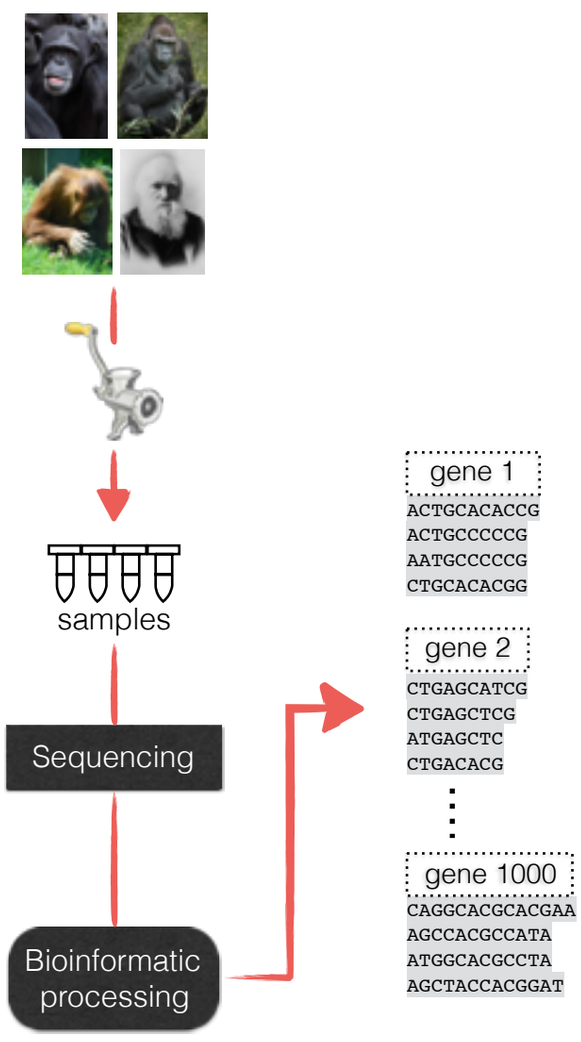
Phylogenetic methods for taxonomic profiling

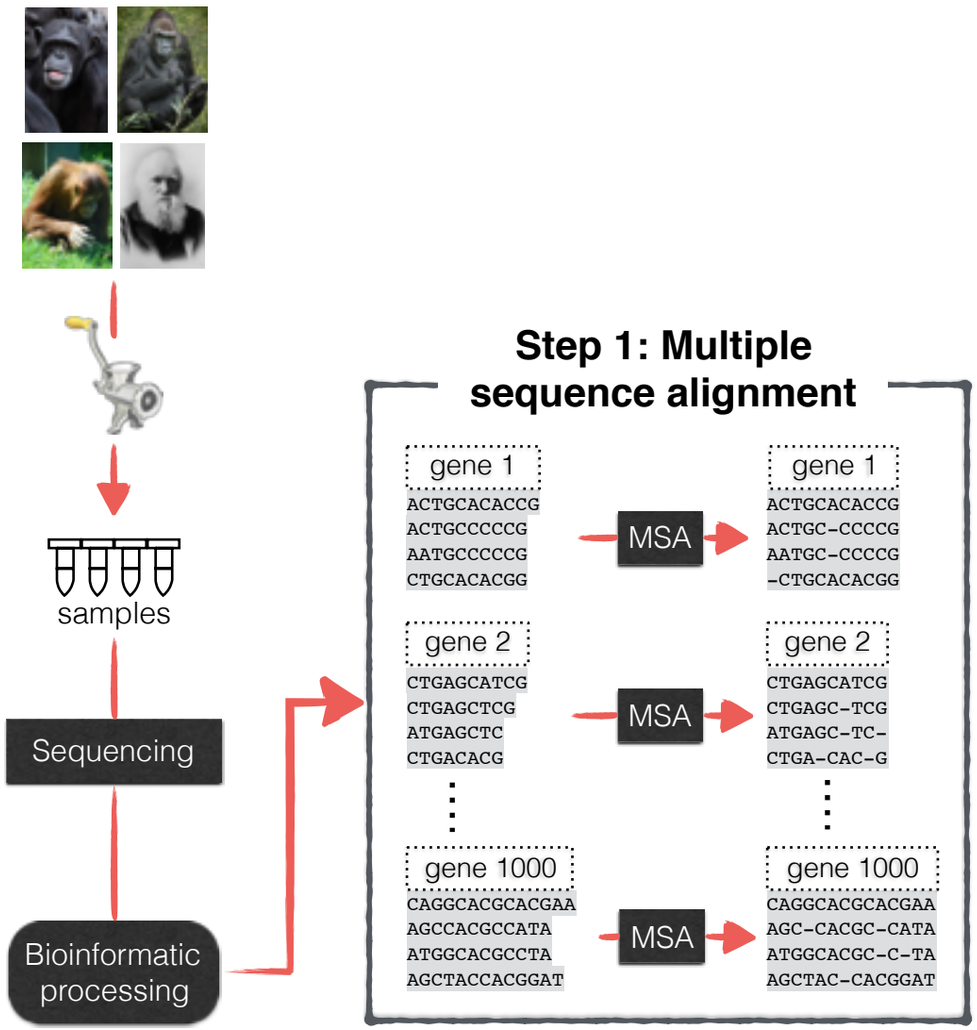
Siavash Mirarab
University of California at San Diego (UCSD)

