PANDEM-2
PANDEMIC PREPAREDNESS AND RESPONSE

Report on laboratory data sources including capture and integration of NGS data for pandemic management. Data sources specification document and database structure for WP3.

Deliverable D2.3

29 October 2021

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 883285

PANDEM-2

Report on laboratory data sources including capture and integration of NGS data for pandemic management. Data sources specification document and database structure for WP3.

Document date: 29 October 2021
Document version: 1.0
Deliverable No: 2.3
Dissemination level: PU (Public)

Full Name	Short Name	Beneficiary Number	Role
NATIONAL UNIVERSITY OF IRELAND GALWAY	NUIG	1	
FRAUNHOFER GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG E.V	FRAUNHOFER	2	
UNIVERSITE CATHOLIQUE DE LOUVAIN	UCL	3	X
PINTAIL LTD	PT	4	
FOLKHALSOMYNDIGHETEN	FOHM	5	
RIJKSINSTITUUT VOOR VOLKSGEZONDHEID EN MILIEU	RIVM	6	
CARR COMMUNICATIONS LIMITED	CARR	7	
TERVEYDEN JA HYVINVOINNIN LAITOS	THL	8	
INSTITUTUL NATIONAL DE SANATATE PUBLICA	NIPH	9	
ROBERT KOCH-INSTITUT	RKI	10	
STICHTING KATHOLIEKE UNIVERSITEIT	RUNMC	11	X
CLARISOFT TECHNOLOGIES ROM SRL	CLARISOFT	12	X
OSTERREICHISCHES ROTES KREUZ	ORK	13	
EPICONCEPT	EPIC	14	X
INSTITUTO NACIONAL DE EMERGÊNCIA MÉDICA	INEM	15	
TRILATERAL RESEARCH LTD	TRI	16	
ISTITUTO PER L'INTERSCAMBIO SCIENTIFICO	ISI	17	
ASSOCIAZIONE DELLA CROCE ROSSA ITALIANA	ITRC	18	
INSTITUTO NACIONAL DE SAUDE DR. RICARDO JORGE	INSA	19	

Authors: Jérôme Ambroise, Julie Hurel, Bertrand Bearzatto, Olga Vybornova and Jean-Luc Gala, Université catholique de Louvain (UCL), Belgium

Coordinator: Máire Connolly, National University of Ireland, Galway (NUIG), Ireland

Project website: https://pandem-2.eu

Grant agreement number: 883285

Table of contents

1 Executive Summary

The objective of Work Package 2 (WP2), Surveillance, is to identify, capture, normalise, and aggregate surveillance data from multiple sources to provide useful surveillance indicators for situational awareness. In WP2 Task 2.5 is dedicated to the identification and mapping of laboratory data for pandemic detection and monitoring, including Next Generation Sequencing (NGS) data and integration with other data sources.

As a result, our first case study focused on laboratory data (conventional methods and NGS data) generated during the current COVID-19 pandemic caused by the SARS-CoV-2 virus. Indeed, a massive amount of laboratory data was generated to characterise this virus, and some of it is now available in a public database. We identified the most commonly used NGS data-flow and its main limitations, namely the lack of NGS data sharing, which includes raw data (fastq), processed consensus sequence (fasta), and contextual metadata. Metadata are facts about data and features thereof. These are used to describe and characterise information about how the data were obtained (i.e., technical metadata) as well as information about the sample (e.g., host sex, host age, date of collection, sample from a vaccinated individual, etc.). The sequencing strategy, protocol, and bioinformatics pipeline used to assemble the raw NGS data are examples of technical metadata. The goal of Task 2.5 is to show that the PANDEM-2 data infrastructure is capable of managing the data flows of the future. Given that the major NGS data-flow constraints (described above) will be addressed in the coming years, we intend to generate dummy (i.e. synthetic) and rich NGS-derived feature and contextual metadata that can be integrated into the PANDEM-2 data infrastructure.

In this report, the first part briefly describes the methods used for pathogen detection, focusing on the diagnostic methods applied in the context of the current COVID-19 pandemic situation. The different SARS-CoV-2 genome sequencing techniques are then described as well as the available applications and challenges in NGS for pandemic surveillance. In a second part, the 7 actions that constitute our approach to this task are listed. These actions include investigating how to transform the massive information contained in raw NGS data into a summarized and tabulated version of the genomic data that can be exploited by partners involved in this task (RUNMC and, in particular, with WP2 leader EPIC) to compute pandemic indicators integrated in the PANDEM-2 dashboard. The third and fourth sections discuss the findings and prospects for the near future, respectively.

2 Introduction & Background

This task investigates several problems of data sharing between laboratories and health authorities. Each institution conducts these analyses using a structure that they create on their own. As a result, information transmission and data accessibility are difficult. Furthermore, the data must be anonymized and safeguarded. In order to achieve a common PANDEM-2 data infrastructure, this task identifies laboratory data sources and data indicators for the PANDEM database and dashboard. The use of Point Of Care Testing (POCT)/Rapid Diagnostic Testing (RDT) results from first responders and primary care settings will be linked to a web service, which will connect them to the laboratory network and health authorities. The integration of NGS data is important as it is a key technology for pandemic management. New diagnostic tests, new case management and contact tracing tools, new ways to

predict and monitor spread will greatly benefit from fine-grained characterization of disease strains. While this technology has not yet reached its full potential in terms of full operational exploitation, our scalable data infrastructure will accept and store current and rapidly evolving NGS data, for future analysis and exploitation after the project.

In the context of the current COVID-19 Pandemic, a huge amount of laboratory data was generated. Accordingly, and in agreement with the WP2 leader EPIC, it was decided to take laboratory data (including conventional methods such as Quantitative Polymerase Chain Reaction (qPCR) as well as NGS data) generated in the context of the current pandemic as a first case study. However, the database structure and model should be as generic as possible and should therefore be easily adaptable to other viral and bacterial pathogens. Regarding bacterial pathogens, special attention will be paid to antimicrobial resistance which is considered a global health and development threat.

2.1 Conventional diagnostic methods for the detection of pathogen

In this subsection, we list and briefly describe conventional diagnostic methods for the detection of pathogens. As said before, a special focus is given to the diagnostic methods applied in the context of the current COVID-19 Pandemic. Different techniques can be used for detecting and identifying pathogenic agents. The European Centre for Disease Prevention and Control (ECDC) factsheets for specific diseases recommend several diagnostic techniques including: Nucleic acid detection by RT-PCR (Reverse Transcription Polymerase Chain Reaction), Serologic testing, Microscopic examination, Culture (bacterial culture or viral isolation on cell culture), Antigen detection, and Bioassays (i.e. subcutaneous inoculation of adult laboratory mice). Serologic testing and PCR are the most commonly used techniques and are usually the first to be developed in case of an outbreak of a new, emerging disease.

Real-time RT-PCR: Since several years, nucleic acid detection-based approaches have become a rapid and reliable technology for viral detection. Among nucleic acid tests, the PCR method is considered as the 'gold standard' for the detection of some viruses and is characterized by rapid detection, high sensitivity and high specificity. As such, real-time RT-PCR is of great interest today for the detection of SARS-CoV-2 due to its adequate sensitivity (1). When RT-PCR is used to detect SARS-CoV-2, for example, the reaction results in a threshold cycle (Ct) value, which is a relative measure of viral load. The lower the viral load, the higher the Ct value.

Serologic testing: The serology tests do not directly diagnose the presence of the virus, but the immune response (IgG/IgM immunoglobulins/antibodies), that are produced as a response to a viral infection. IgG/IgM tests are used to identify individuals who have developed an immune response due to SARS-CoV-2 infection. Although the serology tests are suitable for indirect diagnosis, the low antibody amounts produced on the first few days of the infection may be insufficient for detection (2).

Microscopic examination: Parasites, bacteria and viruses can be detected and identified on the basis of morphology. An integral microscopy study showed that SARS-CoV-2 has a similar image to SARS-CoV (3).

Culture: Bacterial culture or viral isolation on cell culture is a standard for the detection and identification of pathogens. Identification of viruses is usually based on characteristic cytopathic effects in different cell cultures.

