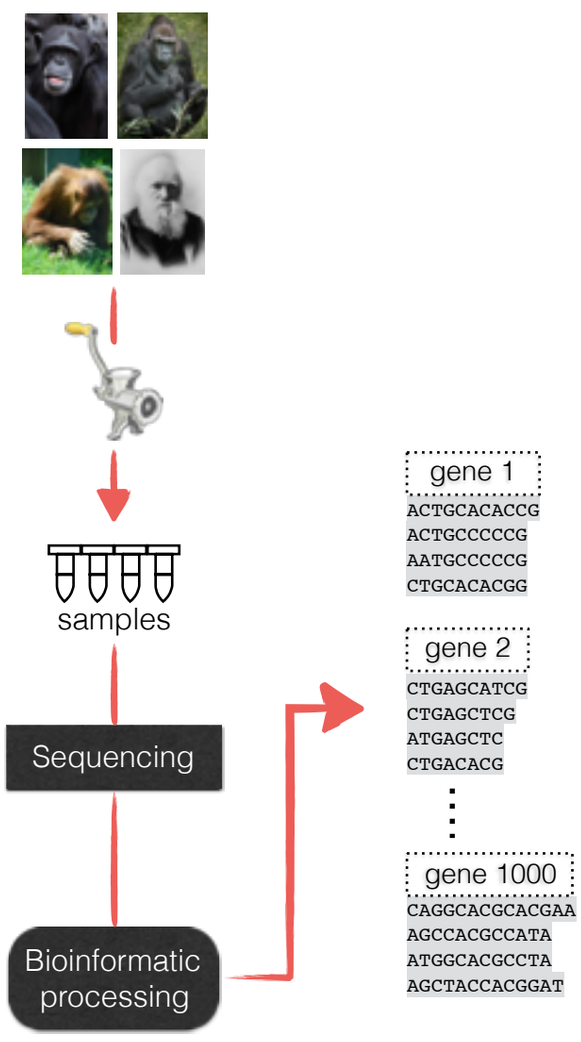
Taxonomic Profiling using Scalable Phylogenetic Placement

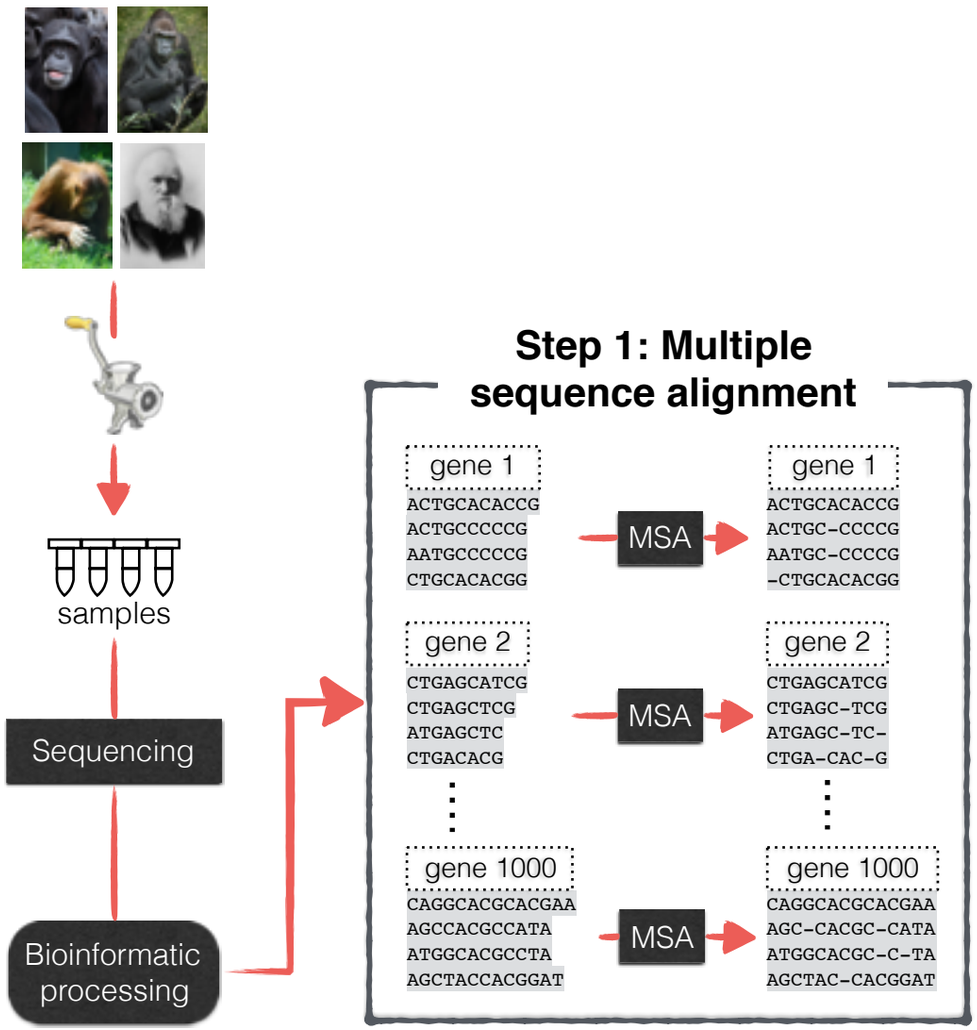
University of California at San Diego (UCSD)
Siavash Mirarab

