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

Report No. 10 – 15 December 2021

Summary:

- Update on mobilisation of raw reads, now totaling sequencing data sets from 2,656,983 viral raw read sets from 85 countries, a 42% 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 Oxford Nanopore data has been deployed on the Google Cloud Platform, allowing us to process the backlog of data. All 213,259 Oxford Nanopore samples have been processed.
- While data mobilisation is progressing, the contribution of countries in Europe is very low. Pathogen sequencing in most countries has been taken up as part of a public health effort, in part supported by HERA/ECDC. Agreements are needed to ensure release of raw data towards global sharing effort.

Background:

This report summarizes mobilisation and analysis of SARS-CoV-2 sequence data submitted to the European COVID-19 Data Platform 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.

Section I: Data mobilisation

The number of read datasets released into the COVID-19 Data Portal up to the current data freeze (7 Dec 2021) 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).

When comparing the number of entries in ENA with the number of entries in GISAID (5,871,503), **assuming** raw reads would be shared for all GISAID entries, it is clear that the number of raw read datasets lags far behind the sharing of assembled genomes in GISAID.

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

Date		**14 June 2021**	**10 July 2021**	**01 Aug 2021**	**21 Sept 2021**	**19 Oct 2021**	**07 Dec 2021**
Raw read datasets	Total	679,693	872,011	1,056,105	1,549,740	1,876,126	2,672,038
	Illumina	575,481	703,104	861,866	1,239,284	1,502,424	2,217,465
	Oxford Nanopore	93,581	106,732	123,021	151,031	172,654	213,259
	Other	7,134	62,175	71,218	159,425	201,048	241,314
Source countries for raw read datasets		66	69	69	75	80	85

