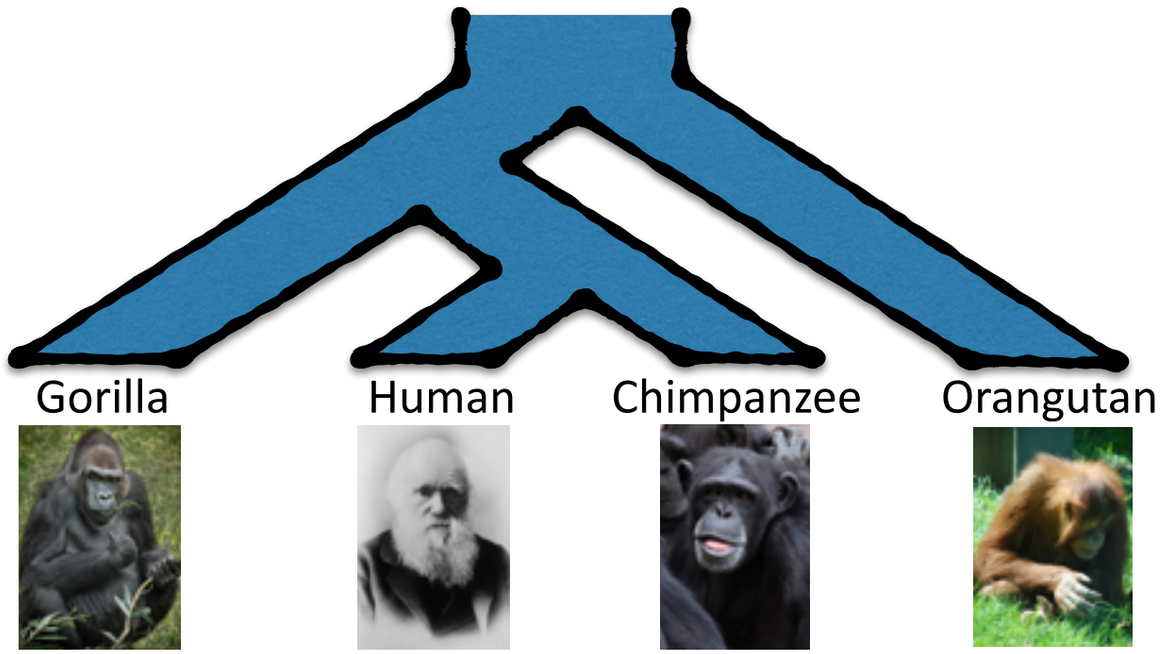
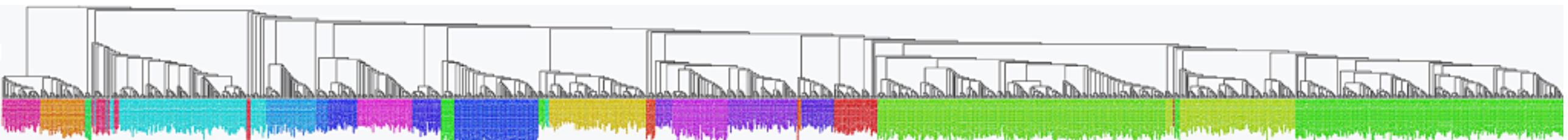
Reconstruction of species histories using genomic data

Siavash Mirarab
University of California San Diego





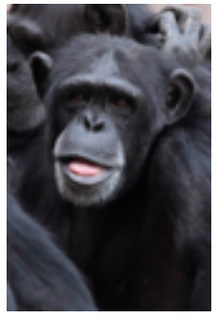
Statistical inference of phylogenies

Gorilla

ACTGCACACCG

Human

ACTGCCCCCG



Chimpanzee

AATGCCCCCG

Orangutan

CTGCACACGG

D

Gorilla ACTGCACACCG

Human ACTGC-CCCCG

Chimpanzee AATGC-CCCCG

Orangutan -CTGCACACGG

$P(D|T)$

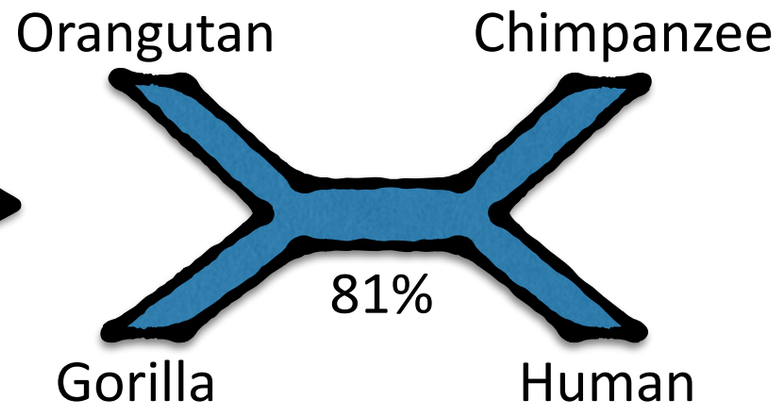
Maximum likelihood

Bayesian

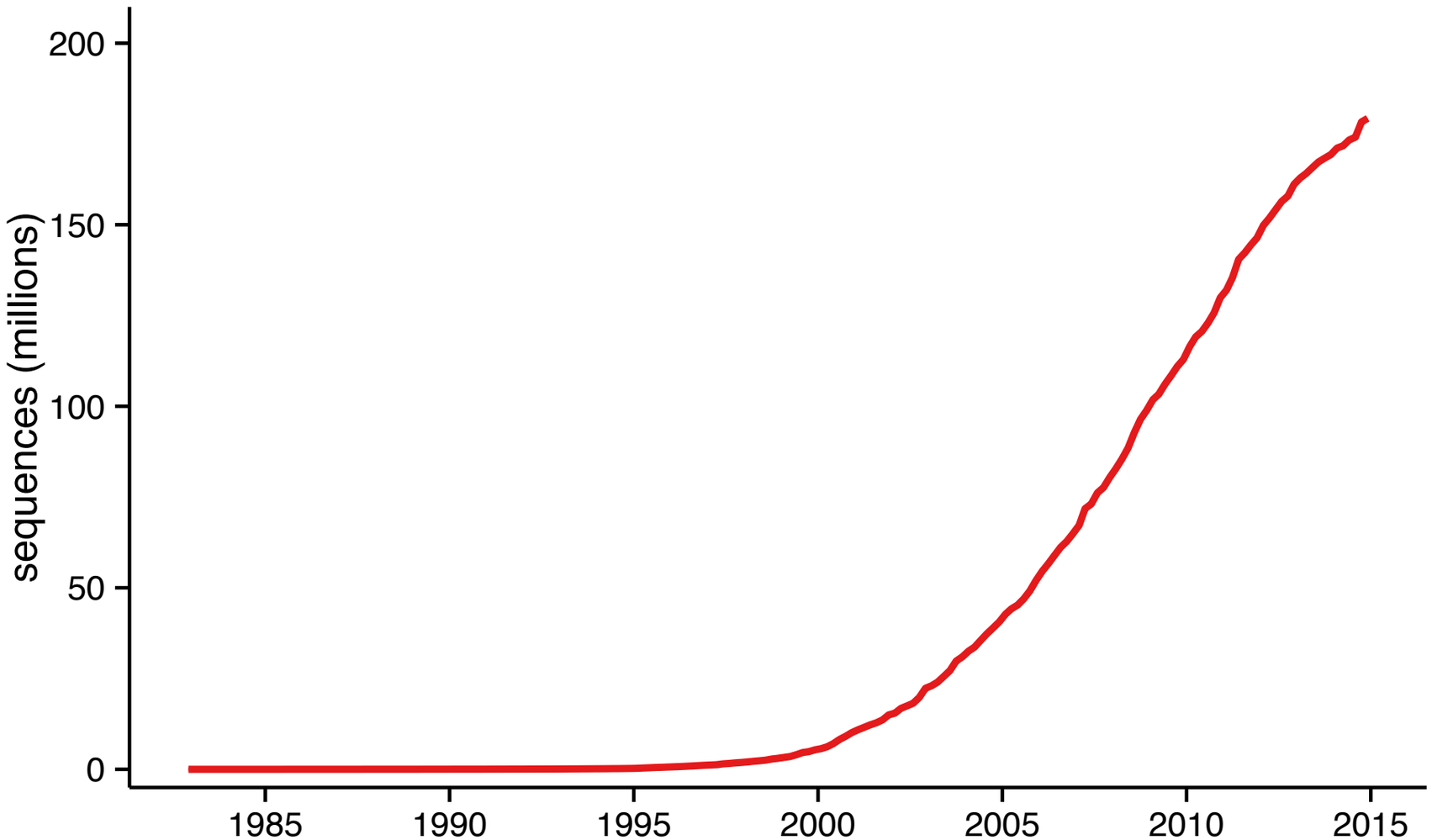
Distance-based

Parsimony

T

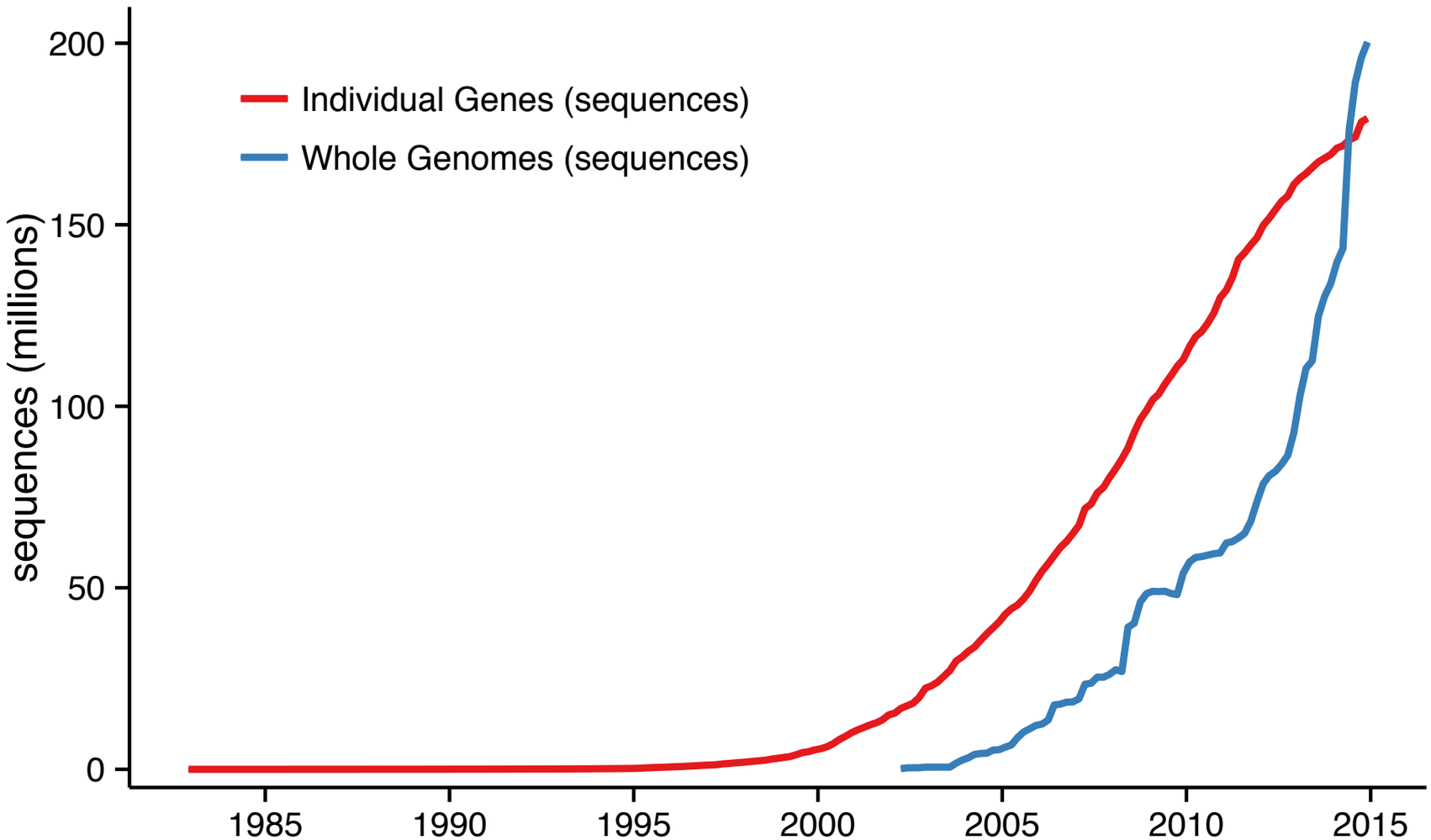


Sequence data growth



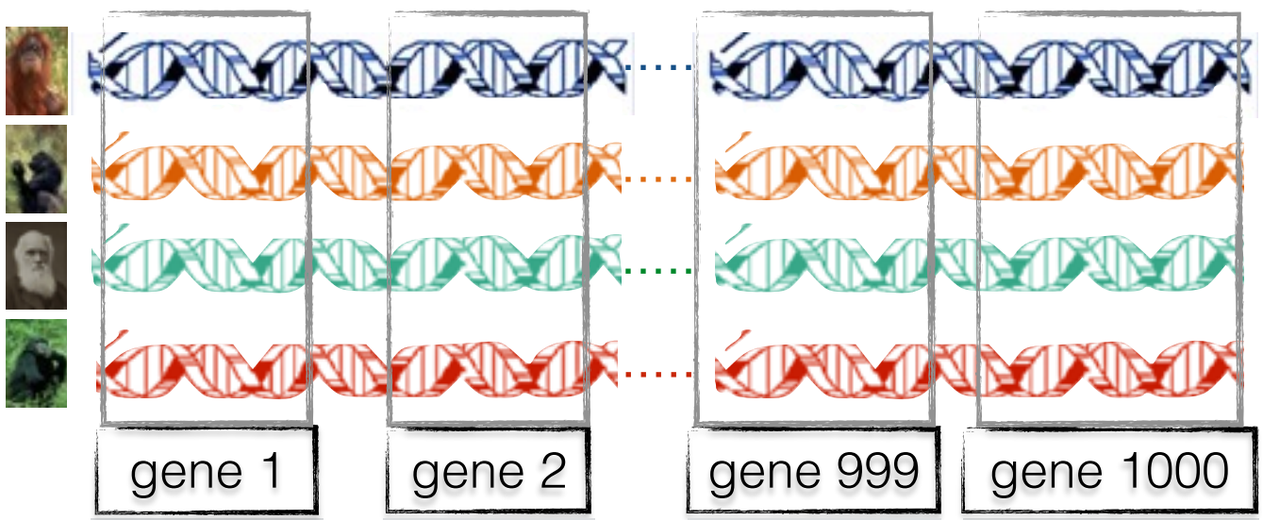
Growth in Genebank sequence data
data source: NCBI (http://www.ncbi.nlm.nih.gov/genbank/statistics)

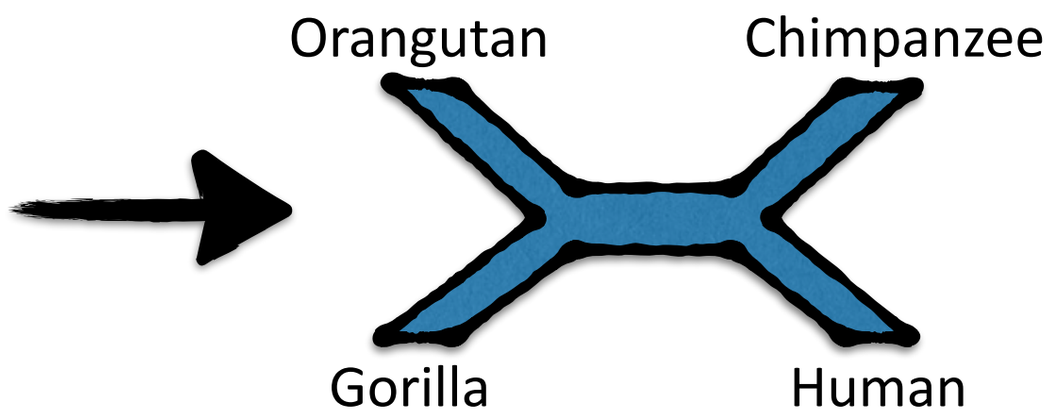
Sequence data growth



Growth in Genebank sequence data
data source: NCBI (http://www.ncbi.nlm.nih.gov/genbank/statistics)

Phylogenomics

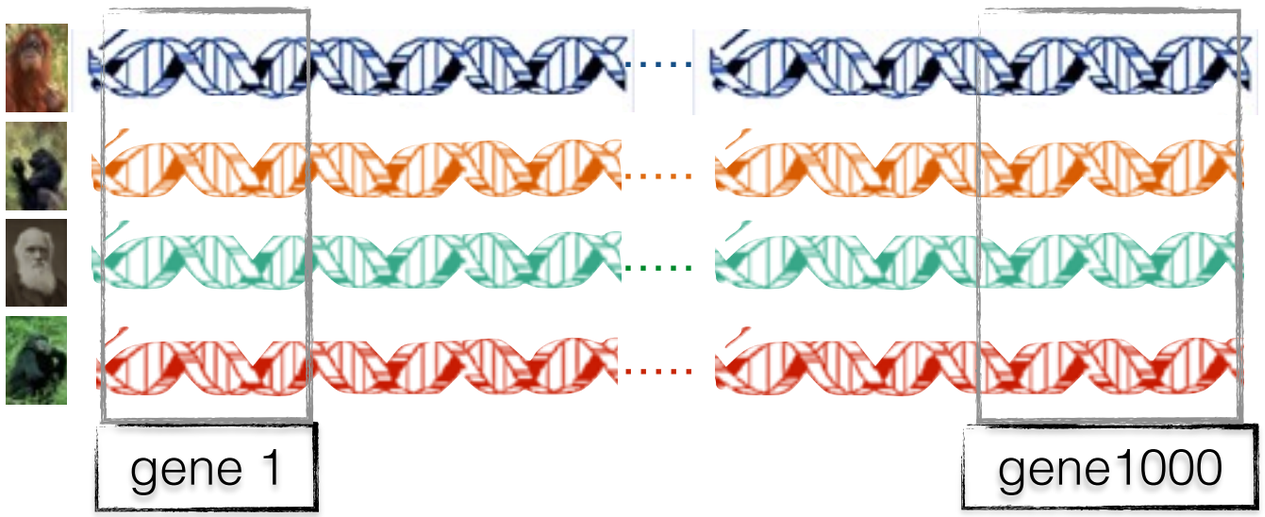




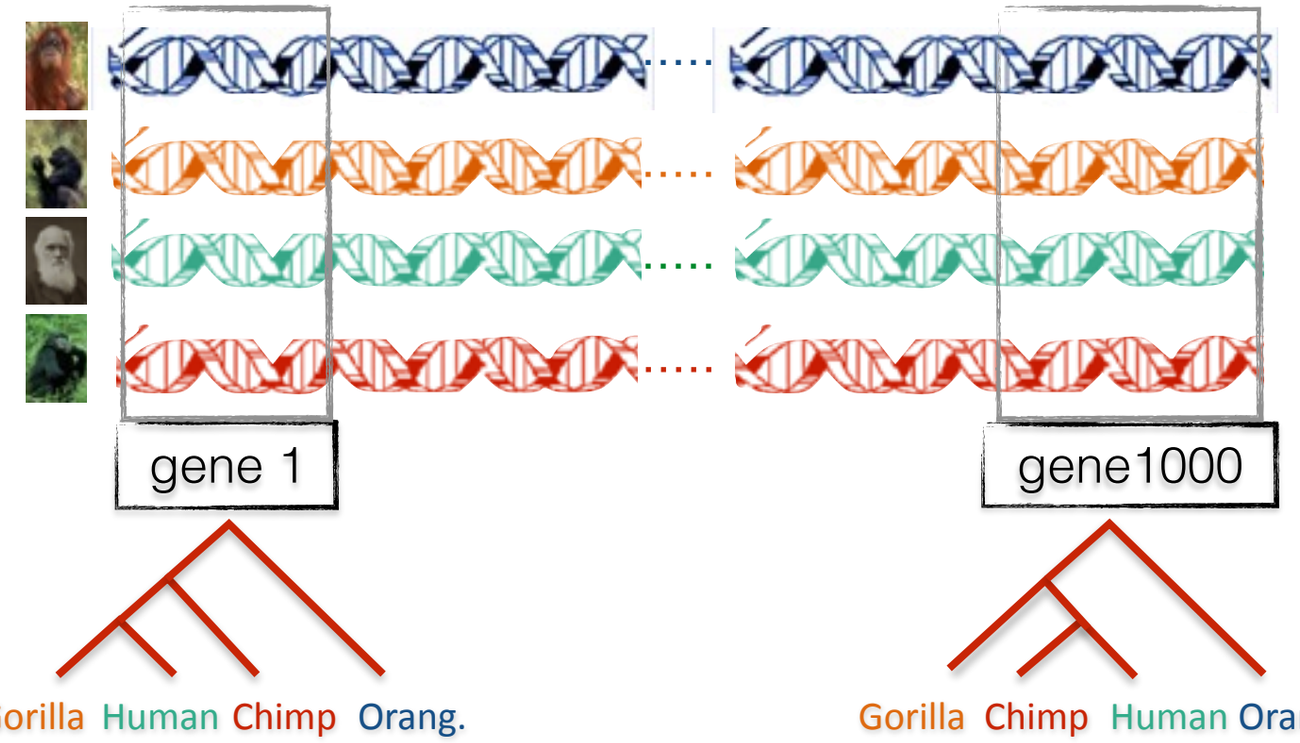

```
ACTGCACACCG   CTGAGCATCG     AGCAGCATCGTG CAGGCACGCACGAA
ACTGC-CCCCG   CTGAGC-TCG ... AGCAGC-TCGTG AGC-CACGC-CATA
AATGC-CCCCG   ATGAGC-TC-     AGCAGC-TC-TG ATGGCACGC-C-TA
-CTGCACACGG   CTGA-CAC-G     C-TA-CACGGTG AGCTAC-CACGGAT
```

