New Phylogenetic Placement and Taxon Identification Methods for Metagenomic Data

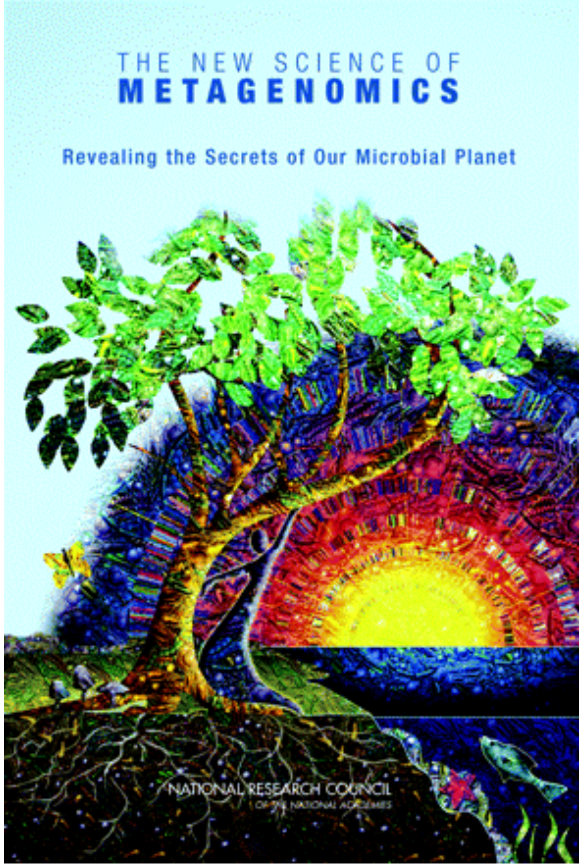
Tandy Warnow

Department of Computer Science

The University of Texas at Austin

Computational Phylogenetics and Metagenomics

Courtesy of the Tree of Life project



THE NEW SCIENCE OF
METAGENOMICS

Revealing the Secrets of Our Microbial Planet

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

NGS and metagenomic data

- Fragmentary data (e.g., short reads):
  - How to align? How to insert into trees?

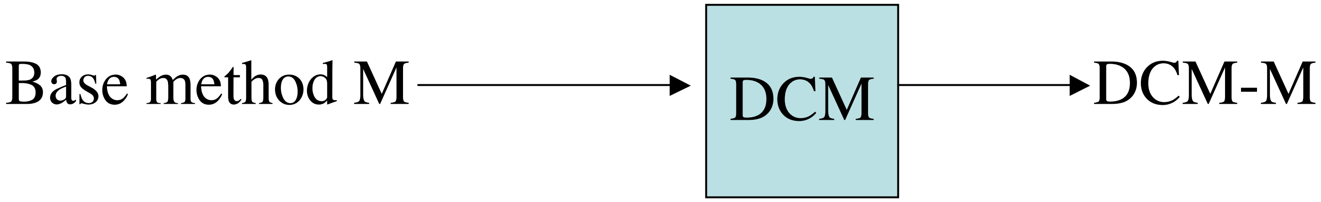
- Unknown taxa
  - How to identify the species, genus, family, etc?

Major Challenges

- Many phylogenetic datasets contain hundreds to thousands of species, some with thousands of genes. *Current alignment and tree estimation methods have poor accuracy or cannot run on large datasets, especially if the data are fragmentary.*
- Metagenomic datasets contain millions of short reads or contigs. *Current taxon identification methods have insufficient sensitivity, and high throughput is essential.*

Disk-Covering Methods (DCMs)

- DCMs “boost” the performance of phylogeny reconstruction methods.

