Introducing mothur: A Computational Toolbox for Describing and Comparing Microbial Communities

N-099

PD Schloss & SL Westcott http://schloss.micro.umass.edu/mothur

Abstract

The field of microbial ecology has recently benefited from the availability of a number of tools for describing and comparing microbial communities using robust statistical methods. These tools have included LIBSHUFF / ∫ –LIBSHUFF, DOTUR, SONS, TreeClimber, and UniFrac. As traditional clone and sequence methods, which generate hundreds of sequences have been supplemented by pyrosequencing techniques, which generate hundreds of thousands of sequences, it has become necessary to refactor these tools to make them more amenable to larger datasets. Our new software, mothur, is designed to be a platform that will enable investigators to align their 16S rRNA gene sequences, calculate pairwise distances, and analyze the resulting distance matrices using algorithms employed in these already popular tools using a single program. Improved programming of these algorithms to accommodate the expanded datasets has radically improved execution times. For example, we have been able to decrease the amount of time required to assign sequences into operational taxonomic units (OTUs) from days to minutes. Within mothur, we have also provided a version of the NAST alignment algorithm used by the greengenes database, which is now more than 100-times faster and considerably more flexible than the online version. In addition to improving the implementation of these algorithms, we have also added functionality such as including the ability to generate Venn diagrams, community trees, heat maps and sample-based rarefaction curves. Using improved object oriented programming strategies in the C++ programming language our goal is for mothur to be easily expandable by other interested programmers as an open source code project.

Example Dataset

In 2005, Sogin and colleagues published the first study using pyrosequencing technology to survey 16S rRNA genes. They isolated DNA from 8 marine samples and obtained 222,291 high quality reads from the bacterial V6 region of the 16S rRNA gene.

North Atlantic Samples

53R (58.3N,-29.1W, 1400 m depth)	**13,040 sequences**
55R (58.3N,-29.1W, 500 m depth)	**10,134 sequences**
112R (50.4N, -25.0W, 4121 m depth)	**16,087 sequences**
115R (50.4N, -25.0W, 550m depth)	**16,651 sequences**
137 (60.9N, -38.5W, 1710 m depth)	**14,147 sequences**
138 (60.9N, -38.5W, 710 m depth)	**13,241 sequences**

Axial Seamount, Juan de Fuca Ridge

FS312 (45.9N, -130.0W, 1529 m depth)	**55,592 sequences**
FS396 (45.9N, -130.0W, 1537 m depth)	**83,399 sequences**

21,903 unique V6 tags (seconds)

