Future Directions

Tandy Warnow
Department of Computer Science
The University of Texas at Austin

Monday's presentations

SATé improves accuracy for large-scale alignment and tree estimation

DACTAL enables phylogeny estimation for very large datasets, and may be robust to model violations

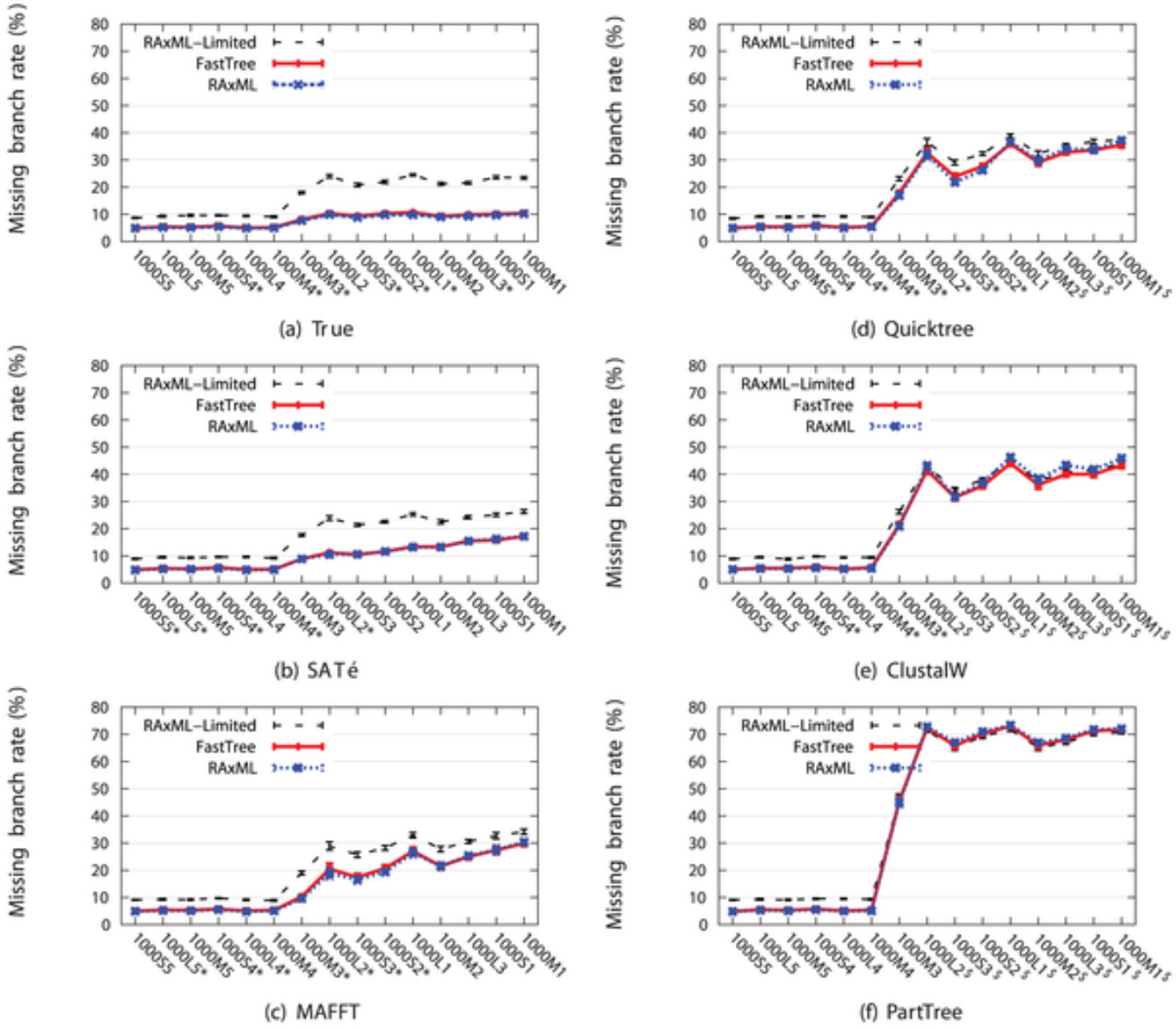
SEPP is useful for phylogenetic placement

See http://www.cs.utexas.edu/users/tandy/warnow-Smithsonian-May20.pdf

Project Software

SATé: Simultaneous Alignment and Tree Estimation
DACTAL: Divide-and-conquer trees (almost) without alignments
SEPP: SATé-enabled phylogenetic placement
TIPP: Taxon insertion using phylogenetic placement
BeeTLe: better treelength (improvement to POY)
SuiteMSA: alignment visualization and comparison tool
SuperFine: supertree estimation

Also partial support to:
Indel-seq-gen: simulation tool
GARLI: Genetic Algorithms for Rapid Likelihood
Various methods for species tree estimation from gene trees



RAxML vs. FastTree, Liu and Warnow, PLoS One, 2012

Possible Future Developments for SATé

GUI: integration for pipelines (model testing, visualization, bootstrapping, site-specific likelihoods)

Optimizing for

- very large datasets or very small datasets
- intron sequences
- coding sequences
- amino-acid sequences

Options for multi-marker analyses (different data types, consideration of gene tree discord)

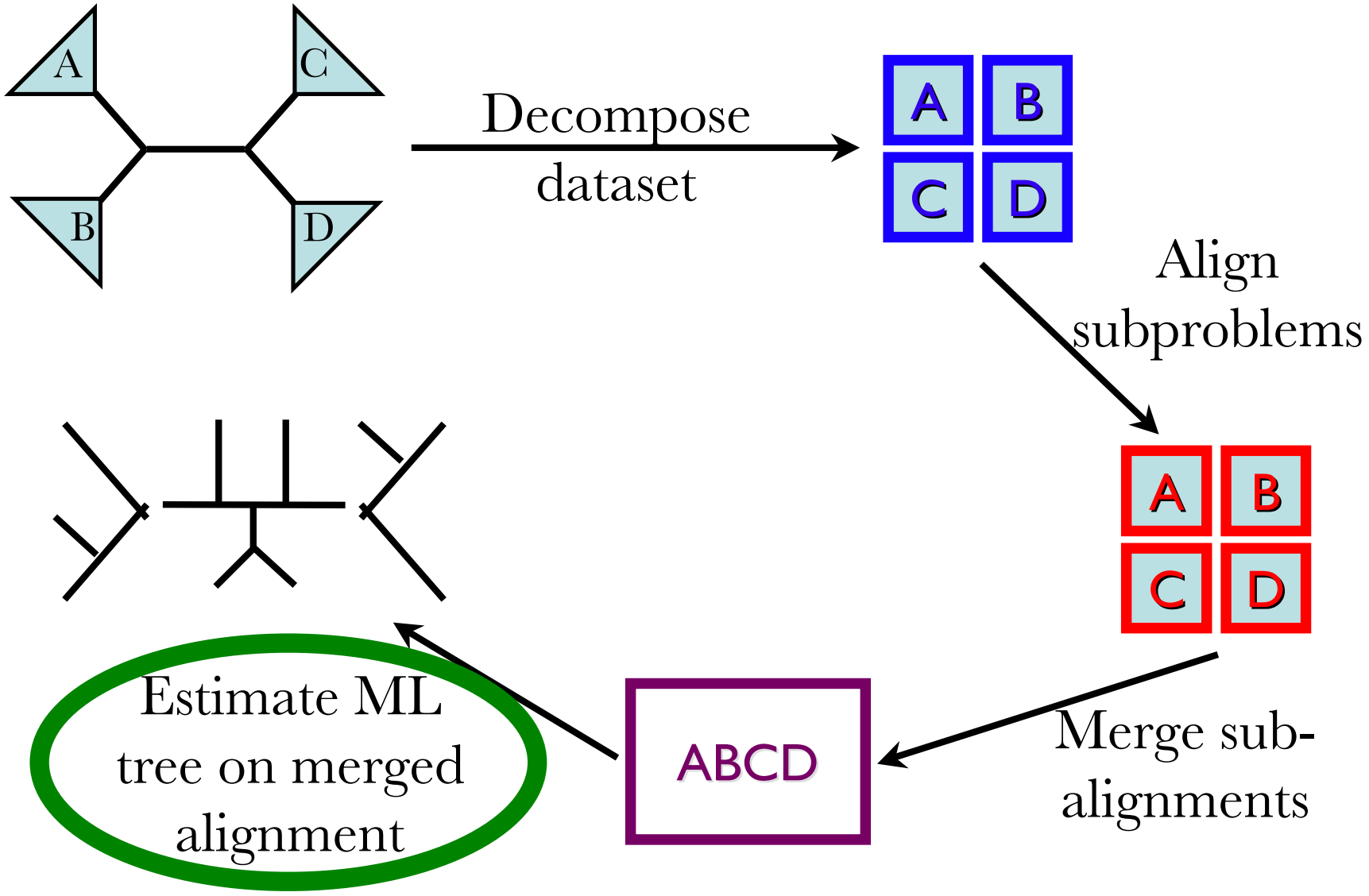
Integration with GARLI and other ML software that can handle non-standard models (e.g., nhPhyml)

Use of BAli-Phy scoring technique

Integration with new alignment methods (e.g., MAFFTash)