Antigen detection: Antigen detection tests detect the presence of SARS-CoV-2 viral proteins in respiratory and salivary samples. These test kits are easy to perform and can be used as laboratory-based tests or POCT. They can provide results in 15-30 minutes and are referred to as RDT. The World Health Organization (WHO) recommended the use of SARS-CoV-2 antigen tests in settings when PCR is unavailable or when prolonged turnaround times preclude clinical utility (within the first 5–7 days following the onset of symptoms) (4). Antigen detection tests can help with the diagnosis of symptomatic SARS-CoV-2 cases. RDT can help to reduce the risk of further virus transmission in this situation.

POCT: Point-of-care tests are easy to use devices that can be used outside the laboratory settings. They can provide immediate diagnostic tests.

In the current document, we will focus on PCR-based diagnostic methods. In the context of SARS-CoV-2 detection, other methods are mainly used to produce binary results (presence or absence of the virus) and may be non-quantitative. Compared to PCR, other methods are less sensitive and less specific. The combination of PCR and NGS technologies is the most commonly used for SARS-CoV-2 detection and characterization. Interconnections between both technologies include (1) sample selection for Whole Genome Sequencing (WGS) characterization (i.e., WGS is applied on samples with low PCR Ct - meaning a high viral load -, especially when shotgun metagenomics are used), and (2) the most common NGS strategy is performed on PCR products. Subsection 2.2.2 discusses both aspects.

2.2 Genomic characterization of the pathogen using Next Generation Sequencing (NGS) technologies

2.2.1 NGS Technologies and platforms

Genome sequencing is based on various biochemical principles and aims to identify the order by which the smaller genomic units (nucleotides) are arranged in a genome. The sequencing platforms are not able to provide the genome information as a single long contiguous sequence. These platforms produce short fragments of nucleotides (called reads) of the sequenced individual or species. Moreover, reads can contain sequencing errors and the error rate varies according to the technologies used. New sequencing techniques have appeared since 2005. These technologies have initiated a new generation of sequencing that allows millions of short reads to be generated in parallel. These second-generation technologies are called NGS. The advantages of these techniques are speed and low cost compared to first generation sequencing (e.g. the Sanger method) (5).

The platforms of second-generation sequencing are Illumina and Ion torrent. Illumina technologies are sequencing by synthesis. Sequencing by synthesis is performed by detecting the nucleotide incorporated by a DNA polymerase. The Illumina platform is the most widely used technology (6). Ion

torrent technology is based on the detection of hydrogen ion released during the incorporation of nucleotides.

However, second generation sequencing technologies have some disadvantages. All these technologies rely on a long and expensive PCR amplification step. In addition, the length of the reads produced by second generation sequencers creates difficulties for the analysis and assembly of complex genomes, due to the difficulty of these reads to resolve repeated regions. Starting in 2011, a new generation of sequencers was developed to solve these problems, the third-generation sequencers. They do not require a PCR amplification step and produce longer reads as compared to second generation sequencers. Third generation reads lengths reach several thousand to several hundred thousand base pair lengths. However, these long reads are noisier than the second generation reads and have higher error rates.

The platforms of third generation sequencing are mostly Pacific Biosciences and Oxford Nanopore Technologies. Pacific Biosciences is a single-molecule real-time (SMRT) sequencing (7). This platform is based on sequencing by synthesis, the signal emitted during the incorporation of a nucleotide is detected in real time (contrary to the Illumina platform where cycles of PCR amplifications are necessary). Oxford Nanopore MinION platform is based on nanopores sequencing. This platform uses an array of pores that read nucleotide identities based on ionic current steps (8).

In the context of the current COVID-19 Pandemic, two sequencing platforms are mainly used to characterize the SARS-CoV-2 genome, namely the Illumina platform (a second-generation platform which generates short precise sequences) and the Oxford Nanopore platform (a third-generation platform which generates longer but less precise sequences).

2.2.2 NGS sequencing strategies and bioinformatics

When NGS (second or third generation) technology is used, several sequencing strategies (e.g. metagenomic, multiple PCR amplicon sequencing) can be selected according to context and the objective of the experiment. Accordingly, the sequencing strategy will not be the same for the diagnosis of infectious diseases of unidentified cause or for the monitoring of the evolution of a well characterized virus.

Metagenomic sequencing: High-throughput sequencing approaches enable genomic analyses of all microbes in a sample. Accordingly, this technology enables the generation of genomic sequences from the pathogens that are not amenable to cultivation. One such method, is named untargeted ('shotgun') sequencing and enables the researchers to sequence all microbial genetic content which is present in the sample (9).

In the context of the COVID-19 Pandemic, metagenomics has proven itself to be an important approach to virus discovery. The metagenomic approach works best when the abundance of the target virus (SARS-CoV-2) is relatively high and other microorganisms in the samples also need to be analysed. During the early phase of the pandemic, many SARS-CoV-2 genomes were obtained using metagenomics sequencing.

Multiplex PCR amplicon sequencing: Thanks to the genomes generated by metagenomic sequencing during the early phase of the pandemic, it was possible to develop a multiplex polymerase chain reaction (PCR) amplification technology targeting SARS-CoV-2. With this method, total RNA was reverse transcribed to synthesize cDNA, and PCR are then run using multiple amplification primer pairs targeting SARS-CoV-2, followed by a ligation reaction to add the indexes/barcodes. The libraries are subsequently sequenced on Illumina, or Nanopore platforms. This multiplex PCR amplification technology proved to be efficient for samples with low viral load.

However, it is worth noting that multiplex PCR amplification sequencing cannot be used to sequence highly diverse or recombinant viruses (e.g. *Lassa* virus) because the primers are designed according to the reference genomes (10). It would also be very difficult to adapt such a sequencing strategy to sequence the complete genome (WGS) of bacterial pathogens, due to the size of the genomes (several million nucleotides).

In the context of the current COVID-19 Pandemic, the most widely used protocol using the nanopore sequencing technology is the ARTIC protocol which was developed by the ARTIC network and which is fully described and freely available (https://www.protocols.io/view/ncov-2019-sequencing-protocol-bbmuik6w.pdf) (11).

Bioinformatics analysis of the NGS data: As NGS will usually generate millions of sequencing reads with or without a priori knowledge of SARS-CoV-2, the efficiency of virus characterization is heavily dependent on the downstream bioinformatics tools employed (12).

Different pipelines have been developed to process the sequencing reads in order to produce a consensus sequence of the SARS-CoV-2 genome. For example, if the sequencing reads are generated using the ARTIC protocol, a dedicated bioinformatics pipeline available on github (https://github.com/artic-network/artic-ncov2019) can be used to process the data.

When a consensus sequence is obtained for the SARS-CoV-2 genome, this sequence should be uploaded to a public repository (such as NCBI/nucleotide database or Global Initiative on Sharing Avian Influenza Data (GISAID)). Importantly, some metadata characterizing the sample but also some technical aspects (such as the sequencing strategy/protocol or the bioinformatics pipeline that was used to generate the consensus sequence) should be associated with each submitted sequence. These aspects are discussed further in the document.

In parallel, the consensus sequence should be analysed with a dedicated software in order to identify the mutations (substitution, insertion, deletion) compared to the reference genome (MN908947.3, i.e. the genome of the Wuhan-Hu-1 isolate). A second objective is to attribute meaningful nomenclatures to the sequence data, based on the genetic relatedness of the sequences (also called clade or lineage). Indeed, this nomenclature will enable a streamlined communication between different actors in the molecular epidemiology field and enables simplified tabulation of the genomic data for integration with classical epidemiological analysis.

The nomenclature which is currently recommended by WHO is using letters of the Greek alphabet. Each key SARS-CoV-2 genetic variant is characterized by a set of mutations, as shown in the following table.

