

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

Report No. 8 – 08 October 2021

Summary:

- Update on mobilisation of raw reads, now totaling sequencing data sets from 1,549,740 viral raw read sets from 75 countries, a 47% 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 78,154 samples of the total 151,031 have been processed so far.
- A new variant was named by WHO, designated Mu (Pango lineage B.1621). In regions with high prevalence of Delta variant, this virus does not seem to increase in prevalence, but it was put under increased surveillance because of dissemination in some countries, particularly in South America.

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 (21 Sep 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>).

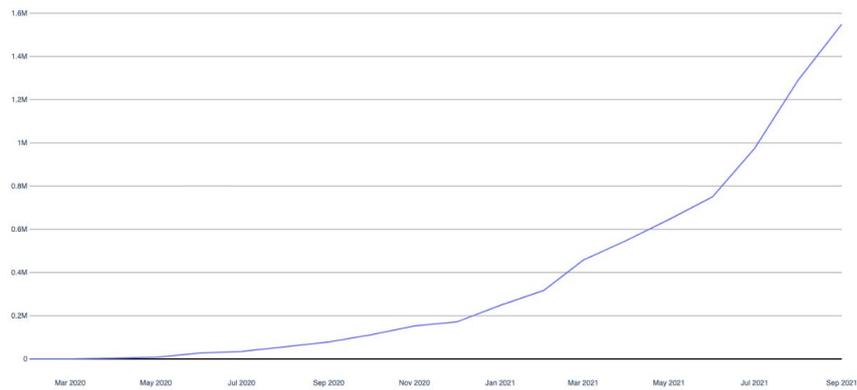


Table I: Update of number of submissions of raw read datasets to the ENA.

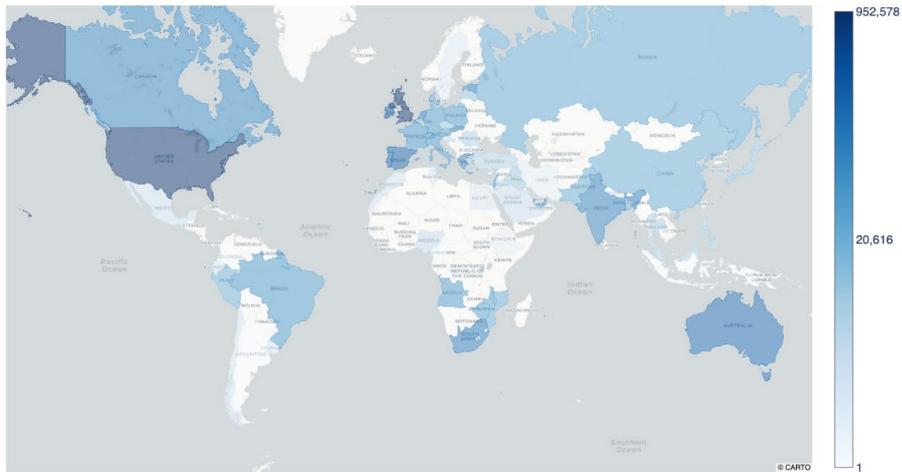
Date		19 Apr. 2021	4 May 2021	14 June 2021	10 July 2021	01 Aug 2021	21 Sept 2021
Raw data sets	Total	525,348	552,185	679,693	872,011	1,056,105	1,549,740
	Illumina	446,375	469,142	575,481	703,104	861,866	1,239,284
	Oxford Nanopore	77,913	81,466	93,581	106,732	123,021	151,031
	Other	1,060	1,577	7,134	62,175	71,218	159,425
Source countries for raw data		61	64	66	69	69	75



A



B



C

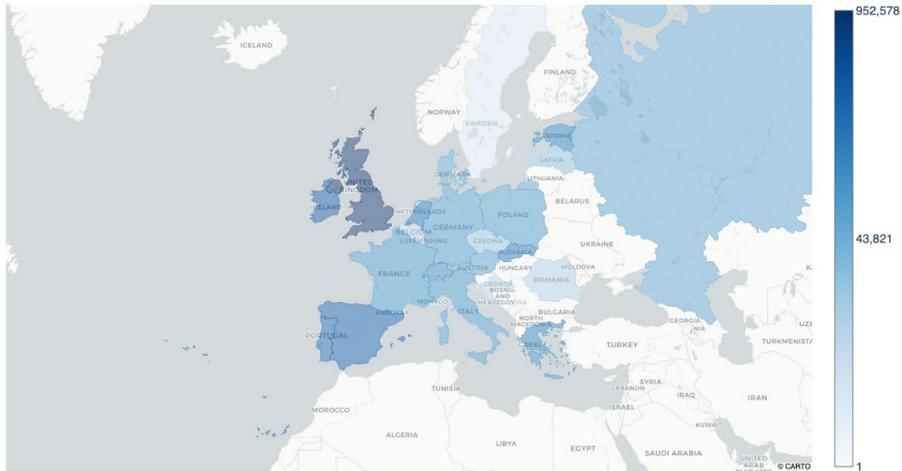


Figure 1: Globally available raw SARS-CoV-2 data and distribution of sources, showing (A) sustained growth in raw data since launch of the mobilisation campaign by cumulative number of data sets, (B) and (C) geographical sources of global and European raw data, respectively, for which 68% of global data have been routed through the SARS-CoV-2 Data Hubs, with the remaining 32% 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).

