MASTER'S THESIS

SARS-COV-2 WHOLE GENOME SEQUENCING DATA ANALYSIS FOR NATIONAL VIRAL MONITORING IN PUBLIC HEALTH.

Author: Pablo Mata Aroco

External supervisor: Isabel Cuesta de la Plaza

Academic tutor: Joaquin Giner Lamia

Submitted: May 2023

ÍNDEX

Abstract.

Relecov is an ongoing spanish project aimed to provide genomic surveillance for SARS-CoV-2 on a national scale. In this work, 2659 samples provided by two laboratories from the Relecov network were analysed using a modified version of a WGS analysis pipeline called Viralrecon. The metadata associated with the samples was processed according to the FAIR principles and uploaded together with the results of the analysis into Relecov platform's database. To achieve this, the Relecov-tools package was provided with several implementations and troubleshooting. The results from Viralrecon showed a major incidence of Lineage B.1.17 among the samples (65.11%), which concurs with the ones found in public health reports at the time the samples were collected. Most of the mutations affected orf1ab and gene S with a 46.77% and a 25.78% of the total respectively. Inside gene S, most mutations were non-synonymous (95.95%) with an acute predominance of missense variants (79.99%). The domain that presented the highest number of variants was the S1 subunit (70.18%) and inside it, the NTD and RBD sub-domains (20.66% and 13.62%). The consensus sequences obtained with the analysis were used along with the ones from a batch of 2827 samples extracted from in Relecov's database to perform a phylogeny analysis. The predicted substitution rate matches with the one described in the literature. Thanks to the metadata with each sample that gave geographical resolution, the evolution of the virus could be observed on a regional level inside Spain, which ultimately matched with the public health reports at the time of collection of the samples.

Resumen.

Relecov es un proyecto español en proceso que tiene como objetivo proporcionar vigilancia genómica para el virus SARS-CoV-2 a nivel nacional. En este trabajo, se analizaron 2659 muestras proporcionadas por dos laboratorios de la red Relecov utilizando una versión modificada de una plataforma de análisis de secuenciación masiva del genoma (WGS) llamada Viralrecon. Los metadatos asociados con las muestras se procesaron siguiendo los principios FAIR y fueron subidos junto con los resultados del análisis a la base de datos de la plataforma Relecov. Para lograr esto, el paquete de herramientas Relecov-tools fue surtido de varias implementaciones y soluciones de errores. Los resultados de Viralrecon mostraron una incidencia significativa del linaje B.1.17 entre las muestras (65.11%), lo cual concuerda con los informes de salud pública en el momento de la recolección de las muestras. La mayoría de las mutaciones afectaron a los genes orf1ab y S, con un 46.77% y un 25.78% del total, respectivamente. Dentro del gen S, la mayoría de las mutaciones descubiertas fueron no-sinónimas (95.95%), con una predominancia de variantes de cambio de sentido o missense (79.99%). El dominio que presentó el mayor número de variantes fue la subunidad S1 (70.18%), y dentro de ella, los subdominios NTD y RBD (20.66% y 13.62%, respectivamente). Las secuencias de consenso obtenidas en el análisis se utilizaron junto con las de un lote de 2827 muestras extraídas de la base de datos de Relecov para realizar un análisis filogenético. La tasa de sustitución predicha coincide con la descrita en la literatura. Gracias a los metadatos asociados a cada muestra que proporcionaron resolución geográfica, se pudo observar la evolución del virus a nivel regional dentro de España, la cual concuerda con los informes de salud pública en el momento de la recolección de las muestras.

Introduction

Genomic surveillance of SARS-CoV-2.

The SARS-CoV-2 outbreak, also known as the COVID-19 pandemic, has been one of the most significant global health crises of the 21st century. Originating in Wuhan, China in late 2019 the outbreak was officially declared a pandemic on March 11, 2020 (1). In order to handle the spread of the virus, various countries introduced different measures which led to significant disruptions in daily life, causing major concern in the population (2).

Globally, as of 10:11am CEST, 12 April 2023, there have been 762.791.152 confirmed cases of COVID-19 reported to WHO across 260 countries and territories, with 6.897.025 fatalities (https://covid19.who.int/). However, As of 9 April 2023, a total of 13.337.964.733 vaccine doses have been administered and 5.100.583.862 persons are fully vaccinated, comprising 63,7% of the global population. The WHO projects that approximately 10% of the world's population (roughly 780 million individuals) have already contracted the disease, given the extensive underreporting of cases worldwide.

Genomic surveillance of SARS-CoV-2 virus, the virus responsible for COVID-19, has become a crucial tool in understanding the epidemiology and evolution of the virus. SARS-CoV-2 virus is a RNA virus with a high mutation rate, leading to the emergence of new variants over time, increasing the probability of increased transmissibility or virulence. Genomic surveillance enables us to identify these variants, determine their origin, and track their path as they spread (3). This information is essential for developing targeted public health interventions to prevent the spread of the virus and to monitor the effectiveness of vaccines against different variants (4).

SARS-CoV-2 virus, molecular insights.

SARS-CoV-2 is a positive-sense single-stranded RNA virus that belongs to the betacoronavirus genus. Its genome is approximately 30 kb in length and encodes for 27 proteins, including the spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins, among others (5).

The S protein is a type I membrane glycoprotein composed of two subunits, S1 and S2, that undergoes a conformational change upon binding to the ACE2 receptor on host cells, allowing for viral entry, reason why it is a major target for host immune responses and therapeutic interventions. The receptor-binding domain (RBD) of the S1 subunit is a critical determinant of viral infectivity. Moreover, several vaccines and monoclonal antibodies targeting the RBD or other domains of the S protein have been shown to be effective in preventing viral entry and reducing disease severity in preclinical and clinical studies (6). The S protein contains several residues that interact with the ACE2 receptor and has undergone multiple mutations during the evolution of the virus, some of which have been associated with increased transmissibility and resistance to immune responses. Recent studies have shown that the S protein of SARS-CoV-2 exhibits high structural plasticity, which allows for conformational changes that are critical for viral entry and

immune evasion. Due to its importance in the virus cycle, this gene has been the main focus of many investigations regarding infectivity and virulence of SARS-CoV-2. (7)

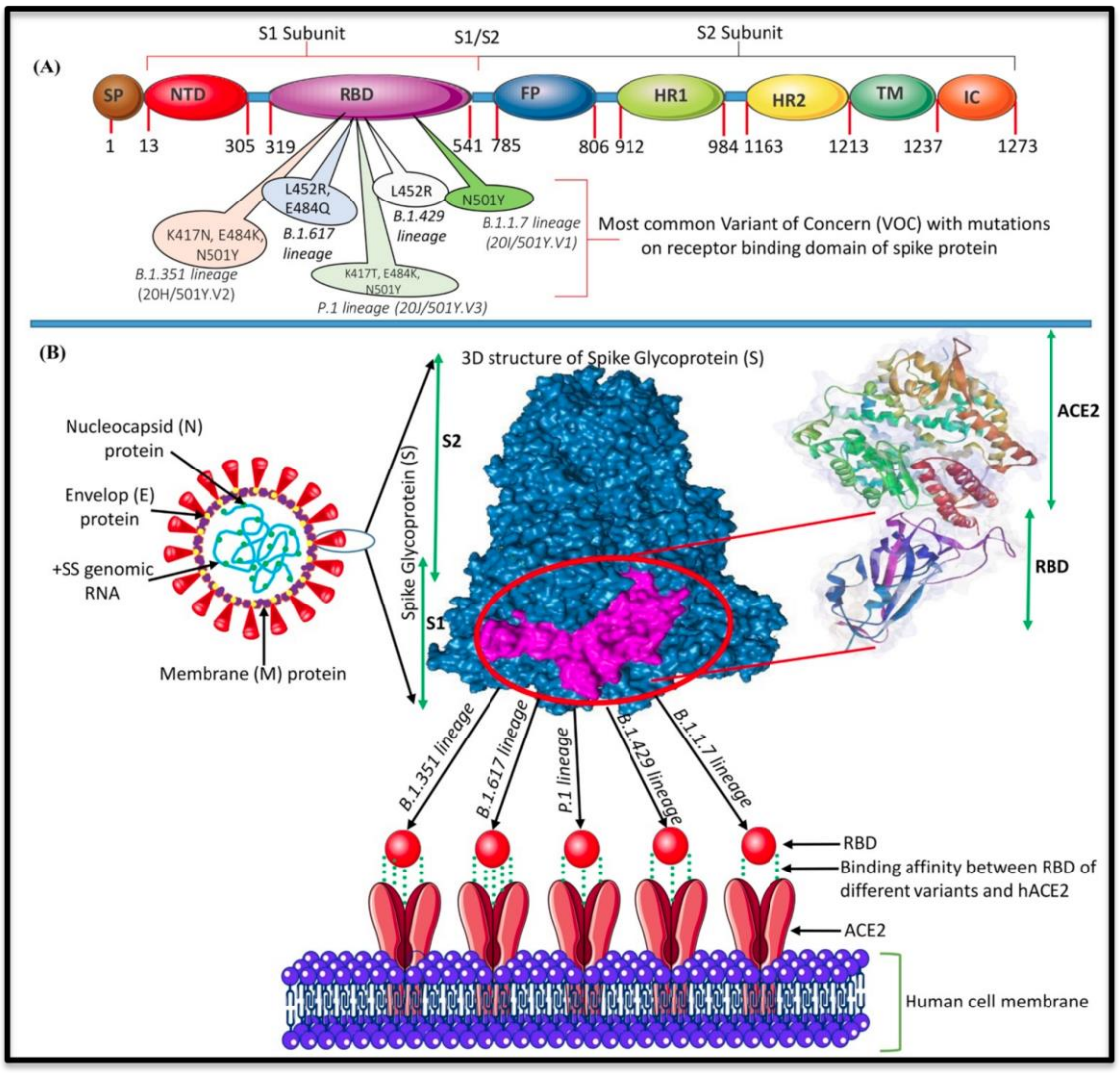

Figure 1. (A) Structure of gene S from SARS-CoV-2 showing relevant positions for mutations in the RBD complex. (B) Molecular structure of SARS-CoV-2 virus, showing the binding centre of the Spike protein and the interactions between ACE2 receptor and RBD. Source: (8).

The E protein is a small transmembrane protein that plays a role in viral assembly and release, while the M protein is the most abundant protein in the viral envelope and has a role in virus assembly and budding. On the other hand, the N protein is involved in genome packaging and plays a role in the regulation of viral transcription and replication. The genome of SARS-CoV-2 also contains other domains, such as the open reading frames (ORF) 1a and 1b, which encode for the viral replicase complex, and ORFs 3a, 6, 7a, and 8, which encode for accessory proteins that are involved in modulating the host immune response and promoting viral replication (9).

Molecular and genetic characterization of SARS-CoV-2 has revealed several mutations and variants, some of which have been associated with increased transmissibility and virulence, especially those located in the coding region of the S gene (10). The rate at which these variants can rise is determined by multiple parameters such as the mutation rate, substitution rate and recombination processes among others.

The mutation rate is the probability of a mutation in the replication process of the virus. In the case of SARS-CoV-2 it is estimated to have around 1×10^{-6} to 2×10^{-6} mutations per nucleotide per replication cycle which is consistent with previous estimates in other betacoronaviruses (11). These mutation rates lie below the range of rates that are typical for other RNA viruses such as hepatitis C virus (HCV) or the human immunodeficiency virus (HIV). Although replication errors are the most basic source of insertions and deletions, another agent capable of introducing direct mutations in the virus genome is the host-mediated genome editing that is produced by the innate cell defence mechanism. This mechanism is coordinated by multiple proteins like the APOBEC (12) family and some antiviral adenosine deaminases (13).