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

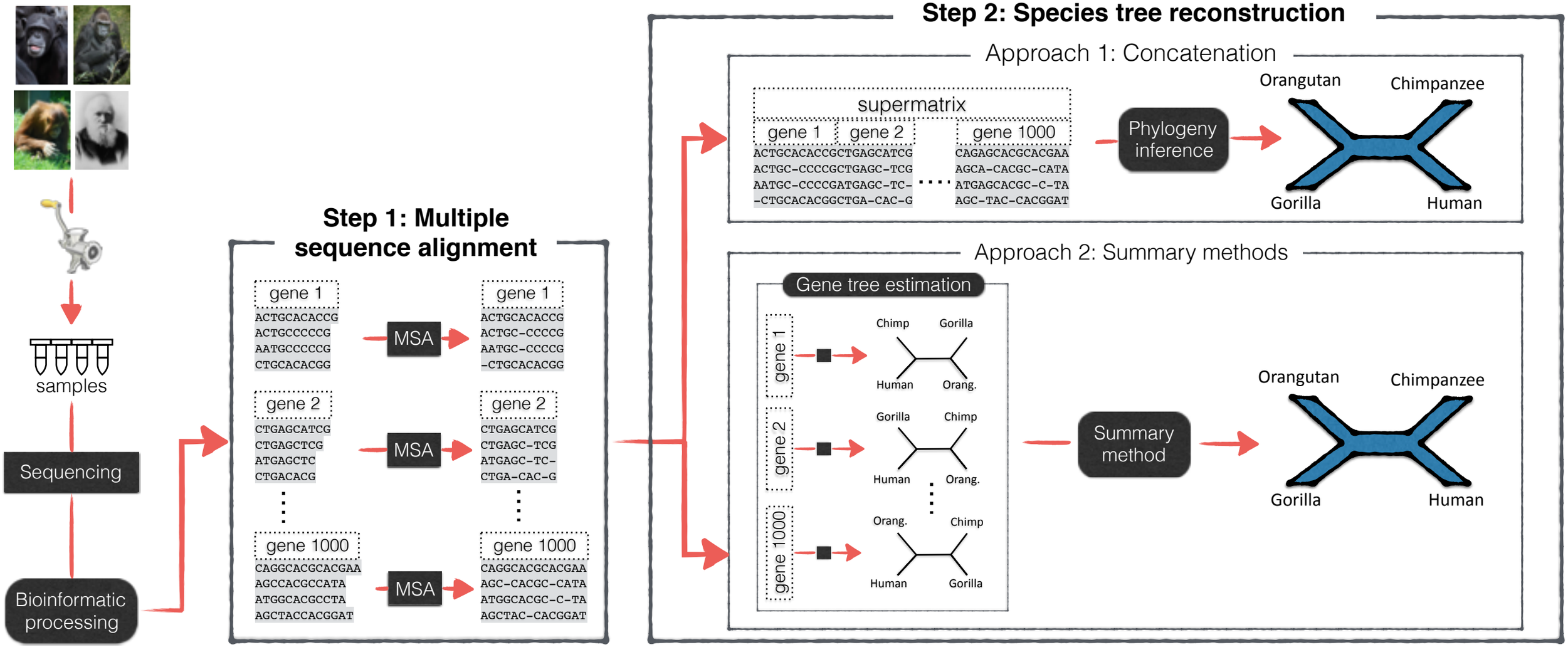
Phylogeny reconstruction pipeline



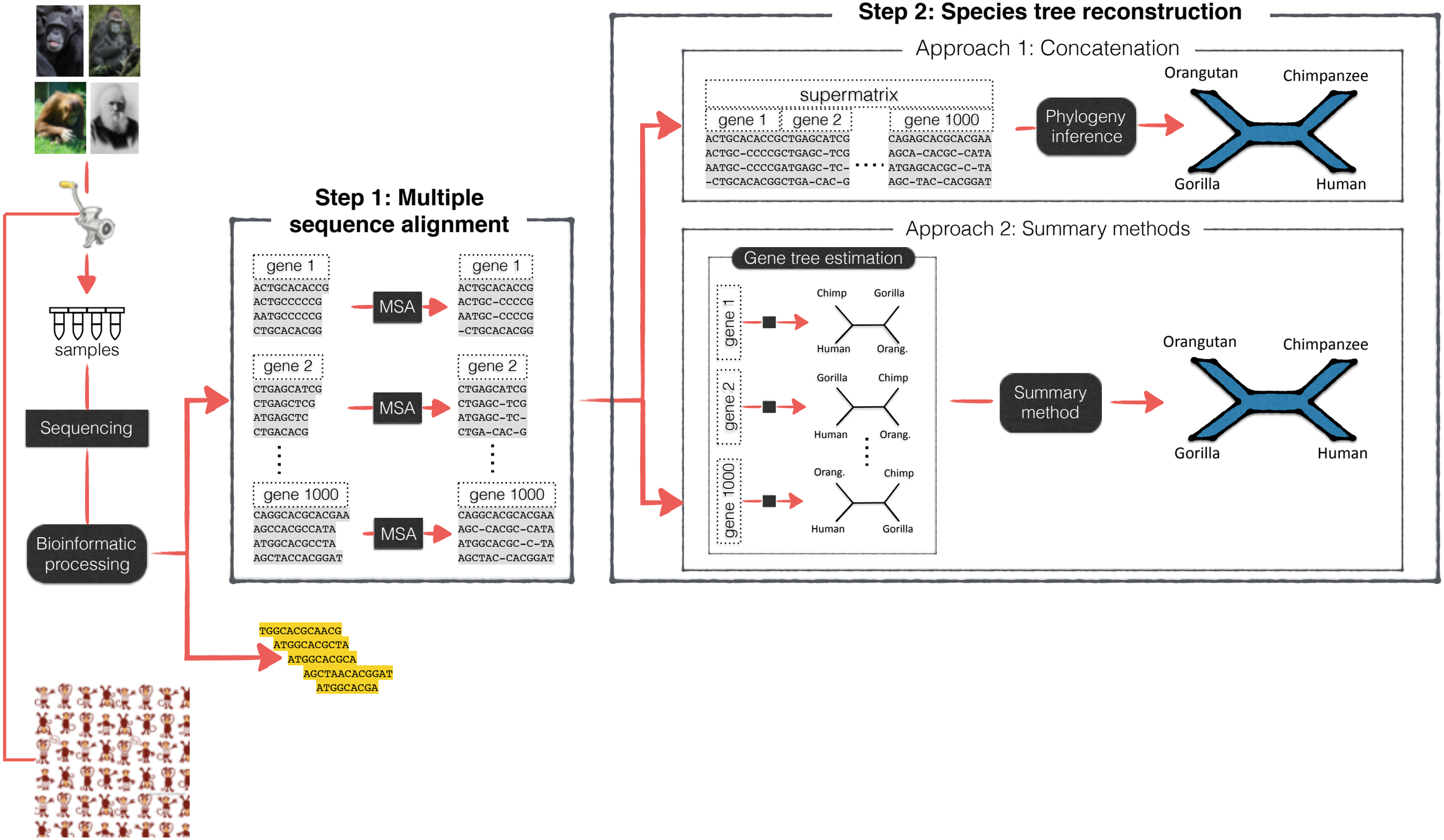
Phylogeny reconstruction pipeline



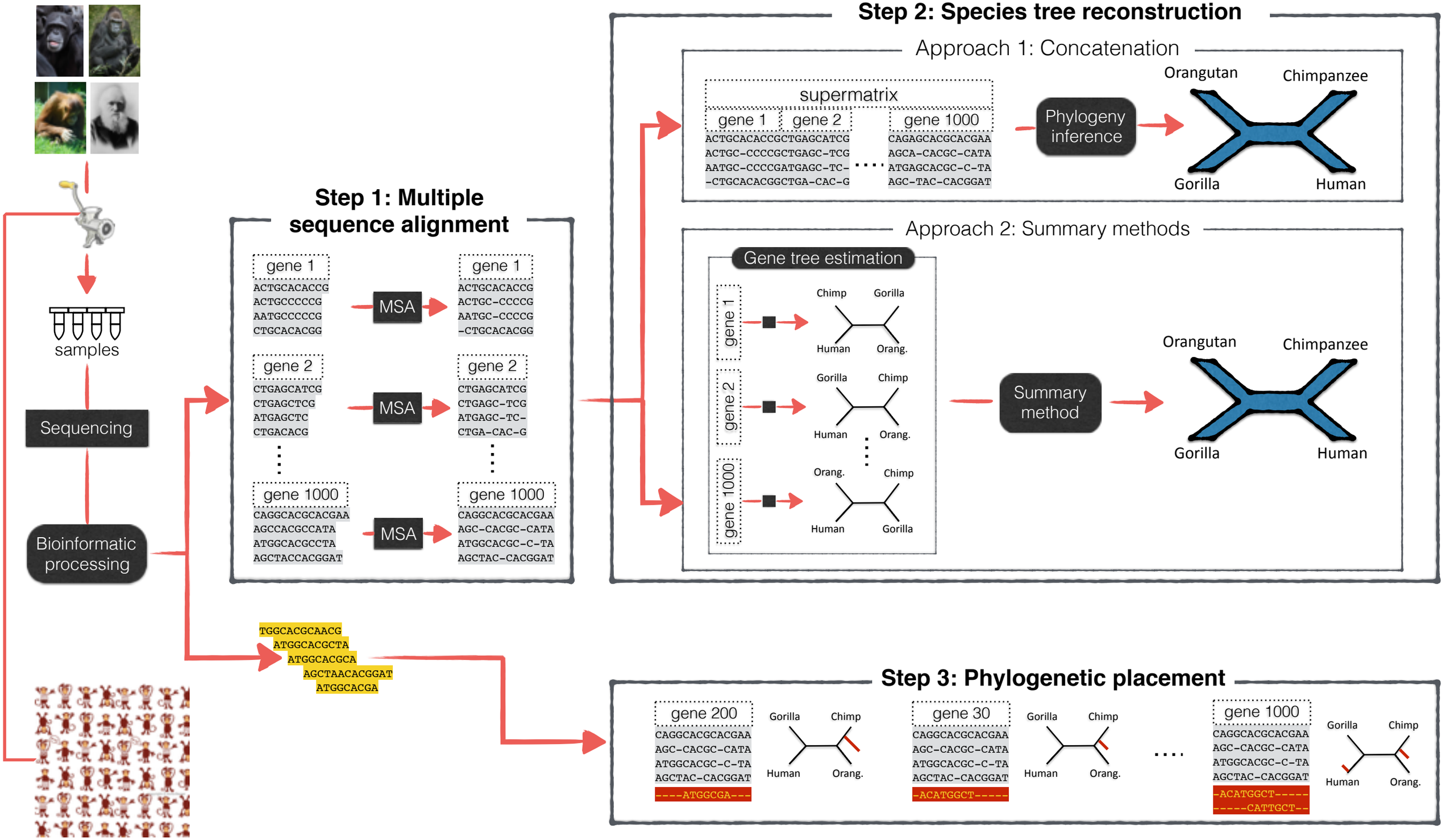
Phylogeny reconstruction pipeline



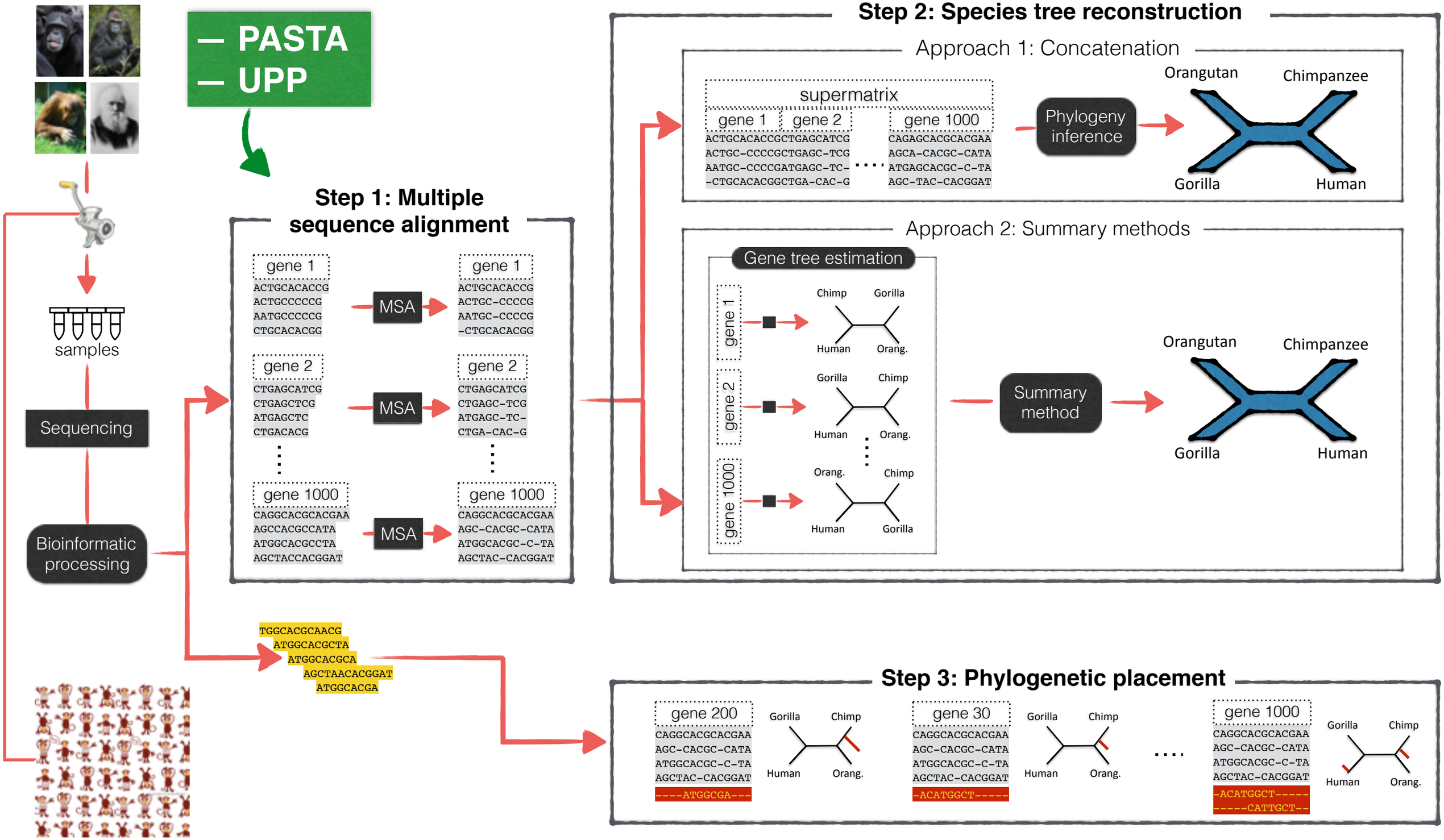
Phylogeny reconstruction pipeline



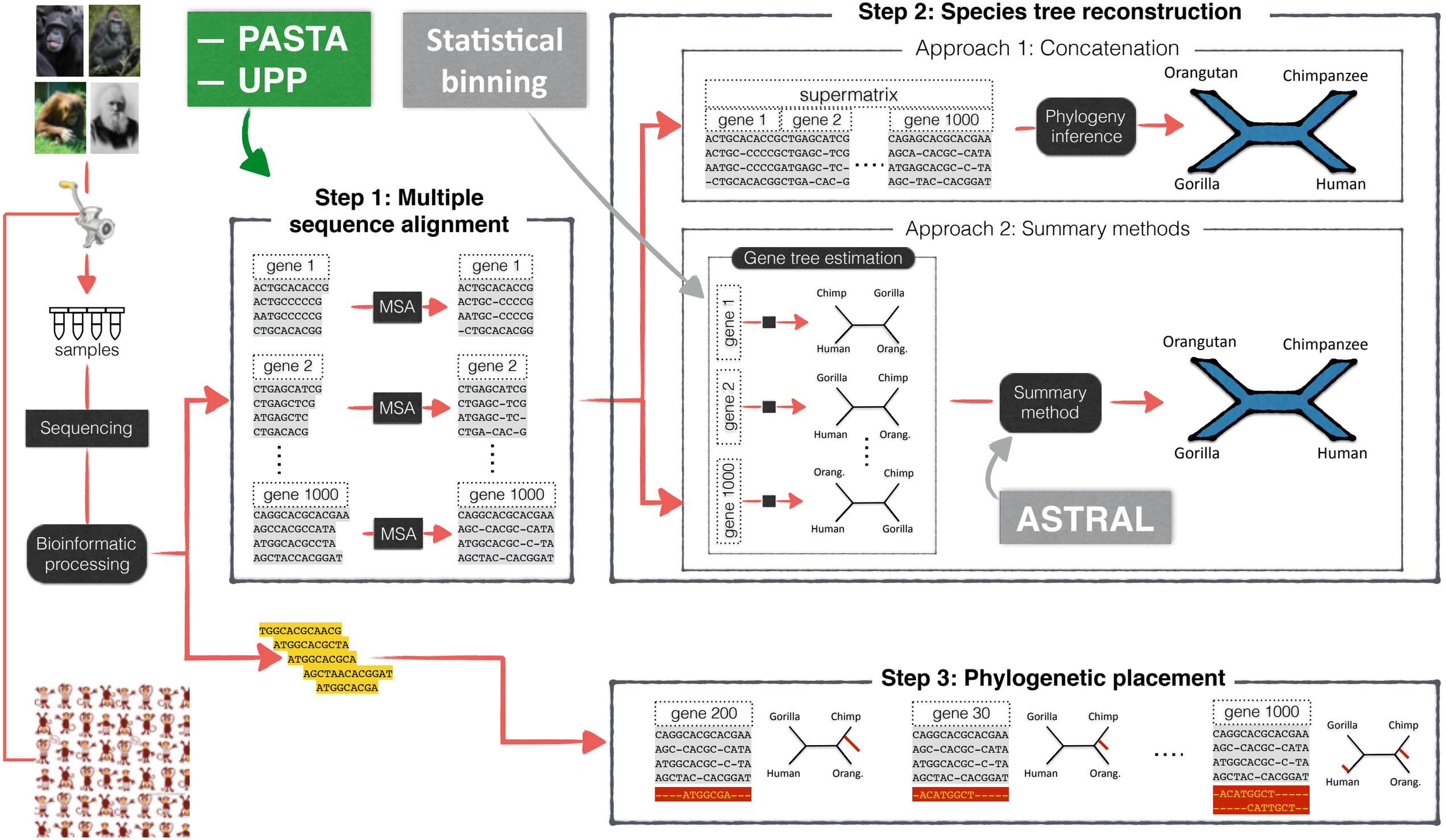
Phylogeny reconstruction pipeline



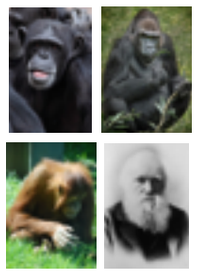
Phylogeny reconstruction pipeline



Phylogeny reconstruction pipeline



Phylogeny reconstruction pipeline



samples

Sequencing

Bioinformatic processing

— PASTA
— UPP

Statistical binning

Step 1: Multiple sequence alignment

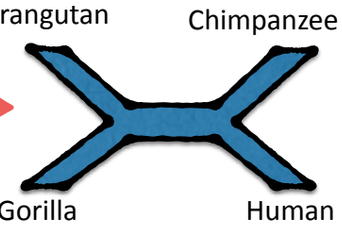
gene 1		MSA		gene 1
ACTGCACACCG				ACTGCACACCG
ACTGCCCCCG				ACTGC-CCCCG
AATGCCCCCG				AATGC-CCCCG
CTGCACACGG				-CTGCACACGG
gene 2				**gene 2**
CTGAGCATCG		MSA		CTGAGCATCG
CTGAGCTCG				CTGAGC-TCG
ATGAGCTC				ATGAGC-TC-
CTGACACG				CTGA-CAC-G
gene 1000		MSA		**gene 1000**
CAGGCACGCACGAA				CAGGCACGCACGAA
AGCCACGCCATA				AGC-CACGC-CATA
ATGGCACGCCTA				ATGGCACGC-C-TA