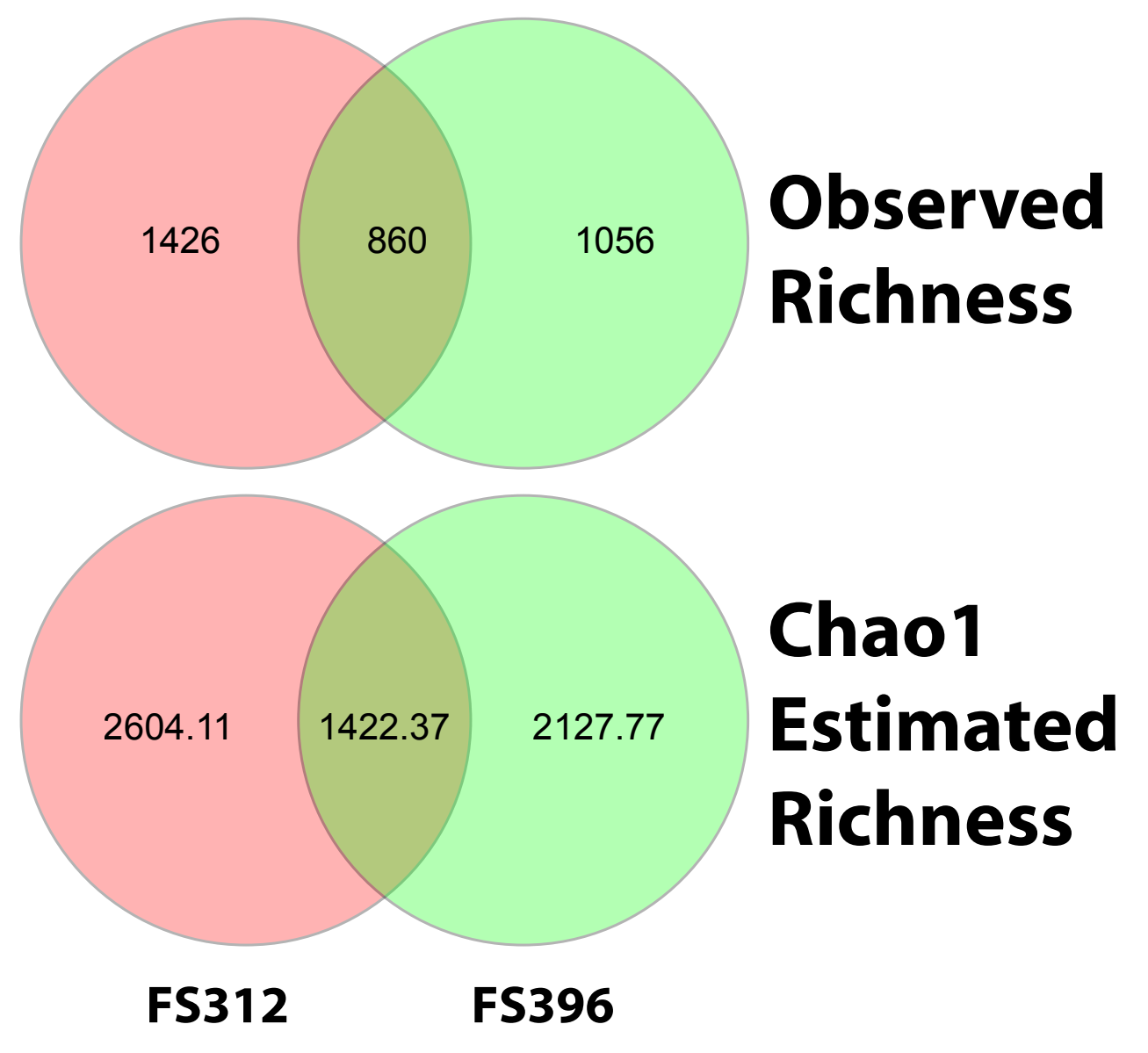
68 seconds, with quality as good as SINA

5 hours, distance counts gaps; matrix occupies 1.31 GB. Options for parallelization

