# Using “HMM Families” in Bioinformatics (or SEPP, TIPP, and UPP)

Tandy Warnow

The Department of Computer Science

The University of Texas at Austin

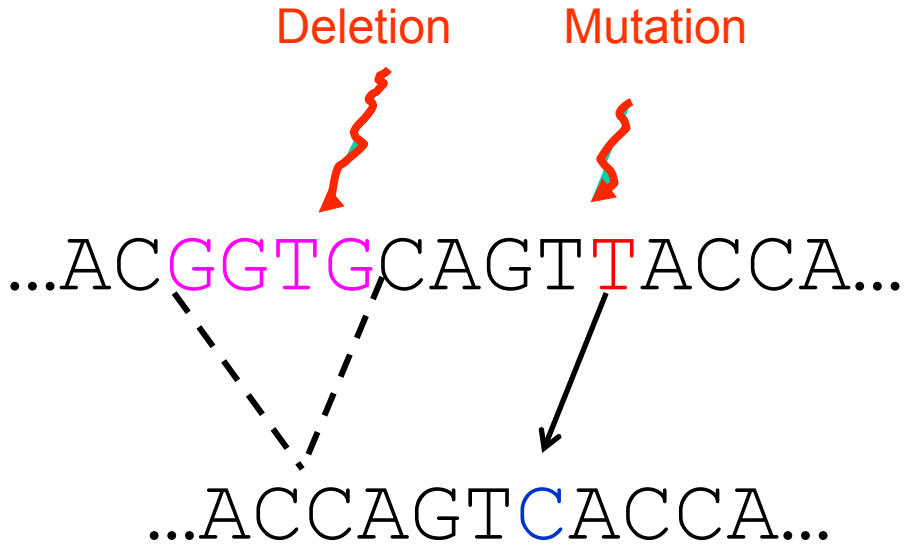
# SEPP, TIPP, and UPP

- SEPP: SATe-enabled Phylogenetic Placement (Mirarab, Nugyen and Warnow, PSB 2012)
- TIPP: Taxon identification and phylogenetic profiling (Nguyen, Mirarab, Liu, Pop, and Warnow, submitted)
- UPP: Ultra-large multiple sequence alignment using SEPP (Nguyen, Mirarab, Kumar, Wang, Guo, Kim, and Warnow, in preparation)

# Multiple Sequence Alignment

- Indels, and why we need to align sequences
- Poor performance of standard methods on large datasets
- SATé (Liu et al., Science 2009 and Systematic Biology 2012)
- Limitations for large datasets
- Additional challenges on real datasets

# Indels (insertions and deletions)




**The true multiple alignment**

- **Reflects historical substitution, insertion, and deletion events**
- **Defined using transitive closure of pairwise alignments computed on edges of the true tree**

# Input: unaligned sequences

```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```

# Phase 1: Alignment

```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```



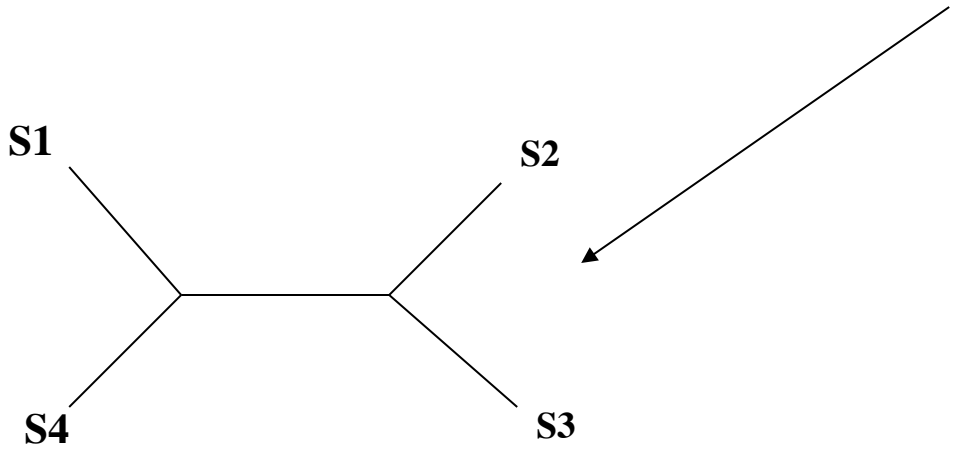
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

# Phase 2: Construct tree

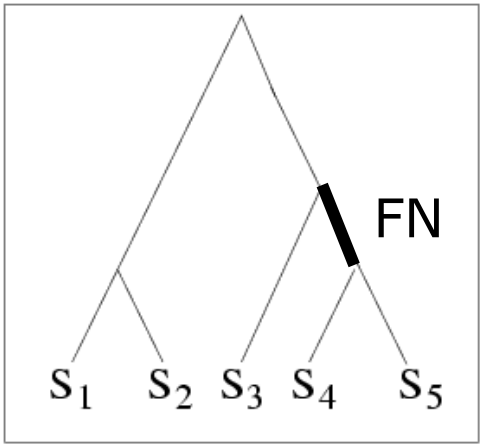
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```

