WP	15									Lead Beneficiary								EMBL			
WP Title	COVID-19 response																				
Participant No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
Participant short name	EMC	DTU	FLI	EMBL	APHA	ELTE	RIVM	IFREMER	AUTH	AMC	IP	CSIC	EPFL	CBS	UEDIN	UNIBO	UNIPD	UU	UH	UC	
Person months	38	25	0	89	0	24	24	0	0	0	0	0	0	0	0	0	0	0	0	0	
Start Month			6							End Month					26						

Objectives

WP15 covers project objective 8: To provide a suite of analytical tools, storage, and data sharing workspace to facilitate the sharing, analysis and reuse of raw and annotated SARS-CoV-2 genomic data.

15.1 Mobilise analysis;
15.2 Mobilise data;
15.3 Enhance access.

Description of Work and role of partners

EMBL (lead), **EMC** (co-lead), RIVM, DTU, ELTE

Task 15.1 Mobilise analysis (months 6 – 26)
Partners: EMBL, EMC, RIVM, DTU, ELTE

In this Work Package, we respond in a timely way to the challenges brought by the COVID-19 pandemic and specific requirements from scientists in the area of virus research. This is granted by the active partnership between the members of this work package who are currently producing and analysing genomics data from the ongoing pandemic and through their direct links to the public health sector (Figure X). This work constitutes a part of the European COVID-19 Data Platform (https://www.covid19dataportal.org/).

Sequence data form an essential foundation for broad and full investigation of infectious disease outbreaks. Currently, due to the lack of a single platform, supporting all sequence data types - raw sequencing data as well as consensus sequences - and the inclusion of various sequence processing and analysis tools used in this sector, different parts of this process happen in different places with an impact on timing of the research, surveillance and outbreak response. In this work package, we will deliver "SARS-CoV-2 Data Hubs", built upon the established "COMPARE Data Hubs", for which ongoing development and delivery is scheduled within work package 1, but here we will have a specific focus on enabling rapid SARS-CoV-2 data analysis driven by VEO partners to respond to the ongoing COVID-19 pandemic and to deliver a platform that could be used in potential future scenarios. The current default is the release of annotated whole genomes through GISAID, a platform that also has a wide network of curators. Here, we deploy to store raw data for re-use for the wider research community which could be complementary to the GISAID platform. However, we note that adequate cross-references between the databases would be required to link associated samples for the raw and consensus sequences served from these databases. Therefore, we aim for the COVID-19 response platform to enable users to provide both raw data and consensus sequences into the Data Hubs. In addition, we provide access to analytics, visualisation and statistical tools for collaborative

projects, and work on scalability of the rapidly expanding data collections. We will integrate some existing and new analytical tools developed and used by partners and offer SARS-CoV-2 Data Hubs to groups responsible for generating SARS-CoV-2 sequence data. Through integration of analytical tools that are known to be useful to the virology community we will create an environment that we endeavor will promote an open data sharing mentality, helping with Task 15.2 (mobilise data). An important aim of this partnership is to develop a platform that would also cater for countries and institutions with less established bioinformatics infrastructure. A critical component here will be to improve ease of data submissions to the European Nucleotide Archive (EMBL-EBI) for institutions without bioinformatics experience and/or support (mobilise data globally), and incorporation of tools to remove human genome sequences during the upload process. Offering computational resources through which raw sequence data can be processed also bypasses the need for essential bioinformatic skills for the processing of sequence data, lowering the barrier for smaller labs and countries lacking required infrastructure to process their data. We envisage that this would lead to a better geographical distribution of SARS-CoV-2 sequences in the database.

The SARS-CoV-2 Data Hubs will provide systematic data processing and analysis based on existing workflows such as Jovian, with its newly developed SARS-CoV-2 extension, currently covering data from the Illumina sequencing platform, developed by the National Institute for Public Health and the Environment (RIVM) and a pipeline for the analysis of Nanopore data, developed by the Erasmus Medical Centre (Erasmus MC) in the Netherlands. Jovian includes a visualisation component for its analysis results which we will implement into the platform, but additional visualisation tools will be provided through the Kooplex system (curated and operated by Eötvös University in Hungary) as well as the EBI compute infrastructure. Further workflows from partners will be developed and added as appropriate according to incoming requirements for the data. Linked phylogenetic analysis tools will be delivered by the Technical University of Denmark (DTU). We will also maintain an instance of of the NextStrain toolkit (https://nextstrain.org/) within the SARS-CoV-2 Data Hubs. While EMBL-EBI provides almost all of the physical infrastructure that serves the data hubs, our partner institutions are involved in the operation and improvement of the system through access to tenancies in our Embassy cloud compute facility.

