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

Report No. 7 – 07 September 2021

Summary:

- Update on mobilisation of raw reads, now totaling sequencing data sets from 1,056,105 viral raw read sets from 69 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 71,797 samples of the total 123,021 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 (1 Aug 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).

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

Date		25 Mar. 2021	19 Apr. 2021	4 May 2021	14 June 2021	10 July 2021	01 Aug 2021
Raw data sets	Total	438,112	525,348	552,185	679,693	872,011	1,056,105
	Illumina	367,462	446,375	469,142	575,481	703,104	861,866
	Oxford Nanopore	69,921	77,913	81,466	93,581	106,732	123,021
	Other	729	1,060	1,577	7,134	62,175	71,218
Source countries for raw data		58	61	64	66	69	69

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

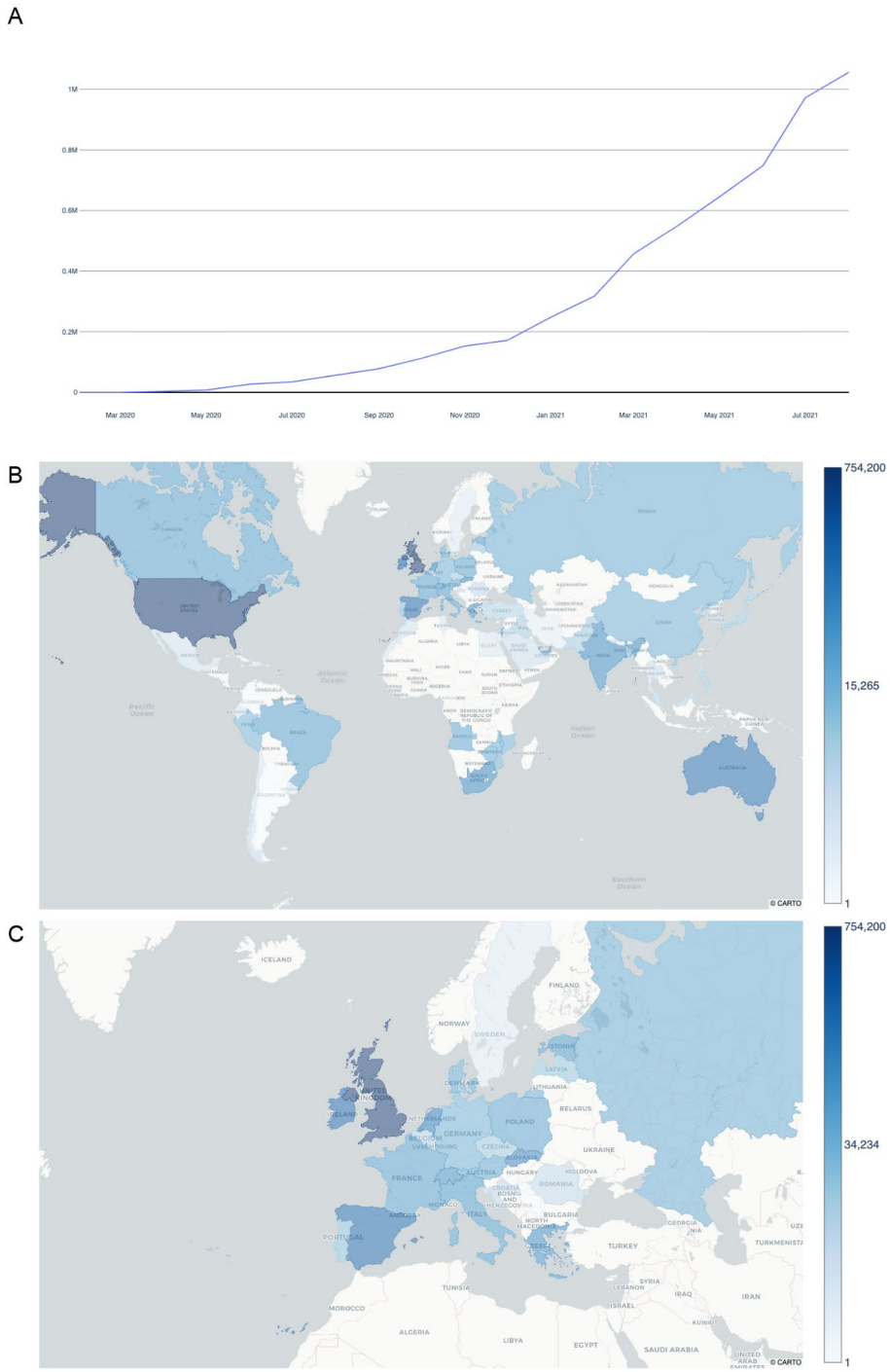


Figure I: Globally available raw SARS-CoV-2 data and distribution of sources, showing (A) sustained growth in raw data since launch of the mobilisation campaign by cumulative number of data sets, (B) and (C) geographical sources of global and European raw data, respectively, for which 78% of global data have been routed through the SARS-CoV-2 Data Hubs, with the remaining 22% 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).

