

# New HMM-based methods for Ultra-large Alignment and Phylogeny Estimation

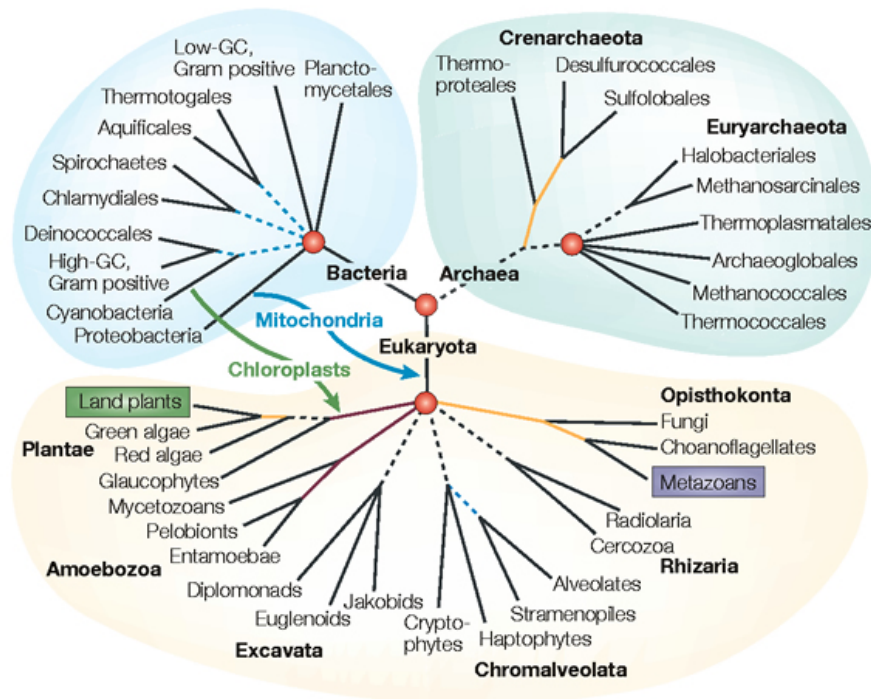
Tandy Warnow

Departments of Bioengineering and  
Computer Science

The University of Illinois at Urbana-Champaign

<http://tandy.cs.illinois.edu>

# Phylogenies and Applications



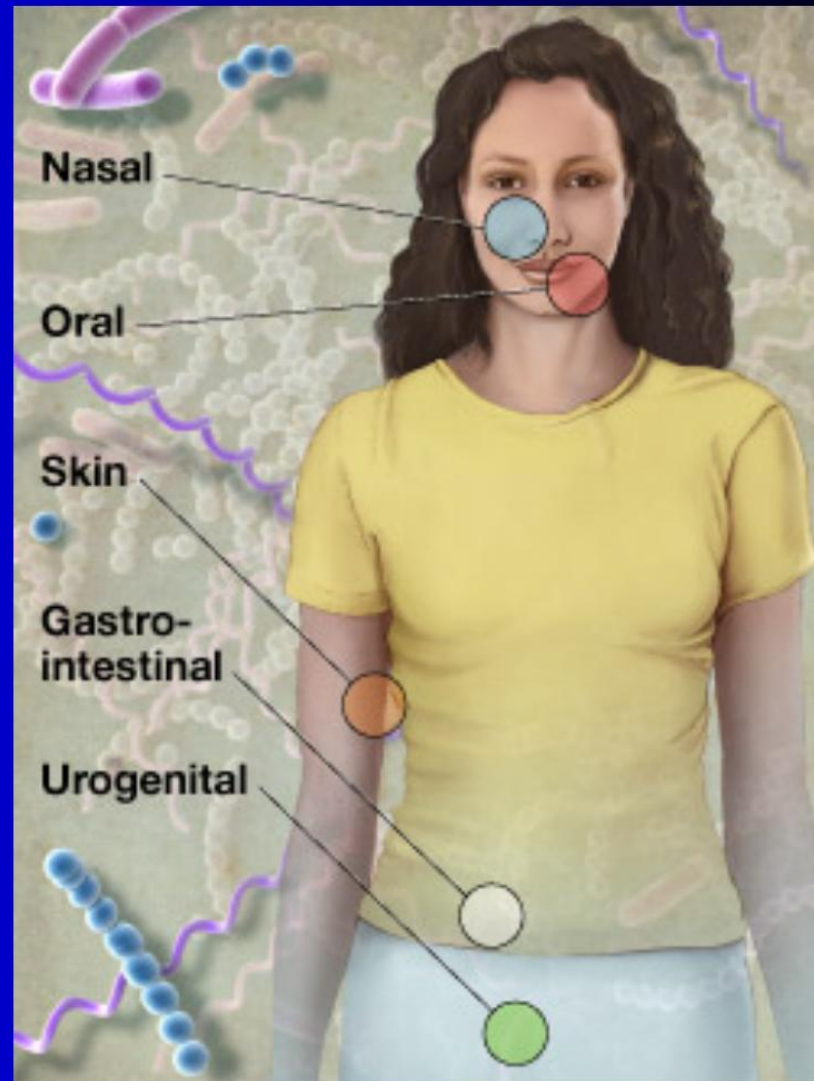
Basic Biology:

How did life evolve?

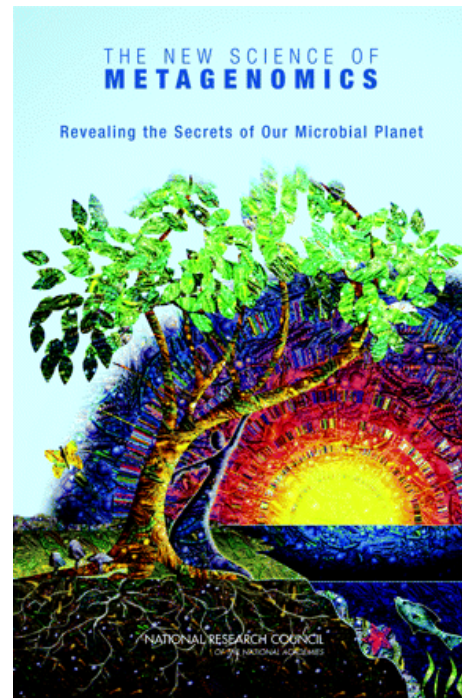
Applications of phylogenies to:  
protein structure and function  
population genetics  
human migrations

# The NIH Human Microbiome Project

25,000 human genes,  
1,000,000 bacterial genes



# Computational Phylogenetics and Metagenomics



Courtesy of the Tree of Life project

# Multiple Sequence Alignment (MSA): *another grand challenge*<sup>1</sup>

S1 = AGGCTATCACCTGACCTCCA	S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC	S2 = TAG-CTATCAC--GACCGC--
S3 = TAGCTGACCGC	S3 = TAG-CT-----GACCGC--
...	...
S <sub>n</sub> = TCACGACCGACA	→ S <sub>n</sub> = -----TCAC--GACCGACA

*Novel techniques needed for scalability and accuracy*

NP-hard problems and large datasets

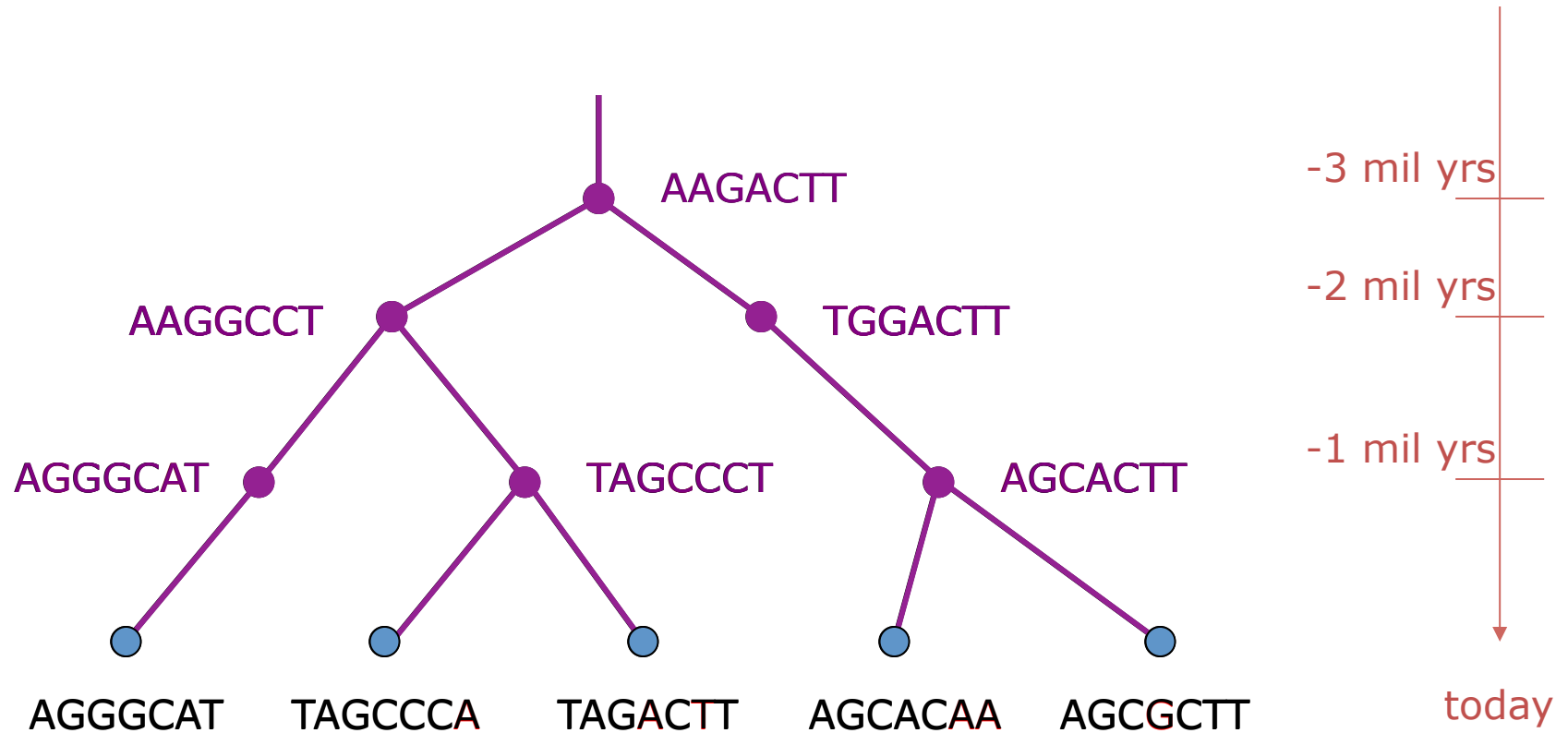
Current methods do not provide good accuracy

Few methods can analyze even moderately large datasets

*Many important applications besides phylogenetic estimation*

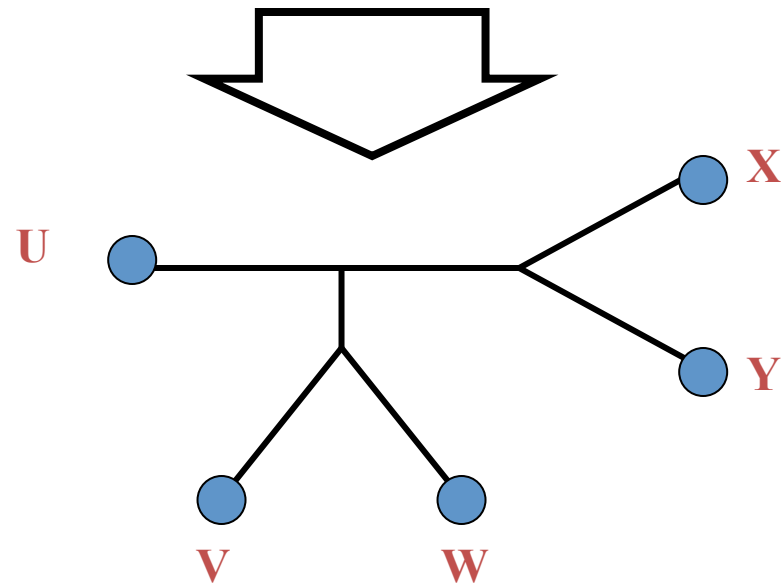
<sup>1</sup> Frontiers in Massive Data Analysis, National Academies Press, 2013

# DNA Sequence Evolution

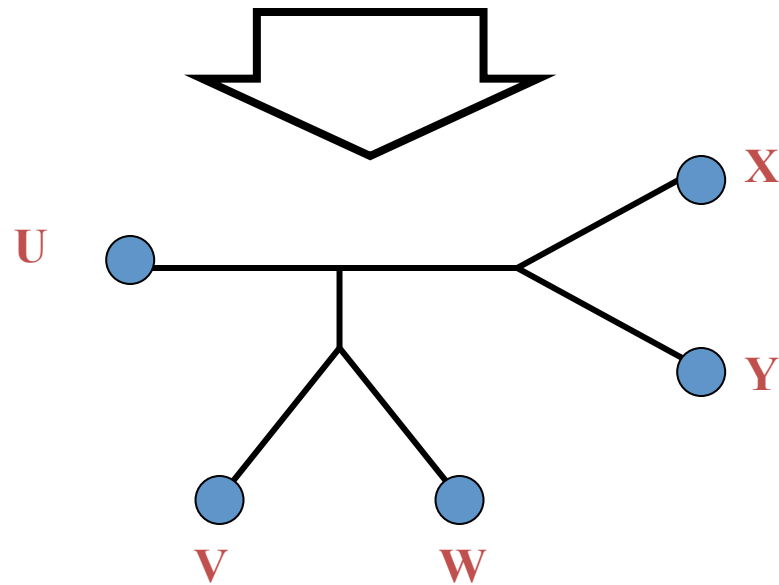
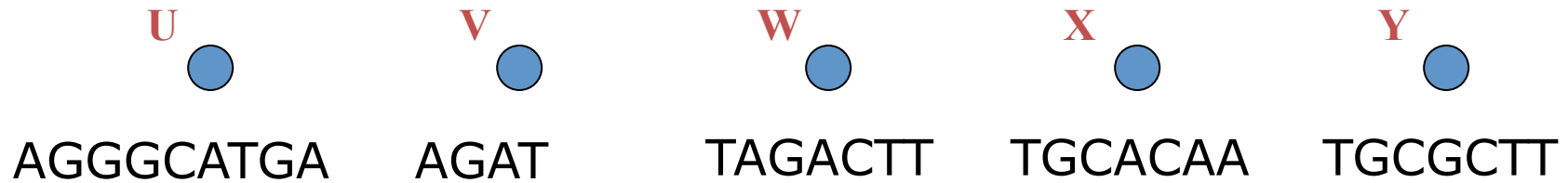


# Phylogeny Problem

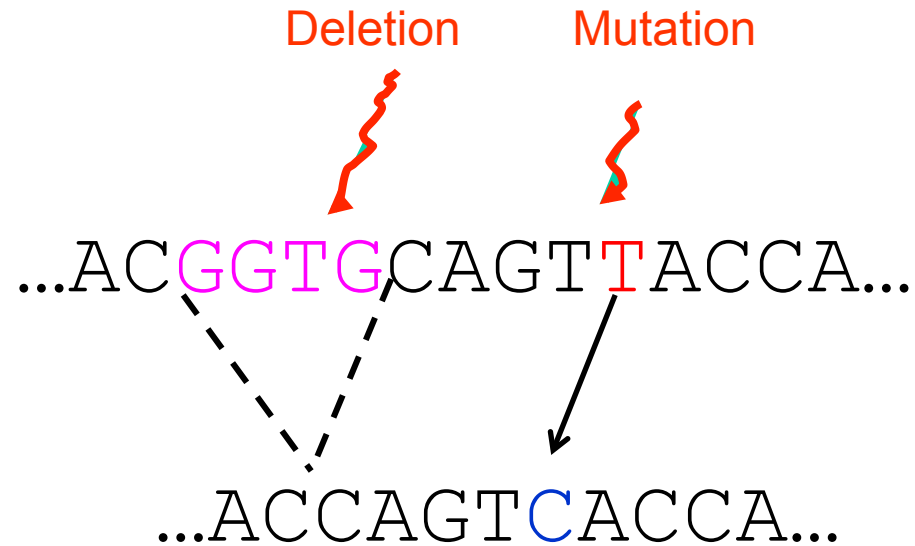
U	V	W	X	Y
●	●	●	●	●
AGGGCAT	TAGCCCA	TAGACTT	TGCACAA	TGCGCTT

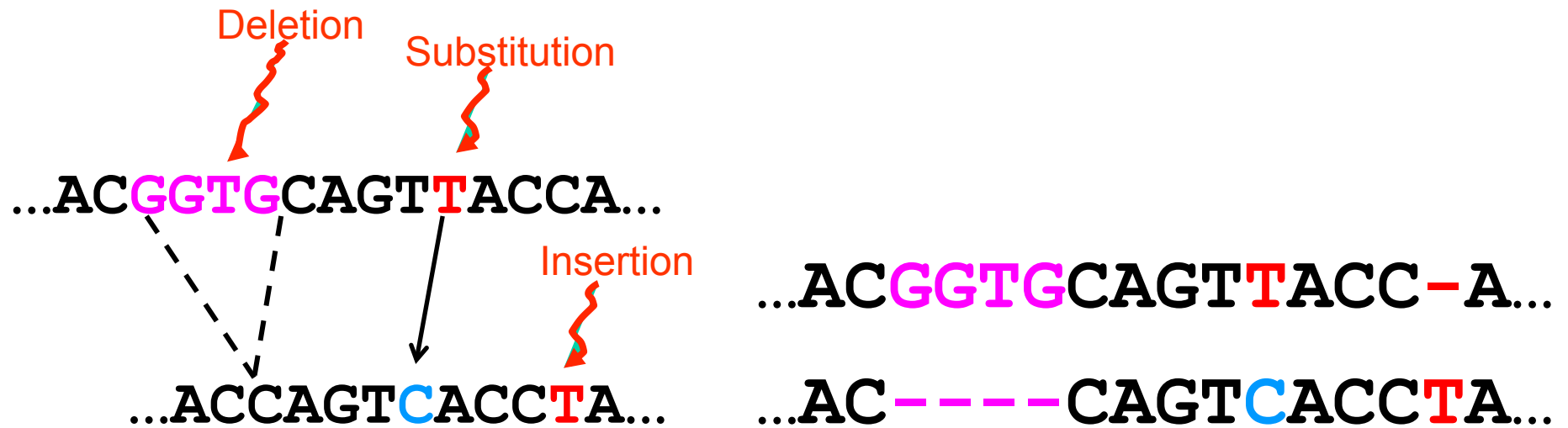


# The “real” problem



# 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

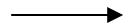
S2 = TAGCTATCACGACCGC

S3 = TAGCTGACCGC

S4 = TCACGACCGACA

# Phase 1: Alignment

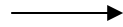
S1 = AGGCTATCACCTGACCTCCA  
S2 = TAGCTATCACGACCGC  
S3 = TAGCTGACCGC  
S4 = TCACGACCGACA



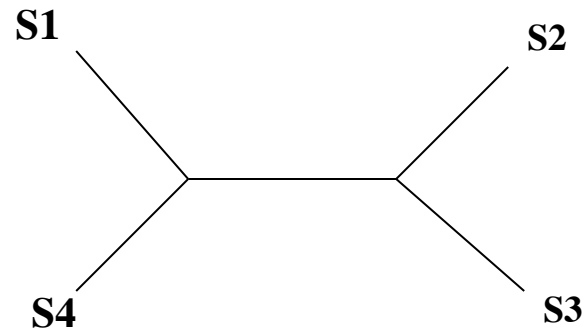
S1 = -AGGCTATCACCTGACCTCCA  
S2 = TAG-CTATCAC--GACCGC--  
S3 = TAG-CT-----GACCGC--  
S4 = -----TCAC--GACCGACA

## 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



# Phylogenomic pipeline

- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- Compute multiple sequence alignments for each locus
- Compute species tree or network:
  - Compute gene trees on the alignments and combine the estimated gene trees, OR
  - Estimate a tree from a concatenation of the multiple sequence alignments
- Get statistical support on each branch (e.g., bootstrapping)
- Estimate dates on the nodes of the phylogeny
- Use species tree with branch support and dates to understand biology

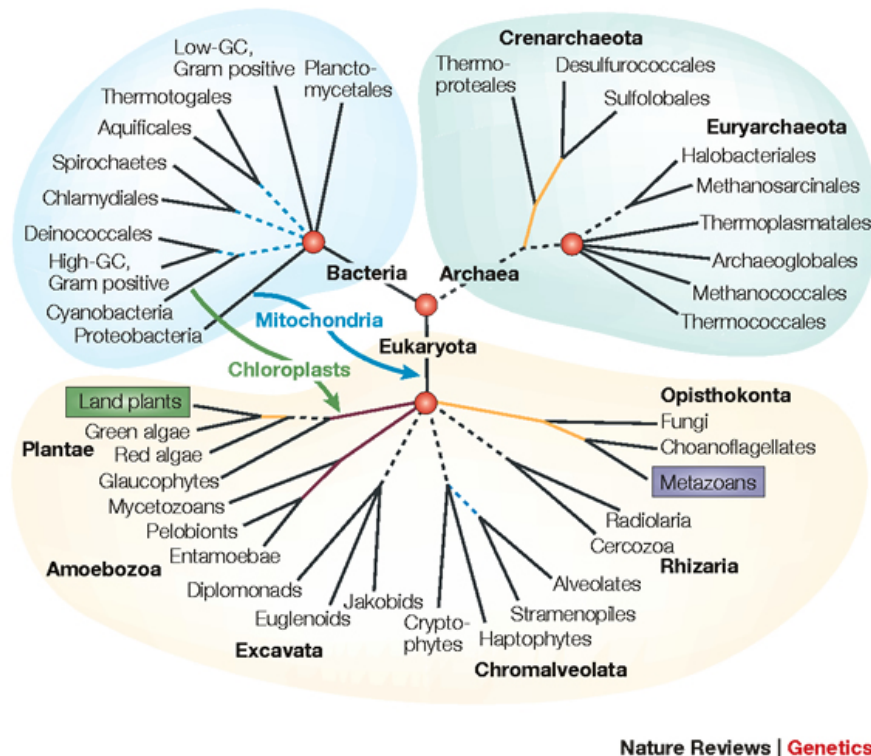
# Phylogenomic pipeline

- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- Compute multiple sequence alignments for each locus
- Compute species tree or network:
  - Compute gene trees on the alignments and combine the estimated gene trees, OR
  - Estimate a tree from a concatenation of the multiple sequence alignments
- Get statistical support on each branch (e.g., bootstrapping)
- Estimate dates on the nodes of the phylogeny
- Use species tree with branch support and dates to understand biology

# Large-scale Alignment Estimation

- Many genes are considered unalignable due to high rates of evolution
- Only a few methods can analyze large datasets
- iPlant (NSF Plant Biology Collaborative) and other projects planning to construct phylogenies with 500,000 taxa

# Hard Computational Problems



NP-hard problems

Large datasets

100,000+ sequences  
thousands of genes

“Big data” complexity:  
model misspecification  
fragmentary sequences  
errors in input data  
streaming data

# 1kp: Thousand Transcriptome Project

G. Ka-Shu Wong  
U Alberta



J. Leebens-Mack  
U Georgia



N. Wickett  
Northwestern



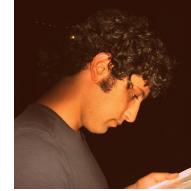
N. Matasci  
iPlant



T. Warnow,  
UIUC/IGB



S. Mirarab,  
UT-Austin



N. Nguyen,  
UIUC/IGB



Md. S. Bayzid  
UT-Austin



Plus many many other people...

