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

**Report No. 13 – 28 September 2022**

## Summary

- This update describes a major step forward in the development of the COVID-19 Data Portal: the portal now provides SARS-CoV-2 sequence data and analysis tools via a new feature, the CoVEO app (https://www.covid19dataportal.org/coveo). The CoVEO app was developed to allow analysis of the incoming raw sequence data for specific mutations or combinations of mutations.
- Update on mobilisation of raw reads, now totaling sequencing datasets from 5,684,416 viral raw read sets from 96 countries, a 21% increase since the previous report.
- The variant nomenclature has been updated, and tables on countries depositing data on VOC and VOI have been included. Information on Omicron is included.
- The variant calling workflow for the Illumina data has been deployed on the Google Cloud Platform, allowing us to process some of the backlog of data. 500,165 Illumina and 12,560 Oxford Nanopore samples have been processed in the period since the last report.

## 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.

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 the 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).

Update on further development of the COVID-19 Data Portal

- Over the past year, EBI has worked hard to encourage such sharing of data, as was reported in previous updates of this variant report. The call for raw read sharing was a success, with the current statistics showing 5,684,416 viral raw read sets from 96 countries.

- As proof of concept for the use of open data through the EU COVID-19 Data Portal, previous versions of this report focused on analyzing and visualizing the variants of interest and variants of concern as they had been annotated through the technical advisory group on virus evolution for the WHO. These analyses were done in three steps: 1) processing of raw read sets submitted to ENA with an automated workflow; ii) generation of a mutation profile in a separate database to be transferred to VEO partners; and iii) analyzing and visualizing the data with specific queries through a collaborative effort of three institutes in VEO (DTU, ELTE, ErasmusMC). This entire process including the variant calling (filtered and unfiltered) and assembly workflows that were developed and fine-tuned by VEO partners, now run on EMBL-EBI's high performance compute infrastructure. The results of these workflows are archived, indexed and available through the COVID-19 Data Portal for browsing and download: https://www.covid19dataportal.org/search/sequences?crossReferencesOption=all&overrideDefaultDomain=true&db=sra-analysis-covid19&size=15

- 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

What does this development mean for the report?

- With the automation of the CoVEO app in the COVID-19 Data Portal, anyone interested can review the status of submissions by country and by variant at their own request. Therefore, variant updates in this report will be referred to the online system.
- In addition to this, we will provide succinct variant updates summarizing key developments from now on, as well as updates on the progress of further development, focusing on the following challenges:
    - Data mobilization. With the increasing COVID fatigue and decreasing testing, the amount of sequencing is going down, and therefore the representativeness of genomic diversity. Continued advocacy for data mobilization will be needed.

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

- Going forward, the packaging of snapshot data for CoVEO (downstream) to use has been altered to support a higher volume of data. This includes splitting out large archives (used by CoVEO) that are stored per analyzed dataset into their separate files. For context, the snapshot being reported on here consists of >500,000 analyzed raw datasets. This new method supports both CoVEO, but also downstream users of the system in downloading files of specific interest, as individual files per analyzed dataset would be available for download. This new method is planned to be demonstrated in the next report.

- Speed of analysis. The unprecedented scale of data generated as part of the pandemic response has highlighted specific challenges in the ability to scale up the data infrastructures in Europe (and globally). Including the global dataset in queries is currently increasing the time to response, thereby decreasing the utility of the query tools for persons interested in fast data querying. We will work on solutions for this problem.

- Improving the current CoVEO app visualizations.

## Section I: Data mobilisation update

The number of read datasets released into the COVID-19 Data Portal up to the current data freeze (14 September 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 | | 19 Oct 2021 | 07 Dec 2021 | 05 Apr 2022 | 22 May 2022 | 10 July 2022 | 14 Sept 2022 |
|---|---|---|---|---|---|---|---|
| **Raw read datasets** | Total | 1,876,126 | 2,672,038 | 4,139,890 | 4,510,859 | 4,844,657 | 5,684,416 |
| | Illumina | 1,502,424 | 2,217,465 | 3,551,782 | 3,893,702 | 4,145,214 | 4,676,550 |
| | Oxford Nanopore | 172,654 | 213,259 | 341,021 | 369,599 | 396,772 | 426,658 |
| | Other | 201,048 | 241,314 | 247,087 | 247,558 | 302,671 | 581,208 |
| **Source countries for raw read datasets** | | 80 | 85 | 92 | 92 | 94 | 96 |