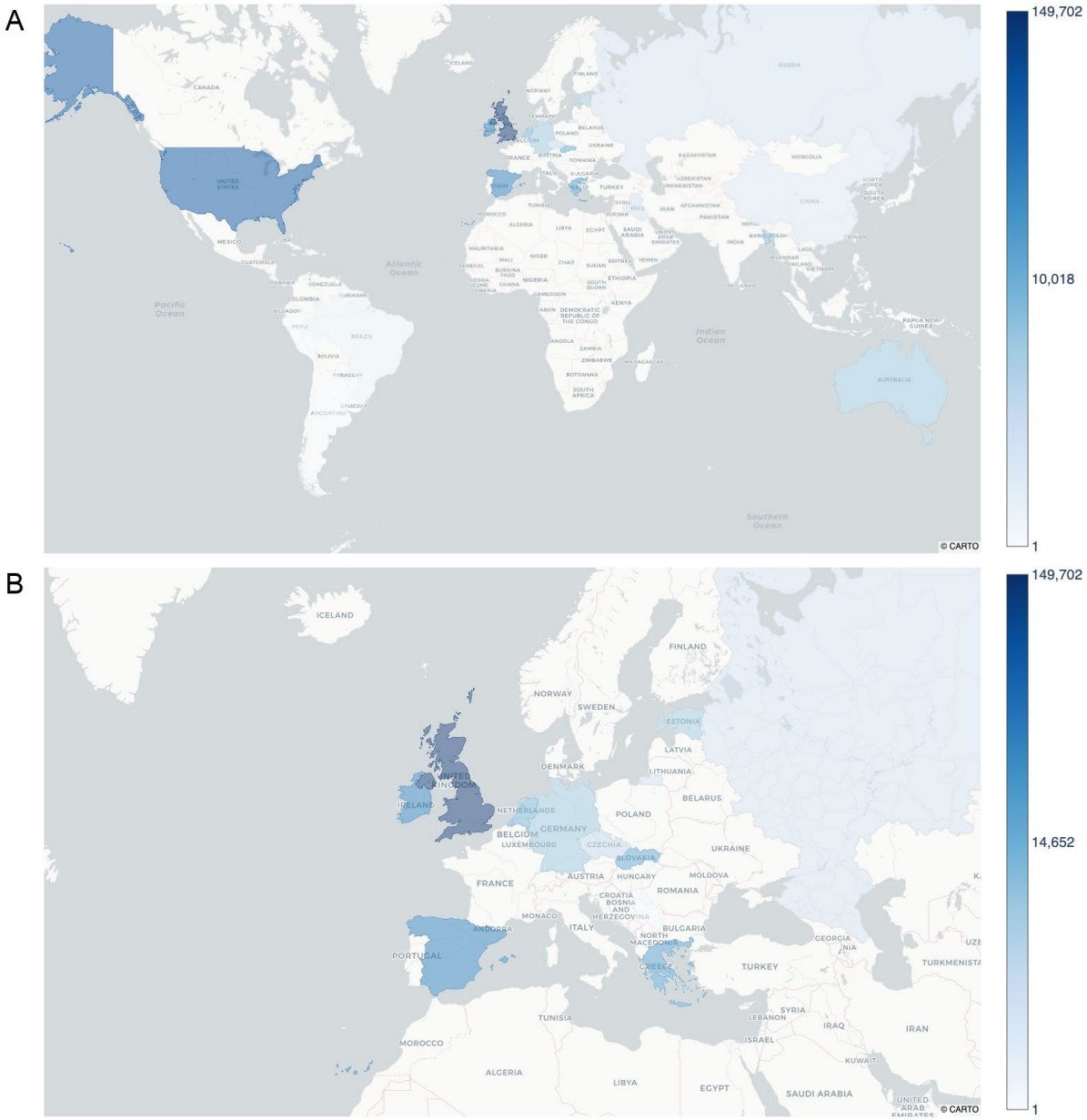


Figure II: New raw SARS-CoV-2 data and distribution of sources at global (A) and European (B) levels mobilised since 10 July 2021. Note that the colour scales are logarithmic best to show the broad range across countries.