Recombination is another important source of diversity in RNA viruses and a common feature of SARS-COV-2 evolution and other betacoronaviruses (14). This process occurs when a host is co-infected with two viruses with enough genetic differences that can recombine to produce a viable progeny. This is the main reason why there could be multiple SARS-COV-2 divergent lineages circulating in the same region at the same time (15).

Whole genome sequencing.

Whole genome sequencing (WGS) of SARS-CoV-2 has been a critical aspect of understanding the virus and its evolution since the beginning of the pandemic as it is the main reason why the emerging variants of the virus could be detected.

Illumina and Oxford Nanopore are two different sequencing technologies that are used for WGS. Illumina is a short-read (50-300 bp) sequencing technology that uses reversible terminators to sequence millions of DNA fragments in parallel. It is fast, accurate, and relatively inexpensive, making it the most widely used technology for WGS. On the other hand, Oxford Nanopore is a long-read sequencing technology that uses nanopores to directly read the DNA sequence as it passes through a membrane. It is slower and more expensive than Illumina sequencing, but it can generate reads that are even 2.3 mb in length, with an average of 10-30 kb (16).

Before samples can be sequenced, it is a must to prepare a library of DNA fragments by fragmentation of the DNA, end repair, adapter ligation, and size selection to ensure that the fragments are of a uniform length. This task is therefore closely related to the sequencing technology. The ARTIC (Adaptive Robust Technology for Identification of COVID-19) protocol is a sequencing-based approach for WGS of SARS-CoV-2 that includes descriptions for viral RNA extraction, reverse transcription, PCR amplification of a described set of amplicons and sequencing using Illumina technology (17). The ARTIC network has also helped to develop sequencing protocols for other emergent pathogens such as Zika or Ebola (18)

After sequencing, the raw data must be processed and analysed in order to obtain valuable information. There are a wide variety of described bioinformatics workflows for WGS analysis like the one described in Figure 2, but most of them include the following steps (19) (20) :

Data pre-processing: One of the major challenges in processing SARS-CoV-2 data is the large number of short reads generated in WGS. The first step in genomic analysis is to process the raw data generated by sequencing machines. This step includes removing low quality reads and trimming the adapter sequences, but may include other processes depending on the analysis.

Genome assembly: This process implies piecing together the short, fragmented reads generated by next-generation sequencing (NGS) technologies into a complete genome sequence. This involves aligning and overlapping the reads to construct longer contiguous sequences, known as contigs, which are further organised into scaffolds and ultimately into a final genome assembly.

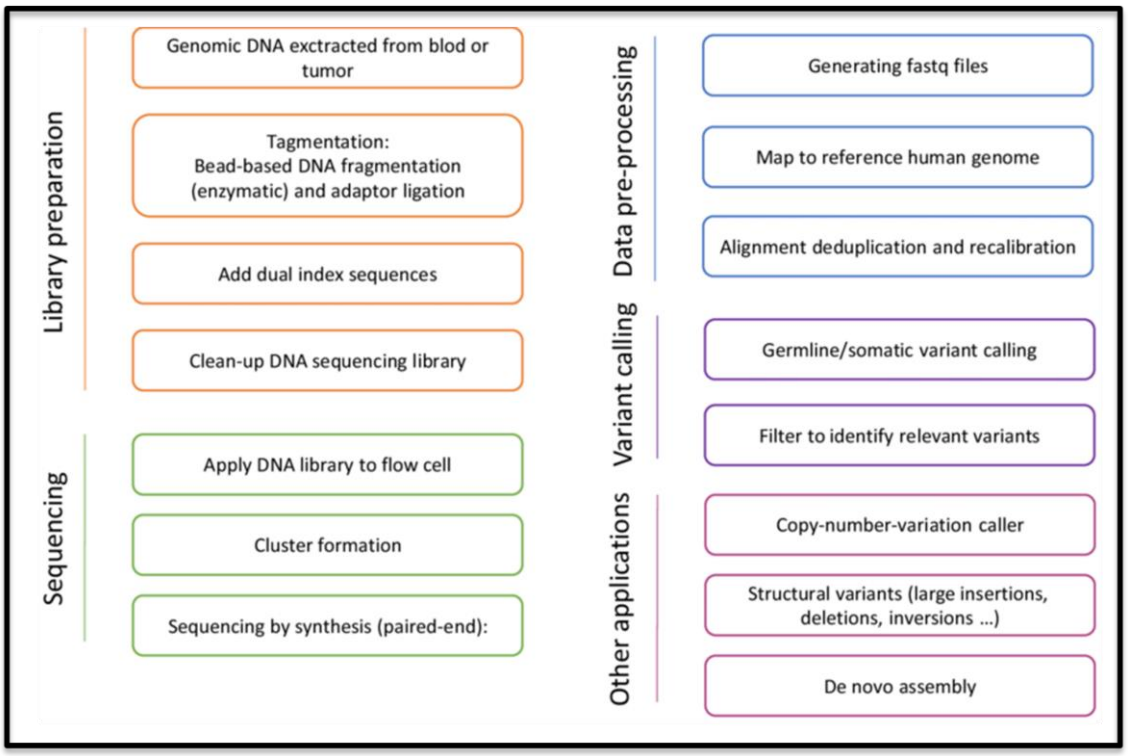

Figure 2. Whole Genome Sequencing workflow schema. Source: (21)

Variant calling: Variant calling is the process of identifying differences between the sequenced genome and the reference genome-sequenced reads are aligned to the reference genome, identifying positions where the sequence reads differ from the reference. Then, frequency and nature of the observed variations are determined (22).

Consensus sequence generation: In cases like the COVID19 pandemic, where the virus was rapidly evolving, determining the genome of a newly found variant becomes a primordial task. The reconstruction of this genome involves aligning the short reads to a reference genome or a closely related sequence, and then calling the most frequently observed nucleotide at each position to generate a final consensus sequence that accurately represents the target genome (23).

Phylogenetic analysis: Phylogenetic analysis involves reconstructing the evolutionary relationships between different genomes. With the emergence of new variants, phylogenetic analysis has become increasingly important to track the spread and

evolution of the virus. New tools like Nextstrain (24), have been developed to improve the accuracy and speed of this process.

FAIR principles.

The FAIR principles are a set of guidelines developed by Wilkinson MD et al. in 2016 (25) to promote the accessibility, interoperability, and reuse of scientific data. FAIR is an acronym that stands for Findable, Accessible, Interoperable, and Reusable. These principles have been widely adopted by the scientific community as a way to make data more discoverable, understandable, and useful.

The first of the FAIR principles is Findable, which means that data should be easy to find and identify. To make data findable, it is important to assign unique and persistent identifiers, such as DOIs or URIs, to each dataset. These identifiers should be searchable using appropriate metadata, such as author name, date, or subject and should be stored or indexed in a repository that can be accessed by public domain (26). There are a lot of public domain repositories for biological data that are findable, such as GenBank's Sequence Read Archive (SRA: https://www.ncbi.nlm.nih.gov/sra) or the European Nucleotide Archive (ENA: http://www.ebi.ac.uk/ena/).

The second FAIR principle is Accessible. This means that data should be openly accessible to all, without any barriers to entry. To make genomic data accessible, these open-access repositories mentioned for the previous principle should allow complete access to all the data and metadata and provide long-term storage and preservation of it (27). By doing so, researchers can access and analyse the data without restrictions, promoting transparency and reproducibility of scientific findings.

The third principle of FAIR is Interoperable. This means that data should be able to be integrated and used with other datasets. To make data interoperable, it is important to use common data standards and formats like FASTQ, BAM, or VCF which are well known and overly managed by the scientific community. Regarding genomic data, the samples should also be linked to other relevant datasets, such as genomic annotations, gene expression data, or clinical data, using stable identifiers like the ones provided by HGNC or ENSEMBL. Not only that, but all the metadata should be properly labelled and described using standard vocabulary and terminology, which can be made using ontology-based denominations (28) .

The fourth principle of FAIR is Reusable. This means that the data should be able to be used and cited for future research. To achieve this, the data should include information regarding the source of the data, its characteristics and any other standardised details that can be relevant for the reproducibility of the experiment or the use of the data in a different task. Licences are also important in this case to clarify the terms of use and attribution of the data (29).

The FAIR principles help to address the challenges associated with new discoveries and technologies by providing a common framework for organising, sharing, and structuring genomic data.

RELECOV Project.

The *Red de Laboratorios de secuenciación de SARS-CoV-2* or RELECOV is a Spanish network that was created in January 2021 aiming to provide the necessary information for genomic surveillance of SARS-CoV-2 variants on a national level. This project is part of the Human and Environmental Risk Assessment (HERA) European initiative from the European Centre for Disease Control and Prevention (ECDC). Through this network, the Spanish laboratories are interconnected with each other and with the entities responsible for COVID-19 surveillance at the regional and national level. RELECOV is composed of at least one laboratory per autonomous community, and there are also associated laboratories that work within the regional networks under the coordination of the Spanish autonomous Public Health authorities (Source: Estructura y funcionamiento de la red RELECOV).

One of the main goals of the project is to create a data repository where sequenced samples along with its metadata are stored together and shared among the laboratories.

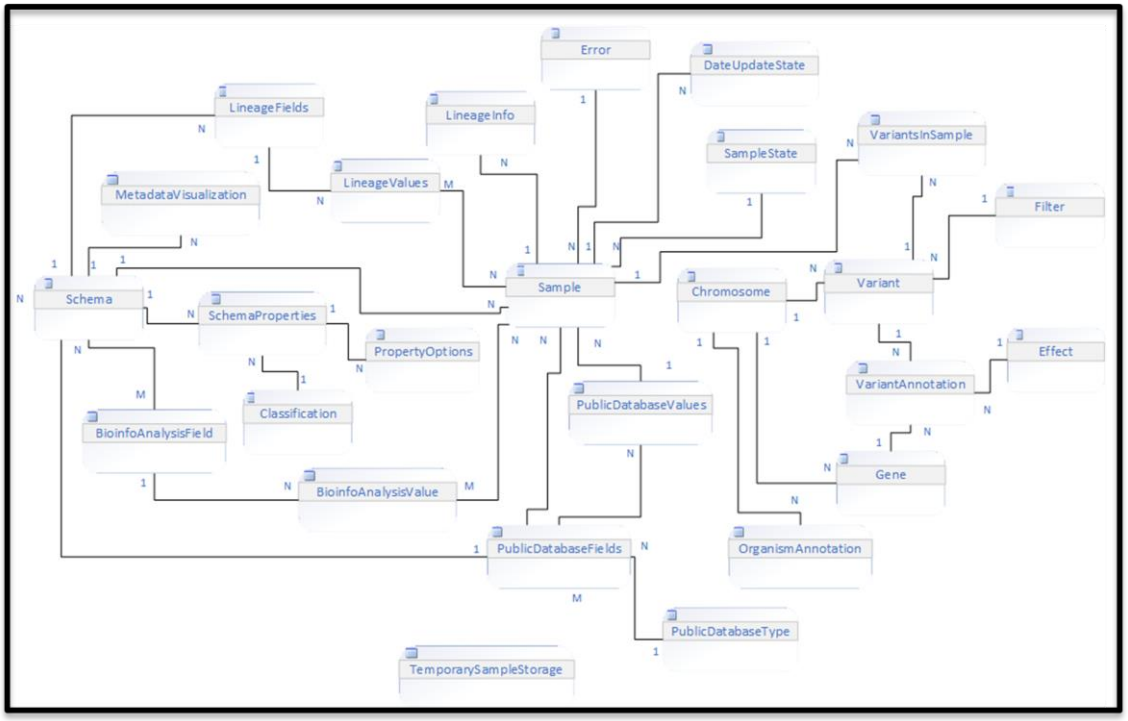

Figure 3. Graphical representation of Relecov platform's database.

