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

Report No. 5 – 23 June 2021

Summary:

- Update on mobilisation of raw reads, now totaling sequencing data sets from 679,693 viral raw read sets from 66 countries, a 23% 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 39,628 samples of the total 93,581 have been processed so far.
- A description of the workflow for generating phylogenetic trees and visualization, as well as how to use this for selecting datasets for individual download, has been included.

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 complement 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 since the previous data freeze (4 May 2021) up to the current data freeze (14 Jun 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		16 Feb. 2021	4 Mar. 2021	25 Mar. 2021	19 Apr. 2021	4 May 2021	14 June 2021
Raw data sets	Total	301,378	354,106	438,112	525,348	552,185	679,693
	Illumina	255,431	302,409	367,462	446,375	469,142	575,481
	Oxford Nanopore	45,222	50,972	69,921	77,913	81,466	93,581
	Other	725	725	729	1,060	1,577	7,134
Source countries for raw data		54	54	58	61	64	66

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

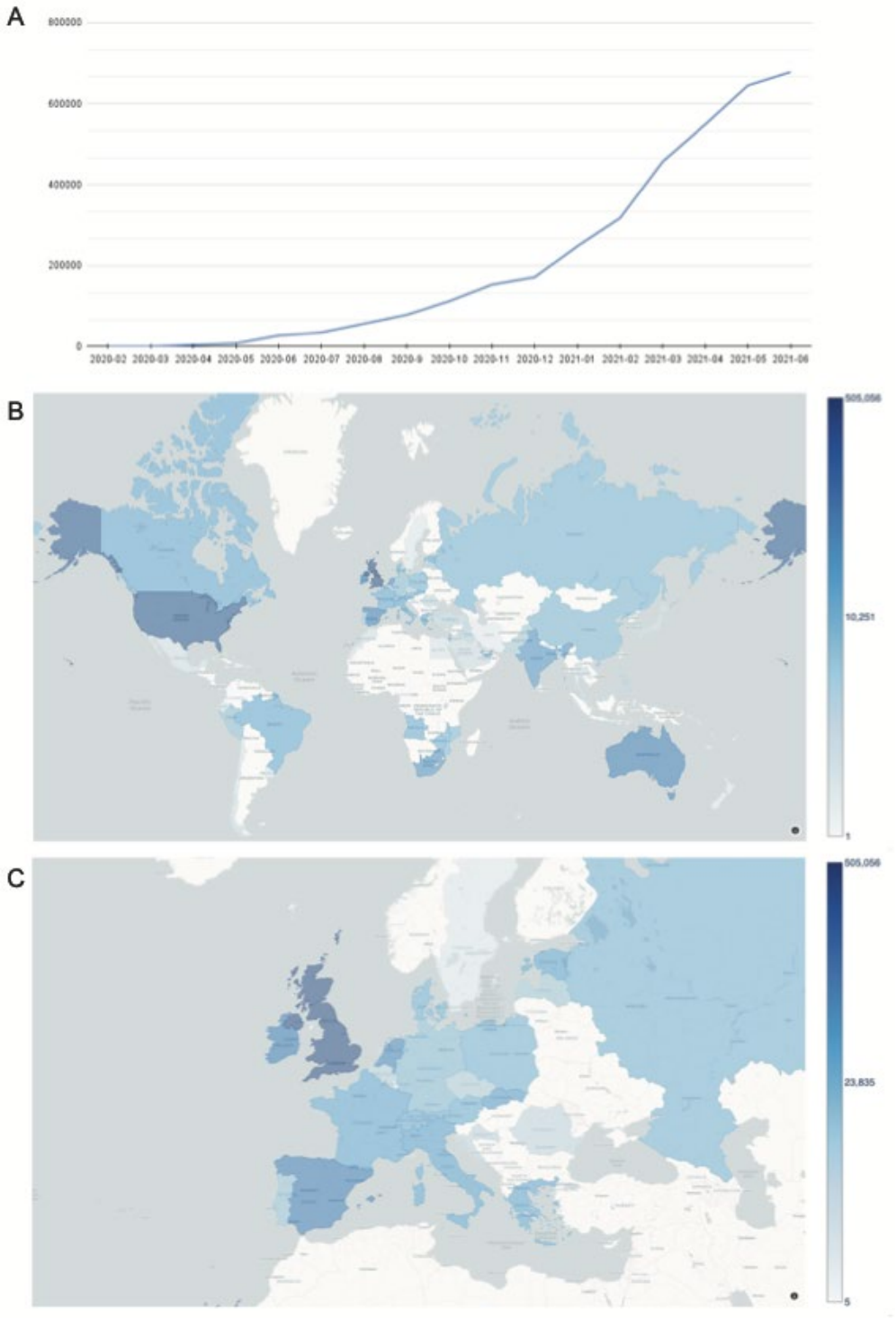


Figure I: Growth of raw SARS-CoV-2 data and distribution of sources, showing (A) sustained growth in raw data since launch of the mobilisation campaign by cumulative number of data sets, (B) and (C) geographical sources of global and European raw data, respectively, for which 81% of global data have been routed through the SARS-CoV-2 Data Hubs, with the remaining 19% 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).

