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

Report No. 9 – 15 November 2021

Summary:

- Update on mobilisation of raw reads, now totaling sequencing data sets from 1,876,126 viral raw read sets from 80 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.
- The variant calling workflow for the Oxford Nanopore data has been implemented and 83,041 samples of the total 172,654 have been processed so far.
- 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 (Figure IV).

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 datasets released into the COVID-19 Data Portal up to the current data freeze (19 Oct 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, the number of raw read sets published is around 4%. This can only be seen as indicative, as it is not possible to link GISAID data with ENA entries. The analysis shows that most countries contributing to GISAID do not seem to share raw read data. It also is indicative of the need for mobilising data from European studies (Figure IV).

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

Date		**4 May 2021**	**14 June 2021**	**10 July 2021**	**01 Aug 2021**	**21 Sept 2021**	**19 Oct 2021**
Raw data sets	Total	552,185	679,693	872,011	1,056,105	1,549,740	1,876,126
	Illumina	469,142	575,481	703,104	861,866	1,239,284	1,502,424
	Oxford Nanopore	81,466	93,581	106,732	123,021	151,031	172,654
	Other	1,577	7,134	62,175	71,218	159,425	201,048
Source countries for raw data		64	66	69	69	75	80

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

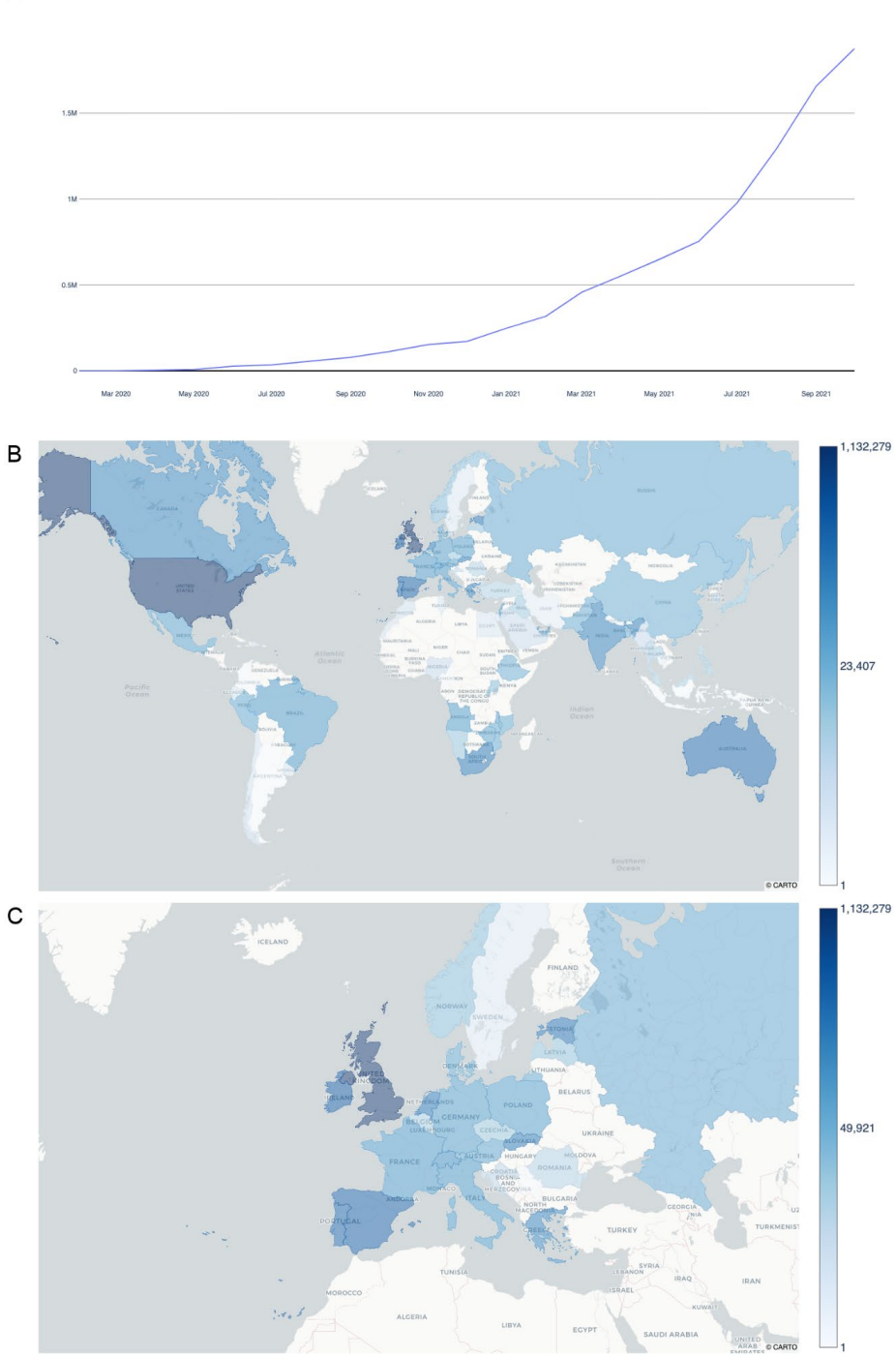


Figure I: Globally available raw SARS-CoV-2 data and distribution of sources, showing (A) sustained growth in raw data since launch of the mobilisation campaign by cumulative number of data sets, (B) and (C) geographical sources of global and European raw data, respectively, for which 66.6% of global data have been routed through the SARS-CoV-2 Data Hubs, with the remaining 33.4% 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).