Joint work with
Tandy Warnow, Nam-Phuong Nguyen
Mike Nute, Mihai Pop, and Bo Liu

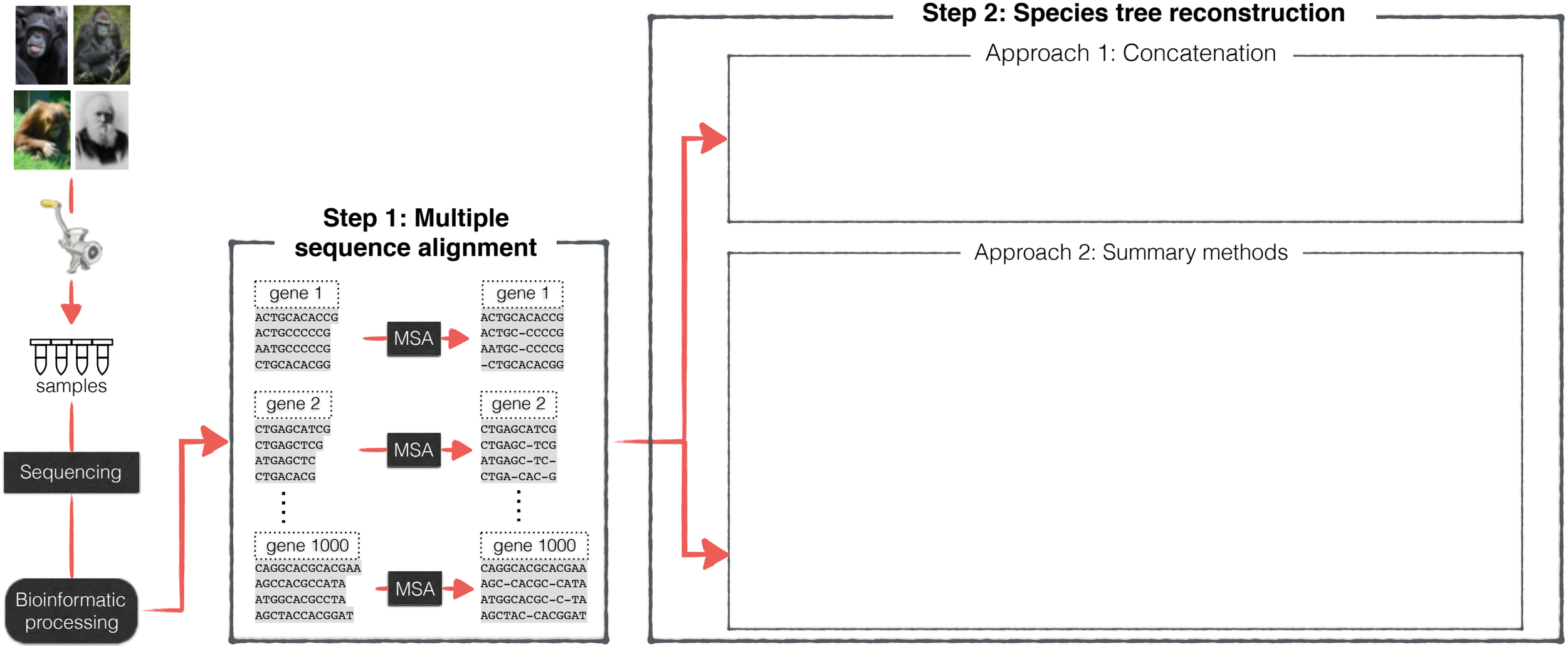
Phylogeny reconstruction pipeline



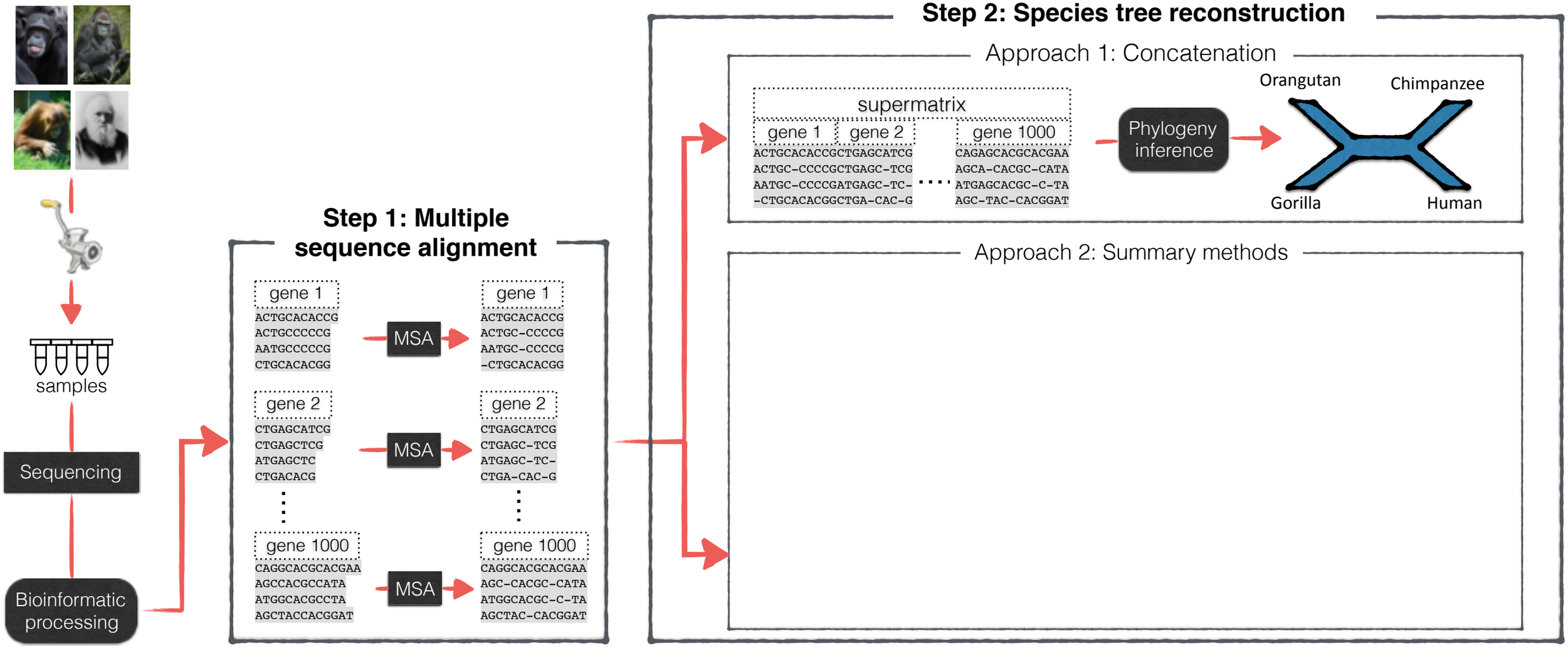
Phylogeny reconstruction pipeline



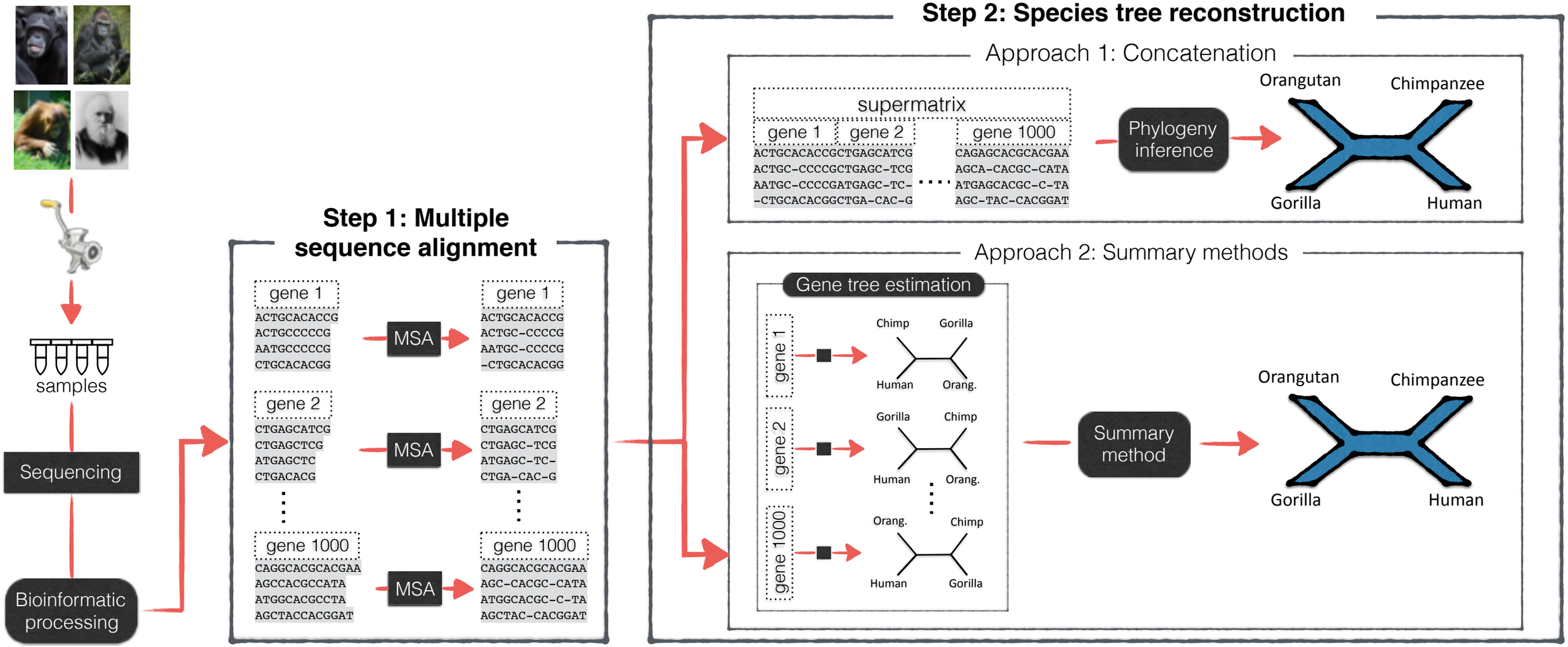
Phylogeny reconstruction pipeline



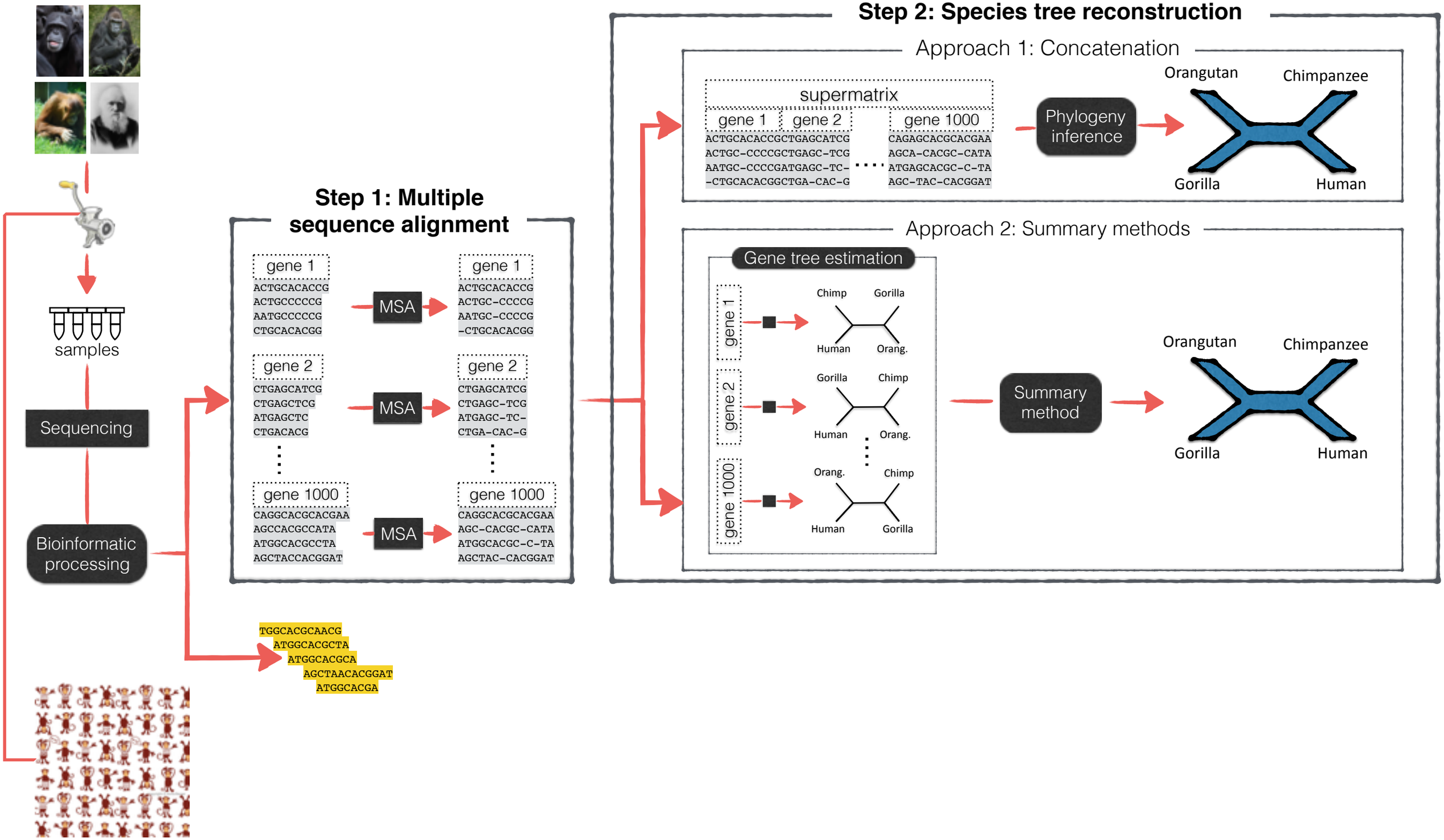
Phylogeny reconstruction pipeline



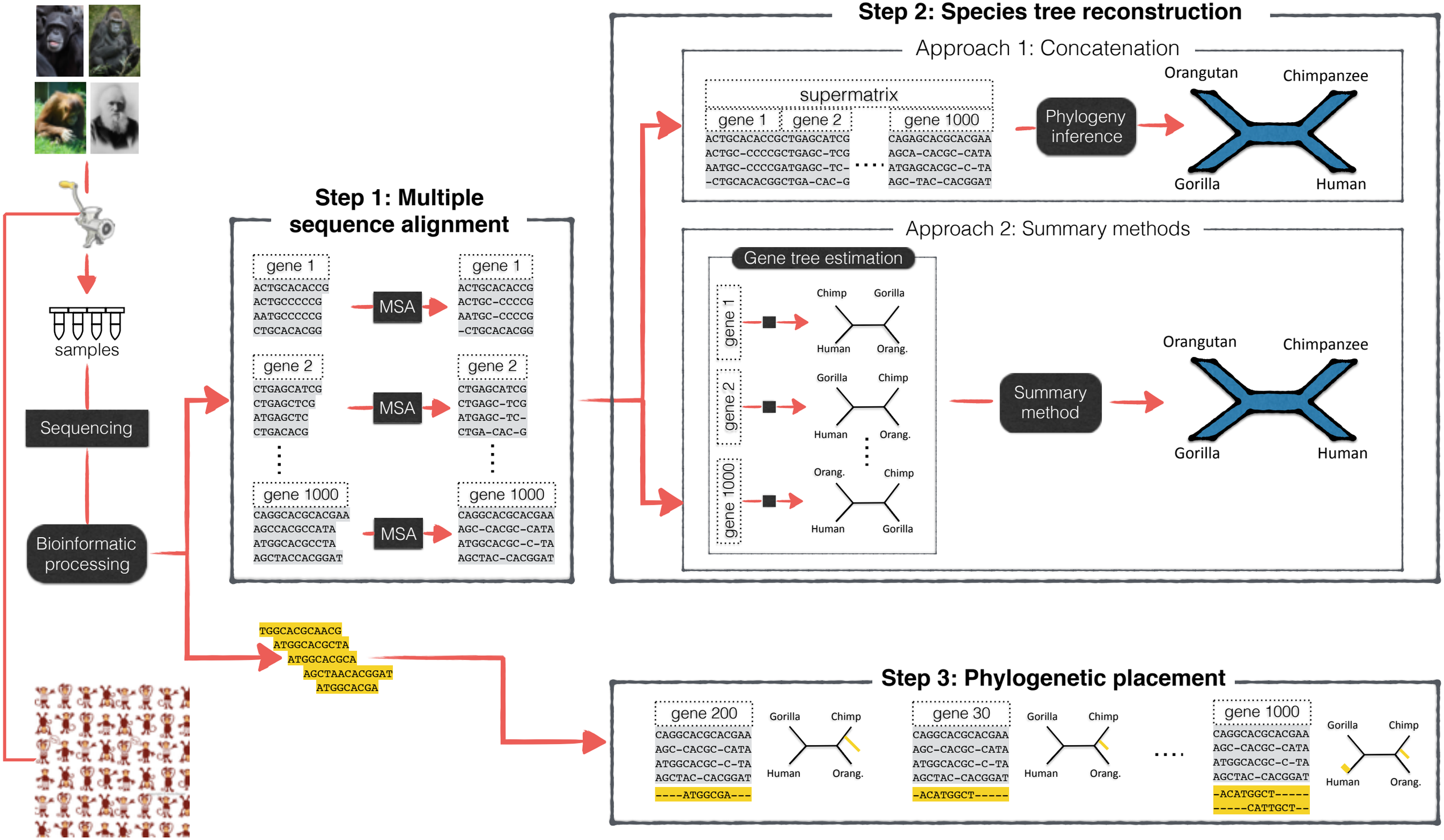
Phylogeny reconstruction pipeline



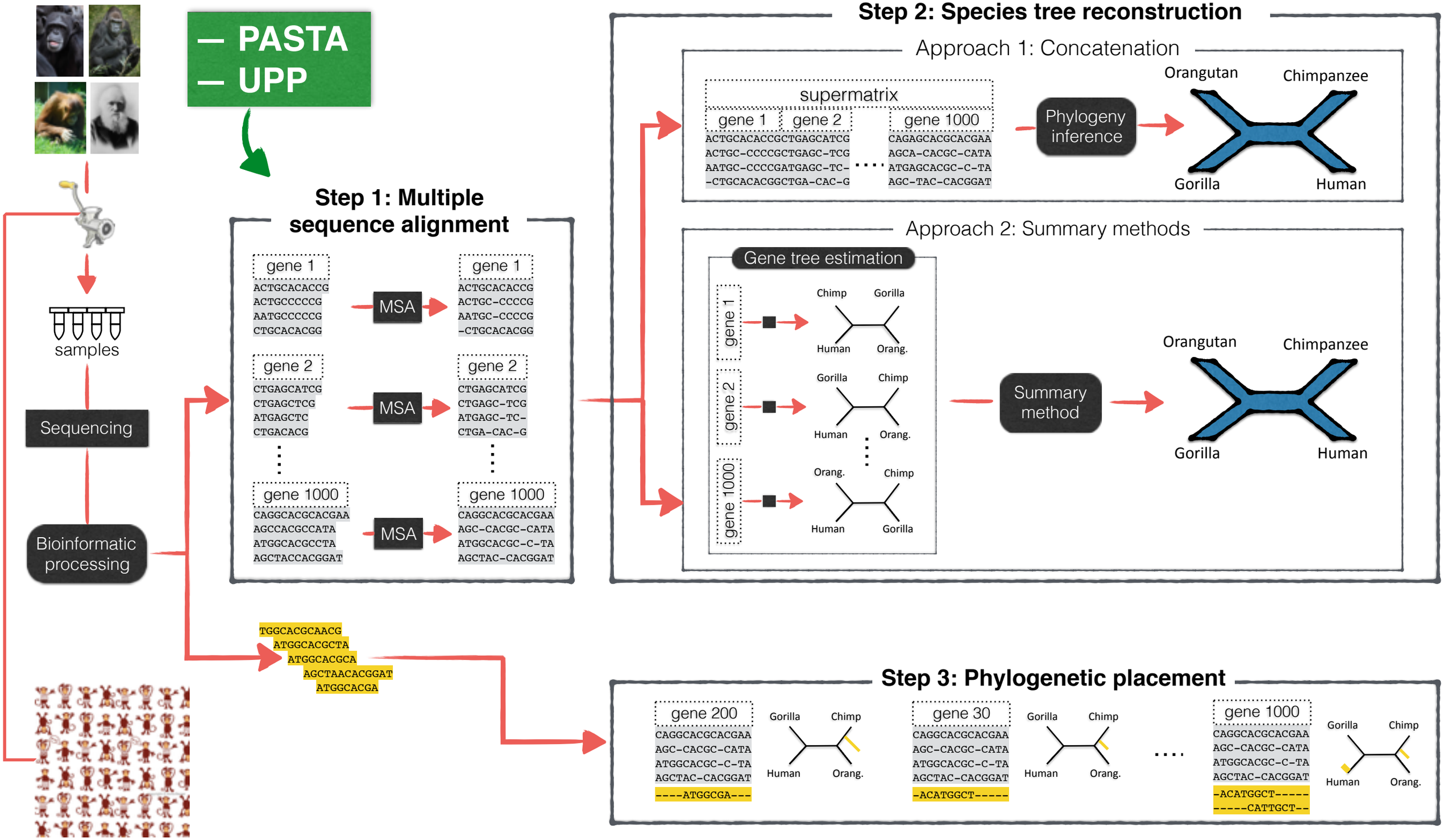
Phylogeny reconstruction pipeline



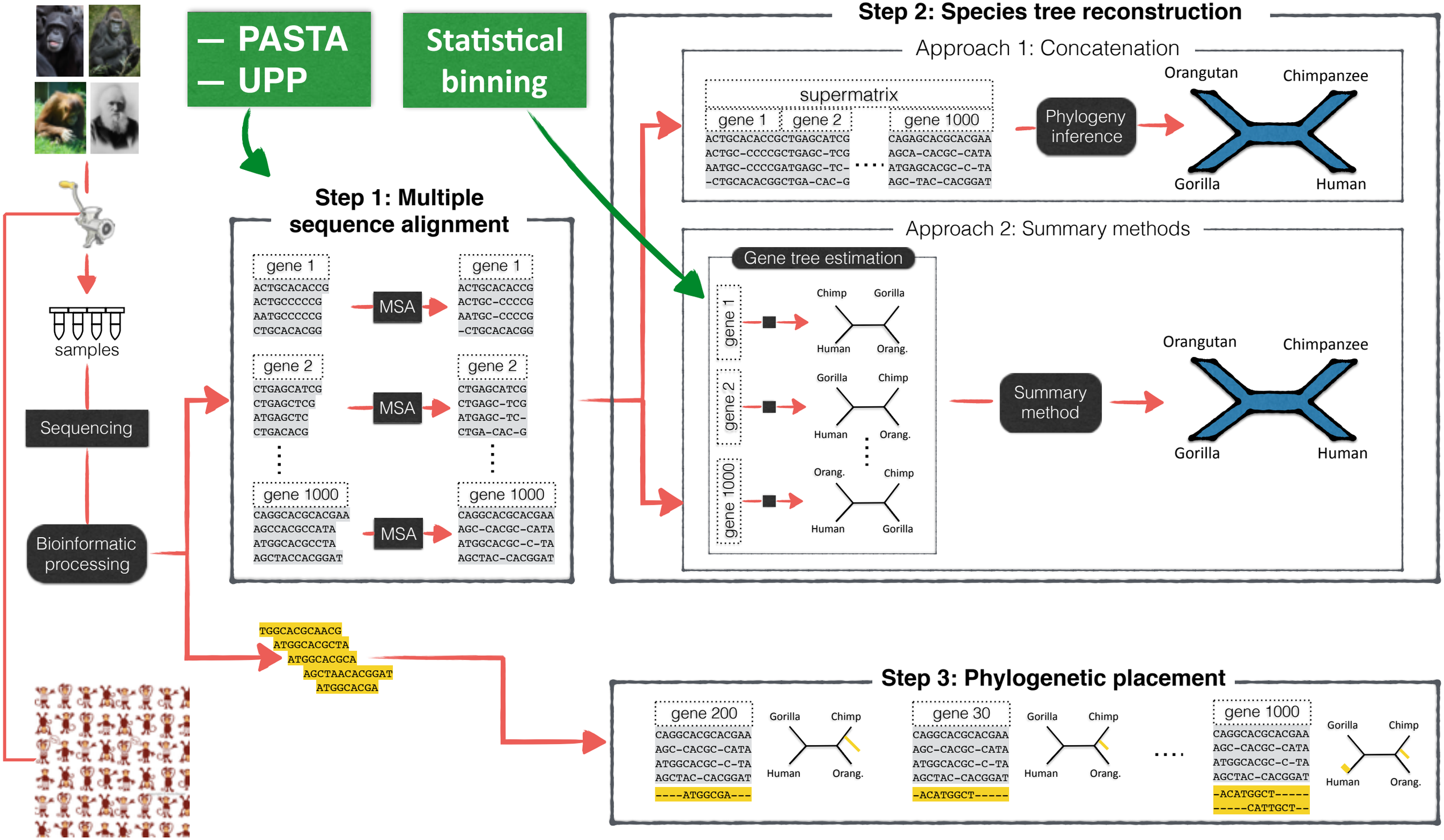
Phylogeny reconstruction pipeline



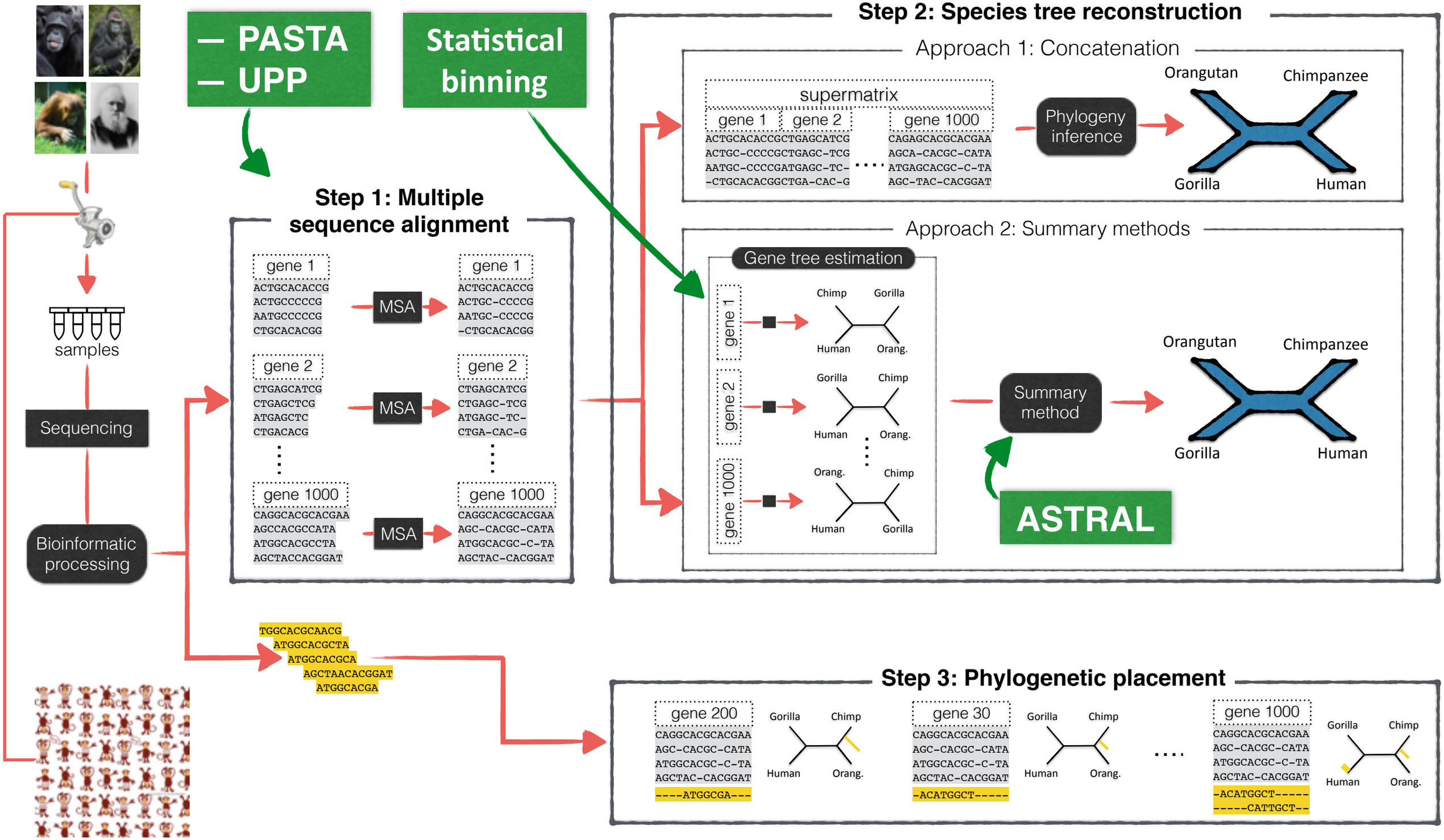
Phylogeny reconstruction pipeline



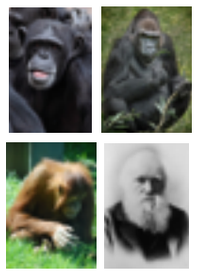
Phylogeny reconstruction pipeline



Phylogeny reconstruction pipeline



Phylogeny reconstruction pipeline



samples

Sequencing

Bioinformatic processing

— PASTA
— UPP

Statistical binning

Step 1: Multiple sequence alignment

gene 1
ACTGCACACCG
ACTGCCCCCG
AATGCCCCCG
CTGCACACGG

MSA

gene 1
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG

gene 2
CTGAGCATCG
CTGAGCTCG
ATGAGCTC
CTGACACG

MSA

gene 2
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G

gene 1000
CAGGCACGCACGAA
