Màster Universitari
Anàlisi de Dades Òmiques /
Omics Data Analysis
FACULTAT DE CIÈNCIES I TECNOLOGIA
UVIC | UVIC·UCC

Master of Science in Omics Data Analysis

Master Thesis

BIOINFORMATICS TOOLS FOR ENVIRONMENTAL DIAGNOSTICS

by

ALBA BURILLO NAVARRO

Supervisor: Mª Adela Yáñez, Innovation manager, LABAQUA.

Academic tutor: Mireia Olivella, Structural Bioinformatics Researcher and Associate Professor, UVIC-UCC.

Biosciences Department

University of Vic – Central University of Catalonia

September 2021

ACKNOWLEDGEMENTS

My most sincere thanks to the company LABAQUA S.A. for allowing me to carry out this work in their facilities. I would also like to thank the company ielab, especially Estibaliz, Mariana, Cesar and Laura for welcoming me, accompanying me at every breakfast and helping me throughout the project. I must include Riccardo, for sharing with me his knowledge of bioinformatics and for always being willing to correct my code errors. But above all, I want to thank Adela, my scientific director, for trusting and always counting on me, for guiding me through the work, and for finding time for me, even when she didn't have it.

Not forgetting my family and friends, for their support, for having the patience to bear with me and for helping me clear my head when necessary.

ABSTRACT

Traditionally, the detection and identification of microorganisms in water samples is done by culture-dependent methods, but these do not allow the detection of all microorganisms. High-throughput sequencing now makes possible to determine the whole microbiome and the viral diversity of water samples.

A 16S ribosomal RNA subunit gene-based method was used to study the bacterial diversity of 24 reclaimed water samples. The phyla Patescibacteria, Bacteroidetes, Proteobacteria, Epsilonbacteraeota or Campylobacteria and Verrucomicrobia dominated the microbiomes of the studied samples. The results provide insight into the complex microbial community of reclaimed water, highlighting the potential environmental and health risks associated with its use.

A set of ten sewage samples was analysed for low-frequency variants within the SARS-CoV-2 genome. The method used was able to distinguish multiple SNPs belonging to different SARS-Cov-2 variants within the same wastewater sample. In addition, several mutations of the Gamma variant (P.1, Brazil) were detected before it was considered a Variant of Concern (VOC). The use of qRT-PCR in conjunction with wastewater sequencing methods has shown the potential to quantify the amount of SARS-Cov-2 present in a sample, and to identify specific mutations or single nucleotide polymorphisms (SNPs) to track SARS-Cov-2 variants.

Thanks to the use of bioinformatics approaches such as those used in this study, it has been shown that the analysis of sequences presents in water samples (reclaimed, wastewater...) should play a key role in environmental decision-making and, increasingly, in the field of public health. Stressing the importance of including these new technologies in the future regulations to improve the quality control of water.

INDEX

LIST OF TABLES

INTRUDUCTION

RESULT AND DISCUSSION

APPENDIX

LIST OF FIGURES

INTRODUCTION

METHODS

RESULTS AND DISCUSSION

GLOSSARY

Polymerase Chain Reaction (PCR): Molecular biology technique consisting of the in vitro amplification of multiple copies of a fragment of specific genetic material (RNA or DNA).

Next Generation Sequencing (NGS): A massively parallel sequencing technology that offers ultra-high throughput, scalability, and speed. The technology is used to sequence whole genomes, sequence target regions of DNA or RNA…

Waste Water Treatment Plants (WWTP): Is a water pollution control point. In these facilities, through a series of treatments (physical, chemical and biological), pollutants are purified and removed from wastewater so that it can be discharged back into the environment.

Open Reading Frames (ORFs): A part of DNA that does not contain stop codons. DNA sequences are read in groups of three base pairs, corresponding to one amino acid. A long open reading frame is probably part of a gene, which encodes a protein.

Variant of Interest (VOI): A SARS-CoV-2 variant with specific genetic markers that have been associated with changes to receptor binding, and phenotypically changed compared to a reference isolate.

Variant of Concern (VOC): A SARS-CoV-2 variant that leads to increased transmissibility, increased severity of the disease, reduced efficacy of treatments or vaccines, or failure to detect the diagnosis.

Single Nucleotide Polymorphisms (SNPs): is a variation at a single nucleotide in a DNA sequence among at least 1% of the individuals.

Reverse Transcription Quantitative PCR (RT- qPCR): Variant of conventional PCR, capable of amplifying and quantifying the mRNA in real time. It allows the detection of rare transcripts and the observation of small variations in gene expression.

Operational Taxonomic Units (OTU): OTUs are clusters of reads used to classify microorganisms based on sequence similarity (The similarity threshold used is normally 97%).

Amplicon Sequence Variants (ASV): Are higher resolution analogues to OTUs, are the individual DNA sequences recovered from a high-throughput marker gene analysis, after removal of erroneous sequences generated during PCR and sequencing.

Cetyltrimethyl Ammonium Bromide (CTAB): CTAB buffer is the lysis buffer for use with the Maxwell® RSC PureFood GMO and Authentication Kit.

Polyethylene Glycol Precipitation (PEG): is a nondenaturing water-soluble polymer, used to precipitate SARS-CoV-2 viral proteins in the concentration step in wastewater samples.

Principal Coordinate Analysis (PCoA): Is a method to explore and to visualize similarities or dissimilarities of data. It starts with a distance matrix and assigns for each item a location in a low-dimensional space.

Candidate Phyla Radiation (CPR): Is a diverse group of uncultured nanobacteria lineages, characterized by their small size, small genome and with poorly understood metabolic functions.

1 INTRODUCTION

Many biological agents are presented in water, and pathogenic microorganisms can pose a risk to public and environmental health. The various environments along the urban water cycle host a complex and diverse microbial community from different sources, such as hospital, domestic or livestock effluents (Hong *et al*., 2020; Dias *et al*., 2020).

The methodologies used in microbiology laboratories for the detection and quantification of these microorganisms in environmental samples are based mainly on traditional methods, such as culture isolation, biochemical reactions, microscopy or immunological methods, since the regulations in the different countries included these methodologies as reference. Nevertheless, they have important limitations, they are laborious and time-consuming. Furthermore, in some cases their levels of specificity are low and the interpretation can be ambiguous. Due to these limitations and the emergence of new molecular techniques, significant changes are occurring in the study of environmental samples (such as in clinical and food industry) (Hong *et al*., 2020; Rompré *et al*., 2002).

Molecular methods based on (polymerase chain reaction) PCR offer greater speed, sensitivity and specificity. Thus, DNA amplification by PCR is one of the most widely used tools for the specific detection and quantification of pathogens. The PCR allows the detection and quantification of small amounts of DNA or RNA in a wide spectrum of environmental samples such as water, soil, mud, feeding, and clinic (Rompré *et al*., 2002). Methods for the detection and quantification of microorganisms in water samples by polymerase chain reaction (PCR), generally are not included in the different regulations, therefore traditional methods are still the most widely used and used as mandatory to demonstrate the quality of waters for the authorities.

Nevertheless, the application of new detection methods based on molecular biology techniques is very promising and will improve the diagnosis in the characterization and diversity of microorganism communities in water samples. For example, the implementation of these methodologies will improve the assessment of health risks, the performance of treatment plants, waste management and even the detection of antibiotic resistant bacterial genes (Isaac & Sherchan, 2020; Numberger *et al*., 2019; Chu *et al*., 2018). The microbiological diagnostic is moving towards the use of non-targeted methods. These allow the simultaneous analysis of many pathogens, studies of phylogenetic diversity or functional potential (Hong *et al*., 2020).

1.1 METAGENOMICS

Metagenomics is the study of the genomes of the microorganisms present in a sample, by direct extraction of their nucleic acids (DNA or DNA/RNA for viruses), sequencing and bioinformatic analysis. Metagenomics helps to understand microbial ecosystems by studying microorganisms in their environment without the need of isolation in pure culture. It allows us to observe and study the presence of the high percentage of microorganisms that cannot be cultured. Metagenomic analysis is not limited to environmental prokaryotes, but also includes eukaryotes and single-cell isolates of bacteria and viruses (Martin, 2020; Li *et al*., 2015; Martinez-Hernandez *et al*., 2017). Depending on the target and/or sample type, sample preparation and nucleic acid extraction should be different to maximize the yield of the procedure and to obtain a good quality and quantity of genetic material for sequencing (Hong *et al*., 2020).

The most widely used sequencing-based approach to understand phylogenetic diversity and relative taxon abundances in a sample, is targeted metagenomics. This approach usually relies on PCR amplification and sequencing of target sequences, such as 16S or 18S ribosomal RNA gene, that act as phylogenetic markers (Li *et al*., 2015).

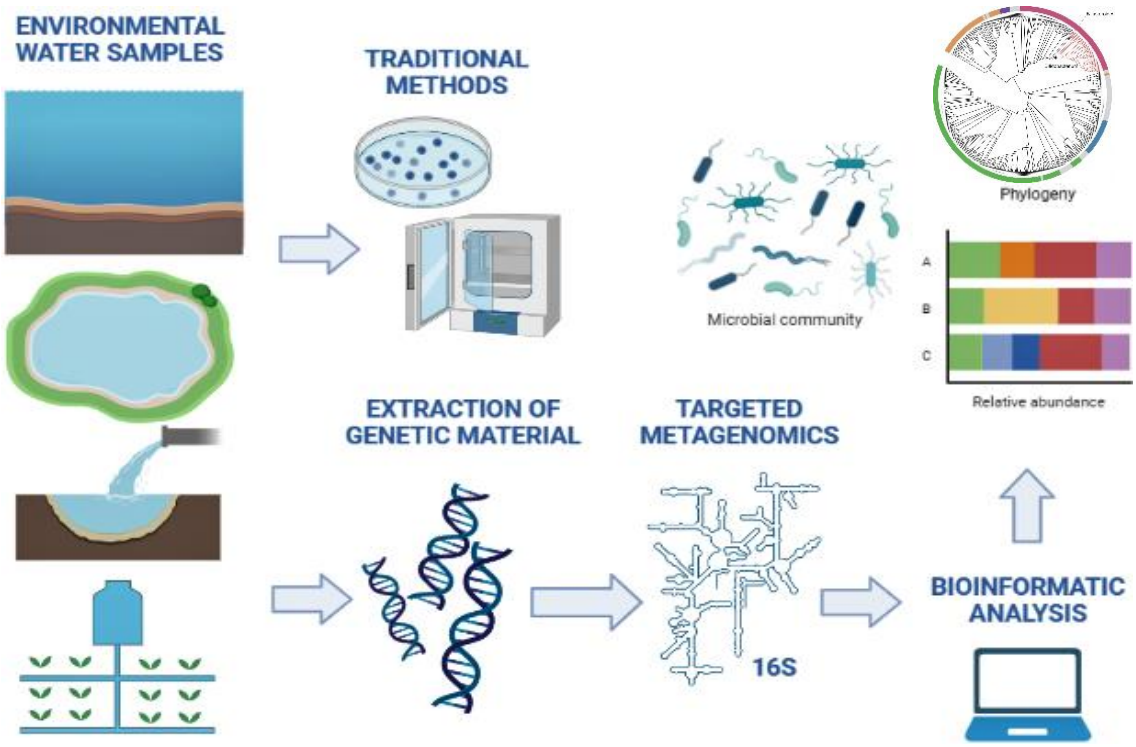


Figure 1.1 Diagram of the processing of water samples (wastewater, reclaimed water, seawater or freshwater) by two strategies: traditional methods or by metagenomic study (extraction of genetic material, targeted sequencing and bioinformatics analysis). Own elaboration.

The lower cost of sequencing, the speed and the amount of data it generates, promote the use and the development of metagenomics in the clinical sector, in the food industry, in industrial production and the environmental sector (Handelsman, 2004).

Focused on environment, the use of metagenomic for remediation has helped to understand the different microbiological communities that carry out these processes (Li *et al*., 2015). The study of direct nucleic acids extraction has also been used to describe the microbiological communities present in drinking water subjected to different disinfection treatments (Techtmann & Hazen, 2016), in biofilms (Alves *et al*., 2018; Gomez-Alvarez *et al*., 2018), in groundwater, marine samples (Li *et al*., 2015) and reclaimed water (Hong *et al*., 2020).

1.2 METAGENOMICS IN RECLAIMED WATER

The Royal Decree 1620/2007, which establishes the legal regime for the reuse of purified water, only contemplates culture isolation methods for the detection of key microorganisms in this matrix. The presence of eggs of intestinal Nematodes (*Ancylostoma, Tnchuns, Ascaris, Taenia saginata* and *Taenia solium), Escherichia coli, Legionella spp.* and *Salmonella spp* (Real Decreto 1620/2007, BOE 2007).

The use of metagenomics could allow better identification of potential pathogenic taxa and other microorganisms or viruses harmful to the environment and with risk in public and environmental health. There are several studies on reclaimed water, studying the microorganisms and also viruses present in the samples (Hong *et al*., 2020; Stüken & Haverkamp, 2020).

Next generation sequencing (NGS) performed a massively parallel or deep sequencing, that allows for rapid sequencing of whole genomes, sequencing of target regions of DNA or RNA. Bioinformatics analyses are used to link fragments by mapping reads against a reference genome (Behjati & Tarpey, 2013). These NGS technologies, which are becoming more affordable, the development of new and faster informatics tools and improved databases will further facilitate the use of metagenomics in the field of reclaimed water (Hong et al., 2020; Rusiñol *et al.,* 2019). The presented project aims to improve the control of reclaimed water that can help to better decisions making in the wastewater treatment plants (WWTP).