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

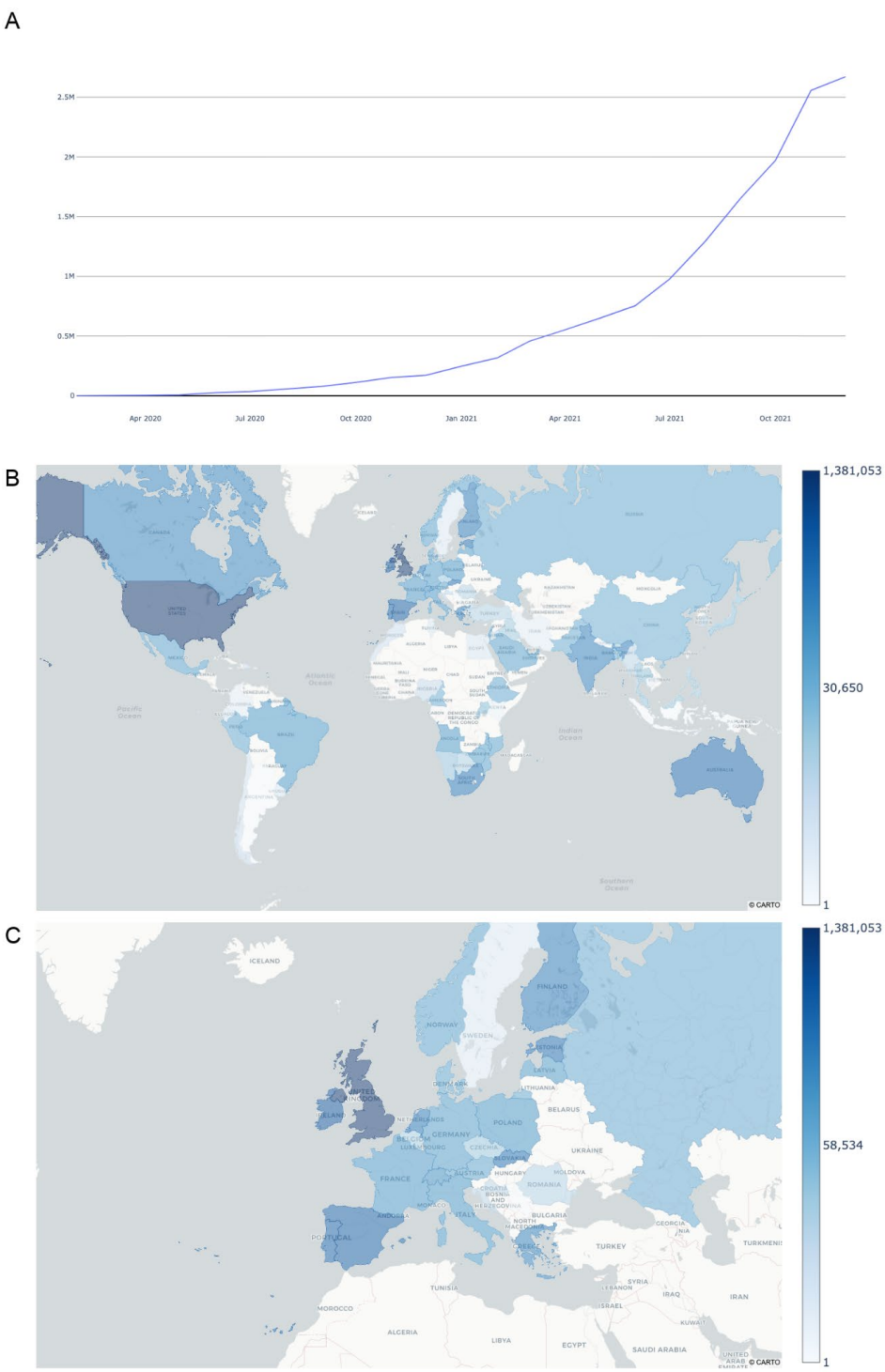


Figure I: Globally available raw SARS-CoV-2 data and distribution of sources, showing (A) sustained growth in raw data since launch of the mobilisation campaign by cumulative number of data sets, (B) and (C) geographical sources of global and European raw data, respectively, for which 59% of global data have been routed through the SARS-CoV-2 Data Hubs, with the remaining 41% 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).

