Taxonomic identification and phylogenetic profiling

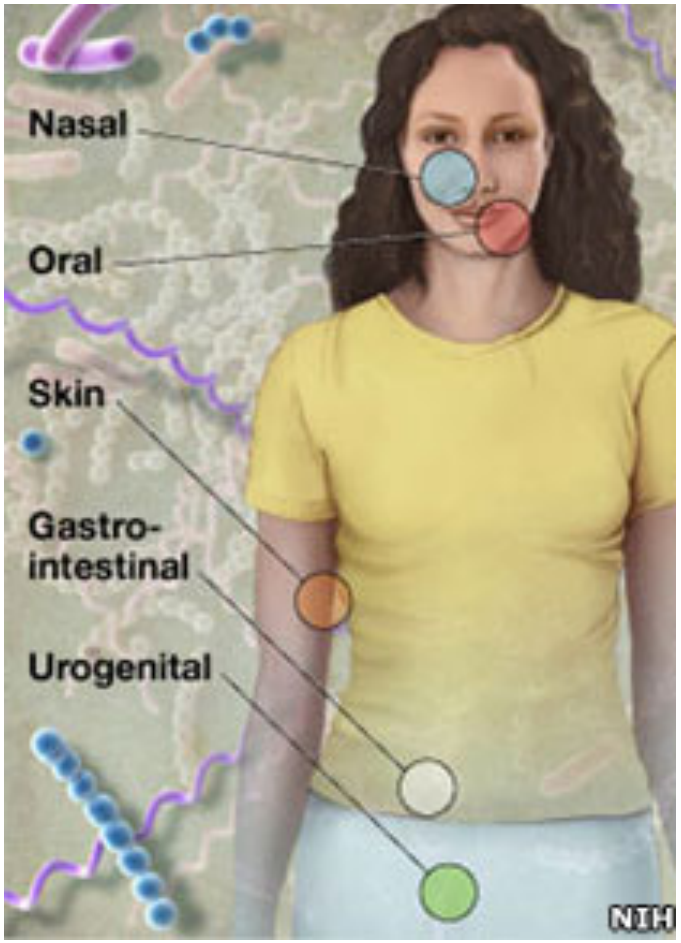
Nam-phuong Nguyen

Carl R. Woese Institute for Genomic Biology

University of Illinois at Urbana-Champaign

Joint work with Siavash Mirarab, Mihai Pop, and Tandy Warnow

Metagenomics



- Culture-independent method for studying a microbiome
- Extract genetic material directly from the environment
- Applications to biofuel production, agriculture, human health
- Sequencing technology produces millions of short reads from unknown species
- Fundamental steps in analysis is identifying taxa of read and estimating a population profile of a sample

Courtesy of Human Microbiome Project

Taxonomic Identification and Profiling

- Taxonomic identification
  - Objective: Given a query sequence, identify the taxon (species, genus, family, etc...) of the sequence
  - Classification problem
- Taxonomic profiling
  - Objective: Given a set of query sequences collected from a sample, estimate the population profile of the sample
  - Estimation problem
  - Can be solved via taxonomic identification

Taxonomic Identification Methods

- Sequence similarity search
  - Classifies by finding most similar sequence
  - Classifies fragments from any region of genome
  - BLAST
- Composition-based methods
  - Typically uses k-mers
  - Classifies fragments from any region of genome
  - PhymmBL, NBC
- Phylogeny-based methods
  - Classifies fragments by using a phylogeny

Phylogeny-based taxonomic identification

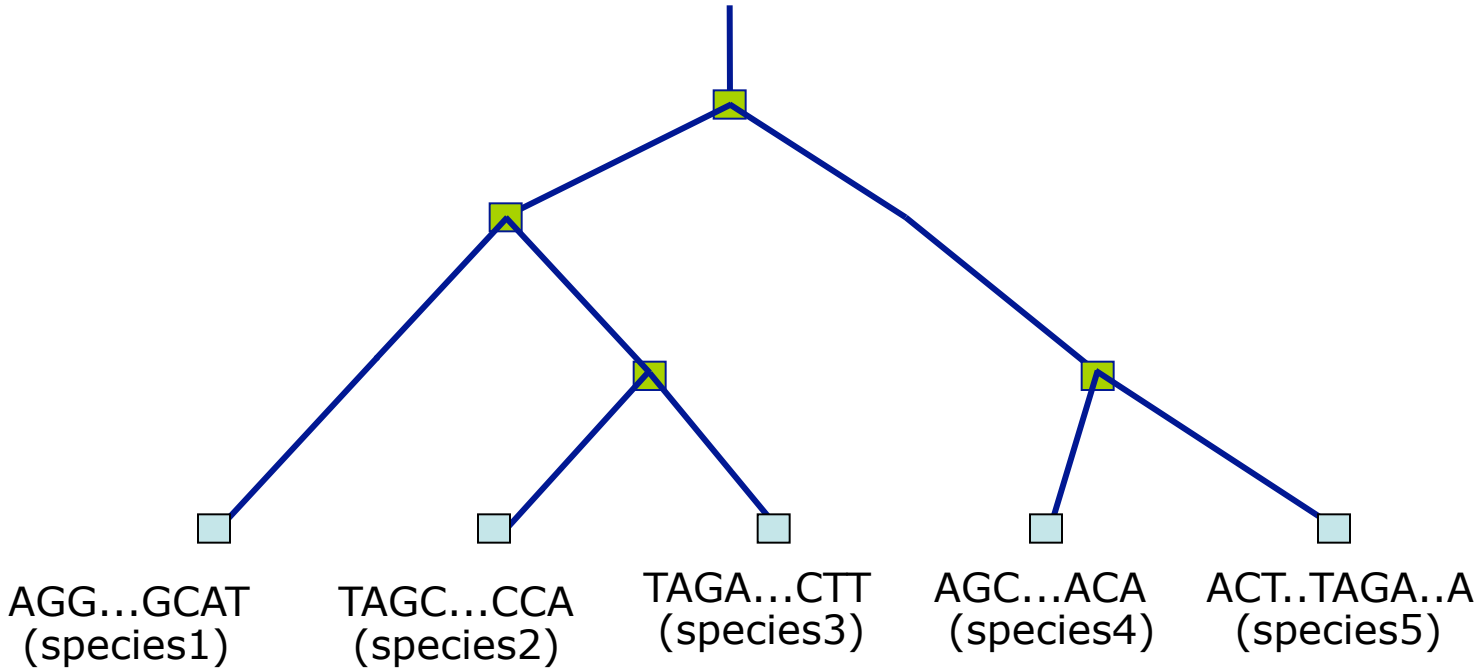
Fragmentary Reads:

(60-200 bp long)

- ACCG
- CGAG
- CGG
- GGCT
- TAGA
- GGGGG
- TCGAG
- GGCG
- GGG
- .
- .
- .
- ACCT

Known full-length gene sequences,
and an alignment and a tree

(500-10,000 bp long)



Phylogeny-based taxonomic identification

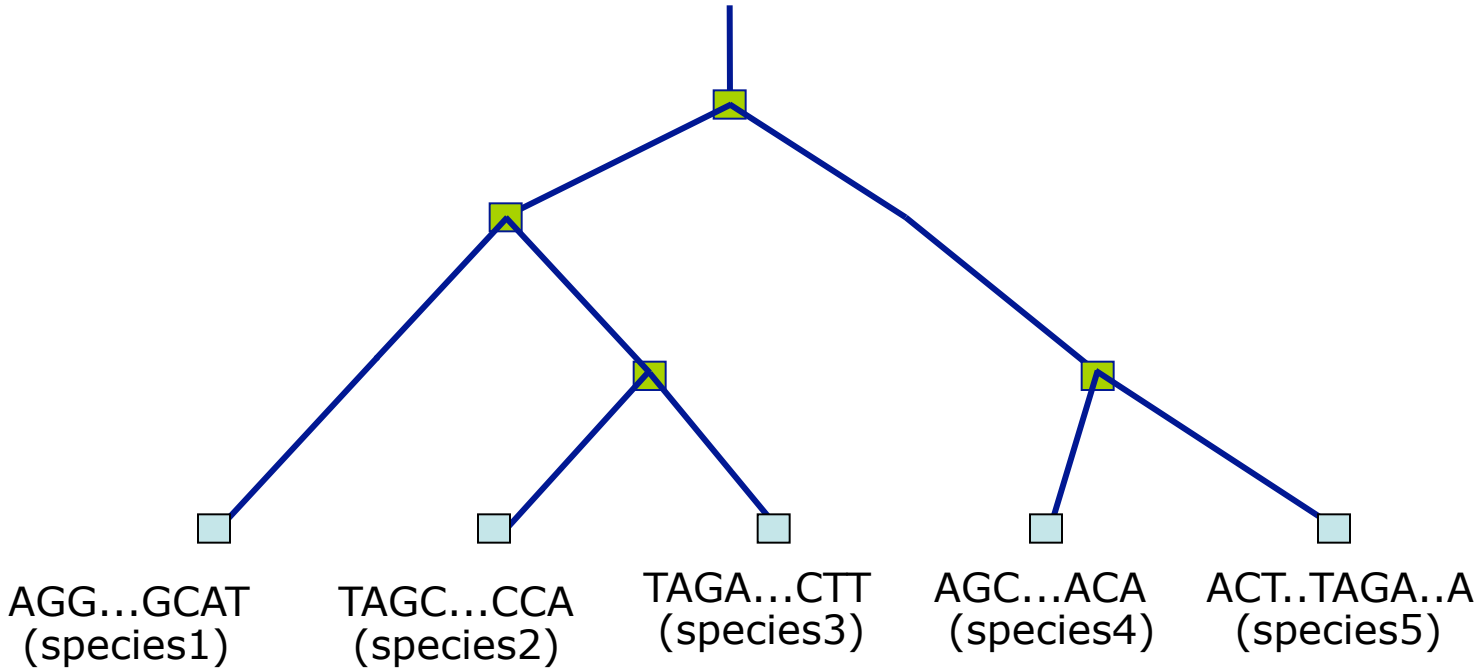
Fragmentary Reads:

(60-200 bp long)

- ACCG
- CGAG
- CGG
- GGCT
- TAGA
- GGGGG
- TCGAG
- GGCG
- GGG
- .
- .
- .
- ACCT

Known full-length gene sequences,
and an alignment and a tree

(500-10,000 bp long)