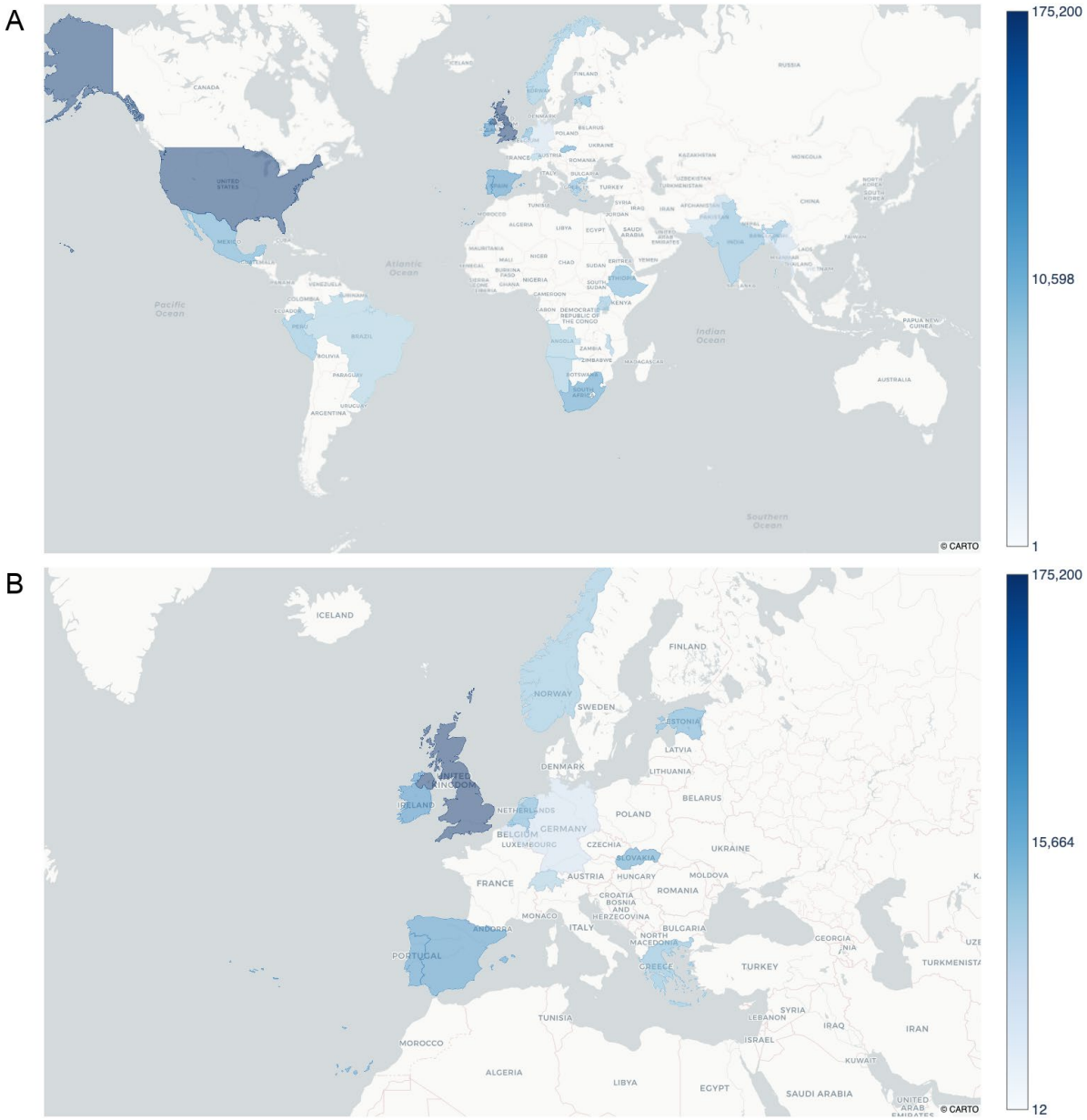


Figure II: New raw SARS-CoV-2 data and distribution of sources at global (A) and European (B) levels mobilised since 21 September 2021. Note that the colour scales are logarithmic best to show the broad range across countries.