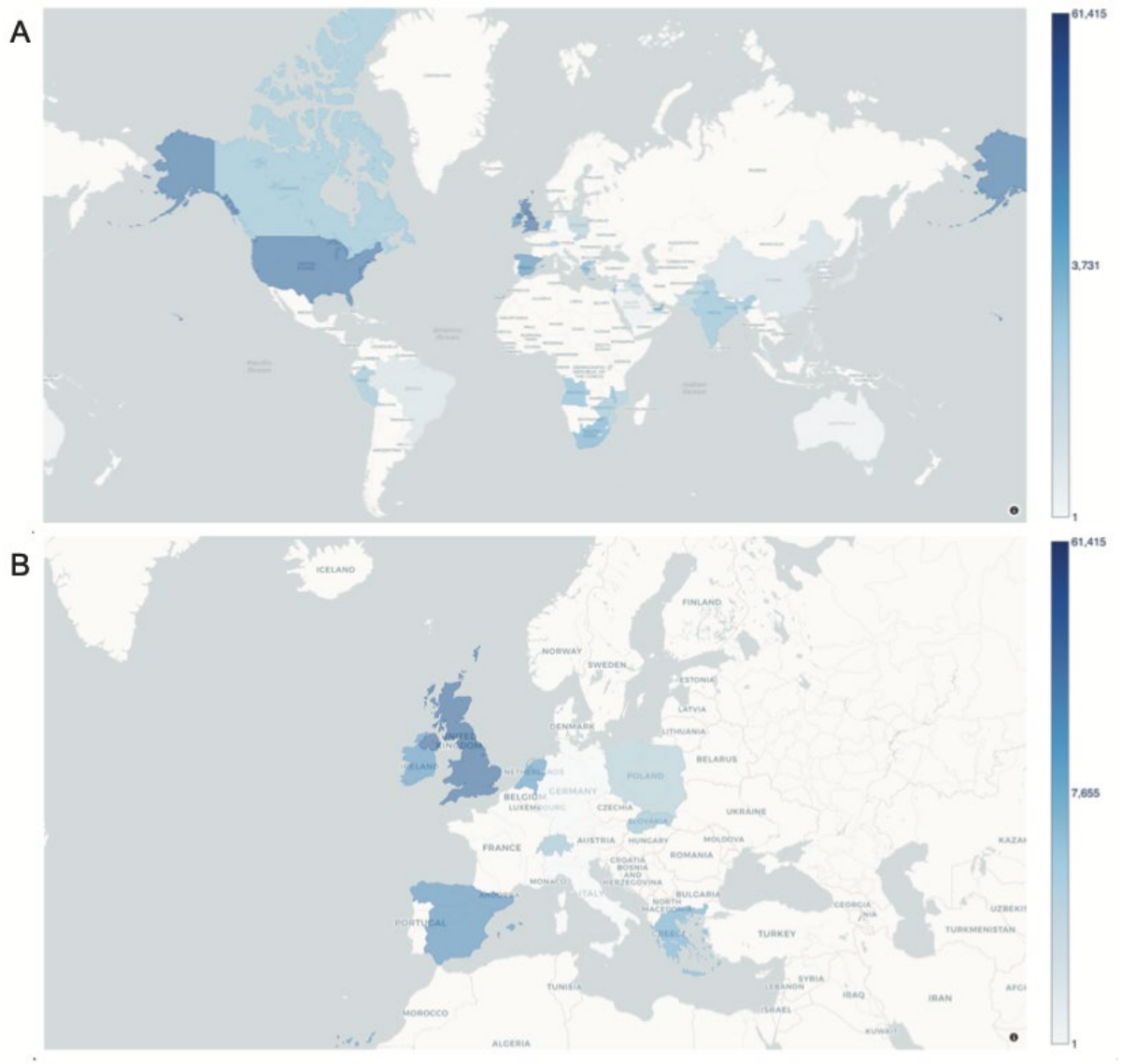


Figure II: New raw SARS-CoV-2 data and distribution of sources at global (A) and European (B) levels mobilised since 4 May 2021. Note that the colour scales are logarithmic best to show the broad range across countries.

