

# Deliverable D1.1

## Extended workflows

<b>Project Title</b> (grant agreement No)	Beyond-COVID Grant Agreement 101046203		
<b>Project Acronym</b> (EC Call)	BY-COVID HORIZON-INFRA-2021-EMERGENCY-01		
<b>WP No &amp; Title</b>	WP1 <b>D1.1 Extended workflows</b> Details of the first wave of extensions and developments upon viral data processing computational workflows operating within the SARS-CoV-2 Data Hubs.		
<b>WP Leaders</b>	Guy Cochrane (EMBL-EBI), Clara Amid (EMC)		
<b>Deliverable Lead Beneficiary</b>	EUROPEAN MOLECULAR BIOLOGY LABORATORY (ELIXIR/EMBL-EBI)		
<b>Contractual delivery date</b>	30/09/2022	<b>Actual Delivery date</b>	28/04/2023
<b>Delayed</b>	[Yes]		
<b>Partner(s)</b> contributing to this deliverable	EMBL-EBI		
<b>Authors</b>	ELIXIR/EMBL-EBI		
<b>Contributors</b>	Nadim Rahman (EMBL-EBI), David Yu Yuan (EMBL-EBI)		
<b>Acknowledgements</b> (not grant participants)	N/A		
<b>Reviewers</b>	BY-COVID Management Board		



## Table of contents

<b>1. Executive Summary</b>	<b>2</b>
<b>2. Contribution towards project objectives</b>	<b>3</b>
Objective 1	3
Objective 2	4
Objective 3	4
Objective 4	5
Objective 5	5
<b>3. Methods and Description of work accomplished</b>	<b>6</b>
3.1 Overview of Infrastructure	6
3.1.1 Current Status	1
3.1.2 Infrastructure Exploration	7
3.2 Use Cases	8
SARS-CoV-2 Public Data	8
Private Data Hubs	8
MPox Data	8
3.3 System Integration	9
<b>4. Results</b>	<b>10</b>
<b>5. Discussion and Next Steps</b>	<b>10</b>
<b>6. Conclusions</b>	<b>11</b>



# 1. Executive Summary

This deliverable is titled ‘Extended Workflows’, which is detailing the first wave of extensions and developments for viral data processing and computational workflows for operation in the SARS-CoV-2 Data Hubs<sup>1</sup>. Described in greater detail below, the SARS-CoV-2 Data Hubs are a workspace, or toolbox, enabling users to share their sequence data, utilising analysis workflows to process, and then visualisation tools to ingest and visualise data for downstream interpretation. The private version of the data hubs enable for private, pre-publication data sharing amongst a group of collaborators.

To enable for analysis of sequence data via integrated workflows, EMBL-EBI, and specifically the European Nucleotide Archive (ENA), has developed and maintains a set of tools, software and utilises infrastructure. Collectively, this is known as the ENA Pathogen Analysis System, which is described in further detail below. The system presents a hybrid-cloud processing system including a Google BigQuery Database<sup>2</sup>, Looker DataStudio, Nextflow, Docker, LSF cluster, Slurm cluster and Google Cloud Life Sciences API. This is setup with support of BY-COVID and has been in-place to analyse data within use-cases from other projects, such as VEO. Use-cases include public SARS-CoV-2 raw read dataset analysis within the COVID-19 Data Platform, but also extends into private data hub analysis and is in place for future workflows to be integrated into the system, following analysis and feasibility testing of those workflows. The first adaptation to the system has taken place in response to the outbreak of Monkeypox virus (Mpox or MPXV). This provides a solid foundation and infrastructure for future developments as part of the Pathogens Platform.

Overall, this deliverable describes the efforts undertaken to analyse an unprecedented amount of data from an infrastructure view, detailing some of the challenges we overcame to achieve this. Furthermore, the deliverable describes the ENA Pathogen Analysis System.