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

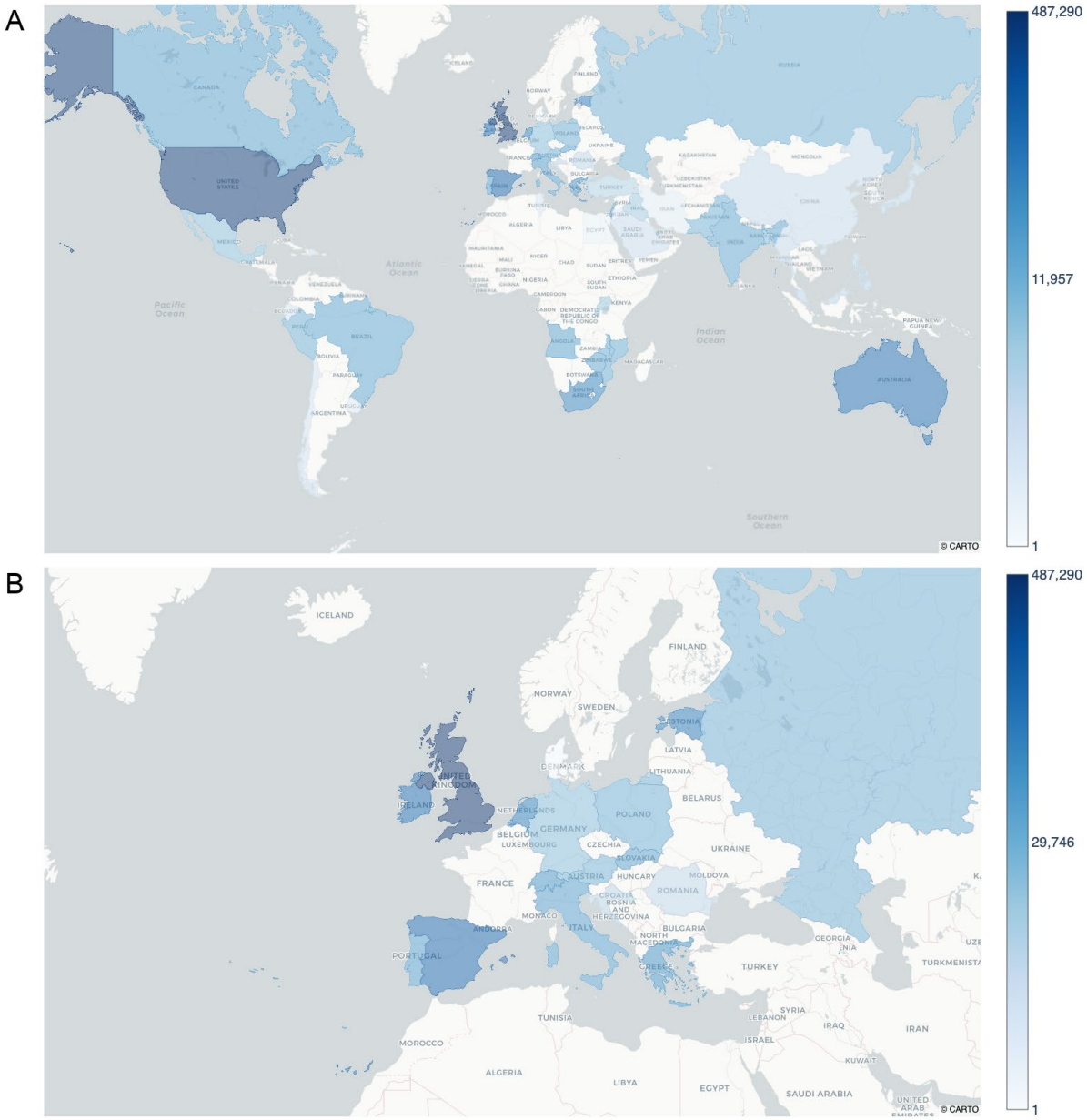


Figure III: Geographical sources of analysed raw data comprising 682,153 data sets spanning the period of data first published from 31 Jul. 2020 to 05 Oct. 2021 globally (A) and within Europe (B). Note that the colour scales are logarithmic best to show the broad range across countries.

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

	ENA	GISAID
Ireland	2595	18891
Greece	281	2162
Angola	4	50
United Kingdom	36857	554410
Estonia	82	2044
Slovakia	98	4092
South Africa	213	9450
Spain	505	23517
Pakistan	9	541
Malawi	3	191
Japan	2	128
USA	7684	673630
Portugal	77	8997
Uruguay	1	175
Netherlands	53	21103
Italy	3	2602
Peru	2	1877
Bangladesh	1	1143
Romania	1	2684
Switzerland	4	26345
Brazil	1	43776
Canada	1	55009

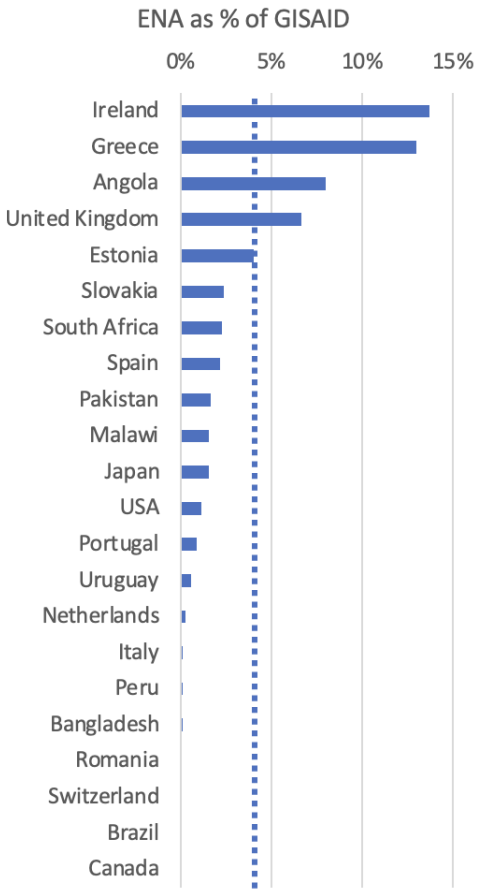


Figure IV: Entries of Delta variant sequences by country in ENA and GISAID, for countries that have deposited data in both databases, showing the rate of sharing of raw reads differs greatly.

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 27 Sept. 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

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

concern, the combination of mutations will be included in the raw read analysis in this report.

Update as of 01 November 2021

In many countries, the VOC Delta is the dominating variant, replacing earlier circulating lineages. As of 01 November 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 sub-lineage of Delta has been identified, AY.4.2, that contains additional mutations of which two in the spike-protein (Y145H, A222V). This sub-lineage showed an increasing frequency on a growing trajectory in the UK. It is not clear where the variant originated and when. Epidemiological and laboratory studies are being performed to assess phenotypic properties of AY.4.2. According to preliminary data of PHE, secondary attack rate for household contacts of cases with AY.4.2 was 12.4% (95% CI: 11.9% to 13.0%) compared with 11.1% (95% CI: 11.0% to 11.2%) for other Delta cases. Crude analysis on death and hospitalisations showed no strong evidence for a difference in risk of hospitalisation or death between AY.4.2 and other Delta cases. These findings are very preliminary, and should be interpreted with caution.

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, four 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) and Delta (B.1.617.2, AY.x). 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; e.g., AY.1 and AY.2 contain the additional mutation K417N when compared with its parent lineage. Recently, sub-lineage AY.4.2 increased in frequency in the UK and seems to be expanding rapidly. AY.4.2 contains additional mutations of which two are located in the spike protein: Y145H and A222V (N-terminal domain - NTD). According to the WHO nomenclature, all of these sub-lineages are still referred to as the same VOC.

Mainly the Delta variant is rapidly spreading globally. More information about how the new variants will affect the efficacy of vaccines in real-world conditions is becoming available and current evidence suggests that most vaccines will still provide protection against symptomatic disease and hospitalisation due to the broad immune response that is induced by vaccination. In this report, data are presented for the analysis of the presence of VOCs Alpha, Beta, Gamma and Delta, and the VOIs Lambda and Mu. A summary of the potential

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

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 and vaccination.	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

