

Assigning Sequences to Taxa

CMSC828G

Outline

- ✦ Objective (1 slide)
- ✦ MEGAN (17 slides)
- ✦ SAP (33 slides)
- ✦ Conclusion (1 slide)

Objective

- ✦ Given an unknown, environmental DNA sequence:
 - ✦ Make a taxonomic assignment by comparing the sample sequence to existing database sequences that have already been taxonomically labeled*

* There is no attempt to characterize *new* species!

MEGAN — Metagenome Analyzer

- ✦ Huson *et al.* 2007
- ✦ Software that enables rapid analysis of large metagenomic data sets
- ✦ MEGAN 3 is the latest released version of the program
- ✦ Available for UNIX, Windows, and Mac OS X

MEGAN Processing Pipeline

- Reads are collected from a sample using any random shotgun sequencing protocol
- A sequence comparison of *all* reads against one or more sequence databases is performed
- MEGAN processes the results of the comparison and assigns each read to a taxon using the **lowest common ancestor (LCA) algorithm**

MEGAN Processing Pipeline

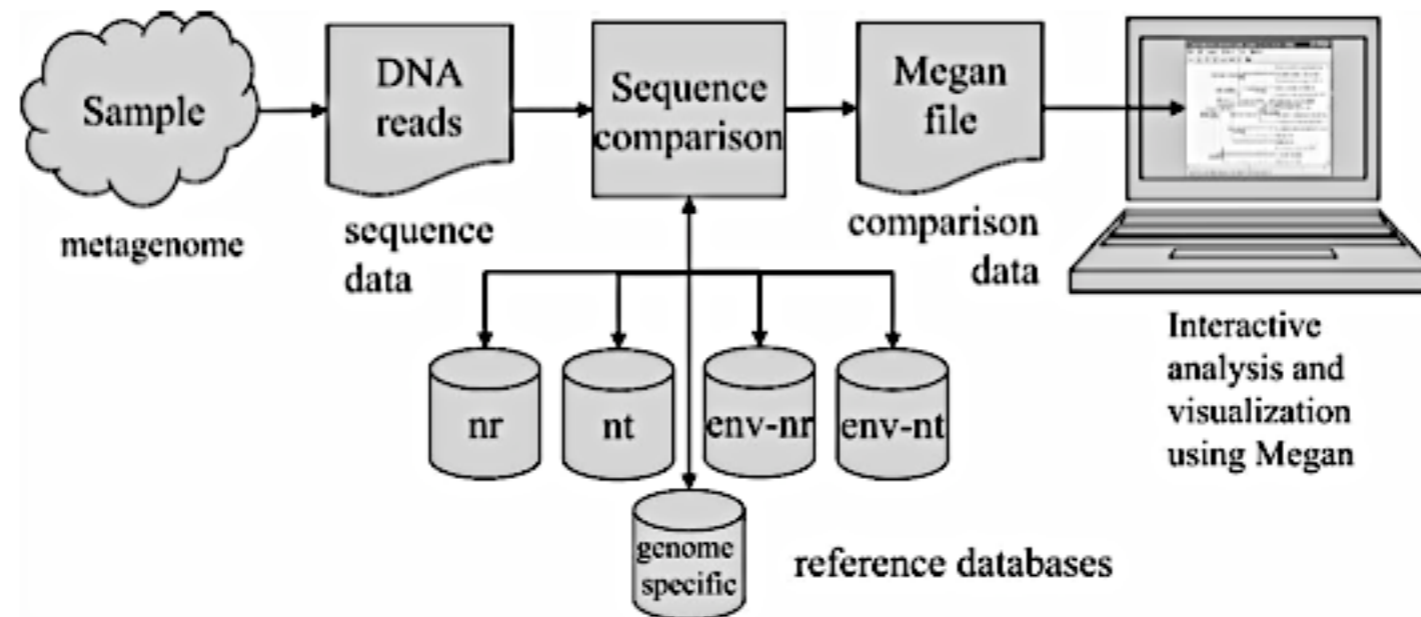


Figure 1. For a given sample of organisms, a randomly selected collection of DNA fragments is sequenced. The resulting reads are then compared with one or more reference databases using an appropriate sequence comparison program such as BLAST (Altschul et al. 1990). The resulting data are processed by MEGAN to produce an interactive analysis of the taxonomical content of the sample.

BLAST Options

- *min-score* — an alignment must achieve *min-score* to be included in the analysis
- *top-percent* — retain only those matches whose score is within *top-percent* of the highest score
- *win-score* — if a match scores above *win-score*, only consider other matches above *win-score*
- *min-support* — at least *min-support* reads must be assigned to a taxon for those assignments to count

LCA Algorithm

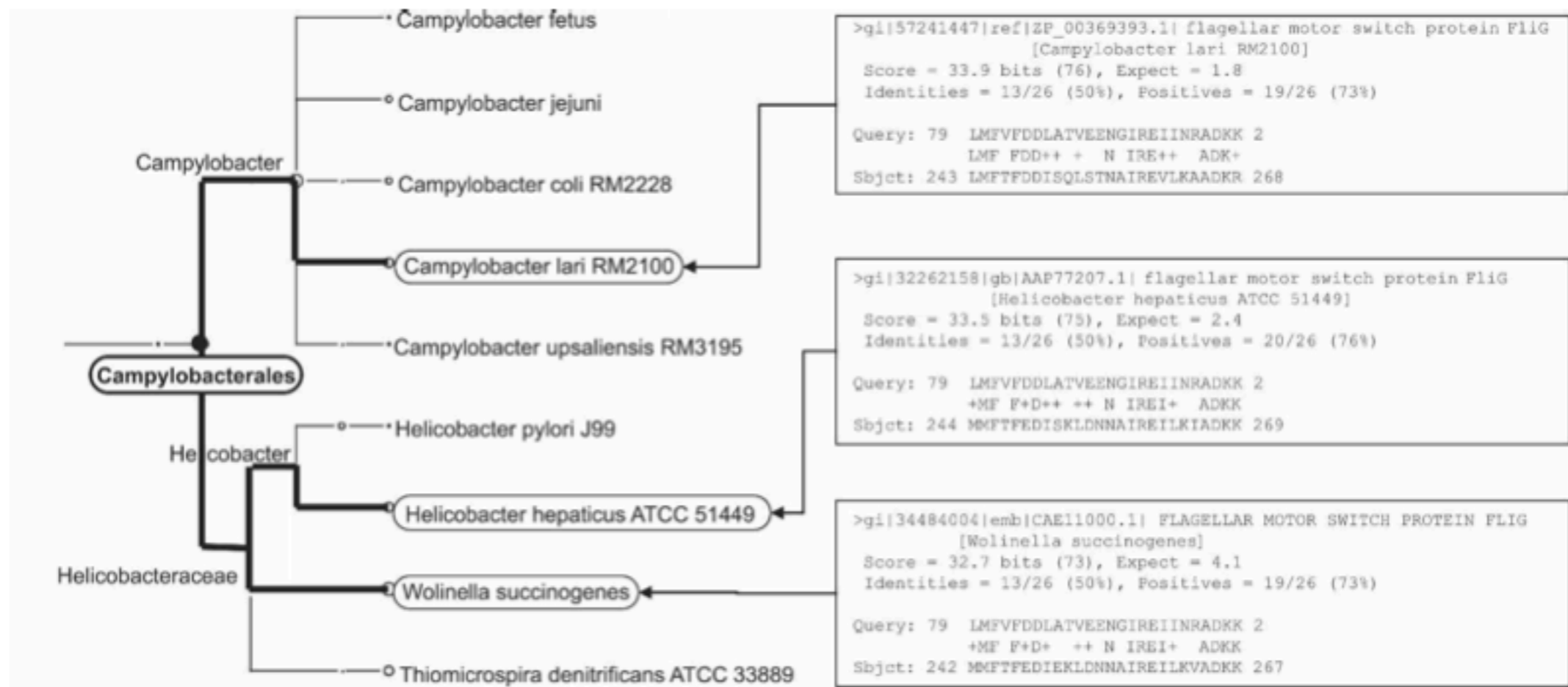


Figure 2. On the *right*, we list the three BLASTX matches obtained for a specific read *r* from the mammoth data set, to sequences representing *Campylobacter lari*, *Helicobacter hepaticus*, and *Wolinella*, respectively. The LCA-assignment algorithm assigns *r* to the taxon *Campylobacterales*, shown on the *left*, as it is the lowest-common taxonomical ancestor of the three matched species.

Data Analyses with MEGAN

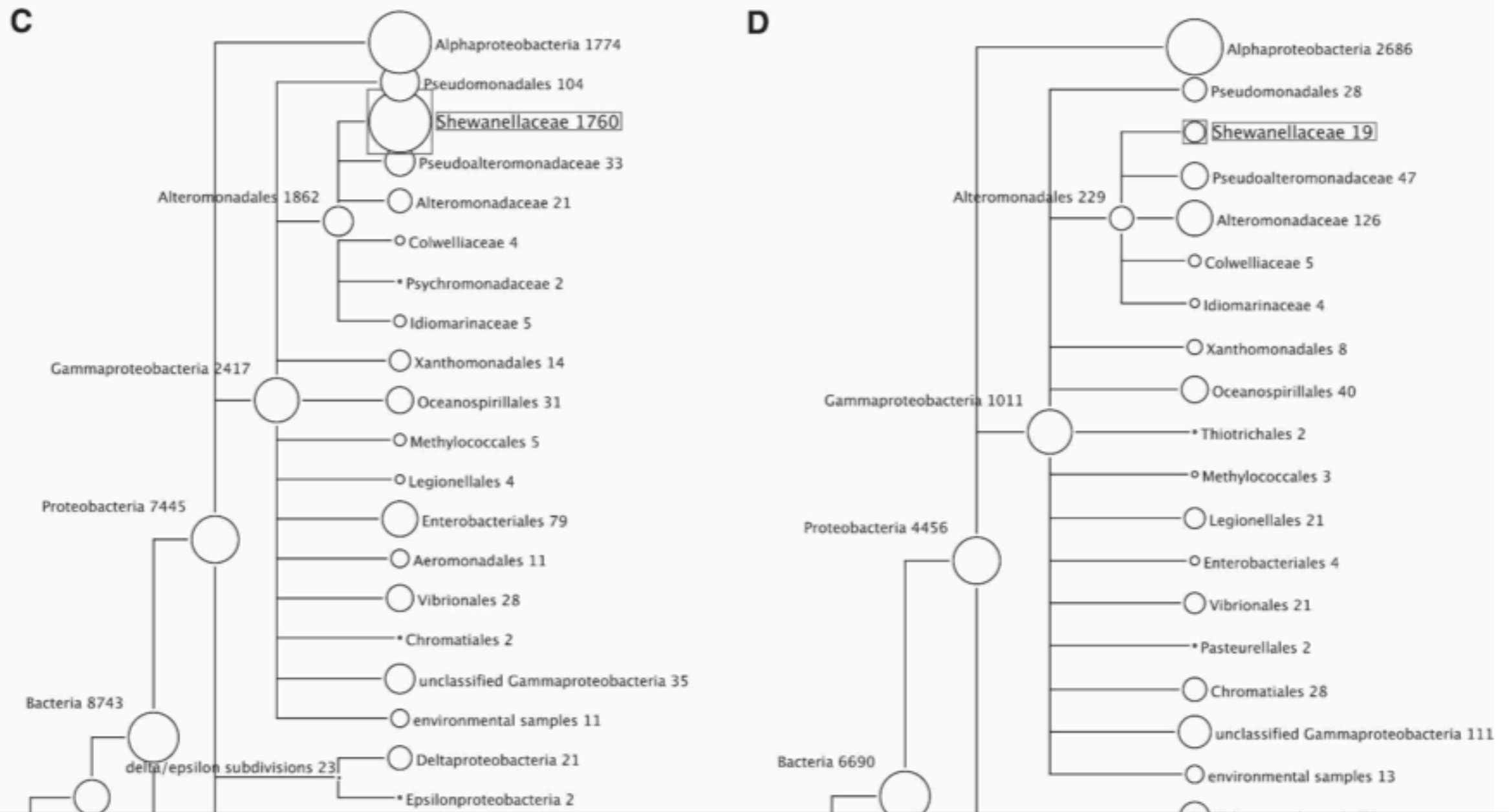
- ✦ Sargasso Sea data set
- ✦ Mammoth data set
- ✦ Species identification from short reads
 - ✦ *E. coli* K12
 - ✦ *B. bacteriovorus* HD100