RELECOV's two main software components are relecov-platform and relecov-tools. Relecov-platform (https://github.com/BU-ISCIII/relecov-platform) is the foundation of the webpage and the different services that are provided to the public domain, such as statistics regarding variants and lineages found in the multiple samples that are uploaded into its database.

Relecov's web platform holds a structured database to store all the data that is collected in the RELECOV project. In order to comply with the FAIR principle of Reusability, each sample has associated metadata with relevant characteristics of the data such as sequencing date, library preparation kit, enrichment protocol, enrichment panel, sequencing instrument model, authors, and information regarding its source.

Relecov-tools (https://github.com/BU-ISCIII/relecov-tools) is a toolkit software built in python that aims to provide help to the different laboratories that want to upload samples to Relecov's platform. In order to comply with the FAIR principles, all the data and metadata related to the samples must conform to a well-defined structure that can be easily accessed by computational systems with none or minimal human intervention. Relecov-tools include a lot of functions that aim to automatize the conversion of the metadata from each sample into a standardised ontology-based structure that can be uploaded to the platform's database.

A)
```
"virus identifier": {
    "examples": [
      ""
    ],
    "ontology": "GENEPIO:0001123",
    "type": "string",
    "description": "Unique laboratory identifier assigned to the virus by the investigator.
    "clasification": "Database Identifiers",
```

B)
```
"collecting_lab_sample_id": {
    "examples": [
        "prov_rona_99"
    ],
    "ontology": "GENEPIO:0001123",
    "type": "string",
    "description": "The name given for the sample by the collecting institution.",
    "classification": "Database Identifiers",
```

Figure 4. Same field with different names in ENA's schema (A) and Relecov's Schema (B), mapped to the same ontology identifier so its value is kept the same for both databases.

As this package is oriented to any kind of researcher, the tools must be not only user-friendly, so it can be used regardless of the programming skills, but robust enough to avoid any introduction of irregular data into the database. For this purpose, the package includes JSON files that hold information regarding each possible field that can go into the metadata into a schema-like structure, with key-value pairs. Each field is mapped to an ontology in order to have a standardised metadata that complies with the FAIR principle of Interoperability (Fig 4). Through the use of these identifiers, if a machine automated process accesses a different database that has this associated ontology value, it will be recognized for what it means by the machine.

Materials and Methods.

Sample source.

The following analyses were performed over two batches of 2658 and 2847 samples respectively, which were collected from January 2020 to January 2022. The samples were provided by two laboratories from the RELECOV network: CNM-ISCIII (https://www.isciii.es/) and FIBAO (https://www.fibao.es/). These laboratories provide the bioinformatic analyses of the sequences that are in turn provided by multiple hospitals and other sequencing centres in the respective region.

Metadata processing

The metadata was provided by both mentioned laboratories in an excel format, following a closed schema provided by BU-ISCIII's unit. Relecov-tools package was used to process this metadata in order to make it suitable to be uploaded to Relecov's database. Relecov-tools is written in Python3 programming language, and therefore, all the additions to the code were made using this language.

Whole genome sequencing analysis

The genomic analysis pipeline that was utilised in order to obtain lineage information from the samples is called **Viralrecon**.

Viralrecon (https://github.com/nf-core/viralrecon) is a bioinformatics analysis pipeline used to perform assembly and intra-host/low-frequency variant calling for viral samples. It's been optimised to perform variant and lineage analysis on SARS-COV-2 but it has demonstrated excellent performance with Monkeypox Virus and Respiratory Syncytial Virus (RSV). Viralrecon is built on top of Nextflow (30), an open-source framework for creating data-intensive workflows that can be easily scaled and executed on a variety of computing infrastructures. One of the key features of Nextflow is its Domain Specific Language (DSL), which allows users to define complex workflows using a simple, easy-to-read syntax. Viralrecon uses DSL2, the second version of Nextflow's DSL, which builds on this foundation by adding new features and improvements that make it even more powerful and flexible such as improved error handling and debugging, better support for parallel execution, and more advanced scheduling and prioritisation algorithms. These features make it easier to create workflows that are reliable, efficient, and scalable, even when dealing with very large datasets or complex computational tasks.

Viralrecon (31) was originally written by members from the Bioinformatics Unit of Carlos III Health Institute (BU-ISCIII) from Spain, with the collaboration of the nf-core community (32), which now keeps it constantly updated. The pipeline has the capability to analyse both Illumina and Nanopore sequencing data. When processing Illumina short reads, the pipeline can handle metagenomics data obtained from shotgun sequencing. This includes data obtained directly from clinical samples, as well as enrichment-based library preparation methods such as amplicon-based techniques like the ARTIC SARS-CoV-2 enrichment protocol or probe-capture-based methods. For Nanopore data, the pipeline is exclusively designed to support amplicon-based analysis that is obtained from primer sets created and maintained by the ARTIC Network. By leveraging Docker/Singularity (33) (34) containers in Nextflow, installation is straightforward, and reproducibility of results is ensured.

Furthermore, the Nextflow DSL2 implementation of this pipeline utilises one container per process, which simplifies software dependency maintenance and updating.

The pipeline consists of multiple steps and it's flexible enough to choose between different tools for certain steps as for variant calling (Fig. 5). Therefore, the output varies depending on the workflow that is selected. All the steps of the Illumina workflow are explained in Supplementary materials S2..

For the purpose of this work, the analysis was performed using only the Illumina workflow for amplicon sequencing data (Fig. 5A), although Viralrecon also supports the usage of Nanopore reads as already mentioned. If the input were nanopore reads the pipeline would be slightly different, specifically in the early stages of the analysis, but most of the tools used with amplicon data remain the same (Fig. 5B).

Software versions used for each process of the analysis are presented in Supplementary materials S1.

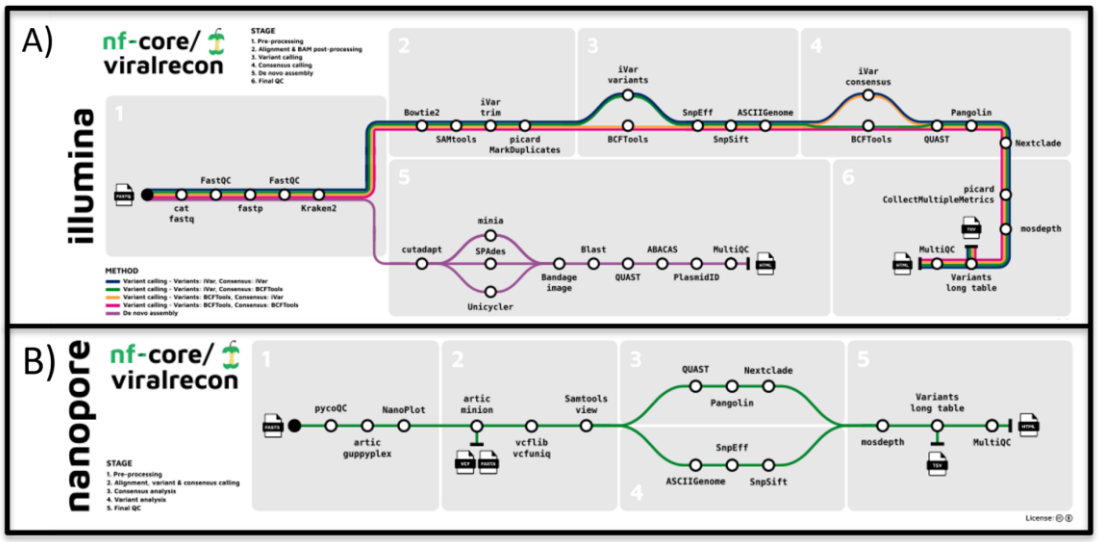

Figure 5. Viralrecon workflow for either Illumina (A) or nanopore data (B). Source: https://nf-co.re/viralrecon/2.6.0.

Analysis automation.

Buisciii-Tools (https://github.com/BU-ISCIII/buisciii-tools) is an auxiliary set of helper tools that can be used to manage the complementary processes that must be carried along with the execution of the different pipelines in the Bioinformatics Unit (BU-ISCIII). These tools are focused on the management of the files that are generated when running these kinds of pipelines. Buisciii-Tools were used to extract the most relevant data from the output that is generated at each of the different stages of the pipeline and to merge them into several tabular files, which are later used as input to relecov-tools in order to upload the data into the RELECOV platform. Only minor implementations were needed in this package, by adding a new column and changing the header of one of the resulting tables.

WGS data processing and visualisation.

The results from viralrecon are mostly condensed in three custom files with tabular format which are created using scripts from the buisciii-tools package.

1. *Long-table.csv* stores all the variant calling information in a table with 19 columns and one row for each variant that was found during the analysis. Each column is assigned to one relevant parameter of the variant calling analysis, including reference sequence, alternative sequence, position, Depth, Coverage, Allele frequency, gene and effect.

2. *Summary_metrics_mqc.csv* summarises the results from MultiQC. Each row is a different sample containing quality metrics such as number of input, trimmed and mapped reads, number of non-host reads, median coverage, number of SNPs and INDELs, number of missense variants or Ns in consensus sequence
3. *Mapping_stats.csv* is a compendium of the most relevant information from both previously mentioned tables but applying a filter to the results from MultiQC to keep only the sequences with a depth superior to 10x. It also incorporates additional information such as reference sequence name, pangolin lineage and nextclade clade.

To filter out low-frequency mutations, the data in the *long_table.csv* was processed to keep only the mutations with an allele frequency superior to 0.75 (75%) as these are the ones included in the consensus sequence. In order to analyse the effects of the mutations in gene S, the specific genomic positions for each domain of the protein S were extracted from (35).

The resulting consensus sequences from Viralrecon are used later to carry out a phylogenetic analysis using Nextstrain (24).

Computation and resource management.

The workflow was executed using the resources from XTutatis, the HTC cluster (High Throughput Computing) computing service of the Carlos III Health Institute (ISCIII). XTutatis is a heterogeneous cluster using SLURM job scheduling system (36). It has:

- Number of nodes: 34
- Total cores/CPUs: 768
- Total memory: ~12.6 TB
- GPGPU: 4 x NVIDIA Tesla P100
- Storage connection: 10 Gbps Ethernet
- Main storage: 100 TB
- Internal node storage: ~800 GB.

Software availability.

The custom scripts used for the analysis and visualisation of the results can be found in https://github.com/Shettland/TFM_Biocomp/, along with configuration files used for Viralrecon and Nextstrain.

All the contributions to the packages were made to the corresponding GitHub repositories (relecov-tools and buisciii-tools), attributed to the user Shettland in this case. Basic contribution guidelines were followed, based on forks, pull requests, and collaborative software reviewing before rebasing.

Results and Discussion.

Relecov-tools implementations.

Although the relecov-tools package was already built and tested in a manually made data model, it failed multiple times on the task of managing the real metadata that was provided from labs inside the Relecov network. Therefore, it was necessary to make a large number of changes to the code. The aim of these changes was, not only to adapt the package to the metadata that came from real laboratories, which included certain irregularities not present in the fictitious data with which it was tested, but also to include multiple fixes in the code in order to prevent program failures and overcome limitations that impeded code automation and functionality.

Relecov-tools has 9 distinct modules that are functional thanks to 17 python files (including the init and main files), 4 configuration files in json format and 4 distinct schemas in json format that store standardised information and ontology values for each possible field.