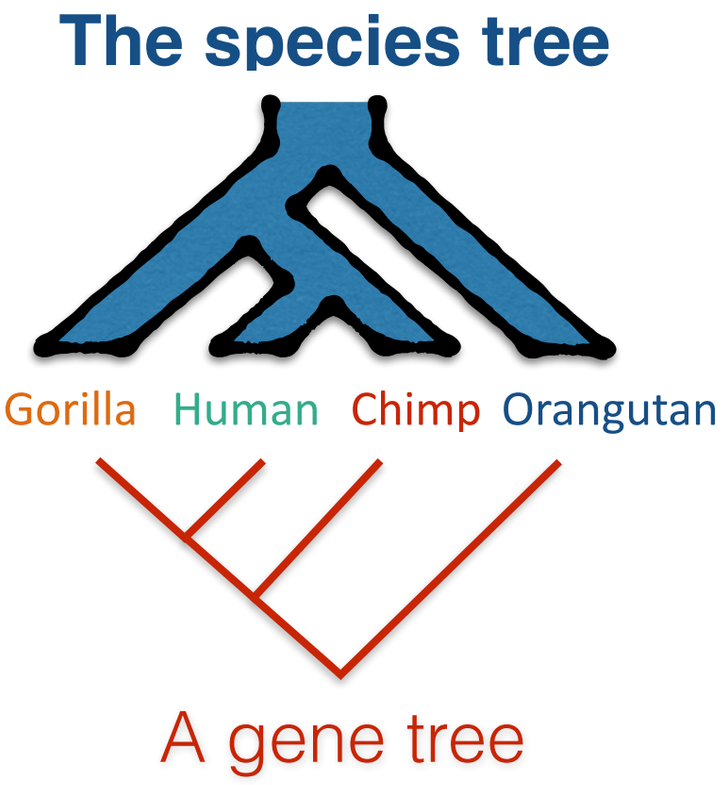
I'll use the term "gene" to refer to "c-genes":
recombination-free orthologous stretches of the genome

Gene tree discordance

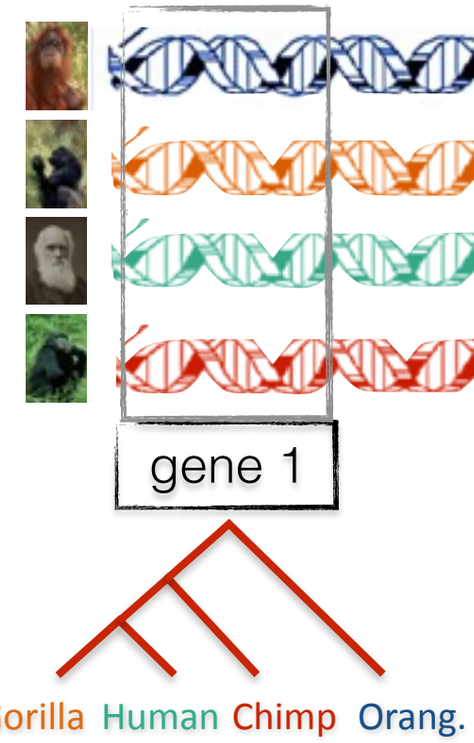


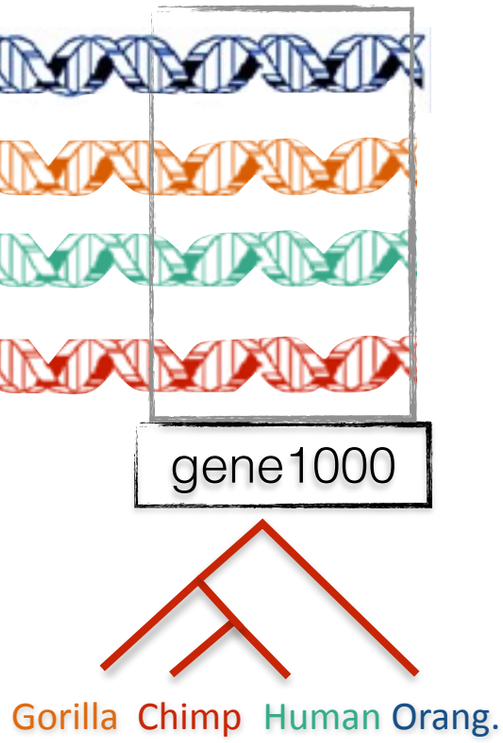
Gene tree discordance

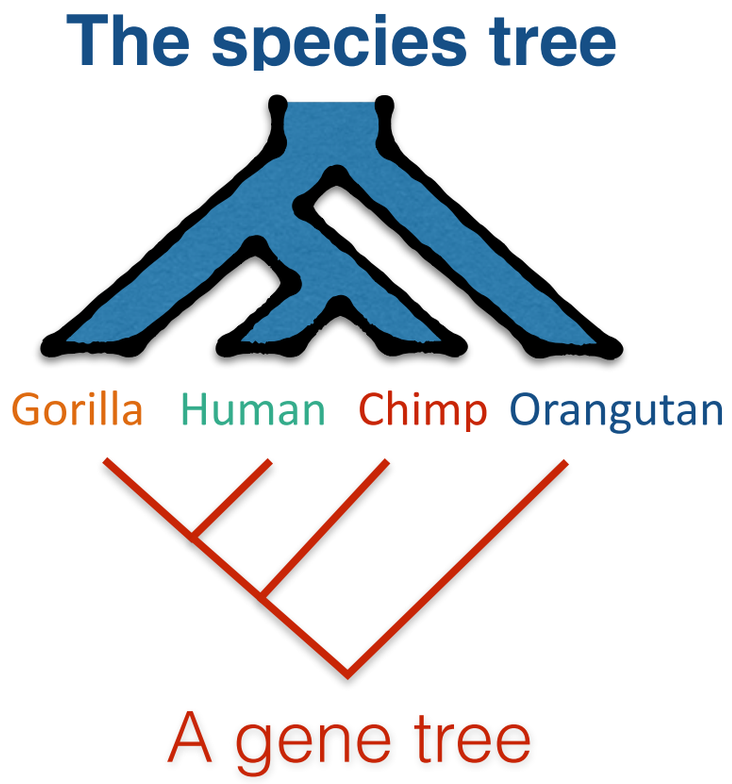




Gene tree discordance





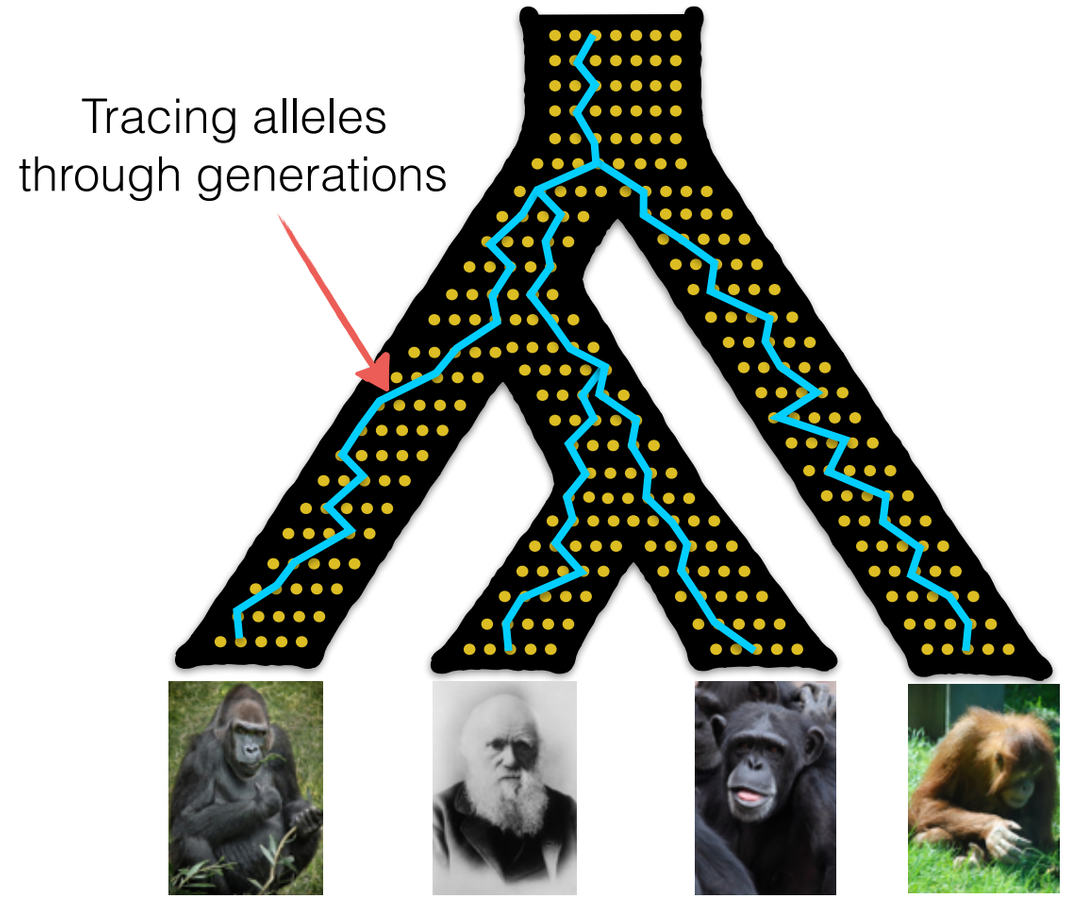


Causes of gene tree discordance include:

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

Incomplete Lineage Sorting (ILS)

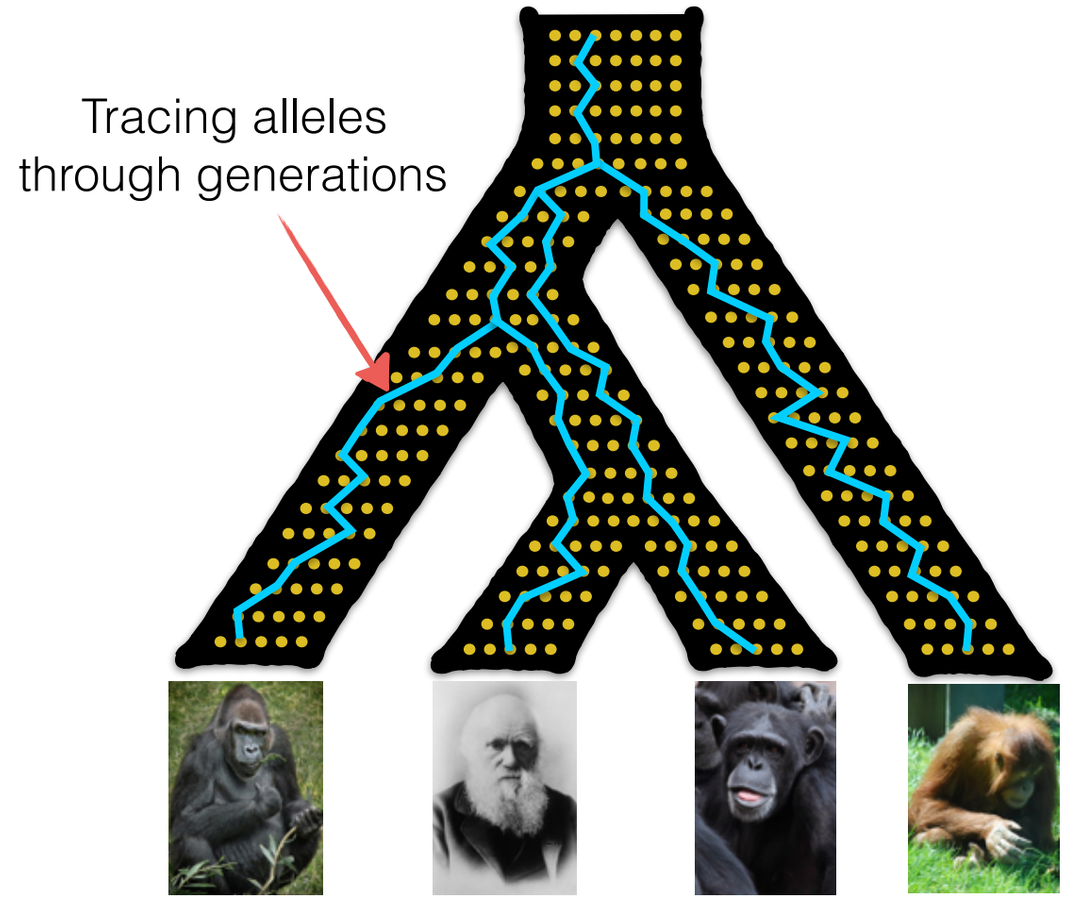
- A random process related to having multiple versions of each gene in a population



Incomplete Lineage Sorting (ILS)

- A random process related to having multiple versions of each gene in a population

- Omnipresent; most likely for short branches and large population size



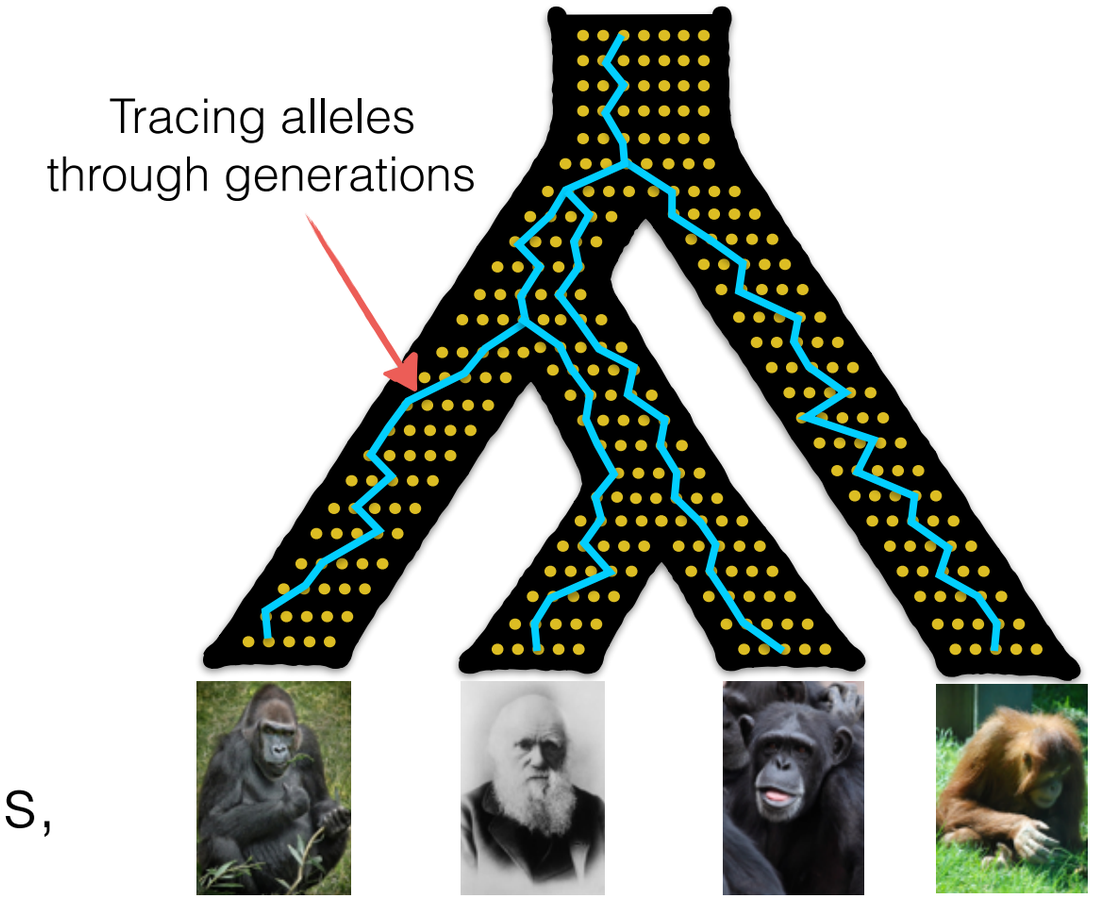
Incomplete Lineage Sorting (ILS)

- A random process related to having multiple versions of each gene in a population

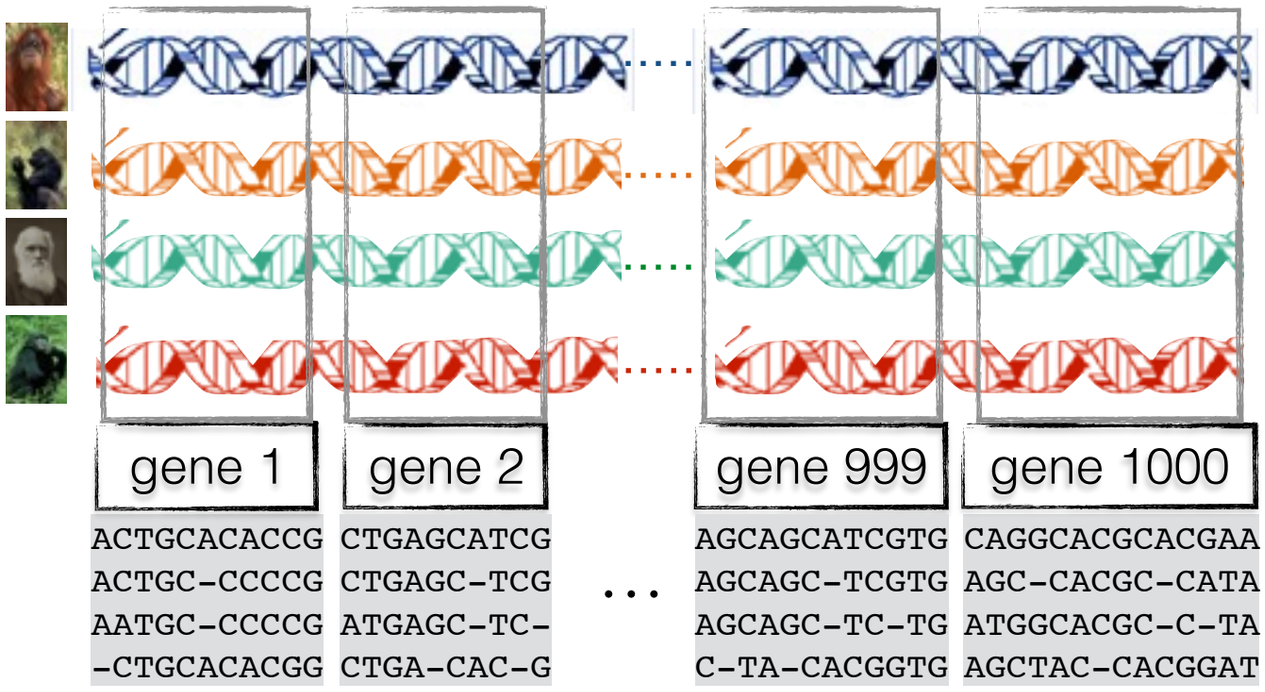
- Omnipresent; most likely for short branches and large population size

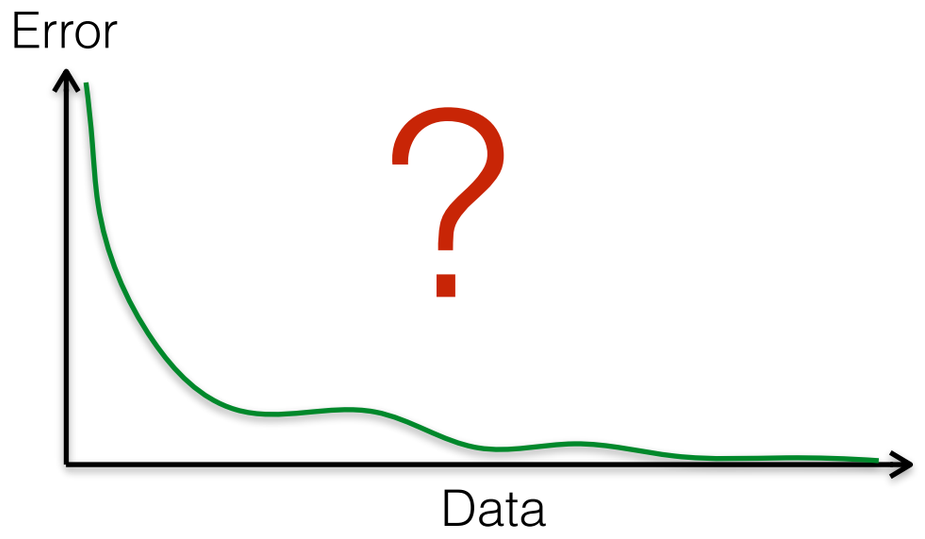
- We have statistical models of ILS (multi-species coalescent)

  - The species tree defines the probability distribution on gene trees, and is identifiable from the distribution on gene trees
    [Degnan and Salter, Int. J. Org. Evolution, 2005]



Increased data —> accuracy?