---

<sup>1</sup> <https://www.covid19dataportal.org/data-hubs>

<sup>2</sup> <https://cloud.google.com/bigquery>



## 2. Contribution towards project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives/key results:

	Key Result No and description	Contributed
<b>Objective 1</b> Enable storage, sharing, access, analysis and processing of research data and other digital research objects from outbreak research	1.A research data management practice in European research infrastructures practice that drives discovery, access and reuse of outbreak data and directly links experimental data from HORIZON-INFRA-2021-EMERGENCY-02 transnational access projects into the COVID-19 Data Portal	No
	2. Workflows and processing pipelines that integrate transparent quality management and provenance and are openly shared.	Yes
	3. Research infrastructures on-target training so that users can exploit the platform.	No
	4. Engagement so that stakeholders (RI, national centres, policy makers, intergovernmental organisations, funders and end-users) incorporate FAIR and open data in infectious disease guidelines and forward planning.	No
<b>Objective 2</b> Mobilise and expose viral and human infectious disease data from national centres	1.A comprehensive registry of available data with established procedures to collate data governance models, metadata descriptions and access mechanisms in a pandemic scenario.	Yes
	2. Mechanisms for the initial discovery across data sources based on available metadata at the reference collection.	No
	3. Demonstrated transnational linking of real-world data from national surveillance, healthcare, registries and social science data that allow the assessment of variants to serve the research needs of epidemiology and public health.	No
	4.Demonstrated assessment of emerging SARS-CoV-2 variants against data generated in the on-going European VACCELERATE clinical trials project to investigate vaccine efficacy.	No



<b>Objective 3</b> Link FAIR data and metadata on SARS-CoV-2 and COVID-19	1. A platform that links normative pathogen genomes and variant representations to research cohorts and mechanistic studies to understand the biomolecular determinants of variant response on patient susceptibility, and disease pathways.	No
	2. An open and extensible metadata framework adopted cross-domain that supports comprehensive indexing of the infectious disease resources based on mappings across resources and research domains.	No
	3. A provenance framework for researchers and policy-makers that enables trust in results and credit to data submitters, workflow contributors and participant resources.	No
<b>Objective 4</b> Develop digital tools and data analytics for pandemic and outbreak preparedness, including tracking genomics variations of SARS-CoV-2 and identifying new variants of concern	1. Broad uptake of viral Data Hubs across Europe deliver an order-of-magnitude increase in open viral variant detection and sharing.	No
	2. Infrastructure and quality workflows mobilised and shared to produce open, normative variant data that is incorporated into national and regional data systems and decision making.	No
<b>Objective 5</b> Contribute to the Horizon Europe European Open Science Cloud (EOSC) Partnership and European Health Data Space (EHDS)	1. Guidelines and procedures for FAIR data management and access will be established, building on work of other guideline producing consortia such as the Global Alliance for Genomics and Health (GA4GH), the 1Mio Genomes Initiative (1MG) and the Beyond One Million Genomes project (B1MG).	No
	2. Services, software, protocols, guidelines and other research objects that are openly accessible for reuse by the EOSC Association and the community at large as a foundation for European preparedness for infectious diseases, leveraging developments in EOSC-Life, SSHOC, EOSC-Future, EGI-ACE and other EOSC	No



	projects.	
	3. Alignment (both policy and implementation routes) will have been achieved between the data governance strategies for routinely collected health data in the EHDS initiative, including the TEHDAS Joint Action and future EHDS Pilot Actions.	No
	4. To empower national centres to build capacity and train platform users and data providers (e.g., from life, social or health sciences), and with experts from across partner institutions collaborating to create training materials for the identified gaps, and to exchange experiences and knowledge.	No

