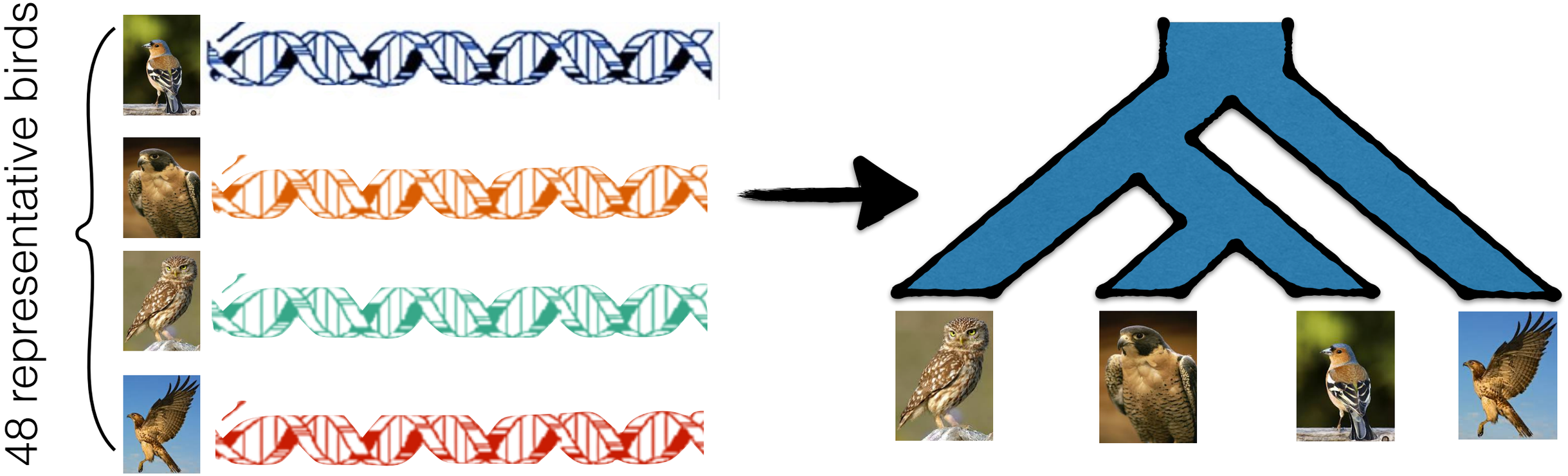
Statistical binning enables an accurate coalescent-based estimation of the avian tree

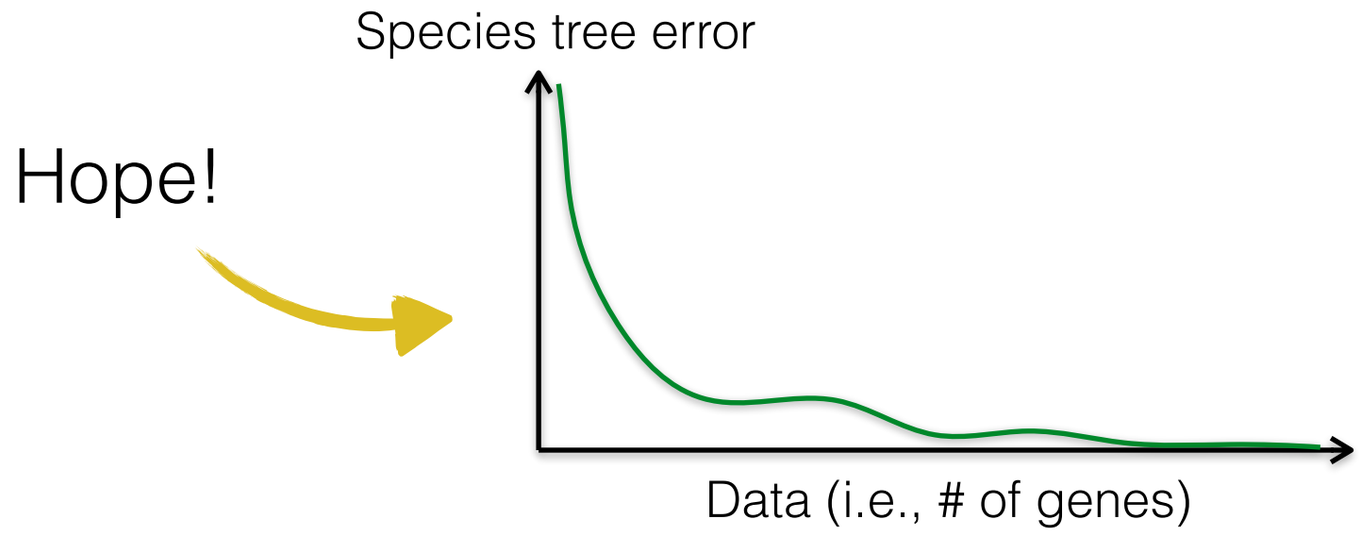
Siavash Mirarab, Md. Shamsuzzoha Bayzid, Bastien Boussau, and Tandy Warnow. Science (2014)

Avian whole genomes phylogenies

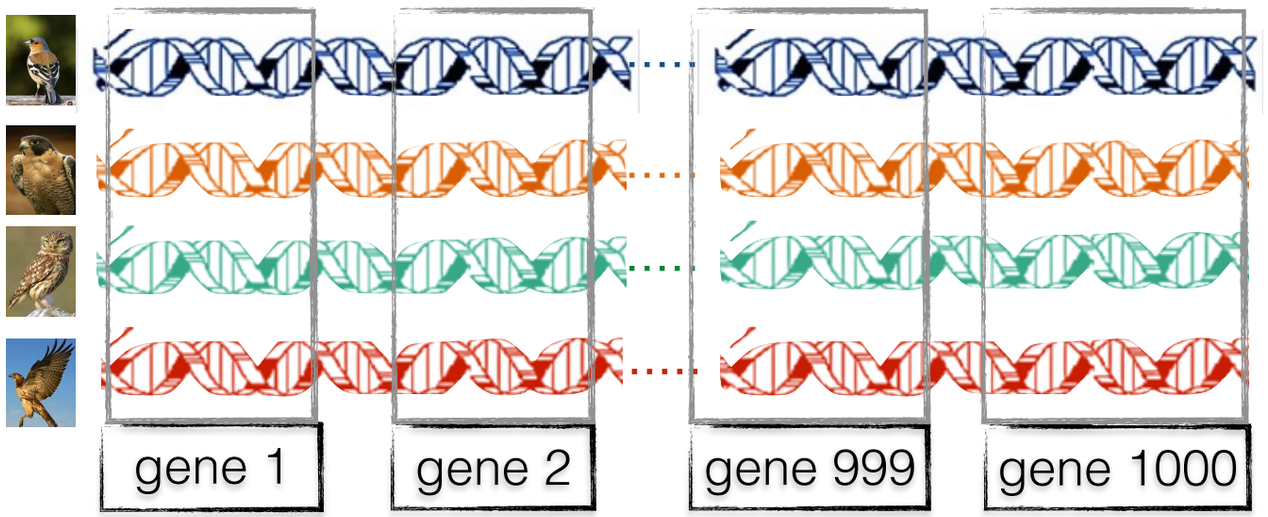
[Jarvis, Mirarab, et al., Science, 2014]

48 representative birds

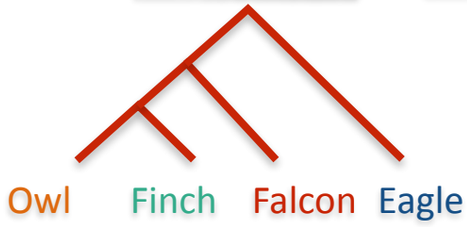


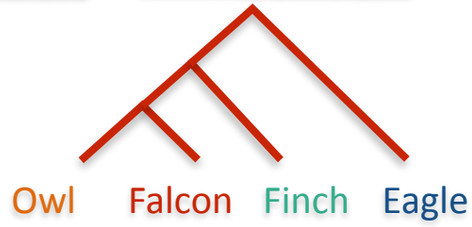


Gene tree discordance

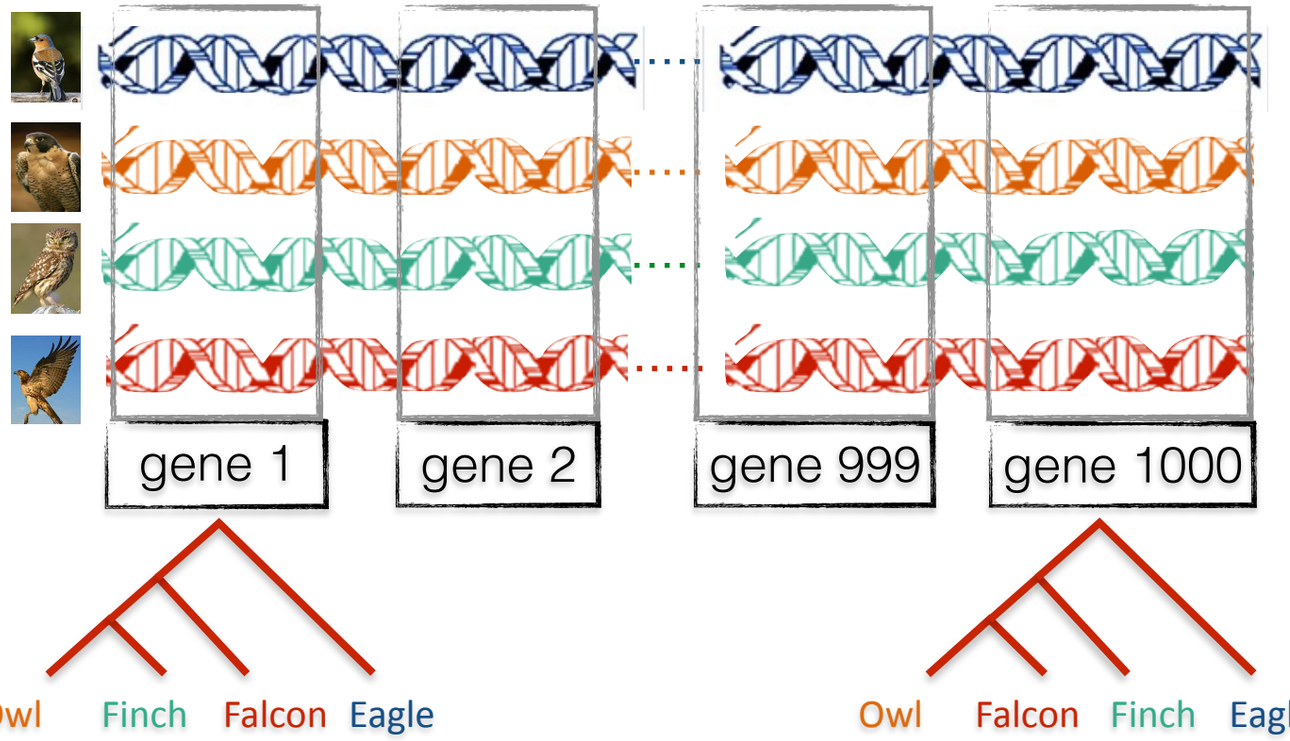


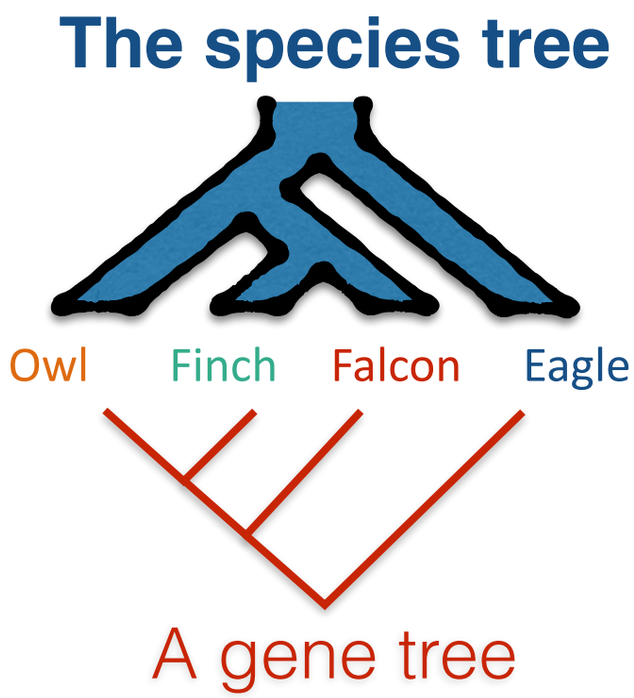
gene:
recombination-free orthologous regions in genomes



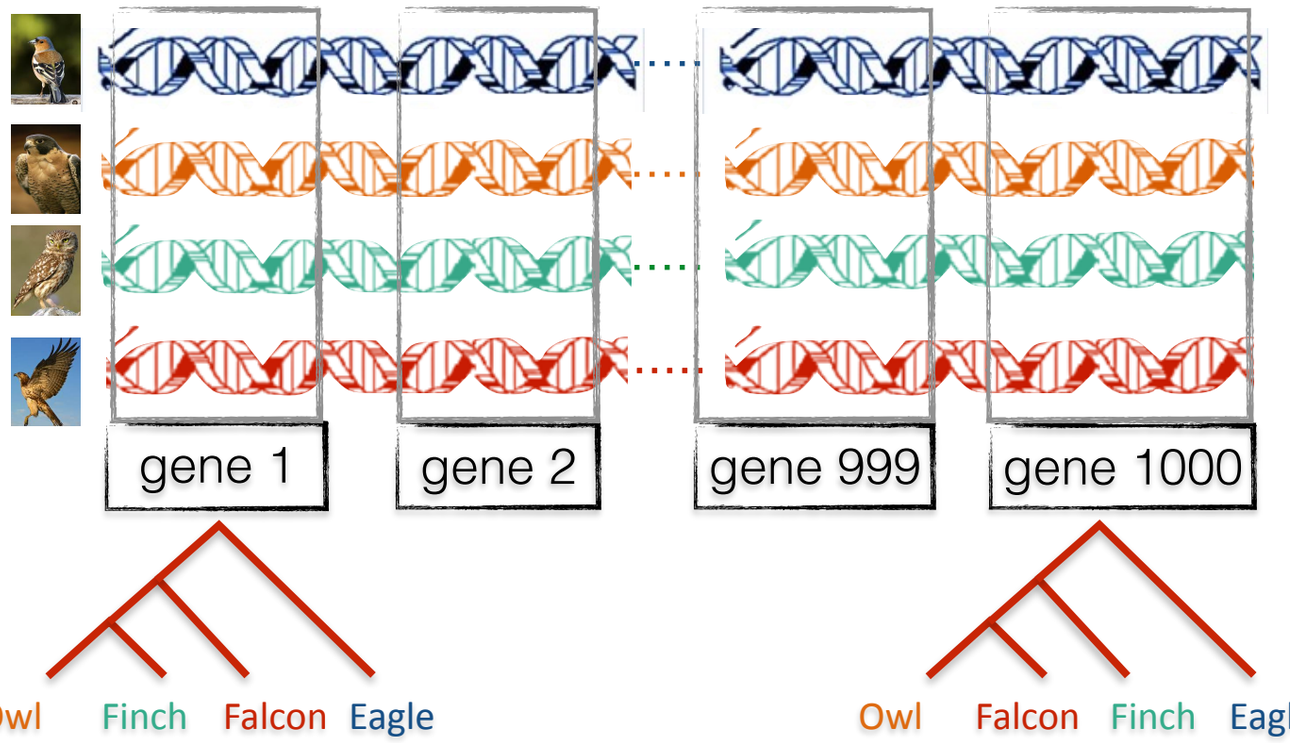


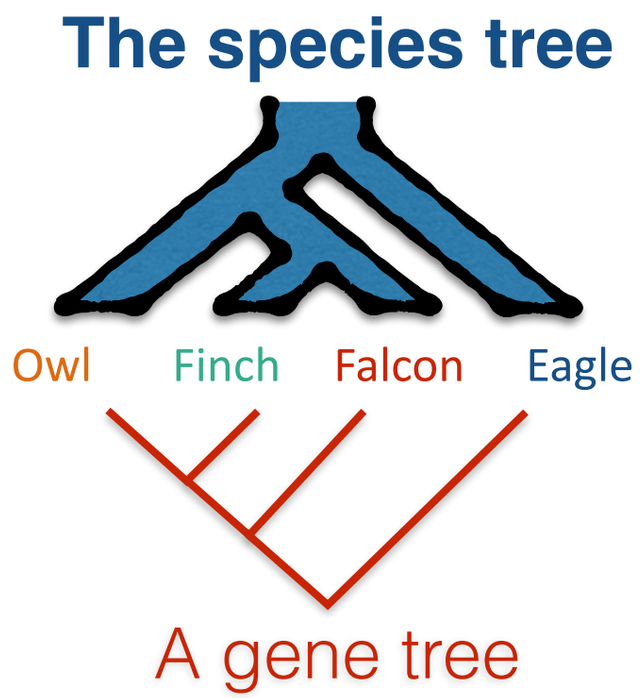
Gene tree discordance





Gene tree discordance



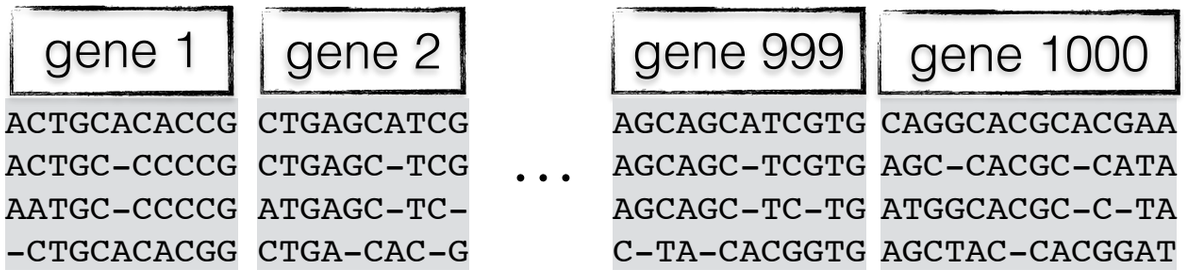


Causes of gene tree discordance:

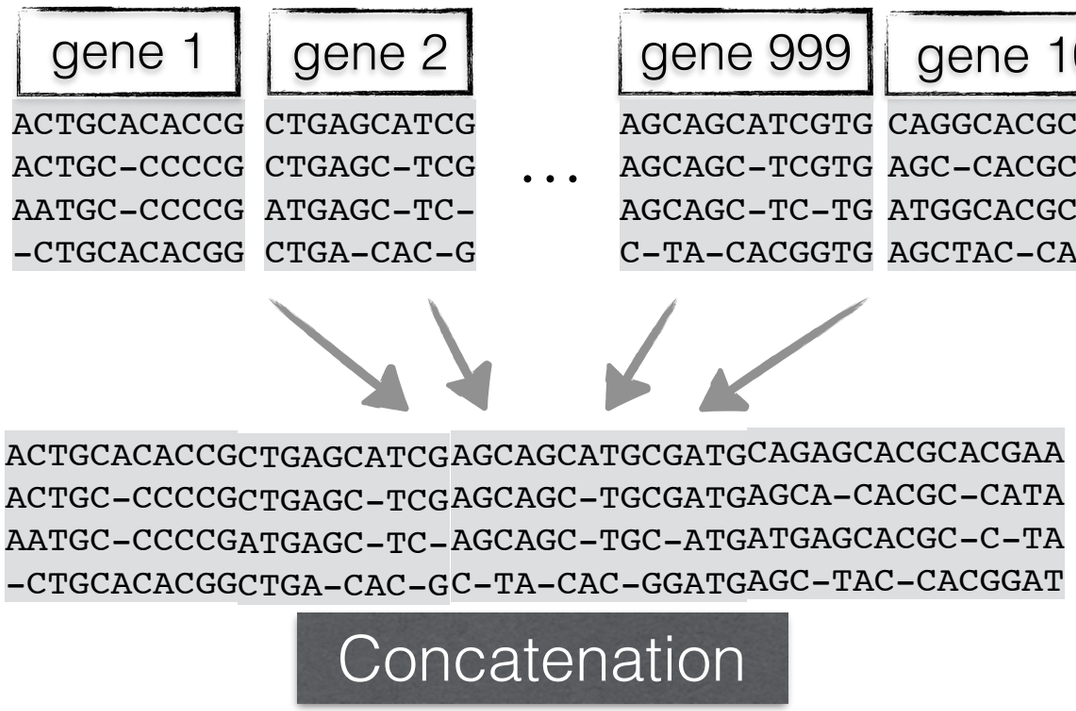
- Incomplete Lineage Sorting (ILS)
- Duplication and loss
- Horizontal Gene Transfer (HGT)

- Modeled by multi-species coalescent
- Highly probable for radiations (e.g., short branches) such as the bird radiation; 60 mya
- The species is identifiable from the gene tree distribution [Degnan and Salter, 2005]

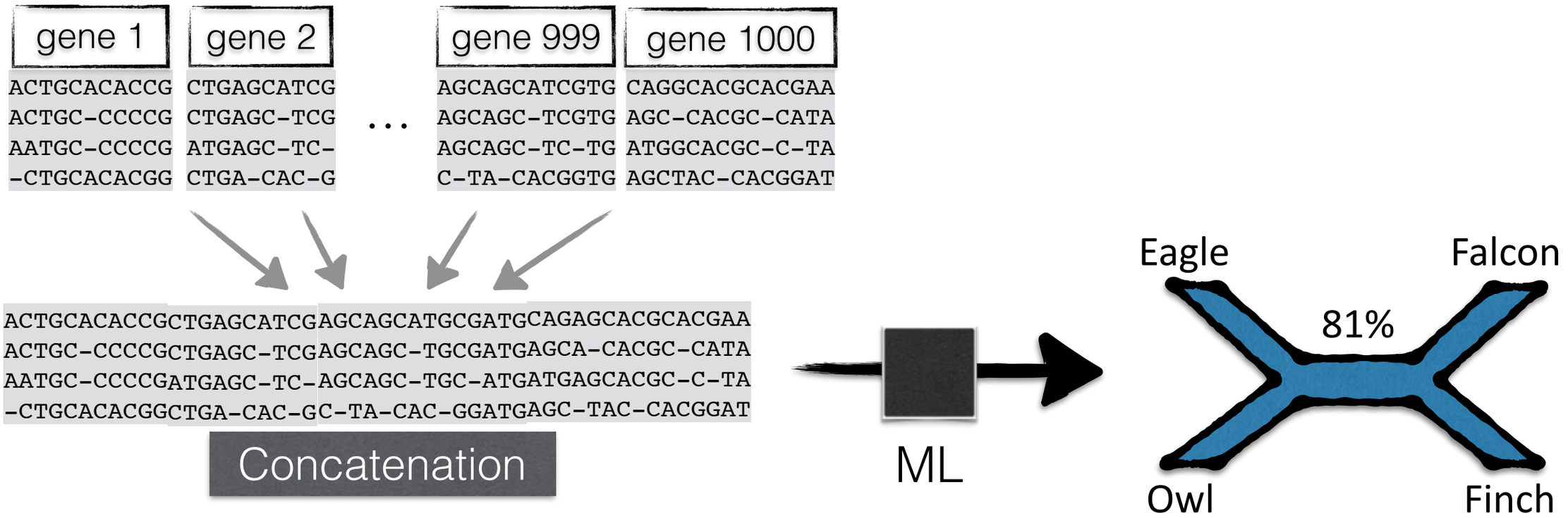
Species tree estimation from phylogenomic data (approach 1: concatenation)



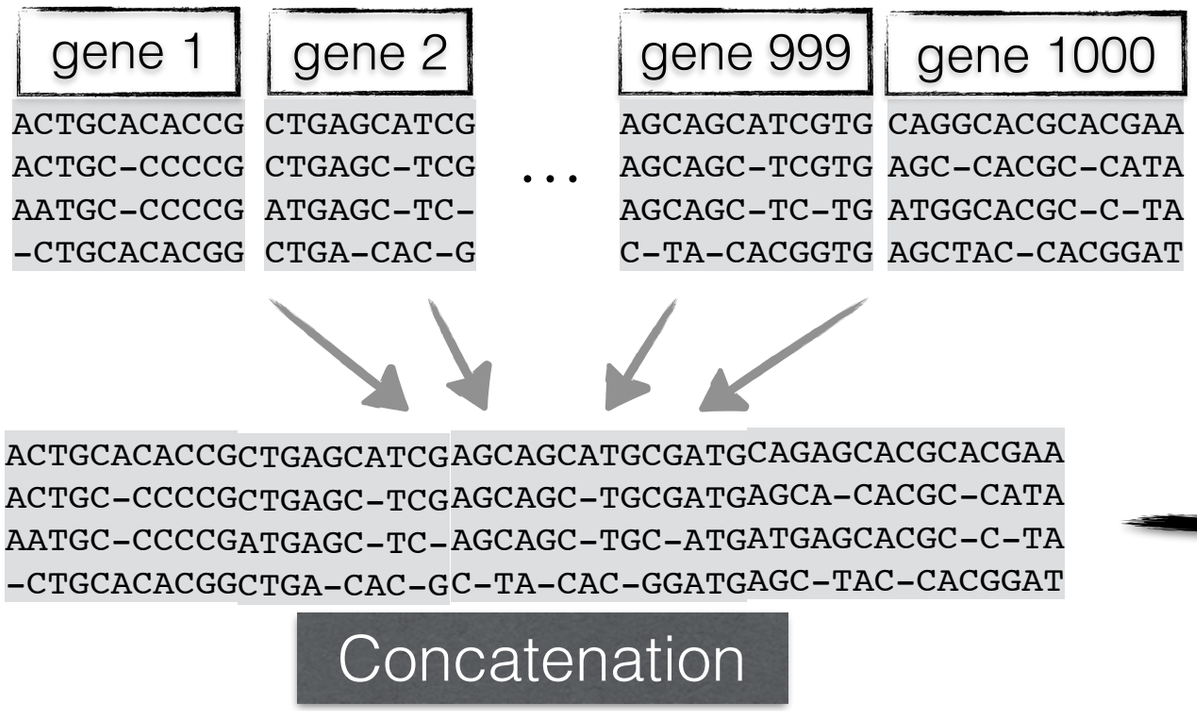
Species tree estimation from phylogenomic data (approach 1: concatenation)



Species tree estimation from phylogenomic data (approach 1: concatenation)

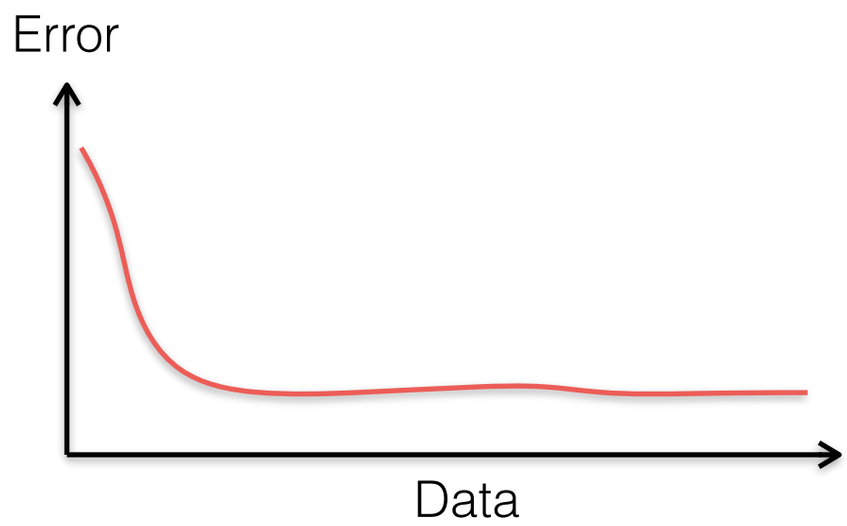


Species tree estimation from phylogenomic data (approach 1: concatenation)

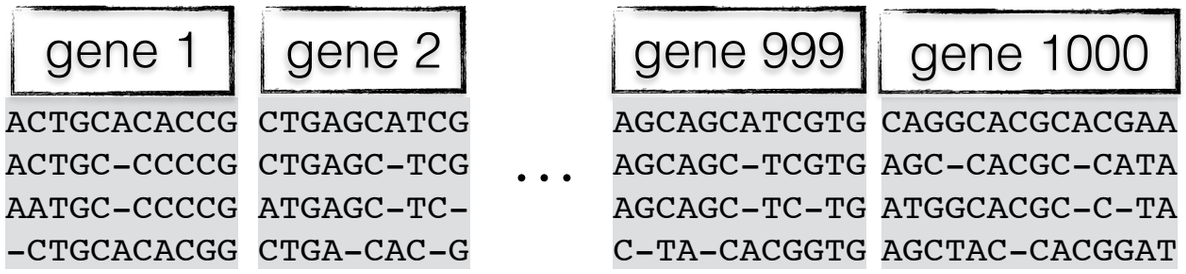


- Statistically <u>inconsistent</u> & positively misleading
  [Roch and Steel, Theo. Pop. Gen., 2014]

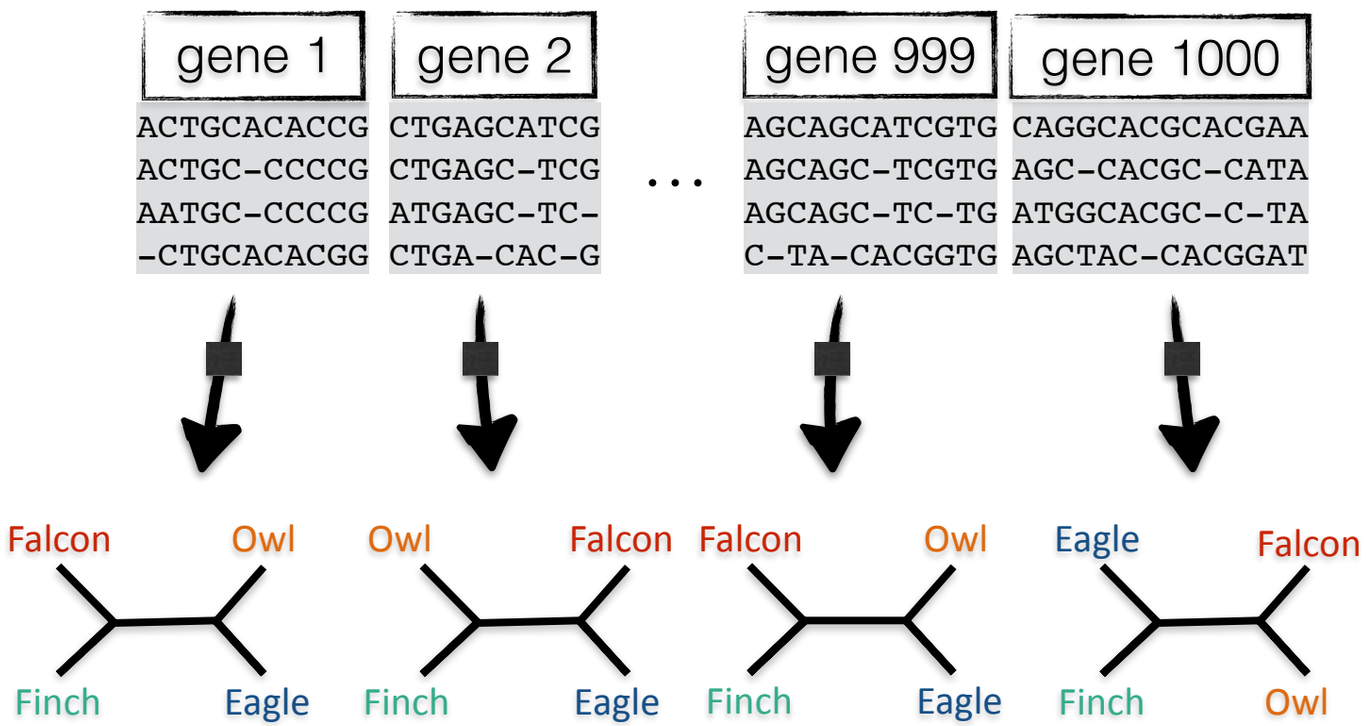
- Mixed accuracy in simulations
  [Kubatko and Degnan, Systematic Biology, 2007]
  [<u>Mirarab</u>, et al., Systematic Biology, 2014]



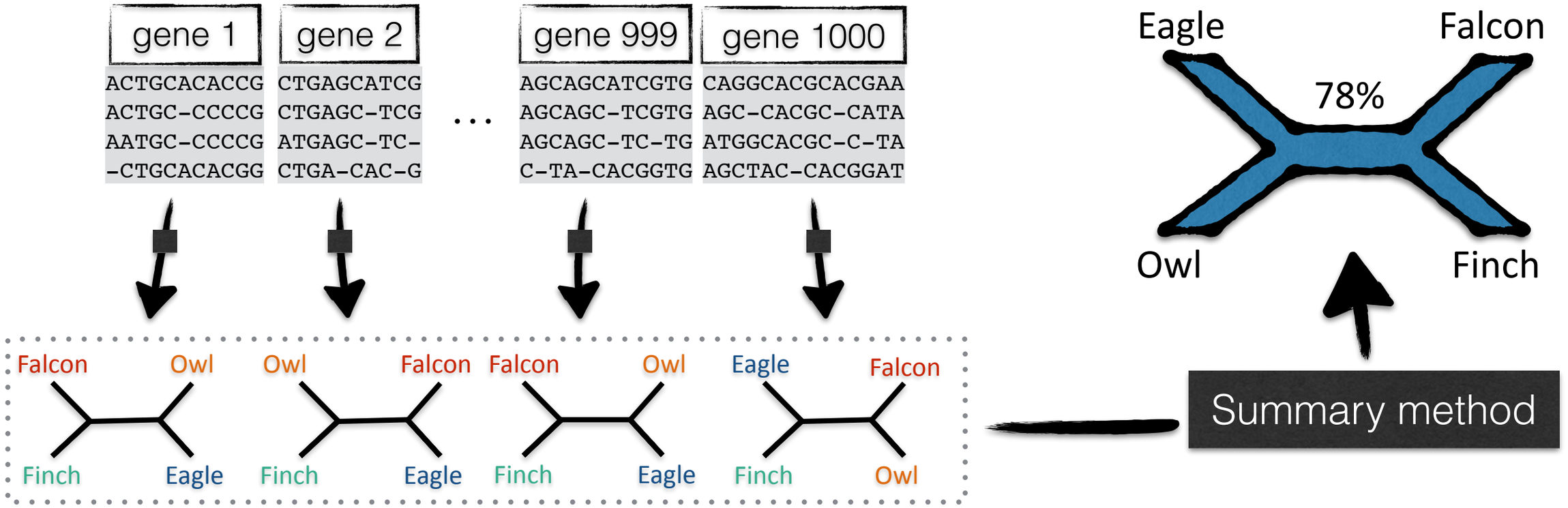
Species tree estimation from phylogenomic data (approach 2: summary methods)



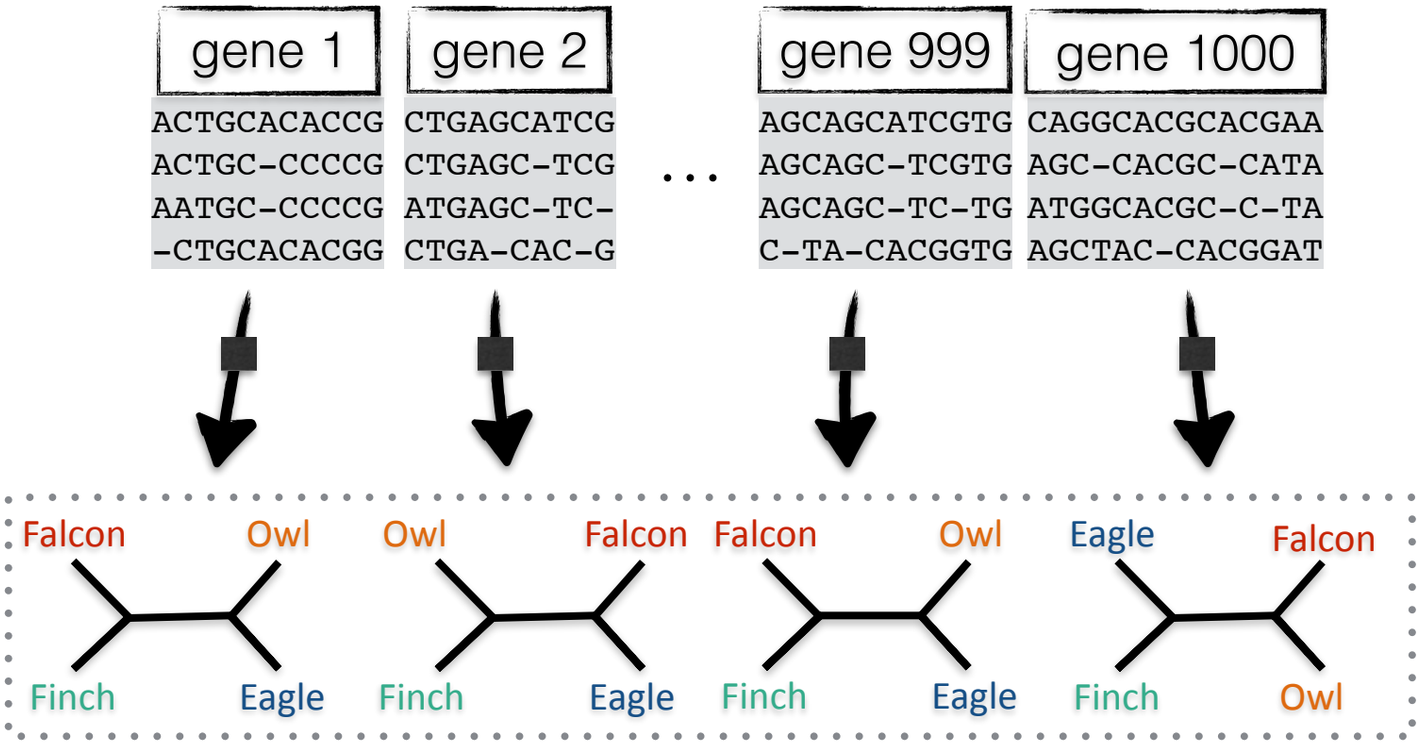
Species tree estimation from phylogenomic data (approach 2: summary methods)