→

```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```



# Quantifying Error



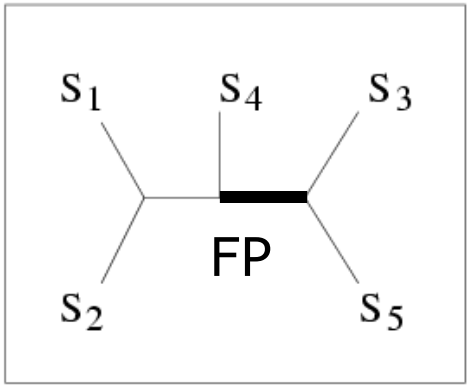

TRUE TREE

| | |
|---|---|
| $S_1$ | ACAATTAGAAC |
| $S_2$ | ACCCTTAGAAC |
| $S_3$ | ACCATTCCAAC |
| $S_4$ | ACCAGACCAAC |
| $S_5$ | ACCAGACCGGA |

DNA SEQUENCES

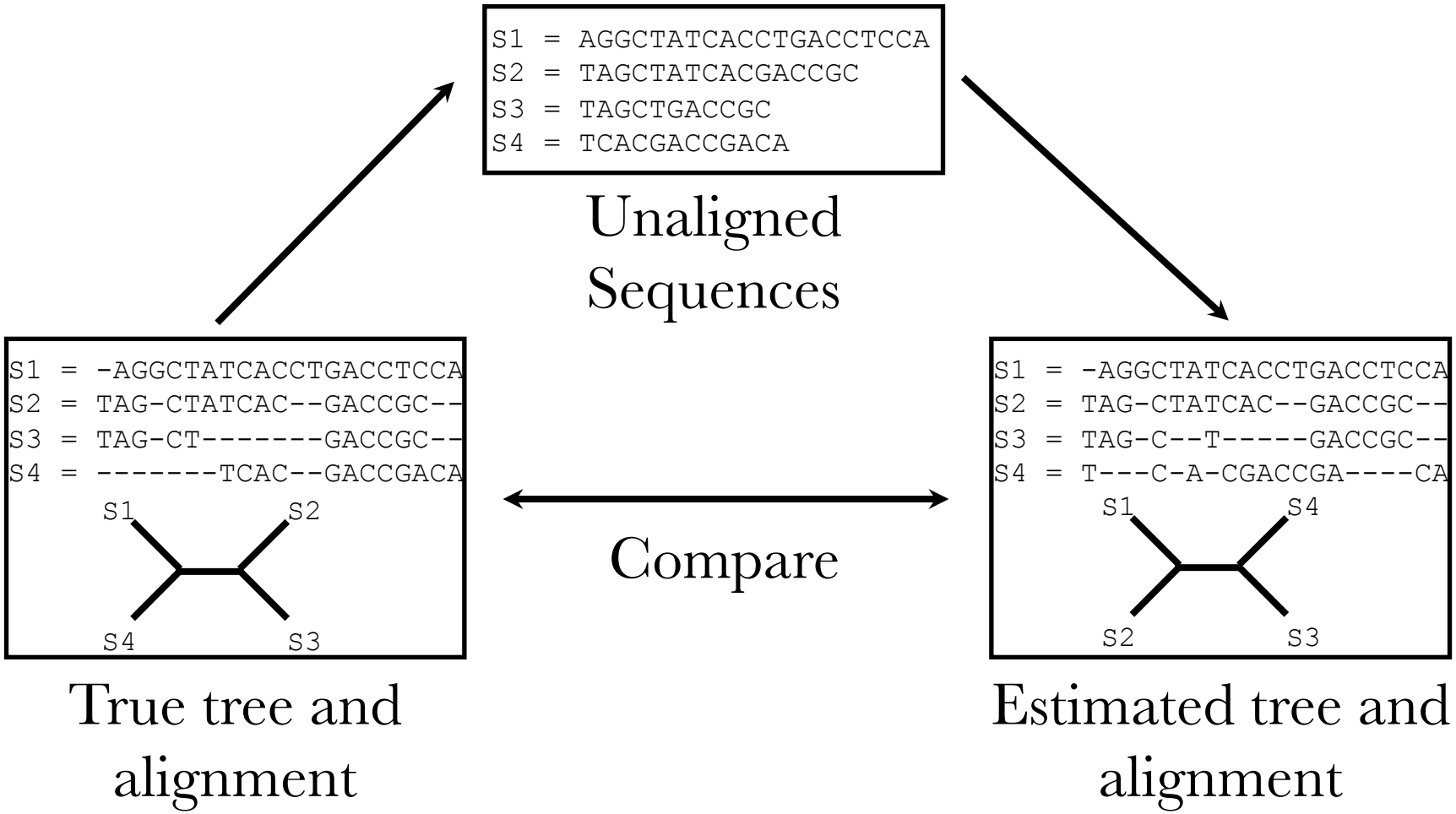
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate




INFERRED TREE

# Simulation Studies



# Two-phase estimation

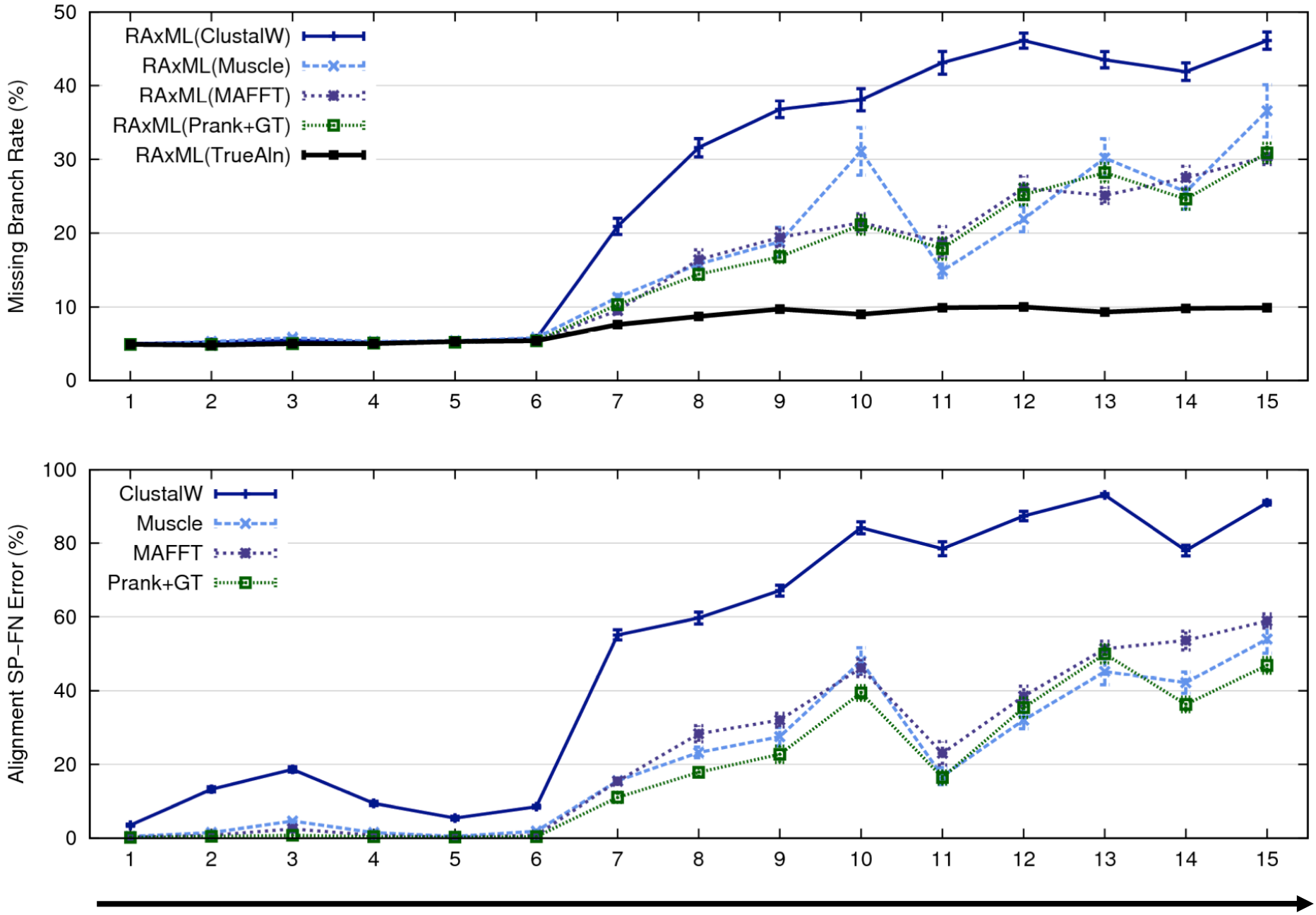
Alignment methods

- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- Probalign
- MAFFT
- Muscle
- Di-align
- T-Coffee
- Prank (PNAS 2005, Science 2008)
- Opal (ISMB and Bioinf. 2007)
- *FSA (PLoS Comp. Bio. 2009)*
- *Infernal (Bioinf. 2009)*
- Etc.

Phylogeny methods

- Bayesian MCMC
- Maximum parsimony
- Maximum likelihood
- Neighbor joining
- FastME
- UPGMA
- Quartet puzzling
- Etc.

***RAxML**: heuristic for large-scale ML optimization*


Missing Branch Rate (%)
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Alignment SP–FN Error (%)
ClustalW
Muscle
MAFFT
Prank+GT
1000-taxon models, ordered by difficulty (Liu et al., 2009)

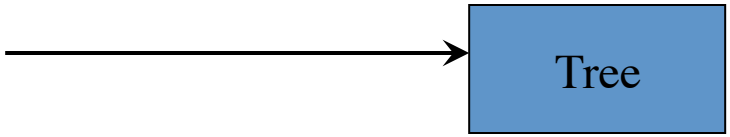
# Problems with the two-phase approach

- Current alignment methods fail to return reasonable alignments on large datasets with high rates of indels and substitutions.
- Manual alignment is time consuming and subjective.
- *Systematists discard potentially useful markers* if they are difficult to align.

This issues seriously impact large-scale phylogeny estimation (and Tree of Life projects)

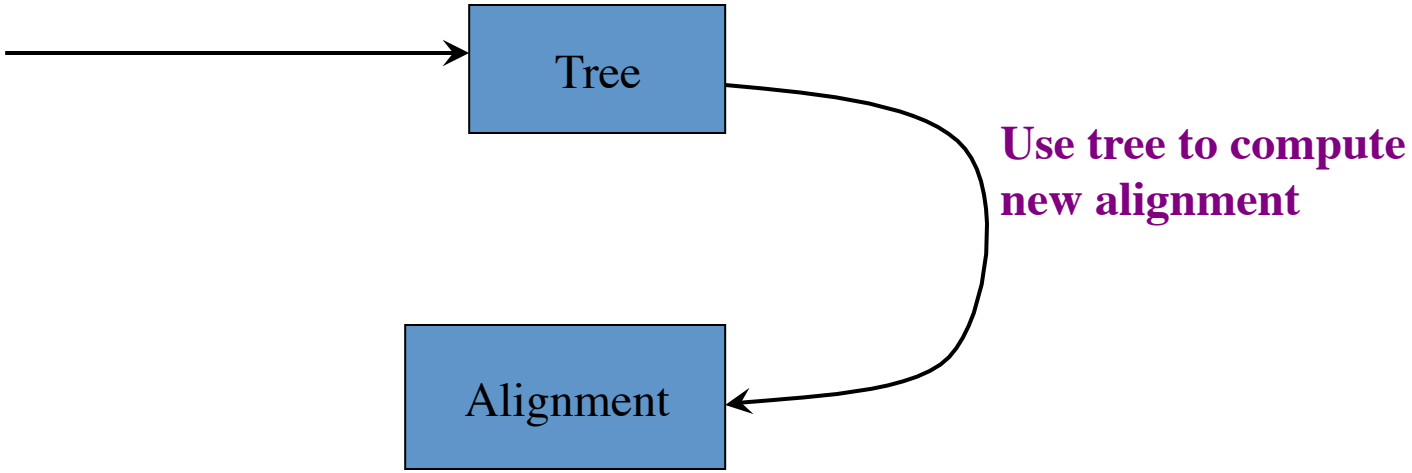
# SATé Algorithm (Science 2009)

Obtain initial alignment and
estimated ML tree



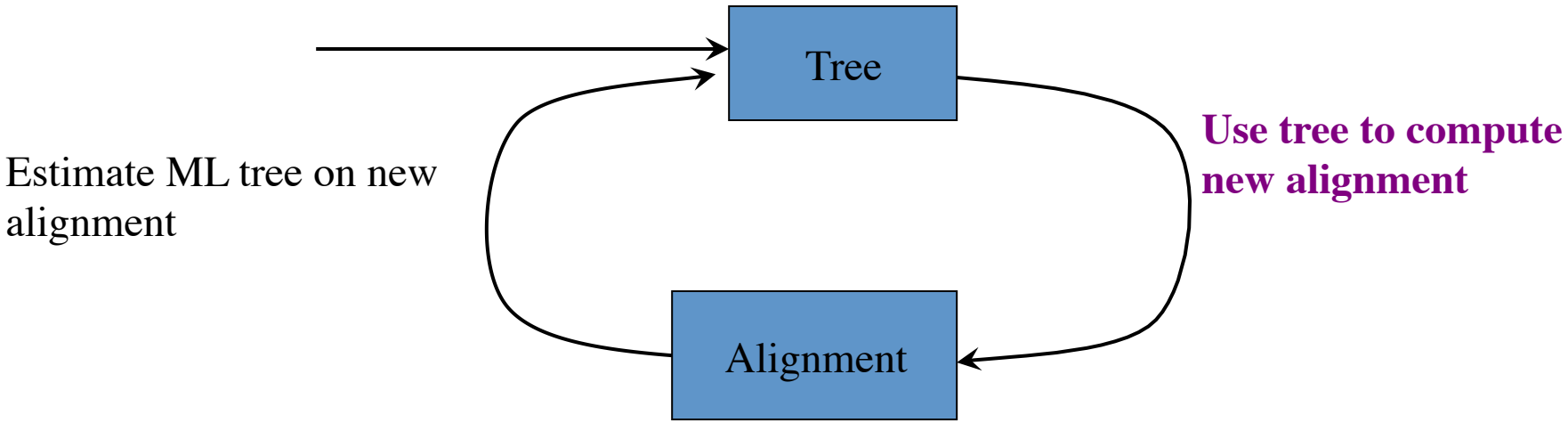
# SATé Algorithm

Obtain initial alignment and
estimated ML tree



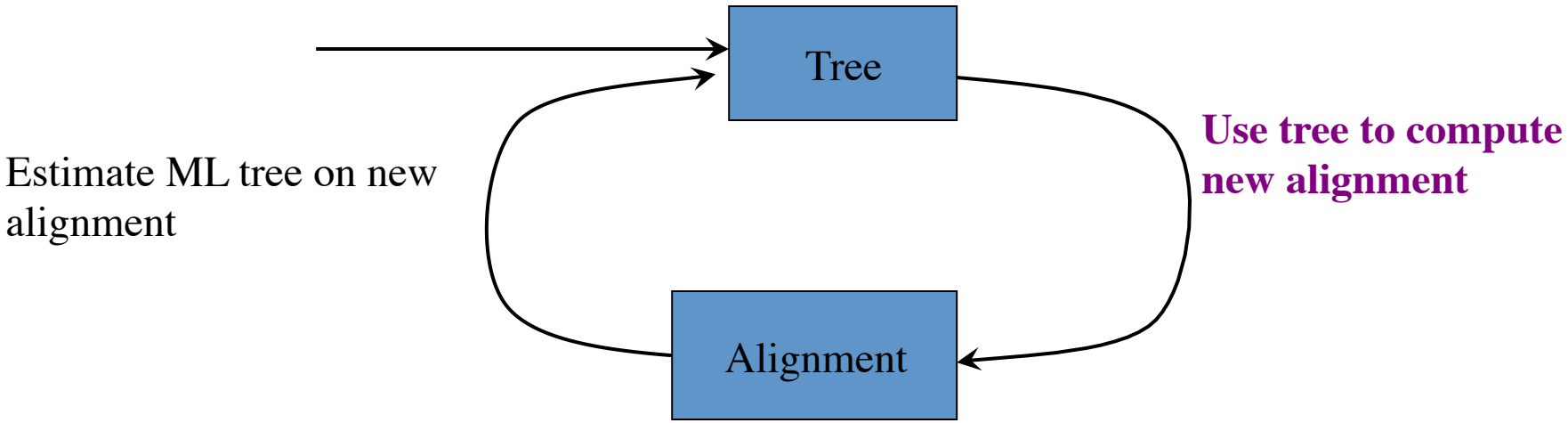
# SATé Algorithm

Obtain initial alignment and estimated ML tree



# SATé Algorithm

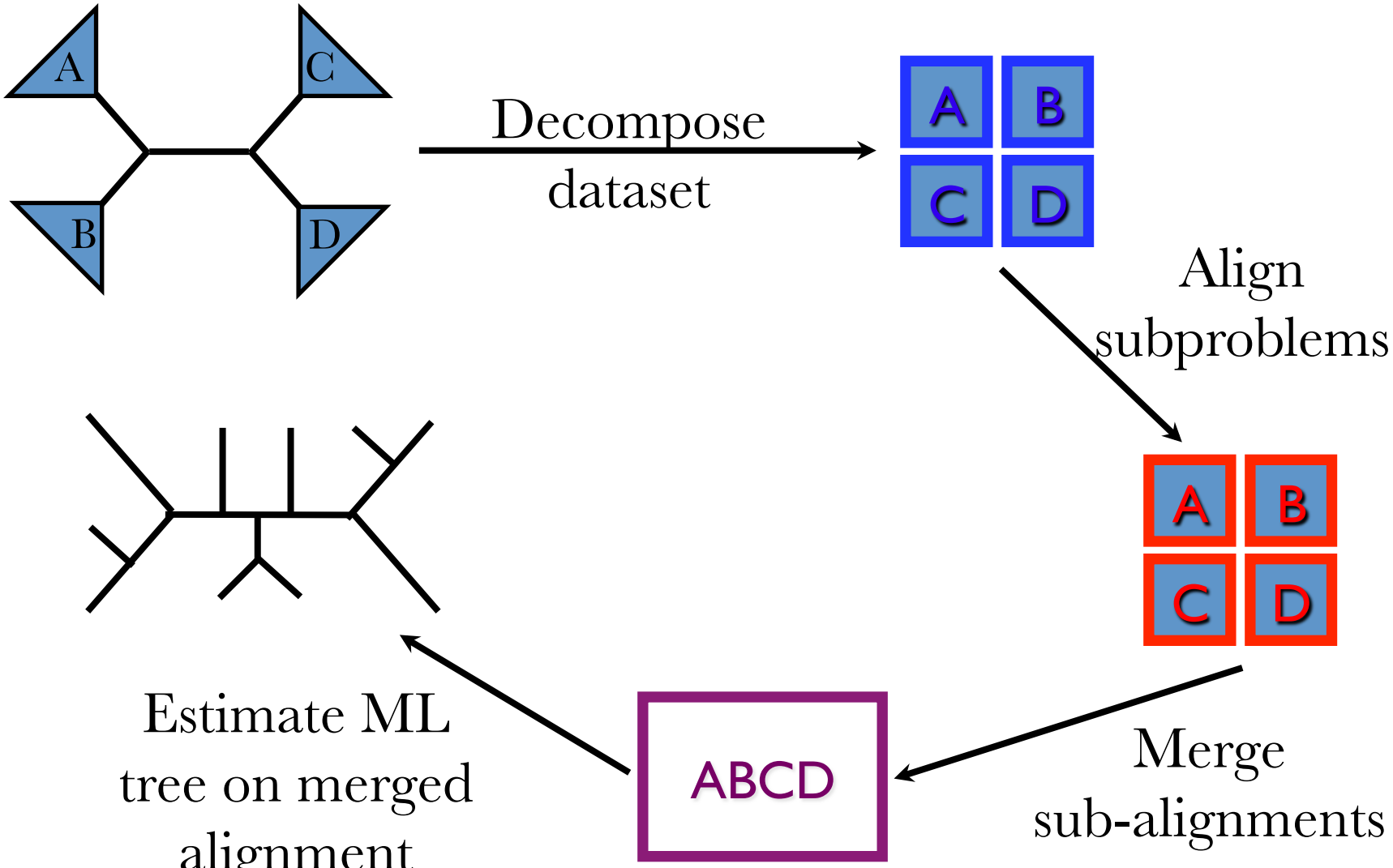
Obtain initial alignment and estimated ML tree

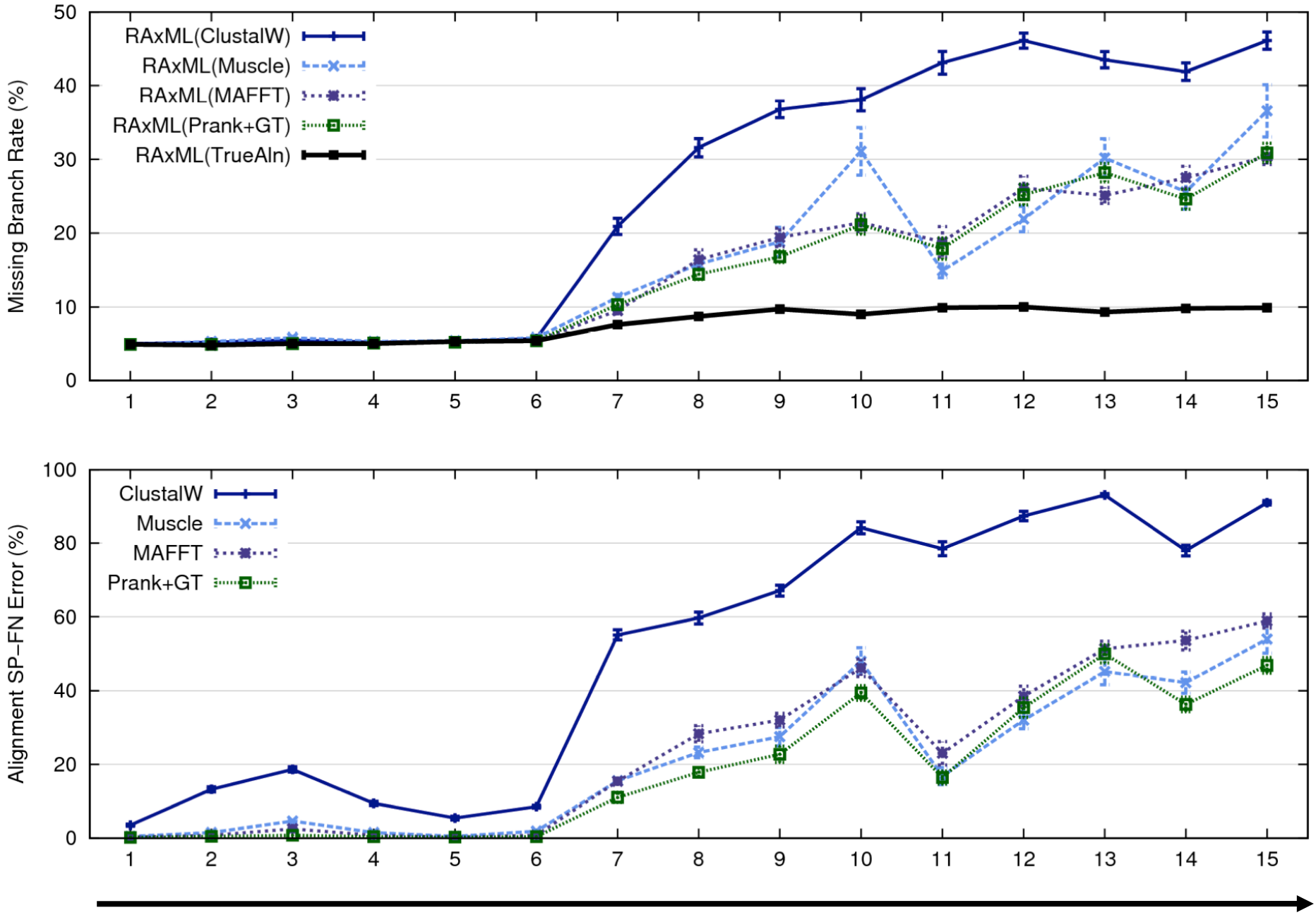


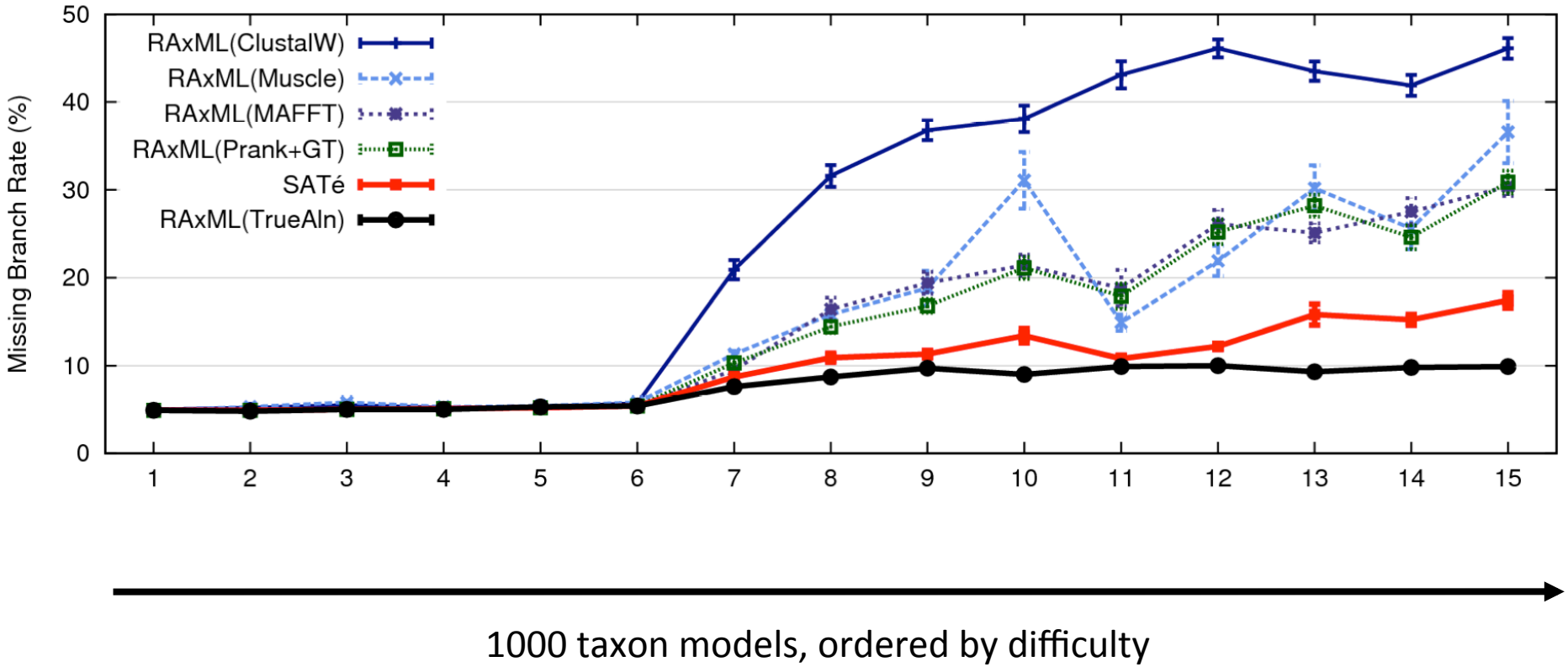

If new alignment/tree pair has worse ML score, realign using a different decomposition

Repeat until termination condition (typically, 24 hours)

# Each SATe iteration



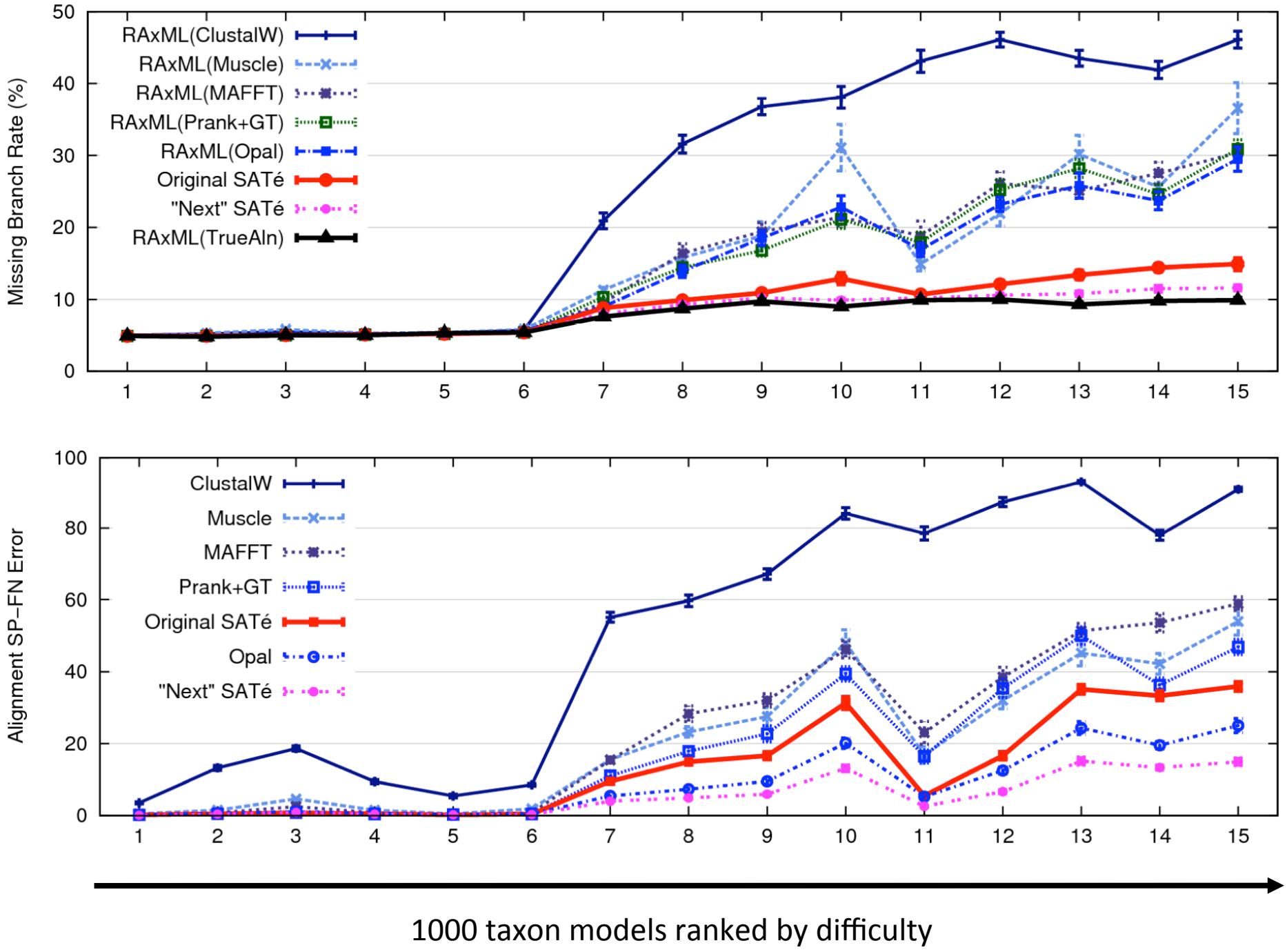

Missing Branch Rate (%)
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Alignment SP–FN Error (%)
ClustalW
Muscle
MAFFT
Prank+GT
1000-taxon models, ordered by difficulty (Liu et al., 2009)



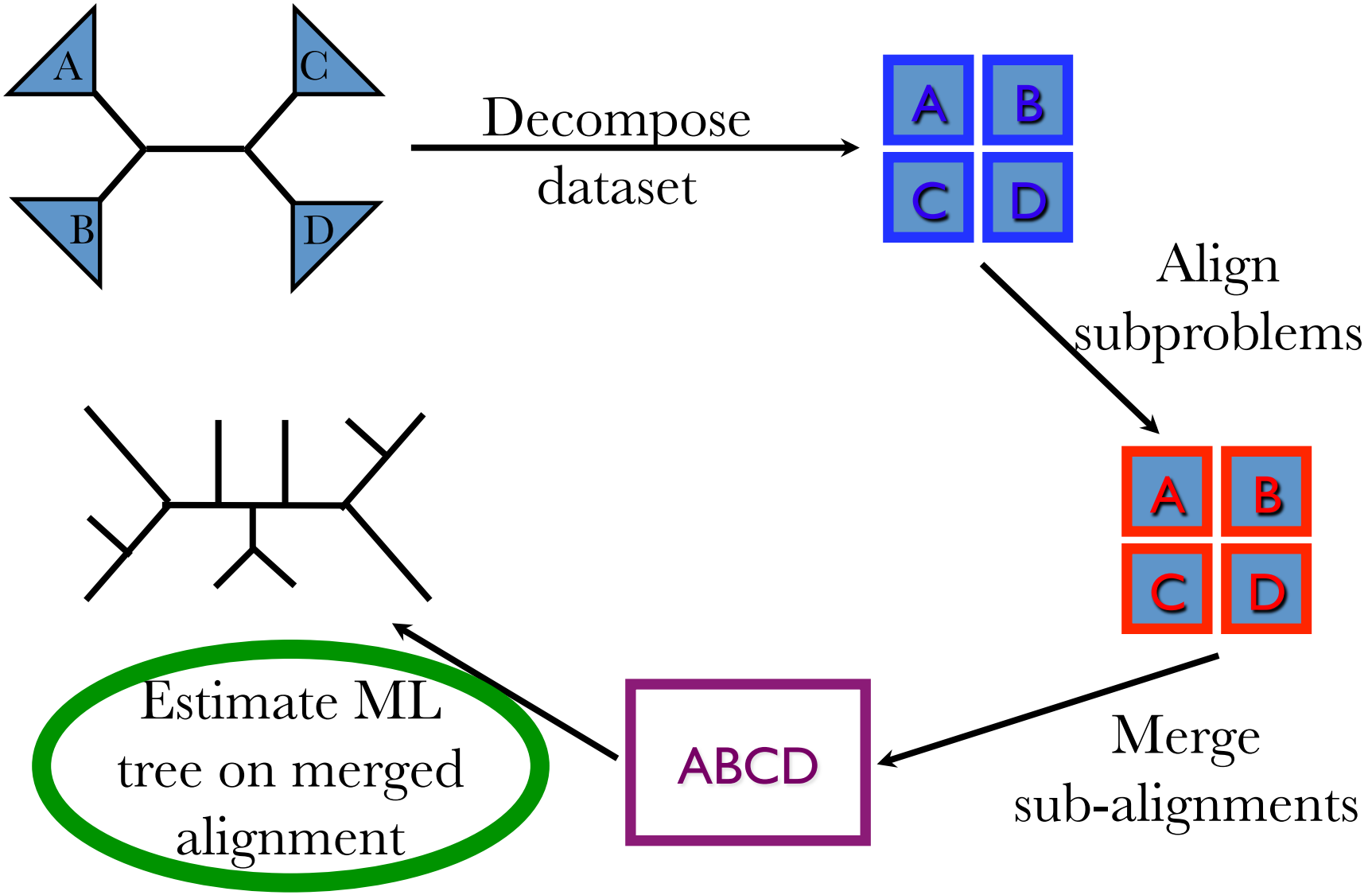

1000 taxon models, ordered by difficulty

24 hour SATé analysis, on desktop machines

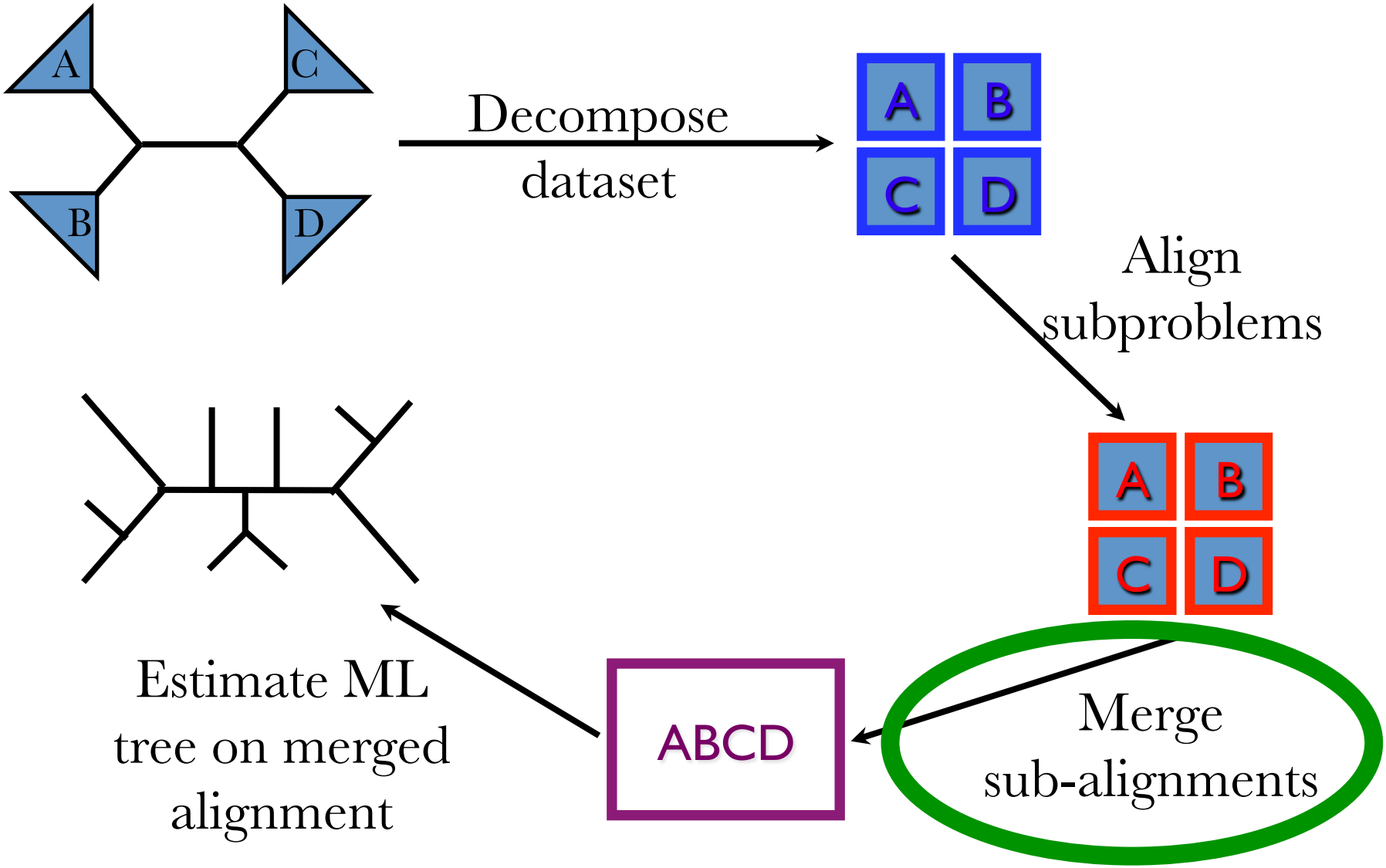
(Similar improvements for biological datasets)


Missing Branch Rate (%)
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(Opal)
Original SATé
"Next" SATé
RAxML(TrueAln)
Alignment SP–FN Error
ClustalW
Muscle
MAFFT
Prank+GT
Original SATé
Opal
"Next" SATé
1000 taxon models ranked by difficulty

# Limitations



# Limitations



# 1kp: Thousand Transcriptome Project

G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

N. Wickett
Northwestern

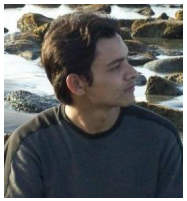
N. Matasci
iPlant

T. Warnow,
UT-Austin

S. Mirarab,
UT-Austin

N. Nguyen,
UT-Austin

Md. S.Bayzid
UT-Austin

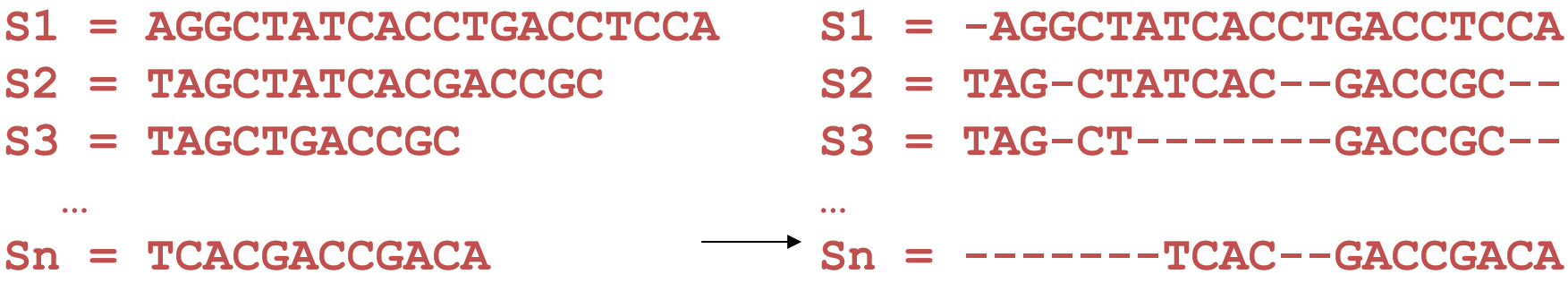


Plus many many other people…

- Plant Tree of Life based on transcriptomes of ~1200 species
- More than 13,000 gene families (most not single copy)