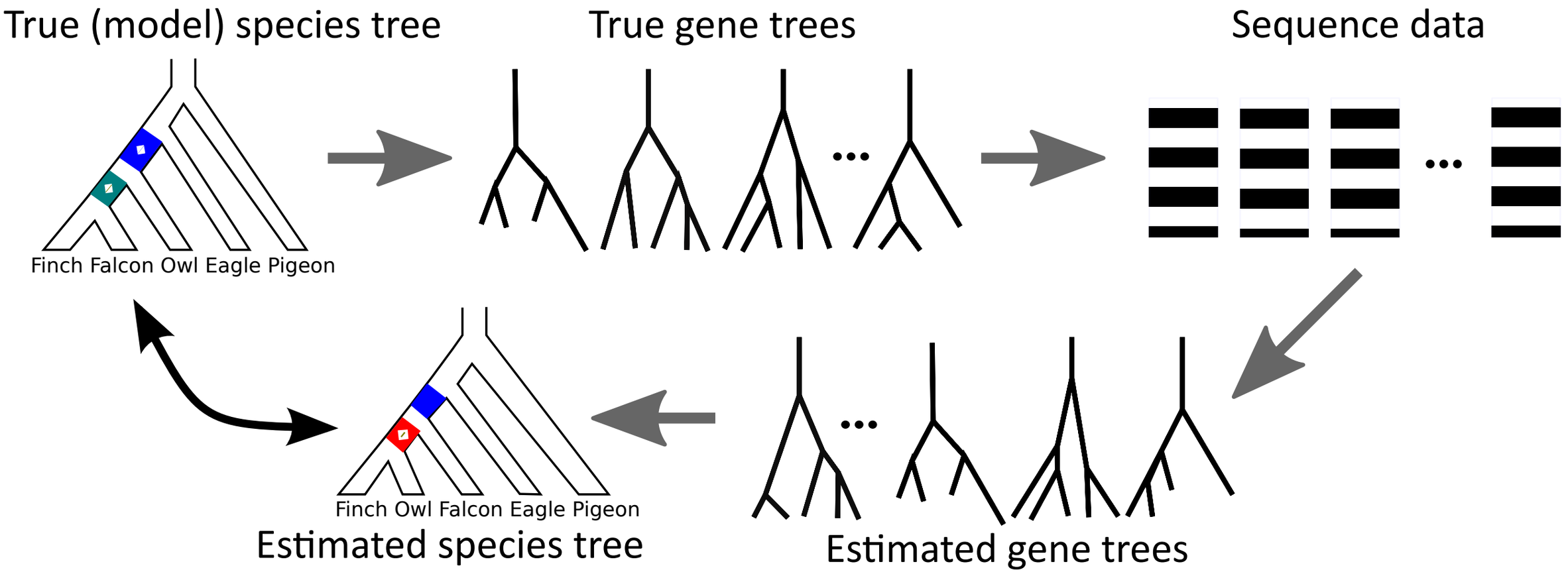




A. **Theoretical statistical guarantees:** assuming data are generated under the multi-species coalescent model, we aspire to statistical consistency

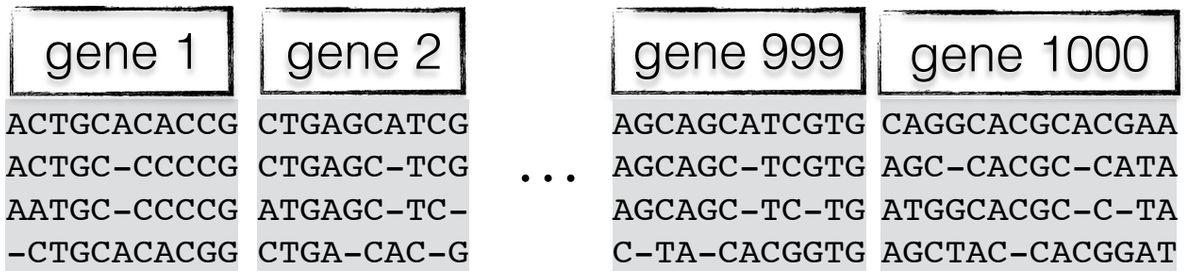
B. **Simulation studies**, generating synthetic data according to the multi-species coalescent model and models of sequence evolution

Simulation studies

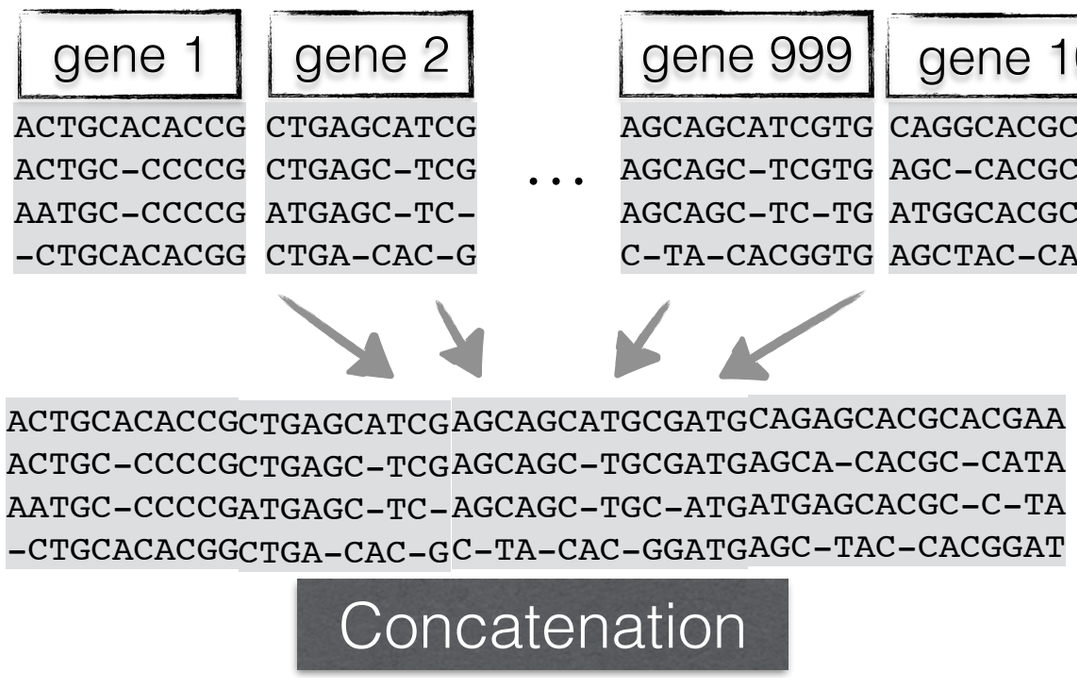


Error metric: percentage of branches in true tree that are missing from the estimated tree

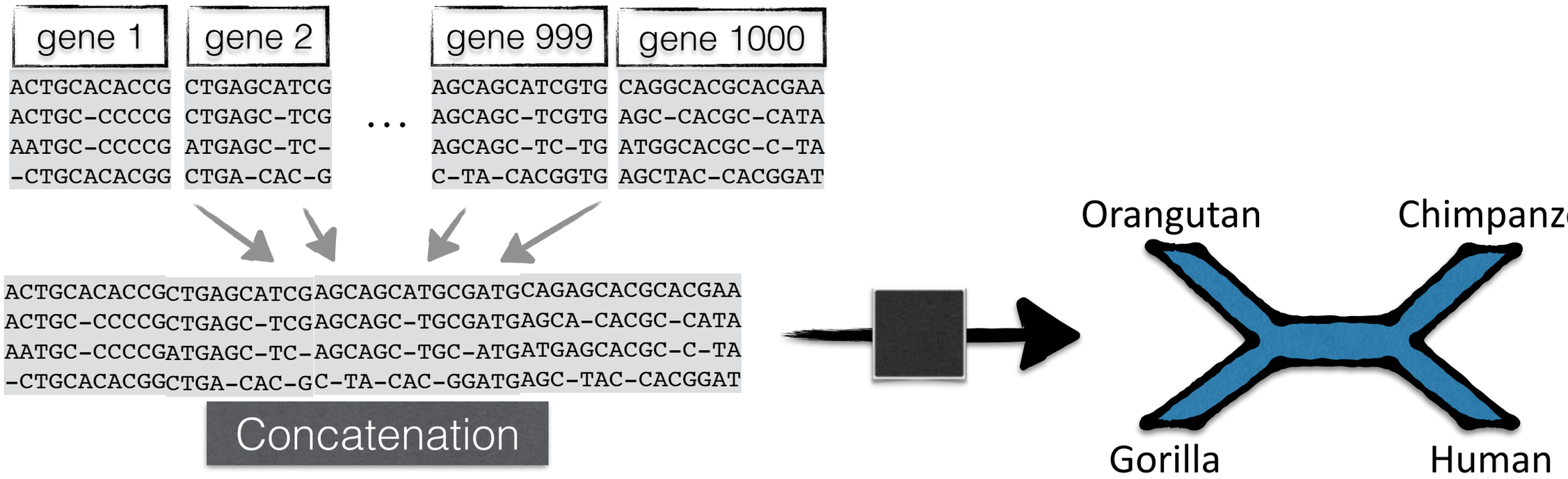
Species tree estimation from phylogenomic data (Approach 1: concatenation)



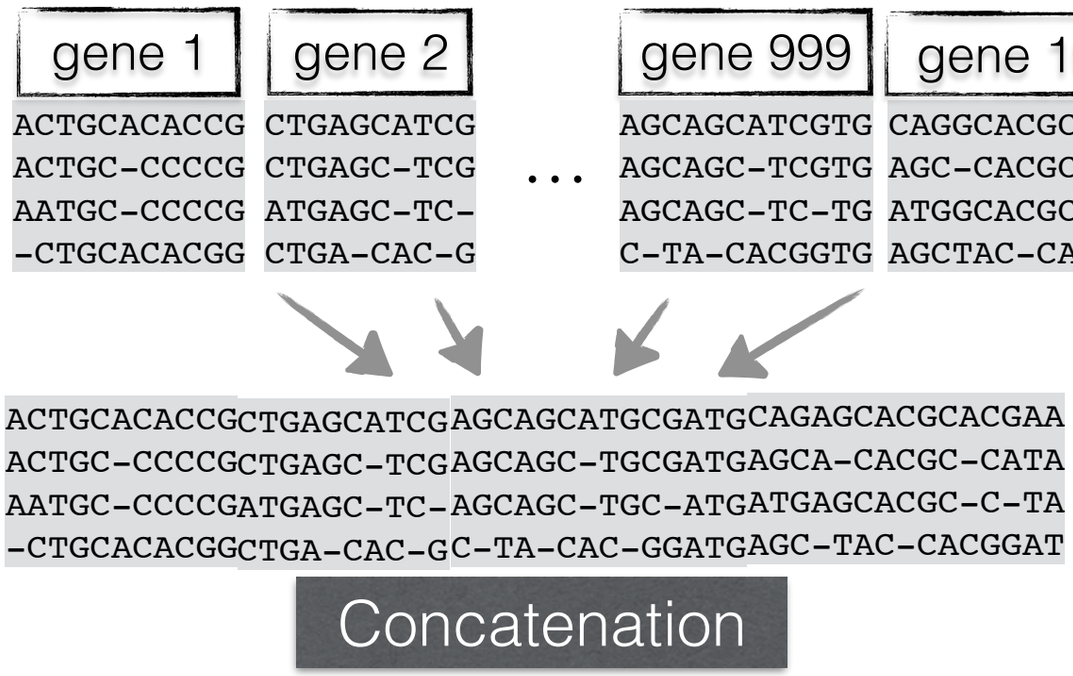
Species tree estimation from phylogenomic data (Approach 1: concatenation)



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



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

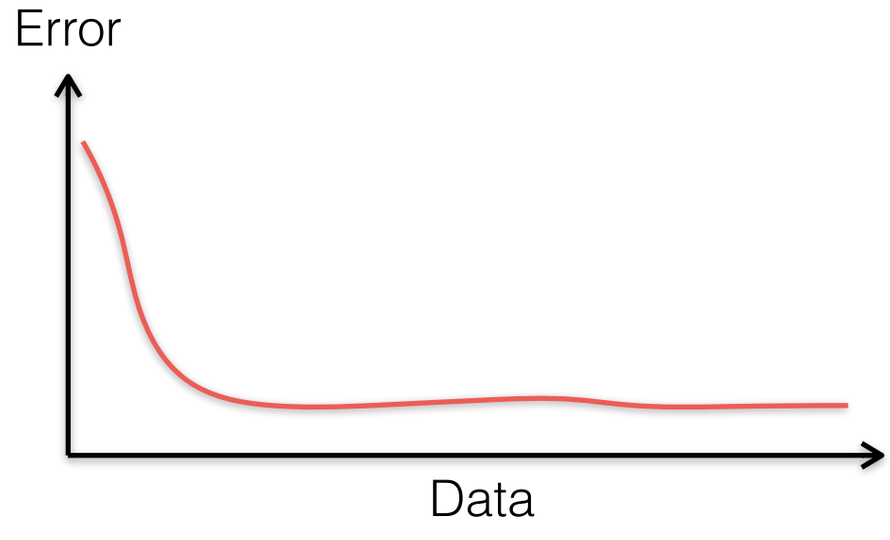


Statistically <u>inconsistent</u> & positively misleading
(proof for unpartitioned maximum likelihood)

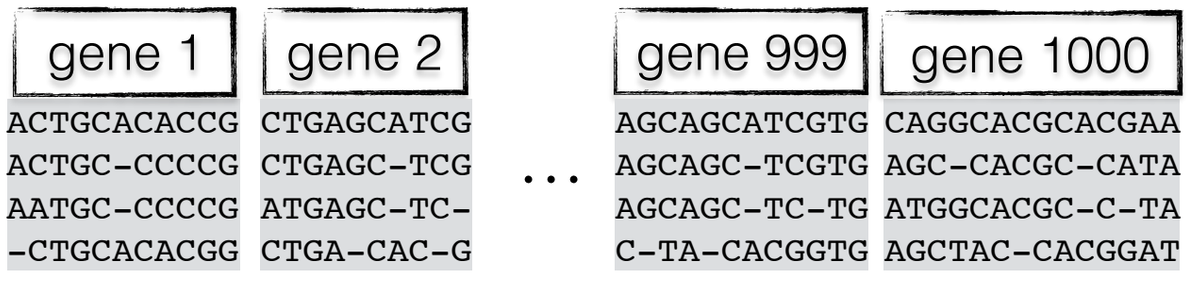
[Roch and Steel, Theo. Pop. Gen., 2014]

Mixed accuracy in simulations

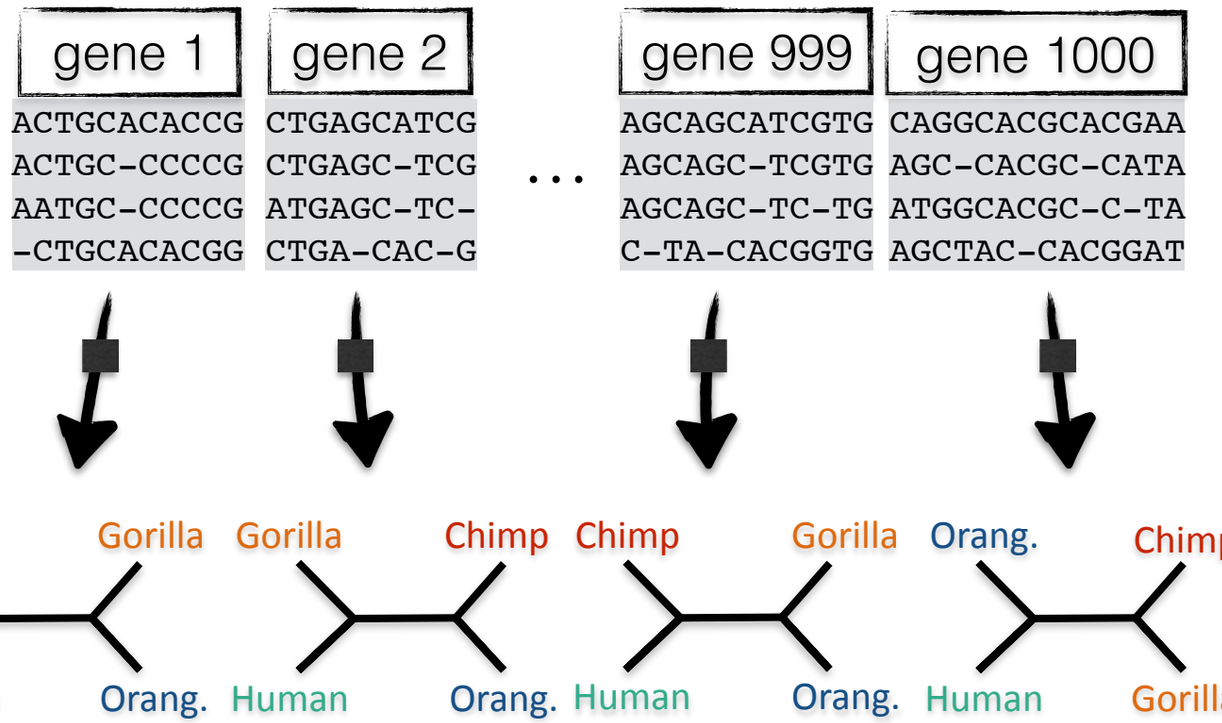
[Kubatko and Degnan, Systematic Biology, 2007]
[Mirarab, et al., Systematic Biology, 2014]



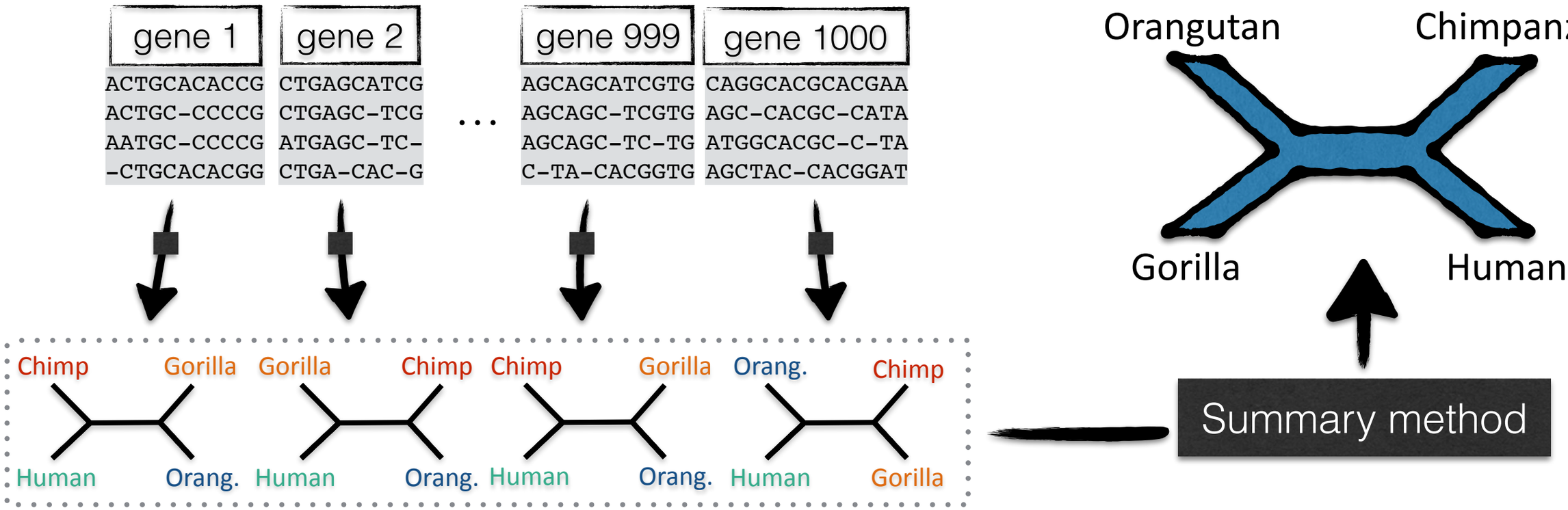
Species tree estimation from phylogenomic data (Approach 2: summary methods)



