

Pipeline – Phase 1: Normalize 16S Copy Number & MicrobiomeAnalyst-compatible file

Preparation of the files

1. Data to analyze – to search for the data, go to the file from Novogene and follow the path: result_[...] > 02.FeatureAnalysis > tables > sample > featureTable_Relative > **featureTable.sample.total.relative.xls** (this is the data file). The file should look like this:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	OTU_num	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR
2	ASV0	0.0070977	0	0	0.000184	0.280455	0.001750	0.180133	0.486667	0.411821	0.456718	0.003660	0.005724	0.306073	0.275837	1	Bacteri	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae					
3	ASV1	0.012588	0.005019	0.005729	0.011033	0.014478	0.300839	0.013561	0.040183	0.052660	0.110212	0.046630	0.027364	0.077476	0.048510	0.070800	1	Bacteri	Firmicutes	Clostridia	Lachnospirales	Lachnospiraceae	Oscillatoria	Oscillatoria		
4	ASV2	0.005099	0.055273	0.123095	0.235782	0	0.032547	0.004198	0.010959	0.005390	0	0.012282	0.004140	0.049020	0.010309	0.507871	1	Bacteri	Firmicutes	Bacilli	Epiphythiales	Epiphythiellaceae	Epiphythiellaceae			
5	ASV3	0.025098	0.016007	0.167056	0.094484	0.040573	0.027046	0.033570	0.077737	0.031613	0.014384	0.015296	0.033568	0.020436	0.024386	0.009070	1	Bacteri	Actinobacteria	Coriobacteriales	Coriobacteriaceae	Coriobacteriaceae	Coriobacteriaceae			

This data file should be adapted to the following .csv format to run the scripts:

[illegible]

Eliminate “k_”, “p_”, “c_”, “o_”, “f_”, “g_”, “s_” from the “Taxonomy” column in the original file. To do this, use the “Substitute” function in Excel.

Localizar e substituir

Localizar Substituir

Localizar:

Substituir por:

Substituir todos Substituir Localizar Tudo Localizar Seguinte Fechar

NOTE: Verify you don't have any blank rows, for this look for "Unassigned" and eliminate that row

2. 16S copy numbers – to normalize the data, use the file **16S_copy_numbers.csv** file which can be obtained from: https://github.com/VaneBR/Project/blob/main/Files/16S_copy_numbers.csv or https://drive.google.com/drive/u/0/folders/18yojrGbjKc_-GAFdOLppHP2UWXFu3f1.

Steps to use the script:

The script described below is available in the following link:

<https://colab.research.google.com/drive/1HEbzgQUZYjxR7zugMDifL0fF4737H6K5> or

https://github.com/VaneBR/Project/blob/main/Scripts/normalize_16s.py.

1. To normalize and create a compatible file to introduce in MicrobiomeAnalyst, upload the `data_file.csv` and `16S_copy_number.csv` into the folder of GoogleColab.
2. After uploading the `data_file.csv`, change its name in the script:

```
# === Read the original CSV file ===
with open("INSERT NAME HERE.csv", 'r', encoding='utf-8') as f:
    lines = f.readlines()
```

3. Run the script and obtain the output file “matrix_normalized_microbiomeanalyst.csv”. This output file contains the normalized data, if this was your goal, you can opt to stop your analysis here. Nevertheless, if you want to perform a statistical analysis in MicrobiomeAnalyst, continue with the next pipeline.