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

Report No. 15 – 16 March 2023

Summary

- For SARS-CoV-2, the number of new entries per month is slowing down, in line with reduced testing and sequencing noted across the globe. There now are 6,214,091 raw read sets from 98 countries, a 5.8% increase since the previous report.
- The automated analysis workflow and a drag and drop submission tool were re-used when the monkeypox outbreak started and is now available in the Pathogens Portal https://www.ebi.ac.uk/ena/pathogens/v2/monkeypox
- With the real-time pressures for SARS-CoV-2 analysis decreasing, we are preparing to evaluate lessons learned from massively scaling up genomic surveillance using the currently developed public infrastructure. The primary focus will be on re-use of data for in depth research. The available genomes have cumulative information on more than 2 billion mutations/minor variants, which is by far the largest pathogen genome dataset ever. Therefore, the next step is to add options for downselecting to facilitate research access.

Background

This report summarises 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 standardised. 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 summarises 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

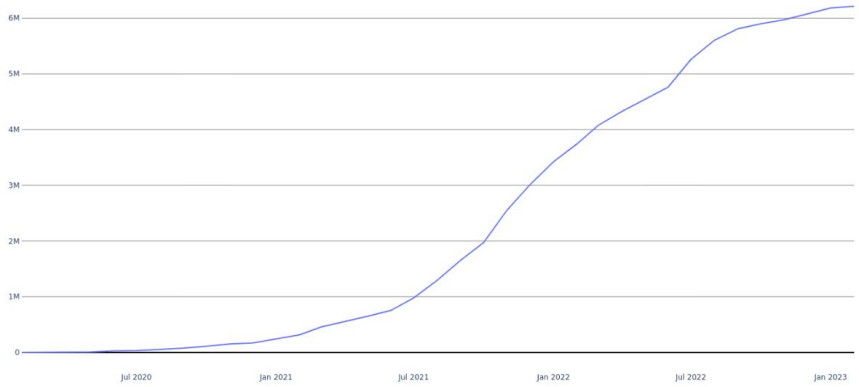
The number of read datasets released into the COVID-19 Data Portal up to the current data freeze (13 February 2023) is shown in Table I. Please note that the sequence dataset is dynamic with options for data owners to update metadata records (such as corrections of geographical annotation and, rarely, 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.

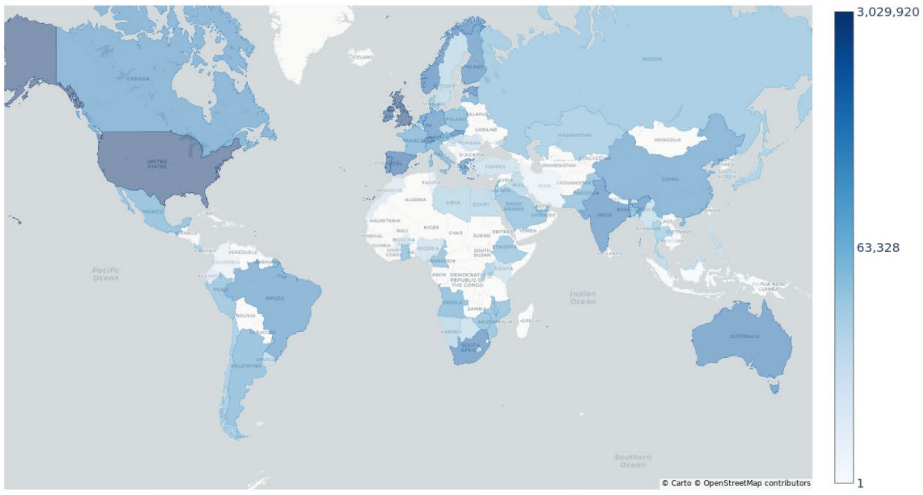
Date		**05 Apr 2022**	**22 May 2022**	**10 July 2022**	**14 Sept 2022**	**20 Oct 2022**	**13 Feb 2023**
Raw read datasets	Total	4,139,890	4,510,859	4,844,657	5,684,416	5,872,867	6,214,091
	Illumina	3,551,782	3,893,702	4,145,214	4,676,550	4,817,454	5,066,612
	Oxford Nanopore	341,021	369,599	396,772	426,658	445,891	494,848
	Other	247,087	247,558	302,671	581,208	609,522	652,631
Source countries for raw read datasets		92	92	94	96	96	98