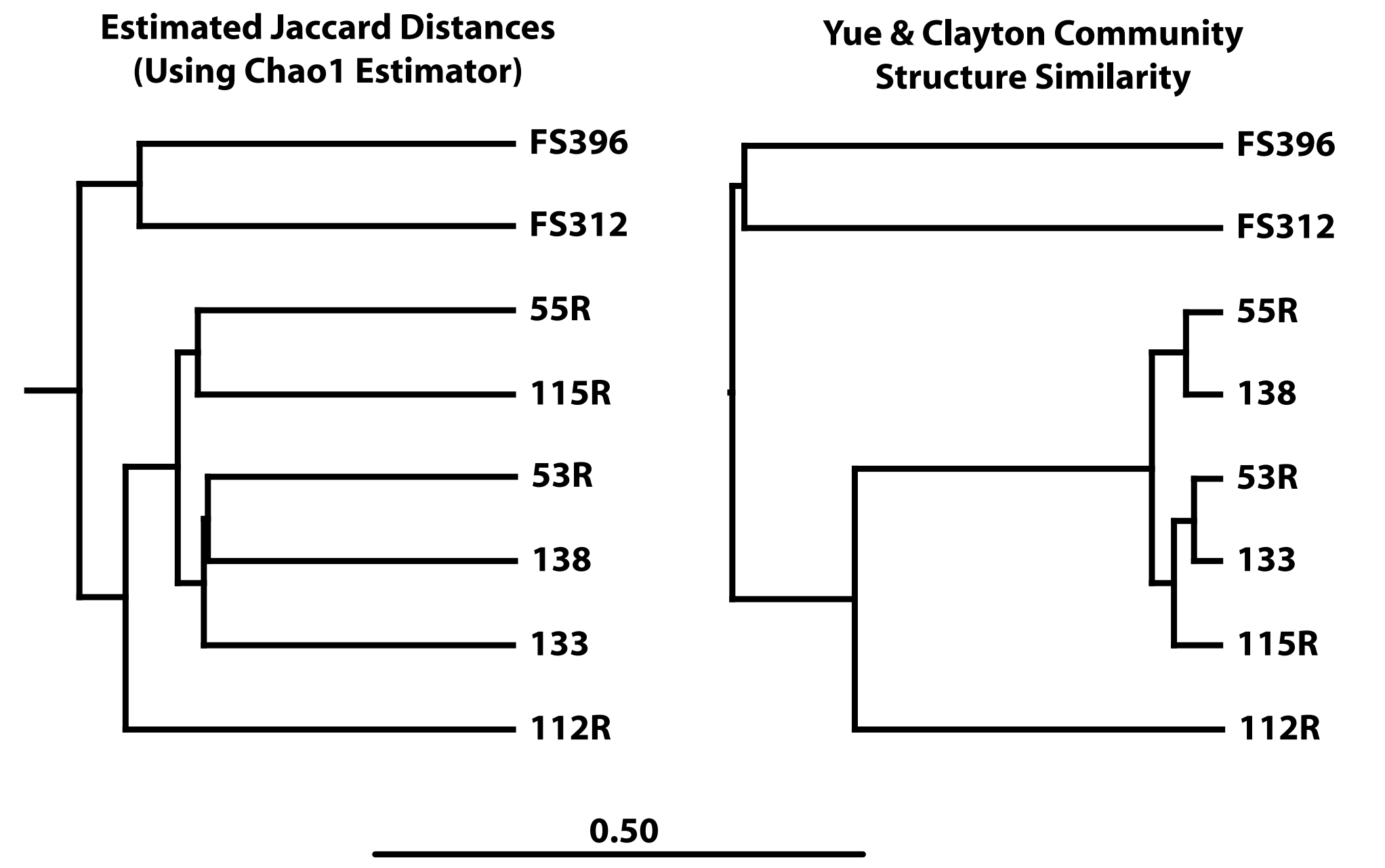
Clustering required ~45 min

```
deconvolute(fasta=sogin.fasta)
align.seqs(database=core_set_aligned.imputed.fasta, candidate=sogin.unique.fasta)
filter.seqs(fasta=sogin.unique.kmer.needleman.nast.align, trump=.)
dist.seqs(fasta=sogin.unique.kmer.needleman.nast.align, cutoff=0.20)
read.dist(column=sogin.unique.dist, name=sogin.names, cutoff=0.20)
cluster()
read.otu(list=sogin.unique.fn.list, group=sogin.groups, label=unique-0.05-0.10-0.20)
venn(groups=FS312-FS396, calc=sharedchao-sharedsobs)
heatmap(scale=log10)
tree.shared(calc=jest-thetayc)
summary.shared()
rarefaction.shared()
read.otu(list=sogin.unique.fn.133.list)
summary.single()
rarefaction.single()
read.otu(list=sogin.unique.fn.138.list)
summary.single()
rarefaction.single()
read.otu(list=sogin.unique.fn.112R.list)
summary.single()
rarefaction.single()
read.otu(list=sogin.unique.fn.115R.list)
summary.single()
rarefaction.single()
read.otu(list=sogin.unique.fn.53R.list)
summary.single()
rarefaction.single()
read.otu(list=sogin.unique.fn.55R.list)
summary.single()
rarefaction.single()
read.otu(list=sogin.unique.fn.FS312.list)
summary.single()
rarefaction.single()
read.otu(list=sogin.unique.fn.FS396.list)
summary.single()
rarefaction.single()
```