AGCTACCACGGAT				AGCTAC-CACGGAT

Step 2: Species tree reconstruction

Approach 1: Concatenation

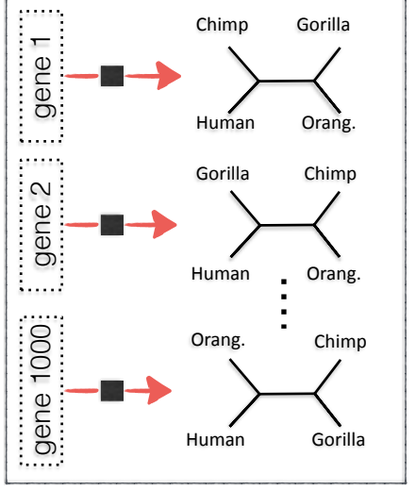
supermatrix			
gene 1	gene 2	⋯	gene 1000
ACTGCACACCGCTGAGCATCG			CAGAGCACGCACGAA
ACTGC-CCCCGCTGAGC-TCG			AGCA-CACGC-CATA
AATGC-CCCCGATGAGC-TC-			ATGAGCACGC-C-TA
-CTGCACACGGCTGA-CAC-G			AGC-TAC-CACGGAT

Phylogeny inference



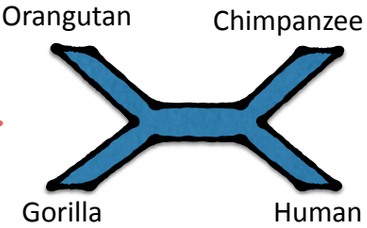
Approach 2: Summary methods

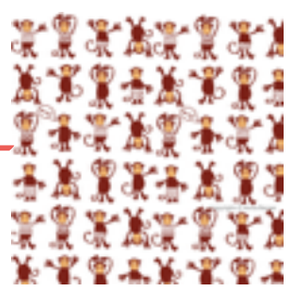
Gene tree estimation



Summary method

ASTRAL



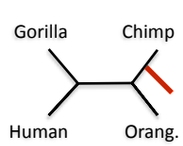


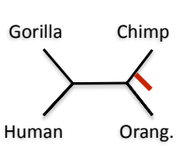
TGGCACGCAACG
ATGGCACGCTA
ATGGCACGCA
AGCTAACACGGAT
ATGGCACGA

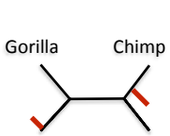
— SEPP
— TIPP

Step 3: Phylogenetic placement

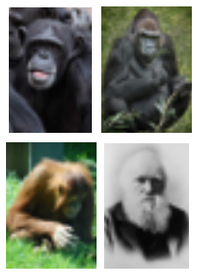
gene 200			gene 30			gene 1000	
CAGGCACGCACGAA			CAGGCACGCACGAA			CAGGCACGCACGAA	
AGC-CACGC-CATA			AGC-CACGC-CATA			AGC-CACGC-CATA	
ATGGCACGC-C-TA			ATGGCACGC-C-TA			ATGGCACGC-C-TA	
AGCTAC-CACGGAT			AGCTAC-CACGGAT			AGCTAC-CACGGAT	
----ATGGCGA---			-ACATGGCT-----			-ACATGGCT-----	
						-----CATTGCT--	







