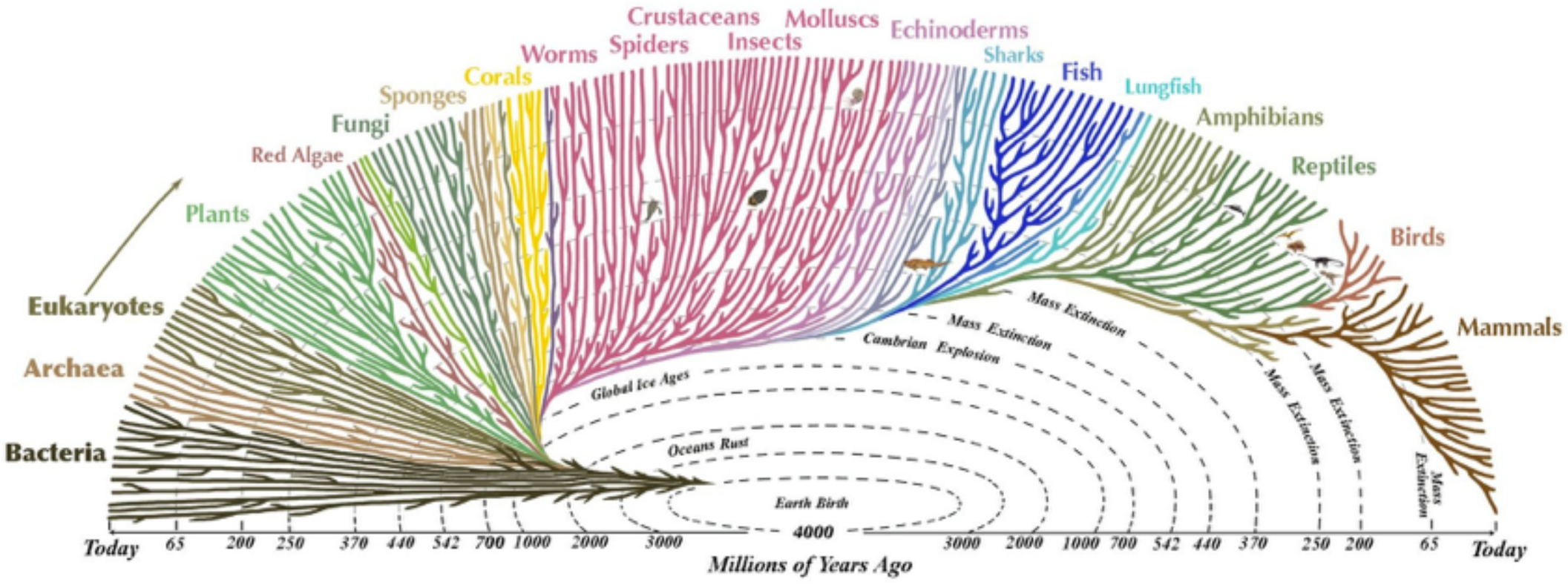
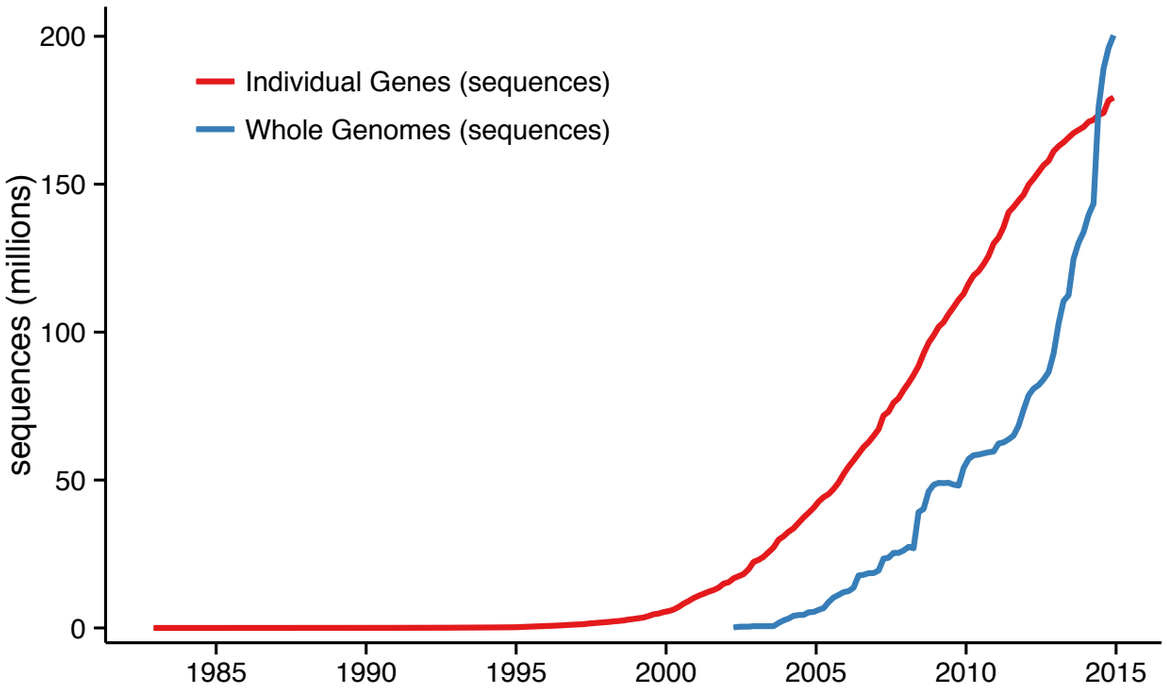


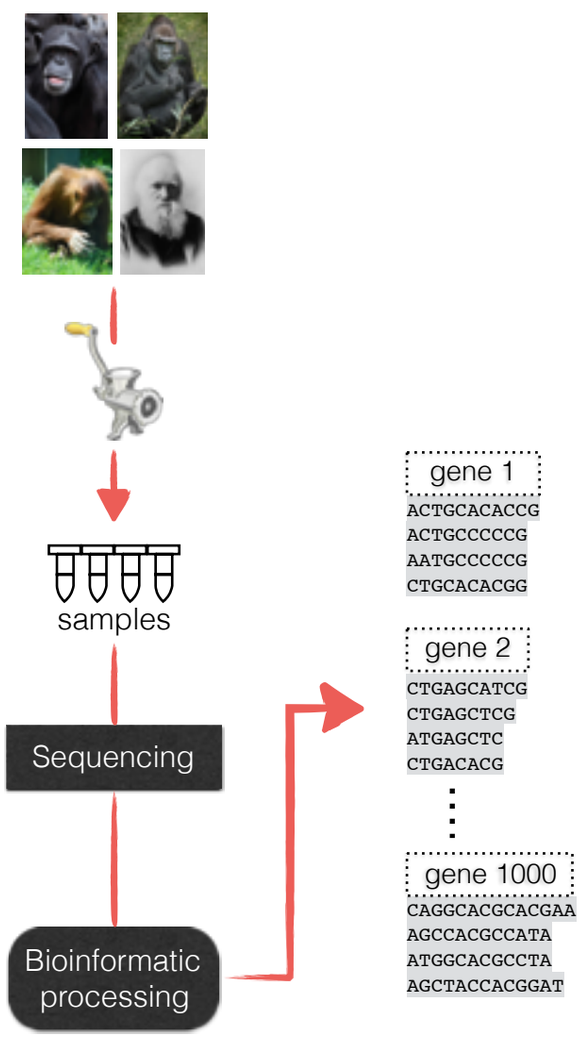
source: http://www.evogeneao.com/



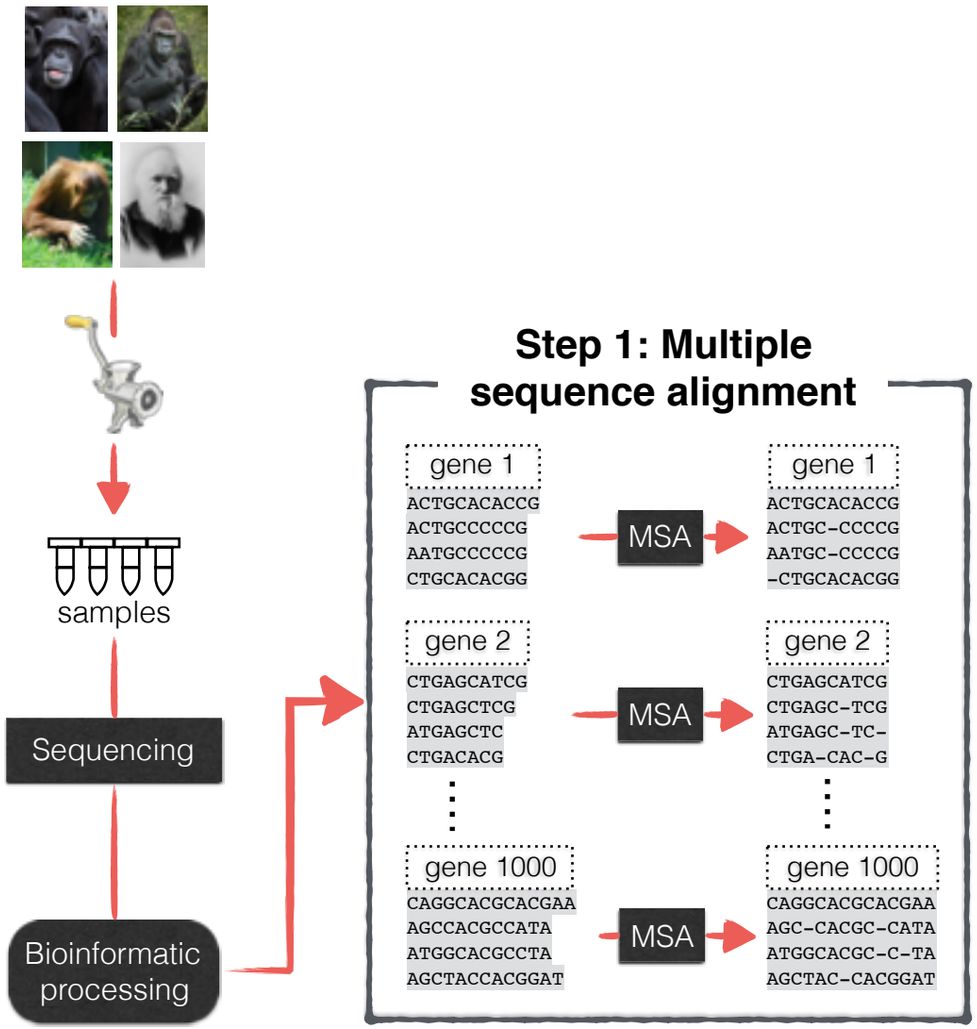
Scalable approaches for phylogenetic analysis of genomes and metagenomes

