

Progress Report

Charles Bond

11/15/2020

Current Status: Stuck, cannot run script

Looking for the tools to run the PCA analysis shown in figure 2 (Bergelson and Horton 2019), I have been unable to run the code so far. I appear to have all the files, code, and data from the two githubs and the data repository provided in the **Availability of Materials and Data** section of the article of interest. The code looks good, but I can't get past this setup error: the script seems to be having trouble accessing the files that came from the repository.

The very first lines of code, which give me an error, in the Rscript file for the PCA analysis are as follows:

```
## This R-script
##
## Author: matt.horton
#####

rm(list=ls());
require(vegan); require(lme4);

## Loading required package: lme4

## Loading required package: Matrix

source(paste0(Sys.getenv('ROOT_MICROBIOME_DIR'), "/hdr.microbiome_methods.R"));

## Warning in file(filename, "r", encoding = encoding): cannot open file '/
## hdr.microbiome_methods.R': No such file or directory

## Error in file(filename, "r", encoding = encoding): cannot open the connection

source(paste0(Sys.getenv('ROOT_MICROBIOME_DIR'), "/hdr.base_methods.R"));

## Warning in file(filename, "r", encoding = encoding): cannot open file '/
## hdr.base_methods.R': No such file or directory

## Error in file(filename, "r", encoding = encoding): cannot open the connection
```

```
#####
## metagenomic variables:
#####
numberOfPCsMicrobes <- 5;
organ <- "root";
phylogeneticMarker <- "ITS";
otuCutoff <- 97;
minimumReads <- 1;
otuThreshold <- 2;

normalizationMethod <- "raw";
analyticalMethod <- "glmer";
ordinationMethod <- "pca";
```

The `source` function cannot access the required file. I checked that the requested files are indeed in the working directory, and I have tried moving them to the main folder and checking the code, I don't know why R can not access these files.

The README says:

To avoid using hardcoded paths, we use R's ability to load environmental variables (using the command syntax `Sys.getenv('variable_name')`). The variables that need to be loaded, by the user, can be done so using the following bash code:

```
export ROOT_MICROBIOME_DIR='<your_directory>/root_microbes/'
```

I have not yet been able to figure out what to do with that information.

I wanted to run some kind of PCA even if only a subset of the data, but the PCA is based on metagenomic data which I have never worked with. Just looking at it I can see that there is a lot of code involved in making the data usable, which I wouldn't know how to do without even the first line of the original code to work.

Previous Update, 11/08/2020

Based on the availability of the data and the posting of the code on github (<https://github.com/bvilhjal/mixmogam> and https://github.com/mahort/root_microbiome.) I have selected a paper on the *Arabidopsis* root microbiome (Bergelson and Horton 2019) as the subject of my next replication attempt. If it goes well there are a lot of figures in this paper that would be worthwhile to replicate, but the one I am most interested in is the PCA shown in Figure 2 of the paper.

Once again, this paper uses the package `vegan` to do the PCA, so I've got that.

Status: Making progress, found PCA file in github and just trying to figure out where to start.

I am confused on how approach this, like I made a github repository for my replication, but they have a huge github repository for the original, and I'm wondering if I should download files individually or try to pull the whole project down in case there are files it's referencing that I miss? I downloaded whole github repository as a zip file and put it here, but I guess I need to alter relative paths or something to make it work? I tried running one of the PCA Rscript and get the error "Error in `getData(cutoff = otuCutoff, combined = FALSE, marker = phylogeneticMarker, : unused arguments (cutoff = otuCutoff, combined = FALSE,`

marker = phylogeneticMarker, tissue = organ, minReadsPerSite = minimumReads, minSamplePerOTU = 1, whetherToGetTaxa = TRUE, rarefy = (minimumReads > 1), numberOfRarefactions = 1)", which idk what that means but I guess I need to tinker with this more.

References

Bergelson, Joy, and Matthew Horton. 2019. "Characterizing Both Bacteria and Fungi Improves Understanding of the Arabidopsis Root Microbiome." Journal Article. *Scientific Reports (Nature Publisher Group)* 9 (1). <https://doi.org/10.1038/s41598-018-37208-z>.