UNIVERSIDADE NOVA DE LISBOA

MScCBBi

MASTER IN
COMPUTATIONAL BIOLOGY
& BIOINFORMATICS

ANDRÉ FILIPE BRUNO DE OLIVEIRA
TOMÁS DOS SANTOS
BSc in Biology

Application and development of viral identification tools for metavirome analysis of wastewaters

September, 2024

iTQB nova
NOVA SCHOOL OF SCIENCE & TECHNOLOGY
NOVA MEDICAL SCHOOL
NOVA IMS Information Management School
DESDE 1902 INSTITUTO DE HIGIENE E MEDICINA TROPICAL UNIVERSIDADE NOVA DE LISBOA

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTAMENT OF
NAME OF THE DEPARTMENT

APPLICATION AND DEVELOPMENT OF VIRAL IDENTIFICATION TOOLS FOR METAVIROME ANALYSIS OF WASTEWATERS

André Filipe Bruno de Oliveira Tomás dos Santos

BSc in Biology

Adviser: Sofia Gonçalves Seabra
Assistant Researcher, IHMT, NOVA University Lisbon

Co-advisers: Ana Barroso Abecasis
Associate Professor, *IHMT,* NOVA University Lisbon

Ricardo Parreira
Associate Professor, *IHMT,* NOVA University Lisbon

Examination Committee:

Chair: Ana Rita Grosso

Rapporteurs:

Adviser:

Members:

MASTER IN COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

NOVA University Lisbon
September, 2024

ACKNOWLEDGMENTS

I am profoundly grateful to my supervisors, Dr. Sofia G. Seabra, Dr. Ana B. Abecasis, and Dr. Ricardo Parreira, whose expertise, guidance, and unwavering support have been invaluable throughout the development of this thesis. Their encouragement and insightful feedback have not only enhanced this work but also enriched my academic journey. Appreciation should also go towards Mónica Nunes from FCUL (Faculdade de Ciências da Universidade de Lisboa) for having shared the sequences from the projects AgriWWAter and VirusFree-Water, and their details so that we could analyze them.

I would also like to extend my heartfelt thanks to Victor Pimentel, Mafalda Miranda, Marta Pingarilho, and Inês Cruz, from IHMT, for their generous assistance, collaborative spirit, and the many enlightening discussions we shared. Their contributions have been instrumental in shaping this research. Additionally, INSa's Team, namely João Santos, Victor Borges and Daniel Sobral, which gave insightful comments to the analysis but also contributed in developing my enthusiasm in the field of metagenomics. I also thank Mónica Cunha for comments in early stage of my work.

I am grateful for the financial support provided by the EUCARE project, which funded my work while I completed my thesis. Being able to contribute to the school study—collecting samples, screening for positivity, and identifying variants—was an invaluable experience that complemented my academic research (Funded by the European Union´s Horizon Europe Research and Innovation Programme under Grant Agreement No 101046016). My sincere appreciation goes to the Instituto de Higiene e Medicina Tropical (IHMT) for providing an inspiring and resourceful environment at the Global Public Health Unit that fostered my growth and enabled the successful completion of this study.

Lastly, I am deeply indebted to my family and friends for their endless love, patience, and encouragement. Their unwavering belief in me has been a constant source of strength and motivation.

Funding was provided by:

- AgriWWAter: "Reclaimed wastewater for agriculture irrigation: mitigating microbiological hazards for a safer food production". Refa PTDC/CTA AMB/29586/2017 (Fundação para a Ciência e Tecnologia, Portugal).
- VirusFreeWater: "Reclaimed water for agricultural irrigation: overcoming the silent threat of waterborne viral infections". Refa 706, Projeto interno IBETXplore 2017.
- Internal exploratory Project GHTM- UID/04413/2020, WasteWaterVir: "Integrating metavirome analysis of wastewaters into tools for surveillance of infectious diseases" (Fundação para a Ciência e Tecnologia, Portugal).

"Music can change the world because it can change people." (Bono).

ABSTRACT

Wastewater surveillance has become an essential method for monitoring viral communities, particularly in understanding the spread of human pathogenic viruses and bacteriophages in the environment. This study aimed to assess the effectiveness and consistency of multiple bioinformatic pipelines in identifying viral communities from wastewater samples. Viral sequences were extracted from wastewater samples collected from two wastewater treatment plants (WWTPs) in Lisbon, Portugal, and analyzed using four bioinformatics tools: Genome Detective, CZ.ID, INSaFLU-TELEVIR, and Kraken2. The comparison focused on viral occurrence, abundance, alpha and beta diversity metrics, and the proportion of viral reads related to the overall sequencing effort.

Our results revealed significant variation between the tools in viral identification performance, with INSaFLU-TELEVIR and CZ.ID showing the highest rates of viral classification. Additionally, we explored the correlation between crAssphage, a marker for human fecal contamination, and other human pathogenic viruses. The findings highlighted the potential of wastewater metagenomics in identifying emerging and known pathogens, offering valuable insights into viral diversity and the reliability of various bioinformatic approaches. The study emphasizes the importance of tool selection and bioinformatics workflow in optimizing wastewater-based epidemiology for public health surveillance.

Keywords: wastewater, metagenomic analysis, environmental surveillance, next generation sequencing

RESUMO

A vigilância de águas residuais tornou-se um método essencial para monitorizar comunidades virais, especialmente na compreensão da propagação de vírus patogénicos humanos e bacteriófagos no ambiente. Este estudo teve como objetivo avaliar a eficácia e consistência de múltiplas pipelines bioinformáticas na identificação de comunidades virais a partir de amostras de águas residuais. As sequências virais foram extraídas de amostras de águas residuais coletadas em duas estações de tratamento de águas residuais (ETARs) em Lisboa, Portugal, e analisadas utilizando quatro ferramentas bioinformáticas: Genome Detective, CZ.ID, INSaFLU-TELEVIR e Kraken2. A comparação focou-se na ocorrência viral, abundância, métricas de diversidade alfa e beta, e a proporção de leituras virais em relação ao esforço total de sequenciamento.

Os resultados revelaram variação significativa entre as ferramentas no desempenho de identificação viral, com o INSaFLU-TELEVIR e CZ.ID apresentando as maiores taxas de classificação viral. Além disso, explorou-se a correlação entre crAssphage, um marcador de contaminação fecal humana, e outros vírus patogênicos humanos. Os resultados destacam o potencial da meta genómica de águas residuais na identificação de agentes patogénicos emergentes e conhecidos, oferecendo informações valiosas sobre a diversidade viral e a fiabilidade de várias abordagens bioinformáticas. O estudo realça a importância da seleção de ferramentas e do fluxo de trabalho bioinformático na otimização da epidemiologia baseada em águas residuais para a vigilância em saúde pública.

Palavras chave: águas residuais, análise metagenómica, vigilância ambiental, sequenciação de nova geração

CONTENTS

List of Figures

LIST OF TABLES

GLOSSARY

Metavirome The collective viral genetic material present in an environmental sample. This includes all viral genomes and genetic fragments – both from known viruses and previously unidentified ones – that can be extracted and analyzed from a specific habitat, such as soil, water or the human gut.

SISPA Sequence-Independent, Single-Primer Amplification (SISPA) is a molecular technique used to amplify viral nucleic acids without prior knowledge of their sequences. It is particularly valuable in metagenomic viral analysis for detecting and characterizing unknown or diverse viral genomes from environmental or clinical samples.

ACRONYMS

BLAST	Basic Local Alignment Tool
DNA	Deoxyribonucleic Acid
HAdV	Human adenovirus
HPyV	Human polyomavirus
ICTV	International Committee on Taxonomy of Viruses
IQR	Interquartile Range
LCA	Lowest Common Ancestor
RNA	Ribonucleic Acid
OTU	Operational Taxonomic Unit
PCoA	Principal Coordinates Analysis
PCR	Polymerase Chain Reaction
SISPA	Sequence-independent single primer amplification
WWTP	Wastewater treatment plant

1

INTRODUCTION

1.1 Viral Monitoring in Wastewater

Viral sequencing of wastewater is a vital tool in public health, allowing for the early detection of pathogenic viruses circulating within communities. This approach provides a non-invasive method to assess the occurrence of viruses and their diversity in human and animal populations, offering insights into the circulation of viruses without individual testing (Hellmér et al., 2014). By allowing the identification of both known and novel viruses, viral sequencing supports tracking outbreaks and understanding viral evolution. For example, it has been employed in monitoring viruses such as noroviruses, the hepatitis A virus, and polioviruses, underscoring its importance in public health strategies (Torres et al., 2016). The recent SARS-CoV-2 pandemic highlighted how wastewater surveillance could reflect viral dynamics within populations (Hasing et al., 2023). In addition to human pathogenic viruses, which were the main focus of this study, wastewaters are highly dominated by bacteriophages (phages) and thus, their host, bacteria (Ballesté et al., 2022), which play a key role in shaping microbial dynamics within wastewater ecosystems. Phages, such as coliphages that infect *Escherichia coli*, are commonly used as indicators of fecal contamination (Kelmer et al., 2023) as they are highly abundant in European and American populations (Morales-Cortés et al., 2024). One particularly abundant phage infecting *Bacteroides* sp., crAssphage, is consistently found in human gut metagenomes, making it a reliable marker for human-derived samples (Dutilh et al., 2014). The presence of crAssphage serves as a strong indicator of human fecal contamination and a quality control measure in metagenomic studies. Its robust detection in human gut samples suggests proper sample handling and sequencing accuracy, while its absence may signal issues in sample processing (Benler & Koonin, 2021 & Wu et al., 2020). In

addition to serving as an indicator of human fecal contamination, its presence can also be strongly correlated with other viruses. McKee & Cruz, 2021 reported an important correlation between the detection of crAssphage and some human viruses in a wastewater treatment process, concluding that crAssphage could be a performance indicator for viral reduction in the wastewater treatment process. In Sabar et al., 2022, crAssphage is described as being found in higher concentrations in sewage than other phages or enteric viruses, has a strong association with human viral pathogens like HPyV (human polyomavirus) and HAdV (human adenovirus) in wastewater, and exhibits greater resilience to environmental stress compared to traditional fecal indicators. With this knowledge, we ought to evaluate the presence of crAssphage in our samples as a mean to address both its presence across the various steps of each wastewater treatment plants and its correlation with some human viruses (also across the various steps of each wastewater treatment plant).

The metagenomic approach to wastewater surveillance provides a detailed view into the viral communities, distinct from targeted methods that are limited to specific known viruses. By analyzing genetic material extracted directly from environmental samples, metagenomics enables the detection of a diverse range of viruses, including novel and previously unrecognized ones, without requiring prior knowledge of their presence (Nooij et al., 2018). This contrasts with targeted approaches, which utilize specific primers or probes (e.g. via qPCR) to identify particular viruses, such as polyomaviruses (Torres et al., 2016; Condez et al., 2021), mastadenoviruses (Cavadas et al., 2022) or dengue virus (DENV) and chikungunya virus (CHIKV) (Monteiro et al., 2024). Other pertinent targets have been SARS-CoV-2, (Maryam et al., 2023), Polio, which has been detected in March of 2022 in New York's sewage water (Castilletti & Capobianchi, 2023) and monkeypox (Lapa et al., 2022).

The use of targeted approaches, although pertinent, can potentially overlook emerging or unusual/divergent viral agents. Concentrating on the metavirome, metagenomics facilitates the discovery of new viruses and enhances our understanding of viral diversity and behavior within wastewater ecosystems (Martínez-Puchol et al., 2020). Although accurately identifying and characterizing these viruses necessitates sophisticated bioinformatics tools and extensive viral databases (Kalantar et al., 2020). The integration of advanced computational techniques is therefore essential for improving the accuracy and effectiveness of metagenomic surveillance in tracking and studying viral populations in wastewater environments.

In metagenomic studies, especially those involving viral assemblages like in wastewater, biases can be introduced at various stages, namely in sample collection, sample

processing and extraction, viral enrichment, sequencing, and sequence analyses. This becomes crucial as the field shifts from descriptive to more quantitative analyses of microbial communities. A study (Parras-Moltó et al., 2018) evaluated biases induced by different virus enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses, observing that biases in amplification can significantly impact the abundance profiles of viruses. These biases are attributed to factors like the preferential amplification of certain genome types over others, with over-amplification of small circular genomes and under-amplification of genomes with extreme GC content being common issues, similar to those found with PCR-based sequence-independent, single-primer amplification (SISPA) methods. Further research (Rosseel et al., 2012) and (Quince et al., 2017) emphasized the challenges and biases associated with next-generation sequencing technologies and bioinformatic processing, highlighting the need for standardized protocols and an understanding of computational biases to improve the accuracy and reliability of microbial diversity assessments. These studies underscore the importance of conducting benchmark studies to understand the nature and impact of biases in metagenomic studies, ensuring more accurate and reliable analysis of viral communities in environmental samples.

1.2 Bioinformatic tools in metavirome analysis

Bioinformatics is essential for analyzing the complex datasets generated by metagenomic sequencing. It facilitates the assembly, annotation, and functional analysis of viral genomes, allowing researchers to uncover novel viruses and assess viral diversity in environmental samples (Luscombe et al., 2001; Edwards et al., 2015).

Metavirome analysis involves the study of viral communities directly from environmental samples using high-throughput sequencing technologies. This approach bypasses the need for culturing viruses, many of which are unculturable with current techniques. Bioinformatic tools play a pivotal role in interpreting the vast amount of data generated. The steps of data preprocessing and quality control, metagenomic assembly, and taxonomic classification are fundamental to ensuring accurate and meaningful results.

1.2.1 Data Preprocessing and Quality Control

In metavirome analysis, where the focus is on viruses—which often have high mutation rates and genetic diversity—starting with high-quality data is essential. Errors at this stage

can propagate through the analysis pipeline, compromising the entire study. Raw sequencing data often contains errors, low-quality reads, and sequencing artifacts (Zhou & Rokas, 2014). These can introduce biases and lead to incorrect conclusions if not addressed. Sequencing adapters and contaminants (like host DNA or other non-viral sequences) must be trimmed or filtered out. Their presence can interfere with downstream analyses like assembly and classification. As such, certain tools have been created to assess metrics such as sequence quality scores, GC content, and duplication levels (Schmieder & Edwards, 2011). This helps in identifying any anomalies or issues with the sequencing run. Ensuring high-quality data is crucial for reliable assembly and accurate taxonomic classification. Poor-quality data can result in fragmented assemblies and misclassification.

FastQC (Simon, 2010) is widely used to check the quality of raw sequencing data, offering visualizations to detect any problems in the dataset. Trimmomatic (Bolger et al., 2014) is often used for removing low-quality bases from sequencing reads, essential in viral metagenomic data analysis.

1.2.2 Metagenomic Assembly

Viral genomes are often underrepresented and highly diverse in environmental samples. Metagenomic assembly is challenging but crucial for piecing together these genomes. Successful assembly can reveal new viral species and provide insights into viral ecology and evolution. Assembly combines overlapping short sequencing reads into longer contigs or scaffolds, reconstructing viral genomes or genome fragments. Many viruses lack reference genomes, assembly allows for the identification of novel viruses by reconstructing their genomes from the data (Roux et al., 2015). Longer contigs are easier to annotate functionally and structurally, aiding in the understanding of viral genes and their potential roles. Assembled sequences reduce data complexity, making downstream analyses more manageable and computationally efficient.

Popular assembly tools help in reconstructing viral genomes from metavirome datasets include MEGAHIT (Li et al., 2015), an assembler for large metagenomic datasets, designed for high memory efficiency, and SPAdes (Bankevich et al., 2012), a versatile assembler also widely used in virome research, due to its flexibility in handling different datasets.

1.2.3 Taxonomic Classification

Viruses lack universal genetic markers, making taxonomic classification challenging. Specialized bioinformatic tools and databases are used to match sequences to known viruses

or predict taxonomy based on genetic similarities. Accurate classification is essential for interpreting the ecological roles and potential impacts of viral populations in the environment (Paez-Espino et al., 2016). The identification of viral species is achieved by assigning assembled sequences to taxonomic groups. Taxonomic classification helps quantifying the diversity and relative abundance of viruses, shedding light on community structure, allowing researchers to study viral-host interactions, transmission pathways, and evolutionary relationships (Edwards & Rohwer, 2005). Additionally, it can be used for comparative analyses, as it enables comparisons between different samples, environments, or conditions to understand factors influencing viral communities.

Taxonomic identification is a critical step in metavirome analysis, where the goal is to determine the types of viruses present in an environmental sample. Two primary categories of bioinformatic tools are used for this purpose: k-mer-based methods and alignment-based methods (McIntyre et al., 2017). Understanding the differences between these approaches is essential for selecting the appropriate tool for a given analysis.

A k-mer is a substring of length 'k' derived from a longer DNA or RNA sequence. For example, the sequence "ATCG" has 2-mers "AT," "TC," and "CG." The k-mer-based tools preprocess a reference database by breaking down known sequences into k-mers and storing them in an indexed format (Wood et al., 2019). The input sequences (reads or contigs) are also broken down into k-mers. The tool then matches these k-mers against the indexed database to find the best taxonomic assignment based on shared k-mers (Moeckel et al., 2024). Examples of taxonomic classification tools which use the k-mer method include: Kraken (Wood et al., 2019), which builds a database of k-mers associated with taxonomic labels and uses exact matches for classification, CLARK (Ounit et al., 2015) which uses discriminative k-mers that are unique to specific taxa for fast and accurate classification or Centrifuge (Kim et al., 2016), an optimized tool that can handle large databases efficiently by compressing the index. These methods are generally faster than alignment-based methods because they rely on exact or near-exact k-mer matching rather than computationally intensive alignments making them suitable for handling large metagenomic datasets due to their speed and lower computational requirements as they optimize memory usage through compressed indices (Ren et al., 2017). On the other hand, short reads may not contain enough informative k-mers, reducing classification accuracy. At the same time, the fact that they rely on existing k-mer representations in the reference database makes it challenging to classify novel or highly divergent sequences. Another limitation can occur when samples with high diversity or closely related organisms, k-mer overlaps can lead to misclassifications (Vázquez-Castellanos et al., 2014).

The alignment-based identification tools perform sequence alignments between the input sequences and reference databases using algorithms like BLAST (Basic Local Alignment

Search Tool). The alignments are scored based on similarity, and statistical measures (like E-values) assess the significance of the matches. The best matches are used to assign taxonomy, often considering multiple hits to account for ambiguity. Tools which employ this method are BLAST (Altschul, 1997), a widely used and known tool for finding regions of local similarity between sequences, MEGAN (MEtaGenome ANalyzer) (Huson et al., 2007), which takes the BLAST output to assign taxonomic labels based on the LCA (Lowest Common Ancestor) algorithm, or DIAMOND (Buchfink et al., 2014), an alignment tool similar to BLAST but optimized for speed, suitable for large datasets. Alignment-based methods are generally more accurate because they consider the entire sequence context, including insertions, deletions, and substitutions. They can identify homologous relationships even when sequences are highly divergent, aiding in the detection of novel viruses (Buchfink et al., 2014). They are also flexible, meaning there is the possibility to handle varying read lengths and are less sensitive to errors in the sequences (Altschul, 1997).