AGCCACGCCATA
ATGGCACGCCTA
AGCTACCACGGAT

MSA

gene 1000
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT

Step 2: Species tree reconstruction

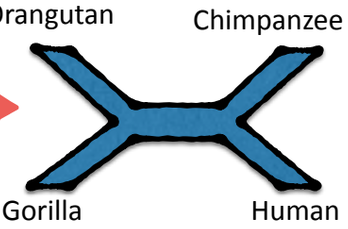
Approach 1: Concatenation

supermatrix

gene 1 | gene 2 | ... | gene 1000

ACTGCACACCGCTGAGCATCG ... CAGAGCACGCACGAA
ACTGC-CCCCGCTGAGC-TCG ... AGCA-CACGC-CATA
AATGC-CCCCGATGAGC-TC- ... ATGAGCACGC-C-TA
-CTGCACACGGCTGA-CAC-G ... AGC-TAC-CACGGAT

Phylogeny inference



Approach 2: Summary methods

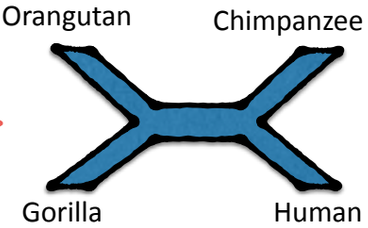
Gene tree estimation

gene 1 → Chimp, Gorilla | Human, Orang.

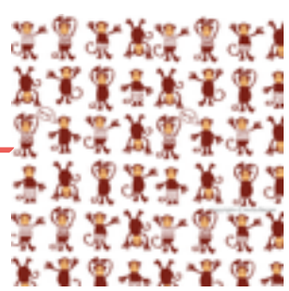
gene 2 → Gorilla, Chimp | Human, Orang.

gene 1000 → Orang., Chimp | Human, Gorilla

Summary method



ASTRAL



TGGCACGCAACG
ATGGCACGCTA
ATGGCACGCA
AGCTAACACGGAT
ATGGCACGA

— SEPP
— TIPP

Step 3: Phylogenetic placement

gene 200
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT
----ATGGCGA---

Gorilla, Chimp | Human, Orang.

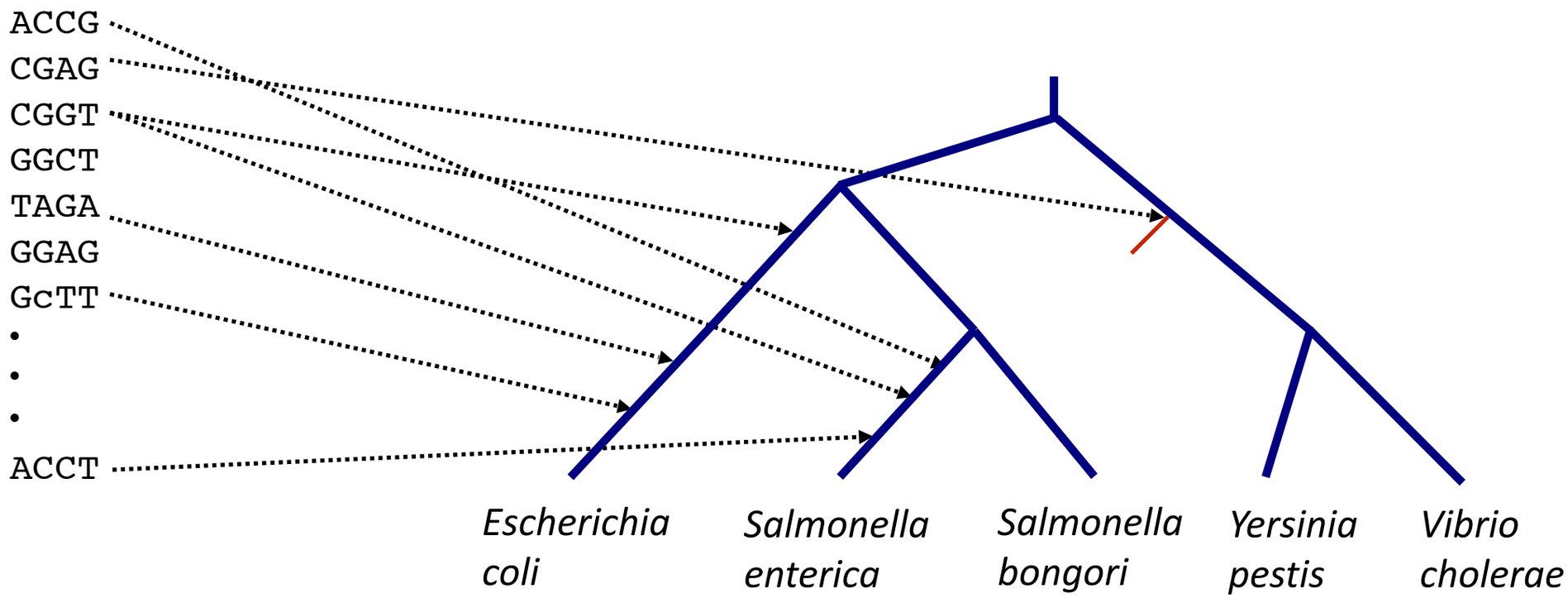
gene 30
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT
-ACATGGCT-----

Gorilla, Chimp | Human, Orang.

gene 1000
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT
-ACATGGCT-----
-----CATTGCT--

Gorilla, Chimp | Human, Orang.