Sargasso Sea Data Set

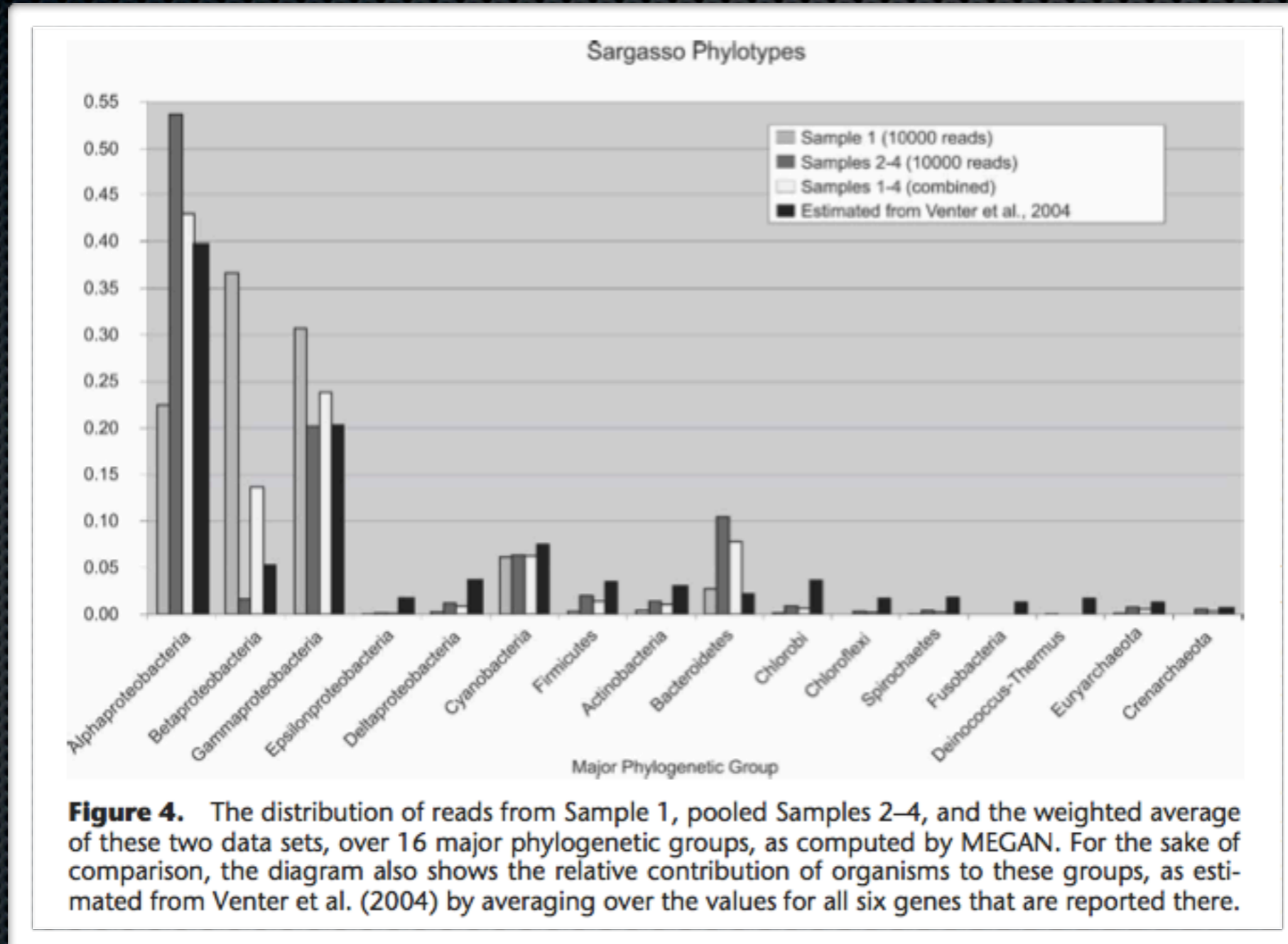


- ✦ Venter *et al.* 2004
- ✦ Samples of seawater were collected, and organisms of size 0.1–3 μm were extracted and sequenced
- ✦ From four individual sampling sites, ~1.66 million reads of average length 818 bp were recovered
- ✦ Biological diversity and abundance were measured using environmental assemblies, and by analyzing six phylogenetic markers (rRNA, RecA/RadA, HSP70, RpoB, EF-Tu, and Ef-G)

Revealing “Microheterogeneity”



Distribution of Species Comparison



Mammoth Data Set



- ✦ Poinar *et al.* 2006
- ✦ 1g bone sample taken from a mammoth that was preserved in permafrost for 28,000 years
- ✦ Obtained 302,692 reads of mean length 95 bp
- ✦ BLASTZ was used to determine reads that came from the mammoth genome, and BLASTX was used to characterize the remaining environmental diversity

Mammoth Data Set Summary

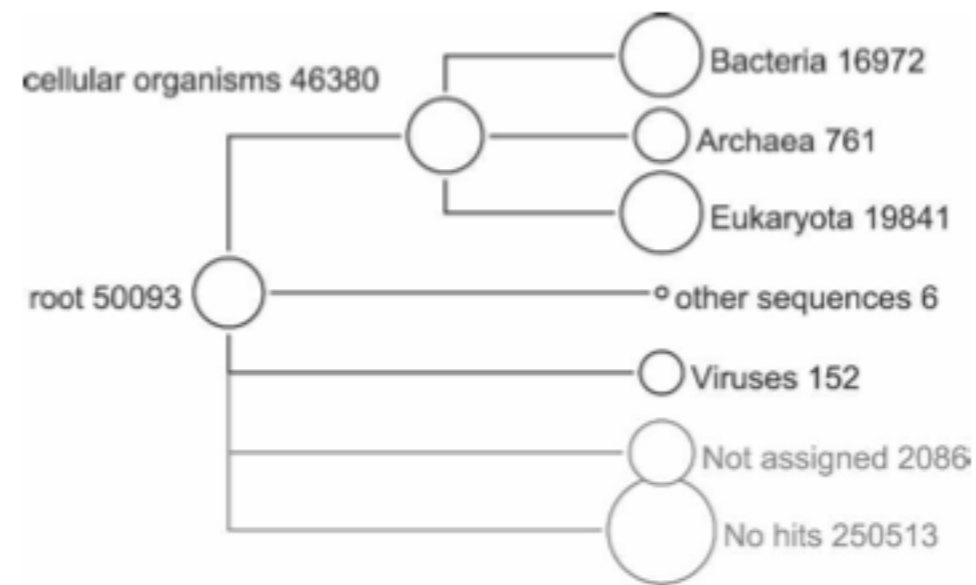


Figure 5. High-level summary of a MEGAN analysis of the mammoth data set, based on a BLASTX comparison of the 302,692 reads against the NCBI-NR database.

Bit score threshold of 30, discarding any isolated assignments

Species Identification from Short Reads

- ✦ What is the minimum read length required to identify species in a metagenomic sample?
- ✦ Idea: simulate short reads from a known genome, and then evaluate accuracy of assignments
- ✦ Two organisms were chosen for this purpose—*E. coli*, and *B. bacteriovorus*
 - ✦ These two organisms were also randomly *resequenced* (and then subsequently analyzed)

E. coli Simulation Results

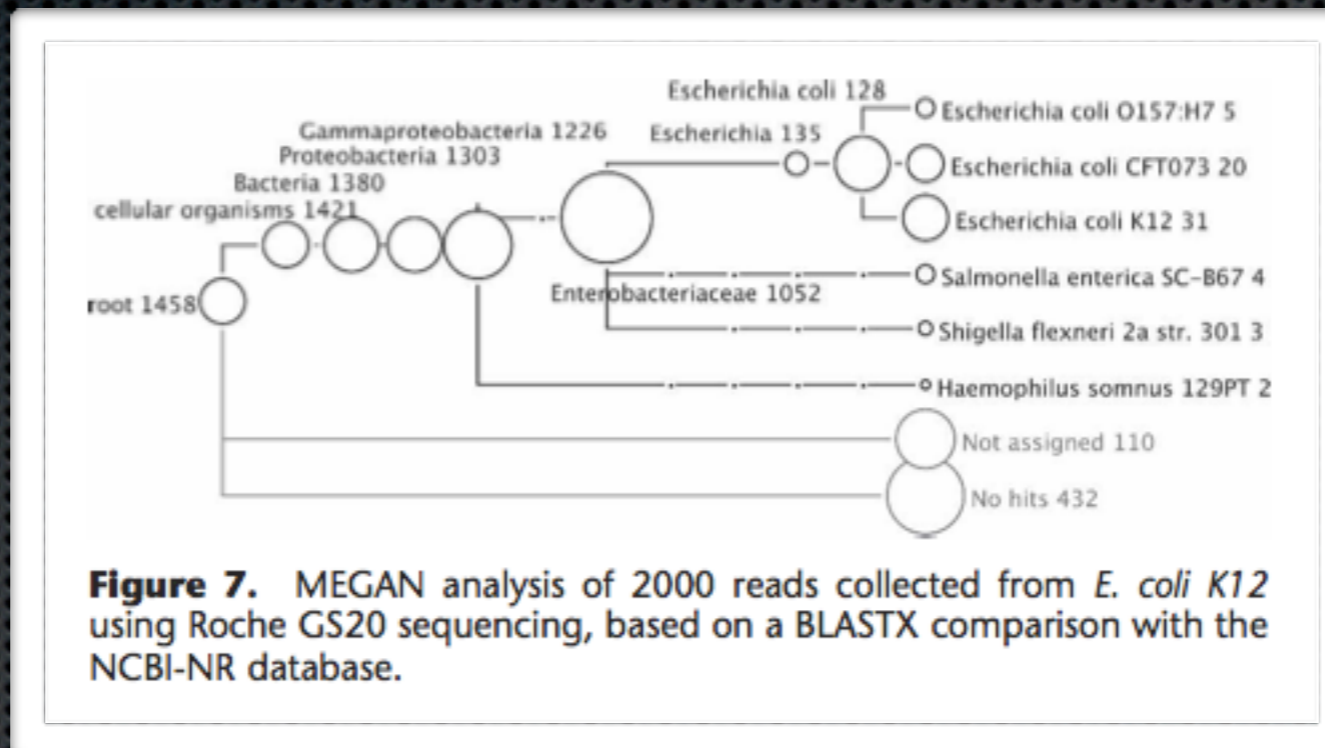
Table 1. Results for *E. coli* simulation

	35 bp	100 bp	200 bp	800 bp
Enterobacteriaceae	22%	64%	73%	85%
Gammaproteobacteria	24%	77%	86%	94%
Proteobacteria	25%	83%	89%	96%

For average read lengths of 35, 100, 200, and 800 bp, we sampled 5000 sequence intervals from random locations in the complete genome sequence of *E. coli* K12 and then processed the reads with MEGAN. Here we report the percentage of reads classified as Enterobacteriaceae, Gammaproteobacteria, and, even more generally, Proteobacteria. The number of false-positive assignments of reads was ~0%.

