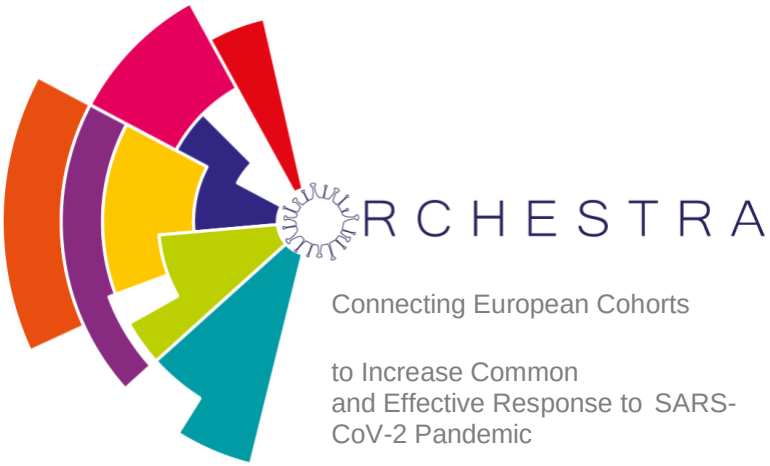


WPN°6 Deliverable N°6.10

Dataset of 16S rRNA sequencing of retrospective cohorts

University of Bologna (UNIBO)

Universiteit Antwerpen (UANTWERPEN)

Project Classification

Project Acronym:	ORCHESTRA
Project Title:	Connecting European Cohorts to Increase Common and Effective Response to SARS- CoV-2 Pandemic
Coordinator:	UNIVR
Grant Agreement Number:	101016167
Funding Scheme:	Horizon 2020
Start:	1st December 2020
Duration:	36 months
Website:	www.orchestra-cohort.eu
Email:	info@orchestra.eu

Document Classification

WP No:	WP6
Deliverable No:	D6.10
Title:	Dataset of 16S rRNA sequencing of retrospective cohorts
Lead Beneficiary:	UNIBO
Other Involved Beneficiaries:	COVID-HOME
Nature:	Report
Dissemination Level:	Public
Due Delivery Date:	31/05/2022
Submission Date:	30/06/2022
Justification of delay:	Inclusion of more cohorts data in the analysis
Status:	Completed
Version:	1.0
Original author(s):	Marco Fabbrini, Patrizia Brigidi

History of Changes

Version	Date	Created/Modified by
0.1	28/06/2022	Patrizia Brigidi, Marco Fabbrini, Matilda Berkell, Surbhi Malhotra-Kumar
0.2		
0.3		
0.4		
0.4		
0.5		

Executive summary

WP and deliverable context

The present report is part of ORCHESTRA project, a three-year international research project aimed at tackling the coronavirus pandemic. ORCHESTRA provides an innovative approach to learn from the pandemic SARS-CoV-2 crisis, derive recommendations to further management of COVID-19 and be prepared for the possible future pandemic waves. The ORCHESTRA project aims at delivering sound scientific evidence for the prevention and treatment of the infections caused by SARS-CoV-2 assessing epidemiological, clinical, microbiological, and genotypic aspects of population, environment, and socio-economic features. The project builds upon existing, and new large scale population cohorts in Europe (France, Germany, Spain, Italy, Belgium, Romania, Netherlands, Luxemburg, and Slovakia) and non-European countries (India, Perú, Ecuador, Colombia, Venezuela, Argentina, Brazil, Democratic Republic of Congo, and Gabon) including SARS-CoV-2 infected and non-infected individuals of all ages and conditions. The primary aim of ORCHESTRA is the creation of a new pan-European cohort applying homogenous protocols for data collection, data sharing, sampling, and follow-up, which can rapidly advance the knowledge on the control and management of the COVID-19. Within ORCHESTRA project, the Work Package 6 (WP6) aims at providing innovative laboratory capabilities combining serology, immunology, viral and human genomes, microbiota and epigenetic analysis. It aims to describe markers and physiopathology of various COVID-19 outcomes including severe cases, long COVID and vaccine efficiency across various patient populations gathered within ORCHESTRA cohorts.

The objectives of WP6 are distributed in two parts: (1) a retrospective part on frozen samples obtained during 2020 and (2) a prospective part starting in 2021. The geal for the analyses of both gut and respiratory microbiota dynamics

Content of the document

The present report describes the data on retrospective cohorts with available biobanking to characterization microbiota dynamics in the gut of patients with a SARS-CoV-2 infection. After mapping all the partners' availability, three cohorts satisfied the inclusion criteria: PREDI-CO, COVID-HOME, UNIVR COVID cohort. The objectives differ slightly across the different cohorts, therefore, are described separately for each analysis.

Dissemination level: Public

Intestinal microbiota dynamics in patients with COVID-19

Cohorts selected for the analysis of human gut microbiota dynamics in patients with COVID-19 in retrospective cohorts within the ORCHESTRA project include:

- PREDI-CO
- COVID-HOME
- UNIVR COVID cohort

The objectives, status of the analyses, and (preliminary) results will be reported separately for each cohort.