**Challenge:**
**Alignment of datasets with > 100,000 sequences with many fragmentary sequences**

# Multiple Sequence Alignment (MSA): *another grand challenge*[1]




*Novel techniques needed* for scalability and accuracy

NP-hard problems and large datasets

Current methods do not provide good accuracy

Few methods can analyze even moderately large datasets

*Many important applications besides phylogenetic estimation*

[1] Frontiers in Massive Data Analysis, National Academies Press, 2013

# SEPP, TIPP, and UPP

- SEPP: SATe-enabled Phylogenetic Placement (Mirarab, Nugyen and Warnow, PSB 2012)
- TIPP: Taxon identification and phylogenetic profiling (Nguyen, Mirarab, Liu, Pop, and Warnow, submitted)
- UPP: Ultra-large multiple sequence alignment using SEPP (Nguyen, Mirarab, Kumar, Wang, Guo, Kim, and Warnow, in preparation)

# Phylogenetic Placement

Input: Tree T and alignment on full-length sequences, and set Q of query sequences

Output: edge in T for each sequence q in Q.

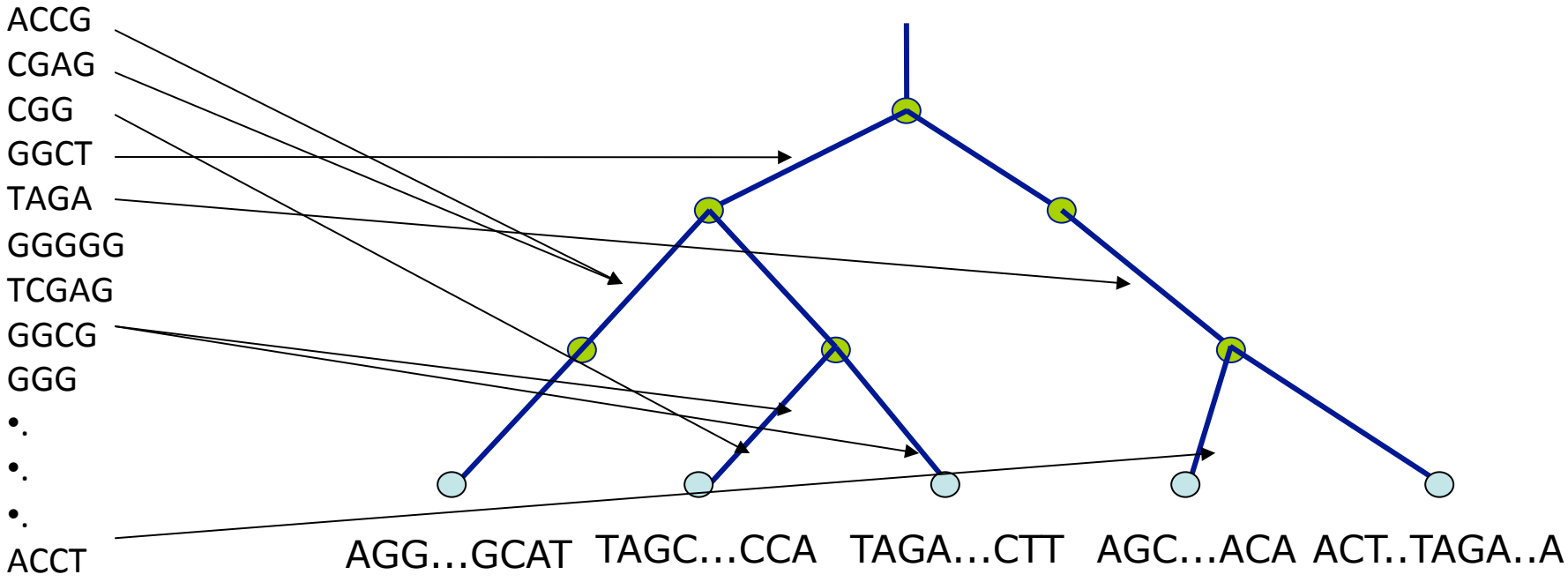
Applications:

- phylogenetic estimation on large datasets with many fragmentary sequences
- classifying shotgun metagenomic data (all from the same gene)

# Phylogenetic Placement

Fragmentary sequences from some gene

Full-length sequences for same gene, and an alignment and a tree



# Phylogenetic Placement

Step 1: Align each query sequence to backbone alignment

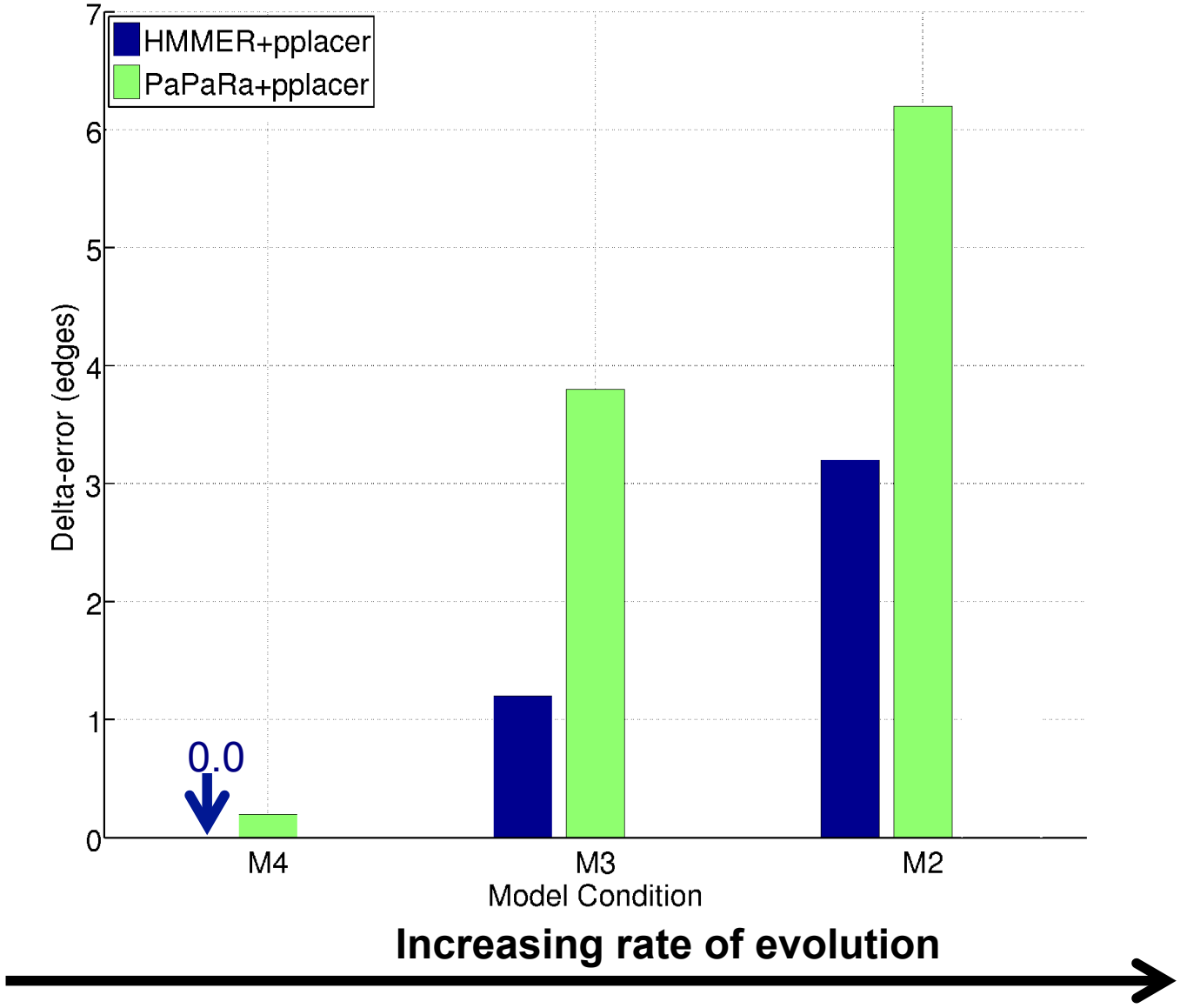
Step 2: Place each query sequence into backbone tree, using extended alignment

# 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
  - Pplacer (Matsen et al., BMC Bioinformatics, 2011)
  - EPA (Berger and Stamatakis, Systematic Biology 2011)

Note: pplacer and EPA use maximum likelihood, and are reported to have the same accuracy.

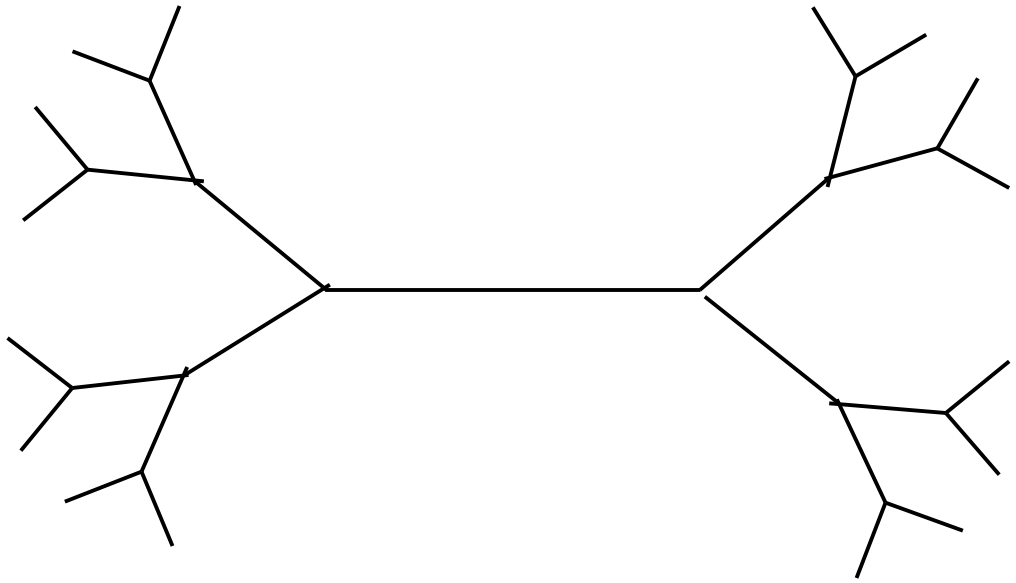
# HMMER vs. PaPaRa placement error



# HMMER+pplacer

Steps:

1) Build one HMM for the entire alignment
2) Align fragment to the HMM, and insert into alignment
3) Insert fragment into tree to optimize likelihood



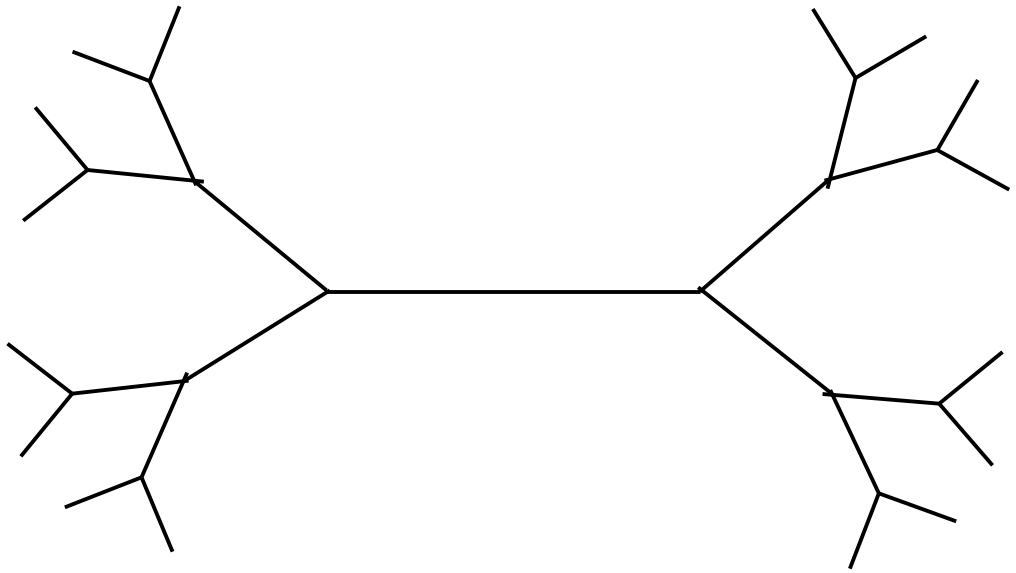
# Using HMMER

Using HMMER works well…

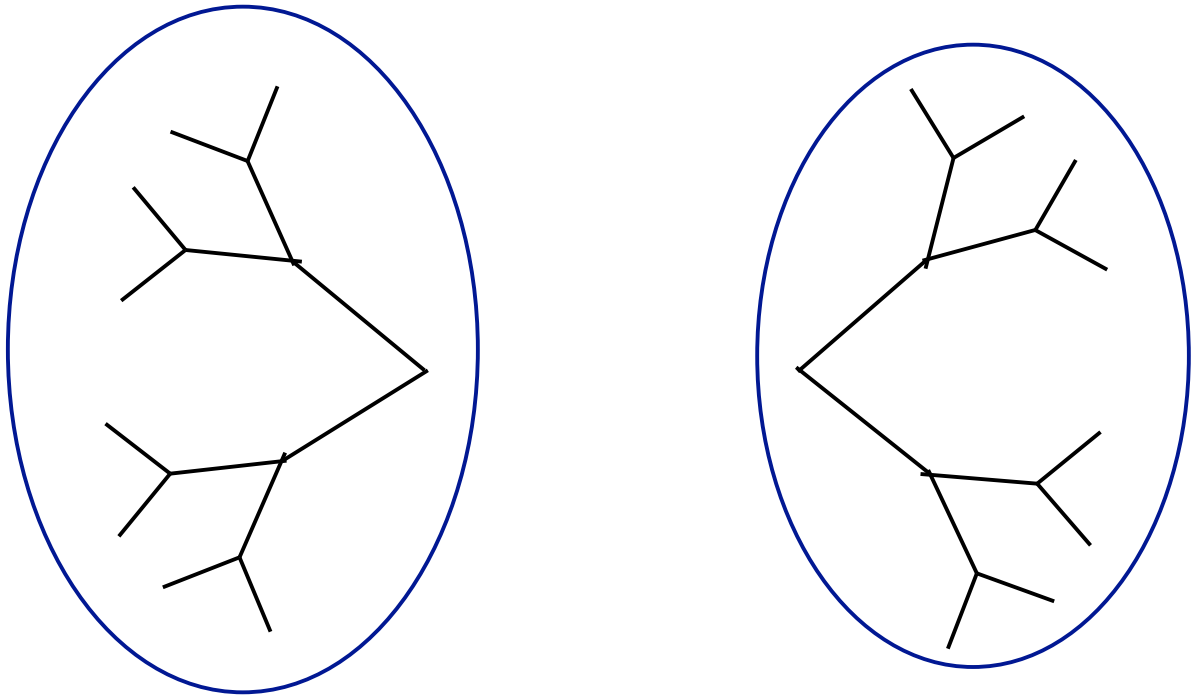
# Using HMMER

Using HMMER works well…except when the dataset is big!

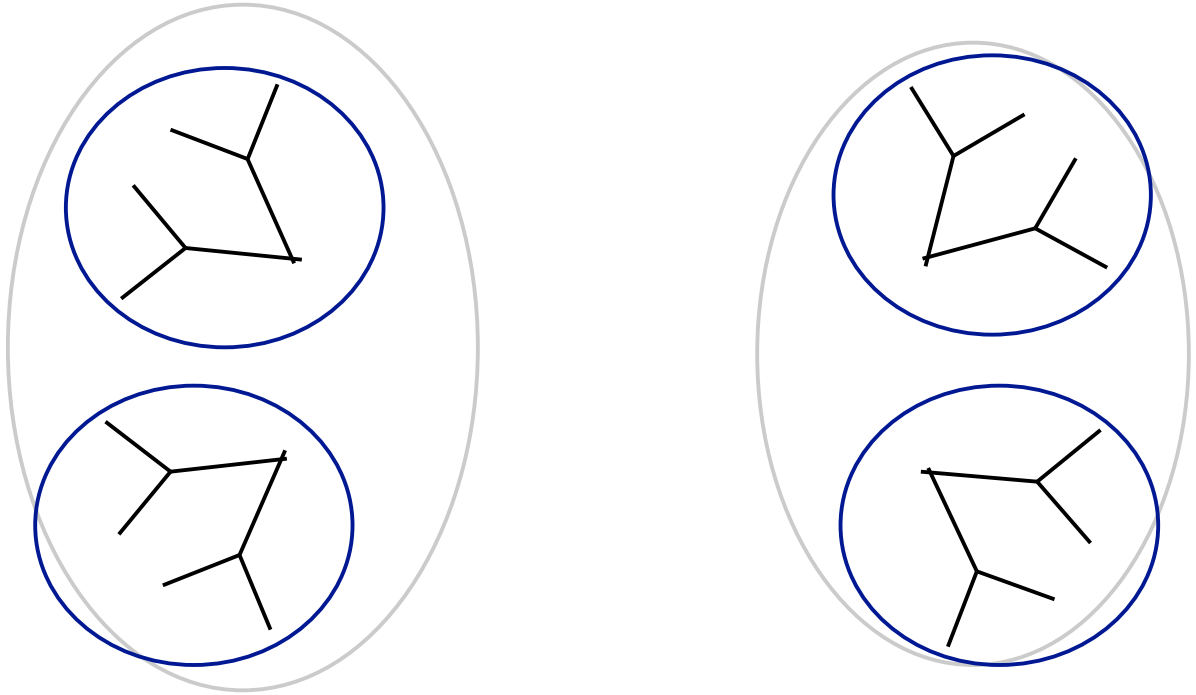
# One Hidden Markov Model for the entire alignment?



# Or 2 HMMs?



# Or 4 HMMs?

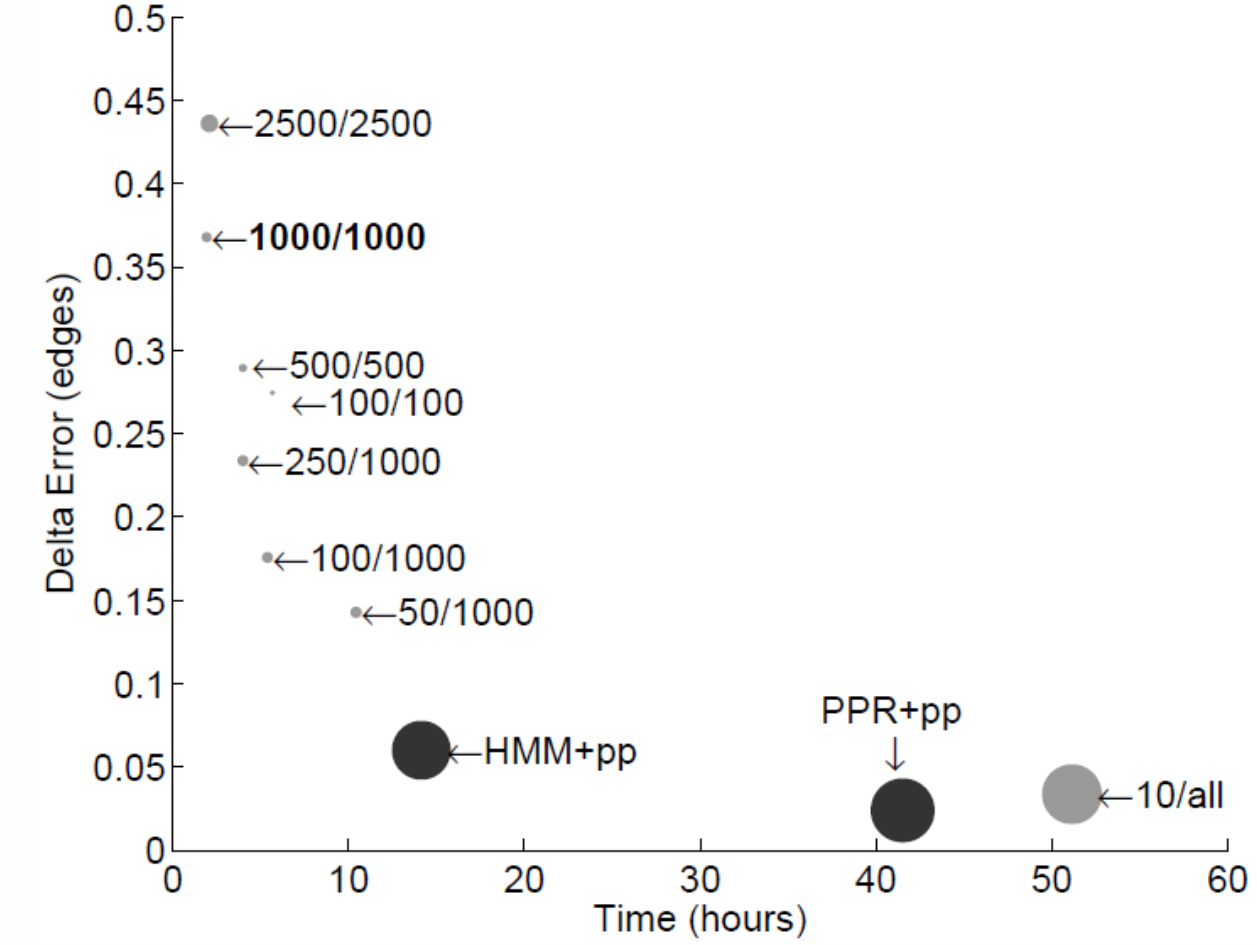


# SEPP

- **SEPP: SATé-enabled Phylogenetic Placement**, by Mirarab, Nguyen, and Warnow

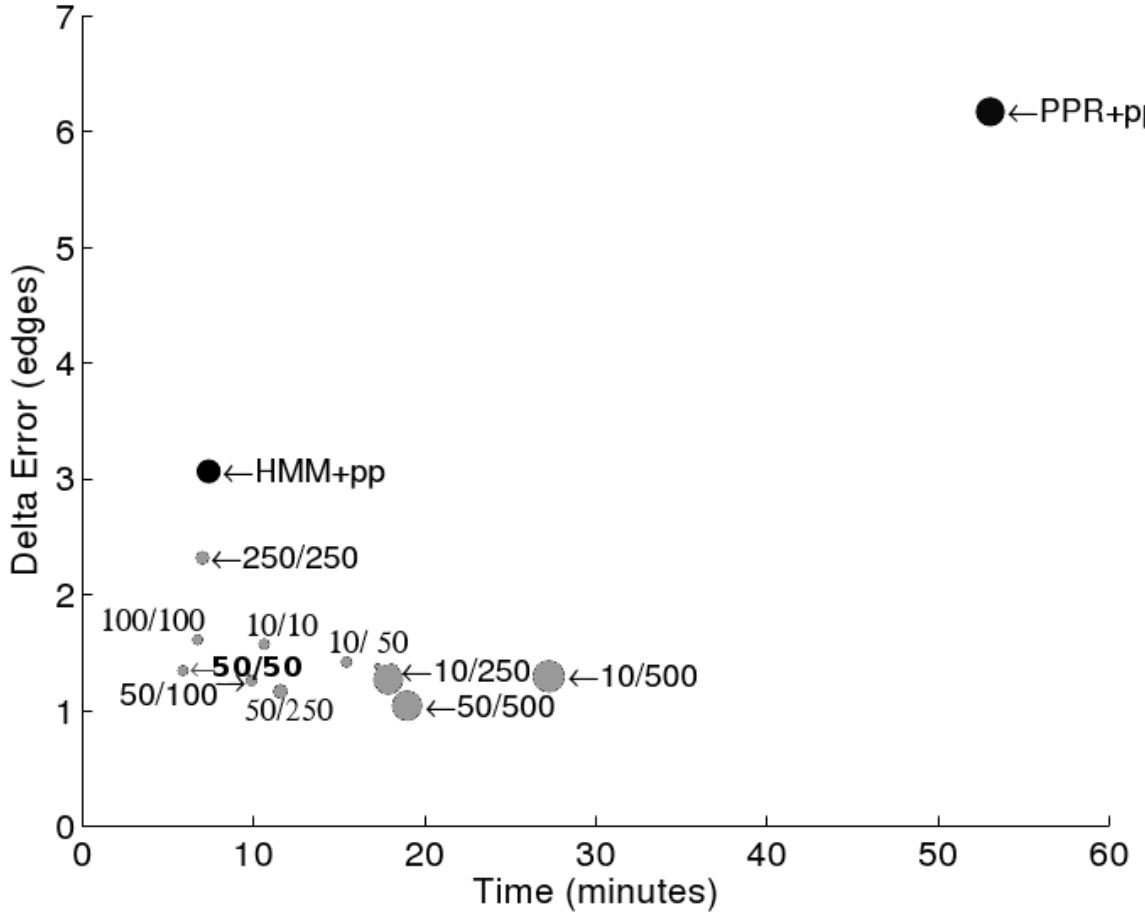
- Pacific Symposium on Biocomputing, 2012 (special session on the Human Microbiome)

# SEPP Parameters – low rate of evolution



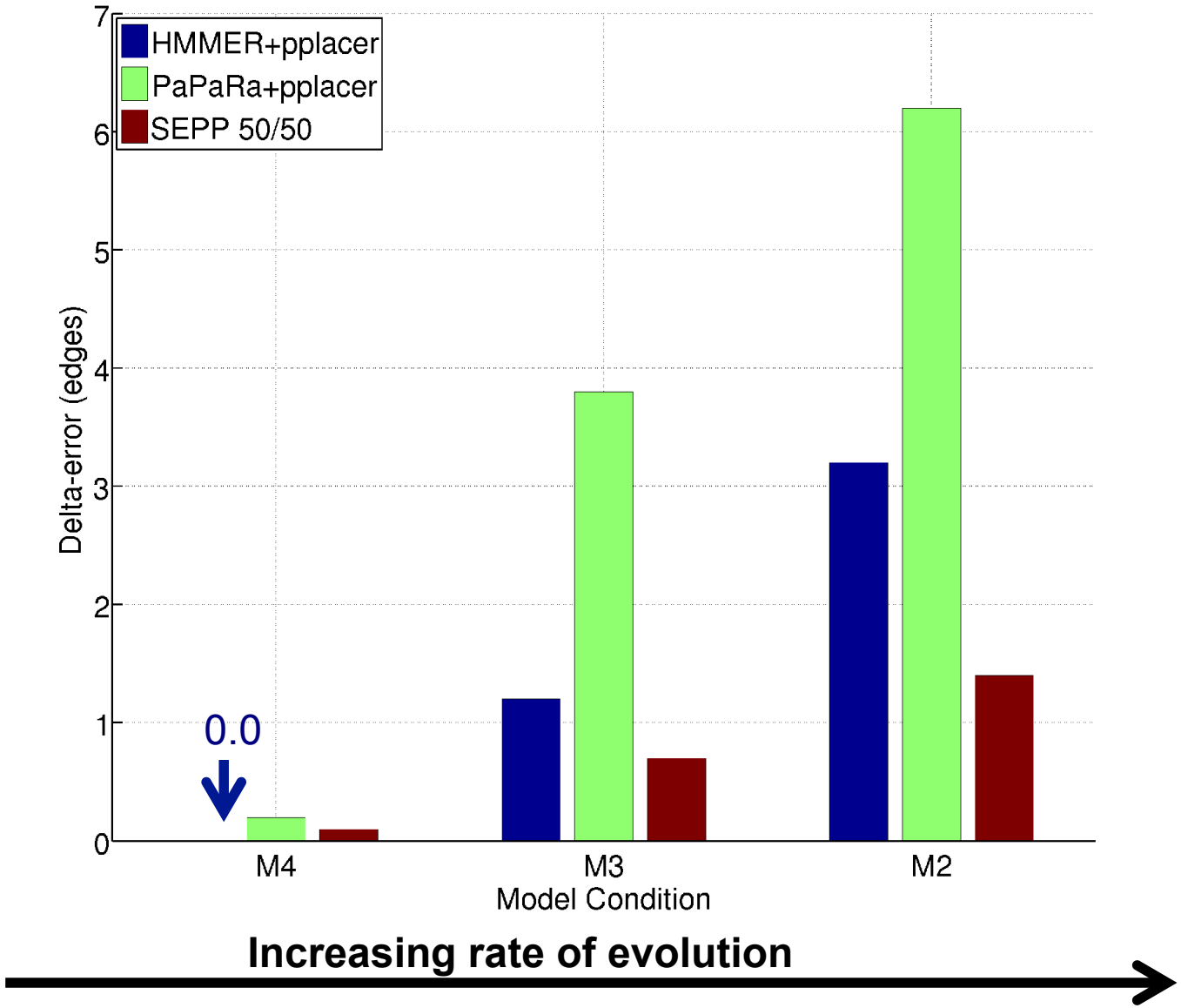

16S.B.ALL dataset, curated alignment/tree

# SEPP Parameters: high rate of evolution

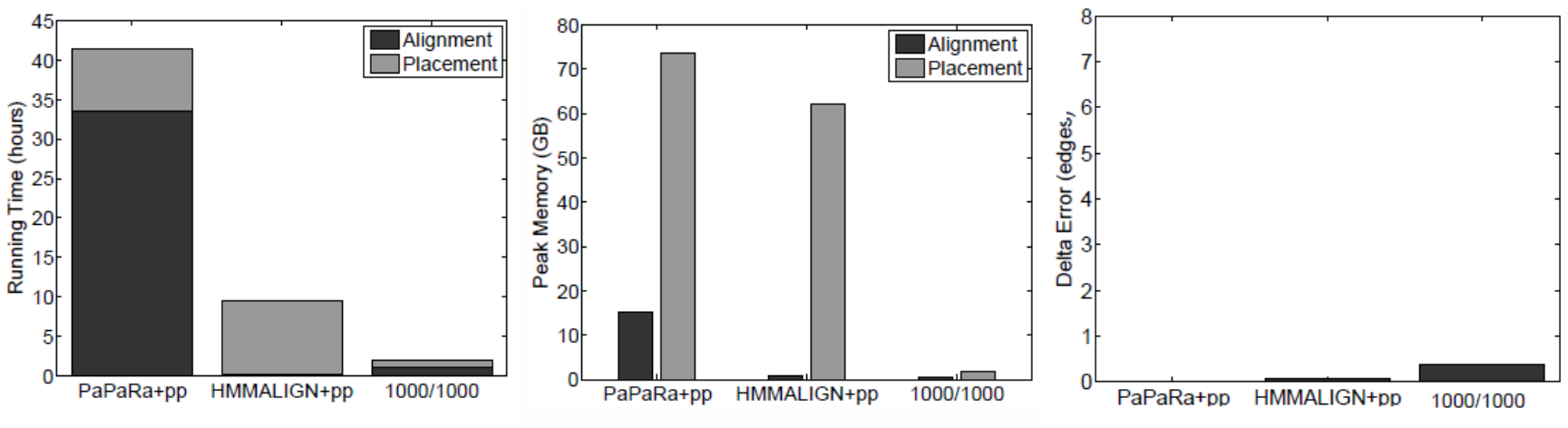



M2 model condition, true alignment/tree

# SEPP(10%), based on ~10 HMMs



# Biological Results




16S.B.ALL dataset, curated alignment/tree, 13k total fragments

For 1 million fragments:

PaPaRa+pplacer: ~133 days

HMMALIGN+pplacer: ~30 days

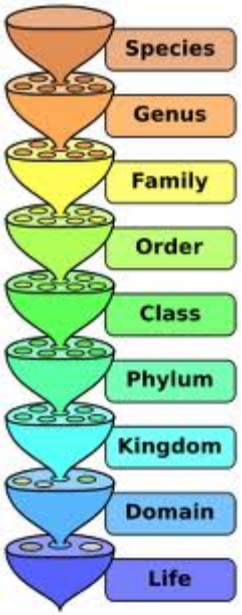
SEPP 1000/1000: ~6 days

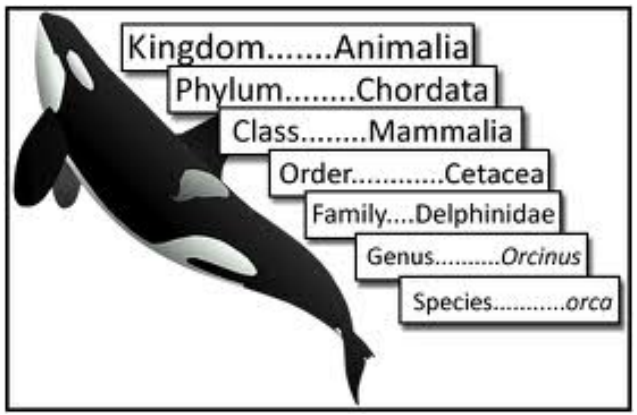
# TIPP

- Submitted for publication
- Developers Nam Nguyen and Siavash Mirarab (PhD students in Computer Science at UTCS)
- Other co-authors: Mihai Pop and Bo Liu (University of Maryland, College Park)

# Metagenomic Taxon Identification

Objective: classify short reads in a metagenomic sample



# Two Basic Questions

1. What is this fragment? (Classify each fragment as well as possible.)