Siavash Mirarab
ECE

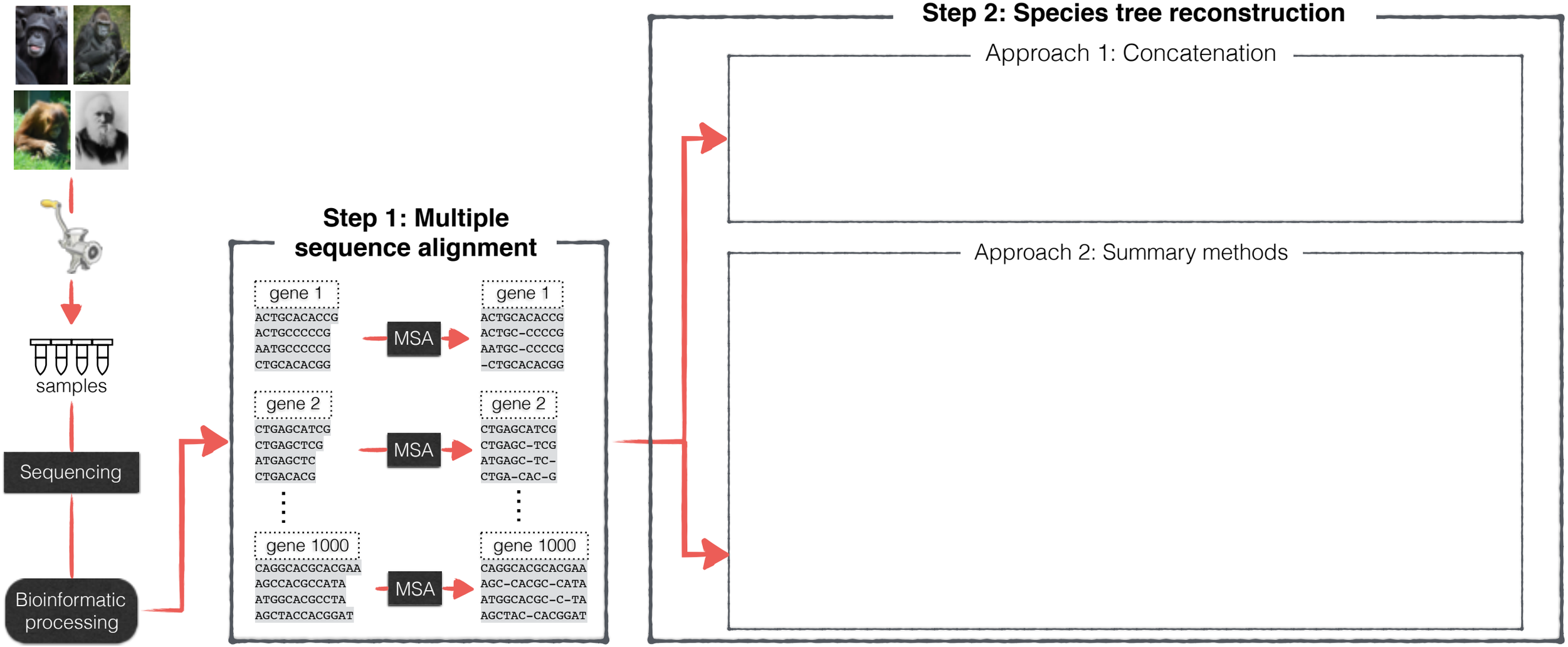
Multi-gene phylogeny reconstruction



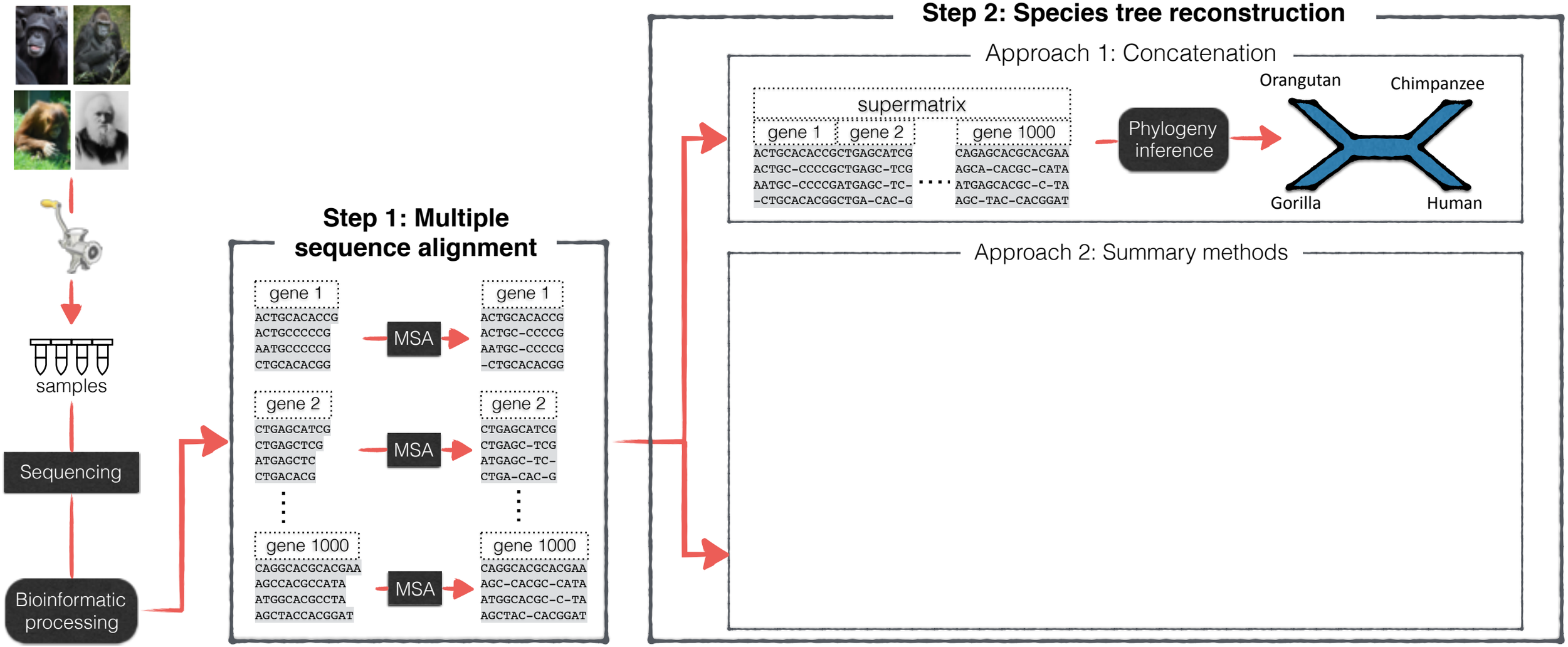
Multi-gene phylogeny reconstruction



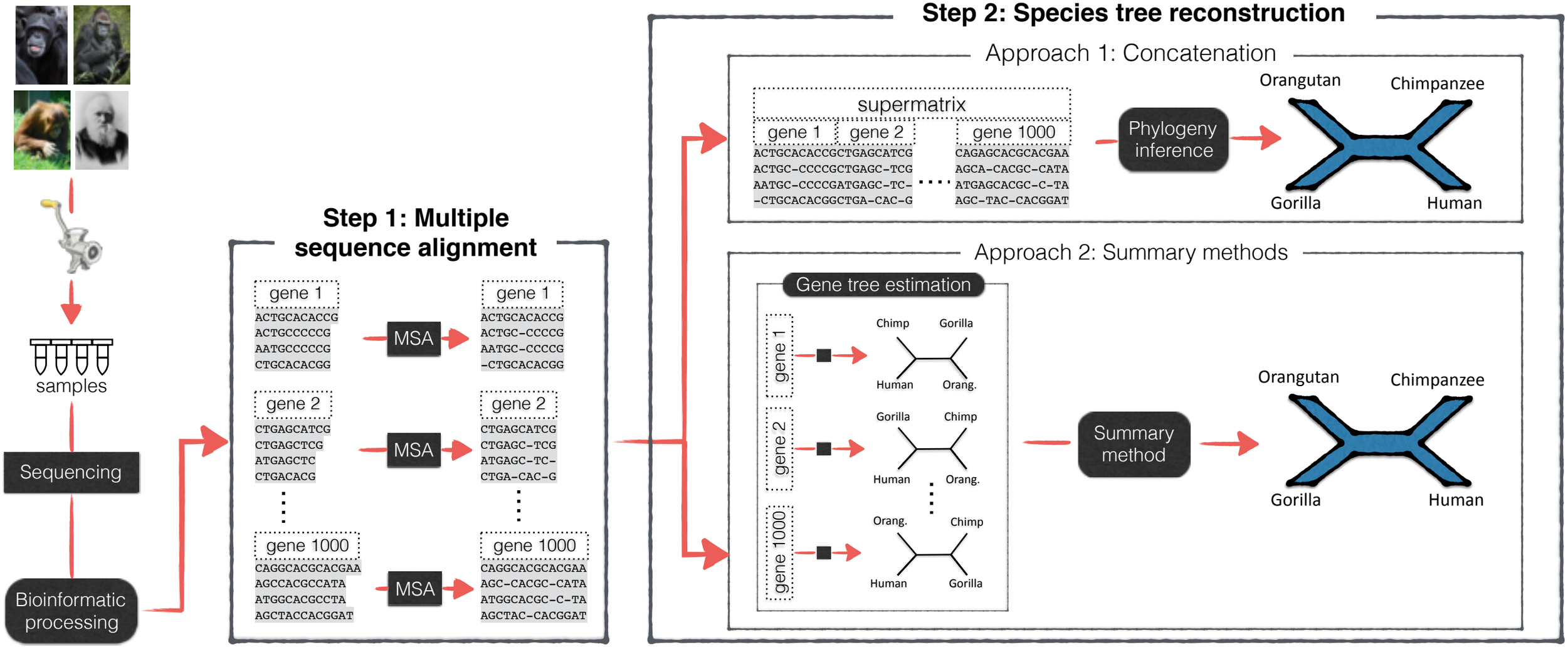
Multi-gene phylogeny reconstruction



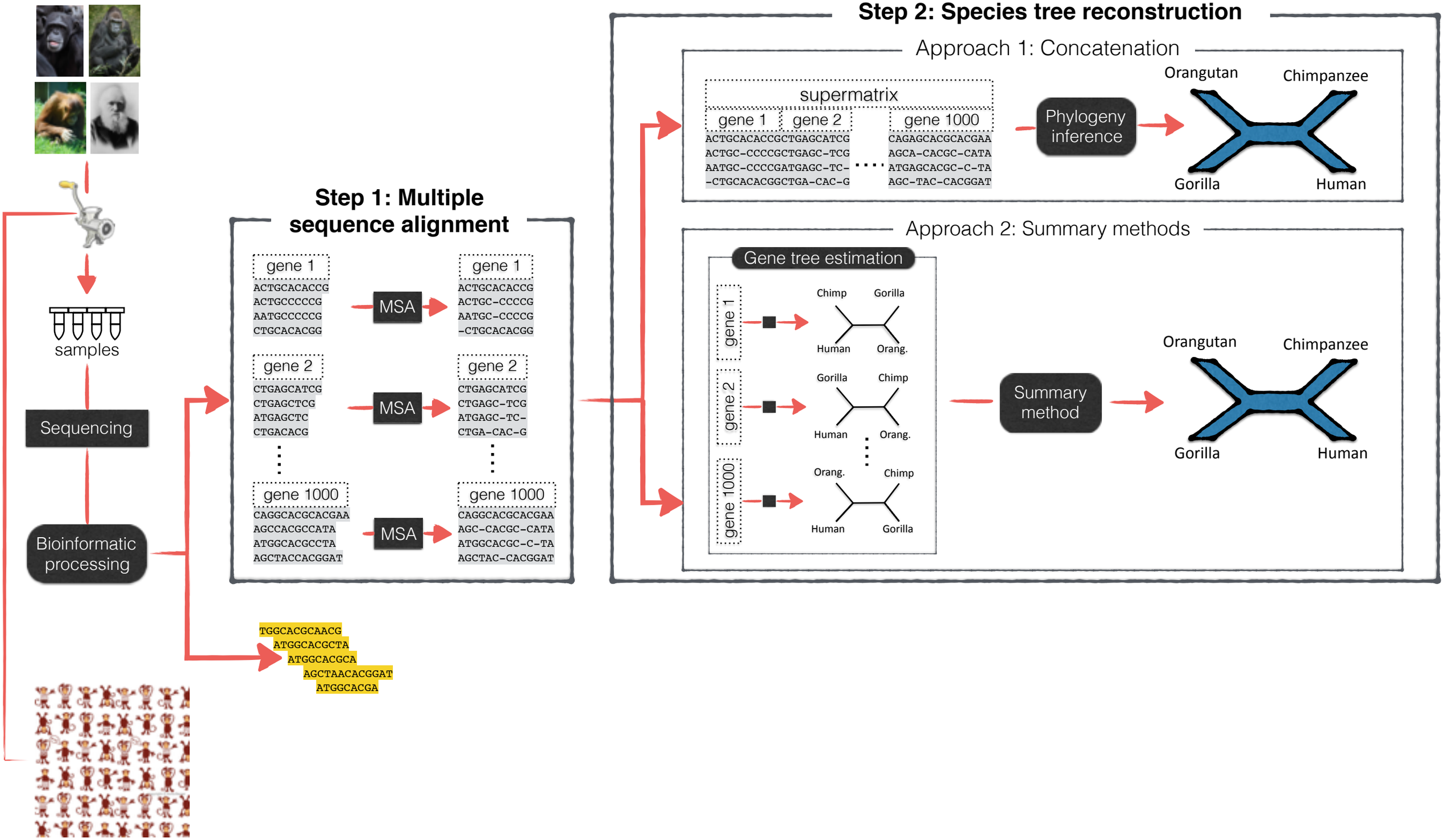
Multi-gene phylogeny reconstruction



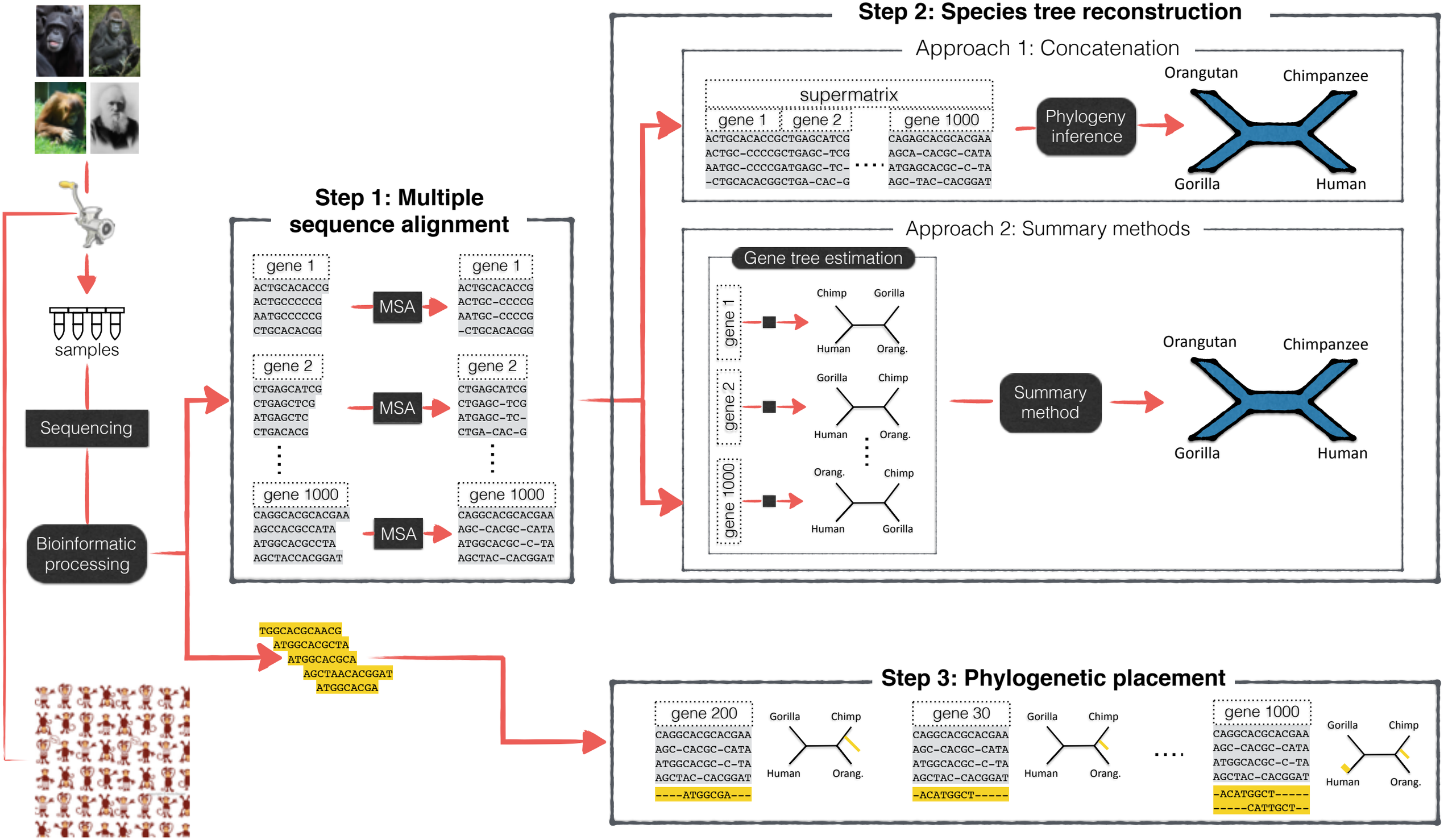
Multi-gene phylogeny reconstruction



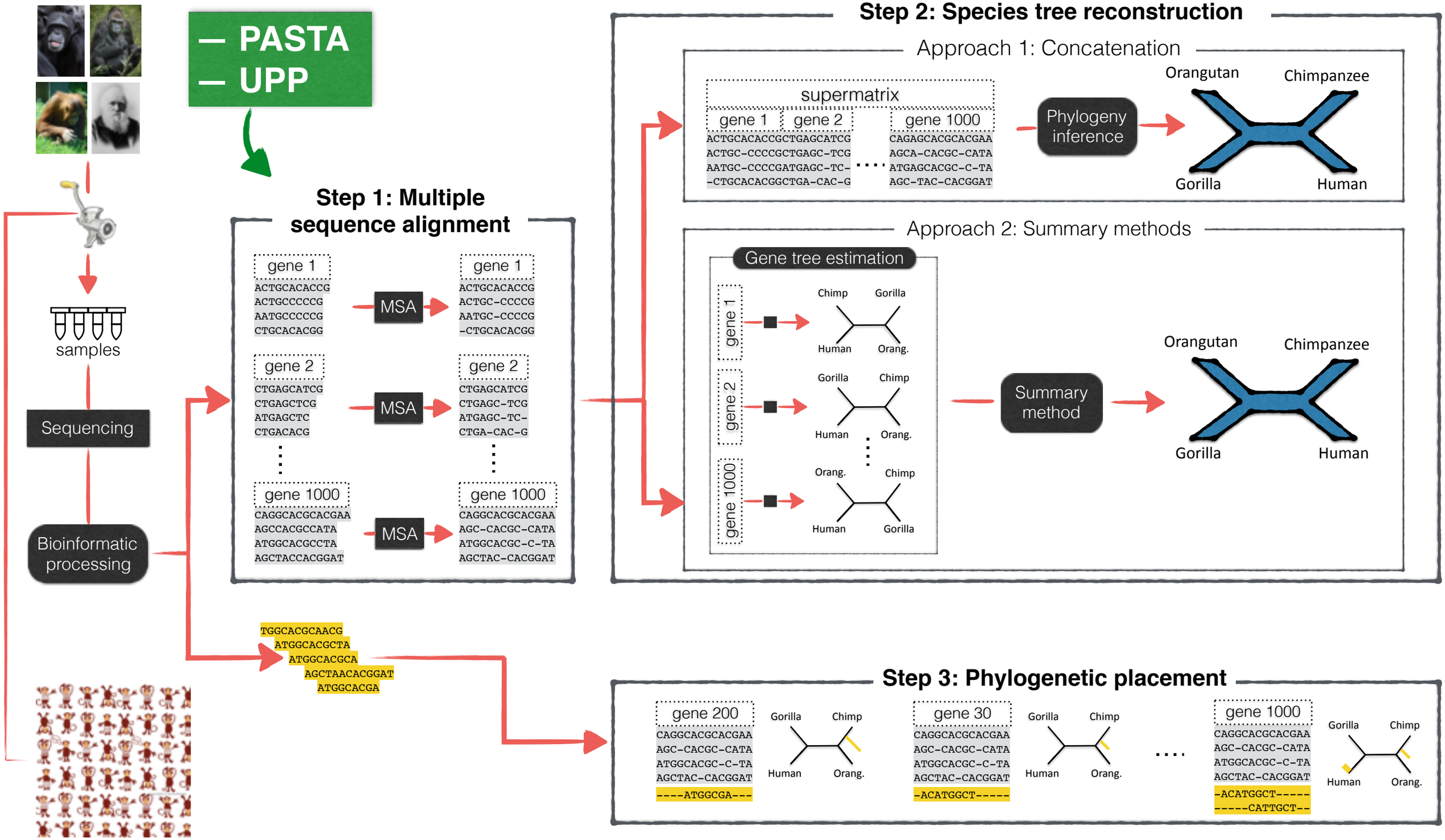
Multi-gene phylogeny reconstruction



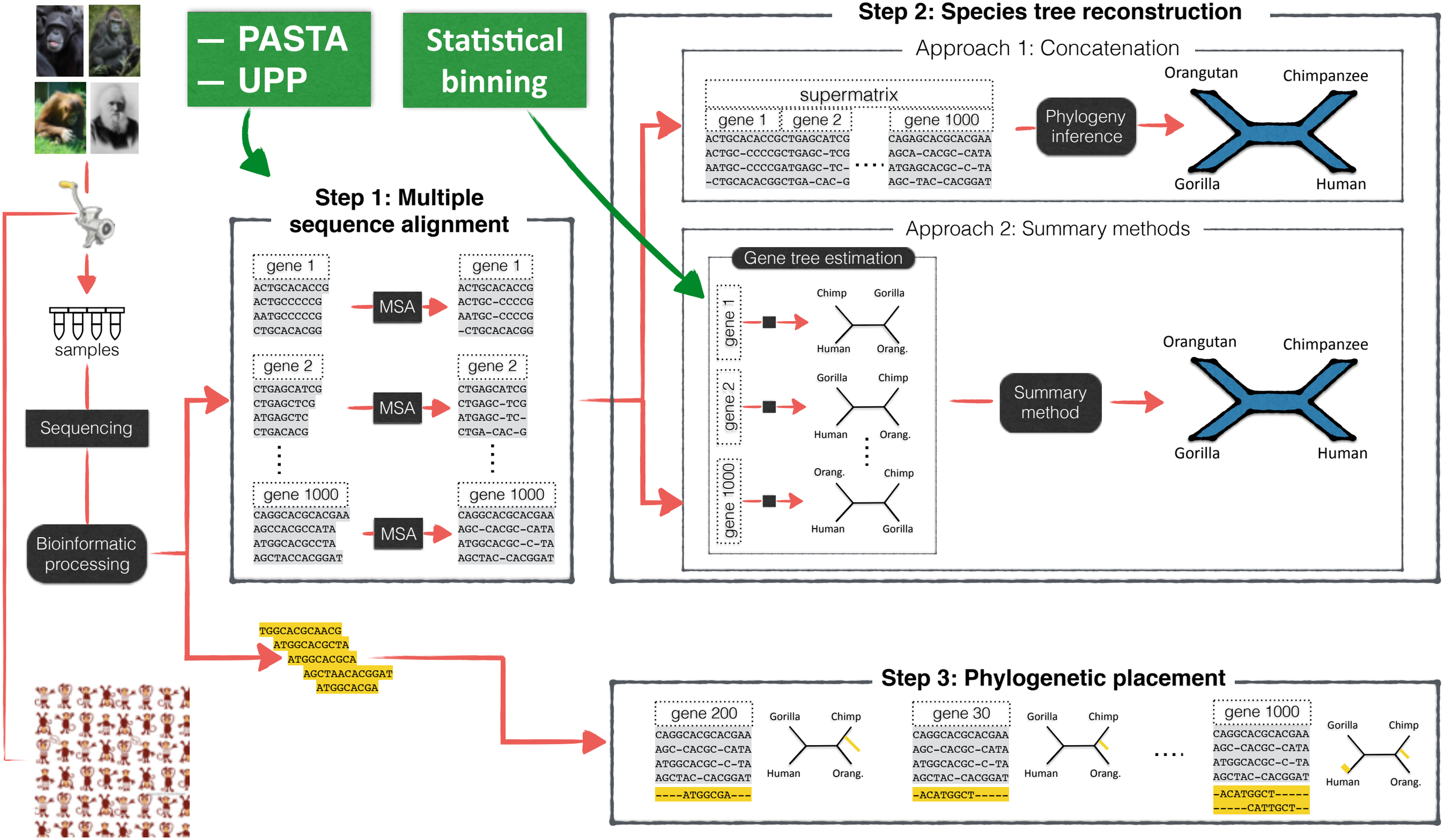
Multi-gene phylogeny reconstruction



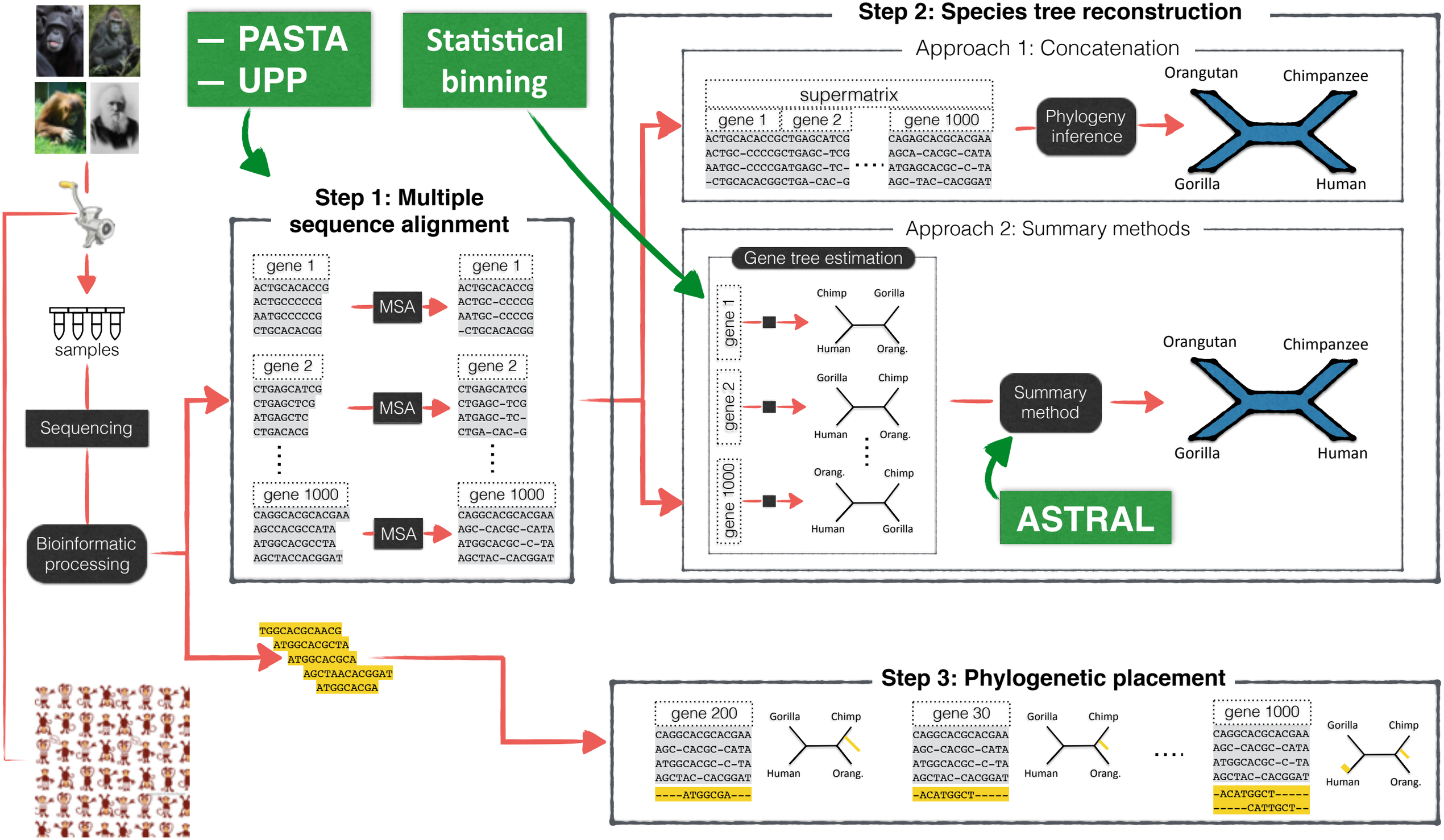
Multi-gene phylogeny reconstruction



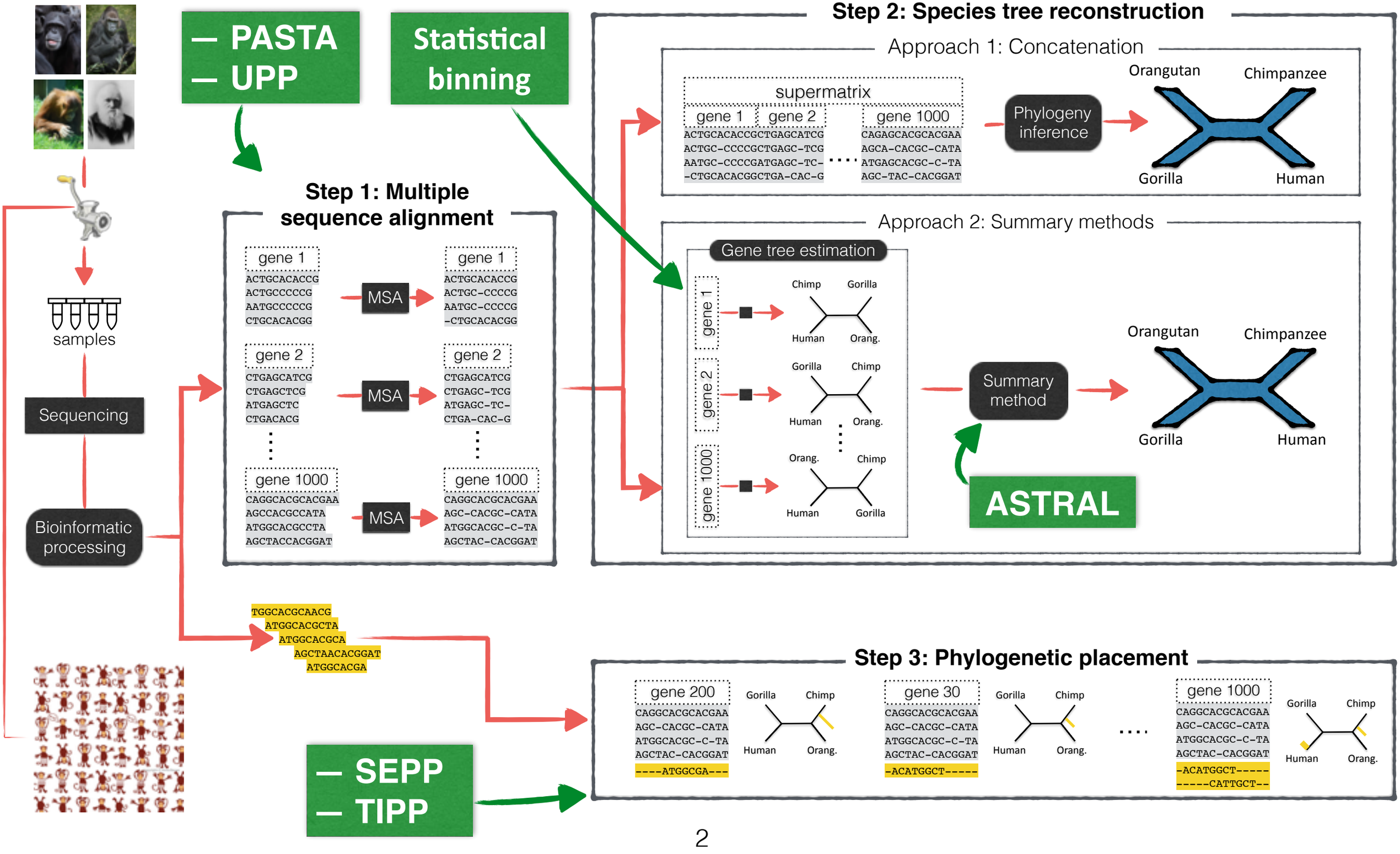
Multi-gene phylogeny reconstruction



Multi-gene phylogeny reconstruction

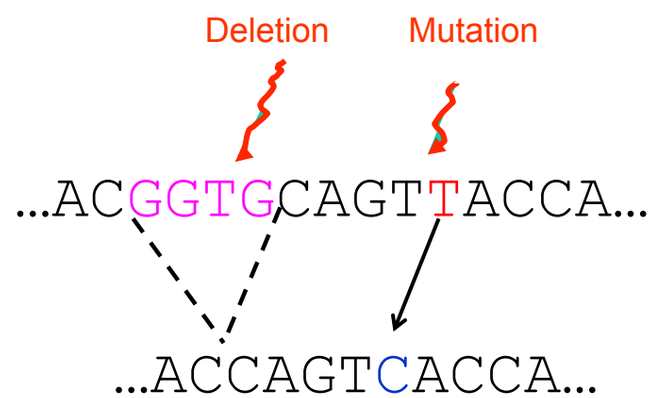


Multi-gene phylogeny reconstruction



Multiple Sequence Alignment (MSA)

A grand challenge in “Frontiers in massive data analysis”, National Academies Press (2013)



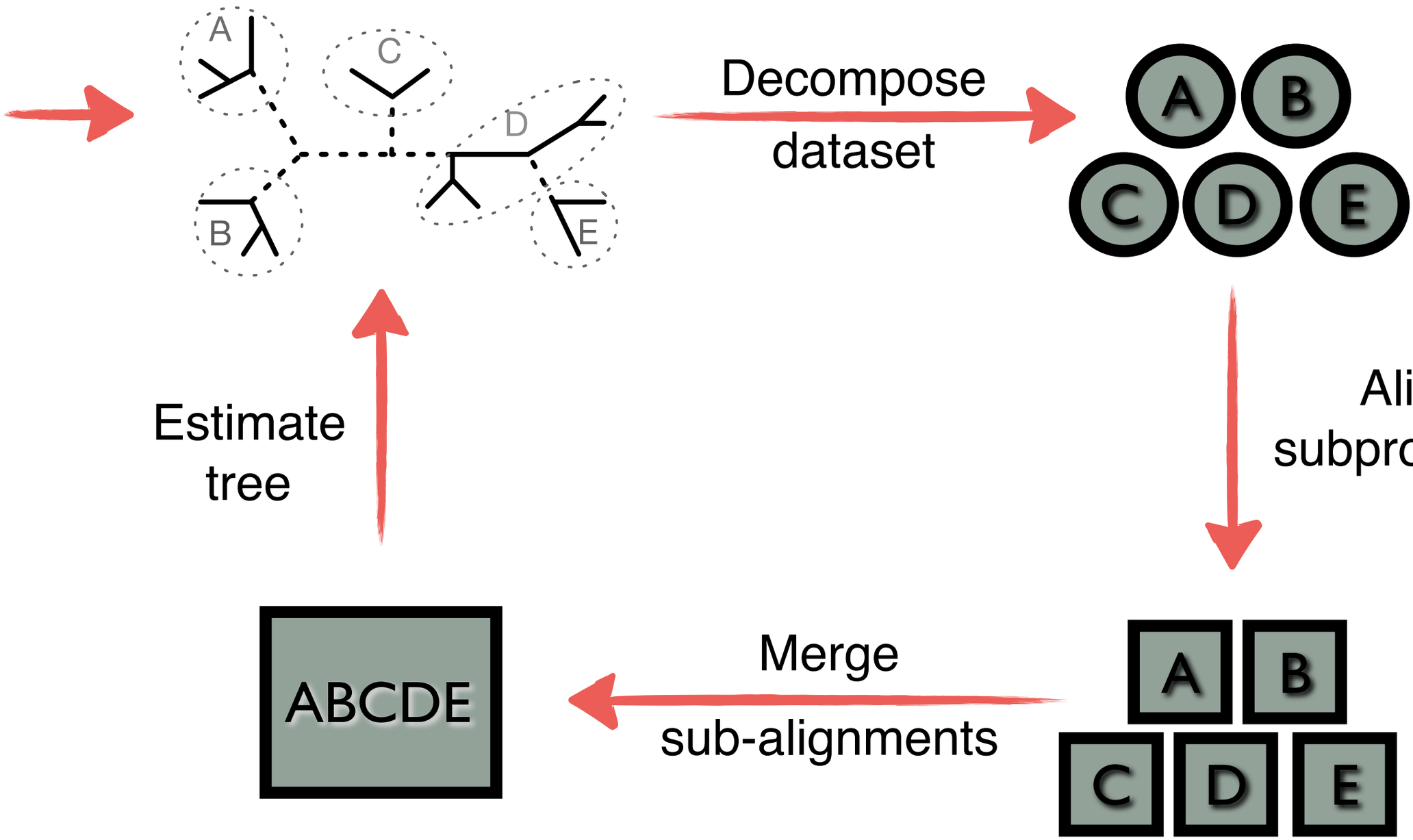
- The accuracy of alignments degrades as the number of sequences increases
- Many alignment methods cannot run on large datasets
- Trouble with mixes of full length and fragmentary data

```
human    = AGGCTATCACCTGACCTACA
chimp    = TAGCTATCACGACCGC
gorilla  = TCGCTGACCTCCA
orang.   = TCACGACCGACA
```

Multiple Sequence Alignment (MSA)

