REMNet Tutorial, R Part 1: Getting Started in R with Microbiome Data

By Jessica Lee Joyner

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

Today we're going to be going through how to work with R to visualize microbiome data. When you are doing a project for microbiomes around the built environment or any specific project, there's a lot of data that comes out of the sequencing.

And it's using R and down the line Excel that you can get an idea of what the story is in your data. So we're going to work with R today and the first step in working with R, we're going to talk about the basics and a lot of the packages that are going to be facilitating your analysis.

Working in R, the first step is to identify what packages you need, and process to download them. We have RStudio downloaded, and using the R version 3.5.1. And the process that we will use is a combination of built-in tools as well as getting Bioconductor and installing step-by-step.

So the first thing I want to point out is when you are writing in R, the way of putting notes for yourself and knowing the steps that you are taking, you can put a pound sign at the first of the line. Anything that follows that pound sign is not going to be read by the computer or interpreted as a code. I have in my script, it starts with the install packages, and then the first package that I want to install is called Phyloseq. I'm going to install it by directing towards the source or the website where we can get the install software, and that is Bioconductor.

So my line 3 is directing the program to the Bioconductor website, so I'm highlighting that script. Or I can just be on that line, and hit run. So the next thing as far as working in Rstudio is to run from our source code. You just have to be on that line, and then you can hit the button to run.

The output is what you will see on the right hand side or where your console is. Here I have a warning. Warnings will mean that it'll still happen, but there was a little bit of an issue. More of a concern is when you have an error, and that's when you really need to troubleshoot. For us, we're just going to go onto the next step and install Phyloseq. So I'm on line 4, and I will hit run.

While this is running, I'll tell you a little bit about how to stay organized. For working in R, we create a source script, and that is a way that you can go back to the same process every time you're working on a project. So my source script is on the left hand side of the screen, and I have named it a project called Bioblitz Five Year Summary.

Now when I'm writing this, I have plenty of places where I have given myself information in these “comments” is what we call where there's that pound sign, and a lot of times I can choose when and which lines I want to run. So after the first time I'm running and setting up this project, I do have a note that it was up to line 10. I won't have to repeat the earlier lines if I am working on the same computer. Another way to stay organized when you're writing, is to have tabs or indents as you see in line 7. This is good practice of being able to identify when different tasks or parts of a task are being completed. Line 6 is using a common command or script conditional statement called “if”, and this is telling the computer that if line 6 is all true, then do line 7.

OK, now after waiting for those programs to, for at least Phyloseq to install, you can look over on the console. And you have a blue carrot again, or at least a carrot. I’m going move my cursor over here and hit a few times so that you see what I'm talking about. That is telling me that the computer is ready for the next command. So now I can go on and complete those 6 and 7 lines all the way to 8. So in this case, I am now highlighting three lines at a time, and I'm going to run all three at once.

While that is again installing, I'll talk you through which packages we need. So so far we have installed Phyloseq, which is one of the heavy lifting packages. We’re going to rely on this the most when it comes to analyzing microbiome data. And then when we did line 6 to line 8, we installed packages called DESeq2, GO.db, impute and preprocessCore.

These are four different packages that are needed for all the functions and the dependencies for some of the key analysis packages. However they’re located in a different place, and we needed to install them in a different way than Phyloseq. So in the second round, we used this BiocManager to install those four packages.

The remaining packages are more straightforward to install, and some of the key ones that I want to highlight are your ggplot2. This is an amazing plotting tool and is necessary for visualizing and creating the graphs that we will for the microbiome data. Another one is vegan. Vegan is a typical ecology package and will use some of the resources and scripts that are within this package to analyze our data. Now I'll give a full list of all of what we have listed on line 11 here, but most of them are working in the background of the overall process that we’re going to go through. So some of these are going to be used in other cases, but for today those are the main ones we’re going to focus on.

To install that cluster of packages, you have two ways that are part of the R system natively. We can do it via this source code line on line 13, and through going Tools > Install Packages…, and you can list all of the packages in the pop-up window.

I'm just going to use what I have written out, so I'm not going to type into that window. A basic R feature that I'm going to take advantage of for this next step is creating a list of objects. So here where I was pointing out the different packages, I have a “c” and an open parentheses. In R this is telling the program to combine. So we have “c”, open parentheses, then a list of all of the packages we’re going to need, and I made it into this variable. So I named that list “necessary_packages_install”.

And you can see honestly in line 12 where I've excluded some packages that were giving some earlier issues. That way when it comes to installing this many packages, I can do it in one step. I can have on line 13 "install.packages" and then the name of that list that I created as a variable.

So you see again on line 11 the combined list of all the packages I want. Naming it that variable where then in line 13 I call upon that variable to install them all at once. You can see the progress of the download based on the red bar as it extends, and so we're still waiting for two out of the four previous ones that we're installing to go through.

OK so all of the previous ones that we ran for line 6 to 8 have loaded, and now I'm going to do the packages listed in lines 11 to 13. For this case I don't need R to restart so I'm going to just select "No". Now that all of those packages have installed, we can do our last one that is going to manage a lot of the unique plotting options that I'll show you, and this is a package called microbiomeSeq.

Now this is important for me to point out that microbiomeSeq is early in its development and still in what is called the beta testing phase. It is really good at visualizing the data, so I'm going to go ahead and use it through our tutorial today, but make sure that you either do a internet search or follow up on this independently down the line if there's been any updates to microbiomeSeq.

Now this package has to be installed from GitHub, which is a source for computer scripts, R codes online, so we're going to need to first install the package to be able to go to GitHub through R. While that's going, I'll show you another tip in R of way of organizing. I did want line 16 both to install and then load the tool for downloading this new package. So I put them on the same line.

So using a semicolon, I can separate out two different tasks that would otherwise be different lines in the source code. Here I have one independent feature of install.packages, and then "library(devtools)". "library(devtools)" is going to tell the computer to load that package. And that's loaded, so now I can install from GitHub. All right. Scrolling down, I'm going to load the rest of the packages, so here I do that same thing where I created a variable in the list. I'm going to highlight and run.

Great. So it looks like we have everything installed, and finishing up, what we need to be able to use R and analyze microbiome sequences, or specifically for us, the bacterial sequences.