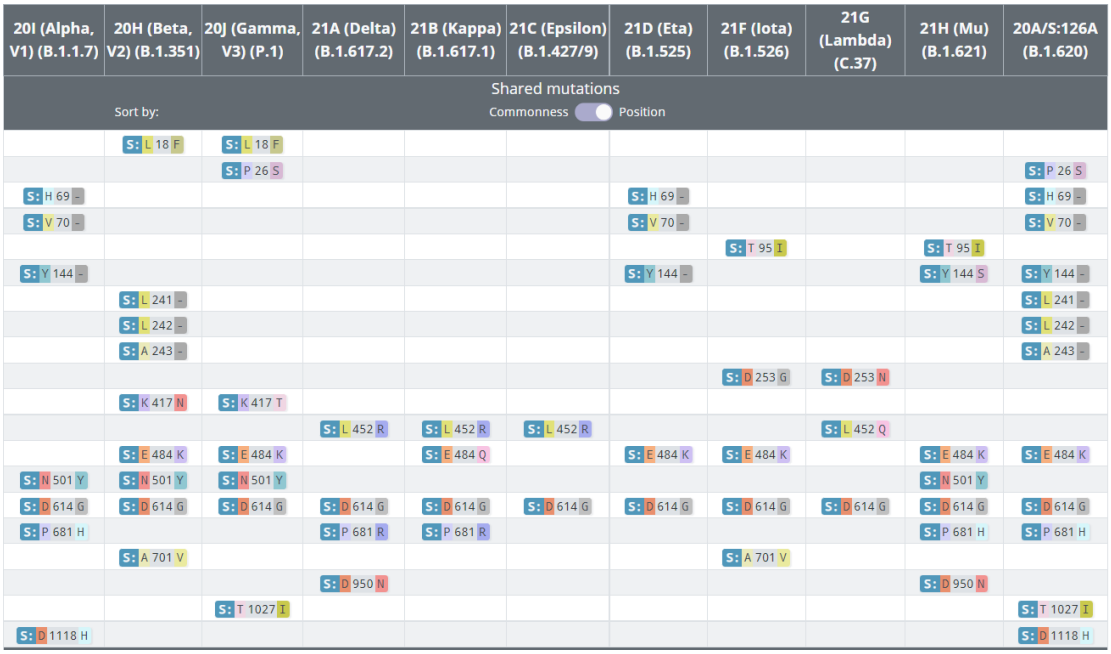

Figure 1 : Lists of the mutations within the S gene which characterize the current key variants (situation at 24/09/2021). source: https://covariants.org/shared-mutations

It is worth nothing that the phylogenetic relation between the key variants can be visualized using a phylogenetic tree that shows the genetic relatedness of the sequences, as illustrated in Figure 2.

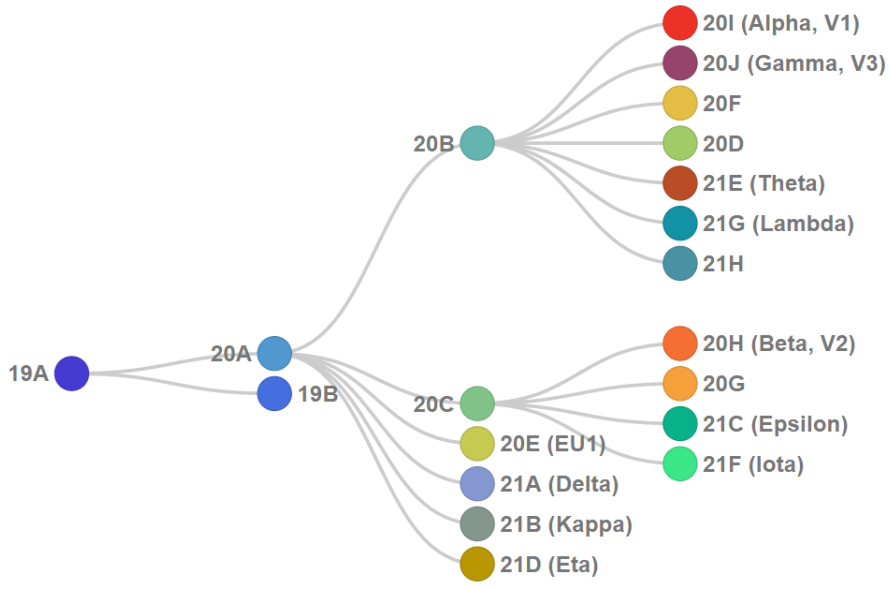

Figure 2 : Illustration of phylogenetic relationships of SARS-CoV-2 clades, as defined by Nextstrain. source: https://clades.nextstrain.org/

The ECDC and WHO regularly assesses new evidence on variants detected through epidemic intelligence, rules-based genomic variant screening, or other scientific sources. The variants are

classified in different categories, including Variants of Concern (VOC), Variants of Interest (VOI), and Variants under monitoring.

For the VOCs, there is clear evidence indicating a significant impact on transmissibility, severity and/or immunity that is likely to have an impact on the epidemiological situation in the EU/EEA. The current (September 2021) VOCs reported by ECDC are reported hereunder.

WHO label	Lineage + additional mutations	Country first detected (community)	Spike mutations of interest	Year and month first detected	Evidence for impact on transmissibility	Evidence for impact on immunity	Evidence for impact on severity	Date of decision	Transmission in EU/EEA
Beta	B.1.351	South Africa	K417N, E484K, N501Y, D614G, A701V	September 2020	Yes (v) (1)	Yes (v) (1, 2)	Yes (v) (3, 4)	2020-12-29	Community
Gamma	P.1	Brazil	K417T, E484K, N501Y, D614G, H655Y	December 2020	Yes (v) (5)	Yes (v) (6)	Yes (v) (4)	2020-12-29	Community
Delta	B.1.617.2	India	L452R, T478K, D614G, P681R	December 2020	Yes (v) (7)	Yes (v) (8-10)	Yes (v) (9, 11)	2021-05-24	Dominant

Figure 3: Current list of VOC reported by ECDC. source: https://www.ecdc.europa.eu/en/covid-19/variants-concern. accessed on september 2021.

Regarding the VOI, there is evidence based on genomic properties, epidemiological or in-vitro experiments that could imply a significant impact on transmissibility, severity and/or immunity, realistically having an impact on the epidemiological situation in the EU/EEA. Some of the current (September 2021) VOIs reported by ECDC are reported hereunder.

WHO label	Lineage + additional mutations	Country first detected (community)	Spike mutations of interest	Year and month first detected	Evidence for impact on transmissibility	Evidence for impact on immunity	Evidence for impact on severity	Date of decision	Transmission in EU/EEA
n/a	C.36+L452R	Egypt	L452R, D614G, Q677H	December 2020		Yes (m) (17)		2021-04-29	Detected (a)
n/a	AT.1	Russia	E484K, D614G, N679K, ins679GIAL	January 2021		Yes (m) (12)		2021-04-29	Detected (a)
n/a	B.1.1.318	Unclear (b)	E484K, D614G, P681H	January 2021		Yes (m) (12)		2021-04-29	Detected (a)

Figure 4 : Three VOIs (among a list of 8 VOIs) reported by ECDC. source: https://www.ecdc.europa.eu/en/covid-19/variants-concern accessed on september 2021.

2.2.3 Applications of NGS Technologies for pandemic surveillance

During the COVID-19 Pandemic, whole genome sequencing (WGS) has been used extensively by laboratories all over the world to characterize the virus. This characterization of the virus genome during the pandemic has proven to be useful for many aspects including (i) the investigation of the transmission dynamics and of the introductions of novel genetic variants, (ii) the investigation of the relationship between clades/lineages and epidemiological data such as transmissibility and disease severity or risk groups to guide public health action, (iii) the understanding of the impact of response measures on the virus population, (iii) the assessment of the impact of mutations on the performance of molecular diagnostic methods and (iv) the assessment of the impact of mutations on the performance of serological methods (13).

2.2.4 Challenges associated with the use of NGS Technology in the context of a pandemic

Bioinformatics: A first challenge associated with the use of the NGS technology in a pandemic is related to the huge amount of data and associated complexity. The bioinformatics expertise that enables us to analyze the raw data in order to produce informative results is lacking in many laboratories.

Linking genomic viral characterization with technical and sample metadata: A second challenge is related to linking the pathogen genomic data to the metadata characterizing the sample. Indeed, for the sequenced genomes to be useful, it is essential to pair them with metadata that contextualizes the time, place and circumstance of the collected sample. This context is what allows public health agencies to use genomic epidemiology to drive an effective intervention as part of a public health response. At research level, there is an urgent need for comprehensive studies linking the viral sequences of SARS-CoV-2 to the phenotype of patients affected by COVID-19.

