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

Report No. 11 – 13 April 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.

- A major step forward is that the COVID-19 Data Portal now provides SARS-CoV-2 sequence data and analysis tools via a new feature, the CoVEO app. That means that the analyses include variant calling (filtered and unfiltered) and assembly 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/sequences?db=sra-analysis-covid19

- 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

Updates on data submissions

- Update on mobilisation of raw reads, now totaling sequencing datasets from 4,139,890 viral raw read sets from 92 countries, a 56% 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. 254,935 Illumina and 170,657 Oxford Nanopore samples have been processed in the period since the last 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.

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. As a consequence, many platforms that provide real-time analysis have started to work with partial datasets, requiring some type of selection of data prior to analysis. We are exploring how to scale up the analysis functions of the European COVID-19 Data Platform and ENA infrastructure to allow real-time analysis without the need for downsizing. We have therefore tested the use of alternative higher-capacity analysis environments (such as a new high-performance internal EMBL-EBI cluster and a commercial cloud service), methods to stream, rather than batch, data through our services and more bandwidth-efficient transfers across the networks between VEO partners involved in the analysis. These explorations have resulted in both a greater ability for the future to process at scale and an unusually large data set for the forthcoming report. We note that up to our 10th report (December 2021), we had processed a total of 925,000 raw datasets over the year up to the 10th report; in the forthcoming report alone we will have processed some 500,000 raw data sets. Henceforth, we will return to monthly reporting and expect that as a result of the advances, we will continue to complete analyses at a greater, and perhaps increasing, rate. Over the next few reports as we use the greater capacity, we will gain a picture of expected processing rates and provide some projections.

Section I: Data mobilisation

The number of read datasets released into the COVID-19 Data Portal up to the current data freeze (5 Apr 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).

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

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

Date		**10 July 2021**	**01 Aug 2021**	**21 Sept 2021**	**19 Oct 2021**	**07 Dec 2021**	**05 Apr 2022**
Raw read datasets	Total	872,011	1,056,105	1,549,740	1,876,126	2,672,038	4,139,890
	Illumina	703,104	861,866	1,239,284	1,502,424	2,217,465	3,551,782
	Oxford Nanopore	106,732	123,021	151,031	172,654	213,259	341,021
	Other	62,175	71,218	159,425	201,048	241,314	247,087
Source countries for raw read datasets		69	69	75	80	85	92