1. PREDI-CO

Methods

Sample collection

The cohort included **faecal samples from 69 COVID-19 patients** from three different hospitals of Bologna (Italy) **enrolled during the 1st wave**, that were later included in the ORCHESTRA framework. All patients were hospitalized and tested positive for SARS-CoV2 infection by means of RT-PCR targeting regions in the N gene following the US CDC protocol. A faecal sample was collected at the time of infectious disease consultation in a sterile plastic container and kept at -80°C until further processing. Of the 69 patients, 23 of them developed severe respiratory failure, 16 were admitted to ICU and 10 were mechanically ventilated.

Sample processing

Starting from 250mg of faecal sample, microbial DNA was extracted using the repeated bead-beating plus column method: 1 mL of lysis buffer (500 mM NaCl, 50 mM Tris-HCl, pH 8, 50 mM EDTA, and 4% SDS), four 3 mm glass beads, and 0.5 g of 0.1 mm zirconia beads (BioSpec Products, Bartlesville, OK, USA) were used to perform chemical and mechanical lysis of the samples in a FastPrep instrument (MP Biomedicals, Irvine, CA, USA) at 5.5 movements/s for 1 min, repeated three times. Samples were incubated at 95 ◦C for 15 min and then centrifuged at 13,000 rpm for 5 min. Subsequently, the supernatant was added with 10 M ammonium acetate and centrifuged for 10 min at 13,000 rpm. The supernatants were then incubated in ice for 30 min with one volume of isopropanol for nucleic acid precipitation. A washing step with 70% ethanol was performed, and the precipitated nucleic acids were resuspended in 100 µL of TE buffer (10 mM Tris-HCl, 1 mM EDTA pH 8.0). Two microliters of 10 mg/mL DNase-free RNase were then added, and the samples were incubated at 37 °C for 15 min. Finally, a column-based method was used for DNA purification using the DNeasy Blood and Tissue Kit (QIAGEN, Hilden, Germany) as per manufacturer's instructions. The yield and the quality of the extracted DNA were assessed with a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). For library preparation, the V3-V4 hypervariable region of the 16S rRNA gene was amplified by using the 341F and 785R primers. The final libraries, indexed and purified, were sequenced on an Illumina MiSeq platform, with a 2 × 250 bp paired-end protocol according to the manufacturer's instructions (Illumina).

Data analysis

Data analysis for these samples has been completed. Raw sequences were processed using a combined pipeline of PANDASeq and QIIME 2. Length and quality-filtered reads were binned into amplicon sequence variants (ASVs) using DADA2 while removing chimeras. Taxonomic assignment was performed using VSEARCH against the Greengenes database (May 2013 release). Publicly available 16S rRNA gene sequences of 69 healthy subjects matched by age, sex, and geography (across Italy) were downloaded from databases and processed as above [33 subjects: NCBI SRA, Bioproject ID SRP042234; 36 subjects: MG-RAST ID mgp17761]. GM sequences of 16 patients

admitted to ICU at St. Orsola Hospital after undergoing liver transplantation in October 2019-February 2020, were also used for comparative purposes (NCBI SRA: PRJNA700830). Alpha diversity (Inverse Simpson Index), beta diversity (Bray-Curtis dissimilarity) and bacterial ecosystem composition were investigated with statistical analyses in R (Wilcoxon test, PERMANOVA, linear discriminant analysis effect size estimation).