Microbiome analyses using evolutionary trees



Fragmentary metagenomic reads

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

Place each fragmentary read independently on a *reference tree* of known sequences

Phylogenetic placement

- **Input**:
  - A backbone multiple sequence alignment for a marker gene, including sequences from known species
  - A backbone ML phylogenetic tree, corresponding to the backbone alignment
  - A collection of (fragmentary, error-prone) query sequences

Phylogenetic placement

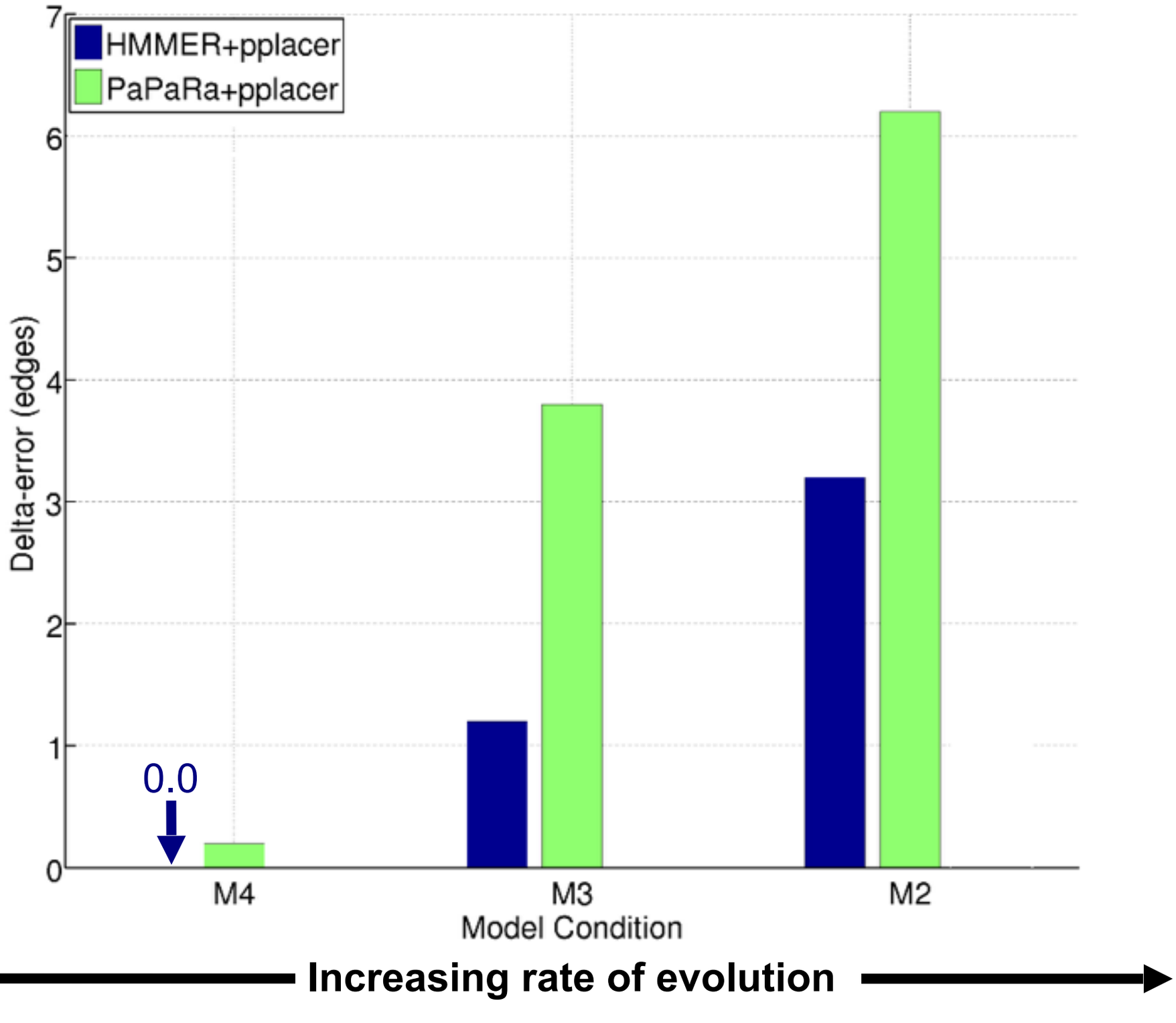
- **Input**:
  - A backbone multiple sequence alignment for a marker gene, including sequences from known species
  - A backbone ML phylogenetic tree, corresponding to the backbone alignment
  - A collection of (fragmentary, error-prone) query sequences
- **Output**: Probabilistic placements of each query sequence on the phylogenetic tree after (locally) aligning the query to the reference

Phylogenetic placement

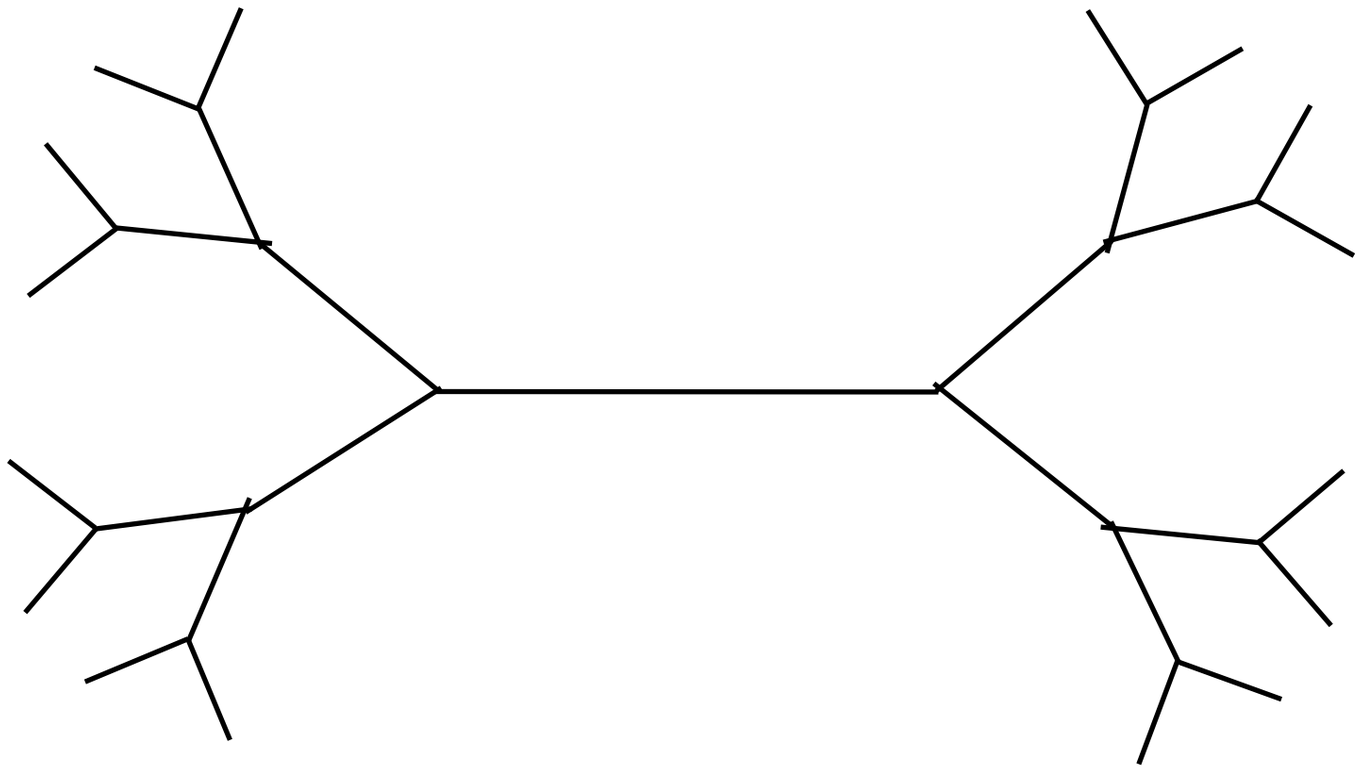
- **Input**:
  - A backbone multiple sequence alignment for a marker gene, including sequences from known species
  - A backbone ML phylogenetic tree, corresponding to the backbone alignment
  - A collection of (fragmentary, error-prone) query sequences
- **Output**: Probabilistic placements of each query sequence on the phylogenetic tree after (locally) aligning the query to the reference
- Tools: — Alignment: HMMER
  — Placement: pplacer (Matsen) and EPA (RAxML)

Phylogenetic placement simulations

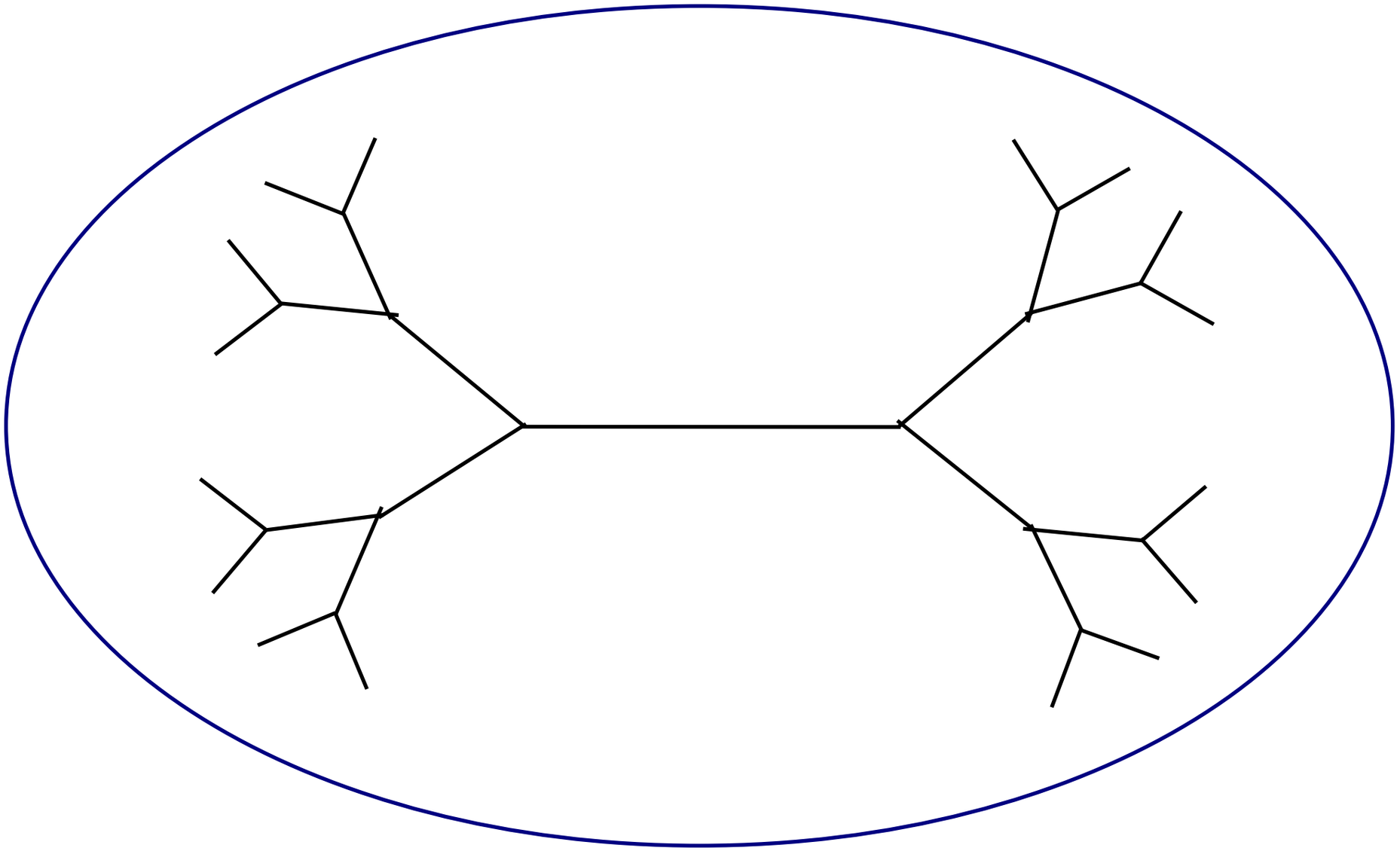
S. Mirarab et al., PSB. (2012).