Phylogeny reconstruction pipeline



samples

Sequencing

Bioinformatic processing

— PASTA
— UPP

Step 1: Multiple sequence alignment

gene 1		MSA		gene 1
ACTGCACACCG				ACTGCACACCG
ACTGCCCCCG				ACTGC-CCCCG
AATGCCCCCG				AATGC-CCCCG
CTGCACACGG				-CTGCACACGG
gene 2				**gene 2**
CTGAGCATCG		MSA		CTGAGCATCG
CTGAGCTCG				CTGAGC-TCG
ATGAGCTC				ATGAGC-TC-
CTGACACG				CTGA-CAC-G
gene 1000		MSA		**gene 1000**
CAGGCACGCACGAA				CAGGCACGCACGAA
AGCCACGCCATA				AGC-CACGC-CATA
ATGGCACGCCTA				ATGGCACGC-C-TA
AGCTACCACGGAT				AGCTAC-CACGGAT

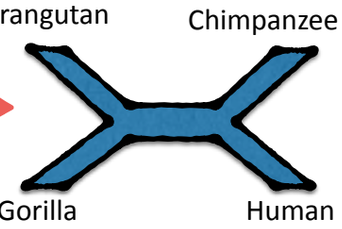
Step 2: Species tree reconstruction

Approach 1: Concatenation

supermatrix

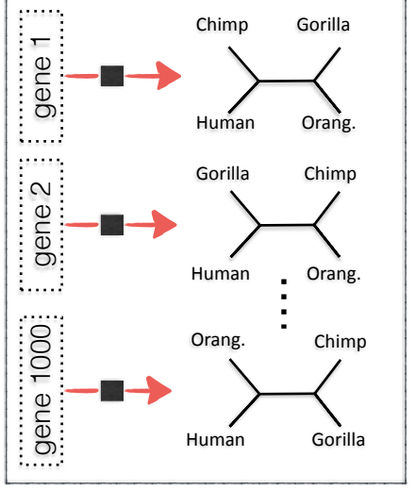
gene 1	gene 2	...	gene 1000
ACTGCACACCGCTGAGCATCG			CAGAGCACGCACGAA
ACTGC-CCCCGCTGAGC-TCG			AGCA-CACGC-CATA
AATGC-CCCCGATGAGC-TC-			ATGAGCACGC-C-TA
-CTGCACACGGCTGA-CAC-G			AGC-TAC-CACGGAT

Phylogeny inference

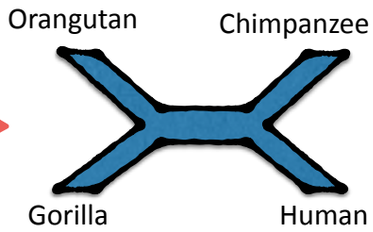


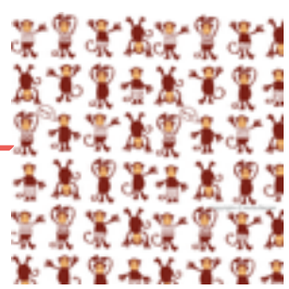
Approach 2: Summary methods

Gene tree estimation



Summary method



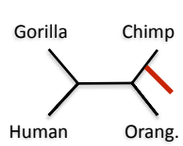


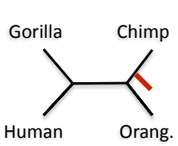
TGGCACGCAACG
ATGGCACGCTA
ATGGCACGCA
AGCTAACACGGAT
ATGGCACGA

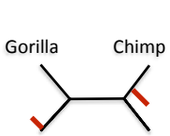
— SEPP
— TIPP

Step 3: Phylogenetic placement

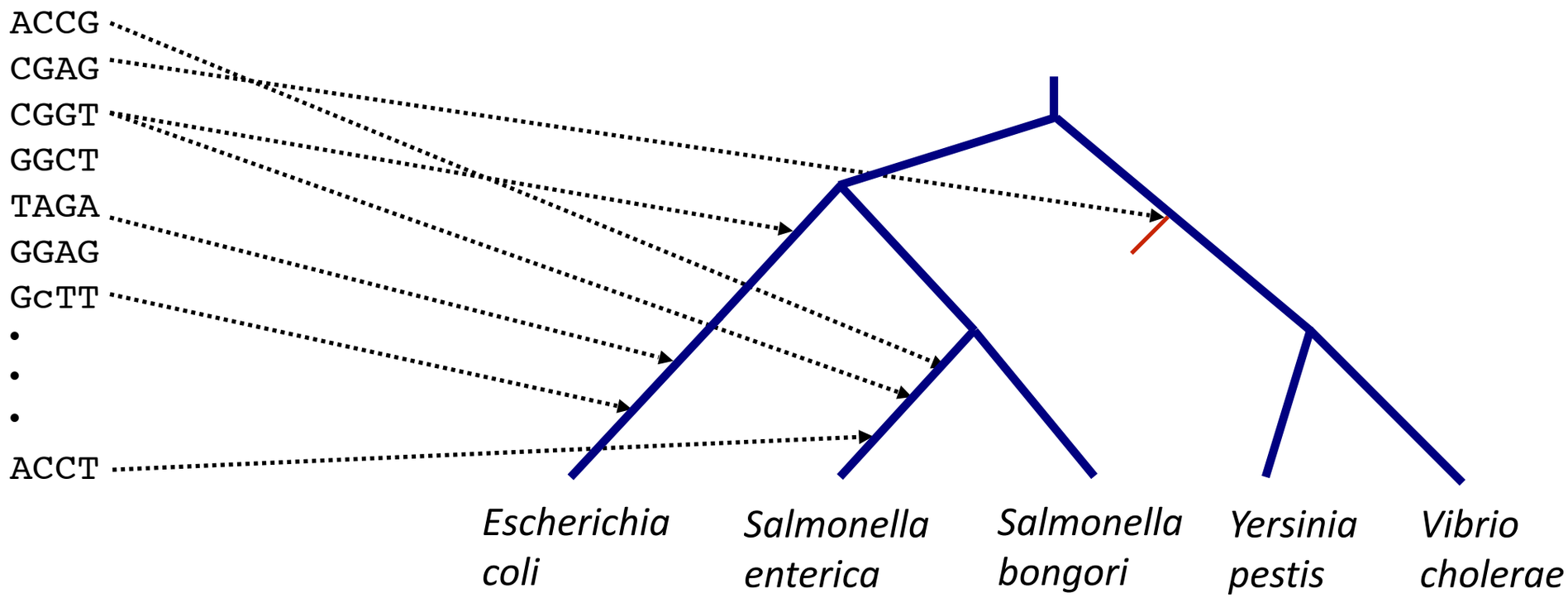
gene 200			gene 30			gene 1000	
CAGGCACGCACGAA			CAGGCACGCACGAA			CAGGCACGCACGAA	
AGC-CACGC-CATA			AGC-CACGC-CATA			AGC-CACGC-CATA	
ATGGCACGC-C-TA			ATGGCACGC-C-TA			ATGGCACGC-C-TA	
AGCTAC-CACGGAT			AGCTAC-CACGGAT			AGCTAC-CACGGAT	
----ATGGCGA---			-ACATGGCT-----			-ACATGGCT-----	
						-----CATTGCT--	







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

- **Input**:
  - A backbone multiple sequence alignment for a marker gene,

SATe-Enabled Phylogenetic Placement (SEPP)

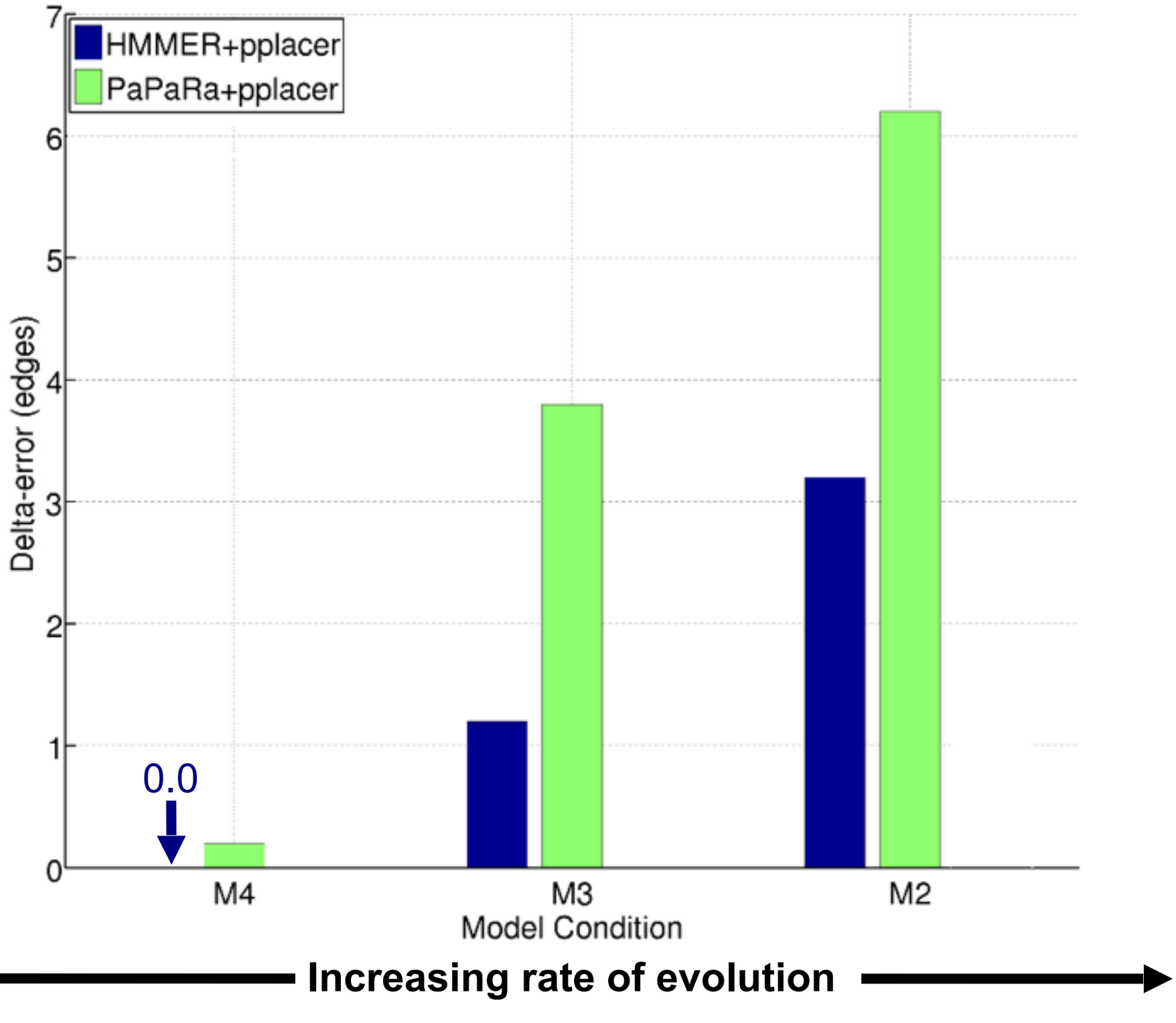
  - A collection of (fragmentary, error-prone) query sequences