The limitations of alignment algorithms are the fact that they are resource-intensive, requiring more processing time and memory, meaning scalability issues, as alignment algorithms may not be practical for very large metagenomic datasets without significant computational resources (Wood et al., 2019).

1.3 Objectives

In this study we used metavirome sequences from wastewater samples processed in triplicate and in pools, to assess the consistency of several bioinformatic analysis pipelines/workflows in characterizing their viral taxonomic composition. We specifically aimed: 1) to compare the viral occurrence and abundance obtained by each software pipeline, analyzing alpha and beta diversity; 2) to analyze the impact of sequencing effort (number of raw reads) on abundance of viral reads; 3) to analyze the correlation between the fecal marker crAssphage and human pathogenic viruses.

2

METHODS

The wastewater collection, laboratory and sequencing procedures were previously carried out on the projects *AgriWWAter*: "Reclaimed wastewater for agriculture irrigation: mitigating microbiological hazards for a safer food production" (PTDC/CTA AMB/29586/2017, FCT, Portugal) and *VirusFreeWater*: "Reclaimed water for agricultural irrigation: overcoming the silent threat of waterborne viral infections" (ref. 706, Internal project IBETXplore 2017).

The bioinformatic and statistical analyses were carried out during this dissertation work.

2.1 Wastewater collection and processing, and nucleic acid extraction

From each of two Wastewater Treatment Plants (WWTP), located in the Lisbon metropolitan region (Portugal), three replicate samples were collected at four different sites of the treatment plant (Table 2.1).

Table 2.1 - Characteristics of the analyzed wastewater treatment stations. UV – Ultraviolet. Adapted from Rodrigues, 2015

WWTP	Population size	Type of residual water	Discharge points	Average flow (m^3/day)	Type of treatment	Biological treatment
1	756 000	Domestic	Rio Tejo	140 000	Tertiary with UV	Biofiltration
2	225 000	Domestic and industrial	Rio Tejo	50 000	Tertiary with UV	Activated Sludge

Both the samples from WWTP1 and WWTP2 were sampled in April 2019. The volumes of water collected ranged from 1 to 10 liters and were stored in sterilized 20 L containers for transportation (Table 2.1). Different volumes of wastewater were collected at each site due to the different characteristics of the collection sites. The samples were immediately taken to the laboratory and kept at 4 °C until processing, done at a maximum of 12h after collection. Concentration of viral-like particles and RNA extraction were done using the methods described

in Appendix A.1. Briefly, viral-like particles were concentrated using organic flocculation, sedimentation, centrifugation of the sediment and resuspension. RNA extraction was done using QIAamp® Viral RNA Mini Kit (QIAGEN®, USA), following the manufacturer's protocol. RNA quantification was performed using NanoDrop One.

2.2 Sequence-independent single-primer amplification (SISPA) and sequencing

The nucleic acid extracts were randomly amplified using the SISPA protocol (Fernandez-Cassi et al., 2018). SISPA is a primer-initiated technique designed for the non-selective logarithmic amplification of heterogeneous DNA populations. The method involves the directional ligation of an asymmetric linker/primer oligonucleotide onto blunt-ended DNA molecules. This common end sequence allows one strand of the linker/primer to be used repeatedly in rounds of annealing, extension, and denaturation, facilitated by thermostable Taq DNA polymerase. SISPA is particularly useful for amplifying and recovering low-abundance genetic sequences, such as those from uncharacterized viral genomes. This was followed by a viral enrichment step with VirCapSeq-VERT capture Panel (Roche, Switzerland). All replicate samples were pooled prior to library preparation. Libraries were prepared for each replicate separately (2 WWTP x 4 sites x 3 replicates) and for pools of replicates (2 WWTP * 4 sites), in a total of 32 samples. The pooled samples were all standardized to a concentration of 50 ng/µl. Sequencing was performed at "Laboratório de Vírus Contaminantes de Água e Alimentos do Departamento de Genética, Microbiologia e Estatística da Universidade de Barcelona" using an Illumina MiSeq (2 × 300 bp) device, producing paired-end reads.

2.3 Bioinformatic analysis

Each triplicate sample (and respective pool) for each of the two WWTP and each of four sites (total of 32 sequenced samples analyzed), was submitted to four different pipelines for viral classification/assignment: Genome Detective (Vilsker et al., 2018), CZ.ID (Kalantar et al., 2020), the TELEVIR module of INSaFLU (INSaFLU-TELEVIR) (Borges et al., 2018) and Kraken2 (Wood et al., 2019), ran on Galaxy Community Hub (Afgan et al., 2022). Each pipeline comprised a different set of tools for each analysis step as well as a variable number of steps in its workflow (Figure 2.1).

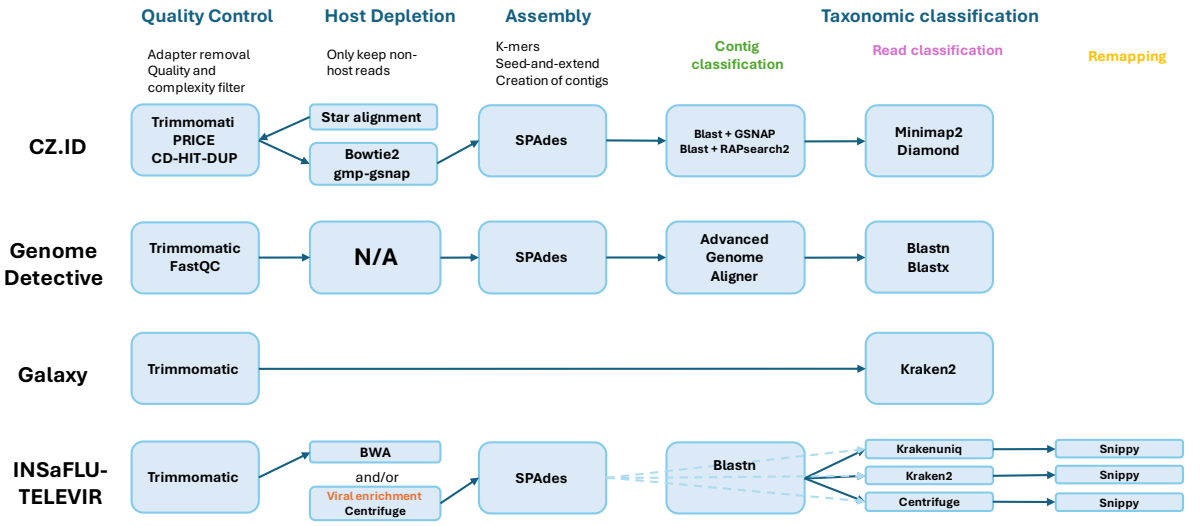

Figure 2.1 - Schematic representation of the workflows of each software. The categories were arbitrarily summarized into four crucial steps: quality control, host depletion, assembly and taxonomic classification. CZ.ID (Kalantar et al., 2020) https://czid.org/, Genome Detective, (Vilsker et al., 2018), https://www.genomedetective.com/app/typingtool/virus/, and Kraken2 (Wood et al., 2019), TELEVIR module of INSaFLU (Borges et al., 2018) https://insaflu.insa.pt/, ran on Galaxy Community Hub, (Afgan et al., 2022), https://usegalaxy.eu/.

CZ.ID initially validates the input files and truncates to 75 million fragments for single-end libraries and to 150 million reads for paired-end libraries so it can then perform an initial host depletion step using STAR alignment (Dobin et al., 2012) to a host-specific database (STAR, RRID:SCR 015899). Following host depletion, there is a plethora of quality control steps, which begin with adapter trimming using Trimmomatic (Bolger et al., 2014), removal of duplicates and low quality and complexity reads using PRICE (Ruby et al., 2013) and the CD-HIT-DUP tool v4.6.8 (Fu et al., 2012). Afterwards, host depletion is done with Bowtie2 (Langmead & Salzberg, 2012), which uses the HG38 reference database, and gmpa-gsnap (Wu & Watanabe, 2005; Wu & Nacu, 2010), which employs a more specific database containing HG38 and chimpanzee sequences. The remaining reads are then fed into an assembly-based alignment where short reads are aligned to the NCBI nucleotide (nt) and non-redundant protein (nr) databases using GSNAPL (Wu & Nacu, 2010) and RAPsearch2 (Zhao et al., 2011). At the same time, a database of pertinent accessions are assigned to each read, using NCBI accession2taxid database and BLAST+ (v2.6.0), in order to have a more stringent comparison (which also makes it faster). Short reads are then assembled into contigs with SPAdes (Bankevich et al., 2012) and raw reads are re-mapped into the resultant contigs with Bowtie2 in order to discover which contig each read belongs to, so that the contigs can then be aligned to the previously generated smaller database.

Genome Detective performs an initial quality control step of low-quality read filtering and adapter trimming using Trimmomatic (Bolger et al., 2014) and FastQC (Simon, 2010), after which uses DIAMOND (Buchfink et al., 2014), a protein-based alignment method used to identify possible viral reads, using as reference databases the latest viral subset of Swissprot UniRef90 protein database and NCBI RefSeq database (our study was performed between October and December of 2023). These potential short viral reads are sorted into groups and each group is submitted to distinct *de novo* assembly step using metaSPAdes (Bankevich et al., 2012). The result is then run on Blastn (Altschul et al., 1990) and Blastx (Altschul, 1997) in order to search for candidate reference sequences in the latest NCBI RefSeq virus database, combining the results for nucleotide and amino-acid level. Genome Detective then groups the results obtained and chooses the five best scoring references for each contig to be used for alignment, which is done using Advanced Genome Aligner (Deforche, 2017).

In INSaFLU-TELEVIR, quality control is performed using Trimmomatic (Bolger et al., 2014) for quality and read length filtering, and removing sequence adapters. An extra filtering step which removes repeated regions is done using Prinseq++ (Cantu et al., 2019). This is followed by viral enrichment step, in which only potential viral leads are kept, based on rapid classification against a viral sequence database using Centrifuge (Kim et al., 2016) coupled with host depletion using BWA (Li & Durbin, 2010). *De novo* assembly is done using SPAdes (Bankevich et al., 2012) and read classification is done using one of three possible options: Krakenuniq (Breitwieser et al., 2018), Kraken2 (Wood et al., 2019) or Centrifuge (Kim et al., 2016). Contigs classification, if any contigs are assembled, are classified using Blastn (Altschul et al., 1990). After the identification of viral sequences, a confirmatory mapping step is done, which takes the viral identifications and maps them to reference viral genomes from NCBI Taxonomy. Finally, Snippy (Tseemann, 2020) is used for confirmatory remapping of the viral reads and contigs against representative genome sequences identified in the previous step. Databases used are NCBI's Refseq Viral Genomes (viral.genome.fna.gz) and viro-saurus90_vertebrate-20200330. As we were interested in the results for the classifiers of each software pipeline, for a fair comparison, the results considered for INSaFLU-TELEVIR were the ones obtained in the intermediate report (the results prior to the confirmatory remapping step), as none of the other software pipelines contained this step and it narrows the results substantially.

Galaxy Community Hub was used to run Trimmomatic (Bolger et al., 2014) for adapter trimming and then read classification was done using Kraken2 (Wood et al., 2019) with the pre-built RefSeq Viral (version 2022-06-07 Downloaded 2022-08-04T105935Z), confidence level 0.1 and minimum hit groups 2. Kraken2 uses a k-mer-based algorithm for taxonomic classification.

2.4 Automating the analyses of viral identification reports

The processing and analyses of reports from the four different software programs used for viral identification in metagenomic samples, was automated using a customized python script (Figure 2.2). It begins by reading each fastq.gz file, for each of the 32 samples in order to count the number of reads. Then, it proceeds to read all the reports from the different software pipelines (examples in Tables 2.2, 2.3, 2.4 and 2.5) generated and stores the number of reads, the name of the organism and the associated taxID in a raw data frame (**df_taxid**).

Table 2.2 - Example report structure for CZ.ID

tax_id	name	category	is_phage	nt_rpm	nt_count	nt_%_identity	nt_alignment_length
2803850	Mammaliicoccus	bacteria	False	9.97596	2	93.548	62.0
643214	Mammaliicoccus stepanovicii	bacteria	False	9.97596	2	93.548	62.0
2782231	Kaistella	bacteria	False	144.65139	29	96.3326	115.96
2487074	Kaistella daneshvariae	bacteria	False	19.95192	4	80.916	131.0
421525	Chryseobacterium haifense	bacteria	False	114.72352	23	99.9623	112.053
266749	Chryseobacterium jeonii	bacteria	False	9.97596	2	92.683	123.0
2782229	Epilithonimonas	bacteria	False	214.48310	43	98.2307	89.9565
2487072	Epilithonimonas vandammei	bacteria	False	204.50714	41	98.1551	88.7619
1241981	Chryseobacterium lactis	bacteria	False	9.97596	2	99.0245	102.5
2782228	Planobacterium	bacteria	False	49.87979	10	95.9287	176.0

Table 2.3 - Example report structure for Genome Detective.

StrainName	NbReads	DepthOfCoverage	NtIdentity	AaIdentity	GenomeCoverage%	TaxonomyId
Punavirus P1	65	21.6267	0.9836	0.0000	0.3871	10678
Mamastrovirus 1	20	3.2944	0.9305	0.9425	10.1304	1239565
Norovirus GII	12	10.1633	0.8980	0.9697	1.3035	122929
Acinetobacter virus Acj61	9	3.7531	0.9383	0.9444	0.1481	2843629
Punavirus RCS47	8	5.1452	0.7823	0.8780	0.1077	2560452
Protoparvovirus rodent1	2	1.9011	0.6423	0.5862	5.1078	3052717
Cosavirus A	2	1.4844	0.9162	0.9531	2.5157	1330491
Aichivirus A	2	1.7619	0.9127	0.9286	1.5271	72149
Beihai picorna-like virus 99	2	1.8287	0.6434	0.6526	4.3980	1922644

Table 2.4 - Example report structure for INSaFLU–TELEVIR.

taxid	counts	description	accid
1239565	2	Mamastrovirus 1 genomic RNA, complete genome, isolate: V1347	KC342249.1
1868658	2	Human astrovirus, complete genome	NC_001943.1
2607629	2041	Physalis rugose mosaic virus isolate Piracicaba, complete genome	NC_055577.1
1239565	12	Mamastrovirus 1 genomic RNA, complete genome, isolate: V1347	KC342249.1
490039	6	Norovirus GII.2 strain Env/CHN/2016/GII.P16-GII.2/BJSMQ, complete genome	MF405169.1
122929	6	Norovirus GII.P7_GII.6, complete genome	KU935739.1
2025359	4	Murmansk poxvirus strain LEIV-11411, complete genome	NC_035468.1
1661062	4	Sclerotinia sclerotiorum fusarivirus 1 strain JMTJ14, complete genome	NC_027208.1
455364	4	Chrysochromulina ericina virus isolate CeV-01B, complete genome	NC_028094.1
1262528	3	Halovirus HVTV-1, complete genome	NC_020158.1
1922499	2	Beihai paphia shell virus 4 strain BWBFG40859 hypothetical protein gene, complete cds	NC_032495.1
2175118	2	Symphysodon discus adomavirus 1 isolate UFL1, complete genome	NC_040698.1
2107709	2	Pandoravirus quercus, complete genome	NC_037667.1
1349409	2	Pandoravirus dulcis, complete genome	NC_021858.1

Table 2.5 - Example report structure for Kraken2

% reads of clade rooted at taxon	Nb reads Clade	Nb reads Taxon	Taxonomic Rank	TaxID	Scientific Name
99.83	24091	24091	U	0	unclassified
0.17	41	0	R	1	root
0.17	41	0	D	10239	Viruses
0.13	32	0	D1	2731341	Duplodnaviria
0.13	32	0	K	2731360	Heunggongvirae
0.13	32	0	P	2731618	Uroviricota
0.13	32	0	C	2731619	Caudoviricetes
0.13	32	0	O	28883	Caudovirales
0.13	32	0	F	10662	Myoviridae
0.12	29	0	G	186789	Punavirus
0.12	29	0	S	10678	Punavirus
0.12	29	29	S1	2886926	Escherichia
0.01	3	0	F1	2842519	Twarogvirinae
0.01	3	0	G	2842819	Lasallevirus

The next step was handling missing data, in which any identifications with 0 reads were filtered out. These happened in Kraken2 due to the way the reports are read, but also in CZ.ID with no explanation given. Then, we retrieved the information about the complete

taxonomy for each viral identification to be used in the following steps. We used the latest ICTV (International Committee on Taxonomy of Viruses) Current Virus Metadata Resource, which is automatically downloaded and integrated into the analysis for up-to-date viral taxonomy, obtained from Virus Metadata Resource (VMR) (https://ictv.global/msl/current). The ICTV taxonomy was also used to create a directed graph which linked each viral taxonomy entry to their host source. This was implemented in a way that allows for taxonomic levels to have gaps. The most specific known taxonomic level was the one used for the assignment of the host source (e.g. for viruses allocated to a kingdom, phylum, class and genus, but for which no order or family is defined, their assigned genus would be the taxonomic level used for the host source search). Since not all viruses were fully classified in terms of their taxonomy, an additional column was created to assign the level of classification: "full" for complete taxonomic assignment, "partial" for cases in which one or more taxonomic levels were missing, and "undefined" if every taxonomic level was undefined (e.g. it was possible for an identification to appear as "uncultured phage" in the NCBI taxonomy database).

The taxID associated with each identification was used to obtain the complete taxonomy using the ETE toolkit, which is a Python library that assists in the analysis, manipulation and visualization of phylogenetic trees. It contains a module called NCBITaxa which is made to efficiently query a local copy of the NCBI Taxonomy database. The class NCBITaxonomy offers methods to convert from taxID to names (and vice versa), to fetch pruned topologies connecting a given set of species, or to download rank, names and lineage track information. As such, taxonomic assignment from kingdom to species level was added to each identification. It is important to note that identifications could stop at a certain taxonomic level, such as phylum, class, order or family, leading to incomplete taxonomic information. This happens because databases often have sequences published, that can be used as a reference for the mapping of reads, which do not yield complete taxonomic classification. Thus, the taxonomic assignment was added to the most specific level possible (sometimes up to species or genus level, sometimes only up to class, order or family).

Furthermore, we incorporated a field about the pathogenic potential for each viral identification (pathogen flag). CZ.ID provides a list of pathogens recognized for their ability to cause disease in immunocompetent individuals, which originated from a selection of eleven papers (https://czid.org/pathogen_list). Since this list was not parsable, we scanned all the reports for all our samples obtained from CZ.ID to extract every organism name which was flagged as a known pathogen. We then created a list of genus names potentially pathogenic to

humans. Taxa identified in our analyses belonging to any genera present in this list, were flagged as potentially pathogenic to humans. All this data was added to the raw data frame of taxa identifications (**taxid_dataframe**) (Figure 2.2).

