Deliverable D8.3 Raw sequence data processing workflow in operation

Project Title (Grant agreement no.):	ELIXIR-CONVERGE: Connect and align ELIXIR Nodes to deliver sustainable FAIR life-science data management services (871075)		
Project Acronym (EC Call):	ELIXIR-CONVERGE (H2020-INFRADEV-2018-2020)		
WP No & Title:	WP8 ELIXIR-CONVERGE European COVID-19 Data Platform		
WP leader(s):	Guy Cochrane (EMBL-EBI)		
Deliverable Lead Beneficiary:	1 - EMBL-EBI		
Contractual delivery date:	30/11/2020	**Actual delivery date:**	27/07/2021
Delayed:	Yes		
Partner(s) contributing to this deliverable:	EMBL-EBI		
Authors: Guy Cochrane (EMBL-EBI) **Contributors:** Nadim Rahman (EMBL-EBI), Alexey Sokolov (EMBL-EBI) **Acknowledgments (not grant participants):**			
Reviewers:	ELIXIR-CONVERGE Management Board (MB) members.		

Log of changes

DATE	Mvm	Who	Description
30/09/2020	0v1	Guy Cochrane (EMBL-EBI)	Initial version
16/07/2021	0v2	Marianna Ventouratou (EMBL-EBI)	Sent to PMU after incorporating internal WP feedback
16/07/2021	0v3	Nikki Coutts (ELIXIR Hub)	Circulated to the MB for final review before submission
27/07/2021	1v0	Nikki Coutts (ELIXIR Hub)	Final version to be uploaded into EC Portal

Table of contents

1. Executive Summary

This deliverable describes raw sequence data processing workflows available for SARS-CoV-2 data hub users and systematic analysis of public data in the COVID-19 Data Portal (DP). The SARS-CoV-2 data hubs are toolboxes and spaces for users to share data (in a pre-publication or public manner), in some cases with collaborators based in different institutes, automatically process the shared data (through the data hub configuration), with resulting analysis returned back to the data hub for interpretation by users and their collaborators.

To carry out systematic analysis of raw datasets held in the COVID-19 DP, workflows developed by collaborators (and shared to GitHub publicly) have been integrated into the ENA analysis management system, a component of the data hubs system involved in coordination and processing of data within a data hub via integrated workflows via Embassy cloud. Two integrated workflows developed under the VEO project provide a means to generate consensus sequences from raw datasets. These are the Nanopore Analysis Workflow (NAW) developed by Erasmus MC (EMC),

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

Netherlands, and Jovian in reference alignment mode developed by National Institute for Public Health and the Environment (RIVM), Netherlands.

A workflow developed under VEO by Eötvös Loránd University (ELTE), Hungary - COVID-19 Sequence Analysis Workflow, has been integrated and adapted to generate variant calls from raw datasets. This workflow is again available to data hub users, however has been involved in systematically producing variant calls from raw datasets, with data products from the processing archived within PRJEB43947.

There are currently five data hubs that have been assigned, and configured to workflows, with discussions for several more ongoing.

2. Contribution toward project objectives

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

Objective no. / Key Result no. Description	**Contributed to:**
Objective 1: Develop a sustainable and scalable operating model for transnational life-science data management support by leveraging national capabilities **(WP1, WP5)**	
Key Result 1.1: Established European expert network of data stewards that connect national data centres and similar infrastructures and drive the development of interoperable solutions following international best practice, including national interpretations of the General Data Protection Regulation (GDPR)	**No**
Key Result 1. 2: Development of joint guidelines and common toolkit that are adopted into funder recommendations, with support available nationally and in local languages	**No**
Key Result 1.3: The catalogue of successful national business models incorporated into national strategies	**No**
Key Result 1.4: The developed “sustainable and scalable operating model for transnational life-science data management support” is adopted into national ELIXIR Node	**No**
Objective 2: Strengthen Europe’s data management capacity through a comprehensive training programme delivered throughout the European Research Area **(WP2, WP6)**	
Key Result 2.1: A comprehensive ELIXIR Training and Capacity building programme in Data Management, directed at both data managers and ELIXIR	**No**

