 5 to 7 genes each

Binning can improve species tree topology estimation

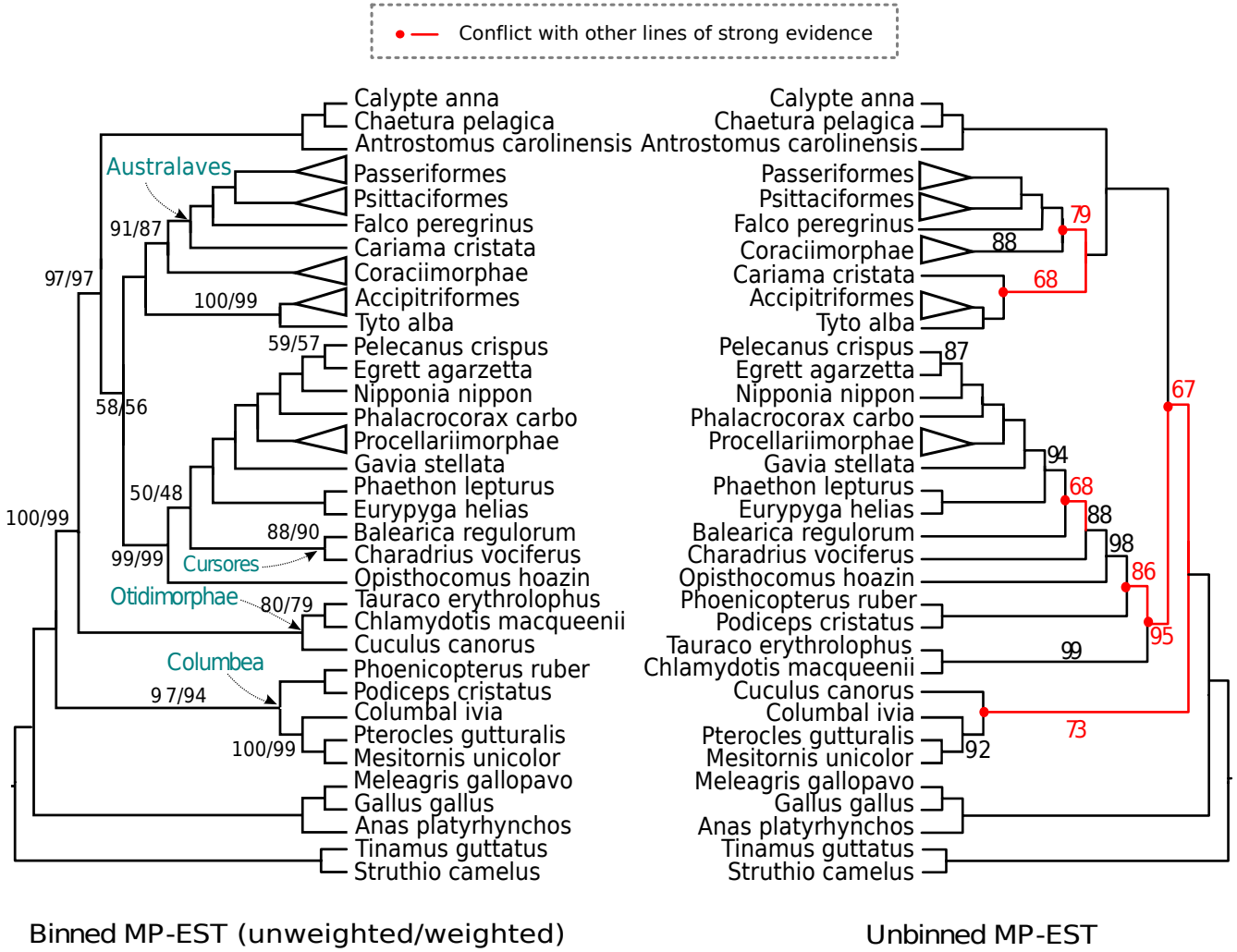


Species tree estimation error for MP-EST and ASTRAL, and also concatenation using ML, on avian simulated datasets: 48 taxa, moderately high ILS (AD=47%), 1000 genes, and varying gene sequence length.

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

PLOS | ONE

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



Unbinned MP-EST strongly rejects Columbea, a major finding by Jarvis, Mirarab,et al.

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

Running Time Comparison

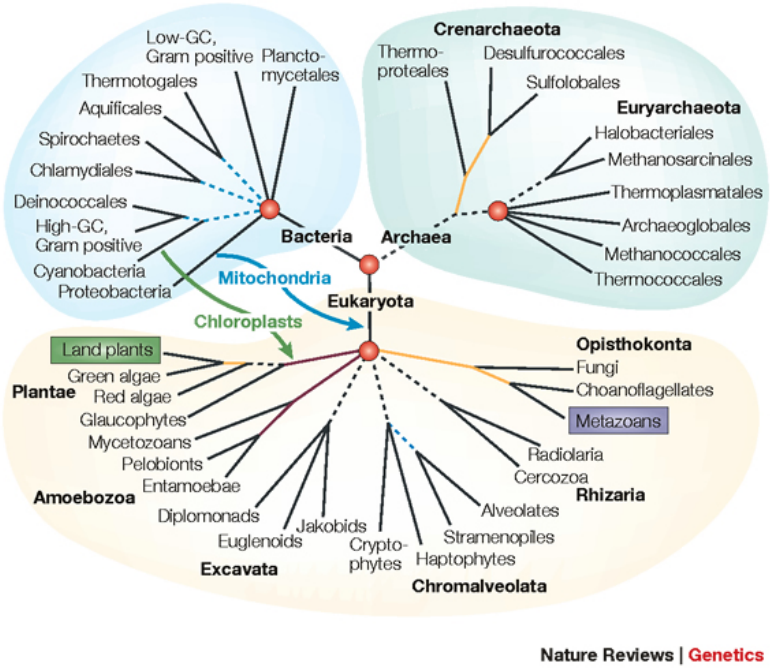
- Concatenation analysis of the Avian dataset:
  - ~250 CPU years and 1Tb memory
- Statistical binning analysis:
  - ~5 CPU years, almost all of which was computing maximum likelihood gene trees, much less memory usage

Species tree estimation using traditional approaches is more computationally expensive, and not as accurate as coalescent-based methods!

Summary

- Species tree estimation is complicated by biological processes that create heterogeneity, and by estimation error.
- Big data issues: NP-hard optimization problems, model-misspecification, and errors (and big datasets, too)
- Constructing phylogenetic trees (and especially the Tree of Life) needs computer scientists, mathematicians, and statisticians to address the multiple challenges.

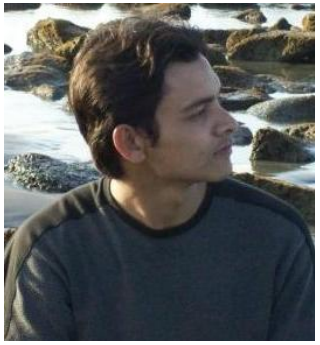
Computational Phylogenomics



NP-hard problems
Large datasets
Complex statistical estimation problems

Metagenomics
Protein structure and function prediction
Medical forensics
Systems biology
Population genetics

Acknowledgments


NSF grant DBI-1461364
http://tandy.cs.illinois.edu/PhylogenomicsProject.html

Other Funding: Grainger Foundation Professorship, David Bruton Jr. Centennial Professorship, and HHMI (to SM)

Computing: Blue Waters and TACC