Basically no false positives

E. coli Resequencing Results



A few false positives

B. bacteriovorus Simulation Results

Table 2. Results for *B. bacteriovorus* simulation

	35 bp	100 bp	200 bp	800 bp
<i>B. bacteriovorus</i>	25%	88%	94%	98%
Deltaproteobacteria	26%	89%	95%	99%
Proteobacteria	26%	90%	97%	~100%

For average read lengths of 35, 100, 200, and 800 bp, we sampled 5000 sequence intervals from random locations in the complete genome sequence of *B. bacteriovorus* HD100 and then processed the reads with MEGAN. Here, we report the percentage of reads classified as *B. bacteriovorus*, Deltaproteobacteria, and, even more generally, Proteobacteria. The number of false-positive assignments of reads was ~0%.

Basically no false positives

B. bacteriovorus Resequencing Results

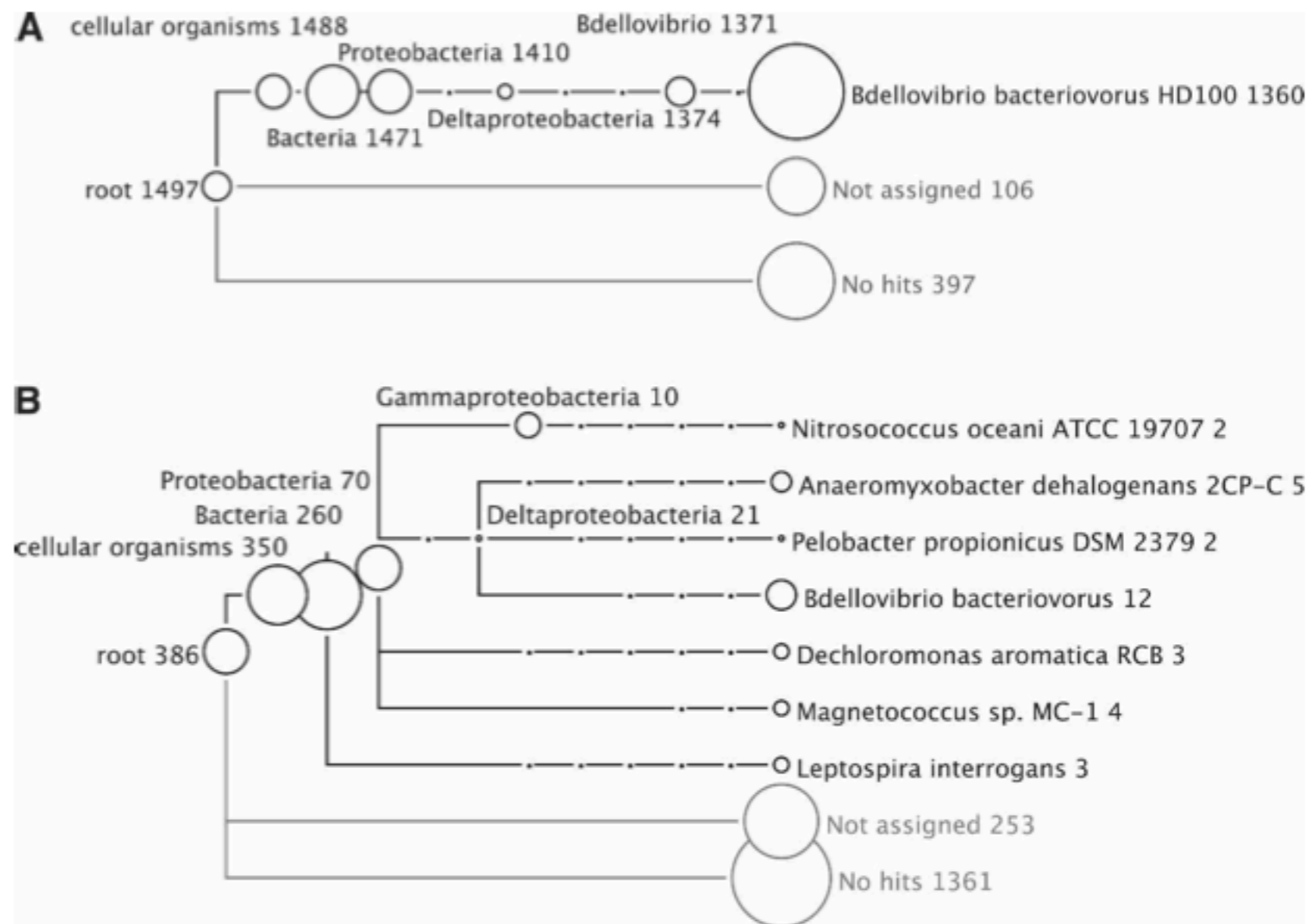


Figure 8. MEGAN analysis of 2000 reads collected from *B. bacteriovorus* HD100 using Roche GS20 sequencing. (A) Analysis based on a BLASTX comparison with NCBI-NR. (B) The same analysis, but with all hits matching database sequences representing the *B. bacteriovorus* HD100 genome removed, mimicking the situation in which the reads originate from a genome that is not represented in NCBI-NR.

MEGAN, in Summary

- ✦ LCA algorithm is simple and conservative
- ✦ Does not make many false positive assignments, even when the unknown sample sequence does not exist in the database
- ✦ Species can be identified from short reads
- ✦ Most of the work has been in developing easy to use software with useful exploratory features and visualizations, many of which were not mentioned

Limitations of BLAST

- ✦ BLAST searches use *local* alignments, not global alignments, which leads to a loss of information
- ✦ BLAST searches do not consider the population genetic and phylogenetic issues associated with species identification
- ✦ The measures of confidence associated with BLAST searches (E-values) represent significance of local similarity, not significance of taxonomic assignment

SAP — Statistical Assignment Package

- ✦ Munch *et al.* 2008
- ✦ SAP is an automated method for [DNA barcoding](#) which includes database sequence retrieval, alignment, and phylogenetic analysis
- ✦ Most importantly, provides statistically meaningful measures of confidence
- ✦ Like MEGAN, does not attempt to identify new species

SAP - An Overview

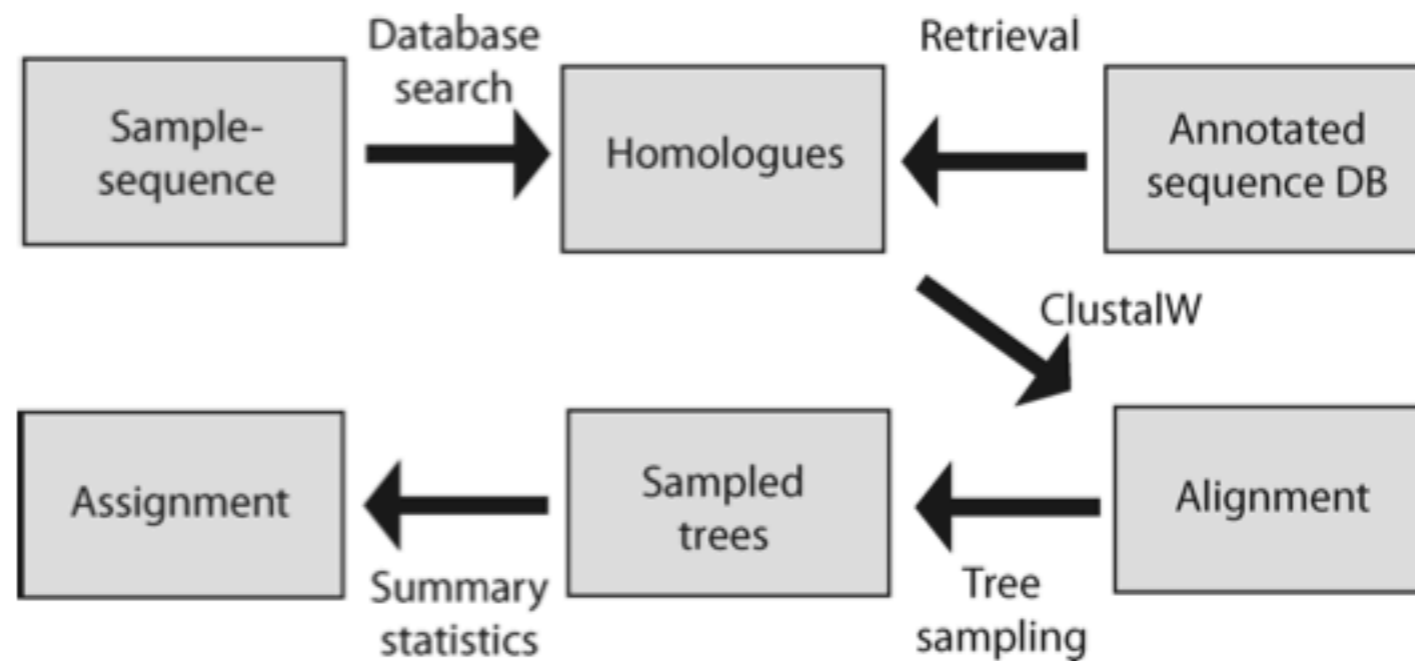


FIGURE 1. Flowchart of the assignment procedure. A set of homologues is compiled using information from Blast searches and annotation from NCBI's Taxonomy database. The relevant sequences are retrieved from GenBank and aligned using ClustalW. Based on the resulting multiple alignment a large number of phylogenetic trees are sampled and these are then used to calculate posterior probabilities of assignment.