Phylogeny-based taxonomic identification

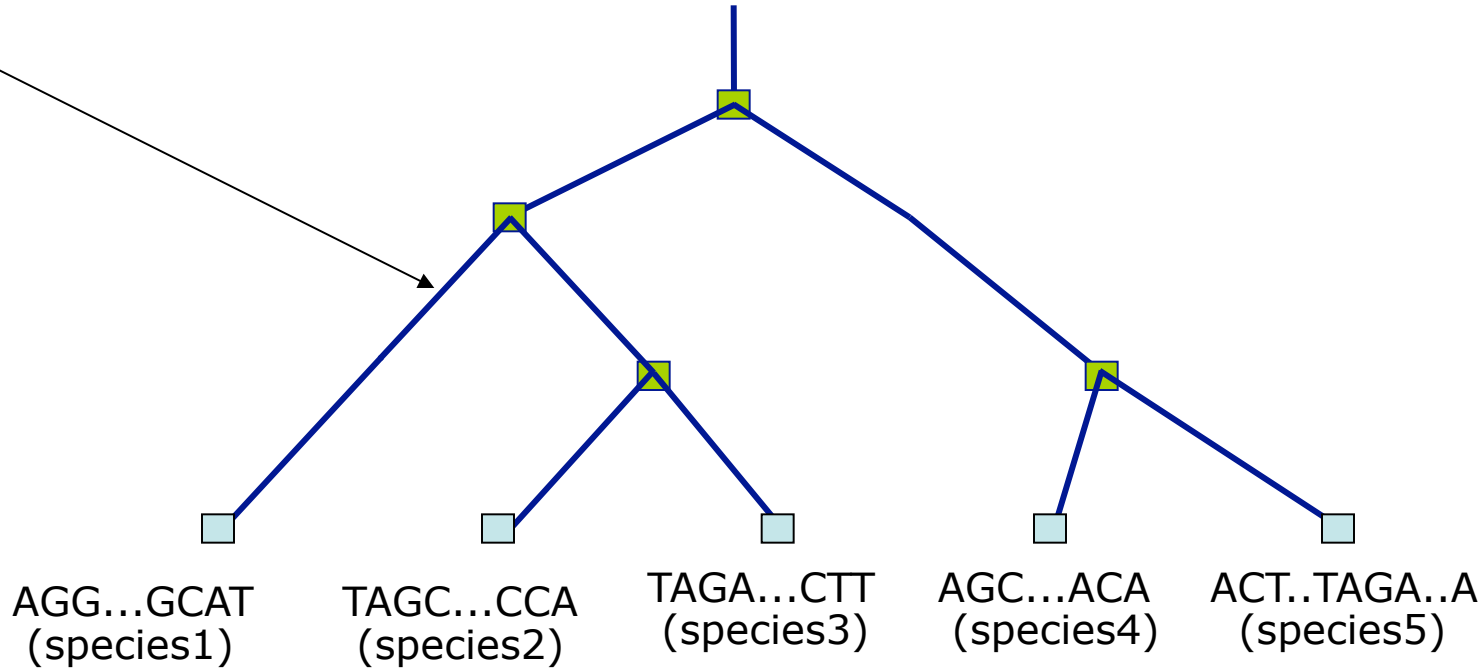
Fragmentary Reads:

(60-200 bp long)

- ACCG
- CGAG
- CGG
- GGCT
- TAGA
- GGGGG
- TCGAG
- GGCG
- GGG
- .
- .
- .
- ACCT

Known full-length gene sequences,
and an alignment and a tree

(500-10,000 bp long)



Phylogeny-based taxonomic identification

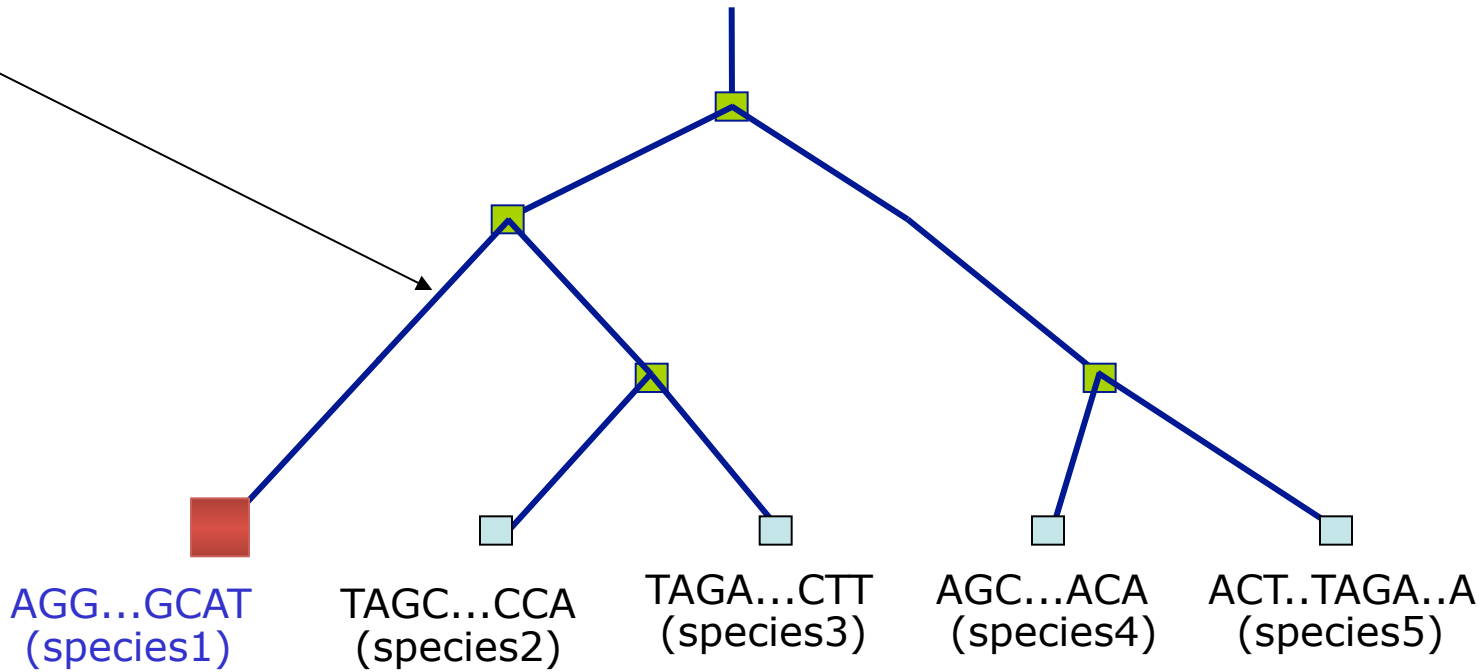
Fragmentary Reads:

(60-200 bp long)

- ACCG
- CGAG
- CGG
- GGCT
- TAGA
- GGGGG
- TCGAG
- GGCG
- GGG
- .
- .
- .
- ACCT

Known full-length gene sequences,
and an alignment and a tree

(500-10,000 bp long)



Phylogeny-based taxonomic identification

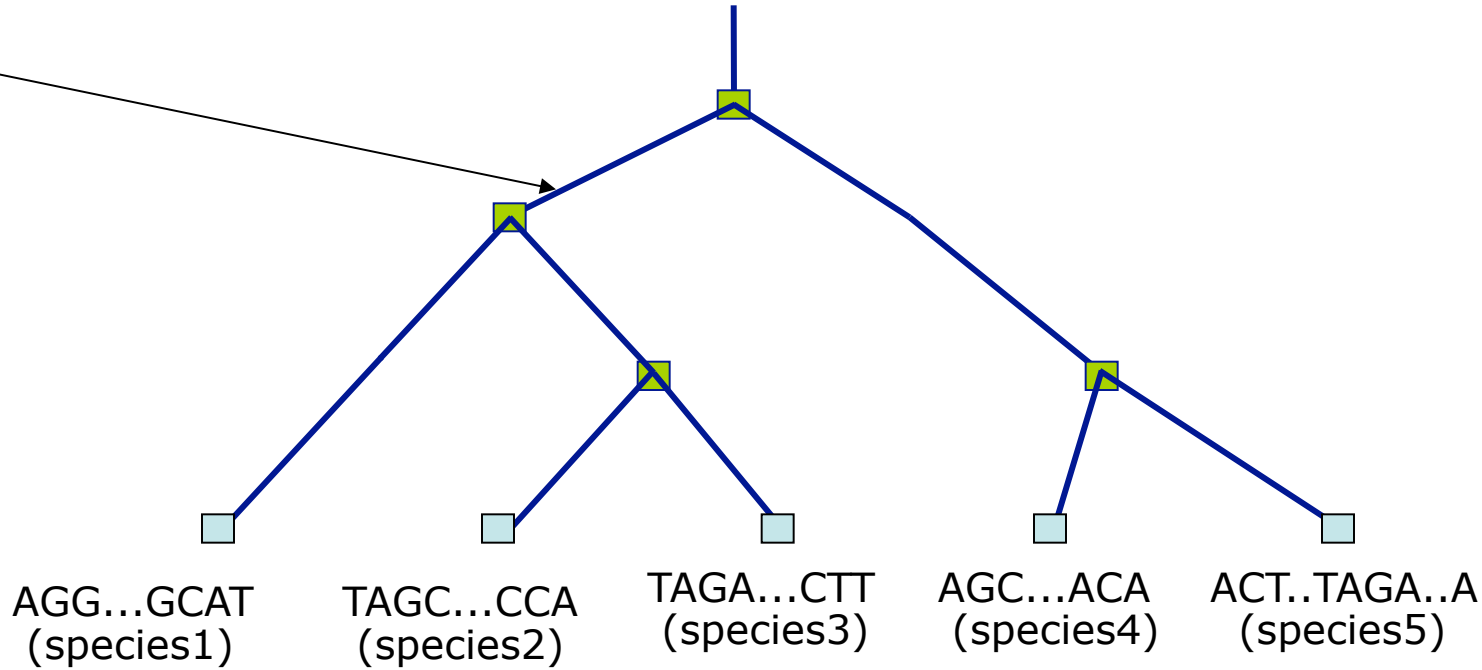
Fragmentary Reads:

(60-200 bp long)

- ACCG
- CGAG
- CGG
- GGCT
- TAGA
- GGGGG
- TCGAG
- GGCG
- GGG
- .
- .
- .
- ACCT

Known full-length gene sequences,
and an alignment and a tree

(500-10,000 bp long)



Phylogeny-based taxonomic identification

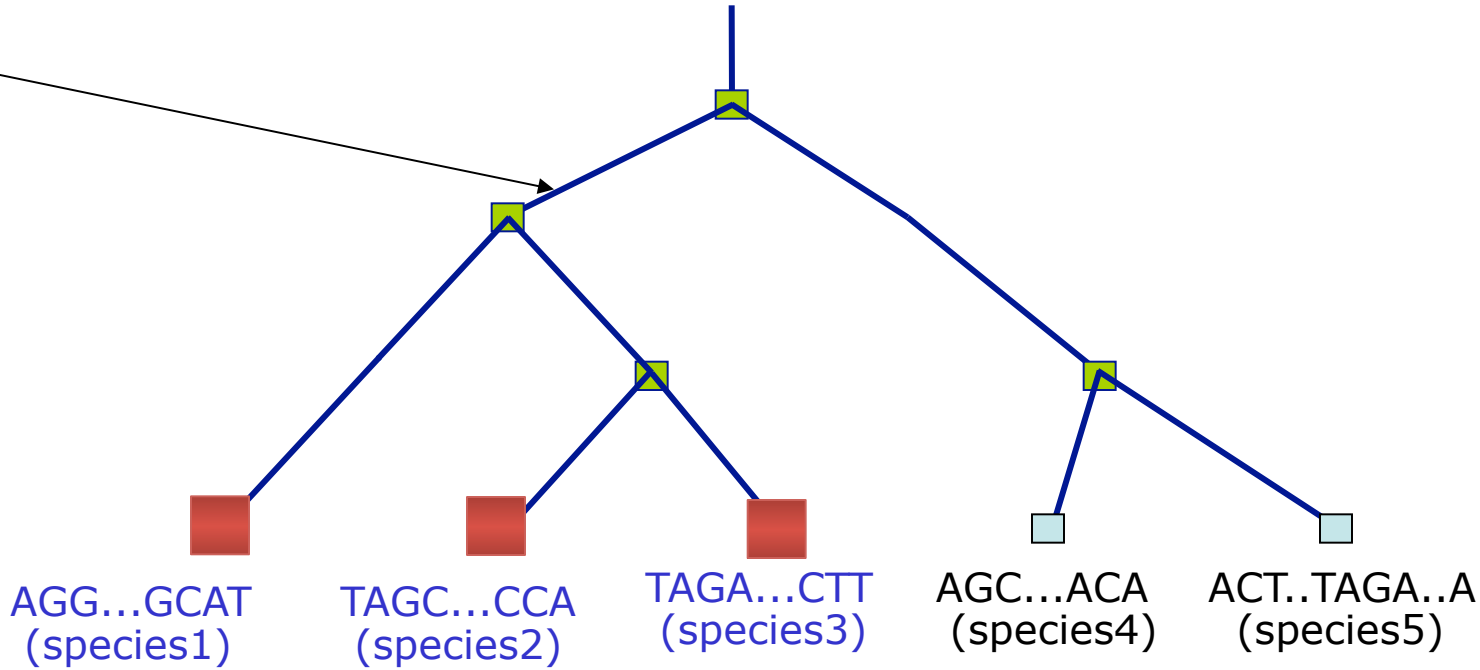
Fragmentary Reads:

(60-200 bp long)

- ACCG
- CGAG
- CGG
- GGCT
- TAGA
- GGGGG
- TCGAG
- GGCG
- GGG
- .
- .
- .
- ACCT

Known full-length gene sequences,
and an alignment and a tree

(500-10,000 bp long)

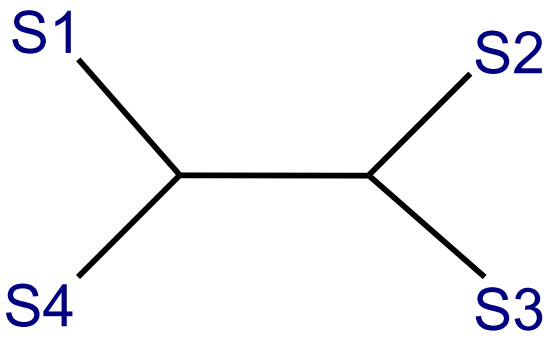


Phylogenetic Placement

- Input: (Backbone) Alignment and tree on full-length sequences and a query sequence (short read)
- Output: Placement of the query sequence on the backbone tree
- Use placement to infer relationship between query sequence and full-length sequences in backbone tree
- Applications in metagenomic analysis
  - Millions of reads
  - Reads from different genomes mixed together
  - Use placement to identify read

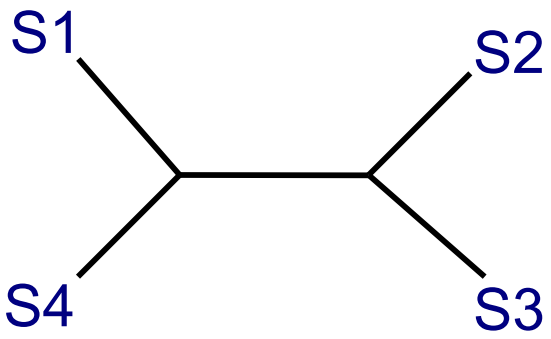
Align Sequence

```
S1   = -AGGCTATCACCTGACCTCCA-AA
S2   = TAG-CTATCAC--GACCGC--GCA
S3   = TAG-CT-------GACCGC--GCT
S4   = TAC----TCAC--GACCGACAGCT
Q1   = TAAAAC
```



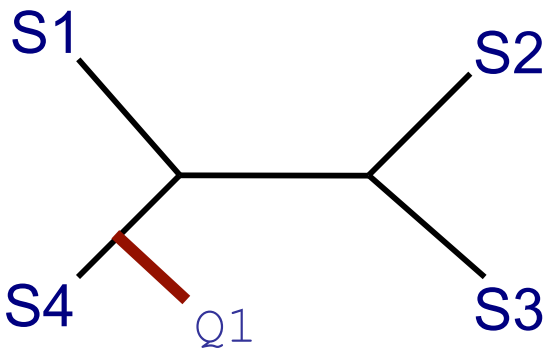
Align Sequence

```
S1  = -AGGCTATCACCTGACCTCCA-AA
S2  = TAG-CTATCAC--GACCGC--GCA
S3  = TAG-CT-------GACCGC--GCT
S4  = TAC----TCAC--GACCGACAGCT
Q1  = -------T-A--AAAC--------
```



Place Sequence

```
S1  = -AGGCTATCACCTGACCTCCA-AA
S2  = TAG-CTATCAC--GACCGC--GCA
S3  = TAG-CT-------GACCGC--GCT
S4  = TAC----TCAC--GACCGACAGCT
Q1  = -------T-A--AAAC--------
```



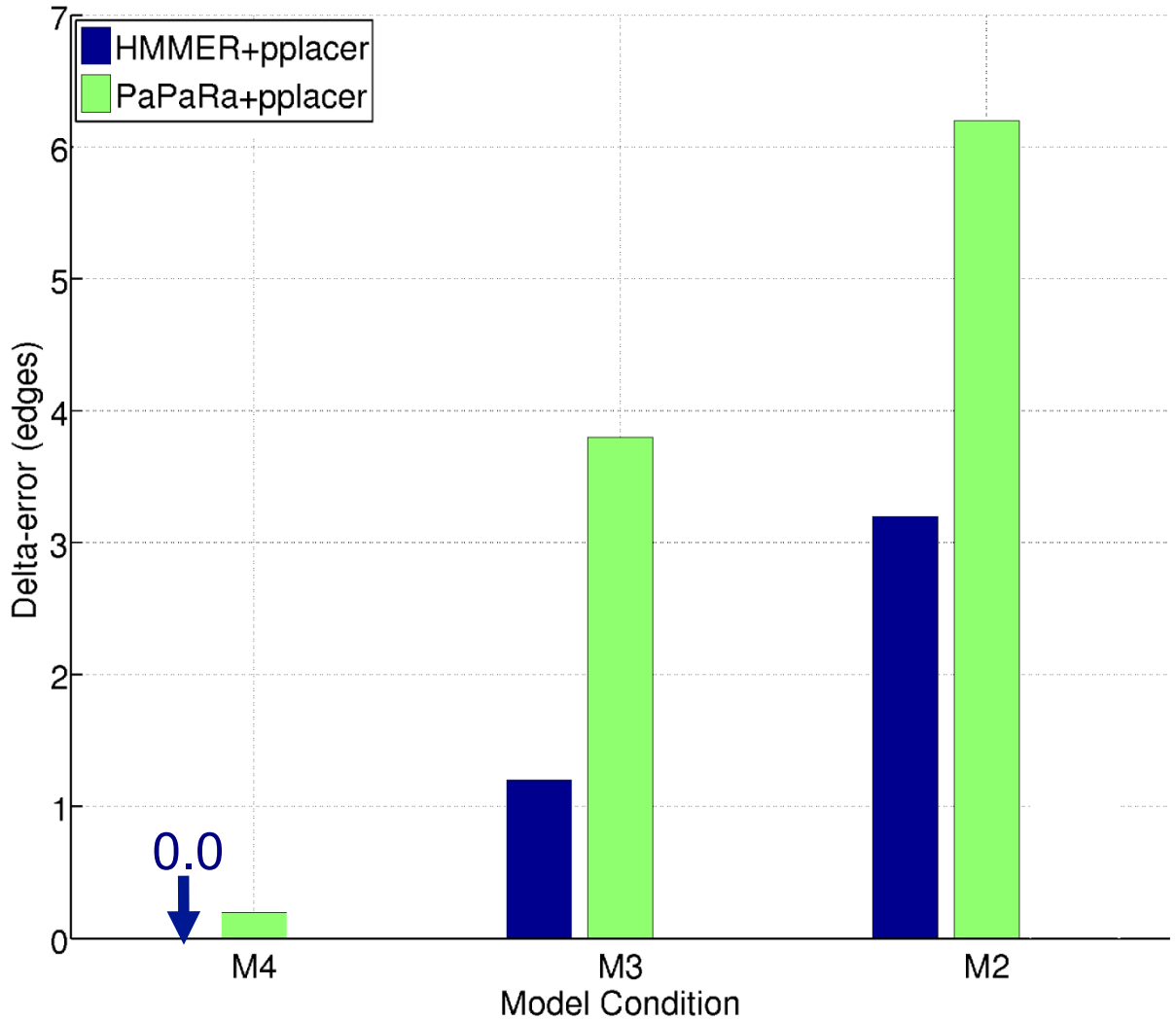
Phylogenetic Placement

- Align each query sequence to backbone alignment:
  - HMMALIGN (Eddy, Bioinformatics 1998)
  - PaPaRa (Berger and Stamatakis, Bioinformatics 2011)
- Place each query sequence into backbone tree, using extended alignment:
  - pplacer (Matsen et al., BMC Bioinformatics 2010)
  - EPA (Berger et al., Systematic Biology 2011)

Phylogenetic Placement

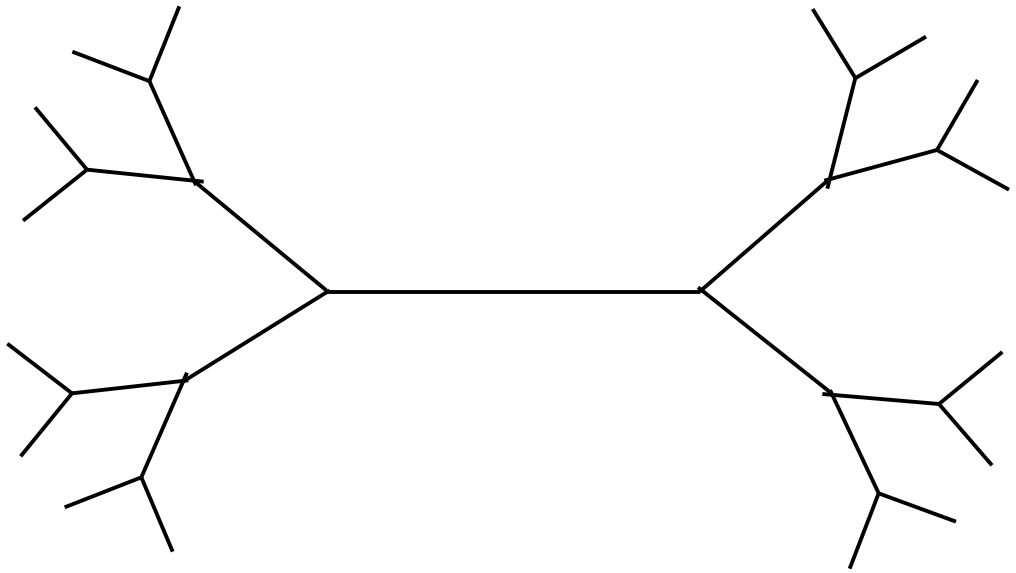
- Align each query sequence to backbone alignment:
  - HMMALIGN (Eddy, Bioinformatics 1998)
  - PaPaRa (Berger and Stamatakis, Bioinformatics 2011)
- Place each query sequence into backbone tree, using extended alignment:
  - pplacer (Matsen et al., BMC Bioinformatics 2010)
  - EPA (Berger et al., Systematic Biology 2011)