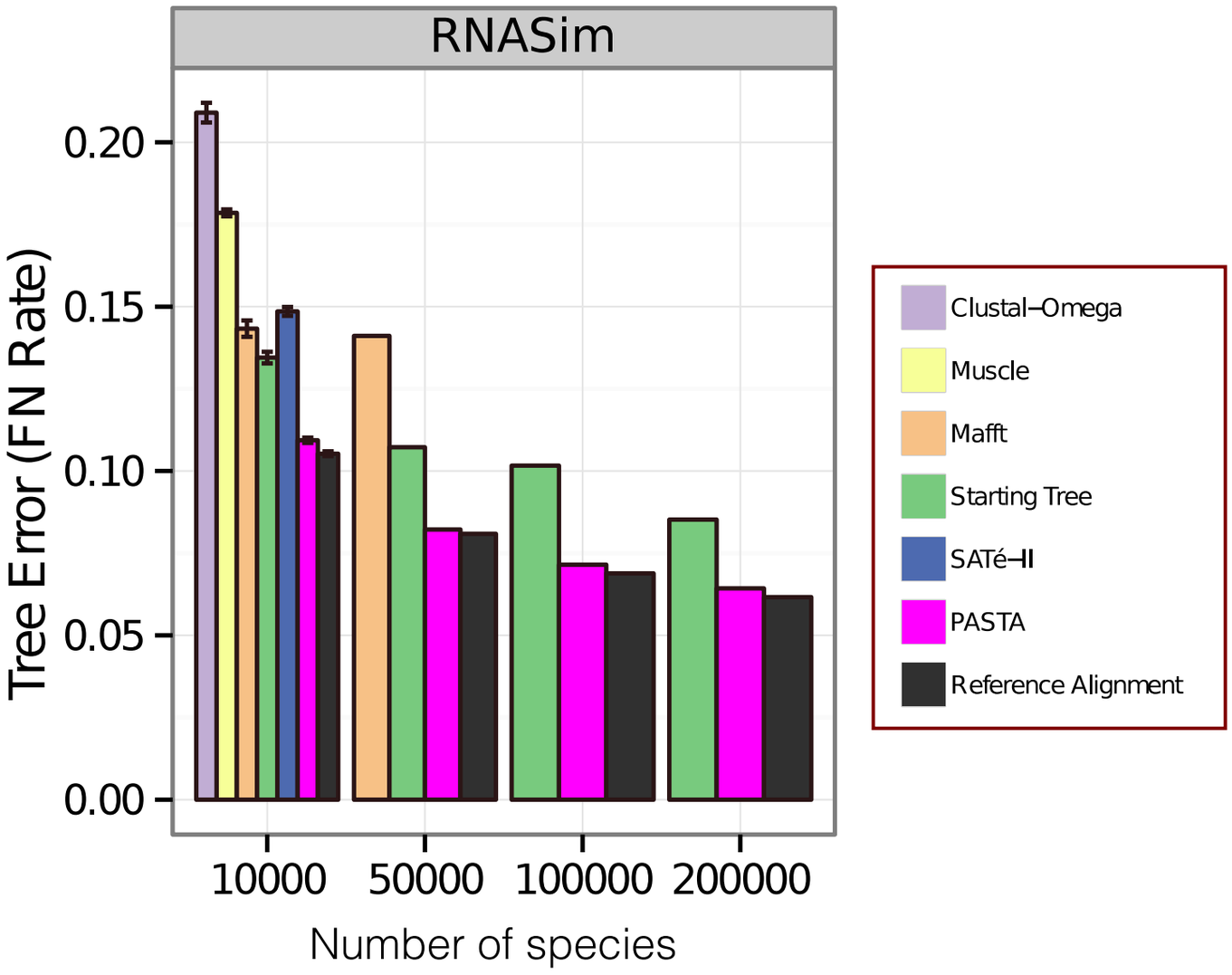
```
human    = -AGGCTATCACCTGACCTACA
chimp    = TAG-CTATCAC--GACCGC--
gorilla  = TCG-CT-------GACCTCCA
orang.   = -------TCAC--GACCGACA
```

PASTA: Iterative divide-and-conquer alignment and tree estimation



S. Mirarab et al., Res. Comput. Mol. Biol. (2014).
S. Mirarab et al., J. Comput. Biol. 22 (2015).

Tree topological accuracy

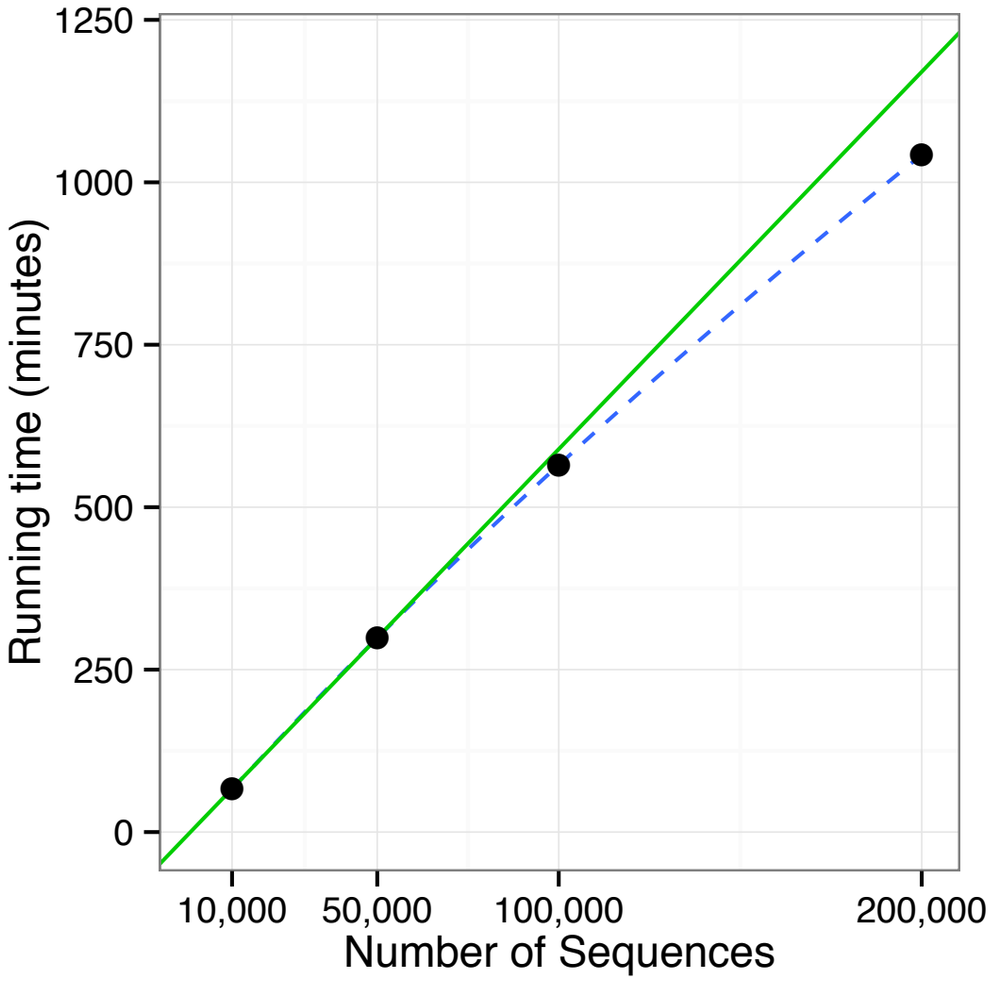


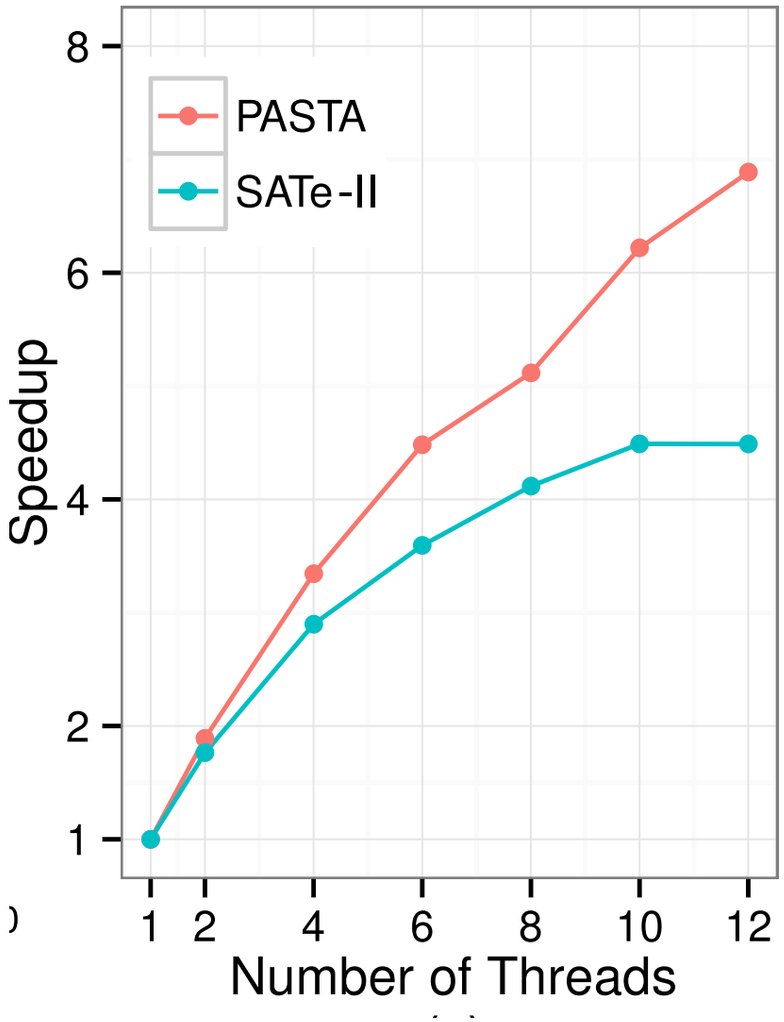
1 million sequences:

- PASTA finished one iteration in 15 days
- PASTA tree had 6% error, compared to 5.6% when using true alignment
- Starting tree had 8.4% error

S. Mirarab et al., Res. Comput. Mol. Biol. (2014).
S. Mirarab et al., J. Comput. Biol. 22 (2015).

Scalability of PASTA





Ongoing collaborations

Daniel McDonald and Uyen Mai

- Testing performance of PASTA for building green genes 16S reference tree
  - Q1: Ability to distinguish samples using unifrac?

	unweighted		weighted	
	GG	PASTA	GG	PASTA
88 soils	0.78	0.78	0.75	0.74
infant-time-series	0.55	0.55	0.37	0.42
moving pictures	728	724	2188	2439
global gut	52.9	51.1	79	72

  - Q2: Taxonomical consistency?
  - Q3: Speed? on Gordon (16 cores) 97% tree (99,322 leaves): 28 hours; 99% tree (203,452 leaves): 49 hours

Ongoing collaborations

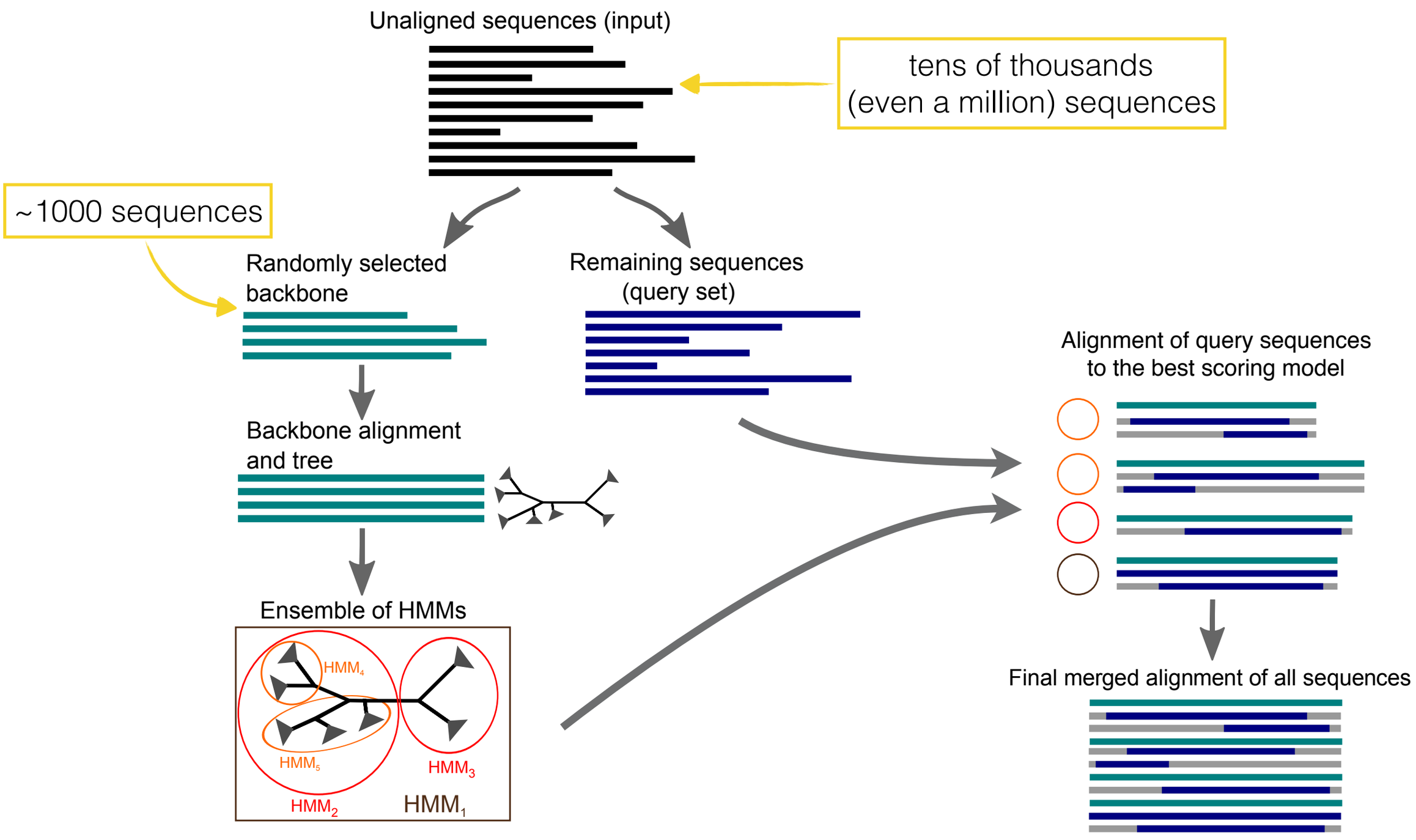
Daniel McDonald and Uyen Mai

- Testing performance of PASTA for building green genes 16S reference tree
  - Q1: Ability to distinguish samples using unifrac?

	unweighted		weighted	
	GG	PASTA	GG	PASTA
88 soils	0.78	0.78	0.75	0.74
infant-time-series	0.55	0.55	0.37	0.42
moving pictures	728	724	2188	2439
global gut	52.9	51.1	79	72

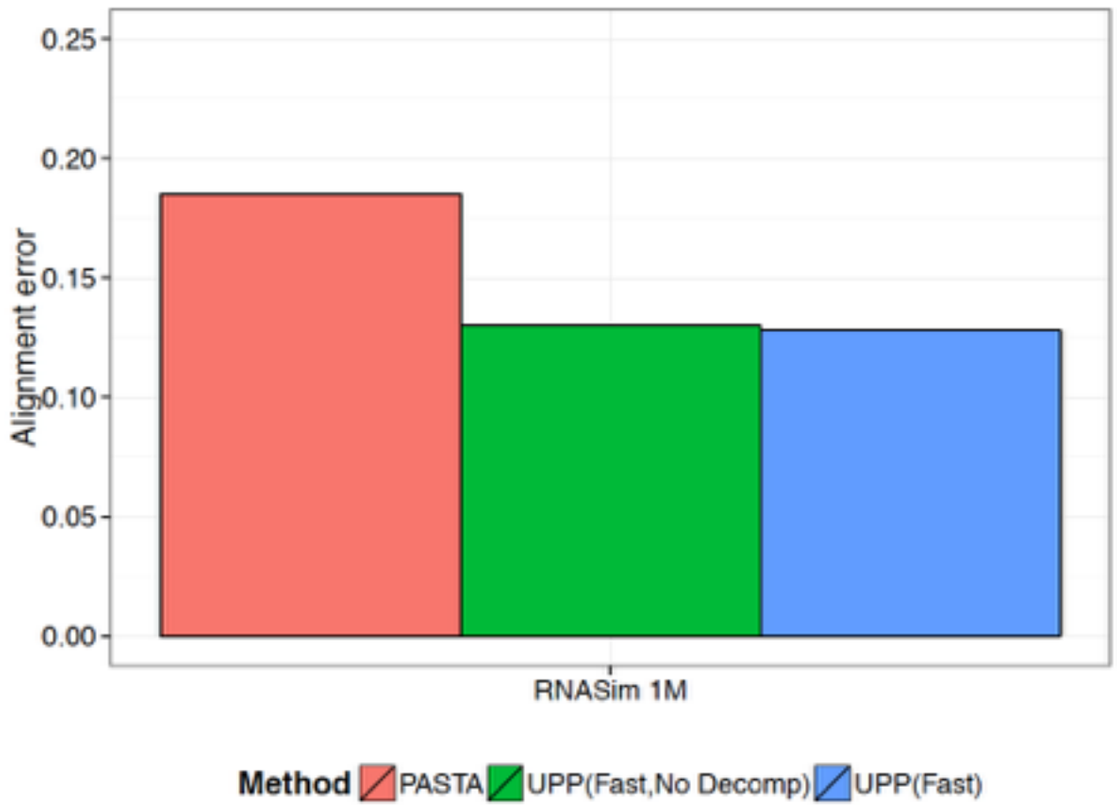
  - Q2: Taxonomical consistency?
  - Q3: Speed? on Gordon (16 cores) 97% tree (99,322 leaves): 28 hours; 99% tree (203,452 leaves): 49 hours
- Can a PASTA-like merge algorithm be used for merging SSU alignments from the three domains of life?

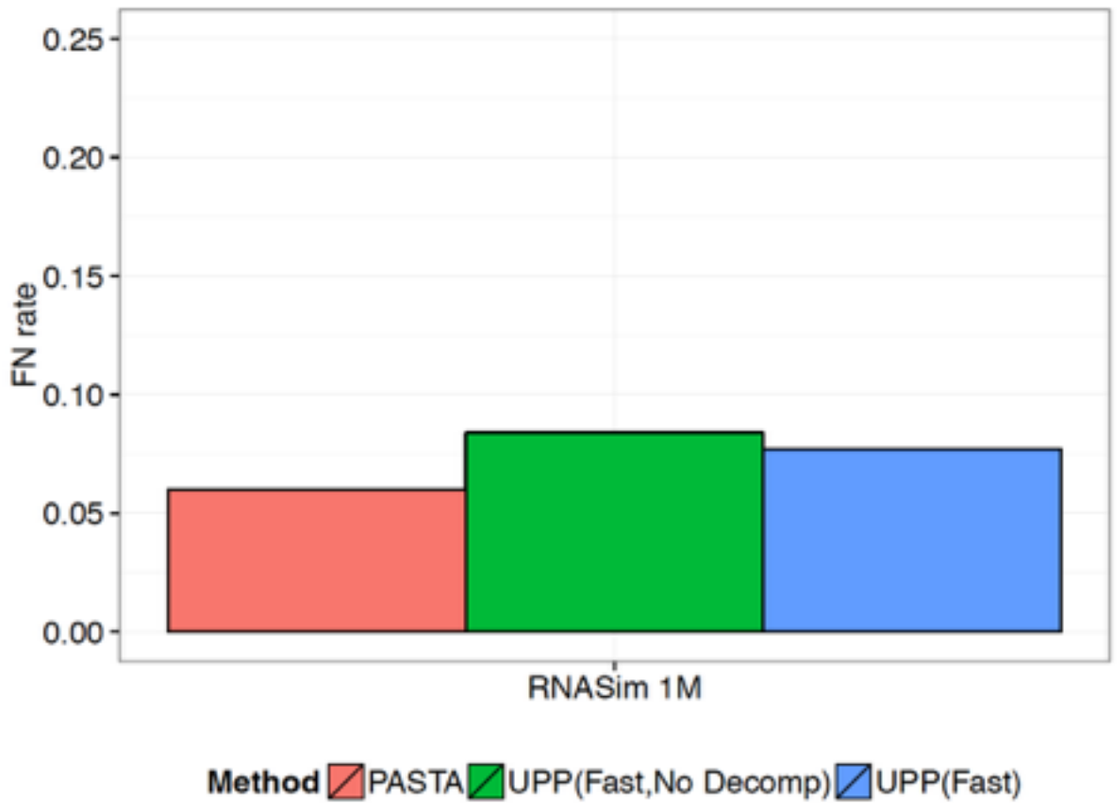
UPP: ensembles of HMMs



N. Nguyen et al., Genome Biol. 16 (2015)

RNASim Million Sequences: alignment error

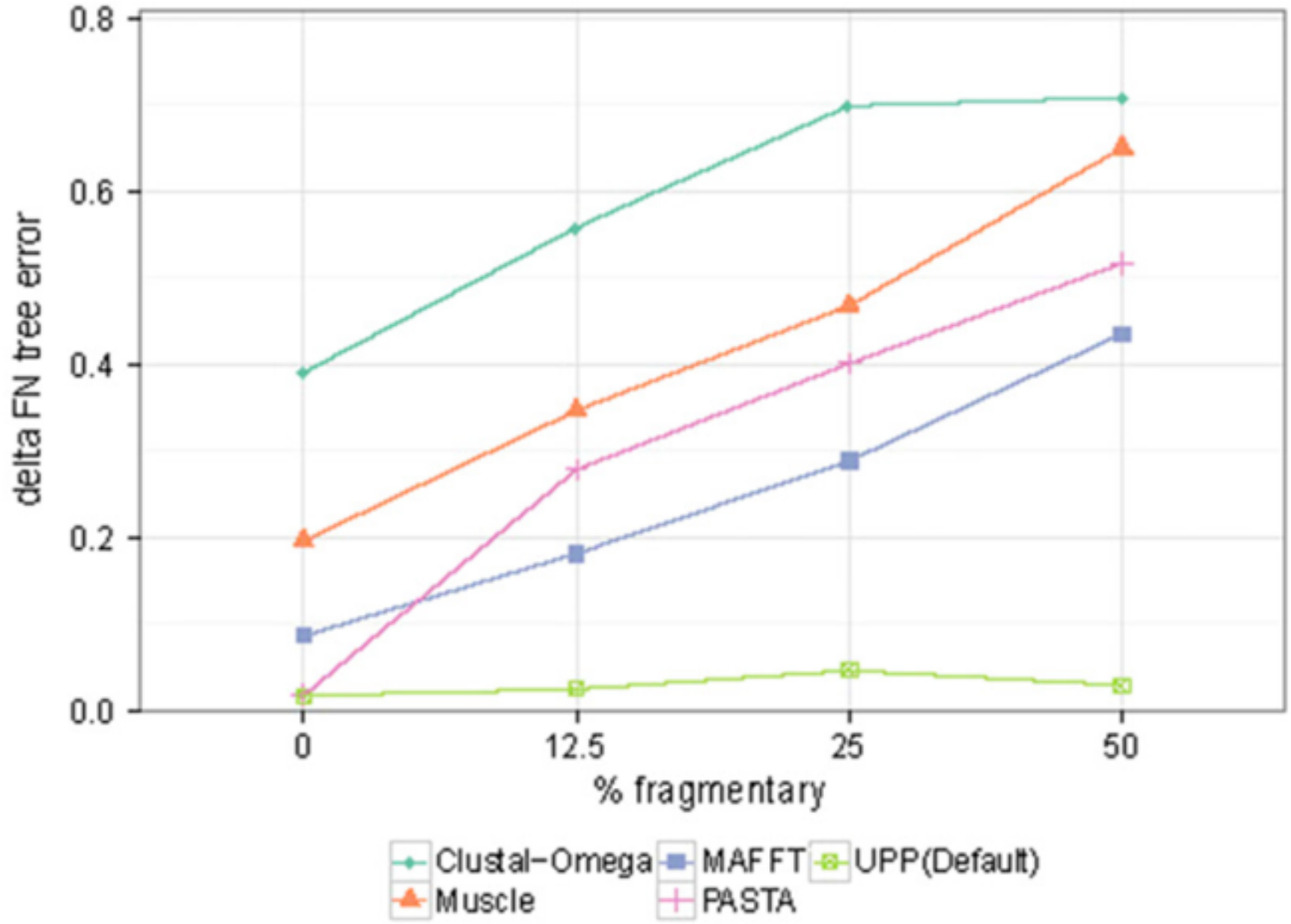




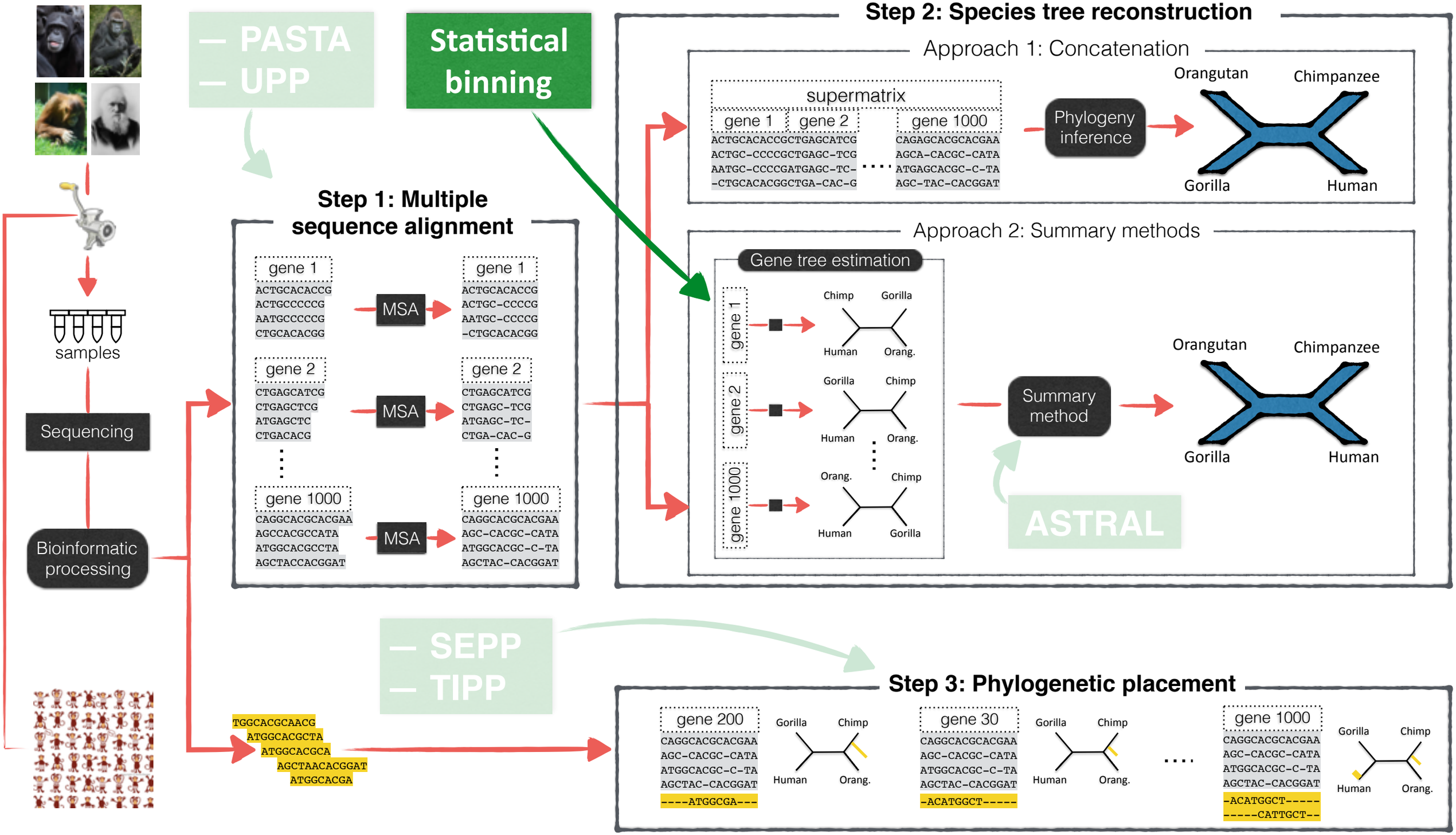
Running time (12 processors):

