Deliverable D8.4 Phylogenetic tools and enhanced results visualisation

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:	28/02/2021	**Actual delivery date:**	27/07/2021
Delayed:	Yes		
Partner(s) contributing to this deliverable:	**EMBL-EBI**		
Authors: Colman O' Cathail (EMBL-EBI), Alexey Sokolov (EMBL-EBI), Nadim Rahman (EMBL EBI), Guy Cochrane (EMBL EBI) **Contributors:** Marianna Ventouratou (EMBL-EBI) **Acknowledgments (not grant participants):** Technical University of Denmark (DTU) team: (Szarvas J, Ahrenfeldt J, Cisneros JLB, et al. Large scale automated phylogenomic analysis of bacterial isolates and the Evergreen Online platform. Communications Biology. 2020 Mar;3(1):137. DOI: 10.1038/s42003-020-0869-5.)			
Reviewers:	ELIXIR-CONVERGE Management Board (MB) members.		

Log of changes

DATE	Mvm	Who	Description
31/12/2020 18/03/2021	0v1	Guy Cochrane Marianna Ventouratou/Nadim Rahman	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	Final version to be uploaded into EC Portal

		(ELIXIR Hub)	

Table of contents

Executive Summary 1

2. Contribution toward project objectives 1

3. Introduction 4

4. Description of work accomplished 4

4.1 SARS-CoV-2 Phylogeny Background 4

4.2 SARS-CoV-2 Phylogeny Architecture 5

5. Results 5

6. Conclusions 6

7. Impact 6

8. Next Steps 6

9. Deviation from Description of Action 6

1. Executive Summary

Deliverable Scope

Deliverable 8.4. "Phylogenetic tools and enhanced results visualisation" aims at improving navigation and visualisation tools for systematic viral data interpretation, including through phylogenetic trees. As part of task 8.2. Data analysis mobilisation (Objective: mobilisation of analysis upon SARS-CoV-2 sequence data), we intended to deploy the data processing and visualisation components of the SARS-CoV-2 Data Hubs system in order to mobilise viral sequence data analysis at scale.

Work accomplished

We produced and integrated into the COVID-19 data portal an interactive phylogenetic tree of SARS-CoV-2 consensus sequences data held in the ENA (and INSDC). This tree is built on the Evergreen/PhyloViz architecture developed by collaborators at the Technical University of Denmark (DTU), and integrated as part of the SARS-CoV-2 data hubs. A sequence TSV file is used by the Evergreen tree back end service. A cron-job retrieves this file daily and adds new samples to the MonoDB database. The system also runs a check to detect suspended samples that should be excluded from the tree, and allows this endpoint to be used by the Evergreen/PhyloViz tree generation service and close the circle of the two services.

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

Conclusion

The tree regularly updates when new SARS-CoV-2 sequences/genomes are shared with the International Nucleotide Sequence Database Collaboration (INSDC), ensuring only public data is included. The tree is accompanied by an appropriate metadata table and a world map illustrating the country of origin of samples represented in the tree. Improvements are being made to the phylogeny include integration of lineage information, variation information and performance updates.

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 users, and connected to the national training programmes in Data Management in the ELIXIR Nodes and prospective ELIXIR Member countries.	**No**
Key Result 2.2: Development of a collective group of trainers that support scalable deployment of Data Management training across ELIXIR Nodes.	**No**

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

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	
O8.1 Data management support for EU projects (Task 8.1)	**No**
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)	**No**

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

3. Introduction

Deliverable 8.4. “Phylogenetic tools and enhanced results visualisation” aims at improving navigation and visualisation tools for systematic viral data interpretation, including through phylogenetic trees.

As part of task 8.2. Data analysis mobilisation (Objective: mobilisation of analysis upon SARS-CoV-2 sequence data), we intended to deploy the data processing and visualisation components of the SARS-CoV-2 Data Hubs system in order to mobilise viral sequence data analysis at scale.

We also envisaged to provide access to a host of computational analysis workflows as required to provide systematic processing according to sequencing library type such as to call variations, provide assembled sequences, provide coding feature annotation or provide phylogenetics analysis. As part of this effort, the scope was to allow for intuitive web data exploration and visualisation environments. This deliverable is closely linked to the VEO project, which has a defined and bounded set of SARS-CoV-2 analyses, including software and visualisation tools already developed as part of VEO's own analysis mobilisation task.

4. Description of work accomplished

4.1 SARS-CoV-2 Phylogeny Background

An interactive phylogenetic tree[1] of SARS-CoV-2 consensus sequences data held in the ENA (and INSDC) has been produced and integrated into the COVID-19 data portal. This tree is built on the Evergreen/PhyloViz architecture developed by collaborators at the Technical University of Denmark (DTU), and integrated as part of the SARS-CoV-2 data hubs.

The Evergreen workflow has been developed and adapted as part of the COMPARE[2] and VEO[3] Europe projects. The workflow has been described in further detail within the publication[4] by Szarvas *et al*, 2020 in Nature, with source code and added documentation in BitBucket[5].