HMMER and PaPaRa results

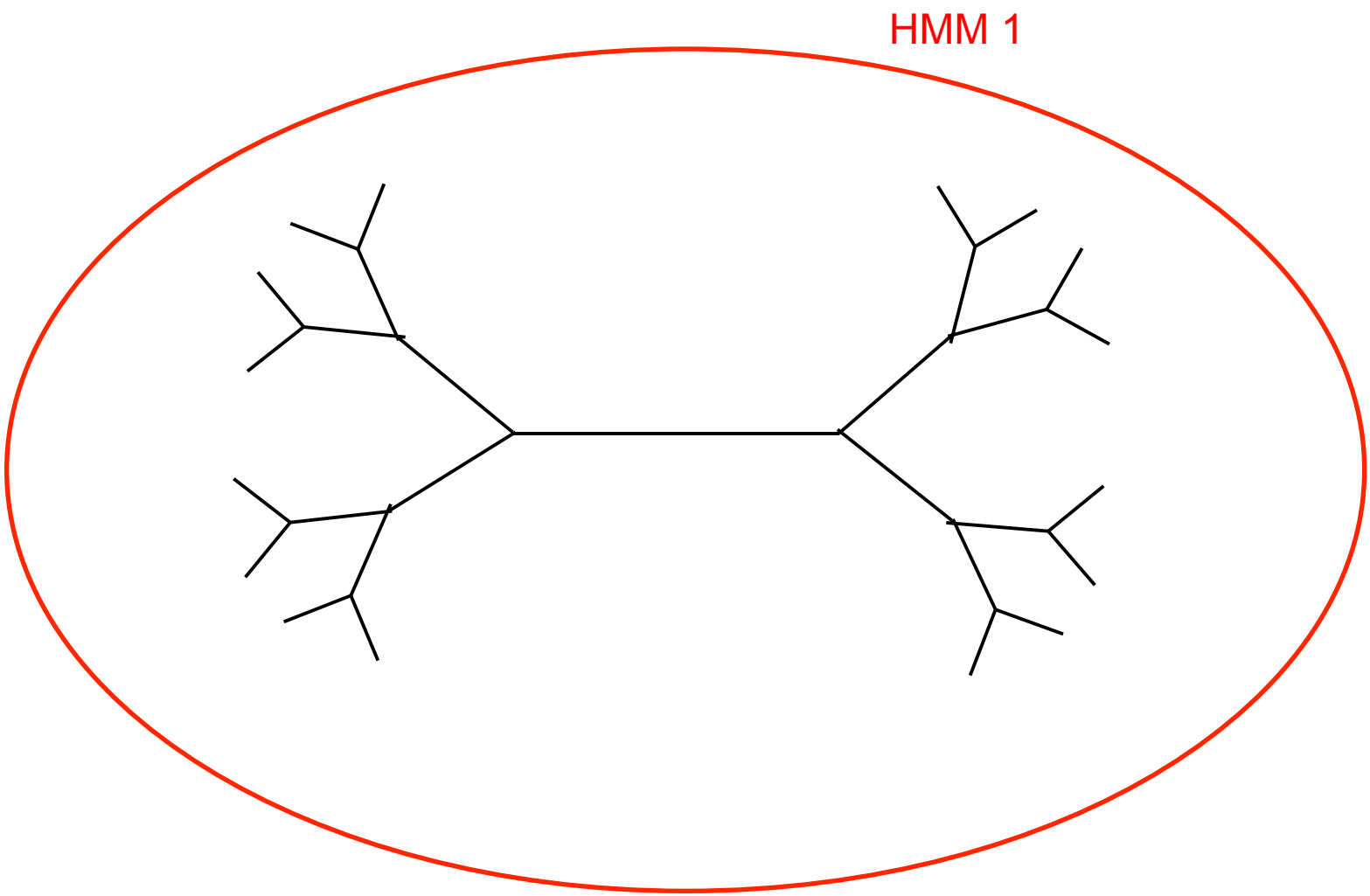


Backbone size: 500
5000 fragments
20 replicates

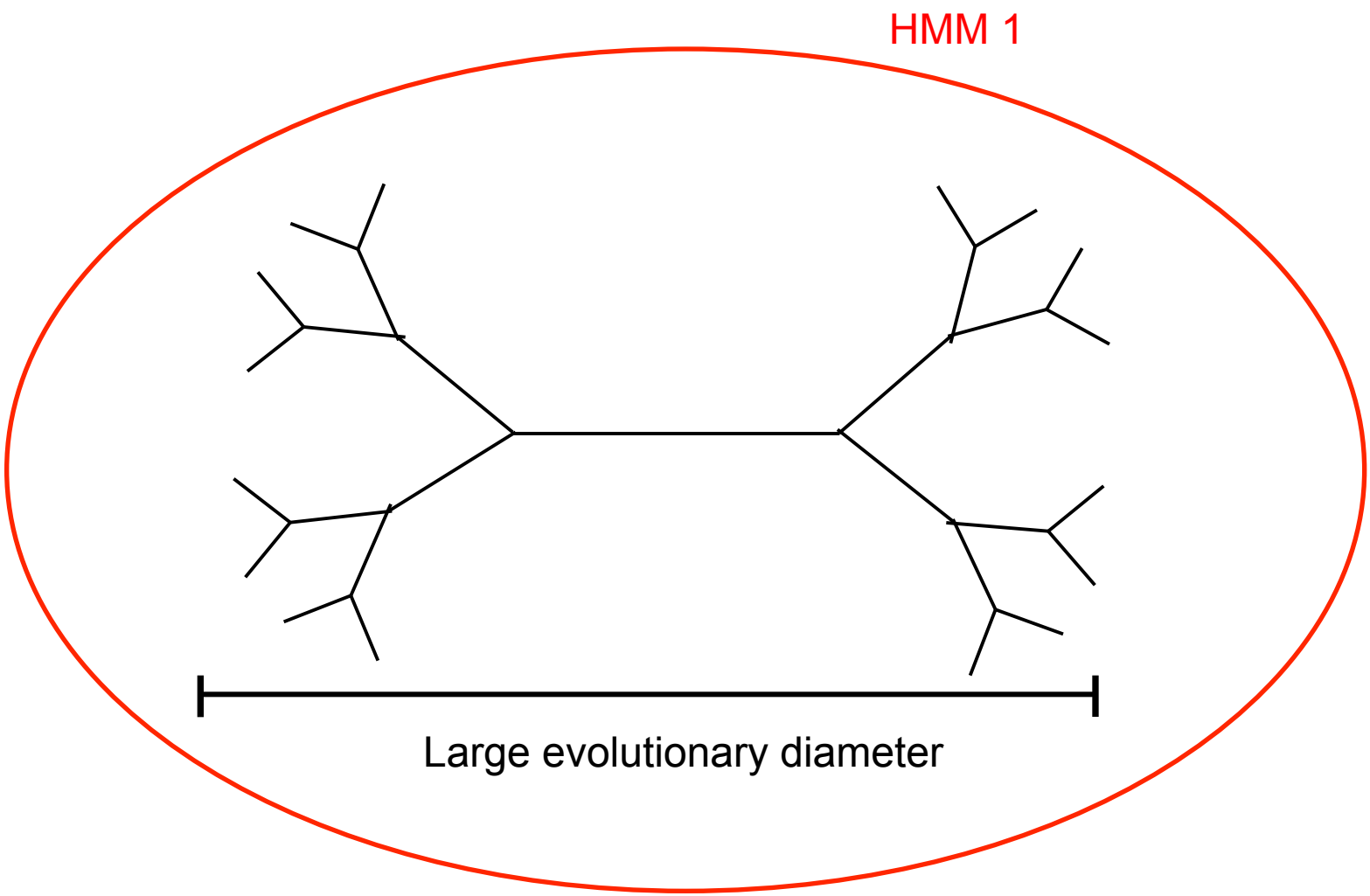
Old approach using single HMM



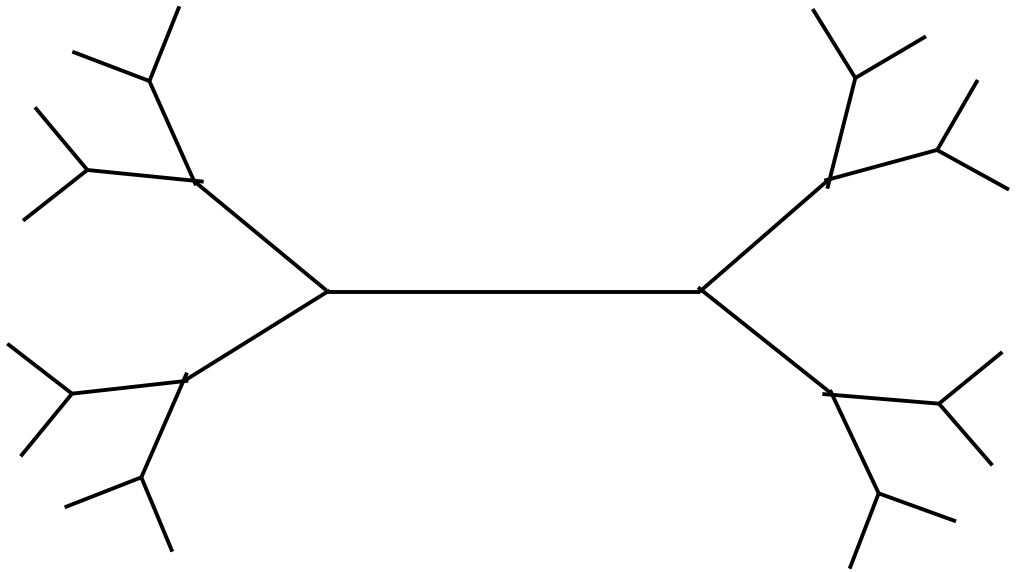
Old approach using single HMM



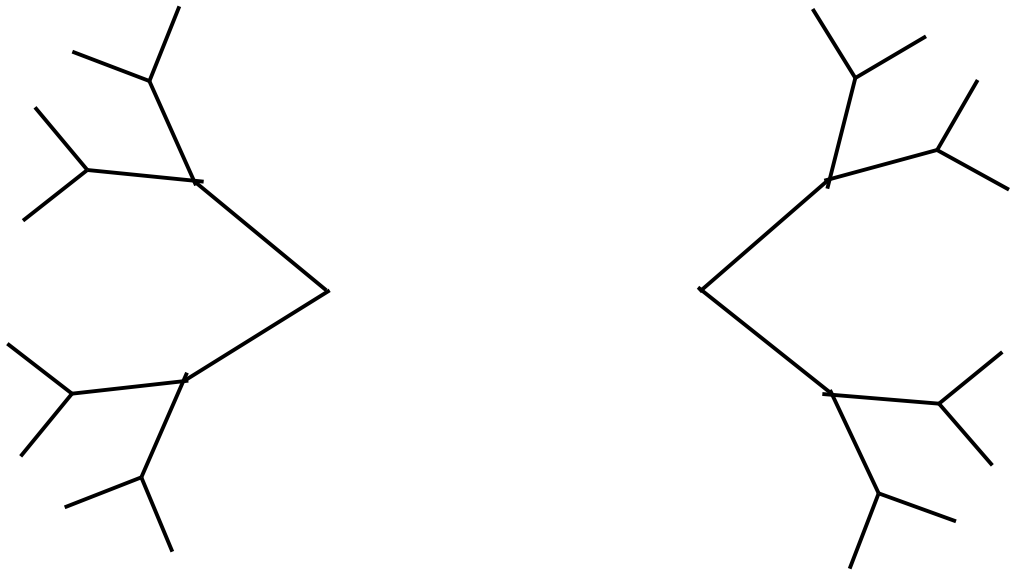
Old approach using single HMM



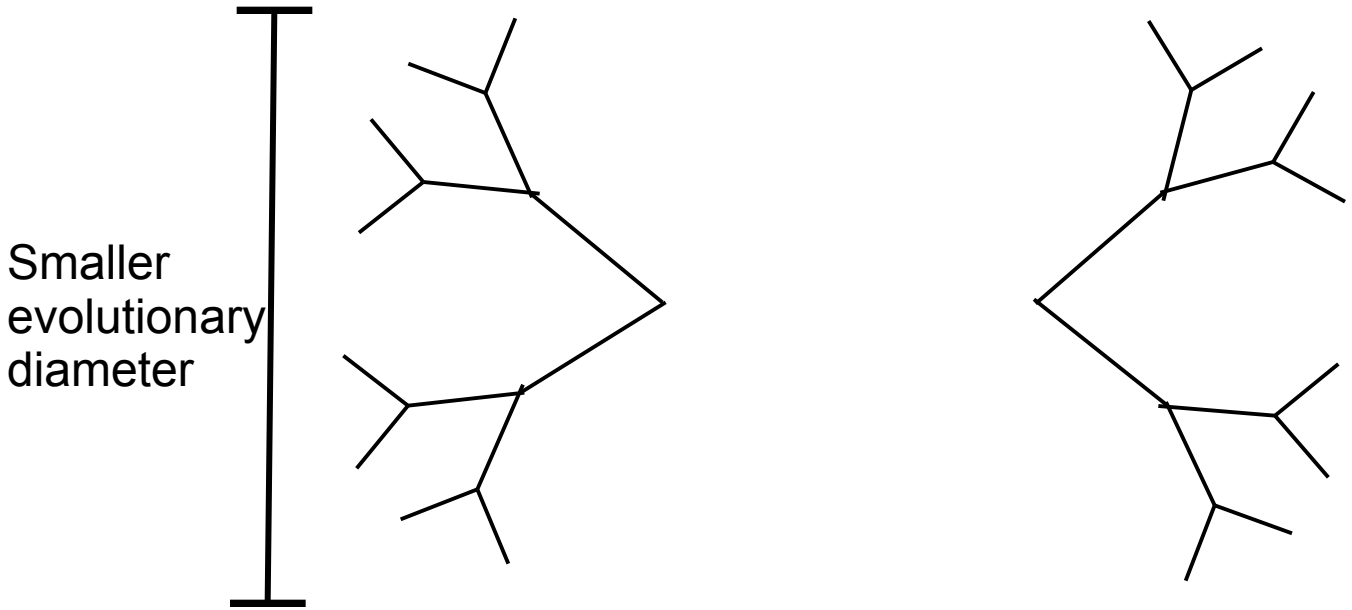
New approach



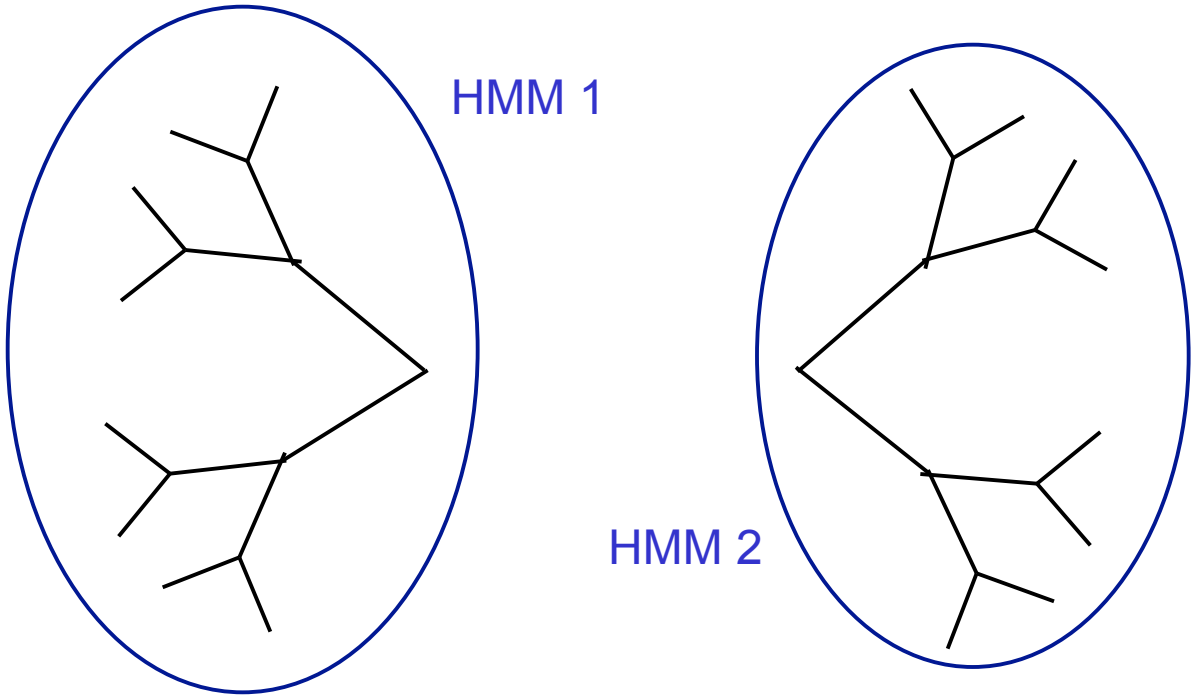
