Albert-Ludwigs University Freiburg
Department of Computer Science
Bioinformatics Group

Master Thesis

Development and evaluation of Galaxy pipelines for detection of SARS-CoV-2 variants by genomic analysis of wastewater samples

Author:
Polina Polunina

Examiner:
Prof. Dr. Rolf Backofen

Second Examiner:
Prof. Dr. Wolfgang R. Hess

Advisors:
Dr. Bérénice Batut, Dr. Wolfgang Maier

Submission date:
07.11.2022

Declaration

I hereby declare that I am the sole author and composer of my thesis and that no other sources or learning aids, other than those listed, have been used. Furthermore, I declare that I have acknowledged the work of others by providing detailed references of said work.
I hereby also declare that my Thesis has not been prepared for another examination or assignment, either wholly or excerpts thereof.

Place, Date

Signature

Acknowledgments

I am grateful so much to Prof. Dr. Rolf Backofen for giving me the great opportunity to work on this master thesis in his Bioinformatics Group at the University of Freiburg. I thank him for his lectures that gave me a basis to start working on my master thesis. I am also thankful for giving me the opportunity to have a hiwi job in the Bioinformatics group.

I would like to thank Prof. Dr. Wolfgang R. Hess for agreeing to be my second examiner.

Words cannot express my gratitude to Bérénice Batut, my supervisor. She was absolutely dedicated to my work on this thesis, and her contribution is invaluable. Her continuous guidance, advice, code review, proofreading, as well as regular feedback led to the completion of this thesis. In addition to academic support, she also provided me with moral support and motivation, which was crucial in some points of this work. With her supervision, I felt extremely comfortable, believed in myself, and never gave up. I am glad to be thankful to Bérénice for giving me the opportunity to have a hiwi job in the Freiburg Galaxy Team that supported me financially.

I am very thankful to my supervisor Wolfgang Maier. He was always willing to answer my academic questions. His knowledge regarding the topic of my master thesis is so immense that I sometimes got lost, but his explanations were always very understandable. My sincere gratitude goes out to him for his contribution not only with knowledge but also with guidance and ideas throughout the research. It was a pleasure to work with him.

I want to thank everyone from the Freiburg Galaxy Team with their great head Björn Grüning. I was treated with such kindness and helpfulness by them while writing my thesis. In addition to our lunches together, I appreciate our time together during the workdays in the lab. Special thanks are to Cristo and Engy, who regularly checked on my condition, and Mira and Simon, who ensured me always being invited to team activities.

I am thankful to Teresa Müller for allowing me to have my first hiwi contract and work on an amazing interactive game development project for Street Science Community that involved me in a bio-related topic. This thesis would not have happened without this involvement.

Thanks to Siyu and Yedil, my Street Science Community mates, that never were greedy to cheer me up, which influenced my master project positively.

Many thanks to my study colleagues and friends from the Computer Science program (and not only) at the University of Freiburg for being friendly and supportive. Special thanks are to Karim, Tidiane, Robin, and Roberto for our study sessions in the library, recommendations, and taking care that I do not lose concentration on my thesis and do not burn out.

I want to thank Oleg Aleksandrov, my supervisor at Ural Federal University, and Julia Budchenko, head of the RT Labs affiliate in Yekaterinburg at the time when I worked there. Without their belief in me and their reference, I would not have come to the University of Freiburg.

It is especially important to me to thank my friend Taya for being proud of me and supporting me financially during the hard times I met while writing my master thesis.

Last but not least, I must mention my parents, my sister Olga, and dogs Teddy and Filya. Although they were far away, they supported me financially and morally.

Abstract

More than 630 million people have been affected by the COVID-19 pandemic nearly two years after the first report of SARS-CoV-2 in Wuhan, China. During the SARS-CoV-2 pandemic, wastewater surveillance has received extensive public attention as a passive monitoring system that complements clinical and genomic surveillance. The detection and quantification of viral RNA in wastewater samples are already possible through several methods and protocols, and viral RNA concentrations in wastewater have been shown to correlate with reported cases.

The Galaxy community has put much effort into a continuous analysis of intra-host variation in SARS-CoV-2, including the development of workflows, on samples of individuals.

In this master thesis, there were investigated existing Galaxy workflows for clinical SARS-CoV-2 data analysis, as well as existing approaches to wastewater surveillance of SARS-CoV-2. Two existing Galaxy workflows and, on the other hand, two existing wastewater surveillance methods (Freyja and COJAC) were used to develop Galaxy pipelines for SARS-CoV-2 wastewater data analysis. The resulting workflows were tested on synthetic and real-world datasets to evaluate efficiency in finding SARS-CoV-2 lineages abundances. The developed workflows were additionally compared with one state-of-the-art approach called Lineagespot.

Results of the Galaxy workflows developed using Freyja and COJAC approaches are encouraging. Developed Galaxy workflows with Freyja showed promising results in detecting expected lineages when single lineage or two lineages were expected. However, the results were biased by detecting other unexpected lineages. Galaxy workflow with COJAC implemented for delineation showed slightly more efficient results as it was able to detect only lineages that were expected in some samples.

Contents

1 Introduction

More than 630 million people have been affected by the COVID-19 pandemic almost 3 years after the first report of SARS-CoV-2 in Wuhan, China. During the COVID-19 pandemic, wastewater surveillance has received extensive public attention as a passive monitoring system that complements clinical and genomic surveillance. The detection and quantification of viral RNA in wastewater samples are already possible through several methods and protocols, and viral RNA concentrations in wastewater have been shown to correlate with reported cases. In the Netherlands, for example, which has had a nationwide wastewater monitoring network for decades, scientists found that fragments of the SARS-CoV-2 virus can accurately reflect its level in the community [4]. The correlation was also shown by tracking wastewater samples in Sweden as well as in Switzerland [5, 6].

The Galaxy community has put much effort into a continuous analysis of intra-host variation in SARS-CoV-2 (`https://galaxyproject.org/projects/covid19/`) [7, 8, 9, 10], including the development of workflows, on samples of individuals.

The purpose of this thesis is an improvement of existing workflows in Galaxy to support the analysis of wastewater data.

1.1 SARS-CoV-2

1.1.1 COVID-19 and pandemic

Coronavirus disease 2019 (COVID-19), a contagious disease, is caused by the severe acute raspiratory syndrome coronavirus 2 (SARS-CoV-2). The first reported case occurred in Wuhan, China, in December 2019 [11], which rapidly evolved into a pandemic.

A public health emergency of international concern was declared by the World Health Organization (WHO) on 30 January 2020 [12], and a pandemic on 11 March 2020 [13, 14]. More than 609 million cases have been reported, and more than 6.51 million deaths have been confirmed as of 14 September 2022 [15].

There are a range of symptoms associated with COVID-19, including fever [16], cough, headache [16], fatigue, breathing difficulties, loss of smell [17, 18], and loss of taste [18, 19]. Virus infection may cause symptoms one to fourteen days after exposure. Infected people develop no noticeable symptoms in at least a third of cases [20]. It has been shown that most people (81%) with symptoms noticeable enough to qualify as patients develop mild to moderate symptoms, while 14% develop severe symptoms (dyspnoea, hypoxia, or more than 50% lung involvement on imaging), and 5% develop critical symptoms (respiratory failure, shock, or multiorgan dysfunction) [21]. There have been reports of damage to organs in people who have suffered a traumatic brain injury that continues months after they have recovered [22].

COVID-19 spreads when airborne droplets and small particles containing the virus are inhaled. Proximity increases the risk of breathing these. In rare cases, contact with contaminated surfaces can also lead to transmission if splashed or sprayed with contaminated fluids. Even if no symptoms appear, people can spread the virus for up to 20 days [23].

1.1.2 SARS-CoV-2 virus

Initially, SARS-CoV-2 was isolated from three people with pneumonia connected to an acute respiratory illness cluster in Wuhan [24]. SARS-CoV-2 virus particles exhibit all structural features found in related coronaviruses in nature [25].

SARS-CoV-2 is closely related to the original SARS-CoV [26]. An animal origin is suspected [27]. Phylogenetic analysis shows that the Coronavirus clusters with the subgenus Sarbecovirus (lineage B) and two bat-derived strains in the genus Betacoronavirus.

It is 96% identical at the whole genome level to other bat coronavirus samples (BatCov RaTG13) [28]. Several structural proteins are associated with SARS-CoV-2, including membrane glycoproteins (M), envelope proteins (E), nucleocapsid proteins (N), and spike proteins (S). There is about 98% homology between the M protein of SARS-CoV-2 and the M protein of bat SARS-CoV, around 98% homology with pangolin SARS-CoV, and 90% homology with the M protein of SARS-CoV; however, the similarity is only 38% with the M protein of MERS-CoV [29].

1.1.3 SARS-CoV-2 mutations

As all viruses mutate, so does SARS-CoV-2. Viruses reproduce (or make copies of themselves) inside host cells and occasionally mutate, e.g., changes appear in the string of 30,000 bases that make up the SARS-CoV-2 genome. These mutations may lead to changes in the amino acids that make up a protein, which would alter its structure. A virus that has been mutated may weaken or die out as a result of these changes. In rare cases, mutations can improve the virus's ability to cause disease, spread, or evade host immunity. If the mutations spread with virus replication and invade the virus population, the corresponding virus can become a "variant" of its wild-type (original) by accumulating advantageous mutations, making it a better infector, disease-causing agent, evading the immune system, or spreading within a population [30]. With over 485 million documented infections during the past two years, and probably many more undocumented, there have been an incredible number of mutation opportunities. Nevertheless, most of them do not cause global surges and do not pose any more significant threat than existing variants.

In the first year of the Coronavirus pandemic, the virus did not change much. Approximately one or two mutations were picked up by SARS-CoV-2 each month. The phylogenetic tree had only one main trunk and a few tiny branches (fig. 1). Pandemic course changed in 2020.

1.1.4 SARS-CoV-2 variants

From approximately summer of 2020 on, SARS-CoV-2's family tree grew increasingly complex. The main trunk sprouted branches for a number of variants. Gamma, lambda, and mu variants appeared (although none spread worldwide). The tree's canopy was formed by dozens of branches (fig. 2). Researchers track the SARS-CoV-2 variants with mutations that are clinically or epidemiologically significant. Detecting variants in a virus usually requires sequencing its complete genome. By sequencing a representative sample of viral specimens in a population, one can determine if new variants are emerging or existing ones are spreading. In particular, variants with the potential or demonstrated ability to be more transmissible, immune evasive,

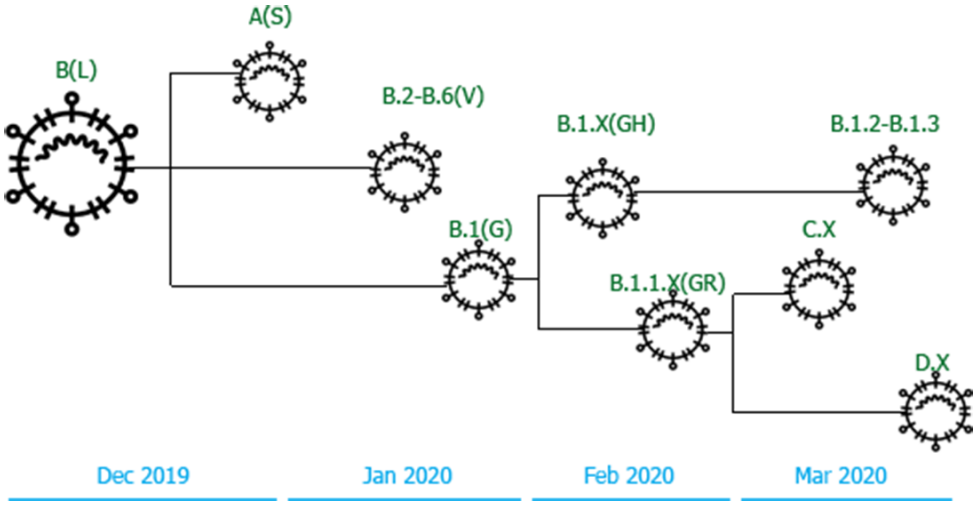


Figure 1: Schematic representation of the major evolutionary events that gave rise to SARS-CoV-2 variants in sequential order (simplified, cf. [31]).

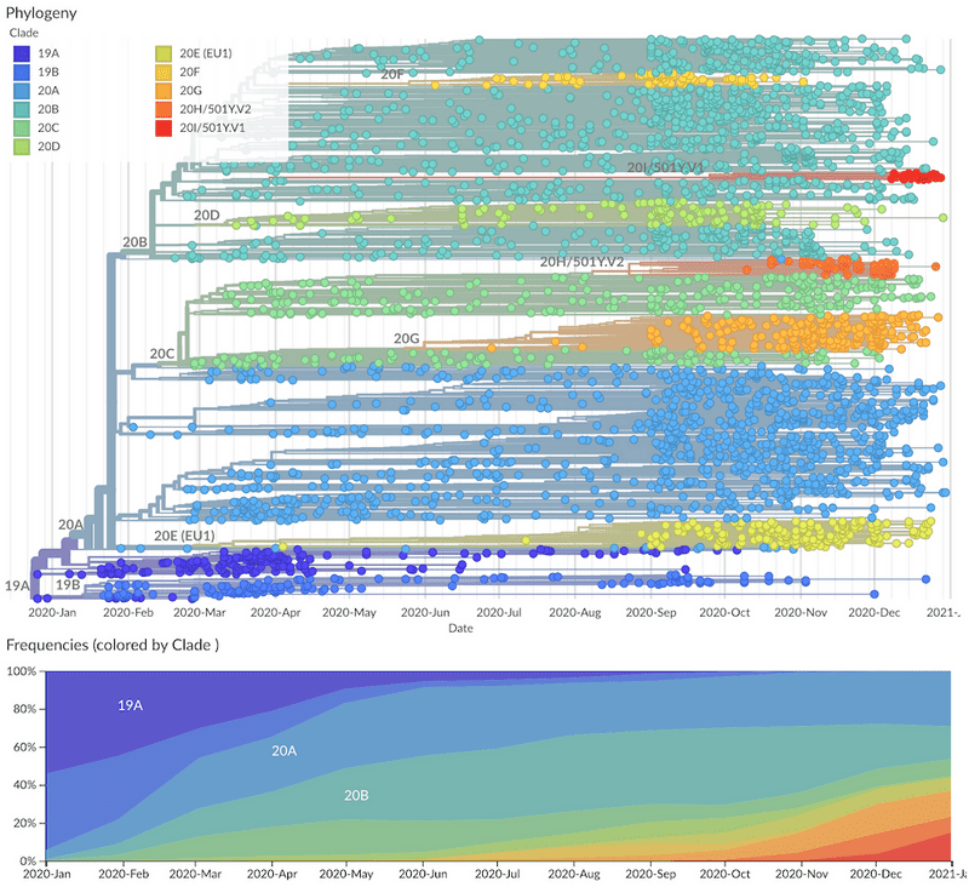


Figure 2: SARS-CoV-2 phylogenetic tree with a timeline from Nextstrain project labeled by clades (image from [32]).

or virulent are closely monitored. These are designated as variants of interest (VOIs) or variants of concern (VOCs).

On-going pandemic control efforts are challenged by the emergence of SARS-CoV-2 variants with greater transmission potential and/or immunity circumvention. In 2020, researchers from South Africa and Great Britain identified two variants of concern [33]: alpha and beta [34]. The virus changed rapidly due to these two variants. According to estimates, each variant had around 20 mutations, compared to a few mutations in previous variants. As a result, they became VOCs instead of VOIs.

The Delta (B.1.617.2) variant was dominant worldwide between June and December 2021 (fig. 3), it has 13 mutations and over 200 known sublineages. Compared to the original virus sequenced in January 2020, the Omicron variant was found to have 50 mutations, when sequenced for the first time in November 2021, and has more than 300 sublineages by now, according to Pango Lineages database `https://cov-lineages.org/`. In the middle of 2020, once Omicron's lineage diverged, it became phylogenetically distinct from other known VOCs or VOIs [35]. However, the Omicron variant reached its peak worldwide in January 2022 and continues to be prevalent compared to other variants of concern.

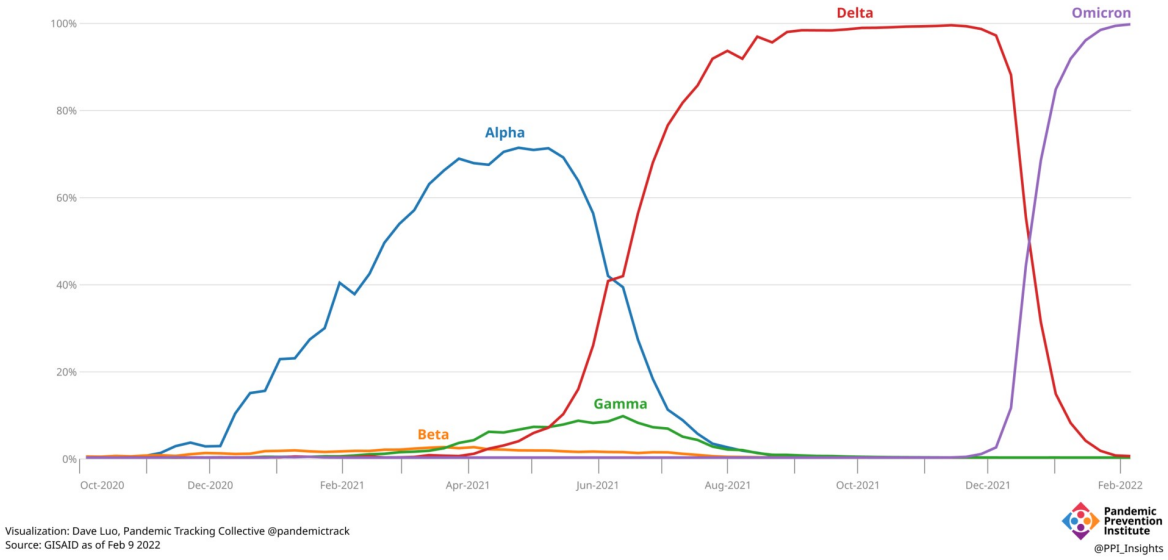


Figure 3: Chart of global SARS-CoV-2 VOC growth from 4 October, 2020, to 21 March, 2022. Graph represents proportions of SARS-CoV-2 genome sequences on GISAID by sample collection week per VOC (image from [30]).

1.1.5 SARS-CoV-2 naming systems

Variants were named with different letters and numbers by different groups of scientists, which can be confusing to the general public. As an example, in the beginning of pandemic, many news organizations and others addressing non-scientific audiences had simplified the naming of variants by referring to the countries in which they first appeared, but this could have resulted in stigma, blame, or prejudice. Therefore, The World Health Organization [36], around 2.5 years

ago, proposed using Greek alphabet letters to label variants of concern and interest in an effort to make their identification easier to pronounce and less stigmatizing. WHO has designated five variants as VOCs and given them the Greek letter names Alpha, Beta, Gamma, Delta, and Omicron (table 1).

WHO	Pango	GISAID	Nextstrain	Earliest date
Alpha	B.1.1.7	GRY	20I (V1)	September 2020
Beta	B.1.351	GH/501Y.V2	20H (V2)	May 2020
Gamma	P.1	GR/501Y.V3	20J (V3)	November 2020
Delta	B.1.617.2 and AY Sublineages	G/478K.V1	21A,21I,21J	October 2020
Omicron	B.1.1.529 and BA Sublineages	GR/484A	21K,21L,21M	November 2021

Table 1: Designated variants of concern named by different naming systems

Based on genomics experts' rules, several commonly used naming systems [37] classify the evolving virus forms. The most widely used naming systems are: i) Pango [35], which uses the nomenclature system and rules outlined in the publication by A. Rambaut et al. [38] and is maintained by the Pango Network (`http://pango.network`), and ii) Nextstrain that initially uses 'year-letter' naming. Nextstrain nomenclature was first used to monitor and document the Ebola epidemic in West Africa in 2013-2016 and the Zika outbreak in America in 2018. A number of components make up Nextstrain: Python scripts maintain a database of sequences and metadata, sourced from public databases such as NCBI (`www.ncbi.nlm.nih.gov`), GISAID (`www.gisaid.org`) and ViPR (`www.viprbrc.org`), as well as GitHub repositories and other genomic data sources. There is a suite of tools available for phylodynamic analysis [39], including subsampling, alignment, phylogenetic inference, and the reconstruction of discrete trait geographic patterns, including the estimation of the probabilities of transmission [40]. As soon as SARS-CoV-2 genomes are shared, Nextstrain incorporates them and provides analyses and reports once they are available. Nextstrain SARS-CoV-2 naming strategy is in sync with Pango. As an example, Nextstrain clade 22A corresponds to Pango lineage BA.4, 22B to BA.5, and 22C to BA.2.12.1 [41]. There may be some differences in the classification of viruses across or within lineages or clades, although the major lineages and clades are generally assigned in a similar manner.

1.2 Surveillance of SARS-CoV-2

In order to respond effectively to SARS-CoV-2, global surveillance is essential for determining which variants require closer monitoring as possible threats to public health. Health professionals and policymakers should have up-to-date information about viral populations present in communities, particularly as new SARS-CoV-2 variants emerge that alter viral fitness and/or pathogenesis.

A variety of surveillance techniques are available for SARS-CoV-2. Clinical testing was the most potent and widespread method at the beginning of the pandemic. Different types of clinical testing are possible now. For example, diagnostic testing is one of the clinical testing types and is intended to identify current infections in individuals who have symptoms consonant with COVID-19 and/or have recently been exposed to SARS-CoV-2. The other type of clinical testing - screening testing - is used for identifying asymptomatic individuals with COVID-19 cases who don't have known, suspected, or reported exposure to SARS-CoV-2 [42].

In addition to clinical testing, public health surveillance entails the systematic collection, analysis, and interpretation of health-related data that are essential for planning, implementing, and evaluating public health practices. An important purpose of surveillance testing in public health is to monitor outbreaks of disease at the community and population level, as well as to characterize disease incidence and prevalence. De-identified specimens are used for surveillance testing, so results are not linked to individuals. To determine if the prevalence of viruses is increasing or decreasing in a particular population, a certain percentage of the population may be sampled for public health surveillance testing [42].

A distinct type of public health surveillance testing is the surveillance of wastewater by high-throughput sequencing. By monitoring wastewater in a community, one can get a better understanding of the types and amounts of viruses and bacteria that are spreading. The virus RNA in stools of people with current or recent SARS-CoV-2 infections is detectable in wastewater samples even if they aren't symptomatic. The cost-effectiveness of wastewater sampling comes from the fact that a single sample can reveal information about an entire building, town, or county. It has been suggested that wastewater surveillance could also be used to detect emerging viral variants in an area as they emerge, in addition to tracking SARS-CoV-2 transmission levels.

Following subsections will generally discuss the process of lineage identification from isolates with a focus on clinical data surveillance techniques. Starting with extraction approaches (e.g. ampliconic-based, metatranscriptomics, etc.) that are used for the extraction of samples with SARS-CoV-2, sequencing techniques will be described in section 1.2.1. It should be said that talking about extraction methods focus will be done on metatranscriptomics and ampliconic-based approaches because these two library preparation techniques are used more often for SARS-CoV-2 datasets, and in this thesis, datasets extracted with the help of these two methods are processed and analyzed. After that, bioinformatics steps of data processing will be discussed. Section 1.2.2 will represent existing tracking projects for clinical data surveillance, following by solutions suggested by Galaxy Project for clinical SARS-CoV-2 data in section 1.2.3.

1.2.1 Methods of SARS-CoV-2 identification

1.2.1.1 Extraction approaches

Choosing the most appropriate extraction strategy depends on the final objectives and the type of biological sample. Currently, four different concepts of library preparation have been applied to SARS-CoV-2 data: i) shotgun metatranscriptomics, ii) hybrid capture enrichment, iii) amplicon-based, and iv) direct RNA sequencing. To make a comparison between library preparation strategies, table 2 shows the main characteristics.

Most of SARS-CoV-2 sequence data is generated using two established library preparation strategies: ampliconic-based sequencing approach and metatranscriptomics sequencing approach. These two approaches will be discussed in more detail in the following two sections.

1.2.1.1.1 Metatranscriptomics approach

In this method, RNA fragmentation is followed by first- and second-strand cDNA synthesis, and then library preparation is carried out based on the high-throughput sequencing techniques

	Shotgun metatranscriptomics	**Amplicon-based**	**Hybrid capture-enrichment**	**Direct RNA sequencing**
Target	SARS-CoV-2, host microbiota, host response to infection	SARS-CoV-2 sequence	SARS-CoV-2 sequence	SARS-CoV-2, host transcriptome, epitranscriptome
Co-infection detection	Yes	No	No/yes (depending on gene panel)	Yes
Min number of reads	20–50 M	5–20 M	5–20 M	0.5 M
Genome Coverage	$\geq 99\%$	$\geq 95 - 99\%$	$\geq 95 - 99\%$	$\geq 99\%$
Accuracy in SNV identification	High	High	Moderate	Moderate
Sample viral load (Ct) requested (ref Xiao)	<24–28	$\geq 24 - 28$	$\geq 24 - 28$	<24–28
Sample RNA input (ng)	10–200	1–50	10–50	≥ 1000
Sample type	Patient specimens	Patient specimens, environmental samples	Patient specimens, environmental samples	Viral cell cultures
Cost	High	Low	Moderate	High
NGS platforms	High- or ultra high-throughput platforms	Mid-throughput platforms	Mid- or high-throughput platforms	ONT

Table 2: Comparison of SARS-CoV-2 library preparation approaches

of choice [43]. Metatranscriptomics technique is able to detect all types of pathogens, which is beneficial. A complete or nearly complete viral genome can be sequenced and reconstructed from available SARS-CoV-2 sequence data that have enough viral loads. However, most preparation protocols involve multiple costly and complicated steps, which inevitably compromises the effectiveness of sequencing of time-sensitive SARS-CoV-2 samples and in identifying other pathogens [44].

Figure 4 shows a simplified schematic representation of the common process for SARS-CoV-2 RNA extraction with metatranscriptomics approach.

In short, the typical workflow of shotgun metatranscriptomics involves breaking RNA into fragments, synthesizing cDNA from the first and second strands, and preparing a library using the

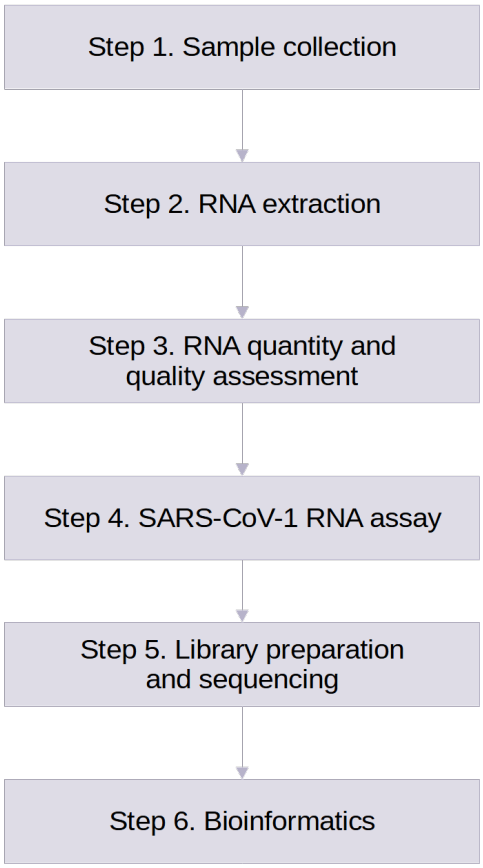


Figure 4: Schematic representation of the common process for SARS-CoV-2 RNA extraction with metatranscriptomics approach (simplified, cf. [43]).

chosen next-generation sequencing technology.

1.2.1.1.2 Ampliconic approach

Amplicon-based method is a highly targeted approach for analyzing genetic variation in specific genomic regions. This extraction method provides information about the targeted sequence. If the targeted sequences are rich in lineage-defining polymorphisms, polymorphisms in the target can be easily linked and lead to easier lineage identification. Deep sequencing of PCR products (amplicons) facilitates the efficient identification and characterization of variants. Using oligonucleotide probes, regions of interest are targeted and captured. The amplicon-based approach is particularly useful for microbiome samples of diverse origins.

The use of amplicon sequencing allows researchers to limit the type and number of sequences that can be analyzed. Despite its high specificity, this method requires significant prior knowledge of the sequence that will be 'targeted'. SARS-CoV-2 genome sequencing by amplicon-based methods employs a workflow in which first-strand cDNA is synthesized, followed by multiplex

PCR. Pools of amplicons, covering the entirety or discrete portions of the viral genome, are produced. For SARS-CoV-2, several different multiplex PCR designs have been proposed, varying in the number and size of amplicons [43].

The ARTIC Network's amplicon-based protocol is one of the most widely used SARS-CoV-2 sequencing protocols [45, 46, 47, 48] The protocol relies on direct amplification of the virus using tiled, multiplexed primers. This approach has been proven to have high sensitivity and work directly from clinical samples.

The amplicon sequencing approach has several advantages. Through a highly targeted approach, researchers can discover, validate, and screen genetic variants efficiently. High coverage can be achieved by multiplexing hundreds to thousands of amplicons per reaction. The amplicon-based method is able to work with difficult-to-sequence areas. It allows for reducing the cost and time of sequencing compared to other approaches, such as whole-genome shotgun sequencing. But most importantly, it reduces the amount of starting material required. For instance, it is impossible to do non-ampliconic sequencing from nasal swabs, so an ampliconic-based approach should be applied in this case.

In spite of the fact that amplicon sequencing is theoretically convenient and inexpensive, it has some limitations that should be taken into account. As it is stated in a quick guide to tiling amplicon sequencing and downstream bioinformatics analysis [49, 50], several challenges may arise during the sequencing and analysis process, including contamination, bar-coding issues, and primer binding issues. As a result of the significantly high sensitivity, even a small amount of contaminating templates (such as amplicons from previous work) can lead to the amplification of sequences not present in the sample. The presence of sequences from other samples can confuse the analysis since there is a small rate of barcode "cross-over". Because PCR relies on synthetic primers, amplicons contain synthetic sequences in their primer binding sites. This is a problem that must be accounted for if the primer sequence contains mismatches compared to the template. Additionally, amplification across the genome can be biased by differences in primer efficiency or possible variations in primer annealing regions. For example, specific genomic regions [51] may have less coverage, and/or 3' and 5' untranslated regions may be missed altogether, resulting in incomplete assembly. Due to the fact that the primers are designed using the reference SARS-CoV-2 genome sequence, this approach may not be capable of identifying long structural variants, and high levels of genomic divergence can pose systematic limitations.

A recent study suggests that, although the amplicon-based approach is highly reliable for reconstructing a viral population's most prevalent genome variant, it has a highly biased representation of minor allele frequencies compared to metatranscriptomics experiments conducted on the same samples [52].

1.2.1.2 Sequencing techniques

The majority of SARS-CoV-2 sequence data is generated using two sequencing platforms: Illumina-based technology (Illumina) and Oxford Nanopore Technologies (ONT).

1.2.1.2.1 Illumina sequencing

Illumina sequencing, second-generation sequencing, detects the sequence of RNA by using reversible dye terminators technology. Solexa company, now part of Illumina company, invented the reversible dye terminators technology and engineered polymerases that are used in this sequencing method.

Illumina sequencing begins with the cleavage of the sample into short sections. A short read or fragment of 100-150bp is created at the start of the Illumina sequencing process. The fragments are then ligated to generic adaptors and annealed to slides. Each fragment is amplified by polymerase chain reaction (PCR). By doing this, the same fragment appears in many copies. Sequencing slides contain fluorescently labeled nucleotides, RNA polymerase, and terminator molecules. As a result of the terminator, only one base is added at a time. It allows the next base to be added to the site after each cycle terminator is removed. In each cycle, the computer detects the base added by relying on fluorescent signals [53].

1.2.1.2.2 Nanopore sequencing

Nanopore sequencing is third-generation sequencing technique that detects the RNA sequence by using a protein nanopore. Nanopore sequencing involves RNA passing through a nanopore and changing its current. There is a relationship between the size, shape, and length of the RNA sequence and how a current changes. In order to obtain the specific RNA sequence, the resultant signal needs to be decoded. There is no need to modify nucleotides, and the method works in real-time.

A number of nanopore sequencing devices are manufactured by Oxford Nanopore Technologies. Flow cells are a common feature of Nanopore sequencing devices. An electro-resistant membrane surrounds a number of tiny nanopores in this flow cell. Each nanopore corresponds to its own electrode. An electrode connects to a sensor chip and a channel. A nanopore's electric current is measured by this electrode. As molecules pass through nanopores, their current is changed or disrupted. A characteristic squiggle results from this disruption. Squiggles are decoded in real-time to determine RNA sequences [53].

1.2.1.2.3 Comparison of Nanopore and Illumina

Nanopore sequencing uses a nanopore to detect the RNA sequence, while Illumina sequencing technique uses reversible dye terminators technology to detect the sequence of RNA. Both techniques have markedly high accuracy. More precisely, Nanopore sequencing has 92-97% accuracy, while Illumina sequencing has 99% accuracy. The following table 3 lists the differences between Nanopore and Illumina sequencing.

1.2.1.3 Data processing with bioinformatics methods

For the next stage after sequencing, bioinformatics analysis should be executed. Tools can differ from one pipeline to another. But the main steps, in general, are more or less the same. After getting raw data reads, they have to be processed with different steps.

	Nanopore	**Illumina**
Generation	third-generation	second-generation
Founded by	Prof. David Deamer	Prof. Shankar Balasubramanian and Sir David Klenerman
Accuracy	92-97%	99%
Long/short read sequencing	long reads	short reads
Read length	2,272,580 bp	75-300 bp
Time taken	Real time	4 to 56 h
Cost	$7-100	$5-150
Advantages	longest individual reads	potential for high sequence yield
Disadvantages	lower throughput than other machines	expensive

Table 3: Comparison of Nanopore sequencing technique (Nanopore) and Illumina-based technology (Illumina) sequencing

Quality control step is often used because there is no perfect sequencing technology, and each instrument will generate different types and amounts of errors, such as incorrect nucleotide calls. Each sequencing platform has technical limitations that result in these incorrectly called bases. Thus, it is important to identify and exclude error types that may affect downstream analysis interpretation. As a result, sequence quality control is an essential first step in the analysis process.

Another step, primer trimming, is a specific step for datasets generated with ARTIC protocol. The auxiliary file is used for this step - a BED file specifying the primers used during amplification and their binding sites on the viral genome. Primer trimmer uses primer positions supplied in a BED file to soft clip primer sequences from an aligned and sorted BAM file. Following this, the reads are trimmed based on a quality threshold. More specifically, some primer trimmers, in order to do quality trimming, use a sliding window approach. The window slides from the 5' end to the 3' end and if at any point the average base quality in the window falls below the threshold, the remaining read is softly clipped. If after trimming, the length of the read is greater than the minimum length specified, the read is written to the new trimmed BAM file. It should be noted, for datasets that were not generated with primer-based protocol like ARTIC, this primer-trimming step is not applicable.

Moreover, adapter trimming step is processed. For instance, upon Illumina sequencing we receive raw reads with adapters at 3' end. The adapters contain the sequencing primer binding sites, the index sequences, and the sites that allow library fragments to attach to the flow cell lawn. This might influence a downstream analysis, thus, adapter trimming is required.

