**MMG 890 Microbial Metagenomics Summer, 2012**

**Evaluating the diversity of microbial communities using molecular surveys**

**Contacts:**

Kevin Theis, [theiskev@msu.edu](mailto:theiskev@msu.edu)

Mario Muscarella, [muscare2@msu.edu](mailto:muscare2@msu.edu)

Arvind Venkataraman, [av299@msu.edu](mailto:av299@msu.edu)

**Objective:** In this exercise, you will use 16S rRNA gene survey data to *evaluate whether standard agricultural land-use practices affect the structure of soil bacterial communities.* We’ve provided you with data from deciduous forest (MB3) and standard agriculture (MA1) soils at the Kellogg Biological Station. Each soil was independently sampled three times and a clone library (~ 90 sequences) was created for each sample. By analyzing the diversity of these six clone libraries, you will gain experience using contemporary community analysis software and statistical tools, forming conclusions based on your results, and effectively communicating your conclusions to peers.

**Approach:** To complete this exercise, you will use two freely accessible software programs. **Mothur** offers a centralized suite of tools for clustering gene sequences and for analyzing the diversities of communities based on sequence data (<http://www.mothur.org/wiki/Main_Page>). **R** is a data analysis package that offers a variety of statistical and graphical tools that are widely used in biology, including contemporary community ecology(<http://www.r-project.org/>).

mothur v.1.25.1

R v.2.15.0

Packages: vegan, MASS

User manuals for mothur and R are available online. They are frequently updated, so we have not provided copies of them here. While the manuals should serve as your primary references in the future, we have included detailed instructions for completing this exercise below. Additionally, we have provided you with the following files:

KBS.dist Phylip distance matrix indicating similarities among all sequences

KBS.groups Indicates community membership of the sequences in the matrix

KBSlumped.groups Same, only samples for each treatment have been lumped

Optional files:

KBStree.ntree Newick phylogenetic tree of all the sequences in the study

KBStree.groups Indicates community membership of the sequences in the tree

**1) Binning 16S rRNA gene sequences into operational taxonomic units (OTUs)**

We have provided a distance matrix that indicates the percentage similarities of all sequences (541) recovered from the six soil samples. In this initial step you are going to have mothur bin the sequences into OTUs across a gradient of sequence-similarity cutoffs, using the average neighbor clustering algorithm. The differences between the furthest, average and nearest neighbor clustering algorithms are explained at: <http://www.mothur.org/wiki/Cluster>

> cluster(phylip=KBS.dist, method=average)

This command yields 3 files:

KBS.an.list Contains data indicating the sequences that cluster together within an OTU

KBS.an.rabund Contains rank-abundance plot data

KBS.an.sabund Contains species-abundance plot data

Having successfully binned sequences into OTUs, you are now going to separate the sequences based on the sample from which they were recovered.

> make.shared(list=KBS.an.list, group=KBS.groups)

This command yields 7 files:

KBS.an.shared Indicates the number of times an OTU is observed in each sample

KBS.an.MA1A.rabund

KBS.an.MA1B.rabund

KBS.an.MA1C.rabund

KBS.an.MB3A.rabund

KBS.an.MB3B.rabund

KBS.an.MB3C.rabund

**2) Evaluating alpha diversity among forest and agricultural soil bacterial communities**

Remember that alpha diversity is the variety and distribution of ‘species’ *within* a given community. Generate rarefaction curves for the sampled forest and agricultural soil bacterial communities at sequence similarity cutoffs of 97, 90 and 80%.

> rarefaction.single(shared=KBS.an.shared, freq=2, label=0.03-0.10-0.20, groupmode=F)

The ‘freq’ component of the command indicates your desired step-distance for the X-axis of your curves The ‘label’ component indicates which sequence similarity cutoffs you would like data for

The ‘groupmode=F” component produces rarefaction files for each individual sampling unit

This command yields 6 files:

KBS.an.MA1A.rarefaction Contains data for constructing rarefaction curves

KBS.an.MA1B.rarefaction

KBS.an.MA1C.rarefaction

KBS.an.MB3A.rarefaction

KBS.an.MB3B.rarefaction

KBS.an.MB3C.rarefaction

**Construct rarefaction curves for each sampling unit using R**

We have provided you with an .r file containing basic scripts that will allow you to

complete each of the R-associated components of this exercise. As you become more

conversant in R, you will likely want to add elements to these analyses so that they

better meet your specific research needs.

**In general, what is the degree of sample coverage at 97, 90 and 80% sequence similarity cutoffs? Given this information, how should you proceed?**

**Based exclusively on the rarefaction curves, do forest and agricultural soils appear to have a dissimilar number of OTUs at a sequence similarity cutoff of 90%?**

Using mothur, generate Chao, Shannon and Simpson indices for both forest and agricultural bacterial communities at a sequence similarity cutoff of 90%.Remember that the Simpson diversity index is typically presented as (1-D) or (1/D).

