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

Report No. 14 – 18 November 2022

Summary

- Update on mobilisation of raw reads, now totaling sequencing datasets from 5,784,221 viral raw read sets from 96 countries, a 5.1% increase since the previous report.
- The variant calling workflow for the Illumina data has been deployed on the Google Cloud Platform. In total, 274,693 Illumina and 22,699 Oxford Nanopore samples have been processed in the period since the last report
- In line with observations elsewhere, the number of new datasets submitted appears to be slowing down as expected given the phase of the pandemic, where the amount of testing and the sequencing effort have been reduced in many countries
- In line with the phase of the pandemic, this report will move to quarterly updates.

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.

Why do we want to encourage and support the analysis of raw sequence data?

The current default in genomic surveillance is the release of assembled full genomes through semi-open (GISAID) or open (ENA/NCBI/INSDC) access. These genomes are generated through locally developed bioinformatic workflows, which are not standardized. This diversity in workflows is not likely to lead to variation in the global strain assignment, but when looking at individual mutations, the choice of workflows may affect outcome of analyses. Therefore, access to the underlying locally produced data (raw reads) is considered to be important for the EU COVID-19 Data Portal.

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 app interprets and summarizes the variation data produced by these pipelines. Here, users can explore the emergence, spread and incidence of SARS-CoV-2 variants across the globe to give a view of the status of the pandemic. This app can be accessed by clicking the 'Variant Browser' links throughout the COVID-19 Data Portal, or by visiting: https://covid19dataportal.org/coveo

Section I: Data mobilisation update

The number of read datasets released into the COVID-19 Data Portal up to the current data freeze (20 October 2022) is shown in Table I. Please note that the sequence dataset 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 dataset 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		**07 Dec 2021**	**05 Apr 2022**	**22 May 2022**	**10 July 2022**	**14 Sept 2022**	**20 Oct 2022**
Raw read datasets	Total	2,672,038	4,139,890	4,510,859	4,844,657	5,684,416	5,872,867
	Illumina	2,217,465	3,551,782	3,893,702	4,145,214	4,676,550	4,817,454
	Oxford Nanopore	213,259	341,021	369,599	396,772	426,658	445,891
	Other	241,314	247,087	247,558	302,671	581,208	609,522
Source countries for raw read datasets		85	92	92	94	96	96