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

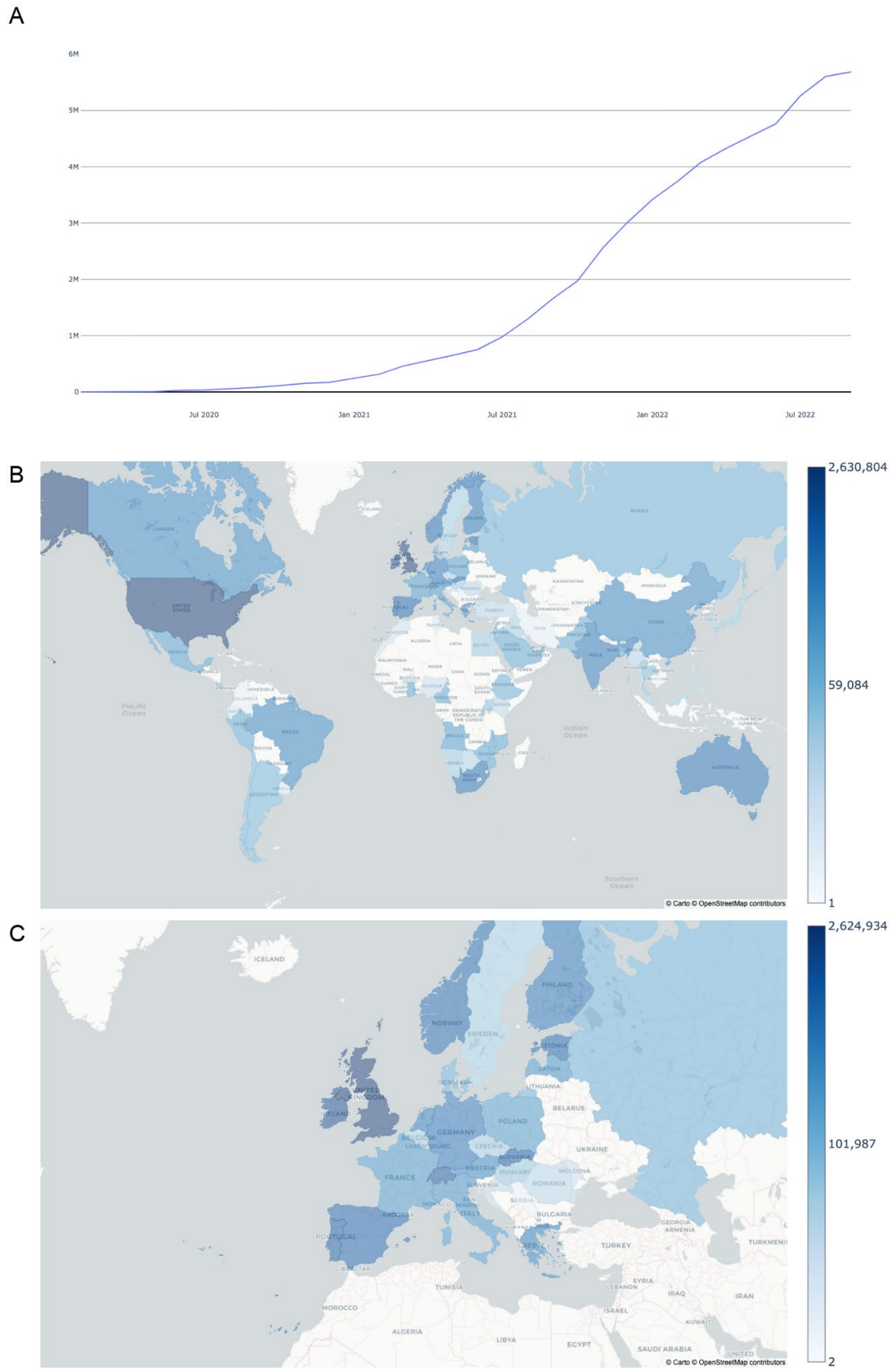


*Figure I:* ***Globally available total number of raw SARS-CoV-2 data*** *and distribution of sources, showing (A) sustained growth in raw data since launch of the mobilization campaign by cumulative number of datasets, (B) and (C) geographical sources of global and European raw data, respectively, for which 52% of global data have been routed through the SARS-CoV-2 Data Hubs, with the remaining 48% 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).