[1] https://www.covid19dataportal.org/phylogeny-tree
[2] https://www.compare-europe.eu/
[3] https://www.veo-europe.eu/
[4] https://www.nature.com/articles/s42003-020-0869-5
[5] https://bitbucket.org/genomicepidemiology/evergreen

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

4.2 SARS-CoV-2 Phylogeny Architecture

EMBL-EBI has provided a virtual machine (UbuntuOS, 16CPU and 128GB of RAM) in Embassy Cloud to support Evergreen/PhyloViz tree generation. On this machine we are running a Nginx server to provide access to all of the samples (correlating to consensus sequences archived at ENA) included in a current tree (http://45.86.170.46/coronavirus_sequence.tsv).

The sequence TSV file is then used by the Evergreen tree back-end service deployed to Kubernetes cluster located on Embassy Cloud. We run a cron-job every night that retrieves this file and adds new samples to the MongoDB database (hosted on Embassy). We expose information from this database through Python/Flask proxy using the following public API end-point - https://phylogeny.covid19dataportal.org/api/phylogeny?size=15. In parallel, the system checks for all of the samples available through EBI-search and compares this list with the list of samples that were added to the tree. This allows us to detect all suspended samples that should be excluded from the tree. We provide a public API-endpoint https://phylogeny.covid19dataportal.org/api/phylogeny_suspended?size=15 with all suspended samples that have to be excluded from the tree. This endpoint is then used by the Evergreen/PhyloViz tree generation service on the virtual machine described above. This closes the circle of two services: Evergreen/PhyloViz tree generation service and Evergreen/PhyloViz tree back-end service, presented below in figure 1.

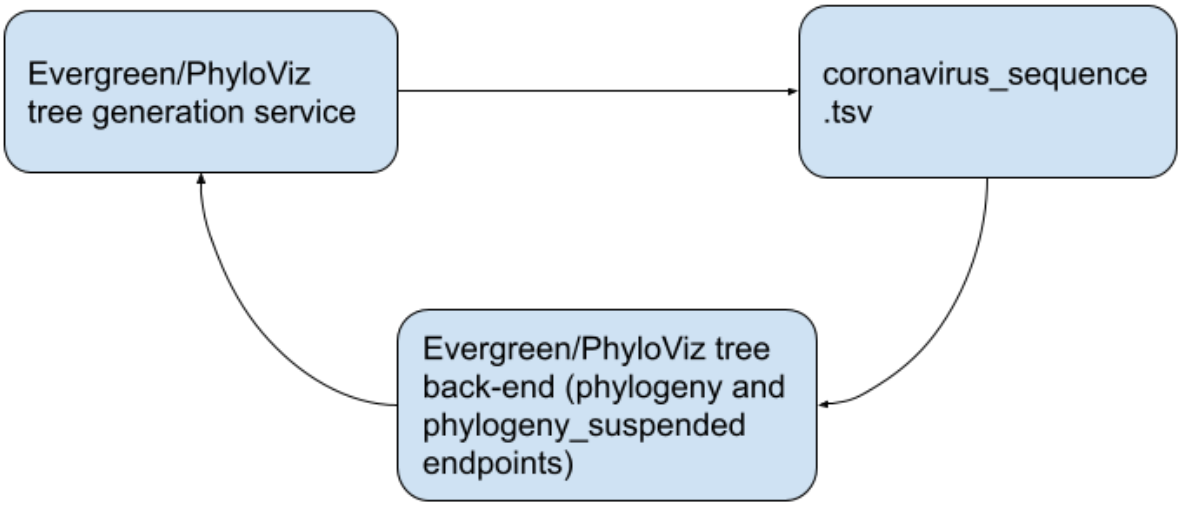

Figure 1. Data flow of integrated Evergreen/PhyloViz phylogeny tree within Embassy cloud infrastructure and SARS-CoV-2 data hubs.

5. Results

The tree regularly updates when new SARS-CoV-2 sequences/genomes are shared with the International Nucleotide Sequence Database Collaboration (INSDC), ensuring only public data is included. The tree is accompanied by an appropriate metadata table and a world map illustrating

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

the country of origin of samples represented in the tree. Improvements being made to the phylogeny include integration of lineage information, variation information and performance updates.

6. Conclusions

In order to mobilise viral sequence data analysis at scale, we produced and integrated into the COVID-19 data portal an interactive phylogenetic tree of SARS-CoV-2 consensus sequences data held in the ENA (and INSDC). The tree is built on the Evergreen/PhyloViz architecture developed and adapted as part of the COMPARE and VEO Europe projects. The tree regularly updates when new SARS-CoV-2 sequences/genomes are shared with the International Nucleotide Sequence Database Collaboration (INSDC), ensuring only public data is included. We will continue to deploy improvements to the phylogeny to include integration of lineage information, variation information and performance updates.

7. Impact

N/A

8. Progress since initial submission

The phylogeny pipeline will be tested for deployment within the SARS-CoV-2 Data Hubs system, initially as a static tree file and metadata with plans to move towards an interactive mode within the Data Hubs.

9. Deviation from Description of Action

Not applicable.

10. References

Szarvas J, Ahrenfeldt J, Cisneros JLB, et al. Large scale automated phylogenomic analysis of bacterial isolates and the Evergreen Online platform. Communications Biology. 2020 Mar;3(1):137. DOI: 10.1038/s42003-020-0869-5.

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