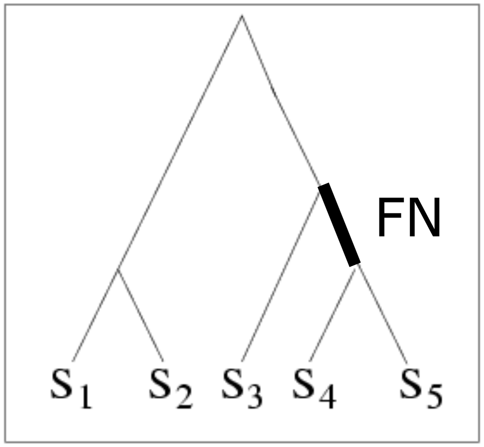


Base method M

DCM

DCM-M

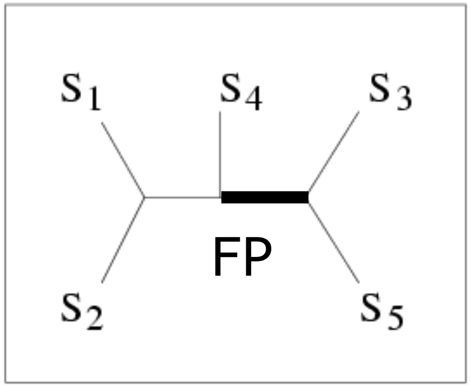
Quantifying Error



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

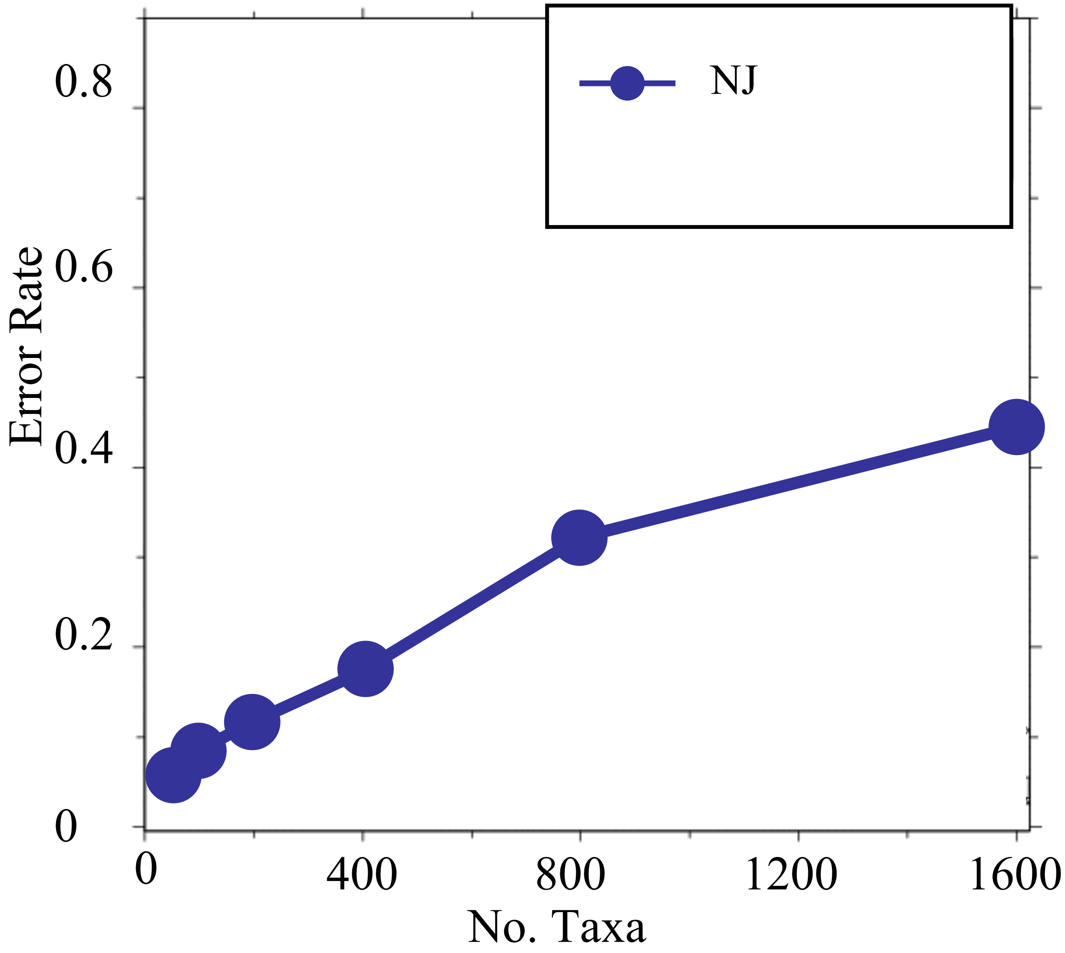


INFERRED TREE

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

50% error rate

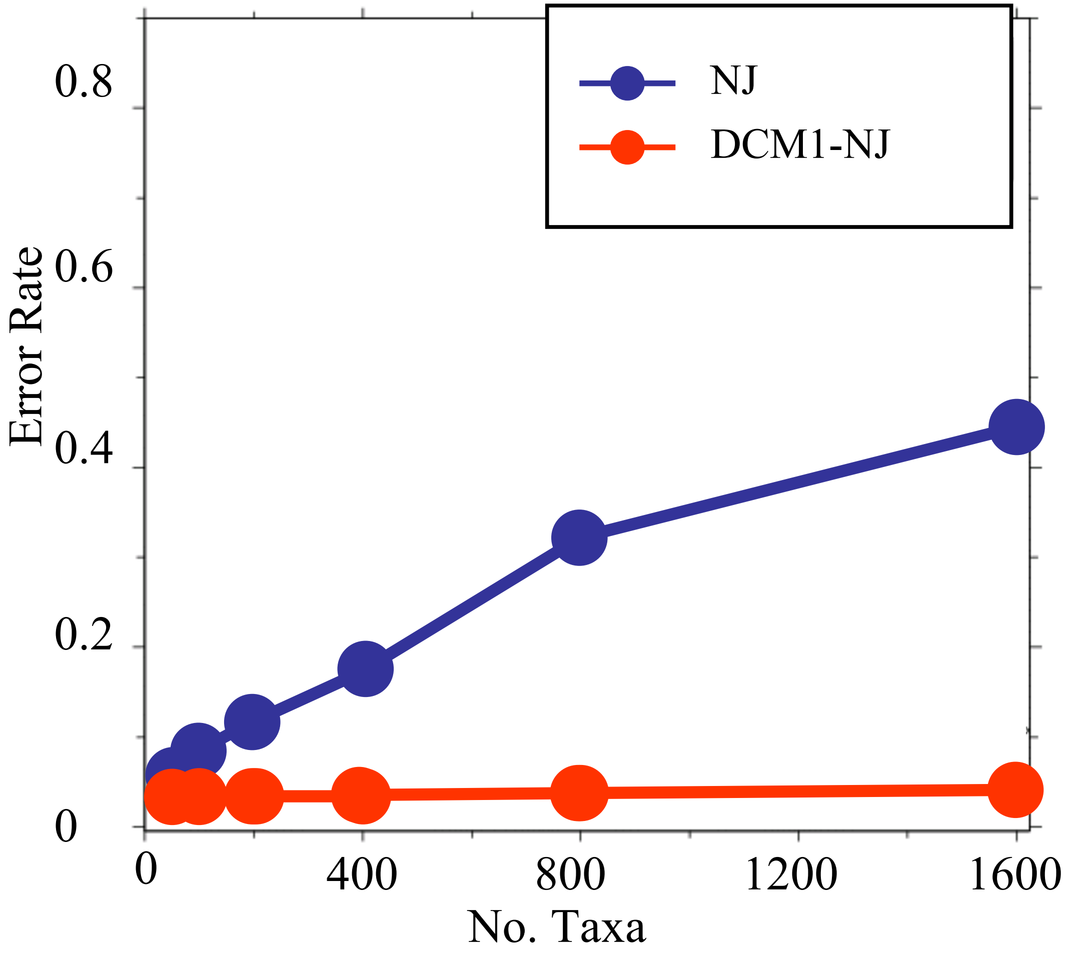
Neighbor joining has poor performance on large diameter trees *[Nakhleh et al. ISMB 2001]*



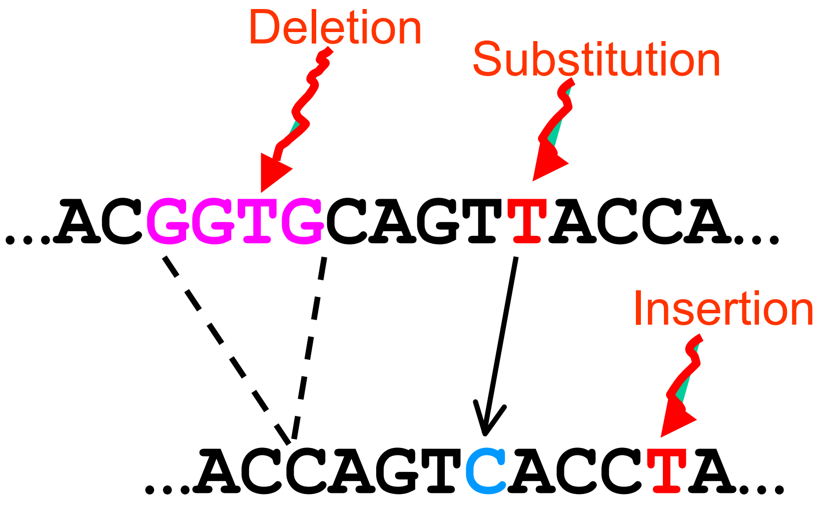
Exponential sequence length requirement for Neighbor Joining (Lacey and Chang, 2006)

DCM1-boosting distance-based methods

[Nakhleh et al. ISMB 2001]



Theorem: DCM1-NJ converges to the true tree from polynomial length sequences (Warnow et al., SODA 2001)



Deletion

Substitution

...ACGGTGCAGTTACCA...

Insertion

...ACCAGTCACCTA...

```
...ACGGTGCAGTTACC-A...
...AC----CAGTCACCTA...
```

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: Multiple Sequence Alignment

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

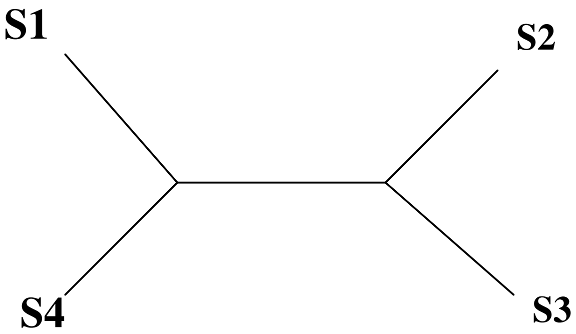
→

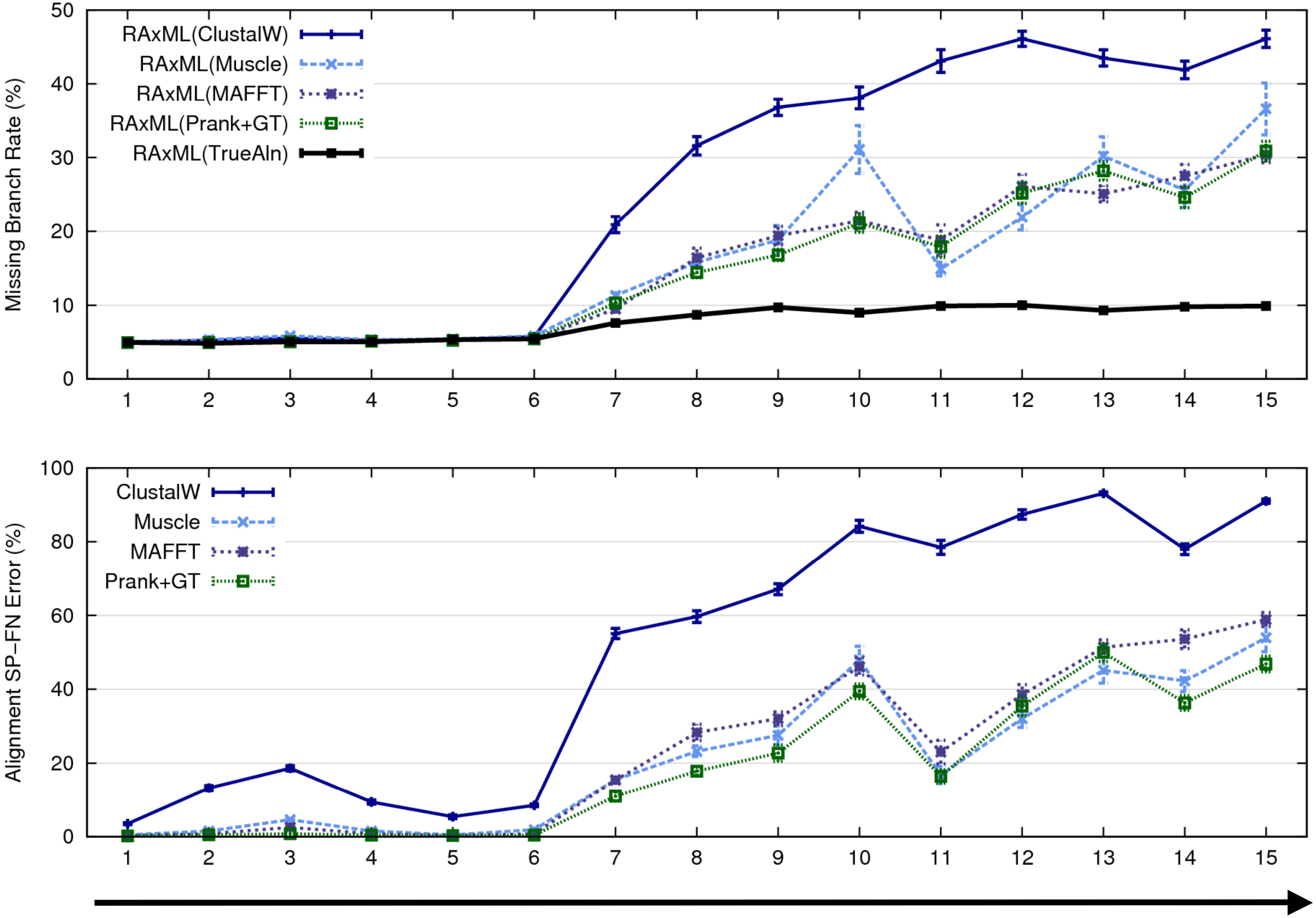
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