## 3. Methods and Description of work accomplished

### 3.1 Overview of Infrastructure

#### 3.1.1 Current Status

The ENA Pathogen Analysis System is currently, mainly automatically, processing SARS-CoV-2 raw reads as part of the systematic analysis of raw reads in the COVID-19 Data Portal<sup>3</sup> with the Horizon 2020 VEO project<sup>4</sup>. Through the same project, the system also provides support for private data hub processing. This provided an initial use-case to test the system's applicability to private data, which resulted in adaptations. Finally, through the MPox outbreak in May 2022, pipelines and infrastructure developed for COVID-19 were repurposed to process MPox data. The analysis system was used to automatically process all MPox raw reads submitted to the ENA and the International Nucleotide Sequence Database Collaboration (INSDC), and like with COVID data, output is archived under a project PRJEB55834<sup>5</sup>, and made available in the Pathogens Portal - Systematic Analyses<sup>6</sup>. Overall, this showcases the system's wide applicability, initially developed for public COVID-19 data, expanded to handle private data, and then data from a different domain.

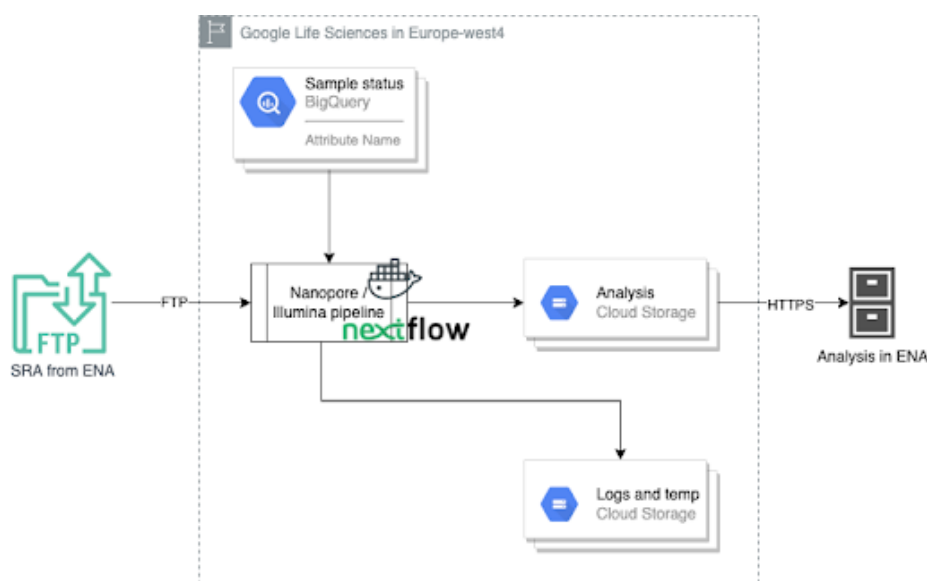
<sup>3</sup><https://www.covid19dataportal.org/search/sequences?crossReferencesOption=all&overrideDefaultDomain=true&db=sra-experiment-covid19&size=15>

<sup>4</sup><https://www.veo-europe.eu/>

<sup>5</sup><https://www.ebi.ac.uk/ena/browser/view/PRJEB55834>

<sup>6</sup><https://www.ebi.ac.uk/ena/pathogens/v2/monkeypox?db=sra-analysis-mpox&size=15#search-content>





**Figure 1.** The main components of the ENA analysis management system. This depicts technologies and general flow of data within the system, culminating in analysis outputs being archived in the ENA on the right hand side.

The system is depicted in the diagram above and consists of Nextflow and Docker as the runtime, BigQuery as the queuing system, Looker DataStudio as the monitoring dashboard, and cloud storage as the intermediate file store. The metadata of raw reads is loaded and tracked in BigQuery. The samples are downloaded from ENA, analysed in the pipelines and submitted back to ENA.

### 3.1.2 Infrastructure Exploration

Processing the wealth of SARS-CoV-2 raw reads in the COVID-19 Data Portal brought several challenges, in particular understanding the most efficient way of processing 5.5 million (at the moment of writing) of publicly submitted datasets and then archiving the results back in the system. Due to the size of the dataset, a backlog of analysis exists, along with the continued submission of datasets on a daily basis.