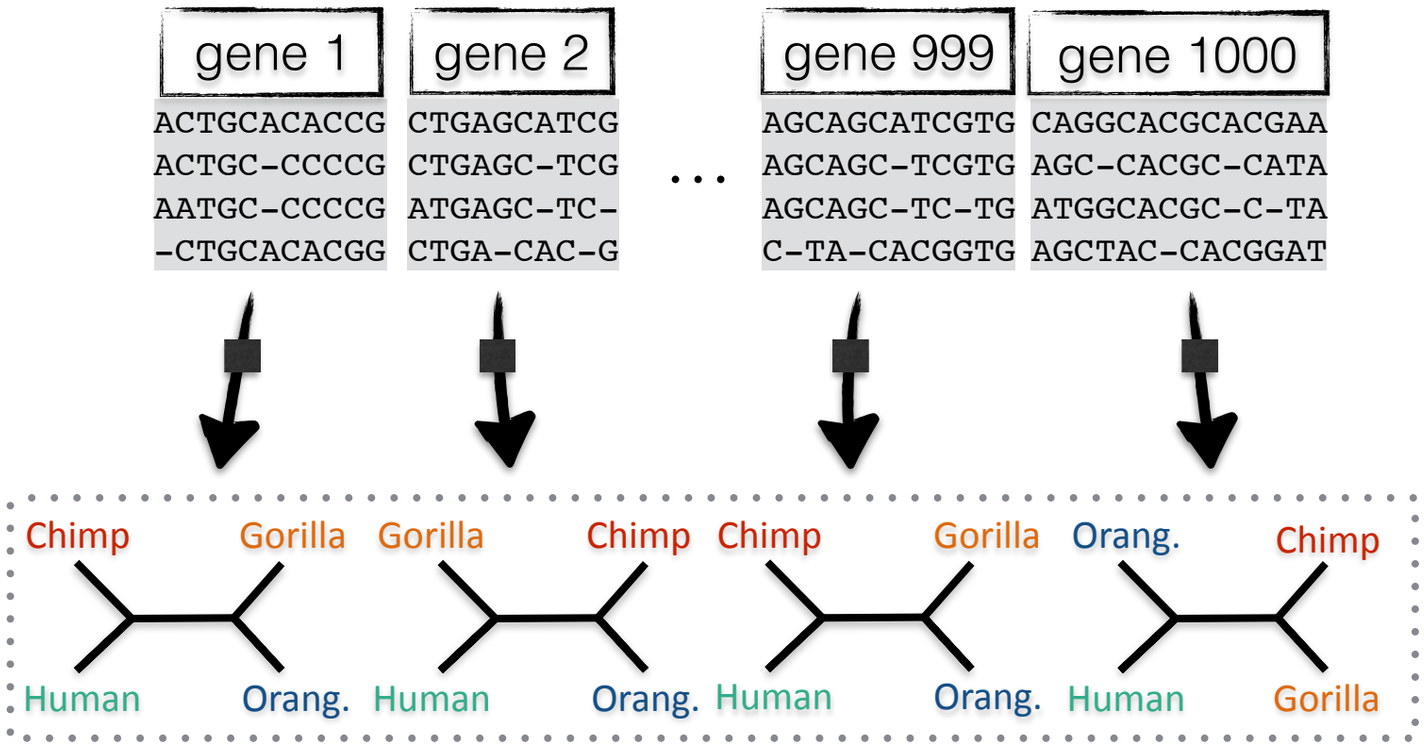
Species tree estimation from phylogenomic data (Approach 2: summary methods)

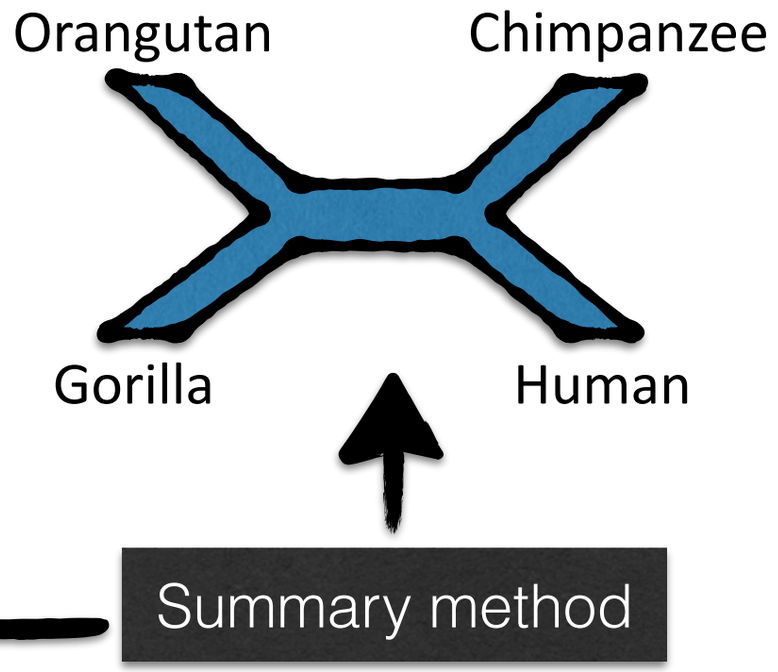


Species tree estimation from phylogenomic data (Approach 2: summary methods)



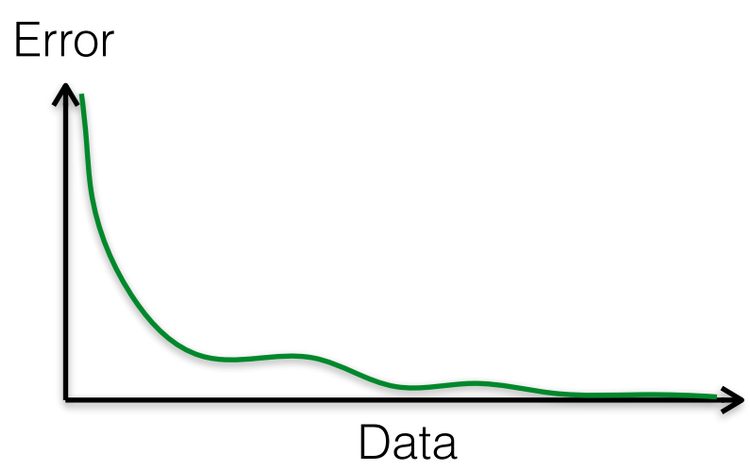
Species tree estimation from phylogenomic data (Approach 2: summary methods)



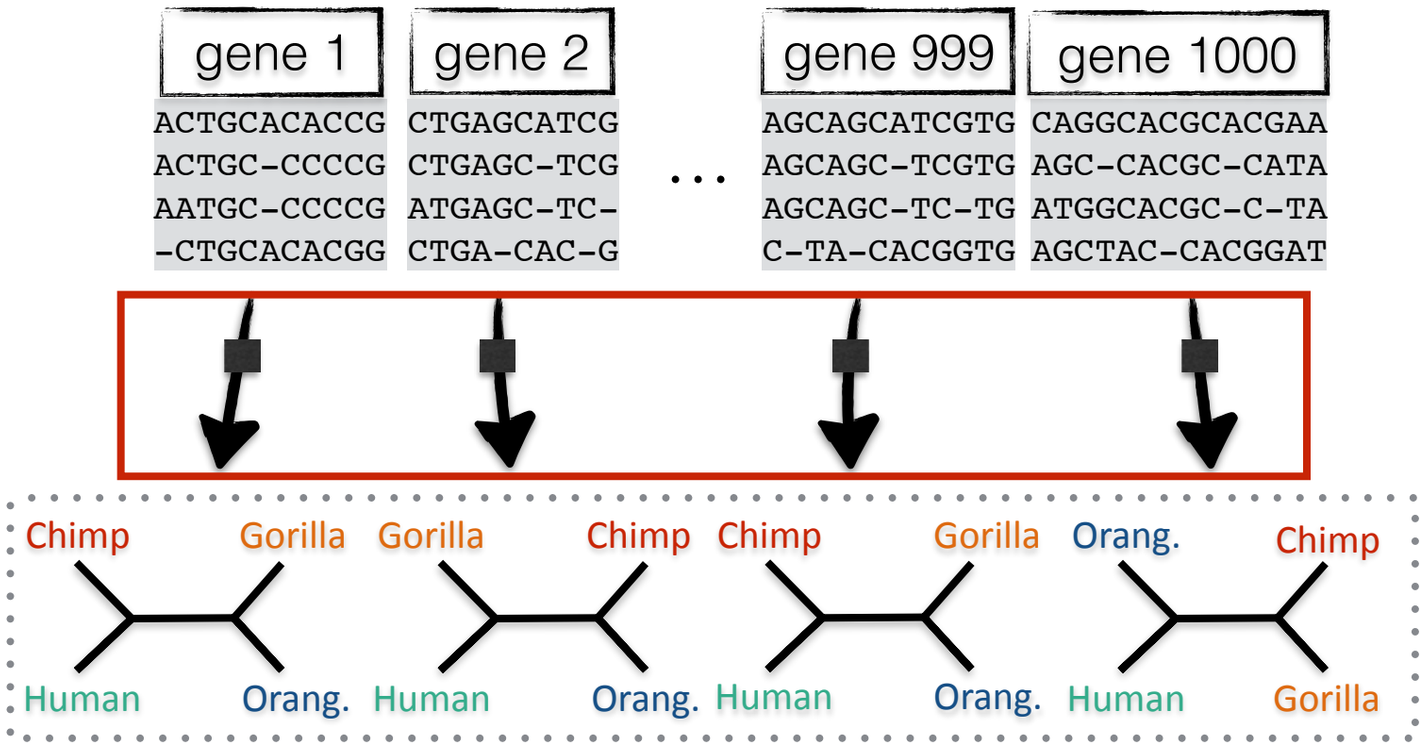


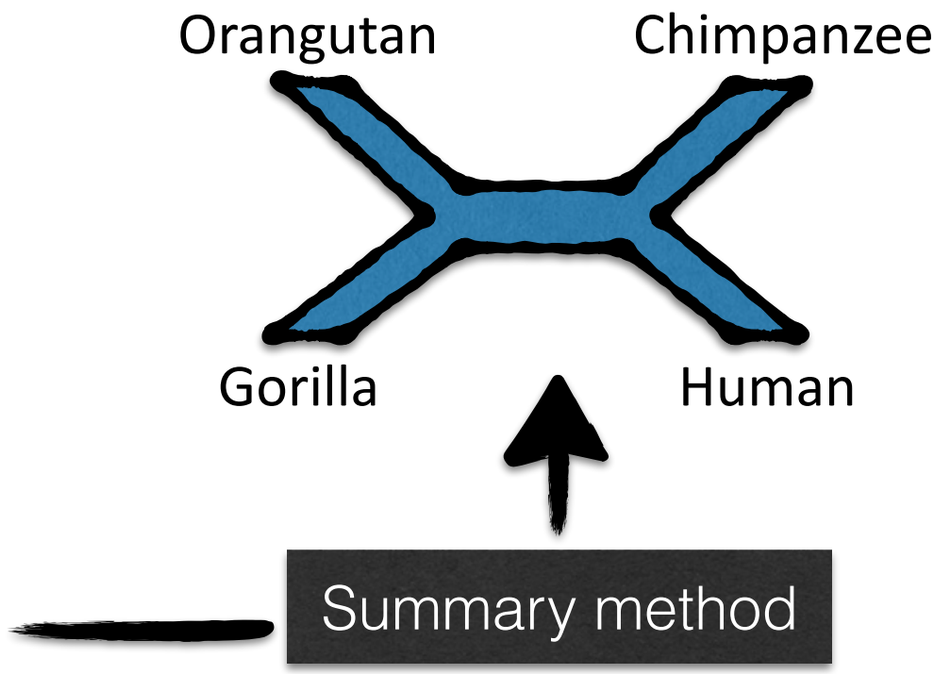
Can be statistically consistent

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



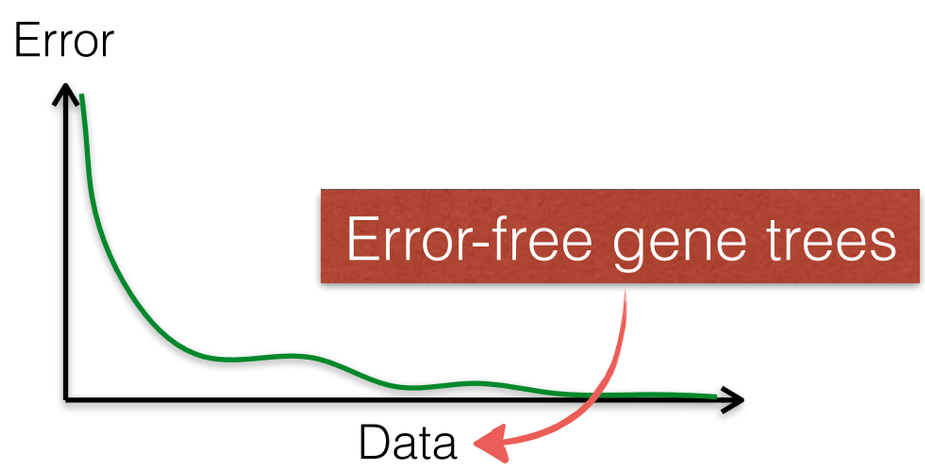
Species tree estimation from phylogenomic data (Approach 2: summary methods)





Can be statistically consistent

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



New methods

- Statistical Binning — improves estimation of gene tree distribution (input to summary methods)
  [Mirarab et al., Science, 2014] [Bayzid et al., PLoS ONE, 2015]

- ASTRAL — a summary method with improved accuracy and scalability compared to other summary methods
  [Mirarab et al., Bioinformatics, 2014 and 2015]

Recent large-scale phylogenomic projects

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

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 Zhang,[4] Hui Li,[4] Long Zhou,[4] Nitish Narula,[7,13] Liang Liu,[14] Ganesh Ganapathy,[1] Bastien Boussau,[15] Md. Shamsuzzoha Bayzid,[2] Volodymyr Zavidovych,[1] Sankar Subramanian,[16] Toni Gabaldón,[17,18,19] Salvador Capella-Gutiérrez,[17,18] Jaime Huerta-Cepas,[17,18] Bhanu Rekepalli,[20] Kasper Munch,[21] Mikkel Schierup,[21] Bent Lindow,[6] Wesley C. Warren,[22] David Ray,[23,24,25] Richard E. Green,[26] Michael W. Bruford,[27] Xiangjiang Zhan,[27,28] Andrew Dixon,[29] Shengbin Li,[30] Ning Li,[31] Yinhua Huang,[31] Elizabeth P. Derryberry,[32,33] Mads Frost Bertelsen,[34] Frederick H. Sheldon,[33] Robb T. Brumfield,[33] Claudio V. Mello,[35,36] Peter V. Lovell,[35] Morgan Wirthlin,[35] Maria Paula Cruz Schneider,[36,37] Francisco Prosdocimi,[36,38] José Alfredo Samaniego,[6] Amhed Missael Vargas Velazquez,[6] Alonzo Alfaro-Núñez,[6] Paula F. Campos,[6] Bent Petersen,[39] Thomas Sicheritz-Ponten,[39] An Pas,[40] Tom Bailey,[41] Paul Scofield,[42] Michael Bunce,[43] David M. Lambert,[16] Qi Zhou,[44] Polina Perelman,[45,46] Amy C. Driskell,[47] Beth Shapiro,[26] Zijun Xiong,[4] Yongli Zeng,[4] Shiping Liu,[4] Zhenyu Li,[4] Binghang Liu,[4] Kui Wu,[4] Jin Xiao,[4] Xiong Yinqi,[4] Qiuemei Zheng,[4] Yong Zhang,[4] Huanming Yang,[48] Jian Wang,[48] Linnea Smeds,[12] Frank E. Rheindt,[49] Michael Braun,[50] Jon Fjeldsa,[51] Ludovic Orlando,[6] F. Keith Barker,[52] Knud Andreas Jønsson,[51,53,54] Warren Johnson,[55] Klaus-Peter Koepfli,[56] Stephen O'Brien,[57,58] David Haussler,[59] Oliver A. Ryder,[60] Carsten Rahbek,[51,54] Eske Willerslev,[6] Gary R. Graves,[51,61] Travis C. Glenn,[62] John McCormack,[63] Dave Burt,[64] Hans Ellegren,[12] Per Alström,[65,66] Scott V. Edwards,[67] Alexandros Stamatakis,[3,68] David P. Mindell,[69] Joel Cracraft,[70] Edward L. Braun,[71] Tandy Warnow,[2,72]† Wang Jun,[48,73,74,75,76]† M. Thomas P. Gilbert,[6,43]† Guojie Zhang[4,77]†

1K Plants (1KP) [Wickett, Mirarab, et al., PNAS, 2014]

PNAS

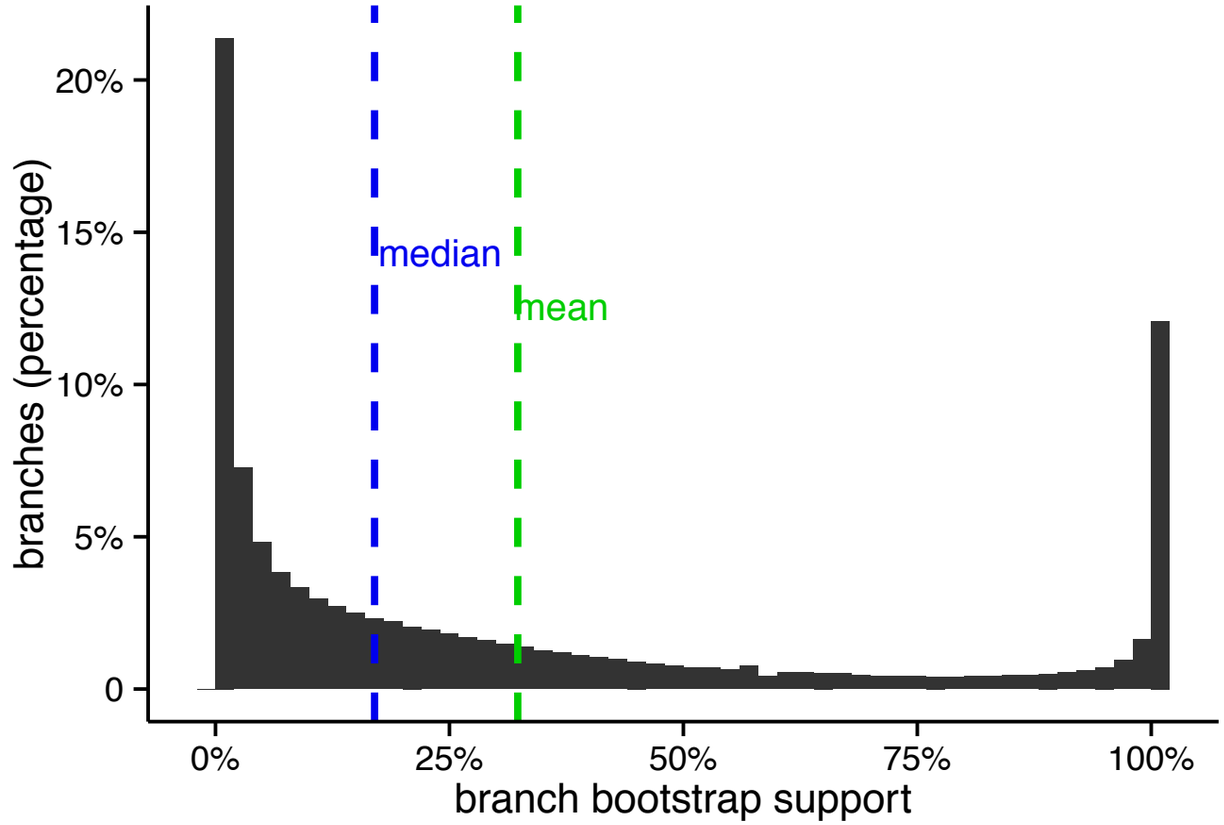
Phylotranscriptomic analysis of the origin and early diversification of land plants

Norman J. Wickett[a,b,1,2], Siavash Mirarab[c,1], Nam Nguyen[c], Tandy Warnow[c], Eric Carpenter[d], Naim Matasci[e,f], Saravanaraj Ayyampalayam[g], Michael S. Barker[f], J. Gordon Burleigh[h], Matthew A. Gitzendanner[h,i], Brad R. Ruhfel[h,j,k], Eric Wafula[l], Joshua P. Der[l], Sean W. Graham[m], Sarah Mathews[n], Michael Melkonian[o], Douglas E. Soltis[h,i,k], Pamela S. Soltis[h,i,k], Nicholas W. Miles[k], Carl J. Rothfels[p,q], Lisa Pokorny[p,r], A. Jonathan Shaw[p], Lisa DeGironimo[s], Dennis W. Stevenson[s], Barbara Surek[o], Juan Carlos Villarreal[t], Béatrice Roure[u], Hervé Philippe[u,v], Claude W. dePamphilis[l], Tao Chen[w], Michael K. Deyholos[d], Regina S. Baucom[x], Toni M. Kutchan[y], Megan M. Augustin[y], Jun Wang[z], Yong Zhang[v], Zhijian Tian[z], Zhixiang Yan[z], Xiaolei Wu[z], Xiao Sun[z], Gane Ka-Shu Wong[d,z,aa,2], and James Leebens-Mack[g,2]