- UPP(Fast,NoDecomp) took 2.2 days
- UPP(Fast) took 11.9 days
- PASTA took 15 days

Fragmentary sequences



Multi-gene phylogeny reconstruction



Large-scale phylogenomics

1K Plants (1KP) [PNAS, 2014]

PNAS

Phylotranscriptomic analysis of the origin and early diversification of land plants

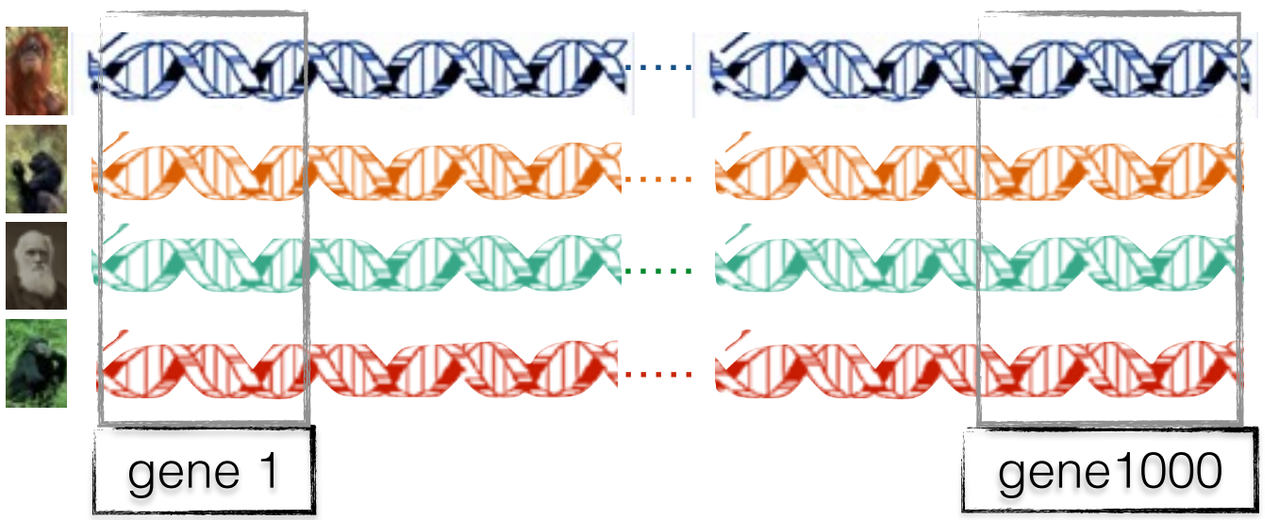
Norman J. Wickett[a,b,1,2], Siavash Mirarab[c,1], Nam Nguyen[c], Tandy Warnow[c], Eric Carpenter[d], Naim Matasci[e,f], Saravanaraj Ayyampalayam[g], Michael S. Barker[f], J. Gordon Burleigh[h], Matthew A. Gitzendanner[h,i], Brad R. Ruhfel[h,j,k], Eric Wafula[l], Joshua P. Der[l], Sean W. Graham[m], Sarah 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]

Avian phylogenomics [Science, 2014]

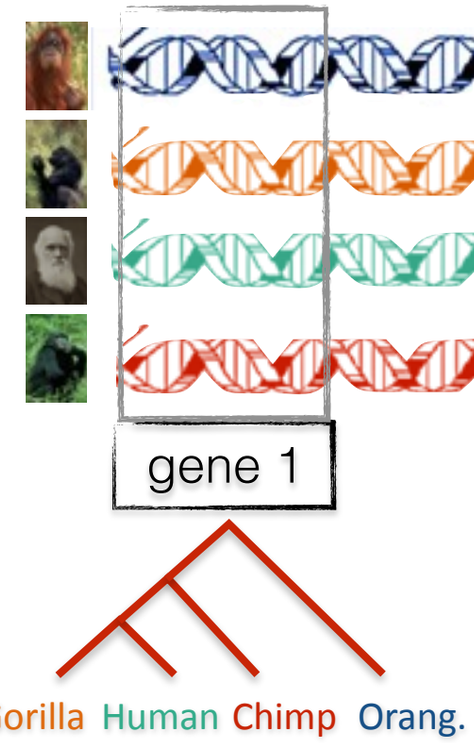
Whole-genome analyses resolve early branches in the tree of life of modern birds

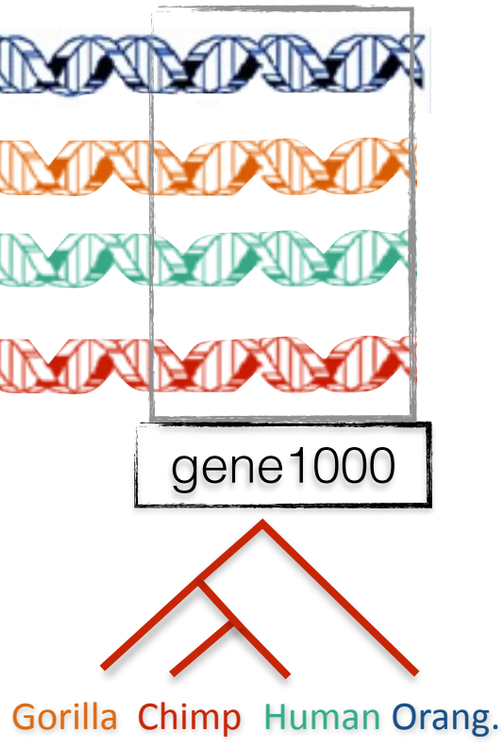
Erich D. Jarvis,[1*†] Siavash Mirarab,[2*] Andre J. Aberer,[3] Bo Li,[4,5,6] Peter Houde,[7] Cai Li,[4,6] Simon Y. W. Ho,[8] Brant C. Faircloth,[9,10] Benoit Nabholz,[11] Jason T. Howard,[1] Alexander Suh,[12] Claudia C. Weber,[12] Rute R. da Fonseca,[6] Jianwen Li,[4] Fang Zhang,[4] Hui Li,[4] Long 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†]

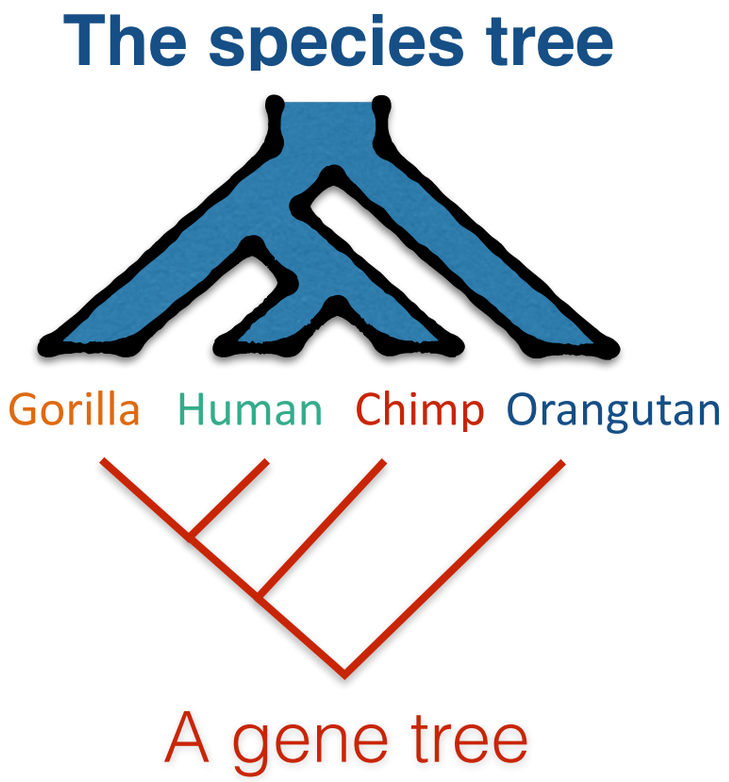
Gene tree discordance



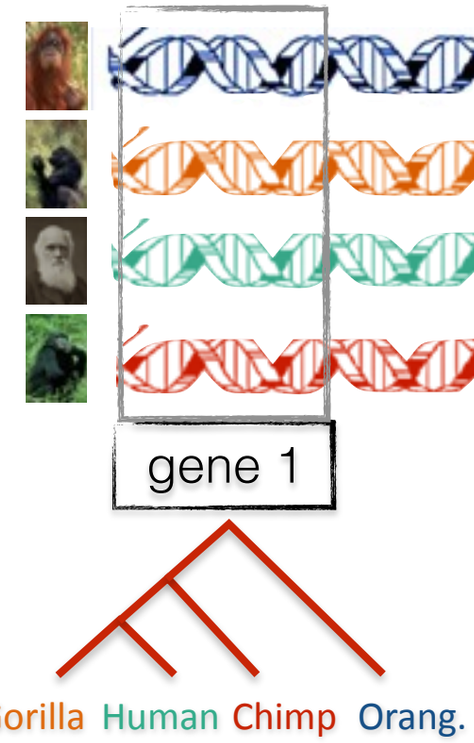
Gene tree discordance

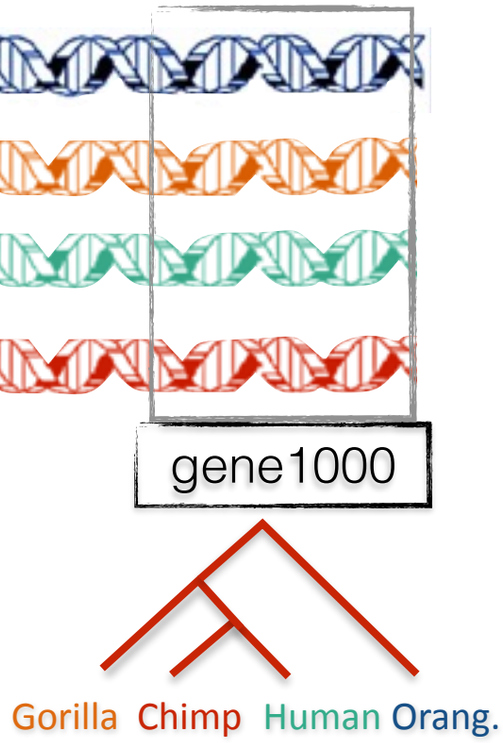


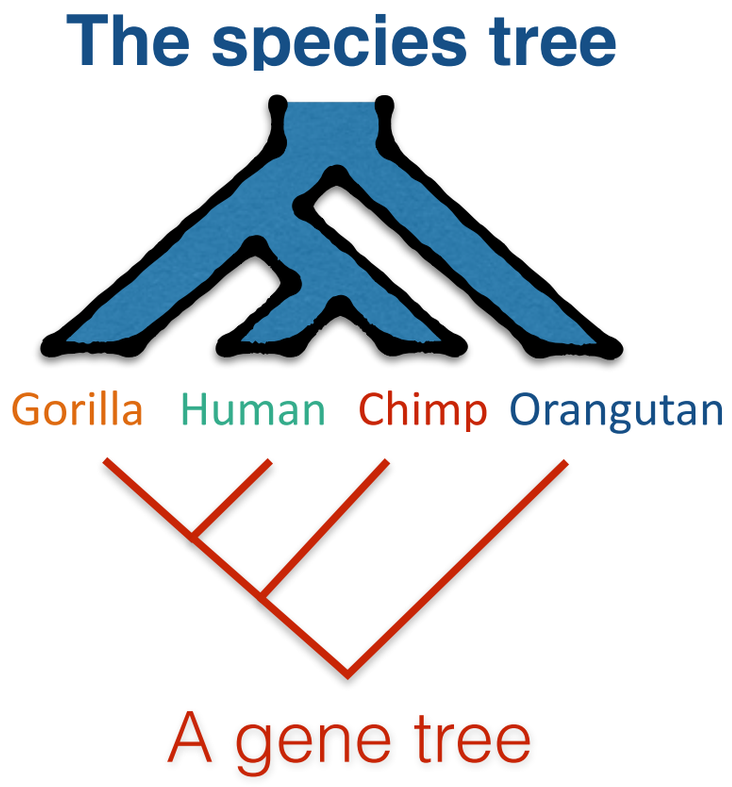




Gene tree discordance





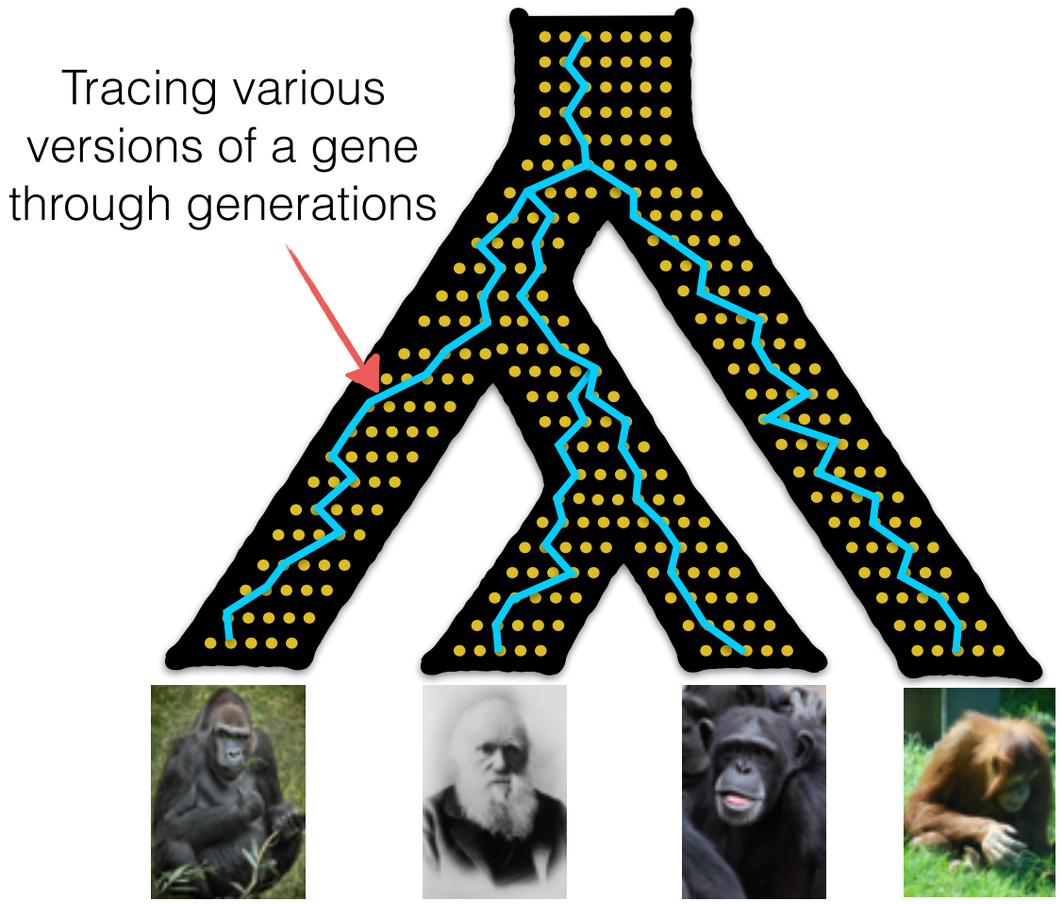


Causes of gene tree discordance include:

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

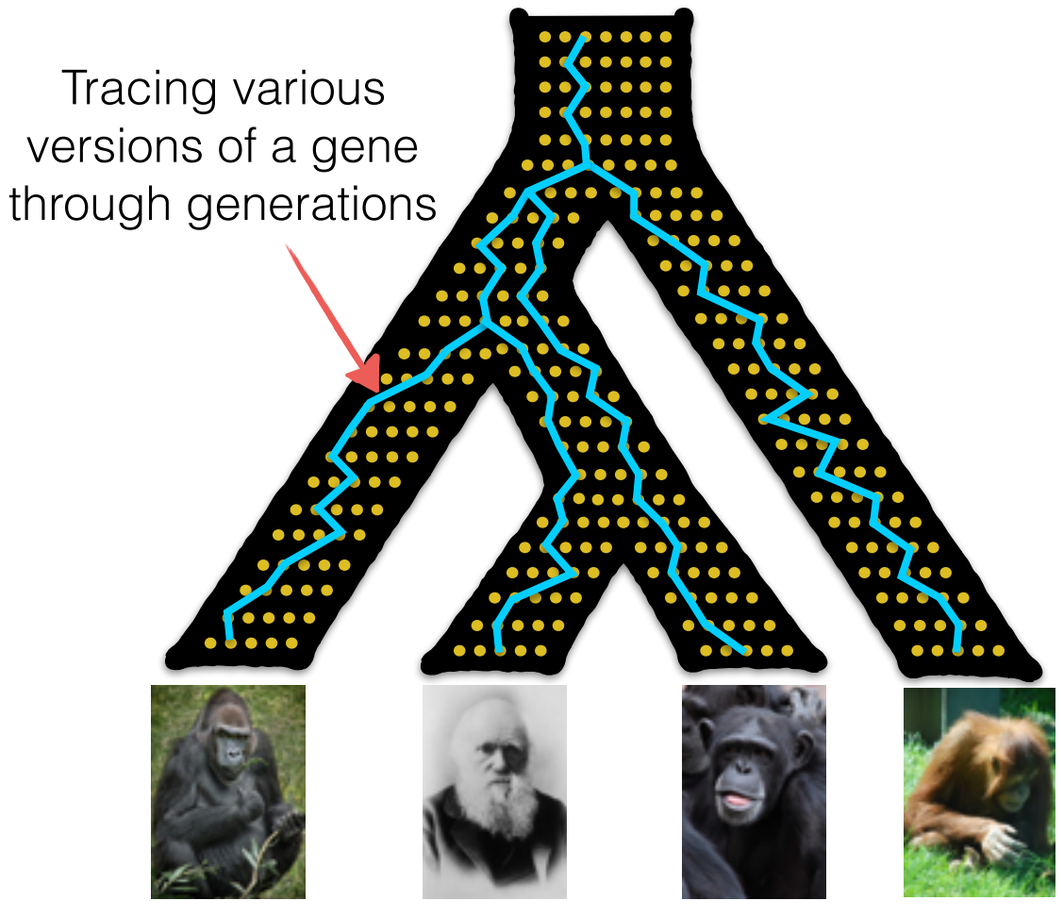
Incomplete Lineage Sorting (ILS)

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



Incomplete Lineage Sorting (ILS)

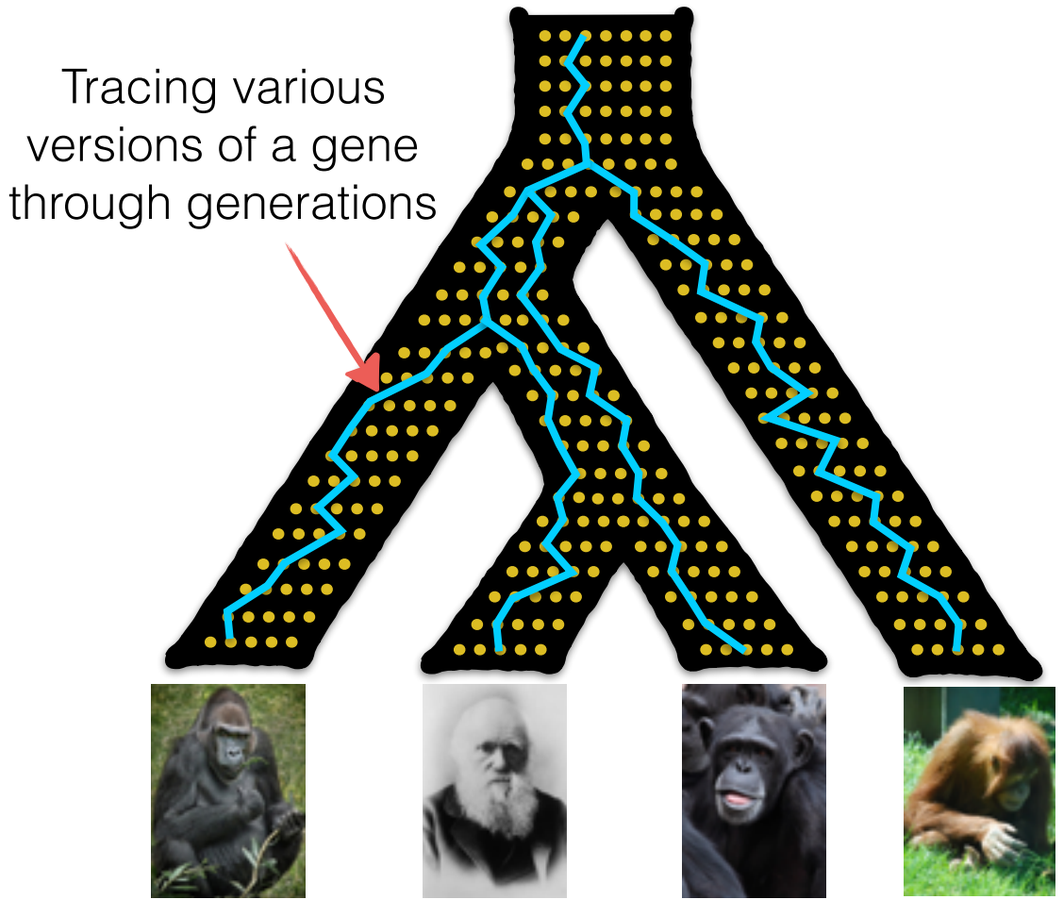
- A random process related to having multiple versions of each gene in a population



Incomplete Lineage Sorting (ILS)

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

- Omnipresent and a major cause of discordance (esp. for short branches).