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

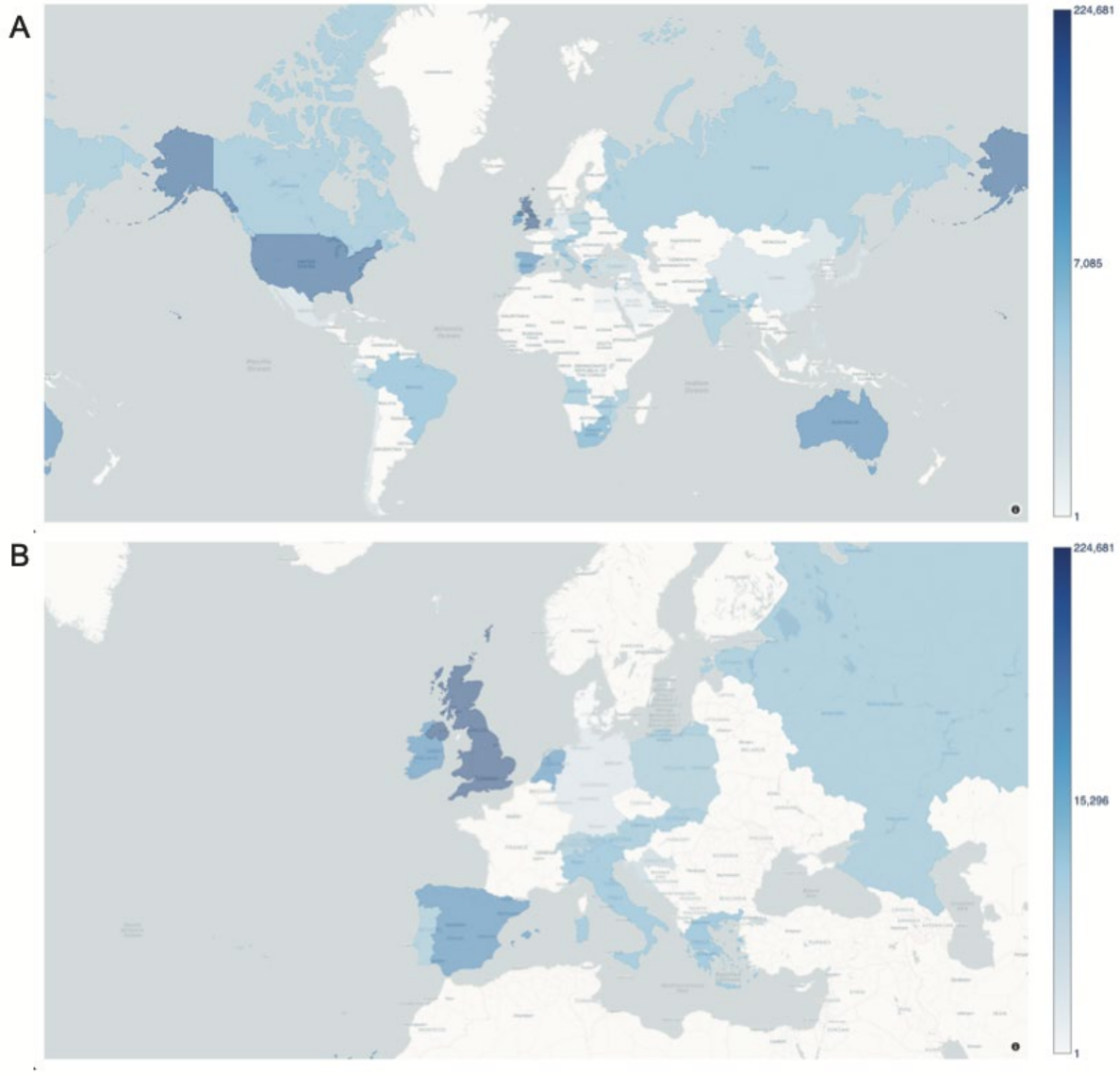


Figure III: Geographical sources of analysed raw data comprising 327,203 data sets spanning the period of data first published from 31 Jul. 2020 to June 14 2021 globally (A) and within Europe (B). Note that the colour scales are logarithmic best to show the broad range across countries.

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 31 Jul. 2020 to 2 Jun. 2021.

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

Mutations and variants

Several variants of concern (VOC) and variants of interest (VOI) have been observed recently. It is important to monitor these variants in time and space and to assess the relevance of these variants. Therefore, a rolling review of literature and reports 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 since 27 May 2021

No new VOCs have been detected since the last update. The Lambda variant (C.37) is added as VOI. As of 18 June 2021, this variant is detected in 26 countries spread over 6 continents. The earliest sequenced samples were reported from Peru in August 2020. Lambda demonstrates increasing prevalence mainly in the South American countries Argentina, Chile and Peru. The Lambda variant contains several mutations in the spike protein that might contribute to reduced susceptibility to neutralizing antibodies and increased transmissibility. Two of these spike mutations, L452Q and F490S, are located in the receptor binding domain (RBD). The mutation F490S is reported to reduce susceptibility against neutralizing antibodies.

The Delta variant has continued to spread globally. There is *in vitro* evidence that this variant replicates more easily in systems that mimic the human airway. Vaccine effectiveness (VE) studies against symptomatic disease and hospitalisation have been performed recently for the Pfizer-BioNtech and Oxford-AstraZeneca vaccines. After one dose, there were significant reductions in VE against symptomatic disease with the Delta variant, but VE against hospitalisation remained high. After two doses, VE against both symptomatic disease and hospitalisation remained high (similar as what was observed for Alpha).

Variants of concern

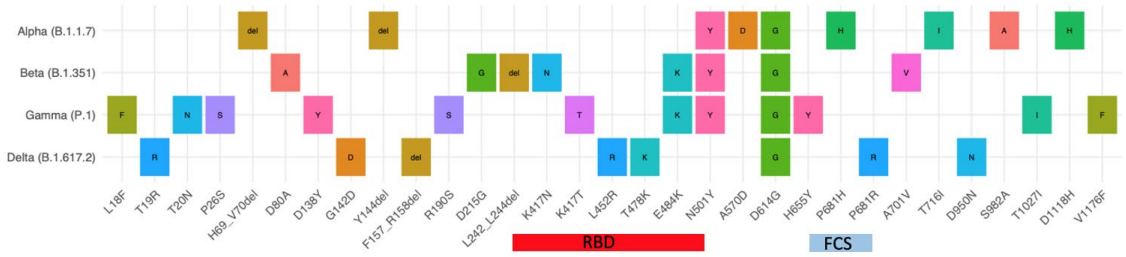
Below is a summary of the analysis of raw read datasets for the presence of the combination of mutations that define the different VOCs.

At the moment, four VOCs have been described: Alpha (B.1.1.7), Beta (B.1.351) , Gamma (B.1.1.28.1 or P1) and the Delta (B.1.617.2) variants. All of these VOCs are defined by a set of mutations and other modifications along the genome and in the spike protein. All variants of concern seem to be more transmissible. 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.

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

In this report, data are presented for the analysis of the presence of variants Alpha (B.1.1.7), Beta (B.1.351), Gamma (P1) and Delta (B.1.617.2).

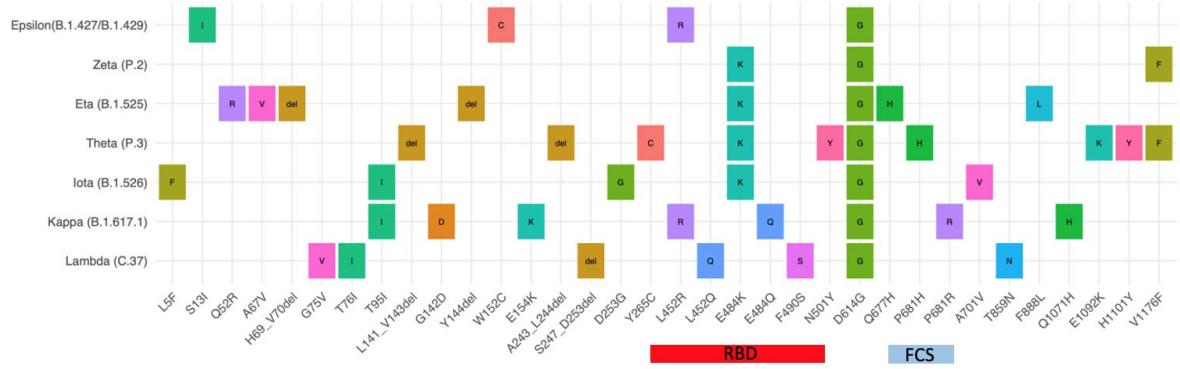
Table II. Overview of the different mutations of several VOCs for the spike gene. Additional mutations are present in other parts of the genome (data not shown). Area in red is the receptor binding domain, and in blue the furin cleavage site.