Results

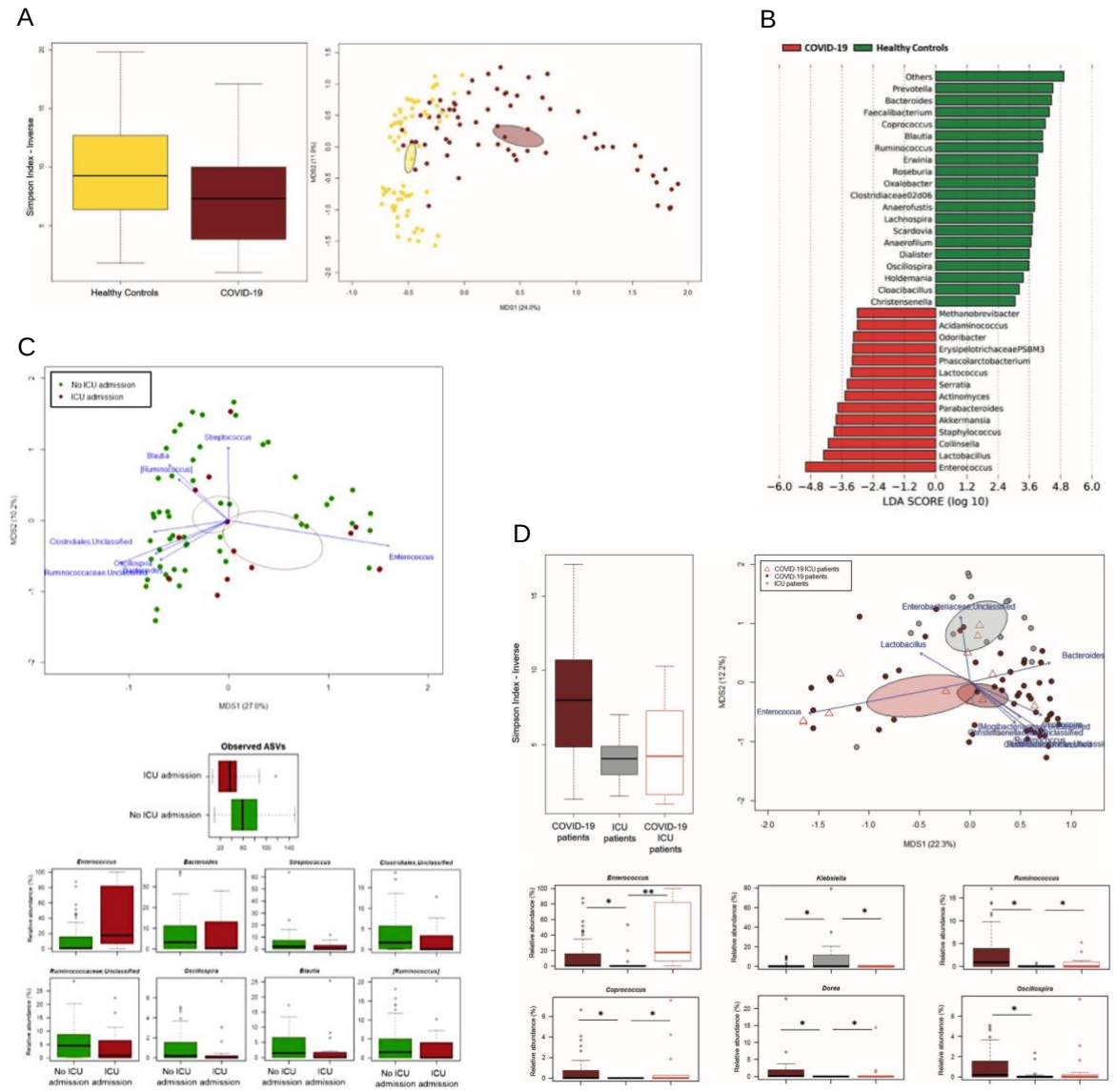


Figure 1: (A) The gut microbiome of healthy controls differs significantly from the one of COVID-19 patients in terms of Inverse Simpson Index alpha diversity and beta diversity (Bray-Curtis dissimilarity index). **(B)** Among the identified genera, *Enterococcus* showed the strongest association with COVID-19 according to linear discriminant analysis effect size (LEfSe). **(C)** COVID-19 patients admitted to ICU showed a different microbial profile compared to the other non-ICU SARS-CoV2 patients enrolled. Among others *Enterococcus* showed to pull the microbial configuration in such dysbiotic direction. **(D)** Compared to non-COVID-19 ICU patients, SARS-CoV2 positive patients - and especially ICU-admitted ones - showed higher abundances of *Enterococcus*, instead of the often-found *Klebsiella*. *, $p < 0.05$; **, $p < 0.01$

The gut microbiota of COVID-19 patients appears **severely dysbiotic, with distinct signatures compared to healthy subjects** (**Figure 1A**). In addition to a loss of diversity, COVID-19 patients show profound GM destruction, with drastic **reduction in the relative abundance of the dominant families** *Bacteroidaceae*, *Lachnospiraceae* and *Ruminococcaceae*, well known to be associated with health and to produce SCFAs, *i.e.*, microbial metabolites with a key, multifaceted role in human metabolic and immunological homeostasis. On the other hand, we found **increased proportions of potential pathobionts, mostly belonging** to *Enterococcaceae*, *Coriobacteriaceae* and *Staphylococcaceae*. Although some of the aforementioned microbial traits are common to other disorders, the remarkable enrichment of Enterococcus seems to represent a distinctive GM footprint of PREDI-CO cohort. In some patients, GM was even almost mono-dominated by *Enterococcus* spp. (**Figure 1B**). It is known that a high abundance of *Enterococcus* in the GM of critically ill patients may be clinically relevant given its pathogenic potential, intrinsic resistance to many commonly used antimicrobials, and the ability to rapidly acquire resistance determinants against virtually all antibiotics.

The severity of COVID-19-related dysbiosis was found to be strongly associated with ICU admission (**Figure 1C**). In particular, the GM of ICU COVID-19 patients was even less diverse and showed a further increase in *Enterococcus* along with a reduction in *Ruminococcaceae* and *Lachnospiraceae* taxa. Furthermore, patients admitted to ICU showed a depletion of *Bacteroides*. In an attempt to further explore the impact of ICU stay, we compared the GM of COVID-19 patients with that of patients admitted to the ICU just before the COVID-19 outbreak (**Figure 1D**). According to our findings, *Enterococcus* was far overrepresented in the GM of COVID-19 patients, especially those admitted to ICU, while almost absent in critically ill non-COVID-19 patients. Conversely, the latter were discriminated by higher proportions of *Enterobacteriaceae* members, especially *Klebsiella*. Although we are aware that we cannot claim that the high proportions of enterococci are specific to COVID-19, these data suggest that *Enterococcus* and the deriving bloodstream infection are somehow related to SARS-CoV- 2 infection.