Bayesian Approach

- Estimate the probability the sample sequence is part of a monophyletic group of database sequences

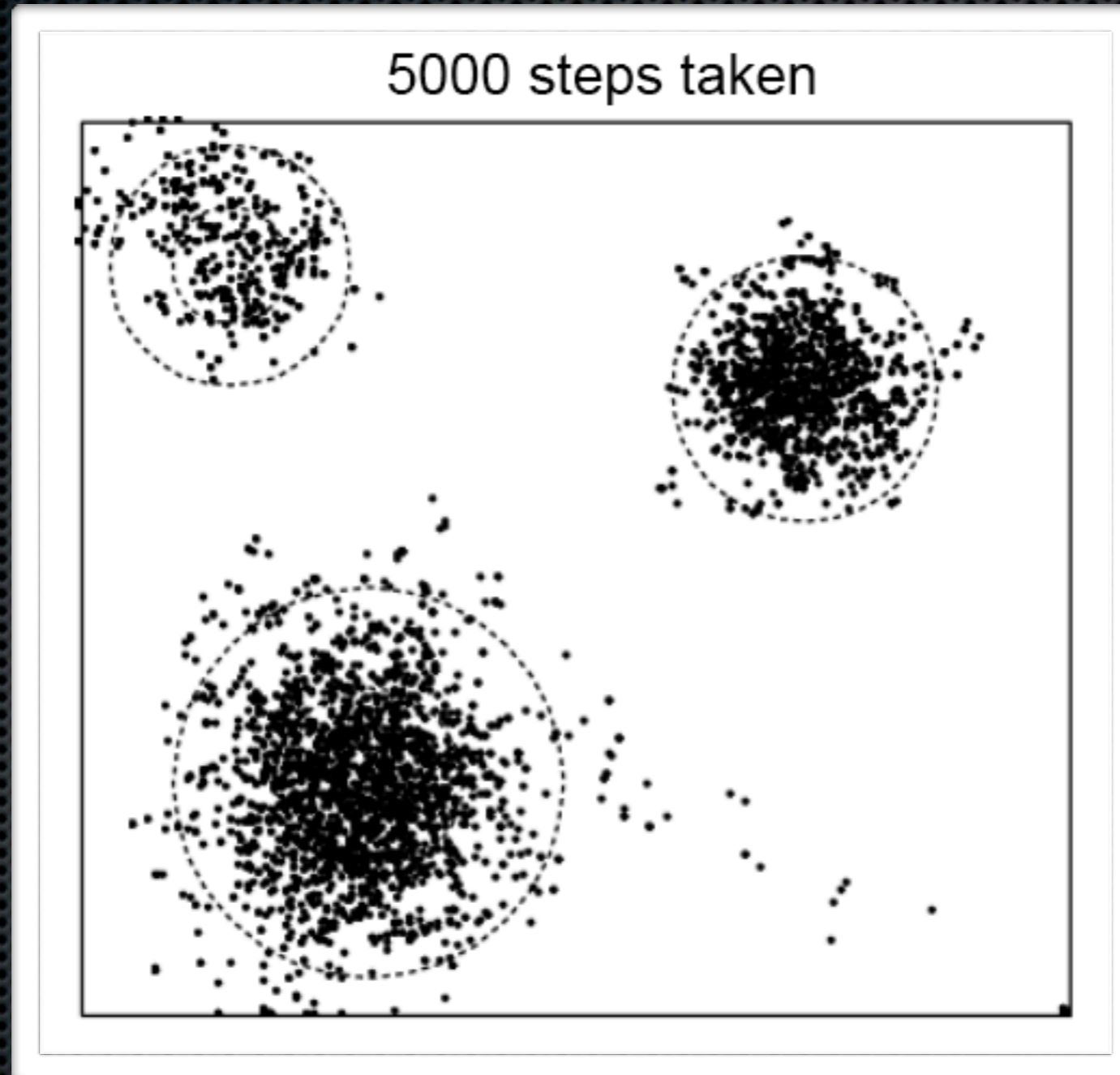
$$P(X \in T_i | X, \mathbf{D}) = \frac{P(X, \mathbf{D} | X \in T_i)P(X \in T_i)}{\sum_{j=1}^k P(X, \mathbf{D} | X \in T_j)P(X \in T_j)}$$

- X is the sample-sequence, T_i is taxon i , and \mathbf{D} is the set of database sequences representing k disjoint groups

Computing the Posterior Probability

- ✦ The posterior probability involves a summation over all possible phylogenetic trees, and for each tree, a multiple integral over all combinations of evolutionary model parameters
- ✦ Hence, the posterior probability cannot be computed analytically, even for small trees
- ✦ However, a method called Markov Chain Monte Carlo (MCMC) can be used to sample trees *in proportion to* their posterior probabilities

Sampling the Posterior Distribution



Finding Homologs

- ✦ Ideally, each sample sequence would be compared with *all* database sequences
- ✦ Instead, a heuristic is required to extract a limited representation of the database
- ✦ Thus, SAP uses BLAST to find database homologs

Finding Homologs, Method

- ✦ Include only matches whose BLAST score is *at least half* that of the best match (*relative cutoff*)
- ✦ Include only the best match from each species
- ✦ Include up to 30 species homologs, 10 genera, 6 families, 5 orders, 3 classes, and 2 phyla
- ✦ If the *relative cutoff* has been reached before 50 homologs have been included, allow other representatives from species already included

MSA and Phylogenetic Analysis

- ✦ The sample sequence and the set of homologs are aligned using ClustalW
- ✦ A program, likely some kind of MrBayes kernel, performs the Bayesian phylogenetic analysis
- ✦ All sequences except the sample sequence are topologically constrained to agree with the NCBI taxonomy
- ✦ 10,000 trees are sampled from the posterior distribution and analyzed to obtain probabilities of assignment to all taxa in the set of homologs

Taxonomic Assignment

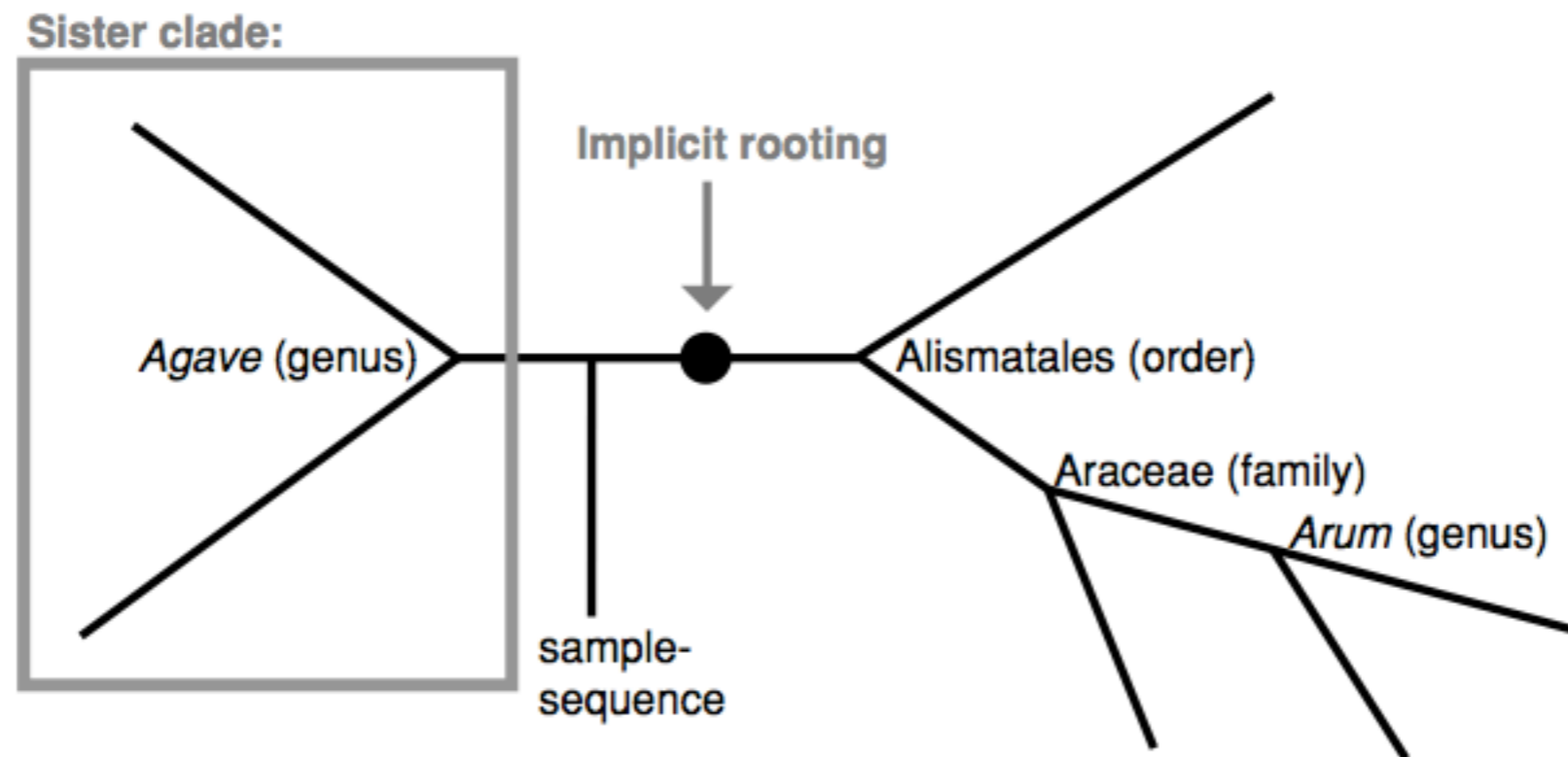


FIGURE 2. Assignment of the sample sequence in each sampled tree is done by assuming the root implied by the taxonomic annotation of homologues and then recording the consensus taxonomy for all members of the sister clade from the highest taxonomic level to the most specific level shared by all clade members.

The probability of forming a monophyletic group with a given taxon is calculated as the fraction of sampled trees where the sister clade to the sample sequence is a member of that taxon.