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

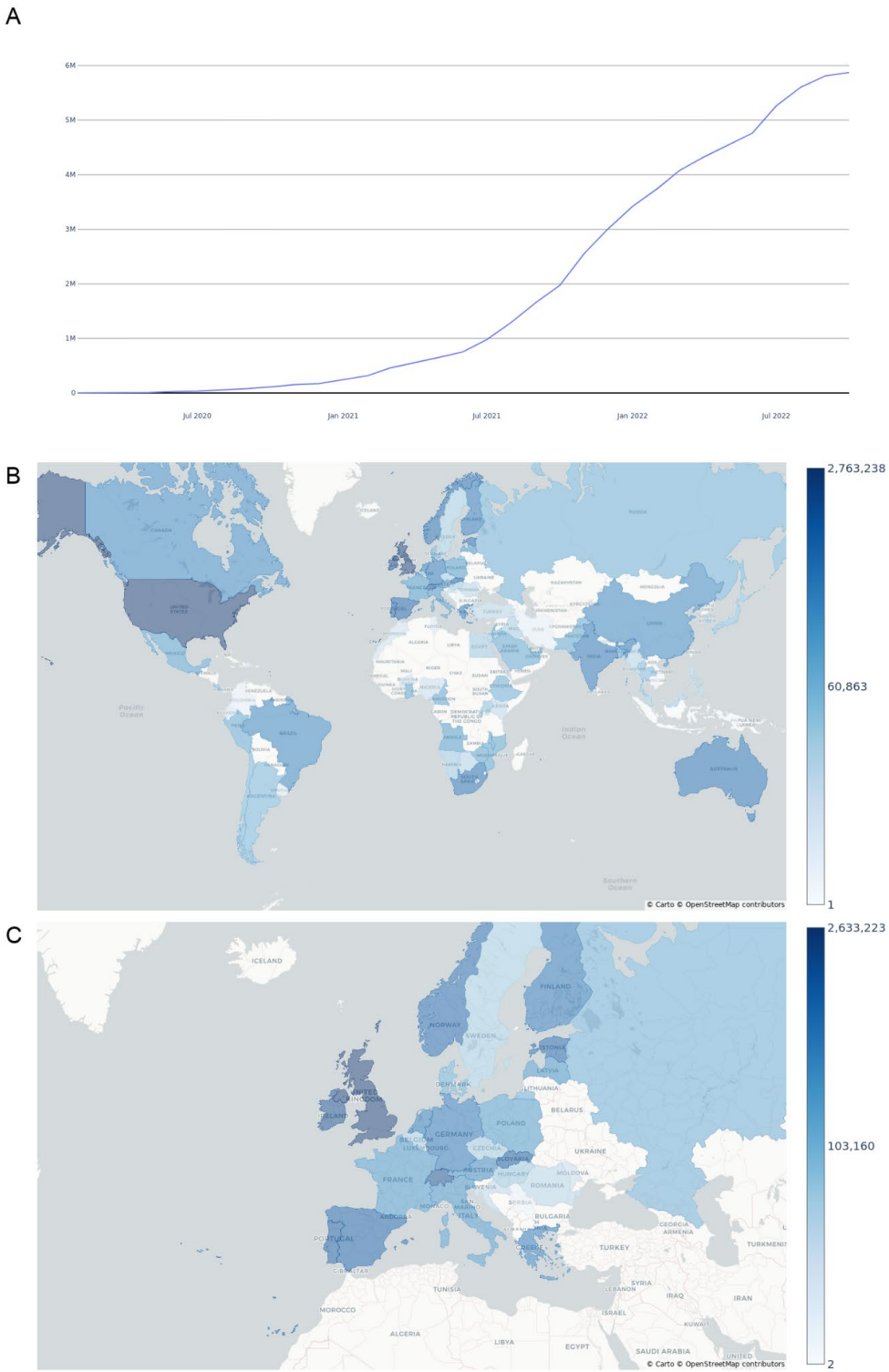

Figure I: ***Globally available total number of raw SARS-CoV-2 data*** *and distribution of sources, showing (A) sustained growth in raw data since launch of the mobilization campaign by cumulative number of datasets, (B) and (C) geographical sources of global and European raw data, respectively, for which 51.2% of global data have been routed through the SARS-CoV-2 Data Hubs, with the remaining 48.8% arriving into the platform from collaborators in the US and Asia. Note that the colour 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).

