Introductory tutorial to help you taxonomically identify your ribosomal sequence data

D. Green & F. De Boever

Introduction

This is a brief introduction and companion tutorial to that provided for the CCAP Bioinformatics Gateway. It's aim is to show you the value and power of the unix/cygwin command line and R to perform many bioinformatics-related functions.

What is required

- R
- unix (for mac/linux users) or cygwin (for Windows users)
- the custom script provided in the course material

Although the script should install the required R packages automatically (dplyr, Biostrings and dada2), you can install them manually if you like:

```
install.packages("BiocManager")
BiocManager::install(c("Biostrings", "dada2"))
install.packages("dplyr")
```

How to access course files?

- https://fdboever.github.io/CCAP_course_2021/
- All files used in this tutorial are accessible via github

Step 1: Download the script

go to the github repository and download the folder

Inside the folder you will find the R script called `Taxonomic_ID_with_Dada2.R`. Don't hesitate to have a look inside, and try to understand how the script is build up and what it may do. You can open it with R, or in any text-editor such as notepad, textEdit or BBedit.

Step 2: Run Taxonomic_ID_with_Dada2.R

Note, that where ever you start the script from, will be where your output files will be saved

NB. The following examples assume you have navigated to the folder that contains the R script and your fasta file.

if using unix command line and R is installed, you can call Rscript to run scripts.

Arguments

1. fasta file (for example 18S.fasta)
2. database name (`pr2` or `silva`)
3. molecule (`18s` or `16s`)

Usage

```
Rscript Taxonomic_ID_with_Dada2.R <fasta file> <database> <molecule>
```

For example, when for a fasta file with name 18S.fasta, we chose to use the PR2 database and specify 18S

```
Rscript Taxonomic_ID_with_Dada2.R 18S.fasta pr2 18s
```