New approach



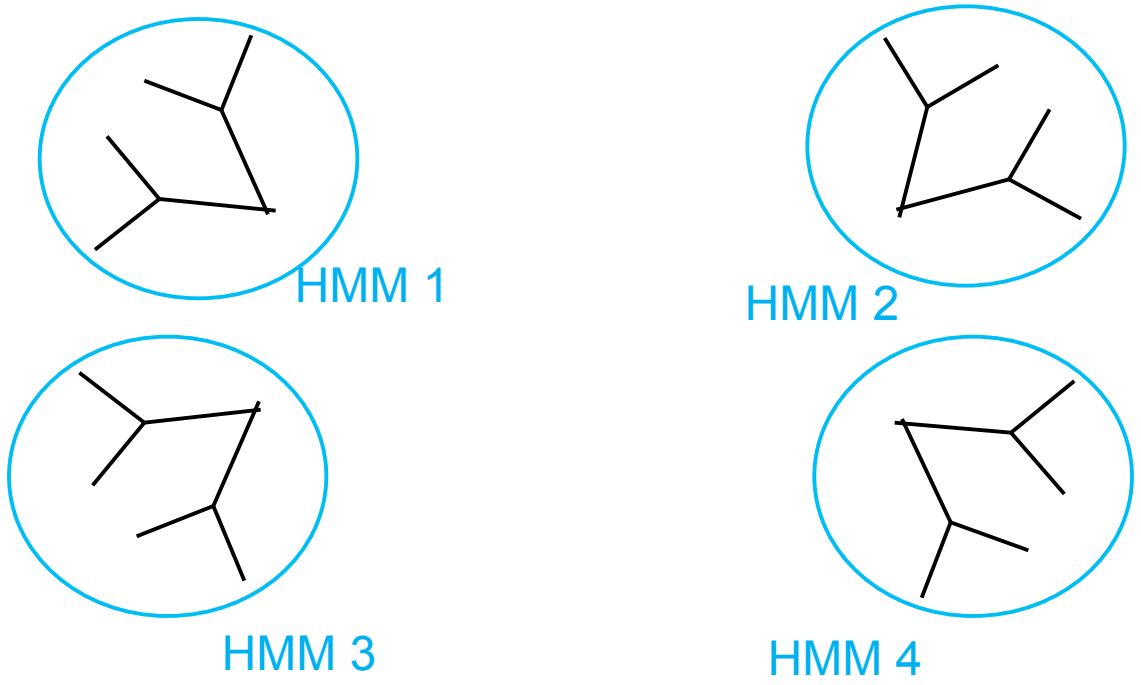
New approach



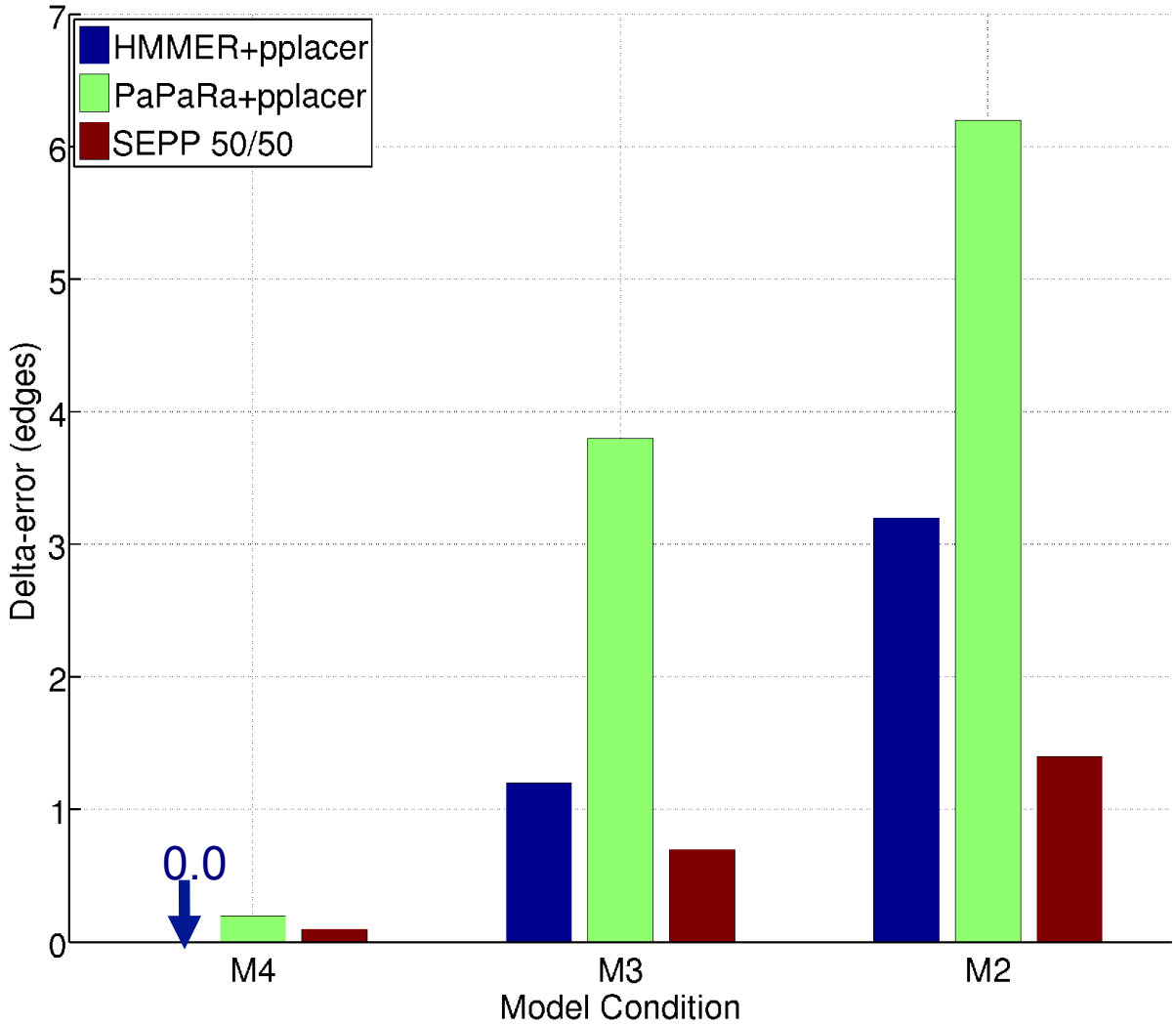
New approach



New approach



SEPP (10% rule) Simulated Results



Backbone size: 500
5000 fragments
20 replicates

Using SEPP

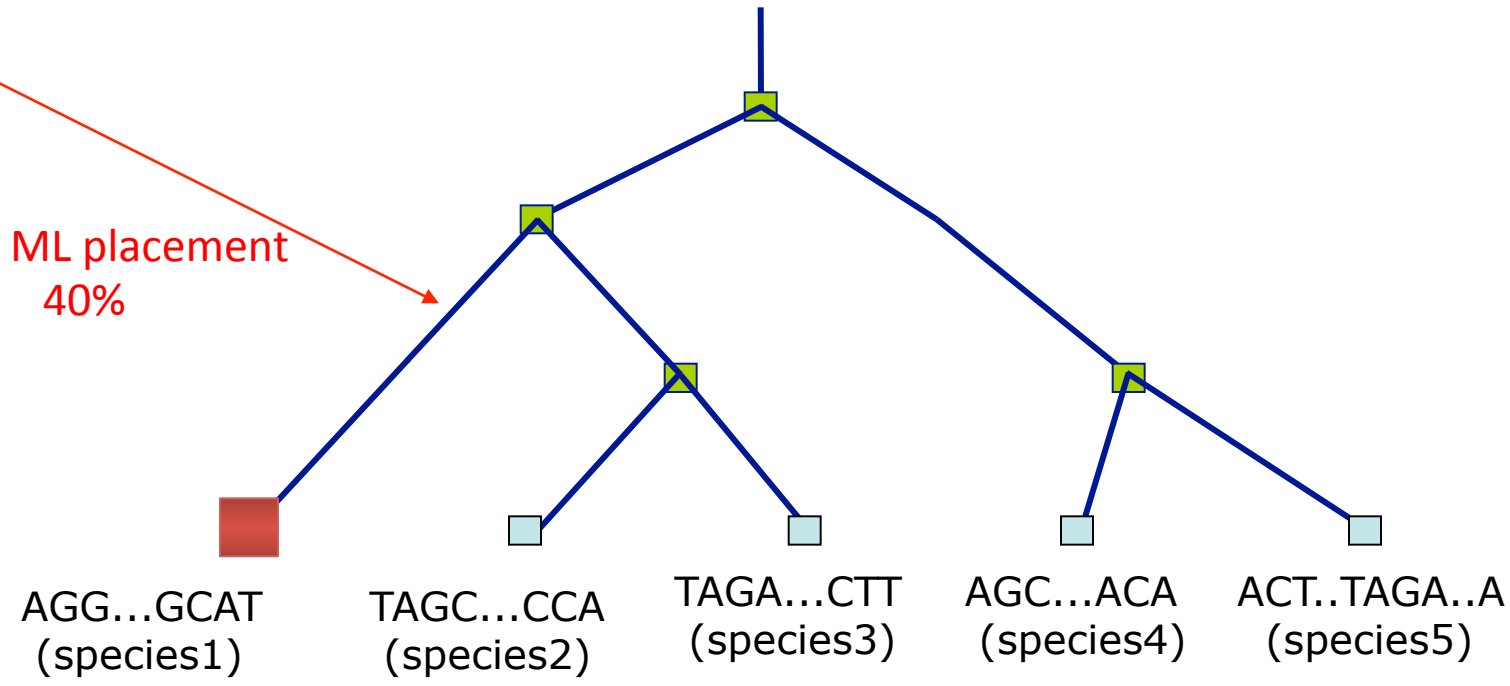
Fragmentary Unknown Reads:

(60-200 bp long)

- ACCG
- CGAG
- CGG
- GGCT
- TAGA
- GGGGG
- TCGAG
- GGCG
- GGG
- .
- .
- .
- ACCT

Known Full length Sequences,