The package is designed to operate with a collection of samples in either Fastq or Fasta format, accompanied by a file with tabular structure that contains metadata for each sample. To ease the metadata file completion process for the researchers in the RELECOV network, an Excel format was selected.

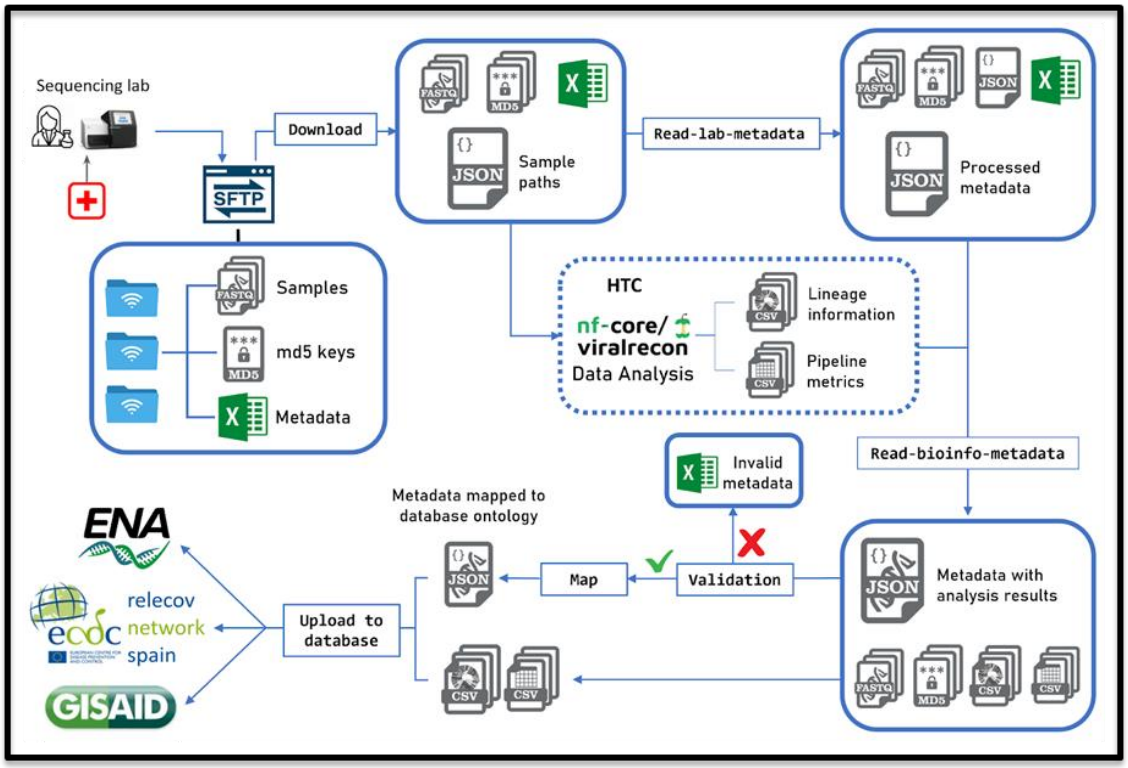

Figure 6. Relecov-tools workflow.

Module 01: download

The *download* module establishes a connection to a transfer protocol, in this case an sftp, to retrieve files within the available folders in a private server using the supplied credentials. Using this module, the platform's developer team can download the samples uploaded to the remote server and store them in a secure data repository. Moreover, the module also performs a comparison check to ensure that the downloaded files correspond to the files listed in the metadata file and verifies the md5sum for each file. In cases where the md5sum is missing, the command generates one before storing the file in the designated repository.

If no configuration file is specified, the command retrieves default values from the *conf/configuration.json* file and prompts the user for their username and password. Users can customise default values by passing a YAML configuration file, allowing them to set parameters such as the username, password, SFTP server, etc.

The *download* module presented many issues:

- The output location for the downloaded files was pre-configured in the config file to */tmp* and had to be changed anytime the user wanted a different location. Furthermore, it caused trouble with the dedicated memory space of the */tmp* folder. As a means to make the process more flexible and to encourage automatization, a new option that could be used to select the output location was introduced.
- The module included a validation of the downloaded files by looking at the metadata excel file within the remote folder and checking if the file names given in the corresponding columns were correct. This slowed down the process considerably, because if there was any discordancy with the given metadata, it would stop after downloading all the files. In order to solve this issue, we designed a new validation process which downloaded the excel file first and checked the files in the remote folder before proceeding with the download.
- The former design downloaded everything in the remote server without any possible user option. Also, it removed every local file in the output location if an error was met, making it impossible to download anything if there were any issues with any of the folders. The re-designed module included a process to select the desired folders to download, giving flexibility and decision-making power to the user.
- The remote sftp server has a limited storage capacity, and there was no functionality in relecov-tools to remove data already downloaded, checked and locally stored. As the network grows, data cleaning is expected to become more and more important, so various cleaning processes were included in the download module. The process was given three possible options, to either proceed only with the download, to delete the contents of the remote folder after download is finished, or to merely delete the contents in the remote folder as a separate cleaning step. These subprocesses are fully functional with the rest of the implementations and options in the module, e.g., the user could delete the contents of a single folder using this function and the one that was previously mentioned.

- Many individual processes lacked error handling, resulting in unintelligible failures with no guidance on resolution. To enhance software robustness, extensive improvements were made regarding this matter.
- In the end, most of the code was refactored to improve performance or to make it more readable.

The *download* process also creates a json file containing information regarding each sample's associated files. This file is then used as input along with the excel metadata file to the next step, *read-lab-metadata.*

Module 02: process lab metadata

Read-lab-metadata module extracts the contents of both the user's excel file and the previously mentioned json file generated using *download* module, and merges them into a json formatted file with all the metadata, adding ontology IDs to each field based on the schema for Relecov database.

Read-lab-metadata also had several issues, including:

- Crashes if there were any empty cells in the excel metadata file. It's not always possible for the researchers to provide all the necessary fields that are asked to be fulfilled, therefore, the issue was solved by introducing the ontology field *Not Provided* (GENEPIO ontology id: 0001668) when this case was met.
- Since RELECOV's database follows the FAIR principles, all the possible metadata values had to be structured, which led to errors when new locations or sequencing centres were given by the laboratories in the metadata excel file. To handle this, there are two auxiliary json files manually created which store information regarding laboratory addresses, contacts, and geographic locations. These files were re-visited and re-structured to remove redundant information and to include new one.
- Other minor issues that were fixed: a) date detection was improved b); error handling regarding the excel header and others were improved; c) some parts of the code were refactored to improve the performance of the process and to ease the comprehension of the code.

Module 03: process bioinformatics analysis metadata

The next step in the workflow is to execute the read-bioinfo-metadata module which extracts information from viralrecon's analysis output and includes it into the previously created json file. The structure of this data is previously described in the WGS analysis section (see WGS data processing).

In the former design, the process was divided into two modules, the proper *read-bioinfo-metadata*, mentioned before, and *long-table-parse*, which transformed viralrecon's output, *variants_long_table.csv*, into json format. As both modules parsed similar outputs, it was auspicious to simplify the workflow by merging both processes into *read-bioinfo-metadata* module. Moreover, *long_table_parse.py* also received remarkable improvements in code optimization and cleanliness, as some functions were completely ad-hoc with no flexibility for future implementations.

Module 04: validation

Once the processing of the metadata is finished, it is supposedly ready to be uploaded to the database. In order to keep RELECOV's database FAIR, all the metadata is validated according to json schema specification (*Draft202012*. Source: https://json-schema.org/) to ensure that no irregular value is uploaded to the database. The validation process reads every field in the metadata json file and checks if it has a valid ontology value associated with it in the given schema in json format; in this case the relecov's schema specification. The improvements in this module were aimed to facilitate the debugging and to clarify the output when a value was not validated.

Module 05: database upload

In order to upload the data to Relecov's database, the upload-db takes the validated json-data and uploads it to the relecov-platform database using the RESTful API from Relecov platform. Similar to the previous step, only implementations in output clarification and debugging were included in this process.

Module 06: data mapping and sharing to public databases

Relecov-tools also includes modules to upload the samples and metadata to ENA or GISAID but, as these databases are also FAIR, they use their own data structures and vocabularies. Map module transforms the metadata from the json files into the corresponding field value found in ENA or GISAID schemas thanks to the ontology annotation, shared between schemas, which is tracked during the whole workflow. After the metadata is correctly mapped to the desired schema, *upload-to-ena* and *upload-to-gisaid* modules can be used to upload both samples and metadata to ENA or GISAID databases respectively.

Along with all the mentioned changes in the code, there were several changes in relation to the configuration files, including new fields and values which helped to reduce the length of the code, to limit ad-hoc processes and to improve the performance of the code overall. Not only that, but most schemas received changes in both structure and content, including missing ontologies and new labels.

Metadata overview and analysis

From the initial 2658 samples, 53 had to be removed because the files were corrupted, resulting in 2605 samples that were processed using relecov-tools and uploaded to relecov's database. Since there were already 2827 samples that were previously uploaded, this led to a total of 5432 samples in RELECOV's database.

Including both lab provided and analysis generated metadata there are 127 possible metadata fields available for each sample, 73 were generated by processing the provided metadata excel file, and 54 were introduced from viralrecon's analysis results. From the 5432 samples, only 70.87% of the fields were fully complete, this means, none of the samples had a missing value for any of those specific fields. If the fields corresponding to the results from the WGS analysis are not considered, then this number is reduced to 60.27%.

Nevertheless, the total percentage of "Not Provided" values across all samples accounts for the 24.90%, which is reduced to 20.61% when including the metadata that was incorporated from viralrecon. There's also a prominent distinction of provided metadata from the two submitting institutions as the ones from FIBAO have a 60% of completeness while the ones from CNM have an 80%.

Despite that, 100% of the minimal fields described in Ph4age that define the essential metadata for any public repository are complete (37). This is also a requirement in the validation process introduced in relecov-tools, as if any of the previously mentioned fields is missing, then the sample will not be accepted to be uploaded into relecov's database.