> summary.single(shared=KBS.an.shared, label=0.10, calc=chao-shannon-simpson)

The ‘calc’ component of the command tells mothur the calculations you want completed. For a thorough

discussion of the calculators available in mothur see: <http://www.mothur.org/wiki/Calculators>

This command yields 1 file:

KBS.an.groups.summary Contains values for the requested diversity indices by sample

Note that, in this particular instance, sampling depth is approximately equal for the six samples. Therefore, we can request and compare single values of our alpha diversity indices. In situations where the alpha diversities of communities sampled to different extents is being compared, sample rarefaction curves for indices of interest should be generated for each community, and the respective curves should then be plotted together for comparison.

> rarefaction.single(shared=KBS.an.shared, freq=2, label=0.10, calc=chao-shannon-simpson)

**Based on these indices, at a 90% sequence similarity cutoff, does alpha diversity appear to differ between forest and agricultural soil bacterial communities? Do the rarefaction curves and the calculated diversity indices tell similar stories?**

**3) Evaluating beta diversity among forest and agricultural soil bacterial communities**

Remember that beta diversity is the extent to which the variety and distribution of ‘species’ varies *among* communities. In this section, you will evaluate beta diversity among agricultural and forest soil bacterial communities using two approaches:

1) OTU-based approach: nMDS, cluster, ANOSIM and SIMPER analyses using R

2) Phylogenetic approach: weighted Unifrac analyses in mothur

**OTU-based approach**

To conduct cluster, nMDS and ANOSIM analyses, you must first generate a distance matrix that describes the similarities (dissimilarities) among all samples. Similarity matrices can be based on numerous beta diversity calculators (<http://www.mothur.org/wiki/Calculators>), each with its own strengths and weaknesses, depending upon the research question at hand. Our research question in this exercise is whether standard agricultural practices influence the structure of soil bacterial communities. Therefore, we will use the Bray-Curtis similarity index to evaluate community diversity. For the nMDS, cluster, ANOSIM and SIMPER analyses, you will again be using a sequence similarity cutoff of 90%.

**Using R, create a community dissimilarity distance matrix based on Bray-Curtis**

**Generate an nMDS plot and a cluster analysis using this distance matrix**

**Conduct ANOSIM and SIMPER analyses as well**

The null statistical hypothesis in ANSOIM is that forest and agricultural soil

bacterial communities do not significantly differ.

SIMPER informs us of the contribution of each OTU to observed differences.

**Taken together, what do the nMDS, cluster and ANOSIM analyses suggest about shared bacterial community structure between forest and agricultural soil bacterial communities? Interpret the R-values and p-values in the ANOSIM analysis. What can you glean from the SIMPER analysis? Can you interpret these findings in the context of experimental design?**

**The visualization and statistical analyses you just completed were co-opted from macroecology and use an OTU-based framework. Why do some microbial ecologists view this as being substantively problematic?**

**Phylogenetic approach**

Recall from our in-class discussion that Unifrac analyses are not limited by OTU definitions, rather they are based on communities’ unique branch lengths within summary phylogenetic trees. To conduct Unifrac analyses, you must first generate a phylogenetic tree of all the sequences in the study. If you are satisfied with the relaxed neighbor joining algorithm used by Clearcut (Sheneman et al. 2006. Bioinformatics, 22, 2823), then you can generate these trees in mothur. If not, you can readily use trees generated using other software. For example, we have provided you with an optional Newick-formatted tree of your data, created using Arb.

**Conduct a weighted Unifrac analysis using mothur**

The null statistical hypothesis is that the structure of bacterial communities does

not significantly differ between agricultural and forest soils.

> clearcut(phylip=KBS.dist)

You may receive a message that HhiKelly is not in your groups file and will be discarded. This is okay, as

HhiKelly is a sequence from an outgroup organism we used to root our phylogenetic tree, and should

not be included in the analysis.

This command yields 1 file:

KBS.tre A neighbor joining tree containing all of your sequences

The weighted Unifrac option only considers two treatments at a time (i.e. no global test

is conducted). Therefore, we have provided you with a second groups file in which the

sequences from each treatment, regardless their sampling unit of origin, have been

lumped.

> unifrac.weighted(tree=KBS.tre, group=KBSlumped.groups, random=t)

This command yields 2 files:

KBS.tre.wsummary

KBS.tre1.weighted

Now conduct a weighted Unifrac analysis where sampling units are considered

individually, and multiple analyses are conducted. To do so, use the original groups file.

> unifrac.weighted(tree=KBS.tre, group=KBS.groups, random=t)

This command yields 2 files:

KBS.tre.wsummary

KBS.tre1.weighted

**What is the challenge of including biological replicates in a Unifrac analysis?**

**What do the weighted Unifrac analyses suggest about shared bacterial community structure between agricultural and forest soils? Are the suggestions the same as those made by the ANOSIM analyses?**

**What preliminary conclusions have you reached about the effect of standard agricultural practices on the structure of bacterial communities?**