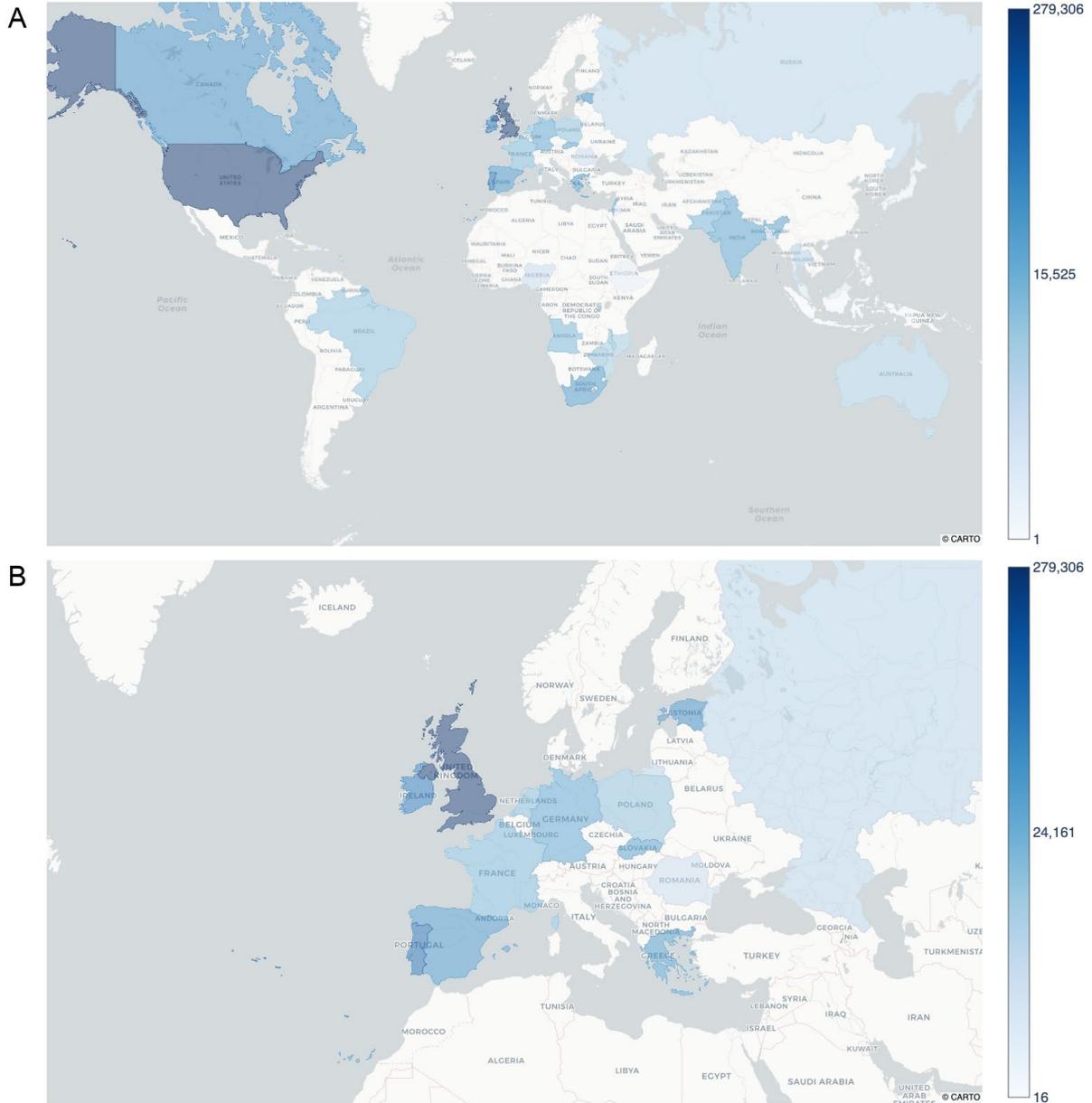


Figure II: New raw SARS-CoV-2 data and distribution of sources at global (A) and European (B) levels mobilised since 1 August 2021. Note that the colour scales are logarithmic best to show the broad range across countries.