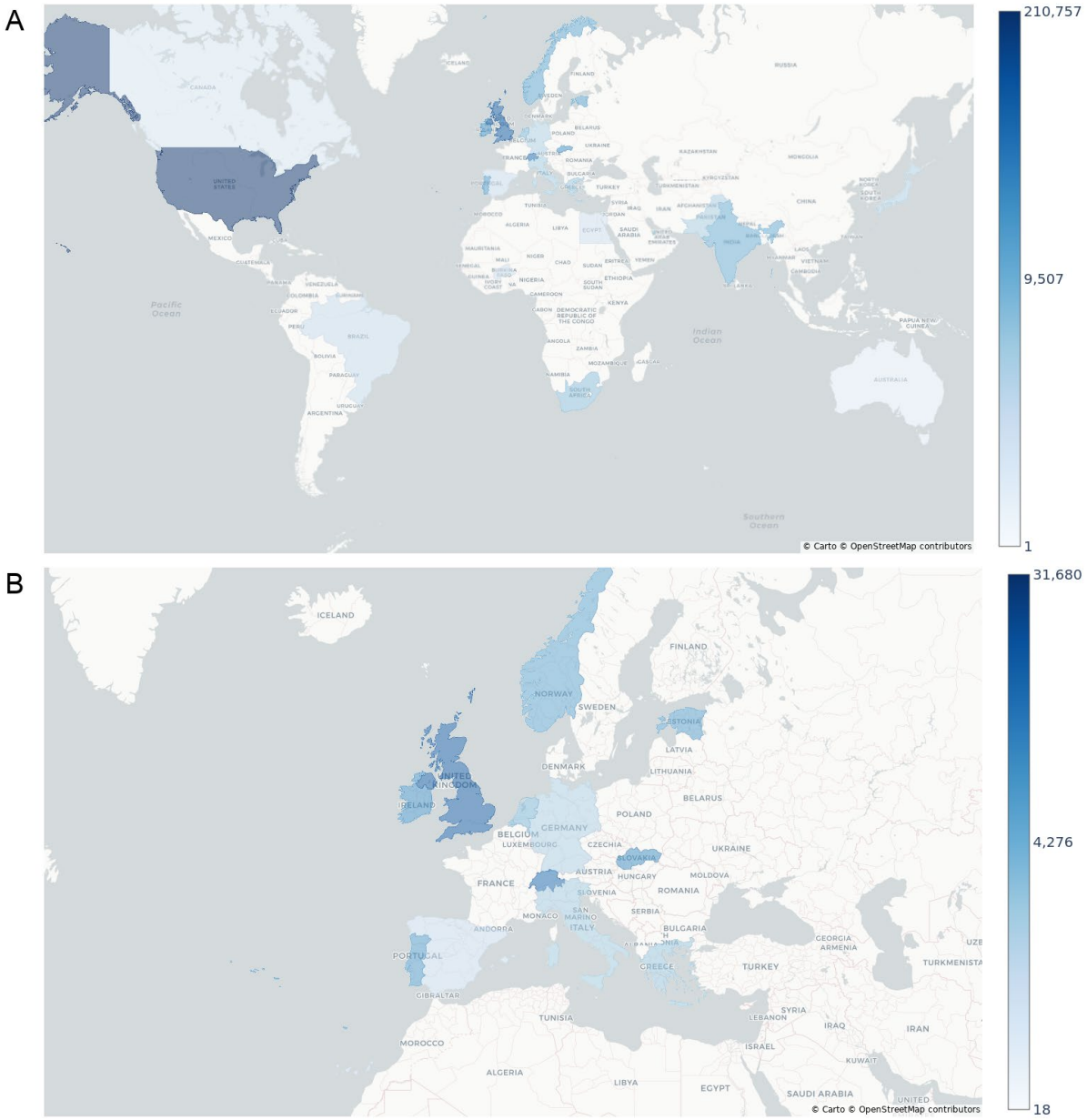

Figure II: ***New raw SARS-CoV-2 data*** *and distribution of sources at global (A) and European (B) levels* ***mobilized since 22 August 2022****. Note that the color scales are logarithmic best to show the broad range across countries.*

