VEO report on mutations and variation in publicly shared SARS-CoV-2 raw sequencing data

Report No. 12 – 20 July 2022

Summary:

- Since the last update, there has been a modification in the workflow of data analyses that are foundational for this report. This was necessary because of the rapidly increasing size of data.

- The CoVEO web-based application that visualizes the SARS-CoV-2 data was embedded to COVID-19 Portal: https://www.covid19dataportal.org/coveo. In this manner, the analyzed datasets that are already uploaded into the CoVEO database are visualized by the app in real time. The application shows basic information about the number of submitted samples visualized by interactive graphs and maps.

- The Pangolin lineage workflow integration has been extended to enable for processing and tagging of Pango lineages for consensus sequences generated: https://www.covid19dataportal.org/sequences?db=sra-analysis-covid19&size=15&crossReferencesOption=all#search-content.

Updates on data submissions

- Update on mobilisation of raw reads, now totaling sequencing data sets from 4,844,657 viral raw read sets from 94 countries, a 14.5% increase since the previous report.

- The variant nomenclature has not been updated in this version of the report as we are currently going through an update of the underlying CoVEO database to increase the throughput and reduce computing time. This was necessary in view of the unprecedented expansion of the sequence data submitted to the public domain. We expect to share updated VOI and VOC information in the next VEO COVID-19 Variants Report.

This work is supported by funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 874735 (VEO).

Background

This report summarizes mobilisation and analysis of SARS-CoV-2 sequence data submitted to the COVID-19 Data Portal in the context of the VEO project (https://www.veo-europe.eu), which aims to develop tools and data analytics for pandemic and outbreak preparedness. VEO data analysis is applied to open data shared through our platform and complements analysis presented upon other data sharing platforms. The platform and analysis tools are in development and are presented in periodic reports.

Update since last report

As was noted in the Report #11, the unprecedented scale of genomic sequencing has challenged the existing infrastructures in terms of storage and compute capacity needed for analysis of the growing datasets. We are exploring how to scale up the analysis functions to allow real-time analysis without the need for downsizing.

For complete archiving and to enable advanced analysis, we keep all of the raw files (FASTQ) and various steps of the processed data (BAM) beyond the final product (VCF) that is rendered in the CoVEO app. As some samples have very deep sequencing depth, the volume of the latest data report became extremely large (185 TB). Due to this volume of data, a new challenge has been created for the data transfer to the CoVEO database.

In order to address this scaling issue, for the next version of the report, the large raw and intermediate files will be packaged separately (while all files remain accessible in the EU open data portal). Also, we have restructured the CoVEO database to take up a smaller volume per sample. In addition, several new constraints were applied to database tables to avoid record repetitions or corrupted fields that occurred from time to time in the previous reports.

EMBL-EBI has also scaled up the data processing workflow and uses Google Cloud to be able to maintain the data ingest rate and gradually work off the backlog of datasets to be analyzed. As a result, the latest analysis batch contains information about as many samples as all of the previous batches added together.

This work is supported by funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 874735 (VEO).

The CoVEO web application (https://coveo.vo.elte.hu/report/report-1/) presents a range of plots, including:

- Maps on the number of raw SARS-CoV-2 sequences submitted and processed from countries across the world or within the EU.
- Graphs on the percentage of sequenced new cases per week, combining data from EMBL-EBI and the European Centre for Disease Prevention and Control (ECDC). Secondly the number of samples submitted within the EU.
- Distribution of variants of concern (VOC) or variants under investigation (VUI) identified in samples from a given country across time.
- Overall number of samples across various countries presenting with various VOCs and VUIs.
- Percentage of samples derived from a given variant within specific countries across the world

Section I: Data mobilisation

The number of read datasets released into the COVID-19 Data Portal up to the current data freeze (10 Jul 2022) is shown in Table I. Please note that the sequence data set is dynamic with options for data owners to update metadata records (such as corrections of geographical annotation and, rarely, suppression); the numbers provided here therefore reflect the currently available data set for the given time windows and thus may differ slightly from those previously reported (https://www.covid19dataportal.org).

Table I: Update of number of submissions of raw read datasets to the ENA.

Date		**21 Sept 2021**	**19 Oct 2021**	**07 Dec 2021**	**05 Apr 2022**	**22 May 2022**	**10 July 2022**
Raw read datasets	Total	1,549,740	1,876,126	2,672,038	4,139,890	4,510,859	4,844,657
	Illumina	1,239,284	1,502,424	2,217,465	3,551,782	3,893,702	4,145,214
	Oxford Nanopore	151,031	172,654	213,259	341,021	369,599	396,772
	Other	159,425	201,048	241,314	247,087	247,558	302,671
Source countries for raw read datasets		75	80	85	92	92	94

This work is supported by funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 874735 (VEO).