Exploration of alignment/tree pairs generated during SATé search

Limitations



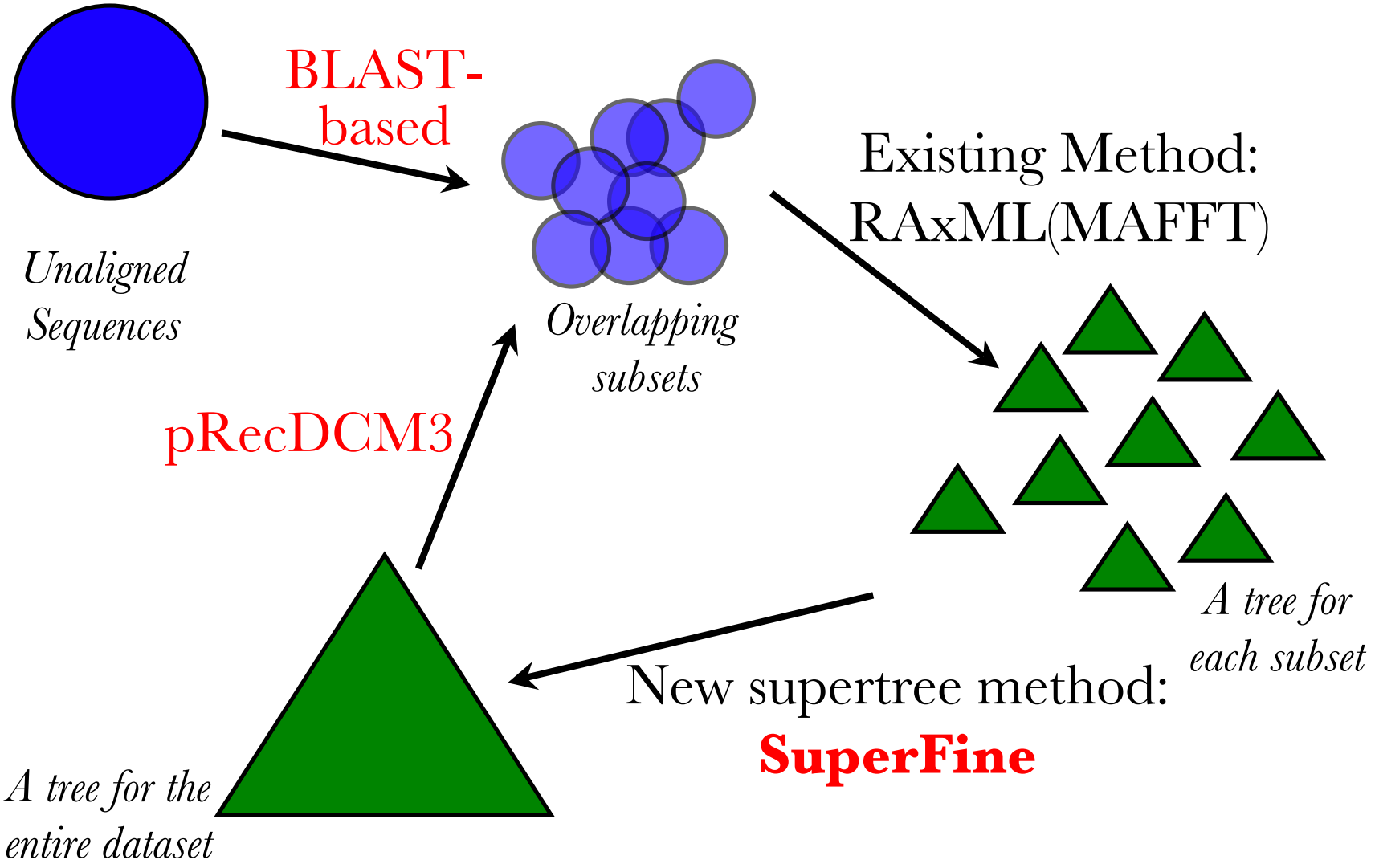
Part II: DACTAL

(Divide-And-Conquer Trees (Almost) without alignments)

- Input: set S of unaligned sequences
- Output: tree on S (but no alignment)

Nelesen, Liu, Wang, Linder, and Warnow, In Press, ISMB 2012 and Bioinformatics 2012

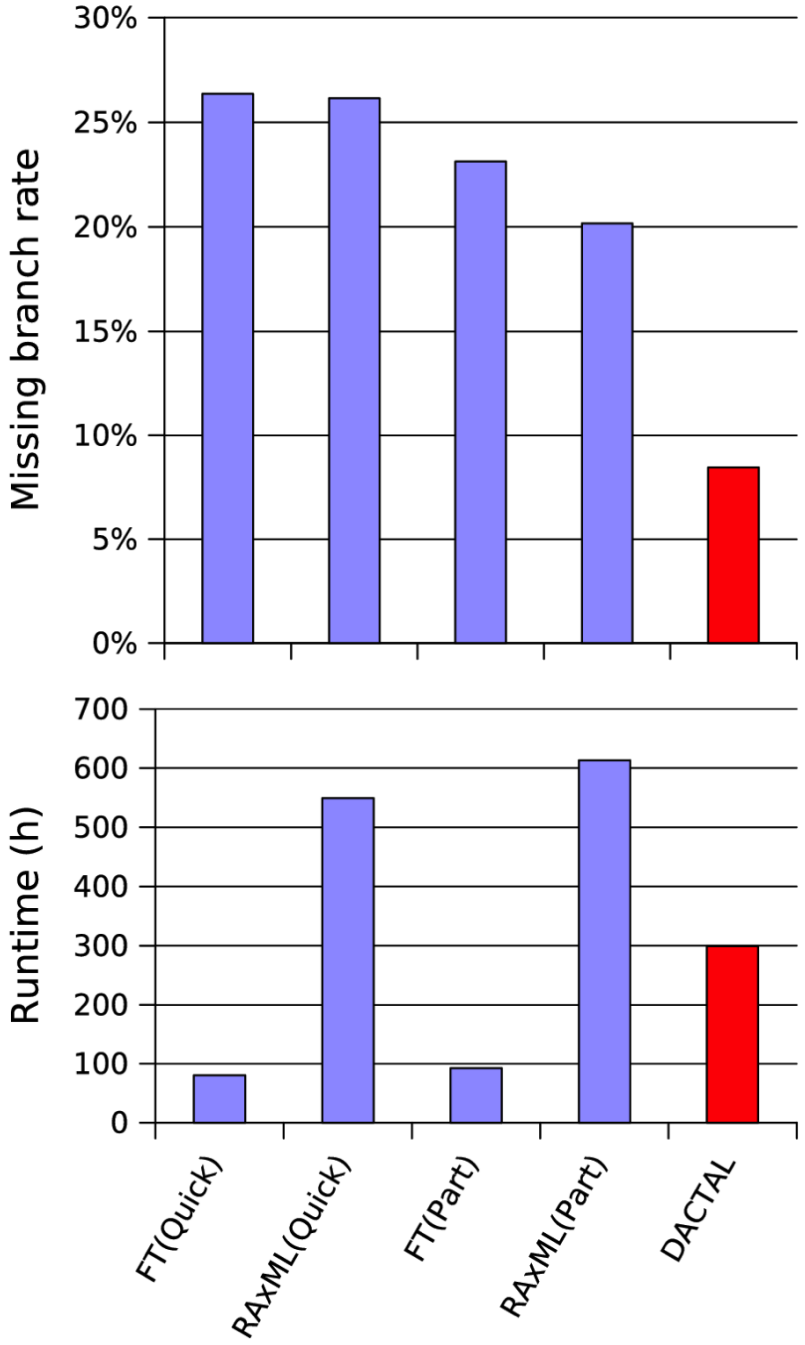
DACTAL


BLAST-based
Unaligned Sequences
Overlapping subsets
Existing Method: RAxML(MAFFT)
A tree for each subset
pRecDCM3
New supertree method:
SuperFine
A tree for the entire dataset

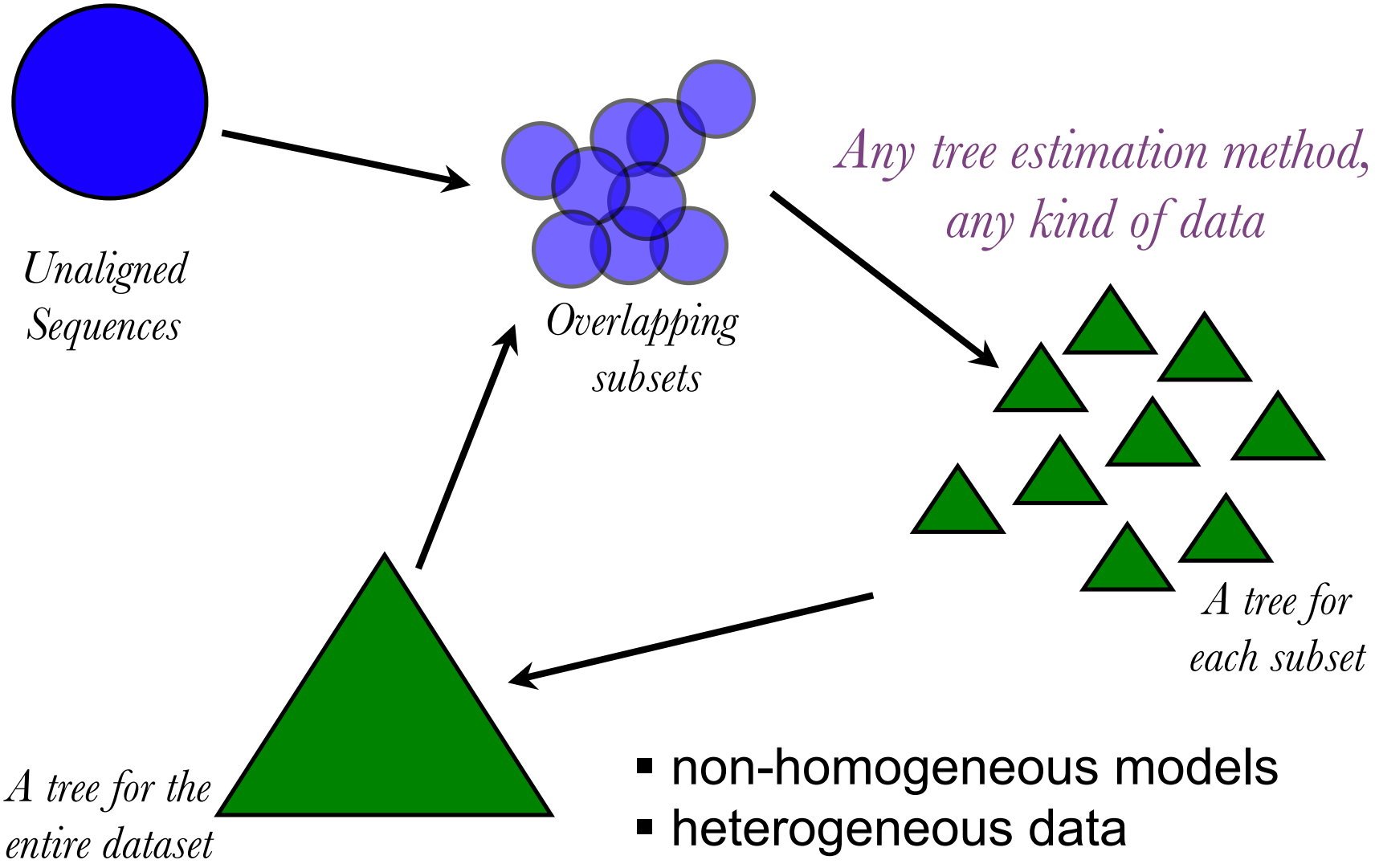
DACTAL: Better results than 2-phase methods

- 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



DACTAL is Flexible



- non-homogeneous models
- heterogeneous data

Part III: SEPP

- SEPP: SATé-enabled phylogenetic placement

- Mirarab, Nguyen, and Warnow. Pacific Symposium on Biocomputing, 2012.