Phase 2: Construct tree

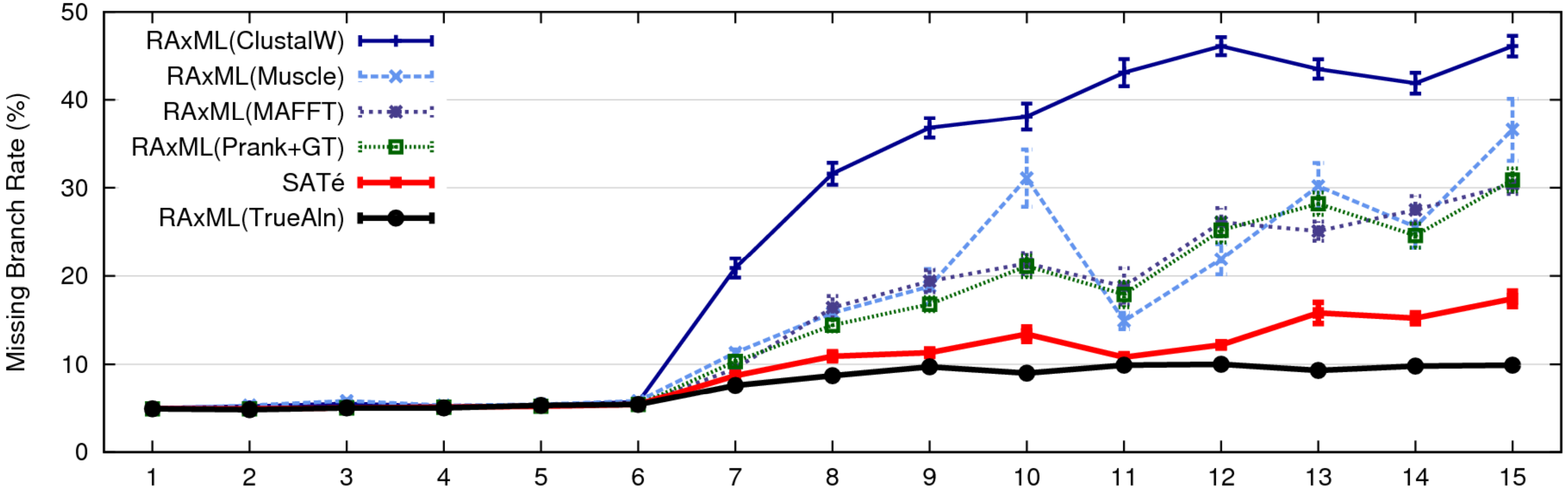
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```

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





1000 taxon models, ordered by difficulty (Liu et al., 2009)

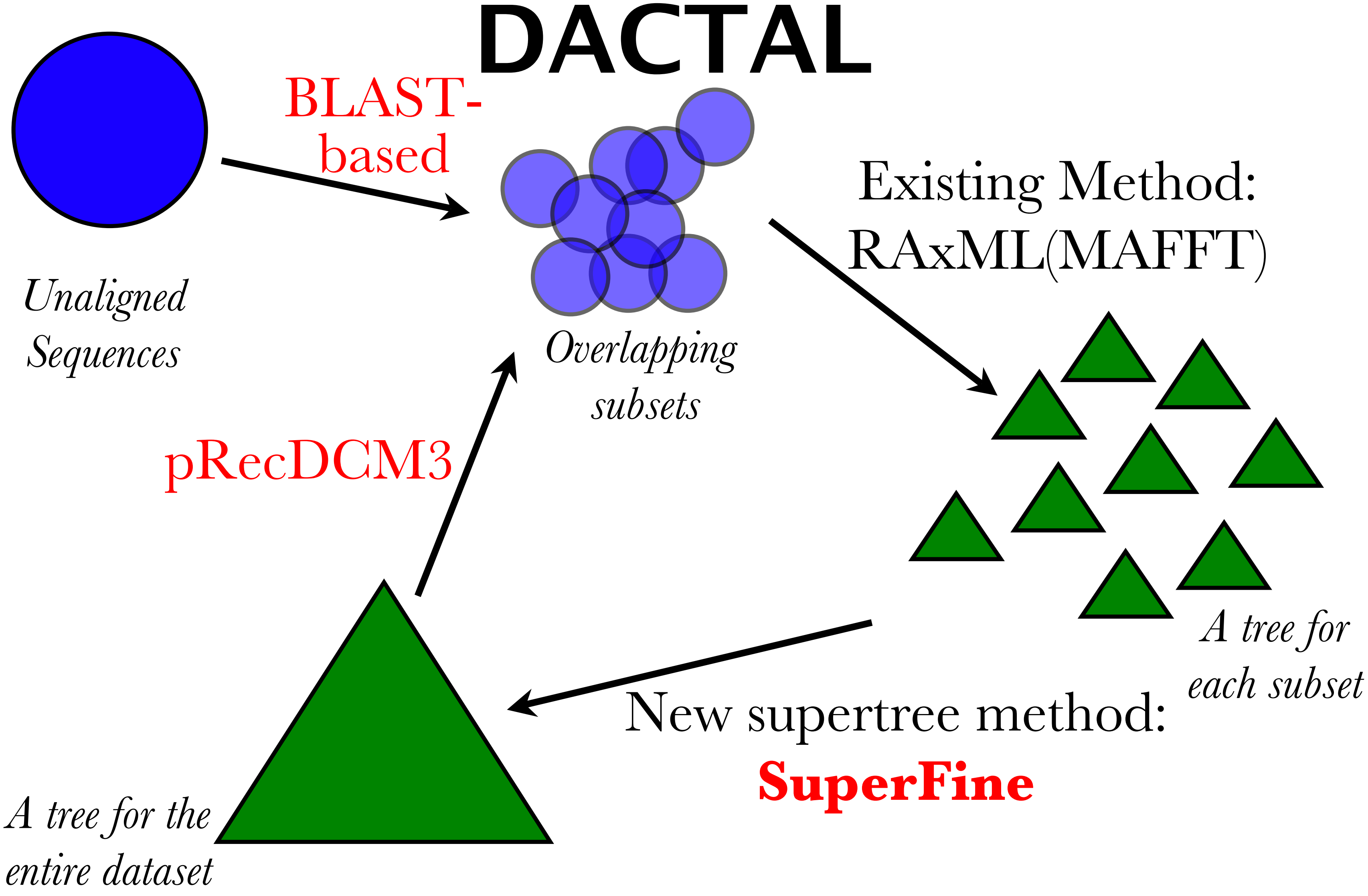


1000 taxon models, ordered by difficulty

24 hour SATé analysis, on desktop machines

(Similar improvements for biological datasets)

DACTAL



Unaligned Sequences

BLAST-based

Overlapping subsets

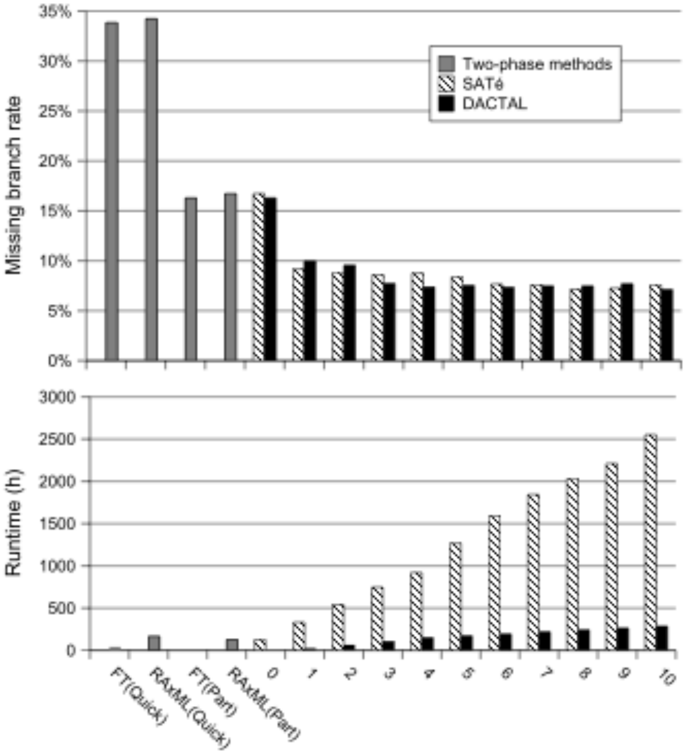
Existing Method: RAxML(MAFFT)

A tree for each subset

New supertree method: **SuperFine**

A tree for the entire dataset

pRecDCM3



DACTAL: as accurate as SATé
(but faster!)