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

A



B



C

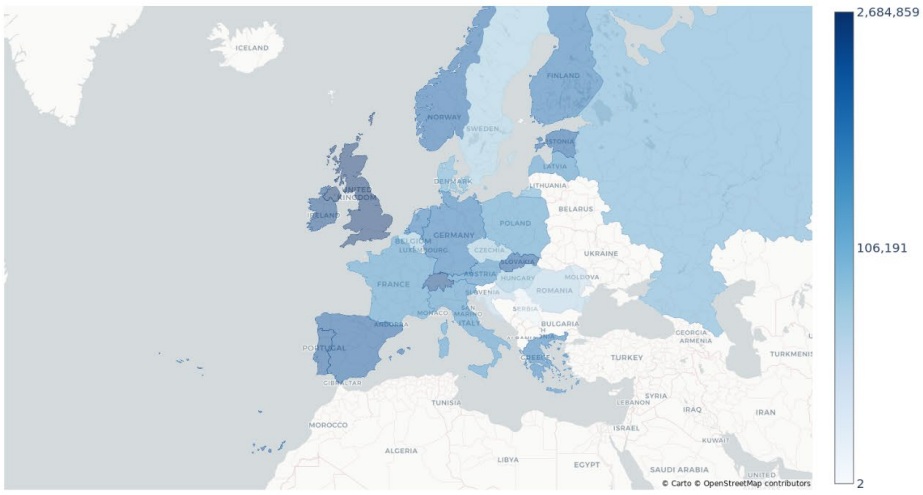


Figure I: ***Globally available total number of raw SARS-CoV-2 data*** *and distribution of sources, showing (A) sustained growth in raw data since launch of the mobilization campaign by cumulative number of datasets, (B) and (C) geographical sources of global and European raw data, respectively, for which 49.6% of global data have been routed through the SARS-CoV-2 Data Hubs, with the remaining 50.4% 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).