2. What is the taxonomic distribution in the dataset? (Note: helpful to use marker genes.)

# Identifying Fragments using Marker Genes

- Approach:
  - Determine the gene for the fragment (if possible), thus producing a set of "bins" (one for each gene, and a bin for "unclassified")
  - For each gene, classify each fragment:
    - Construct a reference alignment and tree for full-length sequences for that gene
    - Place each fragment within the tree
    - Predict taxon identification (species, genus, etc.) from the placement

# TIPP: SEPP + statistics

Using SEPP as a taxon identification technique has high recall but low precision (classifies almost everything)

TIPP: dramatically reduces false positive rate with small reduction in true positive rate, by considering uncertainty in alignment (HMMER) and placement (pplacer)

# Taxonomic Identification

Objective: Identify species/genus/family (etc.) for each fragment within the sample.

Methods:

PhymmBL (Brady & Salzberg, Nature Methods 2009)

NBC (Rosen, Reichenberger, and Rosenfeld, Bioinformatics 2011)

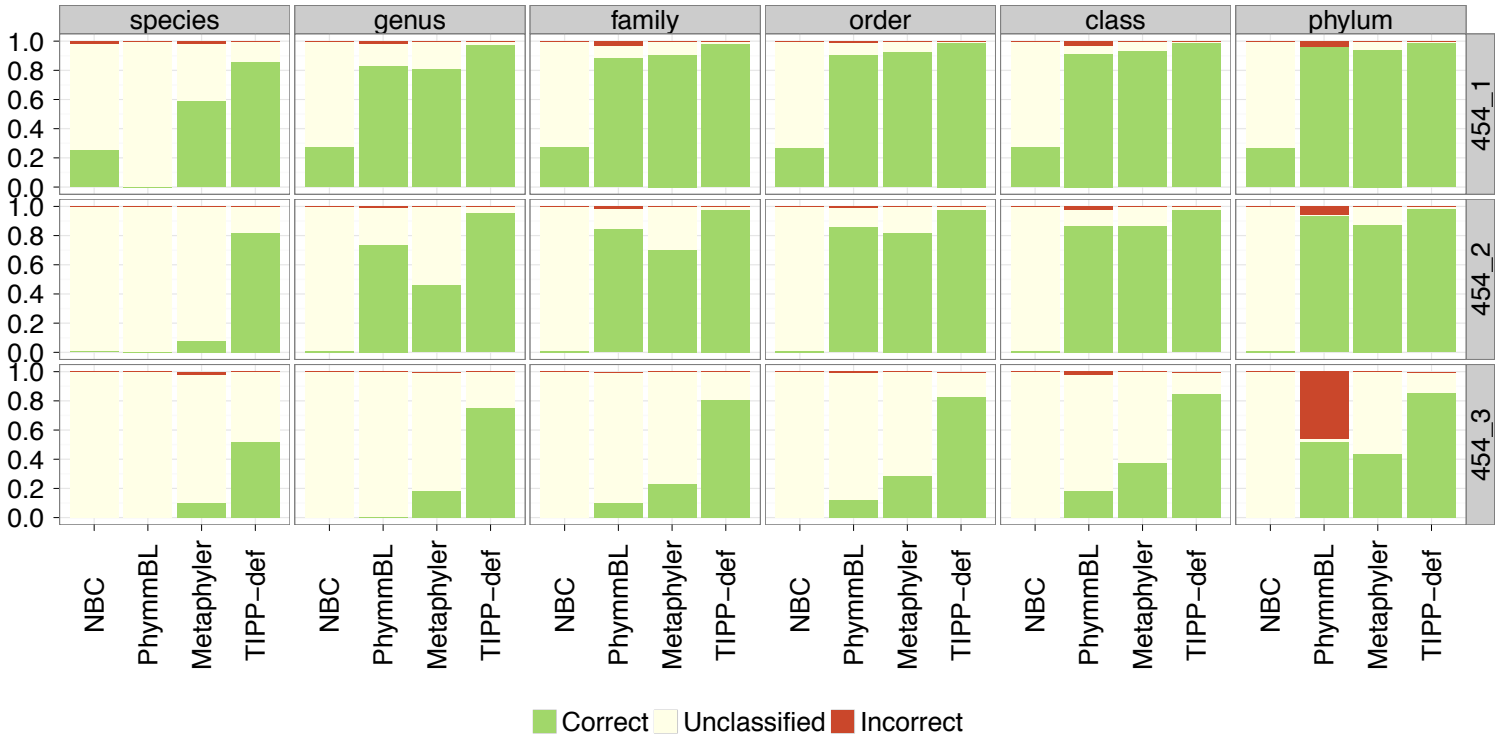
Metaphyler (Liu et al., BMC Genomics 2011), from the Pop lab at the University of Maryland

Metaphyler is a marker-based method.

Marker gene are single-copy, universal, and resistant to horizontal transmission.

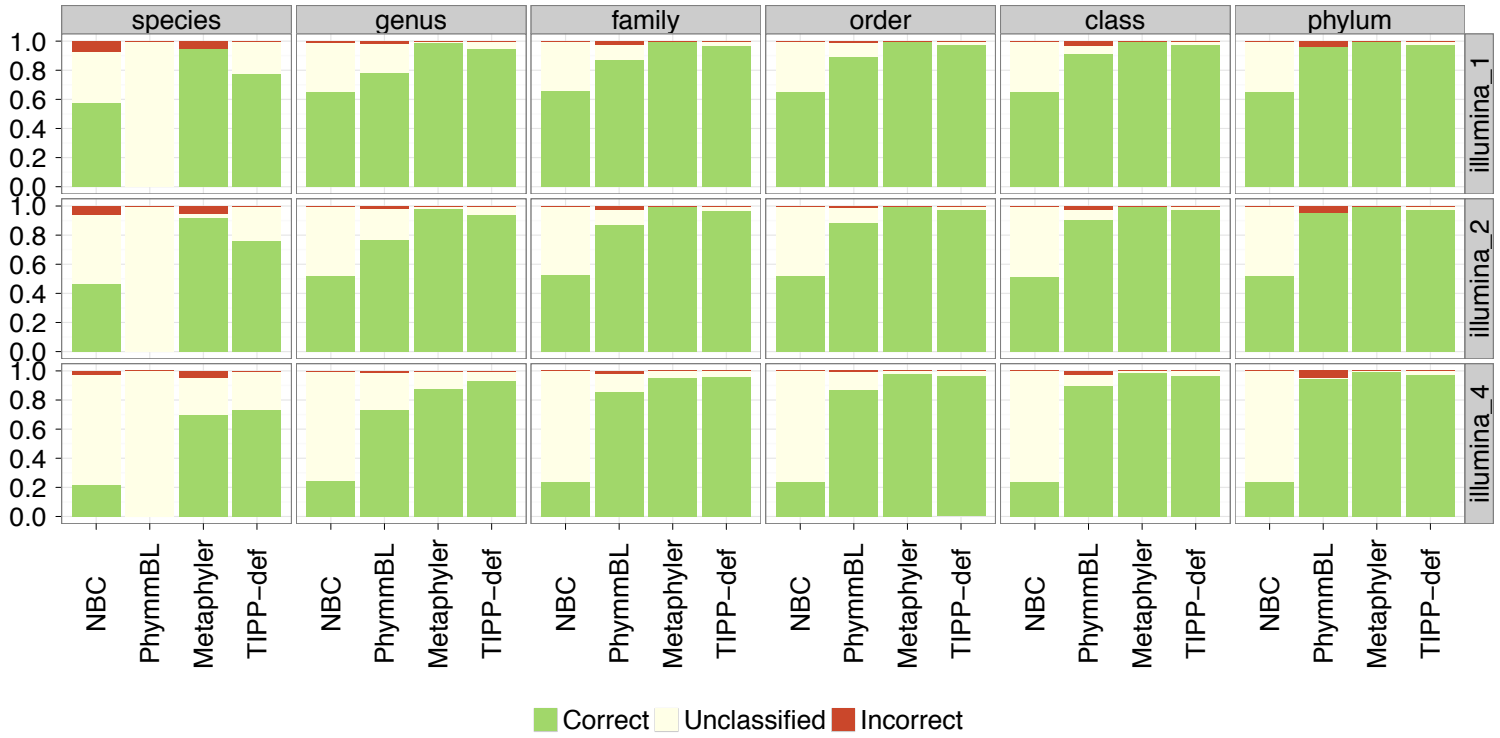
We test this for the 30 marker genes used in Metaphyler.

Criteria: Correct classification, incorrect classification, or no classification, at each level.




(a) 454 error model

Figure: Non-leave-one-out experiments comparing the classification accuracy for NBC, PhymmBL, MetaPhyler and TIPP-default (i.e., TIPP-default refers to TIPP(95%,95%,100)) for fragments simulated from the 30 marker genes under 454-like errors.




(a) Illumina error model

Figure: Non-leave-one-out experiments comparing the classification accuracy for NBC, PhymmBL, MetaPhyler and TIPP-default (i.e., TIPP-default refers to TIPP(95%,95%,100)) for fragments simulated from the 30 marker genes under Illumina-like errors.

# Abundance Profiling

Objective: Distribution of the species (or genera, or families, etc.) within the sample.

For example: The distribution of the sample at the species-level is:

| | |
|---|---|
| 50% | species A |
| 20% | species B |
| 15% | species C |
| 14% | species D |
| 1% | species E |

# Abundance Profiling

Objective: Distribution of the species (or genera, or families, etc.) within the sample.

Leading techniques:

PhymmBL (Brady & Salzberg, Nature Methods 2009)

NBC (Rosen, Reichenberger, and Rosenfeld, Bioinformatics 2011)

Metaphyler (Liu et al., BMC Genomics 2011), from the Pop lab at the University of Maryland

MetaphlAn (Segata et al., Nature Methods 2012), from the Huttenhower Lab at Harvard

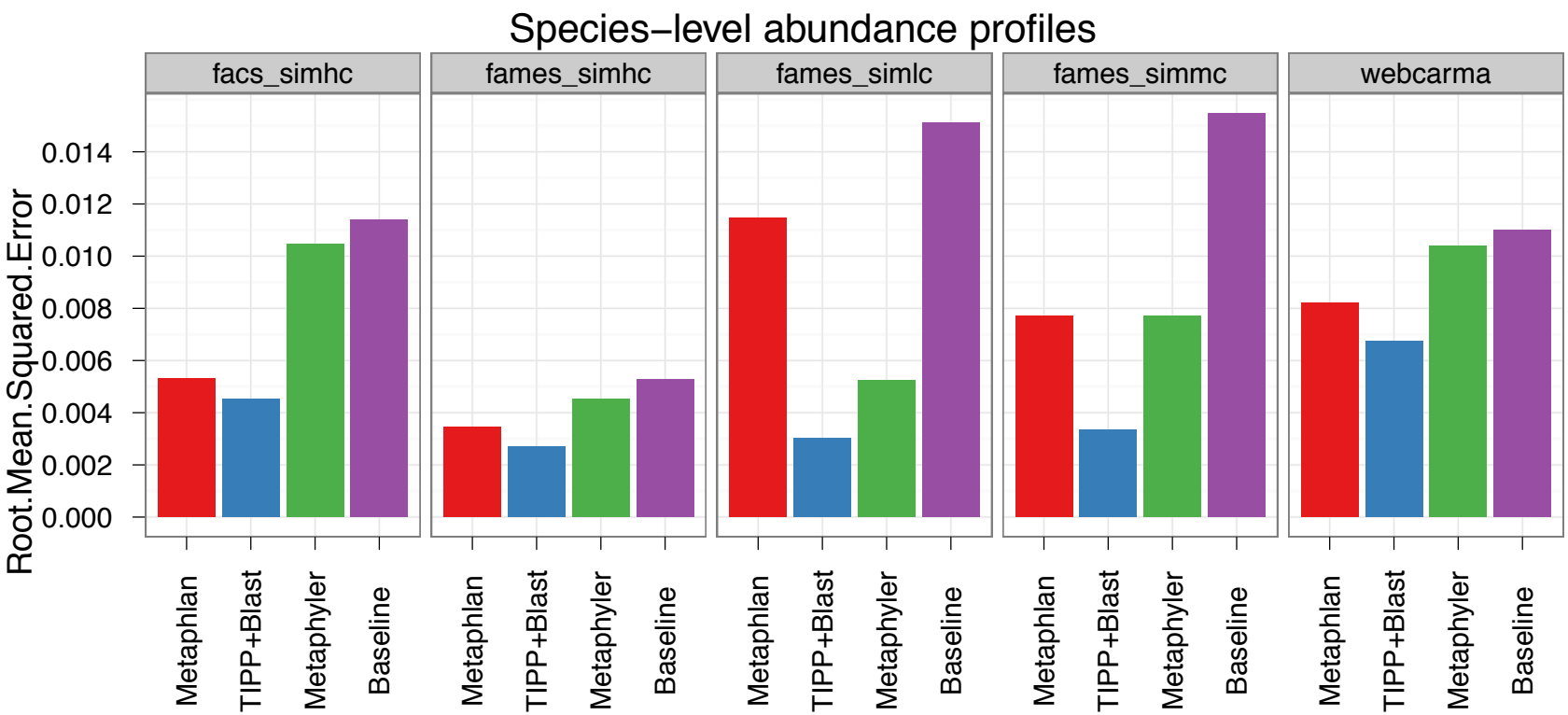
Metaphyler and MetaphlAn are marker-based techniques (but use different marker genes).

Marker gene are single-copy, universal, and resistant to horizontal transmission.

Table: Summary of all simulated abundance datasets. Complexity refers to the distribution of species in the profile: high complexity datasets have an even distribution of species, low complexity datasets have a staggered distribution of species, and medium complexity datasets fall in between.

| Dataset | Genomes | Complexity | Seq. Model | Reads | Avg. length |
|---|---|---|---|---|---|
| MetaPhlAn HC | 100 | High | NA | 1000000 | 88 |
| MetaPhlAn LC | 25 | Low | NA | 240000 | 88 |
| FAMeS HC | 113 | High | DOE-JGI | 116771 | 949 |
| FAMeS MC | 113 | Medium | DOE-JGI | 114457 | 969 |