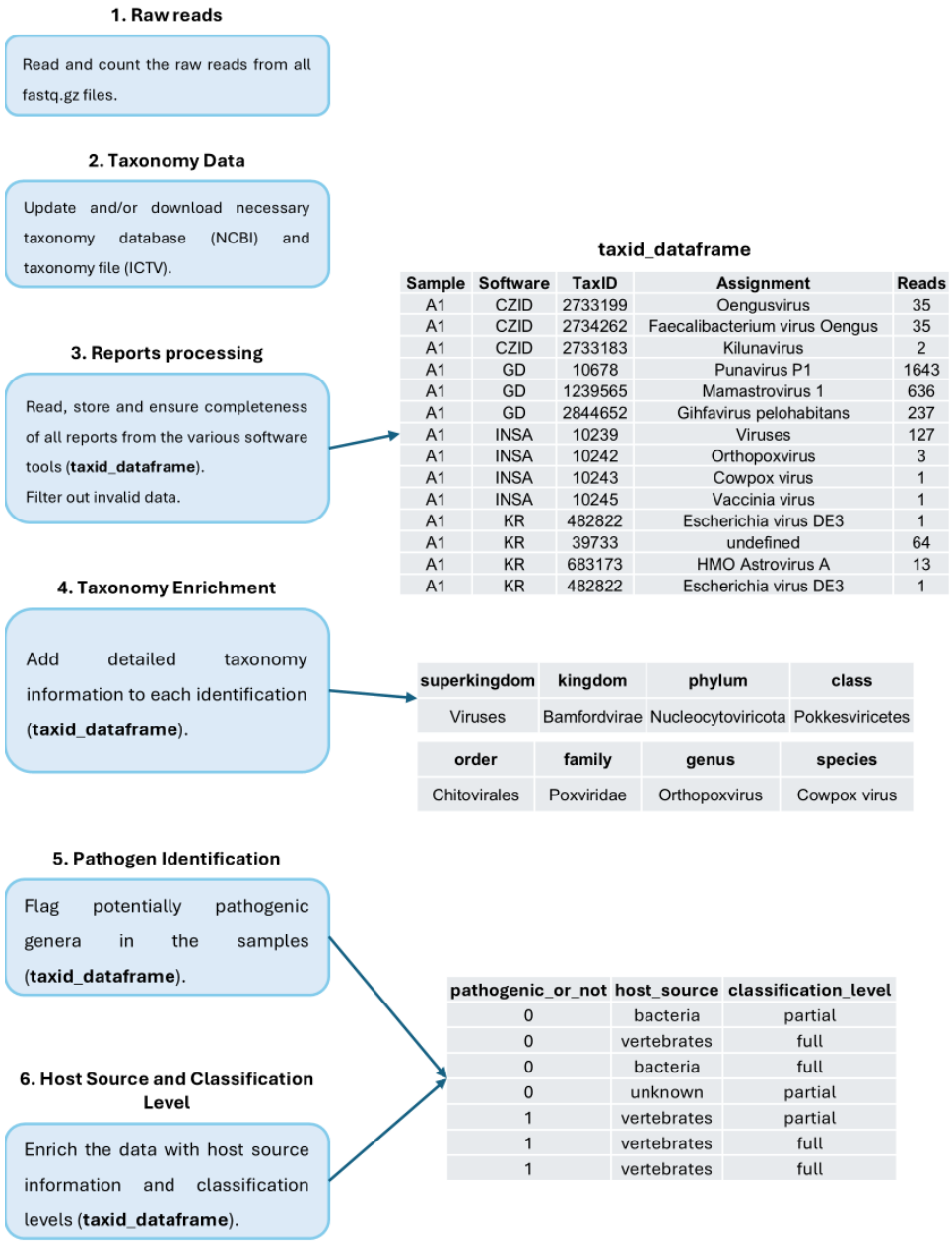

Figure 2.2 - Fluxogram of the script

A new dataframe (**table_samples**) was created with the summary data of the counts of reads of each type for each wastewater sample and each software pipeline as well as the number of raw reads for each sample. Information about each sample, such as the treatment plant

number, the collection step, library concentration was retrieved from the sample metadata file and added to this data frame (Table 2.6).

Table 2.6 - Example of one row in the table_samples dataframe, corresponding to the CZID report for sample A1

Columns	Values
SampleID	A1_CZID
Sample	A1
Replicate	1
Software	CZID
treatment_plant	WWTP 1
collection_step	Step 1
library_concentration (ng/µl)	114.03
volume_collected (L)	1
Raw_reads	3209246
Number_of_all_reads	40058
Number_of_all_viral_reads	7370
Number_of_all_non_viral_reads	31173
Number_of_undefined_reads	1515
Number_of_non_phage_reads	4687
Number_of_only_phage_reads	2683
Number_of_full_id_non_phage_reads	1015
Number_of_partially_identified_non_phage_reads	3672
Number_of_full_id_phage_reads	405
Number_of_partially_identified_phage_reads	2278
Number_of_total_pathogenic_IDs	20
Number_of_viral_pathogenic_IDs	19
Number_of_non_viral_pathogenic_IDs	1
Number_of_non_phage_pathogenic_IDs	19
Number_of_phage_pathogenic_IDs	0
Number_of_crassphages_reads	442
Number_of_poly_reads	4
Number_of_adeno_reads	24
Number_of_boca_reads	0

Following the creation of this summarized table (**table_samples**), the complete data frame is initially split into two new data frames, one containing all the identification data for viruses (taxid_df_viruses), another for non-viruses (taxid_df_non_viruses). The data frame taxid_df_viruses is then split into two data frames, one for all non-bacteriophage data and one for all bacteriophage data. This was accomplished by filtering according to the host_source. The filtering step is customizable when using the script, allowing for any combination of the host source possibilities contained in the ICTV Current Virus Metadata Resource, which contains a vast selection of host sources, both single , viruses which only infect bacteria, and multiple, viruses which can infect bacteria and vertebrates or infecting vertebrates and invertebrates (Table 2.7). For this analysis, the assumption was made that bacteriophages were all the

viral entities which infect either archaea or bacteria (no viruses were described as capable of infecting both).

Table 2.7 - Known categories of hosts infected by the virus, or the source [S] environment from which the virus was isolated. Extracted from ICTV Current Virus Metadata Resource (https://ictv.global/vmr/current)

Possible Host Sources
algae
archaea
bacteria
freshwater
fungi
invertebrates
invertebrates (S)
invertebrates, plants
invertebrates, vertebrates
marine (S)
other (specify)
phytobiome (S)
plants
plants (S)
plants, fungi
protists
protists (S)
sewage (S)
soil (S)
vertebrates

Having the different data frames filtered, an OTU (Operational Taxonomic Unit) table is created for each, which follows the typical OTU table format from other softwares such as QIIME2 (in this case, the samples columns were the sample-software pairs) in order to later perform statistical analysis in R. A long format of the out tables was also created for visualization in R. All the outputs created are described below.

Raw results of the pipeline were stored in **taxid_df.xlsx**, which comprises all identification data, for all samples and all softwares, with no filters. There were 4 different OUT tables: **otu_table_non_viruses.xlsx** comprises all identification data which did not belong to the kingdom of viruses; **otu_table_viruses.xlsx** comprises all identification data specifically belonging to the kingdom of viruses; **otu_table_non_phage.xlsx**, as the name suggests, is a partition of the previous table, with only viruses whose host is different from archaea or bacteria;

otu_table_phage.xlsx, as the name suggests, is a partition of the previous table, with only viruses whose host is archaea or bacteria. Additionally, all the identifications which had no taxonomic information were saved into a separate file, **undefined_data_for_R.xlsx**. The last output was the file **table_samples.xlsx**, which contained the complete metadata data frame previously mentioned in Table 2.6.

The analysis includes data from 8 sampling points, each with 3 replicates and 1 pooled sample, and each analyzed in 4 different softwares, resulting in a total of 128 reports. Each report is read and defined as a "sample" with the following structure <sample_name>_<software>, e.g. A1_CZID, B5_GD. The python script is available on GitHub (https://github.com/xiaodre21/comparison_of_viral_detection_methods) with the following options:

- update_ncbi_taxa (boolean expression), to update the NCBI database which is downloaded upon the first time the script is ran.
- tax_levels_for_classification, to define the taxonomic ranks to consider for the assignment of "full", "partial" or "undefined" classification
- phage_hosts, to define which host sources to consider for the split into the phage dataframe.

We additionally retrieved the number of reads from crAssphage, a recognized marker for detecting human fecal contamination (Sabar et al., 2022), by filtering the dataset containing only phages for those belonging to the order Crassvirales.

2.5 Statistical Analysis

Descriptive statistics (minimum, maximum, median and interquartile range) were obtained for the number of raw reads and nucleic acid extraction concentration of the 32 samples (8 sampling points * 4 replicates and pool) and for the number of viral, phage, non-phage reads of the 128 samples (8 sampling points * 4 replicates and pool * 4 softwares). We also calculated the proportion of viral reads, as well as of the proportion of phage and non-phage reads. The normality of data distributions was tested using Shapiro-Wilk normality tests. Given the deviation from normality of the data, non-parametric hypothesis testing was done subsequently. Correlations between pairs of these quantitative variables were tested using non-parametric Spearman correlation. The number of reads was compared between replicate and pooled samples using Mann-Whitney tests and between softwares using Kruskall-Wallis tests. Dunn's

post-hoc tests with Bonferroni correction were used for pairwise comparisons when Kruskal-Wallis indicated a significant difference between groups.

The R package Phyloseq was used for analysing abundance, viral richness and diversity within samples (alpha diversity). The rarefaction curves were generated using R package vegan, to assess the relationship between species richness and sampling effort across the different samples analyzed by various metagenomic sequencing software tools. Alpha diversity was assessed using the metrics Observed number of taxa (richness), Shannon Index and Simpson Index. Beta diversity was calculated using the same package. To visualize the differences in the viral communities' composition between samples (beta diversity) detected by different software tools, we performed Principal Coordinates Analysis (PCoA) based on the Bray-Curtis dissimilarity matrix.

The PERMANOVA (Permutational Multivariate Analysis of Variance) test was conducted to analyze the influence of different factors (software, treatment plant, and step) on the microbial community composition using the adonis2 R package. Statistical analyses were done in R version 4.3.0 using RStudio version 2024.04.2 Build 764. The significance level considered was 0.05. The R script used is available on GitHub (https://github.com/xiaodre21/comparison_of_viral_detection_methods).

3

RESULTS

3.1 Sequencing Performance and Variability

A total of 32 samples were sequenced (8 sampling points, each with 3 replicates and one pool). The results for RNA quantification and quality ratios for each extraction are summarized in Table 3.1.

Table 3.1 - RNA quantification and quality ratios

WWTP 1	Step Code	Replicates	[RNA] ng/µl	Abs 260/280	Abs 260/230
Influent before biofiltration 1L x 3	A	1	65.56	3.02	0.45
		2	62.08	3.00	0.37
		3	68.85	2.91	0.44
River effluent 10L x 3	B	1	81.65	2.51	0.18
		2	102.46	2.42	0.21
		3	75.71	2.62	0.18
Effluent before hypochlorite 10L x 3	C	1	82.16	2.88	0.52
		2	83.62	2.86	0.37
		3	96.73	2.64	0.53
Reused effluent 10L x 3	D	1	65.41	3.09	0.28
		2	85.54	2.97	0.60
		3	94.55	2.74	0.43
WWTP 2					
Influent before activated sludge 1L x 3	E	1	114.03	2.28	0.48
		2	112.76	2.36	0.61
		3	102.75	2.51	0.66
Effluent after activated sludge 5L x 3	F	1	101.23	2.47	0.61
		2	96.03	2.56	0.24
		3	103.74	2.51	0.24
River Effluent 10L x 3	G	1	70.36	3.05	0.51
		2	80.61	2.62	0.54
		3	61.14	2.69	0.29
Reuse Effluent 10L x 3	H	1	69.9	2.81	0.22
		2	72.55	2.99	0.19
		3	74.5	2.91	0.52

The number of raw reads per sequencing sample ranged from 132,416 to 6,827,732, with a median of 906,068 and interquartile range (IQR) of 1,107,881 (Figure 3.1).

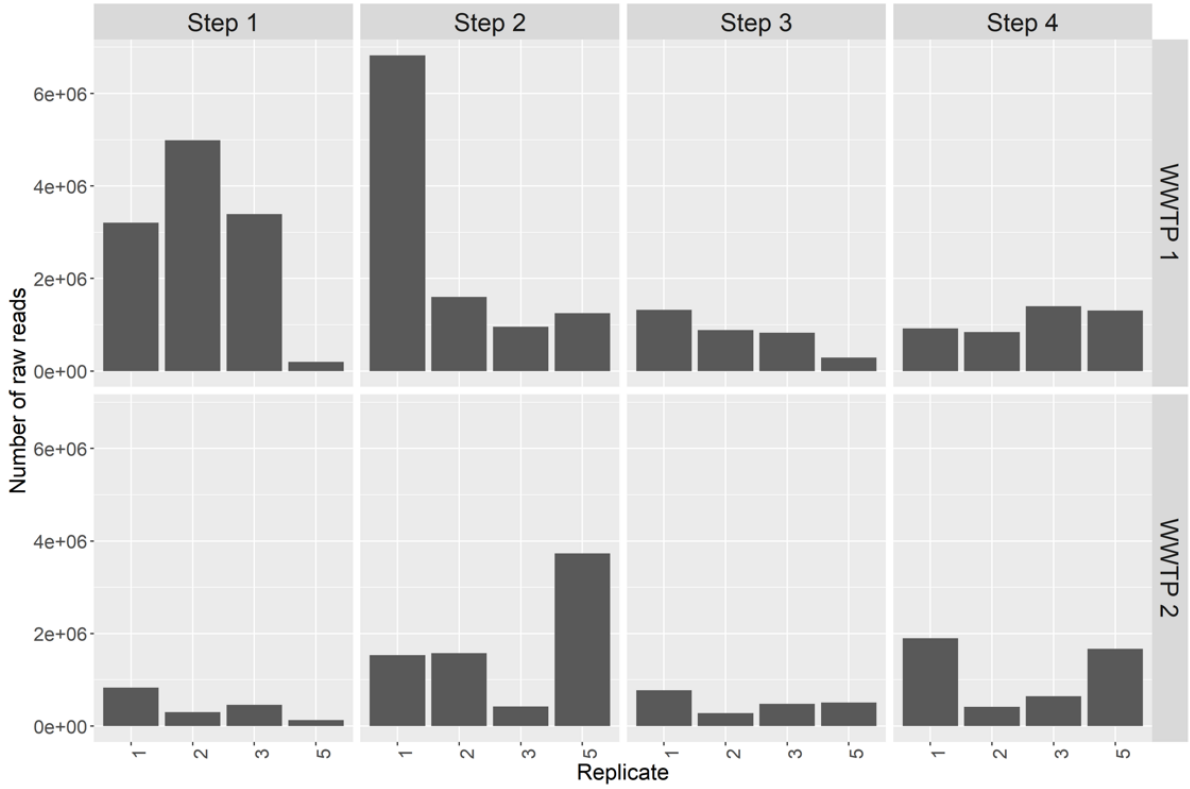

Figure 3.1 - Barplot with number of raw reads per sample (WWTP – wastewater treatment plant; Step –treatment phase; 1,2,3 – replicates; 5 – pool)

The extraction concentrations ranged from 50.00 to 114.03 ng/µl (Table 3.2), with a median of 73.52 ng/µl and an IQR of 36.56 ng/µl. The Shapiro-Wilk normality test indicated that the distribution of the number of raw reads and of the nucleic acid concentration deviated significantly from normality ($W = 0.74$, p-value < 0.001 and $W = 0.92$, p-value $= 0.02$, respectively). The Spearman correlation analysis revealed a positive and significant correlation between the number of raw reads and the extraction concentration of nucleic acid extractions ($rho = 0.39$, p-value < 0.001). No statistically significant difference was found in the distributions of the number of raw reads between replicate and pooled samples (Mann-Whitney test, $U= 76$, p-value $= 0.40$). Pooled samples had a median of 881,004 and IQR of 1,129,542, while replicate samples presented a slightly higher median but lower IQR, at 906,068 and 978,000, respectively.

Table 3.2 - Characterization and number of sequence reads for each sample triplicate and pooled samples

Treament_Plant	Step	Sample	Type	Sample Number	Volume Collected (L)	Concentration	Number of raw reads
WWTP 1	Step 1	A	Replicate	1	1	114.03	3209246
WWTP 1	Step 1	A	Replicate	2	1	112.76	4990332
WWTP 1	Step 1	A	Replicate	3	1	102.75	3396360
WWTP 1	Step 1	A	Pool	5	1	50	200942
WWTP 1	Step 2	B	Replicate	1	10	101.23	6827732
WWTP 1	Step 2	B	Replicate	2	10	96.03	1602524
WWTP 1	Step 2	B	Replicate	3	10	103.74	959348
WWTP 1	Step 2	B	Pool	5	10	50	1253304
WWTP 1	Step 3	C	Replicate	1	10	70.36	1320678
WWTP 1	Step 3	C	Replicate	2	10	80.61	887420
WWTP 1	Step 3	C	Replicate	3	10	61.14	829616
WWTP 1	Step 3	C	Pool	5	10	50	294302
WWTP 1	Step 4	D	Replicate	1	10	69.9	924716
WWTP 1	Step 4	D	Replicate	2	10	72.55	844362
WWTP 1	Step 4	D	Replicate	3	10	74.5	1404672
WWTP 1	Step 4	D	Pool	5	10	50	1311630
WWTP 2	Step 1	E	Replicate	1	1	65.56	829822
WWTP 2	Step 1	E	Replicate	2	1	62.08	299056
WWTP 2	Step 1	E	Replicate	3	1	68.85	458342
WWTP 2	Step 1	E	Pool	5	1	50	132416
WWTP 2	Step 2	F	Replicate	1	10	81.65	1532474
WWTP 2	Step 2	F	Replicate	2	10	102.46	1573236
WWTP 2	Step 2	F	Replicate	3	10	75.71	423008
WWTP 2	Step 2	F	Pool	5	10	50	3733250
WWTP 2	Step 3	G	Replicate	1	10	82.16	776692
WWTP 2	Step 3	G	Replicate	2	10	83.62	279222
WWTP 2	Step 3	G	Replicate	3	10	96.73	477456
WWTP 2	Step 3	G	Pool	5	10	50	508704
WWTP 2	Step 4	H	Replicate	1	10	65.41	1899016
WWTP 2	Step 4	H	Replicate	2	10	85.54	414040
WWTP 2	Step 4	H	Replicate	3	10	94.55	644258
WWTP 2	Step 4	H	Pool	5	10	50	1667124

WWTP: Wastewater treatment plant; Step: treatment phase (The correspondence between Step and type of treatment are presented in Table 3.1)

For the correlation between viral reads and raw reads, CZID, GD and Kraken2 showed a moderate positive correlation between the number of viral reads and the number of raw reads, with statistically significant p-values (Figure 3.2). This suggests that for these software tools, as the number of raw reads increases, the number of viral reads also tends to increase. However, INSaFLU - TELEVIR showed a weaker positive correlation not statistically significant ($p = 0.09$).