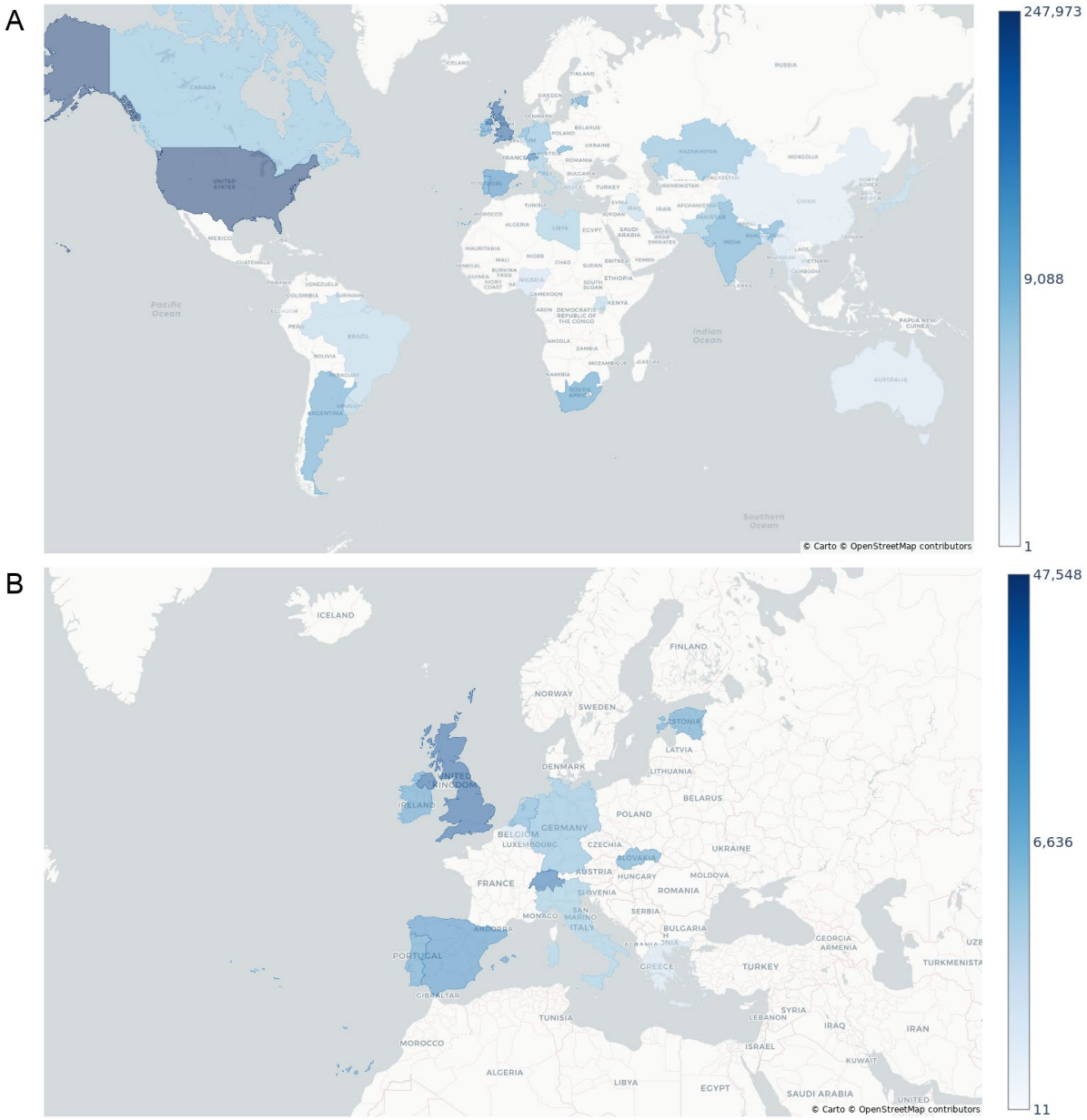


Figure II: ***New raw SARS-CoV-2 data*** *and distribution of sources at global (A) and European (B) levels* ***mobilized since 24 October 2022****. Note that the color scales are logarithmic best to show the broad range across countries.*