| FAMeS LC | 113 | Low | DOE-JGI | 97495 | 951 |
| FACS HC | 19 | High | 454 | 26984 | 268 |
| FACS HC Illumina | 19 | High | Illumina | 300000 | 100 |
| WebCarma | 25 | High | 454 | 25000 | 265 |
| WebCarma Illumina | 25 | High | Illumina | 300000 | 100 |

Short fragment datasets: average length at most 100
Long fragment datasets: average length 265 to 969




- FACs HC: Fragments simulated from 19 bacterial genomes, all in equal abundance (Stranneheim et al. 2010)
- FAMeS: Fragments simulated from 113 bacterial and archael genomes, under 3 different abundance complexity profiles. (Mavromatis et al. 2007)
- WebCarma: Fragments simulated from 25 bacterial genomes, all in equal abundance (Gerlach and Stoye 2011).

Table: The average *RMSE* on the short and long fragment datasets. Note that PhymmBL does not output any species level classifications. We use TIPP(0%,0%,100) for abundance profiling (see SOM for results using other variants). The best results for each level and fragment length are in boldface.

| **Short Fragments** Dataset | Species | Genus | Family | Order | Class | Phylum |
|---|---|---|---|---|---|---|
| NBC | 0.022 | 0.026 | 0.028 | 0.029 | 0.030 | 0.038 |
| PhymmBL | NA | 0.026 | 0.028 | 0.029 | 0.028 | 0.035 |
| MetaPhlAn | **0.012** | **0.012** | 0.012 | 0.014 | 0.017 | 0.020 |
| MetaPhyler | 0.082 | 0.046 | 0.027 | 0.019 | 0.025 | 0.017 |
| TIPP | 0.013 | **0.012** | **0.011** | **0.012** | **0.016** | **0.014** |
| **Long Fragments** Dataset | Species | Genus | Family | Order | Class | Phylum |
| NBC | 0.016 | 0.019 | 0.023 | 0.025 | 0.031 | 0.033 |
| PhymmBL | NA | 0.018 | 0.020 | 0.021 | **0.023** | 0.024 |
| MetaPhlAn | 0.023 | 0.020 | 0.019 | 0.023 | 0.031 | 0.025 |
| MetaPhyler | 0.061 | 0.026 | 0.024 | 0.024 | 0.040 | 0.026 |
| TIPP | **0.013** | **0.014** | **0.018** | **0.020** | 0.032 | **0.017** |

# Observations

- Classification of fragments:
  - TIPP and Metaphyler are methods that use marker genes for taxon identification and phylogenetic profiling. These methods only classify fragments that are assigned to their marker genes. They will fail to classify some fragments.
  - TIPP and Metaphlyler are both more accurate than PhymmBL at classifying fragments from the 30 marker genes (perhaps not surprisingly).
  - Most methods are affected by sequencing errors, and especially by indels (454 errors). TIPP is fairly robust to 454 error (indels).
- Taxonomic profiling:
  - Marker-based profiling can produce more accurate taxonomic profiles (distributions) than techniques that attempt to classify all fragments.
  - Using marker genes from Metaphyler, TIPP produces more accurate taxonomic distributions (profiles) than Metaphyler..
- TIPP uses multiple sequence alignment and phylogenetic placement to improve accuracy. This is probably why TIPP has better robustness to indel errors, and high sensitivity.

# UPP: Ultra-large alignment using SEPP[1]

**Objective: highly accurate multiple sequence alignments and trees on ultra-large datasets**

Authors: Nam Nguyen, Siavash Mirarab, and Tandy Warnow

In preparation – expected submission March 2014

[1] SEPP: SATe-enabled phylogenetic placement, Nguyen, Mirarab, and Warnow, PSB 2012

# UPP: basic idea

Input: set S of unaligned sequences

Output: alignment on S

- Select random subset X of S
- Estimate “backbone” alignment A and tree T on X
- Independently align each sequence in S-X to A
- Use transitivity to produce multiple sequence alignment A* for entire set S

# Input: Unaligned Sequences