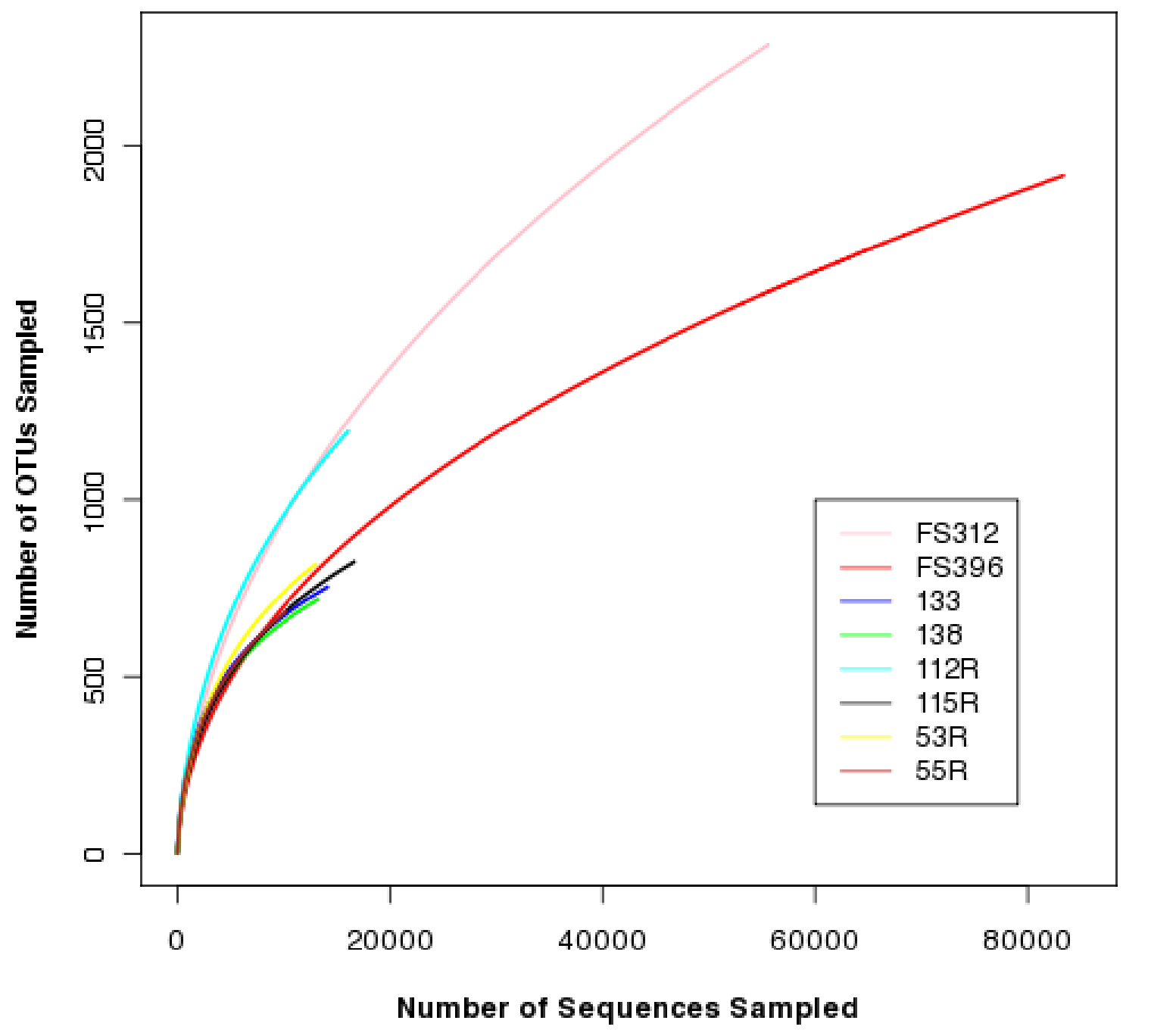
Example batch file: the same commands can be called within mothur