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

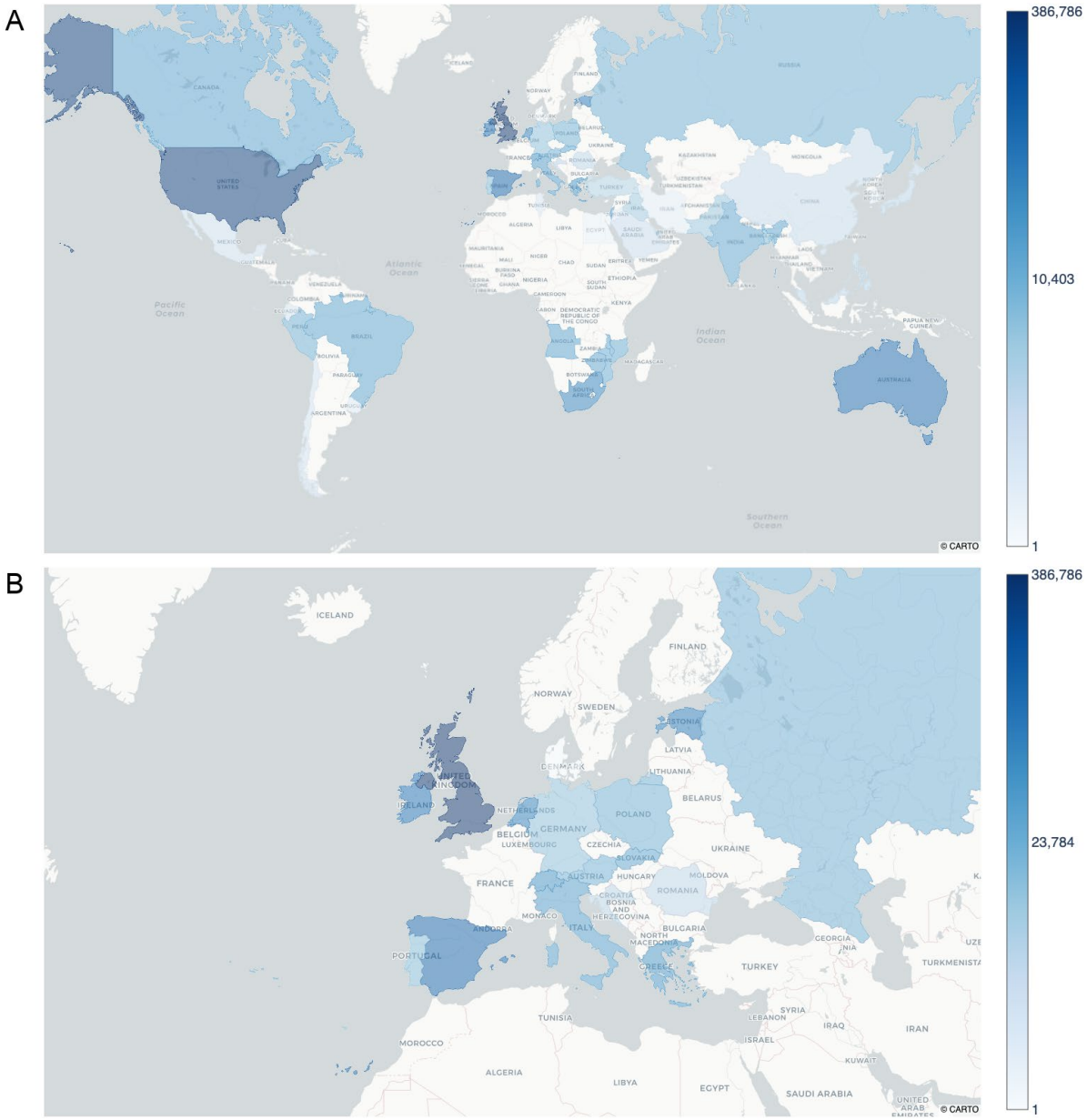


Figure III: Geographical sources of analysed raw data comprising 551,828 data sets spanning the period of data first published from 31 Jul. 2020 to 01 Aug. 2021 globally (A) and within Europe (B). Note that the colour scales are logarithmic best to show the broad range across countries.

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 15 Jul. 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 3 September 2021

The latest VOC Delta is increasing in prevalence in multiple countries and in many countries became dominant, replacing earlier circulating lineages. As of 23 August, the Delta variant is found in at least 130 countries globally. The VOI Lambda (C.37), that mainly circulated in the South American countries (Argentina, Chile and Peru), is found in 33 different countries globally. 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.4), Beta (B.1.351, B.1.351.1-B.1.351.4), Gamma (P.1, P.1.1-P.1.10.2) and Delta (B.1.617.2, AY.1-24). 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.

All VOCs rapidly spread 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

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

the broad antibody response that is induced by vaccination. In this report, data are presented for the analysis of the presence of variants Alpha, Beta, Gamma and Delta. A summary of the potential phenotypic impact based on current available literature 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).