The priorities in Task 15.1 will be as follows:

1. Improvements to basic functionalities: ease of data submission, automatic analysis using Jovian and a nanopore workflow at the start of the WP15 work; future workflows may follow.
2. Develop a way to cater for the viral bioinformatics tool builders, allowing deployment of tools from within the platform.
3. Develop more advanced viral analytics to track SARS-CoV-2

Task 15.2 Mobilise data (months 6 – 26)
Partners: EMBL, EMC, RIVM, DTU, ELTE
The SARS-CoV-2 Data Hubs will support comprehensive sequence data across all platforms/sequencing strategies and along the data life cycle. Raw data, in the form of sequence "reads", will be the primary data input into the system for many public health operations. We will operate high-throughput systematic computational processing and analysis of raw data using VEO partner-developed workflows to make available consensus/assembled sequence including visualisation through the use of evergreen trees. We will support consensus/assembled sequences provided both from our computations and generated directly by data providers, in cases where capacity, interest and expertise exist. Data generated directly by data providers will enter the system through standardised formats that are developed and adopted by the EMBL-EBI system.

As outlined above, we envisage that the enhanced analysis tools mobilisation will have an indirect effect on data mobilisation globally since it will offer added incentive to share data through the platform. In collaboration with existing data brokers to ENA and our partners in the INSDC, our system will be open globally. Here we will also expand the portfolio of ENA's existing submission tools to cater for research facilities without bioinformatics expertise and support.

While the focus of the SARS-CoV-2 Data Hubs will be sequence data, these will be contextualised. We will promote essential metadata, such as sampling tracking identifiers, sampling time, geographical location, method of sampling, health status of host and sequencing platform/strategy, alongside sequence data into ENA and the Data Hubs. Alone, these metadata will enable a great many uses of the data, such as geospatial analysis, evolutionary studies, transmission and hotspot investigations. However, there are national legal barriers which are currently preventing some nations to provide the level of metadata outlined here. For this, we will develop interfaces that allow combined analysis of shared data and private data. This will enable researchers to perform their data analysis with rich metadata for a meaningful interpretation of data, at the same time it will allow researchers to exclude sensitive metadata when releasing data into the public domain.

We will offer intensive user support to reduce barriers to sharing of data through the SARS-CoV-2 Data Hubs. This support will focus on data submissions, for which we will be offering our portfolio of existing submission tools and interfaces as appropriate for the different local contexts, e.g. bioinformatics expertise, volume of data to be transferred, that our data providers will be experiencing. Support work will include help desk, training, adaptation of tools and services better to suit emerging requirements.

Task 15.3 Enhance access (months 6 – 26)
Partners: EMBL, EMC, RIVM, DTU, ELTE
We will enhance access points to data in the system, providing tools and support, for example, for VEO partners and those in their networks around the SARS-CoV-2 Data Hubs. During the lifetime of the project we expect to add new partners with specific expertise to the SARS-CoV-2 Data Hubs environment.

The anticipated specific roles of existing partners will be:

- EMBL-EBI: provision of ENA and its services including data management and user support and full integration of analytical tools developed by partners (see below). The system will be dynamic between users and developers to incorporate improvements in terms of usability.
- RIVM and Erasmus MC: development, improvement of the computational workflow that provides systematic processing and analysis; data standards development; viral alignment methods. EMC and RIVM will test all components of the system and provide feedback to EMBL-EBI. RIVM will help integrate the Jovian pipeline for Illumina paired-end data which has two functions:
    - 1) processing metagenomic data into viral scaffolds with rich annotation and visualisations. This includes sub-species level virus typing of several pathogenic viruses, while this does not yet include SARS-CoV-2, it does help identify off-target pathogens.
    - 2) generating a SARS-CoV-2 consensus sequence, including visualisations for manual curation of the generated consensus sequence.
- Eötvös Loránd University: curation and operation of additional data exploration and visualisation notebooks that appear to users in the Pathogen Portal.
- EMC/RIVM/DTU: development of phylogenetic-analysis software that solves SARS-CoV-2 specific challenges, including detection of deletions and insertions; development and operation of

phylogenetic tree-visualisation software. This will bring together expertise from EMC producing and analysing a large volume of data, RIVM providing the public health perspective with limitations in data access and DTU's resources and experience developing the required software.