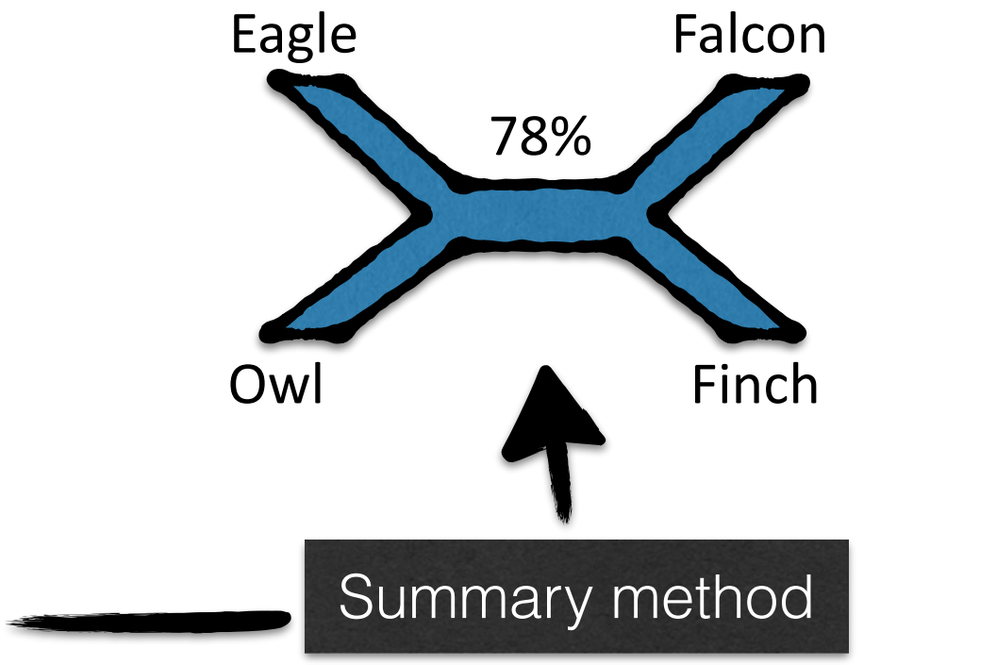


Species tree estimation from phylogenomic data (approach 2: summary methods)



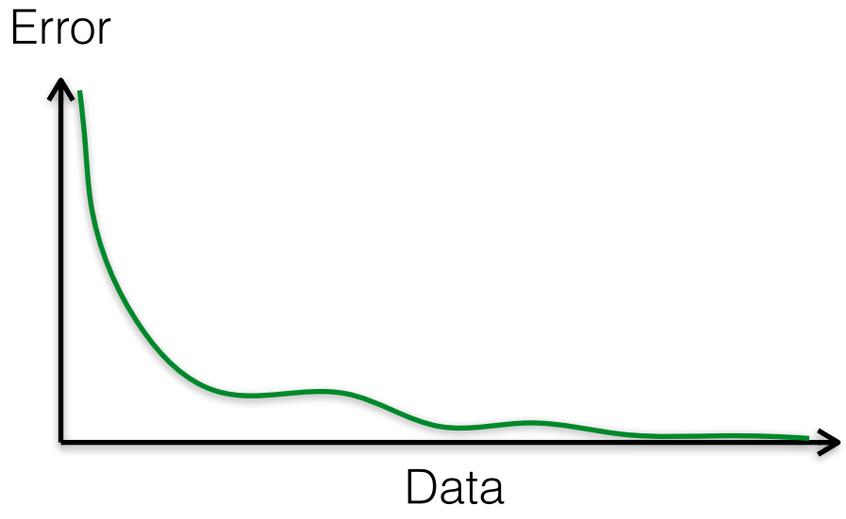
Species tree estimation from phylogenomic data (approach 2: summary methods)



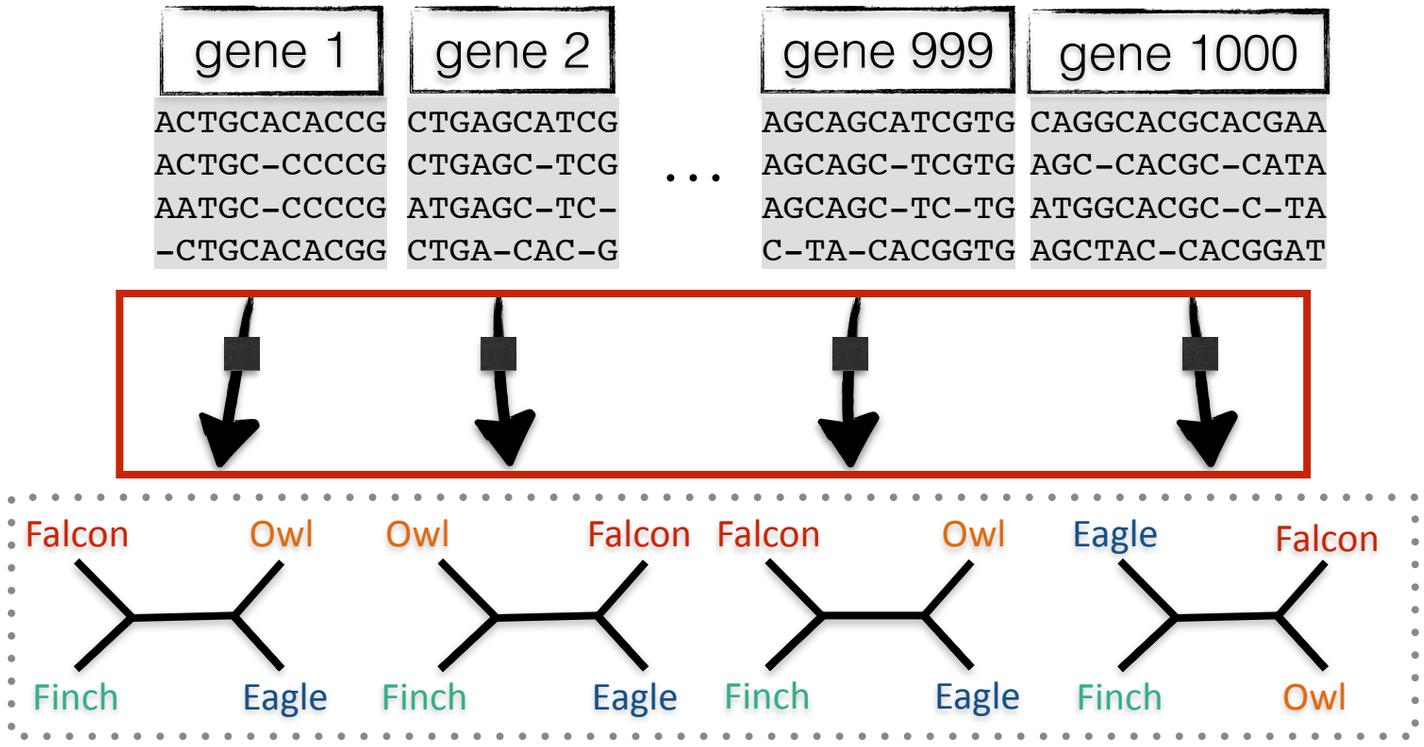


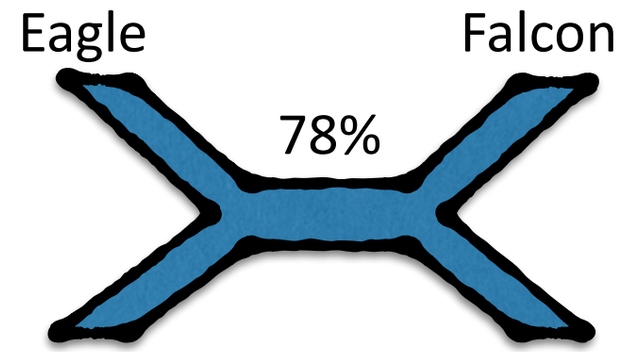
Can be statistically consistent

- **MP-EST** (maximum pseudo-likelihood)
  [Liu, Yu, Edwards, BMC Evol. Bio., 2010]
- BUCKy-pop., NJst, STAR, ASTRAL, …



Species tree estimation from phylogenomic data (approach 2: summary methods)

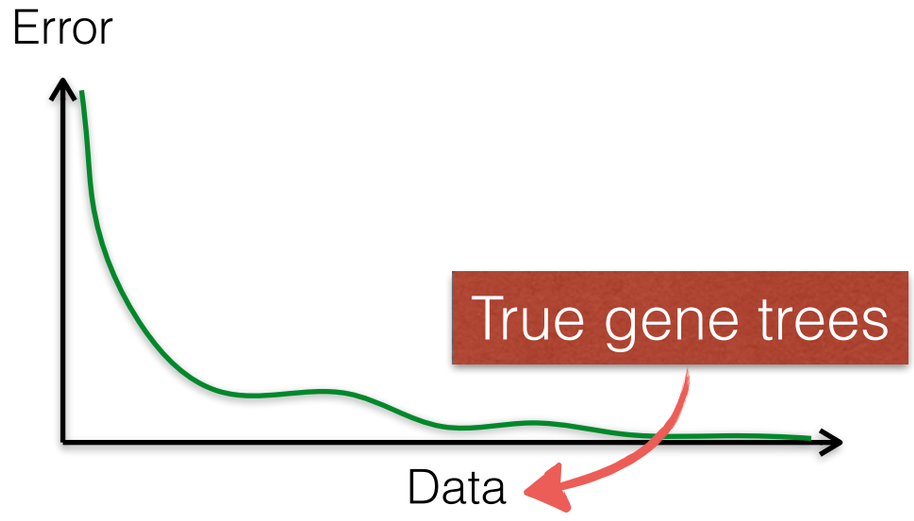




Summary method

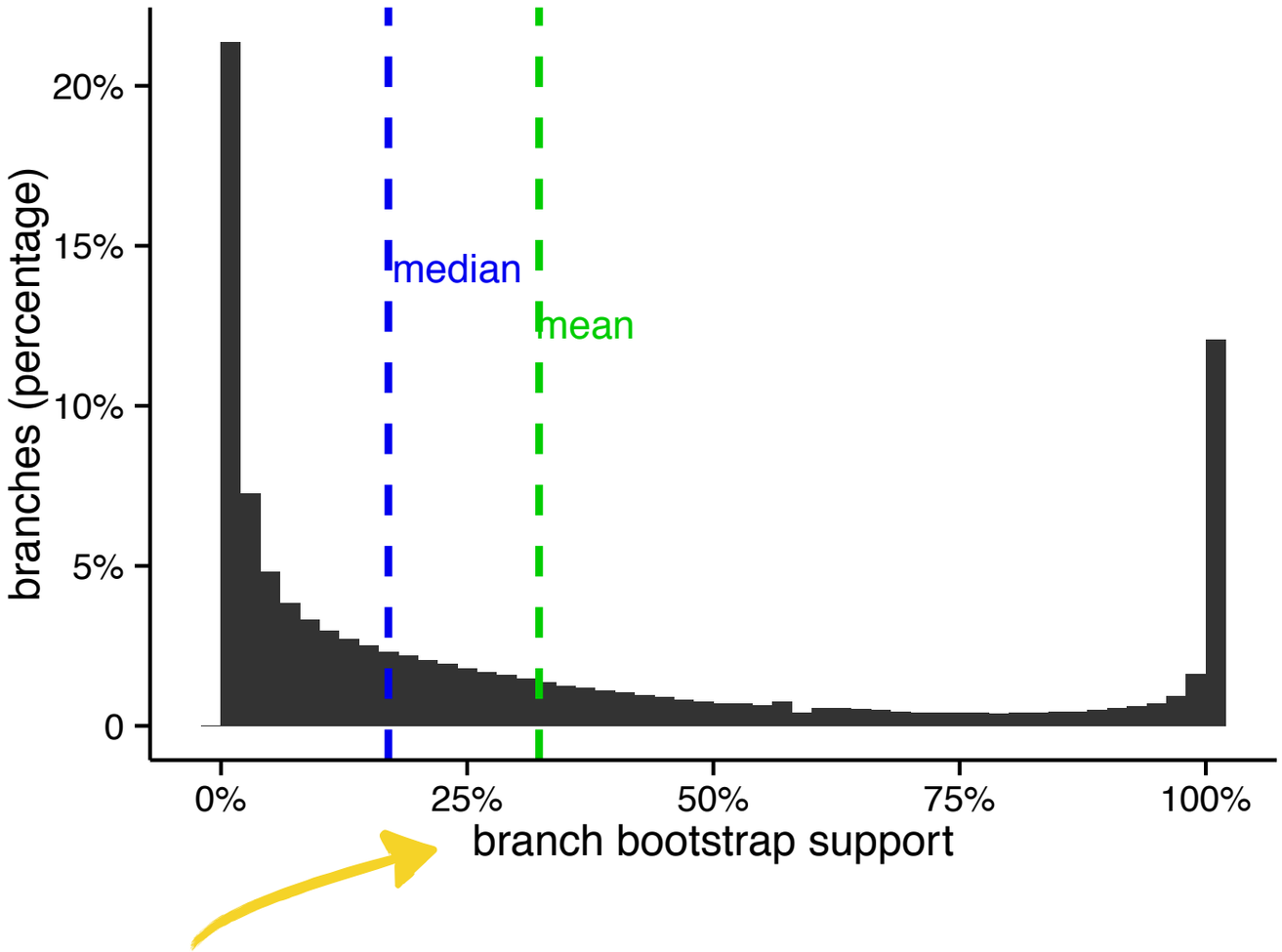
Can be statistically consistent

- **MP-EST** (maximum pseudo-likelihood)
  [Liu, Yu, Edwards, BMC Evol. Bio., 2010]
- BUCKy-pop., NJst, STAR, ASTRAL, …



Gene trees on the avian dataset

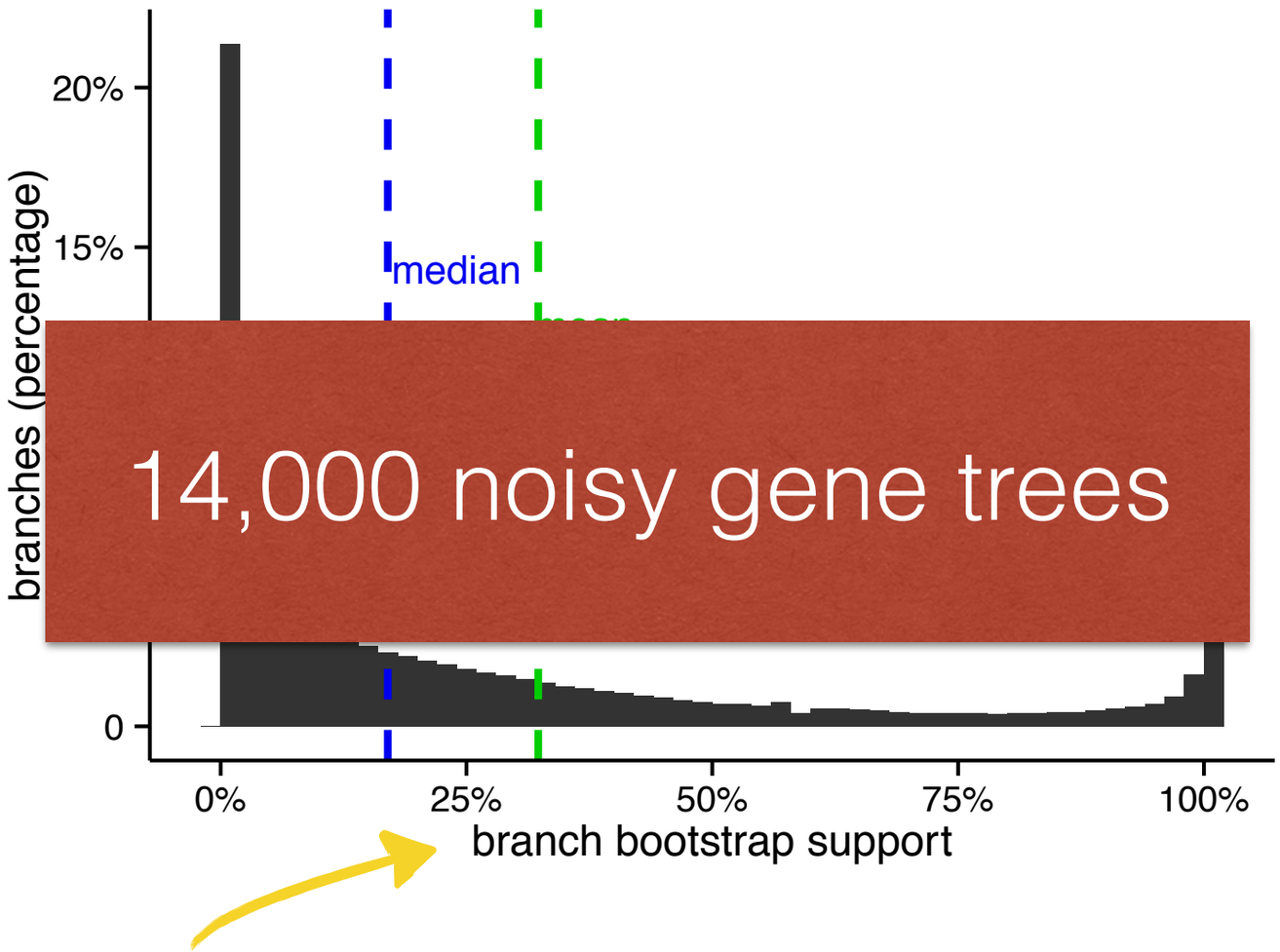
14,000 “genes”: 8,000 exons and 2,500 introns
3,500 Ultra-Conserved Elements