A decontamination step can then be included to remove reads from the human genome, since viral sequence data from clinical samples commonly contain human contamination. Prior to sharing, it needs to be removed for legal and ethical reasons as well as to speed up downstream analysis [54].

The crucial step is mapping with reference SARS-CoV-2 sequence *NC_045512.2* that is publicly

available in NCBI database [55]. A mapping tool of choice can differ from one pipeline to another, depending on read length, sequencing technology, and other factors.

Some pipeline steps are not always included in pipelines, such as removing duplicates. This step can be important for Illumina sequencing reads. During the sequencing process with Illumina sequencing technology, some duplicate reads/sequences can be produced, which can create bias in downstream analyses. It is, therefore, possible to remove duplicates or mark them without removing them. When removing duplicates, one should be certain that they are duplicates and not repeated regions. It can therefore be reasonable to keep duplicates marked rather than remove them, as this can be useful for downstream analysis.

Another step, which is not present everywhere, is helpful due to potential ambiguity, while indels are not parsed when they overlap the beginning or end of alignment boundaries. Input insertions and deletions must be homogenized with left realignment in order to gain a more homogeneous distribution. Left realignment will place all indels in homopolymer and microsatellite repeats at the same position, provided that doing so does not introduce mismatches between the read and reference other than the indel [56]. Basically, this step is considered to correct mapping errors and prepare reads for further variant calling.

Additionally, realigned reads can be taken and checked for the quality of alignment using bioinformatics tools (e.g., Qualimap [57, 58]). Based on the features of the mapped reads, it analyzes SAM/BAM alignment data and provides a global picture of the data that can help detect biases in sequencing and/or mapping of the data and ease decision-making for further analysis.

After mapping and other additional preparation steps, variant calling should be run where variants from sequence data are identified. Variant calling step is followed by mutation annotation where VCF file is used as input and annotated SARS-COV-2 genome - as database, and it is transformed to MAF format. The data is not changed; here, only format is changed to be more readable.

There is a number of publications describing different combinations of bioinformatics steps and tools used to process and analyze raw data [7] and ranging from transparent to opaque [59]. Analytical transparency is crucial for such cases. Now, many organizations provide transparency to data processing and analyzing approaches. For example, the COG-UK datasets, protocols, methods, and techniques for collecting and preparing SARS-CoV-2 virus samples for sequencing, as well as short-read and long-read sequencing, are all publicly available [48].

1.2.2 SARS-CoV-2 tracking projects

SARS-CoV-2 variants began being tracked from the beginning of the COVID-19 pandemic in order to observe global trends of variants of concern (VOCs) (fig. 3).

In the course of the COVID-19 pandemic, many laboratories have developed genomic epidemiology data infrastructures. By January 2022, there were nearly 4000 unique labs submitting data to the GISAID EpiCoV database. A consortium called COG-UK [60] is one of the well-known SARS-CoV-2 tracking projects. Their research built on the availability of SARS-CoV-2 genomes generated throughout the pandemic and spanned across variants, data linkage, and quality. Moreover, there are other complementary projects focused on the emergence of SARS-CoV-2 variants. At the beginning of the COVID-19 pandemic, a problem with available data

to analyze occurred [7]. In this regard, COG-UK is the SARS-CoV-2 data tracking project that reflected quickly to that and was one of the first that shared data for publicity. Another example of a tracking project that is focused on SARS-CoV-2 sequencing is the Swiss SARS-CoV-2 Sequencing Consortium (S3C) [61, 62].

Currently, large amounts of sequencing data have become increasingly available and must be constantly analyzed. As a result of these data, one can monitor the emergence and spread of new variants, as well as understand the viral evolution dynamics. Nevertheless, transparent and freely available infrastructure for such analysis is not present everywhere. It is often the case that infectious disease outbreaks occur in remote areas without adequate infrastructure or in political situations that make unbiased interpretation of results impossible. In response, there is a need for free, open, and robust analytical approaches accessible to everyone worldwide for data analysis, interpretation, and sharing.

1.2.3 Galaxy effort to surveillance data processing

In order to address global health emergencies in an accessible and transparent manner, there is a need for scientific computing infrastructure to help bridge these gaps. Data and data analysis transparency were the primary focus of the Galaxy effort.

Through its graphical web interface, Galaxy allows users to use tools from a variety of domains for FAIR [63] data analysis; run code in interactive environments (RStudio, Jupyter, etc.) along with other tools or workflows; manage data by sharing and publishing results, workflows, and visualizations; ensure reproducibility by capturing information that allows data analysis to be repeated and understood [7].

Galaxy offers powerful public computational infrastructures designed specifically for research purposes in the United States, the European Union, and Australia. In the United States, there is the XSEDE [64] consortium; in the European Union, there is the deNBI [65] and ELIXIR [66, 67] consortiums; and in Australia, there is the Nectar Cloud consortium. Due to their global accessibility, support for diverse configuration schemes (from traditional computational clusters to fully virtualized cloud-like setups), and ability to provide cutting-edge hardware, large-scale public computing resources are well suited to tackling the bioinformatics challenges of the current pandemic.

In combination with open-source software tools, this public computational infrastructure offers a complete solution to the SARS-CoV-2 data analytics challenge. Galaxy provides a glue to bind these into a unified analytics platform that manages users, allocates storage, and pairs analysis tools with computational resources fluidly. The platform provides both a graphical user interface and programmatic access in order to accommodate researchers with different degrees of computational expertise.

1.2.3.1 Four SARS-CoV-2 analysis workflows

Based on Galaxy, four workflows were developed aimed at detecting and interpreting sequence variants in SARS-CoV-2 in clinical data. A special webpage was created to track Galaxy efforts on SARS-CoV-2 analysis, workflows, data, and documentation [68]. These workflows can be accessed immediately and freely in three global Galaxy instances: in the United States (`https:`

/ / usegalaxy . org/), the European Union (https : / / usegalaxy . eu/), and Australia (https://usegalaxy.org.au/).

When creating workflows for SARS-CoV-2 analysis, W. Maier et al. [69] took several goals into account. They aimed to provide continuous analysis of within-host sequence variants in high-quality public read-level databases. The other goal was to provide maintenance of curated workflows for the analysis of SARS-CoV-2 sequence data and free powerful infrastructure to execute them, such as Galaxy infrastructure. They developed a continuously updated analysis page and dashboard summarizing the latest insights from the variant. Last but not least focus was to provide access to all results in raw and aggregated form for immediate use.

As suggested by W. Maier and colleagues, SARS-CoV-2 clinical data analysis is organized into two stages of analysis (fig. 5). Stage 1 takes raw sequencing data as input and produces intermediate summaries such as allelic variants and annotations in mutation annotation format (MAF files) and variant calling format (VCF files). Basically, at Stage 1, depending on library preparation and data sequencing method, the appropriate workflow is taken. Thereby, for datasets prepared by using ARTIC amplicon-based protocol, two workflows are available: for datasets sequenced with Oxford Nanopore Technologies (ONT) and Illumina-based technology (Illumina). As for datasets prepared with metatranscriptomic protocol, there are two workflows available for Illumina-based sequencing technology. The difference is in the types of reads obtained. One workflow works with paired-end reads, while another one works with single-end reads.

Interpreting and visualizing the Stage 1 outputs are the focus of Stage 2. Additional reporting workflow is developed to analyze outputs of Stage 1, which generates tabular and JSON files. These files serve as input for analysis based on two software systems, Jupyter, and ObservableHQ, for visualizing results.

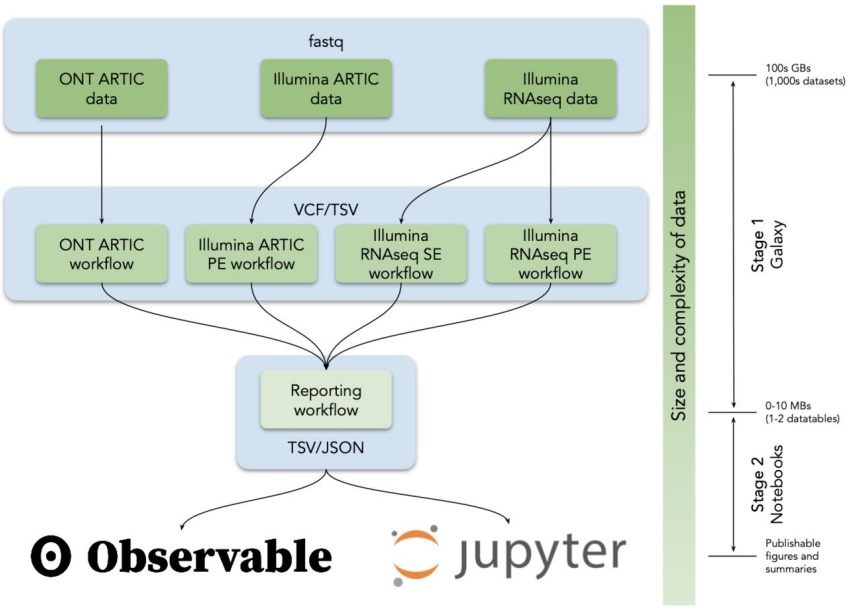


Figure 5: Analysis flow in the analysis system suggested by GalaxyProject (image from [69]).

The choice of bioinformatics tools for these four workflows is made based on the method of sequencing and the type of data obtained by the method. For paired-end reads, the BWA-

MEM [70, 71] mapper is used to map to the SARS-CoV-2 sequence. Bowtie2 [72, 73] mapper is used for single-end read datasets sequenced with Illumina method, whereas minimap2 [71, 74] mapper is most appropriate for datasets sequenced by Oxford Nanopore since Oxford Nanopore sequencer generates very long reads with many errors. As for variant caller, LoFreq is used in all four workflows since it is known as a more ultra-sensitive variant caller for uncovering cell-population heterogeneity [75].

1.2.3.2 Bots for automated SARS-CoV-2 surveillance

Additionally, Galaxy team has developed bots to assist in SARS-CoV-2 surveillance [76], a viable tool for automating the analysis of a large number of SARS-CoV-2 sequences regularly. In essence, it is a set of automation scripts that can be integrated into any scheduling system with a daily throughput of around 1000 samples on a Galaxy instance. Galaxy bots allow to upload newly available data, start a variation analysis workflow, and follow up with downstream workflows for consensus building and reporting once the variation analysis is complete. It is possible to use bots for SARS-CoV-2 surveillance to track National Genome Surveillance projects, such as COG-UK, and reanalyze their data as it becomes available. These scripts are a decent automated sequence data analysis pipeline that works with two Galaxy SARS-CoV-2 surveillance workflows for datasets extracted with ARTIC protocol.

Galaxy workflows developed for SARS-CoV-2 clinical surveillance have shown adequate results. There are, however, some limitations. Currently, Galaxy workflows do not focus on wastewater surveillance. Thus, Galaxy workflows can be improved and repurposed to improve SARS-CoV-2 wastewater surveillance. The current thesis attempts to focus on it.

1.3 Wastewater surveillance

This section will provide the introduction to wastewater surveillance in general and particularly for SARS-CoV-2 as well as show the limitations and benefits of wastewater surveillance over clinical surveillance. Then, the focus will be on wastewater surveillance contribution to global safety and global efforts towards it. After that, a few words about the wastewater surveillance working groups will be said.

Wastewater surveillance has played an important role in controlling outbreaks, for example, poliovirus outbreaks which have been challenging in the past. Israel detected wild poliovirus in wastewater samples between 2013 and 2014, 25 years after the last case. In March 2022, poliovirus was found in sewage samples from Jerusalem and surrounding areas. Both outbreaks were contained with vaccination campaigns, which prevented the further spread of the virus. There was only one case of paralysis in an unvaccinated child in 2022, and no cases of paralysis occurred in the 2013–2014 outbreak in Israel due to rapid detection enabled by wastewater surveillance. In recent two years, increased public awareness of any suspicious virus, including poliovirus, and ubiquitous usage of wastewater surveillance of poliovirus are preventing any cases of paralysis following the recent re-emergence of polio in New York [77].

With the experience in poliovirus wastewater surveillance, it was reasonable to use this method for other virus surveillance. People who are infected with SARS-CoV-2 have the virus in their stools regardless of whether they show symptoms of the disease. Additionally, SARS-CoV-2

RNA may be carried into the sewer system from urine [78] and respiratory secretions (from hand washing, showering, nasal lavages, tissues, and sputum), as indicated by detected SARS-CoV-2 RNAs in washbasins and shower siphons [79].

Wastewater with these secrets inevitably ends up in wastewater treatment plants where samples can be collected. In treatment plants, typically, samples of incoming sewage at an early stage are taken, and the plants are designed so that this can be done safely and effectively [80]. The automated sampling system can also be used to collect wastewater samples directly from pumping stations or at suspected virus hotspots such as hospitals [81], dormitories, residential districts, or confined spaces like cruise ships or passenger airplanes [82].

During a pandemic such as COVID-19, wastewater surveillance [83, 84] can be used to detect both the presence and absence of the virus, as well as the emergence and transmission of new variants that are more infectious or immune-evading. There will be more SARS-CoV-2 variants, regardless of when or where the next major variant emerges. It is imperative to detect them as soon as possible, whether they originate from known variants or appear independently [85].

A method of wastewater genomic surveillance was used by California researchers to detect SARS-CoV-2 infections on the campus of the University of California, San Diego (UCSD), in the midst of the pandemic during a period of 10 months. Researchers found that their approach was effective in identifying viral variants of concern as early as two weeks before they showed up in clinical tests, according to an article published in Nature [86]. Additionally, wastewater samples were analyzed in Sweden. The Rya wastewater treatment plant in Sweden collected samples of wastewater from Gothenburg and surrounding municipalities from mid-February to June 2020. It appeared that the amount of SARS-CoV-2 varied with peaks roughly every four weeks, preceding variations in the number of newly hospitalized patients by 19-21 days [87].

The SARS-CoV-2 virus is already tracked by wastewater surveillance in over 55 countries [80], and various monitoring programs are run worldwide. In the US, in September 2020, the Centers for Disease Control and Prevention (CDC) launched a National Wastewater Surveillance System [88] in conjunction with health departments across the country. SARS-CoV-2 levels in community wastewater are monitored by this program.

1.3.1 Challenges and limitations of wastewater surveillance over clinical surveillance

In spite of the increasing popularity of wastewater surveillance across the world, it can be more challenging to detect viruses accurately in wastewater, compared to clinical testing [89]. Prior to virus concentration, large quantities of sewage sludge typically should be filtered to remove debris, flocculated, precipitated, or centrifuged. Molecular analyses, like PCR, can be hindered by concentration techniques that damage genomic material. Furthermore, sewage contains a wide variety of other microbes and viruses, which may produce false positive results, as well as human DNA [85].

Because it contains human DNA, wastewater data need to be anonymized due to privacy concerns. However, pathogenic surveillance has the task of linking genetic information with the clinical manifestations and immunological status of patients [85]. That means that wastewater surveillance is limited in this regard, and it is able to provide only coarse population-level information.

Another challenge is a great deal of variation in how much virus is shed in feces and urine between individuals, so it is very difficult to quantify the number of people who are infected [89].

Moreover, due to the mobility of populations in any particular region, wastewater surveillance cannot pinpoint infected individuals or trace transmission. As a result, wastewater detection efforts can be hindered by unwittingly spreading a pathogen by infected people passing through a region, often after the index case has moved on [90].

1.3.2 Benefits of wastewater surveillance over clinical surveillance

Nevertheless, the use of wastewater methods is quite beneficial, since they can allow the detection of outbreaks before the first positive clinical tests are reported. Based on the first cases of SARS-CoV-2 wastewater surveillance, it has already been found [91] that virus RNA is detected in sewage even when COVID-19 prevalence is low, and that the correlation between concentration in sewage and reported COVID-19 prevalence indicates that sewage surveillance can be used as a sensitive tool to monitor viral circulation in the population.

Likewise, wastewater surveillance is more economical than clinical testing since it can screen large numbers of people with just a few samples and does not need clinician involvement.

Another advantage is that by using the wastewater surveillance of SARS-CoV-2 method, data can be collected from people who do not have access to healthcare or in places, so-called “sequencing deserts”, around the world where sequencing capacity is limited. Among such "sequencing deserts”, new, potentially dangerous variants, like Omicron, are able to emerge and spread undetected. There is a risk that new variants or subvariants will emerge until representative samples are sufficiently sequenced. Wastewater surveillance is one of the opportunities for covering 'sequencing deserts' for surveillance of variants of SARS-CoV-2. As a result, a very close to a real-time overview of disease prevalence could be provided since it was proved [91] successful enough in revealing infection dynamics earlier than clinical testing. In fig. 6, the schematic diagram shows the process of detecting viruses by wastewater surveillance against clinical surveillance.

1.3.3 Wastewater surveillance contribution to global safety

Viruses are not the only thing that can be monitored in wastewater. This method can be used to detect other microbial pathogens [92], antimicrobial resistance [93], or chemical water contaminants [94]. Further, these tools can be used on samples from other settings, such as transport hubs, hospitals, schools, workplaces, and leisure facilities, in addition to wastewater treatment facilities. Besides public-health applications, wastewater surveillance data may be useful to researchers examining community trends and the efficiency of health policies and non-pharmaceutical interventions [85].

Wastewater surveillance provides valuable information on epidemiological developments at the population level that complements case-based surveillance and aids in resource allocation. Likewise, this can be applied to a global scale. Using the wastewater-based epidemiology method, places and communities with poor healthcare accessibility can shed light on the blind spots of

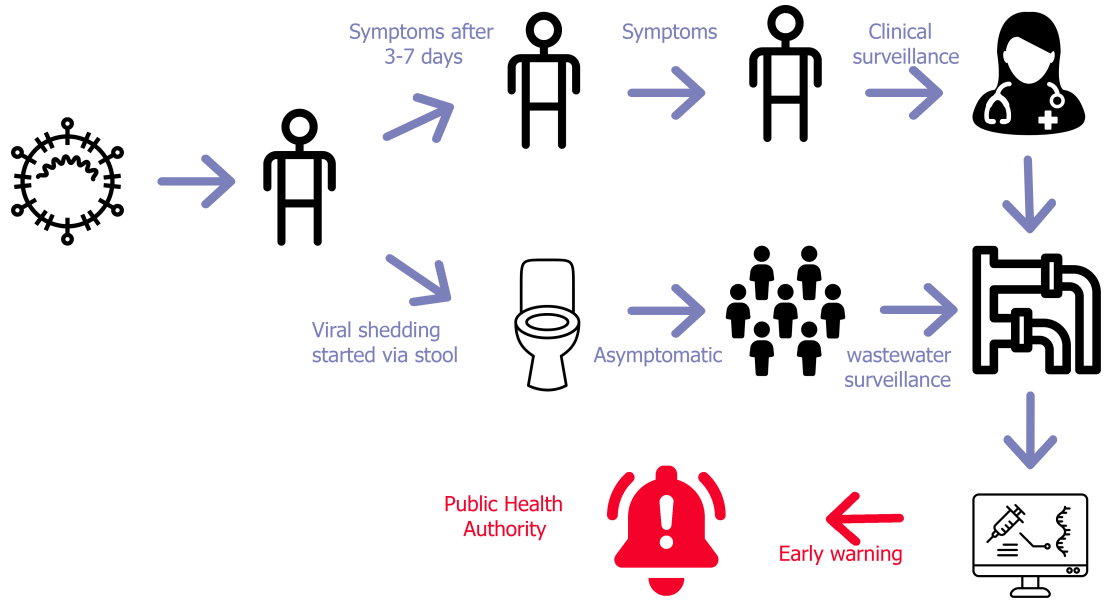


Figure 6: Schematic diagram shows the process of detecting viruses by wastewater surveillance against clinical surveillance.

pathogen surveillance. A wastewater surveillance system for infectious diseases could contribute to global safety if it is carefully set up and used in a respectful manner.

The wastewater surveillance methods seem to fit the bill for enabling early, economical, and efficient detection so that public-health measures can be implemented as soon as they are necessary. Globally, wastewater surveillance is poised to become part of public health strategies [93].

1.3.4 Global effort to wastewater surveillance

Enough efforts have already been made in this direction in the world. Various wastewater tracking projects took place in countries like Estonia [95, 96], Greece [97], and Canada [98]. As a result, COVID-19 trends are visualized within the sewer community, contributing to COVID-19 incidence (both reported and unreported). In response to the COVID-19 pandemic, wastewater surveillance data are used to understand and act, and in public health decision-making. COVIDPoops19, a dashboard developed by Colleen Naughton and colleagues at the University of California (UC), Merced, shows monitoring projects for SARS-CoV-2 have sprung up in at least 70 countries since then [99]. By October 2021, the European Union recommended that all member countries establish monitoring systems for SARS-CoV-2. 26 of 27 countries have adopted this recommendation. In the United States, the National Wastewater Surveillance System includes 400 sites in 19 states. In the U.S., on 2 March 2022, President Joe Biden's administration said the monitoring system would be part of efforts to detect new variants as the Centers for Disease Control and Prevention added a national dashboard of wastewater data. As well, there is a successful project in Bengaluru that has been expanded to half a dozen other cities in India [100].

At the same time, there are several initiatives to coordinate wastewater surveillance efforts via working groups. For example, 219-nCoV WBE is a Slack space in which worldwide participants share protocols, publications, and other state-of-the-art resources, and help each other. At the European level, a focus group meets every month for wastewater surveillance efforts coordination.

1.4 Motivation and the goal of the thesis

As a result of extensive global efforts, numerous pipeline solutions have already been proposed by the scientific community. In terms of accessibility and readiness for use, wastewater surveillance pipelines exist with some limitations. Some pipelines demand an upstream set of tools or sub-pipelines that can take some time to prepare datasets. In addition, they are quite strict when it comes to the tools they use. The best performance is achieved by combining different tools in different workflows. Furthermore, the reproducibility of these workflows is not straightforward. Researchers typically receive instructions of varying depths of detail explaining how to launch workflows in their environments. When comes to repeat analysis several times, can turn to some inconveniences. Additionally, researchers are not provided resources to conduct these workflows and downstream analyses afterward. The platform and resources are crucial to research, even though they are a second priority to some scientists.

In this master thesis, I aim to provide a complete workflow based on Galaxy that assists platforms that can ensure data analysis transparency and reproducibility. To be precise, I intend to adapt the Galaxy workflows developed for clinical data to process wastewater data. In doing so, I integrate existing tools, test these workflows on mock datasets as well as real datasets, and benchmark them against each other and with other solution offered by other researchers.

2 State-of-the-art

From the beginning of the COVID-19 pandemic, several approaches have been presented to determine SARS-CoV-2 in wastewater samples. An overview of existing tools and pipelines will be given in section 2.1. Some tools and pipelines aiming to detect SARS-CoV-2 lineages and their abundance in wastewater samples will be considered. More specifically, section 2.1.1 will show individual tools that need to be implemented into pipelines, while section 2.1.2 will demonstrate standalone pipelines for detecting lineages and their abundances. In section 2.1.3 an overview of the comparisons will be presented both for individual tools and for standalone pipelines. In section 2.2, a short discussion of the existing methods with respect to the goals of this thesis will be given.

2.1 Methods for wastewater surveillance

Existing methods for SARS-CoV-2 wastewater surveillance differ. Some approaches present individual tools that require data preprocessing and/or need to be plugged into pipelines before they can be used, while another methods offer independent pipelines that do all the analysis from raw data to determining lineages and their abundances. This determination can be focused on variants of concern (VOCs) as well as variants of interest (VOIs) and additionally with an intention to detect newly appeared unknown variants. Both types of methods will be discussed in the following sections.

2.1.1 Individual tools

Methods that offer individual tools to discern SARS-CoV-2 lineages and their abundances take preprocessed data. Raw data needs to be preprocessed using a pipeline or other individual tools. The feature of individual tools for wastewater surveillance is their flexibility in being incorporated into workflows.

2.1.1.1 Freyja

One optimized approach for virus concentration from wastewater is proposed by Karthikeyan et al. [86], they developed developed a tool called Freyja that detects variants from SARS-CoV-2 RNA sequencing and estimates the relative abundance of SARS-CoV-2 lineages. To represent each SARS-CoV-2 lineage in the global phylogeny, Freyja uses a barcode library of lineage-defining mutations [85, 101]. These lineage-determining mutational barcodes derived from the UShER [102] global phylogenetic tree as a basis set to solve the constrained (unit sum, non-negative) de-mixing problem. For each lineage-defining mutation, Freyja stores the single nucleotide variant (SNV) frequencies (the proportion of reads that contain the SNV) for each sample. In order to recover relative lineage abundance, Freyja solves a depth-weighted least absolute deviation regression problem, a mixed sample analog to minimize edit distances between sequences and a reference, based on SNV frequencies at positions with greater sequencing depth. In order to guarantee meaningful results, Freyja constrains the solution space such that each lineage abundance value is non-negative and the overall lineage abundance sums to one. Site-specific weighting is applied by Freyja to account for non-constant variance in measured SNV

frequency across sites, which enables prioritizing information according to sequencing depth at each site. Because of using log-transformed read depths, the data is robust to common characteristics of real sequencing data, such as heavily skewed depths across amplicons [86].

As Freyja uses UShER phylogenetic tree, it is a good solution for users, for example, politicians or journalists, that chase only high-level information to know which variants are relevant and the most ubiquitous at this time point. The information about the proportion of specific SARS-CoV-2 lineages (e.g. Alpha, Delta, Omicron, etc.) and their co-occurrences in certain areas could be useful for some rough reports that are needed for political or mass media goals. Since using UShER bar codes, we know the names of lineages without details, that could be a good solution.

2.1.1.2 COJAC

In parallel, based on SARS-CoV-2 RNA amplicon sequencing for wastewater samples in two cities in Switzerland, Jahn and co-authors used a bioinformatics method called Co-Occurrence Adapted Analysis and Calling (COJAC) [6] to detect the local spread of Alpha, Beta and Gamma variants in the virus. For detection, COJAC searches for read pairs with multiple variant-specific mutations.

The COJAC [103, 6] package comprises a set of command-line tools to analyze the co-occurrence of mutations on amplicons. It is useful, for example, for the early detection of viral variants of concern in environmental samples, and has been specifically designed to scan for multiple SARS-CoV-2 variants in wastewater samples [104].

Similarly to Freyja, COJAC is able to detect outbreaks more than two weeks before the first positive clinical tests were reported. Compared to the Freyja tool, COJAC results can be used for downstream analysis that afterward can be imported to CoV-spectrum [105], an interactive platform aiming to assist scientists in investigating and identifying SARS-CoV-2 variants, and it can be interesting for researchers and scientists that need more details about not only the names of SARS-CoV-2 lineages, that are listed in UShER phylogenetic tree but also new variants that are yet unknown. The usage of CoV-Spectrum tool for downstream analysis is supposed to be up-to-date and to show dynamics. This can be used to give information about new variants and new patterns that can be described as suspicious.

2.1.1.3 LCS

Another tool that estimates the relative frequencies of SARS-CoV-2 variants in pooled samples, LCS [106], uses a statistical model. The model takes into account previously defined genomic polymorphisms that characterize SARS-CoV-2 variants. The tool supports both raw sequencing reads for polymorphism-based markers calling and predefined markers in the variant calling format. Similarly to Freyja, LCS tool uses Pango designation, and it is able to use UShER bar codes which is not the case for COJAC.

2.1.1.4 Kallisto

Among other methods, Kallisto [107] was initially developed for quantifying abundances of transcripts from RNA-Seq data, or, more generally, of target sequences using high-throughput sequencing reads. It is based on the idea of pseudoalignment for rapidly determining the compatibility of reads with targets, without the need for alignment.

In spite of tools listed above, Freyja, COJAC, and LCS, Kallisto initially was not developed with focus on SARS-CoV-2. Recently, Kallisto tool was repurposed to analyze SARS-CoV-2 in wastewater samples [108, 109]. Due to the genome's division into regions that code different proteins, it was thought that it should only be sequenced in the region that codes the spike protein, which is used by the virus to attach to its hosts. Sites on the spike and nucleocapsid proteins are mainly affected by adaptive evolution in SARS-CoV-2 [110]. As a viral nucleocapsid protein, it binds to the genomic RNA sequence and enters the host cell. As long as Kallisto is used on those regions, the algorithm is able to distinguish virus variants. Both [108] and [111] provide confirmation of this since the region coding for spike protein seems to fit better in the algorithm than sequencing the whole genome. However, Kallisto has limited capability to detect clinical variants with a frequency of <10%, and background noise was observed to be 0.01-0.09% [108].

2.1.1.5 Alcov

A tool called Alcov [112] is another option for learning the abundance of SARS-CoV-2 variants. For each mutation, Alcov calculates its frequency as the number of single-nucleotide variants divided by the number of reads covering that base. The Alcov algorithm only considers locations with at least 40 reads of coverage in order to reduce the variance of the estimates. Whenever a mutation contains multiple nucleotide polymorphisms, Alcov assigns its prevalence to the nucleotide polymorphism with the highest frequency. Ordinary least squares (OLS), a well-studied problem in statistics that can be solved efficiently, are used to calculate relative abundance for each variant of concern (VOC) lineage. As COJAC, Alcov works with amplicons. When comparing Alcov method with COJAC approach, Alcov considers reads carrying mutations when estimating lineages cooccurrences, while COJAC considers amplicons carrying mutations.

2.1.1.6 SAM Refiner

A further method called SAM Refiner is a part of the workflow developed by Devon A. Gregory et al. [113] based on amplicon sequencing of SARS-CoV-2 spike domains in order to track viral populations in wastewater. In June 2021, Devon A. Gregory and colleagues were able to detect the appearance of two variants of concern, Alpha and Gamma, and their displacement of the D614G B.1 variant in a Missouri sewer shed with the help of SAM Refiner. This tool can report novel variants and remove chimeric sequences generated by PCR. Comparing SAM Refiner with above listed methods, SAM Refiner has a feature of removing errors associated with PCR, such as single-nucleotide polymorphisms (SNPs) and chimeric sequences which is not suggested by Freyja, COJAC, Kallisto, and Alcov tool. Moreover, by using this tool, variant reports are generated that include SNPs, multiple nucleotide polymorphisms (MNPs), insertions, deletions, and downstream amino acid changes, as well as PCR-created chimeric sequences are removed.

2.1.2 Standalone pipelines

In some wastewater surveillance methods, it is possible to set up self-contained pipelines for SARS-CoV-2 wastewater surveillance, including the preprocessing of raw data and the computation of the abundances of lineages. The strength of standalone pipelines for wastewater surveillance is their complete analysis and no need to implement them into other workflows. However, they are not as flexible as individual tools since they use a limited set of tools.

2.1.2.1 Pipeline proposed by Pipes et al.

In the method proposed by Pipes et al. [114], unlikely lineages are initially removed, missing nucleotides are imputed using the global SARS-CoV-2 phylogeny, and the proportion of lineages is estimated using an Expectation-Maximization algorithm [115]. There are two components of the method: estimating proportions of SARS-CoV-2 lineages and imputation of SARS-CoV-2 reference lineages. The imputation component is a specificity of the method proposed by Pipes et al., compared to other methods described above, like Freyja, COJAC, SAM Refiner, etc. This component's purpose is to reduce a problem when estimating the fraction of different SARS-CoV-2 strains. A problem can arise because a large number of SARS-CoV-2 sequences submitted to public databases contain missing data (i.e., bases that are not coded as A, G, C, or T). Thus, when compared to sequencing reads, strains with a high proportion of missing data will contain fewer nucleotide differences. While using an imputation approach, it is possible to allow for a like-to-like comparison of reads against all reference strains. The method was initially applied to wastewater samples collected across the San Francisco Bay Area and showed promising results.

2.1.2.2 Gromstole

Yet another concept for estimating the relative frequencies of different SARS-CoV-2 lineages in wastewater samples is pipeline Gromstole [116]. Gromstole provides a set of scripts, including a Python script for rapidly extracting coverage and mutation frequency statistics from the reference mapping program minimap [74], and an R script that estimates the frequency of a variant of concern from the frequencies of mutations associated with the constellation file, using quasibinomial regression. The other R script is used to visualize variant frequency estimates across a set of samples.

2.1.2.3 AG

Another approach, AG pipeline, is a suite of tools [117]. A pipeline was successfully applied to 936 wastewater samples and thousands of matched clinical sequences collected in Montreal, Quebec City, and Laval between March 2020 and July 2021. To calculate SARS-CoV-2 lineage frequencies within a sample, a linear model was fitted to the signature mutations data to calculate within-sample frequency. This approach is based on the assumption that the frequency of mutations within a sample is a linear combination of the frequency of the lineages and the prevalence of mutations within their consensus sequences.

2.1.2.4 Lineagespot

Pechlivanis et al., in turn, proposed a framework, called Lineagespot [118], for monitoring mutations and detecting SARS-CoV-2 lineages in wastewater samples. Using next-generation sequencing data from 14 wastewater samples covered by a 6-month period collected from the municipality of Thessaloniki, Greece, the method was tested and evaluated.

With Trim Galore [119], Lineagespot performs quality control and adapter trimming on raw FASTQ files. To map reads to reference SARS-CoV-2 sequences, minimap2 [74] is used. Primer trimming is performed using SAMtools [120], BEDTools [121], and minimap2. By using Picard [122], duplicates obtained from the sequencing process can be filtered. Then, Lineagespot uses FreeBayes [56] to call variants and SnpEff [123] to annotate mutations. Then, for the lineage assignment process, in order to retrieve lineage definitions, Lineagespot relies on particular sources. For Lineagespot, two sources can be used: lineage-characteristic mutation profiles derived from (i) outbreak.info [124], and (ii) trained Pangolin [35] models. As a result of the pipeline, a tabular output contains the most probable lineages found. The diagram that describes the lineagespot workflow is presented in fig. 7.

After computing lineages abundances, Lineagespot provides a set of indicators for each of the provided references. Based on these indicators, the final assignment is made. This additional step in Lineagespot is one of the advantages of this pipeline over others since it allows to mitigate the risk of wrong assigned lineages (e.g. in the case of two groups of reads satisfying different rules for the same lineage but reads from both groups never satisfying all rules for this lineage). According to Lineagespot, several different indicators reflect how many rules are satisfied, how many rules are satisfied based on the detected mutations, and how many reads support each rule both as a reference and as an allele. The drawback of this method is that assignment is made by a semi-automated process but not completely automated.