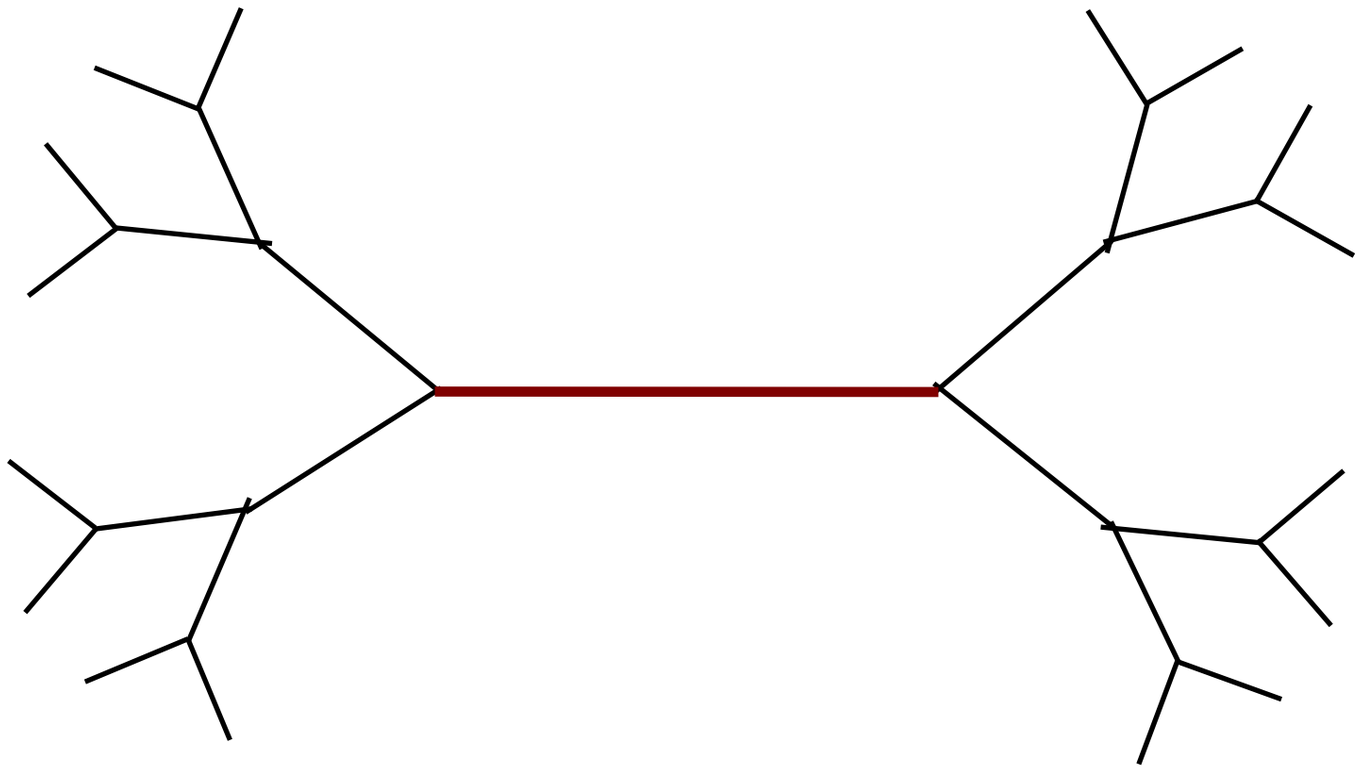
Reference tree



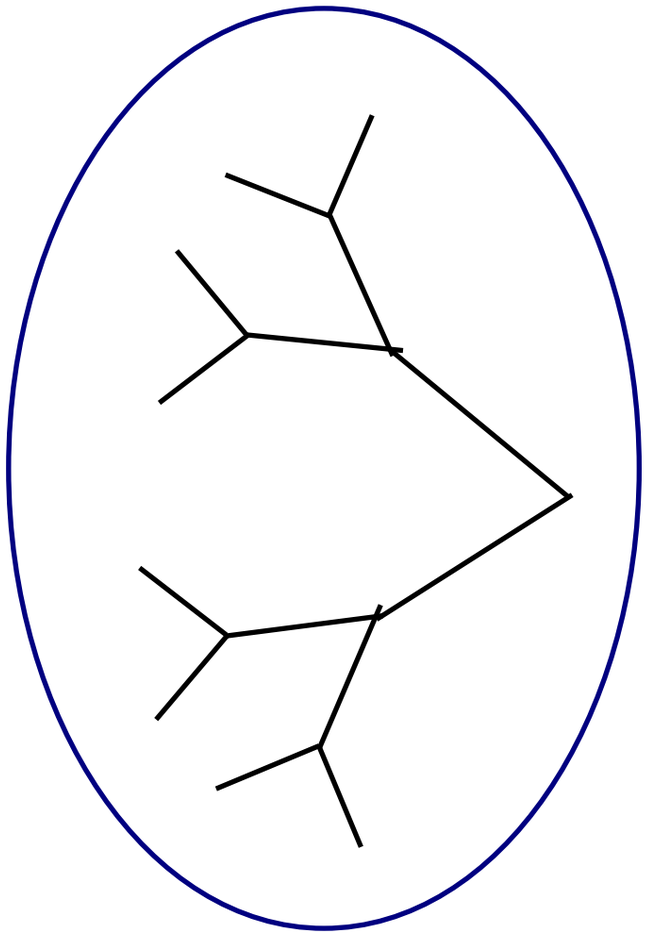
HMM for the alignment step

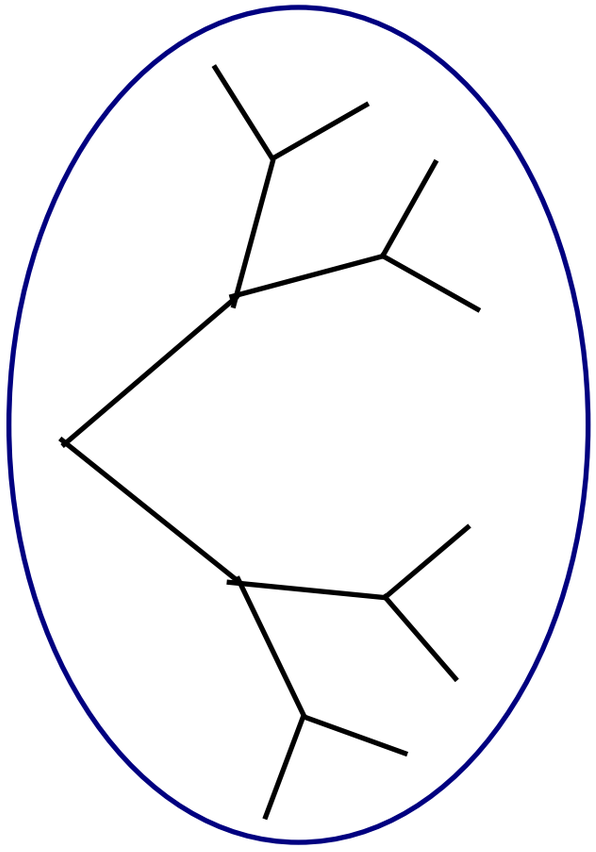


Ensemble of HMMs

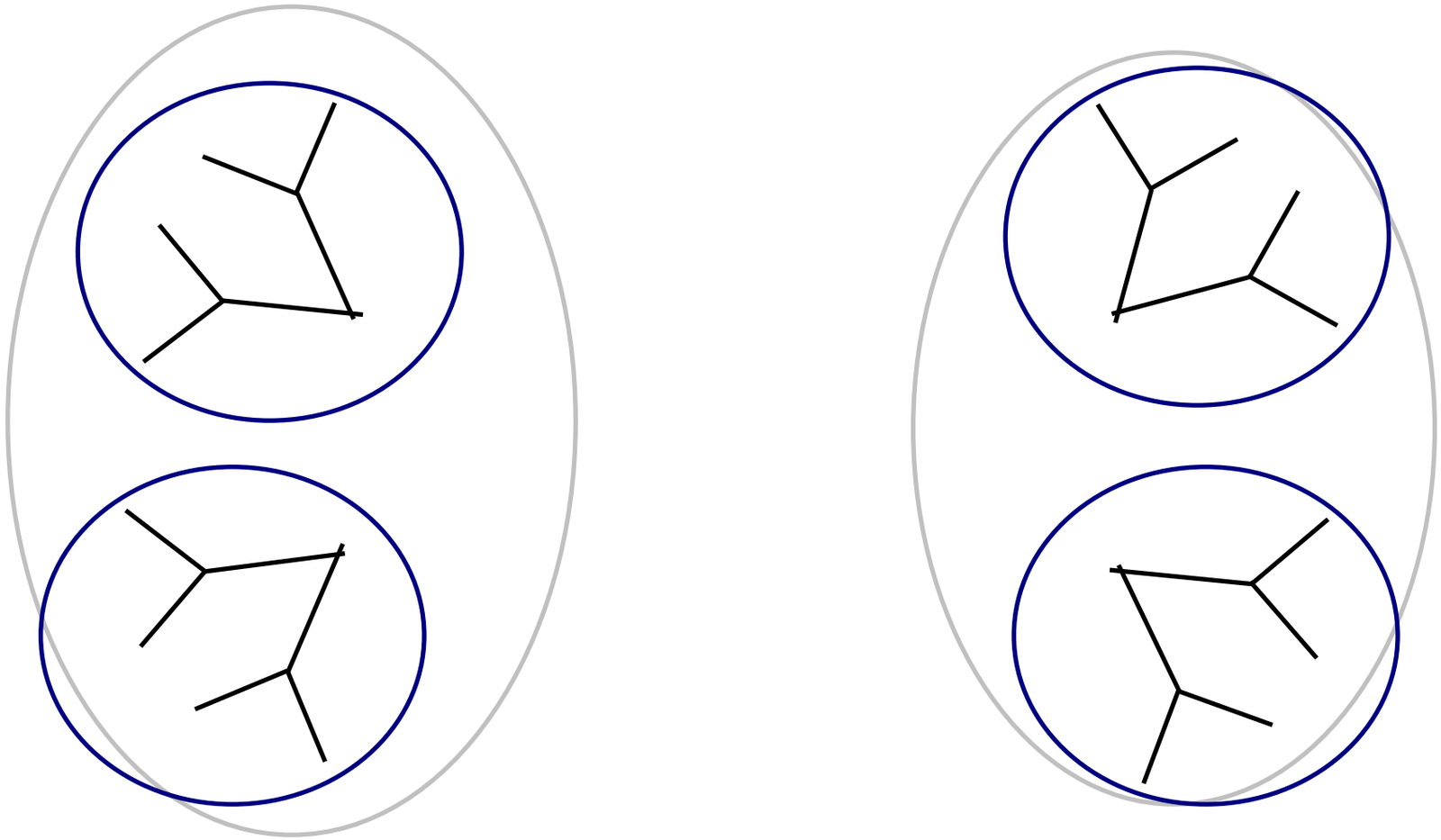


Ensemble of HMMs





Ensemble of HMMs



SATe-Enabled Phylogenetic Placement (SEPP)

Step 1: Align each query sequence to the backbone alignment

- Use an ensemble of disjoint HMMs instead of using a single HMM to improve accuracy.
- The ensemble is created based on the reference tree such that each model better captures details of a part of a tree

S. Mirarab et al., PSB. (2012).

SATe-Enabled Phylogenetic Placement (SEPP)

Step 1: Align each query sequence to the backbone alignment

- Use an ensemble of disjoint HMMs instead of using a single HMM to improve accuracy.
- The ensemble is created based on the reference tree such that each model better captures details of a part of a tree

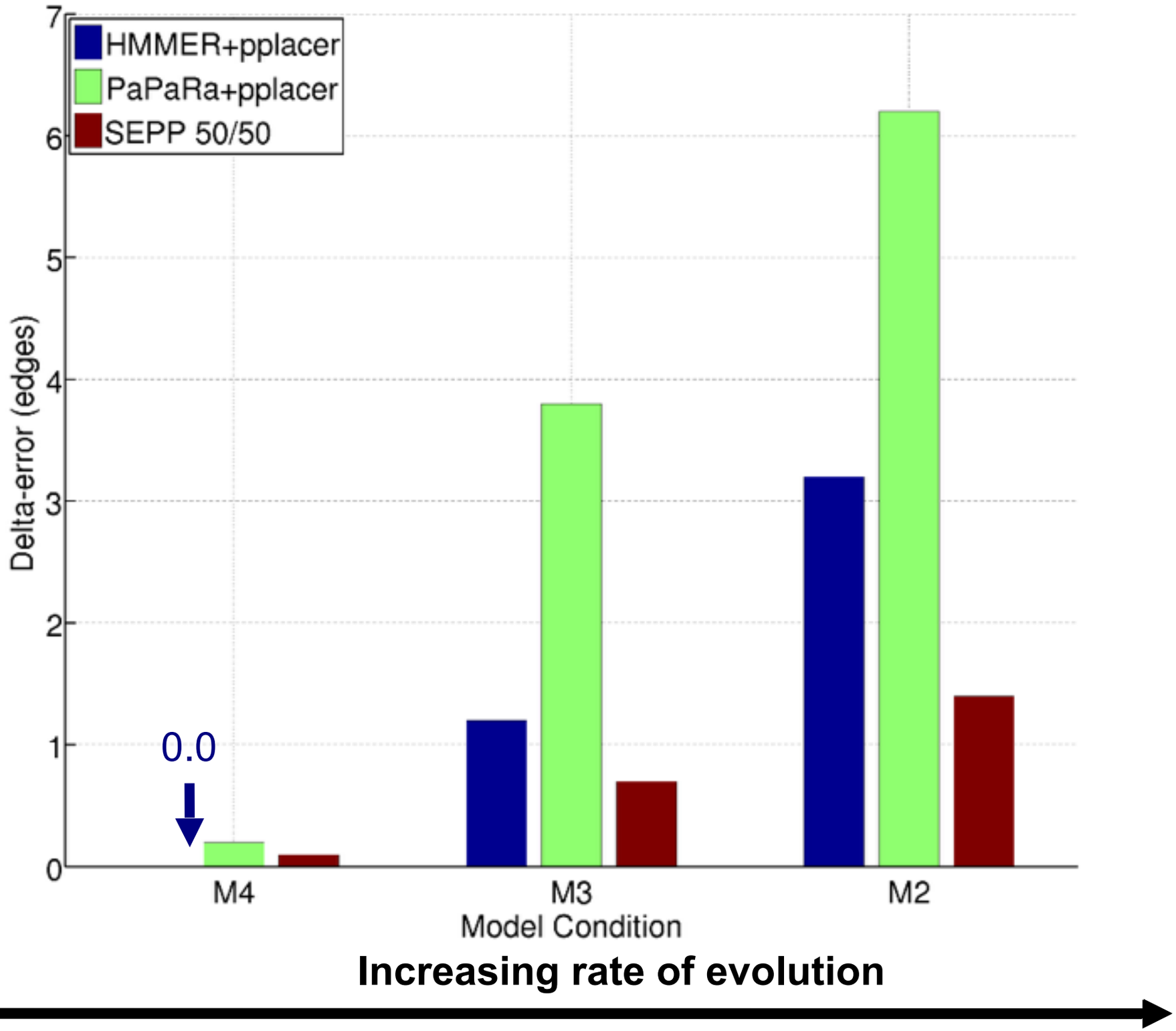
Step 2: Place each query sequence into the backbone tree, using extended alignment