A measure of confidence in estimated gene tree branches

Gene trees on the avian dataset

14,000 “genes”: 8,000 exons and 2,500 introns
3,500 Ultra-Conserved Elements

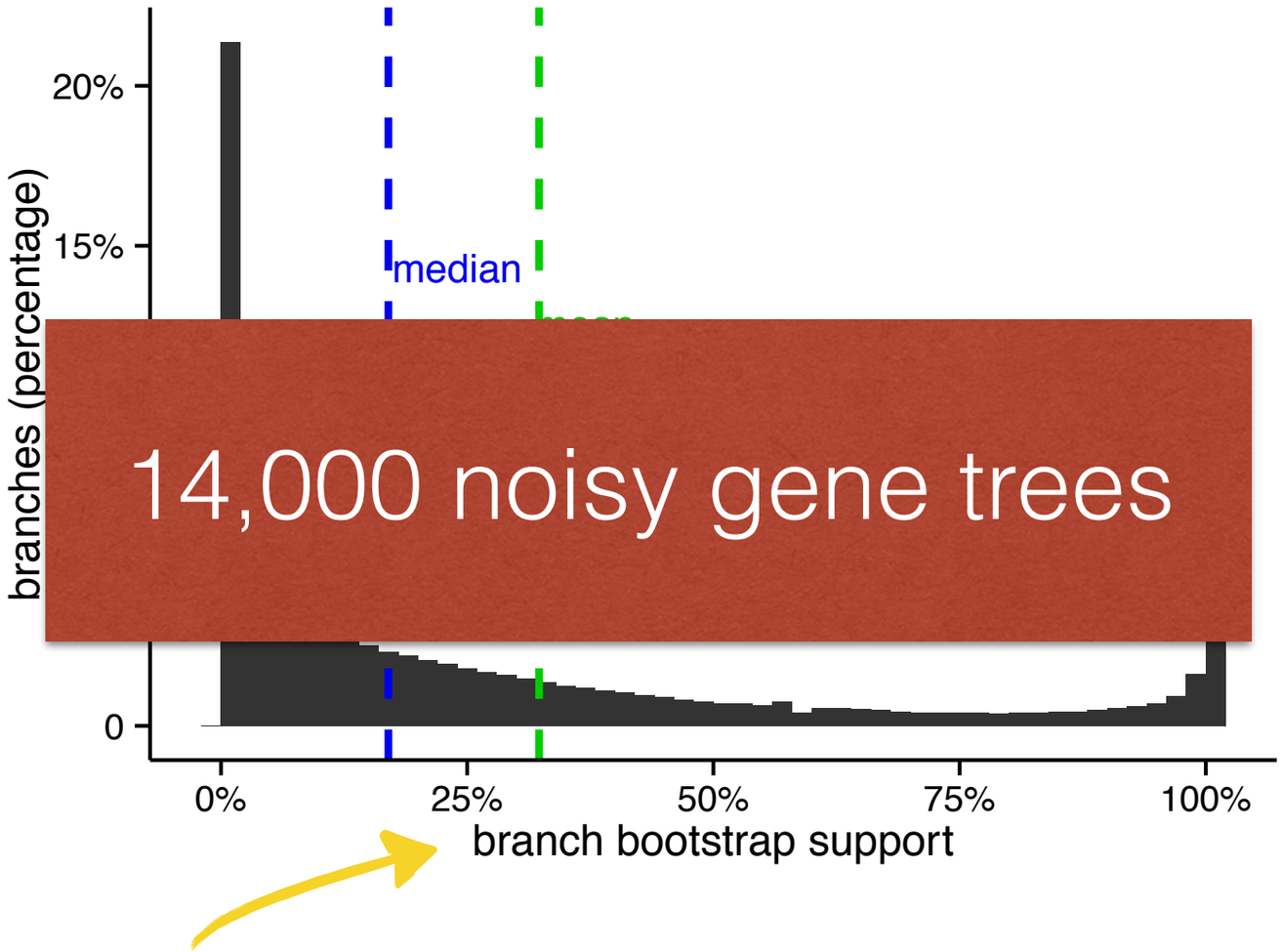


14,000 noisy gene trees

A measure of confidence in estimated gene tree branches

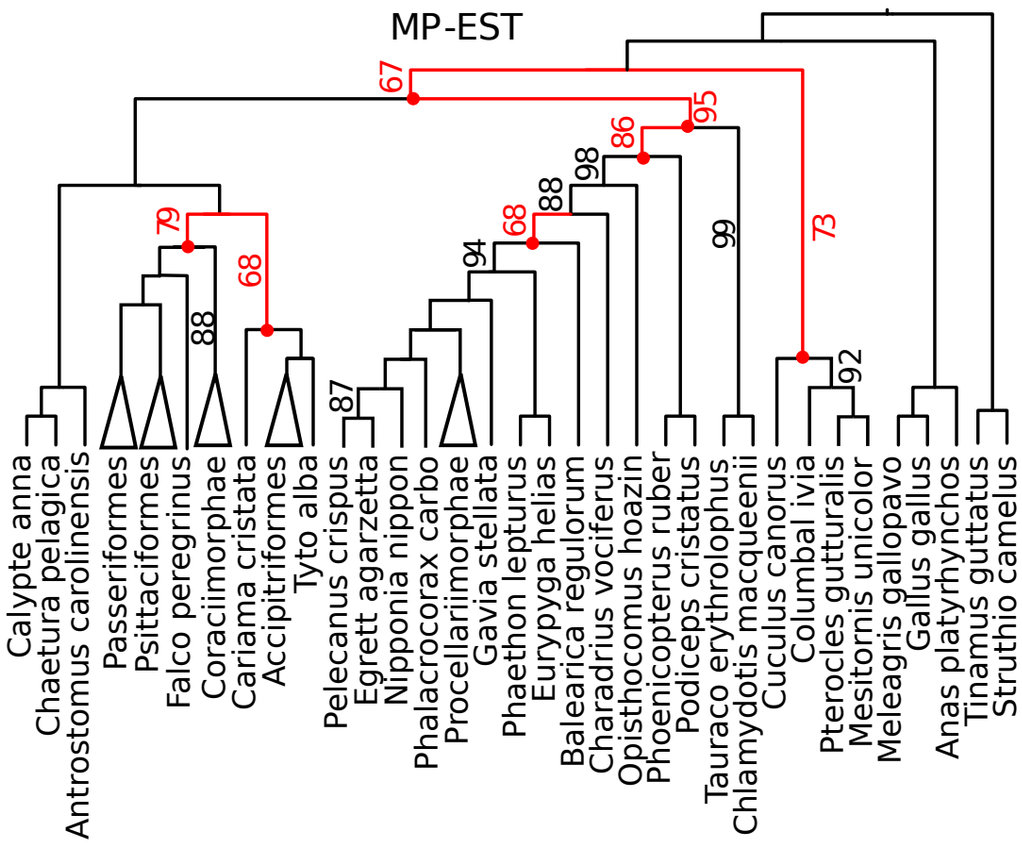
Gene trees on the avian dataset

14,000 “genes”: 8,000 exons and 2,500 introns
3,500 Ultra-Conserved Elements

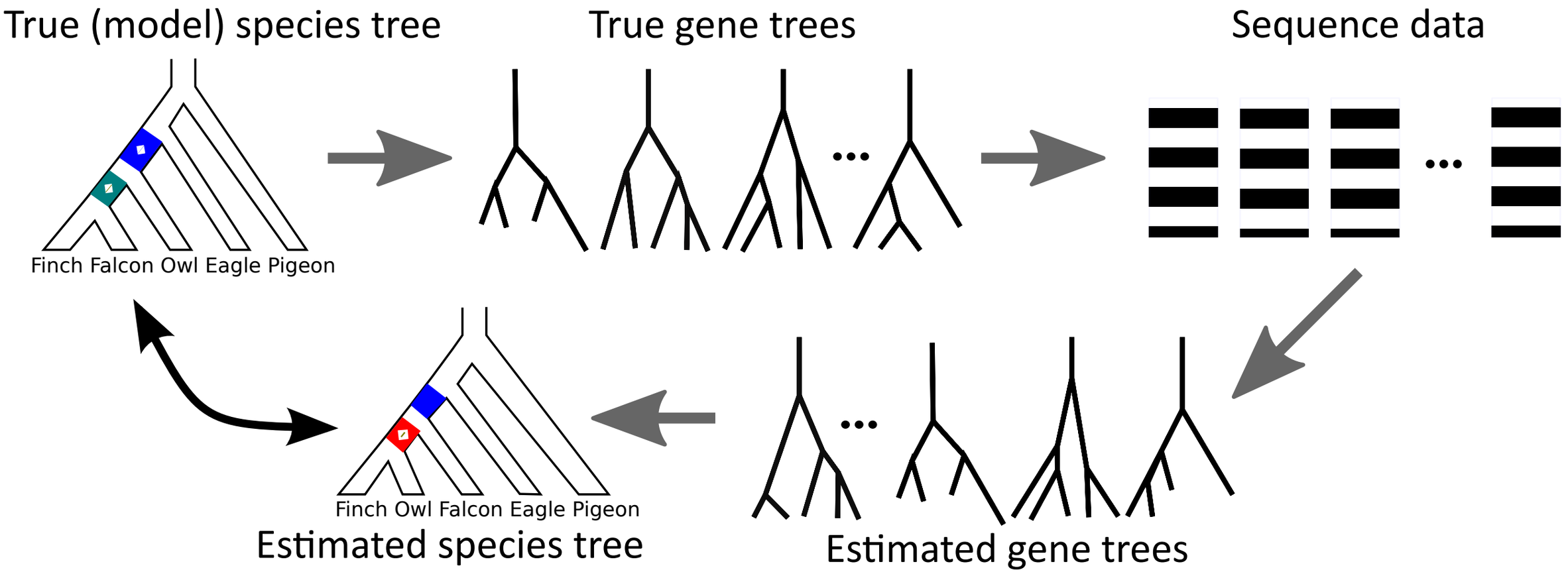


14,000 noisy gene trees

A measure of confidence in estimated gene tree branches



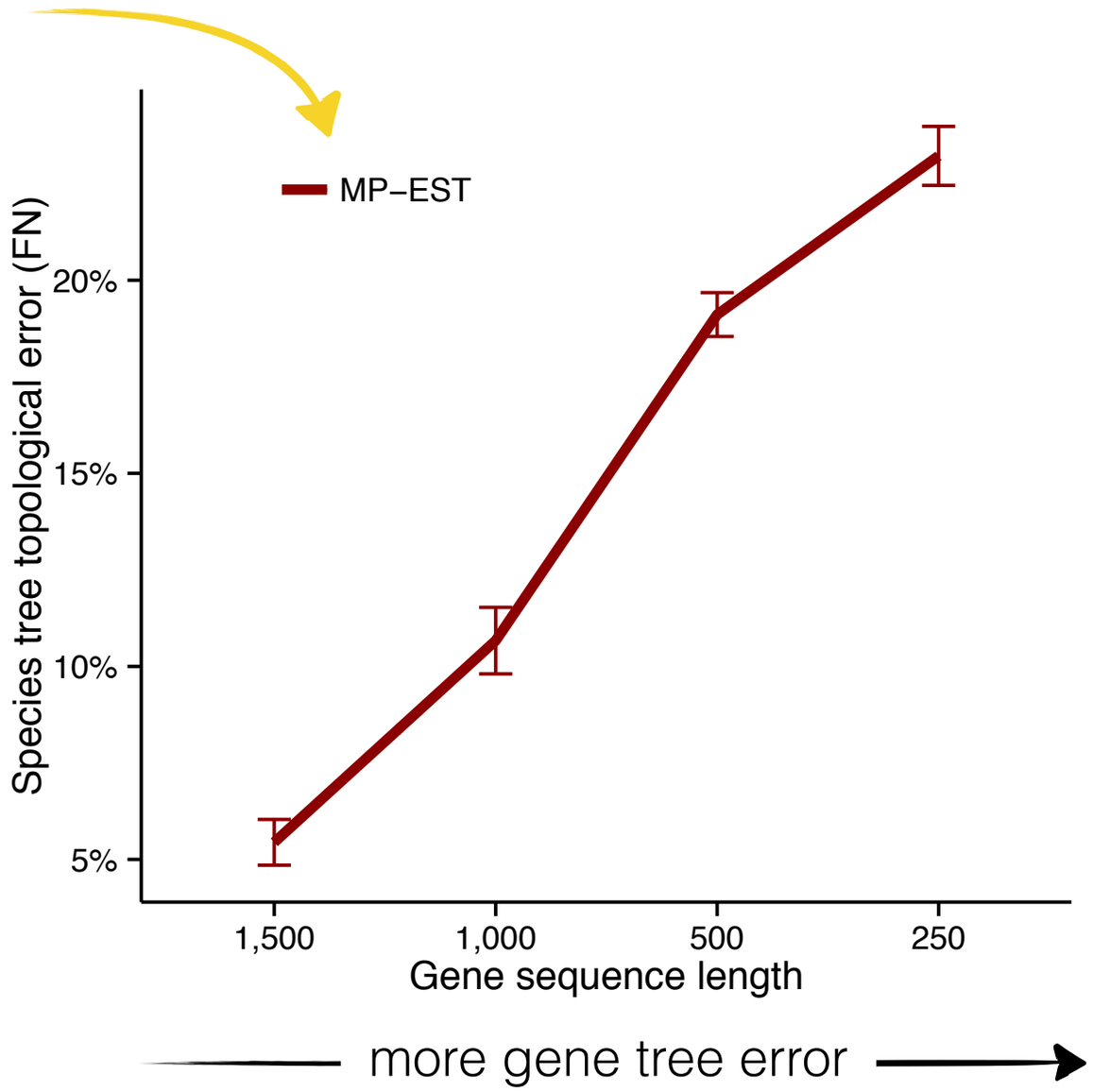
Simulation studies



Error metric: percentage of branches in true tree that are missing from the estimated tree

Gene trees on the avian dataset

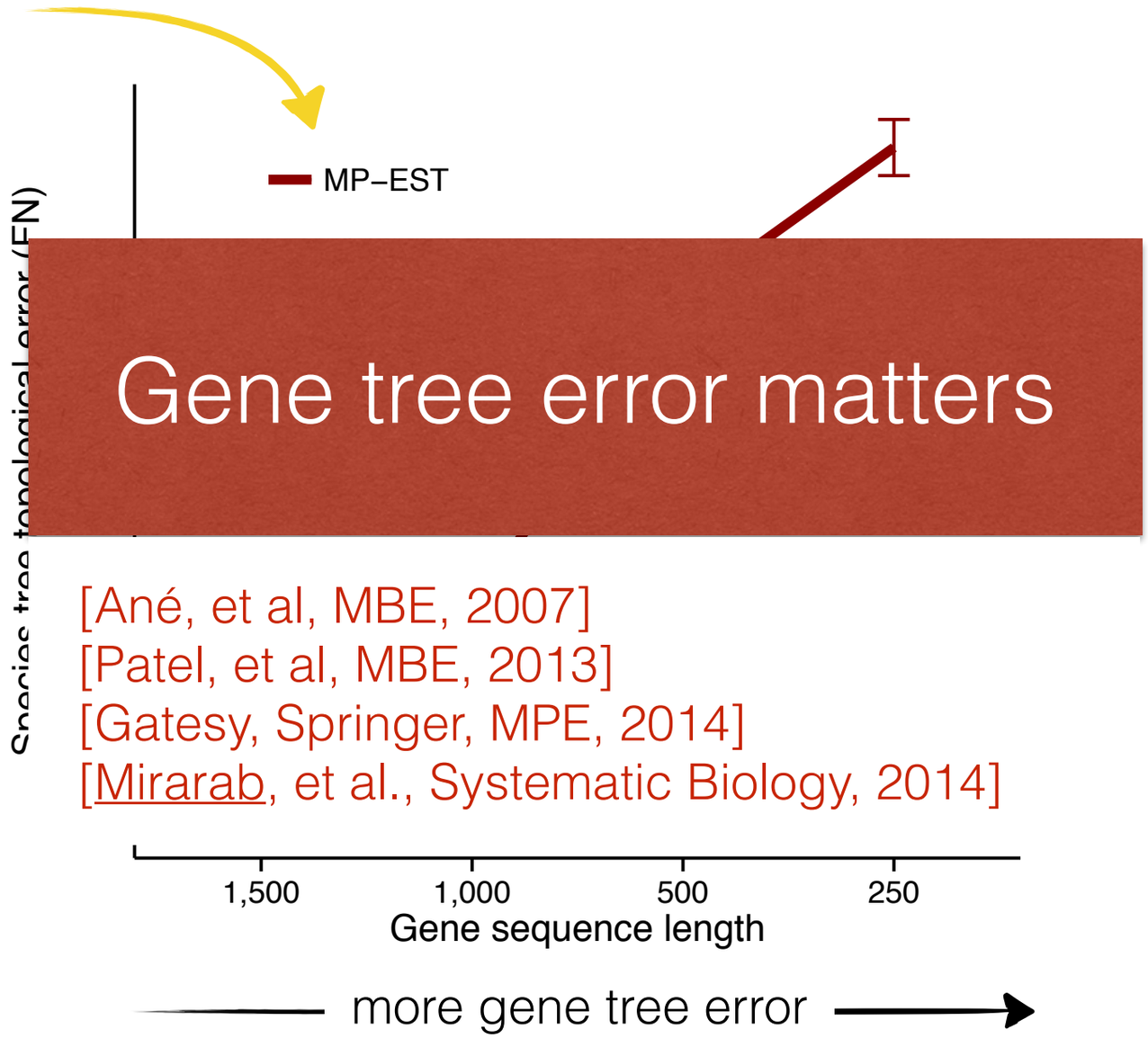
A statistically consistent summary method



Avian-like simulations (1000 genes)
[Mirarab, et al., Science, 2014]