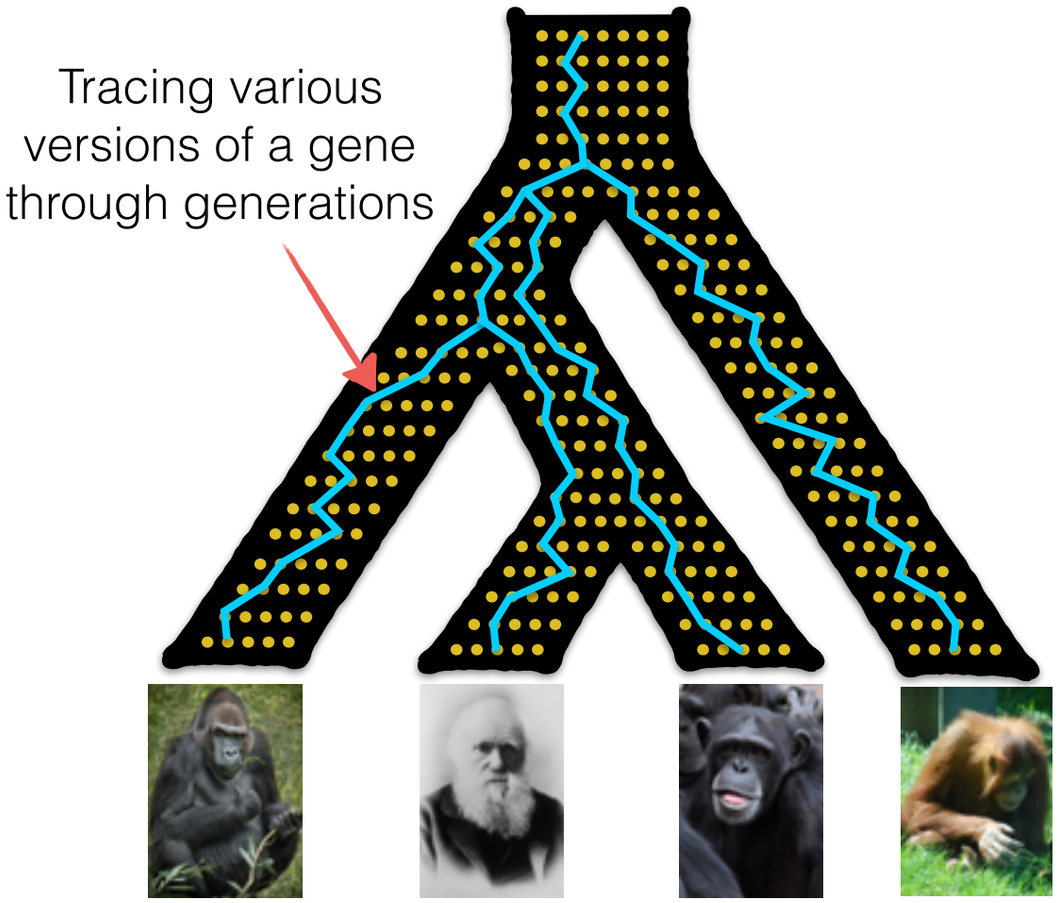
Incomplete Lineage Sorting (ILS)

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

- Omnipresent and a major cause of discordance (esp. for short branches).

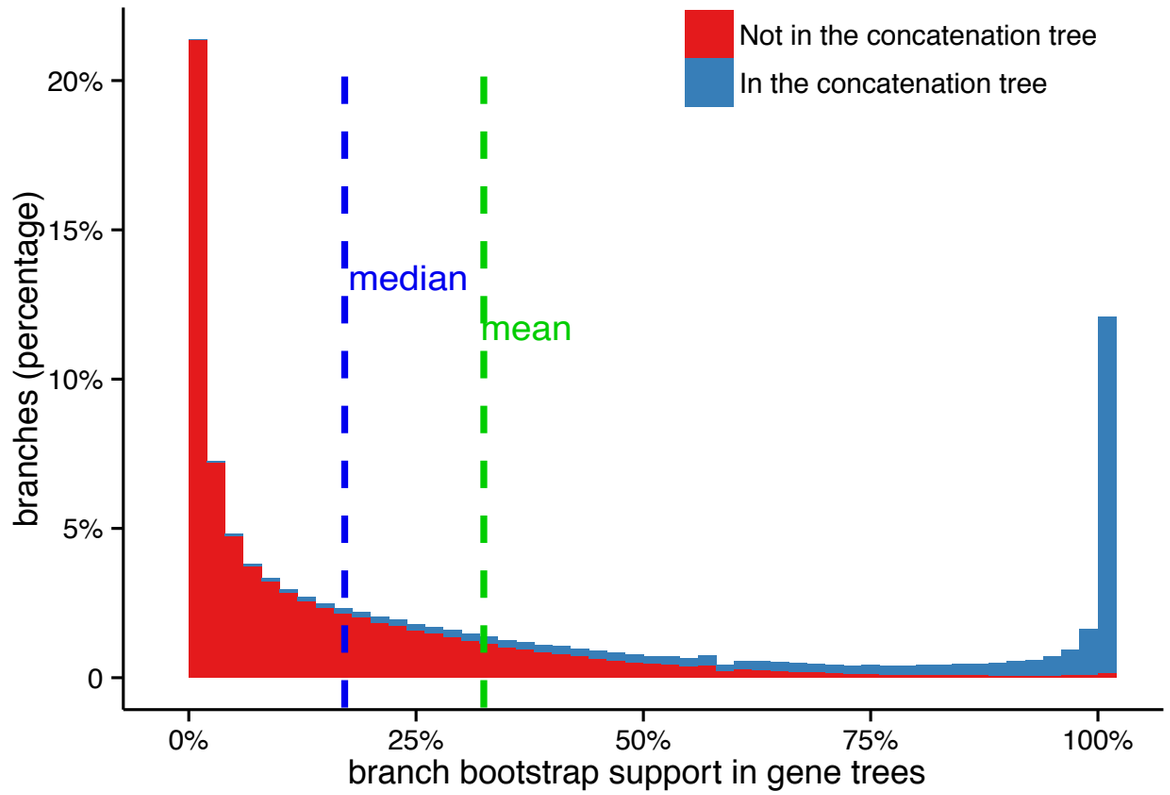
- We have statistical models of ILS (multi-species coalescent)

    - 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]



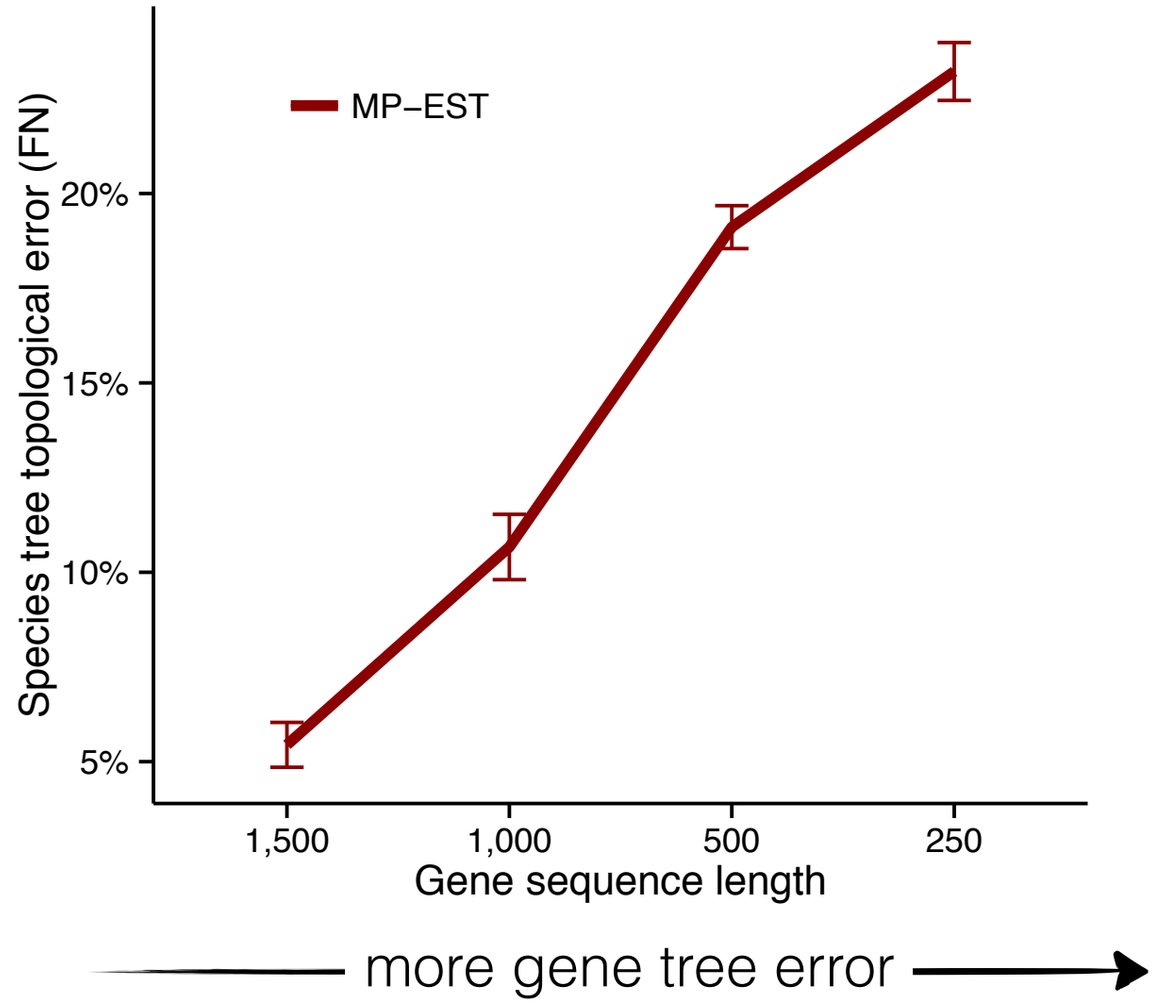
Avian whole genomes analyses

- Whole genomes for 48 bird species (~100 million years of evolution)
- Goal: a phylogeny of major bird lineages
- Extremely challenging due to rampant gene tree incongruence
- 14,000 “noisy” genes with high levels of gene tree estimation error



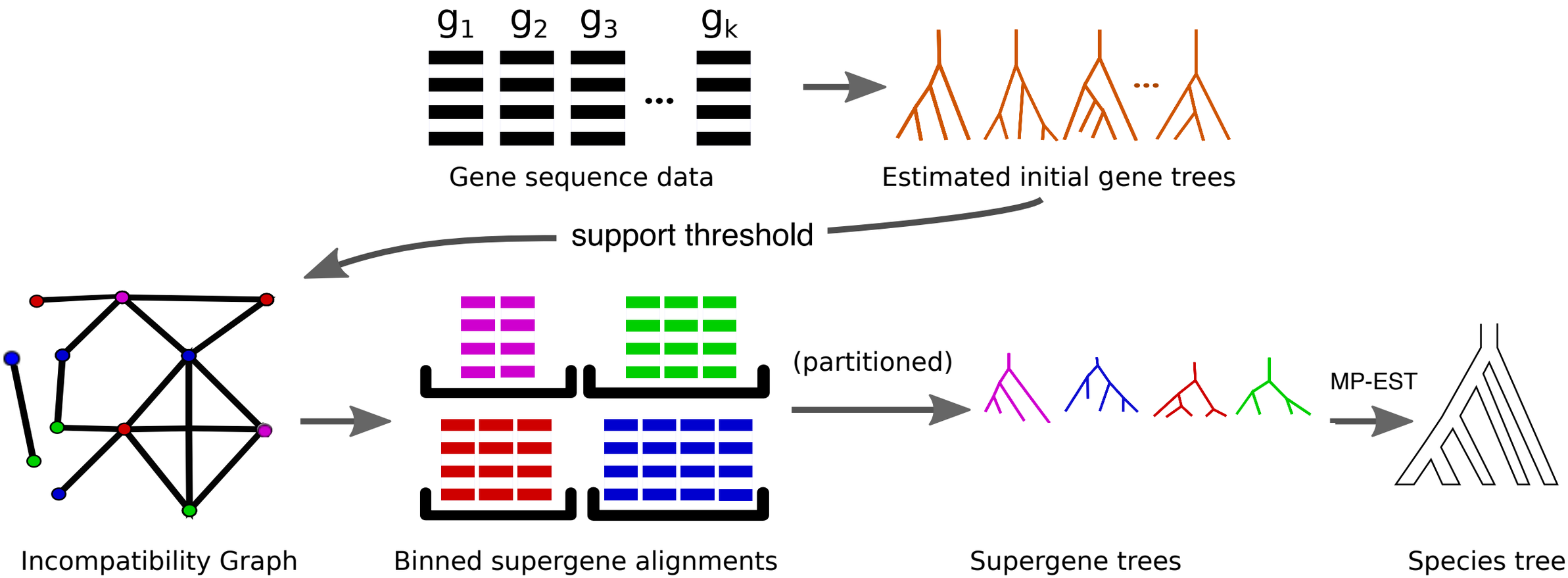
S. Mirarab et al., Science (80). 346 (2014).
E. D. Jarvis et al., Science (80). 346 (2014).

Gene tree estimation error impacts species tree estimation



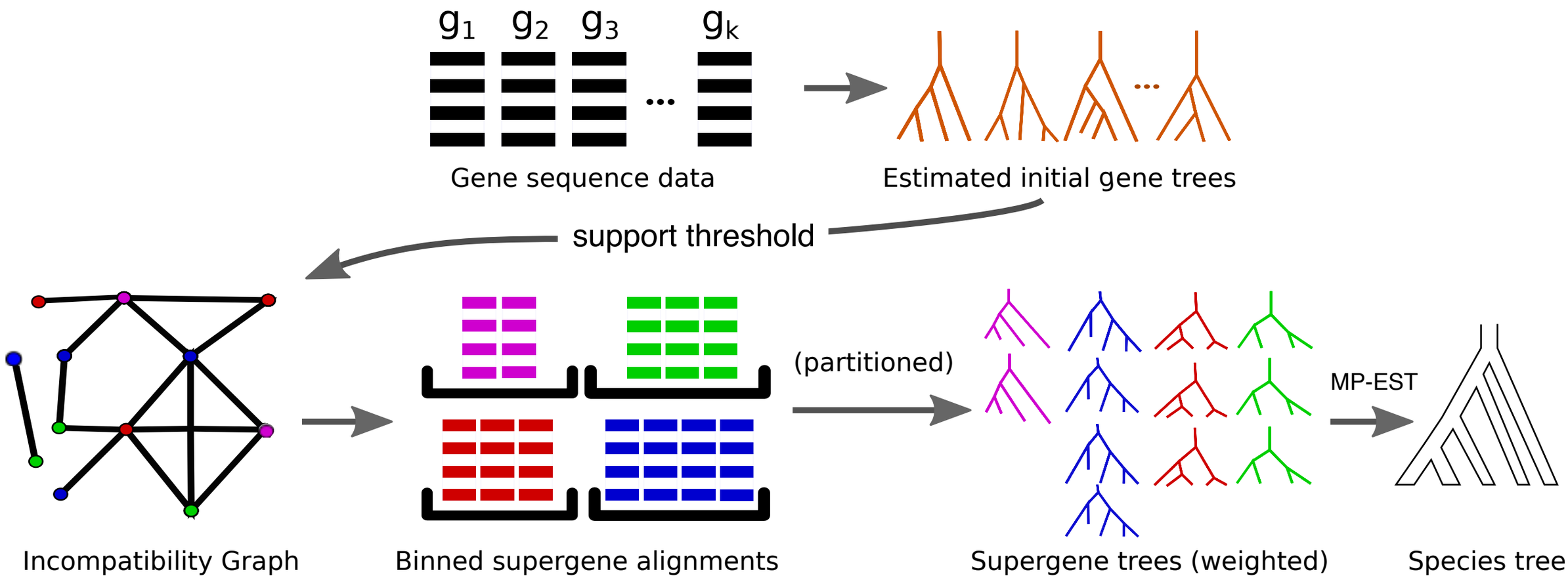
Simulations with 45 avian-like species, 1000 genes
[S. Mirarab et al., Science (80-.). 346 (2014)]

Statistical binning: overview



Original version: unweighted [S. Mirarab et al., Science (80). 346 (2014)]

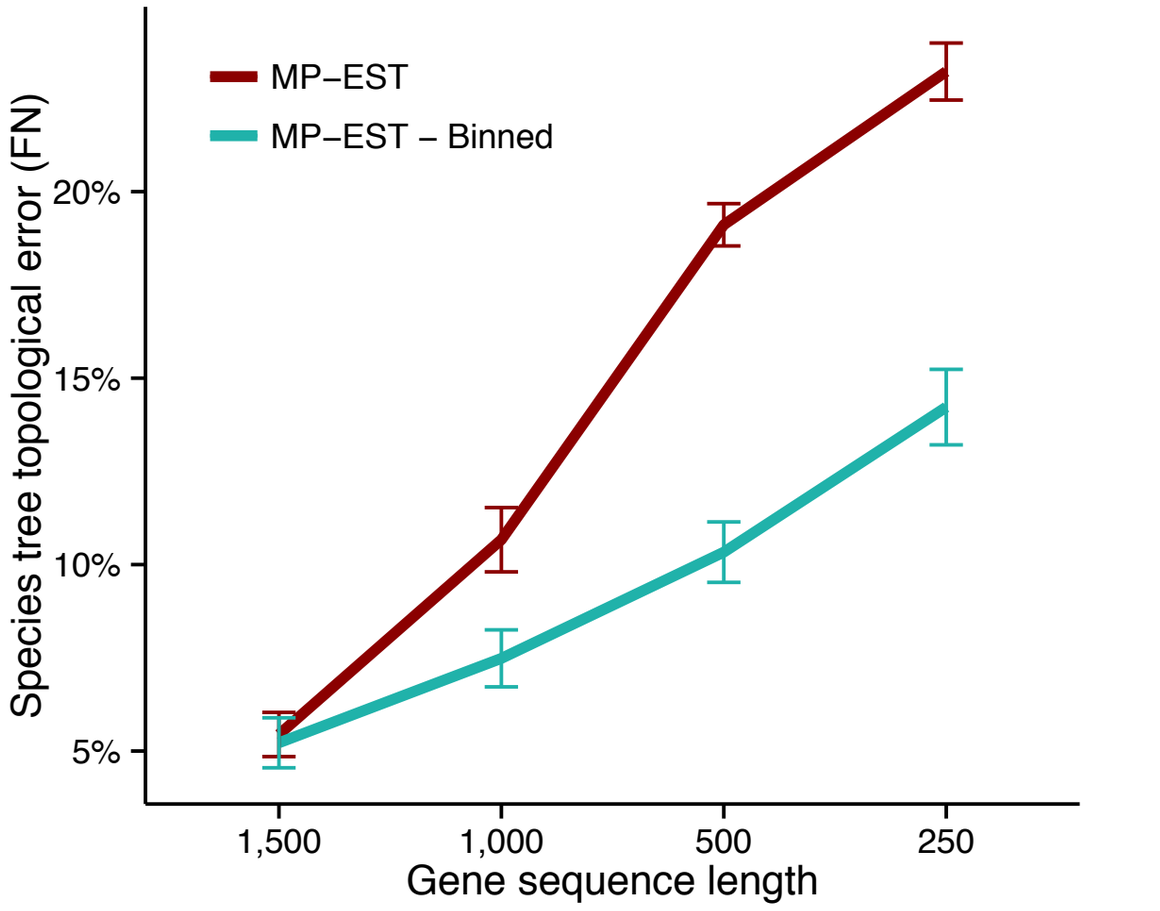
Statistical binning: overview



Original version: unweighted [S. Mirarab et al., Science (80). 346 (2014)]

New version: weighted; statistically consistent [M. S. Bayzid et al., PLoS One. 10 (2015)]

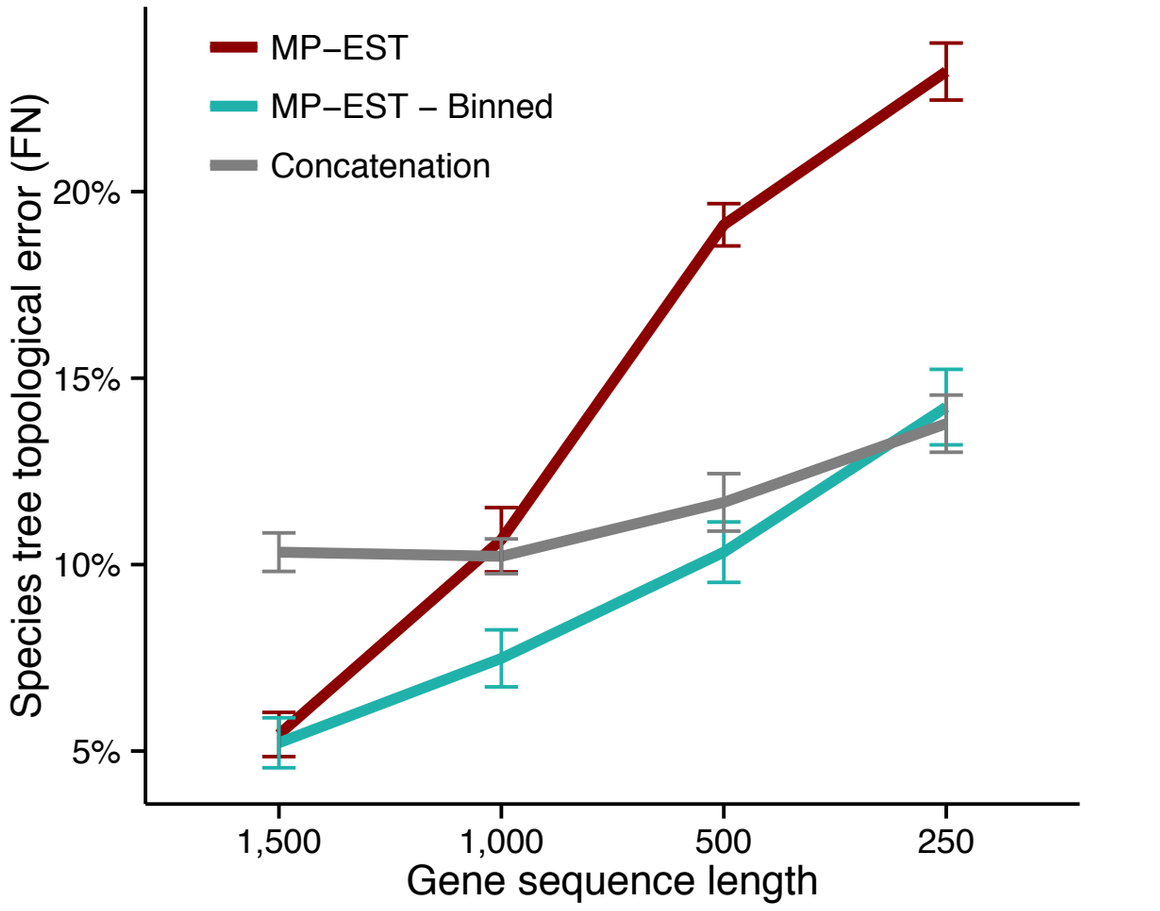
Statistical binning improves species tree estimation



more gene tree error →

45 avian-like species, 1000 genes
[S. Mirarab et al., Science (80). 346 (2014)]