Variants of interest

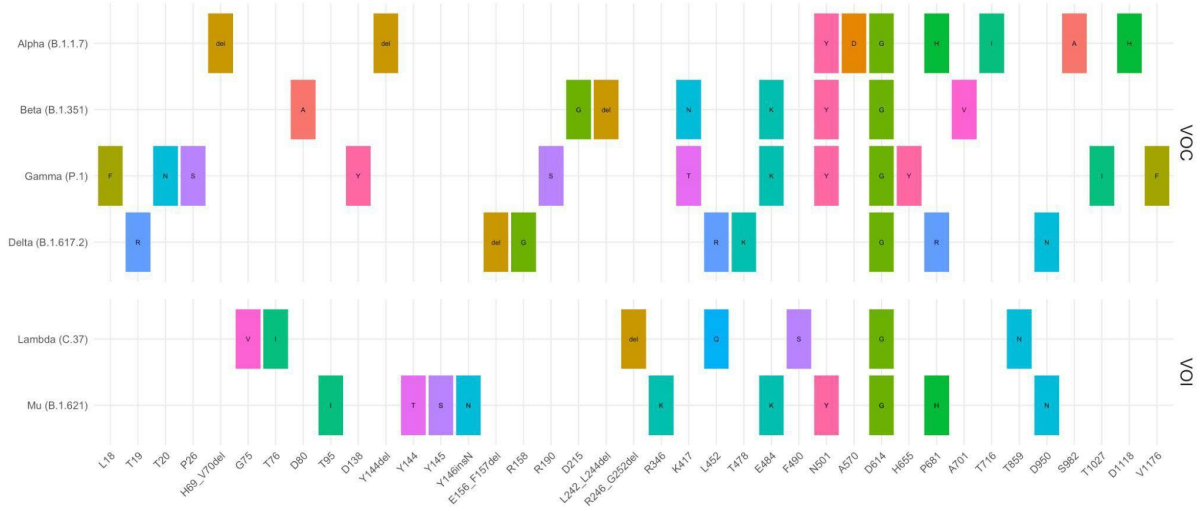
In addition to the VOCs, there have been several reports of Variants of Interest (VOIs) that contain one or more mutations of potential concern and have been found in multiple countries/cause multiple COVID-19 cases. For most of these variants, the potential impact of the combined mutational profile on transmissibility, disease severity, antigenicity, vaccine efficacy and diagnostics is not completely clear. For some VOIs there is evidence for reduction in neutralization capacity.

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

There is evidence that individual mutations may have some effect: for instance, the E484K mutation has been associated with reduced neutralization by convalescent and post-vaccine sera, the N501Y mutation with increased binding affinity to the hACE2 receptor, the L452R mutation with increased infectivity and reduced neutralization by monoclonal antibodies and convalescent/vaccine sera, and the P681H/P681R mutation with enhanced cleavage of the S-protein. An overview of the mutation profiles of the different VOCs and VOIs is given in Tables III and IV.

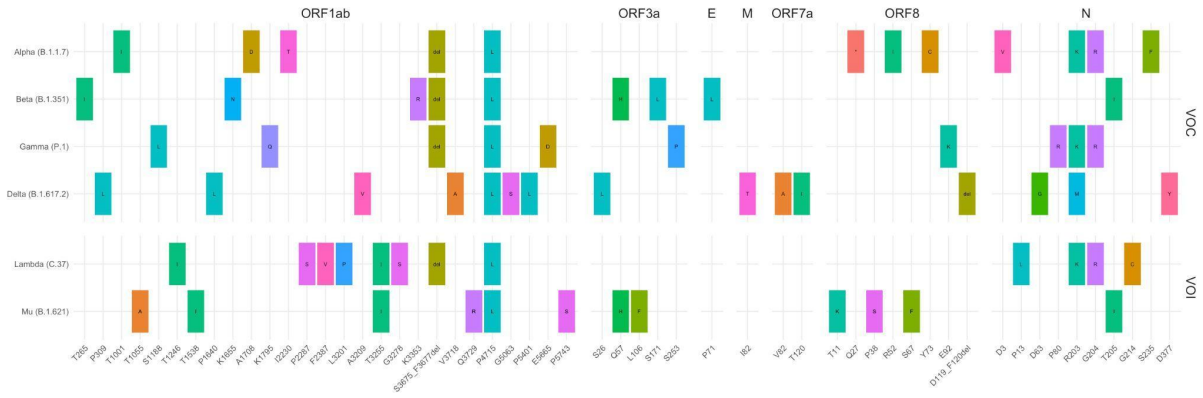
No new variants of interest have been designated since the last report. Currently, Lambda (C.37) and Mu (B.1.621) belong to the VOIs.

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)

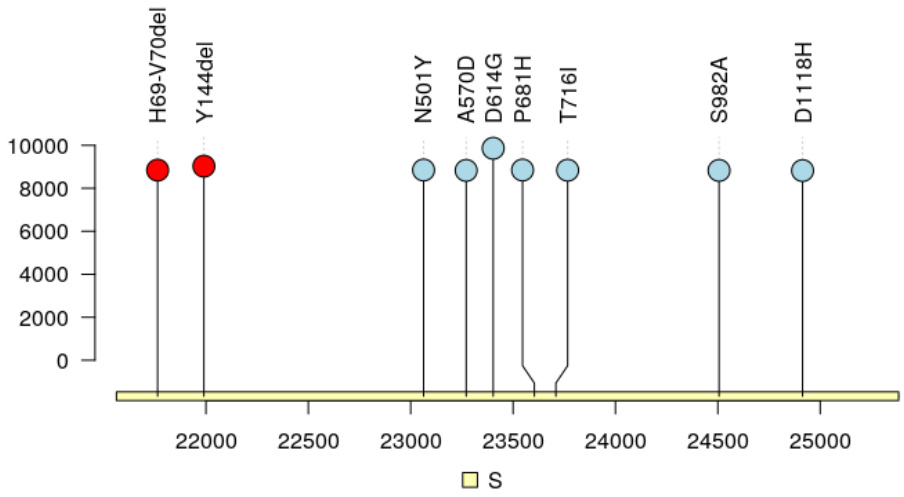


Figure V: Variant of concern Alpha as defined by the mutations in the spike protein.

For the different variants, plots are shown that present the frequency of the different mutations in the spike gene that combined define each variant (e.g. Figures V, VIII, and X). The amino acid mutations are listed on top of the figure. In addition, 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 VI, VII, IX, and XI). This is visible as the plots are shown by date of sampling. The examples show that in the recent release, Variant B 1.1.7 strains are abundantly present for the samples with the most recent release date. The other variants were found sporadically.