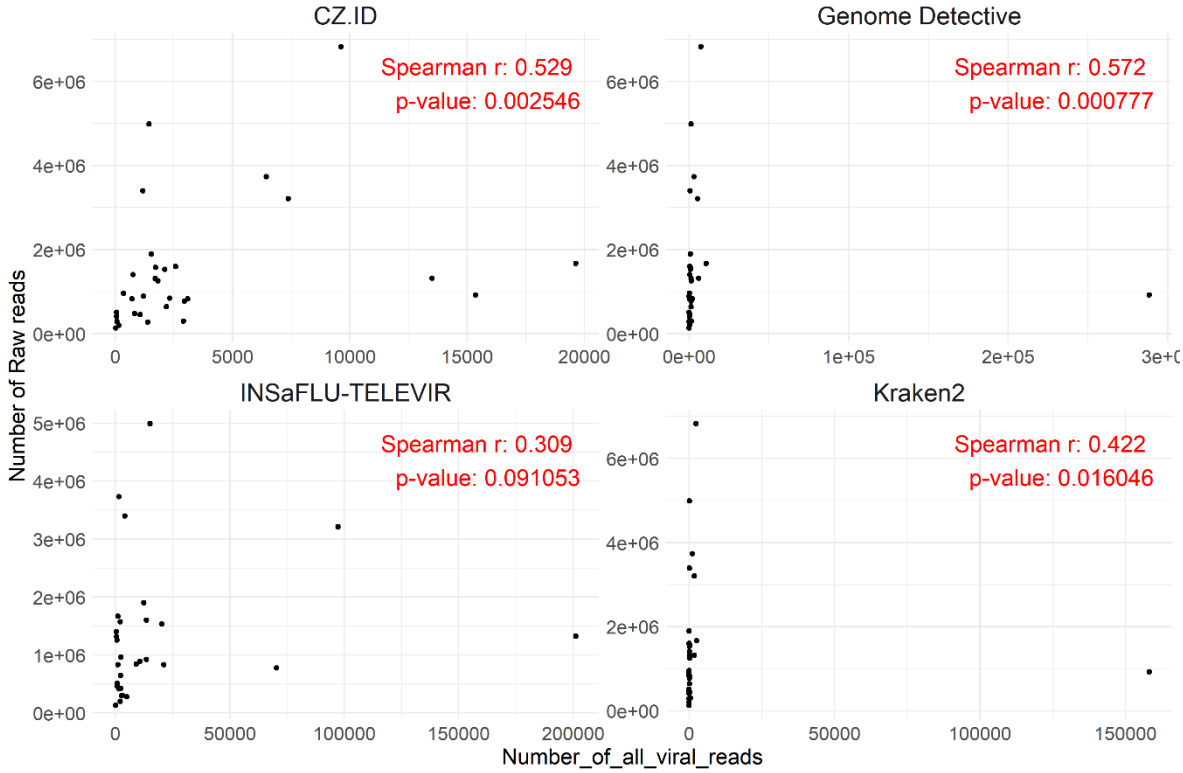

Figure 3.2 - Correlation between number of viral reads and number of raw reads, for each software

3.2 Viral identification results

Considering 128 sample-software reports, the minimum number of viral reads was 3, while the maximum was 288,464, with a median of 1092.5 and IQR of 2449.25. The sample-software with the highest amounts of mapped reads were D1 (WWTP1 – Step 4) from Genome Detective (with the maximum amount, i.e., 288,464), C1 (WWTP1 – Step 3) from INSaFLU – TELEVIR (201,374) and D1 (WWTP1 – Step 4) from Kraken2 (158,295). The sample-software with the lowest number of reads were sample G5 (WWTP2 – Step 3), H2 (WWTP2 – Step 4), both with 3 and E5 (WWTP2 – Step 1), with 4, all from Kraken2. The Shapiro-Wilk normality test indicated that the distribution of number of all viral reads deviated significantly from normality (W = 0.25259, p-value < 2.2e-16).

The proportion of viral reads in relation to the total raw reads per sample ranged from as low as 5.89e-06 to as high as 0.3119. The samples with the highest proportion of viral reads were D1 (WWTP1 – Step 4) from Genome Detective (with the maximum amount, i.e. 0.311949 or 31.19%), D1 WWTP1 – Step 4) from Kraken2 (17.11%) and C1 (WWTP1 – Step 3) from

INSaFLU–TELEVIR (15.23%). On the other end of the spectrum, samples A2 (WWTP1 – Step 1), H2 (WWTP2 – Step 4) and G5 (WWTP2 – Step 3), all from Kraken2, had the lowest proportion of viral reads at 1.34e-05, 7.24e-06 and 5.89e-06, respectively.

However, the number of viral reads did not differ significantly between the softwares in the CZ.ID vs. Genome Detective group and CZ.ID vs. INSaFLU–TELEVIR group (Table 3.3).

Table 3.3 - Median and IQR of viral reads per software

Software	Median	IQR
CZID	1705	2138.5
Genome Detective	761	1140
INSaFLU - TELEVIR	2335	11848.5
Kraken2	77	186.5

The Kruskal-Wallis test ($\chi^2 = 39.059$, df = 3, p-value < 0.001) indicated significant differences in the number of viral reads identified across the four software tools. Pairwise comparisons were tested with Dunn's post-hoc test with Bonferroni correction (Table 3.4), that revealed that the number of viral reads did significantly differ between CZ.ID (median 1705; IQR 2138) and Kraken2 (median 77; IQR 186), between Genome Detective (media 761; IQR 1140) and INSaFLU–TELEVIR (median 2,235; IQR 11,848), between Genome Detective (median 761; IQR 1140) and Kraken2 (median 77; IQR 186) and between INSaFLU–TELEVIR (median 2,335; IQR 11,848) and Kraken2 (median 77; IQR 186).

Table 3.4 - Comparison number of viral reads by pairs of softwares, using Dunn's post-hoc tests with Bonferroni correc-tion. First row value represents the Dunn's pairwise z test statistic; second row represents the p-value associat-ed with the test.

	CZ.ID	Genome Detective	INSaFLU-TELEVIR
Genome Detective	1.5990		
	0.3294		
INSaFLU-TELEVIR	-1.5772	-3.1887	
	0.3442	*0.0043**	
Kraken2	4.3852	2.8086	5.9749
	*0.0000**	*0.0149**	*0.0000**

3.3 Phage vs Non-Phage

The viral identifications were categorized as phage and non-phage. For each sample-software, the number of phage reads showed a wide range of values (Figure 3.3).

The samples with the highest proportion of phage reads were from C5 (WWTP1 – Step 3) from Genome Detective, D1 (WWTP2 – Step 1) from Kraken2 (0.999) and sample D1 (WWTP2 – Step 1) from Genome Detective (0.997). As for non-phage reads, the samples with the highest proportion of non-phage reads were sample C5 (WWTP1 – Step 3), E5 (WWTP2 – Step 1), both from INSaFLU–TELEVIR and sample E5 (WWTP2 – Step 1) from CZ.ID, with 100% non-phage reads. Additionally, some samples were observed to have similar phage percentages across replicates and pools, namely sample A (WWTP1 – Step 1) and B (WWTP1 – Step 2) in CZID. Sample C (WWTP1 – Step 3) in CZID also presented phage percentages congruent across replicates but not in the pool. Genome Detective identified the largest proportion of phages (median = 0.57), with Kraken2 and CZ.ID following with a median proportion of 0.53 and 0.49, respectively. INSaFLU-TELEVIR obtained less than 0.01. For non-phage proportions, the software with the highest median was INSaFLU-TELEVIR (median = 0.99) with CZ.ID having the second highest (median = 0.50) and Kraken2 and Genome Detective having 0.46 and 0.42, respectively.

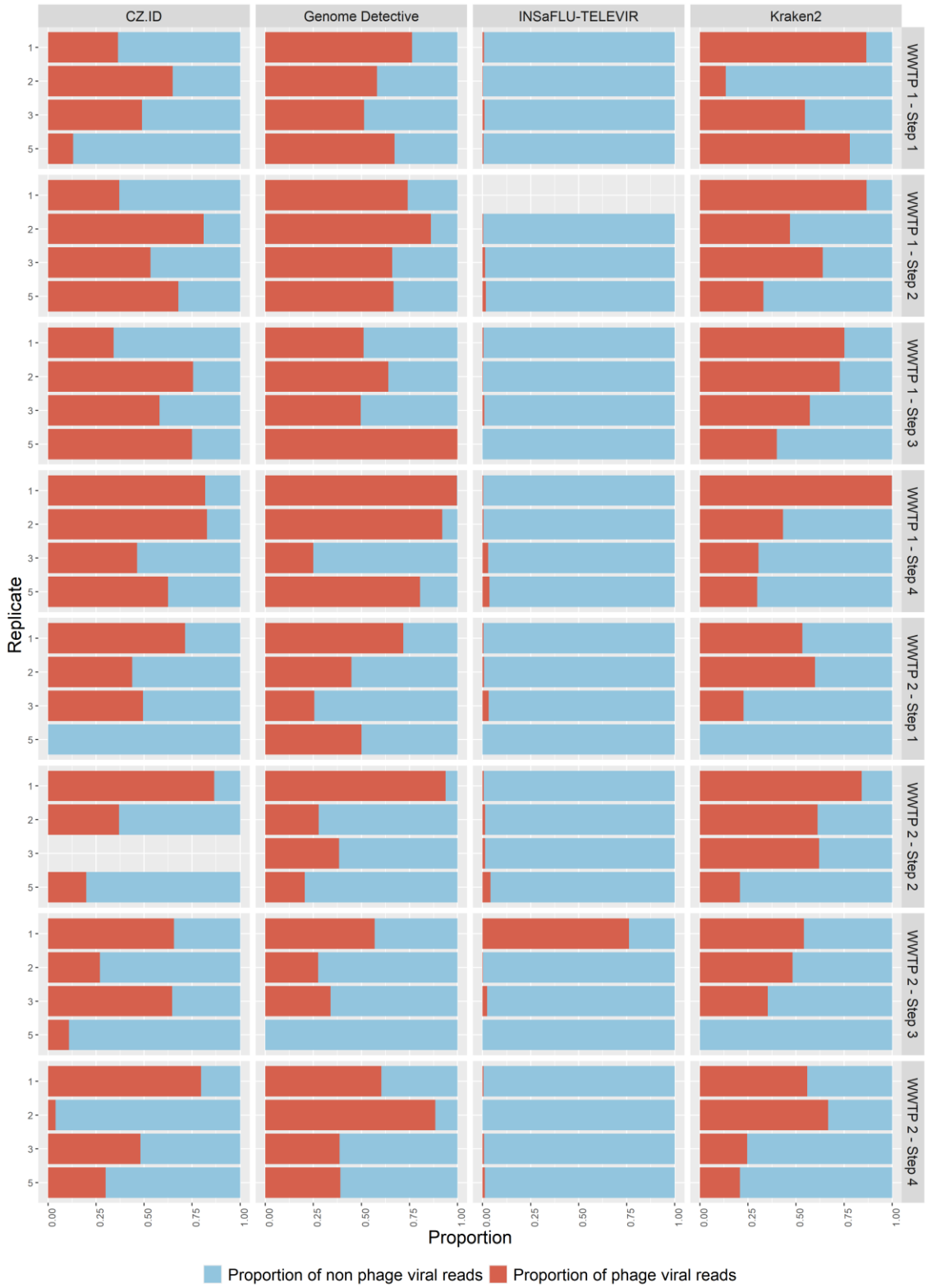

Figure 3.3 - Proportion of phage (red) and non-phage (blue) for each sample type (replicate 1, 2, 3 and pools - 5), by software (columns) and collection step (rows).

3.4 Classification Level

The proportion of reads identified as phage, non-phage, either fully or partially classified showed large differences between the four different bioinformatics software tools (Figure 3.4). Unsurprisingly, the comparison of workflows in the identification of non-phage and phage reads demonstrates significant variability across the different approaches.

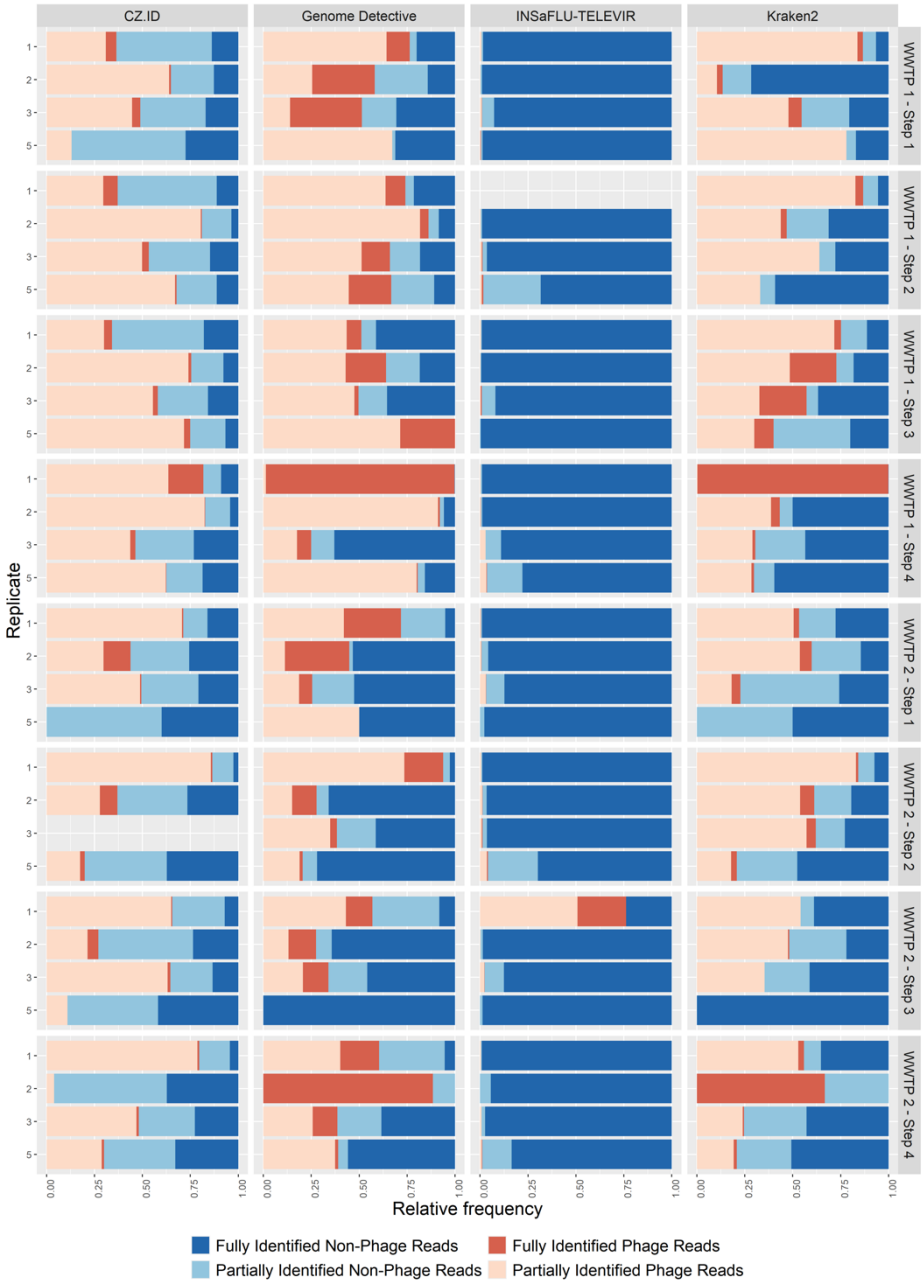

Figure 3.4 - Proportion of fully and partially identified phage and non-phage viral reads for each sample type (replicate 1, 2, 3 and pools - 5), by software (columns) and collection step (row). Fully identified are those for which all the classification levels (from kingdom to species) were available.

CZ.ID contained more partial classifications (for both phages and non-phages), highlighting a possible limitation in its ability to achieve taxonomical resolution for some sequences, which may lead to incomplete identification of viral communities. Genome Detective, while efficient in fully identifying non-phage reads, detected a higher proportion of partially identified phage sequences. INSaFLU–TELEVIR showed the highest proportion of fully identifying non-phage sequences across all samples. This is mainly due to the fact that, as previously stated, it does not include phage identifications in their reports, while the presence of some phage identifications was due to a small error present in this pipeline at the time the samples were submitted, meanwhile already corrected (Victor Borges; personal communication). In contrast, Kraken2 exhibited a moderate proportion of partially and fully non-phage viruses. The Kruskal-Wallis highlighted significant differences in identification performance between software workflows (p-value < 0.001 for all proportions) (Table 3.5).

Table 3.3 - Kruskal Walis tests comparing medians of number of fully identified reads (in relation to fully+partially identified) between softwares.

Viral_Type	Classification_level	Statistic	Value
Non-phage	Fully identified	H	61.817
		df	3
		p-value	*< 0.001*
	Partially identified	H	38.413
		df	3
		p-value	*< 0.001*
Phage	Fully identified	H	34.31
		df	3
		p-value	*< 0.001*
	Partially identified	H	36.46
		df	3
		p-value	*< 0.001*

3.5 Rarefaction Curve

The rarefaction curves (Figure 3.5) for almost all samples have curves showing an exponentially increase of number of taxa with increasing sample size (number of reads) but never

reaching a plateau, meaning the sequence effort was likely not enough and more species would occur if more sequencing reads were produced.

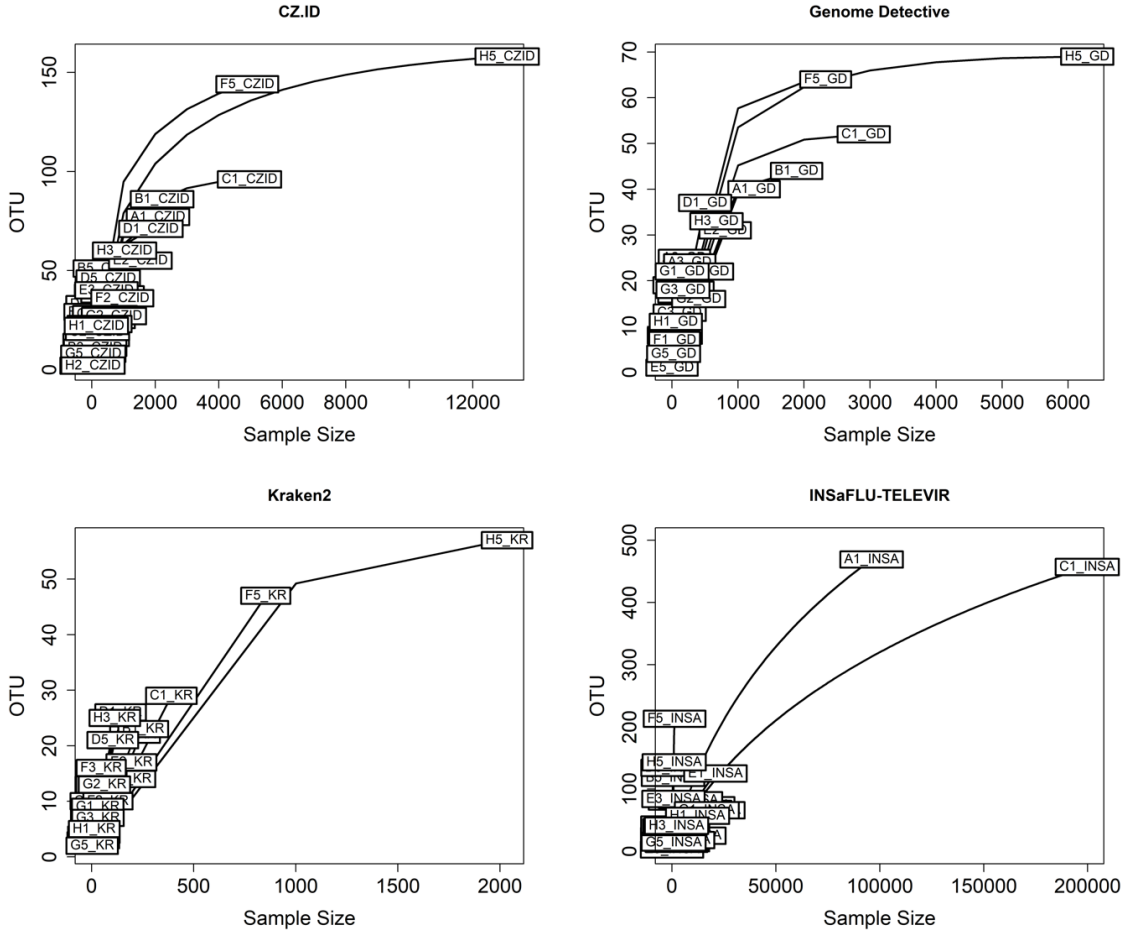

Figure 3.5 - Rarefaction curve for the different software, showing the number of identified taxa (OTU) increasing with higher number of reads (sample size).