1.3 SARS-CoV-2 IN WASTEWATER

In December 2019, an outbreak of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) occurred in Wuhan, (China), that has spread around the world, to this day. At the time of writing (9 August 2021), nearly 203 million cases were confirmed worldwide, with over 4.285.000 deaths, and a total of 4.033.124.099 COVID-19 vaccine doses have been administered (https://www.who.int/docs/default-source/coronaviruse/weekly-updates/wou_2021_9-aug_cleared.pdf?sfvrsn=26deb277_3&download=true).

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) belongs to the beta-coronavirus 2B lineage and is a single-stranded RNA virus. Its genome includes a variable number of open reading frames (ORFs). The 5′ ORF codes for 16 non-structural proteins (nsp1-16), and the 3′ ORF encodes for several structural proteins. The four main structural proteins are: the surface spike protein (S), the envelope protein (E), the matrix protein (M) and the nucleocapsid protein (N) (Giovanetti *et al*., 2021).

SARS-CoV-2 genome is susceptible to mutation, resulting in sequence differences between a virus and its progeny, and genetic evolution. A SARS-CoV-2 isolate is defined as a variant of interest (VOI) if mutations or a phenotypic change are observed compared to the wild-type reference sequence (severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1 (NC_045512)), and has been identified as a cause of multiple cases of COVID-19, or has been reported in several countries. Non-synonymous mutations in different proteins determine functional variants, some of which are considered Variants of Concern (VOC) because of their impact on public health (Cascella *et al*., 2021).

A VOI can be upgraded to VOC if it has been shown to have the potential to cause increased transmissibility, virulence, a damaging change in epidemiology, a change in the clinical presentation of the disease, or a decrease in the effectiveness of available public health, screening and social measures (such as treatment or vaccination) (https://www.who.int/docs/default-source/coronaviruse/situation-reports/20210225_weekly_epi_update_voc-special-edition.pdf).

Table 1.1 Variants of concern (VOC) identified.

WHO LABEL	PANGO LINEAGES	GISAID CLADE	NEXT STRAIN CLADE	ADDITIONAL aa CHANGES MONITORED*	EARLIEST DOCUMENTED SAMPLES	DATE OF DESIGNATION
Alpha	B.1.1.7	GRY	20I (V1)	+S:484K +S:452R	United Kingdom, Sep-2020	18-Dec-2020
Beta	B.1.351 B.1.351.2 B.1.351.3	GH/501Y.V2	20H (V2)	+S:L18F	South Africa, May-2020	18-Dec-2020
Gamma	P.1 P.1.1 P.1.2	GR/501Y.V3	20J (V3)	+S:681H	Brazil, Nov-2020	11-Jan-2021
Delta	B.1.617.2 AY.1 AY.2	G/478K.V1	21A	+S:417N	India, Oct-2020	VOI: 4-Apr-2021 VOC: 11-May-2021

Table 1.1. Variants of concern (VOC) identified. From WHO (https://www.who.int/activities/tracking-SARS-CoV-2-variants. *Notable spike (S) amino acid changes under monitoring, which are currently reported in a minority of sequenced samples.

SARS-CoV-2 RNA is known to be present in the faeces of patients with COVID-19 and therefore in sewage. The study of sewage is being widely used to monitor, estimate the presence and quantify the amount of COVID-19 present in a community, using PCR methods. The most commonly used PCR methods for the detection of SARS-CoV-2 in wastewater samples are reverse transcription polymerase chain reaction (RT-PCR) and reverse transcription quantitative polymerase chain reaction (RT-qPCR). Thanks to these methods we can quantify the amount of SARS-CoV-2 to estimate the prevalence of infected individuals and help to take preventive measures (Farkas *et al*., 2021; Tran *et al*., 2020).

Analysis of SARS-Cov-2 sequences, based on the extraction of genetic material from sewage samples, has been directed to map Single Nucleotide Polymorphisms (SNPs) that have not been reported by clinical approaches before to identify variants of interest or variants of concern (Farkas *et al.,* 2020).

Commission Recommendation (EU) 2021/472 to establish systematic surveillance of SARS-CoV-2 and its variants in wastewater in the European Union underlines the importance of using appropriate RT-qPCR methods to detect and quantify the amount of virus in the water; and the emerging use of Next Generation Sequencing to analyse its variants (EU. Commission Recommendation 2021/472, 2021).

The use of next generation sequencing reveals the genetic diversity and the different SARS-CoV-2 variants present in a wastewater sample, based on the profile of characteristic mutations. Surveillance of SARS-CoV-2 and the emergence of new variants in wastewater is nowadays essential for the public health decision-making process (variant incidence, containment of regions...). It should therefore be included more systematically in national COVID-19 screening and control strategies. In this work, a method for SNP calling and identification of the different variants present in wastewater samples has been developed, following the EU Commission Recommendation (EU. Commission Recommendation 2021/472, 2021).

1.4 BIOINFORMATICS TOOLS FOR RECLAIMED WATER SAMPLES (AMPLICON DATA)

Nowadays different bioinformatics tools for the analysis of microbial amplicon sequencing data are available. Quantitative Insights Into Microbial Ecology (QIIME 2) (Bolven *et al*., 2019) and MOTHUR (Schloss *et al*., 2009) have been the most widely used (Muskesh, 2020), however, we can find more like MAPseq (Matias Rodrigues, 2017), DADA2 (Callahan *et al*., 2016) a package for R environment, MG-RAST (Meyer *et al*., 2008) or Nephele, (Weber *et al*., 2018) that are online tools for automatic analyses. Each one incorporates a different classification algorithm, and can compare the raw sequences against a defined database. The taxonomic assignment of 16S amplicons is based on the comparison with known 16S-rRNA sequences which are available in repositories such as SILVA (Yilmaz *et al*., 2014), Greengenes (McDonald *et al*., 2012), or RDP (Wang *et al*., 2007) databases.

1.4.1 QIIME 2

QIIME 2 is a microbiome analysis package for Mac OS X, Windows or Linux. It performs the analysis from raw DNA sequence data and ends up with publication-quality figures and statistical results, including taxonomic assignment, phylogenetic reconstruction, and more. QIIME 2 is the new redesign version of QIIME 1, that is dismissed (Bolven *et al*., 2019)

1.4.2 MOTHUR

Mothur is an open-source software for Mac OS X, Windows or Linux. Mothur is used to analyze amplicon sequence data, to trim, screen, and align sequences; calculate distances; assign sequences to operational taxonomic units (OTUs); describe the alpha and beta diversity; calculate other diversity indexes; generate rarefaction curves based on the sequence diversity (Schloss *et al*., 2009).

1.4.3 MAPseq

MAPseq is a sequence read classification tool designed to assign taxonomy and OTU classifications to ribosomal RNA amplicon sequences. MAPseq enables a sequence read mapping against hierarchically clustered and annotated reference sequences in a database (Matias Rodrigues *et al.,* 2017).

1.4.4 DADA2

DADA2 is an open-source, fast and accurate software package for modelling and correcting Illumina-sequenced amplicon errors with single-nucleotide resolution. DADA2 is based on inferring exact amplicon sequence variants (ASVs) from amplicon data, instead of using Operational Taxonomic Units (OTUs), resolving the biological problem of the differences of 1 or 2 nucleotides (Callahan *et al*., 2016).

1.4.5 MG-RAST

The open-source metagenomics RAST (MG-RAST) service is extensible, automated, public and free. It is a high-throughput pipeline that produces functional assignments of unassembled or assembled raw sequences in the metagenome by comparing them with different databases (Meyer *et al.,*2008). The analysis and visualization of the results can be done on the same page, where there are many different options and types of graphics.

1.4.6 NEPHELE

Nephele is a free cloud-based microbiome data analysis platform with standardized pipelines and a simple web interface for processing raw data, that includes the data analysis process and different visualization tools (Weber *et al*., 2008). The results can be imported into different platforms to visualize the results and create different plots, like RStudio (to use phyloseq package), QIIME2 and MicrobiomeDB (Oliveira *et al.,* 2018).

1.4.7 BIOINFORMATIC COMPARISON

The use of one program or another can lead to differences in the sensitivity and specificity of the taxonomic assignment of our samples (Prodan *et al.,* 2020). Finding the best taxonomic analysis tools available for metagenome data analysis may improve the results. Several studies compare the different tools for metagenomics analysis trying to evaluate the best option.

In a comparison between QIIME, QIIME2, Mothur and MAPseq the best proportion of sequences classified at the genus and family level and the most precise relative abundances were obtained with QIIME2. Nevertheless, MAPseq showed the highest accuracy, being the closest to the data used in the study. The choice between QIIME 2 or MAPseq for 16S rRNA data should be based on the level of precision and or computational performance required. Being QIIME2 the most demanding tool from a computational point of view, because of higher memory usage and CPU time (Almeida *et al*., 2021).

Six bioinformatics pipelines with two different approaches have been evaluated. Three tools, QIIME, MOTHUR, USEARCH-UPARSE, rely on the definition of OTUs to assign sequences. The other three, DADA2, QIIME2 and USEARCH-UNOISE3 look at ASVs. DADA2 showed the best sensitivity and resolution, being the best tool for studies that require high biological resolution. In addition to this, ASV methods offer more resolution, better specificity, are more precise, reproducible, reusable and show lower error rates compared to the ones that are based on OUTs (Prodan *et al*., 2020; Callahan *et al*., 2017).

The database can cause differences in the taxonomic assignment. Greengenes and SILVA have been the most used databases. The SILVA 128 database performed better results than Greengenes 13_8 in terms of recall at genus and family levels. Greengenes has the smallest taxonomic classification, as it has the smallest number of reference sequences, and is much less diverse than other databases because it comprises fewer taxa (Almeida *et al*., 2018; Balvočiūtė & Huson, 2017).

Metagenomics is a developing field of research. Improvements and new technologies or programs can make metagenomics a competent tool for microbial community identification that is even more comprehensive, versatile, and cost-effective. The development of sequencing techniques, the large amount of metagenomic data that can be obtained from them, and the more complex environmental sequence data pose a challenge at the level of bioinformatics analysis (Mukesh *et al*., 2020; Zhang *et al*., 2019; Simon & Daniel, 2010).

1.5 BIOINFORMATICS TOOLS FOR SARS-COV-2

In the last year, there have been numerous bioinformatics advances related to the study of COVID-19, most of the tools are free and available online, through web pages and/or applications or public code repositories (GitHub), with different levels of computation. From the simplest to the most complex code-based (Hufsky *et al*., 2021).

1.5.1 ARTIC-ncov2019

ARTIC is a bioinformatics pipeline for virus multiplatform sequencing data. A comprehensive tool to perform analyses from raw SARS-COV2 sequencing data to consensus sequence and variant detection, including base calling, demultiplexing, mapping and polishing (Tyson *et al.,* 2020).

1.5.2 V-PIPE

V-pipe is an online tool, available via GitHub (https://github.com/cbg-ethz/V-pipe) that presents a branch to analyze SARS-Cov-2 high-throughput sequencing data from raw reads, in “fastq” file format, to SNV calling (Posada-Céspedes *et al*., 2021).

1.5.3 PANGOLIN

Pangolin (Phylogenetic Assignment of Named Global Outbreak Lineages) (Áine *et al*., 2021) allows the assignment of the most probable lineage to a SARS-CoV-2 sequence of interest. By estimating its similarity to a phylogenetic tree of representative sequences of the different SARS-CoV-2 lineages currently identified, based in the pango linkage (Rambaut *et al.,* 2020). Pangolin is available as a command line tool and as a web application.

2 OBJECTIVES

The purpose of this work was to apply new molecular and bioinformatics tools to study water microbiome to support diagnosis and decision-making in the environmental sector. The work has been focussed on:

1. **Characterize the bacterial population in reclaimed water samples using a target metagenomic approach based on the Next Generation Sequencing of the 16S ribosomal RNA gene.** The main objective of this study was to evaluate the effectiveness of the treatment in the WWTP and the potential sanitary risk of the microorganisms present in the samples. The NGS is shown as an alternative tool with added value for decision-making in industrial plants. To achieve this goal:

- Firstly, the methodology for the concentration, extraction and purification of DNA from reclaimed water samples was optimised.

- Secondly, the aim was to obtain the most complete taxonomic assignment, to describe the microbial population of our samples, the taxa present and the relative abundance of each of them.

- To know the potential risk of reclaimed water, paying special attention to the presence of possible pathogens that could have sanitary risks and are not identified using traditional methodologies.

- Finally, the taxa identified in the different collection points and their relative abundance have been compared.

2. **Characterize SARS-CoV-2 variants in Wastewater samples by performing a genomic analysis following the European recommendation (EU2021/472) to demonstrate that genomic surveillance can help in the monitoring of SARS-CoV-2 variants circulating in a community. To achieve this goal:**

- Processing of SARS-CoV-2 sequences obtained from wastewater samples and identification of single nucleotide variants (SNVs) have been carried out.

3 METHODS

3.1 RECLAIMED WATER SAMPLES PROCESSING

3.1.1 SAMPLE DESCRIPTION

A total of 24 samples were collected from different points of a wastewater treatment plant (WWTP). Two points - point 1 (p1_effluent) and point 2 (p2_tank) reclaimed water distribution tanks - were analysed during four weeks. Each week one sample was collected from the different sampling points. One litre of sample was processed by triplicate.

Point 1 (p1_effluent): Effluent leaving the Waste Water Treatment Plant (WWTP) was the water leaving the plant, after pre-treatment, primary and secondary decantation without chlorination.