- **Output**: Probabilistic placements of each query sequence on the phylogenetic tree after (locally) aligning the query to the reference

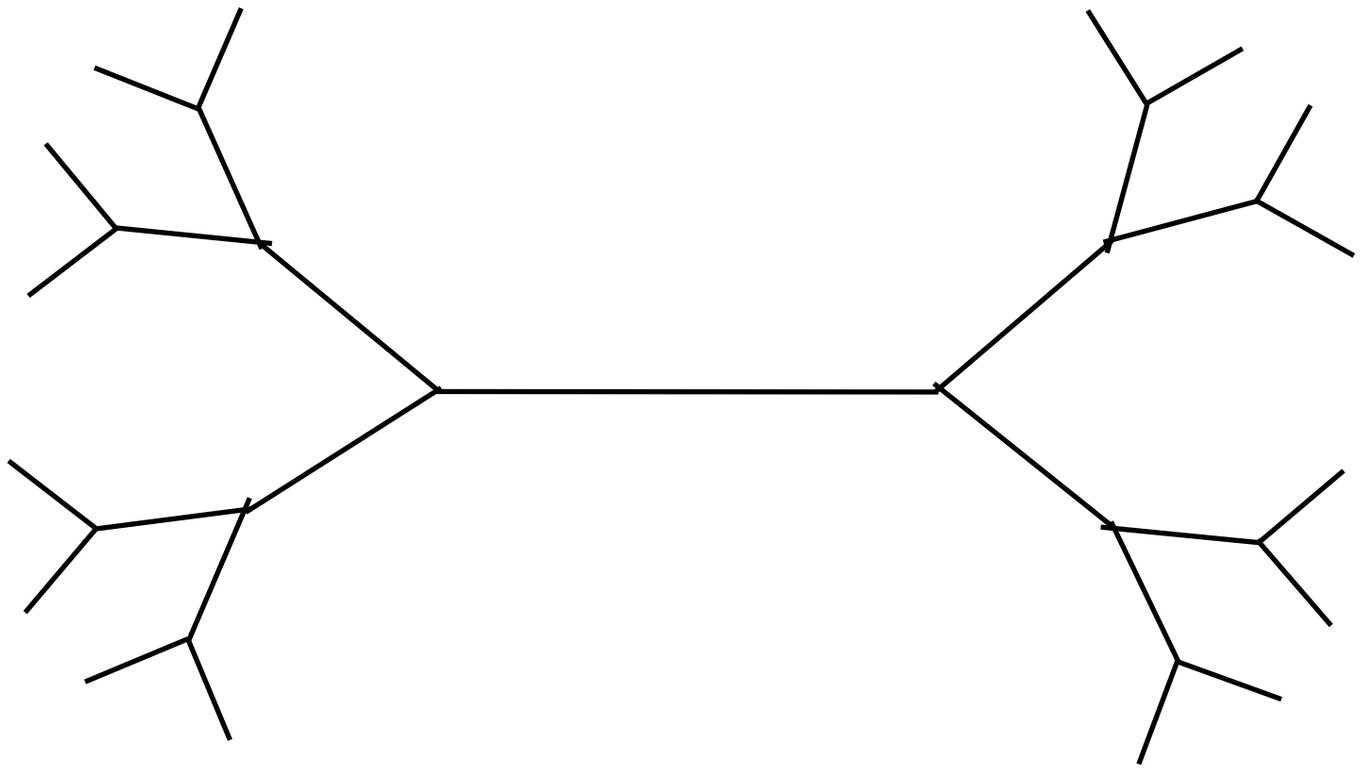
- Tools: — Alignment: HMMER
  — Placement: pplacer (Matsen) and EPA (RAxML)

Phylogenetic placement simulations

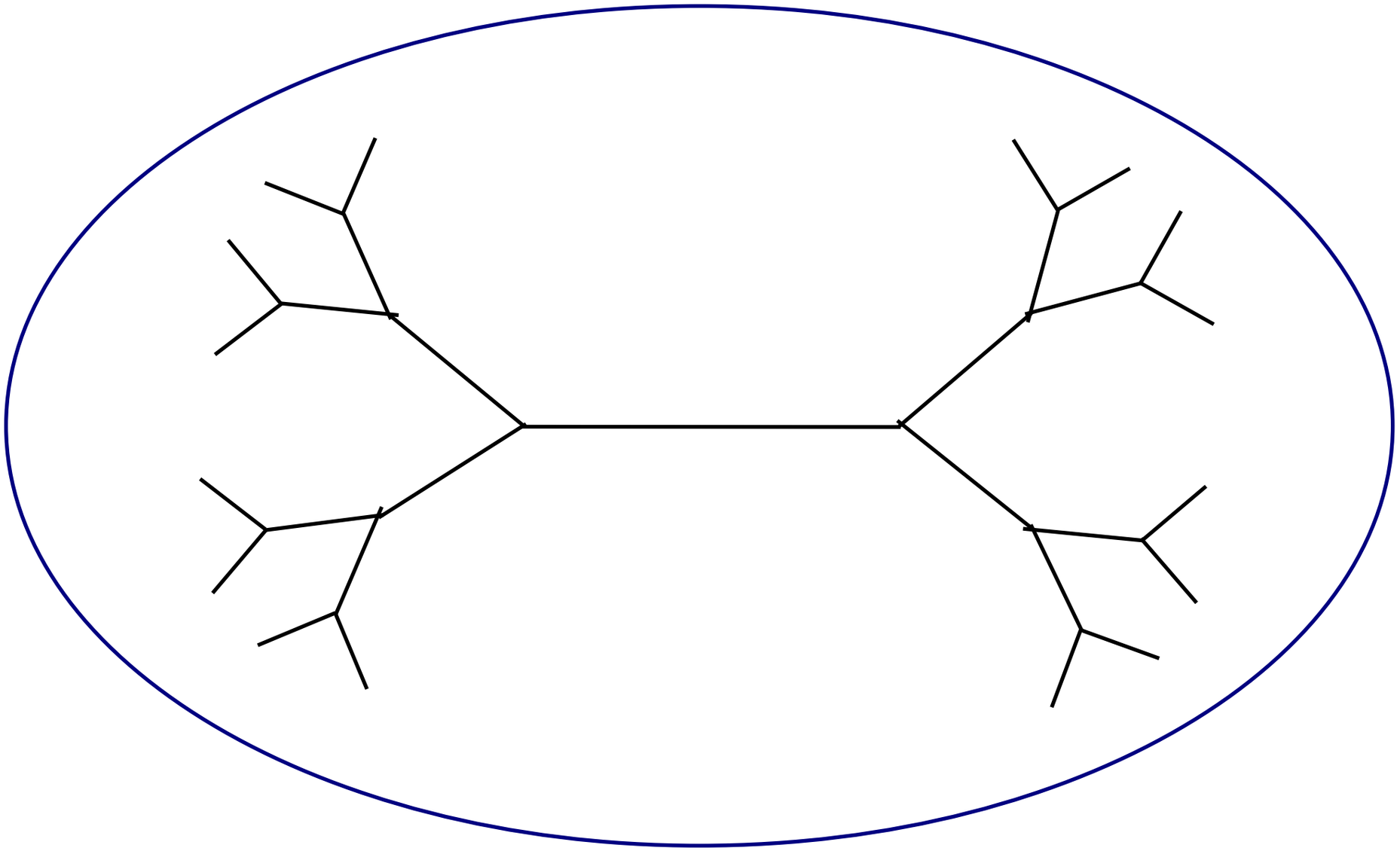
S. Mirarab et al., PSB. (2012).