Section II: Analysis

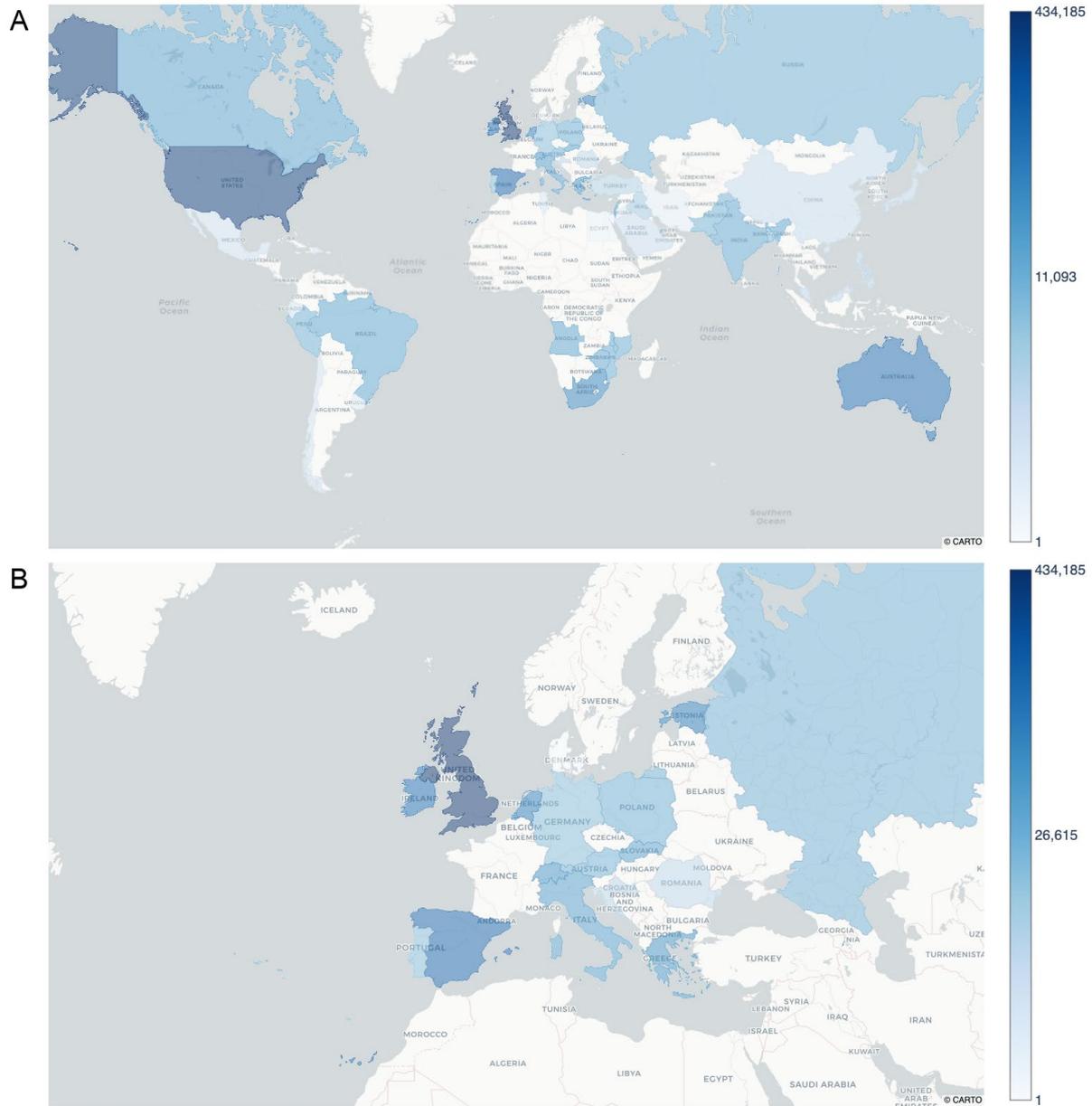


Figure III: Geographical sources of analysed raw data comprising 610,629 data sets spanning the period of data first published from 31 Jul. 2020 to 01 Sept. 2021 globally (A) and within Europe (B). Note that the colour scales are logarithmic best to show the broad range across countries.



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

Update as of 27 September 2021

In many countries, the VOC Delta is the dominating variant, replacing earlier circulating lineages. As of 27 September 2021, the Delta variant is found in at least 163 countries globally. The VOIs Lambda (C.37) and Mu (B.1.621), both originating in South America, are found in 41 and 50 countries, respectively. Some sub-lineages of the VOCs have been identified; these sub-lineages contain additional mutations that might be of biological importance.

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.1.12) and Delta (B.1.617.2, AY.1-AY.34). 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. 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. Evidence is limited on how the new variants will affect the efficacy of vaccines in real-world conditions and current evidence suggests that most vaccines will still provide protection against symptomatic disease and



hospitalisation due to the broad antibody 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 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

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.

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 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. The VOIs Eta (B.1.525), Iota (B.1.526) and Kappa (B.1.617.1) have been downgraded and now belong to the ‘Variants under Monitoring’.

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.

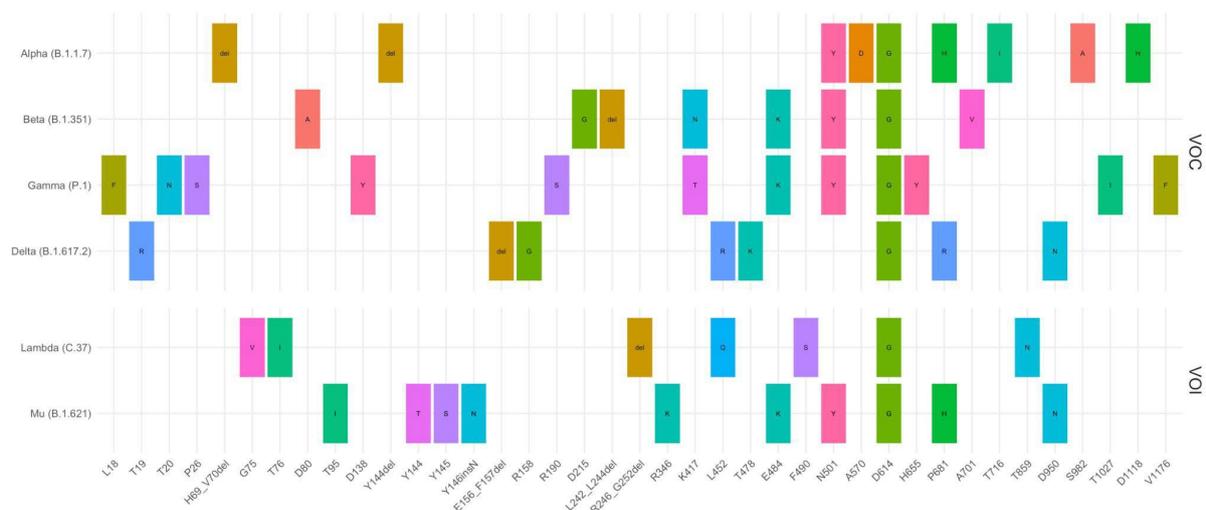
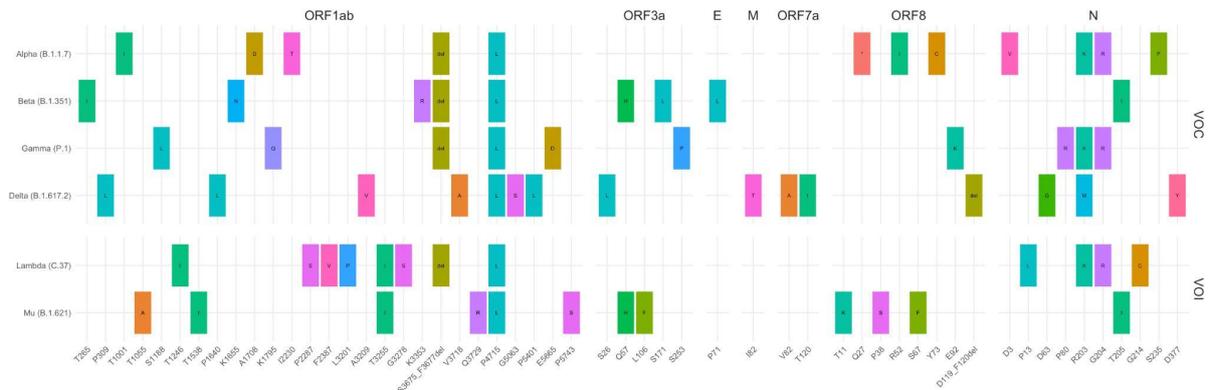