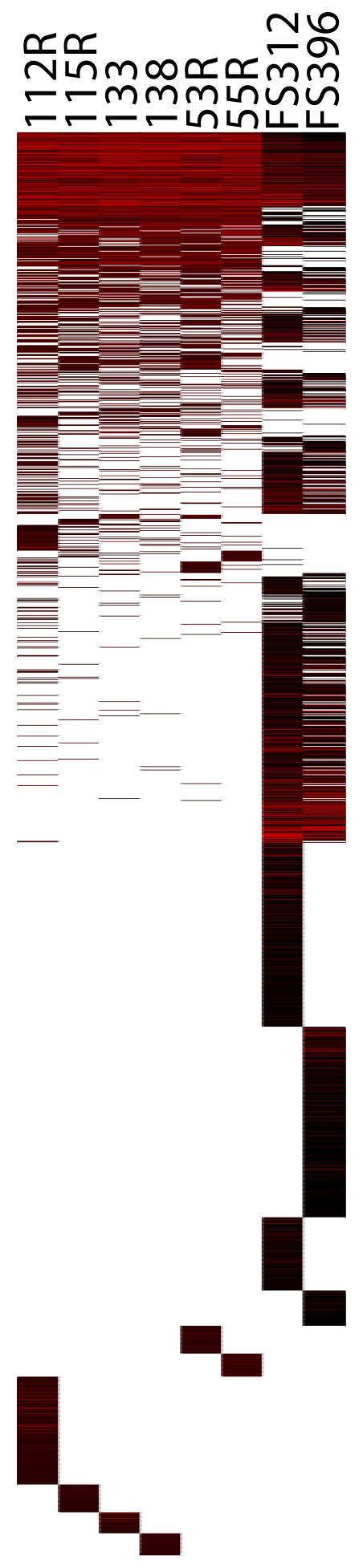
mothur generates SVG-formatted files containing Venn diagrams for up to 4 samples. The circles can be modified by the user to be proportional to the observed or estimated richness. Data shown are for a 0.10 distance cutoff.


Estimated Jaccard Distances (Using Chao1 Estimator)
FS396
FS312
55R
115R
53R
138
133
112R
Yue & Clayton Community Structure Similarity
FS396
FS312
55R
138
53R
133
115R
112R
0.50