Point 2 (p2_tank): Reclaimed water tank, the water is subjected to tertiary physico-chemical treatment and disinfection by chlorination with sodium hypochlorite. This point is the one that is under the legislation of the Royal Decree 1620/2007.

3.1.2 SAMPLE CONCENTRATION

To perform the concentration, sterile polycarbonate membranes of 0.22 µm pore size and 47 mm in diameter (Millipore) for each 250mL of reclaimed water (4 filters per 1L of sample) were used. After filtration, the membrane was transferred to a 50mL Falcon tube with 10mL of sterile water.

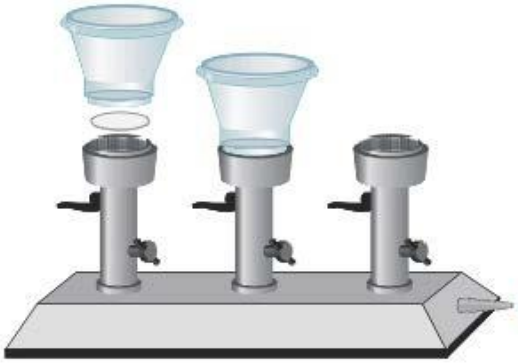


Figure 3.1. Membrane filtration system.

To ensure the correct elution of all bacterial cells from the filter, the tubes were vortexed for 7 minutes. After washing the tubes, the volume obtained was concentrated using 15 mL Amicon® Ultra 100K filter cartridges (Millipore), and centrifuged for 30 minutes at 2.200 rpm until obtaining a final volume of 500 µL. This volume was used for the DNA extraction.

Figure 3.2. Amicon Ultra-15 filter cartridge (Millipore) used for concentration of reclaimed water samples.

3.1.3 EXTRACTION OF DNA

The DNA extraction was carried out with the Maxwell® RSC Instrument (Promega), an automated nucleic acid purification platform. The Maxwell® RSC PureFood GMO and Authentication kit was used. A volume of 200 µL of concentrated sample was used and processed following the provider instructions. Briefly, 400 µL of CTAB Buffer, 40 µL of Proteinase K and 25 µL of PLANT are added to each sample and vortexed. Samples were next placed in a heat block at 64°C for 10 minutes. After incubation, vials were centrifuged at 16,000 g for 5 minutes and loaded into the Maxwell. Total 300 µL of lysis buffer in the first well and the elution tube were added. The sample was eluted in 100 µL of nuclease-free sterile water.

Figure 3.3. The Maxwell Rapid Sample Concentration (RSC) Instrument used in this work.

Finally, the concentration and purity of the genetic material extracted from the samples were verified by using Biophotometer plus (Eppendorf). Concentration was measured in (ng/µL) and purity was based on 260/280 and 260/230 ratios, which should be between 1.6 and 2 to obtain correct purity.

Regions V1-V4 may provide more accurate results than other regions in terms of taxonomic assignment, as there is a greater presence of partial sequences corresponding to these regions in the databases (Hong *et al*., 2020). Therefore, it was decided to use these V1-V4 regions of the 16S rRNA gene as a marker.

Figure 3.4. Biophotometer plus used in this work

3.1.4 SEQUENCING

The extracted genetic material was sent for sequencing by high-throughput amplicon sequencing of the regions V1-V4 of the 16S target gene. The Illumina Miseq sequencing 300bp×2 approach was used. Sequencing was performed after 25 PCR cycles.

3.1.5 BIOINFORMATIC ANALYSIS

The bioinformatic analysis of the reclaimed water samples was performed with DADA2 (23) using R software package version 3.6.0. (R Core Team, 2020)

Raw forward and reverse reads were processed using the DADA2 R package version 3.10 (Callahan *et al*., 2016). The RDP algorithm (Wang *et al*., 2007) was used for sequence classification as implemented in DADA2.

Taxonomic assignment of amplicon reads to phylotypes was performed using Silva Database version 138 (Yilmaz *et al*., 2014)

The differential relative abundance of taxa was tested using the non-parametric Kruskal-Wallis test. Significant threshold was set at 0.05.

The visualization of microbial communities' structure was done through the principal coordinates analysis (PCoA) plots, using the R package ggplot2 (Wickham, 2016). The Weighted Unifrac (phylogenetic quantitative measure) ecological distance between samples was chosen to perform the ordination analysis. The significance of groups present in community structure was tested using the analysis of similarity statistics (ANOSIM), the significant threshold was set at 0.05. ($p < 0.05$)

Other R packages like phyloseq, DECIPHER, DESeq2, vegan… were required to install during the analysis.

```
######################
### DADA2 installation

if (!requireNamespace("BiocManager", quietly = TRUE))
install.packages("BiocManager")
BiocManager::install("dada2", version = "3.10")
library(dada2)

#### Set Working directory where is the data
code_path <-("C:/Users/Albaburillo/Desktop/fastq_16S")
setwd(code_path)

#### Download RawData and Training Data
RawDataPath <- paste(code_path,"/fastq_16S/",sep="")
TrainingPath <- paste0(code_path,"/Training/")

### Sort ensures forward/reverse reads are in same order
fnFs <- sort(list.files(pattern="_R1_001.fastq.gz"))
fnRs <- sort(list.files(pattern="_R2_001.fastq.gz"))

### Infer sample names from file names
sample.names <- sapply(strsplit(fnFs, "_S"), `[`, 1)

fnFs <- file.path(fnFs)
fnRs <- file.path(fnRs)

######################
### Inspect read quality profiles

# Fordward reads
plotQualityProfile(fnFs[1:4])

# Reverse reds
plotQualityProfile(fnRs[1:4])

### Place filtered files in filtered/ subdirectory
filt_path <-file.path(code_path, "DADA2/filtered")
filtFs <-file.path(filt_path,paste0(sample.names,"_F_filt.fastq.gz"))
filtRs <-file.path(filt_path,paste0(sample.names,"_R_filt.fastq.gz"))

###################
### Filter And Trim

out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(250,250),
maxN=0, maxEE=c(2,5), truncQ=2, rm.phix=TRUE,compress=TRUE,
multithread=FALSE,verbose=T)

###################
### Learn the Error Rates
# From filtered reads, we iteratively learn an error profile

errF <- learnErrors(filtFs, multithread=TRUE, nbases=50000000,
randomize=T, verbose=T)
errR <- learnErrors(filtRs, multithread=TRUE, nbases=50000000,
randomize=T, verbose=T)
```

Figure 3.5. Pipeline to analyze the Reclaimed Water samples, in R with DADA2.

```
### Derreplication
derepFs <- derepFastq(filtFs, verbose=TRUE)
derepRs <- derepFastq(filtRs, verbose=TRUE)

# Name the derep-class objects by the sample names
names(derepFs) <- sample.names
names(derepRs) <- sample.names

### DBModule$insertLogs("Removing sequencing errors",job_id,1)
dadaFs <- dada(derepFs, err=errF, multithread=TRUE)
dadaRs <- dada(derepRs, err=errR, multithread=TRUE)

###################
### Merge paired reads
# DBModule$insertLogs ("Merging denoised forward and reverse
reads.",job_id,1)
mergers <- mergePairs(dadaFs, derepFs, dadaRs, derepRs,
verbose=TRUE)

# Inspect the merger data.frame from the first sample
head(mergers[[1]])

####################
### Construct sequence table
seqtab <- makeSequenceTable(mergers)

####################
### Remove chimeras
# DBModule$insertLogs("Removing chimeras",job_id,1)

seqtab.nochim <- removeBimeraDenovo(seqtab, method="consensus",
multithread=TRUE, verbose=TRUE)
seqtab.nochim.pooled <- removeBimeraDenovo(seqtab, method="pooled",
multithread=TRUE, verbose=TRUE)

#####################
### Track reads through the pipeline

dim(seqtab.nochim)
sum(seqtab.nochim)/sum(seqtab)
getN <- function(x) sum(getUniques(x))
track <- cbind(out, sapply(dadaFs, getN), sapply(mergers,
getN),rowSums(seqtab.nochim))
colnames(track) <- c("input", "filtered", "denoised", "merged",
"nonchim")
rownames(track) <- sample.names
head(track)

###################
### Export into a xlsx file
library(openxlsx)
write.xlsx(track,"table16S.xlsx")
```

Figure 3.6. Pipeline to analyze the Reclaimed Water samples, in R with DADA2.

```
####################
### Assing taxonomy

taxa.rdp<- assignTaxonomy(seqtab.nochim,
paste(code_path,"/Training/rdp_train_set_16.fa.gz",sep=""),
multithread=TRUE)

taxa.silva<-readRDS("taxa.silva.rds")

taxa.silva<- assignTaxonomy(seqtab.nochim,
paste(code_path,"/Training/silva_nr_v132_train_set.fa.gz",sep=""),
multithread=TRUE)

taxa.rdp <- addSpecies(taxa.silva,
paste0(code_path,"/Training/rdp_species_assignment_16.fa.gz",""),
verbose=T)

taxa.silva <- addSpecies(taxa.silva,
paste0(code_path,"/Training/silva_species_assignment_v132.fa.gz",
""),verbose=T)

library(DECIPHER)
seqs <- getSequences(seqtab.nochim)
names(seqs) <- seqs
alignment <- AlignSeqs(DNAStringSet(seqs), anchor=NA,verbose=FALSE)

####################
#### Read metadata

metadata<-read.csv(paste0(code_path,"/metadata.csv"))
rownames(metadata)<-metadata$SampleID

####################
#### Create PhyloSeq Object

ps_silva<- phyloseq(otu_table(seqtab,taxa_are_rows=FALSE),tax_table
(taxa.silva))

### Clean phyloseq object from mock controls
sample_data(ps_silva)<-metadata
```

Figure 3.7. Pipeline to analyze the Reclaimed Water samples, in R with DADA2.

3.2 SARS-CoV-2 SAMPLES PROCESSING

A group of ten waste water samples were processed to analyse the different variants of SARS-CoV-2. Samples were collected in different WWTP in the Valencian Community. To total of ten samples were processed (Barcode 14 to Barcode 23). Nine of them positive for SARS-CoV-2 by RT-qPCR; one wastewater sample with Synthetic RNA positives controls for the Gamma and Delta variants (Twist Synthetic SARS-CoV-2 RNA Control, Twist Bioscience, USA) used as positive control and one negative sample. The positive samples were selected based on their quantification results (Ct values between 32 and 37). The samples were processed followed the standard protocol implemented at LABAQUA and accredited by ENAC under ISO 17025.

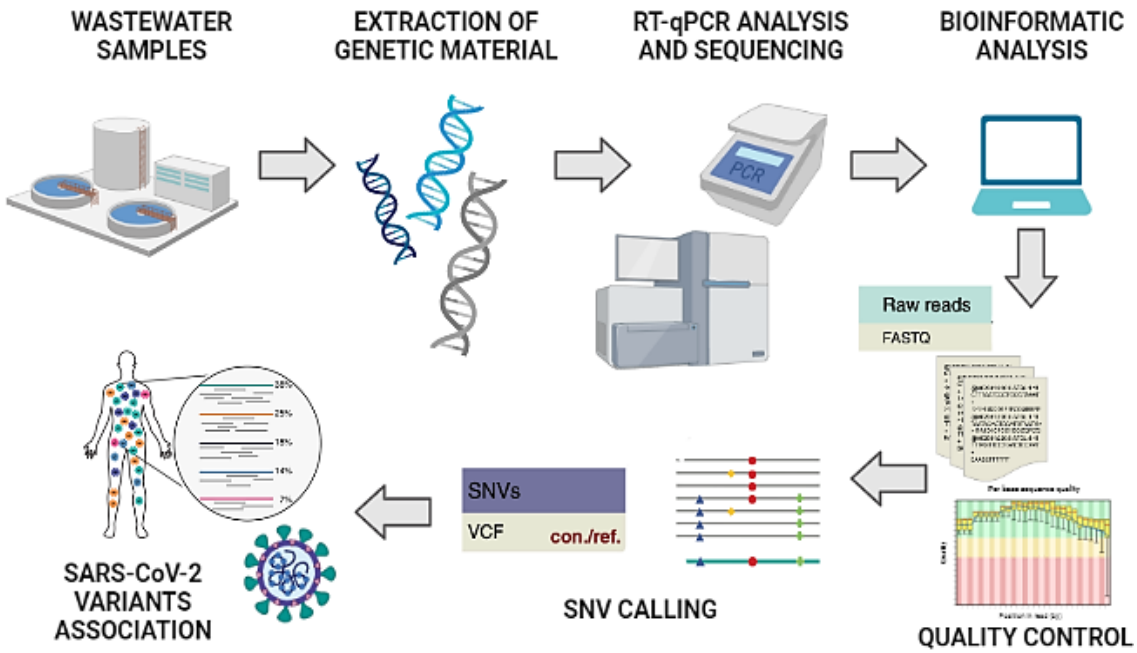


Figure 3.8. Schematic diagram of the working procedure applied to SARS-Cov-2 samples, from sample collection to SPNs identification. Own elaboration, adapted from V-PIPE (Posada-Céspedes et al., 2021).

3.2.1 SAMPLE CONCENTRATION

The concentration step was performed using polyethylene glycol (PEG) precipitation protocol (Hjelmsø et al., 2017). Briefly, 125 mL of glycine buffer (0.05 M glycine, 3% beef extract, pH 9.6) were added to 1 L of sample and shacked for 2 hours at 5 ± 3ºC.