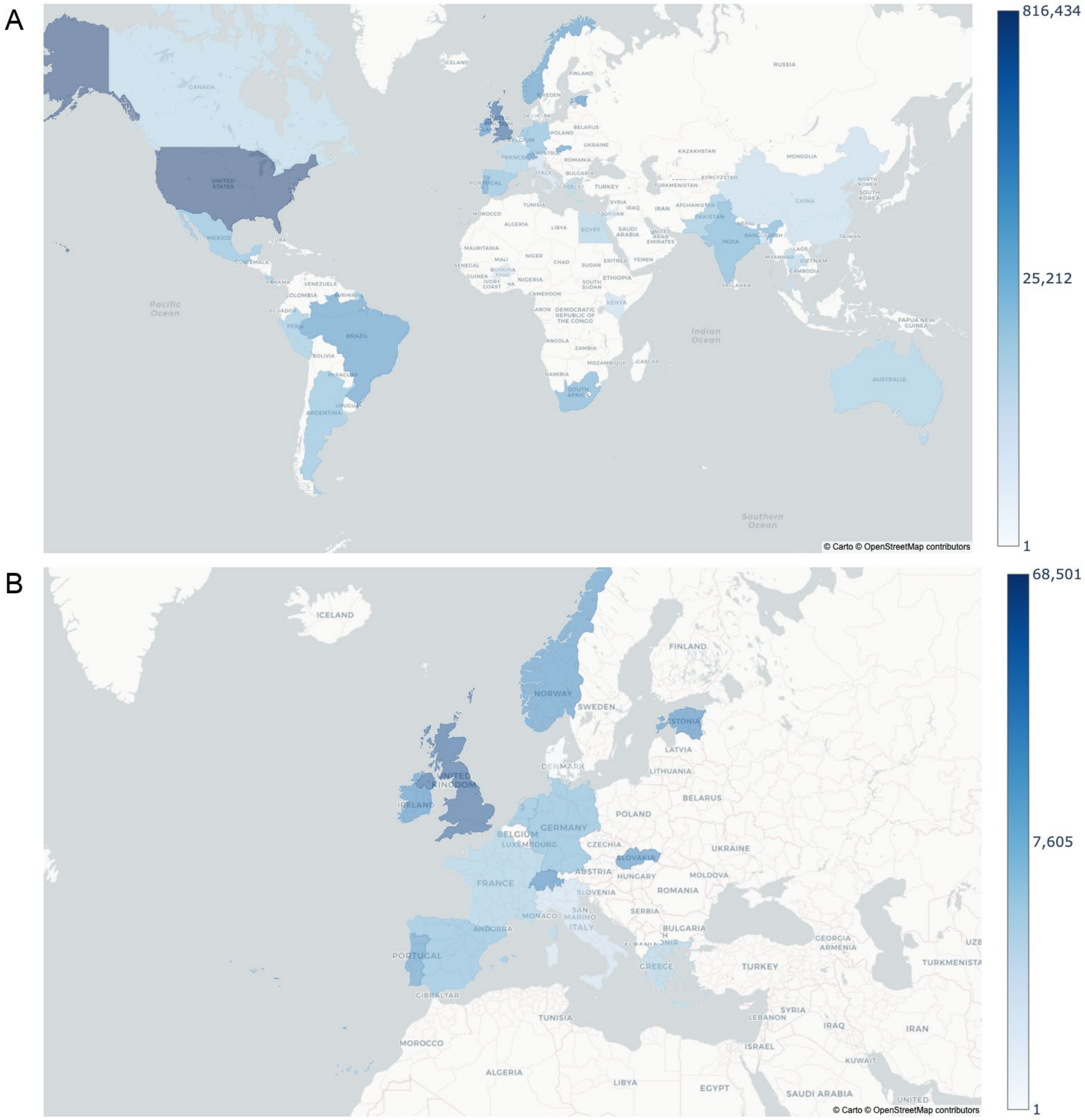


*Figure II: **New raw SARS-CoV-2 data** and distribution of sources at global (A) and European (B) levels **mobilized since 29 June 2022**. Note that the color scales are logarithmic best to show the broad range across countries.*