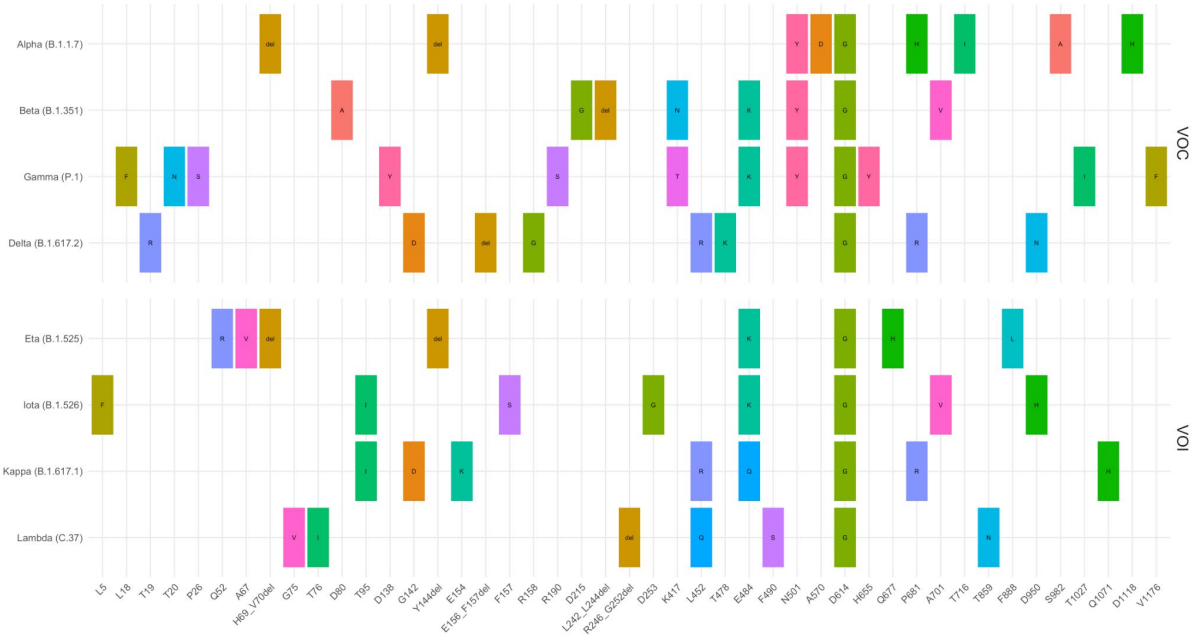
Variants of interest

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, and the L452R mutation with increased infectivity and reduced neutralization by monoclonal antibodies and convalescent sera. An overview of the mutation profiles of the different VOCs and VOIs is given in Tables III and IV.

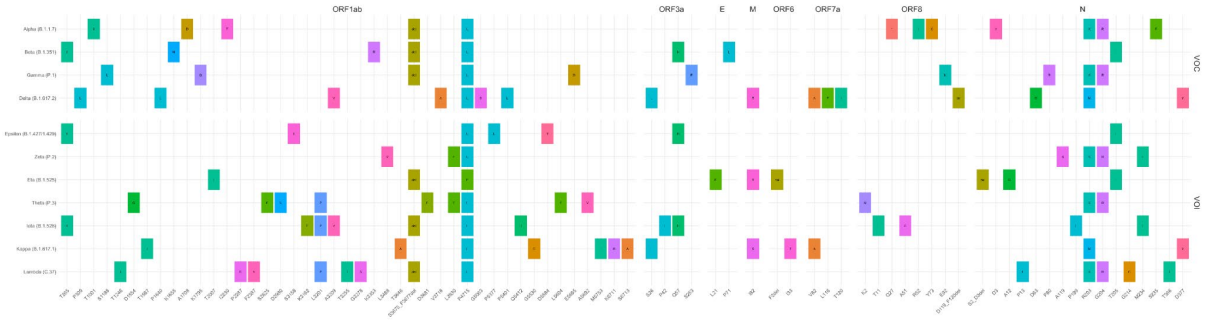
Recently (August 30th), viruses belonging to Pango lineage B.1.621 were classified as a VOI and given the WHO label "Mu". The Mu variant is characterised by a number of amino acid substitutions in the spike protein that indicate potential properties of immune escape (E484K, N501Y, P681H). It was first identified in Colombia in January 2021, and since has been observed in an increasing number of countries, including Europe. However, current evidence suggests that in regions with high levels of circulation of Delta variants, the Mu variant has not been dominating. The figures below will be updated to include the Mu variant in the next report.

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. Area in yellow is the N-terminal domain, in red is the receptor binding domain, and in blue the furin cleavage site.



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.



Alpha variant (B.1.1.7; previously known as the UK variant)

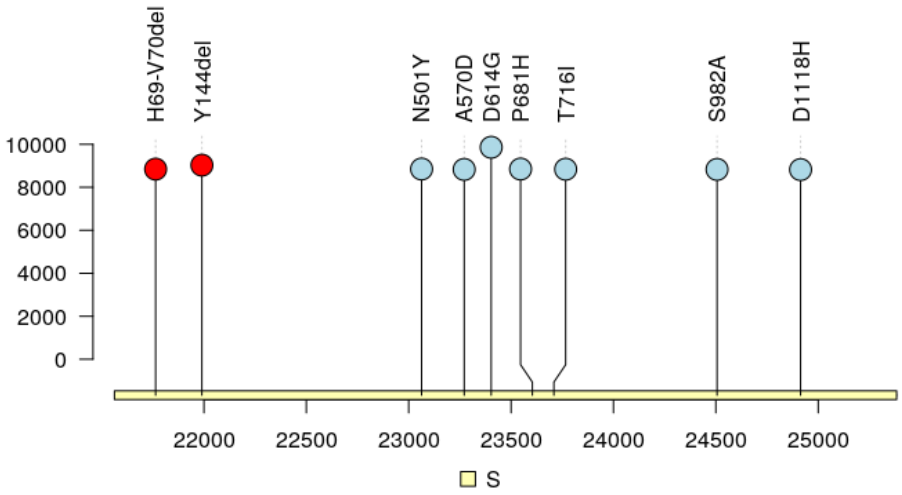


Figure IV: Variant of concern Alpha as defined by the mutations in the spike protein.

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 (B.1.1.7 plus E484K, B.1.1. 28.1).