The sample was then centrifuged at 8,000×g for 30 min, and the supernatant was filtered through 0.45 µm polyethersulfone membrane. The filtrate was collected in a sterile tube and supplied with PEG 8000 (80 g/L) and NaCl (17.5 g/L) to favor the precipitation of viral particles. The sample was homogenized for 12-16 hours at 5 ± 3°C, centrifuged for 90 minutes at 13,000×g. The supernatant was removed by decanting, taking care not to lose the pellet. Finally, the resulting pellet was eluted with 1 mL of phosphate buffered saline (PBS).

3.2.2 EXTRACTION OF RNA

After concentration, RNA contained wastewater samples was extracted using QIAmp Viral RNA mini kit by (QIAGEN) adapted to wastewater samples. To start, 5.6 µL of Carrier were added to 560 µL of AVL buffer and mixed. The above mixture was next supplied to 140 µL sample, homogenized for 15 sec using a vortex and incubated at room temperature for 10 min. After incubation, 560 µL of ethanol (96-100% purity) were added to the sample and vortexed again for 15 sec. Then, 630 µL of the mixture were poured onto the column, a 2 ml tube was inserted and centrifuged for 1 minute at 6000 xg. About 500 µL of AW1 buffer were added to the column and centrifuged again. The column was transferred to a new 2 ml tube and the filtrate was discarded. A total of 500 µL of buffer AW2 were added and centrifuged at 17000 xg for 3 min. The column was retransferred to a new 2 ml tube, the filtrate was discarded and the column was centrifugated again at 17000 xg for 1 min. The column was now transferred to a 1,5 ml RNase-free tube and 60 µL of AVE buffer was added. Finally, the column was incubated for 1 min and centrifuged at 6000 xg for 1 min, resulting in an eluate of about 50 µL containing the total RNA of the sample.

3.2.3 RT-qPCR ANALYSIS

The presence of SARS-CoV-2 was quantified through RT-qPCR using the reaction mix Taqman Virus 1-step Master Mix (Applied Biosystems, Ref. 4444436). The primers and probes for the amplification of three SARS-CoV-2 coronavirus specific targets (the SARS-CoV-2 ORF1ab gene, the S protein gene and the N protein gene) supplied in the "TaqManTM 2019nCoV Assay Kit v1" (Applied Biosystems, Ref. A47532) were also added. As an internal amplification control the human RNase P encoding gene was used. The Applied Biosystems 7500 PCR machine was used.

3.2.4 SEQUENCING

The ARTIC protocol was followed using Nanopore sequencing technology, to characterize low-frequency variants in the SARS-CoV-2 genome, such as single nucleotide polymorphisms (SNPs). High-throughput sequencing was carried out in GridION Mk1 (Oxford Nanopore Technologies).

3.2.5 BIOINFORMATIC ANALYSIS

The sequences of SARS-CoV-2 samples were analyzed following the ARTIC nCoV-2019 novel coronavirus bioinformatics protocol (Tyson *et al*., 2020).

The consensus sequence of each sample was generated with minimap2 (Li, 2018) and the medaka method was used to call the single nucleotide variants (both included in the pipeline). We used medaka because only the base called data is required (.fasta or .fastq). Medaka is faster than Nanopolish, which requires more files produced by the Nanopore sequencing and it is more time consuming.

The medaka model used was: --medaka_model r941_min_fast_g303. This option is suitable for MinION or GridION R9.41 flow cell data using Guppy fast base caller version 3.0.3.

```
########################
### Read filtering

# Because ARTIC protocol can generate chimeric reads, we perform
length filtering. We first collect all the FASTQ files into a
single file for each sample.

artic guppyplex --min-length 250 --max-length 600 --directory
barcode* --rprefix run_name

# we have obtained a run_name_barcode*.fastq file with the
consensus sequence for each sample.

######################
### Medaka pipeline

# Replace samplename (barcode*) as appropriate.

artic minion --medaka --medaka-model r941_min_high_g303 --normalise
200 --threads 4 --scheme-directory ~/artic-ncov2019/primer_schemes
--read-file run_name_barcode*.fastq nCoV-2019/V1 barcode*
```

Figure 3.9. Pipeline to analyze the SARS-Cov-2 sequences following the ARTIC protocol with the Medaka model.

Information such as genome coverage or reference matching for each sample can be viewed online and real time using RAMPART (Read Assignment, Mapping and Phylogenetic Analysis in Real Time) software. Accessible through a Docker container.

```
######################
### To visualize in RAMPART

# Move to the pass directory with all the run_name_barcode*.fastq
files

sudo docker pull ontresearch/artic_rampart:latest

sudo docker run -it -e LOCAL_USER_ID=`id -u $USER` --mount
type=bind,source="$(pwd)",target=/data -p 3000:3000 -p 3001:3001
ontresearch/artic_rampart:latest
```

Figure 3.10. Pipeline to visualize in RAMPART the analysis of the SARS-Cov-2

4 RESULTS AND DISCUSSION

4.1 RECLAIMED WATER SAMPLES

In this study, the bacterial composition of 24 reclaimed water samples was evaluated with a marker-based approach using the 16S ribosomal RNA subunit gene (rRNA16S). Quantity and quality of the genetic material obtained after the processing of the different samples and extraction are shown in Table 4.1.

Table 4.1. Volume of samples concentrated and Quality of the DNA extracted from the reclaimed water samples included in the study*. Respective sample ID, collection date, collection point, initial volume of the sample, filtered volume, DNA concentration in ng/ul, ratio 260/280, ratio 260/230 and final extraction volume in uL.*

	ORIGINAL ID	COLLECTION DATE	COLLECTION POINT	REPLICATE	FILTERED VOLUME	DNA CONCENTRATION (ng/uL)	RATIO 260/280	RATIO 260/230	VOLUME (µL)
WEEK 1 (t1) POINT 1	M1.1.1	05/05/2021	P1_EFFLUENT	R1	1 L	0,164	1,82	1,93	60
	M1.1.2	05/05/2021	P1_EFFLUENT	R2	1 L	0,202	1,85	1,84	70
	M1.1.3	05/05/2021	P1_EFFLUENT	R3	1 L	0,139	1,78	2,00	60
WEEK 1 (t1) POINT 2	M1.2.1	05/05/2021	P2_TANK	R1	1 L	0,193	1,80	2,07	60
	M1.2.2	05/05/2021	P2_TANK	R2	1 L	0,147	1,79	2,17	70
	M1.2.3	05/05/2021	P2_TANK	R3	1 L	0,271	1,81	1,91	70
WEEK 2 (t2) POINT 1	M2.1.1	12/05/2021	P1_EFFLUENT	R1	1 L	0,214	1,74	1,72	70
	M2.1.2	12/05/2021	P1_EFFLUENT	R2	1 L	0,248	1,77	1,92	70
	M2.1.3	12/05/2021	P1_EFFLUENT	R3	1 L	0,167	1,72	2,11	70
WEEK 2 (t2) POINT 2	M2.2.1	12/05/2021	P2_TANK	R1	1 L	0,433	1,81	1,85	70
	M2.2.2	12/05/2021	P2_TANK	R2	1 L	0,332	1,76	2,05	70
	M2.2.3	12/05/2021	P2_TANK	R3	1 L	0,389	1,78	1,90	70
WEEK 3 (t3) POINT 1	M3.1.1	19/05/2021	P1_EFFLUENT	R1	1 L	0,308	1,83	2,16	70
	M3.1.2	19/05/2021	P1_EFFLUENT	R2	1 L	0,213	1,77	2,27	70
	M3.1.3	19/05/2021	P1_EFFLUENT	R3	1 L	0,160	1,71	2,39	70
WEEK 3 (t3) POINT 2	M3.2.1	19/05/2021	P2_TANK	R1	1 L	0,226	1,76	2,09	70
	M3.2.2	19/05/2021	P2_TANK	R2	1 L	0,320	1,83	1,99	70
	M3.2.3	19/05/2021	P2_TANK	R3	1 L	0,229	1,77	2,14	70
WEEK 4 (t4) POINT 1	M4.1.1	26/05/2021	P1_EFFLUENT	R1	1 L	0,102	1,79	1,76	70
	M4.1.2	26/05/2021	P1_EFFLUENT	R2	1 L	0,068	1,80	1,8	70
	M4.1.3	26/05/2021	P1_EFFLUENT	R3	1 L	0,056	1,83	1,87	70
WEEK 4 (t4) POINT 2	M4.2.1	26/05/2021	P2_TANK	R1	1 L	0,219	1,80	1,92	70
	M4.2.2	26/05/2021	P2_TANK	R2	1 L	0,307	1,81	1,84	70
	M4.2.3	26/05/2021	P2_TANK	R3	1 L	0,294	1,83	1,84	70

Table 4.1. also shows the description of the samples. The collection point, collection date and the replicate columns display metadata categories that will be used for comparison tests and visualizations.

The total number of reads per sample, before and after quality control are summarized in Table 4.2. After quality control, 9,038 phylotypes were detected and assigned to species level.

Table 4.2. Number of reads per sample after quality filtering and denoising. *Respective SAMPLE ID, ORIGINAL ID, INPUT (number of reads before quality filtering), FILTERED (number of reads retained after quality filtering and trimming), DENOISED (number of reads retained after denoising steps), MERGED (number of merged reads, only merged reads are kept for further analysis) and NONCHIM (number of reads retained after chimera removal).*

SAMPLE ID	ORIGINAL ID	INPUT	FILTERED	DENOISED	MERGED	NONCHIM
09284AAC-M1-16	M1.1.1	81642	74997	44745	31297	13957
09284AAC-M2-16	M1.1.2	86749	79530	50014	35492	14944
09284AAC-M3-16	M1.1.3	70456	64588	39059	27628	12201
09284AAC-M4-16	M1.2.1	67138	60730	36036	24680	12200
09284AAC-M5-16	M1.2.2	60713	55095	32286	22323	10195
09284AAC-M6-16	M1.2.3	77091	70326	41811	28985	13253
09284AAC-M7-16	M2.1.1	70268	64408	40324	28528	11600
09284AAC-M8-16	M2.1.2	60977	55570	34014	23899	10450
09284AAC-M9-16	M2.1.3	60402	55550	32284	22070	9287
09284AAC-M10-16	M2.2.1	76410	70076	42178	28837	11852
09284AAC-M11-16	M2.2.2	65703	60122	35000	24080	10292
09284AAC-M12-16	M2.2.3	59039	53834	32239	22566	10491
09284AAC-M13-16	M3.1.1	59393	54398	32777	23694	10291
09284AAC-M14-16	M3.1.2	61557	56441	32745	23494	10835
09284AAC-M15-16	M3.1.3	74571	68708	42283	30260	13195
09284AAC-M16-16	M3.2.1	64085	58551	34481	24789	11046
09284AAC-M17-16	M3.2.2	66960	61394	36240	25557	11263
09284AAC-M18-16	M3.2.3	63795	58731	33251	23117	9617
09284AAC-M19-16	M4.1.1	88709	81552	41093	27159	12261
09284AAC-M20-16	M4.1.2	86924	79595	43908	29524	12846
09284AAC-M21-16	M4.1.3	89231	81908	45756	31685	14193
09284AAC-M22-16	M4.2.1	69532	63785	37098	25784	11631
09284AAC-M23-16	M4.2.2	82597	76227	45018	30900	13395
09284AAC-M24-16	M4.2.3	87137	79732	47205	32847	13793

4.1.1 DESCRIPTION OF THE MICROBIAL POPULATION

4.1.1.1 PHYLUM LEVEL

The taxonomic assignments, showed that microbial clades mostly belong to the Bacteria domain. This result was somehow expected, and confirmed the reliability of the used procedure, since the primers used for the amplification of the 16S Ribosomal RNA gene were specifically designed for Bacteria. Nevertheless, the analysis assigned some reads to Archaea, that possess a 16S rRNA which diverges from Bacteria.

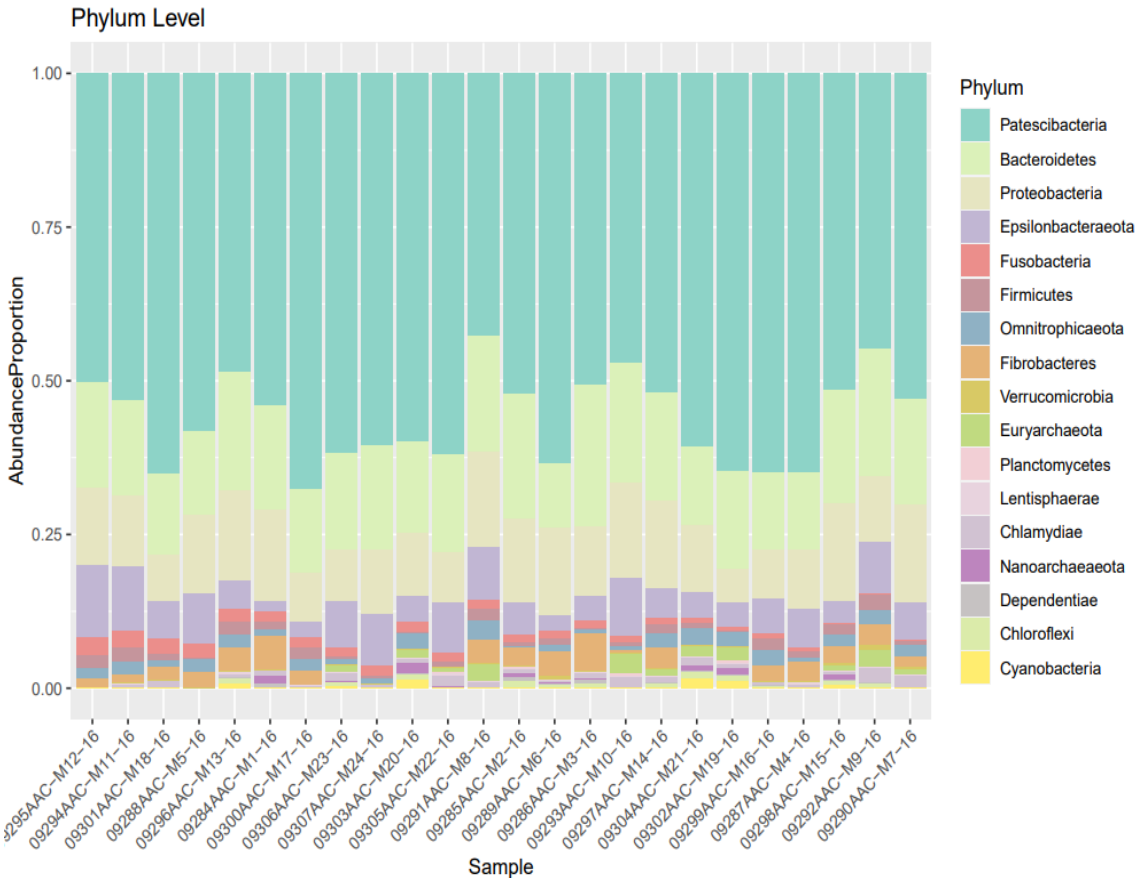


Figure 4.1. Barplot of the Relative Abundance for Phylum Level of the 24 reclaimed water samples.