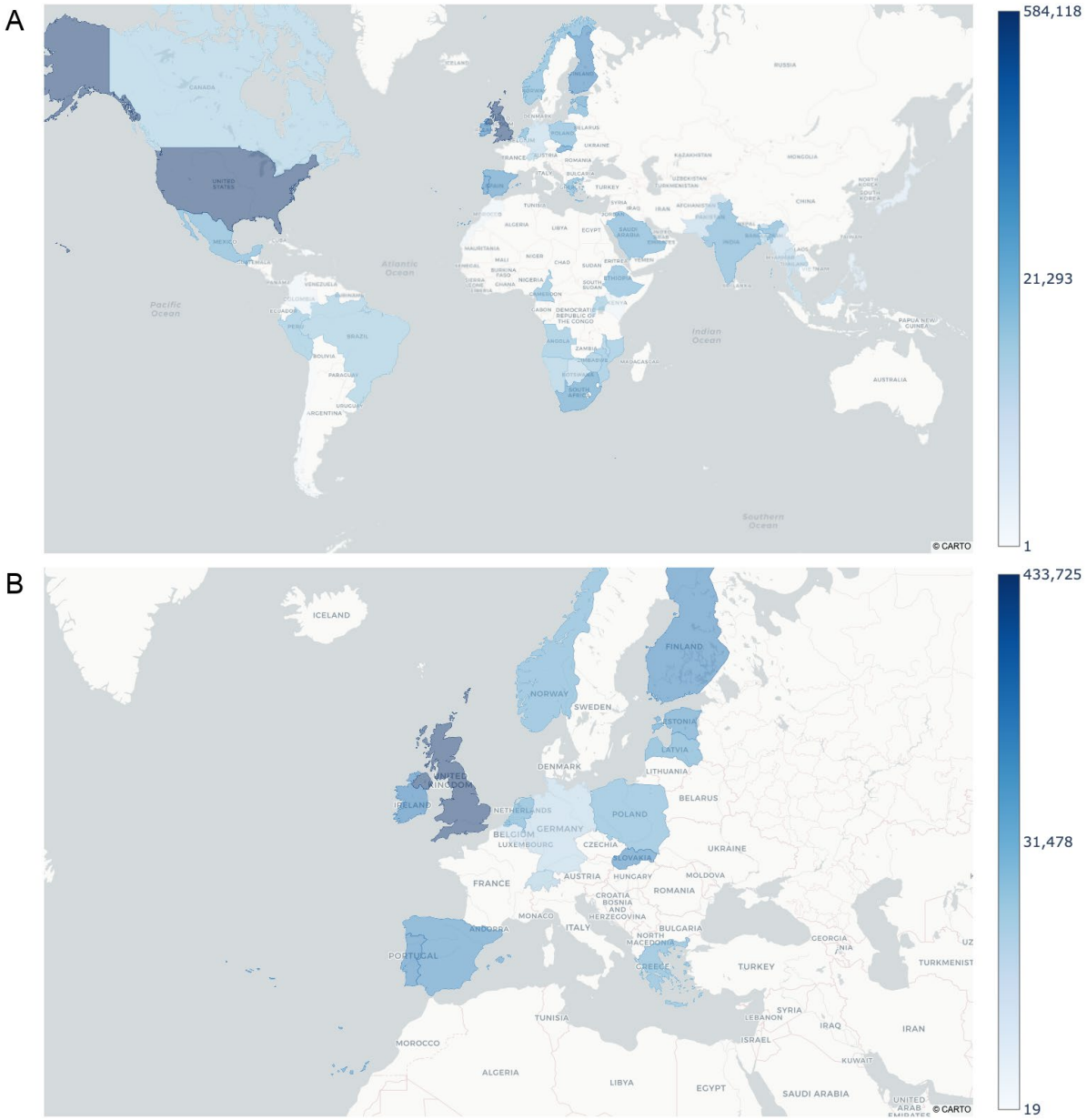


Figure II: New raw SARS-CoV-2 data and distribution of sources at global (A) and European (B) levels mobilised since 19 Oct 2021. Note that the colour scales are logarithmic best to show the broad range across countries.