```
S1  = AGGCTATCACCTGACCTCCAAT
S2  = TAGCTATCACGACCGCGCT
S3  = TAGCTGACCGCGCT
S4  = TACTCACGACCGACAGCT
S5  = TAGGTACAACCTAGATC
S6  = AGATACGTCGACATATC
```

# Step 1: Pick random subset (backbone)

```
S1  = AGGCTATCACCTGACCTCCAAT
S2  = TAGCTATCACGACCGCGCT
S3  = TAGCTGACCGCGCT
S4  = TACTCACGACCGACAGCT
S5  = TAGGTACAACCTAGATC
S6  = AGATACGTCGACATATC
```

# Step 2: Compute backbone alignment

```
S1  = -AGGCTATCACCTGACCTCCA-AT
S2  = TAG-CTATCAC--GACCGC--GCT
S3  = TAG-CT-------GACCGC--GCT
S4  = TAC----TCAC--GACCGACAGCT
S5  = TAGGTAAAACCTAGATC
S6  = AGATAAAACTACATATC
```

# Step 3: Align each remaining sequence to backbone

First we add S5 to the backbone alignment

```
S1  = -AGGCTATCACCTGACCTCCA-AT-
S2  = TAG-CTATCAC--GACCGC--GCT-
S3  = TAG-CT-------GACCGC--GCT-
S4  = TAC----TCAC--GACCGACAGCT-
S5  = TAGG---T-A-CAA-CCTA--GATC
```

# Step 3: Align each remaining sequence to backbone

Then we add S6 to the backbone alignment

```
S1  = -AGGCTATCACCTGACCTCCA-AT-
S2  = TAG-CTATCAC--GACCGC--GCT-
S3  = TAG-CT-------GACCGC--GCT-
S4  = TAC----TCAC--GACCGACAGCT-
S6  = -AG---AT-A-CGTC--GACATATC
```

# Step 4: Use transitivity to obtain MSA on entire set