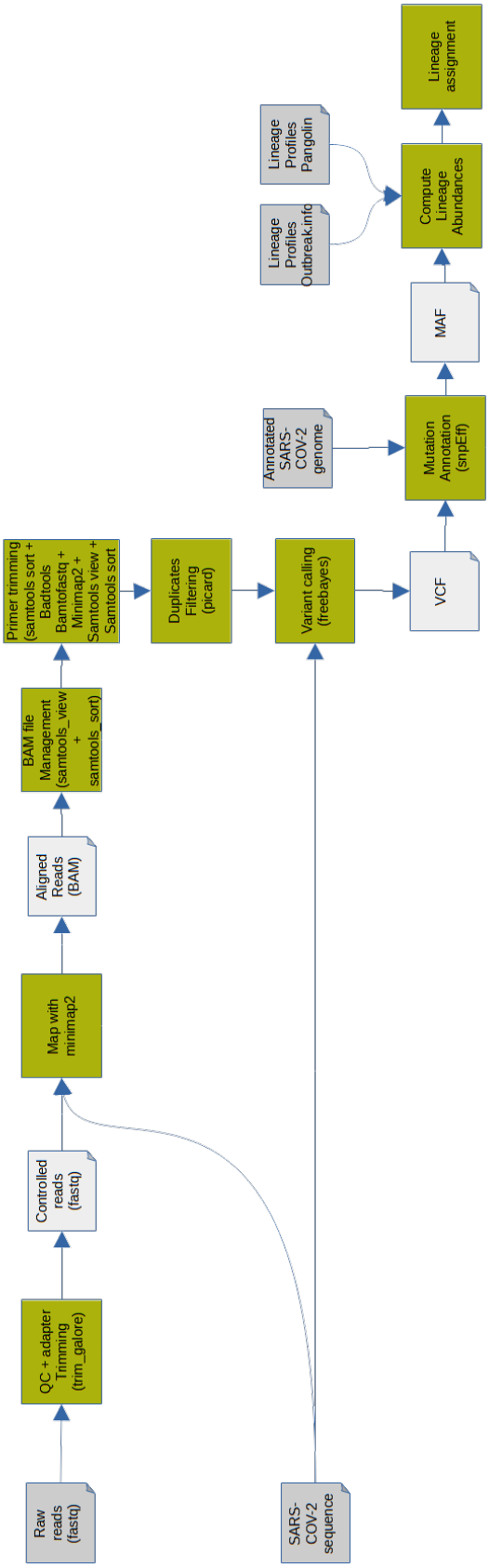


Figure 7: Lineagespot workflow.

2.1.2.5 PiGx

The other pipeline, PiGx SARS-CoV-2 [125], first, performs extensive quality control on the raw reads and information about primers and adapters used. Fastp [126] is used for adapter trimming and filtering, while iVAR is used for primer trimming. Using BWA [70], the trimmed reads are aligned to the reference genome of SARS-CoV-2, resulting in SAM/BAM files with aligned and unaligned reads. After alignment, MultiQC [127] is used to aggregate reports that check the quality of raw and processed reads. Further, PiGx checks samples for genome coverage and how many of the provided signature mutation sites, sites with mutations that characterize variants of concern or variants of interest, are covered by the sample. Every sample is given a quality score based on this. Those samples with genome coverage of less than 90% are discarded from time series analyses and summaries.

With LoFreq [75], variants are called, and SNVs are inferred from aligned reads. The mutations are annotated with VEP [128].

PiGx is capable of deconvoluting the frequencies of variants of concern from pooled sequencing reads. Briefly, the deconvolution method uses signature mutations for each variant of concern and tries to discern the proportions of these variants making up the observed mutation frequencies in the pooled sequencing reads obtained from the wastewater. To infer the proportions of each lineage on each sample, the deconvolution method is applied (based on the frequencies of the signature mutations). On the basis of the observed frequencies of signature mutations for each lineage, the lineage frequencies are predicted using a regression model.

After lineage frequencies prediction, PiGx generates a per-sample report that describes variants of concern and frequency of lineages from each wastewater sample. Additionally, the unaligned reads will be taxonomically classified with Kraken2 [129, 130, 131] to determine the abundance of RNA that matches other existing species found in unaligned reads in the wastewater samples. The diagram of PiGx pipeline is shown in fig. 8.

One of the significant benefits of PiGx method is the addition of different weights of tracked lineages based on how many signature mutations were found for each of them for a given sample. The purpose of this step is to determine which lineages have a low abundance and have only a few or only shared mutations in order to obtain more precise predictions.

PiGx has another advantage - controlling taxonomic classification for other species present in samples but not mapped to the SARS-CoV-2 reference sequence. With Kraken2 usage, this simple step opens opportunities for learning about sample diversity. Due to Kraken2's k-mer classification, it can also identify reads that match SARS-CoV-2 but are not aligned with stringent alignment tools. In addition, it provides insights into the possibility of losing new mutations that weren't captured in the alignment. The user can analyze potential issues here and adjust the alignment stringency if necessary.

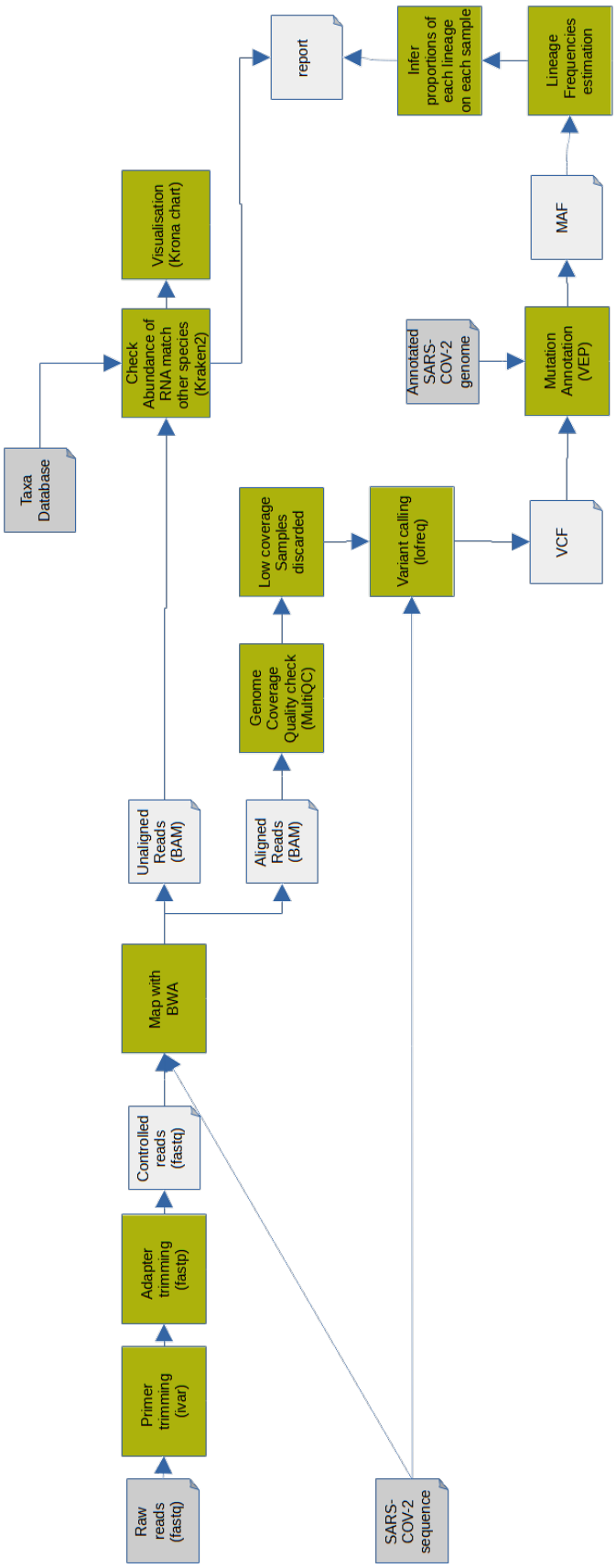


Figure 8: PiGx workflow.

2.1.2.6 Cowwid

In its turn, another approach, Cowwid [103] pipeline uses steps: i) processing raw reads with the help of V-pipe [132] sub-workflow (performing quality controls and read filtering with FastQC and PRINSEQ [133, 134], assembly with VICUNA [135], read alignment with BWA-MEM and ngshmmalign [70, 136] and variant calling with ShoRAH [137]); ii) co-occurrence analysis (detecting mutation co-occurrence) with the help of COJAC [6, 104]; iii) building of the list of signature mutations for the variants of concern, variant mutation table, and visualization including heatmaps and curves plots generation with the help of Jupyter; iv) generation of Python script for upload to CoV-Spectrum. The diagram that describes the Cowwid workflow is presented in fig. 9.

The advantage of Cowwid pipeline is its capability to detect lineages that are unknown as it uses COJAC tool as a co-occurance step. The other bright side of Cowwid is that its output was already experienced to be uploaded to CoV-Spectrum which means the accessibility of results and a variety of visualizations. Though, the Jupyter-based plots as well as the upload script look hard-coding, which is disadvantageous. That means that this pipeline is advanced but requires to be adjusted and generalized in order to use it for other than datasets described by K. Jahn et al. [6].

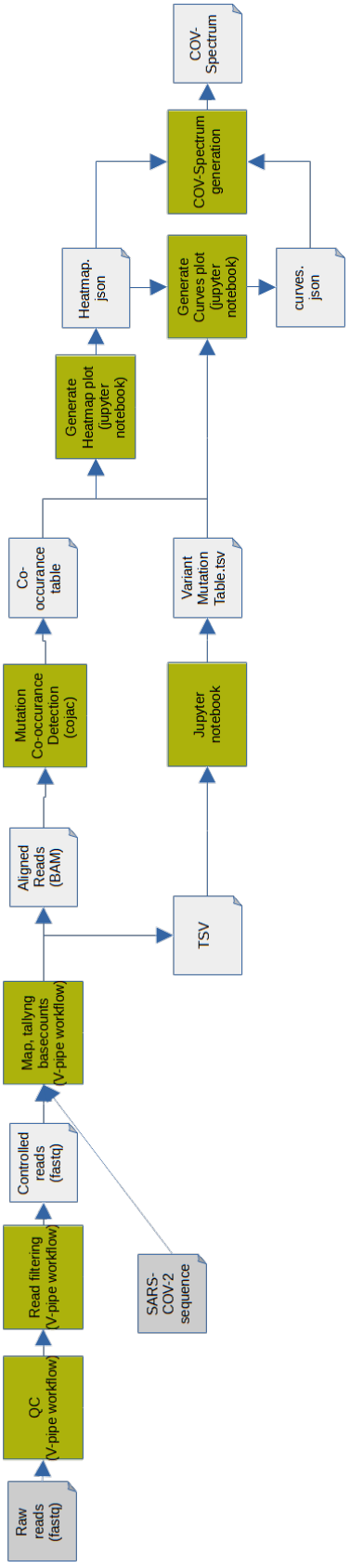


Figure 9: Cowwid workflow.

2.1.2.7 Pipeline described by R. Izquierdo-Lara et al.

Meanwhile, R. Izquierdo-Lara et al. [138] used 2 different workflows using two different workflow managers (Snakemake and Galaxy). For the Nanopore data, they used Snakemake to: i) demultiplex FASTQ raw reads using Porechop [139]; ii) trim of primers using Cutadapt [140]; iii) perform a reference-based alignment against SARS-CoV-2 reference sequence using minimap2; iv) confirm mutations in the genome by manually checking the alignment in UGENE [141] and resolve homopolymeric regions by consulting reference genomes. On the other hand, for the Illumina Sequence Analysis, they used Galaxy to: i) filter raw sequencing reads using Fastp [126]; ii) using BWA-MEM, remove adapter contamination, ambiguous bases, low quality reads (Phred score <30), and fragments <50 nt; then, map reads against GISAID sequence EPI_ISL_412973; iii) realign reads using FreeBayes [56]; iv) generate consensus sequences and variants using iVar [49]; v) finally, confirm variant calling by manual inspection of the aligned reads using UGENE.

For the phylogenetic analysis, sequences were aligned with >75% genome coverage using MAFFT [142], followed by visualization of trees with Figtree [143], and clades assignment with Nextclade tool [144]. The diagram that describes the workflow is presented in fig. 10.

A feature that this pipeline offers that others listed above in this section do not is phylogenetic analysis and visualization of the phylogenetic tree.

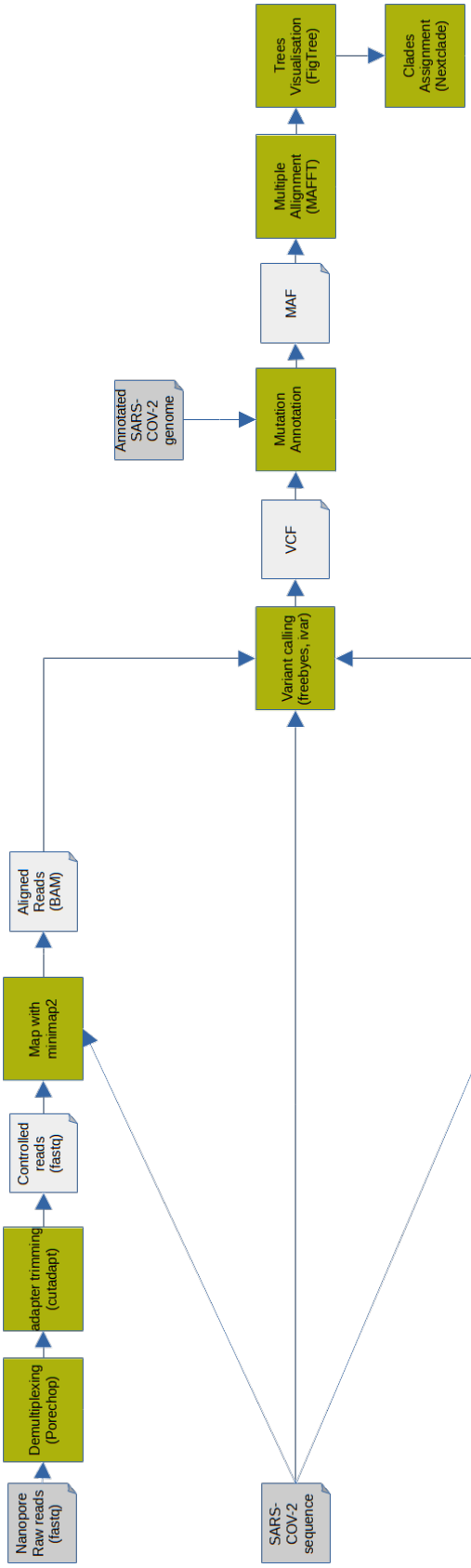


Figure 10: Workflow described by R. Izquierdo-Lara et al.

2.1.3 Comparison of methods for wastewater surveillance

To compare the existing tools and pipelines representing different state-of-the-art approaches, table 4 for individual tools and table 5 for standalone pipelines show some of them accompanied by source code link, license type, external tools used, description of inputs, produced outputs, their availability in Bioconda, availability in Galaxy before starting this thesis, and, finally, their prior usage (applications).

Name	latest version	Source code	License	Goal	External tools plugged	Input	Output	Available in Bioconda	Available in Galaxy	Applications
Freyja	1.3.11	[145]	BSD-2-Clause (open source)	Recover relative lineage abundances from mixed SARS-CoV-2 samples from a sequencing dataset (BAM aligned to the Hu-1 reference)	-	Variant call and sequencing depth information	TSV file that includes the lineages present, their corresponding abundances, and summarization by constellation	+	-	[86, 146, 147, 148]
COJAC	0.2	[104]	GPL-3.0 (open source)	Analyzing co-occurrence of mutations on amplicons	-	BAM/CRAM/SAM and BED file describing the amplified regions	Total count of amplicons carrying the sites of interest, amplicons carrying mutations on all site of interest, amplicons where one mutation is missing, fraction (ratio of number of all amplicons carrying mutations on all sites of interest to total number of amplicons carrying cites of interest)	+	-	[6, 148, 149]
LCS	1.2.0	[106]	-	Lineage decomposition for SARS-CoV-2 pooled samples	-	Variant-groups definitions, raw-fastq files pooled samples, fasta file of the primers used	Lineage decomposition outputs for SARS-CoV-2 pooled samples, plots	-	-	[106, 148]
Kallisto	0.48.0	[107]	BSD-2-Clause (open source)	Quantifying abundances of transcripts from RNA-Seq data, or more generally of target sequences using high-throughput sequencing reads	-	FASTQ (single-end or paired-end reads)	Abundance estimates, bootstrap estimates, and transcript length information length	+	+	[108, 111]
Alcov	-	[112]	-	Abundance learning for SARS-CoV-2 variants.	-	BAM file of reads aligned to the SARS-CoV-2 reference genome	Number of reads with and without each mutation in each sample, heatmap showing the frequencies for all samples, mutations, read depth for each amplicon, plots of amplicon GC content against amplicon depth	-	-	[112]
SAM refiner	1.4.0	[113]	GPL-3.0 (open source)	Gathering variant information from a SAM formatted files	-	SAM formatted files generated from sequencing mapping programs such as Bowtie2 or MiniMap2, FASTA formatted file for a reference sequence	Unique sequences, statistics about removed chimera, covariant deconvolution output	+	-	[113, 150, 151]

Table 4: Comparison table of existing individual tools for wastewater surveillance

Name	latest version	Source code	License	Goal	External tools plugged	Input	Output	Available in Bioconda	Available in Galaxy	Applications
Pipes et al.	1.0	[114]	GPL-3.0 (open source)	Estimating the relative proportions of SARS-CoV-2 strains from wastewater samples	Bowtie2 v2.4.5	FASTA file; MSA FASTA of SARS-CoV-2 reference strains	Estimated proportion of candidate strains, barplot with only those strains with an estimated proportion larger than 1%	-	-	[114]
Gromstole	1.0	[116]	MIT (open source)	To estimate the relative frequencies of different SARS-CoV-2 lineages	Minimap2 v2.24, Cutadapt v4.0	Paired-end reads in separate FASTQ/FASTA files, NC043312.fa as a reference genome	Counts of each mutation of the lineages, coverage at every position on the reference genome, estimate of the proportion (including 95% confidence interval);	-	-	[116]
AG	-	[117]	GPL-3.0 (open source)	Analyzing the within-sample genetic diversity of SARS-CoV-2 in Wastewater samples.	Trim Galore v0.6.7, Minimap2 v2.24, SAMtools v2.0.4, iVar v1.3.1, BEDTools v2.27, Picard v2.18, FreeBayes v1.1.0.46, SnpEff v4.3	Paired-end .fastq files for each sample	Tabular files with scanned variants, common depth report	-	-	[117]
Lineagespot	-	[118]	MIT (open source)	Identify SARS-CoV-2 related mutations based on a single (or a list) of variant(s) file(s) (i.e., variant calling format).	Trim Galore v0.6.7, FastQC v0.73, Minimap2 v2.24, SAMtools v2, BEDTools v2, Picard v2.18, FreeBayes v1.1, SnpEff v4.3	Raw fastq FASTQ files, anf for further analysis - VCF and file containing all lineage-assignment rules, as retrieved from the pangolin tool repository	VCF, MAF files, eport about lineage abundances	-	-	[118]
PiGx	0.0.8	[125]	GPL-3.0 (open source)	Analyzing data from sequenced wastewater samples and identifying given lineages of SARS-CoV-2	iVar v1.3.1, Fastp v0.23.2, BWA v0.7.17.2, MultiQC v1.11, LoFreq v2.1.5, VEP , Kraken2 v2.1.1, Krona v2.7.1	Raw reads and the additional information about used primers and adapters	VCF, visualization of the development of virus over time, variants report, taxonomic classification	-	-	[125]
Cowwid	0.7.1	[103]	-	Surveillance of SARS-CoV-2 genomic variants in wastewater	V-pipe v2.99.2, FastQC v0.73, PRINSEQ v0.20.4, VICUNA, BWA-MEM v0.7.17, ShoRAH, COJAC v0.2	TSV table of the samples, YAML with definition of the variants, BED with amplicon positions for ARTIC, BED with amplicon description	YAML/CSV files with coocurrences, variant mutations table, plts to integrate to CoV-Spectrum	-	-	[103, 152]
Izquierdo-Lara et al.	-	[138]	-	Monitoring SARS-CoV-2 Circulation and Diversity through Community Wastewater Sequencing, the Netherlands and Belgium	Porechop v0.2.4, Cutadapt v4.0, UGENE, Fastp v0.23.2, BWA-MEM v0.7.17, FreeBayes v1.1, iVar v1.3.1	Raw reads and the additional information about used primers and adapters	VCF, MAF files, Trees visualization (Fig tree)	-	-	[138]

Table 5: Comparison table of existing standalone pipelines for wastewater surveillance

2.2 Discussion of existing methods

A number of methods were described above that use different models to determine SARS-CoV-2 lineages and their abundances in wastewater samples. Several of the methods listed can be applied to the purpose of this thesis, i.e., improvement of existing Galaxy workflows for analyzing SARS-CoV-2 clinical samples, repurposing them to detection of SARS-CoV-2 variants in wastewater samples as well as discerning of lineage abundances in these samples. It should be pointed out that, in section 2.1.1, separate tools are listed that focus directly on lineage abundance computation, whereas, in section 2.1.2, complete and almost complete standalone pipelines like PiGx, Lineagespot, Cowwid, etc. are listed that include the entire raw reads bioinformatics analysis. Standalone pipelines are not intended to be improved in this work because they cannot be incorporated into existing Galaxy workflows for SARS-CoV-2 surveillance since Galaxy workflows offer their own independent data preprocessing. Conversely, individual tools that are capable of being plugged into currently available Galaxy workflows and performing lineage abundances computation are the interest of this thesis.

Two tools, Freyja [145, 86] and COJAC [103], were a choice to use in the workflows to achieve the goal of this thesis because their use in various projects (such as wastewater monitoring of SARS-CoV-2 in the Center for Food Safety and Applied Nutrition in US [147], wastewater monitoring of SARS-CoV-2 variants in Switzerland [6] and in England [149]) has shown decent results; and because of their ability to be integrated into existing Galaxy pipelines that use Galaxy workflow manager, which meets all crucial characteristics like transparency, accessibility, and reproducibility. Furthermore, Galaxy pipelines perform decent data preprocessing steps and have shown promising results on clinical data, making them worthy of being repurposed for wastewater data using tools such as Freyja and COJAC.

Interestingly for this thesis, these tools are different in purpose; for example, Freyja's output is easy to interpret without some specific knowledge, while COJAC requires more knowledge to be used. First of all, it demands some knowledge in programming as the output of COJAC is quite raw so far and has to be processed for downstream analysis like plotting and integrating to remote platforms such as CoV-Spectrum. Second of all, the usage of COJAC requires some knowledge of virology. Despite this, COJAC provides more detailed information and is able to detect unknown variants. These two different tools, Freyja and COJAC, target different user groups (e.g., politicians and researchers), which is interesting for this thesis's research purposes.

3 Methods

In the following sections, the technical details of the methods used within the thesis will be discussed. First, an evaluation of needs, including currently available data, will be described in section 3.1.1, followed by an explanation of existing Galaxy workflows in section 3.1.2. Then, section 3.1.3 will discuss improvements of existing steps in these workflows and the addition of new steps. The following section 3.2 will expose metrics for the evaluation of workflows and the characteristics of their comparison to other methods. Particularly, section 3.2.1 will give details of benchmarking with regard to the other state-of-the-art approach. Then, section 3.2.2 will give an explanation of benchmarking results produced by both workflows on chosen real-world datasets.

3.1 Workflow reengineering

3.1.1 Evaluation of the needs

3.1.1.1 Characteristics to be considered for improvement

The purpose of this master thesis is to create Galaxy workflows to detect lineages and sub-lineages of SARS-CoV-2 in wastewater samples as well as the co-occurrence of different lineages in the sample. The workflow development was focused on the following questions to be able to: i) delineate lineages and sublineages; ii) delineate mixtures of at least two lineages of varying frequencies; iii) detect lineages at very low frequencies in mixtures; iv) delineate recombinants [153]; v) detect in low coverage samples; vi) detect with short-read sequencing data; 7) detect with long-read sequencing data.

3.1.1.2 Available data

For the purpose of this thesis, datasets from ENA database [154] were checked and analyzed. The ENA sequence read archive contained 13,893 raw read SARS-CoV-2 in wastewater datasets at the end of May 2022. Real-world data were obtained using library preparation techniques: i) ampliconic-based, and ii) metatranscriptomics (including Targeted-Capture [155], RNASeq, and whole genome sequencing (WGS)); and sequenced using platforms: i) Illumina, ii) Nanopore, iii) ION-Torrent, iv) BGISEQ [156], and v) DNBSEQ [157].

The main statistical numbers are presented in table 6. The amount of data influenced the decision for creating methods to analyze SARS-CoV-2 data with Galaxy workflows. Illumina-based Ampliconic data remain to be the most commonly obtained for wastewater samples (as it was for clinical data [69]).

The heatmap plot in fig. 11 was created with Python (version 3.9.12) using Seaborn visualization library (version 0.11.2) [158] based on Matplotlib (version 3.5.1) [159] to depict the number of wastewater samples received using different sequencing methods and library preparation strategies that are publicly available in ENA. Obviously, Illumina-based Ampliconic is the

Library Strategy	Sequencing platform	Library Layout	Number of accessions	Number of samples
Ampliconic:	BGISEQ	paired-end	1	48
	DNBSEQ	paired-end	1	59
	Illumina	paired-end	21	8,293
	Ion Torrent	single-end	4	65
	Nanopore	single-end	4	169
Ampliconic Total			**31**	**8,634**
Metatranscriptomics:				
RNA-Seq	Illumina	paired-end	1	173
Targeted-Capture	Illumina	paired-end	1	11
WGS	Illumina	paired-end	12	5,146
	Nanopore	paired-end	3	122
Metatranscriptomics Total			**17**	**5,452**
Total			**49**	**14,086**

Table 6: Overall numbers of accessions and samples of different types of data for SARS-CoV-2 in wastewater datasets in ENA. The table was created using a PivotTable in Google Sheets. Metadata about real-world datasets were uploaded from ENA to Google Sheets, then, Pivot table was generated to analyze numerical data. First, data was sorted by Library strategy, following with sorting by Sequencing platform, Labrary Layout. Next, count of accessions and sum of samples were calculated.

most common type of data sequencing, followed by Illumina-based with WGS library preparation strategy. To address this, I developed workflows for Illumina-based data in this master's thesis, both for ampliconic and metatranscriptomic library preparation techniques.

3.1.2 Evaluation of the existing Galaxy workflows

Galaxy was used as a workflow manager for this thesis. The Galaxy facilitates open and accessible platforms that can ensure data analysis transparency and reproducibility. There are three global Galaxy instances where tools and workflows can be accessed immediately and for free. Thousands of users can run hundreds of thousands of analyses per month on each system. Using the service, the user will have access to as much computation as they need (with a limit on the number of simultaneous analyses) and 250GB of disk space, which can be increased based on usage.

Moreover, Galaxy provides infrastructure that already contains loads of existing tools, and tools can be added to Galaxy with the help of Planemo [160]. It serves as a decent basis for workflow development. Additionally, Galaxy provides automated bots [76] for genome data surveillance which opens the opportunity to be used for SARS-CoV-2 wastewater surveillance workflows in order to analyze data regularly.

The four existing workflows are taken as a basis for this thesis. Diagrams that show these workflows step by step are shown in fig. 12 and fig. 13 for the one taken as basis in this thesis, and in fig. 37 and fig. 38 in section 6.8.1 that were not taken as basis in this thesis. They show trustworthy results on clinical SARS-CoV-2 data and are expected to perform well on wastewater data.

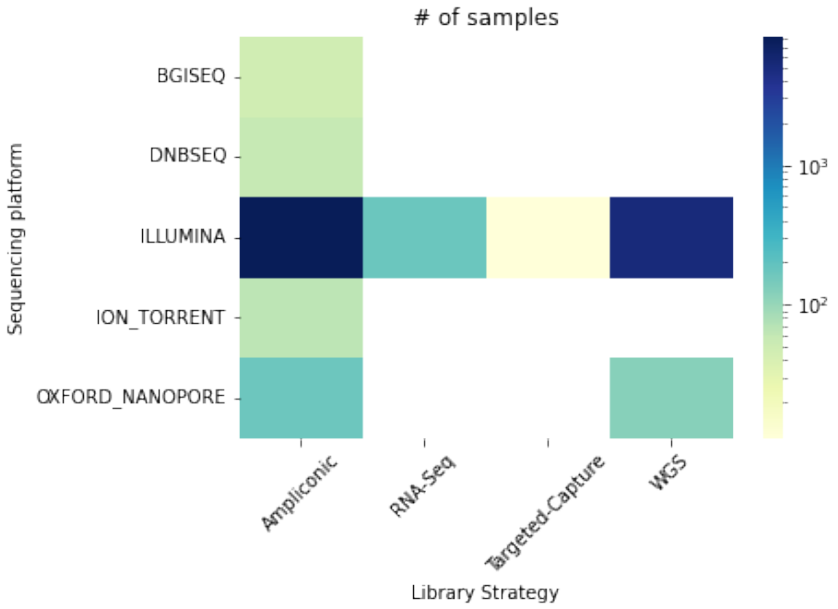


Figure 11: Number of samples available for each sequencing platform and library preparation strategy for SARS-CoV-2 in wastewater according to ENA archive as of May 2022. BGISEQ, DNBSEQ, ILLUMINA, ION_TORRENT, and OXFORD_NANOPORE refer to corresponding sequencing technologies. Ampliconic library strategy refers to library preparation based on amplicons, while RNA-Seq, Targeted-Capture, and WGS refer to types of metatranscriptomic-based library preparation strategy.

Depending on available data and the prevalence of Illumina sequenced data, two out of four Galaxy existing workflows are improved: Illumina-ampliconic oriented (fig. 12) and Illumina-metatranscriptomics oriented (fig. 13).

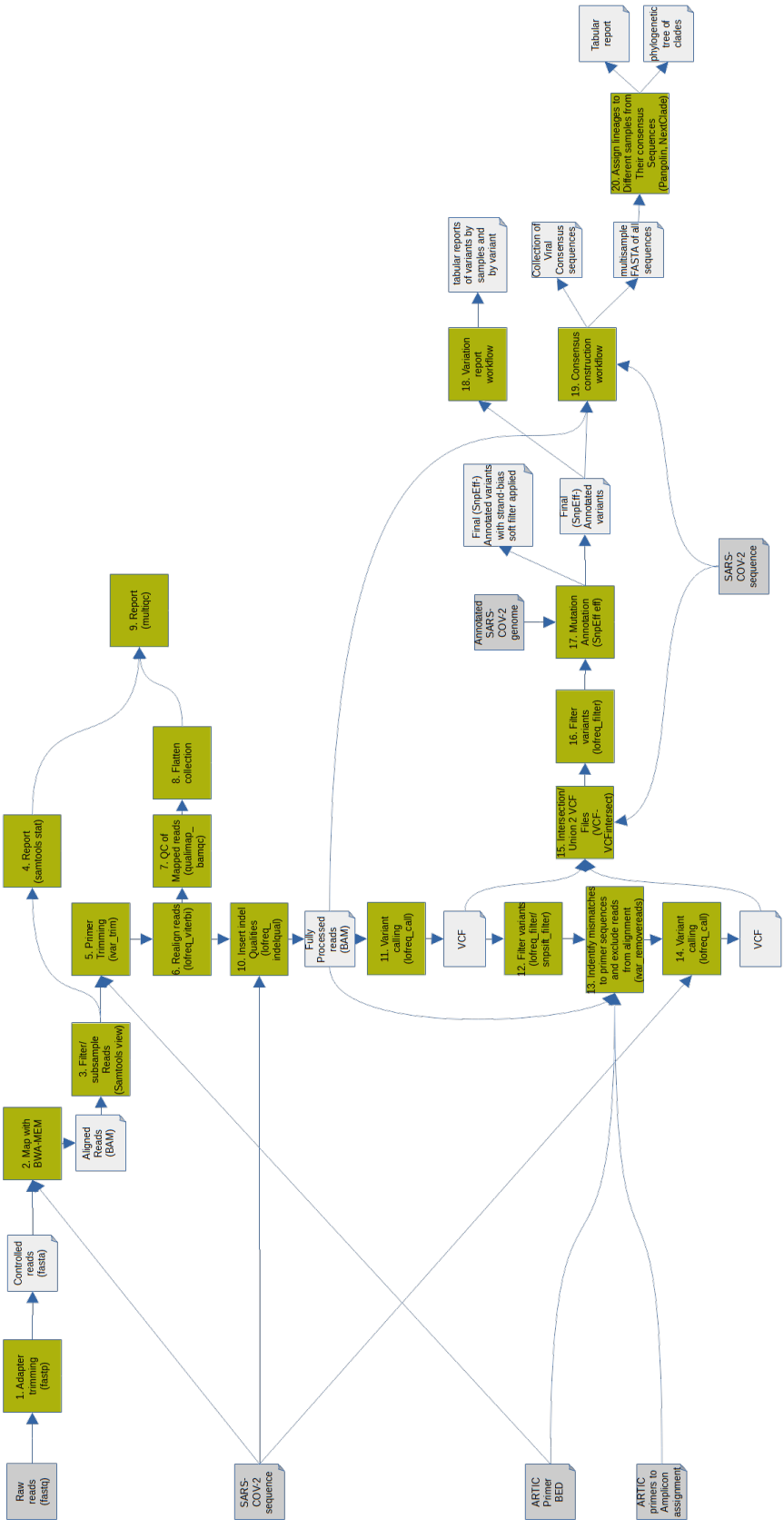


Figure 12: One of four existing Galaxy workflow for SARS-CoV-2 clinical data surveillance for paired-end reads data extracted with amplicon-based technique and sequenced with Illumina sequencing approach.

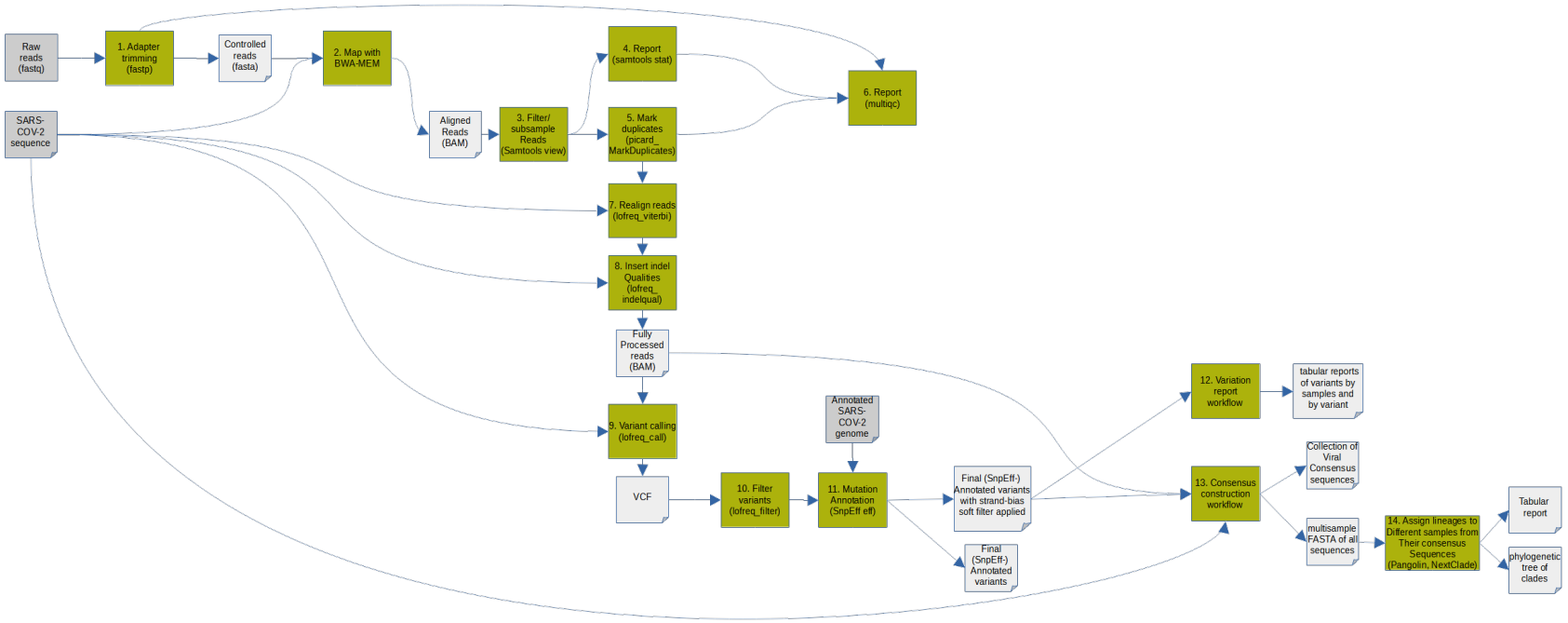


Figure 13: One of four existing Galaxy workflow for SARS-CoV-2 clinical data surveillance for paired-end reads data extracted with ampliconic-based technique and sequenced with Illumina sequencing approach.