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

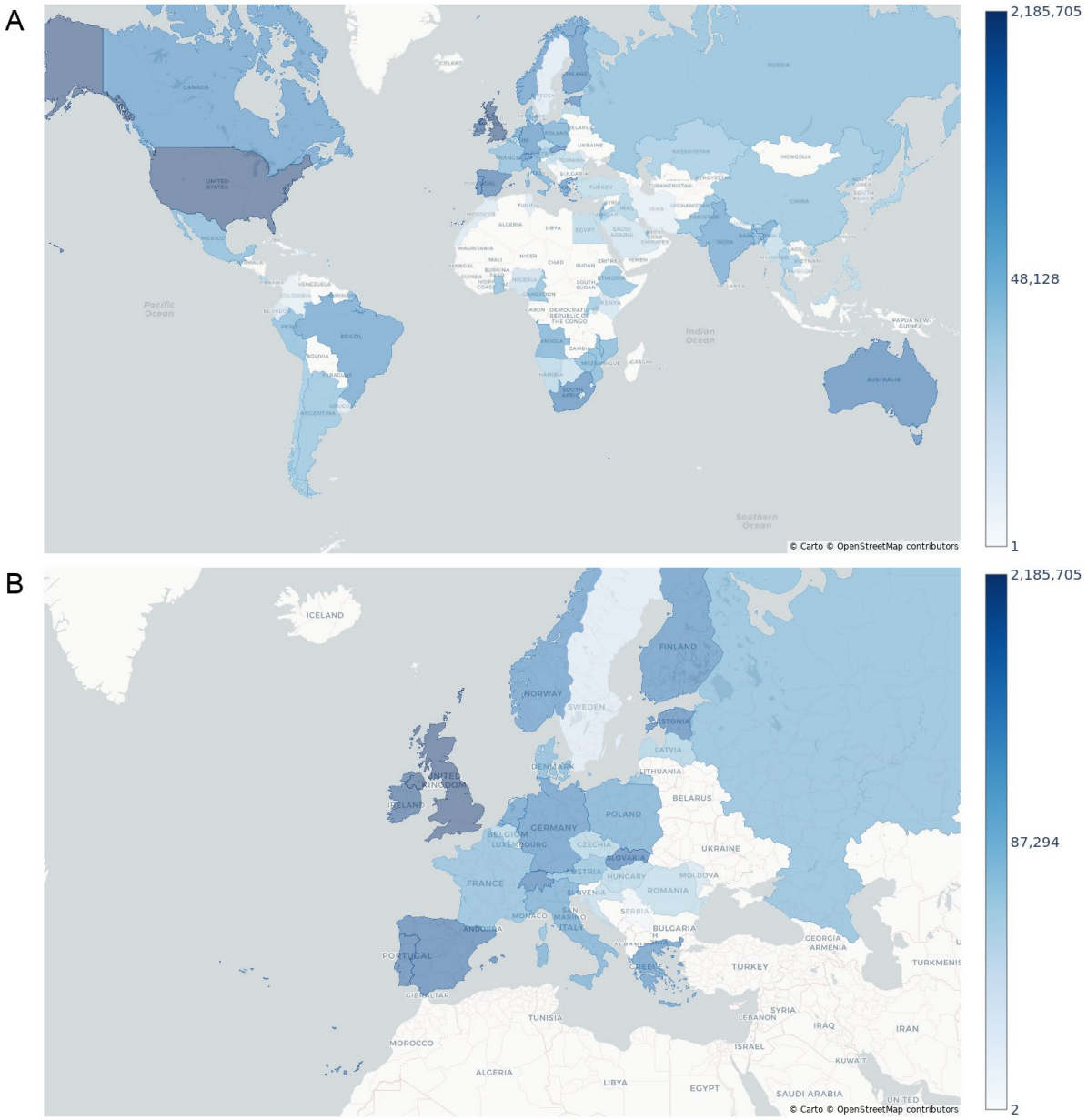


*Figure III: Geographical sources of **raw data processed through the workflow for variant calling**, comprising 4,489,589 datasets spanning the period of data first published from 05 Feb 2020 to 27 Jan 2023 globally (A) and within Europe (B). Note that the color scales are logarithmic best to show the 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, 3,005,983 of the 4,489,589 processed datasets have been made available for variant searching. The output is in files that list all mutations and that can be accessed for in depth studies. Due to the global dissemination of SARS-CoV-2 and the unprecedented sequencing effort, the number of (minor) variants is becoming too large for the interactive searching feature that was developed and has been used to generate near real-time overviews. Currently, we are working on downselecting the variant files for sub analysis.