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

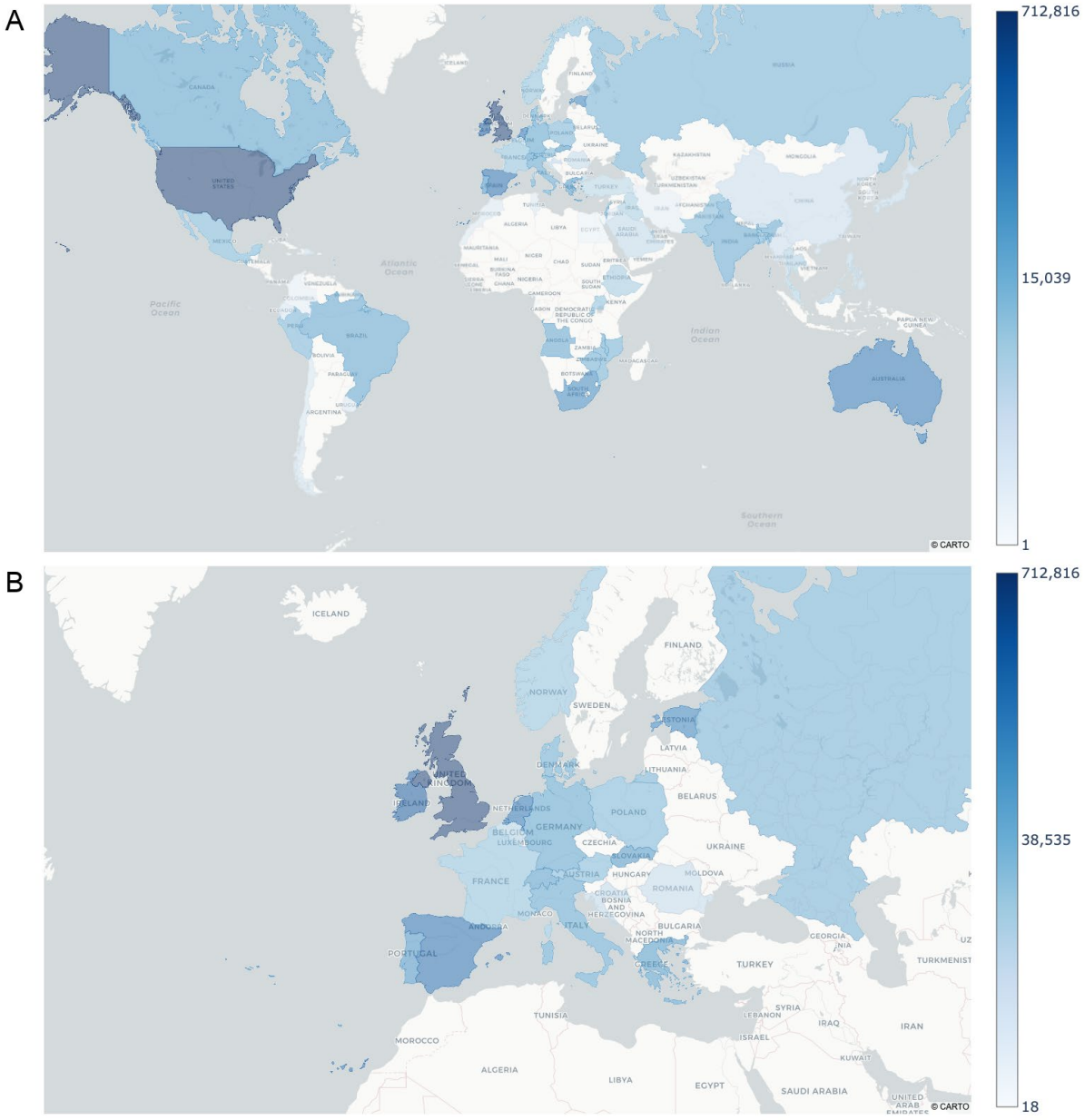


Figure III: Geographical sources of analysed raw data comprising 924,907 data sets spanning the period of data first published from 05 Feb 2020 to 16 Nov 2021 globally (A) and within Europe (B). Note that the colour scales are logarithmic best to show the broad range across countries.

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

Results of variant calling

A workflow to analyse 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 based on the data submitted and/or made public from Jan 2020 until 23 Nov 2021.

Mutations and variants

Several variants of concern (VOC) and variants of interest (VOI) have been identified since late 2020. 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 evolution group, which combines the findings with epidemiological data. Based on review in the evolution working group, 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.

Update as of 06 December 2021

In many countries, the VOC Delta is still the dominating variant. As of 06 Dec 2021, the Delta variant is found in at least 174 countries globally. The VOIs Lambda (C.37) and Mu (B.1.621), both originating in South America, are found in 44 and 72 countries, respectively. Some sub-lineages of the VOCs have been identified; these sub-lineages contain additional mutations that might be of biological importance.

Most recently, a new variant has been identified, Omicron (B.1.1.529/BA.1 + BA.2). It was first identified in a sample from 08 Nov 2021 from the baseline surveillance in South Africa in the province Gauteng. Omicron has an unusually high number of genetic changes when compared with the Wuhan-1 strain, of which 34 are located in the spike protein. Epidemiological and laboratory studies are being performed to assess the phenotypic properties of Omicron.

Variants of concern

Below is a summary of the analysis of raw read datasets for the presence of the combination of mutations that define the different VOCs.

At the moment, five VOCs have been described: Alpha (B.1.1.7, Q.1-Q.8), Beta (B.1.351, B.1.351.1-B.1.351.5), Gamma (P.1, P.1.1-P.1.17.1), Delta (B.1.617.2, AY.x) and Omicron (B.1.1.529, BA.1 + BA.2). All of these VOCs are defined by a set of mutations and other modifications along the genome and in the spike protein. For the Beta, Gamma and Delta variants, some pango sub-lineages have been identified that contain additional mutations;

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

e.g., AY.1 and AY.2 contain the additional mutation K417N when compared with its parent lineage. According to the WHO nomenclature, all of these sub-lineages are still referred to as the same VOC.

Although the Delta variant is the most prevalent globally, Omicron is detected in an increasing number of countries. Current evidence suggests that the neutralising capacity of antibodies from infection or vaccinations against Omicron is far less than against Delta variant viruses. Initial estimates of vaccine efficacy suggest a considerable drop in VE, which may partially be restored through boosting. A summary of the potential phenotypic impact based on current available literature for the VOCs is summarized in Table II.

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)
Alpha	**B.1.1.7**	Increased (+++)	Association with increased hospitalization and mortality	No impact on neutralization capacity	No impact on neutralizing activity VE: no impact
Beta	**B.1.351**	Increased (+)	Possible increased risk of hospitalization and mortality (in-hospital)	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
Gamma	**P.1**	Increased (++)	Possible link with risk of hospitalization and mortality	N: Moderate reduced neutralization capacity against antibodies elicited by infection	N : Reduced neutralization capacity against antibodies elicited by vaccination (--) VE: limited evidence
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