- First study (Wickett, Mirarab, et al., PNAS 2014) had ~100 species and ~800 genes, gene trees and alignments estimated using SATe, and a coalescent-based species tree estimated using ASTRAL
- Second study: Plant Tree of Life based on transcriptomes of ~1200 species, and more than 13,000 gene families (most not single copy)

## Upcoming Challenges:

Species tree estimation from conflicting gene trees

**Alignment of datasets with > 100,000 sequences**

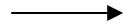
# This talk

- “Big data” multiple sequence alignment
- [SATé](#) (Science 2009, Systematic Biology 2012) and [PASTA](#) (RECOMB and JCB 2014), methods for co-estimation of alignments and trees
- The HMM Family technique, and applications to
  - phylogenetic placement ([SEPP](#), PSB 2012),
  - multiple sequence alignment ([UPP](#), submitted), and
  - metagenomic taxon identification ([TIPP](#), Bioinformatics 2014).

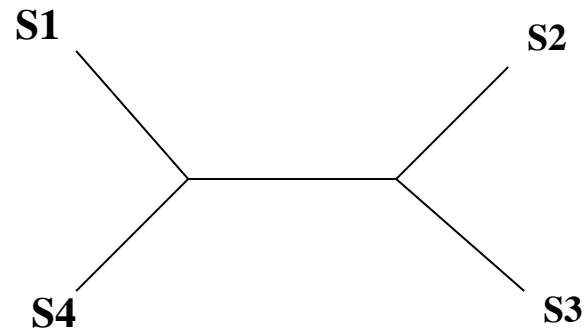
# Multiple Sequence Alignment

# First Align, then Compute the Tree

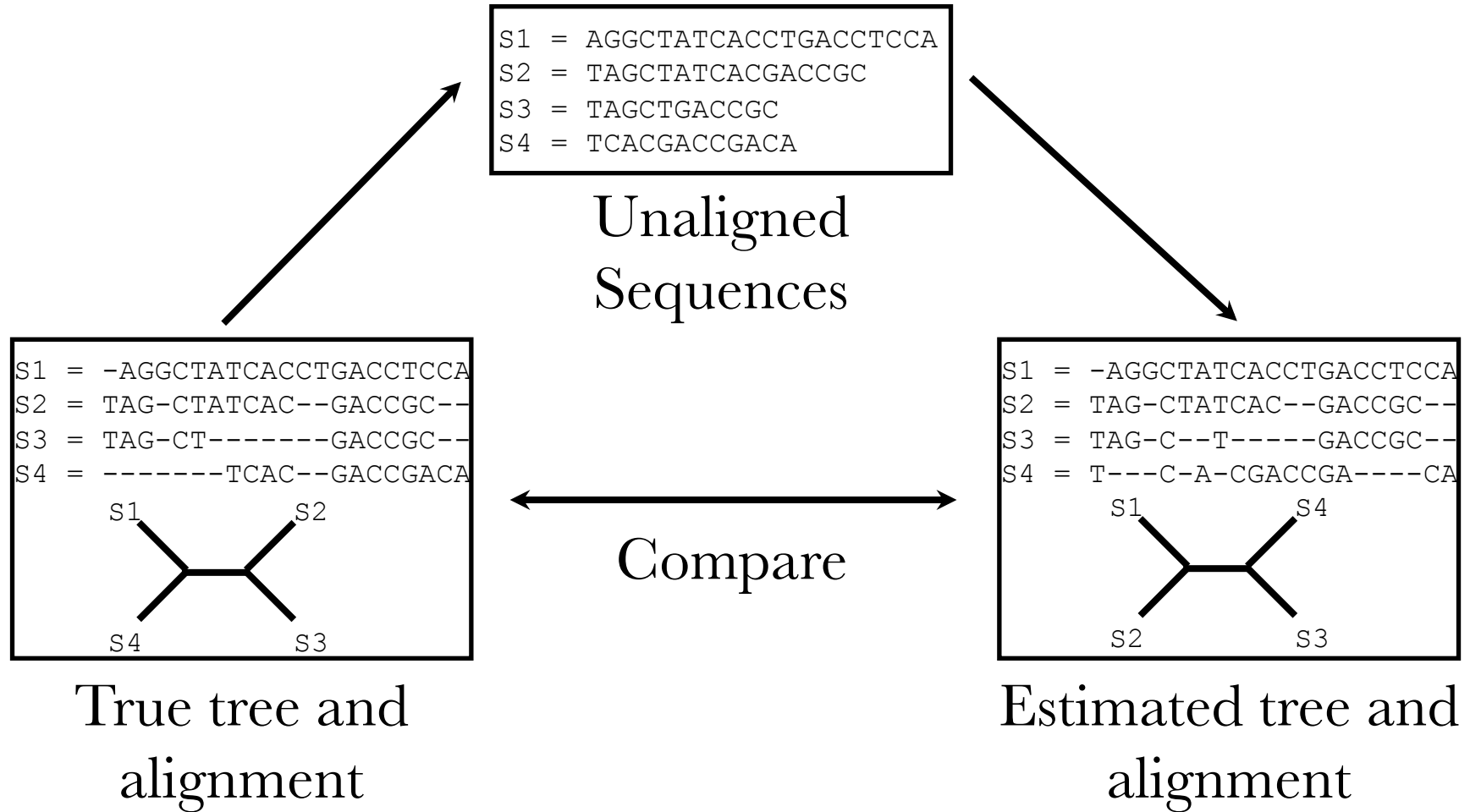
S1 = AGGCTATCACCTGACCTCCA  
S2 = TAGCTATCACGACCGC  
S3 = TAGCTGACCGC  
S4 = TCACGACCGACA



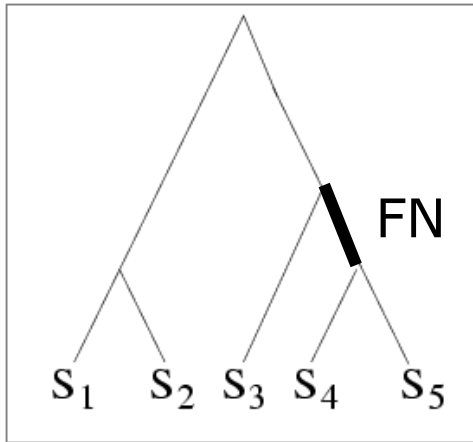
S1 = -AGGCTATCACCTGACCTCCA  
S2 = TAG-CTATCAC--GACCGC--  
S3 = TAG-CT-----GACCGC--  
S4 = -----TCAC--GACCGACA



# Simulation Studies



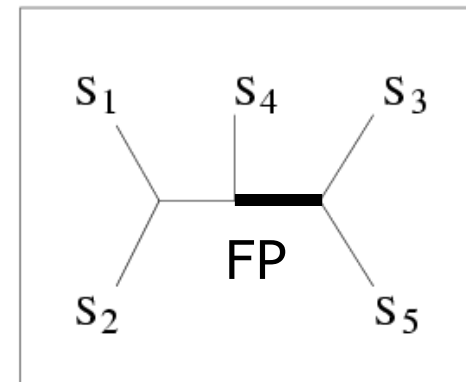
# Quantifying Error



TRUE TREE

S <sub>1</sub>	ACAATTAGAAC
S <sub>2</sub>	ACCCTTAGAAC
S <sub>3</sub>	ACCATTCCAAC
S <sub>4</sub>	ACCAGACCAAC
S <sub>5</sub>	ACCAGACCGGA

DNA SEQUENCES



INFERRED TREE

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

50% error rate

# Two-phase estimation

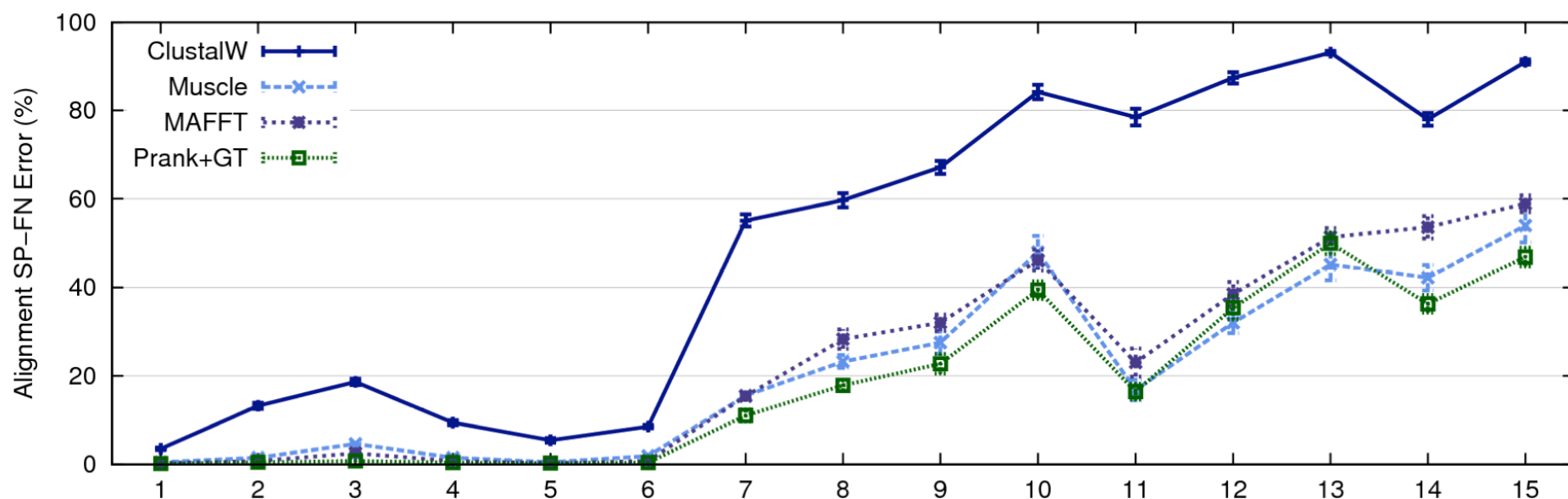
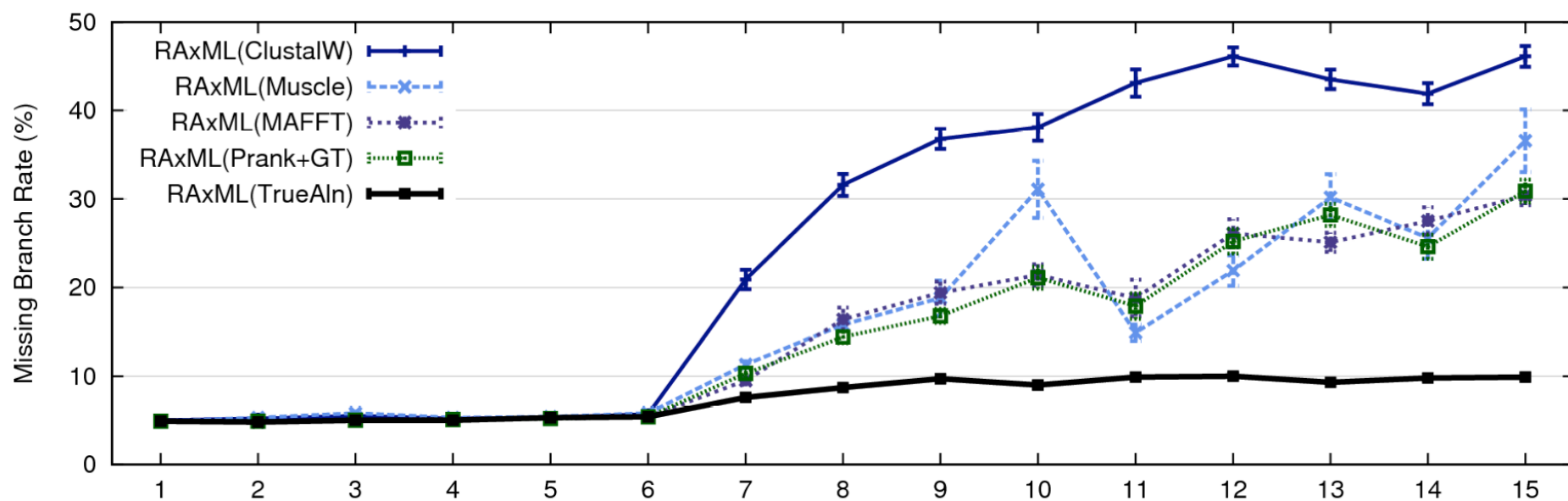
## Alignment methods

- 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

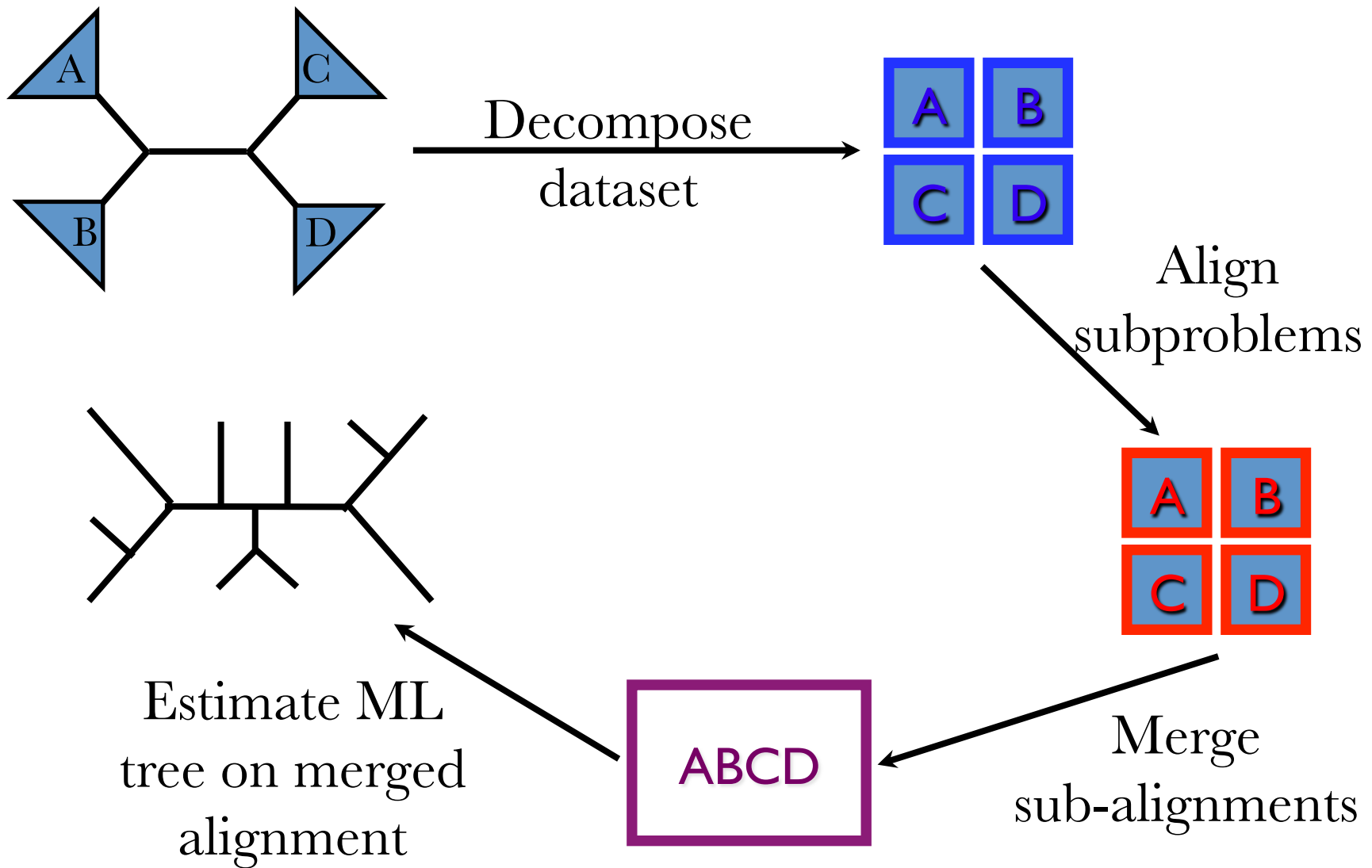
- Bayesian MCMC
- Maximum parsimony
- **Maximum likelihood**
- Neighbor joining
- FastME
- UPGMA
- Quartet puzzling
- Etc.

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



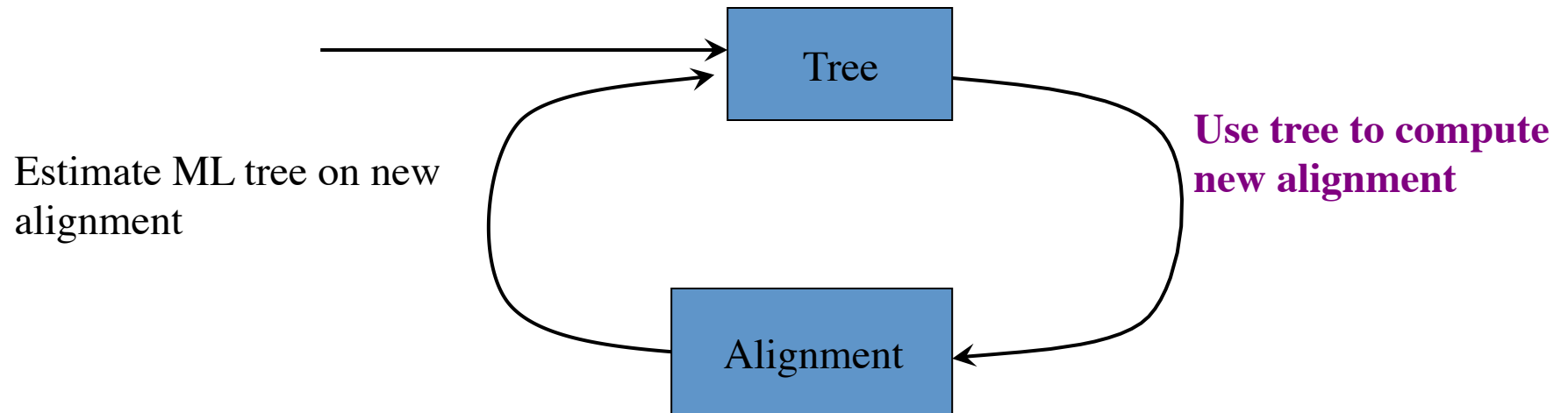
1000-taxon models, ordered by difficulty (Liu et al., 2009)

# Re-aligning on a tree



# SATé and PASTA Algorithms

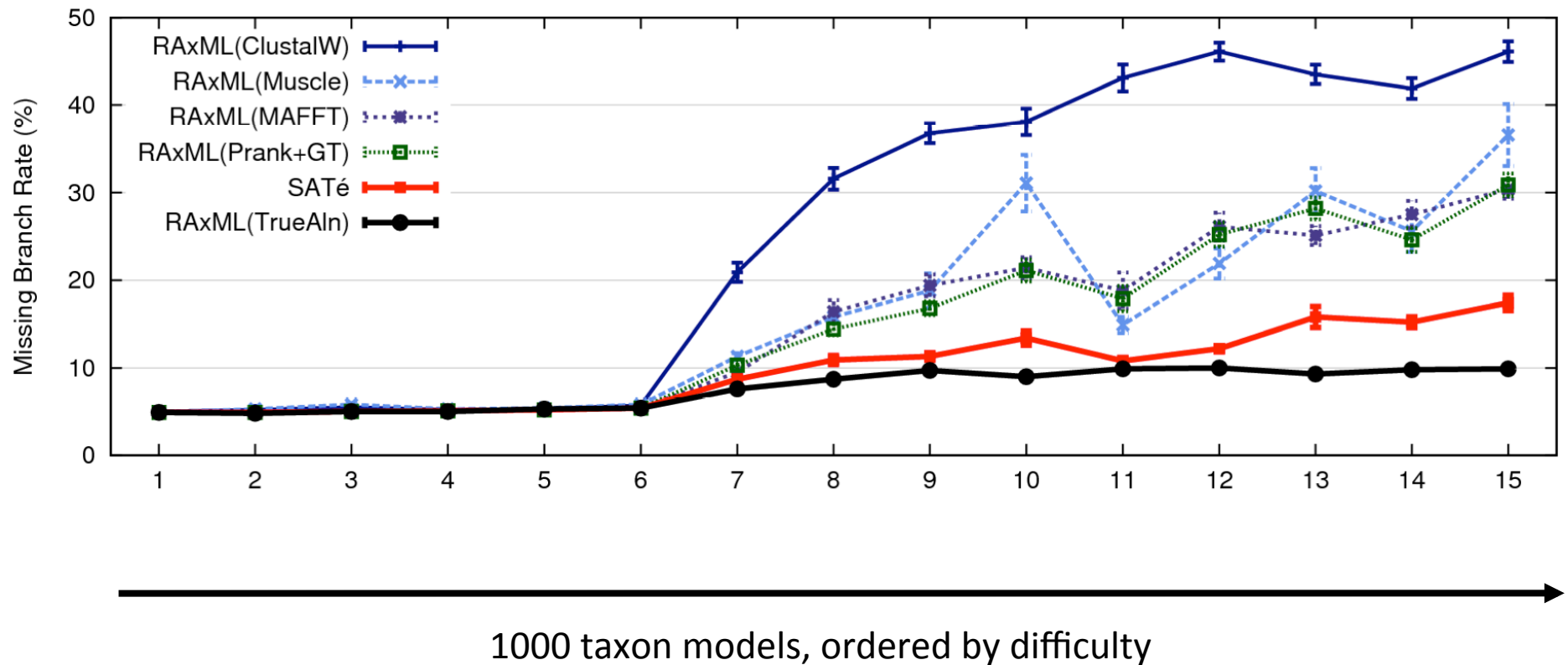
Obtain initial alignment and  
estimated ML tree