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, vaccines efficacy and diagnostics is unknown. The 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. The workflow and visualisation tools presented below can be customized for any new combination of mutations and deletions that is considered to be of interest.

Table III. Overview of the different mutations of several VOIs for the spike gene. Additional mutations are present in other parts of the genome (data not shown). Area 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).

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

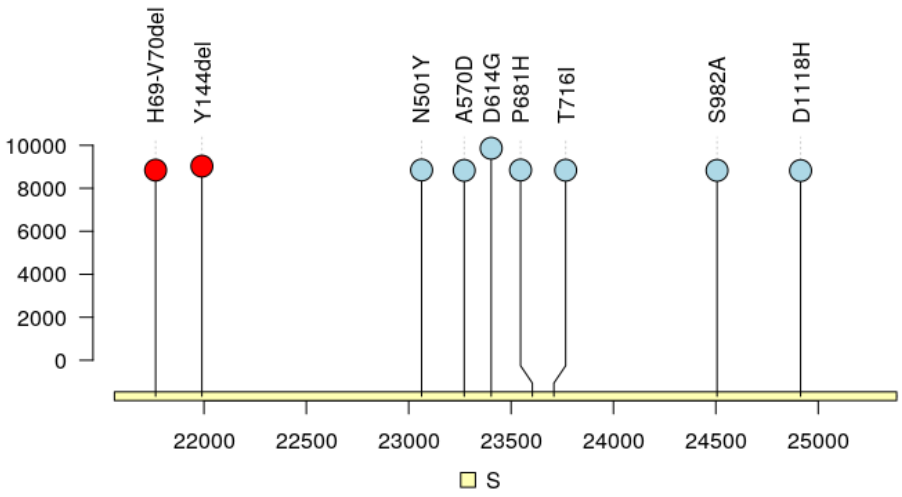


Figure IV: Variant of concern Alpha as defined by the mutations in the spike protein.

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, VIII 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 and VI). 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).

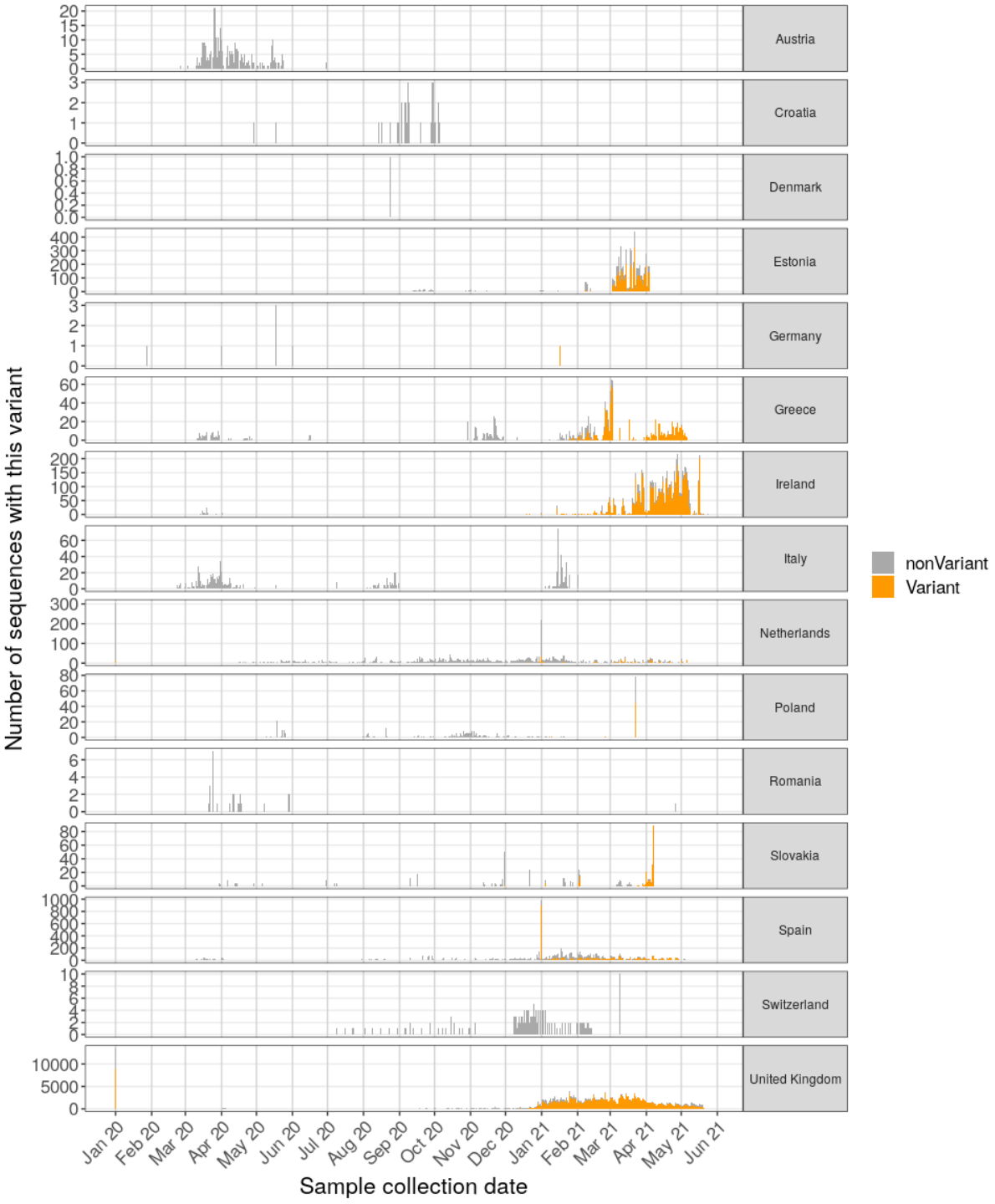


Figure V: Number of sequences by date of sampling for Alpha variant (orange) and non Alpha for countries in Europe and Turkey.