2. COVID-HOME

Methods

Sample collection

The cohort included **833 faecal samples from 223** individuals, including non-hospitalized SARS-CoV2 RT-PCR positively tested patients together with relative household members, enrolled in the University Medical Center of Groningen (Netherlands) during the **1st, 2nd and 3rd waves of SARS-CoV2 pandemics**. Enrolled patients are visited at home soon after initial diagnosis, and then weekly on days 7, 14 and 21 after confirmed infection to obtain clinical data and biological samples (including faeces, blood, nasopharyngeal swab). NP swabs and faeces were tested for SARS-CoV-2 by RT-PCR. If still positive on Day 21 for any of the specimens, the participant is invited to continue weekly sampling for RT-PCR testing for two further weeks.

Sample processing

The experimental approach used was the same illustrated for the PREDI-CO cohort. All the samples underwent DNA extraction and 16S rRNA gene sequencing.

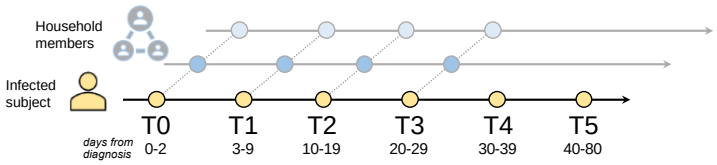
Data analysis

Data analysis for these samples has been completed. Raw sequences were processed using a combined pipeline of PANDASeq and QIIME 2. Length and quality-filtered reads were binned into amplicon sequence variants (ASVs) using DADA2 while removing chimeras. Taxonomic assignment was performed using VSEARCH against the Greengenes database (May 2013 release). Alpha diversity (Shannon's entropy), beta diversity (Jaccard distances) and bacterial ecosystem composition were

investigated with statistical analyses in R. Statistical tests including Kruskal-Wallis and Wilcoxon tests with p values correction using FDR were applied when necessary.

Results

Patients where stratified according to their SARS-CoV2 nose PCR testing (POS and NEG group corresponding to testing results) and the time in which sampling occurred, as detailed below:



We detected **significant differences** (pairwise Wilcoxon rank-sum test, FDR corrected $p < 0.05$ in group 'a' vs group 'b' - **Figure 2A**) in terms of Shannon's **alpha diversity** when comparing negative to positive subjects, with SARS-CoV2 negative ones showing a richer gut ecosystem. At the same time, we detected **significant differences evaluating beta diversity** using Jaccard distances (**Figure 2B**) comparing positive and negative individuals (PERMANOVA $p < 0.05$) from T0 to T3. Such differences can be detected also in relative abundances of the *Peptostreptococcaceae* family and in the genera *Faecalibacterium, Blautia* and *Bifidobacterium* (pairwise Wilcoxon, group 'a' significant vs group 'b' - **Figure 2C**). In particular, during the first weeks of positivity, infected individuals showed lower levels of *Faecalibacterium* and *Bifidobacterium* – two well-known health-promoting commensals – as well as a reduction of *Blautia* in week 2 and 3 compared to the day of diagnosis.

Samples were then clustered according to their genera level similarities using **hierarchical clustering**, validating the approach through the evaluation of clustering silhouettes **(Figure 2D**). We obtained **5 different clusters** (**Figure 2E**) homogenously populated by samples. We implemented the Random Forest-based feature selection algorithm Boruta to detect the peculiar features driving the clustering and we evaluated the relative abundance of the highlighted feature across the clusters (differences were validated with pairwise Wilcoxon test). We found out that cluster #1 was characterized by the genera *Collinsella* and *Roseburia*, whilst cluster #2 appeared to be enriched in *Bacteroides, Akkermansia, Dialister, Gemmiger, Faecalibacterium, Blautia* and *Ruminococcus*. Cluster #3 showed high level of *Collinsella* and *Roseburia* as well, together with *Coprococcus* and *Bifidobacterium*. Cluster #4 showed no particular genera enrichment, but rather average relative abundances values for *Bacteroides*, *Gemmiger* and *Faecalibacterium*, together with a significant lack of *Roseburia* and *Coprococcus*. Interestingly, this cluster appeared to be the most heterogeneous in terms of *Faecalibacterium* relative abundances, with some individuals carrying none of this genus, and others scoring the highest values of the entire cohort. Finally, cluster #5 was characterized by high relative abundances of *Bacteroides, Gemmiger, Dorea, Faecalibacterium, Ruminococcus* and *Blautia* as some other clusters, but specifically by *Pseudobityrivibrio*.