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

Section II: Analysis

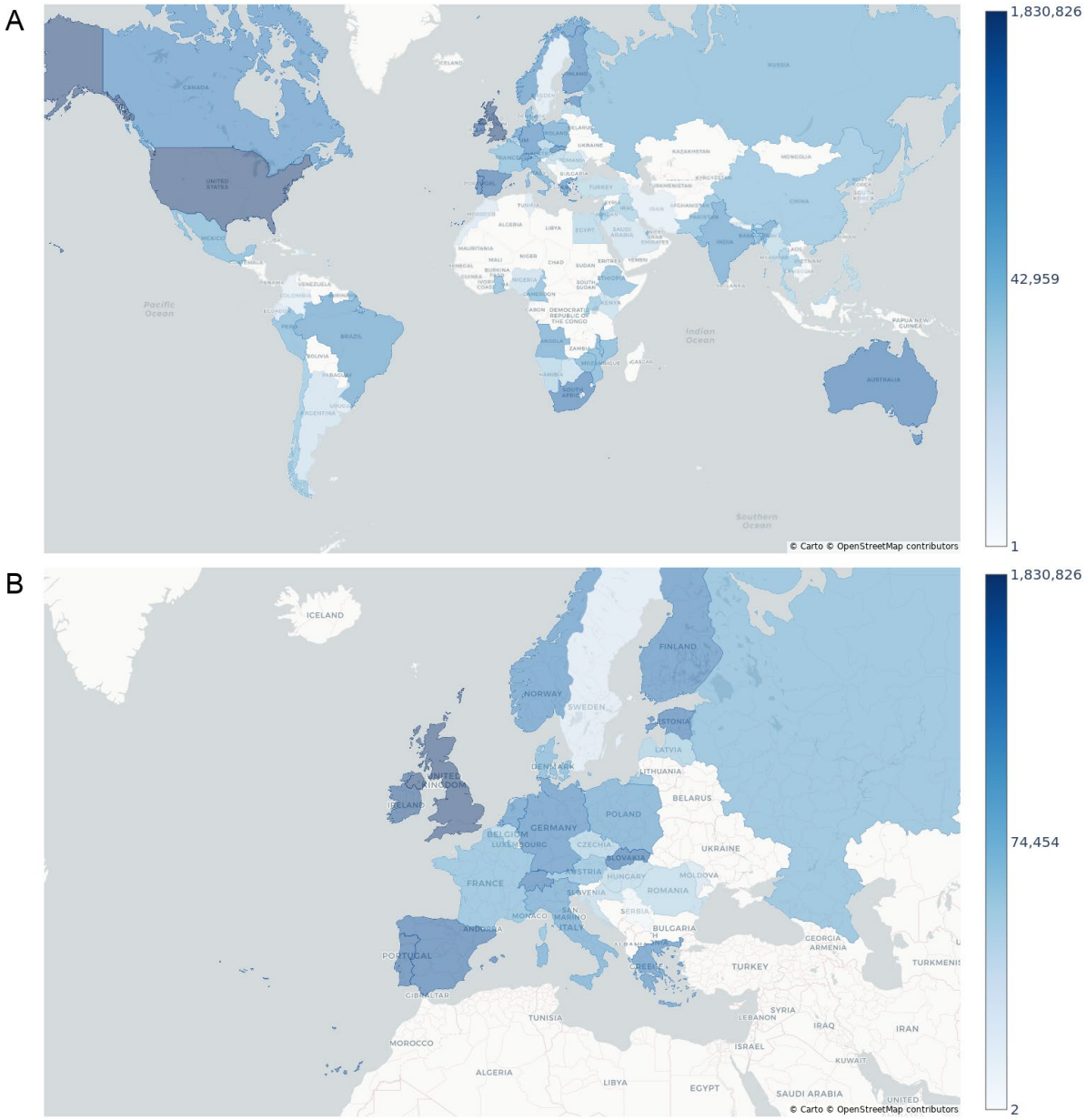

*Figure III: Geographical sources of **raw data processed through the workflow for variant calling**, comprising 3,924,211 datasets spanning the period of data first published from 05 Feb 2020 to 21 Sep 2022 globally (A) and within Europe (B). Note that the color scales are logarithmic best to show the 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).

Results of variant calling

A workflow to analyze the submitted data has been established, and at this stage, full processing of the backlog of data from the start of the pandemic is ongoing. At the moment, 2,019,234 of the 3,924,211 processed datasets have been made available for variant searching. However, because of the enormous volume of (minor) variants that have been processed, interactive searching of constellations of variants within the entire database has become infeasible.

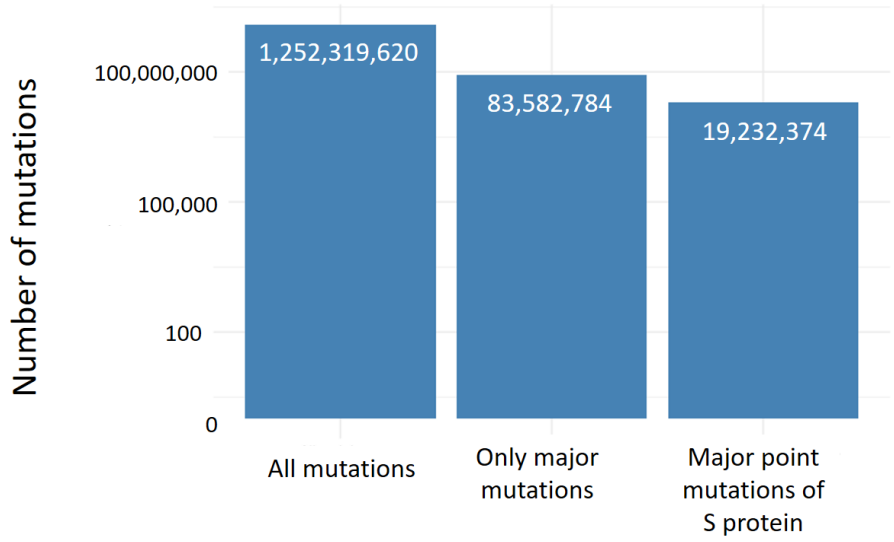

Figure IV. Graph shows how many mutations are present in the current version of the database. The majority of the mutations are minor mutations. The "Custom variant browser", part of the CoVEO app, searches now among only the major point mutations of the S protein and usually gives results in less than a minute.