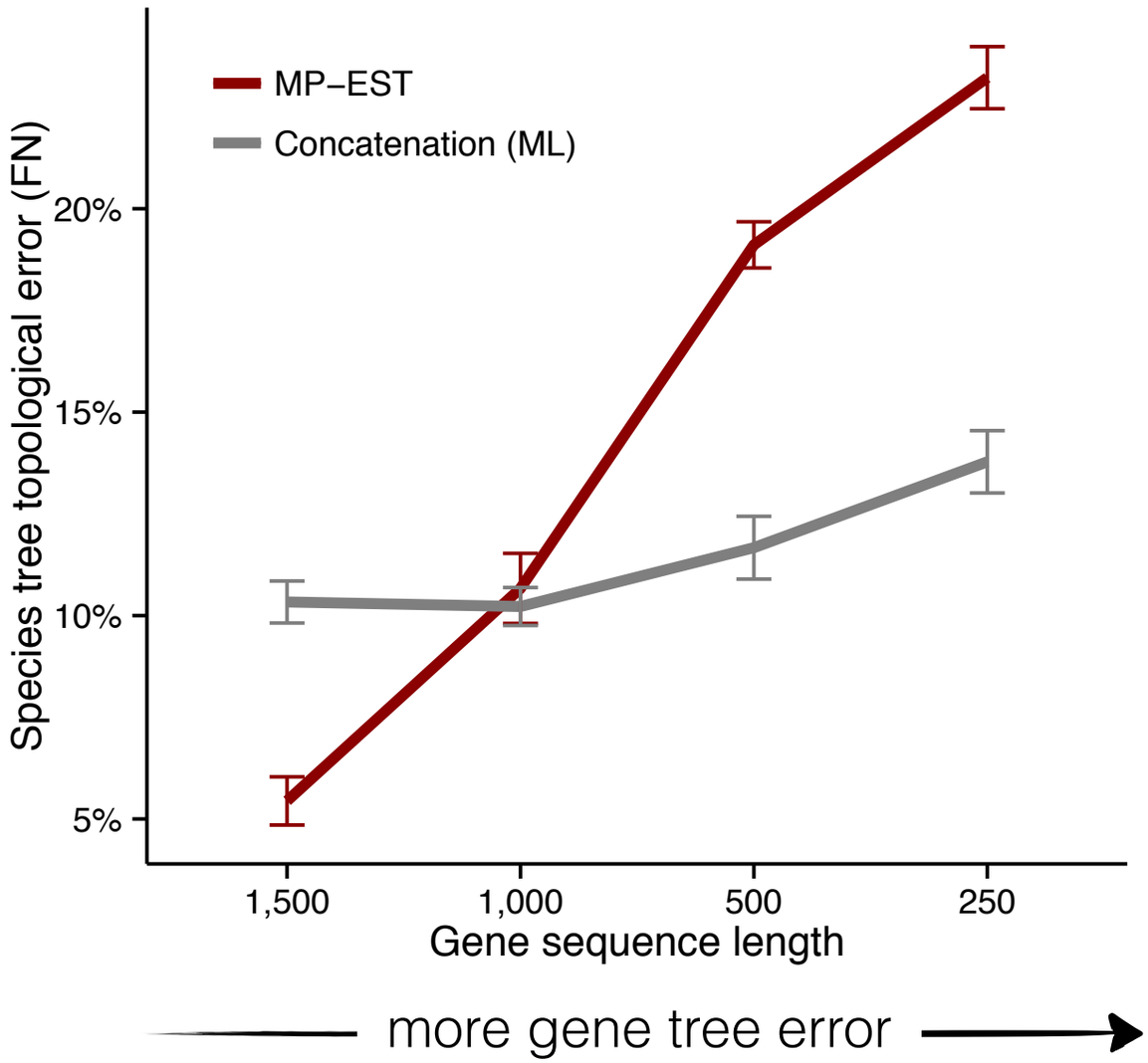
Gene trees on the avian dataset

A statistically consistent summary method



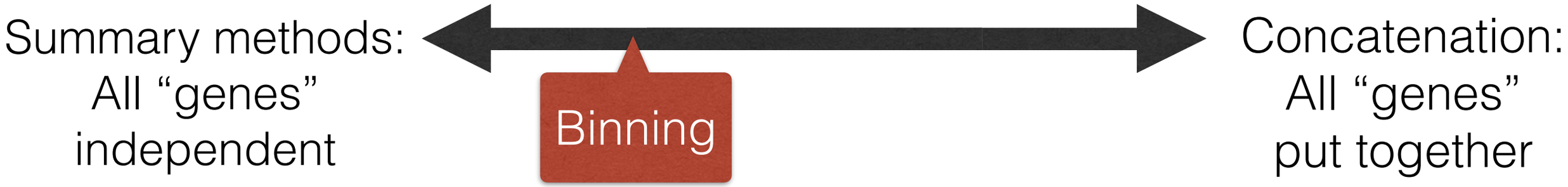
Avian-like simulations (1000 genes)
[Mirarab, et al., Science, 2014]

Gene trees on the avian dataset



Avian-like simulations (1000 genes)
[Mirarab, et al., Science, 2014]

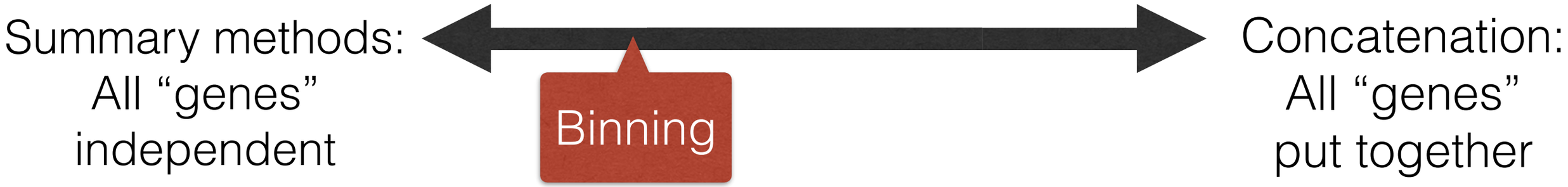
Statistical binning: idea



- Concatenation has good accuracy with low levels of ILS

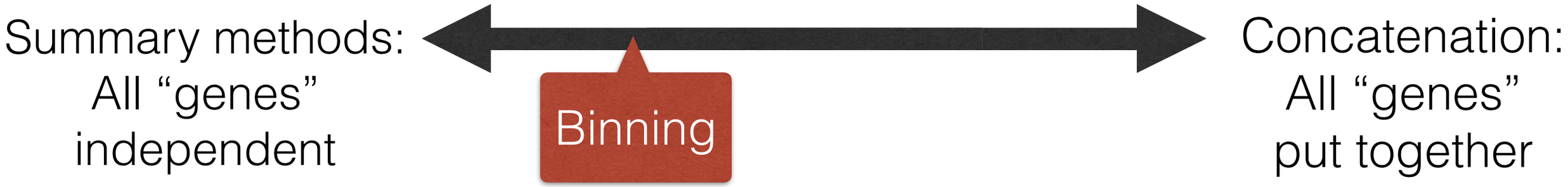
- Some pairs of genes are concordant (at least in topology)

Statistical binning: idea



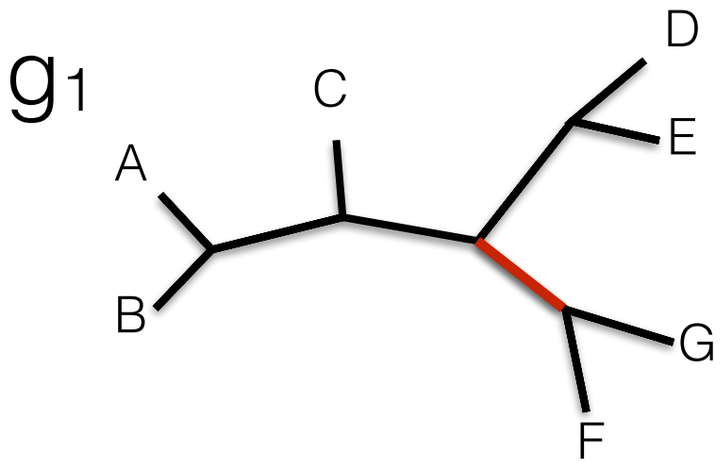
- Concatenation has good accuracy with low levels of ILS
- Some pairs of genes are concordant (at least in topology)
- Concatenate “combinable” sets of genes into “supergenes” to increase the phylogenetic signal

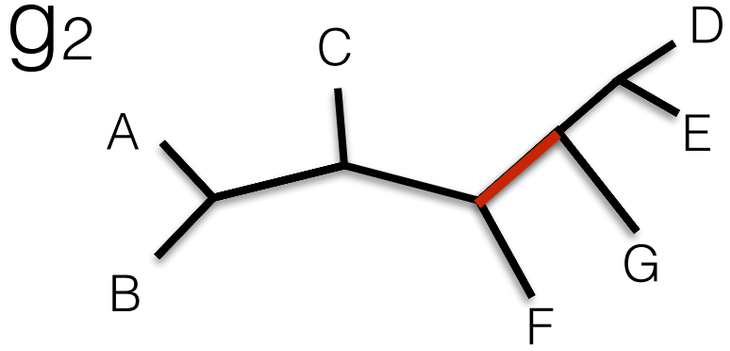
Statistical binning: idea



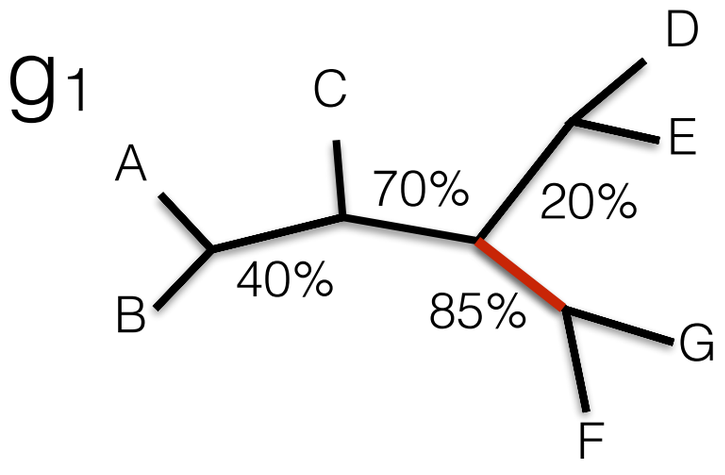
- Concatenation has good accuracy with low levels of ILS
- Some pairs of genes are concordant (at least in topology)
- Concatenate “combinable” sets of genes into “supergenes” to increase the phylogenetic signal
- How combinable genes are found gene tree estimation is hard?

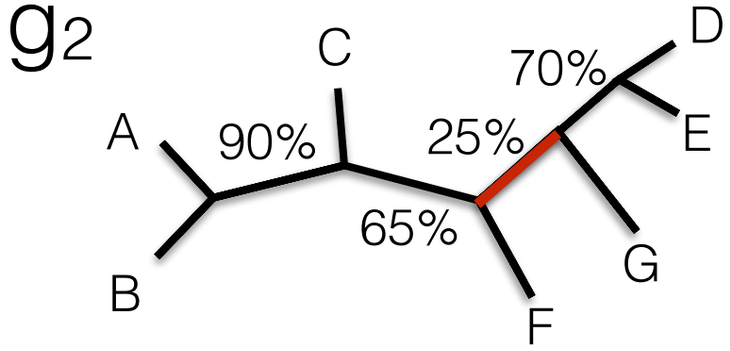
Statistical tests of combinability



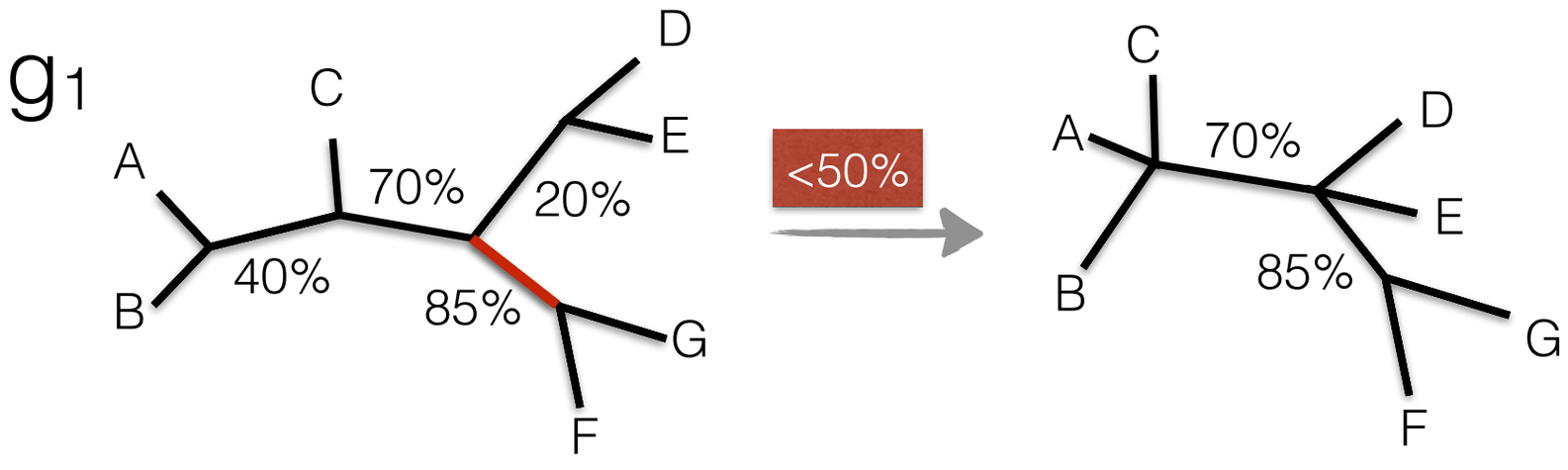


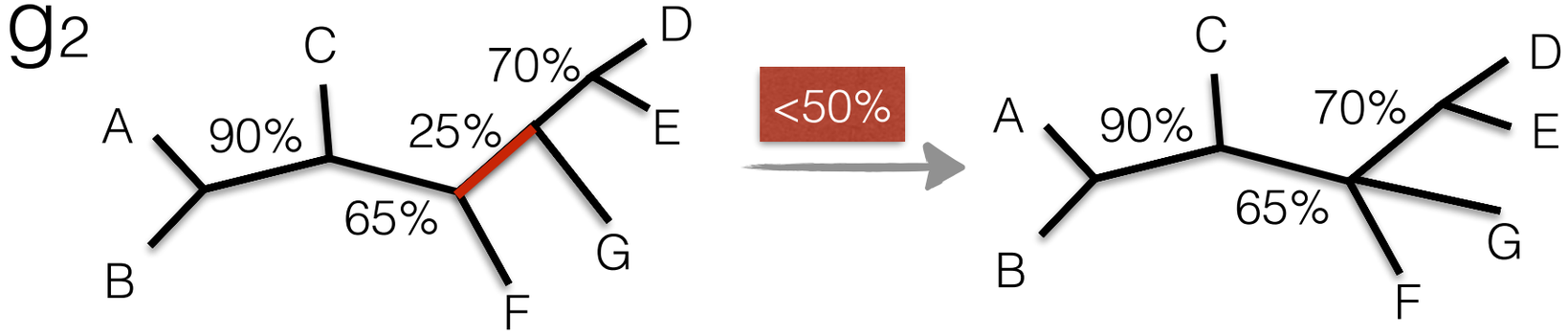
Statistical tests of combinability





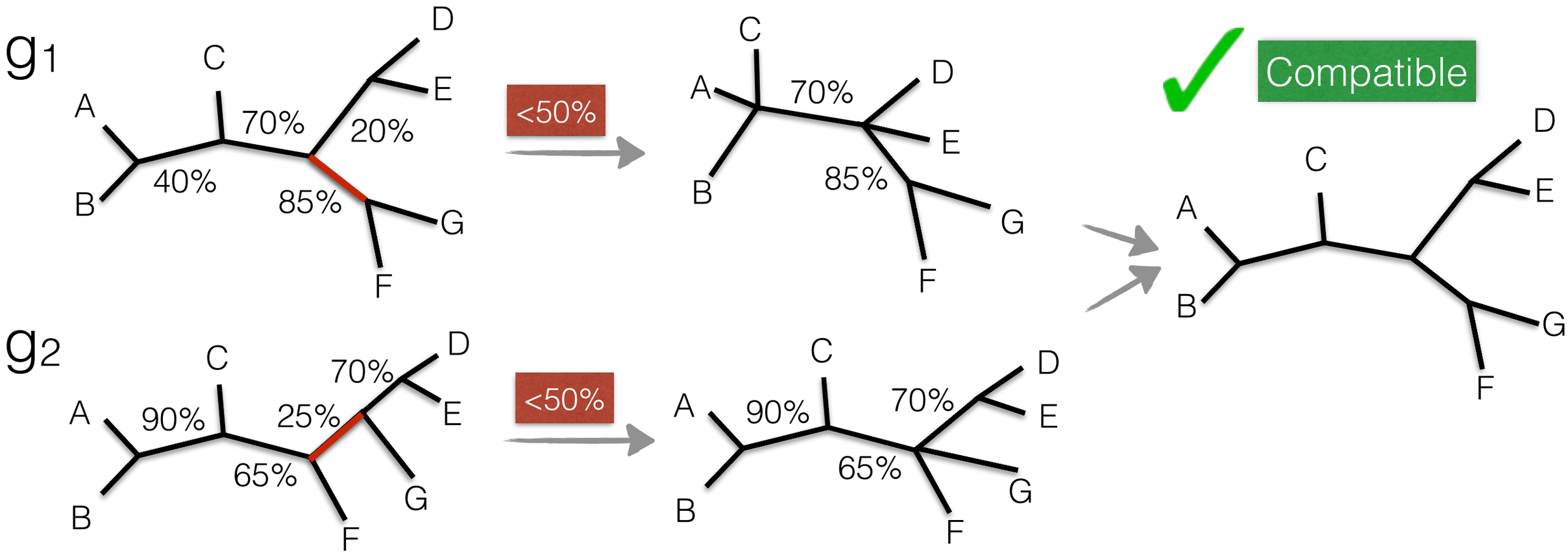
Statistical tests of combinability





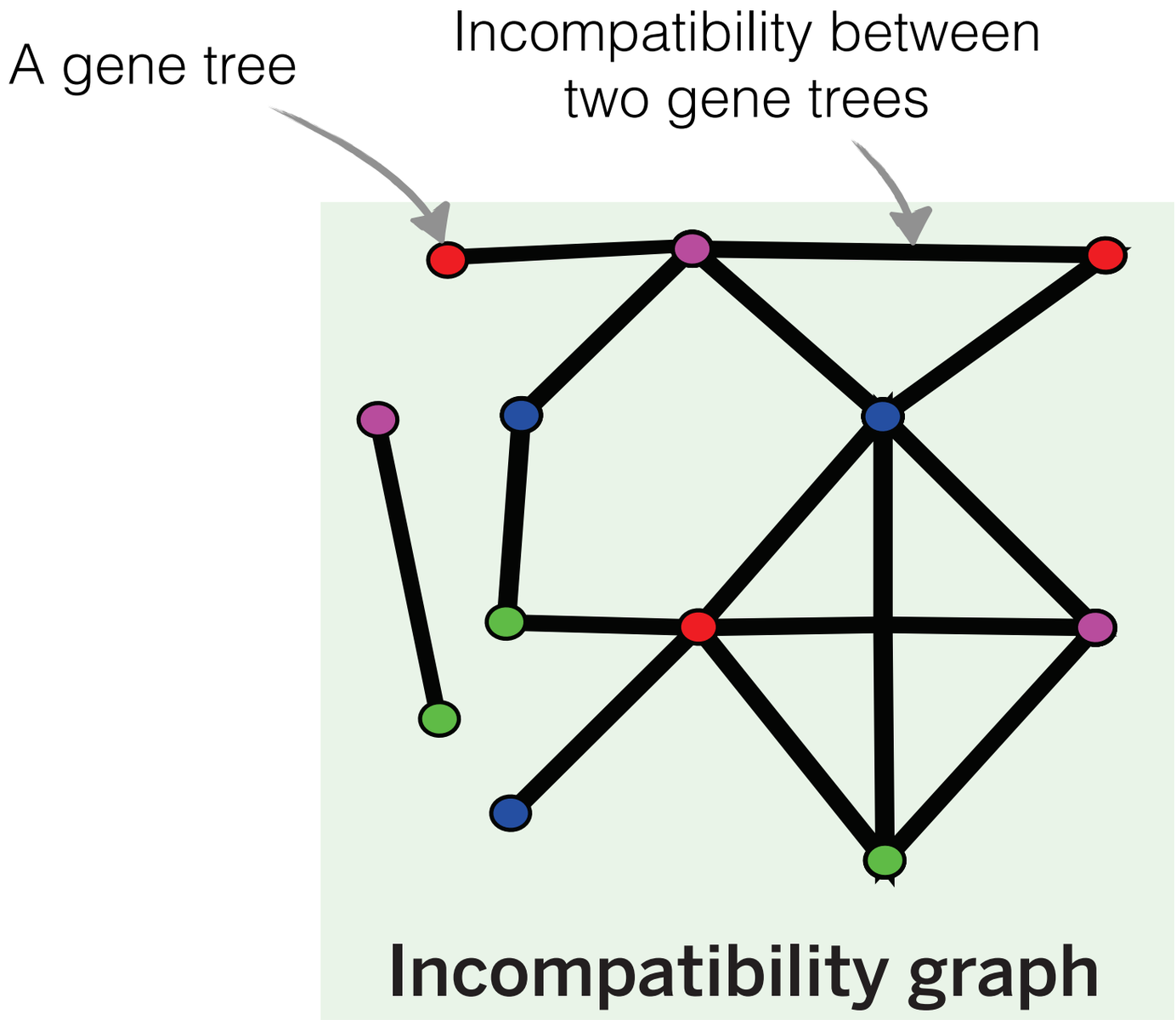
- Restrict genes to parts that have a minimum support