```
S1  = -AGGCTATCACCTGACCTCCA-AT--
S2  = TAG-CTATCAC--GACCGC--GCT--
S3  = TAG-CT-------GACCGC--GCT--
S4  = TAC----TCAC--GACCGACAGCT--
S5  = TAGG---T-A—CAA-CCTA--GATC-
S6  = -AG---AT-A-CGTC--GACATAT-C
```

# UPP: details

Input: set S of unaligned sequences

Output: alignment on S

- Select random subset X of S
- Estimate “backbone” alignment A and tree T on X
- Independently align each sequence in S-X to A
- Use transitivity to produce multiple sequence alignment A* for entire set S

# UPP: details

Input: set S of unaligned sequences

Output: alignment on S

- Select random subset X of S
- Estimate “backbone” alignment A and tree T on X
- Independently align each sequence in S-X to A
- Use transitivity to produce multiple sequence alignment A* for entire set S

# UPP (details) – first version

This is essentially SEPP, but with varying size subsets (depending on the dataset's empirical statistics):

Datasets with at most 1000 sequences, we just run PASTA (improved version of SATe). For larger datasets:

- Backbone alignment and tree: Random subset of 1000 sequences, backbone alignment and tree computed using PASTA.
- Subset size: depends on backbone alignment average p-distance: small p-distances get small subsets of size 10, large p-distances get large subsets of size k/2, where k = |backbone|.
- We decompose the backbone set using SEPP (i.e., the SATe centroid edge decomposition). This creates a collection of disjoint subsets. Each subset gets a HMM (using HMMBUILD). Each sequence is scored against every HMM, and selects the best HMM. The alignment using that HMM is used to add the sequence into the MSA on the full dataset.

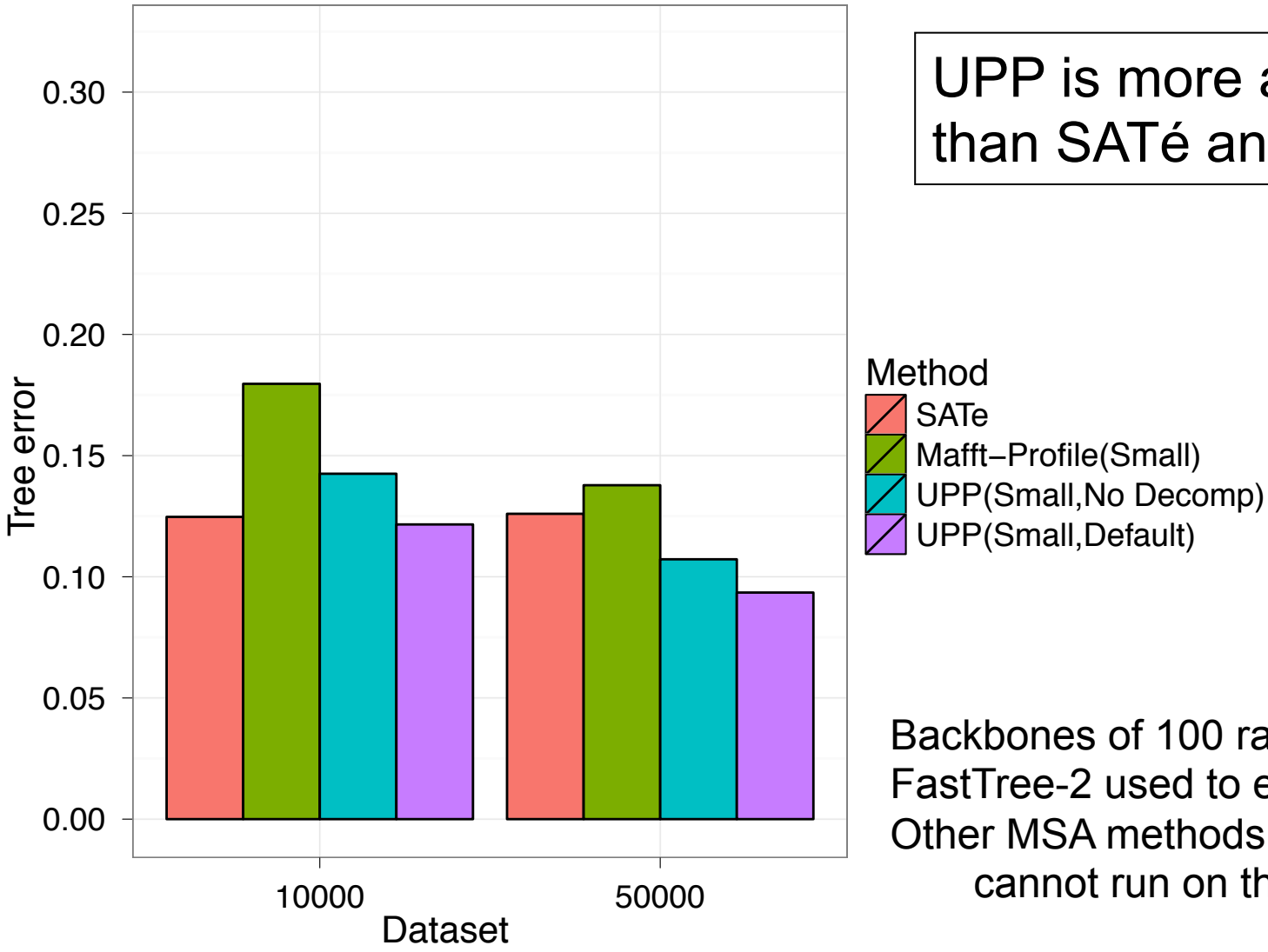
# UPP(x,y)

- Pick random subset X of size **x**
- Compute alignment A and tree T on X
- Use SATé decomposition on T to partition X into small "alignment subsets" of at most **y** sequences
- Build HMM on each alignment subset using HMMBUILD
- For each sequence s in S-X,
  - use HMMALIGN to produce alignment of s to each subset alignment and note the score of each alignment.
  - Pick the subset alignment that has the best score, and align s to that subset alignment.
  - Use transitivity to align s to the backbone alignment.

# Evaluation

- Simulated datasets (some have fragmentary sequences):
  - 10K to 1,000,000 sequences in RNASim (Sheng Guo, Li-San Wang, and Junhyong Kim, arxiv)
  - 1000-sequence nucleotide datasets from SATe papers
  - 5000-sequence AA datasets (from FastTree paper)
  - 10,000-sequence Indelible nucleotide simulation
- Biological datasets:
  - Proteins: largest BaliBASE and HomFam
  - RNA: 3 CRW (Gutell) datasets up to 28,000 sequences

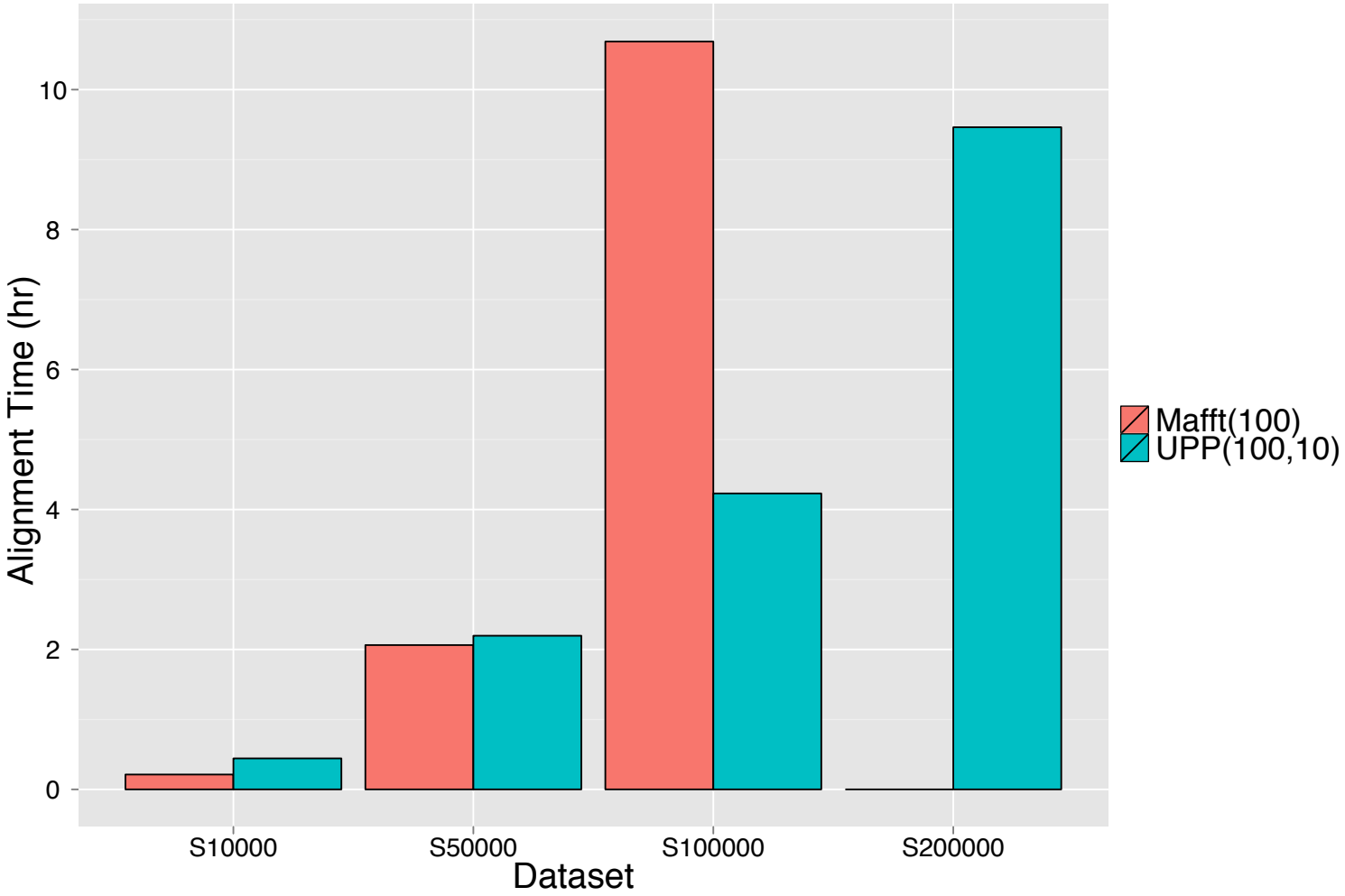
# Tree Error on 10K and 50K RNASim datasets



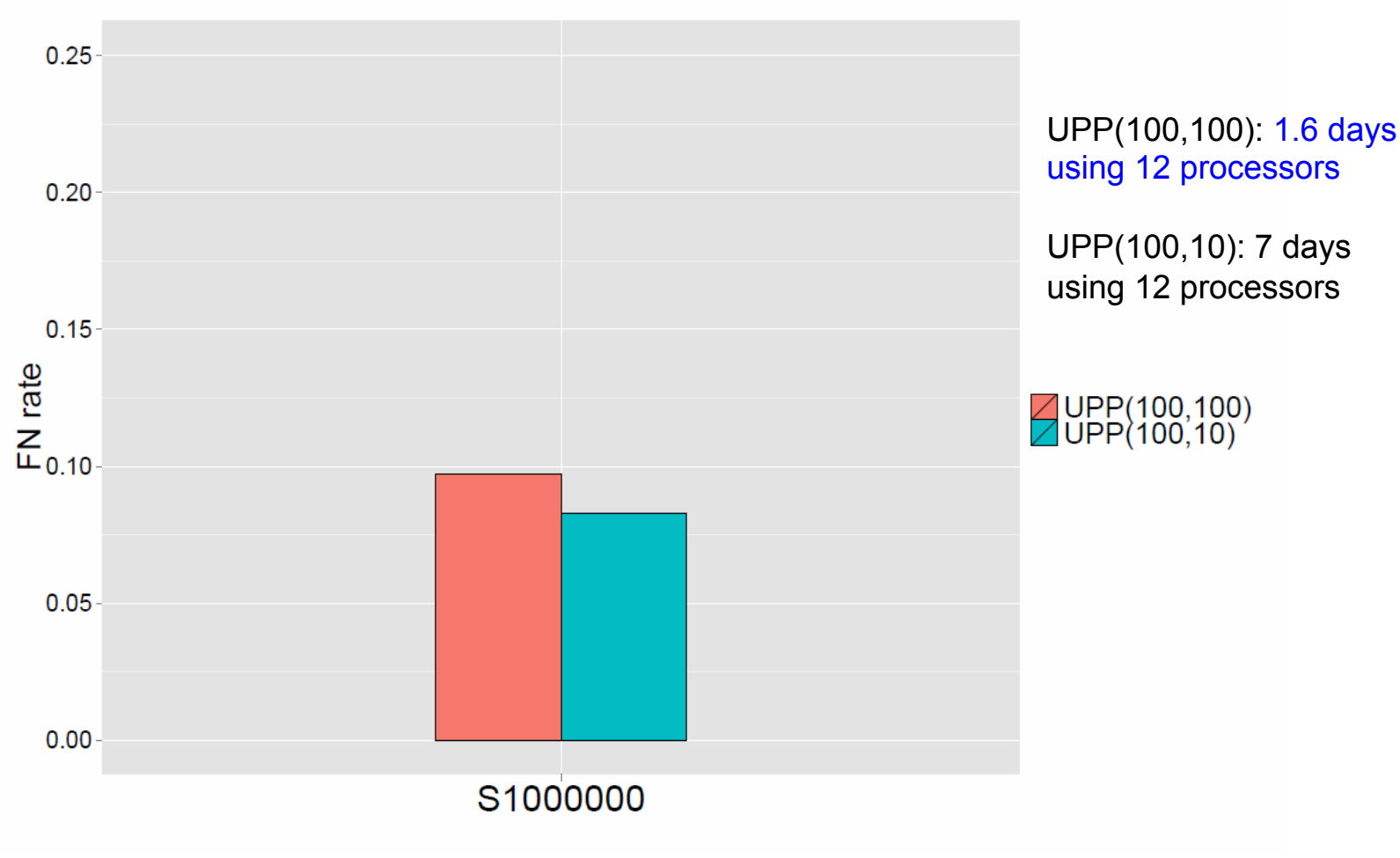

UPP is more accurate than SATé and MAFFT

Backbones of 100 random sequences
FastTree-2 used to estimate ML trees
Other MSA methods less accurate or cannot run on these data

# UPP vs. MAFFT-profile Running Time



# One Million Sequences: Tree Error




Note: UPP Decomposition improves accuracy

# UPP (details) – first version

This is essentially SEPP, but *with varying size subsets (depending on the dataset's empirical statistics):*

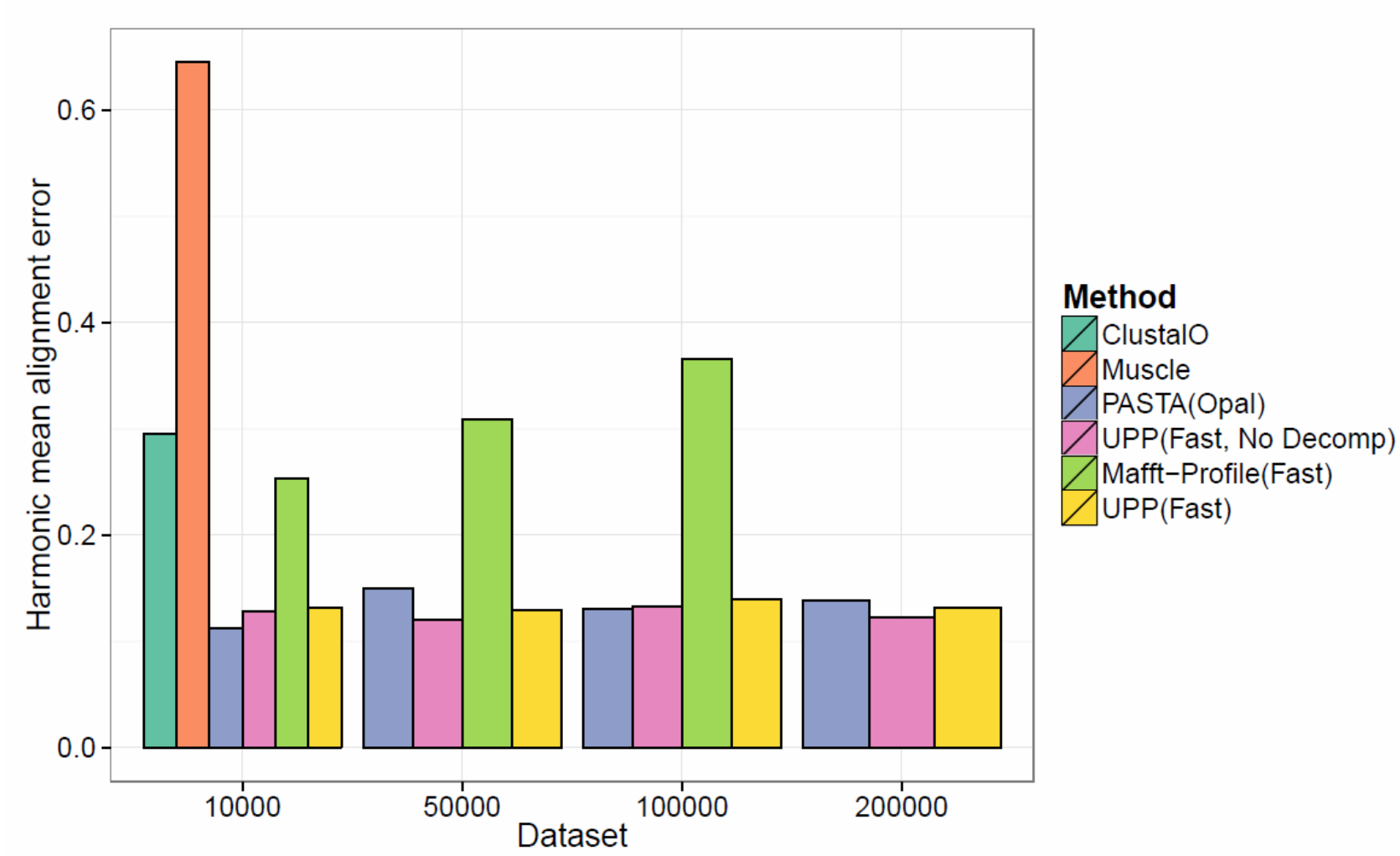
- Backbone alignment and tree: Random subset of 1000 sequences on datasets with more than 1000 sequences, 200 for smaller datasets. SATe alignment and tree on the backbone subset.
- Subset size: depends on backbone alignment average p-distance: small p-distances get small subsets of size 10, large p-distances get large subsets of size k/2, where k = |backbone|.