One of the main influences on the decision to take these workflows as a basis for the work in this thesis is preprocessing steps. Preprocessing steps from these workflows were used: i) quality control, ii) adapter trimming, iii) mapping to reference sequence, iv) primer trimming (for the one ampliconic-based where primers were used), v) steps that remove duplicate reads and realign them afterward with reference sequence to prepare data for vi) variant calling. In other words, the sub-workflow for preprocessing data until receiving the VCF file is taken to be: i) improved and ii) extended with the aim of delineating SARS-CoV-2 variants in wastewater samples.Tools differ for different types of datasets and depend on the library preparation approach and sequencing technique as well as on the types of reads obtained (paired-end or single-end).

In the following section 3.1.3, changes applied to existing Galaxy workflows will be shown, starting with expected input in paragraph 3.1.3.1, expected output in paragraph 3.1.3.2, followed by paragraph 3.1.3.3 and paragraph 3.1.3.4 where new steps added will be described. Finally, paragraph 3.1.3.5 will give a picture of workflows improved and repurposed for wastewater surveillance in accordance with the purpose of this thesis.

3.1.3 Changes

3.1.3.1 Input data

Expected input for both ampliconic-based and metatranscriptomic-based workflows consists of i) a collection of paired-end reads samples of the considered dataset in FASTA/FASTQ format; ii) FASTA file of reference SARS-CoV-2 sequence. Additionally, both workflows recommend providing, optionally, though: iii) UShER bar-codes file for lineage assignment used by Freyja; iv) descriptions of variants of concern in JSON format used by COJAC. Inputs 3 and 4 are provided by a workflow as well. However, they can be outdated because the updates for both tools happen with a delay. It can happen in case of new variant appearance that up-to-date variants description files are required; hence, the opportunity of submitting them by hand directly to workflow is given.

Moreover, for the workflow that is supposed to work with ampliconic-based datasets extracted using ARTIC protocol, the following input data is necessary: v) BED file with ARTIC primers description; vi) description file of ARTIC primers to amplicon assignments that is ued by iVar trim and iVar removereads for assigning primers to amplicons.

3.1.3.2 Output data

In both (ampliconic-based and metatranscriptomic-based) workflows, two distinct branches are developed. The first branch focused on using Freyja tool and its outputs, while the second branch produces results based on COJAC tool usage. The user is provided to use only one of these branches as well as both of them simultaneously. Therefore, depending on the branch run in this workflow, different outputs will be obtained.

The first option is based on Freyja tool for the delineation of SARS-CoV-2. Due to Freyja's use of the UShER phylogenetic tree, the output of the workflow is high-level rough reports, the proportions of specific lineages (e.g. Omicron, Delta, etc.), and their co-occurrences. Reports in this branch can be made in PDF and HTML interactive formats. Additionally, if the metadata

file describing samples metadata (date of collection, viral load, etc.) is provided by a user, various plots show the dynamics of SARS-CoV-2 over time.

The second branch, in turn, is based on COJAC tool for delineation. These tools provide more detailed information about variants of concern. It is possible to describe new variants in COJAC that can turn into VOCs. This feature makes it possible to use this tool and the COJAC-based workflow to connect with CoV-Spectrum. Even though the script for upload to CoV-Spectrum is not automated yet, it can be applied separately based on tabular output from COJAC.

3.1.3.3 Improvement of preprocessing subworkflows

The initial phase of workflows development in this thesis is to take subworkflows from existing workflows responsible for data preprocessing and improve them, taking into account the goal of wastewater surveillance.

One of the improvements made in workflows is the addition of a decontamination step. There were several reasons why a decontamination step was added to the workflow. The first and foremost reason is to eliminate any potential concerns about the anonymity of the host of the virus. Considering that SARS-CoV-2's hosts are predominantly humans and in this master thesis, I am working with wastewater which by default contains the human genome, that seemed to be a good idea to decontaminate samples from the human genome in the beginning. Additionally, this step is added to improve the runtime and accuracy of mapping from the sequence of a virus to a reference virus sequence.

ReadItAndKeep tool [54] was added after adapter trimming (with Fastp tool) for both workflows to only keep reads matching the SARS-CoV-2 genome so that host reads could be removed from sequencing data for SARS-CoV-2.

Another step added to both workflows is a taxonomic analysis of reads that are unmapped to reference SARS-CoV-2. I added two tools, Kraken 2 [129, 130, 131] and Visualization with Krona chart [161, 162].

The tool Kraken 2 is a taxonomic sequence classifier that classifies reads according to their taxonomic labels. In order to accomplish this, it examines the k-mers within a read and queries a database with them. A taxonomic tree of all genomes that contain a given k-mer contains a mapping from Kraken's genomic library to its lowest common ancestor (LCA). To create a taxonomic label for a read, the LCA taxa corresponding to its k-mers are analyzed. Labels can refer to any of the taxonomic nodes. In workflows, the viral genomes database was applied.

The other tool, Krona chart, is an intuitive visualization tool that helps visualize relative abundances and confidences in multiple metagenomic classifications. Radial, space-filling displays are combined with parametric coloring and interactive zooming in polar coordinates in Krona. Using HTML5 and JavaScript, the charts can be interactively explored by any Web browser [161].

It is possible to use these tools to learn about other micro-organisms that can be found in wastewater samples in addition to SARS-CoV-2. A taxonomical analysis, as well, opens the possibility of detecting newly emerged variants of SARS-CoV-2 because otherwise, they would not be able to match reference sequences and, therefore, will not be noticed. By taking this additional step, we close not only the interest in other species that wastewater samples contain

and can be useful outside of SARS-CoV-2 topic but also contribute to the possibility of catching new variants of SARS-CoV-2.

One other step that should be mentioned is computing sequence depths with Samtools depth [120]. It was added into both workflows after mapping to the reference step because the information about sequencing depths is required as input for the demixing command of Freyja tool.

3.1.3.4 Tool integration

Apart from the improvement of the existing preprocessing sub-workflow, new steps were added with the intention to reaching the goal of this thesis: to identify SARS-CoV-2 lineages in wastewater samples. In this regard, I have added delineating steps to both branches of the workflow based on Freyja (version 1.3.8) and COJAC (version 0.2) tools. These tools were chosen as individual tools that can be plugged into Galaxy workflow and represent different types of SARS-CoV-2 lineages abundances analysis and outputs that can be benchmarked afterward with each other as well as with outputs from the outside non-Galaxy standalone workflow for evaluation.

To facilitate this master thesis, I developed Galaxy wrappers for tools that did not have one. For that, I used Planemo [160], a command-line application for creating Galaxy tools, workflows, and training materials. Planemo is also used to deploy tools and workflows (tools to Galaxy ToolShed; workflows to Intergalactic Workflow Commission) as well as execute Galaxy-based analyses by command line if one prefers that interface to Galaxy's graphical interface.

In designing both tools, detailed consideration was given to the user interfaces. A wrapper was designed not only to accommodate the needs of this master thesis research but also to be able to be used by researchers who are interested in using both tools independently on any Galaxy instance and also included in the specific workflow. While designing tool wrappers, enough effort was put in, in order to make the interface convenient as well as functional.

For example, features like samples name autodetection and the option to provide samples names by the user, or the option to use variant descriptions of SARS-CoV-2 variants of concern cached in Galaxy as well as the option to provide the user's preferred list of variant descriptions were included. Another example is the ability to combine different types of output plots for Freyja. Additionally, for COJAC, the possibility of receiving various output tables (e.g. line-oriented, column-oriented, or multilevel-oriented) was provided.

Overall, for Freyja tool, there were four wrappers created: i) "Freyja: Call variants and get sequencing depth information", although, it was not needed for this thesis because Galaxy workflows are elaborated to call variants in a more precise way; ii) "Freyja: Bootstrapping method", which was not used in this thesis either but was designed anyway for the perspective usage by interested users; iii) "Freyja: Demix lineage abundances"; iv) "Freyja: Aggregate and visualize demixed results". Throughout this thesis, the last two tools were applied to both workflows. All Freyja tools are combined in the collection and maintained via macros. All Freyja tool wrappers are available under open source licence (links are in section 6.1).

Withal, for COJAC tools, there were three wrappers designed: i) "Cojac: mutbamscan" scan an alignment file for mutation co-occurrences, where the output is generated in JSON, YAML,

and/or CSV table format; ii) "Cojac: tabmut" export cooccurrence mutations as a table that interprets Cojac mutbamscan into nicer and readable tabular format; iii) "Cojac: pubmut" render a JSON or YAML file to a pretty table that produces output in CSV and/or HTML format which is pretty enough to be included in publications. All COJAC tools are combined in the collection and maintained via macros. All Cojac tool wrappers are available under open source licence (links are in section 6.1). Cojac: pubmut was added to Galaxy only for the case of potential interest in it outside workflows and was not directly used in the workflows developed to achieve the thesis's goal. Cojac: mutbamscan and Cojac: tabmut, in their turn, were used in the workflow corresponding to data obtained with the ampliconic library preparation method. Although COJAC can work with ampliconic data, it was not created to work with metatranscriptomic data.

Once wrappers are ready and tested, I submitted them to Galaxy IUC by contributing to the GitHub repository (available in the fork `https://github.com/PlushZ/tools-iuc`) to make Freyja and COJAC available to Galaxy. The tools were also added to Galaxy Europe.

3.1.3.5 Workflows step by step

The final workflows developed in this master thesis are depicted in fig. 14 and fig. 15 for Illumina ampliconic paired-end SARS-CoV-2 wastewater surveillance and for Illumina metatranscriptomic paired-end SARS-CoV-2 wastewater surveillance, respectively. In green, steps kept from existing Galaxy workflows for clinical SARS-CoV-2 data surveillance are depicted. Newly added steps for the target of this master thesis are indicated in orange. On the diagram, there is also a step called CoV-Spectrum colored in pink. So far, this step is not automated and can be done only manually. Nonetheless, future improvements outside of this thesis could automate this step.

Developed Galaxy workflows are publically available via Galaxy published resources and ready to use. Links to published workflows are in section 6.2.

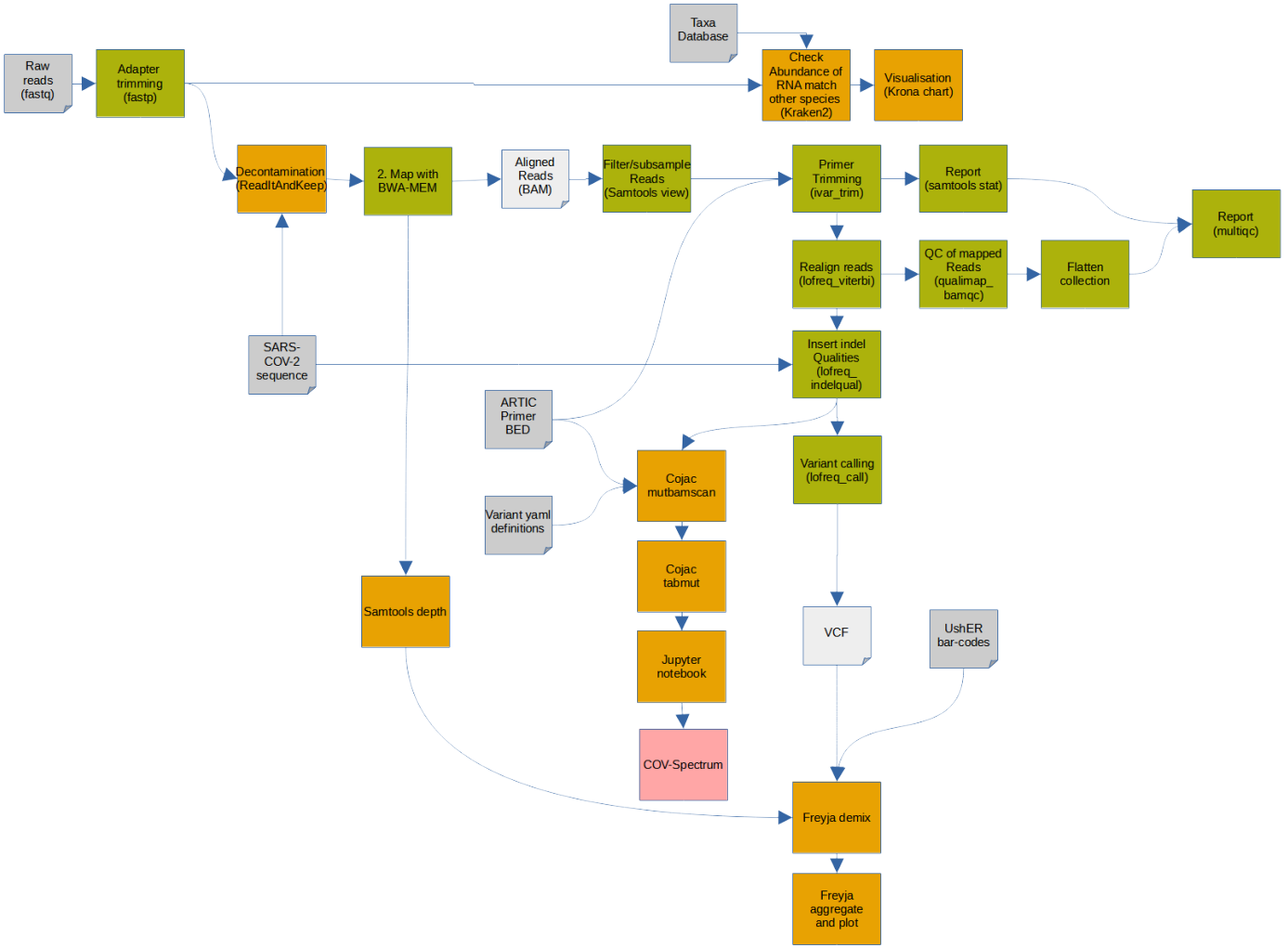


Figure 14: Improved and repurposed to SARS-CoV-2 wastewater surveillance Galaxy workflow for paired-end data extracted with ampliconic-based technique and sequenced with Illumina sequencing approach. Newly added steps for the target of this master thesis are indicated in orange, a step CoV-Spectrum that is not automated and can be done manually is colored in pink.

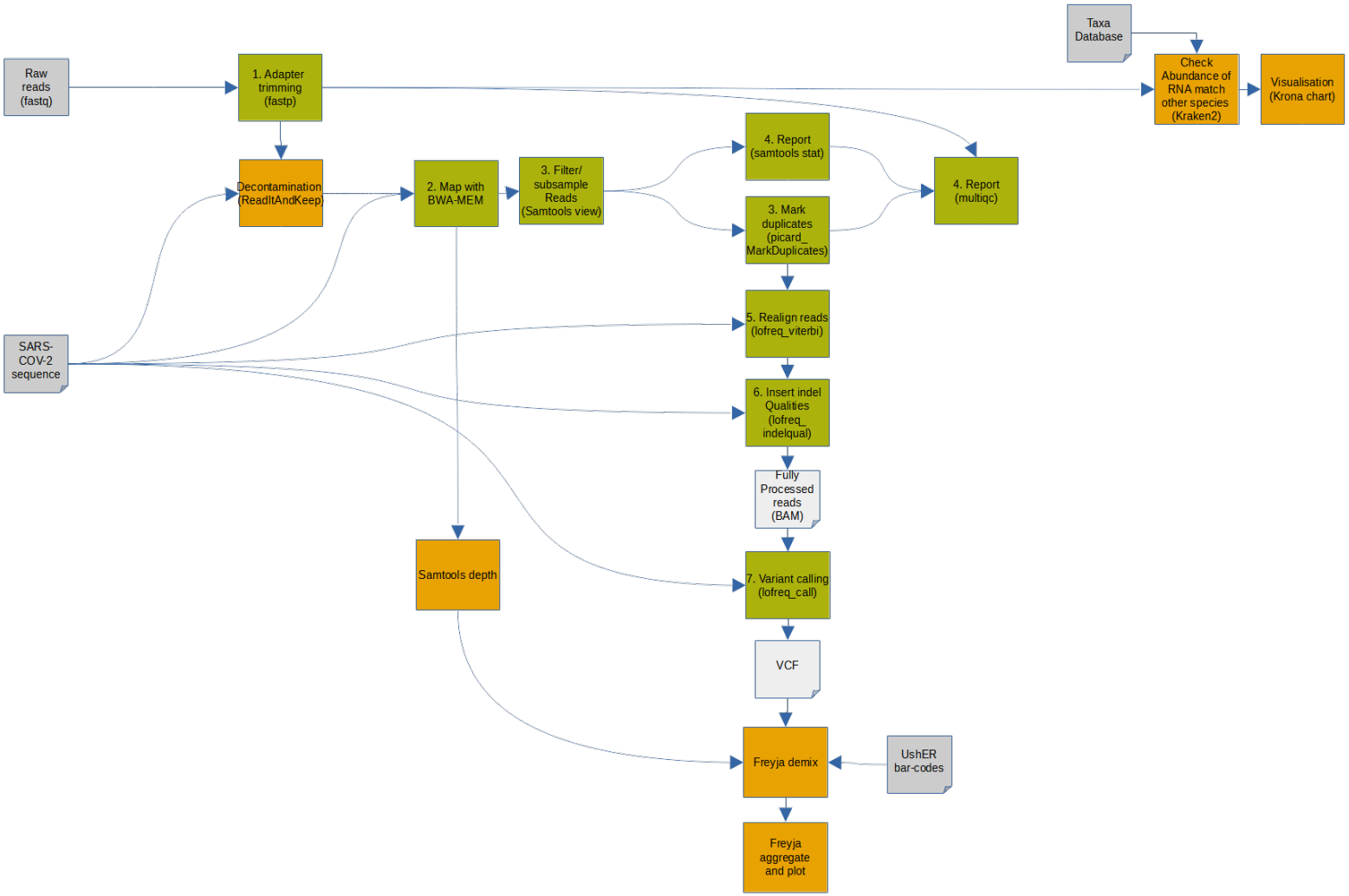


Figure 15: Improved and repurposed to SARS-CoV-2 wastewater surveillance Galaxy workflow for paired-end data extracted with metatranscriptomics-based technique and sequenced with Illumina sequencing approach. Newly added steps for the target of this master thesis are indicated in orange.

3.2 Workflow evaluation

To assess the quality of developed workflows, I used a synthetically generated mock dataset and compared results obtained by the developed Galaxy workflow with expected ground truth results from the information about generated mock dataset. In addition, results from Galaxy workflow were benchmarked with the results from another solution, namely Lineagespot. Afterward, I launched both workflows on four real-world datasets mentioned in section 3.1.1.

3.2.1 Evaluation on mock data

3.2.1.1 Generated mock datasets

Generated synthetic dataset was created by a working group that is interested in evaluation of wastewater SARS-CoV-2 identification methods. The dataset is uploaded in Github repository available under an open source license [163]. Synthetic datasets were created by obtaining genomes from the Pango designation list. While generating synthetic mock data, the focus was on 3 real variants (BA.1, BA.2, Delta), one synthetic 'background' (BG) variant to mimic the effect of an unknown variant, as well as recombinant genomes (Omicron-Delta) [153, 30]. Synthetic 'amplicons' were created using ARTIC v4.1 primers. Synthetic reads were created using error models based on 150 or 250bp NovaSeq reads. There were created samples belonging to groups of different classes such as: i) Single lineage, Two lineages, Three lineages; ii) High coverage, Low coverage; iii) Long reads, Short reads. In table 7, classes used to generate synthetic mock dataset are listed.

Number of lineages	**Lineages**	**Coverage**	**Read length, bp**	**Number of samples**
Single lineage	BA.1	high	150	2
		low	150	2
			250	2
	BA.1 Total			**8**
	BA.2	high	150	1
			250	1
		low	150	1
			250	1
	BA.2 Total			**4**
	Delta	high	150	1
			250	1
		low	150	1
			250	1
	Delta Total			**4**

Number of lineages	**Lineages**	**Coverage**	**Read length, bp**	**Number of samples**
Single lineage	Recombinant		250	1
			250	1
	Recombinant Total			**2**
	Synthetic lineage	high	150	1
			250	1
		low	150	1
			250	1
	Synthetic Total			**4**
Single lineage Total				**22**
Two lineages	BA.1:BA.2	high	250	3
		low	250	3
	BA.1:BA.2 Total			**6**
	BA.1:Delta	high	150	7
			250	14
		low	150	7
			250	14
	BA.1:Delta Total			**42**
Two lineages Total				**48**
Two lineages + Synthetic lineage	BA.1:Delta:Synth	high	250	7
		low	250	7
	BA.1:Delta:Synth Total			**14**
Three lineages	BA.1:BA.2:Delta	high	250	4
		low	250	4
	BA.1:BA.2:Delta Total			**8**
Three lineages + Synthetic lineage	BA.1:BA.2:Delta:Synth	high	250	4
		low	250	4
	BA.1:BA.2:Delta:Synth Total			**8**
Total				**100**

Table 7: Overall numbers of samples, sorted by various groups of samples from the generated synthetically mock dataset. The table was created using a PivotTable in Google Sheets. Metadata about samples in the mock dataset was uploaded to Google Sheets, then, Pivot table was generated to analyze numerical data. First, data was sorted by Number of lineages, followed by sorting by exact Lineages combination, Coverage, and Read length. Next, the sum of samples for every class was calculated.

Overall, the mock dataset contains 100 samples. More specifically, it contains 22 samples with single lineage, 42 with two distant lineages (BA.1 and Delta), 6 samples with two close lineages (BA.1 and BA.2), 14 samples with 2 known and one unknown lineage (synthetically generated), 8 samples with known lineages, and 8 samples with 3 known and one unknown lineages. In terms of coverage, the mock dataset contains 50 samples of low coverage and 50 samples of low coverage. In addition, 24 samples use long reads (250 base pairs) and 76 samples short reads (150 base pairs).

In order to evaluate and draw conclusions about the performance and efficiency of developed workflows, the workflow for ampliconic-based data was run on a mock dataset along both branches: Freyja-based and COJAC-based. Aside from Galaxy, Lineagespot standalone workflow was launched in parallel on the same mock dataset for benchmarking all three methods' results.

3.2.1.2 Comparison of proposed workflows

To compare two branches of the proposed Galaxy workflow (with Freyja and with COJAC) and evaluate Galaxy workflow results, I examined them with the known expected results. Knowing the ground truth results allows me to draw conclusions about the performance of developed workflows and their accuracy.

For further evaluation, resulting data were divided according to the principle of belonging to a particular expected group. The groups of greatest interest were: Single lineage expected to be found in the sample and Two lineages.

Even though both Freyja and COJAC methods found extra lineages that were not expected, I focused on their capability to detect expected lineages (Delta, BA.1, and BA.2) in abundance proportion closer to the expected proportion.

3.2.1.2.1 Preprocessing data

Data obtained by Freyja and COJAC were preprocessed to make them comparable. Data from Freyja were obtained in TSV (tabular) format (as shown for sample1 example in listing 1). To distinguish variants Delta, BA.1, and BA.2, which are focused on the mock dataset, a simple Python script (provided in section 6.4) was used considering that Delta lineage as B.1.617.2 and its sublineages starting with AY according to Pango designation [35, 164]. In COJAC, in its turn, proportions of lineage abundances are computed for every amplicon, and for every lineage, a median of abundance proportion among different amplicons was computed.

```
1 Sample name  summarized  lineages  abundances  resid coverage
2 sample1  [('Omicron', 0.6999889999983451), ('Delta',
     0.22408099987442534), ('Other', 0.07552100018490182)] BA.1.18 AY.4 BA
     .1.19 BA.1.1.13 BA.1.15.1 AY.38 BA.1.9 BA.1.16 B B.1.617.2 B.1.1.529
     XS  0.23943700 0.11764700 0.11363600 0.10000000 0.09667000 0.06944400
     0.06686400 0.06474800 0.06122400 0.03699000 0.01863400 0.01429700
     7.611495978 99.95971667
3
```

Listing 1: Freyja output for sample 1 from mock dataset

To work with resulting output data from both branches of Galaxy workflow, they were stored in google sheets. Python pandas (version 1.4.2) data analysis tool [165] was used to access data and create dataframes from the data obtained. Dataframes were then used for downstream analysis.

3.2.1.2.2 Comparison

Parallel coordinates plot was created for detected proportions of different lineages using Python and Plotly graphing library (version 5.6.0) [166]. Three plots were generated for Delta, BA.1, and BA.2 lineages. First, I plotted all samples in one plot without their division into groups. All 100 samples were included in the first three plots in order to get an overall imagination of lineage proportion distribution among samples and their comparison to expected proportion values. Plots are available in paragraph 6.8.2.1 of section 6. The first axis is the Expected proportion, the second is the result obtained by COJAC, and finally, the third is Freyja's proportion of the considered lineage of SARS-CoV-2. Additionally, I generated one plot per considered lineage (Delta, BA.1, or BA.2) and per group of samples (Single lineage expected, Two lineages expected). Overall, 6 plots were created.

To zoom up, the focus on one sample was done. Hence, there were generated for every sample: parallel coordinates plot and bar plot using Seaborn library (version 0.11.2), and line plot using Matplotlib (version 3.5.1), to look at absolute values of lineage proportion against scaled values in parallel coordinates. The comparison with expected lineage proportions was included in these plots.

To compare Freyja and COJAC branches of workflow and to assess results with expected ones, I took a look at the distribution of Delta lineage proportion in the sample (obtained by Freyja, COJAC, and expected proportion) across all samples. The same comparison was made for BA.1 and BA.2 lineages.

Other angle to look at results is to see a distribution of proportion of all three considered lineages across samples for both Freyja and COJAC. Thus, a distribution plot using Python (version 3.9.12) and Seaborn library (version 0.11.2) was generated. To be said, the similar distribution plot of proportion of all three considered lineages across samples for Lineagespot was generated analogously to plots for Freyja and COJAC.

3.2.1.3 Comparison with lineagespot

For the comparison of proposed Galaxy workflow results with Lineagespot results, the same synthetic mock dataset was used. Lineneagespot results on the mock dataset were obtained following private communication with F. Psomopoulos [118]. It was expected that the results of both branches of Galaxy workflow turned out to be different from the results of Lineagespot. That is because different approaches (sub-workflows) are used for the variant calling steps. In the case of Galaxy workflow, LoFreq variant caller is used preceding several steps for primer trimming, realignment, and insertion of indel qualities (using LoFreq_insertindel). On the other hand, Lineagespot uses duplicates filtering with Picard tool, followed by variant calling with Freebayes tool. For lineage annotation, also different approaches are used. While for Galaxy workflow, Frejya or COJAC are options, Lineagespot pipeline uses its own approach for

delineation. Thus, results from Galaxy workflow, whether Freyja-based or COJAC-based, are expected to differ.

Data obtained from Lineagespot was already preprocessed, proportions of expected lineages were provided as well as assignments to certain lineages. To compare the results of Lineagespot and Galaxy workflows, Freyja-based or COJAC-based, against each other and against expected results, I focused specifically on two groups of samples from the mock dataset: the group where single lineage was expected (22 samples) and the group where two lineages were expected (48 samples).

3.2.1.3.1 Single lineage group of samples

For a single lineage group of samples, four characteristics were used. I intended to learn in which samples were detected: i) expected lineage; ii) expected+unexpected lineages; iii) unexpected lineage; iv) nothing. For each of these characteristics, bar plots were created separately, as well as a combined barplot that shows the number of samples that satisfy each characteristic. Also, a barplot was generated showing the percentage of samples that belong to each characteristic.

Further, considering only samples where expected lineage was detected, the distribution of lineage proportion detected by Freyja and COJAC among samples of the Single lineage group was plotted with Python with the help of Seaborn library (version 0.11.2). Only Freyja and COJAC were compared. I do not consider Lineagespot here because in the results of Lineagespot I do not have a clean proportion distribution between lineages. There is no information about other lineages detected (aside from BA.1, BA.2, Delta), while the sum of proportions of these three is not equal to 1 in Lineagespot.

To better understanding relationships between sets of different methods (Freyja, COJAC, Lineagespot) results, Venn Upset diagram was generated. The goal was to look at intersections between sets of results and which samples were detected correctly (in terms of lineages expected) by which tool and to find similarities in tools performance.

3.2.1.3.2 Two lineages group of samples

As for the group of samples where two lineages are expected to be found, I defined 5 characteristics depending on what was detected in samples: i) both expected lineages; ii) only one expected lineage; iii) one expected+unexpected lineages; iv) both expected+unexpected lineages; v) nothing. Similarly to the single lineage group of samples analysis, for each of these characteristics, barplots were created using Python separately as well as a combined barplot that shows the number of samples that satisfy each characteristic. Additionally, there was a barplot created showing the percentage of samples that belong to each characteristic.

To zoom in, analysis was done for the most interesting sub-group of the group of samples where two lineages are expected to be detected. In this sub-group, there are two types of samples where both expected lineages are detected, and both expected + unexpected lineages are detected. To compare results, it is reasonable to look at the correlation between specific lineages, such as Delta, BA.1, and BA.2, as in this mock dataset, only three known lineages were used. The first option is to look at the ratio between BA.1 and BA.2 lineages. The second option is the relation between BA.1 and Delta lineages. The third is BA.2 and Delta relation. However, the third

option does not make sense in this mock dataset because there are no samples in the group of two lineages where these two appear together at the same time.

Therefore, in total, 6 cases were considered: i) BA1 > BA2; ii) BA1 = BA2; iii) BA1 < BA2; iv) BA1 > Delta; v) BA1 = Delta; vi) BA1 < Delta. I looked at every case where this relation was expected and compared obtained results from Freyja, COJAC, and Lineagespot regarding the relation between two certain lineages (BA.1/BA.2 relation and BA.1/Delta relation). To visualize results, 6 bar plots were generated and explained in section 4.

To better understand relationships between sets of different methods (Freyja, COJAC, Lineagespot) outputs, Venn Upset diagram was generated. The goal was to look at intersections between sets of results and in which samples both lineages were detected correctly (compared to expected ones) by which tool and to find similarities in tools' performance.

All plots are shown in section 4. All links to Jupyter notebooks with visualizations of results on mock dataset created for this thesis are provided in section 6.5. One note should be said, as of the submission of this thesis, metadata for mock dataset about expected results are not published yet. Therefore, this data was used privately and not shared within this master thesis.

3.2.2 Evaluation on real-world data

3.2.2.1 Chosen real-world datasets

In order to provide a fairly comprehensive analysis, real-world datasets for experiments in this thesis were selected in such a way that they cover a variety of locations in the world and different time points of collecting samples. In this way, it was expected to obtain data on the evolution of SARS-CoV-2 over time, as well as data on which variants prevailed in which countries at these time points. This variety of datasets considered in this thesis can be in the future regularly analyzed using automatized Galaxy bot like it is currently managed for clinical data. That is why the choice of my thesis fell on the four datasets: i) one dataset from California (PRJNA661613)[155] where the samples were collected in 2020 at a wastewater interceptor (labeled Berkeley, Berkeley Hills, Oakland, and Marin, according to the municipal areas each serves); ii) a dataset from the UK (PRJEB42191) [167], with data collected in sewage across six major urban centers in the UK (with a total population equivalent of 3 million) around the same time period (late spring - early summer of 2020) as the previous dataset in order to show different proportions of different variants of the virus; iii) a dataset from wastewater treatment facilities across Ontario, Canada collected by Canadian Research Institute for Food Safety (PRJNA824537), which is interesting to analyze because it contains one of the most recent datasets, the last sample was published in June of 2022; iv) a dataset from the US (PRJNA765346) collected by the FDA Center for Food Safety and Applied Nutrition [147], one of the most extensive dataset with more than 340 samples already and regularly new samples are being added (last samples being from October of 2022). The exact data sources will be listed below.

More metadata about datasets can be found in table 8. Additionally, links to public Galaxy histories where datasets were analyzed with the proposed method are listed in section 6.3.

Location	Accession	Project	Description	Center name	Sequencing platform	Library Strategy	Number of samples	% of SARS-COV-2 on 3 randome samples	Collection period
California, US	SRP280174	PRJNA661613	California sewage metatranscriptomes enriched for respiratory viruses	Jill Banfield's Lab at Berkeley	ILLUMINA	Targeted-Capture	11 (PE)	0.74; 0.64; 0	13.05.2020-30.06.2020
Wales and Northwest England, UK	ERP126028	PRJEB42191	Monitoring SARS-CoV-2 in municipal wastewater to evaluate the success of lockdown measures for controlling COVID-19 in the UK	BANGOR UNIVERSITY	ILLUMINA	Ampliconic	340 (170-SE, 170-PE)	86.77; 95.15; 83.69	30.03.2020-12.05.2020
Ontario, Canada	SRP368140	PRJNA824537	Metagenomics analysis of wastewater influents from wastewater treatment facilities across Ontario	Canadian Research Institute for Food Safety	ILLUMINA	Ampliconic	48 (PE)	14.09	29.11.2021-09.02.2022
Washington, US	SRP354651	PRJNA765346	GenomeTrakr wastewater project: Washington State Department of Health	FDA Center for Food Safety and Applied Nutrition	ILLUMINA	Ampliconic	346 (PE)	23.29; 18.79; 14.52	25.11.2021-19.08.2022

Table 8: Real-world datasets used for this master thesis experiments. Datasets were downloaded from ENA database [154] with filters: i) sars-cov-2 and ii) wastewater. Four datasets, presented in this table, were chosen to be analyzed in this thesis considering a variety of locations and sample collection time.

3.2.2.2 Comparison of proposed workflows

Both, Freyja and COJAC, generate a table-formatted file with results. Freyja, additionally, produces non-interactive and interactive plots. Non-interactive plots include 1) bar plots that show a fractional abundance estimate for all aggregated samples; 2) bar plots with sample collection time information with month binning and daily binning (when collection dates are provided)

Freyja's visualization has limitations. Freyja uses the UShER database to find lineages. Freyja creates plots based on WHO designations, like Omicron, Delta, etc. However, in the Californian (PRJNA661613) dataset, there are no exact WHO names defined in UShER. That means that for aggregation tables and for plotting Freyja uses the label "Other". When detecting lineages that are considered by Freyja as "Other" designated name, the result bar plots are non-informative since they contain only one group of lineage - "Other". Moreover, trying to produce an interactive dashboard, Freyja can meet an issue with "too many lineages to plot". Freyja was able to produce interactive plots only for the UK (PRJEB42191) dataset, while for the rest, 3 out of 4 chosen real-world datasets, California (PRJNA661613), Canada (PRJNA824537), and US (PRJNA765346) ones, the issue with "too many lineages to plot" appeared.