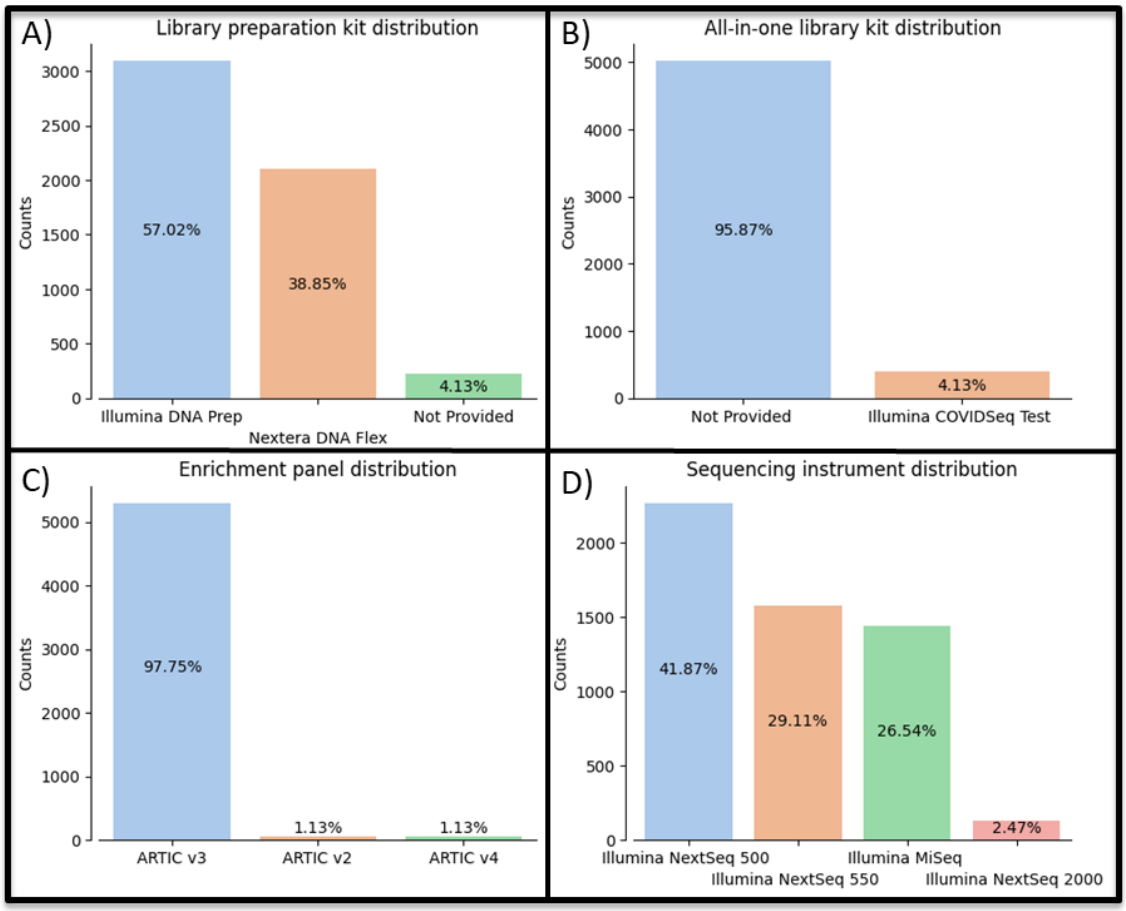

Figure 7. Distribution of some of the labels from the sample's metadata: A) Library preparation kit. B) All-in-one library kit. C) Enrichment panel version, same as amplicon kit. D) Sequencing instrument.

All the provided samples shared common values for some metadata fields:

- The source was always a viral RNA, as all the samples were from SARS-CoV-2.
- The sequences were all paired-end. This is important as it allows to get the sequence of both ends of the RNA fragment; also, thanks to the reducing costs of high-throughput sequencing, it has become a standard in WGS.
- All the samples were sequenced with an Illumina sequencer, which is probably because these have been considered the most cost-efficient in previous years (38).
- Thanks to their high sensitivity even with low viral load, the ARTIC sequencing protocol and enrichment panel have received a very extensive use for sequencing SARS-CoV-2 (39). This may explain why all the samples were sequenced using

this protocol. Furthermore, the library strategy is Amplicon-based in all the samples, probably due to the wide use of ARTIC protocol, as already mentioned.

Regarding the sequencing process of the samples, 57.02% used Illumina DNA prep as the library preparation kit of choice while 38.85% used Nextera DNA Flex, which is the previous version of the same kit, leaving a 4.13% of Not provided (Fig. 7A). Some laboratories used an all-in-one library kit instead, the Illumina COVIDSeq test, which complements the 4.13% of kits that were missing before (Fig. 7B). The vast majority of samples (97.75%) were sequenced using ARTIC v3 enrichment panel, leaving a 1.13% to both v2 and v4 versions (Fig. 7C). The time when the samples were collected might be the reason for this choice, as the third version of this amplicon set was released in March 2020 and the next version did not appear until June 2021 (40). The sequencer of choice was Illumina NextSeq 500 in 41.88% of the samples, followed by Illumina NextSeq 550 (29.11%), Illumina MiSeq (26.54%) and Illumina NextSeq 2000 (2.47%) (Fig. 7D).

Most of the basic metadata, this means, routinary metadata such as sequencing instrument, enrichment panel protocol or library preparation kit are correctly labelled and provided. However, other metadata labels such as host gender, host age, anatomical source or collection method in some cases are severely lacking values. This significantly limits the analysis possibilities that can be made using relecov's database. Certainly, this is one matter that needs to be improved as the project progresses.

Whole genome sequencing analysis

Viralrecon was already fully functional, but it still had room for new implementations. In this case, when running the initial analysis, mosdepth process crashed because some R packages could not plot heatmaps for such a large number of samples. Since there was no need to plot any depth heatmap, as it would most likely be too noisy due to the number of labels, this process could be skipped. Nonetheless, there was no option to do so in viralrecon, and a new parameter to skip this process was introduced in the code.

There are a lot of valuable metrics to ensure the quality of the analysis, but the percentage of mapping, along with the depth of coverage are the most commonly used for this task (41). To measure the quality of the consensus sequence, which will be used later for a phylogeny analysis, the studied metrics were the number of variants and the percentage of unknown nucleotides (N) in the sequence, which were extracted from the previously mentioned *mapping_stats.csv* file.

As it can be appreciated in Fig. 8, there's a clear difference in the quality of the samples provided by CNM, with a much higher percentage of unmapped reads and Ns in consensus sequence. However, both laboratories have similar distributions in the case of median DP coverage of the virus and number of variants in consensus sequence. These similarities might be thanks to the common use of the ARTIC v3 protocol mentioned before. The same happens with the number of variants in consensus sequence, which is intimately related with the coverage of the reads.

In order to assert a genomic surveillance, it is essential to study the time factor. With this goal in mind, the time distribution of the lineages is the most representative way of visualising how different lineages emerge in a certain time period.

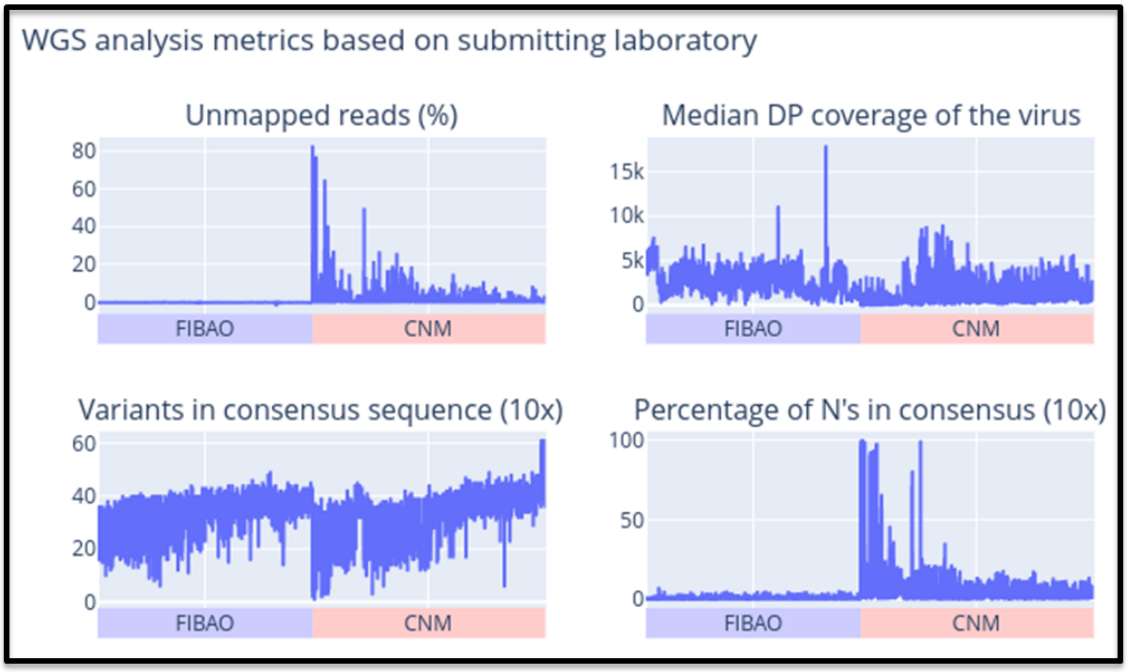

Figure 8. Metric distribution for the analysed samples. From top left to bottom right: Percentage of unmapped reads, Median depth of coverage of the virus, Variants in consensus sequence (with at least 10x depth) and Percentage of Ns in consensus sequence (with at least 10x depth).

The analysed samples were collected between January 2020 up until January 2022, with a major concurrence of collected samples between January 2021 and June 2021 (Fig 9A). In this precise period of time, most of the collected samples were categorised as Lineage B.1.1.7 by pangolin's software, which gives a meaning to why this lineage has a 65.11% of incidence over all the samples (Fig. 9B). The second most prominent lineage is B.1.177, found in 11.40% of the samples, followed by AY.43 (3.27%), B.1 (1.48%), AY.53 (1.22%), AY.4 (1.00%), P.1 (0.96%), AY.5 (0.81%), AY94 (0.79%) and B.1.575.1 (0.79%). The rest of the 95 distinct lineages fall below 1% of the samples. Only a 1.03% of the samples couldn't be assigned to a lineage by pangolin's software.

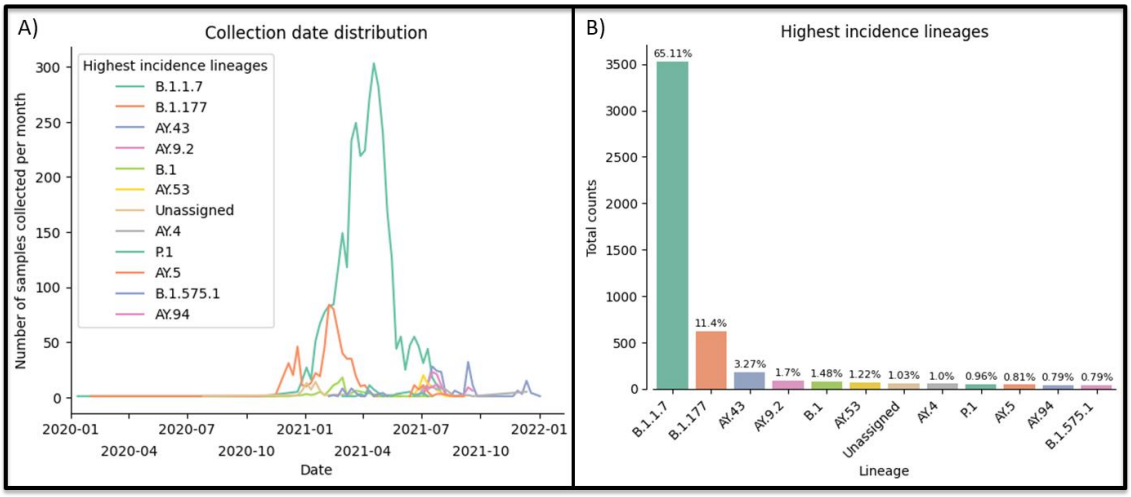

Figure 9. Time distribution (A) and frequency (B) of the most observed lineages in the results.

The assortment of lineages predicted by pangolin concurs with the expected distribution at the time of the collection of the samples as described in the literature (42).

Using the data from *long_table.csv* two plots were generated. The first plot gathered the percentage of total observed mutations affecting each gene of the virus (Fig. 10A), while the second plot contained the number of mutations found in the positions for each domain of the spike protein (Fig. 10B).