The distribution of positive and negative samples among clusters appeared to be homogeneous, however **50% of the individuals exhibiting viral shedding**, with positive stool and negative nose swabs PCR tests, populated **clusters 4** and especially **5** at their last timepoints, possibly suggesting a link between the cluster-enriched features and this phenomenon.

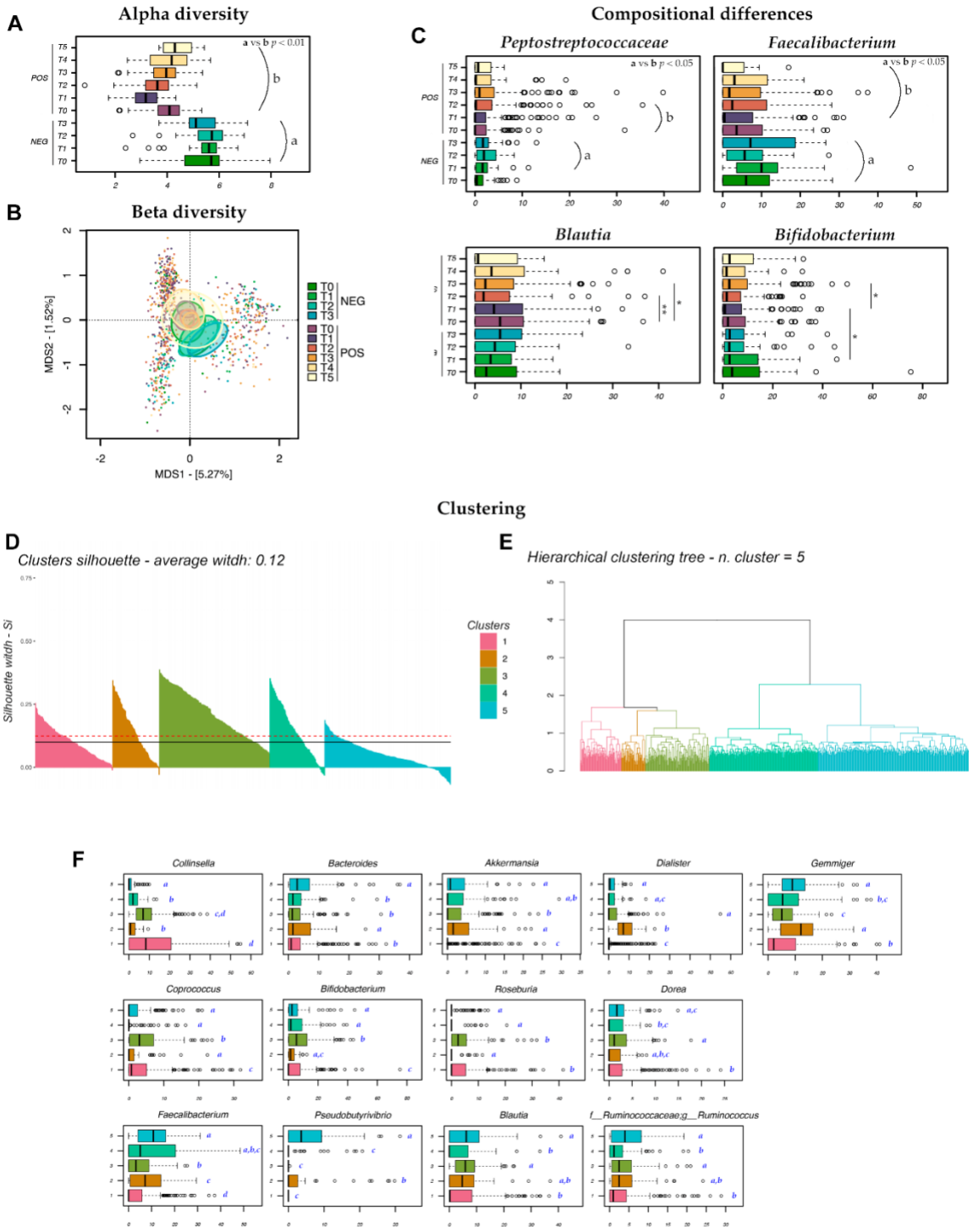


Figure 2: (A) Significant differences in term of microbial richness within an individual (alpha diversity) were detected according to longitudinal sampling between positively and negatively tested subjects. **(B)** Jaccard distances PCoA revealed that significant differences were also detected concerning whole microbial composition. **(C)** Taxonomic differences between groups through timepoints were detected for the family *Peptostreptococcaceae* and the genera *Faecalibacterium, Blautia* and *Bifidobacterium*. **(D)** All samples were clustered according to their microbial composition at the genera level using hierarchical clustering, obtaining 5 different clusters, which were validated according to their silhouette profiles. **(E)** Hierarchical clustering tree of all the samples across the 5 clusters. **(F)** Implementing a Random Forest-based feature classifier algorithm we detected peculiar microbial features descriptive of some clusters.
Different letters report significant ($p < 0.05$) pairwise Wilcoxon tests.