DACTAL: better results than 2-phase methods

Three 16S datasets from Gutell's database (CRW) with

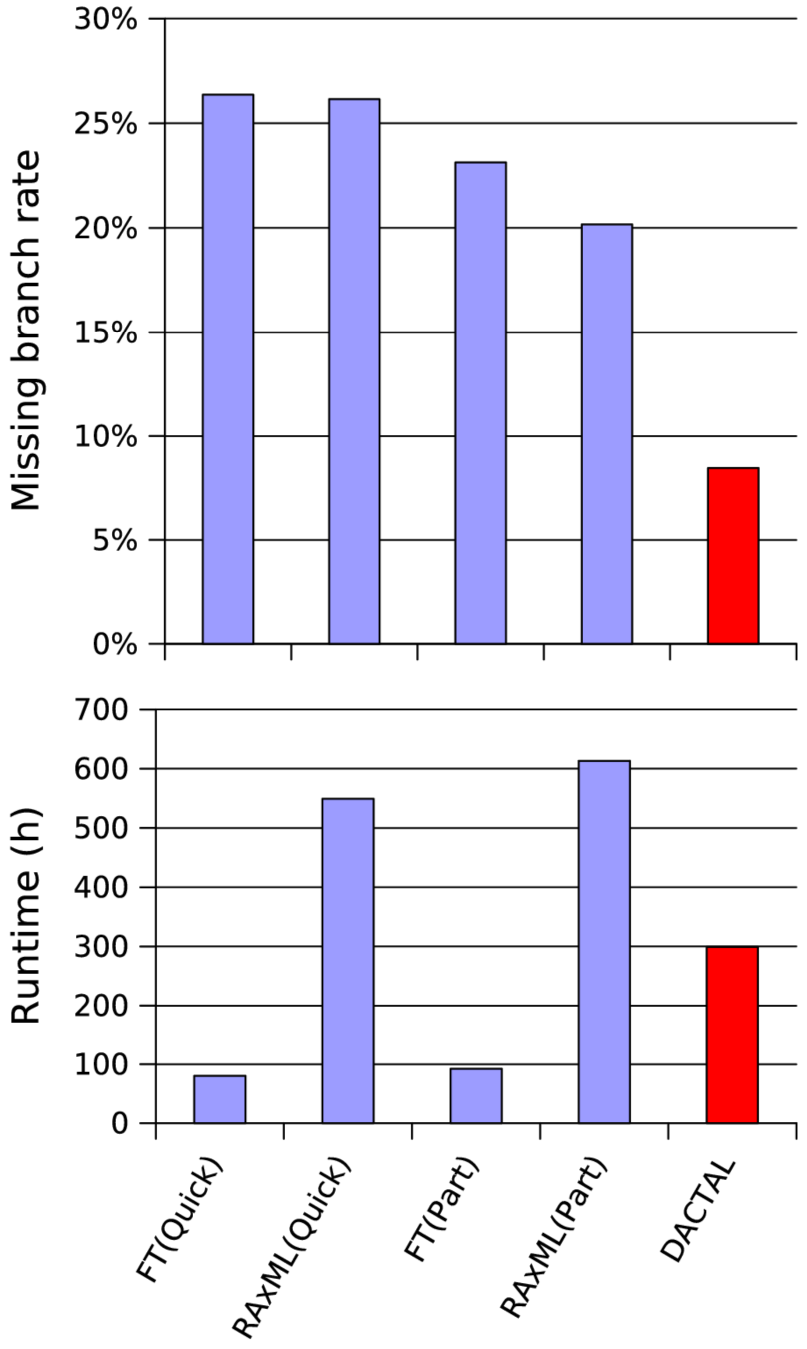
6,323 to **27,643** sequences

Reference alignments based on secondary structure

Reference trees are 75% RAxML bootstrap trees

DACTAL (shown in red) run for 5 iterations starting from FT(Part)

FastTree (FT) and RAxML are ML methods



Phylogenetic “boosters” (meta-methods)

Goal: improve accuracy, speed, robustness, or theoretical guarantees of base methods

Examples:

- DCM-boosting for distance-based methods (1999)
- DCM-boosting for heuristics for NP-hard problems (1999)
- SATé-boosting for alignment methods (2009)
- SuperFine-boosting for supertree methods (2011)
- DACTAL-boosting for all phylogeny estimation methods (2012)
- SEPP-boosting for phylogenetic placement (2012)
- TIPP-boosting for taxon identification (in preparation)

NGS and metagenomic data

- Fragmentary data (e.g., short reads):
  - How to align? How to insert into trees?

- Unknown taxa
  - How to identify the species, genus, family, etc?

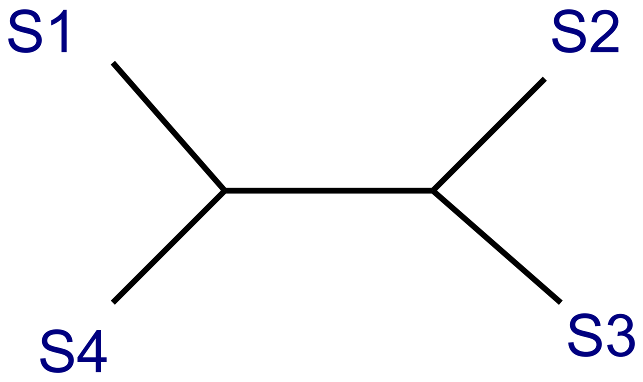
Phylogenetic Placement

Input: Backbone alignment and tree on full-length sequences, and a set of query sequences (short fragments)

Output: Placement of query sequences on backbone tree

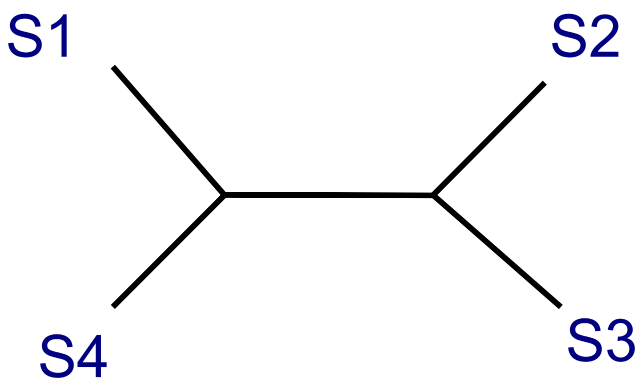
Align Sequence

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



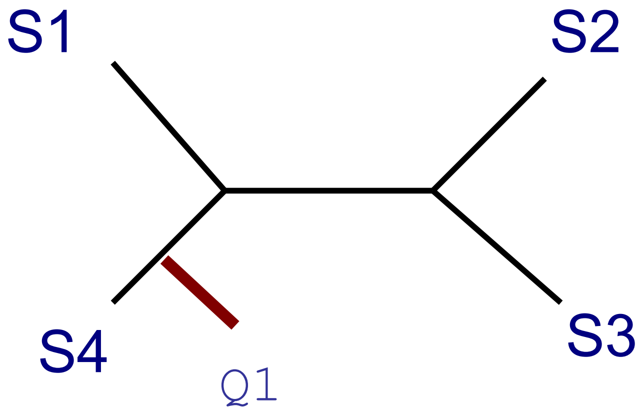
Align Sequence

```
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--------
```