COJAC does not produce plots automatically, so bar plots similar to Freyja bar plots were generated using Python Matplotlib library (version 3.5.1) to make a comparison for this thesis.

4 Results

In the following sections, the results of downstream data analysis will be shown. In section 4.1, results and analysis on a mock synthetic dataset will be presented, followed by section 4.2, where results obtained by running workflows on real datasets will be shown and described.

4.1 Mock dataset results

Mock dataset has been synthetically generated by a working group and is available on GitHub. Synthetic dataset of overall 100 samples was created by obtaining genomes from the Pango designation list using three real variants (BA.1, BA.2, Delta), one synthetic background variant (BG), and recombinant genomes (Omicron-Delta). Synthetic amplicons were created using ARTIC v4.1 primers. Short 150 bp and long 250 bp reads were used while generating. Briefly, the mock dataset can be described by belonging samples to groups of different classes such as: i) Single lineage, Two lineages, Three lineages; ii) High coverage, Low coverage; iii) Long reads, Short reads. Details on the number of mock samples of each group are in table 7.

From 100 mock samples: 22 contain single lineage, 42 - two distant lineages (BA.1 and Delta), 6 - two close lineages (BA.1 and BA.2), 14 samples - 2 known and one unknown lineage (synthetically generated), 8 samples - known lineages, and 8 samples - 3 known and one unknown lineages. On the other hand, the mock dataset contains 50 samples of low coverage and 50 samples of low coverage. The mock samples consist of 24 long reads (250 base pairs) and 76 short reads (150 base pairs).

For the evaluation of workflows developed for ampliconic data, the workflow's Freyja-based and COJAC-based branches as well as the standalone Lineagespot workflow were run on a mock dataset.

4.1.1 Benchmarking Galaxy and Lineagespot workflow results

Galaxy workflows results were compared with Lineagespot results using the same synthetic mock dataset. As expected, Galaxy workflow results in both branches differ from Lineagespot results. That is because different approaches (sub-workflows) are used for data preprocessing steps which were described in paragraph 3.2.1.3 of section 3. In evaluating workflows on the mock dataset, two groups of samples were selected: one with a single lineage and one with two lineages.

4.1.1.1 Single lineage group of samples

The Single lineage group consists of 22 samples. Four characteristics were used for a single lineage group. The goal was to discover in which samples were detected: i) expected lineage; ii) expected lineage plus unexpected lineages; iii) unexpected lineage; iv) nothing. Each of these characteristics is represented by a combined bar plot that shows how many samples meet that characteristic.

From fig. 16 it is obvious that all three tools are effective in detecting expected lineage. Nonetheless, in discerning only expected lineage and nothing more, Lineagespot performed the best,

compared to COJAC and Freyja. Freyja is effective at detecting expected lineage; however, it always detected some unexpected lineages. COJAC's results are close to Freyja's results, but COJAC was able to detect 2 samples with the expected lineage. Another interesting observation is that in 4 samples, COJAC is rather to detect nothing than expected lineage. Figure 16 shows

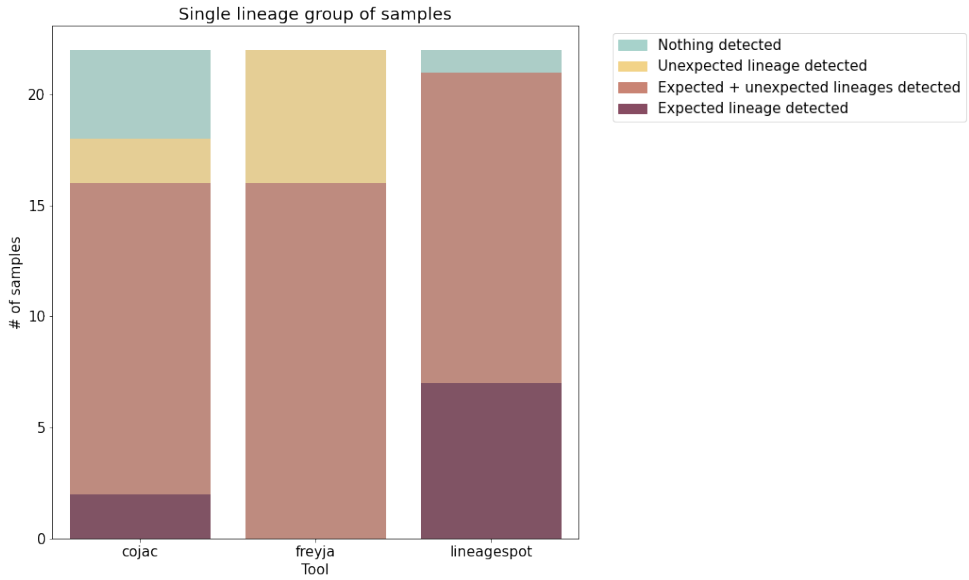


Figure 16: Bar plot describing the number of mock samples of single lineage group that meet one of four characteristics: i) expected lineage detected; ii) expected lineage plus unexpected lineages detected; iii) unexpected lineage detected; iv) nothing detected

that Freyja is an effective tool for detecting expected lineage, even though Freyja always detects other lineages. When it is expected the sample to contain only one lineage but nothing more, Freyja is not that effective. Moreover, in 6 samples for Freyja and in 2 samples for COJAC (out of 22), there were detected unexpected, i.e. incorrect, lineages. On the other hand, Lineagespot is effective in the case of detecting only expected lineages, probably because it has the additional step in its pipeline when the most probable lineages are assigned for the sample. This extra step is made based on several indicators. So Lineagespot detected only expected lineage in 7 samples out of 22 samples from a single lineage group. As for COJAC, it performed quite well and was able to detect only one lineage that was expected in 2 out of 22 samples, however, for 4 samples with expected lineage, COJAC detected nothing.

In order to better understand the interrelationships between different results (Freyja, COJAC, Lineagespot), a Venn Upset diagram was generated. Using this method, I analyzed intersections between sets of results and determined which samples were correctly detected by which tool (in terms of the lineages expected) and how similar the results were between tools.

Venn Upset diagram below (fig. 17) was constructed based on 22 samples in which there was a single lineage expected. Each column corresponds to a set of obtained results from certain tools (COJAC, Lineagespot, Freyja), and bar charts on top show the size of the set of tool's results. The first row in the figure is completely empty, while 1 sample is expected to be detected but was not. This sample75 is expected to contain only unknown synthetic lineage.

The second row corresponds to 5 samples with the expected single lineage only detected by Lineagespot but not by COJAC and Freyja. Finally, the last third row represents 16 samples with the expected single lineage detected by all tools. Freyja and COJAC detected the expected lineage only in 16 samples which are confirmed in the bar plot above (fig. 16).

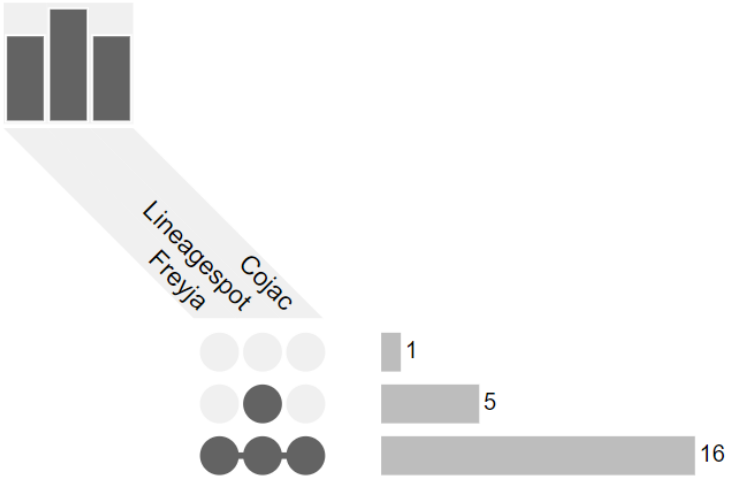


Figure 17: Venn Upset diagram constructed based on 22 samples in which there was single lineage expected.

Moreover, using Python with the Seaborn library, a distribution of the proportion of lineage detected by Freyja and COJAC among samples in the Single lineage group was plotted (fig. 18) by only considering samples where expected lineage was detected. The comparison was limited to Freyja and COJAC. In Lineagespot, I do not have a clean proportion distribution between lineages, so I do not consider it here. One of the reasons for excluding Lineagespot from this distribution analysis: besides BA.1, BA.2, Delta, no information was provided about other lineages that have been detected in Lineagespot, and the sum of proportions of these three does not equal 1. Looking at fig. 18, I conclude that for single lineage detection, the results of lineage proportion from COJAC and Freyja are from 0.9 to 1. However, some differences between Freyja and COJAC results are observed. Freyja showed a lower proportion of expected lineage, while for COJAC the proportion tends to 1 which is expected. Thus, we can guess that COJAC results for the single lineage group are closer to what was expected.

For comparing Freyja and COJAC results with each other and with expected results, parallel coordinates plots were created using Python and Plotly graphing libraries. Samples are plotted as parallel coordinates across different measures using this visualization technique. Samples are represented as connected points along each vertical axis, and each measure corresponds to a vertical axis.

Three plots, focused on the group of samples marked as "single lineage", were generated for Delta (fig. 19), BA.1 (fig. 20), and BA.2 (fig. 21) lineages. Graphs represent different lineages (one lineage per graph). Within the graph, the left axis represents the expected proportion, the middle axis represents proportion of Delta lineage detected by COJAC, while right axis represents the proportion of Delta lineage detected by Freyja.

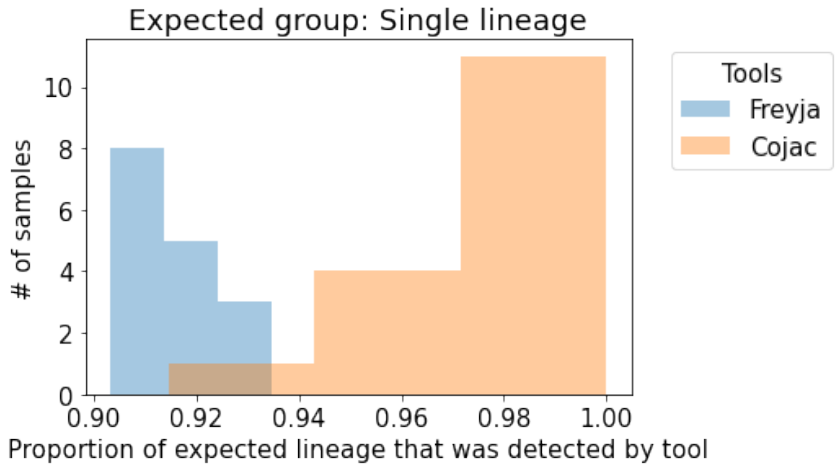


Figure 18: Distplot that represents the distribution of lineage proportion detected by Freyja and COJAC among samples of the Single lineage group, considering only samples in which expected lineage was detected.

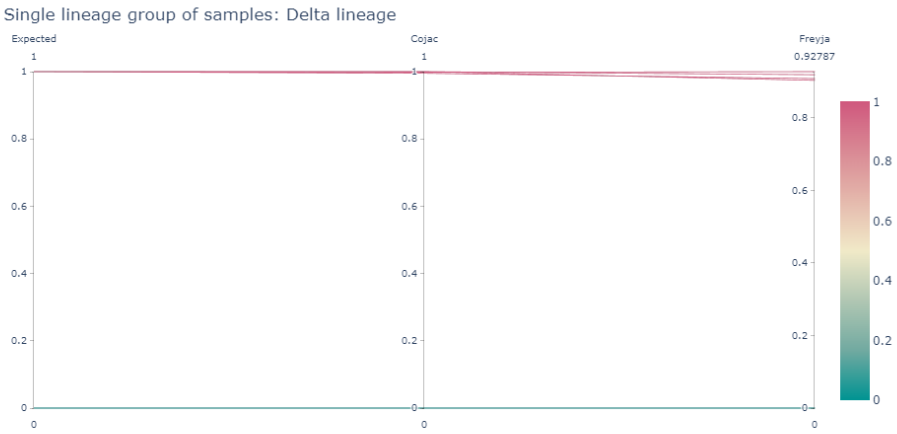


Figure 19: Parallel coordinates plot for a single lineage group of samples that compare Delta lineage proportions detected by Freyja and COJAC with each other as well as with expected proportion. The left axis represents the expected proportion of the lineage, the middle axis represents the proportion of the lineage detected by COJAC, while the right axis represents the proportion of the lineage detected by Freyja.

From fig. 19 and fig. 21, it is observed that for single lineage group of samples, for detecting Delta and BA.2, COJAC performs well, the value of proportion is very close to expected 1, while Freyja performs slightly below expected. However, both results are close to expected for all samples of this group (Single lineage group).

Quite a bit different results are shown in fig. 20, for BA.1 lineage. More specifically, Cojac's results for a few samples are closer to Freyja's results, with a value of 0.9, whereas a value of 1 is expected. In another curious case, for sample83 and sample52 BA.1 lineage was not expected.

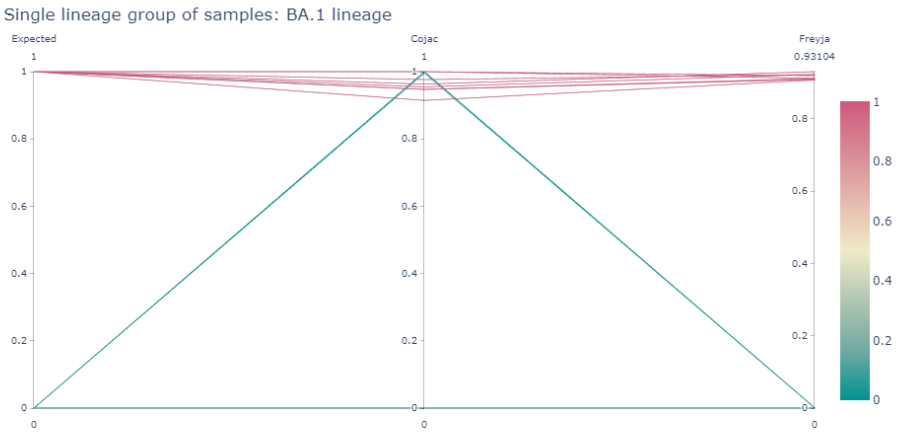


Figure 20: Parallel coordinates plot for a single lineage group of samples that compare BA.1 lineage proportions detected by Freyja and COJAC with each other as well as with expected proportion. The left axis represents the expected proportion of the lineage, the middle axis represents the proportion of the lineage detected by COJAC, while the right axis represents the proportion of the lineage detected by Freyja.

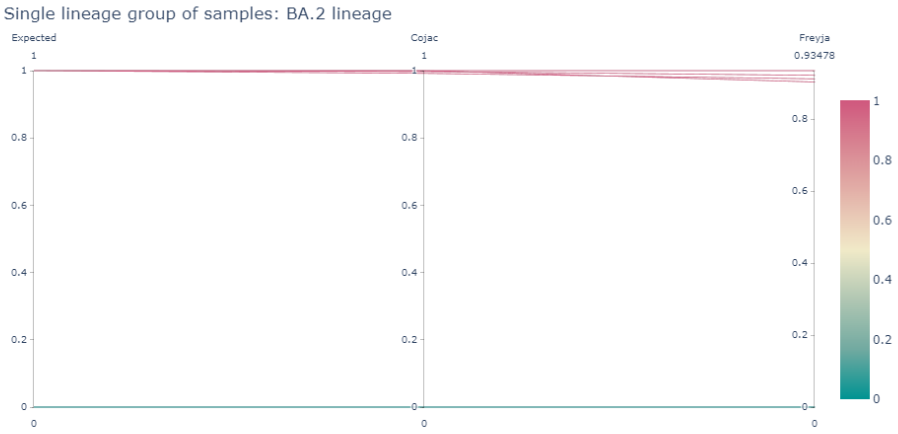


Figure 21: Parallel coordinates plot for a single lineage group of samples that compare BA.2 lineage proportions detected by Freyja and COJAC with each other as well as with expected proportion. The left axis represents the expected proportion of the lineage, the middle axis represents the proportion of the lineage detected by COJAC, while the right axis represents the proportion of the lineage detected by Freyja.

Freyja did not detect this lineage. COJAC, in turn, detected 0.9962962963 proportion of BA.1 lineage for sample83, and 1 for sample52, which is high. The explanation for this is that in the mock dataset, these two samples contain only recombinant (BA.1+BA.2+Delta). It is also possible for viruses to generate diversity through recombination, which occurs when two different

lineages of SARS-CoV-2 infect the same cell. As a result, two different genomes can swap out sections, which is distinct from mutations caused by errors. In this case, the recombinant virus genome occurs. For these 2 samples, it was expected to detect the recombination of three (BA.1, BA.2, and Delta) lineages, but not each of these lineages separately. This recombinant contains BA.1 lineage mutations, and that's why COJAC misinterpreted results for these samples.

4.1.1.2 Two lineages group of samples

As for the group of samples where two lineages are expected to be found, I defined five characteristics based on what was detected in samples: i) both expected lineages; ii) only one expected lineage; iii) one expected plus unexpected lineages; iv) both expected plus unexpected lineages; v) nothing. As with the single lineage group analysis, a combined bar plot, which shows the number of samples that satisfy each characteristic, was generated using Python (fig. 22).

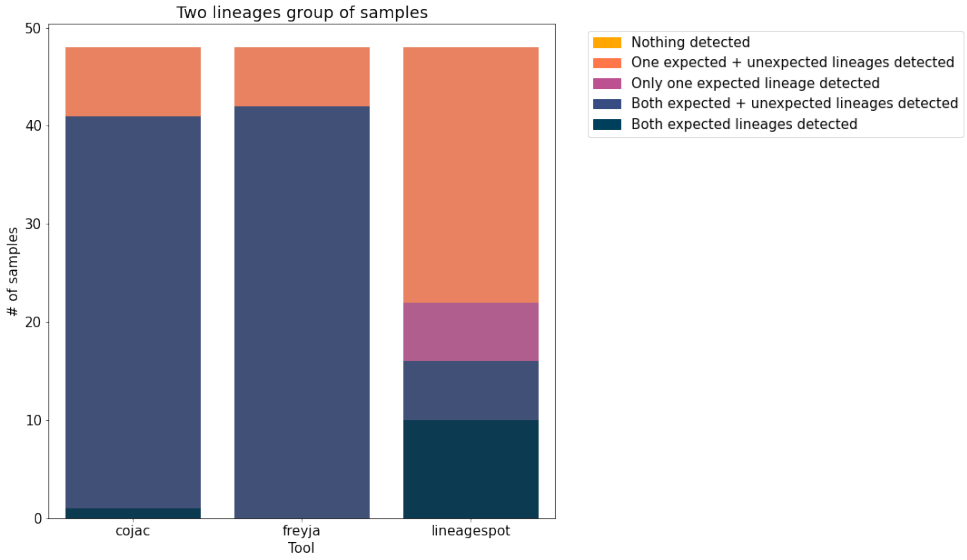


Figure 22: Bar plot describing the number of mock samples of two lineages group that meet one of five characteristics: i) both expected lineages detected; ii) only one expected lineage detected; iii) one expected plus unexpected lineages detected; iv) both expected plus unexpected lineages detected; v) nothing detected.

According to fig. 22, overall, 48 samples are in Two lineages group. From the bar plot, it is obvious that all three tools are effective in detecting both expected lineages. Nonetheless, in detecting only two expected lineages and nothing more, Lineagespot performed the best, compared to COJAC and Freyja, as it is the only tool that was able to detect 10 samples out of 48 with two expected lineages and nothing more. As was already shown for the Single lineage group, Freyja is effective at detecting expected lineages (in 42 samples out of 48); however, it always detected some unexpected lineages. COJAC's results for two lineages are considerably close to Freyja's results; they both detected around 40 samples with two lineages expected. Even more, in one sample, COJAC was able to detect only expected lineages. Interestingly, no tools detected nothing for this group of samples.

Venn Upset diagram below (fig. 23) was constructed based on 48 samples in which there were two lineages expected. Each column corresponds to a set of obtained results of certain tools (COJAC, Lineagespot, Freyja), and bar charts on top show the size of the set of tool's results. The first row in the figure is completely empty, while 3 samples are expected to be detected but were not. These three samples are distinct from other samples by belonging to the "low coverage" group. In addition, there are a few samples in which 2 expected lineages were detected by only one tool: 1,2, and 3 samples corresponding to Lineagespot, Freyja, and COJAC. At the same time, Freyja and COJAC detected two expected lineages in 24 same samples, whereas Lineagespot did not detect both expected lineages, so there is no interaction with Lineagespot set here, but there is one between Freyja and COJAC. That can be explained by the same workflow used for Freyja and COJAC results but the different method used by Lineagespot pipeline. Interestingly, all three tools Lineagespot, Freyja, and COJAC detected the expected two lineages in 15 same samples.

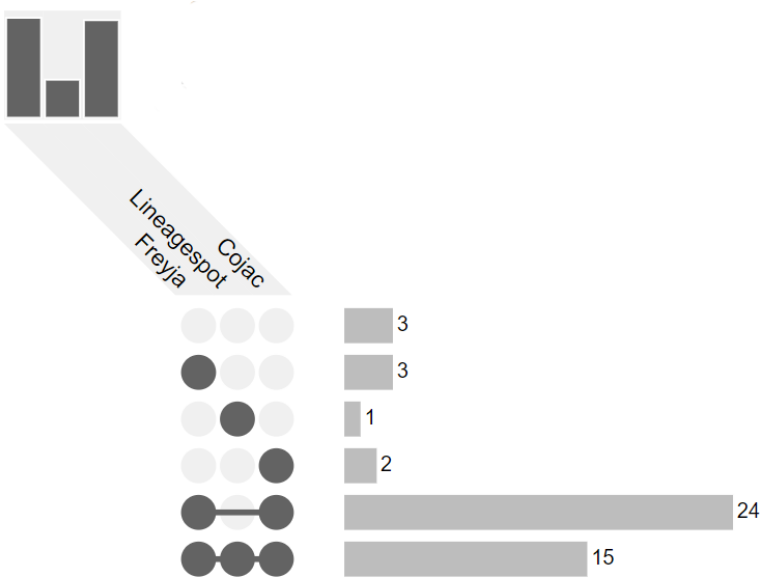
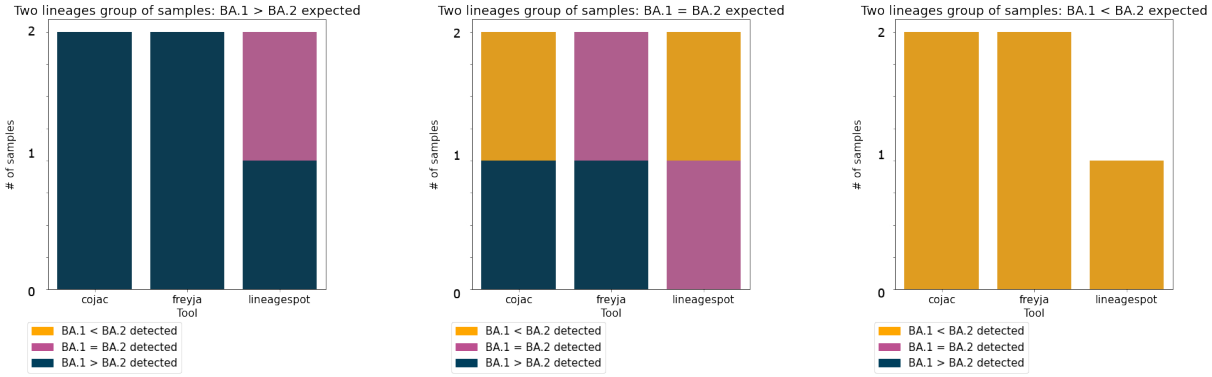
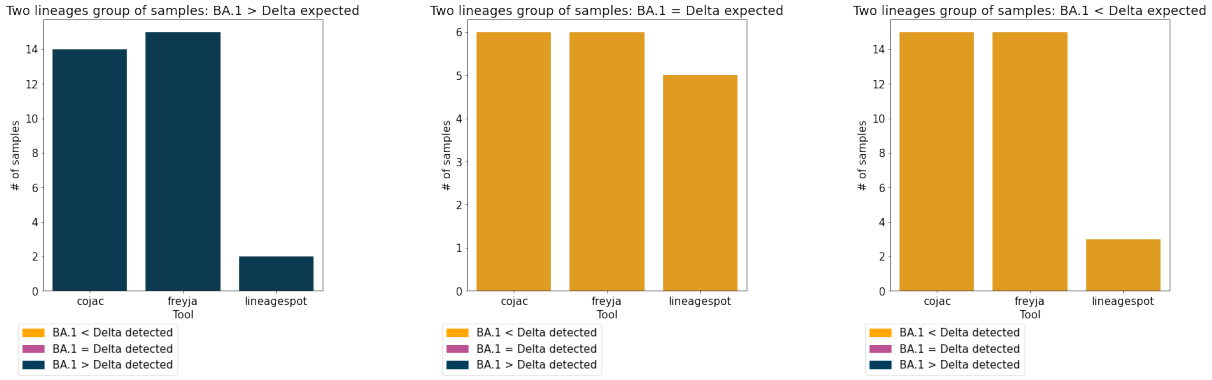


Figure 23: Venn Upset diagram constructed based on 48 samples of the synthetic mock dataset in which there were two lineages expected.

In cases: i) both expected lineages detected and ii) both expected + unexpected lineages detected, there were additionally 6 bar plots generated to evaluate if expected proportion between two expected lineages is correctly defined by different tools. Proportions between expected lineages that were considered are: i) BA1 > BA2; ii) BA1 = BA2; iii) BA1 < BA2; iv) BA1 > Delta; v) BA1 = Delta; vi) BA1 < Delta. These plots show every case where this proportion was expected (one plot per 1 of 6 cases), and the corresponding results from Freyja, COJAC, and Lineagespot are plotted in fig. 24a and fig. 24b.



(a) Bar plots for "two lineages" group of mock samples where following proportions were expected: i) BA1 > BA2; ii) BA1 = BA2; iii) BA1 < BA2.



(b) Bar plots for "two lineages" group of mock samples where following proportions were expected: iv) BA1 > Delta; v) BA1 = Delta; vi) BA1 < Delta.

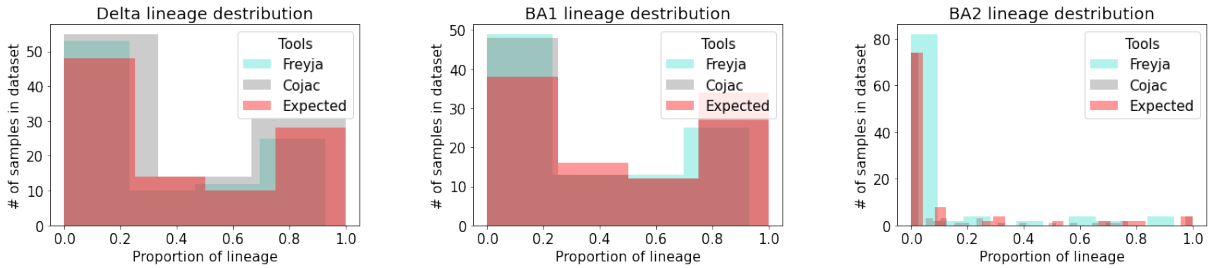
For the “two lineages” expected group of samples, there are only 6 samples where BA.1 and BA.2 lineages were expected together fig. 24a): 2 for every type of proportion. For the case, when the proportion of BA.1 > BA.2 is expected, both COJAC and Freyja detected the same proportion between these two lineages, while Lineagespot only in 1 sample. In case of equal presence of BA.1 and BA.2 in samples, Freyja and Lineagespot detected 1 sample against 2 expected. In the case of BA.1 < BA.2, all tools detected samples correctly, however, Lineagesspot detected only 1 out of 2.

For the “two lineages” expected group of samples, there are 42 samples where BA.1 and Delta lineages were expected together (fig. 24b): 18 for BA.1 > Delta type of relation between lineage proportions, 6 for BA.1 = Delta, and 18 samples for BA.1 < Delta. For the case when the proportion of BA.1 > Delta is expected, all tools detected it correctly but not in all 18 samples. However, Freyja and COJAC performed well, being able to detect 15 and 14 samples with this type of relation, respectively. Lineagespot detected only 2 samples. Likewise, for the case of BA.1 < Delta all tools detected some samples but not all expected samples, Freyja and COJAC both detected 15 out of 18 which is significantly more than Lineagespot with its 3. Finally, when BA.1 = Delta was the expected relation, all tools discerned BA.1 < Delta relation.

4.1.2 Benchmarking Freyja- and COJAC- based workflow results

4.1.2.1 Distribution of linages proportions among tools

To compare Freyja and COJAC branches of workflow in another way, and to assess results with expected ones, in fig. 25a, 25c, and 25c, taking all 100 samples into account, the distributions of every lineage (Delta, BA.1, BA.2) are shown among received results from Freyja, COJAC, and expected proportions. It is seen that for Delta and BA.1 lineages of SARS-CoV-2 there is quite a similar number of samples carrying the same proportion of the considered lineage. Interestingly, for all lineages smaller number of samples with a low proportion of lineage was expected than detected by tools Freyja and COJAC. For BA.1 lineage there were expected more samples with higher lineage proportion than was found by Freyja and COJAC.

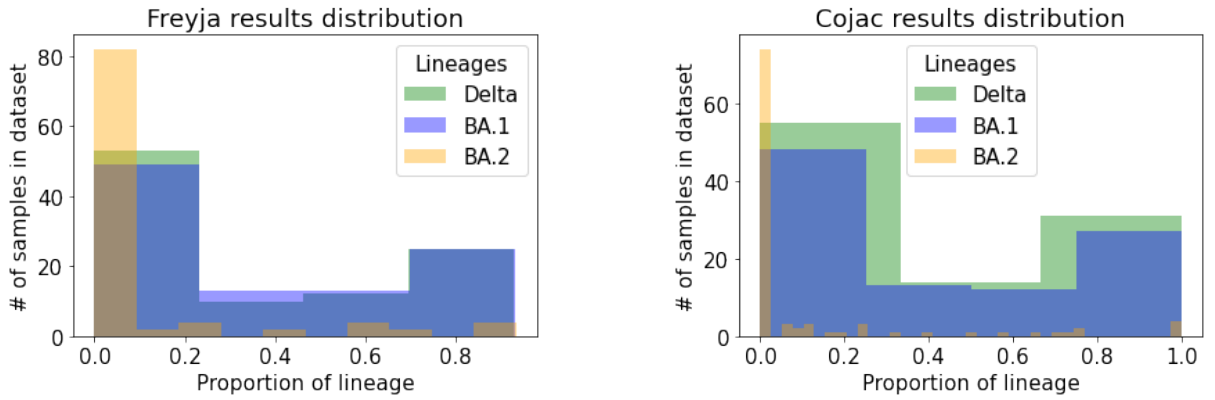


(a) Delta (b) BA.1 (c) BA.2

Figure 25: The lineage proportion distribution among results from Freyja, COJAC on mock dataset (all samples), compared with expected proportions.

4.1.2.2 Distribution of tools results among all mock samples

The other angle to observe results and compare them, is to look at how lineages are distributed in results produced by every tool. In fig. 26a, one can see that Freyja detected BA.2 with a small presence in most of the samples (around 80 out of 100), while Delta and BA.1 were found by Freyja in more or less similar proportions. Looking at fig. 26b, one can see just a bit different distribution when most of the samples contained a very small proportion of BA.2 according to COJAC results. In both graphs, Delta and BA.1 lineages were found in small proportion, before 0.3, in around half of samples, while around 20-30% of samples carry Delta and BA.1 in high presence, from 0.7 to 1 value of the proportion of lineage abundance.



(a) Freyja

(b) COJAC

Figure 26: All three considered lineages (Delta, BA.1, BA.2) proportion distribution among results from the tool on mock dataset.

4.2 Real-world datasets results

For this thesis, the real-world data for experiments were chosen based on the principle that samples were collected over a variety of time periods and locations around the world. In chosen four datasets, I aimed to cover these principles expecting to observe evolution of SARS-CoV-2 over time as well as to show variations in prevalence of a variety of SARS-CoV-2 lineages depending on the location.

Metadata on real-world datasets of choice for this master thesis is shown in table 8. In the following sections, the results of running developed Galaxy workflows will be shown. Section 4.2.1 will represent results of Freyja-based Galaxy workflow on the Californian dataset, section 4.2.2 - Freyja-based and COJAC-based workflows results on the Canadian dataset, section 4.2.3 - Freyja-based and COJAC-based workflows results on US dataset, and, finally, section 4.2.4 - Freyja-based and COJAC-based workflows results on UK dataset.

4.2.1 Dataset (PRJNA661613): California sewage metatranscriptomes enriched for respiratory viruses

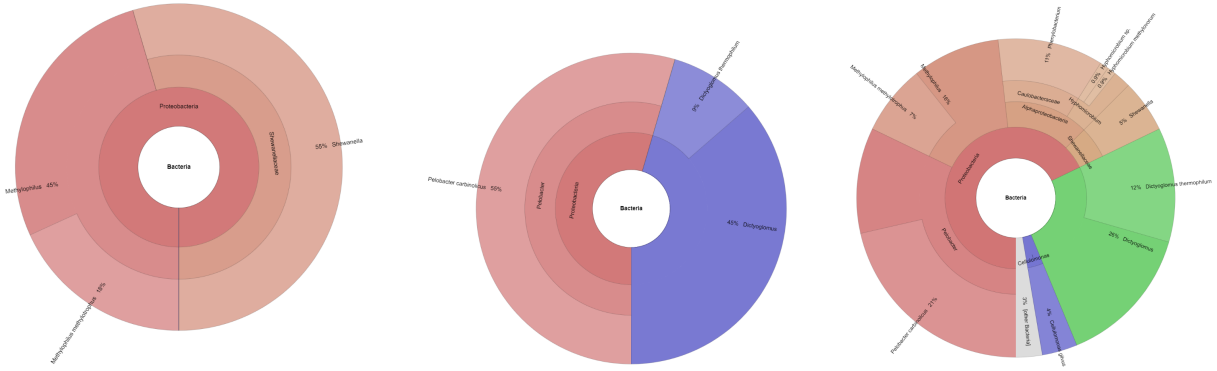
For this dataset, samples of raw sewage were collected from wastewater treatment facilities in Alameda and Marin Counties in Northern California between 13 May 2020 and 30 July 2020 at a wastewater interceptor (labeled Berkeley, Berkeley Hills, Oakland, and Marin, according to the municipal areas each serves).

For this dataset, only the Freyja-based branch of Galaxy workflow was launched because the COJAC-based branch is applicable only to the ampliconic library preparation dataset, while Freyja can work with metatranscriptomics which is the case of the Californian dataset.