To help push through with a burst of analyses, and attempting to reduce the backlog, Google Cloud Platform (GCP) was explored. This resulted in porting of workflows into GCP and setup of processing and storage, requiring some investigative work. Once set up, this enabled EMBL-EBI to process 450,000 datasets in one burst of analysis going into 2022. Due to the low processing time and high threading, the concern then shifted to submission of datasets back into the ENA. As a result, through use-cases in other projects, several bottlenecks in the ENA submission API, Webin, were identified and fixed. Overall, this presented a useful method in analysing a large amount of data, providing a proof of concept to utilise in the future, should it be required.



Due to the higher cost of external cloud processing, the pipelines were then ported and set up on High-Performance Compute Clusters (HPCCs) at EMBL-EBI. With the federated computing architecture, the analysis has been performed in GCP and one or more HPCCs concurrently. This enables the processing of large amounts of data to be processed on a weekly basis, currently standing at roughly 60,000 runs processed per week. It would take 33 weeks to clear a backlog of 2 million datasets, assuming all new submissions had stopped. This highlights the need for cloud funding being made available in crisis times to allow scale-out in hybrid mode. From the infrastructure exploration, the system has become a hybrid-cloud processing system, including HPCC and cloud solutions.

## 3.2 Use Cases

### SARS-CoV-2 Public Data

The system automatically processes SARS-CoV-2 raw reads via an integrated read mapping workflow, developed under the VEO project. The codebase for the integrated pipeline can be found here<sup>7</sup> and has been registered on WorkflowHub:

<https://workflowhub.eu/workflows/105>, it generates unfiltered variant calls, filtered variant calls and consensus sequences for every Illumina or Oxford Nanopore raw read processed. The output is archived at the ENA under the umbrella project PRJEB45555<sup>8</sup>, and fed into the EBISearch indexing system to support for presentation in the COVID-19 Data Portal - Systematic Analyses<sup>9</sup>.

### Private Data Hubs

For the analysis of private data hubs, a specific data hub at the ENA - dcc\_walton was utilised in order to expand the system to handle private data. This resulted in the analysis workflow, available for public COVID-19 data (as mentioned above in the systematic analysis), being tested and available for private data. The resulting output was also fed back to the data hub, to the appropriate projects, that were linked to the data hub.

### MPox Data

As mentioned above, the MPox outbreak provided an opportunity to expand the system beyond COVID-19. Publicly submitted MPox raw reads were analysed automatically, and like with COVID data, output was archived under a project PRJEB55834<sup>10</sup>, and made

<sup>7</sup> <https://github.com/enasequence/covid-sequence-analysis-workflow>

<sup>8</sup> <https://www.ebi.ac.uk/ena/browser/view/PRJEB45555>

<sup>9</sup> <https://www.covid19dataportal.org/search/sequences?crossReferencesOption=all&overrideDefaultDomain=true&db=sra-analysis-covid19&size=15>

<sup>10</sup> <https://www.ebi.ac.uk/ena/browser/view/PRJEB55834>





available via the Pathogens Portal - Systematic Analyses<sup>11</sup>. This presented the flexibility of the system, with some alterations and extensions.

There are two processing modes in the system: batch mode and transaction mode. For a smaller number of samples with the fixed size, up to tens of thousands, such as MPox or private SARS-CoV-2, batch mode is used to process the input. The transaction mode has been developed to handle the extremely large data volume of 5.5 million runs, and ever increasing. The cloud-native data warehousing service BigQuery is used as the queuing system to track the metadata of analysis objects and the change of their status through the pipelines. This solves the scalability and performance problems and has expanded the capacity of the system by a thousand fold.

This global queuing system, together with the streaming input and output via the public APIs by ENA, has enabled distributed computing so that multiple HPCCs can process SARS-CoV-2 genomes concurrently. We have designed and implemented the pipelines portable between GCP and HPCC with Docker and Nextflow. We are able to process samples in the cloud and on HPC concurrently so the federated computing is further expanded into a hybrid cloud implementation.