and an alignment and a tree

(500-10,000 bp long)



Adding Uncertainty

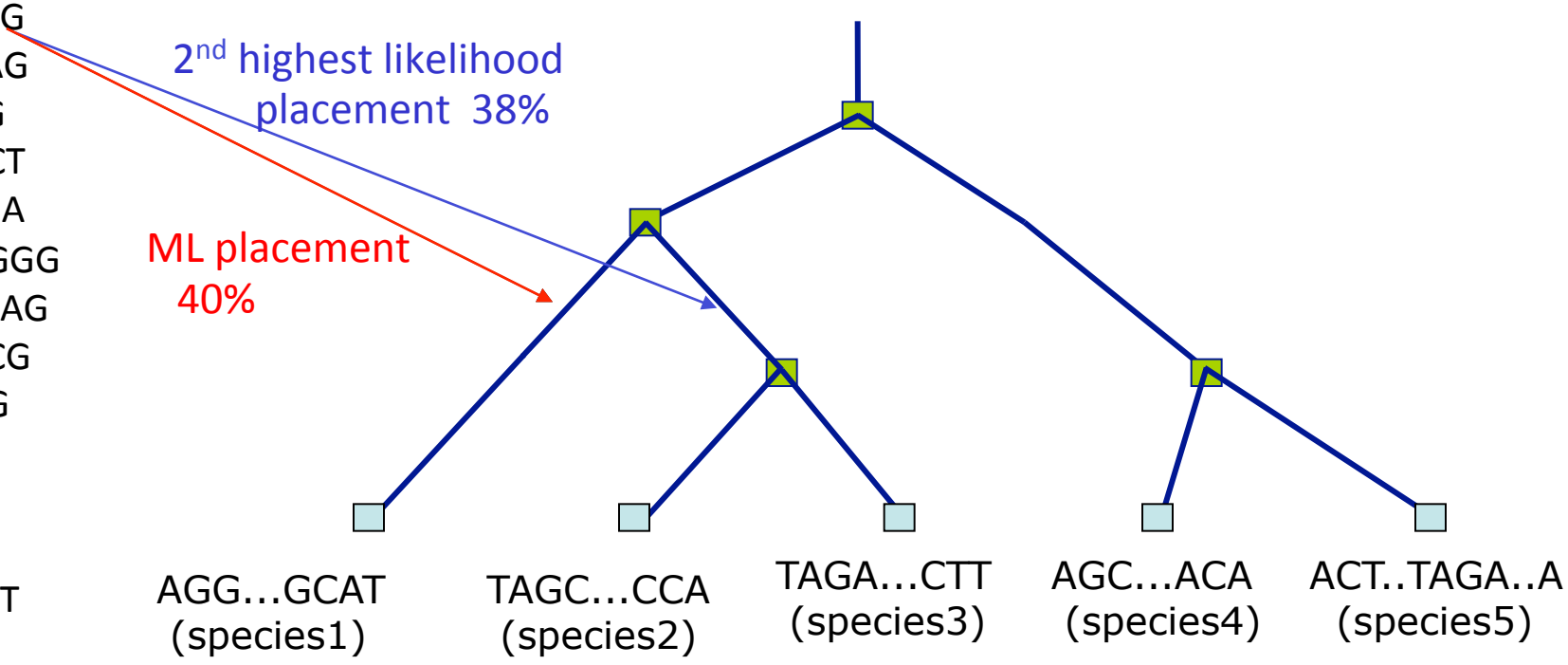
Fragmentary Unknown Reads:

(60-200 bp long)

- ACCG
- CGAG
- CGG
- GGCT
- TAGA
- GGGGG
- TCGAG
- GGCG
- GGG
- .
- .
- .
- ACCT

Known Full length Sequences,
and an alignment and a tree

(500-10,000 bp long)



Adding Uncertainty

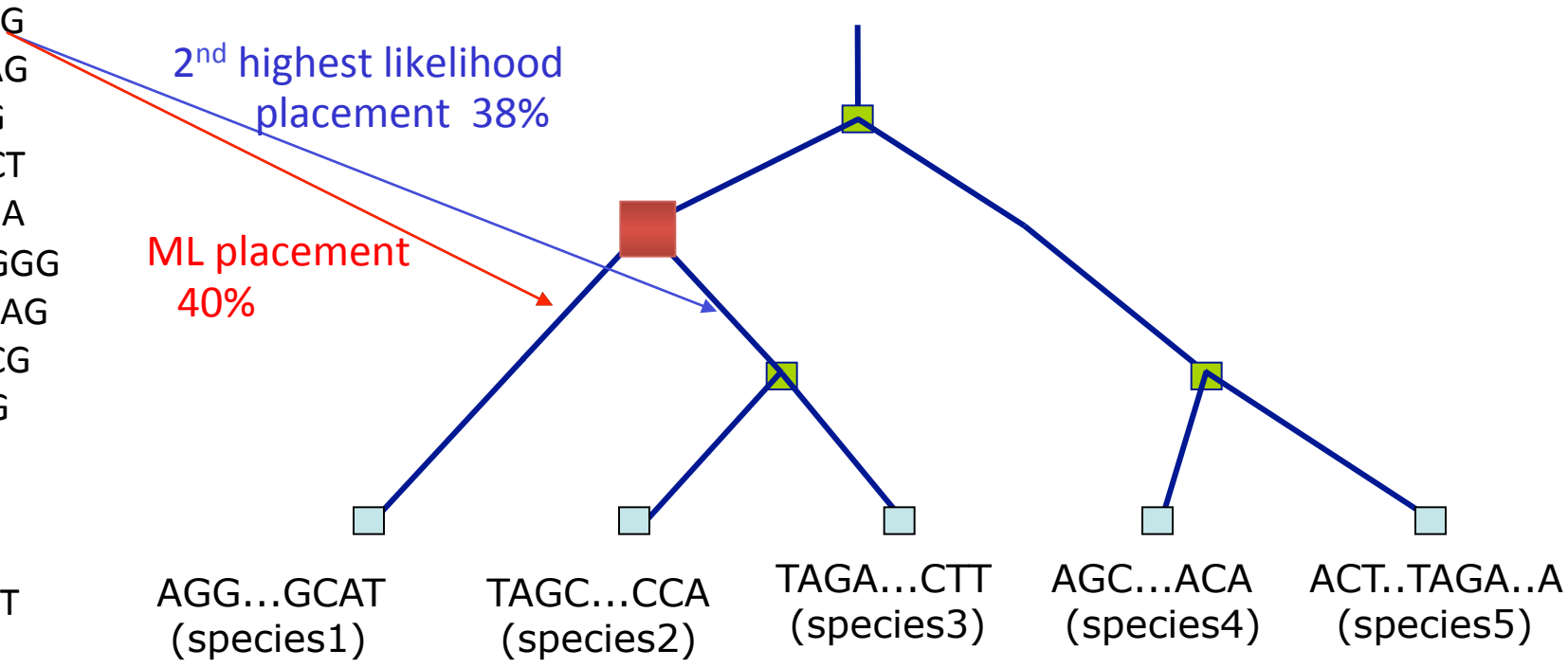
Fragmentary Unknown Reads:

(60-200 bp long)

