Bioinformatic Analysis of 16S rRNA

Demultiplexed sequence reads were processed using QIIME2 2022.8. The sequencing primers and the reads without primers were removed using Cutadapt tool. DADA2 was used to perform paired-end reads filtering, merging and chimeras removal steps after trimming low quality nucleotides from both forward and reverse reads (--p-trunc-len-f 271 and --p-trunc-len-r 177). Hence, ASVs (amplicon sequence variants) were generated and the taxonomic assignments was performed through the Scikit-learn multinomial naive Bayes classifier re-trained on SILVA database (release 138) V3-V4 iper-variable region. An additional check about possibly cross amplified host DNA has been made with Bowtie2 v.2.2.5. For further details about the FASTQ processing see the bash script publicly available at https://github.com/LeandroD94/Papers .

Statistical analyses on 16S rRNA data

The statistical analyses on bacterial communities were performed in R 4.2.1 with the help of the packages phyloseq 1.40.0, vegan 2.6-2, DESeq2 1.36.0, mixOmics 6.20.0 and other packages satisfying their dependencies. The packages ggplot2 3.3.6, ggvenn 0.1.9, dendextend 1.16.0, ggh4x 0.2.2 and ggpubr 0.40 were used to plot data and results. Every ASV associated to genera with average relative abundance under the cutoff of 0.005% have been discarded to minimize sequencing contaminants and improve statistical inferences [**PMID: 23202435** and **PMID: 33510727**]. A rarefaction analysis on ASV was performed on every sample using the function rarecurve (step 100 reads), further processed to highlight saturated samples (arbitrarily defined as saturated samples with a final slope in the rarefaction curve with an increment in ASV number per reads < 1e-5). The observed richness, Shannon and Pielou's evenness indices were used to estimate the bacterial alpha-diversity in each sample using the function estimate_richness from phyloseq. The evenness index was calculated using the formula E = S/log(R), where S is the Shannon diversity index and R is the observed ASV richness in the sample. A Venn Diagramm is used to represent the distribution of the core bacteriota among the groups. Differences in alpha-diversity indices were tested using Kruskal-

Wallis test and Dunnett test. PCoAs was performed using the Hellinger distance on Hellinger transformed ASV abundances. PERMANOVA and Betadisper were used to test the statistical significance of the beta-diversity distances and dispersions. At different taxonomic ranks, the differential analysis (DA) of the abundances has been computed with DESeq2 on raw count data. A minimal log2FC value of 1.5 has been chosen as arbitrary cutoff in DESeq2 algorithm to focus on the stronger effect sizes and therefore on the more robust results. Furthermore, DA taxon with a grand mean count < 150 has been discarded from the displayed results irrespective of their statistical significance to limit noisy results.

A sparse PLS-DA has been further computed using the MixOmics package with settings ensuring the best possible reproducibility and prediction power (details in the R script).

Moreover, potentially expressed Metacyc pathways in each group have been predicted through PICRUST2 v2.5 with SEPP algorithm and then significant differences of those have been explored using LEFSE 1.1.2 (LDA Effect Size) analysis.

For further details about the data analysis see the R script publicly available at https://github.com/LeandroD94/Papers .