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

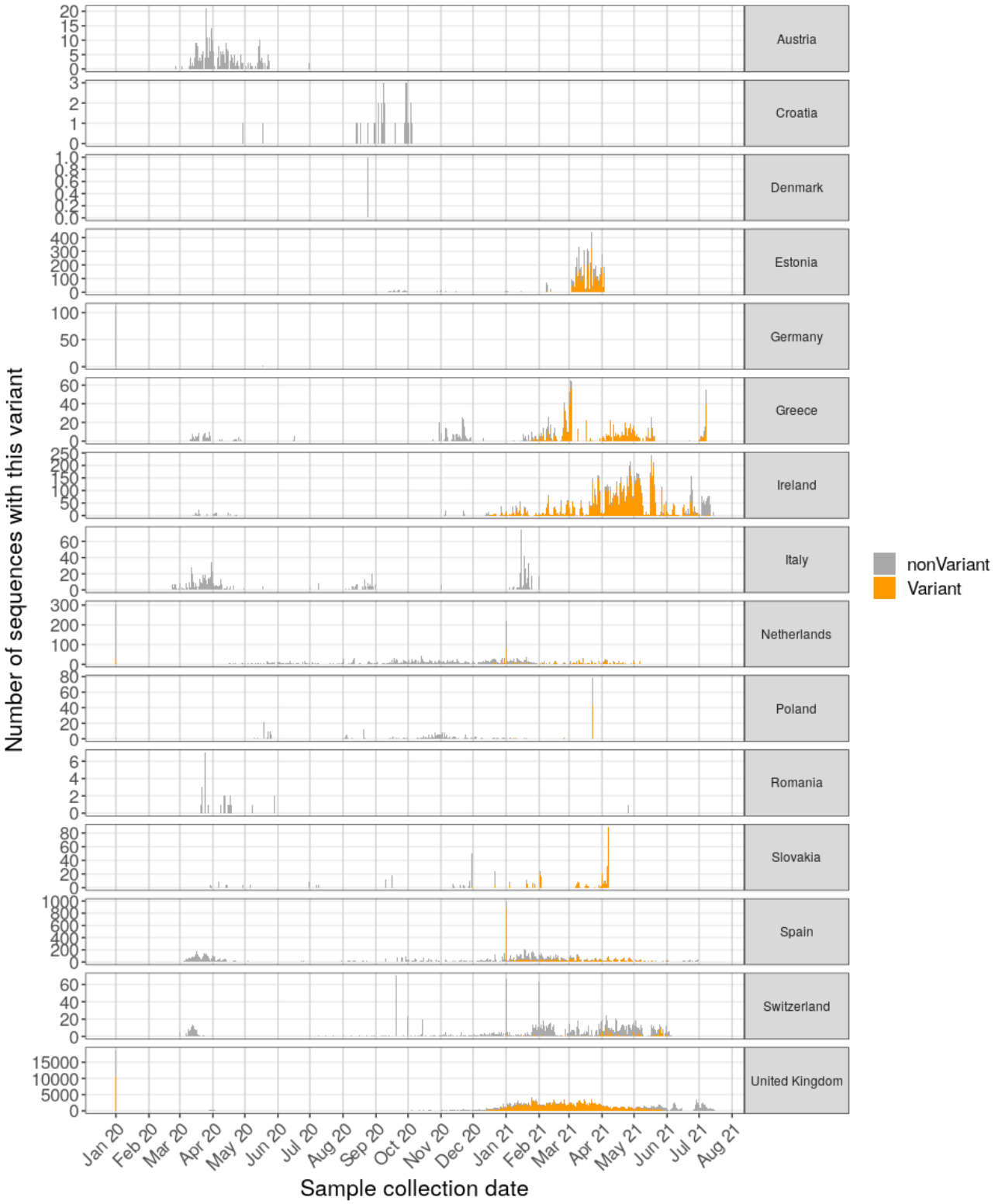


Figure V: Number of sequences by date of sampling for Alpha variant (orange) and non Alpha for countries in Europe and Turkey.