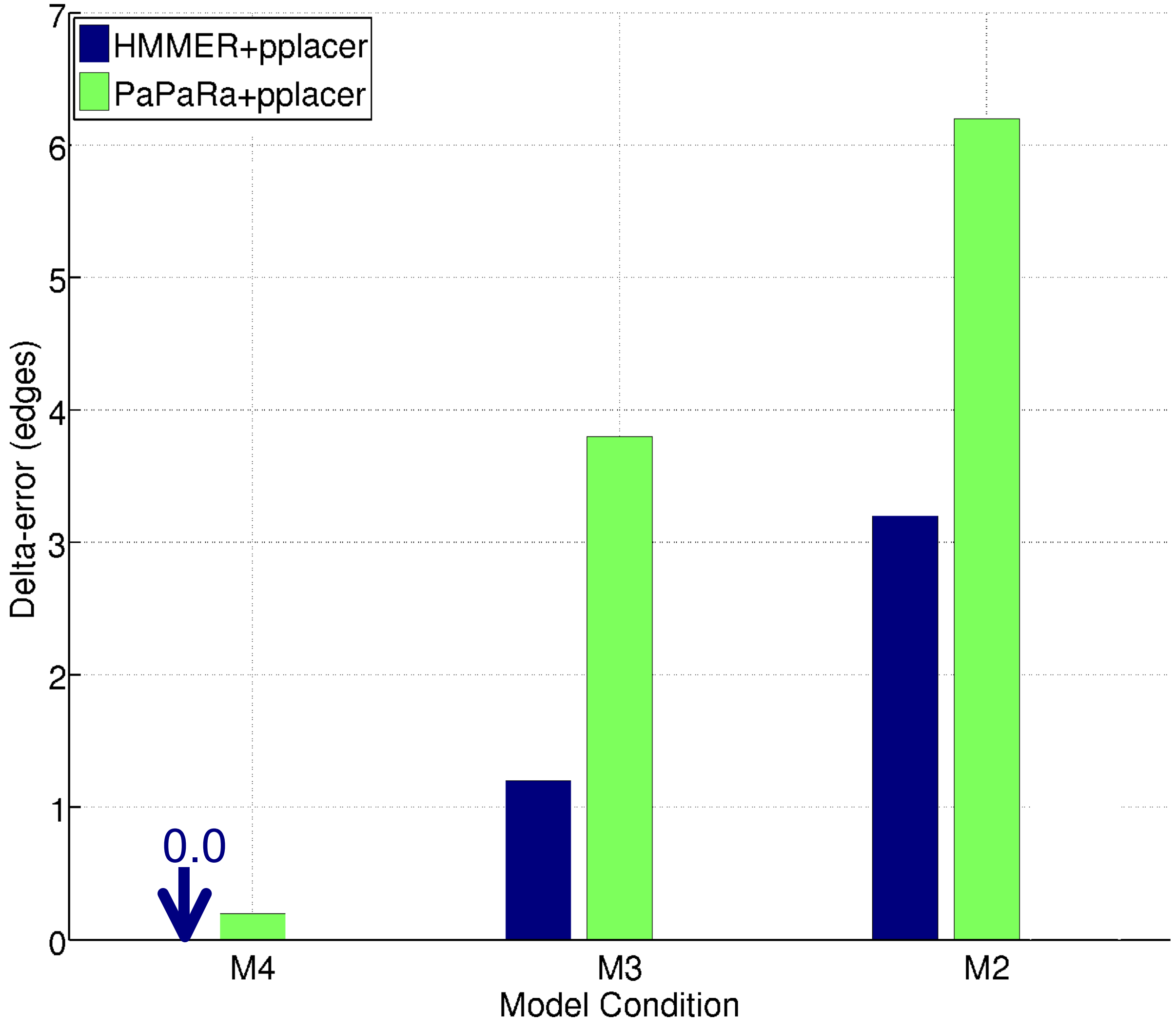


Phylogenetic Placement

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

Note: pplacer and EPA use maximum likelihood

HMMER vs. PaPaRa

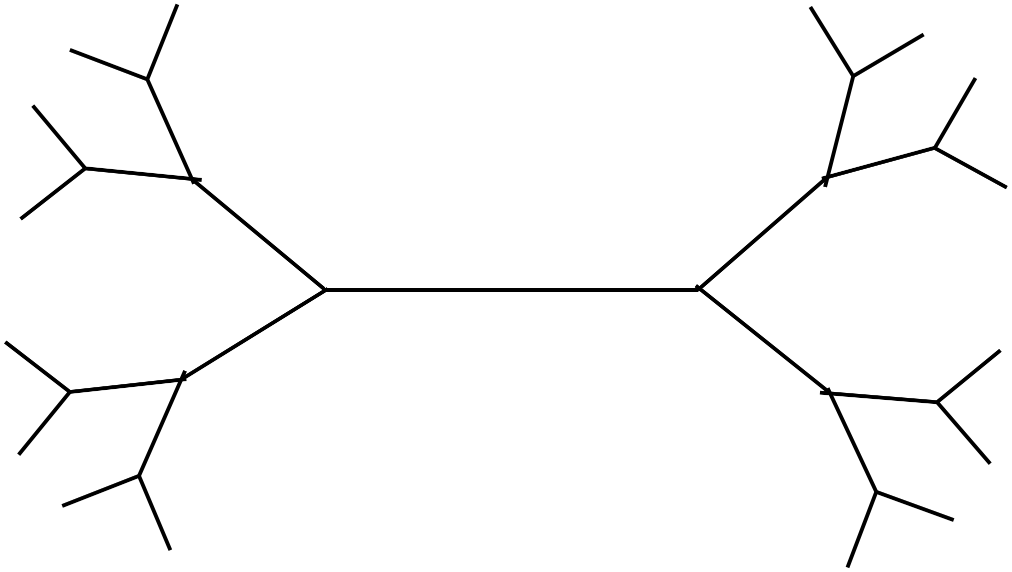


Increasing rate of evolution

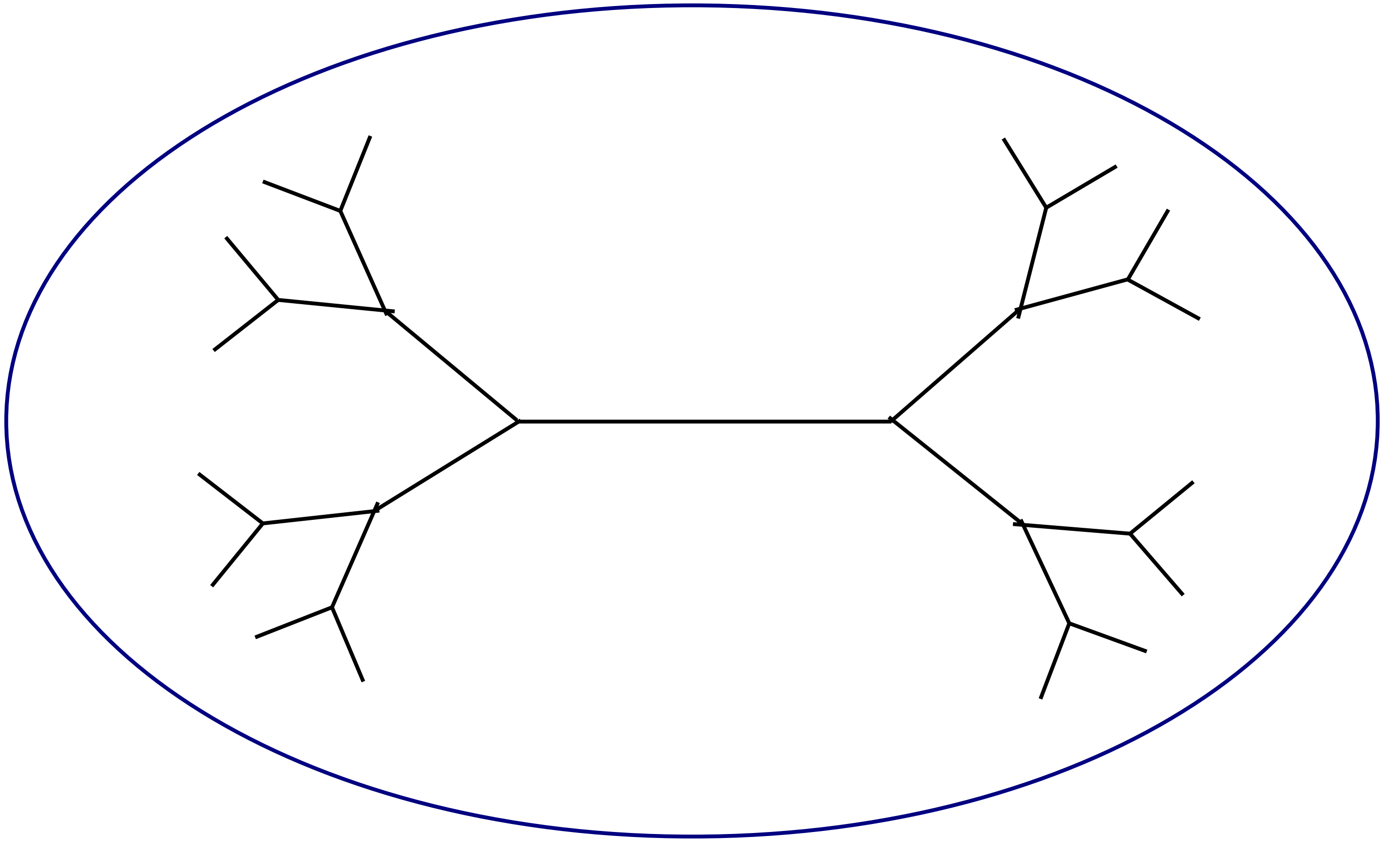
SEPP

- Key insight: HMMs are not very good at modelling MSAs on large, divergent datasets.
- Approach: insert fragments into taxonomy using estimated alignment of full-length sequences, and multiple HMMs (on different subsets of taxa).

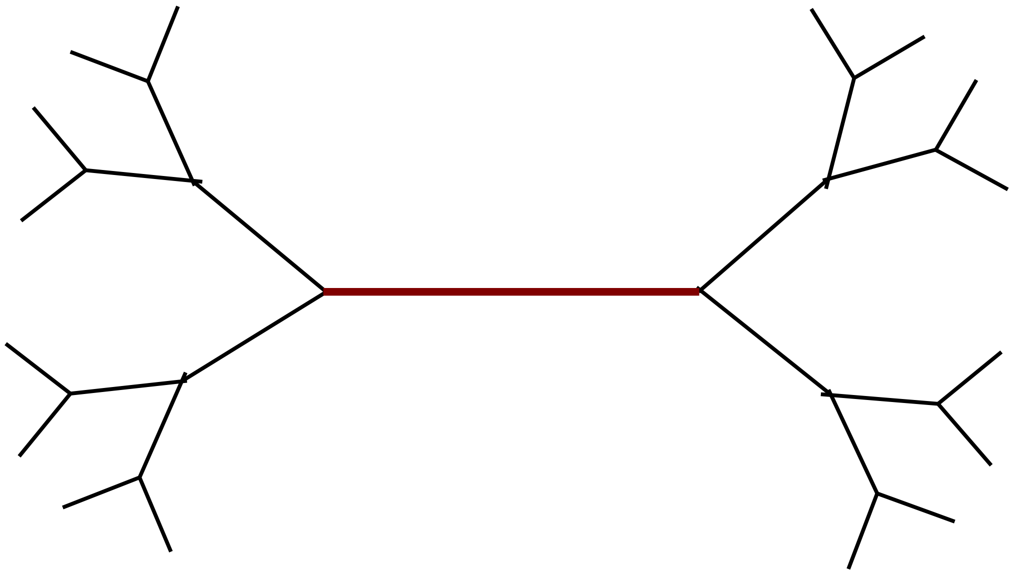
SEPP: SATé-enabled Phylogenetic Placement



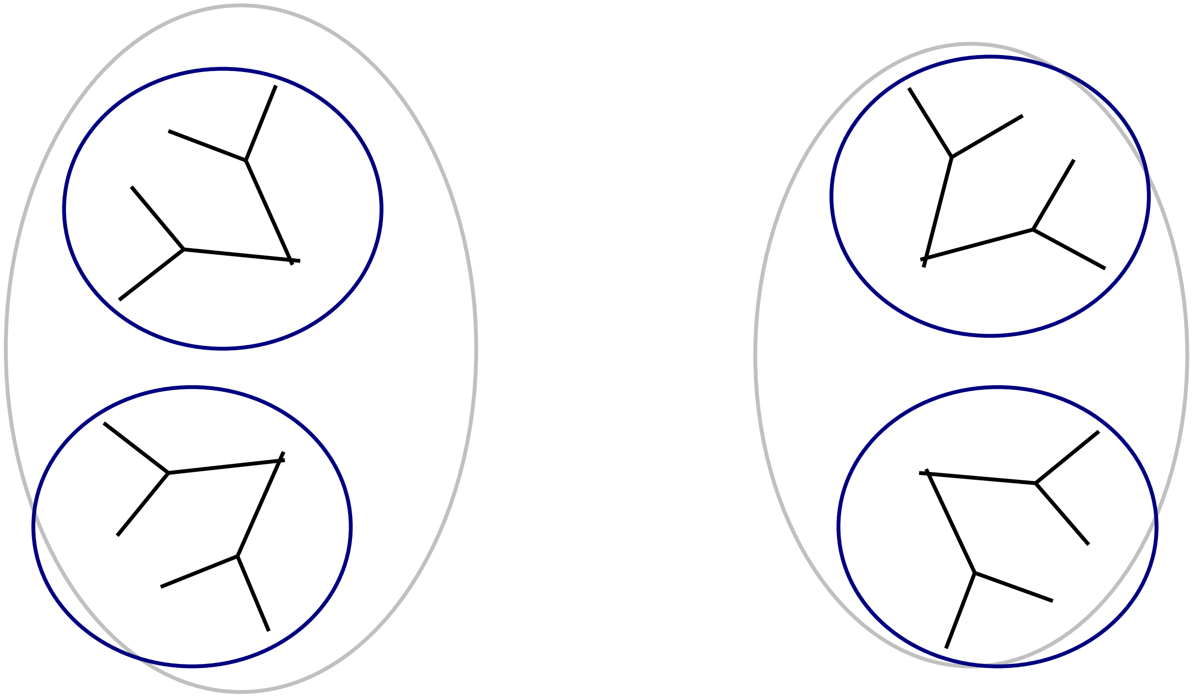
SEPP: SATé-enabled Phylogenetic Placement