3. UNIVR COVID cohort

Methods

Sample collection

The cohort included **134 rectal swabs** sampled from SARS-CoV2 **infected individuals** (tested by RT-PCR) at their **first admission to the hospital during the 1st wave** of SARS-CoV2 pandemics.

Sample processing

The experimental approach used was the same illustrated for the PREDI-CO cohort. All the samples underwent DNA extraction and 16S rRNA gene sequencing.

Data Analysis

Data analyses for these samples has been completed. Raw sequences were processed as previously reported for the COVID-HOME cohort, obtaining taxonomic assignments. Statistical tests including Kruskal-Wallis and Wilcoxon tests with p values correction using FDR were applied when necessary.

Results

Samples have been stratified depending on their genera level compositional differences of the microbiota, using a clustering approach. K-means clustering technique was implemented, first evaluating the ideal number of clusters in order to optimize clusters' average silhouette values. The highest average value of clusters' silhouette (avg.sil=0.2) was achieved with **four clusters** (**Figure 3A**) and the sample's distribution over clusters was homogeneous (**Figure 3B**) and confirmed by a separate hierarchical clustering approach (**Figure 3C**).

We then implemented the feature-selection algorithm Boruta, to detect each cluster's specific taxonomic signatures. Relative abundances of such features were then compared with pairwise Wilcoxon tests followed by *p*-values FDR correction (**Figure 3D**). Cluster #1 appeared to be enriched in *Corynebacterium* and *Finegoldia*, whilst cluster #2 showed higher relative abundances of *Prevotella*. Both cluster #1 and #2 were reported to have increased levels of *Anaerococcus* and *Peptoniphilus*. Cluster #3 was characterized solely by the **pathogenic** genus *Shigella*. Finally, cluster #4 was defined by higher relative abundances of *Eggerthella, Parabacteroides, Ruminococcus, Bifidobacterium, Faecalibacterium, Bacteroides* and *Blautia*.

These results suggest that **cluster's signature can discriminate patients' health status**, due to the high divergence in some well-known health promoting genera (as *Ruminococcus, Faecalibacterium* and *Bifidobacterium*) between cluster #4 and cluster #3, the latter showing frightening levels of the pathogenic genus *Shigella*. Correlations between the microbiota composition and disease severity and general health status will be performed once obtained the patient's metadata and will be included in the following deliverable.