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

Repeat until termination condition (typically, 24 hours)

# SATé-1 (Science 2009) performance

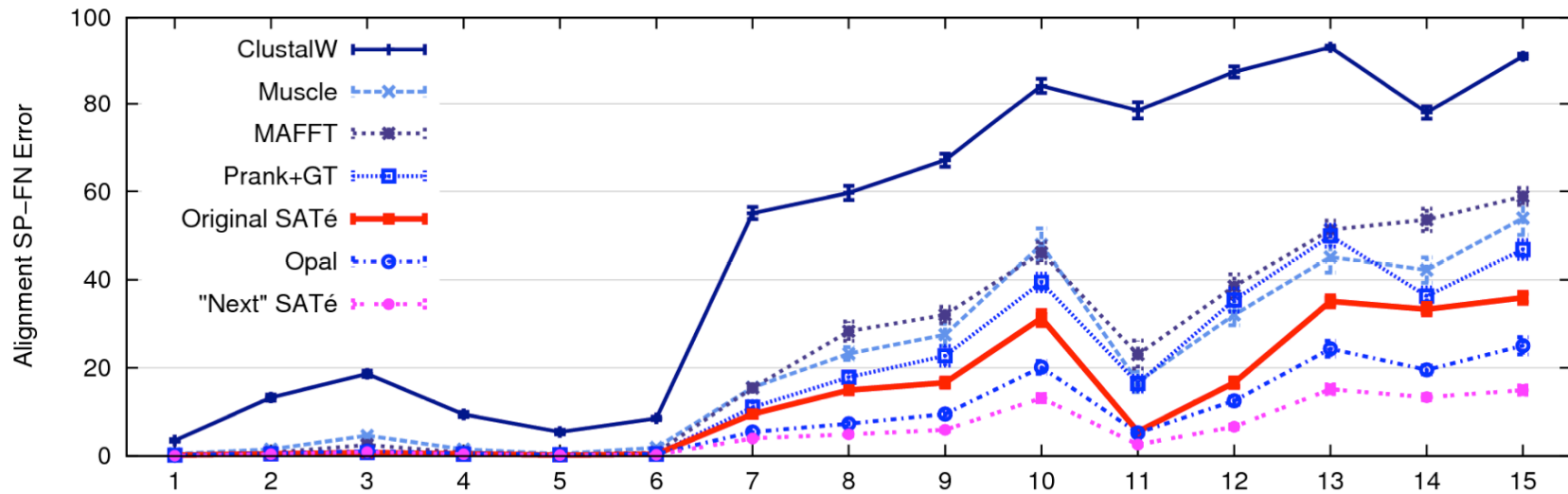
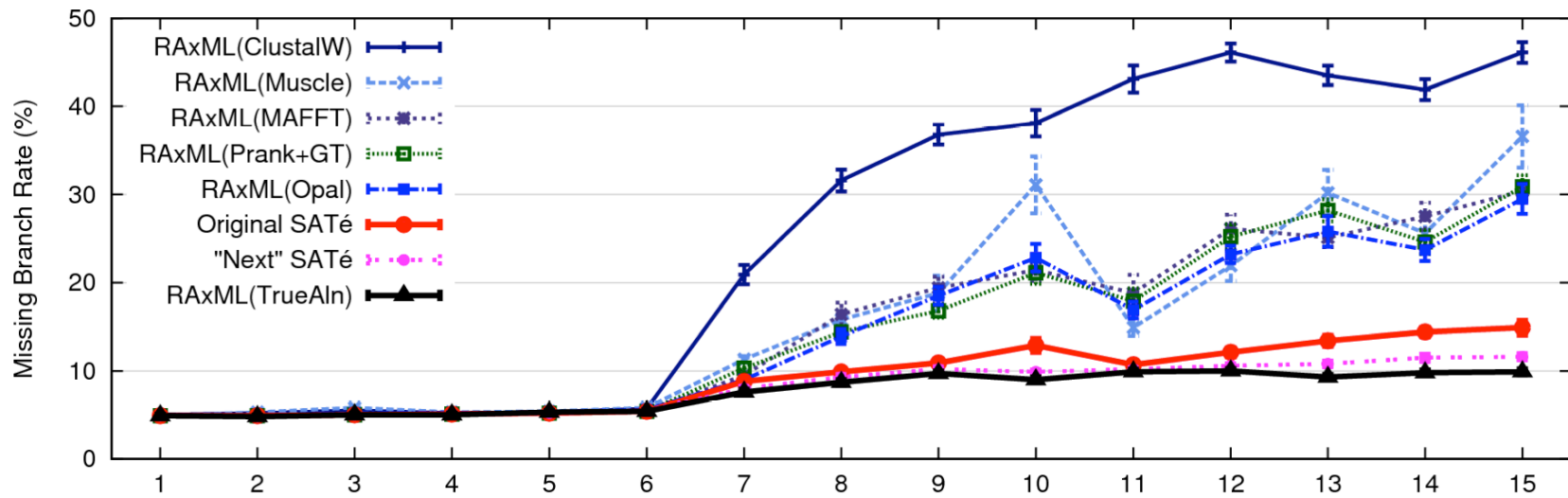


SATé-1 24 hour analysis, on desktop machines

(Similar improvements for biological datasets)

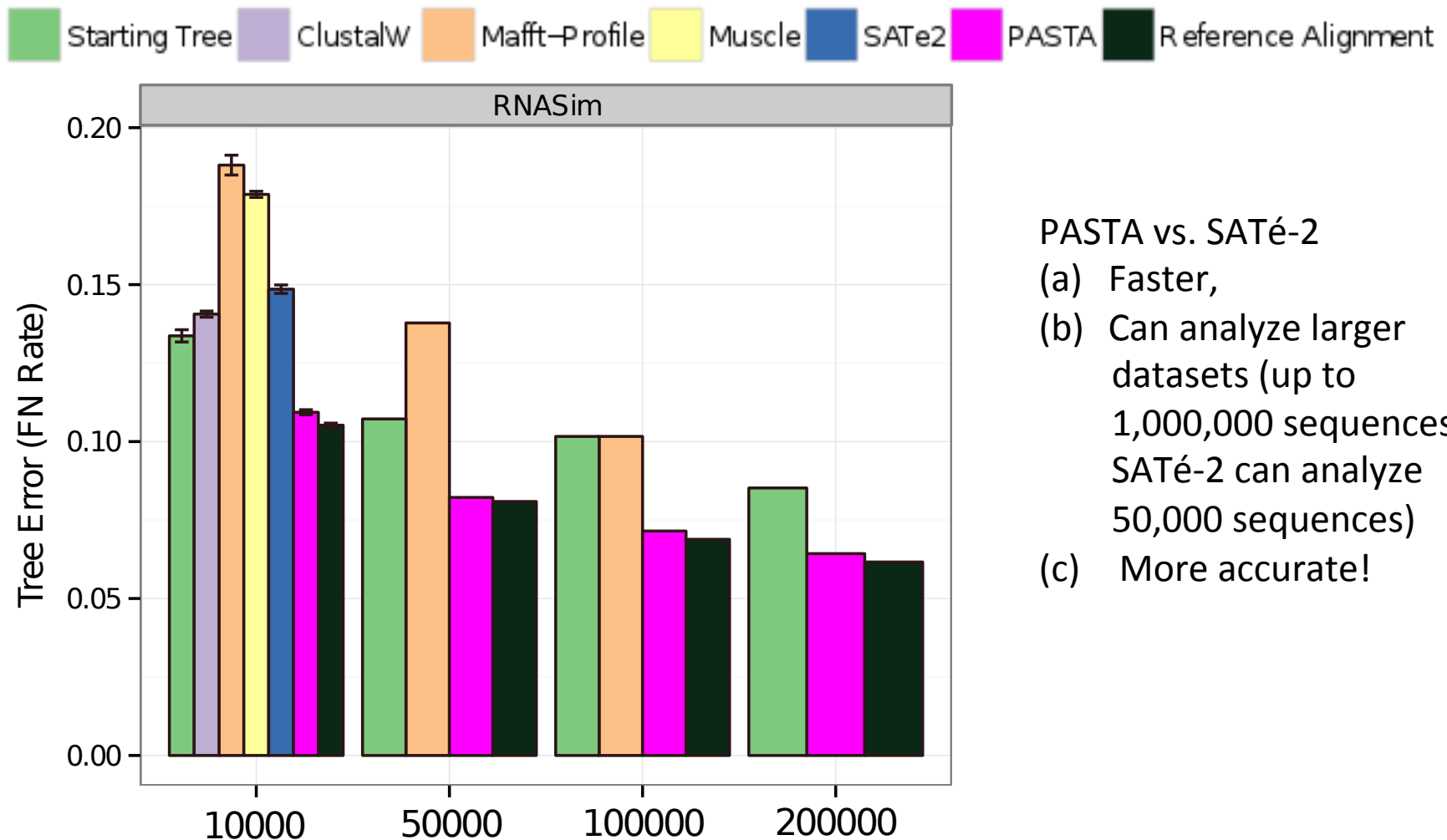
SATé-1 can analyze up to about 30,000 sequences.

## SATé-1 and SATé-2 (Systematic Biology, 2012)



1000 taxon models ranked by difficulty

# PASTA (2014): even better than SATé-2



## PASTA vs. SATé-2

- (a) Faster,
- (b) Can analyze larger datasets (up to 1,000,000 sequences – SATé-2 can analyze 50,000 sequences)
- (c) More accurate!

# 1kp: Thousand Transcriptome Project

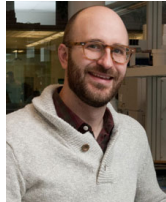
G. Ka-Shu Wong  
U Alberta



J. Leebens-Mack  
U Georgia



N. Wickett  
Northwestern



N. Matasci  
iPlant



T. Warnow,  
UIUC



S. Mirarab,  
UT-Austin



N. Nguyen,  
UT-Austin



Md. S.Bayzid  
UT-Austin

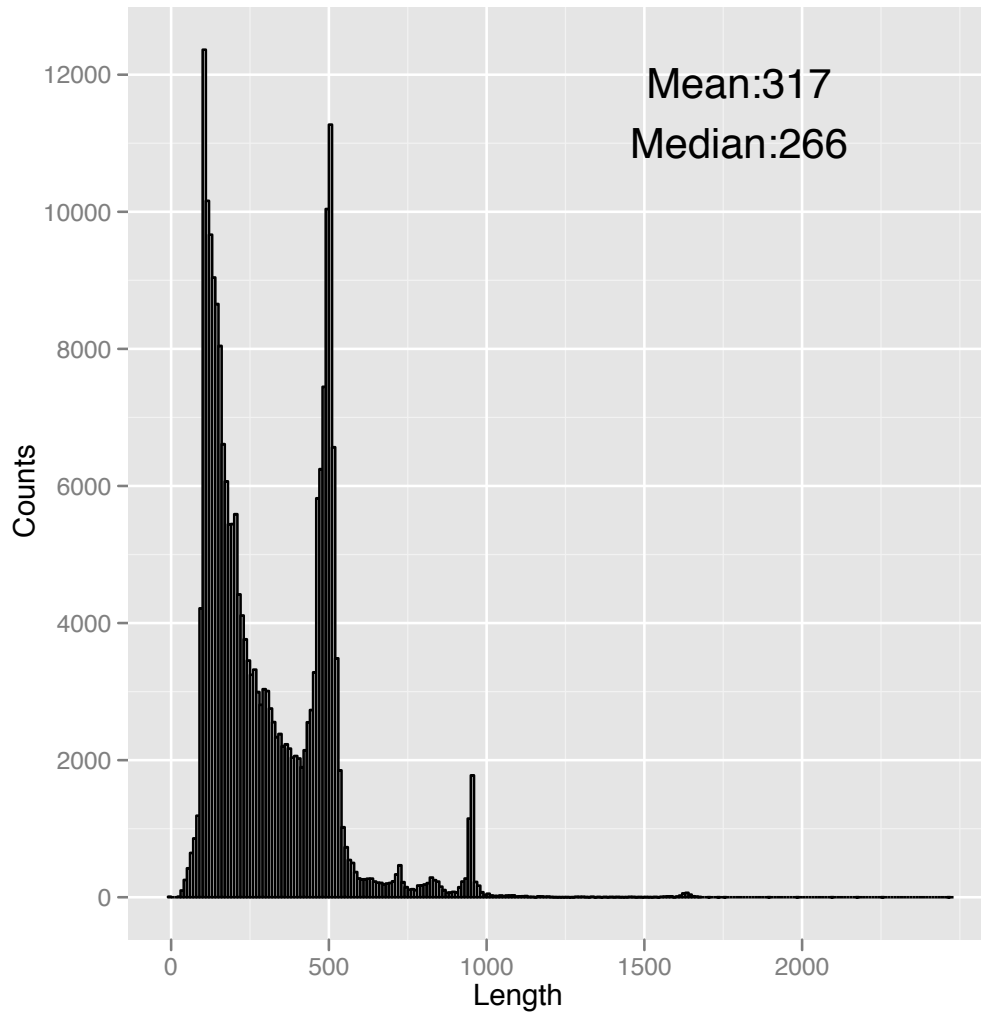


Plus many many other people...

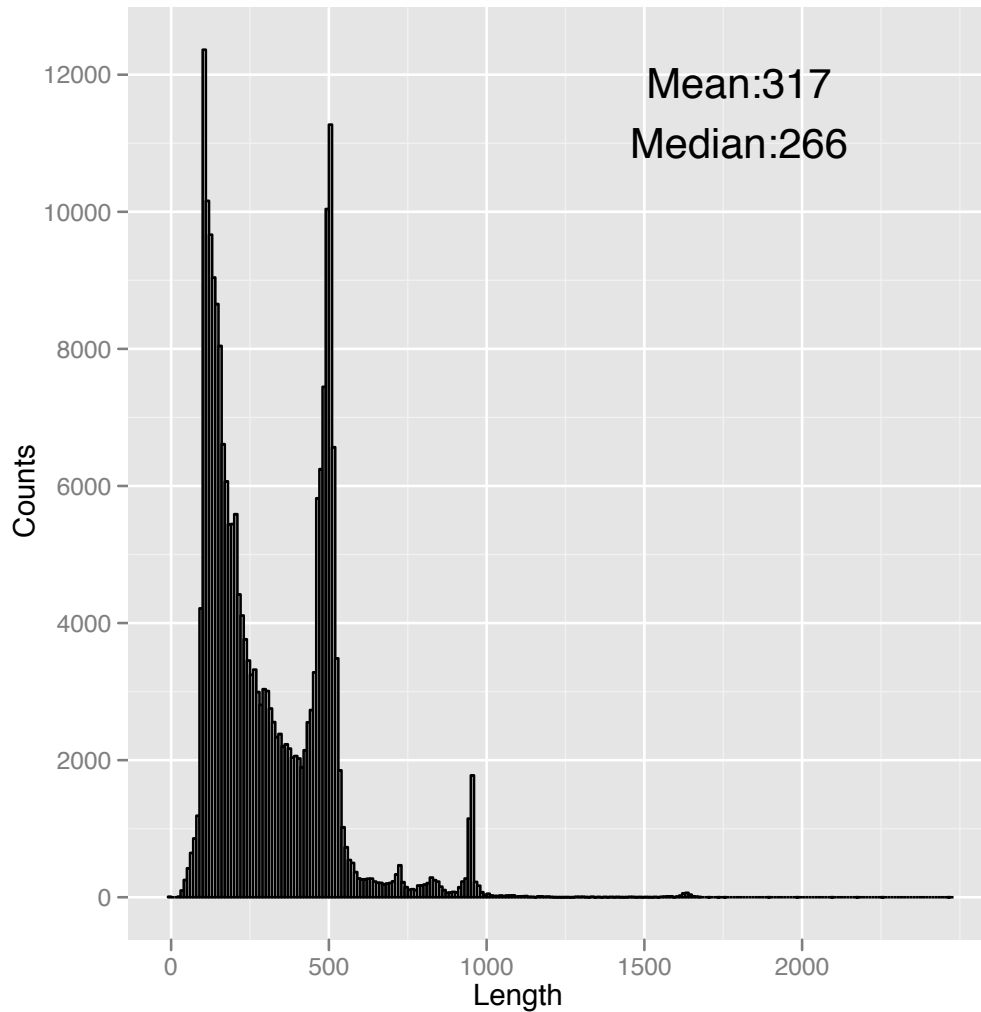
- Plant Tree of Life based on transcriptomes of ~1200 species
- More than 13,000 gene families (most not single copy)

Challenge:

**Alignment of datasets with > 100,000 sequences**



1KP dataset: more than  
100,000 p450 amino-acid  
sequences, many fragmentary



1KP dataset: more than 100,000 p450 amino-acid sequences, many fragmentary

*All standard multiple sequence alignment methods we tested performed poorly on datasets with fragments.*

# 1kp: Thousand Transcriptome Project

G. Ka-Shu Wong  
U Alberta



J. Leebens-Mack  
U Georgia



N. Wickett  
Northwestern



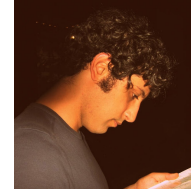
N. Matasci  
iPlant



T. Warnow,  
UIUC



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**

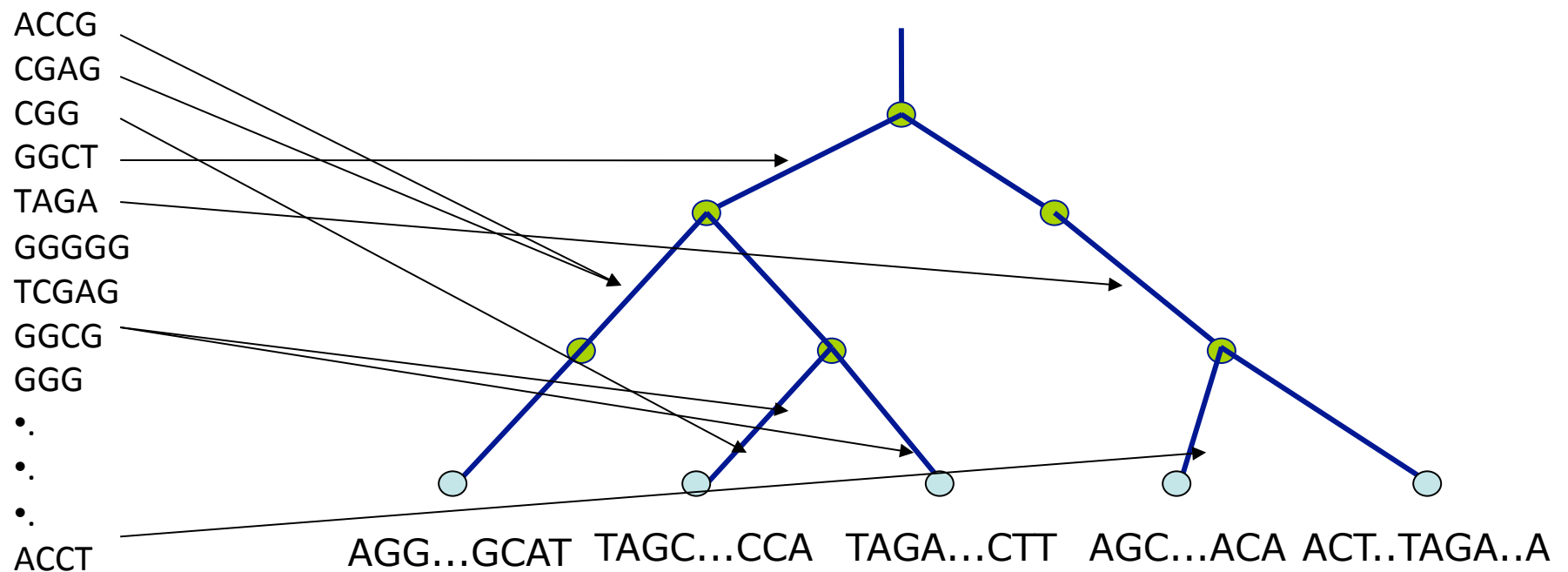
# SEPP

- **SEPP: SATé-enabled Phylogenetic Placement**, by Mirarab, Nguyen, and Warnow. Pacific Symposium on Biocomputing, 2012, special session on the Human Microbiome
- Objective:
  - phylogenetic analysis of single-gene datasets with **fragmentary sequences**
- Introduces “HMM Family” technique