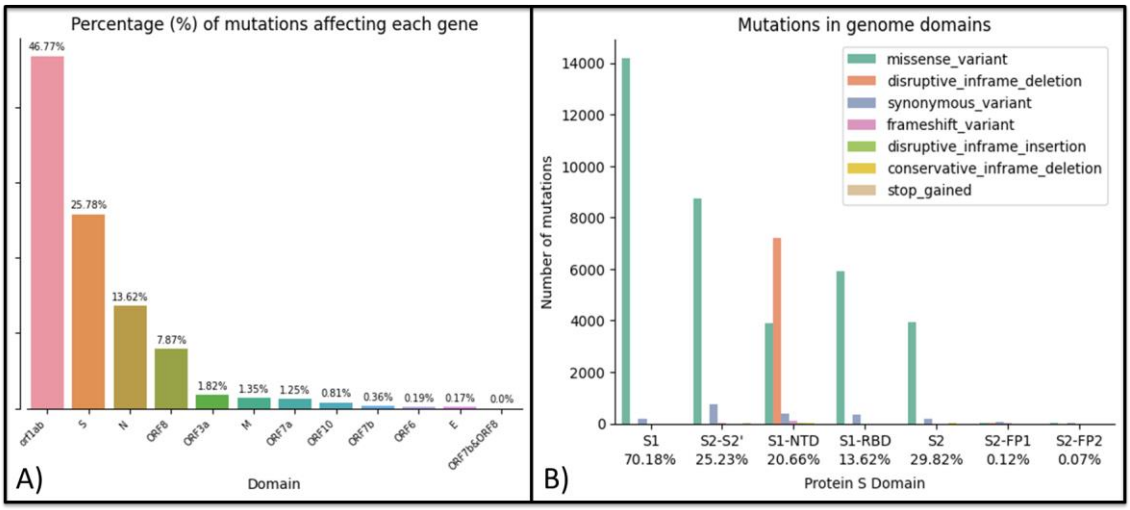

Figure 10. A) Percentage of mutations affecting each gene. B) Distribution of the different type of mutations found in each protein domain according to their genomic position in gene S. The number of mutations displayed for both S1 and S2 are the rest of mutations that don't belong to any of the domains inside those subunits, but the displayed percentage is with respect to the total mutations.

The gene that accumulates most of the mutations is orf1ab (46.77%), followed by gene S (25.78%), gene N (13.78%), orf8 (7.87%), orf3a (1.82%), gene M (1.35%) and orf7a (1.25%). The rest of the genes account for less than 1% of the mutations (Fig. 10A).

As it can be observed in Fig. 10B, there's an acute predominance of non-synonymous mutations in gene S (95.95%), over the synonymous ones (4.05%), something that has already been observed in other studies (43) (44). Most mutations are allocated in the S1 subunit (70.18% of the total mutations found), specifically in the N-Terminal Domain (NTD) and the RBD with 20.66% and 13.62% of the total mutations respectively. The rest of the mutations are allocated in the S2 subunit (29.82%), mostly in the S2' domain (25.23%). In this case the most observed type of mutations are the ones leading to missense variants (79.99%), followed by those that produce a disruptive inframe deletion (15.67%). The N-Terminal Domain allocates 99.99% of the latter.

Virus evolution analysis

The consensus sequences obtained from viralrecon, along with the metadata associated with each sample was used for a phylogenetic analysis. Nextstrain was the tool used for this process, not only because it provides a very insightful analysis with clear visualisations of the results, but also because it was a profitable way of incorporating the geographic location metadata.

264 samples were incorporated from nextstrain's database to provide a reference for other geographic locations and dates rather than Spain in the range of dates of the provided samples, this allows a more distinctive comparison and to reduce bias. Nextstrain's subsampling method was set to take one representative sample for each 100 samples in a

region. This filter, along with the dropping of genomes with a length below 90% of the average length (setting "min_length" parameter to 27200pb) and with the incorporation of the previous samples that were uploaded to Relecov's database, led to a total of 4078 analysed samples.

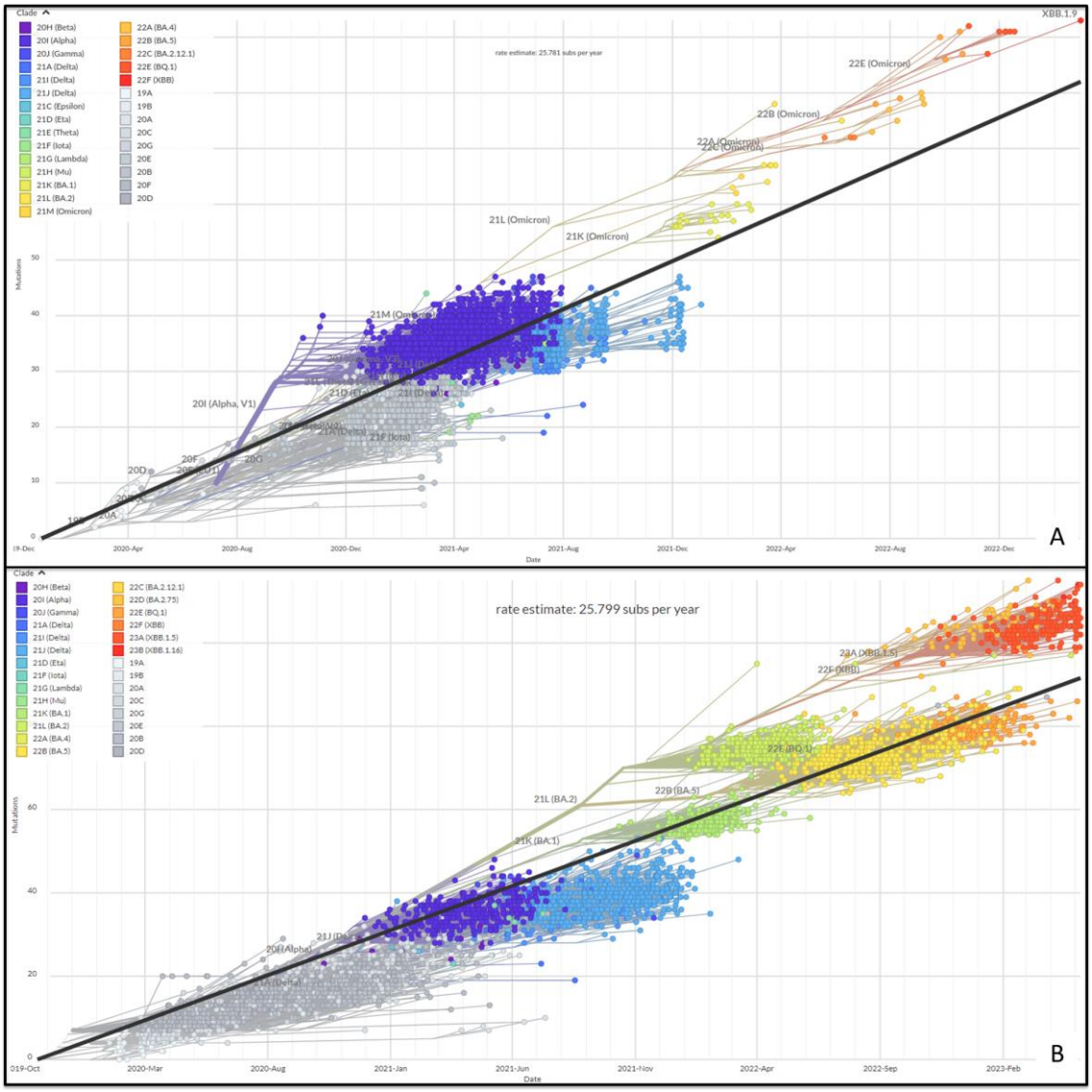

Figure 11. Clade-exploited phylogeny trees using the analysed samples (A) and the Europe samples from Nextstrain database, public in https://nextstrain.org/ncov/gisaid/europe/ (B). Branch length based on datetime, along with the predicted regression and substitution rate. The y value is the accumulation of mutations in the S gene, which in this case was selected as the parameter to determine divergence.

Nextstrain ultimately generates a user-friendly visualisation of the results in a *node.js* webpage, which can be accessed via web-browser. Nextclade and pangolin softwares use different nomenclatures to classify the genetic variants of SARS-CoV-2. Both nomenclatures use alphanumeric denominations for the variants, but Nextclade names their variants as clades (19A, 20E), while Pangolin names them as lineages (B.1.17, AY43). Although not equivalent, both nomenclatures have some similarities, which can be found in this report from the WHO.

The phylogeny tree reconstruction can be observed in Fig. 11A. Most of the provided samples are clustered into three distinct clades, 20I (Alpha variant, purple), 21J (Delta

variant, blue), and clades 20E, 20B and 20A (no assigned variant, grey colour scale). There's also a divergent small cluster formed by samples assigned to clade 21K (BA.1 variant, yellow). All the other observed clades in Fig. 11A correspond to the reference samples from Nextstrain's database.

The estimated substitution rate obtained with the analysis is 25.781 substitutions per year, which is equal to $8.63*10^{-4}$ substitutions per position per year (dividing by the mean length of the consensus sequences incorporated to the analysis). This value is very similar to the one obtained by constructing the tree using European samples from Nextstrain's database (Fig. 11B), which was 25.799 substitutions per year. This rate concurs not only with the one from Nextstrain database, but also with others found in the literature for SARS-CoV-2, that estimate 10^{-3} substitutions per position per year approximately [(45) (46)], although the one presented in this work falls on the lower scale of the spectrum.

When comparing the two phylogenetic trees, it can be observed that there's a slight increase in the estimated divergence for clade 20I in the studied samples with respect to the ones from the Nextstrain database, which might me explained due to the disparity in the number of samples for each clade. Something highly remarkable is the significant reduction of the predicted substitution rate when filtering the construction of the phylogenetic tree to the range of dates of the studied samples. This marks a change in the tendency of the evolution of the virus, which has already been noticed and mentioned in various articles (11) (47) (48).

In order to make the best of the geographical metadata for the samples, Nextstrain tool was used to generate a map that assigns each sample to the given geographic location (Fig. 12). This map varies depending on the specified time range, and therefore, five different time periods were defined to examine the virus's evolution across the Spanish region. There is a small predominance of Lineage B.1.177 in the first period over Madrid. Ceuta and Andalucía regions (Fig. 12A). In the second (Fig. 12B) and third period (Fig. 12C) the most present Lineage is B.1.1.7 with the exception of Granada and Zaragoza in the second period. The first shows a 50.77% of samples assigned to lineage B.1.177 and a 46.44% to B.1.1.7, while the second has 39.86% of the samples as lineage B.1.575.1, 36.24% as B.1.177 and 20.52% of the samples as lineage B.1.1.7. There's a change in this tendency in the fourth period (Fig. 12D), where the lineages are roughly equally distributed between Lineages B.1.1.7, AY.43, AY.9.2, AY.53, AY.4 and AY.5 depending on the region. In the fifth period (Fig. 12E) the most prevalent lineage is AY.43, with the exception of Albacete, where all the samples are associated to lineage BA.1.17; and Madrid which has an even distribution of lineages AY.4, AY.124, AY.122, BA.1 and BA.1.17.

The closest approach to this study regarding the geographic distribution of lineages across Spain in a given time range is found in (49) but it only represents the most prominent lineage in each region. In order to be able to a make proper comparison with the presented results found in the literature, the selected periods of time to analyse are the same as the ones described in the previous report, with the exception of the first period, that was not included because most of the samples in that time period were given "unassigned" as lineage by pangolin's software (Fig. 12).

Remarkably, when comparing the geographical distribution of Pangolin lineages in (49) with the one presented in this analysis, it is observed that the predominant lineages are consistent across most regions and time periods.