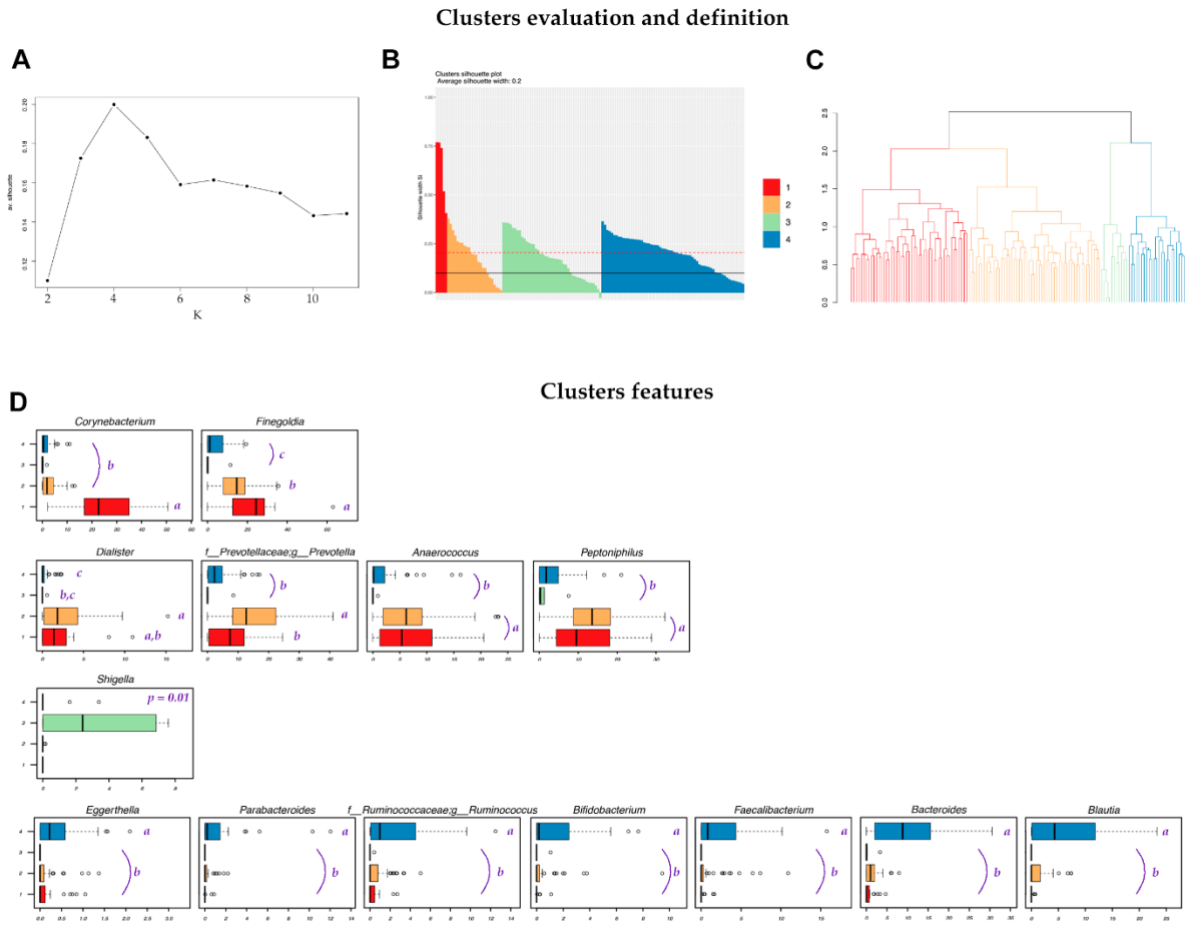


Figure 3: (A) Average silhouette values plotted against several *k* (*i.e.*, the number of clusters used for the K-means clustering) tested. K=4 was scored as the best number of clusters to obtain the highest average silhouette value, thus optimizing samples' distribution across clusters. **(B)** Clustering silhouette values. Each bar represents a sample and their distribution across clusters appears homogeneous. **(C)** Hierarchical clustering tree obtained independently from the K-means clustering, used to validate the clustering approach, confirming the 4 clusters division. **(D)** Implementing a Random Forest-based feature classifier algorithm we detected peculiar microbial features descriptive of the four clusters.
Different letters report significant ($p < 0.05$) pairwise Wilcoxon tests.

4. Comparison of the three cohorts

Methods and Data Analysis

We considered all samples from PREDI-CO and UniVR cohorts, together with all T0 samples (*n*=187) from the COVID-HOME cohort, in order to compare the three cohorts at the same time, *i.e.,* **within 48h from the time of SARS-CoV2 infection diagnosis** through nose swab PCR testing. Thus, a **total of 390 samples** was included in this analysis.

Results

First, we compared the microbial composition at the genus level of the different subjects enrolled in the three cohorts, **stratifying patients according to severity** (*i.e.,* healthy from the COVID-HOME, mild from the COVID-HOME, severe from UniVR, severe from PREDI-CO and critical from PREDI-CO). We detected a **significant separation** (pairwise PERMANOVA $p < 0.05$, **Figure 4A**) between the three cohorts. In particular, significant differences were reported between the healthy and mild participants enrolled in the COVID-HOME, the severe hospitalized patients from UniVR and the severe and critical patients from PREDI-CO. When fitting the compositional variables on the PCoA plot, we detected that the genera *Gemmiger* and *Oscillospira* were significantly associated to healthy and mild groups, *Finegoldia, Peptoniphilus, Prevotella* and *Anaerococcus* with the severe UniVR participants and *Enterococcus* with the PREDI-CO severe and critical patients. These results are in line with the cohort-specific features that have been identified during the single cohort analyses, thus enforcing results reliability. When observing patients' heterogeneity within each group, by means of plotting the distribution of their beta diversity values (**Figure 4B**) we observed that the higher was the severity, the higher was the heterogeneity of the subjects' microbiota, suggesting that the microbial ecosystem is likely to be dysbiotic, thus resulting in a less "healthy like" profile shared across all the individual, but rather in a multifaceted disorganized configuration.

Interestingly, when evaluating alpha diversity (**Figure 4C**) we observed a progressive significant reduction of microbial richness as the severity increases. Together with the evidence reported concerning patients' heterogeneity this suggests that in severe and especially critical patients, each one showed a markedly different microbial ecosystem, characterized by a reduced complexity and individualistic features.

When comparing the taxonomic composition of the subjects enrolled in the cohorts, we highlighted higher relative abundances of the genera *Finegoldia, Anaerococcus, Corynebacterium, Porphyromonas* and *Campylobacter* for the hospitalized patients from the UniVR cohort. The previously reported *Enterococcus* feature for the PREDI-CO cohort – where many patients showed an almost-complete colonization from this genus – was confirmed to be specifically associated to this cohort, and especially to critical ICU patients. Finally, the infected subjects from COVID-HOME with mild symptoms and their negative relatives were discriminated by *Gemmiger, Roseburia, Collinsella, Ruminococcus* and *Oscillospira*.

There results highlight the tight relationship between the gut microbiota and COVID-19 disease severity, providing encouragement towards further analyses, envisaging sound scientific publications in the upcoming months. Further studies need to be carried out in order to derive strong clinical recommendations, but nevertheless the results obtained so far are of great interest to disentangle the relationship between SARS-CoV2, the gut microbiota and COVID-19 disease progression. This 16 rRNA analysis allows to identify for each cluster of the three investigated cohorts, the most representative samples to be further analyzed by shotgun metagenomics approaches.

