

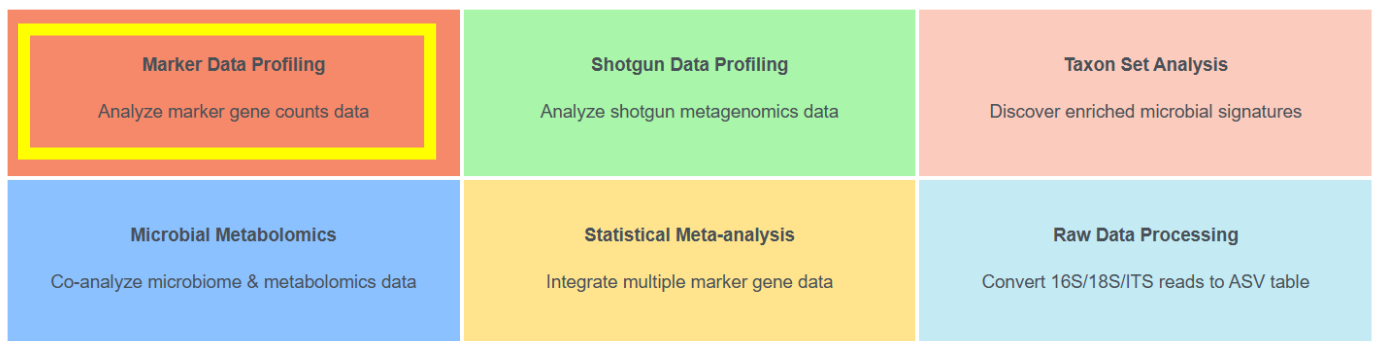
Pipeline – Phase 1: Selection of Microorganisms of Interest (Part 2 – MicrobiomeAnalyst & Selection of microorganisms)

Preparation of files

1. Metadata – to create the **metadata.csv** file, use the following template and fill it in:
<https://github.com/VaneBR/Project/blob/main/Files/metada.csv> or
https://drive.google.com/drive/u/0/folders/18yojrGbjKc_-GAFdOLppHP2UWXFu3f1.
2. OUT/ASV table – this file is the output file, **matrix_normalized_microbiomeanalyst.csv**, obtained from running the **normalize_16S.py** script.

Using MicrobiomeAnalyst

1. Open MicrobiomeAnalyst:
<https://www.microbiomeanalyst.ca/MicrobiomeAnalyst/ModuleView.xhtml> and select



2. Select the following options and upload both files:

The screenshot shows the upload form in MicrobiomeAnalyst. The following options are highlighted with yellow boxes:

- OTU/ASV table (.txt, .csv, or its zip)**: Taxonomy included (checked), Sequences included (unchecked), Normalized data (checked).
- Metadata file (.txt or .csv)**: + Choose button.
- Taxonomy table (.txt or .csv)**: + Choose button.
- (Optional) phylogenetic tree (.tre, .nwk)**: + Choose button.
- Taxonomy labels**: Not Specific / Other (selected).

3. Select and deselect the following options:

Data Integrity Check

Basic data filtering are performed by default, as downstream statistics (especially comparative analysis) may not perform properly due to the presence of singletons or constant values.

The screenshot shows the Data Integrity Check form. The following options are highlighted with yellow boxes:

- Default Filtering**: ? icon.
- Constant features**: unchecked checkbox.
- Singleton**: None (selected).
- One sample occurrence**: unchecked radio button.
- One total count**: unchecked radio button.
- Update**: button.

4. Put the “Minimum count” and “Percentage to remove (%)” to 0:

Low count filter

Minimum count: 0

☒ Prevalence in samples (%) 20

☐ Mean abundance value

☐ Median abundance value

Low variance filter

Percentage to remove (%): 0

☒ Inter-quantile range

Based on: ☐ Standard deviation

☐ Coefficient of variation

5. In the “Data Normalization” click on submit and then proceed.

6. Now you can visualize and make various statistical assays. For the purpose of this script, we are going to need 4 files from this analysis:
- The first one would be “Stacked bar/area plot”

Visual Exploration

[Stacked bar/area plot](#) [Interactive pie chart](#) [Rarefaction curve](#) [Phylogenetic tree](#) [Heat tree](#)

Data overview and general pattern discovery through intuitive visualization techniques

Change the following parameters, hit submit and then download the abundance table:

Data options

☐ Organize samples by SampleType then by None

☒ Merge samples to groups SampleType then by None

☐ View an individual sample V.A.OH

Taxa resolution

Taxonomy level: Genus prepend higher taxa ☐

☒ Merging small taxa with counts < 0 based on: Total

☐ Showing top n taxa, with n = 10

Graph options

Graph type: Stacked Bar [Percentage Abundance]

Color scheme: Set3

Open the file and substitute “.” with “,” and then create a “Mean Abundance (%)” column in the last column, where you calculate the mean between the different samples, such as:

	A	B	C	D	E
1		Baseline	Cellulose	Negative_control	Mean Abundance (%)
2	Acetanaerobacterium	3,33E-05	1,67E-05	0	=MÉDIA(B2:D2)

Finally, substitute “,” with “.” and save the file.

b. In “Core Microbiome” click on the following option:

Community Profiling

[Alpha diversity](#) [Beta diversity](#) [Core microbiome](#)

Quantitative analysis of community profiles using multiple well-established statistical methods

Then change the “Taxonomic level” to “Species”, hit submit and download the result table:

Core Microbiome Analysis

Taxonomic level

Species

Relative abundance (%)

Sample prevalence (%)

View type ☒ Heatmap ☐ Bar plot

Color contrast

Default

View options

☒ All samples together

☐ An experimental factor

SampleType

☐ A particular group

Experimental factor:

SampleType

 group:

Baseline

c. For “Single-factor analysis” click the following option:

Comparison & Classification

[Single-factor analysis](#) [Multi-factor analysis](#) [LEfSe](#) [Random Forest](#)

Identification of significant features or potential biomarkers via statistical and machine learning methods (supervised)

Change the “Taxonomy level” to “Species” and the “Statistical method” to “T-test/ANOVA”, hit submit and download the analysis results:

Taxonomy level

Species

Experimental factor

SampleType

Comparison:

Cellulose

 vs.

Baseline

Statistical method

T-test/ANOVA

Adjusted p-value cutoff

d. For “LefSe” click the following option:

Comparison & Classification			
Single-factor analysis	Multi-factor analysis	LefSe	Random Forest
Identification of significant features or potential biomarkers via statistical and machine learning methods (supervised)			

Change the “Taxonomy level” to “Species”, hit submit and download the analysis results

Taxonomy level	Species
Experimental factor	SampleType
P-value cutoff	0.1 <input type="radio"/> Original <input checked="" type="radio"/> FDR-adjusted
Log LDA score	2.0

Steps to use the script:

The script described below is available in the following link:

https://colab.research.google.com/drive/1Wk8Fq5X1huzqWal-HL_7IOpjKeROAXSg or in github:
https://github.com/VaneBR/Project/blob/main/Scripts/selection_bacterial_taxa.py.

1. Upload the files: **taxa_abund.csv**, **univar_test_output.csv**, **lefse_de_output.csv** and **core_microbiome.csv** in GoogleColab. Notice that you might need to change the thresholds values since this depends on your data:

```
# Define thresholds
lda_cutoff = 1
pval_cutoff = 0.05
abundance_cutoff = 0.001
core_cutoff = 0.1
```

2. Run the script and obtain two output files: **selected_taxa.csv** and **excluded_taxa.csv**
3. Evaluate your data according to your needs.