- We decompose the backbone set using SEPP (i.e., the SATe centroid edge decomposition). This creates a collection of disjoint subsets. Each subset gets a HMM (using HMMBUILD). Each sequence is scored against every HMM, and selects the best HMM. The alignment using that HMM is used to add the sequence into the MSA on the full dataset.

# UPP (details) – second version

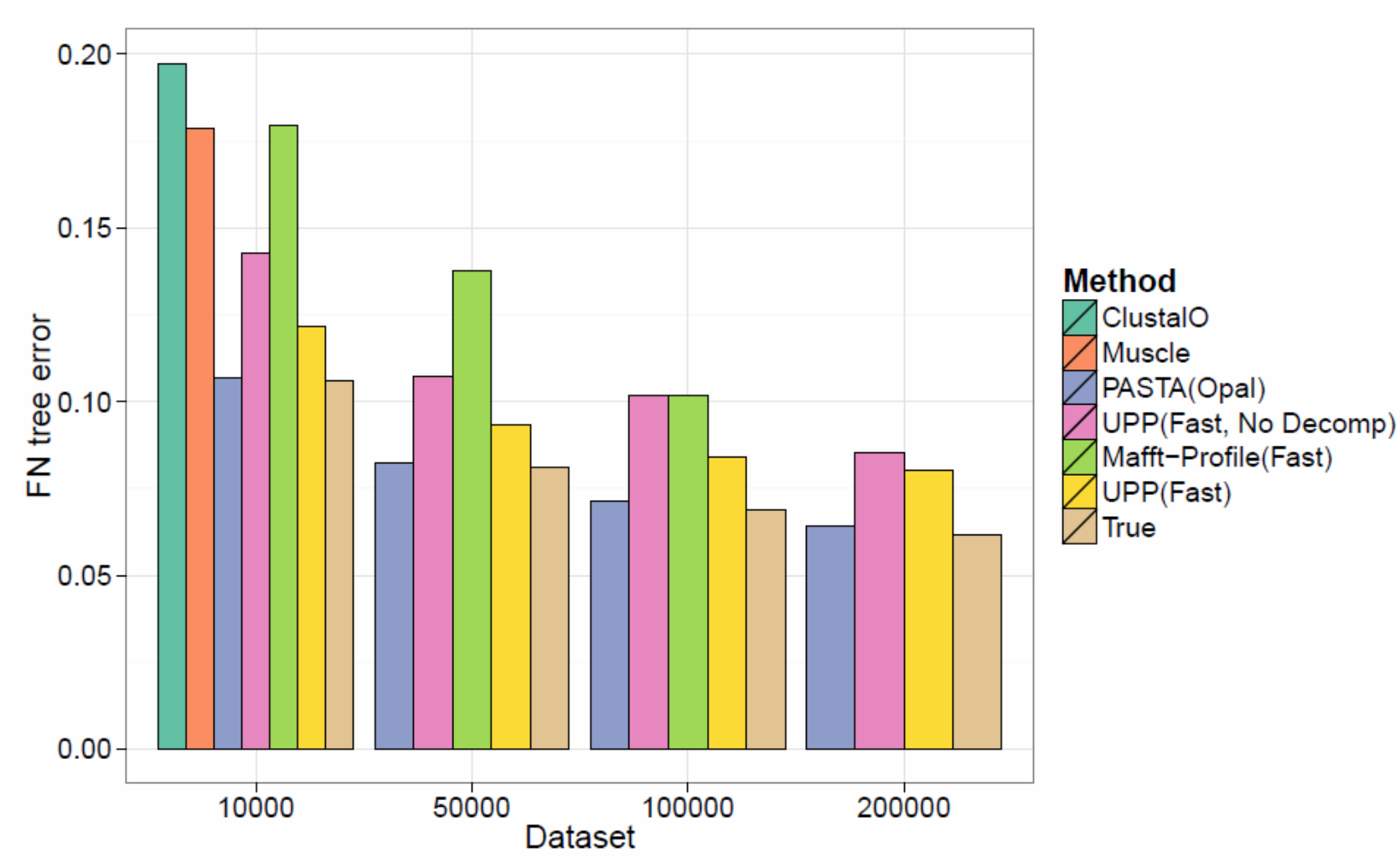
- Backbone alignment and tree: Same as for first version. (Random subset of 1000 sequences on datasets with more than 1000 sequences. PASTA alignment and tree on the backbone subset.)
- **Subset size: We decompose the backbone into many subsets of varying sizes (10, 20, 40, 80, …, 640, 1000).**
- All subsets get HMMs, and the sequence is scored against every HMM – the best scoring HMM is used to align the sequence.

# RNASim: Alignment error

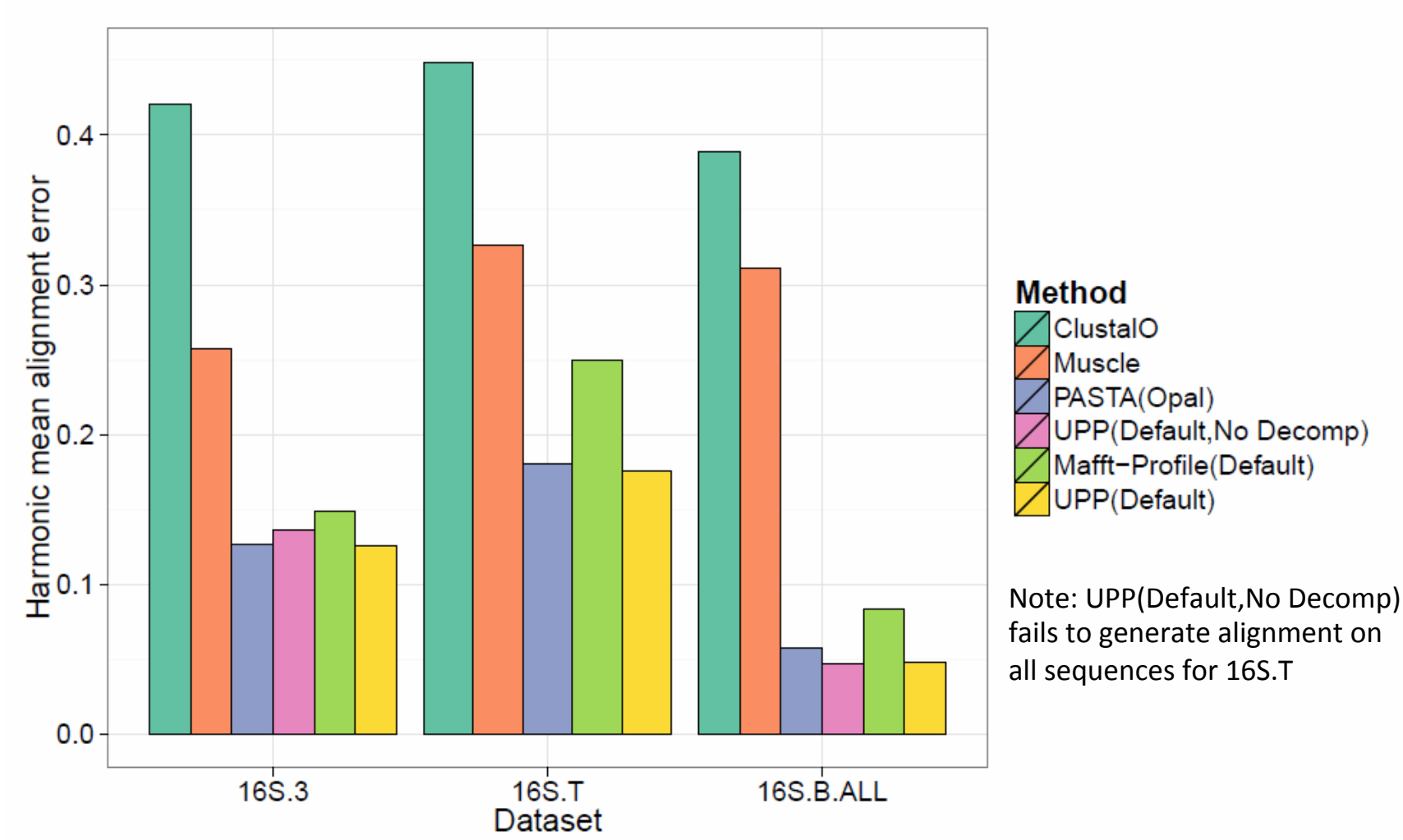
UPP(Fast) uses backbones of size 100



# RNASim: Tree error

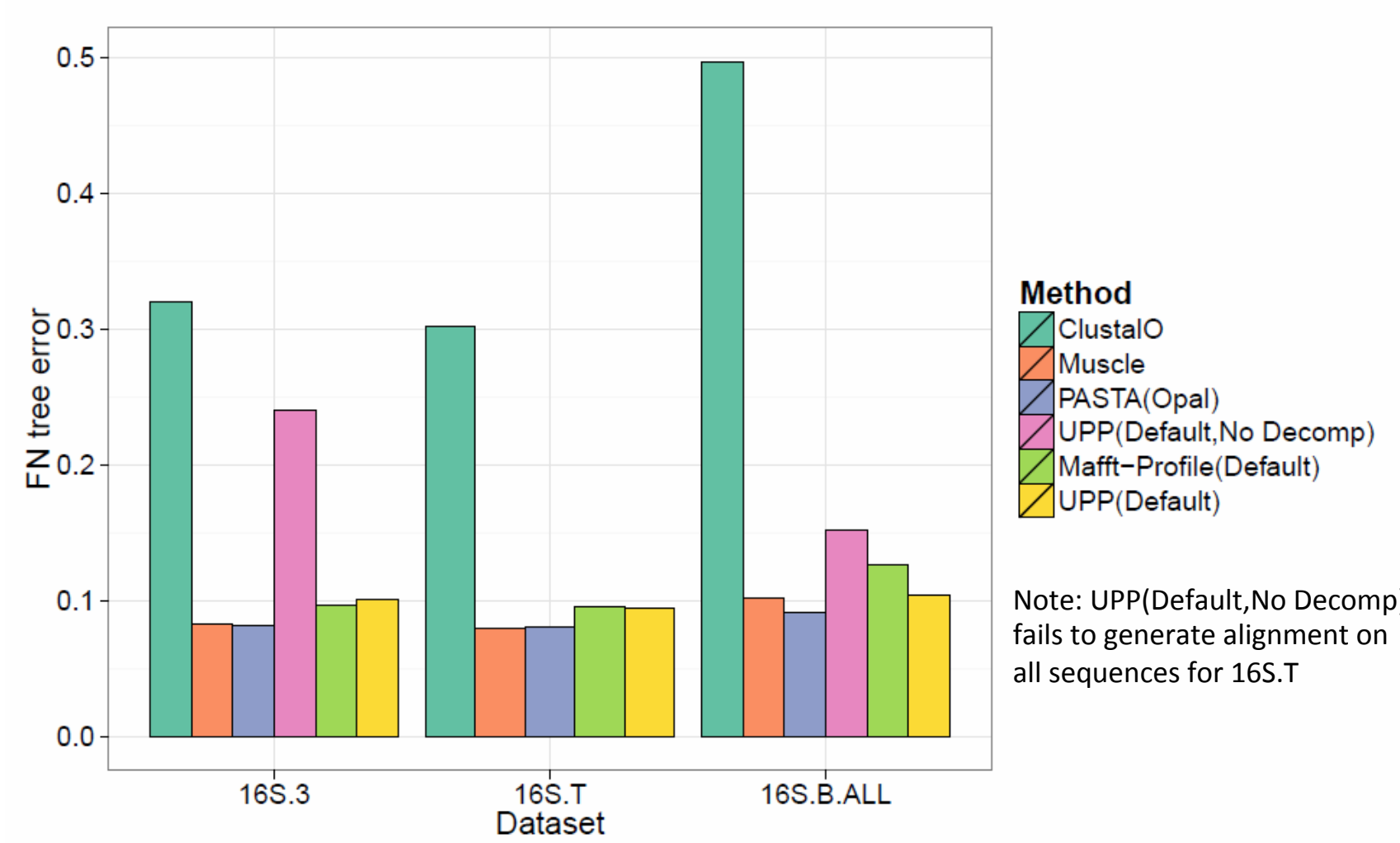


# Gutell 16S: Alignment error



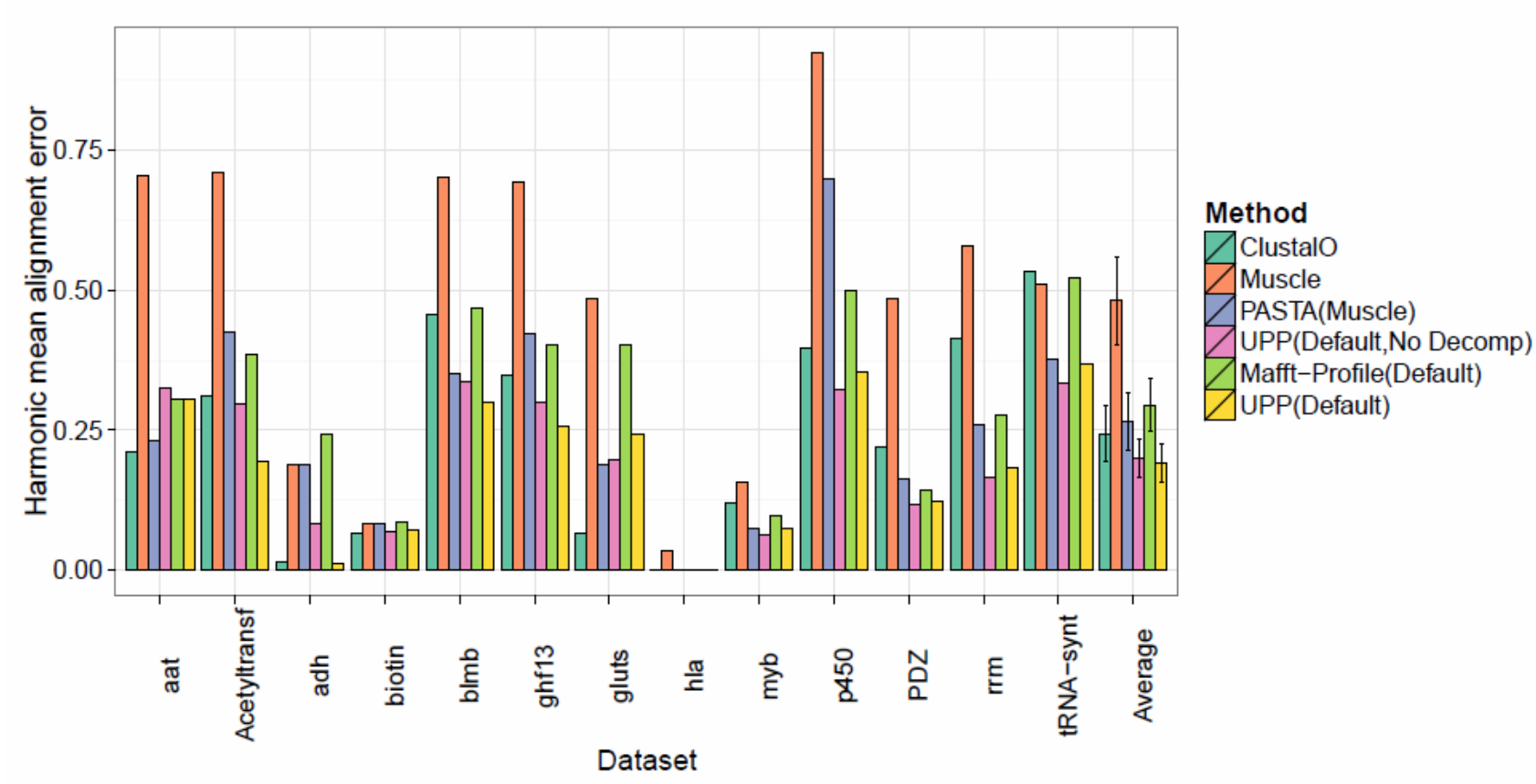

Note: UPP(Default,No Decomp) fails to generate alignment on all sequences for 16S.T

# Gutell 16S: Tree error




Note: UPP(Default,No Decomp) fails to generate alignment on all sequences for 16S.T

# HomFam: Alignment error




Dataset sizes range from 10,000 to 46,000 sequences.

# 1kp: Thousand Transcriptome Project

G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

N. Wickett
Northwestern

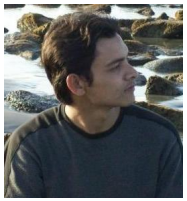
N. Matasci
iPlant

T. Warnow,
UT-Austin

S. Mirarab,
UT-Austin

N. Nguyen,
UT-Austin

Md. S.Bayzid
UT-Austin

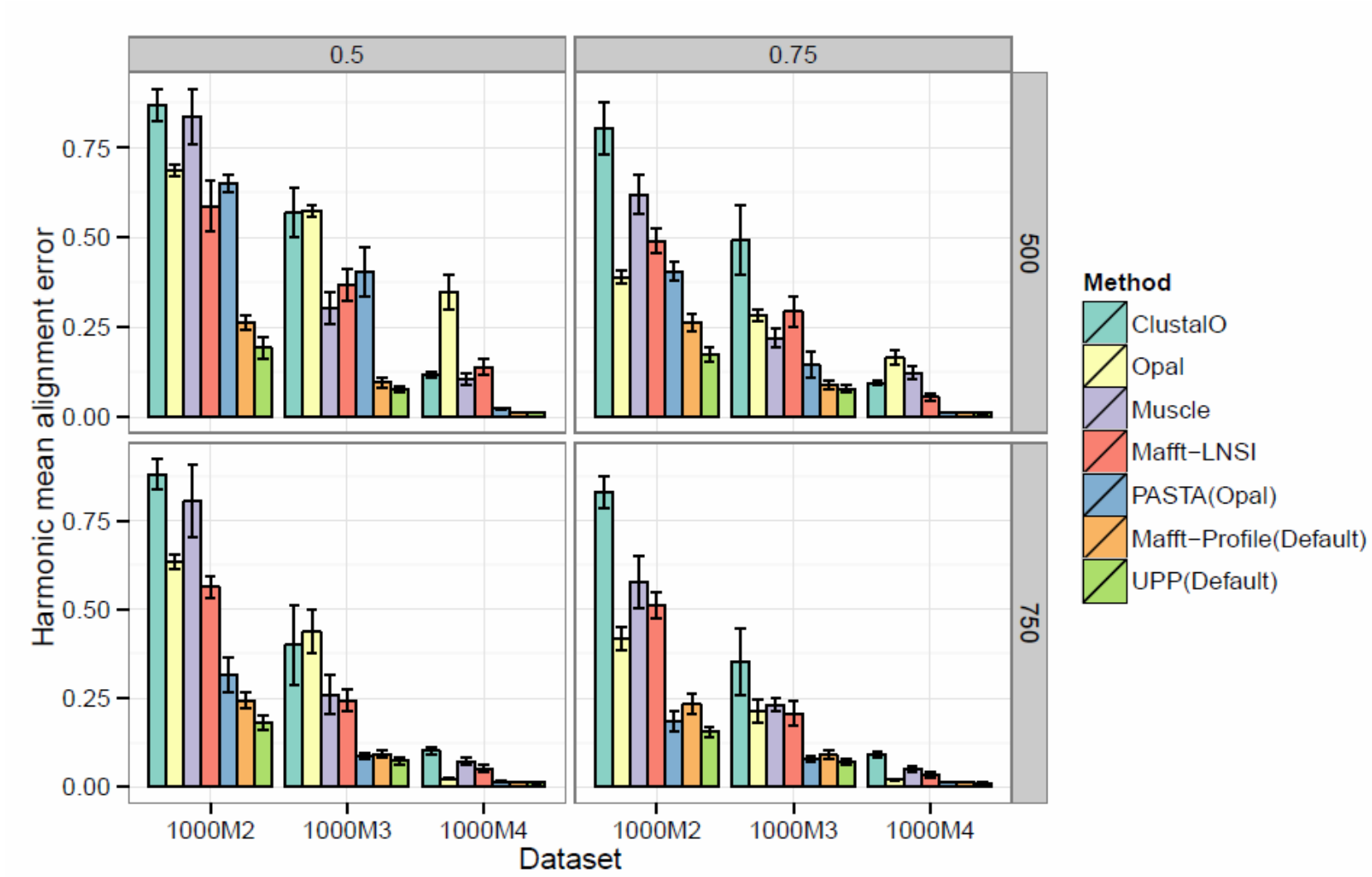


Plus many many other people…

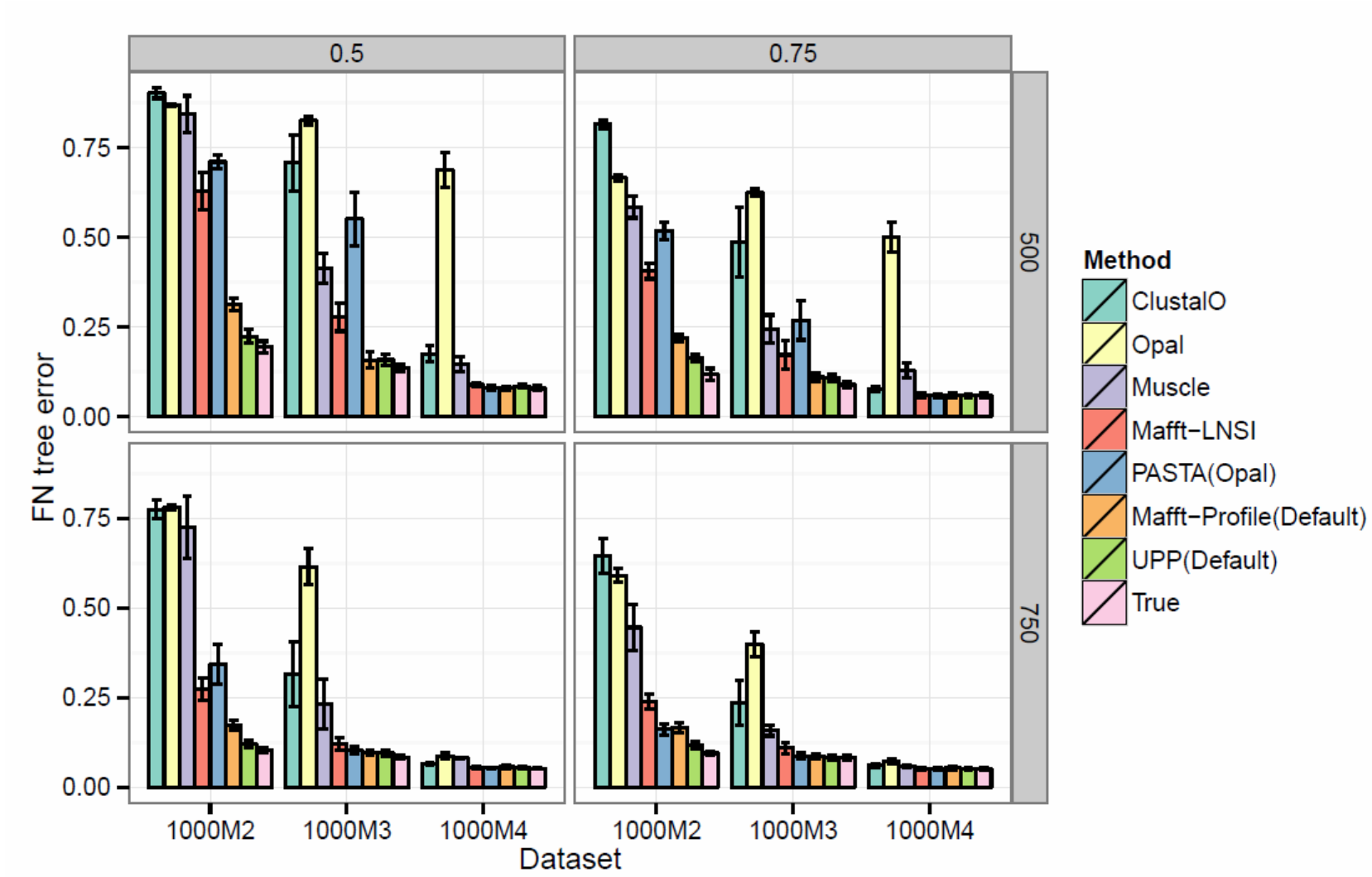
- Plant Tree of Life based on transcriptomes of ~1200 species
- More than 13,000 gene families (most not single copy)

**Challenge:**
**Alignment of datasets with > 100,000 sequences with many fragmentary sequences**

# 1K fragmentary: Alignment error



# 1K fragmentary: Tree error

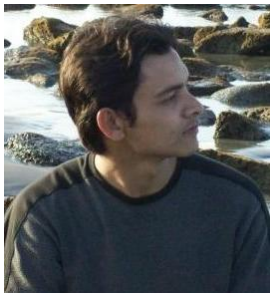


# Future Work

- Extending TIPP to non-marker genes
- Using the new HMM Family technique in SEPP and UPP
- Using external seed alignments in SEPP, TIPP, and UPP
- Boosting statistical co-estimation methods by using them in UPP for the backbone alignment and tree

# Warnow Laboratory





PhD students: Siavash Mirarab*, Nam Nguyen, and Md. S. Bayzid**
Undergrad: Keerthana Kumar
Lab Website: http://www.cs.utexas.edu/users/phylo

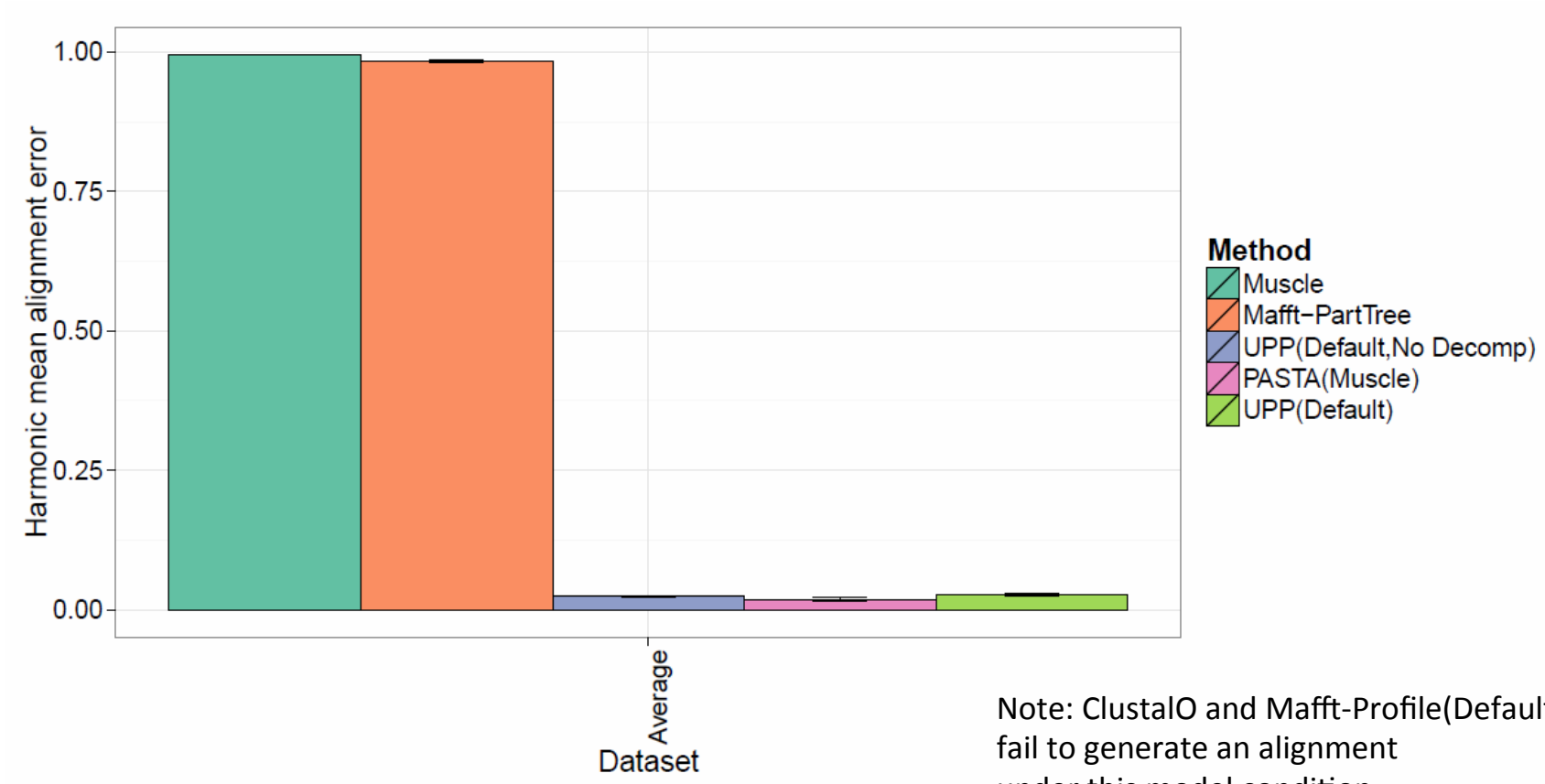
**Funding**: Guggenheim Foundation, Packard, NSF, Microsoft Research New England, David Bruton Jr. Centennial Professorship, TACC (Texas Advanced Computing Center), and the University of Alberta (Canada)

**TACC** and UTCS computational resources

* Supported by HHMI Predoctoral Fellowship

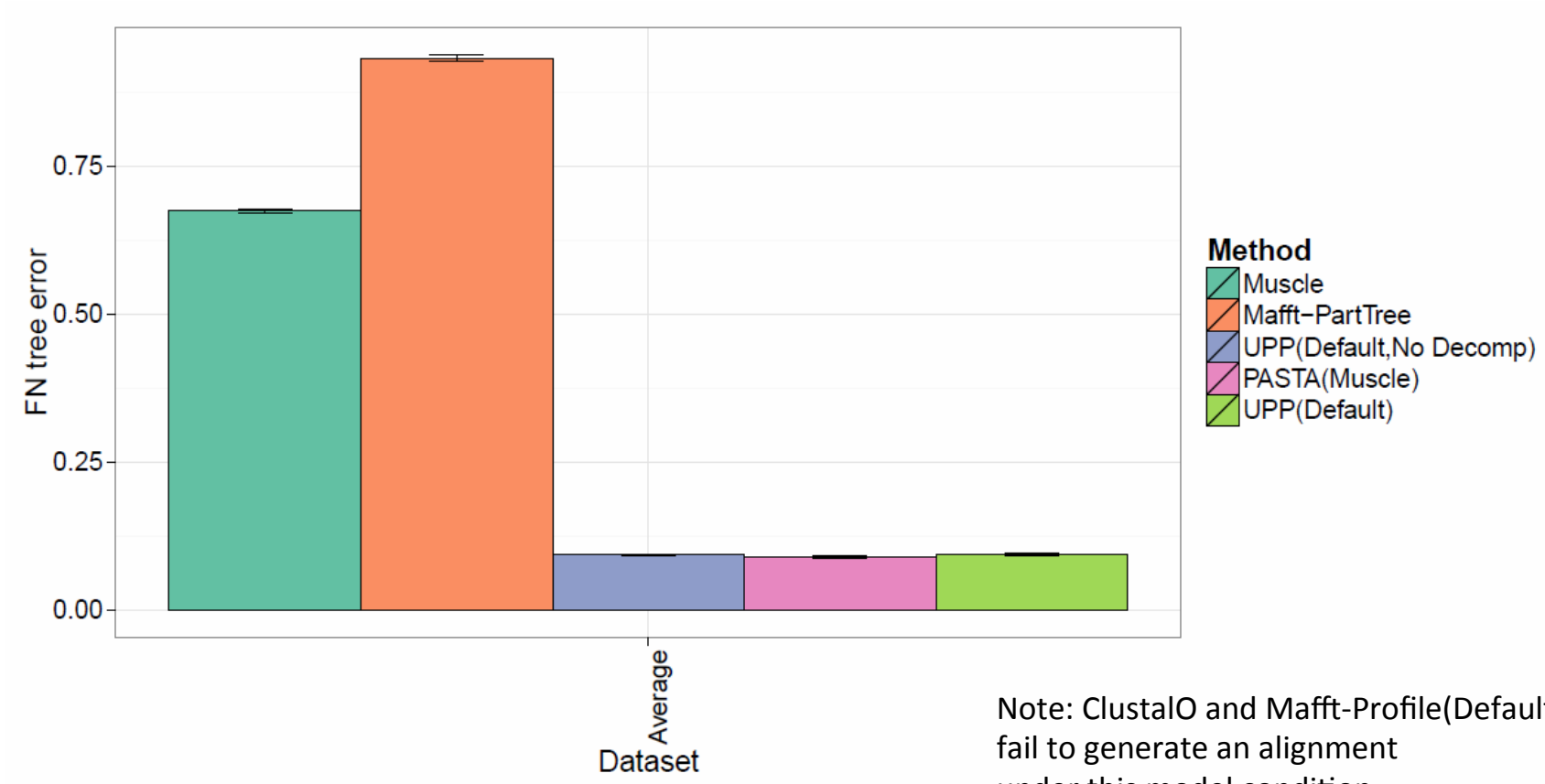
** Supported by Fulbright Foundation Predoctoral Fellowship

# Indelible 10K: Alignment error



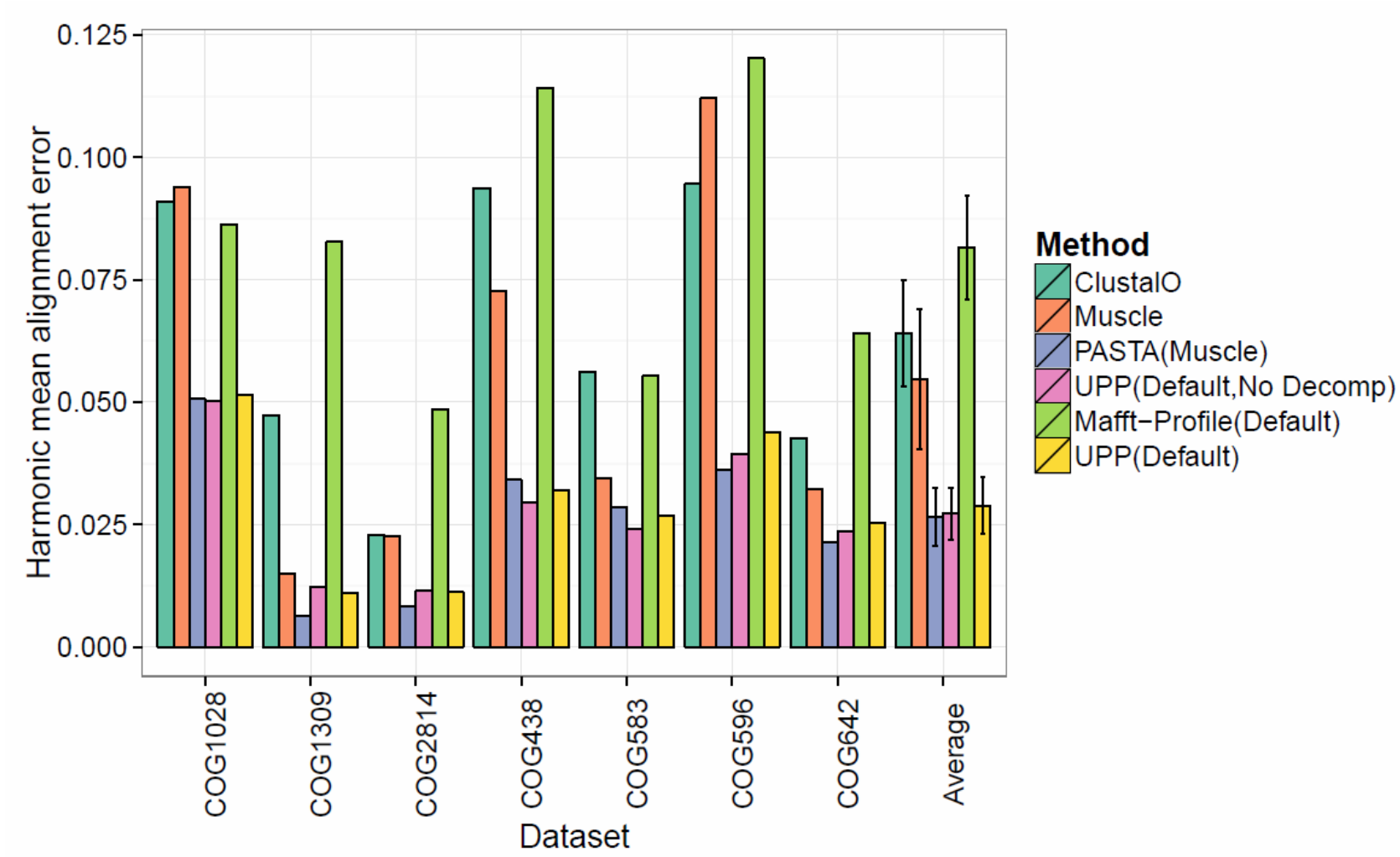

Note: ClustalO and Mafft-Profile(Default) fail to generate an alignment under this model condition.

# Indelible 10K : Tree error

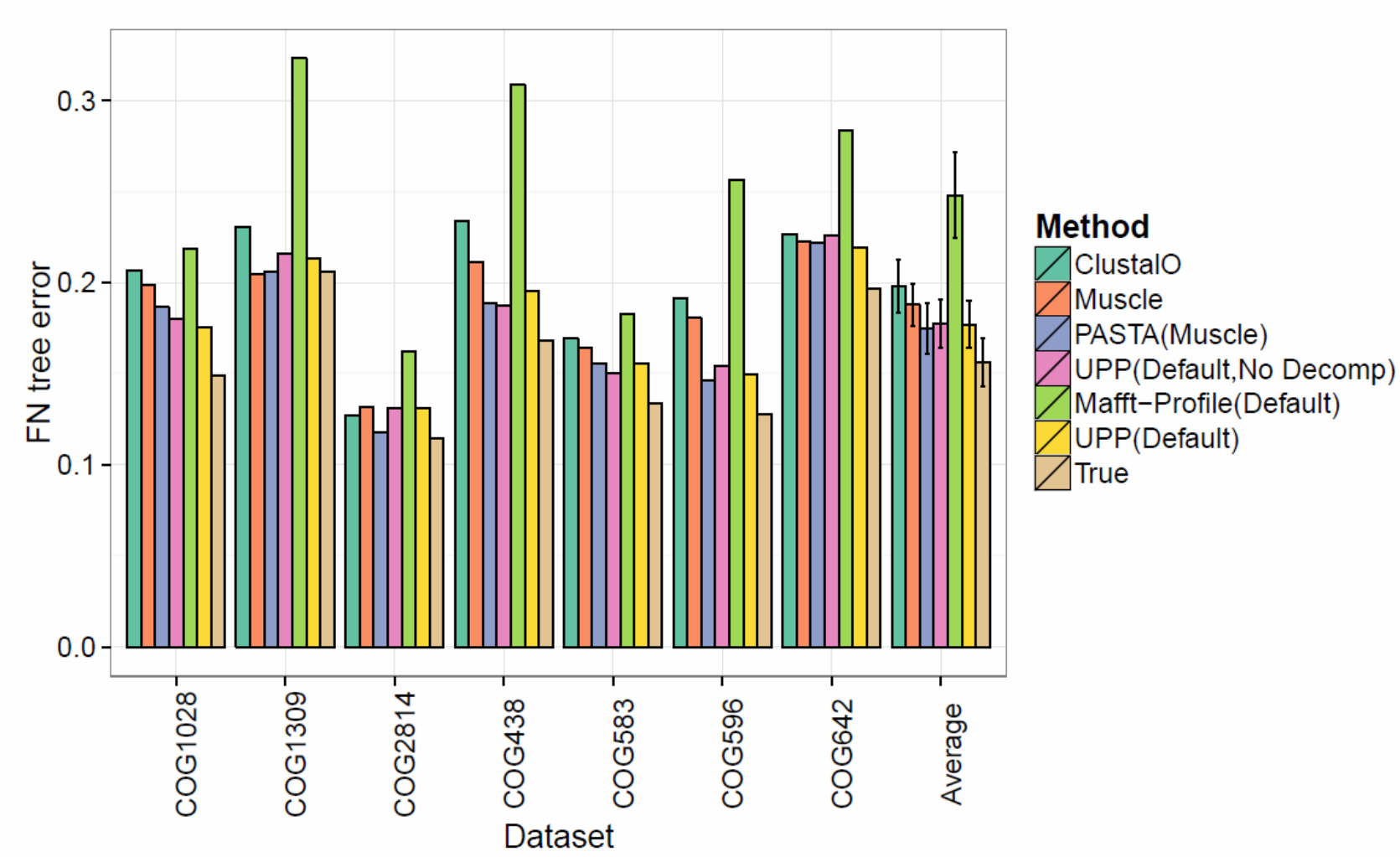



Note: ClustalO and Mafft-Profile(Default) fail to generate an alignment under this model condition.

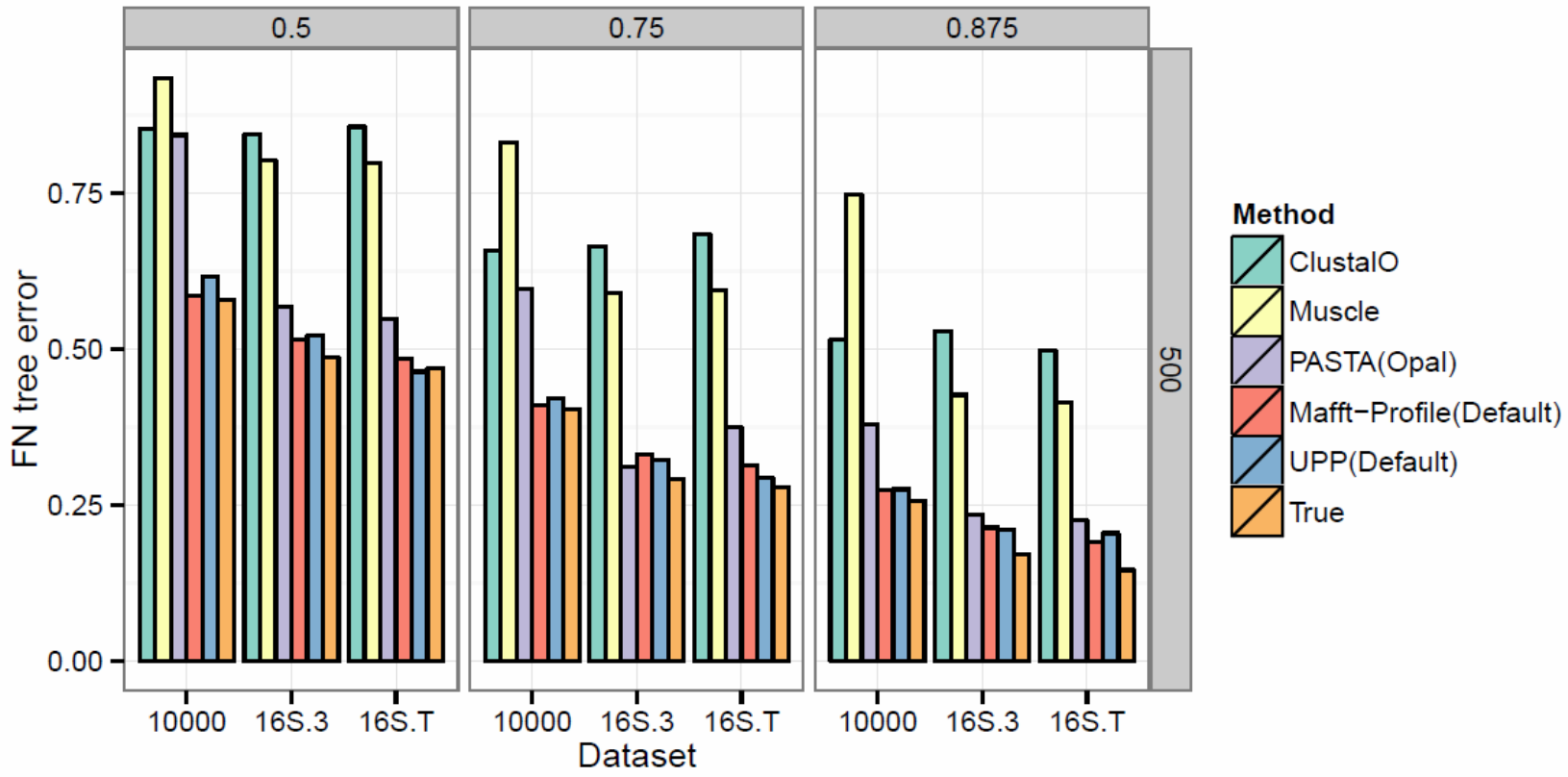
# FastTree AA: Alignment error



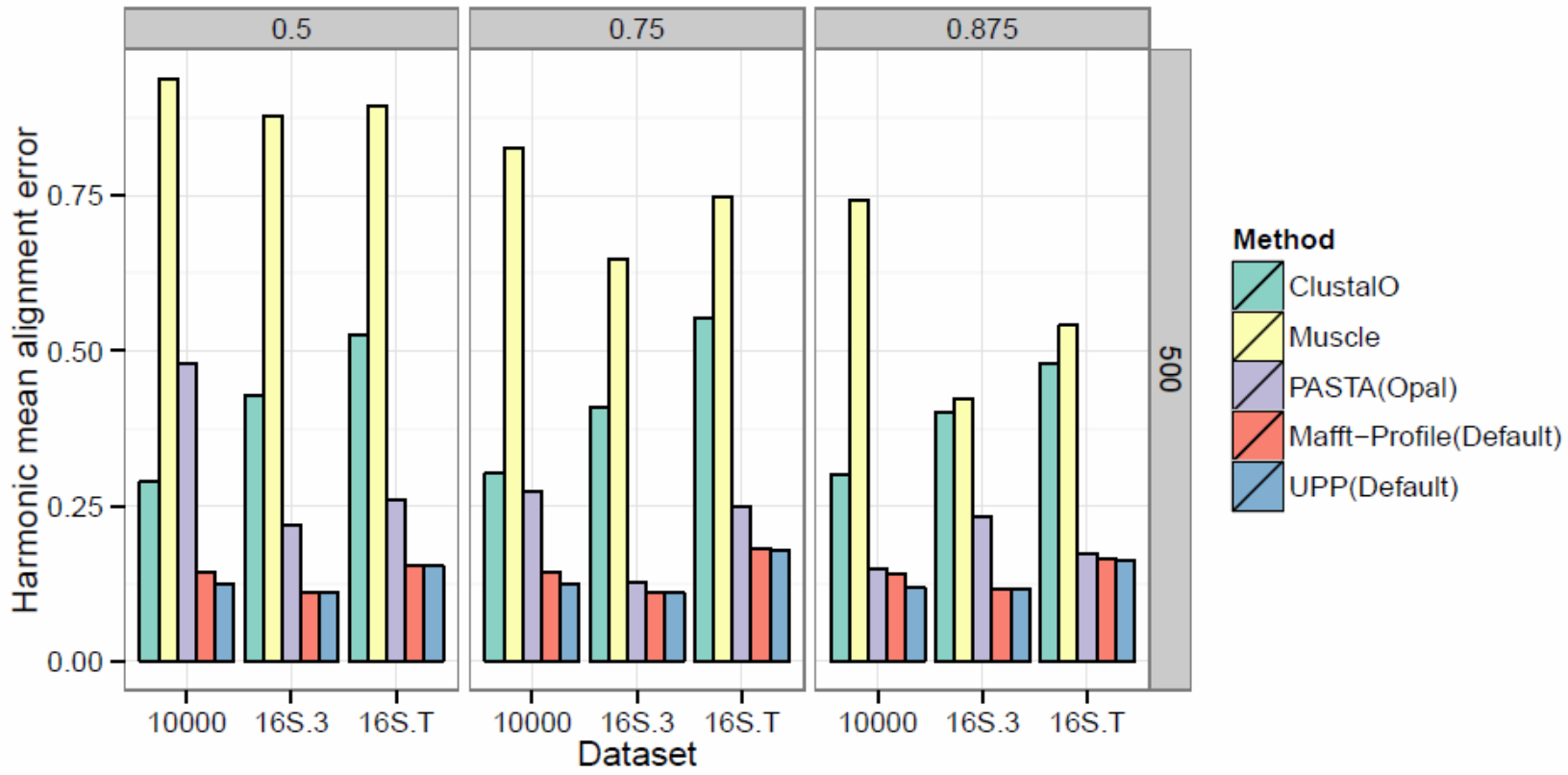
# FastTree AA: Tree error



# Large fragmentary: Tree error



# Large fragmentary: Alignment error




10000-sequence dataset from RNASim
16S.3 and 16S.T from Gutell's Comparative Ribosomal Website (CRW)