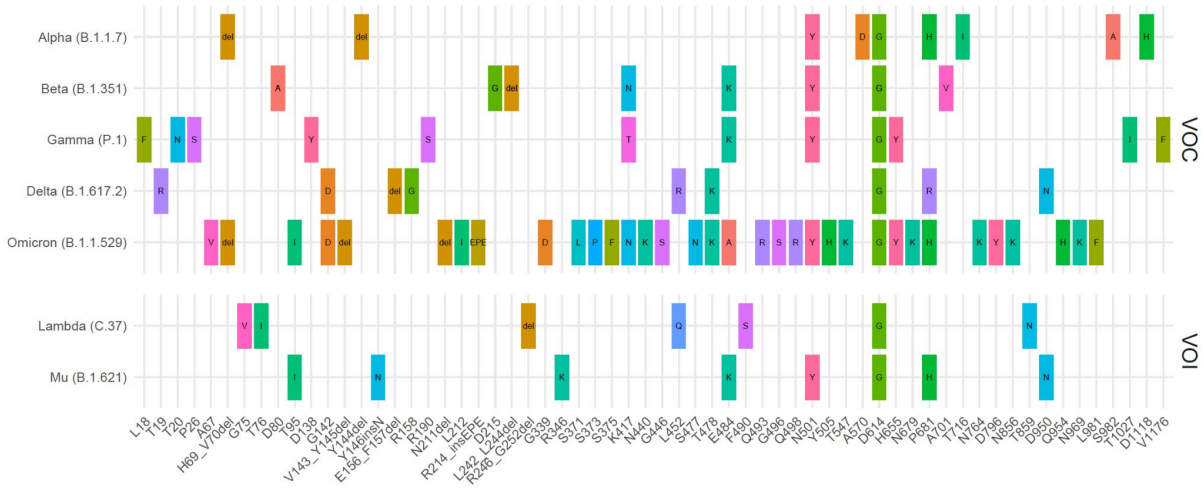
This work is supported by funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 874735 (VEO).

Omicron	**B.1.1.529**	Possibly increased	Unknown	N: Strongly reduced neutralization capacity against antibodies elicited by infection	N: Strongly reduced neutralization capacity against antibodies elicited by vaccination VE: initial estimates suggest significantly reduced protection from symptomatic infection

Variants of interest

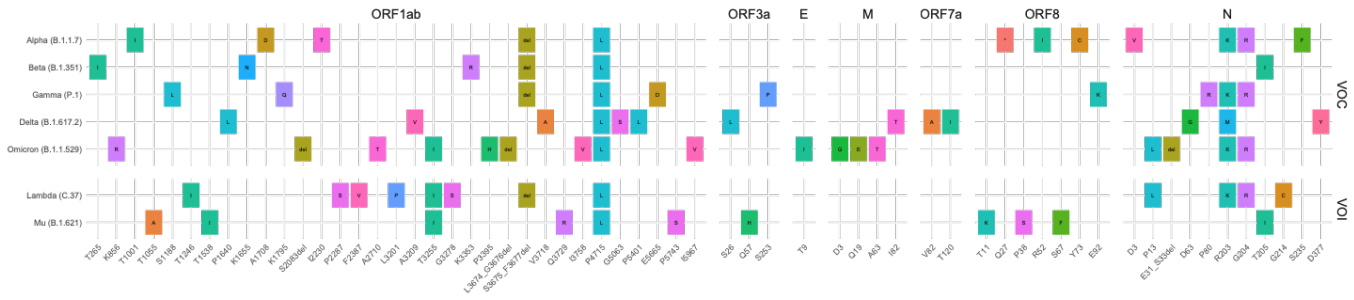
In addition to the VOCs, there have been several reports of Variants of Interest (VOIs) that contain one or more mutations of potential concern and have been found in multiple countries/cause multiple COVID-19 cases. No new variants of interest have been designated since the last report. Currently, Lambda (C.37) and Mu (B.1.621) belong to the VOIs.

Table III. Overview of the different mutations of several 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).

Table IV. Overview of the different mutations of several VOCs and VOIs for the ORF1ab-, ORF3a-, E-, M-, ORF6-, ORF7a- and ORF7b, ORF8 and N-gene.



Variants of Concern

Alpha variant (B.1.1.7), Beta variant (B1.351) and Gamma variant (P1) currently are found infrequently, and therefore will no longer be updated in this report.

For the current, most widely circulating or emerging VOCs, the data submitted since July 2020 have been analysed to determine the frequency of each variant in that dataset. The data are plotted for the countries that have released raw reads since July 2020, even if those were from patients sampled much earlier (Figures IV and V). This is visible as the plots are shown by date of sampling. In the global data collected at GISAID, Alpha variant viruses currently are still found occasionally, but have been replaced by Delta variant viruses.