### 3.3 System Integration

The system utilises Nextflow<sup>12</sup> to orchestrate, process and manage pipeline processing. Therefore, if pipelines are not already written in or wrapped with a Nextflow execution script, then this is an adaptation that is required before integration into the ENA Pathogen Analysis System. So far, the system has integrated pipelines written in Python and natively in Nextflow, and it also allows for pipelines written in other programming languages to be integrated, e.g. Java or C.

Nextflow enables for individual processes to be created, mapping to major tasks that the pipeline executes. This provides an opportunity to define parameters and in particular, compute requirements for individual processes, resulting in a more efficient pipeline, which does not require excess resources. Nextflow also provides a neat way of tracking and logging processing.

The process runtime must be containerized, which should be implemented as Docker containers for the maximum portability. It is not required but highly recommended to integrate with ENA via ENA file downloader and ENA analysis submitter for input and output from and to ENA. For large scale analysis from hundreds of thousands to millions of input samples, the integration with BigQuery as a metadata-based queuing system is also needed.

<sup>11</sup><https://www.ebi.ac.uk/ena/pathogens/v2/monkeypox?db=sra-analysis-mpox&size=15#search-content>

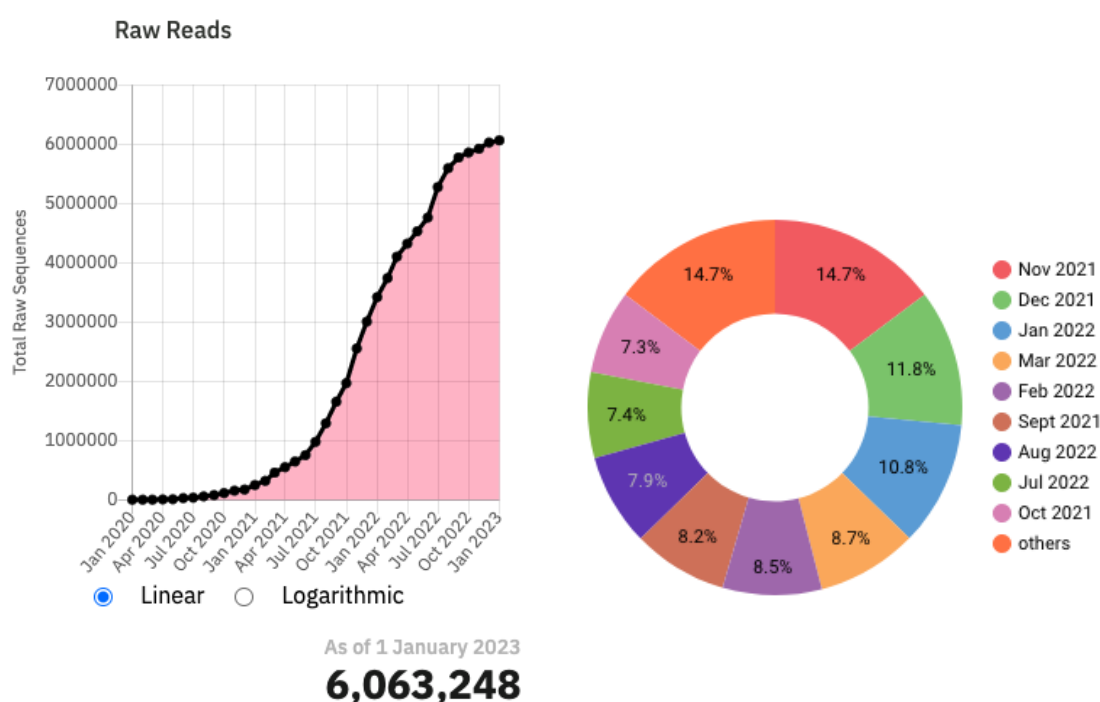
<sup>12</sup> <https://www.nextflow.io/>



There are three execution engines, or executors in Nextflow's terminology supported by the system. They are IBM Load Sharing Facility, Slurm and Google Cloud Life Sciences API. This provides flexibility to execute the workflows in multiple different HPCCs and Google cloud. The open architecture of the system allows the easy addition of other execution engines with configurations externalised to adapt to new HPCCs. There is a collaboration under a separate project to expand the system onto a new ELIXIR node in EuroHPC in Finland.