This project has received funding from the European Union’s Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

users, and connected to the national training programmes in Data Management in the ELIXIR Nodes and prospective ELIXIR Member countries.	
Key Result 2.2: Development of a collective group of trainers that support scalable deployment of Data Management training across ELIXIR Nodes.	**No**
Key Result 2.3: A substantial cohort of data managers, Node coordinators and researchers with specific data management skills, business planning and knowledge of transnational operations across the ELIXIR Nodes	**No**
Objective 3: Align national data management standards and services through a sustainable, scalable and cost-effective data management toolkit **(WP2, WP3, WP5**)	
Key Result 3.1: Assemble a full-stack harmonised common toolkit comprising all aspects of data management: from data capture, annotation, and sharing; to integration with analysis platforms and making the data publicly available according to international standards.	**No**
Key Result 3.2: Provide exemplar toolkit configurations for prioritised demonstrators to serve as templates for future use.	**No**
Key Result 3.3: Establish national capacity in using as well as updating, extending and sustaining the toolkit across the ERA.	**No**
Key Result 3.4: Enable 'FAIR at source' practice for data generation, and analytical process pipeline implementation by flexible deployment of the toolkit in national operations	**No**
Objective 4: Align national investments to drive local impact and global influence of ELIXIR **(WP4,WP6**)	
Key Result 4.1: Development of a Node Impact Assessment Toolkit based on RI-PATHS methodology.	**No**
Key Result 4.2: Adoption of Impact assessment in ELIXIR Nodes, supported by Node coordinators network and feedback on applicability from dialogues with national funders.	**No**
Key Result 4.3: Creation of national public-private partnerships and industry outreach where open life-science data and services stimulate local bioeconomy	**No**
Key Result 4.4: Growth in reach, impact and engagement of stakeholder communication assessed by established ELIXIR Communications metrics	**No**
Key Result 4.5: Initiating and advancing discussions on Membership (EU and international) or strategic partnerships (international countries) following ELIXIR-CONVERGE workshops.	**No**
Objectives - WP8 - ELIXIR-CONVERGE European COVID-19 Data Platform	

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

O8.1 Data management support for EU projects (Task 8.1)	**Yes**
O8.2 Mobilisation of analysis upon SARS-CoV-2 sequence data (Task8.2)	**Yes**
O8.3 Enhanced access to data, tools and support (Task 8.3)	**Yes**

3. Introduction

This deliverable describes raw sequence data processing workflows available for SARS-CoV-2 data hub users and systematic analysis of public data in the COVID-19 Data Portal (DP). Workflows are focused at generating consensus sequences and variant calls from raw read data. This generates a complete data product, with provenance with regards to data owners, data types and processing workflows. The data product is then consumable by the scientific research community, for example epidemiologists who may be interested in the tracking of SARS-CoV-2 infection and specific variants. This deliverable does not cover work on SARS-CoV-2 phylogeny, which has been described in deliverable 8.4, however this is an additional aspect/workflow that is available within the data hubs and utilised for systematically analysis of submitted sequence datasets within the COVID-19 DP.

The SARS-CoV-2 data hubs[1] are toolboxes and spaces for users to share data (in a pre-publication or public manner), in some cases with collaborators based in different institutes, automatically process the shared data (through the data hub configuration), with resulting analysis returned back to the data hub for interpretation by users and their collaborators. In some cases (depending on the workflow used to analyse), interactive Jupyter notebooks are also available for users.

4. Description of work accomplished

4.1 Integrated Workflows

4.1.1 Integration

To carry out systematic analysis of raw datasets held in the COVID-19 Data Portal, workflows developed by collaborators (and shared to GitHub publicly) have been integrated into the ENA analysis management system, a component of the data hubs system involved in coordination and processing of data within a data hub via integrated workflows via Embassy cloud. This system defines the configuration of analysis workflows to data hubs, and in doing so, has been utilised for the systematic analysis of raw datasets. The ENA analysis management system has been described in further detail in deliverable 8.4.