# Phylogenetic Placement

Fragmentary sequences  
from some gene

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



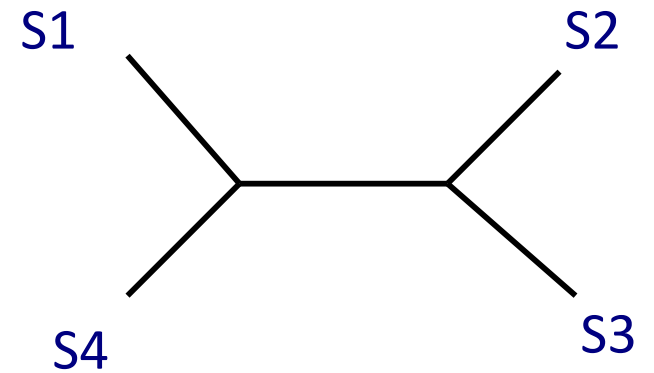
# Phylogenetic Placement

Step 1: Align each query sequence to backbone alignment

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

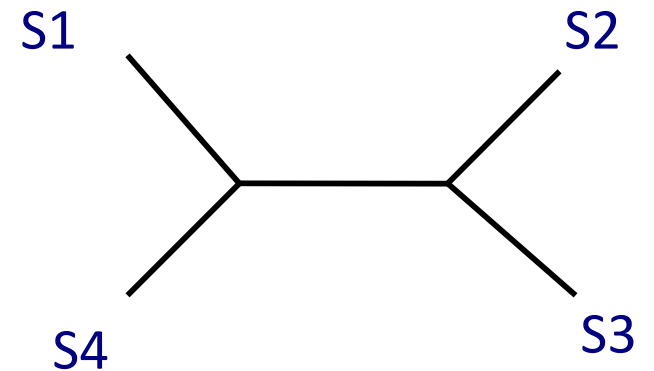
# Align Sequence

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



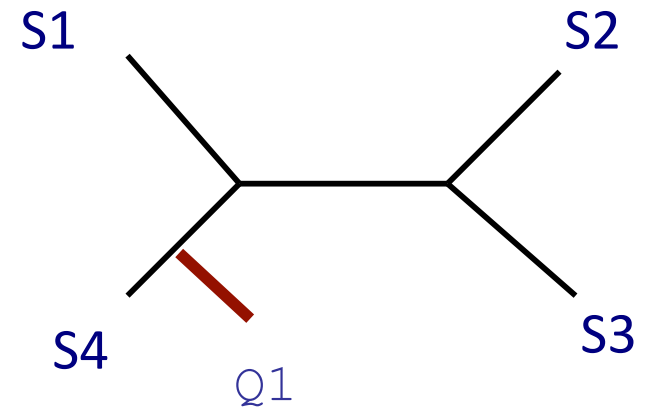
# Align Sequence

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-----

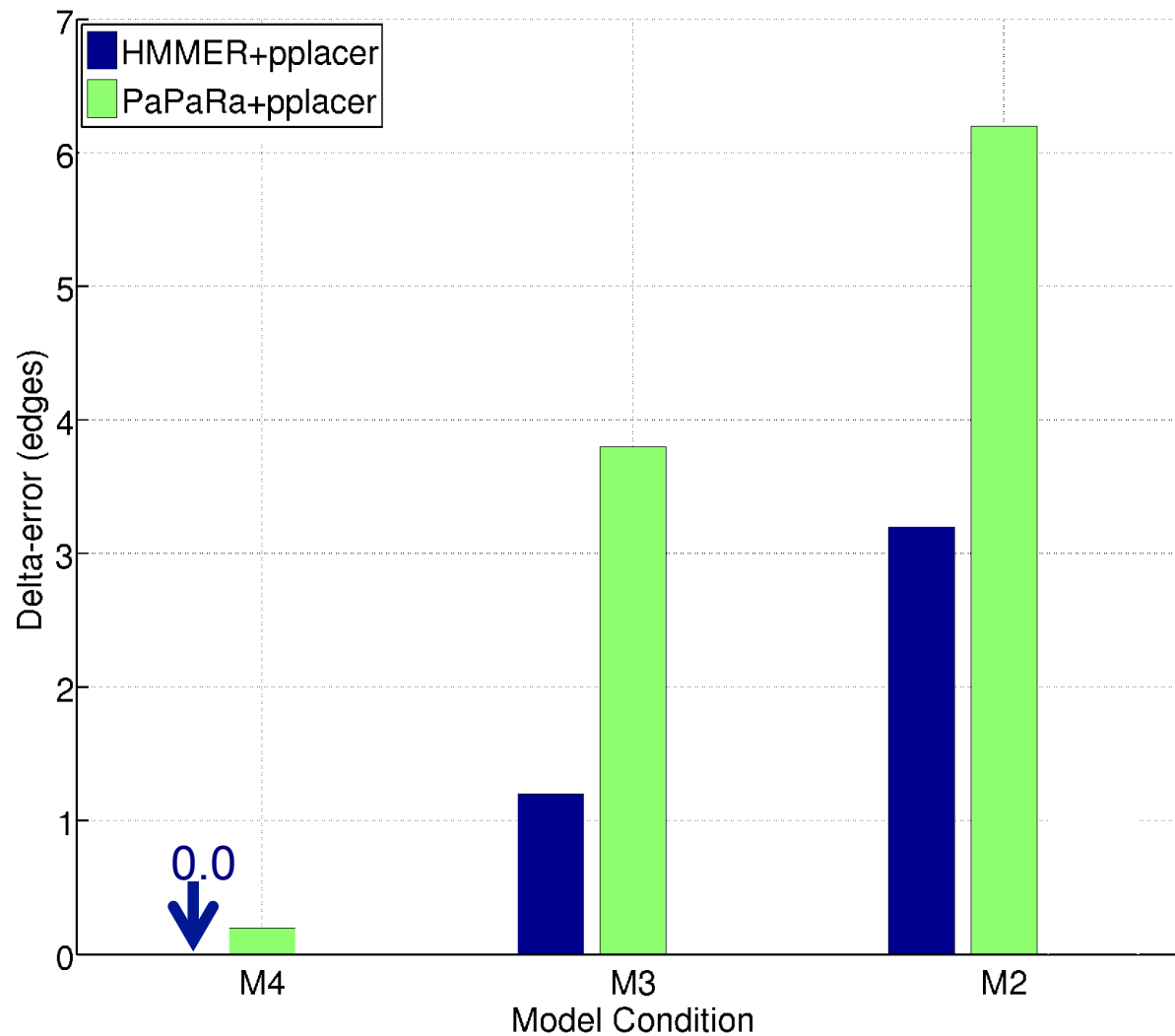


# Phylogenetic Placement

- 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)

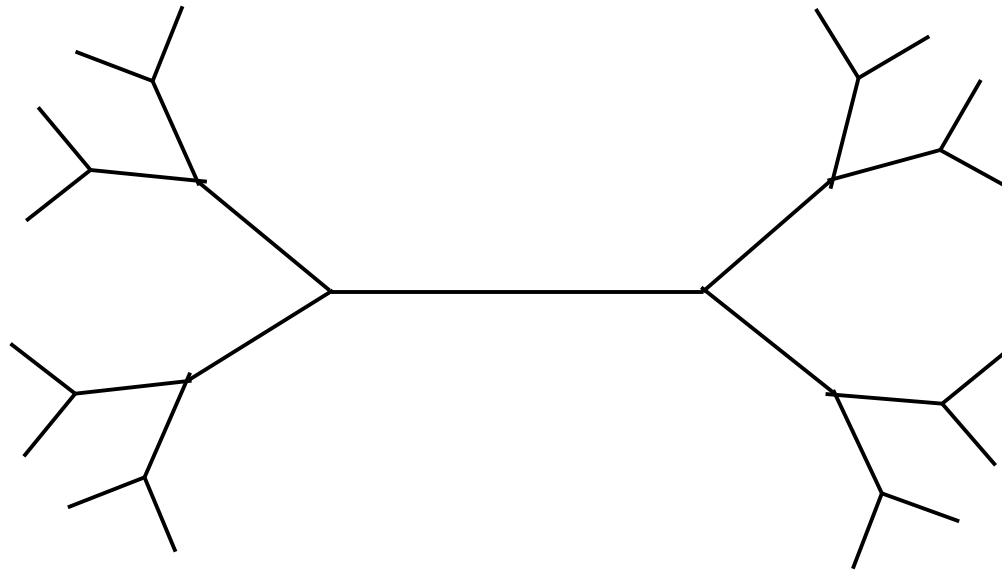
Note: pplacer and EPA use maximum likelihood, and are reported to have the same accuracy.

# HMMER vs. PaPaRa

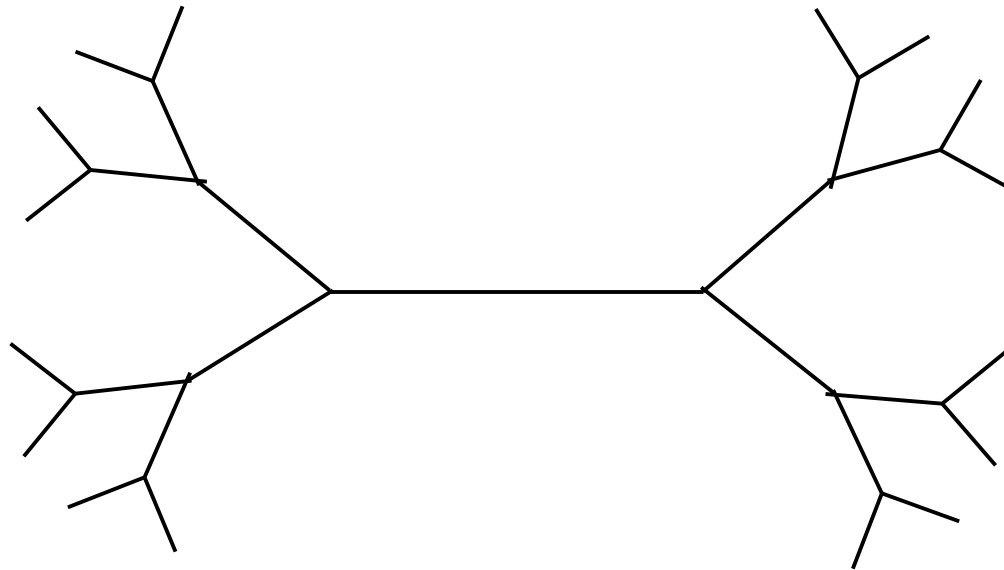


## HMMER+pplacer:

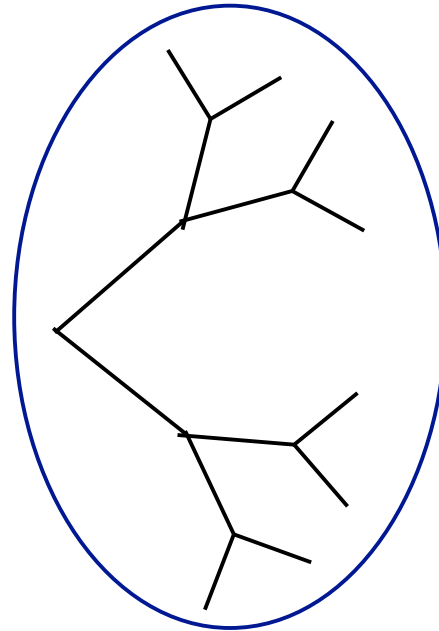
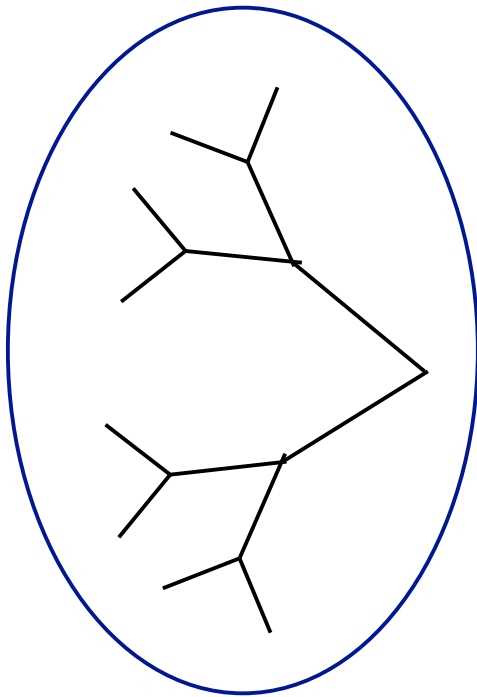
- 1) build one HMM for the entire alignment
- 2) Align fragment to the HMM, and insert into alignment
- 3) Insert fragment into tree to optimize likelihood



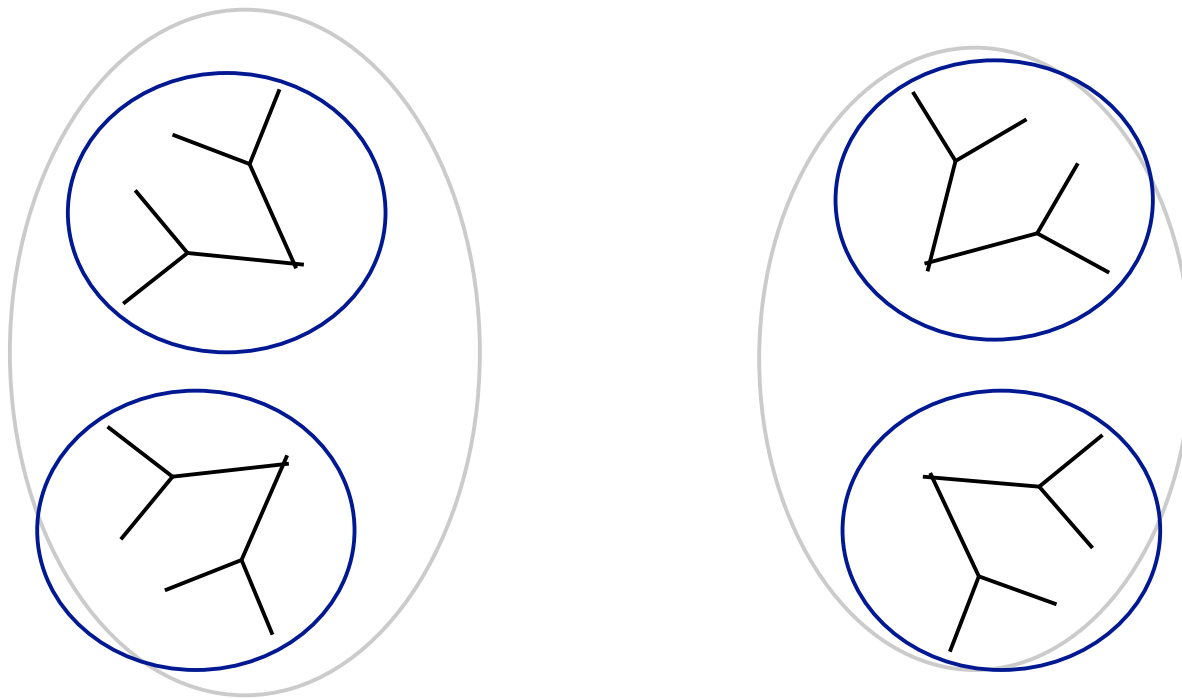
One Hidden Markov Model  
for the entire alignment?



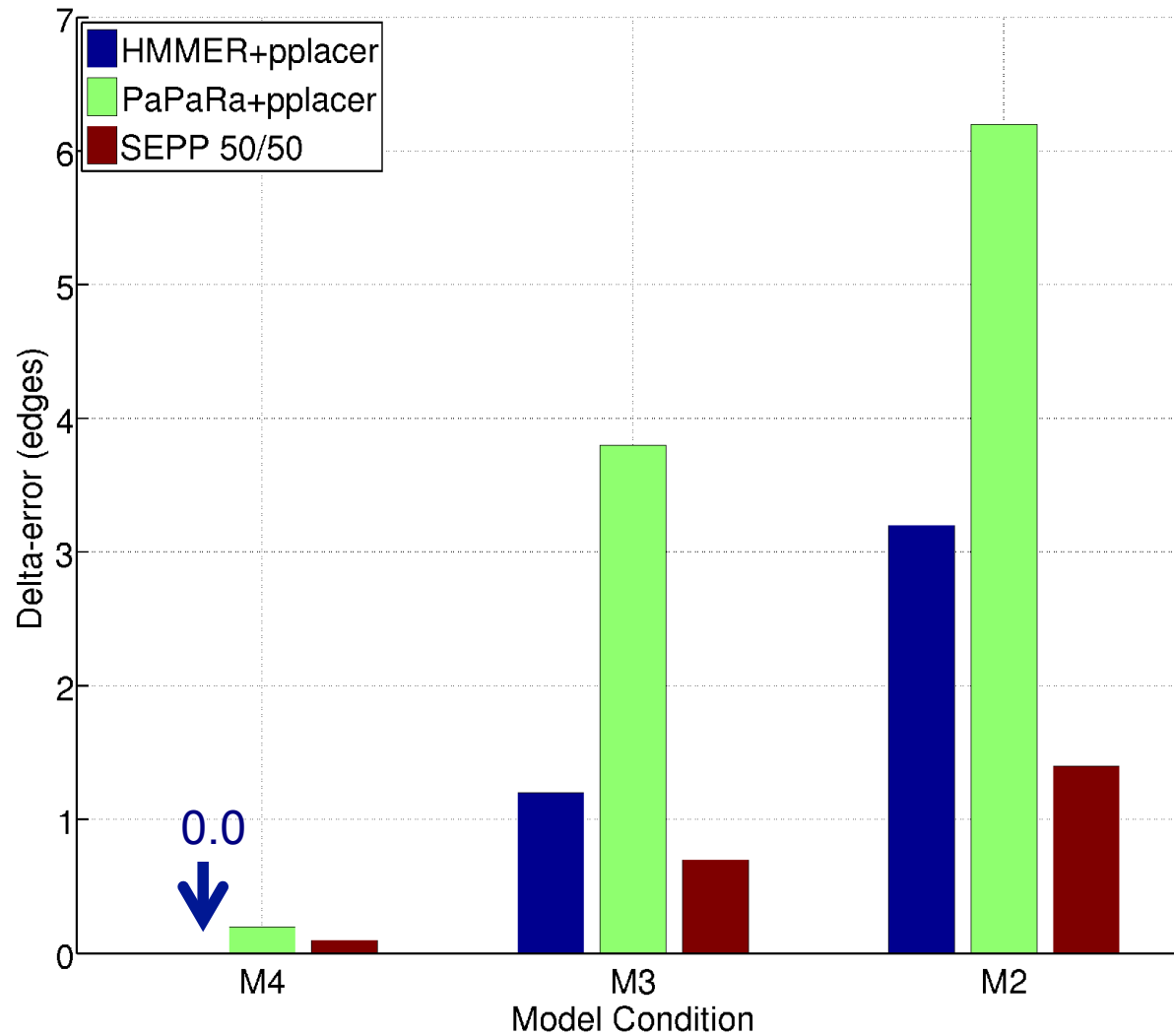
Or 2 HMMs?



Or 4 HMMs?



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



Increasing rate of evolution

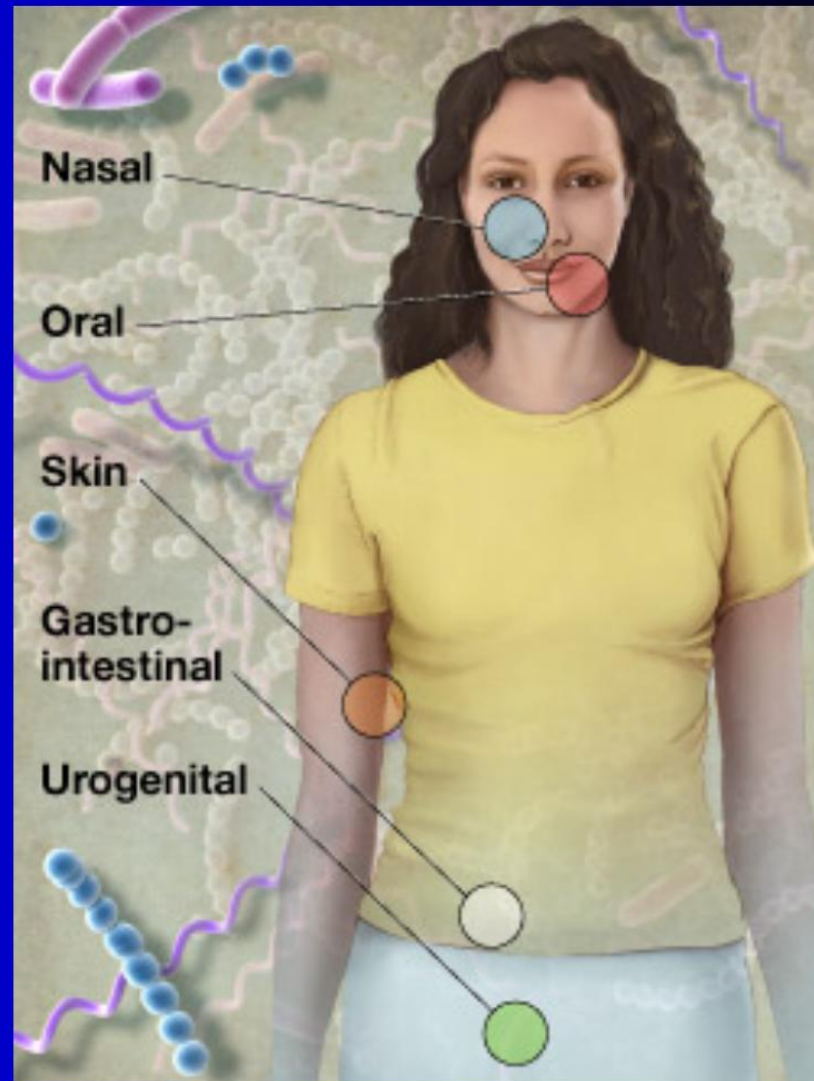


# SEPP vs. HMMER+pplacer

- SEPP produced more accurate phylogenetic placements than HMMER+pplacer.
- The only difference is the use of a Family of HMMs instead of one HMM.
- The biggest differences are for datasets with high rates of evolution.

# The NIH Human Microbiome Project

25,000 human genes,  
1,000,000 bacterial genes



# 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

# **TIPP: SEPP + statistics**

SEPP 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)

TIPP: Taxon Identification and Phylogenetic Profiling. N. Nguyen, S. Mirarab, and T. Warnow. Bioinformatics, 2014.