Probabilities of Assignment

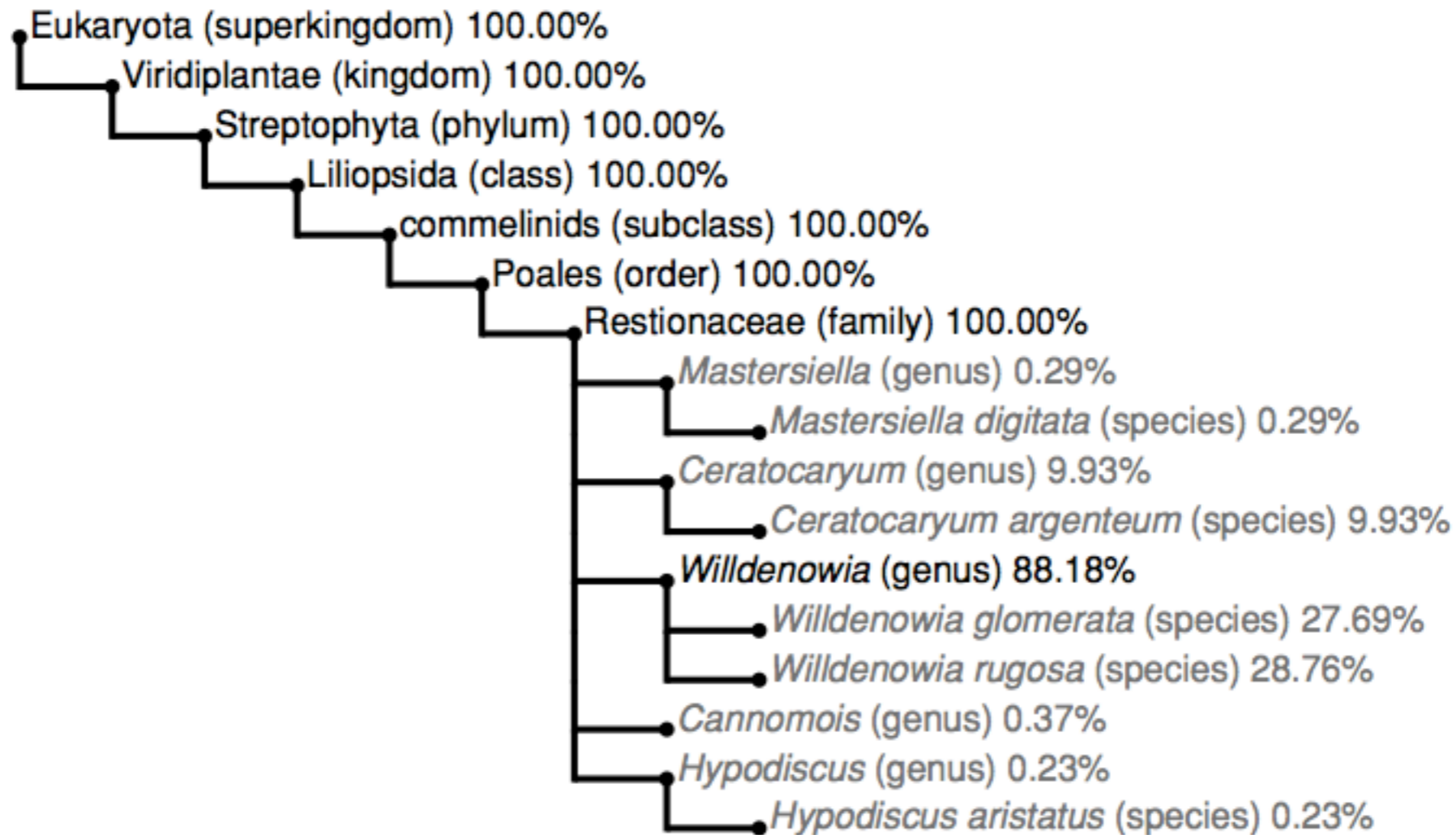


FIGURE 3. Graphic representation of assignment. The taxonomic tree shows all taxa obtaining positive probabilities of assignment. For clarity, assignment probabilities below 50% are shaded. In the example shown, sequence evidence is substantial but too ambiguous to allow a reliable assignment at the species and genus level. The evidence at family level, however, is decisive.

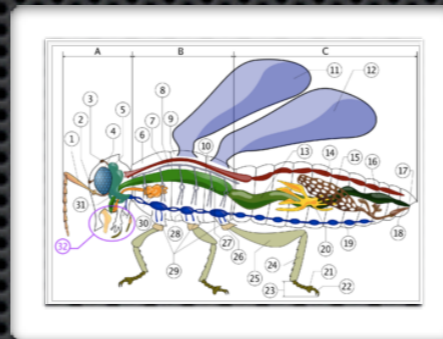
Computational Time

- ✦ Takes time to download sequences from GenBank
- ✦ Multiple alignment is fast, a couple of minutes
- ✦ The MCMC analysis is the bottleneck, averaging 1 hour
- ✦ Post-processing of MCMC output may take 10 minutes
- ✦ (and this is for each sample sequence!)

Benchmark Analyses

- ✦ Cytochrome Oxidase I (COI) gene for the class *Insecta*

- ✦ 10,804 sequences



- ✦ tRNA-Leu (trnL) gene for the class *Liliopsida* (monocots)

- ✦ 640 sequences



Benchmarking Results

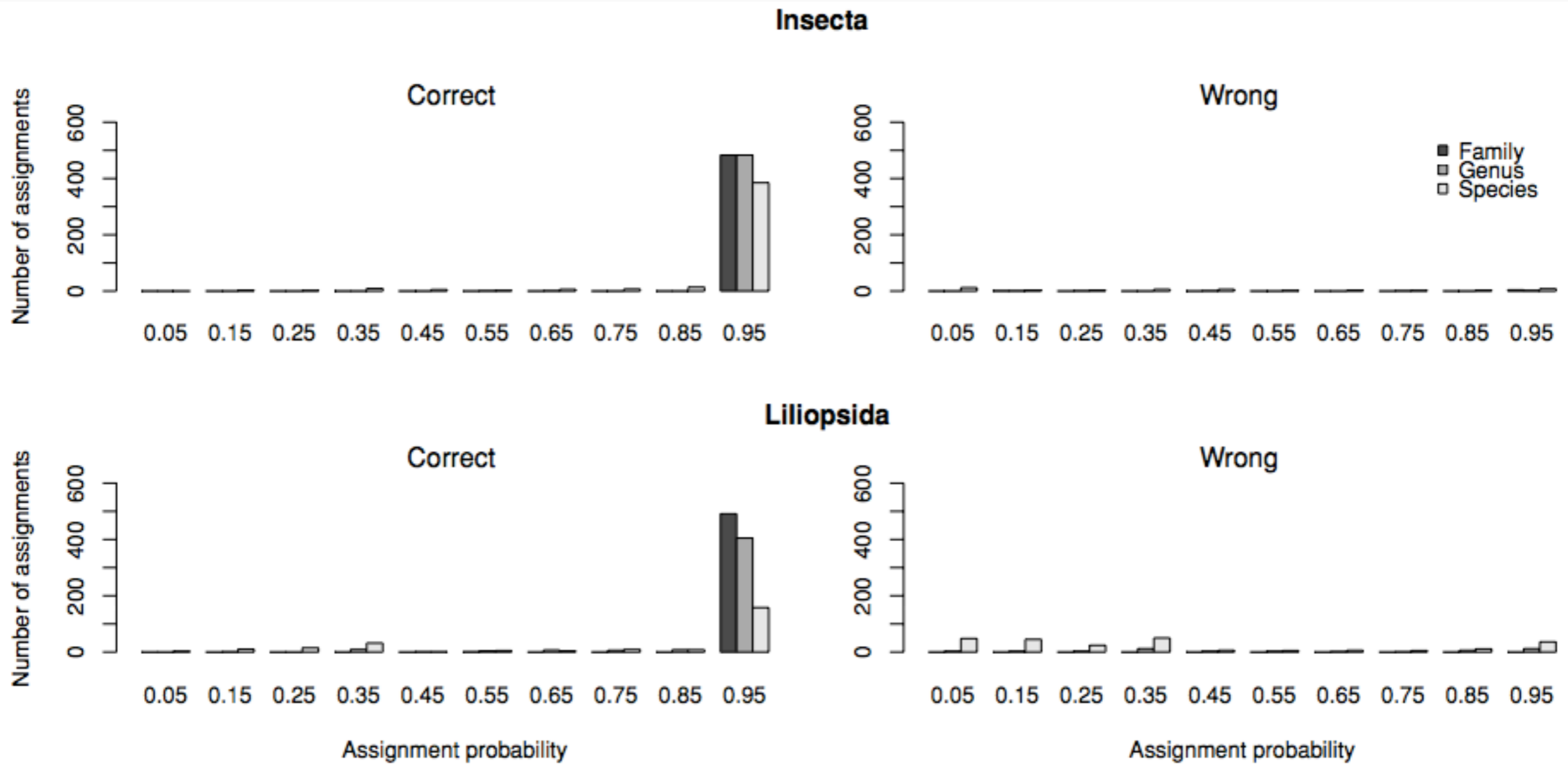


FIGURE 4. Distributions of assignment probabilities for correct and wrong assignments. At the levels of species, genus, and family, 90%, 99%, and 99% of assignments of *Insecta* sequences are correct and 51%, 90%, and 100% of assignments of *Liliopsida* sequences are correct. Wrong assignments are generally associated with low probabilities, whereas most correct assignments achieve probabilities above 95%.