- RIVM/EMC/EMBL-EBI: develop a workflow/legal framework to handle the sensitive metadata required for actionable phylogeny and outbreak tracing. This extends to the problem of removing human (patient) genetic material from any uploaded data.

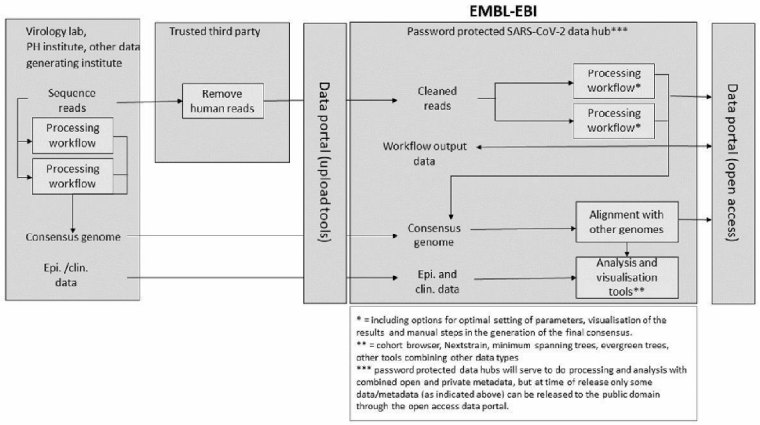

Figure X: shows the envisaged interaction between the Virology, Public Health (PH) and other labs with the EMBL-EBI system, from raw read generation to filtering human reads to data analysis through various pipelines

Deliverables

D15.1 Report on raw sequence data processing workflow (month 7; lead beneficiary EMBL)
D15.2 Report on data mobilisation (month 8; lead beneficiary EMBL)
D15.3 Report on phylogenetic tools and enhanced results visualization (month 9; lead beneficiary DTU)
D15.4 Report on the progress of data mobilisation (month 18; lead beneficiary EMBL)
D15.5 Report on genotype to phenotype translations (month 24; lead beneficiary EMC)
D15.6 Report on data access tools and support (month 24; lead beneficiary EMBL)
D15.7 Report on data mobilisation summary (month 26; lead beneficiary EMBL)

Milestones

MS41 Open source production release of nanopore analysis workflow (month 7; lead beneficiary EMC)
MS42 Second prototype for submission tool to support users without bioinformatics expertise (e.g. a Drag&Drop tool) (month 7; lead beneficiary EMBL)
MS43 Publicly available read filtering software (month 8; lead beneficiary EMBL)

MS44 Open source production release of Jovian including merged SARS-CoV-2 module (month 9; RIVM)

Deliverable description

D15.1 Report on raw sequence data processing workflow (month 7; lead beneficiary EMBL)
Systematic and autonomous processing and analysis of incoming data system, using Jovian, with full results presentation through COVID-19 platform. A nanopore analysis workflow to support the most frequently used platform currently.

D15.2 Report on data mobilisation (month 8; lead beneficiary EMBL)
Summary of mobilised data to date; a tool to show statistics relating to the size of the growing data set.

D15.3 Report on phylogenetic tools and enhanced results visualization (month 9; lead beneficiary DTU)
Improved navigation and visualisation tools for systematic processing and analysis, including multiple alignments and phylogenetic trees.

D15.4 Report on the progress of data mobilisation (month 18; lead beneficiary EMBL)
Summary of mobilised data to date; a tool to show statistics relating to the size of the growing data set.

D15.5 Report on genotype to phenotype translations (month 24; lead beneficiary EMC)
More advanced viral analytics to track SARS-CoV-2.

D15.6 Report on data access tools and support (month 24; lead beneficiary EMBL)
Enhanced access to SARS-CoV-2 data supporting tools and support functions.

D15.7 Report on data mobilisation summary (month 26; lead beneficiary EMBL)
Summary of mobilised data to date, highlighting data platforms, formats, levels of assembly and national contributions.