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

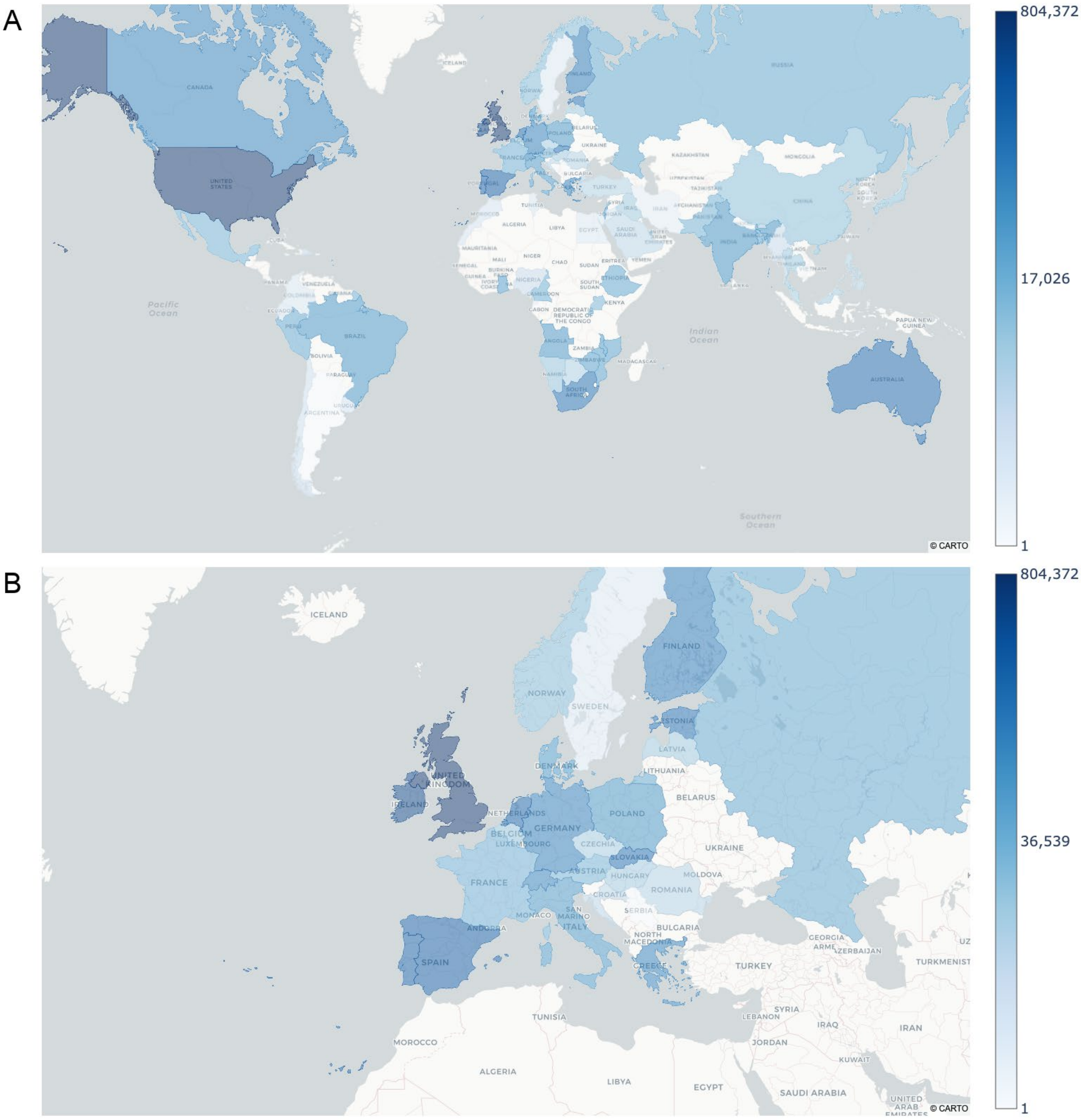


*Figure III: Geographical sources of **raw data processed through the workflow for variant calling**, comprising 3,750,754 datasets spanning the period of data first published from 05 Feb 2020 to 29 Aug 2022 globally (A) and within Europe (B). Note that the color scales are logarithmic best to show the 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, 1,834,210 of the 3,750,754 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. 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. We will report on this in the next update.