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

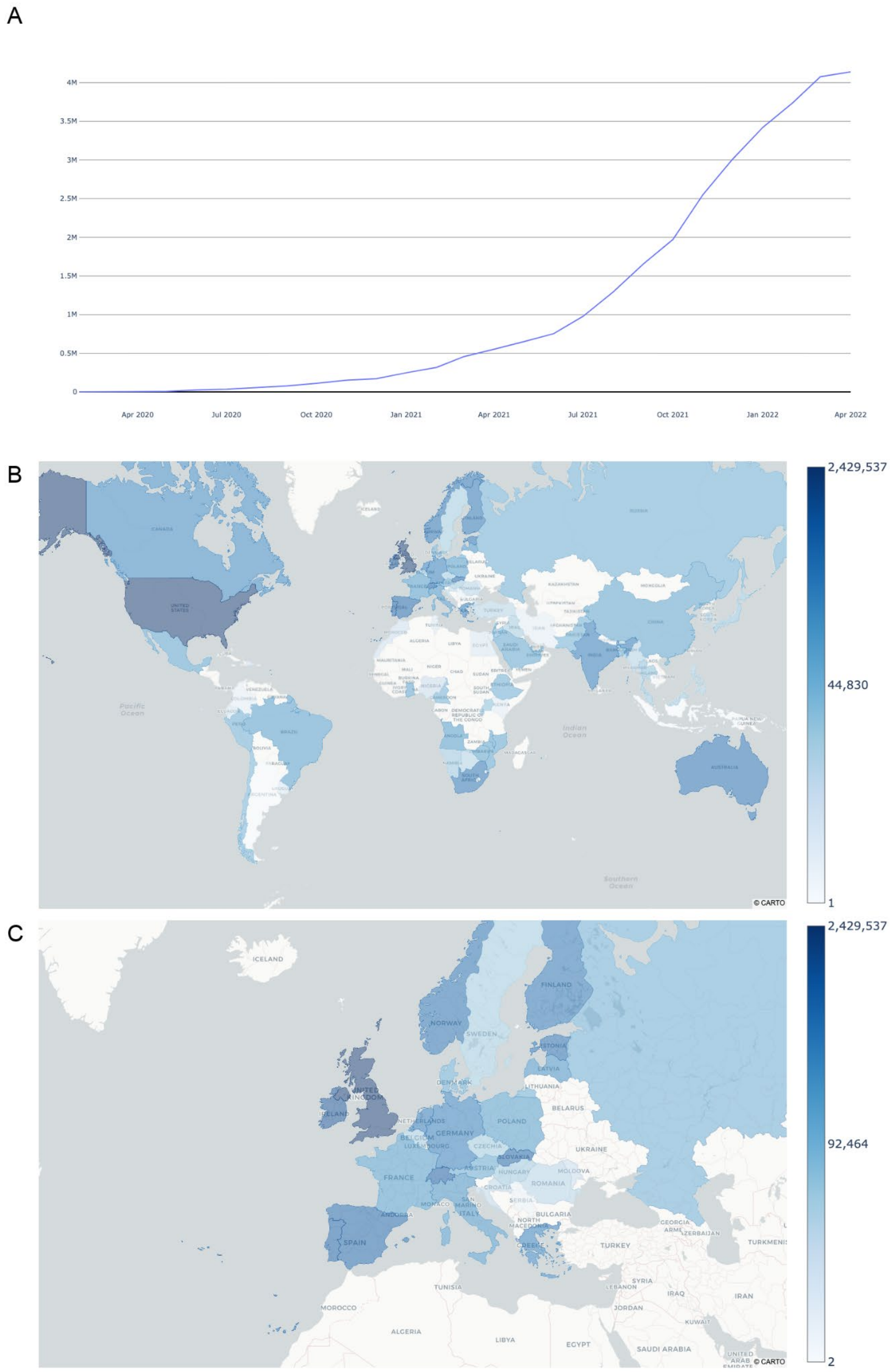


*Figure I: **Globally available total number of raw SARS-CoV-2 data** and distribution of sources, showing (A) sustained growth in raw data since launch of the mobilisation campaign by cumulative number of data sets, (B) and (C) geographical sources of global and European raw data, respectively, for which 65% of global data have been routed through the SARS-CoV-2 Data Hubs, with the remaining 35% 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).