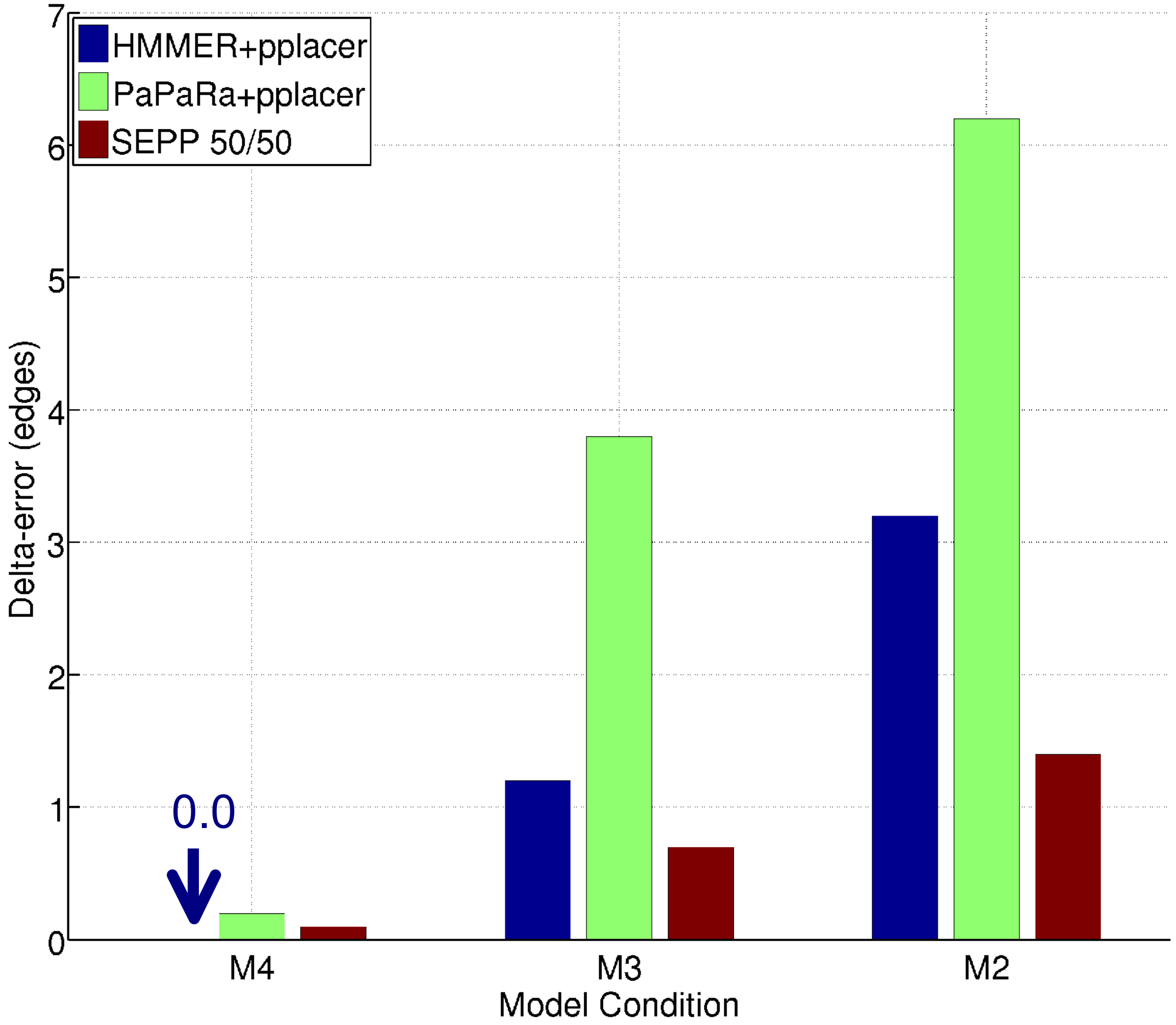
SEPP: SATé-enabled Phylogenetic Placement



SEPP: SATé-enabled Phylogenetic Placement



SEPP (10%-rule) on simulated data



Increasing rate of evolution

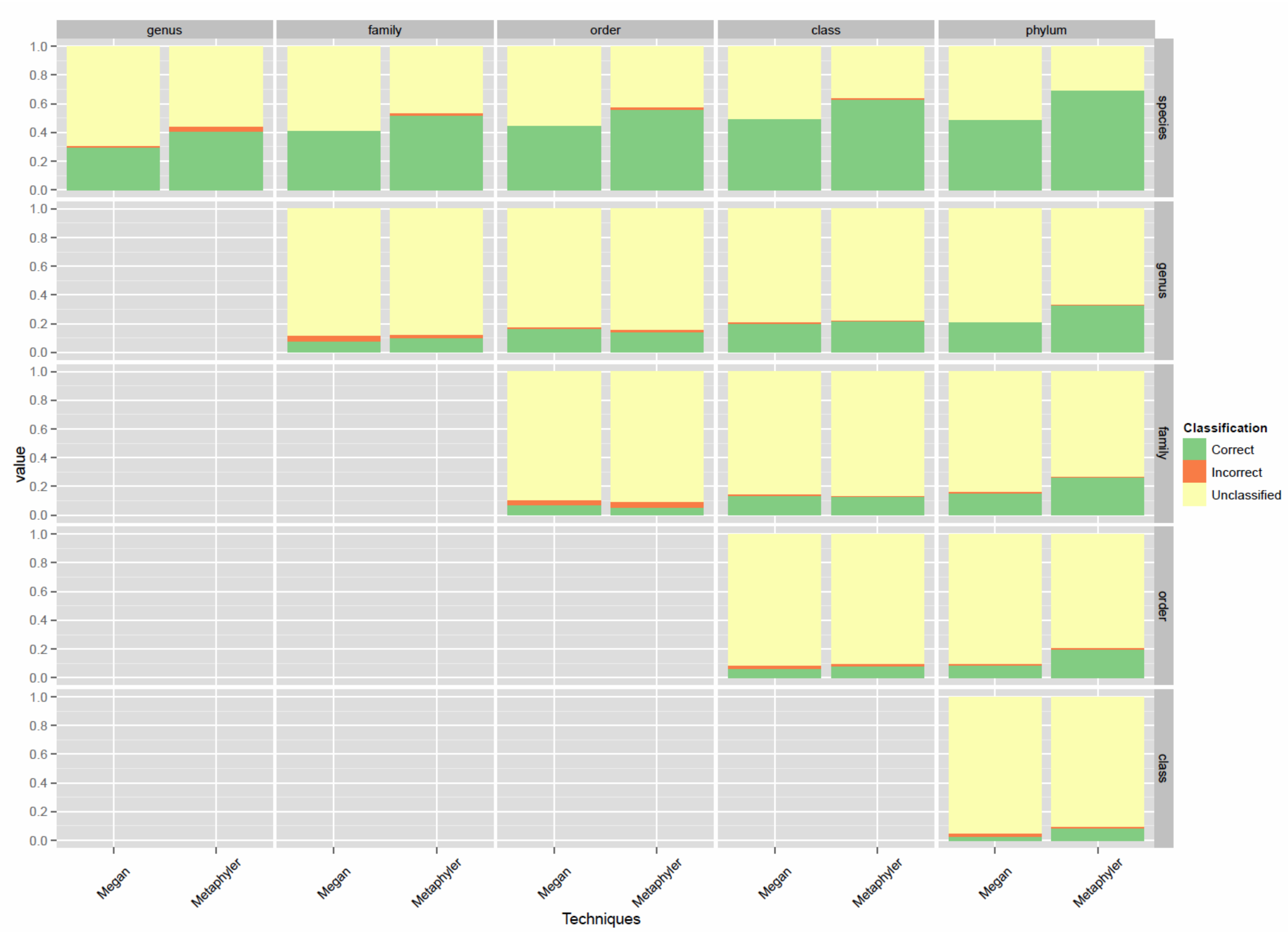
Part IV: Taxon Identification

Metagenomic datasets include short reads from unknown species

Taxon identification: given short sequences, identify the species for each fragment