The most abundant Phylum, Patescibacteria, is present in all samples, in a similar range of abundances assessed around 40-50% of total reads. Other abundant phyla were Bacteroidetes, with a relative abundance of around 20-30%, and Proteobacteria with 15-20% (Fig 4.1.).

The phylum Patescibacteria is very diverse and widespread in terrestrial environments, groundwater, sediments, lakes and other aquifers. They belong to the Candidate Phyla Radiation (CPR), or nanobacteria group, characterized by small size, small genome size and a very simple metabolism; they cannot growth in laboratory conditions and were discovered through culture-independent approaches (Hermann *et al.,* 2019; Jaffe *et al.,* 2020).

The relative abundance of this phylum is consistent with the result described in the studies of Bruno et al., 2017 in groundwater environments and drinking water treatment plants which have been found to contain a particularly high abundance of up to 38% of the total microbiome (Bruno *et al*., 2017).

The presence of the phylum Proteobacteria also stands out, because of their relative abundance throughout all samples and because many common human pathogens are found in this phylum. Including, the genera *Brucella* (Alphaproteobacteria), *Neisseria* (Betaproteobacteria) or *Escherichia, Salmonella* and *Legionella* (Gammaproteobacteria). The presence of these genera belonging to the class Gammaproteobacteria in reclaimed water might be index of fecal contamination, prohibiting its use, as they pose a great risk to health according to Royal Decree 1620/2007 (Rizzatti *et al*., 2017; Royal Decree 1620/2007, BOE 2007).

The phylum Cyanobacteria was also detected in some of the samples. Some among these aquatic bacteria can release toxins which are considered harmful to human and animal health (Kori *et al.,* 2019). Therefore, it may be interesting to further study and take into account the presence of these microorganisms for decision-making and risk control in water treatment plants.

On the other hand, Nanoarchaeaeota, Euryarchaeota, Omnitrophicaeota are the phyla that belong to the Archaea domain. These phyla are not present in all samples, and their relative abundances in the microbial population are very low.

The phylum Nanoarchaeaeota has been described in high-temperature geothermal springs and marine hydrothermal vents, and shows evolutionary convergences with the nanobacteria group. These organisms are obligate symbionts of other Archaea and are also non-cultivable (Jarett *et al.,* 2016). The CPR and the Nanoarchaeaeota phyla have the potential for used bioremediation of pollution and the production of dangerous secondary metabolites (Ludington *et al*., 2017).

4.1.1.2 GENUS LEVEL

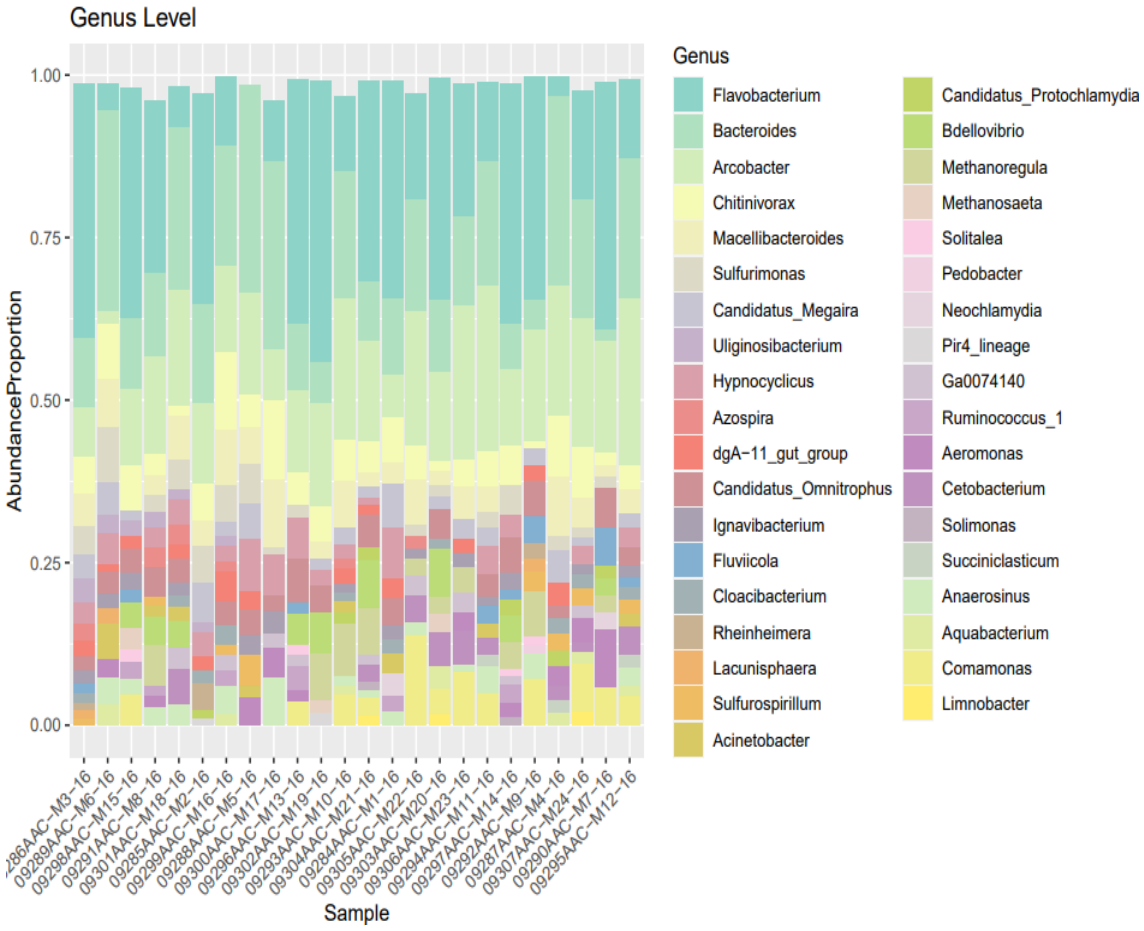


Figure 4.2. Barplot of the Relative Abundance for Genus Level of the 24 reclaimed water samples.

The genera with the highest relative abundances are *Flavobacterium*, *Bacteroides* and *Acrobacter.* (Fig 4.2.). The most abundant genus is *Flavobacterium*, but we can observe that there is a clear decrease in its relative abundance in samples (M4, M5, M6, M10, M11, M12, M16, M17, M18, M22, M23 and M24) belonging to sampling point 2 (p2_tank) reclaimed water distribution tanks. In the remaining samples, which come from sampling point 1, effluent, the relative abundance is around 30%.

Flavobacterium is waterborne and frequently found in water distribution systems, which leads to the conclusion that it could be resistant to disinfection treatments (Moreno-Mesonero *et al*., 2020). In our case, it is observed that the tertiary treatment induced a significant decrease in the relative abundance in the samples taken from the reclaimed water tank.

4.1.2 DIFERENTIAL ABUNDANCE

The taxa with differential abundances between the different sample points, that were significant for the Kruskal Wallis method, included 24 Phyla, 46 Classes, 90 Orders, 123 Families and 262 Genera of bacteria.

Beta diversity based on Weighted UniFrac distance metrics is represented in Fig 4.3. PCoA was used to visualize the differences in bacterial community composition among the samples. The PCoA shows a clustering according to sampling points 1 and 2, therefore suggests that the microbial communities differ between the two points. It is also observed that samples are grouped according to replicates and different sampling times (t1-t4). This was confirmed by the ANOSIM test, which confirmed that there were significant differences (p = 0.001) among the groups.

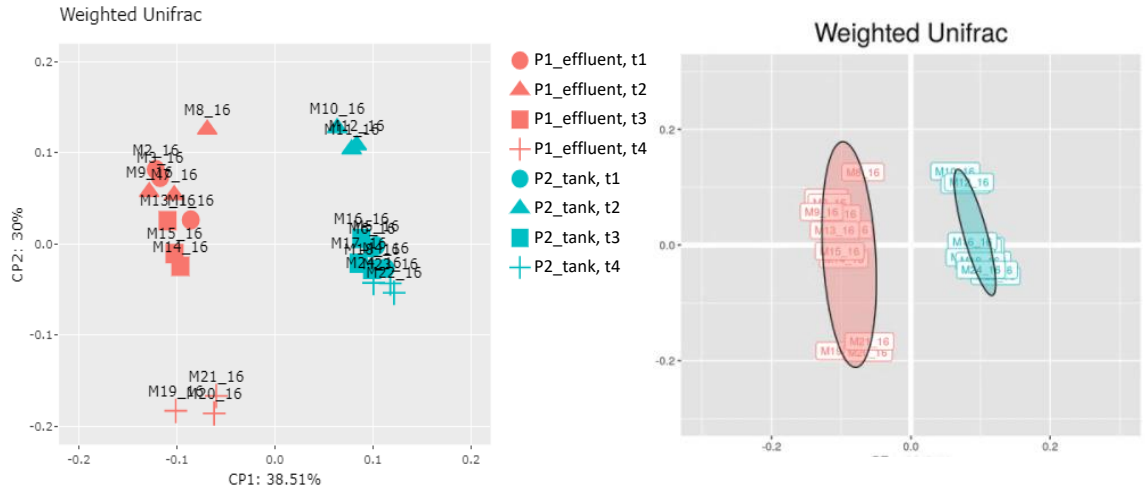


Figure 4.3. Two-dimensional principal coordinate analysis (PCoA) plots based on Weighted UniFrac distance matrices of the 24 reclaimed water samples.

4.1.3 IDENTIFICATION OF POTENTIAL PATHOGENS

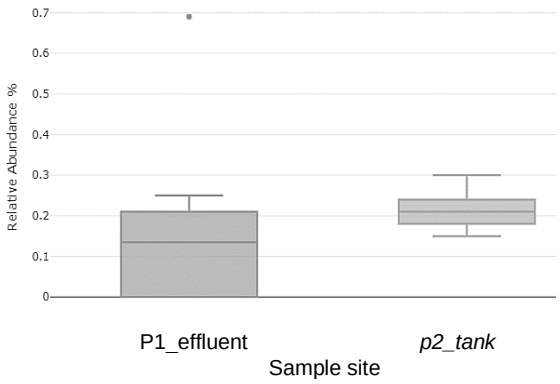
The Regulation (EU) 2020/741 of the European Parliament and of the Council of 25 May 2020 on minimum requirements for water reuse only requires the control of *Escherichia coli, Legionella spp, Clostridium perfringens* spores and intestinal nematodes (helminth eggs) or *Campylobacter, and Cryptosporidium* (EU Regulation 2021/472, 2021).

The presence of key microorganisms regulated by the Royal Decree 1620/2007 and the European Regulation 2020/741 was not found. This indicates that the treatment used in the WWTP is adequate and that the reclaimed water analysed complies with the legislation and is suitable for use (Royal Decree 1620/2007, BOE 2007; EU Regulation 2021/472, 2021).

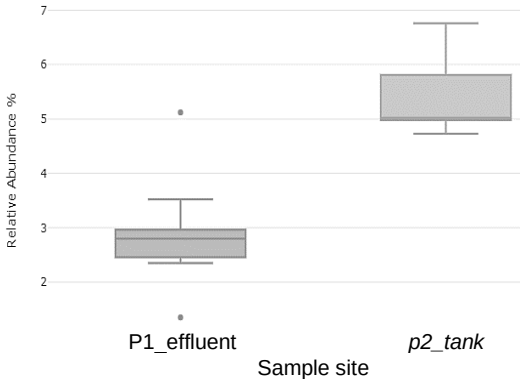
However, due to the use of this 16S method it was possible to detect the presence of all bacteria present in the sample. Potential pathogens that cannot be identified using specific media and are not taken into account in the European standards were discovered. Candidate pathogenic bacteria such as *Arcobacter, Bacteroides, Aeromonas, Acinetobacter* and *Comamonas* were detected in the analysis.