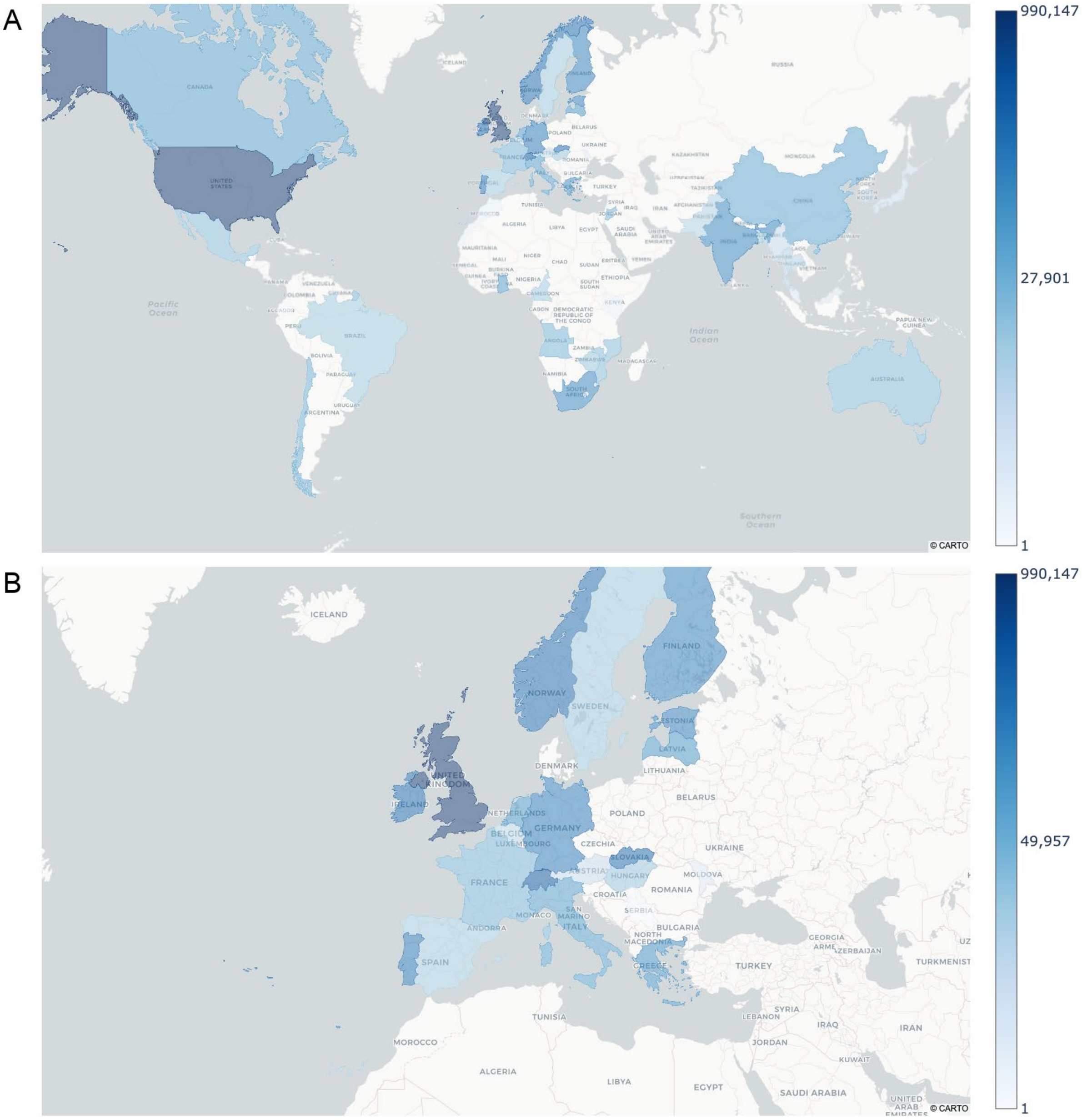


*Figure II: **New raw SARS-CoV-2 data** and distribution of sources at global (A) and European (B) levels **mobilized since 7 Dec 2021**. Note that the color scales are logarithmic best to show the broad range across countries.*