Challenges of avian phylogenomics

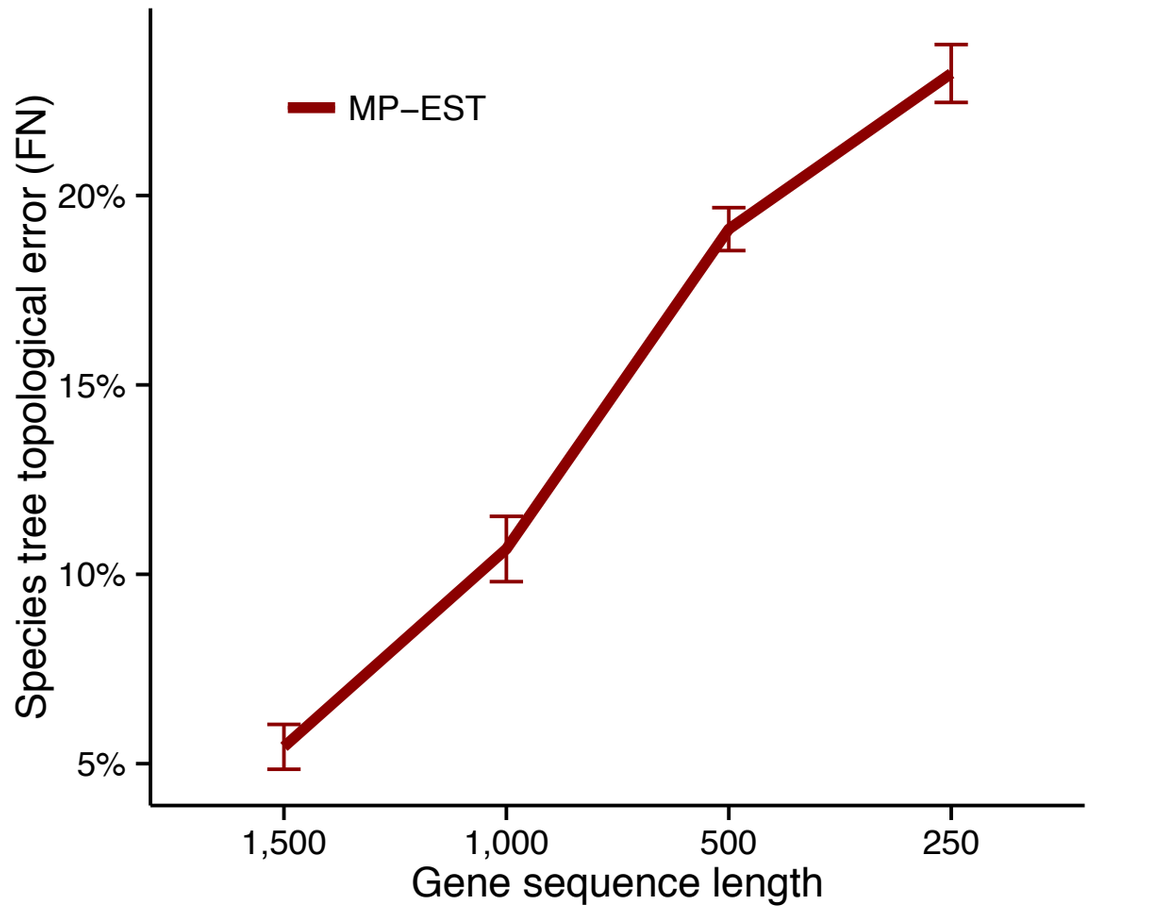
[Jarvis, Mirarab, et al., Science, 2014]

- Whole genomes for 48 bird species (~100m years of evolution)
- Goal: a phylogeny of bird orders
- Extremely challenging due to rampant gene tree incongruence
- 14,000 “noisy” genes (i.e., high gene tree error)



bootstrap support of all branches
in all estimated gene tree

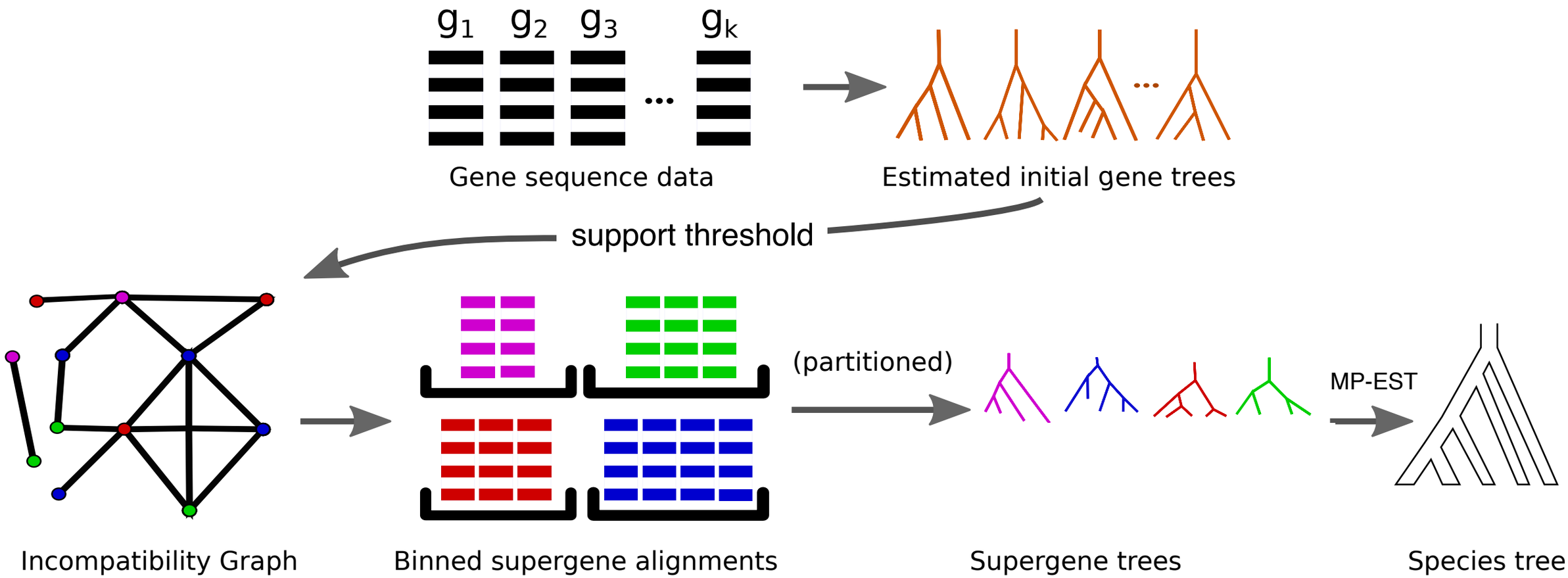
Gene tree estimation error impacts species tree estimation



more gene tree error →

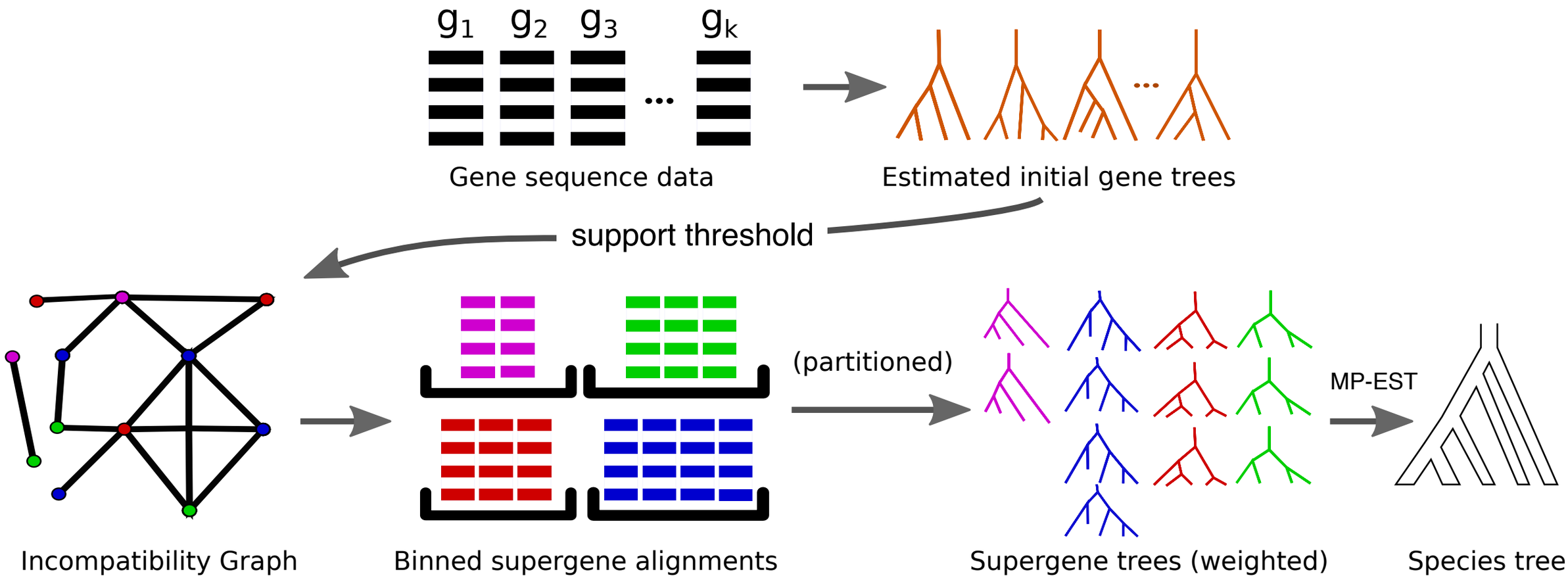
45 avian-like species, 1000 genes
[Mirarab, et al., Science, 2014]

Statistical binning: overview



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

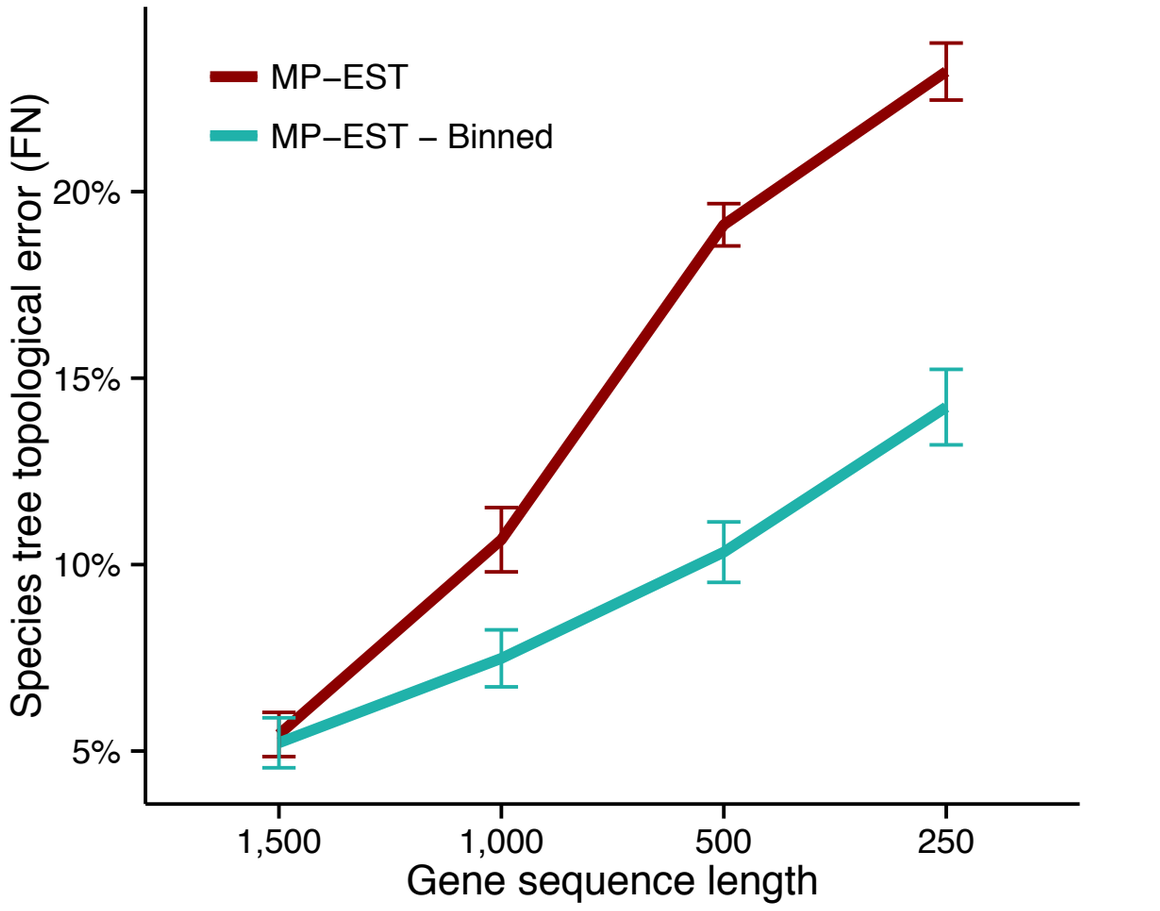
Statistical binning: overview



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

New version: weighted; statistically consistent [Bayzid, Mirarab, Warnow, PloS ONE, 2015]

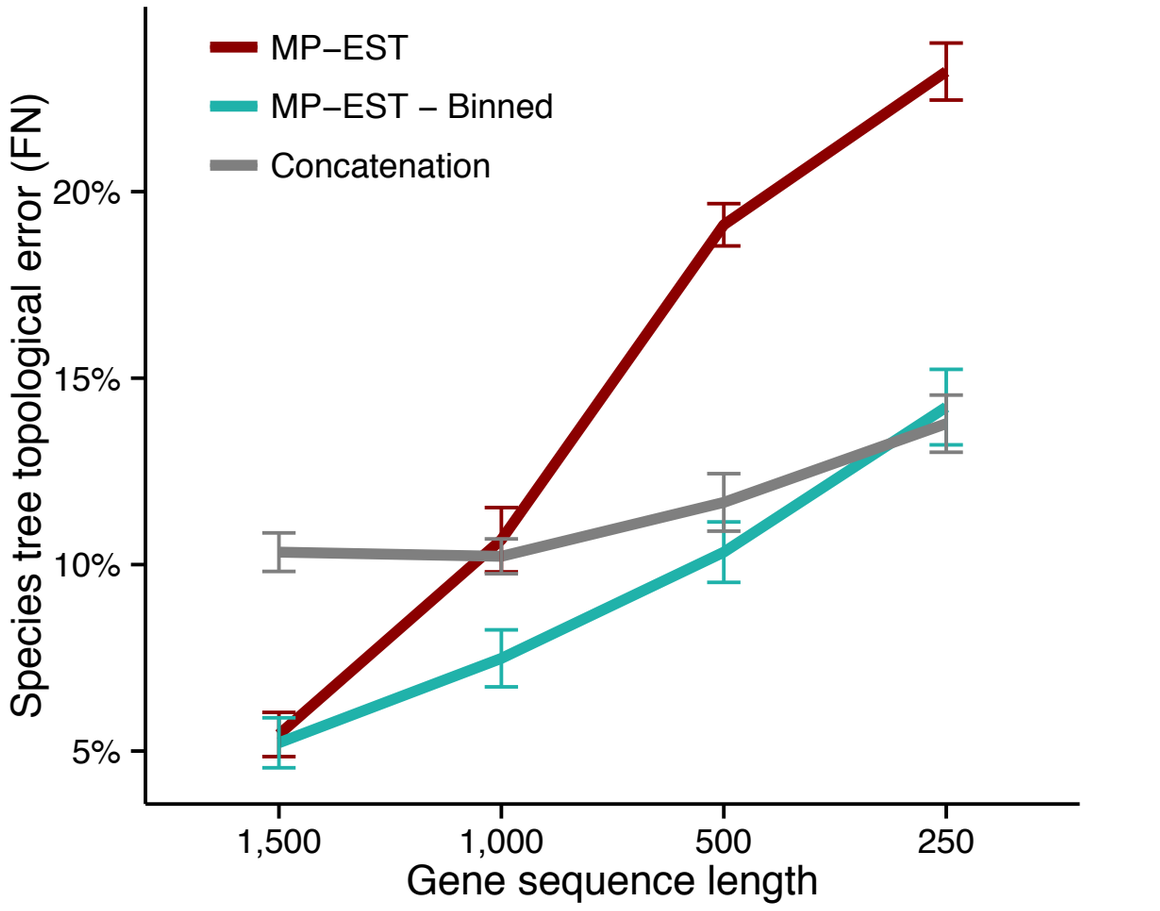
Statistical binning improves species tree estimation



more gene tree error →

45 avian-like species, 1000 genes
[Mirarab, et al., Science, 2014]

Statistical binning improves species tree estimation



more gene tree error →

45 avian-like species, 1000 genes
[Mirarab, et al., Science, 2014]

Binning also improves other measures of accuracy

- More accurate gene tree distributions

Binning also improves other measures of accuracy

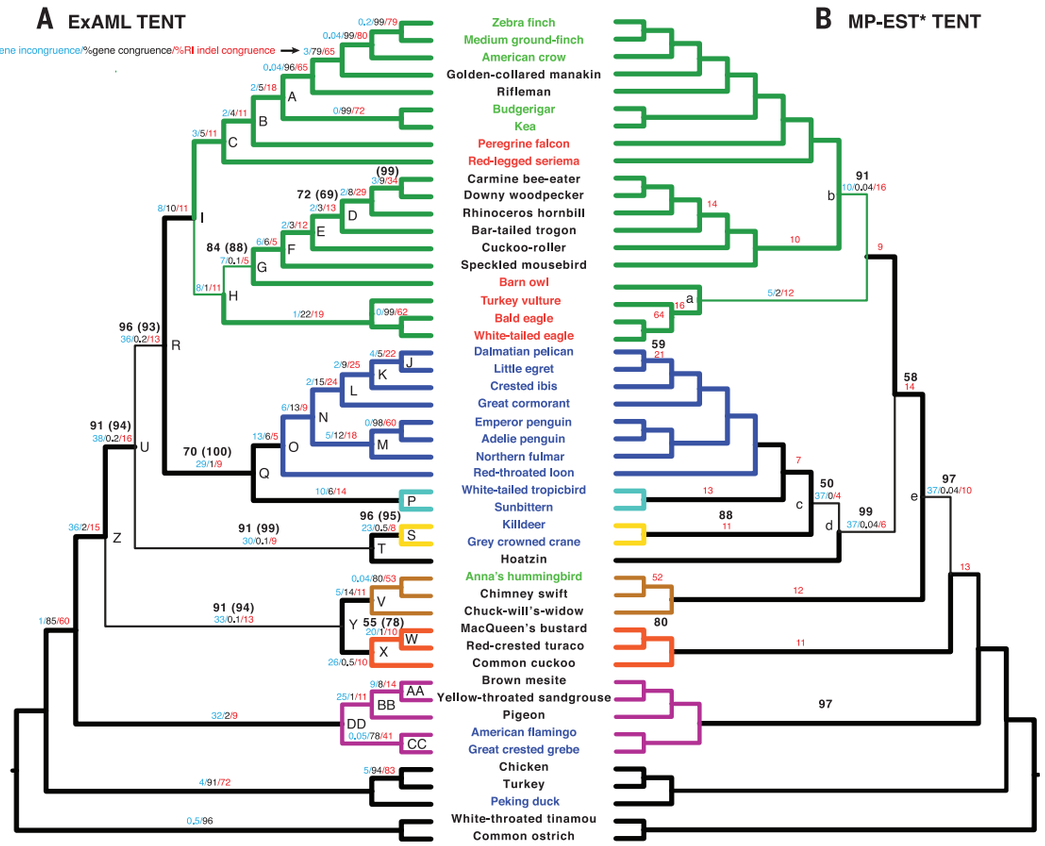
- More accurate gene tree distributions
- Better Species tree bootstrap support
  - Higher support for true positives
  - Fewer false positives with high support

Binning also improves other measures of accuracy

- More accurate gene tree distributions
- Better Species tree bootstrap support
  - Higher support for true positives
  - Fewer false positives with high support
- More accurate species tree branch lengths
  - Hence, estimates of ILS levels

Binning on the Avian dataset

Binned analyses results in a tree that was largely congruent with the concatenation

