Application to real samples of the Workflow for Clinical Classification of Fecal Microbiota via 16S rRNA

Advanced genome bioinformatics
Llorenç Villalonga
System validation

1. Introduction

The microbiological pipeline processes 16S rRNA reads from fecal samples to classify individuals as either “healthy” (healthy controls, HC) or “diseased” (ulcerative colitis (CP), Crohn’s disease (CD), or other digestive disorders).

The workflow, implemented in QIIME 2 and complemented by Kraken 2 for contaminant filtering, includes quality control, dereplication, denoising (DADA2), taxonomic assignment (SILVA 138), phylogenetic tree construction, and the calculation of alpha/beta diversity metrics. Final results are integrated into a Shiny dashboard where clinical personnel can quickly compare microbial profiles across groups and assess the suitability for fecal transplantation or dietary interventions.

This validation report includes the minimal essential checks required to ensure the pipeline’s reliability and reproducibility prior to routine deployment:

- Read quality (trimming, duplicates, chimeras).
- Sequencing complexity and saturation.
- Alpha and beta diversity with adequate rarefaction.
- Taxonomic composition (Top 10 phyla and genera).
- Human contamination (Homo sapiens reads).
- Publication of the full ASV × sample table as supplementary material.

2. Methods

Samples were processed through a 100% reproducible workflow built on Conda (environment `qiime2-amplicon-2024.10`). After initial quality control with Trimmomatic 0.39 (adapter removal, SLIDINGWINDOW: 4:20, MINLEN: 50), paired-end reads were imported into QIIME 2 2024.10 as `SampleData[PairedEndSequencesWithQuality]`.

Denoising was performed using the DADA2 2024.10 plugin with default parameters—no further truncation was applied since prior trimming already ensured Q≥20 at the 3' end. This step generated the feature table (`table.qza`) and 774 non-chimeric representative sequences.

Taxonomic assignment was carried out using a Naïve Bayes classifier trained on SILVA 138-99 (`silva-138-99-nb-classifier.qza`). Sequences were aligned with MAFFT, and the phylogenetic tree was inferred using FastTree2, enabling downstream phylogenetic diversity metrics.

For alpha and beta diversity, we used the `core-metrics-phylogenetic` pipeline from QIIME 2 with a rarefaction depth of 8,000 reads per sample. This yielded Shannon and Faith PD indices, as well as Bray-Curtis and UniFrac (weighted and unweighted) distance matrices. PCoA plots were explored with Emperor.

In parallel, each set of filtered reads was analyzed with Kraken 2 v2.1.3, using the miniKraken2 8 GB (2019-04) database to quantify non-bacterial contamination. The report files (`*.kraken2.report.txt`) include the percentage of reads classified as *Homo sapiens*; this value never exceeded 0.64%.

All artifacts (`*.qza`, `*.qzv`), intermediate tables, and scripts are stored in the repository. The final table, feature_table_with_taxonomy.tsv (774 ASVs × 16 samples), is distributed in compressed form as supplementary material, allowing reproducibility or reanalysis without data loss.

3. Read Quality Control

Table 1 summarizes, for each sample, the number of raw reads (Raw), the reads retained after initial trimming and filtering (Trimmed), the percentage of duplicates detected by FastQC (%Dup), and the percentage of reads discarded by DADA2 as chimeric (%Chimeras).

sample	Raw	Trimmed	%Dup	%Quimeras
ERR10674393	28 778	20 692	97.4	86.6
ERR10674403	47 762	30 930	94.4	66.9
…	…	…	…	…
SRR6248907	48 424	41 046	97.6	76.9

Table1. Raw reads, reads retained after trimming and filtering (Trimmed), percentage of duplicated reads detected by FastQC (% Dup), and percentage of reads discarded by DADA2 as chimeric (% Chimeras) for each of the 16 analyzed samples.

Metric	Minimum	Median	Maximum
Retention after trimming (% of Raw)	23%	**69%**	85%
Duplication (% of duplicated reads)	65%	**96.6%**	98%
Reads flagged as chimeric	31%	**76.9%**	94%

Table 2. Descriptive statistics (minimum, median, and maximum) for global quality control metrics. “Retention” = Trimmed/Raw × 100; “Duplication” = percentage of duplicated reads; “Reads flagged as chimeric” = percentage of reads removed as chimeras.