Best current methods: Metaphyler, Phylopythia, and PhymmBL

60bp error free reads on rpsB marker gene

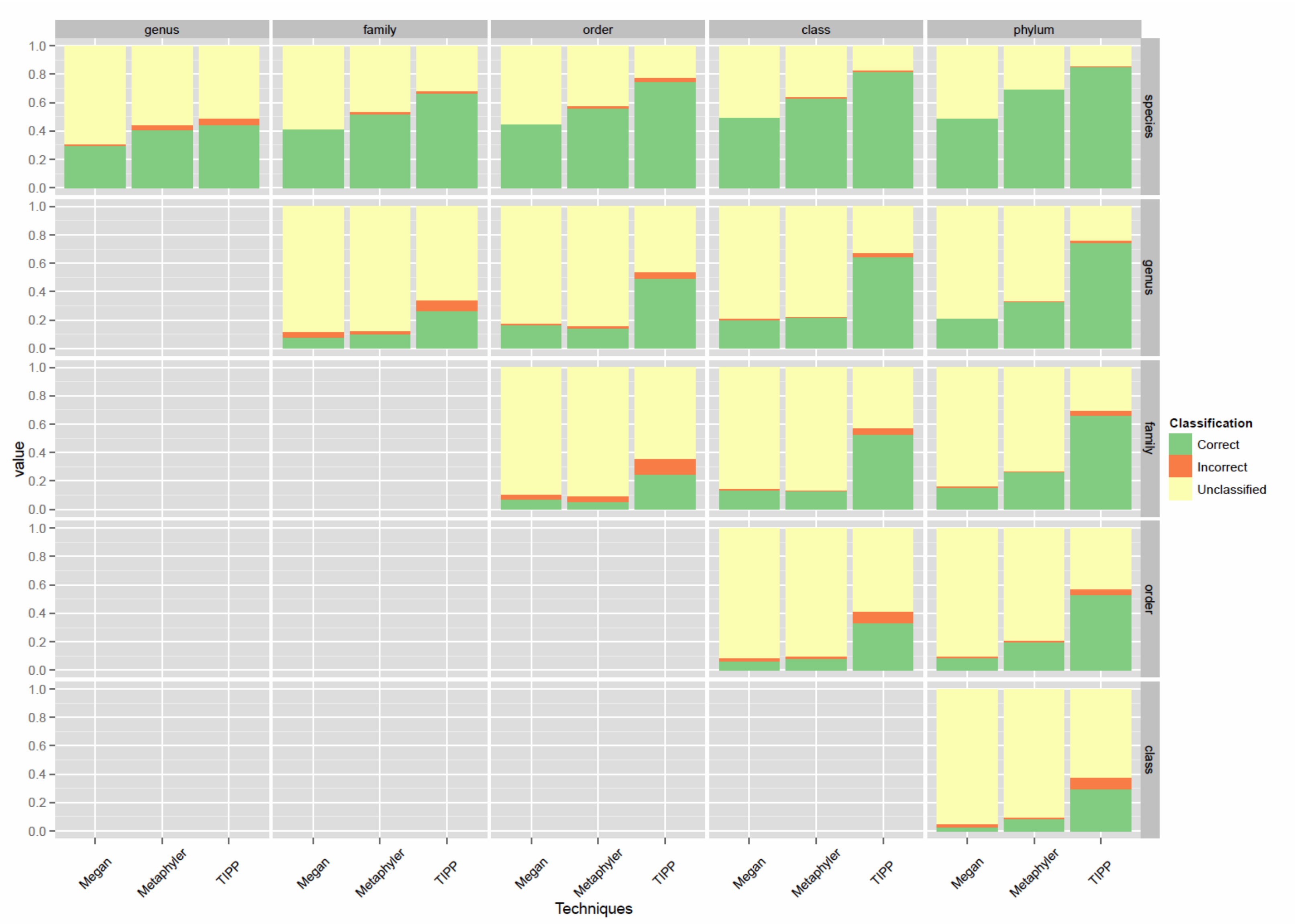


TIPP

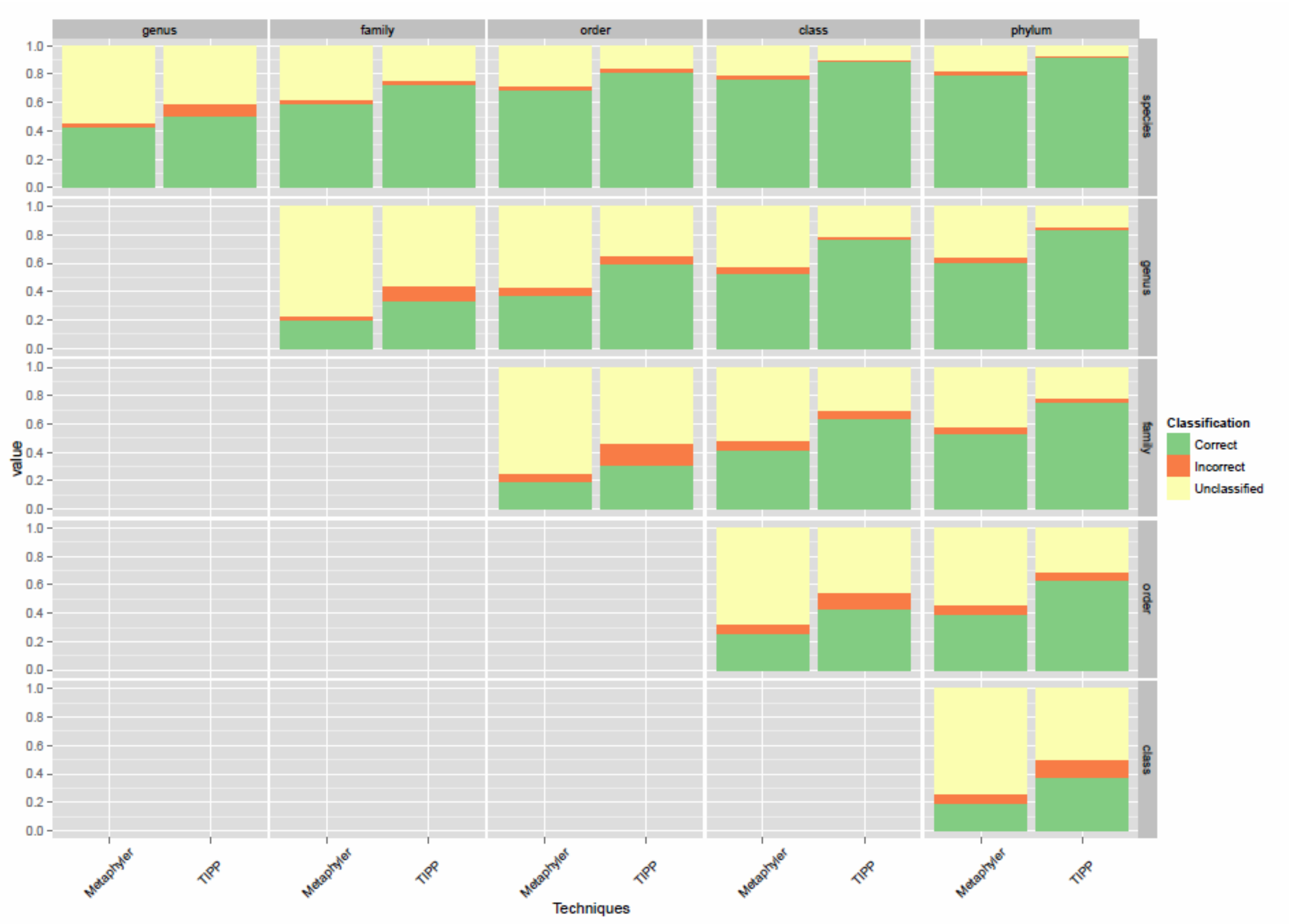
- Taxon Identification using Phylogenetic Placement (Nguyen, Mirarab, and Warnow, in preparation)

- Approach: SEPP, modified to *take statistical uncertainty into account*

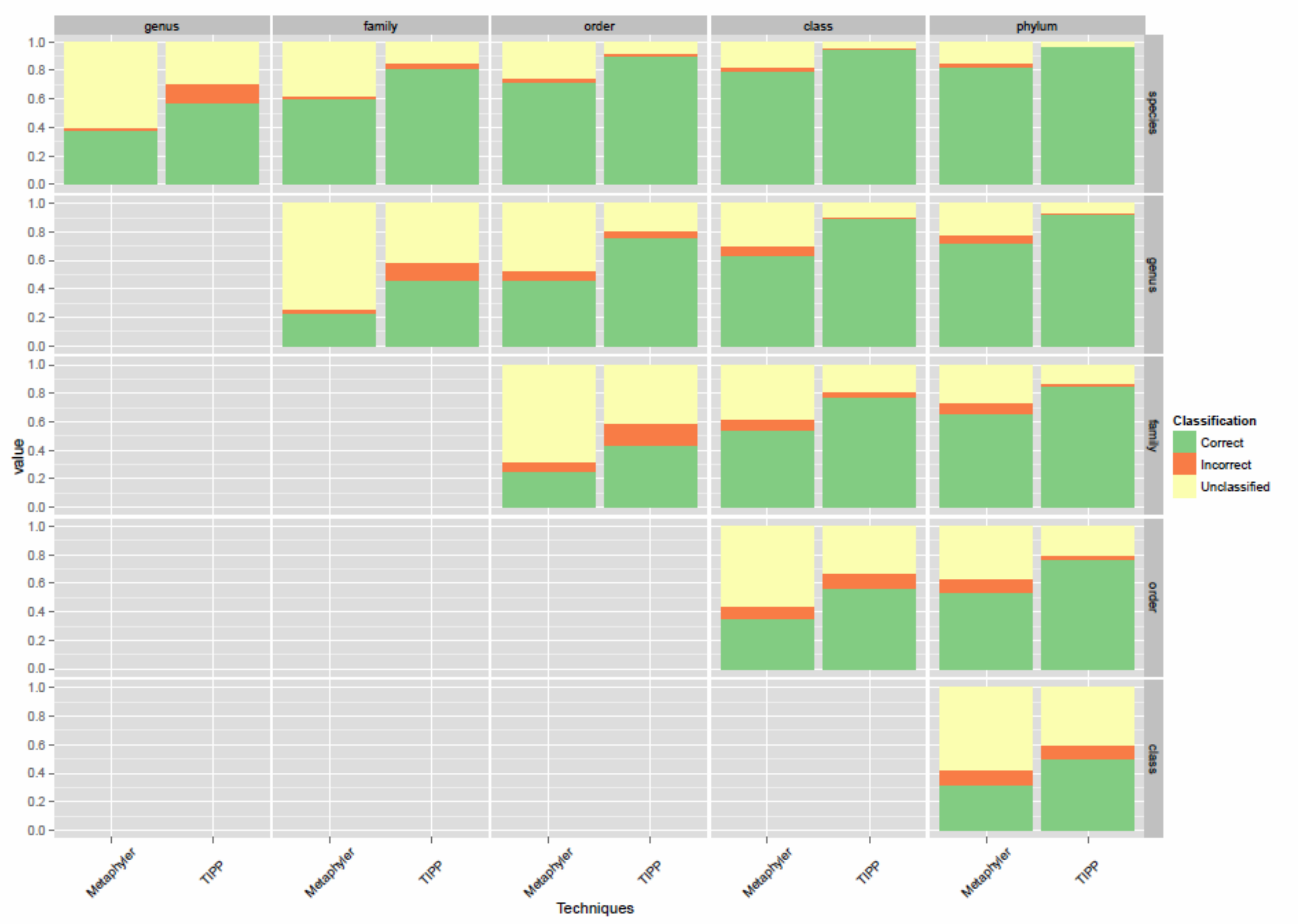
60bp error free reads on rpsB marker gene



MetaPhyler versus TIPP on 100bp Illumina reads across 29 marker genes



MetaPhyler versus TIPP on 300bp 454 reads across 29 marker genes



General Observations

- Relative performance of methods can change dramatically with dataset size.