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

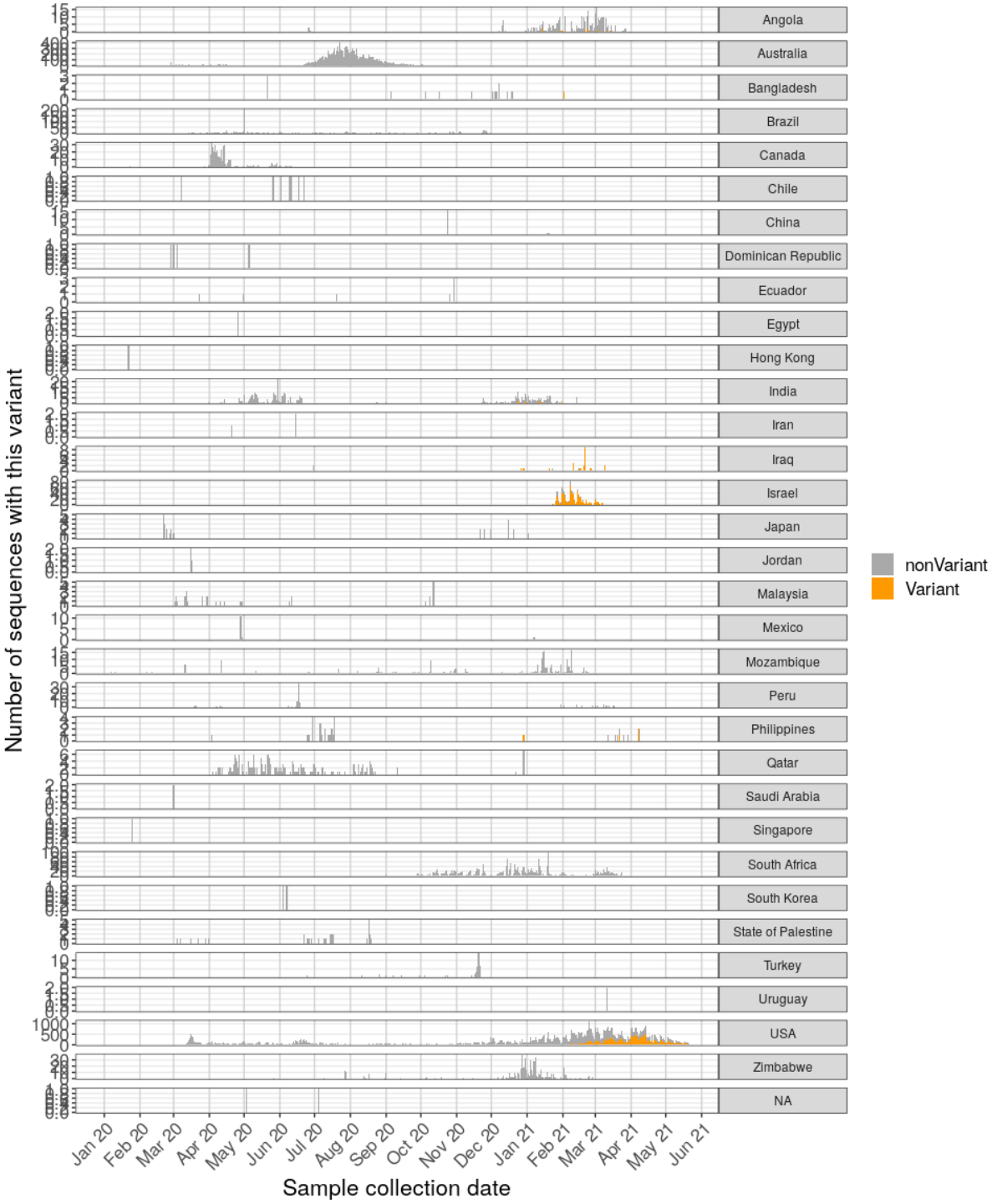


Figure VI: Number of sequences by date of sampling for Alpha variant (orange) for non-European countries.

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

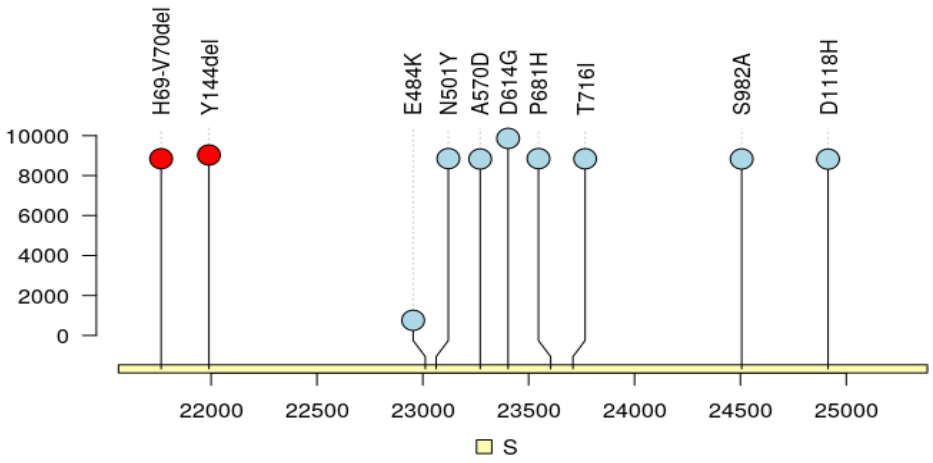


Figure VII: Variant of concern B.1.1.7 + E484K.