- Use divide-and-conquer on the backbone tree to improve scalability to reference trees with tens of thousands of leaves

S. Mirarab et al., PSB. (2012).

SEPP on simulated data

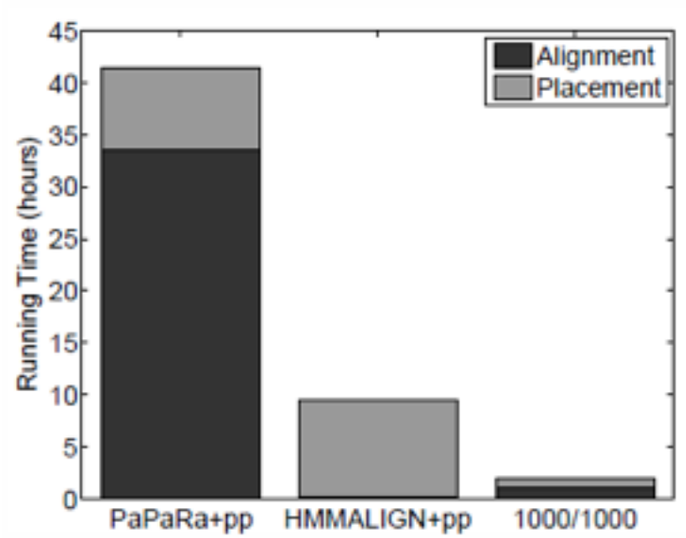
S. Mirarab et al., PSB. (2012).



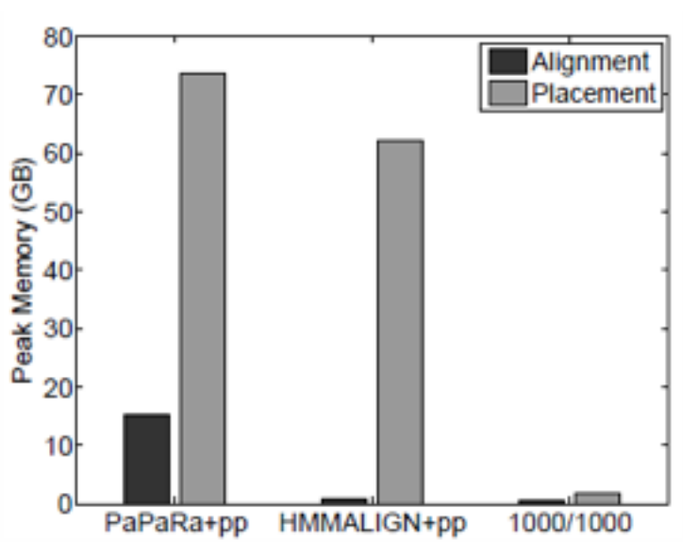
SEPP on large 16S references

Simulations:

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



Running time

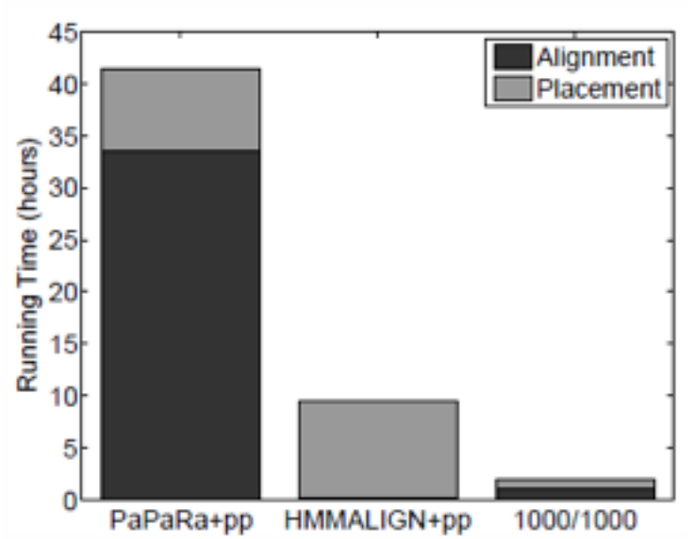


Memory

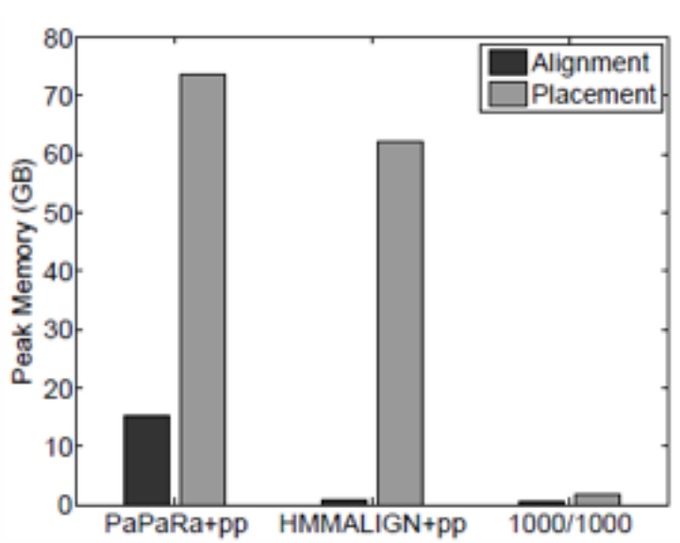
SEPP on large 16S references

Simulations:

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



Running time



Memory

Real data (with Rob Knight's lab):

- **EMP:** placing ~300,000 fragments on the greengenes reference tree with 203,452 sequences
  8 hours (16 cores)

- **AG:** placing ~40,000 fragments on the greengenes reference tree with 203,452 sequences
  10 minutes (16 cores)

Taxonomic Profiling

- **Input**:
  - Reference multiple sequence alignments for a collection of marker genes, each including sequenced species
  - Reference trees for marker genes. We force trees to be compatible with the taxonomy (not necessary).
  - A metagenomic sample: a collection of fragmentary reads from many species with different abundances

Taxonomic Profiling

- **Input**:
  - Reference multiple sequence alignments for a collection of marker genes, each including sequenced species
  - Reference trees for marker genes. We force trees to be compatible with the taxonomy (not necessary).
  - A metagenomic sample: a collection of fragmentary reads from many species with different abundances
- **Output**:
  - The taxonomic profile of the sample

Genus	%
Pseudomonas	16.6
Campylobacter	8.9
Streptomyces	7.6
Pasteurella	6.4
Clostridium	5.1
Alcanivorax	4.5
…	
unclassified	1.2

Phylum	%
Proteobacteria	63.1
Actinobacteria	9.6
Firmicutes	9.6
Euryarchaeota	7.6
Cyanobacteria	4.5
Crenarchaeota	3.8
…	
unclassified	0.0

TIPP: Taxon Identification and Phylogenetic Profiling

Step 1: Find fragments that belong to “marker” genes using BLAST (or a new method based on ensembles of HMMs called HIPPI)

Nguyen et al., Bioinformatics (2014)

TIPP: Taxon Identification and Phylogenetic Profiling

Step 1: Find fragments that belong to “marker” genes using BLAST (or a new method based on ensembles of HMMs called HIPPI)

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

- Take into account uncertainty: for each read, use several alignments and placements on the tree (as many as needed to reach a predefined level of statistical support for both the alignment and the placement)

Nguyen et al., Bioinformatics (2014)

TIPP: Taxon Identification and Phylogenetic Profiling

Step 1: Find fragments that belong to "marker" genes using BLAST (or a new method based on ensembles of HMMs called HIPPI)

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

- Take into account uncertainty: for each read, use several alignments and placements on the tree (as many as needed to reach a predefined level of statistical support for both the alignment and the placement)

Step 3: Summarize results across different genes to get a taxonomic profile

- Each read contributes to each branch and all branches above it proportionally to the probability that it belongs to that branch

- Results from all genes are simply aggregated as counts

Nguyen et al., Bioinformatics (2014)

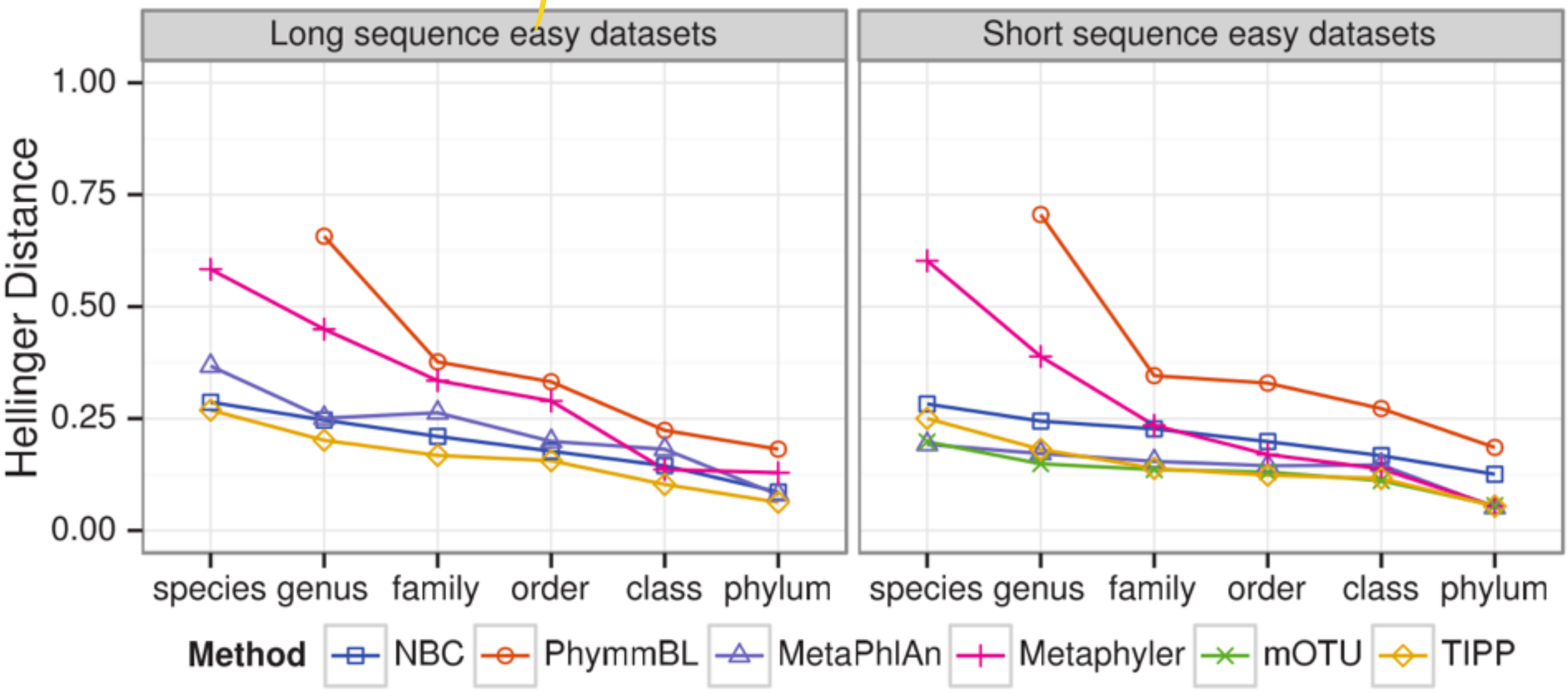
Simulation analysis

- Simulate datasets using MetaSim: vary complexity, sequencing error, length, and whether sequences are novel
- A set of ~30 marker genes (hopefully, single-copy and HGT-free). We filter out all reads from all other genes.
- Measure the distance between the estimated and the true profile