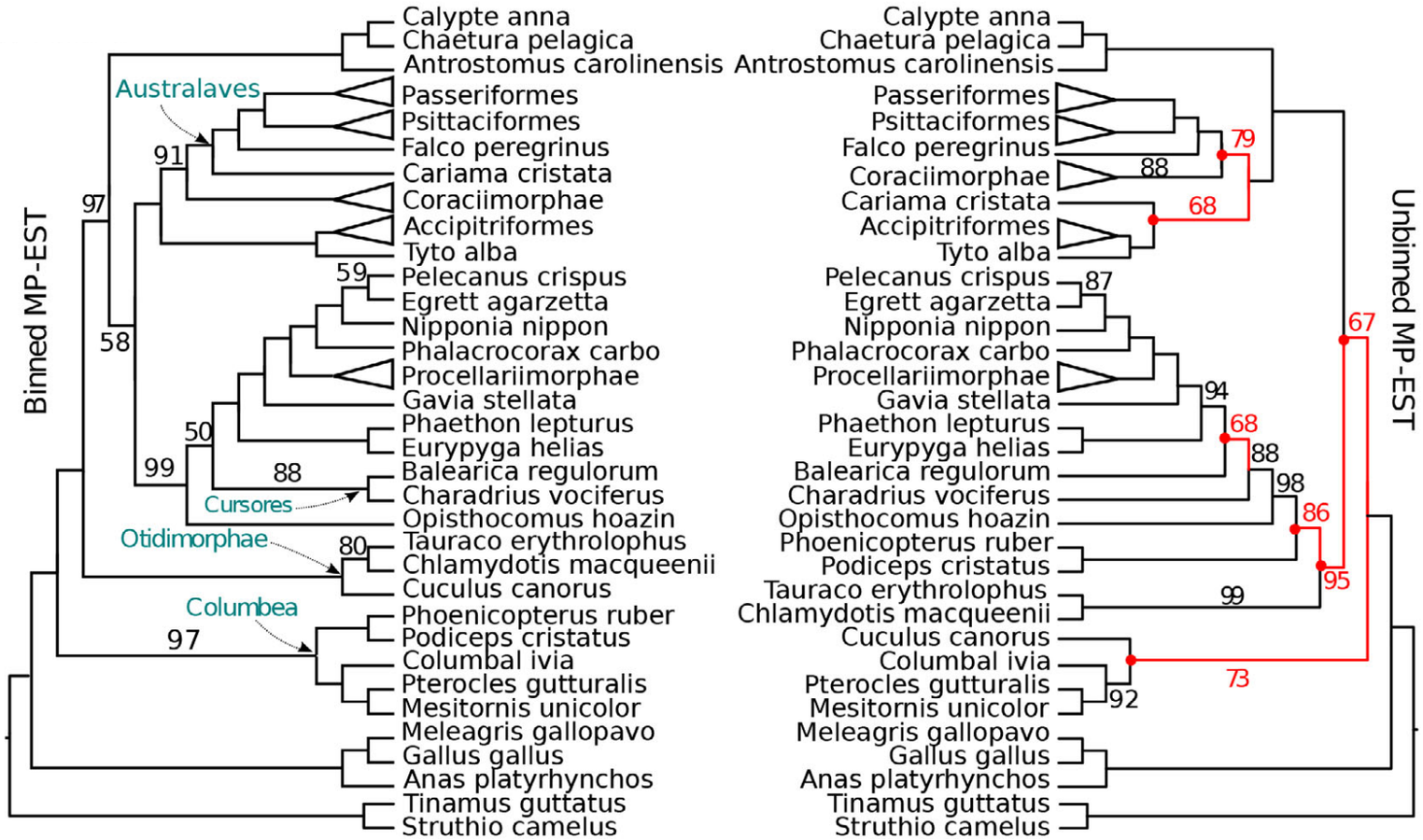
Statistical binning improves species tree estimation



more gene tree error →

45 avian-like species, 1000 genes
[S. Mirarab et al., Science (80). 346 (2014)]

MP-EST on the avian dataset



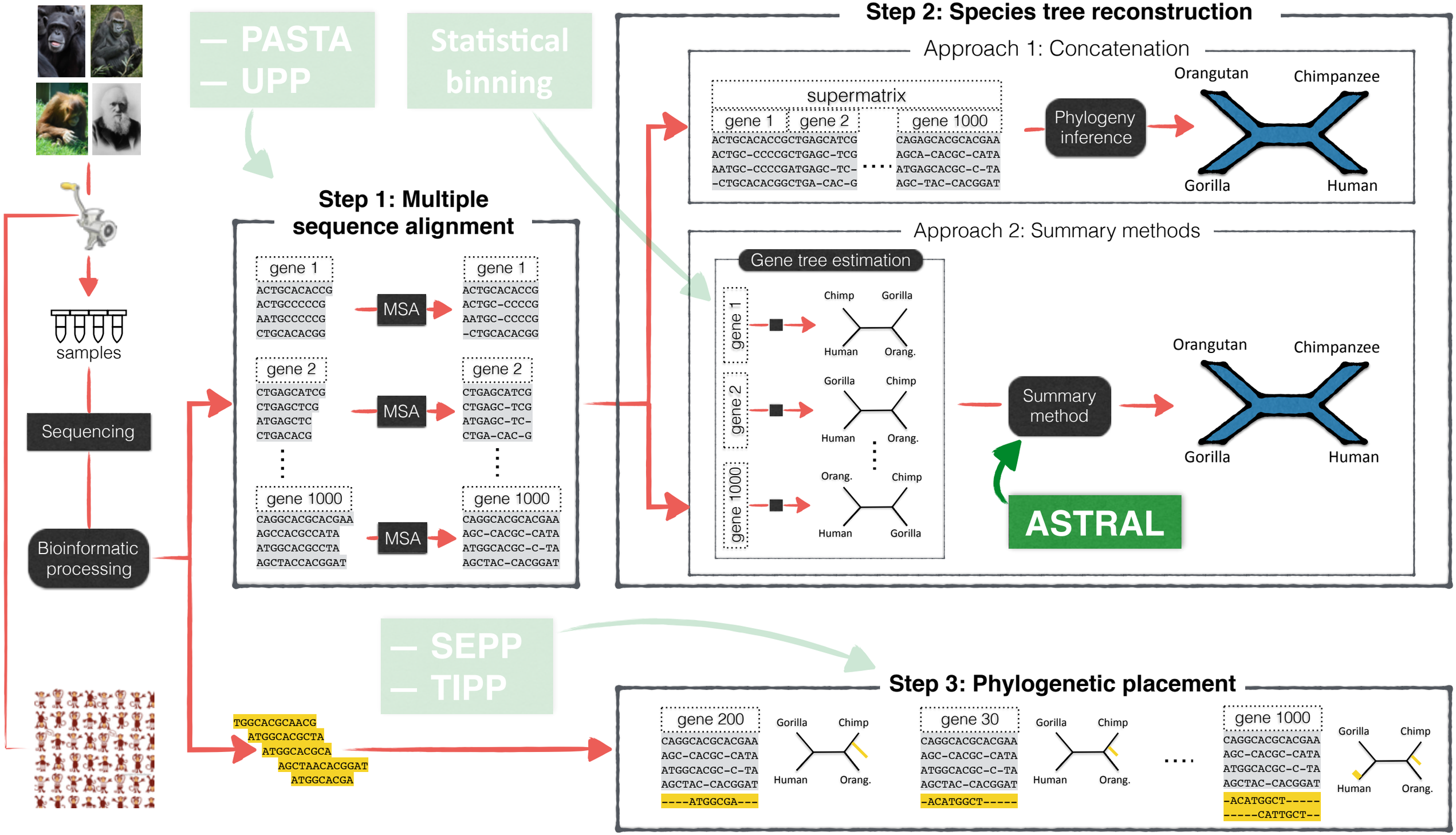
Binned MP-EST:

- highly supported,
- largely congruent with the concatenation,
- one of the two main trees

Unbinned MP-EST:

- low support,
- conflict with strong support from other sources of analyses

Multi-gene phylogeny reconstruction



1KP: Plant whole transcriptomes

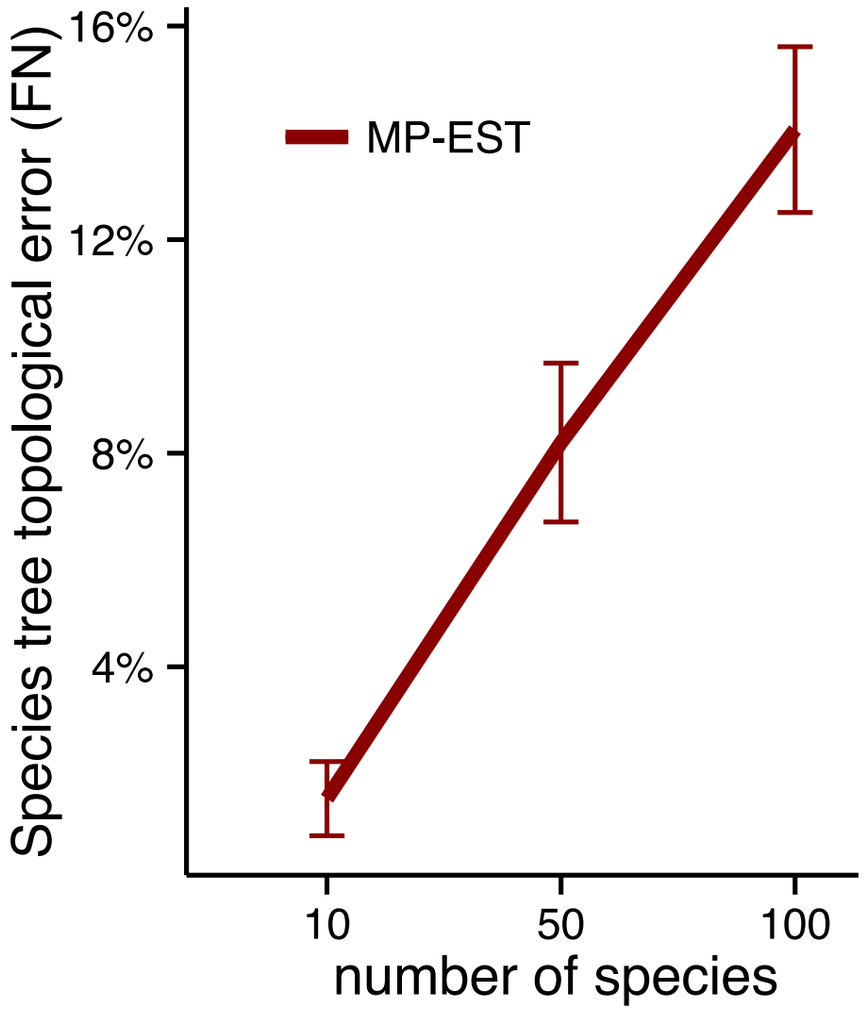
PNAS

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]

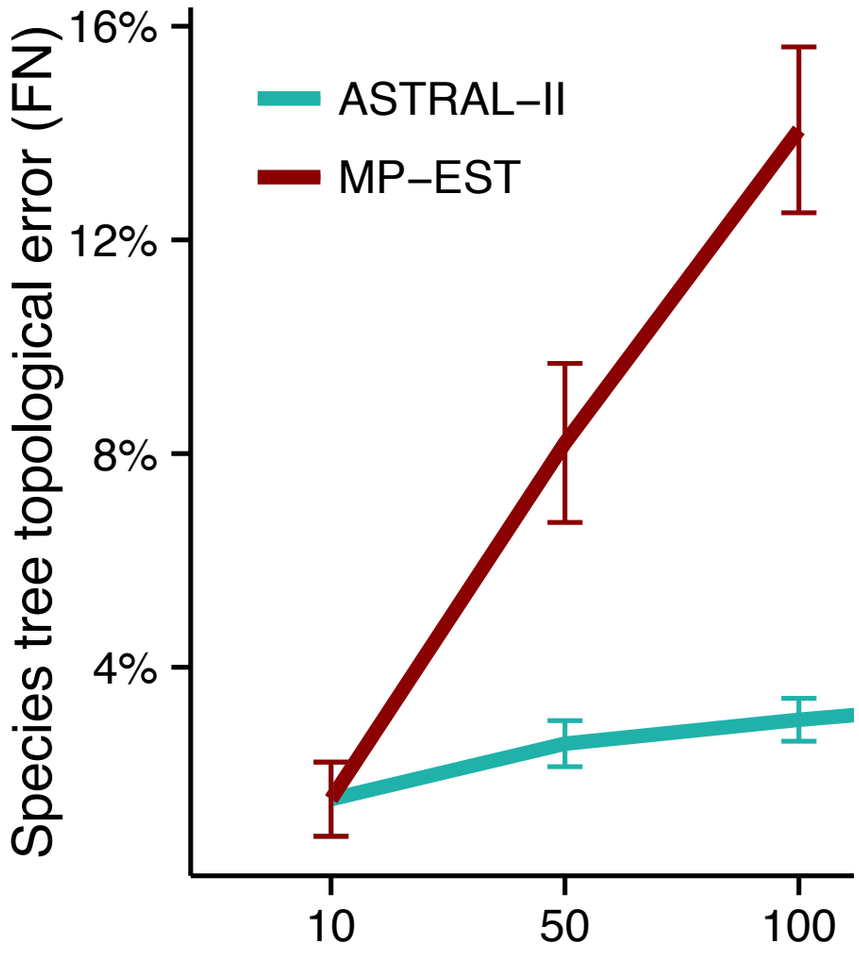
- Whole transcriptomes for 103 plant species
  - 1,200 in the next phase
- 400-800 single copy “genes”
- Spans ~1 billion years of evolution
- Many unanswered questions about plant evolution

Number of species impacts estimation error in the species tree



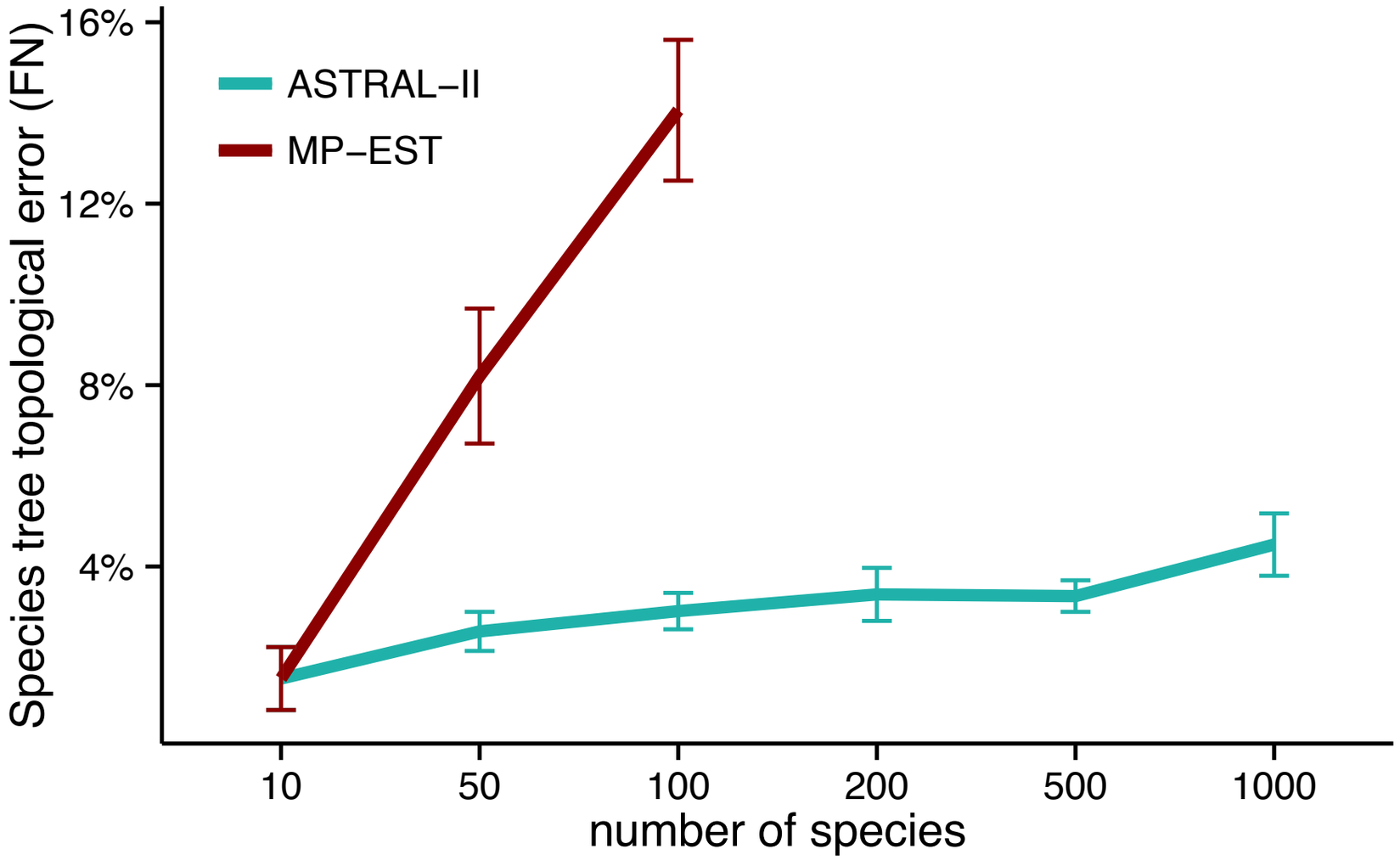
1000 genes, “medium” levels of ILS, simulated species trees
[S. Mirarab, T. Warnow, Bioinformatics. 31 (2015)]

ASTRAL: accurate and scalable



1000 genes, “medium” levels of ILS, simulated species trees
[S. Mirarab, T. Warnow, Bioinformatics. 31 (2015)]

ASTRAL: accurate and scalable



1000 genes, “medium” levels of ILS, simulated species trees
[S. Mirarab, T. Warnow, Bioinformatics. 31 (2015)]

ASTRAL

- **Input:** A set of inferred unrooted gene trees
- **Output:** A species tree with branch lengths in coalescent units and branch support values

[S. Mirarab et al., Bioinformatics. 30 (2014)]

ASTRAL

- **Input:** A set of inferred unrooted gene trees
- **Output:** A species tree with branch lengths in coalescent units and branch support values
- **Approach**: try to find the species tree that shares the maximum number of quartet trees with input gene trees. Proved statistically consistent for incomplete lineage sorting

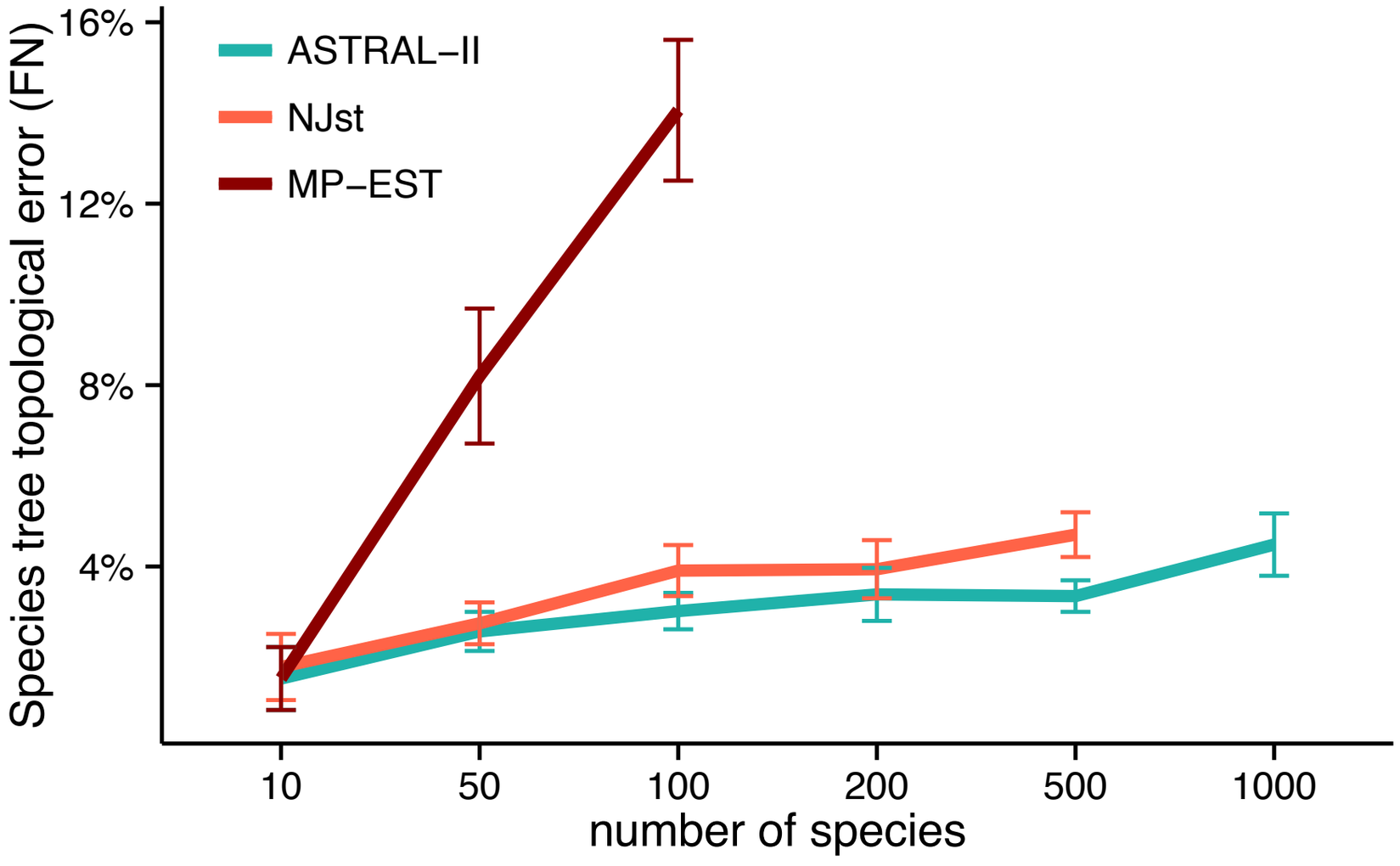
[S. Mirarab et al., Bioinformatics. 30 (2014)]

ASTRAL

- **Input:** A set of inferred unrooted gene trees
- **Output:** A species tree with branch lengths in coalescent units and branch support values
- **Approach**: try to find the species tree that shares the maximum number of quartet trees with input gene trees. Proved statistically consistent for incomplete lineage sorting
- **Designed for:**
  - Accuracy (established in simulation studies)
  - Scalability: the default version runs on a thousand genes from a thousand species in a day — Important for >1000 genomes

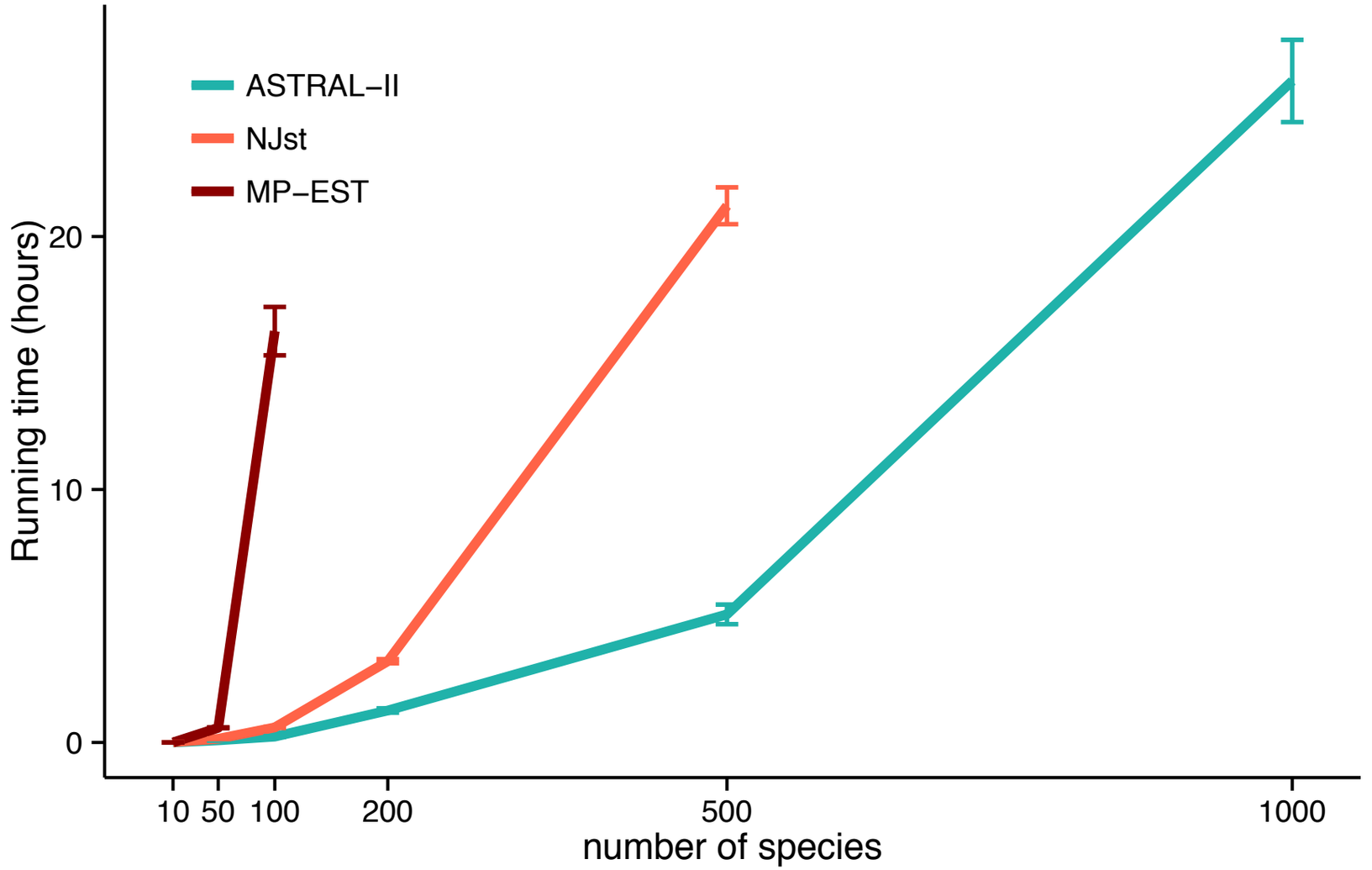
[S. Mirarab et al., Bioinformatics. 30 (2014)]