Comparison with BLAST

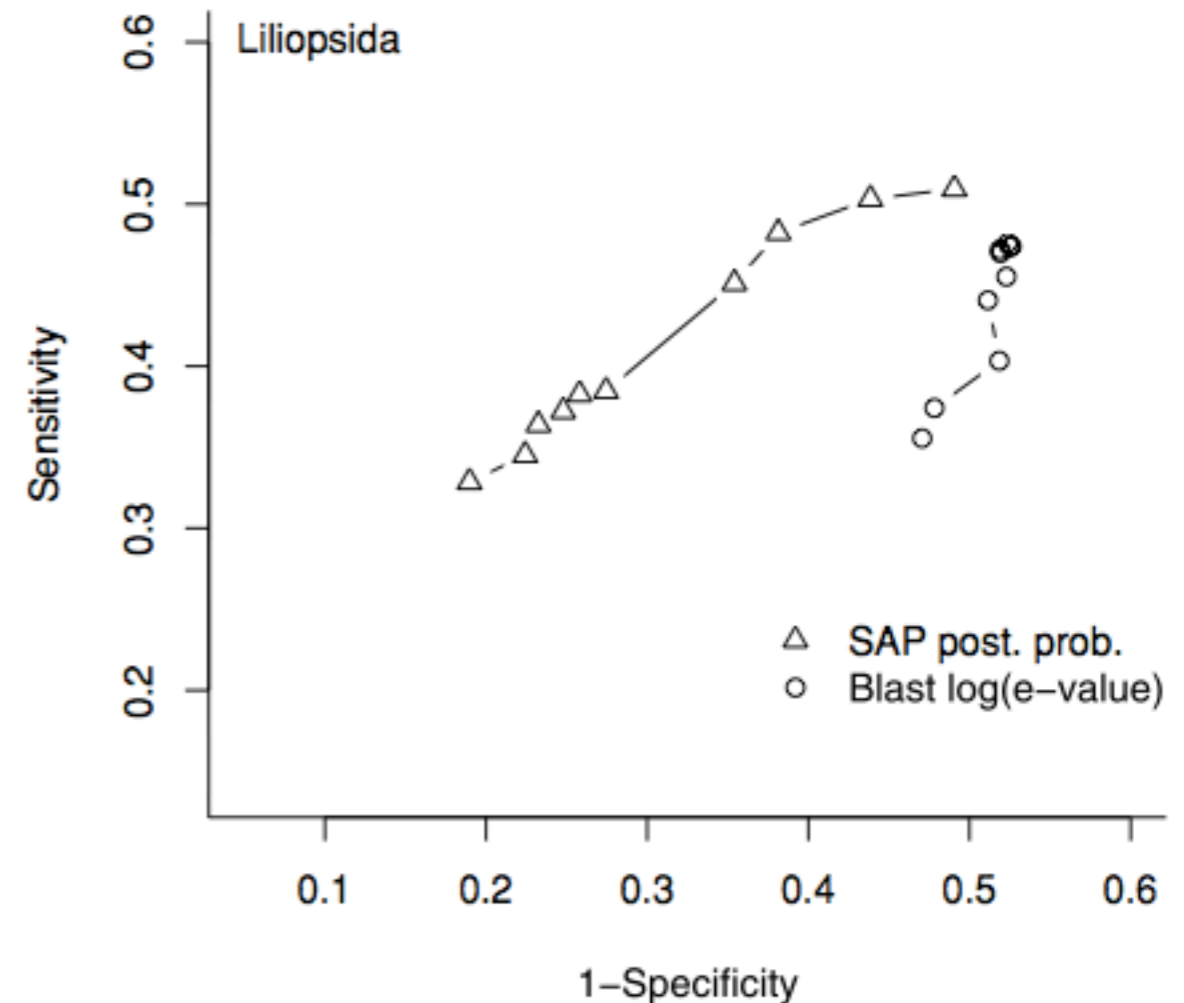
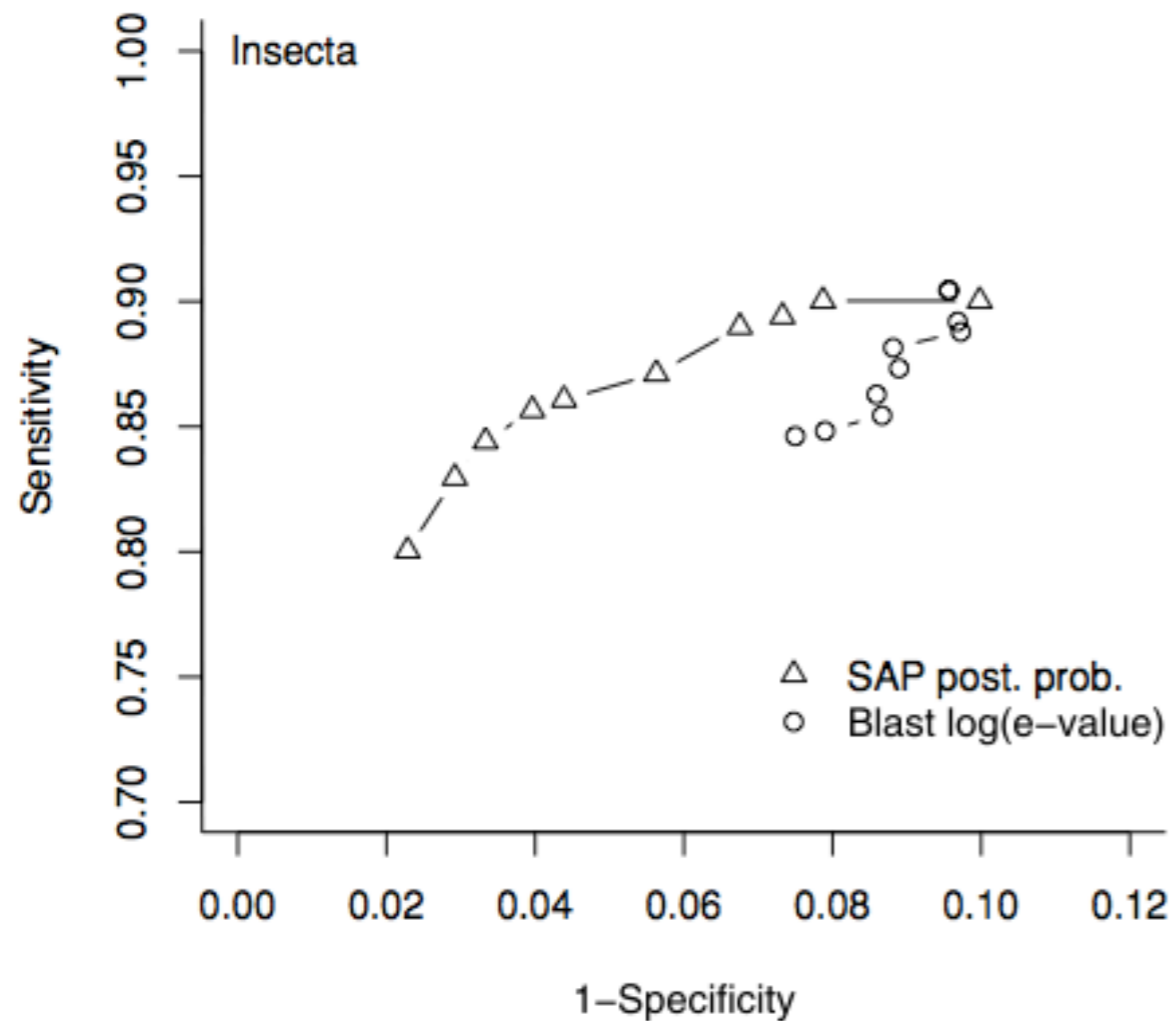


FIGURE 5. ROC (receiver operating characteristic) curves summarizing the tradeoff between sensitivity and specificity in the range of most to least stringent assignment criteria used. Sensitivity is the fraction of all sequences that are correctly assigned, specificity is the fraction of assignments that are correct. The performance of SAP exceeds that of Blast for any sensitivity-specificity combination except when blindly accepting all assignments.

Reanalysis of Neanderthal Sequences

- ✦ In a number of studies, longer ancient DNA sequences were assembled from shorter reads
- ✦ However, what if some of these reads were not of Neanderthal origin?

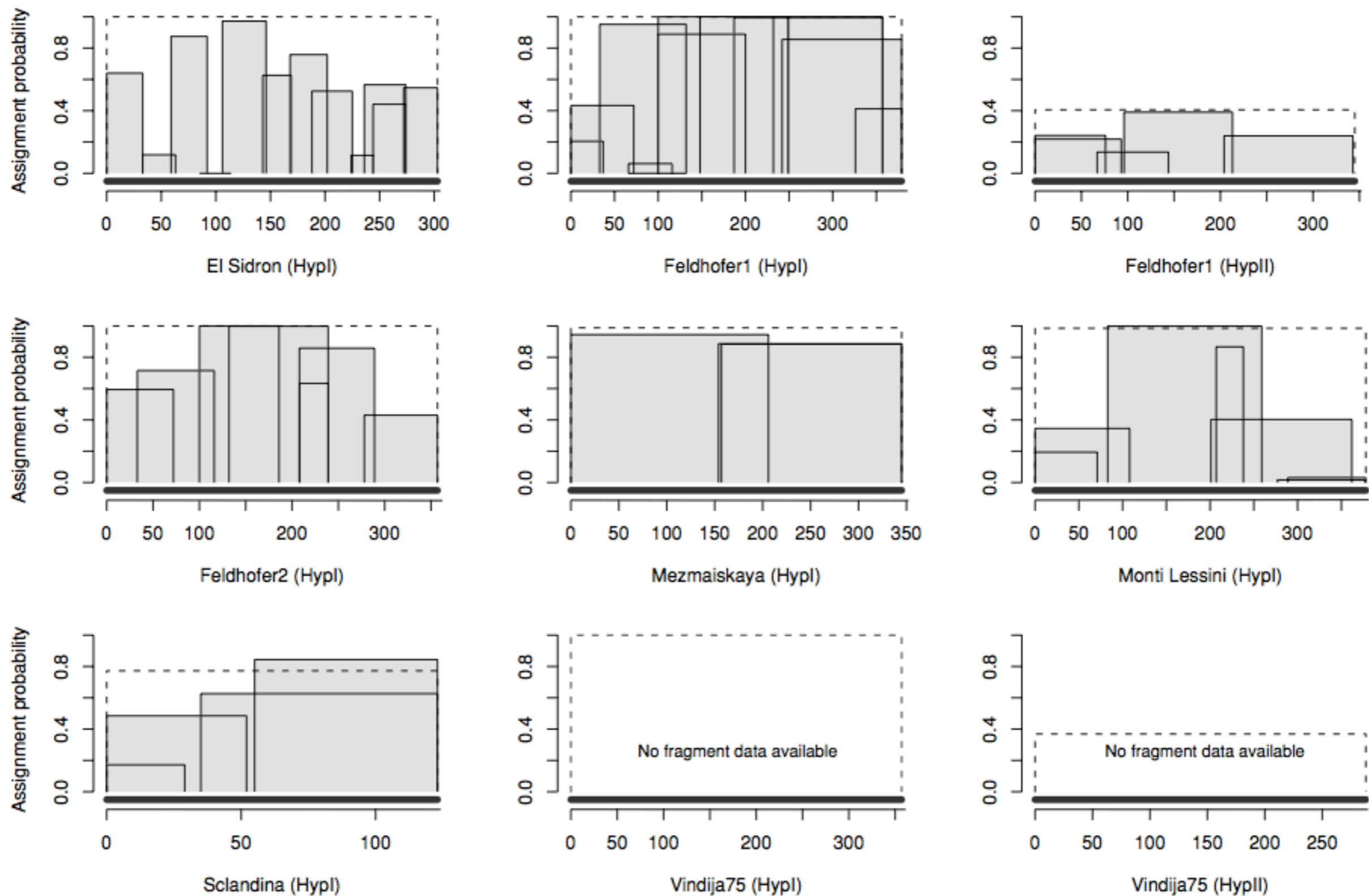


FIGURE 7. Summary of confidence analysis for published Neanderthal sequences. In each sub-figure, a bold bar represents the Neanderthal sequence analyzed. The overlapping boxes above it each represent the assignment probability of the sequence fragment spanned by the box. The dashed box represents the full inferred sequence, whereas shaded boxes represent individual contributing PCR fragments. For the Vindija75 sequences, no information on PCR fragments is available. The five short sequences not in GenBank obtain the following assignment probabilities: Engis2 (HypI): 0.88; LaChapelleAuxSaints (HypI): 0.88; RochersDeVilleneuve (HypI): 0.63; Vindija77 (HypI): 0.87; Vindija80 (HypI): 0.89.