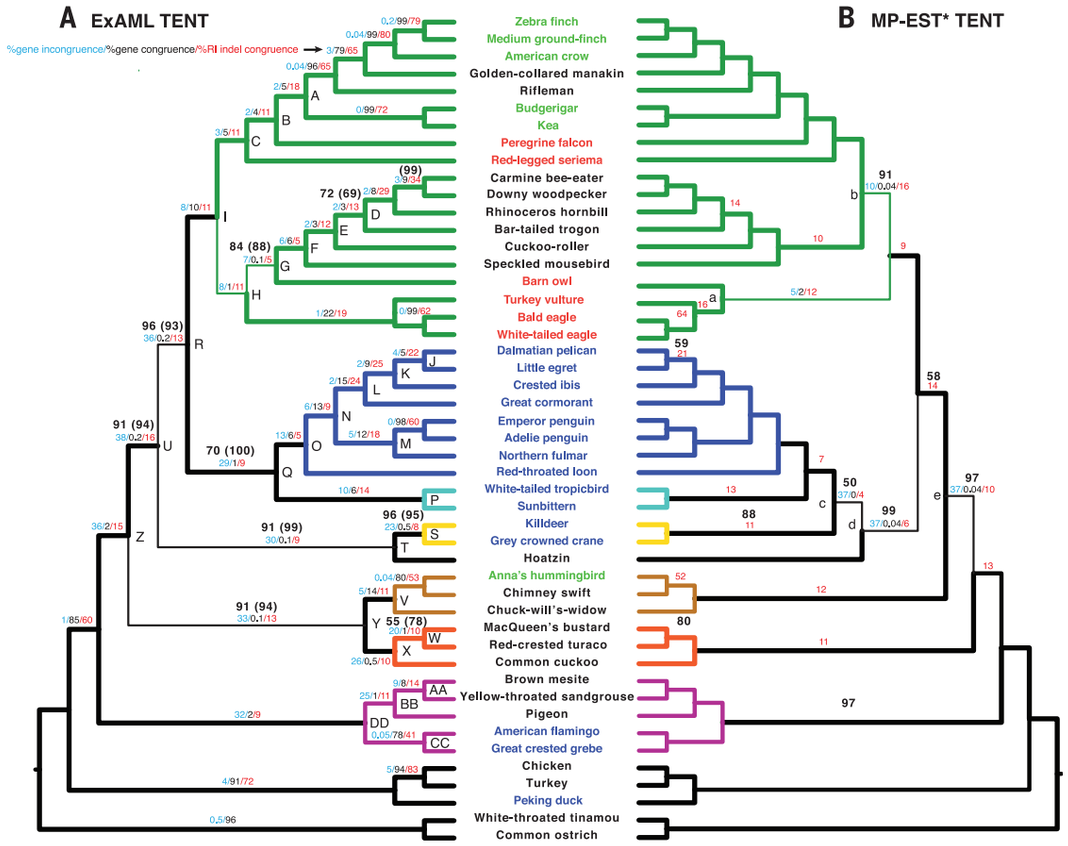


[Jarvis, Mirarab, et al., Science, 2014]

Binning on the Avian dataset

Binned analyses results in a tree that was largely congruent with the concatenation

The binned analyses recovered almost all highly supported gene tree branches



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

New methods

- Statistical Binning — improves estimation of gene tree distribution (input to summary methods)
[Mirarab et al., Science, 2014] [Bayzid et al., PLoS ONE, 2015]

1KP: Plant whole transcriptomes

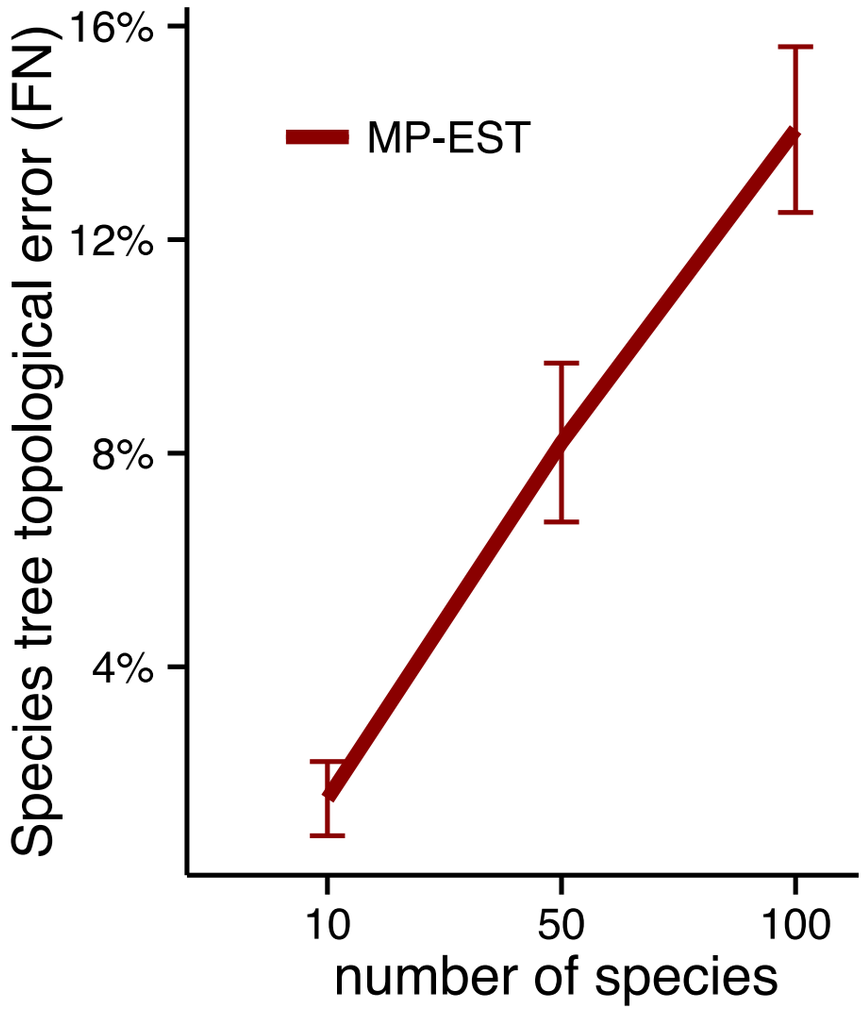
PNAS

Phylotranscriptomic analysis of the origin and early diversification of land plants

Norman J. Wickett[a,b,1,2], Siavash Mirarab[c,1], Nam Nguyen[c], Tandy Warnow[c], Eric Carpenter[d], Naim Matasci[e,f], Saravanaraj Ayyampalayam[g], Michael S. Barker[f], J. Gordon Burleigh[h], Matthew A. Gitzendanner[h,i], Brad R. Ruhfel[h,j,k], Eric Wafula[l], Joshua P. Der[l], Sean W. Graham[m], Sarah Mathews[n], Michael Melkonian[o], Douglas E. Soltis[h,i,k], Pamela S. Soltis[h,i,k], Nicholas W. Miles[k], Carl J. Rothfels[p,q], Lisa Pokorny[p,r], A. Jonathan Shaw[p], Lisa DeGironimo[s], Dennis W. Stevenson[s], Barbara Surek[o], Juan Carlos Villarreal[t], Béatrice Roure[u], Hervé Philippe[u,v], Claude W. dePamphilis[l], Tao Chen[w], Michael K. Deyholos[d], Regina S. Baucom[x], Toni M. Kutchan[y], Megan M. Augustin[y], Jun Wang[z], Yong Zhang[v], Zhijian Tian[z], Zhixiang Yan[z], Xiaolei Wu[z], Xiao Sun[z], Gane Ka-Shu Wong[d,z,aa,2], and James Leebens-Mack[g,2]

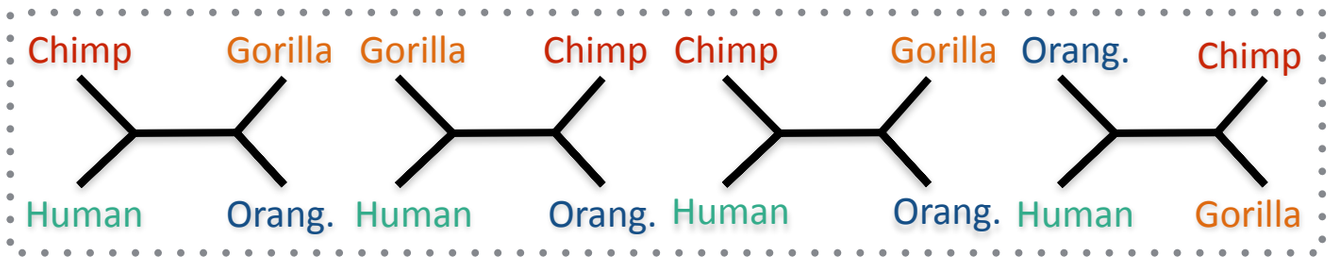
- Whole transcriptomes for <u>103 plant species</u>
  - 1,200 in the next phase
- 400-800 single copy “genes”
- Spans ~1 billion years of evolution
- Many unanswered questions about plant evolution

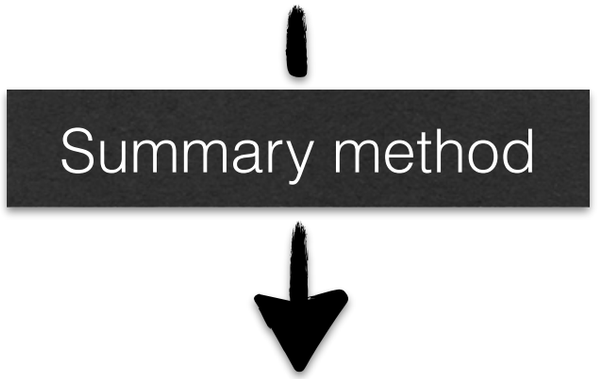
Number of species impacts estimation error in the species tree

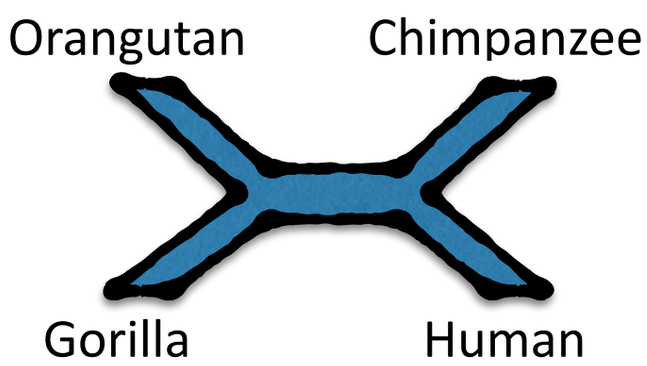


1000 genes, “medium” levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]

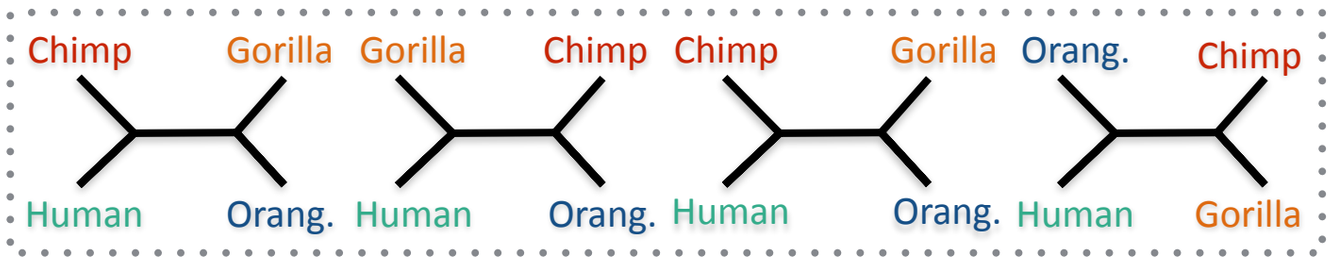
Designing summary methods for quartets of taxa

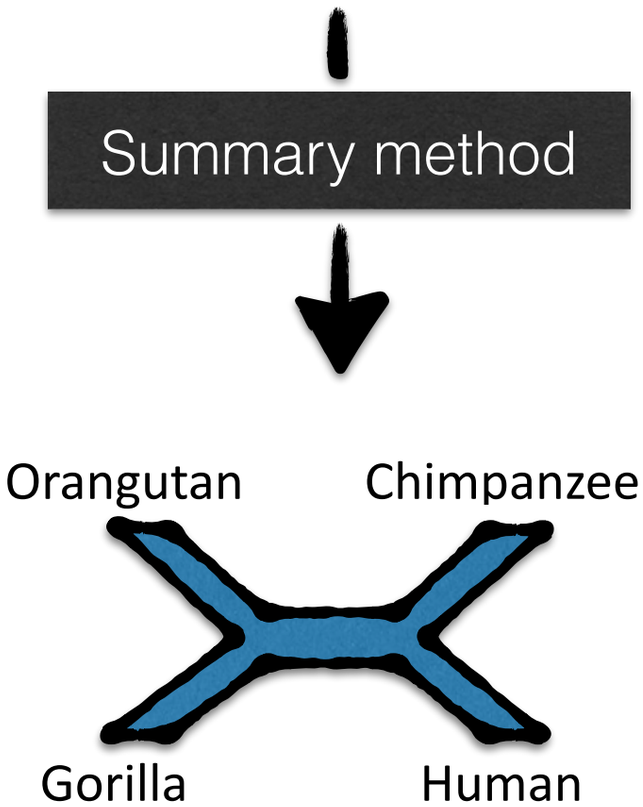




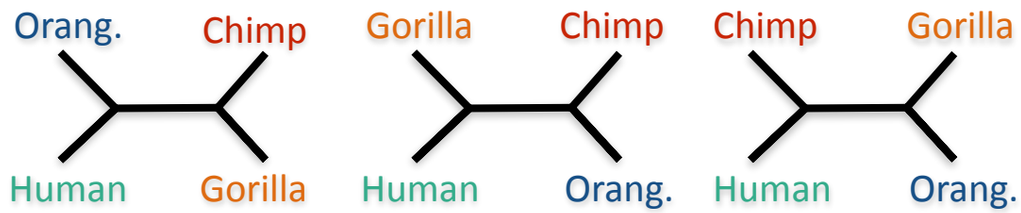


Designing summary methods for quartets of taxa

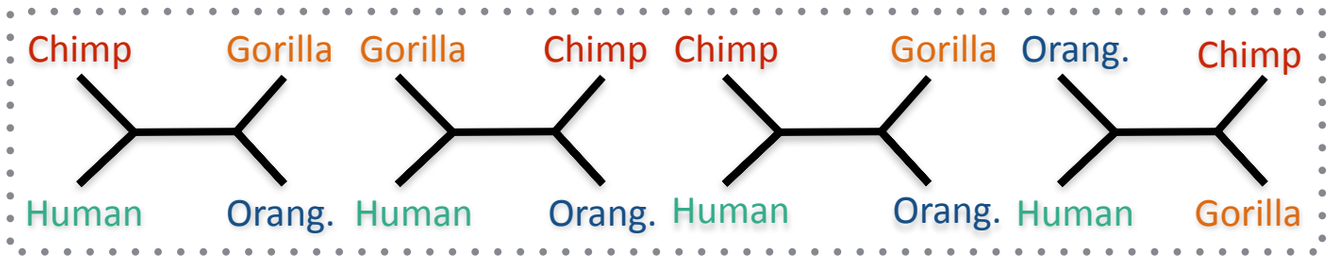


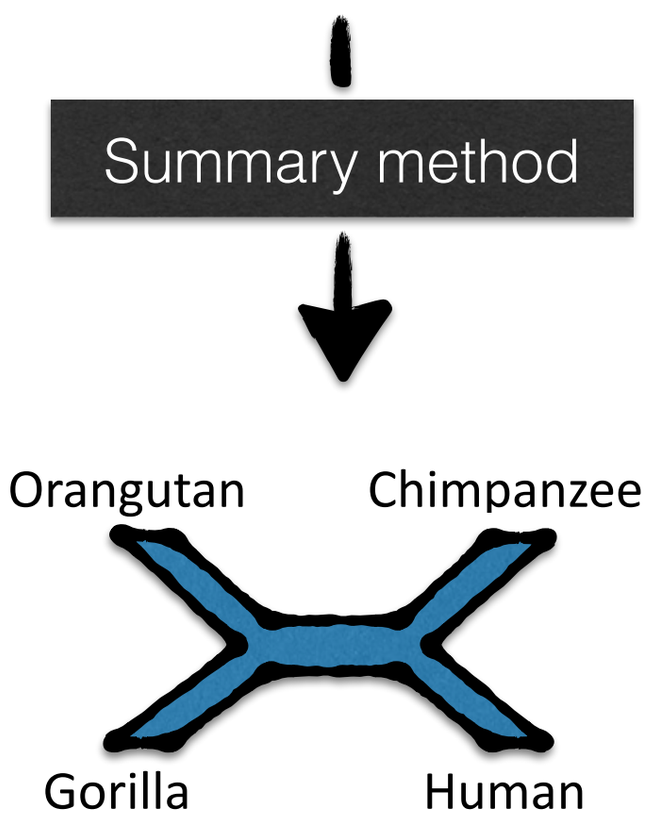


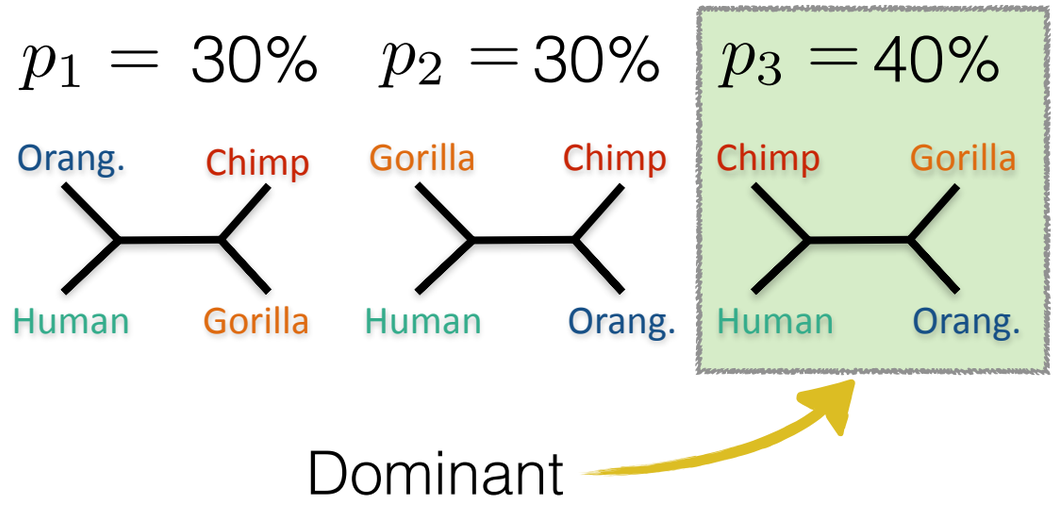
$p_1 = 30\%$ $p_2 = 30\%$ $p_3 = 40\%$



Designing summary methods for quartets of taxa

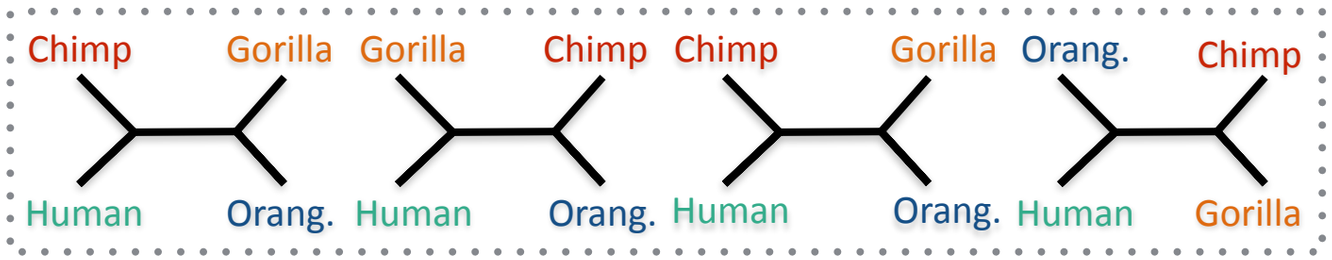


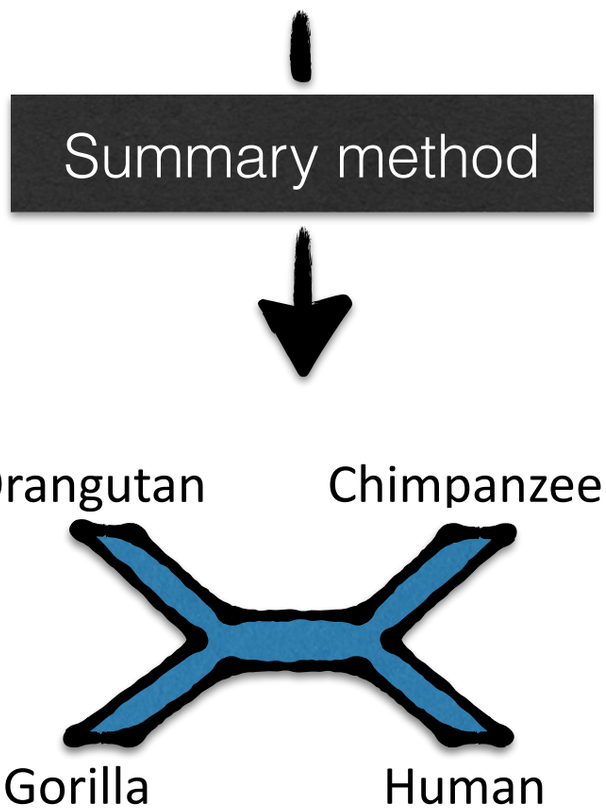


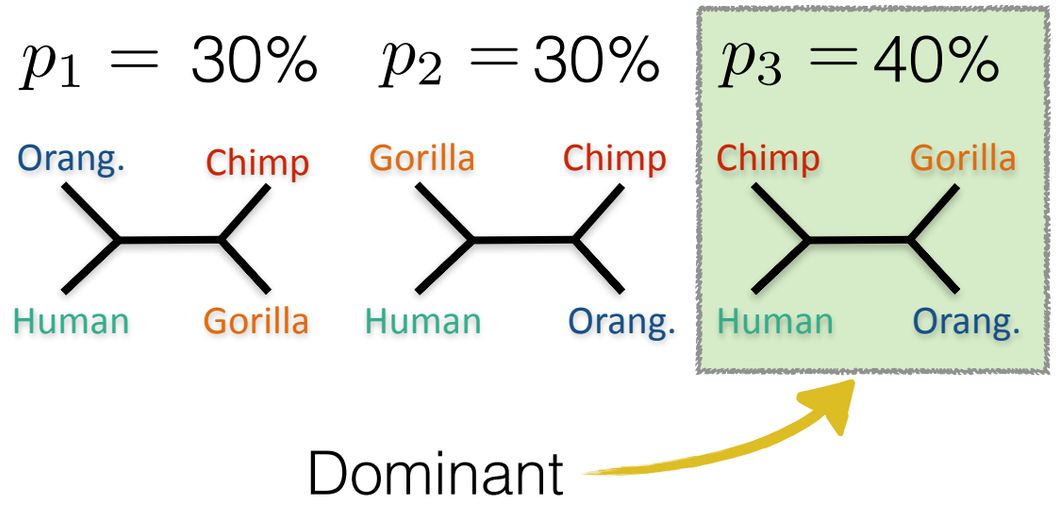


- For 4 species, the dominant quartet topology is the species tree
  [Allman, et al. 2010], [Degnan, 2014]