### Mutations and variants

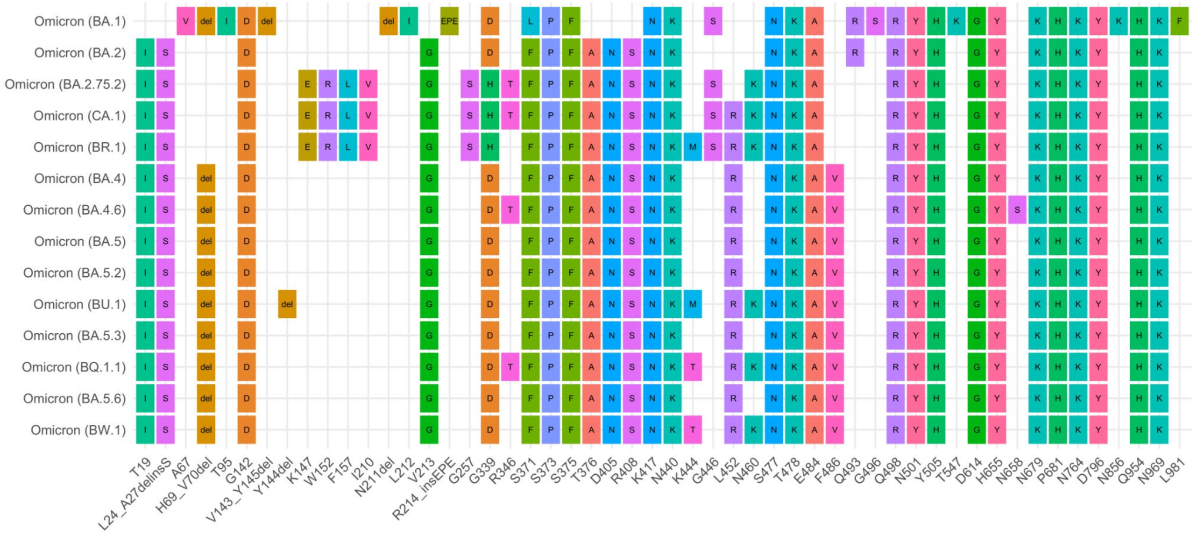
Several variants of concern (VOCs) and variants of interest (VOIs) have been identified since late 2020. All VOCs are defined by a set of mutations and other modifications along the genome and in the spike protein. Currently, only the Omicron variant is classified as VOC, and several pango sub-lineages have been identified that contain additional mutations. According to the WHO nomenclature, all of these sub-lineages are still referred to as the same VOC.

### Update as of 23 September 2022

Currently, three main Omicron lineages and sub-variants are circulating around the globe. These are descendants of the BA.2, the BA.4 and the BA.5 lineages. 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.4, 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.

Currently three different BA.2 variants are circulating: BR.1 (BA.2.75 + R346T + L452R), BR.2 (BA.2.75 + K444M + L452R) and CA.1 (BA.2.75 + L452R + R346T). The difference between BR.1 and CA.1 is that the mutations developed differently and are present in a different backbone. Next to that, BA.4.6 (BA.4 + R346T), BQ.1.1 (BA.5.3 + R346T + K444T + N460K), BU.1 (BA.5.2 + K444M + N460K) and BW.1 (BA.5.6 + K444M + N460K) are circulating among others. A summary of the amino acid changes in the spike protein of the Omicron sub-lineages is shown in Table II.