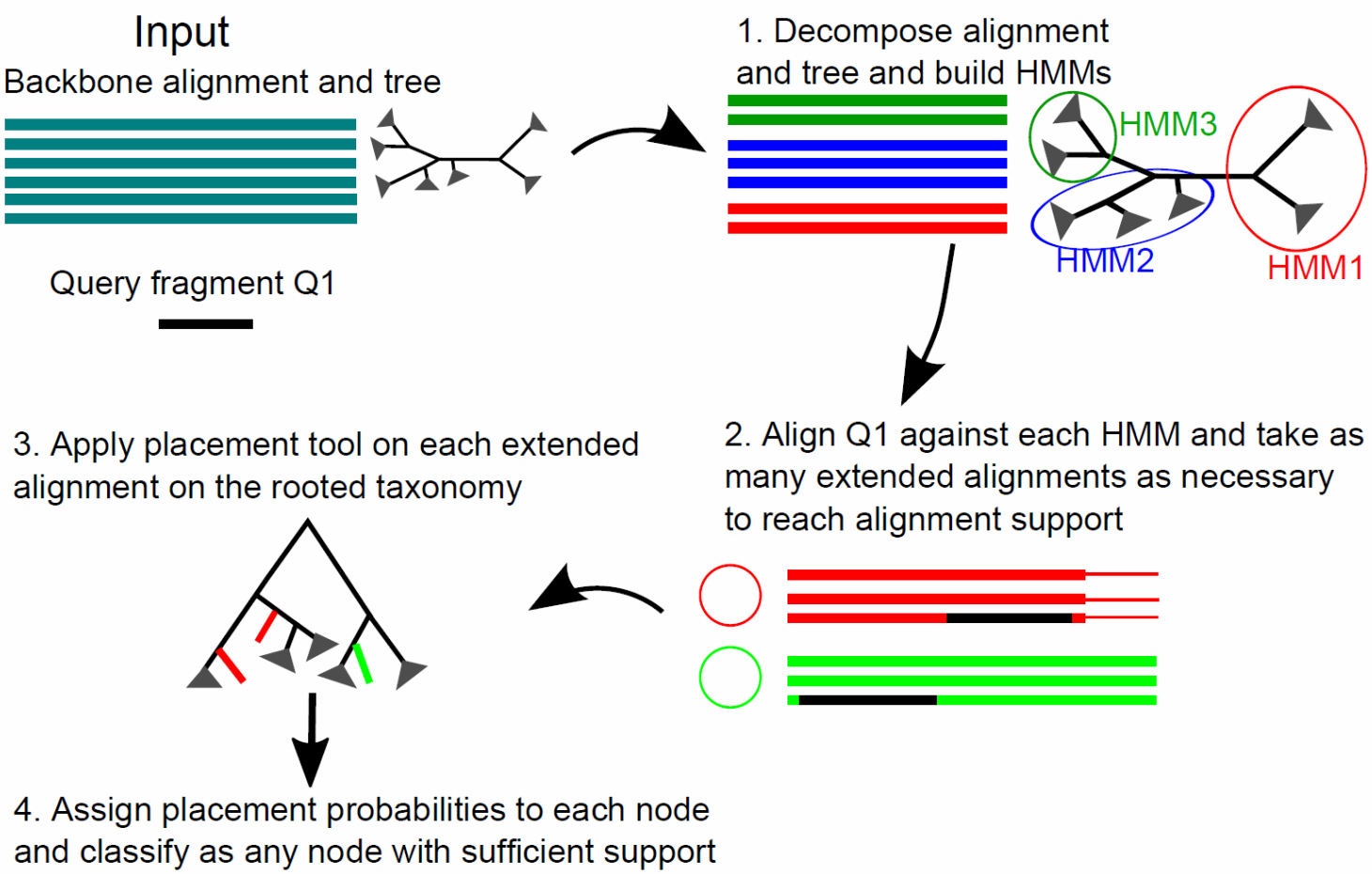
- ACCG
- CGAG
- CGG
- GGCT
- TAGA
- GGGGG
- TCGAG
- GGCG
- GGG
- .
- .
- .
- ACCT

Known Full length Sequences,
and an alignment and a tree

(500-10,000 bp long)



TIPP



Nguyen et al. Bioinformatics 2014

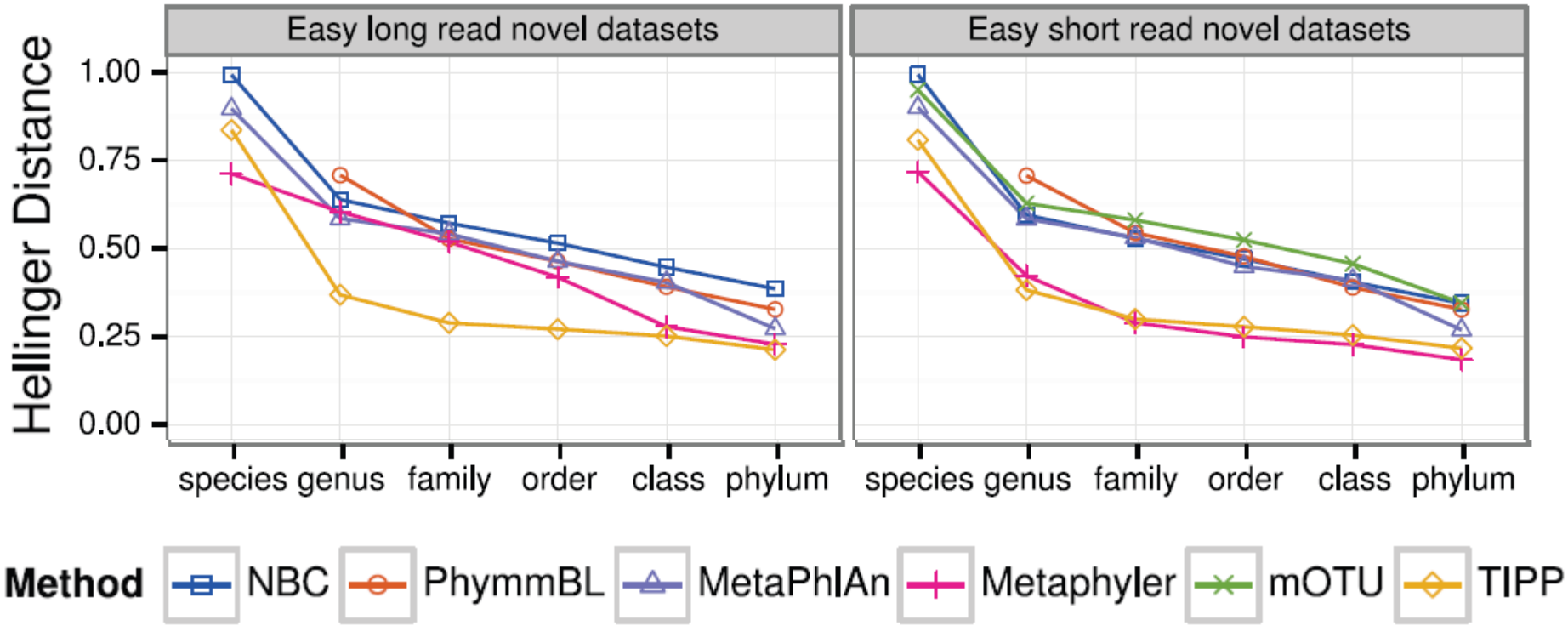
TIPP for Taxonomic Profiling

- Marker-based abundance profiler
- Uses a collection of single copy housekeeping genes
- Only fragments binned to marker genes classified
- Profiling algorithm
  - Bins fragments to marker genes
  - Classify fragments binned to each marker
  - Pool all classified reads
  - Estimate abundance profile on pooled reads

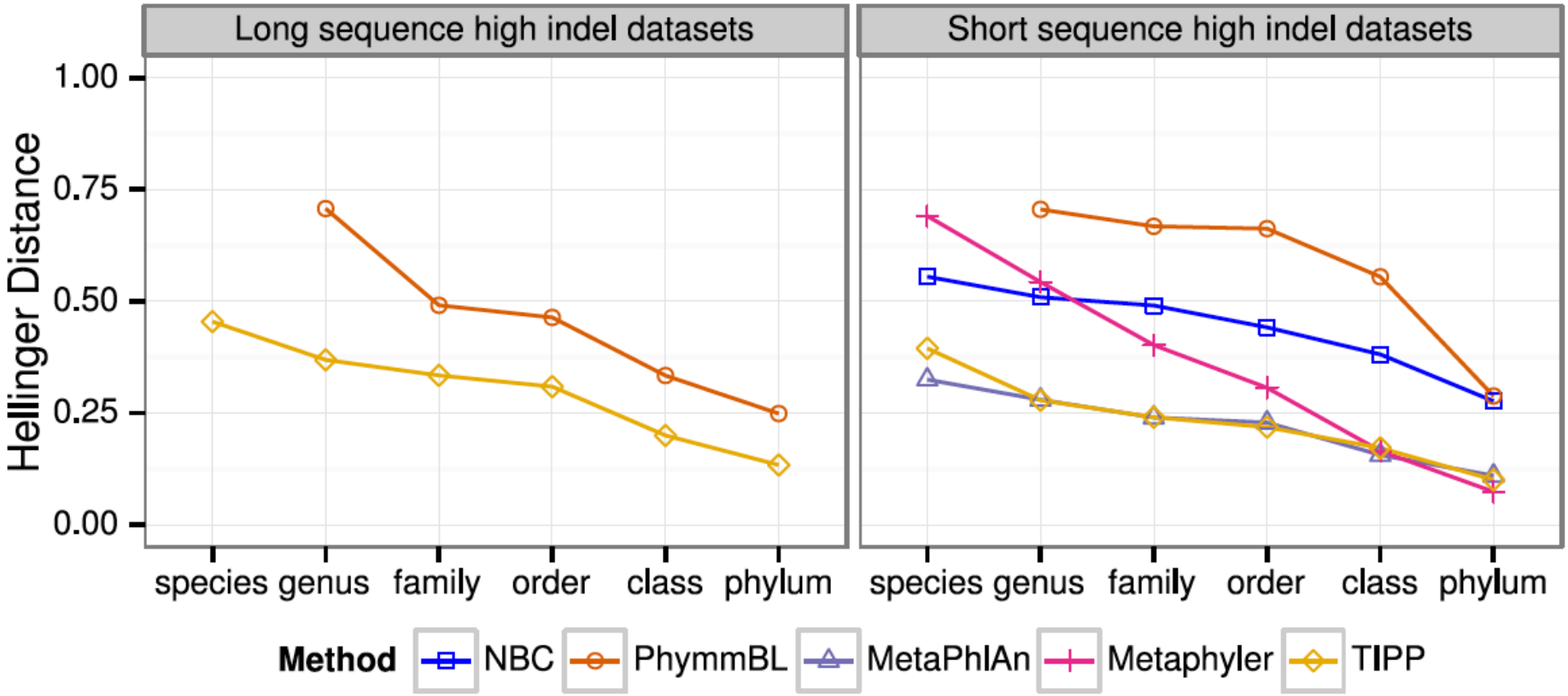
Taxonomic Profiling Experimental Design

- Datasets
  - Easy conditions (low error rates, known genomes)
  - Hard conditions (novel genomes, high error rates)
- Methods
  - Marker-based – TIPP, Metaphyler, mOTU, Metaphlan
  - Genome-based – NBC, PhymmBL
- Measured distance to true profile as error metric