Bayesian MCMC is Slow

- ✦ The Bayesian approach to tree sampling required to obtain a statistically meaningful confidence measure is computationally demanding
- ✦ To use SAP on large datasets, such as environmental samples, faster tree sampling approaches are needed

Fast Phylogenetic DNA Barcoding

- ✦ Munch et. al 2008. (not assigned reading)
- ✦ Tree sampling performed using **neighbor-joining** (Saitou & Nei 1987) and **non-parametric bootstrapping** (Felsenstein 1985).
 - ✦ This method of tree sampling is much faster than MCMC

Neighbor-Joining

- ✦ Neighbor-joining (NJ) selects a pair of taxa from the complete set and constructs a new subtree that joins the pair, iteratively building a tree in this manner
- ✦ Pairs of taxa are selected by minimizing the following criterion:

$$Q(i,j) = (L-2)d(i,j) - \sum_{k=1}^L d(i,k) - \sum_{k=1}^L d(j,k), \quad (2.1)$$

- ✦ Tutorial: <http://artedi.ebc.uu.se/course/sommar/njoin/index.html>

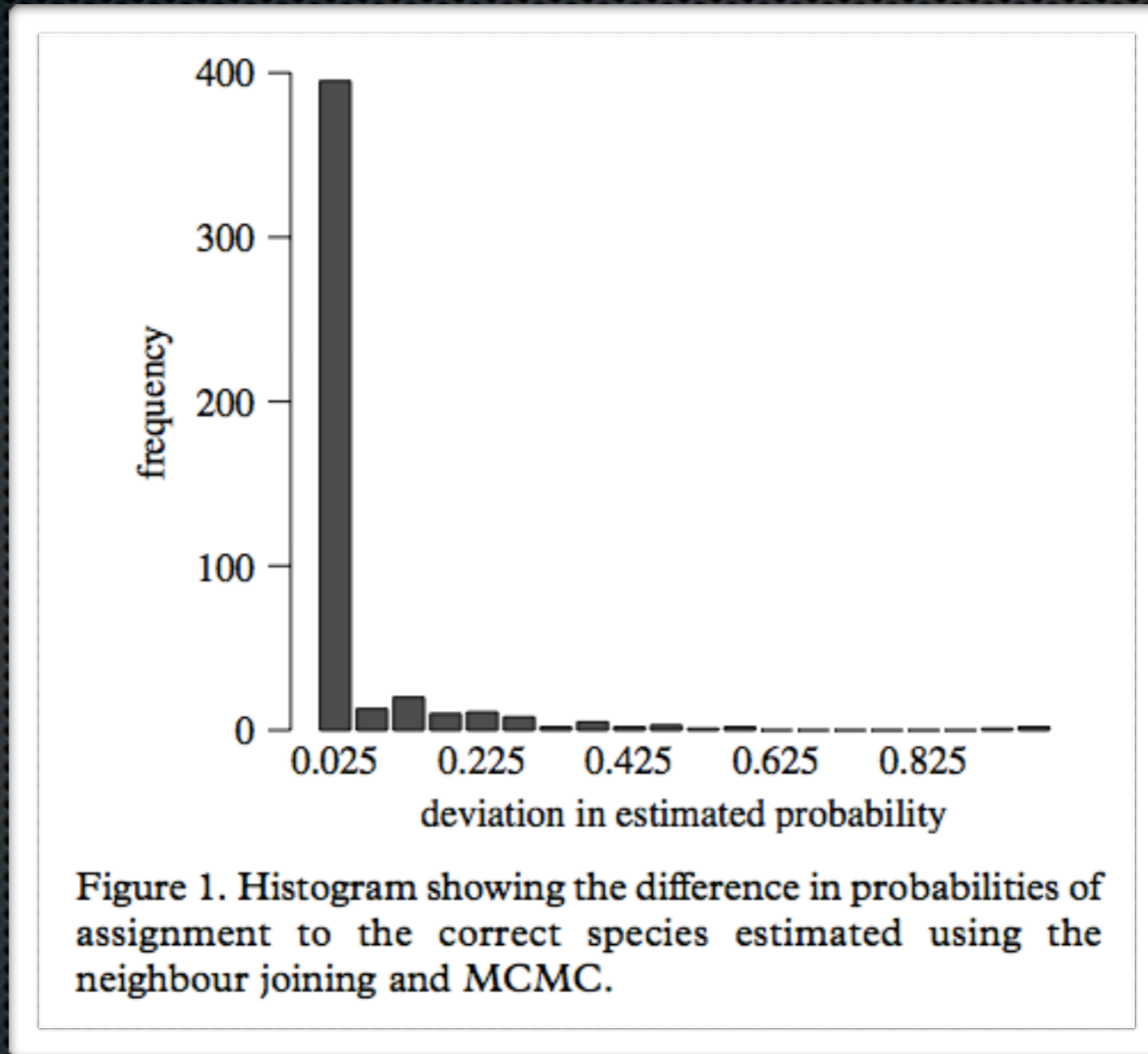
Notes on Neighbor-Joining

- Each iteration step requires only recalculating one row in the Q matrix, leaving the initial calculation of sequence distances and the identification of the minimal entry in Q as the only operations with $O(L^2)$ complexity
- Translates to very fast running times, in practice
- The constrained version of the algorithm simply ensures that the (i, j) taxon pairs chosen are compatible with the taxonomic backbone, and speeds up the algorithm even more - identifying the pair to join is now linear in L

Comparing Bayesian MCMC to Neighbor-Joining with Bootstrapping

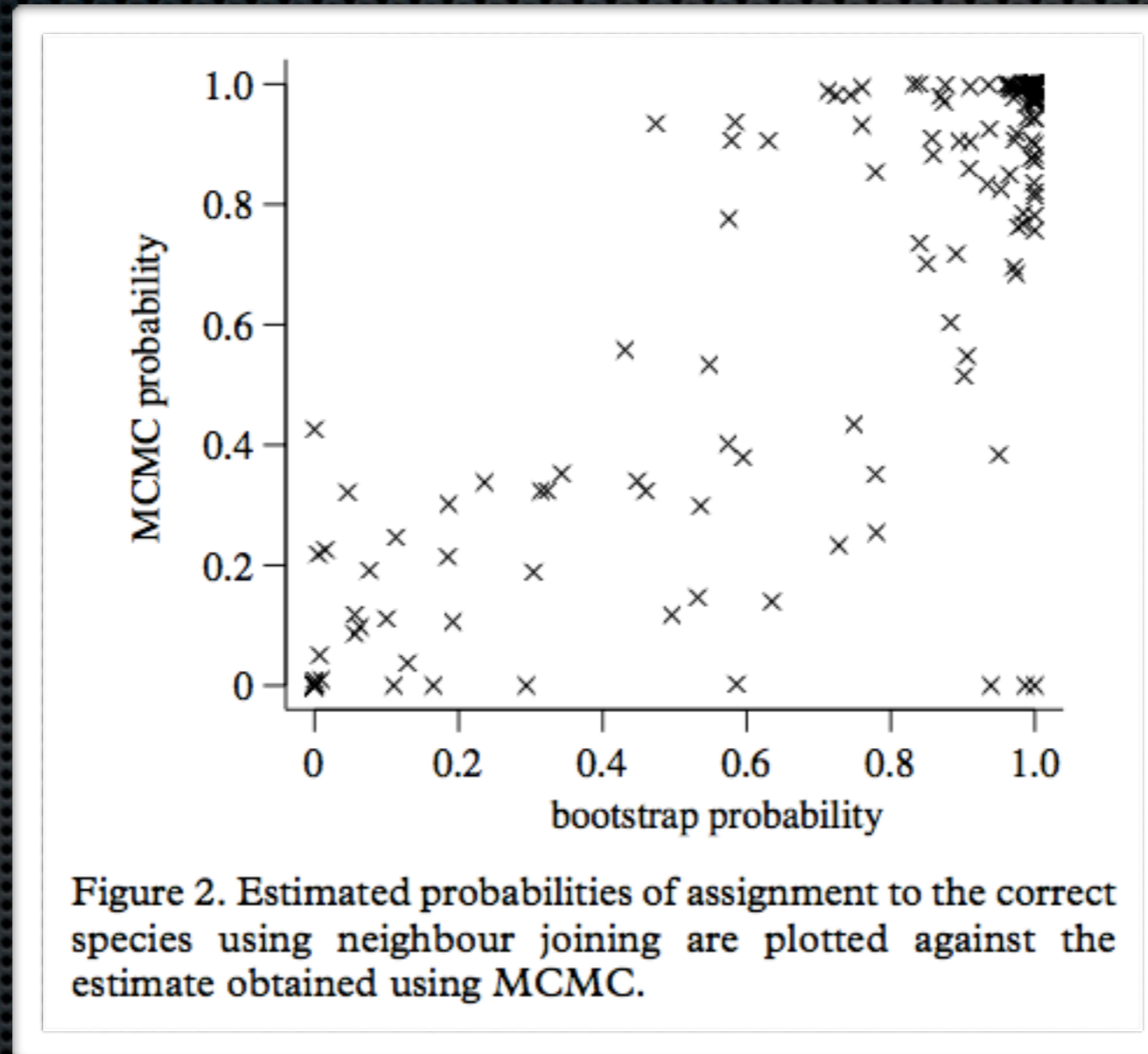
- 10^3 bootstrap samples vs. 10^6 iterations of MCMC
- The average difference between MCMC and NJ assignment probabilities is 5%
- For assignment probabilities between 0.8 and 1.0, the average difference is only 2.6%

MCMC vs. NJ



The majority of the time, the deviation in estimated probability is small

MCMC vs. NJ



There is better agreement when the assignment probability is large

MCMC vs. NJ

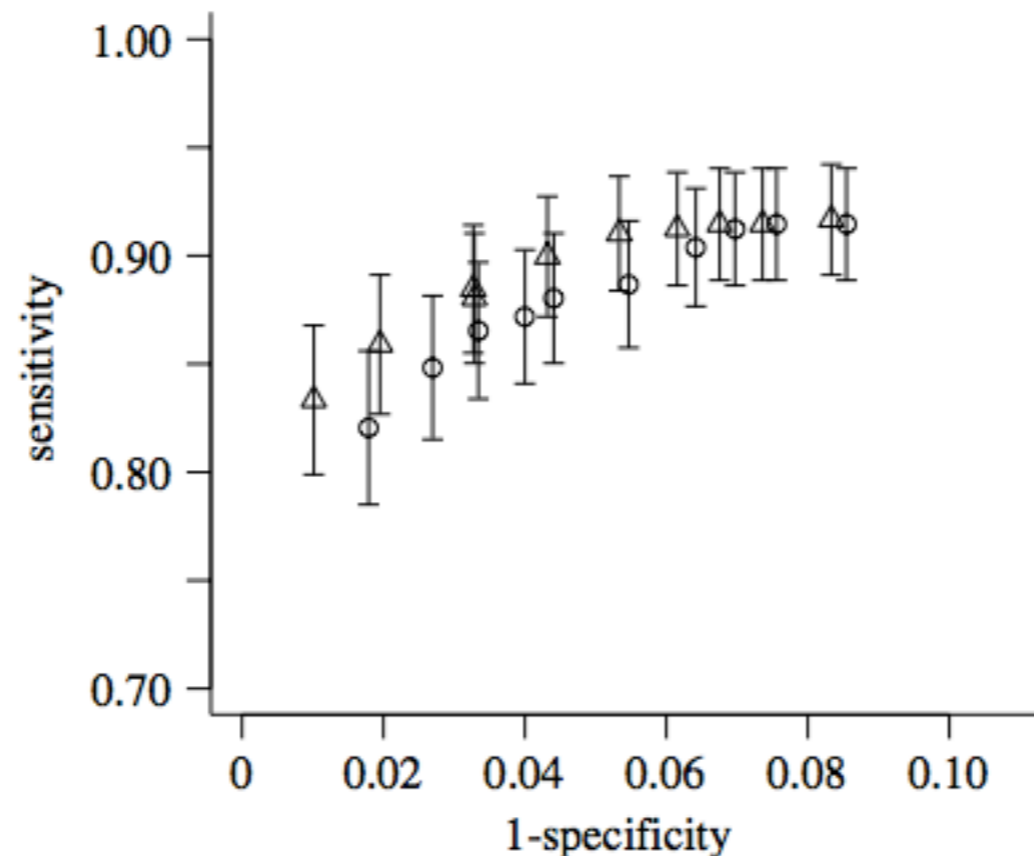


Figure 3. ROC curves summarizing the trade-off between sensitivity and specificity in the range of most to least stringent assignment criteria used. Sensitivity is the fraction of all sequences that are correctly assigned and specificity is the fraction of assignments that are correct. Vertical bars represent confidence intervals of the sensitivity statistic. Triangles, NJ; circles, MCMC.