Delta variant (B.1.617.2 + AY.x)

Samples containing all 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
United Kingdom	113611	961387
Netherlands	104	30724
USA	36410	1088445
Ireland	5968	25536
Switzerland	4	46410

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

Spain	653	30850
South Africa	228	11043
Canada	1	91521
Angola	8	165
Malawi	3	346
Slovakia	1300	10275
Greece	537	2988
Pakistan	9	750
Bangladesh	1	1376
Romania	1	4680
Portugal	1550	13051
Estonia	1093	3021
Bhutan	26	0
Uganda	17	340
Brazil	64	24523
Mexico	7	30220
Thailand	57	5006
Northern Mariana Islands	5	109

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

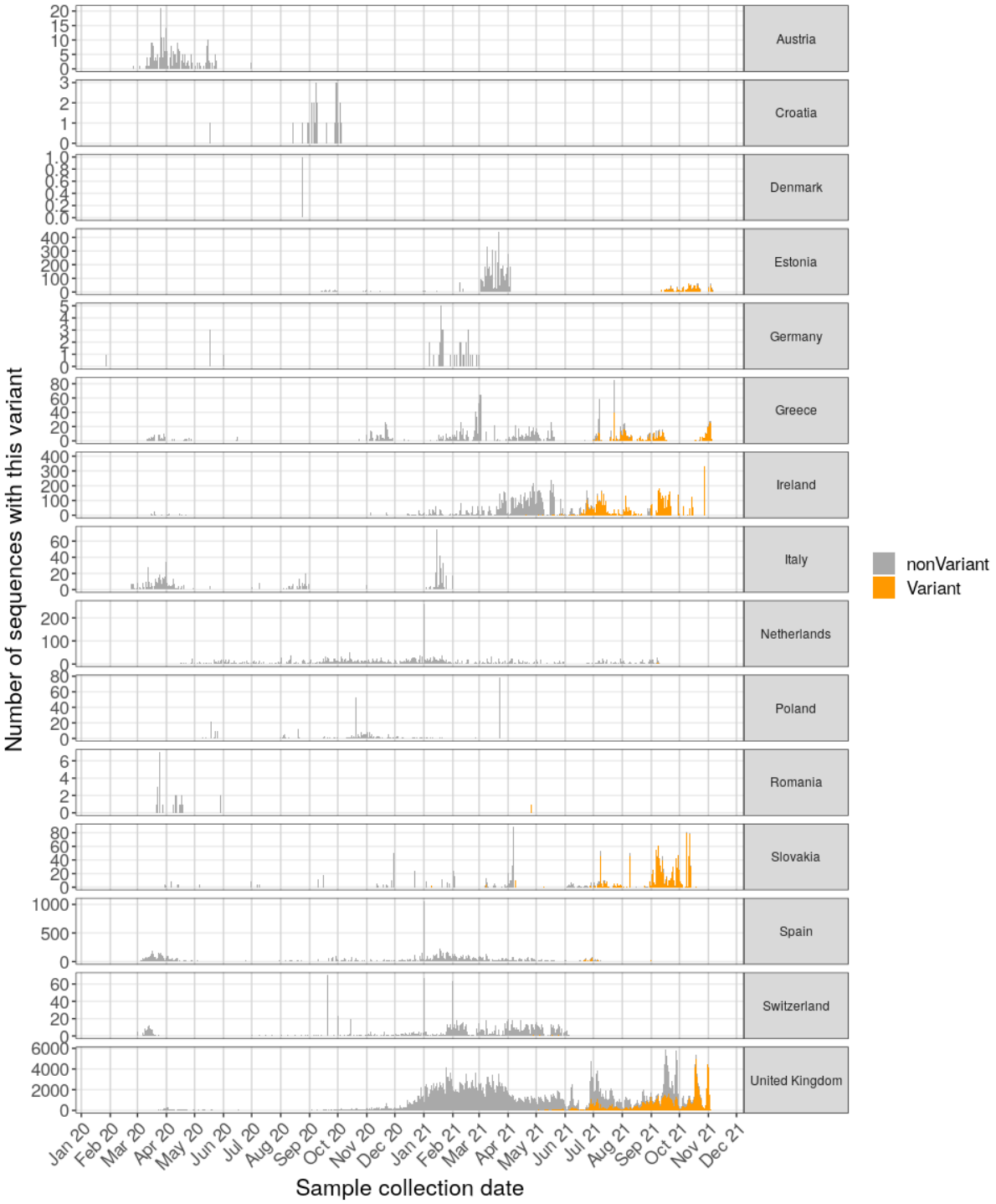


Figure IV: Number of sequences by date of sampling for variant Delta variant (orange) for European countries.