Some alpha variant genomes have an additional mutation that has been identified as potential concern, at position E484K. The number of sequences with this additional mutation is limited, as shown in the Table below.

United Kingdom	207
Netherlands	48
Greece	1
USA	23
Spain	5

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)

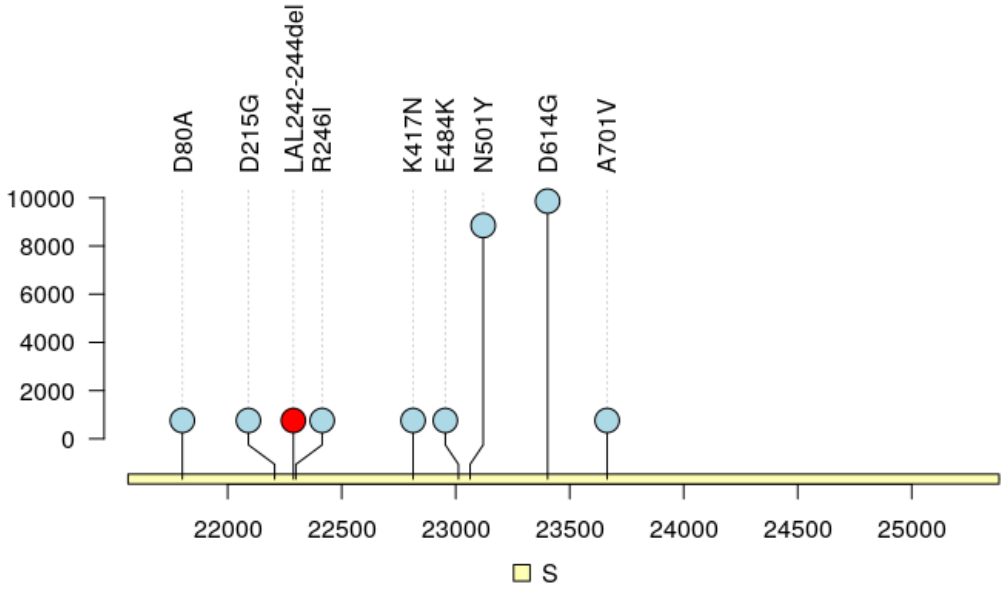


Figure VIII: Mutations in the spike protein defining variant Beta. This variant was not detected in the data uploaded since January.