Therefore, there are, on the one hand, developments ongoing to technically increase the speed of the search process of the entire variant database for in-depth, but time-consuming queries. On the other hand, there are developments ongoing to create a database version with reduced detail and reduced scope, to enable responsive querying of the most recent data (Figure IV).

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

Mutations and variants

Update as of 16 November 2022

Currently, several Omicron lineages and sub-variants are circulating around the globe. These are descendants of the BA.2 and the BA.5 lineages or recombinant viruses. Of these lineages, several different sub-variants are circulating that are all characterized by specific amino acid mutations in the spike protein. Remarkably, these variants acquired similar mutations in a different backbone. The most common mutations found are the R346T and the K444* mutation which can be found independently in BA.5 and BA.2.75 sub-lineages. Especially the R346T mutation is of concern since it has been shown that this mutation is responsible for complete escape from Evusheld, which is one of the main antiviral treatments used at the moment.

At the moment, the variants that are increasingly being detected are the XBB* and the BQ.1* variants. Also several BA.2.75 derived viruses are still being detected, for instance, the BN.1 and CA.2 variant. In the current "zoo" of variants that are circulating around the globe, it is not clear which variant will be dominant in the future. In order to determine the phenotypic effects of the different variants, we have cultured the two main variants (XBB.1 and BQ.1), which we will include in the comparative neutralization data and antigenic maps.

CoVEO web app development

CoVEO, a web-based application (https://www.covid19dataportal.org/coveo), was created to communicate with the database and visualize the content. Earlier, the app showed information only about the official WHO VOC/VUI, but now with the new feature ("Custom variant browser") any user can search for a given virus variant by providing an 'including and excluding' mutation list of the S protein (Figure V). The app will show how many samples contain the custom-selected mutations in each country, and after the selection of a given country in the table, a graph is generated presenting the appearance of these samples in time based on the sampling date. In the future, the selectable mutation will be extended to other regions of the viral genome.

