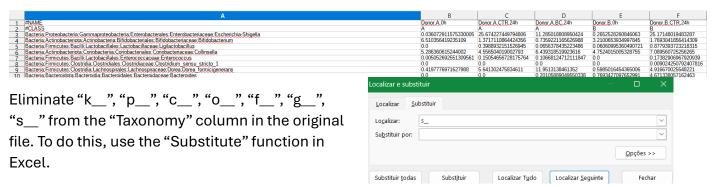# Pipeline – Phase 1: Selection of Microorganisms of Interest (Part 1 – Normalize 16S Copy Number & MicrobiomeAnalyst-compatible file)

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:

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

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_-GAFdOLlppHP2UWXFu3f1.

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 intro the folder of GoogleColab.
2. After uploading the data_file.csv, change its name in the script:

```python
# === 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.