## 4. Results

The system has finished analysing around 3.8 million raw SARS-CoV-2 read datasets, which generated the analysis objects with the total size of 200 TB as of mid January 2023. A truly big-data project, which continues to run autonomously, aiming to cut the backlog of datasets that haven't been analysed as of yet. In total, as of mid-January 2023, there are approximately 2 million additional datasets (that are appropriate for the pipelines) to analyse. Submissions of new datasets continue, but not at the rate that EMBL-EBI had seen during the winter - spring of 2021 - 2022, where multiple submitting centres were sharing large amounts of COVID-19 raw reads consistently on a weekly basis (fig.2).



**Figure 2.** Number of raw SARS-CoV-2 reads shared from the beginning of the COVID-19 pandemic until January 2023.



The system has also processed datasets as part of the private data hub dcc\_walton, this includes >6,500 runs across several projects. Finally, as of mid January 2023, 1,195 raw MPox reads were processed and presented in the Pathogens Portal.

## 5. Discussion and Next Steps

The ENA Pathogen Analysis System (PAS), described above, has been developed to enable high-throughput processing of large and continuously growing datasets. This has been demonstrated from the systematic analysis of SARS-CoV-2 raw read datasets, currently standing at >3.8 million datasets<sup>13</sup>. This system required exploration into big data processing and in particular how a hybrid-based processing approach is key in big data analysis. This has been a crucial step to move forward in cases of future outbreaks, or big data analyses. The engineering work behind these efforts is complex and included many factors that were taken into consideration: number of datasets, size of dataset (in Tera/Peta-bytes), processing memory requirements, storage requirements, process logging, dataset archival, serving data to frontend browser - just to name a few aspects. The analysis system has also been extended to include private data analysis (as part of the private data hubs) and systematic analysis of MPox, which, as mentioned above, demonstrates the flexibility of the system.

Looking forward, should a new outbreak arise, the management system must be flexible and adaptable, to be able to handle new use-cases. With MPox, this was tested and the system was successfully adapted within a relatively short time frame of roughly 2 weeks to systematically analyse data. This aided the definition of the main steps required, including creation of ENA project(s) for analysis products to be stored in. Greater clarity on the analysis tools that can be integrated into the system may be something that is worth considering. This would help detail requirements the system may have for external pipelines. We continue to analyse SARS-CoV-2 raw datasets, with the aim of clearing the backlog of data, following submission of raw datasets. This has already resulted in a large and comprehensive dataset.

The ENA PAS is a component within the SARS-CoV-2 Data Hubs, which handles the analysis processing and distributed compute aspect for both public and private data. This is being expanded under other projects to encompass other types of data, for example Pathogen Data Hubs (to extend further than SARS-CoV-2), as part of the Pathogens Platform<sup>14</sup>.

<sup>13</sup><https://www.covid19dataportal.org/search/sequences?crossReferencesOption=all&overrideDefaultDomain=true&db=sra-analysis-covid19&size=15>

<sup>14</sup><https://www.ebi.ac.uk/ena/pathogens/v2/>



## 6. Conclusions

The ENA Pathogen Analysis System consists of several components, tools and technologies, which together provide a system that pulls, processes, logs and archives data. This system is crucial for the SARS-CoV-2 Data Hubs (private and public), and further Pathogen Data Hubs in general. To achieve the current system, use-cases from the VEO project including systematic SARS-CoV-2 raw read analysis, private SARS-CoV-2 data hub setup and systematic MPox raw read analysis, have driven its development. Overall this is a flexible and reusable piece of infrastructure, which enables future big data analysis.