3.6 Abundance Analysis

Despite the rarefaction curves indicating insufficient sequencing, we here characterize the abundance of families and genus found in each sample, with this caveat in mind, since we were mostly interested in comparing different softwares and not sampling sites. The abundance analysis focused on viral entities capable of infecting humans, meaning Phages were excluded to focus on non-phage viruses. This meant that from the total 128 initial reports (sample-software pair), 123 were retained after the phages were removed. After filtering for a minimum of 10 reads, only 105 sample-software remained. This dataset was used for the rest of the analysis.

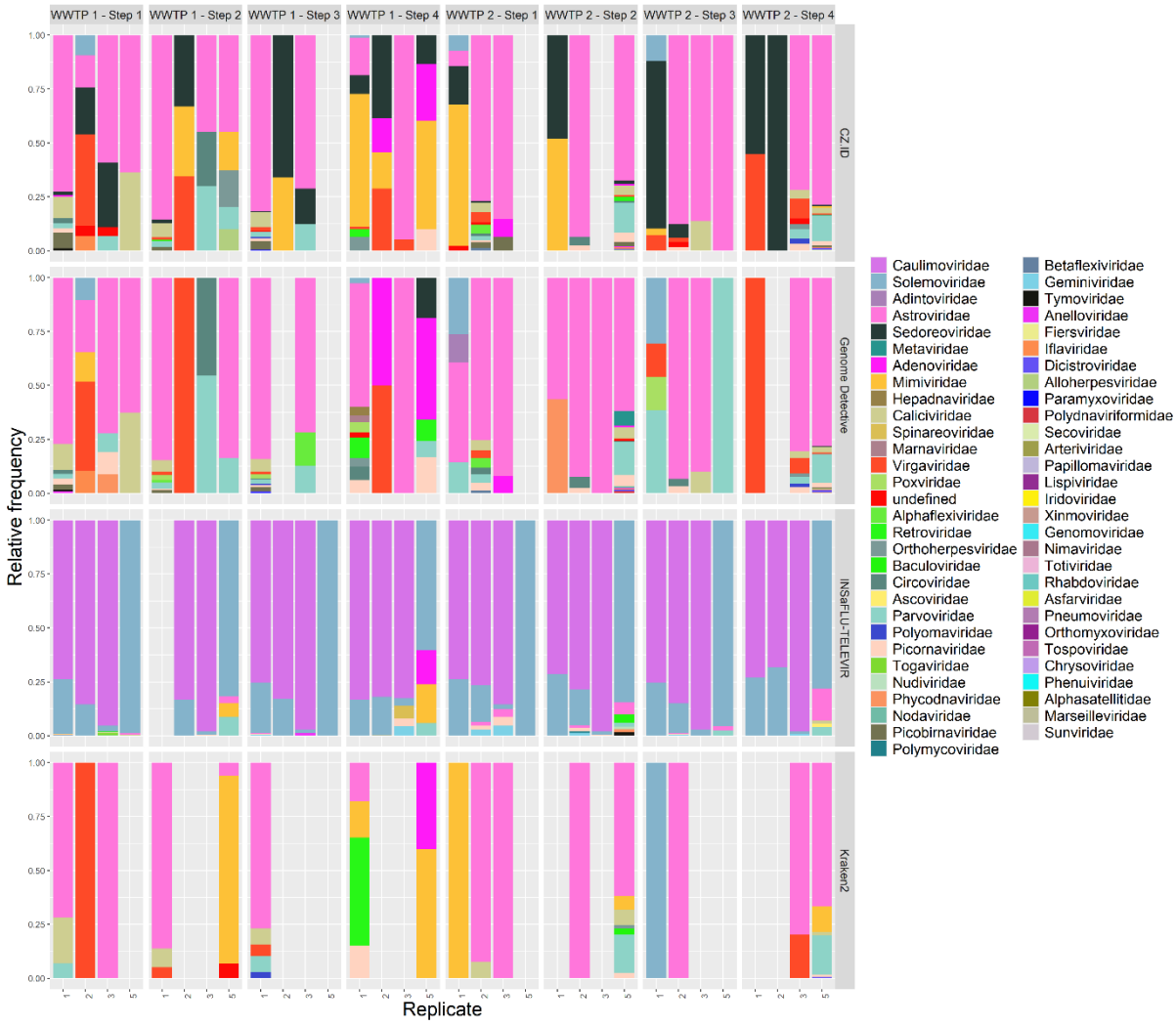

Figure 3.6 - Relative frequency of each viral family for each sample type (replicate 1, 2, 3 and pools - 5), by software (rows)and collection step (columns).

The analysis revealed the presence of 59 viral families (Figure 3.6), with varying levels of abundance across different samples.

Some viral families, such as Astroviridae, Caliciviridae, and Papillomaviridae, appeared consistently across various steps and replicates, indicating their widespread presence and potential resilience or proliferation during the treatment process.

The following list details the most abundant families observed in the study, along with the range of relative abundance values and the specific samples in which these values were recorded.

- Solemoviridae had the highest maximum relative frequency, reaching 99.4% in sample-software C5 (WWTP1 – Step 3) from INSaFLU-TELEVIR. The minimum relative

frequency for this family was notably low at 0.537% in sample-software D1 (WWTP1 – Step 4) from CZ.ID.

- Caulimoviridae also exhibited a maximum relative frequency of 95.1%, observed in sample-software H3 (WWTP2 – Step 4) from INSaFLU-TELEVIR. The minimum relative frequency for this family was at 0.014% in sample-software A1 (WWTP1 – Step 3) from INSaFLU-TELEVIR.
- Mimiviridae showed a maximum relative frequency of 74.7% in sample-software B5 (WWTP1 – Step 2) from Kraken2, with its minimum relative frequency being 0.007% in sample-software C1 (WWTP1 – Step 3) from INSaFLU-TELEVIR.
- Astroviridae had a maximum relative frequency of 64.3% in sample-software C1 (WWTP1 – Step 3) from Genome Detective and a minimum of 0.015% in sample-software C1 (WWTP1 – Step 3) from INSaFLU-TELEVIR.
- Virgaviridae showed a maximum relative frequency of 48.6% in sample-software B2 (WWTP1 – Step 3) from Genome Detective, and a minimum of 0.008% in sample-software C1 (WWTP1 – Step 3) from INSaFLU-TELEVIR.

Other families, such as Parvoviridae, Caliciviridae, Adenoviridae and Circoviridae were also detected with moderately high levels of abundance:

- Parvoviridae exhibited a maximum relative frequency of 37.0% in sample-software G5 (WWTP2 – Step 3) from Genome Detective, with a minimum abundance of 0.08% in sample-software H5 (WWTP2 – Step 4) from CZ.ID.
- Caliciviridae had a maximum relative frequency of 31.6% in sample-software A5 (WWTP1 – Step 1) from Genome Detective, with a minimum of 0.13% in sample-software H5 (WWTP2 – Step 4) from CZ.ID.
- Adenoviridae had a maximum relative frequency of 26.5% in sample-software D5 (WWTP2 – Step 1) from Genome Detective, with a minimum of 0.146% in sample-software A1 (WWTP1 – Step 1) from CZ.ID.
- Cirvoviridae had a maximum relative frequency of 26.3% in sample-software B3 (WWTP1 – Step 2) from Genome Detective, with a minimum of 0.076 in sample-software H5 (WWTP2 – Step 4) from CZ.ID.

The sample-software H5 (WWTP2 – Step 4) from CZ.ID consistently appeared as the one with the lowest relative frequency levels for several families. It identified 14 different families, being the fifth highest number, with C1 (WWTP1 – Step 3) from INSaFLU-TELEVIR having 27

(the maximum amongst all sample-software), while the mean was 4.14. When analyzing the absolute frequency values, there was also a large variation in viral family abundance across different steps of the workflow for each WWTP (Figure 3.7).

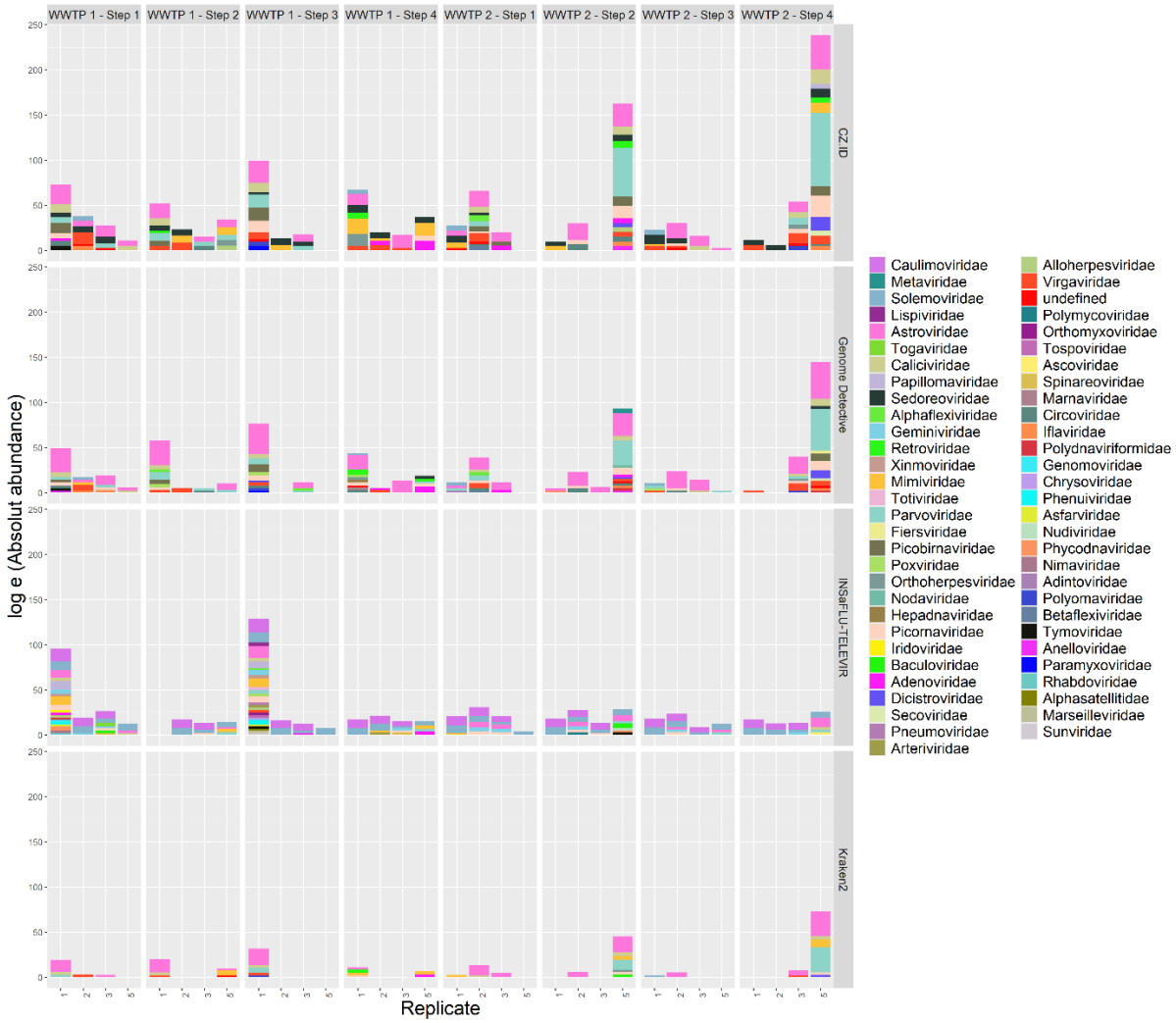

Figure 3.7 - Stacked barplot for log absolut frequency for each sample type (replicate 1, 2, 3 and pools - 5), by software and collection step

The pooled samples in Steps 2 and 4 in WWTP 2 show generally higher log(abundance) values compared to Steps 1 and 3 (in CZ.ID, Genome Detective and KR). This pattern was not visible in the results from INSaFLU–TELEVIR, as it displayed more abundance on replicate 1 of WWTP1, in Step 1. On the other hand, replicate 1 displayed a generally higher log(abundance) values in all steps of WWTP 1 for CZ.ID and Genome Detective. INSaFLU-TELEVIR also displayed this pattern but only in Step 1 and 3 of WWTP 1.

3.7 Alpha Diversity

In this study, we compared the alpha diversity metrics derived from metagenomic sequencing data using the different software tools. The Shapiro-Wilk test for normality on Observed number of taxa (richness), Shannon, and Simpson indices, all had p-values < 0.0001, indicating that none of these distributions are normally distributed. Levene's test was applied to homogeneity of variances across the Software groups, only Shannon and Simpson indexes were homogeneous (p-value = 0.92 and p-value = 0.16, respectively).

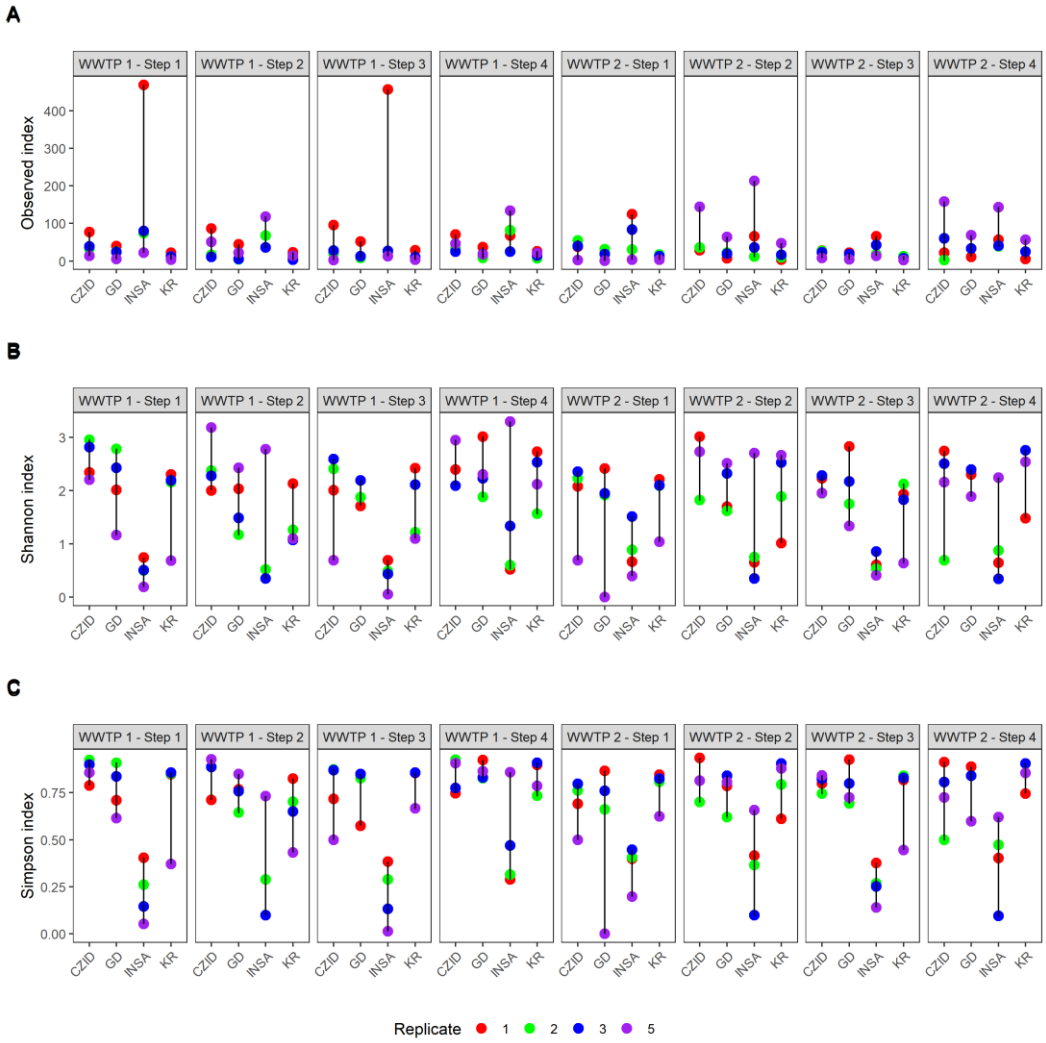

Figure 3.8 - Comparison of the alpha diversity metrics across the different software tools (bottom legends) and collection steps (top legends). (A) Observed number of taxa; (B) Shannon Index; (C) Simpson Index

Kruskal-Wallis test for the three indexes showed highly significant p-values, indicating that there are significant differences in these indices between the Software groups (Observed: p-value = 3.758e-08, Shannon: p-value = 9.194e-09 and Simpson: p-value = 7.793e-12). The

post-hoc tests (Dunn's Test) found differences in richness between all pairs of softwares except between GD and Kraken (Table 3.6). For Shannon and Simpson's index (Table 3.7 and 3.8), the differences were between those involving INSaFLU-TELEVIR.

Table 3.4 - Output of Dunn's Test for richness (observed number of OTUs) across softwares shows the differences in mean ranks between each pair of groups (upper value) along with the adjusted p-values (in italic) from the Holm correction (lower value).

	CZ.ID	Genome Detective	INSaFLU-TELEVIR
Genome Detective	2.228743		
	0.0387		
INSaFLU-TELEVIR	-1.701509	-3.916248	
	0.0444	*0.0002**	
Kraken2	3.994538	1.732917	5.696047
	*0.0002**	*0.0831*	*0.0000**

Table 3.5 - Output of Dunn's Test for Shannon Index across softwares shows the differences in mean ranks between each pair of groups (upper value) along with the adjusted p-values (in italic) from the Holm correction (lower value).