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

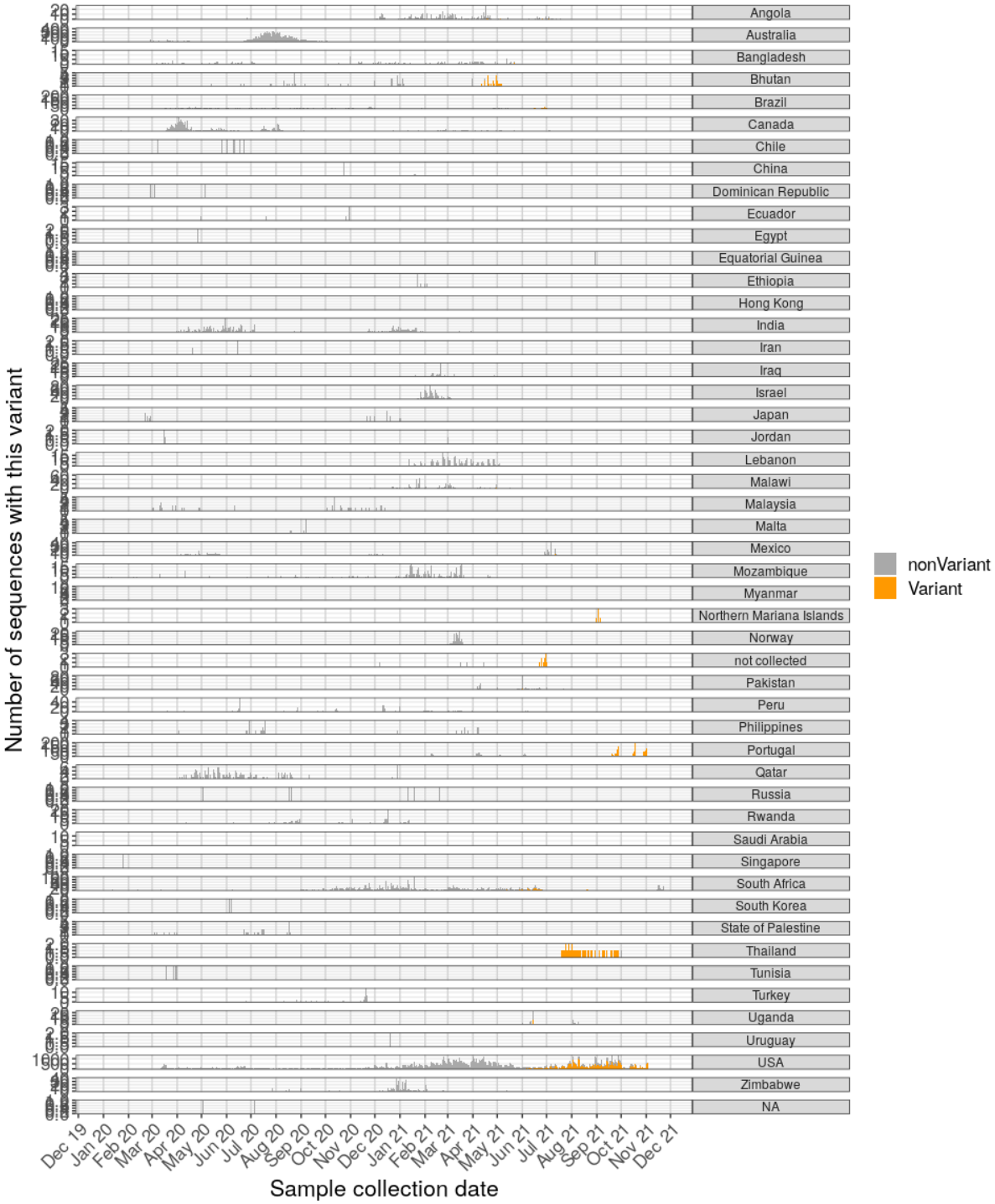


Figure V: Number of sequences by date of sampling for variant Delta variant (orange) for countries outside of Europe.

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

Omicron (B.1.1.529)

On 26 Nov 2021, WHO declared B.1.1.529 as a variant of concern, based on its rapid spread in different regions of South Africa and the unusually high number of mutations. Omicron is characterized by the following 27 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.