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

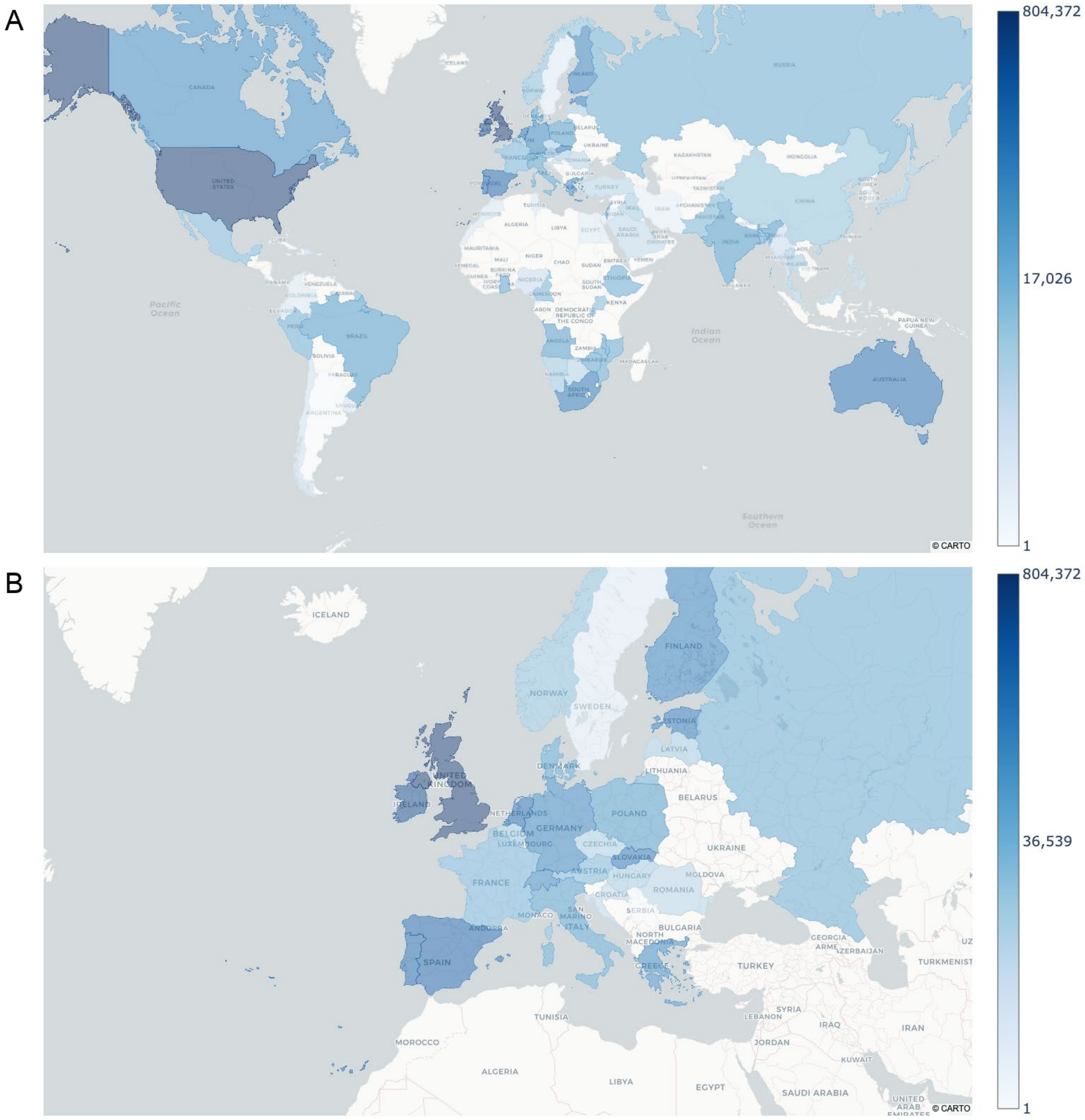


Figure III: Geographical sources of ***raw data processed through the workflow for variant calling****, comprising 1,456,113 datasets spanning the period of data first published from 05 Feb 2020 to 09 Mar 2022 globally (A) and within Europe (B). This represents a 57.4% increase on the previous report. Note that the color scales are logarithmic best to show the broad range across countries.*

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. Below are summaries of the main findings.

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. For the Beta, Gamma, Delta and Omicron variants, some 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.

It is important to monitor these variants in time and space and to assess the relevance of these variants. Therefore, a rolling literature review is performed to summarize studies assessing the virulence, pathogenicity and potential immune escape of these different variants. The updates are provided to the WHO Technical Advisory Group on SARS-CoV-2 Virus Evolution (TAG-VE), which combines the findings with epidemiological data. Based on review in the TAG-VE, variants may be published as variants of concern, and given a name. For each new variant of concern, the combination of mutations will be included in the raw read analysis in this report. VOCs that have not been detected for a certain period of time are declassified as such by WHO. At the moment, two VOCs have been designated by the WHO: Delta (B.1.617.2, AY.x) and Omicron (B.1.1.529, BA.x).