	CZ.ID	Genome Detective	INSaFLU-TELEVIR
Genome Detective	1.468441		
	0.1420		
INSaFLU-TELEVIR	6.066668	4.548294	
	*0.0000**	*0.0000**	
Kraken2	2.041221	0.555979	-4.025446
	0.0618	*0.2891*	*0.0001**

Table 3.6 - Output of Dunn's Test for Simpson Index across softwares shows the differences in mean ranks between each pair of groups (upper value) along with the adjusted p-values (in italic) from the Holm correction (lower value).

	CZ.ID	Genome Detective	INSaFLU-TELEVIR
Genome Detective	0.900333		
	0.5519		
INSaFLU-TELEVIR	6.519095	5.565106	
	*0.0000**	*0.0000**	
Kraken2	0.659034	-0.246723	-5.860061
	0.5099	*0.4026*	*0.0000**

3.8 Beta Diversity

When combined, the two principal coordinates in the Principal Coordinates Analysis (PCoA) explain around 23.76% of the variation in our data. The first coordinate separates INSaFLU-TELEVIR from the others. Kraken2 shows some clustering on one side of the second coordinate. CZ.ID and Genome Detective show no clear distinction (Figure 3.9-A). In Figure 3.9-B, although some samples can be observed as having their replicates grouped, the majority of samples has its replicates scattered, with pools being even more dispersed.

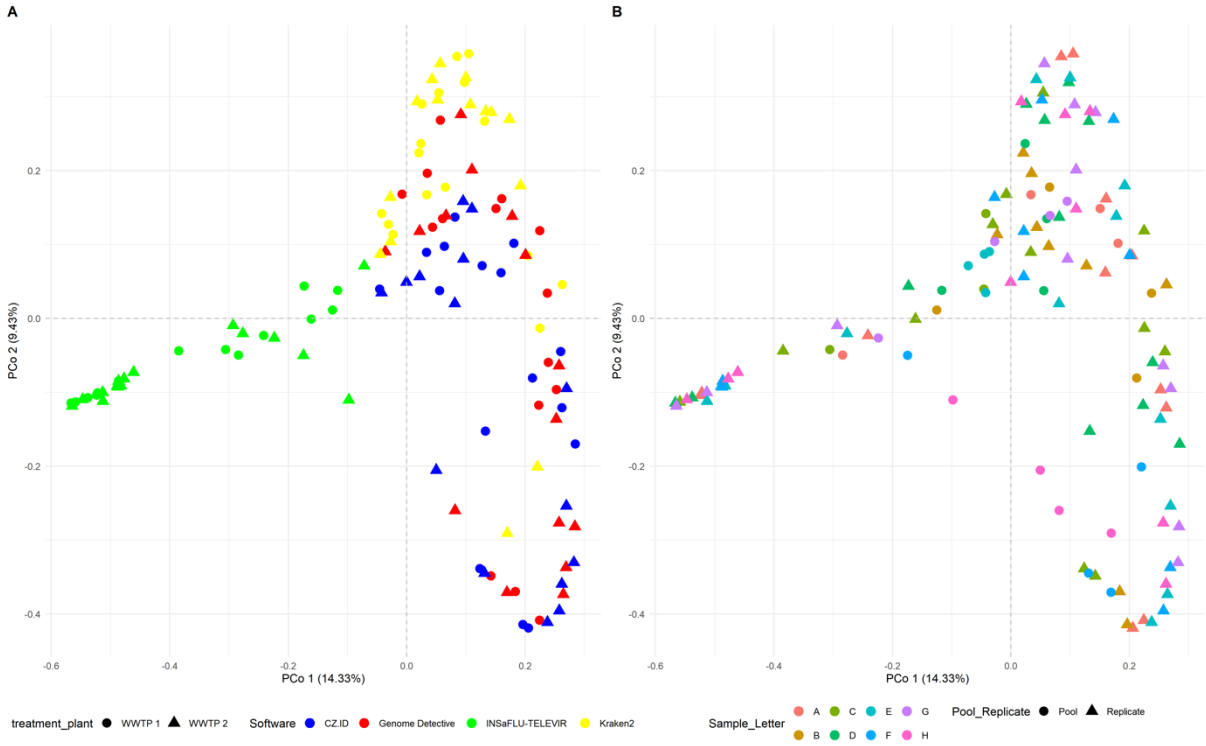

Figure 3.9 –Principal Coordinate Analysis (PCoA) based on the Bray-Curtis dissimilarity matrix. The same scatterplot of PCo1 versus PCo2 is shown with colours and symbols discriminating by: (A) treatment plant and software; (B) Sample and Sample Type (Pool or Replicate).

The factor "software pipeline" explains a significant proportion of the variation in microbial diversity (21.59%), and this effect is highly statistically significant ($F = 10.9846$, $df = 3$, $p < 0.001$). The factor "treatment plant" explains about 0.88% of the variation in microbial diversity, and is not statistically significant ($F = 1.34$, $df = 1$, $p = 0.10$). The collection step explains around 2.17% of the variation in microbial diversity, and this effect is also not statistically significant ($F = 1.10$, $df = 3$, $p = 0.26$). A large portion of the variation (75.35%) remained unexplained by the software, treatment plant, and collection step (Table 3.9). This suggests that

there may be other unmeasured factors contributing to the variation in microbial diversity, or that the remaining variation is due to random noise or individual sample differences.

Table 3.7 - PERMANOVA results

	df	F	p-value
Software	3	10.9846	0.000999***
Treatment Plant	1	1.3414	0.033966*
Step	3	1.1053	0.028971*
Residual	115		
Total	122		

3.9 Potentially Pathogenic Viruses Identified

Analyzing the reads from viruses which are potentially pathogenic to humans, 9 different families were identified, namely, Adenoviridae, Sedoreoviridae, Picornaviridae, Caliciviridae, Parvoviridae, Papillomaviridae, Polyomaviridae, Pneumoviridae and Orthoherpesviridae. Within each of the latter, the genera to which viral pathogens were assigned are those summarized in Table 3.10.

Table 3.8 - Potentially pathogenic families, genera, associated transmission and associated pathologies

Family	Genus
Adenoviridae	Mastadenovirus
Caliciviridae	Norovirus
Caliciviridae	Sapovirus
Orthoherpesviridae	Cytomegalovirus
Orthoherpesviridae	Lymphocryptovirus
Orthoherpesviridae	Roseolovirus
Papillomaviridae	Alphapapillomavirus
Parvoviridae	Bocaparvovirus
Picornaviridae	Enterovirus
Picornaviridae	Hepatovirus
Picornaviridae	Kobuvirus
Picornaviridae	Salivirus
Pneumoviridae	Orthopneumovirus
Polyomaviridae	Betapolyomavirus
Sedoreoviridae	Rotavirus

3.10 CrAssphage

In every sample, there were reads belonging to the order Crassvirales. Four different families and 27 genera within this order were found (Table 3.11) and there were a few identifications which belonged to the order of Crassvirales yet had no defined family nor genus.

Table 3.9 - Families and genera identified from the Crassvirales order

Family	Genus
Crevaviridae	Kahucivirus
Intestiviridae	Carjivirus
Intestiviridae	Delmidovirus
Intestiviridae	Diorhovirus
Intestiviridae	Fohxhuevirus
Intestiviridae	Jahgtovirus
Steigviridae	Kahnovirus
Steigviridae	Kehishuvirus
Steigviridae	Kolpuevirus
Steigviridae	Paundivirus
Suoliviridae	Afonbuvirus
Suoliviridae	Aurodevirus
Suoliviridae	Birpovirus
Suoliviridae	Blohavirus
Suoliviridae	Buchavirus
Suoliviridae	Buorbuivirus
Suoliviridae	Burzaovirus
Suoliviridae	Cacepaovirus
Suoliviridae	Canhaevirus
Suoliviridae	Cohcovirus
Suoliviridae	Culoivirus

The correlation between crAssphage and the number of potentially (human) pathogenic organisms revealed a moderate positive and significant correlation (rho = 0.39, p-value < 0.01). The correlation between crAssphage and Polyomavirus (rho = 0.33, p-value = 0.02) and between crAssphage and Adenovirus (rho = 0.3023, p-value = 0.03) was positive, moderate and significant. The correlation with Bocavirus was weak and not statistically significant (rho = 0.180, p-value = 0.22).

4

DISCUSSION

In this study, we aimed to compare four distinct bioinformatic software tools for analyzing viral sequences from wastewater samples. The motivation for comparing multiple software pipelines stems from the complexity of metagenomic data and the potential biases introduced at different stages of analysis, from sequencing to viral identification. The choice to evaluate Genome Detective, CZ.ID, the INSaFLU-TELEVIR module, and Kraken2 was driven by the need to assess the consistency of results across software with varied algorithms and workflows. Each software employs a different approach to crucial steps such as quality control, host depletion, and taxonomic classification. By comparing these tools, we sought to address key questions regarding the accuracy and diversity of viral detection and the influence of computational choices on final outcomes. The design decisions included analyzing triplicate samples and pooled samples from two wastewater treatment plants at four distinct sites, allowing us to evaluate the reproducibility of each software pipeline. Furthermore, we explored the impact of sequencing effort on viral abundance and examined the correlation between crAssphage, a known fecal contamination marker, and the detection of human pathogenic viruses. This benchmark comparison provides insights into the strengths and limitations of current metagenomic software tools and their applicability in environmental virology.

The comparison of the chosen bioinformatics software tools revealed notable differences in the detection and classification of viral sequences from wastewater samples. Across the 128 sample-software reports, the number of viral reads varied significantly between tools, with INSaFLU-TELEVIR and CZ.ID generally identifying a higher absolute number of viral reads than Kraken2 and Genome Detective. Statistical analysis indicated significant differences in viral abundance and diversity between software, with INSaFLU-TELEVIR showing the highest consistency in detecting non-phage viruses. Correlations between crAssphage and human

pathogenic viruses were moderate and significant, particularly for polyomavirus and adenovirus, further emphasizing the potential of crAssphage as a fecal contamination marker. The results highlight the variability in viral identification across software, emphasizing the importance of software choice in metagenomic studies.

4.1 Sample Processing

In Fernandez-Cassi et al., 2018, it is discussed how untargeted amplification using degenerate primers can offer broader detection of unknown or divergent viruses, although it is often less efficient for viral identification compared to targeted methods. The lower specificity and sensitivity compared to targeted approaches, competition among abundant sequences, amplification biases, all contribute to the challenges associated with untargeted amplification in viral metagenomics. Additionally, in the extraction process, the QIAamp® Viral RNA Mini Kit (QIAGEN®, USA) also extracts DNA, which further contributes to increased noise from other genomic sources such as non-viral sequences.

It has been described that, in wastewater, effects such as dilution and degradation have considerable impact in the level of nucleic acid degradation, which will most likely lead to the variations observed (Fontenele et al., 2021). In the previously citated study, samples were collected in the same treatment plant, two days apart and the difference between each day of collection was immense, as is noted by the total number of reads associated with samples collected on day 1, day 2 and day 3, which contained 11,859, 6,766 and 170,607 reads, respectively.

A total of 32 samples were sequenced with a wide range in the number of raw reads, from 132,416 to 6,827,732, with large variation also between replicates of some samples, suggesting variability in nucleic acid quality, library preparation, or sequencing efficiency, and not only biological variability (e.g., different viral loads in different wastewater samples). The choice of sample volumes (1 L in influent samples and 5 or 10 L in the other three steps) could also be discussed in terms of how it affects both the concentration of nucleic acids and the sequencing depth required to capture viral diversity effectively. Larger sample volumes in more diluted environments (as are the last three steps of wastewater treatment considered in this study) are required to capture viral particles, but it is arbitrary the choice of volume required to achieve comparable detection levels. In this study, there was a positive but non-significant correlation between number of raw reads and the nucleic acid extraction concentration. It is important to denote that even with higher nucleic acid concentration, quality still stands as a highly important factor. When taking into account the quality ratios for the RNA extractions, both absorbance ratios had poor values. For Abs 260/280 the values should ideally sit between 1.8 and 2.0 and our extractions had values between 2.28 and 3.09. For Abs 260/230

the values should be above 2.0 and they were between 0.18 and 0.66. However, this last ratio is known to be reduced due to contamination with guanidine thiocyanate present in the lysis buffer of the extraction kits, and it does not usually affect downstream procedures (Ahlfen & Schlumpberger, 2010) The SISPA protocol and the VirCapSeq-VERT capture Panel that were applied subsequently on these extractions were successful in amplifying and enriching for viruses, but differences in quality of the initial nucleic acid extraction should still be considered. Additionally, we must acknowledge that the wastewater samples collected will be complex matrixes with common human sewage, chemicals, and other substances. Meaning there are lots of different products which can interfere/inhibit the amplification process, resulting in the substantial variability observed in the number of raw reads across different samples. The randomness of the SISPA amplification is also another factor that can influence the sequencing output.

4.2 Pools *versus* individual samples

In metagenomic studies, pooling samples before sequencing is an efficient and valid approach for assessing community-level diversity and gene expression. Pooling allows for the combination of multiple samples into a single sequencing run, thereby significantly reducing the cost and processing time associated with sequencing large numbers of individual samples. Ray et al. (2019) demonstrated that pooling microbiome samples can provide reliable estimates of community-level diversity, showing high correlations with diversity estimates from individually sequenced samples. Their study showed that pooled samples could accurately estimate microbial richness and diversity indices (e.g., Shannon and Simpson's) with minimal loss of information. Similarly, Ko and Van Raamsdonk (2023) provided further validation for pooled sampling in RNA sequencing. They found that RNA sequencing of pooled samples could effectively identify differentially expressed genes, providing comparable results to those obtained from individually sequenced samples. This demonstrates that pooling does not compromise the ability to detect key biological signals, offering a cost-efficient method for large-scale gene expression studies. In summary, pooling samples for sequencing is a pertinent method in metagenomic studies, as it allows for broader coverage of the microbial diversity or gene expression landscape while reducing resource consumption.

With our results, we cannot conclude about the adequacy or not of using pools in comparison to individual samples due to the general wide variation between replicates of each sampling step, suggesting the influence of technical over biological factors in the sequencing outputs. No statistical significant difference was found in the number of raw reads between

pools (median of 906,068 reads) and replicates (median of 881,004 reads), although the pooled samples in Steps 2 and 4 in WWTP 2 showed substantial higher number of reads (and thus higher abundance values for several taxa; Figure 3.7) compared to Steps 1 and 3 (in CZ.ID and Genome Detective), while the replicate samples showed very low abundance values. Considering the within-sample diversity metrics, the values were similar between replicates and pools, with only two pooled samples having higher richness than their corresponding replicate samples, namely the pool from sample F (F5) and the pool from sample H (H5), both in CZ.ID and INSaFLU-TELEVIR.

4.3 Sequencing effort

A moderate positive correlation was observed between raw reads and viral reads using CZ.ID, Genome Detective, and Kraken2. This suggests that increasing sequencing efforts (i.e., generating more reads) could enhance viral detection in wastewater samples, which is important in environmental monitoring and pathogen surveillance. On the other hand, for INSaFLU-TELEVIR, there was a moderate, non-significant correlation. The observed correlations support the hypothesis that deeper sequencing enhances viral detection. This is further supported by the rarefaction curve analysis, which showed that for most samples, the curve increased rapidly but did not reach a plateau. This indicates that the sequencing depth was not sufficient to capture the full viral diversity present in the analyzed samples. Therefore, more sequencing would likely reveal additional viral species, suggesting that the actual number of pathogenic viruses in these wastewater samples could be higher than what was detected. This observation highlights the need for deeper sequencing in future studies to fully capture the viral diversity, particularly for pathogens that may be present in low abundance. Future research could focus on incorporating real time viral detection with sequencing platforms in order to allow for the monitoring of the sequencing depth in real time, contributing to an appropriate sequencing run time.

4.4 Viral Identification

Although a positive and significant correlation was found between the number of viral reads and the number of raw reads, some samples were found to have fewer viral reads while having substantially high numbers of raw reads. Sample F2, as an example, had 1,573,236 raw

reads, however, when submitted to each software, resulted in reports where the number of identified viral reads varied between 206 and 3132. Sample G2, on the other hand, with just 279,222 raw reads, had reports where the number varied between 141 and 5033 viral reads. In other words, despite the significant positive correlation found, some samples displayed a high number of raw reads but low number of viral reads, which can be due to a characteristic of these samples or a lack of efficiency in the viral enrichment step prior to sequencing. In the CZ.ID report for sample G2, around 1,900 reads were identified as not viral, while the number of viral reads was 1373. In Genome Detective the number of viral reads was 560, in INSaFLU-TELEVIR it surpasses 5,000, and in Kraken2 it drops to 141. This fluctuation in the number of viral reads, for the same sample, through different software pipelines shows how much of a crucial difference the software choice plays. In sample D (WWTP1 – Step 4), the number of raw reads in replicates and pools were rather similar, fluctuating between 920,000 and 1,400,00, averaging near 1,000,000 raw reads, while looking at the number of viral reads detected for the different replicates and pool, for each software, we observed massive differences. In CZ.ID, sample D obtained between 1,386 and 72,252 reads assigned, while the number of viral reads only ranged from 748 to 15,366, where the sample with the highest number of reads assigned contained only 1,705 viral reads and 54,914 non-viral reads. This can highlight that even through the usage of a viral enrichment step, DNA and RNA will be amplifying from other organisms such as bacteria, fungi, eukaryotes, etc. Comparing softwares, INSaFLU-TELEVIR identified a higher median number of viral reads compared to the other tools, while Kraken2 detects the lowest median number of viral reads with the smallest variability.

The differences in number of viral reads between all pairs except between CZ.ID and Genome Detective, and CZ.ID and INSaFLU-TELEVIR. Both CZ.ID and INSaFLU-TELEVIR detected high numbers of viral reads compared to Genome Detective and Kraken2, suggesting they may be more sensitive in detecting viral sequences or better optimized for this type of analysis. The marked difference in performance between software tools should be discussed in the context of their underlying algorithms and objectives. For example, Genome Detective, and INSaFLU-TELEVIR may have optimized databases or alignment algorithms for detecting viral sequences, whereas CZ.ID is better suited for a broader metagenomic microbial detection, detecting not just viruses, but bacteria, eukaryotes, fungi, which can lead to a decreased sensitivity to specific characteristics of viral genomes.

The distinction between phages (viruses that infect bacteria) and non-phage viruses (which may infect eukaryotic organisms or other hosts) is important for understanding the viral ecology in WWTPs. The range in the number of phage reads may suggest that certain samples, or specific stages in the WWTP wastewater treatment process are more likely to contain bacteriophages, potentially reflecting the bacterial community structure in the

wastewater. Sample D1 appeared with the highest proportion of phage in two of the softwares, which can hint at the possibly because the microbial communities at these particular steps were primarily composed of bacterial populations. Conversely, in some samples, non-phage viruses (potentially human or animal viruses) were more prevalent. Since the results indicated a wide range in the proportion of phage versus non-phage reads, this can suggest that the viral community composition is highly dynamic and may depend on factors such as the treatment step, microbial population, and environmental conditions within the WWTP. Some of this variation is also attributed to the software pipelines, as for the same sample, the proportions of both phage and non-phage could vary substantially. In sample H2 non-phages were detected as composing 95% of viral reads in CZ.ID, around 5% in Genome Detective, almost 100% in INSaFLU-TELEVIR and around 55% in Kraken2. The lack of consensus between pipelines underlines the potential need for the use of various tools, as the comparison between them could provide a more comprehensive overview of the results.