Table IV. Overview of the different mutations of several VOCs and VOIs for the ORF1ab-, ORF3a-, E-, M-, ORF6-, ORF7a- and ORF7b, ORF8 and N-gene.



Variants of Concern

Alpha variant (B.1.1.7)

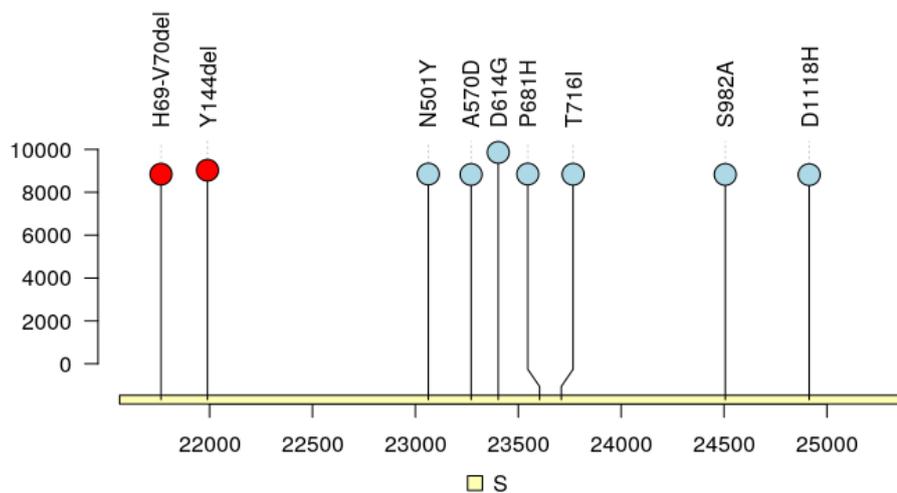


Figure IV: Variant of concern Alpha as defined by the mutations in the spike protein.

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 IV, VII, and IX). 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 V, VI, VIII, and X). 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.