Tree error as a function of # species



1000 genes, “medium” levels of ILS, simulated species trees
[S. Mirarab, T. Warnow, Bioinformatics. 31 (2015)]

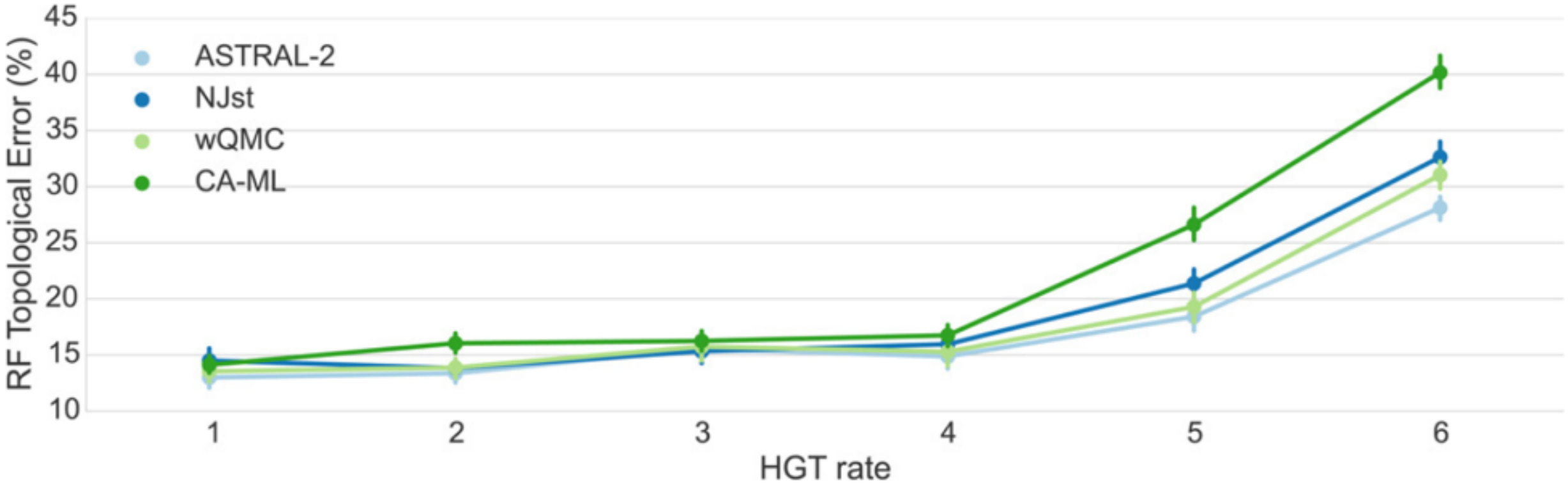
Running time as function of # species

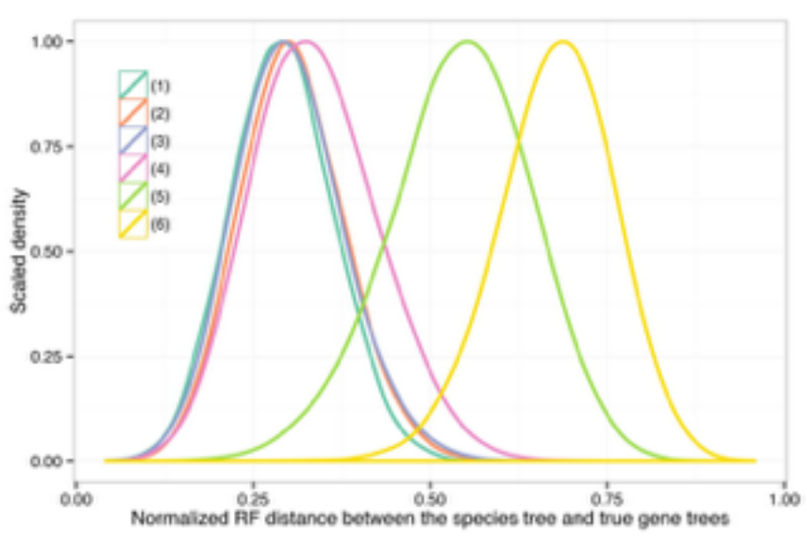


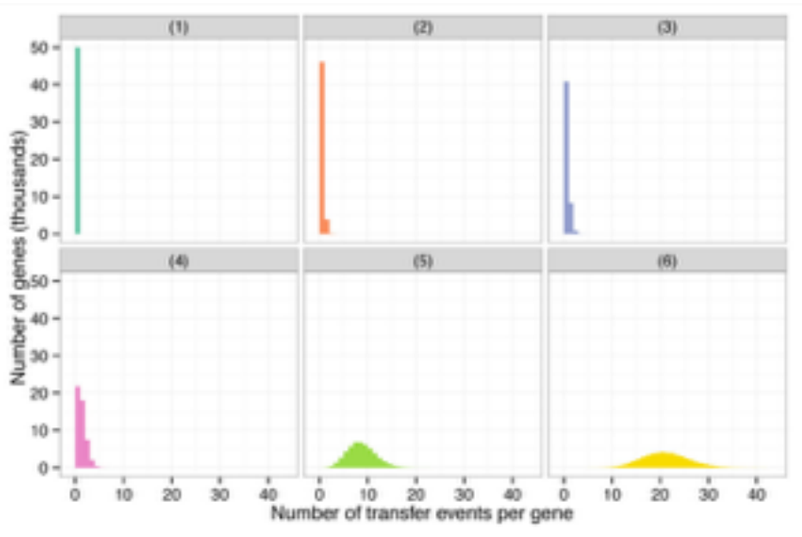
1000 genes, “medium” levels of ILS, simulated species trees
[S. Mirarab, T. Warnow, Bioinformatics. 31 (2015)]

ASTRAL with ILS+HGT

[R. Davidson et al., BMC Genomics. 16 (2015)]







Used in the field

- Plants: Wickett, et al., 2014, PNAS
- Birds: Prum, et al., 2015, Nature
- Xenoturbella, Cannon et al., 2016, Nature
- Xenoturbella, Rouse et al., 2016, Nature
- Flatworms: Laumer, et al., 2015, eLife
- Shrews: Giarla, et al., 2015, Syst. Bio.
- Frogs: Yuan et al., 2016, Syst. Bio.
- Tomatoes: Pease, et al., 2016, PLoS Bio.
- Angiosperms: Huang et al., 2016, MBE
- Worms: Andrade, et al., 2015, MBE

Ongoing improvements to ASTRAL

- John Yin: A GPU implementation is 10-20X faster (less than an hour on 1000 species)

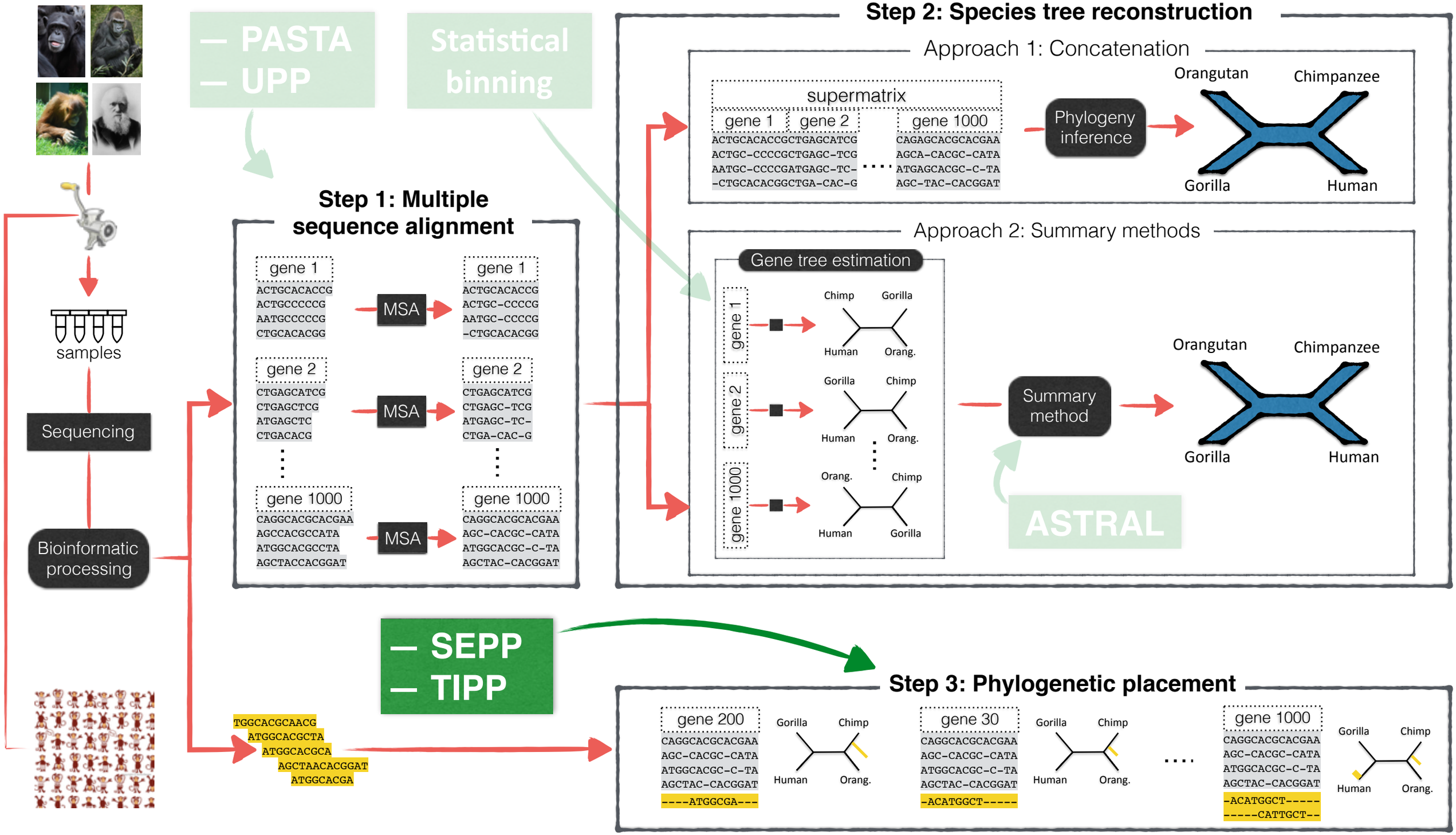
- Erfan Sayyari: Faster and more accurate measures of support, and finding signature of non-ILS-like discordance

- Maryam Rabiee:

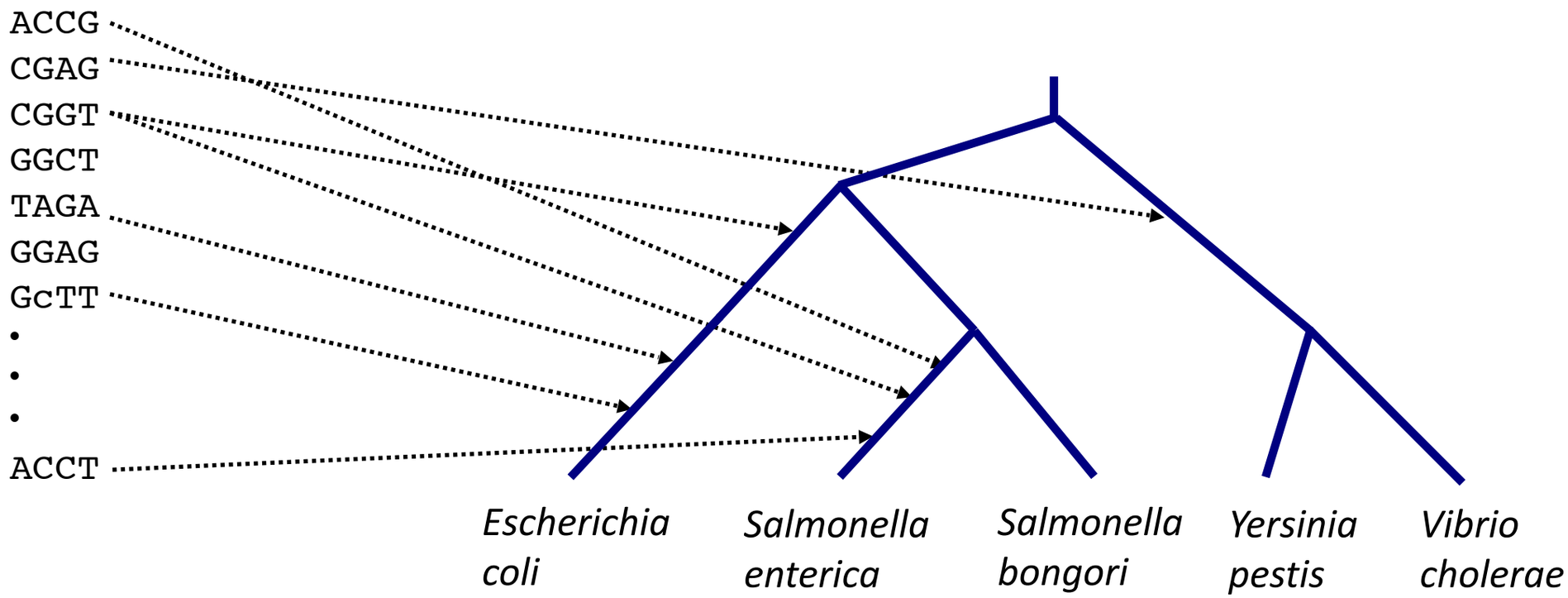
    - Better ways of dealing with multiple individuals from the same species

    - Divide-and-conquer to enable analyzing many tens of thousands of species

Multi-gene phylogeny reconstruction



Microbiome analyses using evolutionary trees



Fragmentary metagenomic reads

A **reference dataset** of full length sequences with an alignment and a tree

- Place fragmentary reads on a *reference tree* from datasets of known sequences

SEPP and TIPP

SEPP:
[S. Mirarab et al., PSB (2012)]

Step 1: Align query sequences to the backbone alignment

- Use a family of disjoint HMMs, created based on the reference tree

Step 2: Place each query sequence into backbone tree, using extended alignment

- Use divide-and-conquer on the backbone tree

SEPP and TIPP

SEPP:
[S. Mirarab et al., PSB (2012)]

Step 1: Align query sequences to the backbone alignment

- Use a family of disjoint HMMs, created based on the reference tree

Step 2: Place each query sequence into backbone tree, using extended alignment

- Use divide-and-conquer on the backbone tree

TIPP:
[N. Nguyen et al., Bioinformatics (2014)]

Step 1: Find fragments from ~30 known “marker” genes

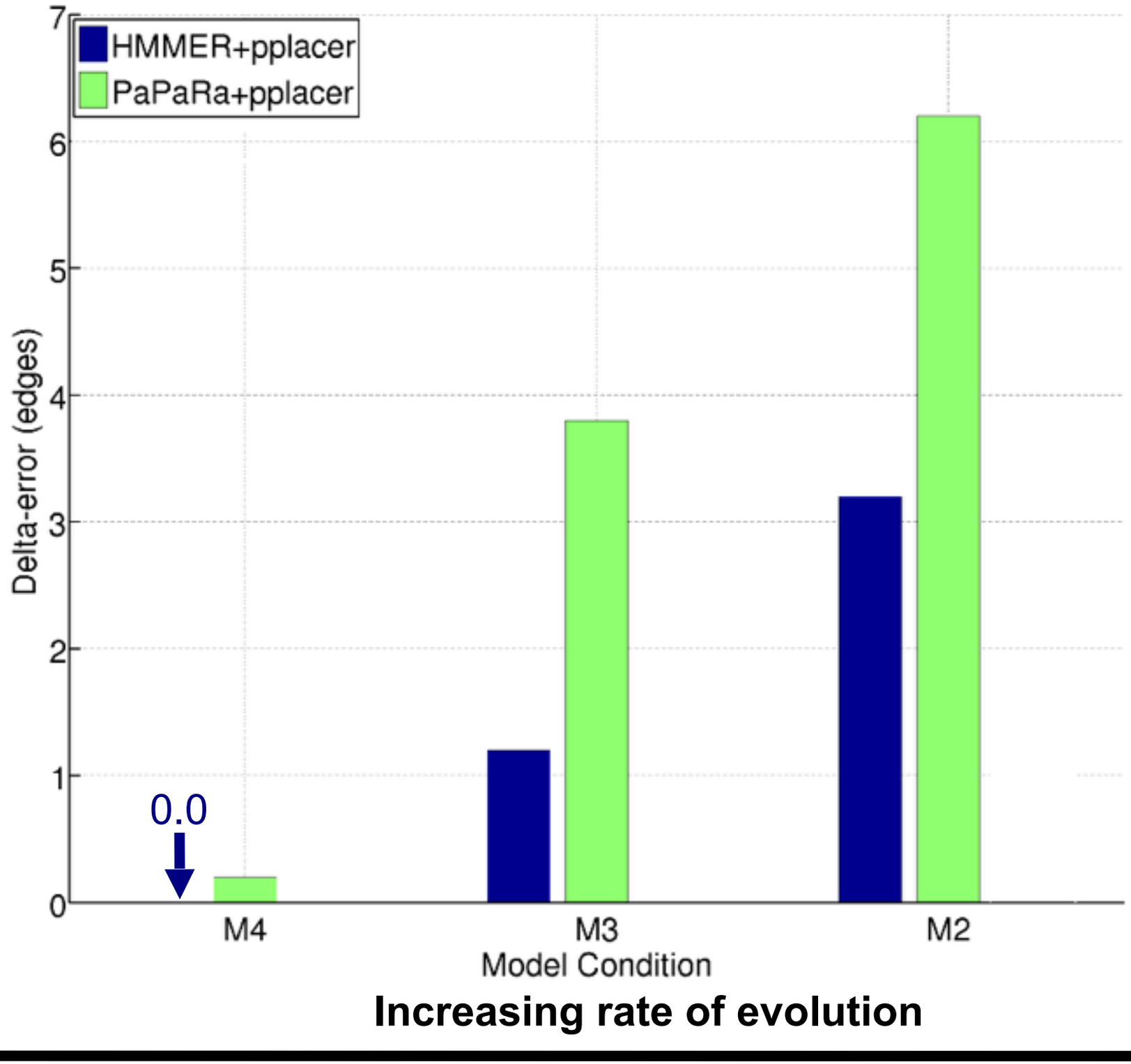
Step 2: Use SEPP to place reads on the marker trees

- Use many alignments and tree placements per read to account for uncertainty

Step 3: Summarize placement results to get a taxonomic profile

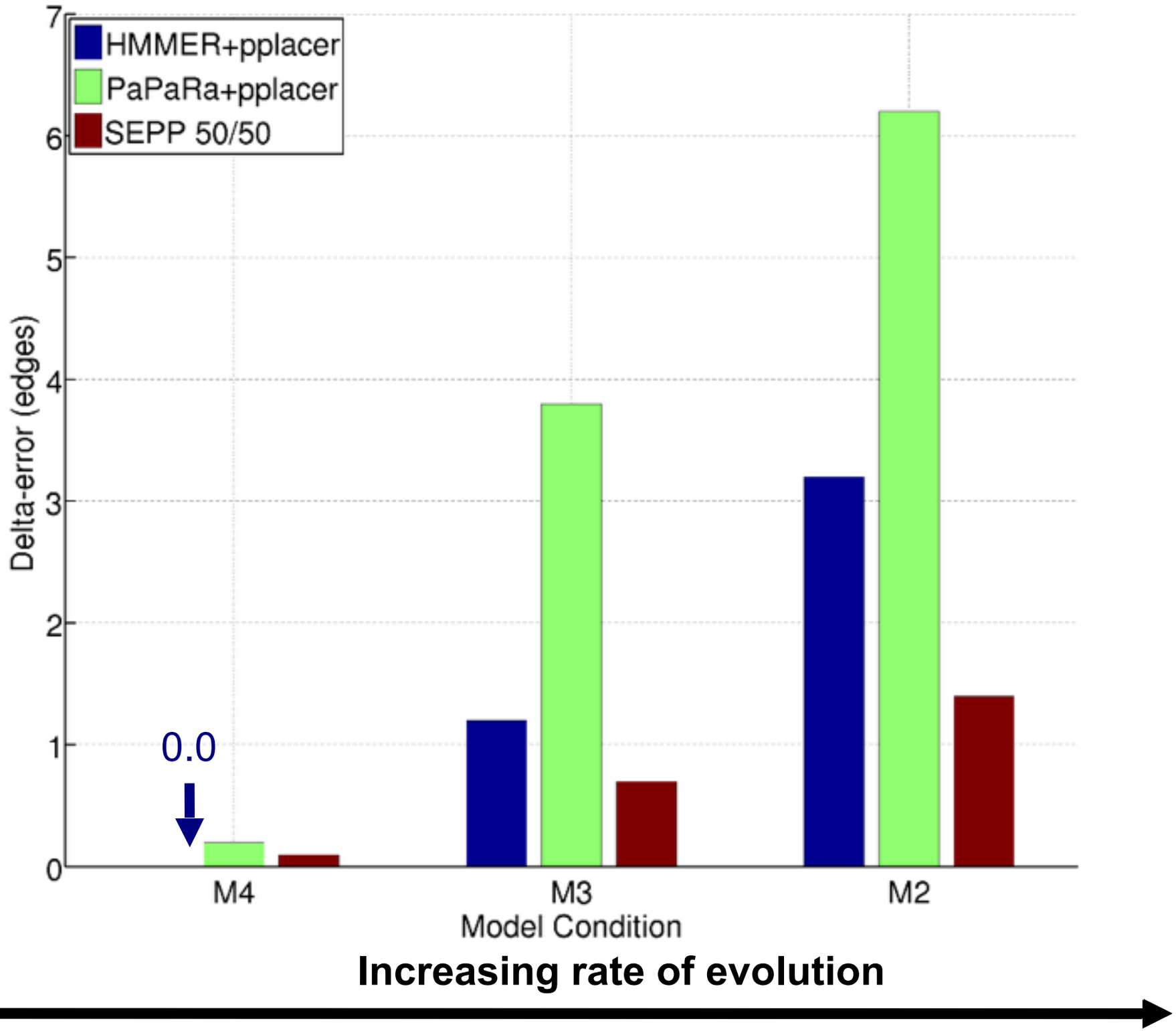
Phylogenetic placement simulations

S. Mirarab et al., Pacific Symp. Biocomput. (2012).



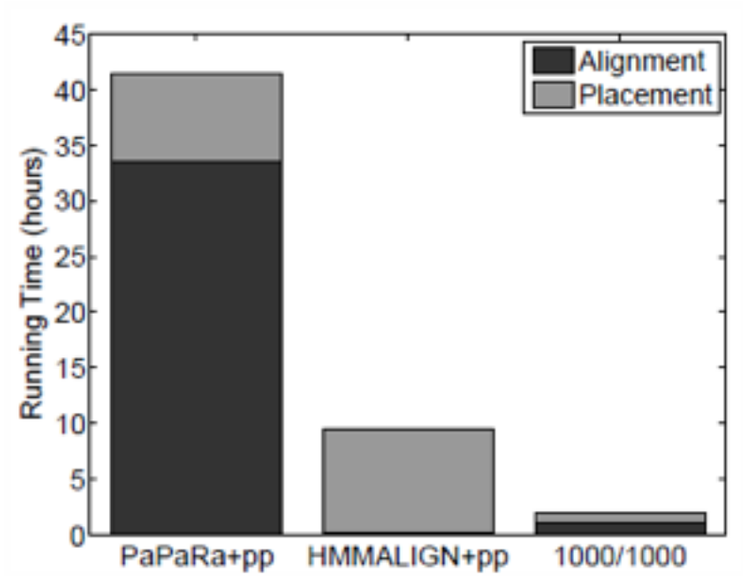
SEPP on simulated data

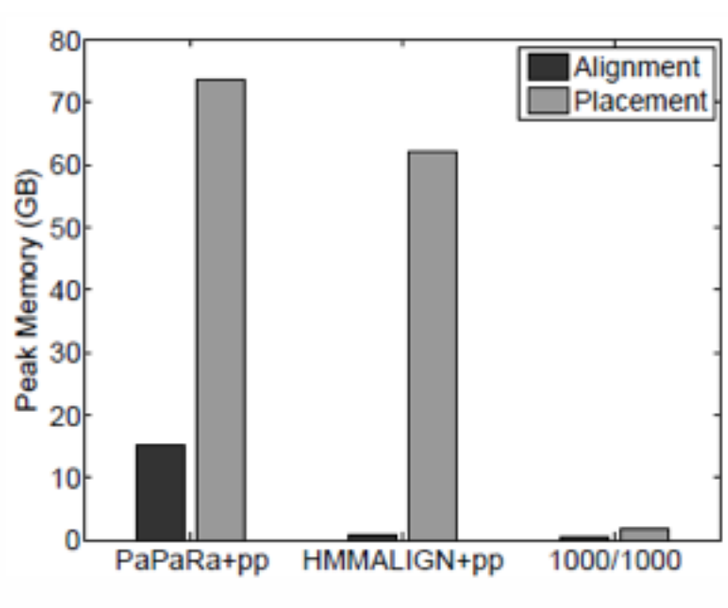
S. Mirarab et al., Pacific Symp. Biocomput. (2012).