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

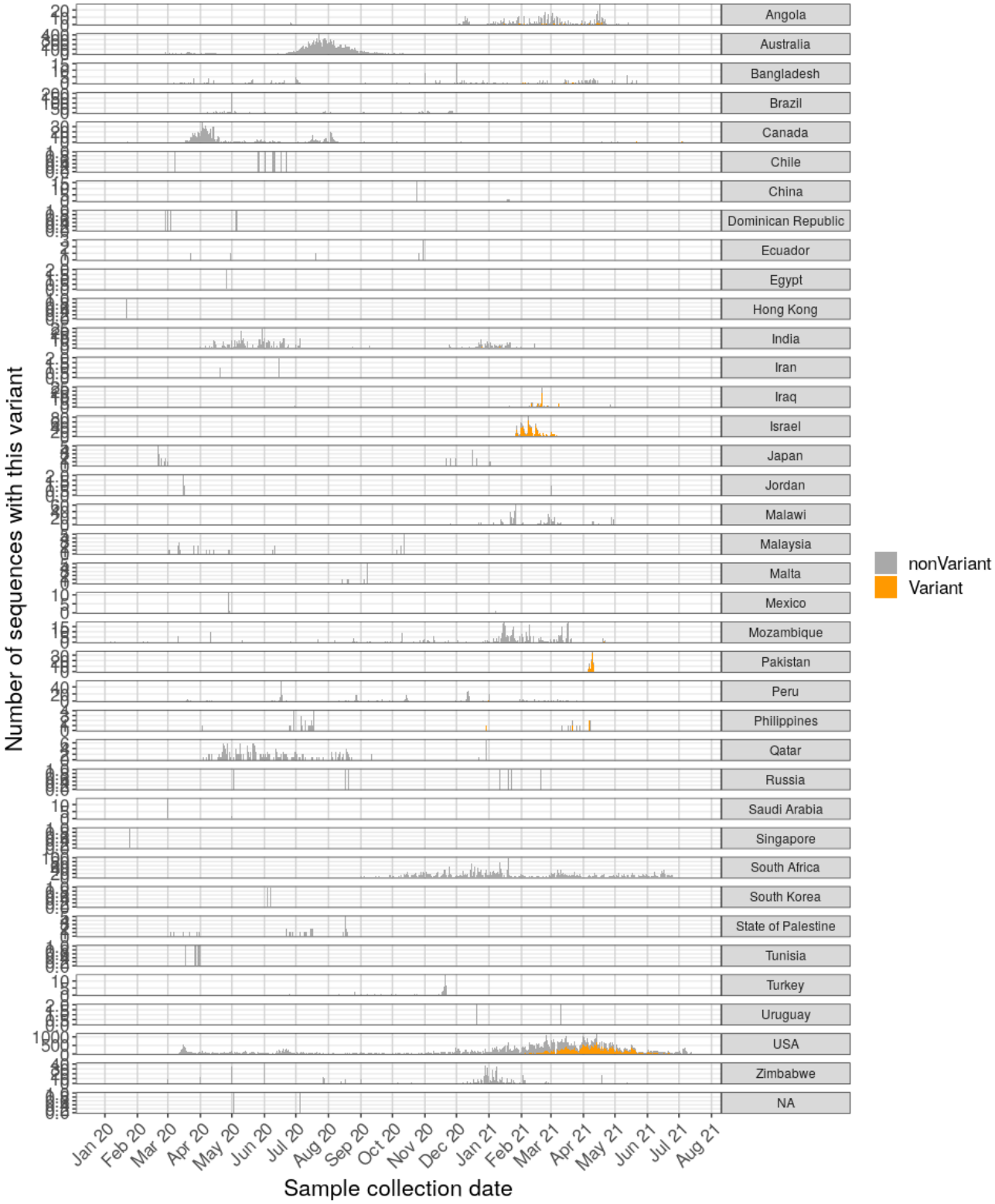


Figure VI: Number of sequences by date of sampling for Alpha variant (orange) for non-European countries.

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

Beta (B.1.351 variant; previously known as the South African variant)

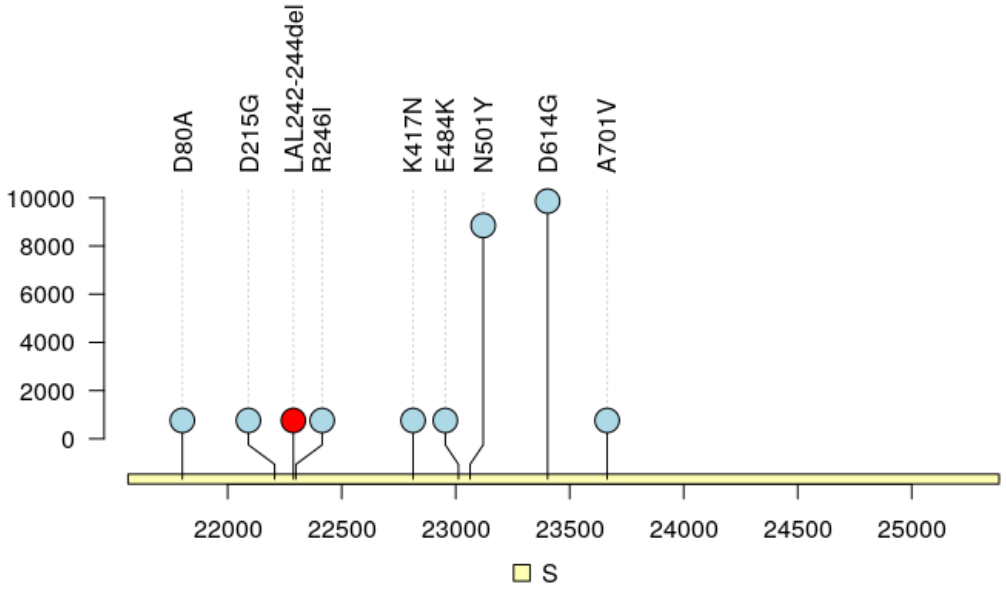


Figure VII: Mutations in the spike protein defining variant Beta. This variant was not detected in the data uploaded since January.