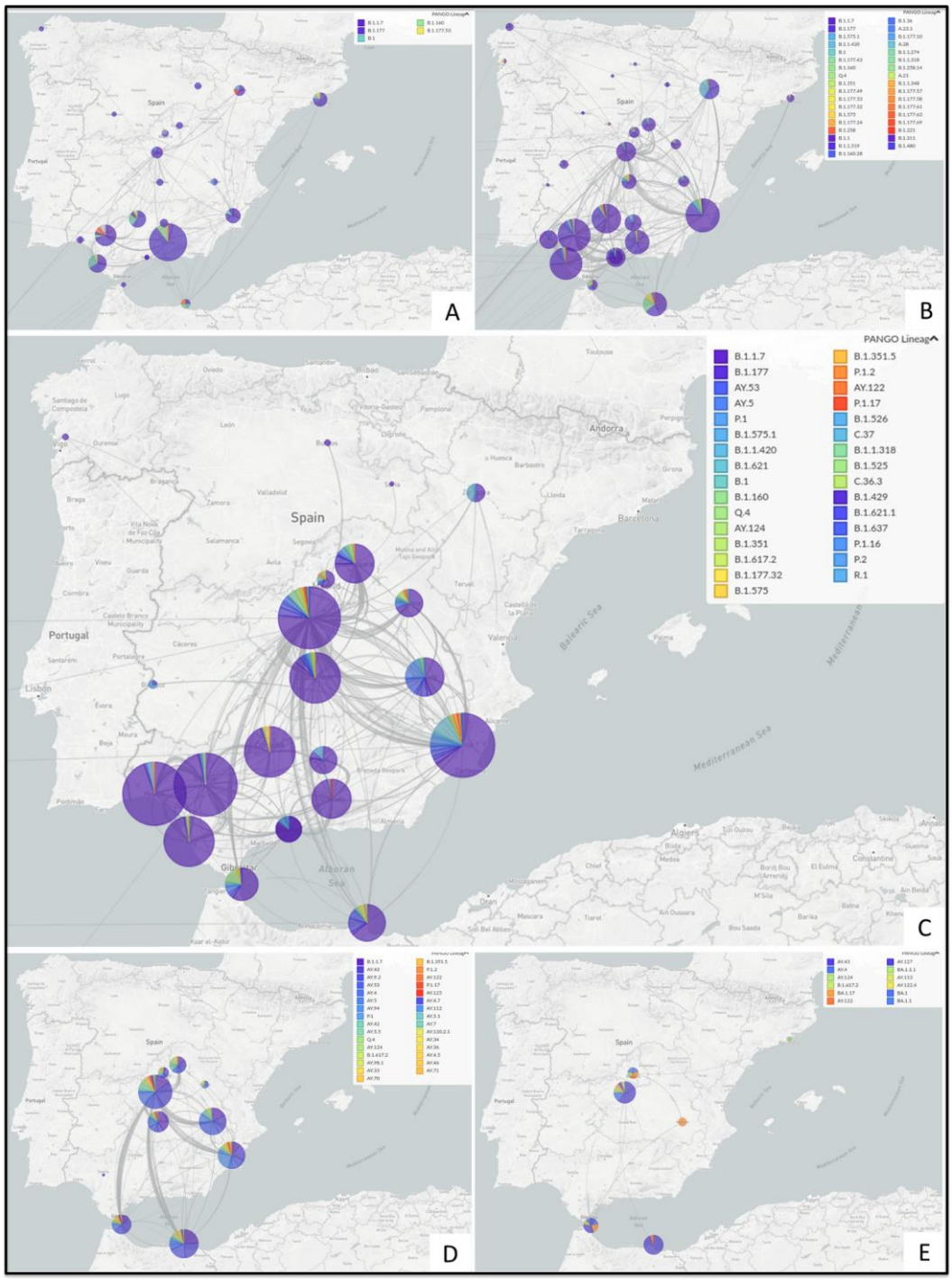

Figure 12. Geographical distribution of the analysed samples across five consecutive time periods A: 2020-06-21 to 2020-12-05; B: 2020-12-06 to 2021-03-13; C: 2021-03-14 to 2021-06-19; D: 2021-06-20 to 2021-10-17; E: 2021-10-18 to 2022-01-31. The width of the edges resembles the predicted lineage transmission between two distinct geographical locations while the size of the coloured areas is proportional to the number of samples from that region.

The maps presented in Fig. 12 are the first of its kind in Spain, as they show the lineage distribution for each region. Due to the lack of samples, specifically in the north area,

there's still not enough data to make a proper analysis of the distinctive evolution of the virus on a regional level.

Although the maps provide an insightful view of how the SARS-CoV-2 evolved differently depending on the region, it is not easy to visualise it smoothly due to the fixed time periods of 4 to 5 months for each map. A timeline of the clade frequencies on a national scale doesn't have that geographic precision but it is more accurate in the time scale, which is the reason why this plot was generated (Fig. 13). The Alpha variant (clade 20I) starts to rise in September 2020, with a major prevalence from January 2021 up to July 2021, when the Delta variant rapidly takes over (clades 21A, 21I and 21J). This variant stays with a frequency close to 100% from August 2021 until January 2022, when the BA.1 and BA.2 Omicron variants arise (clades 21K, 21L and 21M).

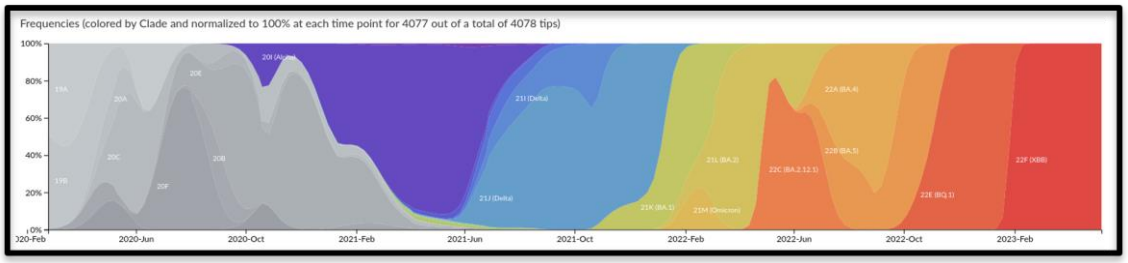

Figure 13. Timeline of the aggregated frequencies of the different clades based on sample collection date.

When looking at the frequency distribution from Fig. 13, it is important to remark that the collected samples fall mostly between January 2020 and January 2022, so the representations provided outside of that range of dates is exclusively due to the global reference samples from Nextstrain's database that were incorporated into the analysis.

However, even with the reduced amount of available data in this analysis, the distribution of variants reported along 2021 by the spanish public health authorities (https://www.sanidad.gob.es/COVID19_Actualizacion_variantes_20230424.pdf) matches in a great way with the one presented here.

Conclusions.

The Relecov project is still in development and not all the functionalities are fully functional, but the implementations into the relecov-tools package mentioned in this work will provide the necessary tools to automate the processing of the metadata provided by the laboratories in the network, ensuring no errors in the process.

So far, most of the laboratories in the network are preparing their available samples and metadata to be sent to the technical coordination team that is in charge of processing and uploading them to the database. This causes Relecov's database to lack the amount of data needed to fulfil the goal of providing an insightful analysis as yet. Despite the project being in its early stage, the analysis results presented in this work align well with the geographic and temporal lineage distributions described in the literature. This supposes a good starting point and increases expectations for when the platform has more available resources. Thanks to the FAIR structure of the database, along with the flexibility incorporated into the processing software, the platform could be expanded with little

effort to monitor not only SARS-CoV-2, but also other emergent pathogens that may appear in the future.

Supplementary materials.

Supplementary materials S1. Software versions for each step of the WGS analysis pipeline. (Note: Python version changed with each step's environment)

Software	Version	Software	Version
bcftools	1.16	ivar	1.4
bedtools	2.30.0	kraken2	2.1.2
bowtie2	2.4.4	mosdepth	0.3.3
pigz	2.6	nextclade	2.13.0
samtools	1.16.1	picard	3.0.0
python	3.9.5	sed	4.7
python	3.11.0	snpeff	5.0e
python	3.9.12	snpsift	4.3
yaml	6	tabix	1.12
getchromsizes	1.16.1	untar	1.3
fastp	0.23.2	Nextflow	22.10.1
fastqc	0.11.9	Viralrecon	2.6.0

Supplementary materials S2. Viralrecon Illumina Workflow.

Viralrecon's Illumina workflow is composed by these processes:

1. Merge re-sequenced FastQ files using shell's "*cat"* command.
2. The quality of the merged FastQ files is checked using **FastQC** for quality control (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

3. Adapter's need to be trimmed to reduce bias of the variant calling. This step is performed using **fastp (50)** by providing the adapter sequences in a FASTA format.
4. **Kraken 2 (51)** can be optionally used to remove the host reads from the sequencing data using k-mer-based algorithm to assign a taxonomic label to each read based on a reference genome database of the host, which in this case is human. These host reads can be a significant source of noise in the metagenomic data, especially if they are abundant.
5. Variant calling is the process in which the variants in the sequences are identified. There are several methods for performing this process, and Viralrecon's approach consists of the following steps.:
    - 5.a Alignment of the reads using **Bowtie 2** (52). Bowtie 2 generates an output file in SAM (Sequence Alignment/Map) format, which contains information about how each read maps to the reference genome. This information includes the read name, reference sequence name, start position, mapping quality score, and CIGAR string, which describes the alignment in terms of insertions, deletions, and mismatches relative to the reference sequence.
    - 5.b **SAMtools** (53) is used to further process the SAM files to generate a sorted and indexed BAM (Binary Alignment/Map) file, which is a compressed version of the SAM file and is more efficient for downstream analyses .
    - 5.c Primer sequence removal is very similar to the adapter trimming, but in this case **iVar (54)** is used as the software of choice, only for amplicon data.
    - 5.d **Picard** tools can be used optionally to mark duplicate reads, which are identical copies of the original DNA fragment that arise during the library preparation step (http://broadinstitute.github.io/picard/) .
    - 5.e ViralRecon's Quality Control (QC) of the alignment can be performed using picard or SAMtool. This step helps to identify potential issues with the alignment, such as poor mapping quality or high error rates.
    - 5.f **Mosdepth** (55) is a QC tool used to measure both genome-wide and per-region coverage and depth and is used to generate plots with these stats.
    - 5.g Variant callers can identify variants using statistical models and heuristics to compare the sequence data to a reference genome or a set of control samples and identify regions of the genome that differ between the sample and the reference. Each variant caller has its own strengths and in ViralRecon there are two options: iVar-variants is default for amplicon data while SAMtools-BCFTools for metagenomics data.
        - I. Variant annotation of the VCF files generated by the variant caller with **SnpEff (56)** and variant filtering based on quality with **SnpSift (57)**.
        - II. Individual variant graphical representations in ASCII format with annotation tracks using **ASCIIGenome (58)**.
    - 5.h In order to continue with the analysis, a consensus sequence is generated from the aligned sequences to represent the most likely true genetic sequence of each sample or organism. ViralRecon supports different

consensus callers: BCFTools and **BEDTools (59)** are default for amplicon and metagenomics data respectively while iVar consensus serves as an alternative for amplicon data. This consensus sequence is then used in the next steps:

I. Consensus quality assessment report with **QUAST (60).**
II. Lineage analysis using **Pangolin (61)**, which assigns each genome to a particular lineage based on its sequence characteristics and evolutionary relationships to other genomes.
III. Clade assignment, mutation calling and sequence quality checks using **Nextclade (62)** as it uses a different notation system than the one provided by pangolin .

5.i All the information from individual variants, Amino acid changes, functional effect prediction and lineage analysis for each sample is collected and merged into a purposely named "long table" in csv format.

6. As this pipeline was used to analyse viruses that had no reference genome in the time being, it also includes De novo assembly as an optional step. This process involves reconstructing the genome of the virus from the reads and includes multiple subprocesses:

6.a Primer trimming using **Cutadapt (63)**, for amplicon data only.