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?

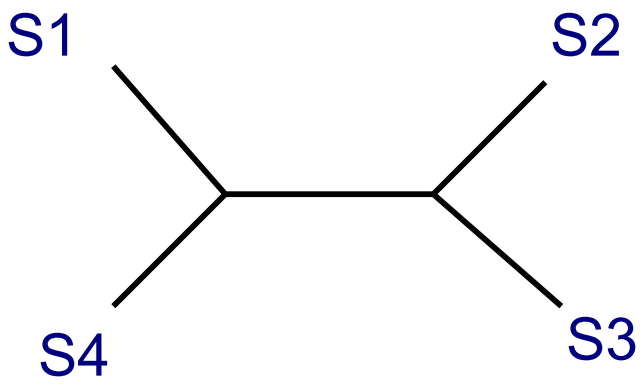
Phylogenetic Placement

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

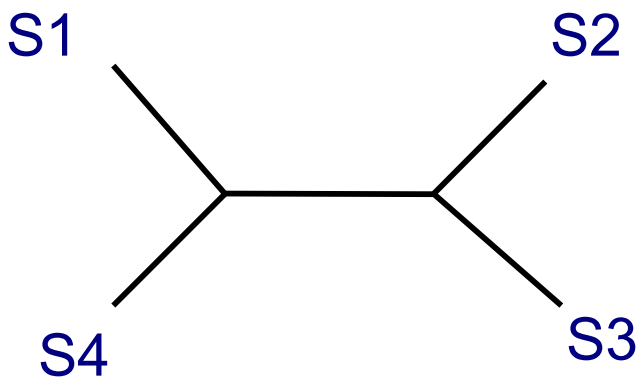
Align Sequence

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



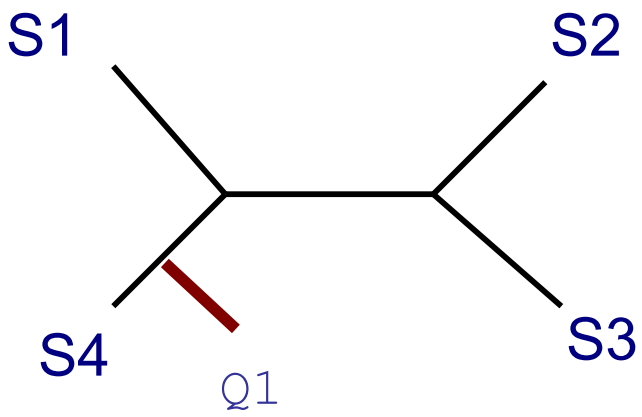
Align Sequence

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

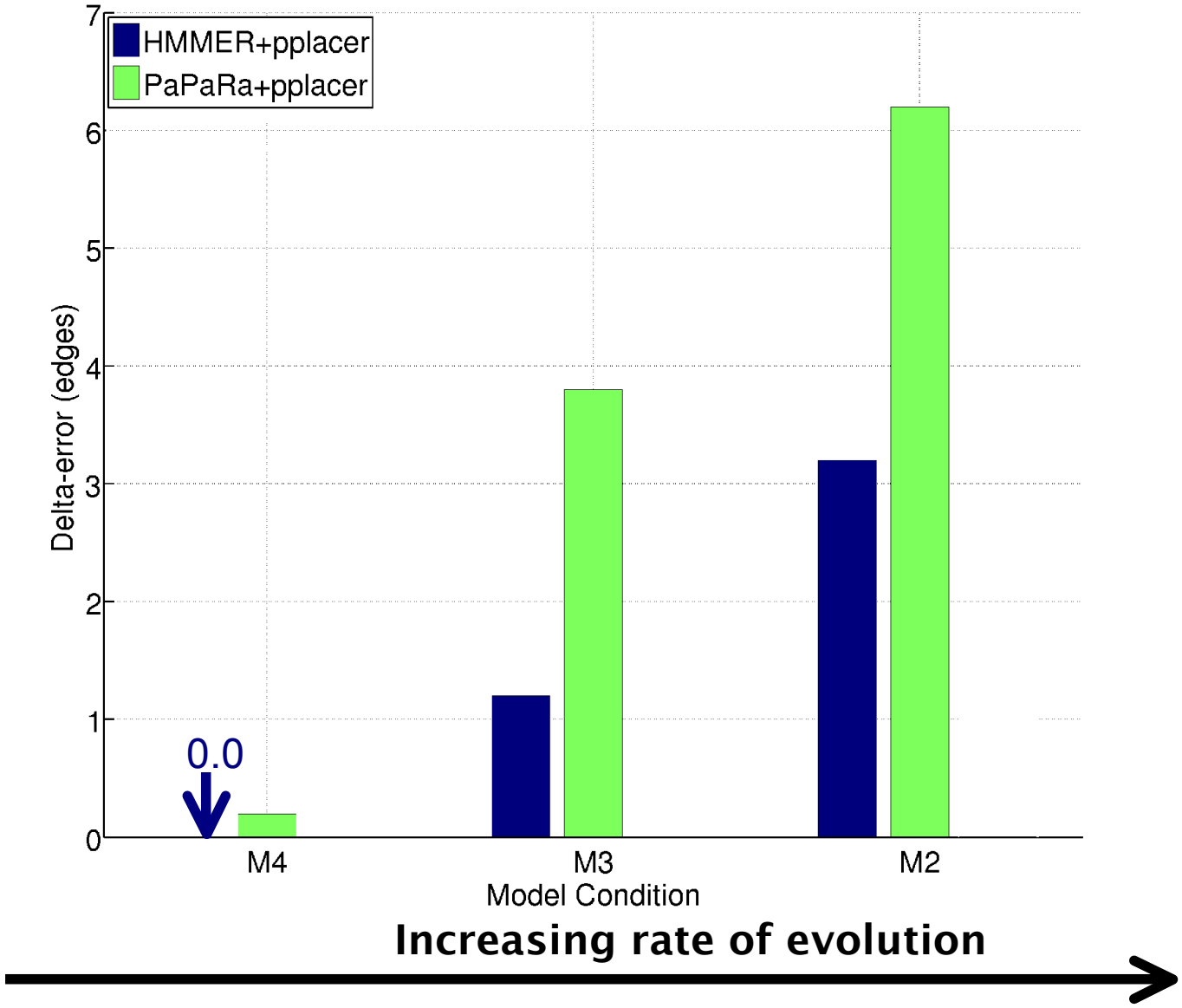


Phylogenetic Placement