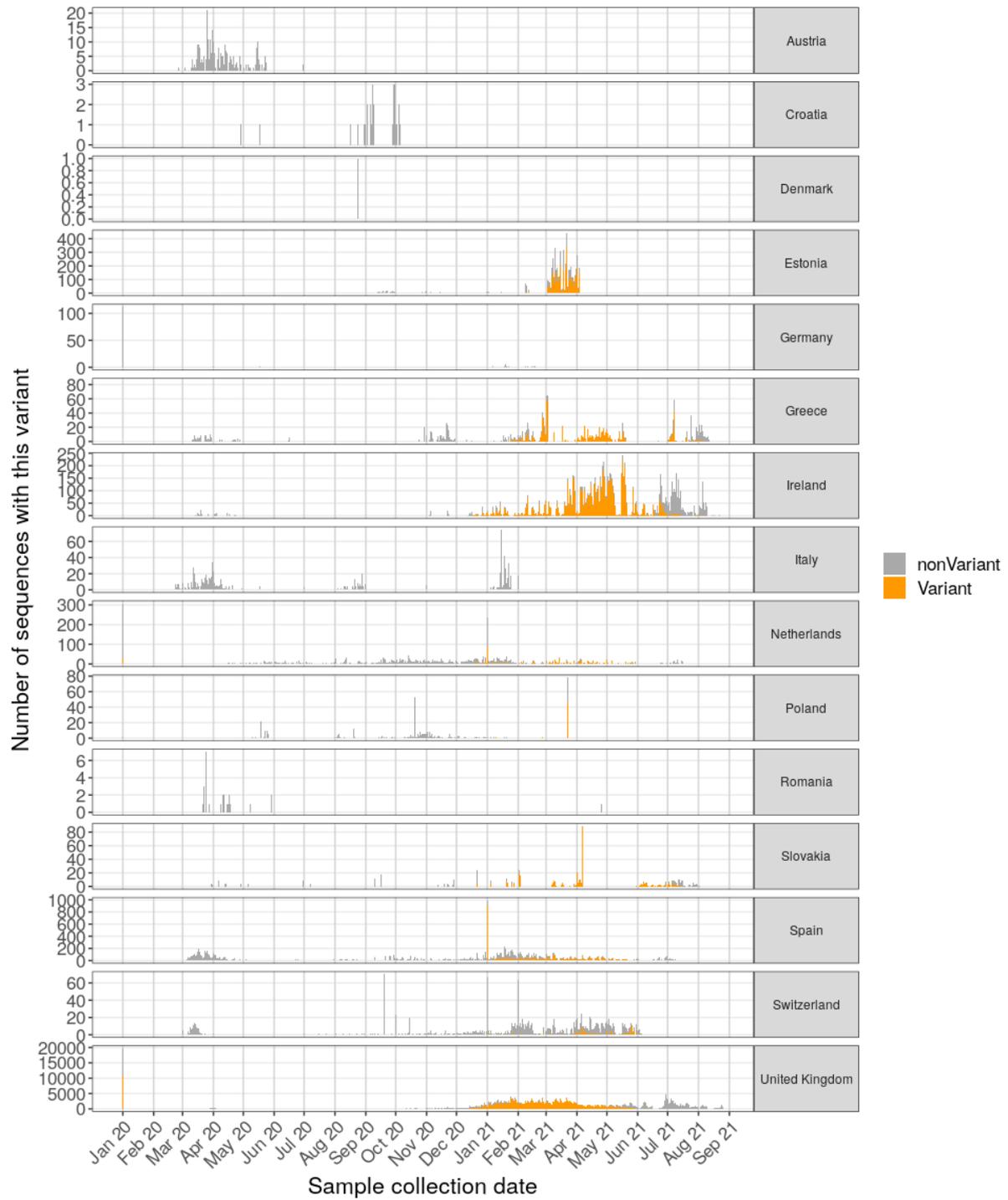


Figure V: Number of sequences by date of sampling for Alpha variant (orange) and non-Alpha for countries in Europe.



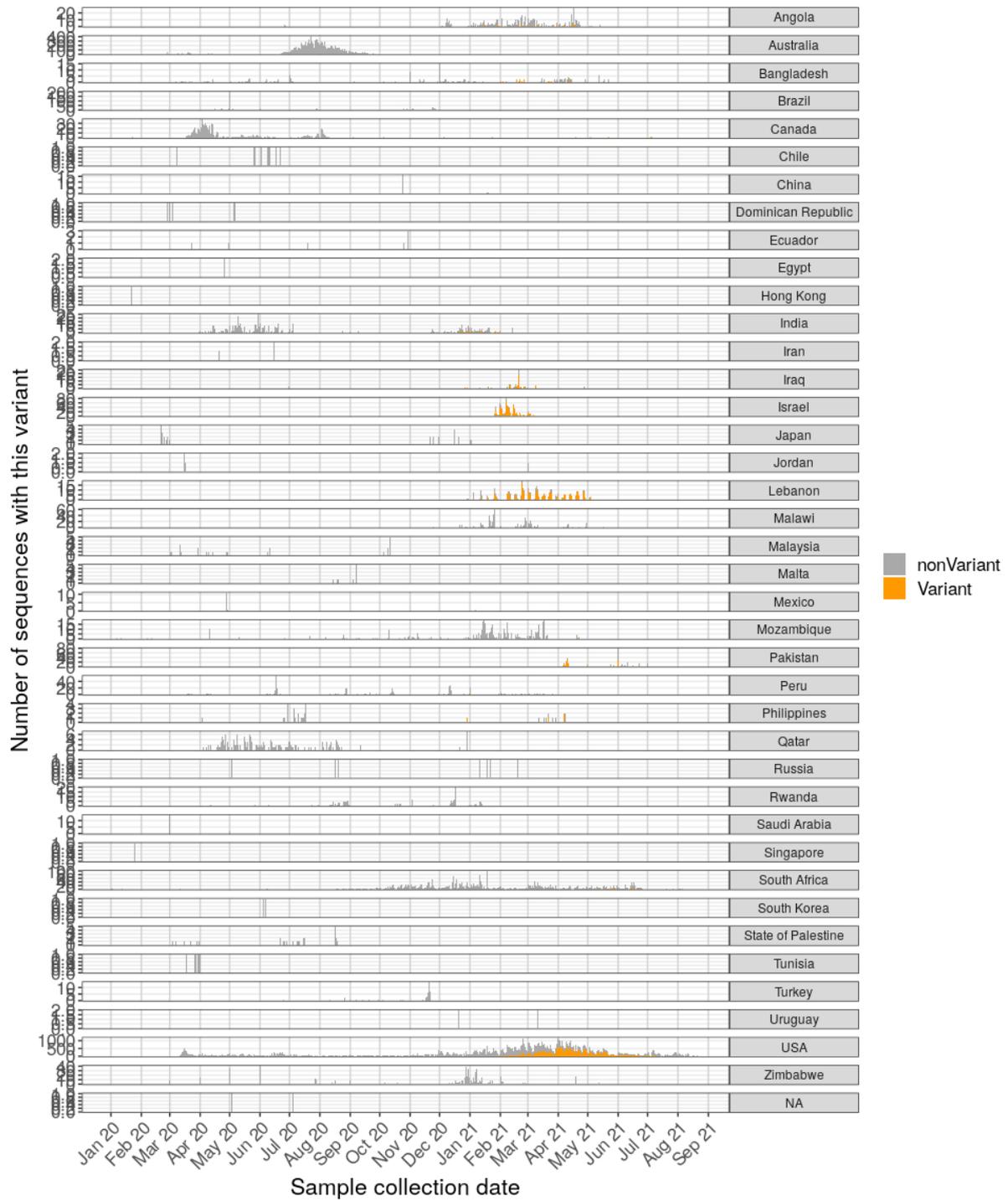


Figure VI: Number of sequences by date of sampling for Alpha variant (orange) for non-European countries.