Dataset	Experiment	# Genomes	Complexity	# Reads	Length
MetaPhlAn HC	1&2	100	High	1000000	88 (s)
MetaPhlAn LC	1&2	25	Low	240000	88 (s)
FAMeS HC	1&2	113	High	116771	949 (l)
FAMeS MC	1&2	113	Medium	114457	969 (l)
FAMeS LC	1&2	113	Low	97495	951 (l)
FACS HC-454	1&2	19	High	26984	268 (l)
TIPP FACS HC-Illum	1&2	19	High	300000	100 (s)
WebCarma-454	1&2	25	High	25000	265 (l)
TIPP WebCarma-Illum	1&2	25	High	300000	100 (s)
TIPP HC novel-Illum	3	100	High	1000000	100 (s)
TIPP LC novel-Illum	3	100	Low	1000000	100 (s)
TIPP HC novel-454	3	100	High	1000000	269 (l)
TIPP LC novel-454	3	100	Low	1000000	269 (l)

Results: known genomes

known genomes, low sequencing error

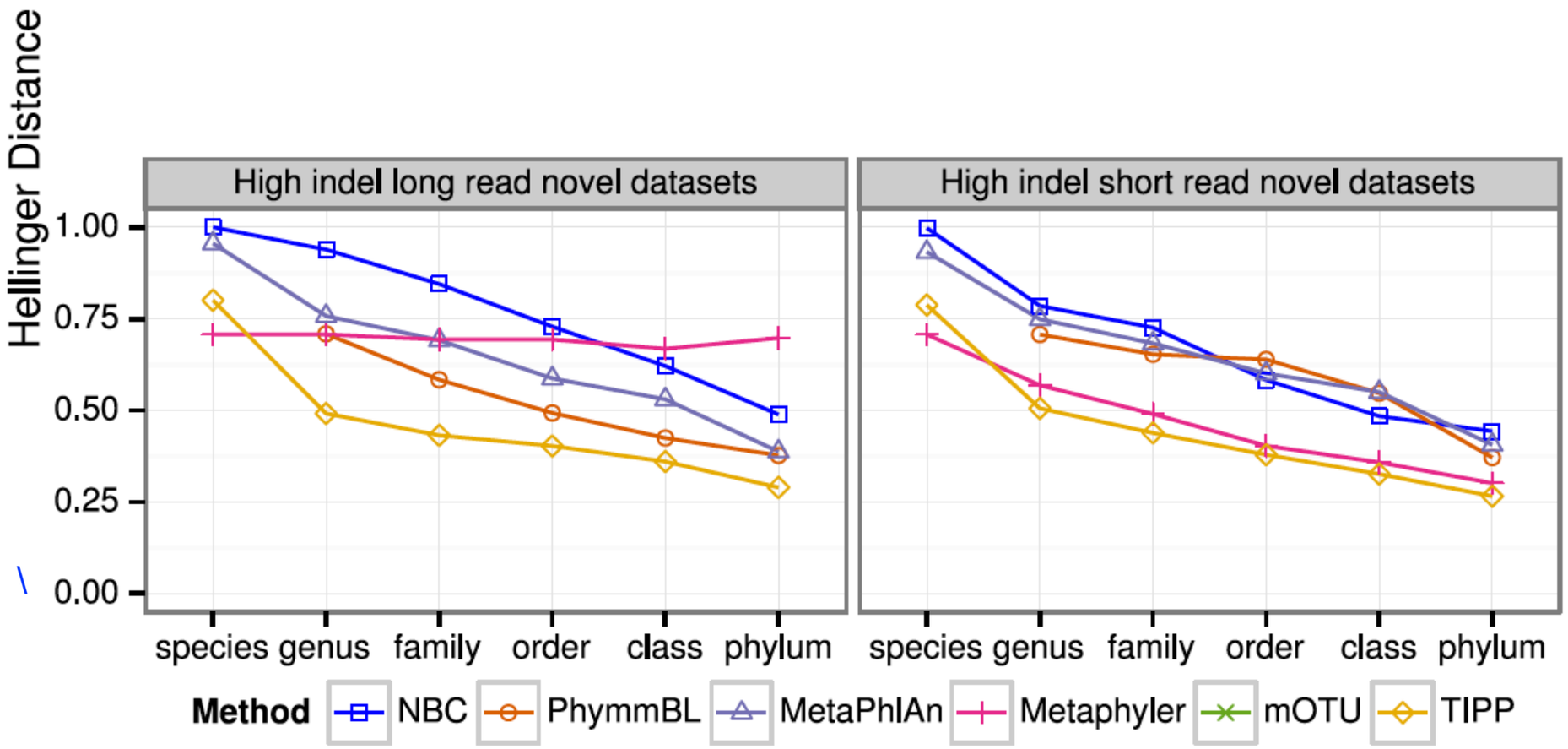


PhymmBL (Brady & Salzberg, Nature Methods 2009)
NBC (Rosen, et al., Bioinformatics 2011)
MetaPhyler (Liu et al., BMC Genomics 2011)

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

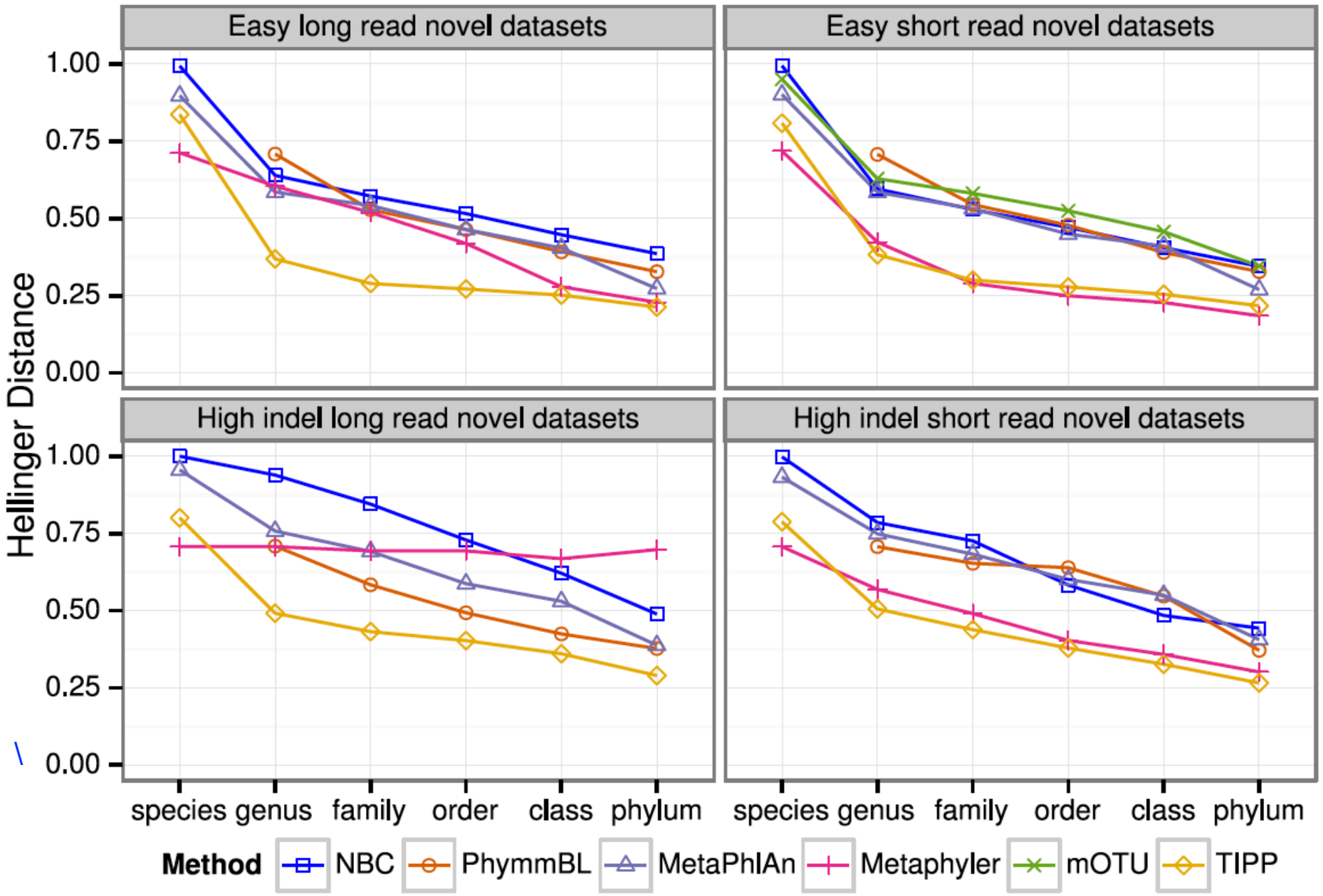
Results:“novel” genomes

Novel —> reads are simulated from a species not used in training any of the methods