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

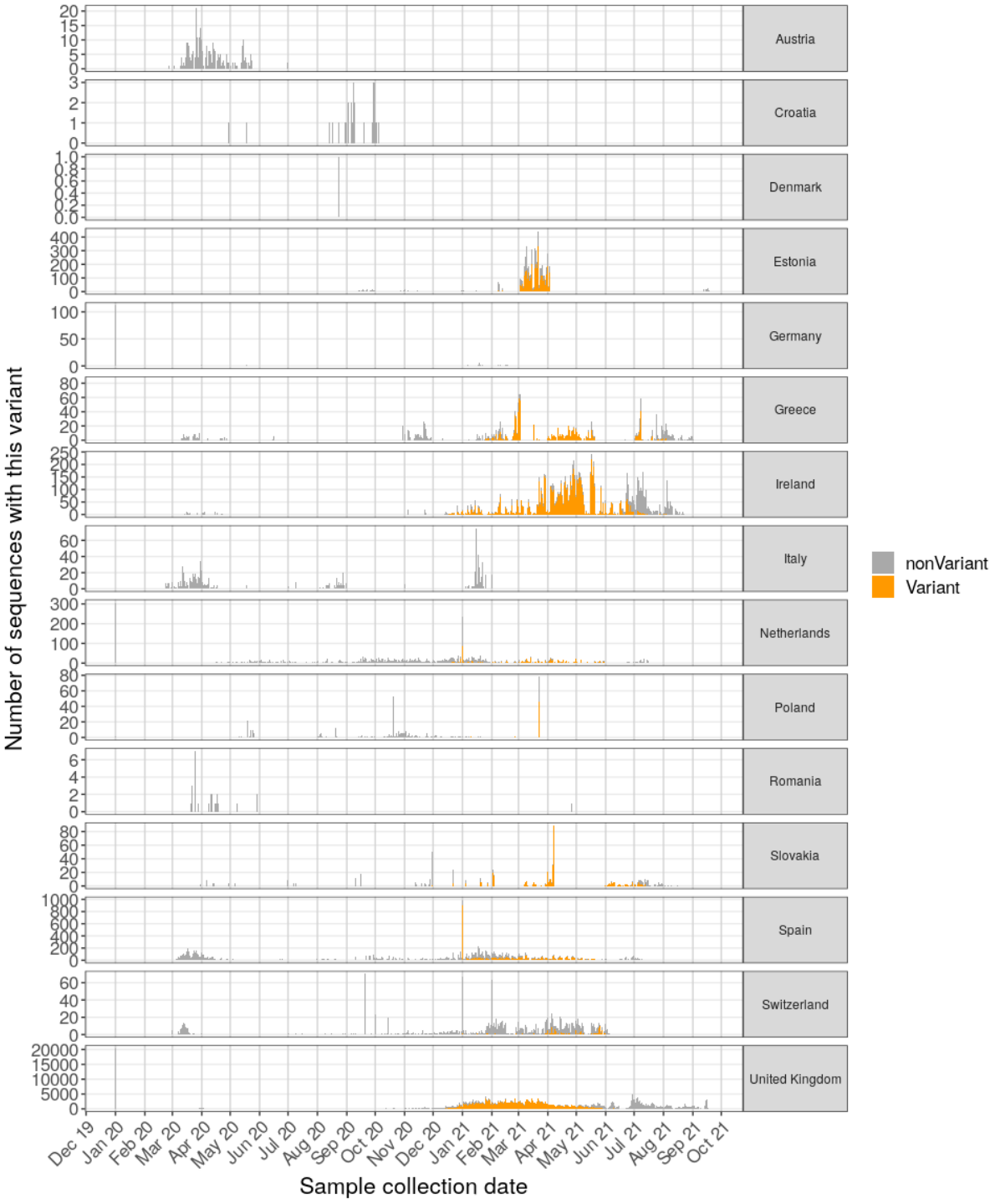


Figure VI: Number of sequences by date of sampling for Alpha variant (orange) and non-Alpha for countries in Europe.

