## REMNet Tutorial, R Part 4: Creating Taxa Plots in R

By Jessica Lee Joyner

Hello, my name is Jessica Lee Joyner, and I teach and do research at the City University of New York.

Many times people are interested in the specific organisms that are present. To do that we can use one of the tools within the microbiomeSeq package to get an idea of the top 20 orders or any taxa level you're interested in, but we're going to do orders for our data. So I have a comment here telling me these are—the following lines are for taxa plots, and that I am going to work at the order level. So my new variable name for this dataset is going to be Mseq.Order, and I'm going to be using the command, or the function, taxa level, which requires the Phyloseq object.

However what's interesting with this function is we have to transform the data. So it's a small, little element of this line, but very important. In this case you have to have "t(BioBlitzdata)" in parentheses. This is going to transform the data frame into the way that can be interpreted by taxa level. And then I'm going to say which taxa level I want to organize this to, and that's order.

So my line is on-- my cursor is on that line, and I'm going to run. Because we're working with such a large dataset, something like this function where we're asking the computer to summarize a lot of the lower taxa, so this is going to summarize all of the species and the genera and the class. We're going to be able to have a smaller dataset to work with and to visualize, but it does take a while to calculate how we bring all that information and the abundances for the taxa up to just the order level.

OK, so we're done calculating. We have that carrot, so I can go on to the next step, and that next step is to plot these taxa that we've now summarized. So we have "plot_taxa", and the first piece of this new Phyloseq objects that's summarized at the order level. We're going to group again by project, and what's fun about this one, and why I like this function is this piece here call "method = "hellinger"".

Here we have a built-in element that does a little bit of the beta diversity for us, the similarity between all of the communities for each of our samples. So the common method that would be appropriate for a data set is called "hellinger", and that's why that's listed here. It has a limit of only listing 21 elements, so we are going to say the number of taxa equals 21. Any other taxa will still be represented except they'll be grouped into a "other" category. So we'll have the top 20 taxa, or top 20 orders identified, and then the remaining will be "other" to group all of the other orders into one, and this really helps with visualizing the data.

You have an option to save the file. We're not going to save the file in this case, so that's null or empty. And here, um, you see where I'll print that plot. It's another way of calling what we have now saved. So we have our new plot named TaxaBar_plot. This is again going to save it so we don't have to recalculate every time, and we will, after everything's typed in and we've checked, hit run.

OK, I am going to just highlight TaxaBar_plot as the variable name, and hit run for it to come up. Going to resize the screen. This is a lot of information we're asking the computer to do at once, so it probably will take a minute or two.

And it's a rather big figure, so if you need to adjust your screen, you can just drag and adjust it, and wait a moment for it to recalculate the figure dimensions. OK, so what you are seeing in this figure are the five sites, five different projects, and color-coded for the top 20 orders. With this indicator on the bottom right corner of the LCBD, and this is the size of the circle is going to tell you how unique the sample is from the rest of the dataset.

Notably we have here over in Alley Pond, one sample that was incredibly different from all the others. You can see a lot more of this pink indicated taxa compared to all the others, which have this gray on the bottom. The gray on the bottom is the group of others, so it's interesting for this sample in Alley Pond that there were less taxa that were rare, or less abundant, so they didn't get grouped into the "others", and it was mostly the Pseudomonadales, which is our pink according to the legend.

I like creating this taxa plot because it is a quick idea of looking at who were some of the major taxa present, and how the samples differed from site-to-site. In my case project and site are going to be very similar. One thing that I can point out as far as guiding and interpreting figures like this, is I do know that some of the samples, or the sites within the locations of Alley Pond, Brooklyn Bridge and Freshkills. All these three on the left had soil samples that came from a wet environment, whether it was a marshy area or storm drain area.

And that might explain why some of them, we see an enrichment in the blue, and some of this green, and then we can look through try to get an idea of where to start digging deeper into the data, whereas Inwood Hill in New York Botanical Gardens had a commonality of this purple.

Now if your computer's like mine, you might still need to adjust to get the screen resolution to list all of your orders, but you can usually just keep adjusting, and it'll load, and you'll have all of them listed. And again a reminder that if you want to edit this in any specific way for your project, you can do that through using the ggplot2 tools.