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)

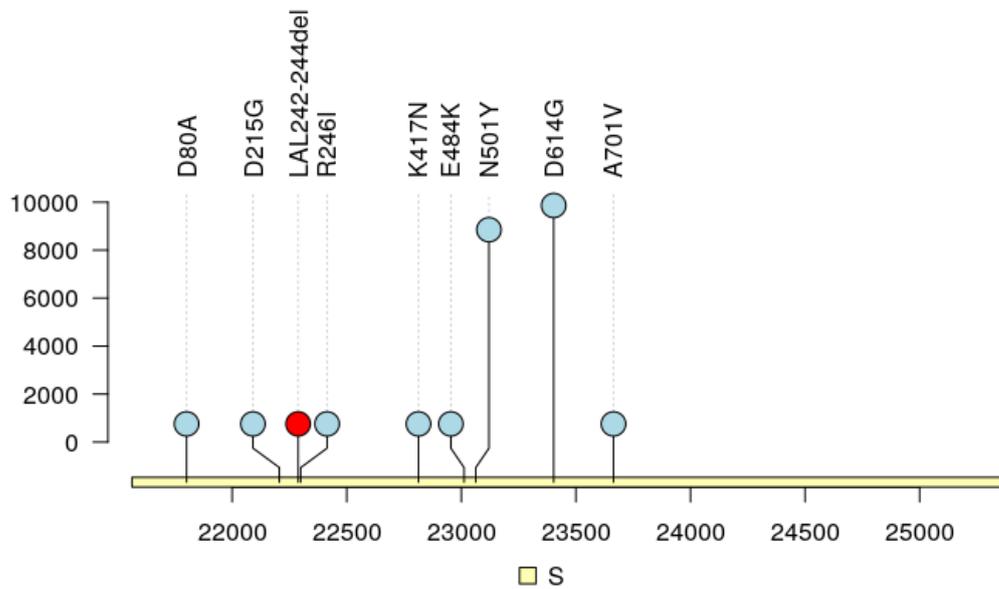


Figure VII: Mutations in the spike protein defining variant Beta. This variant was not detected in the data uploaded since January.

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	817	1081
Netherlands	93	703
Spain	29	1542
Greece	16	59
Ireland	36	81
Estonia	57	37

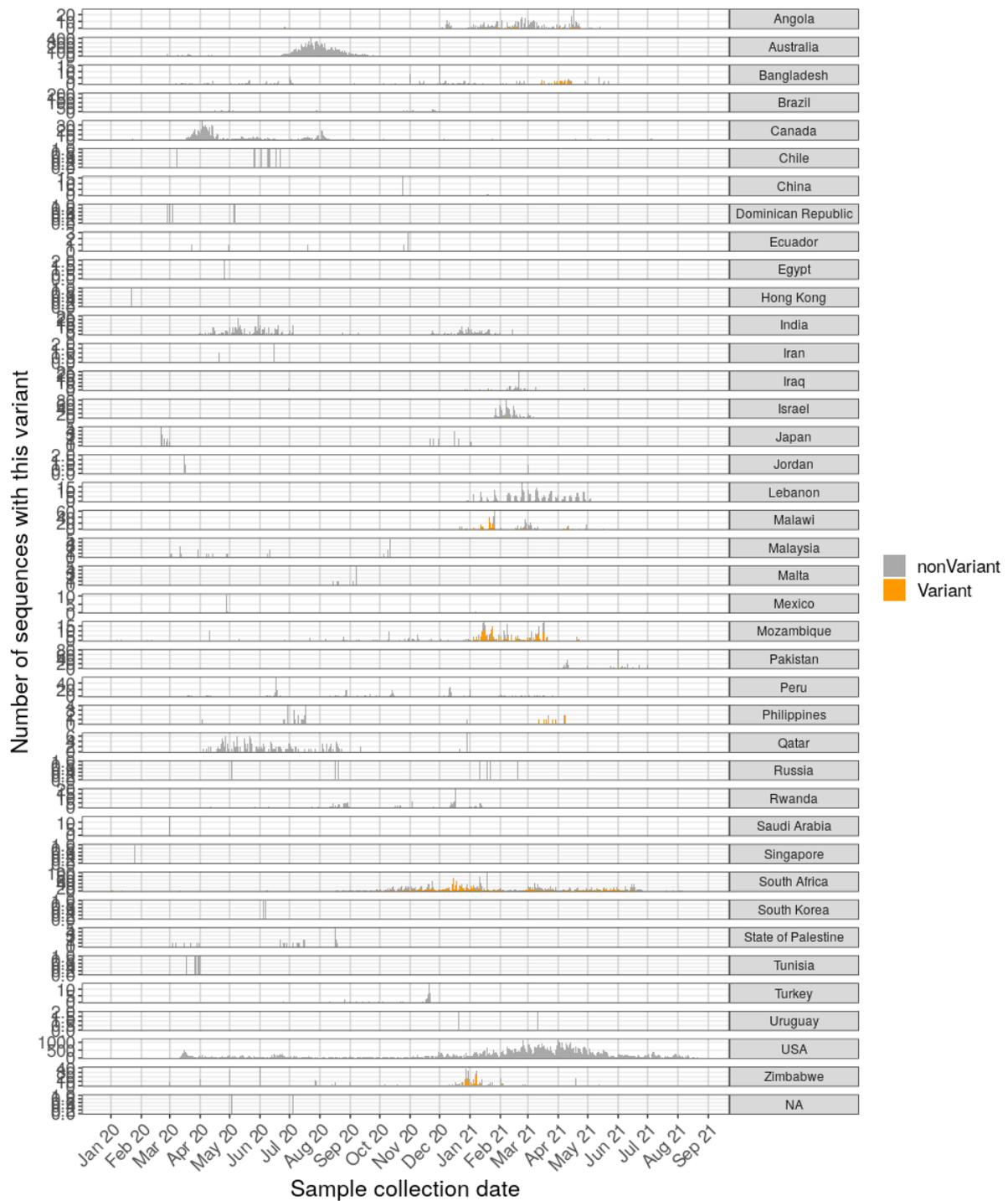


Figure VIII: Number of sequences by date of sampling for variant Beta (orange) for non-European countries.