Currently, there are already several well-defined lists of metadata that are recommended for collection. For example, submissions to the European Nucleotide Archive suggest following the 'ENA virus pathogen reporting standard checklist' (ERC000033), and recently, the Public Health Alliance for Genomic Epidemiology (PHA4GE) drafted a specification for sharing contextual data about SARS-CoV-2 genomes to advocate the openness and reusability of generated data sets. GISAID (the popular database for SARS-CoV-2 genomic sequence described in the results section) is progressively adding information regarding the 'patient status' (e.g. 'intensive care unit, serious'; 'hospitalized, stable'; 'released', 'discharged') to its records.

3 Approach

In order to achieve the objective of the current task and associated deliverable, the approach that we applied can be divided in the 7 actions that we describe below:

Action 1: Regular (every 3 weeks) meetings were organized with partners associated with current tasks. Meetings were held on Tuesday 23th March, Tuesday 13th April, Tuesday 4th May, Tuesday 25th May, Tuesday 15th June, Tuesday 20th July, Tuesday 31th August, Tuesday 21th September. During these

regular meetings, the different diagnostic laboratory methods and sources as well as the current challenges related to the NGS technologies were discussed.

Action 2: We took part in two virtual global workshops, on Friday, March 19th and 26th, organized by WHO to enhance sequencing for SARS-CoV-2 in the context of the elaboration of a globally coordinated plan to increase SARS-CoV-2 genetic sequencing capacities to detect SARS-CoV-2 mutations and variants, and to monitor virus genomic evolution worldwide. The first workshop focused on improving sequencing for SARS-CoV-2, while the second focused on sero-epidemiology of SARS-CoV-2 variants of concern. WHO is indeed working with Member States and partners to increase SARS-CoV-2 sequencing capacities and encourage timely sharing of geographically representative sequences and supporting data.

Action 3: We reviewed the scientific literature for information on the use of laboratory (including NGS) data to track pandemic progression. The most relevant search terms were "genomic data," "Next Generation Sequencing," "monitoring," and "pandemic." To narrow down the search to the COVID-19 Pandemic, search criteria included SARS-CoV-2/COVID-19.

Action 4: We explored different bioinformatics pipelines and tools dedicated to analysing raw NGS data in order to produce relevant features characterizing the genetic evolution of the targeted pathogen. A special focus was paid on the genomic characterization of the SARS-Cov-2 genome.

Action 5: We explored different data sources/databases storing NGS data and conventional diagnostic methods for the detection of pathogens. Again, SARS-CoV-2 related databases were used as a first use case.

Action 6**:** We explored existing NGS data dashboard visualization tools, such as NCBI SARS-CoV-2 nucleotide dashboard (available at https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/sars-cov-2), Nextstrain (https://nextstrain.org/ncov/open/global), and Covariants (https://covariants.org/).

Action 7: We imagined the ideal laboratory data flow based on the currently available open-source ecosystem dedicated to NGS data and we populated a database with synthetic (dummy) sample metadata and laboratory data available in open-access databases (such as NCBI). The objective is to integrate the generated dataset in the PANDEM-2 data infrastructure in order to demonstrate that it is capable of managing the data flows generated by NGS technologies.

4 Results: identification of data sources and database structure for WP3 – Situational awareness

Regular contacts with partners (action 1), as well as attendance at both WHO workshops (action 2) and reviewing the scientific literature (action 3), enabled us to gain information on the use of laboratory data for pandemic detection and monitoring. These tasks, in particular, enabled us to identify the main laboratory data sources and to gain a clear understanding of the current NGS data flow, which aims to transform the massive information contained in raw NGS data into a summarised and tabulated version of the genomic data that can be used for genomic epidemiology. This also enabled us to identify the

main limitations of the existing data-flow, namely (i) the low percentage of NGS raw (fastq) and processed (fasta) data available in public databases, and (ii) the lack of standardised contextual metadata associated with each NGS data. In that respect, we were finally able to design a strategy that will be implemented to demonstrate that the PANDEM-2 data infrastructure is capable of managing the data-flows of tomorrow.

4.1 Identification of data sources

The different data sources that we identified are listed and briefly described below. This includes both data sources for conventional diagnostic methods such as PCR and data sources for NGS data.

4.1.1 Conventional diagnostic methods

The ECDC aggregates and reports the daily number of new reported (i.e. confirmed by PCR) COVID-19 cases and deaths for each country. This information is daily updated on the following website (https://www.ecdc.europa.eu/en/publications-data/data-daily-new-cases-covid-19-eueea-country). Several data formats are available including CSV, XLS, XML and JSON. Each row/entry contains the number of new cases and deaths reported per day and per country in the EU/EEA as shown in Figure 5.

dateRep	year	month	day	cases	deaths	ntriesAndTerritc	geoId	popData2020	untryterritoryCo	continentExp
27/09/2021	2021	09	27	1556	6	Austria	AT	8901064	AUT	Europe
26/09/2021	2021	09	26	2029	4	Austria	AT	8901064	AUT	Europe
25/09/2021	2021	09	25	1803	10	Austria	AT	8901064	AUT	Europe
24/09/2021	2021	09	24	1708	9	Austria	AT	8901064	AUT	Europe
23/09/2021	2021	09	23	2089	12	Austria	AT	8901064	AUT	Europe
22/09/2021	2021	09	22	1242	17	Austria	AT	8901064	AUT	Europe
21/09/2021	2021	09	21	1162	7	Austria	AT	8901064	AUT	Europe
20/09/2021	2021	09	20	1708	7	Austria	AT	8901064	AUT	Europe
19/09/2021	2021	09	19	2072	5	Austria	AT	8901064	AUT	Europe
18/09/2021	2021	09	18	2235	9	Austria	AT	8901064	AUT	Europe
17/09/2021	2021	09	17	2283	8	Austria	AT	8901064	AUT	Europe
16/09/2021	2021	09	16	2638	8	Austria	AT	8901064	AUT	Europe
15/09/2021	2021	09	15	1859	7	Austria	AT	8901064	AUT	Europe
14/09/2021	2021	09	14	1227	4	Austria	AT	8901064	AUT	Europe

Figure 5: Capture of a small part of data from ECDC showing the number of cases and deaths in Austria during a short period. source: https://www.ecdc.europa.eu/en/publications-data/data-daily-new-cases-covid-19-eueea-country

The ECDC also reports data on testing for COVID-19 by week and country. The figures displayed for weekly testing rate and weekly test positivity are based on multiple data sources, the main source being data submitted by Member States to the European Surveillance System (TESSy). When not available, ECDC compiles data from public online sources. It is worth noting that EU/EEA Member States report in TESSy all tests performed (i.e. both PCR and antigen tests). The file is updated weekly and each row contains the corresponding data for a country and a week as shown in Figure 6.

country	country_code	year_week	level	region	region_name	new_cases	tests_done	population	testing_rate	positivity_rate	testing_data_source
Austria	AT	2020-W30	national	AT	Austria	915	99229	8901064	1114,799309	0,922109464	Country website
Austria	AT	2020-W31	national	AT	Austria	823	57416	8901064	645,0464798	1,433398356	Country website
Austria	AT	2020-W32	national	AT	Austria	702	56554	8901064	635,3622443	1,241291509	Country website
Austria	AT	2020-W33	national	AT	Austria	1362	56622	8901064	636,1261979	2,405425453	Country website
Austria	AT	2020-W34	national	AT	Austria	1866	76497	8901064	859,4141105	2,439311346	Country website
Austria	AT	2020-W35	national	AT	Austria	1979	77105	8901064	866,2447546	2,56662992	Country website
Austria	AT	2020-W36	national	AT	Austria	1976	83733	8901064	940,7077626	2,359882006	Country website
Austria	AT	2020-W37	national	AT	Austria	4141	86241	8901064	968,8841694	4,801660463	Country website
Austria	AT	2020-W38	national	AT	Austria	5222	102617	8901064	1152,862175	5,088825438	Country website
Austria	AT	2020-W39	national	AT	Austria	4909	110816	8901064	1244,974758	4,429865723	Country website
Austria	AT	2020-W40	national	AT	Austria	5152	130874	8901064	1470,318605	3,936610786	Country website
Austria	AT	2020-W41	national	AT	Austria	7365	124663	8901064	1400,54043	5,907927773	TESSy
Austria	AT	2020-W42	national	AT	Austria	9574	129647	8901064	1456,533736	7,384667597	TESSy
Austria	AT	2020-W43	national	AT	Austria	16979	158997	8901064	1786,269597	10,67881784	TESSy
Austria	AT	2020-W44	national	AT	Austria	28574	167926	8901064	1886,583447	17,0158284	TESSy
Austria	AT	2020-W45	national	AT	Austria	41398	199567	8901064	2242,057803	20,74391057	TESSy
Austria	AT	2020-W46	national	AT	Austria	50986	215044	8901064	2415,935893	23,70956641	TESSy
Austria	AT	2020-W47	national	AT	Austria	42630	207745	8901064	2333,93446	20,52034947	TESSy
Austria	AT	2020-W48	national	AT	Austria	32058	196461	8901064	2207,163099	16,31774245	TESSy
Austria	AT	2020-W49	national	AT	Austria	22794	163770	8901064	1839,892399	13,91830005	TESSy
Austria	AT	2020-W50	national	AT	Austria	19060	162984	8901064	1831,061994	11,69439945	TESSy
Austria	AT	2020-W51	national	AT	Austria	16186	185766	8901064	2087,008924	8,713112195	TESSy