# 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

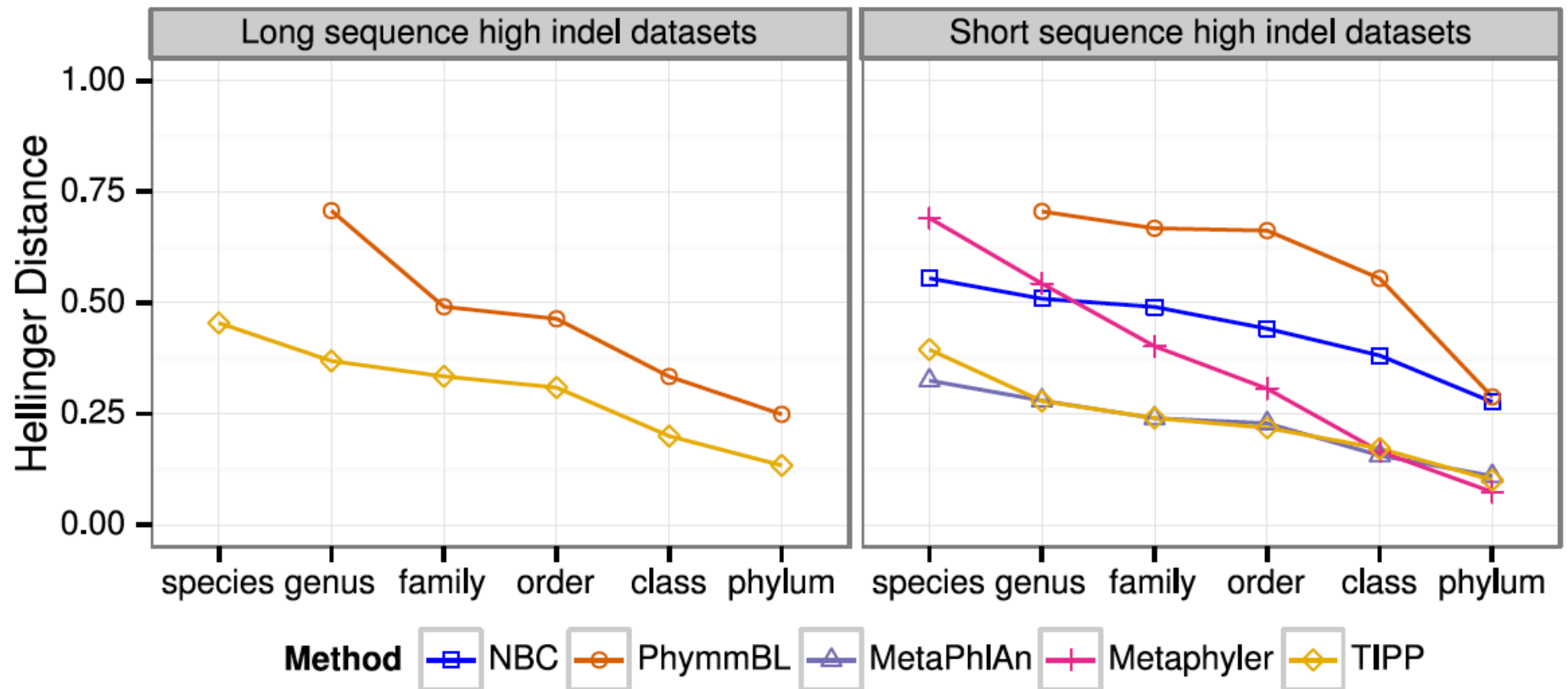
[MetaPhlAn](#) (Segata et al., Nature Methods 2012), from the Huttenhower Lab at Harvard

[mOTU](#) (Bork et al., Nature Methods 2013)

MetaPhyler, MetaPhlAn, and mOTU are [marker-based](#) techniques (but use different marker genes).

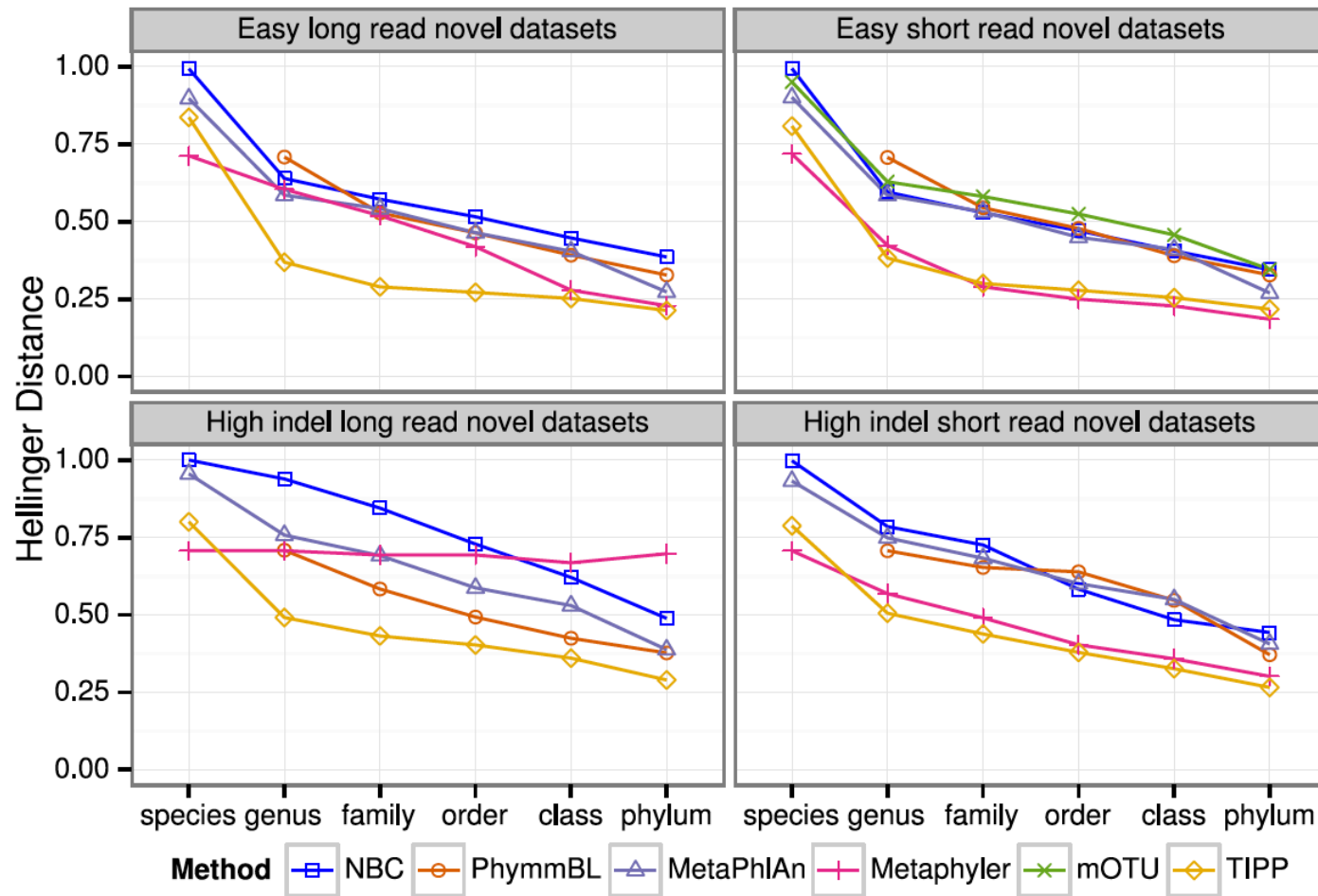
[Marker gene](#) are single-copy, universal, and resistant to horizontal transmission.

# High indel datasets containing known genomes



Note: NBC, MetaPhlAn, and MetaPhyler cannot classify any sequences from at least one of the high indel long sequence datasets, and 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.

# UPP: large-scale MSA estimation

UPP = “Ultra-large multiple sequence alignment using Phylogeny-aware Profiles”

Nguyen, Mirarab, and Warnow. Under review.

Objective: highly accurate large-scale multiple sequence alignments, even in the presence of fragmentary sequences.

*Uses a variant of the HMM Family technique in SEPP*

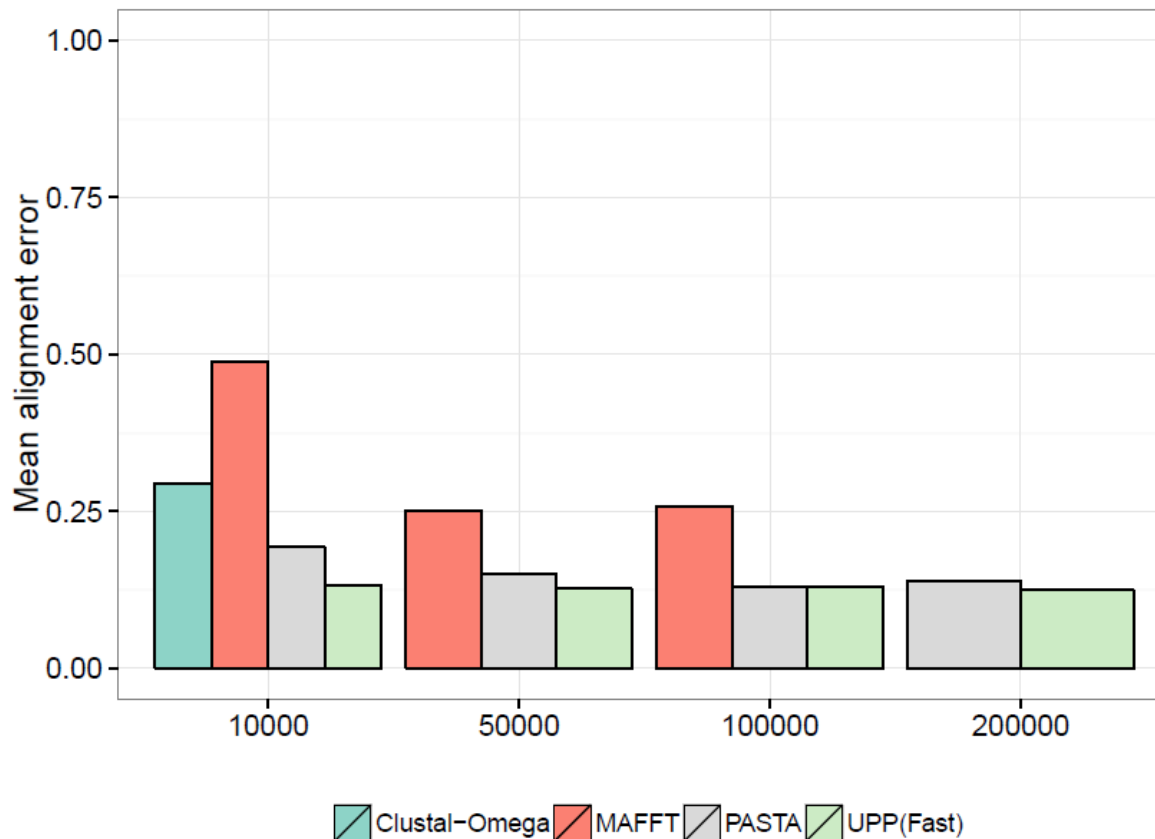
# UPP Algorithmic Approach

- Select random subset of sequences, and build “backbone alignment”
- Construct an “Ensemble of Hidden Markov Models” on the backbone alignment (the family has HMMs on many subsets of different sizes, not disjoint)
- Add all remaining sequences to the backbone alignment using the Ensemble of HMMs

# Evaluation

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

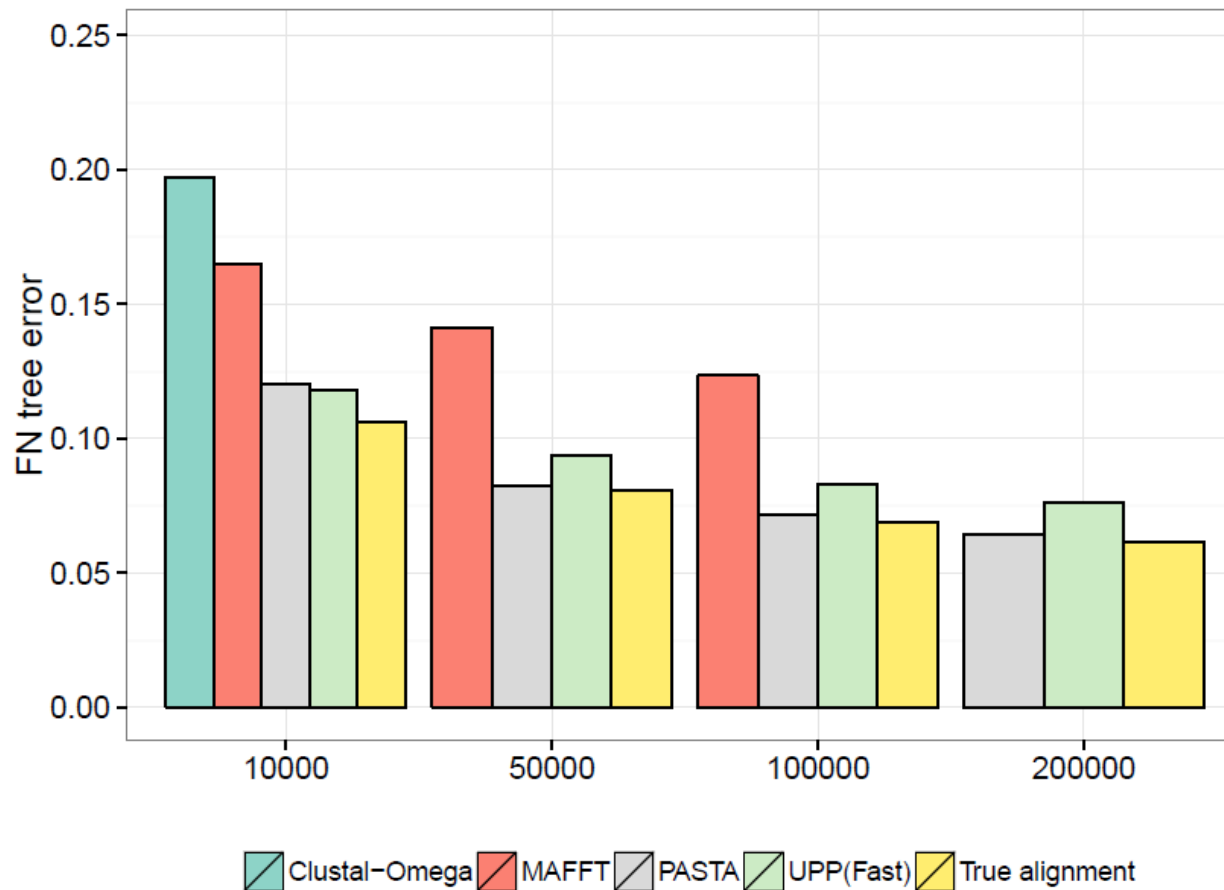
# RNASim: alignment error



All methods  
given 24 hrs  
on a 12-core  
machine

Note: Mafft was run under default settings for 10K and 50K sequences and under Parttree for 100K sequences, and fails to complete under any setting For 200K sequences. Clustal-Omega only completes on 10K dataset.

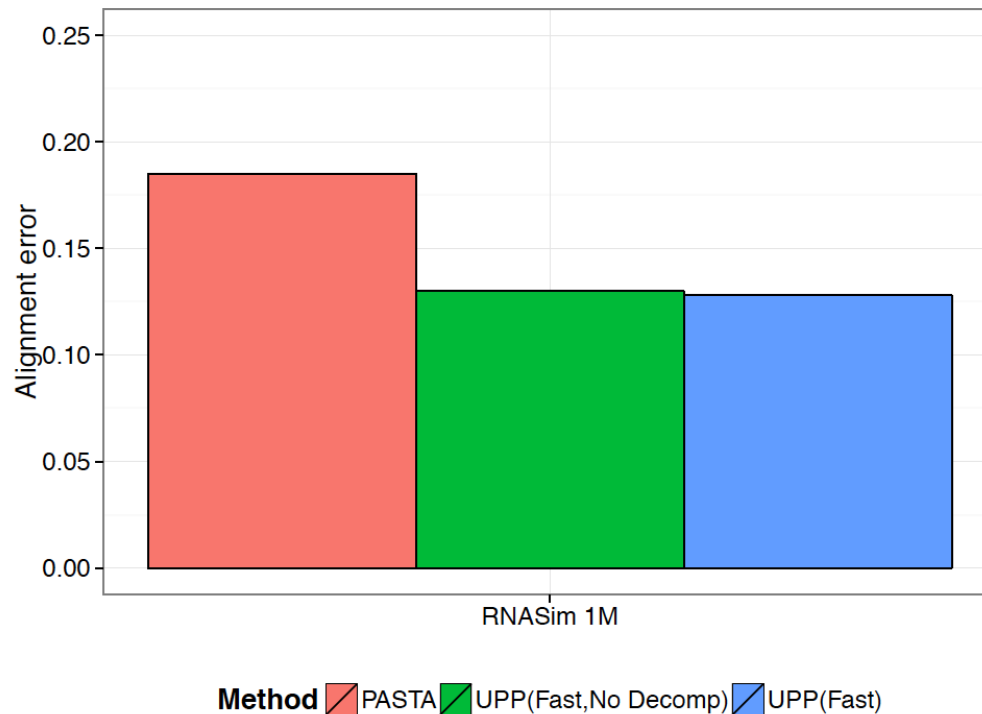
# RNASim: tree error



All methods  
given 24 hrs  
on a 12-core  
machine

Note: Mafft was run under default settings for 10K and 50K sequences and under Parttree for 100K sequences, and fails to complete under any setting For 200K sequences. Clustal-Omega only completes on 10K dataset.

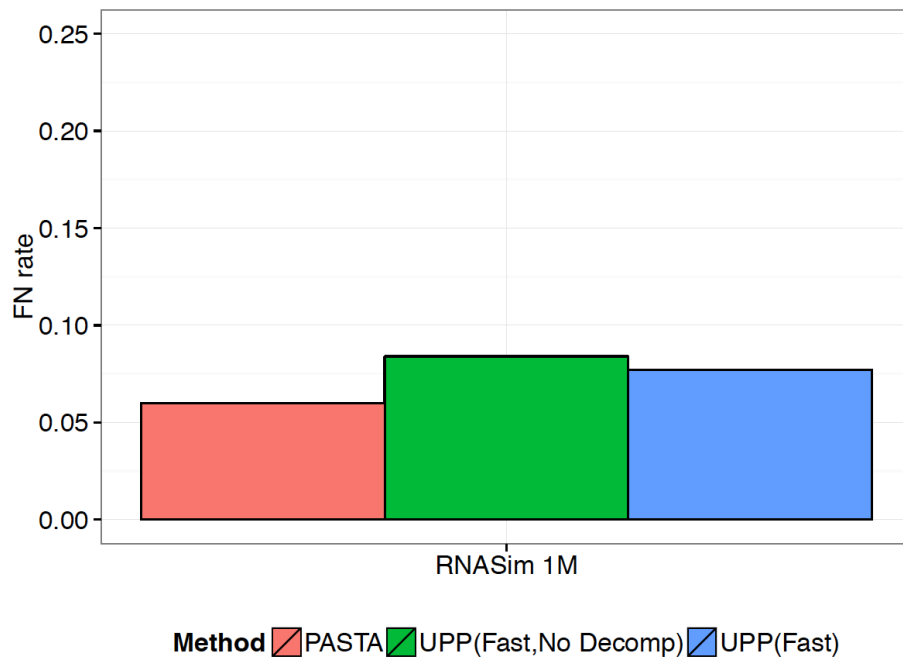
# RNASim Million Sequences: alignment error



## Notes:

- We show alignment error using average of SP-FN and SP-FP. UPP variants have better scores than PASTA.
- But for the Total Column (TC) scores, PASTA is better than UPP: it recovered 10% of the columns compared to less than 0.04% for UPP variants.

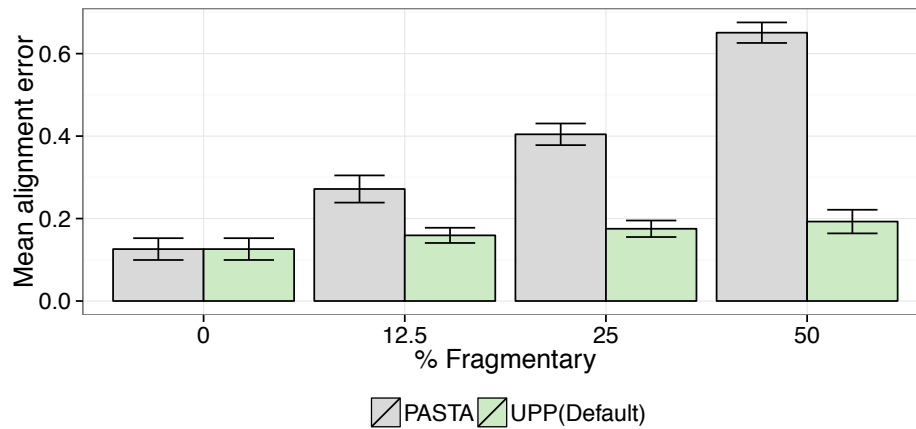
# RNASim Million Sequences: tree error



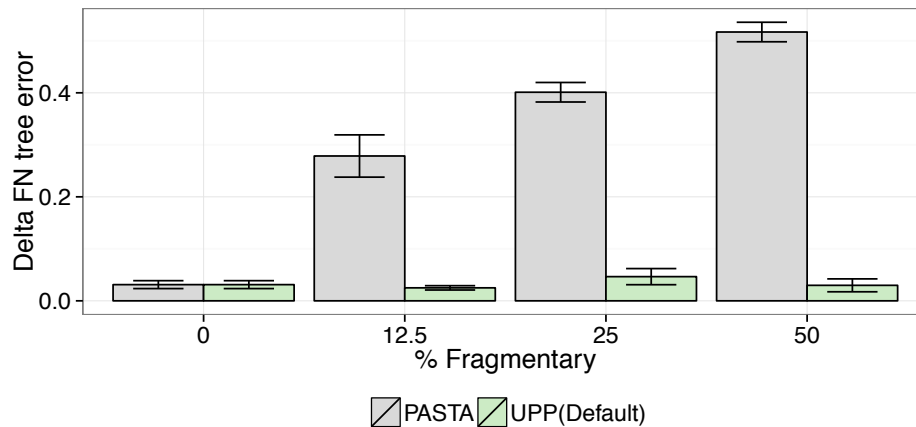
Running time (12 processors):

- UPP(Fast, NoDecomp) took 2.2 days,
- UPP(Fast) took 11.9 days, and
- PASTA took 10.3 days

# UPP vs. PASTA: impact of fragmentation



(a) Average alignment error



(b) Average tree error

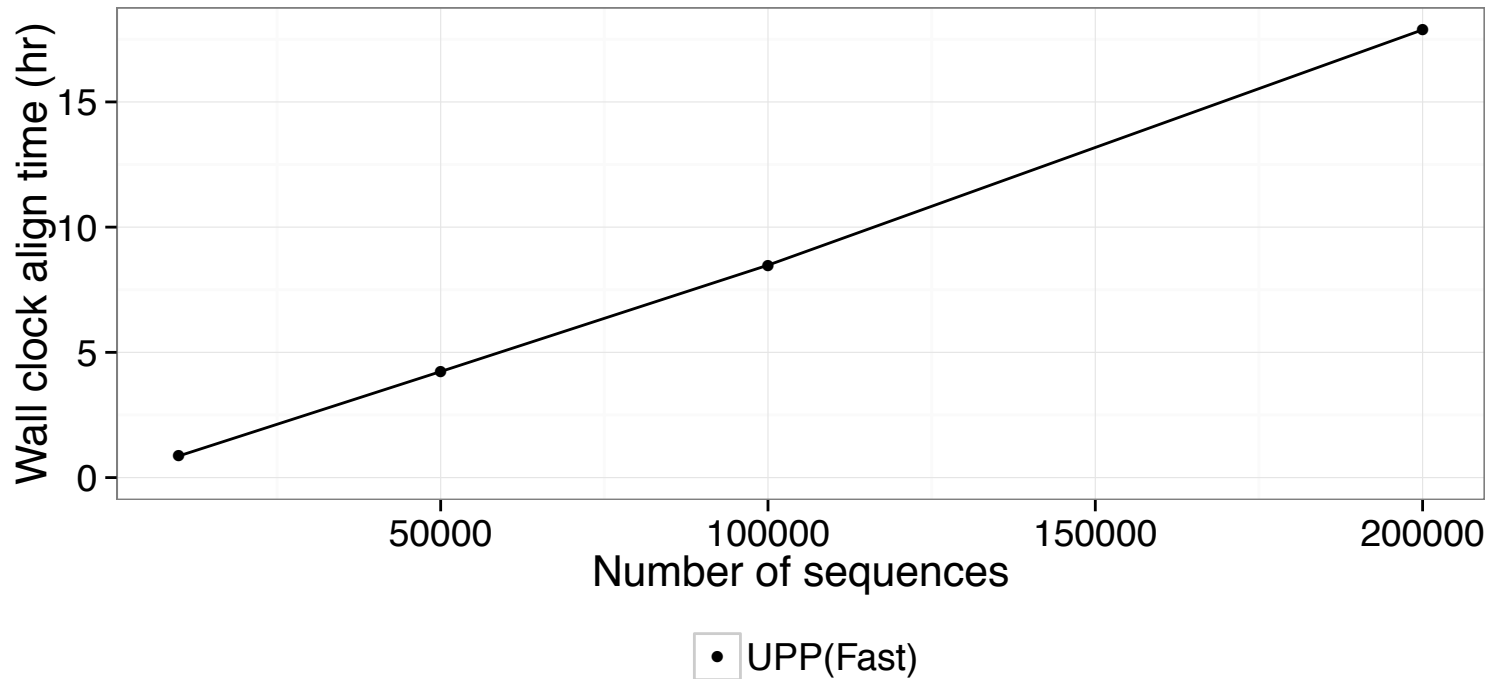
Under high rates of evolution, PASTA is badly impacted by fragmentary sequences (the same is true for other methods).

