

Pipeline – Phase 1: Selection of Microorganisms of Interest (Part 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_nur	V.A.0h	V.B.0h	V.C.0h	V.D.0h	V.E.0h	V.C.N.A.24h	V.C.N.B.24h	V.C.N.C.24h	V.C.N.D.24h	V.C.N.E.24h	CEL.A.24h	CEL.B.24h	CEL.C.24h	CEL.D.24h	CEL.E.24h	Taxonomy									
2	ASV0	0.007697	0	0	0	0.000194	0.293455	0.001750	0.019313	0.468667	0.411921	0.456718	0.003640	0.005724	0.308072	0.275837	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__									
3	ASV1	0.013588	0.005919	0.005279	0.011032	0.014478	0.100819	0.013561	0.040183	0.052660	0.110212	0.048630	0.027344	0.077476	0.046519	0.076893	k__Bacteria;p__Firmicutes;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Dorea;f__									
4	ASV2	0.059969	0.055273	0.132805	0.235792	0	0.032847	0.004196	0.010059	0.003390	0	0.012282	0.004140	0.049020	0.010309	5.557871	k__Bacteria;p__Firmicutes;c__Bacilli;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Catenibacteri									
5	ASV3	0.035598	0.019007	0.147505	0.094984	0.041072	0.026649	0.035570	0.037737	0.031513	0.014394	0.015256	0.033569	0.028456	0.024399	0.009920	k__Bacteria;p__Actinobacteriota;c__Coriobacteriia;o__Coriobacteriales;f__Coriobacteriaceae;g__Collins									

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

	A	B	C	D	E	F
1	#NAME	Donor.A.0h	Donor.A.CTR.24h	Donor.A.BC.24h	Donor.B.0h	Donor.B.CTR.24h
2	#CLASS	A	A	A	B	B
3	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Escherichia;Shigella	0.036072911575330005	25.674227449794806	11.285018808950424	0.2652528260846063	25.17148019483287
4	Bacteria;Actinobacteriota;Actinobacteria;Bifidobacteriales;Bifidobacteriaceae;Bifidobacterium	6.510356419235109	1.3717110864424356	0.7350221165626988	3.2100653934907845	1.7693041856414309
5	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Ligilactobacillus	0.0	0.3988932151526945	0.0656378435223486	0.06060995360490721	0.8779393723218315
6	Bacteria;Actinobacteriota;Coriobacteriales;Coriobacteriaceae;Collinsella	5.296360615244002	4.556504019002793	6.430318510923616	4.7524015005328735	7.08560725256285
7	Bacteria;Firmicutes;Bacilli;Lactobacillales;Enterococcaceae;Enterococcus	0.005052692551309561	0.15054656728175764	0.10668124712111847	0.0	0.17382906967920939
8	Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiaceae;Clostridium_sensu_stricto_1	0.0	0.0	0.0	0.0	0.009024250792407816
9	Bacteria;Firmicutes;Clostridia;Lachnospirales;Lachnospiraceae;Dorea;Dorea_fornicigenans	0.4197776971627988	5.641302475834611	11.9513138461352	0.5985016454365006	4.916670029548221
10	Bacteria;Bacteroidota;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides	0.0	0.0	0.0105880409550338	0.7693427097653991	4.671333567162463

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.

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/1HEbzgQUZYjxR7zugMDifL0f4737H6K5> 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.