Different species of the genus *Arcobacter* are associated with human diseases and may pose a health threat if used to irrigate foods which are usually consumed uncooked, such as fruits and vegetables (Do *et al.,* 2019). *Aeromonas* is a genus widespread in aquatic environments, with a high sanitary risk because they are opportunistic human pathogens that can cause various infectious diseases (intestinal, blood or skin), as well as being able to develop and spread antibiotic resistance (Zdanowicz *et al.,* 2020). In addition, some species of the *Acinetobacter* genus have been also described as opportunistic human pathogens causing infections that are difficult to treat (Do *et al*., 2019). *Bacteroides* are fecal anaerobic microorganism used as indicators of water pollution, identifying human or animal fecal sources, can be potential pathogens (Yang *et al.,* 2019). *Comamonas* species have been isolated from soil, plants and water. They are not usually the cause of disease in humans. However, cases of bacteremia caused by this genus have been reported, due to their ability to grow in hospital devices such as mechanical respirators (breathing machine) (Tiwari & Nanda, 2019).

a) COMAMONAS



b) BACTEROIDES



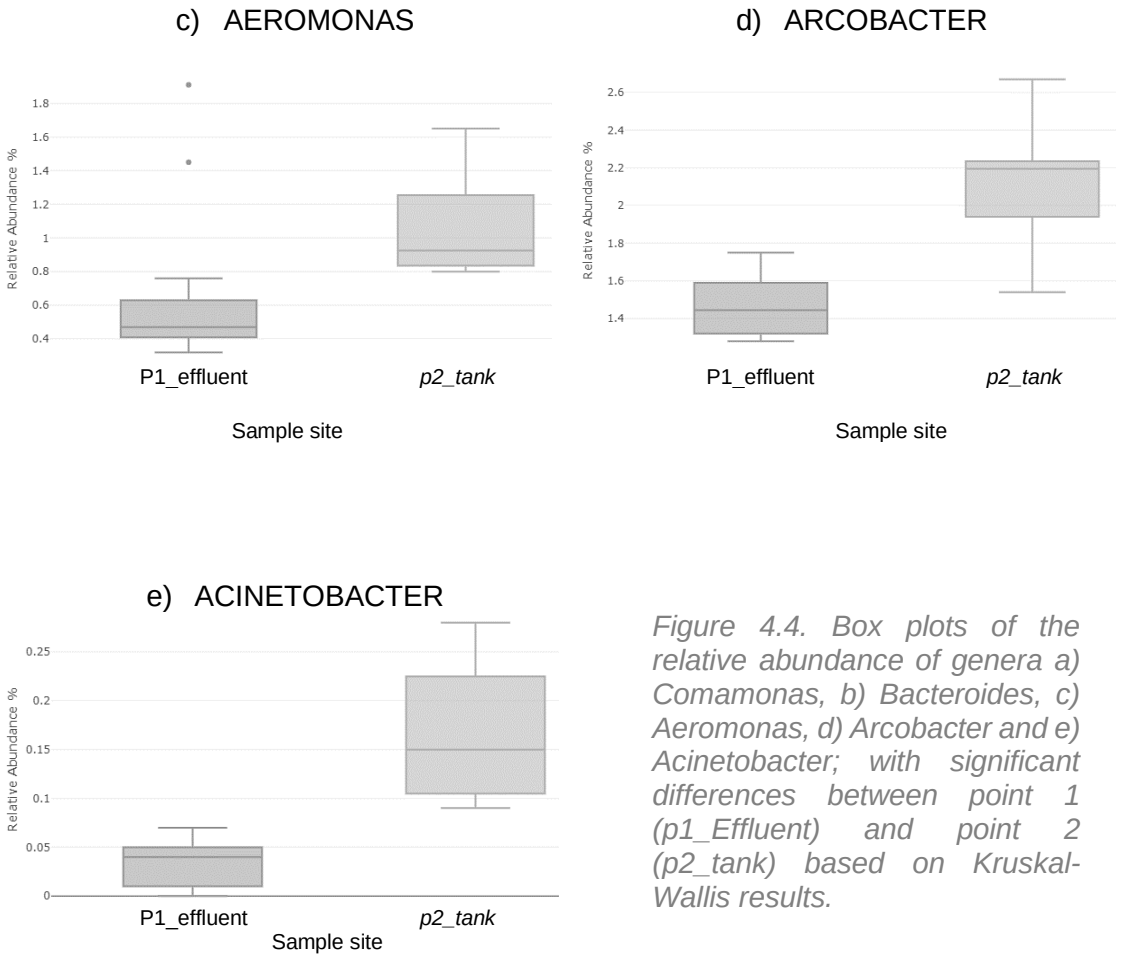


Figure 4.4. Box plots of the relative abundance of genera a) Comamonas, b) Bacteroides, c) Aeromonas, d) Arcobacter and e) Acinetobacter; with significant differences between point 1 (p1_Effluent) and point 2 (p2_tank) based on Kruskal-Wallis results.

These genera which include pathogenic species, have significantly higher relative abundances at sampling point 2. Boxplots of the relative abundance of taxa with significant differences based on Kruskal-Wallis results are shown below. (Fig 4.4.) These genera with pathogenic risk cannot growth in pure cultures but have been identified through metagenomics, which reaffirms that the use of massive 16S sequencing can help us to better characterise and diagnose the quality of reclaimed water.

4.2 SARS-CoV-2 SAMPLES

In this study, the ARTIC nCoV-2019 novel coronavirus bioinformatics protocol was used to analysed the different SARS-Cov-2 genomes present in 10 samples of wastewater.

The average read lengths (bases) of the samples sequenced by Nanopore was 349 bp. The number of processed reads were 319.160, and the total reads mapped to the SARS-CoV-2 Wuhan-Hu-1 reference sequence were 54.498 (17%), resulting in >20 X average coverage across 38,2 % of the genome.

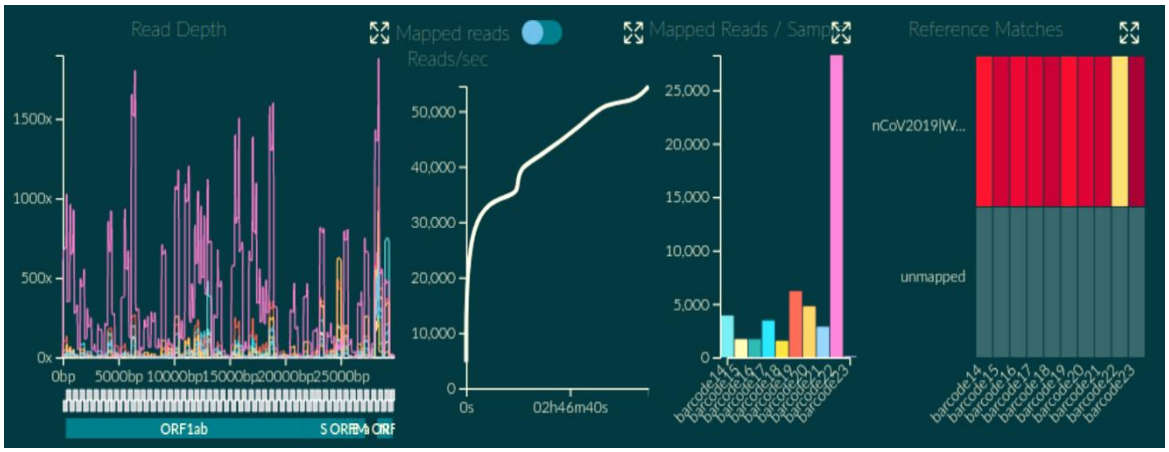


Figure 4.5. Visualization output of the ARTIC pipeline in RAMPART for the SARS-CoV-2 samples sequenced by Nanopore.

After applying all the steps of the bioinformatic analysis, we obtain for each sample different outputs files:

- **Barcode*.rg.primertrimmed.bam:** BAM file for visualization after primer-binding site trimming.
- **Barcode*.trimmed.bam:** BAM file with the primers left on. (Used in variant calling).
- **Barcode*.merged.vcf:** all detected variants in VCF format.
- **Barcode*.pass.vcf:** detected variants in VCF format passing quality filter.
- **Barcode*.fail.vcf:** detected variants in VCF format falling quality filter.
- **Barcode*.primers.vcf:** detected variants falling in primmer-binding regions.
- **Barcode*.consensus.fasta:**consensus sequence.

In wastewater samples there is not a single variant of SARS-CoV-2 as occurs in clinical isolates. In this matrix, different variants might co-occur and there may be a mixture of multiple variants in the same sample. Therefore, the analysis of SARS-Cov-2 sequences in sewage shows the diversity of variants present in the population of a defined geographical area (Izquierdo-Lara *et al*., 2021). To confirm the presence of a variant of SARS-Cov-2 in a wastewater sample, at least three signature mutations per variant should be reported (EU. Commission Recommendation 2021/472, 2021).

Since the samples analysed were not clinical isolates, but wastewater, the genetic material of which may be degraded and fragmented (Jahn *et al*., 2021). In order to identify the different SNPs, the quality filter was bypassed and the file used was the Barcode*.merged.vcf.

All the identified variants are displayed in the variant call format (VFC). The Barcode*.merged.vcf file contains different columns including: chromosome (CHROM) reference, position (POS) within the chromosome, the ID, the reference (REF) sequence, and the alternative (ALT) sequence, which is the SNP that identifies the different SARS-CoV-2 variants, a column with a Phred-scaled quality score assigned by the variant caller (QUAL), the FILTER column, if the SNPs pass (PASS) or not (dp) the quality filter, and an INFO or SAMPLE column, with additional information.

The position of the SNPs obtained in the output files with the positions of the specific mutations for the detection of SARS-Cov-2 variants has been compared. The position of the nucleotides identified where correlated with the amino acids and with the detected variant. All the signature mutations of the different samples (Barcode 14 to Barcode 23) are summarized in tables 4.3 to 4.12.

Table 4.3. Table of all signature SNPs in the sample Barcode14.

SAMPLE	POSITION	REFERENCE	ALTERNATIVE	VARIANT	WHO LABEL	AA
Barcode 14	3267	C	T	UK	ALPHA	T1001I
Barcode 14	5388	C	A	UK	ALPHA	A1708D
Barcode 14	23063	A	T	UK, BRASIL AND SOUTH AFRICA	ALPHA, GAMMA AND BETA	N501Y
Barcode 14	23271	C	A	UK	ALPHA	A570D
Barcode 14	24914	G	C	UK	ALPHA	D1118H
Barcode 14	27972	C	T	UK	ALPHA	Q27
Barcode 14	28111	A	G	UK	ALPHA	Y73C

Table 4.4. Table of all signature SNPs in the sample Barcode15.

SAMPLE	POSITION	REFERENCE	ALTERNATIVE	VARIANT	WHO LABEL	AA
Barcode 15	3267	C	T	UK	ALPHA	T1001I
Barcode 15	23271	C	A	UK	ALPHA	A570D
Barcode 15	24506	T	G	UK	ALPHA	S982A
Barcode 15	24914	G	C	UK	ALPHA	D1118H

Table 4.5. Table of all signature SNPs in the sample Barcode16.

SAMPLE	POSITION	REFERENCE	ALTERNATIVE	VARIANT	WHO LABEL	AA
Barcode 16	3267	C	T	UK	ALPHA	T1001I
Barcode 16	3828	C	T	BRASIL	GAMMA	S1188L
Barcode 16	5388	C	A	UK	ALPHA	A1708D
Barcode 16	23063	A	T	UK, BRASIL AND SOUTH AFRICA	ALPHA, GAMMA AND BETA	N501Y
Barcode 16	23271	C	A	UK	ALPHA	A570D
Barcode 16	24506	T	G	UK	ALPHA	S982A
Barcode 16	24914	G	C	UK	ALPHA	D1118H
Barcode 16	27972	C	T	UK	ALPHA	Q27
Barcode 16	28048	G	T	UK	ALPHA	R52I
Barcode 16	28111	A	G	UK	ALPHA	Y73C

Table 4.6. Table of all signature SNPs in the sample Barcode17.

SAMPLE	POSITION	REFERENCE	ALTERNATIVE	VARIANT	WHO LABEL	AA
Barcode 17	3267	C	T	UK	ALPHA	T1001I
Barcode 17	23271	C	A	UK	ALPHA	A570D
Barcode 17	23604	C	A	UK	ALPHA	P681H
Barcode 17	24506	T	G	UK	ALPHA	S982A
Barcode 17	24914	G	C	UK	ALPHA	D1118H
Barcode 17	28048	G	T	UK	ALPHA	R52I
Barcode 17	28111	A	G	UK	ALPHA	Y73C

Table 4.7. Table of all signature SNPs in the sample Barcode18.

SAMPLE	POSITION	REFERENCE	ALTERNATIVE	VARIANT	WHO LABEL	AA
Barcode 18	3267	C	T	UK	ALPHA	T1001I
Barcode 18	5388	C	A	UK	ALPHA	A1708D
Barcode 18	23271	C	A	UK	ALPHA	A570D
Barcode 18	23604	C	A	UK	ALPHA	P681H
Barcode 18	23709	C	T	UK	ALPHA	T716I
Barcode 18	24506	T	G	UK	ALPHA	S982A
Barcode 18	24914	G	C	UK	ALPHA	D1118H

Table 4.8. Table of all signature SNPs in the sample Barcode19.