- 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

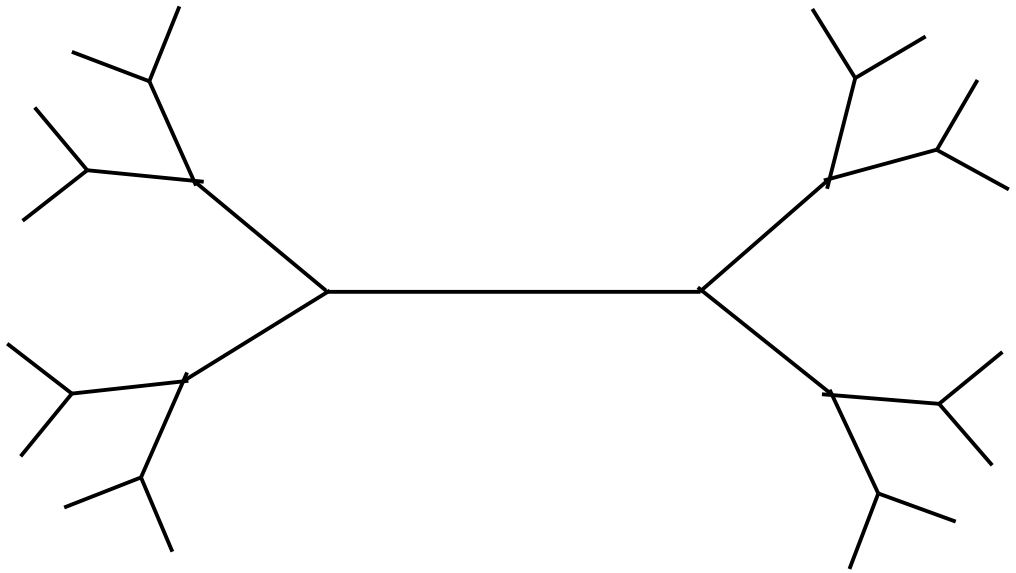
HMMER vs. PaPaRa



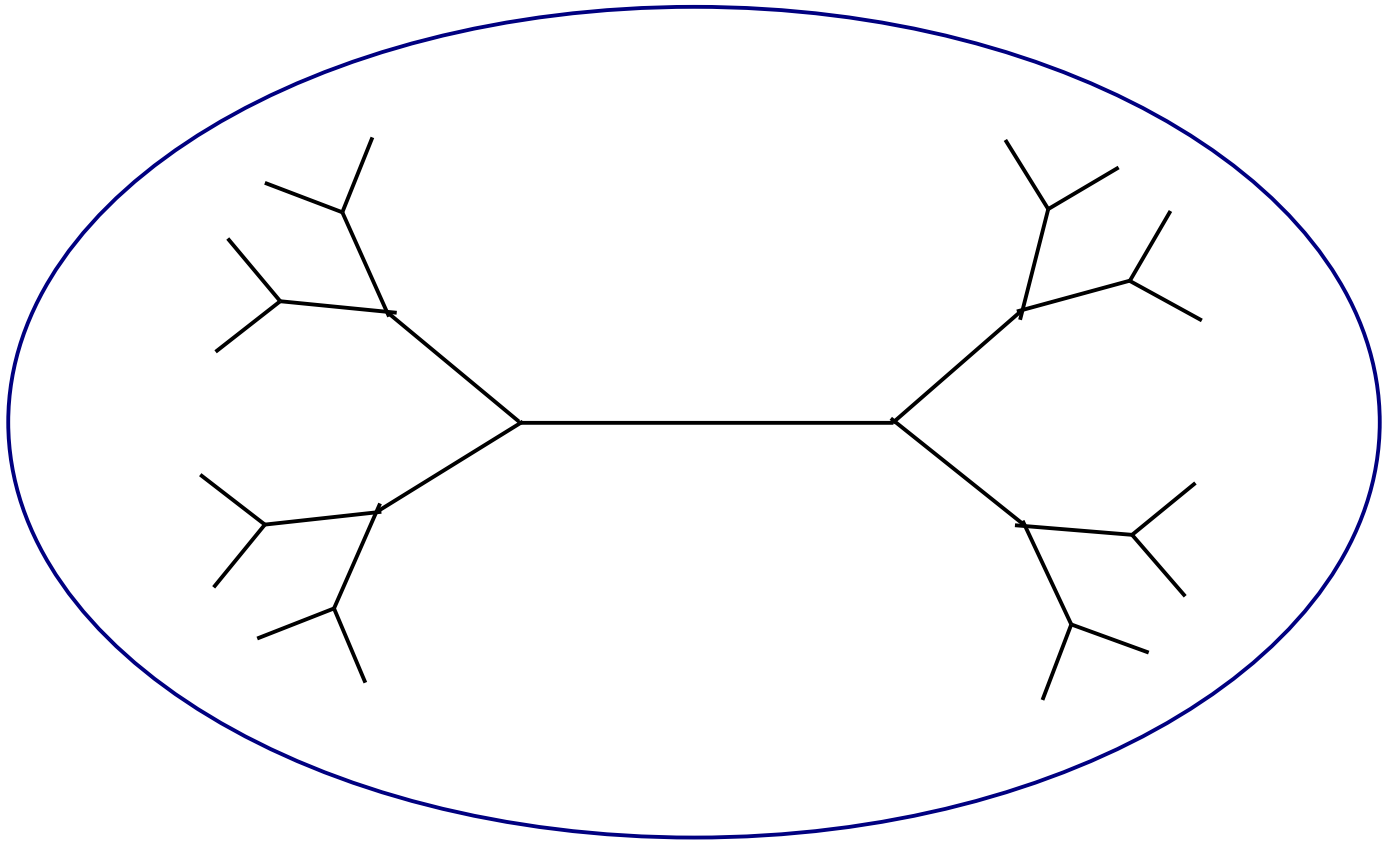
SEPP

- 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).

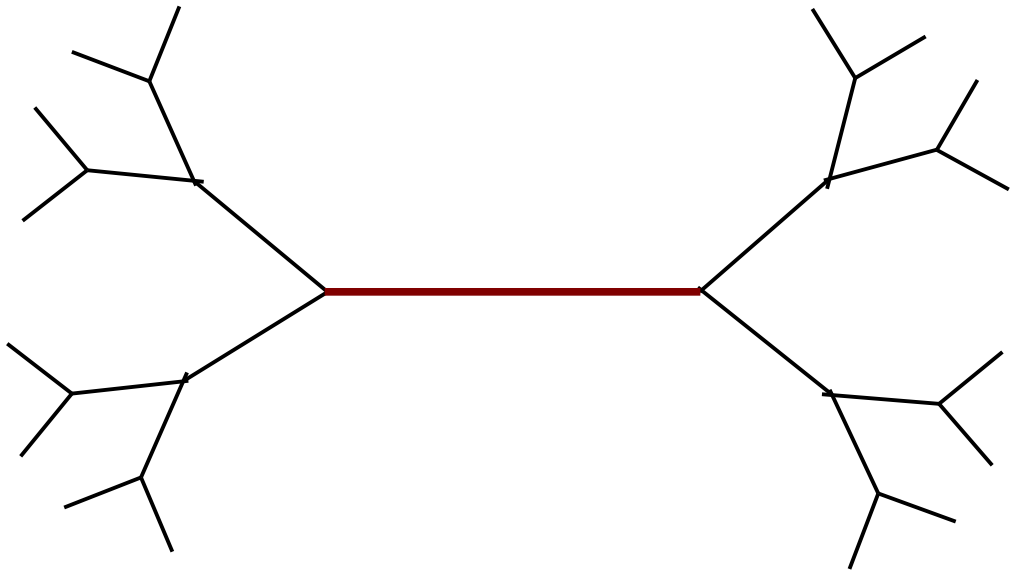
SEPP: SATé-enabled Phylogenetic Placement



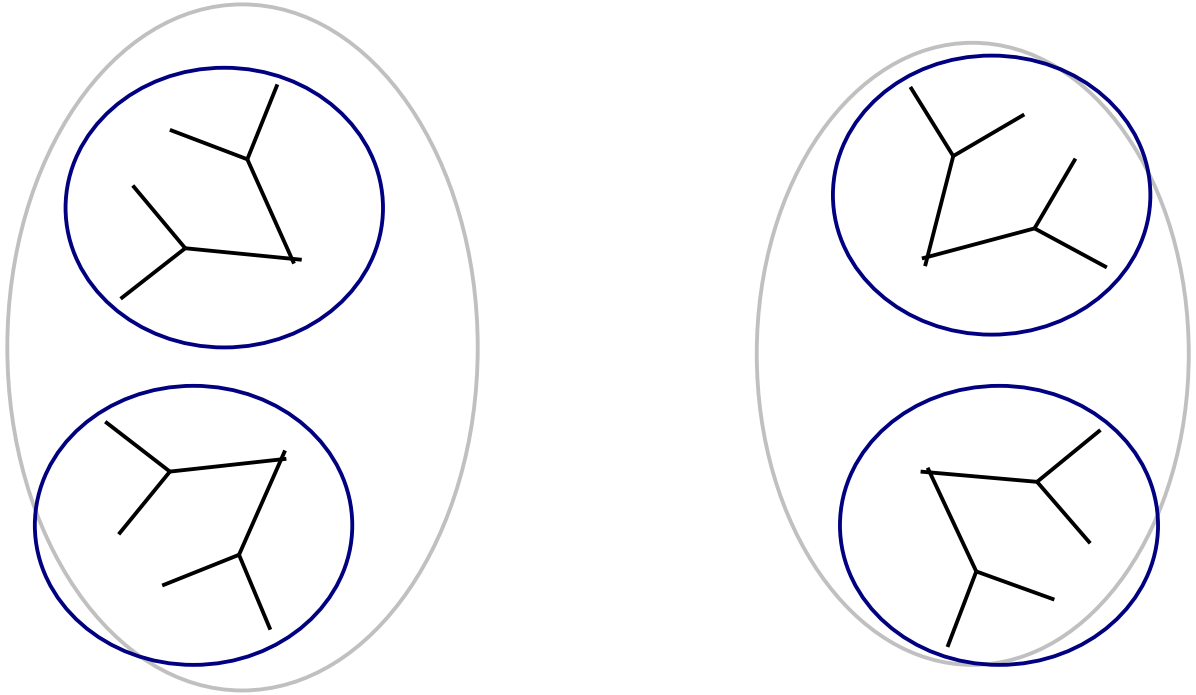
SEPP: SATé-enabled Phylogenetic Placement



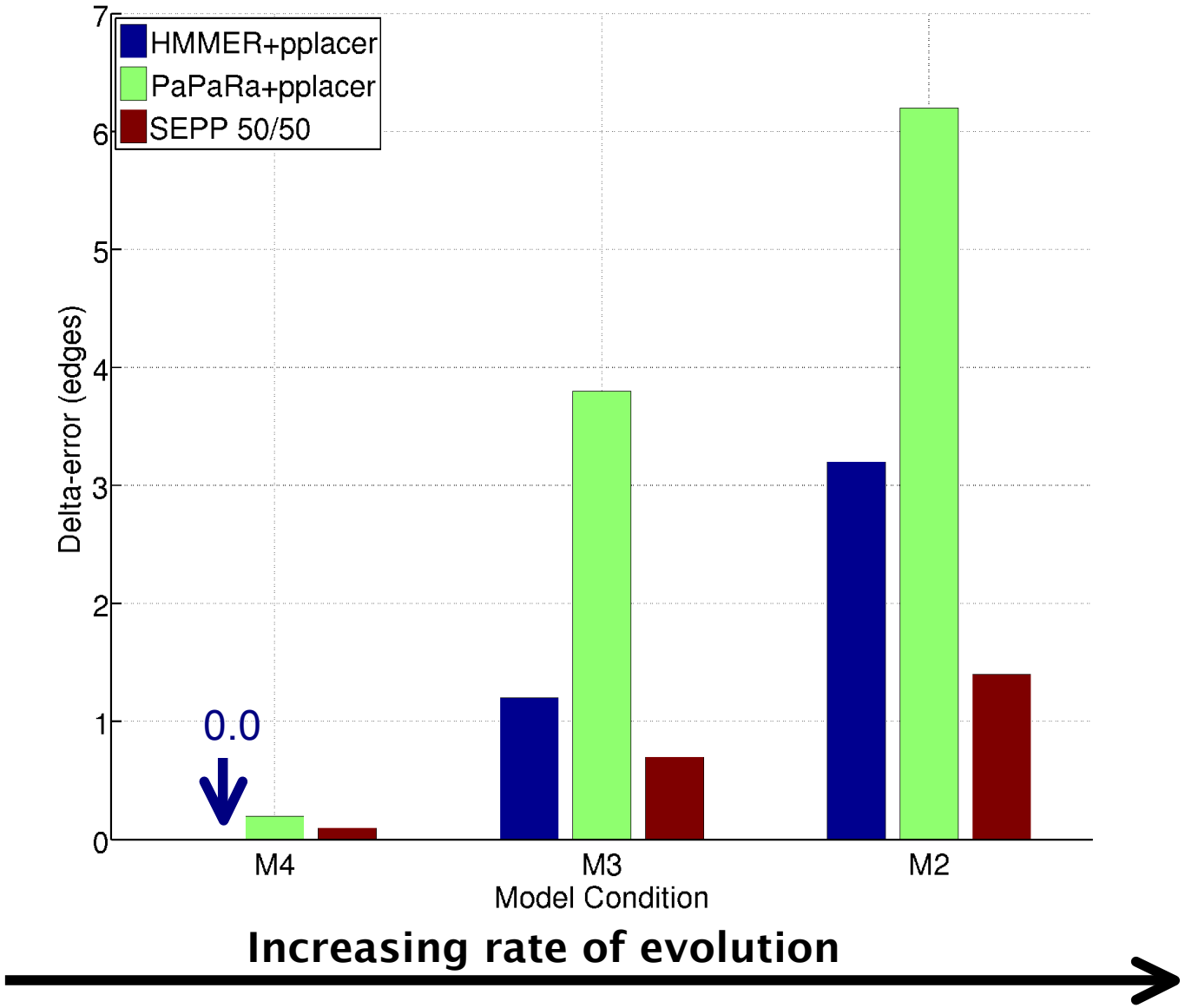
SEPP: SATé-enabled Phylogenetic Placement