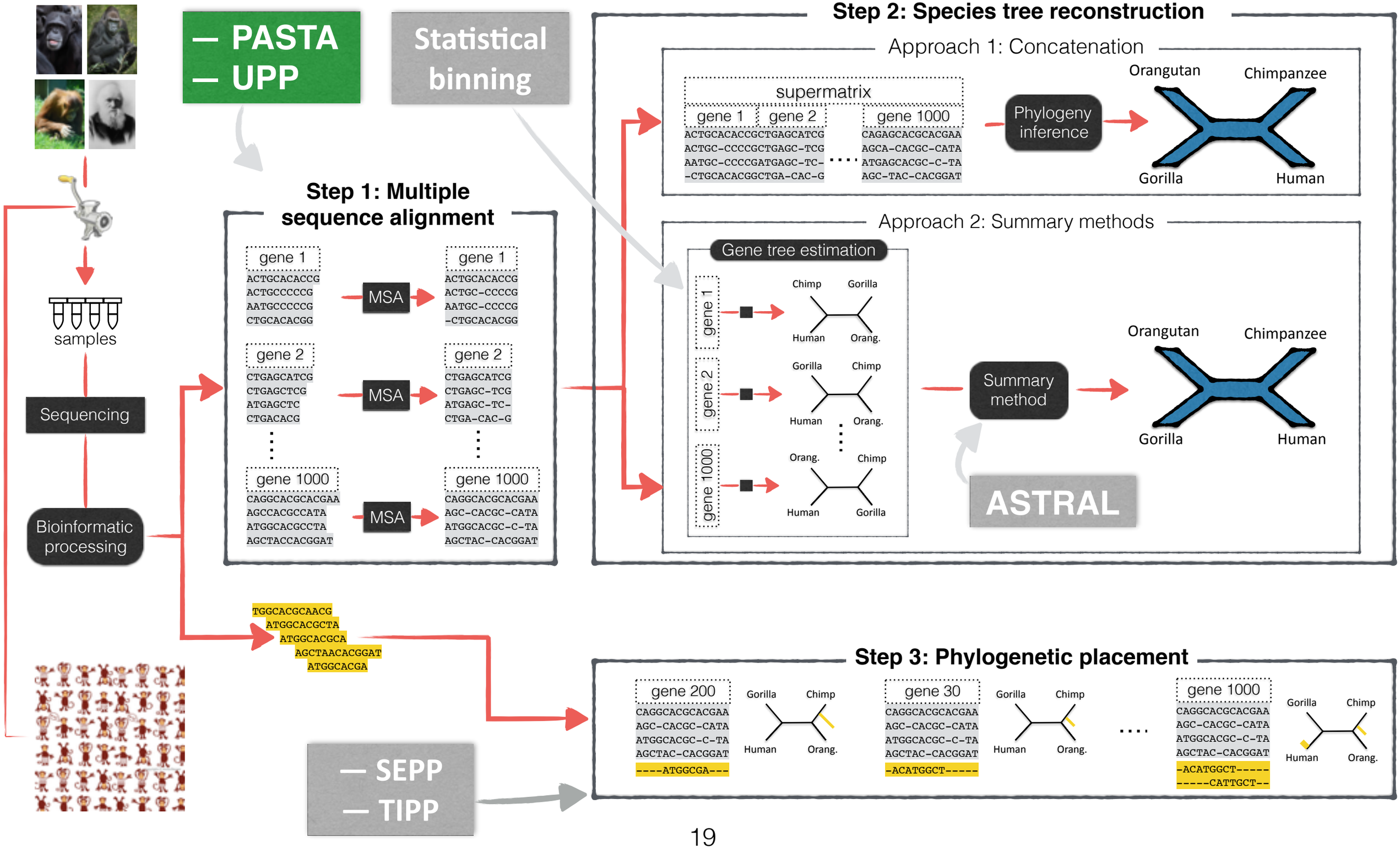


Results:“novel” genomes

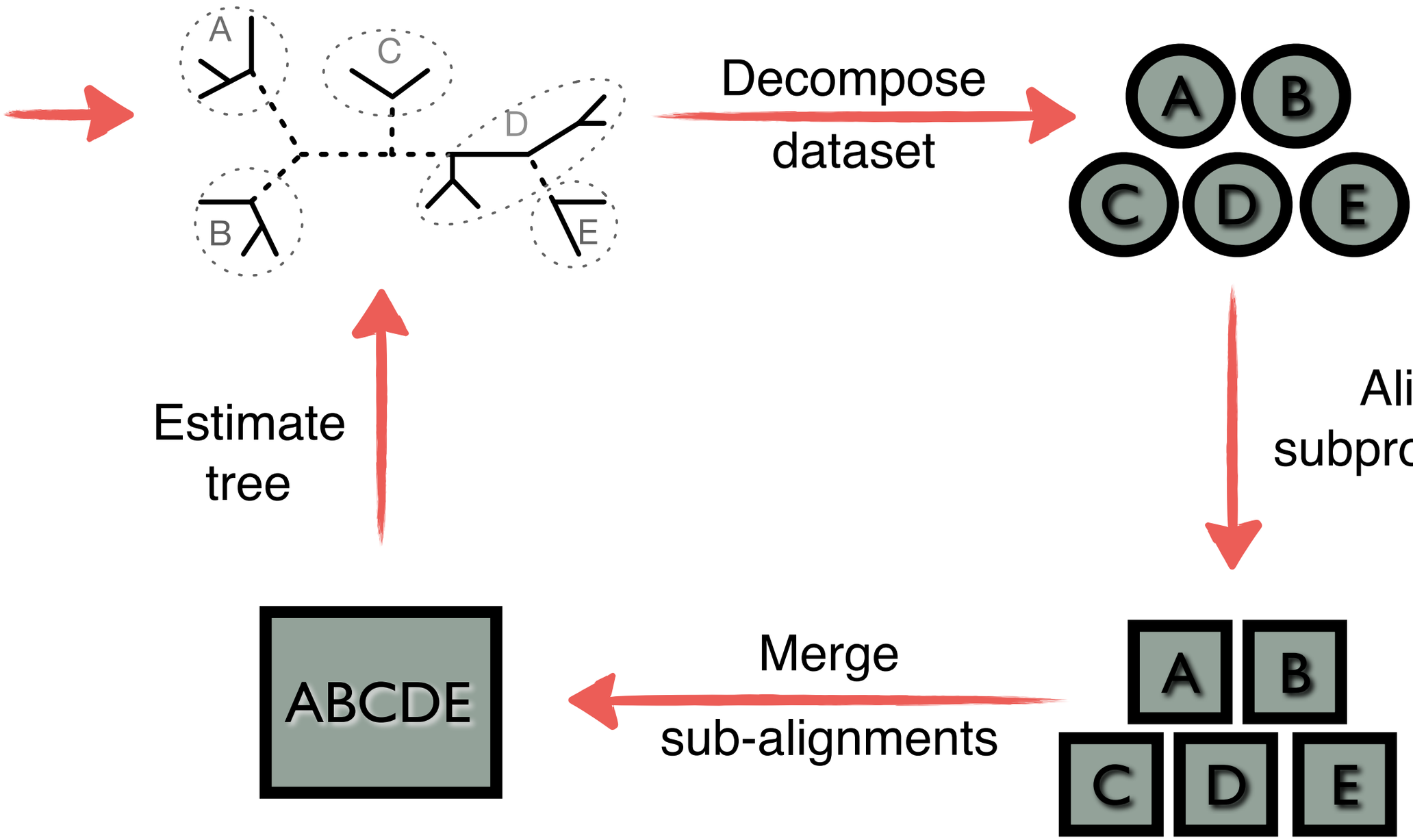
Novel —> reads are simulated from a species not used in training any of the methods



Multi-gene phylogeny reconstruction

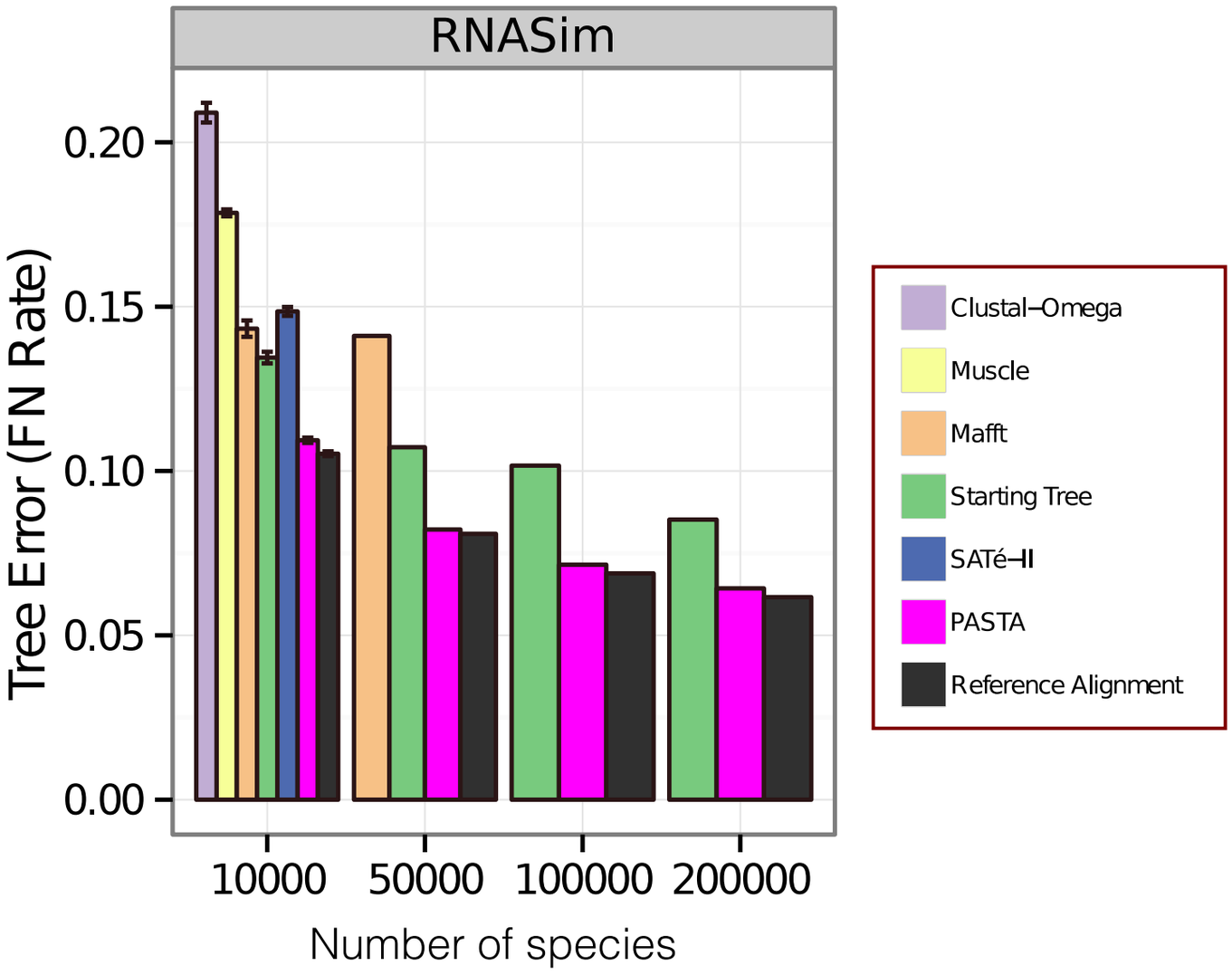


PASTA: Iterative divide-and-conquer alignment and tree estimation



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

Tree topological accuracy

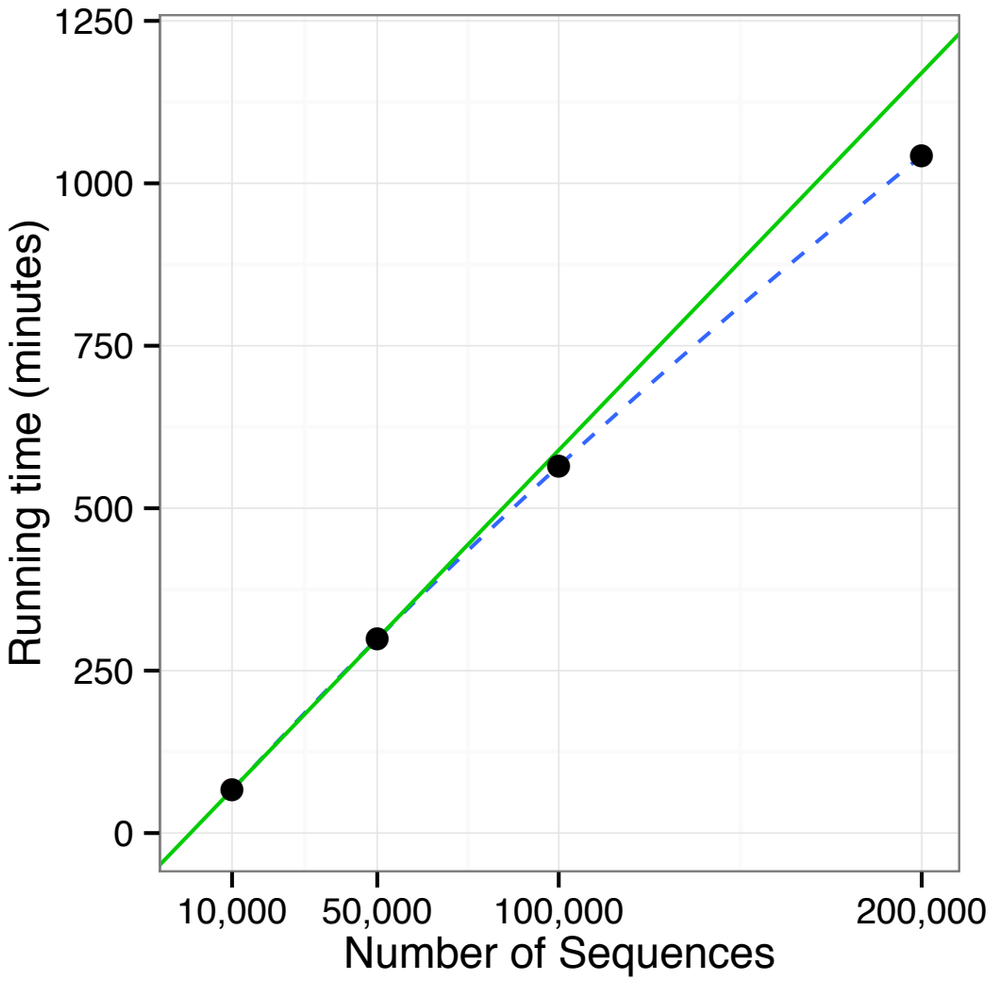


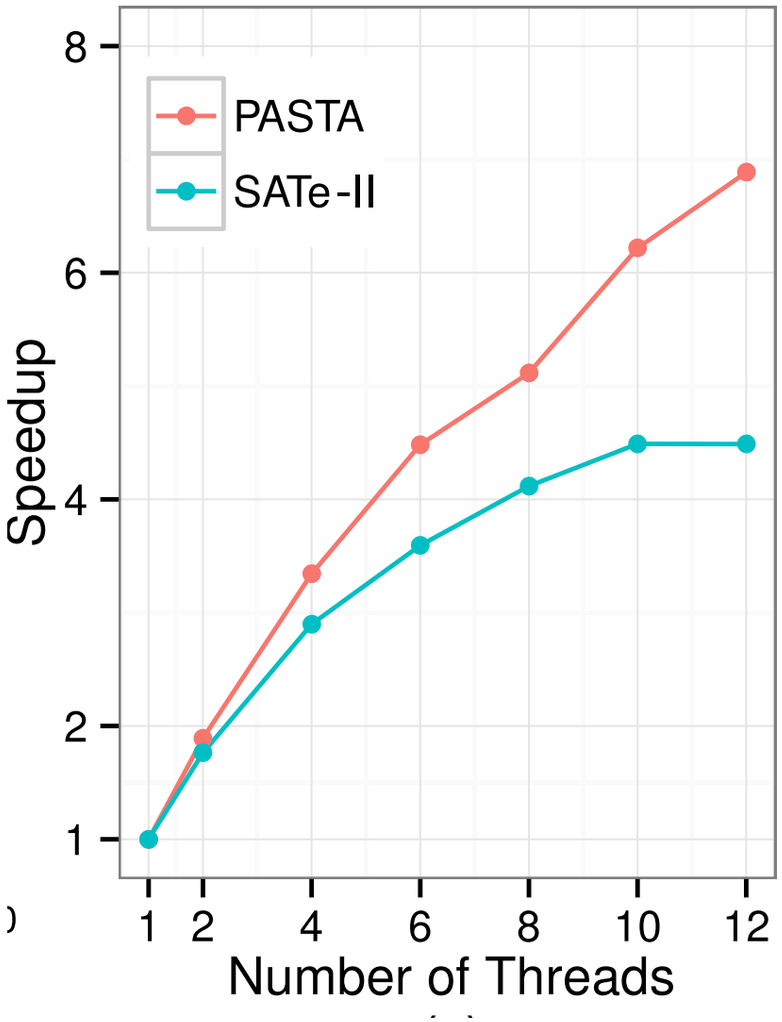
1 million sequences:

- 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





PASTA on Greengenes

- Testing the 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: Speed: (16 cores)
  97% tree (99,322 leaves): 28 hours
  99% tree (203,452 leaves): 49 hours

With Daniel McDonald (Knight's lab) and Uyen Mai

Software availability

- 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)

- Species tree estimation:

  - Statistical binning: https://github.com/smirarab/binning

  - ASTRAL: github.com/smirarab/ASTRAL

Acknowledgments

- Nam-Phuong Nguyen
- Tandy Warnow's lab:
  - Mike Nute
- Mihai Pop's lab:
  - Bo Liu
- Rob Knight's lab
  - Daniel McDownload
- Mirarab lab
  - Uyen Mai

HHMI
HOWARD HUGHES MEDICAL INSTITUTE

NSF