6.b ViralRecon lets the user choose multiple assembly tools, such as **SPAdes (64)**, **Unicycler (65)** or **minia (66)**. Once the assembly is generated, there are still some tasks that have to be undertaken:

I. Blast the contigs to a reference genome using blastn to find structural variations (67).
II. Generate a more contiguous assembly by ordering and orientating the contigs into longer sequences or scaffolds using **ABACAS (68)**.
III. Assembly report using PlasmidID for plasmids.
IV. Assembly assessment report using QUAST for genomes.

7. Finally, **MultiQC (69)** is used to generate a QC report summarising the results from all the multiple analyses of the pipeline, including visualisation for read, alignment, assembly, and variant calling results.

References

1. Adil MT, Rahman R, Whitelaw D, Jain V, Al-Taan O, Rashid F, et al. SARS-CoV-2 and the pandemic of COVID-19. Postgrad Med J. 2021 Feb;97(1144):110–6.
2. OECD O. OECD Policy Responses to Coronavirus (COVID-19): The territorial impact of COVID-19: Managing the crisis across levels of government.
3. Robishaw JD, Alter SM, Solano JJ, Shih RD, DeMets DL, Maki DG, et al. Genomic surveillance to combat COVID-19: challenges and opportunities.

Lancet Microbe. 2021 Sep;2(9):e481–4.

4. Chen Z, Azman AS, Chen X, Zou J, Tian Y, Sun R, et al. Global landscape of SARS-CoV-2 genomic surveillance and data sharing. Nat Genet. 2022 Apr;54(4):499–507.

5. Yao H, Song Y, Chen Y, Wu N, Xu J, Sun C, et al. Molecular Architecture of the SARS-CoV-2 Virus. Cell. 2020 Oct 29;183(3):730-738.e13.

6. Kleanthous H, Silverman JM, Makar KW, Yoon I-K, Jackson N, Vaughn DW. Scientific rationale for developing potent RBD-based vaccines targeting COVID-19. npj Vaccines. 2021 Oct 28;6(1):128.

7. Creech CB, Walker SC, Samuels RJ. SARS-CoV-2 Vaccines. JAMA. 2021 Apr 6;325(13):1318–20.

8. Celik I, Yadav R, Duzgun Z, Albogami S, El-Shehawi AM, Fatimawali, et al. Interactions of the Receptor Binding Domain of SARS-CoV-2 Variants with hACE2: Insights from Molecular Docking Analysis and Molecular Dynamic Simulation. Biology (Basel). 2021 Sep 7;10(9).

9. Wu C-R, Yin W-C, Jiang Y, Xu HE. Structure genomics of SARS-CoV-2 and its Omicron variant: drug design templates for COVID-19. Acta Pharmacol Sin. 2022 Dec;43(12):3021–33.

10. Meng B, Kemp SA, Papa G, Datir R, Ferreira IATM, Marelli S, et al. Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. Cell Rep. 2021 Jun 29;35(13):109292.

11. Markov PV, Ghafari M, Beer M, Lythgoe K, Simmonds P, Stilianakis NI, et al. The evolution of SARS-CoV-2. Nat Rev Microbiol. 2023 Apr 5;

12. Malim MH. APOBEC proteins and intrinsic resistance to HIV-1 infection. Philos Trans R Soc Lond B Biol Sci. 2009 Mar 12;364(1517):675–87.

13. Di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Sci Adv. 2020 Jun 17;6(25):eabb5813.

14. Jackson B, Boni MF, Bull MJ, Colleran A, Colquhoun RM, Darby AC, et al. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. Cell. 2021 Sep 30;184(20):5179-5188.e8.

15. Zhou H-Y, Cheng Y-X, Xu L, Li J-Y, Tao C-Y, Ji C-Y, et al. Genomic evidence for divergent co-infections of co-circulating SARS-CoV-2 lineages. Comput Struct Biotechnol J. 2022 Jul 28;20:4015–24.

16. Slatko BE, Gardner AF, Ausubel FM. Overview of Next-Generation Sequencing Technologies. Curr Protoc Mol Biol. 2018 Apr;122(1):e59.

17. Tyson JR, James P, Stoddart D, Sparks N, Wickenhagen A, Hall G, et al. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome

sequencing using nanopore. BioRxiv. 2020 Sep 4;

18. Ferguson JM, Gamaarachchi H, Nguyen T, Gollon A, Tong S, Aquilina-Reid C, et al. InterARTIC: an interactive web application for whole-genome nanopore sequencing analysis of SARS-CoV-2 and other viruses. Bioinformatics. 2022 Feb 7;38(5):1443–6.

19. Pappas N, Roux S, Hölzer M, Lamkiewicz K, Mock F, Marz M, et al. Virus Bioinformatics. Reference module in life sciences. Elsevier; 2020.

20. Waite DW, Liefting L, Delmiglio C, Chernyavtseva A, Ha HJ, Thompson JR. Development and validation of a bioinformatic workflow for the rapid detection of viruses in biosecurity. Viruses. 2022 Sep 30;14(10).

21. Rossing M, Sørensen CS, Ejlertsen B, Nielsen FC. Whole genome sequencing of breast cancer. APMIS. 2019 May;127(5):303–15.

22. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013 Oct 15;11(1110):11.10.1-11.10.33.

23. Denisov G, Walenz B, Halpern AL, Miller J, Axelrod N, Levy S, et al. Consensus generation and variant detection by Celera Assembler. Bioinformatics. 2008 Apr 15;24(8):1035–40.

24. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018 Dec 1;34(23):4121–3.

25. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018.

26. Jacobsen A, Kaliyaperumal R, da Silva Santos LOB, Mons B, Schultes E, Roos M, et al. A generic workflow for the data fairification process. Data Intelligence. 2020 Jan;2(1–2):56–65.

27. Shanahan H, Bezuidenhout L. Rethinking the A in FAIR data: issues of data access and accessibility in research. Front Res Metr Anal. 2022 Jul 27;7:912456.

28. Guizzardi G. Ontology, ontologies and the “I” of FAIR. Data Intelligence. 2020 Jan;2(1–2):181–91.

29. Boeckhout M, Zielhuis GA, Bredenoord AL. The FAIR guiding principles for data stewardship: fair enough? Eur J Hum Genet. 2018 Jul;26(7):931–6.

30. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017 Apr 11;35(4):316–9.

31. Patel H, Monzón S, Varona S, Espinosa-Carrasco J, Garcia MU, Bot N-C, et al.

nf-core/viralrecon: nf-core/viralrecon v2.6.0 - Rhodium Raccoon. Zenodo. 2023;

32. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020 Mar;38(3):276–8.

33. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS ONE. 2017 May 11;12(5):e0177459.

34. Cook J. Docker. Docker for data science. Berkeley, CA: Apress; 2017. p. 29–47.

35. Xia S, Zhu Y, Liu M, Lan Q, Xu W, Wu Y, et al. Fusion mechanism of 2019-nCoV and fusion inhibitors targeting HR1 domain in spike protein. Cell Mol Immunol. 2020 Jul;17(7):765–7.

36. Yoo AB, Jette MA, Grondona M. SLURM: simple linux utility for resource management. In: Feitelson D, Rudolph L, Schwiegelshohn U, editors. Job scheduling strategies for parallel processing. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003. p. 44–60.

37. Griffiths EJ, Timme RE, Page AJ, Alikhan N-F, Fornika D, Maguire F, et al. The PHA4GE SARS-CoV-2 Contextual Data Specification for Open Genomic Epidemiology. 2020 Aug 9;

38. Carbo EC, Mourik K, Boers SA, Munnink BO, Nieuwenhuijse D, Jonges M, et al. A comparison of five Illumina, Ion Torrent, and nanopore sequencing technology-based approaches for whole genome sequencing of SARS-CoV-2. Eur J Clin Microbiol Infect Dis. 2023 Jun;42(6):701–13.

39. Lam C, Gray K, Gall M, Sadsad R, Arnott A, Johnson-Mackinnon J, et al. SARS-CoV-2 Genome Sequencing Methods Differ in Their Abilities To Detect Variants from Low-Viral-Load Samples. J Clin Microbiol. 2021 Oct 19;59(11):e0104621.

40. Davis JJ, Long SW, Christensen PA, Olsen RJ, Olson R, Shukla M, et al. Analysis of the ARTIC Version 3 and Version 4 SARS-CoV-2 Primers and Their Impact on the Detection of the G142D Amino Acid Substitution in the Spike Protein. Microbiol Spectr. 2021 Dec 22;9(3):e0180321.

41. Wagner DD, Carleton HA, Trees E, Katz LS. Evaluating whole-genome sequencing quality metrics for enteric pathogen outbreaks. PeerJ. 2021 Nov 25;9:e12446.

42. López-Causapé C, Fraile-Ribot PA, Jiménez-Serrano S, Cabot G, Del Barrio-Tofiño E, Prado MC, et al. A Genomic Snapshot of the SARS-CoV-2 Pandemic in the Balearic Islands. Front Microbiol. 2021;12:803827.

43. Magazine N, Zhang T, Wu Y, McGee MC, Veggiani G, Huang W. Mutations and Evolution of the SARS-CoV-2 Spike Protein. Viruses. 2022 Mar 19;14(3).

44. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol. 2021 Jul;19(7):409–24.

45. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol. 2020 Sep;83:104351.

46. Duchene S, Featherstone L, Haritopoulou-Sinanidou M, Rambaut A, Lemey P, Baele G. Temporal signal and the phylodynamic threshold of SARS-CoV-2. Virus Evol. 2020 Jul;6(2):veaa061.

47. Pagani I, Ghezzi S, Alberti S, Poli G, Vicenzi E. Origin and evolution of SARS-CoV-2. Eur Phys J Plus. 2023 Feb 16;138(2):157.

48. Magiorkinis G. On the evolution of SARS-CoV-2 and the emergence of variants of concern. Trends Microbiol. 2023 Jan;31(1):5–8.

49. Troyano-Hernáez P, Reinosa R, Holguín Á. Evolution of SARS-CoV-2 in Spain during the First Two Years of the Pandemic: Circulating Variants, Amino Acid Conservation, and Genetic Variability in Structural, Non-Structural, and Accessory Proteins. Int J Mol Sci. 2022 Jun 7;23(12).

50. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018 Sep 1;34(17):i884–90.

51. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019 Nov 28;20(1):257.

52. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357–9.

53. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

54. Castellano S, Cestari F, Faglioni G, Tenedini E, Marino M, Artuso L, et al. iVar, an Interpretation-Oriented Tool to Manage the Update and Revision of Variant Annotation and Classification. Genes (Basel). 2021 Mar 8;12(3).

55. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. 2018 Mar 1;34(5):867–8.

56. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin). 2012 Jun;6(2):80–92.

57. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet. 2012 Mar 15;3:35.

58. Beraldi D. ASCIIGenome: a command line genome browser for console terminals. Bioinformatics. 2017 May 15;33(10):1568–9.

59. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing

genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2.

60. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013 Apr 15;29(8):1072–5.

61. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol. 2021 Jul 30;7(2):veab064.

62. Aksamentov I, Roemer C, Hodcroft E, Neher R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. JOSS. 2021 Nov 30;6(67):3773.

63. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j. 2011 May 2;17(1):10.

64. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012 May;19(5):455–77.

65. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017 Jun 8;13(6):e1005595.

66. Chikhi R, Rizk G. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. Algorithms Mol Biol. 2013 Sep 16;8(1):22.

67. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec 15;10:421.

68. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics. 2009 Aug 1;25(15):1968–9.

69. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016 Oct 1;32(19):3047–8.