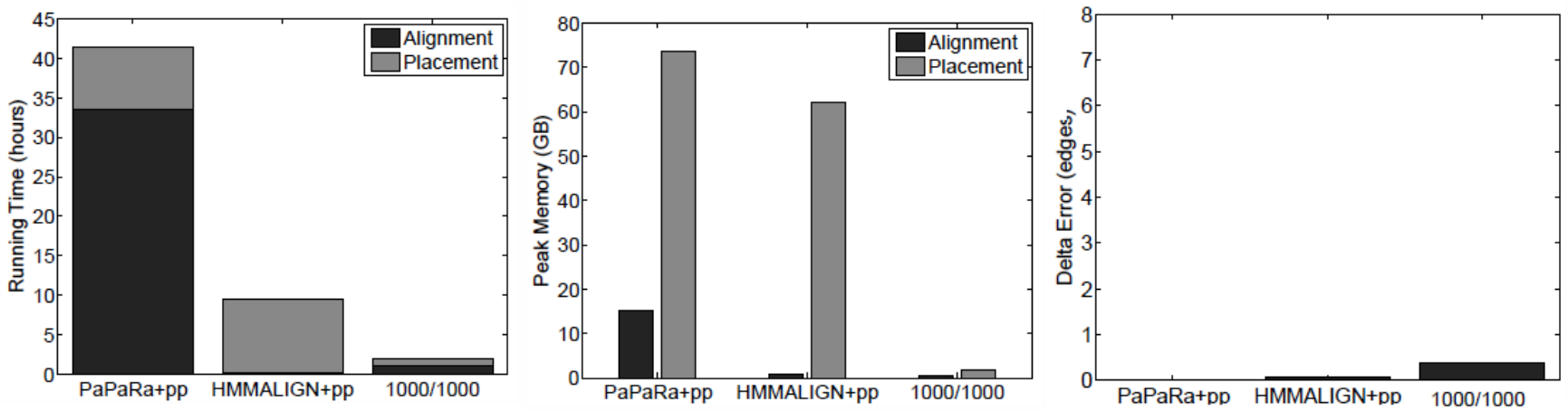
SEPP: SATé-enabled Phylogenetic Placement



SEPP (10%-rule) on Simulated Data

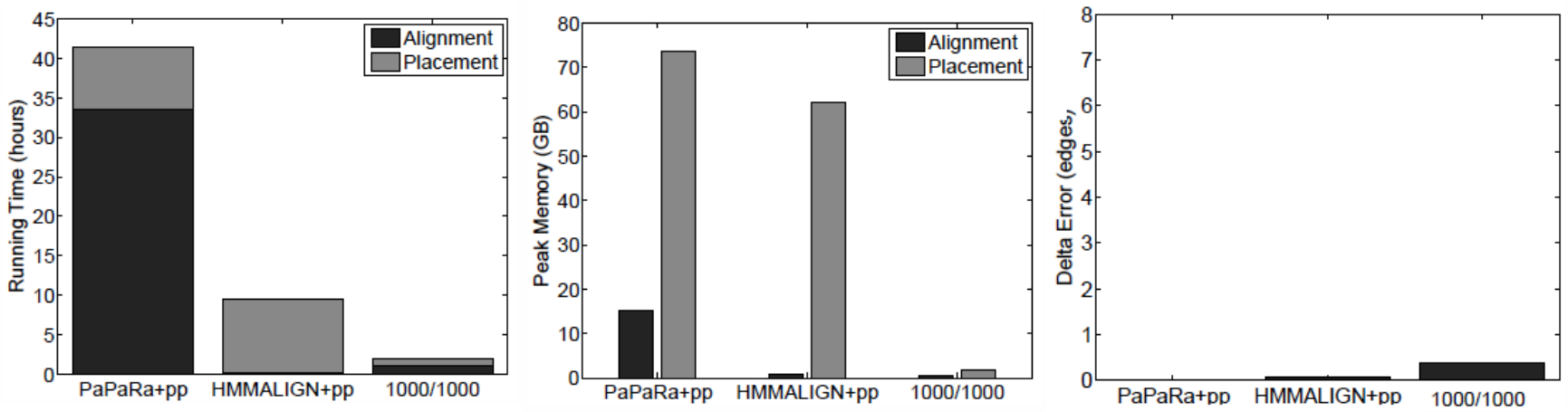


SEPP (10%) on Biological Data



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

SEPP (10%) on Biological Data



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

For 1 million fragments:

PaPaRa+pplacer: ~133 days

HMMALIGN+pplacer: ~30 days

SEPP 1000/1000: ~6 days

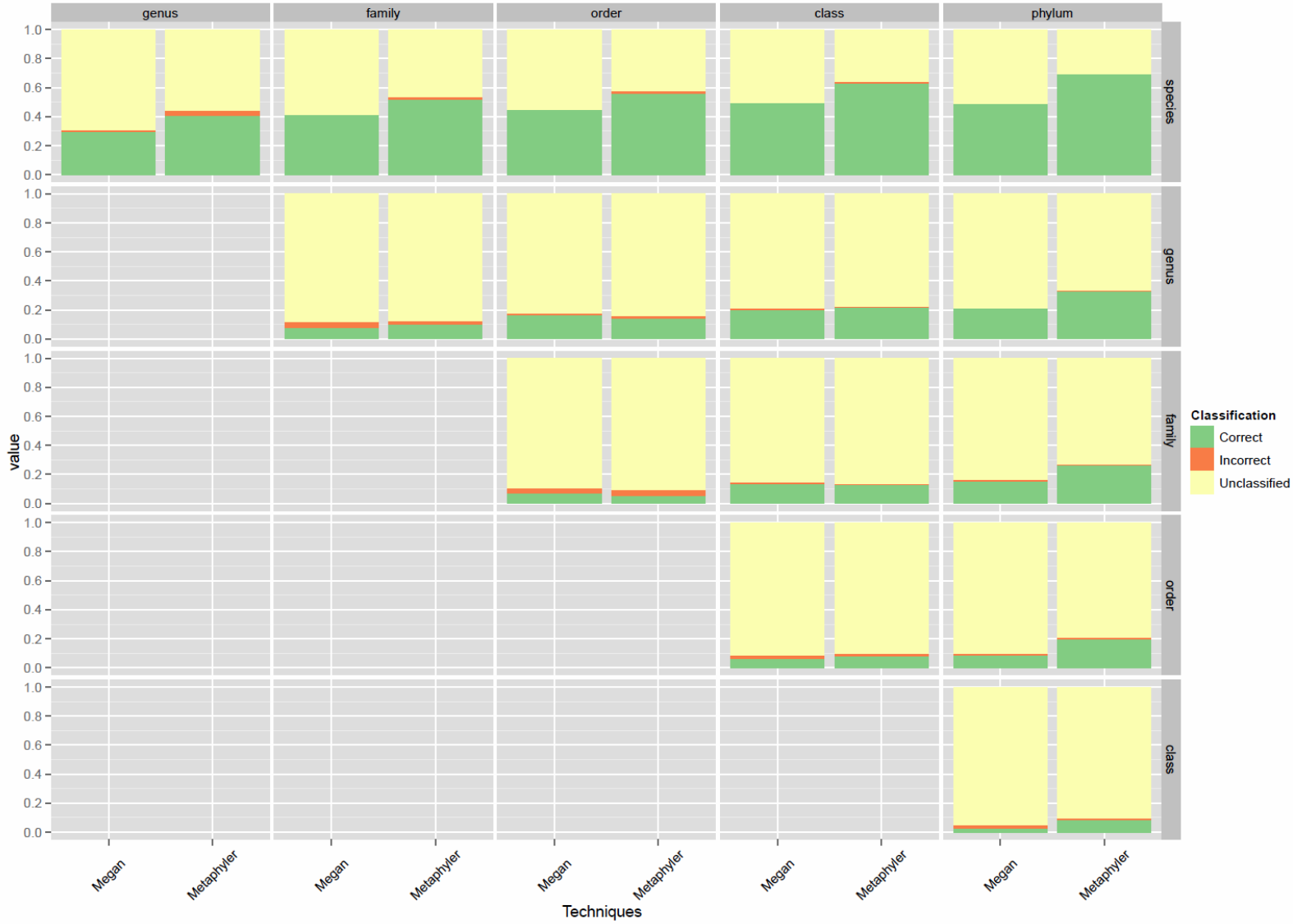
Taxon Identification

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: Megan and Metaphyler

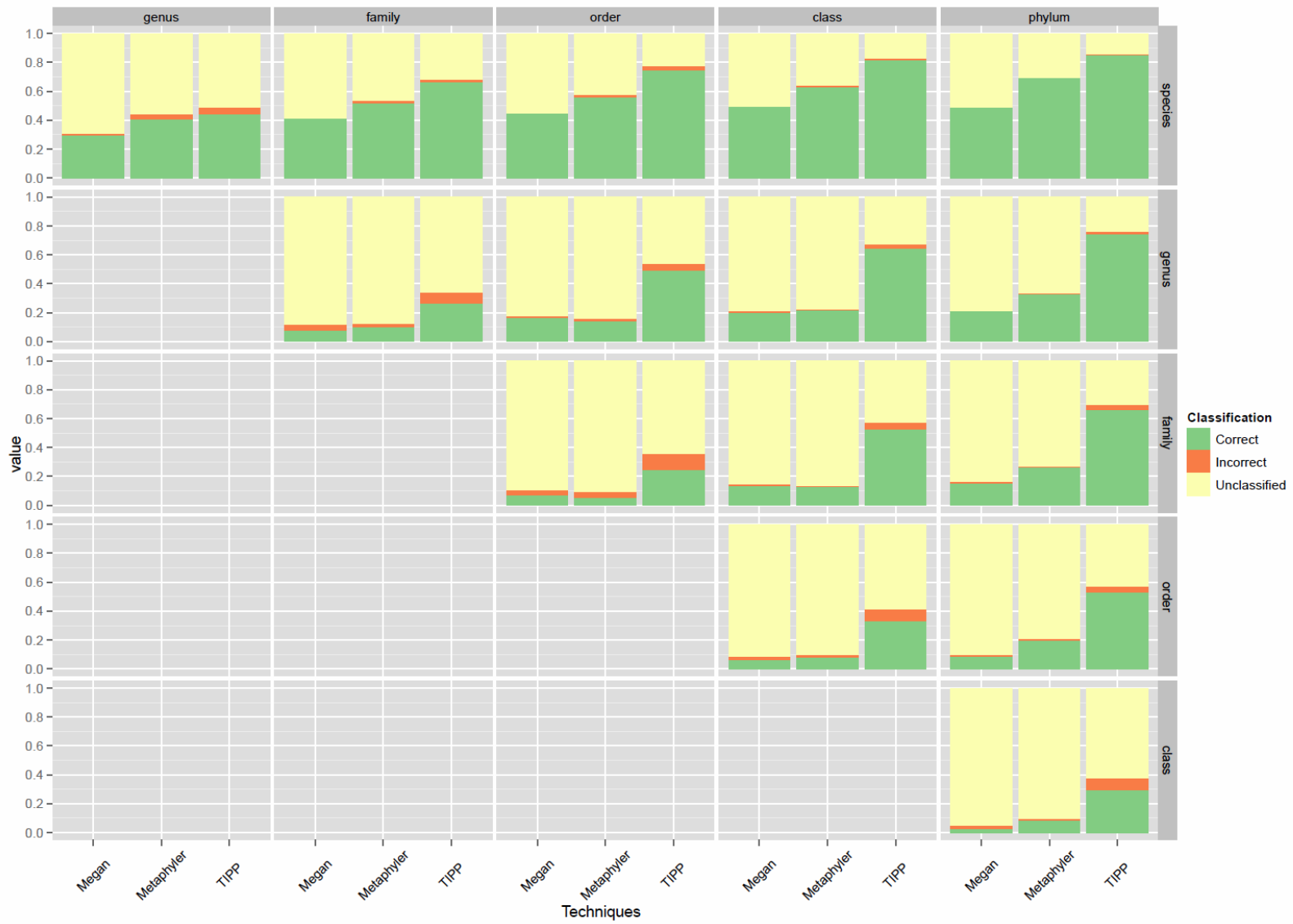


TIPP

- 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



Pipelines

- Site evolution model selection
- Alternative handling of multi-marker datasets

- Statistics on set of tree/alignment pairs
- Bootstrapping

- Visualization of trees and alignments
- Estimating ancestral sequences
- Estimating ancestral dates

Multi-gene analyses

After alignment of each gene dataset:

- Combined analysis: Concatenate (“combine”) alignments for different genes, and run phylogeny estimation methods

- Supertree: Compute trees on alignment and combine gene trees

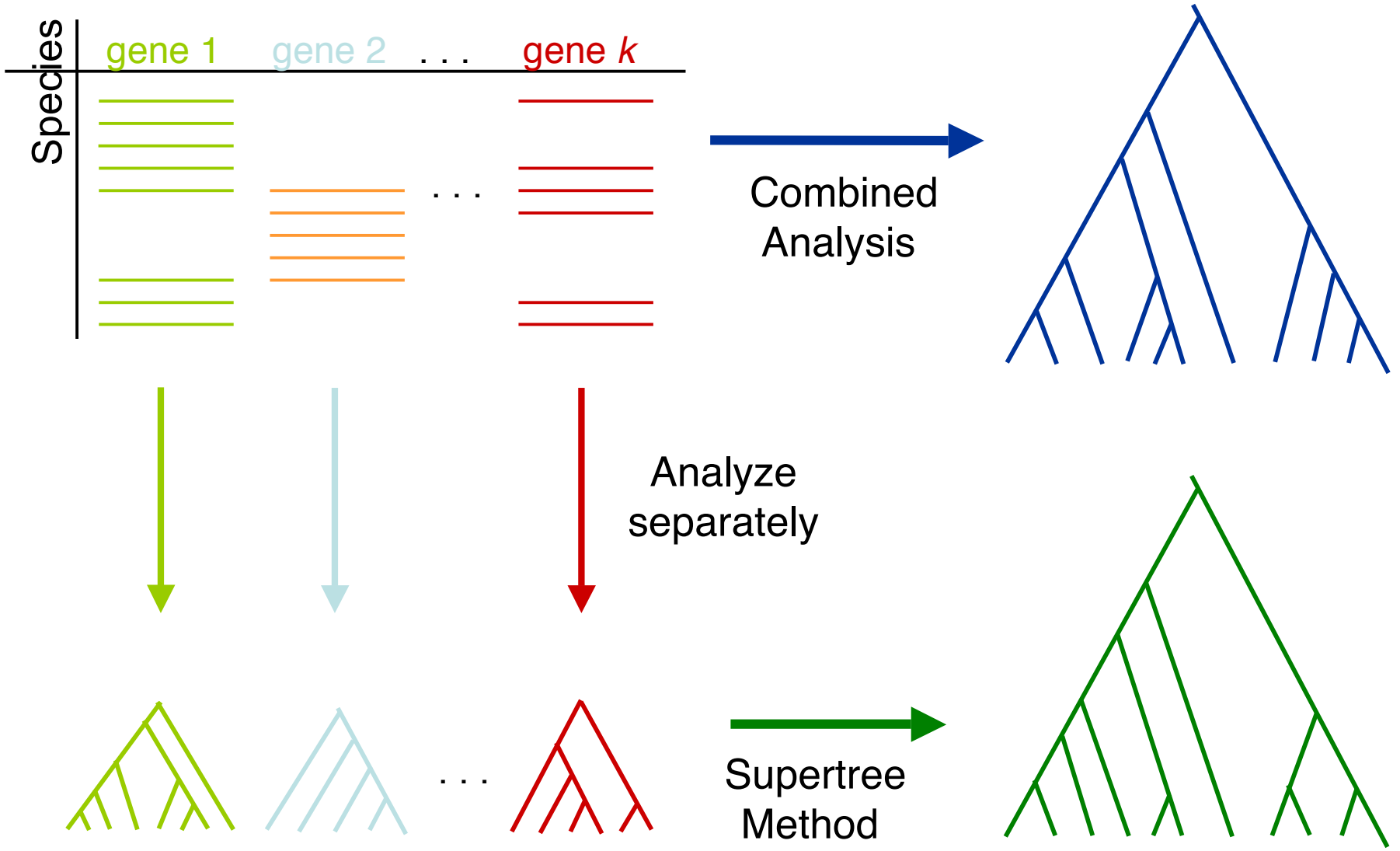
Not all genes present in all species

	gene 1
S_1	TCTAATGGAA
S_2	GCTAAGGGAA
S_3	TCTAAGGGAA
S_4	TCTAACGGAA
S_7	TCTAATGGAC
S_8	TATAACGGAA

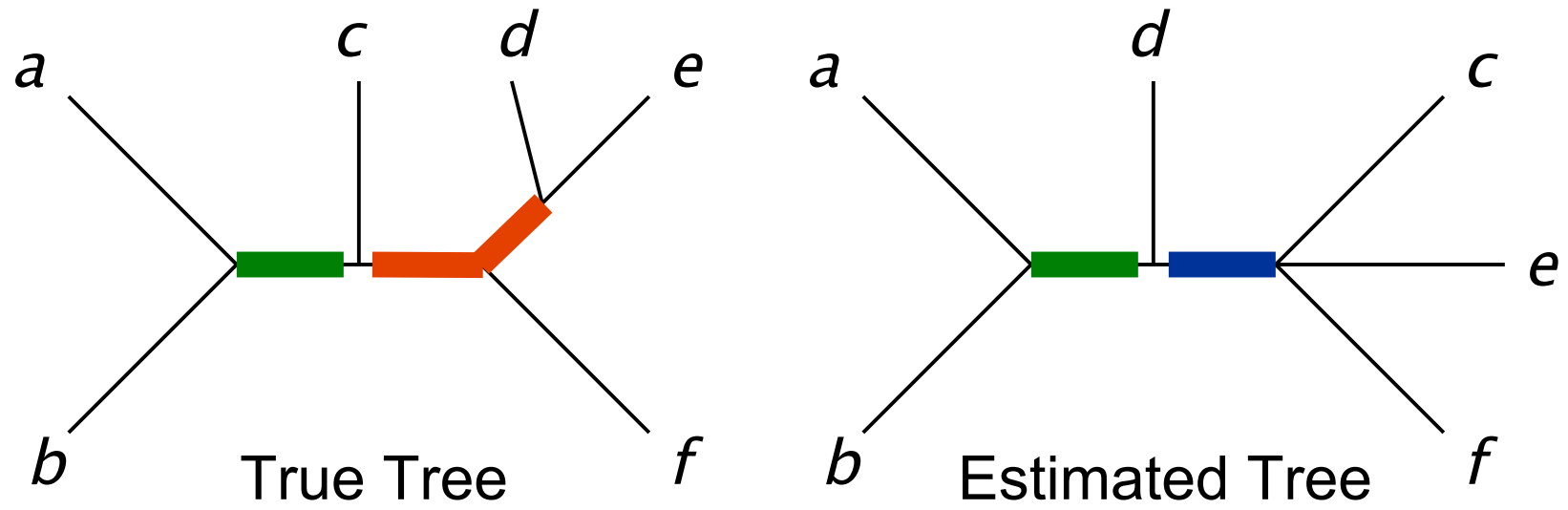
	gene 2
S_4	GGTAACCCTC
S_5	GCTAAACCTC
S_6	GGTGACCATC
S_7	GCTAAACCTC