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	814	1063
Netherlands	92	702
Spain	26	1236
Greece	15	59
Ireland	35	81
Estonia	57	37

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

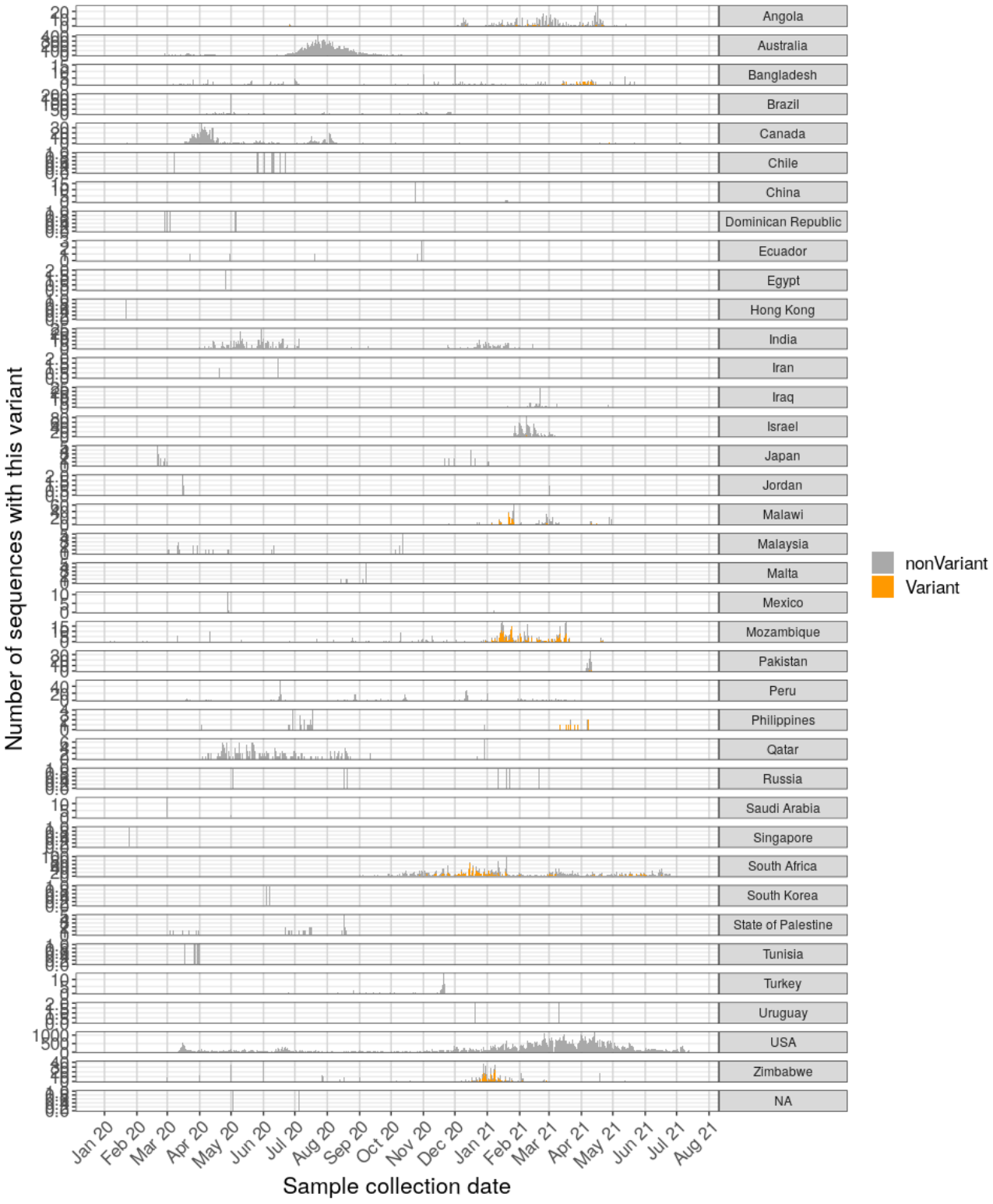


Figure VIII: Number of sequences by date of sampling for variant Beta (orange) for non-European countries.

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

Gamma variant (P1; previously known as the Brazilian variant)

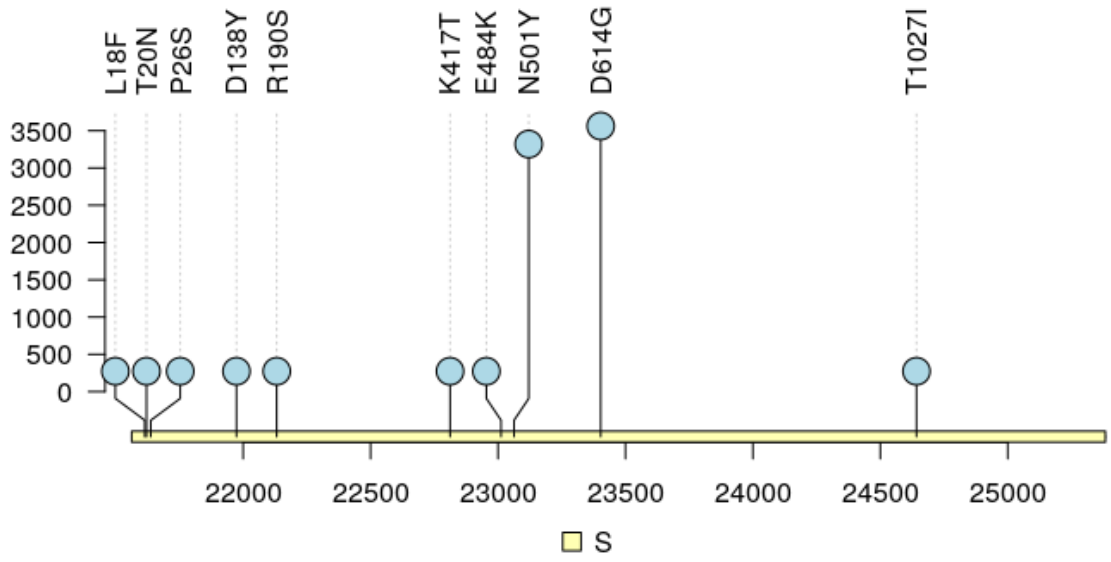


Figure IX: Mutations in the spike protein defining variant Gamma.