In the step of checking the abundance of other species that are unmapped to reference SARS-CoV-2 sequence on the Californian dataset, Kraken2 was run on a dataset collection using viral

genomes taxa database, following Krona pie charts generation. Going across samples', approximately 78% to around 85% of fragments are covered by the clade rooted at unclassified taxon. In a similar manner, from sample to sample percentage, from around 18% to 22%, of fragments are covered by the clade rooted at either viruses, ribovirus, nidovirales, cornidovirineae, or coronaviridae taxon. Slightly smaller proportions of fragments are covered by the clade rooted at Orthocoronavirinae, Betacoronavirus, Sarbecovirus, Severe acute respiratory syndrome-related coronavirus, and SARS coronavirus, around 12-15%. Notable, the number of fragments assigned directly to the SARS coronavirus taxon is non-zero, as it is for most of the taxons.

Krona charts were produced for every sample separately, followed by a Krona pie that represents all samples in one pie chart. Figure 27a, 27b show resulting examples of Krona charts for two separate samples (SRR12596166 and SRR12596175, respectively), fig. 27c illustrates Krona pie for aggregated all samples of dataset.



(a) Krona for sample SRR12596166

(b) Krona for sample SRR12596175

(c) Krona for all aggregated samples

Figure 27: Krona chart visualizing the abundances of species (assigned with viral taxa database) in wastewater samples from the Californian dataset (PRJNA661613).

As for lineages abundances analysis on the Californian dataset, there were no exact names, like Omicron, Delta, etc., defined by Freyja using UShER bar-codes. So, for aggregation tables and for plotting Freyja used the label "Other" (see table 9 in section 6.9.1). When detecting lineages that are considered by Freyja as "Other" designated name, the result bar plots are non-informative since they contain only one group of lineage - "Other". Moreover, trying to produce an interactive dashboard, Freyja can meet an issue with "too many lineages to plot". Due to Freyja's limitations, for the Californian dataset, plots were either non-informative like in fig. 28 or not generated.

4.2.2 Dataset (PRJNA824537): Wastewater influents from wastewater treatment facilities across Ontario, Canada

Wastewater samples were collected by Canadian Research Institute for Food Safety. The dataset contains one of the relatively recent samples, the last sample was published in June of 2022.

Similarly to the Californian dataset analysis, Krona charts were created for each sample of the Canadian dataset separately, followed by a Krona pie representing all samples. Figure 29a,

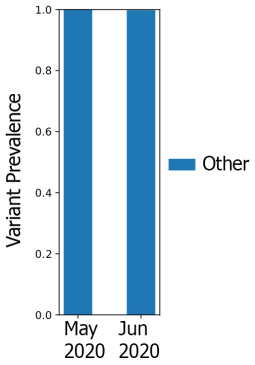
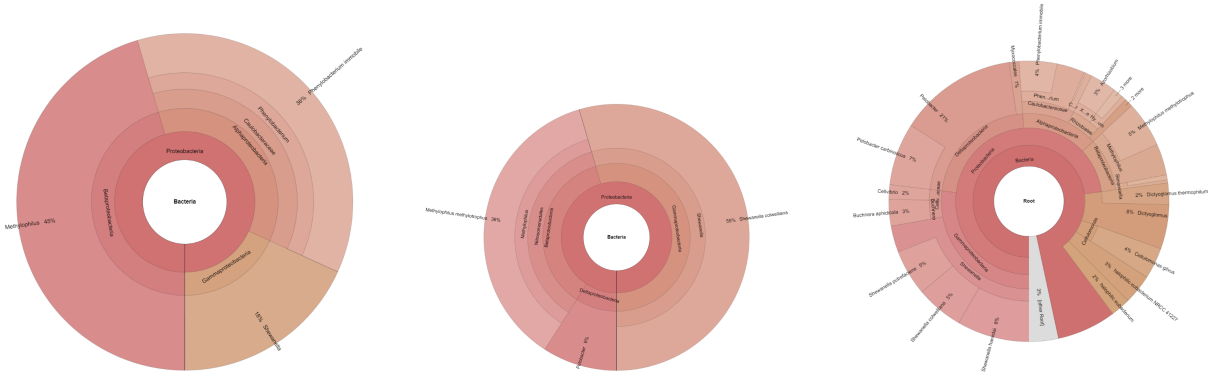


Figure 28: Lineages abundances detected by Freyja for Californian dataset, grouped by months. All lineages were labeled as "Other" which resulting to non-informative plot

29b show resulting examples of Krona charts for two separate samples (SRR18680446 and SRR18680489, respectively), fig. 29c illustrates Krona pie for aggregated all samples of dataset.



(a) Krona for sample SRR18680446

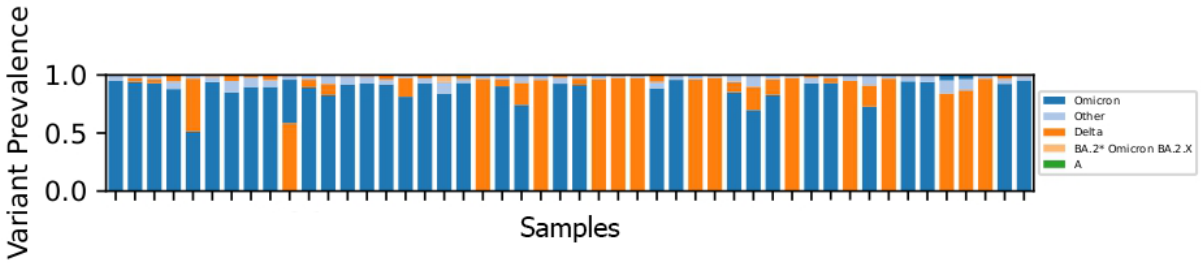
(b) Krona for sample SRR18680489

(c) Krona for all aggregated samples

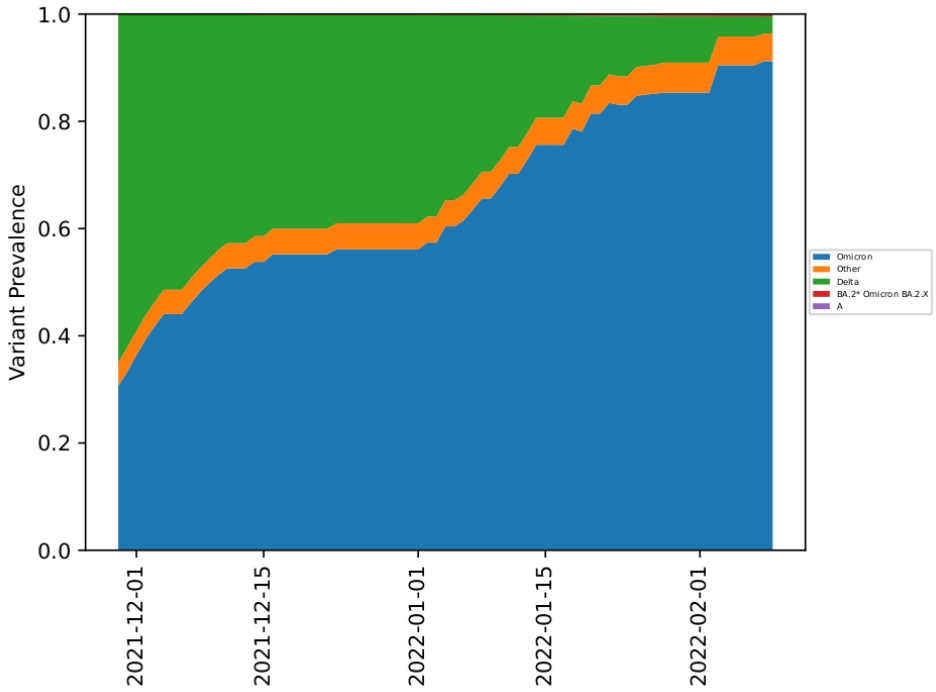
Figure 29: Krona chart visualizing the abundances of species (assigned with viral taxa database) in wastewater samples from the Canadian dataset (PRJNA824537).

Regarding results on SARS-CoV-2 lineages abundances, for the dataset from Ontario, Canada (PRJNA824537), there were readable plots generated by Freyja (fig. 30a,30b). In fig. 30a, one can see all samples represented by bars (one bar per sample). However, they are not in order of collecting date and time, which makes it difficult to interpret. One can see that some of the samples contain abundant Omicron lineage, and some other samples have abundant Delta. Nevertheless, it is the general picture of lineages abundance among samples. Figure 30b, in contrast, represents trends and changes in lineage abundances over time. Delta was a prevalent lineage at the beginning of December 2021, while from the mid of December 2021, Omicron began to be most predominant, and its abundance increased significantly up to around 0.9 lineage abundance proportion in February 2022. It's worth noting that Freyja always detected other lineages in samples, except Omicron and Delta, although in small proportions, around

0.05.



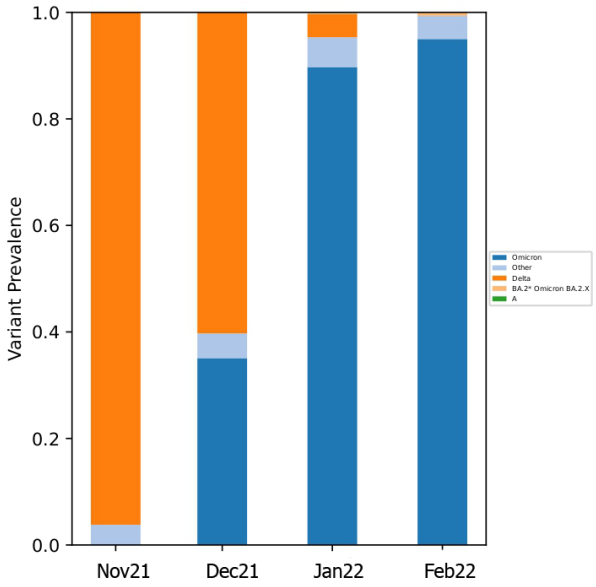
(a) Variant prevalence across samples.



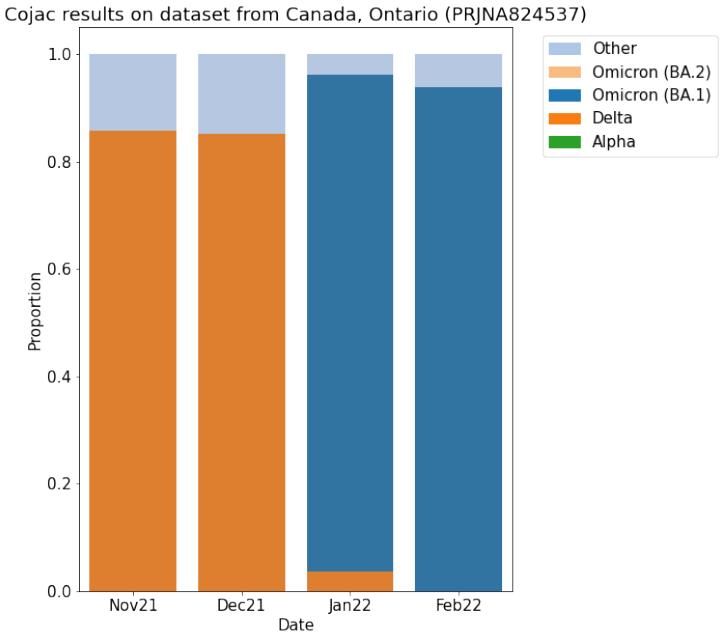
(b) Variant prevalence with daily binning.

Figure 30: Variant prevalence computed by Freyja over samples, produced by Freyja on Ontario, Canada (PRJNA824537).

To compare results on the Canadian (PRJNA824537) dataset from both Freyja and COJAC branches of Galaxy workflow, two bar plots were considered: one generated by Freyja (fig. 31a), one generated from COJAC results using Matplotlib Python library with the same color scheme (fig. 31b). Links to Jupyter notebooks with visualizations of results on real datasets created for this thesis are provided in section 6.7.



(a) Bar plot generated by Freyja.



(b) Bar plot generated from COJAC output using Python.

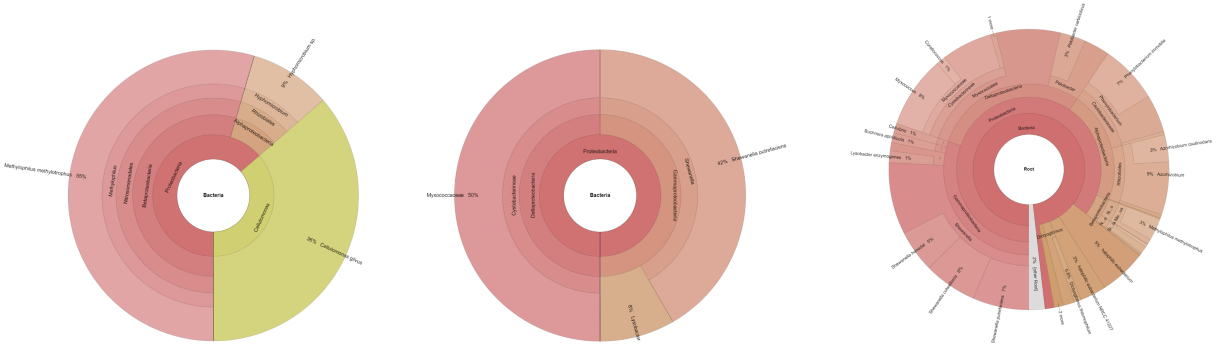
Figure 31: Bar plot with sample collection time information grouped by month, for the Canadian (PRJNA824537) dataset.

Results on the Canadian (PRJNA824537) dataset from Freyja and COJAC differ, as shown in fig. 31a, 31b. However, the prevalence of Delta in November and December 2021 and the prevalence of Omicron in January and February 2022 are discerned by both methods. Another interesting observation is that Freyja found the Omicron variant already in December 2021, while COJAC was able to detect Omicron only in January 2022.

4.2.3 Dataset (PRJNA765346): GenomeTrakr wastewater project of Washington State Department of Health, US

Samples for this dataset were and are being collected by the FDA Center for Food Safety and Applied Nutrition. This dataset is one of the most extensive dataset with more than 400 samples so far, and regularly new samples are being added (the last samples being from October of 2022). This dataset is of research interest also because it can be considered in the future as one of those that will be regularly analyzed by the Galaxy bot that was described in section 1.2.3.

Similarly to the Californian and Canadian datasets, Krona charts were created for each sample of the US dataset separately, followed by a Krona pie representing all samples. Figure 32a, 32b show resulting examples of Krona charts for two separate samples (SRR18680446 and SRR20997582, respectively), fig. 32c illustrates Krona pie for aggregated all samples of dataset.



(a) Krona for sample SRR17578350

(b) Krona for sample SRR20997582

(c) Krona for all aggregated samples

Figure 32: Krona chart visualizing the abundances of species (assigned with viral taxa database) in wastewater samples from the US dataset (PRJNA765346).

As for results on lineage abundances received for the US (PRJNA765346) dataset, the distribution of SARS-CoV-2 lineages over time binned by the day, obtained in Freyja-based branch of the workflow, is depicted in fig. 33.

Figure 33 shows curious results. Delta was found by Freyja in samples of the US dataset only in November 2021. From January to March 2022, Omicron (BA.1 sub-lineage) was prevalent, accompanied by other lineages found in samples but grouped into “Other” lineages. From April to June 2022, the BA.2 sub-lineage of Omicron was common in the US dataset. Finally, by September 2022, Omicron BA.1 took the prevalence back, with about 0.6 proportion against 0.4 for other lineages.

To compare results on the US (PRJNA765346) dataset from both Freyja and COJAC branches of Galaxy workflow, there were two bar plots created: one was generated by Freyja (fig. 34a), while the other was generated from COJAC results using Matplotlib Python library with the same color scheme (fig. 34b)). Links to Jupyter notebooks with visualizations of results on real datasets created for this thesis are provided in section 6.7.

Results from Freyja and COJAC on the US dataset are obviously not the same. The curious fact is detecting the considerable prevalence of the Delta lineage in summer 2022 by COJAC

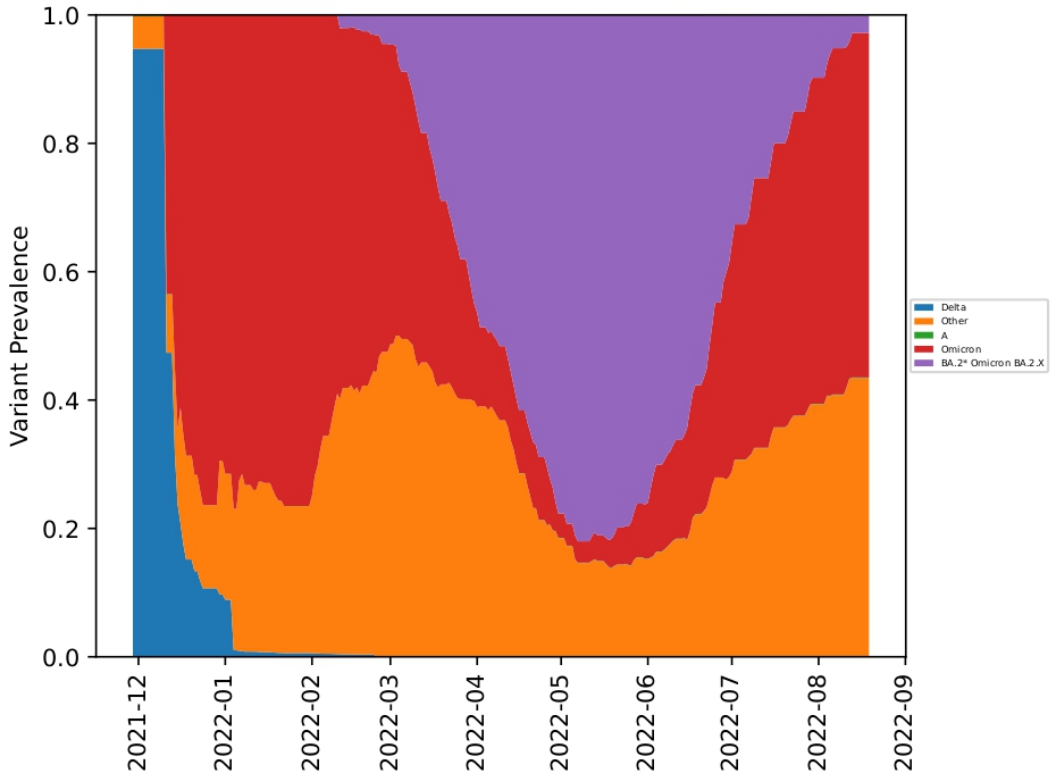


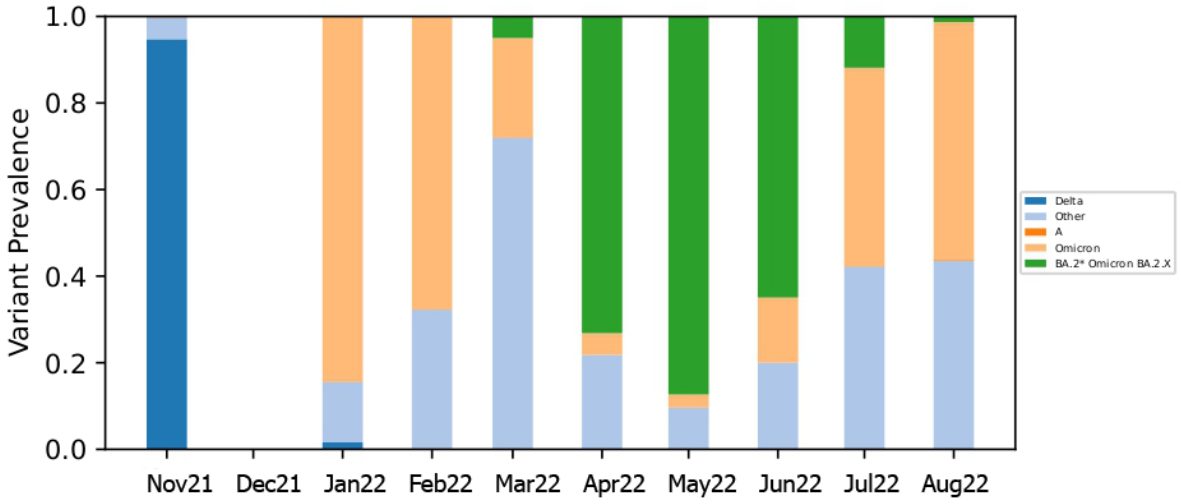
Figure 33: Smooth plot with sample collection time information with daily binning for US (PRJNA765346) dataset.

and not detecting it by Freyja. And more expected that both tools found Delta at the end of 2021 and the prevalence of Omicron (both BA.1 and BA.2 sub-lineages) in 2022. Interestingly, the BA.1 lineage was prevalent from January to March 2022, while BA.2 from April to June 2022.

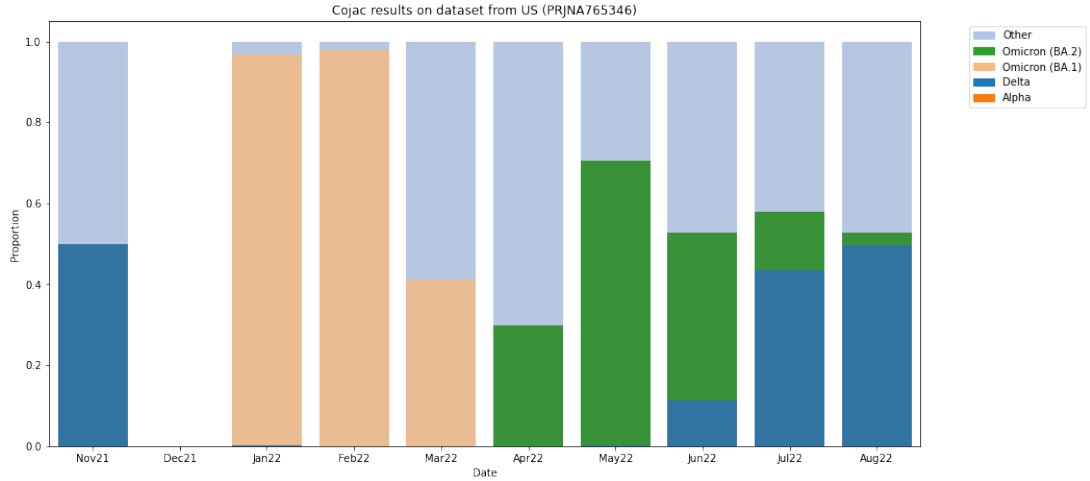
4.2.4 Dataset (PRJEB42191): Monitoring SARS-CoV-2 in municipal wastewater in the UK

Samples for this dataset were collected from six wastewater treatment plants located in Wales and Northwest England. The treatment plants served urban areas in Gwynedd, Cardiff, Liverpool, Manchester, the Wirral and Wrexham, with a combined population equivalent to 3 million people. Weekly samples of untreated wastewater influent were taken from the six treatment plants between March and June 2020.

Similarly to the Californian, Canadian, and US datasets, Krona charts were created for each sample of the UK dataset separately, followed by a Krona pie representing all samples. Figure 35a, 35b show resulting examples of Krona charts for two separate samples (ERR5014633 and ERR5014683, respectively), fig. 29c illustrates Krona pie for aggregated all samples of dataset.



(a) Bar plot generated by Freyja.



(b) Bar plot generated from COJAC output using Python.

Figure 34: Bar plot with sample collection time information grouped by month, for US (PRJNA765346) dataset.

(a) Krona for sample ERR5014633

(b) Krona for sample ERR5014683

(c) Krona for all aggregated samples

Figure 35: Krona chart visualizing the abundances of species (assigned with viral taxa database) in wastewater samples from the UK dataset (PRJEB42191).

The same as to the Californian dataset, for the UK dataset, Freyja labeled all detected variants as “Other”; hence, only non-informative non-interactive plots were generated. Even though Freyja has a limitation while producing an interactive dashboard for too many lineages, for the UK dataset (PRJEB42191), the interactive dashboard was generated (Fig.10). According to the graph, eight lineages were clearly detected by Freyja: i) B.27 that is described in Pango database as “UK lineage, Curacao, Switzerland”; ii) B.23 that was discovered in UK and Portugal; iii) B.12 (Japanese lineage), iv) B.1.14 (USA lineage, California); v) B, one of the two original haplotypes of the pandemic (and first to be discovered); vi) B.1.1.161 (Saudi lineage); vii) B.1.1.514 (US lineage); viii) B.1.1.301 (UK lineage).

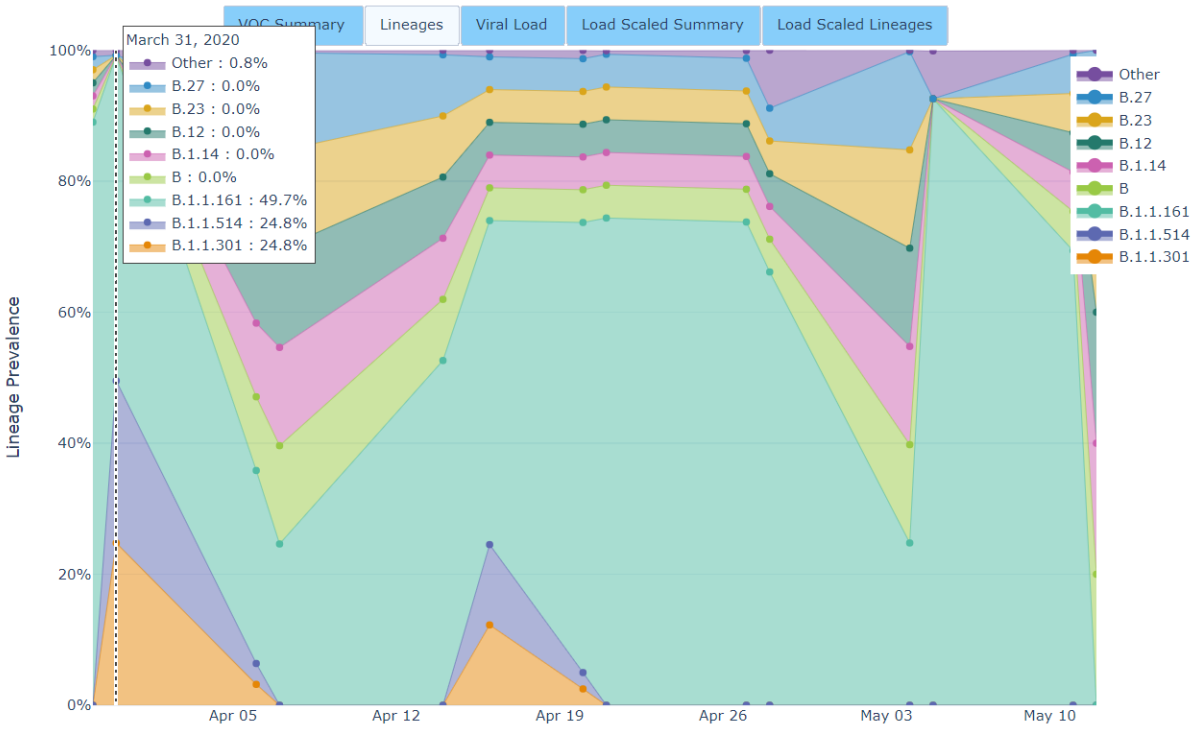


Figure 36: Interactive dashboard for UK (PRJEB42191) dataset that was generated by Freyja tool as a part of Freyja-based Galaxy workflow.

Looking at fig. 36, one can see the prevalence of different lineages over the sample collection time period, from 30 March to 12 May 2020. Detection of B.1.1.514 and B.1.1.301 lineages already at the end of March 2020 (more precisely, in samples collected on 31 March, as well as on 6, 16, and 20 April 2020) is worth mentioning because in Pango database [164] the earliest dates of registration for these lineages are 1 and 22 May 2020, respectively. This fact proves earlier detection within wastewater surveillance over clinical surveillance.

5 Discussion and Outlook

In this thesis, two methods for SARS-CoV-2 wastewater surveillance have been proposed as two workflows. One workflow is assumed to be used for standalone data analysis of wastewater samples extracted using a metatranscriptomics-based library preparation technique and sequenced using Illumina sequencing approach. This workflow has one branch with Freyja tool on the step of SARS-CoV-2 lineages abundances computation. At the same time, the second workflow is developed for standalone data analysis of wastewater samples obtained with an ampliconic-based technique and Illumina sequencing approach. Compare to the first workflow suggested, this workflow has two branches with Freyja and COJAC tools accordingly, on the step of SARS-CoV-2 lineages abundances computation.

Both workflows were examined on synthetic and real-world datasets to evaluate efficiency in finding SARS-CoV-2 lineages abundances and their proportions in samples. The results obtained with developed workflows were additionally benchmarked with one state-of-the-art workflow called Lineagespot. The results acquired with developed Galaxy workflows using Freyja and COJAC approaches are promising.

Developed Galaxy workflows with Freyja-based lineages abundances computation showed promising results in detecting expected lineages in cases of single lineage or two lineages were expected. However, the results were biased by detecting other unexpected lineages. There were detected unexpected lineages in all samples of the mock dataset, while they should not have appeared. The other Freyja-based method limitations are due to the fact that for two out of four examined real-world datasets (more precisely, for Californian and UK ones) all detected SARS-CoV-2 variants were labeled as “Other” when they should have been labeled by WHO designation names. This led to non-informative plots which rely on these labels and the value of SARS-CoV-2 variant proportions corresponding to these WHO designation labels. Nonetheless, lineages were detected for all real-world datasets using Pango naming system.

Galaxy workflow with COJAC implemented for delineation showed slightly more efficient results, compared to Freyja-based workflows. It was able to detect only lineages that were expected while experimenting on mock dataset. On mock samples where only a single lineage was expected, the proposed COJAC-based method showed more effective results, compared to the proposed Freyja-based method, but less effective, compared to Lineagespot approach. Worth it to say, that on mock samples, where only two lineages were expected, COJAC-based method showed better results, compared to Lineagespot, in detecting these two lineages, however, results were biased with detecting unexpected lineages (as it was for Freyja-based). As for other disadvantages of COJAC-based method, it works on amplicons only; thus, this method cannot be used for metatranscriptomes. Moreover, COJAC tool does not provide integrated visualization of its results like Freyja; hence, Jupyter Notebook was used outside of developed workflows to evaluate results on real-world datasets.

Due to both Freyja-based and COJAC-based workflows’ shortcomings, listed above, there is room for improvement. Throughout the thesis, a few starting points for further improvement of both methods were identified.

One starting point for improvement is, because of Freyja limits on WHO variant name designation for further plotting of results, visualization can be implemented in Galaxy workflow

independently. One option for that can be an addition of extra steps in developed Galaxy workflows using Jupyter notebooks in Galaxy [168, 169]. Galaxy opens the opportunity of using Jupyter notebooks in Galaxy workflows.

The other idea for improvement of developed Galaxy workflows focused on COJAC-based branch. The visualization is currently missing, thus, an extra visualization step is needed to be implemented. However, in the case of COJAC the suggestion for further improvement differs from the one for Freyja. COJAC outputs can be visualized in CoV-Spectrum [105] format and uploaded to this platform. This can be accomplished with the help of Python scripts as it was done for Swiss dataset analysis by Jahn et al. [6, 152]. Unfortunately, code is not provided under open source license, thus, currently cannot be used for the improvement of Galaxy workflows.

The potential next step for enhancement could be to create tutorials based on Galaxy training materials [169] that will cover how to use the developed Galaxy workflows for SARS-CoV-2 wastewater surveillance step by step and an explanation of what each step accomplishes. The further step would be to connect them to the automatic Galaxy bot to analyze data on a regular basis in a similar way it is currently managed for clinical data.

Wastewater surveillance has proved to be efficient in the early detection of viruses and its lineages which is crucial for the early detection of newly emerged variants. Even though this thesis focused on SARS-CoV-2 wastewater samples analysis, it is not a limit. As an option for further improvement of proposed pipelines, surveillance can be extended to other viruses. For example, wastewater surveillance is already practiced extensively for detecting poliovirus in the US, more specifically, in New York [77], and plays an essential role in global health. It is desirable to expand wastewater surveillance methods to other pathogens beyond SARS-CoV-2 and validate them with cases-based epidemiological data. To be applicable to locations without centralized sewer infrastructure, current methods must be adapted and optimized [88]. To fully exploit the potential of wastewater surveillance for global pathogen surveillance, international sharing of wastewater-based pathogen sequencing data will be required.

6 Appendix

6.1 Galaxy tool wrappers

Macros for collection of related Freyja tool wrappers:
https://github.com/galaxyproject/tools-iuc/blob/master/tools/freyja/macros.xml
Tool wrapper for Freyja: Call variants and get sequencing depth information:
https://github.com/galaxyproject/tools-iuc/blob/master/tools/freyja/freyja_variants.xml
Tool wrapper for Freyja: Bootstrapping method:
https://github.com/galaxyproject/tools-iuc/blob/master/tools/freyja/freyja_boot.xml
Tool wrapper for Freyja: Demix lineage abundances:
https://github.com/galaxyproject/tools-iuc/blob/master/tools/freyja/freyja_demix.xml
Tool wrapper for Freyja: Aggregate and visualize demixed results:
https://github.com/galaxyproject/tools-iuc/blob/master/tools/freyja/freyja_aggregate_plot.xml

Macros for collection of related Cojac tool wrappers:
https://github.com/galaxyproject/tools-iuc/blob/master/tools/cojac/macros.xml
Tool wrapper for Cojac: mutbamscan:
https://github.com/galaxyproject/tools-iuc/blob/master/tools/cojac/cooc_mutbamscan.xml
Tool wrapper for Cojac: tabmut:
https://github.com/galaxyproject/tools-iuc/blob/master/tools/cojac/cooc_tabmut.xml
Tool wrapper for Cojac: pubmut:
https://github.com/galaxyproject/tools-iuc/blob/master/tools/cojac/cooc_pubmut.xml

6.2 Published Galaxy workflows

The workflow for SARS-CoV-2 wastewater Illumina-sequenced ARTIC data analysis:
https://usegalaxy.eu/u/polina/w/ww-sars-cov-2-artic-pe-workflow

The workflow for SARS-CoV-2 wastewater Illumina-sequenced metatranscriptomics data analysis:
https://usegalaxy.eu/u/polina/w/ww-sars-cov-2-metatranscriptomics-pe-workflow

6.3 Galaxy histories

6.3.1 Galaxy history links for Freyja-based branch on real datasets

California, US (PRJNA661613):
https://usegalaxy.eu/u/polina/h/sf-dataset-prjna661613-covid-19-variation-analysis-on-wgs-pe-data
Wales and Northwest England, UK (PRJEB42191):
https://usegalaxy.eu/u/polina/h/uk-dataset-COJAC-prjeb42191-covid-19-variation-analysis-on-artic-pe-data-1
Ontario, Canada (PRJNA824537):
https://usegalaxy.eu/u/polina/h/ca-dataset-freyja-prjna824537-covid-19-variation-analysis-on-artic-pe-data
Washington, US (PRJNA765346):
https://usegalaxy.eu/u/polina/h/us-dataset-freyja-prjna765346-covid-19-variation-analysis-on-artic-pe-data-1-1-1-1

6.3.2 Galaxy history links for COJAC-based branch on real datasets

Wales and Northwest England, UK (PRJEB42191):
https://usegalaxy.eu/u/polina/h/uk-dataset-COJAC-prjeb42191-covid-19-variation-analysis-on-artic-pe-data-1
Ontario, Canada (PRJNA824537):
https://usegalaxy.eu/u/polina/h/ca-dataset-COJAC-prjna824537-covid-19-variation-analysis-on-artic-pe-data
Washington, US (PRJNA765346):
https://usegalaxy.eu/u/polina/h/us-dataset-COJAC-prjna765346-covid-19-variation-analysis-on-artic-pe-data

6.4 Computation of overall lineage abundances for Freyja output

Python script to compute the overall lineage abundances proportions of considered lineages in mock dataset for Freyja output:
https://github.com/PlushZ/mthesis-sars-ww-galaxy/blob/main/overall-lineage-abundances-freyja.ipynb

6.5 Visualizations of results on mock dataset

Jupyter notebooks for visualizations of results on mock dataset:
https://github.com/PlushZ/mthesis-sars-ww-galaxy/blob/main/mock-benchmark-visualizations.ipynb

6.6 Visualizations of results on real-world dataset

Jupyter notebooks for visualizations of real-world datasets for SARS-CoV-2 wastewater surveillance and results on real-world datasets:
`https://github.com/PlushZ/mthesis-sars-ww-galaxy/blob/main/realworld-benchmark-visualizations.ipynb`

6.7 This document itself in Latex

This thesis itself is written in Latex and is stored in GitHub repository:
`https://github.com/PlushZ/master-thesis-latex`

6.8 Further figures

6.8.1 Existing Galaxy workflows that were not changed

Other existing Galaxy workflows for SARS-CoV-2 clinical data surveillanve that were not taken as a basis for improvement and repurposing in this thesis are presented in fig. 37 and fig. 38.