Designing summary methods for quartets of taxa

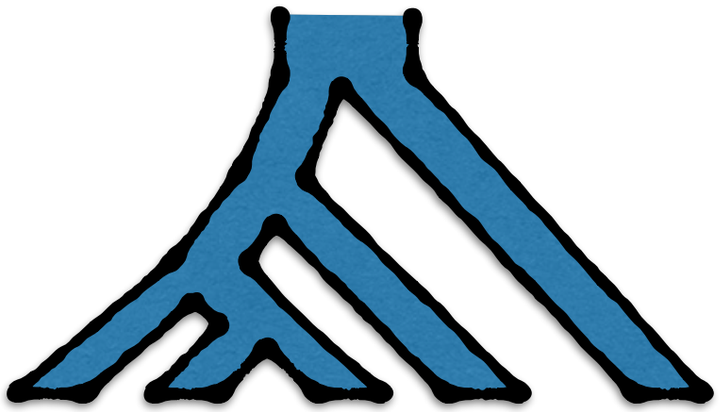






- For 4 species, the dominant quartet topology is the species tree
  [Allman, et al. 2010], [Degnan, 2014]
- For >4 species, the dominant topology may be different from the species tree
  [Degnan and Rosenberg, 2006, 2009]

More than 4 species



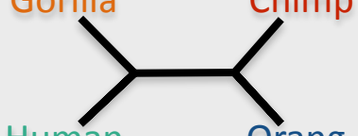
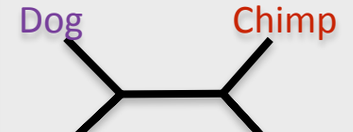
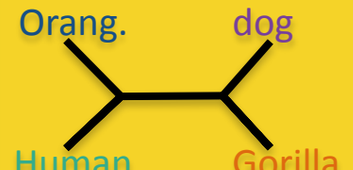
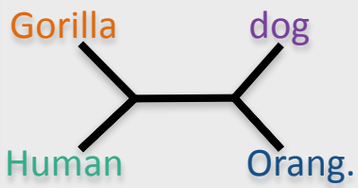
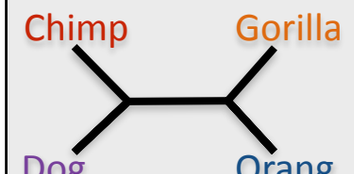
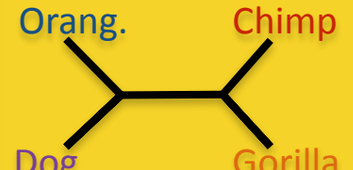
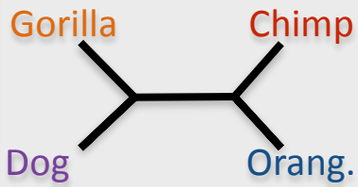
1. Break gene trees into $\binom{n}{4}$ quartets of species

2. Find the dominant tree for all quartets of taxa

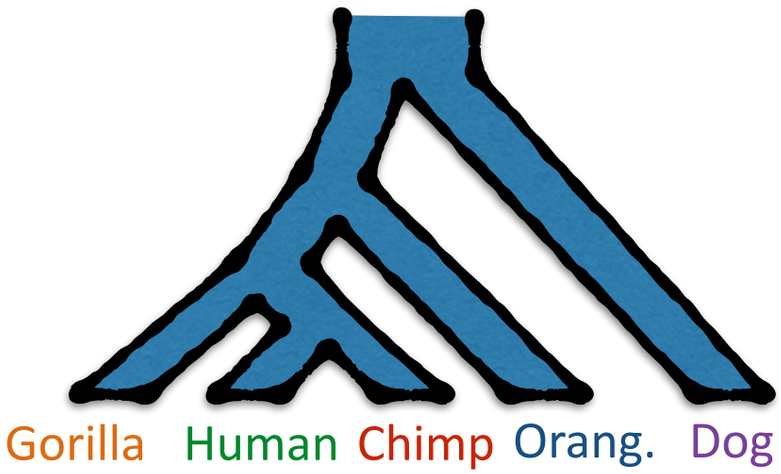
3. Combine quartet trees

Example: BUCKy-pop. [Larget, et al., Bioinformatics, 2010]

(probabilities are made-up just as an example)

Gorilla Human Orangutan Chimp	 50%	 25%	 25%
Gorilla Human Dog Chimp	 55%	 21%	 24%
Gorilla Human Orangutan Dog	 7%	 87%	 6%
Gorilla Dog Orangutan Chimp	 6%	 88%	 6%
Dog Human Orangutan Chimp	 95%	 2%	 3%

ASTRAL: weighting by frequency



1. Break gene trees into $\binom{n}{4}$ quartets of species
2. Find the tree that "satisfies" the maximum number of weighted quartets from gene trees

(probabilities are made-up just as an example)

Gorilla Human / Orangutan Chimp 50%	Orang. Chimp / Human Gorilla 25%	Gorilla Chimp / Human Orang. 25%
Gorilla Human / Dog Chimp 55%	Dog Chimp / Human Gorilla 19%	Gorilla Chimp / Human Dog 26%
Gorilla Human / Orangutan Dog 7%	Orang. dog / Human Gorilla 87%	Gorilla dog / Human Orang. 6%
Gorilla Dog / Orangutan Chimp 6%	Orang. Chimp / Dog Gorilla 88%	Gorilla Chimp / Dog Orang. 6%
Dog Human / Orangutan Chimp 95%	Orang. Chimp / Human Dog 2%	Dog Chimp / Human Orang. 3%

Maximum Quartet Support Species Tree

[Mirarab, et al., ECCB, 2014]

- Optimization Problem (suspected NP-Hard):

Find the species tree with the maximum number of induced quartet trees shared with the collection of input gene trees

$$Score(T) = \sum_{t \in \mathcal{T}} |Q(T) \cap Q(t)|$$

Set of quartet trees induced by T

a gene tree

all input gene trees

- **Theorem**: Statistically consistent under the multi-species coalescent model when solved exactly

ASTRAL

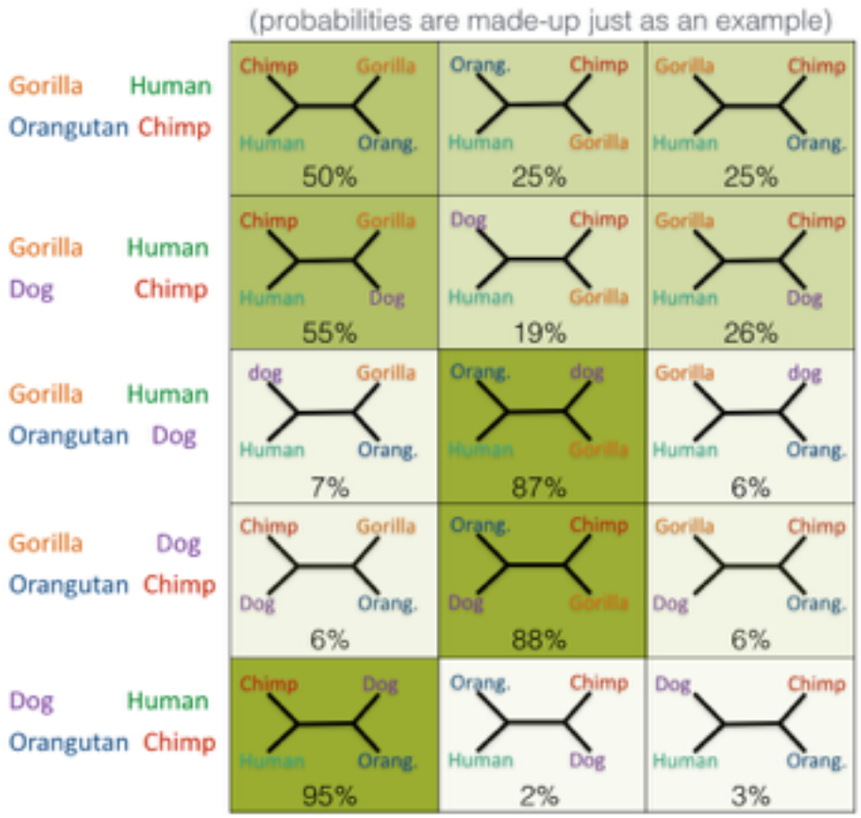
- We developed a dynamic programming algorithm to solve the problem exactly
  - Exponential running time (still feasible for <18 species)
- Developed a constrained version of the problem that can be solved exactly in polynomial time
  - Constraints search to branches that appear in gene trees ($\mathcal{X}$)
  - Runs for 1000 species and 1000 genes in about a day
  - Remains statistically consistent

Asymptotic running time

- $O(nk|\mathcal{X}|^2)$ for *k* genes of *n* species

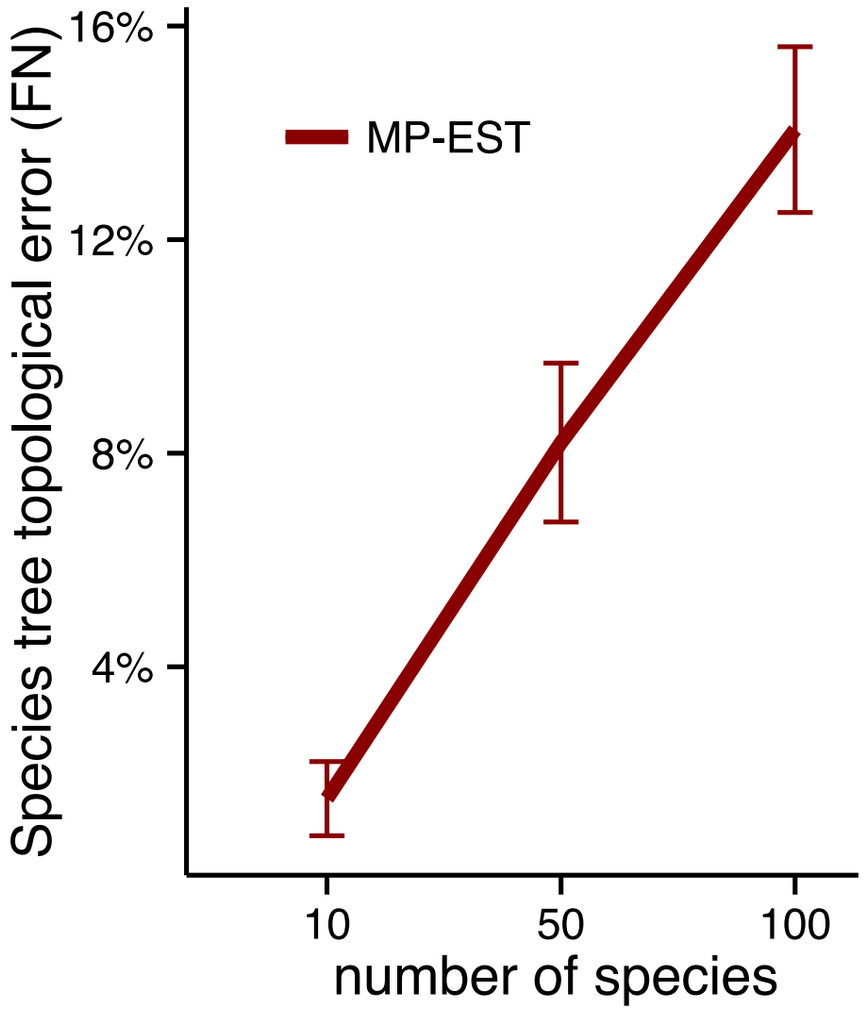
Asymptotic running time

- $O(nk|\mathcal{X}|^2)$ for *k* genes of *n* species



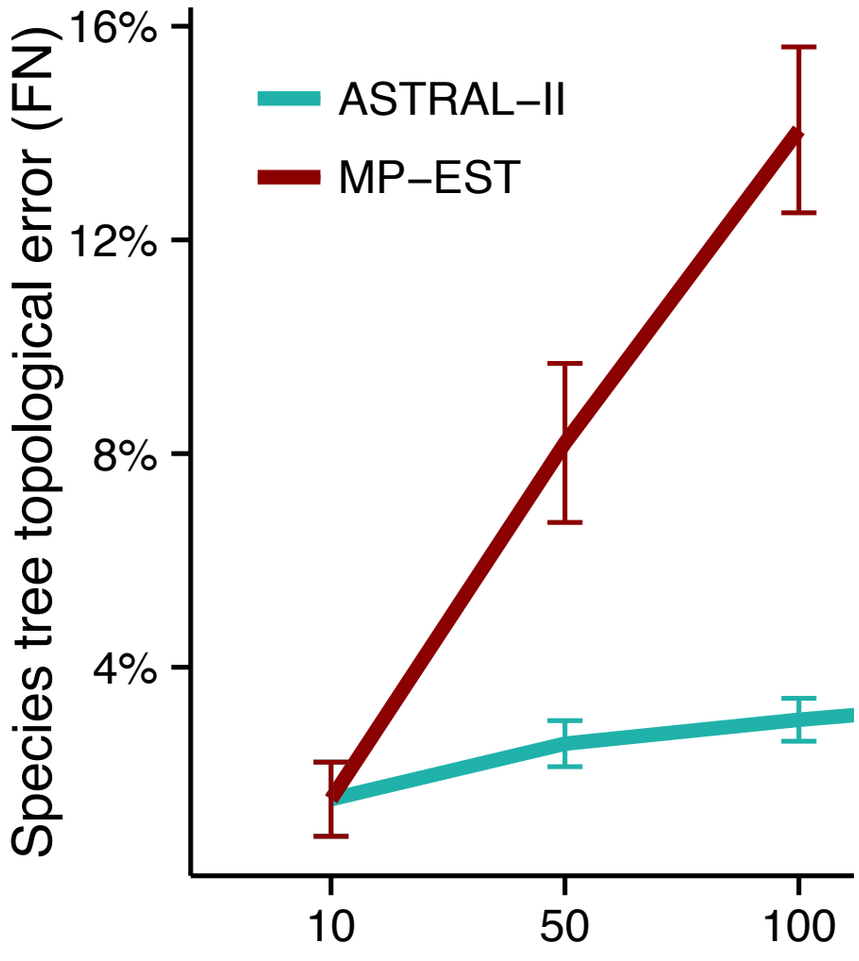
- Surprise: running time is better than $\Theta(n^4)$
- Don't we have to at least list all $\binom{n}{4}$ quartets?
- No! we calculate scores without listing $\binom{n}{4}$ quartets

Number of species impacts estimation error in the species tree



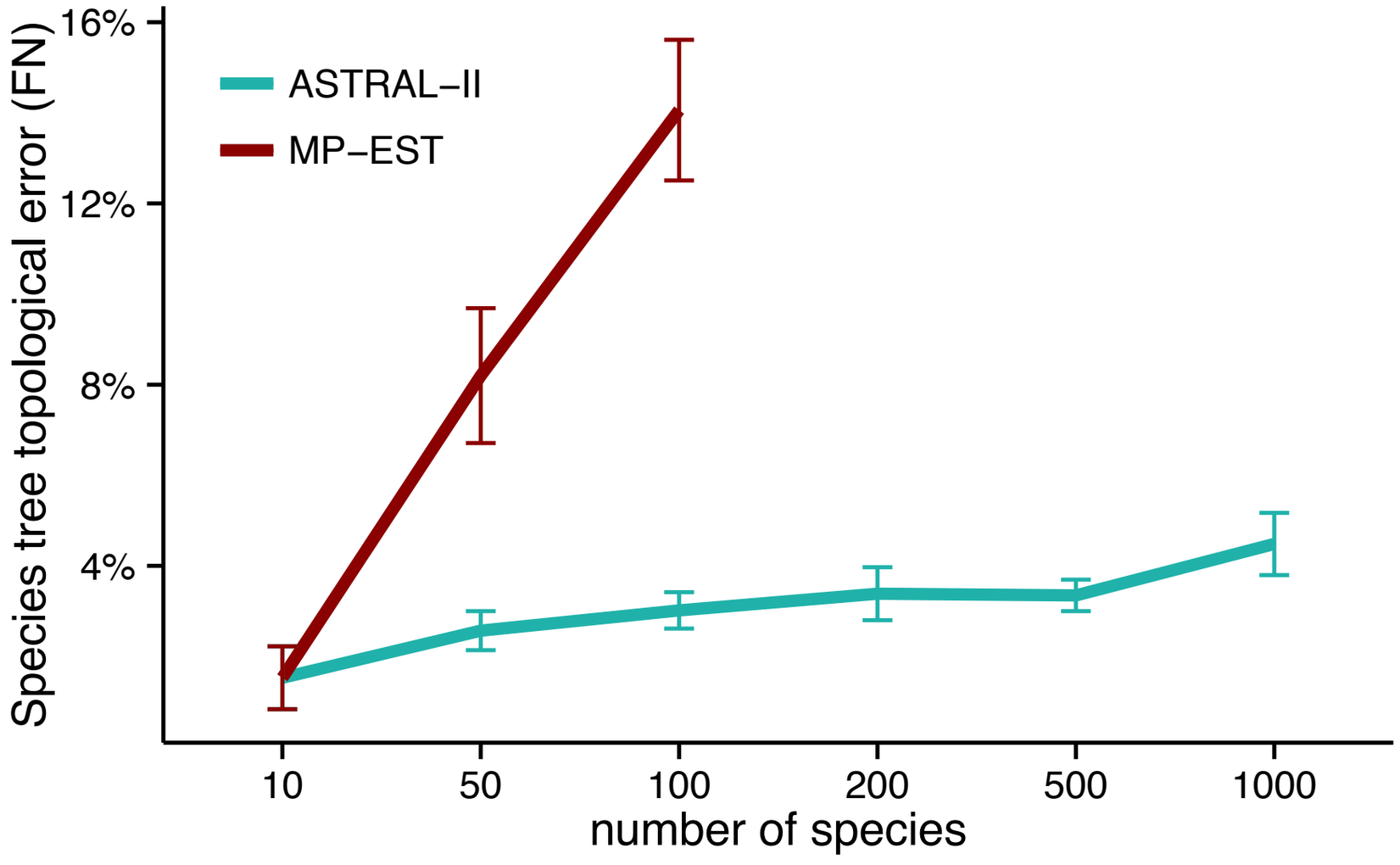
1000 genes, “medium” levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]