Statistical tests of combinability



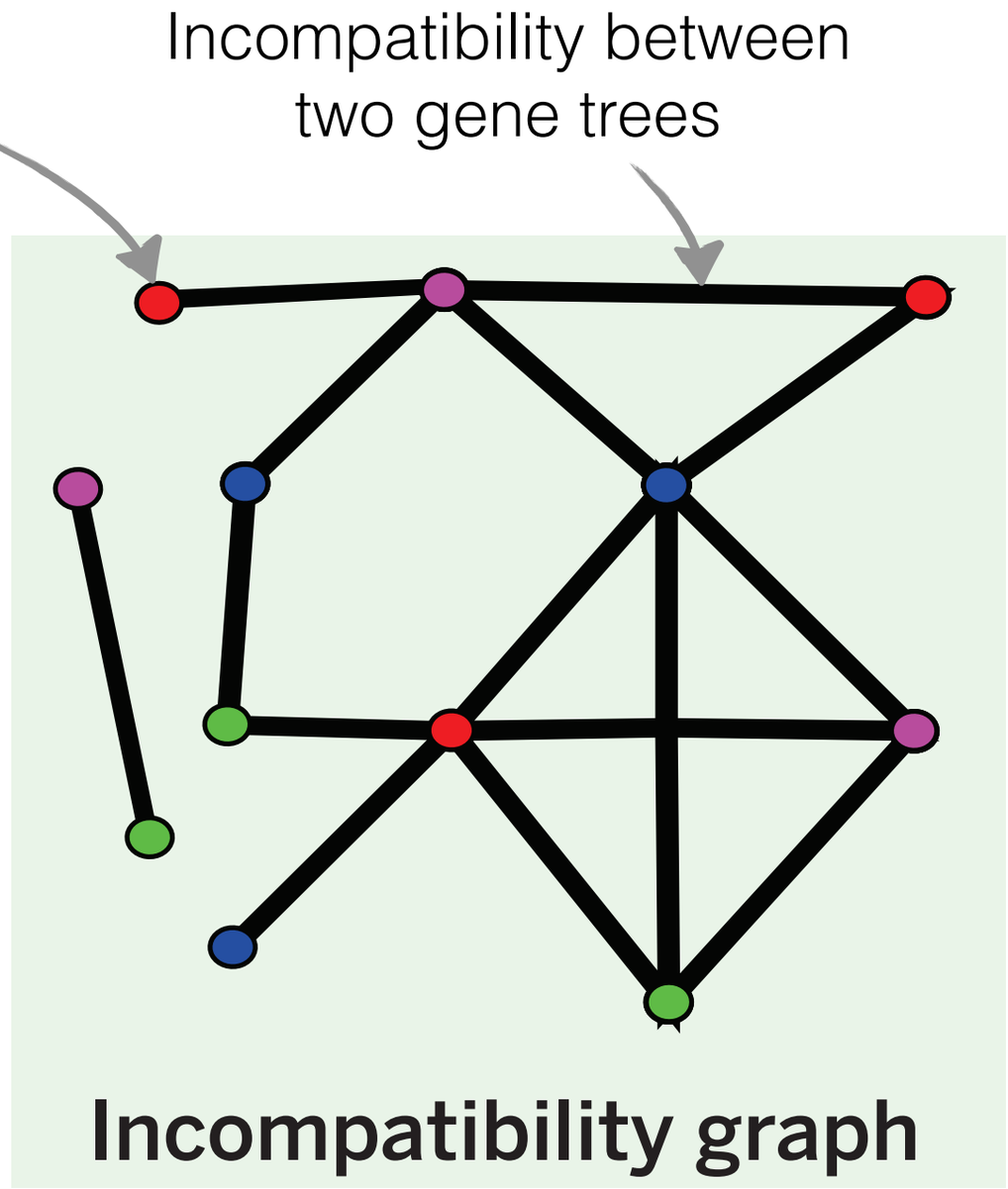
- Restrict genes to parts that have a minimum support
- Test combinability based on the supported parts of gene trees

Incompatibility graph

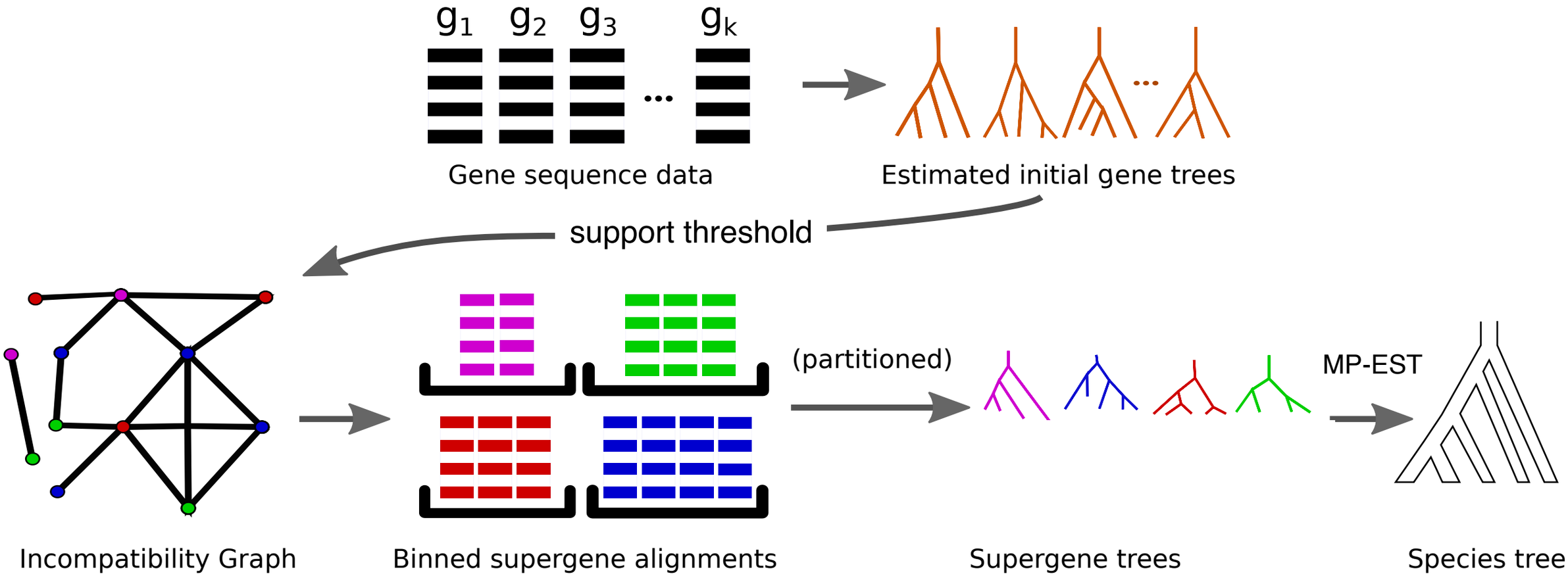


Incompatibility graph

- Find independent sets: sets with no edges between any pairs of nodes
  - Genes in each “bin” are all pairwise compatible
- Minimum vertex coloring (NP-hard)
  - Brélaz heuristics
  - Modified the heuristic to produce balanced bins where possible

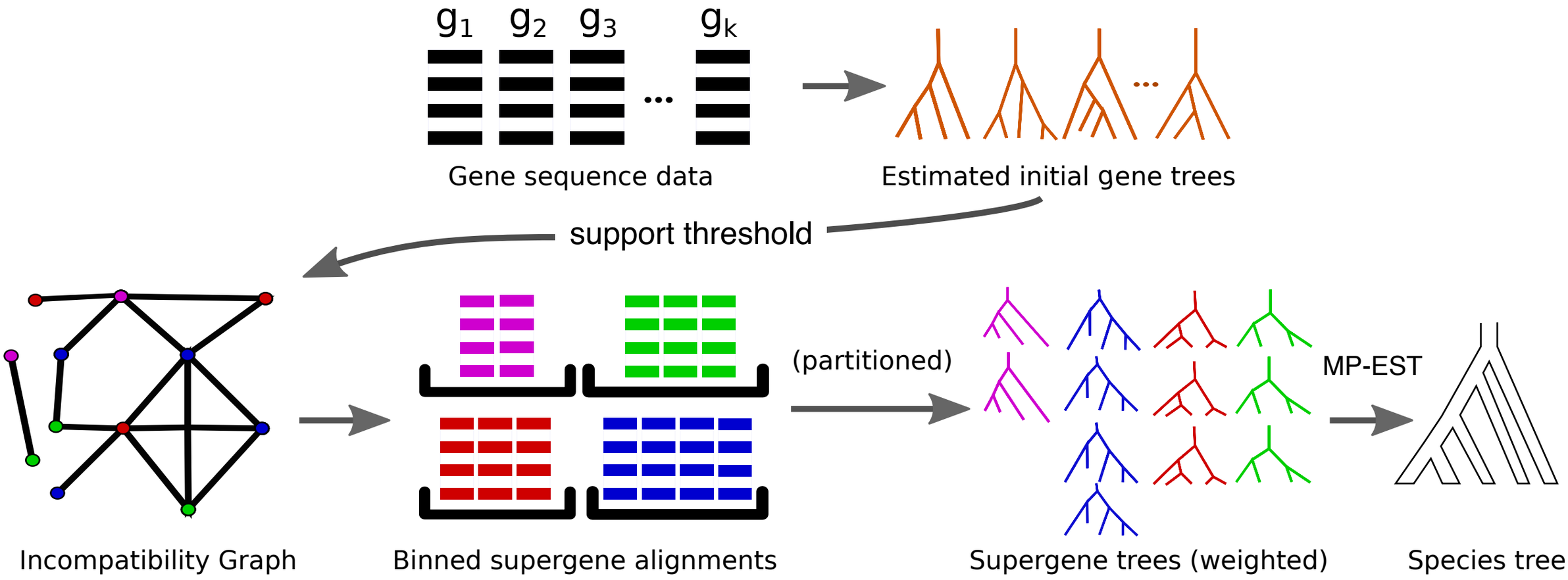


Statistical binning: overview



Original version: unweighted [Mirarab, et al., Science, 2014]

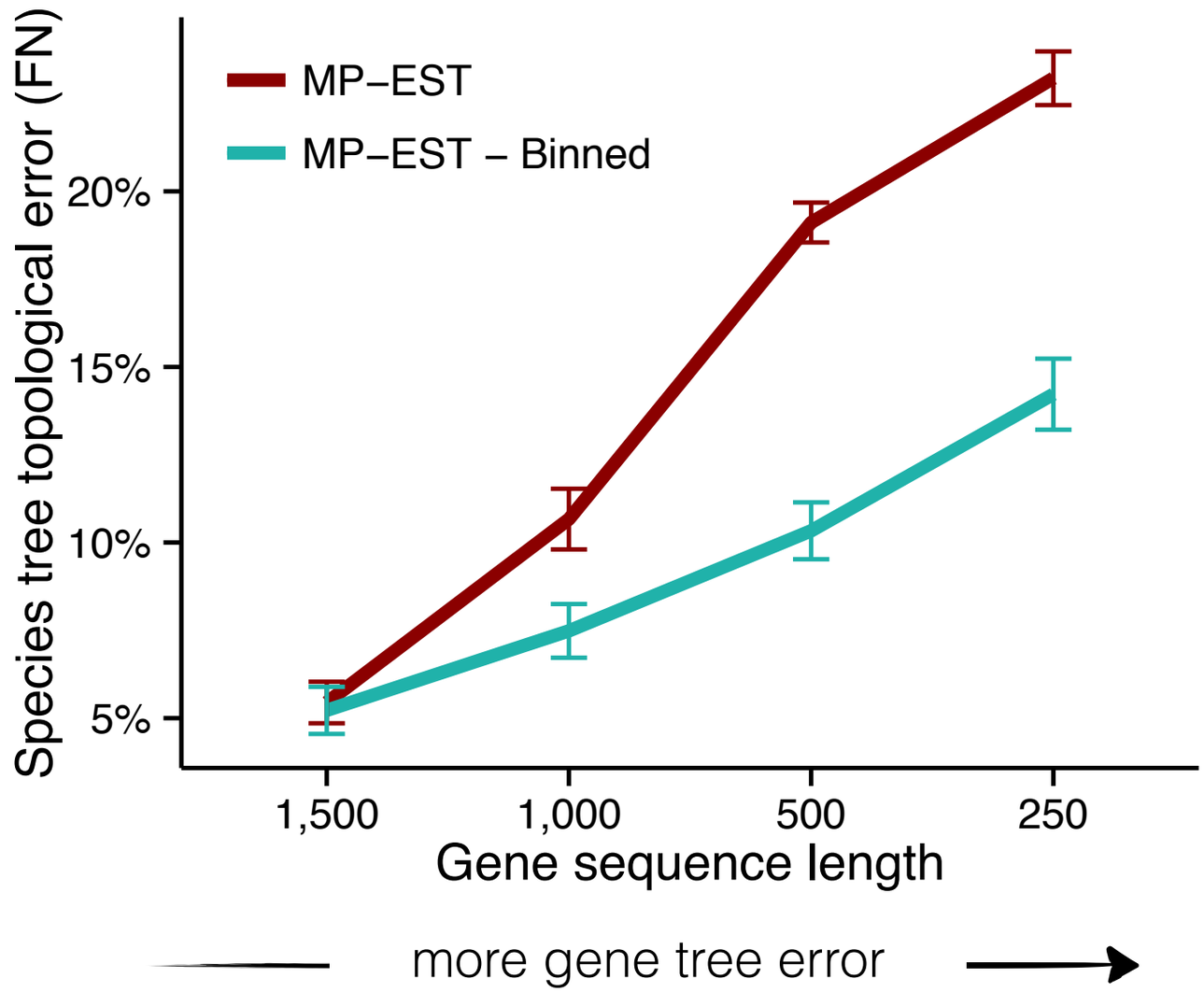
Statistical binning: overview



Original version: unweighted [Mirarab, et al., Science, 2014]

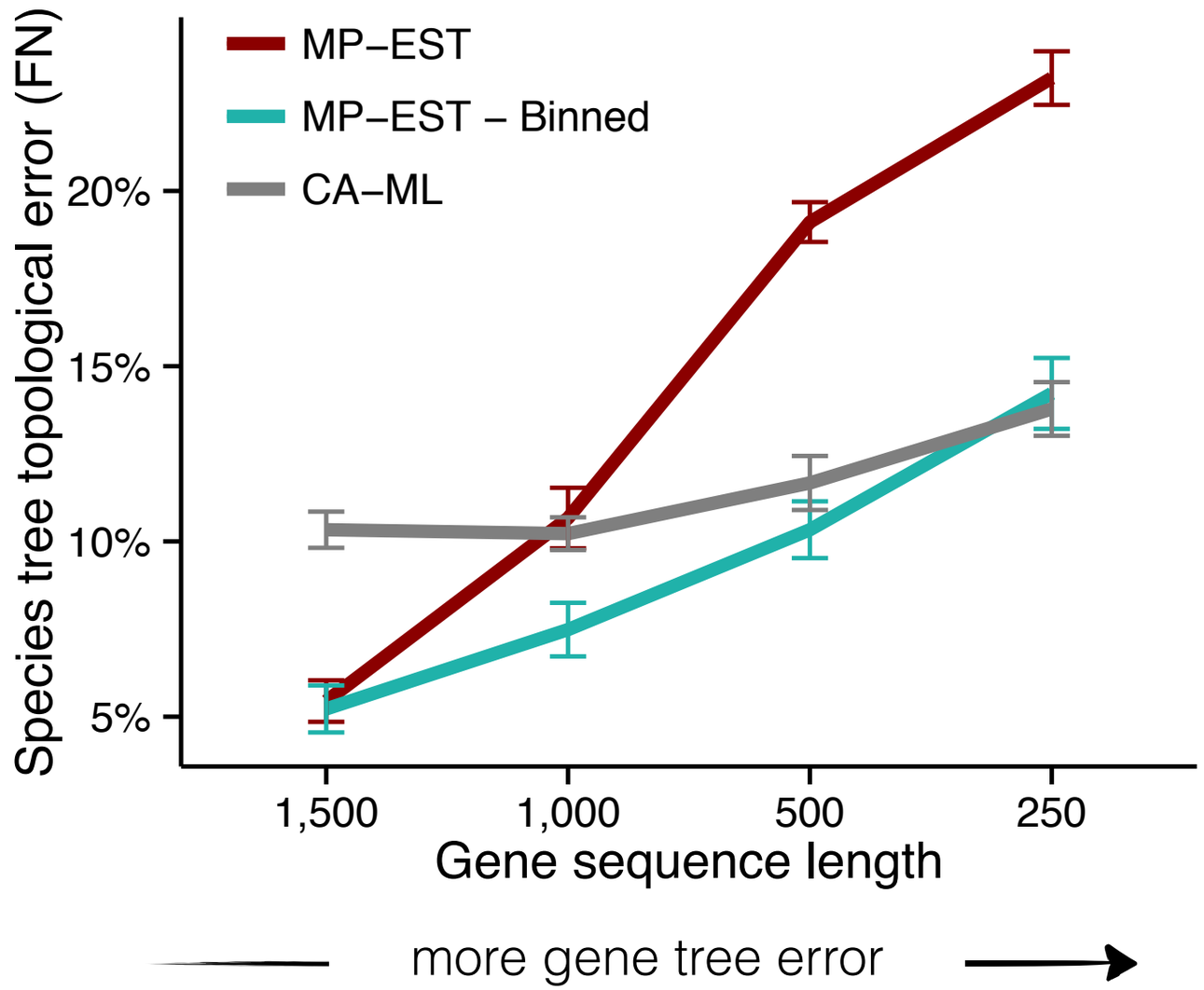
New version: weighted [Bayzid, Mirarab, Warnow, arXiv, 2015]