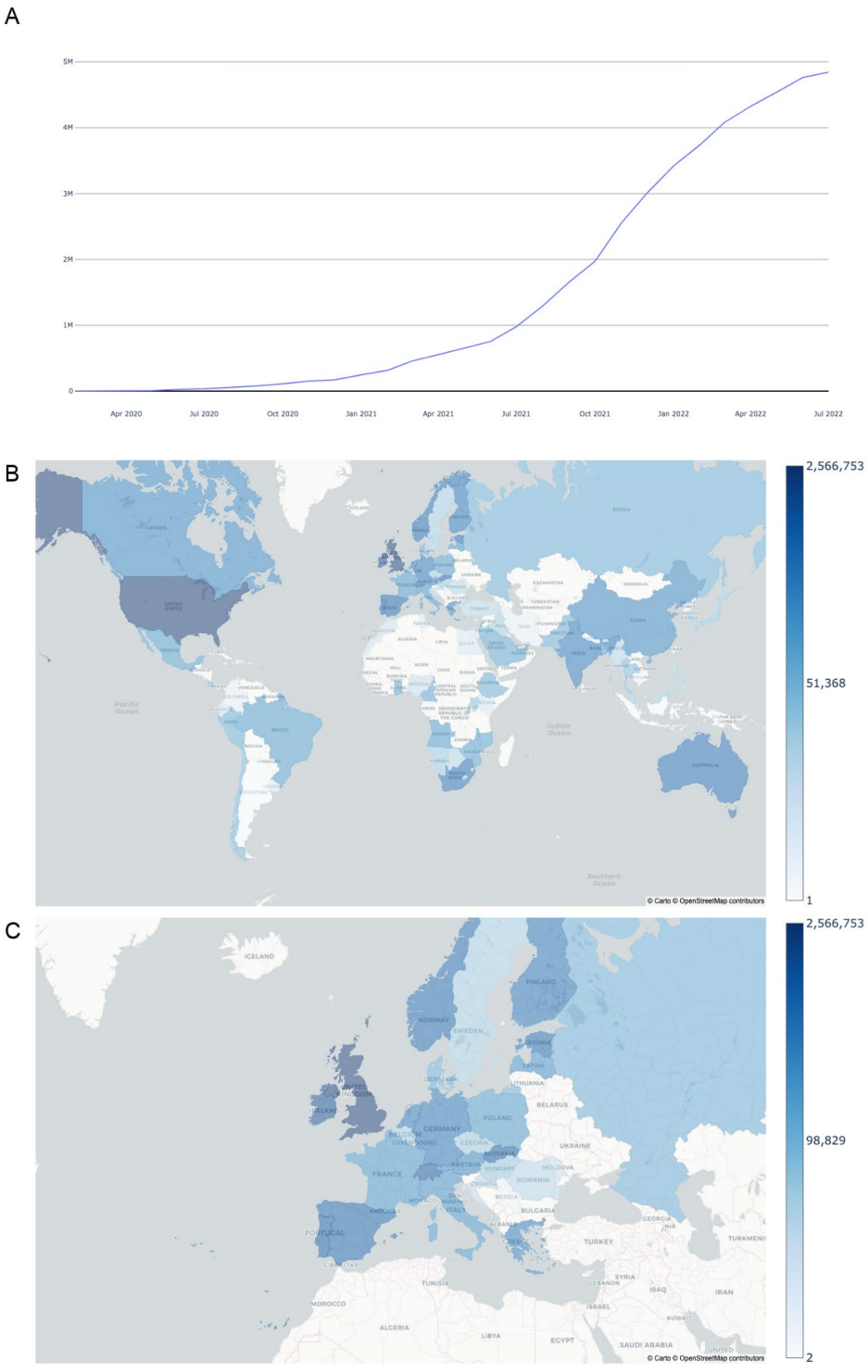

Figure I: ***Globally available total number of raw SARS-CoV-2 data*** *and distribution of sources, showing (A) sustained growth in raw data since launch of the mobilisation campaign by cumulative number of data sets, (B) and (C) geographical sources of global and European raw data, respectively, for which 59.3% of global data have been routed through the SARS-CoV-2 Data Hubs, with the remaining 40.7% arriving into the platform from collaborators in the US and Asia. Note that the color scales are logarithmic best to show the broad range across countries.*

This work is supported by funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 874735 (VEO).