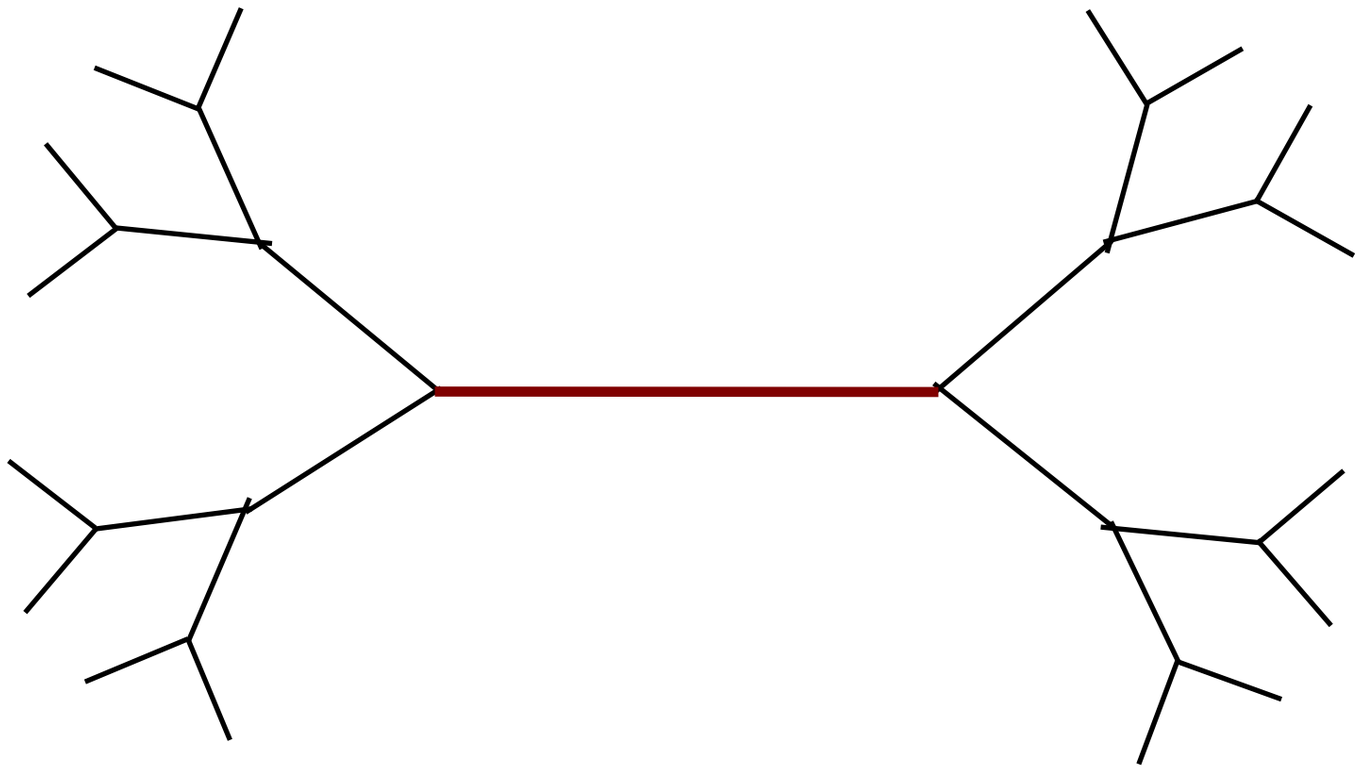
Reference tree



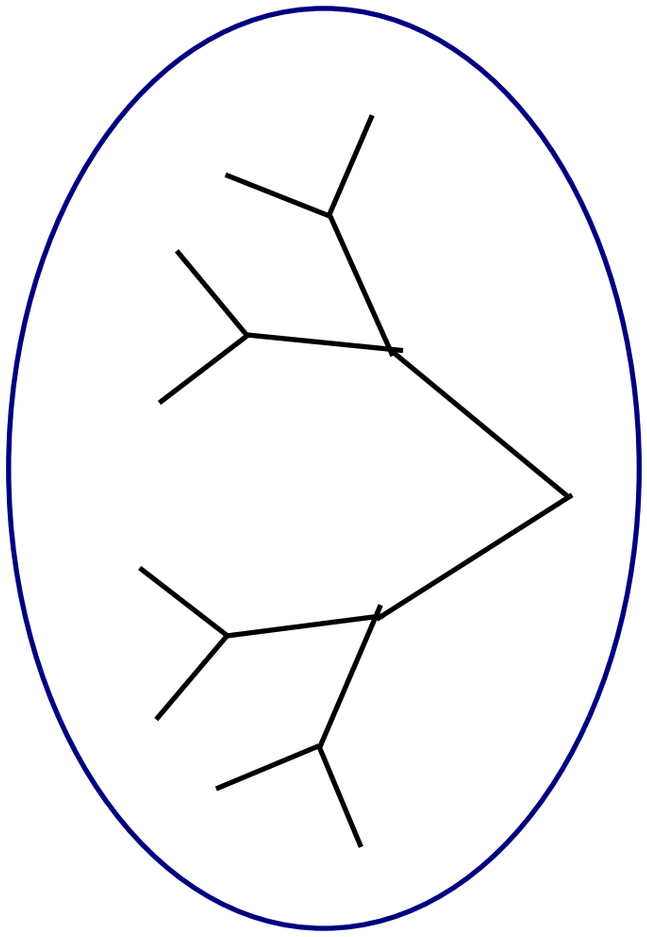
HMM for the alignment step

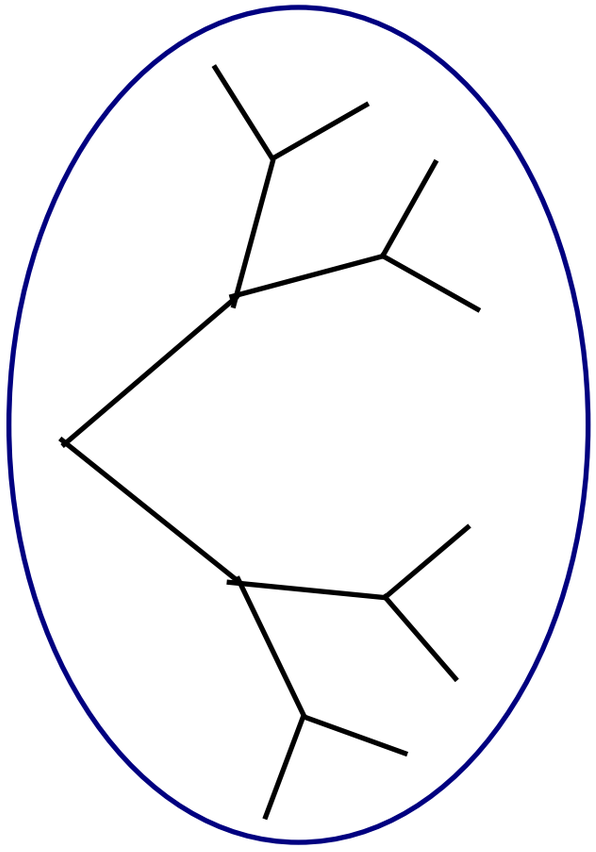


Ensemble of HMMs

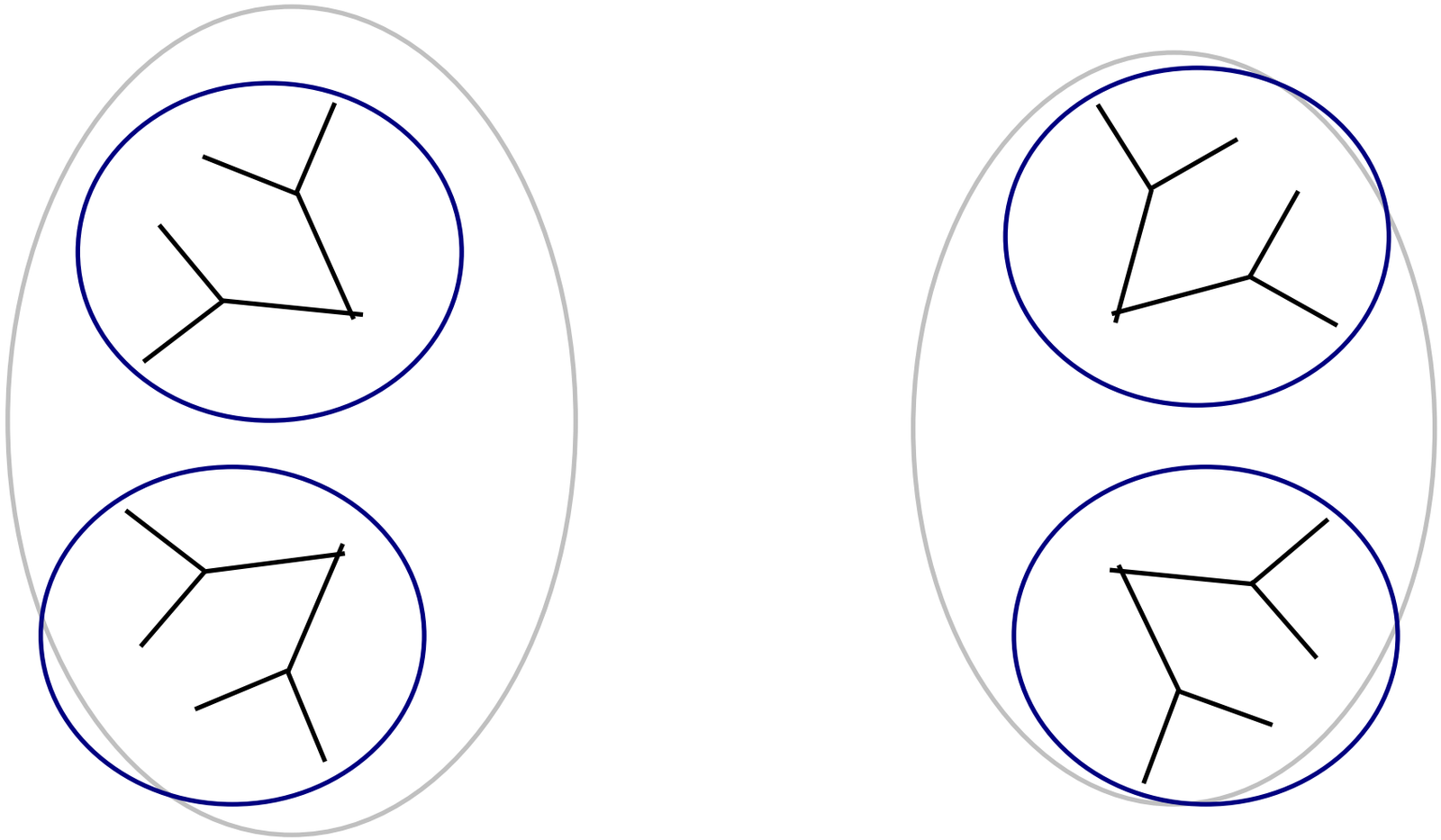


Ensemble of HMMs





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

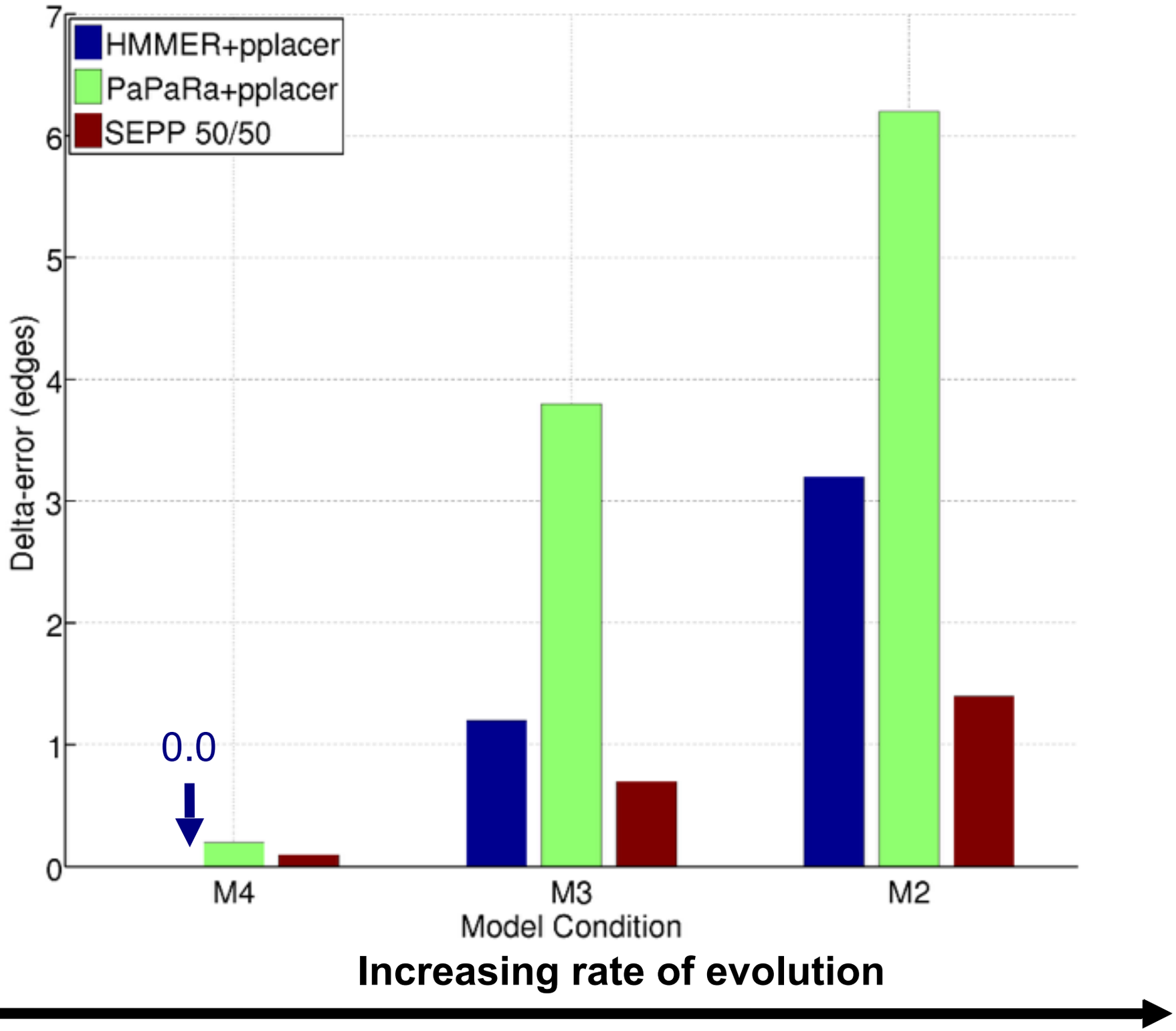
Step 2: Place each query sequence into the backbone tree, using extended alignment

- 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

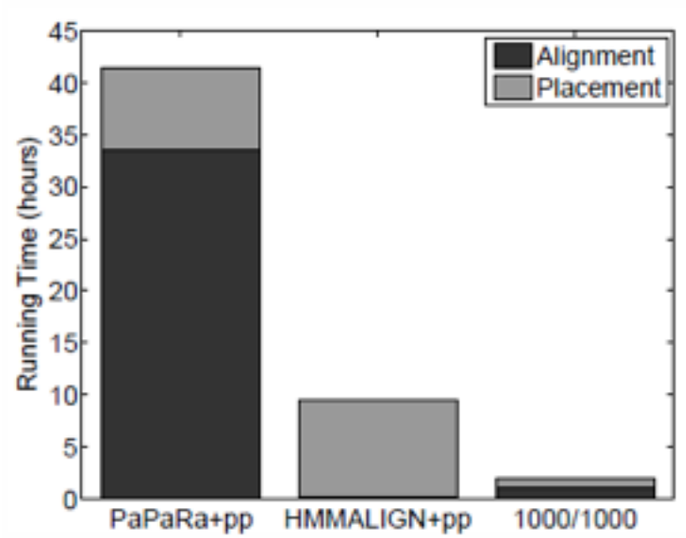
S. Mirarab et al., PSB. (2012).



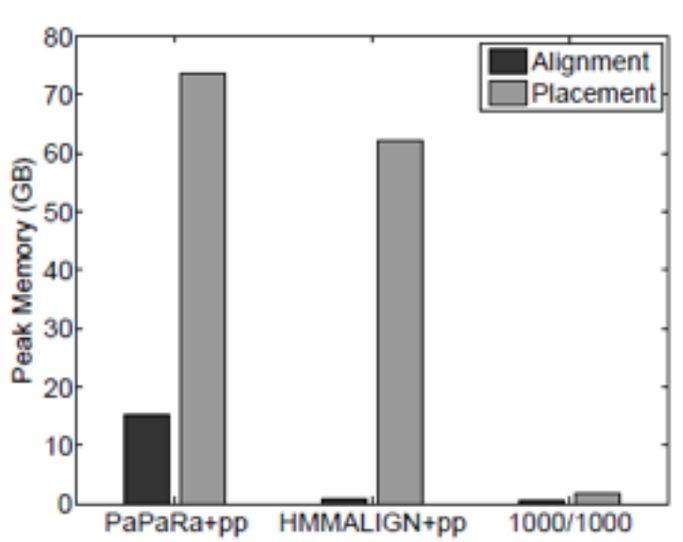
SEPP on large 16S references

Simulations:

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



Running time

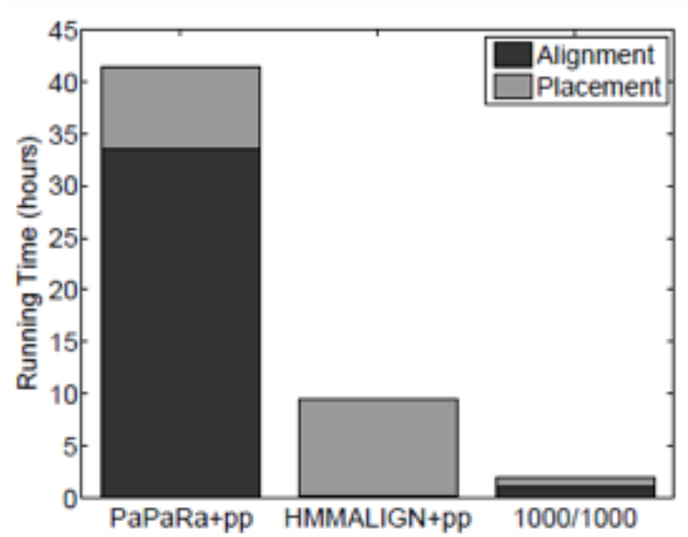


Memory

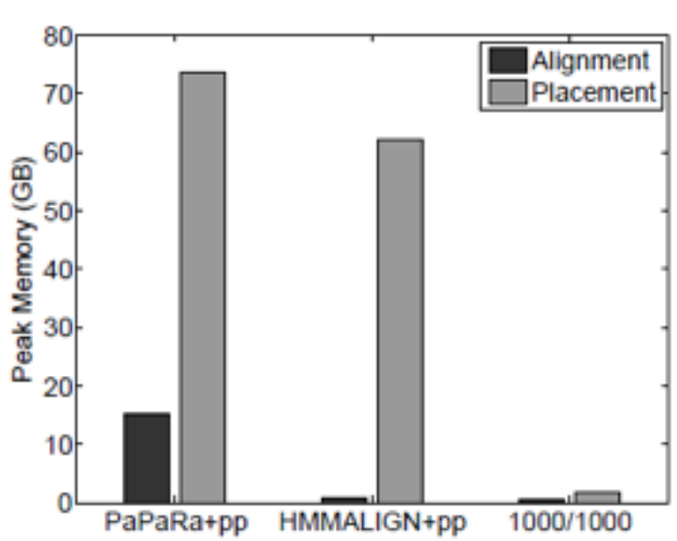
SEPP on large 16S references

Simulations:

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



Running time



Memory

Real data (with **Rob Knight**'s lab; Daniel McDonald):

- **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

Taxonomic Profiling

- **Input**:
  - reads from many species with different abundances
- **Output**:
  - The taxonomic profile of the sample

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

Taxon Identification and Phylogenetic Profiling (TIPP)

Algorithmic steps

Step 1: map fragments to ~30 “marker” genes using BLAST

Nguyen et al., Bioinformatics (2014)

Algorithmic steps

Step 1: map fragments to ~30 “marker” genes using BLAST

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

- Take into account uncertainty: use several alignments and placements on the tree (to reach a predefined level of statistical support)

Nguyen et al., Bioinformatics (2014)