Avian-like simulation results



48 avian-like species, 1000 genes

Avian-like simulation results



48 avian-like species, 1000 genes

Binning also improves other measures of accuracy

- More accurate gene tree distributions

Binning also improves other measures of accuracy

- More accurate gene tree distributions
- Better species tree bootstrap support (i.e., fewer highly supported false positives)

Binning also improves other measures of accuracy

- More accurate gene tree distributions
- Better species tree bootstrap support (i.e., fewer highly supported false positives)
- More accurate species tree branch lengths

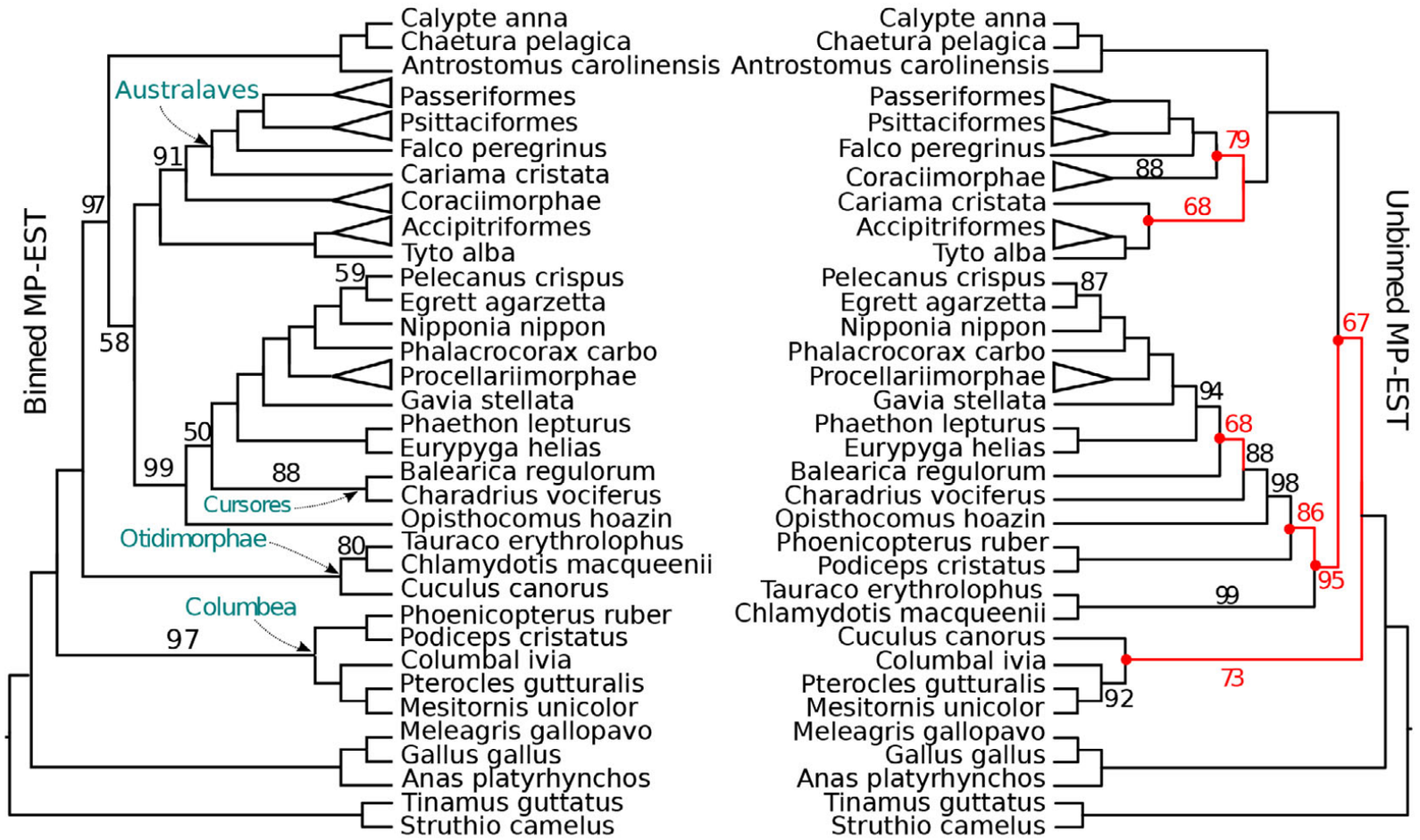
Binning on the avian dataset

RESEARCH ARTICLE

Whole-genome analyses resolve early branches in the tree of life of modern birds

Erich D. Jarvis,[1*†] Siavash Mirarab,[2*] Andre J. Aberer,[3] Bo Li,[4,5,6] Peter Houde,[7] Cai Li,[4,6] Simon Y. W. Ho,[8] Brant C. Faircloth,[9,10] Benoit Nabholz,[11] Jason T. Howard,[1] Alexander Suh,[12] Claudia C. Weber,[12] Rute R. da Fonseca,[6] Jianwen Li,[4] Fang Zha… Ganesh Ganapathy,[1] Ba… Volodymyr Zavidovych,… Salvador Capella-Gutié… Kasper Munch,[21] Mikk… David Ray,[23,24,25] Richa… Andrew Dixon,[29] Sheng… Elizabeth P. Derryberry… Robb T. Brumfield,[33] C… Maria Paula Cruz Schn… Amhed Missael Vargas… Bent Petersen,[39] Thoma… Michael Bunce,[43] Davi… Amy C. Driskell,[47] Beth… Zhenyu Li,[4] Binghang … Yong Zhang,[4] Huanmi… Michael Braun,[50] Jon … Knud Andreas Jønsson… Stephen O'Brien,[57,58] D… Eske Willerslev,[6] Gary … Dave Burt,[64] Hans Elle… Alexandros Stamatakis… Tandy Warnow,[2,72†] W…

[Jarvis, Mirarab, et al., Science, 2014]



The binned tree was highly supported and was largely congruent with concatenation

Summary

- Low phylogenetic signal per gene prevented accurate coalescent-based analyses of the avian dataset

- Statistical binning groups sets of genes based on statistical measures of combinability

- Statistical binning improves accuracy compared to both unbinned summary methods and concatenation

- Statistical binning enabled a coalescent-based analyses of the avian dataset; results were largely congruent with concatenation

More generally …

- Genome-scale data provides a wealth of information
- Yet, reconstruction of species phylogenies remains challenging
  - Limited data per gene
  - Scalability to many species: ASTRAL-II (ISMB 2015)
  - Impact of model violations, missing data, etc.
  - Multiple sources of gene tree discordance
- Many interesting statistical and computational questions and a need for method development

Acknowledgments

Tandy Warnow

Keshav Pingali

Jim Leebens-mack
(UGA)

Norman Wickett
(U Chicago)

Gane Wong
(U of Alberta)

. . .

. . .

Théo
Zimmermann

S.M. Bayzid

Nam Nguyen
(now at UIUC)

Guojie Zhang
(BGI, China)

Tom Gilbert
(U Copenhagen)

Erich Jarvis
(Duke, HMMI)

Bastien Boussau
(Université Lyon)

Ed Braun
(U Florida)

HMMI international student fellowship

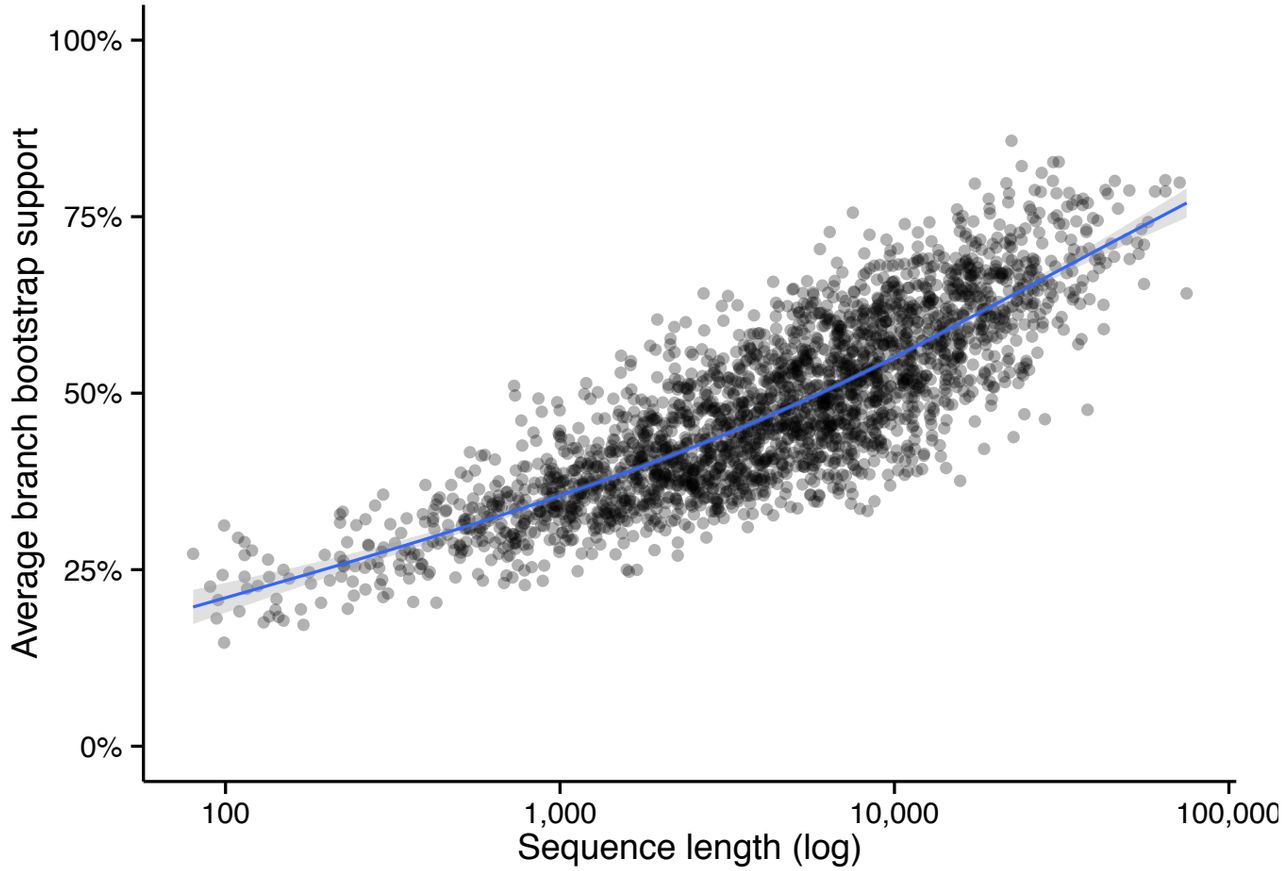
HHMI
HOWARD HUGHES MEDICAL INSTITUTE

NSERC
CRSNG

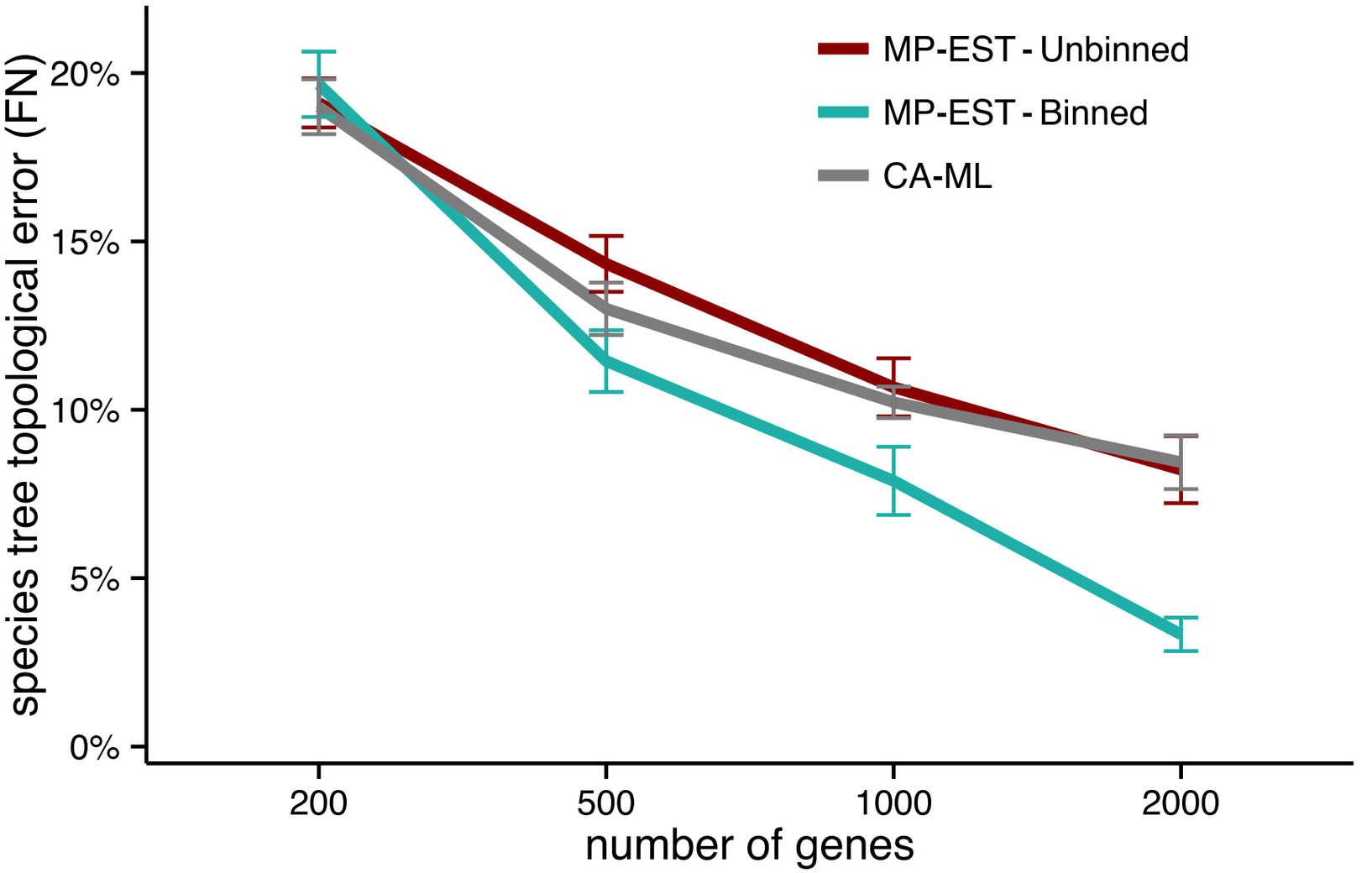
NSF

Lack of phylogenetic signal

1. Limited sequence length for each gene
2. Insufficient variation in each gene



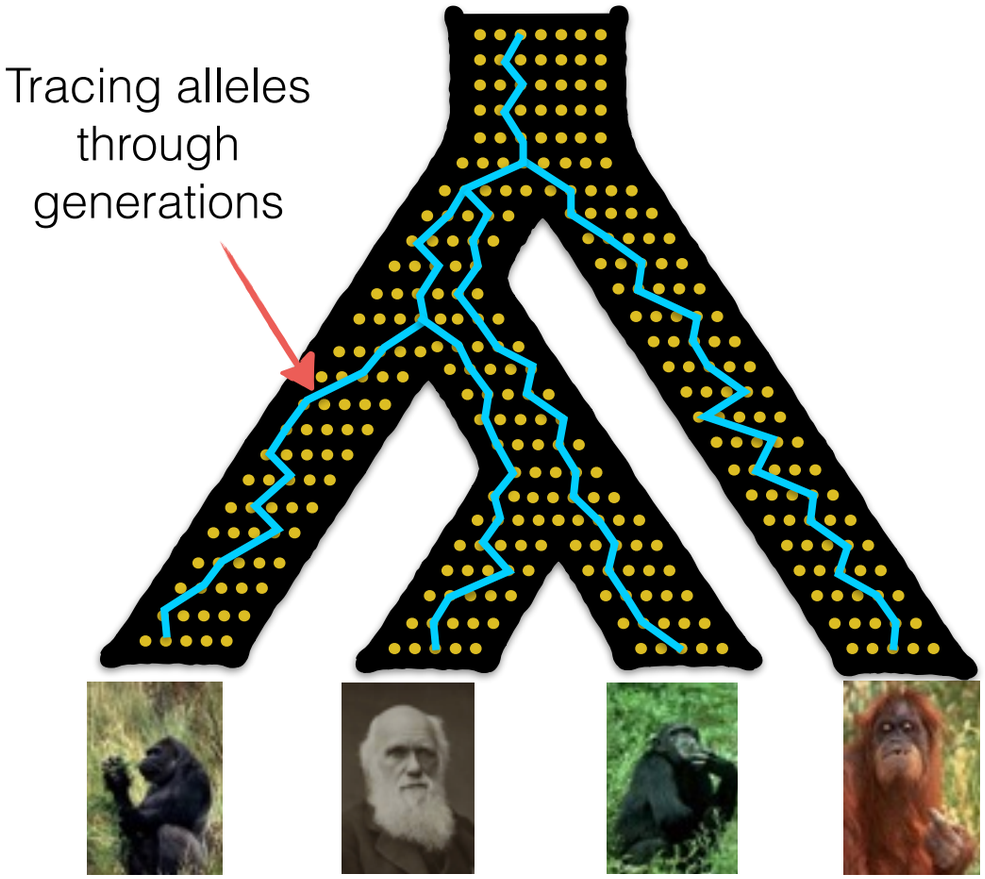
Increasing the number of genes



[Mirarab, et al., Science, 2014]