Update as of 05 April 2022

In most countries, the VOC Omicron is the dominating variant. As of 05 April 2022, the Omicron variant is found in at least 173 countries globally. Some sub-lineages and recombinants of the Omicron have been identified; these sub-lineages contain additional mutations that might be of biological importance. This resulted in a major update of the pangolin classification tool through which the classification of sequences in GISAID might change (e.g. adding sub-lineage like BA.1.17.1, BA.2.3 etc.).

The latest identified sub-lineages are the proposed BA.4 and BA.5 variants, which are at the moment mainly detected in South Africa but seem to have an advantage over the other Omicron variants currently circulating around the globe. These sub-lineages contain three additional mutations in the spike protein compared with BA.2 (69-70del, L452R and F486V) and have reversion mutation to the wild type virus (Q493).

Several recombinant viruses have been detected that are classified as XA-XQ, of which most are recombinants of BA.1 and BA.2 viruses. The detection and identification of these recombinants have shown to be challenging based on consensus sequences only due to potential sequencing and analysis artifacts. This has also complicated the detection and tracing of these variants in GISAID since a quality check has to be performed before these

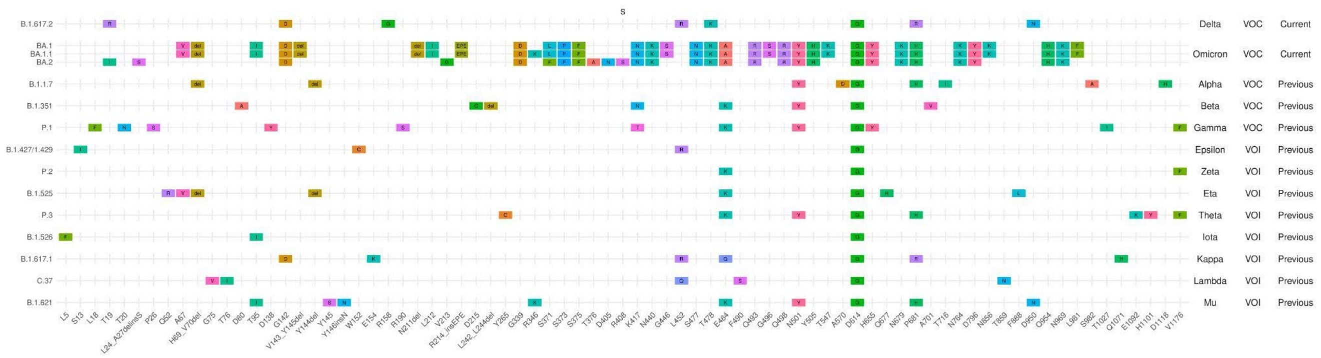
This work is supported by funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 874735 (VEO).

viruses can be accurately classified. A summary of the potential phenotypic impact based on current available literature for the VOCs is summarized in Table II, and the locations of the mutations in the spike protein is shown in Table III.

Table II: Overview of VOCs and their phenotypic impact. N: evidence from neutralization assays; VE: evidence from vaccine effectiveness/efficiency studies.

WHO Label	Pango lineage	Transmissibility	Disease Severity	Immune Escape (natural acquired immunity)	Vaccine Escape (vaccine acquired immunity)
Delta	**B.1.617.2**	Increased	Possible increased risk of hospitalization	N: Reduced neutralization capacity against antibodies elicited by infection	N : Reduced neutralization capacity against antibodies elicited by vaccination (--) VE: Reduced protection against symptomatic disease and infection
Omicron	**B.1.1.529**	Increased	Association with less severe disease	N: Strongly reduced neutralization capacity against antibodies elicited by infection	N: Strongly reduced neutralization capacity against antibodies elicited by vaccination VE: significantly reduced protection from symptomatic infection

Table III. Overview of the different mutations of several current and previous VOCs and VOIs for the spike gene. Additional mutations are present in other parts of the genome.