Figure 6: Capture of a small part of data from ECDC showing the weekly number of diagnostic tests performed in Austria during a short period. source: https://www.ecdc.europa.eu/en/publications-data/covid-19-testing

As said before, ECDC compiles data submitted by Member States. Each country can also publish its own data online and this can be done with a higher level of details. For example, regarding the number of confirmed cases in Belgium, the user can access the information by date, age, sex and province as shown in Figure 7 on the Sciensano (i.e., Belgian public health agency) website.

DATE	PROVINCE	REGION	AGEGROUP	SEX	CASES
2021-01-31	Liège	Wallonia	40-49	F	1
2021-01-31	Liège	Wallonia	40-49	M	1
2021-01-31	Liège	Wallonia	50-59	M	3
2021-01-31	Liège	Wallonia	60-69	M	3
2021-01-31	Liège	Wallonia	80-89	F	1
2021-01-31	Liège	Wallonia	90+	F	2
2021-01-31	Limburg	Flanders	0-9	F	6
2021-01-31	Limburg	Flanders	0-9	M	4
2021-01-31	Limburg	Flanders	10-19	F	11
2021-01-31	Limburg	Flanders	10-19	M	10
2021-01-31	Limburg	Flanders	20-29	F	8
2021-01-31	Limburg	Flanders	20-29	M	9
2021-01-31	Limburg	Flanders	30-39	F	7
2021-01-31	Limburg	Flanders	30-39	M	13
2021-01-31	Limburg	Flanders	40-49	F	7
2021-01-31	Limburg	Flanders	40-49	M	7
2021-01-31	Limburg	Flanders	50-59	F	6
2021-01-31	Limburg	Flanders	50-59	M	9
2021-01-31	Limburg	Flanders	60-69	F	3
2021-01-31	Limburg	Flanders	60-69	M	3
2021-01-31	Limburg	Flanders	70-79	F	4
2021-01-31	Limburg	Flanders	70-79	M	1

Figure 7: Capture of a small part of data from Sciensano showing the daily number of confirmed cases in Belgium by age (age group of 10 years), sex, and province, during a short period. source: https://epistat.wiv-isp.be/covid/

4.1.2 Whole genome characterization

In this subsection, we list the main data sources which store whole genome characterization of the SARS-CoV-2 virus.

The ECDC aggregates information about the volume of COVID-19 sequencing, the number and percentage distribution of VOC for each country, week and variant submitted since 2020-W40 to the GISAID EpiCoV database (https://www.gisaid.org/) and TESSy (as either case-based or aggregate data). As illustrated for Germany for week n°36 of 2021 in Figure 8, almost all sequenced isolates belonged to the B.1.617.2 (i.e., Delta) variant.

country	country_code	year_week	source	new_cases	number_sequenced	rcent_cases_sequenc	valid_denominator	variant	nber_detections_var	percent_variant
Germany	DE	2021-35	GISAID	73284	7489	10,2	VRAI	Other	6	0,1
Germany	DE	2021-36	GISAID	72761	2917	4	VRAI	B.1.1.7	1	0
Germany	DE	2021-36	GISAID	72761	2917	4	VRAI	B.1.1.7+E484K	0	0
Germany	DE	2021-36	GISAID	72761	2917	4	VRAI	B.1.351	0	0
Germany	DE	2021-36	GISAID	72761	2917	4	VRAI	B.1.427/B.1.429	0	0
Germany	DE	2021-36	GISAID	72761	2917	4	VRAI	B.1.525	0	0
Germany	DE	2021-36	GISAID	72761	2917	4	VRAI	B.1.526	0	0
Germany	DE	2021-36	GISAID	72761	2917	4	VRAI	B.1.616	0	0
Germany	DE	2021-36	GISAID	72761	2917	4	VRAI	B.1.617	0	0
Germany	DE	2021-36	GISAID	72761	2917	4	VRAI	B.1.617.1	0	0
Germany	DE	2021-36	GISAID	72761	2917	4	VRAI	B.1.617.2	2915	99,9
Germany	DE	2021-36	GISAID	72761	2917	4	VRAI	B.1.617.3	0	0
Germany	DE	2021-36	GISAID	72761	2917	4	VRAI	B.1.620	0	0
Germany	DE	2021-36	GISAID	72761	2917	4	VRAI	B.1.621	0	0
Germany	DE	2021-36	GISAID	72761	2917	4	VRAI	C.1.2	0	0
Germany	DE	2021-36	GISAID	72761	2917	4	VRAI	C.37	0	0
Germany	DE	2021-36	GISAID	72761	2917	4	VRAI	P.1	1	0
Germany	DE	2021-36	GISAID	72761	2917	4	VRAI	P.3	0	0
Germany	DE	2021-36	GISAID	72761	2917	4	VRAI	Other	0	0

Figure 8: Capture of a small part of data from ECDC showing the weekly number of sequenced genomes in Germany during a short period and the number of detected VOC. source: https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea

GISAID database: As said before, data reported by ECDC is derived from the GISAID database. GISAID is indeed the most used database to share SARS-CoV-2 genomic sequences. To search or download sequences from GISAID, or use the platform's genomic-analysis tools, people must register with their name, and agree to terms that include not publishing studies based on the data without acknowledging the scientists who uploaded the sequences, and even contacting them to ask about collaboration.

In addition to the virus genomic sequence, some metadata are collected for each sample. Some fields such as location, host, gender, and patient age are required but several of them likely constitute personally-identifiable information. In practice, some of these metadata are not provided by the user as reported by Velazquez et al. who showed that more than 68% of GISAID entry (over the 75.000 entries in GISAID when the study was conducted) did not include the 'gender' and 'age' fields (14).

METADATA FIELDS (GISAID)
Virus name
Accession ID
Type
Collection date
Location
Additional location information
Host
Additional host information
Gender
Patient age
Patient status
Specimen source
Outbreak detail
Last vaccinated
Treatment

Figure 9: List of metadata that should be associated with each SARS-CoV-2 genomic sequence. Accordingly, the user which submits a sequence should provide these metadata. source: https://github.com/CDCgov/SARS-CoV-2_Sequencing

NCBI database: Another popular database that reports SARS-CoV-2 genomic characterization is hosted by the NCBI (National Center for Biotechnology Information). This database, based in the USA, is a fully open-source resource which is coordinated by the International Nucleotide Sequence Database Collaboration (INSDC). This consortium also involves two other open-source viral sequences databases including DDBJ (Japan), and EMBL-EBI (Europe).

NCBI has specific resources dedicated to SARS-CoV-2 available here (https://www.ncbi.nlm.nih.gov/sars-cov-2/), including a huge number of nucleotide records (> 1.5 million when accessed in September 2021). By using this resource, it is possible to download both the sequences and associated metadata. The metadata associated is available as a CSV file where each line corresponds to one sequence and each column to a specific field. Figure 10 illustrates some metadata fields that can be downloaded.