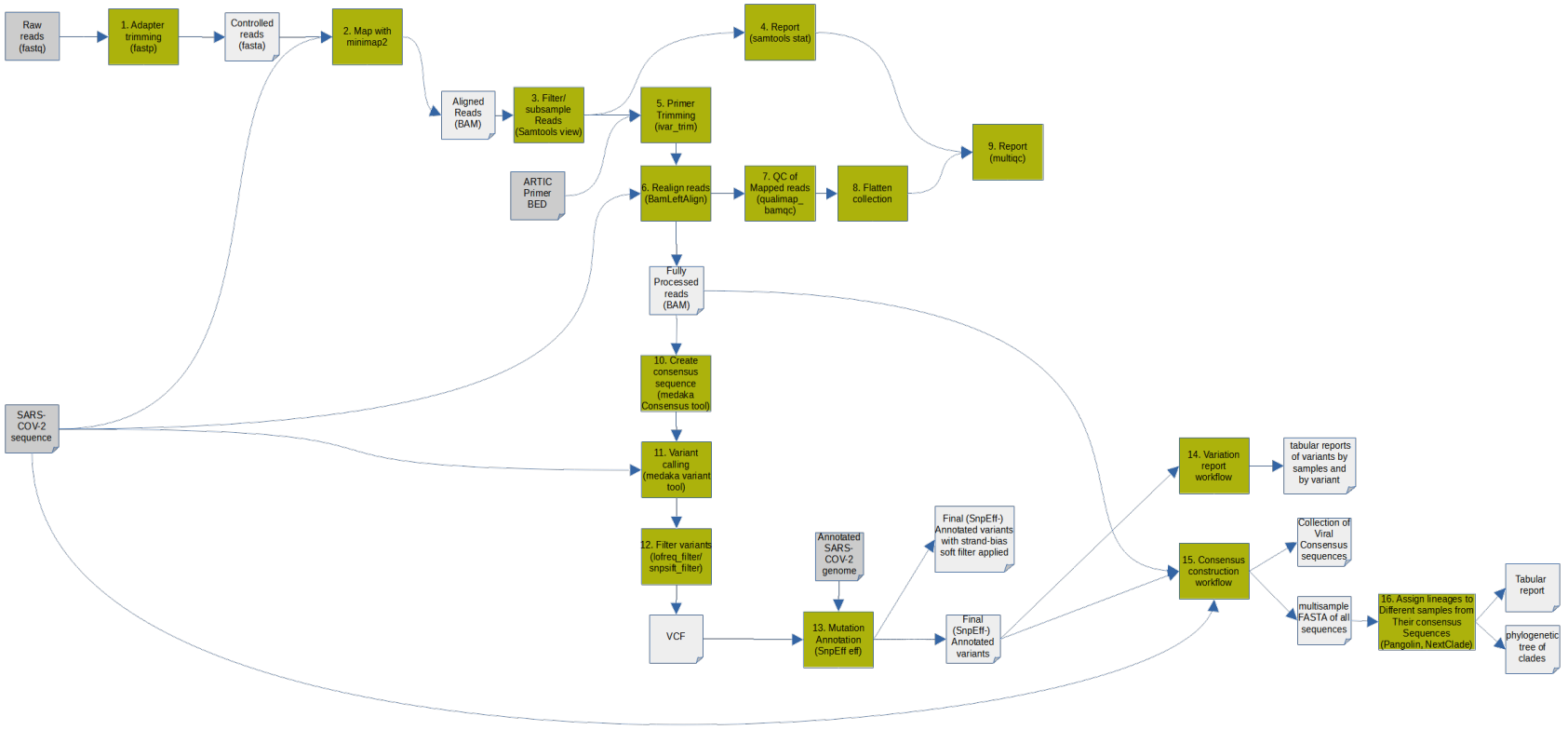


Figure 37: One of four existing Galaxy workflow for SARS-CoV-2 clinical data surveillance for single-end reads data extracted with amplicon-based technique and sequenced with Nanopore sequencing approach.

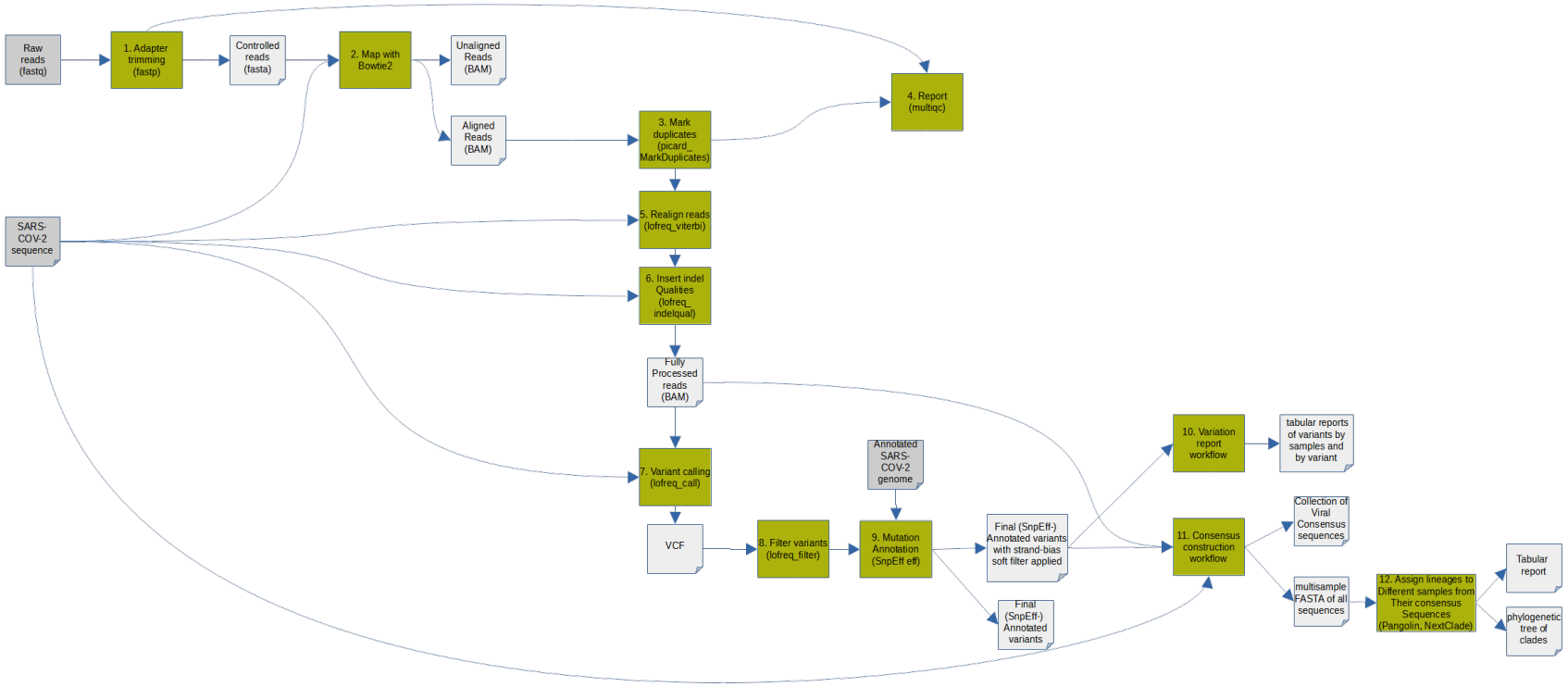


Figure 38: One of four existing Galaxy workflow for SARS-CoV-2 clinical data surveillance for single-end reads data extracted with metatranscriptomic-based technique and sequenced with Illumina sequencing approach.

6.8.2 Comparison of lineage proportions detected by tools with expected proportion

6.8.2.1 Comparison of lineage proportions in all mock samples

Figure 39, fig. 40 and fig. 41 depict parallel coordinates for all samples considering only detection of Delta, BA.1, and BA.2 lineage respectively by Freyja and COJAC. Plots compare the lineage proportions detected by Freyja and COJAC with each other as well as with expected proportion. Left axis represents expected proportion of Delta, middle axis represents proportion of the lineage detected by COJAC, while right axis represents proportion of the lineage detected by Freyja.

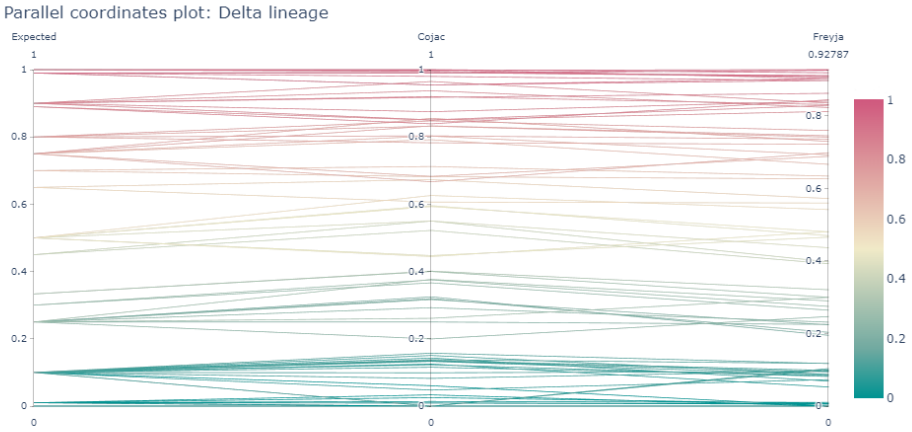


Figure 39: Delta lineage proportions detected by Freyja and COJAC, and expected proportion.

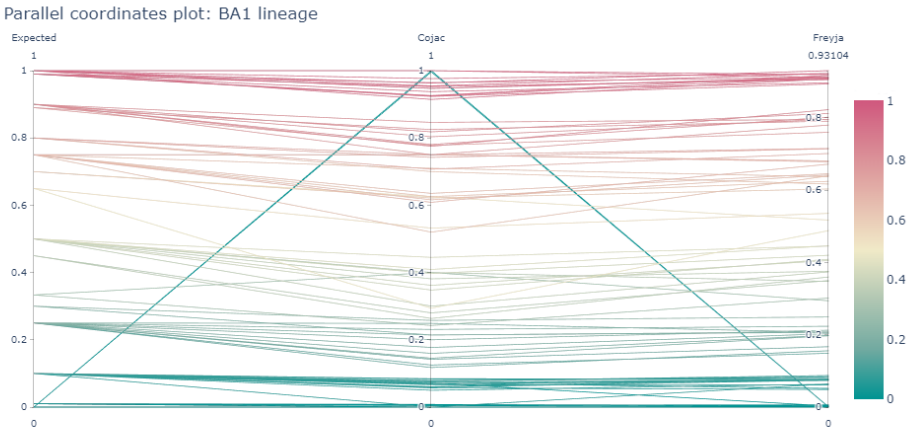


Figure 40: BA.1 lineage proportions detected by Freyja and COJAC, and expected proportion.

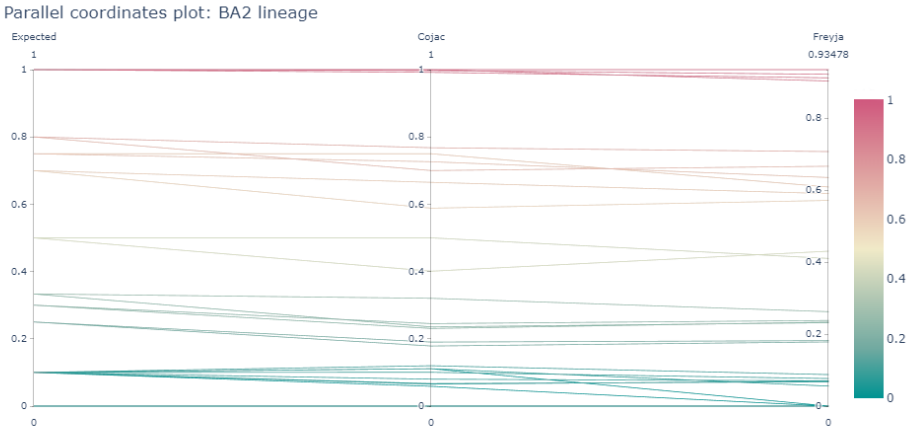


Figure 41: BA.2 lineage proportions detected by Freyja and COJAC, and expected proportion.

6.8.2.2 Comparison of lineage proportions in mock samples with two lineages expected

Figure 42, fig. 43 and fig. 44 depict parallel coordinates for "two lineages" group of samples considering only detection of Delta, BA.1, and BA.2 lineage respectively by Freyja and COJAC. Although, for the "two lineages" group of samples, it is not that reasonable to separate graphs into different lineages but focus on the ratio between two expected lineages. Even though, detected proportion of certain expected lineage was worthwhile to to have a look at. Thus, parallel coordinates graphs were generated in the way of one graph per lineage. Left axis represents expected proportion of Delta, middle axis represents proportion of Delta lineage detected by COJAC, while right axis represents proportion of Delta lineage detected by Freyja

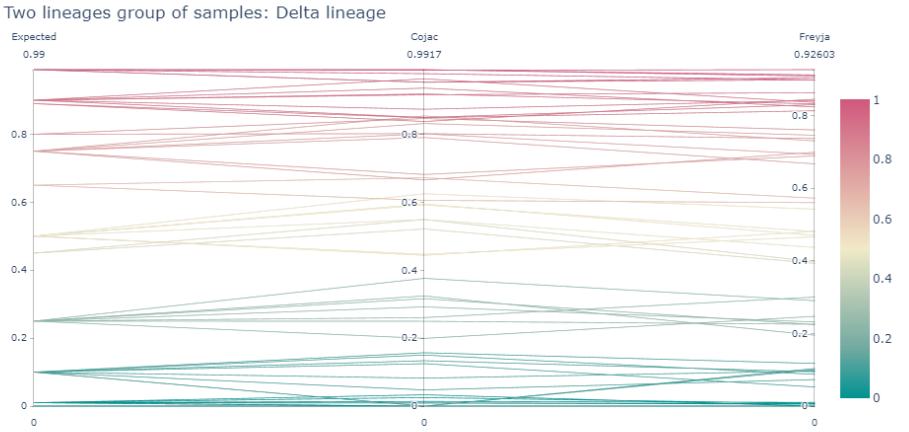


Figure 42: Delta lineage proportions detected by Freyja and COJAC, and expected proportion for samples with two lineages expected.

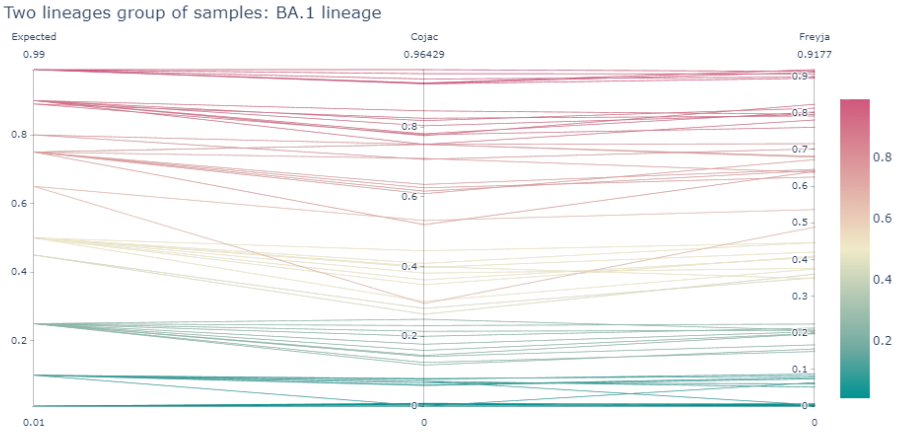


Figure 43: BA.1 lineage proportions detected by Freyja and COJAC, and expected proportion for samples with two lineages expected.

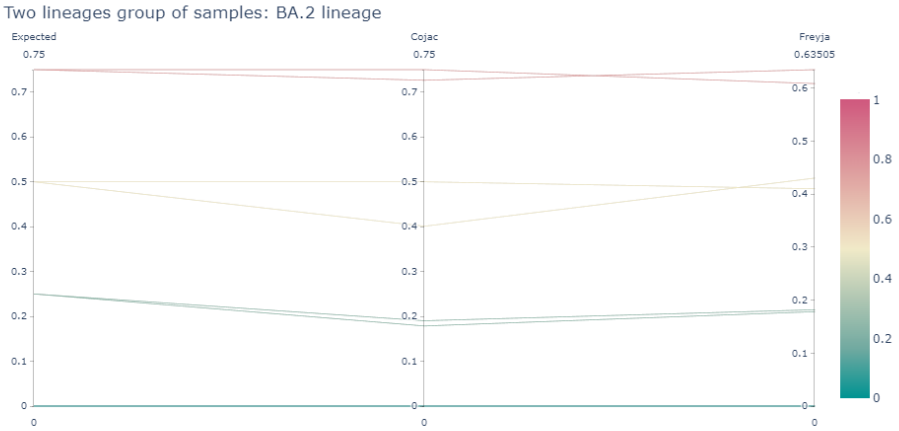
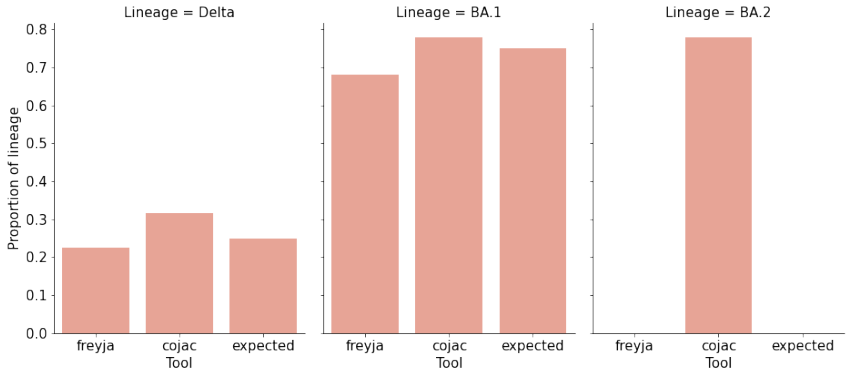


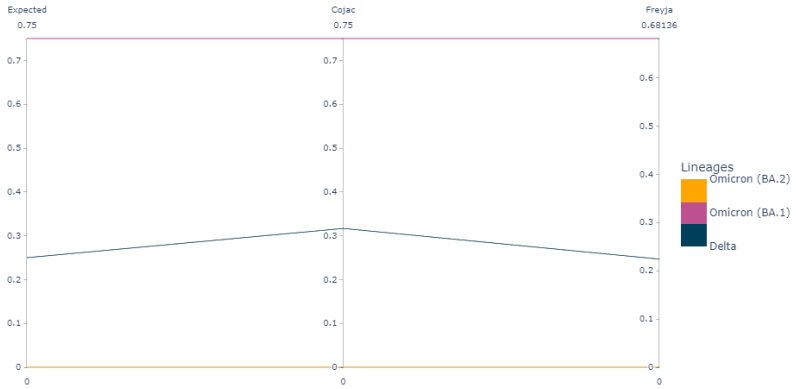
Figure 44: BA.2 lineage proportions detected by Freyja and COJAC, and expected proportion for samples with two lineages expected.

6.8.3 Example of different types of plots for one mock sample

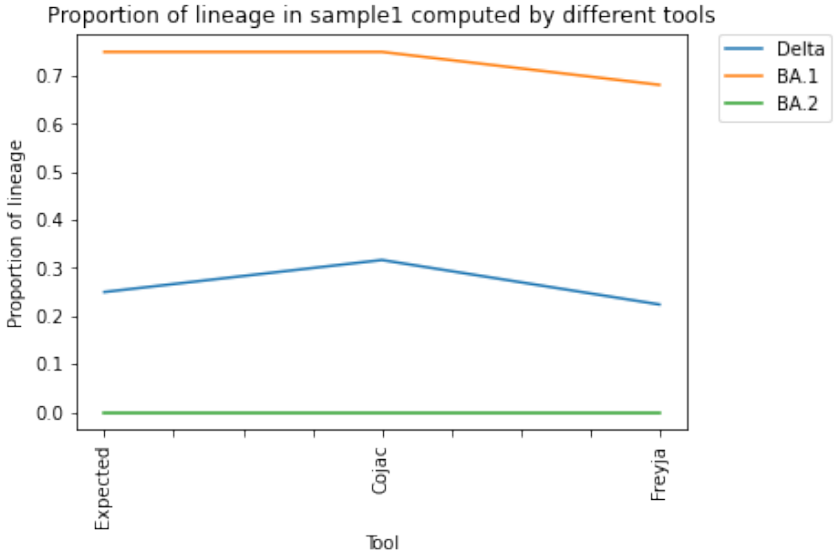
Figure 45a, fig. 45b and fig. 45c show plots generated per sample to make detailed observations. The example for the sample1 is provided below. However, these graphs were constructed for every sample during this master thesis. Types of plots that were generated per one sample: i) bar plot (fig. 45a); ii) parallel coordinates plot (fig. 45b); iii) line plot (fig. 45c), to look at absolute values of lineage proportion against scaled values in parallel coordinates. The comparison with expected lineage proportions was included in these plots. .



(a) Bar plot



(b) Parallel coordinates



(c) Line plot

Figure 45: Example of different types of plots for sample 1 of mock dataset

6.8.4 Distribution of lineages proportions detected by Lineagespot across all mock samples

Figure 46 represents the proportions of every lineage detected by Lineagespot on mock dataset. Observed that most of the samples, more than 50%, according to Lineagespot, carry a small proportion, no more than 0.15, of lineage abundance, while higher proportions of all 3 lineages are present in less amount of samples.

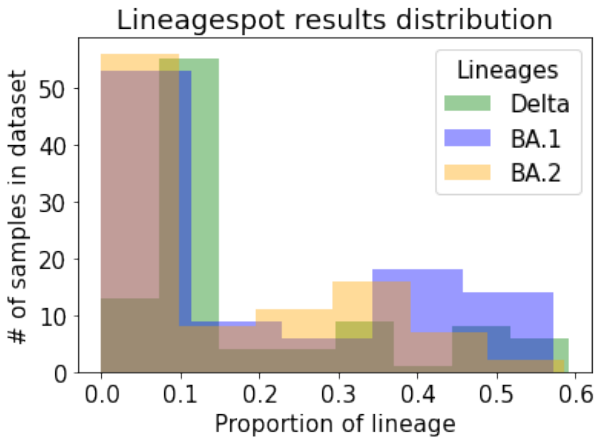


Figure 46: All three considered lineages (Delta, BA.1, BA.2) proportion distribution among results detected by Lineagespot on mock dataset.

6.9 Supplementary tables

6.9.1 Freyja aggregated demixed data

92

	summarized	lineages	abundances	resid	coverage
SRR12596165.fastq	"[('Other', 0.9999999999972105)]"	B.10 B.47 B.23 B.26 B.1.14 B.20 B	0.14285714 0.14285714 0.14285714 0.14285714 0.14285714 0.14285714 0.14285714	2.63E-12	1.926550271
SRR12596166.fastq	"[('Other', 0.9999999999983175)]"	B.1.1.174 B.10 B.47 B.23 B.26 B.1.14 B.20 B	0.55555600 0.06349200 0.06349200 0.06349200 0.06349200 0.06349200 0.06349200 0.06349200	1.162673375	1.926550271
SRR12596167.fastq	"[('Other', 0.9999999999996593)]"	B.10 B.47 B.23 B.26 B.1.14 B.20 B	0.14285714 0.14285714 0.14285714 0.14285714 0.14285714 0.14285714 0.14285714	0.75	1.926550271
SRR12596168.fastq	"[('Other', 0.9999999999582964)]"	B.1.1.174 B.1.1.161 B.1.564 B.1.12 B.1.607 B.1.413 B.1.324 B.1.453 B.1.1.372 B.1.1 B.1.1.92 B.1.1.294 B.1.1.180 B.1.1.43 B.1.1.59 B.1.1.463 B.1.1.402 B.1.1.10 B.1.1.208 B.1.1.61 B.10 B.47 B.23 B.26 B.1.14 B.20 B	0.27027000 0.08080800 0.05555550 0.05555550 0.05555550 0.05555550 0.05555550 0.05555550 0.02377383 0.02377383 0.02377383 0.02377383 0.02377383 0.02377383 0.02377383 0.02377383 0.02377383 0.02377383 0.02377383 0.02377383 0.00432900 0.00432900 0.00432900 0.00432900 0.00432900 0.00432900 0.00432900	0.4968557797	1.926550271
SRR12596169.fastq	"[('Other', 0.9999999999979102)]"	B.1.111 B.1.1.174 B.1.479 B.1.22 B.1.533 B.1.12 B.1.607 B.1.453 B.1.324 B.1.413 B.1.564 B.1.199 B.1.378 B.1.201 B.1.215 B.1	0.14432305 0.14285700 0.09853229 0.06999717 0.05158700 0.04640292 0.04640292 0.04640292 0.04640292 0.04640292 0.04640292 0.04285720 0.04285720 0.04285720 0.04285720 0.04285720	0.7140985235	1.926550271
SRR12596170.fastq	"[('Other', 0.9999999999987905)]"	B.1.533 B.1.370 B.1.301 B.1.424 B.1.111 B.1.479 B.1.201 B.1.378 B.1.199 B.1.215 B.1 B.1.22 B.1.453 B.1.324 B.1.413 B.1.607 B.1.12 B.1.564 B.1.1.161 B.10 B.47 B.23 B.26 B.1.14 B.20 B	0.15773800 0.11717199 0.11641401 0.11641400 0.05154872 0.03906387 0.03904760 0.03904760 0.03904760 0.03904760 0.03904760 0.03189755 0.01718081 0.01718081 0.01718081 0.01718081 0.01718081 0.01718081 0.00892900 0.00892857 0.00892857 0.00892857 0.00892857 0.00892857 0.00892857 0.00892857	0.4778312073	1.926550271
SRR12596171.fastq	"[('Other', 0.9979179208943117)]"	B.1.509 B.1.12 B.1.607 B.1.453 B.1.324 B.1.413 B.1.564 B.10 B.47 B.23 B.26 B.1.14 B.20 B	0.87096800 0.01560282 0.01560282 0.01560282 0.01560282 0.01560282 0.01560282 0.00476186 0.00476186 0.00476186 0.00476186 0.00476186 0.00476186 0.00476186	0.3834278298	1.926550271
SRR12596172.fastq	"[('Other', 0.9999999999983878)]"	B.1.301 B.1.424 B.1.370 B.1.1.174 B.1.453 B.1.12 B.1.607 B.1.324 B.1.413 B.1.564 B.47 B.20 B.1.14 B.26 B.23 B B.10 B.1.533 B.1.111 B.1.22 B.1.479	0.27497564 0.27497564 0.27312572 0.03731300 0.01588226 0.01588226 0.01588226 0.01588226 0.01588226 0.01588226 0.00529100 0.00529100 0.00529100 0.00529100 0.00529100 0.00529100 0.00529100 0.00367100 0.00125773 0.00118732 0.00116342	0.54272957	1.926550271
SRR12596173.fastq	"[('Other', 0.9999999999972105)]"	B.10 B.47 B.23 B.26 B.1.14 B.20 B	0.14285714 0.14285714 0.14285714 0.14285714 0.14285714 0.14285714 0.14285714	2.63E-12	1.926550271
SRR12596174.fastq	"[('Other', 0.999999999979265)]"	B.1.509 B.1.2 B.1.370 B.1.301 B.1.424 B.1.324 B.1.564 B.1.413 B.1.453 B.1.607 B.1.12 B.1.596 B.47 B.20 B.1.14 B.26 B.23 B B.10 B.1 B.1.215 B.1.201 B.1.378 B.1.199	0.28000000 0.20118712 0.06037074 0.06035913 0.06035913 0.04567167 0.04567167 0.04567167 0.04567167 0.04567167 0.04567167 0.02797988 0.00432900 0.00432900 0.00432900 0.00432900 0.00432900 0.00432900 0.00432900 0.00108220 0.00108220 0.00108220 0.00108220 0.00108220	0.8515939213	1.926550271
SRR12596175.fastq	"[('Other', 0.9905735237538151)]"	B.1.509 B.1.2 B.1.382 B.1.111 B.10 B.47 B.23 B.26 B.1.14 B.20 B B.1.479 B.1.22	0.68478300 0.27368400 0.00451200 0.00294756 0.00294557 0.00294557 0.00294557 0.00294557 0.00294557 0.00294557 0.00294557 0.00245515 0.00157281	0.2451972591	1.926550271

Table 9: Aggregated table for Californian real-world dataset (PRJNA661613) - output of 'Freyja: aggregate' tool that aggregate demixed data for all samples in dataset

References

[1] *galaxyproject/tools-iuc: Tool Shed repositories maintained by the Intergalactic Utilities Commission*. en. URL: https://github.com/galaxyproject/tools-iuc.

[2] Polina Polunina. *PlushZ/mthesis-sars-ww-galaxy*. original-date: 2022-11-02T09:15:28Z. Nov. 2022. URL: https://github.com/PlushZ/mthesis-sars-ww-galaxy.

[3] *PlushZ/master-thesis-latex*. en. URL: https://github.com/PlushZ/master-thesis-latex.

[4] Gertjan Medema et al. "Presence of SARS-Coronavirus-2 RNA in Sewage and Correlation with Reported COVID-19 Prevalence in the Early Stage of the Epidemic in The Netherlands". In: *Environmental Science & Technology Letters* 7.7 (July 2020). Publisher: American Chemical Society, pp. 511–516. DOI: 10.1021/acs.estlett.0c00357.

[5] Hao Wang et al. "The amount of SARS-CoV-2 RNA in wastewater relates to the development of the pandemic and its burden on the health system". en. In: *iScience* 25.9 (Sept. 2022), p. 105000. ISSN: 2589-0042. DOI: 10.1016/j.isci.2022.105000.

[6] Katharina Jahn et al. "Early detection and surveillance of SARS-CoV-2 genomic variants in wastewater using COJAC". en. In: *Nature Microbiology* (July 2022). Publisher: Nature Publishing Group, pp. 1–10. ISSN: 2058-5276. DOI: 10.1038/s41564-022-01185-x.

[7] Dannon Baker et al. "No more business as usual: Agile and effective responses to emerging pathogen threats require open data and open analytics". en. In: *PLOS Pathogens* 16.8 (2020). Publisher: Public Library of Science, e1008643. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1008643.

[8] Wolfgang Maier et al. "Ready-to-use public infrastructure for global SARS-CoV-2 monitoring". en. In: *Nature Biotechnology* 39.10 (Oct. 2021). Number: 10 Publisher: Nature Publishing Group, pp. 1178–1179. ISSN: 1546-1696. DOI: 10.1038/s41587-021-01069-1.

[9] Han Mei, Sergei Kosakovsky Pond and Anton Nekrutenko. "Stepwise Evolution and Exceptional Conservation of ORF1a/b Overlap in Coronaviruses". In: *Molecular Biology and Evolution* 38.12 (Sept. 2021), pp. 5678–5684. ISSN: 0737-4038. DOI: 10.1093/molbev/msab265.

[10] Darren P. Martin et al. "The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages". In: *Cell* 184.20 (Sept. 2021), 5189–5200.e7. ISSN: 0092-8674. DOI: 10.1016/j.cell.2021.09.003.

[11] Fan Wu et al. "A new coronavirus associated with human respiratory disease in China". en. In: *Nature* 579.7798 (Mar. 2020). Number: 7798 Publisher: Nature Publishing Group, pp. 265–269. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2008-3.

[12] *Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV)*. en. URL: https://www.who.int/news/item/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov) (visited on 09/15/2022).

[13] World Health Organization. "WHO Director-General's opening remarks at the media briefing on COVID-19-11 March 2020". In: (2020). Publisher: Geneva, Switzerland.

[14] Domenico Cucinotta and Maurizio Vanelli. "WHO Declares COVID-19 a Pandemic". In: *Acta Bio Medica : Atenei Parmensis* 91.1 (2020), pp. 157–160. ISSN: 0392-4203. DOI: 10.23750/abm.v91i1.9397.

[15] *Weekly epidemiological update on COVID-19 - 14 September 2022*. en. URL: https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---14-september-2022 (visited on 09/15/2022).

[16] Md Asiful Islam et al. "Prevalence and characteristics of fever in adult and paediatric patients with coronavirus disease 2019 (COVID-19): A systematic review and meta-analysis of 17515 patients". In: *PLoS ONE* 16.4 (Apr. 2021), e0249788. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0249788.

[17] Jeyasakthy Saniasiaya, Md Asiful Islam and Baharudin Abdullah. "Prevalence of Olfactory Dysfunction in Coronavirus Disease 2019 (COVID-19): A Meta-analysis of 27,492 Patients". In: *The Laryngoscope* 131.4 (Apr. 2021), pp. 865–878. ISSN: 0023-852X. DOI: 10.1002/lary.29286.

[18] Jeyasakthy Saniasiaya, Md Asiful Islam and Baharudin Abdullah. "Prevalence and Characteristics of Taste Disorders in Cases of COVID-19: A Meta-analysis of 29,349 Patients". en. In: *Otolaryngology–Head and Neck Surgery* 165.1 (July 2021). Publisher: SAGE Publications Inc, pp. 33–42. ISSN: 0194-5998. DOI: 10.1177/0194599820981018.

[19] Akosua Adom Agyeman et al. "Smell and Taste Dysfunction in Patients With COVID-19: A Systematic Review and Meta-analysis". In: *Mayo Clinic Proceedings* 95.8 (Aug. 2020), pp. 1621–1631. ISSN: 0025-6196. DOI: 10.1016/j.mayocp.2020.05.030.

[20] Daniel P. Oran and Eric J. Topol. "The Proportion of SARS-CoV-2 Infections That Are Asymptomatic". In: *Annals of Internal Medicine* (Jan. 2021), pp. M20–6976. ISSN: 0003-4819. DOI: 10.7326/M20-6976.