	gene 3
S_1	TATTGATACA
S_3	TCTTGATACC
S_4	TAGTGATGCA
S_7	TAGTGATGCA
S_8	CATTCATACC

Two competing approaches



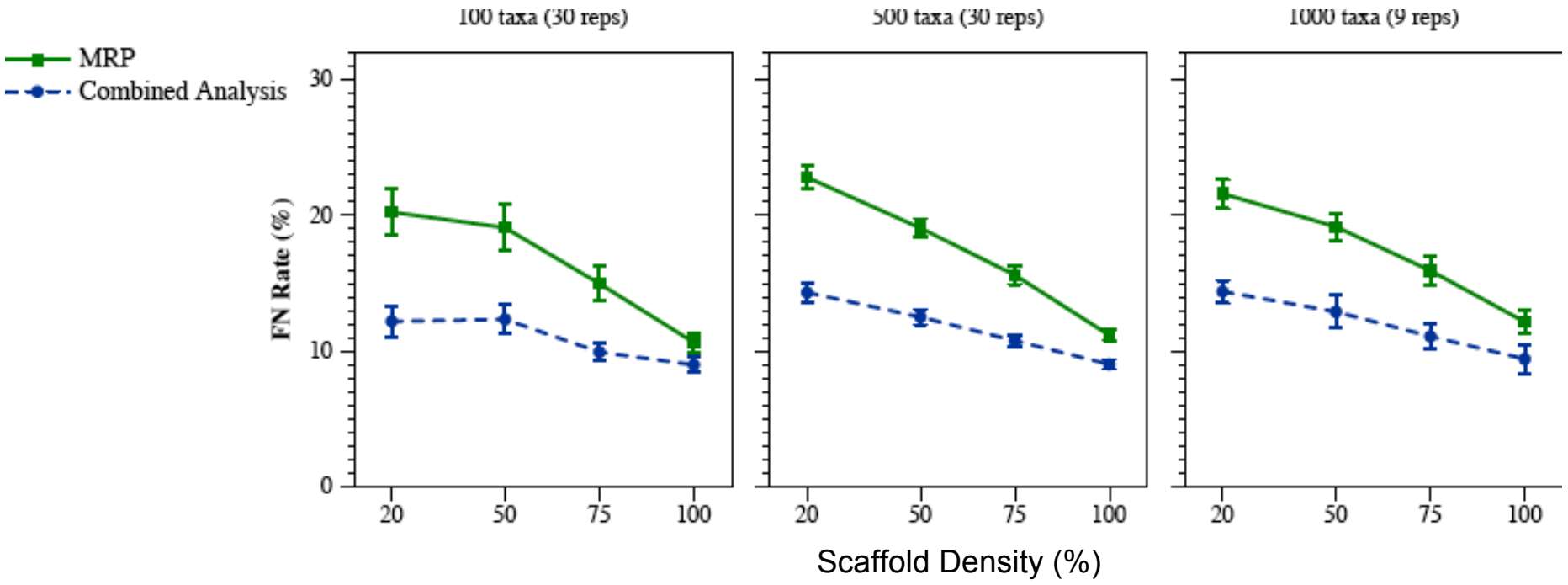
Quantifying topological error



- False negative (FN): $b \in B(T_{true})\text{-}B(T_{est.})$

- False positive (FP): $b \in B(T_{est.})\text{-}B(T_{true})$

FN rate of MRP vs. combined analysis



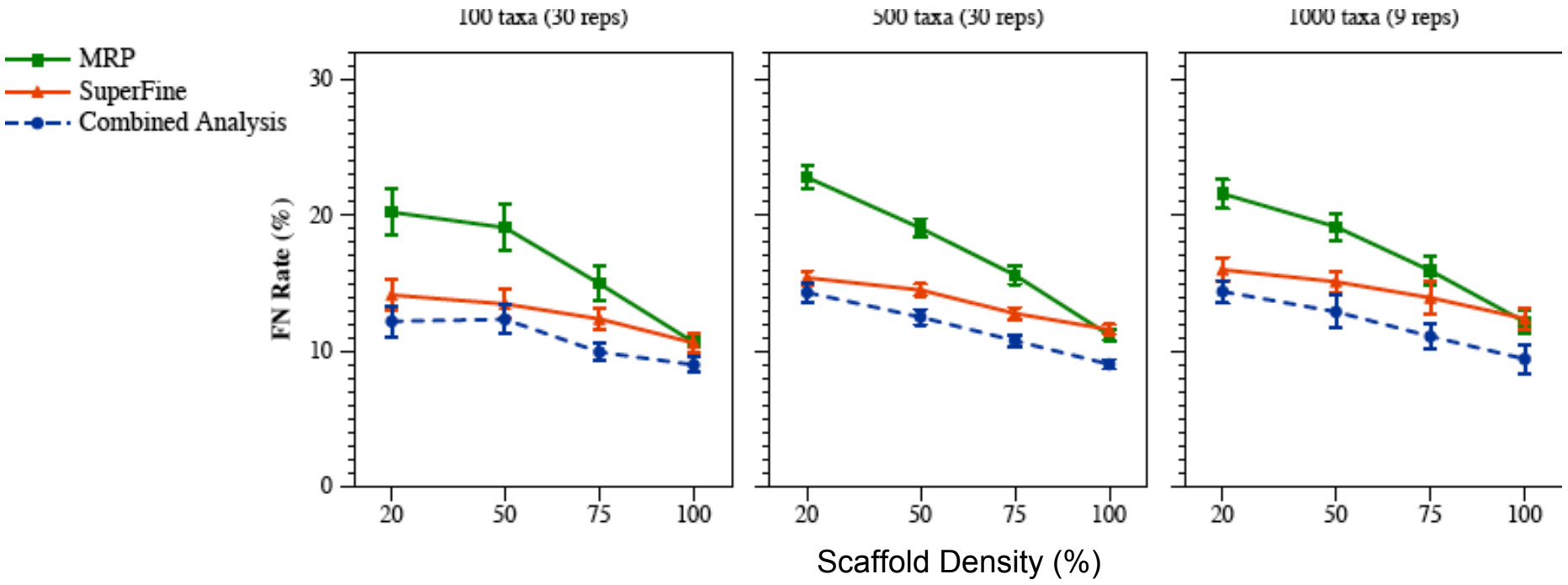
SuperFine: new supertree method

- Step 1: construct a supertree with low false positives (*unresolved*)
- Step 2: *Refine the tree* to reduce false negatives by resolving each high degree node ("polytomy") using a "base" supertree method (e.g., MRP) applied to *recoded source trees.*

Swenson, Suri, Linder, and Warnow, Systematic Biology, 2012.

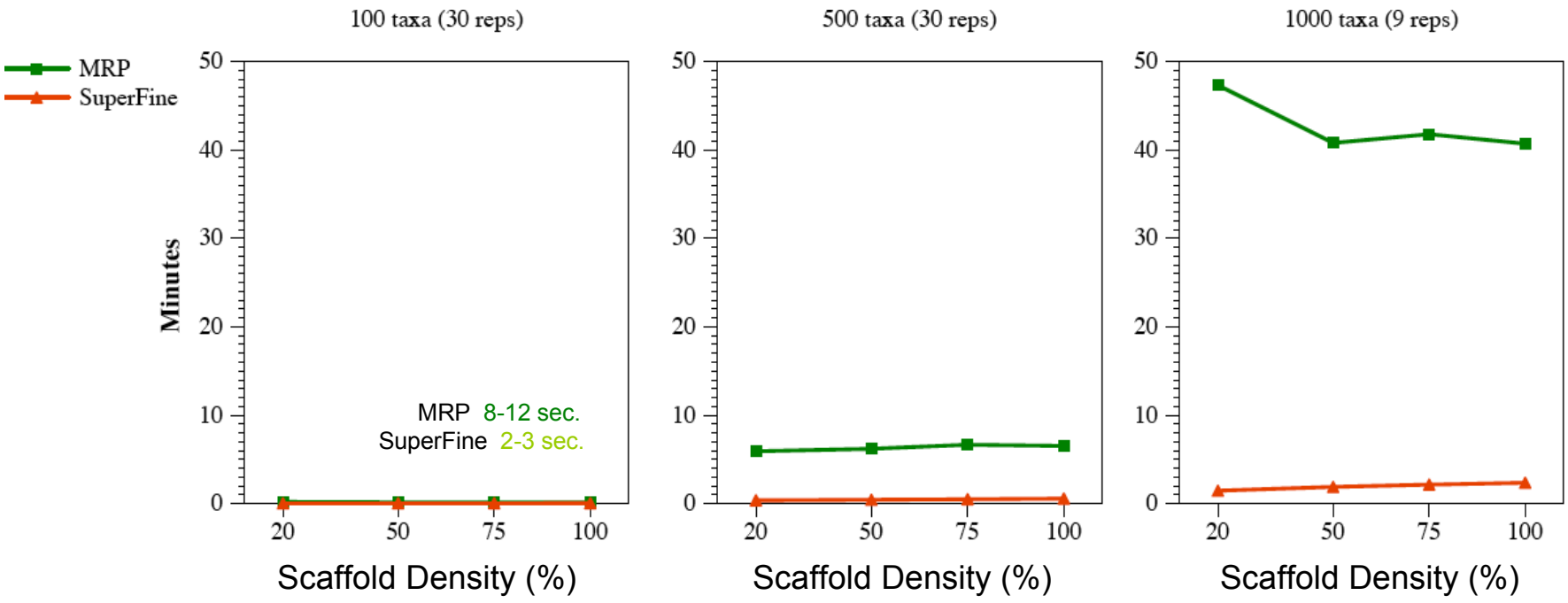
See also Mirarab, Nguyen, and Warnow, J Alg. Molec. Biology.

SuperFine: most accurate supertree method, and very fast



(Swenson et al., Syst. Biol. 2012)

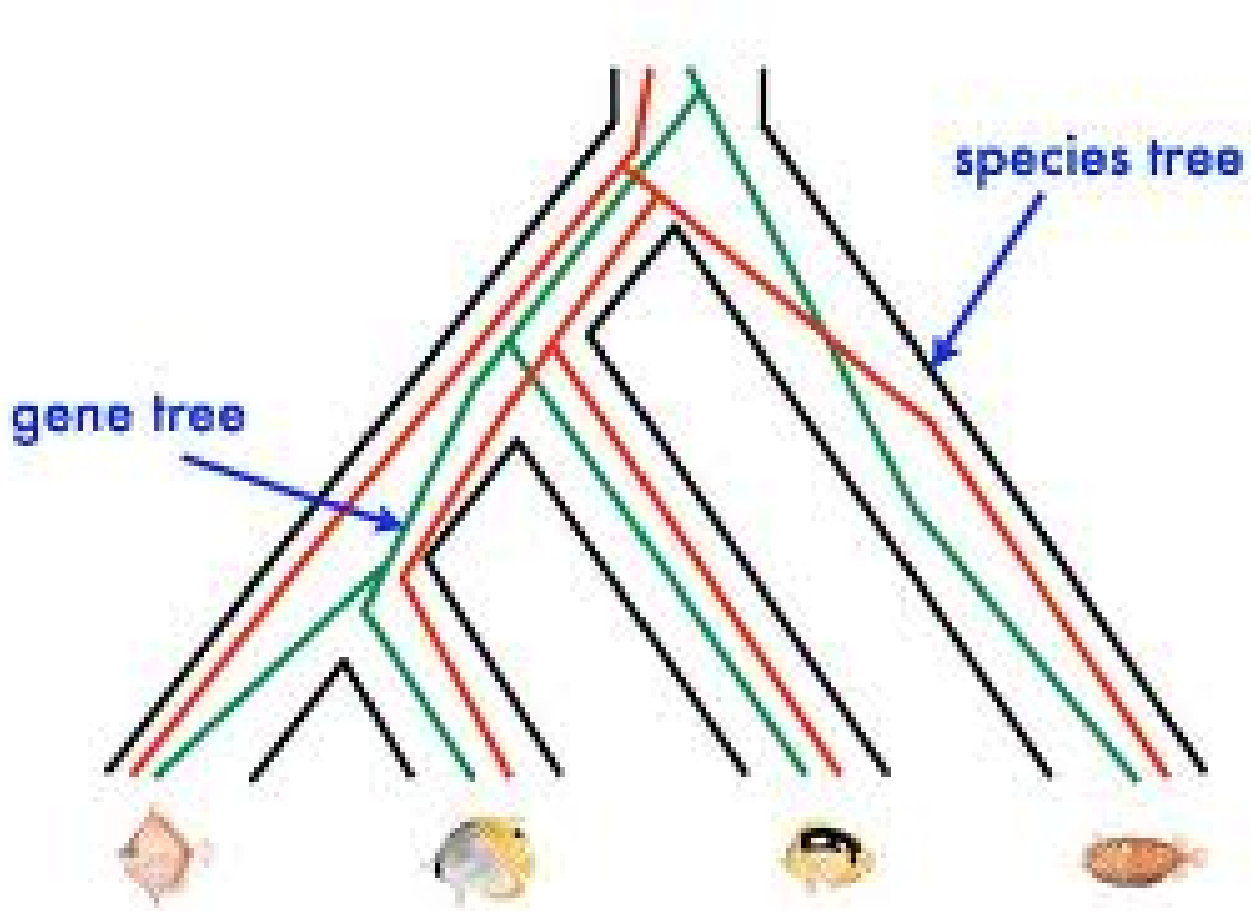
SuperFine is also much faster than combined analysis and leading supertree methods