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

United Kingdom	760
Netherlands	49
Spain	19
Greece	15
Ireland	26
Estonia	57

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

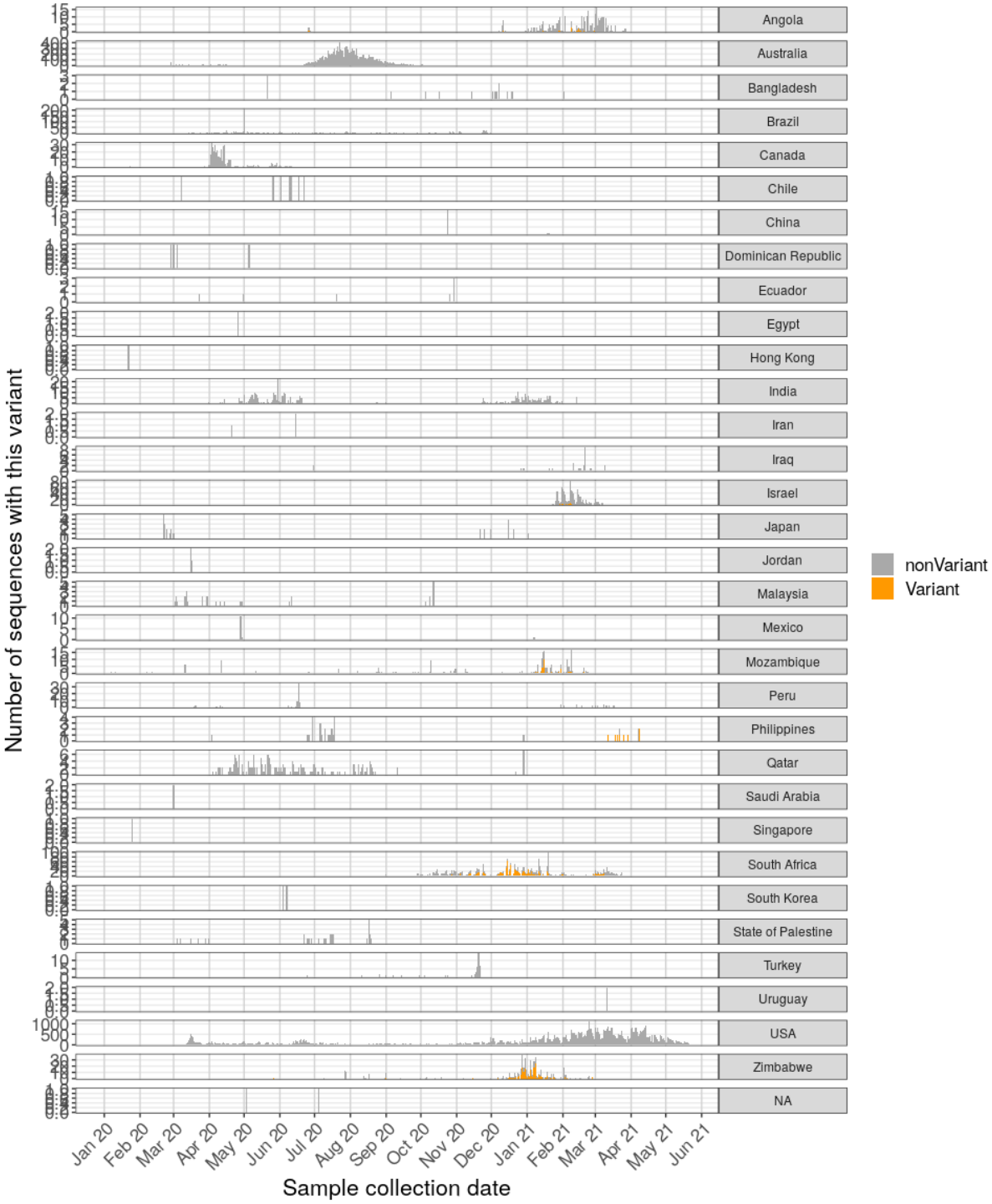


Figure IX: Number of sequences by date of sampling for variant Beta (orange) for non-European countries.

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

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

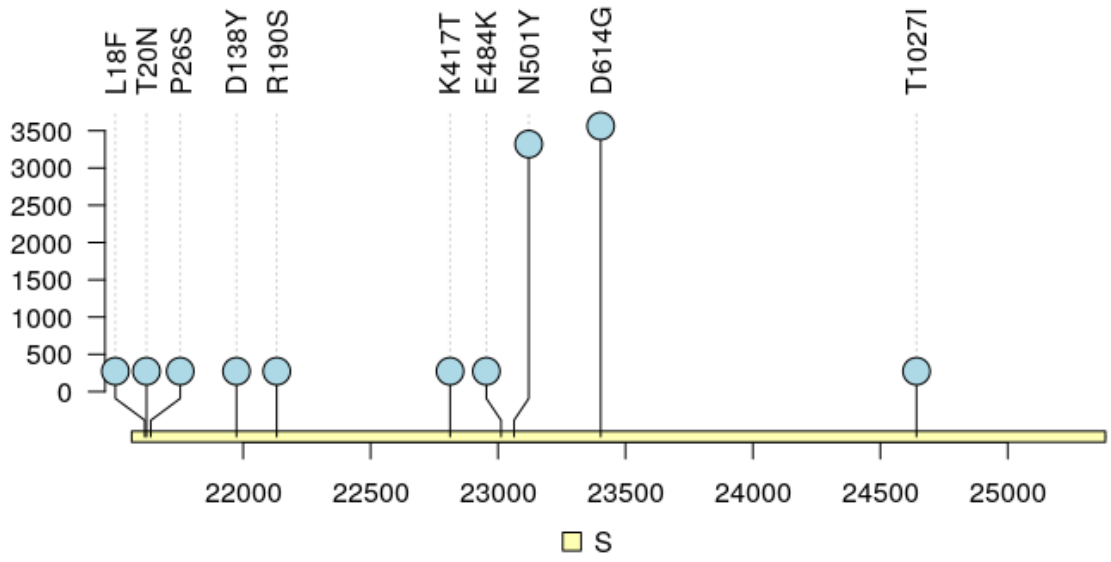


Figure X: Mutations in the spike protein defining variant Gamma.

B.1.1.28.1 was found in countries as listed below. Against the background these numbers are hard to see in the bar chart.

United Kingdom	121
Spain	76
Japan	1
USA	1293
Italy	3
Netherlands	8
Uruguay	1
Ireland	6

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)

Samples containing all Delta variant lineage defining spike protein mutations (T19R, del157/158, 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 do not show up clearly in the bar charts.