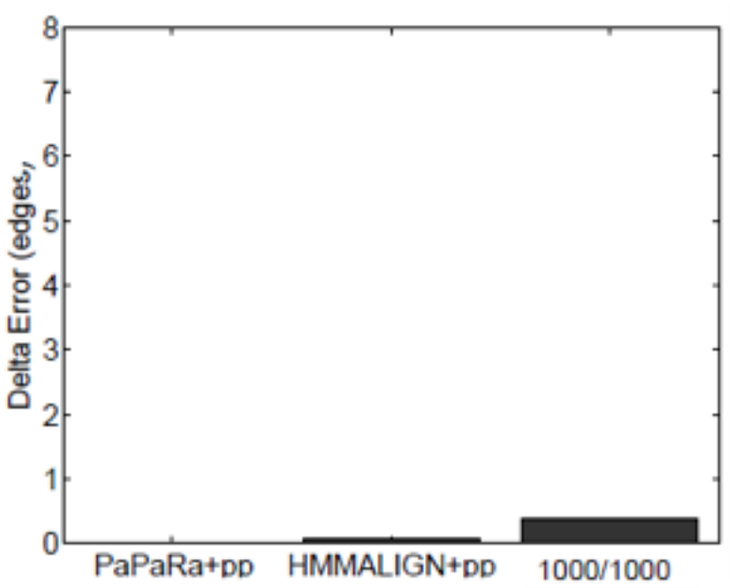


SEPP on 16S

Simulations: 16S bacteria, 13k curated backbone tree, 13k fragments

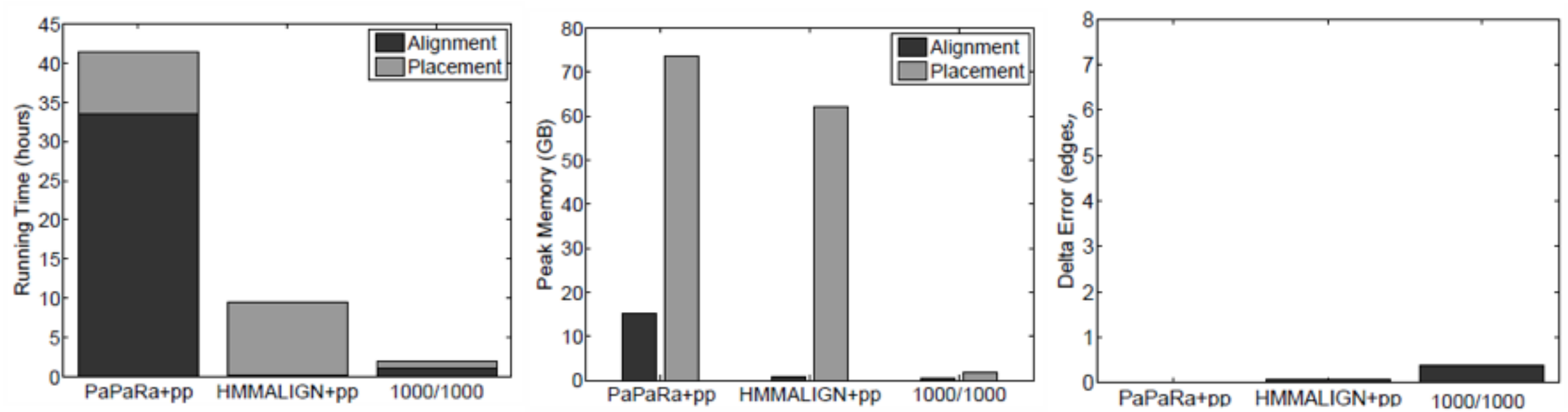






SEPP on 16S

Simulations: 16S bacteria, 13k curated backbone tree, 13k fragments

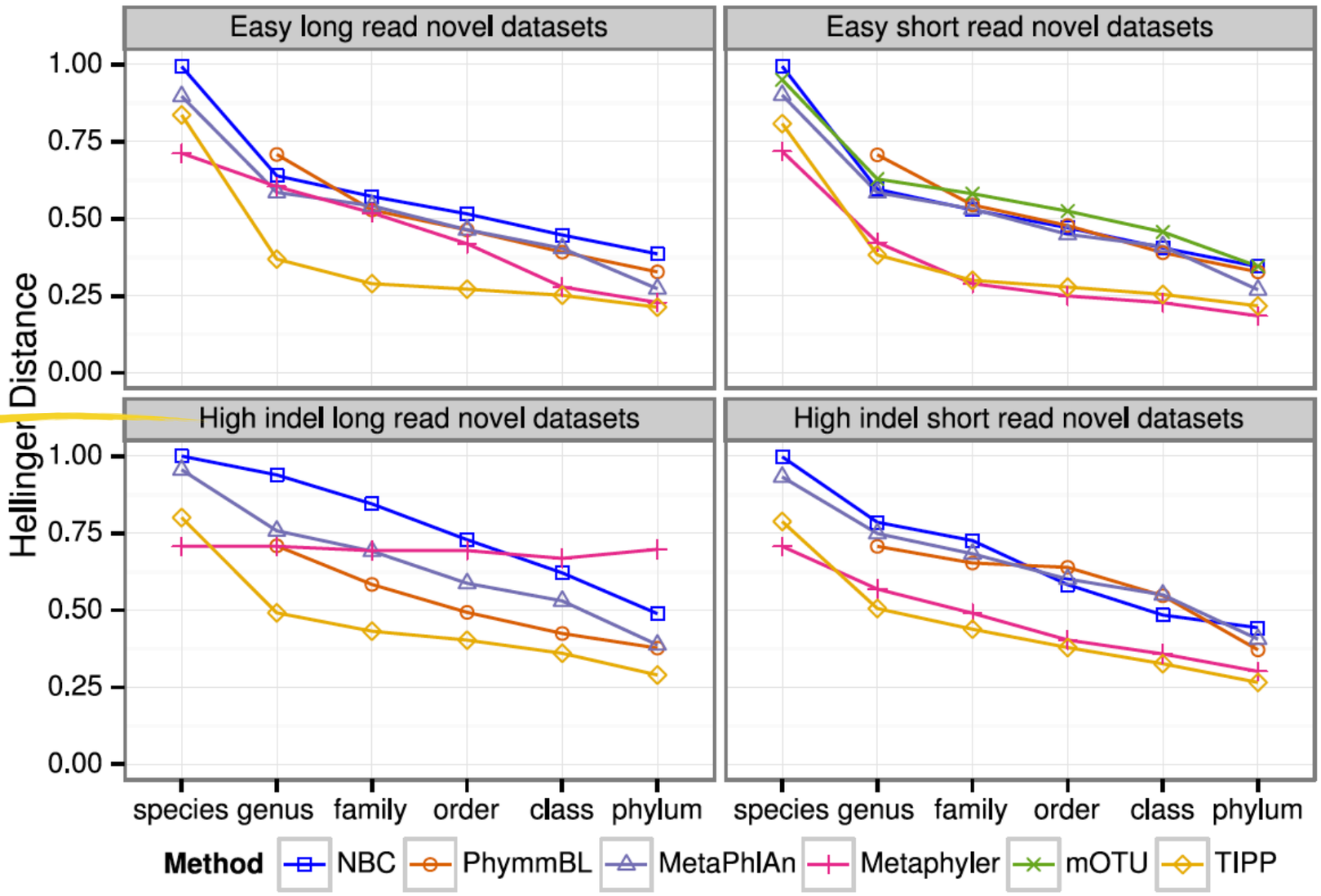


Recent studies (Thanks to Daniel McDonald):

- **EMP:** placing ~300,000 fragments on the 99% greengenes reference tree with ~300,000 sequences: 8 hours (Gordon)

- **AG:** placing ~40,000 fragments on the 99% green genes reference tree with ~200,000 sequences: 10 minutes (Gordon)

TIPP results on “novel” genomes



Long indel rates: “future” generations of sequencing technologies

PhymmBL (Brady & Salzberg, Nature Methods 2009)
NBC (Rosen, et al., Bioinformatics 2011)
MetaPhyler (Liu et al., BMC Genomics 2011)

MetaPhlAn (Segata et al., Nature Methods 2012)
mOTU (Bork et al., Nature Methods 2013)

Software availability

- ASTRAL: github.com/smirarab/ASTRAL
- PASTA: github.com/smirarab/pasta
  (internally uses FastTree, Mafft, HMMER, and OPAL)
- SEPP: github.com/smirarab/sepp
  (internally uses pplacer and HMMER)
- UPP: https://github.com/smirarab/sepp/blob/master/README.UPP.md (internally uses HMMER)
- TIPP: https://github.com/smirarab/sepp/blob/master/README.TIPP.md (internally uses pplacer and HMMER)
- Statistical binning: https://github.com/smirarab/binning

Acknowledgments

Keshav Pingali

Tandy Warnow

S.M. Bayzid

Nam Nguyen
(now at UIUC)

Théo
Zimmermann

Erfan Sayyari
(ECE)

Niema A. Moshiri
(Bioinformatics)

Norman Wickett
(U Chicago)

Jim Leebens-mack
(UGA)

Erich Jarvis
(Duke, HMMI)

Bastien Boussau
(Université Lyon)

Ed Braun
(U Florida)

Guojie Zhang
(BGI, China)

Tom Gilbert
(U Copenhagen)

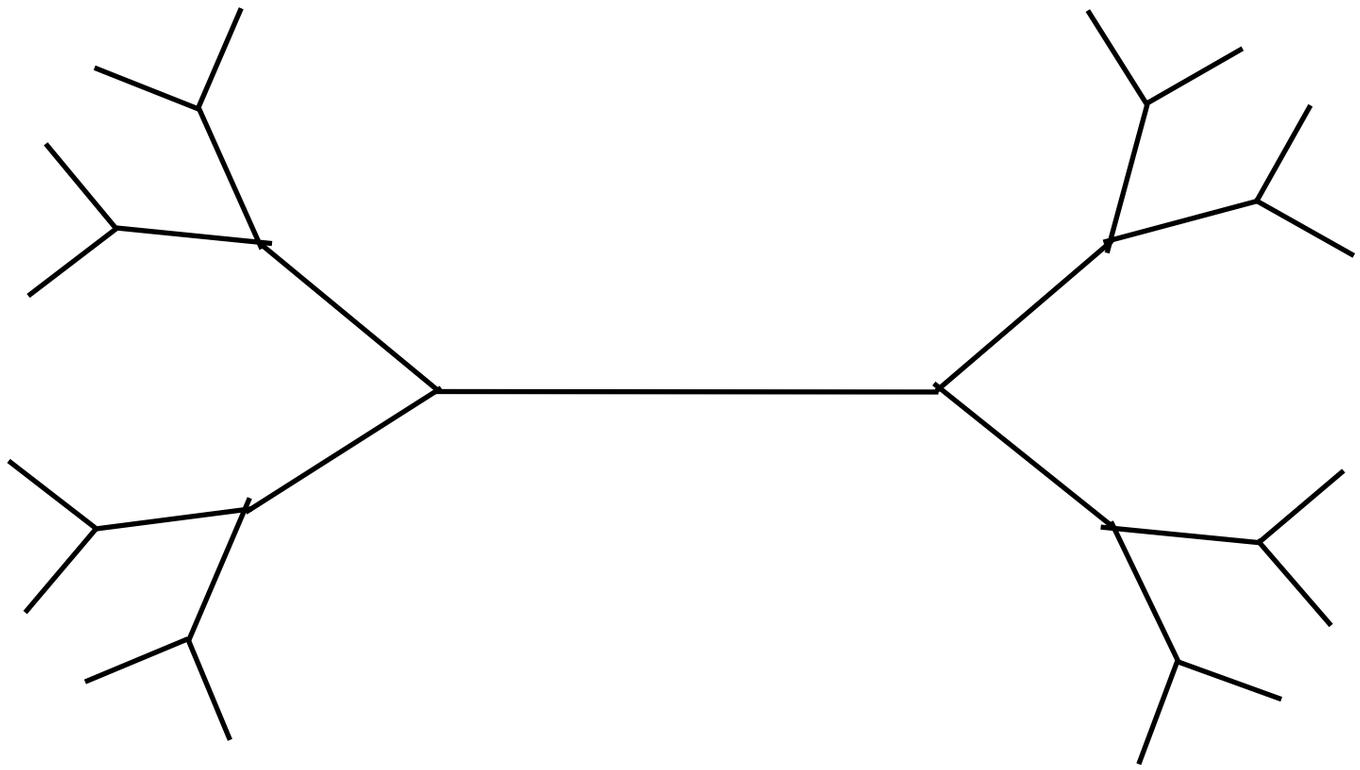
. . .

. . .

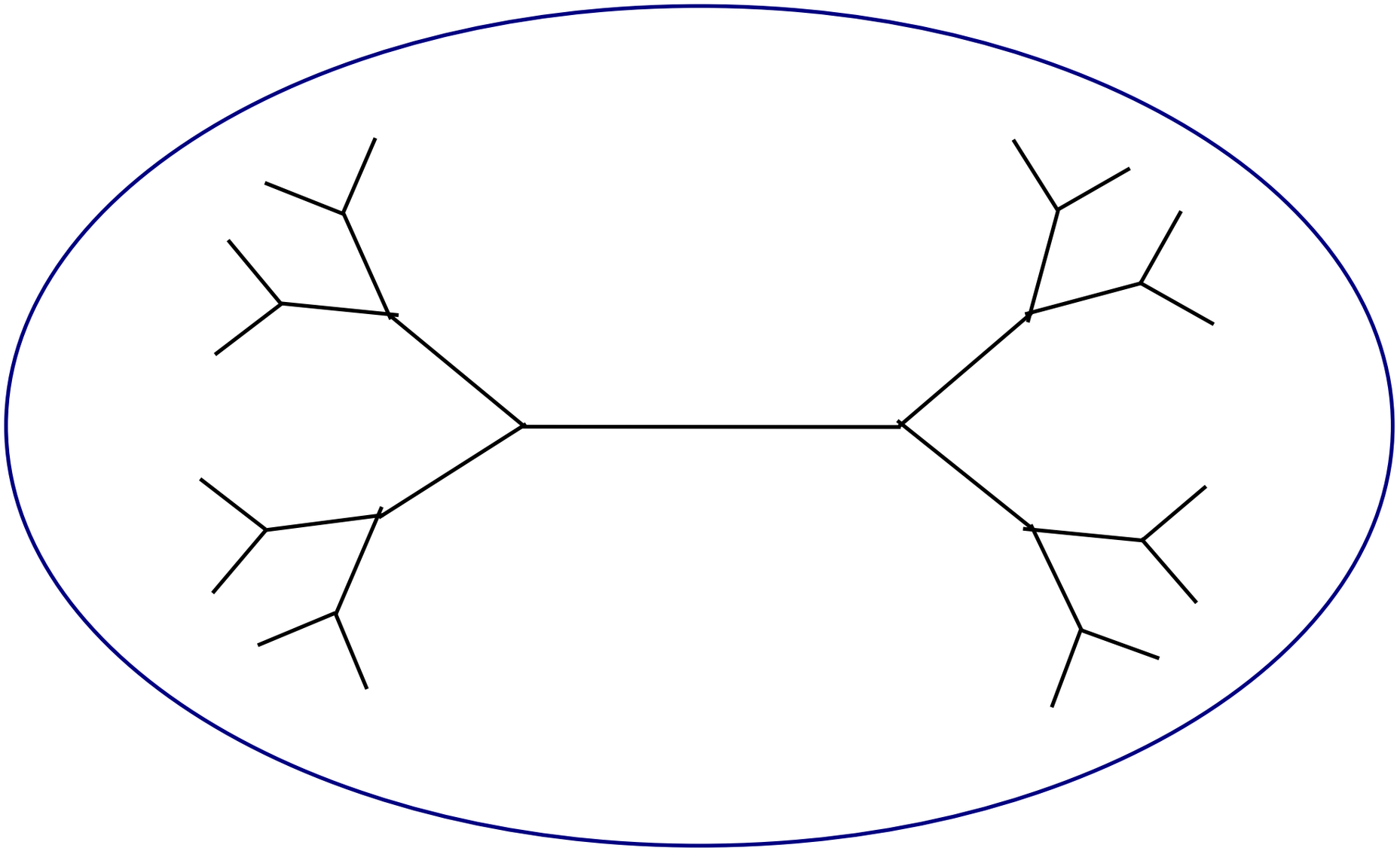
HHMI
HOWARD HUGHES MEDICAL INSTITUTE

NSF

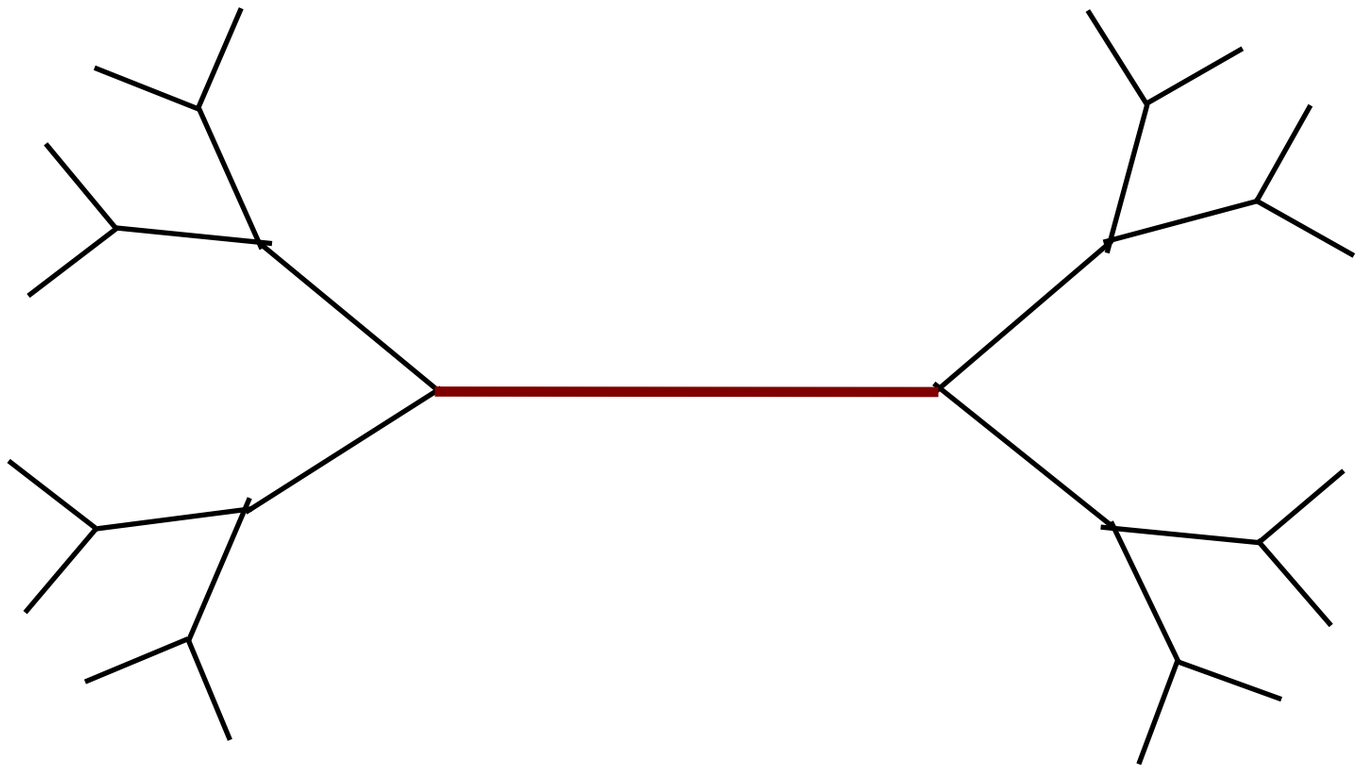
Insights from SATé



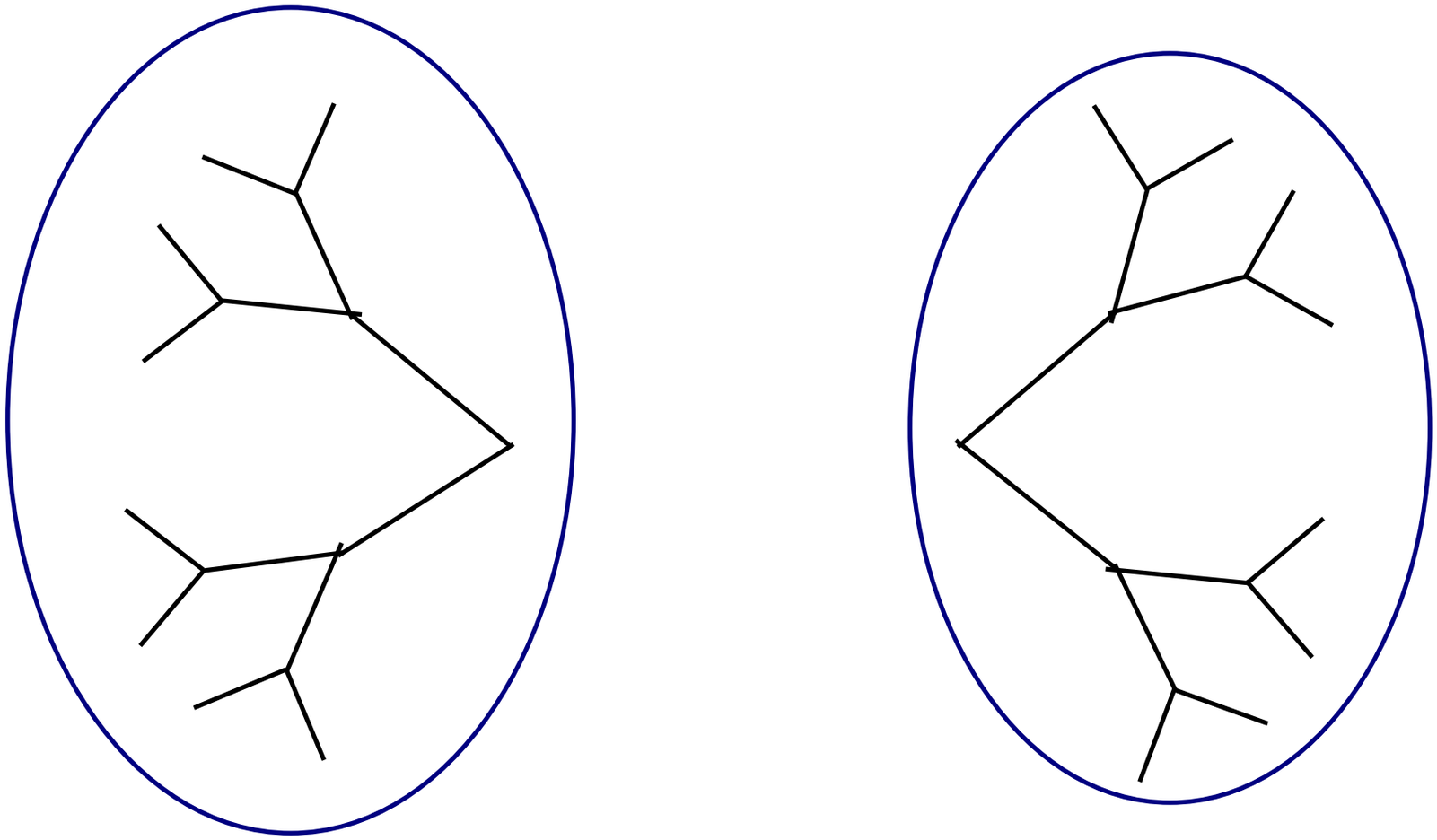
Insights from SATé



Insights from SATé



Insights from SATé



Insights from SATé