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

Delta variant (B.1.617.2 + AY.x)

Samples containing Delta variant lineage defining spike protein mutations (T19R, L452R, T478K, P681R, D950N) have been found in raw reads from the countries as shown in the table below. Due to the many non-variant sequences, they are only clearly visible for some countries in the bar charts.

	ENA	GISAID
Angola	171	276
Bangladesh	1	2475
Bhutan	26	0
Botswana	40	1306
Brazil	64	42450
Cameroon	145	361
Canada	1	120120
Estonia	4109	4246
Ethiopia	101	435
Finland	16	13212
Germany	71	208033
Ghana	73	1138
Greece	1576	5031

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

India	451	114882
Ireland	16583	63194
Malawi	79	438
Mauritius	18	314
Mexico	7	41580
Mozambique	165	417
Namibia	46	144
Netherlands	402	45725
Northern Mariana Islands	46	866
Pakistan	17	1241
Philippines	1	3428
Poland	182	29981
Portugal	12337	15209
Romania	1	6082
Russia	1	8090
Serbia	1	175
Seychelles	80	865
Slovakia	5293	14401

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

South Africa	2187	12079
Spain	2588	45923
Switzerland	1662	60162
Thailand	80	9264
Uganda	17	458
United Kingdom	114863	1157870
USA	140128	1471694
Zimbabwe	146	149