The 'Biosample' field enables the user to access some additional sample metadata. Indeed, BioSample databases at NCBI are used to facilitate capture and organization of metadata (15). It provides a dedicated environment which intends to capture sample metadata in a structured way by promoting use of controlled vocabularies for sample attribute field names and which enable to link sample information to corresponding experimental data across multiple archival databases (e.g. raw NGS data in 'SRA' database and consensus nucleotide sequence in the 'nucleotide' database).

Accession	Release_Date	Pangolin	Country	Host	Isolation_Source	Collection_Date	BioSample
LR963062	25-01-21 01:00	B.1.329	Germany	Homo sapiens			SAMEA7372327
LR963063	25-01-21 01:00	B.1.329	Germany	Homo sapiens			SAMEA7372322
LR963064	25-01-21 01:00	B.1.329	Germany	Homo sapiens			SAMEA7372325
LR963065	25-01-21 01:00	B.1.329	Germany	Homo sapiens			SAMEA7372321
LR963066	25-01-21 01:00	B.1.329	Germany	Homo sapiens			SAMEA7372326
LR963067	25-01-21 01:00	B.1.329	Germany	Homo sapiens			SAMEA7372324
LR963068	25-01-21 01:00	B.1.329	Germany	Homo sapiens			SAMEA7372323
LR963069	25-01-21 01:00	B.1.329	Germany	Homo sapiens			SAMEA7372329
LR963070	25-01-21 01:00	B.1.329	Germany	Homo sapiens			SAMEA7372328
LR963071	25-01-21 01:00	B.1.329	Germany	Homo sapiens			SAMEA7372333
LR963072	25-01-21 01:00	B.1.329	Germany	Homo sapiens			SAMEA7372336
LR963074	25-01-21 01:00	B.1.329	Germany	Homo sapiens			SAMEA7372330
LR963075	25-01-21 01:00	B.1.329	Germany	Homo sapiens			SAMEA7372332
LR963076	25-01-21 01:00	B.1.329	Germany	Homo sapiens			SAMEA7372334
LR963077	25-01-21 01:00	B.1.329	Germany	Homo sapiens			SAMEA7372331
LR963078	25-01-21 01:00	B.1.329	Germany	Homo sapiens			SAMEA7372335
MW514307	25-01-21 01:00	B.1.1.317	Russia	Homo sapiens	oronasopharynx	04-06-20	
MW505982	22-01-21 01:00	B.1.160	France	Homo sapiens	oronasopharynx	24-09-20	
MW491232	19-01-21 01:00	B.1.1.7	Italy	Homo sapiens	oronasopharynx	09-01-21	
MW485795	18-01-21 01:00	B.1.1.70	Serbia	Homo sapiens		04-08-20	
MW485796	18-01-21 01:00	B.1.1.70	Serbia	Homo sapiens		23-07-20	
MW485797	18-01-21 01:00	B.1.1.70	Serbia	Homo sapiens		03-04-20	
MW485798	18-01-21 01:00	B.1.1.70	Serbia	Homo sapiens		23-07-20	
MW485799	18-01-21 01:00	B.1.1.70	Serbia	Homo sapiens		23-07-20	

Figure 10: Some metadata associated with nucleotide entries available on NCBI.

Ideally, each nucleotide entries should be linked to a Biosample and each Biosample should at least report host age and sex. But, as in the case of the GISAID database, such information is frequently missing.

PHA4GE initiative: Regarding the metadata that should be provided by the databases, some recommendations are made by PHA4GE (https://pha4ge.org/), a global coalition that is actively working to establish consensus standards, document and share best practices, to improve the availability of critical bioinformatic tools and resources, and advocate for greater openness, interoperability, accessibility and reproducibility in public health microbial bioinformatics. The metadata proposed in PHA4GE are much more complete compared to other databases. The SARS-CoV-2 Contextual Data Specification - Collection template and associated materials for SARS-CoV-2 metadata are available on github (https://github.com/pha4ge/SARS-CoV-2-Contextual-Data-Specification).

4.2 Database structure for WP3 – Situational awareness

The exploration of the laboratory data sources currently available for European countries highlighted the current lack of harmonization regarding the metadata associated with each sample. This observation applies not only to the definition of the fields but also to the ontologies used to characterize these fields.

The database structure that we propose follows the philosophy of the NCBI ecosystem. Accordingly, each new sample should receive a unique identifier and several required metadata fields should be provided by the laboratory that performs the test. The required metadata fields should at least include patient age, sex, vaccination, location and date with dedicated ontologies associated with each field. It is worth noting that for the PANDEM-2 prototype, we aim to generate synthetic ('dummy') data which has the format and characteristics of sensitive data, but does not in fact describe any real person or persons, as specified in the Grant Agreement. Accordingly, no sensitive data will be collected or stored by the PANDEM-2 prototype system.

When a list of thousands of 'Biosample' identifiers will have been created and associated with synthetic metadata, different laboratory data will be linked to these Biosamples. Laboratory data include conventional (e.g. PCR) and NGS data. Regarding NGS data, while we recommend storing the raw data (e.g. on SRA or ENA), we concentrate here on the consensus (e.g. the assembled) genomic sequence that should be also available (for example on NCBI nucleotide database or GISAID). For populating the database of the PANDEM-2 prototype, SARS-CoV-2 genomic sequences will be downloaded from the NCBI SARS-CoV-2 dedicated database or from GISAID. Each nucleotide sequence will also be characterized by technical metadata including the sequencing strategy and protocol and the bioinformatics pipeline which was used to assemble the raw NGS data. All nucleotide sequences will then be analysed with the Nextclade software (https://clades.nextstrain.org/) in order to produce, for each sequence, the list of mutations (substitution, insertion, deletion, compared to the reference genome) as well as the name of the corresponding clade and lineage.

While some characteristics of the database such the standardized ontology and the level of completeness of metadata fields will not reflect the current observed situation, we are confident that such characteristics could be observed in a future pandemic thanks to initiative such as those initiated by the PHA4GE global coalition and thanks to lessons learned from the current pandemic.

Having a unique Biosample identifier that could be linked to different laboratory data could facilitate some computation such as:

- the percentage of PCR-positive samples that are sequenced
- the correlation between the PCR Ct value (related to the viral load) and the genomic information
- the assessment of the impact of mutations on the performance of molecular diagnostic methods.

Having systematic sample metadata information for each sequenced genome will enable us to create a dashboard such as covariant.org dashboard where the spreading of the variant could not only be visualized according to the country (as performed on covariant.org and illustrated in Figure 11) but also according to other metadata such as host age, sex or vaccination status.

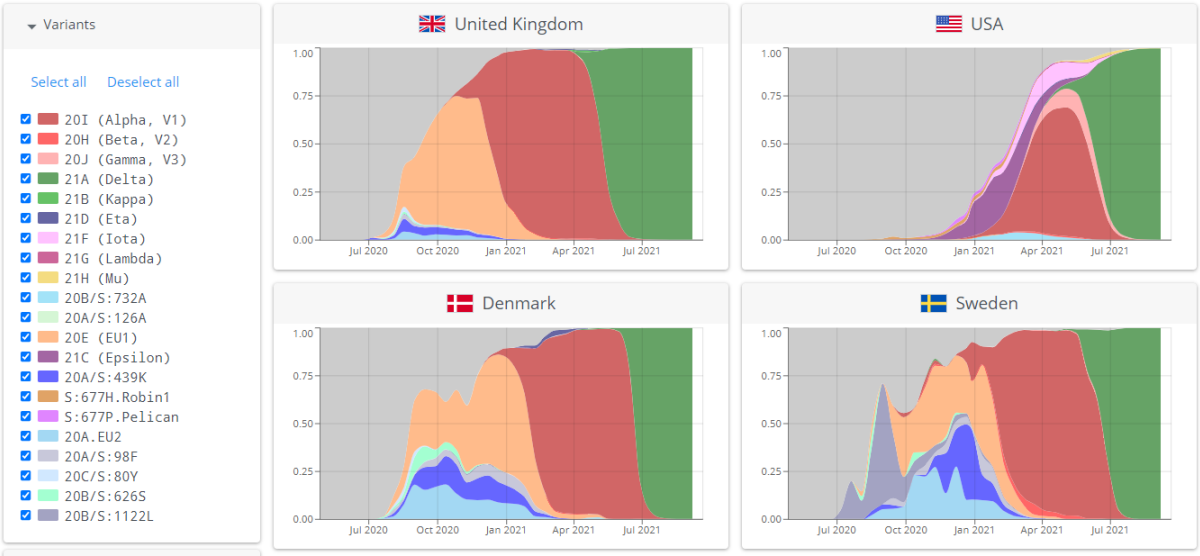

Figure 11: Spreading of the SARS-Cov-2 variants in different countries provided by covariants.org dashboard which is based on data available in the GISAID database.