Both the absolute number of viral reads and the proportion of viral reads relative to the total amount of reads observed need to be taken into consideration. Some samples were observed to have the lowest number viral reads (Appendix A.2) while having the highest number of identified reads, such as sample A5 from CZ.ID, with 153 viral reads out of 33,810 total reads identified, or sample B5 from Kraken2, with 287 viral reads out of 347,885 total reads.

4.5 Taxonomic classification levels

The identifications were assigned a classification level depending on whether the taxid assigned contained full taxonomic information or only partial (at least one taxonomic rank missing). CZ.ID exhibited a higher proportion of partially classified reads for both phage and non-phage sequences. This suggests that while CZ.ID is effective in detecting viral sequences, it may encounter limitations in achieving full taxonomic resolution for certain reads. The prevalence of partial classifications could stem from the breadth of its reference databases or the alignment methods used, which might not fully resolve sequences lacking close matches or representing highly divergent viruses.

Genome Detective demonstrated efficiency in fully classifying non-phage reads, indicating robust performance in identifying known viral sequences within its databases. However, it reported a higher proportion of partially identified phage sequences. This discrepancy could be due to the high diversity and abundance of phages in the samples, coupled with the complexity of phage genomes and their underrepresentation in reference databases. Phages often

possess high genetic variability, making them challenging to classify fully, especially if they are novel or less characterized.

INSaFLU–TELEVIR outperformed the other workflows in fully identifying non-phage sequences, showing the highest proportion of fully classified reads across all samples. This superior performance is primarily attributed to its clinical focus, where it is designed to detect known viruses of human health significance to streamline clinical diagnostics. The presence of some phage identifications in our analysis was due to a minor error in the pipeline at the time of sample submission, which has since been addressed. By concentrating on non-phage viruses, INSaFLU–TELEVIR enhances its ability to fully classify clinically relevant pathogens but may not be suitable for studies where phage detection is important.

Kraken2 showed a moderate proportion of both partially and fully classified non-phage viruses. As a k-mer-based classifier, Kraken2 rapidly assigns taxonomic labels by matching k-mers from sequencing reads to those in its reference database. While efficient, k-mer-based methods can struggle with sequences from organisms not well represented in the database or with high genetic variability, leading to partial classifications. Kraken2's performance reflects a balance between speed and classification depth, making it a versatile tool for general purposes but potentially less effective for comprehensive viral profiling.

Statistical analyses confirmed the significant differences in classification performance among the workflows and subsequent non-parametric tests, highlighted significant differences in identification performance between the workflows for all classification levels.

4.6 Viral families' abundance and diversity

The taxonomic abundance analysis aimed to focus on viral entities capable of infecting humans, and phages were excluded to focus on non-phage viruses. After filtering out identifications with fewer than 10 reads, 105 samples were retained for further analysis, which provided a clearer picture of the viral diversity and abundance in the wastewater samples. Although, having an initial value of 128 samples, 111 after removing phages, we were left with 86% of the initial sample size. This step was essential to eliminate potential noise from low-read viral sequences that might not contribute significantly to the overall viral community or that could represent sequencing artifacts.

A total of 59 viral families were identified, each exhibiting varying levels of abundance across the different samples. Some viral families were consistently detected across various wastewater treatment steps and replicates of sampled water. Their persistent detection suggests that they may be particularly resistant to degradation during the wastewater treatment

process or that the WWTP removal processes to be inefficient, as it was the fact for WWTP 1, where the hypochlorite treatment step was not working at the time of sample collection. Further research could explore how specific environmental factors in the treatment plants influence viral abundance, such as water temperature, chemical treatment, or microbial community dynamics. The detection of these viral families, particularly those known to infect humans like Adenoviridae and Astroviridae, has significant implications for wastewater-based epidemiology. The high abundance of these viral families suggests that wastewater could serve as a valuable resource for tracking viral outbreaks or monitoring the presence of human-infecting viruses in the population. Given that adenoviruses are known to infect humans, their detection aligns with the goal of monitoring human-infecting viruses. Parvoviridae, Caliciviridae, Adenoviridae and Cirvoviridae also demonstrated substantial presence specifically by Genome Detective. Sample H5 (WWTP2 – Step 4) from CZ.ID, had the highest number of viral reads, highest number of families (14 families, while the median number of families identified per sample was 4), However, it had the lowest abundance levels for several of these viral families, such as Parvoviridae, Caliciviridae, Cirvoviridae and others. This low abundance across multiple families might indicate amplification biases or other technical factors that affected viral detection.

The clear variation in viral family abundance across the different steps of the wastewater treatment process, particularly in Steps 2 and 4 of WWTP2, which showed generally higher log(abundance) values compared to Steps 1 and 3 of the same WWTP, or in Steps 1 and 3 of WWTP1 can hint towards the idea that viral abundance may increase as the treatment process progresses. Although counterintuitive, a possible explanation is that as the treatment process continues, viral particles become more concentrated, either due to the breakdown of larger materials or because of other factors like water volume reduction. Another unexpected pattern was that replicate samples higher taxonomic abundance and diversity than pools in WWTP1, while in WWTP2 the opposite happened.

The high viral abundance in early stages, Step 1 and Step 2, could point to the potential use of these stages as critical points for sampling when conducting viral surveillance in wastewater treatment plants. At the same time, the increase in viral abundance in later treatment steps may also be used as a method to monitor the efficiency of the treatment strategies employed.

The alpha diversity analysis suggest that the different bioinformatics tools used in the study yielded significantly different diversity metrics, likely due to variations in their algorithms and databases. INSaFLU-TELEVIR showed the most significant differences compared to the other tools, which may be due to its exclusion of phages or specific features of its algorithm, which could skew the diversity results toward non-phage viruses. Genome Detective

and CZ.ID were not significantly different for the three diversity indexes, reflecting their similarities in the identification of organisms. Kraken2 showed borderline significant differences from Genome Detective for the Observed index, indicating that the two tools may classify the viral communities somewhat differently. INSaFLU-TELEVIR stands out as identifying lower evenness and richness, due to its exclusion of phages, which would reduce overall community complexity. On the other hand, the similarity in results between CZID, GD, and KR suggests that these tools may offer more balanced insights into viral diversity, capturing both phages and non-phage viruses more comprehensively. This could make them more suitable for studies requiring a broad understanding of viral community structure in wastewater samples.

4.7 Comparison of viral communities

The PCoA analysis revealed that the first two principal coordinates (PCo1 and PCo2) explained 23.76% of the variation in the data, which captures a reasonable amount of the differences between samples. However, a substantial portion of variation remains unexplained, indicating that additional principal components or other unmeasured factors might be influencing microbial diversity. CZ.ID and Genome Detective samples exhibited some high overlap, suggesting that they identified similar viral communities. INSaFLU-TELEVIR samples demonstrated some separation, as well as Kraken2 samples. These two clusterings indicates that these tools may have a distinct approach to classifying microbial diversity. The significant influence of software choice on beta diversity emphasizes the need for consistency in bioinformatics pipelines when comparing microbial diversity across studies. Differences in algorithms, databases, and classification methods can lead to divergent results, complicating comparisons between studies using different tools. When looking at the plot which discriminates samples and sample type, we found that replicates of the same sample do not group often, appearing rather overall scattered, which indicates that each replicate has a substantial difference in their viral composition. Wastewater matrices can be quite complex and are often very heterogeneous, which is the primary reason for the result obtained.

The PERMANOVA analysis showed that the choice of software tool explained a significant proportion of the variation in microbial diversity, indicating that the software's algorithm plays a major role in shaping the diversity patterns observed in the data. The WWTP and collection step factors explained a rather small proportion of the variation. The majority of the variation in microbial diversity remains unexplained. This large portion of unexplained variation could be due to unmeasured variables such as environmental conditions (e.g., temperature, chemical treatments), differences in sample handling, or stochastic processes affecting

microbial populations. The high unexplained variation might also indicate that the sequencing depth or other technical factors, like sample contamination or sequencing noise, could be influencing the results. Further research could focus on identifying these additional factors to better explain the variation in microbial communities observed in the wastewater samples and investigate specific factors within treatment plants, such as water flow rates, chemical treatments, or microbial activity, that might explain these differences.

4.8 Potential pathogenic viruses

Nine different viral families were identified as potentially pathogenic to humans, some of which contained viruses known to cause human diseases. As they are associated with a variety of diseases, including respiratory infections, gastroenteritis, hepatitis, and even viral-induced cancers (e.g., alphapapillomaviruses) (Van Doorslaer & Burk, 2010), the detection of these viral families underscores the relevance of wastewater monitoring for public health and demonstrates that wastewater samples can serve as valuable indicators of viral presence in a population, offering insight into potential pathogen transmission within the community. In Itarte et al. (2024) Adenoviridae, Circoviridae, Orthoherpesviridae, Papillomaviridae, and Parvoviridae were found both in a WWTP and a swine farm, further corroborating the presence of these families in our samples. In Hellmér et al. (2014), the authors mention noroviruses, sapoviruses, rotaviruses, astroviruses, and adenoviruses as viruses systematically detected in sewage water. In our study, the two types of adenoviruses detected were aviandenoviruses and mastadenoviruses, which infect birds and bats, respectively. Another paper (Corpuz et al., 2020) mentions adenoviruses, rotaviruses, the hepatitis A virus, as well as some other enteric viruses (noroviruses, coxsackieviruses, echoviruses, and astroviruses) as some of the main viruses transmissible via water media. Coxsackieviruses and echoviruses weren't detected in our analysis, and only one of two genera of astroviruses were detected. Amongst the detected viruses, those who were identified in some genera identified have particular interest, namely, *Mastadenovirus* (from *Adenoviridae*), which includes adenoviruses that are common pathogens associated with respiratory infections, conjunctivitis, and gastroenteritis; *Norovirus* and *Sapovirus* (both from Caliciviridae), both of which are major causes of viral gastroenteritis, commonly responsible for outbreaks in community settings like schools and hospitals; *Enterovirus* and *Hepatovirus* (both from *Picornaviridae*), which include viruses such as polioviruses and hepatitis A virus, both of which have significant public health importance due to their modes of transmission (fecal-oral) and potential for outbreaks; *Rotavirus* (from Sedoreoviridae), another major pathogen that causes severe diarrhea, particularly in children,

and is a leading cause of viral gastroenteritis globally. The identified viral families are associated with a range of transmission routes, primarily via the fecal-oral route, highlighting the importance of monitoring wastewater for pathogens that spread through contaminated water frequently associated with poor sanitation. This diversity of associated pathologies underscores the importance of comprehensive monitoring for a wide range of viral pathogens in wastewater.

The detection of these potentially pathogenic viruses demonstrates the effectiveness of using metagenomics for viral surveillance in wastewater. The identification of genera that include viruses that are known to cause human disease suggests that wastewater-based epidemiology could be a valuable tool for tracking the spread of infectious diseases within a community. Wastewater surveillance could serve as an early warning system for outbreaks of viral diseases, such as noroviruses or enteroviruses associated outbreaks, by detecting these agents before clinical cases are widespread. Similarly, the surveillance of other types of water sources, such as sea or consumption water, could be pertinent as a means of quality control monitoring. Today, both the European (Directive (EU) 2020/2184 of the European Parliament) and Portuguese regulations (Decreto-Lei n.º 306/2007 de 27 de Agosto) only dictate water quality to be evaluated through the monitoring of fecal coliforms, which are bacteria.

The presence of crAssphage in wastewater samples is of particular interest as it is recognized as a strong indicator of human fecal contamination. CrAssphage is not only found in higher concentrations than traditional fecal indicators but also exhibits a strong correlation with other human pathogens. In this study, the positive and significant correlations observed between crAssphage and several human pathogenic viruses (polyomaviruses, and adenoviruses) support its utility as a marker for tracking fecal-oral contamination in wastewater systems. However, the strong correlation with polyomaviruses and adenoviruses highlights the potential of crAssphage to act as an indicator for the presence of pathogenic viruses that can be transmitted via the fecal-oral route. Given crAssphage's resilience to environmental stress and its high abundance in wastewater, it serves as a reliable biomarker for assessing viral contamination and tracking the effectiveness of wastewater treatment processes. This finding reinforces the notion that crAssphage can be an essential tool in monitoring viral contamination, not only as a standalone marker of human fecal pollution but also as a proxy for the presence of other clinically relevant viruses, enhancing the surveillance of public health risks in wastewater treatment environments.

5

Conclusion and future perspectives

This study demonstrated the complexity and variability inherent in viral metagenomic analysis, particularly when different bioinformatic pipelines are applied. The results underline the importance of tool selection based on study objectives, as each software showed strengths and weaknesses in viral classification. INSaFLU-TELEVIR and CZ.ID, for example, excelled in classifying non-phage viruses, while Kraken2 struggled with overall taxonomical resolution. Additionally, the presence of crAssphage served as a useful marker for human fecal contamination, correlating with the presence of potentially pathogenic viruses in certain samples. However, differences in the proportion of viral reads across tools highlight the need for further development and standardization in metagenomic approaches.

The significant differences in the classification performance of the evaluated bioinformatics tools emphasize the critical role of computational methods and database comprehensiveness in viral metagenomic analyses. Fully understanding the capabilities and limitations of each workflow allows researchers to make informed decisions, tailoring their analytical strategies to the demands of their specific research questions. As viral metagenomics continues to evolve, ongoing improvements in bioinformatics tools and databases will enhance our ability to accurately characterize the vast diversity of viral communities across different environments and sample types. The lack of consensus between pipelines underlines the potential need for the use of various tools, as the comparison between them could provide a more comprehensive overview of the results.

Future research could focus on integrating multiple tools to improve sensitivity and specificity in viral detection. The identification of viral diversity in wastewater has profound implications for public health, enabling the monitoring of viral outbreaks, the discovery of novel pathogens, and the assessment of viral evolution. As wastewater-based epidemiology continues to evolve, it will play a pivotal role in disease surveillance and environmental health.

Advanced tools, including machine learning algorithms for virus-host prediction and genome reconstruction, are critical for improving the sensitivity and specificity of wastewater metagenomics, contributing to overcome the complexities of environmental virome studies.

6 BIBLIOGRAFIA

Afgan, E. et al. (2022) 'The Galaxy Platform for Accessible, reproducible and collaborative biomedical analyses: 2022 update', Nucleic Acids Research, 50(W1). doi:10.1093/nar/gkac247.

Ahlfen, S & Schlumpberger, M. 2010. Effects of low A260/A230 ratios in RNA preparations on downstream applications. Qiagen Gene Expression Newsletter, 15/10: 7-8. https://www.qIagen.com/us/resources/faq?id=c59936fb-4f1e-4191-9c16-ff083cb24574&lang=en

Altschul, S. (1997) 'Gapped BLAST and PSI-BLAST: A new generation of protein database search programs', Nucleic Acids Research, 25(17), pp. 3389–3402. doi:10.1093/nar/25.17.3389.

Altschul, S.F. et al. (1990) 'Basic local alignment search tool', Journal of Molecular Biology, 215(3), pp. 403–410. doi:10.1016/s0022-2836(05)80360-2.

Ballesté, E. et al. (2022) 'Bacteriophages in sewage: Abundance, roles, and applications', FEMS Microbes, 3. doi:10.1093/femsmc/xtac009.

Bankevich, A. et al. (2012) 'Spades: A new genome assembly algorithm and its applications to single-cell sequencing', Journal of Computational Biology, 19(5), pp. 455–477. doi:10.1089/cmb.2012.0021

Benler, S., & Koonin, E. (2021). Fishing for phages in metagenomes: what do we catch, what do we miss?. Current opinion in virology, 49, 142-150 . doi.org/10.1016/j.coviro.2021.05.008.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: A flexible trimmer for Illumina sequence data', Bioinformatics, 30(15), pp. 2114–2120. doi:10.1093/bioinformatics/btu170.

Borges, V. et al. (2018) 'INSAFLU: An automated open web-based bioinformatics suite "from-reads" for influenza whole-genome-sequencing-based surveillance', Genome Medicine, 10(1). doi:10.1186/s13073-018-0555-0.

Breitwieser, F.P., Baker, D.N. and Salzberg, S.L. (2018) 'Krakenuniq: Confident and fast metagenomics classification using unique K-Mer counts', Genome Biology, 19(1). doi:10.1186/s13059-018-1568-0.

Buchfink, B., Xie, C. and Huson, D.H. (2014) 'Fast and sensitive protein alignment using diamond', Nature Methods, 12(1), pp. 59–60. doi:10.1038/nmeth.3176.

Cantu, V.A., Sadural, J. and Edwards, R. (2019) Prinseq++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets [Preprint]. doi:10.7287/peerj.preprints.27553.

Castilletti, C. and Capobianchi, M.R. (2023) 'Polio is back? the risk of poliomyelitis recurrence globally, and the legacy of the severe acute respiratory syndrome coronavirus 2 pandemic', Clinical Microbiology and Infection, 29(4), pp. 414–416. doi:10.1016/j.cmi.2022.12.005

Cavadas, J. et al. (2022) 'Mastadenovirus molecular diversity in waste and environmental waters from the Lisbon Metropolitan Area', Microorganisms, 10(12), p. 2443. doi:10.3390/microorganisms10122443.

Condez, A.C. et al. (2021) 'Human polyomaviruses (hpyv) in wastewater and environmental samples from the Lisbon metropolitan area: Detection and genetic characterization of viral structural protein-coding sequences', Pathogens, 10(10), p. 1309. doi:10.3390/pathogens10101309.

Corpuz, M.V. et al. (2020) 'Viruses in wastewater: Occurrence, abundance and detection methods', Science of The Total Environment, 745, p. 140910. doi:10.1016/j.scitotenv.2020.140910.

Deforche, K. (2017) An alignment method for nucleic acid sequences against annotated genomes [Preprint]. doi:10.1101/200394.

Diariodarepublica.pt. Available at: https://diariodarepublica.pt/dr/detalhe/decreto-lei/306-2007-640931 (Accessed: 21 September 2024).

Directive - 2020/2184 - en - EUR-lex (no date) EUR. Available at: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32020L2184 (Accessed: 21 September 2024).

Dobin, A. et al. (2012) 'Star: Ultrafast universal RNA-seq aligner', Bioinformatics, 29(1), pp. 15–21. doi:10.1093/bioinformatics/bts635.

Dutilh, B.E. et al. (2014) 'A highly abundant bacteriophage discovered in the unknown sequences of human fecal metagenomes', Nature Communications, 5(1). doi:10.1038/ncomms5498.

Edwards R.A., et al. (2015). Computational approaches to predict bacteriophage-host relationships. FEMS Microbiology Reviews, 40(2), 258-272. doi:10.1093/femsre/fuv048.

Edwards, R.A. and Rohwer, F. (2005) 'Viral metagenomics', Nature Reviews Microbiology, 3(6), pp. 504–510. doi:10.1038/nrmicro1163.

Fernandez-Cassi, X. et al. (2018) 'Metagenomics for the study of viruses in urban sewage as a tool for public health surveillance', Science of The Total Environment, 618, pp. 870–880. doi:10.1016/j.scitotenv.2017.08.249.