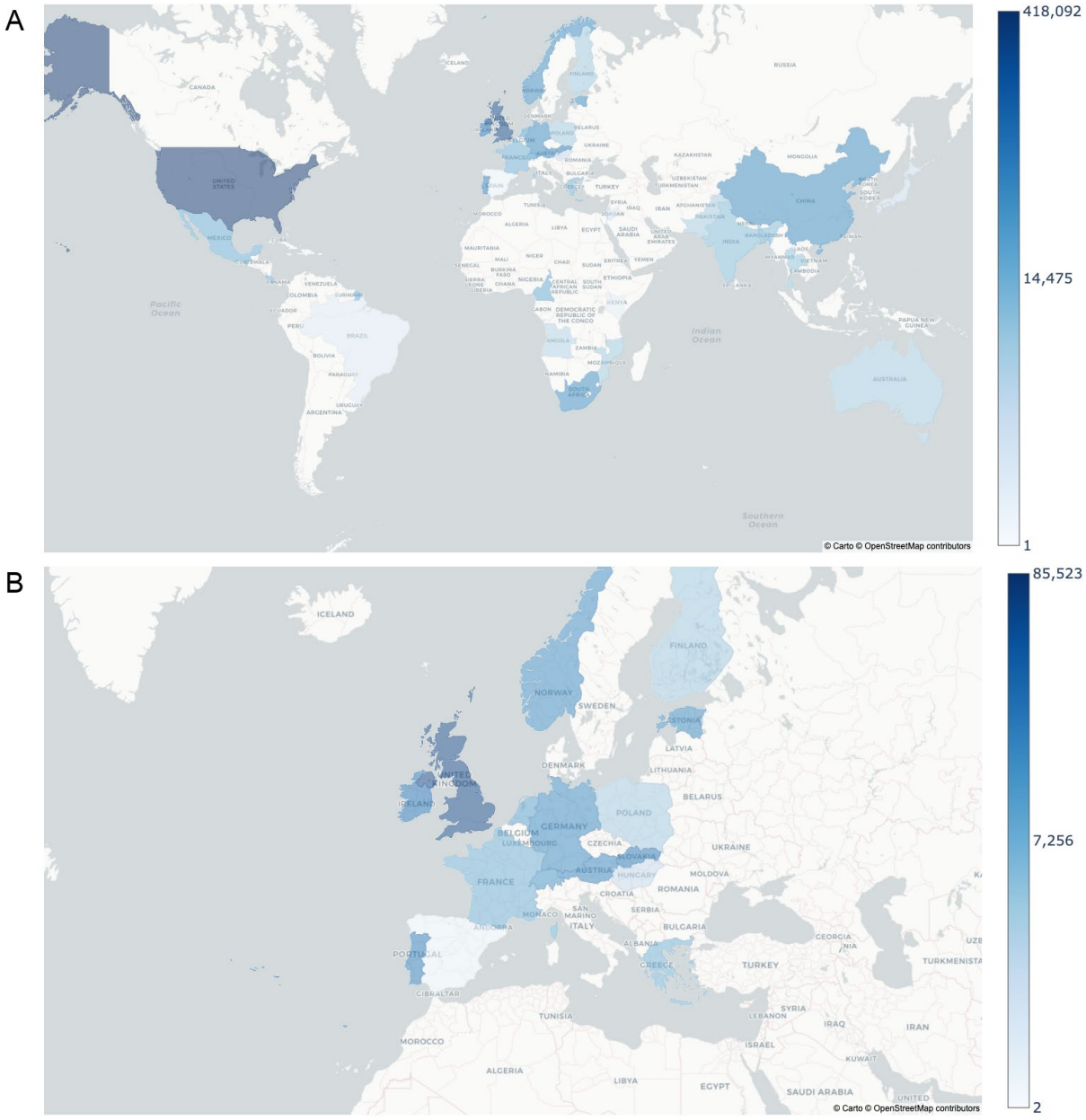

Figure II: ***New raw SARS-CoV-2 data*** *and distribution of sources at global (A) and European (B) levels* ***mobilized since 12 April 2022****. Note that the color scales are logarithmic best to show the broad range across countries.*

This work is supported by funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 874735 (VEO).