Under low rates of evolution, PASTA can still be highly accurate (data not shown).

UPP continues to have good accuracy even on datasets with many fragments under all rates of evolution.

Performance on fragmentary datasets of the 1000M2 model condition

# Running Time



Wall-clock time used (in hours) given 12 processors

# Summary

- [SATé-1](#) (Science 2009), [SATé-2](#) (Systematic Biology 2012), and [PASTA](#) (RECOMB 2014): methods for *co-estimating gene trees and multiple sequence alignments*.

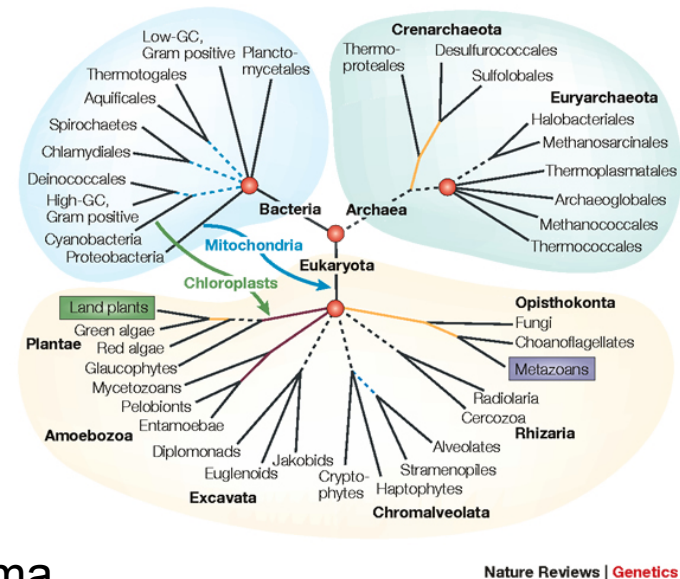
PASTA can analyze up to 1,000,000 sequences, and is highly accurate for full-length sequences. *But none of these methods are robust to fragmentary sequences.*

- [HMM Family technique](#): uses a collection of HMMs to represent a “backbone alignment”. HMM families improve accuracy, especially in the presence of high rates of evolution.
- Applications of HMM Families in:
  - [SEPP](#) (phylogenetic placement), PSB 2012
  - [TIPP](#) (metagenomic taxon identification and abundance profiling), Bioinformatics 2014
  - [UPP](#) (ultra-large multiple sequence alignment), under review

# The Tree of Life: *Multiple* Challenges

## Scientific challenges:

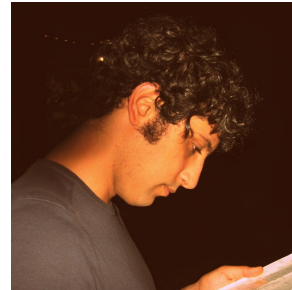
- Ultra-large multiple-sequence alignment
- Alignment-free phylogeny estimation
- Supertree estimation
- Estimating species trees from many gene trees
- Genome rearrangement phylogeny
- Reticulate evolution
- Visualization of large trees and alignments
- Data mining techniques to explore multiple optima
- Theoretical guarantees under Markov models of evolution



## Techniques:

machine learning, applied probability theory, graph theory, combinatorial optimization, supercomputing, and heuristics

# Acknowledgments



PhD students: Nam Nguyen\* and Siavash Mirarab\*\*

Undergrad: Keerthana Kumar

Lab Website: <http://www.cs.utexas.edu/users/phylo>

Personal Website: <http://tandy.cs.illinois.edu>

**Funding:** Guggenheim Foundation, 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

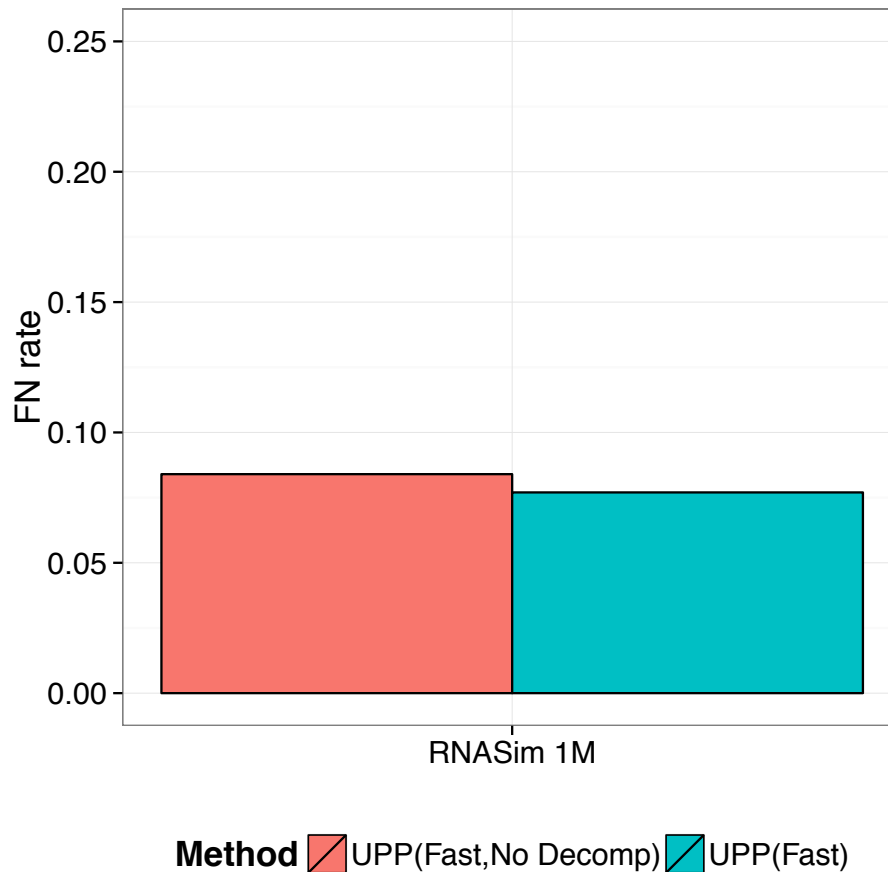
\* Now a postdoc with Becky Stumpf and Bryan White (ICB, UIUC)

\*\* Supported by HHMI Predoctoral Fellowship

# Research Projects

- Using iteration within UPP
- Using other MSA models and methods (not just HMMs) within the “Ensemble”
- Using structural alignments (or sophisticated statistical estimations of alignments) for the backbone
- Re-analyzing biological datasets
- Application to protein structure and function
- Other classification problems

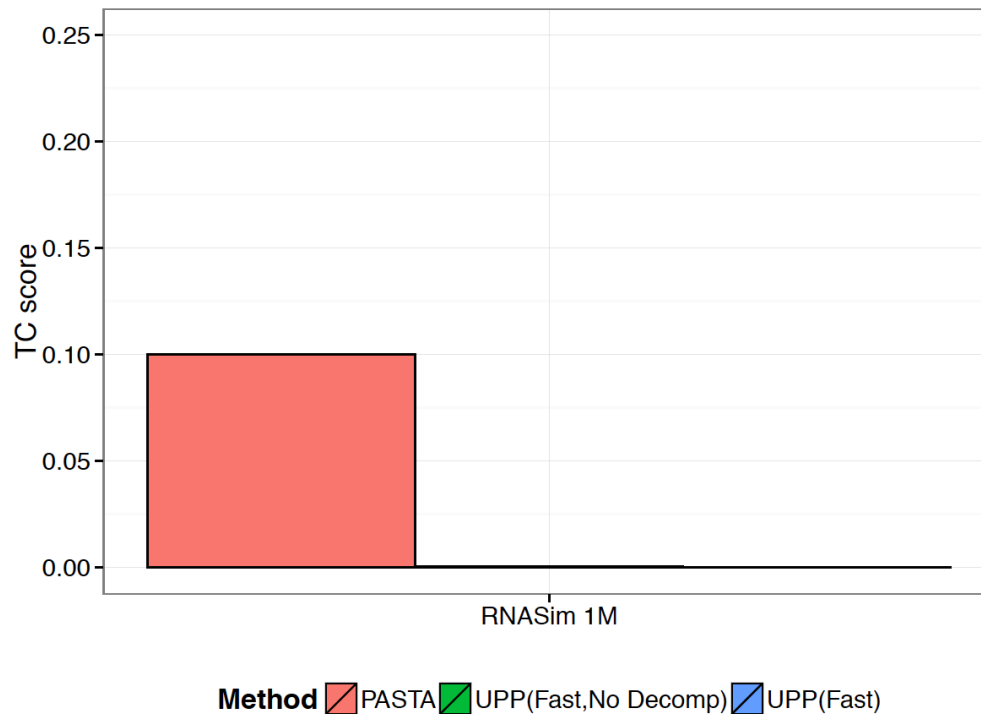
# RNASim Million Sequences: tree error



Using 12 processors:

- UPP(Fast, NoDecomp) took 2.2 days.
- UPP(Fast) took 11.9 days.

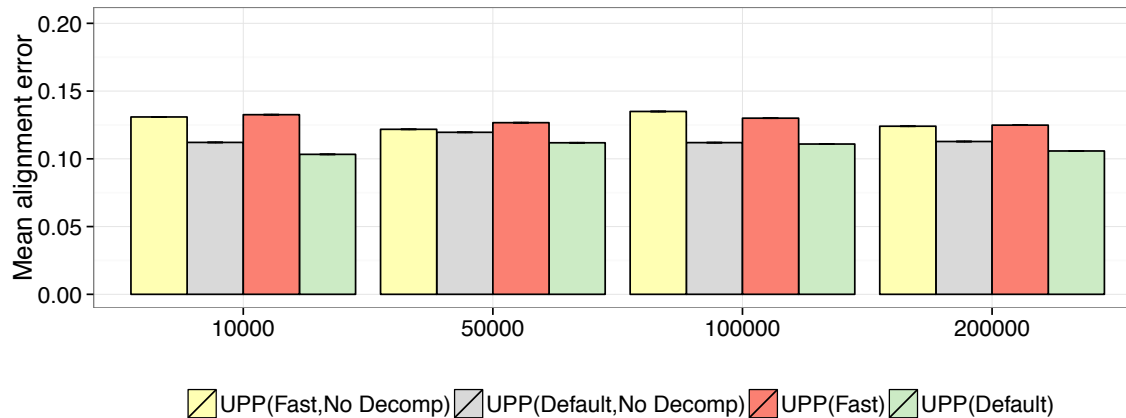
# RNASim Million Sequences: TC score



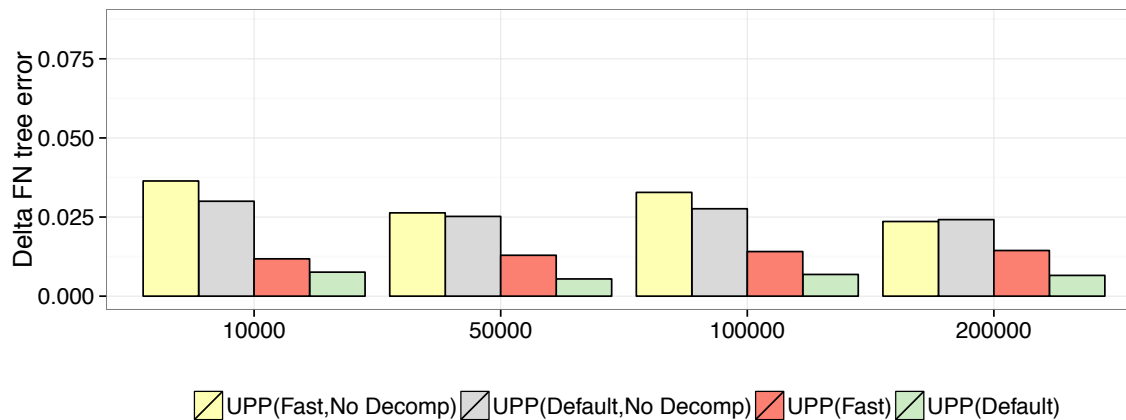
## Notes:

- UPP(Fast, NoDecomp) took 2.2 days,
- UPP(Fast) took 11.9 days, and
- PASTA took 10.3 days (all using 12 processors).

# Impact of backbone size and use of HMM Family technique



(a) Average alignment error



(b) Average tree error

Notes:

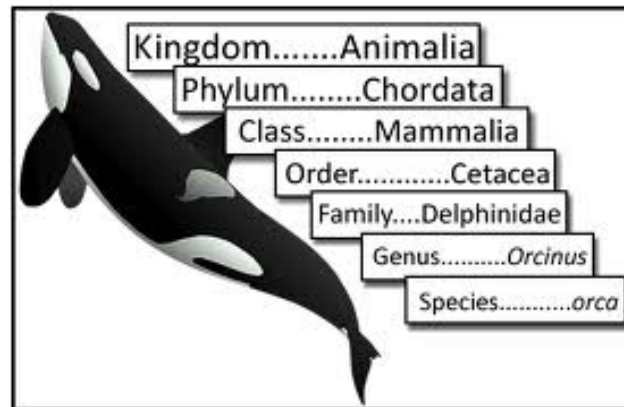
Relative performance under standard alignment criteria is not predictive of relative performance for tree estimation.

For alignment estimation, a large backbone is important.

For tree estimation, the use of the HMM Family is most important, but large backbones also help.

# Metagenomic Taxon Identification

Objective: classify short reads in a metagenomic sample



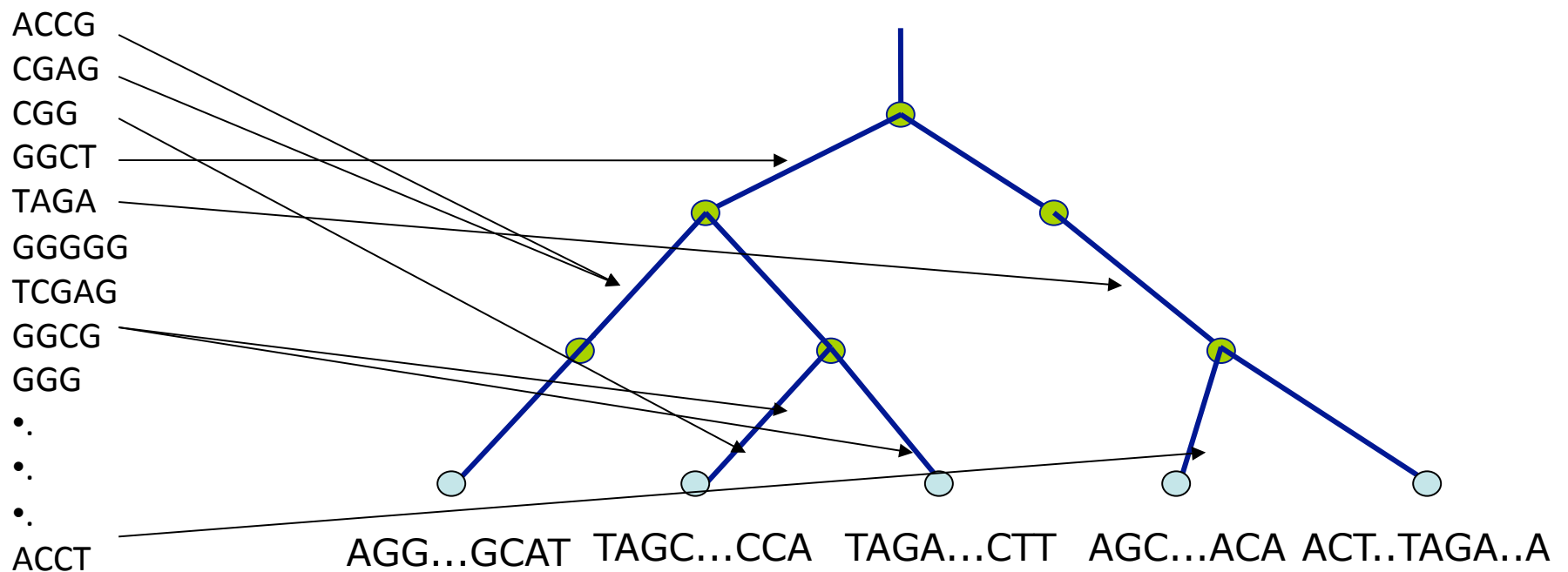
# Basic Questions

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.)
3. What are the organisms in this metagenomic sample doing together?

# Phylogenetic Placement

Fragmentary sequences  
from some gene

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



# TIPP vs. other abundance profilers

- TIPP is highly accurate, even in the presence of high indel rates and novel genomes, and for both short and long reads.
- All other methods have some vulnerability (e.g., mOTU is only accurate for short reads and is impacted by high indel rates).

# Phylogenetic “boosters”

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

Techniques: divide-and-conquer, iteration, chordal graph algorithms, and “bin-and-conquer”

Examples:

- DCM-boosting for distance-based methods (1999)
- DCM-boosting for heuristics for NP-hard problems (1999)
- [SATé- and PASTA-boosting for alignment methods \(2009, 2012, and 2014\)](#)
- SuperFine-boosting for supertree methods (2012)
- DACTAL: almost alignment-free phylogeny estimation methods (2012)
- [SEPP-boosting for phylogenetic placement of short sequences \(2012\)](#)
- [TIPP-boosting for metagenomic taxon identification \(submitted\)](#)
- [UPP-boosting for alignment methods \(in preparation\)](#)
- Bin-and-conquer for coalescent-based species tree estimation (2013 and 2014)

# Algorithmic Strategies

- Divide-and-conquer
- Chordal graph decompositions
- Iteration
- Multiple HMMs
- Bin-and-conquer (technique used for improving species tree estimation from multiple gene trees, Bayzid and Warnow, Bioinformatics 2013)

# 1kp: 1000 Plant Transcriptomes

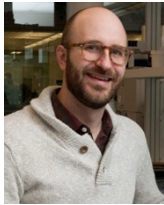
G. Ka-Shu Wong  
U Alberta



J. Leebens-Mack  
U Georgia



N. Wickett  
Northwestern



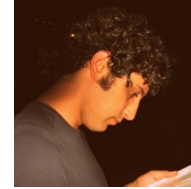
N. Matasci  
iPlant



T. Warnow,  
UT-Austin



S. Mirarab,  
UT-Austin



N. Nguyen,  
UT-Austin



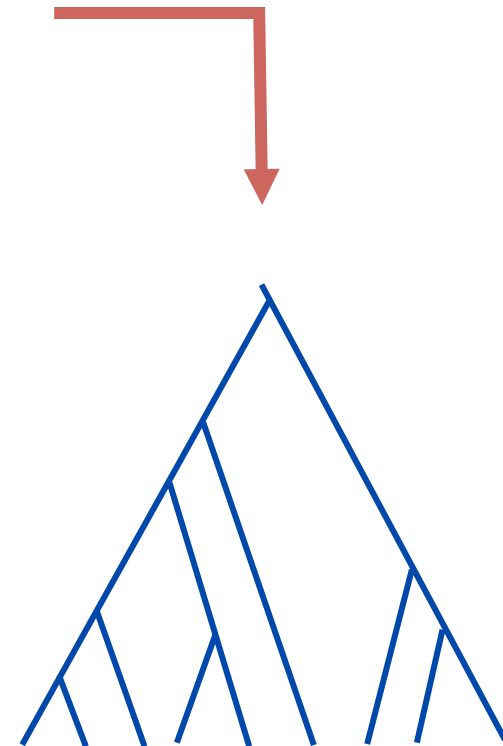
Plus many many other people...

