# VMC 991 – Diseases of Marine Organisms

Sequence Analysis Exercise<br>
Thursday May 29th, 2021

This exercise expands our diagnostic toolbox for evaluating microbes associated with pathology and disease. We are going to start from a completed DADA2 workflow, which uses the Divisive Amplicon Denoising Algorithm to infer true biological sequences and taxonomy from sequencer reads. Taxonomy tables produced by DADA2 can then be used in downstream applications analyzing differences in microbial communities among different treatment groups, in our case the different microbial communities associated with oyster gills with and without hemocytosis/erosion across the three sampling sites. We will be using the software package phyloseq in R, which gives a powerful framework for analyzing microbiome data. You should have R and R Studio downloaded on your computer, but if not please do that now. All R scripts shown here are available in the "Data" folder on our course webpage.

Open R and install Phyloseq:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("phyloseq")

library(phyloseq); packageVersion("phyloseq")
library(Biostrings); packageVersion("Biostrings")
library(ggplot2); packageVersion("ggplot2")

theme_set(theme_bw())
```

The first step is to load the sequence and taxonomy tables from DADA2. These files are located in the "Data" folder on our course webpage.

```
seqtab <- readRDS("output/seqtab_cut_final.rds")
taxa <- readRDS("output/tax_cut_final.rds")
```

We will then construct what R calls a data frame, which is the most common way of storing data in R and the data structure most often used for data analysis. We will construct a simple sample data frame from the information encoded in the excel file describing our samples (also found in the "Data" folder):

```
samples.out <- rownames(seqtab)
sites <- read.csv("sites_cut.csv", fill = FALSE, header = TRUE)
samdf <- data.frame(Event=sites$Site,Group=sites$Group,ID=sites$Sample)
rownames(samdf) <- samples.out
```

We now construct a phyloseq object directly from the outputs:

```
ps <- phyloseq(otu_table(seqtab, taxa_are_rows=FALSE),
            sample_data(samdf),
            tax_table(taxa))
```

It is very common to have many rare taxa present in amplicon sequence data. These can often be biologically meaningful, but for now we are going to filter these rare taxa out of our analysis to focus on the most numerous taxa represent in each of our samples. The first step in doing this is to make a new data frame that filters out taxa found below some threshold, here we will apply a 1% threshold. That is, taxa that are found at less than 1% prevalence are not included in our analysis:

```
prevdf = apply(X = otu_table(ps),
            MARGIN = ifelse(taxa_are_rows(ps), yes = 1, no = 2),
            FUN = function(x){sum(x > 0)})
prevdf = data.frame(Prevalence = prevdf,
                TotalAbundance = taxa_sums(ps),
                tax_table(ps))
prevalenceThreshold = 0.01 * nsamples(ps)
keepTaxa = rownames(prevdf)[(prevdf$Prevalence >= prevalenceThreshold)]
ps1 = prune_taxa(keepTaxa, ps)
ps2 = tax_glom(ps1, "Genus", NArm = TRUE) #glom the pruned taxa
```

For this exercise we will only look at the top ten taxa from all samples:

```
top10 <- names(sort(taxa_sums(ps2), decreasing=TRUE))[1:10]
ps2.top10 <- transform_sample_counts(ps2, function(OTU) OTU/sum(OTU))
ps2.top10 <- prune_taxa(top10, ps2.top10)
```

We are now ready to use phyloseq. I like to first have a look at an ordination plot to see how similar our samples are to each other. We will use a Principal Components Analysis here:

```
ordu <- ordinate(ps2.top10, method = "PCoA", distance ="bray")
p = plot_ordination(ps2.top10, ordu, color = "Event", shape = "Group")
p = p + geom_point(size=7, alpha=0.75)
p = p + scale_colour_brewer(type="qual", palette="Set1")
p
```

We will circle back around to what this may mean for differences in microbial communities among the sites. But first let us have a look at the taxa that are driving these patterns. I like looking at stacked bar graphs showing the relative frequency of different taxa among the samples:

```
ps2.top10.BS <- subset_samples(ps2.top10, Site=="BS")
ps2.top10.SC <- subset_samples(ps2.top10, Site=="SC")
ps2.top10.NR <- subset_samples(ps2.top10, Site=="NR")
names <- taxa_names(ps2.top10.BS)
p1 = plot_heatmap(ps2.top10.BS, taxa.label = "Genus", sample.label = "Group",low="white",
high="#000033",
        na.value = "white", sample.order = "Group",taxa.order = taxa_names(ps2.top10.BS),
        trans = identity_trans())
p1 = p1 + theme(axis.text.x = element_text(size=7, angle=0, hjust=0.5, vjust=0.95)) +
        theme(axis.title.y=element_blank()) +
        theme(legend.position="none") +
        ggtitle("Bogue Sound: Unimpacted Control") +
        theme(axis.title.x=element_blank())
p2 = plot_heatmap(ps2.top10.NR, taxa.label = "Genus", sample.label = "Group",low="white",
high="#000033",
        na.value = "white", sample.order = "Group",taxa.order = taxa_names(ps2.top10.BS),
        trans = identity_trans())
p2 = p2 + theme(axis.text.x = element_text(size=7, angle=0, hjust=0.5, vjust=0.95)) +
        theme(axis.title.y=element_blank(),axis.text.y=element_blank()) +
        theme(legend.position="none") +
        ggtitle("Lower New River: Impacted by Spring Mortality") +
        theme(axis.title.x=element_blank())
p3 = plot_heatmap(ps2.top10.SC, taxa.label = "Genus", sample.label = "Group",low="white",
high="#000033",
        na.value = "white", sample.order = "Group",taxa.order = taxa_names(ps2.top10.BS),
        trans = identity_trans())
p3 = p3 + theme(axis.text.x = element_text(size=7, angle=0, hjust=0.5, vjust=0.95),
        legend.title = element_text(size = 16)) +
        labs(fill = "Relative\nabundance") +
        theme(axis.title.y=element_blank(),axis.text.y=element_blank()) +
        ggtitle("Salter Creek: Impacted by Spring Mortality") +
        theme(axis.title.x=element_blank())
grid.arrange(p1, p2, p3, nrow = 1)
```

Looking at the ordination plot and different taxa among the sampling sites, what can we say about what may be driving springtime oyster mortality? Use your best judgement and any resources to elaborate on taxa present in samples collected at sites experiencing spring mortality.