Fontenele, R.S. et al. (2021) 'High-throughput sequencing of SARS-COV-2 in wastewater provides insights into circulating variants', Water Research, 205, p. 117710. doi:10.1016/j.watres.2021.117710.

Fu, L. et al. (2012) 'CD-hit: Accelerated for clustering the next-generation sequencing data', Bioinformatics, 28(23), pp. 3150–3152. doi:10.1093/bioinformatics/bts565.

Hasing, M.E. et al. (2023) 'Wastewater surveillance monitoring of SARS-COV-2 variants of concern and dynamics of transmission and community burden of covid-19', Emerging Microbes & Infections, 12(2). doi:10.1080/22221751.2023.2233638.

Hellmér M., et al. (2014). Detection of pathogenic viruses in sewage provided early warnings of hepatitis A virus and norovirus outbreaks. Applied and Environmental Microbiology, 80(21), 6771–6781. doi:10.1128/aem.01981-14.

Huson, D.H. et al. (2007) 'Megan analysis of Metagenomic Data', Genome Research, 17(3), pp. 377–386. doi:10.1101/gr.5969107.

Itarte, M. et al. (2024) 'Assessing environmental exposure to viruses in wastewater treatment plant and swine farm scenarios with next-generation sequencing and Occupational Risk Approaches', International Journal of Hygiene and Environmental Health, 259, p. 114360. doi:10.1016/j.ijheh.2024.114360.

Kalantar K.L., et al. (2020). IDseq—an open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. GigaScience, 9(10). doi:10.1093/gigascience/giaa111.

Kelmer, G.A.R., Ramos, E.R. and Dias, E.H.O. (2023) 'Coliphages as viral indicators in municipal wastewater: A comparison between the ISO and the USEPA methods based on a systematic literature review', Water Research, 230, p. 119579. doi:10.1016/j.watres.2023.119579.

Kim, D. et al. (2016) 'Centrifuge: Rapid and sensitive classification of metagenomic sequences', Genome Research, 26(12), pp. 1721–1729. doi:10.1101/gr.210641.116.

Ko, B. and Van Raamsdonk, J.M. (2023) 'RNA sequencing of pooled samples effectively identifies differentially expressed genes', Biology, 12(6), p. 812. doi:10.3390/biology12060812.

Langmead, B. and Salzberg, S.L. (2012) 'Fast gapped-read alignment with bowtie 2', Nature Methods, 9(4), pp. 357–359. doi:10.1038/nmeth.1923.

Lapa, D. et al. (2022) 'Monkeypox virus isolation from a semen sample collected in the early phase of infection in a patient with prolonged seminal viral shedding', The Lancet Infectious Diseases, 22(9), pp. 1267–1269. doi:10.1016/s1473-3099(22)00513-8.

Li, D. et al. (2015) 'Megahit: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph', Bioinformatics, 31(10), pp. 1674–1676. doi:10.1093/bioinformatics/btv033.

Li, H. and Durbin, R. (2010) 'Fast and accurate long-read alignment with Burrows–Wheeler transform', Bioinformatics, 26(5), pp. 589–595. doi:10.1093/bioinformatics/btp698.

Luscombe N.M., et al. (2001). What is bioinformatics? An introduction and overview. Yearbook of Medical Informatics, 83-100.

Martínez-Puchol S., et al. (2020). Characterisation of the sewage virome: comparison of NGS tools and occurrence of significant pathogens. Science of the Total Environment, 713. doi:10.1016/j.scitotenv.2020.136604.

Maryam, S. et al. (2023) 'Covid-19 surveillance in wastewater: An epidemiological tool for the monitoring of SARS-COV-2', Frontiers in Cellular and Infection Microbiology, 12. doi:10.3389/fcimb.2022.978643.

McIntyre, A.B. et al. (2017) 'Comprehensive benchmarking and Ensemble Approaches for Metagenomic classifiers', Genome Biology, 18(1). doi:10.1186/s13059-017-1299-7.

McKee, A.M. and Cruz, M.A. (2021) 'Microbial and viral indicators of pathogens and human health risks from recreational exposure to waters impaired by fecal contamination', Journal of Sustainable Water in the Built Environment, 7(2). doi:10.1061/jswbay.0000936.

Moeckel, C. et al. (2024) 'A survey of K-Mer methods and applications in bioinformatics', Computational and Structural Biotechnology Journal, 23, pp. 2289–2303. doi:10.1016/j.csbj.2024.05.025.

Monteiro, S. et al. (2024) 'Detection of dengue virus and chikungunya virus in wastewater in Portugal: An exploratory surveillance study', The Lancet Microbe, p. 100911. doi:10.1016/s2666-5247(24)00150-2.

Morales-Cortés, S. et al. (2024) 'Crass-like phages are suitable indicators of antibiotic resistance genes found in abundance in fecally polluted samples', Environmental Pollution, 359, p. 124713. doi:10.1016/j.envpol.2024.124713.

Nooij S., et al. (2018). Overview of Virus Metagenomic Classification Methods and Their Biological Applications. Frontiers in Microbiology, 9, 749. doi:10.3389/fmicb.2018.00749.

Ounit, R. et al. (2015) 'Clark: Fast and accurate classification of metagenomic and genomic sequences using discriminative K-MERS', BMC Genomics, 16(1). doi:10.1186/s12864-015-1419-2

Paez-Espino, D. et al. (2016) 'Uncovering Earth's virome', Nature, 536(7617), pp. 425–430. doi:10.1038/nature19094.

Parras-Moltó, M. et al. (2018) 'Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses', Microbiome, 6(1). doi:10.1186/s40168-018-0507-3.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. Nature Biotechnology, 35(9), 833-844. doi:10.1038/nbt.3935.

Ray, K.J. et al. (2019) 'High-throughput sequencing of pooled samples to determine community-level microbiome diversity', Annals of Epidemiology, 39, pp. 63–68. doi:10.1016/j.annepidem.2019.09.002.

Ren, J. et al. (2017) 'Virfinder: A novel K-mer based tool for identifying viral sequences from assembled metagenomic data', Microbiome, 5(1). doi:10.1186/s40168-017-0283-5.

Rodrigues, F. V. Assessment of jc virus in portuguese wastewaters: impact on public health. (Faculdade de Farmácia da Universidade de Coimbra, 2015).

Rosseel, T. *et al.* (2012) 'DNase Sispa-next generation sequencing confirms Schmallenberg virus in Belgian field samples and identifies genetic variation in Europe', *PLoS ONE*, 7(7). doi:10.1371/journal.pone.0041967.

Roux, S. et al. (2015) 'Virsorter: Mining viral signal from microbial genomic data', PeerJ, 3. doi:10.7717/peerj.985.

Ruby, J.G., Bellare, P. and DeRisi, J.L. (2013) 'Price: Software for the targeted Assembly of components of (Meta) genomic sequence data', G3 Genes|Genomes|Genetics, 3(5), pp. 865–880. doi:10.1534/g3.113.005967.

Sabar, M.A., Honda, R. and Haramoto, E. (2022) 'Crassphage as an indicator of human-fecal contamination in water environment and virus reduction in wastewater treatment', Water Research, 221, p. 118827. doi:10.1016/j.watres.2022.118827.

Schmieder, R. and Edwards, R. (2011) 'Quality Control and preprocessing of metagenomic datasets', Bioinformatics, 27(6), pp. 863–864. doi:10.1093/bioinformatics/btr026.

Andrews, S. (2010) FastQC: A Quality Control Tool for High Throughput Sequence Data [Online], Babraham Bioinformatics - FastQC a quality control tool for high throughput sequence data. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed: 23 September 2024).

Torres C., et al. (2016). High diversity of human polyomaviruses in environmental and clinical samples in Argentina: Detection of JC, BK, Merkel-cell, Malawi, and human 6 and 7 polyomaviruses. Science of The Total Environment, 542, 192–202. doi:10.1016/j.scitotenv.2015.10.047.

Tseemann (2020) Tseemann/snippy: :scissors: Rapid haploid variant calling and core genome alignment, GitHub. Available at: https://github.com/tseemann/snippy (Accessed: 13 November 2023).

Van Doorslaer, K. and Burk, R.D. (2010) 'Evolution of human papillomavirus carcinogenicity', Advances in Virus Research, pp. 41–62. doi:10.1016/b978-0-12-385034-8.00002-8.

Vázquez-Castellanos, J.F. et al. (2014) 'Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut', BMC Genomics, 15(1). doi:10.1186/1471-2164-15-37.

Vilsker, M. et al. (2018) 'Genome detective: An automated system for virus identification from high-throughput sequencing data', Bioinformatics, 35(5), pp. 871–873. doi:10.1093/bioinformatics/bty695.

Wood, D.E., Lu, J. and Langmead, B. (2019) 'Improved metagenomic analysis with Kraken 2', Genome Biology, 20(1). doi:10.1186/s13059-019-1891-0.

Wu, T.D. and Nacu, S. (2010) 'Fast and SNP-tolerant detection of complex variants and splicing in short reads', Bioinformatics, 26(7), pp. 873–881. doi:10.1093/bioinformatics/btq057.

Wu, T.D. and Watanabe, C.K. (2005) 'GMAP: A genomic mapping and alignment program for mrna and EST sequences', Bioinformatics, 21(9), pp. 1859–1875. doi:10.1093/bioinformatics/bti310.

Wu, Z. et al. (2020) 'Comparative fate of crassphage with culturable and molecular fecal pollution indicators during activated sludge wastewater treatment', Environment International, 136, p. 105452. doi:10.1016/j.envint.2019.105452.

Zhao, Y., Tang, H. and Ye, Y. (2011) 'Rapsearch2: A fast and memory-efficient protein similarity search tool for next-generation sequencing data', Bioinformatics, 28(1), pp. 125–126. doi:10.1093/bioinformatics/btr595.

Zhou, X. and Rokas, A. (2014) 'Prevention, diagnosis and treatment of high-throughput sequencing data pathologies', Molecular Ecology, 23(7), pp. 1679–1700. doi:10.1111/mec.12680.

A

APPENDIX A

A.1 Experimental Protocol for viral concentration and extraction

Viral-like particles were concentrated using organic flocculation with pre-flocculated powdered milk solution (1% (m/v)), where 10 g of powdered milk (Conda Pronadisa, Madrid, Spain) was dissolved in 1 L of artificial sea water (Sigma Aldrich St. Louis, Missouri, EUA), with the pH adjusted to 3.5 using 1M HCL. Ten mL and 100 mL of the previous solution were added to 1 L and 10 L wastewater samples, respectively (final concentration of powdered milk at 0.01% (m/v)), and the pH was adjusted to 3.5 with 1M HCL. Subsequently, the water samples were kept under constant stirring for 8 hours at room temperature, and then for another 8 hours without stirring for the flocculation sediments to settle. The supernatants were carefully removed using a vacuum pump so as not to disturb the sediment. This sediment (about 500 mL) was then transferred to a centrifuge bottle and concentrated by centrifugation at 5500 RCF for 45 minutes at 12 °C in the J-26 Avanti® XPI Beckman Coulter centrifuge. Afterwards, the supernatant was discarded and the concentrate, which contains the viral particles, was resuspended in 8 mL of 0.2 M phosphate buffer with a pH of 7.5 (1:2 (v/v) mixture of 0.2 M Na2HPO4 and 0.2 M NaH2PO4). The concentrate was stored at -20 °C until further use. RNA extraction was carried out from 140 µl of the concentrated sample with QIAamp® Viral RNA Mini Kit (QIAGEN®, USA), following the manufacturer's protocol. In summary, each sample was lysed under denaturing conditions to inactivate RNAses and ensure the isolation of viral RNA. The conditions of the various buffers were adjusted to allow the RNA to bind to the extraction column membrane. After this binding, contaminants were removed in two stages with different washing buffers, and the RNA was eluted in an RNase-free buffer, becoming free of proteins, nucleases, other contaminants, and inhibitors. RNA was quantified and quality ratios were obtained for each extraction using NanodropOne.

A.2 Complete summary table (tables_samples dataframe)

Sample	Raw_reads	Num of all reads	Num of viral reads	Num of non viral reads	Num of non phage reads	Num of phage reads
A1_CZID	3209246	40058	7370	31173	4687	2683
A1_GD	3209246	5317	5317	0	1259	4058
A1_INSA	3209246	97240	97235	0	96518	717
A1_KR	3209246	1759	1759	0	236	1523
A2_CZID	4990332	19785	1442	17755	506	936
A2_GD	4990332	1277	1277	0	534	743
A2_INSA	4990332	15182	15182	0	15148	34
A2_KR	4990332	67	67	0	58	9
A3_CZID	3396360	20650	1166	19166	595	571
A3_GD	3396360	662	662	0	322	340
A3_INSA	3396360	4079	4079	0	4039	40
A3_KR	3396360	117	117	0	53	64
A5_CZID	200942	33810	153	32531	133	20
A5_GD	200942	122	122	0	40	82
A5_INSA	200942	2116	2116	0	2107	9
A5_KR	200942	24132	41	0	9	32
B1_CZID	6827732	54851	9619	43071	6052	3567
B1_GD	6827732	7429	7403	0	1922	5481
B1_KR	6827732	2323	2323	0	309	2014
B2_CZID	1602524	6876	2561	3480	486	2075
B2_GD	1602524	368	368	0	51	317
B2_INSA	1602524	13519	13519	0	13466	53
B2_KR	1602524	32	32	0	17	15
B3_CZID	959348	2078	347	1627	162	185
B3_GD	959348	197	197	0	67	130
B3_INSA	959348	2335	2335	0	2306	29
B3_KR	959348	36	36	0	13	23
B5_CZID	1253304	65712	1819	52526	584	1235
B5_GD	1253304	1380	1380	0	460	920
B5_INSA	1253304	638	638	0	627	11
B5_KR	1253304	347885	287	0	192	95
C1_CZID	1320678	59427	13506	43011	8892	4614
C1_GD	1320678	6040	6040	0	2951	3089
C1_INSA	1320678	201201	201198	0	200220	978
C1_KR	1320678	1787	1787	0	444	1343
C2_CZID	887420	12629	1196	11082	293	903
C2_GD	887420	114	114	0	41	73
C2_INSA	887420	10628	10628	0	10601	27
C2_KR	887420	33	33	0	9	24
C3_CZID	829616	3020	714	2168	300	414
C3_GD	829616	282	282	0	142	140
C3_INSA	829616	1081	1081	0	1072	9
C3_KR	829616	49	49	0	21	28
C5_CZID	294302	1964	60	1592	15	45
C5_GD	294302	14	14	0	0	14
C5_INSA	294302	2690	2690	0	2690	0
C5_KR	294302	34106	10	0	6	4

D1_CZID	924716	61909	15366	42373	2794	12572
D1_GD	924716	288464	288464	0	828	287636
D1_INSA	924716	13519	13519	0	13466	53
D1_KR	924716	158295	158295	0	154	158141
D2_CZID	844362	7651	2319	4474	401	1918
D2_GD	844362	995	995	0	79	916
D2_INSA	844362	9082	9082	0	9043	39
D2_KR	844362	44	44	0	25	19
D3_CZID	1404672	1385	748	487	401	347
D3_GD	1404672	329	329	0	247	82
D3_INSA	1404672	337	337	0	327	10
D3_KR	1404672	69	69	0	48	21
D5_CZID	1311630	72252	1705	54914	640	1065
D5_GD	1311630	1170	1170	0	227	943
D5_INSA	1311630	553	553	0	533	20
D5_KR	1311630	327212	151	0	106	45
E1_CZID	829822	18262	3081	14031	884	2197
E1_GD	829822	1982	1982	0	558	1424
E1_INSA	829822	21076	21076	0	20986	90
E1_KR	829822	105	105	0	49	56
E2_CZID	299056	5636	2897	1930	1628	1269
E2_GD	299056	1492	1492	0	823	669
E2_INSA	299056	3032	3032	0	3008	24
E2_KR	299056	516	516	0	207	309
E3_CZID	458342	2373	1061	1109	537	524
E3_GD	458342	439	439	0	327	112
E3_INSA	458342	814	814	0	788	26
E3_KR	458342	66	66	0	51	15
E5_CZID	132416	448	5	420	5	0
E5_GD	132416	4	4	0	2	2
E5_INSA	132416	46	46	0	46	0
E5_KR	132416	19738	4	0	4	0
F1_CZID	1532474	9964	2109	7188	285	1824
F1_GD	1532474	758	758	0	47	711
F1_INSA	1532474	20299	20299	0	20183	116
F1_KR	1532474	82	82	0	13	69
F2_CZID	1573236	3132	1718	1012	1084	634
F2_GD	1573236	764	764	0	552	212
F2_INSA	1573236	2150	2150	0	2123	27
F2_KR	1573236	206	206	0	80	126
F3_GD	423008	172	172	0	106	66
F3_INSA	423008	2335	2335	0	2306	29
F3_KR	423008	145	145	0	55	90
F5_CZID	3733250	661278	6440	589078	5156	1284
F5_GD	3733250	3103	3103	0	2464	639
F5_INSA	3733250	1545	1545	0	1481	64
F5_KR	3733250	1268878	1104	0	874	230
G1_CZID	776692	14341	2949	10503	1017	1932
G1_GD	776692	1262	1262	0	544	718
G1_INSA	776692	70403	70403	0	16802	53601
G1_KR	776692	72	72	0	33	39
G2_CZID	279222	3513	1373	1891	1001	372
G2_GD	279222	560	560	0	406	154
G2_INSA	279222	5033	5033	0	5019	14

G2_KR	279222	141	141	0	73	68
G3_CZID	477456	2338	821	1232	291	530
G3_GD	477456	286	286	0	189	97
G3_INSA	477456	638	637	0	622	15
G3_KR	477456	51	51	0	33	18
G5_CZID	508704	2080	55	1919	49	6
G5_GD	508704	27	27	0	27	0
G5_INSA	508704	809	809	0	809	0
G5_KR	508704	3	3	0	3	0
H1_CZID	1899016	8794	1540	6696	314	1226
H1_GD	1899016	886	886	0	351	535
H1_INSA	1899016	12427	12427	0	12373	54
H1_KR	1899016	34	34	0	15	19
H2_CZID	414040	805	51	754	49	2
H2_GD	414040	226	226	0	26	200
H2_INSA	414040	1498	1498	0	1498	0
H2_KR	414040	3	3	0	1	2
H3_CZID	644258	7443	2171	4727	1127	1044
H3_GD	644258	1239	1239	0	760	479
H3_INSA	644258	2253	2253	0	2235	18
H3_KR	644258	159	159	0	120	39
H5_CZID	1667124	529036	19636	488942	13740	5896
H5_GD	1667124	10679	10677	0	6501	4176
H5_INSA	1667124	1168	1168	0	1155	13
H5_KR	1667124	2614	2614	0	2070	544

UNIVERSIDADE NOVA
DE LISBOA

2024

Application and development of viral detection tools for metavirome analysis of wastewaters

ANDRÉ FILIPE BRUNO DE OLIVEIRA TOMÁS DOS SANTOS

MASTER IN **COMPUTATIONAL BIOLOGY & BIOINFORMATICS**

SPECIALIZATION Multi-Omics for Life and Health Sciences

UNIVERSIDADE NOVA DE LISBOA