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	133	236
Spain	118	1105
Japan	2	122
USA	2160	25634
Italy	3	2545
Netherlands	15	586
Uruguay	1	174

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	34
Bangladesh	1	1
Canada	1	14434

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	9733	313439
Netherlands	5	58233
USA	1785	866009
Ireland	42	35968
Switzerland	4	841
Spain	25	2999
South Africa	195	16327
Canada	1	113675
Angola	4	846
Malawi	1	511

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

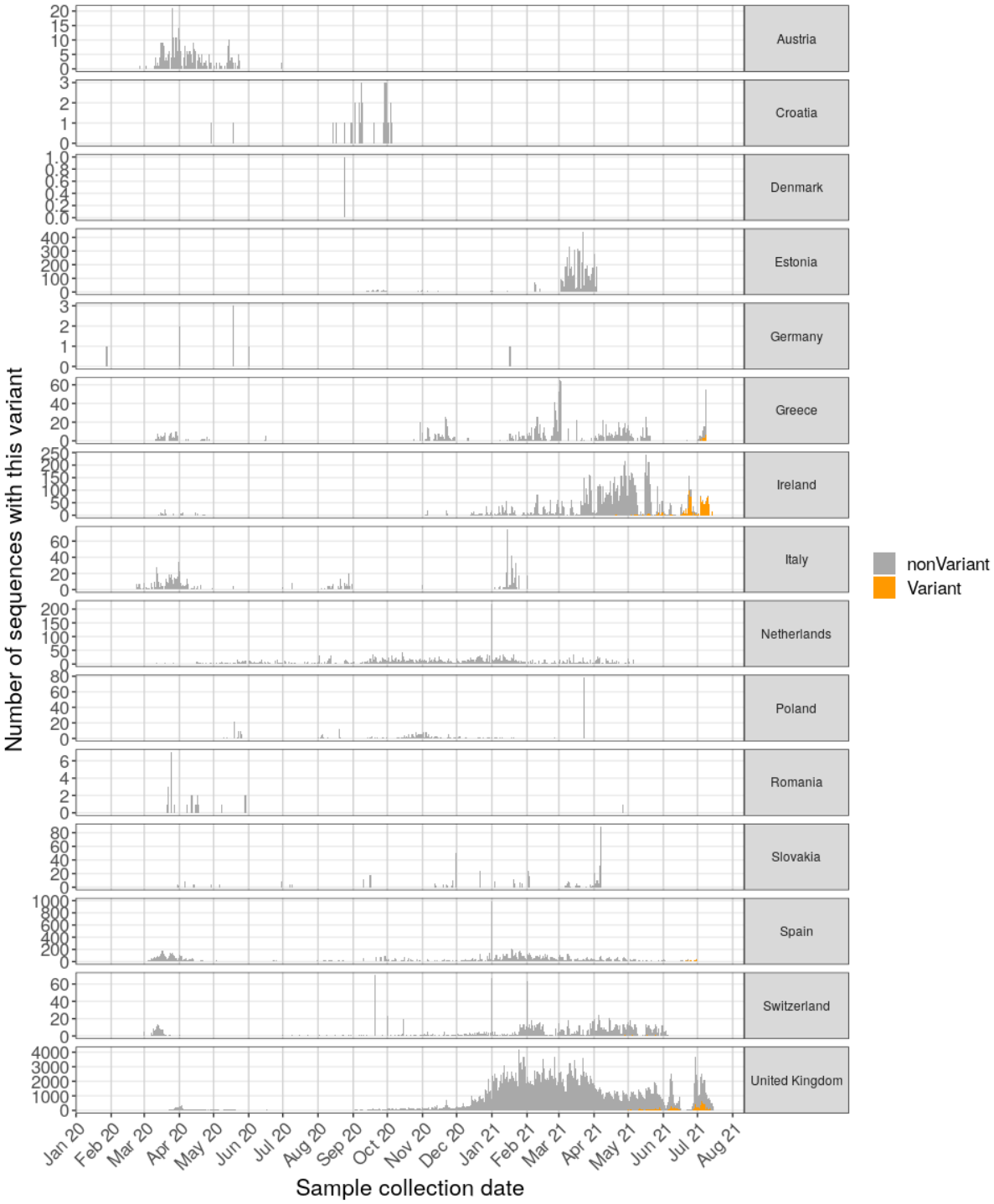


Figure X: Number of sequences by date of sampling for variant Delta variant (orange) for European countries.

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

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.

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

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).