[1] https://www.covid19dataportal.org/data-hubs

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

4.1.1 From raw read to consensus

Two integrated workflows developed under the VEO project provide a means to generate consensus sequences from raw datasets. These are the Nanopore Analysis Workflow (NAW)[2] developed by Erasmus MC (EMC), Netherlands, and Jovian[3] in reference alignment mode developed by National Institute for Public Health and the Environment (RIVM), Netherlands.

As the names suggest, NAW is available to process raw Nanopore amplicon datasets, whereas Jovian has been integrated to process Illumina datasets and generate consensus sequences from SARS-CoV-2 reference-based alignment. Both of these workflows have been utilised in the data hubs system and initially were part of the systematic analysis of raw datasets, before shifting to variant calling, described below.

4.1.1 From raw read to variants and consensus

A workflow developed under VEO by Eötvös Loránd University (ELTE), Hungary - COVID-19 Sequence Analysis Workflow[4], has been integrated and adapted to generate variant calls from raw datasets. Following SARS-CoV-2 reference-based mapping, VCFs are generated via LoFreq (Wilm *et al*, 2021). The workflow processes paired-end Illumina reads, which form the majority of data submissions to the COVID-19 Data Portal.

This workflow is again available to data hub users, however has been involved in systematically producing variant calls from raw datasets, with data products from the processing archived within PRJEB43947[5]. This includes unfiltered VCFs, for users to apply an appropriate cut-off. Further plans include introduction of a consensus sequence generation step within the workflow to generate a complete data set for consumption. Additionally introduction of an allele frequency threshold for the VCFs to generate a filtered and unfiltered set for consumption by users. Finally, adaptations to the NAW are expected to include variant calling in the future.

4.2 Data Hub Examples

There are currently five data hubs that have been assigned, and configured to workflows, with discussions for several more ongoing. These include data hubs for national collaborators who are sharing raw data for processing via the integrated workflows mentioned above. In addition, a data hub of all public data (dcc_grusin) was assigned to support systematic processing of raw datasets.

5. Results

Two integrated workflows from the VEO project were adapted and installed consensus sequences and further information from raw datasets: Nanopore Analysis Workflow (developed by Erasmus MC, The Netherlands) and Jovian (developed by RIVM, The Netherlands). Both of these workflows

[2] https://github.com/dnieuw/ENA_SARS_Cov2_nanopore
[3] https://github.com/DennisSchmitz/Jovian
[4] https://github.com/enasequence/covid-sequence-analysis-workflow
[5] https://www.ebi.ac.uk/ena/browser/view/PRJEB43947

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

have been utilised in the data hubs system and initially were part of the systematic analysis of raw datasets, before shifting to focus on variant calling. A further workflow available to data hub users is the one developed, again under VEO by Eötvös Loránd University (ELTE), Hungary , the COVID-19 Sequence Analysis Workflow, which has been integrated and adapted to generate variant calls from raw datasets.

6. Conclusions

7. Impact

N/A

8. Next Steps

We will disseminate the outputs of the computational workflows through ELIXIR Deposition Databases, including consensus sequence through ENA and variations through the European Variation Archive. In addition, we will make this content available from the COVID-19 Data Portal. We will continue to tune workflow choices and parametrisation and communicate with other analysis groups to move towards harmonised approaches. Finally, we will continue to provide workflows to SARS-CoV-2 Data Hubs, established and new.

9. Deviation from Description of Action

N/A

10. References

Wilm, A., Aw, P. P., Bertrand, D., Yeo, G. H., Ong, S. H., Wong, C. H., Khor, C. C., Petric, R., Hibberd, M. L., & Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic acids research, 40(22), 11189–11201. https://doi.org/10.1093/nar/gks918

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.