[21] *Management of Patients with Confirmed 2019-nCoV | CDC*. Mar. 2020. URL: https://web.archive.org/web/20200302201644/https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html (visited on 09/15/2022).

[22] CDC. *Post-COVID Conditions*. en-us. Sept. 2022. URL: https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html (visited on 09/15/2022).

[23] CDC. *Healthcare Workers*. en-us. Feb. 2020. URL: https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-care/clinical-considerations-index.html (visited on 09/15/2022).

[24] *Risk assessment: Outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): increased transmission beyond China – fourth update – European Sources Online*. en-GB. URL: https://www.europeansources.info/record/risk-assessment-outbreak-of-severe-acute-respiratory-syndrome-coronavirus-2-sars-cov-2-increased-transmission-beyond-china-fourth-update/ (visited on 09/15/2022).

[25] Kristian G. Andersen et al. "The proximal origin of SARS-CoV-2". en. In: *Nature Medicine* 26.4 (Apr. 2020). Number: 4 Publisher: Nature Publishing Group, pp. 450–452. ISSN: 1546-170X. DOI: 10.1038/s41591-020-0820-9.

[26] Na Zhu et al. "A Novel Coronavirus from Patients with Pneumonia in China, 2019". In: *New England Journal of Medicine* 382.8 (Feb. 2020). Publisher: Massachusetts Medical Society, pp. 727–733. ISSN: 0028-4793. DOI: 10.1056/NEJMoa2001017.

[27] Peng Zhou et al. "A pneumonia outbreak associated with a new coronavirus of probable bat origin". In: *Nature* 579.7798 (2020), pp. 270–273. ISSN: 0028-0836. DOI: 10.1038/s41586-020-2012-7.

[28] Jitendra Singh Rathore and Chaitali Ghosh. "Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), a newly emerged pathogen: an overview". eng. In: *Pathogens and Disease* 78.6 (Aug. 2020), ftaa042. ISSN: 2049-632X. DOI: 10.1093/femspd/ftaa042.

[29] Sunil Thomas. "The Structure of the Membrane Protein of SARS-CoV-2 Resembles the Sugar Transporter SemiSWEET". en. In: *Pathogens and Immunity* 5.1 (Oct. 2020), pp. 342–363. ISSN: 2469-2964. DOI: 10.20411/pai.v5i1.377.

[30] *Variants, Sublineages, and Recombinants: The Constantly Changing Genome of SARS-CoV-2*. en-US. URL: https://www.rockefellerfoundation.org/case-study/variants-sublineages-and-recombinants-the-constantly-changing-genome-of-sars-cov-2/ (visited on 08/03/2022).

[31] Jobin John Jacob et al. "Evolutionary Tracking of SARS-CoV-2 Genetic Variants Highlights an Intricate Balance of Stabilizing and Destabilizing Mutations". In: *mBio* 12.4 (July 2021). Publisher: American Society for Microbiology, e01188–21. DOI: 10.1128/mBio.01188-21.

[32] *Updated Nextstrain SARS-CoV-2 clade naming strategy*. en. URL: https://nextstrain.org//blog/2021-01-06-updated-SARS-CoV-2-clade-naming (visited on 09/16/2022).

[33] *SARS-CoV-2 variants of concern as of 4 August 2022*. en. URL: https://www.ecdc.europa.eu/en/covid-19/variants-concern (visited on 08/05/2022).

[34] Delphine Planas et al. "Sensitivity of infectious SARS-CoV-2 B.1.1.7 and B.1.351 variants to neutralizing antibodies". en. In: *Nature Medicine* 27.5 (May 2021). Number: 5 Publisher: Nature Publishing Group, pp. 917–924. ISSN: 1546-170X. DOI: 10.1038/s41591-021-01318-5.

[35] Áine O'Toole et al. "Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool". In: *Virus Evolution* 7.2 (Dec. 2021), veab064. ISSN: 2057-1577. DOI: 10.1093/ve/veab064.

[36] *Tracking SARS-CoV-2 variants*. en. URL: https://www.who.int/activities/tracking-SARS-CoV-2-variants (visited on 08/05/2022).

[37] Estee Cramer and Nick Reich. *A modeler's primer for working with SARS-CoV-2 genomic data*. en. URL: http://reichlab.io//2022/02/15/genbank-data.html (visited on 08/05/2022).

[38] Andrew Rambaut et al. "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology". en. In: *Nature Microbiology* 5.11 (Nov. 2020). Number: 11 Publisher: Nature Publishing Group, pp. 1403–1407. ISSN: 2058-5276. DOI: 10.1038/s41564-020-0770-5.

[39] Erik M. Volz, Katia Koelle and Trevor Bedford. "Viral Phylodynamics". en. In: *PLOS Computational Biology* 9.3 (2013). Publisher: Public Library of Science, e1002947. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002947.

[40] James Hadfield et al. "Nextstrain: real-time tracking of pathogen evolution". In: *Bioinformatics* 34.23 (Dec. 2018), pp. 4121–4123. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty407.

[41] *SARS-CoV-2 clade naming strategy for 2022*. en. URL: https://nextstrain.org//blog/2022-04-29-SARS-CoV-2-clade-naming-2022 (visited on 09/16/2022).

[42] CDC. *Labs*. en. Feb. 2020. URL: https://www.cdc.gov/coronavirus/2019-ncov/lab/resources/sars-cov2-testing-strategies.html (visited on 09/16/2022).

[43] Matteo Chiara et al. "Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities". In: *Briefings in Bioinformatics* (Dec. 2020), bbaa297. ISSN: 1467-5463. DOI: 10.1093/bib/bbaa297.

[44] Yao Meng et al. "An efficient metatranscriptomic approach for capturing RNA virome and its application to SARS-CoV-2". In: *Journal of Genetics and Genomics* 48.9 (Sept. 2021), pp. 860–862. ISSN: 1673-8527. DOI: 10.1016/j.jgg.2021.08.005.

[45] Arnold W. Lambisia et al. "Optimization of the SARS-CoV-2 ARTIC Network V4 Primers and Whole Genome Sequencing Protocol". In: *Frontiers in Medicine* 9 (2022). ISSN: 2296-858X. URL: https://www.frontiersin.org/articles/10.3389/fmed.2022.836728 (visited on 07/31/2022).

[46] DNA Pipelines R&d et al. *COVID-19 ARTIC v3 Illumina library construction and sequencing protocol*. en. May 2020. URL: https://www.protocols.io/view/covid-19-artic-v3-illumina-library-construction-an-bgxjjxkn (visited on 07/31/2022).

[47] DNA Pipelines R&d et al. *COVID-19 ARTIC v4.1 Illumina library construction and sequencing protocol - tailed method*. en. Mar. 2022. URL: https://www.protocols.io/view/covid-19-artic-v4-1-illumina-library-construction-b4myqu7w (visited on 07/31/2022).

[48] Josh Quick. *nCoV-2019 sequencing protocol*. en. Jan. 2020. URL: https://www.protocols.io/view/ncov-2019-sequencing-protocol-bbmuik6w (visited on 07/31/2022).

[49] Nathan D. Grubaugh et al. "An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar". en. In: *Genome Biology* 20.1 (Dec. 2019), p. 8. ISSN: 1474-760X. DOI: 10.1186/s13059-018-1618-7.

[50] Nicholas Loman. "A quick guide to tiling amplicon sequencing and downstream bioinformatics analysis". In: (). URL: https://artic.network/quick-guide-to-tiling-amplicon-sequencing-bioinformatics.html.

[51] Kentaro Itokawa et al. "Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR". In: *PLoS ONE* 15.9 (Sept. 2020), e0239403. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0239403.

[52] Minfeng Xiao et al. "Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples". In: *Genome Medicine* 12 (June 2020), p. 57. ISSN: 1756-994X. DOI: 10.1186/s13073-020-00751-4.

[53] *Difference Between Nanopore and Illumina Sequencing*. en. Section: Molecular Biology. July 2021. URL: https://www.differencebetween.com/difference-between-nanopore-and-illumina-sequencing/ (visited on 10/18/2022).

[54] Martin Hunt et al. *ReadItAndKeep: rapid decontamination of SARS-CoV-2 sequencing reads*. en. Pages: 2022.01.21.477194 Section: New Results. Jan. 2022. DOI: 10.1101/2022.01.21.477194.

[55] "Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome". en-US. In: (July 2020). Type: dataset Version: 2. URL: http://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2.

[56] Erik Garrison and Gabor Marth. *Haplotype-based variant detection from short-read sequencing*. arXiv:1207.3907 [q-bio]. July 2012. DOI: 10.48550/arXiv.1207.3907.

[57] *Qualimap: evaluating next-generation sequencing alignment data | Bioinformatics | Oxford Academic*. URL: https://academic.oup.com/bioinformatics/article/28/20/2678/206551?login=false (visited on 10/29/2022).

[58] *Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data | Bioinformatics | Oxford Academic*. URL: https://academic.oup.com/bioinformatics/article/32/2/292/1744356?login=false (visited on 10/29/2022).

[59] Michelle L. Holshue et al. "First Case of 2019 Novel Coronavirus in the United States". In: *The New England Journal of Medicine* 382.10 (Mar. 2020), pp. 929–936. ISSN: 0028-4793. DOI: 10.1056/NEJMoa2001191.

[60] *COVID-19 Genomics UK Consortium*. URL: https://www.cogconsortium.uk/ (visited on 09/17/2022).

[61] Chaoran Chen et al. "Advancing genomic epidemiology by addressing the bioinformatics bottleneck: Challenges, design principles, and a Swiss example". en. In: *Epidemics* 39 (June 2022), p. 100576. ISSN: 1755-4365. DOI: 10.1016/j.epidem.2022.100576.

[62] *Swiss SARS-CoV-2 Sequencing Consortium (S3C)*. en. URL: https://bsse.ethz.ch/cevo/research/sars-cov-2/swiss-sars-cov-2-sequencing-consortium.html (visited on 09/05/2022).

[63] Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". en. In: *Scientific Data* 3.1 (Mar. 2016). Number: 1 Publisher: Nature Publishing Group, p. 160018. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18.

[64] John Towns et al. "XSEDE: Accelerating Scientific Discovery". In: *Computing in Science & Engineering* 16.5 (Sept. 2014). Conference Name: Computing in Science & Engineering, pp. 62–74. ISSN: 1558-366X. DOI: 10.1109/MCSE.2014.80.

[65] *ELIXIR-DE*. URL: https://www.denbi.de/elixir-de (visited on 09/30/2022).

[66] *ELIXIR: providing a sustainable infrastructure for life science data at European scale | Bioinformatics | Oxford Academic*. URL: https://academic.oup.com/bioinformatics/article/37/16/2506/6310171 (visited on 10/08/2022).

[67] *ELIXIR*. en. URL: https://elixir-europe.org/ (visited on 09/30/2022).

[68] *GalaxyProject SARS-CoV-2 analysis effort*. en. URL: https://galaxyproject.org/projects/covid19/ (visited on 10/03/2022).

[69] Wolfgang Maier et al. *Freely accessible ready to use global infrastructure for SARS-CoV-2 monitoring*. en. Pages: 2021.03.25.437046 Section: New Results. Mar. 2021. DOI: 10.1101/2021.03.25.437046.

[70] Heng Li. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. arXiv:1303.3997 [q-bio]. May 2013. DOI: 10.48550/arXiv.1303.3997.

[71] *Fast and accurate short read alignment with Burrows–Wheeler transform | Bioinformatics | Oxford Academic*. URL: https://academic.oup.com/bioinformatics/article/25/14/1754/225615?login=false (visited on 09/30/2022).

[72] *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome | Genome Biology | Full Text*. URL: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-3-r25 (visited on 09/30/2022).

[73] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2". en. In: *Nature Methods* 9.4 (Apr. 2012), pp. 357–359. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.1923.

[74] Heng Li. "Minimap2: pairwise alignment for nucleotide sequences". In: *Bioinformatics* 34.18 (Sept. 2018). arXiv:1708.01492 [q-bio], pp. 3094–3100. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/bty191.

[75] *LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets | Nucleic Acids Research | Oxford Academic*. URL: https://academic.oup.com/nar/article/40/22/11189/1152727?login=false (visited on 09/17/2022).

[76] *Bots for SARS-CoV-2 genome surveillance*. original-date: 2021-03-08T12:09:19Z. Oct. 2022. URL: https://github.com/usegalaxy-eu/ena-cog-uk-wfs (visited on 10/20/2022).

[77] Grace B Russo et al. "Re-emergence of Poliovirus in the United States: Considerations and Implications". en. In: *Annals of Neurology* n/a.n/a (). ISSN: 1531-8249. DOI: 10.1002/ana.26504.

[78] Vinson Wai-Shun Chan et al. "A systematic review on COVID-19: urological manifestations, viral RNA detection and special considerations in urological conditions". In: *World Journal of Urology* 39.9 (2021), pp. 3127–3138. ISSN: 0724-4983. DOI: 10.1007/s00345-020-03246-4.

[79] Manuel Döhla et al. "SARS-CoV-2 in Environmental Samples of Quarantined Households". en. In: *Viruses* 14.5 (May 2022). Number: 5 Publisher: Multidisciplinary Digital Publishing Institute, p. 1075. ISSN: 1999-4915. DOI: 10.3390/v14051075.

[80] Paola Foladori et al. "SARS-CoV-2 from faeces to wastewater treatment: What do we know? A review". eng. In: *The Science of the Total Environment* 743 (Nov. 2020), p. 140444. ISSN: 1879-1026. DOI: 10.1016/j.scitotenv.2020.140444.

[81] Jie Wang et al. "SARS-CoV-2 RNA detection of hospital isolation wards hygiene monitoring during the Coronavirus Disease 2019 outbreak in a Chinese hospital". en. In: *International Journal of Infectious Diseases* 94 (May 2020), pp. 103–106. ISSN: 1201-9712. DOI: 10.1016/j.ijid.2020.04.024.

[82] Warish Ahmed et al. "Detection of SARS-CoV-2 RNA in commercial passenger aircraft and cruise ship wastewater: a surveillance tool for assessing the presence of COVID-19 infected travellers". eng. In: *Journal of Travel Medicine* 27.5 (Aug. 2020), taaa116. ISSN: 1708-8305. DOI: 10.1093/jtm/taaa116.

[83] *New versions of Omicron are masters of immune evasion*. en. URL: https://www.science.org/content/article/new-versions-omicron-are-masters-immune-evasion (visited on 09/17/2022).

[84] Charles Schmidt. "Watcher in the wastewater". en. In: *Nature Biotechnology* 38.8 (Aug. 2020). Number: 8 Publisher: Nature Publishing Group, pp. 917–920. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0620-2.

[85] "Wastewater monitoring comes of age". en. In: *Nature Microbiology* (Aug. 2022). Publisher: Nature Publishing Group, pp. 1–2. ISSN: 2058-5276. DOI: 10.1038/s41564-022-01201-0.

[86] Smruthi Karthikeyan et al. "Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission". en. In: *Nature* (July 2022). Publisher: Nature Publishing Group, pp. 1–4. ISSN: 1476-4687. DOI: 10.1038/s41586-022-05049-6.

[87] Fredy Saguti et al. "Surveillance of wastewater revealed peaks of SARS-CoV-2 preceding those of hospitalized patients with COVID-19". en. In: *Water Research* 189 (Feb. 2021), p. 116620. ISSN: 0043-1354. DOI: 10.1016/j.watres.2020.116620.

[88] CDC. *National Wastewater Surveillance System*. en-us. May 2022. URL: https://www.cdc.gov/healthywater/surveillance/wastewater-surveillance/wastewater-surveillance.html (visited on 09/16/2022).

[89] Kata Farkas et al. "Wastewater and public health: the potential of wastewater surveillance for monitoring COVID-19". en. In: *Current Opinion in Environmental Science & Health*. Environmental Health: COVID-19 17 (Oct. 2020), pp. 14–20. ISSN: 2468-5844. DOI: 10.1016/j.coesh.2020.06.001.

[90] Warish Ahmed et al. "Minimizing errors in RT-PCR detection and quantification of SARS-CoV-2 RNA for wastewater surveillance". en. In: *Science of The Total Environment* 805 (Jan. 2022), p. 149877. ISSN: 0048-9697. DOI: 10.1016/j.scitotenv.2021.149877.

[91] Qianqian Zhang et al. "Molecular mechanism of interaction between SARS-CoV-2 and host cells and interventional therapy". en. In: *Signal Transduction and Targeted Therapy* 6.1 (June 2021). Number: 1 Publisher: Nature Publishing Group, pp. 1–19. ISSN: 2059-3635. DOI: 10.1038/s41392-021-00653-w.

[92] Karrie K. K. Ko, Kern Rei Chng and Niranjan Nagarajan. "Metagenomics-enabled microbial surveillance". en. In: *Nature Microbiology* 7.4 (Apr. 2022). Number: 4 Publisher: Nature Publishing Group, pp. 486–496. ISSN: 2058-5276. DOI: 10.1038/s41564-022-01089-w.

[93] Rene S. Hendriksen et al. "Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage". en. In: *Nature Communications* 10.1 (Mar. 2019). Number: 1 Publisher: Nature Publishing Group, p. 1124. ISSN: 2041-1723. DOI: 10.1038/s41467-019-08853-3.

[94] Jaeyoung K. Jung et al. "Cell-free biosensors for rapid detection of water contaminants". en. In: *Nature Biotechnology* 38.12 (Dec. 2020). Number: 12 Publisher: Nature Publishing Group, pp. 1451–1459. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0571-7.

[95] *Detecting coronavirus in waste water*. ru. May 2022. URL: https://ut.ee/en/node/113141 (visited on 09/17/2022).

[96] *Reoveeseire kaardirakendus | Terviseamet*. et. URL: https://www.terviseamet.ee/et/reoveeseire-kaardirakendus (visited on 09/17/2022).

[97] ismart. *Covid-19*. en-US. URL: http://trams.chem.uoa.gr/covid-19/ (visited on 09/17/2022).

[98] admin. *COVID-19 Wastewater Coalition Maps*. en-CA. URL: https://cwn-rce.ca/covid-19-wastewater-coalition/covid-19-wastewater-coalition-maps/ (visited on 09/17/2022).

[99] Colleen C. Naughton et al. *Show us the Data: Global COVID-19 Wastewater Monitoring Efforts, Equity, and Gaps*. en. Pages: 2021.03.14.21253564. Nov. 2021. DOI: 10.1101/2021.03.14.21253564.

[100] *Pandemic signals from the sewer—what virus levels in wastewater tell us*. en. URL: https://www.science.org/content/article/pandemic-signals-sewer-what-virus-levels-wastewater-tell-us (visited on 09/17/2022).

[101] Jakob McBroome et al. "A Daily-Updated Database and Tools for Comprehensive SARS-CoV-2 Mutation-Annotated Trees". In: *Molecular Biology and Evolution* 38.12 (Dec. 2021), pp. 5819–5824. ISSN: 1537-1719. DOI: 10.1093/molbev/msab264.

[102] *UShER Wiki — usher_wiki 0.0.1 documentation*. URL: https://usher-wiki.readthedocs.io/en/latest/# (visited on 10/06/2022).

[103] Katharina Jahn et al. *Detection and surveillance of SARS-CoV-2 genomic variants in wastewater*. en. Tech. rep. Type: article. medRxiv, July 2021, p. 2021.01.08.21249379. DOI: 10.1101/2021.01.08.21249379.

[104] *COJAC - CoOccurrence adJusted Analysis and Calling*. original-date: 2021-02-02T16:36:03Z. July 2022. URL: https://github.com/cbg-ethz/cojac (visited on 10/17/2022).

[105] Chaoran Chen et al. "CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants". In: *Bioinformatics* 38.6 (Mar. 2022), pp. 1735–1737. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab856.

[106] Renan Valieris et al. "A mixture model for determining SARS-Cov-2 variant composition in pooled samples". In: *Bioinformatics* 38.7 (Apr. 2022), pp. 1809–1815. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btac047.

[107] Nicolas L. Bray et al. "Near-optimal probabilistic RNA-seq quantification". en. In: *Nature Biotechnology* 34.5 (May 2016). Number: 5 Publisher: Nature Publishing Group, pp. 525–527. ISSN: 1546-1696. DOI: 10.1038/nbt.3519.

[108] Jasmijn A. Baaijens et al. "Variant abundance estimation for SARS-CoV-2 in wastewater using RNA-Seq quantification". In: *medRxiv* (Sept. 2021), p. 2021.08.31.21262938. DOI: 10.1101/2021.08.31.21262938.

[109] *SARS-CoV-2 variant quantification using kallisto code*. en. Comp. software. Publisher: 4TU.ResearchData. Jan. 2022. DOI: 10.4121/18532973.v1.

[110] Nash D. Rochman et al. "Ongoing global and regional adaptive evolution of SARS-CoV-2". In: *Proceedings of the National Academy of Sciences* 118.29 (July 2021). Publisher: Proceedings of the National Academy of Sciences, e2104241118. DOI: 10.1073/pnas.2104241118.

[111] Matei Anton. "Kallisto Repurposed: Using sequencing reads from the spike, nucleocapsid, and a middle region of nsp3 in the kallisto pipeline to better predict SARS-CoV-2 variants in wastewater". en. In: (2022). URL: https://repository.tudelft.nl/islandora/object/uuid%3A990e42c3-e79f-4ff9-85ff-a614554269bb (visited on 06/27/2022).

[112] Isaac Ellmen et al. *Alcov: Estimating Variant of Concern Abundance from SARS-CoV-2 Wastewater Sequencing Data*. en. Pages: 2021.06.03.21258306. June 2021. DOI: 10.1101/2021.06.03.21258306.

[113] Devon A. Gregory et al. "Monitoring SARS-CoV-2 Populations in Wastewater by Amplicon Sequencing and Using the Novel Program SAM Refiner". en. In: *Viruses* 13.8 (Aug. 2021). Number: 8 Publisher: Multidisciplinary Digital Publishing Institute, p. 1647. ISSN: 1999-4915. DOI: 10.3390/v13081647.

[114] Lenore Pipes et al. *Estimating the relative proportions of SARS-CoV-2 strains from wastewater samples*. en. Pages: 2022.01.13.22269236. Jan. 2022. DOI: 10.1101/2022.01.13.22269236.

[115] A. P. Dempster, N. M. Laird and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1 (1977). Publisher: [Royal Statistical Society, Wiley], pp. 1–38. ISSN: 0035-9246. URL: https://www.jstor.org/stable/2984875 (visited on 10/17/2022).

[116] *PoonLab/gromstole*. original-date: 2021-09-08T02:27:03Z. Mar. 2022. URL: https://github.com/PoonLab/gromstole (visited on 10/06/2022).

[117] Arnaud N'Guessan et al. *Detection of prevalent SARS-CoV-2 variant lineages in wastewater and clinical sequences from cities in Québec, Canada*. en. Pages: 2022.02.01.22270170. Feb. 2022. DOI: 10.1101/2022.02.01.22270170.

[118] Nikolaos Pechlivanis et al. "Detecting SARS-CoV-2 lineages and mutational load in municipal wastewater and a use-case in the metropolitan area of Thessaloniki, Greece". en. In: *Scientific Reports* 12.1 (Feb. 2022). Number: 1 Publisher: Nature Publishing Group, p. 2659. ISSN: 2045-2322. DOI: 10.1038/s41598-022-06625-6.

[119] Felix Krueger et al. *FelixKrueger/TrimGalore: v0.6.7 - DOI via Zenodo*. July 2021. DOI: 10.5281/zenodo.5127899.

[120] Heng Li et al. "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16 (Aug. 2009), pp. 2078–2079. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp352.

[121] *BEDTools: a flexible suite of utilities for comparing genomic features | Bioinformatics | Oxford Academic*. URL: https://academic.oup.com/bioinformatics/article/26/6/841/244688?login=false (visited on 10/06/2022).

[122] *Picard Tools - By Broad Institute*. URL: http://broadinstitute.github.io/picard/ (visited on 10/06/2022).

[123] *Full article: A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff*. URL: https://www.tandfonline.com/doi/full/10.4161/fly.19695 (visited on 10/06/2022).

[124] *outbreak.info SARS-CoV-2 data explorer*. en. URL: https://outbreak.info/ (visited on 10/06/2022).

[125] Vic-Fabienne Schumann et al. *COVID-19 infection dynamics revealed by SARS-CoV-2 wastewater sequencing analysis and deconvolution*. en. Tech. rep. Type: article. medRxiv, Dec. 2021, p. 2021.11.30.21266952. DOI: 10.1101/2021.11.30.21266952.

[126] Shifu Chen et al. *fastp: an ultra-fast all-in-one FASTQ preprocessor*. en. Pages: 274100 Section: New Results. Apr. 2018. DOI: 10.1101/274100.

[127] *MultiQC: summarize analysis results for multiple tools and samples in a single report | Bioinformatics | Oxford Academic*. URL: https://academic.oup.com/bioinformatics/article/32/19/3047/2196507 (visited on 09/01/2022).

[128] William McLaren et al. "The Ensembl Variant Effect Predictor". In: *Genome Biology* 17.1 (June 2016), p. 122. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0974-4.

[129] Derrick E. Wood and Steven L. Salzberg. "Kraken: ultrafast metagenomic sequence classification using exact alignments". In: *Genome Biology* 15.3 (Mar. 2014), R46. ISSN: 1474-760X. DOI: 10.1186/gb-2014-15-3-r46.

[130] Derrick E. Wood, Jennifer Lu and Ben Langmead. "Improved metagenomic analysis with Kraken 2". en. In: *Genome Biology* 20.1 (Nov. 2019), p. 257. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1891-0.

[131] Jennifer Lu and Steven L. Salzberg. "Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2". en. In: *Microbiome* 8.1 (Aug. 2020), p. 124. ISSN: 2049-2618. DOI: 10.1186/s40168-020-00900-2.

[132] Susana Posada-Céspedes et al. "V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput data". In: *Bioinformatics* 37.12 (June 2021), pp. 1673–1680. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab015.

[133] *Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*. URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (visited on 09/01/2022).

[134] *PRINSEQ @ SourceForge.net*. URL: https://prinseq.sourceforge.net/ (visited on 10/18/2022).

[135] *VICUNA*. en. July 2012. URL: https://www.broadinstitute.org/viral-genomics/vicuna (visited on 10/18/2022).

[136] *ngshmmalign*. original-date: 2016-03-11T09:59:24Z. Oct. 2021. URL: https://github.com/cbg-ethz/ngshmmalign (visited on 10/18/2022).

[137] *Shorah*. URL: http://cbg-ethz.github.io/shorah/ (visited on 10/18/2022).

[138] Ray Izquierdo-Lara et al. "Monitoring SARS-CoV-2 Circulation and Diversity through Community Wastewater Sequencing, the Netherlands and Belgium - Volume 27, Number 5—May 2021 - Emerging Infectious Diseases journal - CDC". en-us. In: (). DOI: 10.3201/eid2705.204410.

[139] Ryan Wick. *Table of contents*. original-date: 2017-02-13T04:20:00Z. Oct. 2022. URL: https://github.com/rrwick/Porechop (visited on 10/18/2022).

[140] Marcel Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". en. In: *EMBnet.journal* 17.1 (May 2011). Number: 1, pp. 10–12. ISSN: 2226-6089. DOI: 10.14806/ej.17.1.200.

[141] Konstantin Okonechnikov et al. "Unipro UGENE: a unified bioinformatics toolkit". eng. In: *Bioinformatics (Oxford, England)* 28.8 (Apr. 2012), pp. 1166–1167. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts091.

[142] *MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability | Molecular Biology and Evolution | Oxford Academic*. URL: https://academic.oup.com/mbe/article/30/4/772/1073398 (visited on 10/18/2022).

[143] *FigTree*. URL: http://tree.bio.ed.ac.uk/software/figtree/ (visited on 10/18/2022).

[144] *Nextclade*. original-date: 2020-06-12T13:01:14Z. Oct. 2022. URL: https://github.com/nextstrain/nextclade (visited on 10/18/2022).

[145] joshuailevy et al. *andersen-lab/Freyja: 1.3.7*. 2022. DOI: 10.5281/zenodo.6585068.

[146] Jocelyn Solis-Moreira. "Study: New Wastewater Surveillance Method Detected SARS-CoV-2 Variants of Concern Up to 2 Weeks Before Clinical Tests". In: *JAMA* 328.10 (Sept. 2022), pp. 914–915. ISSN: 0098-7484. DOI: 10.1001/jama.2022.12563.

[147] Center for Food Safety and Applied Nutrition. "Wastewater Surveillance for SARS-CoV-2 Variants". en. In: *FDA* (Jan. 2022). Publisher: FDA. URL: https://www.fda.gov/food/whole-genome-sequencing-wgs-program/wastewater-surveillance-sars-cov-2-variants (visited on 09/04/2022).

[148] Smruthi Karthikeyan et al. *Wastewater sequencing uncovers early, cryptic SARS-CoV-2 variant transmission*. en. Pages: 2021.12.21.21268143. Apr. 2022. DOI: 10.1101/2021.12.21.21268143.

[149] *JBC and Defra: Wastewater monitoring of SARS-CoV-2 variants in England: demonstration case study for Bristol (December 2020 to March 2021), 8 April 2021*. en. URL: https://www.gov.uk/government/publications/jbc-and-defra-wastewater-monitoring-of-sars-cov-2-variants-in-england-demonstration-case-study-for-bristol-december-2020-to-march-2021-8-april-20 (visited on 11/02/2022).

[150] Devon A. Gregory et al. "Genetic diversity and evolutionary convergence of cryptic SARS- CoV-2 lineages detected via wastewater sequencing". en. In: *PLOS Pathogens* 18.10 (2022). Publisher: Public Library of Science, e1010636. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1010636.

[151] Hayley D. Yaglom et al. "One health genomic surveillance and response to a university-based outbreak of the SARS-CoV-2 Delta AY.25 lineage, Arizona, 2021". en. In: *PLOS ONE* 17.10 (2022). Publisher: Public Library of Science, e0272830. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0272830.

[152] *Surveillance of SARS-CoV-2 genomic variants in wastewater*. en. URL: https://bsse.ethz.ch/cbg/research/computational-virology/sarscov2-variants-wastewater-surveillance.html (visited on 11/01/2022).

[153] Etienne Simon-Loriere and Edward C. Holmes. "Why do RNA viruses recombine?" en. In: *Nature Reviews Microbiology* 9.8 (Aug. 2011). Number: 8 Publisher: Nature Publishing Group, pp. 617–626. ISSN: 1740-1534. DOI: 10.1038/nrmicro2614.

[154] *ENA Browser*. URL: https://www.ebi.ac.uk/ena/browser/home (visited on 10/31/2022).

[155] Alexander Crits-Christoph et al. "Genome Sequencing of Sewage Detects Regionally Prevalent SARS-CoV-2 Variants". In: *mBio* 12.1 (). Publisher: American Society for Microbiology, e02703–20. DOI: 10.1128/mBio.02703-20.

[156] Chao Fang et al. "Assessment of the cPAS-based BGISEQ-500 platform for metagenomic sequencing". In: *GigaScience* 7.3 (Mar. 2018), gix133. ISSN: 2047-217X. DOI: 10.1093/gigascience/gix133.

[157] Junhua Rao et al. "Performance of copy number variants detection based on whole-genome sequencing by DNBSEQ platforms". en. In: *BMC Bioinformatics* 21.1 (Nov. 2020), p. 518. ISSN: 1471-2105. DOI: 10.1186/s12859-020-03859-x.

[158] Michael L. Waskom. "seaborn: statistical data visualization". en. In: *Journal of Open Source Software* 6.60 (Apr. 2021), p. 3021. ISSN: 2475-9066. DOI: 10.21105/joss.03021.

[159] John D. Hunter. "Matplotlib: A 2D Graphics Environment". In: *Computing in Science & Engineering* 9.3 (May 2007). Conference Name: Computing in Science & Engineering, pp. 90–95. ISSN: 1558-366X. DOI: 10.1109/MCSE.2007.55.

[160] *Welcome to Planemo's documentation! — Planemo 0.74.11 documentation*. URL: https://planemo.readthedocs.io/en/latest/ (visited on 10/20/2022).

[161] Brian D. Ondov, Nicholas H. Bergman and Adam M. Phillippy. "Interactive metagenomic visualization in a Web browser". In: *BMC Bioinformatics* 12.1 (Sept. 2011), p. 385. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-385.

[162] Gianmauro Cuccuru et al. "Orione, a web-based framework for NGS analysis in microbiology". In: *Bioinformatics* 30.13 (July 2014), pp. 1928–1929. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu135.

[163] *ww_benchmark/samples at main · suskraem/ww_benchmark*. en. URL: https://github.com/suskraem/ww_benchmark (visited on 10/20/2022).

[164] *Cov-Lineages*. URL: https://cov-lineages.org/index.html (visited on 10/22/2022).

[165] The pandas development team. *pandas-dev/pandas: Pandas*. Oct. 2022. DOI: 10.5281/zenodo.7223478.

[166] *Plotly*. URL: https://plotly.com/python/ (visited on 10/22/2022).

[167] Luke S. Hillary et al. "Monitoring SARS-CoV-2 in municipal wastewater to evaluate the success of lockdown measures for controlling COVID-19 in the UK". In: *Water Research* 200 (July 2021), p. 117214. ISSN: 0043-1354. DOI: 10.1016/j.watres.2021.117214.

[168] *Galaxy Training: Use Jupyter notebooks in Galaxy*. en. URL: https://training.galaxyproject.org/archive/2021-12-01/topics/galaxy-interface/tutorials/galaxy-intro-jupyter/tutorial.html (visited on 11/01/2022).

[169] Bérénice Batut et al. "Community-Driven Data Analysis Training for Biology". English. In: *Cell Systems* 6.6 (June 2018). Publisher: Elsevier, 752–758.e1. ISSN: 2405-4712. DOI: 10.1016/j.cels.2018.05.012.

Acronyms

BAM Binary Alignment Map.

BED tab-delimited text file that defines a feature track.

COVID-19 Coronavirus disease 2019.

CSV comma-separated values.

FAIR Findability, Accessibility, Interoperability, and Reusability foundational principles.

FASTA text-based format for representing either nucleotide sequences or peptide sequences, in which base pairs or amino acids are represented using single-letter codes.

FASTQ text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores.

HTML HyperText Markup Language.

Illumina Illumina-based technology.

JSON JavaScript Object Notation.

LCA lowest common ancestor.

MAF mutation annotation format.

MSA multiple sequence alignment.

Nanopore Nanopore sequencing technique.

NGS next-generation sequencing.

OLS Ordinary least squares.

ONT Oxford Nanopore Technologies.

PCR polymerase chain reaction.

PE paired-end.

SAM Sequence Alignment Map.

SARS-CoV-2 severe acute raspiratory syndrome coronavirus 2.

SE single-end.

SNV single nucleotide variant.

SNVs single nucleotide variants.

TSV tab-separated values.

VCF variant calling format.

VEP Variant Effect Predictor.

VOC variant of concern.

VOCs variants of concern.

VOIs variants of interest.

WBE wastewater based epidemiology.

WGS whole genome sequencing.

WHO World Health Organization.

YAML yet another markup language.