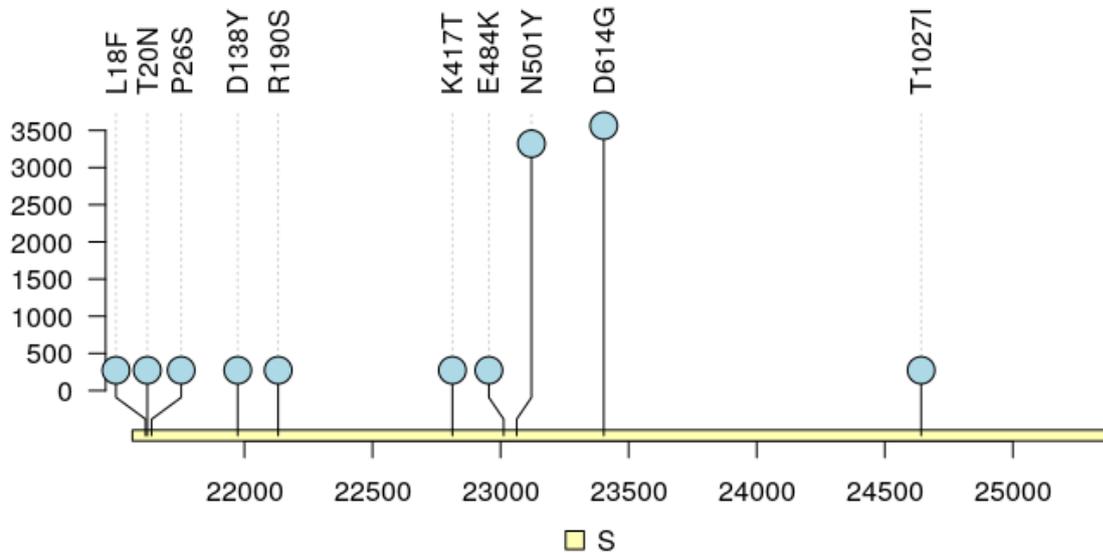


Figure IX: Mutations in the spike protein defining variant Gamma.

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	248
Spain	125	1175
Japan	2	120
USA	2278	27589
Italy	3	2582
Netherlands	15	588
Uruguay	1	174

Ireland	9	34
Bangladesh	1	1
Canada	1	14961
Peru	2	1435
Brazil	1	27841

Delta variant (B.1.617.2 + AY.x)

Samples containing all Delta variant lineage defining spike protein mutations (T19R, del156/157, R158G, 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	22640	490359
Netherlands	53	17556
USA	4465	435494
Ireland	2520	15174
Switzerland	4	20292
Spain	503	19692
South Africa	213	6866
Canada	1	33950
Angola	4	33



Malawi	3	147
Slovakia	95	2077
Greece	246	1051
Pakistan	9	309
Bangladesh	1	953
Romania	1	974

Note: The number of delta variant submissions to GISAID has decreased in this report for some countries because of an error in the previous report, resulting in all GISAID entries for some of the countries to be reported as Delta variant.