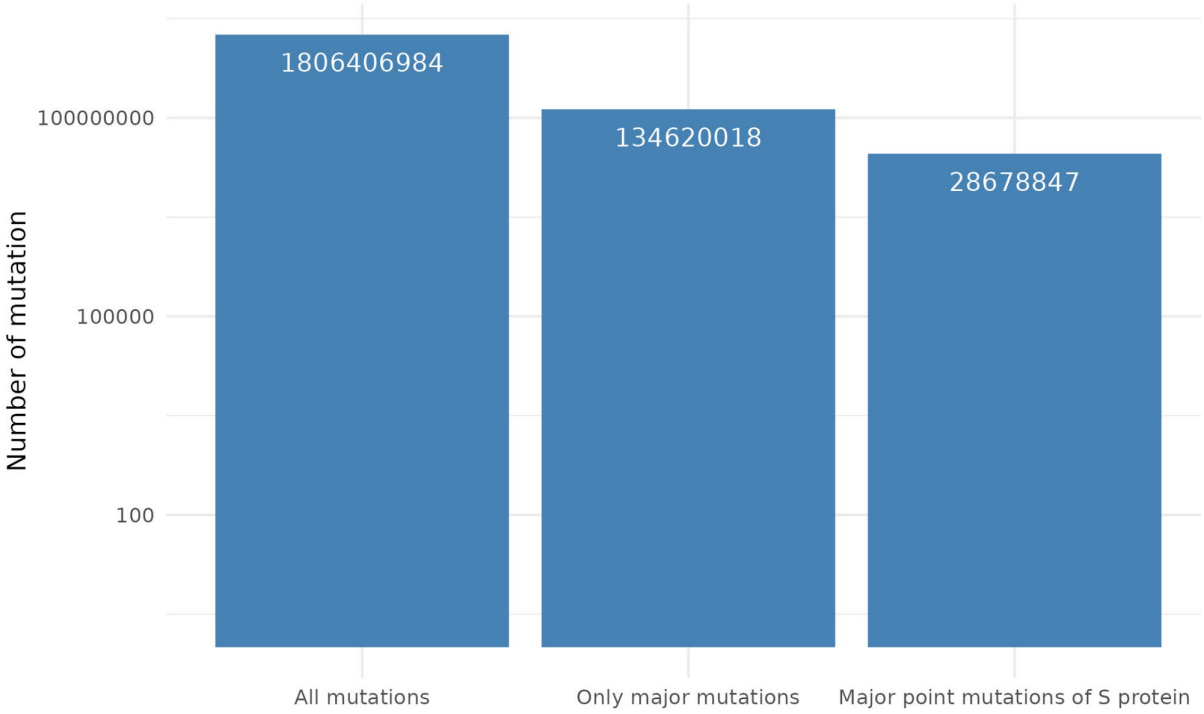


Figure IV. Graph shows how many mutations are present in the current version of the database. The majority of the mutations are minor mutations. The "Custom variant browser", part of the CoVEO app, searches now among only the major (AF>0.5) point mutations of the S protein and usually gives results in less than a minute.

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 March 2023

Since November 2022, several Omicron lineages and sub-variants have been circulating around the globe. These were 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 characterised 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. Novel SARS-CoV-2 Omicron variants, including BM.1.1.1, BQ.1.1, and XBB.1, continue to emerge at an unprecedented rate, evading pre-existing immunity from vaccination and previous infection. Also several BA.2.75 derived viruses are still being detected. The phenotypic effects of the XBB*, BQ.1 and BM1.1.1 variants have been published in a recent manuscript (https://doi.org/10.1016/S2666-5247(22)00384-6).

CoVEO web app development

CoVEO, a web-based application (https://www.covid19dataportal.org/coveo), was created to communicate with the database and visualise the content. Earlier, a “Custom variant browser” tab was added so the user could search for a given virus variant by providing an ‘including and excluding’ mutation list of the S protein. This functionality was extended by adding the option to select only those samples with sequencing depth >30 by each of the selected mutation positions. This allows exclusion of low quality reads from the analysis.