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



| | T19 | L24_A27delinsS | A67 | H69_V70del | T95 | G142 | V143_Y145del | Y144del | K147 | W152 | F157 | I210 | N211del | L212 | V213 | R214_insEPE | G257 | G339 | R346 | S371 | S373 | S375 | T376 | D405 | R408 | K417 | N440 | K444 | G446 | L452 | N460 | S477 | T478 | E484 | F486 | Q493 | G496 | Q498 | N501 | Y505 | T547 | D614 | H655 | N658 | N679 | P681 | N764 | D796 | N856 | Q954 | N969 | L981 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Omicron (BA.1) | | | V | del | I | D | del | | | | | | del | I | | EPE | | D | | L | P | F | | | | N | K | | S | | | N | K | A | | R | S | R | Y | H | K | G | Y | | K | H | K | Y | K | H | K | F |
| Omicron (BA.2) | I | S | | | | D | | | | | | | | | G | | | D | | F | P | F | A | N | S | N | K | | | | | N | K | A | | R | | R | Y | H | | G | Y | | K | H | K | Y | | H | K | |
| Omicron (BA.2.75.2) | I | S | | | | D | | | E | R | L | V | | | G | | S | H | T | F | P | F | A | N | S | N | K | | S | | K | N | K | A | | | | R | Y | H | | G | Y | | K | H | K | Y | | H | K | |
| Omicron (CA.1) | I | S | | | | D | | | E | R | L | V | | | G | | S | H | T | F | P | F | A | N | S | N | K | | S | R | K | N | K | A | | | | R | Y | H | | G | Y | | K | H | K | Y | | H | K | |
| Omicron (BR.1) | I | S | | | | D | | | E | R | L | V | | | G | | S | H | | F | P | F | A | N | S | N | K | M | S | R | K | N | K | A | | | | R | Y | H | | G | Y | | K | H | K | Y | | H | K | |
| Omicron (BA.4) | I | S | | del | | D | | | | | | | | | G | | | D | | F | P | F | A | N | S | N | K | | | R | | N | K | A | V | | | R | Y | H | | G | Y | | K | H | K | Y | | H | K | |
| Omicron (BA.4.6) | I | S | | del | | D | | | | | | | | | G | | | D | T | F | P | F | A | N | S | N | K | | | R | | N | K | A | V | | | R | Y | H | | G | Y | S | K | H | K | Y | | H | K | |
| Omicron (BA.5) | I | S | | del | | D | | | | | | | | | G | | | D | | F | P | F | A | N | S | N | K | | | R | | N | K | A | V | | | R | Y | H | | G | Y | | K | H | K | Y | | H | K | |
| Omicron (BA.5.2) | I | S | | del | | D | | | | | | | | | G | | | D | | F | P | F | A | N | S | N | K | | | R | | N | K | A | V | | | R | Y | H | | G | Y | | K | H | K | Y | | H | K | |
| Omicron (BU.1) | I | S | | del | | D | | del | | | | | | | G | | | D | | F | P | F | A | N | S | N | K | M | | R | K | N | K | A | V | | | R | Y | H | | G | Y | | K | H | K | Y | | H | K | |
| Omicron (BA.5.3) | I | S | | del | | D | | | | | | | | | G | | | D | | F | P | F | A | N | S | N | K | | | R | | N | K | A | V | | | R | Y | H | | G | Y | | K | H | K | Y | | H | K | |
| Omicron (BQ.1.1) | I | S | | del | | D | | | | | | | | | G | | | D | T | F | P | F | A | N | S | N | K | T | | R | K | N | K | A | V | | | R | Y | H | | G | Y | | K | H | K | Y | | H | K | |
| Omicron (BA.5.6) | I | S | | del | | D | | | | | | | | | G | | | D | | F | P | F | A | N | S | N | K | | | R | | N | K | A | V | | | R | Y | H | | G | Y | | K | H | K | Y | | H | K | |
| Omicron (BW.1) | I | S | | del | | D | | | | | | | | | G | | | D | | F | P | F | A | N | S | N | K | T | | R | K | N | K | A | V | | | R | Y | H | | G | Y | | K | H | K | Y | | H | K | |

*Table II. Overview of the different mutations of several sub-lineages of Omicron for the spike gene. Additional mutations are present in other parts of the genome.*

**CoVEO web app development**

CoVEO, a web-based application (https://coveo.vo.elte.hu/report/report-1/), was created that communicates with the database and 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 John Hopkins Coronavirus Resource Center (https://coronavirus.jhu.edu/map.html). PLEASE NOTE that the statistics on the number of cases are not very reliable given the current differences in policies.
- 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.

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 of the release of this variant report.

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.

**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).