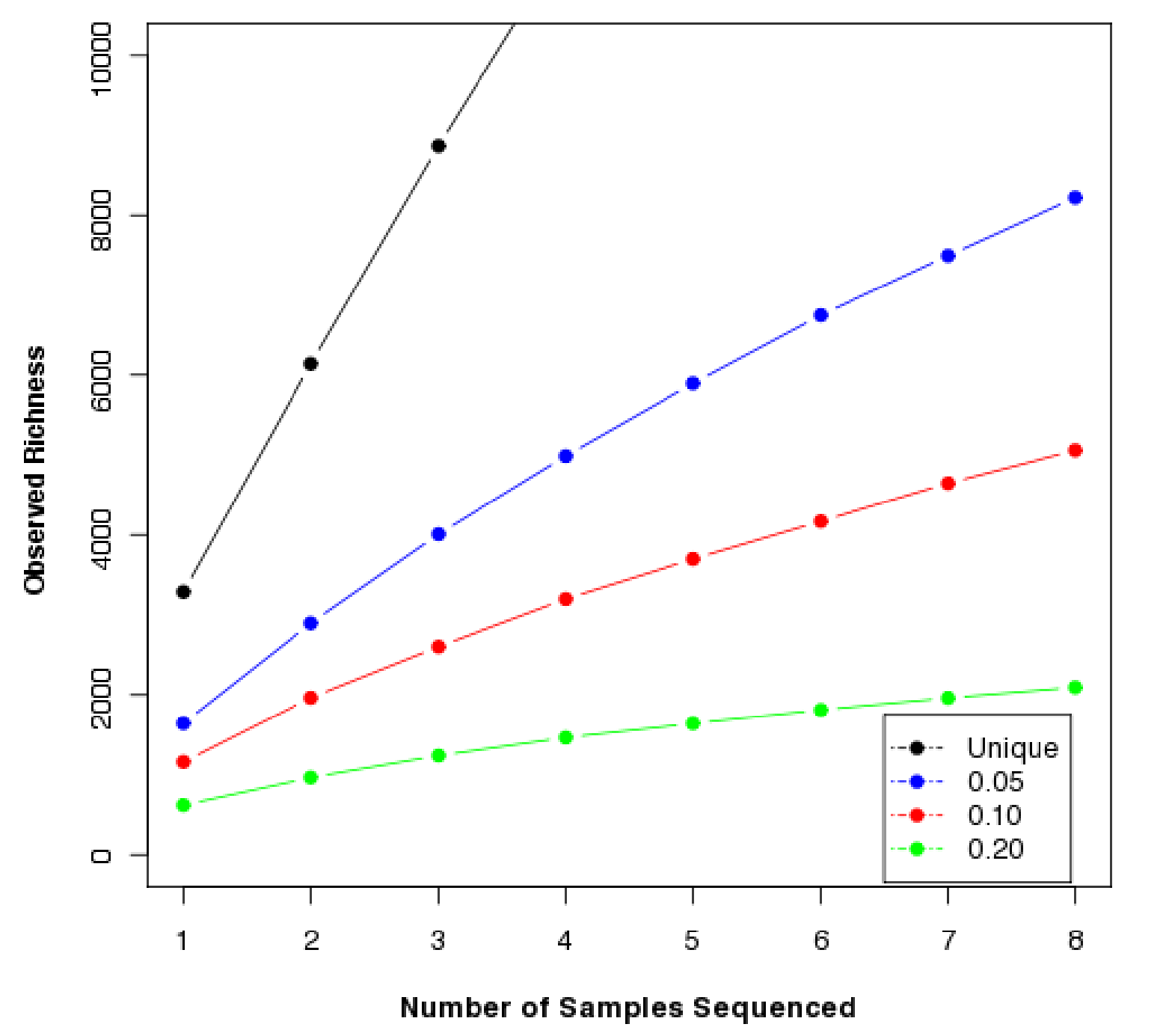
Newick-formatted files are generated in mothur using a variety of distance calculators based on the membership or structure of the samples. Data shown are for a 0.10 distance cutoff.



Like DOTUR, mothur can calculate rarefaction curves for individual samples. Here are the rarefaction curves for the 8 marine seqeunce collections for a distance cutoff of 0.10.



With the heatmap command, mothur will geenrate heat maps representing the relative abundance of each OTU for your different samples. This heatmap depicts the relative abundace for OTUs defined at an 0.20 distance level. The colors are log-scaled between red and black.



Sample-based rarefaction curves can be generated in mothur to assess the relationship between sampling effort and observed richness. Here are four rarefaction curves for OTU definitions based on uniqueness and distance of 0.05, 0.10, and 0.20.

Join the mothur community!

- mothur is freely available as source code for compiling in Linux and Mac OS X or as an executable for Windows
- Website is a wiki for documentation including example workflows that you are encouraged to contribute to
- Feel free to make suggestions for new features or to even contribute your own source code
- We are available as an open source resource to come and train you and your colleagues

Acknowledgements

We thank Thomas Ryabin for helping with programming and previous users of our software for useful feedback and encouragement.

This work was funded through grants received from the Sloan Foundation, the National Science Foundation, and the College of Natural Resources and the Environment at the University of Massachusetts.