SAMPLE	POSITION	REFERENCE	ALTERNATIVE	VARIANT	WHO LABEL	AA
Barcode 19	3267	C	T	UK	ALPHA	T1001I
Barcode 19	5388	C	A	UK	ALPHA	A1708D
Barcode 19	23063	A	T	UK , BRASIL AND SOUTH AFRICA	ALPHA, GAMMA AND BETA	N501Y
Barcode 19	23271	C	A	UK	ALPHA	A570D
Barcode 19	23709	C	T	UK	ALPHA	T716I
Barcode 19	24506	T	G	UK	ALPHA	S982A
Barcode 19	24914	G	C	UK	ALPHA	D1118H
Barcode 19	27972	C	T	UK	ALPHA	Q27
Barcode 19	28048	G	T	UK	ALPHA	R52I
Barcode 19	28111	A	G	UK	ALPHA	Y73C

Table 4.9. Table of all signature SNPs in the sample Barcode20.

SAMPLE	POSITION	REFERENCE	ALTERNATIVE	VARIANT	WHO LABEL	AA
Barcode 20	3267	C	T	UK	ALPHA	T1001I
Barcode 20	5388	C	A	UK	ALPHA	A1708D
Barcode 20	23063	A	T	UK , BRASIL AND SOUTH AFRICA	ALPHA, GAMMA AND BETA	N501Y
Barcode 20	23271	C	A	UK	ALPHA	A570D
Barcode 20	23604	C	A	UK	ALPHA	P681H
Barcode 20	23709	C	T	UK	ALPHA	T716I
Barcode 20	24914	G	C	UK	ALPHA	D1118H
Barcode 20	27972	C	T	UK	ALPHA	Q27
Barcode 20	28048	G	T	UK	ALPHA	R52I
Barcode 20	28111	A	G	UK	ALPHA	Y73C

Table 4.10. Table of all signature SNPs in the sample Barcode21.

SAMPLE	POSITION	REFERENCE	ALTERNATIVE	VARIANT	WHO LABEL	AA
Barcode 21	3267	C	T	UK	ALPHA	T1001I
Barcode 21	5388	C	A	UK	ALPHA	A1708D
Barcode 21	6954	T	C	UK	ALPHA	I2230T
Barcode 21	12778	C	T	BRASIL	GAMMA	-
Barcode 21	23271	C	A	UK	ALPHA	A570D
Barcode 21	23604	C	A	UK	ALPHA	P681H
Barcode 21	23709	C	T	UK	ALPHA	T716I
Barcode 21	24506	T	G	UK	ALPHA	S982A
Barcode 21	24914	G	C	UK	ALPHA	D1118H
Barcode 21	27972	C	T	UK	ALPHA	Q27
Barcode 21	28048	G	T	UK	ALPHA	R52I
Barcode 21	28111	A	G	UK	ALPHA	Y73C

Table 4.11. Table of all signature SNPs in the sample Barcode22.

SAMPLE	POSITION	REFERENCE	ALTERNATIVE	VARIANT	WHO LABEL	AA
Barcode 22	3828	C	T	BRASIL	GAMMA	S1188L
Barcode 22	5648	A	C	BRASIL	GAMMA	K1795Q
Barcode 22	22206	A	G	SOUTH AFRICA	BETA	D215G
Barcode 22	23012	G	A	NIGERIA, BRASIL AND SOUTH AFRICA	BETA AND GAMMA	E484K
Barcode 22	23063	A	T	UK, BRASIL AND SOUTH AFRICA	ALPHA, BETA AND GAMMA	N501Y
Barcode 22	23664	C	T	SOUTH AFRICA	BETA	A701V
Barcode 22	24914	G	C	UK	ALPHA	D1118H
Barcode 22	28048	G	T	UK	ALPHA	R52I
Barcode 22	28111	A	G	UK	ALPHA	Y73C

Table 4.12. Table of all signature SNPs in the sample Barcode13.

SAMPLE	POSITION	REFERENCE	ALTERNATIVE	VARIANT	WHO LABEL	AA
Barcode 23	-	-	-	-	-	-

In the analysed samples, 70 unique mutations for one SARS-CoV-2 variant across samples, and 6 mutations shared among different variants, have been detected. A total of 64 signature mutations were associated with the Alpha variant (B.1.1.7, United Kingdom), 4 were associated with the Gamma variant (P.1, Brazil), and 2 with the Beta (B.1.351, South Africa).

Five nucleotide variations shared between the Alpha, Beta and Gamma variants (Position 23,063, corresponding to the amino acid variation N501Y) were found. The N501Y is a signature mutation in these three variants of concern (VOCs), while one resulted shared between the Beta and Gamma (Position 23,012, amino acid E484K). The genetic marker that determines the variation E484K, is present in several variants, two of them VOCs (Beta and Gamma).

There were more than three specific mutations in the different samples, with which we can confirm that all residual water samples (Barcode 14 – Barcode 21) were positive for the Alpha variant (B1.1.7, United Kingdom), while the control sample (Barcode 22) was positive for Alpha (B1.1.7, United Kingdom), Beta (B.1.351, South Africa), and the Gamma (P.1, Brazil) variants.

The presence of the Alpha variant in most samples corresponds to the epidemiological situation in the Valencian Community at the time of sampling. At this time, only this variant was considered at high risk of transmission.

In March 2021, 54 cases of the Beta variant had been detected in Spain, of which only 18 were confirmed by sequencing. Two isolated cases of the P1 variant (not yet called Gamma) from Brazil, and three outbreaks with a total of 15 cases were detected, only 6 positive cases were confirmed by sequencing. It was recommended to monitor the epidemiological situation of the Beta, the P1 variant and other emerging variants in Spain.
(https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/20210304-EER.pdf).

Two specific SNPs for the Gamma variant (P1, Brazil) were detected in two of the wastewater samples. In Barcode 16, position 3,828, amino acid variation S1188L, and in Barcode 21, position 12,778. The detection of these mutations associated with the Gamma variant before it is considered as a VOC should focus future research in genetic surveillance.

5 CONCLUSION

This work has demonstrated that the use of sequencing approaches based on 16S rRNA amplicons together with the informatic analysis of the results could offer a reliable and detailed picture about the bacterial taxa composition in environmental samples, such as reclaimed water. With this study, it has been demonstrated that:

a) Genera which presence is regulated in both, European and Spanish legislation (*Escherichia, Legionella, Salmonella, Clostridium, Campylobacter* and *Cryptosporidium)* have not been detected. This confirms that the treatment carried out at the plant is adequate for the management of the WWTP and that it complies with the legal regulation.

b) Nevertheless, using this tool other genera are identified that are not in legislation but can be considered potentially pathogenic such as *Arcobacter, Bacteroides, Aeromonas, Acinetobacter*, *Comamonas* and phyla Patescibacteria or Nanoarchaeota has been revealed. These last two are classified as CPR, and their abundance in different environments was unknown until now because they could not be cultured, and their role could be important in water treatment in wastewater treatment plants.

These results could help in the management of the WWTPs to know the impact of the reclaimed water from a sanitary and environmental point of view. In order to make a complete study of the diversity of reclaimed water and to be able to establish the risk of using this water, it would be necessary to carry out an 18S study to extend the range of species to Eukaryotes.

The use of bioinformatics tools for the processing of high volume of sequences has been highlighted the power and ability of sampling and sequencing of wastewater samples to identify alternative genotypes and the different variants of SARS-Cov-2 present in a community.

We have been able to identify the different variants found in a population with a total of 76 SNPs, 70 unique mutations, and fulfilling the requirements of the European recommendations by reporting at least three signature mutations per variant. Detecting thanks to the pipeline used in this work the presence of the Alpha, Beta and Gamma variants of SARS-CoV-2 in wastewater samples.

In conclusion, this study underscores the importance of performing NGS analysis of wastewater to understand the diversity circulating in a community, not only of SARS-CoV-2, but of other viruses, which may cause more outbreaks in the future.

This work is a small probe of concept to demonstrate that the use of new technologies and bioinformatics tools could be applied not only for general knowledge about the populations of microorganisms in a sample to study the environmental impact but also as a robust environmental diagnostic tool in water samples (both reclaimed and wastewater). Nevertheless, a major effort is needed to simplify the procedures and reduce the time and cost of the analysis to have success in their application as routine analysis.

6 BIBLIOGRAPHY

Áine O'Toole, Emily Scher, Anthony Underwood, Ben Jackson, Verity Hill, John T McCrone, Rachel Colquhoun, Chris Ruis, Khalil Abu-Dahab, Ben Taylor, Corin Yeats, Louis Du Plessis, Daniel Maloney, Nathan Medd, Stephen W Attwood, David M Aanensen, Edward C Holmes, Oliver G Pybus, Andrew Rambaut (2021) Virus Evolution DOI:10.1093/ve/veab064

Almeida, A., Mitchell, A., Tarkowska, A. and Finn, R., 2018. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. GigaScience, 7(5).

Alves L de F, Westmann CA, Lovate GL, de Siqueira GMV, Borelli TC, Guazzaroni M-E. Metagenomic approaches for understanding new concepts in microbial science. Int J Genomics. 2018 Aug 23; 2018:2312987.

Balvočiūtė, M. and Huson, D., 2017. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? BMC Genomics, 18(S2).

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, and Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37: 852–857.https://doi.org/10.1038/s41587-019-0209-9

Bruno et all, 2017. Bruno, A., Sandionigi, A., Rizzi, E., Bernasconi, M., Vicario, S., Galimberti, A., et al. (2017). Exploring the under-investigated "microbial dark matter" of drinking water treatment plants. Sci Rep. 7:44350. doi: 10.1038/srep44350

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA and Holmes SP (2016). "DADA2: High-resolution sample inference from Illumina amplicon data." Nature Methods, 13, pp. 581-583. doi: 10.1038/nmeth.3869.

Callahan, B., McMurdie, P. and Holmes, S., 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal, 11(12), pp.2639-2643.

Cascella, M., Rajnik, M., Aleem, A., Dulebohn, S. C., & di Napoli, R. (2021). Features, Evaluation, and Treatment of Coronavirus (COVID-19).

Chu BTT, Petrovich ML, Chaudhary A, Wright D, Murphy B, Wells G, Poretsky R. Metagenomics Reveals the impact of wastewater treatment plants on the dispersal of microorganism and genes in aquatic sediments. Appl. Environ. Microbio. 2018; 84:1-15.

Dias, M., da Rocha Fernandes, G., Cristina de Paiva, M., Christina de Matos Salim, A., Santos, A. and Amaral Nascimento, A., 2020. Exploring the resistome, virulome and microbiome of drinking water in environmental and clinical settings. Water Research, 174, p.115630.

Do, T., Delaney, S. and Walsh, F., 2019. 16S rRNA gene based bacterial community structure of wastewater treatment plant effluents. FEMS Microbiology Letters, 366(3).

EU. Commission Recommendation (EU) 2021/472, Official Journal of the European Union. 17 March 2021. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32021H0472&from=EN (accessed on 1 July 2021)

European Parliament and European Council, 2020. Regulation (EU) 2020/741 of the European Parliament and of the Council of 25 May 2020 on minimum requirements for the reuse of water.

Farkas, K., Hillary, L., Malham, S., McDonald, J. and Jones, D., 2020. Wastewater and public health: the potential of wastewater surveillance for monitoring COVID-19. Current Opinion in Environmental Science & Health, 17, pp.14-20.

Farkas, K.; Hillary, L.S.; Thorpe, J.; Walker, D.I.; Lowther, J.A.; McDonald, J.E.; Malham, S.K.; Jones, D.L. Concentration and Quantification of SARS-CoV-2 RNA in Wastewater Using Polyethylene Glycol-Based Concentration and qRT-PCR. Methods Protoc. 2021, 4, 17. https://doi.org/10.3390/mps4010017.

Giovanetti, M., Benedetti, F., Campisi, G., Ciccozzi, A., Fabris, S., Ceccarelli, G., Tambone, V., Caruso, A., Angeletti, S., Zella, D. & Ciccozzi, M. 2021, "Evolution patterns of SARS-CoV-2: Snapshot on its genome variants", Biochemical and biophysical research communications, vol. 538, pp. 88-91.

Gomez-Alvarez V, Revetta RP, Santo Domingo JW. Metagenomic analyses of drinking water receiving different disinfection treatments. Appl. Environ. Microbiol. 2012;78(17):6095–6102.

Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. Microbiol. Mol. Biol. Rev. 2004; 68 (4): 669–685.

Herrmann M, Wegner C-E, Taubert M, Geesink P, Lehmann K, Yan L, Lehmann R, Totsche KU and Küsel K (2019) Predominance of Cand. Patescibacteria in Groundwater Is Caused by Their Preferential Mobilization From Soils and Flourishing Under Oligotrophic Conditions. Front. Microbiol. 10:1407. doi: 10.3389/fmicb.2019.01407

Hjelmsø, M., Hellmér, M., Fernandez-Cassi, X., Timoneda, N., Lukjancenko, O., Seidel, M., Elsässer, D., Aarestrup, F., Löfström, C., Bofill-Mas, S., Abril, J., Girones, R. and Schultz, A., 2017. Evaluation of Methods for the Concentration and Extraction of Viruses from Sewage in the Context of Metagenomic Sequencing. PLOS ONE, 12(1), p.e0170199.

Hong P-Y, Mantilla-Calderon D, Wang C. 2020. Metagenomics as a tool to monitor reclaimed-water quality. Appl Environ Microbiol 86:e00724-20. https://doi.org/10 .1128/AEM.00724-20

Hufsky, F., Lamkiewicz, K., Almeida, A., Aouacheria, A., Arighi, C., Bateman, A., Baumbach, J., Beerenwinkel, N., Brandt, C., Cacciabue, M., Chuguransky, S., Drechsel, O., Finn, R. D., Fritz, A., Fuchs, S., Hattab, G., Hauschild, A. C., Heider, D., Hoffmann, M., Hölzer, M., … Marz, M. (2021). Computational strategies to combat COVID-19: useful tools to accelerate SARS-CoV-2 and coronavirus research. Briefings in bioinformatics, 22(2), 642–663. https://doi.org/10.1093/bib/bbaa232.