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

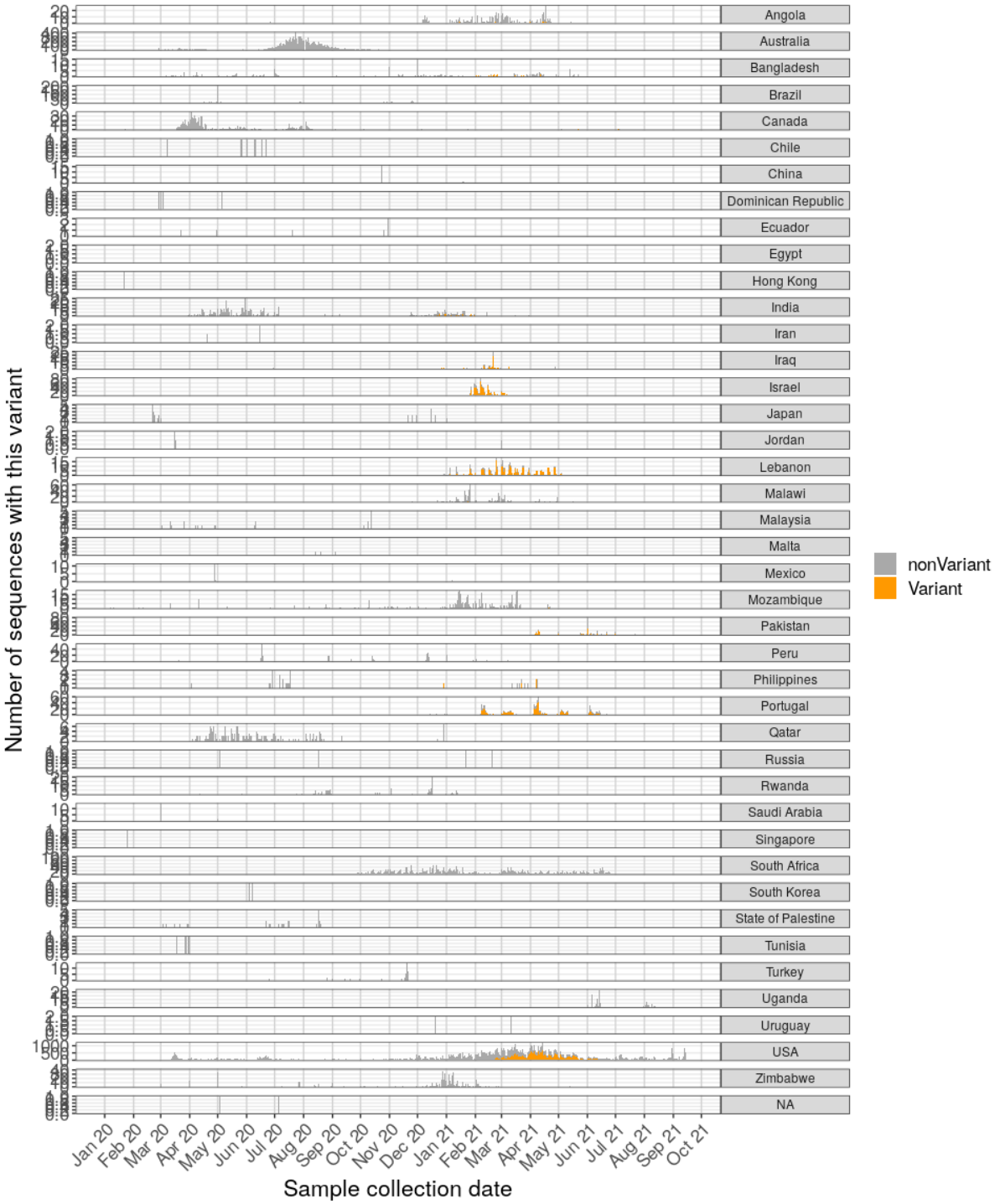


Figure VII: Number of sequences by date of sampling for Alpha variant (orange) for non-European countries.

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

Beta variant (B.1.351)

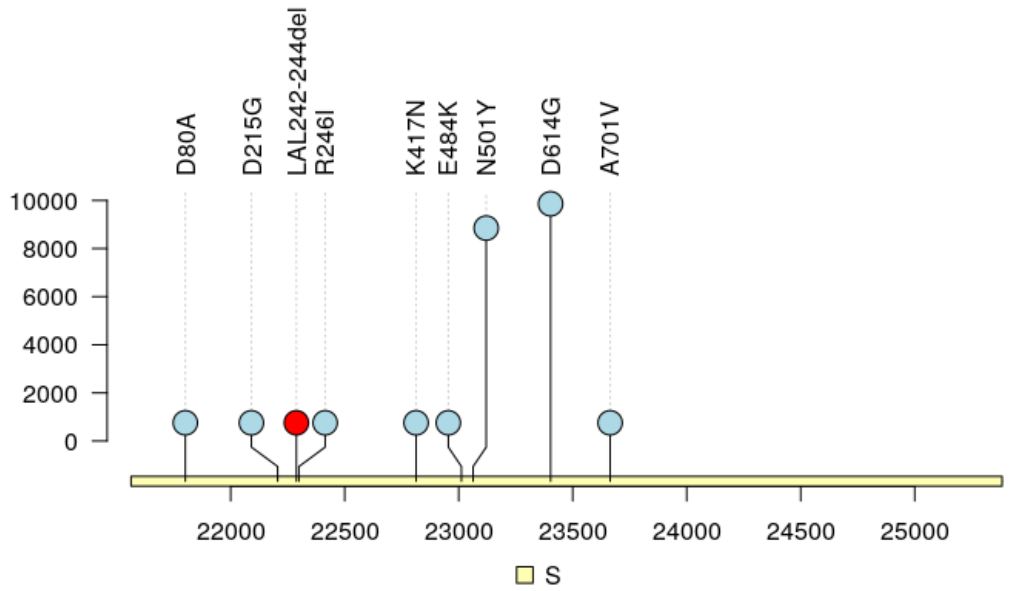


Figure VIII: Mutations in the spike protein defining variant Beta. This variant was not detected in the data uploaded since January.

The only Beta lineage samples from Europe are listed below, but these are hardly visible against the large number of background sequences.

ENA		GISAID
United Kingdom	818	1081
Netherlands	93	703
Spain	29	321
Greece	16	58
Ireland	36	81
Estonia	57	37
Portugal	8	118