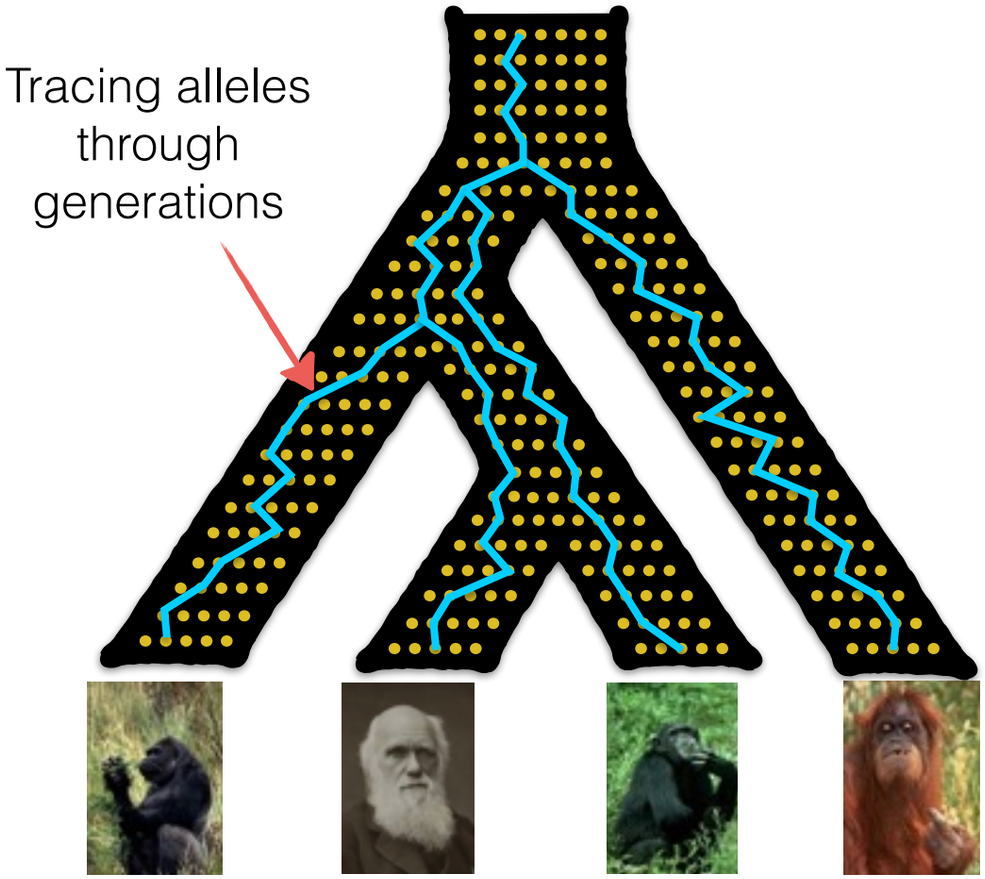
Incomplete Lineage Sorting (ILS)

- A population level process related to inheritance and maintenance of alleles
  - Omnipresent; most likely for short times between speciation events and/or large population size



Incomplete Lineage Sorting (ILS)

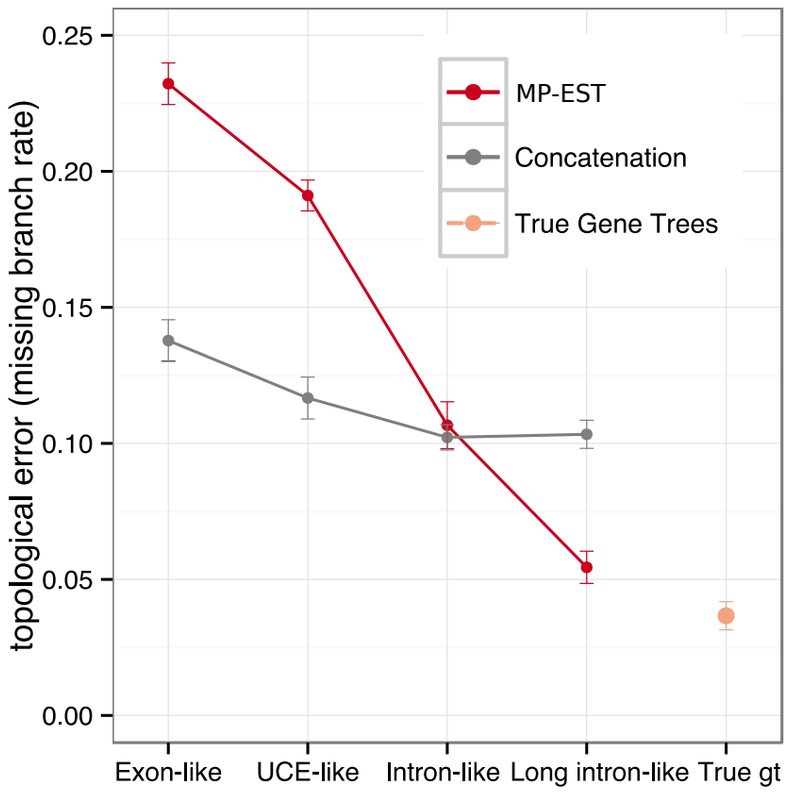
- A population level process related to inheritance and maintenance of alleles
  - Omnipresent; most likely for short times between speciation events and/or large population size
- We have statistical models of ILS (multi-species coalescent)
  - The species tree defines a probability distribution on the gene trees, and is identifiable from the distribution on gene trees
    [Degnan and Salter, Int. J. Org. Evolution, 2005]



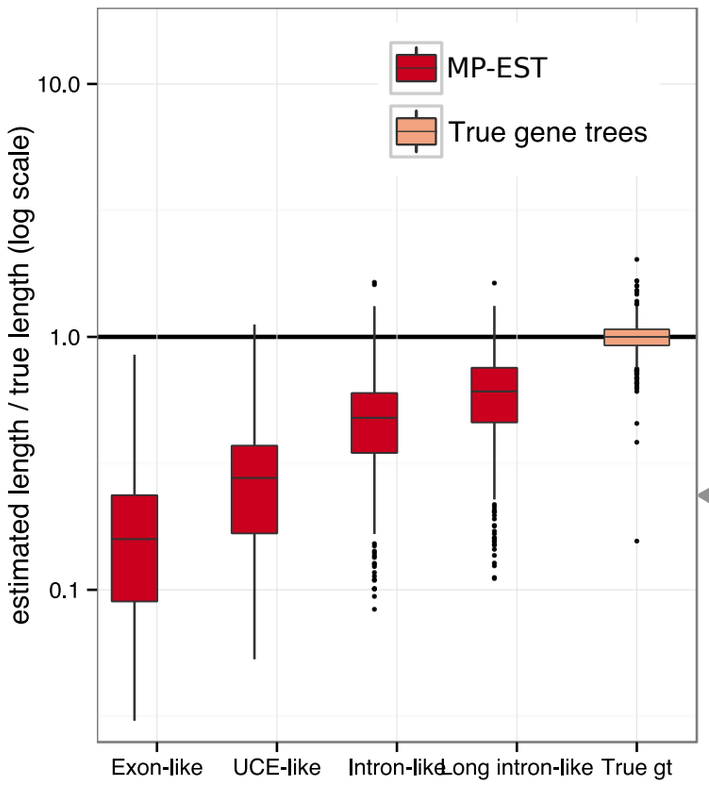
Avian-like simulation results

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

[Mirarab, et al., Science, 2014]



More information per gene

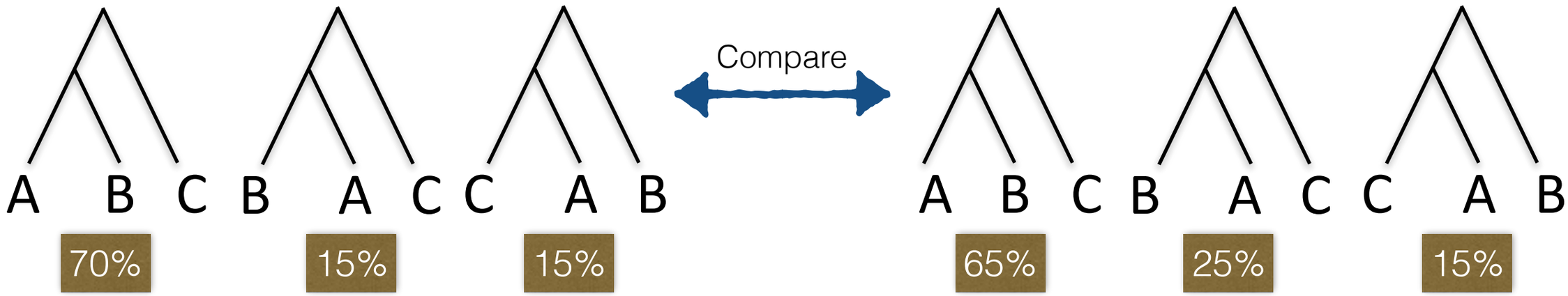


More information per gene

Branch length accuracy

Gene tree distribution error

- We can quantify gene tree distribution error using triplet frequency:



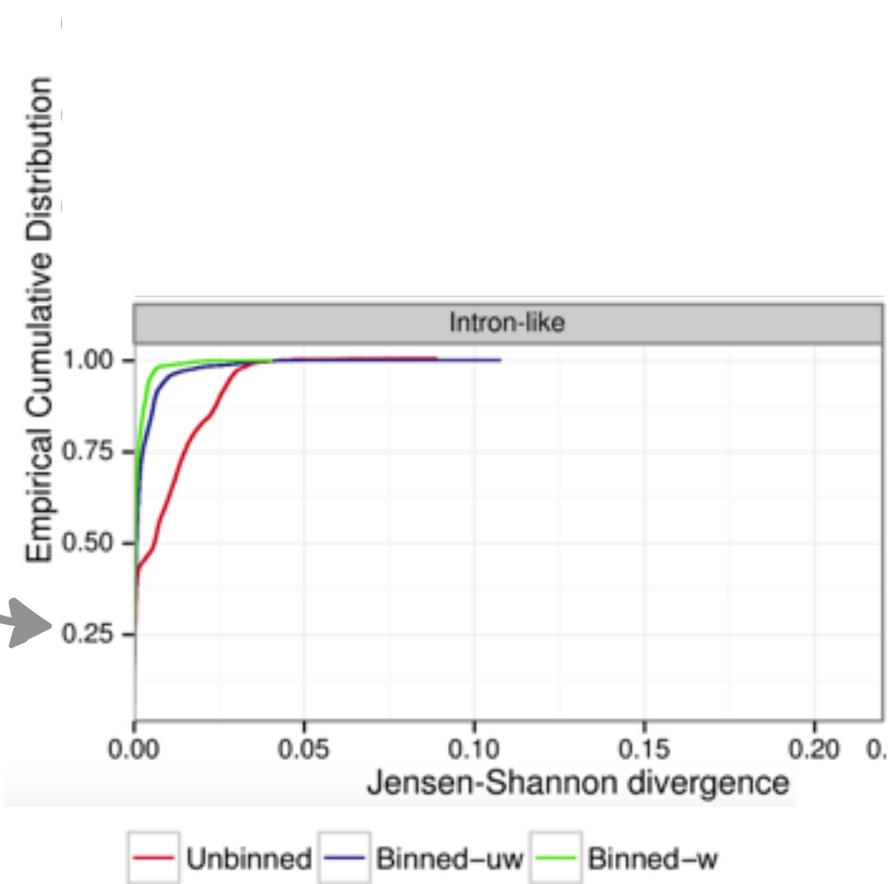
true distribution

estimated distribution

- We can compare triplet frequencies obtained from true gene trees and from the estimated gene trees (for all triplets of taxa)

Binning improves gene tree distribution

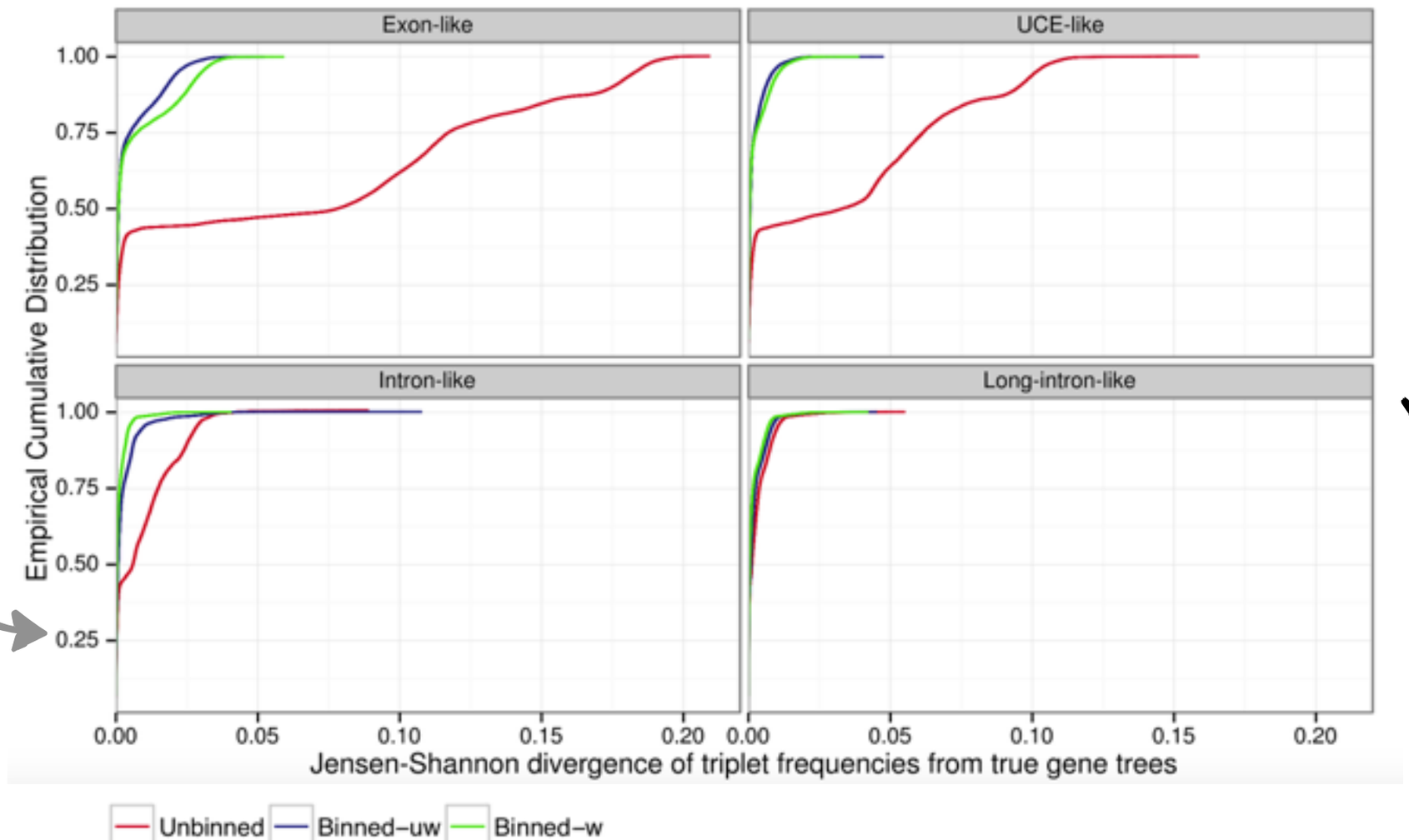
Empirical commutative distribution



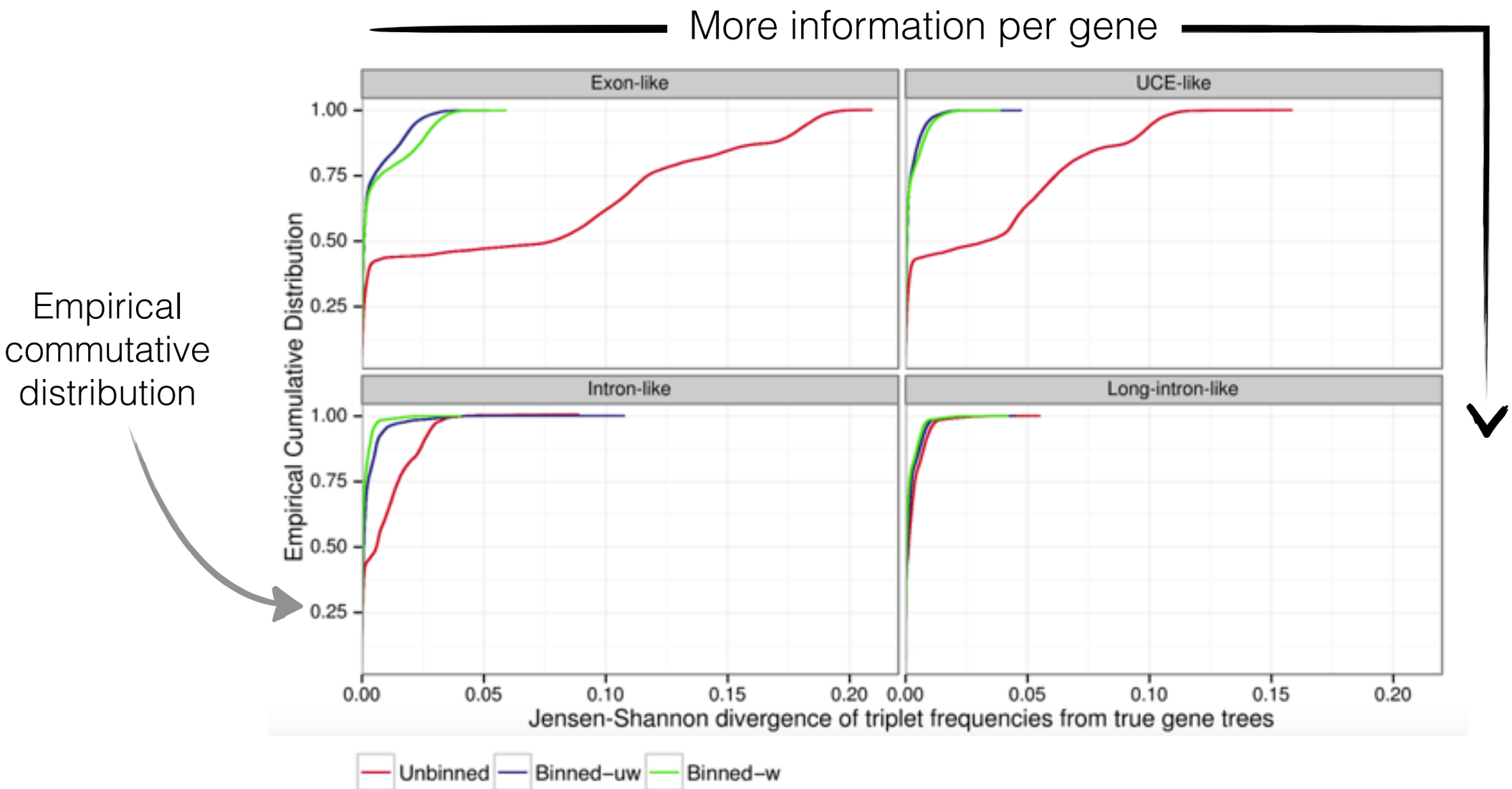
Binning improves gene tree distribution

Empirical commutative distribution

More information per gene



Binning improves gene tree distribution



Supergene trees represent the true gene tree distribution much better than the estimated gene trees without binning.