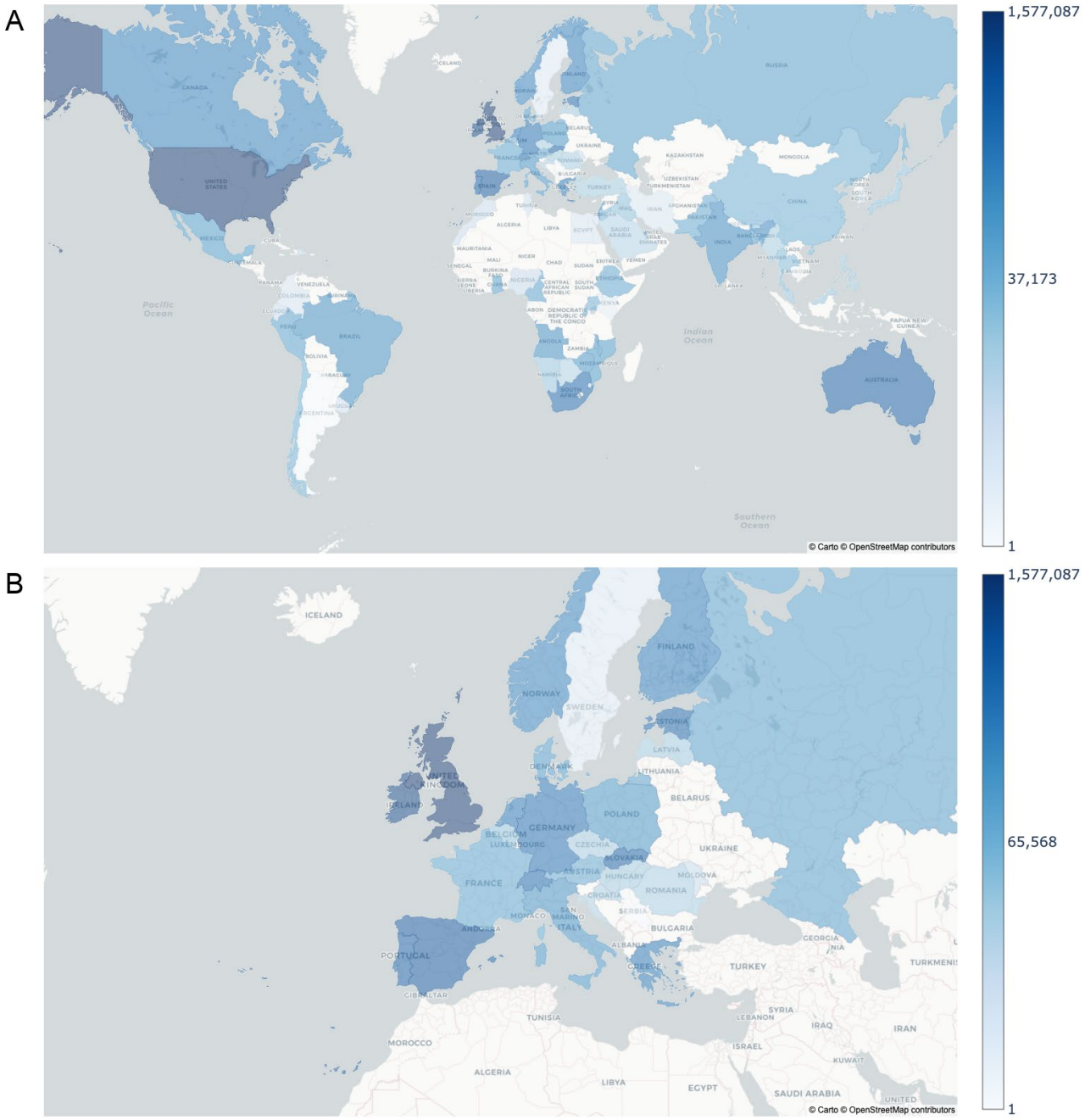

*Figure III: Geographical sources of **analysed raw data** comprising 3,284,367 data sets spanning the period of data first published from 2020-02-05 to 2022-06-29 globally (A) and within Europe (B). Note that the color scales are logarithmic best to show the broad range across countries.*

This work is supported by funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 874735 (VEO).

Recommendations and next steps:

The incorporation of the CoVEO web-app into the EU COVID-19 Data Portal allows rapid scrolling through all publicly submitted data that have been processed through the same workflow to increase comparability of data. With the next update, we expect that the delays in data processing that currently are increasing due to the growing size in data will be reduced with the revisions implemented. However, the advantage of reduced delays in data visualization needs to be met with reduction of delays in public sharing of data. Public health and research centers should be encouraged to share the raw sequencing data as soon as possible after they are generated.

Distribution of the Report

To be added to the distribution list of this report, please send an email to veo.europe@erasmusmc.nl with 'VEO COVID-19 Report' in the subject line. These reports are posted on the www.veo-europe.eu website as well as the www.covid19dataportal.org website.

Contributing to this report from the VEO Consortium:

Erasmus Medical Center

Eötvös Loránd University

EMBL European Bioinformatics Institute

Technical University of Denmark

This work is supported by funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 874735 (VEO).