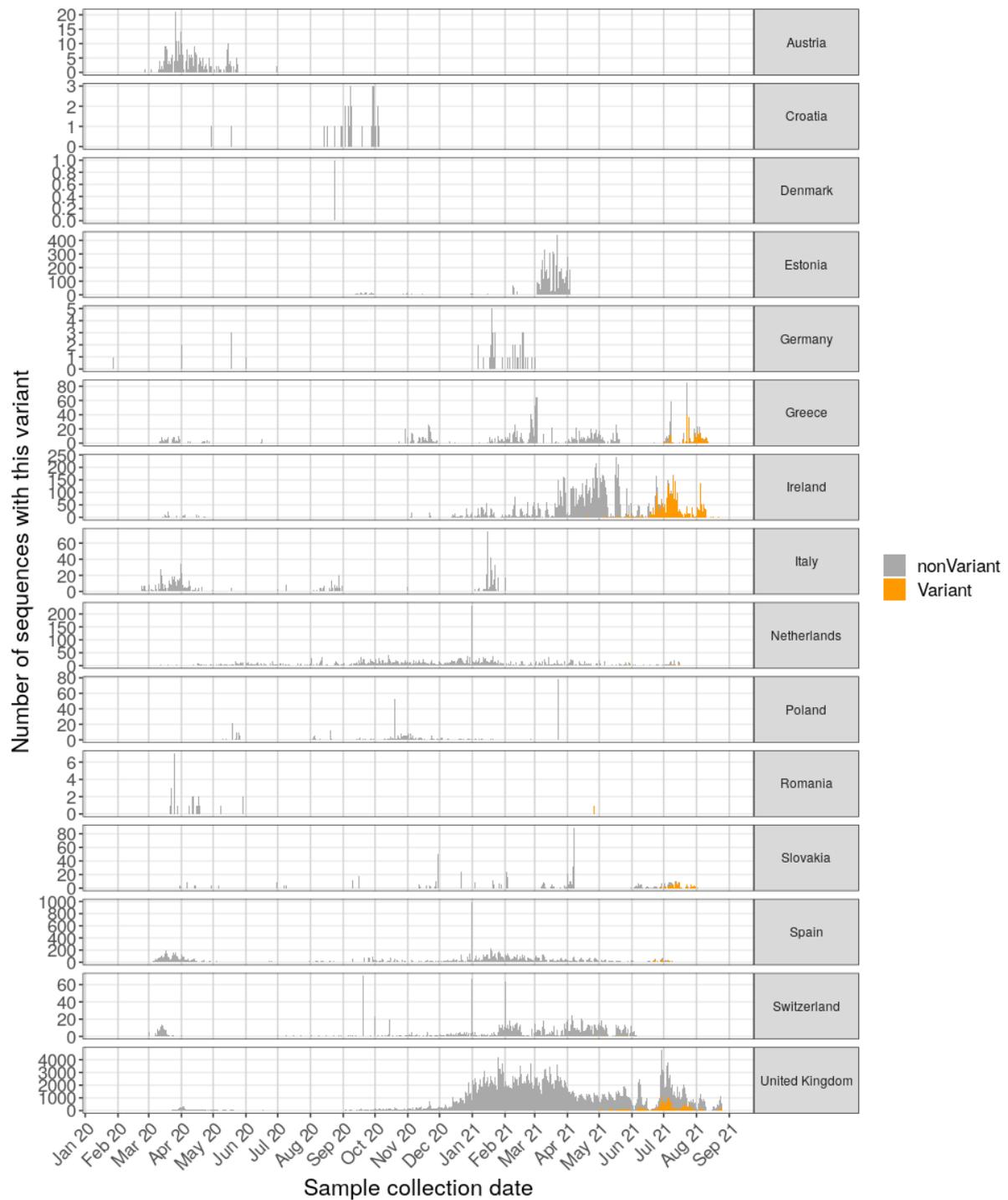


Figure X: Number of sequences by date of sampling for variant Delta variant (orange) for European countries.



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

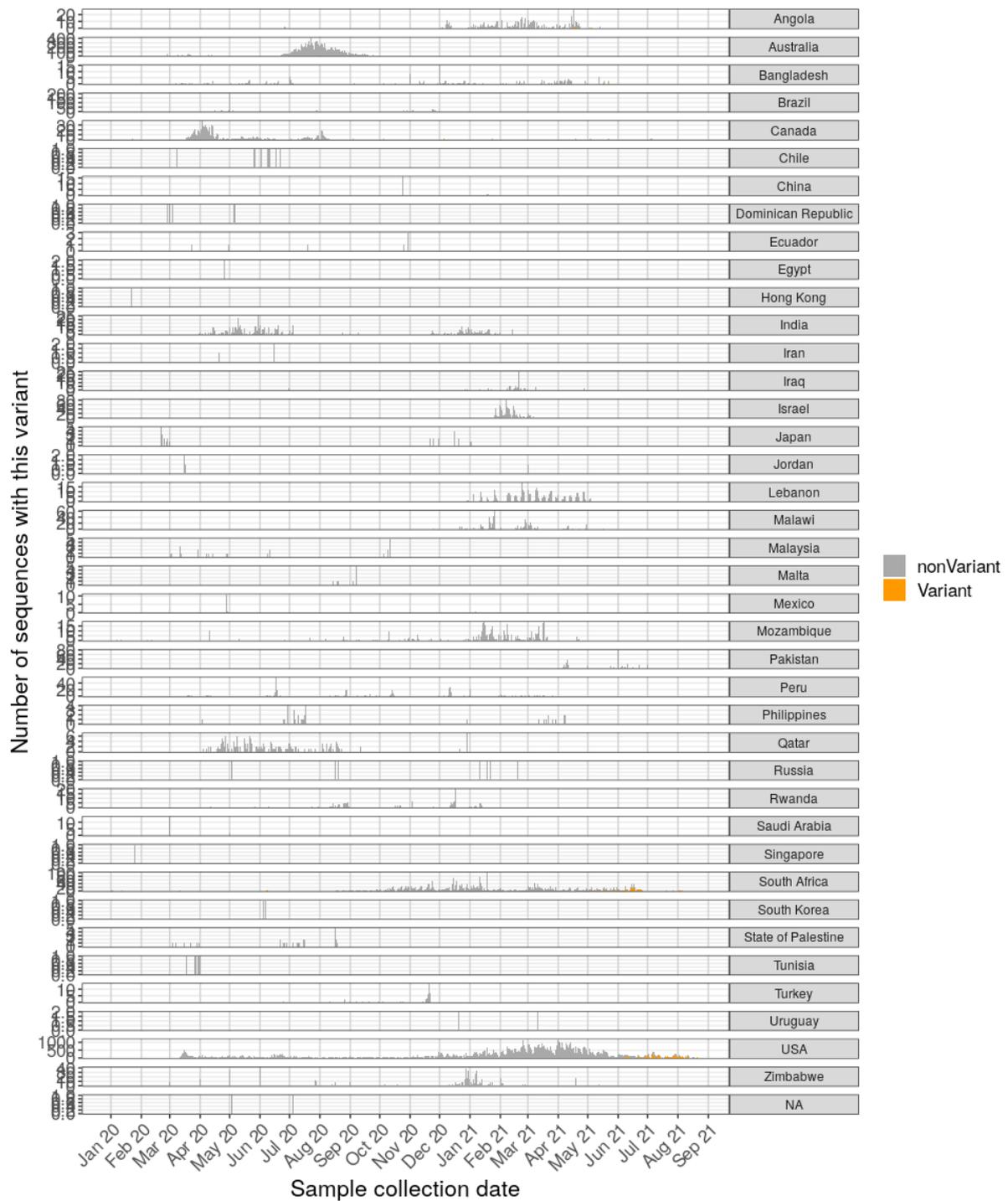


Figure XI: Number of sequences by date of sampling for variant Delta variant (orange) for countries outside of Europe.



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	126	1170
Ireland	4	5
Spain	9	220
Peru	22	3082

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	14	63
Netherlands	1	72
USA	442	4870
Ireland	2	6



Switzerland	2	48
Spain	75	648

Oxford Nanopore sequencing data

The 128,108 Nanopore read datasets in the ENA public database are being processed.

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 is still 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.

VEO will continue to analyse all publicly shared Illumina data for presence of variants. In addition, an Oxford Nanopore VCF calling workflow has been implemented and has started to process the backlog of data. In combination with more data hopefully being shared by member states and some targeted sampling, this will improve our understanding of the pandemic and our ability to identify the emergence of major and minority variants of concern for epidemiology and immunology in a timely way.

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