- Whole Transcriptomes of 103 plant species and 850 single copy loci (1200 taxa in next phase)
  - Most accurate summary methods **cannot handle this size**
- Common ancestor about 1 billion years ago and so gene trees are hard to root
  - Most summary methods **need rooted gene trees**
- Pre-existing summary methods do not provide reasonable results on this dataset

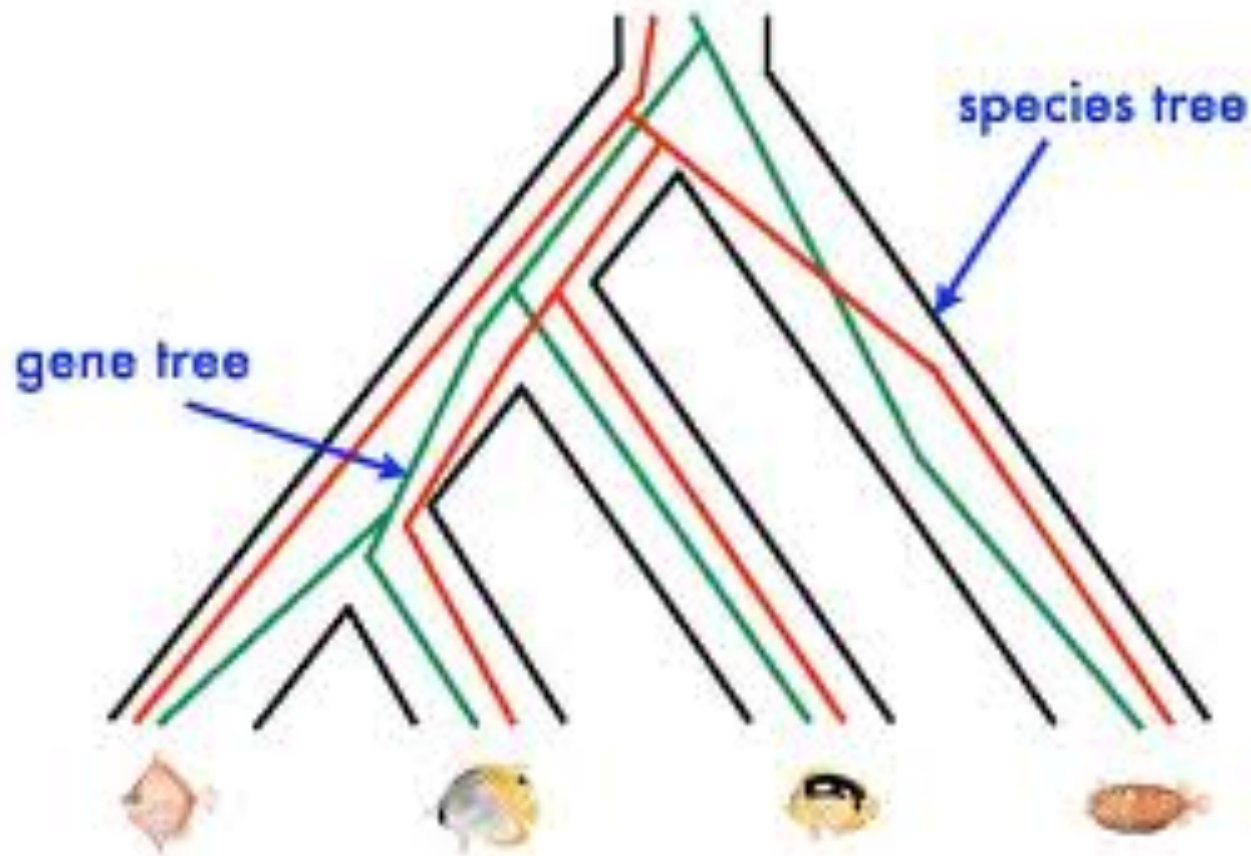
[Wickett et al. (under review), 2014.]

# Combined analysis

	gene 1	gene 2	gene 3
S <sub>1</sub>	TCTAATGGAA	??????????	TATTGATACA
S <sub>2</sub>	GCTAAGGGAA	??????????	??????????
S <sub>3</sub>	TCTAAGGGAA	??????????	TCTTGATACC
S <sub>4</sub>	TCTAACGGAA	GGTAACCCTC	TAGTGATGCA
S <sub>5</sub>	??????????	GCTAAACCTC	??????????
S <sub>6</sub>	??????????	GGTGACCATC	??????????
S <sub>7</sub>	TCTAATGGAC	GCTAAACCTC	TAGTGATGCA
S <sub>8</sub>	TATAACGGAA	??????????	CATTCATACC

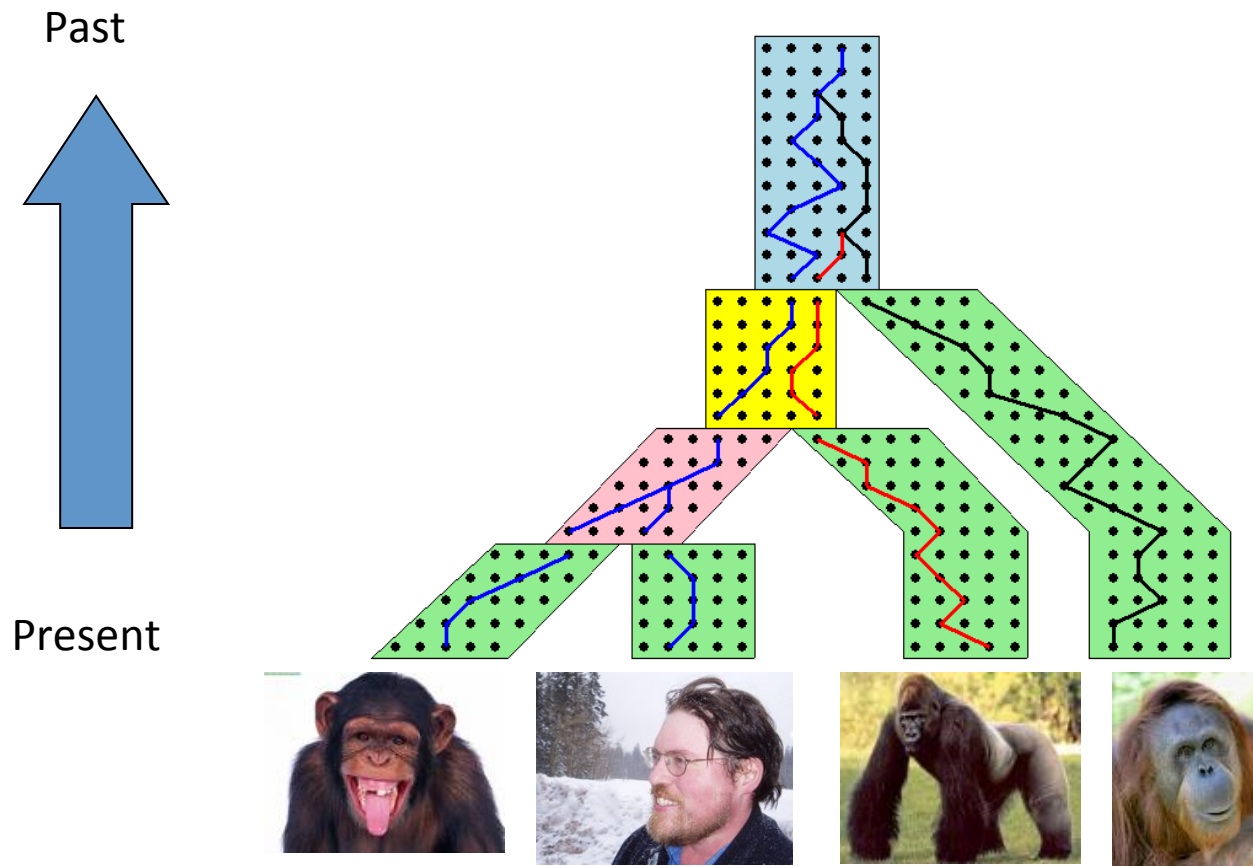


Red gene tree  $\neq$  species tree  
(green gene tree okay)



# The Coalescent

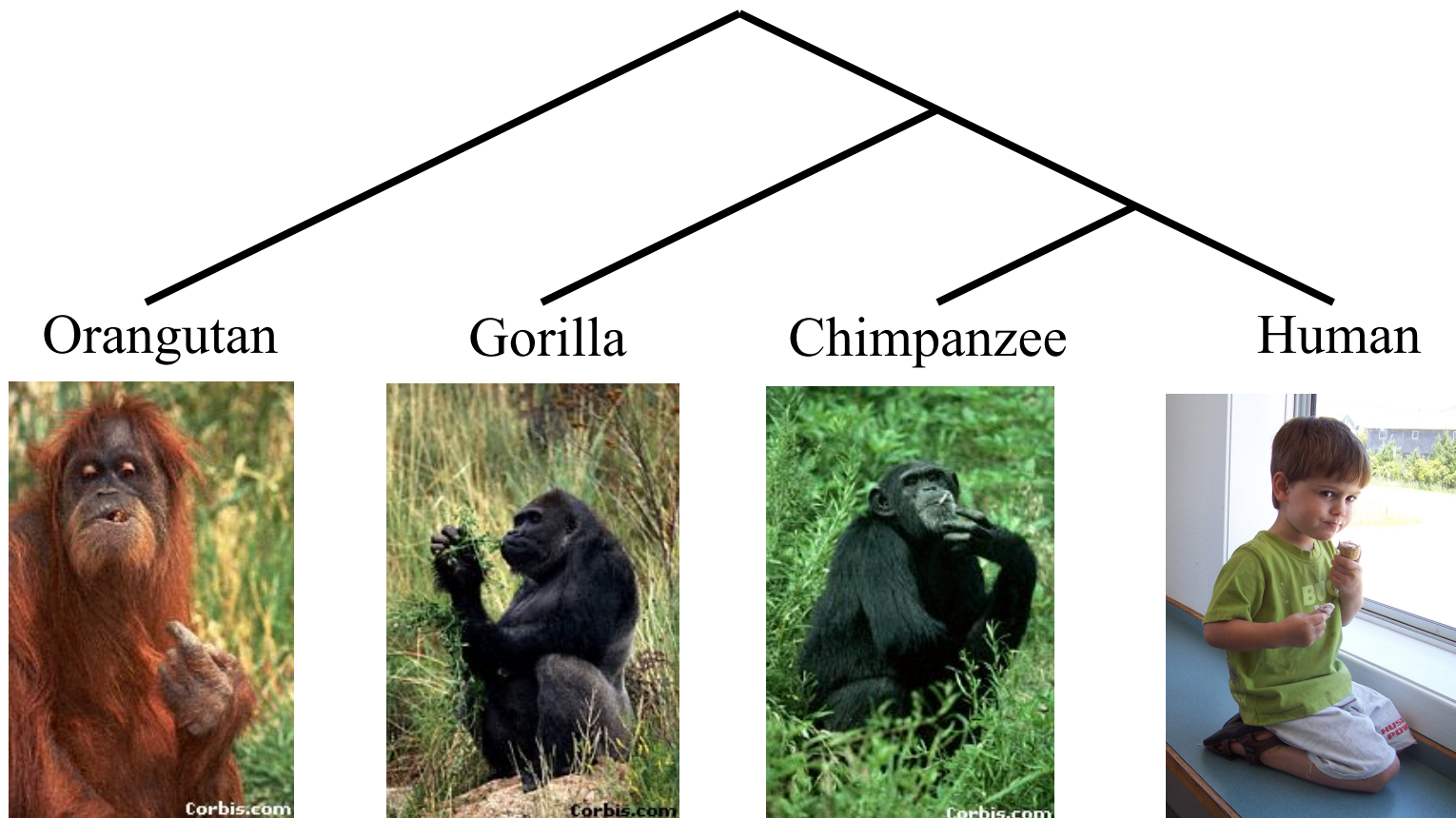
Courtesy James Degnan



# Incomplete Lineage Sorting (ILS)

- 1000+ papers in 2013 alone
- Confounds phylogenetic analysis for many groups:
  - Hominids
  - Birds
  - Yeast
  - Animals
  - Toads
  - Fish
  - Fungi
- There is substantial debate about how to analyze phylogenomic datasets in the presence of ILS.

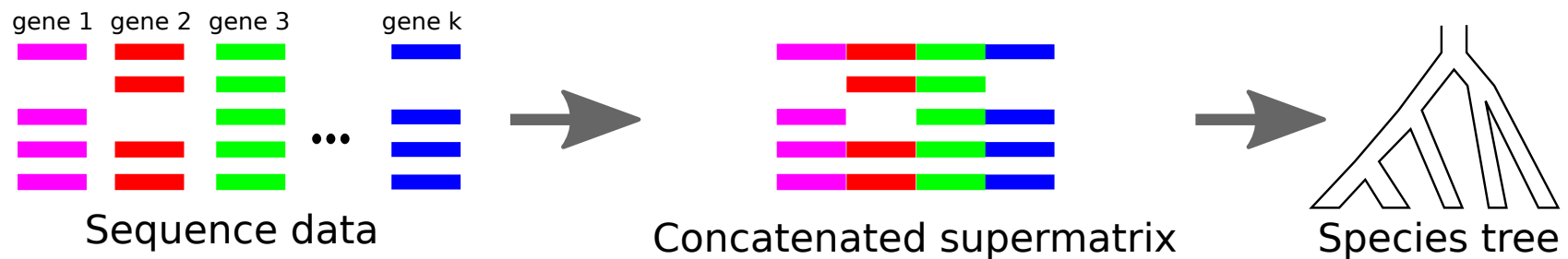
# Species tree estimation: difficult, even for small datasets



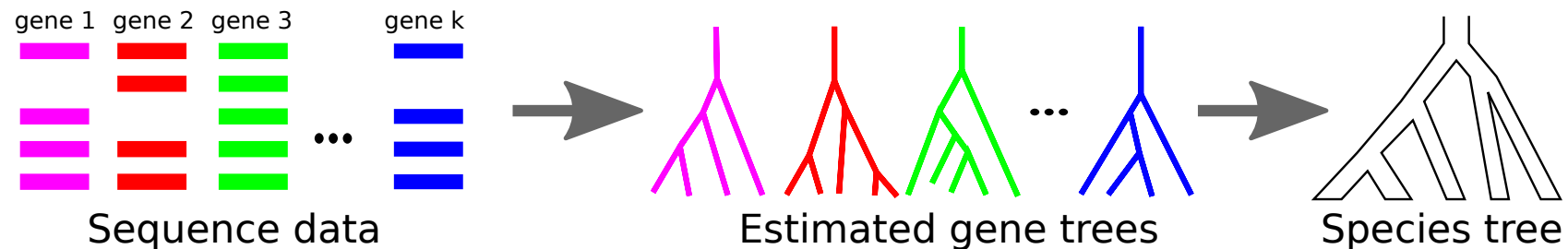
*From the Tree of the Life Website,  
University of Arizona*

# Species tree estimation

**1- Concatenation:** statistically inconsistent (Roch & Steel 2014)



**2- Summary methods:** can be statistically consistent



**3- Co-estimation methods:** too slow for large datasets

# Is Concatenation Evil?

- Joseph Heled:
  - YES
- John Gatesy
  - No
- Data needed to help understand existing methods and their limitations
- Better methods are needed

# Avian Phylogenomics Project (100+ people)

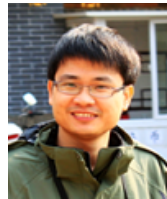
Erich Jarvis,  
HHMI



MTP Gilbert,  
Copenhagen



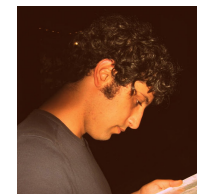
G Zhang,  
BGI



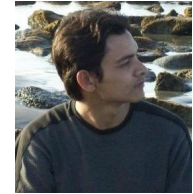
T. Warnow  
UT-Austin



S. Mirarab  
UT-Austin



Md. S. Bayzid  
UT-Austin



- Approx. 50 species, whole genomes
- 8000+ genes, UCEs
- Gene sequence alignments and trees computed using **SATé** (Science 2009, Systematic Biology 2012)
- Concatenation analysis (multi-million site) using ExaML (new version of RAxML for very long alignments)
- Massive gene tree incongruence suggestive of incomplete lineage sorting -- **coalescent-based species tree estimation computed using "Statistical Binning"** (Science, in press)

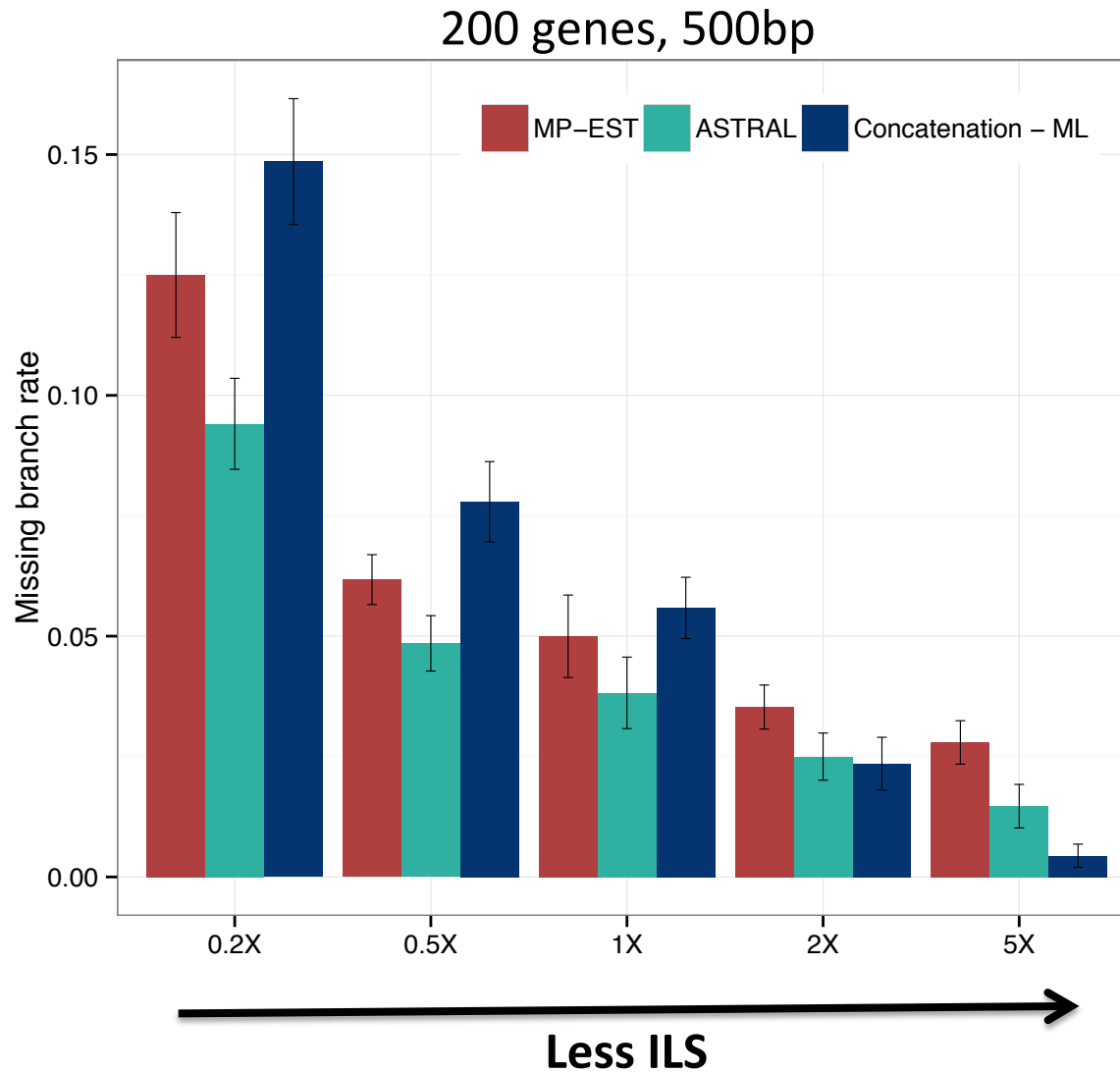
# Our methods

- “Statistical Binning” (accepted): uses a statistical method to determine sets of “combinable” gene sequence alignments, improves coalescent-based species tree estimation accuracy when gene trees have poor resolution (used for Avian Phylogenomics Project).
- ASTRAL (Bioinformatics 2014): polynomial time statistically consistent method, can run on very large datasets of unrooted gene trees.