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

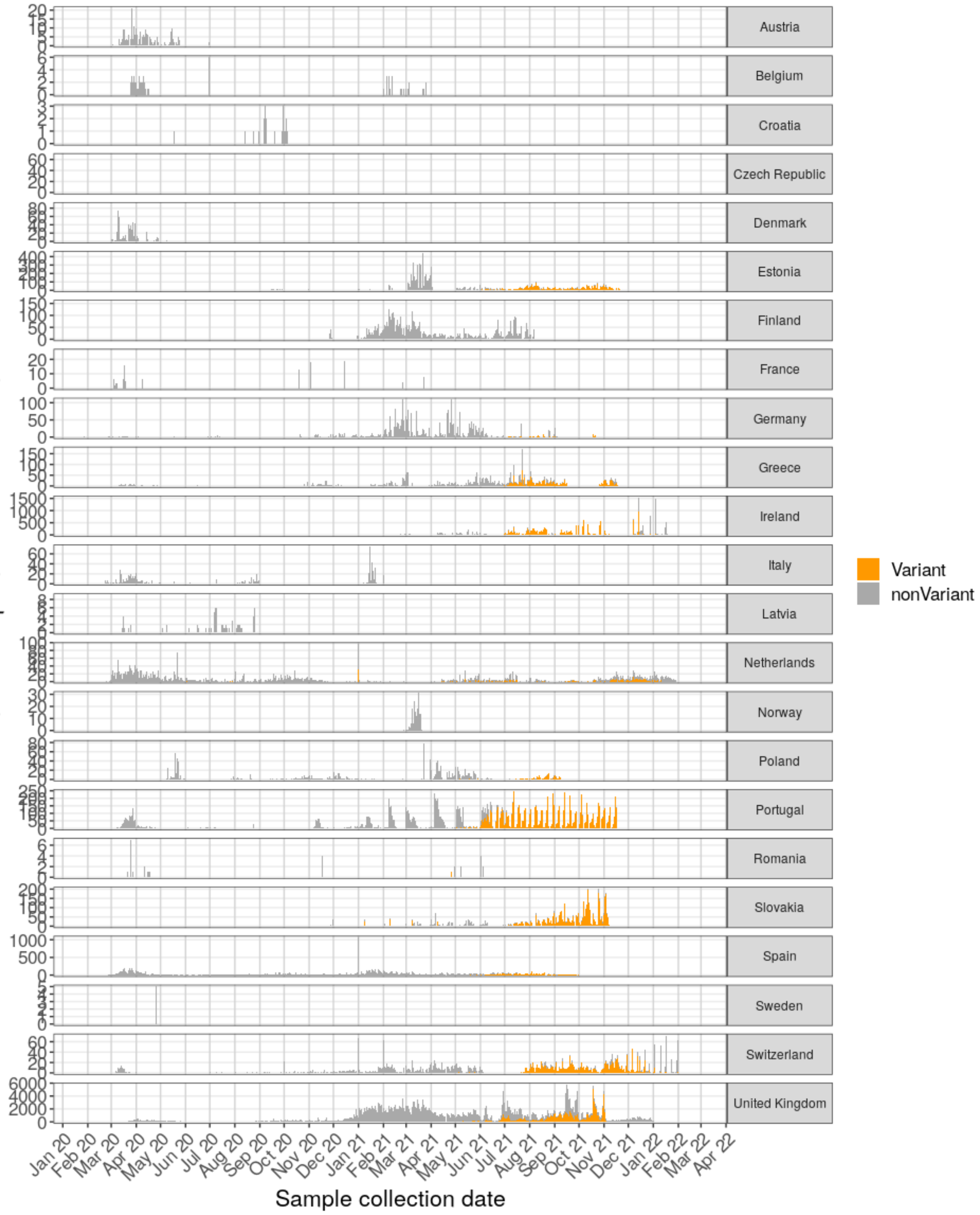


Figure IV: Number of sequences by date of sampling for variant Delta variant (orange) for European countries.