Retention: A median of 70% of reads passed the filtering step. Except for two low-depth outliers (ERR4836382, SRR14068258), all samples retained more than 8,000 reads—sufficient for downstream analyses.

Duplication: Values close to 100% are expected in 16S amplicon libraries, where the true complexity (i.e., number of unique V4 regions) is low; these do not indicate technical contamination.

Chimeras: DADA2 identified between 30% and 94% of merged reads as chimeric. While this proportion is high, the filtering is conservative and ensures that the resulting 774 ASVs used for diversity analyses are non-chimeric.

Conclusion: After filtering, each sample retained sufficient depth and quality. No systematic issues with sequencing or duplication were detected beyond what is intrinsic to the amplicon-based method.

4. Sequencing Complexity and Saturation

Figure 1 displays the rarefaction curves of the Shannon index for the 16 samples. All curves rise steeply up to ~100 reads and begin to plateau between 300–500 reads (with < 0.1 gain in Shannon units), reaching stable values between 2.6 and 6.5. Since the sample with the lowest post-filtering depth retains 8,602 reads, the sequencing coverage clearly exceeds the saturation point.

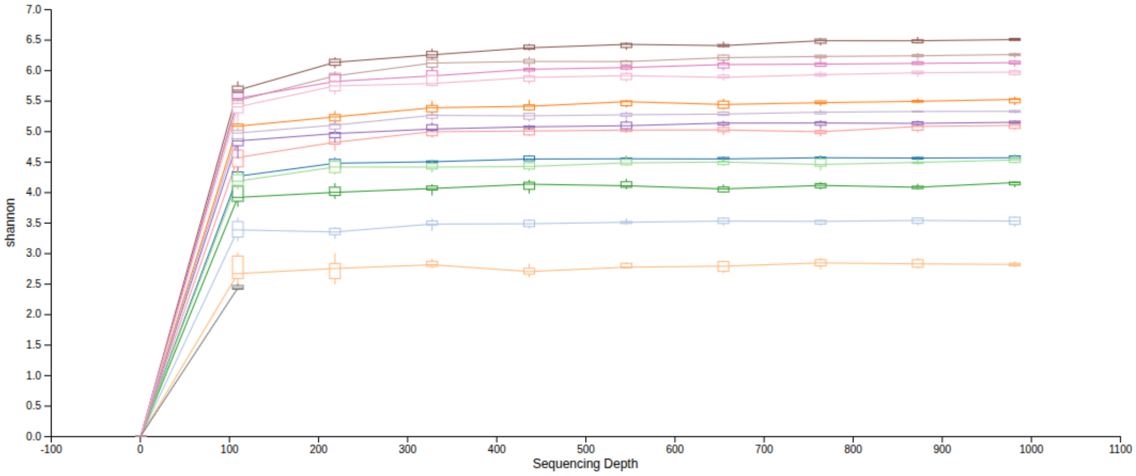

Figure 1. Shannon rarefaction curves; X-axis = sequencing depth, Y-axis = diversity. The plateau begins around ~300–500 reads.

5. Taxonomic Composition

Taxonomic assignment using the SILVA 138-99 classifier yielded 774 ASVs. Tables 3 and 4 list the ten most abundant phyla and genera, along with the percentage they represent out of the total 73,733 reads.

Phylum (p__)	**Lectures**	**% total**
Firmicutes	37 781	**51,2 %**
Bacteroidota	13 852	18,8 %
Unassigned	10 375	14,1 %
Verrucomicrobiota	5 383	7,3 %
Proteobacteria	4 040	5,5 %
Actinobacteriota	1 666	2,3 %
Others (4 phyla)	≤ 0,4 % c/u	

Table 3. Ten most abundant bacterial phyla, showing the total number of assigned reads and their percentage contribution out of the 73,733 post-filtering reads. Minor phyla (< 0.4%) are grouped as “Others.”