- Standard statistical inference techniques often do not scale well.

- Divide-and-conquer and iteration can improve accuracy and speed of base methods.

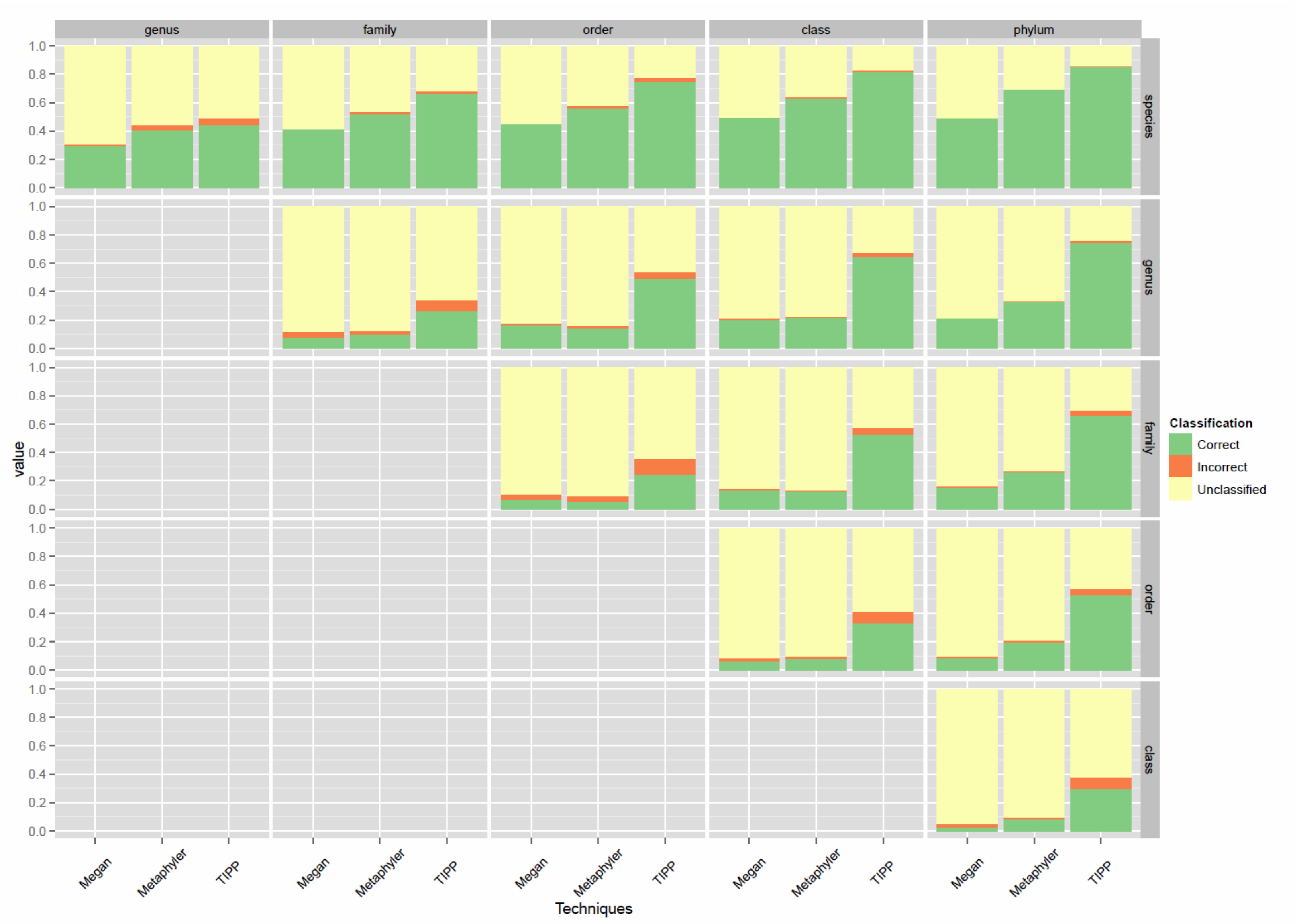
Acknowledgments

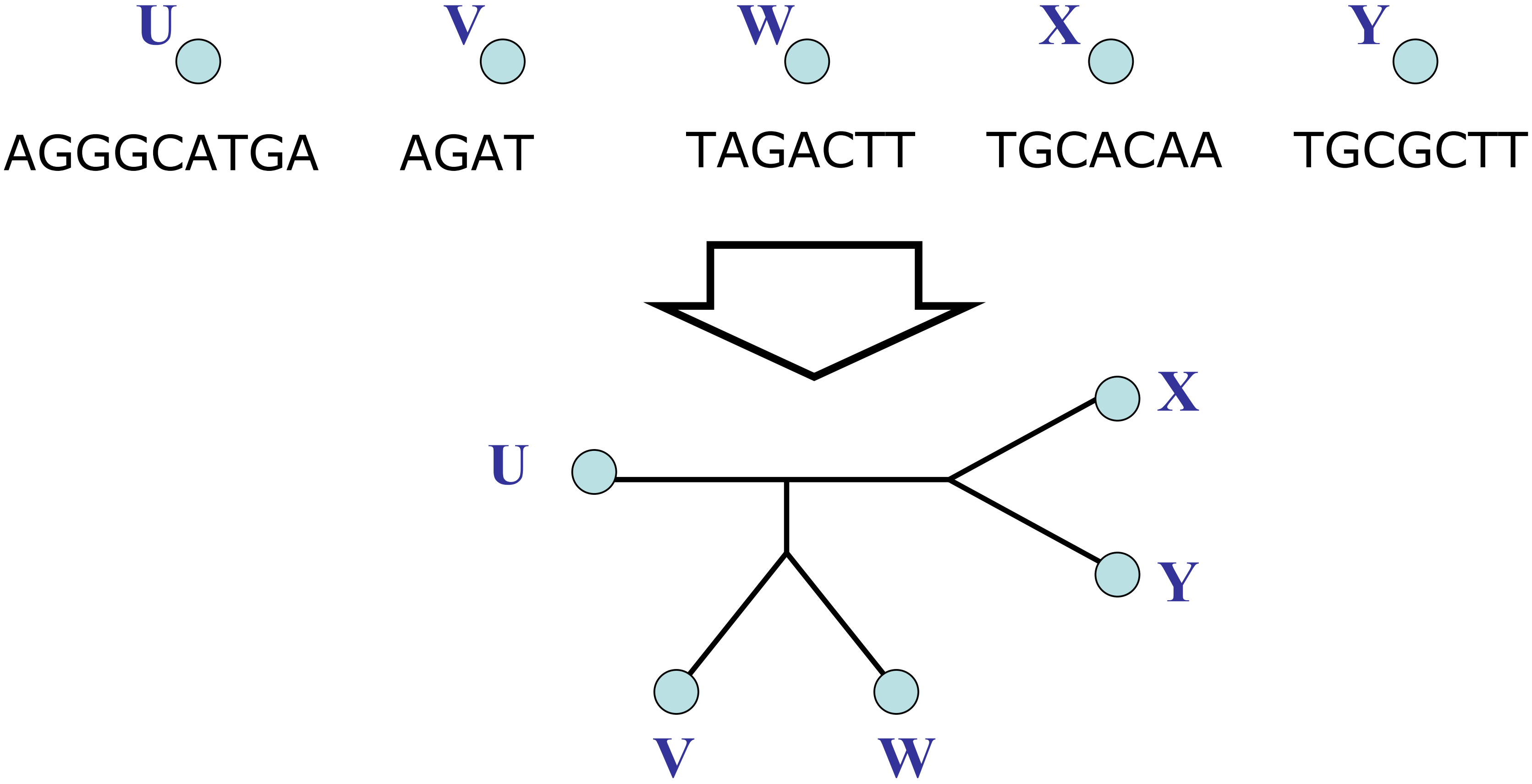
- Guggenheim Foundation Fellowship, Microsoft Research New England, National Science Foundation: Assembling the Tree of Life (ATOL), ITR, and IGERT grants, and David Bruton Jr. Professorship

- Collaborators:
  - DCM-NJ: Bernard Moret and Katherine St. John
  - SATé: Kevin Liu, Serita Nelesen, Sindhu Raghavan, and Randy Linder
  - DACTAL: Serita Nelesen, Kevin Liu, Li-San Wang, and Randy Linder
  - TIPP: Siavash Mirarab and Nam Nguyen

Happy Birthday!

60bp error-free reads on rpsB marker gene





AGGGCATGA AGAT TAGACTT TGCACAA TGCGCTT