MCMC vs. NJ, MCMC vs. BLAST

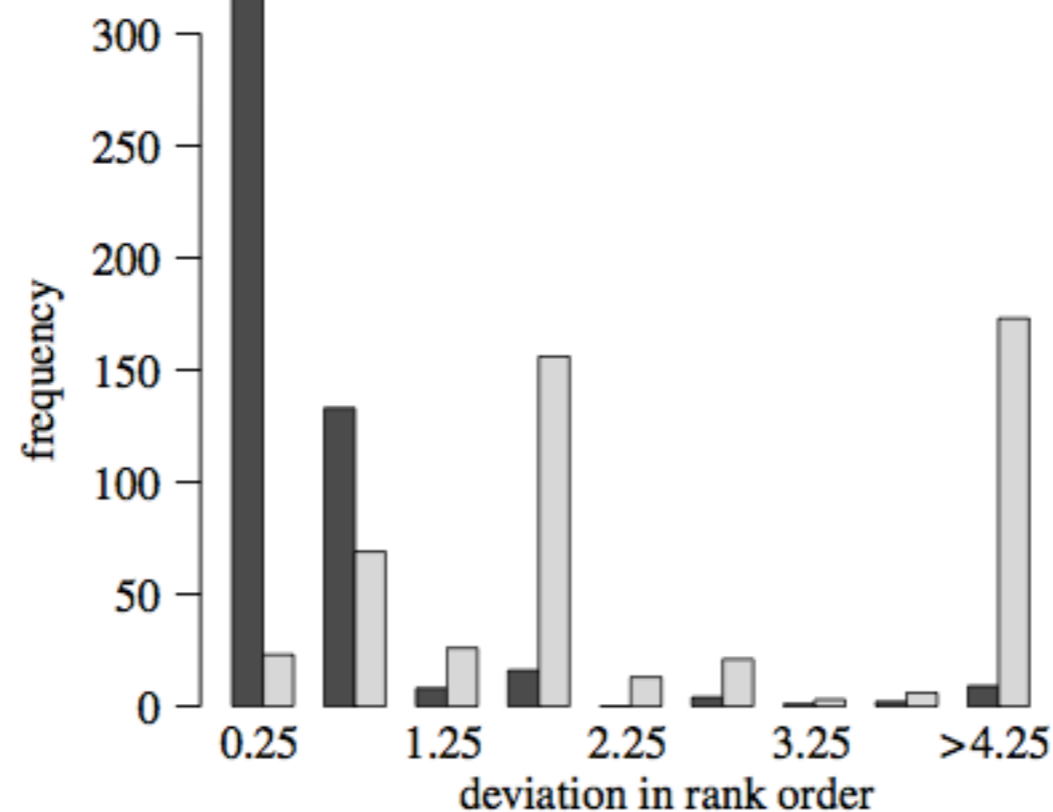


Figure 4. Histogram illustrating the agreement in terms of rank order obtained by sorting the set of homologues by the assignment probability associated obtained neighbour joining with bootstrapping and maximum likelihood and MCMC. The histograms show the average difference in rank order for neighbour joining and BLAST from the one obtained using MCMC.

MCMC vs. NJ

- ✦ Posterior probabilities and bootstrap proportions are not expected to match closely
 - ✦ They measure different quantities
 - ✦ Use different models of nucleotide substitution
 - ✦ High variance in estimates due to relatively small number of bootstrap replicates and MCMC iterations
- ✦ However, it is clear (i.e., the authors are convinced) that for high posterior probabilities, NJ can be considered a fast approximation of MCMC

Reanalysis of Ancient DNA Environmental Samples

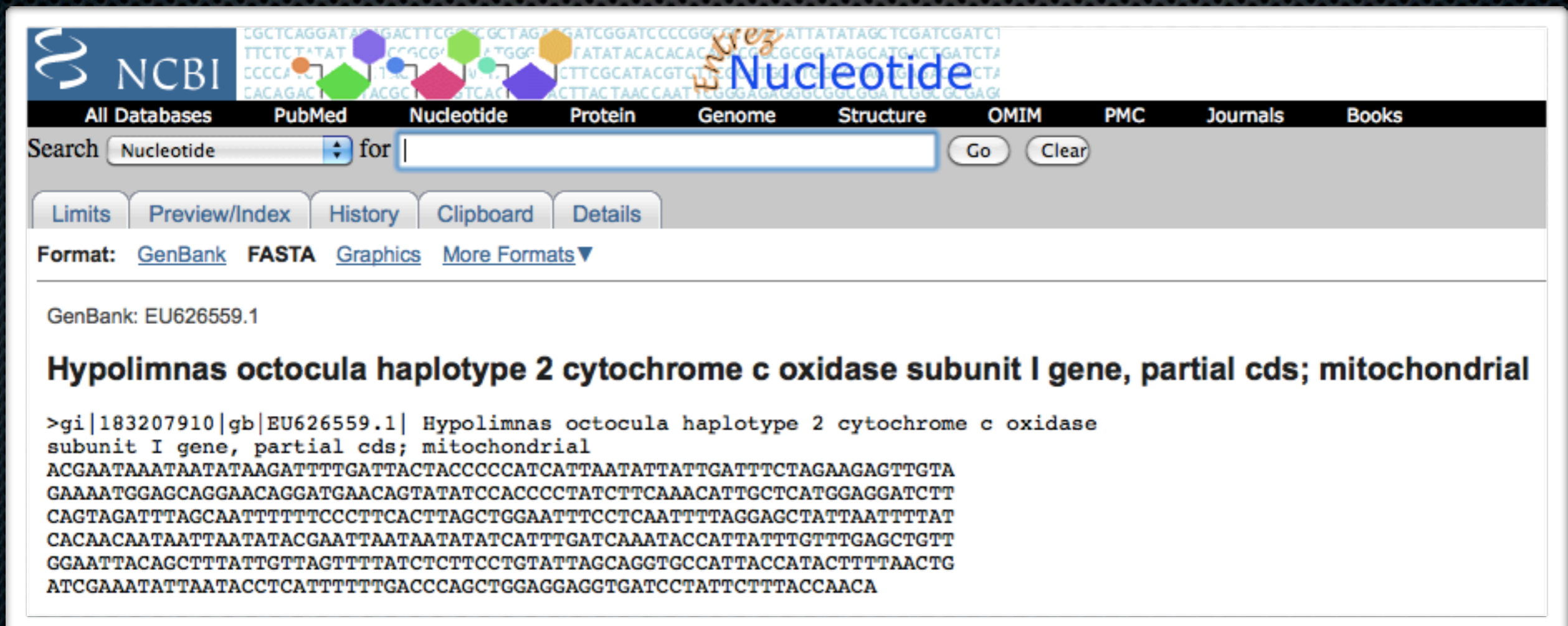
- Previously published analysis of permafrost samples from Siberia and temperate sediments from New Zealand (Willerslev *et al.* 2007)
- 130 bp fragments of the chloroplast *rbcL* gene and 100-280 bp fragments of the vertebrate mitochondrial 16S, 18S, cytochrome *b*, and control region genes were obtained using PCR
- These data were originally analyzed using BLAST along with consensus NJ trees for the vertebrate genes

Reanalysis of Ancient DNA Environmental Samples: Results

- ✦ For the animal species, SAP assignments overlap with original ones, but are not in complete agreement
- ✦ SAP was able to make some assignments to a lower taxonomic level
- ✦ Results emphasize the value of a confidence measure, allowing some assignments to be rejected
- ✦ Also shows that SAP allows for greater sensitivity and resolution than a conservative approach using BLAST

SAP Trial Run

- ✦ Installed SAP (version 1.0.8) and dependencies
- ✦ Downloaded an *Insecta* COI sequence from GenBank



The screenshot shows the NCBI Nucleotide search interface. The search bar contains the text "Nucleotide" and "for". Below the search bar, there are buttons for "Limits", "Preview/Index", "History", "Clipboard", and "Details". The "Format" section is set to "GenBank". The search results display the GenBank entry for EU626559.1, which is the "Hypolimnas octocula haplotype 2 cytochrome c oxidase subunit I gene, partial cds; mitochondrial". The sequence is shown in FASTA format.

```
>gi|183207910|gb|EU626559.1| Hypolimnas octocula haplotype 2 cytochrome c oxidase subunit I gene, partial cds; mitochondrial
ACGAATAAATAATATAAGATTTTGATTACTACCCCCATCATTAAATATTATTGATTTCTAGAAGAGTTGTA
GAAAATGGAGCAGGAACAGGATGAACAGTATATCCACCCCTATCTTCAAACATTGCTCATGGAGGATCTT
CAGTAGATTTAGCAATTTTTTCCCTTCACTTAGCTGGAATTTCTCAATTTTAGGAGCTATTAATTTTAT
CACAACAATAATTAATATACGAATTAATAATATATCATTTGATCAAATACCATTATTTGTTTGAGCTGTT
GGAATTACAGCTTTATTGTTAGTTTTATCTCTTCTGTATTAGCAGGTGCCATTACCATACTTTTAACTG
ATCGAAATATTAATACCTCATTTTTTGACCCAGCTGGAGGAGGTGATCCTATTCTTTACCAACA
```


SAP Trial Run

- ✦ Invoked SAP with default parameters
- ✦ Found 48 significant homologs



48 homologs in set:

1 phyla: Arthropoda

1 classes: Insecta

1 orders: Lepidoptera

6 families: Papilionidae HesperIIDae Nymphalidae Pieridae Lycaenidae SpHINGidae

27 genera: Protogoniomorpha Glaucopsyche SAlamis Auca Albulina Hypolimnas Hyles
Mechanitis Melitaea Sevenia Antanartia Plebejus Joanna Eumorpha RImisia
Aricia Yoma Precis Asterocampa Kallimoides Polygonia Dymasia Luehdorfia
Junonia Pieris Chilades Xylophanes

Last accepted E-value is 8.974690e-111

Ratio of lowest to highest bit score is: 0.497886595867

WARNING: Diversity goal not reached.

Relative bit-score cut-off (0.50) at level: genus

SAP Trial Run

- Results: <http://serine.umiacs.umd.edu/files/saphtml>

SAP, in Summary

- Statistical approaches provide measures of confidence in assignment
- SAP is a modular framework with different options for BLAST searches, alignment, and phylogenetic analysis
- More work would need to be done to make SAP truly feasible for analysis of large metagenomic datasets

Conclusion

- ✦ MEGAN provides a number of useful features for metagenomic analysis, but only uses BLAST for taxonomic assignment
- ✦ SAP is a more sophisticated framework for taxonomic assignment, but requires more computation
- ✦ Suggestion: combine features of MEGAN and SAP