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

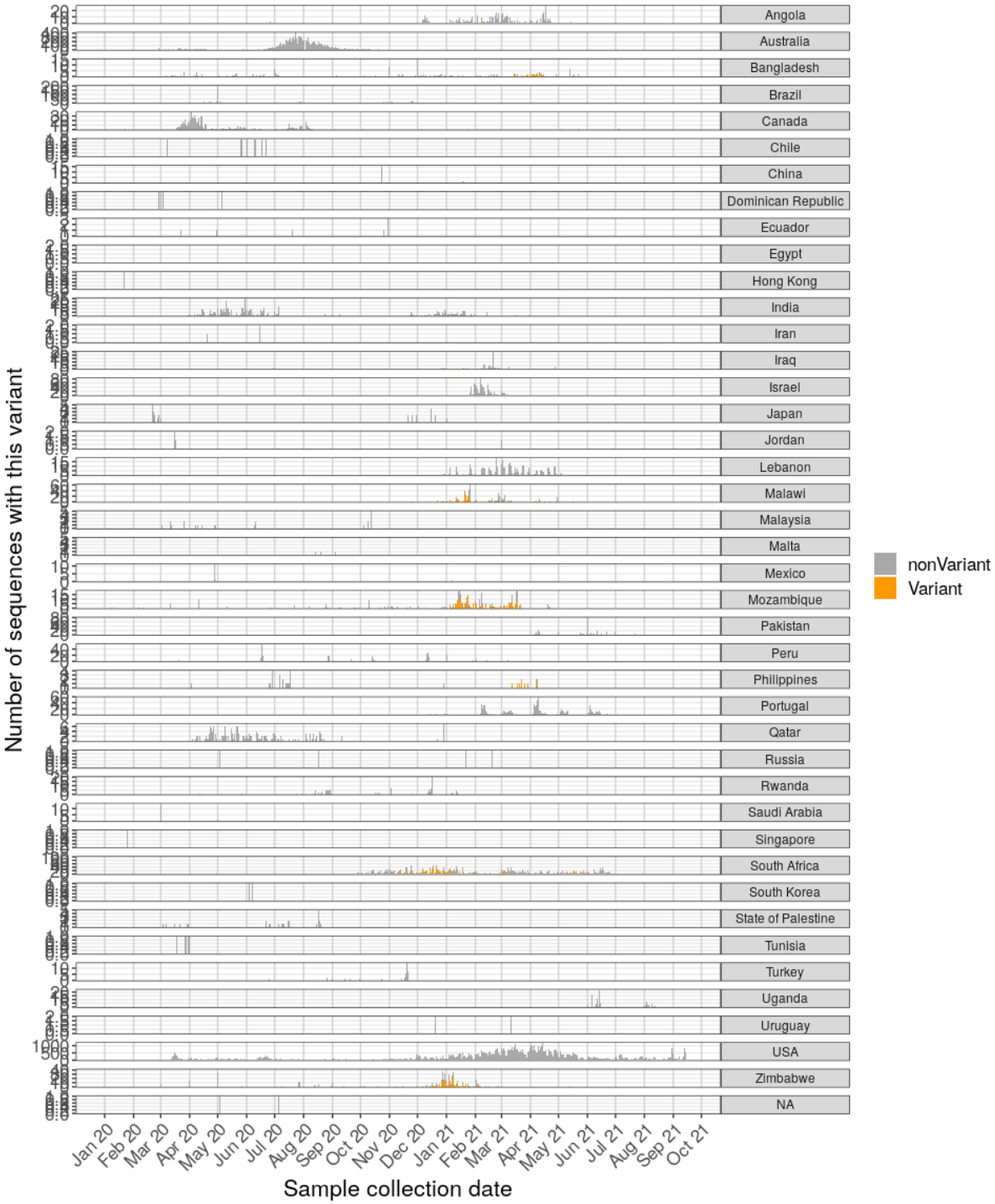


Figure IX: Number of sequences by date of sampling for variant Beta (orange) for non-European countries.

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

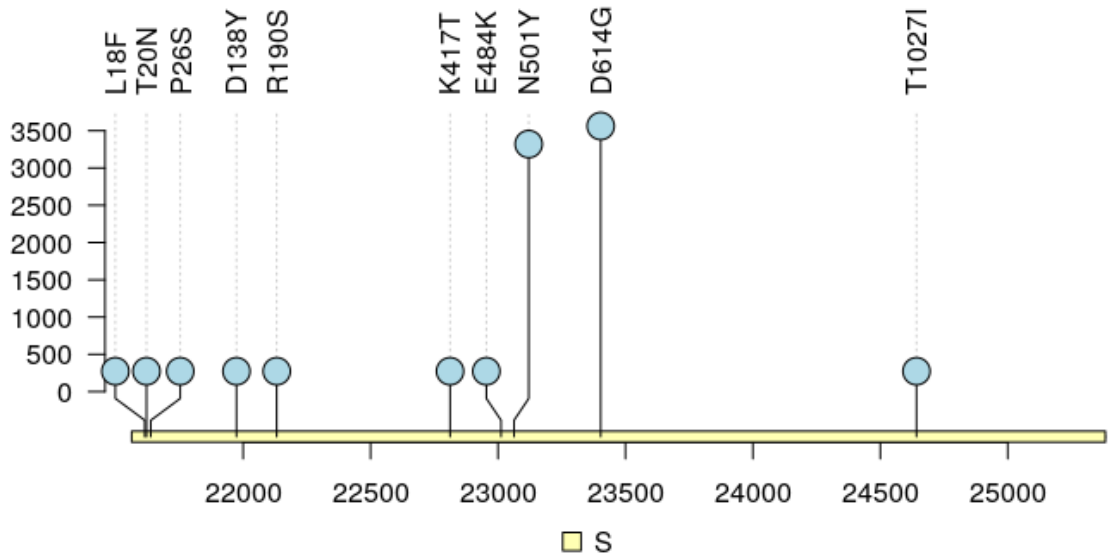


Figure X: Mutations in the spike protein defining variant Gamma.

The Gamma variant was found in countries as listed below. Against the background, these numbers are hard to see in the bar chart.

ENA		GISAID
United Kingdom	138	253
Spain	125	1228
Japan	2	128
USA	2330	28468
Italy	3	2602
Netherlands	15	591
Uruguay	1	175

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

Ireland	9	37
Bangladesh	1	1
Canada	1	15655
Peru	2	1877
Brazil	1	43776

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	36857	554410
Netherlands	53	21103
USA	7684	673630
Ireland	2595	18891
Switzerland	4	26345
Spain	505	23517
South Africa	213	9450
Canada	1	55009
Angola	4	50

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

Malawi	3	191
Slovakia	98	4092
Greece	281	2162
Pakistan	9	541
Bangladesh	1	1143
Romania	1	2684
Portugal	77	8997
Estonia	82	2044