Such metadata-stratified visualization would enable us to monitor if a specific variant present in a country is leading to vaccine escape mutations reducing the effectiveness of vaccination.

The database structure and content are illustrated hereunder (in Figure 12) and use the notion of Biosample as in the NCBI and EBI databases. Accordingly, several assays (e.g. a PCR and NGS experiment) can be connected to one Biosample. The NGS data and associated metadata are stored in 3 different databases. The first database stores the raw NGS data (e.g. SRA/NCBI) while the second stores the assembled NGS data (i.e., the consensus genome, e.g. nucleotide/NCBI). We can imagine a third database which stores genomic information that can be directly used by epidemiologists. In the case of the SARS-CoV-2 genomic characterization, this database includes the list of mutations (substitution and deletion/insertion compared to the reference genome) as well as the clade/lineage of the virus.

Importantly, the NGS raw data and assembled data should not be imported in the PANDEM-2 database. Indeed, the volume of the data is an important consideration for each sample (several Mb for raw NGS fastq files) and in addition it cannot be directly integrated with other epidemiological data.

On the other hand, NGS-derived data (list of mutations and clades) are imported in the database and also stores accession numbers both for the raw sequences (fastq file) and the processed assembled sequence (fasta file). By this way, if a new version of the annotation algorithm (e.g. Nextclade https://clades.nextstrain.org/) is released, the sequence can be retrieved and re-analysed with the new version. Similarly, if a new version of the assembly method (e.g. ARTIC) is released, the raw NGS data can be retrieved to generate an improved consensus sequence that can be annotated with Nextclade.

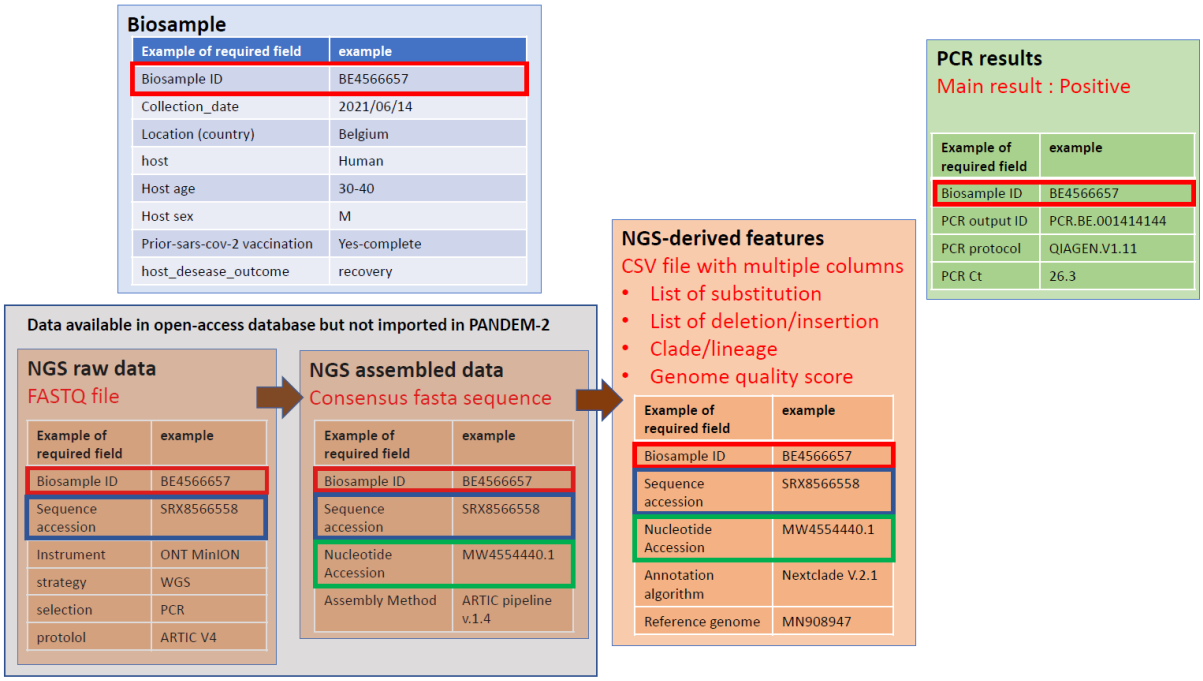

Figure 12: In the database structure that we propose, each sample is associated with a sample (e.g. Biosample) ID which is used to be connected with different laboratory data (e.g. PCR and NGS consensus nucleotide sequence).

5 Perspectives, impact and conclusion

5.1 Adaptation of the database structure and content to other pathogens

In this subsection, we explain how the database structure and content proposed hereunder in the case of the COVID-19 Pandemic should be adapted for the characterization of other pathogens.

If the pandemic agent is a virus (e.g., Ebola, Monkeypox, Lassa, Marburg, Zika), the global database structure and content can directly be adapted from the solution presented in section 4.2.

If pandemic agents are bacteria, the global structure can be applied but the set of NGS-derived features will be different. In such cases, the set will include the presence/absence of virulence genes. The list of virulence genes for many bacterial species are stored in dedicated databases such as VFDB (http://www.mgc.ac.cn/VFs/) (16). NGS-derived features will also include Antimicrobial Resistance (AMR) characterization such as presence/absence of resistance genes (located on plasmid) and

presence/absence of some specific mutations. Such features can be extracted from the assembled genome (fasta files) using dedicated bioinformatics tools such as AMRFinder tool developed by the National Center for Biotechnology Information (NCBI) (17).

An important source of whole genome characterization of bacterial pathogens is the new (currently still in Beta version) NCBI Pathogen Detection Isolates Browser (https://www.ncbi.nlm.nih.gov/pathogens/) which is a web-based portal that integrates the genomic sequence, metadata, antibiotic susceptibility and resistance gene information as well as the SNP cluster information.

5.2 Adaptation of the database structure and content for other type of samples (e.g. wastewater)

In the context of the current COVID-19 Pandemic, it has been shown that the surveillance of SARS-CoV-2 in wastewater can provide important complementary and independent information to public health authorities. However, it is not a replacement for existing COVID-19 testing approaches and strategies.

While the database structure and content are initially designed to handle data from patient samples, it can be adapted to other types of sample such as wastewater samples. This adaptation is facilitated by the fact that the two main laboratories technologies are the same as for patient samples. (i.e., qPCR for detection and NGS for genomic characterization). In the case of a wastewater sample, the Biosample metadata should not refer to host sex and age, but rather to wastewater sample metadata such as collection date, collection localization and water temperature.

Wastewater surveillance is a tool to observe trends and not an absolute means to draw conclusions about the prevalence of COVID-19 in the population. It can serve different purposes at different stages of an epidemic. As a consequence, the European Commission has strongly encouraged Member States to put in place as soon as possible and no later than **1 October 2021** a national wastewater surveillance system targeted at data collection of SARS-CoV-2 and its variants in wastewaters (18). Accordingly, we expect that new laboratory data sources characterizing wastewater will be published in the upcoming months and we will investigate how to integrate such data in PANDEM-2 database/dashboard.

5.3 Connection to EpiPulse – the European surveillance portal for infectious diseases

Most data sources described in the Results sections and characterizing laboratory data in Europe are currently stored and exchanged via TESSy, the current European surveillance portal. As an example, data available on ECDC website for Sweden about the number of tests performed is illustrated in Figure 13.