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

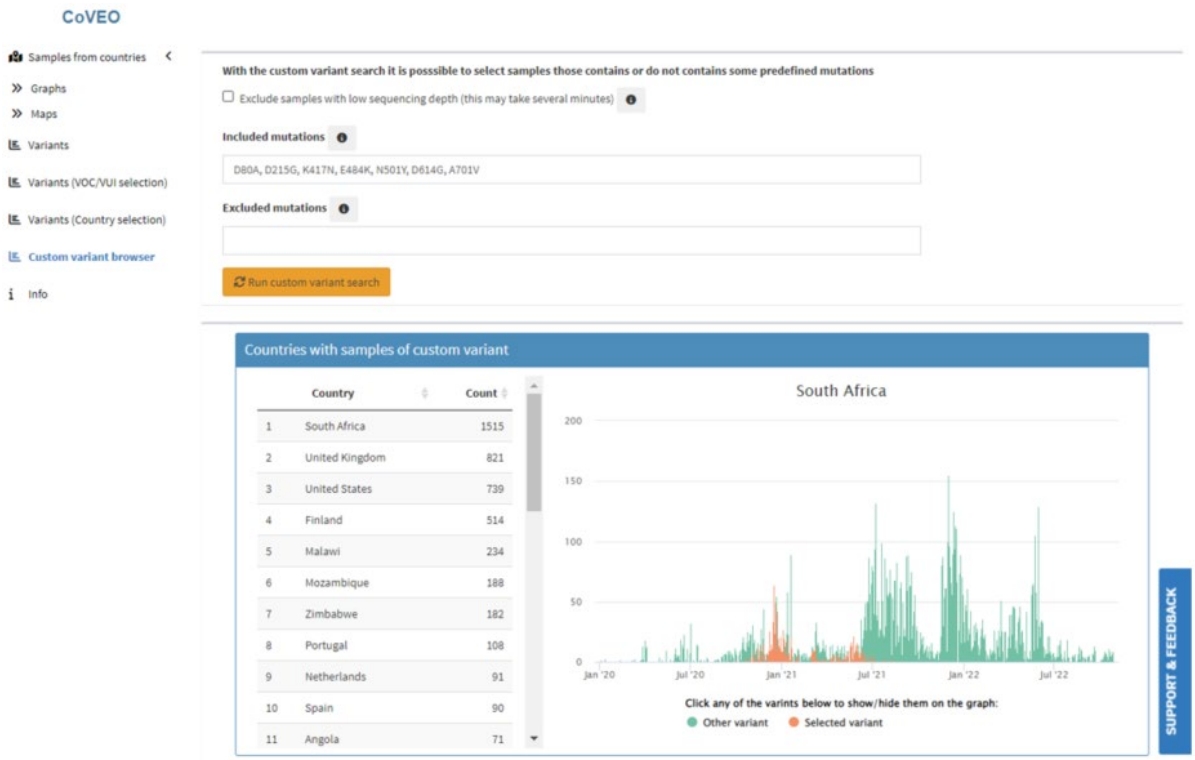


Figure V: Custom variant browser, inclusion and exclusion SARS-CoV-2 spike amino acid mutation spectra can be filled in and searched for in the CoVEO database and are visualised by sampling date per country of origin. It is possible to narrow the search for only those samples where enough information is available by the selected positions.

Repurposing of VEO Analysis Tools

In response to the Monkeypox virus (MPXV) outbreak, a dedicated view of publicly archived MPXV sequences and reads has been added to the new Pathogens Portal (https://www.ebi.ac.uk/ena/pathogens/v2/monkeypox). This allows users to quickly find relevant MPXV sequences and read data and filter through it based on key metadata fields such as “Country” and “Collection date”.

Alongside the deployment of a dedicated MPXV outbreak tab on the portal, dedicated submission guidance (https://docs.google.com/viewer?url=https://github.com/enasequence/ena-content-dataflow/raw/master/docs/Monkeypox%20virus%20ENA%20Submission%20Guidance.pdf) for users looking to submit data has been created and a drag and drop data submission tool (https://ebi-ait.github.io/monkeypox-data-upload/) has been repurposed to allow MPXV submissions, instead of just SARS-CoV-2. In parallel, a targeted outreach campaign is ongoing to identify data holders and encourage public submission of data to the ENA.

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

Since 01 January 20222, 13 distinct submitters have submitted more than 845 MPXV samples to the ENA. This is against a backdrop of 2176 samples submitted to the INSDC as a whole from 44 unique submitters. 38% of INSDC MPXV samples are flowing through ENA.

Private data sharing has been set up through the data hubs system and is currently being used by Erasmus Medical Center (EMC) and Friedrich-Loeffler-Institut (FLI) to collaborate on MPXV analysis.

Analysis workflows for the generation of variation data as well as consensus sequences, developed for SARS-CoV-2, were evaluated for their utility in MPXV analysis. Results indicated that the pipelines perform well at calling important variants known to be present in the current lineage of MPXV circulating in Europe. The next step is to add the Nextstrain (https://nextstrain.org/) visualisation of INSDC MPXV sequences.

Analysis of public MPXV using VEO analysis pipelines has been deployed, with the resultant analysis items being archived at ENA, matching the format of the VEO SARS-CoV-2 analyses, and displayed in the Pathogens Portal MPox Outbreak tab (https://www.ebi.ac.uk/ena/pathogens/v2/monkeypox?db=sra-analysis-mpox&size=15#search-content).

ELTE included the MPXV data into an SQL database and repurposed the CoVEO app that now can visualise the MPXV data. With the "Custom Variant Browser" feature of the app, the whole MPXV genome is searchable for mutation patterns. Overall, this provides a good indication as to the repurposing of infrastructure and tools developed as part of the COVID-19 Data Platform, for other infectious disease outbreaks.

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 visualisations 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.

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

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).