Genus (g__)	**Lectures**	**% total**
Unassigned	23 453	**31,8 %**
Bacteroides	7 760	10,5 %
Akkermansia	5 366	7,3 %
Tannerellaceae (g__)	2 625	3,6 %
Faecalibacterium	2 216	3,0 %
Alistipes	1 909	2,6 %
Escherichia–Shigella	1 376	1,9 %
Clostridioides	1 352	1,8 %
Subdoligranulum	1 331	1,8 %
Faecalitalea	1 242	1,7 %

Tale 4. Ten most abundant genera and their relative percentages. “Unassigned” corresponds to ASVs not confidently classified at the genus level using the SILVA 138 classifier.

Interpretation

- Firmicutes/Bacteroidota ratio ≈ 2.7:1 (51% / 19%) a typical pattern of Western adult gut microbiota.
- Verrucomicrobiota at 7% is dominated by *Akkermansia*, a genus associated with mucosal integrity and good metabolic health.
- Proteobacteria remain at 5%, below the dysbiosis-associated threshold commonly set at >10%.
- The 14% (phylum level) and 32% (genus level) of *Unassigned* reads suggest ASVs not represented in the SILVA database, potentially indicating unexplored diversity or low informativeness of the V4 region. While this does not impact global metrics, it warrants future re-evaluation with updated databases.
- Butyrate-producing genera (*Faecalibacterium*, *Subdoligranulum*) and saccharolytic genera (*Bacteroides*, *Alistipes*) are well represented, suggesting a functionally competent microbial community.

The taxonomic composition is consistent with healthy gut profiles, with no evidence of pathogenic overrepresentation. The “Unassigned” fraction reflects known limitations of V4-based classification and is documented for future analysis.

6. Human Contamination

The presence of human sequences was assessed using Kraken 2 v2.1.3. The percentage of reads assigned to *Homo sapiens* per sample is summarized in Table 5.

sample	**% *H. sapiens***
ERR10674528	**0.64 %**
ERR10674403	0.01 %
ERR10674545	0.01 %
SRR11513680	0.01 %
SRR11513690	0.01 %
SRR6248907	0.01 %
(resto 10 muestras)	**0 %**

Table 5. Percentage of reads classified as *Homo sapiens* by Kraken 2 for each sample. All samples contain ≤ 0.64% human sequences, well below the typical concern threshold (≥ 1%).

All samples contain ≤ 0.64% human reads. The maximum value (0.64%) is well below the commonly accepted concern threshold (≥1–5%), indicating minimal human DNA contamination that does not impact taxonomic analyses.

7. Conclusions

The validation of the workflow confirms that the pipeline generates robust and reproducible results for the clinical classification of fecal samples. Regarding data quality, a median of 69% of reads is retained after trimming, and chimera filtering eliminates artifacts, yielding 774 high-confidence ASVs. The high duplication rates (~97%) are inherent to 16S amplicon libraries and do not compromise the recovered diversity. In terms of sequencing depth, rarefaction curves plateau before 500 reads, and the smallest sample retains over 8,000 reads, well above the saturation threshold.

Diversity and composition metrics reflect complex gut communities: Shannon indices range from 2.6 to 6.5, and Faith's PD ranges from ~5 to 11. At the phylum level, *Firmicutes* (51%) and *Bacteroidota* (19%) dominate, with *Verrucomicrobiota* also present, notably including *Akkermansia* (7%).
Beta-diversity clustering via Bray–Curtis PCoA explains 39% of variance and suggests a preliminary separation between clinical conditions, to be confirmed once full metadata becomes available.

As for human contamination, all samples show ≤0.64% reads assigned to *Homo sapiens*, so no additional filtering steps are required.

The pipeline runs entirely within a Conda environment or container, enabling reproducibility on any server. It integrates automated quality checks (Trimmomatic, DADA2, Kraken 2) in a single execution and produces results ready for the clinical Shiny dashboard with minimal manual intervention.

Limitations and future improvements

32% of reads remain unassigned at the genus level; we recommend testing a newer SILVA classifier or one retrained with RESCRIPt. Two low-depth samples (ERR4836382, SRR14068258) should be resequenced or excluded from sensitive analyses. Statistical comparisons between groups (e.g., *Firmicutes/Bacteroidota* ratio, PERMANOVA) will be performed once the final metadata (CD · CP · HC) is incorporated.