Algorithmic steps

Step 1: map fragments to ~30 “marker” genes using BLAST

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

- Take into account uncertainty: use several alignments and placements on the tree (to reach a predefined level of statistical support)

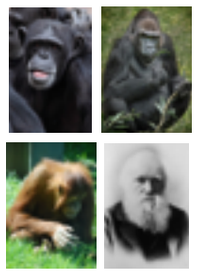
Step 3: Summarize results across genes to get a taxonomic profile

- 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)

Phylogeny reconstruction pipeline



samples

Sequencing

Bioinformatic processing

— PASTA
— UPP

Step 1: Multiple sequence alignment

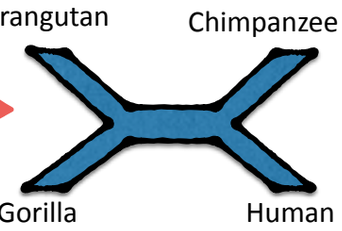
gene 1		MSA		gene 1
ACTGCACACCG				ACTGCACACCG
ACTGCCCCCG				ACTGC-CCCCG
AATGCCCCCG				AATGC-CCCCG
CTGCACACGG				-CTGCACACGG
gene 2				**gene 2**
CTGAGCATCG				CTGAGCATCG
CTGAGCTCG		MSA		CTGAGC-TCG
ATGAGCTC				ATGAGC-TC-
CTGACACG				CTGA-CAC-G
gene 1000				**gene 1000**
CAGGCACGCACGAA				CAGGCACGCACGAA
AGCCACGCCATA		MSA		AGC-CACGC-CATA
ATGGCACGCCTA				ATGGCACGC-C-TA
AGCTACCACGGAT				AGCTAC-CACGGAT

Step 2: Species tree reconstruction

Approach 1: Concatenation

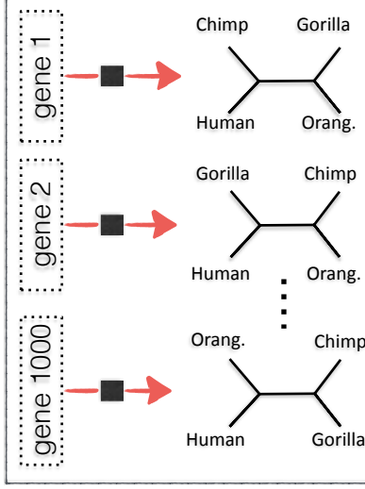
supermatrix			
gene 1	gene 2	...	gene 1000
ACTGCACACCGCTGAGCATCG			CAGAGCACGCACGAA
ACTGC-CCCCGCTGAGC-TCG			AGCA-CACGC-CATA
AATGC-CCCCGATGAGC-TC-			ATGAGCACGC-C-TA
-CTGCACACGGCTGA-CAC-G			AGC-TAC-CACGGAT

Phylogeny inference

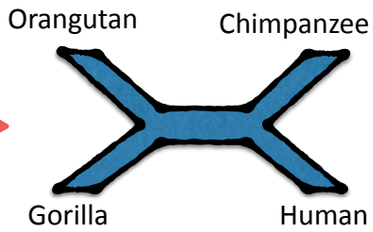


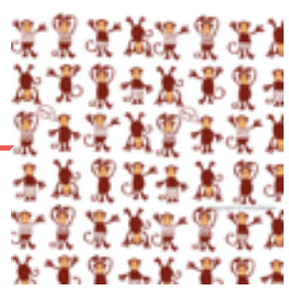
Approach 2: Summary methods

Gene tree estimation



Summary method



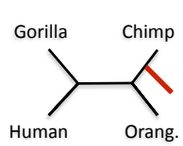


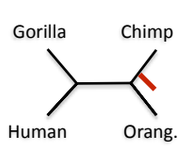
TGGCACGCAACG
ATGGCACGCTA
ATGGCACGCA
AGCTAACACGGAT
ATGGCACGA

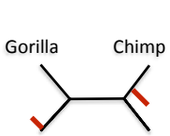
— SEPP
— TIPP

Step 3: Phylogenetic placement

gene 200			gene 30			gene 1000	
CAGGCACGCACGAA			CAGGCACGCACGAA			CAGGCACGCACGAA	
AGC-CACGC-CATA			AGC-CACGC-CATA			AGC-CACGC-CATA	
ATGGCACGC-C-TA			ATGGCACGC-C-TA			ATGGCACGC-C-TA	
AGCTAC-CACGGAT			AGCTAC-CACGGAT			AGCTAC-CACGGAT	
----ATGGCGA---			-ACATGGCT-----			-ACATGGCT-----	
						-----CATTGCT--	







Multiple sequence alignment

- **Input**: a (potentially ultra-large) set of input sequences from a single gene
  - sequence may be full or fragmentary
- **Output**: a multiple sequence alignment
  - Optional: co-estimate the alignment and tree
- **Relevance**: useful to get very large reference alignments and trees with up to hundreds of thousands of leaves

Multiple sequence alignment

- **Input**: a (potentially ultra-large) set of input sequences
- **PASTA**
  - Great for trees
  - Not good for fragmentary data
- **Relevance**: useful to get very large reference alignments and trees with up to hundreds of thousands of leaves

Multiple sequence alignment

- **Input**: a (potentially ultra-large) set of input sequences
- **PASTA**
  - Great for trees
  - Not good for fragmentary data
- **UPP**
  - Good for fragmentary data
- **Relevance**: useful to get very large reference alignments and trees with up to hundreds of thousands of leaves

UPP Steps

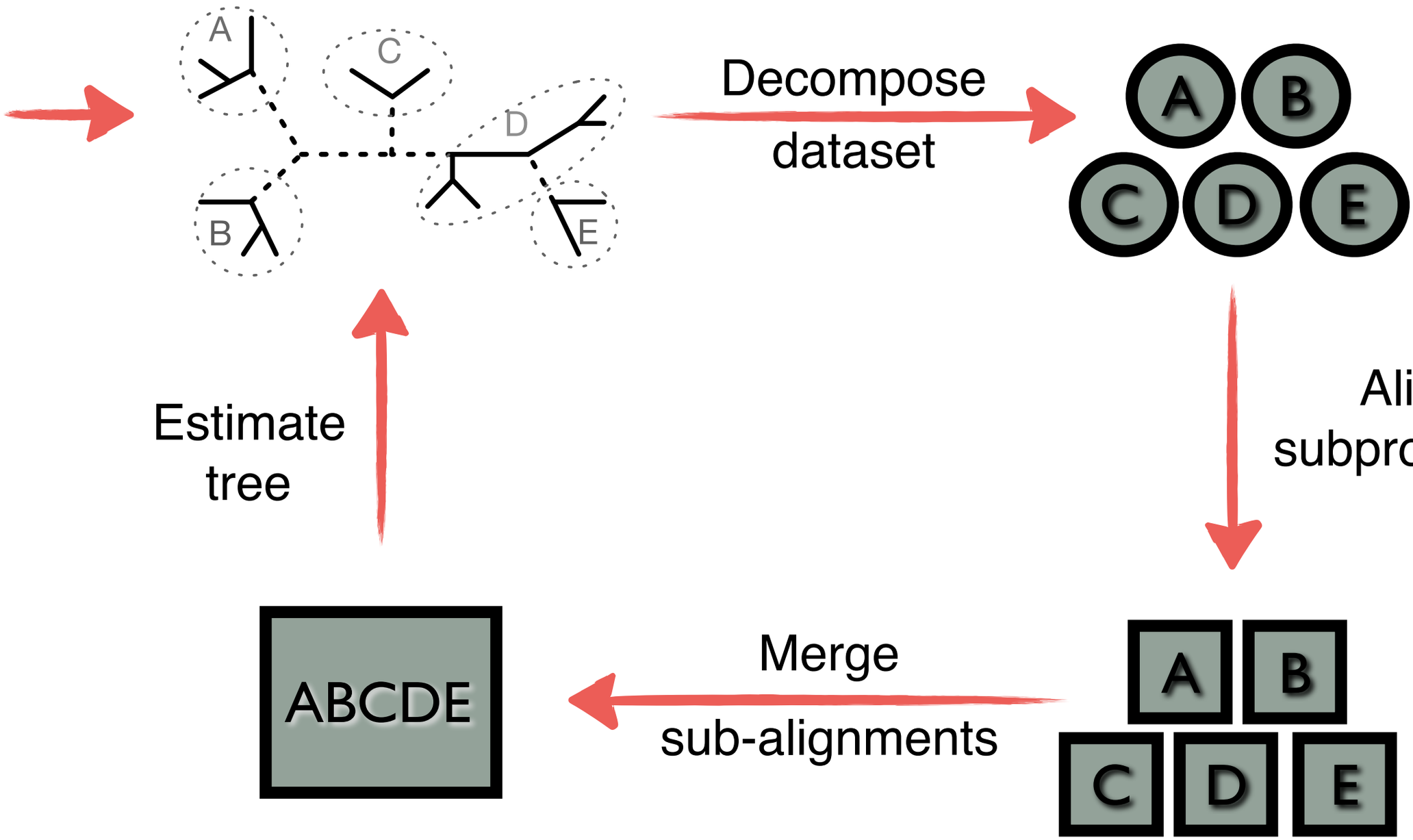
- **Step 1:** randomly select a “small” subset of full length sequences (e.g., 1000) as backbone.

- **Step 2:** align the backbone using other tools (e.g., using PASTA)

- **Step 3:** Use a SEPP-like approach to align the remaining sequences into the reference

- Note: leaves some “insertion” sites unaligned

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).

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 McDonlad
- Mirarab lab
  - Uyen Mai

XSEDE
Extreme Science and Engineering Discovery Environment

HHMI
HOWARD HUGHES MEDICAL INSTITUTE

SDSC SAN DIEGO SUPERCOMPUTER CENTER