ASTRAL: accurate and scalable



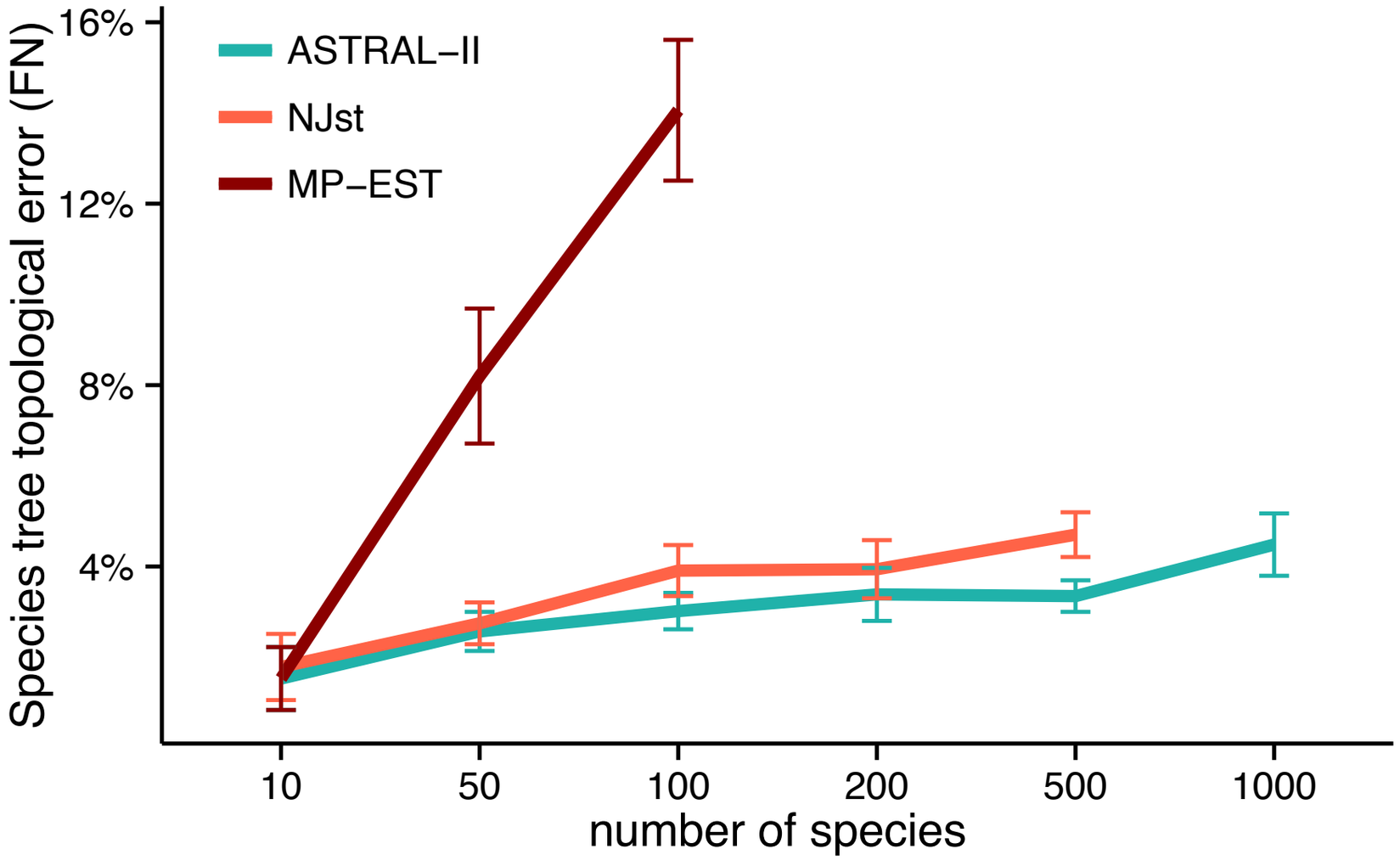
1000 genes, “medium” levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]

ASTRAL: accurate and scalable



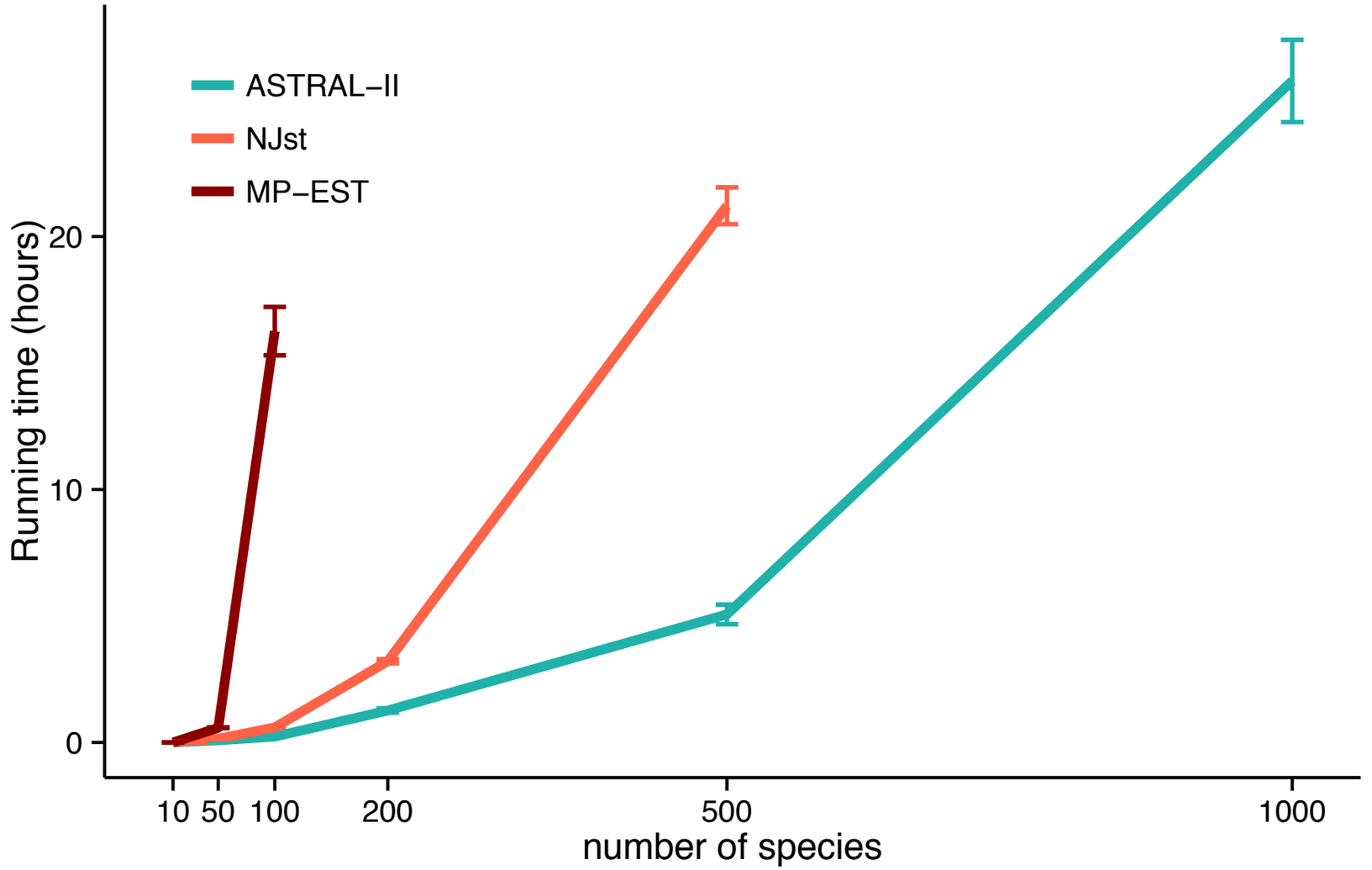
1000 genes, “medium” levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]

ASTRAL: accurate and scalable



1000 genes, “medium” levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]

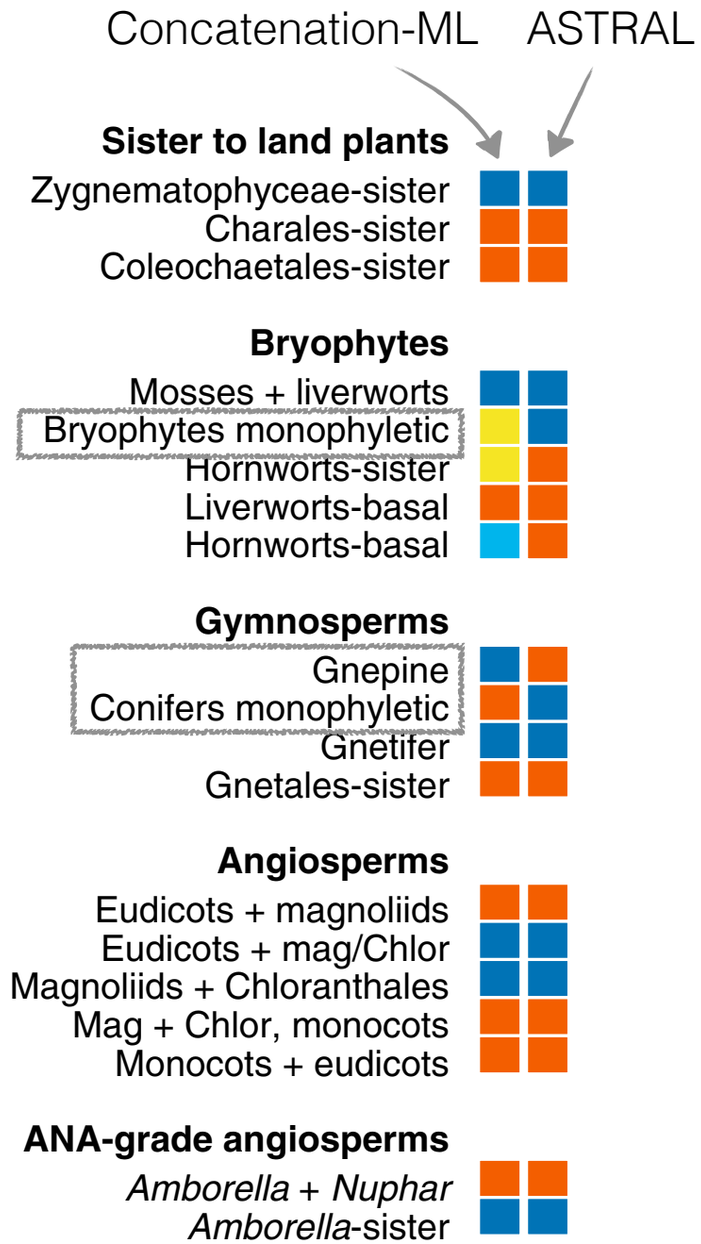
Running time as function of # species



1000 genes, “medium” levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]

ASTRAL on plants dataset

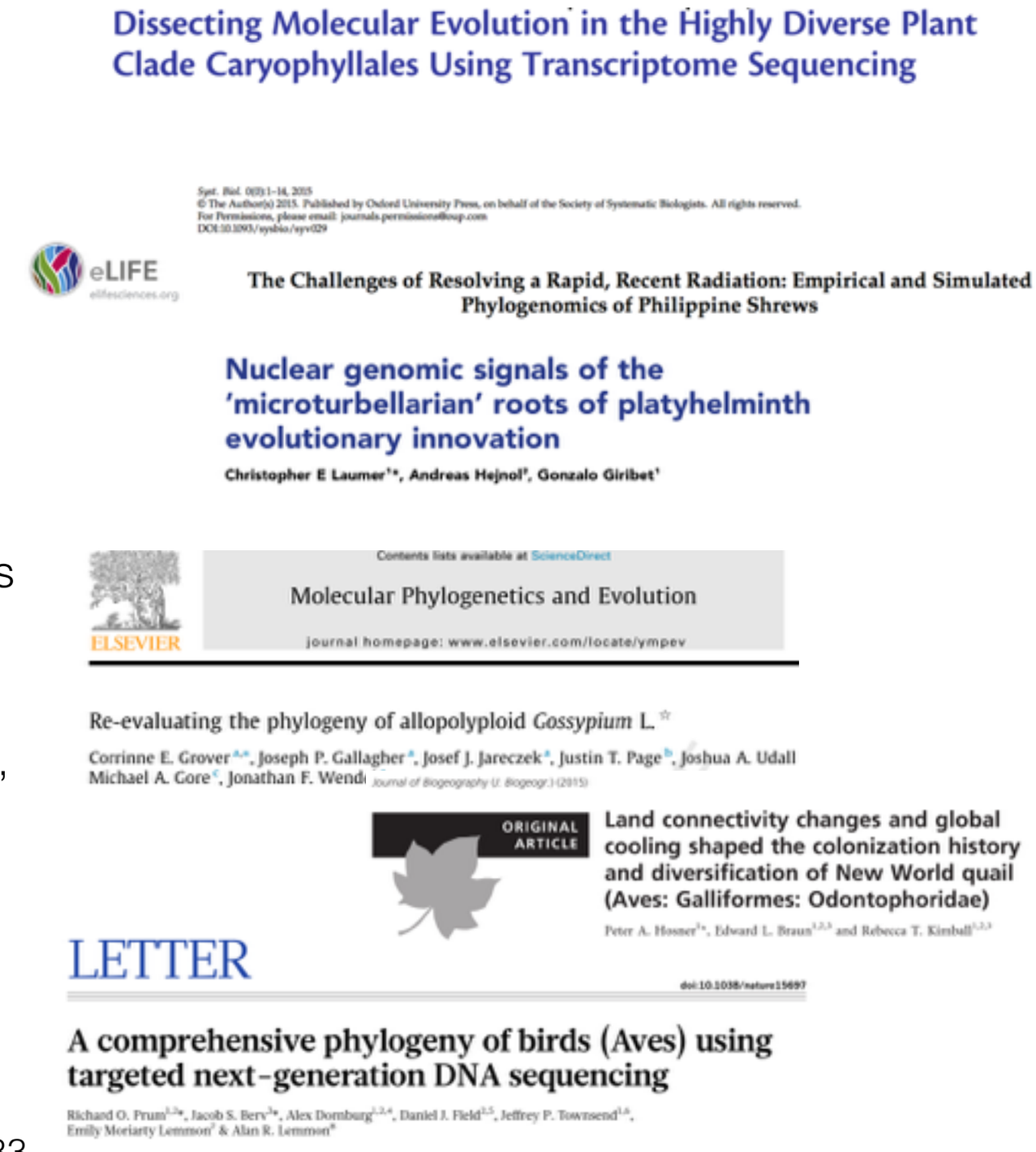
- The ASTRAL tree:
  - High support
  - Similar to concatenation with some interesting differences (e.g., recovered bryophytes)
- ASTRAL took only about 10 minutes (serial running time) on 103 taxa and 400 genes



[Wickett, Mirarab, et al., PNAS, 2014]

ASTRAL on biological datasets

- 1KP: **103** plant species, 400-800 genes
- Yang, et al. **96** Caryophyllales species, 1122 genes
- Dentinger, et al. **39** mushroom species, 208 genes
- Giarla and Esselstyn. **19** Philippine shrew species, 1112 genes
- Laumer, et al. **40** flatworm species, 516 genes
- Grover, et al. **8** cotton species, 52 genes
- Hosner, Braun, and Kimball. **28** quail species, 11 genes
- Simmons and Gatesy. **47** angiosperm species, 310 genes
- Prum et al, **198** avian species, 259 genes



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)

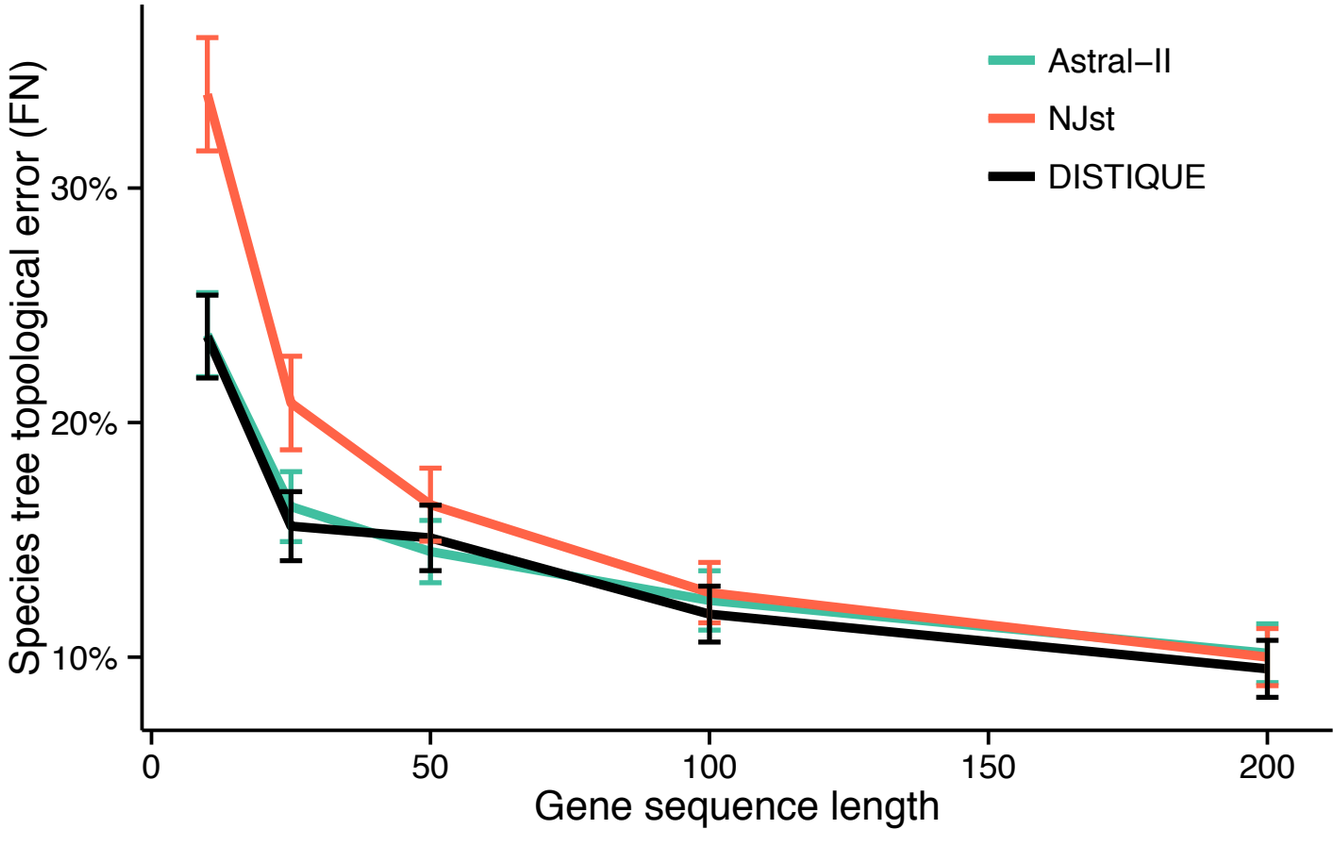
New methods

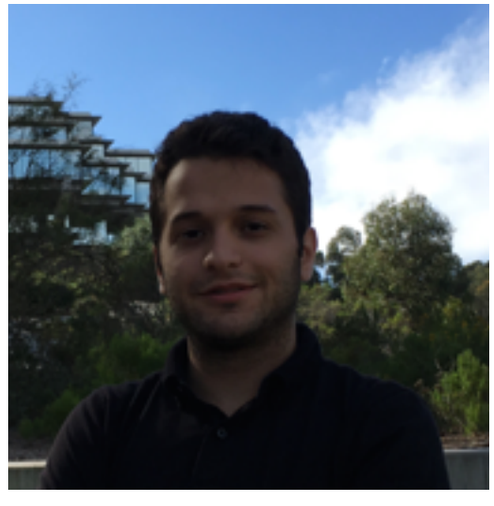
- Statistical Binning — improves estimation of gene tree distribution (input to summary methods)
  [Mirarab et al., Science, 2014] [Bayzid et al., PLoS ONE, 2015]

- ASTRAL — a summary method with improved accuracy and scalability compared to other summary methods
  [Mirarab et al., Bioinformatics, 2014 and 2015]

Ongoing work

DISTIQUE — a family of summary method, some with quadratic running time and interesting theory
[unpublished and unfinished work; with Erfan Sayyari]





Summary

- Genome-scale sequence data provides a wealth of information
- Yet, reconstruction of species phylogenies remains challenging
  - Scalability to many species: ASTRAL
  - Limited data per gene: statistical binning
  - Impact of model violations
  - Recombination
  - Missing data
  - Multiple sources of gene tree discordance
- Many interesting statistical and computational questions and need for method development

Acknowledgments

Kes **Tandy Warnow**

Jim Leebens-mack
(UGA)

Norman Wickett
(U Chicago)

Gane Wong
(U of Alberta)

. . .

S.M. Bayzid

Nam Nguyen
(now at UIUC)

Théo
Zimmermann

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