Limitations

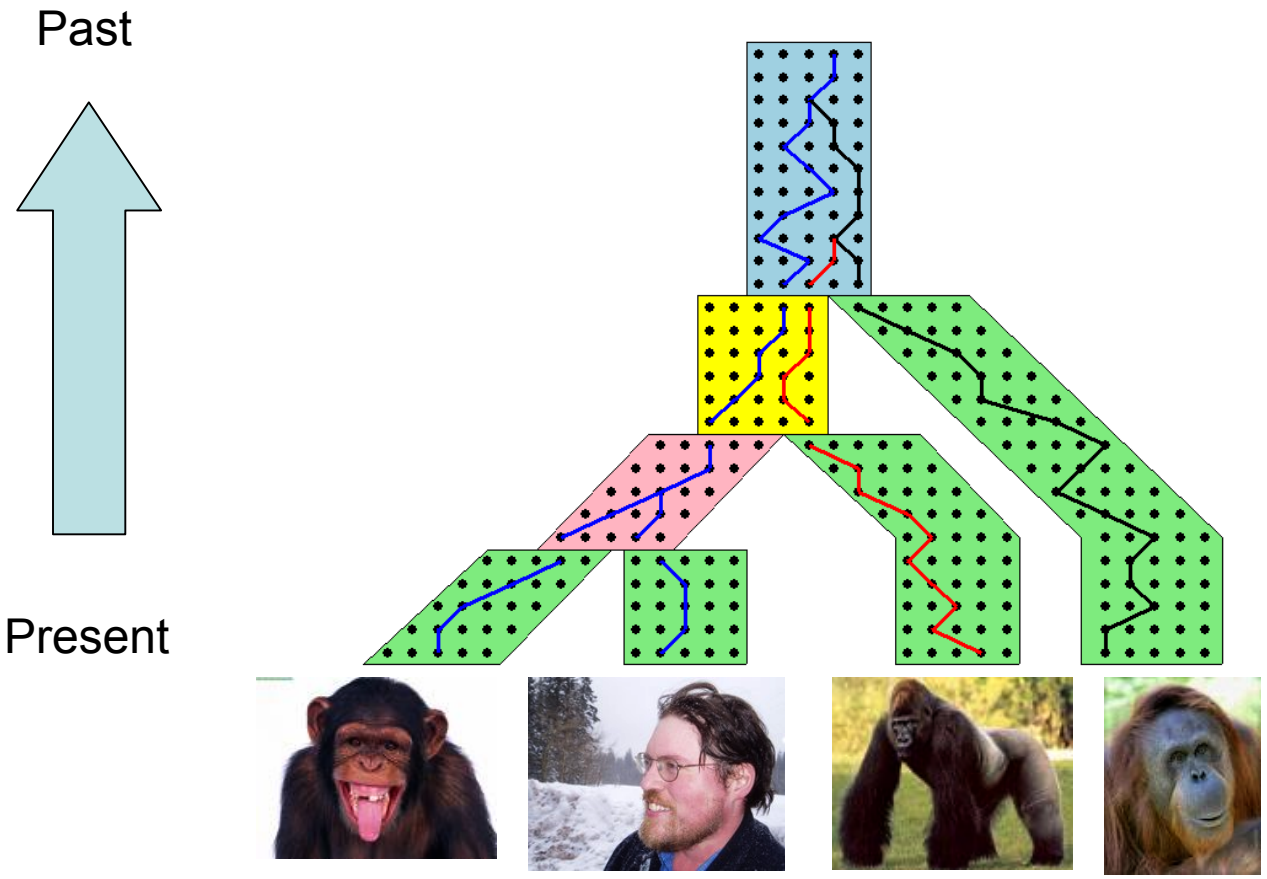
- All these methods assume that the gene trees match the species tree.
- This is known to be unrealistic in some situations, due to processes such as
  - Deep Coalescence
  - Gene duplication and loss
  - Horizontal gene transfer

Red gene tree ≠ species tree (green gene tree okay)



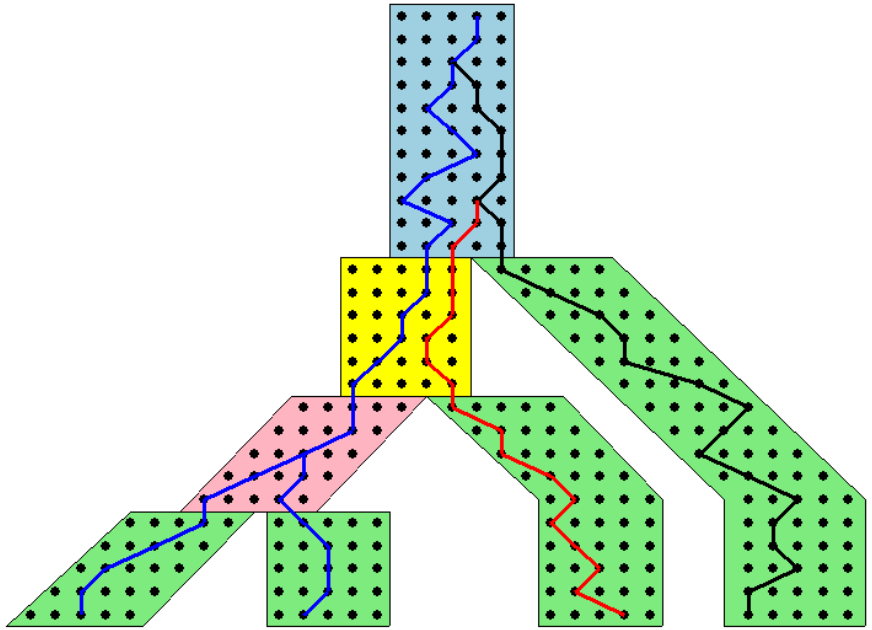
Multiple populations/species

Courtesy James Degnan



Gene tree in a species tree

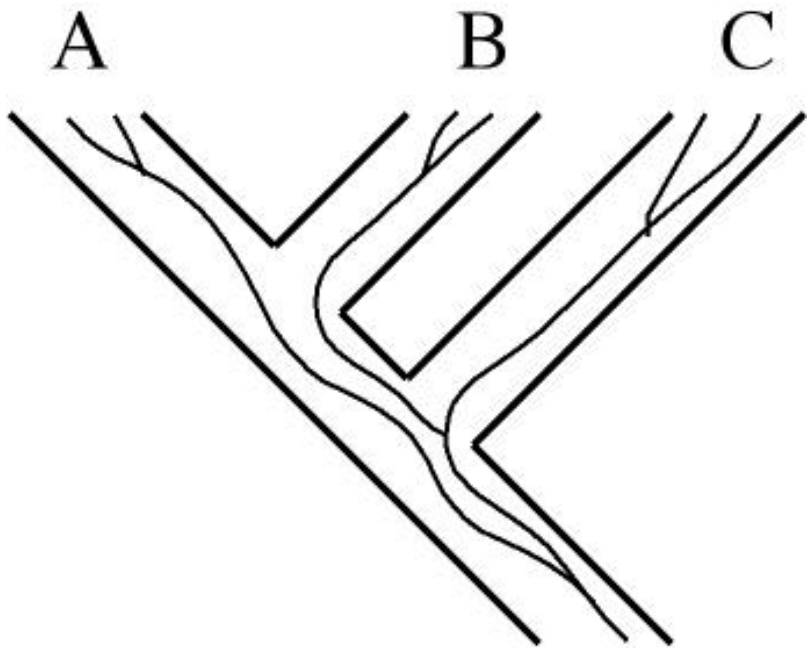
Courtesy James Degnan



Deep Coalescence

- Population-level process
- Gene trees can differ from species trees due to short times between speciation events (population size also impacts this probability)
- MDC (minimize deep coalescence) problem:
  - given set of true gene trees, find the species tree that implies the *fewest deep coalescence events* (Wayne Maddison)

Counting deep coalescences



Limitations

All current methods assume that input gene trees must be correct, binary, rooted trees

Most methods require that all taxa appear in all gene trees

Many methods are extremely expensive

But

- Estimated gene trees are usually partially incorrect and are often unrooted.
- Not all gene trees have all the taxa

Algorithms for ILS (Minimizing Deep Coalescence)

- Phylonet-MDC: software developed by Nakhleh (Rice University). Uses algorithms developed by the Warnow Lab (Bayzid and Warnow), handling incompletely resolved unrooted gene trees, which can also be incomplete.
- iGTP-MDC
- *BEAST: co-estimation of gene trees and species trees (Heled and Drummond)
- BUCKy (Cecile Ané, Bret Larget, and others)
- Consensus methods
- Supertree methods
- Concatenated analysis

And others!

Some Estimation Challenges

- Large datasets
- Long sequences
- Model violations
- Fragmentary sequences
- Estimation of species trees (ILS, duplication/loss, HGT)
- Rearrangements (duplications, inversions, transpositions)

Questionnaire

- Please fill out the two questionnaires (leave on table)
- We will use your feedback to inform our software development!
- We also welcome collaborations with you on your hardest datasets:
    - **DACTAL or SEPP for very large datasets**
    - **SEPP for fragmentary sequences**
    - **DACTAL for datasets with model violations**

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:
  - SATé: Randy Linder, Kevin Liu, Serita Nelesen, Sindhu Raghavan, and Li-San Wang
  - DACTAL: Serita Nelesen, Kevin Liu, Li-San Wang, and Randy Linder
  - SEPP: Siavash Mirarab and Nam Nguyen