Isaac, T. and Sherchan, S., 2020. Molecular detection of opportunistic premise plumbing pathogens in rural Louisiana's drinking water distribution system. Environmental Research, 181, p.108847.

Izquierdo-Lara, R., Elsinga, G., Heijnen, L., Munnink, B., Schapendonk, C., Nieuwenhuijse, D., Kon, M., Lu, L., Aarestrup, F. M., Lycett, S., Medema, G., Koopmans, M., & de Graaf, M. (2021).

Monitoring SARS-CoV-2 Circulation and Diversity through Community Wastewater Sequencing, the Netherlands and Belgium. Emerging infectious diseases, 27(5), 1405–1415. https://doi.org/10.3201/eid2705.204410

Jaffe, A.L., Castelle, C.J., Matheus Carnevali, P.B. et al. The rise of diversity in metabolic platforms across the Candidate Phyla Radiation. BMC Biol 18, 69 (2020). https://doi.org/10.1186/s12915-020-00804-5

Jahn, K., Dreifuss, D., Topolsky, I., Kull, A., Ganesanandamoorthy, P., Fernandez-Cassi, X., Bänziger, C., Devaux, A., Stachler, E., Caduff, L., Cariti, F., Corzón, A., Fuhrmann, L., Chen, C., Jablonski, K., Nadeau, S., Feldkamp, M., Beisel, C., Aquino, C., Stadler, T., Ort, C., Kohn, T., Julian, T. and Beerenwinkel, N., 2021. Detection and surveillance of SARS-CoV-2 genomic variants in wastewater.

Jarett, J.K., Nayfach, S., Podar, M. et al. Single-cell genomics of co-sorted Nanoarchaeota suggests novel putative host associations and diversification of proteins involved in symbiosis. Microbiome 6, 161 (2018). https://doi.org/10.1186/s40168-018-0539-8

Kori, J. A., Mahar, R. B., Vistro, M. R., Tariq, H., Khan, I. A., & Goel, R. (2019). Metagenomic analysis of drinking water samples collected from treatment plants of Hyderabad City and Mehran University Employees Cooperative Housing Society. Environmental science and pollution research international, 26(28), 29052–29064. https://doi.org/10.1007/s11356-019-05859-8

Li J, Cheng W, Xu L, Strong PJ, Chen H. Antibiotic-resistant genes and antibiotic-resistant bacteria in the effluent of urban residential areas, hospitals, and a municipal wastewater treatment plant system. Environ. Sci. pollut. Res. 2015; 22:4587-4596.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34:3094-3100. doi:10.1093/bioinformatics/bty191

Ludington WB, Seher TD, Applegate O, Li X, Kliegman JI, Langelier C, et al. (2017) Assessing biosynthetic potential of agricultural groundwater through metagenomic sequencing: A diverse anammox community dominates nitrate-rich groundwater. PLoS ONE 12(4): e0174930. https://doi.org/10.1371/journal.pone.0174930

Martin, J., 2020. How has the COVID-19 pandemic impacted PCR?. BioTechniques, 69(6), pp.404-405.

Martinez-Hernandez, F. et al. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. Nat. Commun. 8, 15892 doi: 10.1038/ncomms15892 (2017).

Matias Rodrigues JF, Schmidt TSB, Tackmann J, and von Mering C (2017) MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. Bioinformatics. http://doi.org/10.1093/bioinformatics/btx517

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2012;6(3):610–8.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J. and Edwards, R., 2008. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics, 9(1).

Moreno-Mesonero, L., Ferrús, M. and Moreno, Y., 2020. Determination of the bacterial microbiome of free-living amoebae isolated from wastewater by 16S rRNA amplicon-based sequencing. Environmental Research, 190, p.109987.

Mukesh Kumar Awasthi , B. Ravindran , Surendra Sarsaiya , Hongyu Chen , Steven Wainaina , Ekta Singh , Tao Liu , Sunil Kumar , Ashok Pandey , Lal Singh & Zengqiang Zhang (2020) Metagenomics for taxonomy profiling: tools and approaches, Bioengineered, 11:1,356-374, DOI: 10.1080/21655979.2020.1736238.

Mukesh Kumar Awasthi , B. Ravindran , Surendra Sarsaiya , Hongyu Chen , Steven Wainaina , Ekta Singh , Tao Liu , Sunil Kumar , Ashok Pandey , Lal Singh & Zengqiang Zhang (2020) Metagenomics for taxonomy profiling: tools and approaches, Bioengineered, 11:1,356-374, DOI: 10.1080/21655979.2020.1736238.

Numberger D, Ganzert L, Zoccarato L, Mühldorfer K, Sauer S, Grossart HP. Greenwood AD. Characterization of bacterial communities in wastewater with enhanced taxonomic resolution by full-length 16S rRNA sequencing. Nature Scientific Reports 2019; 9:9673.

Oliveira FS, Brestelli J, Cade S, et al.; MicrobiomeDB: A Systems Biology Platform for Integrating, Mining and Analyzing Microbiome Experiments. Nucleic Acids Research 2018.

Posada-Céspedes S., Seifert D., Topolsky I., Jablonski K., Metzner K.J., and Beerenwinkel N. 2021. "V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput sequencing data." *Bioinformatics*, January. doi:10.1093/bioinformatics/btab015.

Prodan A, Tremaroli V, Brolin H, Zwinderman AH, Nieuwdorp M, Levin E (2020) Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PLoS ONE 15(1): e0227434. https://doi.org/10.1371/journal. pone.0227434

R Core Team (2020). R: A language and environment for statistical computing. R foundation for Statistical Computing, Vienna; Austria. URL http://www.R-project.org/

Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, du Plessis L & Pybus OG (2020) Nature Microbiology DOI:10.1038/s41564-020-0770-5

Real decreto B.O.E. 294/2007: Real Decreto 1620/2007, de 7 de diciembre, por el que se establece el régimen jurídico de la reutilización de las aguas depuradas. Pp. 50639-50661.

Rizzatti, G., Lopetuso, L. R., Gibiino, G., Binda, C., & Gasbarrini, A. (2017). Proteobacteria: A Common Factor in Human Diseases. BioMed research international, 2017, 9351507. https://doi.org/10.1155/2017/9351507

Rompré, A., Servais, P., Baudart, J., de-Roubin, M. and Laurent, P., 2002. Detection and enumeration of coliforms in drinking water: current methods and emerging approaches. Journal of Microbiological Methods, 49(1), pp.31-54.

Rusiñol, M., Martínez-Puchol, S., Timoneda, N., Fernández-Cassi, X., Pérez-Cataluña, A., Fernández-Bravo, A., Moreno-Mesonero, L., Moreno, Y., Alonso, J. L., Figueras, M. J., Abril, J. F., Bofill-Mas, S., & Girones, R. (2020). Metagenomic analysis of viruses, bacteria and protozoa in irrigation water. International journal of hygiene and environmental health, 224, 113440. https://doi.org/10.1016/j.ijheh.2019.113440

Schloss, P.D., et al., Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol, 2009. 75(23):7537-41.

Simon, C. and Daniel, R., 2010. Metagenomic Analyses: Past and Future Trends. Applied and Environmental Microbiology, 77(4), pp.1153-1161.

Stüken A, Haverkamp THA. Metagenomic Sequences of Three Drinking Water and Two Shower Hose Biofilm Samples Treated with or without Copper-Silver Ionization. Microbiology Resource Announcements 2020 Jan 16;9(3).

Techtmann SM, Hazen TC. Metagenomic applications in environmental monitoring and bioremediation. J. Ind. Microbiol. Biotechnol. 2016 Aug 24;43(10):1345–1354.

Tiwari S, Nanda M. Bacteremia caused by Comamonas testosteroni an unusual pathogen. J Lab Physicians 2019;11:87-90.

Tran, H.N., Le, G.T., Nguyen, D.T., Juang, R.-S., Rinklebe, J., Bhatnagar, A., Lima, E.C., Iqbal, H.M.N., Sarmah, A.K, Chao, H.-P., SARS-CoV-2 coronavirus in water and wastewater: A critical review about presence and concern, Environmental Research, https://doi.org/10.1016/j.envres.2020.110265

Tyson J.R., James P., Stoddart, D., et al. (2020) Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. bioRxiv [Preprint] 2020.09.04.283077. doi: 10.1101/2020.09.04.283077.

Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73(16):5261–7. doi:10.1128/AEM.00062-07.

Weber N., et al. (2018) Nephele: a cloud platform for simplified, standardized and reproducible microbiome data analysis. Bioinformatics, 34(8): 1411-1413. https://doi.org/10.1093/bioinformatics/btx617.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Yang, K., Li, L., Wang, Y., Xue, S., Han, Y., & Liu, J. (2019). Airborne bacteria in a wastewater treatment plant: Emission characterization, source analysis and health risk assessment. Water research, 149, 596–606. https://doi.org/10.1016/j.watres.2018.11.027

Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucl. Acids Res. 42:D643-D648

Zdanowicz, M. Mudryk, Z.J. & Perlinski, P. Abundance and antibiotic resistance of Aeromonas isolated from the water of three carp ponds. Vet Res Commun 44,9-18 (2020). https://doi.org/10.1007/s11259-020-09768-x

Zhang, L., Loh, K., Lim, J. and Zhang, J., 2019. Bioinformatics analysis of metagenomics data of biogas-producing microbial communities in anaerobic digesters: A review. Renewable and Sustainable Energy Reviews, 100, pp.110-126.

7 APPENDIX

Table A1. Alpha diversity table with the observed Features in the different samples. *Respective sample ID, Sample.time, Replicate, Sample.site (collection point) and observed_features.*

SAMPLE ID	Sample.time	Replicate	Sample.site	observed_features
M1-16	t1	R1	p1_Effluent	760
M2-16	t1	R2	p1_Effluent	832
M3-16	t1	R3	p1_Effluent	715
M4-16	t1	R1	p2_tank	583
M5-16	t1	R2	p2_tank	561
M6-16	t1	R3	p2_tank	670
M7-16	t2	R1	p1_Effluent	575
M8-16	t2	R2	p1_Effluent	523
M9-16	t2	R3	p1_Effluent	500
M10-16	t2	R1	p2_tank	555
M11-16	t2	R2	p2_tank	482
M12-16	t2	R3	p2_tank	461
M13-16	t3	R1	p1_Effluent	536
M14-16	t3	R2	p1_Effluent	544
M15-16	t3	R3	p1_Effluent	656
M16-16	t3	R1	p2_tank	518
M17-16	t3	R2	p2_tank	543
M18-16	t3	R3	p2_tank	505
M19-16	t4	R1	p1_Effluent	588
M20-16	t4	R2	p1_Effluent	615
M21-16	t4	R3	p1_Effluent	648
M22-16	t4	R1	p2_tank	550
M23-16	t4	R2	p2_tank	610
M24-16	t4	R3	p2_tank	661

Table A2. Significant Phylum for the Kruskal Wallis method between sampling site one and two. *Respective, Var1: corresponds to sampling site 2 the reclaimed water tank, Var2: corresponds to sampling site 1 the effluent, the p value and the Taxa).*

Var1	Var2	p_value	Taxa
p2_Tank	p1_Effluent	0	d__Archaea;p__Euryarchaeota
p2_Tank	p1_Effluent	0	d__Archaea;p__Iainarchaeota
p2_Tank	p1_Effluent	0	d__Archaea;p__Nanoarchaeota
p2_Tank	p1_Effluent	0	d__Bacteria;__
p2_Tank	p1_Effluent	0	d__Bacteria;p__Bdellovibrionota
p2_Tank	p1_Effluent	0	d__Bacteria;p__Campilobacterota
p2_Tank	p1_Effluent	0	d__Bacteria;p__Elusimicrobiota
p2_Tank	p1_Effluent	0	d__Bacteria;p__Margulisbacteria
p2_Tank	p1_Effluent	0	d__Bacteria;p__Planctomycetota
p2_Tank	p1_Effluent	0	d__Bacteria;p__SAR324_clade(Marine_group_B)
p2_Tank	p1_Effluent	0	d__Bacteria;p__Verrucomicrobiota
p2_Tank	p1_Effluent	0,0001	d__Bacteria;p__Chloroflexi
p2_Tank	p1_Effluent	0,0001	d__Bacteria;p__WPS-2
p2_Tank	p1_Effluent	0,0004	d__Bacteria;p__Patescibacteria
p2_Tank	p1_Effluent	0,0014	d__Bacteria;p__Synergistota
p2_Tank	p1_Effluent	0,0112	d__Bacteria;p__LCP-89
p2_Tank	p1_Effluent	0,0117	d__Bacteria;p__Actinobacteriota
p2_Tank	p1_Effluent	0,014	d__Bacteria;p__Caldisericota
p2_Tank	p1_Effluent	0,0206	d__Bacteria;p__Fibrobacterota
p2_Tank	p1_Effluent	0,0206	d__Bacteria;p__Spirochaetota
p2_Tank	p1_Effluent	0,0227	d__Archaea;p__Thermoplasmatota
p2_Tank	p1_Effluent	0,0246	d__Bacteria;p__Bacteroidota
p2_Tank	p1_Effluent	0,0478	d__Archaea;p__Micrarchaeota