“Easy” genome datasets

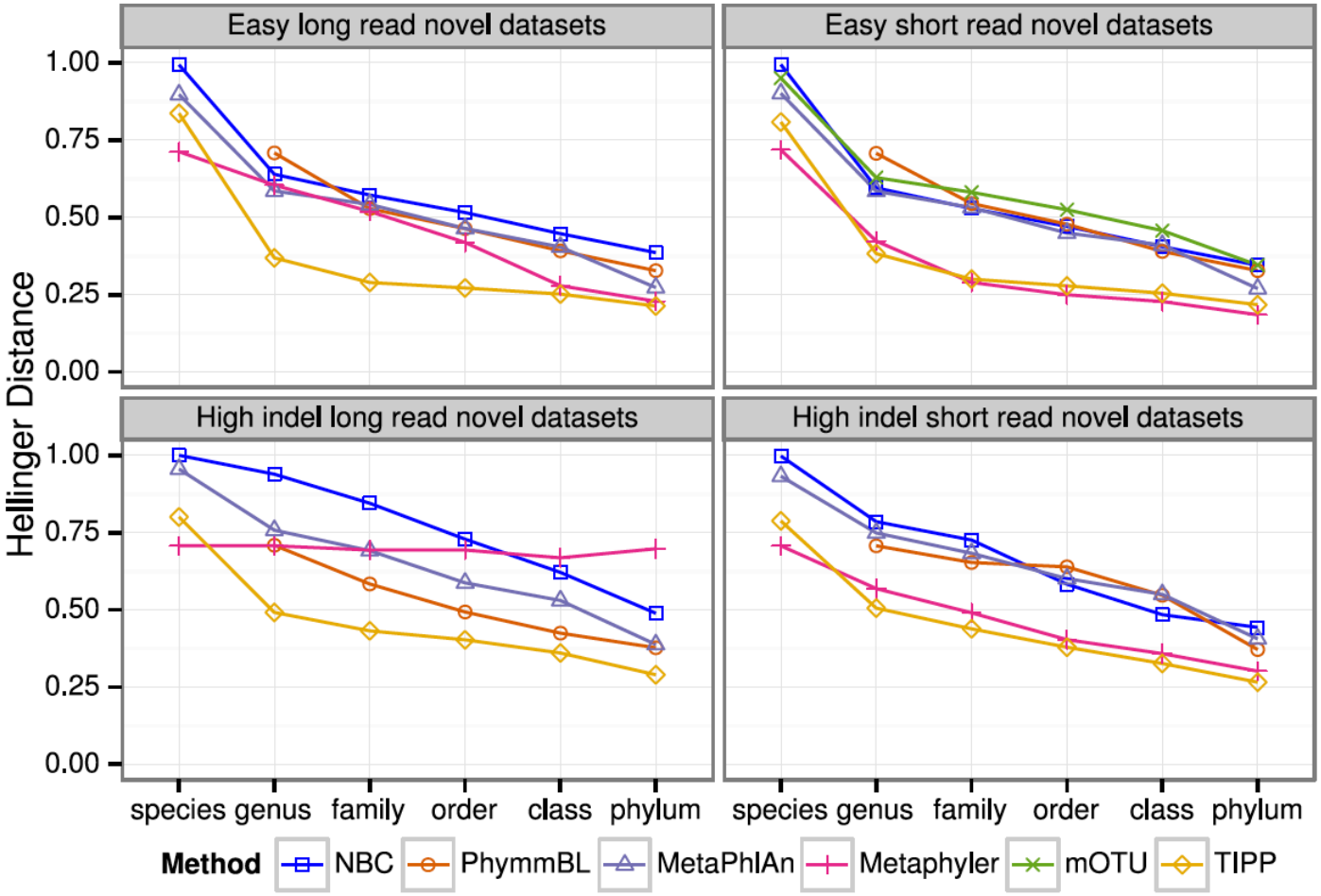


“Hard” genome datasets



Note: NBC, MetaPhlAn, and Metaphyler cannot classify any sequences from at least of the high indel long sequence datasets. mOTU terminates with an error message on all the high indel datasets.

“Novel” genome datasets



Note: mOTU terminates with an error message on the long fragment datasets and high indel datasets.

Summary

- TIPP: marker-based taxonomic identification and classification method through phylogenetic placement
  - Very robust to sequencing errors and novel genomes
  - Results in overall more accurate profiles
- Accurate profiles can be obtained by classifying reads from the marker genes

Acknowledgements

Siavash Mirarab

Bo Liu

Mihai Pop

Tandy Warnow

Supported by
NSF DEB 0733029
University of Alberta

SEPP/TIPP/UPP

SEPP/UPP/TIPP site:
https://github.com/smirarab/sepp/

Instructions for installing UPP:
https://github.com/smirarab/sepp/blob/master/tutorial/upp-tutorial.md

Instructions for installing TIPP:
https://github.com/smirarab/sepp/blob/master/tutorial/tipp-tutorial.md

References:

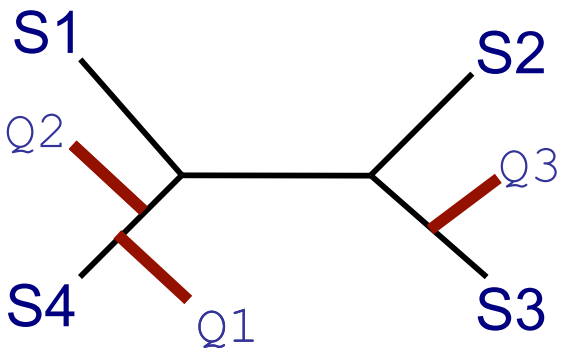
1) N. Nguyen, S. Mirarab, K. Kumar, and T. Warnow. Ultra-large alignments using phylogeny-aware profiles, Proceedings of Research in Computational Biology (RECOMB) 2015 and to appear in Genome Biology 2015.

1) N. Nguyen, S. Mirarab, B. Liu, M. Pop, and T. Warnow. TIPP:Taxonomic Identification and Phylogenetic Profiling. Bioinformatics, 2014, 30 (24): 3548-3555.

2) Mirarab, S., N. Nguyen, and T. Warnow, 2012. SEPP: SATe-Enabled Phylogenetic Placement. Pacific Symposium on Biocomputing.

Place Sequence

```
S1   = -AGGCTATCACCTGACCTCCA-AA
S2   = TAG-CTATCAC--GACCGC--GCA
S3   = TAG-CT-------GACCGC--GCT
S4   = TAC----TCAC--GACCGACAGCT
Q1   = -------T-A--AAAC--------
```

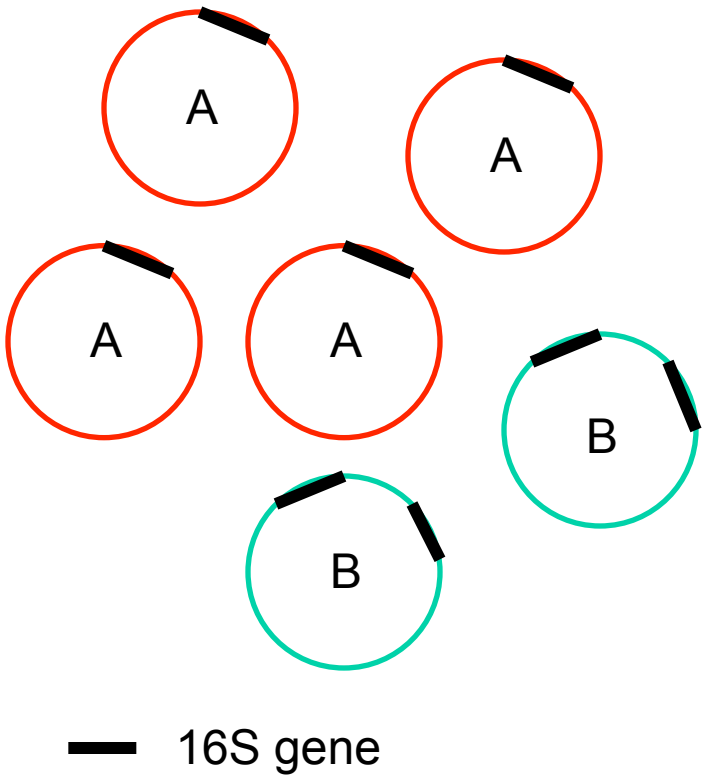


Query sequences are aligned and placed independently

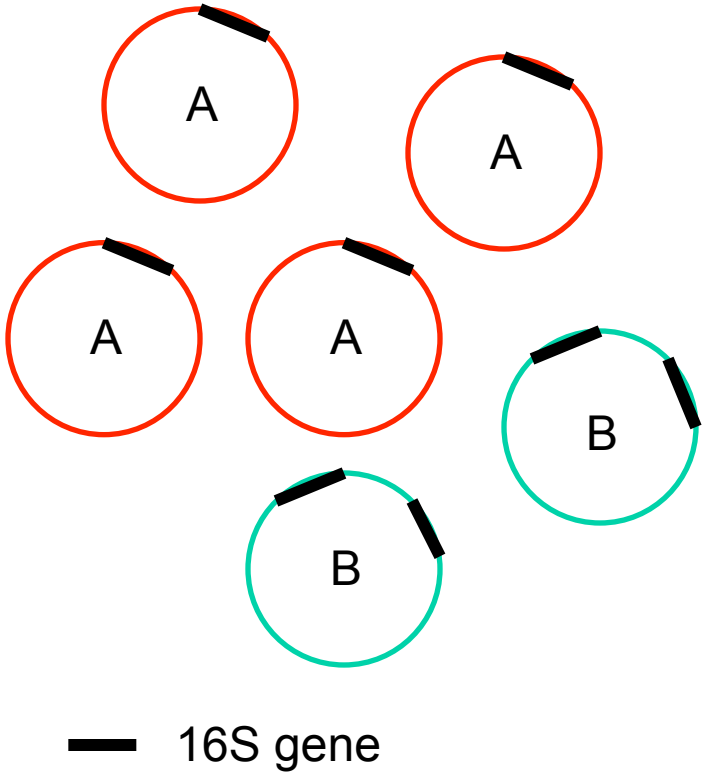
Phylogenetic Placement

- Align each query sequence to backbone alignment:
  - HMMALIGN (Eddy, Bioinformatics 1998)
  - PaPaRa (Berger and Stamatakis, Bioinformatics 2011)
- Place each query sequence into backbone tree, using extended alignment:
  - pplacer (Matsen et al., BMC Bioinformatics 2010)
  - EPA (Berger et al., Systematic Biology 2011)

16S Identification



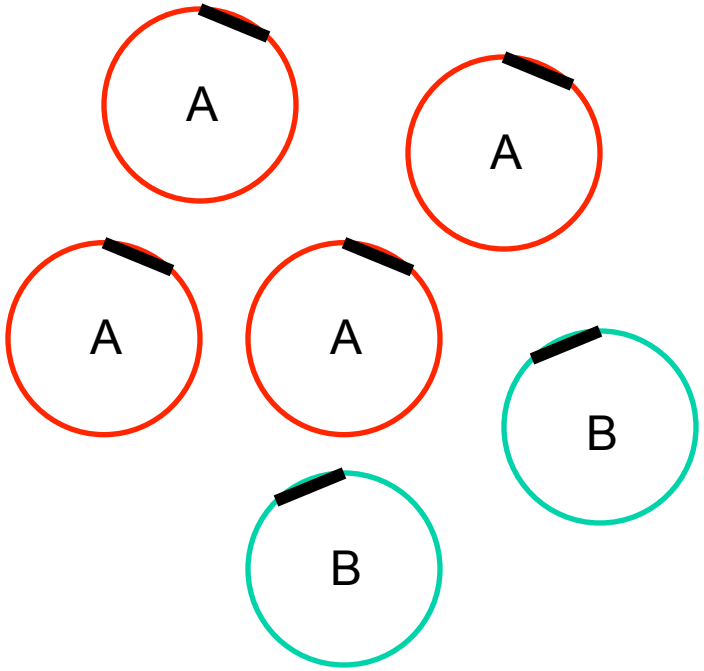
16S Identification



True Abundance
A: 67% B: 33%

Estimated Abundance
A: 50% B: 50%

Single copy gene



True Abundance
A: 67% B: 33%

Estimated Abundance
A: 67% B: 33%

TIPP: Taxonomic identification and Phylogenetic Profiling

- Developers: Nguyen, Mirarab, Pop, and Warnow
- SEPP takes the best extended alignment and finds the ML placement.
- Modify SEPP to use uncertainty:
  - Take as many alignments necessary to reach support alignment threshold
  - Classify query sequence at node with sufficient placement support threshold
- Nguyen et al. Bioinformatics 2014