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

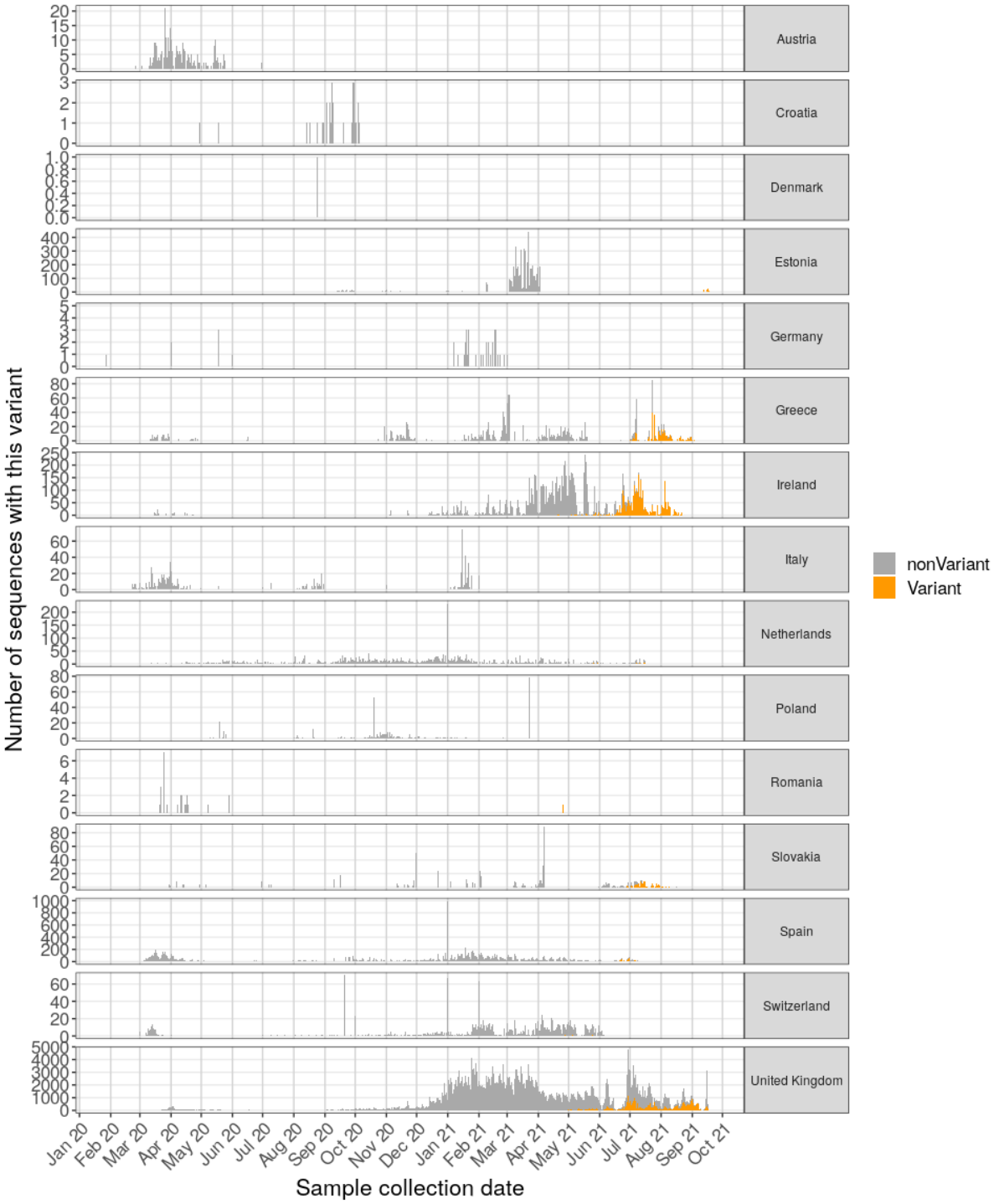


Figure XI: Number of sequences by date of sampling for variant Delta variant (orange) for European countries.

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

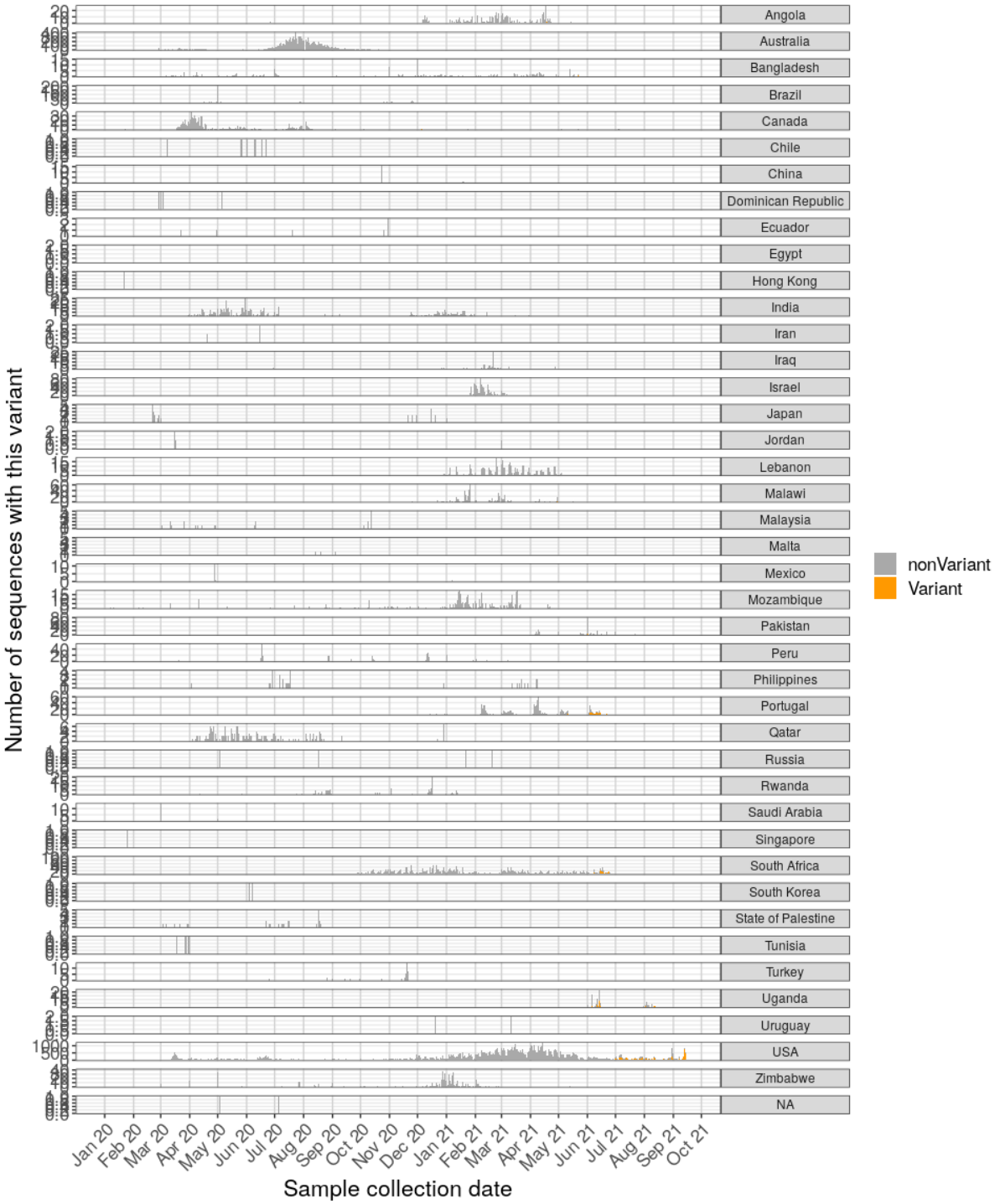


Figure XII: Number of sequences by date of sampling for variant Delta variant (orange) for countries outside of Europe.

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

Variants of Interest

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	130	1223
Ireland	4	5
Spain	9	227
Peru	22	3869

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	15	71
Netherlands	1	75
USA	448	5704
Ireland	2	7

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	67
Spain	78	685

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.

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).