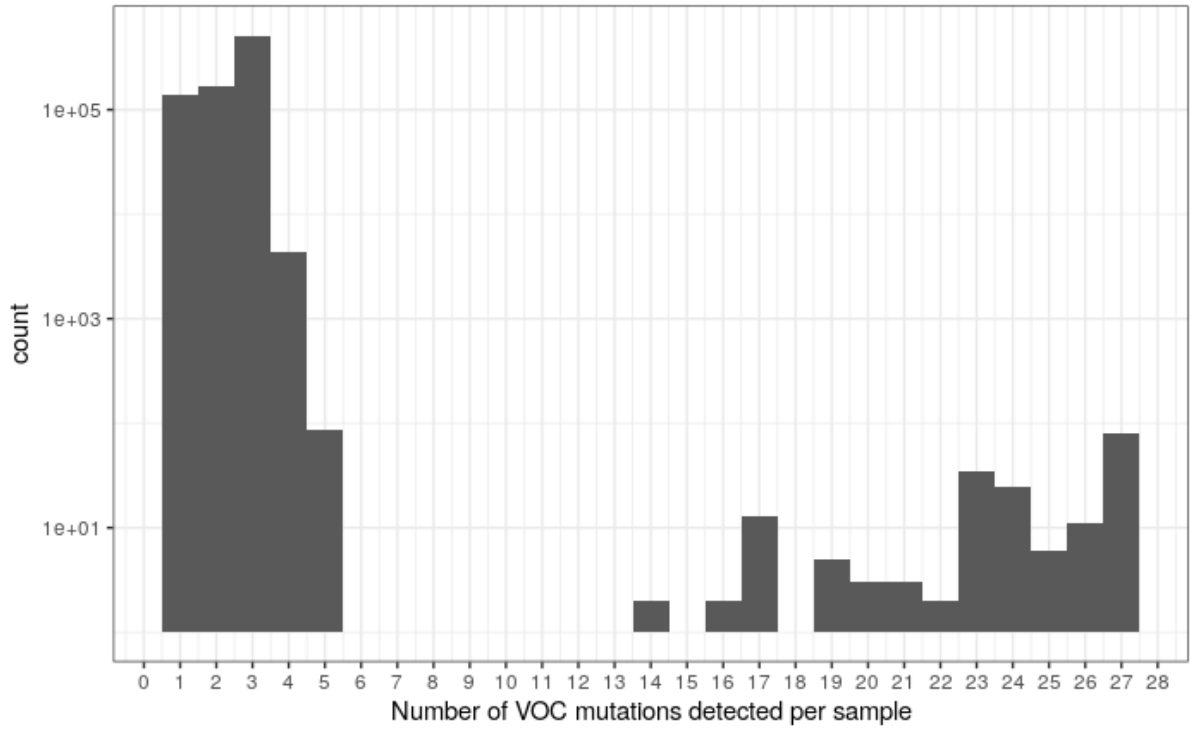


Figure VI: Number of sequences containing one or more of the characteristic Omicron SNPs in the spike region.

As can be seen in Figure VI, some of the read data contain less than the 27 aforementioned spike protein mutations but contain sufficient mutations to be suspected Omicron variants. Upon inspection, these datasets had too low coverage at some of the positions of these mutations due to amplicon failure.

Based on these selection criteria, 186 samples containing Omicron variant viruses could be identified at a threshold of at least 14 characteristic Omicron variant mutations in the spike.

In comparison, 2548 Omicron variant sequences have been deposited in GISAID. However, the samples reported here do represent resequenced data on the Illumina and GridION platform of the same biological specimens.

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

Lambda (C.37)

Lambda was characterized by the following mutations in the spike protein: G75V, T76I, L452Q, F490S, T859N.

ENA		GISAID
United Kingdom	5	8
Netherlands	1	12
USA	148	1264
Ireland	4	5
Spain	9	232
Peru	61	3946
Mexico	3	220

Mu (B.1.621)

Mu was characterized by the following mutations in the spike protein: T95I, R346K, E484K, N501Y, P681H, D950N.

ENA		GISAID
United Kingdom	17	71
Netherlands	1	73
USA	585	5859
Ireland	2	6

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

Switzerland	2	48
Spain	78	687
Mexico	1	473

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. A substantial number of raw reads has been publicly released but the geographical distribution continues to be highly skewed to a few countries, reflecting large-scale sequencing efforts. The number of raw sequencing data that are generated and shared from the EU member states are still limited and delayed, and more and earlier sharing of data is needed to provide a timely overview of circulating variants. 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).