**REMNet Tutorial, R Part 2: Importing Your Microbiome Data into 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.

What we are bringing into R is going to be two key files. We're going to bring in a .biom file and the mapping file, or metadata file. Now for the .biom file, this is a unique file that is formed or has been developed over the years of doing sequencing, and it is a database of sorts.

What is important to know when you're importing a .biom file into R, and into these packages that we've installed, it has to be in the J-S-O-N format, or JSON format. I've already done that and in another video, or you can look and do an internet search, you can convert your .biom file to the necessary format.

The mapping file: typically a mapping file is used in the programs for qiime or other analysis platforms. It's called a mapping file because it is the way that the computer is able to identify all of your sample information to the sequence indicator, which is typically your barcode.

When we're working in R here, we only need these two sources of files because in the .biom file, you're going to have your list of samples, the taxonomy, and the number of sequences for each of the taxa that are present in your samples. You're using your mapping file and bringing it into R here for all of the metadata, all of the supplemental information that helps understand your project.

So my lines 24 and 25, I have my .biom file, and I have named it BioBlitzbiom, and I'm using the command "import_biom" with the file path of my .biome file. You can get to this by getting the info for your file, or you can drag and drop into your source code, and it should copy everything through. But that may vary based on different machines.

It doesn't matter how you do it as long as it's beginning to end for the path or the location of this file, and it is in quotation marks so the computer understands that you're giving it a single piece of information.

The rest of what is needed for this function is to tell the-- tell this function how that taxonomy was identified. Different methods can have different ways of organizing the taxonomy. For this .biom file, I used GreenGenes. So here I have that it's going to "parseFunction" and "parse_taxonomy_greengenes". So this line is now ready: the "import_biome" to import my .biom file. The next step is to import that mapping file. The mapping file is a tab-delimited text file. The key difference between this mapping file versus what was used when analyzing the raw sequences, is the very first line of the mapping file cannot have the pound sign, and it's the same reason that I told you about creating notes to yourself in the script that if that pound sign is inside the file, R will

interpret that as a comment, as just a note. So you have to remove that symbol before you can bring that into R appropriately.

So I've already done that and my mapping file is ready to go, and we're going to use the built-in function of R called read.table. So I'm using read.table, the list of my file path for my mapping file, and then because I do have information on that first line that we want the computer to read, I am saying that the header is true, that it is present.

Depending upon your project, the "import_biom" file, the "import_biom" process could take a while based on the size of your file. OK, so our .biom file is loaded in, and now we can load the mapping file. And you can see just the size difference in the information that are in these two files how it took a bit for the .biom file to come into R and really fast for us to import the mapping file.

All right, so our two key files are now in R. We can now build a Phyloseq object. The way that we're working through this data, the main program is going to be Phyloseq, and this program, or the package set, it pulls all of the information together into one object. And then as we want to visualize our work with the data in different ways, or in different pieces, we can refer back to the single object. So here I'm going to identify the sample data.

I'm calling it by a BioBlitzSampledata. I'm using the function "sample_data", and here is an interesting piece. We have to tell the computer that our BioBlitz metadata, or BioBlitz.meta, is interpreted as a data frame. So this is a built-in R feature that if you just say "as.data.frame" and then you list the information, it'll wrap it in too so that we can identify it as our sample data. So here I have the metadata table that information is coded as rows.

And here is an important piece. An important piece of whenever we're combining different files and different tables is that they're sorted the same way. So we have here, telling the computer how we have our data sorted. So we want the row names to be identified as the sample names of the BioBlitzbiom file.

And that's what I have here. So I have the row names for this data frame are going to equal the sorted list of-- the sorted list of the sample names in the BioBlitzbiom, and that all of this is a string, so all the characters inside are factors.

Now my cursor is on line 27, and I'll hit run. So now we have our sample data identified, and we have now where we can build together the .biom file and a sample data file into our Phyloseq object that will be called BioBlitzdata.

Now here is an important point to check your output. So looking at our console, we have our BioBlitzdata output summarized, that we have our OTU table, the .biom file, and has identified that there are 86,280 taxa present for 33 samples. Our sample data reports that there are 33 samples, and there's 11 different elements, or variables, that can describe these samples, and the taxa table, or the taxonomy table, again repeating that

there's 86,280 taxa across seven taxonomic ranks. So this is going to vary from phyla to species.