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

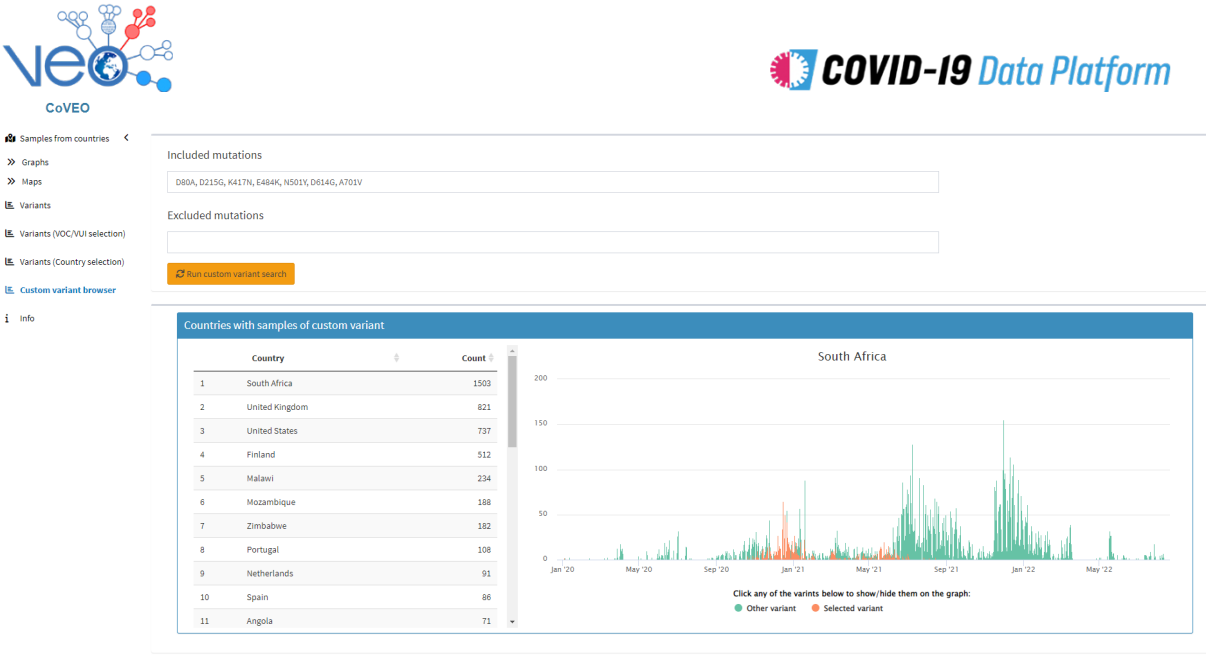

Figure V: Custom variant browser, inclusion and exclusion SARS-CoV-2 spike amino acid mutation spectra can be filled in and searched for in the CoVEO database and are visualized by sampling date per country of origin.

The database and associated web-based application have been integrated at EMBL-EBI, and available from the COVID-19 Data Portal (https://covid19dataportal.org/coveo).

As an example, Figure VI shows the BA 2.75.2 variant on the “Custom variant browser” tab of the CoVEO app. This is an interesting variant due to its immune escape potential.

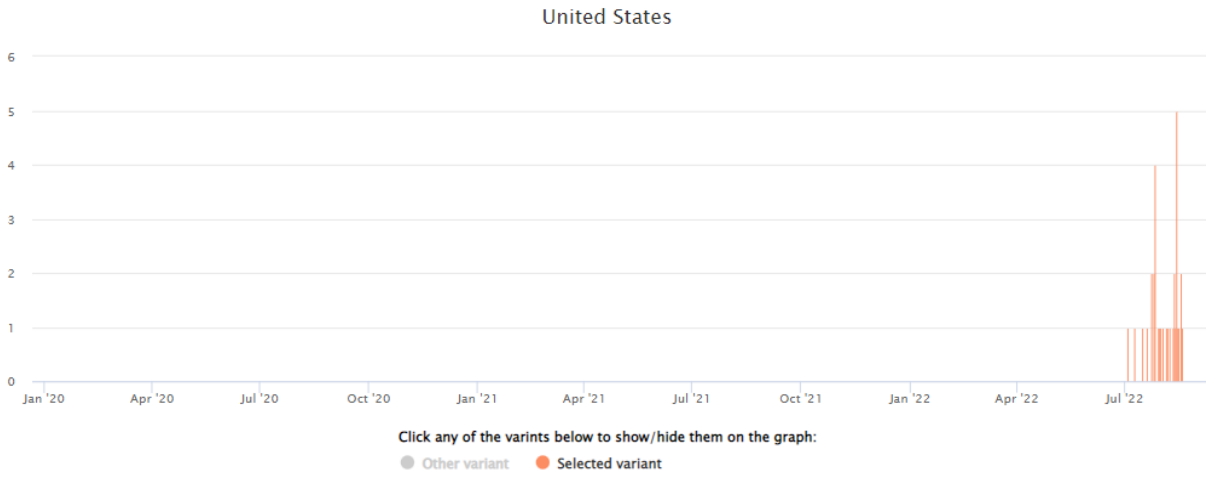

Figure VI. The number of cases with BA 2.75.2 variant characterized by the presence of K147E, W152R, F157L, I210V, G257S, G446S and N460K point mutations of spike protein. The “Other variants” were hidden in the legend below the graph so the presence of this variant is more visible.

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 above report shows the results of the automated mutation analysis on raw read datasets submitted to ENA, as well as visualizations of the data. The number of raw reads continues to increase. We continue to work with potential users to discuss ease of upload to reduce a barrier to sharing of raw reads. Public health and research centers should be encouraged to share the raw sequencing data as soon as possible after they are generated.

The EU member states could consider whether coupling funding to sharing of data should be considered, as has been done in some countries.

The newer Omicron variants are very hard to distinguish without knowing for sure that the mutation is not absent based on too low coverage, so adding that feature to the variant browser will be important going forward. Our database already contains information about the coverage that is reachable programmatically from the database.

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).