United Kingdom	1486
Netherlands	5
USA	45
Ireland	3

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

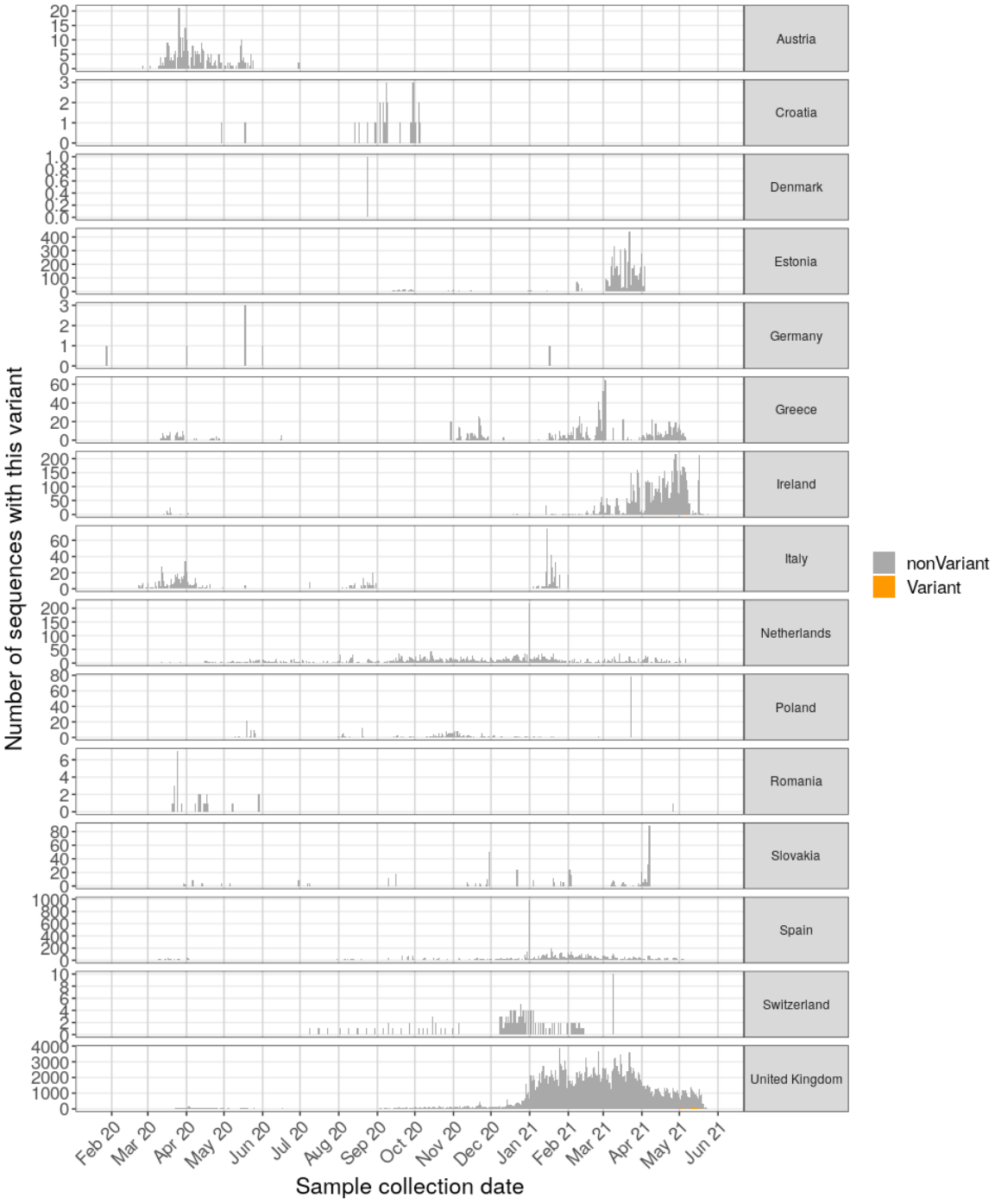


Figure XI: Number of sequences by date of sampling for variant B.1.617.2*(orange) for European countries.*

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

Since the previous report, an additional workflow was added for Oxford Nanopore sequencing. Similar to the Illumina workflow, the Oxford Nanopore sequencing analysis workflow has been updated to produce VCF files via a custom variant calling script. The variants in the 39,628 VCF files are already incorporated in the current report and the remaining 53,953 Oxford Nanopore sequencing datasets will be processed in the coming weeks.

Phylogenetic visualization tool

SARS-CoV-2 is a slowly evolving RNA virus and whole-genome sequencing of its genetic material allows the prediction of the evolutionary path of the virus. An analysis workflow has been developed to infer the genetic relatedness of the submitted consensus sequences. The result of this analysis is supplemented with user-submitted metadata, PANGO-lineage classification and variants detected in the N and S genes of SARS-CoV-2, and displayed in an interactive app in the COVID-19 Data Platform under *Phylogeny*. Through this app, called PhyloDash, the users can search for their sample(s) of interest, or filter on metadata, variant(s) or lineage, and highlight these samples in the phylogenetic tree. By default, the view is of the global tree, but regional trees are available with the drop down menu in the footer of PhyloDash. These are also accessible through the COVID-19 Data Platform: the closest relatives of a given sample in the *Viral sequences* table can be viewed via the button in the *Actions* column. Moreover, it is possible to select and download samples via PhyloDash from the European Nucleotide Archive, facilitating further research on sequences of interest. Further work to improve functioning is ongoing. A help function can be found at the bottom of the metadata section.

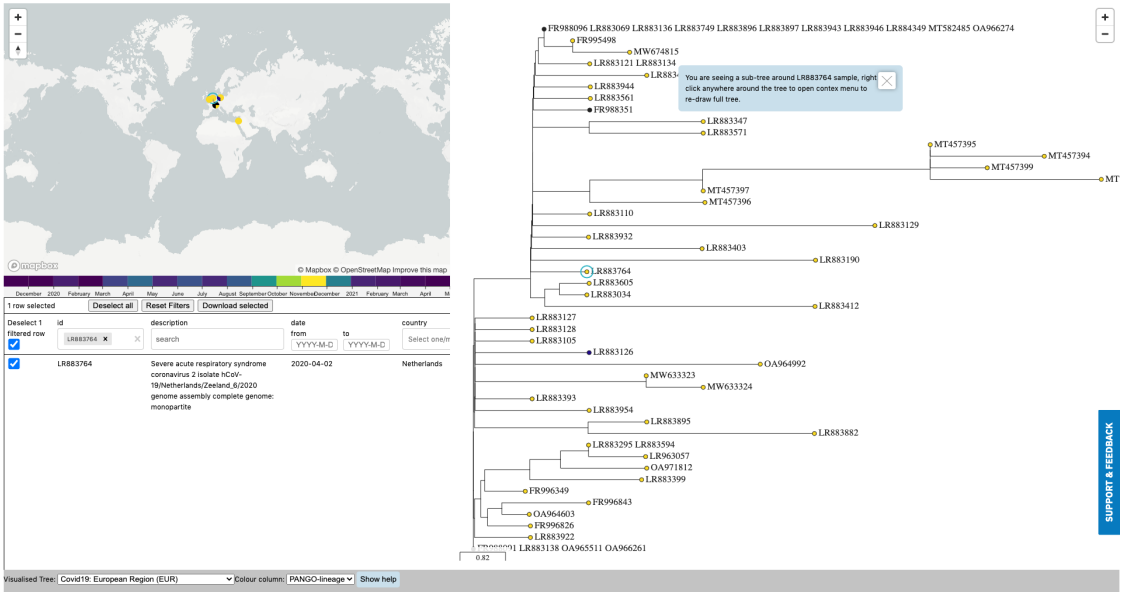


Figure XII: Screenshot from the phylogeny tool of the subtree around sample LR883764 (https://www.covid19dataportal.org/phylogeny-tree?leaf_subtree_id=LR883764&country=Netherlands&no_levels=5&min_leaf_root_length=2)

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

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

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