## **REMNet Tutorial, R Part 3: Creating Alpha Diversity 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.

Now to give you an idea of these files at this point, I'll take a moment and show you what my metadata file is, the converted mapping file. To show you the contents of this file, I am just going to type in the console the name of that table. So we brought it in, and called it BioBlitz.meta.

Capitalization does matter here, so if you notice the two B's are capital, and that's important to remember whenever you're typing straight into your console, and to avoid frustrations.

All right, I'm going to adjust my screen so that we can see a little bit more of what this stable has. All right, it's broken in to three chunks because of the size of our screen. We see the first column is our sample ID. This is the same thing that is in the sequences to identify which samples match into this mapping file. The barcode sequence, linker primer sequence, and barcode name are very specific information that were used during the sequence analysis from the original raw data. To create this dataset, I used various projects, so we have different project names in detail and then summarized site, time, year and date time are all characteristics of my samples that could be an interesting way of looking at the data.

And then the last column that is in this file is a description. For me, I try to make my description be just that, as descriptive as possible. The default is that it's going to be the same as your sample name. What I have here is my project plus site plus day and time. OK, so that's a brief overview of what metadata I am bringing in.

Going to reset my screen. This way we're emphasizing my source code, and here's where we can begin doing some early analyses. A quick idea of getting an overview of the information you have in your samples would be to look at the diversity. To do diversity, we think about it in an ecological context. There's alpha diversity, or the diversity that's within your sample, and then there's beta diversity, the diversity that's between your samples, or the bigger picture.

The first look into this we're going to focus on the alpha diversity. So I've created here a variable called "alpha\_meas", or alpha measures for short, and I again created a list of what I want to look at. So I have combined observed Chao1 and Shannon. I have a comment to the right on this line of other options that I could explore, but this first pass we're just going to leave them as a note at the end of the line.

So I'm going to create this variable by having my line on there and hit run, and now our first piece of visualizing the data. We're going to use the command "plot\_richness" using

our Phyloseq object BioBlitzdata, and we want it to group based on our projects. Note I combined a few projects, and the sites that were within each project to be color-coded.

When you're starting off writing these scripts, sometimes you can get a good help of knowing what each step is that's within that function when you highlight or hover. So you might see that this has popped up a few times. In this case we have "plot\_richness", and in that yellow box it's indicating that the first list-- first item in this list is your Plyloseq object. Then your next list is going to be how it is organized on the x-axes. Then the next would be your color, then shape, and so on, all the information in that yellow box. Here I'm mostly interested in those first three pieces: the Phyloseq object, want my x organized by, and the color, and I want this to be analyzing based on that variable I made "alpha\_meas".

So here I'm using the function. I have it completed out the way that I want. Note that the variables, or the elements of the mapping file, the metadata that I'm referring to, are in quotes. That's important to use in this line, and then I've again used a semicolon because I want to store the plot that I make into the memory of the program, and that way I can refer to it without having to recalculate it every time.

So in this line, and it's our first plot, I've named it plot1. Then after the semicolon, I recall plot1. If I did not have semicolon plot1, it would just store the plot, and we wouldn't see it. So just to show you what that would look like, I'm going to only highlight plot1, and hit run.

And you can see we don't have any output until I highlight the remainder of that line, and hit run. OK, so now we have actually calculated our alpha diversity for the samples to understand more about our data. I've resized the screen so you can get a better perspective of what our plot is, and the reason there is the package ggplot2, is that's the foundation of all of the plots we're going to make. So if you need to make small changes to this plot or any future one, you would look up how do you change a figure or a plot within the package ggplot2.

Here we have three charts: one for observed, one for Chao1, and one for Shannon. Now I'll take a minute here to tell you a little bit about what you're looking at. Our observed, this first chart, is the number of OTUs, or unique organisms, in each of these samples. Our Chao1 is the estimated richness; how many unique organisms could be present in these samples. And then your Shannon is a diversity index that takes into account how the number of unique organisms, or richness, relates to the abundance of each of those organisms.

And here we have that project on the x-axes, and all of the color-coding are based on the different type of samples that we took soil from. Some general conclusions when you're looking at a chart like this is which sample, which site, had greater diversity? Which ones had a high range of diversity? Working with this test dataset, you can see that we have a lot of differences across the samples at least for observed and Chao1. There's much more variation for this middle site Freshkills when you see how the points

are distributed compared to Alley Pond, which except for one have a tighter spread. Because it's an index, we end up seeing a bit more similarity for our sites across Shannon.