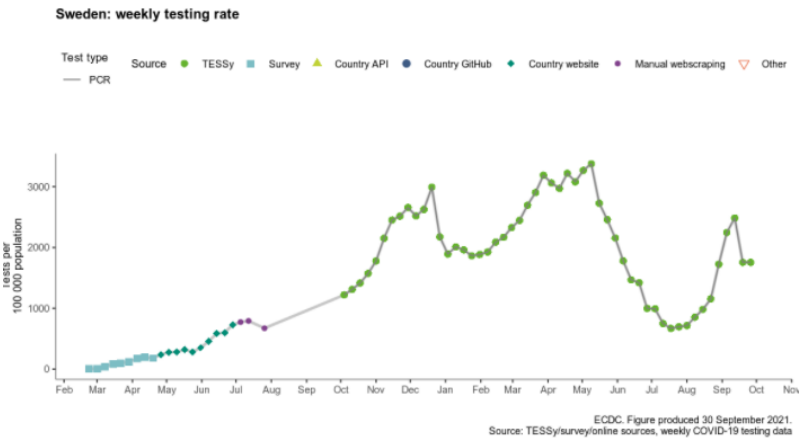

Figure 13: source: https://covid19-country-overviews.ecdc.europa.eu/#35_Sweden

As illustrated in this example, the main source of total tests per country or subnational region per week is aggregate data submitted by Member States to TESSy. However, when not available, as was often the case earlier in the pandemic, ECDC compiled data from public online sources. These data have been automatically or manually retrieved ('web-scraped') daily from national/official public online sources from EU/EEA countries (19).

However, TESSy is planned to be progressively replaced by EpiPulse between 2021 and 2023. This new European surveillance portal for infectious diseases will provide new functionalities and seamless access to data (including whole genome sequencing pathogen characterization) in a single platform. Accordingly, we intend to investigate the data sources that will be available via this new platform and how these sources will be connected to the PANDEM-2 database.

5.4 Connection with Bioconductor ecosystem

Bioconductor uses the R statistical programming language and is an open-source, open-development software project for the analysis and comprehension of high-throughput data in genomics and molecular biology. For many years, the Bioconductor infrastructure has proved to match the requirements for handling omics data (e.g. bulk or single-cell transcriptomics, genomics, proteomics). To this end, Bioconductor employs a flexible object-oriented paradigm that enables users to encapsulate multiple object components into a single instance and preserve the relations between primary (i.e. genomic, transcriptomic, ...) data and metadata.

In the context of research projects that would focus on the impact of SARS-CoV-2 mutations on host phenotype (e.g. disease outcome: asymptomatic, recovery after hospitalization, death), laboratory data could easily be imported in R and connected to the Bioconductor infrastructure. Indeed, we have recently published a paper that demonstrates the many advantages of using the VariantExperiment (i.e. a Bioconductor package) class to store, exchange and analyse SARS-CoV-2 genomic data and associated metadata (20). We are actively investigating the use of a Bioconductor package in PANDEM-2.

5.5 Impact and conclusion

In this report, we discussed the laboratory data sources for pandemic management, including the capture and integration of NGS data. The current COVID-19 Pandemic has generated a large amount of laboratory data, and some of it is accessible in a public database. As a result, we chose to use laboratory data generated during the current pandemic as a first case study (including conventional methods such as qPCR as well as NGS data). The data flow, database structure, and model, on the other hand, are easily adaptable to various viral and bacterial diseases. The exploration of the scientific literature, as well as our participation in PANDEM-2 internal meetings and workshops (organised by WHO and dedicated to SARS-CoV-2 WGS), enabled us to understand the current data-flow, identify its main limitations (lack of raw and processed NGS data in public databases, as well as a lack of standardised contextual metadata associated with NGS data), and forecast future (NGS) data-flows.
In this regard, we may now create a database of laboratory (mainly NGS) data and associated metadata that reflects the data that could be available in the event of a future pandemic. As a result, this database will feed WP3 (Situational Awareness), namely the architecture and data model of the PANDEM-2 display.

6 References

1. Tahamtan A, Ardebili A. Real-time RT-PCR in COVID-19 detection: issues affecting the results. Expert Rev Mol Diagn. 2020 May;20(5):453–4.

2. Yüce M, Filiztekin E, Özkaya KG. COVID-19 diagnosis -A review of current methods. Biosens Bioelectron. 2021 Jan 15;172:112752.

3. Mondeja B, Valdes O, Resik S, Vizcaino A, Acosta E, Montalván A, et al. SARS-CoV-2: preliminary study of infected human nasopharyngeal tissue by high resolution microscopy. Virol J. 2021 Jul 18;18(1):149.

4. Antigen-detection in the diagnosis of SARS-CoV-2 infection using rapid immunoassays [Internet]. [cited 2021 Oct 1]. Available from: https://www.who.int/publications-detail-redirect/antigen-detection-in-the-diagnosis-of-sars-cov-2infection-using-rapid-immunoassays

5. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016 Jan;107(1):1–8.

6. Ravi RK, Walton K, Khosroheidari M. MiSeq: A Next Generation Sequencing Platform for Genomic Analysis. Methods Mol Biol Clifton NJ. 2018;1706:223–32.

7. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015 Oct;13(5):278–89.

8. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics Proteomics Bioinformatics. 2016 Oct;14(5):265–79.

9. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017 Sep;35(9):833–44.

10. Chiara M, D'Erchia AM, Gissi C, Manzari C, Parisi A, Resta N, et al. Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities. Brief Bioinform. 2021 Mar 22;22(2):616–30.

11. Charre C, Ginevra C, Sabatier M, Regue H, Destras G, Brun S, et al. Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. Virus Evol. 2020 Jul;6(2):veaa075.

12. Hu T, Li J, Zhou H, Li C, Holmes EC, Shi W. Bioinformatics resources for SARS-CoV-2 discovery and surveillance. Brief Bioinform. 2021 Mar 22;22(2):631–41.

13. Alm E, Broberg EK, Connor T, Hodcroft EB, Komissarov AB, Maurer-Stroh S, et al. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. Eurosurveillance. 2020 Aug 13;25(32):2001410.

14. Velazquez A, Bustria M, Ouyang Y, Moshiri N. An analysis of clinical and geographical metadata of over 75,000 records in the GISAID COVID-19 database [Internet]. 2020 Sep [cited 2021 Sep 30] p. 2020.09.22.20199497. Available from: https://www.medrxiv.org/content/10.1101/2020.09.22.20199497v1

15. Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, et al. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic Acids Res. 2012 Jan;40(Database issue):D57–63.

16. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, et al. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 2005 Jan 1;33(Database Issue):D325–8.

17. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, et al. Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. Antimicrob Agents Chemother. 2019 Oct 22;63(11):e00483-19.

18. recommendation_covid19_monitoring_wastewaters.pdf [Internet]. [cited 2021 Oct 4]. Available from: https://ec.europa.eu/environment/pdf/water/recommendation_covid19_monitoring_wastewaters.pdf

19. 2021-01-13_Variable_Dictionary_and_Disclaimer_weekly_testing_data_EUEEAUK.pdf [Internet]. [cited 2021 Oct 4]. Available from: https://www.ecdc.europa.eu/sites/default/files/documents/2021-01-13_Variable_Dictionary_and_Disclaimer_weekly_testing_data_EUEEAUK.pdf

20. Ambroise J, Gatto L, Hurel J, Bearzatto B, Gala J-L. On the many advantages of using the VariantExperiment class to store, exchange and analyze SARS-CoV-2 genomic data and associated metadata [Internet]. 2021 Apr [cited 2021 Sep 30] p. 2021.04.05.438328. Available from: https://www.biorxiv.org/content/10.1101/2021.04.05.438328v1

7 List of abbreviations

AMR: Antimicrobial Resistance

DDBJ: DNA Data Bank of Japan

EBI: European Bioinformatics Institute

ECDC: European Centre for Disease Prevention and Control

EMBL: European Molecular Biology Laboratory

ENA: European Nucleotide Archive

GISAID: Global Initiative on Sharing Avian Influenza Data

INSDC: International Nucleotide Sequence Database Collaboration

NCBI: National Center for Biotechnology Information

NGS: Next Generation Sequencing

PCR: Polymerase Chain Reaction

PHA4GE: Public Health Alliance for Genomic Epidemiology

POCT: Point Of Care Testing

qPCR: Quantitative Polymerase Chain Reaction

RDT: Rapid Diagnostic Testing

RT-PCR: Reverse Transcriptase Polymerase Chain Reaction

SRA: Sequence Read Archive

TESSy: Member States to the European Surveillance System

VOC: Variants of Concern

VOI: Variants of Interest

WGS: Whole Genome Sequencing

WHO: World Health Organization

8 List of figures