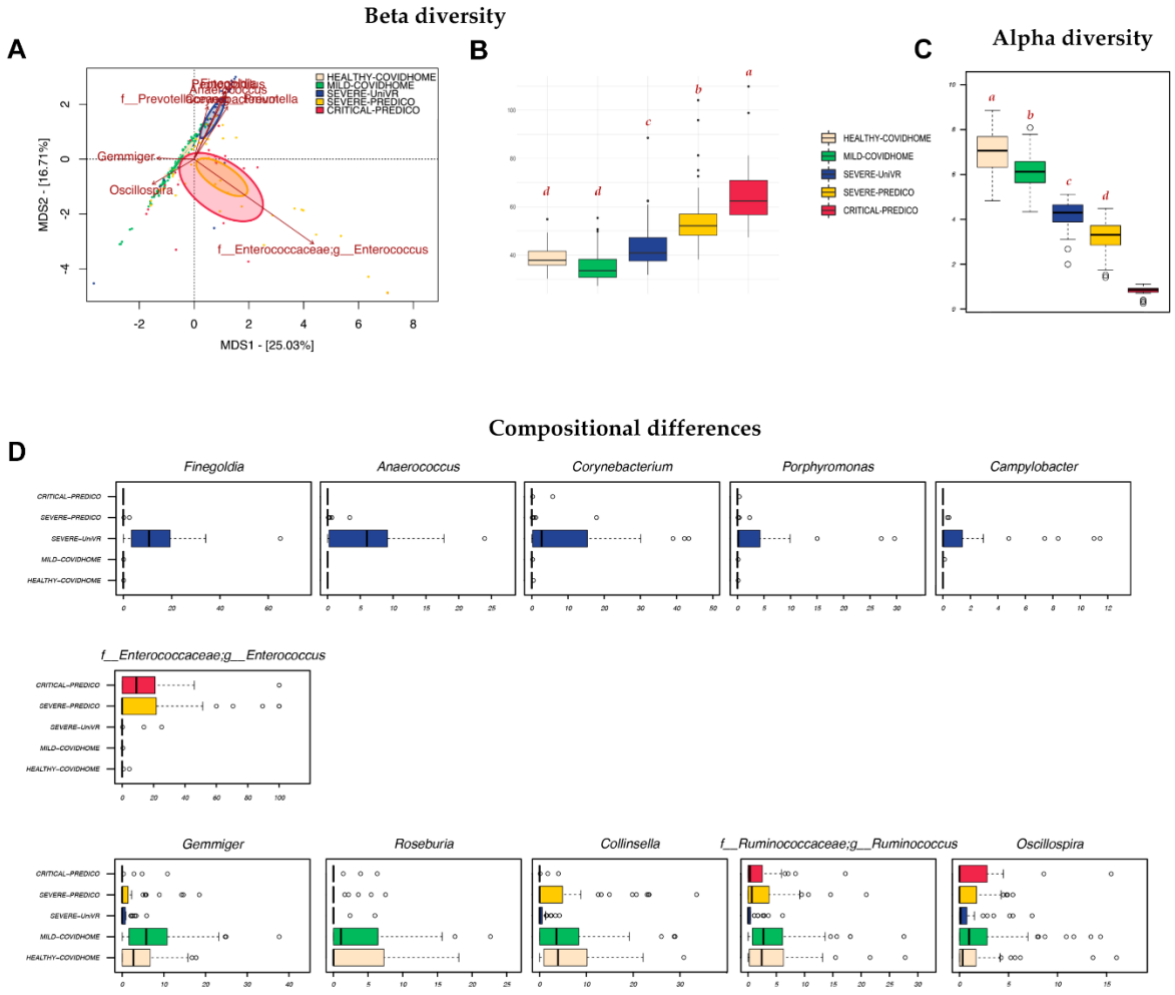


Figure 4: (A) Principal Coordinate Analysis (PCoA) based on Euclidean distances between the genus-level profile of the subjects enrolled in the three cohorts. A significant separation between groups was found with pairwise permutational analysis of variance with pseudo-F ratio. Ellipses include 99% confidence area based on the standard error of the weighted average sample coordinates. **(B)** Within-group sample's heterogeneity derived from beta diversity Euclidean distances. The higher the values, the higher the heterogeneity of the microbial configuration of subjects belonging to the same group. **(C)** Alpha diversity according to Shannon's coefficient. Different letters point to pairwise Wilcoxon test significant *p*-values (FDR-corrected $p<0.05$). **(D)** Compositional differences at the genus level detected between the groups, according to significant ($p<0.05$) Kruskal-Wallis test, followed by pairwise Wilcoxon tests (FDR-corrected Wilcoxon *p*-values < 0.05). In the first line the reported taxa were significantly higher in relative abundance compared to all other groups. In the second line *Enterococcus* showed significant pairwise Wilcoxon *p*-values against UniVR and COVID-HOME cohorts. The bottom line shows the distribution of relative abundance values significantly higher according to pairwise Wilcoxon tests in the COVID-HOME cohort compared to UniVR and PREDI-CO cohorts.