Introduction

The importance of 16S rRNA gene amplicon profiles for understanding the influence of microbes in a variety of environments coupled with the steep reduction in sequencing costs led to a surge of microbial sequencing projects. The expanding crowd of scientists and clinicians wanting to make use of sequencing datasets can choose among a range of multipurpose software platforms, the use of which can be intimidating for non-expert users. Among available pipeline options for high-throughput 16S rRNA gene analysis, the R programming language and software environment for statistical computing stands out for its power and increased flexibility, and the possibility to adhere to most recent best practices and to adjust to individual project needs. Here we present the Rhea pipeline, a set of R scripts that encode a series of well-documented choices for the downstream analysis of Operational Taxonomic Units (OTUs) tables, including normalization steps, alpha- and beta-diversity analysis, taxonomic composition, statistical comparisons, and calculation of correlations. Rhea is primarily a straightforward starting point for beginners, but can also be a framework for advanced users who can modify and expand the tool. As the community standards evolve, Rhea will adapt to always represent the current state-of-the-art in microbial profiles analysis in the clear and comprehensive way allowed by the R language.

Organization

Rhea is composed of 6 steps that can be run independently or as a set.

1. Normalization
2. Alpha-Diversity
3. Beta-Diversity
4. Taxonomic-Binning
5. Serial-Group-Comparisons
6. Correlations

Running them in the given order simplify the process as the output of each step is often the input of the next. There is an extra folder where the original data is recommended to be placed to keep the analysis of one study in a compact and organized structure. Inside the 0.Original-Data folder there is another folder containing the template files used for the Rhea presentation that can be used for exploring the different steps of Rhea. For a quick and effective way to obtain OTU-tables and other required files for further analysis in Rhea, when starting from raw sequencing data, please use the analysis functionality of IMNGS (www.imngs.org). The output of the UPARSE based clustering of sequences to OTUs is fully compatible with Rhea scripts. Before running any script make sure you read and fully understand the corresponding ReadMe file that can be found in each folder.

Script structure

All R scripts in Rhea follow a common structure design to help users orient and accelerate usage. Therefore every script would have a Commentary, Initialization and Main sections, Those are explained in detail bellow.

Commentary section

These are comments lines that add no functionality to the script but they exist to help the user understand the concept of the script and guide him through the script requirements. The user targeting comments start with the hash sign and are followed by a back-tick (#`) to be distinguished from ordinary R commands annotation lines starting with a hash (#) intended for advanced users. We suggest all users to read the introductory and usage commends in all scripts they use. An example of those commends can be seen bellow:

```
#' Version 1.6
#' This script was last modified on 11/02/2016
#' Script Task: Normalize OTU tables
#'
#' Normalize abundance values of the input OTU table
#' calculate relative abundances for all OTUs based on normalized values
#'
#' Input: Please enter following parameters
#' 1. Set the path to the directory where the file is stored
#' 2. Write the name of the OTU table of interest in quotes
#'
#' Output: The script is generating four tab-delimited files
#' 1. Normalized counts without taxonomy information
#' 2. Normalized counts with taxonomy information
#' 3. Normalized relative abundances without taxonomy information
#' 4. Normalized relative abundances with taxonomy information
#'
#' Concept:
#' Normalization via division by the sums of sequences in the given sample
#' and mulitplication by the minimum sum across all samples
#' It is used instead of the classic rarefactioning approach
#' to prevent confounding effects of subsampling sequences with possible consequences on diversity
#'
#' Note:
#' Files are stored in the current folder
#' If a file is needed for downstream analysis it is also stored to the apropriate folder
#' If original folder structure is maintained.
```

Initialization section

In this section the required and optional parameters necessary for the execution of the script are presented to the user and he is requested to change them accordingly. One common setting that needs to change in every script is the path where the script resides. This is important in order to keep all relevant files in the respective folders and avoid overwriting or confusion. Therefore all required files need to be placed within the folder of the script. Unless Rhea users want to modify the scripts no changes are needed after this section and the user can select the whole script and run it. An example of this section can be seen bellow:

```
################################################################################
######             Set parameters in this section manually                  ######
################################################################################

#' Please set the directory of the script as the working folder (e.g D:/Rhea/1.normalize/)
#' Note: the path is denoted by forward slash "/"
setwd("D:/Rhea/1.Normalization") #<--- CHANGE ACCORDINGLY

#' Please give the file name of the original OTU table with taxonomic classification
file_name<-"OTUs-Table.tab"  #<--- CHANGE ACCORDINGLY

######                  NO CHANGES ARE NEEDED BELOW THIS LINE               ######
```

Main Section

This is where the actual R commands start and based on the variables set by the user will be executed and process the input files accordingly. No modifications are required in this section. If the previous section was set correctly the only thing left is to select all (Ctrl-A) and run. See example bellow:

```
################################################################################
######                             Main Script                              ######
################################################################################

# Load the tab delimited file containing the values to be be checked (rownames in the first column)
otu_table <-  read.table (file_name,
                          check.names = FALSE,
                          header = TRUE,
                          dec = ".",
                          sep = "\t",
                          row.names = 1,
                          comment.char = "")

# Save taxonomy information in vector
taxonomy <- as.vector(otu_table$taxonomy)
```

Requirements

In order to run the scripts is important to have the R language and environment installed first (available here). We strongly recommend the usage of the R-studio to simplify usage and enhance productivity (available here). Required packages would be automatically installed the first time the scripts are run so internet connection at least for the first time is expected. The scripts are platform independent and should work in all systems that support R.

Installation

No installation is required for Rhea. After downloading the Rhea project, decompress the files in the desired destination and they are ready to use. We recommend to keep one copy of Rhea files in each study to avoid mix ups and to look from time to time for new and improved versions of Rhea scripts as the project, as all living things, will evolve.

Citation

If you use Rhea in your work please cite/attribute the original publication:

Lagkouvardos I., Fischer S., Kumar N., Clavel T. (2017) Rhea: A transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons. PeerJ 5:e2836 https://doi.org/10.7717/peerj.2836