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

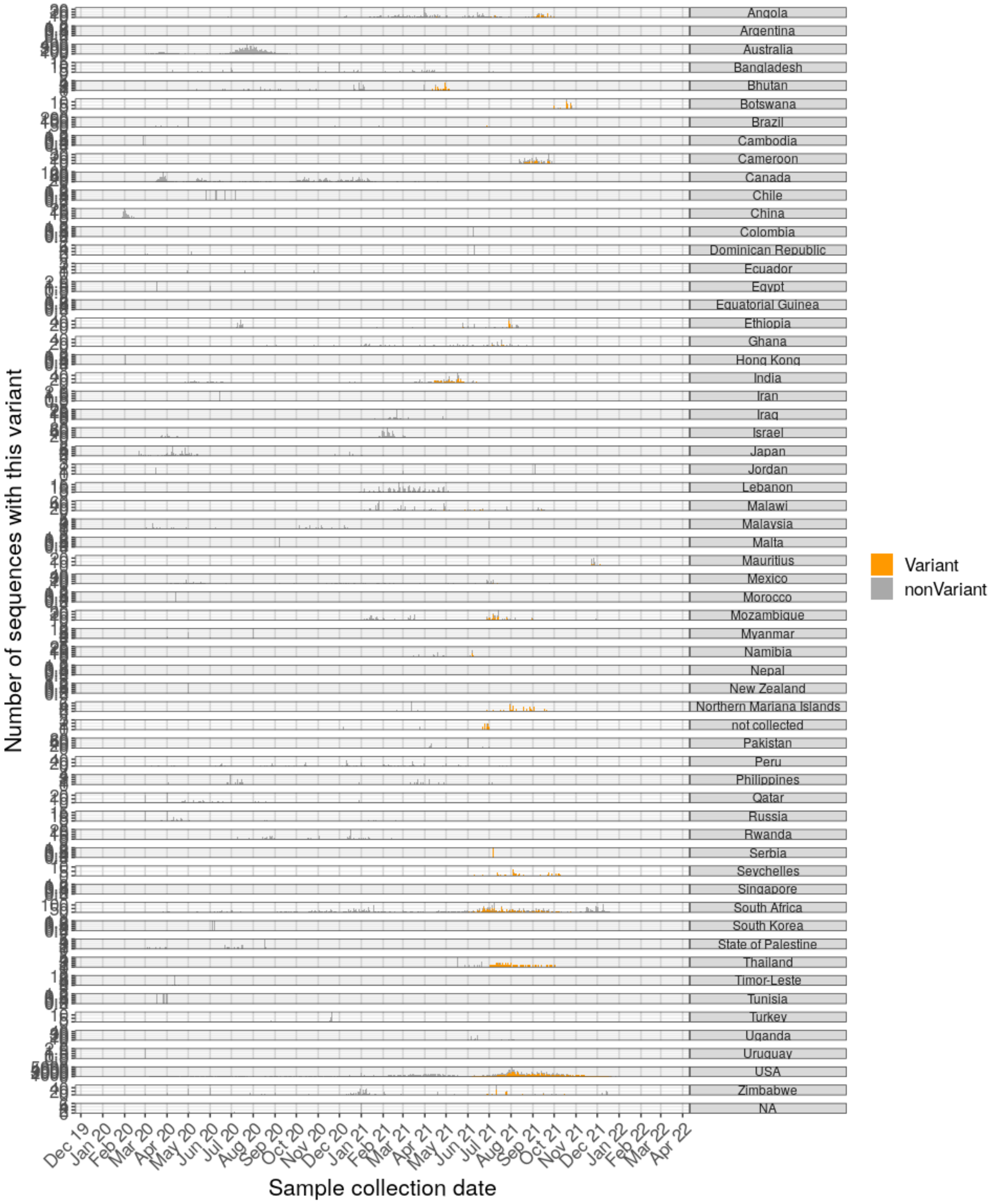


Figure V: Number of sequences by date of sampling for variant Delta variant (orange) for countries outside of Europe.

Omicron (BA.1 and BA.2)

On 26 Nov 2021, WHO declared B.1.1.529 as a VOC, based on its rapid spread in different regions of South Africa and the unusually high number of mutations. The Omicron variant contains multiple sub-lineages of which BA.1 and BA.2 are most frequently found.

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

BA.1

The BA.1 variant is characterized by the following mutations in the spike protein: A67V, T95I, G339D, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, N856K, Q954H, N969K, L981F. An additional 7 characteristic spike indels and SNPs are present but these could not yet reliably be called by the variant calling process and were therefore excluded.

Seven mutations in the spike (A67V, T95I, G446S, G496S, T547K, N856K, L981F) are specific for the BA.1 lineage and do not occur in the BA.2 lineage. The table below shows those samples containing these BA.1 specific mutations.

	ENA	GISAID
United Kingdom	281	369731
Ireland	4615	7961
South Africa	318	4067
Mozambique	16	58
Zimbabwe	23	37
USA	458	497109
Netherlands	7	13193
Switzerland	220	10072

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

BA.2

The BA.2 can be distinguished from the BA.1 variant using the following characteristic mutations in the spike protein: T19I, V213G, S371F, T376A, D405N, R408S. In the table below, the samples containing these six mutations are listed:

	ENA	GISAID
United Kingdom	4	326948
South Africa	3	1819
Ireland	13	7432
Netherlands	12	6930
Switzerland	2	5930

Delta/Omicron recombinants and BA.4/BA.5

With the massive and globally synchronised wave of Omicron infection, additional variants have developed. These are an increasing number of recombinant genomes, with mixtures of BA.1 and BA.2 as well as Delta/Omicron recombinants. Recently, two new Omicron variants have been observed in several different countries including South Africa, Germany, France, Denmark, Austria and Ireland. These two new variants are characterized by having three additional mutations in the spike region compared to the original BA.2 variant (Delta69-70, L452R and F486L). Next to that, one reversion to the wild type virus was observed in the spike (Q493). The two different variants can be distinguished by specific mutations outside the spike region. Although these specific mutations appear to be worrisome, at the moment, the impact of the constellation of these genetic changes is unknown. The spread of these variants will be monitored and more details will be provided in the next report.

CoVEO

To support visualisations and reports for analysed SARS-CoV-2 data, a database was created that is hosted on the servers of EBI. Following systematic analysis of raw data, resulting pipeline output with mutations and metadata information is pulled and stored within this database.

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

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 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.

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. A full announcement is expected shortly.

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.

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).