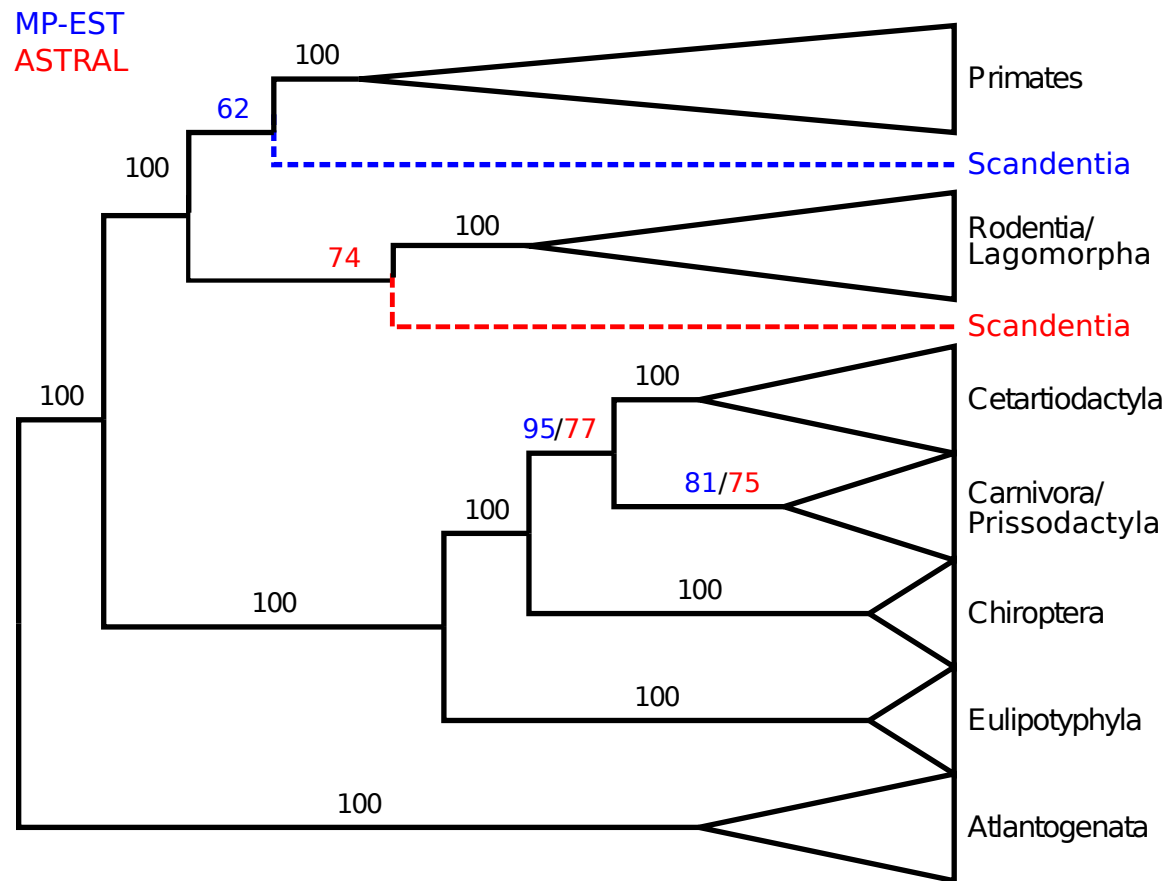
# Mammalian simulations

- Based on a biological mammalian dataset of 37 taxa and 442 genes, published by Song et al., PNAS, 2012.
- In simulations, we vary
  - Levels of ILS
  - Number of genes
  - Alignment length to control gene tree estimation error
- Compare ASTRAL to
  - MP-EST, BUCKy
  - Concatenation
  - MRP, Greedy
- Measure species tree error compared to the known true tree

# ASTRAL vs. Concatenation



# Analyses of the Song et al. Mammalian dataset

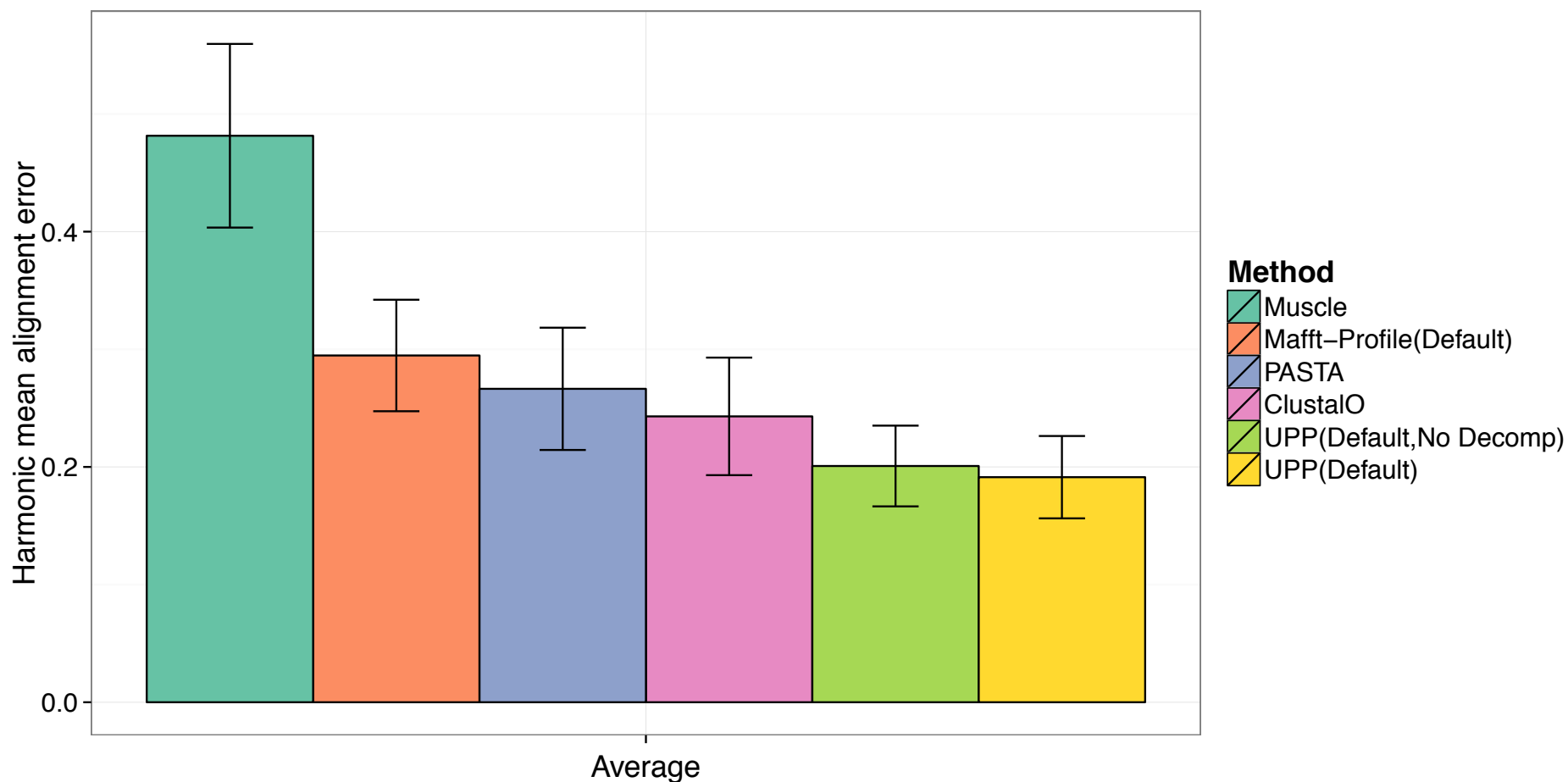


The placement of Scandentia (Tree Shrew) is controversial.  
The ASTRAL analysis agrees with maximum likelihood concatenation analysis of this dataset.

# Summary

- New multiple sequence alignments with improved accuracy and scalability, as well as high robustness to fragmentary data – SATé used in many studies.
- New coalescent-based species tree estimation methods that have better accuracy than current methods, and can analyze large datasets (used in 1KP and Avian Phylogenomics project analyses)
- New methods for metagenomic taxon identification and abundance profiling with improved accuracy (will be used in analyses here at Illinois)
- Method development inspired by collaboration with biologists, and improves biological analyses.

# AA Sequence Alignment Error (13 HomFam datasets)

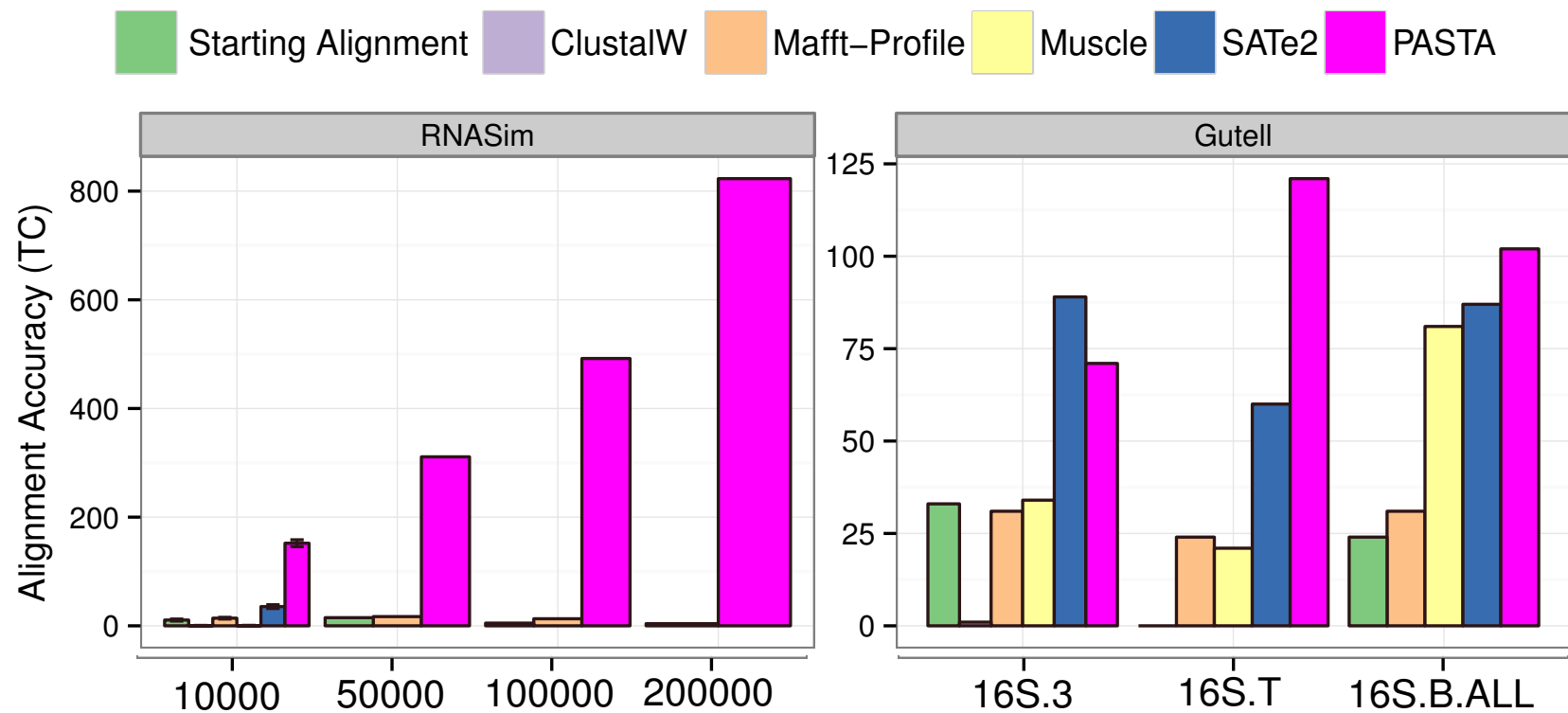


Homfam: Sievers et al., MSB 2011

10,000 to 46,000 sequences per dataset; 5-14 sequences in each seed alignment

Sequence lengths from 46-364 AA

# Alignment Accuracy – Correct columns



# Other co-estimation methods

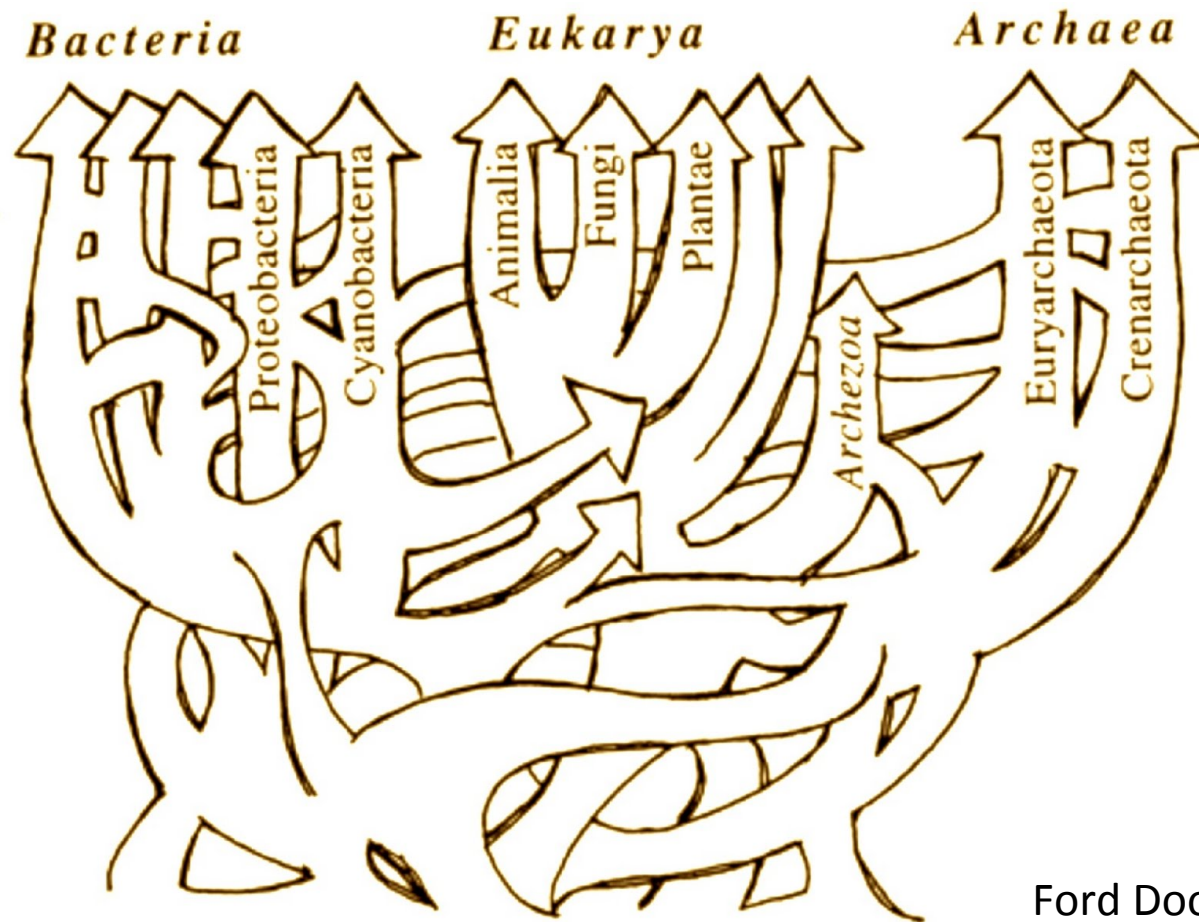
## Statistical methods:

- BAli-Phy (Redelings and Suchard): Bayesian software to co-estimate alignments and trees under a statistical model of evolution that includes indels. Can scale to about 100 sequences, but takes a very long time.
  - <http://www.bali-phy.org/>
- StatAlign: <http://statalign.github.io/>

## Extensions of Parsimony

- POY (most well known software)
  - <http://www.amnh.org/our-research/computational-sciences/research/projects/systematic-biology/poy>
- BeeTLe (Liu and Warnow, PLoS One 2012)

# Horizontal Gene Transfer – Phylogenetic Networks

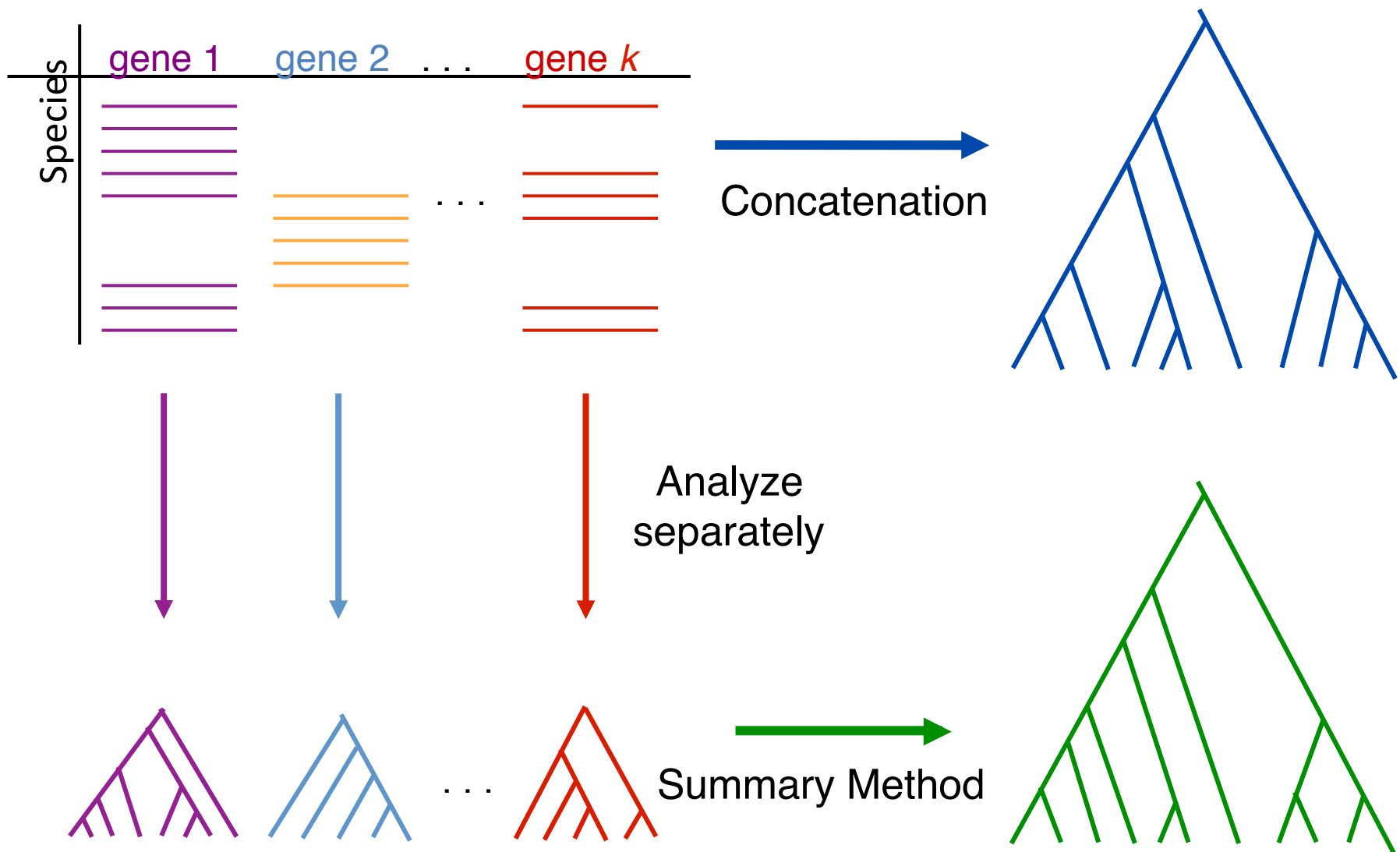


Ford Doolittle

# But...

- Gene trees may not be identical to species trees:
  - Incomplete Lineage Sorting (deep coalescence)
  - Gene duplication and loss
  - Horizontal gene transfer
- This makes combined analysis and standard supertree analyses inappropriate

# Two competing approaches



# How to compute a species tree?



# Statistically consistent under ILS?

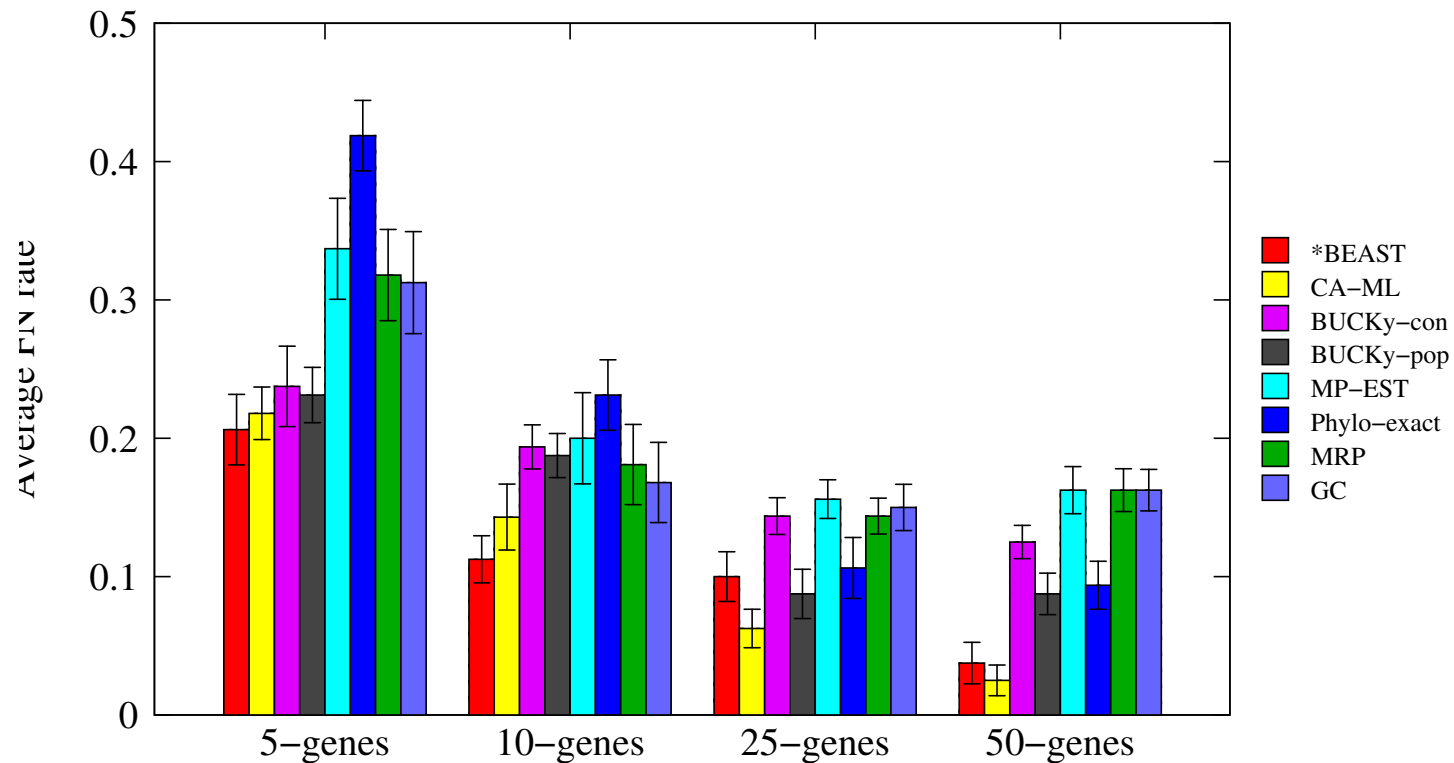
- MP-EST (Liu et al. 2010): maximum likelihood estimation of rooted species tree – YES
- BUCKy-pop (Ané and Larget 2010): quartet-based Bayesian species tree estimation –YES
- MDC – NO
- Greedy – NO
- Concatenation under maximum likelihood – NO
- MRP (supertree method) – open

# The Debate:

## Concatenation vs. Coalescent Estimation

- In favor of coalescent-based estimation
  - Statistical consistency guarantees
  - Addresses gene tree incongruence resulting from ILS
  - Some evidence that concatenation can be positively misleading
- In favor of concatenation
  - Reasonable results on data
  - High bootstrap support
  - Summary methods (that combine gene trees) can have poor support or miss well-established clades entirely
  - Some methods (such as \*BEAST) are computationally too intensive to use

# Results on 11-taxon datasets with strongILS



**\*BEAST** more accurate than summary methods (MP-EST, BUCKy, etc)  
CA-ML: (concatenated analysis) also very accurate

Datasets from Chung and Ané, 2011  
Bayzid & Warnow, Bioinformatics 2013

# Species tree/network estimation

- Methods have been developed to estimate species phylogenies (trees or networks!) from gene trees, when gene trees can conflict from each other (e.g., due to ILS, gene duplication and loss, and horizontal gene transfer).
- Phylonet (software suite), has effective methods for many optimization problems – including MDC and maximum likelihood.
- Tutorial on Wednesday.
- Software available at <http://bioinfo.cs.rice.edu/phylonet?destination=node/3>

# 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.)

# SEPP

- **SEPP: SATé-enabled Phylogenetic Placement**, by Mirarab, Nguyen, and Warnow
- Pacific Symposium on Biocomputing, 2012  
(special session on the Human Microbiome)
- Tutorial on Thursday.

# Other problems

- Genomic MSA estimation:
  - Multiple sequence alignment of very long sequences
  - Multiple sequence alignment of sequences that evolve with rearrangement events
- Phylogeny estimation under more complex models
  - Heterotachy
  - Violation of the rates-across-sites assumption
  - Rearrangements
- Estimating branch support on very large datasets